SESSION

ONTOLOGIES

Chair(s)

TBA

Ontology-centric Source Selection for Meta-querier Customization

Xiao Li, Randy Chow
Department of CISE, University of Florida
Gainesville, FL, 32611, USA
{xl1, chow}@cise.ufl.edu

Abstract

With an increasing number of semi-structured data sources, meta-queriers are introduced to facilitate effective information retrieval from multiple data sources that are accessible through query forms. However, a one-size-fits-all meta-querier cannot cater for various individual needs. In meta-querier customization, source selection is arguably one of the most critical problems. This paper proposes a capability-based source selection to meet user needs in terms of query capabilities. The major challenges include modeling, understanding and matching of the user needs and source capabilities. Our solution is based on a light-weight ontology, M-Ontology, which is generated from a number of verified mappings between heterogeneous query forms of the data sources. With the assistance of the concepts and relations in M-Ontology, user demands and source capabilities are modeled as concept sets, identified through query-form annotation, and matched by an additive utility function. The experiments on real-world data illustrate the potential of this ontology-centric method.

1 Introduction

As an increasing number of semi-structured data sources are available online through HTML query forms [4][7][17], integration of data sources is desirable for improving the efficiency of information retrieval. Meta-queriers are virtual data integration systems that shield users from data heterogeneity and source location. They provide the user with a uniform query form (a.k.a. *global form*) for simultaneously accessing a set of disparate data sources in the same domain. The user does not need to input repetitive information to each source query form (a.k.a. *local form*). Based on the *mappings* between global and local forms, user queries over the global form are respectively reformulated to the queries in terms of the local forms, and then the query results from data sources are presented to the user in an integrated format.

There is a wide divergence in the data sources (e.g., in terms of query capabilities, content qualities and site credibility). A one-size-fits-all meta-querier cannot cater for individual needs [28], even in the same application domain. User-driven selection on data sources is a convenient and straightforward method to customize meta-queriers. From the viewpoints of the user, the contents of global query forms and the selection of data sources are the most critical (or perhaps the only) factors that influence the contents retrieved from the meta-queriers. First, source selection determines the content coverage of meta-queriers. The meta-queriers are virtual data integration systems [13] that do not physically store any information. That is, the returned contents are completely determined by the underlying data sources. Second, modifying the controls in global forms is the only way for the user to express their demands on the results. All the returned contents should conform to the user constraints set by modifications of the *controls* (e.g., clicking radio buttons, dropping down menus, and entering texts). In a sense, the contents of global forms are also decided by the selection of data sources, since the global forms should only consists of the functionalities that are supported by every underlying data sources; otherwise, the results might include some/many records that violate the original user-specified conditions. Therefore, source selection is arguably one of the most critical problems in meta-querier customization.

This paper investigates how to exploit the query capabilities [14][6][25] of semi-structured sources for achieving more accurate selection. We propose an ontology-centric approach to source selection. A domain-based ontology (referred to as *M-Ontology*) was designed for meta-querier customization. In meta-querier customization, various customized meta-queriers are constructed, and thus a potentially large number of mappings between global and local forms need to be stored, managed and discovered. These unordered mappings are organized based on their semantics to form the concepts and relations of M-Ontology. This paper utilizes these concepts and relations to model source capabilities and user demands. Our major contributions can be summarized into the following three aspects:

- *Capability modeling and capture*: Without adequately accurate understanding of source capabilities, the selection in the previous research [10][21][16] is normally coarse-grained and unable to distinguish the functionality difference of the sources. In this paper, we view M-Ontology as a domain-specific capability repository. For each integrated data source, its capabilities can be automatically identified through the association of its query form with the concepts in M-Ontology.

• *Demand modeling and elicitation*: Modeling of user demands is still an open problem in the selection of semi-structured sources. We model the demands by a preference vector, in which each entry corresponds to a concept in M-Ontology. A semi-automatic solution to demand elicitation is also proposed through semantic annotation on the query forms of user-preferred data sources.

• *Demand matching*: The desirability of a specific data source for a particular user needs is quantified by a matching value. The value calculation is treated as a multi-criteria decision making problem. Each criterion corresponds to the desirability of a specific capability. An additive utility function is proposed to combine all the criteria, each of which is calculated on the basis of the similarity of the user's preference vector with the source's capability set.

In the remainder of this paper, we first review the related work on source selection in distributed information retrieval systems. Section 3 introduces the algorithm for capturing user demands and discovering the appropriate resources for the identified user demands. Section 4 details the experimental results on real-world data. Finally, we conclude with directions for future research.

2 Related Work

To distinguish our work from the current solutions to source selection, we discuss the solutions based on source types and selection mechanisms:

• Unstructured and (semi-)structured data sources.

In the field of distributed information retrieval systems, the prior researches on source selection mainly focus on the selection of sources containing unstructured data (a.k.a., texts). To acquire the contents of data sources, randomly generated queries are sent to obtain the sample texts. Through these fetched samples, the data sources can be represented as a single big document (such as in CORI[5], CVV[26] and KL[24]) or a set of big documents (such as, in ReDDE[20], CRCS[19] and SUSHI[22]).

With the rapid growth of the deep Web, (semi-)structured information sources have been experiencing a remarkable increase. Normally, the query interfaces of structured information sources are more complex than those of text sources. First, the query-based sampling becomes impractical since it is difficult to retrieve the sample documents through randomly generated queries. The automatically generated queriers usually cannot satisfy the hidden constraints on the inputs to the query forms. Second, the topics of sources can be directly acquired through the semantics analysis on the query forms. More complex query forms often contain more semantics at the same time. Several researches [10][21][16] have been conducted on *cross-domain source selection*. They cluster structured Web sources based on the query-form similarity so that each cluster corresponds to a single domain. However, the domain-oriented clustering is a coarse classification without considering the capability distinction.

• Selection-per-query and selection-per-engine

Source selection is one of the major research issues in meta-querier customization. The selection can be made when users issue a query (called *selection-per-query*) or when the meta-queriers are constructed (called *selection-per-engine*).

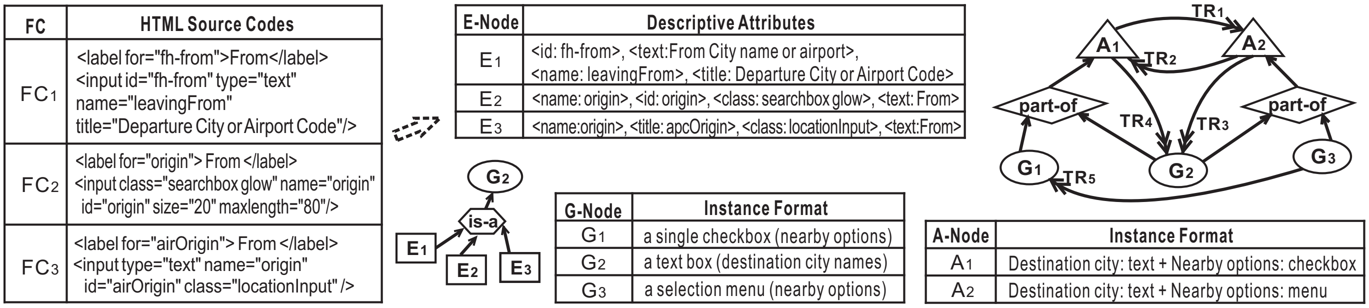
In some domains, the query forms are considerably simple. For example, most news search engines only include a single keyword box and a click button. Thus, the construction of a global query form is not difficult to integrate all the forms in the same domain. In this context, the data sources can be selected based on the user inputs to the global interface (i.e., selection-per-query).

However, in the other domains, it is not practical to build such a single interface (or even a few) to encompass all the functionalities provided by the query interfaces in the same domain. For instance, in the air-ticket booking domain, we can observe that various meta-queriers provide different query interfaces with different functionalities. Most differences are caused by the selection of sources in their construction (i.e., selection-per-engine).

Our work proposes a selection-per-engine strategy in the customization of meta-queriers. Users are allowed to input their preferred data sources. To better reuse the pre-integrated data sources, we provide a capability-based source selection algorithm to recommend the users their potentially desired data sources. Complimentary to the query-based solutions, our approach is based on the query capabilities of data sources, instead of the sampled source contents. Unlike the prior work on query-form clustering, our approach is able to distinguish the query capabilities of the data sources whose query forms have been clustered in the same domain.

3 Capability-based Recommendation

Capability-based source selection is a content-based recommendation problem [3]. In our capability-based recommendation, users can declare their needs of the capabilities by inputting the domains (DM_I) and their preferred data sources (DS_I). Based on the user needs, our system recommends the users a ranked list of the pre-integrated sources (DS_O) from a repository (SR) of data sources. Such a capability-based recommendation problem can be simplified to a *utility maximization* problem. The effectiveness of such a solution relies on the correct understanding of the user demands and data sources. Section 3.1 first explains our ontology-centric model for source capabilities and user preferences, and then Section 3.2 presents the corresponding source-selection algorithms.



FC	HTML Source Codes

FC_1	<label for="fh-from">From</label>
	<input id="fh-from" type="text"
	name="leavingFrom"
	title="Departure City or Airport Code"/>

FC_2	<label for="origin"> From </label> <input class="searchbox glow" name="origin" id="origin" size="20" maxlength="80"/>
FC_3	<label for="airOrigin"> From </label> <input type="text" name="origin" id="airOrigin" class="locationInput" />

E-Node	Descriptive Attributes
E_1	<id: fh-from>, <text:From City name or airport>, <name: leavingFrom>, <title: Departure City or Airport Code>
E_2	<name: origin>, <id: origin>, <class: searchbox glow>, <text: From>
E_3	<name:origin>, <title: apcOrigin>, <class: locationInput>, <text:From>

G-Node	Instance Format
G_1	a single checkbox (nearby options)
G_2	a text box (destination city names)
G_3	a selection menu (nearby options)

A-Node	Instance Format
A_1	Destination city: text + Nearby options: checkbox
A_2	Destination city: text + Nearby options: menu

Figure 1. A fragment of M-Ontology for air-ticket booking

3.1 Modeling

In the context of meta-querier customization, query capabilities refer to the abstract abilities of data sources to retrieve information. The information of these sources can be accessed through their associated query forms. Generally, the contents of query forms decide the possible queries that can be posed by users, and thus they also dictate the capabilities of the meta-queriers and data sources. Understanding the query forms is a cornerstone of the capability-based recommendation. This paper proposes an ontology-centric approach to represent their capabilities by analyzing and understanding the query forms.

Each query form can be regarded as a set of query conditions [27][11], referred to as *functional components*. For HTML forms, each component consists of a control and its associated attributes. The attributes include control type, name/label, descriptive text, instances, data domain, default value, scale/unit (e.g., kg, million, dollar), and data/value types (e.g., date type, time format, char type, etc.).

An individual component or an ordered component list can represent a specific query capability [14][6][25] that a resource possesses. However, the information carried in individual components is very limited. It does not contain the context or knowledge concerning the textual description. For example, Fig.1 shows three functional components, FC_1, FC_2 and FC_3, all of which are extracted from the real-world query forms. Each represents the departure city for booking air-tickets. However, it is difficult for machines to find the capability equivalence. Such naming conflict is very common among the query forms, which are normally created and maintained by different companies and organizations.

M-Ontology: Methodologies of ontology are commonly used to address such representation heterogeneity. Ontologies store well-defined concepts and relations including context knowledge. If query forms are properly annotated using concepts in the same ontology, machines can understand their semantics. In our previous research [15], we designed a domain-specific and light-weight ontologies (named *M-Ontology*). M-Ontology is designed for storage, management and discovery of mappings that are employed to translate query inputs among query forms. We proposed a semantics-based approach to model and organize these mappings, which can be numerous in the context of meta-querier customization due to various user needs and data sources. Each mapping corresponds to a tuple $\langle List_{FC1}, List_{FC2}, Exp\rangle$, where $List_{FC1}$ and $List_{FC2}$ are respectively two ordered lists of functional components, whose potential user inputs can be transformed through the rule Exp. In the real-world scenarios, for a specific mapping, $List_{FC1}$, $List_{FC2}$ or both might include more than a functional component. In M-Ontology, these components (as a whole) are viewed as a single concept, each of which also corresponds to a single concept. Each mapping can be regarded as an instance of a relation between two concepts. We proposed a semi-automatic solution [15] to M-Ontology construction by incremental insertion of the mappings.

As a language-independent ontology, M-Ontology is modeled by a directed acyclic graph, where nodes are the concepts and edges are the relations. In the following, we briefly introduce the nodes/edges and their correspondences with query capabilities.

(1) *E-Nodes, Is-a Edges and G-Nodes*. Each E-Node encapsulates a functional component in a specific query form. For example, as shown in Fig.1, E_1 corresponds to FC_1. Through generalization of a set of E-Nodes (e.g., E_1, E_2 and E_3) that have the same semantics and instance formats, an Is-a Edge can formulate a new concept node, called a G-Node (e.g., G_2). In essence, each G-Node corresponds to an abstract query capability. For describing the semantics of such a query capability, a representative object ro is automatically abstracted from the associated attributes of the inclusive functional components as follows: i) two bags of descriptive words $Set\langle t_i^{DA}\rangle$, $Set\langle t_i^{IST}\rangle$ are generated respectively from the descriptive attributes and instances of all human-verified E-Nodes in gn, which are normalized using NLP techniques [9] such as tokenization, stop-word removal and stemming; ii) a set of descriptive labels $Set\langle t_i^{DL}\rangle$ is determined by selecting the terms with the top-k frequency weight from $Set\langle t_i^{DA}\rangle$. The descriptive attributes and labels can be manually modified by humans. Finally, we generate the representative object ro with a tuple $\langle Set\langle t_i^{DA}\rangle, Set\langle t_i^{DL}\rangle, Set\langle t_i^{INS}\rangle\rangle$.

(2) *A-Nodes and Part-of Edges*. Each A-Node is formed through aggregating a list of G-Nodes. A Part-of Edge is used to represent such an aggregation relation by linking the A-Node (e.g., A_1) to its inclusive G-Nodes (G_1 and G_2). The

A-Node also denotes an abstract query capability, whose semantics can be approximated to a list of representative objects $List_{ro}$.

(3) *T-Edges*. A T-Edge corresponds to a transformation relation between two concept nodes (i.e., G/A-Nodes) which can be used to fetch similar contents. For example, in Fig.1, TR_1 is a T-Edge representing a transformation relation from A_1 to A_2. The transformation relation contains a specific rule to convert the instances of the connected concept nodes. Since their instances are convertible, all these connected nodes represent similar query capabilities. In a sense, M-Ontology is a domain-specific query capability repository, where each connected G/A-Node sub-graph generally corresponds to an abstracted query capability.

Capability modeling: By using the proposed M-Ontology, we model the capabilities of data sources based on their own query forms. Each data source has its own query form, whose inclusive functional components can be clustered into a set of G-Nodes in M-Ontology. Let M-Ontology includes a set of G-Nodes denoted by $GN = \{N_1, N_2, ..., N_n\}$, which are numbered from 1 to n based on their creation time. Since each G-Node denotes an abstract capability in a specific domain, we use a subset of GN (called Capability Set CS) to represent the capabilities that a data source is able to provide. That is, CS can be denoted by a G-Node set $\{N_i | N_i \in GN\}$.

To obtain such a capability set, we need to seek correct G-Nodes to annotate the functional components in the corresponding query form. More precisely, the process of capability capture can be viewed as *schema annotation*. In the context of meta-querier customization, it is straightforward to capture the capabilities of the pre-integrated data sources from their query forms. M-Ontology functions as a mapping repository, and thus the mappings linked to these query forms should have been inserted into M-Ontology. That means, all the functional components have been clustered into the corresponding G-Nodes. These G-Nodes are the components of the corresponding capability sets. The detailed algorithms are presented in the Section 3.2.

Preference modeling: The widely used preference model is based on keywords. However, in the real-world scenarios, a few keywords are often unable to represent the exact semantics. For the purpose of accurately understanding user demands, our solution relies on the mapping-generated M-Ontology. In our solution, a specific user need is modeled by a vector PV, called a *preference vector*. Given that M-Ontology contains G-Nodes $\{N_1, N_2, ..., N_n\}$, the vector PV has n corresponding entries. The i^{th} entry of vector PV, denoted by $PV(i)$, is a preference value that indicates how the user prefers this capability in the target.

3.2 Demand Capture and Matching

Following the proposed capability model and preference model, this section presents an ontology-centric algorithm for source selection. As illustrated in Fig.2, the whole procedure consists of six phases. Based on user selection (i.e., DM_I) from a list of data domains that exist in the system, *Ontology Selection* chooses an appropriate M-Ontology MO for understanding the user-preferred data sources (i.e., DS_I). Only for the data sources that have not been integrated into any meta-querier, *Q-Form Normalization* is invoked to unify their query form representation. By analyzing the normalized query forms, *Q-Form Analysis* can output the capability set CS for each data source. From the analyzed query forms, *Demand Identification* constructs a preference vector PV, while users are able to correct the preference values. *Demand Matching* generates a ranked list of resources for user selection. After user selection, if necessary, *Annotation Verification* verifies the correctness of analysis on query forms of un-integrated data sources.

The details are described as follows.

• **Ontology Selection** is to choose an appropriate M-Ontology MO. In different query forms, there might exist many functional components with the highly similar representation, but, in fact, their semantics are completely different due to the different context. For example, the word “keywords”, which appears in the different domain, has various meanings. In job seeking, it represents job titles or fields, but it also might denote the song name in the music search. Thus, our designed M-Ontology is domain-specific. This assumption is reasonable for the customization of meta-queriers, which combine the data sources in the same domain.

• **Query-form Normalization** is to unify the representation of the query form for each data source in DS_I that is not pre-integrated into our system. The data sources can be categorized as two types, pre-integrated and new data sources. Since the pre-integrated data sources have been analyzed, the results can be directly reused without repetitive normalization. This phase only processes the query forms $qform_I$ that are not included in M-Ontology. For each query form, its inclusive functional components are associated with some descriptive attributes and a set of potential instances (if they exist). We first perform natural language processing techniques on the descriptive attributes and instances in the following orders: tokenization, stop-word removal and stemming. Then, we treat the normalized texts as two unordered bags of words, $Set\langle w_i^{DA}\rangle$ and $Set\langle w_i^{INS}\rangle$. These two bags of words constitute a word-bag pair, which can indicate the semantic characteristics of this component. Fi-

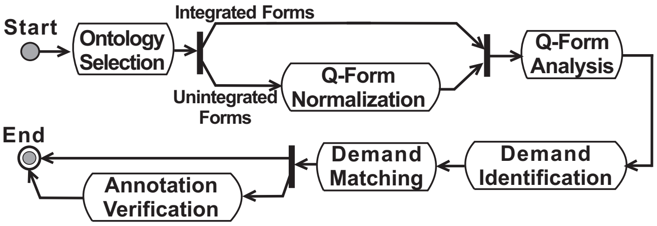


Figure 2. The ontology-centric algorithm for source selection.

nally, each query form corresponds to a set of word-bag pairs, $Set\langle(Set\langle w_i^{DA}\rangle, Set\langle w_i^{INS}\rangle)\rangle$

• **Query-form Analysis** is to understand the query forms by analyzing the semantics of each inclusive functional component fc. Our solution is to annotate each fc by semantics-equivalent concepts (i.e., G-Nodes from the M-Ontology MO). The involved G-Nodes constitute the capability set CS of the corresponding data source. G-Node searching can be divided into two separate types: a) For the pre-integrated data sources, their query forms have been included in the M-Ontology, and thus each functional component of these forms should have been clustered to a certain G-Node. That is, such an association can be directly reused. b) For the new data sources, each functional component in their query forms is normalized to two word bags $Set\langle w_i^{DA}\rangle$ and $Set\langle w_i^{INS}\rangle$. The suitability of a G-Node gn can be measured by the constraint equivalence and semantic similarity between fc and the corresponding representative object ro of gn. The semantic similarity can be calculated as follows.

$$\lambda_1 \sum_{i=0}^{n_1}\sum_{j=0}^{m_1} sim(w_i^{DA}, t_j^{DA}) + \lambda_2 \sum_{i=0}^{n_1}\sum_{j=0}^{m_2} sim(w_i^{DA}, t_j^{DL})$$
$$+ \lambda_3 \sum_{i=0}^{n_3}\sum_{j=0}^{m_3} sim(w_i^{INS}, t_j^{INS})$$

where, λ_i is a scale factor, two word sets $Set\langle w_i^{DA}\rangle$ and $Set\langle w_i^{INS}\rangle$ are the semantic characteristics of fc, and a tuple $\langle Set\langle t_i^{DA}\rangle, Set\langle t_i^{DL}\rangle, Set\langle t_i^{INS}\rangle\rangle$ is the representative object ro. The function sim is to determine the semantic similarity between two terms (respectively from fc and ro). Our implementation relies on the WordNet-synonyms distance, a linguistic-based matcher.

• **Demand Identification** constructs a preference vector PV based on the selected data sources and user interaction. The whole procedure is composed of two steps:

1) *Automatic discovery*: Each G-Node corresponds to an entry $PV(i)$ whose value indicates the preference degree against a specific capability. In M-Ontology, G-Nodes that are connected via a single or multiple T-Edges and Part-of Edges whose directions are ignored constitute a maximal connected G-Node sub-graph. Such a sub-graph corresponds to an abstracted query capability. In the sub-graph, the preference values of these G-Nodes are correlated. Since the conversion through T-Edges and Part-of Edges might lose the semantics, the distance between two G-Nodes i and j indicates their dissimilarity. Here, the distance refers to the minimum hop number from one node to another. The value $1/(1 + distance(i, j))$ is used to represent the potential difference between their capabilities. Assuming that $GSet_M$ is a multiset that contains the G-Nodes encapsulating the functional components of data source in DS_I. It might contain duplicates. The occurrence number of a G-Node in $GSet_M$ is equal to the appearance frequencies of its encapsulated functional components in DS_I. Preference vectors can be decided by two modes: combination and accumulation.

- The *combination mode* is preferred when users want to find the data sources containing all the capabilities (with the same desirability) that are supported by the user-inputted sources DS_I. The values can be obtained through the following procedure: First, all the entries whose G-Nodes are in $GSet_M$ are set to one, i.e., $PV(i)$ = 1 if $i \in GSet_M$; otherwise, the values are initialized to zero. Second, for the G-Nodes that are not in $GSet_M$, their values are calculated based on the distance to the nearest node in $GSet_M$. The procedure can be represented as eq. (1).

$$PV(i) = \begin{cases} 1 & i \in GSet_M \\ \frac{1}{\min\limits_{j\in GSet_M}\{distance(i,j)+1\}} & i \notin GSet_M \end{cases} \quad (1)$$

- The *accumulation mode* assumes the most critical capabilities that user desired are the ones that appear most frequently in the user-inputted sources. For the entries whose G-Nodes are in $GSet_M$, their values of $PV(i)$ are equal to the multiplicity (i.e., occurrence number) of their corresponding G-Nodes in $GSet_M$. For the other entries, their values are accumulated based on the distances to the nodes in $GSet_M$ and their preference values, as shown in eq. (2).

$$PV(i) = \begin{cases} multiplicity(i) & i \in GSet_M \\ \sum\limits_{j\in GSet_M} \frac{PV(j)}{distance(i,j)+1} & i \notin GSet_M \end{cases} \quad (2)$$

2) *Manual correction* (optional): Automatic decision of preference values might not accurately demonstrate the user demands, optional manual correction is necessary to correct the values by the users themselves. However, it often becomes inappropriate or impractical to present the whole preference vector, especially when the vector is very long. In our design, the users are able to modify the values that are not equal to zero. This feature-oriented mechanism is complementary to the initial user inputs (including the preferred sources DS_I and domain DM_I).

• **Demand Matching**: This phase is to generate a ranked list of data sources that best match the identified demands. The ranking of the list is through comparing the numeric values of the utility function u that demonstrates the desirability of a data source for a specific user need. More exactly, in the application domain DM_I, a resource $R1 \in DM_I$ is preferred to a resource $R2 \in DM_I$ if and only if the expected utility of $R1$ is greater than the expected utility of $R2$: $\forall R1, \forall R2, R1 \succsim R2 \Leftrightarrow u(R1) \geq u(R2)$. Such a rational preference relation $\succsim$ is transitive, reflexive and complete.

The preference ranking is a typical multi-criteria decision making problem. In meta-querier customization, each criterion corresponds to a query capability identified in preference vectors (i.e., a G-Node). To make the ranking outcomes manageable by users, we assume *additive independence* exists among the maximal connected subgraphs, which is a normal assumption [12]. The utility of a resource R can be approximated by using an *additive value function* that breaks one n-criteria function into n individual one-criterion functions. Such an approximation not only simplifies the automatic adjustment and manual correction, but also performs well, even if the assumption does not strictly hold [18]. We construct an additive utility function u to aggregate the utility $cu(R[N_i])$ of each individual capability N_i provided by the resource R. The utility $cu(R[N_i])$ is 1 when the capability set CS of R contains N_i; otherwise it is zero. The additive weight of N_i is decided by its preference value $PV(i)$, which are generated from user inputs. For the nodes in any maximal connected subgraph (sg), the sum of their utility values should be less than or equal to δ. δ is 1, if the combination mode is used. δ is set equal to $\sum_{N_i \in V(sg) \cap GSet_M} PV(i)$, if the accumulation mode is used. Let *MCSG* be a set of maximal connected subgraphs, the weighted utility function can be represented as follows,

$$u(R[N_1, N_2, ..., N_n]) = \sum_{i=0}^{n} (PV(i) \times cu(R[N_i])) \qquad (3)$$

subject to the following constraint,

$$\forall sg \in MCSG, \sum_{N_i \in V(sg)} (PV(i) \times cu(R[N_i])) \leq \delta$$

• **Annotation Verification** (after run-time): This phase is to verify whether the new data sources are annotated by the accurate G-Nodes. For those functional components that cannot match with any concept node, the manual annotation is invoked to update and maintain M-Ontology (e.g., by inserting new mappings among the related query forms). As an optional phase, the manual verification can be conducted after the source selection. Although the new data sources might not be supported immediately, the verified annotation can be reused for the future recommendation.

4 Experiments

A prototype of M-Ontology and the related algorithms has been implemented on an open-source mapping management system, Alignment Server[8]. This system provides some basic functionalities on mapping management and discovery.

To show the effectiveness of our approach in real-world scenarios, we design and conduct two sets of experiments in the domain of air-ticket booking. Since the quality of ranking is subjective, it is hard to measure its correctness. Given that our primary goal is to find suitable data sources from a source repository SR to satisfy user demands (their preferred data sources DS_I) on capabilities, the focus of our experiments is to evaluate whether our approach can correctly identify capability matches between user inputs DS_I and the data sources in SR. Specifically, the experiments are designed to evaluate the effectiveness of capability matching, which is the most critical factor that affect the recommendation performance.

Experiment setup: From the UIUC web integration repository[2], we collect 38 query forms for air-ticket booking after eliminating the inactive webpages. First, we manually extract all the query forms from the webpages. They are expressed in Web Ontology Language (OWL) and follow a query-form ontology that was designed based on the HTML specification [1]. Second, we manually classify the functional components of these forms to generate 54 maximal connected G/A-Node sub-graphs, based on their capabilities. Each functional component is associated with a G-Node and a G/A-Subgraph. The manual classification is utilized in the initial construction of M-Ontology and the final evaluation of our algorithm.

Experiment scenarios: Assume that users input three preferred data sources DS_I, n of which are not in the repository SR. Our experiments will examine how well the algorithm can correctly find the data sources with the desired capabilities (that are possessed by the sources in DS_I) from SR.

The first experiment is for our proposed solution (referred to as *MOM*). Except these n non-inclusive data sources in DS_I, all the remaining sources are used to construct a domain-based M-Ontology (for construction details, see our prior work [15]). We assume users can correctly choose an appropriate domain DM_I for their queries. That is, an appropriate M-Ontology *MO* is chosen. With assistance of *MO*, the query forms in DS_I are normalized and analyzed to identify the user demands.

The second experiment evaluates the performance of a reference solution that is a classical nearest neighbor method (referred to as *NNM*). To find the capability correspondences, it compares the query forms in DS_I with the form of each source in SR. For the performance comparison, we use the same algorithms of query-form normalization and similarity calculation in both *NNM* and *MOM*.

To evaluate the performance, we use the following three measures: *precision* measures the proportion of the identified capabilities that are actually desired by users; *recall* measures the proportion of the desired and identified capabilities out of all the desired and identifiable capabilities; *f-measure* is the weighted harmonic mean of precision and recall.

Experiment results: The performance values per n in the Table 1 and 2 are calculated by the average of 100 samples. All the samples are randomly generated. The first and second set of experiments share the same samples.

Table 1. Capability-based matching by MOM

Unintegrated sources (n)	precision	recall	f-measure
3 of 3	90.1%	86.6%	88.3%
2 of 3	92.8%	90.6%	91.7%
1 of 3	96.5%	95.8%	96.1%

The first set of experiment results in Table 1 show promising evidence of effectiveness of *MOM*. Even if all the data sources in user inputs (DS_I) have not been integrated by any existing meta-querier, the f-measure rate also reaches 88%. If only one data source is unintegrated, the capabilities of almost all the data sources can be correctly and completely identified. That indicates a high possibility that our solution can make an accurate recommendation.

Table 2. Capability-based matching by NNM

Unintegrated sources (n)	precision	recall	f-measure
3 of 3	76.0%	35.4%	48.3%
2 of 3	83.9%	56.8%	67.7%
1 of 3	92.0%	78.5%	84.7%

The second set of experiment results are shown in Table 2. Clearly, our method *MOM* outperforms the reference method *NNM* by a large factor, especially when the user-preferred sources have not been integrated. The major reason is the name ambiguity in HTML codes so that it is difficult to find the capability similarity between two individual data sources. Different from *NNM*, our method *MOM* has a better performance by utilizing some regular patterns that are learned from the integrated data sources.

5 Conclusions and Future Work

Meta-querier customization is desired to meet various user needs. In our customization strategy, we design a capability-based solution to source selection based on user inputs. The core is a light-weight ontology, which is generated from a number of existing mappings among query forms. It is viewed as a repository of capabilities. The user demands and source capabilities are modeled, identified and matched based on this ontology. Initial experiments show the potential of this ontology-centric method. Interesting directions for future work include:

1) To reduce the user interaction, ontology/domain selection should be automated by calculating the similarity between user inputs and ontology contents.

2) To implement a practical source selection, the other properties (e.g., popularity, stability and credibility) should be included in the preference model.

3) To support the implicit user preference, the source selection algorithm should be integrated with the classical recommendation algorithms [3][23], e.g., content and collaborative filtering algorithms.

References

[1] HTML 4.01 specification: Forms. http://www.w3.org/TR/html4/interact/forms.html, 1999.

[2] The UIUC Web integration repository. http://metaquerier.cs.uiuc.edu/repository, 2003.

[3] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.*, 17(6):734–749, 2005.

[4] M. J. Cafarella, A. Y. Halevy, Y. Zhang, D. Z. Wang, and E. Wu. Uncovering the relational Web. In *WebDB*, 2008.

[5] J. P. Callan. Document filtering with inference networks. In *SIGIR*, pages 262–269, 1996.

[6] K. C.-C. Chang, H. Garcia-Molina, and A. Paepcke. Boolean query mapping across heterogeneous information sources. *IEEE Trans. Knowl. Data Eng.*, 8(4):515–521, 1996.

[7] K. C.-C. Chang, B. He, C. Li, M. Patel, and Z. Zhang. Structured databases on the Web: Observations and implications. *SIGMOD Record*, 33(3):61–70, 2004.

[8] J. Euzenat. An api for ontology alignment. In *ISWC*, pages 698–712, 2004.

[9] D. A. Grossman and O. Frieder. *Information Retrieval: Algorithms and Heuristics*. The Kluwer International Series of Information Retrieval. Springer, second edition, 2004.

[10] B. He, T. Tao, and K. C.-C. Chang. Organizing structured web sources by query schemas: a clustering approach. In *CIKM*, pages 22–31, 2004.

[11] J. Hong, Z. He, and D. A. Bell. Extracting Web query interfaces based on form structures and semantic similarity. In *ICDE*, pages 1259–1262, 2009.

[12] R. Keeney and H. Raiffa. *Decisions with multiple objectives: Preferences and value tradeoffs*. J. Wiley, New York, 1976.

[13] M. Lenzerini. Data integration: A theoretical perspective. In *PODS*, pages 233–246, 2002.

[14] A. Y. Levy, A. Rajaraman, and J. J. Ordille. Querying heterogeneous information sources using source descriptions. *VLDB*, pages 251–262, 1996.

[15] X. Li and R. Chow. An ontology-based mapping repository for meta-querier customization. In *SEKE*, pages 325–330, 2010.

[16] Y. Lu, H. He, Q. Peng, W. Meng, and C. T. Yu. Clustering e-commerce search engines based on their search interface pages using wise-cluster. *Data Knowl. Eng.*, 59(2):231–246, 2006.

[17] J. Madhavan, S. Cohen, X. L. Dong, A. Y. Halevy, S. R. Jeffery, D. Ko, and C. Yu. Web-scale data integration: You can only afford to pay as you go. In *CIDR*, pages 342–350, 2007.

[18] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*, chapter 16.4, pages 622–626. Prentice Hall, 3 edition, 2009.

[19] M. Shokouhi. Central-rank-based collection selection in uncooperative distributed information retrieval. In *ECIR*, pages 160–172, 2007.

[20] L. Si and J. P. Callan. Relevant document distribution estimation method for resource selection. In *SIGIR*, pages 298–305, 2003.

[21] W. Su, J. Wang, and F. H. Lochovsky. Automatic hierarchical classification of structured deep web databases. In *WISE*, pages 210–221, 2006.

[22] P. Thomas and M. Shokouhi. Sushi: scoring scaled samples for server selection. In *SIGIR*, pages 419–426, 2009.

[23] G. Uchyigit and M. Y. Ma, editors. *Personalization Techniques and Recommender Systems*, volume 70. World Scientific Publishing, April 2008.

[24] J. Xu and W. B. Croft. Cluster-based language models for distributed retrieval. In *SIGIR*, pages 254–261, 1999.

[25] R. Yerneni, C. Li, H. Garcia-Molina, and J. D. Ullman. Computing capabilities of mediators. In *SIGMOD Conference*, pages 443–454, 1999.

[26] B. Yuwono and D. L. Lee. Server ranking for distributed text retrieval systems on the internet. In *DASFAA*, pages 41–50, 1997.

[27] Z. Zhang, B. He, and K. C.-C. Chang. Understanding Web query interfaces: Best-effort parsing with hidden syntax. In *SIGMOD Conference*, pages 107–118, 2004.

[28] P. Ziegler, K. R. Dittrich, and E. Hunt. A call for personal semantic data integration. In *ICDE Workshops*, pages 250–253, 2008.

Issues in Building Ontology for Information Systems

Andrey Soares[1], **Frederico Fonseca**[2]
[1]School of Information Systems and Applied Technologies, Southern Illinois University Carbondale,
Carbondale, IL, USA
[2]College of Information Sciences and Technology, Pennsylvania State University, University Park, PA, USA

Abstract - *Despite a shared understanding that ontology plays a central role in Information Systems, researchers have not yet produced comprehensive guidelines for building ontologies for Information Systems modeling. This paper reports a study on methodologies to build ontologies for Information Systems. We searched major bibliographic databases from which we selected 30 methodologies to investigate. The analysis of the methodologies was formulated around the core components of an ontology and the four issues raised in a preliminary study (i.e. meta-models, procedure knowledge, temporal relations and knowledge acquisition). Our findings confirmed the issues among the methodologies investigated, and uncovered another issue to take into consideration in the process of building ontologies for Information Systems*.

Keywords: Ontology, Information Systems, Methodologies, Issues

1 Introduction

Building domain ontologies for Information Systems (IS) is not an easy task. It requires a great set of skills from the ontology engineer, as well as the involvement of domain experts. Despite three decades of research and a shared understanding that ontology plays a central role in Information Systems [2-5], researchers have not yet produced comprehensive guidelines for building ontologies for Information Systems [6].

This paper describes four issues that should be taken into consideration in the process of building ontologies for IS. The issues are related to meta-model, procedural knowledge, temporal relations, and knowledge acquisition. We report the results of a study on methodologies to build ontologies for IS, we discuss our findings with regard to the issues raised, and we introduce the issue on axiomatization, which emerged from our study.

2 Issues

Figure 1 presents an excerpt from a sample domain ontology that will be used to discuss the issues identified in the process of building ontologies for IS modeling. The domain under investigation refers to the process of adopting a cat from an animal shelter. The typical life-cycle of a cat at the animal shelter proceeds, for example, from the time a person brings a homeless cat to the shelter until the time that a cat leaves the shelter to an adopter's home. To achieve this goal, several activities have to be performed, such as posting information about a cat on the shelter's website, approving and relocating a cat to foster homes until adoption, approving adopters, and publishing the cat's happy end story.

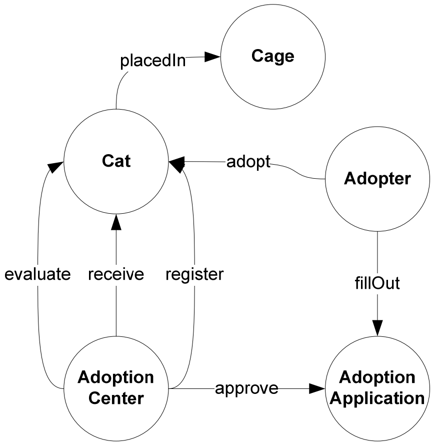


Figure 1: Excerpt from a sample domain ontology

2.1 Meta-model

Meta-models can provide a frame for mapping domain concepts. It should facilitate the integration of domain ontologies that share the same view of a domain [7]. In Figure 1, the concepts Cat (i.e., the animal for adoption) and Cage (i.e., a place to enclose the pet at the animal shelter facility) do not have a clear distinction between them. Both concepts are represented simply as two different concepts connected by a relationship. Although we acknowledge that the representation is correct, we would like to point out that it is also incomplete in terms of identifying the proper roles of the concepts. People should be able to understand these two concepts and the relation between them. However, computer systems would require some extra information to support the semantics of that relation.

We argue that a meta-model ontology could provide a refined understanding of the domain, as well as a frame to map the meta-model ontology with the domain ontology. A meta-model ontology defines the ontological commitment, that is, how we perceive the real world phenomena [8].

We envision a mapping, from the meta-model ontology to the domain ontology, based on scenarios [1]. This approach could increase the understanding of the domain and facilitate its interpretation. For instance, the generic concept of an “agent” in the meta-model ontology is mapped to the concept of an “adopter” in the domain ontology, which can then point to an instance of an adopter called “Joe” (see Figure 2). By connecting the concept “agent” with the concept “adopter”, we see an enhancement on our understanding about an adopter as it can be an agent within a scenario responsible for performing tasks to achieve goals and for triggering events that could change the state of other concepts.

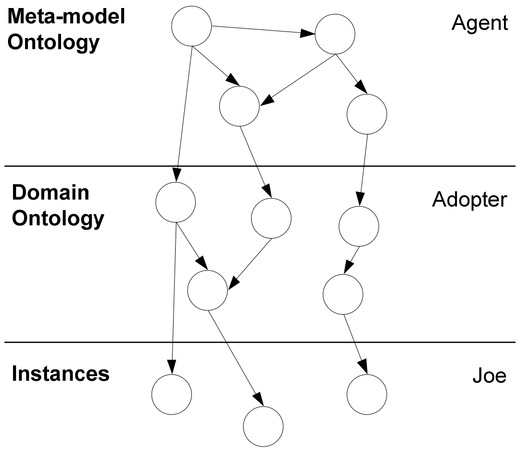


Figure 2: Meta-model and domain level ontologies

2.2 Procedural knowledge

Procedural knowledge refers to a sequence of tasks needed to achieve a goal, that is, the description of how to do things [9]. This is a knowledge about “processes, tasks and activities” [9, p.4]. It refers to the knowledge “generated whenever people refine step-by-step processes for standardizing simple, everyday work processes” [10, p.126].

In the animal shelter domain, the overall goal is to get a cat adopted. This goal depends on the achievement of several sub-goals and tasks. It is possible to identify in Figure 1 at least three distinct clusters of activities that should be performed to achieve the overall goal. First, when someone brings a homeless cat to the animal shelter, second, when a potential person (called adopter) applies for adopting a cat, and finally, when a match between adopter and cat occurs and the process of adoption is completed.

An ontology that will represent the animal shelter domain and be the basis for the design of an information system for the same domain should include a proper representation of the tasks above. These tasks portray the main activities in a domain, and should be used to understand what the system is and how it works.

Procedural knowledge is an important feature for Information Systems. However, it is usually not covered by methodologies to build ontologies.

2.3 Temporal relations

A temporal relation is an approach in knowledge representation involving the events of an application domain as “they exhibit a history of changes through time” [11, p.7]. In Figure 1, the tasks are represented with no particular temporal identification, which makes difficult to define the chronological order of the tasks. Assessing the correct order of the tasks can provide information about the timeline for performing the tasks, which tasks are needed to achieve goals, and how the tasks depend on each other.

From the sample ontology, we cannot identify which domain task comes first or later in the process of adopting a cat, and we cannot say whether the tasks belong to the same cluster of activities (i.e., scenarios). For instance, in the overall process of adopting a cat, a person receiving a cat at the shelter would be one of the first tasks, and an adopter adopting a cat would be one of the last tasks. However, the temporality between these tasks in Figure 1 is nonexistent.

2.4 Knowledge acquisition

Knowledge Acquisition refers to the process of capturing and representing domain knowledge [12]. The lack of appropriate guidelines puts pressure on domain experts and ontology designers, who have to find on their own, ways to identify the relevant knowledge to be represented by the ontology. Breitman & Leite [13] warn that “available methods for ontology construction […] concentrate in the modeling aspects and are either vague or lacking on how concepts and relationships are to be elicited” (p.4). In particular, the function requirements (i.e., actions and their constraints) of Information Systems are overlooked by ontology design [14].

According to Milton [9], knowledge can be described and seen in different ways (i.e., conceptual vs. procedural knowledge, tacit vs. explicit knowledge). Therefore, methodologies proposing knowledge acquisition, should present detailed information about the activities to capture and represent different types of knowledge.

3 Empirical Study

We conducted a Systematic Review [18] to identify methodological guidelines for the process of building ontologies that are suitable to IS modeling. Based on our

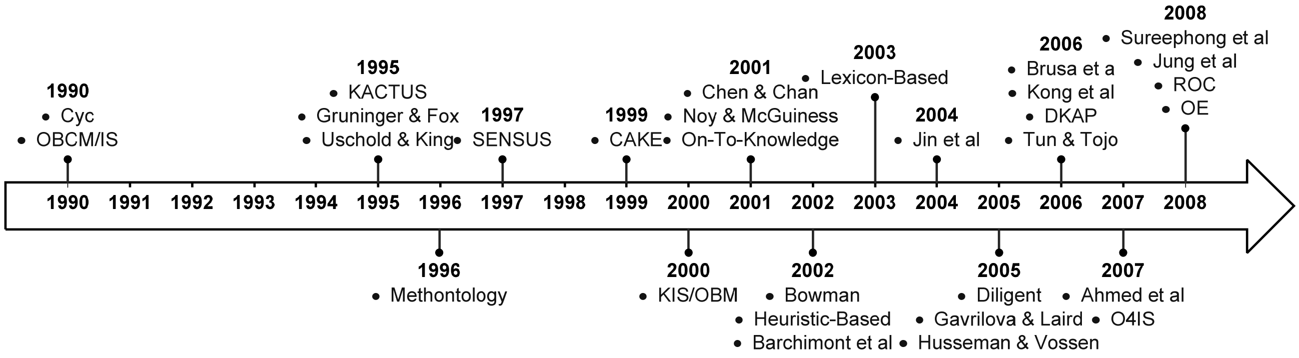


Figure 3: Development timeline of the selected methodologies [1]

search strategy and inclusion/exclusion criteria, we selected 30 methodologies to investigate (see Figure 3) from major bibliographic databases (i.e., ACM, IEEE, Springer, Elsevier, Web of Science, and Proquest). A detailed discussion of the systematic review will be addressed elsewhere.

The analysis of the methodologies is based on the following categories that we developed to investigate methodological approaches to capture domain knowledge, to identify the main components of an ontology (i.e., concept, properties, relations and axioms), and to address the issues above.

1. Knowledge Acquisition: this criterion aims to identify methods that can help in the process of acquiring knowledge about a given domain;
2. Identify Concepts: shows how a methodology supports the identification of domain concepts and their related properties;
3. Identify Relationships: shows how a methodology supports the identification of the relationships between concepts;
4. Identify Tasks: this criterion covers how the methodologies identify and represent the procedural knowledge needed to achieve goals;
5. Identify Temporal Relations: refers to particular ways used to identify and to represent the chronology and dependencies of tasks within the ontology;
6. Identify Axioms: an important feature of ontology is the possibility of representing relevant constraints of the domain. This criterion should provide valuable information on how the methodologies propose the identification and description of theses constraints;
7. Ontology Levels: developing ontologies with the help of a meta-model ontology can provide additional knowledge about the domain. This criterion focuses on the methodologies that are using different levels of ontology;
8. Mapping: if a methodology has adopted different levels of ontologies, it should provide guidelines for identifying the constructs of the higher-level ontologies and for the mapping between levels;
9. Methodological Steps: describes the sequential steps proposed to build an ontology;
10. Examples: provides examples to illustrate how to apply the methodology or some of its steps to build ontologies;
11. Study of Existing Methodologies: this criterion identifies which existing methodologies have been studied or compared with to define the issues to be solved;
12. Use of Existing Methodologies: shows if the methodology incorporates parts of existing methodologies into their own approach.

3.1 Discussion

The overall occurrence of each category is presented in Figure 4. Each category may include descriptions that range from a category not identified within the methodology, to a list of topics without their descriptions, or a list of topics with detail descriptions of an approach. The count of occurrences should not to be considered a measure of quality of the methodologies.

An initial observation of the chart in Figure 4 reveals that some methodologies do not cover all the categories. In fact, only the O4IS (Ontology for Information Systems) methodology [14] was represented in all categories. It not only covers important aspects of ontology design, but also provides the most comprehensive guidelines for building ontologies for Information Systems. The chart also shows that the initial issues discussed in this paper indeed exist among the methodologies analyzed.

With regard to *meta-model*, the adoption of levels of ontology is present in some methodologies investigated, usually with the intention to differentiate levels of abstraction. By agreeing on a meta-model ontology, we establish an ontological commitment to a particular view of the domain under investigation. When a methodology

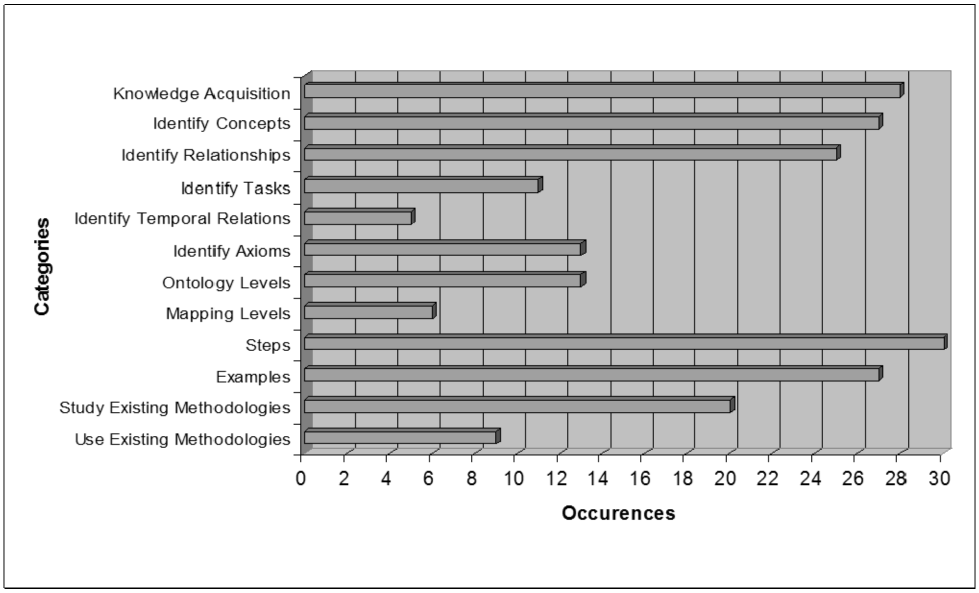


Figure 4: Occurrences per categories

adopts different levels of ontologies, the level above becomes a grammar (i.e., frame) to be mapped into the level below, which requires some guidelines for identifying what constructs to use. Out of 30 methodologies, 13 provided support to Ontology levels, and only 6 methodologies presented some procedures for mapping the concepts between levels.

With regard to *procedural knowledge*, 11 out of the 30 methodologies analyzed included support to representing procedural knowledge in the ontology. Identifying procedural knowledge involves the need to understand how the entities in the domain interact to each other. It would include a description of the behavior (i.e., the dynamics) of a system. From the methodologies providing such support, some describe the constructs used to represent the procedural knowledge (e.g., event, process, state), and others describe the process for identifying the procedural knowledge (e.g., competency questions, scenarios).

With regard to *temporal relations*, of the 30 methodologies investigated, only five provided support to the identification and representation of temporal relations. This relations are identified from the interplay of events within a domain and their dependencies [15]. The O4IS methodology [14] presented the most comprehensive approach for capturing temporal relations. O4IS bases its SAR:Temporal Relationships approach on the linear temporal logic theory, which describes the relation between two events (e.g., event A starts before event B).

With regard to *knowledge acquisition*, 28 out of 30 methodologies discuss some procedures for acquiring domain knowledge. However, there is still a need for more detailed guidelines for knowledge acquisition. Some methodologies discuss knowledge at different levels of detail, such as regarding to the source of knowledge, the relevant knowledge from the domain to be represented, and the structure to be used to represent knowledge.

Some methodologies adopt existing methods for knowledge elicitation, such as interviews with users and observations of their work routines. In addition, existing methodologies for knowledge acquisition are adopted or new ones are created. For instance, the Sureephong et al. methodology [16] suggested the use of the ORSD (Ontology Requirement Support Document) “to guide knowledge engineers in deciding about inclusion and exclusion of concepts/relations and the hierarchical structure of the ontology” (p.6), and the O4IS methodology [14] introduces the Unified Semantic Procedural Pragmatic (USP^2) Design for domain conceptualization, which includes the Semantic Analysis Representations (SAR), a mechanism for identifying structural, functional, temporal, prescriptive and deontic relationships.

In addition to the topics above, *axiomatization* has emerged as another important issue in the process of building ontologies for IS. An axiom is considered a core component of an ontology. However, steps to defining and representing them are not clearly described by many methodologies as only 13 methodologies presented some kind of support to identifying and representing axioms.

Steps for identifying domain restrictions are still unclear and sometimes limited to basic constraints (e.g., cardinality).

Although the methodologies provide descriptions about structures to capture axioms, ways to identify constraints, and constructs to represent the constraints, there is still no prominent methodology combining these approaches into a set of course of actions. Research in this direction has already been proposed. For instance, EZPAL is a tool to help users write axioms in the PAL-Protégé Axiom Language [17].

4 Conclusion

Ontologies have been used across different domains and for different purposes [4, 18], nonetheless, no methodological approach for building ontology has been prominent. We expect that the review of methodologies has produced valuable results for researchers and practitioners by uncovering specific approaches and methodologies to build ontologies for IS as well as by identifying important issues that should be taken into consideration in the process of building ontologies for Information Systems. In particular, we expect to shed light on features that could help improving existing methodologies or designing future ones.

The results of this study should help to enhance not only the process of building ontologies but also the quality of the ontologies created, as they will include more details about the domain being represented, especially with regard to how a system works.

5 References

[1] A. Soares and F. Fonseca, "A Meta-Model Ontology Based on Scenarios," in *AMCIS'11 - Americas Conference on Information Systems*, Detroit, MI, Forthcoming 2011.

[2] J. A. Bubenko, "On the Role of 'Understanding Models' in conceptual schema design," in *Fifth International Conference on Very Large Data Bases*, Rio De Janeiro, Brazil, 1979, pp. 129-139.

[3] F. Fonseca, "The Double Role of Ontologies in Information Science Research," *Journal of the American Society for Information Science and Technology*, vol. 58, pp. 786-793, 2007.

[4] N. Guarino, "Formal Ontology in Information Systems," in *Formal Ontology in Information Systems (FOIS'98)*, Trento, Italy, 1998, pp. 3-15.

[5] Y. Wand and R. Weber, "An Ontological Evaluation of Systems Analysis and Design Methods," in *IFIP WG 8.1 Working Conference on Information Systems Concepts: An In-Depth Analysis*, Namur, Belgium, 1989, pp. 79-107.

[6] B. Yildiz and S. Miksch, "Ontology-Driven Information Systems: Challenges and Requirements," in *International Conference on Semantic Web and Digital Libraries (ICSD-2007)*, Bangalore, India, 2007, pp. 35-44.

[7] I. Davies, *et al.*, "Analyzing and Comparing Ontologies with Meta-Models," in *Information Modeling Methos and Methodologies*, J. Krogs, *et al.*, Eds., ed: Idea Group, 2005, pp. 1-16.

[8] I. Kurtev, "Metamodel: Definitions of Structures or Ontological Commitments?," in *Workshop on Towers of Models*, Zurich, Switzerland, 2007, pp. 53-63.

[9] N. R. Milton, *Knowledge Acquisition in Practice: A Step-by-Step Guide*: Springer-Verlag London, 2007.

[10] V. Allee, *The Knowledge Evolution: Expanding Organizational Intelligence*: Butterworth-Heinemann, 1997.

[11] J. Mylopoulos, *et al.*, "Telos: Representing Knowledge About Information Systems," *Information Systems*, vol. 8, pp. 325-362, 1990.

[12] D. Diaper, "Designing Expert Systems - From Dan to Beersheba," in *Knowledge Elicitation principles, techniques and applications*. vol. 1, D. Diaper, Ed., ed New York, NY: Springer-Verlag New York, 1989, pp. 15-46.

[13] K. Breitman and J. C. S. Leite, "Ontology as a requirement engineering product," in *Eleventh IEEE International Requirements Engineering Conference*, Monterey Bay, California, 2003, pp. 309–319.

[14] K. Vandana, "Ontology for Information Systems (O4IS) Design Methodology: Conceptualizing, designing and representing domain ontologies," Doctor of Technology Doctoral Dissertation, Department of Computer and Systems Sciences, The Royal Institute of Technology, Sweden, 2007.

[15] J. F. Allen, "Maintaining Knowledge about temporal intervals," *Communications of the ACM*, vol. 26, pp. 832-843, 1983.

[16] P. Sureephong, *et al.*, "An Ontology-based Knowledge Management System for Industry Clusters " in *Global Design to Gain a Competitive Edge: An Holistic and Collaborative Design Approach based on Computational Tools*, X.-T. Yan, *et al.*, Eds., ed: Springer London, 2008, pp. 333-342.

[17] C.-S. J. Hou, *et al.*, "EZPAL: Environment for composing constraint axioms by instantiating templates," *International Journal of Human-Computer Studies*, vol. 62, pp. 578-596, 2005.

[18] D. L. McGuinness, "Ontologies Come of Age," in *The Semantic Web: Why, What, and How*, D. Fensel, *et al.*, Eds., ed: MIT Press, 2001.

Clustering of Ontology in Preventive Health Care Through Relational Ontology

Sang C. Suh
Kalyani Komatireddy
Department of Computer Science
Texas A&M University – Commerce
Commerce, TX 75429-3011

ABSTRACT: **Ontology is a rigorous and exhaustive organization of some knowledge domain that is usually hierarchical and contains all the relevant entities and their relations. Clustering simplifies the data into clusters. In this paper we propose to use a model of hierarchical relationship to group related item sets to form clusters which are formed by using relational ontology. Relational ontology represents relationship among attributes and concepts and varies from one domain to another. We show how clustering can be performed on relational ontology to form additional ontology from which further construction of relations can be deduced. Newly generated ontology will give the basis for discovering new rules learned from a database. Preventive health care (PHC) domain was selected as an example domain for demonstrating the methodology. Through this case study on PHC, relational ontology is an effective means to represent ontology and to deduce further relational knowledge that exist in the database.**

Keywords: Health care, Knowledge acquisition, Clustering, Relational ontology, Ontology relation, Concepts.

I. INTRODUCTION

Clustering is to determine the intrinsic grouping in a set of unlabeled data. Clustering has been studied intensively because of its wide applicability in areas such as web mining, search engines, information retrieval, and topological analysis. Clustering is of two types distance-based clustering and conceptual clustering. Two or more objects belong to the same cluster if they are “close” according to a given distance is called distance-based clustering. Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures. Standard clustering techniques such as k-means do not satisfy the special requirements such as high dimensionality, high volume of data, ease for browsing, and meaningful cluster labels for clustering documents. Many existing clustering algorithms require the user to specify the number of clusters as an input parameter and are not robust enough to handle different types of domain in a real-world environment. Traditional clustering algorithms require special handling for high dimensionality, high volume, and ease of browsing become impractical in real-world clustering. Furthermore, incorrect estimation of the number of clusters often yields poor clustering accuracy. Clustering based on relational ontology (CBRO) increases clustering accuracy by forming related item sets.

Health care is one of the most important components in our life. Disease or illness can really mean a down turn in our life. The biggest asset we can have in life is health. One of the most tragic things about many serious health problems is that they are preventable. Many problems that start as small health issues, we all tend to ignore can turn serious if they go undetected and untreated. Preventive health care (PHC) is the area of much needed in life. We want to implement our application on domain which is useful for every kind of people. So here we took PHC as our domain to implement because rather than treating a condition after it has progressed, preventive care focuses on preventing disease and maintaining proper health.

Hierarchy of Attributes and Concepts (HAC) is a hierarchical and conceptual clustering system that organizes data so as to maximize inference ability [8]. HAC accepts a database of structured data (Relationships) as input. In this method, an unstructured data is formed as structured data, forming a relation table maintaining concept for every attribute and concept in the particular domain.

In this paper we propose related item set based clustering, based on the relational ontology(is a, part of, derived from, located in, has agent etc). HAC is a hierarchical model tool for human cognitive concepts. Our application CBRO is the remodel of HAC using relational ontology. Related Item Set (RIS) is a set of related contributed attributes, using RIS and relational ontology we build relationship between contributed attributes and concepts. We also implemented the concept of HAC (Hierarchy of Attributes and Concepts) but here we maintain the hierarchy based on relation. We shown clusters in different levels using relational ontology. We also present PHC domain sample input and the output of implementing our application.

II. RELATED WORK

Clustering algorithms in literature are partitional clustering and hierarchical clustering**.** Partitional clustering algorithm divides the point space into k clusters. Hierarchical clustering is a set of nested clusters organized in a tree [4].

HAC is both a hierarchical and conceptual clustering system that organizes data so as to maximize inference ability. The idea of hierarchical clustering is to begin with each point from the input as a separate cluster. We then build clusters by merging clusters that are close to each other: repeatedly merge two or more clusters that are closest to each other out of all pairs [8]. Visual Data Analytics (VDA) system represents the hierarchy of attributes and concepts graphically. VDA is designed to represents attributes, concepts and their relationships visually [13].

HAC as a hierarchical algorithm constructs a hierarchical description of the structural data by iteration of the substructure. This hierarchy provides varying levels of interpretation that can be accessed based on the specific data analysis goals [12]. HAC accepts a database of structured data as input. In this method, an unstructured data is formed as structured data, forming a relation table for every attribute and concept in the particular domain. Figure 1 shows the HAC of PHC domain, buttons on the circle and buttons in the middle shows the attributes and concepts respectively. Lines show the relationship between concept and attributes. Figure 1 shows that healthy food is high level concept, where all other remaining concepts *diet, restaurant, exercise, habit,* and *books* are the low level concepts comes under this high level concept. The concepts diet, restaurant, exercise, habit, books are the low level concepts which are always under the high level concept Healthy food. Healthy food concept is represented with different color button compare to other concept buttons, which tells the high level concept to all other concepts.

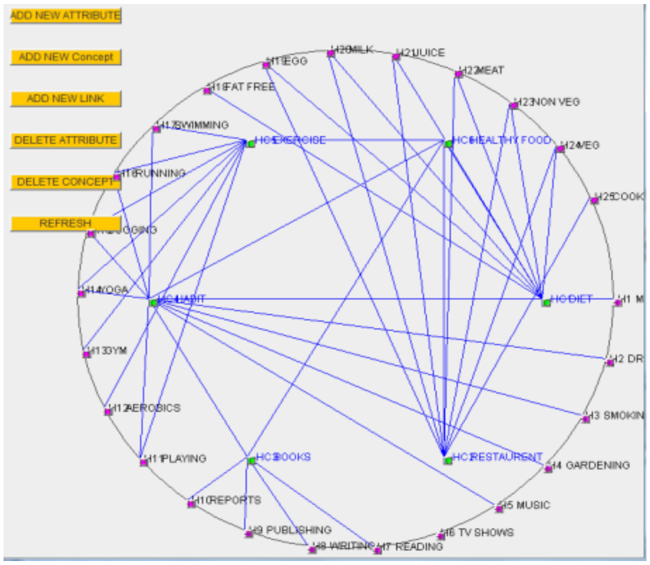


Figure1: HAC of PHC

In this paper, we propose clustering based on the semantics (i.e. meaning), according to which we define ontology relation for every attribute. In a cluster every attribute is related based on that relation such as brain tumor and leukemia is related to the same cluster. For example, since brain tumor is a cancer and leukemia is also a cancer, they are related according to ontology relation (*is-a*).

III. RELATIONAL ONTOLOGY

Relational ontology varies from domain to domain. Ontology, a cornerstone of the semantic web, have gained wide popularity as a model of information in a given domain that can be used for many purposes, including enterprise integration, database design, information retrieval and information interchange on the World Wide Web. AIMS (Agricultural Information Management Standards) define different categories: *traditional thesaurus relationships, concept-to-concept relationship,* and *term-to-term relationship* [7]. FIPA (Foundation for Intelligent Physical Agents) defines ontology relations using extension, identical, equivalent, strongly translatable, weekly translatable, and approx translatable [20]. Gene Ontology defines some ontology relations like *is a, part of, regulates, negatively regulates,* and *positively regulates* [2]. For our work, we describe OBO (Open Biological and Biomedical Ontologies) relational ontology [6].

	Transitive	Reflexive	Anti symmetric
Is a	x	x	X
Part of	x	X	X
Integral part of	X	X	X
Proper part of	X		
Located in	X	X	
Contained in			
Adjacent to			
Transformation of	X		
Derives from	X		
Preceded by	X		
Has participant			
Has agent			
Instance of			

Table 1: Basic relational ontology relations

Is-a relation is transitive. Transitive means "is a sub set of".

Example of transitive: if X is equal to Y and Y is equal to Z then X is equal to Z.

Part-of relation is reflexive means "is equal to".

In this paper we have taken five relational ontology relations from the above mentioned relational ontology. They are *is-a, part-of, derived-from, located-in and has-agent*. Using these ontology relations, contributed attributes are clustered as concepts in PHC domain. We combined two ontology relations and defined a new relation to represent a combined ontology relation.

For example: IS-A⊕ LOCATED-IN $\Longrightarrow$ OCCUR-IN

Hives come under the concept allergy according to ontology relation *is-a* and it also comes under the concept skin according to ontology relation *located-in*. From these two ontology relations, a new ontology relation *occur-in* derived between allergy and skin.

Figure 2 shows the clustering based on the ontology relation *part-of*. This cluster is the first level of cluster, clustering based on existing ontology relation.

For example: Diphtheria is a part of infectious disease, malaria part of infectious disease and Japanese encephalitis, lacrosse encephalitis, St. Louis encephalitis are part of infectious disease so we clustered all these as one cluster that is infectious disease based on the ontology relation *part-of.*

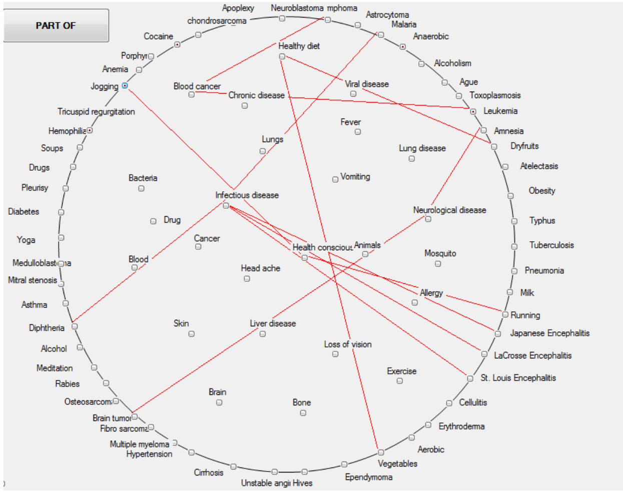


Figure 2: Clusters based on ontology relation “PART-OF”

IV. CLUSTERING ON RELATIONAL ONTOLOGY

In existing clustering methods, there has been no attempt to form clusters using relations. In this paper we build a database on PHC, using word net from which we define definition of each relation for each and every contributing attribute. Word net is a lexical database for English language [14]. It groups English words into sets of synonyms called synset, provides short, general definitions, and records the various semantic relations between these synonym sets.

We build a table on PHC domain which is our input. Sample input table is shown in Table 4 results section. We built contributed attribute table directly from the source table using the fields *ca-id* and *ca-name*. We used an algorithm to build concept table. An algorithm is shown below on how the concepts and relations are generated from the source table. For example, a cancer is a concept and its related attributes are brain tumor, lymphoma, leukemia, neuroblastoma, osteosarcoma, chondrosarcoma according to ontology relation *is-a*.

INPUT: Definition of each contributed attribute of every property
OUTPUT: Concepts generated

```
PROCEDURE:
    1. Initialize property[];
    2. Repeat Until each property from the defined
       properties          {
    3. Repeat until the end of each property (column )
                   {
    4. Repeat Until(j<n)          {
    5. compare the first string with all other strings if
       they are equal consider it as concept            {
    6. Add the String to the group of concepts
    7. make the property of new string as “OR”      }   }
       } }
```

Clustered data is shown in 2 different ways. One way is according to concept, based on ontology relation of first level as illustrated by Figure 4. Figure 4 shows the cluster of concept cancer. For example brain tumor, leukemia and ependymoma come under the concept cancer according to the ontology relation *is-a*. Second way is according to each relation, for example, for ontology relation *located-in* in Figure 3, amnesia and brain tumor are located in brain, so we cluster them as one cluster. Hives, cellulites and erythroderma are located in skin, so we cluster them as another cluster under *located-in* ontology relation. Here we combined two ontology relations and by combining them we defined one new relation between two concepts.

R_1: $a_1 \xrightarrow[IS-A]{} a_2$

R_2: $a_3 \xrightarrow[DERIVED-FROM]{} a_1$

Then $R_3 = R_1 \oplus R_2 \Rightarrow a_3 \xrightarrow[CAUSED-BY]{} a_2$

For example, since brain tumor is a cancer and loss of vision is derived from brain tumor as shown below, from these relations, a new relation, *caused-by* or *causes*, between cancer and loss of vision is derived, as shown below and in Figure 5.

Brain tumor $\xrightarrow[IS-A]{}$ Cancer

Loss of Vision $\xrightarrow[DERIVED-FROM]{}$ Brain tumor

Loss of Vision $\xrightarrow[CAUSED-BY]{}$ Cancer OR Cancer $\xrightarrow[CAUSES]{}$ Loss of Vision

In this way we derived new ontology relations which are shown in combined relation table (Table 2).

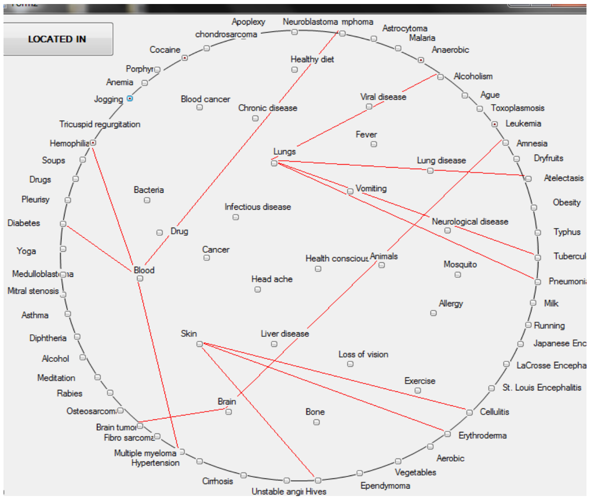


Figure 3: Clusters based on the relation “LOCATED-IN”

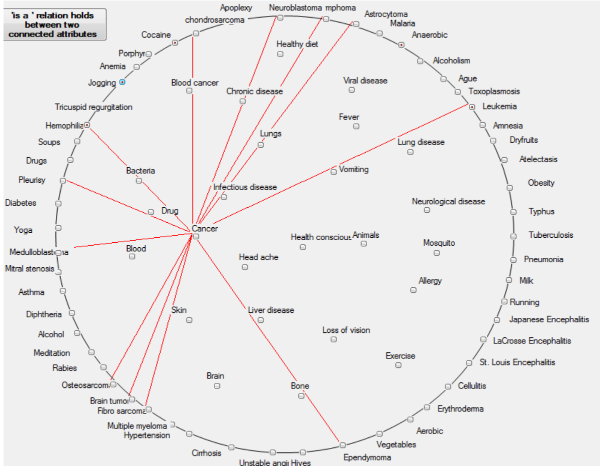
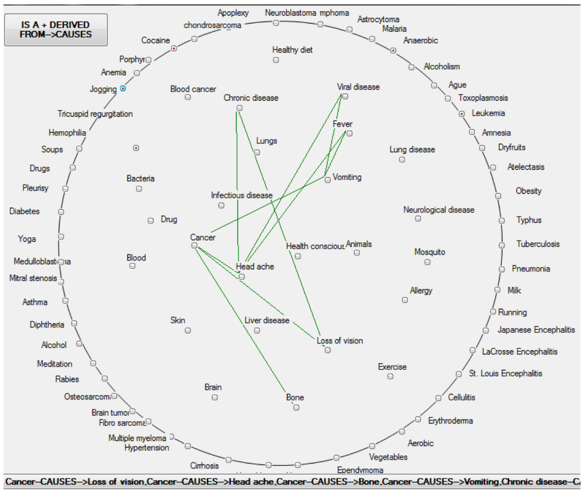


Figure 4: Cluster based on concept

Here we combine two HACs (built based on existing ontology relations) to form a new HAC representing another new relation as shown in Figure 5.



Cancer--CAUSES-->Loss of vision,Cancer--CAUSES-->Head ache,Cancer--CAUSES-->Bone,Cancer--CAUSES-->Vomiting,Chronic disease--C

Figure 5: New relation “Causes”

Combined relation	New relation
Is_A + Part_Of	Part Of
Is A + Located_In	Occur In
Is A + Derived From	Causes
Is A + Has Agent	Has Agent
Part Of + Derived From	Causes
Part Of + Located In	May Occur
Part Of+ Has Agent	May Caused By
Located In + Derived From	May Cause
Located In + Has Agent	May Has Agent

Table 2: Newly generated relations through Relational ontology.

V. IMPLEMENTATION

We have chosen PHC domain as the input database. Figure 6 shows the entire PHC domain using circles with buttons on the circle representing the contributed attributes and buttons inside the circle representing concepts. Preventive health care takes measures to prevent diseases (or injuries) rather than curing them. From our PHC domain, we only consider *causes*, *prevention* and *effects* of any disease among many others. The most important part of preventive health care is in maintaining good health habits.

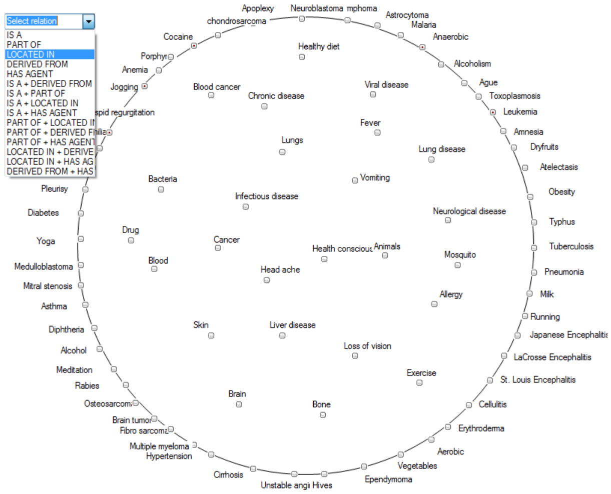


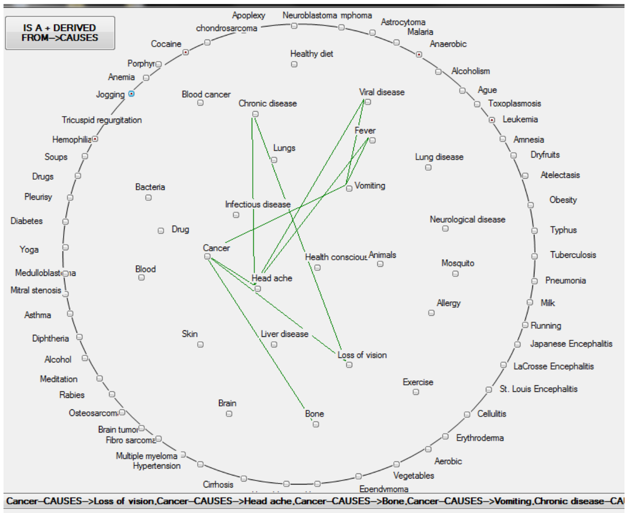
Figure 6: View of PHC domain in CBRO

The database for any domain to implement our application needs four tables: contributed attribute table, concept table, link table and combined relation table. The Contributed Attribute (CA) table contains CA_ID and CA_Name, concept table contains Concept_ID and Concept_Name, link table contains Concept_ID, CA_ID and Relations as shown in Table 3 , combined relation table contains Combined relation and New relations as shown in Table 2 .

Some of the Contributed attributes are Brain tumor, Malaria, Diabetes, Obesity, Lymphoma, Leukemia, Multiple myeloma, Dry fruits, Jogging. Some of the

Concepts are Cancer, Exercise, Healthy diet, Blood cancer, Health conscious, Neurological disease, Allergy, Loss of vision, Brain, Bone, Chronic disease, Fever, Drug, Lung disease.

Using these tables we represent the clusters visually in different ways. Initially it shows the circle with contributed attributes on the circle and concepts inside the circle. When we click on any concept it shows the related attributes by linking each and every related attribute with that concept. CBOR allows the user to select the relation from the dropdown list which they want, and then it shows the clusters based on that relation. CBRO combines two relations to get the new relation and forming next level clusters based on that new relation. CBOR shows the new relation with green links between two concepts like in Figure 7.



Cancer–CAUSES–>Loss of vision,Cancer–CAUSES–>Head ache,Cancer–CAUSES–>Bone,Cancer–CAUSES–>Vomiting,Chronic disease–CA

Figure 7: Combined cluster of "Causes"

When we click on the concept we will get the related contributed attributes of that concept and it will also displays the relation between the concepts and attribute. Figure 8 shows the cluster of concept healthy diet.

Example:
Concept: Healthy diet
Related Contributed Attributes: Vegetables, Dry fruits.
Relational Ontology: *part-of*

User also has an option to select the relational ontology from the drop down list which is shown in Figure 6. When user selects the particular relational ontology then it shows all clusters which are built based on that relational ontology. Figure 9 shows the clusters based on the ontology relation *has-agent*. Rabies and Ague has agent Animals.

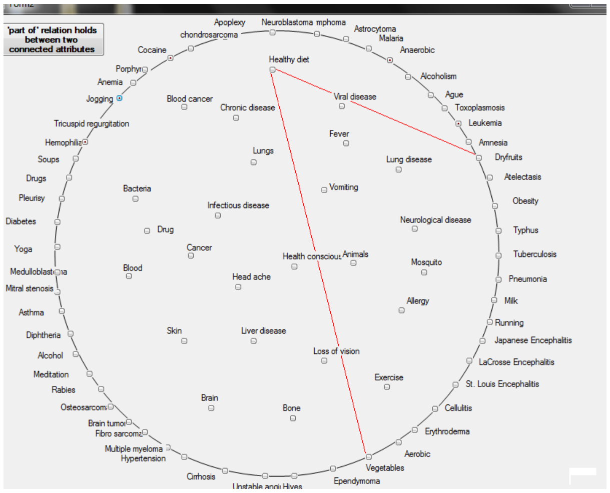


Figure 8: Cluster of Concept Healthy diet

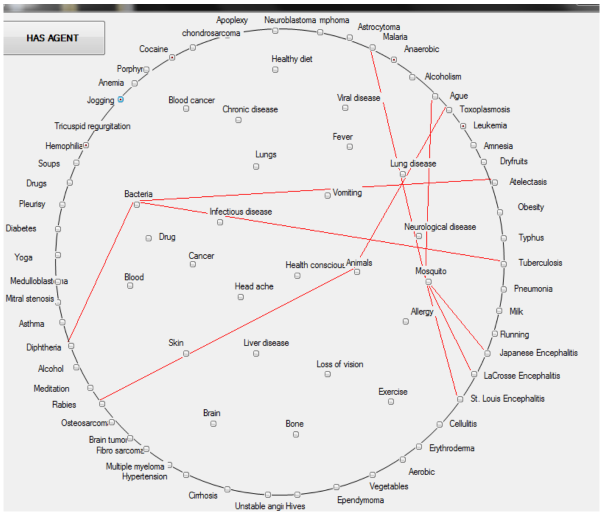


Figure 9: Clusters of ontology relation "HAS-AGENT"

Concept_ID	CA_ID	Relation	Concept_ID	CA_ID	Relation
C1	S1	Is a	C7	S20	Is a
C1	S5	Is a	C7	S21	Is a
C1	S6	Is a	C8	S1	Derived from
C1	S24	Is a	C8	S3	Derived from
C1	S25	Is a	C9	S1	Located in
C2	S11	Is a	C11	S2	Has agent
C3	S8	Part of	C11	S34	Has agent
C3	S10	Part of	C11	S35	Has agent
C4	S5	Part of	C12	S17	Has agent

Table 3: Link table

VI. RESULTS

Here is our sample input table which we used for CBRO. We built this table using information from various websites about the diseases, their symptoms and causes and the factors which affect the health. By considering health issues we built this table as shown in Table 4. To implement CBRO on PHC we derived four tables. Those are contributed attribute table, concept table, link table and combined relation table from the input table.

We used an algorithm to get the concept table and link table from the input table. CBRO derives the new ontology relation between concepts from the existed ontology relation between concept and contributed_attrinute. Output of CBRO contains the derived new ontology relation between two concepts. Diphtheria comes under concept fever based on ontology relation Is-A and it also comes under concept bacteria based on ontology relation *has-agent*, using these two we built a new ontology relation between two concepts fever and bacteria that is *has-agent*. Malaria may *derived-from* head ache and malaria *has-agent* mosquito from this we derived a new ontology relation *caused-by-agent*. Malaria *caused-by-agent* mosquito. Table 5 Sample output table shows some of the newly derived ontology relations.

For example:

Malaria $\xrightarrow[DERIVED-FROM]{}$ Head ache

Malaria $\xrightarrow[HAS-AGENT]{}$ Mosquito

So, Head ache $\xrightarrow[CAUSED-BY-AGENT]{}$ Mosquito.

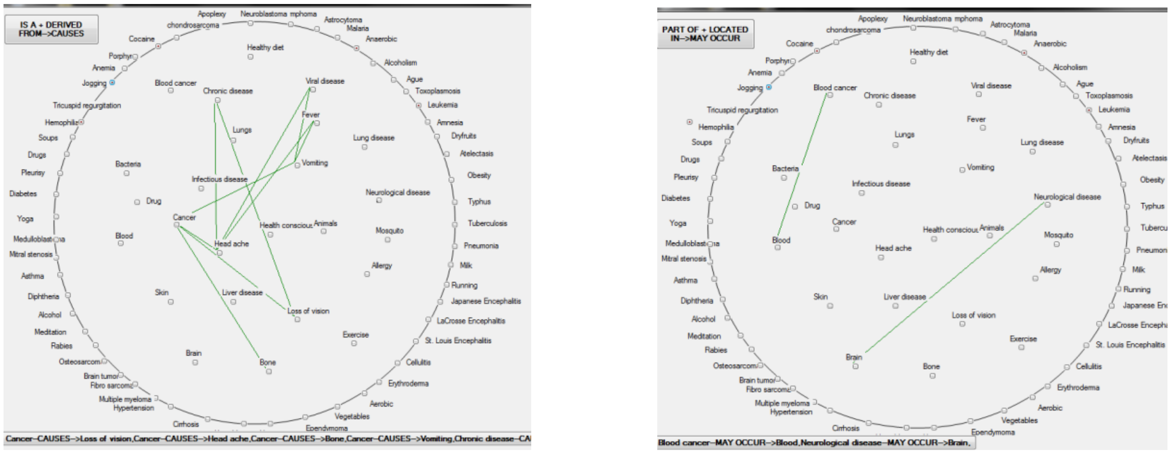
Figure 10 shows the derived ontology relations and their clusters. Figure 10(a) shows ontology relation *causes* which is determined by combining *is-a* and *derived-from.* Examples of derived relation *causes*: Cancer causes headache, cancer causes loss of vision, fever causes vomiting, chronic disease causes loss of vision etc. Figures 10(b) show the ontology relations *may-occur*. Examples of derived relations *may-occur*, *caused-by*, *may-cause*: Neurological disease may occur to brain, head ache caused by mosquito, vomiting caused by mosquito, disorder in brain may cause head ache, blood diseases may cause loss of vision etc.

Concept_ID	**Relation**	**Concept**	**Contributed Attribute**
C1	IS A	Cancer	S1,C4,S54,S53,S52,S44,S43,S24,S25,S26,S27
C2	IS A	Exercise	S9,S11,S12,S55,S56
C3	PART OF	Healthy diet	S8,S10
C4	PART OF	Blood cancer	S5,S6,S7
C5	PART OF	Health conscious	S9,S12
C7	IS A	Allergy	S20,S21

Table 4: Sample input table

Relation	**Concept-Name**
Causes	(Cancer, Loss of vision), (Cancer, Headache)
Part-Of	(Exercise, Health conscious), (Blood Cancer, Cancer)
Occur-In	(Allergy, Skin), (Cancer, Blood), (Cancer, Brain)
Has-Agent	(Fever, Bacteria), (Viral disease, Mosquito)
May-Occur	(Neurological disease, Brain)
May-Caused-By	(Infectious disease, Mosquito), (Infectious disease, Bacteria)

Table 5: Sample output table



Cancer--CAUSES-->Loss of vision,Cancer--CAUSES-->Head ache,Cancer--CAUSES-->Bone,Cancer--CAUSES-->Vomiting,Chronic disease--CA

(a)"Causes"

Blood cancer--MAY OCCUR-->Blood,Neurological disease--MAY OCCUR-->Brain,

(b)"May-Occur"

Figure 10: Clusters of newly generated relations

VII. CONCLUSION AND FUTURE WORK

In this paper, clustering of data using ontology relations in PHC domain was discussed and illustrated. Few tables contributed-attribute table, concept table and link tables were built to implement CBRO for the PHC domain. First, data was clustered in concept level. Second based on the existing ontology relations *is-a, derived-from, part-of, has-agent, located-in* and then HAC's of two related ontology relations were integrated as one based on existing relations. A new ontology relation is generated through this process. Ontology relations *causes, part-of, occur-in, has-agent, may-occur, may-caused-by, may-cause, caused-by, may-cause, caused-by-agent, may-has-agent and derived-from* are derived between two concepts by implementing CBRO application in PHC domain (Data base).

In the future, further work can be done to expand this HAC by allowing more than two ontology relations to be combined as a single relation. While our application CBRO is domain dependent, it will be expanded to be suitable for any domain (domain independent) as future work.

VIII. ACKNOWLEDGMENT

This research has been partially supported by a grant from the US Department of Energy Grant (Funding #: DE-SC0001132). It has also been supported in part by a research enhancement grant from the graduate school of A&M-Commerce under grant #: 751-245876. We would like to express special thanks to Lavanya Gaddam for her help with manuscript editing.

IX. REFERENCES

[1] Benjamin C.M. Fung, Ke Wang, "Hierarchical Document Clustering Using Frequent Itemsets" Martin Ester SIMON FRASER UNIVERSITY, BC, Canada, September 2002.

[2] **http://www.geneontology.org/GO.ontology-ext.relations.shtml**

[3] Sudipto Guha, Rajeev Rastogi, Kyuseok Shims "ROCK: A Robust Clustering Algorithm for Categorical Attributes", Information Systems Volume 25, No. 5,pp 345-366, Elsevier Science Ltd, Britain, 2000.

[4] Cluster Analysis: Basic Concepts And Algorithms.

[5] Ji-Rong Wen, Jian-Yun Nie, Hong-Jiang Zhang "Query Clustering Using Content Words and User Feedback", SIGIR'01, New Orleans, Louisiana, USA, 2001.

[6] **http://www.obofoundry.org/ro/**

[7] **http://aims.fao.org/website/Ontology-relationships/sub**

[8] Sang C. Suh, Sam I. Saffer, Nikhil Goel, Young S. Kwon, "Generating Meaningful Rules Using Attribute Concept Hierarchy", Annie paper, 2006.

[9] E. A. Freigenbaum and H. Simon, "EPAM-like models of recognition and learning", Cognitive Science, Volume: 8, pp. 305 -336, 1984.

[10] R.T.Ng and J.Han "Efficient and effective clustering methods for spatial data mining". In Proceedings of the VLDB conference, 144-155, Santiago, Chile, 1994.

[11] M.Ester, H.P.Kriegel, J.Sander, and X.Xu , "A density-based algorithm for discovering clusters in large spatial database with noise". In International Conference on Knowledge Discovery in Databases and Data Mining, Portland, Oregon, 1996.

[12] Sang C.Suh and Gouthami Vudumula "The Role of Conceptual Hierarchies in the Diagnosis and Prevention of Diabetes", 2010.

[13] Sang C.Suh and Jhansi Baireddy "Visual Representation of Hierarchical Attributes and Concepts as an Ontology for Semantic Reasoning", 2009.

[14] Fellbaum, "WordNet: an electronic lexical database" Cambridge, MIT Press, 1999.

[15] Source available at wordnetweb.princeton.edu/perl/webwn

[16] Source available at en.wikipedia.org/wiki/Ontologies_(computer_science)

[17] Jan Jantzen, "Tutorial on Fuzzy Clustering", Technical University of Denmark.

[18] H. G. Wilson, B. Boots, and A. A. Millward, "A Comparison of Hierarchical and Partitional Clustering Techniques for Multispectral Image Classification", Geoscience and Remote Sensing Symposium, IEEE International, Vol.3 pp 1624 - 1626, 2002.

[19] Holmes Finch, "Comparison of Distance Measures in Cluster Analysis with Dichotomous Data", Ball State University, Journal of Data Science 3, 85-100, 2005.

[20] **http://www.obitko.com/tutorials/ontologies-semantic-web/relations-between-ontologies.html**

Modelling Relationships among Classes as Semantic Coupling in OWL Ontologies

Juan Garcia[1], **Francisco J. Garcia**[2], **and Roberto Theron**[1]
ganajuan, fgarcia, theron@usal.es
[1]Computer Science Department, University of Salamanca
[2]Computer Science Department, Science Education Research Institute (IUCE),
GRIAL Research Group University of Salamanca

Abstract— *In the context of the Web, an ontology provides a shared insight of a certain domain. It basically represents a knowledge base with interrelated concepts called classes. Relationships among classes in an OWL ontology are given by the object properties that are defined as a binary relation between classes in the domain with classes in the range. From a semantic perspective of the knowledge base, these relationships can be described as semantic coupling among concepts. Most of the tools use directed graphs or UML-based visualisations for representing them, but these techniques fail with scalability. In this paper we propose a pair of visualisations from a visual modelling tool focused on visualising the semantic coupling among classes in OWL ontologies.*

Keywords: OWL Ontologies, Modelling Ontologies, Semantic Coupling.

1. Introduction

Ontologies have been defined in many references, nevertheless Gruber's definition [1] later refined by Studer [2] has been the most cited: *An ontology is an explicit and formal specification of a conceptualisation*. In general, an ontology formally describes a domain of discourse; they are explicit representations of domain concepts and provide the basic structure or armature around which knowledge bases can be built. Each ontology is a system of concepts and their relationships, in which all concepts are defined and interpreted in a declarative way. Typically, an ontology consists of a finite list of terms and the relationships between these terms. The terms denote important concepts (classes of objects) of the domain with a hierarchy. Apart from subclass relationships, ontologies may include information such as: properties, value restrictions, disjoint statements, and specifications of logical relationships between objects.

Information modelling is concerned with the construction of computer-based symbol structures that model some part of the real world. Databases and knowledge bases represent the most common examples of these kind of structures. For instance, the Entity-Relationship model represents a conceptual model for a database, whereas a UML class diagram represents a conceptual model of one part of a software system. UML has also been proposed to model ontologies; there are diverse solutions implementing the OMG Ontology Definition Metamodel (ODM) specification. ODM offers a set of metamodels and mappings for bridging the metamodeling approach with ontologies -one example being the SOLERES project[1]-.

In contrast to UML modelling, some tools use graphs in order to represent the knowledge base of an ontology. These tools use graphs based on the natural representation of conceptual maps. Conceptual maps are tools for organising and representing knowledge. Their origin lies in the theories about the psychology of learning, enunciated in the 1960's. Their objective is to represent relations between concepts in the form of propositions. Concepts are included within boxes or circles, whereas the relations between them are explicated by means of lines connecting their respective boxes. The lines, in turn, have associated words describing the nature of the relation that links the concepts. This idea of conceptual maps has been widely used to model ontologies. In this paper we propose a different approach for modelling semantic coupling, based on a visualisation and a concept taken from the analytical visual theory. We discuss our proposal, then evaluate it with some users where the results are presented.

This paper is organised as follows: a brief introduction, some related work and the most important tools for editing or modelling ontologies, a description of our tool design, a discussion of an analysis of the tool with some users, and finally a discussion of conclusions and future work.

2. Related Work

Diverse commercial tools have been proposed for modelling ontologies. The most important currently are: SemanticWorks, TopBraidComposer, IODT and IODE. Some of these tools offer a free version with reduced funcionality. On the other hand, Protege is currently the most widely used tool to edit ontologies.

SemanticWorks[2] is a commercial tool designed to edit RDF documents in a GUI and check its sintaxis, as well as design RDF schema and OWL ontologies using a graphical

[1] http://www.ual.es/acg/soleres/ieee-smca/
[2] http://www.altova.com/semanticworks/owl-editor.html

design view. The modeller uses a representation based on squarified and expandable boxes for properties and classes, where the hierarchy is represented with a line above the box. This representation is based on the conceptual maps approach, intended to convey complex conceptual knowledge bases in a clear, understandable way. The properties associated with classes are also expanded and linked with the classes. This model is based on the expandable / collapsable trees that grow up according to the nodes that are being expanded. The model navigation starts by selecting one element in the ontology, for instance, a class or a property. The selected element is then displayed and all the information related to this element is displayed collapsed. Classes are represented using a squared box with left-side corners rounded, while properties are represented using a normal squared box. The main disadvantage of this modelling tool, is that it represents all the information in the same visualisation, even duplicating nodes. Duplicated elements result in a less efficient and redundant model; this redundancy causes the user to become easily confused navigating the model. The strategy of showing all the information in the same visualisation with redundancy makes it extremely difficult to navigate a large ontology.

TopBraid Composer[3] illustrated on figure 1 is an enterprise-class modelling environment for developing Semantic Web ontologies and building semantic applications. There are three available versions: a Free Edition, Standard Edition and Maestro Edition. TopBraid Composer is a UML-based modelling plug-in eclipse, part of the TopBraid Suite. We tested using TopBraid Composer Free Edition version 3.3.0 which does not support the UML representation that is provided only with paid versions. TopBraid Composer is a fully Protege-based tool that performs the most common operations over ontologies, such as: inference, consistency checking, and the inclusion of SPARQL query engine.

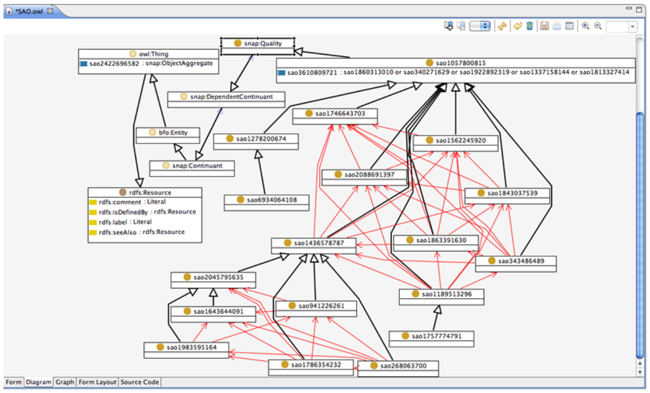


Fig. 1: TopBraidComposer uses a UML-based visualisation to represent both the hierarchy and the relations among classes.

IODT (IBM Integrated Ontology Development Toolkit)[4] is a toolkit for ontology-driven development, including EMF Ontology Definition Metamodel (EODM). EODM is an ontology-engineering environment that supports ontology building, management, visualisation and is also an open source project of Eclipse.org[5]. It has UML-like graphic notions to represent OWL class, restrictions and properties in a visual way. It can also have multiple views to support visualisation of an ontology; these views are independent but synchronized so changes made in one visualisation affect them all. Advantages of this modelling tool include: the widely known UML standard representation of classes, properties and the hierarchy, the facility to create diagrams, and the comprehensibility of them. The main disadvantage of this tool is the scalability to model a large ontology. Large ontologies become difficult to clearly modell and understand due to the surplus quantity of information displayed. Visualising all the information of an ontology in only one visualisation results in an overcrowded view of the knowledge model.

Protege[6] [3] [4] is a free, open source ontology editor and knowledge-base framework. Protege includes diverse plug-ins developed and maintained by the community. One of these plug-ins is OWLViz, a graph-based visualisation that represents classes, properties, hierarchy, and the classical tree of hierarchies view. Classes are represented as nodes in the graph, while properties are represented as edges connecting nodes, where the edges represent is-a relations (hierarchy).

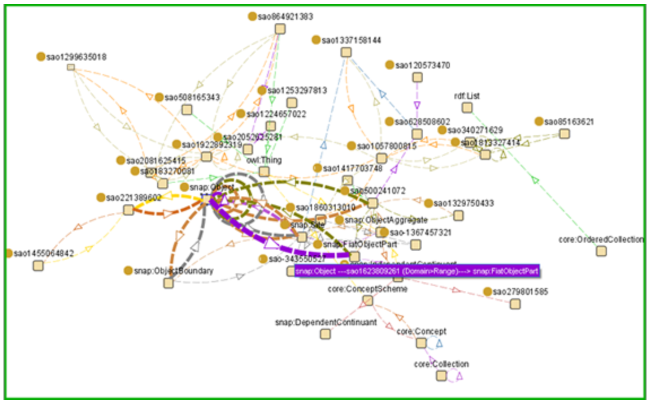


Fig. 2: Jambalaya tool represents relationships as coloured arrows, and the arrowhead indicating the direction of the property, from the domain to the range.

Jambalaya [5] [6] is another plug-in intended to visualise OWL ontologies with Protege. It can be found on its official

[3] http://www.topquadrant.com/

[4] http://www.alphaworks.ibm.com/tech/semanticstk

[5] http://www.eclipse.org/emft/projects/eodm/

[6] http://protege.stanford.edu/

site[7], and basically is a visualisation tool not provided with modelling capabilities. Jambalaya is a complete plugin that visually represents the components of the ontology and its relationships divided into two views. Each view can be displayed using one of six different layouts: grid, radial, spring, sugiyama, tree and treemap. The view shown on figure 2 represents a graph-based visualisation of an ontology. Classes are represented as nodes and the hierarchy is represented by edges connecting nodes. This graph connects classes with classes (is-a relationships and object properties represent coupling relationships among classes), as well as classes with their instances. The properties are represented using dashed lines with an arrowhead in the middle, indicating their direction going from the domain to range. This tool offers a great variety of configuration options - hiding components, changing colours and shapes and filtering data. Although Jambalaya represents a very good tool to visualise an ontology, scalability is the main disadvantage due to the fact that large graph visualisations are well known to become cluttered.

2.1 Analysing the Currently Available Tools

Katifori et. al. [13] provided with a classification of the diverse tools for visualising ontologies. They defined six categories according to the different chracteristics of the presentation, interaction, technique, functionality supported or visualisation dimensions. Nevertheless most of the tools fall in more than one category. We would add another group based on those tools that use UML notation to model ontologies. These seven groups are: indented lists, node-link (graphs) and trees, zoomables, space-filling, focus+context or distortion, 3D information landscapes and the UML-based. Basically, we can distinguish two main modelling approaches for representing the relationships among classes: the first one using the well known graph theory (Protege, SemanticWorks) and the second approach using UML diagrams, such as in the Object Oriented approach (TopBraid Composer, IODT). All the tools based on the use of graphs (node-link) have the same problems. The first is the lack of a layout, and the majority of the time the user moving the visual elements to organise them. The second problem is the scalability. It is well known that graphs are not good to represent a large amount of elements; these problems are illustrated on figure 2. The tools based on UML practically have the same problems which are illustrated in figure 1, even when this figure solely visualises less than twenty classes. Other approaches, such as: indented lists, treemaps or trees are focused on modelling taxonomies not relationships, and they are out of the scope of this paper.

As a result of an analysis of the current diverse tools, we have identified the main problems, including: the symbol redundancy (SemanticWorks), overcrowding of visual elements that difficult the understanding of visualisations, such as those based on directed graphs or UML. This problem is caused due to the majority of the tools saturating the visualisations and putting together the taxonomy with relationships. Therein the user gets easily confused and lost navigating the visualisation. Another detected problem is the lack of layout, in the case of graphs and UML diagrams. A lack of layout makes it difficult to find elements, and create a conceptual map of the knowledge base that is represented.

3. Modelling semantic coupling with OWL-VisMod

The visualisation of ontologies is a particular sub-problem of the field of graph and hierarchy visualisation with many implications due to the various features that an ontology visualisation should present. As [13] implied, there is not one specific method that seems to be the most appropriate for all applications. Consequently, a viable solution would be to provide the user with several visualisations, so as to be able to choose the one that is the most appropriate for his/her current needs. Some ontology management tools already provide combinations of visualisation methods. According to these detected problems, our objective has been to improve the visual expressiveness of our proposals in order to make them a better option than the current ones. Visual expressiveness is defined as the number of visual variables used in a visualisation to perform it efficiently. They are identified in a visualisation as the position (horizontal, vertical), shape (square, circle), size, brightness or opacity, orientation, texture and colours. We have decided to face these problems from a different perspective to improve the visual modelling of ontologies. We have made use of diverse concepts and techniques used in the visual analytics field, such as semantic zooms, radial layouts, the use of edges without arrowheads to avoid cluttering, the use of blurring and more. The first aspect we have considered in our tool is the separation of the diverse aspects to be analysed in an ontology. We decided to separate the representation of the hierarchy of the concepts with their semantic coupling. This is due to the fact that we believe users do not analyse all the aspects at the same time because the cluttered visualised items do not allow it. In this paper we have solely focused on representing the coupling; the taxonomy, however, is out of the scope of the paper.

OWL-VisMod[8] has been conceived to visually model OWL ontologies. This helps designers get a better domain of the knowledge base being modelled and makes modelling processes easier. It aims to provide a free helpful tool for developing ontological engineering. It has been developed in Java and represents an upper modelling tier; it uses Jena API[9] to manage OWL ontologies. Jena framework is respon-

[7] http://webhome.cs.uvic.ca/ chisel/projects/shrimp/demo/jambalaya-applet.html

[8] http://www.analiticavisual.com/juan/OWL-VisMod.html

[9] http://jena.sourceforge.net/

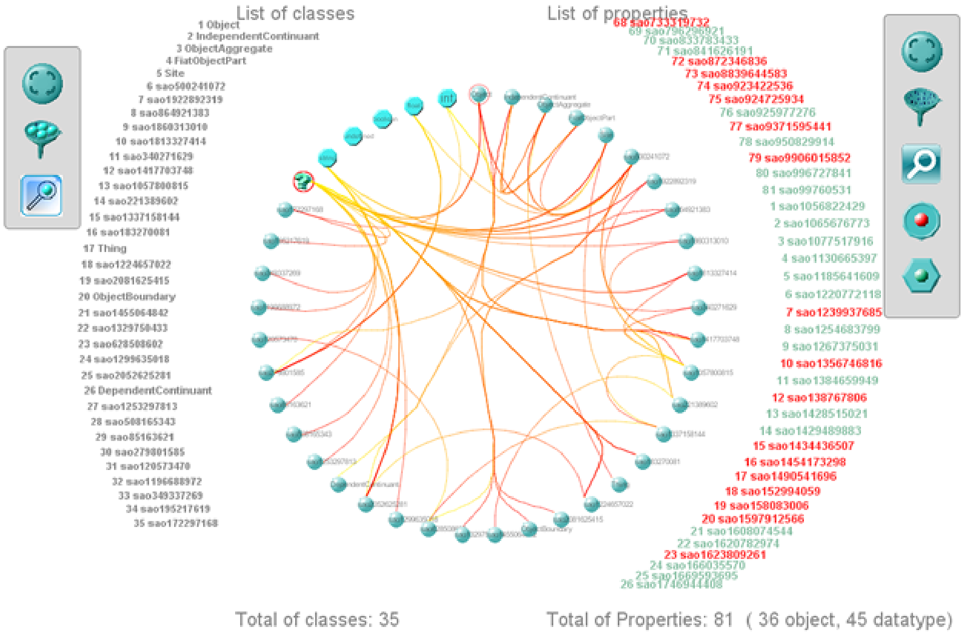


Fig. 3: A general view of the global coupling visualisation, where the classes are represented as spheres in a radial layout, and edges representing the properties as curves connecting them.

sible for loading and managing ontologies. OWL-VisMod queries directly over Jena framework and builds some data structures to manage the information. This information is processed, managed by the framework and finally visualised. This design was intended to get a framework independent from the API to manage ontologies. Nevertheless, we have no implemented any other API to manage ontologies. To describe the visualisations in this paper, we have used the ontology SAO v1.2[10] (Subcellular Anatomy Ontology), freely available and described in [11]. This ontology describes the subcellular anatomy of the nervous system, covering nerve cells, their parts and interaction between these parts. This ontology was built in Protege 3.2.x in OWL 1.0, and conforms to OWL-DL rules. This ontology has also been used in figures 1 and 2 to contrast with other tools.

3.1 Visualising the Global Coupling

As we have shown with diverse analysed tools that the lack of item layout makes the user get lost easily navigating them. We have decided to use a radial layout to organise the visual items, based on the proposals of [9] and [10]. We organised the visual items according to their CBE coupling metric value calculated as defined in [7]. The most coupled classes are located at the top of the visualisation and they are organised in the clockwise sense of the radial. This is done in such a manner in order to easily detect the most coupled classes, which the majority of the time result in the most important classes in the ontology. This organisation also allows the user to easily detect them. The coupling is represented using an edge that connects both coupled classes; this edge is defined by a Bezier curve and directed by the control points nearest to each of both classes. This approach also defined by Holten in [10], is intended to simulate forces that strengthen the edges in order to get a more clear visualisation. In this case, the control points can be modified in order to adjust the strength level of edges. This visualisation is illustrated in figure 3; it can be seen that the properties represented as edges vary from red to yellow, indicating the direction of the property; this representation avoids the use of arrowheads that overcrowds a visualisation. This representation of edges has been widely used in the visual analytics field, and has been shown to improve the visualisations avoiding cluttering.

All the names of the classes are displayed in a semi-circular list to the left of the spheres circle; the properties are also displayed using another semi-circular list on the right side. These lists offer the possibility of displaying an "infinite" number of elements. When there is a huge number of elements to be displayed, they are not all displayed at the same time, but in groups of n elements. To display the next or previous elements, it is only necessary to rotate the list by using the control in the menu or by scrolling up and down with the mouse. The global coupling shown in figure 3 is mainly formed by two parts: a pair of semi-circular lists and at the center of the visualisation the spheres representing the classes while the edges represent the properties or the coupling. It is important to highlight that this visualisation is

[10] http://ccdb.ucsd.edu/SAO/1.2/SAO.owl

completely focused on coupling among classes, so it means that just the coupled classes are represented, not all the classes in the ontology. Datatype properties are also represented in the visualisation because they can be viewed as a certain type of coupling. From a conceptual perspective, just object properties (red-coloured) represent coupling because they relate two concepts (classes) by means of a property. Nevertheless, we decided to represent the datatype properties (green-coloured) in the same visualisation, even though they can be hidden. The interaction includes the possibility of the user creating new properties, and these new properties are aggregated to the visualisation and to the ontological model. Each highlighted property is displayed in detail at the center, showing the class or classes in the domain of the property as well as the class or classes in the range. As illustrated in figure 4, once the user is interested in a specific property, and this property is pointed out with the mouse, a conceptual representation of its domain and range is displayed. In this case, the domain is formed by four classes grouped in a "cluster", while the range is formed by solely one class.

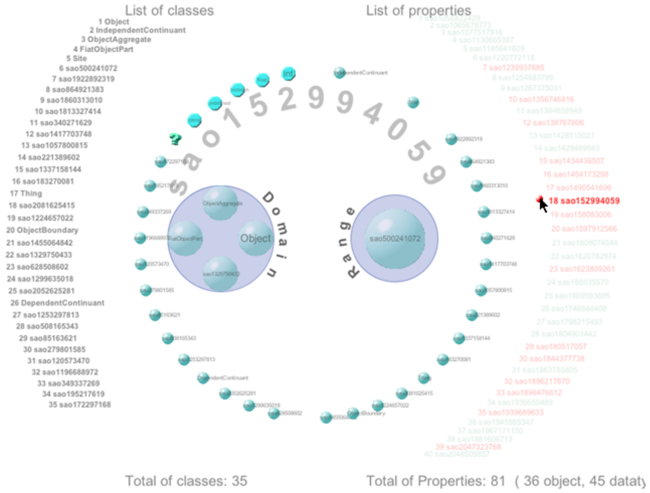


Fig. 4: A view of the property *sao152994059* where the classes in the domain are represented as a cluster with four elements, while there is only one class in the range.

3.2 Using Semantic Zoom to represent the Coupling of a Class

Figure 5 shows the user interaction with the visualisation. The coupling of a class is represented by highlighting the curves representing the properties declared in the class, as well as the coupled classes. The edges representing the coupling of this class are highlighted as previously described above, while the rest of the visual elements are blurred, allowing the user to clearly focus on the coupled classes and relationships.

Once the user has a global view of the coupling of a class, there is another more specialised view invoked with a single mouse click and performed using the semantic zoom technique. This zoom -which is a non-graphical- the user see different amounts of detail in a view, involving changing the type and meaning of the information displayed. In contrast to graphical zoom that only changes the size of the selected element, semantic zoom lets the user see more details that were previously hidden or not shown. To perform this technique a pre-process on the ontology is required. This pre-process consists of querying even the inherited coupling that was not included in the global view, resulting in a specific view where all the classes related to the selected one are displayed, as well as the domain and ranges of these properties. This view has been defined in [8] and depicted in figure 6. As to a brief explanation, it basically represents the selected class at the centre of the visualisation while all its properties defining coupling are located around it with their respective classes. The visualisation is read from left to right indicating the direction of the property, where the domain is represented in the left side opposite the range that is represented in the right side. This means that if the class is located left of the property, the class belongs to the domain of such property, in the opposite case the class would belong to the range of the property. Object properties are represented in red while the datatype properties are in green. This visualisation is focused on a more conceptual perspective of what the domain and range of a property mean. In this concrete example, the selected class *Object* belongs to the range of the property *sao138767806*, while the same class belongs to the domain of the rest of properties.

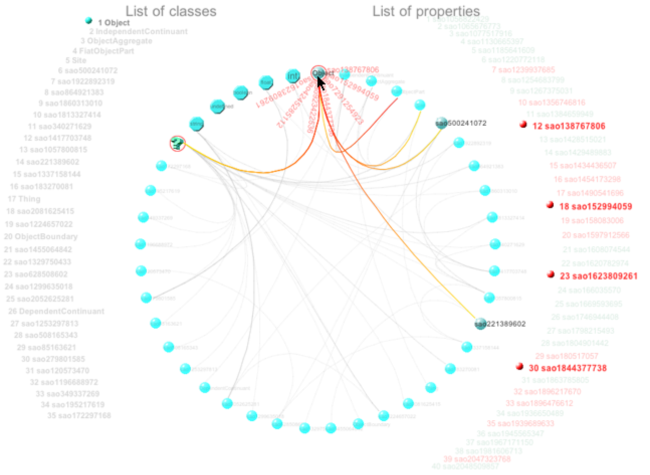


Fig. 5: The coupling of the class *Object* is depicted, all the relationships and coupled classes are highlighted as well as the names of the properties. The rest of edges, properties and classes are blurred.

The characteristics of the properties according to the official OWL specification[11], such as: inverse, symmetric,

[11] http://www.w3.org/TR/2004/REC-owl-features-20040210/

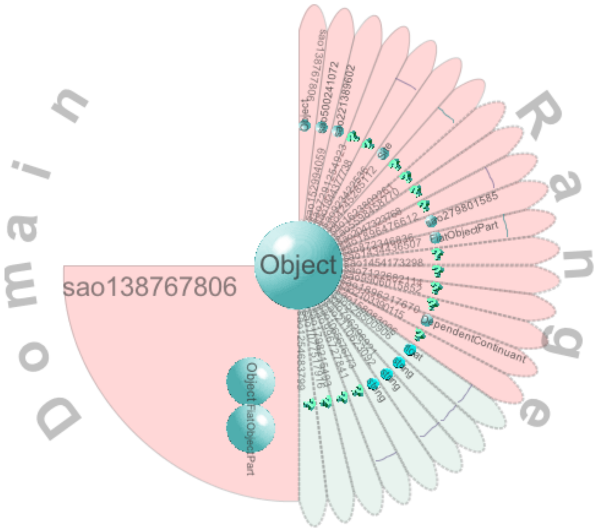


Fig. 6: A semantic zoom view of the most coupled class in the ontology, the class *Object*, where the object properties are red-coloured while the datatype properties are green-coloured.

functional, inverse functional or transitive properties, are also represented in both the semantic zoom visualisation and global coupling visualisation. We have followed the OWL specification, and currently OWL-VisMod is fully compliant with OWL-DL. It has implemented some functionalities of OWL-Lite, and the objective is that they be fully compliant.

4. Contrasting and Evaluating OWL-VisMod

Contrasting figure 3 with figures 1 and 2, and visualising the same ontology with Jambalaya and OWL-VisMod, we strongly feel that our proposal, which includes diverse strategies and techniques borrowed from the visual analytics field, improves the visualisation and representation of the semantic coupling among classes in an OWL ontology. The user navigation is easy to perform and the visualisation is intuitive for anyone, even those that are not experts in the ontologies domain. In order to verify that our tool has reached the goals it has been created for, we evaluated it with some users with different levels of specialisation in ontologies. We developed an evaluation of the whole tool, but only the results involved with the visualisations described in this paper are reported here. Basically, the evaluation consisted of the use of the tool by ten students (some of them specialised in ontologies), that were asked about their personal opinion on the usability of the tool, and whether or not visualisations are clear enough. Generally speaking, the evaluation was focused on assessing the effectiveness, safety, utility and learnability of the tool. It is important to highlight that OWL-VisMod is in the final phase of its development; we are developing testing in order to detect bugs that need to be fixed. The usability evaluation reported here did not include the contrast of our tool with the others analysed. This evaluation will be done once OWL-VisMod has been finished. At this point of our development, we are interested in getting feedback about diverse missing design aspects or things that can be improved.

For starters, users were given a brief introductory explanation about the tool, including some demos. Then, they were asked to create a new small ontology, and to navigate through a larger publicly available ontology, in order to use all the visualisations available. Once the users employed the visualisations, they were asked to answer a brief questionnaire. It included around twenty closed-ended questions to be evaluated in a quantitative manner, and five open-ended question to be evaluated in a qualitative manner. The closed-ended questions were rated from one to five, five being the number five the best evaluation and one the worst. To evaluate each question we used the mean of all the ratings given by the users. Here we report just five questions involved with both visualisations:

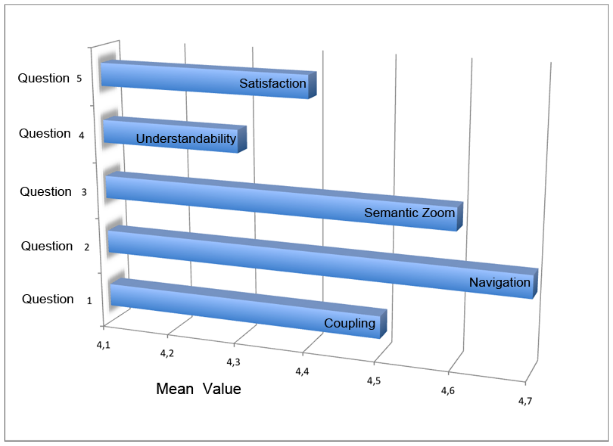


Fig. 7: A graph with the mean values of each question. The fourth question has been qualified with the lowest value.

1) Is the coupling visualisation clear enough to represent relationships like semantic coupling among classes? (1 - 5)
2) Is the navigation easy to perform in the global coupling visualisation? (1 - 5)
3) Is the semantic zoom visualisation clear enough to represent the coupling among classes? (1 - 5)
4) Are these visualisations easy to learn, use and understand? (1 - 5)
5) What is your degree of satisfaction with these both visualisations? (1 - 5)
6) Which extra comments would you add?

The mean value of each question used to evaluate them is graphically represented with a graph shown in figure 7. It can be seen that the second question referring to the easiness of the navigation model has a perfect evaluation, making clear that these visualisations are easy to navigate and are intuitive. The fourth question that was rated with the lowest value, refers to the easiness of learning, using and understanding these visualisations. According to the users comments, they basically said that the semantic zoom visualisation illustrated in figure 6 requires an explanation to be completely understood, due to is not as directly as could be thought, specially for those users not specialist in ontologies. There are some aspects in the visualisation that require an explanation, such as the interpretation of the dashed contour line surrounding the properties, indicating that these properties have been inherited by the class from a superclass. They also suggested the addition of more functionality in this visualisation. For instance, they suggested that when a property is selected, all the information of this property would be displayed. Figure 7 shows that the ratings have been high in general, indicating that in general, users consider that the tool to satisfy the objectives for which it was created. The feedback of comments about new improvements will be considered to be implemented.

5. Conclusions and Future Work

OWL-VisMod is a visual modelling tool that is currently at the final point of its development process. It is intended to provide a useful visual tool targeted to modelling OWL ontologies, based on the principle that they represent a knowledge base, and knowledge bases are better understood, easier edited and maintained using visual representations for the diverse concepts, than the traditional graphic user interfaces. Moreover, we have shown that the common visualisations, such as directed graphs, conceptual maps or UML diagrams, are not enough for effectively dealing with some medium or large sets of information. After analysing the weak points of the current tools, we have implemented diverse techniques and solutions, basically borrowed from the visual analytics approach; these improve the visual models. These techniques and visualisations have also been used in diverse fields with success, especially in the Software Engineering field [12], and now are brought to the ontologies approach. The use of layouts, semantic zooms or blurring, make visualisations more clear and understandable, and facilitate the navigation and modelling processes. We have also implemented a "divide and conquer approach" where diverse visualisations specialised in specific aspects are concerned. This is in order to separate the functionality, avoiding the saturation of visual elements representing all the aspects of an ontology in the same view.

These visualisations are linked to interact each with other and empower the functionalities of the tool. These improvements have been successfully used in other fields, and have now been implemented in the ontologies modelling field. We have discussed two of our visualisations focused on the concept of semantic coupling and we have provided a simple but rich feedback evaluation of them from users with a certain specialisation level in ontologies.

The future work includes the implementation of some of the suggestions of the users, as well as finishing the development of the tool. After finishing the development process, more evaluations are advisable, even contrasting it with the currently available tools.

Acknowledgment

This work was supported by Spanish Government project TIN2010-21695-C02-01 and by the Castile and Lion Regional Government through GR47 and the project FFI2010-16234.

References

[1] Gruber T.:Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal Human-Computer Studies Vol. 43, pp. 907-928, 1995.

[2] Studer R., Decker S., Fensel D. and Staab S.:Situation and Perspective of Knowledge Engineering. Knowledge engineering and agent technology, IOS Press, 2004.

[3] Gennari J., Musen M., Fergerson R., Grosso W., Crubezy M., Eriksson H., Noy N.: The Evolution of Protege: An Environment for Knowledge-Based Systems Development: Stanford Medical Informatics, 2002.

[4] Knublauch H., Fergerson R., Noy N. and Musen M.:The Protege OWL Plugin: An Open Development Environment for Semantic Web Applications, ISWC, 2004.

[5] Storey M.A., Musen M., Silva J., Best C., Ernst N., Fergerson R., and Noy N.: Jambalaya: interactive visualization to enhance ontology authoring and knowledge acquisition in Protege, in Workshop on Interactive Tools for Knowledge Capture (K-CAP-2001), Victoria, British Columbia, Canada, 2001.

[6] Storey M.A, Noy N., Musen M., Best C., and Fergerson R.: Jambalaya: an interactive environment for exploring ontologies: in Proceedings of the International Conference on Intelligent User Interfaces, San Francisco, California, United States, pp. 239, 2002.

[7] Garcia J., Garcia F. and Theron R.: Defining Coupling Metrics among Classes in an OWL Ontology: In Trends in Applied Intelligent Systems. 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2010, Cordoba, Spain, Springer, 2010.

[8] Garcia J., Garcia F. and Theron R.: Visualising Semantic Coupling among Entities in an OWL ontology: Proceedings of ONTOSE 2010, Ontology, Conceptualization and Epistemology for Information Systems, Software Engineering and Service Science, Springer 2010.

[9] Keim D., Mansmann F., Schneidewind J., and Schreck T.: Monitoring Network Traffic with Radial Traffic Analyzer : IEEE Symposium on Visual Analytics Science and Technology, pp. 123-128, 2006.

[10] Holten D.: Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data: IEEE transactions on visualization and computer graphics, vol. 12, no. 5, 2006.

[11] Fong L., Larson S.D., Gupta A., et. al.: An Ontology-Driven Knowledge Environment For Subcellular Neuroanatomy: OWL: Experiences and Directions, Innsbruck, Austria, CEUR Workshop Proceedings, ISSN 1613-0073, http://CEUR-WS.org/Vol-258/, 2007.

[12] Caserta P., Zendra O.: Visualization of the Static Aspects of Software: A Survey: IEEE Transactions on Visualization and Computer Graphics, VOL. 17, 2011.

[13] Katifori A., Halatsis C., Lepouras G., Vassilakis C. and Giannopoulou E.: Ontology visualization methodsâĂŤa survey: ACM Computing Surveys (CSUR), Vol. 39 , 2007.

A domain ontology based approach for analytical requirements elicitation

Fahmi Bargui, Hanene Ben-Abdallah, and Jamel Feki
FSEG, University of Sfax Tunisia, Po Box *1088*
{fahmi.bargui,hanene.benabdallah,jamel.feki}@fsegs.rnu.tn

Abstract - *In recent years, goal-oriented approaches have been used in Data warehouse (DW) projects to elicit the analytical requirements of decision makers. However, these approaches still suffer from a lack of assistance in goal elicitation, and provide little support to generate the suitable information for decision-making from the defined goals. To address these limitations, in this paper we introduce a domain ontology that aims at formalizing the semantic relationships between decision makers' goals, and representing explicitly the semantic links between the decision-making knowledge and the goals. The formal aspect of our ontology allows automated reasoning about the goals, and supports their decomposition which assists the automatic elicitation of these goals. Furthermore, the semantic links stored in the ontology ensure the automatic generation of suitable analytical requirements from the defined goals.*

Keywords: Ontology, Requirements analysis

1 Introduction

A Data Warehouse (DW) is a special type of data repository dedicated for decision-making support, which organizes information into facts and dimensions based on Multidimensional (MD) modeling. Since a DW often integrates data issued from several data sources, the construction of its MD model is often guided by the analysis of these sources [1]. In fact, several approaches have been proposed to automate the construction of a MD model from given data sources, *cf.* [3]. These bottom-up approaches apply a set of heuristics to derive candidate facts from which the decision maker chooses the MD model better reflecting his/her requirements. Although bottom-up approaches can reach a high degree of automation, they ignore the decision-making requirements. As a result, they may produce a MD model that does not meet the decision makers' needs. In addition, the decision maker has to invest a considerable effort in identifying which parts of the produced MD model are pertinent to his/her analysis.

To overcome these limits, several approaches advocate a requirements-driven DW design process. The proposed top-down, *cf.* [5] and Mixed, *cf.* [4] approaches include a requirements analysis phase in order to first elicit the information required by the decision makers, and then derive an adequate MD model. Within the literature, there is a consensus that this phase should be goal-oriented for two main reasons: (i) the DW provides useful information to make decisions contributing to the achievement of the organization goals, and (ii) decision makers often express their requirements in terms of goals that the DW should support [5].

On the other hand, there are several goal-oriented approaches proposed in the literature for the development of information systems (*e.g.,* I*[6]…). Thanks to their graphical languages and tool supports, these approaches gained acceptance both in the academic and industrial communities. However, their widespread usage is hindered by their lack of support in the elicitation of goals. In general, goal elicitation is conducted through the decomposition of high level goals into more concrete sub-goals, which requires domain knowledge and skills. Such knowledge is mastered by domain experts, and to the best of our knowledge, except for the work of Nabli *et al* [11], there is no attempt to formalize it in order to allow its interpretation by both humans and machines. In [11], the authors represent the technical concepts of decision-making in a *decisional ontology.* This latter contains the technical concepts of MD models (facts, dimensions, measures, etc) of a given domain. It can be used by a data warehouse expert to specify his/her analytical requirements. Even though the ontology specifies structural and semantic relationships between its technical concepts, it presents one major limit: it specifies neither the business process to evaluate nor the goals to fulfill, which are things most familiar to decision makers. This limit may hinder the identification of sub-goals and, consequently, the automation of goal elicitation.

Furthermore, our literature review highlighted that requirements-driven approaches provide little or no assistance in identifying appropriate performance indicators to measure the fulfillment degree of the elicited goals. The same shortage is also present in identifying information (*i.e.* data to be stored in the DW) that could be analyzed when the defined/expected goals are not met. In most cases, such information is elicited informally, and without explicitly matching to the goals. Consequently, when decision maker's goals change, it is difficult to trace the parts of the MD model that should be modified.

In our previous work [7], we proposed an approach that addresses the two main limits in current MD modeling: goal elicitation, and the correspondence between the elicited goals and the derived MD model concepts. In this paper, we show how to automate the steps of this approach. As illustrated in Figure 1, the input of our approach is a domain ontology formalizing the decision-making concepts. The ontology assists in the identification of the analytical requirements elements pertinent to the decision maker goals. Furthermore, the semantic relationships between the decision maker goals, stored in the ontology, assist the decomposition of goals and, consequently, the goal elicitation step. Moreover, the formal aspect of our ontology, which makes the domain knowledge machine readable, provides for potentially new requirements to emerge. This contributes to the completeness of the resulting specified requirements. Finally, all elicited elements are organized in order to assist the filling of our template for analytical requirements specification [8]. The resulting template is then used as input of our approach (presented in [10]) which ensures both the validation of the specified requirements with respect to the available data sources, and the design of a DW loadable from these sources. In this approach, the analytical requirements are parsed to extract pertinent terms that could be fact, measures, dimensions or parameters. To decide on the multidimensional type of a term, our approach applies a set of matching and expansion rules on the data source represented through its data dictionary.

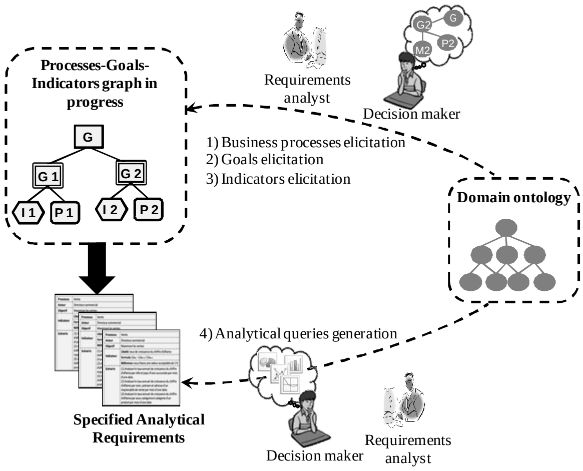

Processes-Goals-Indicators graph in progress
G
G 1
G 2
I 1
P 1
I 2
P 2
Requirements analyst
G2
G
M2
P2
Decision maker
1) Business processes elicitation
2) Goals elicitation
3) Indicators elicitation
Domain ontology
4) Analytical queries generation
Specified Analytical Requirements
Decision maker
Requirements analyst

Fig.1. Overview of our analytical requirements elicitation approach. G (Goal), P (Process) and I (Indicator).

The rest of this paper is organized as follows. In the next section, we present our NL-based requirements specification template. In section 3, the description of the decision-making ontology is given along with an ontology for the commercial domain. Section 4 explains how the ontology is used to automate the elicitation of analytical requirements elements. Finally, Section 5 summarizes our proposal and presents our future work.

2 Analytical requirements definition

In our previous work [8], we have defined a Natural Language (NL) based template for analytical requirements specification. The components of this template were determined through an empirical study covering samples of decision-making processes (*cf.* [9]). As illustrated in Figure 2, in addition to meta-data documenting each template instance, the template identifies the *business process* being analyzed and the *goals* of the analyses. The realization of each goal is measured through an *indicator* monitored through one *formula*. For instance, the performance of the business process "sales" can be analyzed through the achievement of the goal "increase sales". The achievement degree can be measured through the indicator "turnover growth rate". The decision maker (*Actor* in the template) can fix a *target value* (estimated value) that the process must reach for a given goal during a period of time not exceeding an estimated *deadline*. The attained value for a goal is measured by the corresponding indicator.

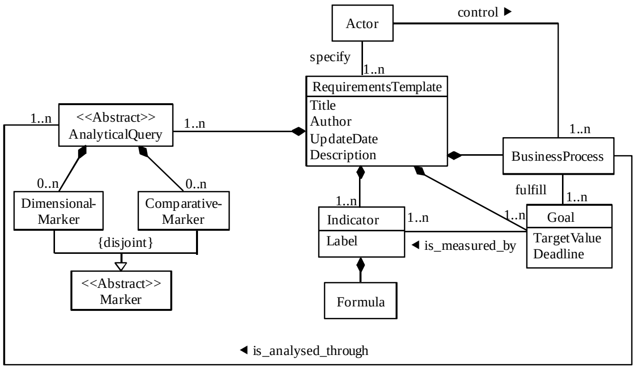


Fig.2. Analytical requirements template metamodel

The analysis of the discrepancy between the attained and target values allows the decision maker to evaluate the realization level of the goal and, hence, to judge the performance of the analyzed process. In the case of a negative variation, the decision maker notes an anomaly and looks for its origins. To do so, he/she examines detailed information retrieved from the DW through *analytical queries*. These queries refer to the indicator and formula terms.

To provide for a flexible requirements specification, our template allows decision makers to express their analytical queries in Natural Language (NL). Note that, In accordance with, a NL is the best means of expressing analytical requirements, mainly because it facilitates communications with the decision maker. However, the diversity of writing styles often causes semantic ambiguities. To overcome this difficulty, we chose to fix an expression style while benefiting from the advantages of NL.

To identify this expression style, we conducted an interview with twenty decision makers at different hierarchical levels (executives, managers…) and belonging to nine different domains (commercial, e-commerce…). In the interview, each decision maker was asked to write a set of queries describing samples of OLAP ("On-Line Analytical Processing) analyses he/she used to perform in decision-making. Our study of the 200 collected queries allowed us to elaborate a *query format* formalizing the recurrent and common components of these queries [8]. Furthermore, through our study, we identified two formats of analytical queries: *simple* and *compound* queries. As illustrated in Figure 3, a simple query includes *one indicator* and several analysis axes each of which is introduced by *one dimensional marker*.

SIMPLE QUERY $\underbrace{\text{Analyze}}_{\textbf{Verb}}\ \underbrace{\text{the turnover}}_{\textbf{indicator}}\ \underbrace{\text{by}}_{\textbf{Dimensional marker}}\ \underbrace{\text{customer's name and address}}_{\textbf{Analysis axis}}$

Fig.3. Format of simple queries

On the other hand, a compound query can always be divided into two or more simple queries. Figure 4 illustrates the decomposition of a compound query into two simple queries.

COMPOUND QUERY (two indicators)	**Analyze the *turnover* by product and the *delivery* by supplier**
SIMPLE QUERY	**Analyze the *turnover* by product**
SIMPLE QUERY	**Analyze the *delivery* by supplier**

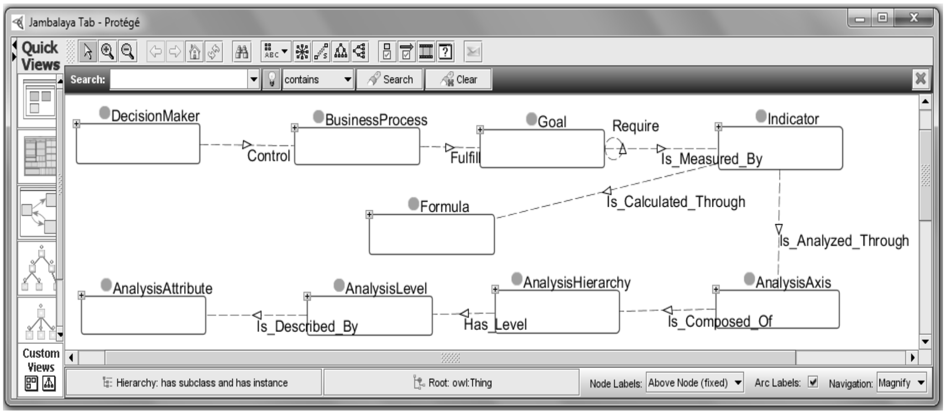
Fig.4. Decomposition of a compound query into simple queries

For the sake of simplicity of query processing, we adopt the simple query format as a means of analytical requirements specification in our template. This query format formalizes frequent writing styles of analytical queries. Moreover, it includes dimensional markers to introduce analysis axes, and comparative markers to specify analyses where the comparison between the realized and target values is significant. In addition, we have demonstrated that analytical requirements specified according to our template can be transformed to a MD model validated with respect to a given data source [10].

3 A decision-making ontology

We have conducted a second interview with decision makers at different hierarchical levels and belonging to different domains. The interview aimed at identifying the *terminology* used in the decision-making process. As illustrated in Figure 5, our study identified the following concepts:

- `DecisionMaker`: a person in the enterprise having the responsibility to evaluate and control the performance of a *business process*.
- `BusinessProcess`: is a collection of related activities that produce a specific service or product (serve a *goal*) for a particular customer.
- `Goal`: a *measurable* goal that a business process must *reach* during a given *period* of time. Goals can be classified into quantitative and qualitative. Quantitative goal is measurable, i.e. its achievement degree is evaluated by comparing an estimated value that the goal must reach (fixed by decision maker) with a realized value of the goal (calculated through an *indicator*). The period designates the latest time for achieving the goal. Qualitative goals do not have these properties and, therefore, must have a textual description. The study of qualitative goals is beyond the scope of this paper.
- `Indicator`: provides a value (calculated through a `formula`) designating the realization level of the corresponding goal.
- `AnalysisAxis`: over time, an indicator produces different values that could be aggregated (SUM, MAX, AVG…) according to an analysis axis. The aggregated value represents a point of view or a perspective that decision makers use in analysis tasks.
- `AnalysisLevel`: an analysis axis is composed of several analysis levels each of which represents a granularity echelon to aggregate indicator values. For instance, SupplierID, its City and Country are three analysis levels.



Hierarchy: has subclass and has instance

Root: owl:Thing

Node Labels: Above Node (fixed) Arc Labels: ☑ Navigation: Magnify

Fig.5. Decision-making Ontology metamodel in Protégé (thesaurus part)

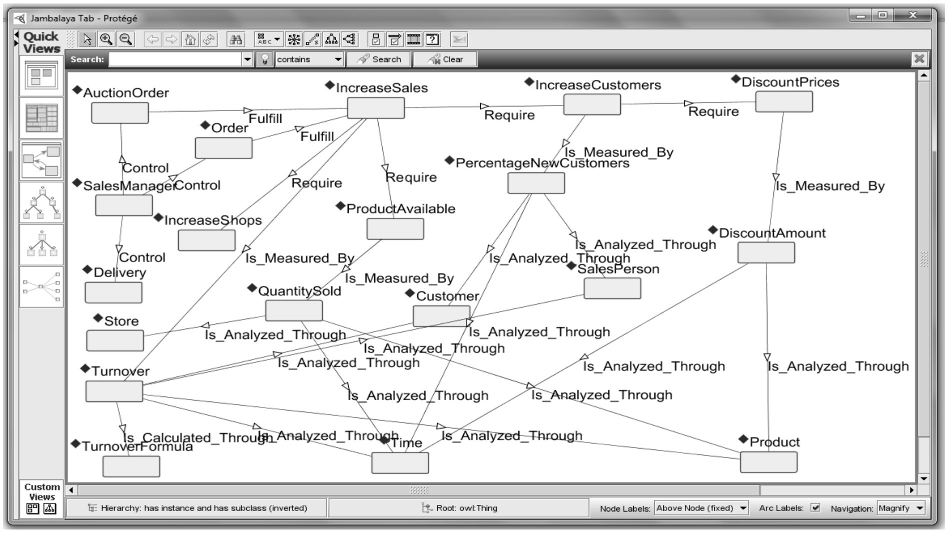
- AnalysisAttribute: an analysis level has a name that may not be significant enough (*e.g.*, SupplierID or a surrogate key). The role of an analysis attribute is to provide a textual description that explains the meaning of this analysis level name; *e.g.*, the supplier name (Sname).
- AnalysisHierarchy: The analysis levels are organized into analysis hierarchies where level names are semantically ordered from the finest to the highest granularity. As an example, the “SupplierID, City, Country” build a hierarchy.

Relationships formalize semantic links that associate concepts in the ontology. For the sake of clarity, we adopt the following abbreviations: P, G, D, I, F, A, H, L and At to designate, respectively, a business Process, a Goal, a Decision maker, an Indicator, a Formula, an Analysis axis, a Hierarchy, a Level and an Attribute. The identified relationships for the ontology are described as follows:

- Control(D, P): The decision maker D controls the performance of the business process P.
- Fulfill(P, G): The business process P is defined to fulfill the goal G.
- Is_Measured_By(G, I): The fulfillment degree of the goal G is measured by the indicator I.
- Is_Calculated_Through(I, F): The value produced by the indicator I is calculated through the formula F.
- Is_Anlyzed_Through(I, A): The indicator I is analyzed through the analysis axis A.
- Is_Composed_Of(A, H): The analysis axis A is composed of a hierarchy H.
- Has_Level(H, L): Hierarchy H has level L.
- Is_Described_By(L, At): The level L is described by the analysis attribute At.
- Require(G, G1): The fulfillment of goal G requires the achievement of G1

Figure 6 illustrates an instance for the commercial domain of our ontology metamodel. In this instance, the concept SalesManager is defined to control the performance of the three business processes: Order, AuctionOrder and Delivery; each of which must fulfill some goals. For instance, the process Order is defined to fulfill the goal IncreaseSales. The realization of this later requires the achievement of three goals: IncreaseCustomers, ProductAvailable and IncreaseShops. Note that relationships among goals are represented in the ontology by means of predicates. For example, the realization of the goal IncreaseSales requires the availability of the products in the stock, and either increasing the customers or shops. This knowledge is represented in the ontology through the predicate :(Require (IncreaseSales, IncreaseCustomers)∨Require (IncreaseSales, IncreaseShops))∧Require (IncreaseSales, ProductAvailable).

On the other hand, the fulfillment degree of each goal (*e.g.*, IncreaseSales) is measured through an indicator (*e.g.*, Turnover). For example, the values produced by the indicator Turnover are calculated through the Formula: Σ (price * quantity sold). This indicator could be analyzed according to the product perspective during a period of time, which allows the decision maker to identify both the most-in-demand and the least-in-demand product. Additionally, when the target value of the indicator is not reached, the decision maker notes an anomaly and looks for its origin by carrying out various analyses of this indicator according to different analysis axes. These axes provide the decision maker with detailed values of the indicator. The comparison of these detailed values with their corresponding target values allows the decision maker to locate the problem at certain levels of the analysis axis.



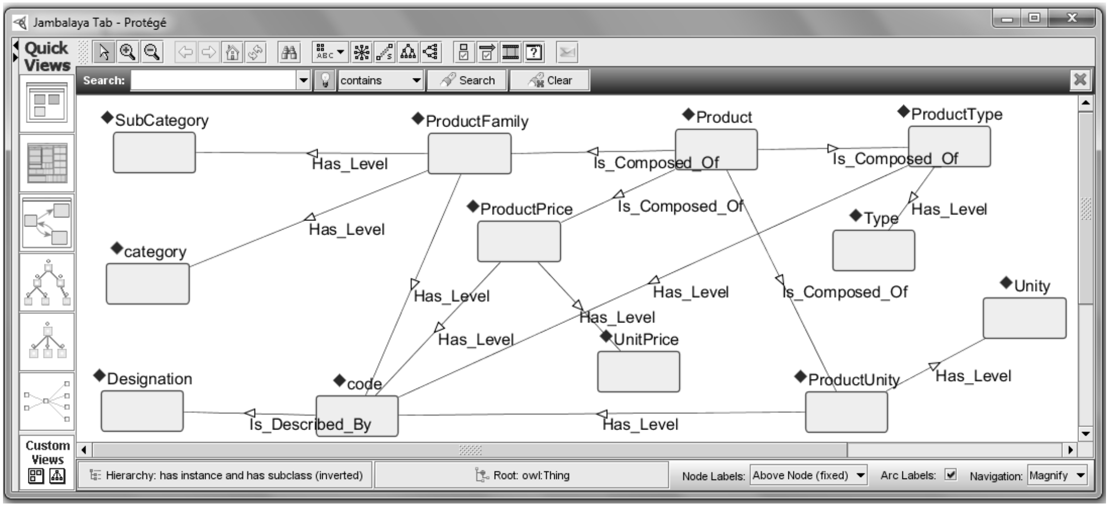
Search: contains

Hierarchy: has instance and has subclass (inverted)

Root: owl:Thing

Node Labels: Above Node (fixed) Arc Labels: ☑ Navigation: Magnify

Fig.6. An extract of the commercial domain ontology (the thesaurus part)



Hierarchy: has instance and has subclass (inverted)

Root: owl:Thing

Node Labels: Above Node (fixed) Arc Labels: ☑ Navigation: Magnify

Fig.7. An extract of the analysis axis concept (product)

Figure 7, shows various levels organized by analysis hierarchies for the analysis axis `product`. For example, for the hierarchy `family`, the indicator `Turnover` could be analyzed according to the level `category` of a product. This gives the decision maker detailed values of the `Turnover` by category of product. The comparison of each realized value with an estimated target value allows the decision maker to judge what category of product is not sold as expected.

Our domain ontology includes also inference rules. These later are defined by the domain experts (*i.e.* decision makers) and identified based on the semantic relationships among concepts of the domain. An inference deduces a set of conclusions from a set of premises, possibly under a given condition. The premises express constraints on the concepts, and designate expressions always known to be true by the domain experts. Therefore, they represent knowledge (stored in the ontology) that cannot be proven or contested. When an inference rule is applied, the reasoning engine matches the premises formulas with the knowledge stored in the ontology to infer a set of *new* knowledge, as conclusions, not explicitly stored.

In the following section, we show how the domain ontology is used to automate the elicitation of analytical requirements. For this, we use eight queries and two inference rules; they are to assist the elicitation of our analytical requirements template's components.

4 Using the ontology to elicit analytical requirements

As shown in Figure 1, our elicitation process uses four steps to extract the following components: the business processes to evaluate (step 1), the goals to fulfill (step 2) and the indicators that measure the achievement degree of the goals (step 3). These components are derived from the ontology, and then they are organized by the decision maker into a Processes-Goals-Indicators (PGI) graph. This graph shows explicit links between the derived components and allows a better understanding of the elicited requirements. The last step (step 4) extracts, for each indicator, all potential analysis axes, and then generates a set of candidate NL analytical queries (see Section 2) from which the decision maker selects a subset reflecting his/her needs as well as possible.

In what follows, we will refer to the ontology depicted in Figure 6 to illustrate each step of our elicitation process.

Step 1: business processes elicitation

In this first step, all business processes, in the ontology, that are associated to the current decision maker `D` with a relationship `control` are retrieved by executing the following query:

Query 1. List all Business Processes controlled by a given Decision Maker.

Suppose that the current decision maker is a `SalesManager`. The execution of Query 1 on our running example (Figure 6) returns three business processes: `Order`, `Delivery` and `AuctionOrder`. The next step involves selecting the adequate process that should be added to the PGI graph. If the sales manager chooses to evaluate the `Order` process, then we will get the part of the graph depicted in Figure 8, annotated with "S1".

The next step of our process elicits the goals that have to be fulfilled by the identified business process.

Step 2: goals elicitation

Retrieving goals is an iterative task. Each iteration input is a process `P` elicited through step 1, whereas its output is a set of goals that the process must fulfill. We identify two categories of goals. The first category, we call *High level* goals, results from the execution of the following query:

Query 2. List all Goals that must be fulfilled by a given Business process.

The decision maker chooses a subset from the resulted goals. Then the selected goals are added to the PGI graph. For example, for the selected process "order", Query 2 returns the high level goal IncreaseSales added to the PGI graph (see figure 8) and annotated with "S2". In the case of a complex goal, it is necessary to decompose it into more concrete sub-goals, we call *operational* goals (second category). These later are deduced using the following inference rule:

Rule 1. If a process P is defined to fulfill a goal G, and G requires G1, then P must fulfill G1 to achieve G.

The application of Rule 1 returns all goals that are indirectly related to the process P via the transitive closure of the relation require obtained by the subsequent Rule:

Rule 2. If a goal G requires G1, and G1 requires G2, then G requires G2.

The first application of Rule 1 gives goals at a second level of abstraction, from which the decision maker D chooses those that better reflect his/her requirement. For example, in the case of the goal IncreaseSales, the application of Rule 1 will propose the following goals: IncreaseCustomers, ProductAvailable and IncreaseShops. Figure 8 shows the two selected goals annotated with "S2.1". Next, for each selected goal, Rule 1 is applied again (second application) to produce goals at a lower level of abstraction. This rule is applied as many times as there are goals in the ontology related by a require relationship. For instance, the result of the second application of Rule 1 is annotated in Figure 8 with "S2.2".

Step 3: retrieve indicators

The achievement of each goal in the PGI graph is measured through the difference between the attained value calculated by the indicator and the target value.

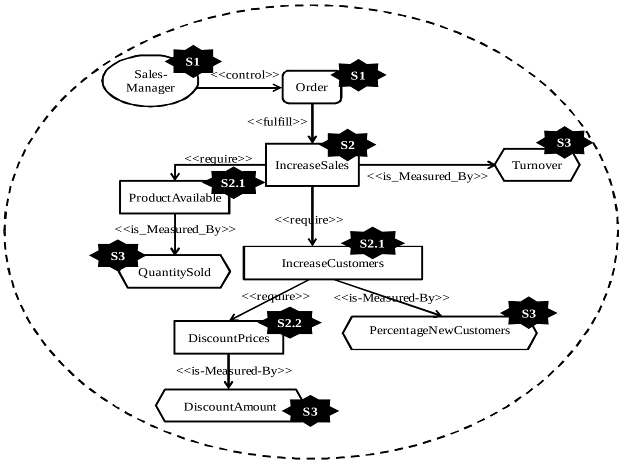


Fig.8. Part of the constructed PGI graph

For each goal G the corresponding indicator I is retrieved from the ontology by executing Query 3:

Query 3. Select the Indicator that measures a given Goal

The formula needed to calculate the attained value of an indicator I is then obtained from the ontology by Query 4:

Query 4. Select the Formula of a given Indicator.

The annotation text "S3" in Figure 8 highlights the output of this step.

Step 4: analytical queries generation

This step identifies, for each tuple (P,G,I) in the PGI graph, the relevant analysis axes, and uses the simple NL query format (*cf.* section 2) to generate the analytical queries. The analysis axes are deduced from the ontology by executing the following query:

Query 5. List all Analysis Axes of a given Indicator.

We propose the template (*Process, Goal, Indicator, Choice of analysis axes*) to represent the result of Query 5. This template assists the decision maker by proposing all possible analysis axes from which he/she selects those relevant to his/her analysis tasks. Table 1 shows the result of Query 5 for the business process Order. Note that the analysis axis Time is checked by default since any analysis must be realized during a given period of time.

TABLE 1. Template for choosing the analysis axes related to the process order. P(Business Process), G (Goal), I (Indicator) and ? (choice)

P	G	I	? analysis axes	
Order	IncreaseSales	Turnover	?	SalesPerson
			?	Customer
			?	Product
			✓	Time
	ProductAvailable	QuantitySold	?	Product
			?	store
			✓	Time
	IncreaseCustomers	PercentageNewCustomers	?	SalesPerson
			?	Customer
			✓	Time
	DiscountPrices	DiscountAmount	?	Product
			✓	Time

The output of this step is a set of tuples of the form T_i = (P, G, I, *A*) where *A* indicates the set of selected analysis axes for the tuple (P, G, I). Suppose that the decision maker has chosen the tuple T_1 = (order, IncreaseSales, turnover, {SalesPerson, Customer, Product, Time}.

The next step involves the derivation of the analytical queries for each tuple T_i. To do so, for each selected analysis axis A ∈ *A* in T_i, the corresponding analysis hierarchies *H* are deduced from the ontology by the following query:

Query 6. List all analysis hierarchies of a given analysis axis.

Then, for each analysis hierarchy H ∈ *H*, the corresponding analysis levels and analysis attributes are derived by the subsequent queries:

Query 7. List all analysis level of a given analysis hierarchy.
Query 8. List all analysis attribute of a given analysis level.

The resulted analysis levels and analysis attributes of the queries 7 and 8 are combined with the analysis axis A ∈ *A* and the indicator I in the tuple T_i, to generate an analytical query. For example, for the case of the analysis axis Product ∈ *A* in T_1, the execution of queries 6 and 7 will generate the analytical queries illustrated by Table 2. Note that, for each hierarchy, the analytical query including the first level of analysis, i.e. code, is selected by default since every analysis hierarchy must contain at least a root called the first analysis level.

The last step of our analytical requirements elicitation process organizes the elicited information according to our requirements template. Figure 9 shows a part of the graphical representation of the resulted requirements.

TABLE 2. Choice of hierarchies and analytical queries for the analysis axis Product ∈ T1

? Hierarchy		Choice of analytical queries	
?	ProductUnity	✓	Analyze the turnover by product's code and designation
		?	Analyze the turnover by product's unity
?	ProductPrice	?	Analyze the turnover by product's code and designation
		?	Analyze the turnover by product's price
?	ProductFamily	✓	Analyze the turnover by product's code and designation
		?	Analyze the turnover by product's category
		?	Analyze the turnover by product's sub-category
?	ProductType	✓	Analyze the turnover by product's code and designation
		?	Analyze the turnover by product's type

5 Conclusion and future work

In this paper, we have proposed an ontology driven approach for analytical requirements elicitation. The formal aspect of the ontology automates the reasoning about the decision-making knowledge, which allows systematic requirements elicitation. In addition, the proposed approach enables a more intuitive requirements analysis process, starting from the identification of the business processes to evaluate, followed by the identification of the goals that must be fulfilled and their decomposition from high-level goals into more concrete sub-goals. In turn, the goals provide for the identification of the indicators and their associated formulas that measure the achievement degree of the defined goals. The last step of our requirements elicitation generates a set of analytical queries that can be used by the decision maker to carry out different analysis tasks when his/her goals are not met. The output of the proposed approach is a template that ensures the traceability between the elicited requirements elements. This traceability ensures a better understanding of the specified requirements, and facilitates the maintenance of the MD model when changes in the goals occur.

Although the issue of completeness of our ontology, for a given domain, still needs more investigation, we can have several semi-automated techniques to do it. More precisely, many existing ontologies for various domains represented in standardized OWL language, can be easily used to improve the population of our ontology. Furthermore, some existing approaches that extract ontological concepts and their relationships from NL documents can be used to extract the concepts goals, indicators and business processes from available organizations' Balanced Scorecard. Furthermore, we are currently working on finalizing a supporting tool for our approach. Our immediate future work comprises the evaluation of the obtained results within a case study. As a long term research axis, we plan to study how the evolution of the decision-making ontology may be handled.

Meta-data	**TITLE**	Order process analysis		
	SUMMARY	This requirement analysis the performance of the process order according to…		
	UPDATE DATE	08/03/2011		
	AUTHOR	Fahmi Bargui		
	ACTOR	Sales Manager		
	PROCESS	Order		
	Goal *1*: Increase the sales	INDICATOR 1	LABEL	Turnover
			FORMULA	Σ(price * quantity sold)
			TARGET	100.000 TND
			ANALYTICAL QUERIES	1) Analyze the turnover by product's code and designation. 2) Analyze the turnover by product's category. 3) Analyze the turnover by product's sub- category. 4) Analyze the turnover by month.

Fig.9. Extract of the specified analytical requirements

6 Reference

[1] W. Inmon, *Building the Data Warehouse.* Wiley & Sons, 2002.

[2] E. Yu, "Towards Modeling and Reasoning Support for Early-Phase Requirements Engineering". In *Proc. 3rd IEEE Int. Symposium on Requirements Engineering*, USA, pp. 226–235, 1997.

[3] J. Feki, Y. Hachaichi, " Conception assistée de MD: Une démarche et un outil ". Journal of Decision Systems, vol. 16, no. 3, pp. 303-333, 2007.

[4] P. Giorgini, S. Rizzi, M. Garzetti, "GRAnd: A goal-oriented approach to requirement analysis in data warehouses". Journal of Decision Support Systems, vol. 45, no.1, pp. 4-21, 2007.

[5] J.-N. Mazón, J. Pardillo, J. Trujillo, "A Model-Driven Goal-Oriented Requirement Engineering Approach for Data Warehouses". ER Workshops, LNCS vol. 4802, pp. 255–264, 2007.

[6] J. Mylopoulos, L. Chung, E. Yu, "From Object-Oriented to Goal-Oriented Requirements Analysis". *Communications of the ACM*, vol. 42, no. 1, pp. 31–37, 1999.

[7] F. Bargui, H. Ben-Abdallah, J. Feki, "Analyse des besoins analytiques : une approche dirigée par les buts et basée processus métiers ". In *Proc. 14th IBIMA*, Turkey, pp. 1725-1733, 2010.

[8] F. Bargui, J. Feki, H. Ben-Abdallah, "A natural language approach for data mart schema design". In the *9th Int. ACIT*, Tunisia, 2008.

[9] M. Mard, R.R. Dunne, E. obsborne, J.S. Rigby, "Driving your company's value: strategic benchmarking for value". Wiley & Sons, 2004.

[10] F. Bargui, H. Ben-Abdallah, J. Feki, "Multidimensional Concept Extraction and Validation from OLAP Requirements in NL". In *Proc.* the IEEE NLP-KE, China. pp. 199-206, 2009.

[11] A. Nabli, J. Feki, F. Gargouri, "An Ontology Based Method for Normalisation of Multidimensional Terminology", SITIS 2006, LNCS 4879, pp. 235–246, 2009.

SESSION

DATA MINING

Chair(s)

TBA

Mining Actionable Knowledge for Domain-Driven and Customer-Centric Decision Support

Tong Sun[1], **Wei Peng**[1], **and Tao Li**[2]
[1]Xerox Innovation Group, Rochester, NY, USA
[2]School of Computer Science, Florida International University, Miami, FL, USA

Abstract – *It has been increasingly critical for businesses to become more customer-centric and more responsive for customer needs. The "voice of customer" (VOC in short) is a general term to describe the stated and unstated customer needs that can be captured through various customer touch-points: sales meetings, surveys, interviews, focus groups, customer support call-center, social media sites, etc.. Most existing approaches in analyzing VOC focus on mining patterns of significant technical interestingness, which are mainly concerned with the data mining method used, but not necessarily meaningful to business problems. To discover knowledge from VOC dataset that can be used for taking actions to business advantages, we develop a hybrid framework that tightly integrates domain knowledge and decision making principles with data-driven approaches. The proposed framework provides following actionable insights: (1) uncover the main themes and the evolving trends of customer needs; (2) identify the gaps between ever-changing customer needs and the service/product provider strategy and development organization structure; (3) effectively prioritize both customer needs and service/product offerings simultaneously by taking into account of customer demography and purchase history, via a novel Semantic Enhanced Link-based Ranking (*SELRank*) algorithm as described in this paper. This analytical framework that inherently embodies domain specific knowledge has been successfully applied on Xerox Office Group semi-structured VOC dataset to support customer-centric business decisions in both short-term and long-term service/product planning, and adaptive transformation of development organizations.*

Keywords: knowledge mining, decision support, domain-driven business applications

1 Introduction

The "voice of customer" (VOC in short) [8] is a general term to describe the stated and unstated customer needs or requirements that can be captured through various customer touch-points: sales meetings, surveys, interviews, focus groups, customer support call-center, product reviews, field reports, social media sites etc.. Traditionally, marketing organization has the sole responsibility for capturing customer needs and converting them into service/product requirements, which tends to isolate development personnel from gaining a first-hand understanding of customer needs. As a result, customer's real needs can become somewhat abstract to the development teams. However, VOC datasets are usually diverse and in large quantity. To maximize the benefits of direct access to customer requirements, it is extremely important to mine VOC datasets with a goal of discovering actionable knowledge that the development teams can act upon to their business advantages. The actionable knowledge is used to influence a decision making process and can be applied to create effective actions that produce desired results (e.g. answer to a question, a good decision, and a valid conclusion) [18].

Several data-driven approaches are available [3,19,7] for synthesizing and mining customer textual data for information retrieval purpose. None of them directly address the decision support process that relevant to these customer needs or requirements due to lack of integrating domain knowledge. Although there is an increasingly focus of integrating the data mining approaches with decision process in recent literatures [12,2,10], some phenomenal difficulties and a large gap still exist between academic deliverables of data mining algorithms/tools and business expectations [3]. A large portion of the mined patterns are neither transparent nor interesting enough to support business decision-making and operation.

In this paper, we develop a hybrid framework that integrates domain knowledge with data-driven mining algorithms to analyze the VOC to support the business decisions. For instance, the decision makers in service/product marketing and planning organization need to understand the main themes and evolving trends in customer needs, then timely identify the emerging new features to meet customer needs, and eventually verify the alignment of its service/product strategy with changing customer needs. For service/product development organization that is usually structured based on functional features, the decision makers need to prioritize the customer requirement to drive engineering development process, and also track changing customer needs and ensure their organization is adaptive to emerging new or cross-organizational. Another key contribution is the novel Semantic Enhanced Link-based ranking algorithm (or SELRank) that we propose in this paper to prioritize customer feature requests by analyzing the explicit and implicit link structures in VOC dataset.

2 The Integrated Framework

Figure 1 illustrates the high level overview of the proposed hybrid VOC analytic framework. It encompasses the following process steps and components.

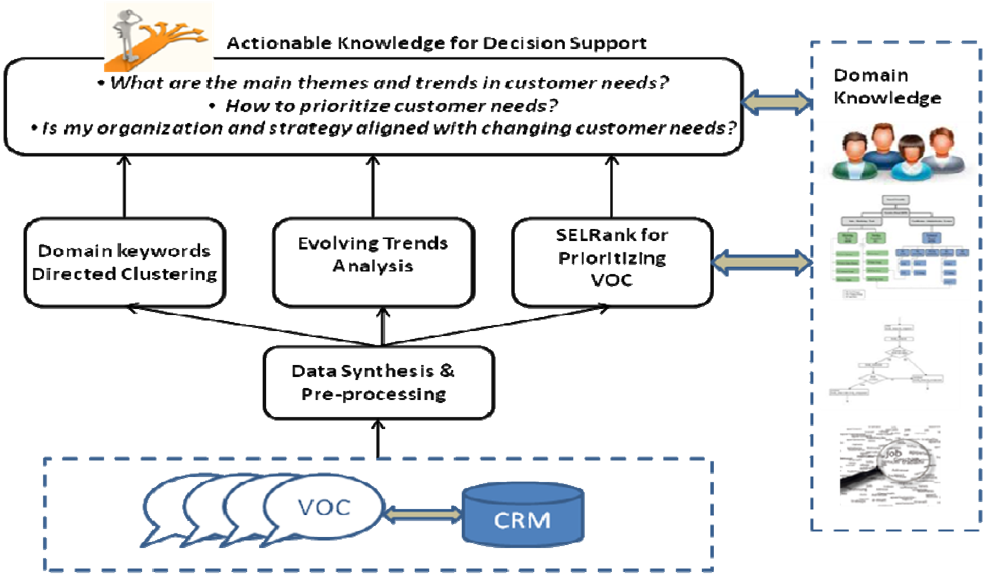

Actionable Knowledge for Decision Support
• What are the main themes and trends in customer needs?
• How to prioritize customer needs?
• Is my organization and strategy aligned with changing customer needs?
Domain Knowledge
Domain keywords Directed Clustering
Evolving Trends Analysis
SELRank for Prioritizing VOC
Data Synthesis & Pre-processing
VOC
CRM

Figure 1. The Integrated Analytical Framework

2.1 VOC data pre-processing

First, most of VOC data are unstructured or semi-structured textual data. Typically, the customer requests in VOC data contain at least the following meta-data fields: a textual description about a desired feature or a current problem, the associated service(s) or product(s), and the customer identification information. Customer demography information (e.g. market segment, business size), and its historical transactions (e.g. service/product purchased or owned) can be retrieved through the customer identification information. These raw VOC textual data needs to be pre-processed using text processing techniques such as skipping the stop words and extracting keywords using TFIDF (Text Frequency Inverse Document Frequency) [22]. The preprocessed VOC data are them transformed into Semantic VOC by mapping into a lower dimensional concept space using LSI (Latent Semantic Indexing) [14].

2.2 Keywords directed clustering

The Semantic VOC data are then categorized or clustered into key feature groups or themes using unsupervised clustering methods [16,23] , or domain knowledge based clustering [15,5]. The domain knowledge is usually captured from domain experts and can be presented in a formal ontology model, such as hierarchical taxonomy, business process rules, organization structures and roles, data labels or keywords. In this paper, the domain knowledge includes hierarchical service/product taxonomy, market segments and their relative importance scores according to business strategy, bags of distinguishing keywords for existing features and organization structures. Here we propose the two-staged *keywords directed clustering* method to obtain feature groups or themes by incorporating domain knowledge.

For instance, the “security” feature group can be described by a collection of keywords, such as “authentication”, “password” and “login”, etc. These domain keywords can help us obtain the preliminary first-stage clusters such that:

$$cid_i^j = \begin{cases} 1 & \text{if } \exists w \in T_i \text{ and } w \in D_j; \\ 0 & \text{otherwise.} \end{cases} \qquad (1)$$

cid_i^j indicates whether the *i*-th requirement belong to the j-the feature group or not, with the value 1 or 0. *w* is a word, T_i is the set of words in the i-th requirement, and D_j is the set of keywords that domain experts provide for the j-th feature group. In this first-stage classification, the prior word distribution among feature groups is formed, and one requirement may belong to multiple feature groups.

In the second-stage, the clustering can be refined by training the first-stage results and then regrouping the requirements. We use the Naïve Bayes classifier [17] in the framework. The words with high information gains are selected from data. The probability that the i-th requirement R_i belongs to j-th feather group C_j is:

$$P(C_j \mid R_i) = \frac{1}{a} P(C_j) \prod_{q=1}^{l} P(w_q \mid C_j), \qquad (2)$$

where a is the joint probability of all words, P(C_j) is the percentage of the j-th feature group in the training data, and $P(\omega_q | C_j)$ is the probability distribution of the word ω_q in C_j . In the case that a large amount of requirements belongs to two feature groups with a high probability, an alert may need to fire especially for adjusting feature-driven development structure to handling the overlapping features. We will discuss the decision support implication in later section.

2.3 Evolving trend analysis

The feature groups or themes resulting from above clustering method are weighted by summing up the importance of the constituent requests, according to the business impact of associated customers (e.g. their market sizes, historical transactions on services/products). If we take into account the time when the requirements emerge, the importance of each feature group varies along the time. The importance trend of all feature groups can be plotted for people to capture the customer's evolving or changing needs and the key themes timely. For plotting the evolving trend, the time is quantified into a number of periods, and the importance of each requirement is provided. Given the requirements *i*, the importance vector *v* with v_i denoting its importance value, the time vector *t* with t_i indicating its reported time, and the feature assignment vector *c* with c_i denoting the feature group that it belongs to. Therefore, the importance of the feature *l* in the period *Q* is:

$$I_{l,Q} = \sum_{\{i | t_i \in Q \text{ and } c_i == l\}} v_i. \qquad (3)$$

2.4 Prioritize VOC Data

In product requirement engineering and design practices, understanding the relative importance of a customer request on certain product features is extremely critical and has a direct impact on the effective prioritization in the development process. By examining the explicit and implicit link structures in VOC data between requests, associated customer, and mentioned products, we propose a novel SELRank (Semantic Enhanced Link-based Ranking) algorithm to prioritize the VOC data. The detail algorithm is described in next section.

2.5 Actionable decision support knowledge

Based on the above data-driven analytics, several data patterns are uncovered with regards to VOC data: the clustering, the evolving trends, and the request's ranking. In order to put these data patterns in a proper business context, we work with domain experts in marketing and planning organization, and service/product feature development team. Three types of relevant business decision questions are identified: (a) What are the main themes and evolving trends in customer needs? (b) How to prioritize the customer needs according to relative importance of both customers and services/products? (c) Are the service/product strategy and development organization structure aligned with ever-changing customer needs?

A decision support interface is developed and includes: VOC Themes and Trending Chart, Alerts for emerging new features, VOC Search Clustering Analysis. The detail description is provided in Section 5.

3 Semantic-Enhanced Link-Based Ranking (SELRank) Algorithm

In Section 2.3, we rank the feature groups of customer requests based on the relative importance of the associated customers. This high-level ranking gives us the overview of the main themes in VOC. However, there are needs to prioritize individual requests either within a feature group or cross feature groups.

We propose a novel semantic enhanced link-based ranking (SELRank) algorithm for rating customer requests by considering not only the importance of the customers who ask for them, but also the importance of the products they are related and how *representative* they are. Meanwhile, the relative importance scores of the products among all customer requests are also calculated according to the importance of their related requests and related products. The *representative* request is the request semantically similar to many other requests. The representative request is important because solving it may help to solve other requests. It is straightforward that the request related to the important products and important customers is important, and the product associated with important requests and important customer is also important. In general, the intuition behind our proposed ranking algorithm is **a request is important if it is strongly linked with many other important requests, targeted to many important products, and asked by many important customers.**

3.1 Link structures in VOC data

As we describe in Section 2.1, each customer request in VOC data is generally about a Customer provides Request(s) for new or enhanced features about certain Product(s). Therefore, the explicit links (similar to hyperlink in a web page) are the associations between Request and Customer, and Request and Product. In addition, other semantics help to provide additional link-based relationships that are implicit links, such as (1) the hidden links between customer requests based on their semantic similarity; (2) the semantic links between services/products based on the hierarchical domain taxonomy; (3) the semantic links between customers based on their related market segments. Figure 2 illustrates a generic conceptual model for the customer requests in VOC data. The link structure embedded in the VOC data exhibits a set of inter-related and intra-related complex networks (such as product-to-product network, request-to-request network, customer-to-customer network, product-to-request network, request-to-customer network), where there are three types of nodes: Product, Customer, and Request. They could have the initial assigned importance scores. In our experiment, we have prior importance scores assigned to different customers based on their market segments and historical transactions.

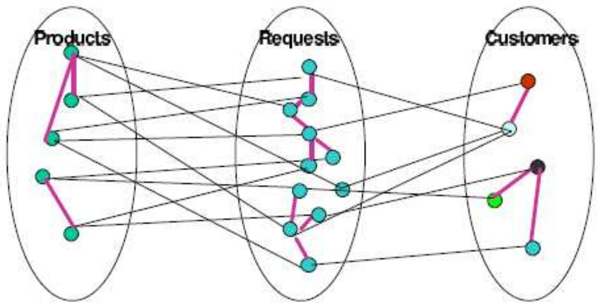


Figure 2. The Link Structures in VOC data

The links between requests are weighted with semantic similarity. The links between products can be weighted with the "conceptual distance" [21]. It can be derived from the domain hierarchical taxonomy graph for the family, types and individual service/product. The conceptual distance between two service/product nodes can be the number of the intervening nodes. These intervening nodes form the '*shortest path*' between the two nodes. We calculate the shortest path by adopting the breadth-first search that iteratively explores one node's neighbors and its neighbors' neighbors etc. until the other node is found. The nodes that are "conceptually closer" to each other in the domain hierarchical taxonomy graph should have more similar functionalities and features. The semantic links between customers can be also weighted by using conceptual distance.

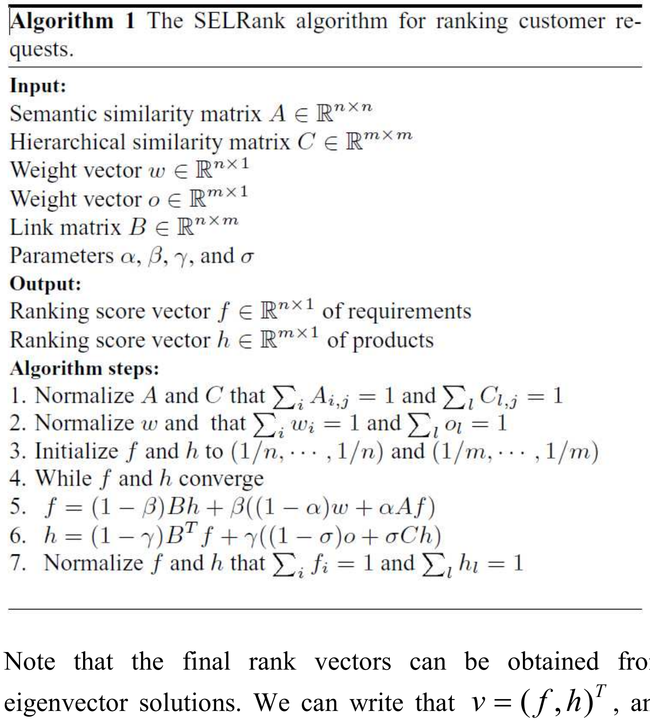
3.2 Algorithm description

In the following, we will describe our SELRank algorithm. Given the similarity matrix *A* with $A_{i,j}$ be the semantic similarity between the requests R_i and R_j; *C* with $C_{l,j}$ be the hierarchical semantic similarity between the products U_l and U_j, which can be easily derived from their conceptual distance; the vector *w* and w_i be the weight of the customer (the importance of customer) who raises the request R_i; the vector *o* with o_l be the weight of U_l (the importance of product defined by product strategy experts); the vector *f* with f_i be the ranking score of the request R_i; the vector *h* with h_l be the ranking score of the product U_l; and the matrix *B* with $B_{i,l}$ be 1 if R_i is linked with U_l and 0 otherwise. Then,

$$f_i = (1-\beta)\sum_{U_l\leftrightarrow R_i} h_l + \beta((1-\alpha)\frac{w_i}{\sum_q w_q} + \alpha\sum_j \frac{A_{i,j}f_j}{\sum_k A_{i,k}}),$$

$$h_l = (1-\gamma)\sum_{U_l\leftrightarrow R_i} f_i + \gamma((1-\sigma)\frac{o_l}{\sum_q o_q} + \sigma\sum_j \frac{C_{l,j}h_j}{\sum_k C_{l,k}}),$$

where $U_l \leftrightarrow R_i$ means that the product U_l is linked with the request R_i. β and γ are parameters to adjust the influence of the importance propagation along inter-links and intro-links. The links between the same types of nodes are called intra-links, for instance, the links among requests. The links between different types of nodes are called inter-links, for instance, the links between requests and products. α and σ are another parameters to adjust the influence of the *stationary preference* (initial importance) [1] of requests and the importance flow along the semantic links between requests, 0 < β, γ, α, σ < 1, which can be defined by domain experts. In our experiments, domain experts give more influence to importance propagation along intra-links than inter-links. More influence is given to semantic links than stationary preference. We can see that SELRank algorithm is similar to HITS [13] in the way that *f* and *h* continuously reinforce each other until they converge. The detail algorithm is listed in Algorithm 1.



Algorithm 1 The SELRank algorithm for ranking customer requests.

```
Input:
Semantic similarity matrix $A \in \mathbb{R}^{n\times n}$
Hierarchical similarity matrix $C \in \mathbb{R}^{m\times m}$
Weight vector $w \in \mathbb{R}^{n\times 1}$
Weight vector $o \in \mathbb{R}^{m\times 1}$
Link matrix $B \in \mathbb{R}^{n\times m}$
Parameters $\alpha$, $\beta$, $\gamma$, and $\sigma$
Output:
Ranking score vector $f \in \mathbb{R}^{n\times 1}$ of requirements
Ranking score vector $h \in \mathbb{R}^{m\times 1}$ of products
Algorithm steps:
1. Normalize $A$ and $C$ that $\sum_i A_{i,j} = 1$ and $\sum_l C_{l,j} = 1$
2. Normalize $w$ and that $\sum_i w_i = 1$ and $\sum_l o_l = 1$
3. Initialize $f$ and $h$ to $(1/n, \cdots, 1/n)$ and $(1/m, \cdots, 1/m)$
4. While $f$ and $h$ converge
5.   $f = (1-\beta)Bh + \beta((1-\alpha)w + \alpha Af)$
6.   $h = (1-\gamma)B^T f + \gamma((1-\sigma)o + \sigma Ch)$
7.   Normalize $f$ and $h$ that $\sum_i f_i = 1$ and $\sum_l h_l = 1$
```

Note that the final rank vectors can be obtained from eigenvector solutions. We can write that $v = (f, h)^T$, and $v = Dv$, where

$$D = \begin{bmatrix} \beta(1-\alpha)w \cdot 1^T + \beta\alpha A & (1-\beta)B \\ (1-\gamma)B^T & \gamma(1-\sigma)o \cdot 1^T + \gamma\sigma C \end{bmatrix}.$$

From the above equation, we can view D as the similarity adjacent matrix of the network composed of both products and requests. The links in the network could be request-to-request links, product-to-produce links and request-to-product links. The entry value in matrix D can be regarded as the corresponding link weight. Therefore, the SELRank algorithm can be regarded as the modified PageRank algorithm [20], where the links are both hyperlinks and semantic links weighted by D. We can observe that v is the eigenvector centrality of a weighted graph D. v can be obtained by calculating the principle eigenvector corresponding to the greatest eigenvalue. The algorithm therefore converges as fast as PageRank algorithm.

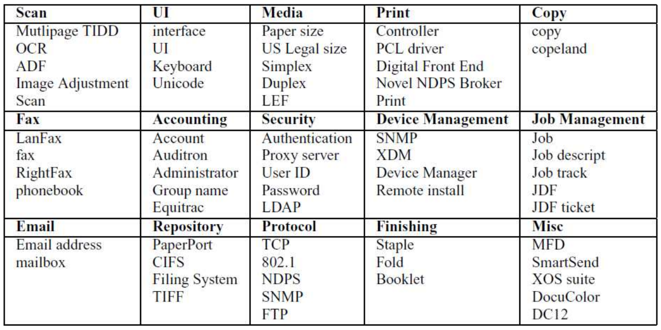
3.3 Comparison with existing algorithms

As we know, link analysis algorithms play key roles in web search systems. The HITS algorithm [13] relies on query-time processing to find the hubs and authorities that exist in a sub-graph of the web consisting of both the results to a query and the local neighborhood of these results. Google's PageRank [20] pre-computes a ranking vector that provides a-priori "importance: estimates for all the pages on the web. This vector is computed once, offline and is independent of the search query. At query time, these importance scores are used in conjunction with query-specific IR (information retrieval) scores to rank the query results. There are several enhanced PageRank algorithms being developed recently, such as a weighted PageRank [11], two-layer PageRank [24], hierarchical PageRank [25], and the topic-sensitive PageRank [9]. All above methods only consider the explicit graph-topological links (either flat or hierarchical networks) residing in a web page, and most of them generate a single page-ranking vector. The linguistic-based topic structure used in [9] is only used for biasing the ranking scores based on different topics, and it does not provide any additional "semantic implicit link" structure into the web page. Although multiple ranking vectors can be computed by [9], these ranking vectors are still for web pages with biasing by different topics. None of these existing ranking algorithms are sufficient to effectively handle the prioritization of customer requests in VOC data. This is because the analysis of VOC data is a very domain-driven problem, and also the link-based relationships embedded in them are well beyond the explicit hyperlinks and involve much more complex inter-related networks.

4 Business Decision Support

In this section, we discuss the decision support application scenarios through our case studies on the Xerox Office Group VOC dataset, which contains 1878 feature enhancement requests from existing customers for the past six years. The total number of Xerox Office Group (XOG) products referenced by these requests is 83. This VOC dataset is a synthesized collection of customer requests captured via multiple touch-points, such as call center records, sales meetings, focus group studies, emails, marketing events or survey. Each data entry with the reported timestamp describes a customer request for enhancement for one or more XOG products. The relevant customer information, such as market segment, historical purchase and number of XOG products owned, is also available.

The domain knowledge we used in our XOG case studies include XOG product taxonomy tree graph and its 15 existing development teams, each of which is structured based on a key "functional feature". For instance, a UI team is responsible for developing all UI-related features. Table 2 illustrates on these 15 teams and associated bags of descriptive keywords. Generally, the feature teams include "Scan", "UI", "Print", etc. and parts of corresponding keywords are listed below for each feature team. Other domain knowledge include the initial importance score of customers based on marketing strategy, market segments and their historical purchase transactions.



Scan	**UI**	**Media**	**Print**	**Copy**
Mutlipage TIDD OCR ADF Image Adjustment Scan	interface UI Keyboard Unicode	Paper size US Legal size Simplex Duplex LEF	Controller PCL driver Digital Front End Novel NDPS Broker Print	copy copeland
Fax	**Accounting**	**Security**	**Device Management**	**Job Management**
LanFax fax RightFax phonebook	Account Auditron Administrator Group name Equitrac	Authentication Proxy server User ID Password LDAP	SNMP XDM Device Manager Remote install	Job Job descript Job track JDF JDF ticket
Email	**Repository**	**Protocol**	**Finishing**	**Misc**
Email address mailbox	PaperPort CIFS Filing System TIFF	TCP 802.1 NDPS SNMP FTP	Staple Fold Booklet	MFD SmartSend XOS suite DocuColor DC12

Table 2: The domain key words of FER tasks for 15 functional teams.

Based on the data patterns uncovered using methods discussed in Section 2 and 3, we work with domain experts and develop the following decision support interface.

4.1 Main themes and trends in VOC data

In order for market analysts and planner, and other decision maker to effectively track and monitor the evolving customer needs, we provide an aggregated view on the main themes and trends based on the data-patterns extracted from the diverse and vast amount of VOC datasets.

Our proposed domain keywords directed clustering method successfully disseminates requests into these known domain feature group as verified by domain experts. The experimental results shown in Table 3 are evaluated to be meaningful in the XOG feature team context. The first column is the feature group id, and second one is the feature name, and the third one is the number of requests in the corresponding feature group while the last one is the normalized importance scores summed to 1. We observe that the "security" feature group obtains the highest important score. It is verified by domain experts that the security is the main issue that they will look into.

Fid	Name	Req Num	Importance
1	Scan	366	0.1173
2	UI	116	0.0877
3	Media	126	0.0487
4	Print	245	0.1417
5	Copy	73	0.0301
6	Fax	77	0.0548
7	Accounting	172	0.0720
8	Security	210	0.1451
9	DeviceManagement	25	0.0342
10	JobManagement	156	0.0615
11	Email	21	0.0154
12	Repository	35	0.0078
13	Protocol	108	0.0774
14	Finishing	32	0.0084
15	Misc	116	0.0996

Table 3: The experimental results on Main Themes with importance score uncovered by our domain keywords directed clustering method

With considering the time, the evolving trend of the above 15 feature groups shown in Table 2 can be obtained and captured in Figure 4. Time is quantified into 14 periods onto X-axis. There are 15 important curves corresponding to feature groups. The sum of importance score of a feature in a periods is the vertical gap between this feature curve and the underneath feature curve in this period. From Figure 4, we can find that for example, "Security" feature and "Misc" feature (they are the curves with * marks) are increasingly becoming more and more important. This insight triggers the decision makers to tap into "Security" and "Misc" feature groups deeper to see what the emerging requirements are.

4.2 Alerts for emerging features

During the clustering process, we not only know how much possibility that a request belongs to a feature group, by also the pair-wise overlap between feature groups can also be calculated. In the case that a large number of requirements belong to two feature groups with a high probability, a "cross-team focus group" may need to be formed especially for handling this overlap feature. To obtain the overlap, a threshold and a requirement-feature matrix with each entry indicating the probability of the corresponding requirement belonging to the corresponding feature group are needed. The matrix can be easily derived from the clustering method used in Section II-B. Given a threshold λ and the requirement-feature probability matrix P, the requirement-feature indication matrix F is a binary matrix where the entry $F_{i,j}$ is 1 if $P_{i,j} \geq \lambda$, and 0 otherwise. The overlap Q between the feature i and j is actually $F^T F$.

The symmetric feature group overlap matrix where the lower triangle is plotted in Figure 5 is obtained from the request-feature probability matrix with threshold λ set to 0.3. In the feature overlap matrix the big values are mostly along the diagonal. They are the number of requirements in 15 feature groups. The bold-line squares marked in the figure present two comparably bigger overlaps. They are the overlap between "Job" and "Print", and the overlap between "Protocol" and "Security". Therefore, this actionable insight alerts and triggers decision makers to form cross-feature task force to tackle these emerging and increasingly important overlapped features. This enables the organization timely adapts to the ever-changing customer needs, and ensure the better alignment to serve customers.

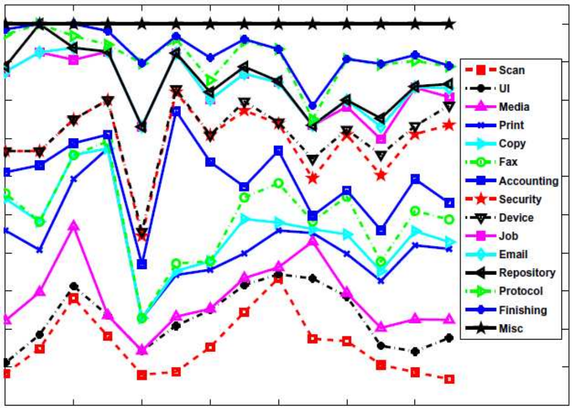


Figure 4. The evolving importance trends of 15 feature groups along 14 time periods

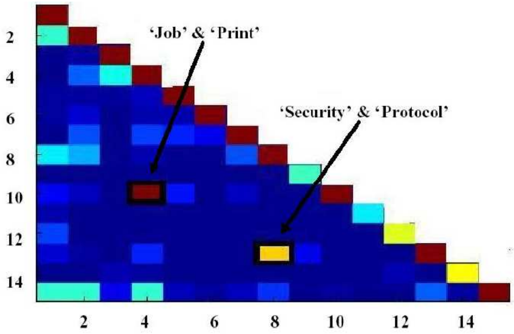


Figure 5. The plot of the feature overlap matrix

4.3 Search VOC applications

The importance of a customer request on product features is inherently related to many factors, such as emerging trends, the marketing strategy and the customer's influence and their historical transaction records. Our proposed SELRank algorithm can provide on-demand insights to support the decision makers in prioritizing the overall product development and verifying their product strategy with changing customer needs. Here we consider two scenarios for searching VOC with our SELRank algorithm: **Scenario 1**: Keyword based search on VOC data, in which a user types in a keyword query, and then a list of ranked customer requests and ranked relevant products are provided. **Scenario 2**: Product-based query on VOC data, in which a user types in one or more product names, and then a list of ranked customer requests linked to these products are provided respectively. Steps in this scenario are very similar to the

Scenario-1 except that the retrieved set only contains the requests that are directly linked with the query product(s).

We conduct user studies on both keyword based query and product based query. The ranking results from our SELRank method are evaluated to be very useful by domain experts and decision makers to facilitate the feature development and market research process and prioritize the handling of customer requirements.

5 Conclusion and Future Works

In this paper, a hybrid framework that integrates domain knowledge with data-driven mining algorithms is developed to analyze VOC data and extract actionable knowledge for decision support purposes. We have successfully applied the integrated framework in analyzing 6 years of VOC data for Xerox Office Group. The analytical frameworks includes domain keywords directed clustering method, trend analysis and in particular a novel SELRank algorithm, that is proposed in this paper to prioritize the customer requirements by exploiting the explicit and implicit link structures in the VOC data. By working with domain experts, we develop several business decision support interfaces that convert the data-patterns discovered from these data mining methods into actionable knowledge, such as identify the key themes, and emerging trends of customer needs, verify the alignment of service/product strategy and development organization structure, and prioritize the customer requirements for development planning process. There are several possible directions in the future. We need to be able to develop a method to easily acquire the domain knowledge and automatically integrate with our data analytical methods, since it is now a mostly time-consuming and error-prone manual process. Although we are able to support several decision support scenarios, additional methods are required in order to generate more complicate business decision rules from the mined data patterns.

6 References

[1] R. Baeza-Yates, P. Boldi, and C. Castillo. Generalizing pagerank: Damping functions for link-based ranking algorithms. *In Proceedings of ACM SIGIR*, pages 308-315. ACM Press, August 2006.

[2] L. Cao, P. Yu, C. Zhang, and Y. Zhao, *Domain Driven Data Mining*. Springer, 2009.

[3] L. Cao, Y. Zhao, H. Zheng, D. Luo, C. Zhang, E.K. Park. Flexible Frameworks for Actionable Knowledge Discovery. *IEEE Transactioins on Knowledge and Data Engineering*, vol. 22, No.9, September 2010.

[4] Clarabridge. Clarabridge content mining platform product. http://www.clarabridge.com/Products/ContentMiningPlatform/tabid/105/Default.aspx, 2007.

[5] A. Dayanik, D. D. Lewis, D.Madigan, V. Menkov, and A. Genkin. Constructing informative prior distributions from domain knowledge in text classification. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'06)*, pages 493-500, 2006.

[6] D. Zhou, S. Orshanskiy, H. Zha, and C. L. Giles. Co-ranking authors and documents in a hetergeneous network. In *IEEE International Conference on Data Mining (ICDM'07)*, pages 739-744, 2007.

[7] S. Gao, W. Wu, C.-H. Lee, and T.-S. Chua. A mform learning approach to robust multiclass multi-label text categorization. In *Proceedings of the 21st International Conference on Machine Learning (ICML'04),* page 42, 2004.

[8] A. Griffin and J. R. Hauser. The voice of customer. *Marketing Science*, 12(1):1-27, winter 1993.

[9] T.H. Haveliwala. Topic-sensitive page rank. In *Proceedings of the 11th International Conference on World Wide Web (WWW'02),* pages 517-526, 2002.

[10] Z. He, X. Xu, and S. Deng. Data mining for actionable knowledge: A Survey.

[11] X. Jiang, G. Xue, W. Song, H. Zeng, Z. Chen, and W. Ma, Exploiting pagerank analysis at different block level. In *Proceedings of Conference of WISE 2004*, pages 241-252, 2004.

[12] H. Kargupta, B. Park, D. Hershbereger, and E. Johnson. Collective Data Mining: A New Perspective toward Distributed Data Mining. *Advances in Distributed Data Mining*, AAAI/MIT Press, 1999.

[13] J. Kleinberg, Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604-632, 1999.

[14] T. Landauer, P.W. Foltz, and D. Laham. Introduction to latent semantic analysis. *Discourse Processes*, 25:259-284, 1998.

[15] J. Liu, W. Wang, and J. Yang. A framework for ontology-driven subspace clustering. In *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*, pages 623-628, 2004.

[16] J. Macqueen. Some methods for classification and analysis of multivariate observations, In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281-297, 1967.

[17] A. McCallum and K. Nigam. A comparison of event models for naïve bayers text classification. In *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*, pages 41-48, 1998.

[18] J. Narducci, What is actionable knowledge? *2002 Narducci Enterprises*.

[19] N.C. Romano, C. Bauer, H. Chen, and J.F. Nunamaker. The mindmine comment analysis, sense-making and visualization. In *Proceedings of the 33rd Hawaii International Conference on Systen Sciences*, page 1036, 2000.

[20] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. *In Proceedings of the 7th International World Wide Web Conference*, pages 161-172, 1998.

[21] R. Rada, H. Mili, E. Bicknell and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17-30, 1989.

[22] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513-523, 1988.

[23] S.C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 2:241-254, 1967.

[24] J. Wu and K. Aberer. Using a layered markov model for distributed web ranking computation. In *Proceedings of the 25th IEEE International Conference on Distributed Computing Systems (ICDCS'05)*, pages 533-542, 2005.

[25] G.-R.Xue, Q. Yang, H.-J. Zeng, Y. Yu, and Z. Chen. Exploiting the hierarchical structure for link analyss. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05),* pages 186-193, 2005.

Adopting Data Mining Techniques on the Recommendations of Library Collections

Shu-Meng Huang[a] , Lu Wang[b] and Wan-Chih Wang[c]
[a] Department of Information Management, Hsing Wu College, Taiwan
(simon@mail.hwc.edu.tw)
[b, c] Graduate School of Management Sciences, Tamkang University, Taiwan
(pheobemimilucky@hotmail.com)
Correspondence: Shu-Meng Huang[a]

Abstract *–In this research, the researchers explored not only the cluster of the readers with similar characteristics, but also the connection between the readers and the book collections of the library by using Data Mining techniques. By doing this, the library will be able to improve the interaction with its readers, and further increase the usage of library collections.*

The Modified Attribute-Oriented Induction (MAOI) method was introduced to deal with the multi-valued attribute table and further sort the readers into different clusters. Instead of using concept hierarchy and concept trees, MAOI method implemented the concept climbing and generalization of multi-valued attribute table with Boolean Algebra and modified Karnaugh Map, and described the clusters with concept description. On the other hand, the Chinese books in the library collections were classified into four groups with New Classification Science for Chinese Libraries (CCL). Not only the attributes of readers, but also the attributes of library collections borrowed by readers are included in the multi-valued attribute table. After the completion of induction, the reading preferences of the readers with the same characteristics can be learned.

Keywords: Data Mining , Recommendations , MAOI , Multi-Valued Attribute

I. INTRODUCTION

Potential readers seeking information in a library often face a daunting, time and energy-consuming task. Given the immense body of data gathered in modern libraries, it can be difficult for these readers to quickly sift through the mass of information to uncover what they need. This difficulty can affect how often, and how willingly, readers make use of library's vast resources. Some studies[1][2], have indicated that one of the key aspects of a library's service and marketing success is how well they actively provide information to readers by means of personal service technology based on readers' personal preferences and needs.

In recent years, many e-commerce websites have adopted personal recommender information systems in an effort to increase their interaction with customers to generate a higher rate of return patronage[3][4]. Take YouTube as an example, the "Recommended for you " column provides viewers information related to the videos they just browsed；And on www.amazon.com, they analyze the pages customers have browsed, then actively recommend the books the customers might be interested. Thus, the application of this concept in the library must be able to improve the relationship between the library and the readers.

In this research, a method of concept description in Data Mining was adopted — Modified Attribute-Oriented Induction[5], to sort the readers into different clusters. Each cluster includes readers with similar characteristics and preferences.

II. RELATED STUDIES

There are mainly two recommender systems applied in online stores [6]: for the merchandise that customers would consume more often, such as books or movies, and for the merchandise that customers would not consume so often, such as cars or computers. The first recommender system analyzes the consumption records of the customers to uncover the customers' preferences, and then provide advices. It usually adopts Data Mining technology, as well as personal service. But for the merchandise that customers don't buy so often, the advices based on the earlier consumption records may not achieve the expected results.

The situation of readers' making use of the collections of the library is similar to the consumption of the merchandise that customers buy more often. Therefore, many studies related to library recommendations adopt Data Mining technology to discover the relationship between readers and books.

2.1 Data Mining

When facing massive data, Data Mining provides powerful and effective tools to transform the data into useful information and knowledge[7][8].

Table 1 is a summary of the functions and technologies frequently used in Data Mining[9].

The Market Basket Analysis is also named Association Rule Analysis. This technology can explore the connections between attributes or objects. And it's an appropriate means to dig the relationships between readers and books.

But there are mass records in the library. When considering about analysis complication, execution efficiency and recommendation results, researchers usually perform the grouping of the readers before analysis. Yu-Ling Cheng (2002)、Jien-Hwa Tsao(2003)、Chang-Ting Yang(2007) classified the readers based directly on reader's department. Ching-Shium Chen (2000) 、Yuan-Jing Zhang (2001)、Chien-Yu Chen (2009) clustered the readers with Cluster Detection. Kuan-Hua Sun(2000) adopted Multilevel Association Rule Mining to discover the different characteristics of the readers who have different preferences. Yu-Ling Cheng (2003) adopted Memory-Based Reasoning to group the readers with the same background.

Some of the methods adopted in the above mentioned studies are complicated , meanwhile , some are brief. But most of them only can deal with single-valued data. When facing multi-valued data, lots of pre-processing work must be done for further analysis. In this research, Modified Attribute-Oriented Induction, which was proposed by Shu-Meng Huang (2010), was adopted, to induct and sort the multi-valued data, such as reader's department and year.

Table 1. Functions and technologies of Data Mining

Function / Technology	Classification	Estimation	Prediction	Affinity grouping	Clustering	Description
Statistics	V	V	V	V	V	V
Market Basket Analysis			V	V	V	V
Memory-Based Reasoning	V		V	V	V	
Genetic algorithm	V		V			
Cluster Detection					V	
Link Analysis	V		V	V		V
Decision Tree	V		V		V	V
Artificial Neural Network	V	V	V		V	

2.2 Modified Attribute-Oriented Induction (MAOI)

Attribute-Oriented Induction Approach (AOI) was proposed in 1991 [10]. It appears in a form of described Data Mining[11], and can deal with different kinds of knowledge rules efficiently, such as characteristic rules, discrimination rules, quantitative rules, and data evolution regularities [12]. This approach is one of the most classification scheme in Data Mining [13].

The basic concept and steps of AOI include [14]：
(1) Concept Hierarchy
(2) Attribute-Removal
(3) Concept-Tree climbing
(4) Vote propagation
(5) Attribute-Threshold Control
(6) Rule transformation

Though it's convenient making use of AOI to induct data into simple rule description, and some complicated procedures for data processing are eliminated as well. But different people make different Concept Hierarchies, and different definition leads to different results. The confidence would be low, if there is no apparent Concept Hierarchy between attributes.。Besides, AOI only can deal with single-valued attribute data [3]. Therefore Shu-Meng Huang (2010) combine the concepts of Boolean bit and simplified Karnaugh map with AOI, named MAOI, to deal with multi-valued attribute data. Figure 1 presents the steps of MAOI.

Table 2. is a database of high-frequent-crime areas in [3]. It's multi-valued attribute database. The researcher explained the steps of MAOI with it.

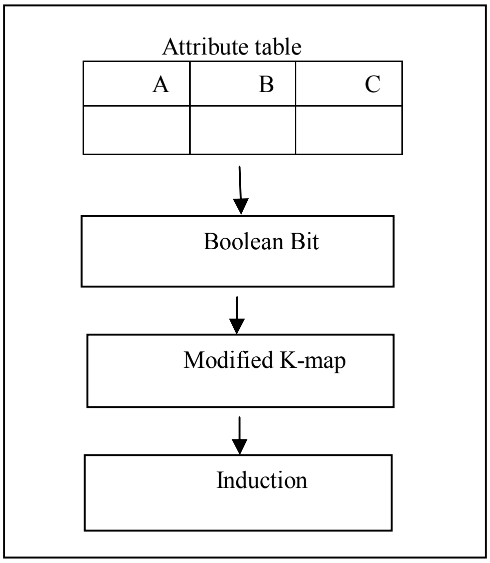


Figure 1. The induction steps of Modified AOI

2.2.1 Boolean Bit Transformation

To decide a value's Boolean bit, a cutting point must be defined first. Taking the mean value of the attribute as the cutting point for that attribute, all the values in that attribute can be transformed. That means, if a value is bigger than or equal to the cutting point, it's Boolean bit is 1. If a value is smaller than the cutting point, it's Boolean bit is 0. Table 3. presents the result after transformation.

Table 2. Database of high-frequent-crime areas

Area ID	Gender	Age	Education
1	<g$_1$,30> <g$_2$,70>	<a$_1$,20><a$_2$,30> <a$_3$,50>	<e$_1$,20><e$_2$,10> <e$_3$,40><e$_4$,30>
2	<g$_1$,45> <g$_2$,55>	<a$_1$,25><a$_2$,35> <a$_3$,40>	<e$_1$,15><e$_2$,10> <e$_3$,35><e$_4$,30>
3	<g$_1$,65> <g$_2$,35>	<a$_1$,35><a$_2$,25> <a$_3$,40>	<e$_1$,30><e$_2$,40> <e$_3$,10><e$_4$,20>
4	<g$_1$,40> <g$_2$,60>	<a$_1$,20><a$_2$,40> <a$_3$,40>	<e$_1$,10><e$_2$,10> <e$_3$,40><e$_4$,40>
5	<g$_1$,35> <g$_2$,65>	<a$_1$,30><a$_2$,20> <a$_3$,50>	<e$_1$,25><e$_2$,5> <e$_3$,40><e$_4$,30>
6	<g$_1$,60> <g$_2$,40>	<a$_1$,25><a$_2$,25> <a$_3$,50>	<e$_1$,20><e$_2$,15> <e$_3$,35><e$_4$,30>
7	<g$_1$,20> <g$_2$,80>	<a$_1$,10><a$_2$,40> <a$_3$,50>	<e$_1$,30><e$_2$,5> <e$_3$,35><e$_4$,30>
8	<g$_1$,70> <g$_2$,30>	<a$_1$,30><a$_2$,40> <a$_3$,30>	<e$_1$,10><e$_2$,40> <e$_3$,40><e$_4$,10>
9	<g$_1$,40> <g$_2$,60>	<a$_1$,20><a$_2$,10> <a$_3$,70>	<e$_1$,20><e$_2$,20> <e$_3$,30><e$_4$,30>
10	<g$_1$,35> <g$_2$,65>	<a$_1$,20><a$_2$,30> <a$_3$,50>	<e$_1$,20><e$_2$,10> <e$_3$,40><e$_4$,30>

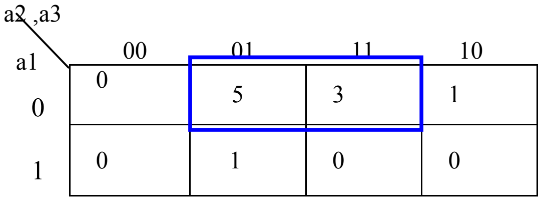
In this table, g1 mean male, g2 means female ; a1means yang man, a2 means adult, a3 means old man ; e1means primary education, e2 means secondary education, e3 means university education, e4 means institute of education

Table 3.Database after Boolean bit transformation

Area ID	Gender	Age	Education
1	01	001	0011
2	01	011	0011
3	10	101	1100
4	01	011	0011
5	01	001	1011
6	10	001	0011
7	01	011	1011
8	10	010	0110
9	01	001	0011
10	01	001	0011

2.2.2 Karnaugh Map Concept

Karnaugh Map presents the simplification of Boolean Algebra in the way of intuitive graph. But to avoid double counting, the researcher simplified it. Only the nearest neighbors that have the largest added value will be combined and simplified. Figure 2. is the Karnaugh Map of Attribute "Age".



a2 ,a3 / a1	00	01	11	10
0	0	5	3	1
1	0	1	0	0

Figure 2. The Karnaugh Map

From Figure 2, it shows that 001 and 011 can be combined. That is,

001,011 → 0_1
"_" means "don't care"

With the same step, the "education" attribute can be simplified:

0011,1011 → _011

2.2.3 Data Replacement

Table 4. presents the feature tha attribute values have been replaced with the simplified values inducted from Karnough Map.

Table 4.Database after Karnaugh Map simplification

Area ID	Gender	Age	Education
1	01	0_1	_011
2	01	0_1	_011
3	10	101	1100
4	01	0_1	_011
5	01	0_1	_011
6	10	0_1	_011
7	01	0_1	_011
8	10	010	0110
9	01	0_1	_011
10	01	0_1	_011

2.2.4 Scan and Recount

Scan the database again, and count the rows with the same attribute values. There are 4 rules in table 5.

Table 5. Database after scan and recount

	Gender	age	Education	vote
1	01	0_1	0_11	7
2	10	101	1100	1
3	10	0_1	_011	1
4	10	010	0110	1

2.2.5 The Descriptive Rules

In table 5, the number 1 rule has the highest vote value . It can be described as :

{<g$_1$,L><g$_2$,H>}^{<a$_1$,L><a$_3$,H>}^{<e$_1$,L><e$_3$,H><e$_4$,H>}→70%

The interpretation of the rule：

70% of high-frequent-crime areas have more females, elderly people, university students and graduate students.

III. RESEARCH METHOD

3.1 Research Process

In this research, all the data came from a library in a college. After data selection, matching, pruning and replacement, the data was inducted by MAOI to generate descriptive rules. Figuer 3. is the research process.

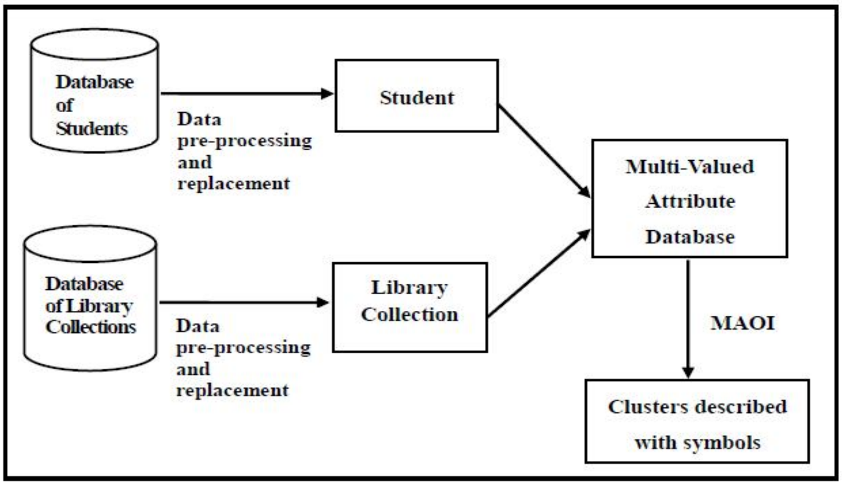

Database of Students
Data pre-processing and replacement
Student
Multi-Valued Attribute Database
Database of Library Collections
Data pre-processing and replacement
Library Collection
MAOI
Clusters described with symbols

Figure 3. Research process

3.2 The Multi-Valued Table

The data selected contains all the college students' library records in 2009. Except some data pre-processing, the twelve departments were divided into three academies, and the library collections were divided into four groups by their book numbers according to New Classification Science for Chinese Libraries(CCL). Table 6. presents a database ready for further analysis. Attribute A stands for academy; a1 is the first academy, including departments of Accounting Information, Business of Administration, International Trade and Business, Marketing and Distribution Management, and Finance; a2 is the second academy, including the departments of Tourism Management, Hospitality Management, Travel Management, and Applied English; a3 is the third academy, including the departments of Information Management, Information Technology and Information Communication. Attribute B stands for student's year in the college; b1 is Freshman; b2 is Sophomore; b3 is Junior; b4 is sinor. Attribute C stands for gender; c1 is male, and c2is female. Attribute D stands for student's grade; d1is 90~100; d2 is 80~89; d3 is 70~79; d4 is under 69.Attribute E stands for the classification of the library collections; e1 is 000~299; e2 is300~499; e3 is 500~799; e4 is 800~999.

Table 6. Database of student's records in the library

Month ID	A	B	C	D	E
1	<a1,101><a2,111> <a3,83>	<b1,2><b2,91> <b3,68><b4,134>	<c1,134> <c2,161>	<d1,9><d2,96> <d3,124><d4,66>	<e1,36 ><e2,98 > <e3,75><e4, 194>
2	<a1,117 ><a2,115> <a3,81>	<b1,1><b2,96> <b3,67><b4,149>	<c1,167> <c2,146>	<d1,28><d2,143> <d3,74><d4,68>	<e1,44><e2,119> <e3,77><e4,201>
3	<a1,207><a2,191> <a3,143>	<b1,3><b2,160> <b3,147><b4,231>	<c1,295> <c2,246>	<d1,45><d2,198> <d3,206><d4,92>	<e1,98><e2,186> <e3,144><e4,380>
4	<a1,203><a2,177> <a3,48>	<b1,3><b2,127> <b3,155><b4,240>	<c1,252> <c2,176>	<d1,38><d2,195> <d3,89><d4,106>	<e1,87><e2,226> <e3,147><e4,294>
5	<a1,156><a2,154> <a3,102>	<b1,1><b2,110> <b3,123><b4,178>	<c1,194> <c2,218>	<d1,56><d2,126> <d3,143><d4,87>	<e1,62><e2,161> <e3,115><e4,248>
6	<a1,136><a2,110> <a3,93>	<b1,3><b2,106> <b3,118><b4,112>	<c1,178> <c2,161>	<d1,31><d2,94> <d3,135><d4,79>	<e1,67><e2,134> <e3,79><e4,226>
7	<a1,11><a2,4> <a3,19>	<b1,0><b2,3> <b3,6><b4,25>	<c1,21> <c2,13>	<d1,0><d2,27> <d3,4><d4,3>	<e1,5><e2,24> <e3,7><e4,6>
8	<a1,6><a2,9> <a3,18>	<b1,0><b2,6> <b3,6><b4,21>	<c1,22> <c2,11>	<d1,2><d2,19> <d3,8><d4,4>	<e1,3><e2,19> <e3,5><e4,8>
9	<a1,147><a2,151> <a3,117>	<b1,92><b2,114> <b3,114><b4,95>	<c1,176> <c2,239>	<d1,48><d2,149> <d3,132><d4,86>	<e1,70><e2,140> <e3,82><e4,282>

10	<a1,170><a2,199> <a3,131>	<b1,78><b2,175> <b3,149><b4,98>	<c1,269> <c2,231>	<d1,70><d2,123> <d3,165><d4,142>	<e1,69><e2,213> <e3,121><e4,302>
11	<a1,177><a2,206> <a3,143>	<b1,76><b2,214> <b3,139><b4,97>	<c1,253> <c2,273>	<d1,63><d2,202> <d3,194><d4,67>	<e1,67><e2,183> <e3,147><e4,331>
12	<a1,195><a2,192> <a3,152>	<b1,121><b2,184> <b3,139><b4,95>	<c1,285> <c2,254>	<d1,76><d2,211> <d3,176><d4,76>	<e1,63><e2,213> <e3,131><e4,324>

3.2.1 Boolean Bit Transformation

In the grid 1A, the data is <a1,101> <a2,111> <a3,83>. The total of these values is 295, and the average number is 295/3=98. When taking 98 as the cutting point, because (101>98.3), (111>98.3), and (83<98.3), the Boolean bit of 1A becomes 110.

Repeat the steps mentioned above, we can transform all the attribute values into Boolean bit, as shown in table 7.

But column E is a special column. Because every student can borrow more than one kind of books, we define the cutting point to be quarter of the number of the students in the month. That means the cutting point of column E equals to the cutting point of column B or D.

Table 7. Database after Boolean bit transformation

M. ID	A	B	C	D	E
1	110	0101	01	0110	0111
2	110	0101	10	0100	0101
3	110	0111	10	0110	0111
4	110	0111	10	0100	0111
5	110	0111	01	0110	0111
6	100	0111	10	0110	0101
7	001	0001	10	0100	0100
8	001	0001	10	0100	0100
9	110	0110	01	0110	0101
10	110	0110	10	0011	0101
11	110	0110	01	0110	0111
12	110	0110	10	0110	0101

3.2.2 Karnaugh Map Concept

The Karnaugh maps of Attribute A, B, D, E are presented in figure4.

Attribute A

a1 \ a2,a3	00	01	11	10
0	0	2	0	0
1	1	0	0	9

100,110 → 1_0

Attribute B

b1,b2 \ b3,b4	00	01	11	10
00	0	2	0	0
01	0	2	4	4
11	0	0	0	0
10	0	0	0	0

0111,0110 →011_

Attribute D

d1,d2 \ d3,d4	00	01	11	10
00	0	0	1	0
01	4	0	0	7
11	0	0	0	0
10	0	0	0	0

0100,0110 →01_0

Attribute E

e1,e2 \ e3,e4	00	01	11	10
00	0	0	0	0
01	2	5	5	0
11	0	0	0	0
10	0	0	0	0

0101,0111 →01_1

Figure 4. The Karnaugh Map of Attribute A,B,D,E

3.2.3 Data replacement

Replace the attribute values with the rules inducted in figure 4 with Karnaugh Map , we complete table8.

Table 8. Database after Karnaugh Map simplification

M. ID	A	B	C	D	E
1	1_0	0101	01	01_0	01_1
2	1_0	0101	10	01_0	01_1
3	1_0	011_	10	01_0	01_1
4	1_0	011_	10	01_0	01_1
5	1_0	011_	01	01_0	01_1
6	1_0	011_	10	01_0	01_1
7	001	0001	10	01_0	0100
8	001	0001	10	01_0	0100

9	1_0	011_	01	01_0	01_1
10	1_0	011_	10	0011	01_1
11	1_0	011_	01	01_0	01_1
12	1_0	011_	10	01_0	01_1

3. 2.4 Scan and Recount

Table 9. Database after scan and recount

	A	B	C	D	E	vote
1	1_0	011_	10	01_0	01_1	4
2	1_0	011_	01	01_0	01_1	3
3	001	0001	10	01_0	0100	2
4	1_0	0101	01	01_0	01_1	1
5	1_0	0101	10	01_0	01_1	1
6	1_0	011_	10	0011	01_1	1

3.2.5 The descriptive rules

From table 9, the sum of the votes for rule number1, 2, and 3 is 9. Thus, rule number 1, 2, and 3 have included 75% of the data. And they are the 3 highest inducted rules in this research. They can be described as the following:

(1){<a1,H><a3,L>}^{<b1,L><b2,H><b3,H>}^{<c1,H><c2,L>}^{<d1,L><d2,H><d4,>L}^{<e1,L><e2,H><e4,H>}→33.3%

It means that there are about 33.3% readers who are males and in the second or third year of the first academy. Their grades are about 80~89. Their reading preference is on book number 300~499 and 800~999.

(2) {<a1,H><a3,L>}^{<b1,L><b2,H><b3,H>}^
{<c1,L><c2,H>}^{<d1,L><d2,H><d4,>L}^{
<e1,L><e2,H><e4,H>}→25.0%

It means that there are about 25% readers who are females and in the second or third year of the first academy. Their grades are about 80~89. Their reading preference is on book number 300~499 and 800~999.

(3) {<a1,L><a2,L><a3,H>}^{<b1,L><b2,L><b3,L>
<b4,H>}^{<c1,H><c2,L>}^{<d1,L><d2,H><d
4,>L}^{<e1,L><e2,H><e3,L><e4,L>}→16.6%

It means that there are about 16.6% readers who are females and in the fourth year of the third academy. Their grades are about 80~89. Their reading preference is on book number 300~499.

IV. Conclusion

To improve library's service and marketing success, the readers needs should be satisfied. There are lots of methods proposed to analyze the relationships between readers and library collections. Most of them only can handle the single-valued attributes. But in our daily life, many information appear as multi-valued attributes. MAOI can induct multi-valued attributes directly, and present the results briefly and descriptively.

In this research, 3 rules were uncovered to explain the characteristics of the readers and their reading preferences. They have accounted for about 75% information. Therefore, it's a successful induction. The library can actively provide readers with appropriate recommendation, and consider the purchase strategy of the collections.

sets of applications and a wider range of multi-valued tables, as the purposes to verify this algorithm and to discover the generalized knowledge from Relational Databases.

REFERENCE

[1]Jun-Rong Huang, “Using clusters to find the most adaptive recommendations of books” *Journal of Educational Media & Library Science,* 43:3， pp309-325, 2006

[2]Ou, J., Lin, S. and Li, J., “The Personalized Index Service System in Digital Library,” *Proc. of the Third International Symposium on Cooperative Database Systems for Advanced Applications*, pp92-99, 2001

[3]J. B. Schafer, J. A. Konstan, and J. Riedl, “ E-Commerce Recommendation Applications,” *Data Mining and Knowledge Discovery,* 5(1), pp115-153, 2001

[4]A. Ansari, S. Essengaier, and R. Kohli, “ Internet Recommendation Systems,” *Journal of Marketing Research*, 37(3), 2000

[5]Shu-Meng Huang, “ A study on the Modified Attributed-Oriented-Induction Algorithm of Mining the Multi-Value Attribute Data”, ICERM, pp62, 2010

[6]W.P. Lee, C.H. Liu, and C.C. Lu, “Intelligent agent-based systems for personalized recommendations in Internet commerce,” *Expert Systems with Applications*, vol. 22, no.4, pp. 275-284, 2002

[7]M. S. Chen, J. Han and P. S. Yu, “Data Mining : An Overview From a database Perspective”, *IEEE, Transactions on Knowledge and Data Engineering*, Vol. 8, No.6, pp866-883,1996

[8]Fayyad, U.M., “ Data Mining and Knowledge Discovery :Making Sense Out of Data,” *IEEE Expert*, Vol.11, Issue 5, pp20-25, 1996

[9]M.J.A. Berry and G. Linoff，*Data Mining Techniques：For Marketing, Sales, and Customer Support,* John Wiley & Sons. 1997

[10]Y. Cai, N. Cercone, and J. Han,” attribute-oriented induction in relational database”, *Knowledge Discovery in Databases* ,Ch 12, AAAI/MIT Press. 1991

[11]Jiawei Han and Micheline Kamber, *Data Mining : Concepts and Techniques* (Second Edition), Morgan Kaufmann Pub, 2006

[12]J. Han, Y. Cai, and N. Cercone, “ Knowledge Discovery in Databases : An Attribute-Oriented Approach,” In Proceedings of the 18th VLDB Conference, Vancouver, British Columbia, Canada. Pp547-559, 1992

[13]Yen-Liang Chen, Ching-Cheng Shen, “ Mining generalized knowledge from ordered data through attribute-oriented induction tecniques.” *European Journal of Operational Research*, 166, pp221-245, 2005

[14]J. Han, Y. Cai and N. Cercone, “Data-Driven Discovery of Quantitative Rules in Relational Database,” *IEEE Transaction on Knowledge and Data engineering*, Vol.5, No.1, February 1993

Mining a Web Security Portal – A Case Study

Suresh Kalathur
Computer Science Department, Boston University Metropolitan College, Boston, MA, USA

Abstract - *Vast amount of information is available on the web which could help an instructor when teaching an information technology course. Various online publishers provide this knowledge in the form of RSS feeds which could be subscribed by individual users through various RSS aggregators. In this paper, we present on object oriented framework which combines those concepts and provides a higher level perspective of web portals by aggregating and maintaining such information. The information gathered from these sources is archived and data mining techniques are used to analyze the aggregated items and examine the similarities between them.*

Keywords: Information retrieval, Data and Knowledge mining, Data processing, Text Extraction, Object oriented analysis and design.

1 Introduction

A framework to set up and store information about web portals is of interest to various users. A typical user is an instructor of a course. The instructor wishes to put together a portal from various sources. The portal serves as a tool for the students to pursue and understand the day to day happenings in those fields of relevance. However, in the absence of a framework to manage this knowledge, such a portal would be of little use to others after the instructor is done with the course. The main purpose of the proposed framework is to maintain information about these portals and let the users create their own portals from existing ones or create new ones as necessary. In this paper, the techniques are shown where the information gathered by these portals can be stored persistently in relational databases for retrieval and analysis for document mining and classification.

2 Object Data Model

In this section, we look into the data model for the web portals from an object-oriented perspective and describe the various classes and the associations among them. Each class represents an entity and thus a table in the relational database. The associations between the classes are captured with the foreign key constraints in the database. Since Java is used in implementing the classes, the object-relational persistence is achieved by making use of the Hibernate [2] framework to maintain the consistency between the Java classes in our implementation and the tables in the relational database. The following are the core classes capturing the data model following the Unified Modeling approach [3].

2.1 Portal

The *portal* represents the main class in the data model. Each portal has a unique name that distinguishes the portal from the others (e.g., Security por*tal,* Java portal, Data mining portal, etc.). A typical portal (also referred to as a concrete portal) will contain a list of sources which provide the items appropriate for that particular portal. On the other hand, a portal may also be modeled as a composite of the existing portals. For example, a programming languages portal can be established from the existing portals – Java portal, C# portal, etc. The following diagram shows the class structure and the composite relationship between the portal and its child portals. The class structure also makes it feasible to have a hierarchical relationship between the composite portals and the concrete portals.

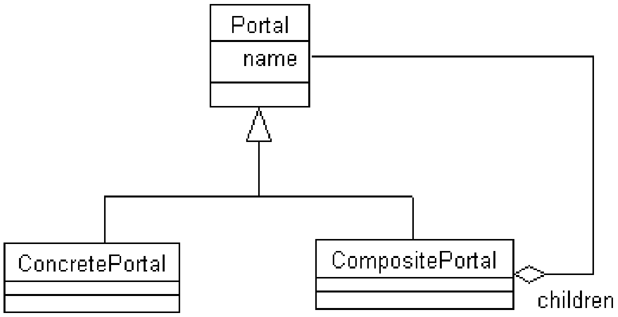


Figure 1. Portal Class Structure

2.2 Sources

Associated with each *concrete portal* is a list of sources that feed the portal. Since the sources could be of arbitrary types, the class structure shown in Figure 2 employs the polymorphism model to accommodate multiple source types.

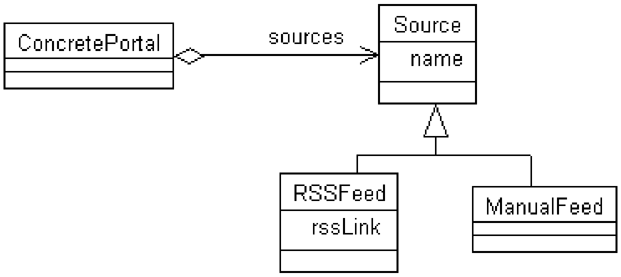


Figure 2. Source Class Structure

The most common type of portal sources are the RSS feeds [6] made available by various news, journal, and web publishers. An RSS feed is an XML document [1] which contains elements for the title, an hyperlink, a brief description (optional), publication date and so on. These individual elements of an RSS feed are captured by the *Item* class structure as shown in Figure 3. In addition to the RSS feeds, manual feeds may also be setup as part of the portal. The user would then post items to the respective manual feed as appropriate.

2.3 Items

The sources of each portal contain a list of items which are displayed with their title and the associated hyperlink. Clicking the hyperlink would take the web browser to the actual content described by that particular item. An optional description along with the publication date is also part of the *item* class structure as shown in the following figure.

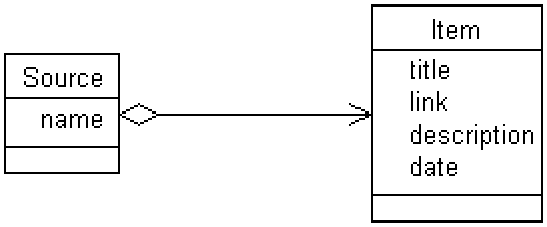


Figure 3. Item Class Structure

2.4 Class Diagram

The complete class diagram modeling the portals is shown in Figure 4. The class structure is simple and elegant in accordance with the fundamental principles associated with object oriented analysis and design resulting in low coupling between the classes and high cohesion within.

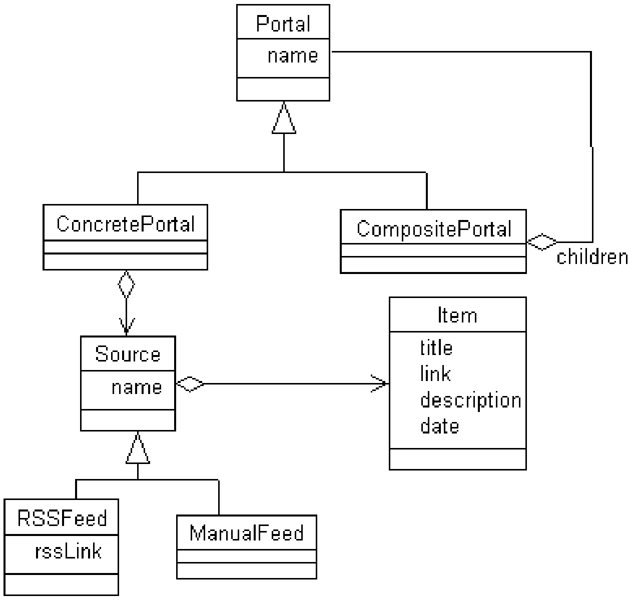
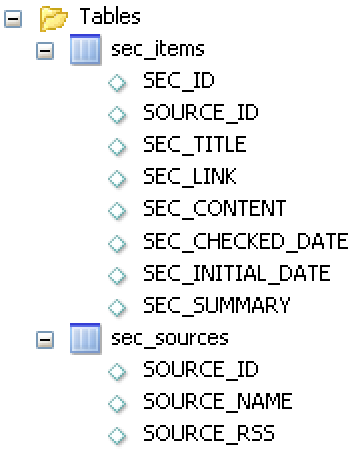


Figure 4. Complete Class Diagram

A snapshot of the data mining portal setup using RSS feeds from various sources is shown in Figure 5. Similarly, Figure 6 shows a security portal set up with RSS feeds.

3 Archival

The web portal is crawled at regular intervals to see if any new items are published by the respective portals. Since the source RSS feeds change frequently and only publish the latest items, a repository is maintained to store the information of the published items. The following relational schema is used for the archival of the items from a particular portal.



- Tables
  - sec_items
    - SEC_ID
    - SOURCE_ID
    - SEC_TITLE
    - SEC_LINK
    - SEC_CONTENT
    - SEC_CHECKED_DATE
    - SEC_INITIAL_DATE
    - SEC_SUMMARY
  - sec_sources
    - SOURCE_ID
    - SOURCE_NAME
    - SOURCE_RSS

Figure 5. Relational Schema

The current items published on a portal are checked against the database to see if there are any new ones. The new items are then inserted in the *sec_items* table. The relevant information about the item is stored which includes the *title*, the *URL*, and the *HTML* content, etc. The text extraction process in the next section describes the extraction and storing of the text summary.

4 Text Extraction

The HTML contents of each item are stored in the repository during the data acquisition phase. The text that is associated with the item needs to be extracted for performing text mining on the acquired data. Since different sources have their own HTML templates, the text extraction process is customized using the factory pattern as shown below.

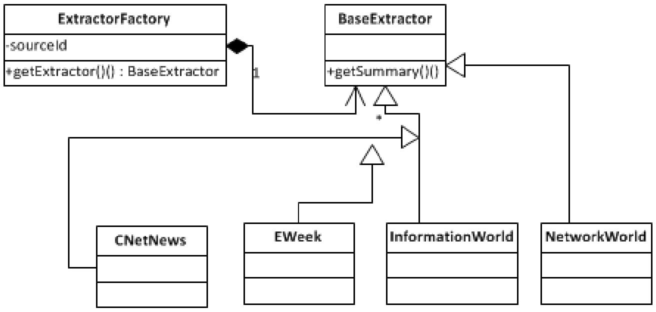


Figure 6. Text Extraction from HTML

Depending upon the portal source, the *getExtractor()* factory method returns the respective HTML extractor which is customized for text extraction of the corresponding portal. An HTML parser is used to parse the content and filter the text summary of the item. The irrelevant parts of the web content are ignored during this process. Since the HTML templates may be changed at any time by the publisher, appropriate strategies need to be developed to detect and tune the extractor if that happens.

The following example shows the analysis of 18,582 items related to the Web Security portal extracted and stored since November 2006. The text mining is done using the SQL Server 2008 R2 Business Intelligence Studio using the Integration Services and Analysis Services components. The first analysis deals with the titles of the items and the second analysis concerns with the text content summary of the items. The portal contents are from 9 sources providing the RSS feeds.

The data flow tasks involved in building the dictionary of the terms occurring in the titles are shown in the figure below.

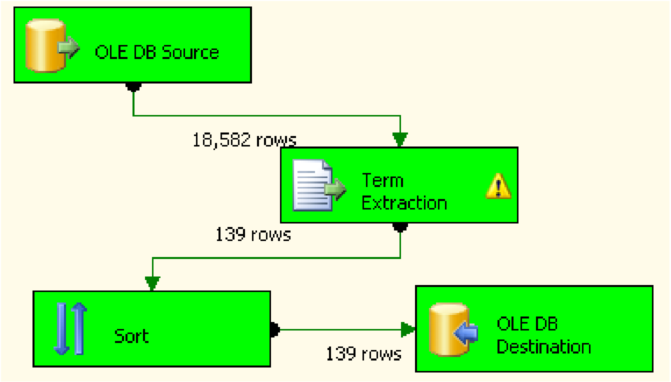
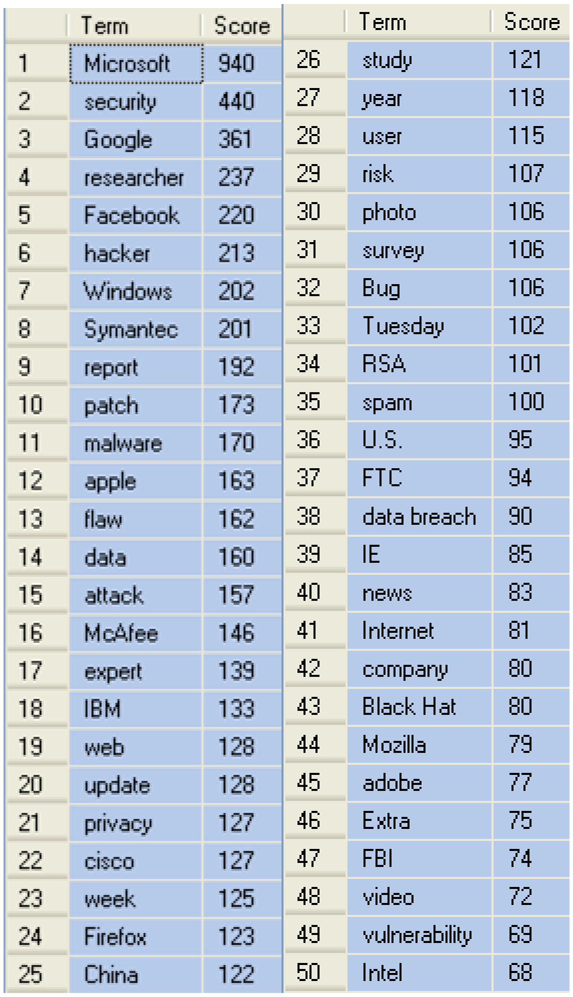


Figure 7. Data Flow for Item Titles Dictionary

The *noun* and *noun phrase* technique with a frequency threshold of 30 is used for this process. The dictionary extraction resulted in 139 terms. The first 50 terms along with their scores are shown in the table below.



	Term	Score		Term	Score
1	Microsoft	940	26	study	121
2	security	440	27	year	118
3	Google	361	28	user	115
4	researcher	237	29	risk	107
5	Facebook	220	30	photo	106
6	hacker	213	31	survey	106
7	Windows	202	32	Bug	106
8	Symantec	201	33	Tuesday	102
9	report	192	34	RSA	101
10	patch	173	35	spam	100
11	malware	170	36	U.S.	95
12	apple	163	37	FTC	94
13	flaw	162	38	data breach	90
14	data	160	39	IE	85
15	attack	157	40	news	83
16	McAfee	146	41	Internet	81
17	expert	139	42	company	80
18	IBM	133	43	Black Hat	80
19	web	128	44	Mozilla	79
20	update	128	45	adobe	77
21	privacy	127	46	Extra	75
22	cisco	127	47	FBI	74
23	week	125	48	video	72
24	Firefox	123	49	vulnerability	69
25	China	122	50	Intel	68

Figure 8. Item Titles Dictionary

Similarly, the term extraction is performed using the summary contents of the items as shown in the figure below.

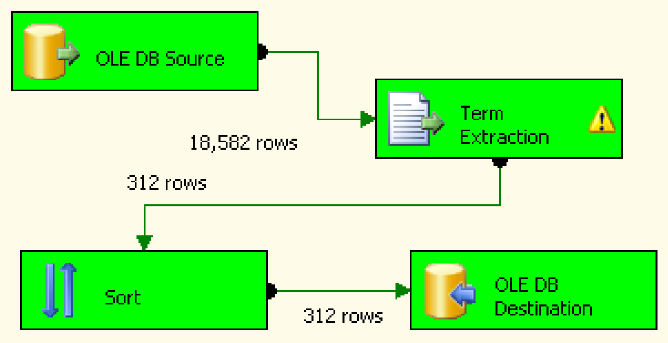
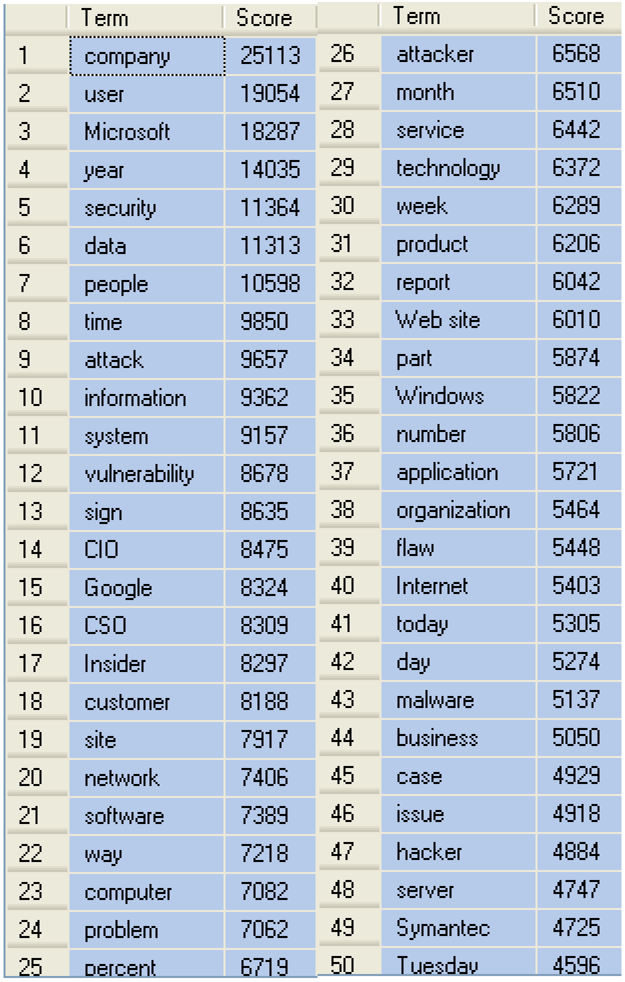


Figure 9. Data Flow for Item Summaries Dictionary

Since the summary length of each item is much larger than the length of the item titles, a frequency threshold of 1000 is used for this process. The extraction resulted in 312 terms in the dictionary. The first 50 terms along with their scores are shown in the table below.



	Term	Score		Term	Score
1	company	25113	26	attacker	6568
2	user	19054	27	month	6510
3	Microsoft	18287	28	service	6442
4	year	14035	29	technology	6372
5	security	11364	30	week	6289
6	data	11313	31	product	6206
7	people	10598	32	report	6042
8	time	9850	33	Web site	6010
9	attack	9657	34	part	5874
10	information	9362	35	Windows	5822
11	system	9157	36	number	5806
12	vulnerability	8678	37	application	5721
13	sign	8635	38	organization	5464
14	CIO	8475	39	flaw	5448
15	Google	8324	40	Internet	5403
16	CSO	8309	41	today	5305
17	Insider	8297	42	day	5274
18	customer	8188	43	malware	5137
19	site	7917	44	business	5050
20	network	7406	45	case	4929
21	software	7389	46	issue	4918
22	way	7218	47	hacker	4884
23	computer	7082	48	server	4747
24	problem	7062	49	Symantec	4725
25	percent	6719	50	Tuesday	4596

Figure 10. Item Summaries Dictionary

Once the dictionary terms are extracted based on their frequencies, the items that need to analyzed have to be represented as a vector of the terms present in the dictionary. This representation is used to compare the similarities between the items. The first analysis on the item titles involves the data flow tasks as shown in the following figure. Since the length of the titles is relatively small, the average vector length of the titles is 30,584÷18,582 = 1.65.

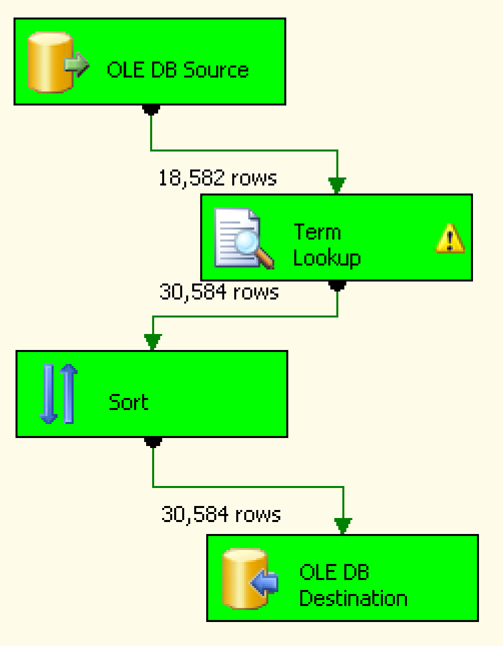
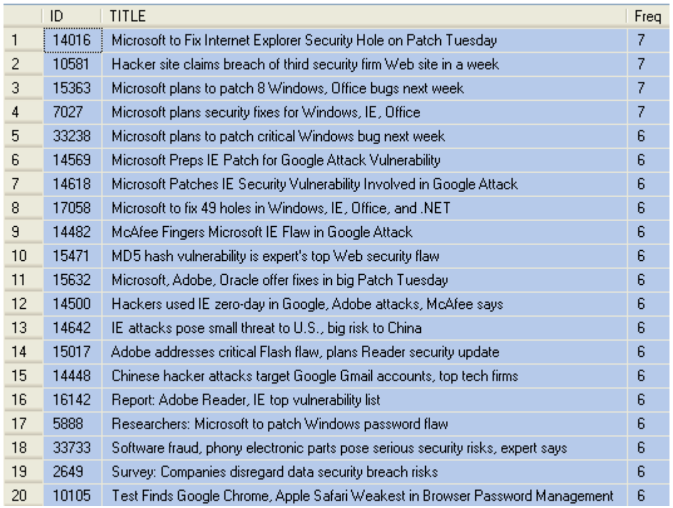


Figure 11. Data Flow for Item Title Vectors

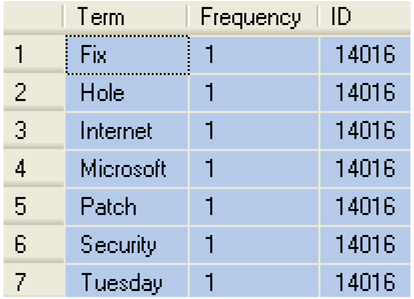
Further analysis of the dictionary terms occurring in the titles of the items is illustrated in the figure below. The first 20 titles sorted by the number of terms are shown. The last column shows the maximum number of dictionary terms occurring in each title. The analysis reveals that the maximum number of dictionary terms in the titles is 7.



	ID	TITLE	Freq
1	14016	Microsoft to Fix Internet Explorer Security Hole on Patch Tuesday	7
2	10581	Hacker site claims breach of third security firm Web site in a week	7
3	15363	Microsoft plans to patch 8 Windows, Office bugs next week	7
4	7027	Microsoft plans security fixes for Windows, IE, Office	7
5	33238	Microsoft plans to patch critical Windows bug next week	6
6	14569	Microsoft Preps IE Patch for Google Attack Vulnerability	6
7	14618	Microsoft Patches IE Security Vulnerability Involved in Google Attack	6
8	17058	Microsoft to fix 49 holes in Windows, IE, Office, and .NET	6
9	14482	McAfee Fingers Microsoft IE Flaw in Google Attack	6
10	15471	MD5 hash vulnerability is expert's top Web security flaw	6
11	15632	Microsoft, Adobe, Oracle offer fixes in big Patch Tuesday	6
12	14500	Hackers used IE zero-day in Google, Adobe attacks, McAfee says	6
13	14642	IE attacks pose small threat to U.S., big risk to China	6
14	15017	Adobe addresses critical Flash flaw, plans Reader security update	6
15	14448	Chinese hacker attacks target Google Gmail accounts, top tech firms	6
16	16142	Report: Adobe Reader, IE top vulnerability list	6
17	5888	Researchers: Microsoft to patch Windows password flaw	6
18	33733	Software fraud, phony electronic parts pose serious security risks, expert says	6
19	2649	Survey: Companies disregard data security breach risks	6
20	10105	Test Finds Google Chrome, Apple Safari Weakest in Browser Password Management	6

Figure 12. Frequent Terms in Titles

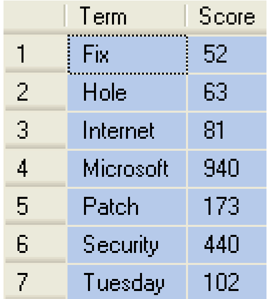
The following table shows the terms present in the first title from the previous figure. The 7 terms from the dictionary are each present once in this item.



	Term	Frequency	ID
1	Fix	1	14016
2	Hole	1	14016
3	Internet	1	14016
4	Microsoft	1	14016
5	Patch	1	14016
6	Security	1	14016
7	Tuesday	1	14016

Figure 13. Terms in a Title

The score of the above terms with respect to all the titles in the collection are shown below.



	Term	Score
1	Fix	52
2	Hole	63
3	Internet	81
4	Microsoft	940
5	Patch	173
6	Security	440
7	Tuesday	102

Figure 14. Corresponding Scores

The frequency distribution (from highest to lowest) of the terms occurring in all the item titles is shown in the figure below.

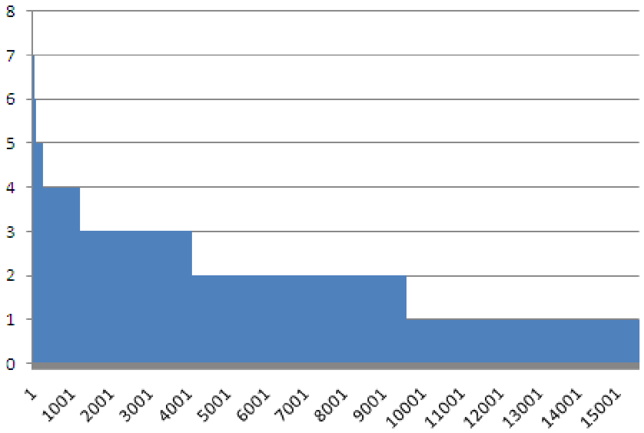


Figure 15. Frequency Distribution of Title Terms

Similarly, the terms occurring in the summaries of the items are analyzed. The data flow tasks involved in building the document vectors of the item summaries are shown in the figure below. The average vector length of the item summaries is 819,993÷18,582 = 44 (approx).

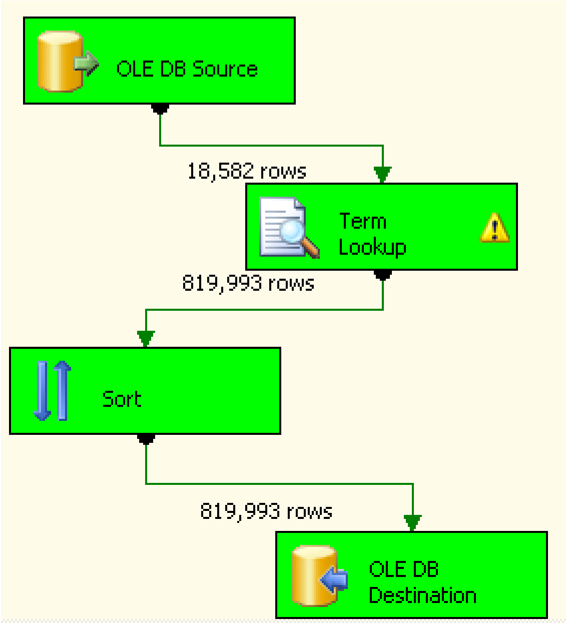


Figure 16. Data Flow of Item Summary Vectors

The maximum number of dictionary terms occurring in an item summary in this case is 161. A distribution of the frequencies of the dictionary terms ordered from the highest to the lowest is shown in the following figure.

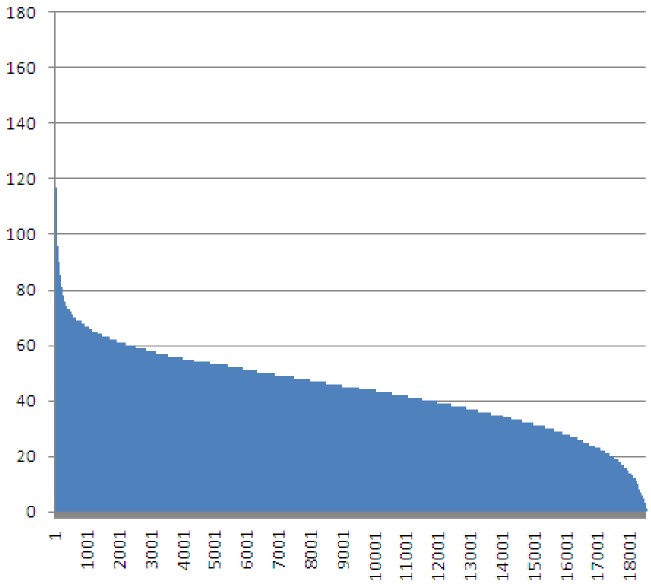


Figure 17. Frequency Distribution of Summary Terms

Note that the dictionaries used for analyzing the item titles is different from the one used for item summary analysis. The dictionary for the item titles has 139 terms while the dictionary for the item summaries has 312 terms. The intersection of these two dictionaries contains 106 terms common between the two.

5 Text Mining

The Analysis Services component of the SQL Server Business Intelligence Studio provides the capabilities for performing data mining tasks. In this scenario, the articles are compared with each other using the dictionaries that were created. The analysis is done using the item summaries. A decision tree approach is used to predict the source of the items as shown in the following figure.

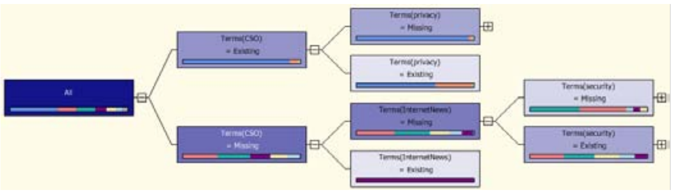

All
Terms(CSO) = Existing
Terms(CSO) = Missing
Terms(privacy) = Missing
Terms(privacy) = Existing
Terms(InternetNews) = Missing
Terms(InternetNews) = Existing
Terms(security) = Missing
Terms(security) = Existing

Figure 18. Decision Tree

Clustering technique is commonly used to group similar documents together. The results of the clustering technique are shown below.

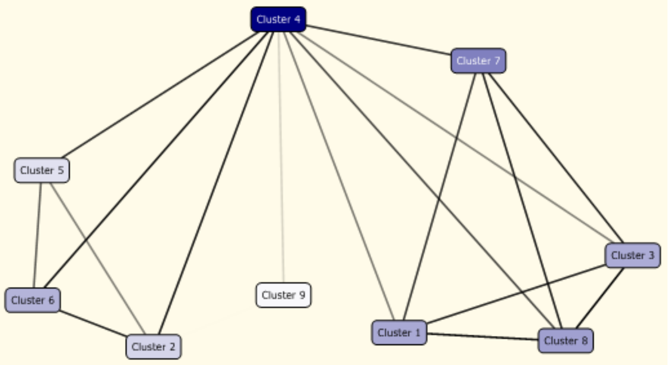


Figure 19. Clustering

The tool identified 9 clusters in the data. The original articles were collected from 9 different sources as shown in the following table.

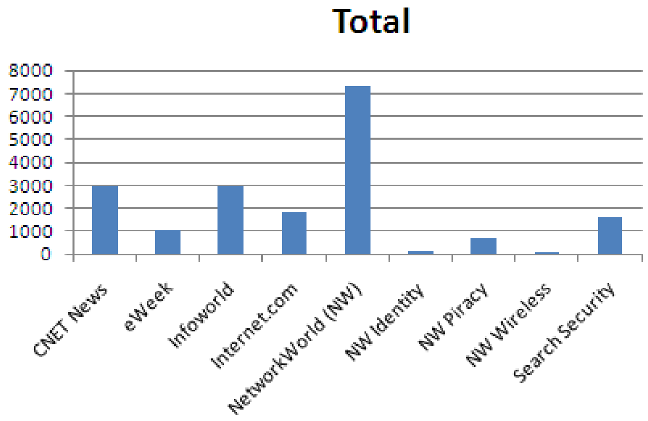


Figure 20. Item Counts

The yearly frequencies of the collected items are shown in the figure below. The dates range from November 2006 to early March 2011.

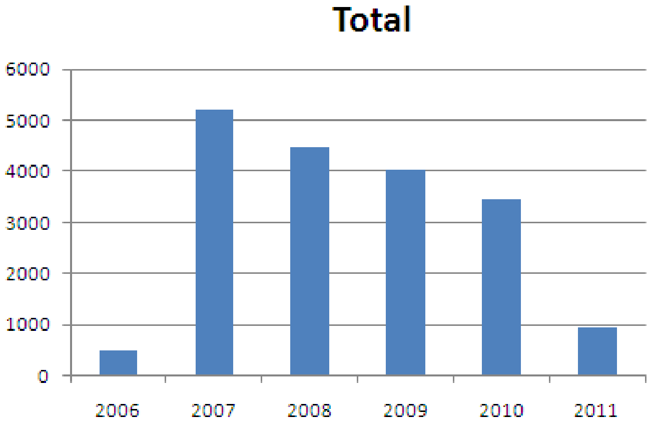
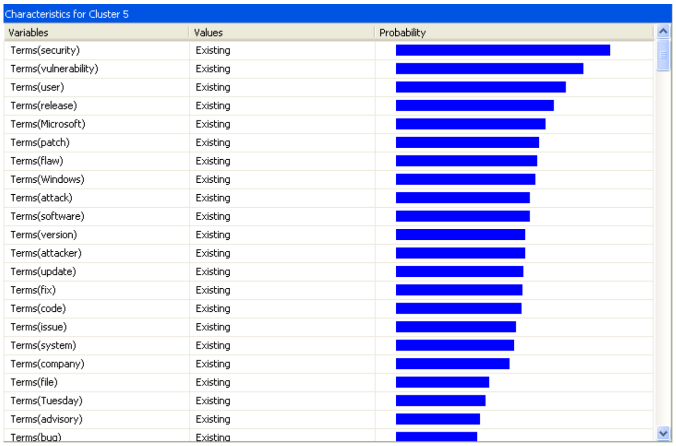


Figure 21. Yearly Frequencies

The relative probabilities of the terms existing in a particular cluster are as shown in the figure below.


Characteristics for Cluster 5

Variables	Values	Probability
Terms(security)	Existing	
Terms(vulnerability)	Existing	
Terms(user)	Existing	
Terms(release)	Existing	
Terms(Microsoft)	Existing	
Terms(patch)	Existing	
Terms(flaw)	Existing	
Terms(Windows)	Existing	
Terms(attack)	Existing	
Terms(software)	Existing	
Terms(version)	Existing	
Terms(attacker)	Existing	
Terms(update)	Existing	
Terms(fix)	Existing	
Terms(code)	Existing	
Terms(issue)	Existing	
Terms(system)	Existing	
Terms(company)	Existing	
Terms(file)	Existing	
Terms(Tuesday)	Existing	
Terms(advisory)	Existing	
Terms(bug)	Existing	

Figure 22. Relative Term Probabilities

Conclusions

Various RSS aggregators exist which let users put together the portals. In our case, we have a unique way of combining such scenarios with an object oriented framework and put together composite portals. An instructor for a course can setup such a portal. The framework can provide usage statistics for the items in the portal and provide feedback on the most viewed ones and those of high relevance to the course. Semantic web techniques [1] may be used to find the relationships between the items. Document mining techniques [6][8] are incorporated into the framework to analyze the content of each item in the portal. Such an analysis may provide information about similar items, provide capabilities for querying the items, and retrieve the similar items.

6 References

[1] C. Patel, et. al., OntoKhoj: A Semantic Web Portal for Ontology Searching, Ranking, and Classification, Proceedings of the 5th ACM international workshop on Web information and data management, 2003.

[2] J. Elliott, Hibernate: A Developer's Notebook," O'Reilly, 2004.

[3] M. Fowler, "UML Distilled: A Brief Guide to the Standard Object Modeling Language," Addison-Wesley, 2004.

[4] E. Frank, "Predicting Library of Congress Classifications from Library of Congress Subject Headings," Journal of the American Society for Information Science and Technology, 55(3), pp. 214-227.

[5] S. Kalathur, "An Object Oriented Dynamic Web Portal Framework – A Case Study," The 2008 International Conference on Information and Knowledge Engineering, IKE'08, Las Vegas, July 14-17, 2008.

[6] Z. Shi, H. Ma, Q. He, "Web Mining: Extracting Knowledge from the World Wide Web", Data Mining for Business Applications, 2009, pp. 197-208.

[7] F. Wolf, T. Poggio, P. Sinha, Human Document Classification using Bag of words, MIT-CSAIL-TR-2006-054 CBCL-263 August 9, 2006.

[8] Q. Zhang, R. Segall, "Web Mining: A Survey of Current Research, Techniques, and Software", International Journal of Information Technology & Decision Making, 2008, pp. 683-720.

Applying the Data Mining Technique for Improving Internal Controls in Financial and Credit Institutes

Golsoom Akbarpour[1], Dr. Mohhammad Ebrahim Mohammad Pourzarandi[2]
[1] Accounting Department, Eslamic Azad University-Mashhad Branch, Mashhad, Iran
[2] Management Depatrnent, eslamic Azad University- Science and Research Branch, Tehran, Iran

Abstract - *Fraud detection is a practical issue for many industries. Although much efforts for detection fraud, hundreds of millions of dollars annually from the investment fraud affected institutions are destroyed. Mainly because cheating in a small number are seen and fraud detection requires a high skill. Academic research has shown how data mining techniques can be valuable in the fight against fraud. In the present study, we seek a solution to improve internal controls systems in banks. So what should techniques for better and faster detection to find out the internal fraud? In this study, the specific fraud that happens in banks, directly effects to revenue and resource in bank, has been studied. Effort has been using the techniques of data mining, clustering, performance of inspectors of banks are better with using the results of this research. In this research, considered cheating by using clustering algorithms modeled K – mean.*

Key Words: Data Mining, Internal Controls, Fraud Detection, Clustering

1 Introduction

Quick development in volume and value of electronic trade show that audit old techniques are less practical and efficient, these techniques lack the ability to ensure the accuracy and integrity of transactions firm. (Onions, 2003) In today's fast business world, information systems without delay have provided accounting system without delay and without delay relationship between trade firms. Current methods of audit spend a lot of time to ensure supply. (Flowerday, 2006)

Complexity and technology of today modern business is the characteristics of firms that will require auditors to create the methods and procedures for auditing. Auditors, both internal and independent auditors, are forced to use new methods and ways. They are fully aware that the current auditing methods will not be able properly to audit and inspect companies, especially banks. Due to the necessity, the purpose of this study is to use new computer technology for detection of fraud. Data mining techniques, new techniques are that these days have been the subject of a lot of researches. In the present study, the effort is that these techniques are used to improve internal controls for banks. In Iran, the necessity of using the data mining techniques and articles are realized in this area but has not been tried on as a research study and as scientific - applied article for this case. Much research in this area have been done including the following: Jans et (2006) believe in your research, at the beginning of the study should be specified what kind of fraud you are considering. They have used data mining techniques to discover misappropriation of assets in a company called Epsilon.

Spathis et.al (2007) in their article has used effective data mining classification techniques to discover the companies which publish fraud financial statements. This research deals with the dependent elements to financial statement. In this article the decision trees and neural networks in discovering the fraudulent statements are considered. The input data are made from the ratios that are taken from the financial statements.

Advanced data mining techniques and the structure of neural networks are combined successfully to gain a record of the hidden forgery with the low wrong warning rate. Jans et.al (2010) in their other article argued the application of data mining to reduce the internal fraud.

2 Significance the objectives of this study

The objective of this study is to find out the fraud and financial crimes by using the new technology of data analysis. The interviews that are carried out with the accountants and the auditors, nowadays in Iran they use the traditional methods and approaches to audit, and do not apply any software and they have to go through a lot of evidence and documents applying the traditional approaches, and this is a demanding and oppressive job, in that many of these workers will lose their energy after some years to continue the job. Therefore, we have to apply the new technology in this field. Managers of financial organization are trying hard to get better methods and approaches to control their unit internally. The manager know well that if there was no internal effective control system, the main objective of the company which is to maintain the profitability of the company and decreasing the number of unpredictable events would be very difficult.

Internal controls will increase the efficiency reduce the risk of losing the assets and gain logical certainty about capability of financial statements and observing the rules. In this research the researcher is going to discuss and analyze the application of data mining, the role of it in improving the internal, and to clarify the necessity of using these techniques.

3 Definition of internal controls

Internal control system includes; policies, methods, financial controls and non-financial controls and organizational arrangements which are defined in an organization / office or a unit. These are assigned to prevent any fraud, wasting or abuse in a company. These rules are applied to create order, increase benefits and interest and efficiency of the practice, increase the capability of reliability and dependability of the document and evidence and financial reports, observing the rules and gaining assurance about keeping and maintaining the assets. (Arbab soleimani, Nafari, 1992)

3.1 Fraud

Any deliberate or deceitful act one or more directors, employees or third parties for having undue or illegal advantage is fraud.

3.1.1 Different types of fraud in banks

In reviews and interviews with special inspectors and auditors in banks following is described. Two types of fraud in banks as a whole can exist:

First division: A type of fraud which does not affect the bank income and the sources. (fraud of violations)

Discover fraud violations:

Inspectors detect this kind of violations in the banks and the documents are compared and the whole operation is done manually. There are not any integration data and some of the information should be taken from the Fed as inquiry.

Second division: Fraud that affects the bank's income and resources. In this type of fraud when the inspectors enter to the bank, they go straight to the heading of banks revenue. There are two types of fraud:

1- That fraud results in the low income.
2- Fraud that causes high bank cost.
The effect of these two is the same, because finally, both of them reduce profits. But in terms of importance, the kind of fraud that will lead to low-income, get priority for investigation. Therefore, this study examines the first case deals. Bank earnings are low in two ways:

- Revenue collected despite of collection is not recorded in the heading of income.
- Revenue collected are recorded in the relevant headings but is gone out with any reason and different types, they is used individually: (case study in this survey)

When the amount as creditor records in heading of income, necessarily debtor circulation is questionable for the inspector. For example, when the installment of the customer is taken to be received into the headlines of income, so received profit account is recorded as a debtor and thus received profits account increase, so when the money is gone out the this account, received profit should be owed and account balance falls.

In financial documents at any branch bank, amounts of debtor funds received from customers including: profits (received- future interest receivable), money pledged and bank fees should have a certain process. The funds received should certainly be taken into account branch manager, and document type must be internal recharges.

Authorized debit trend in earnings heading in the bank is two forms:

- The funds received will be debited and they are transferred to the branch manager account and their document type must be intra bank memo[1]. Means creditor account is the branch manager account.
- Or the funds due to incorrect calculation may be done by officer will be debited that in the process their document type is an internal recharges.

As described, each of the two states has its own specific process. So if this process is not within their normal routine likely there will be cheating here. For example, if the amount of account of funds received been debtor in a date, and opponent of account does not be branch manager account, and computational wrong also does not be, here a question occurs. Where has gone this money gone?

3.1.1.1 How to detect this type of fraud as a traditional

In these cases, inspectors go straight to ledgers in first. They extract the total amount of debtors and the date, that these amounts have been debited, and accounting document number associated with it. Then the individual documents check with adapting to the type of document and their opponent of account. Suspicious cases are extracted and they ask the reason about them from the officials. If not satisfied, they will be announced as fraud cases.

3.1.1.2 How to use data mining approach to fraud detection and discovery

With technology development, as information systems are developed, the fraudulent behavior also will change, and we deal with a dishonest of computer based. This subject is caused that the work of inspector be more difficult. There are a lot of complexity and ambiguity in the large volumes of data and they will be caused feeling of

[1] There are two types of document in banking of Iran: Intra bank memo, Internal recharges

strong needs to techniques that they would can seek into data and finally extract the information; this information is very useful for decision.

4 Data Mining Definition

Data mining is the process in which there is analysis of data from different angle and perspectives and summarizing the same data into the relevant information. This kind of information could be utilized to increase the revenue, cutting the costs or both. Data mining uses artificial intelligence techniques, neural networks, and advanced statistical tools (such as cluster analysis) to reveal trends, patterns, and relationships, which might otherwise have remained undetected. In contrast to an expert system (which draws inferences from the given data on the basis of a given set of rules) data mining attempts to discover hidden rules underlying the data.

4.1 Data Mining Process

Data mining process consists of six stages. The process is shown in Figure 1.

In any project this process of data mining must be done. It is an iterative process that typically involves the following phases:

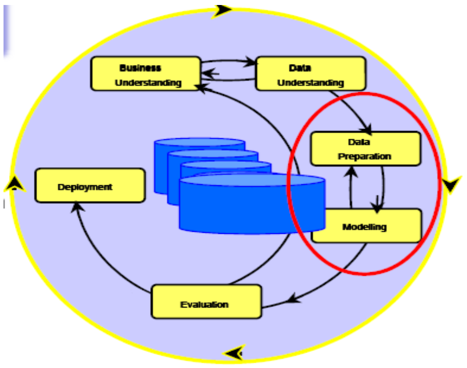


Figure1: data mining process

4.2 How this particular fraud detection using data mining techniques

Implementing Data Mining Process:

STEP ONE: Business Understanding

Understand the project objectives and requirements from a business perspective, and then convert this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

In this case study must find enough knowledge about the fraud that it must be discovered with using data mining techniques. With research conducted on data mining techniques and software to the conclusion we reached that in this study we would use clustering technique and Rapidminer software.

STEP TWO: Data Understanding

Start by collecting data, then get familiar with the data, to identify data quality problems to discover first insights into the data, or to detect interesting subsets to form hypotheses about hidden information.

According to targets were set in the previous stage, information about income heading of customer extracted from database of Mellat bank of Iran.

STEP THREE: Data Preparation

Includes all activities required to construct the final data set (data that will be fed into the modeling tool) from the initial raw data. Tasks include table, case, and attribute selection as well as transformation and cleaning of data for modeling tools.in this study, we need to the attributes of financial data of income heading

In the accounting documents, account credit entry in another row and we need that simulated with the rest of the attributes are brought in columns of table in excel to be able to enter as input to software and the software able to compare all records. Therefore this problem resolved with the use of SQL programming that is explained completely in the next section (extracting the required data and processing them).

STEP FOUR: Modeling

In modeling, the first step is the selection technique that is used. In defined software clustering algorithm (using k-means) is designed. Data mining techniques are different that each of these techniques has a lot of algorithms including:

1- Describing techniques
2- Predictive techniques

4.2.1 Clustering Technique

Clustering Definition:

Cluster analysis is a statistical method that is used to find the real groups of data. Classified based on similarities or similarities do not be. The goal of clustering is that the data divided into several groups and in this classification, the data of different groups should be had maximum difference and data contained in a group should be very similar.

K-mean Algorithm:

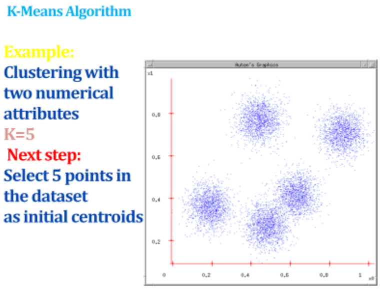
The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori that the k value can be determined by the user. The algorithm works on a set of multidimensional examples (vectors) can be very large.

Category X that includes N samples is considered. The purpose of clustering is to split samples to k- clusters {

$C_1, C_2, \ldots, C_k\}$. Each Cj, nj sample had and each sample exactly be classified in one cluster. In other word, each cluster with the following conditions:

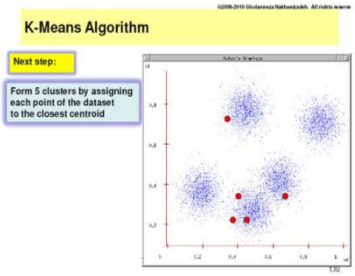
1) $C_1 \cup C_2 \cup \ldots \cup C_k = X$ (1)
2) $\forall\, i, C_i \neq \emptyset$
3) $\forall\, i, j, i \neq j\; C_i \cap C_j = \emptyset$

The main idea is to define k centroids, one for each cluster. These centroids shloud be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more. The figures 3-7 show the steps of k-mean algorithm:2


K-Means Algorithm

Example:
Clustering with two numerical attributes
K=5
Next step:
Select 5 points in the dataset as initial centroids

Figure 2 --step1 of k-means

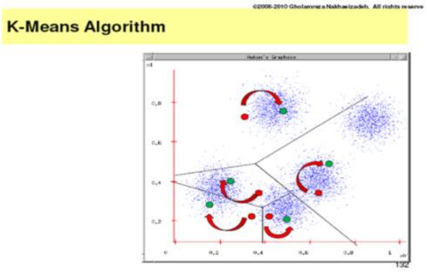

©2006-2010 Gholamreza Nakhaeizadeh. All rights reserve

K-Means Algorithm

Next step:

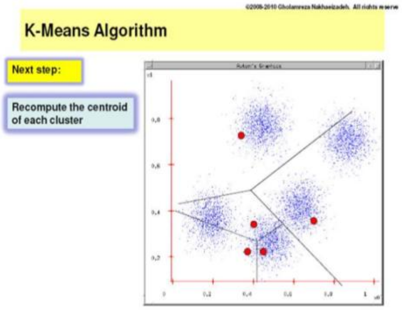
Form 5 clusters by assigning each point of the dataset to the closest centroid

Figure 3- step2 of k-means


©2006-2010 Gholamreza Nakhaeizadeh. All rights reserve

K-Means Algorithm

Figure 4- step3 of k-means

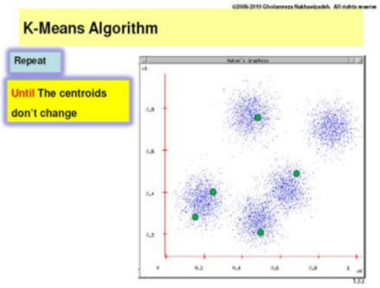

©2006-2010 Gholamreza Nakhaeizadeh. All rights reserve

K-Means Algorithm

Next step:

Recompute the centroid of each cluster

Figure 5 – step4 of k-means

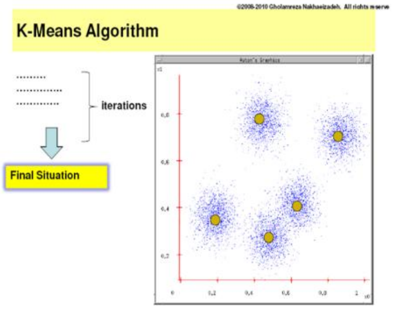

©2006-2010 Gholamreza Nakhaeizadeh. All rights reserve

K-Means Algorithm

Repeat

Until The centroids don't change

Figure 6 – step5 of k-means


©2006-2010 Gholamreza Nakhaeizadeh. All rights reserve

K-Means Algorithm

Final Situation

Figure 7- step6 of k-means

STEP FIVE: Evaluation

Information extracted according to the target user is analyzed and the best results are given. The purpose of this step is not just representation of results (logically or chart), But refining the information presented to the user is goals for this stage. When you get the results, an assessment is done from the whole process of data mining has been done so far. Did we get our results and objectives that were

[2] Figure based on: http//www.autonlab.org/tutorials/kmeans.html

defined in the first stage? Can the system run with 90% confidence?

STEP SIX: DEPLOYMENT

Model making is not the end of the project. Even if more information is the target from modeling the data, the data need to be organized and presented, so that customers can use them.

5 Methodology

In this study, first a theoretical background about the fraud, in particular fraud over banks, internal control, data mining have been explained and the handling of data mining techniques to uncover certain fraud in banks will consider.

To design system, the following steps have been taken:

1- First, state of fraud types of bank was reviewed and analyzed during meetings with the inspector expert of Bank Inspection Department.
2- Knowledge Discovery models were studied and appropriate model is chosen
3- After the conceptual model design, began model of detection fraud and designing its software. The system based on queries a database was designed.

6 Statistical Society

All databases of Mellat Bank of Iran have been available on the units for in the implementation model and test its performance information, information related to heading of revenue from date 2006 to 2010.

7 How to extract the required data and processing them

In databases of the bank, required accounting information was observed by using view 8401. The first, considering the need of project, all the records of revenue heading (with code 5111) were extracted. Note that in each row, the document number and document date together create a unique record, therefore was observed a creditor accounts headlines by using this feature. After preparing the required query by using the tools DTS(Data transfer service) the output of the system was prepared to excel format. This subject is explained more in below paragraph:

There is an option that called view in SQL software. To seeing a series of specific information in the database, writing a query can access the relevant information. In order to avoid writing multiple query and also prevent the user or programmer direct access to data in the database we use the view. So that every query of the database that we need, in the view with a specific name (for example, view 8401) entered, therefore we create the view. Every time it wanted to gain access, rather than re-write the query and direct access to the database then is used the new view that we made. To access to accounting documents 8401 we have used "view 8401".

Technical Description:

1. First, using the following query all the records that was extracted from the headlines of the 5111 Code. These records may be in debt and creditor:

```
SELECT              (convert(nvarchar(30),sndNo)+''+
convert(nvarchar(30),sndDate) )as *into a5111
FROM      VIEW8401
WHERE   KolCode=5111
```

2- Using the following query all records in 5111 headline have been circulating with the opposite side are extracted:

```
select * into ttt  from SELECTKolCode , SndNo , SndDate,
RDbAmnt, RcrAmnt , SndType   SndDesc,
convert(nvarchar(30),sndNo)+''+
convert(nvarchar(30),sndDate) )as  a
FROM      VIEW8401
where               (convert(nvarchar(30),sndNo)+''+
convert(nvarchar(30),sndDate) ) in(
SELECT              (convert(nvarchar(30),sndNo)+''+
convert(nvarchar(30),sndDate)     )as     c     FROM
VIEW8401
WHERE   KolCode=5111
(as a
where a.KolCode<>5111
```

3- Now, using the following query to each flow 5111 is extracted, and at the end of each row of the opponent also be displayed. Technical point is that in stage 1 and 2, in each row there were the 5111 code or its opponent side, in other words, we could not be had in a row at the same time the two sides that in Step 3 this problem was resolved:

```
select a5111.*,ttt.[codeSarfasl] as[ tarafeHesab]' from
a5111
inner join ttt on a5111.c = ttt.a
```

4- Using the tool of DTS in MS Sql Server 2005 the required outputs was provided that 26,000 records are available. As it was explained in the statistical society, research period be 2006 to 2010. At this stage we have data with excel format as the following:

Table 1: Attributes needed in the project

Document Number	Date of Document	Code of Income Heading	Debit Amount	Credit Amount	Type of Document	Code of Opposite Account of Income Heading

In database these data are in another record and with using SQL and appropriate query were located in the same record with anther attributes.

With obtain data above, extracting the information needed starts. As previously described, the possibility of fraud in this heading, there are in amounts of debtor. For reducing the data, attributes must be filter out with capabilities of Excel, and just column of Type of Document and Code of Account of Income Heading must input the software. Then the clustering algorithms can be modeled in Rapidminer software.

Note that parameter of number of cluster in algorithm must be set before running. This value with the run several times and giving different values and See the results for each, finally appropriate number of clusters can be gained.

After cluster analysis, the results obtained in the table is given below:

Table 2: Results of clustering algorithm

Number of Document	Date of Document	Type of Document	Debit Amount	Code of Opposite Account of Income Heading
38	2008/7/21	3	798809	1312
495	2008/08/04	3	597211	1332
548	2007/10/06	3	460274	1335
563	2007/09/16	3	25000000	1393
78	2009/05/19	3	1941340	1511
78	2009/05/19	3	1941340	1511
78	2009/05/19	3	1941340	1511
466	2008/11/01	3	3430135	1521
542	2009/12/08	3	1197671	1521
150	2010/02/09	3	788082	1521
537	2009/03/14	3	191143483	4853
239	2009/08/15	3	364523	4855
285	2010/01/26	3	460591	5311

8 Comparison of model results with the results of manually

In results of manually, in addition to above results, records with Code of Opposite Account of Income Heading 4131 and 1111 are also extracted that they may be suspicious cases in terms of inspectors that in results of programmer were observed.

9 Causes of conflict

In analyzing the clusters, items that were repeated once or at most three, as suspicious areas were introduced. In inconsistent cases in results of inspector each record were repeated about 10 times. As you know, cheating number is basically very low. That is why additional records in the discovery manually were not announced as the record of suspicious by data miner.

10 Conclusion

After implementing the algorithm and create clusters, someone who is familiar with data must analysis of the clusters. So if the inspectors and auditors are trained in data mining techniques, they can do operations of modeling of own business. In the near future, the day will come that all kinds of different fraud are predicted by prediction models of data mining.

11 Reference

[1] Arbab Soleimani, Kamali Zare (2006). Internal auditing, (second edition), Audit Organization

[2] Etemadi. H (2005). Auditor responsibility in connection with fraud and error, Publication 24, Audit publication

[3] Johnson Richard.A, W.Vychrn Din, (2010). Applied multivariate statistical analysis, (Niroman, Translator)

[4] Brause,R.&Langsdorf ,T.&Hepp ,M.(2006). Neural Data mining for Credit card fraud Detection,J.W.Goethe-University,Frankfurt a.m.,Gesellschaftf.Zahlugssysteme GZS,Frankfurt a.m.,Germany.

[5] Kirkos, Efstathios& Spathis, Charalambos, Manolopoulos,Yannis (2007). Data mining techniques for the detection of fraudulent financial statement, Science Direct, (pp 995-1003)

[6] Fraweley,W& Piatetsky-Shapiro,G(1991). Knowledge Discovery Databases,AAAI/MIT Press

[7] Jans, Mieke& Lybaert, Madine& Vanhoof, Koen (2010). Internal fraud risk reduction ,Result of a case study, International Journal of Accounting Information System,(pp 17-41)

[8] Nakhaeizadeh,Gholamreza(2008). Tutorial Data Mining: Advance in Predictive Modeling and Unsupervised Learning .

[9] Onions,R.L. (2003). Towards a Paradigm for Continuous Auditing,university of Salford, United Kingdom.

Design and Analysis of a Dynamic Load Balancing Strategy for Large-Scale Distributed Association Rule Mining

Raja Tlili[1], **Yahya Slimani**[2]

[1, 2] Department of Computer Science, Faculty of Sciences of Tunis, Campus Universitaire, Tunis, Tunisia

Abstract - *Association rule mining is one of the most important data mining techniques. Algorithms of this technique search a large space, considering numerous different alternatives and scanning the data repeatedly. Parallelism seems to be the natural solution in order to be able to work with industrial-sized databases. Large-scale computing systems, such as Grid computing environments, are recently regarded as promising platforms for data and computation-intensive applications like data mining. However, to improve the performance and achieve scalability by using these heterogeneous platforms, new data partitioning approaches and workload balancing features are needed. The focus of this paper is to propose a dynamic load balancing strategy for parallel association rule mining algorithms in the context of a Grid computing environment. This strategy is built upon a distributed model which necessitates small overheads in the communication costs for load updates and for both data and work transfers. It also supports the heterogeneity of the system and it is fault tolerant.*

Keywords: Association rules, Apriori algorithm, Dynamic load balancing, Grid computing, Parallel association mining.

1 Introduction

With the advances in data acquisition and storage technologies, the problem of how to turn large volumes of raw data into useful information becomes a significant one. In order to decrease the gap between data and useful information, a group of architectures and utilities, some of them are new and others exist since a long time, are grouped under the term data mining. Association rule mining which trends to find interesting correlation relationships between items in a large database of sales transactions has become one of the most important data mining techniques [3, 11]. Although algorithms of this technique have a simple statement, they are computationally and input/output intensive. High performance parallel and distributed computing can relieve current association rule mining algorithms from the sequential bottleneck, providing scalability to massive data sets and improving response time.

Grid computing is recently regarded as one of the most promising platform for data and computation-intensive applications like data mining. A Grid can be envisioned as a collection of geographically dispersed computing and storage resources interconnected with high speed networks and effectively utilized in order to achieve performances not ordinarily attainable on a single computational resource [2]. In such computing environments, heterogeneity is inevitable due to their distributed nature.

Almost all current parallel association rule mining algorithms assume the homogeneity and use static load balancing strategies. Thus applying them to Grid systems will degrade their performance. The load imbalance that occurs during execution time is caused by the dynamic nature of these algorithms and also by the heterogeneity of such distributed systems. Because of that we have to develop new methodologies to handle this problem, which is the focus of our research.

In this paper, we develop and evaluate a run time load balancing strategy for mining association rule algorithms under a grid computing environment. The rest of the paper is organized as follows: Section 2 introduces association rule mining technique. Section 3 describes the load balancing problem. Section 4 presents the system model of a Grid and the proposed dynamic load balancing strategy. Experimental results obtained from implementing this strategy are shown in section 5. Finally, the paper concludes with section 6.

2 Mining association rules

Association rules mining (ARM) finds interesting correlation relationships among a large set of data items. A typical example of this technique is market basket analysis. This process analyses customer buying habits by finding associations between different items that customers place in their "shopping baskets". Such information may be used to plan marketing or advertising strategies, as well as catalog design [3]. Each basket represents a different transaction in the transactional database, associated to this transaction the items bought by a customer. Given a transactional database D, an association rule has the form $A=>B$, where A and B are two itemsets, and $A \cap B = \varnothing$. The rule's support is the joint probability of a transaction containing both *A* and *B* at the same time, and is given as *σ(AUB)*. The confidence of the rule is the conditional probability that a transaction contains *B* given that it contains *A* and is given as *σ(AUB)/σ(A)*. A rule is frequent if its support is greater than or equal to a pre-determined minimum support and strong if the confidence is more than or equal to a user specified minimum confidence.

Association rule mining is a two-steps process:

1) The first step consists of finding all frequent itemsets that occur at least as frequently as the fixed minimum support;

2) The second step consists of generating strong implication rules from these frequent itemsets.

The overall performance of mining association rules is determined by the first step which is known as the frequent set counting problem [3].

2.1 Sequential association rule mining

Many sequential algorithms for solving the frequent set counting problem have been proposed in the literature. We can define two main methods for determining frequent itemsets supports: with candidate itemsets generation [11, 13] and without candidate itemsets generation [5].

Apriori [11] was the first proposed effective algorithm. This algorithm uses a generate-and-test approach which depends on generating candidate itemsets and testing if they are frequent. It uses an iterative approach known as a level-wise search, where *k*-itemsets are used to explore *(k+1)*-itemsets. During the initial pass over the database the support of all *1*-itemsets is counted. Frequent *1*-itemsets are used to generate all possible candidate *2*-itemsets. Then the database is scanned again to obtain the number of occurrences of these candidates, and the frequent *2*-itemsets are selected for the next iteration. DCI algorithm proposed by Orlando and others [12] is also based on candidate itemsets generation. It adopts a hybrid approach to compute itemsets supports, by exploiting a counting-based method (with a horizontal database layout) during its first iterations and an intersection-based technique (with a vertical database layout) when the pruned dataset can fit into the main memory. FP-growth algorithm [5] allows frequent itemsets discovery without candidate itemsets generation. First it builds from the transactional database a compact data structure called the FP-tree then extracts frequent itemsets directly from the FP-tree.

2.2 Parallel association rule mining

Association rule mining algorithms suffer from a high computational complexity which derives from the size of its search space and the high demands of data access. Parallelism is expected to relieve these algorithms from the sequential bottleneck, providing the ability to scale the massive datasets, and improving the response time. However, parallelizing these algorithms is not trivial and is facing many challenges including the workload balancing problem. Many parallel algorithms for solving the frequent set counting problem have been proposed. Most of them use Apriori algorithm [11] as fundamental algorithm, because of its success on the sequential setting. The reader could refer to the survey of Zaki on association rules mining algorithms and relative parallelization schemas [7]. Agrawal et al. proposed a broad taxonomy of parallelization strategies that can be adopted for Apriori in [10]. It also exist many grid data mining projects, like Discovery Net, GridMiner, DMGA [9] which provide mechanisms for integration and deployment of classical algorithms on grid. Also the DisDaMin project that deals with data mining issues (as association rules, clustering, etc.) using distributed computing [14].

3 Load balancing: problem description

Work load balancing is the assignment of work to processors in a way that maximizes application performance [4]. A typical distributed system will have a number of processors working independently with each other. Each processor possesses an initial load, which represents an amount of work to be performed, and each may have different processing characteristics (i.e. different architecture, operating system, CPU speed, memory size and available disk space). To minimize the time needed to perform all tasks, the workload has to be evenly distributed over all processors in a way that minimizes both processor idle time and inter-processor communication. Work-load balancing process can be generalized into four basic steps: *(1)* Monitoring processor load and state; *(2)* Exchanging workload and state information between processors; *(3)* Decision making; and *(4)* Data migration. The decision phase is triggered when the load imbalance is detected to calculate optimal data redistribution. In the fourth and last phase, data migrates from overloaded processors to underloaded ones. According to different policies used in the previously mentioned phases, Casavant and kuhl [13] classify work-load balancing schemes into three major classes: Static versus dynamic load balancing, centralized versus distributed load balancing, and application-level versus system-level load balancing.

3.1 Load balancing in parallel association rule mining algorithms

Static load balancing can be used in applications with constant workloads, as a pre-processor to the computation [4]. Other applications require dynamic load balancers that adjust the decomposition as long as the computation proceeds [4, 6]. This is due to their nature which is characterized by workloads that are unpredictable and change during execution. Data mining is one of these applications. Parallel association rule mining algorithms have a dynamic nature because of their dependency on the degree of correlation between itemsets in the transactional database which cannot be predictable before execution. Basically, current algorithms assume the homogeneity and stability of the whole system, and new methodologies are needed to handle the previously mentioned issues.

3.2 Load Balancing in Grid Computing

Although intensive works have been done in load balancing, the different nature of a Grid computing environment from the traditional distributed system, prevent existing static load balancing schemes from benefiting large-scale applications. An excellent survey from Y. Li et al. [16], displays the existing solutions and the new efforts in dynamic load balancing that aim to address the new challenges in Grid. The work done so far to cope with one or more challenges brought by Grid: heterogeneity, resource sharing, high latency and dynamic system state, can be identified by three categories as mentioned in [16]: ***(1)*** Repartition methods focus on calculating data distribution in a heterogeneous way, but don't pay much attention to the data movement in Grid; ***(2)*** Divisible load theory based schemes well model both the computation and communication, but loose validity in case of adaptive application; ***(3)*** Prediction based schemes need further investigation in case of long-term applications.

4 Proposed load balancing approach

4.1 Proposed System Model

In our study we model a Grid as a collection of *T* sites with different computational facilities and storage subsystem. Let $G = (S_1, S_{2,...,} S_T)$ denotes a set of sites, where each site S_i is defined as a vector with three parameters $S_i = (M_i\,,\, Coord(S_i)\,,\, L_i)$, where M_i is the total number of clusters in S_i, $Coord(S_i)$ is the workload manager, named the coordinator of S_i, which is responsible of detecting the workload imbalance and the transfer of the appropriate amount of work from an overloaded cluster to another lightly loaded cluster within the same site (intra-site) or if it is necessary to another remote site (inter-sites). This transfer takes into account the transmission speed between clusters which is denoted $\zeta_{ijj'}$ (if the transmission is from cluster cl_{ij} to cluster $cl_{ij'}$). And L_i is the computational load of S_i. Each cluster is characterized by a vector of four parameters $cl_{ij}=(N_{ij},Coord(cl_{ij})\ ,\ L_{ij},\ \omega_{ij})$, where N_{ij} is the total number of nodes in cl_{ij} , $Coord(cl_{ij})$ is the coordinator node of cl_{ij} which ensures a dynamic smart distribution of candidates to its own nodes, L_{ij} is the computational load of cluster cl_{ij} and ω_{ij} is its processing speed which is the mean of processing times of cluster's nodes. In fact each node nd_{ijk} has its processing speed denoted ω_{ijk} . Thus

$$\omega_{ij} = \text{Average}(\omega_{ijk}) = (\sum_k \omega_{ijk}) / N_{ij} \qquad (1)$$

Figure 1 shows the Grid system model. To avoid keeping global state information in a large-scale system (where this information would be very huge), the proposed load balancing model is distributed in both intra-site and inter-sites.

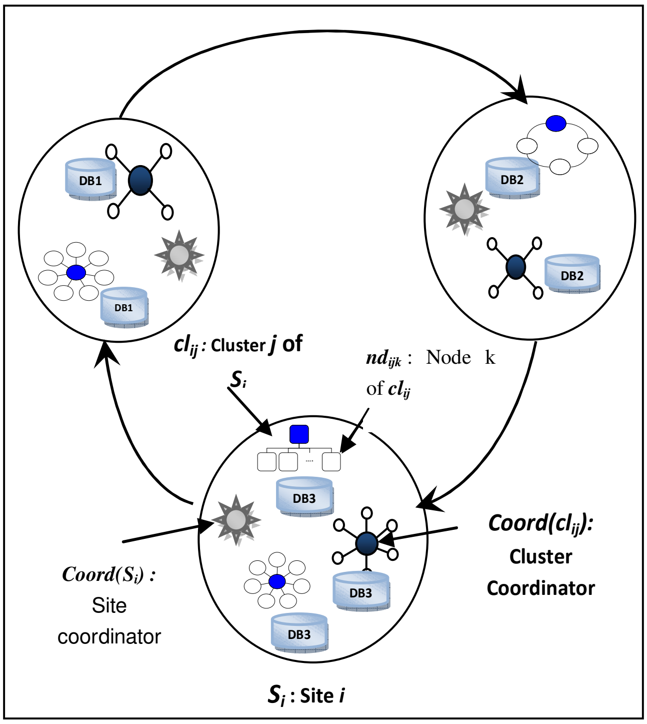

DB1
DB1
DB2
DB2
DB3
DB3
DB3
cl_{ij} : Cluster *j* of S_i
nd_{ijk} : Node k of cl_{ij}
$Coord(cl_{ij})$: Cluster Coordinator
$Coord(S_i)$: Site coordinator
S_i : Site *i*

Fig.1 The system model of a Grid

Each site in the Grid has a workload manager, called the coordinator, which accommodates submitted transactional database partitions and the list of candidates of the previous iteration of the association rules mining algorithm. Each coordinator aims at tracking the global workload status by periodically exchanging a "state vector" with other coordinators in the system. Depending on the workload state of each node, the frequency of candidate itemsets may be calculated in its local node or will be transferred to another lightly loaded node within the same site. If the coordinator cannot fix the workload imbalance locally, it selects part of transactions to be sent to a remote site through the network. The distination of migrated work is chosen according to the following hierarchy : First The coordinator of the cluster $Coord(cl_{ij})$ selects the available node within the same cluster; If the workload imbalance still persists then $Coord(cl_{ij})$ searches for an available node in another cluster but within the same site; Finally, in extreme cases, work will be send to a remote site. The coordinator of the site $Coord(S_i)$ will look for the nearest site available to receive this workload (i.e. least communication cost). Our model is fault-tolerant. In fact, it takes into consideration the probability of failure of a coordinator node. If the coordinator node does not give response within a fixed period of time, an election policy is invoked to choose another coordinator node.

4.2 Proposed Load Balancing Strategy

Dynamic load balancing is necessary for the efficient use of highly distributed systems (like Grids) and when solving problems with unpredictable load estimates (like association rule mining). That's why we chose to develop a dynamic work load balancing strategy. Our proposed load balancing strategy depends on three issues: ***(i)*** Database architecture (partitioned or not); ***(ii)*** Candidates set (duplicated or partitioned); ***(iii)*** network communication parameter (bandwidth).

Our strategy could be adopted by algorithms which depend on candidate itmesets generation to solve the frequent set counting problem. It combines static and dynamic load balancing, and this by interfering before execution (i.e. static) and during execution (i.e. dynamic).

Before execution: To respond to the heterogeneity of the computing system we are using (Grid) the database is not just partitioned into equal partitions in a random manner. Rather than that, the transactional database is partitioned according to the characteristics of the different sites, where the size of each portion is determined according to the site processing capacity (i.e., different architecture, operating system, CPU speed, etc.). It is the responsibility of the coordinator of the site $Coord(S_i)$ to allocate to its site the appropriate database portion according to the site processing capacity parameters stored in its information system.

During execution: Our load balancing strategy acts on three levels: ***(1)*** level one is the migration of work between nodes of the same cluster. If the skew in workload still persists the coordinator of the cluster $Coord(cl_{ij})$ moves to the next level; ***(2)*** level two depends on the migration of work between clusters within the same site; ***(3)*** and finally if work migration of the previous two levels is not sufficient then the coordinator of the overloaded cluster $Coord(cl_{ij})$ asks from the coordinator of the site $Coord(S_i)$ to move to the third level which searches for the possibility of migrating work between sites. Communication between the coordinators of different sites is done in a unidirectional ring topology via a token passing mechanism.

The following workload balancing process is invoked when needed. It is the responsibility of distributed coordinators to detect that need dynamically according to the load status of their relative nodes:

1) From the intra-site level, coordinators of each cluster update their global workload vector by acquiring workload information from their local nodes. From the Grid level, coordinators of different sites periodically calculate their average workload in order to detect their workload state (overloaded or underloaded). If an imbalance is detected, coordinators proceed to the following steps.

2) The coordinator of an overloaded cluster makes a plan for candidates migration intra-site using equation *(2)*. If the imbalance still persist, it creates another plan for transactions migration inter-sites (i.e. between clusters of the Grid) using equation *(3)*.

$$EET_{i,j} > Coefintra * (CCN_{i,j,k} + EET_{i,k}) \qquad (2)$$

$$EET_{i,j} > Coefinter * (CCS_{i,p} + EET_{p,q}) \qquad (3)$$

Where $EET_{i,j}$ is the estimated required processing time for node $N_{i,j}$ of the site S_i, $EET_{i,k}$ is the estimated required time to process the same operations in another node $N_{i,k}$ of the same site S_i in the case of equation (1), $EET_{p,q}$ is the estimated required processing time in another node $N_{p,q}$ of a remote site S_p in the case of equation (2), $CCN_{i,j,k}$ is the communication cost between nodes $N_{i,j}$ and $N_{i,k}$ of the site S_i, $CCS_{i,p}$ is the communication cost between sites S_i and S_p, *Coefintra* is the coefficient of decision of the intra-site migration, and *Coefinter* is the coefficient of decision of the inter-site migration.

The two previously mentioned equations serve in ensuring that before performing any candidate itemsets migration between nodes within the same site, or transactions migration between different sites, the coordinator must guaranty that transactions or candidates migration will improve the performance of the Grid. The processing time at a local node must dominate (by a prefixed threshold) the processing time at a remote node added to it the time spent in communication and transactions (or candidates) movements. Otherwise, it will be better to process transactions (or candidates) locally. The definition of this coefficient of domination (or threshold) depends of the environment of execution and the size of data to be processed. The coefficient of the intra-site migration is smaller than the coefficient for the inter-sites migration, because the communication cost intra-site is much less than the communication cost inter-sites.

3) The concerned coordinator (the coordinator of the overloaded cluster or the coordinator of the overloaded site) sends migration plan to all processing nodes and instructs them to reallocate the work load.

For each site S_i, the coordinator will execute the algorithm displayed in figure 2. Where *Coord* (S_i) is the coordinator node of the site S_i, M_i is the number of computational clusters, GL_i is the global vector of workloads of all nodes in the site S_i, AL_i is the average workload of the site S_i, $LMax_i$ is the threshold of the maximum workload of the site S_i, L_{ij} is the local workload of the cluster cl_{ij} of the site S_i, $Limit_{i,j}$ is the threshold of the maximum workload of the cluster $cl_{i,j}$ in the site S_i , $CCN_{i,j,k}$ is the communication cost between clusters $cl_{i,j}$ and $cl_{i,k}$ of the site S_i, $EET_{i,j}$ is the estimated required time for cluster $cl_{i,j}$ of the site S_i to complete the processing of remaining transactions data, $CCS_{i,p}$ is the communication cost between sites S_i and S_p, V_i is the state vector of all the other coordinators in the Grid. The state of the coordinator of each site is stored in the vector with these information: *Id-site*, $CCS_{i,p}$ and L_i. This vector is sorted by $CCS_{i,p}$ and L_i.

```
Begin
    For All ($cl_{i,j}$ in $S_i$) Do
        Update ($Coord_i, GL_i(L_{i,j})$)      // Update the global vector of workload of each site
    End For
    $AL_i \leftarrow \frac{\sum_{j=1}^{M_i} L_{i,j}}{M_i}$      // Calculate the average workload of the site $S_i$
    If $AL_i > LMax_i$ Then      // The site is overloaded
        Sort $GL_i(L_{i,j})$      // The sort is done  in a decreasing order to detect the more overloaded node.
        Search in $V_i$
        If Exists ($cl_{p,q}$ in $S_p$) Where ($EET_{i,j} > Coefinter$ *( $CCS_{i,p} + EET_{p,q}$)) Then
            Transactions_Migration inter_Site($S_i$ ,$S_p$)      // Workload balancing inter-sites
        End If
    Else
        If ($L_{i,j} > Limit_{i,j}$) Then   //An overload is detected in a node $N_{i,j}$
            If  Exists ($cl_{i,k}$ in $S_i$ ) Such that (($L_{i,k} < Limit_{i,k}$) And ($EET_{i,j} >$ ($Coefintra$ *( $CCN_{i,j,k} + EET_{i,k}$)))) Then
                Candidates_Migration intra_Site ($cl_{i,j}$ ,$cl_{i,k}$)      // Load balancing intra-site
            Else      // Load balancing inter-sites
                // the migration inter-site is more expensive (in time) than the migration intra-site.
                Search in $V_i$
                If Exists ($cl_{p,q}$ in $S_p$) Where ($EET_{i,j} > Coefinter$ *( $CCS_{i,p} + EET_{p,q}$)) Then
                    Transactions_Migration inter_Site($S_i$ ,$S_p$)
                End If
            End If
        End If
    End IF
End
```

Fig. 2: The run time work load balancing strategy

5 Performance evaluation

5.1 Case Study: the Apriori Algorithm

The specific characteristics of the problem of frequent set counting associated with those of the computing environment (Grid) must be taken into account. While association rule mining method is based on global criteria (support, frequencies), we are only disposed by local (partial) data views due to the fact of distribution. The treatment must be done on the entire database, comparing each partition of the base with all the others must be possible in order to be able to obtain global information. Our goal is to limit the number of communications and synchronizations, and to be benefit as much as possible from the available computing power. This could be done by exploiting all possible ways of parallelism and if necessary by using a pipeline approach between dependent tasks in order to be able to parallelize the various stages of the frequent set counting algorithm.

In order to evaluate the performance of our workload balancing strategy we parallelized the sequential Apriori algorithm which is the fundamental algorithm for frequent set counting algorithms with candidates generation. To reduce the number of accesses to the transactional database we used the depth-first Apriori proposed by W. Kosters et al. [15]. This version of Apriori needs only three passes over the transactional database, while classic Apriori needs k-passes (where k is the length of the maximal itemset).

Data parallelism is not sufficient to improve the performance of association rule mining algorithms. Subsets of extremely large data sets may also be very large. So, in order to extract the maximum of parallelism, we applied a hybrid parallelisation technique (i.e. the combination of data and task parallelism). Where we aimed to study parallelism inside the program code. This could be done through searching inside the algorithm procedures for independent segments and analyzing the loops to detect tasks (or instructions) that could be executed concurrently. A hybrid approach between candidate duplication and candidate partitioning is used. The candidate itemsets are duplicated all over the sites of the Grid, but they are partitioned between the nodes of each site. The reason for partitioning the candidate itemsets is that when the minimum support threshold is low they overflow the memory space and incur a lot of disk I/O. Hence, the candidate itemsets are partitioned into equivalence classes based on their common *(k-2)* length prefixes. A detailed explanation of candidate itemsets clustering could be found in [8]. We can resume the important basic concepts of our parallelization method in what follows:

Site:

- The transactional database is partitioned between sites according to the capacity of treatment of each site.
- Candidate itemsets are duplicated between sites (in order to reduce the communication cost between sites).

Cluster:

- Every database partition is shared between nodes of the same site if they have the same storage subsystem, otherwise it will be duplicated.
- Candidates are partitioned between site's clusters (according to the capacity of treatment of the cluster)

Node :

- Receives a group of candidates from the coordinator of the cluster.
- Calculates their supports.
- Sends local supports to cluster's coordinator which performs the global supports reduction.

Cluster's coordinator:

- Distributes candidate itemsets between nodes according to their capacities. Candidates are distributed by their (k-1) common prefix.
- Performs the global reduction of supports to obtain global frequencies.
- Responsible for workload balancing operation of his cluster

Site's coordinator :

- Search for the maximum loaded cluster (or site) and the minimum loaded cluster (or site).
- Migration of the necessary amount of work (candidates or transactions or both) from the maximum to the minimum loaded clusters or sites.

5.2 Experimental Platform

The performance evaluations presented in this section were conducted on Grid'5000 [1], a dedicated reconfigurable and controllable experimental platform featuring 13 clusters, each with 58 to 342 PCs, interconnected through Renater (the French Educational and Research wide area Network). It gathers roughly 5000 CPU cores featuring four architectures (Itanium, Xeon, G5 and Opteron) distributed into 13 clusters over 9 cities in France (Bordeaux, Grenoble, Lille, Lyon, Nancy, Orsay, Rennes, Sophia-Antipolis, and Toulouse).

We used heterogeneous clusters in order to generate the maximum workload imbalance. All the nodes were booted under Linux on Grid'5000. Nodes were reserved by the reservation system which ensures that no other user could log on them during the experiments. We conducted several experiments, by varying the number of sites, clusters and computational nodes. Due to pages limitation, we will present in what follows only the results obtained by using two sites, each site containing two clusters and with 16 computational nodes distributed as follows: three nodes/cluster1, two nodes/cluster2, four nodes/cluster3 and seven nodes/cluster4. We allocated clusters with different sizes to show the effectiveness of our approach in dealing with the heterogeneity of the system.

The datasets used in tests are synthetically generated. Table 1 shows the datasets characteristics.

Table 1: Transactional databases characteristics

	DataSet1	**DataSet2**
Database size	70 MB	100 MB
Transactions number	1000000	1300000
Items number	4000	4000
Average transaction size	20	25

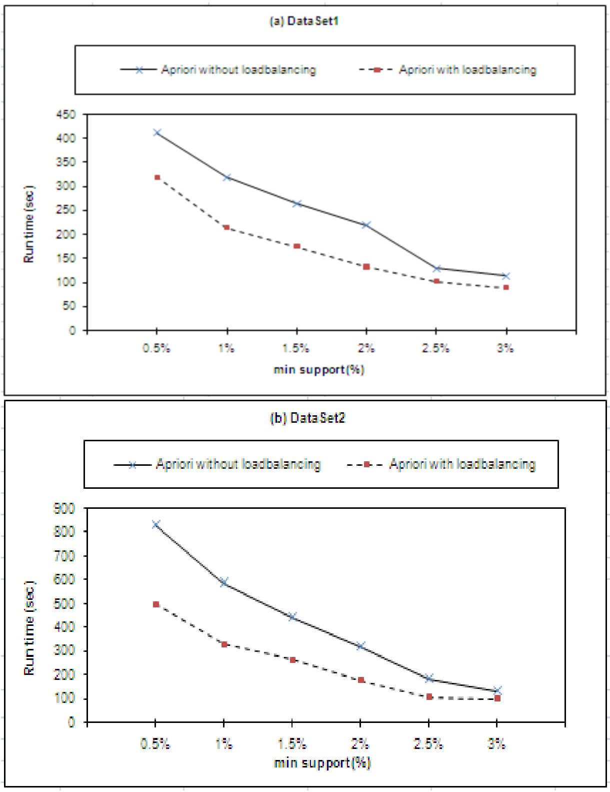


Fig. 3: Run time with and without load balancing for different support values.

Figure 3 displays the execution time obtained from running the parallel version of Apriori without the work load balancing strategy and the time obtained when the strategy is embedded in the parallel implementation. The database is initially partitioned over different sites, where the size of different portions depends on the site's capacity (CPU speed, memory size, available disk space …). We can clearly see that the parallel execution time with work load balancing outperforms the time needed for the parallel execution without taking care to the load imbalance that may occur during the execution of the association rule mining algorithm. Our work load balancing strategy has reduced the execution time of DataSet2 about 40% for very small supports and 20% for larger supports. This shows that as the size of the dataset is large and with small support values (which increases the size of generated candidate itemsets), the amount of work increases. Therefore the probability of work load imbalance between processors is bigger.

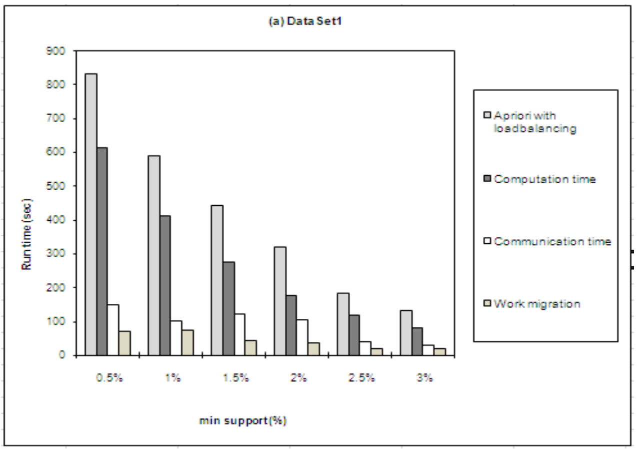


Fig. 4: Run time, communication time and workload balancing time for dataSet1.

Figure 4 shows the time needed for workload balancing (work migration and communication). It is clear that computation time dominates the time needed for communication and work migration, which means that the overhead caused by the proposed workload balancing strategy could be negligible. We also tried to use small data sets (like mushroom and chess with Kb sizes) and the results from the experiments show us one important issue of data mining on grids. When the data set used is not big enough, the number of parallel nodes used should be decreased. That is because the communication overhead is too big when compared with computing time needed for smaller sets.

6 Conclusion

Data mining algorithms have a dynamic nature during execution time which causes load-imbalance between the different processing nodes. Such algorithms require dynamic load balancers that adjust the decomposition as the computation proceeds. Numerous static load balancing strategies have been developed where dynamic load balancing still an open and challenging research area. In this article we developed a dynamic load balancing strategy for association rule mining algorithms, with candidate itemsets generation, under a Grid computing environment. Experimentations showed that our strategy succeeded in achieving better use of the Grid architecture assuming load balancing and this for large sized datasets. In the future, we plan to study the effect of the database type (dense and sparse) on our strategy. We also aim to adapt our strategy to association rule mining algorithms without candidate itemsets generation.

7 References

[1] F. Cappello, E. Caron, M. Dayde, F. Desprez, Y. Jegou, P. Vicat-Blanc Primet, E. Jeannot, S. Lanteri, J. Leduc, N. Melab, G. Mornet, B. Quetier, O. Richard, Grid'5000: a large scale and highly reconfigurable grid experimental testbed, in: SC'05: Proc. The 6th IEEE/ACM International Workshop on Grid Computing, pp. 99–106, Washington, USA, 2005.

[2] I. Foster and C. Kesselman, The Grid2: Blue print for a New Computing Infrastructure. Morgan Kaufmann, 2003.

[3] J. Han and M.. Kamber. Data Mining : concepts and techniques. Maurgan Kaufman Publishers, 2000.

[4] K. Devine, E. Boman, R. Heaphy and B. Hendrickson, "New Challenges in Dynamic Load Balancing". Appl. Num. Maths, Vol.52, issues 2-3, 133-152, 2005.

[5] K. Wang, L. Tang, J. Han and J. Liu, "Top Down FP-Growth for Association Rule Mining". In Proc. Of the 6th Pacific-Assia Conf. on Advances in Knowledge Discovery and Data Mining, Taipei, pp. 334-370, 2002.

[6] M. H. Willebeek-LeMair and A. P. Reeves, "Strategies for Dynamic Load Balancing on Highly Parallel Computers". IEEE Transactions on Parallel and Distributed Systems, Vol. 4, No. 9, pages 979-993, 1993.

[7] M. J. Zaki, "Parallel and Distributed Association Mining: a Survey". IEEE Concurrency, 7(4): pp14-25, 1999.

[8] M. J. Zaki, S. Parthasarathy, M. Ogihara and W. Li. "New Algorithms for Fast Discovery of Association Rules". University of Rochester, Technical Report 651, 1997.

[9] M.S. Perez, A. Sanchez, V. Robles, P. Herrero, J. Pena, Design and Implementation of a data mining grid-aware architecture, Future Generation Computing Systems 23 (1), pp 42–47, 2007.

[10] R. Agrawal and J. C. Shafer. "Parallel Mining of Association Rules". IEEE Transactions on Knowledge and Data Engineering , 8:962-969, 1996.

[11] R. Agrawal and R. Srikant. "Fast Algorithms for Mining Association Rules in Large Databases". In Proc. of the Int'l Conf of VLDB'94, pp 478-499, 1994.

[12] S. Orlando, P. Palmerini and R. Perego, "A Scalable Multi-Strategy Algorithm for Counting Frequent Sets". In Proc. Of the 4th International Conference on Knowledge Discovery and Data Mining (KDD), New York, USA, 2002.

[13] T. L. Casavant and J. G. Kuhl, "Taxonomy of Scheduling in General Purpose Distributed Computing Systems". IEEE Transactions on Software Engineering, 14(2): 141, 1988.

[14] V. Fiolet, B. Toursel, Distributed data mining, Scalable Computing: Practice and Experiences 6 (1), pp 99–109, 2005.

[15] W. Kosters and W. Pijls. Apriori, A Depth First Implementation. In Proceedings of the FIMI Workshop of Frequent Itemset Mining Implementation, Melbourne, Florida, USA, 2003.

[16] Y. Li and Z. Lan, "A Survey of Load Balancing in Grid Computing". Computational and information Science, First International Symposium, CIS 2004, Shanghai, China, 2004.

SESSION

KNOWLEDGE MANAGEMENT AND PERFORMANCE

Chair(s)

TBA

Analyzing Academic Communities' Collaboration and Performance

Alireza Abbasi, Liaquat Hossain
Centre for Complex Systems Research, Faculty of Engineering and Information Technologies
The University of Sydney, NSW 2006, Australia

Abstract - *Recently, there has been a sharp increase on the scholars' collaborations and there are pros and cons by the effect of scientific collaboration on each scholar' performance. Most of previous researches study the micro-level collaboration network to investigate the effects of scholars' collaboration network structure on their performance but to our knowledge there so few macro-level collaboration network studies to evaluate the association between academic communities network structure and the communities' academic performance. In this study, we analyze scientific collaboration network structure and network attributes of five information schools and test if there is any link between them and their academic performance. Analysis of collected data shows that the communities' which are lower density and lower network degree centrality (more decentralized) have higher performance. This could be as a result of share more redundant knowledge in the dense and centralized scientific collaboration networks, which is an obstacle for innovation and new ideas.*

Keywords: Scientific collaboration networks, co-authorship, academic community performance, social network analysis

1 Introduction

In recent years, there has been a sharp increase in the number of collaborations between scholars especially internationally. An explanation for the rapid growth of international scientific collaboration has been provided by Luukkonen et al. as well as Wagner and Leydesdorff [1-3]. By jointly publishing a paper, researchers show their knowledge sharing activities, which are essential for knowledge creation. As most scientific output is a result of group work and most research projects are too large for an individual researcher to perform, it often needs scientific cooperation between individuals across national borders [4]. Thus, being more conscious of collaborations in science has led to a sharpened focus on the collaboration issue [5]. There are many studies, such as [6, 7], have been confirmed the importance of scientific collaboration of researchers' performance.

Since scientific collaborations are defined as "interactions taking place within a social context among two or more scientists that facilitates the sharing of meaning and completion of tasks with respect to a mutually shared, super-ordinated goal" [8], those collaborations frequently emerge from, and are perpetuated through, social networks. Social network analysis has produced many results concerning social influence, social groupings, inequality, disease propagation, communication of information, and indeed almost every topic that has interested 20th century sociology [9].As social networks may span disciplinary, organizational, and national boundaries, they can influence collaboration in multiple ways [8].

In this paper, we are aiming to find: How to identify and evaluate the research collaboration structure of academic communities? Can communities' collaboration network structure and attributes explain their performance (productivity)? Thus, for our analysis, we use scholars' publication information that is available on five information systems schools (iSchools) in US. After processing the publication information (e.g., title, authors, publish year), we store the data in a relational database and extract and shaped co-authorship networks of each school. Then by calculating collaboration (co-authorship) network measures and their performance measures (through number of citations each publication received), we test to find if there is relation between scholars' collaboration and their performance.

Following reviewing literature on social network analysis by focusing on the (whole) network level (marco-level) analysis and measures, our data collection method and dataset details has been explained followed by methods and measure that we will use in this study. Then, results of calculation of schools' collaboration network citation-based performance and network measures have been shown. The paper ends with conclusions and talking about research limitations and highlighting contribution.

2 Background

Social networks operate on many levels, from families up to the level of nations. They play a critical role in determining the way problems are solved, organizations are run, markets evolve, and the degree to which individuals succeed in achieving their goals [10]. Social networks have been analyzed to identify areas of strengths and weaknesses within and among research organizations, businesses, and nations as well as to direct scientific development and funding policies [8, 11].

A social network is a set of individuals or groups each of which has connections of some kind to some or all of the others. In the language of social network analysis, the people or groups are called ‘‘actors’’ or “nodes” and the connections ‘‘ties’’ or “links”. Both actors and ties can be defined in different ways depending on the questions of interest. An actor might be a single person, a team, or a company. A tie might be a friendship between two people, collaboration or common member between two teams, or a business relationship between companies [9]. In scientific collaboration network actors (nodes) are authors and ties (links) are co-authorship relations among them. A tie exists between each two actors if two scholars have at least a coauthored paper. Constructing collaboration (co-authorship) networks of scholars is widely studied so far [6, 8, 12-19] for different fields of study.

2.1 Social cohesion

Social cohesion is often used to explain and develop sociological theories. Members of a cohesive subgroup tend to share information, have homogeneity of thought, identity, beliefs, behavior, even food habits and illnesses [20]. Social cohesion is also believed to influence emergence of consensus among group members [21]. “Examples of cohesive subgroups include religious cults, terrorist cells, criminal gangs, military platoons, sports teams and conferences, work groups etc“ [21].

Modeling a cohesive subgroup mathematically has long been a subject of interest in social network analysis. One of the earliest graph models used for studying cohesive subgroups was the clique model [22]. A clique is a sub graph in which there is a link between any two actors (vertices). However, the clique approach has been criticized for its overly restrictive nature [20, 23] and modeling disadvantages [24, 25]. Clique models idealize three important structural properties that are expected of a cohesive subgroup, namely: familiarity (each node has many neighbors and only a few strangers in the group), reachability (a low diameter, facilitating fast communication between the group members) and robustness (high connectivity, making it difficult to destroy the group by removing members) [21].

2.2 Network Density

Density describes the general level of linkage among the points (actors) in a network (graph) [26]. The more points connected to one another, the denser the network is. So, the densest network is the one which all points are connected with each other but such a networks are very rare.

2.3 Clustering Coefficients

Mainly networks are clustered which means they possess local communities in which a higher than average number of people know one another. One way to check the existence of such clustering in network data is to measure the fraction of ‘‘transitive triples’’ (also called clustering coefficients) in a network [13]. The clustering coefficients of a network is the fraction of ordered triples of nodes A, B, C in which edges AB and BC are present that have edge AC present. In other words, it is the probability that two neighbors of a vertex adjacent to each other. In other words, clustering coefficient is an important property of networks which is “the probability that two of a scientist’s collaborators have themselves collaborated” [27, 28]. Thus, higher clustering coefficients value means it is significantly common for scientists to broker new collaborations between their co-authors.

2.4 The Giant Component

In small networks (few actors and connections), all individuals belong to small group of collaboration or communication. As the total number of connections increases, however, there comes a point at which a giant component forms, "a large group of individuals who are all connected to one another by paths of intermediate acquaintances" [13]. It is important to realize that the collaboration network is fragmented in many clusters. There are several reasons for this. First, in every field there are scientists that do not collaborate at all, that is they are the only authors of all papers on which their name appears. In most research fields, apart from a very small fraction of authors that do not collaborate, all authors belong to a single giant cluster from the very early stages of the field [28].

2.5 Network Centralities

A method used to understand networks and their participants is to evaluate the location of actors in the network. Measuring the network location is about determining the centrality of its actors. Actors centrality measures help determine the importance of the actor in the network. Bavelas [29] was the pioneer who initially investigates formal properties of centrality and proposed several centrality concepts. Later, Freeman [30] found that centrality has an important structural factor influencing leadership, satisfaction, and efficiency.

To examine if a whole network (graph) has a centralized structure. “The concept of density and centralization refer to differing aspects of ‘compactness’ of a graph (network). Density describes the general level of cohesion in a graph; centralization describes the extent to which this cohesion is organized around particular focal points” [26]. The important node centrality measures are:

2.5.1 Degree Centrality

The degree centrality is simply the number of other points connected directly to a point. Necessarily, a central point is not physically in the centre of the network. Degree of an actor is calculated in terms of the number of its adjacent actors.

2.5.2 Closeness Centrality

Freeman [30, 31] proposed closeness in terms of the distance among various points. Sabidussi [32] used the same concept in his work as ‘sum distance’, the sum of the ‘geodesic’ distances (the shortest path between any particular pair of points in a network) to all other points in the network. A point is globally central if it lies at the shortest distance from many other points which means it is ‘close’ to many of the other points in the network. So, simply by calculate the sum of distances of a point to others we will have ‘farness’, how far the point is from other points and then we need to use the inverse of farness as a measure of closeness. As in unconnected networks every point is at an infinite distance from at least one point, closeness centrality of all points would be 0. Thus, Freeman proposed another way for calculating closeness of a point by “*sum of reciprocal distance*” of that point to any other points.

2.5.3 Betweenness Centrality

Freeman [30] yet proposed another concept of centrality which measures the number of times a particular node lies ‘between’ the various other points in the network (graph). Betweenness centrality is defined more precisely as “the number of shortest paths (between all pairs of points) that pass through a given point” [33].

2.5.4 Eigenvector Centrality

Eigenvector centrality assigns relative scores to all actors in the network based on the principle that connections to high-scoring actors contribute more to the score of the node in question than equal connections to low-scoring actors. Based on the idea that an actor is more central if it is in relation with actors that are themselves central [34], it is arguable that the centrality of some node does not only depend on the number of its adjacent actors, but also on their value of centrality. Bonacich [34] defines the centrality c(vi) of a node vi as positive multiple of the sum of adjacent centralities.

2.6 Community (group) Performance

A community can be any group of individuals. In the research context, an individual (i.e. scholars) can belong to different communities. For example, at a university, we can distinguish, in hierarchical order, research groups, departments, schools, colleges, and the entire university. Such a hierarchical classification allows comparing the performance of communities at different levels but here we consider each school as a community (group) and compare their performance.

To assess the performance of scholars, many studies suggest quantifying scholars’ publication activities as a good measure for the performance of scholars. The general idea is that a researcher gets a high visibility in the research community, if the researcher publishes and her publications get cited. The number of citations qualifies the quantity of publications [35]. Hirsch [36] introduced the h-index as a simple measure that combines in a simple way the quantity of publications and the quality of publications (i.e., number of citations). A scholar with an index of h has published h papers, which have been cited by others at least h times. The h-index is also being used by many academic databases (e.g., Web of Science and Scopus) to measure the performance of scholars. Furthermore, the h-index became also the basis for a wide range of new measures [37-42].

There are some studies that suggest measures for evaluating the output of research communities by extending the previously mentioned indices to groups [37, 40, 43-45]. For instance, Prathap (2006) used h-index basic and defined h1-index (h1 papers which have at least h1 citation) and h2-index (h2 researchers who have at least h-index of h2) to quantify the performance of the institutes. Also, some more successive h-indices were defined for measuring journal, publishers and countries, level-wise (Schubert, 2007; Braun et al., 2005). Also Tol (2008) defined g1-index as a successive g-index factor (g1 department members that have a g-index of at least g1 on average).

3 Data and Methods

3.1 Data Sources

For this study, we collected data on five information schools (iSchools): University of Pittsburgh, University of Berkeley, University of Maryland, University of Michigan, and Syracuse University. These schools have been chosen, since they offer similar programs in the area of information management and systems and, because of the fact, that the topic of these schools is new within the university landscape.

The data sources used are the school reports, which include the list of publications of researchers, DBLP, Google Scholar, and ACM portal. Citation data has been taken from Google Scholar and ACM Portal, using AcaSoNet [10]. AcaSoNet is a Web-based application for extracting publication information (i.e., author names, title, publication date, publisher, and number of citations) from the Web. It also extracts relationships (e.g., co-authorships) between researchers and stores the data in the format of tables in its local database.

For its citation counting service, Google Scholar considers a variety of publication databases, which belong to different publishers and list different types of publications. Thus, it produces a higher publication count per researcher and a higher citation count per publication than other citation counting services (e.g., Web of Science of Thomson Reuters, and Scopus) [46]. Consequently, the calculation of the the g-index, if based on Google Scholar, results in higher values than for the other citation counting services. However, Ruane and Tol show that rankings based on Google Scholar have a high rank correlation with rankings based on Web of Science or Scopus [47].

For our analysis, we followed Google Scholars approach and did not differentiate between the different types of publications. Our data covered a period of five years (2001 to 2005), except for the University of Maryland iSchool, which had no data for the year 2002 in their report. To resolve this issue, we substituted the missing data with data of the year 2006.

Despite AcaSoNet, much data cleansing has become necessary in order to allow processing of the extracted publication data. Most of the cleansing was due to the lack of a standard format used for listing publications (e.g., the order of first name and family name of authors, the order of title and publication year and the inaccuracy in writing journal and conference names). After the cleansing of the publication data of the five iSchools, 2122 publications which have received totally 31100 citations, 1806 authors, and 5310 co-authorships were finally available for our analysis.

3.2 Methods and Measures

In this study, we will evaluate networks (groups) measures for five different information schools and compare their group productivity (performance). We apply network (group) level of social network analysis by measuring some quantities of collaboration networks: cohesion measures (e.g., density, clustering coefficient, and distance), region measures (e.g., number of components and blocks) and group centrality measures (e.g., degree, closeness, betweenness and eigenvector centralities). we also measuring some other quantities of collaboration networks using UCINET [48].

3.2.1 Network Density

Simply density of a network is the proportion of exiting links to the maximum possible distinct links that could be exists.

3.2.2 Clustering Coefficients

The clustering coefficient, a quantitative measure of this phenomena, C, can be defined as follows [49]: if node i that has links to k_i other actors in the system. If these k_i actors form a fully connected clique, there are k_i $(k_i - 1)/2$ links between them, but in reality we find much fever. Let us denote by N_i the number of links that connect the selected k_i actors to each other. The clustering coefficient for node i is then $C_i = 2N_i / k_i (k_i - 1)$. The clustering coefficient for the whole network is obtained by averaging C_i over all actors in the system. In simple terms the clustering coefficient of a node in the co-authorship network tells us how much a node’s collaborators are willing to collaborate with each other, and it represents the probability that two of its collaborators wrote a paper together [28].

3.2.3 Network Centrality Measures

A network centralization measure indicates how tightly the network is organized around its most central points. So, the general view is finding differences between most central points’ centrality scores and others’. Then, centralization calculated as a ratio of sum of these differences to the maximum possible sum of differences. So, to calculate network centrality measures first step is to find all actors measures and then find the whole network centralities measures.

3.2.4 Community (Group) Performance

In order to measure community performance, we use a variant of g-index (g1-index) by Tol (2008) which defined as a successive g-index factor (g1 community members that have a g-index of at least g1 on average). It can be calculate using similar formula of Egghe’s [38] g-index but considering on the g-index of scholars in a community rather than citations of publications. For communities having equal g1-index value, we use the average of g-index indices of g1 scholars in the community (g1a-index).

4 Analysis and Results

Table 1 includes detail statistics of the performances measures (e.g., g1 and g1a), number of publications, sum of citations, number of authors and number of collaborations for each school separately. The results based on number of publication, authors and citation indicate that Michigan’s scholars had published more than others following by Pittsburg and Berkeley with just few differences (490, 477 and 468 respectively) while in terms of number of citations those publications received Michigan has the most (10962) with a big gap following by Berkeley (7544) and others. As it is shown the number of authors of Michigan is much higher than others. So, considering average number of citations received per author, Berkeley will be ranked first and then Michigan while considering average number of citations per publications Michigan is ranked first and then Berkeley (similar to total number of citations).

From seventh row to the end of the Table 1, the value of different network measures, which has been discussed in section 3.2, have been shown. Based on the number of actors and links for each school, Michigan and Pittsburg are the densest co-authorship networks and Berkeley and Syracuse have the least. As it is expected, Michigan has the most number of blocks, components and cliques and also average distances among the connected actors (due to having high number of actors and links). Considering clustering coefficient, Pittsburg and Berkeley authors' co-authors are more willing to collaborate with each other (the probability that two of its co-authors wrote a paper together is high). Thus, it will lead to a denser network.

The communities (schools) have been ranked based on their citation-based performance index (g1-index and g1a-index), the same order from left to right in Table 1. While the number of communities (schools) that we analyze is not enough to infer a statistical conclusion about the association between collaboration networks aspects and communities' performance but we find that density have almost a negative relation, the less dense the school the better performance (except for Berkeley which has the second highest density value while they are in the second order). Network degree centrality also shows almost a negative relation, as almost schools with lowest degree centralities have better performance and vice versa (except for Pittsburg).

The results confirmed that the networks with more collaboration (links) have more blocks, components and cliques. Also, the networks with higher number of actors and links have higher network eigenvector and betweenness centrality measures. But interestingly, Berkley with fewest scholars and collaborations (actors and links) are ranked second in terms of performance.

Maryland iSchool is the most central network in terms of degree and closeness network centrality measures and Michigan is the most central in terms of betweenness and eigenvector centralities. Figure 1 and 2 shows Berkeley and Michigan collaboration networks (the least dense and densest networks respectively) as an example. The components are differentiated with different colors. As we can see in the two sample networks there are some broker scholars who connect sub-groups of scholars which otherwise disconnected.

Table 1. Schools' network measures

Measures	Michigan	Berkeley	Syracuse	Pittsburg	Maryland
Performance (g1-index)	**19**	**17**	**15**	**15**	**14**
g1a-index	212.78	124.42	141.40	136.87	130.93
Number of papers	490	468	375	477	312
Sum of citations	**10962**	**7544**	**4917**	**4410**	**3267**
Number of authors (actors)	603	262	280	358	303
Number of collaborations (links)	1486	864	873	1147	907
Density	.016	.032	.024	.025	.035
Average Distance	4.567	3.341	4.077	4.447	3.258
Clustering Coefficients	.814	.821	.664	.898	.696
Number of Blocks	136	64	102	107	57
Number of Components	17	8	11	12	9
The Giant Component	472	130	242	273	252
Number of Cliques	201	84	93	110	118
Network Centrality Measures					
Degree	.012	.022	.026	.018	.036
Closeness	.407	.309	.380	.326	.536
Betweenness	.410	.107	.284	.249	.345
Eigenvector	.841	.031	.025	.783	.036

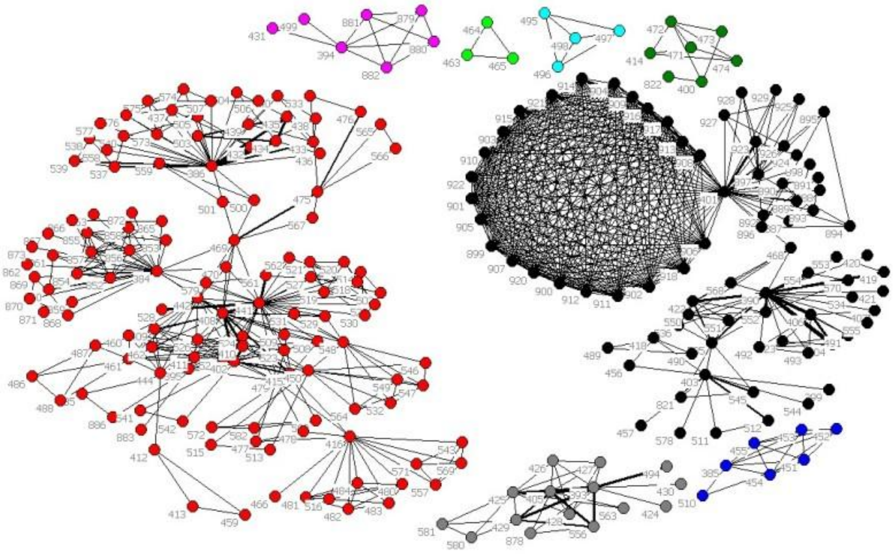


Figure 1. Berkeley co-authorship network and its components

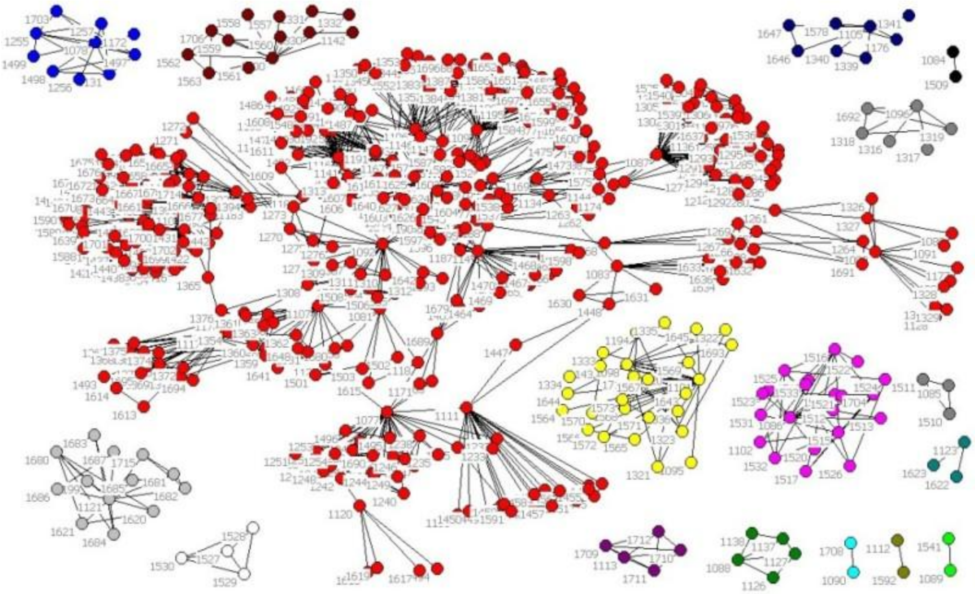


Figure 2. Michigan co-authorship network and its components

5 Conclusion

We analyzed five information schools collaboration network structure in terms of density, number of cliques, blocks, components and network centrality measures. Schools with more authors and collaborations show better eigenvector and betweenness centrality measures. Calculating schools' citation-based performance output using bibliometric indicators (g1-index); we find that network density and degree centrality have almost negative relation with communities' performance. This could be as a result of share more redundant knowledge in the dense and centralized scientific collaboration networks, which is an obstacle for innovation and new ideas. However, with just few communities we have analyzed their performance and network attributes we cannot infer a general conclusion about the relation between network performance and network measures and attributes. This can be considering as our research limitations.

We know that the extent to which researchers co-authors vary among scientific fields and it is usually assumed that this is caused by variation in the level of collaboration. To investigate how scientific field (due to their different collaboration characteristics) influence on the association of researchers' collaborations activities and their performance, it is a need to do similar analysis for several research collaboration groups from different fields as a future work. As this study evaluate the static collaboration network, another extension of this work could be studying dynamicity of the collaboration network and investigate the networks evolutionary changes on their performance.

6 References

[1] T. Luukkonen, *et al.*, "Understanding patterns of international scientific collaboration," *Science, Technology & Human Values,* vol. 17, p. 101, 1992.

[2] T. Luukkonen, *et al.*, "The measurement of international scientific collaboration," *Scientometrics,* vol. 28, pp. 15-36, 1993.

[3] C. S. Wagner and L. Leydesdorff, "Network structure, self-organization, and the growth of international collaboration in science," *Research Policy,* vol. 34, pp. 1608-1618, 2005.

[4] M. Leclerc and J. Gagné, "International scientific cooperation: The continentalization of science," *Scientometrics,* vol. 31, pp. 261-292, 1994.

[5] G. Melin, "Pragmatism and self-organization: Research collaboration on the individual level," *Research policy,* vol. 29, pp. 31-40, 2000.

[6] A. Abbasi and J. Altmann, "On the Correlation between Research Performance and Social Network Analysis Measures Applied to Research Collaboration Networks," in *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual.*, Waikoloa, HI, 2011.

[7] A. Abbasi, *et al.*, "Identifying the Effects of Co-Authorship Networks on the Performance of Scholars: A Correlation and Regression Analysis of Performance Measures and Social Network Analysis Measures," *Journal of Informetrics,* under review.

[8] D. Sonnenwald, "Scientific collaboration: a synthesis of challenges and strategies," *Annual Review of Information Science and Technology,* vol. 41, pp. 643-681, 2007.

[9] M. E. J. Newman, "Scientific collaboration networks. I. Network construction and fundamental results," *Physical review E,* vol. 64, p. 16131, 2001.

[10] A. Abbasi and J. Altmann, "A Social Network System for Analyzing Publication Activities of Researchers," in *Symposium on Collective Intelligence (COLLIN 2010), Advances in Intelligent and Soft Computing*, Hagen, Germany, 2010.

[11] J. Owen-Smith, *et al.*, "A comparison of US and European university-industry relations in the life sciences," *Management Science,* vol. 48, pp. 24-43, 2002.

[12] G. Melin and O. Persson, "Studying research collaboration using co-authorships," *Scientometrics,* vol. 36, pp. 363-377, 1996.

[13] M. E. J. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 98, p. 404, 2001.

[14] M. E. J. Newman, "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality," *Physical review E,* vol. 64, p. 16132, 2001.

[15] J. Moody, "The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999," *American Sociological Review,* vol. 69, p. 213, 2004.

[16] V. Suresh, *et al.*, "Discovering mentorship information from author collaboration networks," *Discovery Science,* vol. 4755, pp. 197-208, 2007.

[17] Y. Jiang, "Locating active actors in the scientific collaboration communities based on interaction topology analyses," *Scientometrics,* vol. 74, pp. 471-482, 2008.

[18] A. Abbasi, *et al.*, "Evaluating scholars based on their academic collaboration activities: two indices, the RC-index and the CC-index, for quantifying collaboration activities of researchers and scientific communities," *Scientometrics,* vol. 83, pp. 1-13, 2010.

[19] F. J. Acedo, *et al.*, "Co Authorship in Management and Organizational Studies: An Empirical and Network Analysis*," *Journal of Management Studies,* vol. 43, pp. 957-983, 2006.

[20] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*: Cambridge Univ Pr, 1994.

[21] B. Balasundaram, *et al.*, "Clique relaxations in social network analysis: The maximum k-plex problem," *Manuscript,* 2008.

[22] R. D. Luce and A. D. Perry, "A method of matrix analysis of group structure," *Psychometrika,* vol. 14, pp. 95-116, 1949.

[23] R. D. Alba, "A graph-theoretic definition of a sociometric clique," *The Journal of Mathematical Sociology,* vol. 3, pp. 113-126, 1973.

[24] L. C. Freeman, "The sociological concept of" group": An empirical test of two models," *American journal of sociology,* vol. 98, pp. 152-166, 1992.

[25] S. B. Seidman and B. L. Foster, "A graph-theoretic generalization of the clique concept," *The Journal of Mathematical Sociology,* vol. 6, pp. 139-154, 1978.

[26] J. Scott, *Social network analysis: a handbook.*: Sage, 1991.

[27] J. W. Grossman, "The evolution of the mathematical research collaboration graph," *Congressus Numerantium,* pp. 201-212, 2002.

[28] A. L. Barabási, *et al.*, "Evolution of the social network of scientific collaborations," *Physica A: Statistical Mechanics and its Applications,* vol. 311, pp. 590-614, 2002.

[29] A. Bavelas, "Communication patterns in task-oriented groups," *Journal of the Acoustical Society of America,* vol. 22, pp. 725-730, 1950.

[30] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks,* vol. 1, pp. 215-239, 1979.

[31] L. C. Freeman, "The gatekeeper, pair-dependency and structural centrality," *Quality and Quantity,* vol. 14, pp. 585-592, 1980.

[32] G. Sabidussi, "The centrality index of a graph," *Psychometrika,* vol. 31, pp. 581-603, 1966.

[33] S. Borgatti, "Centrality and AIDS," *Connections,* vol. 18, pp. 112-114, 1995.

[34] P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," *Journal of Mathematical Sociology,* vol. 2, pp. 113–120, 1972.

[35] S. Lehmann, *et al.*, "Measures for measures," *Nature,* vol. 444, pp. 1003-1004, 2006.

[36] J. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences,* vol. 102, p. 16569, 2005.

[37] J. Altmann, *et al.*, "Evaluating the productivity of researchers and their communities: The RP-index and the CP-index," *International Journal of Computer Science and Applications,* vol. 6, pp. 104–118, 2009.

[38] L. Egghe, "Theory and practise of the g-index," *Scientometrics,* vol. 69, pp. 131-152, 2006.

[39] B. Jin, "H-index: an evaluation indicator proposed by scientist," *Science Focus,* vol. 1, pp. 8-9, 2006.

[40] R. Tol, "A rational, successive g-index applied to economics departments in Ireland," *Journal of Informetrics,* vol. 2, pp. 149-155, 2008.

[41] A. Sidiropoulos, *et al.*, "Generalized Hirsch h-index for disclosing latent facts in citation networks," *Scientometrics,* vol. 72, pp. 253-280, 2007.

[42] P. Batista, *et al.*, "Is it possible to compare researchers with different scientific interests?," *Scientometrics,* vol. 68, pp. 179-189, 2006.

[43] G. Prathap, "Hirsch-type indices for ranking institutions' scientific research output," *Current Science-Bangalore,* vol. 91, p. 1438, 2006.

[44] T. Braun, *et al.*, "A Hirsch-type index for journals," *Scientometrics,* vol. 69, pp. 169-173, 2006.

[45] A. Schubert, "Successive h-indices," *Scientometrics,* vol. 70, pp. 201-205, 2007.

[46] K. Kousha and M. Thelwall, "Google Scholar citations and Google Web/URL citations: a multi-discipline exploratory analysis," *Journal of the American Society for Information Science and Technology,* vol. 58, pp. 1055-1065, 2007.

[47] F. Ruane and R. Tol, "Rational (successive) h-indices: An application to economics in the Republic of Ireland," *Scientometrics,* vol. 75, pp. 395-405, 2008.

[48] S. Borgatti, *et al.*, "Ucinet for windows: Software for social network analysis (version 6)," *Harvard, MA: Analytic Technologies,* 2002.

[49] D. Watts and S. Strogatz, "Collective dynamics of ‘small-world’networks," *Nature,* vol. 393, pp. 440-442, 1998.

A recommender system to share knowledge in software organizations

Juan Pablo Soto[1], Aurora Vizcaíno[2], María de Guadalupe Cota[1], Carlos A. Soto[3], and Roberto Núñez-González[1]

[1]Departmento de Matemáticas, Universidad de Sonora, Hermosillo, Sonora, México
[jpsoto, lcota, ronunez]@gauss.mat.uson.mx

[2]ALARCOS Research Group, University of Castilla – La Mancha, Ciudad Real, Spain
aurora.vizcaino@uclm.es

[2]CETINIA, University Rey Juan Carlos, Móstoles, Spain
carlos.soto@urjc.es

Abstract - *Nowadays knowledge management is a highly important issue for organizations since it helps to improve their competitive advantage. A means to encourage employees to manage knowledge is that of communities of practice in which employees can exchange knowledge and experience. However, members of these communities are often geographically distributed. This fact decreases the feeling of trust between their members and there is consequently less knowledge sharing. In order to avoid both this problem and others which occur in software organizations this paper describes a recommender system designed to support this kind of organizations.*

Keywords: Knowledge Sharing, Software Agents, Intelligent knowledge-based system.

1 Introduction

Knowledge management is, at present, a topic of special interest since knowledge has become organizations' most valuable asset. Suitable knowledge management therefore improves employees' learning and encourages them to share information. A technique which helps organizations to attain the goal of sharing knowledge is that of Communities of Practice (CoP) [1], which can be defined as groups of people who share a concern, a set of problems, or a passion about a topic, and who extend their knowledge and expertise in this area by interacting on an ongoing basis [2].

A further important aspect of CoP is that of membership feeling. Developing community membership implies a clear role, responsibility and the development of trust. Moreover, many authors consider that trust facilitates problem solving by encouraging information exchange and the influence of team members, in the absorption of knowledge, in formulating a sense of self-identity and as the basis of political soundness [3]. Trust has a silent presence in all social interaction [4]. However, the development of trust in a virtual setting may be more difficult than in co-located meetings [5]. At present, CoP are frequently distributed and are supported by technology. This is, in some respects, advantageous as it facilitates communication between people who work or live in different places. However, distribution also makes face-to-face communication more difficult. Nevertheless, this kind of communication is important as it is arguably the "surest way to establish and nurture the human relationships which are grounded in social bonding and are symbolic expressions of commitment" [6]. Because trust is important in all virtual relationships [7], we propose a multi-agent recommender system, based on the trust concept, whose goal is to help software organization teams to take advantage of the knowledge contained within this knowledge recommender system.

The remainder of this paper is organized as follows. In Section 2 we describe the concepts of trust and reputation and how we have attempted to formalize the concept of trust in the domain of virtual communities. Section 3 explains the multi-agent recommender system to support the exchange of knowledge in software organizations. Finally, conclusions are described in Section 4.

2 Trust and reputation

There is no universal agreement on the definition of trust, despite the fact that trust is the basis of economic activities, and without it things such as credit agreements, business contracts and customer confidence would not be possible. One possible definition is that presented by [8] in which the authors define trust as confidence in the ability and intention of an information source to deliver correct information. Wang and Vassileva in [9] define trust as a peer's belief in another peer's capabilities, honesty and reliability based on his/her own direct experiences. In [10] trust is defined as a subjective expectation that one agent has in another's future behavior based on the history of their encounters.

Another concept highly related to trust is that of reputation which is described in the following paragraphs.

Mui et al in [10] define reputation as a perception that one agent has of another's intentions and norms. Barber and Kim [8] define this concept as the amount of trust an agent has in an information source, created through interactions with information sources. Wang and Vassileva in [9] define reputation as a peer's belief in another peer's capabilities, honesty and reliability based on recommendations received from other peers.

In our work we shall adhere to Wang's definitions in considering that the difference between both concepts depends on who has previous experience. Thus, if a person has direct experience of, for example, a knowledge source we can say that this person has a trust value in that knowledge. However, if another person has had the previous experience and recommends a knowledge source to us, we can say that this source has a reputation value.

Our aim is to provide a trust model based on real world social properties of trust in CoPs. Other trust models which take into account social aspects exist, such as the Marsh trust model [11], which has strong sociological foundations. However, the author introduces a large number of variables into the model, making it large and complex. Another model is that of Abdul-Rahman and Hailes [12] in which previous experience, either from the agent itself or from a recommender, are the only factors considered. Therefore, as the previous authors claim, it could be said that an effective practical trust model for virtual environment does not yet exist.

Since the concept of trust in CoPs influences many social properties, this concept has been defined through the consideration of various factors (*position*, *level of expertise, previous experience and intuition*). Some of these factors are objective and some subjective, since when making personal decisions we frequently take into account both types of factors. This issue is that which creates the difficulty in modelling concepts narrowly related to human behaviour such as trust or reputation.

As will later be explained, it is possible decide to give more importance to one factor or to another according to the setting in which the trust model is to be used, and for this reason we have pondered each factor with a weight which either emphasizes a factor or decrease its importance.

3 Multi-agent recommender system

Workers and employers often complain about the time that is spent searching for information. In fact the International Data Corporation in (IDC) [13] estimates that knowledge workers spend between 15 and 35% of their time performing this task. What is worse, the IDC has discovered that 90% of a company's accessible information is used only once. The amount of time spent reworking or re-creating information because it has not be found is increasing at an alarming rate [14]. This section describes a multi-agent recommender system which is used in global software development teams.

In order to understand how the system works, it is necessary to explain the difference between two concepts that will be used: Knowledge Source (KS) and Knowledge Object (KO). A KS is a generator of knowledge which may be: a person, a book, etc., and various KOs can be obtained from it. Therefore, a KO is a piece of knowledge that comes from a KS. In a CoP the main KSs are its members so the tool also considers people as being key KSs. The tool represents each CoP member with an agent called the "User Agent". Each time a person uses a KO his/her user agent reminds him/her that s/he should rate the KO. Four values should be rated, some of which have been adapted from the Wiig Model [15] or from Rao [16], who describes certain attributes that should be considered when analysing knowledge.

- *Importance*: how relevant is this KO for you? In order to discover whether a KO is related to the topic at hand.
- *Useful*: how useful is the KO for the CoP?
- *Time of relevance*: how long that knowledge will be useful, since a piece of knowledge may sometimes be relevant over a certain period of time and may later become obsolete.
- *Granularity*: this indicates whether the knowledge is very general or specific.

The first two values will be used by a User Agent when it needs to evaluate a KS. The third will be used by the system to control whether a KO has become obsolete and the last will categorize the KO.

Each User Agent can assume three types of behaviour or roles similar to the tasks that a person may carry out when working with knowledge management. The User Agent will play one role or another depending upon whether the person that it represents carries out one of the following actions:

- The person contributes new KO to the communities in which s/he is registered. In this case his/her User Agent plays the role of Provider.
- The person uses a KO previously stored in the community. The User Agent will therefore be considered as a Customer.

The person helps other users to achieve their goals by, for instance, giving an evaluation of certain KO. In this case the role is that of Partner. Figure 1 shows that in Community 1 there are two User Agents playing the role of Partner, one User Agent playing the role of Consumer and another playing that of Provider.

The second type of agent within a community is called the *Manager Agent* (represented in black in Figure 1) which must manage and control its community.

In order to facilitate the search for a KO (i.e. documents) the users in a community can choose one topic from those which are available in the community and the User Agent will attempt to discover a KS related to this topic.

The general idea is to consider those KOs which come from trustworthy KSs according to the user's opinion or needs.

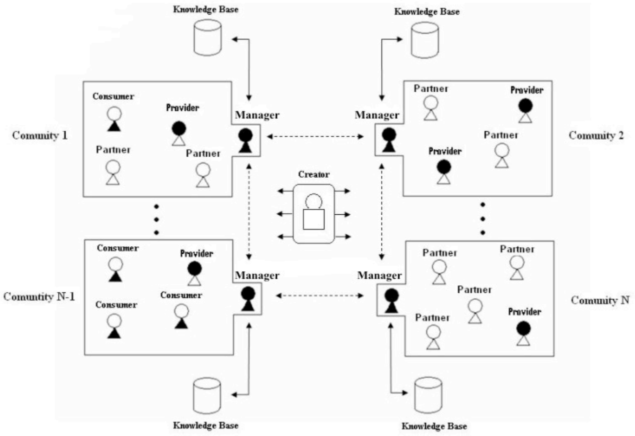


Fig. 1. Communities of agents

In order to discover which KSs are trustworthy the user agents will use the *Trust Formula* (1) which implements the trust model as follows. According to the amount of previous experience, some factors will or will not be used, as will be explained in this section.

$$T_{ij} = wp*P_{\ j} + we*LE_j + wi*I_{ij} + PE_{ij} \qquad (1)$$

Where P_j is the Position of the Agent "*j*" in the CoP or in the organization in which the CoP exists. LE_j is the Level of Expertise that the person represented by the Agent "*j*" has in a particular domain. I_{ij} is the intuition that the User Agent "*i*" has with regard to the Agent "*j*" and finally PE_{ij} is the value of Previous Experience that the Agent "*i*" has had with the Agent "*j*". Finally, w_p w_e, and w_i are weights with which the trust value can be adjusted according to the degree of knowledge that one agent has about another. Therefore, if an Agent "*i*" has had frequent interactions with another Agent "*j*", then Agent "*i*" will give a low weight (or even zero) to w_i since, in this case, Previous Experience is more important than Intuition. The same may occur with w_e, $w_{p.}$ So the weights may have the value between 0 or 1 depending on the previous experience that an agent has.

In order to illustrate how this formula is used, let us imagine that an Agent "*i*" must evaluate how trustworthy another Agent "*j*" is. Agent "*i*" will therefore use Formula (1) in which $\mathbf{T}_{ij}$ is the value of *j*'s trust in the eyes of *i*. We shall now describe how each factor of the formula is calculated.

Position: When a new member joins a community that person must indicate his/her position within the organization and his/her software agent will calculate the Position (P) value of that person by using the following formula:

$$P = UPL/NL \qquad (2)$$

Where UPL is the User's Position Level. NL is the Number of Levels in the community. Therefore, if a community, for instance, has 5 possible position levels then NL=5, and if the new member has a level of UPL=2 then the value of P will be 2/5=0.4. Thus, the different values of P for a community with five levels will be those shown in Table 1.

Levels	Values
1	0.2
2	0.4
3	0.6
4	0.8
5	1

Table 1. Example of Position values

The P values will always be between 0 and 1. Moreover, situations may exist in which P will not been taken into account, for instance in those CoPs in which all the members have the same level or whose members do not wish to consider this criterion. In these cases w_p (weight of position) will be zero and position will not be considered in the formula. A further situation exists in which w_p is equal to zero. This occurs when the value of the Previous Experience PE > U (U being a threshold which is chosen when creating the community). In this case, the agent will use the following formula to calculate the w_p value:

w_p = floor (U/PE_{ij}) being $PE_{ij} > 0$

U = Threshold of Previous Experience.

PE_{ij} = Value of Previous Experience of an agent *i* with another agent *j*.

Thus, when PE_{ij} is greater than a particular threshold U, wp will be 0, and the position factor will consequently be ignored. However, when one agent does not have enough Previous Experience (PE) of another it may use other factors to obtain a trust value. On the other hand, when the agent has had a considerable amount of PE with this agent or with the knowledge that it has provided then it is more appropriate to give more weight to this factor, since previous experience is the key factor in all trust models, as will be described in Section 4. Therefore, although an Agent "*j*" has a high value of position but most of Agent *i*'s previous experience of *j* has not been successful then the position will be ignored. This thus avoids the situation of, for instance, a boss who does not contribute with valuable documents but is considered trustworthy solely because s/he is a boss.

Level of Expertise (LE): This factor is used to represent the level of knowledge and know-how that a person has in a particular domain. In this prototype this factor may change since a person may become more expert in a topic as time goes by.

In this tool, when creating a community the levels of expertise considered are also indicated - for instance: novice, beginner, competent, expert and master. Each time a new member joins a community s/he will indicate the level of expertise that s/he considers him/herself to have. If the members of the community and their LE are known to the creator of the community then that person can introduce them in the tool. Once the LE has been introduced, the user agent will calculate the value for this level by using the following formula:

$$LE=L/NT+AV_j \qquad (3)$$

where L is the level of expertise that was introduced, and NT is the number of levels in the community. The term AV_j is the Adjustment Value for Agent "*j*". This term is extremely important since it will be used to adjust the experience of each user. This term was introduced with the goal of avoiding two situations:

- That a person either deliberately or mistakenly introduces a level of expertise (L) that is not his/her level.
- That, whilst in the community, a person becomes more expert leading to the situation that his/her level of expertise should be adjusted.

Initially AV_j will be 0, and each time a member interacts with a document or information provided by *j* the member will rate this document or information and send this evaluation to the Manager agent in charge of managing the community. The manager agent will verify whether the evaluation is negative or positive. If it is positive, then Agent *j*'s LE can be modified by calculating AV_j as:

$AV_j = (VL_n – VL_{n-1})/PT \quad (n \neq 1)$

If it is negative, then:

$AV_j = - (VL_n – VL_{n-1})/PT \quad (n \neq 1)$

where VL_n is the value of a particular Level of expertise. PT is the Promotion Threshold which is used to determine the number of positive rates necessary to promote a superior Level of expertise. Let us illustrate this with an example. In a community there are four levels with the values shown in Table 2.

Labels	Level (n)	Values (VL)
Beginner	1	0.25
Competent	2	0.50
Expert	3	0.75
Master	4	1.0

Table 2. Position Labels

In this case, the difference between the levels is 0.25, since: $VL_n – VL_{n-1}=0.25$.

In this version of the tool it is assumed that at least 5 rates are necessary to change the level, so PT will be 5, and AV_j will be 0.25/5=0.05. This is therefore the value that will be added when a positive rate is received or that will be subtracted when this rate is negative. With five positive rates (5*0.05=0.25) there is thus a level promotion. In other words, an agent whose position was, for instance, beginner will be promoted to competent.

Intuition: This term is used when the Previous Experience is low and it is necessary to use other factors to calculate a trust value. This is one contribution of our work, since most of the earlier trust models are based solely on previous experience. The agents attempt to emulate human behaviour, as people often trust more in people who are similar to themselves. For instance a person who has to choose between information from two different people will normally choose that which comes from the person who has the same background, same customs etc. as him/her. By following this pattern, the agents compare their own profiles with those of the other agents in order to decide whether a person appears to be trustworthy or not. Therefore, the more similar the profiles of two agents are, for instance *i* and *j*, the greater the I_{ij} value in formula (1) will be. We could say that an agent 'thinks' "I do not know whether I can trust this agent but it has similar features to me so it seems trustworthy". The agents' profiles may alter according to the community in which they are working. In our case, as the data stored in the agents' profiles are 'position' and 'expertise', both these features will be taken into account. Therefore, the factors that the tool compares are: Expertise Difference (ED) and Position Difference (PD).

Thus, the Intuition value of an Agent *i* about *j* (I_{ij}) is:

$$I_{ij}=ED_{ij}+PD_{ij} \qquad (4)$$

where $ED_{ij}=LE_i-LE_j$ and $PD_{ij}=P_i-P_j$

This formula is based on the idea that a person normally has a greater level of trust in people who have a higher level of expertise or who are in a higher position than that person him/herself. Hence, when an agent compares its profile with that of another agent with higher values, the value of intuition will be positive. Let us consider the case of Agent "*i*" which has values of $LE_i=0.5$ and $P_i=0.5$. This agent wishes to know how trustworthy another Agent "*j*" is. In this case the agent will use Formula (1) and, depending on the information that it has about *j*, it will or will not be necessary for it to calculate the intuition factor. In this situation we shall suppose that there is little previous experience and that this must be calculated. The values for the Agent "*j*" are $LE_j = 0.2$ and $P_j = 0.6$.

$I_{ij} = 0.2$ as $ED_{ij} = 0.3$ and $PD_{ij} = -0.1$.

As with position, intuition will or will not be calculated depending on the level of PE (Previous Experience). Thus, the weight of intuition, (see Formula (1)) w_i will be calculated as follows:

$w_i = \text{floor}(U/PE_{ij})$ with $PE_{ij} \neq 0$.

Previous Experience: This factor is the most decisive of all the factors in Formula (1). In fact, all the previous factors depend on it as an agent will decide whether or not to use the remaining factors according to the value of Previous Experience (PE). PE is obtained through the interactions that the agent itself has, so this is direct experience. Each time one agent interacts with another (by interacting we mean that one agent uses a document provided by another), the first agent asks its user to rate that document in order to discover whether the document was: important, useful, up-to-date, very general or very specific.

The agent then labels this interaction with a label from Table 2. A value for Current Experience (CE) is thus obtained which will modify the previous value of PE in accordance with the following formula:

$$PE_{ij}(x) = PE_{ij}(x-1) + CE_{ij}(x) \qquad (5)$$

where $PE_{ij}(x)$ is the value of Previous Experience that the Agent "*i*" has about another Agent "*j*" in an interaction *x*.

$PE_{ij}(x-1)$ is the value of Previous Experience that the Agent "*i*" had about another Agent "*j*" before the interaction *x*.

$CE_{ij}(x)$ is the value of the experience that *i* has had with *j* in the interaction *x*.

For instance, if an Agent "*i*" has just taken part in an interaction with another Agent "*j*", and this is labeled as "bad", but the value of $PE_{ij}(x-1)$ was 0.8, then the value of $PE_{ij}(x)$ will be 0.6 obtained from (0.8 + (-0.2)) where -0.2 is the value of the label 'bad' in Table 3.

Moreover Agent "*i*" will send the manager agent the value of $CE_{ij}(x)$ since as is explained in the Level of Expertise these values can alter the Level of Expertise initially indicated increasing or decreasing it.

Label	PE level
very bad	-0.3
bad	-0.2
medium	+0.1
good	+0.2
very good	+0.3

Table 3. PE Labels

As has previously been explained, the Position and Intuition factors depend on the PE value. When an agent has sufficient PE then Position and Intuition can be ignored, and only the PE and the Level of Expertise will be considered. The latter is also included to ensure that an agent takes advantage not only of its own previous experience but also of that of the other agents since LE is adjusted by the AV_j which comes from other agents' previous experience.

In order to illustrate how the prototype works, let us look at an example. If a user selects a topic and wishes to search for documents related to that subject. his/her User Agent will follow this algorithm:

The input of this algorithm will be a set of KOs. Each KO may or may not have been evaluated previously, so a KO may already have a list of evaluations (along with the identity of each person who evaluated it), or it may appear without any evaluation. This aspect will be taken into account by the algorithm which therefore distinguishes two groups:

Group 1 (G1): This group is formed of the KOs that have been evaluated. This is the most important group since if there are previous evaluations about a KO the agent has more information about it in order to know whether it is advisable to recommend it or not.

Group 2 (G2): these KOs have not been used previously so the tool does not have any evaluations about them. Let us now observe how each group is processed by the algorithm.

In G1 the KOs will be ordered by a Recommendation Rate (RR) which is calculated for each KO. Hence RR_k signifies the Recommendation Rate for a particular KO called *k*, and is obtained from:

$$RR_k = w_1 * TE_{ik} + w_2 * T_{ik} \qquad (6)$$

where TE_{ik} is the weighed mean of the evaluations determined by the trust that an Agent "*i*" has in each evaluator (the person who has previously evaluated that KO "*k*"). TE_i is calculated as:

$$TE_{ik} = \frac{\sum_{j=1}^{n} E_{jk} * T_{ij}}{\sum_{j=1}^{n} T_{ij}} \qquad (7)$$

Therefore, T_{ij} is the trust value that the User Agent "*i*" has in the knowledge source "*j*", and E_{jk} is the evaluation that an Agent "*j*" has made about a particular KO "*k*".

The parameter T_{ik} used in Formula (6) similarly indicates the trust that an Agent "*i*" has in a knowledge source "*k*". Both w_1 and w_2 are weights which are used to adjust the formula. The sum of w_1 and w_2 should be 1. One advantage of formula 6 is that it permits us to change these weights in accordance with the CoPs' preferences, since some CoPs may prefer not to take the T_{ik} into consideration, and in this case w_2 would be zero. Other CoPs might wish to give a little weight to this factor and more weight to TE_i, so w_1 could be 0.8 and w_2 0.2. These weights therefore give more importance (more weight) to the trust obtained by taking into account previous evaluations.

The algorithm would then calculate the RR of each KO related to a topic that a user is interested in and would later show a list with the KOs ordered according to the RR. In the case of there being a high quantity of KOs, then only those with a higher RR would be shown.

Group 2 will use another formula to calculate the RR for each KO since in this case there are no results of previous evaluations of the KOs. The formula used is, therefore:

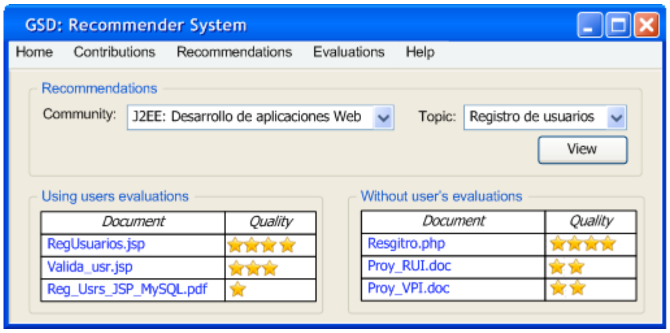
$$RR_k = w_1 * T_{ix} + w_2 * Re_x \qquad (8)$$

where T_{ix} is the Trust that the User Agent "*i*" has in the KS "*x*" which provides the KO "*k*", and Re_x is the reputation that the KS has (according to other member's agents' opinion). This Re_x value is calculated by asking those agents with a higher trust value in the eyes of Agent "*i*" about the KS and this value is obtained by using formula 9.

$$Re_x = \frac{\sum_{j=1}^{n} T_{jx} * T_{ij}}{\sum_{j=1}^{n} T_{ij}} \qquad (9)$$

where T_{jx} is the trust that an Agent "*j*" has in the KS "*x*" and T_{ij} is the trust value that the Agent "*i*" has in Agent "*j*". Therefore, the agent's opinion about KS "*x*" is adjusted by the opinion that the Agent "*i*" has with regard to the agent which is giving its "opinion" (trust value in the KS "*x*").

Figure 2 shows the results of a search sorted by the trust values and divided into two windows in the first one the documents that have evaluations are shown (Group 1) and in the second one those documents what are not been evaluated yet (Group 2).



Community: J2EE: Desarrollo de aplicaciones Web

Topic: Registro de usuarios

Document	Quality
RegUsuarios.jsp	★★★★
Valida_usr.jsp	★★★
Reg_Usrs_JSP_MySQL.pdf	★

Document	Quality
Resgitro.php	★★★★
Proy_RUI.doc	★★
Proy_VPI.doc	★★

Fig. 2. Showing and sorting results

This manner of rating trust helps companies to detect a problem which is increasing in those companies or communities in which employees introduce information which is not valuable because they are rewarded if they contribute knowledge to the community. Thus, if a person introduces a KO that is not related to the community with the aim of obtaining rewards, the situation can be detected, since when another person evaluates that KO, its rate will be low and the 'contributor's value of previous experience will likewise become very low. The community agent is thus able to detect whether there is a “fraudulent” member in the community.

In addition, the recommender system facilitates the exchange and reuse of information, since the most suitable documents are recommended. For this reason, the prototype can also be understood as a knowledge flow enabler, which encourages knowledge reuse in software organizations.

4 Conclusions and future work

The multi-agent system proposed use trust values to recommend documents, which may imply a reduction in users’ overload since it is not necessary for them to search for the most appropriate documents as this task is carried out by their software agents. The aim of the system is to decrease this effect in CoP. One of the main goals of this system, therefore, is help community members to find the most “trustworthy” piece of knowledge and that which is most suitable for each person. We are currently searching for other functionalities that could be added to this tool, such as the detection of experts in a topic, since people who contribute with the most useful documents could, at first sight, be considered as experts in that topic.

Acknowledgements

This work has been funded by CONACYT project (6464001) and the PEGASO/MAGO project (Ministerio de Ciencia e Innovación MICINN and Fondo Europeo de Desarrollo Regional FEDER, TIN2009-13718-C02-01).

5 References

[1] H. Gebert, M. Geib, L. Kolbr, and G. Riempp, "*Knowledge-enabled Customer Relationship Management - Integrating Customer Relationship Management and Knowledge Management Concepts*". In Journal of Knowledge Management, Vol. 7 (5): Pp. 107-123, 2003.

[2] E. Wenger, R. McDermott, and W. Snyder, "*Cultivating communities of practice: a guide to managing knowledge*", Harvard Business School Press, 2002.

[3] A. Abdul-Rahman, and S. Hailes, (2000), "*Supporting Trust in Virtual Communities*". In Proceedings of the 33rd Hawaii International Conference on Systems Sciences (HICSS), IEEE Computer Society, Vol. 6: Pp. 1769-1777.

[4] B. Misztal, "*Trust in Modern Societies*", Polity Press, Cambridge MA., 1996.

[5] S. Jarvenpaa and D. Leidner, "*Communication and trust in global virtual teams*". In Organization Science, Vol. 10 (6), Pp. 791-815, 1999.

[6] P. Hinds and C. McGrath, "*Structures that work: social structure, work structure and coordination ease in geographically distributed teams*", in CSCW: Proceedings of the 20th anniversary conference on computer supported cooperative work, Pp. 343-352, 2006.

[7] D. Paul and R. McDaniel, "*A Field Study of the Effect of Interpersonal Trust on Virtual Collaborative Relationship Performance*". In MIS Quarterly, Vol. 28 No.2, Pp. 183-227, 2004.

[8] K. Barber and J. Kim, "*Belief Revision Process Based on Trust: Simulation Experiments*". In 4th Workshop on Deception, Fraud and Trust in Agent Societies, Pp. 1-12, 2001.

[9] Y. Wang and J. Vassileva, "*Trust and Reputation Model in Peer-to-Peer Networks*". In Proceedings of IEEE Conference on P2P Computing, Pp. 150-157, 2003.

[10] L. Mui, M. Mohtashemi, C. Ang, P. Szolovits, A. Halberstadt, "*Ratings in Distributed Systems: A Bayesian Approach*", in 11th Workshop on Information Technologies and Systems (WITS), 2001.

[11] S. Marsh, "*Formalising Trust as a Computational Concept*". PhD Thesis, University of Stirling, 1994.

[12] A. Abdul-Rahman, and S. Hailes, (2000), "*Supporting Trust in Virtual Communities*". In Proceedings of the 33rd Hawaii International Conference on Systems Sciences (HICSS), IEEE Computer Society, Vol. 6: Pp. 1769-1777.

[13] S. Feldman and C. Sherman, "*The High Cost of not Finding Information*". KM World, Vol. 13(3), 2004.

[14] K. Dalkir, "*Knowledge Management in Theory and Practice*", Elsevier, 2005.

[15] K. Wiig, "*Knowledge Management Foundation*", Schema Press, 1993.

[16] M. Rao, "*Knowledge management tools and techniques: Practitioners and experts evaluate km solutions*", Elsevier, 2005.

Performance Evaluation of a Density-based Clustering Method for Reducing Very Large Spatio-temporal Dataset

Nhien-An Le-Khac
School of Computer Science and Informatics
University College Dublin
Belfield, Dublin 4, Ireland.
an.lekhac@ucd.ie

Michael Whelan
School of Computer Science and Informatics
University College Dublin
Belfield, Dublin 4, Ireland.
michael.whelan@ucd.ie

M-Tahar Kechadi
School of Computer Science and Informatics
University College Dublin
Belfield, Dublin 4, Ireland.
tahar.kechadi@ucd.ie

Abstract—**Spatio-temporal datasets are often very large and difficult to analyse. Today, a lot of interest has arisen towards data-mining techniques to reduce very large spatio-temporal datasets into relevant subsets as well as to help visualisation tools to effectively display the results. Cluster-based mining methods have proven to be successful at reducing the large size of raw data by retrieving its useful knowledge as representatives. As a consequence, instead of dealing with a large size of raw data, we can use these representatives to visualise or to analyse without losing important information. Recently, there is a new approach for reducing large spatio-temporal datasets in the literature. This approach is based on the combination of density-based and graph-based clustering. In this paper, we analyse this approach in detail, especially the impact of initial parameters on the quality of the results. We also present and discuss the application of this approach for different time-steps of hurricane datasets.**

Keywords-spatio-temporal datasets; data reduction; centre-based clustering; density-based clustering; shared nearest neighbours.

I. INTRODUCTION

Today, many natural phenomena present intrinsic spatial and temporal characteristics. Besides applications concerned with climate change, the threat of pandemic diseases, and the monitoring of terrorist movements are some of the newest reasons why the analysis of spatio-temporal data has attracted increasing interest. Spatio-temporal datasets are often very large and difficult to analyse [1][2][3]. Although visualisation techniques are widely recognised to be powerful in analysing these datasets [4], the techniques currently provided in the existing geographical applications are not adequate for decision-support systems when used alone [5]. Data Mining (DM) techniques have been proven to be of significant value for analysing spatio-temporal datasets [6][7]. It is a user-centric, interactive process, where DM experts and domain experts work closely together to gain insight on a given problem. However, several open issues have been identified ranging from the definition of techniques capable of dealing with the huge amounts of spatio-temporal datasets to the development of effective methods for interpreting and presenting the final results. An approach for dealing with the intractable problem of learning from datasets is to reduce this huge set to a smaller subset before analyzing it [2]. It would be convenient if large datasets could be replaced by a small subsets of representative patterns so that the accuracy of estimates (e.g., of probability density, dependencies, class boundaries) obtained from such reduced sets should be comparable to that obtained using the entire dataset.

As there are many reduction techniques presented in literature such as sampling [8], data compression [9], scaling [10], etc., most of them are concerned with data reducing size without paying attention to their geographic properties. Data reduction using cluster-based methods [11][17] have been proposed as a feasible approach to reducing very large dataset by representing large groups of data with different cluster properties such as cluster centres, cluster representatives, etc. Clustering is one of the fundamental techniques in DM. It groups data objects based on the characteristics of the objects and their relationships. It aims at maximising the similarity within a group of objects and the dissimilarity between the groups in order to identify interesting patterns in the underlying data. Some of the benefits of using clustering techniques to analyse spatio-temporal datasets include:

- the visualisation of clusters can help with understanding the structure of spatio-temporal datasets,
- the use of simplistic similarity measures to overcome the complexity of the datasets including the number of attributes, and
- the use of cluster representatives to help to filter (reduce) datasets without losing important and/or interesting information.

Furthermore, we want to exploit the important aspect of spatio-temporal data (i.e., objects that are physically and temporally close tend to be "similar").

In [11][17], the authors presented a cluster-based data reduction strategy when incorporated in a system of exploratory spatio-temporal data mining [10][12], to improve its performance on analysing very large spatio-temporal datasets. In [11] the popular centre-based clustering method known as K-medoids was implemented. This algorithm was chosen for its simplicity, its representatives (medoid points) cannot however reflect adequately all important features of the datasets because this technique is not sensitive to the shape of the datasets (convex). In [17] the authors presented a new solution based on a combination of density-based (DBSCAN [13]) and graph-based clustering (Shared Nearest Neighbour [14]). This solution was named snnDBS. The use of DBSCAN core (or specific core) points as representatives performed much better than centre-based representatives with regards to the shape of the data in spatial dimensions. Preliminary results of this approach were presented in [17] where the authors compared its performance against the scaling and K-medoids based approaches for reducing the Hurricane Isabel datasets. However, these comparisons focused only on the shape of the datasets before and after the reduction process. In this paper, we analyse this snnDBS approach in details, especially the impact of choosing initial parameters on the quality of dataset reduction.

The rest of the paper is organised as follows. We résumé the different approaches for reducing spatio-temporal datasets as the background in Section II. In Section III, we summarise the snnDBS algorithm that is based on the Share Nearest Neighbour (SNN) degree [17]. We show different criteria for evaluating our approach in Section IV. We also evaluate the results of this data reduction technique in the context of applying data mining for analysing spatio-temporal datasets. In Section V, we discuss future work and conclude the paper.

II. BACKGROUND

In spite of much research in the areas of spatio-temporal data analysis and data reduction such as [6][7][10][15], there is very little on data reduction based on DM techniques. Until now, to the best of our knowledge, there are two approaches in this paradigm that were presented [11][17]. In these papers, the authors studied the feasibility of using DM techniques for reducing the dataset size. As mentioned in the previous section, most of the current reduction techniques do not pay attention to geographic properties of spatio-temporal datasets. Hence, they propose to apply DM technique to reduce the dataset size without losing important geographic information by retrieving essential knowledge from these datasets. The main idea is to reduce the size of the data by producing a smaller, knowledge-oriented representation of the dataset, as opposed to compressing the data and then uncompressing it later for reuse. The reason is that authors want to reduce and transform the data so that it can be managed and mined interactively.

In the first approach [11] the centre-based clustering technique that was used is K-Medoids. The k-medoids algorithm chooses the closest data object to the centre of the cluster as the cluster representative. This is very important as the authors use the cluster medoid point's spatial and temporal attributes to visualise the clusters with their representatives (medoid points). This was the main advantage offered by k-medoids algorithm over other centre-based algorithms such as k-means, which would create new values for the cluster centre based on all the members of its cluster but would have no spatial or temporal attributes associated with it. So the goal was to find data objects where each object represents one cluster of raw data. The experimental results showed that knowledge extracted from the mining process can be used as efficient representatives of huge datasets. The advantage of this technique is simple, however its representatives (medoids points) cannot reflect adequately all important features of the datasets. The reason is that this technique is not sensitive to the shape of the datasets (specifically convex shapes).

In the second approach [17], the authors have implemented a combination of density-based and graph-based clustering. They have chosen a density-based method rather than other clustering method such as centre-based because it is efficient with spatial datasets as it takes into account the shape (convex) of the data objects [13]. However, it would be a performance issues when a simple density-based algorithm applied on huge amount of spatial datasets including the differences in density. Indeed, the execution times as well as the choice of suitable parameters have performance impacts on complex density-based algorithms [2]. In this approach, a modified version of DBSCAN [13] is used because it is simple; it is also one of the most efficient density-based algorithms, applied not only in research but also in real applications. In order to cope also with the problem of differences in density, the authors combine DBSCAN with a graph-based clustering algorithm. The experimental results [17] showed that it can preserve the visual shape of large datasets.

Another approach with a combination of SNN and DBSCAN was proposed in [16]. However, it is not in the context of data reduction and it did not take into account the problem of the huge size of the datasets in the context of memory constraint.

III. DENSITY-BASED APPROACH FOR REDUCING VERY LARGE SPATIO-TEMPORAL DATASETS

In this section, we summarise firstly the environment of spatio-temporal data mining where our reduction method will be applied. Then, we present briefly the snnDBS algorithm.

A. Spatio-temporal data mining framework

As described in [10][12], the spatio-temporal data mining framework consists of two layers: mining and visualisation. The mining layer implements a mining process along with the data preparation and interpretation. The visualisation layer contains different visualisation tools that provide complementary functionality to visualise and interpret mined results. One of the main challenges for this framework is how to deal with the very large size of spatio-temporal datasets as they are too large for any traditional mining algorithms to process. Therefore in the mining layer the authors applied a two-pass strategy, were the goal of this strategy is to reduce the size of that data by producing a smaller representation of the dataset so that it can be managed and mined efficiently. The purpose of the first pass is to group the data according to their close similarity and represent these groups without losing any relevant information. Then for the second pass the objective is to apply a mining technique such as clustering, association rules, etc., on the new data representatives to produce new knowledge and prepare for evaluation and interpretation.

B. Cluster-based data reduction

Data reduction using clustering [11] [17] has shown to be a promising method for reducing such datasets. Clustering is one of the fundamental techniques in DM. It aims at maximising the similarity within a group of objects and the dissimilarity between the groups in order to identify interesting structures in the underlying data. Some of the benefits of using clustering techniques to analyse spatio-temporal datasets were listed in section 1, particularly exploiting the fact that objects that are spatially and temporally close tend to be similar. Density-based rather than other clustering methods such as centre-based has proven to be efficient at analysing spatial datasets as it can take into account the shape of the data objects [1]. In this algorithm, DBSCAN was modified because it is simple; it is also one of the most efficient density-based algorithms, applied not only in research but also in real applications for spatial datasets. The authors make the modification so that they can cope with the problem of differences in density, as a result they combine DBSCAN with a SNN similarity. The advantage of SNN is that it addresses the problems of low similarity and differences in density. This is achieved by calculating a new similarity between points based on the number of neighbours they share. This similarity is known as the Shared Nearest Neighbour Degree (SNN_{degree}).

Fig. 1 shows how authors can calculate SNN_{degree} for point P1 with respect to two of its' neighbours P2 and P3. P1 shares three neighbours with P2 and three neighbours with P3 so its' SNN_{degree} is 6. By giving (*Minpts*, ε), in the SNN-DBSCAN algorithm [17], the point x is a core point if:

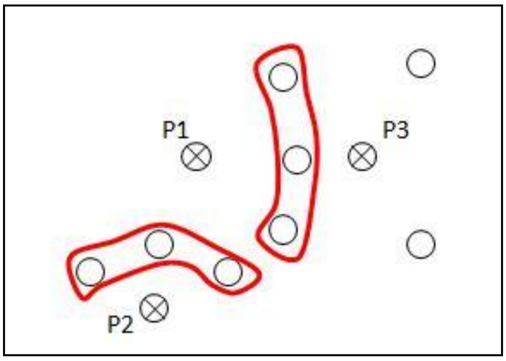


Figure 1. Shared Nearest Neighbour Degree is 6 for point P1.

- For N^k: k-nearest neighbours of x, there exists a neighbour of x, $n_i \in N^k(x)$.
- Given that $SNN_{degree}(n_i) \geq \varepsilon$ and the number of n_i is greater than *MinPts*, $count(n_i) \leq MinPts$.

Besides, authors also used the following concepts: *core point, specific core point, density-reachable points, density-connected points* defined in [13][18].

Briefly, the reduction method includes the four steps: (1) a distance matrix is built for all datasets. (2) An SNN graph is built from this distance matrix. Similarity degree of each data object is also computed in this step. The two parameters ε and *Minpts* are selected based on these similarity degrees. (3) snnDBS algorithm is carried out on the datasets to determine core objects, specific core objects, density-reachable objects, density-connected objects. These objects are defined as above. Clusters are also built based on *core objects* and *density-reachable* features in this step. Data objects which do not belong to any cluster will be considered as noise objects. (4) *Core objects* or *specific core objects* are selected as cluster representatives that form a new (meta-) dataset. This dataset can then be analysed and produce useful information (i.e. models, patterns, rules, etc.) by applying other DM techniques (second pass of our framework). It is important to note that data objects that have a very high similarity between each other can be grouped together in the same clusters. As a result of this pass, the new dataset is much smaller than the original data without losing any important information from the data that could have an adverse effect on the result obtained from mining the data at a later stage.

IV. Evaluation and Analysis

A. Criteria

In the previous work [17], the authors evaluated their snnDBS approach by two main criteria: the shape of the datasets and the ratio of data size before and after the reducing process. They also compared this approach with their two other approaches: scaling and K-Medoids clustering. Based on these criteria, the authors showed that their snnDBS approach gains the reduction ratio: the size of the reduced datasets is about 10% of whole datasets. Besides, they also showed that their approach is efficient in terms of preserving the shape of the datasets before and after the reduction. However, their analysis was only carried out for some locations of the datasets such as hurricane eyes, few positions on the borders, etc.

As mentioned in Section I, spatio-temporal datasets are very large and complex, they are therefore difficult to analyse and visualise. Current research such as [10][11] proposed two-pass strategy to reduce the size of these datasets and then carry out data mining techniques in order to retrieve useful knowledge as well as facilitate the visualisation. Consequently, the reduction process should guarantee not only an efficient reduction ratio but also preserve the important information of the datasets. In this paper, we continue this work by analysing the snnDBS method in details. We propose three criteria:

- The optimal parameters for the algorithm.
- The number of clusters that the algorithm produces.
- The shape of the clusters produced by this algorithm.

The first criterion is very important. As discussed in [2][13], the performance of all density-based clustering is based on the choice of initial parameters. These parameters depend on different algorithms, and the methods for selecting efficient parameters are always a challenge. As the snnDBS method is based on the DBSCAN algorithm, the two parameters that are considered in this section are (*Minpts*,ε)[17]. On the other hand, the second criterion is only focused on the snnDBS method. Although this method only takes into account representatives objects (*specific core points*) rather than the clusters themselves, the number of clusters also affects the representatives selected. If this number is too high, then there are some clusters of which sizes are relatively small compared to the whole dataset. This means that the representatives of these clusters would be noise objects. Indeed, the number of clusters created depends also on the choice of initial parameters. The last criteria relates to the shape of clusters created by the snnDBS approach. This criteria is used to guarantee that representatives selected from clusters preserve the important information (the shape in this case) of the spatio-temporal datasets. The experimentation details and a discussion are given below.

B. Experiments

The dataset is the Isabel hurricane data [19] produced by the US National Centre for Atmospheric Research (NCAR). It covers a period of 48 hours (time-steps). Each time-step contains several atmospheric variables. The grid resolution is 500×500×100. The total size of all files is more than 60GB (~ 1.25 GB for each time-step). Datasets of each time-step include 13 non-spatio attributes, so-called dimensions. In this evaluation, QCLOUD is chosen for analysis; it is the weight of the cloud water measured at each point of the grid. The range of QCLOUD value is [0…0.00332]. Totally, the testing dataset contains around 25 million data points of four dimensions X, Y, Z, QCLOUD for each time step. The evaluation is carried out on six different time-steps: 2, 10, 18, 26, 34 and 42. We also filter the NULL value and land value in the testing data.

C. Analysis

The first issue that has to be dealt with is deciding on what would be good values for the parameters for the algorithm snnDBS. In general, as the snnDBS approach is a DBSCAN-based algorithm, a brief test with the original DBSCAN where a sample of datasets is taken into account can be carried out to estimate the initial values of the two parameters (*Minpts*,ε). However, these two parameters of snnDBS are based on the SNN_{degree} (cf. III.B) not the Euclidean metric. Therefore, we cannot apply methods in the literature such as [13] to determine these initial values. Consequently, we use three steps to evaluate these parameters. In the first step, we execute the experiments with all potential pairs of parameters on the sampling set of QCLOUD. Then we test some candidates for the time-step 2 in the second step. Finally, the evaluation is carried out for all time-steps of the QCLOUD in the last step. The main factors that affected the choice of optimal parameters are the minimum number of noise points so that very little information would be lost, the maximum number of core points so that the shape of the data will be as close as possible to the original in terms of preserving important knowledge of spatio-temporal datasets; and the minimum number of representatives so that the maximum amount of reduction can be achieved.

Table I shows the results of the step 2 where values for the pair (*Minpts*,ε) are varied from (3,4) to (12,8). Note that the selection of this range is based on the first step mentioned above. By observing this table, we recognise that:

- There is a small difference in the number of *specific core points* (chosen as representatives, cf. III.B) among these pairs of (*Minpts*,ε) with the exception of the case (12,8). These representatives cover approximately 10% of the whole datasets. Although the smaller value of *specific core points* is better in term of reducing the size of large datasets, we should also take into account other impact factors such as the number of *noise points*, *core points*, and clusters created.
- The number of *core points* depends on both parameters ε and *Minpts*. It is decreasing when both parameters are increasing. The reason is that more data points have the same core point when these parameters increased.
- The number of noise points depends on either ε or *Minpts*. When either of them is increased, the number of noise points is increasing too. For example, when the pair (*Minpts*,ε) are (3,12), (12,4), (12,8), the noise points gain 23%, 46% and 72% respectively.
- ε and/or *Minpts* also affect(s) to the number of clusters created. Concretely, it is linear to these parameters.

By combining all these observations, we can conclude that the optimal pair of parameters (*Minpts*,ε) is (3,4). With this pair, we obtain the minimum number of noise points (47), the maximum number of core points (939292, around 99% of the whole datasets) and an appropriate number of representatives (103902, around 10% of the whole datasets). Besides, the number of the clusters created (1552) is also minimum compared to other pairs of parameters evaluated.

Table II shows the results of the evaluation of snnDBS algorithm for different time-steps of QCLOUD with (*Minpts*,ε) is equal to (3,4). We also obtain appropriate values of noise points (only 0.02% of the whole datasets), of core points (97% of the whole datasets) and representatives (10% of the whole datasets). These results also confirm that (3,4) is the optimum parameters for this snnDBS approach.

Fig.2 shows the 3-dimensional views of the 20 largest clusters for three of the chosen QCLOUD time step (2, 18, 42) datasets produced by the snnDBS algorithm with its best parameters ε=4 and minPts=3. These 20 clusters are in different colours from blue to grey and the rest is in black. By observing this figure, we notice that the shape and movement of the hurricane is clearly visible. This is indicated by the distinct swirling shape in each of the graphs. In Fig.3 the 2-dimensional views of Fig.2 is shown. From our results we can clearly see the movement of the hurricane from time step 2, 18 and 42. Also we can identify key hurricane features such as the hurricane eye, eye wall and the swirling rainbands.

In order to evaluate the quality of the visual shape of the clusters produced by this snnDBS algorithm, we also carried out the DBSCAN algorithm [13] for the whole QCLOUD datasets. Fig.4 is the 3-dimensional views of the 20 largest clusters for three of the QCLOUD time step (2, 18, 42) datasets produced by the DBSCAN algorithm. By observing this figure as well as comparing it to the Fig.2, we recognize that the snnDBS algorithm can preserve the important information of QCLOUD. Concretely, it preserves most of the important visual shapes comparing to the DBSCAN. Note that snnDBS algorithm produces datasets of which size is only 10% of the whole QCLOUD datasets. Furthermore, it is obviously easier to view than the DBSCAN results due to this small size of datasets.

However, there is a performance issue on the number of clusters created by snnDBS algorithm. Instead of this number being optimal, it is still relatively high compared to the whole datasets. For example, in the case of time-step 2, the number of clusters is 1552 of 946569 data points. This means that one cluster could have around 600 points (0.06% of whole datasets). For the other time-steps, one cluster contains on average around 0.02% of whole datasets.

This issue can lead to the problem of fragmentation where some clusters could contain only noise points (not important information for the further analysis by experts) as mentioned above. It is not easy to determine the original cause of this issue. In our case, by studying the datasets, carrying out more experiments and analyzing the snnDBS in details, we can conclude that this issue is caused by the large variety in density of the datasets. Concretely, the snnDBS does not take into account the Euclidean distance between two data points but only their SNN_{degree}. Consequently, some data points have very far neighbours compared to some of the other points. In this case, a fixed pair of parameters (*Minpts*,ε) would be an issue of performance as in our case. Furthermore, it is not flexible enough in merging sub-clusters [13] to build main clusters in the application. Normally, a multi-pair of parameters which are applied to different areas of datasets would be considered as a solution. However, the complexity of the algorithm and the running time costs would cause other performance issues.

V. CONCLUSION AND FUTURE WORK

In this paper, we evaluated in detail the performance of a density-based clustering algorithm for reducing very large spatio-temporal datasets: snnDBS described in [19]. It is a combination of density-based (DBSCAN [13]) and graph-based clustering (Shared Nearest Neighbour [14]). It uses *specific core points* as representatives to build a reduced datasets. We proposed firstly different criteria that are used to evaluate this approach. We also explained the reasons of choosing these criteria. Next, we showed the evaluation of snnDBS by each criterion from choosing optimal parameters to the creation of quality visual shapes. Furthermore, we compared snnDBS with DBCAN algorithm to show that this new approach is efficient in terms of data size reduction while preserving their important information. We also show that this snnDBS approach is efficient for cases with small or medium density variations of different investigating areas.

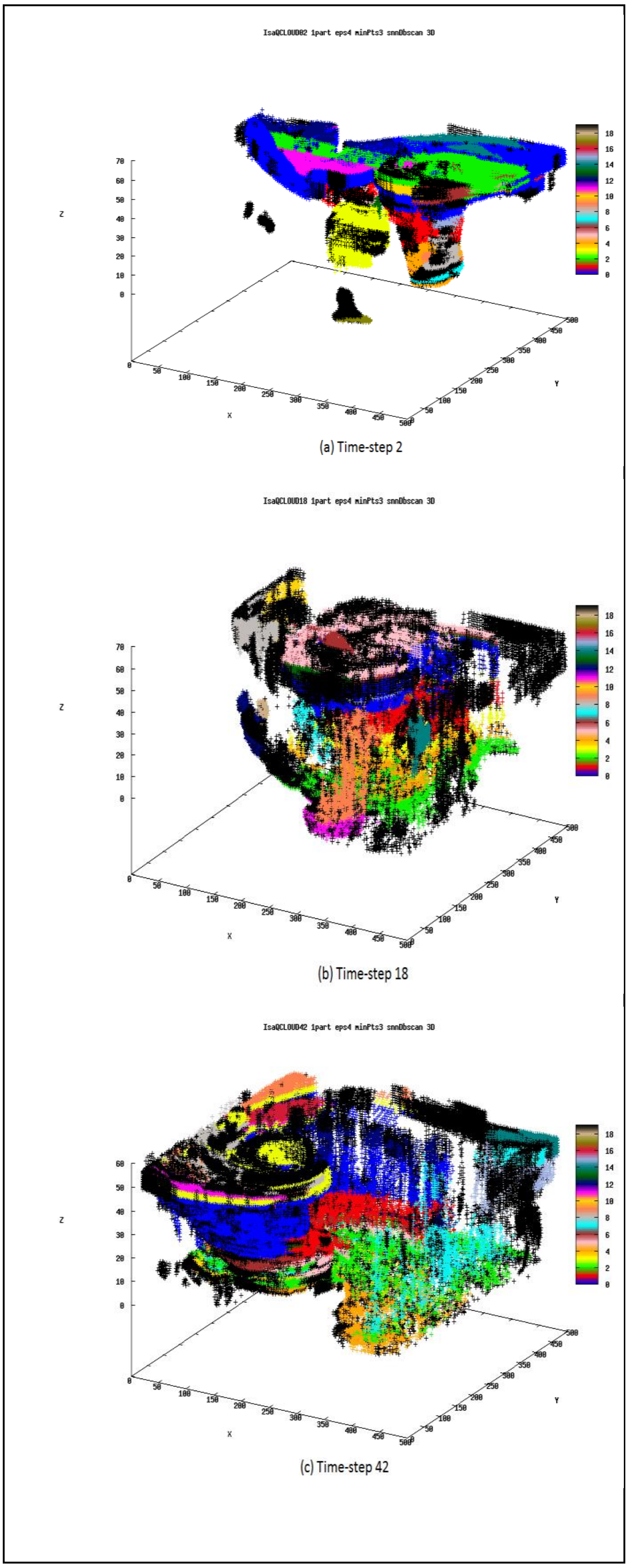
In the future we intend to provide a more efficient algorithm to take into account the problem of large variation in density of datasets. Besides, we continue to carry out an extensive evaluation involving the analysis of more dimensions that have larger datasets than QCLOUD. Furthermore parallel and distributed techniques will also be studied to carry out our approach on both multi-core and distributed architectures in order to prove its robustness.

TABLE I. THE RESULTS OF SNNDBS FOR QCLOUD TIME STEP 2 WITH DIFFERENT PAIRS OF PARAMETERS

Algorithms	Parameters and results						
	Minpts	ε	*size*	*noise*	*core*	*specCore*	*clusters*
snnDBS	3	4	946569	47	939292	103902	1552
	12	4	946569	434433	126298	103902	3093
	3	12	946569	216393	432936	87829	4489
	4	6	946569	673	922787	103835	2210
	8	8	946569	111215	386247	95570	1894
	12	8	946569	687307	42066	34284	3676

TABLE II. RESULTS FOR SNNCDBS ON DIFFERENT QCLOUD TIME STEPS

time steps	Parameters and results						
	Minpts	ε	*size*	*noise*	*core*	*specCore*	*clusters*
QCLOUD10	3	4	809244	223	790649	87843	3470
QCLOUD18	3	4	775150	213	755753	82122	3683
QCLOUD26	3	4	967519	273	944834	103174	4178
QCLOUD34	3	4	1134109	308	1108795	121351	4816
QCLOUD42	3	4	1243338	334	1219233	133355	4701

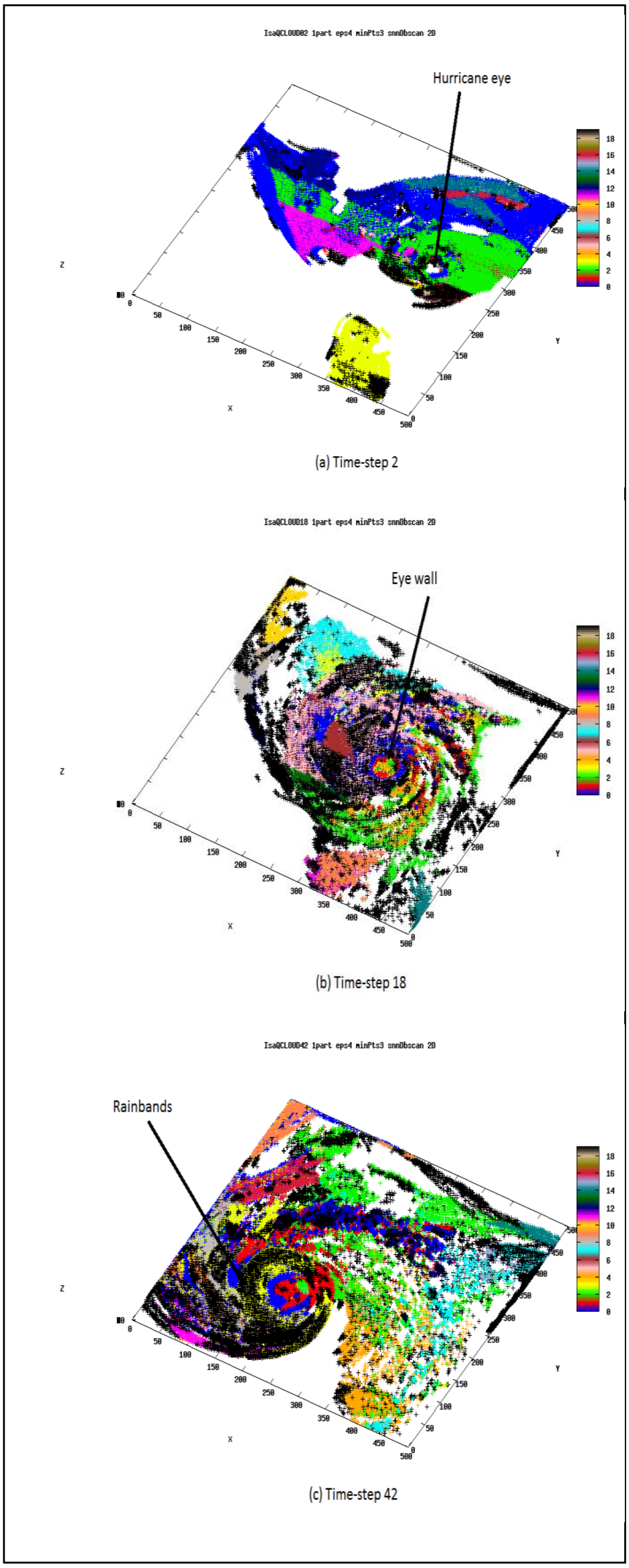


(a) Time-step 2

(b) Time-step 18

(c) Time-step 42

Figure 2. 20 largest clusters produced by snnDBS.

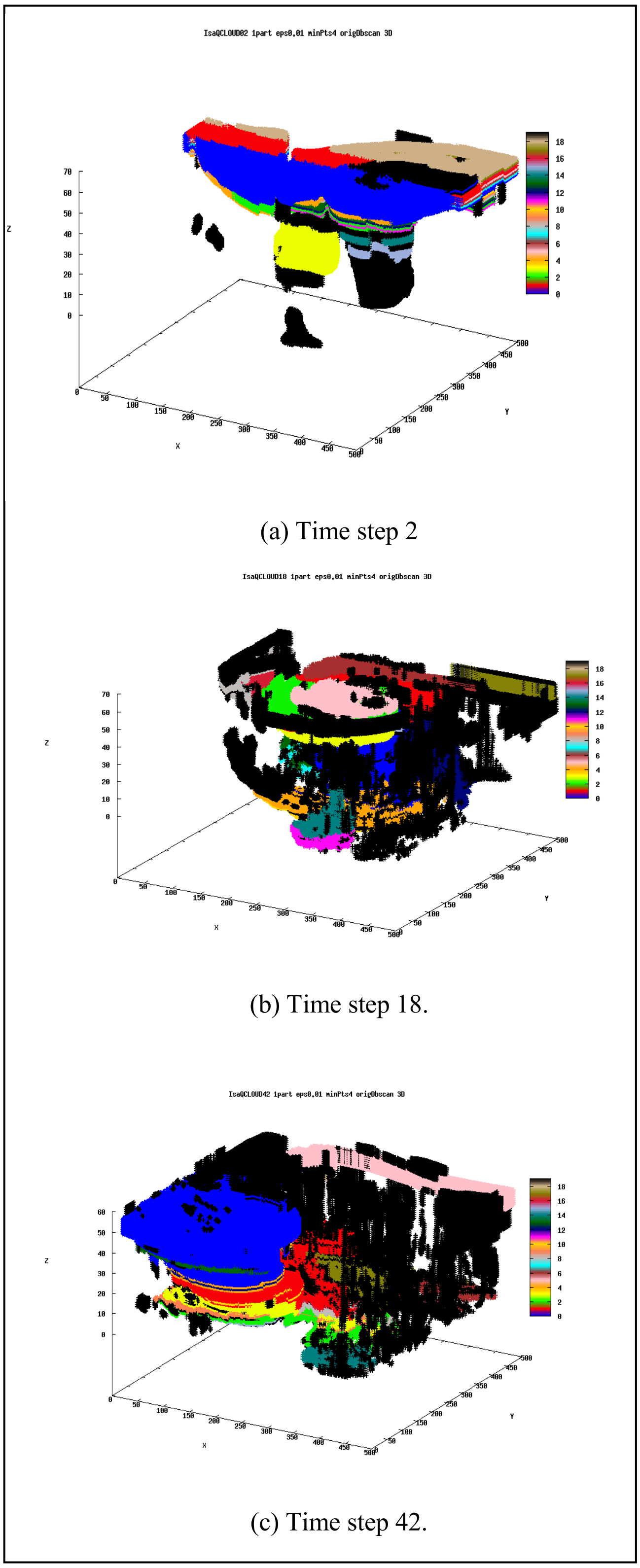


(a) Time-step 2

(b) Time-step 18

(c) Time-step 42

Figure 3. 2-dimensional view of Fig. 3 showing movement over time.



(a) Time step 2

(b) Time step 18.

(c) Time step 42.

Figure 4. 20 largest cluster produced by DBSCAN.

REFERENCES

[1] M. H. Dunham, Data Mining: Introductory and Advanced Topics, Prentice Hall, 2003.

[2] P-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, Addison Wesley, 2006.

[3] N. Ye, (ed), The Handbook of Data Mining. Lawrence Erlbaum Associates Publishers, Mahwah, New Jersey, USA, 2003.

[4] W. L. Johnston, “Model visualisation,” in: Information Visualisation in Data Mining and Knowledge Discovery, Morgan Kaufmann, Los Altos, CA, pp. 223–227, 2001.

[5] N., Andrienko N., G., Andrienko, and P. Gatalsky, “Exploratory Spatio-Temporal Visualisation: an Analytical Review”, Journal of Visual Languages and Computing, special issue on Visual Data Mining. December, v.14 (6), pp. 503-541, 2003.

[6] J. F. Roddick, K. Hornsby, and M. Spiliopoulou, “An updated bibliography of temporal, spatial, and spatio-temporal data mining research,” in Proceedings of the First International Workshop on Temporal, Spatial, and Spatio-Temporal Data Mining-Revised Papers, pp.147-164, September 12, 2000.

[7] J. F. Roddick, and B. G. Lees, “Paradigms for spatial and spatio-temporal data mining,” In Geographic Data Mining and Knowledge Discovery. Miller H. and Han J. (Eds), Taylor & Francis, 2001.

[8] J. Kivinen, and H. Mannila, “The power of sampling in knowledge discovery,” Proceedings of the ACM SIGACT-SIGMOD-SIGART, pp.77-85, Minneapolis, Minnesota, United States, May 24 - 27, 1994.

[9] K. Sayood, Introduction to Data Compression, 2nd Ed., Morgan Kaufmann, 2000.

[10] P. Compieta, S. Di Martino, M. Bertolotto, F. Ferrucci and T. Kechadi, “Exploratory spatio-temporal data mining and visualization,” Journal of Visual Languages and Computing, 18, 3, pp.255-279, June, 2007.

[11] M. Whelan, N-A. Le-Khac and M-T. Kecahdi, “Data reduction in very large spatio-temporal datasets,” IEEE International Workshop On Cooperative Knowledge Discovery and Data Mining 2010 (WETICE 2010), Larissa, Greece, June 2010.

[12] M. Bertolotto, S. Di Martino, F. Ferrucci and T. Kechadi, “Towards a framework for mining and analysing spatio-temporal datasets,” International Journal of Geographical Information Science, 21, 8, pp.895-906, July 2007.

[13] M. Ester, H-P. Kriegel, J. Sander and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD'96), pp.226-231, Portland, OR, USA, 1996.

[14] R. A. Jarvis and E. A. Patrick, “Clustering using a similarity measure based on shared nearest neighbours,” IEEE Transactions on Computers, C-22(11) pp.1025-1034, 1973.

[15] D. R. Wilson and T. R. Martinez, “Reduction Techniques for Instance-based Learning Algorithm,” Machine Learning 33, 3, 257–286, 2000.

[16] L. Ertöz, M. Steinbach and V. Kumar, “Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data,” in Proceedings of Second SIAM International Conference on Data Mining, 2003.

[17] N-A. Le Khac, M. Bue, M. Whelan, and M-T. Kechadi, “A knowledge-based data reduction for very large spatio-temporal datasets,” International Conference on Advanced Data Mining and Applications, (ADMA’2010), Springer Verlag LNCS/LNAI, Chongquing, China, November 19-21, 2010.

[18] E., Januzaj, H-P., Kriegel, M., Pfeifle, “DBDC: Density-Based Distributed Clustering”, Proc. 9th Int. Conf. on Extending Database Technology (EDBT) pp.88-105, Greece, 2004

[19] National Hurricane Center, Tropical Cyclone Report: Hurricane Isabel, http://www.tpc.ncep.noaa.gov/2003isabe l.shtml, 2003.

Database module of Vijjana, a Pragmatic Model for Collaborative, Self-organizing, Domain Centric Knowledge Networks

Amara Satish Kumar, R. Reddy, L. Wang, S. Reddy
Lane of CSEE Department
West Virginia University
Morgantown , West Virginia

Abstract - ***To develop the Database module of Vijjana and visualize the data in the form of some standard Views. According to the standards of IEEE 1484.12.1, a main and well structured relational database of Vijjana is built using MySQL. The database is populated by transferring data from an Open Directory Project (ODP). Data is extracted out of it in the form of an XML and is viewed in the form of a Tree View and a Radial View using visualization techniques.***.

Keywords: Vijjana, Database, Jan Structure, ODP, Prefuse.

1 Introduction

Web plays a vital role in our daily life, a person x will surf through the web daily when he needs information about books, tutorial, health, news, sports shopping and entertainment. Etc. Information on a particular topic is scattered all over the web at various sources each millions of pages away from one another, to cope up with this people use search engines to get all the information on the topic displayed at one place.

There are many search engines like Google, Yahoo, MSN, AOL, EBAY, Netscape etc. Some search engines relies on searching data over the web using keywords specified for specific web pages, but search engines like Google will search the pages through the description of it rather than the specified keywords. This is the reason Google is well used by many users. Even though these search engines are able to provide large amounts of pages to the user, the user still has to surf through the large number of results to find the most appropriate result for his problem. The same person will end up in finding the result sooner if he has a general idea about each and every result or the result which worked the best for his friends working on the same topic etc. On the whole this problem can be solved by developing an agent which could successfully handle the results bound to particular topics or domains and which can also provide the personal opinions of users who have worked or are working on the same topics by saving their comments or rating and enabling more collaboration between them. This framework called Vijjana[2] is going to break through all these problems by providing Pragmatic mechanism for collaboratively building useful knowledge networks in well-bounded domains.

2 Problem Statement

This paper illustrates the way we develop the Database module of Vijjana and visualize the data in the form of some standard Views. According to the standards of IEEE 1484.12.1, a main and well structured relational database of Vijjana is built using MySQL. The database is populated by transferring data from an Open Directory Project (ODP) [3]. Data is extracted out of it in the form of an XML and is viewed in the form of a Tree View and a Radial View using visualization techniques.

3 The Vijjana Model

We define the Vijjana model as:

Vijjana-X = { J, T, R, dA, oA, cA, vA, sA, rA}
where
X = the domain name
J= the collection of Jans in the Vijjana-X
T = the Taxonomy used for classification of Jans
R= the domain specific relations
dA = the discovery agent which find relevant Jans
oA = the organizing agent which interlinks the Jans based on R
cA = the consistency/completeness agent
vA = the visualization agent
sA = the search agent
rA = the rating agent

The markup agent is a sub-agent of the discovery agent. Similarly, the validation agent is a sub-agent of the consistency/completeness agent. We now examine the underlying concepts followed by the markup process and the validation process.

4 VIJJANA DATABASE IEEE 1484.12.1

This is a multipart standard that specifies learning object metadata. This part specifies a conceptual data schema that defines the structure of a metadata instance for a learning object. For this standard, a learning object is defined as any entity—digital or non-digital—that may be used for learning, education, or training. For this standard, a metadata instance for a learning object describes relevant characteristics of the learning object to which it applies. Such characteristics may be grouped in general, life cycle, meta-metadata, educational, technical, educational, rights, relation, annotation, and classification categories.

Metadata is information about an object, be it physical or digital. As the number of objects grows exponentially and our needs for learning expand equally dramatically, the lack of information or metadata about objects places a critical and fundamental constrain on our ability to discover, manage, and use objects. This standard addresses this problem by defining a structure for interoperable descriptions of learning objects. A data element for which the name, explanation, size, ordering, value space, and data type are defined in this standard is known as a Learning Object Metadata (LOM) [7].

Data elements describe a learning object and are grouped into categories. The LOMv1.0 base schema:

The general

Category groups the general information that describes the learning object as a whole.

The lifecycle

Category groups the features related to the history and current state of this learning Object and those who have affected this learning object during its evolution.

The meta-metadata

Category groups information about the metadata instance itself (rather than the learning object that the metadata instance describes).

The technical

Category groups the technical requirements and technical characteristics of the learning object.

The educational

Category groups the educational and pedagogic characteristics of the learning object.

The rights

Category groups the intellectual property rights and conditions of use for the learning object.

The relation

Category groups features the relationship between the learning object another related learning object.

The annotation

Category provides comments on the educational use of the learning object and provides information on when and by whom the comments were created.

The classification

Category describes this learning object in relation to a particular classification system.

5 data layer

5.1 Source Data for Database

As we discussed earlier, information or knowledge is scattered all over the web at different places. As users search on a particular domain or topic we need to have a collection of URL’s for each and every domain. So a resource with a predefined taxonomy with collection of Jan’s for each and every particular domain is necessary. Here Open Directory project (ODP) on Dmoz.org, the largest and most comprehensive Human-Edited Directory of the web with well structured and predefined taxonomy, is used as the core dataset for the database of Vijjana. The Open Directory follows in the footsteps of some of the most important editor/contributor projects of the 20th century. Just as the Oxford English Dictionary became the definitive word on words through the efforts of volunteers, the Open Directory follows in its footsteps to become the definitive catalog of the Web. The Open Directory was founded in the spirit of the Open Source movement, and is the only major directory that is 100% free. There is not, nor will there ever be, a cost to submit a site to the directory, and/or to use the directory's data. The Open Directory data is made available for free to anyone who agrees to comply with our free use license.

The Open Directory [3] powers the core directory services for the Web's largest and most popular search engines and portals, including Netscape Search, AOL Search, Google, Lycos, HotBot, Direct Hit, and hundreds of others. The source of Data is directly form the user himself, when as we discussed whenever a user Mark-Up’s a Jan then that Jan is automatically added to the database if it Is not present in the database and will be added in his Jan’s list. Also the Consistency Agent will periodically check the Jan’s present in the Database and removes dead links so the database shrinks and grows from time to time..

The database of Vijjana consists of:

1. Information or properties of the JAN’s and the
2. Information of the user and the
3. Relation of Each Jan with user like
   a) Who added it
   b) How many times it has been marked up by users.
   c) When it has been last modified
   d) Rating

It also shows the Present Status of JAN whether it is an alive or a dead link.

5.2 JAN Structure

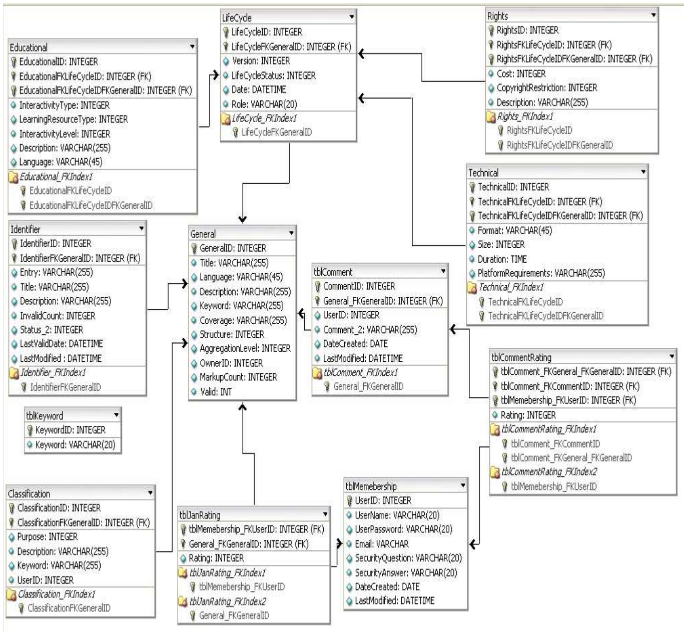


Figure 1: Jan Structure

The Above picture shows the typical schema or structure of the Vijjana Database.

1. The tables General, Identifier and Classification are related to the Collection of the JAN's form the Open Directory Project.
2. The Properties of the JAN are given by the tables:
   a) Educational
   b) Technical
   c) Rights
3. The Status of the Jan is given by the Table Life Cycle.
4. The User Information is given in the table Membership
5. The relation between the user and JAN is given by the tables:
   a) Comment
   b) CommentRating
   c) JanRating
   d) Keyword

5.2.1 Educational

This table describes the followings:

1. The educational and pedagogic characteristics of the learning object. The pedagogical information is useful for users who are involved in achieving a quality learning experience. The users for this metadata include teachers, managers, authors, and learners.
2. The degree of interactivity to which this learning object is categorized. Interactivity in this context refers to the degree to which the learner can influence the aspect or behavior of the learning object. The Interactivity type can be "Active' Learning, " Expositive' Learning or "Mixed" learning "Active" learning (e.g., learning by doing) is supported by content that directly induces productive action by the learner. "Expositive" learning (e.g., passive learning) occurs when the learner's job mainly consists of absorbing the content exposed to him when a learning object blends the active and expositive interactivity types, then its interactivity type is "mixed."

5.2.2 Rights

This table describes the followings:

1. The intellectual property rights and conditions of use for this learning object.
2. Whether copyright or other restrictions apply to the use of this learning object.

5.2.3 Technical:

This table describes

1. The technical requirements and characteristics of this learning object. Technical data type of this learning object.
2. To identify the software needed to access the learning object. Information about other software and hardware requirements.

5.2.4 Lifecycle

This table describes the completion status or condition of this learning object.

5.2.5 Membership:

This table completely relates to the personal information of the user. It contains the registration and login details of every user. Apart form the field of password, the Security Question and Answer will enable more secure login. It can also be used when user forgets his password.

5.2.6 Comment:

This table stores all the comments given by the user about respective JAN's. The field UserID shows the particular user, the field CommentID says the related comment of that user and the field GeneraID shows the particular JAN. It also shows the date the comment is created and last modified.

5.2.7 Comment Rating:

This table shows the user ratings for each and every comment given by different users. For a particular comment given by a User, many other users can rate it according to their wish. The UserID shows the user who has rated it and the CommentID shows the comment on which he has rated it.

5.2.8 JanRating:

This table shows how much the users have rated for each and particular Jan. The UserID gives the information of the user who rated that JAN and the GeneralID gives the information of that particular JAN. Generally a user can rate any Jan irrespective of he has created it or not.

5.3 Mysql DATABASE and Phpmyadmin

All above are implemented using MYSQL[9] database which is an open source database tool. We use its workbench tool to draw the ORM model and another open source tool, phpmyadmin, as our web interface to monitor the database

5.4 Populating the Database

As we have already chosen to use the taxonomy of the Open Directory Project (Dmoz.org) as the core data for the Vijjana Database, the task now is to transfer the data from ODP to our Database [8]. The data in the ODP is in the format of RDF [4] known as Resource Description Framework. Files in this form will use tags to distinguish different resource categories. To transfer these files into a database, the editor of Open Directory Project created a set of tool named ``ODP/dmoz". A famous tool for this is called ``PhpODPWorld"[5].

The script used to import the RDF file into database is developed by Steve who is one of open directory project editor. The following procedure explains in steps to do the RDF to Database transformation:

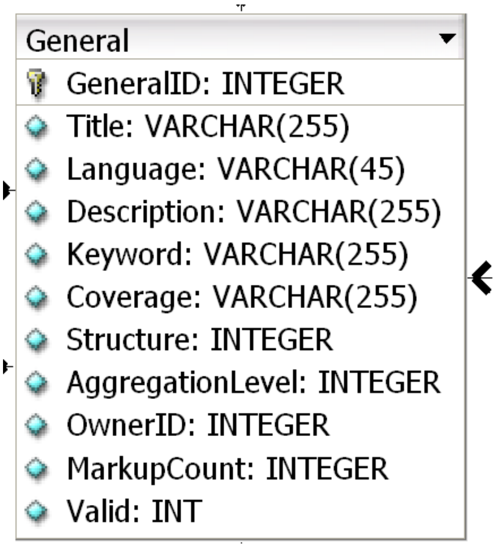
1. To the wanted directory on the web server, extract and move all the unpacked files of downloaded PhpODPWorld package.
2. Complete the following steps
3. Create "logs" directory, if you enabled logging in the config file and this directory must be writable by the web server
4. Create "smarty/cache" directory and "smarty/compiled", if you enabled smarty in the config file and this directory must be writable by the web serve
5. Create a database and a user either with password or without.
6. The database table defined in "tools/db.sql" is created.
7. Wanted categories and references, database settings are reflected by editing "config.inc.php" and "tools/config.pl".
8. Downloaded either the complete content or structure RDF.
9. Use the Perl script "tools/extract.pl" to extract your categories from the RDFs. (Perl module DMOZ-ParseRDF-0.14 should be installed now)
10. Using Perl scripts insert your categories (from the RDFs) into the database.
11. Do the following for initial run to update the count (of sites) for each category
    a) Turn on maintenance mode in "config.inc.php".
    b) Turn off maintenance mode in "config.inc.php".

5.5 Data Structure

After populating the database the data has been moved into three main tables known as:

- Classification
- Identifier
- General

5.5.1 General



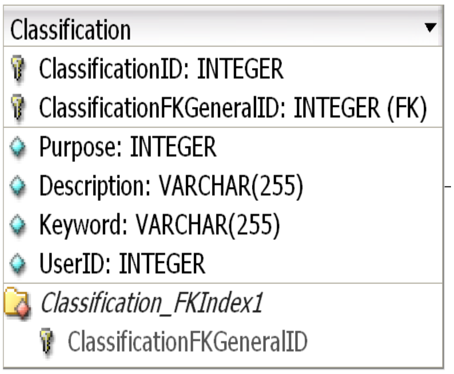
General
GeneralID: INTEGER
Title: VARCHAR(255)
Language: VARCHAR(45)
Description: VARCHAR(255)
Keyword: VARCHAR(255)
Coverage: VARCHAR(255)
Structure: INTEGER
AggregationLevel: INTEGER
OwnerID: INTEGER
MarkupCount: INTEGER
Valid: INT

Figure 1: General Table Structure

The most important fields of this table are General ID, Title, Coverage and Aggregation Level. General ID describes or gives the General ID for each and every node in the tree or graph and at the end of the taxonomy it gives the ID for the leaf. Title Gives the Name of the node of the graph and at the last level it gives the name of the leaf. Coverage is the most Important field of the table and it shows the taxonomy through which this node or particular Level is achieved. Aggregation Level shows the number of Childs a particular node (Root in this level) has.

Owner ID gives the ID of the person who has added this JAN. Also the no of times that particular JAN has been marked up is given by Markup Count. Valid field shows the status of the JAN whether it is alive or a dead link. The keyword field stores the keyword given by the USER for his own convenience.

5.5.2 Classification

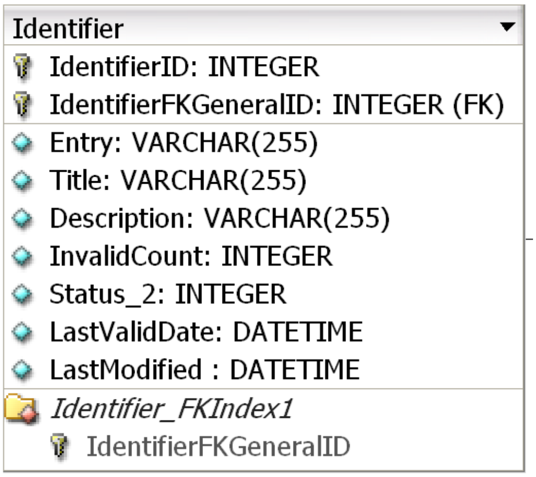


Classification
ClassificationID: INTEGER
ClassificationFKGeneralID: INTEGER (FK)
Purpose: INTEGER
Description: VARCHAR(255)
Keyword: VARCHAR(255)
UserID: INTEGER
Classification_FKIndex1
ClassificationFKGeneralID

Figure 2: Classification Table Structure

The Classification ID gives the ID into which that Particular NODE has been classified. GeneralID refers to the general Id of this node or leaf. User Id gives the ID of the user who classified the JAN. KEYWORD stores the keywords given by that particular User.

5.5.3 Identifier



Identifier
IdentifierID: INTEGER
IdentifierFKGeneralID: INTEGER (FK)
Entry: VARCHAR(255)
Title: VARCHAR(255)
Description: VARCHAR(255)
InvalidCount: INTEGER
Status_2: INTEGER
LastValidDate: DATETIME
LastModified : DATETIME
Identifier_FKIndex1
IdentifierFKGeneralID

Figure 3: Identifier Table Structure

This is the table where the Value of the JAN or the URL will be stored. Identifier ID is a special ID here which is given for the END leaf or the URL of the whole taxonomy, This ID is different from the GeneralID and will not have any values for the nodes. The foreign Key GeneralID gives the general id for this Node (Leaf in this case, since the last level)

Note: Identifier ID gives an ID for the end URL's or Jan's (Leafs) only in the Taxonomy (Tree) and GeneralID gives the ID for both Leaf's and Nodes.

Entry gives the value of the URL or JAN. Title here gives the Title for the JAN and description gives about the information of the JAN. Invalid Count gives the no of times this JAN has been Invalid or Dead. Status gives the present condition of the JAN if it is ALIVE or DEAD. LastModifiedField gives the last time this JAN has been modified.

6 Visualization Data

6.1 Introduction

Wiki's Definition of Visualization is any technique for creating images, diagrams, or animations to communicate a message. Visualization through visual imagery has been an effective way to communicate both abstract and concrete ideas since the dawn of man. Visualization [6] of Vijjana is the most important part of all the agents that we have in Vijjana model. It is the users interface to interact with the knowledge base to find the semantics [1] and also to obtain relevant information in a particular field by means of user friendly navigation.

Now for the Visualization of Vijjana we have to first connect to the database of Vijjana and generate graph XML files, which are given as an input to the Prefuse and Hypergraph tools and thereby we will get corresponding User Controllable visuals. Before dealing with generation of XML, I will introduce XML, XML schemas and the XML schema's used for different visual views.

The raw data from the database is to be given in an abstract form to the Prefuse & Hypergraph toolkits, so an XML having the structured information about the Vijjana semantic net or taxonomy is required. There are several XML schemas that are related to the corresponding visual views. For example for Hypergraph we use an XML that has the schema of GraphML. So based on the type of Visualization that we generate, first we have to study the XML schema and then we have to transform the raw data to that form of XML.

The XML file is easy to transfer because of its small size but, hard to interpret because of its flexibility. Actually, what we are going to implement is to create the XML data schema based on the database table we have, and create an automatically intelligence system to generate the XML file based on the search result. The result is large XML file which can be interpret based on several of data schema, and it will return different small xml files based on them. This XML file can be interpreted as the raw data file for the visualization and other purposes.

Now for the visualization of Vijjana, in the first stage, we have to define the logic in retrieving data to follow the rules of XML schema's of corresponding visualizations.

6.2 MySQL Query Browser

The MySQL Query Browser is a graphical shell where you can execute queries and develop SQL scripts, with several features to help you improve your productivity. The MySQL Query Browser interface tries to mimic the interface of a web browser. [9]

Here we are using MySQL query browser to show the way we query the data from database. Queries are executed on this local database to extract data in the desired manner for the generation of the XML's. These XML's are used as an input for the Prefuse to generate the Visualization techniques. Since the MySQL query browser interface almost acts as the original browser, our task is to fill the local database with the data present in the original database over the server. After the data is ready in this database, several operations can be done to manage it and it also can be accessed in different programming languages for generating the XML's.

The whole process in creating a Database and populating it with data is described below:

1. Install MySQL Server 5.0
2. Open MySQL query browser
3. On the top of the page you can see a small window (Query Execution Space), execute a query for creating a new database.

Create Database Vijjanadatabase;

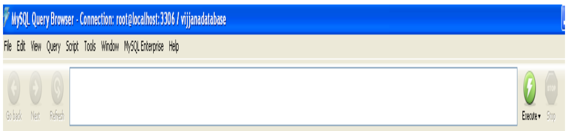

MySQL Query Browser - Connection: root@localhost:3306 / vijjanadatabase
File Edit View Query Script Tools Window MySQL Enterprise Help
Go back
Next
Refresh
Execute
Stop

Figure 4: Query Execution Window of MySQL Query Browser[9]

4. The created database will be shown at the schemata window. Select the database by double clicking it; this means that all the future operations like queries etc will be done on that database since you have selected it.

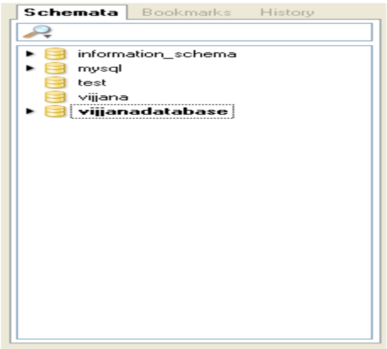


Figure 5: Database Is Selected

5. Now create a new Table by name general in the database (Vijjanadatabase) by executing the whole SQL script for the Table General at the query window.
6. The created table will not be shown immediately, so right click on the database Vijjanadatabase and then click refresh. This can also be done by double clicking on the Vijjanadatabase.
7. Likewise all other corresponding Tables which are present in vijjanadatabase are created using their corresponding scripts.

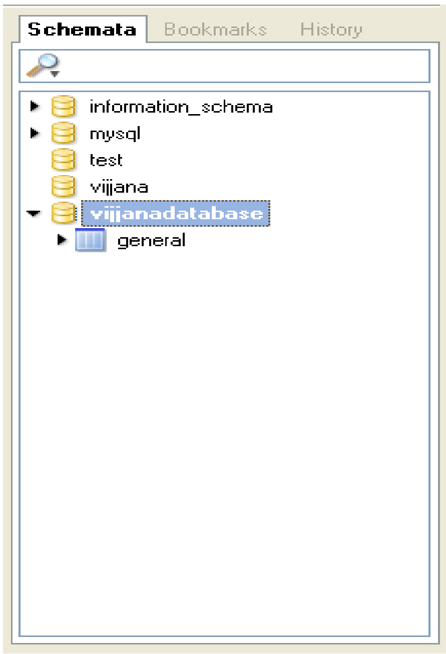
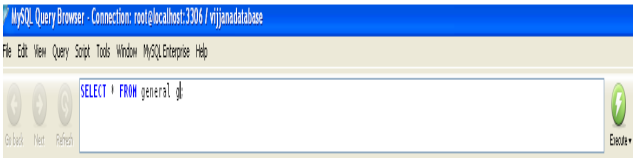


Figure 6: Showing the Contents present in Vijjanadatabase

8. Now select the general table by double clicking it. This also generates the query automatically in the query browser window. At the start the General table will not show any values because it not having any data with it. This is similar with all other tables so created.



```
SELECT * FROM general g;
```

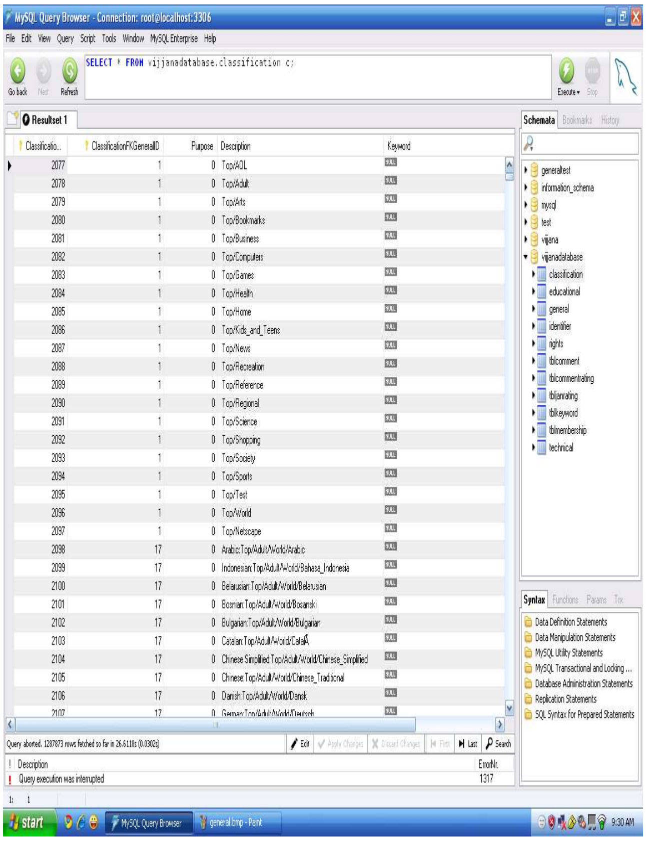
MySQL Query Browser - Connection: root@localhost:3306 / vijjanadatabase
File Edit View Query Script Tools Window MySQL Enterprise Help
Go back
Next
Refresh
Execute

Figure 7: Query Automatically Generated When General Table Is Double-clicked

9. Now our task is to populate this local database with the data present in the server. To do this we have to first download all the data from the server from eksarva.csee.wvu.edu/vijjanadata. .
10. The above site would let you download over 1GB of data file with name “vijjanadata”, Change the name of this file to “vijjanadata.sql” so that it will become an executable SQL file format.

11. To import this data into the database on your machine (vijjanadatabase), we use cmd prompt. *”mysqldump -u root -p vijjanadatabase < vijjanadata.sql.”*
12. The above command will dump the data vijjanadata.sql in the vijjanadatabse created on your local machine. A dialog box will appear showing Dump is complete. This means that the data has been imported into the local database and is available for usage.

After the data has been imported into the local database (Vijjanadatabase), it can be used for further execution of queries and for generating the XML’s required for the visualization[6].



```
SELECT * FROM vijjanadatabase.classification c;
```

Classificatio...	ClassificationFKGeneralID	Purpose	Description	Keyword
2077	1	0	Top/AOL	NULL
2078	1	0	Top/Adult	NULL
2079	1	0	Top/Arts	NULL
2080	1	0	Top/Bookmarks	NULL
2081	1	0	Top/Business	NULL
2082	1	0	Top/Computers	NULL
2083	1	0	Top/Games	NULL
2084	1	0	Top/Health	NULL
2085	1	0	Top/Home	NULL
2086	1	0	Top/Kids_and_Teens	NULL
2087	1	0	Top/News	NULL
2088	1	0	Top/Recreation	NULL
2089	1	0	Top/Reference	NULL
2090	1	0	Top/Regional	NULL
2091	1	0	Top/Science	NULL
2092	1	0	Top/Shopping	NULL
2093	1	0	Top/Society	NULL
2094	1	0	Top/Sports	NULL
2095	1	0	Top/Test	NULL
2096	1	0	Top/World	NULL
2097	1	0	Top/Netscape	NULL
2098	17	0	Arabic:Top/Adult/World/Arabic	NULL
2099	17	0	Indonesian:Top/Adult/World/Bahasa_Indonesia	NULL
2100	17	0	Belarusian:Top/Adult/World/Belarusian	NULL
2101	17	0	Bosnian:Top/Adult/World/Bosanski	NULL
2102	17	0	Bulgarian:Top/Adult/World/Bulgarian	NULL
2103	17	0	Catalan:Top/Adult/World/CatalÃ	NULL
2104	17	0	Chinese Simplified:Top/Adult/World/Chinese_Simplified	NULL
2105	17	0	Chinese:Top/Adult/World/Chinese_Traditional	NULL
2106	17	0	Danish:Top/Adult/World/Dansk	NULL
2107	17	0	German:Top/Adult/World/Deutsch	NULL

Query aborted. 1287873 rows fetched so far in 26.6118s (0.0302s)

!	Description	ErrorNr.
!	Query execution was interrupted	1317

1: 1

Figure 8: Local Database

7 Conclusions

Thus according to the standards and format of IEEE 1484.12.1, a database schema has been designed. The user database has been integrated into the schema and a complete Jan Structure has been developed for Vijjana. The corresponding script for this schema has been generated. The scripts were executed over the server using the PHP-MyAdmin tool and a database (vijjanadatabase) has been created. Data is imported into this Database by using “PhpODPWorld” and PEARL scripting language. A local Database has been created by using MySQL query browser interface and is populated with a few amounts of data.

Now programming and query execution can be done on this local database rather than working on the large some of data over the server.

8 References

[1] Ivan Herman, (W3C) Semantic Web Activity Lead. W3C Semantic Web. W3C. [Online] 1994-2008. http://www.w3.org/2001/sw/.

[2] Vijjana: A Pragmatic Model for Collaborative, Self-organizing, Domain Centric. Ramana Reddy 2008.

[3] Corporation, Netscape Communications. About the Open Directory Project. dmoz. [Online] Netscape Communications Corporation, 1998-2005. http://www.dmoz.org/about.html.

[4] W3C. Resource Description Framework (RDF). W3C. [Online] http://www.w3.org/RDF/.

[5] Hansfn, Srainwater, ODP Editors. phpODPWorld. sourceforge.NET. [Online] http://phpodpworld.sourceforge.net/.

[6] Visualization module of Vijjana, a Pragmatic Model for Collaborative, Self-organizing, Domain Centric Knowledge Networks. Sasanka Babu Gottipati.

[7] IEEE 1484.12.1-2002, 15 July 2002 Draft Standard for Learning Object Metadata. [PDF] http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf

[8] Wolf, Boris. Storing RDF Metadata in a Relational Database V1.2. 2001, [Online] http://www.kbs.uni-hannover.de/Arbeiten/Studienarbeiten/01/Wolf/olr_documentation.pdf.

[9] MySQL Query Browser cross-platform GUI client program [Online] http://www.mysql.com/products/tools/query-browser/

Knowledge Representation and Annotation for Semantic Web Library

Hadeel S. AL-Obaidy[1] **and Amani Al Heela**[2]

[1]Computer Engineering Department Ahlia University, Manama, Kingdom of Bahrain

[2] Information Technology Department Arabian Gulf University, Manama, Kingdom of Bahrain

Abstract - *The information contained in the World Wide Web or the web content is increasing every day. In this paper, we describe the semantic annotation process for university's library semantic web application. The step in developing the semantic web application that adds the effectiveness and reality to it is the semantic annotation for the documents published and distributed throughout the Web. The semantic annotation in this paper concerns about the research papers of the university's faculty. Semantic annotation is nothing but tagging the instances data of ontology already created with classes then map in to the related ontology classes. In this paper, two tools are going to be used for the annotation: OntoMat and OntoStudio.*

Keywords: Annotation, Semantic Web, Ontology, Knowledge Representation.

1 Introduction

The World Wide Web (WWW) is a service that needs Internet to work. It allows users to read and write information that is displayed in computers connected to the internet. What is used in the proposed system is the second and third generation of the WWW. The second generation of World Wide Web (Web 2.0) concentrates mainly on collaboration, interaction and social networking. Examples of Web 2.0 are blogs, RSS, wikis, web applications.

Tim Berners-Lee [19] has described the Semantic Web as a component of Web 3.0. The Semantic web allows for accessing information based on its meaning. A new Semantic Library model (SWLib) is considered an important need for any university. The new library website with a new design, updated information and Semantic Web model is going to change the way visitors experience the website. Currently, only few of the library websites are integrated with Semantic Web. Integrating Semantic Web with e-library is a major shift for any university's library website and allows it to be one of the leading library websites.

SWLib enables Arabian Gulf University (AGU) faculty to have their research papers published in one centralized place and makes it easy for them to find the research papers that belong to their colleges. This research papers is added to the SWLib using a Semantic Web model and through annotation process this papers are stored in an RDF store that in turn compose a knowledge base. The Semantic Web is, as mentioned above, a component of Web 3.0 which is a major intelligent addition to the Web.

This paper discusses the step in developing the Semantic Web application that adds the effectiveness to it which is the semantic annotation. This step applied for the documents published and distributed throughout the Web. Semantic annotation in this paper concerns about the research papers of the university's faculty.

2 Literature Review

Nicola Guarino from National Research Council and Pierdaniele Giaretta [3] from the University of Padova in their paper "Ontologies and Knowledge Bases" have clearly defined the Ontology from technological and philosophical views. They made careful analysis of Gruber's definition of ontology as a specification of a conceptualization.

Design and Implementation of Semantic Community Web Portal is the title of the paper written by Ching-Long Yeh and Chang-Gang Chen from Tatung University in Taiwan [4]. They built a semantic web portal using the RDF technology used to represent the contents of the portal. They discuss the semantic web technologies and the steps they follow to build the semantic web portal.

Another paper discusses the Extensive Markup Language (XML) and Resource Descriptive Framework (RDF) standards in depth. The paper title is "The Semantic Web - on the respective Roles of XML and RDF" and it was written by Stefan Decher, Frank van Harmelen, Jeen Broekstra, Michael Erdmann, Dieter Fensel, Ian Horrocks, Michel Klein and Sergery Melnik [15]. These standards are used as part of this dissertation.

3 Web 3.0 and Semantic Web

The reporter John Markoff says in an article in The New York Times that the idea of adding meaning which is used in Web 3.0 or Semantic Web is just now emerging [6].

As Tim O'Reilly has defined Web 2.0, Nova Spivack has also defined Web 3.0 as connective intelligence that is applied through embedding intelligence in the connected data, concepts, applications and people. He rejected the view of considering Web 3.0 as Semantic Web; he includes Semantic Web is part of Web 3.0 [20].

Semantic Web has also been defined by Tim Berners-Lee, the director of World Wide Web Consortium W3C, as “a web of data that can be processed directly and indirectly by machines” and as "the extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation" [7].

The Semantic Web can be easily defined as making the machines understand the meaning of the content by applying a collection of technologies. These technologies are Resource Description Framework (RDF), RDF Schema and Web Ontology Language (OWL).

Each of these technologies will be discussed in detail in the next sections of this chapter. The building blocks of the semantic web presented by Berners Lee at the Conference XML-2000 are illustrated in the next Figure 1.

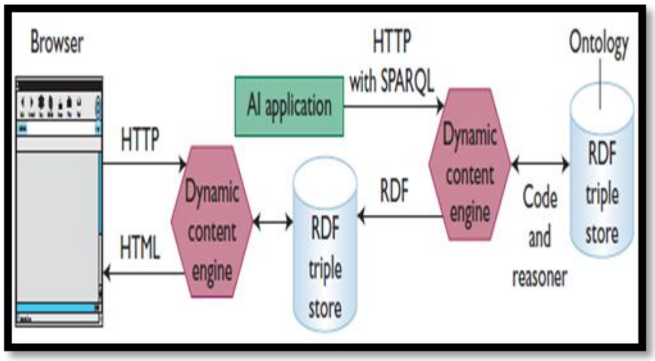


Figure 1: Semantic Web Building Blocks, Source: [8].

There are four principles which must be taken into account during developing a semantic web application:

1. All the data and entries that share the same information should be identified by Uniform Resource Identifier (URI) references.
2. The data must be provided in RDF format.
3. The URI in Hypertext Transfer Protocol (HTTP) should be linked to the RDF that belongs to it.
4. The data should be interlinked with each other.

Architecture of a sample of semantic web application (see Figure 2) has the following components:

- RDF triple store.
- Dynamic content engine.
- Artificial Intelligence (AI) application.
- Browser.

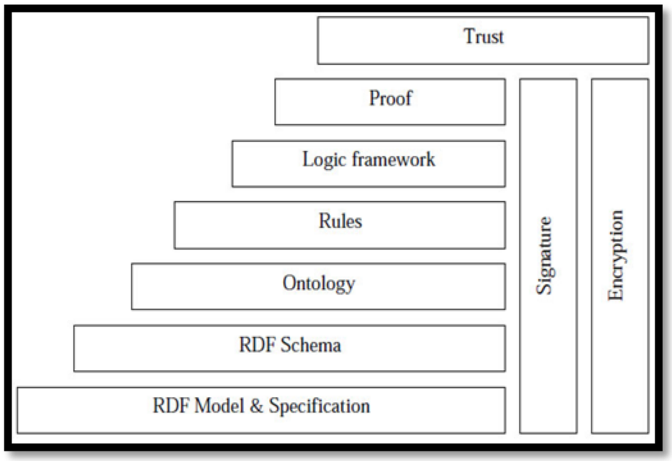


Figure 2: Architecture of a sample of Semantic Web application, Source: [17].

4 Annotation Process & RDF

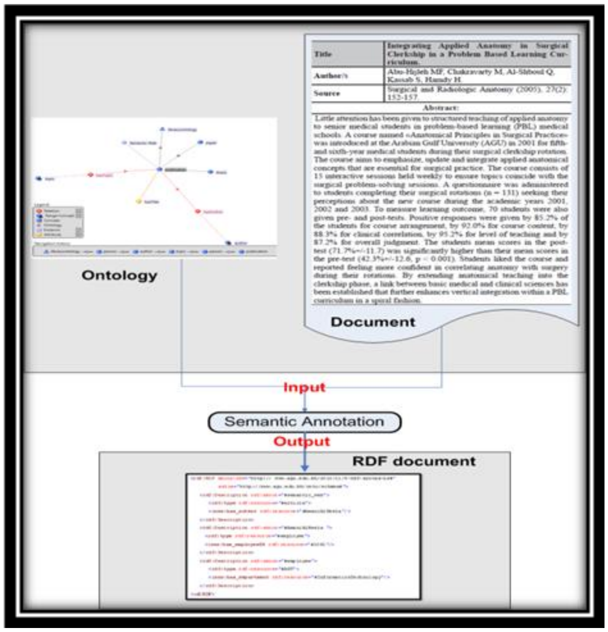
The information contained in the World Wide Web or the web content is increasing every day. The latest survey conducted by the Internet System Consortium was on October 2010 and found that the number of hosts advertised in Domain Name Server (DNS) was 777,994,517. These hosts are the one whose responsible for serving the Web pages, one host can serve up to millions of Web pages and now imagine how many web pages with its information is available in the World Wide Web! It is a very huge number that makes it very difficult for a person to search and find the needed information from it. For that reason, Semantic Web has existed and presented to solve the problem of finding the wanted information in the World Wide Web. Its main idea is to search based on the semantics of information that use a technology to make the machines understand the information and that is obviously leading to easing and fastening the search process and overcome the problem that the World Wide Web was only provide the information to people who are the only one that understand those information.

As any process, annotation process has input and output. The input is the documents and ontology, and the output is a Resource Description Framework (RDF) document.

The first input is the documents and in this paper, the documents used as input in the annotation process are the research papers that wrote by the university's faculty.

The second input is the ontology. The ontology is the brain component of the semantic web application. It's providing the application or the machine the understanding capability. Thomas Gruber defines the ontology as "explicit specification of conceptualization" [10] while it also can be defined as the relationships that connect concepts, nodes or entities to each other.

The output is the RDF document and it is explained in details in the next paragraphs.



Title	Integrating Applied Anatomy in Surgical Clerkship in a Problem Based Learning Curriculum.
Author/s	Abu-Hijleh MF, Chakravarty M, Al-Shboul Q, Kassab S, Hamdy H.
Source	Surgical and Radiologic Anatomy (2005). 27(2): 152-157.

Abstract:

Little attention has been given to structured teaching of applied anatomy to senior medical students in problem-based learning (PBL) medical schools. A course named «Anatomical Principles in Surgical Practice» was introduced at the Arabian Gulf University (AGU) in 2001 for fifth- and sixth-year medical students during their surgical clerkship rotation. The course aims to emphasize, update and integrate applied anatomical concepts that are essential for surgical practice. The course consists of 15 interactive sessions held weekly to ensure topics coincide with the surgical problem-solving sessions. A questionnaire was administered to students completing their surgical rotations (n = 131) seeking their perceptions about the new course during the academic years 2001, 2002 and 2003. To measure learning outcome, 70 students were also given pre- and post-tests. Positive responses were given by 85.2% of the students for course arrangement, by 92.0% for course content, by 88.3% for clinical correlation, by 95.2% for level of teaching and by 87.2% for overall judgment. The students mean scores in the post-test (71.7%+/-11.7) was significantly higher than their mean scores in the pre-test (42.3%+/-12.6, p < 0.001). Students liked the course and reported feeling more confident in correlating anatomy with surgery during their rotations. By extending anatomical teaching into the clerkship phase, a link between basic medical and clinical sciences has been established that further enhances vertical integration within a PBL curriculum in a spiral fashion.

Figure 3: Diagram illustrating the annotation process

RDF is a W3C standard model used for the purpose of data interchange in the Web. It's providing the semantic web application with interoperability feature because RDF is readily for any program and facilitates data merging, no matter what schema used. Storing knowledge using this standard done by decomposing it into (3 tuples) triples. One triple is composed of object, attribute and value. In another way it composed of a resource (object), named property (attribute) and value for the property (value).

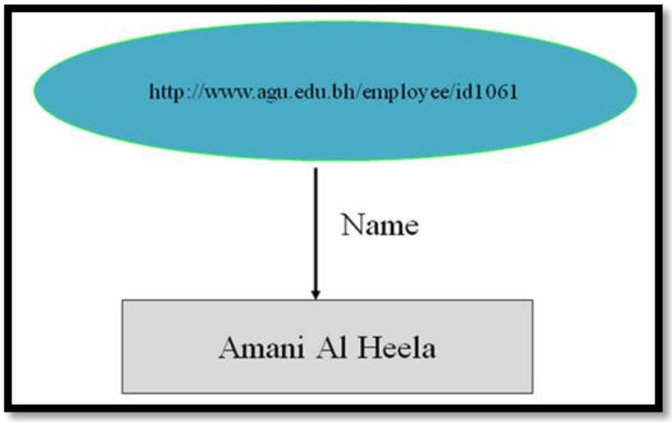
RDF allows structured and semi structured data to be exchanged between applications by using URI to identify each relationship between data in a triple.

The triples can be expressed in three ways: tables, xml files and graphs. The easiest view is the graph view. Let's take this example:

Name ('http://www.agu.edu.bh/employee/id1061", "Amani Al Heela"). This example has three views table (see Table 1), xml and graph (see Figure 4).

Table 1: Table View of RDF example

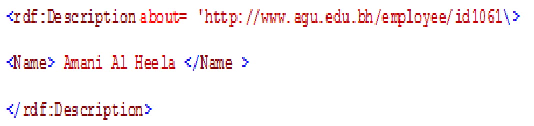
Object	Attribute	Value
http://www.agu.edu.bh/employee/id1061	Name	Amani Al Heela



The xml expression of the example is:

Figure 4: Graph View of RDF example

The graph view of the example is:



```
<rdf:Description about= 'http://www.agu.edu.bh/employee/id1061\>
<Name> Amani Al Heela </Name >
</rdf:Description>
```

Simple Protocol and RDF Query Language (SPARQL) is a query language just like Standard Query Language (SQL) which is used to perform manipulations such as insert, update and delete the native graph stored in RDF stores. The results of the executed query using SPARQL are a set of RDF graphs, XML, JSON and HTML.

The query of SPARQL is composed o the following:

- Declaring Prefix using URIs
- Defining RDF dataset and specifying the graph to be queried.
- Identify which information should be returned as a result of the query.
- Decide what the information to query for.
- The arranging query like ordering the resulted data.

This is an example of SELECT query in SPARQL:

To execute SPARQL query via HTTP, the SPARQL endpoint must be used for querying from RDF stores that can be accessed through Web.

5 THE PROPOSED SYSTEM DESIGN

The proposed system implemented by completing the following the steps (Figure 5):

- The first step is to prepare a good design for the AGU library website that satisfies the standards.
- The web developer then converts this design to a developed website. The development language used in the proposed system is the ASP.NET language and the program used is the visual studio 2008.
- The proposed website integrated with web 2.0 applications.
- The Web 3.0 integrated to the proposed system through developing a semantic web service. The three major processes are:
  - o Engineer ontology: the OntoStudio software is used for engineering the ontology of the proposed system. Ontology is the core of the proposed semantic web service that is defines the data schema for which the data will be entered.
  - o Annotation: in the annotation step, OntoStudio and OntoMat are used to enrich the ontology with data. The AGU faculty papers are collected to be used in this step. This step is what this paper discusses.
  - o Indexing: the RDF document that produced from the previous steps is stored in the SQL database to build RDF store.

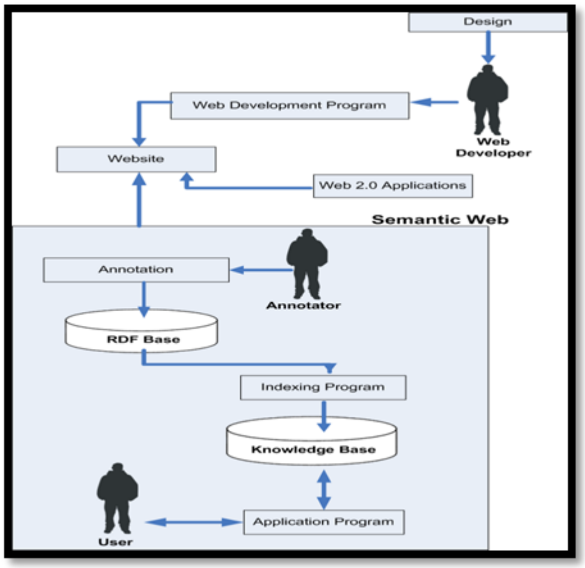


Figure 5: The conceptual architecture of the proposed website of AGU's library

The steps needed in annotation process for the proposed system are described in details in the following algorithms besides the sequence diagram in Figure 6.

Algorithm for Annotation Process (entering paper's information) using OntoStudio
Input: paper information
Output: Knowledge Base – RDF document

```
Begin
        Step 1: Collect the paper information: paper
        title, abstract and author from AGU faculty.
        Step 2: Open ontology that created previously
        to start enriching it with papers information.
        The annotation is done by creating instances
        for the concepts.
  Begin

    Step 4: Open the OntoStudio software.
Loop
    Step 5: Click the class that wanted to enter
    information to it.
    Step 6: Right click the class and create new
    instance.
    Step 7: The related attributes and relations values
    are entered for each instance.
End Loop
    Step 8: Save the ontology with the entered
    information as an RDF document.
     Step 9: Close OntoStudio software.
  End
```

Algorithm for Annotation Process (entering paper's information) using OntoMat
Input: paper information
Output: Knowledge Base – RDF document

```
Begin
    Step 1: Open OntoMat software.
    Step 2: Open ontology file.
    Step 3: Open the online paper file in the specified
    area.
    Step 4: Start annotating by creating instances for
    each class and drag the related information in each
    one.
    Step 5: Save ontology file with entered data.
    Step 6: Save the ontology file as RDF.
End

    Step 7: Save the RDF documents into RDF store in
    SQL database.

Begin

    Step 8: Open SQL express 2005.
    Step 9: Create a database for storing the RDF
    documents.
    Step 10: Execute dotNetRDF store manager.
    Step 11: Click file then New SQL Store Manger.
    Step 12: Enter the connection details: database
    name, username and password.
    Step 13: Import the RDF document.
    Step 14: Click create data store.
End
```

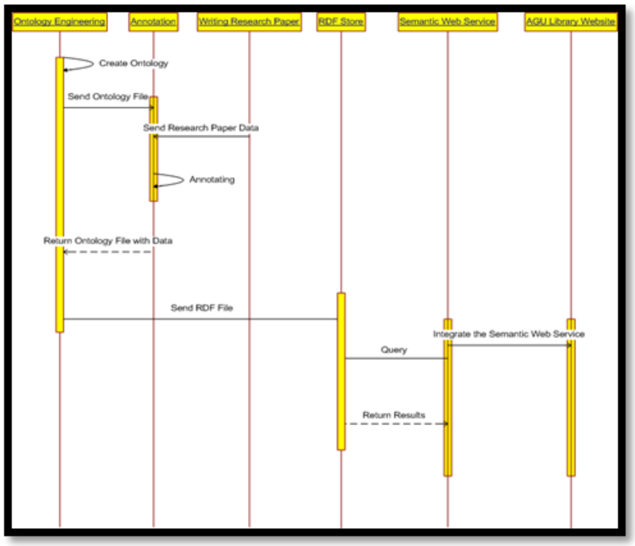


Figure 6: The sequence diagram of the proposed system

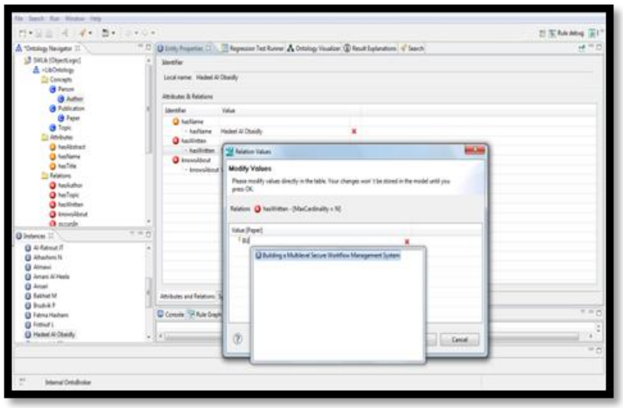
The processes used to develop the Semantic Web Service for e-library are summarized in figure 6. The annotation process this paper most concern about is shown in figure 6 after engineering ontology process; the annotating process looping to enter the research papers information and after completing it, the ontology becomes rich of data and so the knowledge base is created.

5.1 Annotation using OntoStudio

"The OntoStudio is an engineering environment for ontologies and for the development of semantic applications, with particular emphasis on rule-based modeling. It is the successor of OntoEdit which was distributed worldwide more than 5000 times. OntoStudio was originally developed for F-Logic but now also includes some support for OWL, RDF, and OXML. It also includes functions such as the OntoStudio Evaluator. The Evaluator is used for the implementation of rules during modeling; this procedure has been recently patented" [1].

The data of the documents (research papers) is mapped to the ontology that engineered previously in the OntoStudio. The annotation process is the process of creating new instances and entering data to it.

To feed the ontology with knowledge, the annotation step takes this role and enriches the ontology with knowledge. The following figures describe in details how the annotation process done. There are two ways of implementing the annotation step. The first one is by using the OntoStudio software and the second is by using OntoMat. Using OntoStudio, the following figures shows the steps for annotation process.



Identifier

Local name: Hadeel Al Obaidy

Attributes & Relations

Identifier	Value
hasName	
hasName	Hadeel Al Obaidy
hasWritten	
knowsAbout	

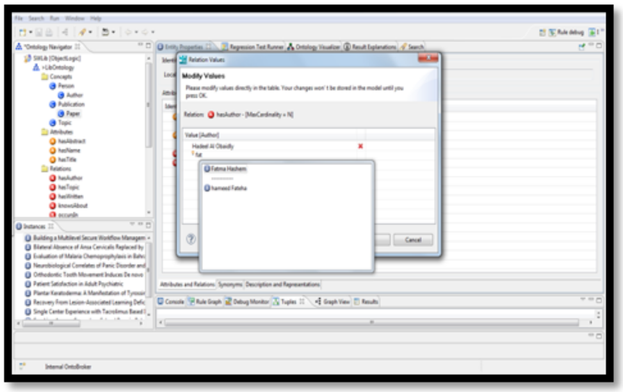
Please modify values directly in the table. Your changes won't be stored in the model until you press OK.

Relation: hasWritten - [MaxCardinality = N]

Value [Paper]

Building a Multilevel Secure Workflow Management System

Figure 7: Annotation step – create new instance



Please modify values directly in the table. Your changes won't be stored in the model until you press OK.

Relation: hasAuthor - [MaxCardinality = N]

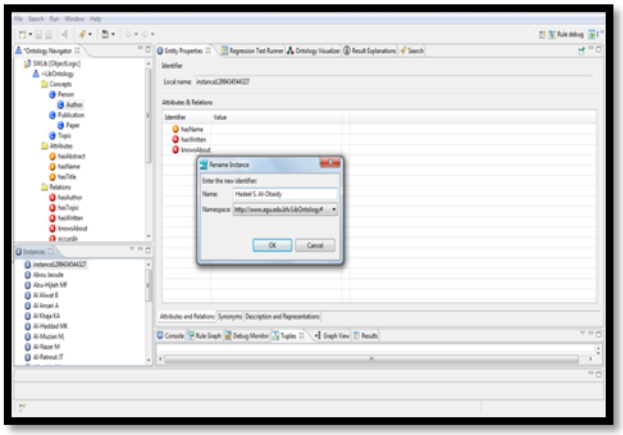
Value [Author]

Hadeel Al Obaidy

Fatma Hashem

hameed Fataha

Figure 8: Annotation step – enter data for the new instance attribute



Enter the new identifier

Name: Hadeel S. Al-Obaidy

Namespace: http://www.agu.edu.bh/LibOntology#

Figure 9: Annotation step – enter value for the new instance relation

5.2 Annotation using OntoMat

The annotation can be done by using the interactive webpage annotation tool OntoMat. It is a user-friendly and easy tool that can be used by any person. Once the OntoMat is open, the next step is to import the ontology into OntoMat so it becomes possible to maintain the ontology and create instances, attributes and relationships. OntoMat composes of two browsers, ontology browser for viewing the ontology and instances and HTML browser that display the document that is wanted to be annotated.

The annotation process in OntoMat is just about drag and drop. Drag the part of the document and drop it to the instance of the relevant ontology's class.

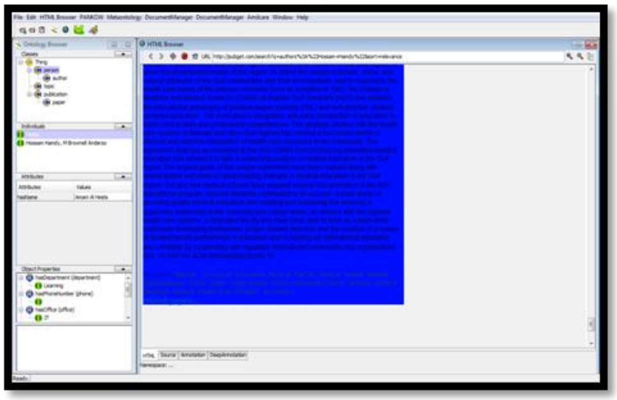


Figure 10: Annotation using OntoMat

6 Conclusions

The library is a very essential unit in any university; it is the unit that provides the knowledge to help the members of the university. Nowadays in the information era, the need for a website that reflects the university's library and provides access to the knowledge it holds is increasingly becoming more important. AGU is like any another global university and needs an electronic gateway to the library, which is a website. It is not an exaggeration to say that planning for developing a library website should be given the same planning and care as the library itself.

This paper supports using annotation process and RDF Semantic Web techniques for adding more value and functionality features to e-library. This can be achieved by start creating an archive of knowledge that allows the visitors to access easily. These features definitely are definitely increasing the number of visitors to the e-library and increase user satisfaction. In addition, it affects the e-library to get a higher ranking among universities which allows it to compete successfully with other the leading library websites.

This paper discussed the annotation process that is one the processes used to develop a Semantic Web Service for e-library and how important is to enrich ontology with knowledge. The input for annotation process is the documents and ontology, and the output is an RDF document that is then ready to be used by any Semantic Web Service.

1. FUTURE WORK

Future researches are needed to include: Developing and testing phase for the Semantic Web service for e-library. Enhancing the proposed system and adding more functionality to e-library and extending the use of Semantic Web to other services in e-library, Maximize the benefits of Semantic Web by reusing it for presenting other type of knowledge, and expand the use of library Semantic Web application to the mobile technology and develop a web application that working in WAP.

7 References

[1] OntoStudio. (2011). Retrieved March 5, 2001, from http://semanticweb.org/wiki/OntoStudio.

[2] ALEXANDER, B. 2006. Web 2.0: A new wave of innovation for teaching and learning. EDUCAUSE Review. Vol. 41, No. 2, March/April 2006, pp. 32–44. EDUCAUSE: Boulder, USA. Updated version available online at: http://www.educause.edu/apps/er/erm06/erm0621.asp [last accessed 14/01/2011]

[3] Nicol Guarino, Pierdaniele Giaretta. Ontologies and Knowledge Bases: Towards a Terminological Clarification. In *Towards Very Large Knowledge Bases*, N.J.L. Mars, Ed. Amsterdam: IOS Press; 1995.

[4] Yeh, Ching-Long and Chen, Chang-Gang. Design and Implementation of Semantic Community Web Portal. Available at http://www.cse.ttu.edu.tw/chingyeh/papers/DATFPortal.pdf. [Last accessed 29/02/2011]

[5] O'REILLY, T. 2005a. What is Web 2.0: Design Patterns and Business Models for the next generation of software. O'Reilly website, 30th September 2005. O'Reilly Media Inc. Available online at:http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html [last accessed 17/01/2011].

[6] J. Markoff, "Entrepeneurs See a Web Guided by Commonsense," The New York Times, Business, 12 Nov. 2006.

[7] Berners-Lee, Tim; James Hendler and Ora Lassila (May 17, 2001). "The Semantic Web". Scientific American Magazine. http://www.sciam.com/article.cfm?id=the-semantic-web&print=true. [Last accessed 29/02/2011].

[8] Berners-Lee, T., J. Hendler and O. Lassila: Semantic Web Scientific American, May 2000.

[9] Ora Lassila & James Hendler: "Embracing 'Web 3.0'", IEEE Internet Computing 11(3):90-93, May/June 2007

[10] A guide to Future of XML, Web Services and Knowledge Management by Michael C.Daconta, Leo J. Obrst, Kevin T.Smith, 2003

[11] Holger Lausen, Ying Ding, Michael Stollberg, Dieter Fensel, Ruben Lara Hernandez and Sung-Kook Han, (2005), "Semantic web portals:state of the art survey" Jornal of knowledge Management, Vol. 9, No. 5 (2005): 40-49.

[12] G Kück, (2004), "Tim Berners-Lee's Semantic Web" South African Journal of Information Management, Vol.6 (1) March (2004). http://www.sajim.co.za/index.php/SAJIM/article/download/297/288. [Last accessed 06/03/2011].

[13] Nenad Stojanovic, Alexander Maedche, Steffen Staab, Rudi Studer, York Sure. SEAL: a framework for developing SEmantic PortALs, In K-CAP, pp. 155-162, 2001

[14] Siddharth Gupta1 and Narina Thakur, (2010), "Semantic Query Optimisation with Ontology Simulation"

[15] Stefan Decker, Frank van Harmelen, Jeen Broekstra, Michael Erdmann, Dieter Fensel, Ian Horrocks, Michel Klein, Sergey Melnik, (2000), "The Semantic Web - on the respective Roles of XML and RDF"

[16] K Srinivas, S I Ahson, T A V Murthy, (2006), "Builiding a Semantic Web for Academic Networks: a conceptual architecture"

[17] Ora Lassila and James Hendler, (2007), "Embracing Web3.0"

[18] Leonardo Magela Cunha, (2007), "A Semantic Web Application Framework"

[19] LÉGER, A. et al. D2.2 Successful Scenarios for Ontology-based Applications V1.0 ,2002/05/31. ,2002p. 100. Available at:http://ontoweb.org/Members/huro/MyPublications/OntoWeb%20Deliverable%202.2/view. [Last accessed 27/02/2011].

[20] Definition of web 3.0. (2011). Retrieved March 1, 2011, from http://www.webopedia.com/TERM/W/Web_3_point_0.html.

SESSION

ALGORITHMS

Chair(s)

TBA

Fast Ordered Tree Matching for XML Query Evaluation

Yangjun Chen, Yibin Chen
Dept. Applied Computer Science, University of Winnipeg
Winnipeg, Manitoba, Canada R3B 2E9
y.chen@uwinnipeg.ca

Abstract– *An XML tree pattern query, represented as a labeled tree, is essentially a complex selection predicate on both structure and content of an XML. Tree pattern matching has been identified as a core operation in querying XML data. We distinguish between two kinds of tree pattern matchings: ordered and unordered tree matching. By the unordered tree matching, only ancestor/descendant and parent/child relationships are considered. By the ordered tree matching, however, the order of siblings has to be taken into account besides ancestor/descendant and parent/child relationships. While different fast algorithms for unordered tree matching are available, no efficient algorithm for ordered tree matching for XML data exists. In this paper, we discuss a new algorithm for processing ordered tree pattern queries, whose time complexity is polynomial.*

Key words: XML documents; tree pattern queries; tree matching; tree encoding; XB-trees

1 Introduction

Xpath [16, 17] is a language for matching paths and, more generally, patterns in tree-structured data and XML documents. These patterns may use either just purely the tree structure of an XML document or data values occurring in the document as well. For example, the XPath expression:

book[title = 'Art of Programming']//author[firstName = 'Donald' and lastName = 'Knuth']

matches *author* elements that (i) have a child subelement *firstName* with content *Knuth*, (ii) have a child subelement *lastName* with content *Donald*, and (iii) are descendants of *book* elements that have a child *title* subelement. It can be represented by a tree structure as shown in Fig. 1.

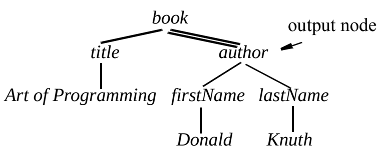


Fig. 1. An Xpath tree

In Fig. 1, there are two kinds of edges: child edges (/-edges for short) for parent-child relationships, and descendant edges (//-edges for short) for ancestor-descendant relationships. A /-edge from node *v* to node *u* is denoted by *v* → *u* in the text, and represented by a single arc; *u* is called a */-child* of *v*. A //-edge is denoted by *v* ⇒ *u* in the text, and represented by a double arc; *u* is called a *//-child* of *v*.

Many different strategies have been proposed to efficiently evaluate such kind of queries [1, 3 - 9, 12, 14, 15]. But most of them take only ancestor/descendant and parent/child relationships into consideration. No attention is paid to the left-to-right order of the nodes.

However, in many applications, such as the natural language processing [2], the video content-based retrieval [13], the scene analysis, as well as some problems in the computational biology (such as RNA structure matching [11]) and the data mining (such as tree mining [18]), the order of the nodes is significant. As an example, consider querying grammatical structures as shown in Fig. 2, which is the parse tree of a natural language sentence.

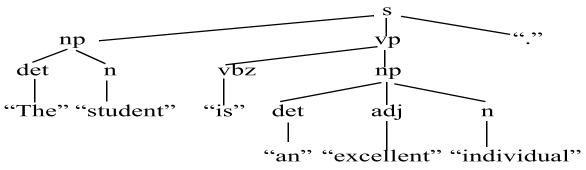


Fig. 2. The parse tree of a sentence

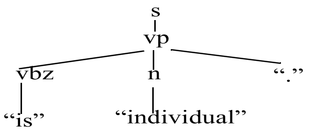


Fig. 3. A query tree which matches a subtree of the parse tree shown in Fig. 2

One might want to locate, say, those sentences that include a verb phrase containing the verb "is" and after it a noun "individual" followed by ".". This is exactly the sentences whose parse tree can be matched to a subtree of the tree shown in Fig. 2. (See Fig. 3 for illustration.) But the left-to-right ordering must be followed.

In this paper, we discuss an efficient algorithm to solve this kind of problems.

The remainder of the paper is structured as follows. In section 2, we give some basic definitions, which are needed for the subsequent discussion. In Section 3, we present the main algorithm. In Section 4, we analyze the computational complxities. Finally, the paper concludes in Section 5.

2 Basic definitions

We concentrate on labeled trees that are ordered, i.e., the order between siblings is significant. Technically, it is convenient to consider a slight generalization of trees, namely forests. A forest is a finite ordered sequence of disjoint finite trees. A tree T consists of a specially designated node *root*(T) called the root of the tree, and a forest $<T_1, ..., T_k>$, where $k \geq 0$. The trees $T_1, ..., T_k$ are the subtrees of the root of T or the immediate subtrees of tree T, and k is the outdegree of the root of T. A tree with the root t and the subtrees $T_1, ..., T_k$ is denoted by $<t; T_1, ..., T_k>$. The roots of the trees $T_1, ..., T_k$ are the children of t and siblings of each other. Also, we call $T_1, ..., T_k$ the sibling trees of each other. In addition, $T_1, ..., T_{i-1}$ are called the left sibling trees of T_i, and T_{i-1} the immediate left sibling tree of T_i. The root is an ancestor of all the nodes in its subtrees, and the nodes in the subtrees are descendants of the root. The set of descendants of a node v is denoted by *desc*(v). A leaf is a node with an empty set of descendants.

Sometimes we treat a tree T as the forest $<T>$. We may also denote the set of nodes in a forest F by $V(F)$. For example, if we speak of functions from a forest G to a forest F, we mean functions mapping the nodes of G onto the nodes of F. The size of a forest F, denoted by $|F|$, is the number of the nodes in F. The restriction of a forest F to a node v with its descendants *desc*(v) is called a subtree of F rooted at v, denoted by $F[v]$.

Let $F = <T_1, ..., T_k>$ be a forest. The preorder of a forest F is the order of the nodes visited during a preorder traversal. A preorder traversal of a forest $<T_1, ..., T_k>$ is as follows. Traverse the trees $T_1, ..., T_k$ in ascending order of the indices in preorder. To traverse a tree in preorder, first visit the root and then traverse the forest of its subtrees in preorder. The postorder is defined similarly, except that in a postorder traversal the root is visited after traversing the forest of its subtrees in postorder. We denote the preorder and postorder numbers of a node v by *pre*(v) and *post*(v), respectively.

Using preorder and postorder numbers, the ancestorship can be easily checked. If there is path from node u to node v, we say, u is an ancestor of v and v is a descendant of u. In this paper, by 'ancestor' ('descendant'), we mean a proper ancestor (descendant), i.e., $u \neq v$.

Lemma 1 Let v and u be nodes in a forest F. Then, v is an ancestor of u if and only if *pre*(v) < *pre*(u) and *post*(u) < *post*(v).

Proof. See Exercise 2.3.2-20 in [10] (page 347). □

Similarly, we check the left-to-right ordering as follows.

Lemma 2 Let v and u be nodes in a forest F. The node v is said to be to the left of u if they are not related by the ancestor-descendant relationship and u follows v when we traverse F in preorder. Then, v is to the left of u if and only if *pre*(v) < *pre*(u) and *post*(v) < *post*(u).

Proof. The proof is trivial. □

In the following, we use the postorder numbers to define an ordering of the nodes of a forest F given by $v \prec v'$ iff *post*(v) < *post*(v'). Also, $v \preceq v'$ iff $v \prec v'$ or $v = v'$. Furthermore, we extend this ordering with two special nodes $\perp \prec v \prec \top$. The *left relatives*, lr(v), of a node $v \in V(F)$ is the set of nodes that are to the left of v and similarly the *right relatives*, rr(v), are the set of nodes that are to the right of v.

Based on the above concepts, we give the definition of ordered tree matching.

Definition 1 An embedding of a tree pattern P into an XML document T is a mapping $\varphi: P \to T$, from the nodes of P to the nodes of T, which satisfies the following conditions:

(i) Preserve node label: For each $u \in P$, *label*(u) = *label*($\varphi(u)$) (or say, u matches $f(u)$).

(ii) Preserve *parent-child*/*ancestor-descendant* relationship: If $u \to v$ in P, then $\varphi(v)$ is a child of $\varphi(u)$ in T; if $u \Rightarrow v$ in Q, then $\varphi(v)$ is a descendant of $\varphi(u)$ in T.

(iii)Preserve *left-to-right order*: For any two nodes $v_1 \in P$ and $v_2 \in P$, if v_1 is to the left of v_2, then $\varphi(v_1)$ is to the left of $\varphi(v_2)$ in T. □

If there exists such a mapping from P to T we say, T includes P, T contains P, T covers P, or say, P can be embedded in T.

Fig. 4 shows an example of an ordered tree embedding.

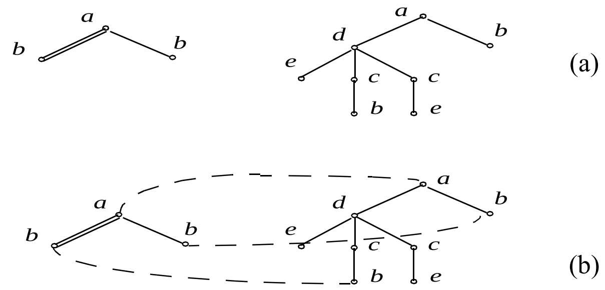


Fig. 4: (a) The tree on the left can be matched to a subtree in the tree on the right. (b) The dashed lines show a tree embedding.

Let P and T be two labeled ordered trees. An embedding φ of P in T is said to be *root-preserving* if φ(*root*(P)) = *root*(T). If there is a root-preserving embedding of P in T, we say that the root of T is an occurrence of P.

Fig. 4(b) also shows an example of a root preserving embedding. Obviously, restricting to root-preserving embedding does not lose generality. In fact, what can be found by the top-down algorithm to be discussed is a root-preserving tree embedding.

Throughout the rest of the paper, we refer to the labeled ordered trees simply as trees.

3 Algorithm

In this section, we give our algorithm. For simplicity, we consider only the case that a query tree contains only //-

edges. But it is an easy task to extend the algorithm for general cases.

Let G = <P_1, ..., P_l> ($l \geq 1$) be a forest. Consider a node v in G with children v_1, ..., v_j, ordered from left to right. We will use <v_k, i> ($1 \leq k \leq j$; $1 \leq i \leq j - k + 1$) to represent an ordered forest containing i subtrees of v: <$G[v_k]$, ..., $G[v_{k+i-1}]$>. Let v be a node on the left-most path in P_1. We call <v, i> a *left corner* of G. Denote by p_j the root of P_j in G = <P_1, ..., P_l> ($j = 1, ..., l$). Then, the left corner <p_1, i> represents the forest <P_1, ..., P_i> ($i \leq l$). In addition, we use $\delta(v)$ to represent a link from a node v in G to the left-most leaf node in $G[v]$, as illustrated in Fig. 5.

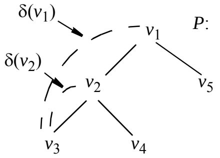


Fig. 5. A pattern tree and illustration for $\delta(v_1)$

Let v' be a leaf node in G. $\delta(v')$ is defined to be a link to v' itself. So in Fig. 5, we have $\delta(v_1) = \delta(v_2) = \delta(v_3) = v_3$. We also denote by $\delta^{-1}(v')$ a set of nodes $\mathbf{x}$ such that for each $v \in \mathbf{x}$ $\delta(v) = v'$. Therefore, in Fig. 5, $\delta^{-1}(v_3) = \{v_1, v_2, v_3\}$. The out-degree of v in a tree is denoted by $d(v)$ while the height of v is denoted by $h(v)$, defined to be the number of edges on the longest downward path from v to a leaf. The height of a leaf node is set to be 0.

Our algorithm mainly contains two functions: *top-down*(T, G) and *bottom-up*(F, G) to check tree matching, where T is a tree, and F and G are two forests. Each of the two functions returns a left corner <v, i> of G (*i.e.*, v is a node on the left-most path of P_1) such that

- <$G[v_1]$, ..., $G[v_i]$> can be embedded in T or in F, where $v_1 = v$, v_2,..., v_i are consecutive siblings; and
- there is no other left corner <v', j> with v' being an ancestor of v, which can be embedded in T or in F. (In other words, <v, i> is the highest left corner in G such that it can be embedded in T.)

If $v = p_1$ (the root of P_1), it shows that P_1, ..., P_i can be embedded in T or in F.

If the target (a document tree) is a tree and the pattern (a query tree) is a forest, we call the function *top-down*. If both the target and the pattern are forests, we call the function *bottom-up*. But during the computation, they will be called from each other.

In addition, each time a call *top-down*(T, G) returns a pair <v, i>, the root t of T is associated with that pair, referred to as $\kappa(t)$. Initially, each $\kappa(t)$ is set to ϕ. $\kappa(t)$ is mainly used in *bottom-up*( ) to avoid redundancy.

Let T = <t; T_1, ..., T_k>. Denote by t_s the root of T_s ($s = 1, ..., l$). We use *top-down*(t, <p_1, l>) to represent *top-down*(T, G), which is designed to check T and G top-down. For a given G, two cases are recognized:

Case 1: G = <P_1>; or G = <P_1, ..., P_l> ($l > 1$), but $|T| \leq |P_1| + |P_2|$. (That is, G is a forest containg only a single tree or a proper forest but the size of the first two subtrees is equal to or larger than the size of T.)

Case 2: G = <P_1, ..., P_l> ($l > 1$), and $|T| > |P_1| + |P_2|$.

In *Case* 1, what we can do is to find a left corner within P_1, which can be embedded in T. This is done as follows:

i) If t is a leaf node, we will check whether label(t) = label($\delta(p_1)$) (note that p_1 is the root of P_1.) If it is the case, set $\kappa(t)$ to be a triplet [$\delta(p_1)$, 1] and return <$\delta(p_1)$, 1>. Otherwise, set $\kappa(t)$ to be [$\delta(p_1)$, 0] and return <$\delta(p_1)$, 0>.

ii) If $|T| < |P_1|$ or $h(t) < h(p_1)$, we will make a recursive call *top-down*(t, <p_{11}, j>), where p_{11} is the left-most child of p_1 and $j = d(p_1)$. So <p_{11}, j> represents a forest of the subtrees of p_1: <P_{11}, ..., P_{1j}>. The return value <v, i> of *top-down*(t, <p_{11}, j>) is used as the return value of *top-down*(t, <p_1, l>).

iii) If $|T| \geq |P_1|$ and $h(t) \geq h(p_1)$, we further distinguish between two cases:

- label(t) = label(p_1). In this case, we will call *bottom-up*(<t_1, k>, <p_{11}, j>), by which <P_{11}, ..., P_{1j}> will be checked against <T_1, ..., T_k>.
- label(t) $\neq$ label(p_1). In this case, we will call *bottom-up*(<t_1, k>, <p_1, 1>), by which P_1 will be checked against <T_1, ..., T_k>.

In both cases, assume that the return value is <v, i>. A further checking needs to be conducted:

- If label(t) = label(v's parent) and i = d(v's parent), the return value should be <v's parent, 1>. Set $\kappa(t)$ to be [v's parent, 1].
- Otherwise, the return value remains <v, i>. Set $\kappa(t)$ to be [v, i].

In *Case* 2, we try to find a left corner within G = <P_1, ..., P_l>, which can be embedded in T. This is done by calling *bottom-up*(<t_1, k>, <p_1, l>). Assume that the return value is <v, i>. The following checkings will be continually conducted.

iv) If $v = p_1$, the return value of *top-down*(t, <p_1, l>) is the same as <v, i>.

v) If $v \neq p_1$, check whether label(t) = label(v's parent) and $i = d(v)$. If it is the case, the return value will be changed to <v's parent, 1>, and $\kappa(t)$ is set to be [v's parent, 1]. Otherwise, the return value remains <v, i>, and $\kappa(t)$ is set to be [v, i].

The following is the formal description of the algorithm *top-down*(t, <p_1, l>), in which we assume that each node v has a link to its direct sibling, making a sibling chain. Starting from p_1, we can access p_1, ..., p_l along the sibling chain.

Function *top-down*(t, <p_1, l>)

```
input: $t$ - stands for $T = <t; T_1, ..., T_k>$, $<p_1, l>$ - for $G = <P_1, ..., P_l>$.
output: $<v, i>$ specified above.
begin
1.  if ($l = 1$ or $|T[t]| \le |G[p_1]| + |G[p_2]|$)
2.  then { let $p_{11}$ be the left-most child of $p_1$; let $j$ be $d(p_1)$;
                                        (*Case 1*)
3.      if $t$ is a leaf then {if label($t$) = label($\delta(p_1)$)
                                    then $i := 1$ else $i := 0$;
4.                                  $\kappa(t) := [\delta(p_1), i]$; return $<\delta(p_1), i>$;}
5.      if ($|T[t]| < |G[p_1]|$ or $h(t) < h(p_1)$)
6.      then {$<v, i>$ := top-down($t$, $<p_{11}, j>$); return $<v, i>$;}
7.      if label($t$) = label($p_1$)  (*$|T| \ge |P_1|$ and $h(t) \ge h(p_1)$*)
8.      then {if $p_1$ is a leaf then {$v := p_1$; $i := 1$;}
9.              else {$<v, i>$ := bottom-up($<t_1, k>$, $<p_{11}, j>$);
10.                 if label($t$) = label($v$'s parent) and
                       $i = d$($v$'s parent)
                   then {$v := v$'s parent; $i := 1$;}
11.             }
12.             else $<v, i>$ := bottom-up($<t_1, k>$, $<p_1, 1>$);
                   (*If label($t$) $\ne$ label($p_1$), call bottom-up( ).*)
13.             $\kappa(t) := [v, i]$; return $<v, i>$;
14.         }
15.     else {$<v, i>$ := bottom-up($<t_1, k>$, $<p_1, l>$);
                                        (*Case 2*)
16.             if $v \ne p_1$ then { $p := v$'s parent;
17.                                 if (label($t$) = label($p$)) and $i = d(p)$
18.                                 then {$v := p$; $i := 1$; }
19.                                 $\kappa(t) := [v, i]$;
20.              }
21.              return $<i, v>$;
22.         }
end
```

The above algorithm mainly consists of two parts: lines 2 - 14 for *Case* 1, and lines 15 - 22 for *Case* 2. In the first part, we first handle the case that T contains only a single node (see lines 3 - 4); and then the case that $|T| < |P_1|$ or $h(t) < h(p_1)$ (see lines 5 - 6). The lines 7 - 14 are devote to the case that $|T| \ge |P_1|$ and $h(t) \ge h(p_1)$. If label(t) = label(p_1), we need to check whether p_1 is a leaf node. If it is the case, return $<p_1, 1>$ (see line 8). Otherwise, the bottom-up procedure will be invoked to check $<P_{11}, ..., P_{1j}>$ against $<T_1, ..., T_k>$ (see line 9). If label(t) $\ne$ label(p_1), the bottom-up procedure is invoked to check P_1 against $<T_1, ..., T_k>$ (see line 12).

In the second part, the bottom-up procedure is invoked to check $<T_1, ..., T_k>$ against $<P_1, ..., P_l>$ (see line 15). Finally, We notice that each time a node t is checked $\kappa(t)$ is changed to a new value, which is the return value of the current *top-down* execution (see lines 4, 13, and 19).

bottom-up(F, G) is designed to handle the case that both F and G are forests with each containing some subtrees rooted at a set of consecutive siblings in the target and the pattern, respectively. Let $F = <T_1, ..., T_k>$. We use *bottom-up*($<t_1, k>$, $<p_1, l>$) to represent *bottom-up*(F, G). In *bottom-up*($<t_1, k>$, $<p_1, l>$), we will make a series of calls of the form *top-down*(t_i, $<p_{j_i}, l - j_i + 1>$), where $j_1 = 1$, and $j_1 \le j_2 \le ... \le j_h \le l$ (for some $h \le k$), controlled as follows.

1. Two index variables s, j are used to scan $t_1, ..., t_k$ and $p_1, ..., p_l$, respectively. (Initially, s is set to 1, and j is set to 0.) They also indicate that $<P_1, ..., P_j>$ has been successfully embedded in $<T_1, ..., T_s>$.
2. Let $<v_s, i_s>$ be the return value of *top-down*(t_s, $<p_{j+1}, l - j>$). If $t_s = p_{j+1}$, set j to be $j + i_s$. Otherwise, j is not changed. Set s to be $s + 1$. Go to (2).
3. The loop terminates when all T_s's or all P_j's are examined.

See Fig. 7. for illustration.

If $j > 0$ when the loop terminates, *bottom-up*($<t_1, k>$, $<p_1, l>$) returns $<p_1, j>$.

Otherwise, $j = 0$. In this case, we will continue to search for a left corner $<v, i>$ in G, which can be embedded in F, as described below.

i) Let $<v_1, i_1>, ..., <v_k, i_k>$ be the return values of *top-down*(t_1, $<p_1, l>$), ..., *top-down*(t_k, $<p_1, l>$), respectively. Since $j = 0$, each v_f ($f = 1, ..., k$) must be a descendant of p_1 and on the left-most path in P_1.

ii) If each $i_f = 0$ ($f = 1, ..., k$), return $<\delta(p_1), 0>$. Otherwise, there must be some $<v_f, i_f>$'s such that $i_f > 0$. We call such a v_f a *non-zero point*. Find the first non-zero point v_f such that v_f is not a descendant of any other non-zero point. Let $w_1, ..., w_h$ be the right siblings (in this order) of v_f. We will further check $<T_{f+1}, ..., T_k>$ against $<G[w_{i_f}], G[w_{i_f+1}]..., G[w_h]>$. This can be done in the same way as described above. But it is not necessary to record the highest non-zero point. If it is found that $<T_{f+1}, ..., T_k>$ embeds the first q subtrees in $<G[w_{i_f}], G[w_{i_f+1}]..., G[w_h]>$, the return value of *bottom-up*($<t_1, k>$, $<p_1, l>$) is set to be $<v_f, i_f + q>$. Otherwise, the return value is $<v_f, i_f>$.

In this process, a node t in F may be checked multiple times due to the second checking described in (ii). In order to avoid any possible redundancy, we define a simple function as below.

Let v, v' be two nodes in G. Define

$$\beta(v, v') = \begin{cases} \textit{true}, & \text{if } v = v'\text{, or } \delta(v) = \delta(v') \text{ and } v' \text{ is} \\ & \text{an ancestor of } v; \\ \textit{false}, & \text{otherwise.} \end{cases}$$

During the execution of *bottom-up*( ), this function will be used each time we make a call of the form *top-down*(t, $<p, l>$) for a node t in F. Let $\kappa(t) = [v, i]$. If $\beta(v, p) = \textit{true}$, we simply set the return value of *top-down*(t, $<p, l>$) to be $<v, i>$ and *top-down*(t, $<p, l>$ is not actually executed. It is because $<v, i>$ is the highest left corner of some forest in G that can be embedded in $F[t]$, and therefore for any ancestor p of v with $\delta(v) = \delta(p)$ a call of the form *top-down*(t, $<p, l>$) will definitely return $<v, i>$.

Obviously, if p is a descendant of v and $i > 0$, the return value should be $<p, l>$. But if $i = 0$, the return value is $<p, 0>$.

In terms of the above discussion, we give the following algorithm to implement the bttom-up procedure, in which a subprocedure *td-checking*( ) is invoked to check a T_i against a forest $<P_{j_i}, ..., P_l>$, including the redundancy checking by using $\kappa(t)$'s.

```
Function bottom-up(<t_1, k>, <p_1, l>)
input: <t_1, k> - stands for F = <T_1, ..., T_k>,
       <p_1, l> - for G = <P_1, ..., P_l>.
output: <v, i> specified above.
begin
1.   $s := 1; j := 0;\ t := t_1; p := p_1;\ \tau_f := 1; v_f := \phi; i_f := 0;$
         (* $\phi$ is considered to be a descendant of any node.*)
2.   while ($j < l$ and $s \leq k$) do(*first checking*)
3.   { $<v, i> :=$ td-checking($t, p, j, l$);
4.     if ($v = p$ and $i > 0$) then {$j := j + i; p := p_{j+1}$;}
          (*navigate along the sibling chain to find $p_{j+i+1}$.*)
5.     else if $v$ is an ancestor of $v_f$ then {$v_f := v;\ i_f := i;\ \tau_f := s$;}
          (*record the highest non-zero point.*)
6.     $s := s + 1; t := t_s$;
          (*navigate one step along the sibling chain to find $t_{s+1}$.*)
7.   }
8.   if $j > 0$ then return $<p_1, j>$;
9.   if $i_f = 0$ then return $<\delta(p_1), 0>$
10.  let $d(v_f$'s parent$) = c$; find $v_f$'s $(i_f + 1)th$ right sibling $w_{i_f}$;
          (*Let $w_1, ..., w_c$ be the right siblings of $v_f$.*)
11.  $x := \tau_f + 1; y := i_f; t := t_{\tau_f + 1}; p := w_{i_f}$;
12.  while ($y < c$ and $x \leq k$) do(*second checking*)
13.  {    $<v, i> :=$ td-checking($t, p, y, c$);
14.       if ($v = p$ and $i > 0$) then {$y := y + i; p := w_{y+1}$;}
15.       $x := x + 1; t := t_x$;
16.  }
17.  if $y > 0$ then return $<v_f, i_f + y>$ else return $<v_f, i_f>$;
18.}
end
```

```
Function td-checking(t, p, j, l)
input: t - a node in F; p - a node in G; j, l - two integers with $j \leq l$.
output: <v, i> specified above.
begin
1.   let $\kappa(t) = [\gamma, \eta]$;
2.   if $\beta(\gamma, p) = true$ then {$v := \gamma,\ i := \eta$;}
3.   else {if p is a descendant of $\gamma$
           then {$v := p$; if $\eta = 0$ then $i := 0$ else $i := l - j$;}
4.         else $<v, i> :=$ top-down($t, <p, l - j>$);
5.         }
6.    return $<v, i>$;
end
```

In *bottom-up*( ), the variables *s* and *j* are used to scan T_1, ..., T_k and P_1, ..., P_l, respectively, while the variables *t* and *p* are used to store the roots of the current T_s and P_{j+1} (see line 1). The variables v_f and i_f are for storing the highest non-zero point, and τ_f is for the root of the corresponding T_f.

As described above, the algorithm involves two times of checkings. The first checking is done in lines 2 - 7 while the second checking is conducted in lines 10 - 16. Whether the second checking will be carried out depends on the checking result performed in lines 8 and 9.

First, in lines 2 - 7, we do a series of checkings of T_i against $<P_{j_i}, ..., P_l>$ ($i = 1, ..., h$, $1 \leq h \leq k$) and each is done by calling *td-checking*( ) (see line 3), in which $\kappa(t)$'s are checked to eliminate redundancy (see lines 2 - 3 in *td-checking*( )). Line 5 is devoted to the computation of the highest non-zero point $<v_f, i_f>$.

If $j > 0$, the return value of *bottom-up*($<t_1, k>$, $<p_1, l>$) is $<p_1, j>$ (see line 8). If $j = 0$ and $i_f = 0$, the return value is $<\delta(p_1), 0>$ (see line 9). In both cases, the second checking will not carry on. Therefore, we call the following condition the *second-checking* condition:

$j = 0$ and $i_f > 0$.

If the above condition holds, the second checking will be conducted (see lines 10 - 16). This is almost the same as line 2 - 7. But no computation is arranged to record the highest non-zero point. In line 17, we calculate the return value for the case of $j = 0$.

Example 1 Consider the tree *T* and the forest *G* shown in Fig. 6. As indicated by the dashed lines, we have an ordered embedding of a subtree of *G* in *T*.

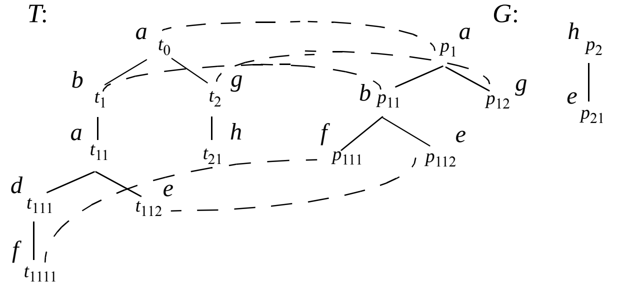


Fig. 6. A target tree and a pattern tree

In Fig. 6, each node in *T* is identified with t_i, such as t_0, t_1, t_{11}, and so on; and each node in *G* is identified with p_j. Besides, each subtree rooted at t_i (p_j) is represented by T_i (resp. P_j). In Fig. 7, we trace the computation process when applying the algorithm to *T* and *G*. In this figure, a solid arrow represents a subprocedure call while each dashed arrow represents a return value. Associated with a solid arrow is the condition under which the subprocedure is invoked.

The return value of the whole procedure is $<p_1, 1>$, showing that *T* contains P_1.

From the sample trace, we can see that a node in *T* can be checked multiple times, but against different nodes in *G*. For instance, t_{112} is first checked against p_{111}, and then against p_{112}. t_2 is also checked two times, against p_{111} and p_{12}, respectively.

4 Computational complexities

In this section, we analyze the computational complexities of the algorithm.

In the algorithm discussed in the previous section, a node *t* in *F* may be involved in multiple calls of the form *top-down*(t, $<p, l>$) due to a possible second checking in *bottom-up*( ).

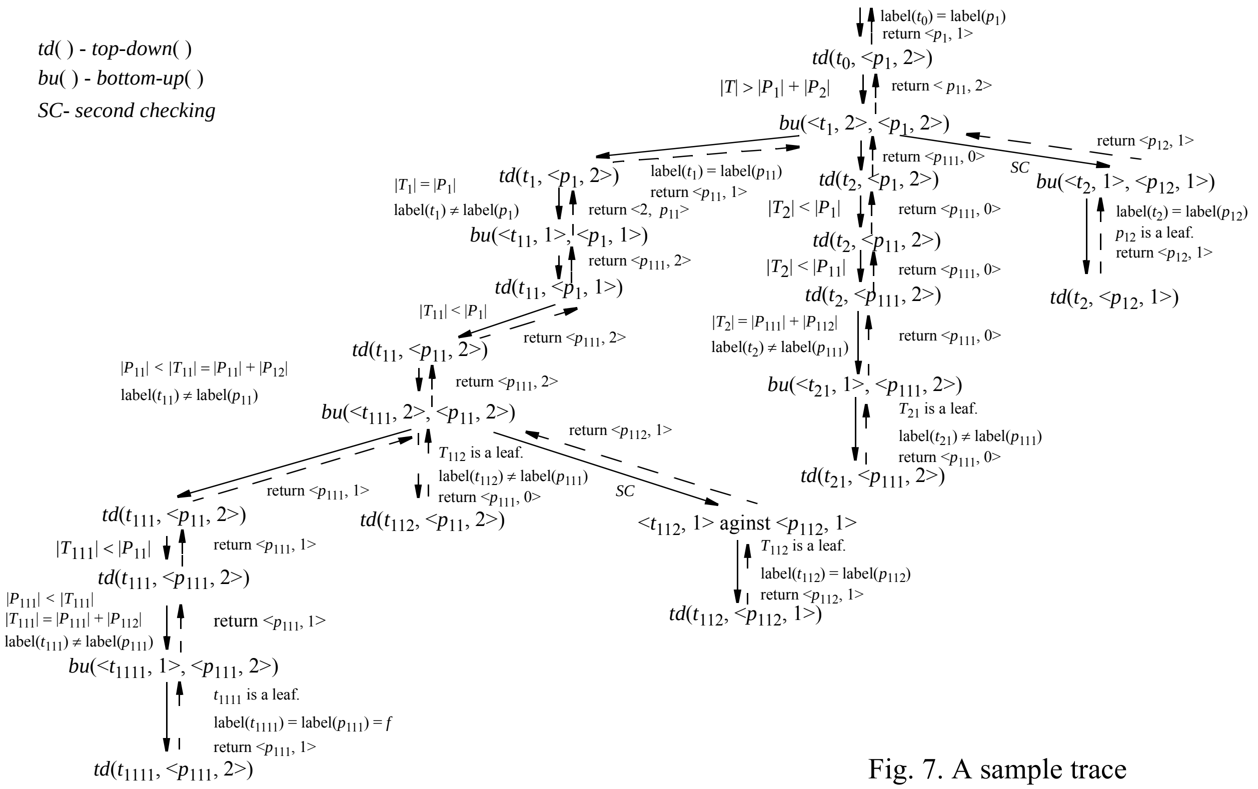


Fig. 7. A sample trace

In the algorithm discussed in the previous section, a node t in F may be involved in multiple calls of the form *top-down*(t, <p, l>) due to a possible second checking in *bottom-up*( ). We denote by [t, p] each of such calls for simplicity. We further distinguish two kinds of [t, p]'s. During a [t, p] of the first kind, t is checked against a node in G, which is done in line 3, line 7, or in line 17 in *top-down*( ).

During a [t, p] of the second kind, we navigate to the left-most child of p if p is not a leaf node (see line 6.)

First, we estimate the number of the calls of the first kind. Without loss of generality, assume that the first [t, p] is invoked by executing line 3 in *bottom-up*( ) to check <P_1, ..., P_l> against <T_1, ..., T_k>. It is possible for t to be involved in a second subprocedure call [t, p'] (see line 13 in *bottom-up*( )). Obviously, p' must be a descendant of p. Also, p' cannot be a node on the left-most path in $G[p]$ due to the second-checking condition: $j = 0$ and $i_f > 0$, where <v_f, i_f> is the first highest non-zero point and $j = 0$ indicates that even P_1 cannot be embedded in <T_1, ..., T_k>.

Since $j = 0$, v_f must be a node on the left-most path in $G[p]$. But its (i_f + 1)*th* right sibling is definitely not on such a path (see line 10 in *bottom-up*( )). So p' is not on the left-most path in $G[p]$.

Now we consider a child t_j of t. Clearly, during the execution of [t, p], t_j can also be involved in two subprocedure calls [t_j, u_1] and [t_j, u_2] while during the execution of [t, p'] t_j can be involved in another two subprocedure calls [t_j, u_1'] and [t_j, u_2']. As discussed above, u_2 cannot be on the left-most path in $G[u_1]$, and u_2' cannot be on the left-most path in $G[u_2]$. Concerning u_2 and u_1', we claim that

u_1' is a node appearing in a subtree to the right of u_2.

Below we show this property.

Consider all the left siblings t_s of t. Let <v_s, i_s> be the return value of the corresponding *top-down*(t_s, <p, l>). Let <v, i> be the return value of *top-down*(t, <p, l>). We distinguish among three cases:

i) For any <v_s, i_s>, v_s is a descendant of v.

ii) There is at least one non-zero point v_s (*i.e.*, $i_s > 0$), which is an ancestor of v and not a descendant of any other non-zero point.

iii) There is at least one non-zero point $v_s = v$, which is not a descendant of any other non-zero point.

In case (i), t will not be checked for a second time at all since by the second checking it must be a forest (or a tree) with the first subtree rooted a node to the right of t against a forest (or a tree) in G.

In case (ii), p' must be a node appearing in a subtree to the right of v_s while u_2 is definitely a node in the subtree rooted at v or at a jth right sibling of v with $j \leq i - 1$, and therefore a descendant of v_s. Since u_1' is in the subtree rooted at p', it is to the right of u_2. for illustration.)

In case (iii), we have $v_s = v$. If $i_s \geq i$, p' is definitely to the right of u_2, and so is u_1'. (See Fig. 10(b) for illustration.) In the following, we analyze the case when $i_s < i$.

Let $t_1, ..., t_{j-1}$ be all the left sibling of t_j. Consider $v_1 = v$ and all its right siblings $v_2, ..., v_l$. If u_2 is a node in a subtree rooted at v_q with $q \leq i_s$, u_1' must be a node to the right of u_2. Otherwise, assume that u_2 is a node in a subtree rooted at $v_{q'}$ with $i_s < q' \leq i$. Then, we have $<F[t_1], ..., F[t_{j-1}]>$ embedding $<G[v_1], ..., F[v_{q'-1}]>$. Therefore, $<F[t_1], ..., F[t_{j-1}]>$ must embed $<G[v_{i_s+1}], ..., F[v_{q'-1}]>$. Thus, p' can be $v_{q'}$ or to the right of $v_{q'}$. If p' is $v_{q'}$, u_1' can be an ancestor of u_2, equal to u_2, or a descendant of u_2. (Also, see Fig. 10(c) for illustration.) In any case, the corresponding checking is skipped by using $\kappa(t_j)$. If p' is to the right of $v_{q'}$, u_1' must be to the right of u_2.

The above discussion shows that the claim concerning u_2 and u_1' holds.

Mapping u_1 (u_1') to a node on the left-most path in $G[u_1]$ ($G[u_1']$), we think that t_j is involved in four $[t, v]$'s with each v on a different path in G. So we claim that the number of the first kind of calls is bounded by $O(|T|\cdot|\text{leaves}(G)|)$.

Now we consider the second kind of *top-down* calls. For each t in T, corresponding to a checking of it against a node in G, a downward segment in G may be searched; and for any of its children a segment following that segment may also be searched. So corresponding to a path in T, for all the checkings of the nodes on that path with each checked once, a path in G may be navigated. According to the above analysis, however, a node in T may be checked against different nodes on different paths in G. So the number of the second kind of calls is bounded by $O(|\text{leaves}(T)|\cdot|P|)|$.

Proposition 3 The time complexity of the algorithm is bounded by $O(|T|\cdot|\text{leaves}(G)| + |\text{leaves}(T)|\cdot|P|)|)$.

Proof. See the above analysis. □

Since in the working process no extra data structure is used, we have the following proposition.

Proposition 3 The space complexity of the algorithm is bounded by $O(|T| + |G|)$.

Proof. It is trivially true. □

5 Conclusion

In this paper, a new algorithm is proposed to evaluate XML queries based ordered tree matching, by which not only the ancestor/descendant and parent/child relationships, but also the left-to-right order of nodes are considered. The algorithm mainly contains two functions: *Top-down*( ) and *Bottom-up*( ). Each of them returns a left corner to indicate a subtree (subforest) embedding. This arrangement enables us to use a simple data structure to record intermediate results to avoid redundancy. The time complexity of the new algorithm is bounded by $O(|T|\cdot|\text{leaves}(P)| + |P|\cdot|\text{leaves}(T)|)$ while the space requirement is bounded by $O(|T| + |P|)$, where T and P are a target and a pattern tree, respectively.

References

[1] N. Bruno, N. Koudas, and D. Srivastava, Holistic Twig Joins: Optimal XML Pattern Matching, in *Proc. SIGMOD Int. Conf. on Management of Data*, Madison, Wisconsin, June 2002, pp. 310-321.

[2] B. Catherine and S. Bird, Towards a general model of Interlinear text, in *Proc. of EMELD Workshop*, Lansing, MI, 2003.

[3] T. Chen, J. Lu, and T.W. Ling, On Boosting Holism in XML Twig Pattern Matching, in: *Proc. SIGMOD*, 2005, pp. 455-466.

[4] Y. Chen, A time optimal algorithm for evaluating tree pattern queries, *SAC 2010*, ACM, 1638-1642.

[5] Y. Chen, Donovan Cooke: XPath query evaluation based on the stack encoding, *C3S2E 2009*, IEEE, 43-57

[6] Y. Chen: Unordered Tree Matching and Tree Pattern Queries in XML Databases, *ICSOFT (2) 2009*: 191-198.

[7] Y. Chen, An Efficient Streaming Algorithm for Evaluating XPath Queries. in *Proc. WEBIST*, 2008, pp. 190-196.

[8] Y. Chen, S.B. Davison, Y. Zheng, An Efficient XPath Query Processor for XML Streams, in *Proc. ICDE*, Atlanta, USA, April 3-8, 2006.

[9] Sayyed Kamyar Izadi, Theo Härder, Mostafa S. Haghjoo, S^3: Evaluation of Tree-Pattern Queries Supported by Structural Summaries, *Data & Knowledge Engineering*, 68, pp. 126-145, Elsevier, Sept. 2008.

[10] D.E. Knuth, *The Art of Computer Programming, Vol. 1 (1st edition)*, Addison-Wesley, Reading, MA, 1969.

[11] R.B. Lyngs, M. Zuker & C.N.S. Pedersen, Internal loops in RNA secondary structure prediction, in *Proceedings of the 3rd annual international conference on computational molecular biology (RECOMB)*, 260-267 (1999).

[12] Q. Li and B. Moon, Indexing and Querying XML data for regular path expressions, in: *Proc. VLDB*, Sept. 2001, pp. 361-370.

[13] Y. Rui, T.S. Huang, and S. Mehrotra, Constructing table-of-content for videos, *ACM Multimedia Systems Journal, Special Issue Multimedia Systems on Video Libraries*, 7(5):359-368, Sept 1999.

[14] L. Qin, J.X. Yu, and B. Ding, “TwigList: Make Twig Pattern Matching Fast,” In *Proc. 12th Int’l Conf. on Database Systems for Advanced Applications (DASFAA)*, pp. 850-862, Apr. 2007.

[15] H. Wang, S. Park, W. Fan, and P.S. Yu, ViST: A Dynamic Index Method for Querying XML Data by Tree Structures, *SIGMOD Int. Conf. on Management of Data*, San Diego, CA., June 2003.

[16] World Wide Web Consortium. XML Path Language (XPath), W3C Recommendation, 2007. See http://www.w3.org/TR/xpath20.

[17] World Wide Web Consortium. XQuery 1.0: An XML Query Language, W3C Recommendation, Version 1.0, Jan. 2007. See http://www.w3.org/TR/xquery.

[18] M. Zaki. Efficiently mining frequent trees in a forest. In *Proc. of KDD*, 2002.

Decision tree strategy for web service composition in WeSCo_CBR approach

IKE'11 conference

Soufiene Lajmi[1], **and Khaled Ghedira**[2]
[1,2]SOIE Research Unit, Institut Supérieur de Gestion, University of Tunis, Tunis, Tunisia
[1]soufiene.lajmi@ensi.rnu.tn, [2]khaled.ghedira@isg.rnu.tn

Abstract— *Composing web services by using prior cases within a case base to provide value-added services has not been as much studied. In this paper, we present an approach called WeSCo_CBR based on case based reasoning (CBR) to compose web services. In our proposal, web services are organized into service classes while prior cases are classified into case classes. Web services within the same service class have the same functionality when prior cases belonging the same case class are the most similar. Moreover, we use CBR techniques for selecting the related web service composition experiences according to the user query. Therefore, the decision tree is utilized for browsing the case base to find the case class containing the most similar cases to user query.*

Keywords: Web Services (WS), Web Service Composition, Case-Based Reasoning (CBR), clustering, browsing, Decision tree

1. Introduction

Web services are autonomous software components advertised and invoked over Internet. The use of Internet technologies makes applications able to access to remote, autonomous and heterogenous systems without regard to the operating system and implementation language. We can classify web services into two categories: simple services and composite services. In this work, we are interested in the later category.

The most important problem to provide more complex and more useful solutions is to discover, and then compose web services. In the last few years, several initiatives in web service composition have been implemented to help the user in finding and composing web services within large repositories. Nevertheless, web service composition still remains complex. Moreover, it is already beyond the human ability to make the composition manually.

In [3], we have proposed an approach for web service composition called WeSCo_CBR and founded on case based reasoning (CBR). We use the CBR techniques to build a composition diagram of a composite web service. A composition diagram is defined as follows:

Definition *1:* A composition diagram is a service specification represented by a set of orchestrated service classes. □

In this paper, we are interested in the way to organize the case base in WeSCo_CBR. Then, we present a strategy based on decision tree and utilized in the retrieval process of a similar case to a given query. The adaptation step of the retrieved similar case is out of the scope of this paper.

After the composition diagram building, it remains to carry out for each class in the diagram the web service fulfilling the best user query.

However, the composition diagram building appears the most crucial step and is a complex task to do. We aim to address it by a best and suitable way. So, we propose to apply the Case Based Reasoning at a such step. This type of reasoning consists of finding in the case base, the similar cases to a new user query. The composition diagrams of the found cases are useful to build a new composition diagram for a composite service which can fulfill user query.
The next step is to select within classes contributing in the composition diagram the most relevant services. Eventually, having composition diagram and web services concerned by this composition, we can then build a composite web service.

The outline of the paper is as follows: Section 2 gives an overview of our proposal. Section 3 presents the way to use CBR based approach. Then, section 4 takes an interest in the classification process and similarity computation. Next, section 5 presents the strategy used in the retrieval process. An illustrating example in section 6. Section 7 discusses related work. Immediate development and future work will be presented in section 8.

2. WeSCo-CBR overview

This section gives an overview of our proposal [3] to understand how it works. WeSCo_CBR [3] consists of five components and each of them has a fundamental function. Figure 1 presents the architecture overview.

1) Query reformulation module. It is an interface component in WeSCo_CBR. It accepts the query from the user and reformulates it in a comprehensible and computer interpretable form. This module allows to transform the query into ontological concepts. The query is transformed into three main concepts as presented in Figure 2.The result returned by this

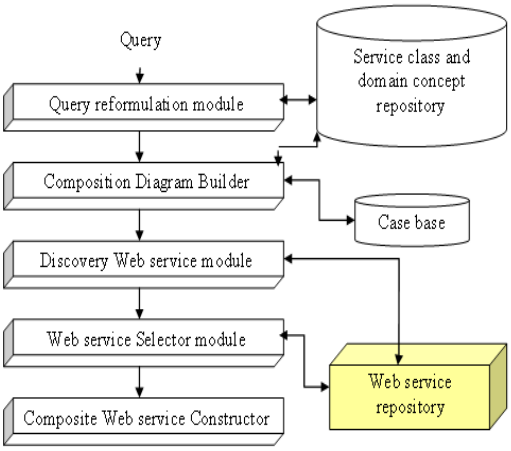

Query
Query reformulation module
Service class and domain concept repository
Composition Diagram Builder
Case base
Discovery Web service module
Web service Selector module
Web service repository
Composite Web service Constructor

Figure 1: WeSCo_CBR architecture

module is an OWL[15] document describing different components of a user query.

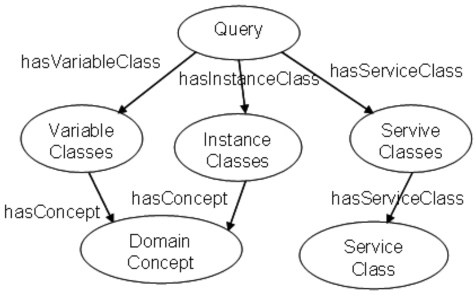


Figure 2: Query representation

- Instance classes: they represent the instantiated and pertinent concept set in a given user query. Indeed, a query may contain a data set. The data set is considered as values for object proprieties. All concepts instantiated by these objects constitute the "Instance Classes" for the query.
- Variable classes: they contain the set of the non-instantiated and pertinent concepts in a user query. Indeed, a query may have variables (proprieties without values in the query). These variables may be proprieties for one or more concepts. Such concepts constitute the "Variable Classes" for the query.
- Service classes: they are a set of pertinent service classes in a given query.

2) Composition diagram builder. The role of this component is to build the composition diagram. Several researchers invest themselves in web service composition field. However, the majority of these initiatives are based either on a static workflow (EFlow [9], METEOR-S [10]), or the composition is made by the functional attributes, pre-conditions and post-conditions (SWORD [11]). In general, the type of compositions considers the requirements of web services but does not fulfil the user needs. Therefore, we propose an approach allowing to respond the user needs by starting from his query. In our approach, the composition diagram builder is the crucial component to elaborate a composite web service. Essentially, it uses the Case Based Reasoning technique.
3) Discovery web service module. Web service discovery is a crucial step in web service composition process. Discovery web service module allows to find web services which can substitute each service class in the composition diagram. Indeed, a service class in a composition diagram is a node in the hierarchical tree of service classes. So, this module aims to find the service class which is a descendant of this node (i.e the service class which is a sub-class of the service class contributing on the composition diagram) and it is the deepest one in the hierarchical tree.
4) Web service selector module. For each service class in a composition diagram, the Web service Discovery Engine provides a set of web services superseding it. We need to select one web service within each service class. The Web Service Selector module achieves this goal basing on the input and output parameters of the available web services, quality of service and execution time cost.
5) Composite web service constructor. After determining actual web services, the "Composite web service constructor" module carries out the task of the actual process construction. It allows the transformation of the composition diagram built by the "Composition diagram binder" in an executable actual process. It is the component which generates an executable OWL-S[2] process.

After presenting a light description of WeSCo_CBR architecture, next we will be interested in the use of case based reasoning.

3. How to use CBR based approach

Case-Based Reasoning (CBR) is one of the preferred problem-solving strategies and machine learning techniques in complex and dynamically changing situations [17]. It consists of four basic steps:
1. Find similar case (Remind)
2. Adapt the solution to the specific problem (Adapt)
3. Verify that this solution works or is reasonable (Evaluate)
4. Save this new case/solution pair for future use (Store).

The primary focus of this paper is on the first step (Find similar case). Moreover, the use of the case based reasoning requires the identification of a case. Therefore, we should represent a case in a best way. Regardless the application domain, a case has always the same components [13]: a *problem* representing the case, a *solution* for this case and its *valuation*. In our approach, we identify the following case components:

- *Problem.* The problem of a case is represented like a user query. It is composed of three parts: service classes, variable classes and instance classes.
- *Solution.* The solution of a case is the composition diagram.

  We identify three types of relations between service classes contributing in a composition diagram:

  - Relation of type $\approx$: this relation implies that the two services of the correspondent classes may be executed in parallel independently each to other.
  - Relation of type $\trianglerighteq$: this relation implies that the two services of the correspondent classes may be executed in parallel but the second service requires a result from the first.
  - Relation of type $\triangleright$: this relation implies that the second service can not start until the first finishes its execution.

  Each relation R is defined by four elements: two service classes (Sc_1, Sc_2) participating in the relation, a relation type RT and a constraint $Cstr$. The constraint indicates the parameters provided by the first service to the second.

  $$R = \{Sc_1, Sc_2, RT, Cstr\}$$

  Each composition diagram CS of a case C is a set of service classes $\mathcal{SC}$ and a set of relations $\mathcal{R}$.
  $CS = (\mathcal{SC}, \mathcal{R})$
  $\mathcal{SC} = \{(Sc_i, InputParameters(Sc_i),$
  $OutputParameters(Sc_i))\}$
  where Sc_i is a service class in the composition diagram and $\mathcal{R} = \{R_i\}$, R_i is a relation.
- *Valuation.* It is the relevance ratio of the solution. Due to the existence of irrelevant cases which do not fulfil the user's needs, we proposed a valuation criteria of the user to express his satisfaction degree to the suggested composition diagram. This valuation will be associated to the new memorized case.

  After defining the case components, it remains to define the similarity computation methods and case classification algorithm.

4. Case classification process and similarity computation

The size of the case base increases progressively, thus the exploration becomes more complex and the retrieval time grows. To reduce the retrieval space, we propose to classify the cases into classes. Each case class contains the most similar cases.

On the beginning, all cases are placed to an initial and unique class. The farthest case from its class, will be used to create a new class. Next, all cases which are near to the new class will migrate to it. This procedure is repeated until having a given number of classes. Besides, classification process requires to define similarity and pertinence computing methods between cases.

Let CB be the case base containing prior cases $\{C_i\}_{i \in 1..N}$. Each case C_i is characterized by three components:

$$C_i = \{Pb_i, CS_i, Ev_i\}$$

Pb_i: is the problem of the case C_i
CS_i: is the solution (composition diagram) of the case C_i
Ev_i: is the valuation of the case C_i

Each problem Pb_i of a case C_i consists of three parts:

$$Pb_i = (VC_i, IC_i, SC_i)$$

VC_i : is a set of variable classes of the problem Pb_i
IC_i : is a set of instance classes of the problem Pb_i
SC_i : is a set of service classes of the problem Pb_i

Let VC_i be the variable class part of a problem Pb_i of a case C_i:

$$VC_i = \{Vc_{ij}\}_{j \in 1..nv_i}$$

nv_i is the number of the variable classes of a case C_i
Vc_{ij} is the j^{th} variable class of a case C_i

Let IC_i be the instance classs part of a problem Pb_i of a case C_i:

$$IC_i = \{Ic_{ij}\}_{j \in 1..ni_i}$$

ni_i is the number of instance classes of a case C_i
Ic_{ij} is the j^{th} instance class of the case C_i

Let SC_i be the service class part of a problem Pb_i of a case C_i:

$$SC_i = \{Sc_{ij}\}_{j \in 1..ns_i}$$

ns_i is the number of service classes of the case C_i
Sc_{ij} is the j^{th} service classe of the case C_i

However service classes are organized as a tree. Thus, the number of branches separating two service classes has an effect on the similarity between these classes. So, we define the similarity sim_{sc} between two service classes Sc_1 and Sc_2 by the following formula:

$$sim_{Sc}(Sc_1, Sc_2) = \begin{cases} 1 & \text{if } Sc_1 = Sc_2 \\ (\frac{1}{d+1})^d & \text{if } Sc_1 \text{ is-subclass-of } Sc_2 \\ (\frac{1}{d+1})^{2d} & \text{if } Sc_2 \text{ is-sub-class-of } Sc_1 \\ 0 & \text{else.} \end{cases}$$

d is the number of branches between the two service classes.

The pertinence $per_{Sc}(Sc, C_i)$ of a service class Sc in a case C_i is the similarity between Sc and the most similar service class to Sc in C_i:

$$per_{Sc}(Sc, C_i) = max(sim_{Sc}(Sc, Sc_{ij})), \forall j \in [1..ns_i]$$

The pertinence PER_{SC} of a service class part of a case C_i in a case C_j is the sum of the pertinence of each service class Sc_{ik} of the case C_i in the case C_j:

$$PER_{SC}(C_i, C_j) = \sum_{k=1}^{ns_i} per_{Sc}(Sc_{ik}, C_j)$$

The similarity SIM_{SC} between two service class parts of two cases C_i et C_j is the sum of the pertinence values of each case in the other divided by the sum of the service classes in the two cases:

$$SIM_{SC}(C_i, C_j) = \frac{PER_{Sc}(C_i, C_j) + PER_{Sc}(C_j, C_i)}{ns_i + ns_j}$$

Variable and instance classes are concepts in an ontology domain. To compare two cases, we need to compute the similarity between their variable and instance classes. Thus, we define the similarity sim_{cp} between two domain concepts cp_1 and cp_2 as follows:

$$sim_{cp}(cp_1, cp_2) = \begin{cases} 1 & \text{if there is the same concept;} \\ x^p & \text{if } cp_1 \text{ is a sub-class of } cp_2; \\ x^{2p} & \text{if } cp_2 \text{ is a sub-class of } cp_1. \\ x^{3p} & \text{else.} \end{cases}$$

p is the maximum of distances of cp_1 and de cp_2 to their common parent.

$$x = \frac{2 * \ SharedProNum(cp_1, cp_1)}{PropNum(cp_1) + PropNum(cp_2)}$$

$SharedProNum(cp_1, cp_1)$ is the number of the shared proprieties between the tow concepts cp_1 and cp_1. $PropNum(cp)$ is the propriety number of the concept cp.

The pertinence per_{Vc} (resp. per_{Ic}) of a concept cp in the variable class part (resp. instance class part) of a case C_i is the maximum of the similarity between the concept cp and each variable class (resp. instance class) of the case:

$$per_{Vc}(cp, C_i) = max(sim_{cp}(cp, Vc_{ij})), \forall j \in [1..nv_i]$$

$$per_{Ic}(cp, C_i) = max(sim_{cp}(cp, Ic_{ij})), \forall j \in [1..ni_i]$$

The pertinence PER_{VC} (resp. PER_{IC}) of a variable class part (resp. instance class part) of a case C_i in a case C_j is the sum of pertinence values of each variable class (resp. instance class) of the case C_i in the case C_j defined as follow:

$$PER_{VC}(C_i, C_j) = \sum_{k=1}^{nv_i} per_{Vc}(Vc_{ik}, C_j)$$

$$PER_{IC}(C_i, C_j) = \sum_{k=1}^{ni_i} per_{Ic}(Ic_{ik}, C_j)$$

The similarity SIM_{VC} (resp. SIM_{IC}) between two variable class parts (resp. instance class parts) of two cases C_i and C_j is the sum of the variable class (resp. instance class) pertinence values of each case in the other:

$$SIM_{VC}(C_i, C_j) = \frac{PER_{VC}(C_i, C_j) + PER_{VC}(C_j, C_i)}{nv_i + nv_j}$$

$$SIM_{IC}(C_i, C_j) = \frac{PER_{IC}(C_i, C_j) + PER_{IC}(C_j, C_i)}{ni_i + ni_j}$$

The global similarity SIM_g between two cases C_i and C_j is an aggregation of the three similarity types SIM_{SC}, SIM_{VC} and SIM_{IC}:

$$SIM_g(C_i, C_j) = \alpha.SIM_{SC}(C_i, C_j) + \beta.SIM_{VC}(C_i, C_j) + \gamma.SIM_{IC}(C_i, C_j)$$

where $\alpha,\ \beta\ and\ \gamma$ are the importance of the similarity type verifying the next propriety: $\alpha + \beta + \gamma = 1$.

Using the similarity computation methods, we can present the algorithm utilized for case classification.
Let CLN be the suitable case class number and CN_k be the case number within the class CL_k. We define so the distance $dist(C_i^k)$ between the i^{th} case C_i^k of the class CL_k and its class as the reverse of the sum of similarities between the case C_i^k and all the cases belonging to its class CL_k.

$$dist(C_i^k)) = \frac{CN_k - 1}{\sum_{j=1, j \neq i}^{CN_k} SIM_g(C_i^k, C_j^k)}$$

This distance is used to choose the farthest case from its case class to create a new class containing this case.
For the classification, we need to compute the distance between a case C_i^k of a case class CL_k and a class CL_p where $CL_p \neq CL_k$.

$$DIST(C_i^k, CL_p)_{p \neq k} = \frac{CN_p}{\sum_{j=1}^{CN_p} SIM_g(C_i^k, C_j^p)}$$

This distance is used to know the class to which a case is nearest for a possible migration.

Thus, we define the minimal distance $minDIST(C_i^k)$ between a case C_i^k and the set of case classes as follow:

$$minDIST(C_i^k) = min\{DIST(C_i^k, CLp), p \in [1..n] - \{k\}\}$$

The case whose the difference between $dist(C_i^k)$ and $minDIST(C_i^k)$ is positive and maximum, is a case which must migrates from its class CL_k to the nearest class $CL_{\tilde{k}}$ verifying $DIST(C_i^k, CL_{\tilde{k}}) = minDIST(C_i^k)$. We deduce then the classification algorithm presented table 1.

This algorithm allows to classify the cases into case classes. The target of this classification is to avoid to browse the whole case base. The retrieval process is a crucial step to deal with. In the next section, we present the retrieval process within the case base.

```
Algorithm CaseClassification($CL_1$, $CLN$)
Input: $CL_1, CLN$
{$CL_1$ is the start class that contains the whole of cases in the case base and $CLN$ is the desired Class Number}
Output: ClassSet {Set of Classes}
Begin
n Integer=1 {we have a unique class $CL_1$}
repeat
$n \longleftarrow n+1$
{create a new class $CL_n$}
$\forall j \in [1..CN_p]$ et $\forall p \in [1..CLN]\}$
select the case $C_i^k$ / $dist(C_i^k) = max\{dist(C_j^p)\}$
{Migrate the case $C_i^k$ from $CL_k$ to $CL_n$ class}
migrate($C_i^k$, $CL_k \rightarrow CL_n$)
    while $\exists$ a case $C_i^k$ / $dist(C_i^k)$ - $minDIST(C_i^k) > 0$
    and $dist(C_i^k)$ - $minDIST(C_i^k) =$
    max{$dist(C) - minDIST(C), \forall C \in CB$}
    Do
    Let's $\tilde{k} \in [1..n]/minDIST(C_i^k) = DIST(C_i^k, CL_{\tilde{k}})$
    {Migrate the case $C_i^k$ from $CL_k$ to $CL_{\tilde{k}}$ class}
    migrate($C_i^k$, $CL_k \rightarrow CL_{\tilde{k}}$)
    end while
until $n = CLN$
End
```

Table 1: Case Classification algorithm

5. Retrieval process strategy

After classifying the cases into case classes, we need now to define a strategy to browse the classification tree. Therefore, we propose to use a strategy based on a decision tree. The result of the classification algorithm is a class hierarchy which will be then browsed to find the most similar case to a user query. Furthermore, a small set of cases named class core is associated to each class. It is useful in the retrieval process of a class in which we can find the most similar case and it helps us to memorize a new case in the case base.

The classification algorithm presented in section 4 allows to create case classes. Among them, we identify those we name intermediate classes or abstract classes. They are classes whose cases are migrating to (or coming from) another class at the execution of the classification algorithm. In contrast, there are classes which we name concrete classes. They are classes which have not been modified since their creation. Furthermore, we find in the case base only the concrete classes. However, intermediate classes are used to improve the class retrieval process. For example, in the beginning we have a unique CL_1 containing all cases. After the execution of the first algorithm classification step to create the second case class CL_2, we have at least a case migrating from the class CL_1 to the class CL_2 (it is a principle of the algorithm functionality) and the rest of cases which does not leave the class CL_1 constitute a new class CL_3. Moreover, the started class CL_1 is always an intermediate class. Figure 3 presents an example of a binary classification tree. For each class CL, the left child CL_l is the class created by migrating the first case from its parent CL and the right child is the class containing the remaining cases after case migration.

Classes which are leaves are concrete classes (In figure 3, the classes $CL8$, $CL9$, $CL5$, $CL6$, $CL10$ and $CL11$ are concrete classes).

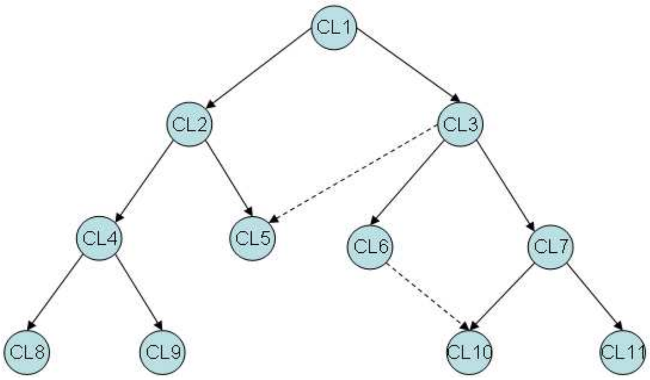


Figure 3: An example of a case classification tree

origin → traget: the sold arrow denotes that the origin of the target class is the class origin (target class may be a left or a right child).

origin ⇢ target: the dashed arrow denotes that there are cases migrating from the class origin to the class target.

The retrieval aim is to identify a concrete class in which we can find the most similar cases to a query. Now, we must define a strategy to browse the classification tree and retrieve within it. A trivial solution is to compare distances between a query and each child class of the current node and to choose the class whose distance is minimal. Nevertheless, computing this distance requires to explore at least one time the whole case base. So, it makes an important execution time. In order to reduce this time, each CL_k is represented by its center which we name core $\mathcal{N}cl_k$, consisting of a small set of class cases which minimize the distance criteria. The distance computing will be then reduced to this set.

For each case C_i^k within the core, its distance $dist(C_i^k)$ is low to all the distances of the other cases without the core: $\forall C_i^k \in CL_k$, $C_i^k \in \mathcal{N}cl_k \Rightarrow dist(C_i^k) < dist(C), \forall C \in CL_k - \mathcal{N}cl_k$

The number of cases to be included within the core (i.e the size of the core) depends on the computer performance. After creating the cores for all classes in the classification tree, each new query will be processed as follows:

1) The tree running begins from the root node.
2) At each current node, we compute distances between the query and the cores of child nodes.
3) The node having a minimal distance will be then explored and becomes the current node.
4) Step 2 and 3 are repeated until the current node becomes a leaf (a concrete class).

In WeSCo_CBR, we distinguish between two types of retrieval: case class retrieval containing the most pertinent case and the most similar case retrieval in the founded case class. Once the case class is found, it remains to select the most similar case within it to the query. The retrieval is done by filtering. The case having the highest global similarity will be presented to the user. The latter may use other cases to adapt the solution.

6. Implementation and evaluation

This section presents the implementation highlights of WeSCo_CBR. At the first step, we have applied the classification algorithm on a set of cases. Then, we have applied the retrieval algorithm for some queries. Table 2 presents the similarity measures between cases (values are given in rate).

Cas	1	2	3	4	5	6	7	8	9	10
1		22	87	35	42	11	24	73	31	08
2	22		19	28	15	68	32	41	27	91
3	87	19		16	38	13	29	69	25	33
4	35	28	16		81	44	77	06	70	20
5	42	15	38	81		36	67	14	82	23
6	11	68	13	44	36		40	29	15	74
7	24	32	29	77	67	40		18	75	04
8	73	41	69	06	14	29	18		24	41
9	31	27	25	70	82	15	75	24		30
10	08	91	33	20	23	74	04	41	30	

Table 2: Table of similarity rate between cases

The execution of the classification algorithm to get four case classes gives the classification tree presented in Figure 4. Classes $CL2$, $CL6$, $CL7$ and $CL5$ are the concrete classes because they are the tree leaves.

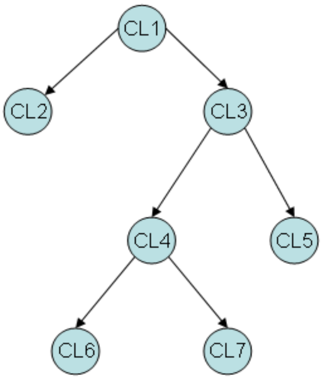


Figure 4: Classification tree corresponding to similarity measures in table 2

The contents of concrete classes CL_i after the execution of the classification algorithm are:
CL_2 ={C_1, C_3, C_8}, CL_6 ={C_2, C_{10}}
CL_7 ={C_6}, CL_5 ={C_4, C_5, C_7, C_9}

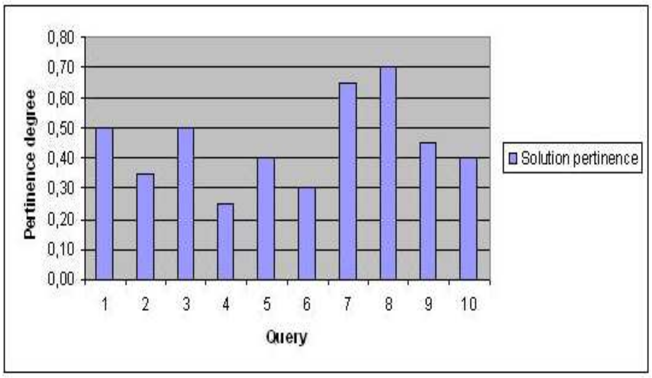


Figure 5: Similarity between retrieved cases and queries

In this example, we assume that the core size is low or equal to 2. Table 3 represents the distance values $dist(C_i^k)$ of each case from its class.

The core building is done by selecting cases having the minimal distances $dist(C_i^k)$ from their class. On each column in table 3, we select two cases having the minimal values. These cases constitute the class core of the corresponding column:
$\mathcal{N}cl_1 = \{C_5, C_9\}$, $\mathcal{N}cl_2 = \{C_1, C_3\}$, $\mathcal{N}cl_3 = \{C_4, C_5\}$
$\mathcal{N}cl_4 = \{C_6, C_{10}\}$, $\mathcal{N}cl_5 = \{C_4, C_5\}$, $\mathcal{N}cl_6 = \{C_2, C_{10}\}$,
$\mathcal{N}cl_7 = \{C_6\}$

	CL_1	CL_2	CL_3	CL_4	CL_5	CL_6	CL_7
C_1	2,70	1,25	X	X	X	X	X
C_2	2,62	X	2,29	1,96	X	0,98	X
C_3	2,73	1,28	X	X	X	X	X
C_4	2,38	X	1,87	X	1,31	X	X
C_5	2,26	X	1,97	X	1,30	X	X
C_6	2,72	X	2,16	1,48	X	X	0,00
C_7	2,45	X	2,03	X	1,36	X	X
C_8	2,85	1,40	X	X	X	X	X
C_9	2,37	X	2,00	X	1,32	X	X
C_{10}	2,77	X	2,47	1,21	X	1,09	X

Table 3: Table representing the distance $dist(C_i^k)$ of each case within its class

The evaluation of our approach is done by computing the similarity of the retrieved case solution to the user query. Figure 5 presents the pertinence result PER_{SC} (presented in section 4) of service classes of the found case $foundCase$ for each query $Query$: $PER_{SC}(foundCase, Query)$.

7. Related work

Web service composition field has been drawing a lot of attention over the last few years. For example, EFlow [9] uses a static workflow generation method. A composite service is modelled by a graph that defines the order of execution among the nodes in the process.

In Meteors [10], authors propose a framework to support advertisement, discovery and the semantic service composition. This approach consists of adding the semantic to the current standards such as UDDI [16], WSDL [16] and BPEL[1]. It allows to capture high level specifications with an abstract process containing abstract services. Templates can be built for abstract services to define their functionality and other attributes. However, the major disadvantage of EFlow and Meteors is the requirement of a predefined workflow.

Other work is based on the artificial intelligence planning. For example, SWORD [10] is a set of tools for the web service composition. In SWORD, a service is represented by a rule. This rule indicates that its preconditions involve its postconditions. SWORD uses Entity-Relation (ER) model to specify a web service. In contrast, SWORD does not use the emerging service description standards such as WSDL [16], SOAP [16], RDF and OWL-S.

There have been a few work done in web service composition using the case based reasoning technique. To our knowledge, the first approach using this technique is proposed by Limthanmaphon and Zhang [13]. They have presented a framework which applies the case based reasoning in the service discovery process. case bases of web services store a set of prior cases. Each case is a composite service and is identified by a service name and a service description. However, the retrieve in this initiative is made by keyword (such as the service name). A recent approach [14] has appeared which use the case based reasoning too. It is based on functional and non functional parameters to retrieve web services. Moreover, cases are actual web services and the huge number of services makes the retrieve process more complex. In contrast, in our approach we build a composition diagram by service classes which improves the retrieval process.

8. Conclusion

In this paper, we have presented our proposal WeSCo-CBR based on case based reasoning. We have opted for this technique to propose an approach to build a composition diagram to a given query. At the first time, we have organized web services into classes. Finding service classes which can contribute to the composition diagram allows to reduce the retrieval process to these classes. Then, we have applied a case classification algorithm within the case base. This classification avoids to browse the whole case base to retrieve the most similar cases to a given query. Once the case class is found, the retrieval process is reduced to this class. Next, we have used the decision tree to browse this classification.

According to the evaluation results, we can deduce the importance of the adaptation process. Besides, some queries need a light adaptation which can be done by the user and others require an important adaptation. As a future work, we aim at enhancing the proposed solution using an adaptation system and defining an heuristic to find the best number of classes that can improve the browsing time.

References

[1] Jordan Diane and Evdemon John. (2007). Web Services Business Process Execution Language Version 2.0. [Online]. Available: http://docs.oasis-open.org/wsbpel/2.0/CS01/wsbpel-v2.0-CS01.html

[2] D. Martin and all. (2006). OWL-S 1.2, OWL-based Ontology for Services. [Online]. Available: Available at http://www.daml.org/services/owl-s/

[3] S. Lajmi, C. Ghedira, K. Ghedira, and D. Benslimane, "WeSCo_CBR: How to compose web services via case based reasoning," *IEEE International Symposium on Service-Oriented Applications, Integration and Collaboration held with the IEEE International Conference on e-Business Engineering (ICEBE 2006).*, pp. 618–622, Oct. 2006.

[4] S. Lajmi, C. Ghedira, and K. Ghedira, "How to apply CBR method in web service composition," *ACM/IEEE International Conference on Signal-Image Technology and Internet-Based Systems (SITIS 2006).*, pp. 230–239, Dec. 2006.

[5] M. Stollberg, C. Feier, D. Roman, and D. Fensel. "*Semantic Web Services - Concepts and Technology.*" Kluwer, 2006.

[6] M. Stollberg, "Reasoning Tasks and Mediation on Choreography and Orchestration in WSMO," *The 2nd International WSMO Implementation Workshop (WIW 2005).*, Jan. 2005.

[7] D. Benlismane, Z. Maamar, and C. Ghedira, "A view-based approach for tracking composite Web Services," *The European Conference on Web Services (ECOWS-05).*, pp. 170–179, Nov. 2005.

[8] R. Aggarwal, K. Verma, J.A. Miller, and W. Milnor, "Constraint Driven Web Service Composition in METEOR-S," *IEEE International Conference on Services Computing (SCC 2004).*, pp. 23–30, Nov. 2004.

[9] B. Limthanmaphon, Y. Zhang, "Web service composition with case-based reasoning," *The 14th Australasian Database Conference.*, pp. 201–208, Feb. 2003.

[10] S.R Ponnekanti, and A. Fox, "Sword: A developer toolkit for web service composition," *The World Wide Web Conference.* May. 2002.

[11] D. LEAKE. "*CBR in Context: The Present and Future.*" MIT Press, 1996.

[12] A. Aamodt , and E. Plaza. "Case-based reasoning: foundational issues, methodological variations, and system approaches," *AI Communications.*, vol. 7, pp. 39–59, May. 1994.

[13] J.L Kolodner. "*Case-based reasoning.*" Morgan Kaufman, San Mateo, CA, 1993.

[14] D. Thakker, T. Osman, and D. Al-Dabass. "Semantic-Driven Matchmaking and Composition of Web Services Using Case-Based Reasoning," *The Fifth European Conference on Web Services.*, pp. 67–76, Nov. 2007.

[15] L. Deborah, and M. Frank van. (2004). OWL Web Ontology Language Overview. [Online]. Available: http://www.w3.org/TR/owl-features/

[16] J.M Chauvet. "*Services Web avec SOAP, WSDL, UDDI, ebXML.*" Première édition. Paris: Eyrolles, 2004.

[17] D. Pi-Sheng. "Using Case-Based Reasoning for Decision Support," *The Twenty-Seventh Annual Hawai International Conference on Systems Sciences.*, pp. 552–561, 1994.

Time Series Predictability Analysis with Modified Dynamic Gray Systems

Junfeng Qu, Hamid R. Arabnia, Yinglei Song, Khaled Rashed, Yong Wei

***Abstract*— Prediction of time series data has been used extensively in engineering, economics, and many other areas, however, more precise models are always sought by scientists. In this research, the predictability of dynamic gray systems is studied on forecasting time series data. Modified dynamic gray system models are also proposed and compared. The criteria such as MAE, MSE, directional accuracy, and the Theil's inequality coefficient are used to analyze the predictability of dynamic gray models with relationship to parameters of models, and compared with random walk model.**

***Index Terms*—Time series, dynamic system, gray model, random walk**

I. Introduction

TIME series data has been studied widely as a case of prediction problem and find ways to mine information from the date. However, these efforts have has less than successful results compared with the random walk rule. Traditional time series analysis models, such as ARIMA model, also known as Box-Jenkins method [1] are limited by the requirement of stationary property of the time series and normality and independence of the residuals.

Osborne (1959)[2] proposed the random walk characteristic of stock market. Later, the random walk model has been widely considered as a statistical model for the movement of the logged stock price. Under such a model, the stock price is not predictable or mean reversing. A time series is a random walk if it satisfies

$P_t = P_{t-1} + a_t$

Where P_t and P_{t-1} are data value at time *t* and *t-1*, and a_t is a white noise.

Artificial intelligence techniques such as artificial neural network (ANN) and genetic algorithms (GA) have been applied to forecast time series data as the computation power increased dramatically[3, 4]. Those approaches are based on the training time data that includes those far away from the present to train the model and thus produce prediction. Thus the data are not fully considered as in the time series because the all data are treated without any preference. Kim and Han[9] also showed that ANNs had some limitations in learning the patterns because some time series data such as financial data has tremendous noise and complex dimensionality and the sheer quantity of financial time series data sometimes interferes with the learning of patterns.

The technical difficulty of the financial forecasting problem is due to low signal-to-noise ratio, non-Gaussian noise distribution, non-stationary, and non-linearly[5]. The other problem with predicting stock prices is that the volume of data is too huge to influence the ability of using information [6]

Due to these controversies and difficulties in the stock market time series forecasting, a different approach is desirable that does not depends on the stationary and Gaussian distribution of the data and owns a certain accuracy and ability to forecast.

The gray system theory is based on the assumption that a system is uncertain and that the information regarding the system is insufficient to build to construct a model to depict the evolution of the system exactly. The gray system was first introduced by Deng[7,8]. The gray predicting model is the essential of the gray system theory and it has been successfully used in geography, hydrology, management, engineering, agriculture, ecology, medicine, and social science because of its computational simplicity and effectiveness. The advantage of the gray predicting system is that only a few discrete data are sufficient to characterize an unknown system that depicted by the first-order differential equation. Thus, the gray predicting system is suitable for predicting the system that historical data is limited and a quick and reliable resolution for the decision-makers reference.

The gray theory avoids the inherent defects of conventional statistic methods such as regressive analysis or traditional time series analysis that requires the certain constrains on the data. It provides an opportunity for the time series data analysis to establish a non-function model with a limited amount of data to estimate the behavior of gray system. In this paper, we use

Junfeng Qu, Department of Information Technology, Clayton State University, Morrow, GA 30260. Email: jqu@clayton.edu

Hamid R. Arabnia is with the University of Georgia, Athens, GA 30602. Email: hra@cs.uga.edu

Yinglei Song is with the University of Maryland Eastern Shore, Princess Anne, MD 21853; email: ysong@umes.edu

Khaled Rashed is with the University of Georgia, Athens, GA 30602. Email: khaled@cs.uga.edu

Yong Wei is with the North Georgia College & State University, email: ywei@northgeorgia.edu

the gray theory to analyzed original gray model and modified gray models with Fourier residual correction and their predictability of time series data while assume that the evolution of time series is a gray system.

The remainder of this paper is organized as follows: Section two describes and discusses the theories and models of the traditional gray model and modified gray models. Section three discussed the evaluation criteria and methodologies used for comparing the performance and predictability of these gray models. The experimental results are presented in the Section four. Section five sets forth the authors' conclusions.

II. DYNAMIC GRAY MODELS

A. *GM11 Model*

The gray system[30, 31] uses a gray predictor to predict the system behavior and feed the predicting information back to the decision-making mechanism to indicate an appropriate control action. There are three basic procedures:

- Accumulated generation operation (AGO)
- Gray modeling(GM)
- Inverse accumulated generation operation (IAGO)

The gray prediction system uses accumulated generation to build differential equations.

The most commonly used is the GM(1,1) , i.e. , a single variable first-order gray model. The algorithm is summarized as follows [31].

Assume that $x^{(0)}$ represents raw time series, and $x^{(1)}$ represents accumulated generated series, and $z^{(1)}$ represents the series obtained from the average of two consecutive data points from $x^{(1)}$; then the GM(1,1) model is written as:

$$x^{(0)}(k)+az^{(1)}(k)=b\,, \qquad (1)$$

where $x^{(0)}(k)$ and $z^{(1)}(k)$ represent the k^{th} element of the series respectively, a and b are model parameters, where a is called the developing coefficient and b is the gray input, that are calculated from given data in order to minimize errors.

In order to solve for *a* and *b* easily, the equation (1) is whitened by a first-order differential equation:

$$\frac{dx^{(1)}}{dt}+ax^{(1)}=b\,, \qquad (2)$$

the solution of this whitened differential function can be easily solved as:

$\hat{x}^{(1)}(k+1)=\frac{b}{a}+(x^{(0)}(1)-\frac{b}{a})e^{-ak}$, and raw sequence can be solved as:

$$\hat{x}^{(0)}(k+1)=x^{(1)}(k+1)-x^{(1)}(k)$$
$$=(x^{(0)}(1)-\frac{b}{a})(1-e^{a})e^{-ak} \qquad (3)$$

In the discussion above, the $x^{(0)}$ is defined as

$$x^{(0)}=\left\{x^{(0)}(1),x^{(0)}(2),...,x^{(0)}(n)\right\} \qquad (4)$$

$x^{(1)}$ is defined as:

$x^{(1)}=\left\{x^{(1)}(1),x^{(1)}(2),...,x^{(1)}(n)\right\}$ and

$$x^{(1)}(k)=\sum_{i=1}^{k}x^{(0)}(i) \qquad (5)$$

and

$$z^{(1)}(k)=\frac{x^{(1)}(k)+x^{(1)}(k+1)}{2}\,,$$

for k = 1, 2,…,n-1 (6)

B. *GM1XN Model*

What has been done so far to improve on the gray system prediction model is to take the first term of $x^{(1)}$ as the initial condition of the first ordinary differential equation, thereby incorporating the error terms to the model, or, alternatively, to optimize the parameters via other methods. In the time series date, if data value is the reflection of information, then new information contained in terms other than the first term of $x^{(1)}$ is more significant; i.e., the recent value $x^{(1)}(n)$ is of greater interest to the researchers. In our newly developed gray model, we chose the n^{th} term of $x^{(1)}$ as the starting condition to solve the gray differential equation. Therefore, the new information is more adequately incorporated into the gray model; as a result, the accuracy of prediction should be improved and residual errors should decrease correspondingly. We have proven our assumption and given the formula to for creating the gray model GM1XN as follows:

The discretized solution of the first-order differential equation

$$\frac{dx^{(1)}(t)}{dt}+ax^{(1)}(t)=b$$

at the initial value at $x^{(1)}(t)\big|_{t=n}=x^{(1)}(n)$ is

$$x^{(1)}(k)=(x^{(1)}(n)-\frac{b}{a})e^{-a(k-n)}+\frac{b}{a}$$

The original raw data can be restored as

$$\hat{x}^{(0)}(k+1)=(x^{(1)}(n)-\frac{b}{a})(e^{-a}-1)e^{-a(k-n)}$$

C. *GM11 and GM1XN Models with Fourier Feedback*

Let the residual time series R_r be defined as

$$R_r=\{R_r(2),R_r(3),...,R_r(n)\}^T$$

where

$R_r(k) = x(k) - \hat{x}(k)$, for k = 2, 3,…,n

Assume that residual time series is described by discrete Fourier series as:

$$R_r(k) = a_0/2 + \sum_{i=i}^{k_a} \left[a_i \cos(2\pi ik/T) + b_i \sin(2\pi ki/T) \right]$$

where $T = n - 1$, and $k_a = \lfloor (n-1)/a \rfloor - 1$. The parameters a_0, a_i and b_i for $i=1,2,\ldots,k_a$ can be estimated by the least-squares method. The coefficients a_0, a_i and b_i for $i=1, 2, \ldots k_a$ can be estimated as $C = (M^T M)^{-1} M^T E_r$, where

$C = \left[a_0, a_1, b_1, a_2, b_2 \cdots a_{k_a}, b_{k_a} \right]^T$ and M

$$M = \begin{bmatrix} \frac{1}{2} & \cos\left(\frac{2\pi\times2}{T}\right) & \sin\left(\frac{2\pi\times2}{T}\right) & \cos\left(\frac{2\pi\times2\times2}{T}\right) & \sin\left(\frac{2\pi\times2\times2}{T}\right) & \ldots & \cos\left(\frac{2\pi\times2\times k_a}{T}\right) & \sin\left(\frac{2\pi\times2\times k_a}{T}\right) \\ \frac{1}{2} & \cos\left(\frac{2\pi\times3}{T}\right) & \sin\left(\frac{2\pi\times3}{T}\right) & \cos\left(\frac{2\pi\times3\times2}{T}\right) & \sin\left(\frac{2\pi\times3\times2}{T}\right) & \ldots & \cos\left(\frac{2\pi\times3\times k_a}{T}\right) & \sin\left(\frac{2\pi\times3\times k_a}{T}\right) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{2} & \cos\left(\frac{2\pi\times n}{T}\right) & \sin\left(\frac{2\pi\times n}{T}\right) & \cos\left(\frac{2\pi\times n\times2}{T}\right) & \sin\left(\frac{2\pi\times n\times2}{T}\right) & \ldots & \cos\left(\frac{2\pi\times n\times k_a}{T}\right) & \sin\left(\frac{2\pi\times n\times k_a}{T}\right) \end{bmatrix}$$

The new Fourier-corrected residual gray model can be rewritten as:

$$\hat{x}^{(0)}(k+1) = \hat{x}^{(0)}(k+1) + R_r(k+1) \text{ for } k=2,3,\ldots,n,..$$

with the initial condition being $\hat{x}(1) = x^{(0)}(1)$.

In the algorithm, only a small amount of data is needed to estimate the only two parameters in the model for prediction. The workflow of the gray modeling procedure is shown in figure 1.

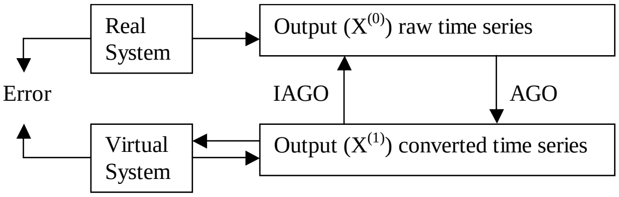

Real System
Output (X$^{(0)}$) raw time series
Error
IAGO
AGO
Virtual System
Output (X$^{(1)}$) converted time series

Figure 1. Workflow of Gray Modeling

III. Evaluation and Experiments

A. Evaluation Measures

There are many ways to evaluate the forecasting performance of a model, ranging from directional measures to magnitude measures to distributional measures. The selection of evaluation criteria for forecasting techniques was conducted by Yokum and Armstrong [9]. Accuracy was their most important criterion, followed by the cost savings generated from improved decisions. In particular, execution issues such as ease of both interpretation and use were also highly rated.

Three measures of magnitude were incorporated into the research. The first measure is the mean square error (MSE), which measures the overall performance of a model. The calculation for MSE is

$$MSE = \frac{1}{n} \sum_{k=1}^{n} [x(k) - \hat{x}(k)]^2$$

where $\hat{x}(k)$ is the predicted value for time k. $x(k)$ is the actual value at time k, and n is the number of data used for prediction.

The second measure is the mean absolute error (MAE). It is a measure of the average error for all points and is computed as

$$MAE = \frac{1}{n} \sum_{k=1}^{n} \left| x(k) - \hat{x}(k) \right| \quad (14)$$

where the meaning of the $\hat{x}(k)$ and $x(k)$ is as MSE.

If we express the MAE into percentile format, we call it Mean Absolute Percentage Error (MAPE).

$$MAPE = \frac{1}{n} \sum_{k=1}^{n} \left| \frac{x(k) - \hat{x}(k)}{x(k)} \right| \quad (15)$$

The definition of $\hat{x}(k)$ and $x(k)$ is the same as that of MSE. MAPE is the more objective statistic indicator because the measure is in relative percentage and will not be affected by the unit of the forecasting series. The closer MAPE approaches zero, the better the forecasting results. According to Lewis [10], the performance of the model is categorized in Table 1:

Table 1. Performance measure by MAPE

MAPE (%)	Performance
<10	High precision forecast
10-20	Good forecast
20-50	Reasonable forecast
>50	Imprecise forecast

Although these three measures are fairly accurate for deriving the deviations of the predicted values from the actual values, they do not provide much information about the power of the models in predicting the turning points or direction of the predicted value. For many applications of time series analysis, the direction is as important as the magnitude; e.g., in the financial market, traders and analysts market direction and turning points besides the value of predictability. In these markets, money can be made simply by knowing the direction in which the time series moves. A correct directional prediction requires that

$$sign(\hat{x}(k+1) - x(k)) = sign(x(k+1) - x(k))$$

where $x(k)$ is the time series data point at time k, $x(k+1)$ is the time series data point at time $k+1$, and $\hat{x}(k+1)$ is the estimated time series data point at time $k+1$. Therefore, the direction accuracy (DA) is computed as

$$DA = \frac{1}{n} \sum_{i=1}^{n} a_i$$

where

$$a_i = \begin{cases} 1, if\ (x(k+1)-x(k))(\hat{x}(k+1)-x(k)) > 0 \\ 0, otherwise \end{cases}$$

This means that a_i takes the value 1 if the actual change and the predicted change have the same sign and 0 if they have opposite signs. If $(S_{i+1}-S_i)(\hat{S}_{i+1}-S_i) > 0$ for all *i*, then the value of DA will be 1, implying that the model predicts the accurate change on all occasions.

Theil's inequality coefficient [11] measures the forecasting power of a model relative to the random walk model. This relationship is calculated by dividing the RMSE of the model by the RMSE of the random walk model. Hence,

$$U = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(S_{i+1}-\hat{S}_{i+1})^2}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(S_{i+1}-S_i)^2}}$$

The denominator is calculated from the actual change in the exchange rate between *i* and *i+1*. The numerator is derived from the difference between the actual change and the predicted change, which results in

$$(S_{i+1}-S_i)-(\hat{S}_{i+1}-S_i) = (S_{i+1}-\hat{S}_{i+1})$$

where $(S_{i+1}-S_i)$ is the actual change and $(\hat{S}_{i+1}-S_i)$ is the predicted change. Another version of Theil's inequality coefficient is based on relative changes. Hence it is calculated as:

$$U = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\frac{S_{i+1}-\hat{S}_{i+1}}{S_i})^2}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\frac{S_{i+1}-S_i}{S_i})^2}}$$

The implications of the numerical values of U are summarized in table 2.

Table 2. Theil's Inequality Coefficient range

Value	**Implication**
U=0	The model produces perfect forecasts
0<U<1	The model provides less than perfect forecasts but out performs the random walk model
U=1	The model is as good as the random walk model

B. *Experiments with Mackey-Glass time series data*

One of the most commonly used chaotic time series generation function, the Mackey-Glass chaotic time series, is defined by the following delay-differential equation [13].

$$\frac{dx(t)}{dt} = \frac{0.2x(t-\tau)}{1+x^{10}(t-\tau)} - 0.1x(t)$$

where τ is an adjustable delay term.

This delay differential equation can be extremely chaotic and display a wide variety of behaviors because its value at any time may depend on its entire previous history. Farmer and Sidorowich [14] presented a forecasting technique for chaotic time series. They introduced nonlinear mapping using a local approximation after embedding a time series in a state space using delay coordinates. In order to evaluate the ability of the algorithm to identify structural change of chaotic time series that periodicity is not apparent or might not exist at all, a Mackey-Glass chaotic time series is used with *τ =30* in the simulated evaluation.

The following experiments are conducted based on the Mackey-Glass chaotic time series to analyze the forecast precision of these four dynamic gray models.

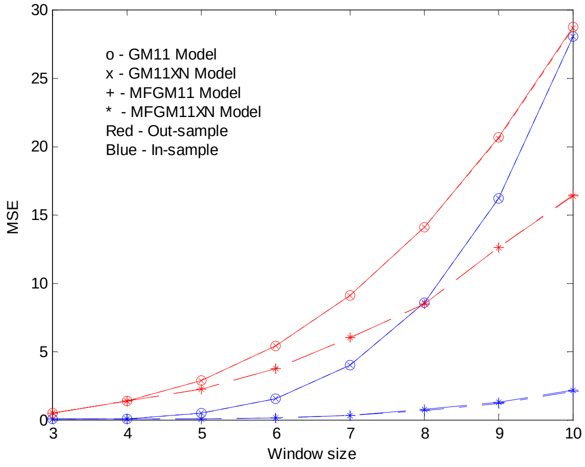


Figure 2. Mackey-Glass MSE vs. Window Size

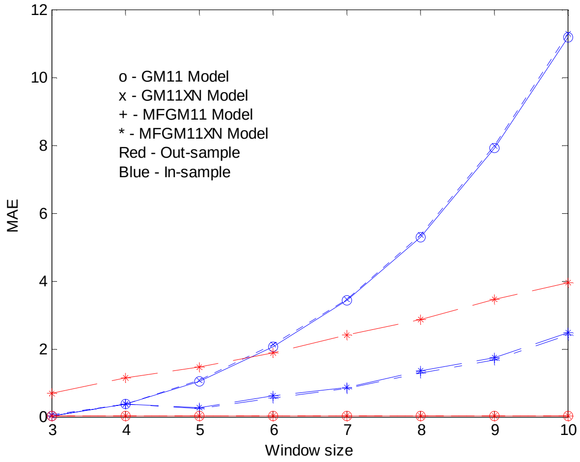


Figure 3 Mackey-Glass MAE vs. Window Size

Figure 2 shows that the MSE increases as window size used by models increases; i.e., the more historical data used to build the model, the higher the MSE will be. The result confirms the research [12] that the short-term memory is around 7±2. Fourier Modified Gray Models tend to have a smaller MSE than these of non-Fourier error corrections. Also, MSE from out-sample is higher than that of in-sample. The latest data values used as the initial condition to solve differential equation do not seem to improve the MSE. The bigger the window size used to build the model, the worse is the model based on MSE.

As shown in figure 3, the MAE changes with the change of window size used by the models. It is clear that our sample MAE is smaller than that of an in-sample. The Fourier-corrected error models decrease MAE compared to non-Fourier modified errors on in-samples. The out-sample MAEs are smaller than those of the in-sample.

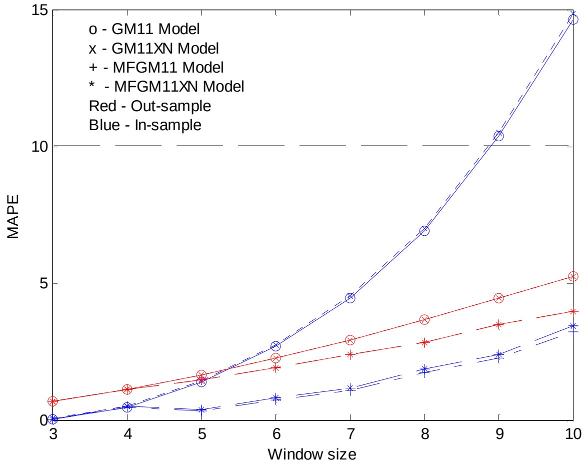


Figure 4. Mackey-Glass MAPE vs. Window Size

The change of MAPE against the windows size of the model is shown in figure 4. Models with MAPE values of less than ten can be classified as high-precision models. We can see that when the window size is less than nine, all these four models can be categorized as high-precision models. In contrast, models modified with Fourier transform have lower MAPE values than those of non-Fourier modified models in the in-sample comparison. However, the out-sample comparisons reveal that all four models achieve similar MAPE values. As window size increases, the MAPE value increases also.

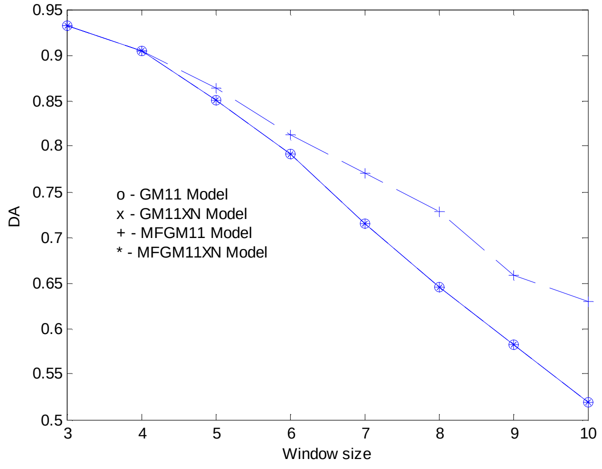


Figure 5. Mackey-Glass DA vs. Window Size

Figure 5 shows that the increasing of window size used by the models tends to decrease the model's ability to forecast directional moves. Fourier modified models do not change the DA significantly, and usage X_n as initial conditional value does not cause major improvement. It is interesting to see the Fourier-modified gray model produce better DA predictability as window size increases.

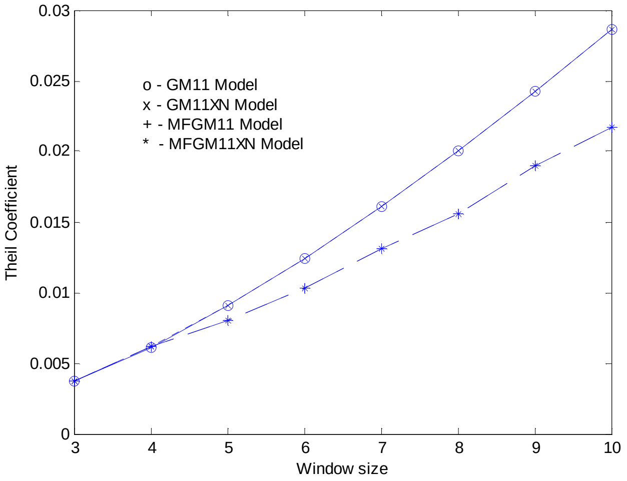


Figure 6 Mackey-Glass Theil's Inequality Coefficient vs. Window Size

Figure 6 illustrates the change of Theil's Inequality Coefficient with the change of window size. Clearly, the increasing of window size also increases the Theil's Inequality Coefficient. Fourier modification of models decreases the Theil's Inequality Coefficients, while initial condition change does not affect the Coefficient very much.

IV. Conclusion

A smaller window size is better suited to building a high precision forecast model. All of four models exhibit this discovery based on MAPE. GM1XN outperforms GM11 both in-sample and out-sample prediction. With Fourier residuals correction of gray model, MAPE decreased about 50% and shows consistently high precision predictability. A smaller windows size also generates a better directional forecast accuracy on all of four models especially when windows size is less than five, the directional accuracy is greater than 85%. In our experiments, all of these models have more than 50% directional forecast rate. This is also confirmed from the Theil's Inequality Coefficient indicator, the smaller windows size builds a model that outperforms a random walk model better. It also indicated that Fourier residual corrected gray model outperformed non-corrected models too. Overall, all four models perform very well, and are categorized into high-precision models based on these measures outperform the random walk model.

REFERENCES

[1] G.E.P. Box, G.N. Jenking, and G.C. Reinsel, Time series analysis: forecasting and control, 3rd ed. Englewood Cliffs, N.J., Prentice Hall, 1994

[2] M. Osborne, “Brownian Motion in the Stock Market” Operations Research, vol 7, pp 707-712, 2003

[3] E. Kalyvas, “Using Neural Networks and Genetic Algorithms to Predict Stock Market Returns”, University of Manchester, 2001.

[4] R. J. Frank, N. Davey, and S.p. Hunt, “Time Series Prediction and Neural Networks” Journal of Intelligent and Robotic Systems, pp.91-103, 2000.

[5] B. Kovalerchuk and E. Vityaev, Data Mining in Finance: advances in relational and hybrid methods, Kluwer Academic Publisher, 2000.

[6] Ul Fayyad, G. paatesky-Shpiro, and P. Smith, “Process for extracting useful knowledge from volumes of data” ACM Comm, vol. 39, pp. 27-34, 1996

[7] J. Deng, Control Problems of Gray Sysems, Systems & Control Letters, Vol. 1., pp. 288-294, 1982

[8] J. Deng, Introduction to Grey System Theory, The Journal o fGrey System, Vol. 1, pp 1-24, 1989

[9] J. T. Yokum and J.S. Amrstrong, Beyond accuracy: Comparison of criteria used to select forecasting methods, International Journal of Forecasting, pp. 591-599, 1995

[10] C. D. Lewis, Industrial and Business Forecasting Method. London: Butterworth Scientific, 1982.

[11] H. Theil, Economic Forecasts and Policy. Amsterdam: North-Holland Publishing Company, 1961

[12] D. Wang, "Temporal Pattern Processing," in The Handbook of Brain Theory and Neural Networks, 2nd. 1163-1167: MIT Press, Cambridge, MA, 2003.

[13] A. Lapedes and R. Farber, "Nonlinear signal processing using neural networks: prediction and system modelling," Report LA-UR-87-2662, 1987.

[14] J. D. Farmer and J. J. Sidorowich, "Predicting Chaotic Time Series," Physical Review Letters, vol. 59, pp. 845-848, 1987.

CUDA BASED MULTI OBJECTIVE PARALLEL GENETIC ALGORITHMS:

ADAPTING EVOLUTIONARY ALGORITHMS FOR DOCUMENT SEARCHES

Jason P. Duran, Sathish AP Kumar

Computer Science and Information Systems Department

National University, San Diego, CA 92123

Email: Jason_Duran@acm.org, sathish.ap@gmail.com

ABSTRACT:

This paper introduces a Multi Objective Parallel Genetic Algorithm (MOPGA) using the Compute Unified Device Architecture (CUDA) hardware for parallel processing. The algorithm demonstrates significant speed gains using affordable, scalable and commercially available hardware. The algorithm implements a document search using techniques such as Term Frequency Inverse Document Frequency (TF-IDF), Latent Semantic Analysis (LSA), Multi Objective Algorithms (MOA), Genetic Algorithm (GA), and Quad Tree Pareto Dominance techniques.

The objective of the proposed algorithm is to assemble an adaptable and scalable search mechanism to efficiently retrieve highly relevant document for a given search query. TFIDF and LSA vector space searches are two of the more common approaches to text mining. We have demonstrated that by combining results from both operations the number and quality of results could be improved. Evolutionary algorithms, specifically Genetic Algorithms have long been used to efficiently optimize multi-objective problems and so provide a natural starting point for our approach.

Keywords: Multi-Objective Parallel Genetic Algorithms, CUDA, TF-IDF, LSA, Pareto Quad Tree, Text Mining.

1 INTRODUCTION

The proposed algorithm searches through multiple documents looking for relevant matches to a given search string. The algorithm begins by converting the search string to a query vector in the TF-IDF, and LSA document search spaces. Chromosomes in this algorithm are composed of ten alleles that are each a direct encoding of a term. Using TFIDF domain knowledge a heap is constructed for each search term and any document that the term is relevant to. The algorithm exploits this domain knowledge to select unevaluated documents that have the strongest relation to the term. Iterative generations benefit from the principle of survival of the fittest in an attempt to discover the documents that are most closely related to the search query. The algorithm allocates parallel processes to CUDA enabled graphics devices to distribute the work across multiple processors, dramatically reducing the processes run time. Tabularized results for various search keys are presented along with corresponding execution times.

Multi-Objective Algorithms require ranking systems to properly evaluate tradeoffs between the search domains. Pareto Dominance ranking provides the ability to objectively analyze tradeoffs between both the different domain results that originate from different input. Genetic Algorithms represent an efficient heuristic search of some data set that can often retrieve near optimal results in fewer operations when compared to linear ranking methods. To further improve performance problems associated with the TFIDF and LSA search space models, the algorithm has been adapted to utilize massively parallel processors. The final result is a Multi-Objective Parallel Genetic (MOPGA) Algorithm that utilizes TFIDF, and LSA vector searches, implemented on the Compute Unified Device Architecture (CUDA) framework.

1.1 ALGORITHM ORGANIZATION

This algorithm runs on both the Linux PC host and on available CUDA devices. On the PC Side, some initial loading and search parameters are done in parallel where possible. This preloaded data is stored in a Read-Only database called data space. Before the search begins, data space is divided amongst the MOPGA Agents that are to be run. While running each MOPGA Agent has its own, database local PKnown where Pareto Quad Tree and final document evaluations are stored. Mathematical operations take advantage of CUDA acceleration where possible. Finally, once a predetermined time or a number of cycles have elapsed the results are merged and sorted for display.

2.0 Domain Search Spaces

This implementation of MOPGA uses two vector search spaces for its Domain. The relevance of a given search vector is the angular difference between a search vector and a document vector. The two-domain spaces are “TFIDF” and “LSA” . The actual data that composes the matrices are book synopsis taken from user comments on Amazon.com. The mathematical explanations are as follows.

2.1.0 Term Frequency - Inverse Document Frequency (TF-IDF)

Term Frequency – Inverse Document Frequency provides the MOPGA a statistical method of determining how important a term is within a document with respect to a collection of documents. The MOPGA constructs a matrix where rows represent terms, columns represent documents, and individual cells record the number of occurrences of a given term in a document. Within this matrix often repeated terms are considered important terms in the document. This is not entirely true, as some words with such highly repeated frequency do not convey uniquely searchable concepts. Words such as “The” are an example. It will appear with high frequency in all documents and as such, it is not a relevant search term. Extending this concept if we have a set of documents all on a single subject such as databases, because all documents in the collection contain this word, a search for documents with that term would return the whole collection. For this reason, TF-IDF does not consider it a relevant search term. TF-IDF itself is a combination of two ratios, Term Frequency (TF), and Inverse Document Frequency (IDF).

2.1.1 Term Frequency (TF)

Term Frequency is the statistical process of determining how important a term is in the context of a document. It does this by counting the occurrences of a particular term in the document, which divided by the overall count of terms in the document, which establishes a ratio of the terms statistical importance to the document [1].

2.1.2 Inverse Document Frequency (IDF)

Inverse Document Frequency is used to determine how unique a given term is within a collection of documents. By taking the ratio of the total number of document with respect to the number of documents a term appears in, the relative importance of a term as a unique identifier can be determined. By taking the log of this ratio, a dampened value that is suitable for combination with TF is derived [1].

2.1.3 Weighted Search Vectors

The MOPGA searches for documents in the TF-IDF search space by creating a query vector. A query vector is similar to a standard document vector in the search space, however, the TF portion is calculated slightly different. The MOPGA replaces the raw frequency of the term within the query with a user provided bias weight. In this manner, the MOPGA can create bias in favor of a particular term within a query. The MOPGA scores query vector to document vector comparisons by performing the following mathematical operations:

$$tf_{i,j} = \frac{|n_{i,j}|}{\sum_k n_{k,j}} \qquad (1)$$

$$idf_i = \log_2 \frac{|D|}{|\{j : t_i \in d_j\}|} \qquad (2)$$

$$tfidf_{i,j} = tf_{i,j} \times idf_i \qquad (3)$$

$$score = sim_{d,q} = \frac{D \cdot Q}{|D||Q|} \qquad (4)$$

2.1.4 Advantages and Limitations

TF-IDF search vectoring will never assign scores to documents that the exact search terms do not appear in, regardless of how informative a particular term may be. TFIDF does not connect words with their synonyms [2]. Further TF-IDF does not equate words to their conjugated forms, as an example a search for database would be unrelated to a search for databases. This lack of knowledge to make these connections makes the requirement of a strong lexical analyzer to process the information highly desirable, adding to the complexity of the task. The main advantage TFIDF gives is that it is simple to compute and easily parallelized.

2.2.0 Latent Semantic Analysis

Latent Semantic Analysis provides the MPOGA with the ability to find documents based on the relationship of words as they appear in context to one another. The MOPGA LSA function utilizes the same matrix of term relations to documents that are constructed for TF-IDF vector searches [3]. The MOPGA LSA function decomposes these relationships into different matrices that represent relations of terms and documents. Finally, the MOPGA LSA function filters noise in these relations to find better matches. These filtered relations do a good job at finding synonyms and other related words. As an example if a user was searching for "dog training" it would recognize that terms like leash, pet, and treat are related and possibly return documents with these terms in them. The individual steps used in the MOPGA LSA functions are presented next.

2.2.1 SVD decomposition and K Reduction

LSA depends on constructing a term by document rectangular matrix M x N, where cell values contain weighted TF-IDF values. This matrix is then decomposed via a Singular Value Decomposition into three different matrices where A=USVT. Where U is decomposed into an M x N matrix and S is an N x N diagonal matrix containing the ranked singular values, and V is an N x N matrix. Both U and V are unitary so that

$$U \cdot UT = I \qquad (5)$$

$$V \cdot VT = I \qquad (6)$$

K reduction is performed by selecting a K x K portion of the S matrix, M x K of the U matrix, and M x K of the V matrix such that [4].

$$A_k = U_k S_k V_k^t \qquad (7)$$

2.2.2 Pseudo Documents

With these, K reduced matrices the MOPGA LSA function can compare two terms to one another, or documents to documents. To compare a chromosome to a query the MOPGA LSA function first projects the query into the K dimensional space and treats it as if it were any other document in the collection. Because chromosomes themselves are M x 1 matrix representations of their contained terms they are easily converted into these pseudo documents. The MOPGA LSA function accomplished these projections into the K reduced space via the equation [4].

$$D_k = QtU_k S_k^{-1} \quad (8)$$

2.2.3 Similarity

Given the resulting two pseudo documents, the MOPGA LSA function then compares the level of similarity between them by examining the angles between their vector representations with the following well-known equation

$$score = sim_{d,q} = \frac{D \cdot Q}{|D||Q|} \qquad (4)$$

This comparative process gives a range between 1 and -1 where 1 would be a very similar document and -1 as dissimilar as possible [2].

2.2.4 Advantages and Limitations

"Folding in" pseudo documents in this manner does not affect the underlying relationships, a new SVD decomposition is required to incorporate new relations into the process. LSA does not perform well with words that have different meanings based on their contextual usage. LSA is also conducted under the assumption that words and documents follow a Gaussian distribution when it is possible that a Poisson distribution exists [8]. Despite these limitations, the MOPGA LSA function is able to perform the SVD decompositions on the same underlying data that is constructed for the TF-IDF vector search. The ability to recognize relations of the terms to other terms is also crucial to returning relevant information. The actual generation of pseudo documents by the MOPGA requires only a few matrix multiplications once the SVD decompositions are complete.

3 Multi Objective Algorithms

A Multi Objective algorithm seeks to maximize/minimize two or more distinct functions. As is often the case one set of input may increase the desirable results from one function while degrading the results from the other functions. To accomplish the goal of returning relevant information quickly the MOPGA must know how to balance trade offs between one resulting function score with respect to the other function. To accomplish this task the MOPGA uses Pareto Dominance [5].

3.1 Quad Tree Pareto Dominance Based Ranking

A Quad Tree is a tree based data structure that stores different chromosomes. Each chromosomal node is a vector with two elements that reflect the TFIDF and LSA scores. All nodes residing in the tree are non-dominated. This is accomplished by assigning scores to the TFIDF and LSA values of particular candidate. A 0 indicates that this chromosomes score is worse than the comparison node. A score of 1 indicates that the candidate's score is superior to the comparison's node. Each candidate thus scores one of four possible values. 0/0 indicates that it is completely dominated by the comparison node and thus subsequently discarded. A 0/1 or 1/0 indicates that one of the search values scores better than the comparison node and thus is not considered dominated by it. In this case if a child node exists further comparisons will be evaluated against that node. If there are no child nodes then the candidate is inserted into the tree at the appropriate position. In the final case 1/1 indicates that the candidate dominates the existing node. In this case the candidate will take the place of the existing node, and all child nodes of the existing node to be discarded will be evaluated for insertion. The Quad tree is thus guaranteed to contain only non-dominated nodes, and new candidates can be efficiently compared for insertion [6].

4 GENETIC ALGORITHMS

GAs has been successfully applied to a wide variety of problems. In particular, GAs excels at optimization problems where optimal solution execution times are exponentially dependent on the size of the data search space. GA's encode, directly or indirectly different input values for a chromosome. These chromosomes undergo genetic processes where the current best-known chromosomes contained in PKnown are used to generate new chromosomes, PCurrent. The best of these new PCurrent chromosomes replace less fit chromosomes in PKnown. Over many generations, chromosomes converge to optimal solutions [7]. MOPGAs offer a wide number of solutions to overcome problems associated with premature convergence.

4.1 ENCODING DOCUMENTS

Encoding is the process of mapping a problem and its possible solutions to a chromosome. These chromosomes can be a direct encoding of values, it can represent different states of a state machine, or even an order arrangement of items. Because of the number of possible terms that can be included in a relatively small set of documents it is not possible to represent all terms in a single chromosome. The MOPGA presented here uses a direct encoding scheme to limit the number of alleles in a chromosome to ten. Each individual allele is a unique term identifier that is assigned to terms as they are read into the dictionary.

The dictionary is an M x N matrix, where M is the number of searchable terms in all documents. The actual position number of a term in this dictionary is the real encoded value. As each document is read into, the dictionary the number of time a term appears in the document is recorded in the matrix. Once all documents are read into the dictionary the matrix is used to generate TFIDF values. This domain information is exploited later to intelligently select strong candidates for evaluation. This is accomplished by constructing a series of term heaps for each MOPGA Agent that include only documents within an agents assigned search space.

4.2 Selection

Each generation of the MOPGA generates a predefined number of chromosomes called PCurrent to evaluate for insertion into the Quad tree based set called PKnown. To generate the PCurrent set two PKnown nodes are selected for crossover, which generates four new chromosomes. The selection phase is the process of randomly picking two parents for the crossover phase. In terms of this process, the selected parent nodes represent the local space that will be further explored.

4.3 SINGLE POINT CROSSOVER

The process by which the MOPGA generates new chromosomes from the Pknown set is called Single Point Crossover. Using two chromosomes A, and B from the Pknown set both chromosomes are split in halves resulting in A1,A2 and B1,B2. The second halves are then exchanged resulting in A1 B2, A1 B1, A2 B1, A2 B2, B1 A1, B1 A2, and B2 A1, B2 A2. The order of the alleles in the resulting chromosomes affects the local search space that is evaluated during the subsequent evaluation phase.

4.4 RANDOM MUTATION

One drawback to this type of search occurs when the MOPGA continually selects the same parents to generate the same results. When the MOPGA finds a series of solutions that are strong candidates many very similar results begin to be found in PKnown. From an evolutionary standpoint, these chromosomes represent the fit possibilities found so far, but they may in fact represent only local optima.

To force the MOPGA to explore chromosome that cannot be formed by the combination of PCurrent solutions, a pre-known percentage is used to determine if a random mutation occurs in the PCurrent chromosomes generated by the MOPGA. A mutation is accomplished by selecting a random allele and then generating a random term value to encode. This mutation is similar to annealing concepts used in other algorithms. It forces the algorithm to explore other possible locations in the search space [8].

4.4 Evaluation

Evaluation is the process the MOPGA uses to determine if a given chromosome is fit for insertions into Pknown and ultimately presented to the user. The MOPGA selects a document and generates TFIDF and LSA vector similarity scores by comparing the appropriate document vectors to the search vectors. These scores are then used by the Pareto quad tree ranking mechanism to determine if the chromosome should be retained in the PKnown set.

The nature of a document search implies it is a finite search of available documents that are contained in a library. The MOPGA conforms to this constraint by selecting an allele in a chromosome and then using the term heaps to select a document. This process ensures that processing time can never exceed the processing time of a straight linear search of the same library. As documents are evaluated, they are inserted into a hash map that stores the evaluations and enable the evaluation to rapidly determine if an evaluation has already been performed against the selected document.

5 Parallelization

Many of the mathematical operations involved in TF-IDF and LSA lend themselves to parallelization, and in particular to CUDA based implementations [9]. When the algorithm first begins computing TF-IDF scores, it is beneficial to use reduction techniques when summing term counts. When computing a singular value decomposition there are several parallel techniques to apply. Beyond the mathematical applications, there is the process of exploring a Pareto front. By dividing the data space among the MOPGA Agents, different parts of the whole can be run in different threads. This organization conforms too many of the common parallel patterns in use today [10].

5.1 Databases

This MOPGA implementation uses two types of databases, data space and PKnown. Data space contains all of the “static” data that does not change while a search is underway. It contains all of the TFIDF and LSA domain information as well as document excerpts and indexed terms lists. This data space also contains a series of heaps for each MOPGA Agent that is scheduled to run. These heaps represent the assigned search area each MOPGA Agent will concern itself with. These heaps are copied into each MOPGA Agents PKnow database. There is only one copy of data space and it is read only once a search begins.

The second type of database is called PKnown. Each MOPGA Agent that is tasked to run has its own PKnown database that initially contains its assigned term heaps and an empty Pareto quad tree. As chromosomes are evaluated, they are translated into a document via the term heaps and the results are stored. Using these evaluation results a chromosome is then considered for insertion into the Pareto quad tree.

5.2 MOPGA Agent

The MOPGA Agent is tasked with attempting to find non-dominated chromosomes to insert into its own Pknown. When there are insufficient chromosomes in the Pareto quad tree the MOPGA Agent will generate random chromosome to help explore the Pareto front. After two chromosomes are generated or selected, single point crossover and mutations are finished a random allele is selected. Using the term heaps assigned to this agent a document in its assigned sections of data space is then evaluated. In the case that a chromosome has an allele that has no more documents in its heap then the allele is mutated and a different existing allele is selected. If the resulting document evaluation is non-dominated then the chromosome is inserted into Pknown for the MOPGA Agent to exploit.

5.3 Limitations

While the CUDA devices have the ability to greatly speed up some of the more complex mathematical operations, there are only a limited number of these devices on a given machine. This limitation means that access to the devices has to be shared with semaphore type locks. This algorithm has the ability to find the Maximum optima for a given search there is no guarantee that it will be found if the search constrained by time or cycles.

6 Algorithms

6.1 Mopga Algorithm (Steps 2 – 5 in are in parallel and agents run in parallel)

1. Read in Data, build term dictionary and document library matrices.
2. Perform TFIDF / LSA
3. Divide Search Spaces
4. Build term heaps per MOPGA Agent
5. Gather search terms
6. Launch MOPGA agents.
7. Wait for time, cycle, or generation end condition
8. Present results from agent PKnown.

6.2 MOPGA AGENT

1. Select parent chromosomes from Pknown or generate random chromosome if insufficient chromosomes.
2. Perform single point crossover.
3. Mutate chromosomes
4. Determine document – term heaps.
5. Evaluate – perform TFIDF/LSA.
6. Insert chromosomes into Pknown.

7 EXPERIMENTS

A series of experiments were conducted to demonstrate that combined TFIDF/LSA ranking returns both more and superior results than using either of the technique alone would generate. Experimentations also includes execution times of the different techniques. All experiments were conducted on a intel i7-970 and 12 GB of ram and 4 cpu cores. Additionally the test machine has two GTX480 Nvidia video cards, each having 1.5 GB of memory and 480 CUDA cores. The code was written in C++ and C for CUDA version 2.5. Mathematical libraries provided by NVIDA and EM Photonics CULA were used were possible. Serial tests relied on TNT JAMA linear algebra libraries and templates.

7.2 EXPERIMENT 1

The first experiment was conducted to gather data that could be used to determine if combining TFIDF and LSA would yield better results. The MOPGA algorithm was allowed to iterate over all data as well as by cyclic restriction. The data set size was also altered. LSA K reductions were held constant at 100 for both the MOPGA and the LSA serial version. It should be noted that the TNT JAMA linear algebra libraries could not process a matrix larger than 10,000 X 300 so the final two runs have no results recorded.

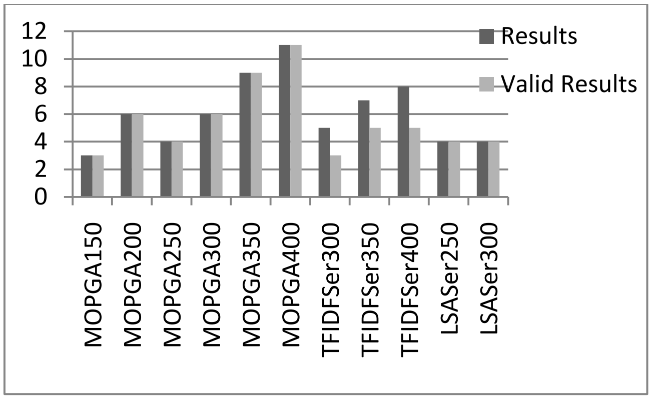


Fig 1.

Figure 1 shown here demonstrates that both the LSA and the MOPGA algorithm returned only highly relevant data, while TFIDF seemed to return some items that were not truly related to the search but contained a term in the search vector. Both the MOPGA algorithm and the TFIDF returned more results than did the LSA. The MOPGA algorithm did return the most relevant documents.

7.2 EXPERIMENT 2

The second experiment is very similar to the first experiment, except that execution times recorded in an attempt to show how well the proposed MOPGA algorithm performed against the other serial implementations. The time shown in these charts is both the initial time spent processing a data set before a query and the search time spent finding the results. It should be noted that while the data set size was increased from 8860 X 150 to 11882 X 400 for the two serial versions, only the larger dataset was used for the recorded MOPGA time shown here.

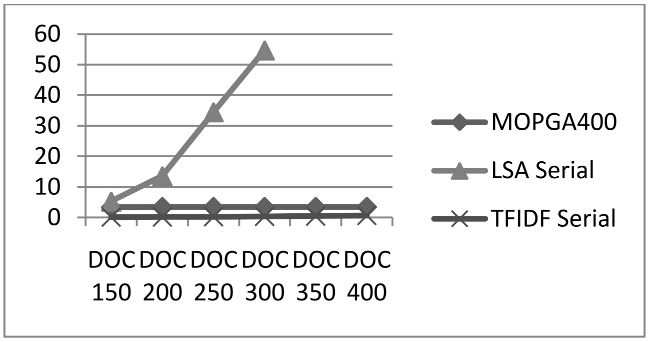


Fig 2.

Figure 2 shows that as data is added, the serial linear algebra libraries in the LSA Algorithm scales very poorly and even fails to execute with matrices that exceed some 10,000

X 300 cells. The MOPGA algorithm also has some initial load time due to the SVD decomposition in LSA but scales more like the TFIDF algorithm as the number of cycles increases.

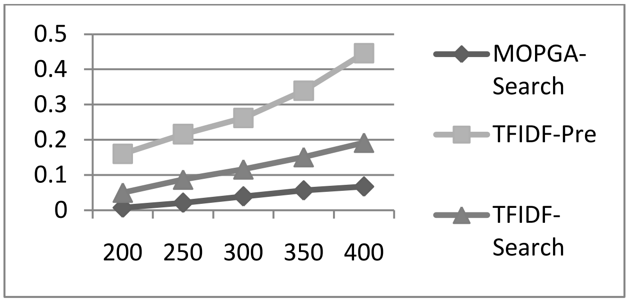


Fig 3.

Taking into consideration the total processing time and breaking up the processing times into pre-search and vector similarity computations demonstrate that LSA pre-compute times are several orders of magnitude greater than the search times so those values do not appear..

8 Conclusion

By combining TFIDF and LSA the MOPGA algorithm demonstrates that both the relevance and number of results returned to a user submitted search are improved and the scaling implications are much more favorable for the MOPGA algorithm in comparison to the LAS and TFIDF. The greatly reduced SVD decomposition time means that if necessary MOPGA could be re-indexed frequently lending itself to much more volatile databases. The MOPGA shows promise with much larger datasets where iterating over all possible results may not be possible. The Pareto Quad Tree is easily adapted to include other search domains lending itself to further refinement and experimentation.

Throughout the testing of this algorithm a constant K dimensional reduction of 100 was applied to the results. Many experiments show that altering this value can have a great effect on the quality and quantity of LSA searched. Some experimentation should also be conducted with the frequency with which genetic mutations are introduced for evaluation. Currently this is a constant 5% and this may not be optimal.

When the initial data set size grows beyond what fits on a single CUDA device, the data dpace database will need to be split before genetic operations are conducted. MOPGA Agents currently deal with an assigned piece of the data space database so further scaling experiments should be conducted to determine what the scaling limits are.

References

[1] Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Term frequency and weighting*. from http://nlp.stanford.edu/IR-book/html/htmledition/tf-idf-weighting-1.html Retrieved March 23, 2010.

[2] Ramos, J. *Using TF-IDF to Determine Word Relevance in Document Queries.* Piscataway, NJ: Rutgers University, 2001.

[3] Landauer, T., Foltz, P., & Laham, D. An Introduction to Latent Semantic Analysis. *Discourse Processes , 25* (2 & 3), pp. 259 – 284, 1998.

[4] Papdimitriou, C., Tamaki, H., Raghavan, P., & Vempala, S. "Latent semantic Indexing: a probabilistic analysis", In *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems* (pp. 159-168). Seattle, Washington, United States: ACM, 1998.

[5] Coello, C. A., Lamont, G. B., & Van Veldhuizen, D. A. "*Evolutionary Algorithms for Solving Multi-Objective Problems 2nd Ed.* New York: Springer Science + Buisness Media., 2007.

[6] Mostaghim, S., Teich, J., & Tyagi, A., "Comparison of Data Structures for Storing Pareto-sets in MOEAs" In *Proceedings of the 2002 World on Congress on Computational Intelligence. 1*, pp. 843-848. Honolulu, HI, USA: WCCI, 2002.

[7] Bhattacharya, M "Exploiting Landscape Information to Avoid Premature Convergence in Evolutionary Search" In *Proceedings of the 2006 IEE Congress on Evolutionary Computation* (pp. 560 - 564). Vancouver, BC: IEEE, 2006.

[8] Xu, Y., Deli, Y., & Yu, L., "Efficient Annealing - Inspired Genetic Algorithm for Information Retrieval from Web-Document", In Proceedings of the first ACM/SIGEVO Summit on Genetic and Evolutionary Computation, 2009.

[9] Mattson, T. G., Sanders, B. A., & Massingill, B. L. *Patters for Parallel Programming.* Boston: Addison-Wesley, 2005.

[10] NVIDIA Corporation. *CUDA Zone - Documentation.* Retrieved November 24, 2009, from CUDA Zone: http://developer.download.nvidia.com/compute/cuda/2_3/toolkit/docs/NVIDIA_CUDA_Programming_Guide_2.3.pdf, accessed 2009, August 26.

Web-Algorithm with Optimal Traffic Control for Location-Information Contents by Sampling Duration

Bong Joon Choi, Jung Min Lee
Dept. of Health Administration, Masan University, Korea
E-mail: bjchoi@masan.ac.kr, min30@masan.ac.kr

Abstract— **To analyze web-browser users' propensity of using contents, the web market has passed diverse changes for the past several years, thereby now coming to extract the degree of concern for the optimal analysis of propensity as well as the locational information according to users' preference for contents. The existing analysis on degree of concern was carried out centering simply on click event, but was difficult for the optimal traffic acceptance by sampling duration, and had the limitation of analysis targeting just the propensity on users' information inquiry. Accordingly, through the proposed algorism, there is a suggestion for the optimally Traffic Control Algorithm in the high sampling duration as well as for extraction in the degree of concern according to location.**

Keywords: *Locational information, Click event, Optimized data packet, sampling cycle, User propensity*

I. Introduction

For the past several years, a series of process according to collecting, accumulating and analyzing these data had required the huge cost to implement these systems, and have suffered difficulty for being wasted additional manpower and time in order to efficiently perform such data. To solve this problem, many researches on Web Log Data Mining were progressed. As it leads now to analysis on web-site visitors' degree of concern in contents, there came to be concern in analysis of users' degree of concern and propensity by using the existing page view and Hits algorism and in the optimal algorism of location-information contents centering on users' preference. The conventional method, namely, the analysis with the focus simply on click event, which occurs on web browser, had limitation of analysis targeting just users' propensity on information inquiry. Also, the browser, in which diverse educational information contents like portal learning site and EduMall have the restricted size, is difficult to be expressed all. The algorism is implemented aiming at the optimally data collection necessary for information transmission and the optimized data generation for packet through diverse experiments on information acquisition and loss in the location-information contents by sampling duration, along with a case of transmitting it to log server with real time by correctly and swiftly extracting information on cursor movement and page scroll and by analyzing this, as well as expressing contents even in the location of surpassing the fixed resolution and conforming it through scroll.

This study aims at the fundamental research for implementing Traffic Control Algorithm in order to be changed with real time into the intelligent information site, which recognizes the optimal traffic acceptance in line with a goal to which the site points, by collecting information on contents of prioritizing users' preference given the educational-information site with the aim of diverse contents.

II. Related work

A. VIPS(Vision based Page Segmentation)

People view a web page through a web browser and get a 2-D presentation which provides many visual cues to help distinguish different parts of the page, such as lines, blanks, images, colors, etc. Jiawei Han proposed a vision-based page segmentation method called VIPS in [2].

This method simulates how a user understands web layout structure based on his or her visual perception. The DOM structure and visual information are used iteratively for visual block extraction, visual separator detection and content structure construction. Finally a vision-based content structure can be extracted.

The vision-based content structure of sample page is illustrated. Visual blocks are detected as shown in Figure 1(b) and the content structure is shown in Figure 1(c). It is an approximate reflection of the semantic structure of the page.

Authors are interesting in Web-Algorithm or similar field, however in some ways, we (authors) are eager to study and pursue with medical informatics as well. This work is funded by Masan University.

Bong Joon Choi is with Masan University, #100 YongDam-Ri, NaeSeo-Up, ChangWon, GyeongNam, Korea
(phone: +82-10-9345-1114 or +82-55-230-1284; fax: +82-55-230-1445; e-mail: bjchoi@ masan.ac.kr).

Jung Min Lee is with Masan University, #100 YongDam-Ri, NaeSeo-Up, ChangWon, GyeongNam, Korea
(phone: +82-10-9655-2900 or +82-55-230-1364; fax: +82-55-230-1445; e-mail: min30@ masan.ac.kr).

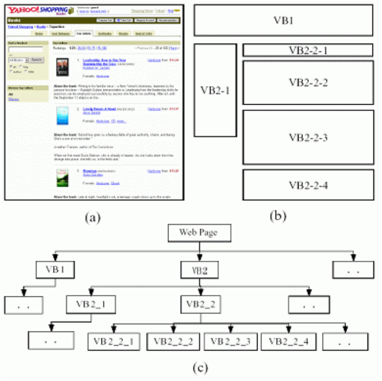


Figure 1. Vision-based content structure of sample page

B. HITS

HITS calculate attribute values of Authorities and Hubs by analyzing links between web documents using the algorithm suggested by Kleinberg[ ]. It comes to determine Authorities values by the frequency number of interconnected links between web documents and on the contrary, gets Hubs values known. In reviewing the operation process, first of all, when the number of web document sets of the searched result obtained by applying users' inquiries to the meaning-based search engine is N, it creates subgraphs of pages related to N pieces of inquiries and regards them as input values of the HITS algorithm. Next, by using created subgraphs it turns to the stage of calculating Hubs and Authorities.

$$Hub(n) = \sum_{u \in S | n \rightarrow u} Auth(u)$$

$$Auth(n) = \sum_{u \in S | n \rightarrow u} Hub(v)$$

[FORMULA 2.1]

In [Formula 2.1], Auth[n] is the Authorities-score to web documents-n which is n∈N, and Hub[n] is the Hub-score. Operation repetition in [Formula 2.1] is operated until it is converged, and 5 repetitions is regarded as a standard through the empirical experiment. Therefore Auth[n] and Hub[n] are finally converged as the Authorities-score and Hub-score. This HITS algorithm fundamentally has a problem that it is dependent on characteristics of web document sets organized in the earlier stage by expanding web document sets in the upper class calculated by the meaning-based search engine on pertinent inquiries.

C. Proposed Algorithm

The Traffic Control Algorithm in the location-information contents of prioritizing users' preference, which is proposed in this study, was implemented a site available for always delivering comfortable contents and was developed the approximate optimal algorithm, on the basis of analyzing users' demand, degree of concern in main contents, and propensity with real time, by embedding the visually analytical system in educational contents of analyzing information on the visible representation location and size in educational contents with pixel unit, and the new algorism in educational-information contents through traffic control by sampling duration along with information on window cursor and information on page scroll on the web browser.

In the proposed algorithm, we analyze by collecting information on the actually browser users' behavior with escaping from stage of analyzing log, which was generated in the existing web log server, and then by transmitting this to log server. In [Fig. 2], there is an analysis as [Table 1] by dividing it into (T1), which is time of being seen as B, (T2), which is time of having been seen as C, and (T3), which is time of disappearing as D, according to the recognition ratio in web contents.

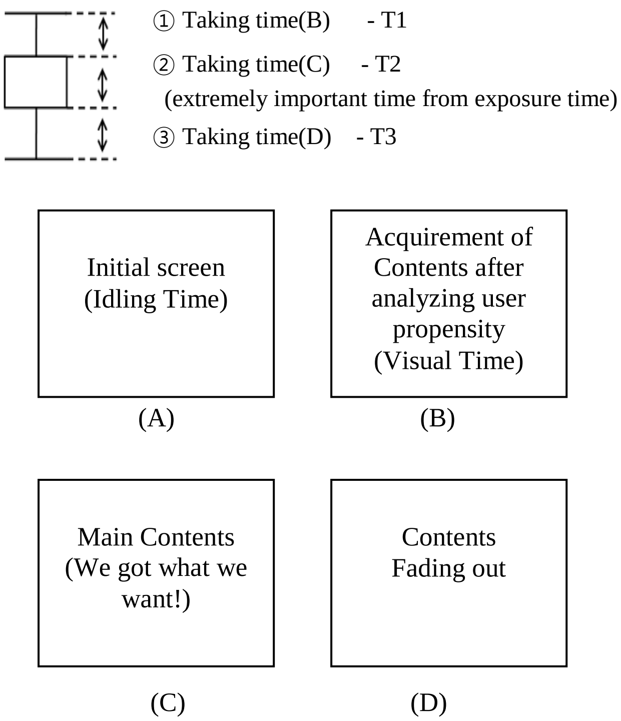

① Taking time(B) - T1
② Taking time(C) - T2
(extremely important time from exposure time)
③ Taking time(D) - T3
Initial screen (Idling Time)
Acquirement of Contents after analyzing user propensity (Visual Time)
(A)
(B)
Main Contents (We got what we want!)
Contents Fading out
(C)
(D)

Figure 2. Dividing by the recognition Ratio for web contents

TABLE 1.ANALYSYS OF THE RECOGNITION RATIO

Entry	Total exposure number	Total exposure time	T1	T2	T3
Texts	965	4,350	-	-	-
Contents	825	2,750	60	2,440	156
....	.	.	.	.	.

In the meantime, there is necessity for enabling Contents on demand in real while minimizing traffic through analyzing appropriate degree of concern in order to prevent a case of going over to log analyzer by being exposed to the users' actual desired contents and mouse and by being recognized as pseudo-degree on coordinates, given the educational-information site [Fig. 3] shows the configuration diagram in the analytical system for the degree of concern.

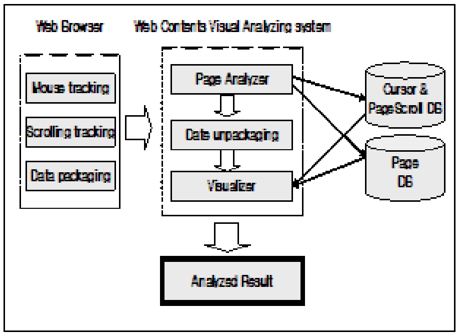


Figure 3. System component for the degree of concern

It uses the DOM analyzer for analyzing educational contents' location and size in the inside of web browsers where their layouts and hyperlinks are expressed and storing them in the database. [Figure 4] shows an algorithm extracting hyperlinks included in web pages and it analyzes DOM information through the web browser's interface after web pages are expressed on web browsers completely. It analyzes tags-expressed location and size like [Figure 4] concerning some tags such as AREA and TABLE related to web pages' layouts and hyperlinks through analyzing DOM objects.

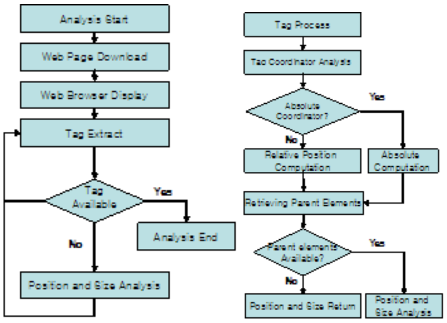


Figure 4. Extracting algorithm

When it brings a page list being analyzed from the page database and completes downloading through the Page Downloader, the Page Presenter expresses downloaded pages using the DOM of the web browser control. The Page Analyzer analyzes contents' location and size by extracting contents included in pages by the DOM Parser. Analyzed information offers visualized information using the Visualizer and stores it in the database.

Next, as for analyzing traffic information by sampling duration, there is an analysis in totally 10 times from 0.1 second to 1.0 second with 0.1 sec. unit, respectively. [Table 2] shows the traffic volume for 10 seconds when Java Script transmits it to log server by having the first 0.1 sec. as duration.

TABLE 2. TRAFFIC VOLUME

Collecting duration(sec.)	Traffic(bps)
0.1	150
0.2	78
0.3	46.5
0.4	38
0.5	32
0.6	25
0.7	22.5
0.8	18
0.9	16.5
1.0	14

[Fig. 5] shows the proposed algorism that induces the optimal traffic on the location-information contents of prioritizing users' preference, it has basis as the traffic volume by collection duration in the above.

```
 Optimized traffic sorter (subject, starting_urls, frequency) {
 user_inform (starting_urls);
 foreach link (starting_urls) {
 enqueue (frontier, link);
 }
 while (visited < MAX_PAGES) {
  user=Dequeue(user_queue);
  concern_list=Loadconcern(user);
  Enqueue(concern_queue, concern_list);
while (not empty(concern_queue))
concern=GetconcernData();
link := dequeue_top_link(frontier);
 doc := fetch(link);
 concern_WM := CWC(subject, doc);
 score_sim := sim(subject, doc);
 enqueue(buffered_pages, doc, score_sim);
 if (#buffered_pages >= MAX_BUFFER) {
 dequeue_bottom_links(buffered_pages);
 }
 merge(frontier, extract_links(doc), concern_WM);
 if (#frontier > MAX_BUFFER) {
 dequeue_bottom_links(frontier);
 }
 }
```

Figure 5. Traffic control algorithm

D. *Implementation and Evaluation*

As a result of experimenting with the proposed algorism, in the conventional system as shown in [Fig. 6], the lengthier sampling duration from 0.1 second to 1.0 second led to the more gradual reduction in traffic volume per second. However, a result of the proposed algorism as shown in [Fig.7] the traffic volume per second is stable and has no big change as exceeding 95% in the same ratio as the original data.

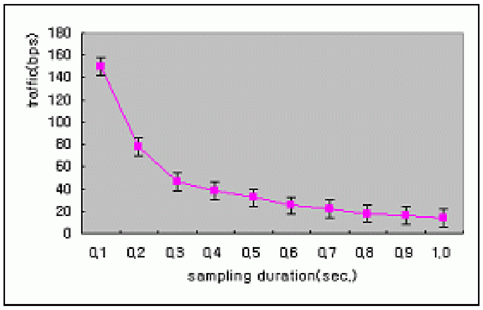


Figure 6.

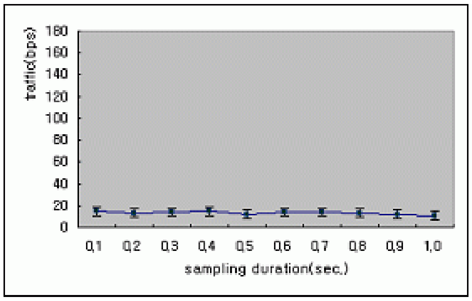


Figure 7.

[Table 3] shows the original data and traffic volume by collection duration in the existing system through log analyzer.

TABLE 3. ORIGINAL DATA & TRAFFIC VOLUME

Collecting Duration (sec)	Comparing to original data (%)	Traffic quantity per second (bps)
0.1	99.7	150
0.2	51.4	78
0.3	32.1	46.5
0.4	25.1	38
0.5	19.8	32
0.6	17.0	25
0.7	13.9	22.5
0.8	12.2	18
0.9	11.0	16.5
1.0	10.3	14

[Table4] shows the traffic volume by collection duration after accepting the proposed algorism.

TABLE 4. TRAFFIC VOLUME AFTER APPLYING THE PROPOSED ALGORITHM

Collecting duration(sec.)	Traffic(bps)
0.1	15
0.2	13
0.3	14
0.4	15
0.5	12
0.6	14
0.7	14
0.8	13
0.9	12
1.0	11

Meanwhile, as [Fig. 8], when the educational information mall, which deals with complex contents, considers traffic volume by figuring out users' preference, the stable mechanism can be confirmed with the proposed algorism even in the high sampling duration(per 0.2 sec.).

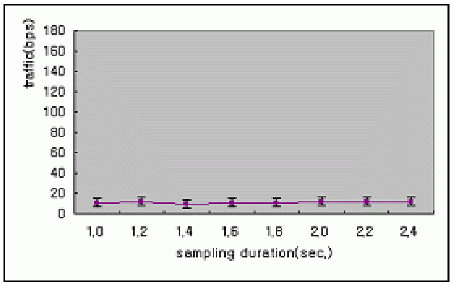


Figure 8.

III. CONCLUSION

This study proposed a mechanism that can acquire contents and can control traffic according to prioritizing preference, by using the proposed mechanism, with judging that there is a little limitation to the analysis of having the basis as the propensity of information inquiry or frequency centering on the click event in the conventional system. A continuous research is needed hereafter on implementing the intelligent algorism that can develop a system of analyzing propensity available for accepting proper traffic in the higher sampling duration, and that can analyze actual degree of concern of the collected information.

REFERENCES

[1] Barker D., Carey M. s., "Factors underlying mouse pointing performance", Contemporary 2006.

[2] Jiawei Han, Miceline Kamber, Data Mining: Concepts & Techniques. Simon Fraser Univ.

[3] Laura A. Granka, "Eye-Tracking Analysis of User Behavior in WWW-Search", Cornell Univ., 2002.

[4] PhilipHeller & SimonRoberts, "Inside Secrets Java2 Developer's Handbook", pp. 645-715. 2007

[5] www.nanet.go.kr & www.netthru.com

[6] D. MacKay. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, 2003.

[7] J. Cho and H. Garcia-Molina, Parallel algorithm. In Proc. of the 11th International World-Wide Web Conference. 2002.

[8] M. Yin, D. Goh, E. Lim and A., Discovery of Concept Entities from Web Sites using Web Unit Mining International Journal of Web Information Systems, Troubador Publishing, UK, vol. 1 no. 3, 2005.

A new initialization method for the Fuzzy C-Means Algorithm using Fuzzy Subtractive Clustering

Thanh Le, Tom Altman
Department of CSE, University of Colorado Denver, Denver, CO, USA

Abstract - *Fuzzy C-means (FCM) is a popular algorithm using the partitioning approach to solve problems in data clustering. A drawback to FCM, however, is that it requires the number of clusters and the clustering partition matrix to be set a priori. Typically, the former is set by the user and the latter is initialized randomly. This approach may cause the algorithm get stuck in a local optimum because FCM depends strongly on the initial conditions. This paper presents a novel initialization method using fuzzy subtractive clustering. On both artificial and real datasets, this algorithm is able, not only to determine the optimal number of clusters, but also to provide better clustering partitions than standard algorithms.*

Availability: The supplementary documents and the method software are at http://ouray.ucdenver.edu/~tnle/fzsc.

Keywords: fuzzy c-means; fuzzy subtractive clustering

1 Introduction

In data mining, clustering is used to group data points based on their similar properties. Data points within a cluster are highly similar to each other and can be discriminated from data points within other clusters. Successful clustering, therefore, maximizes both the compactness within clusters and the discrimination between clusters. Approaches to clustering include partitioning and hierarchical methods. Of the former, a popular algorithm is Fuzzy C-Means (FCM, Bezdek 1981) that uses fuzzy cluster boundaries and fuzzy sets to associate every data point with at least one cluster. An advantage of FCM is that it converges rapidly, however, like most partitioning clustering algorithms, it depends strongly on the initial parameters and the estimate of the number of clusters. For some initial values, it will converge rapidly to a global optimum, however, for others, but it may become stuck in a local optimum.

To address the limitations of FCM, researchers recently have integrated the algorithm with optimization algorithms, such as, the Genetic Algorithm, Particle Swarm Optimization, and Ant Colony Optimization. Alternatively, a Subtractive Clustering (SC) method has been used with FCM, where SC is used first to determine the optimal number of clusters and the location of cluster centers [1-4] and FCM is then used to determine the optimal fuzzy partitions [1], [5-7]. While this approach can overcome the problem of initialization of parameters, it still requires a priori specification of the parameters of the SC method: the mountain peak and the mountain radii. SC uses these two parameters to compute and amend the value of the mountain function for every data point while it is looking for the cluster candidates. Most approaches using the traditional SC method use constant values for these parameters [1], [3-7]. However, different datasets have different data distributions, and, therefore, these values need to be adjusted accordingly. Yang and Wu [2] proposed a method to do this automatically. However, because the data densities are not always distributed equally within the dataset, the automatic values may be appropriate only for some data points.

In general, use of the SC method to determine the optimal number of clusters is based on the ratio between the amended value and the original value of the mountain function at every data point. The data points at which the ratios are above some predefined cutoff are selected and their number is used as the optimal number of clusters [1], [3], [6]. This approach, however, requires the specification of the cutoff value which will differ among datasets.

In this study, we combine FCM with a new SC method to automatically determine the number of clusters in a dataset. FCM randomly creates a set of fuzzy partition solutions for the dataset using a maximum number of clusters; SC then uses the fuzzy partition approach to search for the best cluster centers and the optimal number of clusters for each solution. The FCM algorithm then rapidly determines the best fuzzy partition of every solution using the optimal number of clusters determined by SC. The best solution from this solution set is the final result.

This work was supported in part by the Vietnamese Ministry of Education and Training (TL).

Thanh Le is a doctoral student in the Department of Computer Science and Engineering, University of Colorado Denver, Denver, CO 80217-3364, USA (email: lntmail@yahoo.com).

Tom Altman is a professor in the Department of Computer Science and Engineering, University of Colorado Denver, Denver, CO 80217-3364, USA (phone: 303-556-3434; email: Tom.Altman@ucdenver.edu).

2 Fuzzy C-Means and Subtractive Clustering method

2.1 Fuzzy C-Means algorithm (FCM)

The FCM algorithm uses fuzzy set memberships to associate every data point with at least one cluster.

Given a dataset X = {$x_i \in R^p$, i=1..n}, where n>0 is the number of data points and p>0 is the dimension of the data space of X, let c, $c \in N$, $2 \le c \le n$, be the number of clusters in X.

Denote V = {$v_k \in R^p$, k=1..c} as the set of center points of c clusters in the fuzzy partition; U = {$u_{ki} \in [0,1]$, i=1..n, k=1..c} as the partition matrix, where u_{ki} is the fuzzy membership degree of the data point x_i to the k^th cluster, and

$$\sum_{k=1}^{c} u_{ki} = 1, \ i = 1..n. \tag{1}$$

The clustering problem is to determine the values of c and V such that:

$$J(X \mid U, V) = \sum_{i=1}^{n}\sum_{k=1}^{c} u_{ki}\|x_i - v_k\| \to \min, \tag{2}$$

where ||x-y|| is the distance between the data points x and y in R^p, defined using Euclidean distance as:

$$\|x - y\|^2 = \sum_{i=1}^{p}\left(x^i - y^i\right)^2. \tag{3}$$

By using fuzzy sets to assign data points to clusters, FCM allows adjacent clusters to overlap, and therefore provides more information on the relationships among the data points. In addition, by using a fuzzifier factor, m, in its objective function (4), the clustering model from FCM is more flexible in changing the overlap regions among clusters.

$$J(X \mid U, V) = \sum_{i=1}^{n}\sum_{k=1}^{c} u_{ki}^{m}\|x_i - v_k\| \to \min, \tag{4}$$

where m, $1 \le m < \infty$, is the fuzzifier factor.

Equation (4) can be solved using Lagrange multipliers with respect to (1).

$$v_k = \sum_{i=1}^{n} u_{ki}^{m} x_i \Big/ \sum_{i=1}^{n} u_{ki}^{m}, \tag{5}$$

$$u_{ki} = \left(\frac{1}{\|x_i - v_k\|^2}\right)^{\frac{1}{1-m}} \Bigg/ \sum_{j=1}^{c}\left(\frac{1}{\|x_i - v_j\|^2}\right)^{\frac{1}{1-m}}. \tag{6}$$

To estimate the solution of the system of equations (5) and (6), FCM uses an iteration process. The values of U are initialized randomly. The values of V are estimated using (5). The values of U are then re-estimated using (6) with the new values of V. This process is iterated until convergent where

$\exists \varepsilon_u > 0$, T > 0: $\forall t > T$,

$$\|U_{t+1} - U_t\| = \max_{k,i}\{\|u_{ki}(t+1) - u_{ki}(t)\|\} < \varepsilon_u. \tag{7}$$

Or, $\exists \varepsilon_v > 0$, T > 0: $\forall t > T$,

$$\|V_{t+1} - V_t\| = \max_{k}\{\|v_k(t+1) - v_k(t)\|\} < \varepsilon_v. \tag{8}$$

The FCM algorithm has the advantage of converging quickly, however, it may get stuck in local optima and be unable to detect the global optimum.

2.2 Subtractive Clustering method (SC)

SC is commonly used to find cluster centers and provide the optimal number of clusters. To detect the center point of a new cluster, SC considers each data point as a potential cluster center. The mountain function is used to measure the potential for each data point to be a cluster center.

$$M(x_i) = \sum_{j=1}^{n} e^{-\frac{\|x_i, x_j\|^2}{(\alpha/2)^2}}, \tag{9}$$

where α>0, is a predefined constant that represents the mountain peak or the neighborhood area radius of each data point. The greater the number of close neighbors a data point has, the higher the magnitude of its mountain function. The mountain function can be viewed as a measure of data density in the neighborhood of each data point. Data points with large values for the mountain function have more chances to become cluster centers, and the data point with the maximum value of the mountain function is the first choice for the center of the cluster.

Once a data point is selected as center point of a cluster, its neighbors are affected. Let x* be the selected data point and M* be the mountain value at x*. The mountain function values of its neighbors are then amended as:

$$M_t(x_j) = M_{t-1}(x_j) - M^* e^{-\frac{\left\| x^* - x_j \right\|^2}{\left(\beta/2 \right)^2}}, \qquad (10)$$

where β>0, a predefined constant, represents the mountain or the affected neighborhood area radius at each data point. Usually, β > α and quite often β = 1.5α.

SC stops increasing the number of clusters when it is equal to a predefined constant C_{max}, or the ratio between the value of M* at time t, M_t*, and its original value, M_0*, is less than a predefined constant δ,

$$\frac{M_t^*}{M_0^*} < \delta. \qquad (11)$$

While SC can help to determine the number of clusters in the dataset, the constants α, β, and σ must be specified, and will differ among datasets.

3 The proposed algorithm

3.1 Fuzzy subtractive clustering (SC) method

We first propose a novel fuzzy SC method that uses a fuzzy partition of the data instead of the data themselves. Our method addresses the drawback of the traditional SC in that it does not require a priori the values of the mountain peak and mountain radii.

3.1.1 Fuzzy mountain function

Given {U,V}, a fuzzy partition on X, the accumulated density at v_k, k=1..c, is calculated [8] as

$$Acc(v_k) = \sum_{i=1}^{n} u_{ki}. \qquad (16)$$

The density at each data point x_i is estimated, using a histogram based method, as

$$dens(x_i) = \sum_{k=1}^{c} Acc(v_k) \times u_{ki}. \qquad (17)$$

The density estimator in (17) is more true at the centers {v_k} than at other data points [8]. Therefore, if we try to find the most dense data points, c, using (17), we will obtain the centers of c clusters of the fuzzy partition. To address this problem, we use the concept of strong uniform fuzzy partition [8]. We define a new fuzzy partition {U',V} using {U,V},

$$u'_{ki} = \left(e^{\frac{\left\| x_i - v_k \right\|}{\sigma_k^2}} \right)^{-1} \Bigg/ \sum_{l=1}^{C} \left(e^{\frac{\left\| x_i - v_l \right\|}{\sigma_l^2}} \right)^{-1}, \qquad (18)$$

where

$$\sigma_k^2 = \sum_{i=1}^{n} P(x_i \mid v_k) \left\| x_i - v_k \right\|^2. \qquad (19)$$

Because {U',V} is a strong uniform fuzzy partition [8], the density at every data point can then be estimated as

$$dens(x_i) = \sum_{k=1}^{c} Acc(v_k) \times u'_{ki}. \qquad (20)$$

3.1.2 Fuzzy mountain function amendment

Each time the most dense data point is selected, the mountain function of the other data points must be amended to search for new cluster centers. In the traditional SC method, the amendment is done using the direct relationships between the selected data point and its neighborhood, i.e., using a pre-specified mountain radius (10). In our approach, this is done using the fuzzy partition and no other parameters are required. First, the accumulated density of all cluster centers is revised using

$$Acc^{t+1}(v_k) = Acc^t(v_k) - M_t^* \times P(v_k \mid x_t^*), \qquad (21)$$

where x_t^* is the data point selected as the new cluster center and M_t^* is the mountain function value, at time t. The density at each data point is then re-estimated using (20) and (21) as follows,

$$\begin{aligned} dens^{t+1}(x_i) &= \sum_{k=1}^{c} Acc^{t+1}(v_k) \times u'_{ki} \\ &= \sum_{k=1}^{c} \left[Acc^t(v_k) - M_t^* \times P(v_k \mid x_t^*) \right] \times u'_{ki} \\ &= \sum_{k=1}^{c} Acc^t(v_k) \times u'_{ki} - M_t^* \sum_{k=1}^{c} u'_{ki} \times P(v_k \mid x_t^*) \\ &= dens^t(x_i) - M_t^* \sum_{k=1}^{c} u'_{ki} \times P(v_k \mid x_t^*). \end{aligned}$$

Hence, the estimated density at each data point is amended as:

$$dens^{t+1}(x_i) = dens^t(x_i) - M_t^* \sum_{k=1}^{c} u'_{ki} \times P(v_k \mid x_t^*). \qquad (22)$$

To determine the optimal number of clusters in the dataset, we first locate the most dense data points, $\lfloor \sqrt{n} \rfloor$. Only the data points where the ratio between the mountain function values at time t and time 0 is above 0.95 are selected. This number is considered the optimal number of clusters in the dataset.

3.1.3 Fuzzy subtractive clustering algorithm

- Input: Fuzzy partition {U,V} on X.
- *Output: Optimal number of clusters, c*, and c* cluster candidates.*

Steps

1. Construct a new fuzzy partition {U',V} using (18)
2. Estimate the density in X using (20)
3. Search for the most dense data point, say x*
4. If dens^t(x*) > 0.95*dens^0(x*) then select x* as a cluster candidate
5. Amend all mountain function values using (22)
6. If the number of such visited x* is less than $\lfloor\sqrt{n}\rfloor$ then go to step 3
7. Return the set of selected cluster candidates.

Regarding computational complexity, the evaluation of formula (18) is equal to that of (6); and (19) is the same as (5). The running time of (16), (20), (21) and (22) is O(c×n). Hence, the computational complexity of our fuzzy SC method is O(c×n) which is equal to an iteration of the FCM algorithm. This is in contrast to the traditional SC method, as in (9) and (10), which is O(n^2). Thus, for large datasets, while n increases significantly, c may become large, but is reasonably limited by √n. This illustrates another advantage of our method over the traditional SC.

3.2 Model selection method

In order to decide the optimal number of clusters in the dataset, we randomly generate multiple clustering solutions and choose the best one. This approach is a novel method for model selection using the log likelihood estimator with fuzzy partition. Each clustering solution is modeled with θ = {U,V}. Our model selection method is based on the likelihood of the solution model and the dataset as

$$\begin{aligned} L(\theta \mid X) &= L(U,V \mid X) \\ &= \prod_{i=1}^{n} P(x_i \mid U,V) = \prod_{i=1}^{n}\sum_{k=1}^{c} P(v_k) \times P(x_i \mid v_k). \end{aligned} \tag{23}$$

The log likelihood estimator is computed as

$$\log(L) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{c} P(v_k) \times P(x_i \mid v_k)\right) \rightarrow \max. \tag{24}$$

Because our clustering model is a possibility based one, a possibility to probability transformation is needed before applying (24) to model selection. For each fuzzy partition, u_{ki} represents the possibility of x_i given the cluster v_k. We used the method of Florea et al. [9] to approximate the probability of x_i given v_k, $P^a(x_i|v_k)$. The prior probability $P(v_k)$ is then estimated as in (25),

$$P(v_k) = {\sum_{i=1}^{n} P^a(x_i \mid v_k)} \Big/ {\sum_{l=1}^{c}\sum_{i=1}^{n} P^a(x_i \mid v_l)}. \tag{25}$$

The probability of x_i given v_k is computed using $P^a(x_i|v_k)$ and the Gaussian distribution model as

$$P(x_i \mid v_k) = \max\left\{ P^a(x_i \mid v_k), \left((2\pi)^{1/n} \times \sigma_k \times e^{\frac{\|x_i - v_k\|^2}{2\sigma_k^2}} \right)^{-1} \right\}. \tag{26}$$

Equation (26) represents the data distribution better if it is not Gaussian. Our model selection method is based on (24) with (25) and (26).

3.3 The proposed algorithm

We combine the FCM algorithm with our fuzzy SC method for a novel clustering algorithm (fzSC) that automatically determines the optimal number of clusters and the fuzzy partition for the given dataset.

- Input: The data to cluster X={x_i}, i=1..n.
- Output: An optimal fuzzy partition solution,
  - c: Optimal number of clusters.
  - V = {v_i }, i =1..c: Cluster centers.
  - U={u_{ki}}, i=1..n, k=1..c: Partition matrix.

Steps

1. Randomly generate a set of L fuzzy partitions where the number of clusters is set to √n
2. For each fuzzy partition,
   - Use the fuzzy SC method to determine the optimal number of clusters
   - Run the FCM algorithm using the optimal number of clusters and the partition matrix found by the fuzzy SC
   - Compute the fitness value of the fuzzy partition solution using (24)
3. Select the 'best' solution from the partition set based on the fitness values
4. Return the 'best' solution.

4 Experimental results

To evaluate the performance of the fzSC algorithm, we generated one dataset manually in a 2D space with 171 data points in 6 clusters (labeled from 1 to 6 (Fig.1)) and 200 datasets using the method based on a simple finite mixture model [12]. Datasets are distinguished by the dimensions and cluster number, and we generated (5-2+1)*(9-5+1)=20 dataset types. For each type, we generated 10 datasets, for a total of 200. For the real datasets, we used Iris, Wine, Glass, and Breast Cancer Wisconsin datasets from the UC Irvine

Machine Learning Repository [10]. These real datasets contain classification information. They are therefore helpful for clustering algorithm evaluation.

4.1 Evaluation of candidate cluster centers initialization

We first used the manually created (MC) and Iris datasets to demonstrate the ability of the fzSC algorithm to determine the centers of the candidate clusters.

4.1.1 MC dataset

For the MC dataset, we first ran the standard FCM algorithm with the number of clusters set to 13 which is equal to the square root of 171, and the partition matrix randomly initialized. fzSC was then used to search for the six (the known number of clusters) most dense data points as candidate cluster centers. The 13 cluster centers found by the FCM algorithm and the six candidate cluster centers found by fzSC were plotted as in Fig.1.

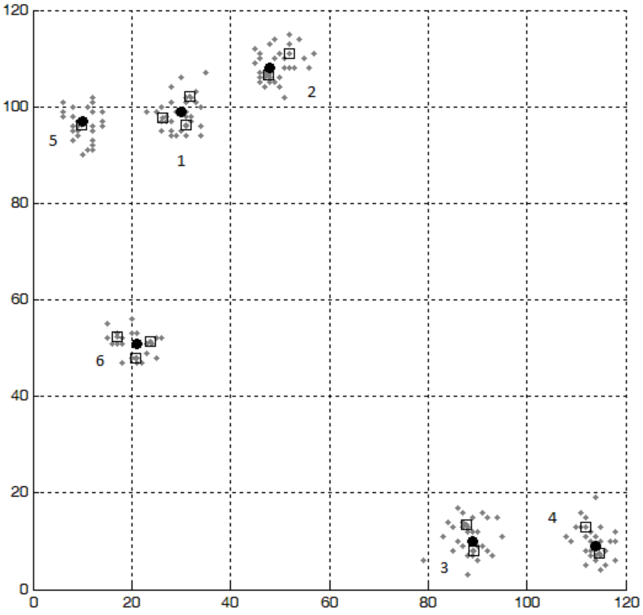


Fig.1. Candidate cluster centers in the MC dataset found using fzSC. Squares, cluster centers from FCM; dark circles, cluster centers found by fzSC. Classes are labeled by numbers from 1 to 6.

Fig.1 shows that fzSC successfully detected the best centers for the six cluster candidates in the MC dataset.

4.1.2 Iris dataset

For the Iris dataset, we first ran the standard FCM algorithm with the number of clusters set to 12 which is equal the square root of 150, and the partition matrix initialized randomly. We then used fzSC to search for the three candidate cluster centers. Cluster centers found by FCM and fzSC are plotted in Fig.2. fzSC successfully detected the best centers for the three cluster candidates in the Iris dataset.

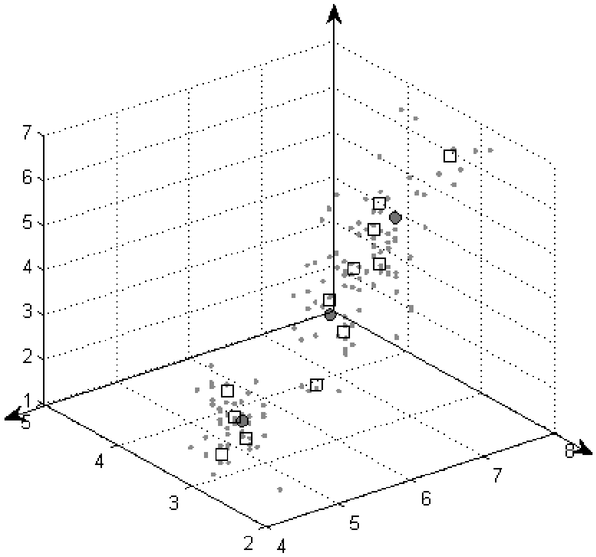


Fig.2. Candidate cluster centers in the Iris dataset found using fzSC. Squares, cluster centers from FCM; dark circles, cluster centers found by fzSC.

4.2 Evaluation of partition matrix initialization

To evaluate the effectiveness of fzSC in initializing the partition matrix for FCM, we compared the performance of fzSC with that of k-means, k-medians and FCM using the MC dataset with 6 clusters. The partition matrices of the three standard algorithms were initialized randomly at each runtime. We ran each of these algorithms three times and recorded their best solutions. For the fzSC algorithm, the standard FCM algorithm was run once with the number of clusters set to 13. The fuzzy SC method was then applied to search for the six candidate cluster centers. The standard FCM algorithm was then rerun with the values of c and V set to the six cluster centers found by the fuzzy SC method. We repeated this experiment 20 times and averaged the performance of each algorithm (Table 1).

Table 1

The algorithm performance on the MC dataset where the number of clusters is set to 6

Algorithm	Correctness ratio by class						Avg. Ratio
	1	2	3	4	5	6	
fzSC	1.00	1.00	1.00	1.00	1.00	1.00	1.00
k-means	0.97	0.87	1.00	1.00	1.00	0.75	0.93
k-medians	0.95	0.82	1.00	1.00	1.00	0.62	0.90
FCM	0.97	1.00	0.95	1.00	1.00	0.96	0.98

Table 1 shows that fzSC outperformed the three standard algorithms on the MC dataset. While fzSC correctly assigned all data points to their corresponding classes, the k-means and k-medians algorithms misclassified some members of classes #1, 2 and 6. The FCM algorithm misclassified some members of classes #1, 3 and 6.

4.3 Evaluation of the number of clusters determination

In this section we evaluate the ability of the fzSC algorithm to determine the optimal number of clusters in a given dataset.

4.3.1 Artificial datasets

We ran the fzSC algorithm 20 times on each of the 200 artificial datasets with the value of L set to 5. The number of clusters found was compared with the known number of clusters in the dataset, and then averaged across datasets of the same type to determine the correctness ratio on each dataset type.

Table 2 shows that fzSC correctly detected the number of clusters on most of the artificial datasets. Its lowest correctness ratio is 0.87; this was obtained with the dataset type of dimension 2 and 9 clusters. Overall, the fzSC algorithm performed well in detecting the optimal number of clusters on the artificial datasets.

Table 2

The performance of the fzSC algorithm in determining the optimal number of clusters in artificial datasets

The number of clusters generated in the dataset	The dataset dimension			
	2	3	4	5
5	0.97	1.00	1.00	1.00
6	1.00	0.98	0.90	1.00
7	1.00	1.00	1.00	1.00
8	1.00	0.99	0.97	1.00
9	0.87	0.99	1.00	0.96

4.3.2 Real datasets

The number of clusters in the Iris, Wine, Glass and Breast Cancer Wisconsin datasets are 3, 3, 6, and 6 respectively. The data in the Wine and Glass datasets were normalized by attributes before clustering. The Pearson Correlation distance, which is the most appropriate for real datasets [11], was used to measure the similarity between data points.

We ran the fzSC algorithm 20 times on each of these datasets with the value of L set to 15. The number of clusters found was compared with the known number in the dataset. As shown in Table 3. fzSC successfully detected the number of clusters in Iris, Wine, and Glass datasets. While its correctness ratio on the Breast Cancer Wisconsin dataset is only 0.65, it's the alternative solution is five clusters, which is close to the optimal one. The results therefore show that fzSC achieved a significant performance on the real datasets.

Table 3

The performance of fzSC in determining the optimal number of clusters in real datasets

Dataset	# data points	known #clusters	predicted #clusters	ratio
Iris	150	3	3	**1.00**
Wine	178	3	3	**1.00**
Glass	214	6	6 5	**0.95** 0.05
Breast Cancer Wisconsin	699	6	6 5	**0.65** 0.35

5 Conclusions

We have presented a novel algorithm to address the problem of parameter initialization of the standard FCM algorithm. We have proposed a new subtractive clustering method that uses fuzzy partition of the data instead of the data themselves. The advantages of fzSC are that, unlike traditional SC methods, it does not require specification of the mountain peak and mountain radii, and, with a running time of O(c×n) compared to O(n^2) for the traditional SC method, it is more efficient for large datasets. In addition, our method can be integrated easily with fuzzy clustering algorithms to search for the best centers of cluster candidates. Finally, we have proposed a new method of model selection using the data likelihood estimator based on fuzzy partitions. The experimental results show that fzSC performs effectively on both artificial and real datasets. In future work, we will integrate fzSC with optimization algorithms for new clustering algorithms that can effectively support clustering analysis on real datasets.

6 References

[1] W. Y. Liu, C. J. Xaio, B. W. Wang, Y. Shi, S. F. Fang, "Study On Combining Subtractive Clustering With Fuzzy C-Means Clustering", in Machine Learning and Cybernetics, 2003 International Conference, Xi'an, 2003, pp. 2659–2662.

[2] M.S. Yang, K.L. Wu, "A modified mountain clustering algorithm", Pattern Anal Applic, Vol. 8, pp. 125–138, 2005.

[3] J.Y Chen, Z. Quin, J. Jia, “A weighted mean subtractive clustering algorithm”, Information Technology, Vol. 7, pp. 356–360, 2008.

[4] C.C. Tuan, J.H. Lee, S.J. Chao, “Using Nearest Neighbor Method and Subtractive Clustering-Based Method on Antenna-Array Selection Used in Virtual MIMO in Wireless Sensor Network”, in Tenth International Conference on Mobile Data Management: Systems, Services and Middleware, Taipei, 2009, pp. 496–501.

[5] J.C. Collazo, F.M. Aceves, E.H. Gorrostieta, J.O. Pedraza, A.O. Sotomayor, M.R. Delgado, “Comparison between Fuzzy C-means clustering and Fuzzy Clustering Subtractive in urban air pollution”, in lectronics, Communications and Computer (CONIELECOMP), 2010 20th International Conference, Cholula, 2010, pp. 174–179.

[6] Q. Yang, D. Zhang, F. Tian, “An initialization method for Fuzzy C-means algorithm using Subtractive Clustering”, Third International Conference on Intelligent Networks and Intelligent Systems, Vol. 10, 2010, pp. 393–396.

[7] J. Li, C.H. Chu, Y Wang, W. Yan, “An Improved Fuzzy C-means Algorithm for Manufacturing Cell Formation”, Fuzzy Systems, Vol. 2, pp. 1505–1510, 2002.

[8] K. Loquin, O. Strauss, “Histogram density estimators based upon a fuzzy partition”, Statistics and Probability Letters, Vol. 78, pp. 1863–1868, 2008.

[9] M.C. Florea, A.L. Jousselme, D. Grenier, E. Bosse, “Approximation techniques for the transformation of fuzzy sets into random sets”, Fuzzy Sets and Systems, Vol. 159, pp. 270–288, 2008.

[10] A. Frank, A. Asuncion, (2010) Machine Learning Repository. [Online]. http://archive.ics.uci.edu/ml.

[11] B.P.P Houte, J. Heringa, “Accurate confidence aware clustering of array CGH tumor profiles”, BioInformatics, Vol. 26(1), pp. 6–14, 2010.

[12] L. Xu, M.I. Jordan, “On convergence properties of the EM algorithm for Gaussian mixtures”, Neural Computation, Vol. 8, pp. 129–151, 1996.

A Methodology to Conceal QR Codes for Security Applications

Akshay Choche and Hamid R. Arabnia

Department of Computer Science, University of Georgia, Athens, GA 30602
{choche, hra}@cs.uga.edu

Abstract. Steganography is a technique used to conceal information in such a way that only the communicating party would know about the existence of the information. In this paper, Steganography is used to embed an encrypted QR Code into an image. The QR Code is encrypted using Triple DES algorithm in order to increase the level of security. The overall approach presented in this paper could potentially be used for secure communication and even for embedding signature/copyright information into an image.

1 Introduction to Steganography

Steganography is a technique used to conceal information in such a way that only the communicating party would know about the existence of the information. The literally meaning of the word Steganography is "covered writing"[6]. This paper provides a brief overview of Steganography and how it can be used to embed information into an image in an efficient manner. The main focus of the paper is to embed an encrypted QR Code into an image.

A brief overview of Steganography is provided in the following paragraph. Steganography uses a **Cover Image**[7]. A Cover Image is the one that would have information embedded in it. Then there is the actual information itself, this information could be anything like plain ASCII text or another image. After the information has been embedded into a cover image using any of the Steganography techniques(such as Least Significant Bit Insertion) a **Stego Image** [7] is obtained. It is also possible to encrypt the information before embedding it, thereby adding another level of security.

Least Significant Bit Insertion is one of the most effective and widely used technique in order to achieve Steganography. In a byte the right most bit is called Least Significant Bit as changing it has the least effect on the value of the Byte. In this technique modifies the last bit of the color components of pixels of the chosen cover image to embed the bit stream corresponding to the information to produce a stego image.

Embedding the data:If you consider a 24-bit color image each pixel has three components(Red, Green and Blue) each having 8-bits, thus its possible to store 3 bits of information in every pixel, on contrary if you use a Grey scale image you can only store 1bit of information per pixel. Now in order to embed a letter A whose ASCII value is 65(i.e. 01000001) in an image would require 3 Pixels. Let the 3 pixels be as below

Red	Green	Blue
11000100	10010100	00100100
10000100	11010101	01101100
11111101	11110100	01100100

Before Embedding the Data.

After Embedding the data the resulting pixels would be.

Red	Green	Blue
11000100	1001010**1**	00100100
10000100	1101010**0**	01101100
1111110**0**	1111010**1**	01100100

After Embedding the Data.

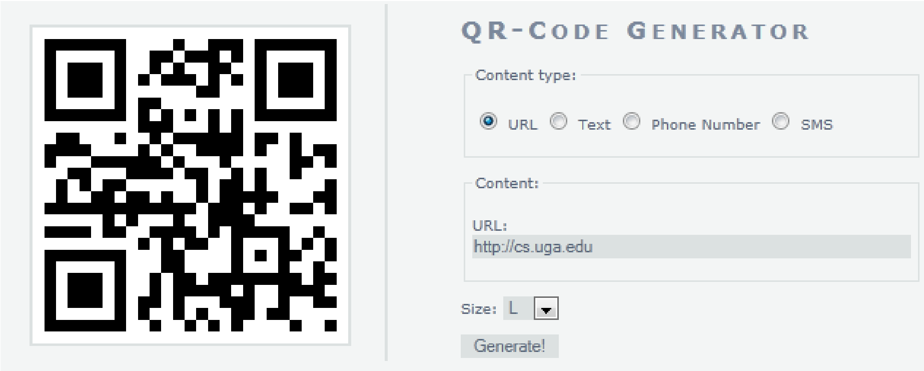
In the above example it can be seen that only 3 bits were flipped. Normally when using this LSB insertion technique on an average 50% of the bits in an image are flipped[8].

Recovering the data: In order to retrieve the information from an image the 8-bit binary equivalent of each RGB color component of pixels is obtained. The LSB of this binary number represents a one bit of the hidden information that was embedded. Each of such bits are then stored in an output file.
One downside of LSB insertion is that it is susceptible to image processing operation such as cropping and compression. Another drawback of this technique is that if the original image was in GIF or BMP file format (i.e. which uses lossless compression technique) and was converted to JPEG file format (i.e. which uses lossy compression technique) and then converted back to the original format then the data in the LSBs would be lost.

2 Introduction to QR Codes

Quick Response Code or better know as QR Code is a two dimensional barcode that allow high speed data encoding and decoding capabilities. It was invented by Denso-Wave[3] a Toyota subsidiary in 1994 in order to track the various parts during the vehicle manufacturing. Generally QR Codes are used for distributing small information like URL, a phone number or even small text. The Government of Canada uses QR Codes for efficient and faster processing of the Passport application forms. A QR Code is embedded on the first page of their application form and the code gets updated as the form is being filled[11]. Also, a Dutch poet Chielie published a collection of 12 poems, QRCode that fits in one sheet of A4 paper.

The most effective and efficient way of decoding a QR Code is using a smart phone equipped with a camera and a compatible decoding application. There are many freeware decoding applications that are available on the Internet one of which is zxing[2]. If users don't have access to smart phones they can access websites such as Xzing[10] in order to decode the QR Codes, all they need to do is upload the QR Code image to the website and they website decode the information and display it to the user. Similarly there are multiple sites such as Kaywa[9] that can be used for generating the QR Code. Figure 1 shown below has a QR Code in it which was generated from Kaywa[9].



QR-CODE GENERATOR

Content type:

● URL ○ Text ○ Phone Number ○ SMS

Content:

URL: http://cs.uga.edu

Size: L

Figure 1: Generating a QR Code using Kaywa[9].

There has been increase in the use of QR Codes and the reason for this increase is due to the various features offered by the QR Codes. One of the most desirable feature is its readability from any direction, also other features provided by QR Codes are high capacity encoding of data, small printout size, Dirt and Damage Resistant and so on[4].

3 Introduction to Symmetric Key Encryption ”Triple DES”

Symmetric Key Encryption techniques are the one in which the keys used for encryption and decryption purpose are the same. In this paper a Triple DES Algorithm[5] is used for the encrypting and decrypting QR Codes. Triple DES is a block cipher, which applies the Data Encryption Standard(DES) cipher algorithm three times to each data block(64-bit long). The advantage of using this algorithm is its ability to have a larger length keys, without designing a new block cipher algorithm. Triple DES utilizes three 56-bits key $Key_1, Key_2andKey_3$. In order to encrypt the data it performs following sequence of operations,

$$ciphertext = Encrypt_{Key3}(Decrypt_{Key2}(Encrypt_{Key1}(plaintext)))$$

And to decrypt the data it performs following sequence of operations,

$$plaintext = Decrypt_{Key1}(Encrypt_{Key2}(Decrypt_{Key3}(ciphertext)))$$

plaintext = The original information.
ciphertext = The encrypted information.
$Encrypt_{Ki}$ = Encrypting using the key K_i.
$Decrypt_{Ki}$ = Decrypting using the key K_i.

4 Understanding the Methodology for concealing the QR Code

The Figure 2 shown below provides the big picture of the methodology used for embedding and extracting encrypted QR Codes into a cover image. This process can be divided into two subsections one which elaborates the embedding phase and the other which elaborates the extraction phase.

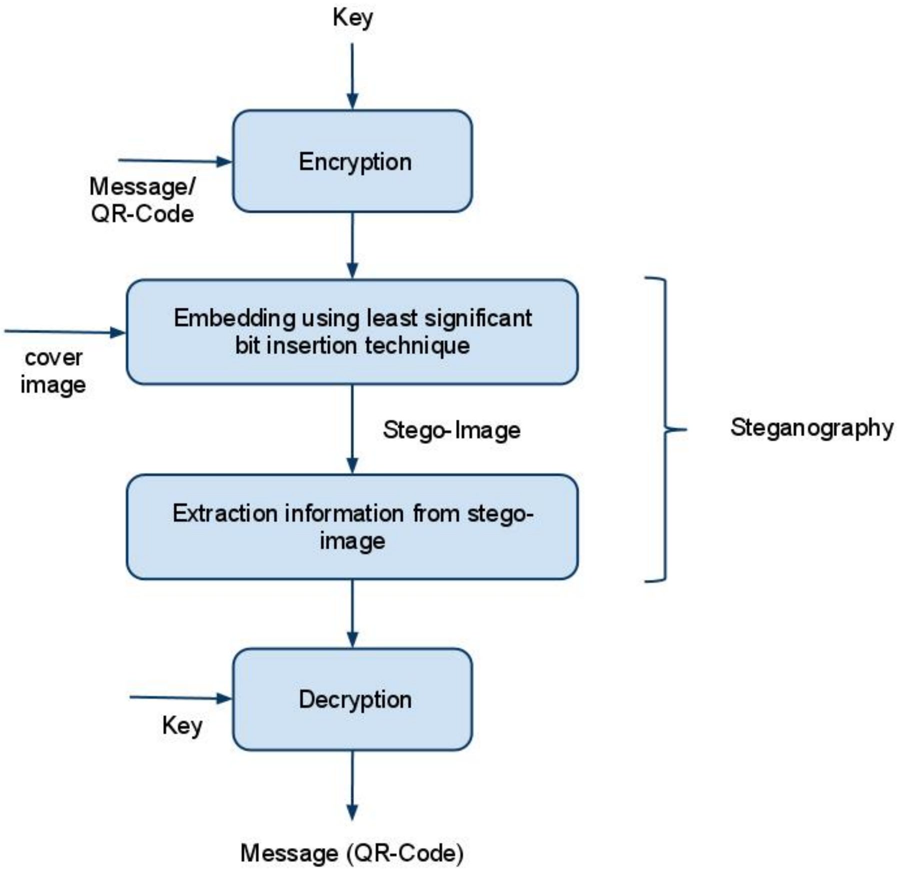

Key
Encryption
Message/
QR-Code
Embedding using least significant bit insertion technique
cover image
Stego-Image
Steganography
Extraction information from stego-image
Decryption
Key
Message (QR-Code)

Figure 2: Embedding & Extracting the QR Code using Steganography.

Once the QR Code has been embedded into a cover image, only the communicating parties will be aware about the existence of the QR Code, and even if some how the existence of the QR Code is discovered using Steganalysis it will be very difficult to decrypt it without the key. Thus this methodology can be used in security applications such as applications for exchanging confidential information or for embedding signature/copyright information in an image. Consider a situation where the owner of an image embeds an encrypted QR Code(which represents his ownership information) into the image. This will allow the owner to claim his ownership in case some one copies the image, also it will not be possible for the offender to modify the QR Code since it has been encrypted using a key that only the owner knows.

4.1 The Embedding Phase

The first step in the embedding phase is to generate the desired QR Code. There are number of QR Code generator available on the Internet, the one that was used in this paper was Kaywa[9]. All you need to do visit ***http://qrcode.kaywa.com*** and use their online application to generate the QR-Code.

The Next step is choosing a cover image. A complete paper can be written on this topic. Here are few pointers for choosing an effective cover image. Choosing a JPEG image as a cover image is probably not a good idea as JPEG use a lossy compression technique. A better candidate would be an image which has Bitmap(BMP) or Graphics Interchange Format(GIF) format. Apart from the file format it is also required that the chosen image has a palette that contains lots of variances in color as an image with less variance would have uniform patches with same color. If a QR Code is embedded in an image with less variance the distortions caused due to the embedding process would be quite visible.

QR Code as described in the Section 2 is a black and white image. Thus any QR Code can be represented as a byte array in Java with '0' representing 'Black' and '1' representing 'White'. This byte array is encrypted using a DES Algorithm. Java has a built in library ***javax.crypto*** that allows efficient encryption and decryption to be performed. In this paper the DESede "Triple DES" Algorithm is used for encrypting the byte array representing the QR Code.

In the last step of this phase the encrypted byte array is embedded[7] into a cover image. In order to achieve this the individual bytes from the byte array are converted into their equivalent 8-bit binary representation using the shift operators provided in Java and these individual bits are embedded using the Least Significant Bit(LSB) insertion technique as mentioned in Section 1. At the end of Embedding Phase a Stego Image[7] is obtained which contains an encrypted QR Code embedded in it.

4.2 The Extraction Phase

In order to extract a QR Code from a Stego Image following operations are performed. The first step involves going over Pixels and fetching the LSBs from the Red, Green and Blue component of the pixel and arranging them in a group

of 8-bits to form a byte which would be part of the byte array. Here it is assumed that the amount of data present in the image is known before hand, however this can easily be implemented dynamically by first embedding the length of the data in the image. Once this encrypted byte array is obtained the next step involved is to decrypt the byte array using the key used during the embedding phase. In this step the DESede "Triple DES" Algorithm is used for decryption purpose to obtain the original byte array which contained only 1's and 0's.
This byte array represents the QR Code, which can be painted on a canvas(i.e. 0 corresponds to a Black pixel and 1 corresponds to a White pixel) in order to obtain the QR Code. The last step of this phase involves decoding the QR Code, this can be done using either a smart phone enable with a camera and a compatible decoding application or using the Internet as discussed in Section 2. This section provided a broad picture of the entire methodology discussed in this paper.

5 Summary and Conclusion

This paper presented a methodology that can be used for concealing an encrypted QR Code in a cover image using Steganography. Triple DES algorithm was used for encrypting the QR Code which made it difficult to decrypt it without the key. Apart from that, due to the use of Steganography it is difficult for any third person to notice the existence of the QR Code. The methodology discussed in this paper can be utilized for embedding signatures/copyright information(in form of QR Codes) in an efficient manner into an image, apart from that it can also be utilized in security applications such as the applications used for exchanging confidential information.

References

1. Jose Rouillard, "Contextual QR Codes," iccgi, pp.50-55, 2008 The Third International Multi-Conference on Computing in the Global Information Technology (iccgi 2008), 2008
2. BarcodeContents: A rough guide to standard encoding of information in barcodes Updates June 28, 2010 http://code.google.com/p/zxing/wiki/BarcodeContents
3. Denso-Wave a Toyota subsidiary & The Inventors of QR Code. http://www.denso-wave.com/qrcode/qrstandard-e.html
4. QR Code Features. http://www.denso-wave.com/qrcode/qrfeature-e.html
5. Triple DES Algorithm: http://tools.ietf.org/html/rfc2420
6. Investigator's Guide to Steganography Gregory Kipper Auerbach Publications 2004 Print ISBN: 978-0-8493-2433-8 eBook ISBN: 978-0-203-50476-5
7. Birgit Pfitzmann, Information Hiding Terminology, First Workshop of Information Hiding Proceedings, Cambridge, U.K. May 30 - June 1, 1996. Lecture Notes in Computer Science, Vol.1174, pp 347-350. Springer-Verlag (1996).
8. Neil F. Johnson, Sushil Jajodia, Exploring Steganography: Seeing the Unseen
9. QR-Code Generator, http://qrcode.kaywa.com
10. QR-Code Decoder, http://zxing.org/w/decode.jspx

11. Canadian Government Passport Application Form (see page 6), http://www.ppt.gc.ca/form/pdfs/pptc153.pdf

SESSION

DATABASES, TOOLS, AND VISUALIZATION

Chair(s)

TBA

Three Visualization Tools to Grasp Dynamism in the Global Economy: PRISM, TRADE MAPPER and EMERGENT

Erik Noyes

Babson College
Entrepreneurship Division
Arthur M. Blank Center for Entrepreneurship
Babson Park, MA, USA
001-781-239-5495
enoyes@babson.edu

Deligiannidis Leonidas

Wentworth Institute of Technology,
Department of Computer Science and Systems
550 Huntington Avenue
Boston, MA, USA
001-617-989-4142
deligiannidisl@wit.edu

Abstract - *In this interactive demo-session we present how information visualization can be used as an innovative pedagogical tool in business education to help students grasp dynamism in the global economy. We will demonstrate how data-rich visualizations enable students' understanding of foundational business content, including global competition, innovation-based competition, and industry evolution. The growing field of information visualization examines how data-rich representations of complex phenomena can drive new insights and hypotheses. In the words of visualization researcher Ben Schneiderman, "Information visualization gives you answers to questions you didn't know you had." In a highly visual presentation, participants will be engaged to discuss, explore and critique a visual approach to teaching about changing business environments, the global economy, and especially multi-decade change which suggests new opportunities and risks for business leaders. This research was undertaken to examine ways visual knowledge discovery can improve learning outcomes in business education including technology-enabled learning and distance learning.*

1. Introduction

We will demonstrate three highly interactive visualization tools and present our formal findings. A successful visualization tool is one that enables a user to analyze and query the data efficiently and, as a result, it helps them comprehend the data faster and easier. To accomplish this, one can convert data into a visual representation that allow users to dynamically explore the data so that they can comprehend and explore the data. As others have observed, humans possess great pattern recognition skills especially when it comes to visual representations [1]. Restrictions such as the visual "real estate" of a computer monitor or a sheet of paper can become an obstacle in presenting data in such a way that humans, who are the primary judges of a visualization technique, can observe, query, and comprehend data. Because of this challenge, many visualization techniques present the data to the user as an *overview* of the dataset with drill-down capabilities to see details on demand [2].

Most visualization techniques present the data as an overview of the dataset [3] and include functionalities which enable the user to *zoom* in and out to study the detail of the data. Other visualization techniques offer simultaneous *macro-micro* levels of information, allowing the user to choose varying levels of abstraction. Small multiples--reoccurring visual structures which facilitate a "visual language" within and across visualizations--can provide users grounding points for broader exploration, theorizing, or question framing for a particular visualization. Yet other approaches to visualization give the user full control to explore the dataset one step at a time [4]. With this last technique, users can explore any subgraphs of the dataset but are not able to visualize a full overview of the dataset.

Generally it is best for a visualization to be built around the chief structure, or domain, of the data. For example, data that contains geographic information should generally be presented over a map [5] [6]. Even though the idea to add more dimensions to visualize a dataset seems attractive (i.e., 3-D versus 2-D), research shows that caution must be taken when adding dimensions since the added complexity—of interpretation and/or navigation— can confuse users. However, thoughtfully implemented and tested for effectiveness, visualizations with additional dimensions can create added flexibility and introduce new richness [7].

Below we give a short overview of three visualization tools to grasp dynamism in the global economy – PRISM, TRADE MAPPER and EMERGENT-- which we developed for use in business education. Particularly, each visualization and its interactive capabilities are intended to enable innovative teaching about changing business environments, the global economy, and especially multi-decade change which suggests new opportunities and risks for business leaders.

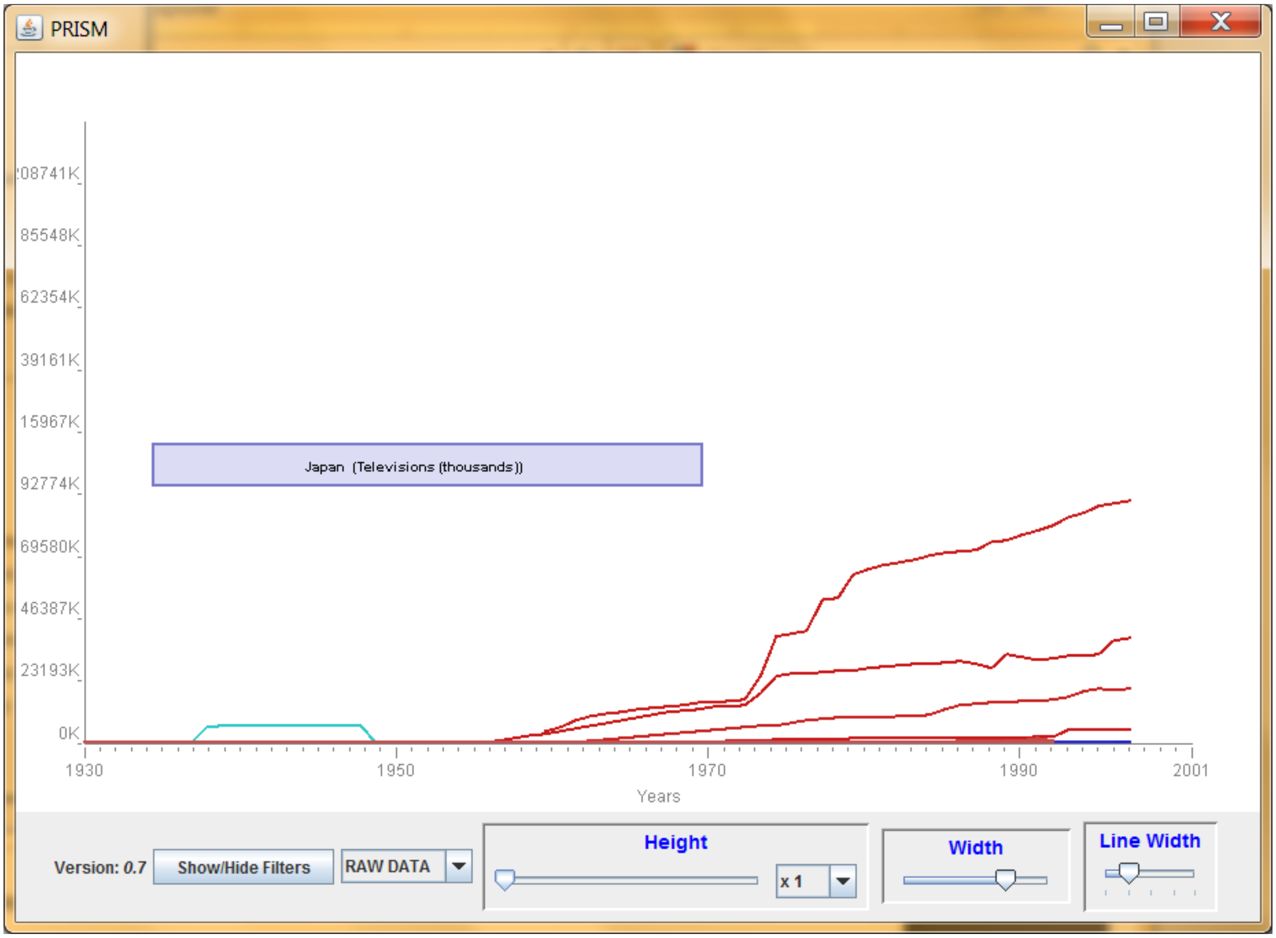


Fig. 1. PRISM is a visual knowledge discovery tool to visualize and understand International Technology Adoption and the innovation “S-curve”

2. PRISM

PRISM, shown above in figure 1, is a knowledge discovery tool for exploring and understanding International Technology Adoption and the innovation “S-curve”. A central challenge in business education and entrepreneurship education specifically, is conveying the dynamism of industries and innovation. Students of business must appreciate the waves of creative destruction which birth industries, redistribute wealth and alter the basis of competition. Core to this objective is understanding patterns of innovation and technology adoption in the global economy (e.g., the rate at which

Mobile Phones and Personal Computers penetrate different world markets). Namely, an accurate historical grasp of the diffusion of innovation in one's industry is central to making accurate business assumptions about the risks and timing of entrepreneurial opportunities, competition, return on investment and the chances of venture success.

This research exploited the largest known international database on technology adoption to develop and test innovative teaching tools for entrepreneurship education, a specialty within business education. The overarching goal was to create interactive visual interfaces to improve teaching on technology adoption and the diffusion of innovation with the aim of improving new venture planning. In entrepreneurship education, technology adoption (also known as the S-curve) is generally taught qualitatively as a cornerstone concept in innovation, competitive analysis and new venture planning. Accordingly, students are rarely, if ever, provided rich data to examine the tempo, implications and varying patterns of technology adoption.

Creating an interactive visual interface, we enabled entrepreneurship students to browse and compare international technology adoption data across the leading 25 industrialized countries from 1788-2001. This included national trends in the adoption of *telegraphs*, *AM radios*, *private cars*, *televisions*, *personal computers*, *mobile phones* and *even industrial robots*. Our head-to-head evaluation of PRISM, versus an Excel spreadsheet with identical data showed that entrepreneurship students make more accurate reflections and future forecasts about technology adoption for a wide range of technologies with our tool. The research was undertaken to examine ways visual knowledge discovery can improve learning outcomes in entrepreneurship education and strengthen entrepreneurship students' conception of innovation dynamics when planning for and launching a new technology venture.

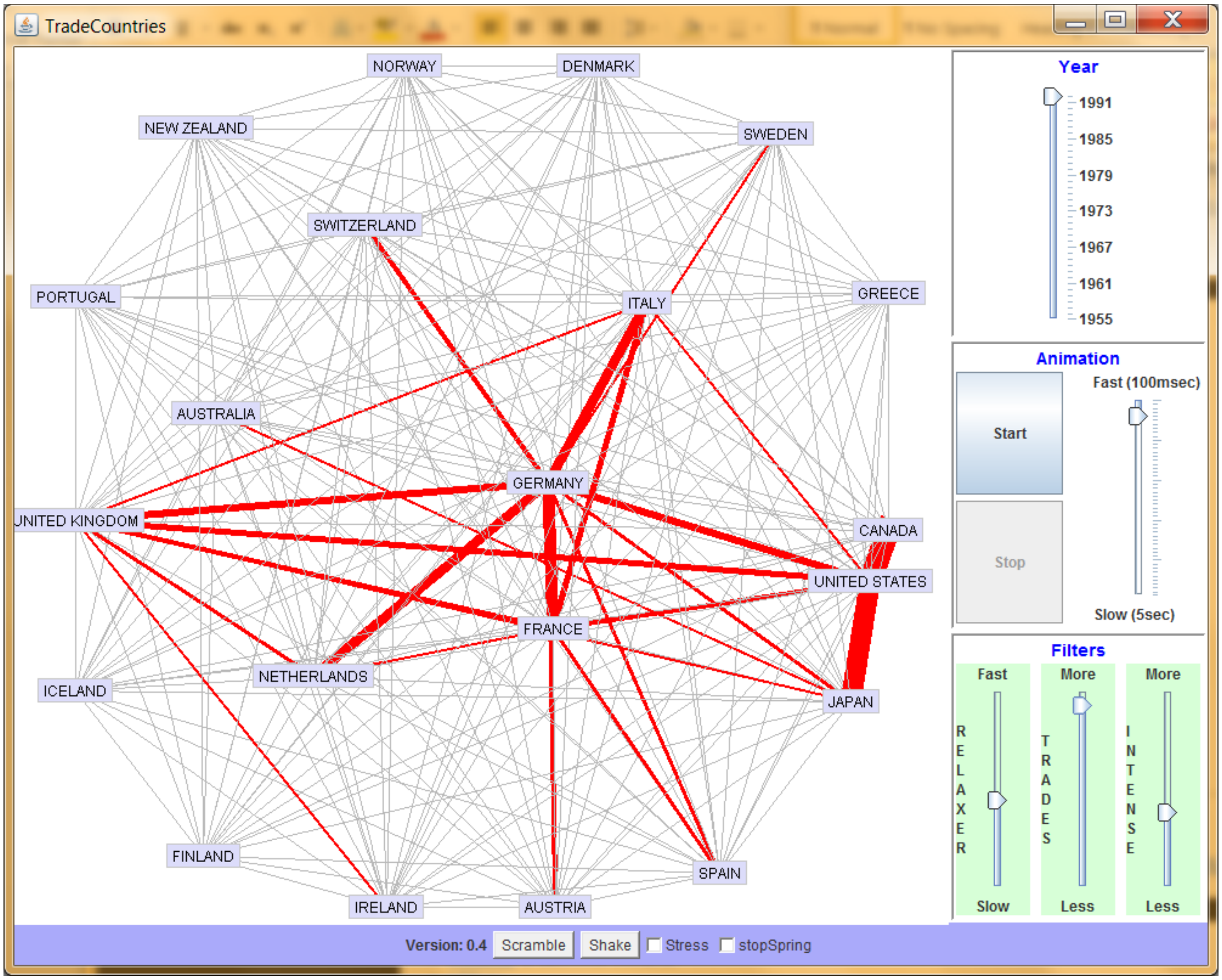
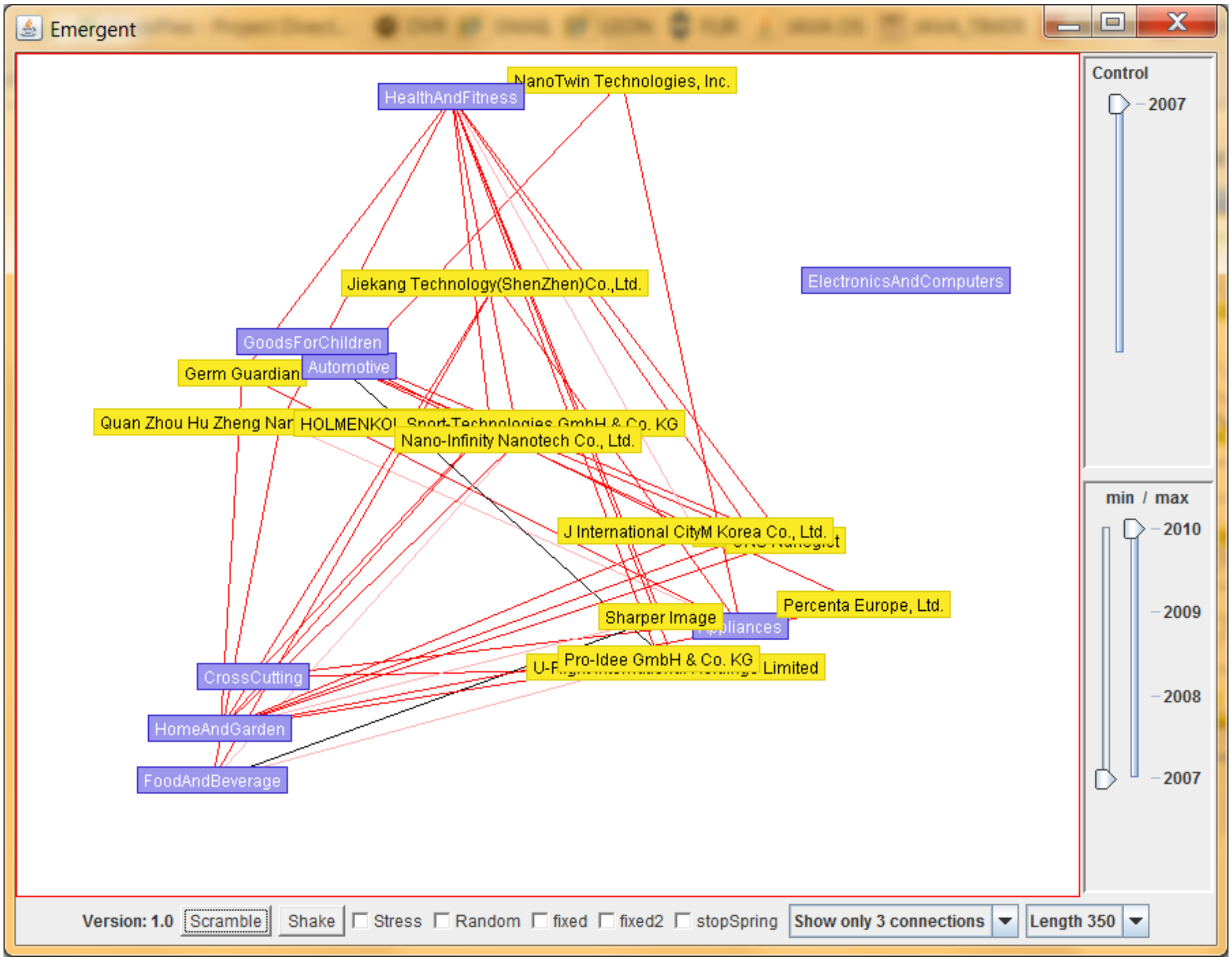


Fig. 2. TRADE MAPPER is a visual knowledge discovery tool to understand shifting trade flows, and the criticality of world actors, in global trade between WWII and the last decade.

3. TRADE MAPPER

TRADE MAPPER, shown above in Figure 2, is a visual knowledge discovery tool to understand shifting trade flows, and the criticality of world actors, in global trade between WWII and the last decade. It provides business students an interactive and intuitive interface to see changing world trade patterns between the top 25 industrialized countries. By applying a network layout algorithm, TRADE MAPPER shows the bi-directional flows and value of trade between the top 25 industrialized countries, where countries which move to the center of the visualization are those rising in importance in world trade in a certain time period. The thickness of the lines, or “ties”, between countries illustrates the total combined value of imports and exports between any two countries. The user can “play a movie” and see shifting patterns of world trade including the emergence of trading blocs for the past 60+ years. Additionally, using sliders, the user can drill down and focus on a snap shot or changing moment in world trade.

This tool was created to give business students an intuitive sense of the dynamism of world trade and show, for example, the shifting importance of the United States in world trade. Where in the 1960’s and 1970’s the United States was unarguably a world hub for trade, now massive flows of trade are moving among specialized trade blocs and the United States is diminishing in overall importance and centrality to the flow of world trade. A network view – as opposed to a standard non-relational line chart of simple trade volumes – adds critical dimensions and interactivity to the typical presentation of trade data. In a classroom or remotely on the web, business faculty and students can explore, discuss and reveal interesting longitudinal patterns in world trade as part of discussion about the changing business environment for industry and entrepreneurial opportunities.



Version: 1.0

Show only 3 connections

Length 350

Fig. 3. Emergent is a visual knowledge discovery tool for understanding emerging industry structure.

4. EMERGENT

EMERGENT [7], shown above in Figure 3, is a visual knowledge discovery tool for understanding emerging industry structure. Particularly, Emergent is an interactive tool to understand the emergence and structure of the nanotechnology industry. According to the U.S. National Science Foundation (NSF), the global nanotechnology industry is expected to grow to $1 trillion in 2010 and drive dramatic innovation and wealth creation in industries as diverse as energy, computing and biotechnology. Nanotech entrepreneurs, analysts and investors alike need tools to understand the emerging structure of the industry because firm competitive positions in the industry impact ventures' survival, growth and profitability.

Overall, the value of the tools is that it shows the relations between all known consumer-focused nanotech ventures in the forming industry and their positions vis-à-vis each other amidst seven subareas of technology commercialization. EMERGENT provides and interactive, dynamic map to see how the industry is taking shape and what the areas of innovation and competition are. Specifically, nanotech ventures compete from different initial "strategic footprints" in the industry which ease or complicate entry into new growth businesses. In fact, the emergence of the nanotechnology industry is the story of interweaving with—and penetration into—different adjoining industries. EMERGENT is able to show these emerging connections which require a network layout of nanotech ventures, nanotech products, and forming markets of the industry.

EMERGENT is interactive, scalable, and adaptable to relational data from other industries. Business applications of EMERGENT include industry analysis, strategic planning and entrepreneurial opportunity identification. As shown in figure 3, EMERGENT consists of the main visualization renderer window and several controls. EMERGENT uses a spring-embedded algorithm to layout the graphs. The controls are used to query the data and render animations of the data based on range of years.

5. Results and Conclusion

Through our user studies we found that students are able to understand and analyze data more accurately with our three visualizations than by using other conventional tools such as Excel. Namely, Excel, the standard data visualization tool of business students and business professionals, cannot render relational data in network diagrams, nor does it facilitate visual knowledge discovery in any fluid sense as it is a time-consuming process to query and render different parts of a data set. Specifically, users of PRISM, we found, were able to answer a series of questions about international technology adoption with far greater accuracy in *less than half the time* of those provided identical data in Excel. While in-class results have been strong, we are in the process of conducting more formal evaluations of TRADE MAPPER and EMERGENT. Overall, while information visualization is a relatively new field, we have compelling evidence that visualization has an important role to play in business education and particularly to enable students' understanding of foundational business content, global competition, innovation-based competition, and the changing global business environment. Given the changing nature and technologies of business education, we are exploring how visualization can expand the reach and impact of business education through technology-enabled learning and distance education.

References

[1] Thomas A. DeFanti, Maxine D. Brown and Bruce H. McCormick, Visualization: Expanding Scientific and Engineering Research Opportunities, IEEE Computer, 22 (8), 1989, pp 12-25.

[2] Card S. K., Mackinlay J. and Shneiderman B., Readings in Information Visualization Using Vision to Think, Morgan Kaufmann, San Francisco, CA, 1999.

[3] Taowei David Wang and Bijan Parsia, CropCircles: Topology Sensitive Visualization of OWL Class Hierarchies, 5th International Semantic Web Conference, (ISWC) 2006.

[4] Leonidas Deligiannidis, Krys J. Kochut, Amit P. Sheth, "RDF Data Exploration and Visualization". In Proc. of the ACM first Workshop on CyberInfrastructure: Information Management in eScience (CIMS'07), pp.39-46, Nov. 9th 2007, Lisboa, Portugal.

[5] Leonidas Deligiannidis, Farshad Hakimpour, Amit P. Sheth, "Event Visualization in a 3D Environment". In Proc. of Human System Interaction (HSI'08), pp.158-164, May '08 Krakow Poland.

[6] T. Kapler and W. Wright, GeoTime Information Visualization, In Proc. of IEEE InfoVis, 2004.

[7] Erik Noyes, Leonidas Deligiannidis, "Emergent: A Knowledge Discovery Tool for Understanding Emerging Industry Structure". In Proc. of the 4th International Conference on Human System Interaction (HSI'11). May 19-21 2011, Yokohama, Japan.

Enhancing the SET Based Data Modeling Method with Context Meta Descriptors

Gregory Vert
Texas A&M Central Texas and
Center for Secure Cyberspace, LSU University
(206) 409-1434
gvert12@csc.lsu.edu

Anitha Chennamaneni
Texas A&M University Central Texas
(254)519-5463
Chennamaneni@tarleton.edu

S.S Iyengar
Center for Secure Cyberspace Computer Science
Louisiana State University
(222) 578-1252
iyengar@csc.lsu.edu

ABSTRACT

Contextual processing is a new emerging field based on the notion that information surrounding an event lends new meaning to the interpretation of the event. Data mining is the process of looking for patterns of knowledge embedded in a data set. The process of mining data starts with the selection of a data set. This process is often imprecise in its methods as it is difficult to know if a data set for training purposes is truly a high quality representation of the thematic event it represents. Contextual dimensions by their nature have a particularly germane relation to quality attributes about sets of data used for data mining. This paper reviews the basics of the contextual knowledge domain and then proposes a method by which context and data mining quality factors could be merged and thus mapped. It then develops a method by which the relationships among mapped contextual quality dimensions can be empirically evaluated for similarity. Finally, the developed similarity model is utilized to propose the creation of contextually based taxonomic trees. Such trees can be utilized to classify data sets utilized for data mining based on contextual quality thus enhancing data mining analysis methods and accuracy.

Keywords
Contextual processing, data mining, taxonomies, data mining quality data sets

1. INTRODUCTION

1.1 Background

(Anitha)

2. CONTEXTUAL PROCESSING

2.1 Background

Contextual processing starts with the notion that data that is not shared often has uncorrelated inferences of meaning and criticalities of information processing in a fashion that truly serves various perspectives needs. Context driven processing is driven by the environment and semantics of meaning describing an event. Often this type of processing requires a context which may contain meta data about the events data. Meta descriptive information of leads to previously unknown insights and contextually derived knowledge. Such meta data usually has a spatial and temporal component to it but is actually much more complicated. The key is that contextual meta data describes the environment that the event occurred in such as the collection and creation of data sets for knowledge mining.

The concept of context has existed in computer science for many years especially in the area of artificial intelligence. Application of this idea has also been applied to robotics and to business process management [1]. Some preliminary work has been done in the mid 90's. Schilit was one of the first researchers to coin the term context-awareness [2,3]. Dey extended the notion of a context with that of the idea that information could be used to characterize a situation and thus could be responded to [4]. In the recent past more powerful models of contextual processing have been developed in which users are more involved [5]. Most current and previous research has still largely been focused on development of models for sensing devices [6] and not contexts for information processing.

Initial development of the context paradigm was based on analysis of the natural disasters of the Indian Ocean tsunami, Three Mile Island nuclear plant and 9/11. Analyses of these events lead to the definition of the dimensions of context. These were categories of contextual data that could uniquely identify a context and determine how it was processed. They are:

temporality – the span of time and characterization of time for an event

spatiality – the spatial dimension of an event

impact – the relative degree of the effect of the event on surrounding events

similarity – the amount by which events could be classified as being related or not related.

Each one of the dimensions can be attributed with meta characterizers which can be utilized to drive processing of context. The temporal and spatial dimensions contain geospatial and temporal elements.

Sample example attributes for these dimensions that might affect the quality of a data set used for data mining could be:

- time period of information collection

- criticality of importance,
- impact e.g. financial data and cost to humans
- ancillary damage of miss classification
- spatial extent data set coverage
- proximity to population centers spatially or conceptually to other related data sets.

Other factors affecting quality classification might be based on the *quality of the data* such as:

- currency, how recently was the data collected, is the data stale and smells bad
- ambiguity, when things are not clear cut – e.g. does a degree rise in water temperature really mean global warming
- contradiction, what does it really mean when conflicting information comes in different sources
- truth, how do we know this is really the truth and not an aberration
- confidence that we have the truth

2.2 Defining a Context

Contextual processing is based on the idea that information can be collected about events and objects such as data sets and that meta information about the object can then be used to control how the information is processed by a data mining methods. In its simplest form, a context is composed of a feature vector

$$F_n<a_1,..a_n>$$

where the attributes of the vector can be of any data type describing the object. Feature vectors can be aggregated via similarity analysis methods, still under investigation, into super contexts $S_{c.}$ Some potential methods that might be applied for similarity reasoning can be any of the methods utilized in data mining that were discussed earlier. The goal of context similarity analysis is to be able to state the following:

$$R(A|B)$$

where:

A, B - are sets of contextual vectors F_n about a data set

R() - is a relation between A and B, s.t. they can be said to be similar in concept and content

Similarity analysis based on data mining methods facilitates the aggregation into super sets of feature vectors describing attributes of a data set based on contextual dimensions describing the data set. This is done to mitigate collection of missing or imperfect information and to minimize computational overhead when processing contexts.

definition: A context is a collection of attributes aggregated into a feature vector describing a abstract event, object or concept.

A super context (set of aggregated contexts) for contextual processing is described as a triple denoted by:

$$S_n = (C_n, R_n, S_n)$$

where:

C_n - is the context data of multiple feature vectors

R_n - is the meta-data processing rules derived from the event and contexts data

S_n - is controls security processing.

However, in definition of super context for data mining purposes there is really not a need for the S_n vector because the sets used for data mining are assumed to be a single location where they are being analyzed.

definition: A super context for contextual data mining is defined by the feature vectors describing the contextual dimensions of the set and the data mining methods applied to the set.

The definition for data mining set quality and selection then becomes:

$$S_n = (C_n, M_n\,)$$

where:

C_n - is the context data of multiple feature vectors

M_n - contain the meta-data processing rules, data mining quality attributes and application methods for analysis quality for the data defined later in this paper.

3. DEFINING SET META CHARACTERIZATION AND QUALITY CHARACTERZATIONS

Contexts describe themes just as data sets do. A theme of a context is the event {tsunami, cancer cluster, etc} that a context describes. Multiple contextual theme may be related and thus candidates for aggregation into a super context. Themes are referred to as thematic objects. Themes may exist with relationship to any of the dimensions of context discussed previously.

definition: A thematic event object (Teo) is the topic of interest for which event objects are collecting data. An example of a Teo would be the center of a tsunami.

As previously defined contexts are defined by four dimensions those of temporality, spatiality, impact and similarity. Contextual objects thus can have *descriptive meta characterizations* based on any of these areas. The quality of a data set used for data mining can also have contextual meta characterization. Some spatial and temporal dimension meta characterizations that can be applied to SET modeling are:

- Singular – an event that happens a point in time, at a singular location
- Regional
- Multipoint Regional
- Multipoint Singular – events that occur at a single point in time but with multiple geographic locations
- Episodic – events the occurs in bursts for given fixed or unfixed lengths of time
- Regular – as suggested these events occur at regular intervals
- Irregular – the time period on these type of events is never the same as previous t
- Slow Duration - a series of event(s) that occupy a long duration, for example the eruption of a volcanoes
- Short Duration – example an earthquake
- Undetermined
- Fixed Length
- Unfixed Length
- Bounded
- Unbounded
- Repetitive - these types time events generate streams of data – graph of attributes change in value over time

In previous work the above meta-characterization of context were classified into semantically based categories that could be utilized in mapping context to data mining quality. The classifications can be utilized with quality metrics in data mining to point the way to methods that might classify quality of data sets based on context. The previously developed categories for the above meta characterization are:

Event Class < abstract, natural>

Event Type < spatial, temporal>

Periodicity < regular, irregular>

Period < slow, short, medium, long, undeterminable, infinite, zero >

Affection<regional, point, global, poly nucleated, n point>

Activity < irregular, repetitive, episodic, continuous, cyclic, acyclic>

Immediacy < catastrophic, minimal, urgent, undetermined >

Spatiality < point, bounded, unbounded >

Dimensionality <1, 2, 3, n>

Bounding < Fixed Interval, Bounded, Unbounded, Backward Limited, Forward Limited, continuous>

Directionality < linear, point, polygonal >

Figure 1. Semantic characterization classification of contextual data that might be utilized to describe sets in the SET model.

The above were developed into a semantic grammar that could can be utilized to form complicated queries for the SET model. Such a grammar could also be developed to classify the quality of data mining sets in a high level qualitative fashion. The syntax of the grammar is of the following:

R1: <event class>, <event type>, <R2>

R2: (<periodicity> <period>) <R3>

R3:(<affection><activity>) <spatiality> <directionality> <bounding> <R4>

R4: <dimensionality> <immediacy>

Figure 2. Syntax for application data meta-characterizations of sets in the SET model.

The above grammar could be developed into complex queries for data from the SET model paradigm. For example, the following might be a semantic descriptor of a data set about an event in statistics, perhaps a cancer data set:

Q' = abstract, spatial & temporal, regular-slow, episodic urgent

In this case *abstract* defines the fact that data may be derived from naturally observable data, that the area of the cancer cluster occurs in "*spatial and temporal*" areas of the country, that the development of the cancers occur regularly (such as skin cancers in the southwest, and that the *urgency* of the data needs to be considered. Considering a semantic rule describing another data set, it might be described in the following fashion:

Q'' = abstract, singular event, point, undetermined

One can deduce that the rules could have very different relations with SET or the way data sets are processed or selected based on SET. While these are mappings of contextual processing concepts onto SET suggest that this technique can be done, it does not really incorporate the issues of quality of data being managed in the model.

We have defined measures SET quality defined as the following [cite dm book]:

- Relevance (Re) - degree of relationship of data to a theme
- Timeliness (Ti)- temporal proximity of the data to it T_{eo}
- Noise (No) – the degree to which data is observed versus injected by observational equipment.
- Outliers (Ou) – observed actual data that exceeds the norm

- Sparsity (Sp)– a binary representation of known data versus unknown data in a matrix of observed data
- Dimensionality (Di) – the number of observed attributed about a T_{eo}, of which some may be more relevant for analysis
- Freshness (Fr)– proximity of collection to a point in time
- Accuracy (Ac)– the degree to which the data in the set reflects reality
- Sequentiality (Se) –
- Bias (Bi) – qualitative versus quantitative
- Duplication (Du) – a characterization of some data appearing to be the same and representing different objects, other times being the same and representing the same object
- Aggregation (Ag) – the degree to which data is combined where increased aggregation produces better stability in the data but loses granularity.

In considering the development of a method that could be used build contextually based taxonomic trees for SET, we mapped the above defined SET attribute onto the dimensions of the contextual model paradigm. This produced the following mapping:

$$f\,(Re) \rightarrow (T,\ Sp,\ Im,\ Si)$$
$$f\,(Ti) \rightarrow (T,Im)$$
$$f\,(No) \rightarrow (\ Im,Sp,\ Am)$$
$$f\,(Ou) \rightarrow (\ Im,\ Sp,\ Am)$$
$$f\,(Sp) \rightarrow (\ Im,\ Si,\ Am)$$
$$f\,(Di) \rightarrow (\ Sp,\ Im,\ Am)$$
$$f\,(Se) \rightarrow (T,\ Si)$$
$$f\,Bi) \rightarrow (Am)$$
$$F\,(Du) \rightarrow (Am,\ Si,\ Im,\ Sp)$$
$$F\,(Ag) \rightarrow (Am)$$

where:

Sp	- spatiality dimension
T	- temporality dimension
Si	- similarity dimension
Im	- impact dimension
Am	- the ambiguity dimension

$f\,(x) \rightarrow (y1,\ y2..)$ – is a function that maps SET attributes onto multiple dimensions of the context model

In development of this analysis, a new contextual dimension became necessary in order consider an aspect of SET, that of ambiguity.

The above mapping supports the next step in potentially useful method for contextual sets. This method requires understanding the relationship among contextual dimensions based on the mapping of SET attributes to sets of contextual dimensions. Reversing the mapping of SET onto contextual dimensions produce the following:

$$f\,(T) \rightarrow (Re,\ Ti,\ Sei)$$
$$f\,(Am) \rightarrow (No,\ Ou,\ Sp,\ Di,\ Du)$$
$$f\,(Sp) \rightarrow (Re,\ No,\ Ou,\ Di,\ Du)$$
$$f\,(Si) \rightarrow (Re,Sp,\ Du)$$
$$f\,(Im) \rightarrow (Re,\ Ti,Ou,Sp,\ Di,Du)$$

(gv why ?)

In the reverse mapping the SET attributes are mapped onto the four original dimensions of context including the new one of *Ambiguity*. Of note, the evaluation of how any single SET in the mapping is evaluated quantitatively is the subject of future research.

4. Extending the SET Model with the Context Model

With this mapping, the model for contextual management of SET models is presented. The original model [cite vert] was developed with notions that time and space would be descriptors of the sets managed in the set model. It is a standard ERD presented as shown in figure n.

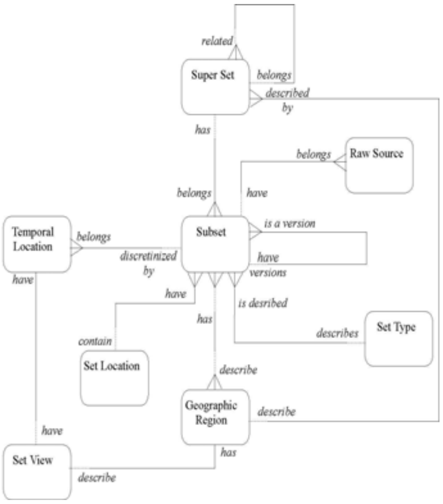
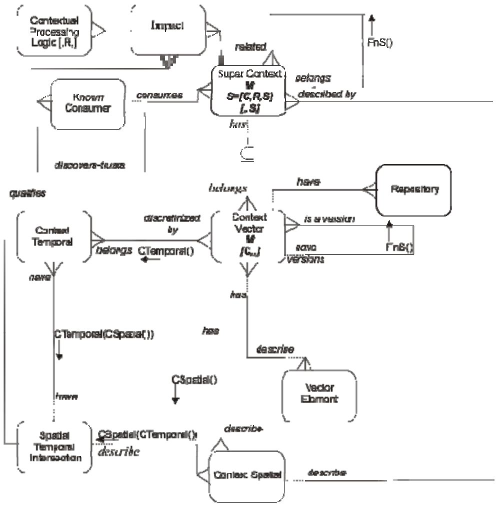


Figure n Original model for SET management of spatial data

Because the sets can have members whose extents overlap spatially and in time, the above model had to be extended with operators using fuzzy set theory to describe relations among overlap for the purposes of query selection Cspatial() and Ctemporal() [vert cite]. Arrows in the model indicate the direction a selection function is applied.

Upon the development of the contextual model, this old model has been revised to include the notions of the dimensions of context as they apply to the SET model

The first addition to the original model of note is that of the notation of **M** being added to the model. This notation could have been place in many entities but is particularly relevant to the Context Vector entity. **M** means that the entity *primarily* contains meta-data, in this case data about a given Context Vector. The logic is that the *Vector Element* and *Repository* entities contain the actual raw data therefore Context Vector only contains descriptions of this data. Because *Super Context* is a collection of *Context Vectors*, the logic follows that it also is a meta-data entity. Of note, in the listed super context tuple [C,R,S], this entity also becomes the architectural location for security information which is discussed in subsequent chapters.



The relationship between *SuperSet* and *Set* in the original model had a subset notation and the name of the entities have been changed in the new contextual model. Additionally a weighting operator (OWA) represented by a summation sign has been removed because a super context only consists of all the context vectors that aggregate to define the super context. The SuperSet entity has now become *Super Context* and the original Set entity is now Context Vector. *Context Vectors* by definition are not unique entities. When considered with the entities with which they have relations with, they can become unique.

A subset of a S*uper Context* data has an unusual property in that this entity is not unique in itself. It becomes unique when the fuzzy temporal, spatial and impact relations around it are considered. There can be multiple physical files containing context data and rules that a subset may represent

This relation can be characterized by the existence of multiple *Vector Element*s of heterogeneous data formats that cover the same area but may be of different scale or perspective. Each one of the Context Vector may cover or be relevant to a minute area part of the spatial coverage a *Super Context* and is therefore a subset. Additionally, the *Context Vector* coverage's may not be crisply defined in the spatial sense. They may also cover other areas defined as part of other partitions in the superset. This leads to the property of sometimes being unique and sometimes not being unique.

The *Context Spatial* entity has a $C_{Spatial}()$ relationship with Context Vector. The rationale is that because context *Vectors* of data can be overlapping in spatial coverage for a given point in space, selection of a subset becomes an ambiguous problem. The $C_{Spatial}()$ symbol then implies that selection of *Context Vectors* covering a geospatial point needs to done using some type of fuzzy selection. Of note, the directional indicator means that given a *Context Vectors* data, the $C_{Spatial}()$ function is executed on *Context Spatial* to find all the spatial context data that may apply to a Context Vector.

The *Temporal Region* entity exists because there is a need for given data sets to map to various locations in time and spatial coverages. The relationship "discretized" was originally defined to map a value to a continuous field. In this case the discretenized relation has been extended to represent a discretized function where the continuous field is time. Because this function can relate a data set to various points in time and coverage of several different spaces, this function is a fuzzy function, $C_{temporal}()$ that selects on time and spatial definitions for a *Context Vector*.

New in the model is that the dimension of Similarity is applied as a directed fuzzy function on relationships. This allows Super Contexts and Context Vectors to be associated. Of particular use is that Super Contexts can be similar to other instances thus multiple types of thematic objects (mentioned previously) can be supported in the model.

Vector Element is a new entity in the model. Because a Context Vector can be composed of infinite numbers of elements, there is a need to describe each of the elements and potential processing of idiosyncrasies of its processing. *Vector Element* is where this information would be placed.

Repository is a location where the physical storage structure might be described. Such a structure maps the abstraction of set management onto a physical system. Systems that might be described here are the type of database (ORACLE, DB2) or file or file locations.

The entity Spatial Temporal location is an intersection entity that is meant to capture the notion that contextual data can only be uniquely identified by its intrinsic descriptors of the time and the location, space, that the thematic object described by a *Super Context* existed in. Such a tuple can be assigned something like a GUID in this entity but it is the temporal and spatial identifiers that localize the *Super Context*s theme.

Known Consumer is the place where hyper-distribution information for contextual data would be stored. On the recursive relation, the concepts described in the chapter about the discovery and trust of consumers in their process of getting to know each other for hyperdistribution, is supported

Fuzzy theory literature [2] defines a membership function that operates on discrete objects. This function is defined as similar and has the following property:

$$
\begin{array}{ll}
 & \{\ 1 \quad |\ \text{if a domain(A)}\ \} \\
FnS() = & \{\ 0 \quad |\ \text{if a domain(A)}\ \} \\
 & \{[0,1]\ |\ \text{if a is a partial member of domain(A)}\}
\end{array}
$$

This function expresses a dimension of contextual processing that of similarity and thus provides a framework to reason about impact. The function is also useful in contexts and super contexts where overlapping coverages of the same event space or time may exist, but some coverage for a variety of reasons may be more relevant to a particular concept such as a desire to perform editing of surrounding regions. Similarity can also be none spatio-temporal, thus this function on a relationship can also model thematic similarity of what might appear to be disjointed super contexts. The actual definition of how partial membership if calculated has been the subject of much research including the application of Open Weighted Operators (OWA) [3,10] and the calculation of relevance to a concept [6]

Finally, the end product of most of the discussion about contextual processing is the processing part. With the dimensions of contextual processing supported in the model, and tied together through relations, the Contextual Processing Logic entity is where the semantic rules for a given Impact would be contained. This like the *Repository* is where there is a physical interface to the outside world where users define the specific actions for there systems for a given set of contextual reasoning and semantic meaning.

5. CONCLUSIONS

The ideas and concepts introduced in this chapter offer a one potential method to manage contextual information. They should be examined and enriched with more rigorous definition. Additionally there is potential to define metrics that might express such things as:

- Confidence in relationships
- Trust in relationships
- Model quality metrics

The model proposed in this chapter is architectural in nature and thus not attributed. The process of determining the meta-data to attribute the model could be an active and vibrant research area pursued with the question of how correct attribution will be determined. One potential approach to this question is to use an established standard such as the geographic and military community currently have as a starting place.

Development of the retrieval mechanism might be another strong area for research. In this vein, definition and refinement of the fuzzy selection and relation mechanisms could be pursued. There will be a need for development of the mathematics behind the concepts. Such equations probably should have an adaptive mechanism making them possible to change behavior based on the context of environmental data. This could lead to the research areas of contextual set management and contextual retrieval of information.

6. REFERENCES

1. Rosemann, M., & Recker, J. (2006). "Context-aware process design: Exploring the extrinsic drivers for process flexibility". T. Latour & M. Petit *18th international conference on advanced information systems* engineering. proceedings of workshops and doctoral consortium*: 149-158, Luxembourg: Namur University Press.*
2. Schilit, B.N. Adams, and R. Want. (1994). "Context-aware computing applications" (PDF). *IEEE Workshop on Mobile Computing Systems and Applications (WMCSA'94), Santa Cruz, CA, US*: 89-101.
3. Schilit, B.N. and Theimer, M.M. (1994). "Disseminating Active Map Information to Mobile Hosts". *IEEE Network* **8** (5): 22–32. doi:10.1109/65.313011.
4. Dey,Anind K. (2001). "Understanding and Using Context".*Personal Ubiquitous Computing* **5** (1): 4–7. doi:10.1007/s007790170019.
5. Cristiana Bolchini and Carlo A. Curino and Elisa Quintarelli and Fabio A. Schreiber and Letizia Tanca (2007). "A data-oriented survey of context models" (PDF). *SIGMOD Rec.* (ACM) **36** (4): 19--26. doi:10.1145/1361348.1361353. ISSN 0163-5808. http://carlo.curino.us/documents/curino-context2007-survey.pdf.
6. Schmidt, A.; Aidoo, K.A.; Takaluoma, A.; Tuomela, U.; Van Laerhoven, K; Van de Velde W. (1999). "Advanced Interaction in Context" (PDF). *1th International Symposium on Handheld and Ubiquitous Computing (HUC99), Springer LNCS, Vol. 1707*: 89-101.
7. Tan, Pang-Ning, Steinbach, M., Kumar, V., Introduction to Data Mining, pp 27-43, Addison Wesley, 2006.
8. J. Han and M. Kamber, *Data Mining: Concept and Techniques*, Morgan Kaufmann Publishers, 2001.
9. V. Barnett. *Outliers in Statistical Data*. John Wiley, 1994.
10. J.W. Tukey. *Exploratory Data Analysis*. Addison-Wiley, 1977.
11. Edwin M. Knorr and Raymond T. Ng. Algorithms for mining distance-based outliers in large datasets. In *VLDB*, pages 392-403, 1998.
12. Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and JÄorg Sander.LOF: Identifying density-based local outliers. In *SIGMOD Conference*, pages 93-104, 2000.
13. Martin Ester, Hans-Peter Kriegel, JÄorg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise.In *KDD*, pages 226{231, 1996
14. R. Agrawal, T. Imielinski and A. Swami, Data Mining: A Performance Perspective, *IEEE Transactions on Knowledge and Data Engineering* **5**(6) (1993), 914–925.

XMLattes – A Tool for Importing and Exporting Curricula Data

Gustavo de O. Fernandes [1], Jonice de O. Sampaio [2], and Jano M. de Souza¹

[1]PESC/COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Rio de Janeiro, Brazil

[2]DCC/IM, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Rio de Janeiro, Brazil

Abstract *- The Brazilian scientific scenario has been substantially modified in recent decades with the emergent new agencies for promotion and development of new research areas. Which research projects would be awarded with funding is a question that can only be answered through an (quantitative and qualitative) analysis of publicly available data. Such data is available in the Lattes web platform, but it can not be easily manipulated by software. This paper presents an approach to retrieve the textual data from researchers registered in the Lattes platform, preprocessing it and exporting it in XML format, making easier their subsequent manipulation. In order to validate this tool, a study case was applied and its results are also listed in this paper.*

Keywords: Lattes Curriculum, Importing, Exporting, XML, Database.

1 Introduction

New relationships between academic entities are created or changed every day. The need to understand the dynamics of how this process occurs, as well as to make important decisions (such as scholarships selection to encourage scientific research), respect the analysis of this scenario, which includes researchers, teachers and students across the country.

This analysis, both quantitative and qualitative, must be based on publicly available and easily accessible data, and a large body of data that fulfills these requirements is the Lattes curriculum.

The Lattes curriculum is part of Lattes Platform, an initiative of the Brazilian federal government to centralize data about researchers and teaching institutions, in order to analyze quantitatively and qualitatively the entities involved. However, the data available in a Lattes curriculum is available in only in two ways:

- Through a HTML formatted file, available publicly on the Internet;

- Through a XML formatted file, available via direct agreement with the CNPq (*Conselho Nacional de Desenvolvimento Científico e Tecnológ*ico - National Council for Scientific and Technological Development).

A manual analysis from the HTML document retrieved from the former can bring, undoubtedly, a quick and easy result. However, by doing this same operation with tens or hundreds of documents is a clear obstacle to any project. Moreover, XML-formatted files are available only to CNPq cooperating institutions - and one can only request data from the researchers of his same institution. Therefore we conclude that the first method is not a 100% easy way of doing this operation, while the second one does not allow access to all of the data available.

This paper presents the XMLattes tool, which aims to import data from selected researchers within Lattes curricula platform and make it available in XML format, thus making manipulation by third-party programs easier.

The sections are organized as follows: section 2 describes related work with the same purpose of the tool here presented; the third section summarizes the development effort and the technology used; section 4 concerns about the tool itself, emphasizing the module that actually filters the imported data; section 5 reports the case study to evaluate the XMLattes and, finally, section 6 presents the conclusions of this work.

2 Related works

The proposed tool takes into account the effort already expended by other researchers and students who want (and need) to import data from Lattes curricula.

Recently, many works were based on data taken from the Lattes Platform ([4], [5] and [6]) and some researchers, due to the absence of an auxiliary tool, reported a delay of weeks, months and even years to manual retrieval of information [7].

The data extraction tool developed by CNPq, LattesExtractor [11], is available on the web interface and exports the selected fields of in XML format, but is only accessible and licensed to institutions which have signed CNPq cooperation agreements. The researchers from these institutions, however, do not have access to all curricula of the platform: only the data from researchers, teachers, students and employees of their own institution can be read [4].

Aiming to fill this gap in provision of data, some independent initiatives have sprung up around the country, highlighting ScriptLattes [12], PPGI-GSTP [9] and API LattesMiner [3], described below.

2.1 ScriptLattes

The ScriptLattes [8] is a PERL script to extract the bibliographical productions, guidelines and general data from selected researchers. Developed by the Department of Computer Science / University of São Paulo (DCC/USP) under the GNU-GPL [13], it has as its main feature the reports with academic production numbers. Therefore, no files are generated for each researcher - just an HTML page containing the all information consolidated.

2.2 PPGI-SGPC

The PPGI-SGPC was produced by a group of researchers from the Graduate Program in Computer Science (under Center for Computing and Electronic / UFRJ) and designed to facilitate the task of managing the development of post graduate programs and allow greater sharing and dissemination of scientific production [9]. The operation of the system, already finished and available to the academic community, requires as an input the XML files containing the data of researchers monitored (provided by LattesExtractor). Thus, this tool has as its main focus the intra-department management and does not consider departments outside an institution.

2.3 LattesMiner

The latest work found in the literature is LattesMiner API, which is also the most similar to the one proposed here. Developed in Java, provides an API (provides advanced users the ability to expand and customize the initial functions) to import any Lattes curriculum (available on the web in HTML format) in an XML format. Moreover, it generates a social network illustrating the relationship of academic researchers already imported. However, at the time of this tool design, the INPE / ITA group (responsible for this software) had not finished it nor announced further details.

2.4 Comparisons

As noted, XMLattes and LattesExtractor show themselves as auxiliary tools for generating reports, as they propose to provide the formatted data in an XML file used as input into other systems responsible for the consolidation of information and the analysis itself. The API LattesMiner stands by the promise of producing, addition to the XML file containing curriculum vitae, information concerning the formation of social networks.

3 Development

The development of this tool occurred in Database Laboratory (LabBD), which integrates the Systems Engineering and Computing Program (PESC) / Federal University of Rio de Janeiro (UFRJ). Also in this laboratory, we developed the independent web system “Scientific Knowledge Management” (GCC) [1], which has some entities (database tables) in common with XMLattes.

The following summarizes the choice of some technologies necessary to implement this project.

3.1 The Platform

Today we come across several development platforms composed by a wide variety of programming languages. Among the main, we can name. NET (C #, C, VB.NET, etc.)., Java, Python and Ruby. The choice of development platform, however, was not a difficult task, considering the advantages / disadvantages of each. Next, is justified the choice of platform chosen.

The option for the Java platform is justified in several ways. The first is that all development is free. As Oracle / Sun [14] provides free both the Java VM (JRE) as the Platform Development Kit (JDK), the cost of development was restricted to the cost of infrastructure (provided by the LABBD).

The second point is that this technology has a model already successful and widespread. It is unknown *a priori* which hardware and software will be available for the end-user, so the advantage of Java being platform independent was decisive.

Besides these, Java also has a third and very important advantage: its vast collection of libraries and API's - mostly free. Used by the development stage, APIs like Swing [15] (used for rapidly building graphical interfaces) and Log4J [16] (organization to source code, bug tracking and logging customizable) proved to aid excellently the development, making efficient testing processes and interface design.

Furthermore, the NetBeans IDE (as Eclipse IDE) is a consolidated, stable and an excellent choice as it is maintained by a giant software manufacturer (Sun Microsystems) that invests significant resources for its maintenance [2].

3.2 Technologies

Besides Java platform, already justified, other technologies were alto helpful to develop this work. In particular, XML and regular expressions were of utmost importance and will be briefly reviewed here.

3.2.1 XML

According to W3C [17], XML describes a class of data objects called XML documents and describes the behavior of software that will process them. Generally speaking, XML documents are, actually, SGML (Standard Generalized Markup Language) defined by ISO 8879 [10].

Being a neutral and open format, widely and widespread used for interface systems, the XML standard was chosen to be the primary input / output data format.

Furthermore, it is also the format used by the exporter of CNPq (LattesExtractor), and use the very same data format that such a system represents, undoubtedly, is a good idea, since its output data should be used as input in XMLattes.

The data model [18], defined by the community LMPL [19], however, proved to be much more complex than necessary for the scope of this work. We decided then to simplify this model [20], since the HTML pages of the curriculum (the main source of data XMLattes), for example, underlies many of the data that are part of the DTD.

3.2.2 Regular Expressions

A deficiency of Lattes (appointed by related work [3], [9]) is the lack of standardization in filling some of its text fields. Thus, researchers refer to awards, entities or papers almost freely. For example, the following set of expressions (taken from an HTML page Lattes) represents the same institution:

- <td>Universidade Federal do Rio de Janeiro</td>;
- <td>Univ. Fed. do Rio de Janeiro</td>;
- <td>Universidade Fed. do RJ</td>;

Identifying the semantic similarity of these expressions is trivial for a human being, but may not be for a computer program. Moreover, one should also consider that these expressions retrieved from Lattes platform are formed by a set of data and metadata (HTML tags and <td> </ td> represent the meta-data in this example).

The use of regular expressions as a mechanism for comparison of strings during the import process data was essential to concept of the tool, since regular expressions apply in practice concept of regular languages [23].

The basic idea is, given an initial string, try to find a substring using a character pattern. By applying a mask (ie: the regular expression in itself), a set of characters can be found and reveal the pattern in the original string of characters.

Considering the example given at the beginning of this section, the expressions representing an institution would be easily identified by the regular expression "<td> .*?</ td>", where the HTML tags delimit the beginning and the end of an expected substring representing the name of an institution.

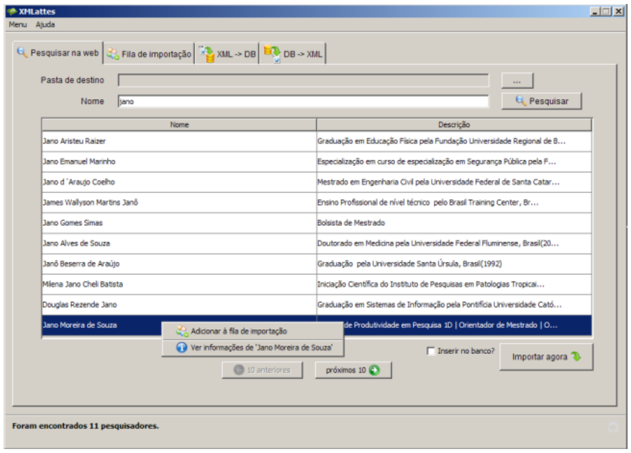
4 The XMLattes

As already stated, the main purpose of the tool is to provide an easy and efficient way to extract Lattes curricula data to DBs and file systems. For the latter, the XML format was chosen because its recognized importance for data portability and being multiplatform. Through XMLattes is possible:

- Look for researchers;
- Download the data from their curricula to the local file system (individually or in groups);
- Extract the data from DB to local file system;
- Insert data from local file system to a database.

Aiming to increase end-user’s usability, the tool is divided into graphical modules. However, it is important to note that these modules do not represent the code division into packages (Java way of grouping code and files). Namely, the packages are:

- DataRetriever: imports and searchs resumes - the main tool package;
- DBManager: manages database connection;
- Entity: entities that represent a logical abstraction of the tables used in the GCC;
- Util: common methods to other modules within XMLattes;
- GUI: GUI through which user access the features to XMLattes. **Erro! Fonte de referência não encontrada.** illustrates one of the tabs available in the GUI - the search screen.



Pasta de destino

Nome: jano

Nome	Descrição
Jano Aristeu Raizer	Graduação em Educação Física pela Fundação Universidade Regional de B...
Jano Emanuel Marinho	Especialização em curso de especialização em Segurança Pública pela F...
Jano d´Araujo Coelho	Mestrado em Engenharia Civil pela Universidade Federal de Santa Catar...
James Wallyson Martins Janô	Ensino Profissional de nível técnico pelo Brasil Training Center, Br...
Jano Gomes Simas	Bolsista de Mestrado
Jano Alves de Souza	Doutorado em Medicina pela Universidade Federal Fluminense, Brasil(20...
Janô Beserra de Araújo	Graduação pela Universidade Santa Úrsula, Brasil(1992)
Milena Jano Cheli Batista	Iniciação Científica do Instituto de Pesquisas em Patologias Tropicai...
Douglas Rezende Jano	Graduação em Sistemas de Informação pela Pontifícia Universidade Cató...
Jano Moreira de Souza	de Produtividade em Pesquisa 1D | Orientador de Mestrado | O...

☐ Inserir no banco?

Foram encontrados 11 pesquisadores.

Figure 1 - GUI tool (search screen)

In Figure 2, you can see how these modules communicate with each other through the established architecture.

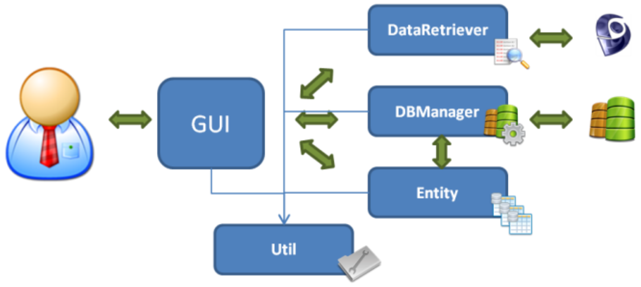


Figure 2 - Modules communication

It is described next how DataRetriever (the main module) works.

4.1 DataRetriever

DataRetriever implements the search for a researcher (restricted by a name), imports the data from a curriculum (identified by a Lattes key) in HTML format and parses the data. In this context, the term "parsing" (and its derivatives) does not take into account the formal definition of the context of computation (parsing a string in order to determine its grammatical structure in accordance with a predefined formal grammar). In other words, the term "parsing" as used in this context considers the analysis and extraction of data from a document (more specifically a text document in HTML format) using methods of string manipulation, without a predefined formal grammar. In Figure 3, follows a graphical simplification of the parsing process used in this work, which will be described throughout this section.

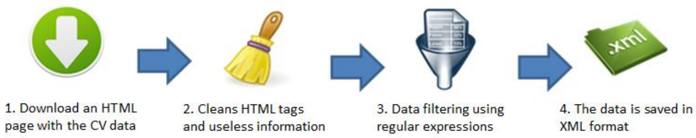

1. Download an HTML page with the CV data
2. Cleans HTML tags and useless information
3. Data filtering using regular expressions
4. The data is saved in XML format
xml

Figure 3 - Parsing process of HTML formatted Lattes curriculum

The DataRetriever package is critical to the overall efficiency of the tool, since the data of a given researcher will be extracted through it. Such data is inserted into the database for further analysis and therefore this step would generate inconsistencies analyzes incorrect or ill-founded. However, considering the lack of standardization of data entry in the Lattes platform, people from different research areas (as well as different geographic regions) have different standards for data entry. Ideally, it should provide a process of parsing data that would, in most times, correctly extract the desired data.

The structure of this package is based on three classes: ResearcherFinder, DataRetriever and Filter. The interaction between them is given as follows:

1) User performs a search by name using the ResearcherFinder;

2) DataRetriever downloads and filter page one of the researchers returned by the search in (1);

3) Through static methods defined in Filter, data is filtered and structured.

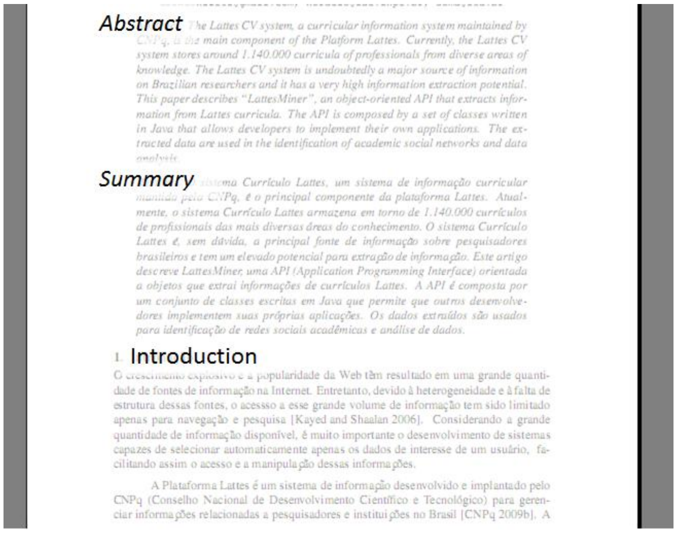
Next is explained how each class works.

The first, ResearcherFinder, aims to address the functionality to search by a specific name. The web-search of researchers maintained by CNPq is used as data source and the main method of this class uses this interface to mediate the searches made by XMLattes. As might be expected, the user provides a name as input and the search method returns a list of the researchers found. As the web interface returns CNPq groups of up to 10 results per page, the list returned by the class is also limited to 10 items. Moreover, the class maintains the total number of results and sends the desired page number, enabling paging general result of our search for GUI XMLattes. Each of the pages returned (in HTML format) with the search result is a process of parsing and only the names, keys and descriptions of the researchers are returned by search method.

The second, DataRetriever, which downloads the HTML document which contains the data of selected researcher and it cleans up unnecessary data (thus reducing the processing required in the next step) and converting HTML entities (ie HTML codes representing stress and some characters punctuation). This removal of unnecessary data proved to be very efficient by eliminating over 40% of the size of the original HTML page - in some cases, over 70%. Naturally, the time required to process the researcher's data is greatly reduced, since the parsed text size is much smaller than the original.

After cleaning, DataRetriever indexes desired data types ("Personal Data", "Academic Data", "Bibliographic production", etc..): the main idea is to divide the entire page into small pieces and then apply specific filters in each particular piece. This concept can be better understood by viewing Figure 4.

The image represents a document (precisely one of the articles used as reference for this work). In this document, consider the need to extract the fields Abstract, Summary and Introduction - which are known to be in order. We can assume that the contents of each field ends when the next field contents begin. When we find the occurrence of these fields in the original text, the positions **x**, **y** and **z**, respectively, we assign indexes to these coordinates. So logically, first field contents is delimited by [x, y-1], second one by [y, z-1] and so on.



Abstract The Lattes CV system, a curricular information system maintained by CNPq, is the main component of the Platform Lattes. Currently, the Lattes CV system stores around 1.140.000 curricula of professionals from diverse areas of knowledge. The Lattes CV system is undoubtedly a major source of information on Brazilian researchers and it has a very high information extraction potential. This paper describes "LattesMiner", an object-oriented API that extracts information from Lattes curricula. The API is composed by a set of classes written in Java that allows developers to implement their own applications. The extracted data are used in the identification of academic social networks and data analysis.

Summary O sistema Currículo Lattes, um sistema de informação curricular mantido pelo CNPq, é o principal componente da plataforma Lattes. Atualmente, o sistema Currículo Lattes armazena em torno de 1.140.000 currículos de profissionais das mais diversas áreas do conhecimento. O sistema Currículo Lattes é, sem dúvida, a principal fonte de informação sobre pesquisadores brasileiros e tem um elevado potencial para extração de informação. Este artigo descreve LattesMiner, uma API (Application Programming Interface) orientada a objetos que extrai informações de currículos Lattes. A API é composta por um conjunto de classes escritas em Java que permite que outros desenvolvedores implementem suas próprias aplicações. Os dados extraídos são usados para identificação de redes sociais acadêmicas e análise de dados.

1 Introduction

O crescimento explosivo e a popularidade da Web têm resultado em uma grande quantidade de fontes de informação na Internet. Entretanto, devido à heterogeneidade e à falta de estrutura dessas fontes, o acessso a esse grande volume de informação tem sido limitado apenas para navegação e pesquisa [Kayed and Shaalan 2006]. Considerando a grande quantidade de informação disponível, é muito importante o desenvolvimento de sistemas capazes de selecionar automaticamente apenas os dados de interesse de um usuário, facilitando assim o acesso e a manipulação dessas informações.

A Plataforma Lattes é um sistema de informação desenvolvido e implantado pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) para gerenciar informações relacionadas a pesquisadores e instituições no Brasil [CNPq 2009b]. A

Figure 4 - Example of a document being index-partitioned

That is the main idea used to index the plain text (HTML) obtained by downloading one of Lattes. However, some curricula are not fully satisfied (e.g.: a researcher who had not received awards will not have the "Awards and Titles" field filled), which would represent an obstacle to our initial assumption. In order to maintain the algorithm's flexibility, a modification is required, which was implemented with the following piece of algorithm:

1)Make up a list containing the keywords (terms) that surround the fields to be filtered.

2)For each item in this list, is verified if it occurs in the original text.

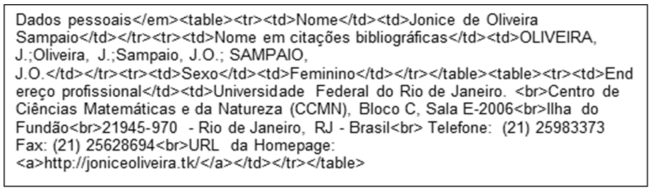
   a) If so, is recorded the position (index) of this occurrence.
   b) Else, it is considered the position (index) of the next term - if there is no next term, is considered the end of the document.

By this reasoning, we can get the contents of each of the fields (or sections) that compose the Lattes curriculum. But we still have not extracted the specific data from each of these fields.

The third and last class of this package, the class Filter, implements the necessary filters to extract specific data from Lattes curriculum fields. All methods of this class are static and they are responsible for applying filters based on regular expressions to the data retrieved from the HTML page. By using independent filters arranged sequentially, the parsing process is not interrupted even if one of the types of data is not found or filtered properly.

We decided to create this third class (instead of adding its methods to class DataRetriever) because we believe this approach would facilitate further code maintenance. Moreover, such architecture would encourage the idea that the filters would work as plugins: if a third-party application makes use of XMLattes, the developers could adapt the Filter class at their will, making the use of the search module more flexible. As LattesMiner API[3], XMLattes allows its users to create custom methods to extract the data by themselves.

It is exactly at this phase of the import process that regular expressions prove its usefulness. Each field is composed by a singular form, but each of the elements that make these items up follows a pattern. This remark by itself would justify the use of regular expressions, but there's also the fact that these elements are filled with metadata from the HTML of the original document. It is worth noting that some of the expressions are simply words or literal phrases, since some of the fields are pre-defined data (e.g.: the type academic training may be "Basic Education", "High School", "Graduate", etc.). Take for an example the personal data field of a researcher, after the cleanup phase.



```
Dados pessoais</em><table><tr><td>Nome</td><td>Jonice de Oliveira
Sampaio</td></tr><tr><td>Nome em citações bibliográficas</td><td>OLIVEIRA,
J.;Oliveira, J.;Sampaio, J.O.; SAMPAIO,
J.O.</td></tr><tr><td>Sexo</td><td>Feminino</td></tr></table><table><tr><td>End
ereço profissional</td><td>Universidade Federal do Rio de Janeiro. <br>Centro de
Ciências Matemáticas e da Natureza (CCMN), Bloco C, Sala E-2006<br>Ilha do
Fundão<br>21945-970 - Rio de Janeiro, RJ - Brasil<br> Telefone: (21) 25983373
Fax: (21) 25628694<br>URL da Homepage:
<a>http://joniceoliveira.tk/</a></td></tr></table>
```

Figure 5 - Personal data field from a curriculum Lattes

The occurrence of the substring "<td> Nome </ td> <td" e "</ td>" respectively before and after the actual name of the researcher, for example, is crucial to the successful implementation of regular expressions. Again, the indexing process by parts is applied (this time with the keywords "Nome", “Nome em citações bibliográficas”, “Sexo”, “Endereço profissional”, “Telefone” e “URL da Homepage”) to filter the desired information. Then, a regular expression removes the unwanted characters and reduces the initial portion of each of the new sub-strings, each with only the desired data. To obtain the name of the researcher (in this case, "Jonice Sampaio de Oliveira”), for example, would be sufficient to use the expression "\ \ <[^ \>] * \> | Nome", which matches any HTML tags (<...> formats or </...>) and the literal “Nome”. Finally, the process of parsing the data from the HTML page is finished writing to a local file system in XML format file containing the data imported from the researcher.

5 Case Study

In order to validate this development, a form was elaborated [21] and it was released to a group of managers of selected research projects that work with data of Lattes. The projects involved are FALE [24], BRINCA and i9com[22], and all of them are based directly on data from the Lattes platform. These are ongoing projects and their research topics are of great importance to the scientific scenario, as social networking, mapping skills, autonomic computing, scientific knowledge management, knowledge management, innovation and innovation networks.

Their coordinators positively evaluated the use of this tool and testified that:

- The XMLattes achieves its goal (ie: search people at the CNPq database, import their data in XML format, load this data into a database and export this data from database to XML files);
- The use of the tool should make their tasks easier;
- The tool is practical, efficient and has a good user interface.

Thus we conclude that XMLattes reached its goal of making easier the access to the curriculum data available in HTML format, transforming them to XML format (for easy manipulation by third-part systems).

6 Conclusions

This paper presented the XMLattes tool, which aims to fulfill the need to importing data from Lattes curricula in academic scenario nationwide. We presented the scenario involved, the related works, the architecture of XMLattes, the development process and how the search process / data filtering works. We also illustrated a case study in ongoing scientific research projects, which ultimately justify the design and development of XMLattes. The positive evaluation by these managers indicates the fulfilled of the previously identified needs.

Therefore, the XMLattes can be used as auxiliary to other IT systems and does not depend on a single platform. Its ties with the GCC are only limited to the data model used for insertion of imported curricula - particularly at a system that will use the tool. The addition of a data access object (DAO), for example, can eliminate this dependence. It is noteworthy that due to development have been done with Java, the tool is also independent of operating systems (ie: it can run on any personal computer that has installed the Java Virtual Machine).

Acknowledgment

The authors of this paper would like to thank Brazilian entitie CNPq for all the financial and technical support.

[1] Sampaio, Jonice de Oliveira. Methexis: Uma abordagem de apoio à Gestão do Conhecimento para Ambientes de “e-Science”. 2007.

[2] Jaccheri,Letizia; Østerlie, Thomas. Open Source Software: A Source of Possibilities for Software Engineering Education and Empirical Software Engineering. FLOSS’2007

[3] Alexandre D. Alves; Horacio H. Yanasse; Nei H. Soma. Extração de Informação na Plataforma Lattes para Identificação de Redes Sociais Acadêmicas. 2009

[4] Cardoso, O. N. P., Machado, R. T. M. Gestão do conhecimento usando data-mining: estudo de caso na Universidade Federal de Lavras. 2008

[5] Pacheco, R. C. S., Forcellini, F. A., Kern, V. M., Gonçalves, A. L., Igarashi, W. Uma análise da pesquisa em engenharia e ciências mecânicas no Brasil a partir dos dados da plataforma lattes. Associação Brasileira de Engenharia e Ciências Mecânicas. 2007.

[6] Silva, A. B. O., Matheus, R. F., Parreiras, F. S., Parreiras, T. A. S. Estudo da rede de co-autoria e da interdisciplinaridade na produção científica através de métodos de análise de redes sociais: Avaliação do caso do ppgci/ufmg. 2006.

[7] CAVALCANTE, Raika Augusta et al . Perfil dos pesquisadores da área de odontologia no Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Rev. bras. epidemiol., São Paulo, v. 11, n. 1, Mar. 2008 .

[8] MENA-CHALCO, Jesús Pascual; CESAR JUNIOR, Roberto Marcondes. ScriptLattes: an open-source knowledge

extraction system from the Lattes platform. J. Braz. Comp. Soc., Campinas, v. 15, n. 4, Dec. 2009

[9] Miguel G. P. Carvalho, Ruben P. Albuquerque, Marcos R. S. Borges, Vanessa Braganholo, PPGI-SGPC Sistema Para Gestão da Produção Científica. 2009

[10] ISO (International Organization for Standardization). ISO 8879:1986(E). Information processing — Text and Office Systems — Standard Generalized Markup Language (SGML). First edition — 1986-10-15. [Geneva]: International Organization for Standardization, 1986.

[11] LATTESEXTRACTOR, 2011. http://lattesextrator.cnpq.br/lattesextrator/; [online] Available at: 13/02/2011.

[12] SCRIPTLATTES, 2011. http://scriptlattes.sourceforge.net/; [online] Available at: 13/02/2011.

[13] GPL, 2011. http://www.gnu.org/licenses/gpl.html; [online] Available at: 13/02/2011.

[14] ORACLE, 2011. http://www.oracle.com/us/sun/index.html; [online] Available at: 13/02/2011.

[15] SWING, 2011. http://java.sun.com/products/jfc/download.html; [online] Available at: 13/02/2011.

[16] LOG4J, 2011. http://logging.apache.org/log4j/; [online] Available at: 13/02/2011.

[17] XML, 2011. http://www.w3.org/TR/REC-xml/#sec-intro; [online] Available at: 13/02/2011.

[18] DTDLATTES, 2011. http://lmpl.cnpq.br/lmpl/Gramaticas/Curriculo/DTD/Fontes/LMPLCurriculo.DTD; [online] Available at: 13/02/2011.

[19] LMPL, 2011. http://lmpl.cnpq.br/lmpl/; [online] Available at: 13/02/2011.

[20] DTDXMLATTES, 2011. http://methexis.cos.ufrj.br/Modelo-XMLattes.dtd; [online] Available at: 13/02/2011.

[21] GDOCS 2, 2011. http://methexis.cos.ufrj.br/AvaliacaoXMLattes.asp; [online] Available at: 13/02/2011.

[22] NETO, B. H. ; OLIVEIRA, J. ; SOUZA, J. M. . Technological and Knowledge Diffusion Through Innovative Networks. In: Samuel Chu, Waltraut Ritter, Suliman Hawamdeh. (Org.). MANAGING KNOWLEDGE FOR GLOBAL AND COLLABORATIVE INNOVATIONS. Hong Kong: World Scientific, 2009, v. 8, p. -.

[23] REGEXP, 2011. http://www.regular-expressions.info/; [online] Available at: 13/02/2011.

[24] GIORDANI, N.; OLIVEIRA, J. ; MONCLAR, R. S.; SOUZA, J. M.; SUEMITSU, W. . Connecting Experts in a Knowledge-intensive Organization: The FALE Project. In: 12th International Symposium on the Management of Industrial and Corporate Knowledge, 2008, Niteroi. Proceedings of 12th International Symposium on the Management of Industrial and Corporate Knowledge, 2008.

A Comprehensive Security Model for Patient Data Warehouse

Paramasivam. Suraj thyagarajan [1], **Neeraj Harikrishnan Girija Paramasivam** [2]
[1]Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA
[2]Department of Electrical and Computer Engineering, Michigan Technological University, Houghton, MI, USA

Abstract— *This paper deals with implementation of a three pronged security model for data warehouse, which is comprehensive and fool proof. The three pronged model implements security at the ETL level, by introducing aggregation and data masking and establishes database level security by implementing the virtual private database. This model prevents any direct access to the actual data warehouse, and all accesses are provided only through the VPD. This ensures that the database is secured from unauthorized access. The standard methods of data masking and aggregation are used to prevent any attacks on the data warehouse. The combined power of these methods, give a comprehensive security model, that can be implemented on enterprise wide or public data warehouse.*

Keywords: Datawarehouse Security, ETL Security, Virtual Private Databases, Three pronged security to Datawarehouses, patient data security

1. Introduction

In modern day data warehouses, the implementation of security has not been given a great deal of importance and this sometimes leads to data theft at various levels. This project aims at preventing the data theft from the data warehouses. The project proposes a three pronged approach to implement security in datawarehouses. We also discuss the other related work in this area, distinguishing them from the current work. This paper looks into a real time experiment conducted on a large dataset of patient genomic data and compares the performance parameters with a regular system.

A datawarehouse is a large collection of data in a large specialized database. These datawarehouses are used for a variety of purposes including data mining, business intelligence and many more. The databases generally used for these purposes are large scale machines with high reliability and scalability. The best examples of such specialized systems are Oracle Warehouses, Teradata, and netezza systems. However, these systems follow security through obscurity principle.That is they consider the system is secure because limited users have access to it. The security models employed for most of the corporate data warehouses are not comprehensive in providing security.

In terms of security implementation, the process of data loading, defined by ETL, has a lower security priority for most organizations. The process of ETL is defined as Extraction, Transformation and Loading, which is the process of loading data into the Datawarehouses. Extraction refers to the process of extracting data from multiple sources. Transformation refers to the process of cleaning data or applying business rules to the data. Loading is the actual process of loading data into the warehouse. There are other kinds of processes which are used to load data into the warehouse, which include ELT(Extract Load and Transform) etc. However, none of these loading techniques seem to implement data security at the ETL level albeit there are a few works suggesting security at this level. However, these papers do not propose a model that encompasses both database and ETL security.

Another level of security that is implemented by some of the datawarehouses is the provision of controlled access to the DW users. The access restriction is usually based on a simple authentication mechanism, which authenticates users based on their credentials. However, most data warehouse set ups do not have rigorous implementations of authentication mechanisms for these. This provides with a small amount of security to the databases, which unfortunately can be broken with not too much effort.

This paper aims at combining some of these well known security practices to put forward a comprehensive model that aims to secure a datawarehouse in the health sector. The issue of security in the health related data is one of the most important, yet one of the most overlooked of all. This paper aims in providing comprehensive security to a patient genomic datawarehouse. However, our model can be implemented across any domain and can be extended to any generic data warehouse.

2. The Three Pronged Model

We here propose a three pronged approach to security, for providing comprehensive security for the patient datawarehouses. This is a generic model that can be applied to any data warehouse per se. However, we have handled patient datawarehouse as a case. We also have presented an analysis of the performance of the loading process of data. We can safely presume with numerous examples that the performance of a warehouse is more dictated by the ETL process rather than by any downstream processes.

In this model, we look at three different areas where the data resides or traverses. We divide these into high risk areas, low risk areas and presumably safe areas. In the figure [1] , we provide the complete overview of the model that has been devised. In the figure [1], the first area, containing the sources and the staging area, are considered as the safe areas. These areas have least security threat due to the corporate nature of their maintenance. These areas where data reside can hence be presumed to be safe. The next important area where the data resides and traverses is the public datawarehouse area, which is prone to attacks of various kinds. However we can further subdivide this area into smaller fragments which have different risks. The target data warehouse is where the data is originally housed. This area is prone to different kinds of attacks. The data in this area is to be protected since this is the final set of data, that contains sensitive information. To mitigate the risks in this area, we suggest a secure approach, in the way data is being populated into this database. The principles of aggregation and generalization is well known in the industry for two purposes. The concepts of aggregation is used both for improvements in the overall security of the warehouse and also for improving performance of the joins during mining of data. However, we propose to use aggregation in conjunction with data masking, to provide additional security to the warehouse. Aggregating data will let us hide some parts of data from anyone using the data warehouse. Also the usage of data masking obfuscates the most important columns from any one accessing the warehouse. However, the aggregate tables and data masking transformations together can provide improved security to the entire set up. The aggregate tables are set up in such a way that only the VPD engine can access it. The same applies to the Data Masking tables. This provides enhanced security,since these tables contain vital information which can be misused when joined with the data from the actual data warehouse.

We also introduce the concept of virtual private database in this context, making it the only engine which will have access to both the aggregate and data masking tables along with the actual warehouse tables. This makes it impossible for any intruder to steal any meaningful data from the warehouse. The virtual private database is a concept prevailing for a long time and is one, which gives different views of data to different users based on their roles and privileges. We, in our experiment have used an oracle virtual database on a patient genomic data system.

In order to ensure user authentication, most of the existing systems use built in security mechanisms of the databases or other Business intelligence or data mining tools. However these tools are sometimes not very efficient in providing ample amounts of security and can be easily broken. To mitigate this, we propose to use a separate authentication server, which handles a DES keystore, to authenticate users. This authentication will be a two way authentication and will provide a more fool proof method for user authentication.

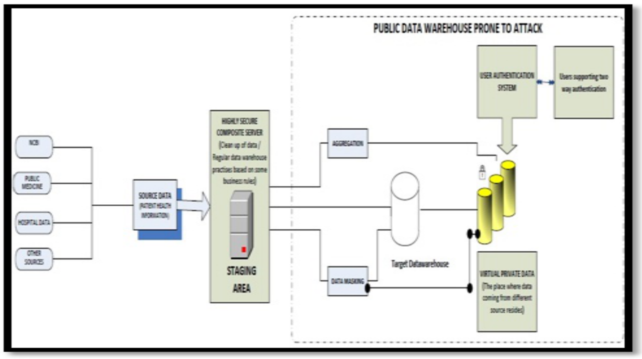

PUBLIC DATA WAREHOUSE PRONE TO ATTACK
USER AUTHENTICATION SYSTEM
Users supporting two way authentication
HIGHLY SECURE COMPOSITE SERVER (Clean up of data / Regular data warehouse practises based on some business rules)
NCBI
PUBLIC MEDICINE
HOSPITAL DATA
OTHER SOURCES
SOURCE DATA (PATIENT HEALTH INFORMATION)
STAGING AREA
AGGREGATION
Target Datawarehouse
DATA MASKING
VIRTUAL PRIVATE DATA (The place where data coming from different source resides)

Fig. 1: The block diagram for the three pronged model. Refer Fig 1 in Appendix for a bigger picture

3. A Comparison with some Related Work

Though there are several related works in the field of database security, the implementations of security to ETL is studied in a very few works. One such work was presented at the Third UK Symposium on Computer Modelling and simulation by M Mrunalini et all [1]. The authors in this piece of work look at implementing ETL security and also provide appropriate UML based use cases for a model to implement ETL security. However, the authors do not concentrate on any of the database security options. Also the paper deals with the methods of user authentication and implementing security, which are generic in nature and are applied to the principles of datawarehousing. In our work, we have developed a specific model for ETL , which are based on the well known concepts of ETL. Also we here provide a more comprehensive model, which includes database security and ETL security within the same model, along with user authentication systems. This makes our system unique with a more comprehensive nature. There are other related material available on the internet and other sources, which talk about security in ETL or security in databases. Also the use of virtual private databases is being extensively referenced by various literature. However none of these material seem to suggest a comprehensive model for security of data through the entire process of data warehousing. A virtual private database for example, is usually not accompanied by any security measures at the ETL level. The security at ETL level is ignored by most authors, since they assume ETL to be inherently secure and restricted to a corporate sector. However, due to increasing threats to corporate data, which includes sensitive data, by insiders, the need to implement ETL security has gained extreme importance. Another related work by Kimmo Palletvuori from Helsinki University of technology titled "Security of Data Warehousing server" [2] details the methods of implementing security to the physical servers for the datawarehouse. This deals more with the security of the server at a physical level rather than at a logical level. The author in this paper also talks about concealing data at a logical level and preventing unauthorized access at the ETL level. However, this paper does not detail any methods of providing security at the ETL level. This makes the process less secure by not providing security for the entire process. Our work in this project provides a method to mitigate this gap in providing security to the data in the data warehouse. The project gives a method of providing ETL security and security to the datawarehouse, using industry standard methods. This project details a comprehensive method of providing data security to the data warehouse. Another major work related to the area of datawarehouse security is the white paper presented by Oracle in April 2005 [3]. This white paper provides complete details of the virtual private database that oracle is capable of providing. However, this white paper also does not provide any details on the possible ETL level security. This provides a gap in the security arrangement provided inherently by oracle. This differentiates our work from the existing white paper published by oracle. Our project provides a more comprehensive model which fills some of these existing gaps in providing security to the data warehouse.

4. Experiments and Results

Table 1: Table indicating the run times using various ETL tools.

Tool Used	Number of Records	Run Time without Security features implemented	Run Time with Security features.
Informatica	11104	3.00 minutes	4.02 minutes
Informatica	4509237	15.08 minutes	20.43 minutes
TalenD	11104	3.24 minutes	4.29 minutes
TalenD	4509237	16.23 minutes	23.07 mins

This project was carried out on a patient data warehousing system, with industry standard databases and ETL tools. The project was carried out on a large dataset obtained from the NCBI. Due to privacy concerns, we created hypothetical data using the genomic accessions available in the NCBI database. We created a source dataset, with hypothetical names and SSNs, which did not reflect actual data, however reflected the volume of data found typically on a patient datawarehouse. Additional columns were introduced to create a more logical dataset, and this dataset was loaded into the data warehouse. Two sets of loading were performed on the data warehouse. The first set of loading was done without any security implemented on ETL or the database. The second set of loading was done after implementing the security model on the ETL and databases. This provided us a clear picture of the performance degradation on the entire process. In this case, we used a dataset of 11104 records in the source data. The first run was conducted and this took an overall run time of 3 minutes to load the entire set of data into the warehouse. The second run, with the security features implemented took a total time of 4 minutes and 02 seconds. This was achieved with a total throughput of 1110 records using an industry standard tool for ETL, Informatica. Also all these runs were conducted on Oracle 10g and a machine with Intel Core 2 Duo processor running with a RAM of 2 GB. We also performed the same experiment with an open source ETL tool, TalenD, to ascertain the results were not biased by the tool and we could get accurate results. Below given is a table of values obtained during the experiments using both the ETL tools.

The entire experiment showed that the degradation in performance due to the additional implementation of security was very minimal and hence this model of implementation of security can be implemented without too much overhead on the entire system. The screenshots of the execution of our project is available in the appendix.

5. Conclusion

The project has proposed a means of implementing security using a new model, which has been proposed to fill any gaps in security of data warehouse. The notion of security in data warehouses have so far been restricted to security in databases. A very few related works have been related to security in ETL. However, none of these models are comprehensive, providing both ETL and database level security. This project provides a total security for the entire data warehouse along with the process of loading data using ETL. The scope of future work in this project can be extended to creating models implementing cryptographic protocols at the ETL level, which can provide more efficient security. The framework we provided here is open and can be extended to different levels or different domains. However, this existing domain of health care and security of patient data is extremely important and it is important to provide more secure ways of data access in these data warehouses. Also the work could be extended to provide a model of security for the data mining paradigm.

References

[1] M Mrunalini, T V Suresh Kumar, K Rajani Kant, "Simulating Secure Data Extraction in Extraction Transformation Loading (ETL) Processes," *Third UKSim European Symposium on Computer Modeling and Simulation*, 2009.

[2] Kimmo Palletvuori, " Security of Data Warehousing Server,"

[3] *Security and the Data Warehouse, An Oracle White Paper*,2005.

[4] Wiley Publications 2006 *Ralph Kimball, The Datawarehouse ETL Toolkit*.

[5] A. Simitsis, P. Vassiliadis, and T. K. Sellis, "Optimizing ETL processes in data warehouses,"*In Proc. ICDE, pages 564-575*, 2005.

[6] D.W. Embley, D.M.Campbell, Y.S.Jiang, S.W. Liddle,D.Wlonsdale, Y.-K.Ng, and R.D. Smith,“Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages,” *Data and Knowledge Engineering 31,no. 3* pp. 227-251 Nov 1999

[7] S. Jajodia and D. Wijesekera, “Securing OLAP data cubes against privacy breaches,” *In Proc. IEEE Symp. on Security and Privacy* pages 161-178 2004.

[8] A. Simitsis, “ Mapping conceptual to logical models for ETL processes,” *In Proc. DOLAP*, pages 67-76, 2005

[9] T. Priebe and G. Pernul,“. A pragmatic approach to conceptual modeling of OLAP security,” *In Proc. ER, pages 311-324* 2000

6. Appendix

The Architecture Block Diagram

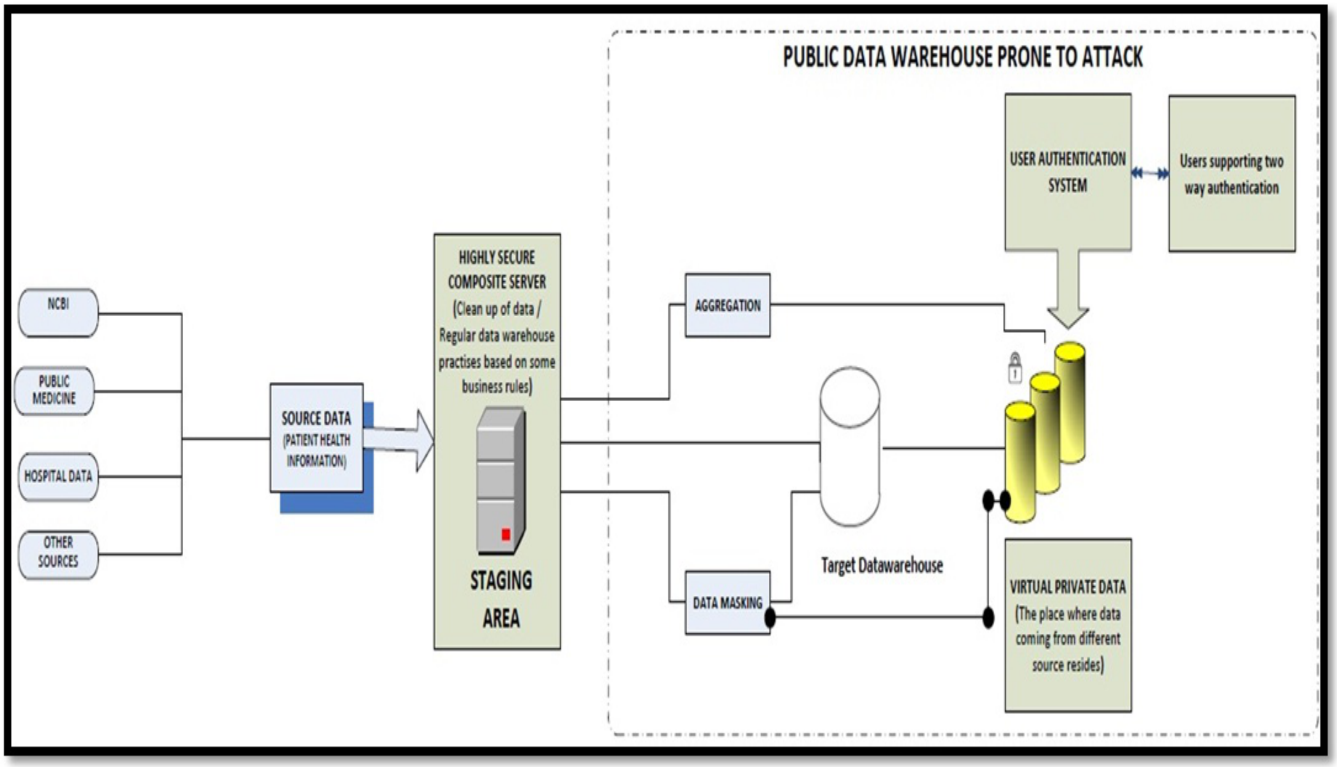
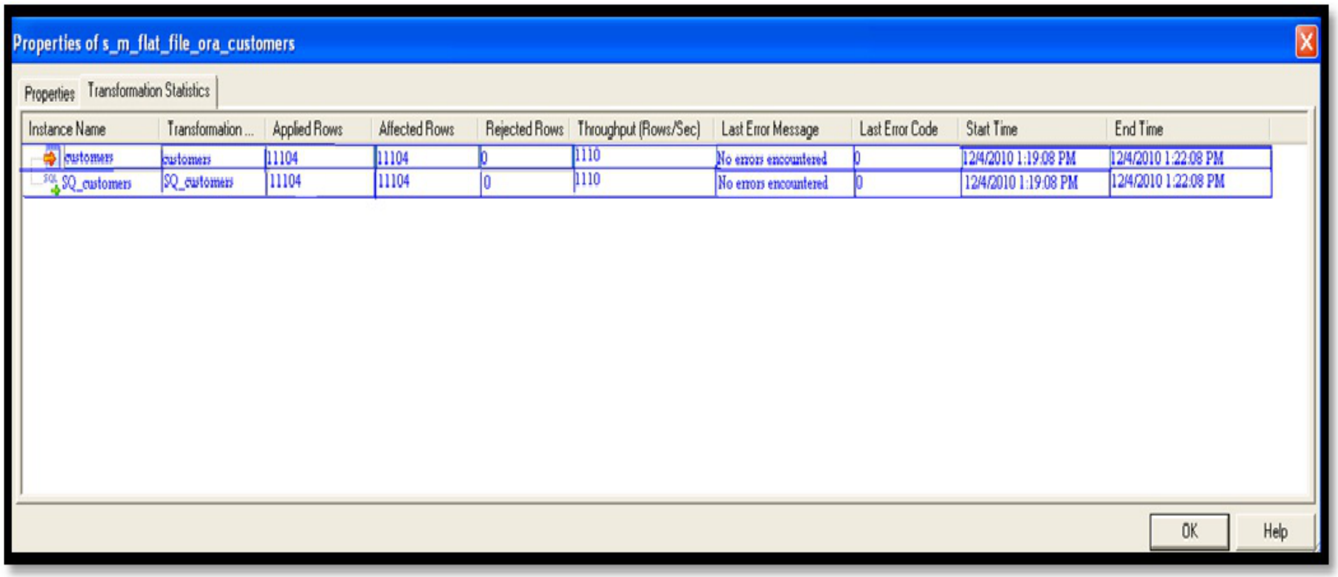

PUBLIC DATA WAREHOUSE PRONE TO ATTACK
USER AUTHENTICATION SYSTEM
Users supporting two way authentication
NCBI
PUBLIC MEDICINE
HOSPITAL DATA
OTHER SOURCES
SOURCE DATA (PATIENT HEALTH INFORMATION)
HIGHLY SECURE COMPOSITE SERVER (Clean up of data / Regular data warehouse practises based on some business rules)
STAGING AREA
AGGREGATION
DATA MASKING
Target Datawarehouse
VIRTUAL PRIVATE DATA (The place where data coming from different source resides)

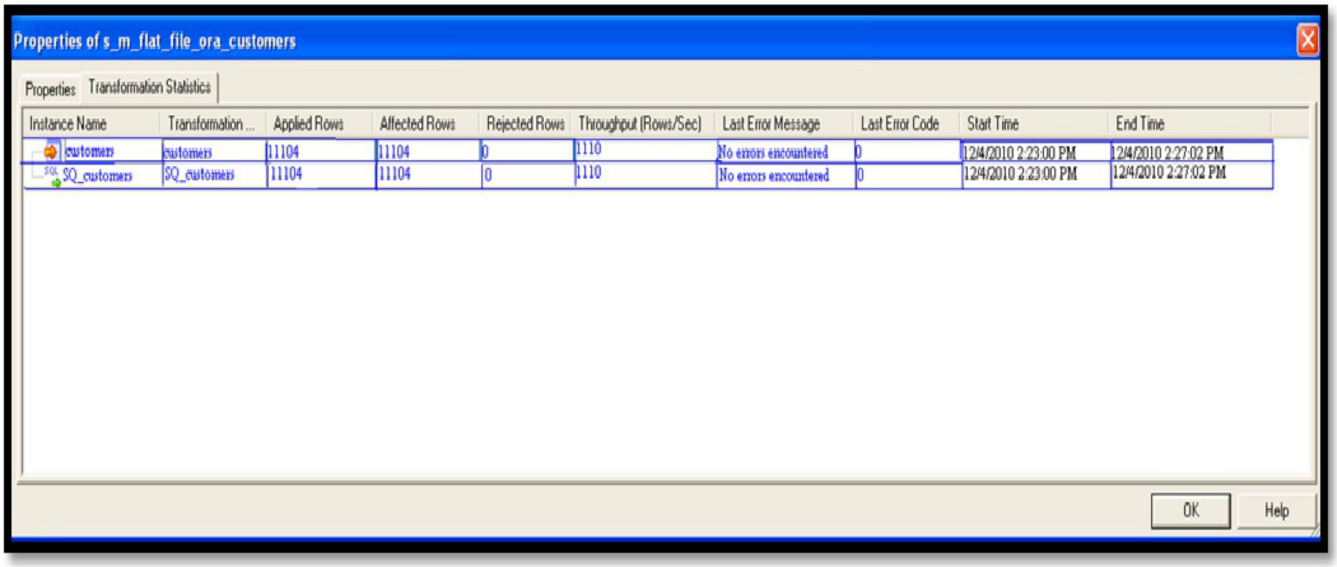
Fig. 2: Enlarged Version of architecture.

Screenshot of Experiment Results:


Properties of s_m_flat_file_ora_customers

Instance Name	Transformation ...	Applied Rows	Affected Rows	Rejected Rows	Throughput (Rows/Sec)	Last Error Message	Last Error Code	Start Time	End Time
customers	customers	11104	11104	0	1110	No errors encountered	0	12/4/2010 1:19:08 PM	12/4/2010 1:22:08 PM
SQ_customers	SQ_customers	11104	11104	0	1110	No errors encountered	0	12/4/2010 1:19:08 PM	12/4/2010 1:22:08 PM

Fig. 3: The result of experiment on Informatica showing the time required to load 11104 records without security module implemented


Properties of s_m_flat_file_ora_customers
Properties
Transformation Statistics
OK
Help

Instance Name	Transformation ...	Applied Rows	Affected Rows	Rejected Rows	Throughput (Rows/Sec)	Last Error Message	Last Error Code	Start Time	End Time
customers	customers	11104	11104	0	1110	No errors encountered	0	12/4/2010 2:23:00 PM	12/4/2010 2:27:02 PM
SQ_customers	SQ_customers	11104	11104	0	1110	No errors encountered	0	12/4/2010 2:23:00 PM	12/4/2010 2:27:02 PM

Fig. 4: The result of experiment on Informatica showing time required to load 11104 records after implementing security model

SESSION

NOVEL TECHNIQUES

Chair(s)

TBA

Enhanced Biomedical Taxonomy Mapping Through Use of A Semantic Measure of Proximity

Jeffery L. Painter
Computer Science Department, North Carolina State University, Raleigh, NC, USA
GlaxoSmithKline, Statistical & Quantitative Sciences, RTP, NC, USA

Abstract— *By employing a notion of semantic closeness to create a multi-phase mapping between one biomedical taxonomy and another, we are able to determine various levels of proximity for mapping term based structures which may fail to map using traditional mapping techniques. The multi-phase approach allows for defining the appropriateness of applying the maps to various domain problems such as the investigation of high level system organ or class effects versus a problem that requires a higher level degree of specificity in the analysis.*

Keywords: semantic mapping, meaning hierarchy, ICD-10 codes, MedDRA codes, UMLS Metathesaurus, terminology alignment

1. Introduction

For many exercises in biomedical data analysis, the need arises to map one medical terminology to another. Over the years, several methods (1) have been developed which attempt to solve the problem of accomplishing this goal in some automated procedure to reduce the effort necessary to process these sometimes large taxonomies which would otherwise require hundreds if not thousands of man hours to complete and would be prone to human error and fatigue.

In another paper (2), it was shown that it might be possible to also introduce a structure to previously unstructured terminologies by making use of "extra" information found within already structured terminologies.

In this paper, the purpose is to demonstrate how taking a *multi-phase* approach; the mapping from one terminology to another may be drastically enhanced to help solve any number of inferential analysis problems by looking at broader meaning relations among various terminologies in order to capture more relationships among the codes they each contain.

It is a *multi-phased* approach in that we attempt various methods for aligning, or matching, one terminology to another where many previous attempts at ontology alignment seek to only exploit one method or another. It is by the combination of several different methods, and developing a model of measure for relative closeness of mappings, that this enhanced procedure produces much more effective results.

By assigning a weight to indicate a sense of "closeness" between two terms or concepts, the mapping which we produce could then be applied in a manner described by Wang, Gong and Zhou (3) to create a composite model for ontology mapping.

Mappings of these types are being developed to enhance the data analysis of large observational databases such as electronic health records, claims data and other medical history databases used in both public and private systems such as those being developed by the Observational Medical Outcomes Partnership (OMOP) [1] initiative and the SafetyWorks (4) project which was initiated by GlaxoSmithKline. A common data model (5) allows for a standardization of the records across disparate data sources. However, the first step in creating these models is to *normalize* the data to a common, or reference, terminology. It is only by making use of ontologies and methods such as those described in this paper that we are able to achieve any level of success in *fitting* the data to the common data model.

The problem first encountered with the Read/OXMIS codes in (2) is re-evaluated with these enhanced methods in addition to mapping the ICD-10 codes to the Medical Dictionary for Regulatory Affairs (MedDRA)[2].

Our goal for mapping ICD-10 stemmed from the fact that the latest effort relating to the SafetyWorks project is to incorporate the IMS Germany database which is itself coded in ICD-10, German language variant. In order to incorporate this new database into the system, a mapping between ICD-10 and MedDRA (our reference terminology) needed to be constructed.

2. Term Based Matching

In previous work, we have taken advantage of many techniques previously described for aligning two ontologies, taxonomies or terminologies based on the terms occurring in each of them – from here on we will just call them terminologies since we are typically making use of the terms identified as representing the concepts in each. However, *term-based* mapping can be prone to errors due to the nature

[1] See http://omop.fnih.org/

[2] By 'ICD-10' we mean to refer to the International Classification of Diseases, 10th Revision, which is maintained by the World Health Organization. MedDRA® (Medical Dictionary for Regulatory Activities) is a registered trademark of the International Federation of Pharmaceutical Manufacturers Association. The Clinical Terms Version 3 (Read Codes)© are maintained by the (UK) National Health Service Information Authority.

of the representation employed within any given terminology system.

When one investigates the meaning being represented by an individual term, there may be some contextual information found in the hierarchy which does not appear directly in the term itself. For example, ICD-10 has several codes in the Yxx family which indicate possible poisoning or adverse effect of a particular agent, however the code's term may only list the agent and not explicitly state that it should have any adverse indication in that term. If you were trying to align ICD-10 to a terminology which also included a drug or medicament hierarchy, matching solely on the term occurence could place ICD-10 codes which are meant to represent an adverse reaction in the wrong position in the target terminology where the same terms represent only the substance and no mention of reaction, positive or negative.

Therefore, while we still employ term-based mapping within our *multi-phase* mapping process, those codes that can only be linked by a term-based match alone will be given the lowest level of "closeness" in our *hierarchy for semantic proximity*.

2.1 Hierarchy for Semantic Proximity

There are essentially four levels of "closeness" in our model which include:

1) Conceptual Level
2) Boosted Level
3) Nearest Neighbor Level
4) Term Level
    a) Direct Match
    b) Fuzzy Match

The conceptual level is considered as having the highest (or most relevant) degree of closeness between any two given entries in the two terminologies we are attempting to align, followed by the boosted level, the nearest neighbor level, and finally the term level (possessing the lowest degree of closeness) discussed briefly above.

After creating the mapping file between any two terminologies, we retain the semantic proximity information embedded within the mapping to enable the end user to filter mappings based on the particular need of precision in a given analysis. To speed the multi-phase mapping process (and to insure the highest level of semantic proximity is held between any two given terms), we execute the mapping in order from highest level to lowest level, with the exception of the term level.

The term level maps are generated first and are used in addition to the other methods. If a mapping can be found at any higher level, then the term map will be eliminated in favor of one which has a higher degree of semantic proximity.

Once a term from the source terminology is mapped at one of the higher levels, it is then removed from the set of terms to be evaluated for the lower level maps to follow.

2.1.1 Conceptual Level

The conceptual level of the hierarchy takes advantage of the UMLS Metathesaurus[3] in order to create a mapping between two terminologies. If the source and target terminologies are both found within the UMLS, this process is accomplished quite easily, and we can associate any two term entries by the presence of a shared concept unique identifier (CUI) which indicates synonymy within the UMLS conceptual model. If one terminology is not found within the UMLS, then an attempt can be made to identify "potential" conceptual maps by string matching. If two terms have the same exact string, in most instances, these are in fact representative of the same concept.

One way we were able to improve the mapping of the ICD-10 and Read/OXMIS codes was by incorporating mutliple sources from the UMLS. It may be the case that a term found in one terminology is present in another, even if it is not present in the target terminology. By matching the terms to those known to exist in the UMLS, we are still able to find a CUI which would link back to the source terminology and give us the necessary information in order to make a conceptual level match to the target terminology (in this case MedDRA).

When mapping the IMS Germany database, which is coded in ICD-10 (German), to MedDRA, the first step is to link each of the ICD-10 code entries to the version of ICD-10 found in the UMLS. Only one code from the IMS Germany database was unable to be mapped directly to ICD-10 in this way.

After converting the IMS Germany codes to their ICD-10 equivalents. We looked at the UMLS entries for each ICD-10 code that was applicable and identified it's CUI in the UMLS. If there existed a corresponding CUI for MedDRA, then we subsequently link the IMS Germany code to the MedDRA code. Any given CUI may link to several MedDRA entries, and from those, we choose a single MedDRA code based on some simple rules related to the term type associated with a particular MedDRA code.

Those selection rules favor MedDRA codes at the PT/LT level and if none are found, proceeds to climb the MedDRA hierarchy until a match is found. We do take care to note the problems commonly attributed to the MedDRA hierarchy (6) by eliminating duplicate LT entries when a PT entry is found. The selection criteria also involves multiple passes of the potential MedDRA concepts identified by CUI association to identify term-type (TTY) matches at various levels of the MedDRA hierarchy. If a preferred term (PT) is found, then preference is first given to that concept (since it exists at a more specific level of the MedDRA hierarchy). If the PT level is exhausted, the code proceeds to look further at the potential matches by trying to associate to a high-level term

[3] UMLS Metathesaurus is a project of the (US) National Library of Medicine, Department of Health and Human Services. Available at: http://www.nlm.nih.org/research/umls/

(HT), followed by a group term (HG) and lastly by looking for obsolete terms found in MedDRA which may be useful for determining levels of semantic proximity at a lower level than the conceptual level.

The entries identified by this first step in the *multi-phase* match are given a CONCEPT_MAP identifier in the semantic proximity embedding of our mapping file. 44% of the IMS Germany codes are mapped to the MedDRA target by the conceptual level, while only 16% of the Read/OXMIS codes are mapped to MedDRA at the conceptual level alone.

2.1.2 Boosting

Boosting is one of the more contraverisal methods for mapping codes from one terminology to another. Many of the biomedical terminologies have some inherent structure identifiable in the manner which the codes are constructed or by some external hierarchical structure that is annotated by the codes and terms found within that system.

The idea of boosting is to simply take advantage of this structure in an attempt to incorporate knowledge about *surrounding* codes that may have been mapped by either the conceptual map or a term level map. This idea is modified for terminologies, but deeply rooted in the notion of discovering proximity by relatedness such as discussed in (7).

When looking at the IMS Germany database, each data reference to an ICD-10 code is typically in the form of a 5 digit code.

The ICD-10 codes themselves are arranged in a hierarchy such that the first 3 and 4 digits of each code form various levels of a "family" of related codes. Boosting will atttempt to extract from an unmapped 5-digit code both it's 3 and 4-digit family levels and attempt to search for a conceptual mapping which was found at those higher levels. If a match iss found, then that boosted node will have a CUI associated with it which also occurs in the overall MedDRA CUI set. The process is then to associate the unmapped 5-digit code with the boosted relative's CUI and map it into the target terminology.

An additional 775 ICD-10 codes were mapped using this method. From the Read/OXMIS codes, an additional 11,648 mappings (approximately 14.9%) were created to the MedDRA target by using this method.

The maps produced by *boosting* generally provide a broader concept map to a lower level term than one might generally hope for, but it still allows many codes to be included which might not have otherwise been captured using traditional methods. For many of the types of analysis used in data mining observational databases, it is often the case that we only care about a higher level of generality (such as *liver disease, diabetes, etc*) anyway, rather than searching for individual lower-level conditions or diagnoses.

In some cases, we were able to create additional links by climbing the hierarchy beyond the parent level to the grandparent or great-grandparent in order to find a link back to the target terminology. These links were only applied if the level of the target terminology was still at the PT or LT level of the MedDRA hierarchy. In the case of the Read/OXMIS codes, it is not advisible to proceed past the grandparent level for boosting. The mappings produced beyond this level possessed numerous inaccuracies due to the fact that the Read terminology is multi-hierarchical and the linkages between higher levels of Read and MedDRA lead to inconsistencies in annotating the maps between code instances at a lower level to those in the target terminology which are found at a higher level.

From the GPRD Read/OXMIS terminology, an example of a code captured by the boosting method is: "K44..00 - Female gonococcal pelvic inflammatory disease" which was succesfully mapped to the broader MedDRA code "10034254 - Pelvic inflammatory disease". And from IMS Germany, an example boost match includes "E02 - Sub-clinical iodine-deficiency hypothyroidism" mapped to the MedDRA code "10043709 - Thyroid disorder".

Each of the mappings produced by the *boosting* method were evaluated manually for relevance of match, and were all deemed successful aside from those at the great-grandparent level for the Read codes. These entries are identified with a BOOST_MAP identifier in the semantic proximity embedding of our mapping file.

2.1.3 Nearest Neighbor

The last method developed for the multi-phase mapping process is the concept of using a *nearest neighbor* match. Both the IMS Germany codes and those found in Read possess an inherent hierarchical structure used to organize the content in each terminology. The nearest neighbor level matching attempts to look at codes relatively "close" to an unmapped code, that is they share at least the first three characters in their code designation, or by their hierarchical structure reside as siblings in the tree structure. If any of the neighboring codes were successfully mapped to MedDRA, then those links are used to enhance the mapping process to create additional links to the unmapped siblings.

To illustrate the success of this method, an IMS Germany example is given:

"D61.3 - Idiopathic aplastic anaemia" had no direct term or conceptual match to any particular MedDRA code. But through the nearest neighbor level, the MedDRA code "10002037 - Anaemia aplastic" was associated with this particular ICD-10 code and given an embedding of a nearest neighbor match for the level of semantic proximity in the mapping file.

Again, the results were manually reviewed to determine goodness of fit in the overall map file. Most of the mappings produced via this level occur again at the PT or LT level of MedDRA. An additional gain of slightly more than 400 IMS Germany codes were mapped using this method while the GPRD Read/OXMIS maps were enhanced with an addition

355 mappings at this level.

2.2 External Sources

2.2.1 ICD-9 CrossMap

While mapping the IMS Germany database to MedDRA, our first step was to investigate the existence of any maps already created from ICD-10 to another terminology such as MedDRA. While we found no freely available crossmaps between these two coding schemes, the existence of crossmaps between ICD-10 and ICD-9[4] were readily available.

The UMLS itself possesses a much higher degree of concept coverage between ICD-9 and MedDRA than it does between ICD-10 and MedDRA. Therefore, the idea was to incorporate the use of the external crossmap files available for ICD-10 to ICD-9 in order to further enhance our mapping process to provide a crossmap between ICD-10 and MedDRA.

The algorithm gives mappings produced at this level the same degree of semantic proximity as that of the conceptual level. Since most of these external sources have been validated for collecting statistics for large national health services, the level of rigor in creating these maps is of similar caliber as that of the concept maps created for the UMLS itself. The algorithm still gives preference to a UMLS concept map above all others, then if a crossmap reference was found, it was given preference over any lower level mapping produced in the multi-phase approach.

This process did prove extremely useful in gaining an additional 2,684 code mappings between ICD-10 and MedDRA that the UMLS itself failed to reveal in the conceptual mapping phase.

The sources we made use of were:

1) New Zealand Health Information Service (8)
2) National Center for Health Statistics (GEMS) map (9)

The cross maps for the most part seemed highly correlated to the MedDRA mappings that were produced. There were however a small number of "Y" codes (e.g. Y51.1, Y54.5, and Y58.9) from the New Zealand file which were not correct (again, this may be attributed to the contextual placement of terms within ICD-10 that do not completely relate to those same terms in the MedDRA coding scheme).

3. Evaluation

The multi-phase approach to mapping terminologies provides an extremely diverse set of target matches between one terminology and another. It is highly improbable to ever produce exact one-to-one mappings (1) between any two biomedical terminologies due to the fact that most are developed independently in order to serve the needs of a particular domain problem such as recording surgical procedures, diagnostics and conditions for an electronic health records system or for medical claims and billing data. Each terminology has a specific context from which it was developed and meant to be applied toward. However, the hope still exists to be able to align these terminologies with one another to the greatest extent possible.

The UMLS goes a long way towards reaching this goal, by providing an overarching conceptual model through which concept synonymy is expressed by means of shared CUIs among terms in each vocabulary. However, it still is not inclusive of every possible biomedical terminology, and still suffers from a lack of resources to maintain and evaluate the changes which frequently occur in biomedical terminologies from one release to the next.

It is also still very much the case that the UMLS does not claim to have any conceptual hierarchy relating one concept to another with any distance measure that can be used with any consistency when moving between one terminology and another. Therefore, by providing a hierarchy for semantic proximity such as the one proposed in this paper, we hope to encourage the development of similar measures which can then be embedded within a system such as the UMLS.

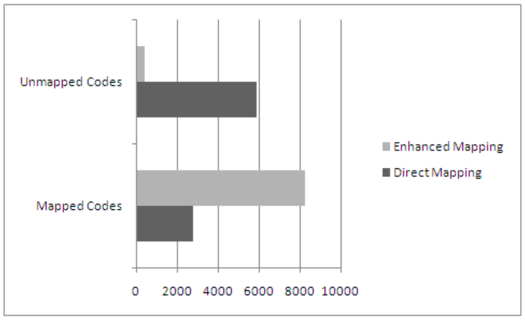


Figure 1: ICD-10 Code Map Coverage

Traditional mapping methods still fall short in producing as many maps are as possible with the multi-phased mapping approach demonstrated here. Term matching alone can produce a high number of potential map targets between any given two terminologies such as the *fuzzy* matching methods described in (2). However, term based matching alone is still subject to a lack of contextual information as noted in the examples stated previously.

The conceptual, boosted and nearest neighbor methods provide additional maps which can still be useful in many data mining activities. The content coverage between IMS German and the MedDRA terminology was drastically improved as shown in Figure 1. The enhanced multi-phase map reduced the total number of unmapped codes by almost 95%.

The number of maps produced between the Read/OXMIS and MedDRA terminologies using the multi-phase approach reduced the total number of unmapped codes from 38,825 to 34,878. While this may not seem all that significant, the

[4] By 'ICD-9' we mean to refer to ICD-9-CM, the International Classification of Diseases, 9th Revision, Clinical Modification, which is maintained jointly by the National Center for Health Statistics and the Health Care Financing Agency.

actual clinical data coverage that was increased by these mappings jumped from slightly less than 50% previously, to more than 60% in this update. The confidence in the mappings produced is also boosted by the fact that we now can provide the semantic proximity as an additional aid in determining the usefulness of these maps to the scientists who will be working with them in the future.

4. Conclusion

Through the use of a multi-phase mapping process, many more potential maps between two terminologies can be realized than by using traditional methods alone. The use of a semantic measure of proximity gives valuable insight into the mappings produced and discretion in how they may be applied to various data mining problems.

It is often the case that the mappings produced between two terminologies must be validated or verified before inclusion in applications such as those used to assess drug safety issues. But the results of this effort show that in many instances, we can reduce the overall volume of concepts which need manual review to those only occurring at lower levels of semantic proximity. Thus, saving countless man hours and valuable resources which would be better suited to the actual investigation of data rather than simply reviewing mapping files between one terminology and another.

Shortly after producing the map files between Read/OXMIS and MedDRA and the IMS Germany codes and MedDRA, we were tasked with creating yet another map between the MeSH (Medical Subject Heading) and Read/OXMIS terminology. The result of the work in this paper allowed us to leverage mappings produced between Read/OXMIS and MedDRA to produce a mapping to MeSH in relative short order achieving similar results and enabling scientists to proceed with mining literature by way of the Read/OXMIS terminology and the translation to MeSH. Again, the embedding of semantic proximity helps to provide valuable clues as to the level of specificity to be deemed necessary when conducting a literature review and being able to cast a broader or narrower net as necessary by means of the filtering now available in the maps this method is capable of generating.

Future work will investigate the possibility of refining our hierarchy for semantic proximity even further and to investigate the applicability of this method to general ontology matching methods.

References

[1] Y. Kalfoglou and M. Schorelmmer, “Ontology mapping: the state of the art”, *Knowledge Engineering Review*, vol. 18, no. 1, pp. 1–32, 2003.

[2] Jeffery L. Painter, “Toward automating an inference model on unstructured terminologies: Oxmis case study”, in *Advances in Computational Biology*, Hamid R. Arabnia, Ed., vol. 680, pp. 645–651. Springer New York, 2011.

[3] Ying Wang, Jianbin Gong, Zhe Wang, and Chunguang Zhou, “A composite approach for ontology mapping”, in *Flexible and Efficient Information Handling*, David Bell and Jun Hong, Eds., vol. 4042 of *Lecture Notes in Computer Science*, pp. 282–285. Springer Berlin / Heidelberg, 2006.

[4] G.H. Merrill, P.B. Ryan, and J.L. Painter, “Construction and annotation of a UMLS/SNOMED-based drug ontology for observational pharmacovigilance.”, in *Proceedings of the Intelligent Data Analysis for bioMedicine and Pharmacology*, Washington, DC, 2008.

[5] Stephanie J Reisinger, Patrick B Ryan, Donald J O’Hara, Gregory E Powell, Jeffery L Painter, Edward N Pattishall, and Jonathan A Morris, “Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases”, *Journal of the American Medical Informatics Association*, vol. 17, no. 6, pp. 671–674, November 2010.

[6] G.H. Merrill, “The MedDRA paradox”, in *AMIA Annual Symposium Proc*, Washington, DC, 2008, pp. 470–474.

[7] M. A. Merzbacher, “Discovering semantic proximity for web pages”, in *Proceedings of the 11th International Symposium on Foundations of Intelligent Systems*, London, UK, 1999, ISMIS ’99, pp. 244–252, Springer-Verlag.

[8] “New zealand health information service”, 2011, Available online: http://www.nzhis.govt.nz/moh.nsf/pagesns/254.

[9] “National center for health statistics (gems)”, 2011, Available online: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications /ICD10CM/2010/2010_DiagnosisGEMs.zip.

Cyclic Association Rules: Coupling Multiple Levels and Parallel Dimension Hierarchies

Eya Ben Ahmed, Ahlem Nabli and Faïez Gargouri

Abstract— **The data warehouses contain massive volumes of historicized data defined over a set of dimensions and aggregated through multiple levels of granularities. Although the extensive analysis tools aiming to navigate through those granularity levels, few works exploit the multidimensional model features to derive regular fitting knowledge. In this paper, we highly take advantage of the different dimensions and their parallel levels of granularity to propose a new mining method for cyclic patterns extraction from data cubes. Hence, the innovative definitions and dedicated algorithm are extended from ordinary cyclic patterns to this particular context. Experiments are reported, showing the significance of our approach.**

I. Introduction

In the last decade, several works were interested in mining association rules from data cubes to explain the relationships amongst the multidimensional data. Since their extraction, most of the generated association rules benefit from the multidimensional data features, *i.e.*, dimensions, measures, concept hierarchies. However, deriving strong associations among data at low levels of abstraction seems to be in the multidimensional space an effortful task due to the sparsity of data. Thus, providing capabilities to mine association rules at multiple levels of abstraction and traverse easily among different abstraction spaces are efficiently carried out using the Multi-level association rules (MLAR). To mine MLAR, concept hierarchies should be provided for generalizing primitive level concepts to high level ones.

Unfortunately, only simple hierarchy is mainly used in such a mining of association rules from data cubes. In fact, the simple hierarchy describes the relationship between the members of the dimension can be represented by a tree. Nevertheless, in real situations, the dimension can be aggregated using several relationship analysis. So that, the granularity levels can form more than one hierarchy. Hence, investigating this analysis context on the mining process may efficiently explore such variety of dimensional analysis views leading to more specific rules fitting the user expectations.

In this paper, we focus on cyclic patterns which aim to discover rules that occur in user-defined intervals at regular periods. Our main claim is to generalize the use of concept hierarchies for dimensions during the mining process. The main idea behind our approach is to combine the multiple-levels forming the concept hierarchies and the parallel concept hierarchies which are employed to express several granularities of given dimension depending on the analysis context. Hence, we provide a comprehensive framework for the multi-level hybrid cyclic patterns extraction.

The remainder of the paper is organized as follows. The section 2 introduces a motivating example illustrating our contribution. In section 3, we present a survey of some related works. We briefly define the foundations of our method in section 4. We describe our algorithm MIHYCAR for multi-level hybrid cyclic patterns mining in section 5. Through extensive carried out experiments performed on real data warehouse, we stress on the performance of our approach in section 6. Finally, section 7 presents a conclusion resuming the strengths of our contribution and sketches future research directions.

II. Motivating example

In order to illustrate our contribution, we assume the sales data cube depicted by the figure 1 and defined over three dimensions, namely: the `Time` T of the transactions, the `Item` I which was bought, the `Point Of Sale` *POS* where the item is bought.

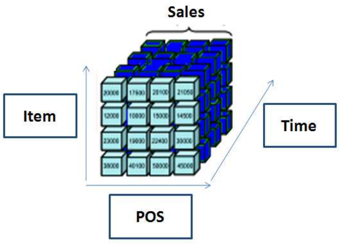


Fig. 1. Sales data cube.

We provide the dimensional concept hierarchies of the data cube in the following. Figure 2 illustrates the concept hierarchies for both `Time` dimension and `Item` dimension. Such concept hierarchies are known as simple hierarchy because their members can be represented using only one tree. Nevertheless, the `Point of sale` dimension is described using two hierarchies as depicted by figure 4 : the first hierarchy is composed of `POS -> City -> Country -> All`, and the other is represented by `POS -> Sales Group Division -> Sales Group Region->All`. These hierarchies on `Point of sale` dimension account for different analysis criteria, for example, the member values of `Point of sale` can be analyzed by geographic location or organization structure criteria. Apparently, such hierarchies are mutually non-exclusive, *i.e.*, it is possible to compute the aggregates grouped by both geographic location and/or organization structure (see figure 2).

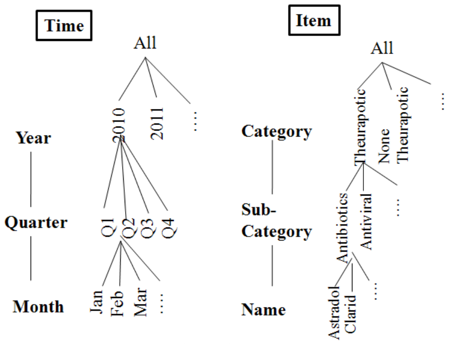


Fig. 2. Concept hierarchies of the `time` and `item` dimensions.

The expert in such a context needs to analyze the cyclic correlation existing between the `item` such as *Astradol* and its `point of sale` provided through its sales group division such as *SGDIV1* and its geographic position such as *Tunis*. Such correlation is cyclic and it is repeated every month in the sales data cube (see Table I).

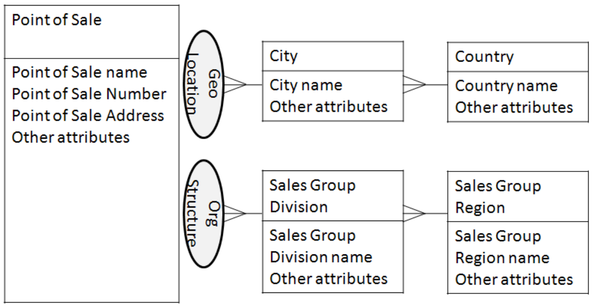


Fig. 3. Parallel concept hierarchies associated to the `point of sale` dimension.

We aim at building rules combining several dimensions while each dimension is formed using simple or parallel hierarchies depending on the user analysis requirements.

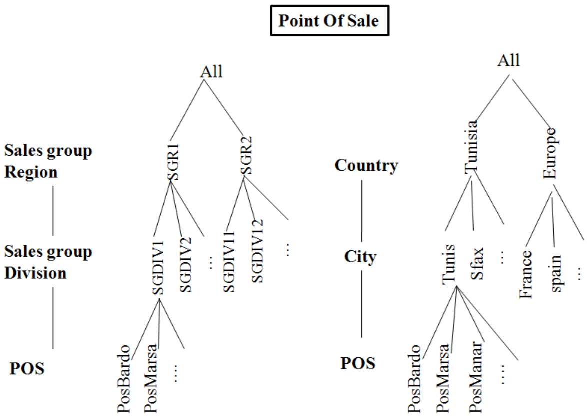


Fig. 4. Parallel concept hierarchies of the `point of sale` dimension.

Thus, the derived patterns answer different analytical purposes, and it makes sense to explain the correlation within the multi-faced data.

`Time` **T**	`Item` **I**	`Point Of Sale` **POS**
Jan 2010	Astradol	PosBardo
Feb 2010	Astradol	PosBardo
Mar 2010	Astradol	PosBardo
Apr 2010	Astradol	PosBardo
May 2010	Clarid	PosMarsa
Jun 2010	Clarid	PosMarsa

TABLE I

TABLE $\mathscr{T}$

III. RELATED WORKS

In this section, we focus on the various research works closely related to the cyclic pattern extraction and multidimensional association rules mining.

A. *Cyclic patterns*

The extraction of the CAR is a major issue in the data mining field. It was introduced by Ozden et *al.* (1998). It involves the association rules mining from articles characterized by their regular variation over time. Indeed, these association rules can highlight the daily, weekly, quarterly, or annual regular variation which is naturally cyclic. Discovering such regularities on the behavior of association rules allow marketers, for example, to better identify sales trends and provide a relevant prediction of future requests. The transactional data for analysis are time-stamped and that time intervals are specified by the user to divide the data into disjoint segments. Generally, users opt for ”natural” data segmentation based on the months, weeks, days, etc. Indeed, users are the ablest to make such a decision based on their data comprehension. We present briefly the basic concepts related to cyclical patterns. The databases which are based on the cyclical pattern extraction data have three closely related problem of the consumer basket, the first is an identifier on the client, the second is a list of products and the third represents the date that this customer bought this product package. The database is composed of itemsets identified by date and customer ID. A *cycle* is a period in time characterized by its length (*a month in our case*). The database is therefore considered as a set of cycles of fixed length specified by the user. A *cyclic item* is an assigned value for the attribute that is repeated cyclically according to the length of the cycle (*Astradol* occurs each month of 2007). A *Cyclic itemset* is a set of cyclic items. For example (*Astradol*, *Clarid*) is a cyclic itemset if it appears during the first and the second quarter of 2007. The crucial challenge of CAR mining algorithms is the best extraction of the frequent cyclic patterns. Several algorithms were proposed such as INTERLEAVED and SEQUENTIAL introduced by [11] or MTP presented by Thuan [14], [15] or the Chiang's method to combine cyclic and sequential patterns [5] or PCAR, proposed by [3]. These propositions rely on *generate and prune paradigm* where candidates are generated then unfrequent ones are pruned.

B. *Multi-dimensional association rules mining*

We shed light on the hierarchical aspect on the survey of multidimensional association rules.

Method	Temporality			Dimension			Hierarchy		Constraint	
	Non-temporal	Sequential	Cyclic	Intra-dimensional	Inter-dimensional	Hybrid	Single-level	Multi-level	constraint-based	Without constraints
(Kamber et *al.*,1997)	x				x		x			x
(Zhu,1998)	x			x	x	x	x			x
(Odzen et *al.*,1998)			x	x			x		x	
(Imielinski et *al.*,1999)	x				x		x		x	
(Thuan,2004,2008)			x	x			x		x	
(Tjioe and Taniar,2005)	x				x			x	x	
(Ben Messaoud et *al.*,2006)	x				x		x			x
(Chiang et *al.*,2009)		x	x	x			x		x	
(Plantevit et *al.*,2010)		x			x			x	x	
(Ben Ahmed and Gouider,2010)			x	x			x			x
(Ben Ahmed and Gargouri,2010,2011)			x		x		x		x	
(Our approach,2011)			**x**			**x**		**x**	**x**	

Fig. 5. Comparison of cyclic and multidimensional association rules approaches.

Based on this criterion, we can distinguish two types of rules: *(i)* Single-level association rules, *(ii)* Multi-level association rules.

1) Single-Level association rules: Most of related works neglect the multiple dimensional granularities levels. Kamber et *al.* introduced the mining of association rules from data warehouses [10]. In [16], underlying the number of involved dimensions and predicates in the association rule, Zhu introduces three classes of association rules, *i.e.*, *(i)* intra-dimensional (association within one dimension), *(ii)* inter-dimensional (association among a set of dimensions), *(iii)* and hybrid association mining (association among a set of dimensions with some items belonging to the same dimension). Ben Ahmed and Gargouri study the CAR mining from several dimensions [2]. After that, Ben Ahmed et *al.* involve the measures during the CAR extraction from data cubes [1].

2) Muliple-levels association rules: The approach of Imielinski et *al.* is the first work dealing with the multi-level association rules over data warehouses. Then, Tjioe and Taniar present a method for association rules extraction from multiple dimensions whithin several levels of abstraction [13]. Plantevit et *al.* take advantage of the different dimensions and levels of granularities to mine sequential patterns [12]. However, all the multi-level association rules consider only one concept hierarchy associated to each involved dimension. Nevertheless, some dimensions associate several hierarchies according to different analysis criteria. Such hierarchies are very frequent and called parallel hierarchies. Only few works handle the association rules mining from parallel dimensional association rules. To overcome this drawback, we investigate an evolving of such hierarchies to derive patterns rules within repetitive predicates.

IV. Formal background

In this section, we introduce the basic notions then we present our innovative key concepts that will be of use in the remainder.

A. Dimensions and hierarchies

Definition 1: (***Concept Hierarchy for dimension***)
A `concept hierarchy` for dimension is a tree whose nodes are elements belonging to the domain of this dimension [9]. It is a set of binary relationships between dimension levels. A dimension level participating in a hierarchy is called `hierarchical level` or in short `level`. The sequence of these levels is called a `hierarchical path` or in short path. The number of levels forming a path is called the `path length`. The first level of a hierarchical path is called `leaf` and the last is called `root` generally denoted by `ALL`. The `root` represents the most generalized view of data. The `edges` are considered as `is-a` relationships between members. Given two consecutive levels of a hierarchy, the higher level is called `parent` and the lower level is called `child`. Every instance of a level is called `member`.

Example 1: The concept hierarchy of the `Time` dimension is depicted by the figure 2. The *ALL* attribute is the root, the *Month* is the child and 2011 is the member.

Several types of concept hierarchies for dimension may be underlined. In our context, we focus on the parallel concept hierarchies.

Definition 2: (***Parallel concept hierarchies for dimension***)
Parallel hierarchies arise when a dimension has associated several hierarchies accounting for different analysis criteria. Such hierarchies can be independent or dependent. In a parallel independent hierarchies, the different hierarchies do not share levels, *i.e.*, they represent non-overlapping sets of hierarchies.

Example 2: An example of parallel concept hierarchies is depicted by figure 4. In the first concept hierarchy of `point of sale`, each *POS* is mapped into corresponding *city*, which is finally mapped into a corresponding *country*.

And the second concept hierarchy, each *POS* is mapped into *sales group division*, which is mapped into *sales group region*.

In this setting, we propose our key concepts.

B. Dimensions Partition

We consider that all is set in a multidimensional context. The three necessary data for cyclic mining drawn from classic context (Customer, Product, Date) become in a multidimensional context sets.
We consider that the table T, related to the sales data issued by customers, defined on a set $\mathscr{D}$ of n dimensions is partitioned into two sets:

- *Context dimensions* $\mathscr{D}_C$ which concern the investigated dimensions;
- *Out of context dimensions* $\mathscr{D}_{\overline{C}}$ related to the rest of uninvestigated dimensions or the complementary dimensions.

The context dimensions can be divided into three subcategories: (i) *Temporal dimension* $\mathscr{D}_T$: introducing a relation of temporal order (date in classical context), (ii) *Reference dimensions* $\mathscr{D}_R$: the table is segmented according to the reference dimensions values (customer in classical context), and (iii) *Analysis dimensions*: $\mathscr{D}_A$= $\{\mathscr{D}_1$,..., $\mathscr{D}_m\}$ with $\mathscr{D}_i$ $\subset$ Dom($\mathscr{D}_i$) corresponding to products in the classic context and relative to dimensions from which will be extracted the cyclic correlations.

Example 3: In our running example shown by table 1, we consider the whole table as our context composed of : *(i)* context dimensions $\mathscr{D}_C$=$\{T, I, POS\}$ with the temporal dimension $\mathscr{D}_T$=$\{T\}$, the reference dimension $\mathscr{D}_R$= $\emptyset$ and the analysis dimensions $\mathscr{D}_A$=$\{I, POS\}$.

C. Concept Hierarchies Partition

The analysis dimensions may be organized using one or more concept hierarchies. The latter can be partitioned into two sets:

- *Context concept hierarchies* $\mathscr{H}_C$ concern the set of involved concept hierarchies related to the analysis dimensions $\mathscr{D}_A$;
- *Out of context concept hierarchies* $\mathscr{H}_{\overline{C}}$ which report the set of unexplored concept hierarchies related to the analysis dimensions $\mathscr{D}_A$.

Let $\mathscr{T}_{1\mathscr{D}A}$ = $\{\mathscr{T}_{\mathscr{D}A1}$, . . . , $\mathscr{T}_{n\mathscr{D}Am}$ $\}$ the set of the n concept hierarchies associated to the m analysis dimensions. The elements of the analysis dimension $\mathscr{D}_{A1}$ are summarized using k concept hierarchies organizing the hierarchical relationships between the elements of this dimension :
$\mathscr{T}_{\mathscr{D}A1}$= $\{\mathscr{T}_{1\mathscr{D}A1}$,..., $\mathscr{T}_{k\mathscr{D}A1}\}$.

We assume that the k concept hierarchy of the i analysis dimensions $\mathscr{T}_{k\mathscr{D}Ai}$ is an oriented tree; $\forall$ node n_i $\in$ $\mathscr{T}_{k\mathscr{D}Ai}$, label($n_i$) $\in$ Dom($\mathscr{D}_{Ai}$).

Example 4: In our running example in the respect of the concept hierarchies shown by figures 2 and 4, we consider $\mathscr{T}_{\mathscr{D}A}$ = $\{\mathscr{T}_I$, $\mathscr{T}_{1POS}$, $\mathscr{T}_{2POS}\}$, with the $\mathscr{T}_I$ is illustrated by figure 2, $\mathscr{T}_{1POS}$ is depicted in the left side of figure 4 and $\mathscr{T}_{2POS}$ is shown by the right side of figure 4.

D. Generalization / Specialization in the concept hierarchies

We denote by ▲ x (respectively ▼ x) the set containing x along with all generalizations (respectively specializations) of x with respect to $\mathscr{T}_{\mathscr{D}A1}$ that belong to Dom($\mathscr{D}_{A1}$). Each analysis dimension $\mathscr{D}_{Ai}$ is instantiated using only one value d_{Ai} considered as node having the leaf label in the k concept hierarchy associated to the dimension $\mathscr{D}_{kAi}$.

Example 5: In our running example shown by figure 4, we consider x =*Tunis* $\in$ $\mathscr{T}_{2POS}$; the specialization of x is *i.e.*, ▼ x= ▼ *Tunis* =*PosBardo* and the generalization of x is *i.e.*, ▲ x= ▲ *Tunis* = *Tunisia*.

E. Multi-level Dimensional Cyclic Item and Multi-level hybrid Cyclic Itemset

Definition 3: (**Multi-level Dimensional Cyclic Item**)
Let the analysis dimensions $\mathscr{D}_A$ = $\{\mathscr{D}_1$,...,$\mathscr{D}_m$ $\}$ and a cycle length l. A multi-level dimensional cyclic item α is an item belonging to one of the analysis dimensions, namely $\mathscr{D}_k$ and having a value of d$_k$ for the date t and the date $t+l$ with d_k $\in$ $\{\mathscr{T}_{\mathscr{D}k}\}$ and such that $\forall$ k $\in$ [1,m], d$_k$ $\in$ Dom($\mathscr{D}_k$).

Unlike the transactional databases, a multi-level dimensional cyclic item can be generalized using any value node associated to d_i in the k concept hierarchy without necessarily being a leaf.

Example 6: Typical example of multi-level dimensional cyclic item, considered in the multidimensional context, shown by the table 1 and the delimitation of the context considered previously, is α= (*PosBardo*) because it belongs to the `POS` dimension, being a part of analysis dimension and its value *PosBardo* belongs to the `POS` domain and is repeated each month of the first quarter of 2010.

Definition 4: (**Multi-level Hybrid Cyclic Itemset**)
A multi-level hybrid cyclic itemset F defined on $\mathscr{D}_A$ = $\{\mathscr{D}_1$,...,$\mathscr{D}_m$ $\}$ is a nonempty set of multi-level dimensional cyclic items F = $\{\alpha_1$,...,α_m $\}$ with $\forall$ j $\in$ [1, m], α_j is a multi-level dimensional cyclic item defined on $\mathscr{D}_j$ at the date t and it is repeated at each date $t+l$ with $\forall$ j,k $\in$ [1, m], $\alpha_j \neq \alpha_k$.

Example 7: An example of multi-level hybrid cyclic itemset is F=[*Astradol*, *PosBardo*] because it is composed of two multi-level hybrid cyclic items *i.e.*, α_1=(*Astradol*), α_2=(*PosBardo*). It is repeated monthly during the first quarter of 2010.

F. Connectivity of multi-level hybrid cyclic itemsets

We study the connectivity by scrutinizing the different relationships that may exist between the multi-level hybrid cyclic itemsets. Let two multi-level hybrid cyclic itemsets F=(d_1,..., d_m) and G=(d'_1,..., d'_m), two types of connectivity between those itemsets are considered:

1) Connected multi-level hybrid cyclic itemsets;
2) Disconnected multi-level hybrid cyclic itemsets.

Definition 5: (***Disconnected multi-level hybrid cyclic itemsets***)
F and G are disconnected iff they do not belong to the same concept hierarchies.

Example 8: *F=SGDIV1* and *G=Tunisia* are disconnected because they do not belong to the same concept hierarchies, *F=SGDIV1* $\in$ $\mathscr{T}_{1POS}$ and *G= Tunisia* $\in$ $\mathscr{T}_{2POS}$.

Definition 6: (***Connected multi-level hybrid cyclic itemsets***)
F and G are connected iff they belong to the same concept hierarchies.

Example 9: *F=PosBardo* and *G=Tunisia* are connected because they belong to the same concept hierarchy.

If the multi-level hybrid cyclic itemsets are connected, two classes of relationships may be outlined:

1) Covered multi-level hybrid cyclic itemsets;
2) Uncovered multi-level hybrid cyclic itemsets.

Definition 7: (***Covered multi-level hybrid cyclic itemsets***)
F is covered by G iff $\forall$ d_i, d_i = ▲ d'_i or d_i = d'_i.

Example 10: F=[*Astradol*,*PosBardo*] is covered by G=[*Antibiotic*,*Tunis*] because *Tunis*= ▲ *PosBardo* and *Antibiotic*= ▲ *Astradol*.

Definition 8: (***Uncovered multi-level hybrid cyclic itemsets***)
F is un covered by G iff $\forall$ d_i, ▲ d_i $\neq$ d'_i.

Example 11: F=[*Astradol*,*PosBardo*] is covered by G=[*Antiviral*,*Tunis*] because ▲ $Astradol \neq Antiviral$.

If two multi-level hybrid cyclic itemsets are covered, two eventual relationships may be highlighted:

1) Adjacent multi-level hybrid cyclic itemsets;
2) Non adjacent multi-level hybrid cyclic itemsets.

Definition 9: (***Adjacent multi-level hybrid cyclic itemsets***) F belongs to the n hierarchical level, G is considered as its adjacent multi-level hybrid cyclic itemset iff G belongs to $n-1$ level or $n+1$ level of the same concept hierarchy.

Example 12: F=[*Astradol*,*PosBardo*] belonging to the 1-level is adjacent to G=[*Antibiotic*,*Tunis*] because G belongs to the 2-level in the concept hierarchies depicted by both figure 2 and figure 4.

Definition 10: (***Non adjacent multi-level hybrid cyclic itemsets***)
F belongs to the n hierarchical level, G is considered as a non adjacent multi-level hybrid cyclic itemset of F iff G does not belong to $n-1$ level or $n+1$ level of the same concept hierarchy.

Example 13: F=[*Astradol*,*PosBardo*] belongs to the 1-level and is not adjacent to G=[*Africa*,*Therapeutic*] because G belongs to the 3-level in the concept hierarchies which is not the 2-level in the concept hierarchies.

G. Support of multi-level hybrid cyclic itemset

Definition 11: : (***Support of multi-level hybrid cyclic itemset***)
- The support of multi-level hybrid cyclic itemset, denoted $Supp(F)$ is the number of tuples that contain the itemset; $Supp(F) = COUNT(F)$.

Example 14: Consider the context shown by the table I and the delimitation already presented. The multi-level hybrid cyclic itemset F=(*Antibiotics,PosBardo,SGDIV1*) has an absolute support related to the sales of the products considered as *Antibiotics* and which are sold in the first sales group division *SGDIV1* in *PosBardo*:

$$\mathbf{Supp}(\mathit{Antibiotics,PosBardo,SGDIV1}) =$$

$$\mathbf{COUNT}(\mathbf{I} = \mathit{Antibiotics,}\mathbf{POS} = \mathit{PosBardo} \bigwedge \mathit{SGDIV1}) = 4$$

H. Support and Confidence Computing of Multi-level Hybrid Cyclic Rule

Definition 12: : (***Support of multi-level hybrid cyclic rule***)
- The rule support $R : F \Rightarrow G$, denoted $Supp(R)$, is equal to the ratio of the number of tuples that contain F and G to the total number of tuples in the sub-cube.

$$Supp(R) = \frac{COUNT(F \cup G)}{COUNT(ALL,ALL)};$$

The support of de R, $Supp(R) \in [0, 1]$.

Definition 13: : (***Confidence of multi-level hybrid cyclic rule***)
- The rule confidence $R : F \Rightarrow G$, denoted $conf(R)$, is equal to the ratio of the number of tuples that contain F and G to the number of tuples that contain F in the sub-cube.

$$conf(R) = \frac{Supp(R)}{Supp(F)};$$

The confidence of R, $conf(R) \in [\ 0,\ 1\ ]$.

Example 15: In our running example, the rule R: *Antibiotics*,*PosBardo* $\Rightarrow$ *SGDIV1* has :

- $Supp(R) = \mathbf{COUNT}(\mathbf{I} = \mathit{Antibiotics}, \mathbf{POS} = \mathit{PosBardo} \bigwedge \mathit{SGDIV1}) = 4$
- $conf(R) = \frac{\mathbf{COUNT}(\mathbf{I}=\mathit{Astradol},\mathbf{POS}=\mathit{PosBardo} \bigwedge \mathit{SGDIV1})}{\mathbf{COUNT}(\mathbf{I}=\mathit{Astradol},\mathbf{POS}=\mathit{PosBardo})} = \frac{4}{4} = 1$

I. MIHYCAR: A Method for MultI-level HYbrid Cyclic Association Rules

A method for mining multi-level hybrid cyclic association rules is introduced in this section, which uses a hierarchy-information encoded multi-dimensional data cube instead of the classical data cube. Indeed, it is advantageous to encode the relevant data. Such encoded predicate string is composed as follows '[*d*-*h*-*l*-*k*]' with d is the dimension, h is the concept hierarchy of the d dimension, h indicated the abstraction level in the concept hierarchy, and finally k represents the number of itemsets. For example, *Tunisia* is encoded using the following string [3-2-3-1] with 3 represents the dimension Point of Sales, 2 represents the second hierarchy concept and 3 describes the level of abstraction in the concept hierarchy, finally 1 represents 1-item. Indeed, such encoding requires fewer bits than the corresponding object-identifier or bar-code. To illustrate the encoding method, we present an abstract example which simulates the real life example illustrated by the table II.

Item	Encoded Item	POS	Encoded POS
Astradol	[2-1-1-1]	*PosBardo*	[3-*-2-1]
Clarid	[2-1-1-1]	*PosMarsa*	[3-*-2-1]

TABLE II

ENCODED DATA CUBE *T*

The process of mining of multi-level hybrid cyclic rules is performed using our algorithm MIHYCAR which proceeds as follows.

Notation	Description
SC	: Sub-Cube
lc	: Length of Cycle
d	: Current dimension
nd	: Number of dimensions
l	: Current level
h	: Current hierarchy of dimension
depth	: Depth of the current concept hierarchy
$\mathcal{D}_t$	: Date *t*
$\mathcal{M}insupp$	: Minimum Support Threshold
$\mathcal{C}[d,h,l,k]$	: Set of candidates from the dimension *d* belonging to the hierarchy *h* and the level *l* having *k* itemsets
(resp.$\mathcal{F}[d,h,l,k]$)	: Set of frequents from the dimension *d* belonging to the hierarchy *h* and the level *l* having *k* itemsets
s	: nonempty subset *s* of $\mathcal{F}_i$
Supp($\mathcal{C}$)	: Support of the multi-level hybrid cyclic itemset $\mathcal{C}$

TABLE III

LIST OF USED NOTATIONS IN THE MIHYCAR ALGORITHM.

Algorithm 1: MIHYCAR: MultI-level Hybrid Cyclic Association Rules

```
Data: $SC$, $\mathscr{M}insupp$
Result: Multiple-levels frequent itemsets.
begin
    // initialisation $d$=1; $h$=1;$l$=1;
    $\mathscr{F}[d,h,l,1]$= Find 1-frequent cyclic itemsets($SC$, $l$,$\mathscr{D}_t$, $\mathscr{M}$inSupp) ;
    for ($d$=1; $d <= nd$ ; $d$++ ) do
        //scan of dimensions
        for ($h$=1; $h < depth$; $h$++) do
            //scan of concept hierarchies of each dimension
            for ($l$=1; $\mathscr{F}[d,h,l,1] \neq \emptyset$; $l$++ ) do
                //scan of concept hierarchies levels of each dimension
                for ($k$=2; $\mathscr{F}[d,h,l,k-1] \neq \emptyset$; $k$++ ) do
                    $\mathscr{C}[d,h,l,k]$ = CandidatGeneration ($\mathscr{F}[d,h,l,k$-1]);
                    if $\mathscr{C}[d,h,l,k]$ is a hybrid cyclic itemset then
                        foreach transaction $\mathscr{T} \in SC$ at date $\mathscr{D}_t$ do
                            $\mathscr{C}[d,h,l,t]$=subset($\mathscr{C}[d,h,l,k]$, $\mathscr{T}$)
                            foreach candidat $\mathscr{C} \in \mathscr{C}\mathscr{C}[d,h,l,t]$ do
                                $\mathscr{C}$.support = SupportComputing($SC$, $l$,$\mathscr{D}_t$, $\mathscr{C}$);
                            $\mathscr{F}[d,h,l,k]$ = { $\mathscr{C} \in \mathscr{C}[d,h,l,k]$, $\mathscr{C}$.support > $\mathscr{M}insupp$ }
                Return $\mathscr{F}[d,h,l,k] = \cup_k \mathscr{F}[d,h,l,k]$ ;
end
```

Starting at dimension 1, we scan all the concept hierarchies related to the first dimension. Thus, we derive for each level l, the frequent multi-level dimensional cyclic items $\mathscr{F}[1,h,l,1]$. In fact, the multi-level dimensional cyclic item is frequent if it is cyclic otherwise in the respect of the length of cyclic specified by the use, its cyclic occurrences exceed the minimum support threshold (see procedure *ComputingSupport*).

For each level l, we extract the frequent multi-level hybrid cyclic itemsets. In fact, only the descendants of frequent multi-level hybrid cyclic itemsets at level l are considered as candidates in the level-l+1 frequent itemsets. A scan of dimensions is performed. In fact, we apply the anti-monotony property which states that for each non frequent itemset, all its super-itemsets are drastically not frequent. This property is projected in the multi-level granularities space in order to enable an outstanding reduction of the search space.

After finding the frequent multi-level hybrid cyclic itemsets, the set of multi-level hybrid cyclic association rules can be derived according to the minimum confidence threshold $MinConf$. An example of generated rule is R : $Antibiotics, PosBardo \Rightarrow SGDIV1$.

```
Function Find 1-frequent cyclic itemsets ($SC$, $l$,$\mathscr{D}_t$, $\mathscr{M}$inSupp)
Result: $\mathscr{F}_1$
begin
    while (!End of tuples in SC) do
        foreach transaction $\mathscr{T} \in SC$ do
        foreach item $\alpha \in \mathscr{T}$ do
        foreach transaction $\mathscr{T}' \in SC$ at date $\mathscr{D}_{t+l}$ do
        $Supp(\alpha)$=COUNT($\alpha$);
        if ($Supp(\alpha) > \mathscr{M}inSupp$ ) then
            $\mathscr{F}[d,h,l,1] = \mathscr{F}[d,h,l,1] \cup \alpha$;
    Return $\mathscr{F}[d,h,l,1]$ ;
end
```

```
Function SupportComputing ($SC$, $l$,$\mathscr{D}_t$, $\mathscr{C}$)
Result: $Supp(\mathscr{C})$
begin
    NoMoreCyclic: Boolean;
    NoMoreCyclic = false;
    while ((!End of tuples in SC) and (!NoMoreCyclic)) do
        $\mathscr{C}[d,h,l,k]$ = CandidatGeneration ($\mathscr{C}[d,h,l,k-1]$);
        foreach transaction $\mathscr{T} \in SC$ at date $\mathscr{D}_{t+l}$ do
        if $\mathscr{C}$ exists in $\mathscr{T}$ then
            $Supp(\mathscr{C})$= $Supp(\mathscr{C})$+1;
        NoMoreCyclic = true;
    Return $Supp(\mathscr{C})$ ;
end
```

V. Experimental Study

All experiments were carried out a PC equipped with 1.73 GHz and 1 GB of main memory.

In the following, we report experiments performed on a real sales data warehouse [1], which contains three dimensions (e.g., `Time` dimension, `Item` dimension, `point of sale` dimension) and one sales fact table.

The data warehouse is built using relational OLAP (ROLAP) and is modeled in a star schema, which contains dimension tables for the hierarchies and a fact table for the dimensional attributes and measures.

Our objective is to show, through our extensive experimental study: (i) the performance of our algorithm according to the length of cycle and the number of analysis dimensions; (ii) the assessment of the hierarchical aspect in respect of the number of involved concept hierarchies and the average depth of those hierarchies.

Figure 6.(a) plots the runtime needed to generate multi-level hybrid cyclic association rules with the respect of the length of cycle. Clearly, in efficiency terms, it can be seen from this figure that the running time decreases proportionally to the length of cycle.

[1] The data warehouse is related to pharmaceutical listed company. It is built using the available information at http ://www.bvmt.com.tn/companies/ ?view=listed.

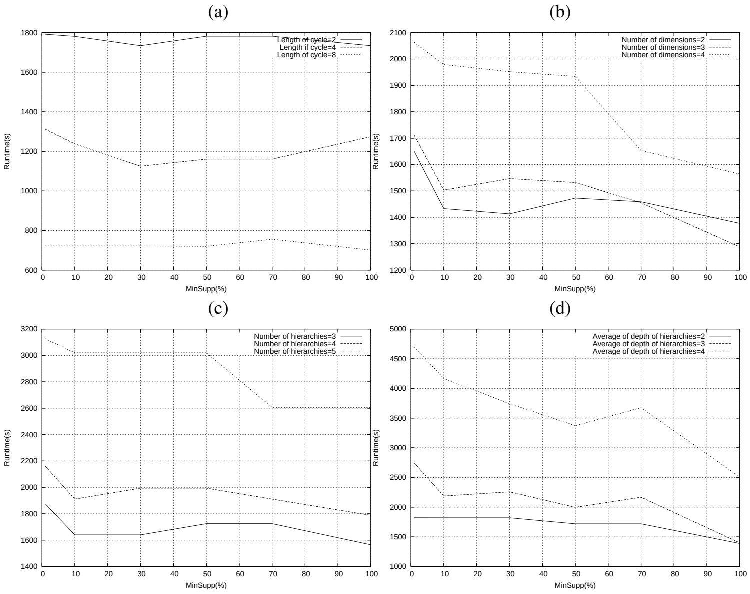


Fig. 6. Performance of our algorithm in respect of the (a) **length of cycle**, (b) **dimensions number**, (c) **number of concept hierarchies**, (d) **average depth of the concept hierarchies**.

Figure 6.(b) describes the behavior of our approach in terms of runtime according to number of analysis dimensions. Obviously, we observe that the slopes of the three plots are increasing when the number of analysis dimensions increases. In fact, having more analysis dimensions, more concept hierarchies will be included. So that, the number of generated patterns will highly increase.

Through the last experiments, we compare the runtime needed to generate multi-level hybrid cyclic rules over the number of involved concept hierarchies and the average depth of the concept hierarchies, also called the specialization level.

Moreover, as shown in figure 6.(c), the number of concept hierarchies related to the analysis dimensions radically influences the performance of our algorithm. Taking more hierarchies into account through parallel hierarchies involving, the runtime of our algorithm significantly increases.

Figure 6.(d) shows the number of generated multi-level hybrid cyclic association rules over the depth of the concept hierarchies. In fact, increasing the size of the concept hierarchies brings additional specialization level. Accordingly, our algorithm mines less frequent patterns until it cannot mine any more knowledge.

VI. Conclusions and future works

We have extended the scope of the study of mining cyclic association rules from single level to multiple concept levels and studied methods for mining multiple-level hybrid cyclic association rules from data warehouses. Mining such patterns may lead to progressive mining of refined knowledge from data. Exclusively, our method extracts patterns in the respect to parallel concept hierarchies of dimensions which organize the attribute values into different levels of abstraction related to different analysis criteria. The performance study underlines the utility of our approach. The extension of our method for future works addresses the following issues: (i) It is interesting to develop efficient algorithms for mining multiple-level hybrid cyclic rules under crossing levels on concept hierarchies; (ii) Involving the independence of the parallel concept hierarchies, an extension of our work by considering both of dependent and independent concept hierarchies; (iii) Reducing the redundant rules and filtering the uninteresting patterns.

References

[1] E.Ben Ahmed, A.Nabli and F.Gargouri, “Usage Des Mesures Pour La Gnration Des Rgles d’Associations Cycliques“, *7 me confrence francophone sur les entrepts de donnes et l’analyse en ligne (EDA’11)*,2011, *To appear.*

[2] *E.Ben Ahmed and F.Gargouri, “Règles d’association cycliques dans un contexte multidimensionnel“, Atelier des Systmes Décisionnels (ASD’10), Tunisia, 2010.*

[3] *E.Ben Ahmed and M.S.Gouider, “Towards a new mechanism of extracting cyclic association rules based on partition aspect“, IEEE International Conference on Research Challenges in Information Science, 2010, pp 69–78,2010.*

[4] *R.Ben Messaoud, O.Boussaid, S.L Rabasda and R.Missaoui, “Enhanced mining of association rules from data cubes“, Proceedings of the 9 th ACM International Workshop on Data Warehousing and OLAP (DOLAP 2006), pp 11–18, 2006.*

[5] *D.Chiang, C.Wang, S.Chen and C.Chen, “The Cyclic Model Analysis on Sequential Patterns“, IEEE Trans. on Knowl. and Data Eng., pp 1617–1628, 2009.*

[6] *G.Dong, J.Han, J.Lam, J.Pei, K.Wang and W.Zou, “Mining Constrained Gradients in Large Databases“, IEEE Transactions on Knowledge Discovery and Data Engineering, 2004.*

[7] *J.Han, W.Gong and Y.Yin “Mining Segment-Wise Periodic Patterns in Time-Related Databases“, KDD, pp 214–218, 1998.*

[8] *J.Han, W.Gong and Y.Yin “Efficient Mining of Partial Periodic Patterns in Time Series Database“, ICDE, pp 106–115,1999.*

[9] *C.S.Jensen, T.B.Pedersen, C.Thomsen, “Multidimensional Databases and Data Warehousing“, 2010.*

[10] *M.Kamber, J.Han and J.Y.Chiang “Metarule-guided mining of multidimensional association rules using data cubes“, Proceedings of the 1997 International Conference on Knowledge Discovery and Data Mining (KDD’97) pp 207–210,1997.*

[11] *B.Ozden, S.Ramaswamy and A.Silberschatz , “Cyclic Association Rules“, Proceedings of the Fourteenth International Conference on Data Engineering pp 412–421, 1998.*

[12] *M.Plantevit, A.Laurent, D.Laurent, M.Teisseire and Y.Choong “Mining multidimensional and multilevel sequential patterns“, ACM Transactions on Knowledge Discovery from Data, pp 155–174,2010.*

[13] *H.C.Tjioe and D.Taniar, “Mining Association Rules in Data Warehouses“, IJDWM, pp 28-62, 2005.*

[14] *N.D.Thuan, “Mining Cylic Association Rules in Temporal Database “, The Journal Science and technology developement, Vietnam National University , pp 12–19, 2004.*

[15] *Thuan, N.D., “Mining Time Pattern Association Rules in Temporal Database“, SCSS, pp 7-11, 2008.*

[16] *H.Zhu, “On-line analytical mining of association rules“, Master’s thesis,Simon Fraser University, Burnaby, British Columbia, Canada, 1998.*

Application of Non Parametric Regression Network to model Risk Parameters for ranking countries to carry out business in Water, Electronics, Education, Pharmaceuticals, and Infrastructure Sectors

Jon Tong-Seng Quah and Prerna Mishra
Nanyang Technological University, 50 Nanyang Avenue, Republic of Singapore

Abstract-

Globalization creates investment opportunities having varying degree of risk. Companies are aggressively pursuing strategy to get into overseas business. There is uncertainty in identification and quantification of risks associated with each opportunity. The decision to select any country is contingent upon several aspects like country risk parameters (economic, social, and political), degree of corruption, ease of doing business, economic well being and status of development of the select industries. We felt a requirement to model these risk parameters to rank countries in terms of these factors by employing a methodology which would not be restricted by the analyst's conviction. We exercised nonparametric regression invoking Alternating Conditional Expectation (ACE) and General Method of Data Handling (GMDH) to rank countries in terms of specific risk parameters which may impact the global opportunities like water resources, electronic, education, infrastructure, and pharmaceutical. Present work is an attempt to demonstrate these models.

Keywords: Country risk, rank, industry, ACE, GMDH, Neuroshell

1 Introduction

Global business environment provides companies with myriads of business options which are beset with environmental, economic, social, and political or business risk. Country risk parameters play a pivotal role in strategic decisions to select overseas destination for investment. Some sort of country favorability matrix incorporating the risk elements is required to analyze any overseas business opportunity. Several institutions and agencies have developed ranking of countries in respect of key parameters considered to impact the attractiveness of any country. Earlier attempts mostly require quantification of the impacting parameters by assigning some weights and invoking mathematical model which required user intervention. On the contrary, the objective of our modeling technique is that data itself should lead to a mathematical model depicting the functional relationship between various modeling parameters by using modern nonparametric methods e.g. Alternating Conditional Expectation (ACE), and General Methods of Data Handling (GMDH).

2 Literature Review

Chng (2000) [3] has provided a detailed review on works in the area of global business strategy and international management and hase provided insight into focus for future research. Harvey et al (1996) [7] has discussed about country risk parameters (economic, social and political) and methodology to rank countries on the basis of these risk. An index of globalization has been developed by Dreher (2003) [5] taking into consideration economic, social and political aspects of countries.

A model to rank country by a matrix of weighted risk factors has been discussed by Surya et al (2008) [15]. The effort identified risk factors and assigned relative weights to these parameters. However, there is an element of uncertainty and analyst's bias in assigning weights to risk. In order to overcome such limitation we envisage invoking technique to develop rank of countries which overcome these limitations by integrating risk factors affecting global investment decisions. Nonparametric techniques viz. Alternating Conditional Expectation (ACE) [2] and Group Method of Data Handling (GMDH) have been widely applied to solve problems where it is always not possible to identify the relationship between dependent and independent variables.

Duolao Wang et al (2004), [6] Guoping et al (1997) [9,10] and Veugelers (2001) [16] have discussed the ACE algorithm. Sum et al (1995) [13] has attempted to model the effects of a service guarantee on perceived service quality using ACE. Sum et al (1995) [14] has applied ACE to model benefits of Material Requirements planning. GMDH has been used to automatically determine the network size and connectivity, and coefficients for network model (Osman, 2002) [11].

It has been observed that no previous research has focused on development of such model. Therefore, we thought it prudent to model the country specific parameters invoking nonparametric technique. The model may find practical application in deciding global business opportunities.

3. Factors of Consideration

The decision to enter any country requires addressing two aspects - how to globalize and where to globalize and the selection of any country is contingent upon the degree of attractiveness the country in question may hold to provide return on investment at profit.

3.1 Modeling Parameters

We distinguish two broad categories of modeling parameters: country specific variables and problem specific variable (table 1a). The "country specific parameters" include characteristics of countries, and the ranks/ index derived by various agencies taking into consideration these parameters have been considered. The "problem specific parameters" are those which may have significant bearing on the selection criteria. In the lack of any direct "industry or problem specific" indicator suitable proxies have been used (Table 1, 2).

Table 1 Predictor variables for modeling

Country specific parameters			
Risk	**Data Required**	**Data represented by**	**Data used**
Economic	Economic flows, Population density	FDI, FII, GDP, EFW, Population	FDI [19], EFW [8], GDP [19], Population [12]
Social	Status of growth of social institutions	Ease of doing business, Corruption Index, Population	EODB [4], Population [15], Corruption Perception Index (CPI) [18]
Political	Political institutions	Ease of doing business, Corruption Index	EODB, CPI
Attracting Global investment	Amount of FDI, Confidence of Global investors	Foreign Direct Investment (FDI)	FDI Index [19]
Economic restrictions	Imposed import barrier	Foreign Direct Investment	FDI Index
Taxes	Fiscal system	Fiscal policy	FDI Index
Problem (industry) specific parameters			
Industry	**Data required**	**Data represented by**	**Data used**
Water	Water availability	Per capita expenditure on drinking water	Water coverage [25]
Electronics	Growth of electronics industry	Intensity of application and usage of Electronics and computer	Number of internet users [19]
Education	Literacy, Growth and degree of development of educational system in any country	- Literacy rate, - Per capita number of educational institutions, - Per capita expenditure on education	Public spending on education, [19] Literacy rate [24]
Infrastructure	Growth of infrastructure (road, rail, airways, telecommunication etc) in any country	- Consumption of steel and cement, - Length of railways, road network, - Number of civilian airport, - Number of telecom subscribers. - Per capita expenditure on infrastructure	Infrastructure index [20]
Pharmaceuticals		- Per capita expenditure on education	World health systems, [21] Physicians per 1000 population, [22] Total health expenditure as percent of GDP [23]

Table 2. Input parameters for modeling country risk parameters

Abbreviations used in the table- CPI= Corruption Perception Index, EFW = Economic Freedom of world, EODB= Ease of Doing Business, KOF GI = KOF Globalization Index, POP= Population, GDP = Gross Domestic Product, FDI = Foreign Direct Investment, NINU=No. of internet users, PSE = Public Spending on education, WC = Water Coverage, WHS = WHO World Health System Index, LR = Literacy Rate, EOH = Expenditure on health as % of GDP, PHPTP = Physician per 1000 population, INFRI = Infrastructure Index (NOTE – In all cases 1=highest)

NO	COUNTRIES	PROXY FOR RESPONSE VARIABLE	PREDICTORS													
			A						B							
									B1		B2	B3	B4			B5
			Country Specific General Parameters						Industry Specific Parameters used as proxy for different industries							
		KOFGI	EFW	EODB	CPI	POP	GDP	FDI	WC	NINU	PSE	LR	WHS	EOH	PHPTP	INFRI
		The parameters in Part A have been used along with select parameters in part B to derive rank of countries for specific industry as per following scheme							Proxy for Water resource	Proxy for Electronics	Proxy for Education		Proxy for Pharmaceutical			Proxy for Infrastructure
		Water Resources = A + B1, Electronics = A + B2, Education = A + B3, Pharmaceuticals = A + B4, and Infrastructure = A + B5														
4	Australia	13	7	8	5	32	12	7	1	17	32	7	18	14	21	11
12	Cameroon	74	65	78	43	36	57	71	24	62	77	49	73	53	66	68
13	Canada	4	6	7	6	25	9	9	1	10	25	7	17	7	26	7
16	China	44	42	44	30	1	2	2	7	1	75	28	71	59	33	16
26	Egypt	47	40	49	37	12	18	24	4	19	49	44	37	42	57	66
34	India	69	45	65	31	2	4	10	17	4	56	53	62	57	54	29
35	Indonesia	57	49	59	37	4	11	29	22	8	79	30	53	81	67	47
38	Israel	28	39	17	15	55	37	23	1	45	9	16	15	27	5	13
40	Japan	33	16	11	10	10	3	12	1	3	58	7	4	19	29	4
58	Pakistan	65	57	42	41	6	19	31	11	14	71	56	63	82	52	71
64	Portugal	5	25	26	16	46	32	34	6	33	16	21	6	5	11	21
66	Russia	31	43	57	43	9	5	4	2	7	57	5	67	55	2	54
69	Singapore	12	1	1	3	62	35	13	1	39	53	25	1	75	36	3
79	United Kingdom	16	7	4	10	17	6	3	1	6	21	7	11	21	25	10
80	United States	19	4	3	11	3	1	1	1	2	17	7	22	1	22	5
	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	

P.S. Total 82 countries have been taken for Modeling Purpose and data for only few is exhibited in table 2. However, the figure 1 depicts the observation for 82 countries.

3.2 Predictor Variables

The following set of variables has been considered for modeling (Table 1 and 2). The Corruption Perceptions Index (CPI) rank shows how one country compares to others included in the index [17]. Economic Freedom of World (EFW) takes in to consideration security of property, money growth, inflation, freedom of nationals of any country to acquire property without the force, fraud and theft (protection of property rights); and the right to protection of their property from invasion by others, own foreign currency bank accounts and trade internationally, regulation of credit, labor, and business, integrity of the legal system, business regulations, tax rate, regulatory trade barriers, hiring and firing regulations, price controls, licensing restrictions etc [8]. The Ease of Doing Business (EODB) [4] refers to the degree of ease by which a business entity can start and do business in a country. We have used the Population data provided by World Bank [11]. GDP has been used to represent economic health of a country, and country's standard of living [19]. Foreign Direct Investment (FDI) indicates that the country with superior FDI has made possible a competitive and conducive atmosphere for the overseas investors to do business. The proxy for industry specific parameters is as described in Table 1.

3.3 Response Variables

We have used the index provided by KOF (Dreher, 2006) [5] referred to as KOF Globalization Index (KOF GI) as a proxy for response variable. (Table 2).

3.4 Linear Regression Model

Linear regression of KOF GI versus individual predictor variables result in varying correlation R2 as provided in brackets viz. CPI (0.65), EFW (0.53), EODB (0.56), Population (0.06), GDP (0.13), FDI (0.16), literacy rate (0.60), public spending on education (0.14), infrastructure index (0.57), water coverage index (0.55), number of internet users (0.16), world health index (0.45), physician per thousand population (0.63) and expenditure on health (0.34) respectively.

4. Experiment Design

We hypothesize that in our problem the considered predictor variables do not hold linear relationship with response variable. Thus, we intend to model these parameters by invoking nonparametric regression techniques such as ACE and GMDH. The rational behind preferring nonparametric to parametric lies in the fact that in parametric regression a model is fitted to data by assuming a functional relationship. But, it may not be possible to recognize the underlying functional relationship between dependent (response) and multiple independent (predictors). Nonparametric regression explores such relationship without *a priori* knowledge of the dependencies and the functional form is derived from data.

5. Experiment Results

The data in tables 2 has been subjected to modeling by invoking ACE and GMDH (ANN and Neuroshell). The output obtained from these methods compare well as depicted in figure 1 and table 3.

Table 3 Output of ACE and Neuroshell

		Electronics	Education	Water	Pharmaceuticals	Infrastructure
Output of ACE	Predicted Std dev	8.24	6.76	8.50	7.68	9.23
	Fitted Std dev	9.33	8.43	9.08	8.48	9.85
	Optimal Regression Correlation	0.94	0.96	0.94	0.95	0.92
	Optimal Inverse Trans. R^2	0.99	1.00	0.99	0.99	0.99
Output of Neuroshell	Correlation coeff.	0.92	0.94	0.93	0.83	0.90
	R squared	0.85	0.87	0.87	0.70	0.83

6. Analysis of Results

The results corroborate our hypothesis that there is nonlinear relationship between response and multiple predictor variables. It has been observed that the value of KOF GI has been impacted and the degree of impact varies from country to country.

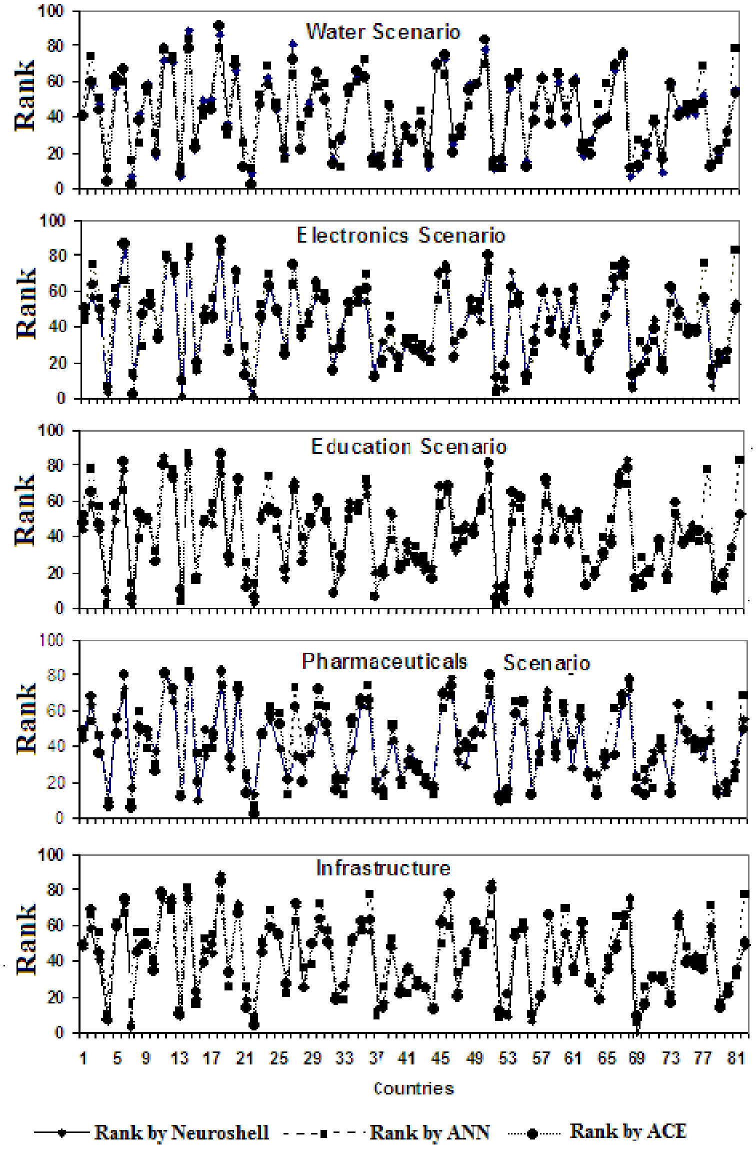


Figure 1 Ranks obtained by three Methods

7. Conclusions

Countries differ in degree of attractiveness for different business opportunities because of inherent risk. Companies desirous to pursue global business must evaluate and rank each opportunity. We propose a numerical modeling approach having multiple criteria decision analysis to evaluate and weigh risk factors for country selection. We have scanned novel modeling techniques like ACE and GMDH to model critical risk parameters to assess 82 countries. The research work has made an attempt to derive rank of countries for carrying out business in select target industry sectors viz. water, education, electronics, pharmaceuticals, infrastructure by overseas business entities. It has been observed that other numerical methods like assigning weights to different risk parameters, regression, and sensitivity analysis may provide comparative evaluation of countries. But, these methods require user intervention. However, the modeling of country risk parameters using ACE and GMDH does not require a priori assumptions of a functional form and the results are derived solely based on the data set.

We observe that both the country and industry specific risk factors have influence on ranking of countries and thus on global business scenario. Thus, industry specific parameters should not be the only criteria for evaluating countries but, on the country, it should take into consideration the country specific parameters like GDP, population, degree of corruption, FDI etc. The comprehensive model thus derived will facilitate in assessment of countries and help decision makers to prefer overseas business opportunities. Despite all these the work has some limitations. Any numerical modeling requires robust data. We have developed the model taking into consideration only 82 countries for which uniform data for all risked parameters could be obtained from public domain data base. The model may have given some other output has it been subjected to more populated dataset. This leaves a room for future work

References

1. Ang, S. K., Quek, S. A., Teo, S. H., & Liu, Modeling is planning benefits using ACE, Decision Sciences, Vol. 30 (2), pp. 533-562, 1999.

2. Briemann, L. and Friedman, J. H., (1985); Estimating Optimal Transform for multiple Regression and correlation, J. Amer. Statistical Asso., Sept 1985, 1985.

3. Chng, P., Research on global strategy, International Journal of Management Reviews, Mar. 2000, Vol. 2(1).

4. Doing Business 2007: How to Reform, A co publication of the World Bank and the International Finance Corporation, 2007.

5. Dreher, Axel, Does Globalization Affect Growth? Empirical Evidence from a new Index, Applied Economics, Vol. 38(10), pp 1091-1110, 2006.

6. Duolao, Wang, Estimating optimal transformations for multiple regression using the ACE Algorithm, Journal of Data Science 2, pp. 329-346, 2004.

7. Harvey, C. B. Erb, and T. E. Viskanta, Political risk, Economic Risk and Financial Risk, 1996. http://www. Duke.edu/charvey/country_risk/pol/pol.htm.

8. Gwartney, James and Robert Lawson with Herbert Grubel, Jakob de Haan, Jan-Egbert Sturm, and Eelco Zandberg (2009). Economic Freedom of the World: 2009 Annual Report. Vancouver, BC: The Fraser Institute. Data retrieved from www.freetheworld.com.

9. Guoping Xue, Optimal Transformations for multiple regression: Application to permeability estimation from well logs; SPE Formation Evaluation, pp 85-93, 1997.

10. Guoping Xue, Datta-Gupta, A., Valkó, P. and Blasingame, T. A. (1997), Optimal Transformations for Multiple Regression, Application to Permeability Estimation from Well Logs, SPE Formation Evaluation, Vol. 12(2), pp 85-93, 1997.

11. Osman, E. A., Abductive Networks: A new Modeling Tool for the Oil and Gas Industry, (SPE 77882), pp. 1-7, 2002.

12. Population 2008, World Development Indicators database, World Bank, 19 April, 2010

13. Sum, C.C, Yang-Sang Lee, Julie M Hays, and Arthur V Hill, Modeling the effects of a service guarantee on perceived service quality using Alternating Conditional Expectations (ACE), 1995.

14. Sum, C. C., Yang, K. K., Ang, J. S. K., & Quek S. A., An analysis of materials requirements planning (MRP) benefits using alternating conditional expectations (ACE). Journal of Operations Management, Vol. 13, pp. 35-58, 1995.

15. Surya, Ranjan, Simplified Country entry risk assessment model for global petroleum investments, SPE paper no. 112932, 2008

16. Veugelers, R., (2001); " Locational Determinants and Ranking of Host Countries: An Empirical Assessment," KYKLOS, Vol. 44, Fasc.3, pp. 363-382, 2001.

17. Wong, J., Determinants of marketing adaptation/ globalization practices of Australian exporting firms, World Review of Science, Technology and sustainable Development, Vol. 1(1), pp. 81-92, 2004.

18 www.transparency.org/policy_research/surveys_ indices/cpi/accessed 19.11.2010

19. http://data.worldbank.org/data-catalog accessed on 19.11.2010

20. https://www.cia.gov/library/publications/the-world-factbook/geos/xx.html/accessed on 19.11.2010

21. UNIDO, Industrial Statistics Database, 2007.

22. WHO world health systems http://www.photius.com/rankings/healthranks.html/accessed on 20.11.2010

23. Physicians per 1000 population http://www.geographic.org/country_ranks/physicians_per_capita_country ranks_2009.html/accessed 20.11.2010

24. Total health expenditure as % of GDP http://www.photius.com/rankings/total_health_expenditure_as_pecent_of_gdp_2000_to_2005.html/accessed 20.11.2010

25 Literacy rate http://www.photius.com/rankings/population/literacy_total_2010_1.html/ accessed 21.11.2010

26. Drinking water availability (most recent) by country, http://www.nationmaster.com/graph/hea_dri_wat_ava-health-drinking-water-availability/accessed 20.11.2010

Positional Uncertainty Estimation of a Parametric Surface

CAI Jianhong[1], LI Deren[2], and ZHU Daolin[1]

[1]the College of Resources and Environmental Sciences, China Agricultural University, Beijing, China

[2]the National Laboratory for Information Engineering of Surveying, Mapping and Remote Sensing , Wuhan University, Wuhan , China

Abstract - *When the complex real world is modeled, modeling errors appear inevitably for abstraction and simplification, which should be an important part in uncertainty theory, while ignored in the GIS literature. In this paper, we take a bilinear Beizer surface for example and use error promulgation law to estimate positional uncertainty of parametric surface considering modeling errors into uncertainty models.*

Key words: parametric surface ；modeling error; error promulgation law; uncertainty visualization.

1 Introduction

Uncertainty plays an important role in many subjects, and an impressive number of scientific articles concerning to uncertainty also emerges in relation to GIS. With the rapid development of digital earth, digital city and 3D representation etc., 3D uncertainty research is also urgent. The real world is so complex that when modeling it with various mathematical models, modeling errors are unavoidable for abstraction and simplification, which is an important part in uncertainty theory. However, they are ignored by all uncertainty models developed for use in GIS (see, e.g., Dutton 1992, Caspary and Scheuring 1993, Shi ,1998, Shi, etc., 2000, Cai etc., 2004, Shi ，2008,Meidow J. etc.，2009). Alesheikh (1998) realizes the importance of modeling errors, but doesn’t build a successful uncertainty model.

In this paper we consider modeling errors into uncertainty models and take bilinear Beizer surface for example to estimate positional uncertainty of parametric surfaces using error promulgation law.

The remainder of this paper is organized as follows. In Section 2, we analyze modeling errors and build an uncertainty model of parametric surface, which fully accounts for random errors of data and modeling errors. A simplified example of the unified model will be discussed and positional uncertainty is visualized in Section 3. Finally, some conclusions will be summarized in Section 4.

2 An uncertainty model of parametric surface

To visualize 3D geological objects which represent complex entities in the real world, we may always use parametric surfaces, such as plane, conicoid, ruled surface, Bezier Curved Surface, Coons Curved Surface, B-spline surfaces, Non-Uniform rational B-spline surfaces etc.. Any surface fitting, smoothing and/or interpolating is to abstract and simplify the complexity of them with simple models. During the process, modeling errors burst, and they are always much larger comparing with measurement errors of sample data, and cannot be ignored.

When modeling uncertainty models, modeling errors should be considered, and we treat them random errors. Taking bilinear Bezier surface for example, we study the positional uncertainty of parametric surfaces.

Given three sample points P_1, P_2, P_3 and P_4 with the coordinates $(x_1, y_1, z_1), (x_2, y_2, z_2), (x_3, y_3, z_3)$ and (x_4, y_4, z_4), respectively. Taking modeling errors into account, equation of bilinear Bezier surface can be more properly written as:

$$\begin{cases} x_{u,w} = (1-w)(1-u)x_1 + w(1-u)x_2 \\ \qquad + u(1-w)x_3 + uwx_4 + \xi_x \\ y_{u,w} = (1-w)(1-u)y_1 + w(1-u)y_2 & 0 \le u \le 1 \\ \qquad + u(1-w)y_3 + uwy_4 + \xi_y & 0 \le w \le 1 \\ z_{u,w} = (1-w)(1-u)z_1 + w(1-u)z_2 \\ \qquad + u(1-w)z_3 + uwz_4 + \xi_z \end{cases} \quad (1)$$

where ξ_x, ξ_y and ξ_z are the modeling errors of the x, y and z components, respectively. Using error theory, we can readily represent the errors of $(x_{u,w}, y_{u,w}, z_{u,w})$ in terms of the errors of $(x_1, y_1, z_1), (x_2, y_2, z_2)$, (x_3, y_3, z_3) and (x_4, y_4, z_4) and the modeling errors (ξ_x, ξ_y, ξ_z) as follows:

$$\begin{cases} \Delta x_{u,w} = (1-w)(1-u)\Delta x_1 + w(1-u)\Delta x_2 \\ \qquad + u(1-w)\Delta x_3 + uw\Delta x_4 + \xi_x \\ \Delta y_{u,w} = (1-w)(1-u)\Delta y_1 + w(1-u)\Delta y_2 & 0 \le u \le 1 \\ \qquad + u(1-w)\Delta y_3 + uw\Delta y_4 + \xi_y & 0 \le w \le 1 \\ \Delta z_{u,w} = (1-w)(1-u)\Delta z_1 + w(1-u)\Delta z_2 \\ \qquad + u(1-w)\Delta z_3 + uw\Delta z_4 + \xi_z \end{cases} \quad (2)$$

Equation (2) can be rewritten in matrix form as

$$X_{u,w} = AX \qquad (3)$$

where $X_{u,w} = \begin{bmatrix} \Delta x_{u,w} & \Delta y_{u,w} & \Delta z_{u,w} \end{bmatrix}^T$,

$$X = \begin{bmatrix} \Delta x_1 & \Delta y_1 & \Delta z_1 & \Delta x_2 & \Delta y_2 & \Delta z_2 & \Delta x_3 & \Delta y_3 & \Delta z_3 & \Delta x_4 & \Delta y_4 & \Delta z_4 & \xi_x & \xi_y & \xi_z \end{bmatrix}^T$$

$$A = \begin{bmatrix} (1-w)(1-u) & 0 & 0 & w(1-u) & 0 & 0 & u(1-w) & 0 & 0 & uw & 0 & 0 & 1 & 0 & 0 \\ 0 & (1-w)(1-u) & 0 & 0 & w(1-u) & 0 & 0 & u(1-w) & 0 & 0 & uw & 0 & 0 & 1 & 0 \\ 0 & 0 & (1-w)(1-u) & 0 & 0 & w(1-u) & 0 & 0 & u(1-w) & 0 & 0 & uw & 0 & 0 & 1 \end{bmatrix}$$

Applying the error propagation law to (3), we obtain the uncertainty of any arbitrary point on the bilinear Bezier surface, which is denoted by the variance-covariance matrix $D_{X_{u,w}X_{u,w}}$ and simply given below

$$D_{X_{u,w}X_{u,w}} = AD_{XX}A^T = \begin{bmatrix} \sigma_{x_{u,w}}^2 & \sigma_{x_{u,w}y_{u,w}} & \sigma_{x_{u,w}z_{u,w}} \\ \sigma_{y_{u,w}x_{u,w}} & \sigma_{y_{u,w}}^2 & \sigma_{y_{u,w}z_{u,w}} \\ \sigma_{z_{u,w}x_{u,w}} & \sigma_{z_{u,w}y_{u,w}} & \sigma_{z_{u,w}}^2 \end{bmatrix} \qquad (4)$$

where

$$D_{XX} = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1y_1} & \sigma_{x_1z_1} & \sigma_{x_1x_2} & \sigma_{x_1y_2} & \sigma_{x_1z_2} & \sigma_{x_1x_3} & \sigma_{x_1y_3} & \sigma_{x_1z_3} & \sigma_{x_1x_4} & \sigma_{x_1y_4} & \sigma_{x_1z_4} & \sigma_{x_1\xi_x} & \sigma_{x_1\xi_y} & \sigma_{x_1\xi_z} \\ & \sigma_{y_1}^2 & \sigma_{y_1z_1} & \sigma_{y_1x_2} & \sigma_{y_1y_2} & \sigma_{y_1z_2} & \sigma_{y_1x_3} & \sigma_{y_1y_3} & \sigma_{y_1z_3} & \sigma_{y_1x_4} & \sigma_{y_1y_4} & \sigma_{y_1z_4} & \sigma_{y_1\xi_x} & \sigma_{y_1\xi_y} & \sigma_{y_1\xi_z} \\ & & \sigma_{z_1}^2 & \sigma_{z_1x_2} & \sigma_{z_1y_2} & \sigma_{z_1z_2} & \sigma_{z_1x_3} & \sigma_{z_1y_3} & \sigma_{z_1z_3} & \sigma_{z_1x_4} & \sigma_{z_1y_4} & \sigma_{z_1z_4} & \sigma_{z_1\xi_x} & \sigma_{z_1\xi_y} & \sigma_{z_1\xi_z} \\ & & & \sigma_{x_2}^2 & \sigma_{x_2y_2} & \sigma_{x_2z_2} & \sigma_{x_2x_3} & \sigma_{x_2y_3} & \sigma_{x_2z_3} & \sigma_{x_2x_4} & \sigma_{x_2y_4} & \sigma_{x_2z_4} & \sigma_{x_2\xi_x} & \sigma_{x_2\xi_y} & \sigma_{x_2\xi_z} \\ & & & & \sigma_{y_2}^2 & \sigma_{y_2z_2} & \sigma_{y_2x_3} & \sigma_{y_2y_3} & \sigma_{y_2z_3} & \sigma_{y_2x_4} & \sigma_{y_2y_4} & \sigma_{y_2z_4} & \sigma_{y_2\xi_x} & \sigma_{y_2\xi_y} & \sigma_{y_2\xi_z} \\ & & & & & \sigma_{z_2}^2 & \sigma_{z_2x_3} & \sigma_{z_2y_3} & \sigma_{z_2z_3} & \sigma_{z_2x_4} & \sigma_{z_2y_4} & \sigma_{z_2z_4} & \sigma_{z_2\xi_x} & \sigma_{z_2\xi_y} & \sigma_{z_2\xi_z} \\ & & & & & & \sigma_{x_3}^2 & \sigma_{x_3y_3} & \sigma_{x_3z_3} & \sigma_{x_3x_4} & \sigma_{x_3y_4} & \sigma_{x_3z_4} & \sigma_{x_3\xi_x} & \sigma_{x_3\xi_y} & \sigma_{x_3\xi_z} \\ & & & & & & & \sigma_{y_3}^2 & \sigma_{y_3z_3} & \sigma_{y_3x_4} & \sigma_{y_3y_4} & \sigma_{y_3z_4} & \sigma_{y_3\xi_x} & \sigma_{y_3\xi_y} & \sigma_{y_3\xi_z} \\ & & & & & & & & \sigma_{z_3}^2 & \sigma_{z_3x_4} & \sigma_{z_3y_4} & \sigma_{z_3z_4} & \sigma_{z_3\xi_x} & \sigma_{z_3\xi_y} & \sigma_{z_3\xi_z} \\ & & & & & & & & & \sigma_{x_4}^2 & \sigma_{x_4y_4} & \sigma_{x_4z_4} & \sigma_{x_4\xi_x} & \sigma_{x_4\xi_y} & \sigma_{x_4\xi_z} \\ & & & & \text{symmetric} & & & & & & \sigma_{y_4}^2 & \sigma_{y_4z_4} & \sigma_{y_4\xi_x} & \sigma_{y_4\xi_y} & \sigma_{y_4\xi_z} \\ & & & & & & & & & & & \sigma_{z_4}^2 & \sigma_{z_4\xi_x} & \sigma_{z_4\xi_y} & \sigma_{z_4\xi_z} \\ & & & & & & & & & & & & \sigma_{\xi_x}^2 & \sigma_{\xi_x\xi_y} & \sigma_{\xi_x\xi_z} \\ & & & & & & & & & & & & & \sigma_{\xi_y}^2 & \sigma_{\xi_y\xi_z} \\ & & & & & & & & & & & & & & \sigma_{\xi_z}^2 \end{bmatrix}$$

From this point of view, our uncertainty model is more general, since even the correlations between $(x_1, y_1, z_1), (x_2, y_2, z_2), (x_3, y_3, z_3)$ and (x_4, y_4, z_4) are fully taken into account.

We have the variance components of $x_{u,w}, y_{u,w}$ and $z_{u,w}$, respectively, as

$$\begin{aligned}
\sigma^2_{x_{u,w}} &= (1-w)^2(1-u)^2\sigma^2_{x_1} + w^2(1-u)^2\sigma^2_{x_2} \\
&+ u^2(1-w)^2\sigma^2_{x_3} + u^2w^2\sigma^2_{x_4} + \sigma^2_{\xi_x} \\
&+ 2(1-w)(1-u)\sigma_{x_1\xi_x} + 2w(1-u)\sigma_{x_2\xi_x} \\
&+ 2u(1-w)\sigma_{x_3\xi_x} + 2uw\sigma_{x_4\xi_x} \\
&+ 2w(1-w)(1-u)^2\sigma_{x_1x_2} + 2u(1-w)^2(1-u)\sigma_{x_1x_3} \\
&+ 2uw(1-w)(1-u)\left(\sigma_{x_1x_4} + \sigma_{x_2x_3}\right) \\
&+ 2uw^2(1-u)\sigma_{x_2x_4} + 2u^2w(1-w)\sigma_{x_3x_4}
\end{aligned}$$

$$\begin{aligned}
\sigma^2_{y_{u,w}} &= (1-w)^2(1-u)^2\sigma^2_{y_1} + w^2(1-u)^2\sigma^2_{y_2} \\
&+ u^2(1-w)^2\sigma^2_{y_3} + u^2w^2\sigma^2_{y_4} + \sigma^2_{\xi_y} \\
&+ 2(1-w)(1-u)\sigma_{y_1\xi_y} + 2w(1-u)\sigma_{y_2\xi_y} \\
&+ 2u(1-w)\sigma_{y_3\xi_y} + 2uw\sigma_{y_4\xi_y} \\
&+ 2w(1-w)(1-u)^2\sigma_{y_1y_2} + 2u(1-w)^2(1-u)\sigma_{y_1y_3} \\
&+ 2uw(1-w)(1-u)\left(\sigma_{y_1y_4} + \sigma_{y_2y_3}\right) \\
&+ 2uw^2(1-u)\sigma_{y_2y_4} + 2u^2w(1-w)\sigma_{y_3y_4}
\end{aligned}$$

$$\begin{aligned}
\sigma^2_{z_{u,w}} &= (1-w)^2(1-u)^2\sigma^2_{z_1} + w^2(1-u)^2\sigma^2_{z_2} \\
&+ u^2(1-w)^2\sigma^2_{z_3} + u^2w^2\sigma^2_{z_4} + \sigma^2_{\xi_z} \\
&+ 2(1-w)(1-u)\sigma_{z_1\xi_z} + 2w(1-u)\sigma_{z_2\xi_z} \\
&+ 2u(1-w)\sigma_{z_3\xi_z} + 2uw\sigma_{z_4\xi_z} \\
&+ 2w(1-w)(1-u)^2\sigma_{z_1z_2} + 2u(1-w)^2(1-u)\sigma_{z_1z_3} \\
&+ 2uw(1-w)(1-u)\left(\sigma_{z_1z_4} + \sigma_{z_2z_3}\right) \\
&+ 2uw^2(1-u)\sigma_{z_2z_4} + 2u^2w(1-w)\sigma_{z_3z_4}
\end{aligned}$$

and the covariance between $x_{u,w}, y_{u,w}$ and $z_{u,w}$ as

$$\begin{aligned}
\sigma_{x_{u,w}y_{u,w}} &= (1-w)^2(1-u)^2\sigma_{x_1y_1} + w^2(1-u)^2\sigma_{x_2y_2} \\
&+ u^2(1-w)^2\sigma_{x_3y_3} + u^2w^2\sigma_{x_4y_4} + \sigma_{\xi_x\xi_y} \\
&+ w(1-w)(1-u)^2\left(\sigma_{y_1x_2} + \sigma_{x_1y_2}\right) \\
&+ u(1-w)^2(1-u)\left(\sigma_{y_1x_3} + \sigma_{x_1y_3}\right) \\
&+ uw(1-w)(1-u)\left(\sigma_{y_1x_4} + \sigma_{x_1y_4} + \sigma_{y_2x_3} + \sigma_{x_2y_3}\right) \\
&+ uw^2(1-u)\left(\sigma_{y_2x_4} + \sigma_{x_2y_4}\right) + u^2w(1-w)\left(\sigma_{y_3x_4} + \sigma_{x_3y_4}\right) \\
&+ (1-w)(1-u)\left(\sigma_{y_1\xi_x} + \sigma_{x_1\xi_y}\right) + w(1-u)\left(\sigma_{y_2\xi_x} + \sigma_{x_2\xi_y}\right) \\
&+ w(1-u)\left(\sigma_{y_3\xi_x} + \sigma_{x_3\xi_y}\right) + uw\left(\sigma_{y_4\xi_x} + \sigma_{x_4\xi_y}\right)
\end{aligned}$$

$$\begin{aligned}
\sigma_{x_{u,w}z_{u,w}} &= (1-w)^2(1-u)^2\sigma_{x_1z_1} + w^2(1-u)^2\sigma_{x_2z_2} \\
&+ u^2(1-w)^2\sigma_{x_3z_3} + u^2w^2\sigma_{x_4z_4} + \sigma_{\xi_x\xi_z} \\
&+ w(1-w)(1-u)^2\left(\sigma_{z_1x_2} + \sigma_{x_1z_2}\right) \\
&+ u(1-w)^2(1-u)\left(\sigma_{z_1x_3} + \sigma_{x_1z_3}\right) \\
&+ uw(1-w)(1-u)\left(\sigma_{z_1x_4} + \sigma_{x_1z_4} + \sigma_{z_2x_3} + \sigma_{x_2z_3}\right) \\
&+ uw^2(1-u)\left(\sigma_{z_2x_4} + \sigma_{x_2z_4}\right) + u^2w(1-w)\left(\sigma_{z_3x_4} + \sigma_{x_3z_4}\right) \\
&+ (1-w)(1-u)\left(\sigma_{z_1\xi_x} + \sigma_{x_1\xi_z}\right) + w(1-u)\left(\sigma_{z_2\xi_x} + \sigma_{x_2\xi_z}\right) \\
&+ w(1-u)\left(\sigma_{z_3\xi_x} + \sigma_{x_3\xi_z}\right) + uw\left(\sigma_{z_4\xi_x} + \sigma_{x_4\xi_z}\right)
\end{aligned}$$

$$\begin{aligned}
\sigma_{y_{u,w}z_{u,w}} &= (1-w)^2(1-u)^2\sigma_{z_1y_1} + w^2(1-u)^2\sigma_{z_2y_2} \\
&+ u^2(1-w)^2\sigma_{z_3y_3} + u^2w^2\sigma_{z_4y_4} + \sigma_{\xi_z\xi_y} \\
&+ w(1-w)(1-u)^2\left(\sigma_{y_1z_2} + \sigma_{z_1y_2}\right) + u(1-w)^2(1-u)\left(\sigma_{y_1z_3} + \sigma_{z_1y_3}\right) \\
&+ uw(1-w)(1-u)\left(\sigma_{y_1z_4} + \sigma_{z_1y_4} + \sigma_{y_2z_3} + \sigma_{z_2y_3}\right) \\
&+ uw^2(1-u)\left(\sigma_{y_2z_4} + \sigma_{z_2y_4}\right) + u^2w(1-w)\left(\sigma_{y_3z_4} + \sigma_{z_3y_4}\right) \\
&+ (1-w)(1-u)\left(\sigma_{y_1\xi_z} + \sigma_{z_1\xi_y}\right) + w(1-u)\left(\sigma_{y_2\xi_z} + \sigma_{z_2\xi_y}\right) \\
&+ w(1-u)\left(\sigma_{y_3\xi_z} + \sigma_{z_3\xi_y}\right) + uw\left(\sigma_{y_4\xi_z} + \sigma_{z_4\xi_y}\right)
\end{aligned}$$

We can also compute the positional uncertainty for surface features by using the variance-covariance matrix, which is simply to put the terms of variance together and is given by

$$\begin{aligned}
\sigma^2_{u,w} &= \sigma^2_{x_{u,w}} + \sigma^2_{y_{u,w}} + \sigma^2_{z_{u,w}} \\
&= (1-w)^2(1-u)^2\left(\sigma^2_{x_1} + \sigma^2_{y_1} + \sigma^2_{z_1}\right) + w^2(1-u)^2\left(\sigma^2_{x_2} + \sigma^2_{y_2} + \sigma^2_{z_2}\right) \\
&+ u^2(1-w)^2\left(\sigma^2_{x_3} + \sigma^2_{y_3} + \sigma^2_{z_3}\right) + u^2w^2\left(\sigma^2_{x_4} + \sigma^2_{y_4} + \sigma^2_{z_4}\right) + \left(\sigma^2_{\xi_x} + \sigma^2_{\xi_y} + \sigma^2_{\xi_z}\right) \\
&+ 2(1-w)(1-u)\left(\sigma_{x_1\xi_x} + \sigma_{y_1\xi_y} + \sigma_{z_1\xi_z}\right) + 2w(1-u)\left(\sigma_{x_2\xi_x} + \sigma_{y_2\xi_y} + \sigma_{z_2\xi_z}\right) \\
&+ 2u(1-w)\left(\sigma_{x_3\xi_x} + \sigma_{y_3\xi_y} + \sigma_{z_3\xi_z}\right) + 2uw\left(\sigma_{x_4\xi_x} + \sigma_{y_4\xi_y} + \sigma_{z_4\xi_z}\right) \\
&+ 2w(1-w)(1-u)^2\left(\sigma_{x_1x_2} + \sigma_{y_1y_2} + \sigma_{z_1z_2}\right) + 2u(1-w)^2(1-u)\left(\sigma_{x_1x_3} + \sigma_{y_1y_3} + \sigma_{z_1z_3}\right) \\
&+ 2uw(1-w)(1-u)\left(\sigma_{x_1x_4} + \sigma_{x_2x_3} + \sigma_{y_1y_4} + \sigma_{y_2y_3} + \sigma_{z_1z_4} + \sigma_{z_2z_3}\right) \\
&+ 2uw^2(1-u)\left(\sigma_{x_2x_4} + \sigma_{y_2y_4} + \sigma_{z_2z_4}\right) + 2u^2w(1-w)\left(\sigma_{x_3x_4} + \sigma_{y_3y_4} + \sigma_{z_3z_4}\right)
\end{aligned} \tag{5}$$

If we set modeling errors

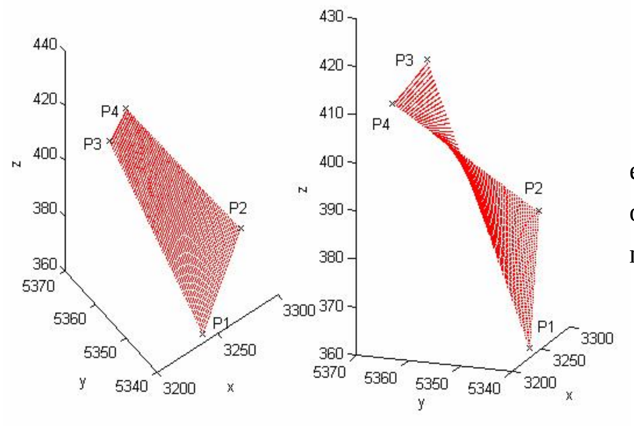
$\sigma_{x_1\xi_x}, \sigma_{y_1\xi_y}, \sigma_{z_1\xi_z}, \sigma_{x_2\xi_x}, \sigma_{y_2\xi_y}, \sigma_{z_2\xi_z}, \sigma_{x_3\xi_x}, \sigma_{y_3\xi_y}, \sigma_{z_3\xi_z}, \sigma_{x_4\xi_x},$

$\sigma_{y_4\xi_y}, \sigma_{z_4\xi_z}, \sigma^2_{\xi_x}, \sigma^2_{\xi_y}$ and $\sigma^2_{\xi_z}$ all to zero, then the positional uncertainty (5) becomes

$$
\begin{aligned}
\sigma^2_{u',w'} = &(1-w)^2(1-u)^2\left(\sigma^2_{x_1}+\sigma^2_{y_1}+\sigma^2_{z_1}\right)\\
&+w^2(1-u)^2\left(\sigma^2_{x_2}+\sigma^2_{y_2}+\sigma^2_{z_2}\right)\\
&+u^2(1-w)^2\left(\sigma^2_{x_3}+\sigma^2_{y_3}+\sigma^2_{z_3}\right)\\
&+u^2w^2\left(\sigma^2_{x_4}+\sigma^2_{y_4}+\sigma^2_{z_4}\right)\\
&+2w(1-w)(1-u)^2\left(\sigma_{x_1x_2}+\sigma_{y_1y_2}+\sigma_{z_1z_2}\right)\\
&+2u(1-w)^2(1-u)\left(\sigma_{x_1x_3}+\sigma_{y_1y_3}+\sigma_{z_1z_3}\right)\\
&+2uw(1-w)(1-u)\left(\sigma_{x_1x_4}+\sigma_{x_2x_3}+\sigma_{y_1y_4}+\sigma_{y_2y_3}+\sigma_{z_1z_4}+\sigma_{z_2z_3}\right)\\
&+2uw^2(1-u)\left(\sigma_{x_2x_4}+\sigma_{y_2y_4}+\sigma_{z_2z_4}\right)\\
&+2u^2w(1-w)\left(\sigma_{x_3x_4}+\sigma_{y_3y_4}+\sigma_{z_3z_4}\right)
\end{aligned}
\tag{6}
$$

3 Uncertainty visualization

Uncertainty visualization is an important part in GIS study of uncertainty. Since the uncertainty of a point can be readily visualized either by using an ellipsolid, a sphere or a cuboid, we will show how the uncertainty of a surface changes with the uncertainty of the several given points, the modeling errors.



(a) (b)

Figure 1 Bilinear Bezier Surface

Perspective angles are (69.50°, 8°) and (-42.5°,-28°) in Figure 1(a) and Figure 1(b)respectively.

Coordinates and variance-covariance matrix of P_1, P_2, P_3, P_4 are represented in Table 1, and a bilinear Bezier surface is produced, shown in Figure 1。

Here, modeling errors $\sigma^2_{\xi_x}, \sigma^2_{\xi_y}, \sigma^2_{\xi_z}$ are assumed as random numbers.

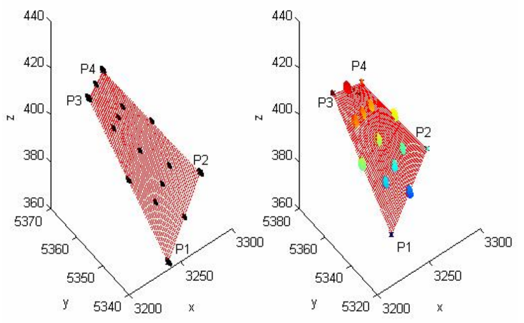
Tab 1. Known data

Coordinates of P_1, P_2, P_3, P_4 (m)

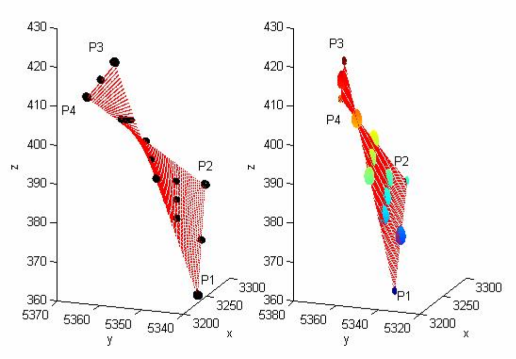
P1=(3240.77,5341.13,361.12);P2=(3278.55,5343.63,385.76);
P3=(3203.28,5356.83,422.55);P4=(3246.56,5368.32,408.23)

variance-covariance matrix of P_1, P_2, P_3, P_4 (m^2)
.43 .16 .12 .09 .10 .08 .10 .13 .11 .09 .10 .08 .11 .12 .10 .16 .35 .14 .11 .12 .09 .12 .14 .09 .05 .07 .10 .09 .11 .07 .12 .14 .39 .09 .16 .12 .09 .12 .11 .16 .18 .14 .06 .10 .09 .09 .11 .09 .32 .04 .09 .08 .10 .09 .12 .14 .11 .14 .12 .13 .10 .12 .16 .04 .34 .19 .08 .10 .09 .10 .06 .09 .16 .12 .10 .08 .09 .12 .09 .19 .39 .08 .13 .09 .12 .15 .07 .11 .07 .15 .10 .12 .09 .08 .08 .08 .42 .11 .07 .15 .14 .13 .10 .07 .18 .13 .14 .12 .10 .10 .13 .11 .40 .06 .12 .13 .12 .12 .10 .12 .11 .09 .11 .09 .09 .09 .07 .06 .39 .13 .07 .13 .11 .09 .07 .09 .05 .16 .12 .10 .12 .15 .12 .13 .37 .12 .11 .09 .06 .07 .10 .07 .18 .14 .06 .15 .14 .14 .07 .12 .42 .13 .09 .06 .14 .08 .10 .14 .11 .09 .07 .13 .12 .13 .11 .13 .37 .07 .12 .11 .11 .09 .06 .14 .16 .11 .10 .12 .11 .09 .09 .07 rand(1)*2.2 .12 .11 .12 .11 .10 .12 .12 .07 .07 .10 .09 .06 .06 .12 .12 rand(1)*2.3 .13 .10 .07 .09 .13 .10 .15 .18 .12 .07 .07 .14 .11 .11 .13 rand(1)*2.5

Let us assume that the coordinates $(x_1, y_1, z_1), (x_2, y_2, z_2), (x_3, y_3, z_3)$ and (x_4, y_4, z_4) are error-free. In other words, we will show the significance of uncertainty of the modeling errors in our uncertainty model (5).



(a) (b)



(c) (d)

Figure 2 Positional uncertainty of Arbitrary Points on the Bilinear Bezier Surface

The radius of error sphere are enlarged 100 times.

Modeling errors do not exist in Figure 2(a) and Figure 2(c). By contrast, Figure 2(b) and Figure 2(d) have combined effect of both types of errors, modeling errors and measurement errors of sample points.

It is clearly seen from Figure 2(b), Figure 2(d) and Table 2 that inner points on the surface is more uncertain. This should be realistic and readily understandable, because the surfaces of geographic entities are so complex that when simple models try to describe them, modeling errors come into being. Thus inner points should be more uncertain.

Table 2. Radius of Error Sphere of Arbitrary points on the Bilinear Bezier Surface

(u,w)	$\sigma_{u',w'}$ (cm)	$\sigma_{u,w}$ (cm)	(u,w)	$\sigma_{u',w'}$ (cm)	$\sigma_{u,w}$ (cm)	(u,w)	$\sigma_{u',w'}$ (cm)	$\sigma_{u,w}$ (cm)
(0,0)	1.08	1.08	(.75,.25)	0.83	2.17	(.25,.25)	0.81	2.18
(0,1)	1.02	1.02	(.75,.5)	0.77	2.15	(.25,.5)	0.75	2.17
(1,0)	1.10	1.10	(.75,.75)	0.80	2.16	(.25,.75)	0.78	2.18
(1,1)	1.04	1.04	(.5,0)	0.88	2.41	(1,.5)	0.88	2.39
(.5,.5)	0.73	2.14	(0,.5)	0.84	2.43	(.5,1)	0.83	2.13

In contrast, from Figure 2(a) , Figure 2(b) and Table 2 we may find out that inner points on the surface is precise. Actually, this is exactly what has been always discussed and seen in the literature of GIS uncertainty (see, e.g., Dutton 1992, Caspary and Scheuring 1993, Cai 2004, Shi 2008).

4 Conclusion

Modeling errors are of great consequence in uncertainty estimation, which denote degree of simplification of mathematical models.

In this paper, based on the assumption of taking modeling errors as a kind of randomized phenomenon we establish an uncertainty model to estimate parametric surface in GIS. The new uncertainty model automatically takes both modeling errors and measurement errors into account. And using visualization techniques we show the significance of uncertainty, especially modeling errors. It is easy to find that interpolating and fitting points are more uncertain, which is contrary to the available research results.

This work was supported in part by the Nature Science Foundation of China (41001372).

5 References

[1] Alesheikh A. A.. "Modeling and managing uncertainty in object-based geospatial information systems". PHD Thesis, the university of Calgary,1998

[2] Blakemore M.. "Generalization and error in spatial data bases[J].Cartographic", 1984, 21(2), pp.111-139.

[3] Cai Jianhong, Wen Hongyan, Chen Dake. "Error estimation of spatial curved surface of 3D GIS". Journal of Guilin Institute of technology, 2004, Vo1. 24 No. 2,pp.183-187.

[4] Caspary W. and Scheuring R.. "Error-band as measures of geographic accuracy", Proceedings of EGIS'92, 1992, pp. 226-233.

[5] Dutton G.. “Handling Positional Uncertainty in Spatial Databases”. In: Proceedings of the 5th International Symposium on Spatial Data Handling, 1992, vol. 2, 460–469.

[6] Meidow G., Beder C. and Förstner W.. “ Reasoning with uncertain points, straight lines, and straight line segments in 2D”, ISPRS Journal of Photogrammetry and Remote Sensing, 2009, 64(2), pp.125–139.

[7] Shi W. Z.. “A generic statistical approach for modelling error of geometric features in GIS”. International Journal of Geographical Information Science, 1998, 12 (2), pp.131–143.

[8] Shi W. Z.and Liu W. B.. “A stochastic process-based model for the positional error of line segments in GIS”. International Journal of Geographical Information Science, 2000, 14 (1), pp. 51–66.

[9] Shi W. Z.. “Modeling uncertainty in geographic information and analysi”s. Science in China Series E: Technological Sciences, 2008 (51), pp.38-47.

An enhanced Index Structure for a Digital Library Search Engine

M. Shahriar Hossain
Dept. of Computer Science, Virginia Tech,
Blasksburg, VA 24060, USA

Abstract – *The focus of this paper is to design an efficient indexing structure for spatial-temporal-textual data in a digital library environment. Along with traditional document collections, modern digital libraries contain forum discussions, bolg-posts, user provided URLs, and many other digital artifacts. All of these artifacts have time stamps, most of these artifacts have textual description, and some of them might have metadata regarding geographical location. Moreover, the textual descriptions sometimes provide idea regarding time and location. Management and retrieval of information from such multimodal data become very challenging due to the growing nature of digital libraries. Since spatial-temporal queries may not be adequately resolved by conventional text based search engines, we propose an enhanced mechanism to index digital artifacts using a spatial-temporal indexing mechanism called an R*B-tree. We discuss different strategies found in the literature and accept some of them for the proposed search engine.*

Keywords: Spatial-temporal-textual data, digital library, index structure, hybrid index, R*B-tree.

1 Introduction

The necessity of spatial and temporal access is increasing due to the diversity of digital artifacts in digital archives. Spatial-temporal access is traditionally used for location-aware moving objects (e.g., GPS) in the *D*-dimensional space [1]. In this paper we propose spatial-temporal access method for motionless digital artifacts each containing a timestamp on it that expresses the temporal scope of the object in a digital library. Examples of such motionless objects are web documents, forum discussions, blog-posts, tweets, etc. In the rest of the paper, we use either the term "document" or the term "page" to refer to any of these textual digital artifacts of a digital library. The aim of this paper is to propose an index structure for a digital library that would help a search engine for quick retrieval of digital artifacts. Each of the artifacts in the proposed design is indexed separately by a spatial-temporal index and by an inverted index. Along with efficient index structures combining space, time and text, we also need to efficiently process spatial-temporal-textual queries. We assume that each digital artifact in the library contains textual keywords, geographic locations, and time periods which are part of the corresponding text.

The rest of this paper is organized as follows. Section 2 gives an overall idea of the proposed system. Section 3 describes the indexing mechanism for spatial-temporal-textual data. Section 4 contains a discussion about some enhancements for personalization. We conclude this paper in section 5.

2 Overall Design

The overall structure of the indexing mechanism and the search engine proposed in this paper is motivated by the location-based search engine described by Zhou et al. [2]. The main difference between that search engine and the search engine proposed in this paper is that the proposed search engine is designed for spatial-temporal-textual information of digital libraries, where the previous system was designed for spatial-textual queries only. The framework of our proposed system is depicted in Fig. 1. The proposed system has four major components: (1) query processor, (2) spatial-temporal-textual index, and (3) extractor.

(1) Query processor: The query processor has two main modules, one is the parser and the other is the ranker. The parser is responsible for tokenizing the user query and retrieving useful keywords from it. It extracts three types of keywords from the query: textual, geographical (spatial) and temporal. The query processor takes the help of the gazetteer and timekeeper to convert the geographical and temporal terms to suitable format for searching the spatial-temporal index, in our case, minimum bounding rectangle (MBR) and timestamp respectively.

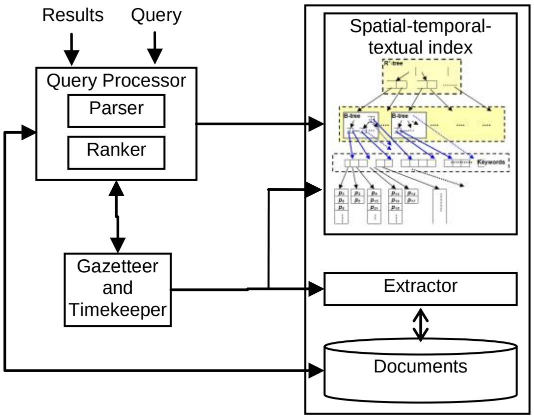


Fig. 1. Framework of the proposed search engine.

The ranker returns important documents relevant to the text keywords, geographic location and time related information provided by the user. There have been several schemes to combine two types of relevance scores, textual and geographical into a final ranking score. The most common scheme is the weighted sum of individual scores [5-7]. We incorporate temporal ranking in this:

$$R(q,d) = w_T \times R_T(q,d) + w_G \times R_G(q,d) + w_P \times R_P(q,d)$$

where d is a document and q is a query. R_T, R_G and R_P respectively refer to the functions to calculate the textual, geographic and temporal relevance. w_T, w_G and w_P are weights of these three individual relevance scores.

A very common way to find the textual relevance R_T between a query and a document is based on comparing textual keywords of the document and the query. Formally, given a query q, a ranking function R_T assigns a document a score. This score reflects the relevance of the document with the query. For a collection of documents D, the function $F(q, D)$ returns k documents with highest such scores. Most popular measures to find the relevance between a query and a document are Cosine measure [8], Jaccard measure, Okapi method [9], Pivoted Normalization method [9], etc.

Zhou et al. [2] describe some geographical ranking strategies supporting four spatial query types: contain, overlap, inside, and nearby:

i. Contain: $grank(Q_G, W_G) = W_G / Q_G$

ii. Inside: $grank(Q_G, W_G) = Q_G / W_G$

iii. Overlap:

$$grank(Q_G, W_G) = (Q_G \cap W_G) / (Q_G + W_G - Q_G \cap W_G)$$

iv. Nearby: nearby query is transformed to an overlap query.

where Q_G is the extent of the spatial query region and W_G is the extent of the scope of the web page.

Similar to the geographical ranking, we propose to use the same four spatial functions in the temporal context for temporal ranking:

i. Contain: $trank(Q_T, W_T) = W_T / Q_T$

ii. Inside: $trank(Q_T, W_T) = Q_T / W_T$

iii. Overlap:

$$trank(Q_T, W_T) = (Q_T \cap W_T) / (Q_T + W_T - Q_T \cap W_T)$$

iv. Nearby: nearby is transformed to an *overlap* query.

where Q_T is the extent of the temporal query and W_T is the temporal scope of the web page.

(2) Spatial-temporal-textual index: We use R*B-tree with inverted index as our spatial-temporal-textual index. MBRs of geographic locations are indexed by an R*-tree. Each of the leaf level MBRs is linked with a separate B-tree that indexes the timestamps. In this paper, we propose to use geographical scopes as MBRs of coordinates and *day* as the timestamp. Each timestamp is linked to a group of keywords of the inverted index. We describe our hybrid index structure in section 3.

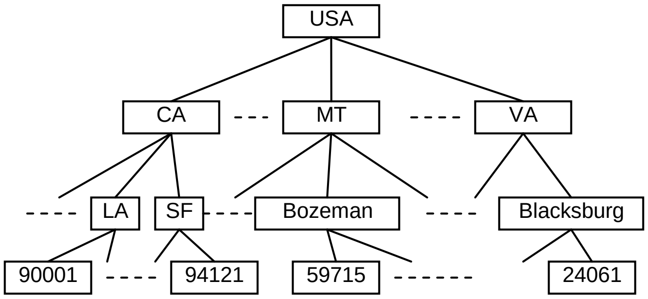


Fig. 2. A sample geographic ontology.

(3) Extractor: The extractor is responsible for retrieving the geographic and temporal scope of the web pages. We propose the use of a geographic ontology (e.g., Fig. 2) during the extraction of geographical scope of a document. We can use the gazetteer combined with the ontology for better extraction of geographic scope. Our extractor works in collaboration with the gazetteer and the timekeeper. The gazetteer is able to translate any geographic scope to MBRs of coordinates and the timekeeper is able to translate time to corresponding timestamps. The time ontology just like the geographic ontology would help the timekeeper to retrieve the timestamps easily. Moreover, this type of ontology can be very useful for query expansion both in geographic and temporal dimensions.

The query processor sends the geographic and temporal information found in the query to the gazetteer and timekeeper for translation. The query processor can traverse the spatial-temporal-textual index only when it gets translated geographic and temporal scopes from the gazetteer and the timekeeper.

Wang et al. [3] describe a method to extract geographic scope from a webpage. They define three types of geographic location of a web resource: provider location, content location and serving location. They provide an instance of these three types of locations in their paper taking MSN site as an example. On MSN, their algorithms found that provider location of the site is “Microsoft Corporation One Microsoft Way Redmond, Washington 98052, USA”. The serving location computed was “the Globe” because msn.com is a general web site with world-wide user reach. They found the content location of MSN’s New York local page [4] “New York, NY, USA”. Their paper defines content location as *the geographic location that the content of the web resource is about* and serving location as *the geographic scope that the web resource reaches*. The closest match of the geographic scope of our design is with the *serving location*. Since, we have the assumption that all our documents contain geographic locations as contents, it would be hard to find the actual geographic scope of a webpage if it contains multiple locations in its content. In that case we can also include geographical distribution of hyperlinks and user logs to analyze and extract the correct geographic scope of the web page. This is the strategy used in [3]. We propose a similar technique for our extractor.

We have a similar problem with the temporal scope. If multiple dates are provided in a document, the extractor needs to find a range of time period that illustrates temporal scope of

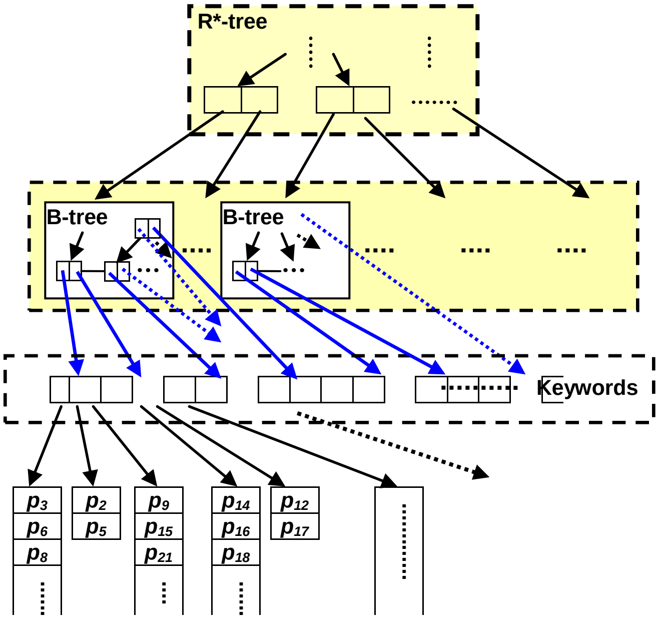


Fig. 3. The illustration of first R*B-tree and then inverted index.

the webpage. The time ontology can resolve this problem. We propose to use days as timestamps. The timestamps will be kept as integers. Suppose, we consider the first day of the year *2000* as *0*. Any day forward will be a positive integer and backward would be negative. The timekeeper is able to do this conversion between date and the corresponding integer form of timestamp. *Now, what would happen if a document talks about pre-mediaeval age?* This type of document will be rare. We can give these documents a particular timestamp, say *-10000*, which is around 27 years prior to year 2000 when digital artifacts were rare.

3 Index structure

A conventional approach to handle queries in multiple dimensions involves consecutive applications of several single key structures, one for each dimension. But, this type of independent traversal of every index becomes very inefficient for large datasets [10]. There is no easy and universal way to extend single key designs to handle multiple dimensions. The design of a multidimensional access method is highly dependent on the nature of the dimensions, in our case: spatial, temporal and text.

This paper proposes a solution to the problem described in the question by introducing a hybrid index structure for spatial-temporal-textual web search. The solution is motivated by the method described by Zhou et al. [2]. Zhou et al. focus on hybrid index structures for location-based web search. We incorporate an additional dimension, time, since our aim is to design an index structure for spatial-temporal-textual data.

In this section, we assume that the extractor already extracted the geographic and temporal scopes of the documents of the dataset. We also have documents and their corresponding keywords. Additionally, we have the inverted index, a structure that lists documents against the textual keywords. We aim to build a hybrid index structure to integrate text, location and temporal information of web pages. R-tree family, quad-tree and grid structures are generally used to index spatial information. We propose to use an R*-tree for the spatial indexing and B-tree for timestamp indexing, combined will be a R*B-tree. The R*B-tree will be used on top of the inverted index.

Fig. 3 illustrates the use of the R*B-tree as a spatial-temporal index with inverted file. An R*-tree is used to index all the MBRs found in the geographic scope of the web pages. Each lowest level MBR is linked to a B-tree. Each B-tree represents a geographic location and indexes the timestamps of the corresponding documents. Every node of the B-trees is linked to a group of keywords (or keyword-IDs) of the inverted index. The inverted index contains document-IDs against keywords. An example of the execution of a query, "Stores selling books on discount in Christiansburg today" is described as follows:

(Step 1) Traverse the R*-tree for MBRs covered by "Christiansburg". Each of the MBRs is linked with individual B-trees. Retrieve these B-trees.

(Step 2) Traverse each of the B-trees found in the previous step to get data with timestamps of "today", i.e., today's date. Each of the satisfied timestamps is linked with a group of keywords of the inverted index. These keywords are collected and stored. Final keywords are selected by the resultant intersection between these keywords and the textual keywords of the query.

(Step 3) Use the final keywords to retrieve the documents from the inverted file.

Execution Order: In the proposed spatial-temporal -textual index, any execution starts with the spatial-temporal index. It uses the R*B-tree as a spatial-temporal filter. The inverted index is accessed after this filtration.

Enhancements: One enhancement of this approach can be using B+ tree instead of B-tree to index the temporal data. In B+ tree, index is built with a single key per block of data records rather than with one key per data record. So, the B+ tree index is smaller than the B-tree. Therefore, the use of B+ tree can reduce the number of I/O operations required to find an element in the tree.

Another enhancement is by reducing the size of the MBR. [11] uses reduced size of MBR called toeprints instead of footprint MBRs and shows that toeprints work better. The paper also discusses enhancements using a *k-sweep* and a *tile index algorithm* for efficient query processing.

A lot of conventional search engines use parallelism where each machine or node of a cluster contains a subset of the documents and possesses its own inverted index [11]. This type of distributed indexing mechanism can provide faster query processing.

3.1 Cost Model

TABLE I describes the symbols used in this section. We name the triplet of keyword, MBR and timestamp a *geo-temporal-*

keyword. The main storage in the disk includes the page lists whose entry is a geo-temporal keyword and the R*B-tree itself. The storage cost is:

$$Storage = B_R + B_B + B_{List}$$
$$= O(M) + O(M \times S) + O(\sum_{g=1}^{G} P_G(g))$$

The experimental dataset used by [2] contains a total of *1,053,111* web pages and *26,090* MBRs. Let us consider that we have timestamps for all the days of last *10* years, i.e., a total of *3650* timestamps. In the worst case, for each of the MBRs of the R*-tree there will be *3,650* timestamps. Let us assume that each of the timestamps require *4* bytes to store an integer and another *4* bytes for any additional information, a total of *8* bytes. To store the R*B-tree, the system would require *26,090*×*3,650*×*8* bytes or around *727* MB. Conventional digital library servers contain several gigabytes of physical memory. So, the spatial-temporal part of the spatial-temporal-textual index can be placed in the main memory for faster retrieval even in a worst case scenario.

The online computation involves times to: (1) search R*-tree and get resultant MBRs, (2) for each of the MBRs, retrieve necessary B-trees satisfying the temporal part of the query, (3) from each of the retrieved B-trees get keywords and then corresponding page lists, (4) merge the page lists. The time complexity is:

$$Time = T_R(M) + M \times T_B(S)$$
$$+ \sum_{i=1}^{g(Q)} T_{disk} O(\frac{P_G(i)}{B_{section}}) + O(\sum_{i=1}^{g(Q)} P_G(i))$$

An enhancement with caching is used by conventional search engines. Frequently accessed inverted lists are cached in main memory [12]. Therefore, search for an old query will not involve the time to retrieve the inverted index.

Update Cost: Updating an entry in the inverted index using the proposed indexing mechanism involves the following three steps: (1) find the location in the R*-tree, (2) find corresponding timestamp from the corresponding B-tree, and (3) update the page lists associated with the timestamp. The first step should operate in $O(\log m)$ time where *m* indicates the number of MBRs. However, it can be $O(m)$ in the cases where the query point can appear in multiple MBRs. Later, we show that this part of the update cost can be ignored because it is relatively small compared to the other parameters. Let us consider that there are a total of *p* timestamps. In the worst case, the MBR selected in the first step points to a B-tree that contains all *p* timestamps. Therefore retrieving the timestamp from a B-tree takes $O(\log p)$ time. Updating the corresponding inverted index would take $O(n)$ time in the worst case where *n* is the total number of documents in the inverted index. After summing all of them we get the complexity for an update operation:

TABLE I
DESCRIPTION OF SYMBOLS

Symbol	Description
M	Total number of MBRs in the gazetteer
S	The number of timestamps in the dataset
G	The number of geo-temporal keywords in the lexicon.
K	The number of keywords in the lexicon
$g(Q)$	The number of geo-temporal keywords for a query
B_R	Storage of an R*-tree
B_B	Storage of All the B-trees followed by an R*-tree
$B_{section}$	The size of a section of the disk (depends on the file system)
B_{List}	Storage of page lists
$P_G(g)$	The length of the page list of a geo-temporal keyword g
$T_R(x)$	The time cost to retrieve an R*-tree of x elements
$T_B(x)$	Time cost to retrieve a B-tree of x elements
$\mathrm{T}_{\mathrm{disk}}$	The time cost of one disk access

$$U = O(\log m) + O(\log p) + O(n)$$

Since *m* would be much smaller than the number of timestamps and documents, we can ignore the $O(\log m)$ part. We can also ignore $O(\log p)$ as well since we proposed *day* as a timestamp. As an example, number of timestamps for the last 10 years would be $365 \times 10 = 3650$ which would be much smaller than the number of millions of documents in the inverted index. Therefore, the logarithmic parts of *m* and *p* will be surely very small compared to the linear function of *n*. Hence, the update cost using the proposed indexing will be $U \approx O(n)$.

3.2 Dynamic Ordering

In the proposed indexing mechanism, inverted index access can only start after the outputs from the R*B-tree has been generated. Let us assume another indexing mechanism where inverted index is placed first and then R*B-trees. That is, keywords are filtered first and then it traverses the R*B-trees accordingly. Now, we have two indexing mechanisms. The first one is, R*B-tree first and then inverted index, let us call this INDEX1. The second is, inverted index first and then R*B-trees, let us call this INDEX2.

If the geographic content of the query is large, then INDEX1 will perform less filtration. On the other hand, if textual terms of the query are very common then INDEX2 will perform less filtration at the top level because common terms are directed to a large number of R*B-trees. Therefore, the performance of the search engine will depend on such execution orders: inverted index first or R*B-tree first. We can propose a dynamic solution to this problem by providing a scoring mechanism that will decide which execution order is to perform by looking at the query terms. If the query contains too many common terms, the execution will go for INDEX1, because INDEX2 will be slow in this case. If the geographic scope of the query is too large, then the system will choose INDEX2 execution order, because INDEX1 will not provide sufficient filtration.

4 Enhancements

User profiles and interests can be incorporated in the proposed digital library indexing structure. Indexing or filtering techniques are not typically personalized to individual users or their prevailing context [13]. We can use a demographic filtering for personalized searching. We can use descriptions of people (such as salary range, age, and gender) to learn the relationship between a webpage and the type of people who like it [14]. For example, an educator is likely to browse for sophisticated course materials, whereas a teenage student might surf only to learn a particular topic. However, personalization depends on the user behaviors. Since user interest changes over time, socio-economic situation and many other factors, taking demographic information as indexing criteria would not be a good choice. Due to the dynamic nature of user behavior, it would be a good choice to keep the spatial-temporal-textual indexing intact and incorporate some demographic or usage factor to the ranking function during the search. For the same reasons, demographic filtering is rarely used independently of indexing or filtering techniques in the recommender systems [15].

We propose to enhance the ranking function based on age, gender and income level categories:

$$\begin{aligned} R(q,d) &= w_T \times R_T(q,d) + w_G \times R_G(q,d) + w_P \times R_P(q,d) \\ &\quad + w_A \times R_A(q,d) + w_S \times R_S(q,d) + w_I \times R_I(q,d) \end{aligned}$$

where R_X indicates the function that gives a relevance score between a query q and a document d, w_X is the corresponding weight for category X.

5 Conclusion

In this paper, we propose a spatial-temporal indexing mechanism for textual data. The focus of this paper is to study and logically analyze the possible indexing options of a specialized digital library search engine. The paper discusses different strategies that can be used by digital libraries to incorporate geo-spatial searching in text corpora. In the recent future, we would design an efficient query parser for efficient retrieval of data using the proposed indexing structure. Additionally, analyses of the proposed solution in a real digital library environment remain as a future task.

References

[1] M. F. Mokbel, T. M. Ghanem ,and W. G. Aref, "Spatio-temporal Access Methods", IEEE Data Engineering Bulletin, vol. 26, 2003, pp. 40-49.

[2] Y. Zhou, X. Xie, C. Wang, Y. Gong, and W. -Y. Ma, "Hybrid index structures for location-based web search", Proceedings of the 14th ACM international conference on Information and knowledge management, Germany, 2005, pp. 155-162.

[3] C. Wang, X. Xie, L. Wang, Y. Lu, and W. -Y. Ma, "Web resource geographic location classification and detection", 14th international conference on World Wide Web, Japan, 2005, pp. 1138-1139.

[4] MSN New York local page. http://local.msn.com/NewYork/

[5] B. Yu, G. Cai, "A query-aware document ranking method for geographic information retrieval", Proceedings of the 4th ACM workshop on Geographical information retrieval, Portugal, 2007, pp. 49-54.

[6] B. Martins, M. J. Silva, and L. Andrade, "Indexing and Ranking in Geo-IR Systems", Proceedings of the workshop on Geographic Information Retrieval, CIKM 05, Bremen, Germany, 2005, pp. 31-34.

[7] L. Andrade, and M. J. Silva, "Relevance Ranking for Geographic IR", Proceedings of the workshop on Geographic Information Retrieval, SIGIR 06, Seattle, USA, 2006.

[8] I. H. Witten, A. Moffat, and T. C. Bell. Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kaufmann, second edition, 1999.

[9] H. Fang, T. Tao, C. X. Zhai, "A formal study of information retrieval heuristics", Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 2004, pp. 49-56.

[10] V. Gaede, and O. Günther, "Multidimensional access methods", ACM Computing Surveys (CSUR), vol. 30, no. 2, 1998, pp. 170-231.

[11] Y. -Y. Chen, T. Suel, and A. Markowetz, "Efficient query processing in geographic web search engines", Proceedings of the 2006 ACM SIGMOD international conference on Management of data, USA, 2006, pp. 277-288.

[12] P. Saraiva, E. de Moura, N. Ziviani, W. Meira, R. Fonseca, and B. Ribeiro-Neto, "Rank-preserving two-level caching for scalable search engines", Proc. of the 24th Annual SIGIR Conf. on Research and Development in Information Retrieval, 2001, pp. 51–58.

[13] B. Sheth, and P. Maes, "Evolving agents for personalized information filtering", Proceedings of the 9th Conference on Artificial Intelligence for Applications, USA, pp. 345–352.

[14] B. Krulwich, "Lifestyle finder: Intelligent user profiling using large-scale demographic data", AI. vol. 18, no. 2, 1997, pp. 37–45.

[15] Y. Z. Wei, L. Moreau, and N. R. Jennings, "A market-based approach to recommender systems", ACM Transactions on Information Systems (TOIS), vol. 23, no. 3, 2005, pp. 227-266.

Comparison Analysis for Editorials by Reversible FACT-Graph

Ryosuke Saga[1], Seiko Takamizawa[2], Hiroshi Tsuji[3] and Kazunori Matsumoto[2]
[1]Faculty of Information and Computer Science, Kanagawa Institute of Technology, Atsugi, Kanagawa, Japan
[2]Graduate School of Engineering, Kanagawa Institute of Technology, Atsugi, Kanagawa, Japan
[3]Graduate School of Engineering, Osaka Prefecture University, Sakai, Osaka, Japan

Abstract - *This paper describes the method of comparison analysis between two targets using FACT-Graph which is trend-visualized graph. FACT-Graph shows the trend for time-series text data originally and it consists of class transition analysis and co-occurrence transition in an analysis period. To apply FACT-Graph to comparison analysis between targets, we attach pseudo time data to each target and regard class transition through analysis period as comparison between targets. Also, FACT-Graph shows the information from only one side period so that we carry out comparison analysis by using two FACT-Graphs like a reversible graph. In experiment for 122 editorials in two newspapers, we can find the differences from the reversible FACT-Graphs.*

Keywords: A Maximum of 6 Keywords

1 Introduction

Nowadays several business organizations have begun to focus on knowledge management to create business value and sustain competitive advantage by using data in data-warehouses [1][2]. They attempt to recognize their strong points, develop a strategy, and make effective investments from the data warehouses.

To recognize advantages, comparison analysis is often done by using cross-tabulation and visualization analysis. The aim of comparison analysis is basically to recognize the difference between two or more objects and the analysis is relatively easy when the comparative data are expressed quantitatively. However, most significant data often occur in text data and are difficult to obtain from pre-defined attributes. Therefore, text data in questionnaires, reports, and so on must be analyzed.

Text mining is useful for analyzing text data to obtain new knowledge [3]. In text mining, the applicable areas are wide-ranging such as visualization, keyword extraction, summarization of text, and so on. We have developed the Frequency and Co-occurrence Trend (FACT)-Graph for trend visualization of time-series text data [4]. FACT-Graph is used to visualize the trends in politics and crimes to extract important keywords that look unimportant at a glance.

This paper describes a method to compare two targets by using FACT-Graph. However, FACT-Graph targets time-series data, so we cannot apply it for comparison analysis. Therefore, we change data for analysis on the basis of class transition analysis to enable FACT-Graph to carry out comparison analysis.

The rest of this paper is organized as follows: Section 2 describes the overview and underlying technologies of FACT-Graph. Next, Section 3 describes how to apply FACT-Graph for comparison analysis. After that, Section 4 performs a case study of two Japanese newspapers. Finally, we conclude this paper.

2 FACT-Graph

FACT-Graph is a method and visualized graph for time-series text data from the viewpoint of large-scale trends. It shows the graph embedded co-occurrence graph and keyword class transition information. It allows us to show the hint of a trend, which is used for analyzing trends of politics and crime [5].

FACT-Graph uses nodes and links. It embeds the change in

Table 1. Transition of Keyword Classes;
Class A (TF: High, DF: High), Class B (TF: High, DF: Low),
Class C (TF: Low, DF: High), and Class D (TF: Low, DF: Low)

		After			
		Class A	**Class B**	**Class C**	**Class D**
Before	**Class A**	Hot	Cooling	Bipolar	Fade
	Class B	Common	Universal	-	Fade
	Class C	Broaden	-	Locally Active	Fade
	Class D	New	Widely New	Locally New	Negligible

a keyword's class transition and co-occurrence in nodes and edges. It has two essential technologies: class transition analysis and co-occurrence transition.

2.1 Class Transition Analysis and Co-occurrence Transition

Class transition analysis shows the transition of keyword class between two periods [6]. This analysis separates keywords into four classes (Class A to D) on the basis of term frequency (TF) and document frequency (DF) [7]. The results of the analysis detail the transition of keywords between two time-periods (before and after) as shown in Table 1. For example, if a term belongs to Class A in a certain time period and moves into Class D in the next time period, then the trend regarding that term is referred to as "fadeout". FACT-Graph identifies these trends by the node's color. For example, red means fashionable, blue unfashionable, and white unchanged. In convenience, we call the fashionable patterns Pattern1, the unchanged patterns Pattern 2, and unfashionable patterns Pattern 3.

Additionally, a FACT-Graph visualizes relationships between keywords by using co-occurrence information to show and analyze the topics that consist of multiple terms. As a result, useful keywords can be obtained from their relationship with other keywords, even though that keyword seems to be unimportant at a glance, and the analyst can extract such keywords by using FACT-Graph. Moreover, from the results of the class-transition analysis, the analyst can comprehend trends in keywords and in topics (consisting of several keywords) by using FACT-Graph. Also, FACT-Graph pays attention to the transition of the co-occurrence relationship between the keywords. This transition is classified into the following types;

(a) Co-occurrence relation continues in both analytical periods.
(b) Co-occurrence relation occurs in later analytical period.
(c) Co-occurrence relation goes off in later analytical period.

The relationship in type (a) indicates that these words are very close together, so we can consider them to be essential elements of the topic. On the other hand, relationships in types (b) or (c) indicate temporary topical changing.

2.2 Output FACT-Graph

The overview of the steps for outputting FACT-Graph is shown in Figure 1. In order to output FACT-Graph, the analyzer configures analysis periods, thresholds of TF, DF and co-occurrence, and the number of keyword to show the graph as parameters. From the parameters, the steps for generating a FACT-Graph are passed through as follows:

1. Separate time-series text data according to the analysis periods: The user sets up the parameters such as analysis span, filter of documents/terms, thresholds used in the analysis, and so on. Then, the term database is divided into two databases, first half period and second half period, in accordance with the analysis span.

2. Extract keywords in each period by morphological analysis and TF-IDF algorithm [7]: It is necessary to make a morphological analysis for the text data to output FACT-Graph. A morpheme is the smallest unit that has meaning in a sentence. Text data is divided into morphemes. Each term's frequency is aggregated in the respective database, and keywords are extracted from terms under the established conditions.

3. Carry out class transition analysis and extract co-occurrence relations: These keywords go through procedures concerning the transition of keyword classes

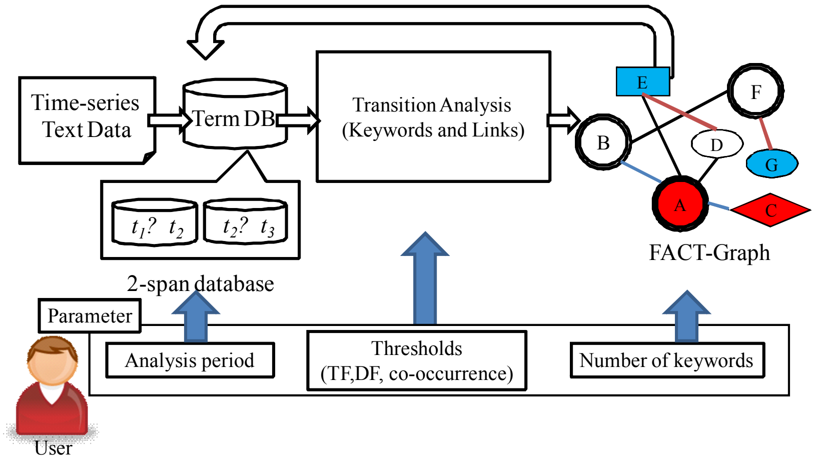


Figure. 1. Overview of Outputting FACT-Graph

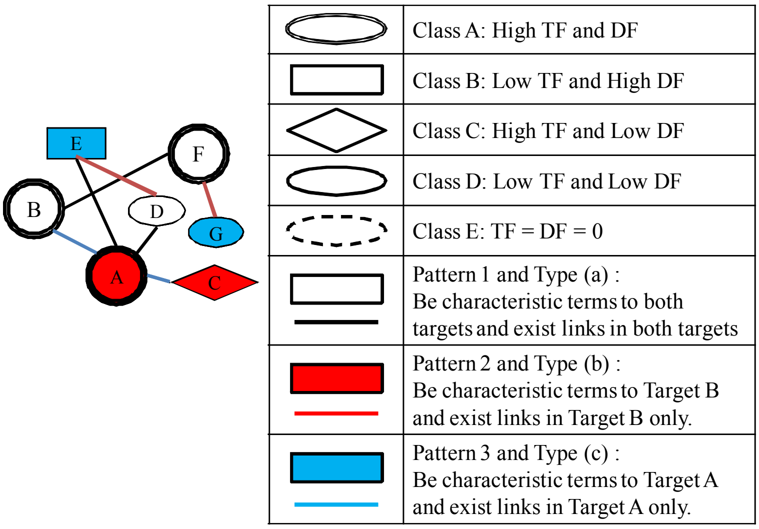

E
F
B
D
G
A
C
Class A: High TF and DF
Class B: Low TF and High DF
Class C: High TF and Low DF
Class D: Low TF and Low DF
Class E: TF = DF = 0
Pattern 1 and Type (a) : Be characteristic terms to both targets and exist links in both targets
Pattern 2 and Type (b) : Be characteristic terms to Target B and exist links in Target B only.
Pattern 3 and Type (c) : Be characteristic terms to Target A and exist links in Target A only.

Figure. 2. FACT-Graph for Comparison Analysis

and co-occurrence.

4. Visualize keywords and relations: The output chart that reflects the respective processing results is FACT-Graph.

3 Comparison Analysis by FACT-Graph

3.1 Approach

To apply FACT-Graph to compare information, we pay attention to class transition analysis in FACT-Graph. As we mentioned before, class transition analysis is carried out on the basis of two time periods, and FACT-Graph shows the changes between them. The other side of the coin is that FACT-Graph shows the results of the comparison between the periods, and the periods are simply regarded as the categories “Before” and “After”. In other words, we can comprehend that FACT-Graph performs a comparison analysis between two categories “Before” and “After” although it treats time-series text data. By replacing the periods with targets for comparison, we can compare them by using FACT-Graph.

However, applying FACT-Graph to comparison analysis has three problems: processing target data, explaining class transition analysis in comparison analysis, and how to express co-occurrence relationships. To apply FACT-Graph to comparison analysis, we need to convert date data. In FACT-Graph, the time data must be included in target data because the data are necessary and help to separate all target data into two periods. On the other hand, for comparison analysis the time data are not necessarily, and the time data do not exist in target data from the very first.

Therefore, we attach pseudo time data to target data as a category that belongs to either a period between t_1 and t_2 like the “Before” period or a period between t_2 and t_3 like the “After” period. Therefore, it is possible to perform comparison analysis by using FACT-Graph.

3.2 Explanation of Class Transition Analysis and Co-occurrence

The interpretation of comparison analysis by FACT-Graph is different from that of trend analysis, but the essential idea is same.

The concept of the comparison between two targets is the same as that of class transition analysis although the meanings of a FACT-Graph change, and we can compare the two targets in the same way we analyze FACT-Graph. For example, if two targets have the same keywords that belong to Class A (high TF and high DF), these targets have the equivalent features about topics that the keywords indicate. Let one target have Class A keywords and another Class B, Class C, Class D keywords. Then the former target is characteristic of the topic.

FACT-Graph has three types of co-occurrence. For comparison analysis, the co-occurrence means that one or more target uses the terms together. That is, the co-occurrence of type (a) means that a co-occurrence relationship exists in both targets. The other types mean that a co-occurrence relationship exists in alternative targets.

By the way, for trend analysis by using FACT-Graph, flux and reflux of the tides of terms are important, so we classify classes into four classes, Class A to D, by the height of TF and DF. However, for comparison analysis, knowing whether a term exists or not is necessary to find features of comparison targets in comparison analysis. Therefore, we add a new class, Class E, which expresses a term existing in only one side of comparison targets.

FACT-Graph for comparison analysis visualizes terms as shown in Figure. 2. FACT-Graph for trend analysis allocates four shapes according to Class A to D, and the color corresponds to the incremental or decremented trend based on class transition analysis. Also, the size of node is based on TF in after period due to the same reason, that is, the higher TF the keywords have, the larger nodes are. For comparison analysis, moreover, FACT-Graph expresses class E newly by a circular broken line in a visualized graph.

Here, in FACT-Graph for trend analysis, the allocated shapes to terms are based on after period class because we assume that the important information about trends occur in after period. However, for comparison analysis, it is difficult to understand the features of both targets. Because FACT-Graph mainly shows the features of one-side target (eg. “After”) which cover over another side (eg. “Before”). We should output both sides FACT-Graphs, that is, not only from “Before” to “After” but also from “After” to “Before” and we carry out comparison analysis. Therefore, in this following experiment, we analyze the trend by reversible FACT-Graphs.

4 Experiment

4.1 Environment and Data

By using FACT-Graph, we carried out an experiment to verify whether comparison analysis can be performed. In this study, we used editorials published in The *Mainichi* and The *Yomiuri* newspapers, two of Japan’s major newspapers, between 2006 and 2008.

Editorials are used because they pick up on important issues and are often written on the basis of interviews or opinions. Generally, these articles are written from several viewpoints, and the assertions are characteristic of and different for each publisher. Note that we regard few frequent words as unnecessary terms because there is a probability that

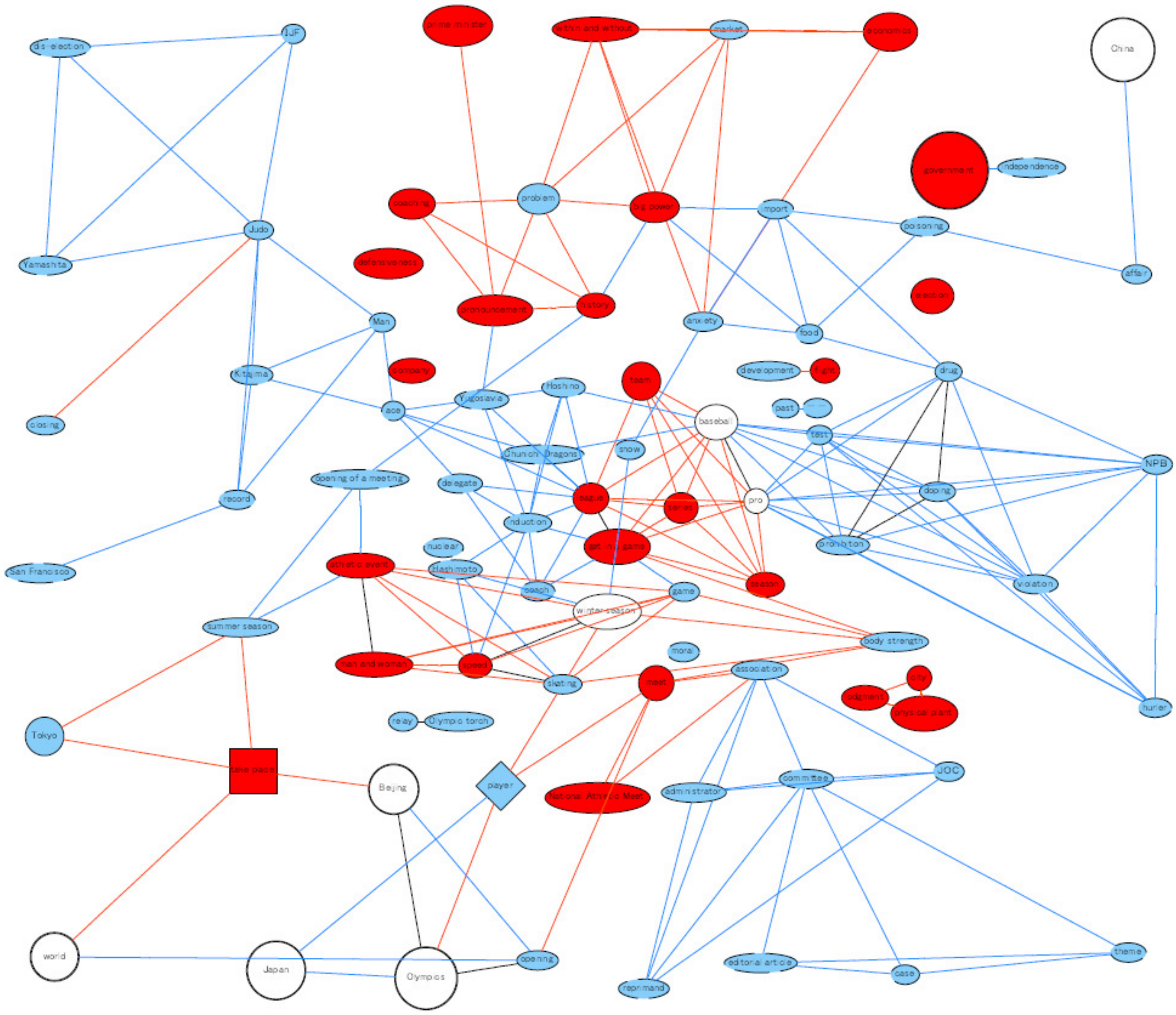


Figure. 3. Mainichi-Yomiuri FACT-Graph

they are noise and error words. Therefore, we removed the terms for which TF is less than 2 and DF is equal to 1.

In this case study, we limited 122 editorial articles (*Mainichi*: 64, *Yomiuri: 58)* to those on the topic of the Olympic Games. We apply Jaccard coefficient as co-occurrence and adopt the relationships whose co-occurrence is over 0.3. To carry out class transition analysis in FACT-Graph, we configure the threshold into the top 20% ranked termed.

4.2 Result of Analysis

Figure 3 shows the results of FACT-Graph when *Mainichi* is regarded as "Before" period and Yomiuri is regarded as "After" period (in convenience, this graph is called *Mainichi-Yomiuri* FACT-Graph). In this graph, blue nodes and links indicate the features in The *Mainichi* and red nodes and links The *Yomiuri*. Also, Figure 4 shows the results *Yomiuri-Mainichi* FACT-Graph.

When we take a global view of FACT-Graph, the term "Olympic", which is the most important word, is bigger than other nodes and belongs to Class A and Pattern 1 in these graphs. Also, there are "Beijing", "Japan", and "China", which have much the same pattern and class as "Olympic". Therefore, in this analysis period, the biggest topic in this graph concerns the Beijing Olympics.

Also, the nodes of Pattern 2 in Figure 4 connected by type (c) have existed in several parts. These nodes are a lot of words that are relevant to the games themselves, such as "Kitajima" (a Japanese gold medal winning swimmer), "Judo", and "Skating". Also, we can see the Class A nodes of Pattern 2 such as "Player" and "Game" in Figure 4. For these reason, we can say that The *Mainichi* describes the Olympic Games

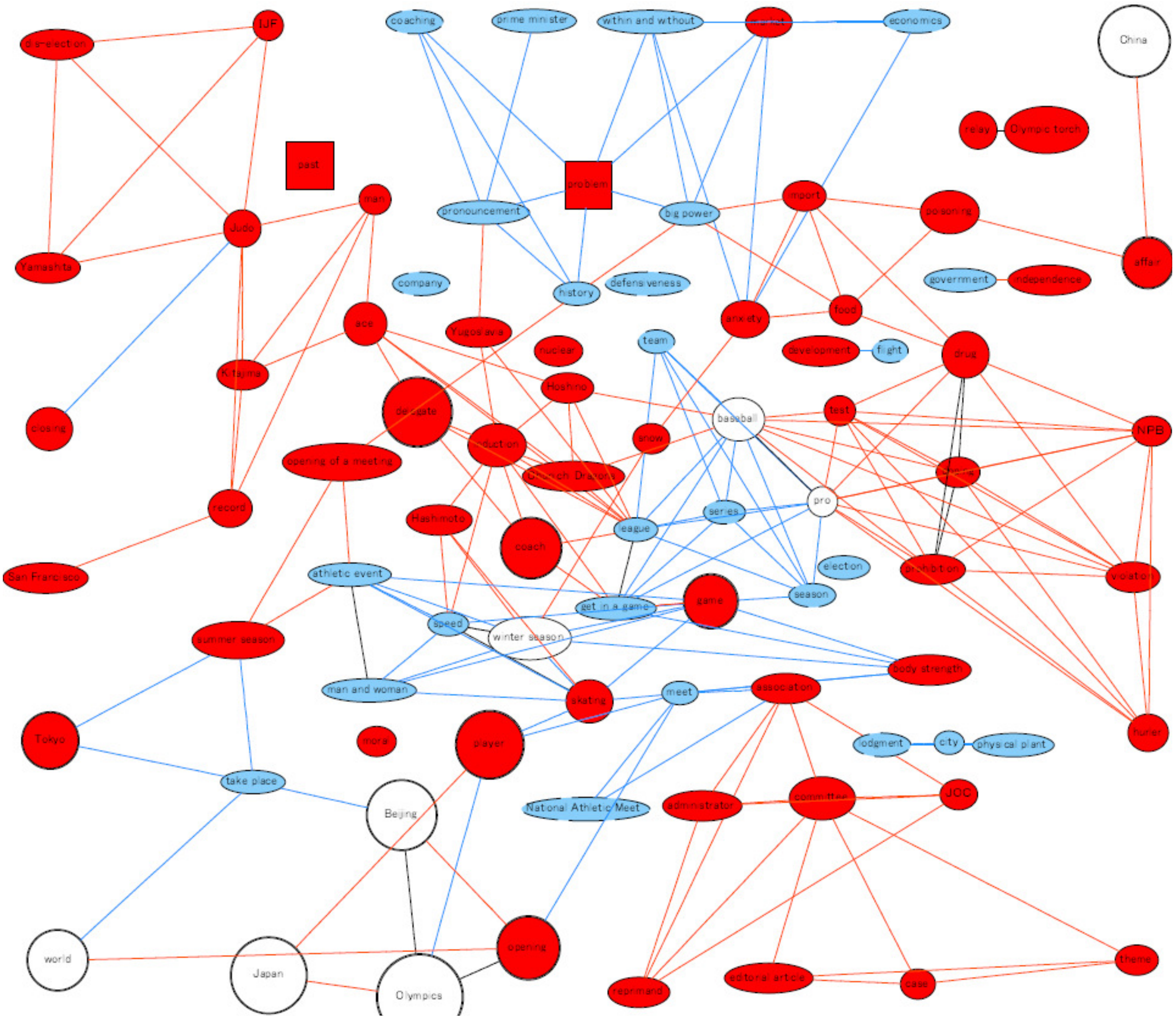


Figure. 4. Yomiuri-Mainichi FACT-Graph

without referencing anything else. In Figure 3, *Yomiuri* describes the topics for economics, pro baseball, and politics because the red nodes which indicate the features of *Yomiuri* are connected each other about them. However, these nodes belong to Class D and shows low level of attention. Therefore, the *Yomiuri* describes these topics involved in Olympics than the *Mainichi* but does not have a stronger tone.

5 Conclusions

This paper described a method to compare two targets by using both sides of FACT-Graph which can visualize the trends for time series text data. To apply FACT-Graph to comparison analysis, we interchanged target data with time series on the basis of class transition analysis. Also, we explained the two essential technologies (class transition analysis and co-occurrence transition) for comparison analysis and performed comparison analysis.

To validate the usability of FACT-Graph, we compared the features of The *Yomiuri* and The *Mainichi* newspapers by using editorials from two reversible FACT-Graphs. From the results of comparison analysis targeting the word “Olympic”, we found that The *Yomiuri* tended to write more political articles than The *Mainichi* and showed that the proposed method could be used for comparison analysis between two targets.

Acknowledgement

This research was supported by The Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Young Scientists (B), 21760305, 2009.4-2011.3

References

[1] A. Tiwana. “The Knowledge Management Toolkit: Orchestrating IT, Strategy, and Knowledge Platforms”. Prentice Hall, 2002

[2] W.H. Inmon. "Building the Data Warehouse". John Wiley & Sons, Inc., 2005

[3] R. Feldman, J. Sanger. "The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data". Cambridge University Press, 2007.

[4] R. Saga, M. Terachi, Z. Sheng, H. Tsuji. "FACT-Graph: Trend Visualization by Frequency and Co-occurrence"; Lecture Notes in Artificial Intelligence, Vol. 5243, pp. 308-315, Oct 2008.

[5] R. Saga, H. Tsuji, T. Miyamoto, K. Tabata. "Development and case study of trend analysis software based on FACT-Graph"; Artificial Life and Robotics, Vol. 15, No. 2, pp. 234-238, Oct 2010.

[6] M. Terachi, R. Saga, H. Tsuji. "Trends Recognition in Journal Papers by Text Mining"; Proceeding of IEEE International Conference on Systems, Man & Cybernetics (IEEE/SMC 2006), pp. 4784-4789, Oct 2006

[7] G. Salto. "Automatic text processing". Addison-Wesley Longman Publishing Co., Inc., 1988.

Time Series modeling of visitors' type on web analytics

Mohammad Amin Omidvar[1], **Vahid Reza Mirabi**[2], **and Narjes Shokry**[3]
[1]Information Technology Management, IAU E-Campus, Tehran, Iran
[2]Faculty of Management, IAU Central Branch, Tehran, Iran
[3]Faculty of Public Administration, IAU Central Branch, Tehran, Iran

Abstract - *The aim of this paper is to develop a flexible methodology to analyze the effectiveness of different variables on various dependent variables which all are times series specifically visitors' type on page views. This survey shows how to use a time series regression on one of the most important and primary index (page views per visit) on Google analytic and in conjunction it shows how to use the most suitable data to gain a more accurate result. There are too many data available on web analytics which are overwhelming for data analyzer. With this method, more accurate results are available. This methodology is critical for effective website monitoring and benchmarking that may lead to better website strategies. The value of this paper relies on introducing and using a systematic flexible methodology to analyze visitors' behavior and their impact on page views. Additionally this methodology can be used to analyze other time series variable.*

Keywords: worldwide web, Systems analysis, Data mining, visitors' behavior, web analysis, web metric, Google Analytics

1 Introduction

The internet is growingly rapidly and has a great impact on many businesses. Thousands of companies now own a website and websites have become an integrated part of the business. Furthermore many companies have employed many technologies which are available through the web such as online services. With web information, web developers and designers can improve user interfaces, search engines, navigation features, online help and information architecture and have happier visitors/ costumers [14]. One of the most popular ways which most frequented websites use to collect data and information about their websites is through web analytic. Web analytic collects a large amount of data from users such as browser type, connection speed, screen size, visitors' type, and etc. The collected data are usually large in quantity and type that need to be further processed to become useful information or knowledge.

1.1 Profile of the website

In 1998 an Iranian visual artist website was launched (http://www.omidvar.net). This website has many pages with images and few texts. The Google Analytics traffic overview showed that all traffic sources sent a total of 27,422 visits from 1 June 2008 to 31 March 2011. The total page views during this period were 145,874.

2 Impact of the internet on business

The internet has been playing the important role of corporate marketing during the past ten years [30]. With its combination of rich text, multimedia and user involvement, the internet contains more information than any other media [18], [25]. The internet offers speed, reach, and multimedia advantages, and has changed the way in which businesses interact with their customers, suppliers, competitors, and employees [8].

Nearly all businesses now have a website [12]. A corporate website enriches the image of a business and provides direct benefits in terms of electronic commerce (e-commerce) sales [22] and indirect benefits in terms of information retrieval, branding, and services [23]. Recognition of the internet is driving marketers in traditional companies to conduct transactions on the internet [10]. Barua, Konana, B.Whinston, & yin (2001) found that e-business operational excellence results in financial performance [5]. Thousands of companies have a fear to be left behind by their competitors if they do not use online technologies.

The total number of internet users and the number of websites are increasing significantly which will result in the rapid growth of the use of the World Wide Web for commercial purposes[16] [11] [9].

On a five year forecast by Forrester, e-commerce sales in the U.S. will grow 10% annually from 2014 and online retail sales will be nearly $250 billion, up from $155 billion in 2009 [27].

The increasing amount of internet users, websites and retail sales implies that web developing should be carried out in a competent, professional manner to increase profit.

However, systematic analysis of costumers' behaviors has not kept pace with the rapid growth in e-commerce. Without quantifiable metrics which is available through web analytics software, website optimization (WSO) is a guessing game, therefore a majority of e-commerce companies cannot afford this risk given their huge amount of money. Above 70% of the most frequented websites use web analytic tools but with their large amount of data, it is difficult to use them effectively [1]. Therefore, it is important to understand what kind of data and knowledge are required for successful website development work.

Web Analytics Association Standards (2006) committee defined the three most important metrics as Unique Visitors, Visits/Sessions, and Page Views; and, also categorized search engine marketing metrics through counts (visits…), ratios (page views per visits….), and key performance indicators (KPIs) [3].

The main reasons for measuring Search engine marketing (SEM) successes are related to traffic measurement and the return on investment come on 4th [28]. The top for reasons are as follows:

Increased traffic volume (76%)
Conversion rates (76%)
Click-through rates or CTRs (70%)
Return on investment (67%)

Progressive improvement of SEM campaigns, conversion rates, and website performance are available through web metrics, which would results in an increase in profits, happier customers, and higher return on investment (ROI) by tracking progress over time or against the competition [4].

Online technology collects large amounts of detailed data on visitor traffic and activities on websites, which would cause a plethora of metrics [13], and on the other hand this variety of measures can be overwhelming. Developing a website is a dynamic ongoing process which is guided by knowledge of its visitors.

3 Methodology

Google Analytics allows users to export report data in Microsoft Excel format, which when transformed can be analyzed with time series statistical programs. The software EViews is used to compute time series regression [7]. Initially a data set with 27,422 entries for 34 months drawn from Google Analytics was employed to analyze the performance of page views or page views per visits which is defined as one of the three most important metric [2]. Monthly data series was the most suitable series among daily and weekly because the accuracy and credibility of the regression was higher than those of other series [6] [26].

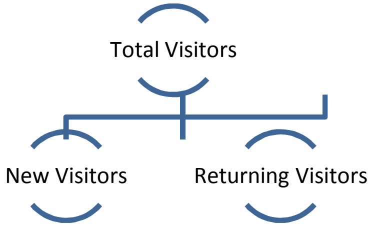


Figure 1 Independent Variables

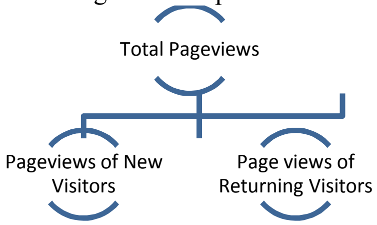


Figure 2 Dependent Variables

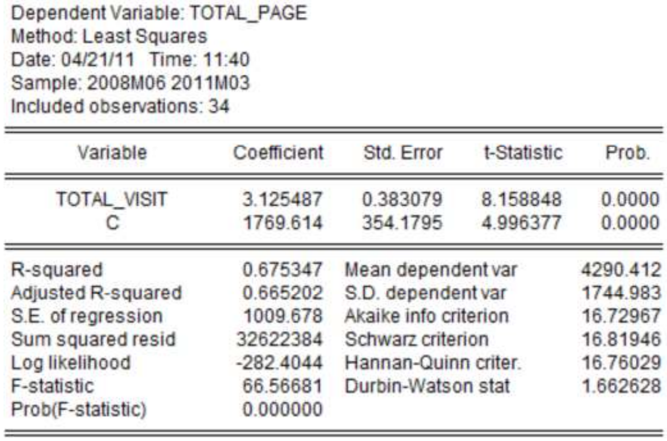
Before creating the model, some statistical matters with regards to the use of Google Analytics data in combination with time series methodology must be considered. For this reason all independent variables were processed by Augmented Dickey-Fuller Test to see if they were stationary or not. If the variables had a unit root they would be transformed to stationary by Difference-Stationary Process (DSP) [21].

Table 1 Augmented Dickey-Fuller test statistic (Level 10%)

Independent Variable	Has a Unit Root	Stationary After DSP
Total Visitors	No	
New Visitors	No	
Returning Visitors	No	

The first model is created with Autoregressive Moving Average (ARMA) model which consider total visitors as independent variable and total page views as dependent variable.

Table 2 Regression of Total Visitors



```
Dependent Variable: TOTAL_PAGE
Method: Least Squares
Date: 04/21/11   Time: 11:40
Sample: 2008M06 2011M03
Included observations: 34

      Variable        Coefficient   Std. Error   t-Statistic   Prob.

     TOTAL_VISIT       3.125487     0.383079     8.158848     0.0000
          C            1769.614     354.1795     4.996377     0.0000

R-squared              0.675347   Mean dependent var      4290.412
Adjusted R-squared     0.665202   S.D. dependent var      1744.983
S.E. of regression     1009.678   Akaike info criterion   16.72967
Sum squared resid      32622384   Schwarz criterion       16.81946
Log likelihood         -282.4044  Hannan-Quinn criter.    16.76029
F-statistic            66.56681   Durbin-Watson stat      1.662628
Prob(F-statistic)      0.000000
```

Table 2 represents the following model and the coefficient of total visits can be considered as the average impact of visitors (3.13).

PAGEVIEWS = 3.13*Total VISITS + 1769.61 (1)

The new model credibility and reliability is checked following these seven steps [19] [20] [15] [24].

1. Regression line must be fitted to data strongly (R^2>0.6).
2. Independent variables should be jointly significant to influence or explain the dependent variable (i.e. F-test, Anova)
3. Most of the independent variables should be individually significant to explain dependent variable (i.e. T-test).

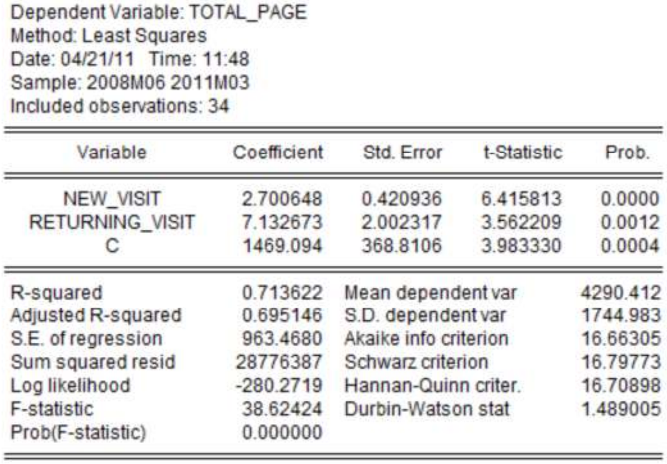
4. The sign of the coefficients should follow economic theory or expectation or experiences or intuition.
5. No serial or auto-correlation in the residual (Breusch-Godfrey serial correlation LM test : BG test)
6. The variance of the residual (u) should be constant meaning that homoscedasticity (Breusch-Pegan-Godfrey Test)
7. The residual (u) should be normally distributed (Jarque Bera statistics).

Table 3 credibility and reliability of total visitors' regression

Independent Variable	Total Visitors
Dependent variable	Total Page views
R^2>0.6	Yes
Independent variables are jointly significant	Yes
Independent Variables are individually significant	Yes
The sign of the coefficients follow economic theory	Yes
No serial in the residual	Yes
The variance of u is constant	No
Normal distribution of u	Yes

For further information about the impact of other visitors such as new visitors and returning visitors on page views, we would normally consider total page views as dependent variable and the two mentioned variables as independent variables.

Table 4 Ordinary Regression for total page views



```
Dependent Variable: TOTAL_PAGE
Method: Least Squares
Date: 04/21/11   Time: 11:48
Sample: 2008M06 2011M03
Included observations: 34

        Variable          Coefficient   Std. Error   t-Statistic   Prob.

       NEW_VISIT           2.700648     0.420936     6.415813     0.0000
    RETURNING_VISIT        7.132673     2.002317     3.562209     0.0012
           C               1469.094     368.8106     3.983330     0.0004

R-squared              0.713622    Mean dependent var     4290.412
Adjusted R-squared     0.695146    S.D. dependent var     1744.983
S.E. of regression     963.4680    Akaike info criterion  16.66305
Sum squared resid      28776387    Schwarz criterion      16.79773
Log likelihood        -280.2719    Hannan-Quinn criter.   16.70898
F-statistic            38.62424    Durbin-Watson stat     1.489005
Prob(F-statistic)      0.000000
```

Table 5 Credibility and reliability of ordinary total visitors' regression model

Independent Variable	Visitors' type (new visitors and returning visitors)
Dependent variable	Total Page views
R^2>0.6	Yes
Independent variables are jointly significant	Yes
Model	2.70*New visitors + 7.13*Returning Visitors + 1469.09

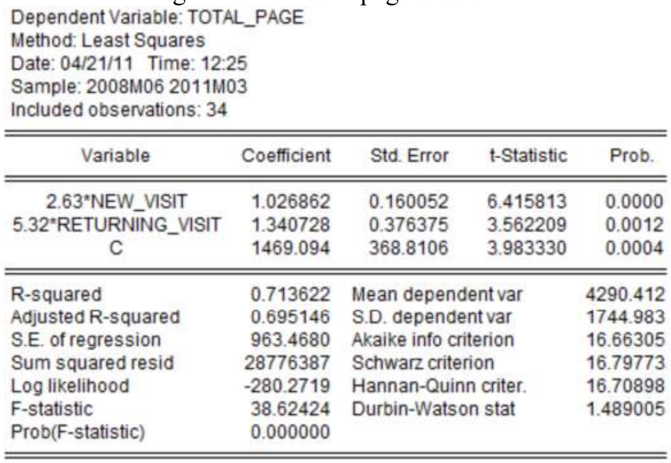
However, this method is not logical for web analytic because the detailed information of each visitor's page views is available [26]. Therefore total page view is broken down to its elements (page views of new visitors and returning visitors). Then, Autoregressive Moving Average (ARMA) model were created with their independent variables and their dependent variables. Since these models were describing part of total page views, only there fitness to data and significance were checked.

Ultimately, a new ARMA model of total page views was created with the sum of its related fundamental model. With the combined use of both dependent variables and independent variables, more accurate results were achieved.

Table 6 Regression model for each independent variable

Independent Variable	New visitors	Returning Visitors
Dependent variable	Page views of Search visitors	Page views of Direct visitors
R^2>=0.6	yes	Yes
Independent variables are jointly significant	Yes	Yes
Model	2.63* New visitors + 1458.77	5.32* Returning Visitors + 321.1

Table 7 new Regression for total page views



```
Dependent Variable: TOTAL_PAGE
Method: Least Squares
Date: 04/21/11   Time: 12:25
Sample: 2008M06 2011M03
Included observations: 34

        Variable            Coefficient   Std. Error   t-Statistic   Prob.

    2.63*NEW_VISIT           1.026862     0.160052     6.415813     0.0000
  5.32*RETURNING_VISIT       1.340728     0.376375     3.562209     0.0012
           C                 1469.094     368.8106     3.983330     0.0004

R-squared              0.713622    Mean dependent var     4290.412
Adjusted R-squared     0.695146    S.D. dependent var     1744.983
S.E. of regression     963.4680    Akaike info criterion  16.66305
Sum squared resid      28776387    Schwarz criterion      16.79773
Log likelihood        -280.2719    Hannan-Quinn criter.   16.70898
F-statistic            38.62424    Durbin-Watson stat     1.489005
Prob(F-statistic)      0.000000
```

The new model for total page views is as follow:

$$\text{Total page views} = 2.63*\text{New Visitors} + 5.32*\text{Returning Visitors} + 1469.1 \quad (2)$$

Table 8 Credibility and reliability of new total visitors' regression model

Independent Variable	Visitors' Type
Dependent variable	Total Page views[1]
R^2>=0.6	Yes
Independent variables are jointly significant	Yes
Independent Variables are individually significant	Yes
The sign of the coefficients follow economic theory	Yes
No serial in the residual	Yes
The variance of u is constant	Yes
Normal distribution of u	Yes

4 Results and Analysis

What can be manifest from close analysis of the information in the Table 4 and Table 7 is that there is significant difference on returning visitors' impacts and their method. Initially, the impact of returning visitors was 7.13 and on new model it decreased to 5.32. The latest model is more appreciated because it measures the correct impact of independent variables on dependent variables.

Many websites have measured their success with their competitors' behavior and many others have measured that with their own website success. This survey had introduced a methodology to measure the success of the website with its time series data. It also focused on one of the most primary and important variables which are page views and page views per visits, and showed how to use the most suitable data for measurment. This method can be used on all websites and time series variables.

5 Recommendation

One of the most difficult variables in webanalytic is returning visitors because it requires users to enable cookie on their browser, not delete the past cookie files, and also avoid changing browser. Some sites has used Ip address, username, and cookies all together to overcome this problem, but there are still problems in considering them as returning visitors. For Example, if a new visitor use some one else computer which has not logged out from the site, with all efforts that is done to categorise him correctly on returning visitors, ironicaly he will be labeled as returning visitors because he has used the same user name, IP address and cookie. However, it would make sense to consider direct visitors as returning visitors, since they have some simmilarities in behavior [26] .

This methodology can be used on more detailed variables, such as new visitos from refferal sites with T1 speed to get more detail information about users behaviour. Furthermore, since Google Analytics is integrated with Adwords and Adsense and their time series data's are available on Google Analytics, so further studies on time series data of Adsense Revenue or Adwords campaigns might have interesting results; because they bring traffic and revenue to the site.

Although this methodology is not recommended for search visitors' behavior and some semantic term is suggested [17], combining some other attributes such as visitors speed, terriority and type might help to perdict search visitors behaviour even without knowing the semantic terms of quarries.

6 Discussion

Many website owners or developers have used web traffic elements for their website performance, but the important thing is the knowledge gained about visitors and their behavior to keep them happy and satisfied with the website. Most of the websites have used web analytics to collect data about visitors with no systematic way to convert these data into tangible knowledge. The amount of these data which is available through web analytic is immense, and the developer may get lost in it. So a systematic way is needed to analyze these data.

7 References

[1] Google Analytics Usage Statistics. (2010, Jun). Retrieved from Web and Internet Technology Usage Statistics: http://trends.builtwith.com/analytics/Google-Analytics

[2] Association, W. A. (n.d.). "Big Three Definitions" Ver. 1.0. 2300 M Street, Suite 800, Washington DC 20037, United States.

[3] Association, W. A. (2006). Web Analytics "Big Three" Definitions., (p. 2). Washington DC 20037.

[4] B.king, A. (2008). Website Optimization. Orielly.

[5] Barua, A., Konana, P., B.Whinston, A., & yin, F. (2001.). Driving e-business excellence. Sloan Management, Rev. 34(1) 36–44.

[6] Batchelor, P. R. (n.d.). EVIEWS tutorial:Cointegration and error correction. City University Business School, London.

[7] Beatriz Plaza Faculty of Economics, U. o. (2009). Monitoring web traffic source effectiveness with Google Analytics An experiment with time series. Emerald, 9.

[8] Bodily, S., & Venkataraman, S. (2004). Not walls, windows: capturing value in the digital age. Journal of Business Strategy, Vol. 25 No. 3, pp.15-25.

[9] capital, I. u. (n.d.). ITU World Telecommunication/ICT indicators database.

[10] Chakraborty, G. L. (2002). An empirical investigation of antecedents of B2B Websites' effectiveness. Journal of Interactive Marketing, 16: 51–72. doi: 10.1002/dir.10044.

[1] Total page views= page views of New Visitors + page views of Returning Visitors

[11] Consotium, I. S. (2010). The ISC Domain Survey | Internet Systems Consortium. Retrieved from Internet Systems Consortium | Internet Systems Consortium: http://www.isc.org/solutions/survey

[12] Cotter, S. (1993). TAKING THE MEASURE OF E-MARKETING SUCCESS. Journal of Business Strategy, Vol. 23 Iss: 2, pp.30 - 37.

[13] FMI Group. (2001). Web site Visitors Analysis-Statistics or Intelligence? Basinstoke: FMI Group.

[14] Fourie, I., & Bothma, T. (2007). Information seeking: an overview of web tracking and the criteria for tracking software. Aslib Proceedings, ISSN: 0001-253X, 24.

[15] Garson, G. D. (2010, 2009, 2008). Logistic Regression. Retrieved July 2010, from http://faculty.chass.ncsu.edu: http://faculty.chass.ncsu.edu/garson/PA765/logistic.htm

[16] Global Number of Internet Users,total and per 100 inhabitants 2000-2010. (n.d.). Retrieved from ITU: Committed to connecting the world: http://www.itu.int/ITU-D/ict/statistics/material/excel/2010/Internet_users_00-10.xls

[17] Gupta, Siddharth; Thakur, Narina;. (2010). Semantic Query Optimisation with Ontology. International journal of Web & Semantic Technology (IJWesT).

[18] Hoffman, D. L., & Novak, T. P. (October 1996). Marketing in Hypermedia Computer-Mediated Environments. journal of marketing, 50-68.

[19] Hossain, S. (2006, 6 16). An Investigation into Regression Model using EVIEWS. Retrieved from http://www.sayedhossain.com/: www.sayedhossain.com/files/Lec1.Regression.ppt

[20] Hossain, S. (n.d.). An Investigation into Regression Model using EVIEWS. Lecturer for Economics.

[21] http://www.hkbu.edu.hk/~billhung/econ3600/application/app01/app01.html. (n.d.). Dickey-Fuller Unit Root Test. Retrieved from Hong Kong baptist university: http://www.hkbu.edu.hk/~billhung/econ3600/application/app01/app01.html

[22] Inge Geyskens, K. G. (2002). The Market Valuation of Internet Channel Additions. The Journal of Marketing, Vol. 66, No. 2 (Apr., 2002), pp. 102-119.

[23] Lederer , A., Mirchandani, D., & Sims, K. (2001). The Search for Strategic Advantage from the World Wide Web. International Journal of Electronic Commerce, Volume 5 , Issue 4 Pages: 117-133 .

[24] Ludlow, E. (n.d.). Multiple Regression: Fitting Models for Multiple Independent Variables.

[25] Okazaki, S., & Rivas, J. A. (2002). A content analysis of multinationals' Web communication strategies: cross-cultural research framework and pre-testing. Internet Research, Vol. 12 No. 5, pp.380-90.

[26] Omidvar, Mohammad Amin; Mirabi, Vahid Reza; Shokri, Narjes. (2011). Analyzing the impact of visitors on page views with Google Analytic. International Journal of Web & Semantic Technology (IJWesT).

[27] Schonfeld, E. (2010, March 8). Forrester Forecast: Online Retail Sales Will Grow To $250 Billion By 2014. Retrieved from TechCrunch: http://techcrunch.com/2010/03/08/forrester-forecast-online-retail-sales-will-grow-to-250-billion-by-2014/

[28] Search Engine Marketing Professional Organization, S. (2008, February). Search engine marketing 2007.

[29] SEMPO. (2006). Search Engine Marketing Professional Organization survey of SEM agencies and advertisers.

[30] Welling, R., & White Macquarie, L. (2006). Web site performance measurement: promise and reality. Managing Service Quality, SSN: 0960-4529.

SESSION

KNOWLEDGE DISCOVERY AND LEARNING

Chair(s)

Dr. Raymond A. Liuzzi
Peter M. LaMonica
Todd Waskiewicz

Entity Disambiguation for Multi-Source Knowledge Discovery

T. Darr, S. Ramachandran, P. Benjamin, T. Ramey and R. Mayer
Knowledge Based Systems, Inc., College Station, TX, USA

Abstract - *This paper describes the experimental results of applying a new approach for entity disambiguation in large Resource Description Framework (RDF) triple stores. Our approach combines data mining algorithms and the application of domain rules to reduce the ambiguity in large RDF graphs derived from real-world observational event data. Initial results have shown that significant reductions in ambiguity are achievable when combining data mining along multiple dimensions with domain rules that narrow that entity space that is necessary to disambiguate. The data mining algorithms incorporate relational, social network and process feature sets to compute proximity scores between pairs of entities. The application of domain rules is used to identify high-confidence distinct entities. Combining these methods will improve the quality, fidelity and utility of large-scale RDF graphs derived from real-world data. Our approach avoids the traditional manually intensive and time-consuming process of developing a rule base for inferring missing relationships and resolving ambiguous entities.*

Keywords: Knowledge discovery, data mining, entity disambiguation, RDF.

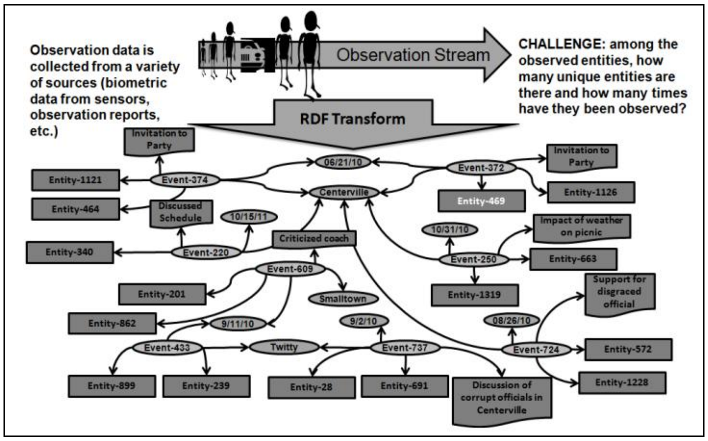
1 Introduction

Sensor networks are now capable of producing large quantities of high quality surveillance data. Data stores of surveillance data quickly become ungainly in size and create computational challenges for intelligence analysis. In the surveillance world the observer has little or no control over the events being observed, the data collected, or even the circumstances under which observations are made. Encoding such surveillance data using variants of the RDF formalism has become increasingly popular.

Encoding in RDF provides relatively high flexibility with relatively low expressive efficiency because of the minimalist representation model it employs: *subject-relation-object* triples. RDF is an effective means of encoding data from dynamic or poorly understood domains, where more efficient or effective expressions may not yet be available. Technology has been developed that supports substantial flexibility in "reasoning" over a data store encoded in RDF. The simplicity and flexibility of RDF are proving to be of significant value in intelligence analysis.

A significant challenge facing intelligence analyses is establishing the inter-relationships between real-world objects and the observations recorded in sensor data stores. Because the mapping from observation to coded entity is generally a localized process, the same object may be encoded with a different identity in different observation encodings. Ambiguity often occurs in the identification of the objects being represented.

A significant opportunity exists to increase the information content of such an observation-based RDF store through resolution of those object referents. Figure 1 is an illustration of this challenge. The figure shows a number of observation events, each of which involves one or more "entities." A wealth of information would be made available if we could determine which of the entities (e.g., Entity-572, Entity-1319, etc.) are actually on observation of the same individual. Persistent surveillance data would become *persistent surveillance information*.


Observation data is collected from a variety of sources (biometric data from sensors, observation reports, etc.)

CHALLENGE: among the observed entities, how many unique entities are there and how many times have they been observed?

Figure 1 – Context

The challenge, then, is this: Given an observation stream that consists of observed entities from a variety of sources and sensors (biometric, human reporting, etc.), the time at which the entity was observed, the location of the observation, and some information about the observation (the theme that appears in a message), how many unique identities are there among the observed entities and how many times has each individual been observed?

Figure 2 shows information fusion for a situation awareness and knowledge discovery application that is enabled by this capability. If we were able to correctly determine that Entity-1319, Entity-899, Entity-340 and Entity-691 from Figure 2 are the same individual ("Fred"), we could link up the events "discussed schedule," "discussion of corrupt officials," and "impact of weather on picnic"[1]. This would allow analysts and operatives to focus attention on the individual "Fred" as a suspected operational planner.

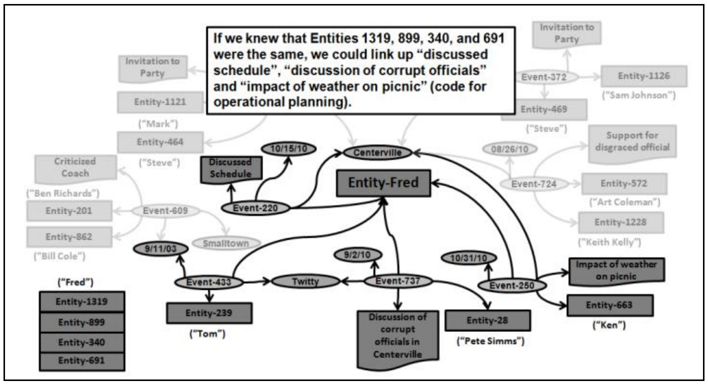


Figure 2 – Application

1.1 Benefits

The benefits of the approach presented in this paper include the following:

- The data mining and domain rule-based disambiguation increases the information content of the entity space (same data, more intelligence).
- Allows the decision maker/analyst to focus attention on entities of interest
  - More rapidly acquire situational awareness
  - More effectively manage limited intelligence resources
  - More effectively analyze existing data.
- Simplifies and reduces the manually intensive and time-consuming process of developing a domain rule base for inferring missing relationships and resolving ambiguous entities.

2 Data Mining Analytic Framework for Entity Disambiguation

Entity disambiguation is an essential part of situational awareness within a variety of modern day intelligence applications, especially where the intelligence data involves a vast variety of observations on objects/entities/artifacts that are distributed in time and space. Given the dynamic and evolving nature of situational awareness within this setting, knowledge discovery and machine learning methods [1, 2], coupled with expert knowledge offer the potential to quickly navigate through large accumulating volumes of spatio-temporal observations and aid in piecing together the partially known knowledge patterns and templates [3].

Entity disambiguation (or entity resolution) is a common problem that is studied, particularly in the field of natural language processing (NLP), to resolve references to the same entity using different language forms (such as references of "President Bush," "43rd," and "W" referring to the same underlying entity) [4]. The most widely used approach relies on feature-based matching [5], which involves first extracting various properties of entities from the natural language description that could potentially help in disambiguating, and then using a similarity-based approach to resolve or disambiguate the entities. The other approach is relational, which exploits the strength of relationships between entities to disambiguate. Very little research results have been reported that exploit multiple sources; such as sensor, natural language, relationship data, and user or domain knowledge.

We first describe the key analytical aspects and requirements for an operationally viable entity disambiguation system, followed by the analytical framework that was developed as part of this proof-of-concept for modeling the entity disambiguation problem from disparate types of spatio-temporal data. In real-world applications, facts and observations about entities are made or discovered along multiple attributes through multiple means, ranging from physical attributes (from sensor readings or human observations), human observed facts ("was seen fidgeting or loitering"), relational attributes ("was seen with entity B"), to process-based attributes ("the entity has been seen doing task-A followed by task-B"). There exist two different taxonomies for the information: one representing the types of attributes (for example, *physical* versus *relational* versus *process*) and a second taxonomy regarding the way that the information is captured (sensor, human reporting, induced, etc).

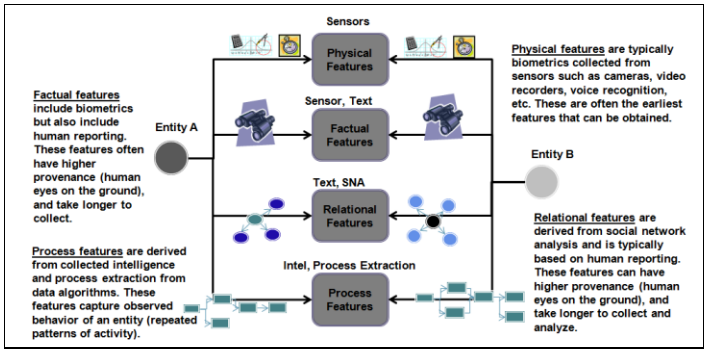

Sensors
Physical Features
Sensor, Text
Factual Features
Text, SNA
Relational Features
Intel, Process Extraction
Process Features
Entity A
Entity B
Factual features include biometrics but also include human reporting. These features often have higher provenance (human eyes on the ground), and take longer to collect.
Process features are derived from collected intelligence and process extraction from data algorithms. These features capture observed behavior of an entity (repeated patterns of activity).
Physical features are typically biometrics collected from sensors such as cameras, video recorders, voice recognition, etc. These are often the earliest features that can be obtained.
Relational features are derived from social network analysis and is typically based on human reporting. These features can have higher provenance (human eyes on the ground), and take longer to collect and analyze.

Figure 3 – Entity Disambiguation Framework

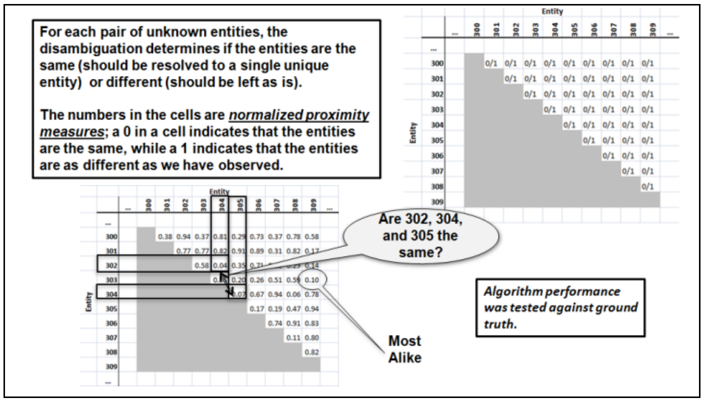
The goal of an entity disambiguation framework in such a setting, as depicted in Figure 3 becomes one of continually leveraging the full range of emerging information along these multiple taxonomies to match and disambiguate entities. From an operational setting, the entity disambiguation problem can be viewed as a dynamic task of continuously reasoning and matching observed entities within the field to a collection of known or unknown entities. Practically, the solution concept should recognize three important aspects – first, that the knowledge about an entity is evolving, starting with early observations more on the physical and factual dimensions, and as the entity is observed over time, additional attributes such as the social networks and process-based features for an entity is observed; second, that multiple knowledge types (or templates) have to be fused to achieve a tighter entity resolution, and third, that there is an underlying feedback process in that the knowledge gained also drives the

[1] This could be code for an operational plan.

seeking of knowledge; in other words, that the analytical process should also drive the intelligence requirements.

Given n entities in the dataset, the overall analytical problem involves evaluating $(n^2 – n)/2^2$ links for possible entity match. Figure 4 shows an example evaluation of a subset of the possible links. A matrix is created to assess each pair of links (since the entity proximity values are reflexive, only the upper diagonal needs to be evaluated). In reality, each cell can contain a 0 or 1 (shown in the upper right), which is a normalized measure of proximity or distance. If the contents of a cell are 0, then the corresponding entities are the same; if the contents of a cell are 1, then the corresponding entities are as different as possible for the observation.

The matrix in the lower left of the figure shows a matrix populated with data. In the row labeled with entity 303, the column labeled with entity 309 has the lowest proximity score (0.1), meaning that these two entities are the most alike (for the entities shown). Similarly, the entities labeled 303, 304 and 305 have similar proximity scores, meaning that they are possibly the same as well.



For each pair of unknown entities, the disambiguation determines if the entities are the same (should be resolved to a single unique entity) or different (should be left as is).

The numbers in the cells are *normalized proximity measures*; a 0 in a cell indicates that the entities are the same, while a 1 indicates that the entities are as different as we have observed.

Entity	...	300	301	302	303	304	305	306	307	308	309	...
...												
300			0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	
301				0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	
302					0/1	0/1	0/1	0/1	0/1	0/1	0/1	
303						0/1	0/1	0/1	0/1	0/1	0/1	
304							0/1	0/1	0/1	0/1	0/1	
305								0/1	0/1	0/1	0/1	
306									0/1	0/1	0/1	
307										0/1	0/1	
308											0/1	
309												

Entity	...	300	301	302	303	304	305	306	307	308	309	...
...												
300			0.38	0.94	0.37	0.81	0.29	0.73	0.37	0.78	0.58	
301				0.77	0.77	0.82	0.91	0.89	0.31	0.82	0.17	
302					0.58	0.04	0.35	0.71	[illegible]	[illegible]	0.14	
303						[illegible]	0.20	0.26	0.51	0.59	0.10	
304							0.07	0.67	0.94	0.06	0.78	
305								0.17	0.19	0.47	0.94	
306									0.74	0.91	0.83	
307										0.11	0.80	
308											0.82	
309												
...												

Algorithm performance was tested against ground truth.

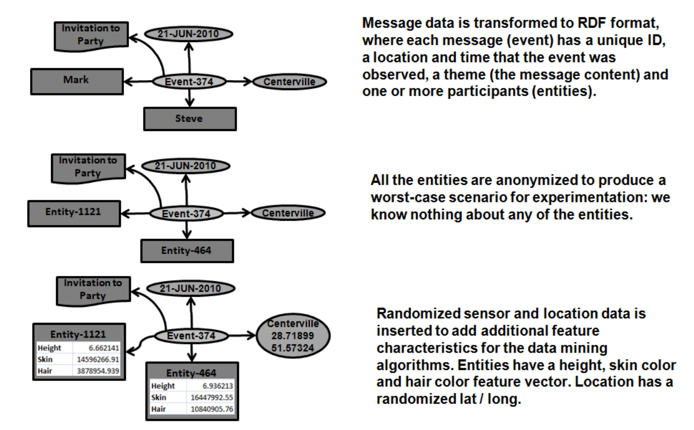
Figure 4 – Example Entity Match Calculation

3 Experimental Dataset

The dataset that was used for the demonstration consists of a set of synthetic messages regarding a scenario about a covert group planning an operation against a significant landmark. The dataset has the essential characteristics needed to support research into the entity disambiguation problem: observation of entities over a period of time, and includes ground truth so that experiments can be performed and accurately evaluated. An example message from the dataset is “Source 7 in location 1 claims to have known persons of interest John Doe and Fred Smith.” Each message has at least one entity (known or unknown) with some relationship, location or event that ties the entities together. In our proof-of-concept, the data will be anonymized (the named entities will be blacked out) and features to support the data mining will be extracted.

There are 670 reports with 1319 participating entities and 634 unique entities. In addition, 231 of the unique entities appeared multiple times and 403 entities appear only once (403 = 634 – 231). Of the set of 1319 participating entities, there are 916 which contain the unique, multiple occurring entities (916 = 1319 – 403). The average multiplicity of the multiple occurring entities is ≈ 4 (916 / 231).

Some modifications to the dataset were necessary to allow a better demonstration of the capability. This is shown in Figure 5. At the top of the figure is an RDF representation of the data without modification. Events have a unique identifier, a theme, a timestamp, a location where the event occurred and one or more named entities (some of the entities in the original dataset are anonymous: “Source 1,” “Source 2,” etc.). The first modification, shown in the middle of the figure, is to anonymize all of the entities. Even though the original dataset contained the names of some of the entities, we decided to start with a worst-case scenario where nothing was known about the identities of any of the entities. The second modification, shown at the bottom of the figure, is to add some sensor data for each of the entities, and a lat/long for each of the locations where the events were observed. We assumed the existence of biometric sensors located at or around the location that the event occurred. These sensors recorded the height, skin color, and hair color of the entity. Assuming ground truth for each of the entities, we added random noise taken from a uniform distribution to each of the sensor readings. In addition, we added a random noise taken from a uniform distribution to each location in the original dataset. These realistic modifications allowed us to highlight the performance of the data mining algorithms.



Entity-1121	
Height	6.662141
Skin	14596266.91
Hair	3878954.939

Entity-464	
Height	6.936213
Skin	16447992.55
Hair	10840905.76

Message data is transformed to RDF format, where each message (event) has a unique ID, a location and time that the event was observed, a theme (the message content) and one or more participants (entities).

All the entities are anonymized to produce a worst-case scenario for experimentation: we know nothing about any of the entities.

Randomized sensor and location data is inserted to add additional feature characteristics for the data mining algorithms. Entities have a height, skin color and hair color feature vector. Location has a randomized lat / long.

Figure 5 – Dataset Preparation

4 RDF Representation of Data Mining Results

Provenance statements are asserted about triples in the RDF graph based on the results of data mining. This allows us to incorporate probabilistic knowledge into the entity RDF graph. Figure 6 shows the end result of a rule that inserts a hypothesized association between two entities into the RDF graph, based on the results of data mining. Recall that the data mining algorithms compute proximity scores between pairs of

[2] To eliminate comparing x with y and y with x, and to eliminate comparing x to x itself. This count of cells in the lower triangle of n by n matrix, with the diagonal elements eliminated = (n*n – n)2.

entities. The end result is a “hypothesized association” provenance statement, with a normalized proximity value, that links the two entities.

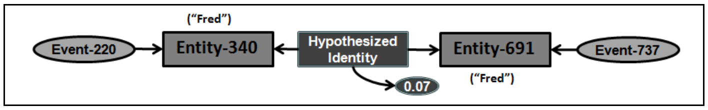
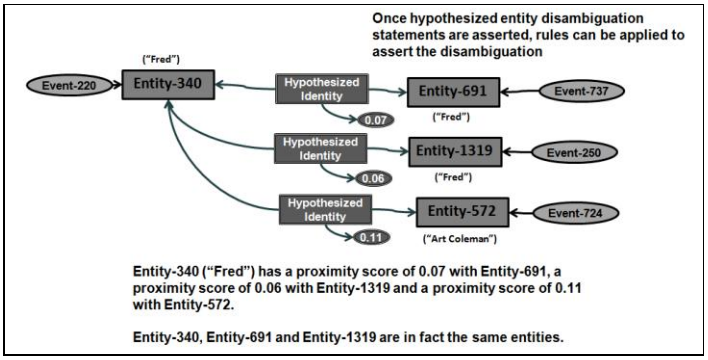


Figure 6 – Representing Hypothesized Entity Identity

Once the rules as illustrated in Figure 6 are applied, the RDF looks something like what is shown in Figure 7. The previously disconnected observations are now linked at the provenance level via statements that hypothesize possible associations, each of which has the respective proximity score associated with it. In this particular example taken from the demonstration dataset, we have three entities for the person “Fred” that have been linked together correctly, and one entity for a person that is not “Fred” (i.e., “Art Coleman”) that has been incorrectly linked. This is expected due to the probabilistic nature of the data mining and the uncertainty of the underlying data.



Once hypothesized entity disambiguation statements are asserted, rules can be applied to assert the disambiguation

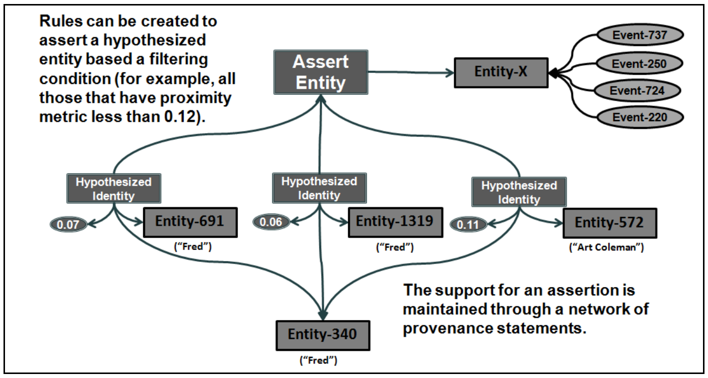
Entity-340 ("Fred") has a proximity score of 0.07 with Entity-691, a proximity score of 0.06 with Entity-1319 and a proximity score of 0.11 with Entity-572.

Entity-340, Entity-691 and Entity-1319 are in fact the same entities.

Figure 7 – Augmented Entity Disambiguation RDF Graph

Suppose that we have decided that 0.12 is “good enough” proximity for disambiguating entities. Our goal is to remove redundant entities and add links that more accurately reflect the ground truth. In this particular example, this means, based on the results of data mining, that Entity-340, Entity-691, Entity-1319 and Entity-572 should be removed and replaced with an entity, (called Entity-X for illustration purposes) that represents the “ground truth.” In addition, the links from Event-220, Event-737, Event-250 and Event-724 to Entity-340, Entity-691, Entity-1319 and Entity-572, respectively, should be removed and these same links added to the new entity. Figure 8 shows the disambiguated graph.

The new entity is linked to a provenance assertion statement (Assert Entity); in turn, this statement has links from the individual provenance statements that asserted the associations. In this way, a network of provenance statements provides support for each assertion. This makes it possible to retract information if future intelligence requires it. In this example, suppose that the identity of Entity-572 was determined to be “Art Coleman.” It is straightforward to write a rule that retracts the hypothesis that Entity-572 is the same as Entity-340. This retraction would propagate through the support network, updating links and other support along the way.



Rules can be created to assert a hypothesized entity based a filtering condition (for example, all those that have proximity metric less than 0.12).

The support for an assertion is maintained through a network of provenance statements.

Figure 8 – Disambiguated RDF Graph

5 Experimental Metrics

For each entity, multiple analytical processing pipelines produce a ranked list of entities which is ordered in terms of the respective proximity metric. The ranked lists provide the means to measure the performance of the disambiguation approach. In the example below, for the entity E_1, the closest entity based upon the feature-based proximity metric is E_{i1}, the next closest is E_{i2}, followed by E_{i3} and so on. If the match is perfect, then entity E_1 could be E_{i1}, and if E_1 happens to be $E_{\ i2}$, then the match has a misalignment of 1 (ranking distance = 1), and if E_1 happens to be E_{i3} then the match has a misalignment of 2 (ranking distance = 2), and so on.

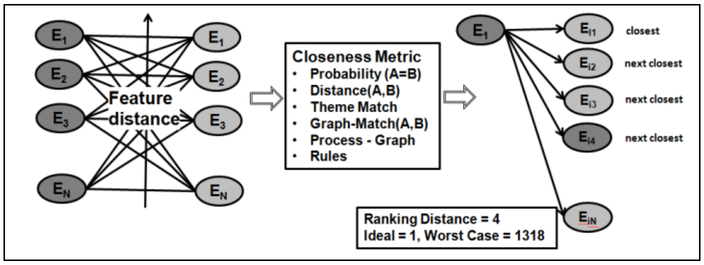


Figure 9 – Measurement Framework

A number of such metrics were defined to evaluate the performance of the disambiguation approach. These include:

- *Average Ranking Distance* (Figure 10) – this measures the average of the ranked distance over all the entities that appear more than once in the dataset. For the two entities below, average distance = (4+3)/2 = 3.5. The lower this number, the better the disambiguation performance.

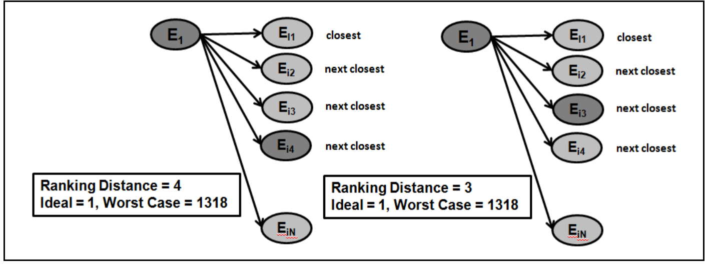


Figure 10 – Average Ranking Distance Metric

- *Matched Percentage N* (Figure 11) – This metric considers the closest N matches for each entity, and evaluates for each entity if a match was found in the closest N set, and computes the percentage for which a match was found in the closest N. The Figure below shows two cases, one for which a match was found at distance 4, and once for which a match was not found in the top N.

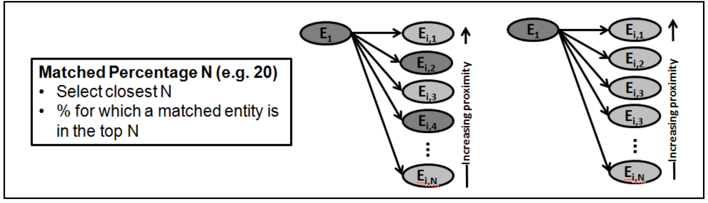


Figure 11 – Matched Percentage N Metric

- *Top-Matched Percentage* (Figure 12) – this measures the percentage of entities for which the top ranked entity was a match, in other words a match was found at a ranking distance of 1 (left-figure below, as opposed to the right figure)

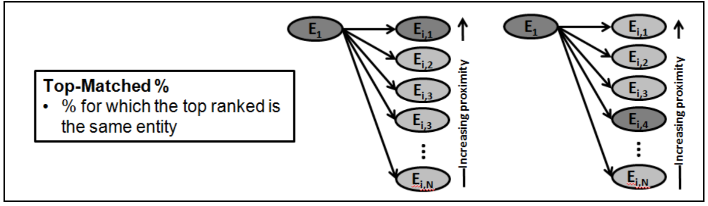


Figure 12 – Top-Matched Percentage Metric

6 Experimental Results

A series of analytical data processing pipelines were setup and evaluated – pipelines that extract features from the raw data, reduce and map the features to an analytical space, and then derive entity-entity proximity metrics to extract a ranked set of entity matches. The proximity metrics are distances between entities on a normalized and orthogonal space – the dimensions of the space depend on the features under consideration. The following four such pipelines were evaluated:

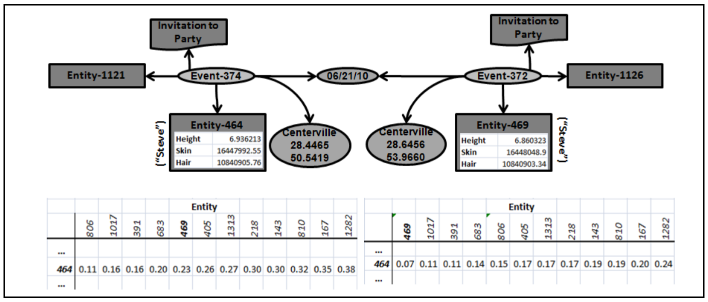
a) biometric feature based proximity metric – here, the biometric sensor features (color height, hair color) were processed (normalized using z-transform, and reduced using PCA [2]) and proximity metric was derived as a simple Euclidean distance between entities.
b) a theme based proximity metric – where features extracted from the text messages as parsed and part-of-speech-tagged tokens, and theme vectors were constructed as bag-of-words of verbs and participating nouns in the themes, and proximity metric defined as an Euclidean distance between the bag-of-word vectors [6].
c) A spatial proximity metric – based upon the distance between the locations (lat-long) – this measures how spatially close the entities operate.
d) A relational proximity metric – where entities are first represented in terms of their social networks, and defining the proximity metric that measures the similarity between the graphs (graph matching metrics).

While four data mining processing pipelines were evaluated, based upon data completeness and quality, two of these metrics (biometric feature based and theme based) were included in the detailed performance analysis. For both of these metrics, we defined the performance results using the metrics that were defined above.

The performance metrics for biometric features only are

- Top Matched Percentage = 27%
- Matched Percentage 20 = 64%

Next, the theme proximity metric was evaluated as a means to enhance the disambiguation process. While multiple schemes were considered (voting schemes), the best performance was achieved by applying a sequential sort – by sorting based upon feature proximity metric, and then resorting based upon theme proximity. In essence, this approach pushed entities that are similar in features and involved the same type of activity to be more proximal. The concept is illustrated in Figure 13. Consider the ranked list of potential matches for Entity-464 (Entity Truth Value = “Steve”), based upon the feature proximity metric, ordered in the most proximal to least proximal (shown at left). The closest match for Entity-464 occurs at a rank distance of 5: Entity-469 (Entity Truth Value = “Steve”). The knowledge that these two entities Entity-464 and Entity-469 were involved in the same activity of “recruiting to hear a corrupt official” can be used to re-order the initial sorted list of matched entities, pushing Entity-469 to the top, shown at the right of the figure.



Entity-464	
Height	6.936213
Skin	16447992.55
Hair	10840905.76

Entity-469	
Height	6.860323
Skin	16448048.9
Hair	10840903.34

	Entity											
	806	1017	391	683	469	405	1313	218	143	810	167	1282
...												
464	0.11	0.16	0.16	0.20	0.23	0.26	0.27	0.30	0.30	0.32	0.35	0.38
...												

	Entity											
	469	1017	391	683	806	405	1313	218	143	810	167	1282
...												
464	0.07	0.11	0.11	0.14	0.15	0.17	0.17	0.17	0.19	0.19	0.20	0.24
...												

Figure 13 – Data Mining Fusion

Employing the theme-based proximity metric to apply a sequential sort over the first sort improves the overall performance of the entity disambiguation task. The performance metrics after the second proximity based sort are:

- Top Matched Percentage = 33% (from 26% - improvement of 27%)
- Matched Percentage 20 = 75% (from 64% - improvement of 17%)

The performance enhancements produced by combining multiple sources of information reinforces the analytical framework and data processing pipelines that were researched and prototyped. Each of the processing streams is set up to provide a ranked list of matches, which can be combined using multiple means; for example a voting scheme, or any customized sequence of sorts as shown above.

We also compared the performance of the data mining algorithms with a random selection experiment from the demonstration dataset. This experiment was chosen because there are no benchmarking problems with which to assess our results, and the selected dataset contains ground truth that we can use to evaluate. The statistics for the random selection problem are as follows:

- Top Matched Percentage = 0.11%
  - For each of the 916 observed entity occurrences that contain the entities with multiple occurrences, there is a 1/915 chance of a match
- Matched Percentage 20 = 8.47%
  - The probability of there being a match in the top 20 is (1 – the probability that there is not a match in the top 20)
  - The probability that there is not a match in the top 20 is:
    - (912/916)*(911/915) … (repeat 20 times)

Table 1 summarizes the results:

- Three orders of magnitude improvement over a random selection for the top matched % metric
- Order of magnitude improvement over a random selection for the matched top 20 metric
- 22% improvement for the top matched % metric when adding theme features to sensor features
- 17% improvement for the matched top 20 metric when adding theme features to sensor features

Table 1 – Summary of Results

	Random	Sensor Features	+ Theme Features
Metric 1 (Best Match)	0.11%	27.00%	33.00%
Metric 2 (Top 20 Matches)	8.47%	64.00%	75.00%

7 Rule-Based Entity Disambiguation

The data mining entity disambiguation previously described can be augmented via domain rule disambiguation. The objective in adding this capability is two-fold: (i) to apply knowledge to "fill in the gaps" where data mining falls short, and (ii) to develop an approach to rule development that is as simple and easy to understand as possible.

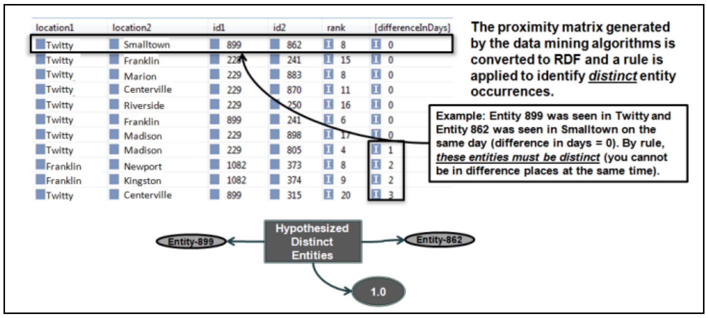
Example rules for each data mining feature group include:

- Physical feature rule – The same person can't be at different locations at the same time
- Factual feature rule – If a person is observed at Location X at Time T1, the same person is unlikely to be present at Location Y at time T2, for (T2 – T1) being 'small' and Distance (Y, X) being 'large' (rationale: infeasible to travel a 'large' distance in 'small' time)
- Relational feature rule – A person is unlikely to 'jump' more than two levels in hierarchy while communicating with other members of the network (so a member at Level 4 is unlikely to directly communicate with a member at Level 1)
- Process feature rule – In the context of an activity that requires several participants with specialized skillsets, if a person is known to be a specialist at one specialized activity type (say "Logistics"), it is unlikely that he will participate in an activity type requiring a vastly different skill say 'Recruitment')

Each of these example rules, and the class of rules that they represent, are generic in nature and can easily be parameterized using modern rule languages and representations such as SPIN[3]. This minimizes the number of rules that must be written, reducing the maintenance burden of the rule base.

Domain rules can be used to identify *distinct entities*; that is entities that are believed to be distinct from one another, with a high degree of confidence, using a handful of common sense principles. This is a counterintuitive result as typically rules are used to identify entities that are the same. This section illustrates this concept using a rule that "two people cannot be in different locations at the same time."

Figure 14 shows an example application of this rule. The table in this figure shows information from the data mining proximity calculation and observations from the original data. Each row represents a pair-wise association between the entities with IDs as shown in the column id1 and id2 with ranking as determined by the data mining algorithms. For example, in the first row, the data mining algorithms determined that the Entity-862 is the 8th closest entity to Entity-899. In addition, each of these entities appeared in a separate observation where Entity-899 was seen in Twitty and Entity-862 was seen in Smalltown. Finally the column differenceInDays represents the different in the time (in days) of the two observations.



location1	location2	id1	id2	rank	[differenceInDays]
Twitty	Smalltown	899	862	8	0
Twitty	Franklin	228	241	15	0
Twitty	Marion	229	883	8	0
Twitty	Centerville	229	870	11	0
Twitty	Riverside	229	250	16	0
Twitty	Franklin	899	241	6	0
Twitty	Madison	229	898	17	0
Twitty	Madison	229	805	4	1
Franklin	Newport	1082	373	8	2
Franklin	Kingston	1082	374	9	2
Twitty	Centerville	899	315	20	3

The proximity matrix generated by the data mining algorithms is converted to RDF and a rule is applied to identify *distinct* entity occurrences.

Example: Entity 899 was seen in Twitty and Entity 862 was seen in Smalltown on the same day (difference in days = 0). By rule, *these entities must be distinct* (you cannot be in difference places at the same time).

Figure 14 – Example Application of Identify Distinct Entity Rule

To summarize, the first row represents the following information:

> On the same day, Entity-899 was observed in Centerville and Entity-862 was observed in Smalltown, based on data mining using sensor and theme features, the data mining algorithms predict that Entity-862 is the 8th closest entity to Entity-899.

Common sense tells us that Entity-899 and Entity-862 must be distinct entities, since one cannot be in two different places

[3] **SP**ARQL **I**nferencing **N**otation (http://spinrdf.org).

on the same day[4]. The benefit of this approach is the following: by identifying distinct entities, it is possible to improve the performance of the data mining algorithms. Any time that a rule identifies a distinct pair of entities, it effectively sets the ranking of those two entities to ∞. This allows other entities to "move up" in the ranking, some of which might be the actual disambiguated entity. For example, in the data shown in Figure 14, suppose that the entity in position 9 for the Entity-899 is a match. When we remove Entity-862 from consideration, then the real entity will move up a position to rank 8. This will improve the overall performance of the algorithm.

By tuning the filter to take into account greater differences in observations (see last four rows in the table), further disambiguation is possible.

8 Conclusions

The approach described in this paper was demonstrated to achieve an orders of magnitude increase in entity resolution on a realistic dataset. This approach includes a data mining pipeline that performs entity occurrence disambiguation on observation-based event streams. The domain rule application approach is a simple way to identify unique entities.

Our hybrid approach combines text analytics, data mining, machine learning, multi-sensor information fusion, process mining, event extraction, provenance analysis and knowledge-based (domain rule) methods for entity occurrence disambiguation.

The approach presented in this paper can be used for knowledge discovery by revealing information hidden in the data. With this approach, you will avoid the uncontrolled growth of data with its commensurate exponential reduction in the ability to derive useful information for the decision maker.

Acknowledgements

The research described in this paper was supported by the Office of Naval Research; contract number N00014-10-M-0438.

9 References

[1] Nemati, H., Phelps, M., Stoeffler, D. I. 1996. "Knowledge Discovery Through Data Mining," Advances in Knowledge Discovery and Data Mining, pp. 2- 16, AAAA/MIT.

[2] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurasamy, "Advances in Knowledge Discovery and Data Mining," AAAI Press/ The MIT Press, 1996.

[3] Bishop, Christopher M., (1996). "Neural Networks For Pattern Recognition." New York. Oxford University Press.

[4] Chris Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999.

[5] Dmitri V. Kalashnikov and Sharad Mehrotra, "A probabilistic model for entity disambiguation using relationships," TR-RESCUE-04-12 Jun 1, 2004.

[6] Mehrnoush Shamsfard, Maryam Sadr Mousavi, "Thematic Role Extraction Using Shallow Parsing," International Journal of Information and Mathematical Sciences 4:2 2008.

[4] Note that this is an approximation for illustration purposes because it may be possible for the same person to appear in difference places on the same day, if the two places are close enough together, the person can get from one place to another, etc. Future work will involve refining these rules.

A Framework for Ontology Life Cycle Management

Perakath Benjamin, Nitin Kumar, Ronald Fernandes, and Biyan Li
Knowledge Based Systems, Inc., College Station, TX, USA

Abstract - *This paper describes a method and automation approach for ontology life cycle management. First, the challenges associated with knowledge creation are summarized. The problems associated with ontology capture, analysis, and maintenance are used to motivate the need for ontology life cycle management. Then a semi-automated method for ontology life cycle management is described. The method is comprised of multiple, inter-related activities including ontology extraction and learning, ontology refinement, ontology analysis, ontology harmonization and integration, and ontology validation and deployment. Next, a layered approach for semi-automated ontology extraction and learning is described. The ontology extraction is performed using a combination of statistical methods and rule based methods. Finally, the paper summarizes the status of current research and suggests directions for future research.*

Keywords: Ontology Management, Knowledge Discovery, Ontology Learning, Natural Language Processing, Information Extraction

1 Motivations

1.1 Knowledge discovery and creation

The discovery of knowledge is one of the most desirable end products of computing [1]. The knowledge based systems approach to AI was motivated by the observation that knowledge is an essential characteristic of human intelligence. Human agents acquire and use knowledge to perform intelligent tasks, but the goal of producing artificially intelligent machines that reason with huge amounts of high quality knowledge has remained elusive so far. One of the main impediments to producing truly intelligent knowledge based systems (KBS) has been the *knowledge acquisition bottleneck:* the intrinsic complexity and huge cost of capturing knowledge directly from experts. Another impediment to KBS development has been an inability to *reliably and rapidly convert data to knowledge*. A third problem inhibiting KBS development is the *dynamic requirements of knowledge based systems*. Needed are systems that continuously revise their knowledge content and the mechanisms that apply that knowledge to new and emerging problem situations.

A central barrier to the cost-effective development of knowledge based systems is direct Knowledge Acquisition (KA) from humans. In spite of well-documented KA techniques (interviews, questionnaires, protocol analysis, facilitated meetings, QFD, etc.) and sophisticated KA tools, KA is slow and expensive. Moreover, the knowledge about a domain is often very poorly documented and exists primarily in the minds of a few domain (or subject matter) experts. The KA activity is therefore highly dependent on the availability of (scarce) subject matter experts and knowledge engineers. KA development also suffers from the intrinsic inability of humans to clearly articulate what they know – a difficulty that is expected to continue in the future.

A prominent characteristic that distinguishes successful organizations is the ability to respond quickly, proactively, and aggressively to unpredictable change (‘agility’). Organizational agility requires, as a prerequisite, agility in the decision support systems used to support the operation of the enterprise. This requirement, in turn, necessitates agility in the knowledge bases and in the automated reasoning tools that use these knowledge bases. Needed are tools that will facilitate the dynamic update and enhancement of knowledge bases and the dynamic reconfiguration of applications that use these knowledge bases. That is, robust knowledge based systems must provide mechanisms for continually updating their internal models about the world.

1.2 Instructions ontology development: a mechanism for knowledge creation

Knowledge about that world is often captured, stored, and maintained as *Ontology* Models. Moreover, it is widely accepted that ontological analysis is an important first step in the construction of robust *knowledge based systems* [3]. The term ‘ontology” is used to refer to a catalog of terms used in a domain, the rules governing how those terms can be combined to make valid statements about situations in that domain, and the “sanctioned inferences” that can be made when such statements are used in that domain [4]. In every domain, there are phenomena that the humans in that domain discriminate as (conceptual or physical) objects, associations, and situations.

Ontology Development focuses on extracting the essential nature of the concepts in any domain and representing this knowledge in a structured manner. The construction of an ontology differs from traditional information capture activities in the *depth* and *breadth* of the information captured. Thus, an ontology development exercise will reach beyond asserting the mere *existence* of relations in a domain: the relations are “axiomatized” within an ontology (i.e., the *behavior* of the relation, in terms of the sanctioned inferences that can be made with the relation, is explicitly documented).

An ontology development project results in three products: (i) a catalog of the terms, (ii) the constraints that govern how those terms can be used to make descriptive statements about the domain, and (iii) a model that, when provided with a specific descriptive statement, can generate the “appropriate” additional descriptive statements. By appropriate descriptive statements we mean (i) because there are generally a large number of possible statements that could be generated, the model generates only that subset which is “useful” in the context; and (ii) the descriptive statements that are generated represent facts or beliefs that would be held by an intelligent agent in the domain who had received the same information. The model is then said to embody the sanctioned inferences in the domain. It is also said to “characterize” the behavior of objects and associations in the domain.

1.3 Ontology life cycle management

Ontologies, once created, are analyzed, stored, validated and updated to reflect the dynamics of real world phenomena and application situations. Ontologies thus have a ‘life cycle’ that needs to be managed in a systematic manner (Figure 1).

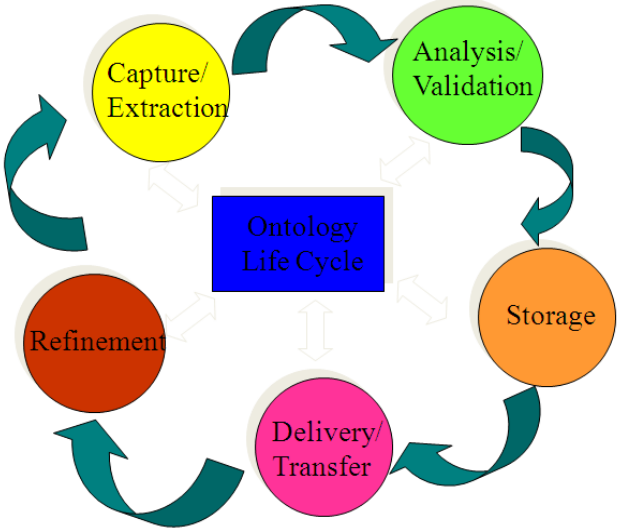


Figure 1. The Ontology Management Life Cycle Concept

This paper will provide an overview of a ‘life cycle management’ method. Then, we will focus our attention on one key activity in our method: an approach to performing semi-automated ontology learning and extraction. In order to understand the method, we define the term “Ontology Life Cycle Management (OLCM)” to refer to the collection of activities that are undertaken in order to create (elicit and extract), analyze, validate, store, deliver, and refine ontologies (Figure 1). A structured method that was formulated to address the technical and pragmatic challenges associated with OLCM is described in the next section.

2 Ontology life cycle management method

A method for OLCM is described in the following paragraphs. The ontology life cycle management activities are shown in Figure 2.

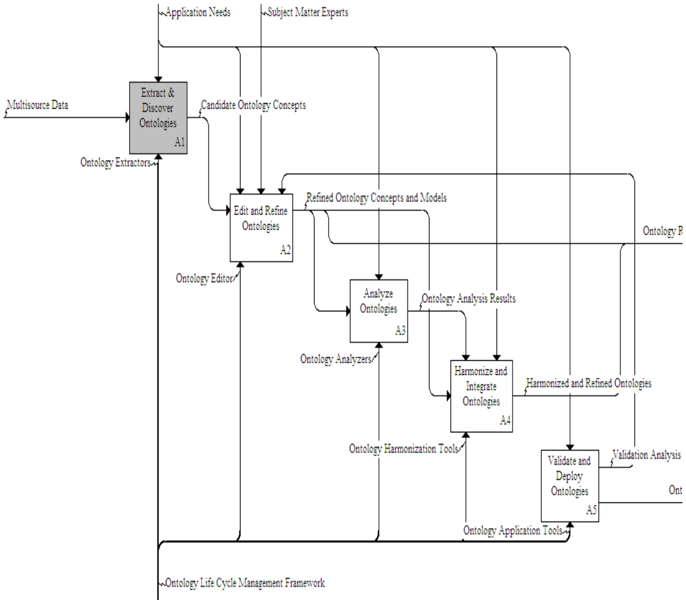

Application Needs
Subject Matter Experts
Multisource Data
Extract & Discover Ontologies
A1
Candidate Ontology Concepts
Ontology Extractors
Edit and Refine Ontologies
A2
Refined Ontology Concepts and Models
Ontology R
Ontology Editor
Analyze Ontologies
A3
Ontology Analysis Results
Ontology Analyzers
Harmonize and Integrate Ontologies
A4
Harmonized and Refined Ontologies
Ontology Harmonization Tools
Validate and Deploy Ontologies
A5
Validation Analysis
Ont
Ontology Application Tools
Ontology Life Cycle Management Framework

Figure 2. Ontology Life Cycle Management Method

We now briefly describe the activities supported by the ModelMosaic® Ontology Management Toolkit.

2.1 Discover and extract ontologies

This activity uses a layered approach to ontology extraction that discovers and extracts the elements of an ontology (concepts, relationships, etc.) from multi-source text data. The extracted ontology information “candidates” are presented to the end user for review for possible inclusion into the evolving ontology repository. Automated ontology comparison methods are used to provide the user with guidance in understanding the differences and similarities between the discovered ontology elements and the pre-existing ontology repository.

2.2 Edit and refine ontologies

This activity provides automation support for ontology elicitation (direct capture from subject matter experts) and ontology editing. Automated support for this activity includes (i) graphical and tree view navigation and editing, (ii) text editing, and (iii) ontology visualization. Visual aids are particularly useful in navigating and understanding an ontology model.

2.3 Analyze ontologies

This activity refers to the use of different comparative analysis techniques to analyze ontologies and ontology collections. The analysis methods are motivated by the observation that ontologies grow through the re-use and modification of pre-existing and often ill-formed ontology fragments. The ontology comparison and mapping methods enable the mediation, harmonization, and integration of ontologies to support advanced, knowledge-intensive applications.

2.4 Harmonize and integrate ontologies

The Harmonization activity refers to the resolution of potential ontology mismatches and inconsistencies that result from the ontology comparison analyses. Integration is the activity of composing an ontology model from two or more ontology models or "sub models."

2.5 Validate and deploy ontologies

This activity refers to evaluating the merit of an ontology model relative to the objectives for which it was designed. The ontology design process often requires that the ontology application address a set of "competency questions" that are identified early in the ontology development process. Once the ontology model has been validated, it is packaged and deployed to support one or more knowledge-intensive applications.

The next section will describe an approach to providing automated support for the "Extract and Learn Ontologies" activity.

3 Ontology extraction and learning approach

3.1 Ontology extraction: a layered strategy

A layered strategy is used to augment 'deep' ontology extraction with 'surface' concept extraction. The extracted 'surface' concepts provide a 'filtering' mechanism that allows end users an effective means for navigating the (potentially large number of) concepts within the text corpora. This idea is illustrated in Figure 3.

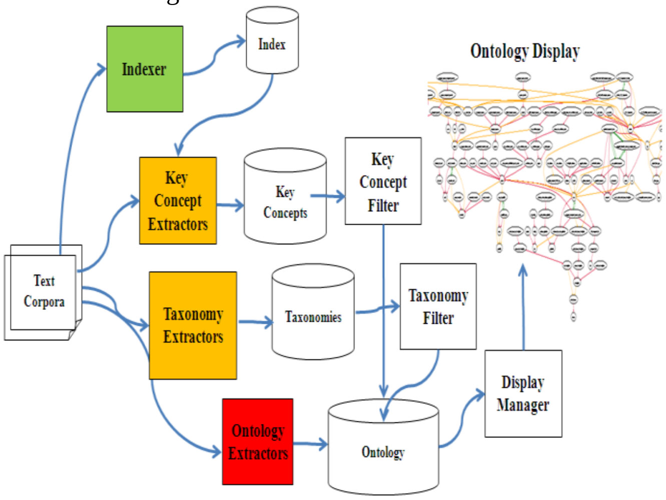


Figure 3. A Layered Approach to Knowledge Extraction From Text

Figure 3 illustrates a 'staged' extraction process: first, a key concept extraction activity, followed by a taxonomy extraction activity, and finally a deep ontology extraction activity. The results of the activities are combined and presented to the user. The ontology extraction and learning process comprises the following information processing tasks: format conversion, tokenization, sentence boundary detection, part of speech tagging, key concept extraction, key concept-based filtering, relation extraction, relation clustering, and ontology composition.

3.2 Parsing, format conversion, and tokenization

This activity includes parsing, format conversion (to the internal text representation format of the extractor tools), and filtering of non-value adding symbols and words.

3.3 Tokenization and paragraph boundary recognition

The purpose of *tokenization* is to separate the text stream into 'words' or tokens. Here 'word,' or token, is a word in the traditional sense and also numbers, punctuation marks, and other items that may prove useful in extracting the semantics. *Paragraph boundary recognition* refers to the identification of paragraph separators such as carriage returns, line feeds, vertical tabs, etc.). They are performed using a heuristic approach.

3.4 Sentence Boundary Detection

Sentence boundary detection methods are used determine the occurrence sentence boundaries tokens. The method is accomplished using a Hidden Markov Modeling (HMM) approach. The HMM uses contextual information about features, such as (i) the probability of part of speech of the surrounding words (one word to the left and one word to the right of the punctuation mark); (ii) abbreviation or an honorific (Dr., Mr., etc.) to the left of the punctuation mark; and (iii) the number of words between consequent sentence boundary candidates.

3.5 Part of speech (POS) tagging and POS disambiguation

This activity identifies the Parts Of Speech (POS) for the terms in the document. Multiple methods are used to perform POS identification, including (i) using 'look-ups' such as the WordNet® lexical database, (ii) Hidden Markov Modeling (HMM), and (iii) Transformational Learning. It is important to note that a given word may have (i) multiple POS uses (e.g., launch may be used as a noun or a verb), and (ii) multiple meanings for each POS use [e.g., the noun use of launch may be "take-off" (for instance, for space vehicles) or "presentation" (for instance, a sales briefing)]. POS tagging helps with the initial disambiguation relative to the multiple meanings of a term, and it also helps with interpreting the roles of the terms in relationships (e.g., noun – noun relations vs. noun –verb relations). Part of Speech Disambiguation narrows the scope of the possible meanings of a term. Techniques such as Hidden Markov Models (HMM) are used to facilitate POS Disambiguation.

3.6 Key concept extraction

The ontology extraction activity seeks to extract a large number of 'candidate' entity classes, the attributes of these classes, and the relationships between these classes. The extracted 'ontology model' is then presented to the end user of an ontology tool. In situations where the input data corpus is large, the extracted ontology models are relatively 'large.' Consequently, the end user is overwhelmed in trying to understand the ontology model. The extraction of key concepts is an early step in the knowledge discovery process and provides a mechanism for addressing the ontology management 'usability/information overload' issue. The idea of a 'key concept filter' is employed to reduce the end user information overload. In particular, a key concept filter helps filter the displayed information from the extracted ontology model.

The Key Concept Extraction process uses two corpuses of different domains to extract the key concepts of a particular domain. These two types of corpuses are the contrast corpus and domain corpus. The contrast corpus is a general corpus, and the domain corpus contains documents from a specific domain. The whole extraction process includes four major steps: indexing, text parsing, frequency generation and term weighting (**Error! Reference source not found.**).

The domain and contrast corpus are processed using the first three steps separately and merged on the final step to assign weight to the terms extracted and generate the key concept list along with the weight value. The assumption in this approach is that the domain concepts will occur relatively more frequently in the domain corpus than in the contrast corpus.

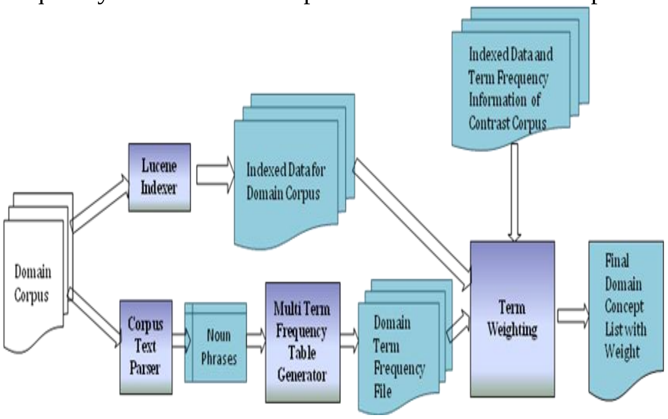


Figure 4. Key Concept Extraction Process

The indexing is accomplished by using a Lucene Index [5] to obtain a term's document frequency, which will be used later in term weighting. The Corpus Text Parser is used to generate noun and verb terms, which are passed into the table generator to generate a frequency table. The final weighting of terms is performed using an adaptation of the functions designed by Jiang and Tan [6]. The whole process can be accomplished in the run time, but it might take a significant period of time to finish the indexing and preprocessing. We pre-processed the first three steps of our chosen contrast corpus: 2000 documents from TREC data set [7]. We also allow for the indexing of the domain corpus to be done at either run time or as an off-line activity.

3.7 Key concept-based ontology filtering

The notion of the relative importance of key concepts and document terms is used as part of the ontology filtering strategy. Figure 5 illustrates the enhanced key concept-based filtering strategy. As shown in the figure, the filtering helps focus the output ontology with more important ontology concepts and more useful relationships in the ontology.

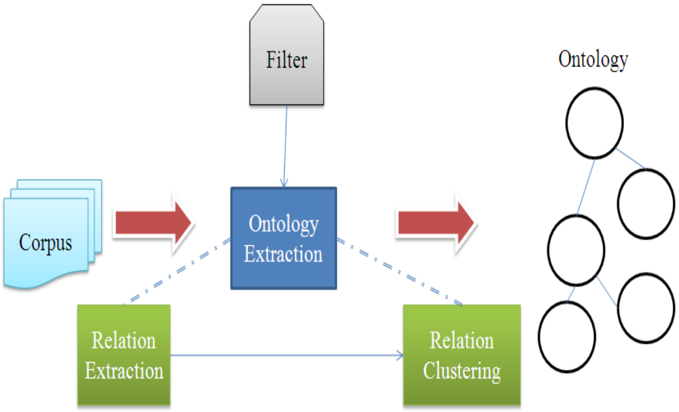


Figure 5. Key Concept-based Ontology Filtering Strategy

The filter module uses the key concept list generated by a Key Concept Extraction module. The filtering function checks whether the terms in the extracted relations are in the key concept list. If they are, then the relation is deemed to be 'important' and is passed through the filter as shown in Figure 5. A variant of the above approach is to include relation terms that are 'similar' to one or more key concepts.

3.8 Relation extraction

A 'pattern matching' approach is used to discover relationships embedded in text data. We use machine learning methods to learn (linguistic) patterns that are indicative of the occurrence of relations in natural language text. The patterns are organized and stored in a 'pattern library' (Figure 6).

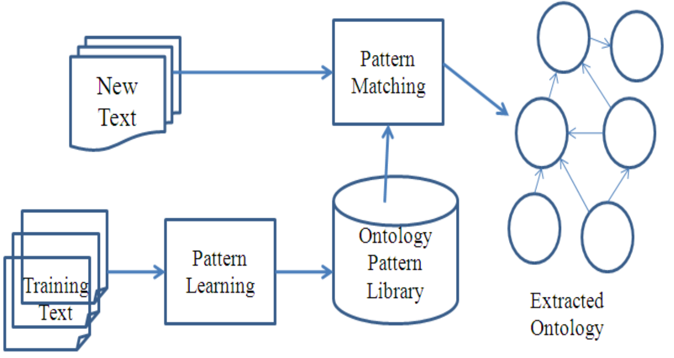


Figure 6. Pattern Matching Approach to Ontology Extraction

Relation extraction is performed by comparing the occurrence of terms within a sentence with the patterns in the pattern library. The relation extraction process produces a list of 'discovered' relations in the next based on the degree of match found during pattern matching.

3.9 Relation clustering and Ontology Composition

Once the relations have been extracted, statistical clustering methods are used to group and organize the collection of concepts. The concept clusters and relations are represented as graphs. The vertices are used to denote the concepts and the edges are used to denote the relations between concepts. Concept clusters that have a greater number of concepts than others are determined to be the "key ontology concepts" within the text copora. The clusters are presented for review and evaluation by the end user. The end user may accept/reject the discovered relationships and also assert additional relations between concepts (based on his or her domain knowledge). Multiple graphs are merged by identifying the relationships between them. Automated graph merging requires the algorithm to analyze the relationships between concepts in one graph and concepts in another graph. The relation clusters are then used to compose integrated ontology models from component ontology/relation clusters.

4 Current status and future research directions

This section outlines the current status of this research and identifies areas that are the focus of ongoing and future work.

4.1 Current research status

An automated software toolkit had been developed that supports the ontology life cycle management method described in this paper. This toolkit has been tested and validated by multiple customer communities of interest. Multiple technology application areas have been investigated, including (i) requirements management for complex systems, (ii) collaborative missile defense planning, (iii) cross domain secure information sharing, and (iv) cyber security management and assurance.

4.2 Ongoing and future research directions

Areas of ongoing and planned research are focused on the following areas.

- *Automated Ontology Extraction and Maintenance Methods:* techniques that use Natural Language Processing methods for automated ontology extraction and maintenance from text data sources.

- *Ontology-Assisted Semantic Mapping Methods:* semantic analysis techniques for mapping / comparing ontologies.

- *Ontology Alignment and Harmonization Methods:* rule based techniques for aligning and harmonizing ontologies.

5 Acknowledgements

The research described in this paper was partly supported by projects from the Office of Naval Research [8] and Rome Air Force Research Laboratory [9].

6 References

[1] G. Wiederhold. "On the Barriers and Future of Knowledge Discovery", In *Advances in Knowledge Discovery and Data Mining*, Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (Eds), AAAI/MIT Press, Menlo Park, CA., 1996, pp. vii-xii.

[2] P. Benjamin, Darr. T., and Corlette, D. "A Semantic Technology Framework for Knowledge-Intensive Applications", In *Proceedings of IC-AI 2009*, Arabnia, H.A., de la Fuente, D., and Olivas, J. A. (Eds), CSREA Press, pp. 416~421. July 2009.

[3] J. Hobbs, Croft, W., Davies, T., Edwards, D., and Laws, K., 1987. *The TACITUS Commonsense Knowledge Base*, Artificial Intelligence Research Center, SRI International.

[4] P. Benjamin, Menzel, C., and Mayer, R. "Towards a Method for Acquiring CIM Ontologies", *International Journal of Computer Integrated Manufacturing*, 8 (3), 1995, pp. 225-234.

[5] http://lucene.apache.org/java/docs/index.html.

[6] X. Jiang and A.-H. Tan, "Mining Ontological Knowledge from Domain-Specific Text Documents", In *Proceedings of the Fifth IEEE Int'l Conf. Data Mining (ICDM '05)*, pp. 665-668, 2005.

[7] National Institute of Standards and Technology (NIST) TREC Document Database: Disk 5. http://www.nist.gov/srd/nistsd23.cfm.

[8] Knowledge Based Systems, Inc. "Toolkit for Agent-based Knowledge Extraction (TAKE)", *Office of Naval Research (ONR) Contract Number N00014-05-C-0072*, 2005.

[9] Knowledge Based Systems, Inc. "Adaptive Toolkit for Pattern Discovery (ATPD)", *Air Force Contract Number FA8750-07-C-0045*, 2007.

XSD To OWL: A Case Study

Aaron Wheeler, Jim Dike, and Michael Winburn
3 Sigma Research, Indialantic, Florida, USA

Abstract – *This paper addresses the challenge of applying Semantic Web technologies like OWL-DL reasoners to XML documents to find implied relationships not stated directly in the documents. The challenge comes because the structure of XML schema often contains implicit assumptions about taxonomy and relationships. This makes it difficult to implement a completely general and automated mechanism to transform XML schemas to OWL ontologies. We describe our efforts to transform XML schemas to OWL ontologies and discuss our rationale for particular mappings. This work contributes to this area of research by providing new interpretations of XSD terminology in the context of OWL and shows how to map XSD structures to more complex OWL structures.*

Keywords: knowledge representation, ontologies, semantics, OWL, XML

1 Introduction

In this paper we address the challenge of applying Semantic Web technologies like OWL-DL reasoners to XML documents to find implied relationships not stated directly in the documents. The challenge comes because the structure of XML schema often contains implicit assumptions about taxonomy and relationships [1]. This makes it difficult to implement a completely general and automated mechanism to transform XML schemas to OWL ontologies.

We describe our efforts to transform XML schemas to OWL ontologies and discuss our rationale for particular mappings. This work contributes to this area of research by providing new interpretations of XSD terminology in the context of OWL and shows how to map XSD structures to more complex OWL structures.

Section 2 of this paper provides background information about relevant structures in both the XML Schema language and in OWL. Section 3 describes examples of related work, our approach, and how our work differs from previous efforts. We also illustrate a how the proposed transformation fits in to a larger system architecture for automated reasoning about XML documents. Section 4 concludes with a brief description of a current project that we intend to apply these transformations, and also a discussion of future research directions.

2 Background

The Extensible Markup Language (XML) provides language for describing the structure of information to support automated processing [2]. The XML Schema Definition (XSD) language contains type and element definitions that describe characteristics of well-formed elements and attributes of XML documents [3]. However, the XSD language does not express semantics [1] and so creates difficulties for Semantic Web technologies.

An ontology provides a formal description of a domain of discourse [4] and represents a core component of the Semantic Web. The World Wide Web Consortium (W3C) has recommendations for expressing ontologies using RDF [5], RDFS [6], and OWL [9].

The Resource Description Framework (RDF) represents a W3C standard for the Semantic Web that makes statements using subject-predicate-object triples [5]. RDF Schema (RDFS) extends RDF by allowing for the definition of classes and properties that describe other classes and properties [6].

Both RDF and the OWL 2 Web Ontology Language provide a formal semantics for interpreting ontology structures. OWL extends RDF and RDFS with terminology to express ontologies using description logic, a decidable fragment of first order logic [7]. OWL DL ontologies belongs to subset of OWL ontologies that satisfy the expressive requirements of a description logic. The expressiveness, completeness, and decidability of OWL DL makes it possible for automated reasoning engines to discover new information implied by ontology structures. OWL 2 has terminology to express ontologies semantically equivalent to the SROIQ description logic [8].

2.1 XML Schema Definition (XSD) Language

The XSD language has a large vocabulary for describing the structure of valid XML documents. However, we do not need necessarily all of these descriptors to make semantic inferences about the documents. In this section we make note of those XSD entities that we choose to use for automated reasoning. See [3] for more details.

The xsd:attribute defines attributes of elements using build-in types like xsd:integer and xsd:string. The xsd:simpleType construct includes built-in types like string and integer, as well as derived types. We can restrict values using xsd:enumeration, xsd:minInclusive, xsd:minExclusive, xsd:maxInclusive, and xsd:maxExclusive.

The xsd:complexType element may have any combination of attributes, elements, and content. Complex types with simple content have attributes and content, with complex content have attributes and elements, and with mixed content have attributes, elements, and content. We can extend the content of complex types or place restrictions on existing content.

The xsd:element declares elements used to create entities. The substitutionGroup attribute for elements indicates that all elements belonging to this group may replace one another.

Several elements describe collections of elements. The xsd:group element allows for easy reference of collections of elements, while xsd:attributeGroup does the same for collections of attributes. For defining a particular ordering of child nodes we use xsd:sequence, for an exclusive choice we use xsd:choice, and for no particular ordering we use xsd:all.

Lastly, the elements xsd:annotation and xsd:documentation provide human-readable descriptions of XSD constructs.

2.2 Web Ontology Language (OWL)

In OWL we declare sets of individuals with owl:Class. We relate individuals to data values with owl:DatatyeProperty and to other individuals with owl:ObjectProperty. We specify the equivalence of classes with owl:equivalentClass and superclass/subclass relationships with rdfs:subClassOf. OWL allows for set definitions through the existential (some) and universal (only) quantifiers owl:someValuesFrom and owl:allValuesFrom. We represent intersections and unions of set restrictions with, respectively, owl:intersectionOf and owl:unionOf. We restrict datatype property values in our set definitions using owl:oneOf for enumerations; and owl:minInclusive, owl:minExclusive, owl:maxInclusive, and owl:maxExclusive for numerical ranges [9].

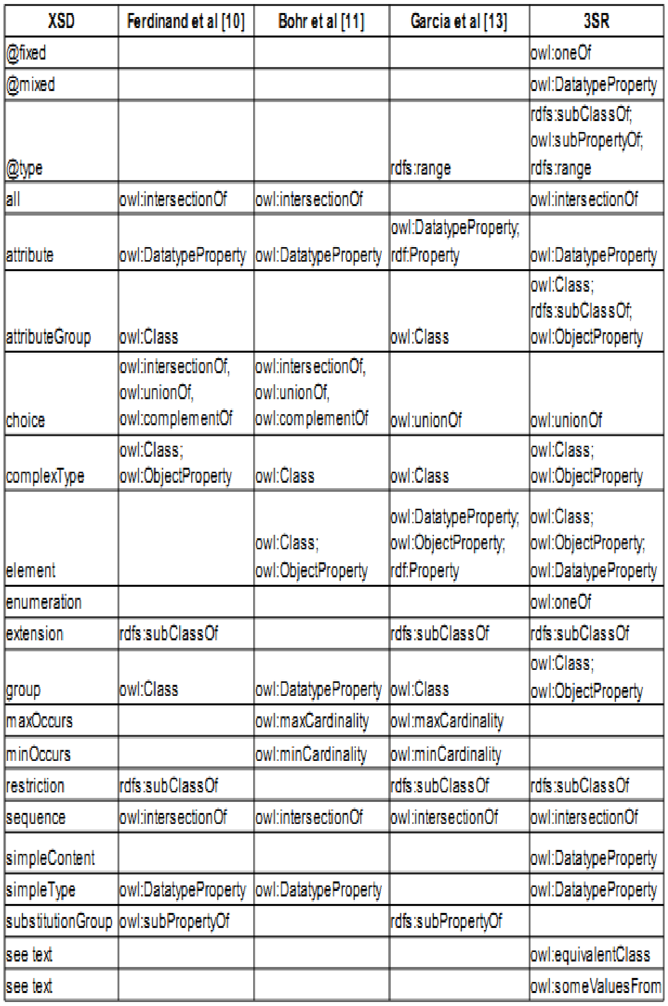
3 XSD to OWL

In this section we introduce similar work done by others to transform XSD to OWL, and then describe our methodology in detail.

3.1 Related Work

Table 1 shows the mapping between XSD and RDF/OWL proposed by Ferdinand et al [10]. They also use an intersection of owl:allValuesFrom and a cardinality restriction (owl:minCardinality or owl:maxCardinality) to express local XSD entities having a type and cardinality. They do not handle complex types defined by simple type restrictions[10]. Bohr et al [11] have a similar mapping (Table 1).

Table 1: Comparison of XSD to OWL mappings.



XSD	Ferdinand et al [10]	Bohr et al [11]	Garcia et al [13]	3SR
@fixed				owl:oneOf
@mixed				owl:DatatypeProperty

@type			rdfs:range	rdfs:subClassOf;
				owl:subPropertyOf;
				rdfs:range

all	owl:intersectionOf	owl:intersectionOf		owl:intersectionOf
attribute	owl:DatatypeProperty	owl:DatatypeProperty	owl:DatatypeProperty; rdf:Property	owl:DatatypeProperty

attributeGroup	owl:Class		owl:Class	owl:Class;
				rdfs:subClassOf;
				owl:ObjectProperty

choice	owl:intersectionOf, owl:unionOf, owl:complementOf	owl:intersectionOf, owl:unionOf, owl:complementOf	owl:unionOf	owl:unionOf
complexType	owl:Class; owl:ObjectProperty	owl:Class	owl:Class	owl:Class; owl:ObjectProperty

element		owl:Class;
		owl:ObjectProperty
			owl:DatatypeProperty;
			owl:ObjectProperty;
			rdf:Property
				owl:Class;
				owl:ObjectProperty;
				owl:DatatypeProperty

enumeration				owl:oneOf
extension	rdfs:subClassOf		rdfs:subClassOf	rdfs:subClassOf
group	owl:Class	owl:DatatypeProperty	owl:Class	owl:Class; owl:ObjectProperty
maxOccurs		owl:maxCardinality	owl:maxCardinality	
minOccurs		owl:minCardinality	owl:minCardinality	
restriction	rdfs:subClassOf		rdfs:subClassOf	rdfs:subClassOf
sequence	owl:intersectionOf	owl:intersectionOf	owl:intersectionOf	owl:intersectionOf
simpleContent				owl:DatatypeProperty
simpleType	owl:DatatypeProperty	owl:DatatypeProperty		owl:DatatypeProperty
substitutionGroup	owl:subPropertyOf		rdfs:subPropertyOf	
see text				owl:equivalentClass
see text				owl:someValuesFrom

We need to preserve as much as possible the semantic relationships expressed in the XML schema when we convert the XSD to OWL. We found a tool that comes close to our needs contained in the XSD2OWL package in the ReDeFer Project from the Rhizomik Initiative led by researchers from the GRIHO (Human-Computer Interaction and Data Integration) research group at the University of Lleida in Spain [12]. Table 1 shows the XSD to OWL mappings used by ReDeFer. The transformations may produce an OWL-Full ontology because it creates an rdf:Property for elements with both datatype and object ranges [13].

The commercial product TopBraid Composer [14] has an Eclipse plugin called XsdImport that converts XSD to OWL. XsdImport converts complex and simple types, attributes, attribute groups, as well as anonymous and named declarations to OWL classes and properties. However, XsdImport does not fully support OWL restrictions [15].

None of the efforts described above seem to provide us with sufficient support to create OWL representations of XML schema that we need for our particular automated reasoning applications. For example, we would like more automated support to create class restrictions from attributes and sub-elements for entities in the XSD. Our methodology, that we describe in the next section, for mapping XSD to OWL differs from others in at least two significant respects. Our methodology differs first in our handling of datatype properties and second in our use of restrictions for class definitions.

3.2 Methodology

In this section we describe our mapping strategy from XSD to OWL. The inspiration for our mapping strategy comes from the kinds of competency questions [16] we would like to answer with SPARQL queries over the inferred relationships generated by an OWL-DL reasoner.

Generally, if an XML document has particular a kind of element with some attribute value or some sub-element then we want to infer that this element belongs to some other class or has some particular property value or relation to some other specific entity or type of entity. Note that our strategy focuses on a transformation that allows for automated reasoning about the semantics of the XML documents and not on an equivalent representation of all document structure described in the schema. This means we can select from the schema only those aspects that we need for reasoning and leave out those related to validation. We assume that we reason only on valid XML documents.

Essentially, the entity in an XML document represents an individual. Attributes become datatype property relations and sub-elements become object property relations. Attributes may have a fixed value or enumerated values.

Table 1 shows our proposed mapping from XSD to RDF/OWL. The following sections provide details about particular decisions we made for the mappings.

owl:DatatypeProperty

We create DatatypeProperty definitions for element content and for attributes. Elements and complex types with mixed content also have a datatype property relation to the content. When a complex type allows for content, we create an owl:DatatypeProperty for this content. Complex types and elements with only simple content also receive owl:DatatypeProperty definitions.

owl:ObjectProperty

The XSD entities attributeGroup, group, complexType, and element that produce OWL class definitions also generate OWL object property definitions.

owl:Class

We create OWL classes for the XSD entities attributeGroup, group, complexType, and element because all of these could contain attributes and/or sub-elements. An attributeGroup represents a class with restrictions on the OWL datatype properties for each of the attributes in the group. We transform the group entity similarly, but using the OWL object property.

Complex type and element entities with only name and type attributes we could interpret as datatype properties. Instead, for consistent treatment of complex types and elements in general, we define these as classes having a datatype property relation to the type.

We model xsd:complexType in OWL as owl:Class defined with an owl:equivalentClass expressed as an owl:intersectionOf restrictions on the XSD sub-components listed in the complex type declaration.

In one respect that our approach differs from that implemented by ReDefer [12] lies in the interpretation of the XSD element. ReDeFer maps element to owl:DatatypeProperty, owl:ObjectProperty, or rdf:Property depending on the details of the element description. We interpret the XSD element as an owl:Class with an owl:DatatypeProperty relation to its content (if any).

owl:equivalentClass

We use an equivalent class of the intersection of restrictions on elements and attributes contained in the entity. The XSD terms all and sequence generate an intersection of all entities they contain. The any and choice entity produces an union of its component entities.

Attributes and sub-elements in an equivalent class definition allows us to infer that individuals of any class having these property relations also belong to this class.

owl:someValuesFrom

This restriction applies to enumerated datatype properties and object properties. We chose to express these properties with existential rather than universal restrictions because universal restrictions prevent a reasoner from inferring class membership of other classes and individuals not also having the universal restriction. This applies in particular to individuals of unknown type but with certain known properties.

Elements with attributes and content have relations to attribute data and content data, so in OWL they become classes defined as an equivalent class to the intersection of datatype properties for the attributes and content. Elements that can contain other elements have object property restrictions to these other elements.

We want to make explicit the relationships between classes and their property relations to other objects and datatype values. We do this with an owl:equivalentClass using an owl:intersectionOf of all owl:Restriction declarations created from the attributes and elements in the class.

rdfs:subClassOf and owl:subPropertyOf

Both the base and type attributes result in subclass and subproperty statements in the OWL representation.

xsd:choice

Other approaches to mapping XSD to OWL kept the exclusive-or interpretation of the XSD choice entity. We believe this creates combinatorial issues should a type have many choice elements with many alternatives. We choose instead to use a logical union represented by owl:unionOf on all the sub-entities. This interpretation of xsd:choice makes explicit in OWL the possibilities so that later one might add explicitly more specific exclusive-or restrictions for some combinations.

3.3 System Architecture

Figure 1 shows a system architecture for ontological reasoning about XML document semantics. One or more data sources provide data components that get collected into an XML document that validates to an XML schema(s). This XML schema we have previous transformed to an OWL ontology using the transformation rules described in the preceding section. We transform the XML document to an OWL individual and load it into a reasoning engine, such as Pellet [17], FaCT++ [18], or HermiT [19]. The reasoning engine generates an ontology model that contains the inferred relationships. We apply SPARQL [20] queries to this model to retrieve facts explicit in or implied by the data. These facts then become components of new data products.

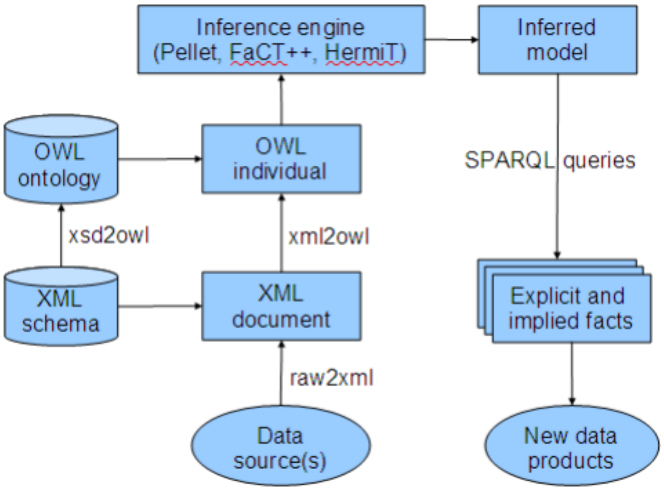


Figure 1. System architecture for ontological reasoning about document semantics.

4 Conclusions

We have shown a way to interpret key components of the XSD language in a way that allows their semantic representation as an OWL ontology. Our strategy for mapping XSD to OWL seems to offer to us better handling of datatype values, element content, and more elaborate restrictions on class definitions.

The representation of XML schema as ontologies makes it possible to apply automated reasoning engines infer additional relationships implied by the XML structures expressed in the schema.

We will continue to develop and refine the mapping in several ways. For example, we\do not currently create owl:minCardinality or owl:maxCardinality restrictions on property relations. However, we expect that a need for inference on cardinality will arise.

5 References

[1] OWL Web Ontology Language Overview. W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-owl-features-20040210/
[2] Extensible Markup Language (XML) 1.1 (Second Edition). W3C Recommendation 16 August 2006, edited in place 29 September 2006. http://www.w3.org/TR/2006/REC-xml11-20060816/
[3] XML Schema Part 0: Primer Second Edition. W3C Recommendation 28 October 2004. http://www.w3.org/TR/xmlschema-0/
[4] T.R. Gruber. A translation approach to portable ontology specifications. Knowledge Acquisition 5(2):199–220, 1993. http://tomgruber.org/writing/ontolingua-kaj-1993.pdf
[5] RDF Primer. W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-rdf-primer-20040210/

[6] RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation 10 February 2004. http://www.w3.org/TR/rdf-schema/

[7] From SHIQ and RDF to OWL: The Making of a Web Ontology Language by Ian Horrocks, Peter F. Patel-Schneider, and Frank van Harmelen. Journal of Web Semantics, 1(1):7-26, 2003. http://www.comlab.ox.ac.uk/people/ian.horrocks/Publications/download/2003/HoPH03a.pdf

[8] OWL 2 Web Ontology Language Document Overview. W3C Recommendation 27 October 2009. http://www.w3.org/TR/owl2-overview/

[9] OWL 2 Web Ontology Language Primer. W3C Recommendation 27 October 2009. http://www.w3.org/TR/owl2-primer/

§[10] Matthias Ferdinand, Christian Zirpins, D. Trastour. Lifting XML Schema to OWL. Web Engineering - 4th International Conference, ICWE 2004, Munich, Germany, July 26-30, 2004. http://vsis-www.informatik.uni-hamburg.de/getDoc.php/publications/204/fzt-lxs-04.pdf

[11] Hannes Bohring and Soren Auer. Mapping XML to OWL Ontologies. Leipziger Informatik-Tage, volume 72, 2005. pp. 147-156. http://www.informatik.uni-leipzig.de/~auer/publication/xml2owl.pdf

[12] ReDeFer. http://www.rhizomik.net/html/redefer/

[13] Roberto Garcia (2007). A Semantic Web Approach to Digital Rights Management. Dissertation. Department of Technologies, Universitat Pompeu Fabra, Barcelona, Spain. http://www.rhizomik.net/html/~roberto/thesis/

[14] TopBraid Composer. http://www.topquadrant.com/products/TB_Composer.html

[15] XsdImport – Convert XSD schemas to OWL. http://www.incunabulum.de/projects/it/xsdimport

[16] N.F. Noy and D.L. McGuinness. Ontology Development 101: A Guide to Creating Your First Ontology. http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html

[17] Pellet: OWL 2 Reasoner for Java. http://clarkparsia.com/pellet

[18] FaCT++. http://owl.man.ac.uk/factplusplus/

[19] Hermit OWL Reasoner. http://hermit-reasoner.com/

[20] SPARQL Query Language for RDF. W3C Recommendation 15 January 2008. http://www.w3.org/TR/rdf-sparql-query/

Automatic Generation of a Grounding Framework for Information Extraction

Anthony Stirtzinger, Steve Gorczyca, Joshua Powers and Matthew Dyke
Securboration Inc., Melbourne, FL, USA

Abstract - *The research described in this paper presents an approach to automatically generating a framework, called the Grounding Framework, which is applicable to any existing domain ontology and positions that ontology to be used as a model for information extraction. This approach allows a domain ontology to accurately reflect the domain while not compromising its structure and style in efforts to prepare it for use in information extraction. It also promotes the reuse and extension of existing ontologies for the purpose of information extraction without costly manual intervention. The research in this paper builds on existing research we have conducted for our Ontology Generation and Evolution Processor (OGEP) and Lexicalizing an Ontology.*

Keywords: Information Extraction, Knowledge Extraction, Ontology, Semantic, Framework

1 Introduction

Model-driven strategies for artificial intelligence have been successful in several problem spaces[1][2][3]. Common in these strategies is the idea that a top down approach to problem solving can lead to solutions gained from the use of common frameworks as opposed to implementation-specific solutions. These approaches leverage mechanisms that are guided (or driven) by an accurate representation of the environment (both problem space and solution space) in the form of models.

Several information extraction efforts have used ontological domain modeling to direct or improve extraction results. Most of these efforts require a combination of natural language processing (NLP), ontology development, and algorithms to match NLP discoveries with ontological concepts. Ontology plays the role of model in these model-driven applications. Generally there are two approaches for dealing with the ontology component of these solutions.

One approach is to generate the ontology either automatically or semi-automatically as in [4]. Strategies like these are tied closely to the NLP component, resulting in ontologies derived from the NLP mentions. Another approach is to build the ontology manually and then leverage the ontology during information extraction[5][6]. This type of approach provides more flexible separation from the NLP component and allows for domain declarations external to the particulars of a parsed corpus. However, there is a time and effort cost associated with building the ontology. Hybrid approaches[7] may start with a minimally sized ontology, then evolve that ontology from the NLP findings.

While the first approach greatly reduces the time required in the manual process of ontology building, the completeness and accuracy of the ontology remains in question. It is also difficult, if not impossible, to leverage assets of existing ontologies in combination with the generated ontology. The second approach requires the burden of manual construction, but usually results in a more accurate domain ontology. However, conventional ontology construction does not necessarily lend itself to the purpose of information extraction.

The research described in this paper presents an approach to automatically generating a framework, called the Grounding Framework, which is applicable to any existing domain ontology and positions that ontology to be used as a model for information extraction. This “automatic generation” approach allows the domain ontology to accurately reflect the target subject matter while not compromising efforts to prepare it for use in information extraction. It also promotes the reuse and extension of existing ontologies for the purpose of information extraction without costly manual intervention.

2 Motivation

As we have applied our Ontology Generation and Evolution Processor (OGEP) [8] more to information extraction as opposed to strictly ontology generation/evolution, critical needs have been uncovered. OGEP requires our Semantic Grounding Mechanism (SGM) constructs to exist in order to process NLP output and align it with ontological concepts. We have always created SGM constructs manually, which is time consuming, often taking weeks to complete. When a domain ontology already existed, but was not in the SGM structure, a manual alignment process was required. Therefore, to reduce the time of generating the SGM constructs and also allowing widespread reuse of existing ontologies, we initiated the effort described in this paper to automatically generate the Grounding Framework. This paper describes both the constructs and approach for generating the Grounding Framework from an existing domain ontology.

Since the focus of this research is based on the automatic generation of an SGM-compatible framework, our results will be shown in the context of successfully creating the Grounding Framework from an existing ontology. We present the results of a simple information extraction in this paper to demonstrate that the Grounding Framework can work with our existing SGM information extraction services.

3 Related Work

The Bank Ontology used throughout this paper was built using an upper level ontology, specifically the Basic Formal Ontology (BFO)[9]. BFO is an ontology that defines general universal types which can then be extended through sub-typing to define more specialized universals which comprise the content of a domain of interest. BFO is differentiated from other upper level ontologies such as DOLCE and SUMO by being narrowly focused on supporting the task of building domain ontologies for areas of scientific research. Along with BFO, the Bank Ontology made use of the Relation Ontology (RO)[10]. RO is similar to BFO in scale and purpose, but where BFO defines upper level objects and events, RO defines a set of core relations between objects, events and objects and events. In its Web Ontology Language (OWL) serialization, these relations are expressed using object properties. Finally, the RO_BFO_Bridge[11] was used. RO_BFO_Bridge is a serialization of the union of BFO and RO (in either OBO or OWL format) created by Chris Mungall.

The research described in this paper is dependent on the lexicalization of the domain ontology. [12] describes an automated approach for generating mapping candidates between WordNet synsets and target ontology objects. We use this approach to create mappings for our domain ontology to prepare it for automatically generating the Grounding Framework.

The WordNet lexical database[13] provides sense definitions for words (categorized into Noun, Verb, Adjective, and Adverb), called Synsets, along with a natural language descriptions, synonyms and relations to other senses. The sense relationships are categorized into hypernyms, hyponyms, holonyms and troponyms along others. These relationships allow checking whether a word belongs to a particular group (e.g. *dentist* is a *person*).

4 Approach

Our approach to automatically generating a grounding framework that provides an alignment mechanism between information extraction technologies and ontology is based on the following pre-exiting components: 1) a domain model representing the target for which to align the information; and 2) a lexical basis to relate the domain concepts to language.

The domain model is specified in an ontology language. For the tests performed in this research, the ontologies were expressed using Web Ontology Language (OWL). No other requirements are enforced on the domain model, however the quality of the lexicalization process (described later) increases when naming and/or label annotation of ontological components (classes and object properties) closely parallels domain definitions in the language that will be used for information extraction. During our research we tested with two basic ontology styles: 1) a domain ontology that did not make use of an upper level ontology; and 2) a domain ontology that made use of upper-mid-level ontology.

Our research made use of an automated process for generating mapping candidates between WordNet synsets and target ontology objects[12]. This research is intended to build upon [12] so that we may subsequently build a grounding framework for information extraction.

4.1 Grounding Framework Namespace

The Grounding Framework namespace is specified in OWL and is used as a base framework for representing both the domain ontology structures and the WordNet annotations in a way that they can be used to align unstructured text. Our original OGEP development used the SGM_0 ontology to define semantic grounding constructs that were recognized by the Semantic Grounding Mechanism (SGM) processor. The Grounding Framework is an evolution of the SGM_0 ontology that is more compatible with auto-generation. Both the SGM_0 and Grounding Framework ontologies are designed primarily to accomplish two functions.

First, SGM provides a way to extend the OWL Restrictions in the domain ontology such that they can be used to reason with partial evidence. This is done using the node-dependency-capability relationships contained with SGM as documented in [14].

Second, SGM is used to correlate the concepts in the domain ontology to the English language. This is accomplished through the use of the WordNet annotations as previously described along with verb-noun-qualifier attributions available within the SGM construct[7][8][14].

4.1.1 SGM Pattern in the Grounding Framework

Figure 1 shows the basic Node, Dependency, Capability pattern that supports the original SGM pattern. [7][8]and[14] describe more detail about the SGM pattern. The essence of the pattern is that a domain concept can be defined two ways. First, a domain concept can be defined by what it produces. We refer to this as a Capability. For example, the concept “exploded bomb” may have the capabilities of “smoke”, “fire” and “shrapnel”. Second, a domain concept can be defined by what must exist in order for the concept itself to exist. We refer to this as a Dependency. For example, the concept of “bicycle” may have dependencies of “wheel”, “seat”, “handle bar”, “frame”, and “pedal”. The SGM pattern provides a mechanism for defining the relationship of Domain Concept, Dependencies and Capabilities.

In the Grounding Framework ontology this pattern is realized as a gf:Node class and associated annotations. The three annotations are gf:groundsTo, gf:providesCapability and gf:requiresDependency. The values of the annotation are fully qualified ontology classes. Descriptions of the classes and annotations comprising this pattern follow.

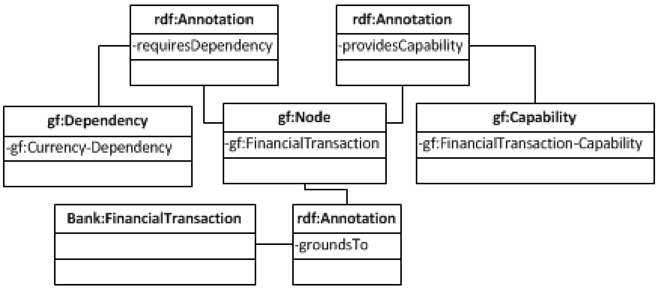


Figure 1 Basic SGM Pattern

gf:Node - Nodes are the ontology representations of the classes in the target ontology. Each node class will have a one to one correspondence to a class in the target ontology, and the class will be annotated with the 'groundsTo' annotation that points to the target class. Nodes require one or more Dependency and provide one or more Capability.

gf:groundsTo - The subject of a groundsTo rdf:annotation must be a gf:Node subclass. The value of the annotation must be the fully qualified name of the domain class that the gf:Node subclass grounds to, meaning that any individual of the gf:Node subclass is believed to be an instance of the domain class with some confidence level.

gf:Capability - Capabilities define the domain concepts that are produced (or left behind) if the targeted domain concept exists or existed.

gf:providesCapabaility – Annotation that is used to relate a gf:Node to a gf:Capability. This is accomplished by attaching the annotation to gf:Node with the value of the annotation being the fully qualified name of the gf:Capability.

gf:ConceptDependency - Dependencies indicate pre-requisites for a node; domain concepts that must exist in order for the target domain concept to exist. Dependencies are generally derived from the domain ontology owl:Restrictions on the target domain concept.

gf:requiresDependency - Annotation that is used to relate a gf:Node to a gf:Dependency. This is accomplished by attaching the annotation to gf:Node with the value of the annotation being the fully qualified name of the gf:Dependency.

4.1.1.1 Subclassing in Grounding Framework

The Grounding Framework ontology has defined subclasses for gf:Node, gf:Dependency, and gf:Capability. This subclassing parallels logic in the SGM ontology and the reasoning algorithms that rely on the SGM ontology. Subclasses for gf:Node include gf:Actor, gf:Concept and gf:Physical. Subclasses for gf:Dependency and gf:Capability mimic this subclassing pattern with gf:ActorDependency, gf:ConceptDependency, gf:PhysicalDependency, gf:ActorCapability, gf:ConceptCapability, and gf:PhysicalCapability.

This subclassing supports the ability to classify domain concepts into the three subclass categories to simulate a very simple upper level ontology. The research described in this paper classified all domain concepts as gf:Concept for purposes of simplification.

4.1.2 NLP Annotations

The Grounding Framework is used to support information extraction. During this process the Grounding Framework is used by the consumer of NLP output. Therefore, the Grounding Framework includes provisions for storing the NLP output. Storage of NLP output insures traceability between the information source and the domain ontology that is the subject of the extraction effort. Figure 2shows the ontology structures used to store the NLP output followed by a brief description of each.

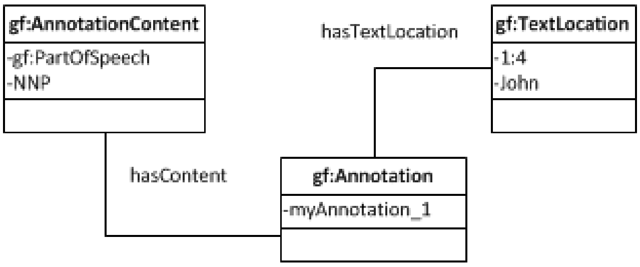


Figure 2 NLP Output in Grounding Framework

gf:Annotation - An ontology class that is used to store the information extracted from an NLP annotation (e.g. PartOfSpeech, Entity, etc.).

gf:TextLocation – An ontology class that is used to store the text location information extracted from an NLP annotation (e.g. PartOfSpeech, Entity, etc.). This information includes source document name and offset information of the NLP annotation.

gf:AnnotationContent – An ontology class that is used to store NLP annotation type-specific information. The Grounding Framework currently supports the following subclasses of gf:AnnotationContent: DTContent, FacilityContent, GPEContent, NamedEntity, OrganizationContent, PartOfSpeech, PersonContent, PosContent, StanfordDependency and WordNetContent.

4.1.3 WordNet Annotations in Grounding Framework

The Grounding Framework also contains some custom rdf:Annotations. These are used to provide the linkages for gf:Nodes (Dependencies, Capabilities) and the association of both domain ontology concepts and WordNet lexical groundings. For this research, the decision was made to represent this critical SGM information as rdf:Annotations as opposed to ontology classes or individuals. While this may potentially limit the ability to reason across this information using DL Implementation Group (DIG) reasoners, this strategy was chosen to encourage a loosely coupled relationship with regards to the original domain ontology structures. Following is a short description of each of the rdf:Annotations.

gf:wordnet3.0SenseKey - The subject of a wordnet3.0SenseKey annotation is associated with the SenseKey that is the value of the annotation.

gf:wordnet3.0SynSet - The subject of a wordnet3.0SynSet annotation is associated with the SynSet that is the value of the annotation.

gf:requiresDependency - Annotated on subclasses of gf:Node to indicate that all instances of that node must have a dependency of the class named in the annotation value.

gf:matchesDependencyClass - Annotated on a subclass of gf:Capability to indicate the dependency class that instances of the annotated class may fill.

gf:matchesCapabilityClass - Annotated on a subclass of gf:Dependency to indicate the capability class that instances of the annotated class may be filled by.

The following sections illustrate examples of instantiations of Grounding Frameworks using a small ontology that we refer to as the Bank Ontology.

4.2 Grounding Framework Instantiation

We created the Bank ontology to test and validate the algorithms that automatically generate an instance of the Grounding Framework from the prerequisites described in Section 4. This ontology will be used to show examples of a generated Grounding Framework in the following sections.

4.2.1 Bank Ontology to Grounding

Bank Ontology is a small ontology that contains a few concepts about the banking domain. These concepts are connected to the BFO and RO upper ontologies via subClassing. Figure 3 shows a segment of the Bank Ontology classes with domain-specific classes in bold and Upper/Mid-Level ontology classes in normal font.

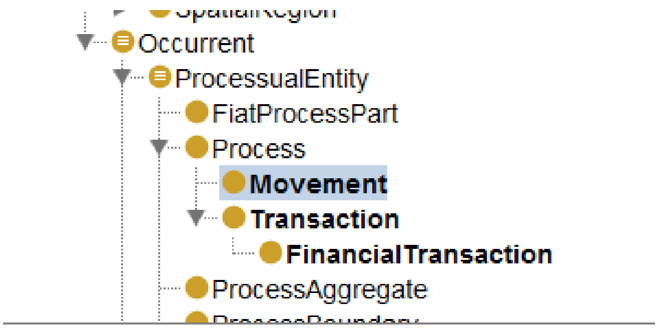


Figure 3 Bank Ontology Classes

When creating the Grounding Framework, we reflect the class hierarchy of the domain ontology via gf:ConceptNode classes and the owl:subClassOf property between them. Maintaining the domain ontology's class hierarchy inside the Grounding Framework allows for two significant capabilities. First, the inheritance relationships defined in RDF can be reasoned about using DIGs reasoners. Second, during the information extraction process of matching NLP discoveries to Grounding Framework concepts, the traversal of hierarchical relationships provides a convenient mechanism for evaluating hypernym and hyponym hierarchies in WordNet. Figure 4 shows as sample of the domain classes instantiated as gf:Concept nodes in the Grounding Framework.

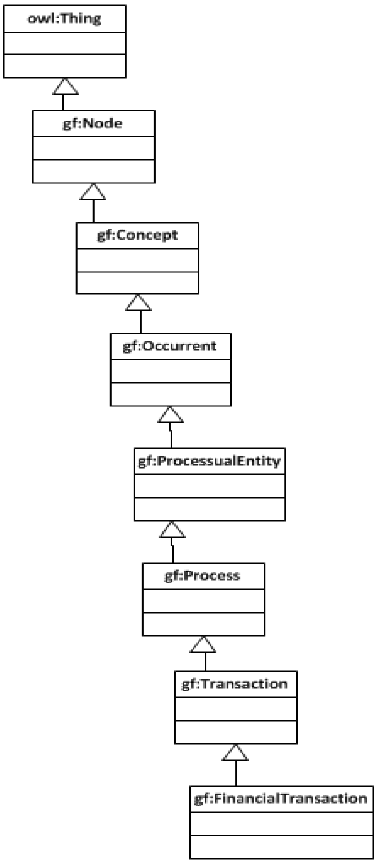


Figure 4 Bank Ontology Classes as gf:Concept Nodes

4.2.2 Bank Ontology Restrictions to Grounding

The Grounding Framework creation process also preserves the owl:Restrictions associated with each class. However, this information is not represented as an owl:Restriction, rather it is accomplished through the structures that exist within the Grounding Framework to support the SGM concepts. These structures include: 1) a set of custom annotations that are associated with each gf:Node type class; and 2) the use of gf:Dependency and gf:Capability class types.

The reason that owl:Restrictions are not used directly in the Grounding Framework is that we want to provide the ability to partially reason about the existence of domain entities. For example, in the Bank Ontology domain ontology, restrictions on a FinancialTransaction require the existence of a Transaction along with a has_participant relationship to some Currency. Unless both of these conditions are satisfied, the reasoner will not assert the existence of a FinancialTransaction. When performing information extraction, it is often beneficial to make partial assertions, or come to conclusions using partial evidence and associate a confidence or probability value to that conclusion. In the case where a transaction has been detected in the context of information extraction, but no currency is currently available from the corpus, there may be value in asserting the “possible” existence of a FinancialTransaction to some degree of certainty that is less than 100%. Figure 5 shows the domain restrictions for the FinancialTransaction class.

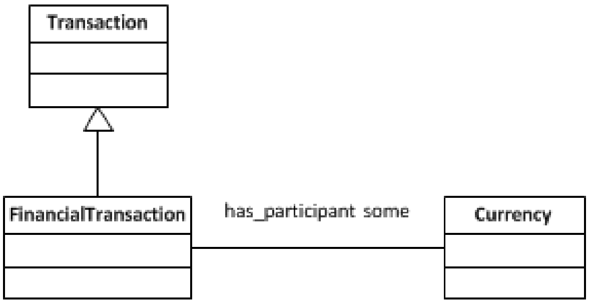


Figure 5 FinancialTransaction Restrictions

The custom annotations that are associated with each of the gf:Node class types to preserve the owl:Restriction declarations. These annotations consist of gf:groundsTo, gf:providesCapability, and gf:requiresDependency. These structures are converted to runtime graph structures during the information extraction processing. In this runtime structure, spread activation techniques can be applied to reason about the existence of domain concepts in unstructured text corpus. Below is a list of the custom annotations:

groundsTo:Bank#FinancialTransaction

providesCapability:GFl#FinancialTransaction-Capability

providesDependency:GF#Currency-Dependency

providesDependency:GF#FinancialTransaction-Dependency

4.2.3 Bank Ontology WordNet Grounding

The primary purpose of the Grounding Framework is to position the domain ontology for use during information extraction processing. To carry out this purpose, the Grounding Framework needs relationships that bind the domain ontology concepts to natural language. WordNet Synset and SenseKey information is used for this purpose. Leveraging the lexical annotations provided by [12], the Grounding Framework maintains the lexical groundings by attaching them to gf:Dependency classes.

As previously discussed, each of the owl:Restriction declarations from the domain ontology are captured through the use of gf:Dependency classes and custom annotations on gf:Node classes. The linkage of the Dependency classes to the English language is specified by gf:wordnet3.0SenseKey and gf:wordnet3.0Synset annotations. These annotations are attached to the Dependency class representative of its domain class counterpart.

During the SGM information extraction process, these annotations are used to match NLP mentions with domain ontology concepts through the use of synonym, hypernym, and hyponym comparisons. The WordNet annotations for the FinancialTransaction-Dependency class in the Grounding Framework are listed below:

wordnet3.0SenseKey "(dealing%1:04:02::)"

wordnet3.0SenseKey "(dealings%1:04:00::)"

wordnet3.0SenseKey "(financial%3:01:00::)"

wordnet3.0SenseKey "(fiscal%3:01:00::)"

wordnet3.0SenseKey "(transaction%1:04:00::)"

wordnet3.0Synset "(01106808n)"

wordnet3.0Synset "(02847894a)"

5 Results

To test our approach we focused on two areas. First, can we successfully generate the Grounding Framework from existing ontology? Second, once the Grounding Framework is created, is it compatible with the existing SGM matching algorithms? Each of these tests along with their results is discussed in the following sections.

5.1 Generation from Existing Ontology

Three different domain ontologies were used to test the generation of the Grounding Framework.

5.1.1 Bank Ontology

This ontology was built expressly for the purposes of the research. As previously discussed, this ontology is a small ontology that also makes use of the BFO and RO upper level ontology. The Bank ontology has 56 classes and 24 object properties using 6 namespaces. Generation time for the lexicalization was 6.14 seconds while generation time for the Grounding Framework was 2.98 seconds.

5.1.2 Wine Ontology

The Wine Ontology[15] is a popular sample ontology used in the OWL specification documents. The Wine Ontology has 138 classes and 16 object properties using 2 namespaces. Generation time for the lexicalization was 6.5 seconds while generation time for the Grounding Framework was 3.34 seconds.

5.1.3 Large BFO Ontology

In order to test a larger ontology and one that incorporated the concepts of BFO but was not developed with the intentions of testing the Grounding Framework, we used what we refer to as the Large BFO Ontology. The Large BFO Ontology has 14 ontology files, 1530 classes and 95 object properties. Generation time for the lexicalization was 114 seconds while generation time for the Grounding Framework was 61 seconds.

5.1.4 Grounding Framework Generation Conclusions

Based on these tests, it is evident that significant time savings can be realized when compared to manually creating SGM compliant structures. Our testing showed that an experienced developer can manually create a Grounding Framework at approximately 2 minutes per domain ontology class and 2 minutes per object property. The automatically generated times per class/object property for each ontology are below:

Bank Ontology	0.114 seconds per class/object property
Wine Ontology	0.063 seconds per class/object property
Large BFO	0.107 seconds per class/object property

Our tests also demonstrated that multiple types of ontology can be utilized for generating a Grounding Framework.

In comparing the manually created Grounding Framework versus the automatically generated versions, two shortcomings are recognized. First, the depth of the SGM definition is completely driven by the domain ontology. This often limits the richness in the Dependency and Capability definitions, thus limiting information extraction results. Second, the total reliance on the lexicalizing of the ontology using WordNet also reduces the richness of the semantic definitions. This approach excludes domain specific terms, slang and derivatives not found in WordNet. Both of these shortcomings are addressed in Section 6 and in [12].

5.2 Testing SGM Compatibility

The final testing involved parsing test sentences with the SGM information extraction algorithm using the automatically generated Grounding Framework. The results were compared against those obtained using a manually generated Grounding Framework over the same sentences.

5.2.1 Test Sentence One

Results of matching the second test sentence to the domain ontology using the Grounding Framework follows.

‘John Smith’ is identified as a person’s name, and a personCapability individual is asserted. ‘entered’ is identified as ‘movement’ verb through WordNet, and a movementCapability individual is asserted. ‘Summit Bank’ is not identified as a named entity. These results are the same as the original OGEP SGM processing.

5.2.2 Test Sentence Two

Results of matching the first test sentence to the domain ontology using the Grounding Framework are as follows.

‘John Smith’ is identified as a person’s name, and a personCapability individual is asserted. ‘withdrew’ is identified as ‘transaction’ verb through WorldNet, and a transactionCapability individual is asserted. ‘$10,000’ matches a currency format and is assigned the monetaryValueCapability and the currencyCapability. These results are the same as the original OGEP SGM processing.

5.2.3 SGM Compatibility Conclusions

Based on the simple parsing tests as described, we have concluded that the automatically generated Grounding Framework is completely compatible with the SGM matching algorithms.

6 Future Work

The original objectives of our research were accomplished; primarily, develop a mechanism to automatically generate the Grounding Framework from existing, lexicalized ontology. One of the limitations encountered when using existing ontology was the classic model-based issue which is the fact that the Grounding Framework is completely reliant on the composition of the domain ontology (i.e. the model).

Since domain ontology is generally not developed with information extraction in mind, there are potential weaknesses implicit between the ontological construction and what is most beneficial for information extraction. The primary area of concern was the lack of depth in the owl:Restrictions within the domain ontology. SGM has been shown to produce the best results when a large number of restriction-derived Dependencies are present for the ontological concepts that are to be matched. This is not a typical means for constructing domain ontology.

Future work is planned to introduce supplementary logic into the Grounding Framework generation process to attempt to dynamically enrich current ontological concepts as information extraction targets. One potential strategy is to leverage lexical databases such as WordNet, VerbNet, or FrameNet to elaborate the owl:Restriction entities in the context of language derivations and relationships.

7 Acknowledgements

This material is based in part upon work supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory contract number

FA8650-10-C-7062. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL or the U.S. Government.

8 References

[1] A. Deshpande, C. Guestrin, S. Madden, J. Hellerstein and W. Hong, “Model-driven data acquisition in sensor networks”; VLDB '04 Proceedings of the Thirtieth international conference on Very large data bases - Volume 30.

[2] Songnian Zhou, “A Trace-Driven Simulation Study of Dynamic Load Balancing”; IEEE Transactions on Software Engineering, Vol. 14, No. 9. (September 1988), pp. 1327-1341.

[3] K.Kiili, “Digital game-based learning: Towards an experiential gaming model”; The Internet and Higher Education, Vol. 8, No. 1. (MarchJanuary 2005), pp. 13-24.

[4] Hoifung Poon and Pedro Domingos “Unsupervised Ontology Induction from Text”; Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 296–305, Uppsala, Sweden, 11-16 July 2010.

[5] Maria Vargas-vera , Enrico Motta , John Domingue , Simon Buckingham Shum , Mattia Lanzoni, “Knowledge Extraction by using an Ontology-based Annotation Tool”; In K-CAP 2001 workshop on Knowledge Markup and Semantic Annotation.

[6] Adrian Silvescu, Jaime Reinoso-castillo, Vasant Honavar, “Ontology-Driven Information Extraction and Knowledge Acquisition from Heterogeneous, Distributed Autonomous Biological Data Sources”; In Proceedings of the IJCAI-2001 Workshop on Knowledge Discovery from Heterogeneous, Distributed, Autonomous, Dynamic Data and Knowledge Sources.

[7] C. Anken, A. Stirtzinger and B. McQueary. “Goal-Driven Semi-Automated Generation of Semantic Models”; In Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defence Applications (CISDA), July, 2009.

[8] A. Stirtzinger, C. Anken. “Semi-automated ontology generation and evolution”; In Proceedings of the SPIE, Volume 7347, pp. 734706-734706-10, April, 2009.

[9] Grenon P, Smith.B., Goldberg L.: Biodynamic ontology: applying BFO in the biomedical domain. Stud Health Technol Inform. 102 (2004) 20-38

[10] Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall CJ, Neuhaus F, Rector A, Rosse C Relations in Biomedical Ontologies.Genome Biology, 2005, 6:R46

[11] Downloadable from: http://www.obofoundry.org/ro/

[12] Joshua Powers and Anthony Stirtzinger, “Lexicalizing an Ontology”; conference proceedings 2011 International Conference on Information and Knowledge Engineering (IKE) July 18-22, 2011.

[13] G. A. Miller. “WordNet: A Lexical Database for English”; Communications of the ACM, Vol. 38 No. 11: 39—41, December, 1995.

[14] Attila Ondi, Anthony Stirtzinger: Information Discovery Using VerbNet: Managing Complex Sentences. IC-AI 2010: 268-276.

[15] W3C. Wine Ontology; http://www.w3.org/TR/owl-guide/wine.rdf, December, 2003.

Lexicalizing an Ontology

Joshua Powers and Tony Stirtzinger
Securboration Inc., Melbourne, FL, USA

Abstract – *Rich lexica such as WordNet are valuable resources for information extraction from unstructured text. When extraction techniques have formal ontologies as their targets, a mapping from the lexicon to the ontology has been shown to be beneficial in sense disambiguation and usability of the extracted knowledge. Such mappings are generally established manually, which can be a costly procedure if either the lexicon or the ontology is large. This paper describes an approach to accelerate this mapping process via automation using WordNet as the lexicon and a variety of standard ontologies. Times required to create useful mappings are measured across various parameterizations.*

Keywords: Ontology, WordNet, lexicography, sense disambiguation

1 Introduction

Accurate mappings from a rich lexicon such as WordNet[1] to formal ontology objects are of significant benefit to information extraction from unstructured text. Such extraction techniques may be used to grow an ontology or to align extracted knowledge to a formal context for later reasoning and analysis. The mapping of the lexicon to the ontology prior to processing text aids with sense disambiguation and accurate selection of extraction targets within the ontology. This mapping process, while only necessary once per version of the lexicon and target ontology, is a time-consuming manual job.

A particularly useful characteristic of the WordNet lexicon is that it has a graph-like structure, with meaningful links between lexical entries. These links are not as formally defined as in a Web Ontology Language (OWL)[2] ontology or full first order logic language, but they provide a useful means for finding associations between synsets.

Many ontology objects, particularly classes, are compositional in nature, combining multiple senses and terms in a single class name. Linguistic resources such as WordNet seldom do this. An example is the class 'DryRedWine' found in the popular Wine Ontology, which should be mapped to the WordNet lexical entries for both 'red wine' and 'dry (as in liquor)'.

This paper describes an automated approach for generating mapping candidates between WordNet synsets and target ontology objects. We use this approach to create mappings for three ontologies, and compare the approach to a fully manual mapping process. Finally, we compare different parameterizations of the mapping process to determine optimal use of the automation.

2 Motivation

Our efforts in aligning information extraction results to formal ontologies have resulted in both an automated ontology growth mechanism - Ontology Generation and Evolution Processor (OGEP)[3], and a mechanism for detecting instances of subgraphs of an ontology within extraction results, even if the concept represented by the subgraph is not explicitly mentioned in the source text – the Semantic Grounding Mechanism (SGM)[4]. Each use of these technologies with a different target ontology requires a new set of mappings from the WordNet lexicon to classes of the ontology. We estimate that it takes an average of nearly 2 minutes to produce a mapping for each ontology object when an expert lexicographer uses an ontology editing tool such as Protégé[5] and the online WordNet browser[1].

3 Related Work

The ontologies used for this paper include two popular 'example' ontologies – the Wine Ontology[6] and the Pizza Ontology[7]. A third ontology is really a collection of related ontologies under the Basic Formal Ontology (BFO)[8] foundational upper level, including a number of The Open Biological and Biomedical Ontologies (OBO)[9]. This BFO collection is an 'industrial strength' ontology, built very carefully and used in numerous academic, commercial and government applications.

As referenced earlier, we use the WordNet lexical database. WordNet is a semantic lexicon for the English language. WordNet groups English terms, called 'lemmas,' into sets of synonyms called 'synsets', provides short, general definitions called 'glosses', and records some light semantic relations between these synsets. The purpose is twofold: to produce a combination dictionary and thesaurus that is understandable to a language user, and to support automatic text analysis and information extraction applications.

4 Approach

Our automated lexicalization approach is to compare the terminological scopes of candidate mappings, exploiting the

[1] http://wordnetweb.princeton.edu/perl/webwn

graph natures of both the lexicon and the target ontology. We have a preference for high-recall, low-precision mapping suggestions, as accounting for false negatives is a much more costly operation in the review process than accounting for false positives, as will be shown in our results in Section 5.

4.1 Preparing Ontology Object Terminology

OWL ontology objects have unique identifiers which are expressed as Universal Resource Identifiers (URI) composed of a namespace and identifier. The identifier is very often a camel case string approximating an English expression with a meaning close to that of the named object. In addition, the Resource Description Framework Schema (RDFS) annotation 'label' is often used to provide a natural language expression whose meaning is close to that of the labeled object.

Buitelaar, et. al.[10] have noted that the use of a natural language label annotation is not the same as a true linguistic lexicalization of an ontology object, since a label alone cannot provide sense disambiguation, part of speech or other linguistic contextual information. For our purposes, a label annotation is a useful piece of evidence which will contribute to the generation of a correct mapping using the linguistically richer WordNet lexicon.

Camel case URIs are tokenized, as are terms whose individual words are separated with underscores and hyphens. These methods of creating multi-word terms are frequently seen in ontology naming standards and even in RDFS label annotations which should generally be more natural language compatible. Single- and two-word terms are extracted from the resulting tokenized names and labels using a very simple subphrase algorithm which ignores stop words, but otherwise generates all one- and two-word terms which it is possible to form from the larger term. These terms are stored together as representing direct match candidates for the ontology object. Terms are converted to lower case and stemmed to their singular forms. This collection of terms is called the object's terminological Scope 0.

4.1.1 Example

In the Wine Ontology, the class which is defined logically as the intersection of Dry Wines and Red Wines has the URI:

http://www.w3.org/TR/2003/PR-owl-guide-20031209/wine#DryRedWine

It has no RDFS label.

After de-camel-casing, the term 'dry red wine' is added as a candidate. Further, the terms 'dry', 'red' and 'wine' become candidates, as do the two-word terms 'dry red' and 'red wine'.

4.2 Preparing Synset Terminology

WordNet synsets are easier to prepare terminologically, since their member lemmas are already lower case and stemmed to singular form. WordNet lemmas do not use camel case, but multi-word lemmas are expressed with underscores rather than spaces, and hyphenation is occasionally present, both of which are treated as with ontology objects. We also extract single- and two-word terms from synset lemmas as for ontology object labels, as described in Section 4.1. This collection of terms is called the synset's terminological Scope 0.

4.2.1 Example

In WordNet, the synset for red wine contains only one lemma, 'red_wine.' This is converted to 'red wine'. Further, the terms 'red' and 'wine' become members of the scope.

4.3 Preparing Ontology Object Extended Terminological Scopes

We then create extended terminological scopes for each ontology object. We do this by following the ontology objects' ObjectProperties for n 'hops' and recording the URIs of ontology objects found at the other end of these hops. Since ObjectProperties are 'directional,' and do not always possess inverse properties, we follow links both pointing to and from the object at hand. Once these URIs are collected, we add their Scope 0 terms to the Scope n terms of the object at hand. We proceed this way to create Scope 1 and Scope 2 environments for each ontology object. Anonymous ontology objects such as Restrictions and Collections contribute no terms to a Scope, but are still considered as a 'hop' in the process. If duplicates are present between Scopes, these may be removed from the 'greater' or 'lesser' scope, or left as duplicates, depending on the parameterization of the lexicalization automation.

4.3.1 Example

In the Wine Ontology, the DryRedWine class is linked to an anonymous Collection. Thus, it has no Scope 1 terms. This anonymous Collection is linked to two classes: RedWine and DryWine. The Scope 2 terms for DryRedWine are 'red', 'dry', 'wine', 'red wine', and 'dry wine'.

4.4 Preparing Synset Extended Terminological Scopes

In much the same way, we then create 2 additional terminological Scopes for each synset. WordNet does not contain anonymous synsets, so each Scope contains some terms. WordNet is also complete in its use of inverse relations, so if there is a link R from synset A to synset B, there will nearly always be a link 'inverse(R)' back from synset B to synset A. We check to avoid duplication of synsets in multiple Scopes. Duplicate terms, however, may be present between Scopes, and as for ontology objects, these can be removed from the 'greater' or 'smaller' scope, or left as duplicates, depending on the parameterization of the lexicalization automation.

4.4.1 Example

The WordNet synset for red wine has Scope 1 terms: 'wine,' 'zinfandel,' 'sangria,' 'red bordeaux,' 'beaujolais,' 'merlot,' 'rioja,' 'sangaree,' 'chianti,' 'medoc,' 'cabernet sauvignon,' 'vino,' 'claret,' 'cabernet,' and 'pinot noir'.

Scope 2 terms for red wine are too numerous to list fully here, but they include: 'inebriant,' 'vermouth,' 'fortified wine,' 'cotes de provence,' 'mulled wine,' 'sparkling wine,' and 'alcoholic drink'.

As can be seen in the selected example, WordNet usually has a much more richly connected graph of senses and terms than a domain ontology. This difference held for each of our test ontologies.

4.5 Filtering Candidate Synsets

The full WordNet lexical database has over 140,000 synsets. In order to prune this to include only synsets which are likely to match ontology objects, we select all synsets which share a Scope 0 term with any of the target ontology objects, and all synsets which contribute to their extended terminological scopes, constructed as described in Section 4.4. The Wine Ontology has 74 ontology targets, which select 1714 synsets for consideration in mapping.

4.6 Comparing Ontology Objects and Synsets

All candidate pairs of ontology objects and synsets are scored according to the size of the intersections of their various terminological scopes. This score calculation is as follows:

$$\mathrm{score}(o, s) = \sum_{i=0}^{2} \sum_{j=0}^{2} w_{i,j} \left| \bigcap o_i, s_j \right|$$

Where o_i is the set of terms in the ontology object’s Scope i, s_j is the set of terms in the synset’s Scope j, and $w_{i,j}$ is a weight for matches between Scopes i and j.

4.6.1 Example

There are 126,836 potential mappings between the Wine Ontology objects and the set of filtered synsets as described in Section 4.5. The similarity score between the DryRedWine class in the Wine Ontology and the red wine synset in WordNet is 24 based on weighted intersections of their terminological scopes involving the terms ‘red’, ‘wine’ and ‘red wine’.

4.7 Reviewing Candidate Mappings

Once all pairs of ontology objects and synsets have been scored, the candidate mappings passing a given threshold are presented to a human user for review. Remembering that the intent of the automation was to assist this human process, and not to fully automate the mappings, the threshold and weights are set such that very many candidates may be presented, although ranked by their scores, so that the human reviewer may have access to as many sensible choices as possible.

4.7.1 Example

In addition to suggesting that the ‘red wine’ synset be mapped to the DryRedWine ontology object, synset mappings are proposed for: ‘wine’, the color ‘wine red’, the adjective ‘dry (as in liquor)’, and the adjective ‘dry (as in prohibition)’. There is no synset for the concept of ‘dry wine’.

5 Results

To test our approach, we hand-mapped a sample of ontology objects from our three test ontologies using an ontology editor and the online WordNet lookup function. This gave us a baseline of fully manual performance against which to compare the automation-assisted review process. A summary of these results is given in Table 1:

Metric	*Measurement*
Ontology Objects Mapped	75
Mappings per Object	1.6
Time to Map one Object	110 seconds
Total Estimated Time	59 hours

Table 1: Manual Mapping Measurements

Where ‘Total Estimated Time’ is for all mappable ontology objects from our three test ontologies, totaling 1932 such objects.

We then generated automated mapping suggestions for each of the test ontologies using the same parameterization of weights and thresholds. A human reviewer then selected good candidates from these suggested mappings, and, where a mapping was not suggested but the reviewer believed one existed in WordNet, the reviewer added this ‘missing’ mapping manually. The time to perform this review of suggested mappings was recorded to compare to the fully manual process. The resulting fully reviewed mappings were used as ‘ground truth’ from which to generate receiver-operator characteristics (ROC) against the unthresholded results of the automated runs. The results of this approach for each of the three test ontologies follow.

5.1 Wine Ontology Results

Table 2 presents the time spent generating the automated suggestions for the Wine Ontology, and reviewing the results, broken into accounting for bad suggestions (false positives) and missing suggestions (false negatives).

Metric	*Measurement*
Ontology Objects	275
Mappable Ontology Objects	74
Automated Mappings	1968
Time to Generate Mappings	6 seconds
Mappings Rejected by Review	1878
Mappings Added by Review	3
Mappings per Object	1.3
Time to Reject a Mapping	0.3 seconds
Time to Add a Mapping	50 seconds
Total Time to Create Mappings	13 minutes

Table 2: Review Metrics for Wine Ontology

Table 3 presents the ROC when the fully human-reviewed mappings were compared to the original fully automatic mappings with no thresholding.

Metric	*Measurement*
Precision	4.6%
Recall	96.8%
Accuracy	93%
Specificity	93%

Table 3: ROC for Wine Ontology

As mentioned earlier, a heavily recall-biased technique alleviates the most onerous human review task (dealing with a false negative), thus minimizing the total time spent generating mappings.

5.2 Pizza Ontology Results

The Pizza Ontology, although not used in the examples in this paper, was tested as well. Table 4 presents the human review metrics for it:

Metric	*Measurement*
Ontology Objects	331
Mappable Ontology Objects	99
Automated Mappings	4182
Time to Generate Mappings	8 seconds
Mappings Rejected by Review	4034
Mappings Added by Review	6
Mappings per Object	1.6
Time to Reject a Mapping	0.2 seconds
Time to Add a Mapping	50 seconds
Total Time to Create Mappings	18 minutes

Table 4: Review Metrics for Pizza Ontology

The version of the Pizza Ontology used had a large number of Portuguese labels in it which made it difficult to match those against the English WordNet successfully. This will be addressed in future work, but since the class names themselves were generally English camel case forms, overall performance was not greatly different than for the Wine Ontology.

Metric	*Measurement*
Precision	3.5%
Recall	96.1%
Accuracy	95%
Specificity	95%

Table 5: ROC Metrics for Pizza Ontology

5.3 Basic Formal Ontology (BFO) Results

The BFO that we used actually encompasses 15 OWL files, covering the foundational SNAP and SPAN upper ontologies of BFO, as well as several mid-level ontologies which are in wide use in the biotechnology field, but whose subject matter is not extremely domain-specific. We attempted to use the Ontology for Biomedical Investigations (OBI), since it is quite large, but it is very domain-specific in areas for which WordNet has little or no coverage. Table 6 presents the review metrics for BFO:

Metric	*Measurement*
Ontology Objects	2457
Mappable Ontology Objects	1759
Automated Mappings	47149
Time to Generate Mappings	1.9 minutes
Mappings Rejected by Review	44407
Mappings Added by Review	100
Mappings per Object	1.6
Time to Reject a Mapping	0.5 seconds
Time to Add a Mapping	50 seconds
Total Time to Create Mappings	8 hours

Table 6: Review Metrics for BFO

The BFO contains a large number of compositional classes such as ‘CitizenOfDemocraticYemenRole,’ which may map inexactly to the single WordNet synset for ‘Yemeni’ or more precisely to a combination of the synset for ‘citizen’ and the synset for ‘Yemen.’ In general, automated mapping included all three among its suggestions for classes such as this. However, this type of class requires additional effort in sifting through proposed mappings to obtain the correct combination, and hence a longer time spent per false positive. Still, the overall time savings over manual mapping is significant. Table 7 presents the ROC for the BFO.

Metric	*Measurement*
Precision	5.8%
Recall	96.5%
Accuracy	98%
Specificity	98%

Table 7: ROC for BFO

6 Conclusions and Future Work

Based on the results from the three test ontologies, the overall performance improvement over fully manual mapping comes out to an average of about 90 seconds saved per mappable ontology object. There are several areas of future work which may improve precision without significantly degrading recall.

6.1 Exploring Parameterizations

The weights given to terms found in Scope intersections were set using an ad-hoc system biasing toward near-scope matches. Additional weighting favoring different parts of speech, or multi-word terms may result in better ranked candidates, which would in turn allow a higher match threshold and result in fewer false positives which the human reviewer would have to process.

6.2 Mining Glosses

In addition to label annotations, ontology objects often have natural language 'comment' or 'definition' annotations describing the meaning of the object. Identifying useful terms within these glosses may improve the sparse nature of many ontology objects' terminological Scopes.

WordNet also provides glosses for synsets, and the eXtended WordNet[11] project replaces the keywords in those glosses with pointers to the synsets they represent. While the terminological Scopes of synsets are not typically sparse, additional synsets which are not explicitly linked via the fixed set of WordNet relations may add value in matching.

6.3 OntoWordNet

OntoWordNet[12] is an attempt to create a formal ontology encompassing WordNet lexical knowledge. Such a lexical ontology may make comparison to other domain-oriented ontologies easier, since their semantics would be more closely aligned. A danger exists in comparing ontologies with significantly different upper-level bases, however, as foundational assumptions may defeat easy mapping.

6.4 Accounting for Multilingual Terms

The RDFS label annotation contains an optional language attribute which was not used in the results shown in this paper, and for the Pizza Ontology, caused a number of mismatches which would otherwise have been avoided if language information were used. Likewise, WordNet lexica exist in many languages, often with equivalencies between synsets across languages which would extend mapping to non-English-based and/or multi-lingual ontologies.

7 Acknowledgements

This material is based in part upon work supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory contract number FA8650-10-C-7062. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL or the U.S. Government.

8 References

[1] G. A. Miller. "WordNet: A Lexical Database for English"; Communications of the ACM, Vol. 38 No. 11: 39—41, December, 1995.

[2] W3C. "OWL 2 Web Ontology Language

Document"; http://www.w3.org/TR/owl2-overview/, October, 2009.

[3] A. Stirtzinger, C. Anken. "Semi-automated ontology generation and evolution"; In Proceedings of the SPIE, Volume 7347, pp. 734706-734706-10, April, 2009.

[4] C. Anken, A. Stirtzinger and B. McQueary. "Goal-Driven Semi-Automated Generation of Semantic Models"; In Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defence Applications (CISDA), July, 2009.

[5] J. Gennari, M. Musen, R. Ferguson, W. Grosso, M. Crubézy, H. Eriksson, N. Noy and S. Tu. "The evolution of Protégé: an environment for knowledge-based systems development"; In International Journal of Human-Computer Studies, Volume 58, Issue 1, January, 2003.

[6] W3C. "Wine Ontology"; http://www.w3.org/TR/owl-guide/wine.rdf, December, 2003.

[7] N. Drummond, M. Horridge, R. Stevens, C. Wroe, and S. Sampaio. "Pizza Ontology"; http://www.co-ode.org/ontologies/pizza/2007/02/12/pizza.owl, February, 2007.

[8] A. Spear. "Ontology for the Twenty First Century: An Introduction with Recommendations"; Under Review. http://www.ifomis.org/bfo/documents/manual.pdf

[9] B. Smith, et. al. "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration"; In Nature Biotechnology Volume 25, pp. 1251-1255, November, 2007.

[10] P. Buitelaar, P. Cimiano, P. Haase, and M.Sintek. "Towards Linguistically Grounded Ontologies"; In The Semantic Web: Research and Applications, Springer Lecture Notes in Computer Science, Volume 5554, pp. 111-125, 2009.

[11] S. Harabagiu, G. Miller, and D. Moldovan. "WordNet 2 - A Morphologically and Semantically Enhanced Resource"; In Proceedings of the ACL SIGLEX Workshop, Standardizing Lexical Resources, pp. 1-8. 1999.

[12] A. Gangemi, R. Navigli, P. Velardi. "The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet"; In Proceedings of Cooperative Information Systems 2003. pp. 820-838.

Data Profiling Using Attribute Clustering

M. Heidi McClure
The University of Sheffield, and
Intelligent Software Solutions, Inc
5450 Tech Center Dr., Suite 400
Colorado Springs, CO 80919

Abstract— *Finding trends in database data is hard when presented with data sets containing many attributes (columns). The difficulty is increased when the data is in text fields and may include large summary or remarks fields. This paper discusses an approach that uses attribute level clustering in order to discover trends or profiles in the data. This is different from traditional uses of clustering in that each attribute is clustered separately and then the results are combined to define profiles. For example, in a case study of the Global Terrorism Database (GTD) data set, there are 98 columns (attributes) in the data. A profile might be defined by a particular group, attack type, weapon type and by specific information found in larger remarks-type fields. The profiles will show the values of these attributes along with all the records that matched that profile.*

Keywords: attribute-clustering, WebTAS, clustering, GTD, visualization, data-profiling

1. Introduction[1]

A requirement from a customer is to discover profiles of related objects in a database. The objects are from a table that may have many attributes and may have one-to-one or one-to-many joins to other tables. Data from attributes[2] from the top level table and all joined in tables may be considered for profiling. As noted in the abstract, finding trends in database data is hard when the data contains many columns and when that data includes large text fields like summaries, notes or remarks fields.[1] The solution presented in this paper brings together clustering algorithms and link analysis displays to discover profiles in the data.

The system on which to build this profiling discovery capability is the Web-enabled Temporal Analysis System (WebTAS) - a US government off-the-shelf tool used for data integration, visualization and analysis [2]. Attribute clustering allows for the discovery of profiles. Profiles help users (customers) make sense of their data.

This paper discusses the details of an attribute clustering implementation built on the WebTAS platform and it discusses areas for further research. It also presents a visualization using clustering in a link analysis chart display that includes visual clustering algorithms.

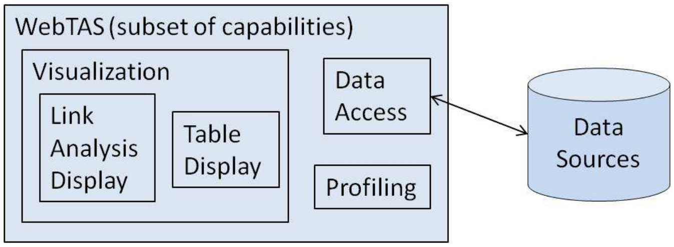


Fig. 1: General Architecture

2. Background

This section will describe the building blocks of the profiling system.

2.1 WebTAS

WebTAS accesses data from any traditional relational database like SQLServer, Oracle, etc and WebTAS access many other sources of data - such as file system data, live streams of data and web services. When accessing these other sources, a custom data source capability found in WebTAS is used. WebTAS's strength lies in its ability to visualize data from disparate sources on tables, graphs, timelines, grids and link analysis charts. For profiling, a custom data source has been added to WebTAS that allows a specific query to be performed that produces a result set and then attribute clustering to be performed on that set. Once profile results are available, link analysis is used to display the results. Link Analysis contains some visual clustering algorithms which further group like profiles together providing another level of clustering of the results.[2], [3]

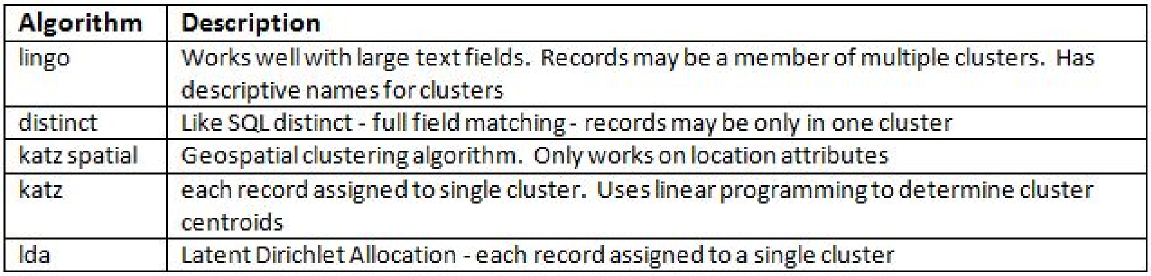
2.2 Attribute Clustering (Profiling)

Data mining (or text mining) clustering algorithms are usually applied to documents, bodies of text or other collections of text and provide results that group or cluster documents or records into buckets. Clustering of this type applies one

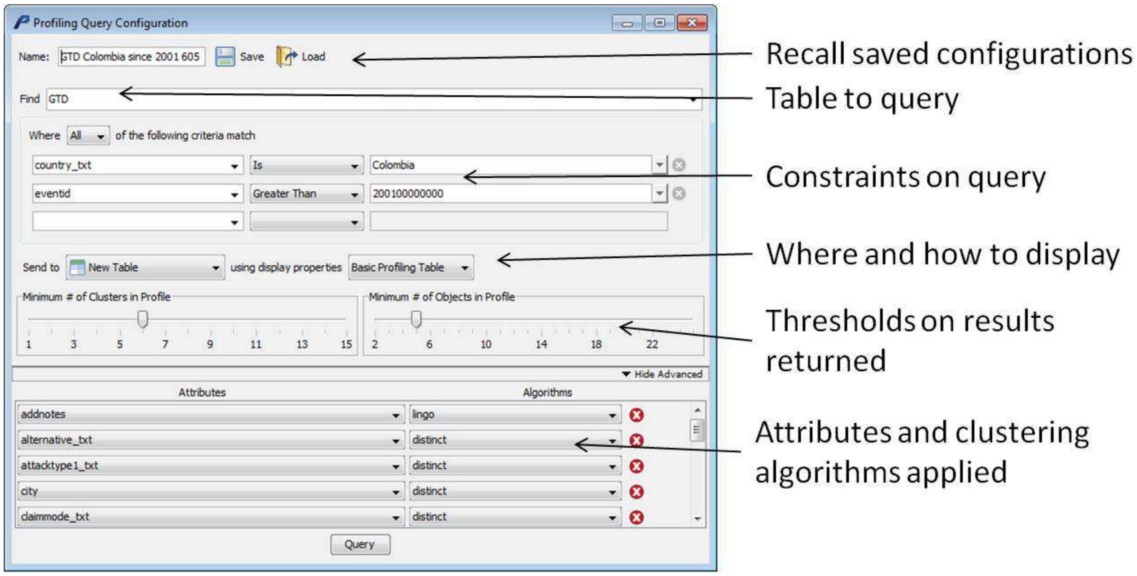
[1] Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

[2] In this paper, attributes, fields and columns are used interchangeably

Table 1: Clustering Algorithms Available



Algorithm	Description
lingo	Works well with large text fields. Records may be a member of multiple clusters. Has descriptive names for clusters
distinct	Like SQL distinct - full field matching - records may be only in one cluster
katz spatial	Geospatial clustering algorithm. Only works on location attributes
katz	each record assigned to single cluster. Uses linear programming to determine cluster centroids
lda	Latent Dirichlet Allocation - each record assigned to a single cluster



Name: GTD Colombia since 2001 605

Find GTD

Where All of the following criteria match

country_txt	Is	Colombia
eventid	Greater Than	200100000000

Send to New Table using display properties Basic Profiling Table

Minimum # of Clusters in Profile: 1 3 5 7 9 11 13 15

Minimum # of Objects in Profile: 2 6 10 14 18 22

Attributes	Algorithms
addnotes	lingo
alternative_txt	distinct
attacktype1_txt	distinct
city	distinct
claimmode_txt	distinct

Recall saved configurations

Table to query

Constraints on query

Where and how to display

Thresholds on results returned

Attributes and clustering algorithms applied

Fig. 2: Profiling Configuration UI

algorithm per pass to the set of data [4], [5]. Attribute clustering applies clustering algorithms to columns or attributes of data, usually pulled from databases. The clustering algorithms applied may be customized to best suit the type of data being clustered. To form profiles, records are grouped based on membership in the same attribute level clusters.

Attribute clustering seeks to discover profiles[3]. Clustering groups similar objects or records into groups or buckets. While an exact definition of clustering is not available, a range of clustering techniques may be found in Xu and Wuncsh's survey of clustering algorithms[5]. As described there, clustering may place records into one and only one group or clustering may place records into multiple buckets.

When data consists of many attributes (that is, many features or columns), one clustering algorithm may be more appropriate than another for a specific attribute. For example, when data consists of short pick-list driven data, a simple grouping algorithm like an SQL distinct function or group by call [6] may be all that is needed. If an attribute is a geographical coordinate, a geo-clustering algorithm is appropriate which will group objects based on how close they are physically to each other. For large text fields, an algorithm like lingo [7] that allows a record to be a member in one or more clusters is appropriate. Numerical data requires yet another clustering algorithm [5].

Attribute clustering is a way to cluster each attribute separately using specialized clustering algorithms and then to bring the results together based on membership in same attribute clusters. In the approach presented in this paper, a minimum number of matching clusters is selected along with a minimum number of matching objects in the profile.

2.2.1 Clustering Algorithms

The clustering algorithms used in this case study are briefly described in Table 1. For more information on their specifics, please see [8], [5], [4].

[3] the verb profiling is used interchangeably with attribute clustering in this paper

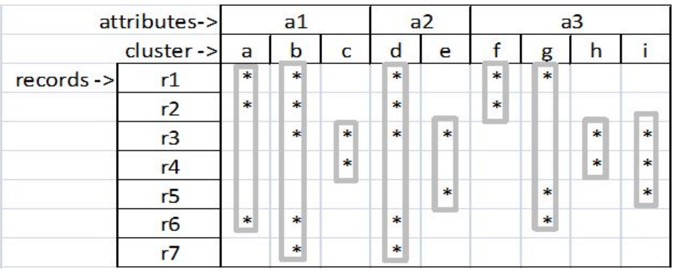
2.2.2 Profiling Configuration Dialog

Using the WebTAS infrastructure, a custom data source is used for profiling. This customer data source uses a very large and complicated query string - similar to a large SQL statement, but customized for use with WebTAS and with profiling. In order to hide the ugliness of the large query, a user interface has been created to allow easier selection of profiling criteria. See Figure 2.

2.2.3 Example of Attribute Clustering

As a simple example, consider seven records that are sent to attribute clustering. Attribute one (a1) generates three clusters (a, b, c); Attribute two (a2) generates two clusters (d, e); Attribute three (a3) generates four clusters (f, g, h, i). The gray boxes around the results is to highlight the records in those clusters. '*' character indicates the record fell into the specified cluster for the specified attribute. See Table 2.

Table 2: Clusters Found by Attribute



	attributes->	a1			a2		a3			
	cluster ->	a	b	c	d	e	f	g	h	i
records ->	r1	*	*		*		*	*		
	r2	*	*		*		*			
	r3		*	*	*	*			*	*
	r4			*					*	*
	r5					*		*		*
	r6	*	*		*			*		
	r7		*		*					

Examining pairs of records, we find the following and could conclude that r1-r2 and r1-r6 pairs are strongly related. See Table 3.

By looking at the records that match each set of matching clusters, we find that clusters b and d together are found to match five records. This, in addition to the "a,b,d,f" and the "a,b,d,g" cluster sets, form the highest strength clusters shown highlighted. See Table 4.

2.3 Visualization

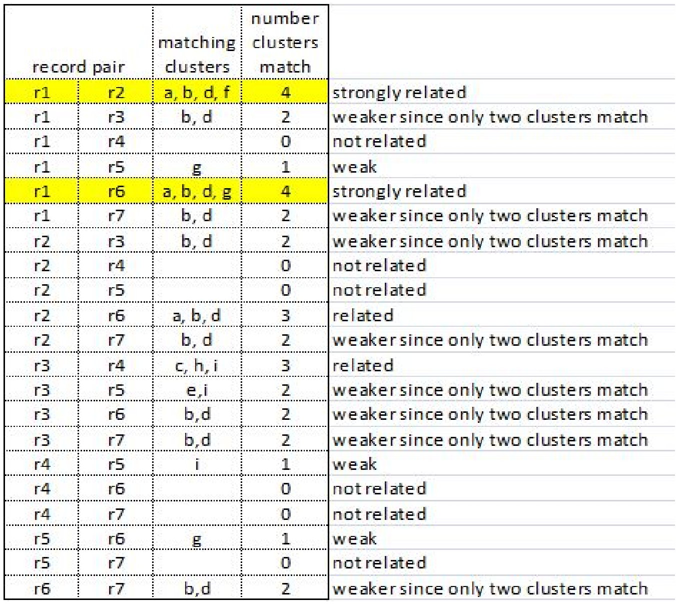
Link analysis of the profiles and their member objects presents a good visualization of how objects are related. WebTAS link analysis has four visual clustering algorithms[2]

- Springs and Repulsion (nodes repel, links attract)
- Clustering - Self Organizing (Fruchterman-Rheingold Algorithm)
- Filling ISOM (Inverted Self Organizing Map)
- Balanced (Kamada-Kawai Algorithm)

The Clustering - Self Organizing visual clustering algorithm works best for visualizing profiles. See Figures 3 and 4.

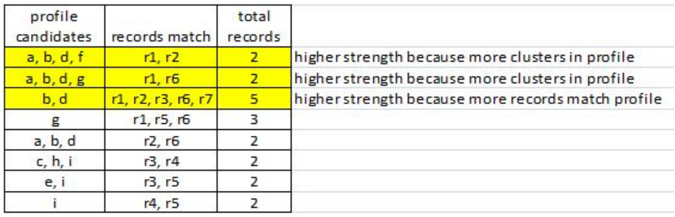
Details of the profiling results may also be displayed to a table. See Figure 5.

Table 3: Record Pairs



record pair		matching clusters	number clusters match	
r1	r2	a, b, d, f	4	strongly related
r1	r3	b, d	2	weaker since only two clusters match
r1	r4		0	not related
r1	r5	g	1	weak
r1	r6	a, b, d, g	4	strongly related
r1	r7	b, d	2	weaker since only two clusters match
r2	r3	b, d	2	weaker since only two clusters match
r2	r4		0	not related
r2	r5		0	not related
r2	r6	a, b, d	3	related
r2	r7	b, d	2	weaker since only two clusters match
r3	r4	c, h, i	3	related
r3	r5	e,i	2	weaker since only two clusters match
r3	r6	b,d	2	weaker since only two clusters match
r3	r7	b,d	2	weaker since only two clusters match
r4	r5	i	1	weak
r4	r6		0	not related
r4	r7		0	not related
r5	r6	g	1	weak
r5	r7		0	not related
r6	r7	b,d	2	weaker since only two clusters match

Table 4: Discovered Data Profiles

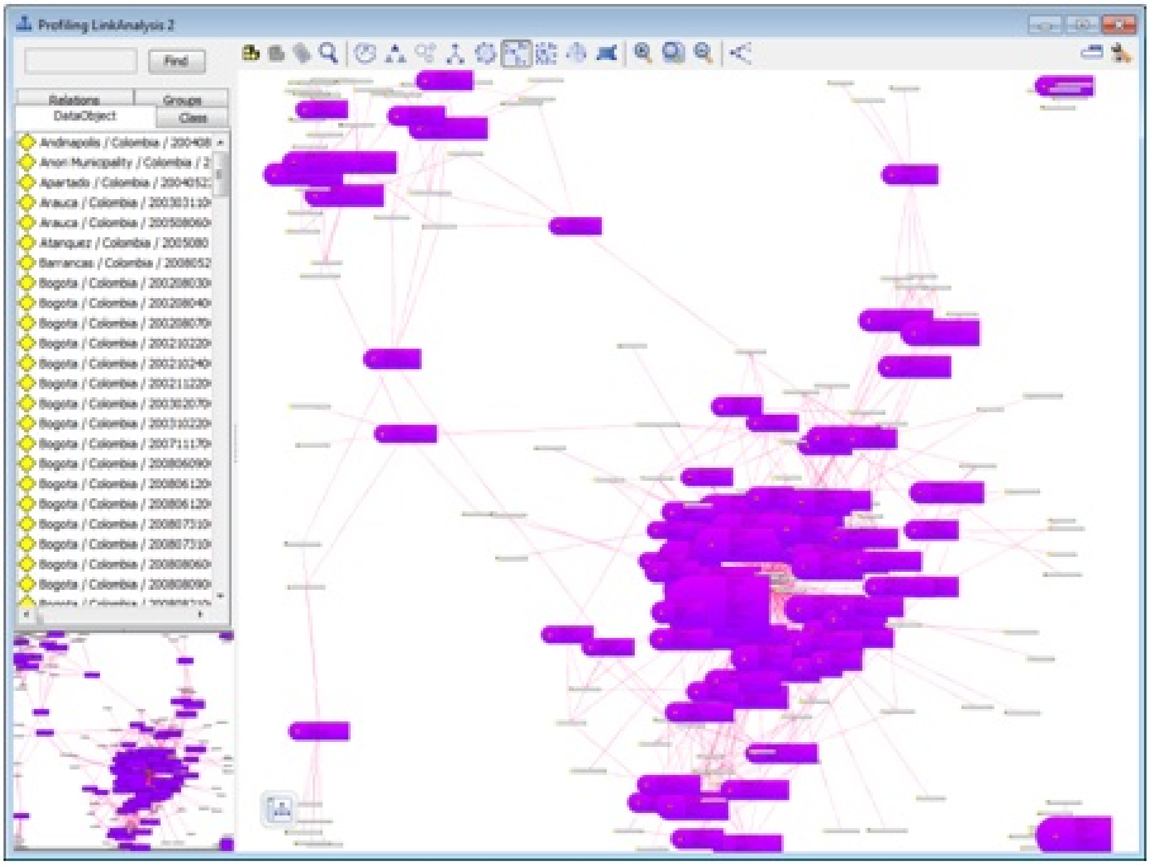


profile candidates	records match	total records	
a, b, d, f	r1, r2	2	higher strength because more clusters in profile
a, b, d, g	r1, r6	2	higher strength because more clusters in profile
b, d	r1, r2, r3, r6, r7	5	higher strength because more records match profile
g	r1, r5, r6	3	
a, b, d	r2, r6	2	
c, h, i	r3, r4	2	
e, i	r3, r5	2	
i	r4, r5	2	

3. Results

Tests have been performed on the Global Terrorism Database (GTD)[9]. The case study presented here is for the country of Colombia from the year 2001 thru 2008 (includes approximately 600 records). Seventeen (17) attributes are used for profiling the data - the attributes were chosen if they contained data in the records selected. The limits of a minimum of 6 clusters and 6 records are used since they presented a manageable number of profile results (approx 150) for the link analysis display. The goal is to see if data profiling can discover knowledge in the data not easily found in other ways.

A sample of the data sent to a table may be seen in Figure 5. The Profile Summary field shows attribute names and their values for all attributes that matched. The second column shows all the records or data objects that matched the data profile discovered. They are links to the details of the records. The summary information for the GTD records include city, country, event id and province or state.



Andinapolis / Colombia / 200408
Anori Municipality / Colombia / 2
Apartado / Colombia / 2004052
Arauca / Colombia / 2003031
Arauca / Colombia / 2005080
Atanquez / Colombia / 2005080
Barrancas / Colombia / 2008052
Bogota / Colombia / 2002080
Bogota / Colombia / 2002080
Bogota / Colombia / 2002080
Bogota / Colombia / 2002102
Bogota / Colombia / 2002102
Bogota / Colombia / 2002112
Bogota / Colombia / 2003020
Bogota / Colombia / 2003102
Bogota / Colombia / 2007111
Bogota / Colombia / 2008060
Bogota / Colombia / 2008061
Bogota / Colombia / 2008061
Bogota / Colombia / 2008073
Bogota / Colombia / 2008073
Bogota / Colombia / 2008080
Bogota / Colombia / 2008080

Fig. 3: Clustered Link Analysis Display

Initial display to link analysis may be seen in Figure 6. (The purple or dark icons are the profiles, the light gray are the member records.) By default, data is displayed to the link analysis charts using a radial tree layout. Link analysis shows objects and their related objects - for example, sending profile results to a link chart will show the profile objects (purple or dark) in the first ring of objects and linked to each profile object in the 2nd radial row are the report objects (gray) that are in that profile. Using the radial tree layout, you see that there are some report objects that are members of many profile objects.

Performing Fruchterman-Rheingold Algorithm for visual, self organizing clustering is shown in Figure 3. The Fruchterman-Rheingold algorithm is an example of a force directed layout algorithm where nodes repel and attract - the result is that the profiles and their member objects cluster visually based on how related the profiles are and which objects are members of the profiles.

Notice the separated purple (dark) nodes - these are profiles that don't share many objects with the nodes in the large grouping in the center of the chart. They describe profiles which have distinct characteristics which are not shared by other profiles in the link chart display.

The details of the profile in the upper left of the link chart may be seen in Figure 4. The profile has grouped records for bombing/explosion attack types by the National Liberation Army of Colombia (ELN) that reported injuries and was targeting a business.

Table 5 shows descriptions of some of the other profiles discovered using attribute clustering.

4. Conclusion

The discovered profiles enhance understanding of the data but also allow the customer to categorize new data and more quickly know if the new data matches a pattern which has been seen before. They may also use data in the discovered profiles to know how events are related. Discovered profiles may be examined to know how to prevent the same events from happening again.

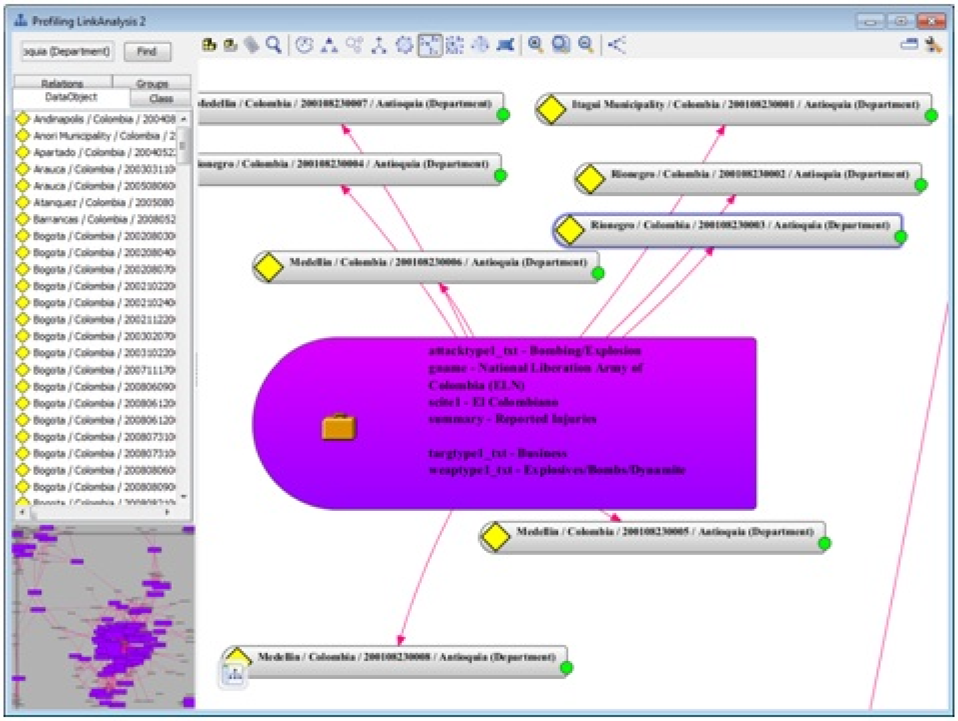
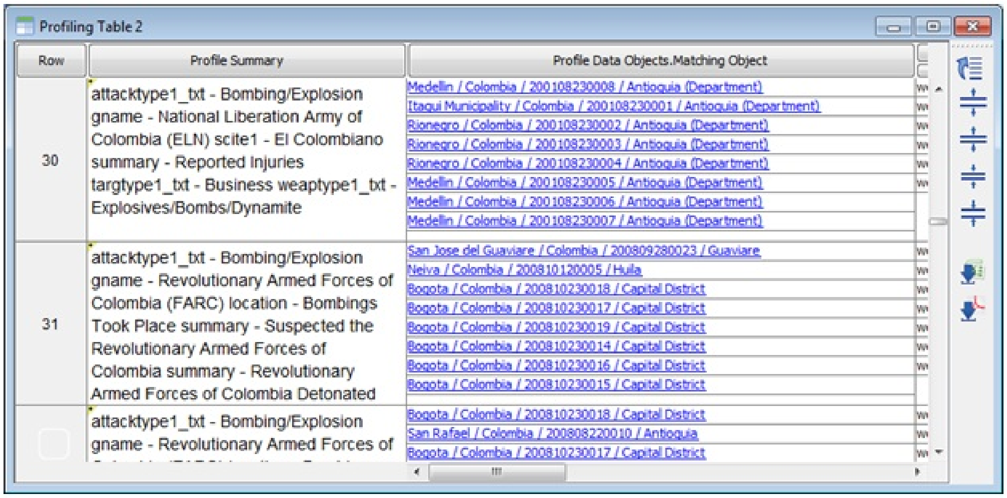
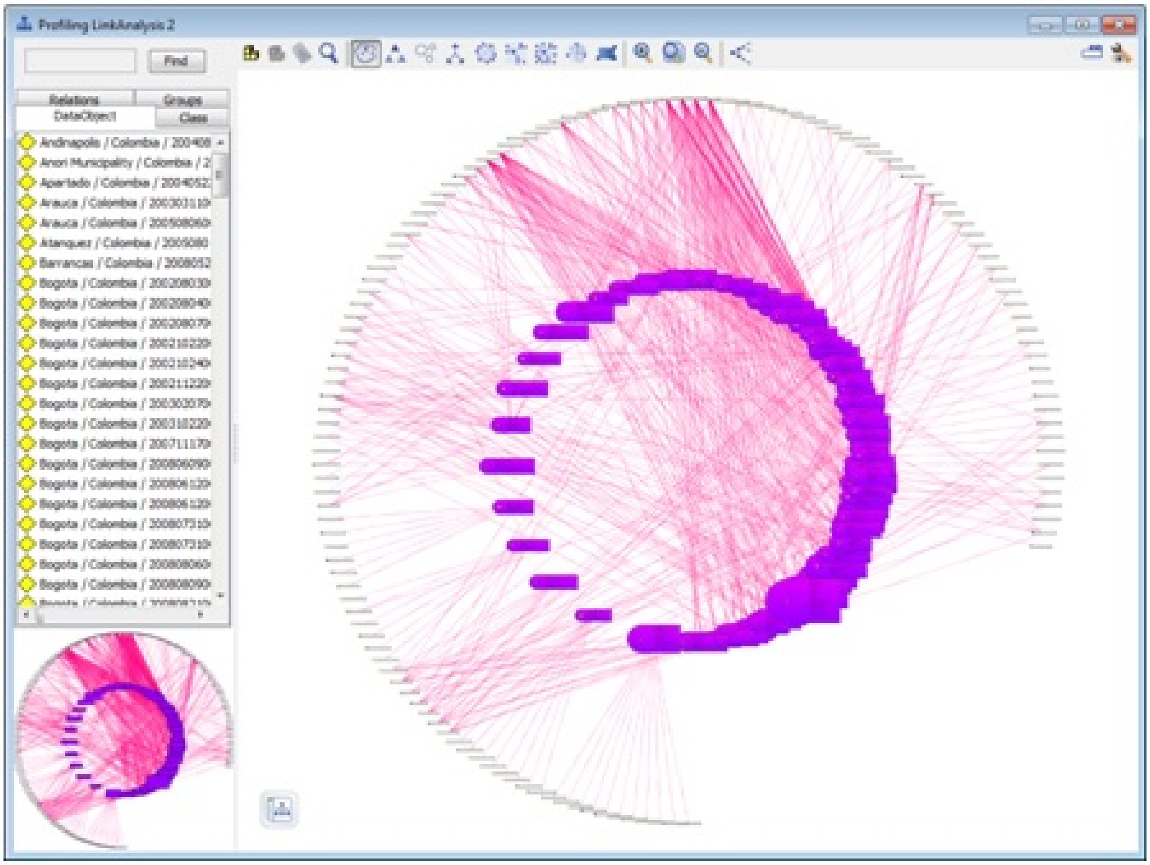

Profiling LinkAnalysis 2
oquia (Department)
Find
Relations
Groups
DataObject
Class
Andinapolis / Colombia / 200408
Anori Municipality / Colombia / 2
Apartado / Colombia / 2004052
Arauca / Colombia / 2003031
Arauca / Colombia / 2005080
Atanquez / Colombia / 200508
Barrancas / Colombia / 200805
Bogota / Colombia / 2002080
Medellin / Colombia / 200108230007 / Antioquia (Department)
Itagui Municipality / Colombia / 200108230001 / Antioquia (Department)
Rionegro / Colombia / 200108230004 / Antioquia (Department)
Rionegro / Colombia / 200108230002 / Antioquia (Department)
Rionegro / Colombia / 200108230003 / Antioquia (Department)
Medellin / Colombia / 200108230006 / Antioquia (Department)
attacktype1_txt - Bombing/Explosion
gname - National Liberation Army of
Colombia (ELN)
scite1 - El Colombiano
summary - Reported Injuries
targtype1_txt - Business
weaptype1_txt - Explosives/Bombs/Dynamite
Medellin / Colombia / 200108230005 / Antioquia (Department)
Medellin / Colombia / 200108230008 / Antioquia (Department)

Fig. 4: Zoomed In Link Analysis


Profiling Table 2

Row	Profile Summary	Profile Data Objects.Matching Object
30	attacktype1_txt - Bombing/Explosion gname - National Liberation Army of Colombia (ELN) scite1 - El Colombiano summary - Reported Injuries targtype1_txt - Business weaptype1_txt - Explosives/Bombs/Dynamite	Medellin / Colombia / 200108230008 / Antioquia (Department) Itagui Municipality / Colombia / 200108230001 / Antioquia (Department) Rionegro / Colombia / 200108230002 / Antioquia (Department) Rionegro / Colombia / 200108230003 / Antioquia (Department) Rionegro / Colombia / 200108230004 / Antioquia (Department) Medellin / Colombia / 200108230005 / Antioquia (Department) Medellin / Colombia / 200108230006 / Antioquia (Department) Medellin / Colombia / 200108230007 / Antioquia (Department)
31	attacktype1_txt - Bombing/Explosion gname - Revolutionary Armed Forces of Colombia (FARC) location - Bombings Took Place summary - Suspected the Revolutionary Armed Forces of Colombia summary - Revolutionary Armed Forces of Colombia Detonated	San Jose del Guaviare / Colombia / 200809280023 / Guaviare Neiva / Colombia / 200810120005 / Huila Bogota / Colombia / 200810230018 / Capital District Bogota / Colombia / 200810230017 / Capital District Bogota / Colombia / 200810230019 / Capital District Bogota / Colombia / 200810230014 / Capital District Bogota / Colombia / 200810230016 / Capital District Bogota / Colombia / 200810230015 / Capital District
	attacktype1_txt - Bombing/Explosion gname - Revolutionary Armed Forces of	Bogota / Colombia / 200810230018 / Capital District San Rafael / Colombia / 200808220010 / Antioquia Bogota / Colombia / 200810230017 / Capital District

Fig. 5: Table Display of Profiles



Andinapolis / Colombia / 200408
Anori Municipality / Colombia / 2
Apartado / Colombia / 2004052
Arauca / Colombia / 2003031
Arauca / Colombia / 2005080
Atanquez / Colombia / 2005080
Barrancas / Colombia / 2008052
Bogota / Colombia / 2002080
Bogota / Colombia / 2002080
Bogota / Colombia / 2002080
Bogota / Colombia / 2002102
Bogota / Colombia / 2002102
Bogota / Colombia / 2002112
Bogota / Colombia / 2003020
Bogota / Colombia / 2003102
Bogota / Colombia / 2007111
Bogota / Colombia / 2008060
Bogota / Colombia / 2008061
Bogota / Colombia / 2008061
Bogota / Colombia / 2008073
Bogota / Colombia / 2008073
Bogota / Colombia / 2008080
Bogota / Colombia / 2008080

Fig. 6: Link Analysis Circular Display

Table 5: Details of Link Analysis Chart

Where in link chart	Details of profile
Upper left	attacktype1_txt - Bombing/Explosion gname - National Liberation Army of Colombia (ELN) scite1 - El Colombiano summary - Reported Injuries targtype1_txt - Business weaptype1_txt - Explosives/Bombs/Dynamite
Upper right	attacktype1_txt - Armed Assault gname - Revolutionary Armed Forces of Colombia (FARC) summary - Killed by the Revolutionary summary - Suspected the Revolutionary Armed Forces of Colombia summary - Members of the Revolutionary Armed Forces weaptype1_txt - Firearms
Lower left	addnotes - Marked the 40th Anniversary of its Founding addnotes - Founding addnotes - Police attacktype1_txt - Bombing/Explosion gname - Revolutionary Armed Forces of Colombia (FARC) weaptype1_txt - Explosives/Bombs/Dynamite
Lower right	attacktype1_txt - Bombing/Explosion city - Puerto Colon corp1 - Colombian Petroleum Enterprise (Ecopetrol) gname - Revolutionary Armed Forces of Colombia (FARC) location - Bombings Took Place location - Villages of Puerto Colon and San Miguel scite1 - El Tiempo scite1 - January 2 summary - Villages of Puerto Colon and San Miguel summary - Fuerzas Armadas Revolucionarias de Colombia FARC Guerrillas targtype1_txt - Utilities weaptype1_txt - Explosives/Bombs/Dynamite

Although the results shown in this paper are from the GTD, there is nothing preventing profiling from being run on any data that WebTAS can see. As an additional experiment, profiling was run on a table where similar or even duplicate records exist. Profiling was able to identify and link these related records.

5. Future

5.1 Classification (Categorization)

Once interesting profiles are discovered, these related records may be used to train a classifier (categorization in WebTAS). Then when records are inspected by the system, they may be classified into buckets based on the profiles trained. [4]

5.2 Entity Extraction

Entity extraction in this context is the process of pulling entity information out of text. Term extraction may be a better way to think of this kind of entity extraction [8]. An enhancement is to perform entity extraction on data and then apply attribute clustering on the extracted entities. (this has been prototyped, but refinement to the entity extraction grammars has yet to be done.)

5.3 SEER

Once characteristics of a profile are discovered, these characteristics may be incorporated into a WebTAS SEER model for detection of new records matching defined profiles. Situation Exploitation Engine Real-time (SEER) is a component of WebTAS which allows detection of patterns in data both on historical and in a near real-time manner.[3]

Acknowledgments

Work described in this paper was funded at various times by Air Force Research Laboratory (AFRL) at Rome, NY (AFRL Contract FA8750-06-D-0005-0012) and the Defense Advanced Research Projects Agency (DARPA). Some of the work came from the Personalized Assistant that Learns (PAL) program, a DARPA program that is currently in military transition.

The author thanks her advisor, Dr. Mark Stevenson, of The University of Sheffield for being a sounding board during the development of this solution. The author also thanks her brother, Dr. John McClure, for his helpful reviews of this paper and Intelligent Software Solutions, her employer, for presenting such an interesting problem to solve.

The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

[1] S. Džeroski, "Multi-relational data mining: an introduction," *SIGKDD Explor. Newsl.*, vol. 5, pp. 1–16, July 2003. [Online]. Available: http://doi.acm.org/10.1145/959242.959245

[2] (2011) Webtas overview. Intelligent Software Solutions. Intelligent Software Solutions - http://www.issinc.com/solutions/webtas-overview.html. [Online]. Available: http://www.issinc.com/solutions/webtas-overview.html

[3] M. Gerken, R. Pavlik, C. Houghton, K. Daly, and L. Jesse, "Situation awareness using heterogeneous models," in *Collaborative Technologies and Systems (CTS), 2010 International Symposium on*, may 2010, pp. 563 – 572.

[4] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Burlington, MA: Morgan Kaufmann, 2011.

[5] R. Xu and D. W. II, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, pp. 645–678, 2005.

[6] K. Kline, D. Kline, and B. Hunt, *SQL in a Nutshell*, 3rd ed. O'Reilly Media, Inc., 2008.

[7] S. Osinski and D. Weiss, "Conceptual clustering using lingo algorithm: Evaluation on open directory project data," in *In IIPWM04*, 2004, pp. 369–377.

[8] H. Marmanis and D. Babenko, *Algorithms of the Intelligent Web*. Greenwich, CT: Manning Publications Co., 2009.

[9] Global Terrorism Database, START, accessed on 9 December 2010. [Online]. Available: http://www.start.umd.edu/gtd/

SESSION

MINING OF DATA RICH SOURCES

Chair(s)

Prof. Ray Hashemi

An E-Health Model: A Technical and Economic Perspective

Hassan Makhmali, , Simeon Yates, Babak Akhgar , Naser Aniba, Mazen Qeitishat, and Richard Wilson
C3RI, Sheffield Hallam University, Sheffield, United Kingdom

Abstract - E-health in the developed countries has increased significantly in recent years, and yet these healthcare services are rapidly reaching a point of inflection. The rise in health expenditure, the burden of the ageing population, and the growing expectations of citizens are all contributing towards large-scale restructuring in the 'way' that healthcare is provided and supported via use of ICT. Our proposed conceptual economic e-health model focuses not only in technical but also economic areas. Furthermore, we have attempted to show some economic perspectives of e-health, in particular, related to alternative markets. Competition in an ideal and is engaged with many non-markets problems, however, quasi competition market is more efficient than monopolist governments and therefore we think that the main responsibility of government is planning and controlling continuously not engaging directly with e-health deployment scenarios.

Keywords: E-health, Market, Monopolies, Models

1 Introduction

E-health is an extremely broad topic, encompassing the many different domains of information and communication 'technologies' (ICT) responsible for supporting many aspects of healthcare provision. Adoption of such technologies, in the developed countries, has increased significantly in recent years, and yet these healthcare services are rapidly reaching a point of inflection. The rise in health expenditure, the burden of the ageing population, and the growing expectations of citizens are all contributing towards large-scale restructuring in the 'way' that healthcare is provided and supported via use of ICT [1]. Whilst previous adoption of ICT in healthcare has been largely in the implementation of distributed systems, it is now apparent that a more holistic approach to 'e-health strategies' is required (both locally and on national levels) in order to move toward a new successful model of healthcare [2]. Moreover, governments and healthcare providers around the world are considering patient satisfaction and healthcare costs. Arguably, economics within micro and macro environment are controversial in proving or disproving the mechanism to maximising, minimising and optimising healthcare services.

2 Global perspective of e-health

According to the World Health Organisation (WHO), e-health relates to three important functionalities in supporting three main areas of healthcare services. These are digital data transmission, data storage and data retrieval, to support clinical, educational, and administrative purposes [3].

In order to explore the area of e-health fully and to determine how these 'strategies' can be developed, it is essential to understand what exactly e-health is? Why do we need e-health? and what do we hope to achieve by implementing it? In a focused review in 2001, Gunther Eysenbach described the 10 e's of e-Health, which are aimed at providing some resolution to the question 'why do we need e-Health?' In a more focused review, Gunther Eysenbach in 2001 has noted ten of the most important advantages of e's in e-Health in response to the question 'why eHealth?' these are:

1. 'Efficiency';
2. Enhancing quality of care;
3. Evidence based;
4.Empowerment of consumers and patients;
5. Encouragement of a new relationship between the patient and health professional (Information Society);
6. E-Learning;
7. Enabling information exchange and communication in a standardized way between health care establishments;
8. Extending the scope of health care beyond;
9. Ethics; and
10.Equity. [4]

In 2005, the European Commission described e-health as "the use of modern ICT to meet the needs of citizens, patients, healthcare professionals, healthcare providers, as well as policy makers" [5]. At the same time, Bruno Salgues published a paper describing e-health as an umbrella term that encompasses all of the ICT domains, commonly associated with health informatics, and extends these by incorporating modern tele-services, medical virtual learning, and medical science applications [6].

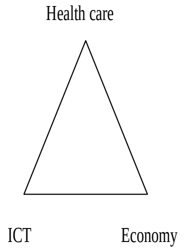


Figure 1- triangle figure as a basic efficient model.

In the context of the multidimensional of e-health, as shown figure 1, it is noted that, that there are three main components. In the proceeding sections, we will explore healthcare services in the context of ICT and economy.

3 ICT Infrastructure related to e-health

Over the past couple of decades using ICT has increased dramatically. Currently, rapid development in several ICT sectors is becoming very notable [7]. Telecommunications, mobile and wireless service, internet technologies and software developments become more attractive to many governments, investors and customers in developing countries. Therefore, many governments especially in developed countries have used ICT applications to enhance efficiency, decrease cost, prevent time wasting and decentralisation [7]. One of the most important sub-domains of e-government is e-health [8]. Indeed, ICT in this context is divided in to two sectors, hardware and software. Hardware is needed for the development of the infrastructure, initially for connectivity, for remote areas. For example, landline, mobile, computer, installing antenna or satellite and developing communication cable or light fibre to achieve wideband and high speed in the Internet connection. Moreover, in the software sector this could be addressed by using portals and websites. Portals are the main sectors of 'e-paths' or 'Networks' in order to provide services such e-health services in healthcare system. They can enable each user to access initial health information. In addition, websites like portals could be addressed with some services to support patients within internet engine research. Portals and websites can store and share a wide range of information services, through access of health information based on WHO's definition [3]. In addition, some portal's advantages could be offered within research and e-learning facilities, Information society utilities, accessing business offerings and the other individual productivity applications [9].

At a glance, the e-health domain covers a wide range of applications. For example, some applications include deploying Management Information Systems (MIS) in clinics or hospitals, a national health monitoring system, computerisation by the evidence-based of primary care services. Further application include, the linking of ministry of health and healthcare insurance schemes, issuing smart cards for patients, making electronic appointments and health records, standardisation of health terminology, categorising and codify of diagnoses and accessing mortality databases to determine a cause of death and medical production are considered [10]. Furthermore, some technical facilities are based on remote access such as telemedicine, internet and robotics [11].

In addition to this, developed countries such as North America and Europe are also beginning to develop health information platforms and infrastructures as part of their national strategies (e.g. UK NHS ICT Platforms). Today, their ideas are emerging as a unique approach toward building more efficient and effective healthcare services based on the Internet. Also electronic health records are playing an important role in improving clinical practice, hospitals, research, and policy and service management [12].

One of the most important premises of e-health is forming an information society. Information society could be formed through combination of professional human resources based on ICT infrastructures as technical parts.

4 Economic perspective of e-health

According to Samuelson, economics is "*the study of how individuals and societies choose to employ scarce resources that could have alternative uses in order to produce various commodities and to distribute them for consumption, now or in the future, among various persons and groups in society*"[13]. Briefly, this definition notes that economics is a scientific way of optimising the use of scarce resources..Decreasing the costs in health services, for healthcare providers, and increasing the demand for patients will be considered as an essential value. Indeed, governments should investment in some e-health projects which will have economical evaluation components such as opportunity cost, marginal analysis, time preference and economic efficiency. E-health economic evaluation means the methodical way to assess resources and whether they are used or allocated efficiently within an explicit criterion. At a glance, microeconomic covers firm or organisation's situation and markets, which they involve to give e-health services in order to describe minimising cost or maximising benefits based on market's type. Furthermore, macroeconomics are about government policies with regards to how the government obtains tax and how it will be distributed. This could be distributed as an investment, expenditure or subsidies to increase welfare or reduce poverty [14]. There are two important factors in the e-health economic context these can affect many economic variables such as government and market. Additionally, they have two crucial and famous mechanisms in order to action in the real world such as cost and price factors [15].

In addition, to Williamson (1981), one of the most basic approaches in all organisations is transaction cost. He has noted economising is compatible by allocating transactions to government structures within an internal organisational environment. Moreover, it has needed to divide some application details and recognising government structure. Williamson believes that this approach can determine domains of efficiency from organisation to market and it could be applicable for inside transaction in order to plan employment details [16].Moreover, Mahoney has noted about behavioural approach and he believes "resource learning" theory. In addition, resources are categorised within financial, human, physical, technical and organisational capabilities as a capital [17]. As it can be seen in e-health definitions, it can cover these elements as a multidimensional

5 Theoretical analysis of market mechanism

Modern technology such ICT beside health care services can shift demand and supply's curves upwards because of it can increases efficiency. For the reason, costs of providing e-health services are decreased significantly. It means that customers and suppliers can consume and produce more goods and services in the same price and cost [18]. This figure shows the shifting demand and supply curves of e-health services because of increasing technology as below.

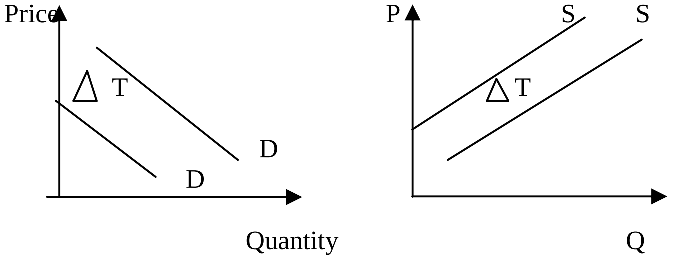


Figure 2 - Increasing Technology and shifting Demand and Supply Curves

However, many governments make huge investments in order to facilitate the move from traditional to modern healthcare (e-health) in non-profit organisations. Nevertheless, this situation looks like a governmental monopoly, and a monopoly in goods and services is not efficient [16]. Since, governments are not efficient in allocating goods and services to people such inefficiencies are the main factors in reducing welfare and satisfying of consumers [19]. Thus, they have to release there duties to private sector within price and market mechanism, although some goods and services are necessary and urgently needed by many people such as food and healthcare services.

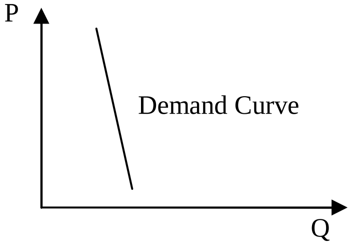


Figure 3 - inelastic demand for necessity goods and services

The figure above shows, that in the context of health services, increasing the demand for a health service adds to the price of that healthcare service. Adam Smith believes an “Invisible Hand” is in the market in order to equal demand and supply automatically [20]. Moreover, it seems that this amazing hand should be in ideal market within an automatic mechanism is well suited in perfect competition market. According to economics theory, a perfect competition market has some classic assumptions that can be efficient in minimising price for consumers. These assumptions are; “the agents have no market power, they are price takers, competition market will drive the market price down to the competitive level (equal to the marginal costs) and consumer’s welfare.

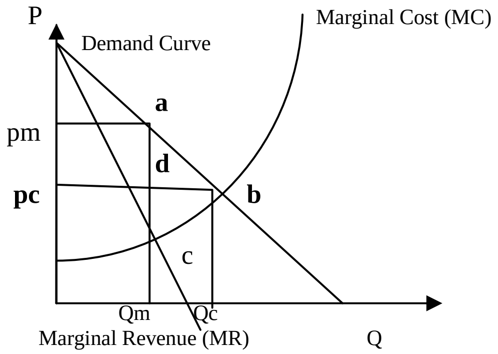


Figure 4 - presenting extreme price in perfect competition and perfect monopoly market

The figure above shows, perfect competition price is taken from cross marginal cost and demand curve and it is lower than a perfect monopoly. Meanwhile, monopolist determine price from cross marginal revenue and marginal cost. Therefore, high pricing in a monopoly market is decreased social welfare exponentially. The determine area under the demand curve in figure 4, above (A,B,C,D), states the deadweight, social gain and proves inefficiency in a perfect monopoly market [21][22].

6 Conceptual designing for private e-health model

According to the three sides of the triangle literature review, as introduced above, it is understood that e-health services could be given on the internet. Moreover, modern facility such as ICT is well-suited platform to support e-health advantages within strategic planning. In addition, economic knowledge was addressed new technology and price mechanism through competition market could be provided more satisfying for consumers or patients. Healthcare and medication are essential for patients and they do not have any considerable substitute goods and services. Therefore, we providing the following comments to design an e-health private model:

➢ Advantages of e-health in private section within competition market

- High market efficiency to decrease price and cost.
- Lower price than monopoly as it is mentioned in figure 4.
- Queue less and saving time because patient satisfaction will be increased.
- More social benefit than social cost due to the reductions in price for patients.
- Increased patient satisfaction brought about by low prices and improvement in services.
- High Quality because of changing technology and the re-distribution of e-health.

➢ Disadvantages of privatise e-health within ICT companies

- Concerns for the security of patients records
- Centralised companies and inequity. Many people do not have computer and internet and they are unable to gain access to computers.
- Office corruptions, which is conterevertial between public and private sections.
- Weakness in government control as they remain responsible for them.
- Creating trust

➢ Clearing some contexts for e-health services based on ICT Companies

- Making e-health consulting centres, some places that professionals can give consulting to patients.
- Building e-health database centres these could store health information in order to share and retrieve between stakeholders.
- E-learning centres regarding e-health, virtual centres can improve health based knowledge repositories.
- E-doctor centres, health services within intelligence algorithm.
- Invites to SMEs (Small and Medium Enterprise) to carryout e-health projects
- Encouraging Public Private Partnership as successful way to accept heavy government responsibilities.
- Supporting co-operative company as a kind of private organisation to do some of the activities.
- Supporting NGOs (Non Government Organisation), as a charity, in providing e-health services.

➢ Offering some solutions for promoting ICT companies by government

- Tax exemption, which is an economic policy to support e-health providers.
- Buying guaranteed services which promote a way to create a secure path by predicting techniques.
- Anti-Trust law against monopoly events as a non-market.
- Inviting from NGOs to control failing market and support some disabled people.
- Changing culture by advertising, as an effective tool to state new technology and a way to deal with e-health services.
- Giving speed memory by ministry of ICT as a platform to improve e-health companies.
- *Supporting by free domain, it could be helpful beside speed memory.*
- *Training human resources, which is first player in this case.*
- *Giving speed process by multi-computers for assessing some projects which is needed for processing.*

7 Conclusion

This paper stated there are other important factors to performance e-health planning that governments should be considering seriously. It means that e-health is not only a fixed and linear model but also has a complex various and multidimensional based on real world. This conceptual economic e-health model is focused on two important efficiencies in technical and economic areas. Furthermore, we have attempted to show some economic perspectives of e-health, in particular, related to alternative markets. Because the competition market is an Ideal market and is engaged with many non-markets problems. Although, quasi competition market is more efficient than monopolist governments but the main responsibility of government is planning and controlling continuously not engaging directly with e-health deployment scenarios.

8 References

[1] *Car, J., Black, A., Anandan,C., Cresswell, K. (2008),The Impact of eHealth on the Quality &Safety of Healthcare, A Systematic Overview & Synthesis of the Literature, Report for the NHS Connecting for Health Evaluation Programme, Imperial College London, p.9-10*

[2] Pagliari, C., Sloan,D., Gregor, P., Sullivan, F., Detmer, D., Kahan, J.P., Oortwijn, W. and MacGillivray, S. *(2005), What Is eHealth* (4):A Scoping Exercise to Map The Field, Journal Medical Internet Research, 7(1):e9

[3] WHO Regional Office for the Eastern Mediterranean, authors (2005) E-Health in the Eastern Mediterranean. Eysenbach, G. (2001), What is eHealth? Med Internet Res, 3(2):e20

[4] Eysenbach, G. (2001), What is eHealth? Med Internet Res, 3(2):e20 Available at: http://ec.europa.eu/information_society/activities/health/whatis_ehealth/index_en.htm Last accessed [15th April 2010]

[5] Salgues, B., (2005) E-health, Integrated systems versus incompatible data stock, Proceedings of the 9th Congresso Mudial de Informacao em Saude e Bibliotecas, Salvador – Bahia, Brazil.

[6] Nazi, K.M. (2002), "The journey to e-health: VA healthcare network upstateNew York", Journal of Medical Systems, Vol. 27 No.1, pp.35-45.

[7] eGovernment and eHealth in Cyprus, (2009)

[8] Ridley, G. and Young, J. (2006), Towards Evaluating Health Information Portals: A Tasmania E-health Case Study . University of Tasmania

[9] Kokkinaki, Angelika I., Socrates Mylonas, Stalo Mina, (2005), E-GOVERNMENT INITIATIVES IN CYPRUS, *e*Government Workshop '05 (*e*GOV05), September 13 2005

[10] CHRISTODOULOU, E. (2008), The Development of eServices in an Enlarged EU

[11] Pagliari, C., et al (2005), Risk A. *What is e-health? Email sent to the SIM mailing list.* 2001. [2004 Jul 13]

[12] MACHLUP, Fritz (1964). Professor samuelson on theory and realism. *American economic review*, 54 (5), 733.

[13] KALTER, Robert J. and STEVENS, Thomas H. (1971). Resource investments, impact distribution, and evaluation concepts. *American journal of agricultural economics*, **53** (2), 206.

[14] ROBINS, James A. (1992). Organizational considerations in the evaluation of capital assets: Toward a resource-based view of strategic investment by firms. *Organization science*, **3** (4), 522-536.

[15] Williamson, O.E., (1981), The Economics of Organization: The Transaction Cost Approach, The American Journal of Sociology, Vol.87, No. 3, P. 548, 577

[16] MAHONEY, Joseph T. (1995). The management of resources and the resource of management. *Journal of business research,* **33** (2), 91-101.

[17] Bolton, D. J. (1937), M.Sc, Member. ELECTRICITY DEMAND AND PRICE

[18] Hunt,S.D. and Morgan, R.M., (1995), Resource-Advantage Theory, Jornal of Marketing, Vol. 61, P.77

[19] Mankiw, N.G., (2003), Principles of Economics. Mechanical Industry Press, 2003.8:172

[20] Gottinger H. W. (2003), Economise of Network Industries Network economics, Democracy and Efficiency in the Economic Enterprise, P. (1)

[21] Walker, M. (2000),Competition Law, Anti-Competitive Behaviour, and Merger Analysis: Economic Foundations

Reference Architecture for Knowledge Management Strategy (KMS) Deployment

Babak Akhgar[1], Mohammad Hassazadeh[2], Simeon Yates[1] and Richard Wilson[1]
[1]C3RI, Sheffield Hallam University, Sheffield, United Kingdom
[2]Faculty of Humanities, Tarbiat Modares University, Teheran, Iran

Abstract - *This paper explores the relationship between Enterprise Strategy and Knowledge Management (KM) deployment and argues that successful KM deployment requires a holistic architecture. Based on an industrial consulting project and an action research cycle, a three tier Reference Architecture (RA) for KM deployment is recommend, which addresses the strategic, core and deployment components of KM realisation. The latter can be used for strategic KM initiative of enterprises.*

Keywords: Knowledge Management, Architecture, Knowledge Management Strategy

1 Introduction

It is widely recognised that knowledge has become as a key component in sustainable socio-economic development. In this regard, the Organisation for Economic Co-Operations and Development (OECD) has emphasized the importance of the knowledge economy and the role of knowledge for economic development in the third millennium [11]. The World Development Report published by the World Bank emphasized that: "*For countries in the vanguard of the world economy, the balance between knowledge and resources has shifted so far toward the former that knowledge has become perhaps the most important factor determining the standard of living – more than Land, than tools, than* labor" [17].

World Bank again emphasized the importance of knowledge assets to nations and viewed them as an effective tool through which developing countries could participate effectively in the global economy[18]. From this argument, it follows that we are within or about to enter the Knowledge Society - a global social, economic and political configuration in which there is a special emphasis on the creation, dissemination and management of knowledge. Broadly speaking, the term Knowledge Society refers to any society where knowledge is the primary production resource instead of capital and labour. It may also refer to the use a certain society gives to information. A Knowledge society "creates shares and uses knowledge for the prosperity and well-being of its people". In a knowledge society the majority of social and economic relations are dependent upon the knowledge base rather than say capital. In such societies knowledge workers therefore play a vital role. It also follows that to be a successful organisation it is crucial to deploy an appropriate knowledge management strategy. Achieving a competitive edge in a knowledge society needs some necessary enablers that are deployable in a knowledge strategy. In this paper we strive to develop a RA of appropriate knowledge strategy deployment.

2 An Evolutionary Approach

Emphasis on knowledge in societal affairs and a focus on social capital in organizations are important indications of our era. Daniel Bell’s [6] focus on theoretical knowledge as a new core component of society reminds us that “knowledge” is considered by some thinkers to consist largely of symbolic representations of the natural world. Such understanding now mediates our relationship with the natural world and provides the basis for social understanding in modern societies and organisations [14]. It can be argue, these are, of course, contested positions. If knowledge is symbolic, then the capital embodying such knowledge may be thought to be symbolic also. This school of thought, often associated with Pierre Bourdieu, has led to the consideration of human capital, social capital and cultural capital as important features of contemporary social and economic organisation [7].

Social capital as a consistent set of ideas in relation to organisational cooperation is usually thought of in terms of social networks and the norms of reciprocity and trust that arise from them, and the application of these assets in achieving mutual objectives [12]. It seems that without an appropriate architecture for conductive components of social capital in organisations, it is difficult to achieve expected objectives. What organisations really do is to arrange the combination, transfer and exchange of knowledge, and social capital plays an essential tool in these processes. These are complex social activities, and have led to interesting studies of how organisations actually work (e. g. [16]). Bringing these elements together in an exhaustive framework will enable organisations to define appropriate data model and provide required physical and structural infrastructures to meet changing expectations of customers.

A holistic view of the strategies through which the production, consumption and management of knowledge function is needed. This need derives from the fact knowledge itself can to be “controversial” (see [8]), and its production is also considered to be different from previous economic goods. The production of knowledge in current circumstances also differs from traditional methods, with a focus on the context of application, the growth in the number of sites of production, and other features. Information and communication technologies provide all individuals with facilities to create and disseminate knowledge. Firms themselves are viewed by some writers in terms of their abilities to create, organise and transfer knowledge. The firm exhibits combinative capabilities to use learning from inside and outside the organisation, as Senge [13] named it a “learning organisation”, and its effectiveness is related to how well it performs this combined task [10].

Haggie & Kingston [9] argued that, "one fact that does seem to be agreed on is that different situations require different knowledge management strategies". A clear KM Strategy is needed to ensure the proper development of the Critical Success Factors (CSFs) for any KM System. Such a strategy also provides the basis for identifying the impact of KM on Business Processes, Human Resources, Organisational Extended Value Chain, Enterprise data models and any desired product / services development. These arguments provide us with an inevitable need of exhaustive reference architecture to realise necessary elements and depict relation among them.

3 Knowledge Management Strategy

Organisations often regard deployment of their KM cycles as one of the key strategic initiatives, which can potentially provide the necessary basis for sustainable competitive advantage. There are numerous methodological approaches for KM Deployment (KMD), predominantly driven from KM software vendors or consulting firms promoting their services. The lack of an end-to-end holistic approach for KM deployment that is grounded in a sound methodological foundation is clearly evident in the current KM market place ([3]). Akhgar and Mosakhani [2] elaborated on two types of KM deployment approaches - epistemological and ontological. He argued that although the ontological approaches for KM deployment provide substantive benefit and advantages, in order to fully realise KM potential, organisations need to adapt epistemological approaches. One of the critical success factors identified by Akhgar et al [4] for an epistemological approach for KM deployment is the formulation of a KM strategy. In this section of the paper we will elaborate on this and provide a detailed roadmap for the realisation of KM strategy deployment. However, before we address our proposed architectural roadmap, it is necessary to define KM strategy and the KM strategy formulation process. Conceptually we define KM strategy as “*a term that reflects an evaluatiable framework for a complex matrix of thoughts, visions, ideas, insights, learning processes, experiences, goals, expertise, values, perceptions, and expectations or collective mental constructs of individuals that provides specific guidance for specific actions in pursuit of particular ends by utilising knowledge within organisational extended value systems. [Thus, KM strategy formulation also can be defined as] a pragmatic, action-oriented and goal driven process of transforming organisational knowledge utilisation from current status (AS IS) to the desired status (TO BE) based on KM life cycle processes which include knowledge; collection, creation, transformation and collaboration, visualisation, storage, evaluation, business models refinement and assessment.*”

In order to develop a generic road map, we have critically evaluated a number of KM strategic initiatives. Intentionally we have chosen a wide range of industries and project sizes in order to extract a common set of factors contributing to successful deployment of a Knowledge Management Strategy (KMS). The industries chosen were as follows: Oil, Gas and Petrochemical, Automotive, Finance and Security. In total we have looked at 12 large KMS projects. We have used a conceptual template for the construction of a methodology (CTCM) (see Akhgar [1]) to identify and elaborate the critical success factors needed for our purpose. Based on the elements of CTCM we have identified 3 core perspectives as our viewpoints (Bashar [5]). Our viewpoints were; Business, Core KM and IT perspectives.

4 Proposed Architectural Roadmap

Based on previously identified perspectives we have constructed our RA as a three tier model (Figure1). The proposed architecture has 3 interrelated perspectives:

1- Strategic Propositions (Business View)

2- KM Tactical models (Core KM View)

3- IT Perspectives (IT View)

The foundation of the model is based on organisational strategy (Business View). It is identified that a KMS becomes successful only if at first the following strategic considerations are meet:

1) KM Strategic proposition (Why, What, How, When)

2) Development of the KMS CSFs including the project KPIs.

3) Identification of KM impact on Business Processes, Human Resources, Organisational Extended Value Chain, Enterprise data models and product / services offering

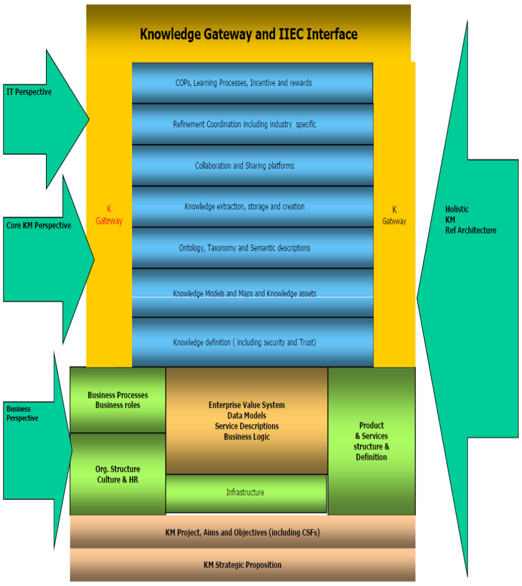

Knowledge Gateway and IIEC Interface
IT Perspective
Core KM Perspective
Business Perspective
COPs, Learning Processes, Incentive and rewards
Refinement Coordination including industry specific
Collaboration and Sharing platforms
Knowledge extraction, storage and creation
Ontology, Taxonomy and Semantic descriptions
Knowledge Models and Maps and Knowledge assets
Knowledge definition (including security and Trust)
K Gateway
K Gateway
Holistic KM Ref Architecture
Business Processes Business roles
Org. Structure Culture & HR
Enterprise Value System Data Models Service Descriptions Business Logic
Infrastructure
Product & Services structure & Definition
KM Project, Aims and Objectives (including CSFs)
KM Strategic Proposition

Figure 1 Reference Architectural

At the core our model are the KM life cycle elements. It addresses all the necessary components needed for tactical realisation of a KM project (Akhgar and Mosakhani, [2]) from knowledge definition and model formation of CommuCOPs.

The outer layer of our model (IT perspective) address the notion of knowledge gateways where the elements of KM life cycle can be visualised and communicated throughout an Enterprise's extended value system (via Internet, Intranet, Extranet or Cloud Computing platform) based on the criteria identified and justified at the business perspective layer.

5 Conclusions

Based on critical evaluation of 12 KM project deployments, we have put forward a RA for KM deployment. We have argued that the current KM solutions are inadequate for creating real competitive advantage for organisations as they are not addressing the holistic perspective of KM. We have argued that the presented RA is holistic enough to address ALL of KM deployment requirements from concept to code. Our proposed RA will be accessible through open wiki for practitioners to comment on its suitability.

6 References

[1] Akhgar , B, Strategic Information Systems, From Concept to Code, 2003, SHU Thesis

[2] Akhgar, B and Mosakhani , M, Epistemological Perspective of KM , 2010, 2d IRI Conference on KM.

[3] Akhgar, B and Yates (2011); Holistic Architecture for KM deployment, 3d Int Conf on KM. KNS.

[4] Akhgar, B, Mossakhanie, M and (2009); Epistomological perspective of KM, KM WP 2009.

[5] Bashar, N; An Abductive Approach for Analysing Event-Based Requirements Specifications. ICLP 2002: 22-372002

[6] Bell, D. The Coming of Post-Industrial Society: A Venture in Social Forecasting. New York, NY: Basic Books. 1973

[7] Bourdieu, P. 'The Forms of Capital'. In J. G. Richardson (ed.), Handbook of Theory and Research for the Sociology of Education. Westport, CT: Greenwood Press. 1986

[8] Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P. and Trow, M. The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies. London: SAGE Publications. 1994

[9] Haggie, Knox & Kingston, John (2003). Choosing Your Knowledge Management Strategy. Journal of Knowledge Management Practice. June Issue. [Online]. Avilabe at: http://www.tlainc.com/articl51.htm . Visited: 2011-03-06

[10] Kogut, B. and Zander, U. 'Knowledge of the Firm: Combinative Capabilities, and the Replication of Technology (reprinted from 'Organizational Science', 3(3), 1992)'.

[11] OECD. (1996). the knowledge- based economy. Paris: OECD. April Report. PP. 24-26.

[12] Putnam, R. D. Bowling Alone: The Collapse and Revival of American Community. New York, NY: Simon and Schuster. 2000

[13] Senge, P. The Fifth Discipline: The Art and Practice of The Learning Organization. The art and practice of the learning organization, London: Random House. 1990

[14] Stehr, N. 'Modern Societies as Knowledge Societies'. In G. Ritzer and B. Smart (eds), Handbook of Social Theory. London: SAGE Publications. 2001

[15] Bashar, Requirements Engineering, View Point 2001

[16] Wenger, E. Communities of practice: Learning, meaning, and identity. Cambridge: Cambridge University Press. 1999

[17] World Bank. (1998). World development report: Knowledge for development. Oxford press.

[18] World Bank. (2002). The knowledge assessment methodology and scorecards. [Online]. Available at: http://worldbank.org/gdln/programs/kam2002/methodology.htm. Visited: 2010-10-12

The Role of Trust in E-CRM: An Empirical Study

Naser Aniba, Hassan Makhmali, Mazen Qteishat, Jawed Siddiqi and Babak Akhgar
Faculty of Arts, Computing, Engineering and Sciences, Sheffield Hallam University, UK
Naniba@my.shu.ac.uk, a9039260@my.shu.ac.uk, mazenqteishat@hotmail.com, J.I.Siddiqi@shu.ac.uk, B.Akhgar@shu.ac.uk

Abstract-*The paper reports is on a major piece of research investigating the role of certain key factors of e-CRM in customers' use of airlines. It investigated these through an empirical study involving a large scale survey that gathered data from Afriqiyah Airways customers and the data collected was analysed using exploratory factor analysis. The report here focuses on trust and shows empirically the importance of trust as primary factor in explaining and predicting e-CRM.*

1 Introduction

Various definitions exist for e-CRM. Ab-Hamid and McGrath (2005) use the term to describe elements of CRM that are delivered through the Internet. The Internet functions as the channel for communication between the customer and the firm, with many of the processes automated to personalise the experience for the customer. Sanayei et al. (2010) consider e-CRM as a system for creating knowledge from process automation and the collection of information through Internet and information technology-based interactions between a company and its customers. Harris and Goode (2010) argued that the online environment is substantially different from the physical environment, and requires firms to adopt practices tailored to the environment to create an effective e-CRM system. Al-Momani and Noor (2009) suggested that the only difference between CRM and e-CRM is the use of Internet technology as a medium for communications with the customer. In contrast, Chen et al. (2007) noted that e-CRM systems 'can stand alone as web-based collaborative communication systems, or may be connected to powerful CRM analytics, or may serve as an interaction engine for enterprise-wide CRM systems.' This suggests that the design of e-CRM is flexible, with firms using multiple e-CRM strategies.

1.1 Determinant of E-CRM

We are conducting a major study involving the exploration of the key determinants of e-CRM in Airline usage .The five key determinants from a extensive literature review are Trust, Pre-sales services, After-sales services, Perception and Attitude ; for further details see [Aniba20011].

The investigation was empirical in nature in that we conducted a survey at Afriqiyah Airways (AAW).It explored the five key determinants which we mentioned above .However, in this study we focus exclusively on trust because it is one of the most important factors in e-CRM.

1.2 Trust

Trust is generally defined as the 'reliance on the integrity, ability or character of a person or thing' (Lilien and Bhargava, 2008). In many online transactions, trust is implicit, with a user obtaining information or providing data to a website based on an assumption that the entity operating the website is trustworthy. Research examining the perception of trust among online users has determined that the presentation and content of a website produced by an unknown entity positively influence the belief that a website is trustworthy (Chang and Chen, 2009). Online users make inferences about the unknown based on the information and cues available in the environment, with factors such as website usability functioning as surrogate factors to assess trustworthiness.

Factors such as the level of security provided by a firm and any history of breach of security can also influence the trusting belief (Fjermested and Romano, 2009). Trust not only affects the interaction of a user with a website, but also influences the loyalty of the user to that website which results in repeat visits (Flavian and Guinalu, 2006). From our extensive literature review we have classified the nine attributes into three categories: security, fairness and privacy & confidentiality. The nine attributes of Trust correspond directly to the nine questions see Appendix.

Security The literature indicates that security remains a very critical issue with which customers are concerned (Lai et al., 2010). In regards to the development in e-commerce, Nasir et al. (2007) stated that customers' initial concerns regarding online trust mainly focused on the issues of security and privacy on the Internet; that is because perceived security control and perceived privacy control are essential features of online transactions, which affect the development of online customers' confidence in e-commerce. The goal of security is to protect the confidentiality of customers' data collected by a firm (Charney, 2008). The usage of the most recent security technologies is crucial to improving the level of trust between a firm and its customers (O'Reilly and Finnegan, 2005).

Fairness The literature suggests that the perception of fairness in an organisation is a contributing factor to the amount of trust various stakeholders, including customers, place in an organisation. An individual who perceives the organisation as fair and equitable is more likely to use that organisation's services, including its e-CRM system. To some degree the perceptions of the power imbalance when using an e-CRM system relate to the perception of fairness. For example, users often consider an e-CRM system unfair if it does not provide transparent information concerning prices Zhang and Feng (2009) determined that the relationship of price to the perception of fairness occurs in two dimensions in e-CRM. A customer will perceive the price as fair if it is reasonable when compared to the prices charged by competitors, and if it represents value for the amount charged, which is related to the status or quality provided by the firm.

Privacy and Confidentiality The literature discussing prior research into trust also indicates that privacy and confidentiality are factors influencing trust in online communications methods. Internet users are increasingly sensitive to the protection of personal information, which affects their perceptions of and attitude towards a website. Privacy can be broadly defined as 'the right of an entity, acting on its own behalf, to determine the extent that it will interact with its environment, including the degree to which the entity is willing to share information about itself with others' (Lilien and Bhargava, 2008). A passive visitor to a website can remain anonymous because the website host is unaware of the identity of the visitor even if a tracking cooking is lodged in the visitor's computer. Once the visitor supplies information to the website, however, anonymity may be lost. As a result, a website has to provide the user with adequate assurances of confidentiality to support a decision concerning the degree of personal information the user is willing to provide (Fjermested and Romano, 2009). Confidentiality is related to the technical security procedures used by a firm operating a website and is a separate construct from privacy (Flavian and Guinalu, 2006).

2 Data Collection and Analysis

A self-administered survey questionnaire was used for the data collection element of the research design. In this study, the survey questionnaire was used to collect data about the perspectives of customers of Afriqiyah Airways concerning use of the firm's e-CRM system. Because no existing survey questionnaire was available to test the specific variables of Trust, Pre-Sales Services, After-Sales Services, Perception, and Attitude in the context of e-CRM for airlines, the research design required the development of a questionnaire to collect the data.

2.1 Survey questionnaire

The survey questionnaire was designed in six sections. The first section obtained information on the demographic variables of age, gender, and income relevant to the study.

The remaining five sections of the survey consisted of fifty-one questions intended to obtain data concerning perceptions of the respondents in the five dimensions of Pre-Sales Services, Trust, After-Sales Services, Perception, and Attitude, which were related to the variables under consideration in the study. The five sections assessing the dimensions related to the variables used a 5-point Likert scale that asked respondents to rate their level of agreement with statements, with a possible level of agreement ranging from 'strongly disagree' to 'strongly agree'.

2.2 Data collection procedure

The data collection procedure produced 306 usable responses, which was sufficient for the study. The participants completed the survey questionnaire either face-to-face or by email. No attempt at follow-up was made for non-respondents. The procedure resulted in the administration of 415 questionnaires and the return was a total of 306 usable questionnaires. This represented a response rate of usable questionnaires of 73.7%, which is very high for survey research. The response rate is high enough to conclude that the sampling was not substantially skewed by self-selection bias, which occurs when a high percentage of the study population does not take part in the research.

2.3 Exploratory factor analysis for Trust

Data provided by the survey questionnaire were analysed with descriptive statistics for the demographic information. Cronbach's alpha was used to establish the reliability of the survey questionnaire, and exploratory and exploratory factor analysis was used to establish the validity of the survey questionnaire.

Exploratory factor analysis was used to assess the dimension of Trust. This test assesses to see whether the items in the questionnaire provide a basis for measuring Trust as well as whether it is a single component and what proportion of the variation is explained by Trust. Prior to this Bartlett's Test for Sphericity was used to examine the homogeneity of variances to ensure that multi-collinearity was not present. Additionally, the Kaiser-Meyer-Olkin (KMO) test was used to assess the validity of the correlations among the items in each scale. The coefficient of correlation was used to assess inter-item correlations, with the minimum level necessary to establish a correlation set at 0.30.

Questions 13 through 21 of the survey questionnaire (see Appendix) assessed items related to the variable of Trust. The correlation coefficients as presented in Table 1 below indicated that Question 21 should be eliminated from the survey questionnaire because it had a correlation coefficient below 0.30.

Table1: Correlation Matrix for Trust

		Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21
Correlation	**Q13**	1.000	.485	.512	.477	.364	.458	.518	.377	.278
	Q14	.485	1.000	.739	.624	.390	.668	.591	.570	.426
	Q15	.512	.739	1.000	.667	.381	.641	.622	.489	.405
	Q16	.477	.624	.667	1.000	.491	.620	.482	.468	.496
	Q17	.364	.390	.381	.491	1.000	.552	.502	.445	.549
	Q18	.458	.668	.641	.620	.552	1.000	.618	.591	.486
	Q19	.518	.591	.622	.482	.502	.618	1.000	.645	.363
	Q20	.377	.570	.489	.468	.445	.591	.645	1.000	.568
	Q21	.278	.426	.405	.496	.549	.486	.363	.568	1.000

a. Determinant = 0.06

The correlation matrix was revised to account for the elimination of Question 21 from the survey instrument. This revision was necessary to determine whether the elimination of Question 21 produced changes in the correlation coefficient sufficient to result in the elimination of additional questions. The correlation matrix for the variable of Trust as presented in Table 2 below shows that Questions 13 through Questions 20 should be retained in the survey questionnaire.

Table 2: Revised Correlation Matrix for Trust

		Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20
Correlation	Q13	1.000	.485	.512	.477	.364	.458	.518	.377
	Q14	.485	1.000	.739	.624	.390	.668	.591	.570
	Q15	.512	.739	1.000	.667	.381	.641	.622	.489
	Q16	.477	.624	.667	1.000	.491	.620	.482	.468
	Q17	.364	.390	.381	.491	1.000	.552	.502	.445
	Q18	.458	.668	.641	.620	.552	1.000	.618	.591
	Q19	.518	.591	.622	.482	.502	.618	1.000	.645
	Q20	.377	.570	.489	.468	.445	.591	.645	1.000

a. Determinant = .011

The eigenvalues as shown in Table 3 indicate that the first component accounts for the greatest amount of variance in the eight items remaining for the Trust scale (i.e. nearly 60%).

Table 3: Explanation of Total Variance for Trust Scale

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	**Total**	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**
1	4.781	59.764	59.764	4.781	59.764	59.764
2	.746	9.324	69.087			
3	.638	7.981	77.068			
4	.611	7.638	84.706			
5	.379	4.741	89.447			
6	.335	4.182	93.629			
7	.291	3.639	97.268			
8	.219	2.732	100.000			

Extraction Method: Principal Component Analysis.

The scree plot for the Trust scale shown in Figure 1 also showed only one component accounting for the majority of variance to the left of the elbow.

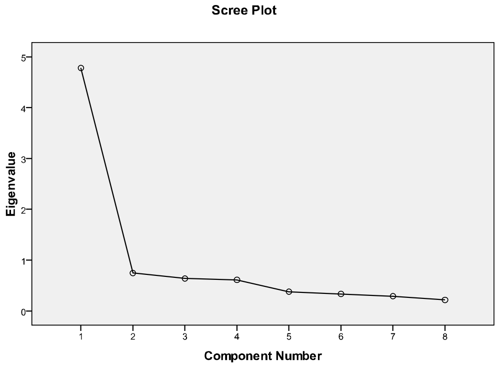


Figure 1: Scree Plot for Trust Scale

The factor loading for the eight questions in the Trust scale were all above 0.40, as shown in Table below ,with all items retained in the survey questionnaire, thereby confirming that all of the retained questions are relevant and that a single component accounts for the majority of the variance.

Table 4: Component Matrix[a]

	Component
	1
Q18	**.841**
Q14	**.832**
Q15	**.829**
Q19	**.809**
Q16	**.785**
Q20	**.743**
Q13	**.668**
Q17	**.653**

Extraction Method: Principal Component Analysis.

3 Concluding Discussion

The paper has reported on a major piece of research investigating the role of certain key factors of e-CRM namely: Trust, Pre-sales services, After-sales services, Perception and Attitude (see Aniba 2011) as they relate to customers' use of airlines. The report here focuses specifically on the key determinant of Trust; through the literature it establishes the importance of this determinant. The novelty of this paper is that it investigated Trust through an empirical study involving a large scale survey of 415 questionnaires that gathered data from Afriqiyah Airways customers. The data collected was analysed using exploratory factor analysis to establish that the items in the questionnaire provide a basis for measuring Trust as well as confirming that Trust is a single component and that the proportion of the variation explained by Trust is nearly 60%. The outcome of the investigation establishes empirically the importance of trust as primary factor in explaining and predicting e-CRM.

Appendix

These were the actual questions used regarding trust in the survey questionnaire. Responses to these questions on a five point Likert scale: *strongly disagree*, *disagree, neutral, agree* and *strongly agree* were collected; see table 5 for Q13 to Q21.

Table 5 questions related to Trust

13	I think that prices of products/services provided by the company are always lower compared to other airline companies.
14	I think that the terms and conditions laid out by the company are customer friendly and fair.
15	I think that the return/cancellation policies of the company should be customer-friendly and fair.
16	I think that the reputation of the company in terms of security is important.
17	I think that the company should always send a confirmation of secure payment and transmission.
18	I think that providing third party verification (e.g. seal of approval) to verify the company's website authenticity for customers is vital.
19	I think that providing a privacy statement to guarantee customer information is kept confidential is necessary.
20	I feel comfortable when providing sensitive information (e.g., credit card/debit card numbers) for online purchase.
21	I think that the online service of the company does not share customers' personal information with other sites.

4 References

[1] Ab Hamid, N. & McGrath, G. (2005). The diffusion of internet activity on retail web sites: A customer relationship model. *Communications of the IIMA,* **5**(2), 35-46.

[2] Al-Momani, K. & Noor, A. (2009). E-service quality, ease of use, usability and enjoyment as antecedents of e-CRM performance: an empirical investigation in Jordan mobile phone services. *The Asian Journal of Technology Management*, **2**(2), 11-25.

[3] Aniba, N. (2011). An investigation into factors of e-CRM influencing customer retention in Afriqiyah Airways Ph.D. Thesis, Sheffield Hallam University.

[4] Chang, H. & Chen, S. (2009). Consumer perception of interface quality, security and loyalty in electronic commerce. *Information Management,* **46**, 411-417.

[5] Charney, S. (2008) *Establishing end to end trust.* MicrosoftCorp. Available at http://download.microsoft.com/download/7/2/3/723a663c-652a-47ef-a2f5-91842417cab6/Establishing_End_to_End_Trust.pdf [Accessed 13 October 2010].

[6] Chen, Q., Chen, H., & Kazman, R. (2007). Investigating antecedents of technology acceptance of initial e-CRM users beyond generation X and the role of self-construal. *Electronic Commerce Research*, **7**, 315-339.

[7] Fjermested, J. & Romano, N. (2009). An integrated model for personalization, privacy and security in e-commerce. *Proceedings of the Fifteenth Americas Conference on Information Systems* (1-10).San Francisco California, August 6th-9th 2009.

[8] Flavian, C. & Guinaliu, M. (2006). Consumer trust, perceived security, and privacy policy: three basic elements of loyalty to a website. *Industrial Management and Data Systems*, **106**(5), 601-620.

[9] Harris, L. & Goode, M. (2010). Online servicescapes, trust, and purchase intentions. *Journal of Services Marketing*, **24**(3), 230-243.

[10] Lai, I. K. W., Tong, V. W. L. and Lai, D. C. F. (2010). Trust factors influencing the adoption of internet-based interorganizational systems. *Electronic commerce research & applications,* 1-9 .

[11] Lilien, L. & Bhargava, B. (2008). *Trading privacy for trust in online interactions* (1-35) Idea Group.

[12] Nasir, R. M., Ponnusamy, V., and Wazeer, M. W. (2007) *An exploratory study on the level of trust towards online retails among consumers in the United Kingdom and Malaysia*. Paper No. 8252, Munich Personal RePEc Archive, 2007. Available at http://mpra.ub.uni-muenchen.de/8252/ [Accessed 12 October 2010].

[13] O'Reilly, P., and Finnegan, P. (2005) Performance in electronic marketplace: theory in practice. *Electronic Markets*, **15**(1), 23–37.

[14] Sanayei, A., Ansari, A. & Ranjbarian, B. (2010). A hybrid technology acceptance approach for the E-CRM information system in the clothing industry. *International Journal of Information Science and Management*, Special Issue, 15-25.

[15] Zhang, X. & Feng, Y. (2009). *The impact of customer relationship marketing tactics on customer loyalty*. Master's Thesis, Halmstad University

Users'Evaluation of Public E- Services: Jordanian context

Mohammad Hjouj Btoush, Mazen Qteishat
Al-Balqa' Applied University-Jordan
m.hujooj@bau.edu.jo

Jawed Siddiqi and Babak Akhgar
Sheffield Hallam University-UK
{ j.siddiqi, b.akhgar}@shu.ac.uk

Abstract— The rapid rise in the delivery of e-services which involve communications and transactions between government, at various levels, and citizens has lead to a pressing need to develop models of users' satisfaction that will gauge the extent to which e-service meet the users' needs and expectations. This research aims to contribute to a growing major domain in e-government research, specifically in developing countries, that examines e-services delivery to users. The model that is used in this research, the 6I Model, has been developed by the researchers in an earlier stage. The results in this research which was conducted within the Jordanian context suggest that there is a value in utilizing a robust measure of users' perception of the level of satisfaction with the e-services presented to them.

Keywords— Public E-services, Jordan, 6I Model, Current Status, Desired Status

1 Introduction

E government has been defined in various ways, yet there is no watertight definition that researchers can refer to as the only one. However, definitions of e-government that has emerged range from narrow to broad ones, the narrow ones tend to focus on the provision of services through Internet, as for example this definition provided by West [1] in which e-government is: "*a delivery of government information and services online through the Internet or other digital means*". However, Defining e-government is not merely attaching it to technology or delivery of services; it has a more profound task of addressing the transformation of the methods and means by which governments interact with stakeholders [2]. Thus, e-government is better defined as way for governments to provide stakeholders with a more convenient and transparent access to government information and services, and to provide greater opportunities to participate in democratic institutions and processes.

A major service domain in the e-governemt involves the provision of e-services to citizens in the form of government to citizen (G 2 C) services. A common theme in the e-government discourse is the improved relations with citizens. Therefore, most governments present their approach to the adoption of e-government initiatives as being customer-centric approach, which means that services are designed and provided to meet or satisfy customers' or citizens' needs and expectations– leading them to be *customer-centric* or *citizen-centric* [3]. The main customers of the government are the citizens. According to this, meeting citizens' demands and maintaining users' satisfaction should remain the aims of e-services [4].

However, many researchers refer to this realm as G2C and C2G to highlight the reciprocal relationships, interactions and transactions between government agencies and citizens [5], [6]. In this case the governments offer their citizens with information and versatile services ranging from simple ones like provision of benefits, welfare, public health information, to more complicated services like renewing driving licenses and obtaining permits, income taxes, notification of assessment, and social security services [7], [8]. Furthermore, this interactive manner of services provision and use could enhance citizens' participation in the debates and forums that would reinforce the transparency and accountability of the government agencies, leading eventually to a practice of democracy where citizens share in decision making and policy shaping [6], [9]. Table 1 summarises the main objectives of G2C services.

However, this study attempts to contribute to understanding of about citizen interactions with e-governmental services within the Jordanian context through the use of a developed framework: the 6I Model [10].

THE MAIN OBJECTIVES OF G2C SERVICE
●Provide users with more effective, efficient and versatile e-services. ●Improve interactive communication between government and remote users. ●Create premium personalized and integrated e-services. ●Enhance user involvement, participation and contribution into e-services.

TABLE 1

2 RESEARCH MODEL

In developing the 6I maturity model, we utilise a qualitative meta-synthesis methodology to synthesize different e-government stage- models. Meta-synthesis is a research method that is used to integrate multiple studies and examine them critically in order to produce comprehensive and interpretative findings through discovering underlying themes and metaphors, so as to advance the current knowledge and produce a broad and comprehensive view [11]. The approach adopted by Siau & Long (2005) in arriving to their synthesised model was broadly followed; however, we arrived at our own synthesis independently.

We use meta-synthesis in this research to compare, interpret, translate, and synthesize different e-government maturity models to produce a new model: the 6I model. However, most maturity model stages were established depending on qualitative studies, and many by picking up on the literature because the same terms apply throughout all the different models without having empirical evidence or quantitative studies to build up the maturity model or underpin it. In addressing this gap in literature, we considered taking the building up of the model a step further by presenting an empirical evidence to underpin the proposed model through a quantitative research.

We divided the e-services within the research context into two categories, as presented in the previous section. However, here we explain what is meant by each category. The current status of the e-services is part of the 6I model and here it represents the actual state of the e-services within the research context. It identifies the available facilities that each eservice provides to the user, whereas the desired status, which is also part of the 6I model, represents what is not available in any of the provided e-services, and what users aspire to have in the future.

A summary of the characteristics of each stage within the two categories is presented in table 2.

State	6I Model Stages	Characterization of e-services
Current	Inform	Provides content that informs the user, ranges from formal, limited static content to dynamic specialized and regularly updated information.
	Interact	Two way communication in which interaction flows between government and users via ICT features range from downloading information to email communication possibility using security technique like keys password...etc.
	Intercommunicate	Carry out and complete transaction online. This may range from filling and updating forms electronically to processing payments and issuing of certificate. A complete chain of activities or transaction.
	Individualize	Allows users to be identified and /or services to be personalized, so that services that are offered are tailored to the individual's needs.
Desired	Integrate	Combine different separate services ranging from clustering of common services to a unified and seamless service (So that the parts are hidden from the user)
	Involve	Promotion of citizens' participation and empowerment. This can range from survey to voting to focus groups, and opinion polls. This could have either direct or indirect influence on decision-making and policy shaping.

TABLE 2
E-SERVICES' CHARACTERISTICS ACCORDING TO THE 6I MODEL

3 DATA COLLECTION

This research is addressing the users' true needs and expectations within the Jordanian context. Jordan is a developing country that has adopted and implemented e-government strategies to facilitate delivery and access to government services. The introduction of the e-services under the paradigm of e-government was stated in the e-government mission to "manage the transformation of the government towards a more "citizen-centric" approach in the delivery of services by means of appropriate technology, knowledge management and skilled staff" [12]. In its endeavouring to present e-services to the different stakeholders, Jordan launched many e-services since 2000. Nevertheless, an evaluation of e-services development is needed to help making them more efficient and effective. By evaluating e-services, we do not mean measuring the number of services provided online; rather the evaluation should contribute in understanding users- related issues. Therefore, we believe that using our proposed conceptual model, the 6I model, as a benchmark to evaluate these e-services from users' perceptions would contribute in understanding the true needs and expectations of users.

A questionnaire was used to gain insights of users' perceptions of the level of satisfaction of the e-services in the Jordanian context. In developing this questionnaire; certain procedures were followed. The research questions to be answered through this part were formulated clearly in order to understand the objectives of the questionnaire. Then, the specifications of the target population were identified. This was followed by a review of the relevant literature which enabled identifying the dimension of the investigated subject. Next, questions were selected based on consideration of research environment, participants, and objectives. The questionnaire was divided into six subscales that resemble the facilities of conceptual framework; the 6I Model, and these were further divided in to current status of public services, which include (Inform, Interact, Intercommunicate, Individualize) and the desired status of the public e-services including (Integrate and Involve). A subscale for the usability was added. So at this phase, this questionnaire aims to explore the relationships between these subscales and the five demographic characteristics, which provides in-depth analysis of users' perceptions of the public e-services.
Moreover, since the main purpose of the questionnaire, which was adapted from WebQual 4.0, was to evaluate the public e-services depending on a conceptual framework, the 6I model, some modifications on the WebQual 4.0 was undertaken. WebQual 4.0 is a well established technique and highly validated instrument that can provide both wide- and fine-grained measurements of organizational Web sites. However, we do not think, to the best of our knowledge that it has been used in terms of factor analysis. It is based on quality function deployment and has been extensively used [13], [14]. Basically it is looking at the quality of websites, but we extended it or modified it to be used for investigating users' perceptions of e-services. WebQual 4.0 originally consists of 23 questions. However, some questions were added depending on extensive literature review, and, more specifically, depending on the developed conceptual framework. The final modified version of the users' questionnaire that was used in this research consists of 31 items. Moreover, the original instrument recorded responses on a seven-point Likert scale. This study used a five-point scale to encourage a complete response and provide more flexibility in analysing the data.

The sample for users was drawn from a list of Jordanian universities, community colleges and Internet Cafes because it was believed that the staff and the students at these educational institutions, and the citizens at the Internet Cafes have the skills of the ICT that enable them to use the public e-services, and, therefore, to answer the research questions. Moreover, the slogan of the MoICT (The Ministry of Information and Communications Technology), which is responsible for fostering e-services' adoption, is customer-centric or citizen-centric approach to e-services' provision [12]. Yet, the potential users' perceptions of the public e-services have not been taken into consideration. This fact provides a strong motivation to investigate the perceptions of the users to whom these e-services are directed.

4 TRENDS OF ANALYSIS

The analysis is conducted on two levels. The first is based on the use of factor analysis, and
aims to identify dimensions that are related to e-service usability and the different dimensions that constitute the proposed 6I maturity model. The second level is based on bivariate analysis, which provides an indication of the relationship or the correlation between the independent (demographic) variables, and the dependent variables

Three cohorts of users were found to be standing out as follows:

- Those with high level of education
- Those with greater ICT expertise
- The group of users, who are more mature

However, it should be noted that the strength of the associations that we determined were small to moderate for what follows:

The first cohort (i.e., those with higher levels of education) had a negative level of satisfaction with usability. In relation to the detailed analysis of the current stages they had a positive level of satisfaction with *Inform*; however, they had negative levels of satisfaction with *Interact* and *Intercommunicate*

In relation to the desired stages this cohort had a favourable attitude towards high quality/*Integrate* stage as well as towards more participation and engagement.

I. The second cohort (i.e., those with higher levels of ICT expertise) had no significant association with usability, but like the previous cohort had a positive association with *Inform* but a negative association with *Interact* and *Intercommunicate*. However, unlike the "educated cohort", they did not have any significant preference for the desired stages (*Integrate* & *Involve*)

II. The third cohort (i.e., the more mature group) shared with the "educated" cohort a negative level of satisfaction as measured by usability. However, they did not share the trend of the other two in relation to *Inform*, *Interact*, or *Intercommunicate*, but did share with the "educated" cohort the favourable level of satisfaction for *Integrate.*

Table 3 summarises the suggested trends that appear in users' perception of the e-services, where upward or ascending arrows refer to the users' favourable perceptions of the e-services, and the downward or descending arrows to the users' unfavourable perceptions of the e-services in the research context.

Demographics	Broad Brush	**6I** Maturity Model				
		Current Status **4Is**			Desired Status **2Is**	
	Usability	**Inform**	**2Is** (**I**nteract, **I**ntercommunicate)	**Individualize**	**Integrate**	**Involve**
Education Level	− ↘	+ ↗	− ↘		+ ↗	+ ↗
ICT		+ ↗	− ↘			
Age	− ↘				+ ↗	− ↘

TABLE 3

A SUMMARY OF THE ANALYSIS' TRENDS

Our first analysis revealed that the users with higher levels of education and those who are more mature perceive the levels of satisfaction with e-services as measured by our broad brush measure of usability decreases; i.e. There is a negative trend.

Detailed analysis with the current e-services offerings of Inform, Interact, and Intercommunicate show that the above mentioned negative trend is reflected in relation to the Interact and Intercommunicate stages, but for higher levels of education and ICT expertise the more detailed analysis clearly provides a richer and more accurate picture of this earlier negative trend in that the more mature; and those with higher levels of ICT expertise and education level has a positive trend with the Inform stage, but have a negative trend with Interact and Intercommunicate.

Further detailed analysis of the desired e-services offerings of Integrate and Involve shows once again that higher levels of education have higher expectation on a quality of service (i.e. an integrated service) as well as a desired to be more engaged in the decision making. Yet, interestingly, the more mature appeared to be less willing to engage or to be involved.

The results about the Inform stage are consistent with the findings of the few studies that have investigated the public e-services in developing countries, for example (Akman et al., 2005; AlAwadhi & Moriss, 2008; Bouaziz**,** 2008)[15], [16], [17] have reported that users who have greater level of education and ICT expertise are more likely to use public e-services to obtain information. However, the findings about Interact and Intercommunicate contradict the same previous studies that have also reported that the same group of individuals i.e. those with greater education and ICT expertise are the ones to Interact and Intercommunicate more with government using public e-services. The contradiction could be justified by the fact that none of the studies have considered the nature and characteristics of the Interact or Intercommunicate of public e-services and the perceived level of satisfaction of the different users with these e- services. The findings might also indicate potential concerns of those with greater level of education and ICT over risks of security and privacy, which are usually seen as crucial and salient for the implementation and adoption of the Intercommunicate stage of the e-services in many studies, for example, (Hiller & Bélanger, 2001; Warkentin et al., 2002; Holden & Millet, 2005; Irani et al., 2006) [18], [19], [20], [21].

In relation to the desired status of e-services, no studies, to the best of our knowledge, have examined users' perceptions of the Integrate or Involve stages. However, Siddiqi et al. (2006) and Grimsley & Meehan (2007) [22], [23] consider these as new potentials that users are seeking from their engagement with e-services; mainly: to feel more empowered to take charge of the e- services they use and influence policies that affect them.

5 Conclusion

The results reported in this paper revealed that users of the public e-services in Jordan were satisfied with some of the current status of the e-services and not happy with others. Yet, that did not prevent them from wanting better public e-services. Some users were also found to have negative evaluation of the usability of the e-services.
A review of the available literature confirms that this is the first in-depth study within the Arab countries, and more specifically within the Jordanian context, A report by OECD (2007: 10) [24] points out that there is "a lack of evaluation culture" of e-services within the Arab countries. This study takes a step towards that direction of creating an evaluation culture by accounting for the users' evaluation of the public e-services; according to their stages, which are suggested by the proposed 6I model. To the best of our knowledge, this study would be the first to investigate what is affecting the e-services' adoption and use, not for the organizations but for the users themselves. In doing so we are tackling how in reality the e-services are designed and provided for citizen-centric approach.

Moreover, this study is also concerned with the usability as the overall way of evaluating the level of satisfaction from again the users' perspectives. Although, there are some studies that address the issue of usability of public e-services within the Arab countries for example (Abanumy et al., 2005) [25], nevertheless, this study differs in that it accounts for the users' evaluation rather than depending on different methods and tools to evaluate the public e-services and to see whether these e-services meet the requirements of these tools. Therefore, the findings provide a valuable contribution to the body of knowledge within the Jordanian context. They suggest some practical and theoretical implications for public e-services implementation and development.

References

[1] West, D.M. (2004). 'E-government and the transformation of servicedelivery and citizen attitudes', *Public Administration Review*, **64** (1), 15-27
.

[2] Lofstedt, U. (2005). 'E-government assessment of current research and some proposals for future directions', *International Journal of Public Information Systems*, **1**(1), 39-52.

[3] Horan, T.A., Abhichandani, T., Rayalu, R. (2006). 'Assessing user satisfaction of e-government services: Development and testing of quality-in-use satisfaction with Advanced Traveler Information Systems (ATIS)'. In the *Proceedings of the 39th Hawaii International Conference on System Sciences, IEEE Computer*
.

[4] Ho, A.T. (2002) 'Reinventing local governments and the e-government Initiative', *Public Administration Review*, **62** (4), 434-444.

[5] Fang, Z. (2002). 'E-government in digital era: Concept, practice and Development', *International Journal of the Computer, the Internet and Management,* **10** (2), 1-22.

[6] Halachmi, A. (2004). 'E-government theory and practice: The evidence from Tennessee (USA)'. In Holzer, M., Zhang, M. & Dong, K. (Eds.). *Proceedings of the Second Sino-U.S. International Conference*: *—Public Administration in the Changing World*□ Beijing, China, 24-36.

[7] Riley, B.T. (2001). 'Electronic Governance and Electronic Democracy:Living and Working in the Connected World', *Commonwealth Centre For Electronic Governance*, Brisbane, Australia, from: <http://www.electronicgov.net/pubs/research_papers/eged/contents.asp.> (15Nov. 2006)

[8] Sagheb-Tehrani, M. (2007). 'Some steps towards implementing egovernment', *SIGCAS Computers and Society*, **37** (1), 22-29.

[9] Ndou, V. D. (2004). 'E–Government for developing countries:
Opportunities and challenges', *The Electronic Journal of Information Systems in Developing Countries*, **18** (1), 1-24.

[10] Hjouj Btoush, M., Siddiqi, J., Grimsley, M., Akhgar, B., & Alqatawna, J.(2008). 'Comparative review of e-service maturity models: 6I Model'. In the *Proceedings of The 2008 International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government (EEE'08).* Las Vegas,Nevada, USA.

[11] Siau, K. & Long Y. (2005) 'Synthesizing e-government stage models-ameta-synthesis based on meta- ethnography approach', *Industrial Management& Data Systems*, **105** (4), 443-458.

[12] MoICT (2006). _e-Government program overview‘. From: <http://www.moict.gov.jo/MoICT/MoICT_program_overview.aspx#Executive_summary> (27 Oct 2006)
.

[13] Barnes, S.J. & Vidgen R. (2003). 'Measuring web site quality improvements: A case study of the forum on strategic management knowledge exchange', *Industrial Management & Data System*, 103 (5), 297-309.

[14] Loiacono, E. T., Watson, R.T. & Goodhue, Dale L. (2007). 'WebQual: An instrument for consumer evaluation of web sites', *International Journal ofElectronic Commerce*. **11** (3), 51-87.

[15] Akman, I., Yazici, A. Mishra, A. & Arifoglu, A.(2005) 'E-government: A globalview and an empirical evaluation of some attributes of citizens', *Government Information Quarterly,* **22** (2), 239-257.

[16] Alawadhi, S. & Morris, A. (2008). 'The use of the UTAUT model in the adoption of e-government services in Kuwait'. In *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, 219-229.

[17] Bouaziz, F. (2008). 'Public administration presence on the web: A cultural Explanation', *Electronic Journal of E-Government,* 6 (1), 11-22.

[18] Hiller, J.S. & Bélanger, F. (2001). 'Privacy strategies for electronic Government'. In Abramson, M.A. & Means, G. (Eds.) *E- Government 2001.*Lanham, (MD) :Rowman & Littlefield.

[19] Warkentin, M., Gefen, D., Pavlou, P.A.& Rose, G.M. (2002).
'Encouraging citizen adoption of e-government by building trust', *ElectronicMarkets,* **12** (3), 157-162.

[20] Holden, S.H. & Millett, L.I. (2005). 'Authentication, privacy, and the federal egovernment',*The Information Society,* 21 (5), 367-377.

[21] Irani, Z., Al-Sebie, M. & Elliman, T. (2006). 'Transaction stage of egovernment systems: Identification of its location and importance'. In the *Proceedings of the 39th Annual Hawaii International Conference on System Sciences*, IEEE Computer Society Press, 1-9.

[22] Siddiqi, J., Akhgar, B., Gamble, T. & Zaefarian, G. (2006). 'A
framework for increasing participation in e-government', in the *Proceedings of The 2006 International Conference on E-Learning, E-Business, Enterprise Information Systems, E-Government & Outsourcing*. Las Vegas, 60-66.

[23] Grimsley, M. & Meehan, A. (2007). 'E-government information systems:Evaluation-led design for public value and client trust', *European Journal of Information Systems,* 16 (2), 134-148.

[24] OECD (2007). 'Measuring and evaluating e-government in Arab
countries', *Fourth High Level Seminar on Measuring and Evaluating EGovernment*. Dubai. From; http://www.dsg.ae/cms/_data/global/Publications/EGov%20in%20Arab%20Countries.pdf (1 Jun. 2007).

[25] Abanumy, A., Al-Badi, A. & Mayhew, P. (2005). 'E-government website accessibility: In-depth evaluation of Saudi Arabia and Oman', *The Electronic Journal of E-government,* 3 (3), 99-106.

Entity Resolution for Longitudinal Studies in Education using OYSTER

E. D. Nelson [1] **and J. R. Talburt** [2]
[1] Department of Information Quality, University of Arkansas, Little Rock, AR, USA
[2] Department of Information Quality, University of Arkansas, Little Rock, AR, USA

Abstract - *This paper describes the application of Oyster, an open source, general purpose entity resolution (ER) system, to the problem of conducting multi-year longitudinal studies of student achievement. Although originally designed to support ER instruction and research, this paper demonstrates that OYSTER can be used in practical applications with processing requirements on the order of one million records, a range that includes many existing small-scale information systems. The paper also discusses an enhancement to the basic R-Swoosh algorithm implemented in OYSTER that allows higher performance in processing student files with high duplication rates.*

Keywords: Entity Resolution, OYSTER Open Source System, R-Swoosh Algorithm, Longitudinal Studies

1 Introduction

Entity Resolution (ER) is the process of determining whether two references to real world objects are referring to the same, or to different, different objects [1]. This is done by successively locating and merging multiple records [2], [3], [4], [5]. References that refer to the same entity are said to be "equivalent references." In ER it is important to understand that entities are real world object that do not exist within information systems. Information systems only contain representations of these objects called entity references [1].

ER can be applied to any type of entity reference. When the underlying object is a customer, whether the customers are individuals or other businesses, is called customer data integration (CDI). CDI allows companies to maintain an accurate, timely, complete and comprehensive representation of a customer across multiple channels, lines of business, and enterprises. Often times, the ability to recognize a customer at any point within the enterprise allows improved customer experience and the ability to up-sale or cross-sale new services or products. If the data is product in nature it is called Product Information Management (PIM). This type of process is integral to large reseller and wholesale companies that often times product but with different descriptive attributes.

Newcombe, Kennedy, Axford, & James [6] state that ER consists of two steps locating or prospecting the records to be used in the merge step, then a comparison step to match the records. They specifically look at ways to optimize both the location and merge steps. Fellegi and Sunter [7] give the statistical basis for linking data. Newcombe, Fair, and Lalonde [8] provide a history of probabilistic record linkage and show empirical results. The SERF project [9] looks at a generic description of ER. They treat the match and merge processes as generic black boxes and describe a series of algorithms to resolve sets of references.

2 Oyster

OYSTER is an open source project sponsored by Center for Advance Research in Entity Resolution and Information Quality (ERIQ) and the University of Arkansas at Little Rock. OYSTER was originally designed to support instruction and research in ER by allowing users to configure its entire operation through XML scripts executed at run-time. The resolution engine of the current version (3.0) can be configured to run in several modes including record-linking/merge-purge, identity resolution, identity capture, and identity update. In addition the system's attributes and identity rules are defined with run-time scripts, as well as, the location and type of each reference source to be processed. OYSTER source code and documentation is freely available from the ERIQ website (1).

3 Problem Definition

We are working with an organization that will be providing data to researchers who request data. The organization receives multiple types of files from different state agencies. The initial files that were a part of this study are the student enrollment files for all of the public schools within the state. There is a total of seven years of data totaling 3.8 million records. There are also test scores, ACT records, BMI records and other records. Within these records a student can be seen across multiple years, i.e. students entering, exiting, and being promoted to higher grades. Student can also often be seen multiple times within a year, i.e. student moves for school district A to district W, the students parents get married or divorced necessitating a name change, etc. All these issues create a large level of duplication between the files. It is the implications of this large amount of duplication on the effects of Entity Resolution processing that is the focus of this paper.

In resolving the data files 11 simple rules (Table 1) were used to determine equivalence. It was also decided to use R-Swoosh [10] to ensure that all possible merges where processed. When each file was run alone R-Swoosh worked very well. But it was noticed that when the files where run together there was a larger number of R-Swoosh iterations. The multiple processing or churn was causing excessively long run times. Initially it was thought that the implementation of the R-Swoosh algorithm was incorrect but it was determine that it was correct and that the problem was a combination of how R-Swoosh processed records and the fact that there was a large percentage of duplication between the files. This was further compounded by the distance between many of the duplications, i.e. one file would have to be completely processed before the duplicates in the next file would be seen.

	First Name	Last Name	First Middle Name	Full Name	Date of Birth	SSN
Rule 1	Exact Ignore Case	Exact Ignore Case			Exact	Exact
Rule 2				Exact Ignore Case	Exact	Exact
Rule 3	QTR(0.25)	Exact Ignore Case			Exact	Exact
Rule 4			QTR(0.25)	Exact Ignore Case	Exact	Exact
Rule 5		Exact Ignore Case			Exact	Exact
Rule 6	Exact Ignore Case				Exact	Exact
Rule 7	SUBSTRLEFT(5)	SUBSTRLEFT(5)			Exact	Exact
Rule 8	Exact Ignore Case	Exact Ignore Case				Exact
Rule 9	Exact Ignore Case	Exact Ignore Case			Exact	
Rule 10		Hyphenated			Exact	Exact
Rule 11					Exact	Exact

Table 1. Identity Rules used for all Runs

4 R-Swoosh

According to the SERF project, R-Swoosh uses two sets R and R'. R is the set of all input records and R' is the set of all non matched records. Records from R are compared to R' in a pair wise fashion. If a match is found the records are merged. The matched record is removed from R' and the merged record is placed on the bottom of R. This continues until all records have been removed from R. In this respect R-Swoosh is easy to implement and it avoids some unnecessary comparisons because of the merge and replacement of the initial entity references with the merged value.

Supposed that you have the records shown in Table 2, for a record to match at least four attribute fields must be an exact match. Merging is a simple merge of each attribute field into an attribute field set. Using R-Swoosh and processing each record in order none of the records match until record 7 is read. It then takes an additional 6 matches to determine that all 7 references refer to the same entity. In fact it doesn't matter which order the records are in it will always take 7 matches. A walk through in Listing 1 shows the entire progression.

	Field1	Field2	Field3	Field4			Field1	Field2	Field3	Field4
1	A	B	C	D		Rule 1	Exact	Exact	Exact	
2	D	E	F	G		Rule 2	Exact	Exact		Exact
3	A	B	F	G		Rule 3	Exact		Exact	Exact
4	D	E	C	D		Rule 4		Exact	Exact	Exact
5	A	E	F	D						
6	D	B	C	G						
7	A	E	F	G						

Table 2. Example Data and Rules

1. Read in record 1, No match, New Entity
2. Read in record 2, No match, New Entity
3. Read in record 3, No match, New Entity
4. Read in record 4, No match, New Entity
5. Read in record 5, No match, New Entity
6. Read in record 6, No match, New Entity
7. Read in record 7, Matches Record 2 on Rule 4, Merge records and put on bottom of the list as RS1.
8. Read in record RS1, Matches Record 5 on Rule 1, Merge records and put on bottom of the list as RS2.
9. Read in record RS2, Matches Record 4 on Rule 2, Merge records and put on bottom of the list as RS3.
10. Read in record RS3, Matches Record 6 on Rule 3, Merge records and put on bottom of the list as RS4.
11. Read in record RS4, Matches Record 1 on Rule 1, Merge records and put on bottom of the list as RS5.
12. Read in record RS5, Matches Record 3 on Rule 1, Merge records and put on bottom of the list as RS6.
13. Read in record RS6, No match, New Entity
14. No more records, end

# of Consolidation Steps: 6

Listing 1. R-Swoosh applied to example data

5 R-Swoosh Enhanced

The enhanced version uses three sets but instead of reading in all the records into R, records are read one at a time from the input stream (R). This initially reduces the memory requirements needed since R is transient by nature. Records that are not matched are placed in R’. Because ER is expensive [10] a simple inverted index is used to reduce any unnecessary comparisons. R-Swoosh compares records in a pair wise fashion. In the R-Swoosh algorithm, when a match is found the records are combined and they are both placed back into the record queue at the end to be recheck at some later time for additional matches. Matching continues until the queue is emptied. The enhanced version makes one slight change; with the use of the index it pulls back a candidate list of all possible matches to R’. If multiple records are found to match they are all merged with the input record, removed from R’ and placed on R’’. R’’ is used to hold all records that have been matched and merged. In R-Swoosh these are placed on the end of R but since R in the enhanced version is a stream R’’ represents the end of the stream. Once the stream is extinguished R’’ takes the place of R.

A simple example shows the benefits of this slight change. In the enhanced R-Swoosh method, again no records are matched until record 7 is read. At that time records 2, 3 & 5 are all found to match. At that point they are all merged with record 7 and put on the bottom of the list as RS1. The next record read is RS1 since it is the only record in the list and it is found to match to 1, 4, & 6 (Listing 2).

1. Read in record 1, No match, New Entity
2. Read in record 2, No match, New Entity
3. Read in record 3, No match, New Entity
4. Read in record 4, No match, New Entity
5. Read in record 5, No match, New Entity
6. Read in record 6, No match, New Entity
7. Read in record 7, Matches Record 2, 3 & 5, Merge records and put on bottom of the list as RS1.
8. Read in record RS1, Matches Record 1, 4 & 6, Merge records and put on bottom of the list as RS2
9. Read in record RS2, No match, New Entity
10. No more records, end

# of Consolidation Steps: 2

Listing 2. Enhanced R-Swoosh Applied to Example Data

This slight change was able to reduce the number if iterations from 6 to 2. But here order is important, move the glue record (record 7) to a different point in the file can change the number of iterations required to fully resolve the entity. Even then the number of consolidation steps is less than regular R-Swoosh. If we take the permutation of the order of these 7 records we get 5040 different dataset. Running R-Swoosh on each one shows that order is unimportant as they all return 6 consolidation steps. If we run the enhanced R-Swoosh on the same data sets we find that in 3,168 (62.9%) of the data sets it takes 3 consolidation steps and 1,872 (37.1%) it takes 2 consolidation steps. But in all cases this requires fewer steps than R-Swoosh to produce the same results.

6 Results

To determine how the Enhanced R-Swoosh compares to R-Swoosh, three files containing First and Last Name, SSN and DOB were processed with the Oyster system. The files were run separately and then together using both the R-Swoosh and Enhanced R-Swoosh methods. File A contained 588,279 records and file B contains 463,405 records and file C contains 585,409 records. Each test is run on a Dell Server running Linux (CentOS release 5.5) using 8 GB of main memory and Java 1.6 VM.

7 Conclusion and Future Work

In using the Enhanced version of R-Swoosh during the project we have seen several gains in processing speed while producing the same result. Based on the empirically results that we are seeing, Enhanced R-Swoosh seems to be a viable alternative to the standard R-Swoosh ER algorithm. But as can be seen in the results section there are some differences that are not completely understood. Future work includes determining in what circumstances Enhanced R-Swoosh should be used, determining the amount of duplication that is necessary for efficient ER with Enhanced R-Swoosh and refining the implementation to allow for a more efficient matching.

Run	Total Records	Clusters	Max Cluster Size	Min Cluster Size > 1	Average Cluster Size	Num. of Duplicate Records	Duplication Rate	Elapsed Seconds	
								R-Swoosh	Enhanced R-Swoosh
File A	588,279	523,068	9	2	1.12467	120,301	11.085%	4,818	5,129
File B	463,405	462,815	3	2	1.00127	1,177	0.127%	2,254	2,286
File C	585,409	528,714	8	2	1.10723	105,821	9.685%	5,094	4,981
File A & B	1,051,684	581,569	9	2	1.8084	893,750	44.701%	13,521	13,351
File A & C	1,173,688	654,135	12	2	1.7943	945,006	44.267%	17,193	17,238
File B & C	1,048,814	558,915	9	2	1.8765	930,615	46.710%	13,369	13,679
All Files	1,637,093	655,861	13	2	2.4961	1,477,165	59.937%	30,134	27,634

Table 3. Comparison of Regular and Enhanced R-Swoosh Performance

Running each file individually using R-Swoosh, reasonable consistent times when the duplication rate is fairly low. Duplication rate is calculated by dividing the clusters by the total records and subtracting from 1. But when the duplication rate increases, the run times also increase. This is due to the fact that R-Swoosh only matches one record at a time, and when it finds a match, puts the merged record back on the bottom of the set. The extended R-Swoosh, does much better on the higher duplication rates as can be seen in Table 3. There was a difference in run time by 2500 seconds a time savings of 8.3%. Interestingly, the Enhanced R-Swoosh under performs when the duplication is low actually taking longer to run. It is believed that the cost of the more complex coding is what is causing this increase. Two, other anomalies where noticed when Files AC and CB were processed it was found that the Enhanced R-Swoosh took longer to resolve these records. It is unknown what caused these two ER processes to run slightly longer. We are not running on a dedicated machine so there is the possibility that there was resource contention issue.

8 Acknowledgment

The research described in the paper was supported in part by a grant from the Arkansas Department of Education.

9 References

[1] Talburt, J. R. (2011). *Entity Resolution and Data Quality.* Morgan Kaufman.

[2] Benjelloun, O., Garcia-Molina, H., Gong, H., Kawai, H., Larson, T.E., Menestrina, D., Thavisomboon, S., (2007). "D-Swoosh: A Family of Algorithms for Generic, Distributed Entity Resolution", *Proceedings of the 27th International Conference on Distributed Computing Systems*; Retrieved from HYPERLINK "http://infolab.stanford.edu/serf/" http://infolab.stanford.edu/serf/

[3] Bhattacharya, I., Getoor, L., (2007). "Collective entity resolution in relational data", *ACM Transactions on Knowledge Discovery from Data (TKDD)*; 1(1)

[4] Bilgic, M., Licamele, L., Getoor, L., Shneiderman, B., (2006). "D-Dupe: An Interactive Tool for Entity Resolution in Social Networks", *IEEE Symposium on Visual Analytics Science and Technology*; Retrieved from HYPERLINK "http://infolab.stanford.edu/serf/" http://infolab.stanford.edu/serf/

[5] Garcia-Molina, H., (2006). "Pair-Wise entity resolution: overview and challenges", *Proceedings of the 15th ACM international conference on Information and knowledge management;* 1(1)

[6] Newcombe H. B., Kennedy J. M., Axford S. J. and James A. P., (1959). "Automatic Linkage of Vital Records", *Science New Series*; 130(3381):954-959

[7] Fellegi, I.P., & Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210.

[8] Newcombe H.B., Fair M.E., Lalonde, P. (1992). "The Use of Names for Linking Personal Records", Journal of the American Statistical Association; 87(420)

[9] Benjelloun O., Garcia-Molina H., Kawai H., Larson T.E., Menestrina D., Su Q., Thavisomboon S., Widom J., (2006). "Generic Entity Resolution in the SERF Project", *IEEE Data Engineering Bulletin*; Retrieved from HYPERLINK "http://infolab.stanford.edu/serf/" http://infolab.stanford.edu/serf/

[10] Benjelloun, O., Garcia-Molina, H., Menestrina, D., Su, Q., Whang, S. E., & Widom, J. (2009). "Swoosh: A generic approach to entity resolution". *The VLDB Journal , 18* (1), 255-276.

The Role of Asserted Resolution in Entity Identity Information Management

Yinle Zhou and John R. Talburt
Information Science Department, University of Arkansas at Little Rock, Little Rock, Arkansas, USA

Abstract – *This paper introduces the concept of asserted resolution as a technique for entity resolution. In asserted resolution trusted information sources are used to force the equivalence (or non-equivalence) of entity references and identity structures regardless of matching conditions. The paper proposes five specific forms of assertion to support entity identity information management, the process of storing and maintaining identity information in an information system so that references to an entity can be recognized and labeled with the same identifier over time. The paper also gives a description of the way in which asserted resolution is being implemented in the OYSTER open source entity resolution system.*

Keywords: entity resolution, entity identity information management, entity identity structure, asserted resolution, OYSTER open source system

1 Background

Entity Resolution (ER) is the process of determining when two records in an information system refer to the same or to different real-world objects [1]. When two records are determined to refer to the same object they are said to be *equivalent references*. ER is foundational to several important information system processes where different data sources or different information systems provide information for a common set of entities such as entity-based data integration (EBDI) [2], master data management [3], and information exchange systems [4].

The term entity in ER describes the real-world object, such as a person, product, place, or event that has an individual identity. The identity of an entity is a set of attribute values for that entity along with a set of distinct rules with which the entity could be distinguished from all other entities of the same class in a given context [5].

ER is sometimes called record matching because all ER systems to some extent rely upon the method of direct matching to resolve references. Direct matching rules consider the degree of similarity between the values of corresponding identity attributes in two entity references. The simplest form of direct matching is deterministic matching, a method in which two references are considered equivalent if and only if all corresponding pairs of identity attributes have identical values [6] otherwise they are judged to be references to different entities. Deterministic matching is often not effective because it tends to produce too many false negatives, i.e. equivalent references that do not satisfy deterministic matching rules because of errors or variations in identity attribute values.

For this reason another form of direct matching called probabilistic matching is often employed. Probabilistic matching rules allow some attributes to have different values as long as the values of certain other attributes are the same. Probabilistic matching incurs a certain amount of risk of a false positive resolution, i.e. determining that references are equivalent when they in fact refer to different entities. A common extension of probabilistic matching which is also called Approximate String Matching (ASM) [7] that allows for intermediate levels of similarity between rules, i.e. accounting for the fact that attributes values may not be the same but are somewhat similar. For example, allowing names values to differ by a limited number of characters or strings judged to be phonetically similar.

In addition to direct matching, there are also three other resolution techniques that can be used to support equivalence decisions. These are

- Transitive resolution
- Relationship resolution
- Asserted resolution

Transitive resolution is a way to establish equivalence between references that do not satisfy direct matching rules. It does this by building a chain of intermediate equivalent references. Transitive resolution says that if reference A is equivalent to reference B, and reference B is equivalent to reference C, then reference A must also be equivalent to reference C. Transitivity is a necessary property of an ER system for it to be a consistent ER process, i.e. produces a unique result [8].

Most commercial and open source ER systems rely entirely upon direct matching and transitive resolution to determine which records in an input file comprise equivalent references. However recently there has been growing interest two additional methods, relationship resolution and asserted equivalence. Relationship resolution is a method for resolving a group of two or more references that conform to a particular pattern of relationships. Often borrowing from graph and network theory, algorithms such as SCAN [9] have been developed to perform ER using relationship resolution. A pattern of relationships does not necessarily require all of the relationships between pairs of references to be equivalence. In many cases they can simply be a partial match perhaps only involving a single attribute. For example a pattern of relationships among four customer records where two pairs share the same address and two pairs share the same name might be used to establish the equivalence of pairs sharing the same name even though they do not agree on address.

Direct matching, transitive resolution, and relationship resolution are all characterized by the fact that the evidence for establishing equivalence is internal, i.e. it comes from references being resolved, either from the values of reference attributes or from relationships between the references. These three forms of resolution are sometimes called inferred resolution or inferred linking [10].

2 Asserted Resolution

Asserted resolution is the use of a trusted information source which asserts or declares that references or identity structures are equivalent (or not equivalent). Asserted resolution is distinguished by the fact that it employs a priori knowledge from an external source to determine whether two references are equivalent or not. For this reason asserted resolution is also called knowledge-based resolution, and ER systems that use this method of resolution are sometimes called knowledge-based ER systems [1].

There are many sources of asserted information. Often it is self-reported, but may also be obtained from public records or commercial data providers, such as magazine subscription services. Asserted resolution is particular important for ER systems that create and store identity information in persistent identity structures as they resolve references.

Asserted resolution is essentially a method for overriding other resolution methods, especially direct matching rules. If two references are known to be equivalent, then they can be asserted as equivalent even though they do not necessarily satisfy a particular matching rule. For example two customer records with different last names and different addresses, but know to be equivalent based on information reported directly by the customer.

Conversely, if it is known that two references are not equivalent, they can be (negatively) asserted as non-equivalent even though they satisfy a matching rule. For example twins attending the same school with the same last name, date-of-birth, and first names similar enough that that they meet an approximate string matching threshold.

3 Identity Information Management

Entity identity information management (EIIM) is the process of creating and maintaining persistent data structures which represent the identities of the external entities. In this way, references to the same entity can be recognized and labeled with the same identifier over time, i.e. able to maintain persistent entity identifiers. ER systems that perform EIIM store and manage identity information in an entity identity structure (EIS).

ER systems that support EIIM typically run in one of three modes, Identity Capture Mode, Identity Resolution Mode, and Identity Update Mode. Identity capture is used to create an initial set of EIS from input references. Identity resolution begins with a given set of identities to which input references are to be resolved, but does not create new identities or alter existing identities. Identity update starts with existing identities, but can create new identities and update existing identities with information from the references being processed.

Knowledge-based ER systems can use asserted equivalences to enhance their EIIM processes. Asserted knowledge gives a much higher confidence to the identities created and updated from these sources and also allows for "fine-tuning" of the system by overriding direct matching rules in exceptional cases.

Knowledge-based ER systems also have the sequence neutral property [11]. Because with reference to reference assertion, asserted knowledge is acquired and provisioned to the EIM process prior to processing other input references, references resolved through assertions can be recognized as equivalent at the time they are processed, regardless of the order in which they were received.

The main disadvantage of asserted resolution is the increased storage and processing effort required to use it effectively. A typical ER process supporting asserted resolution is usually divided into two parts, a foreground process for resolving equivalence and a background process of integrating assertion into the entity identity management system. Figure 1 shows the principal components of an ER system that supports asserted resolution for EIIM.

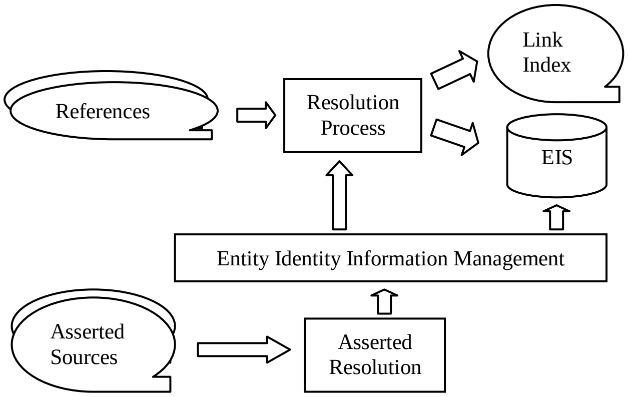


Figure1: An ER system Using Asserted Resolution

Assertion sources often have different layouts and attributes than typical reference files. If the ER system requires that all input sources must have a common layout and common attributes, then it is more difficult to implement assertion. Most ER systems can be made to handle simple reference-to-reference assertion by simply defining a separate field with an assertion link value. The potential problem in this arrangement is that without separate logic for assertion and inferred resolution there can be interference between them, i.e. inferred resolutions overriding asserted resolutions and vice versa. It also doesn't address asserted resolution to and among EIS that are essential to efficient EIIM.

There are a few of ER systems that are designed to handle certain type of assertion directly. AbiliTec® [12] developed by Acxiom Corporation is an example of a commercial ER system based on asserted resolution. AbiliTec® is a Customer Data Integration (CDI) platform which has made reference to reference assertion as it primary resolution process. AbiliTec® manages billions of assertions for U.S. customers alone.

4 Entity Identity Structure Model

The following is a proposed model for describing EIS that builds on the algebraic models for ER [13] and Entity-Based Data Integration [14] [2].

Definition 1: An Entity-Based Information System is an ordered triple S= (E, Q, D) where

- E is a finite set of Entities
- Q={q_1, q_2 , … , q_n} is a finite set of n attributes that describe the entities in E
- D = {D_1, D_2, …, D_n}where each D_j is the domain of values for the attribute q_j for j=1,…,n.

Definition 2: The Span of Attribute q_j represented by P_j is the set of all attribute-value pairs for attribute q_j i.e. $P_j = \{q_j\} \times D_j = \{(q_j, v) \mid v = D_j\}$

Definition 3: The Span of S represented by P is the set all attribute-value pairs in S, i.e.

$$P = \bigcup_{j=1}^{n} P_j = \{(q_j, v) \mid q_j \in Q \text{ and } v \in D_j\}$$

Definition 4: An Entity Identity Fragment f in S is any non-empty subset of P, i.e. f ⊆ P, where P ≠ ∅.

Note that a fragment is only required to contain at least one attribute-value pair from P. There is no requirement that every attribute in Q is represented in f, and there is nothing to prevent the same attribute from occurring more than one time with different values.

Definition 5: An Entity Identity Structure (EIS) λ in S is any non-empty set of Identity Fragments, i.e. λ = {f_1, f_2, …, f_m} where each $f_k \subseteq P$ and $f_k \neq \emptyset$, for k=1,…,m.

To illustrate these concepts, consider the following example in the context of magazine subscribers described by their first and last name and address. In this case let S= (E, Q, D) where

E= {Customer_1, Customer_2, Customer_3}

Q= {First, Last, Addr}

D= {D_{First}, D_{Last}, D_{Addr}}

D_{First} = {Mary}, D_{last}= {Smith, Jones}, D_{Addr} = {3 Pine St, 1 Oak St, 2 9th St}

P_{First} = {(First, Mary)}

P_{Last} = {(Last, Smith), (Last, Jones)}

P_{Addr} = {(Addr, 3 Pine St), (Addr, 1 Oak St), (Addr, 2 9th St)}

P= $P_{\text{First}} \cup P_{\text{Last}} \cup P_{\text{Addr}}$

Note that the total number of possible entity identity fragments in S is 2^8-1 = 255. The inputs to the system are the following reference fragments

f_1= {(First, Mary), (Last, Smith), (Addr, 3 Pine St)}

f_2= {(First, Mary), (Last, Smith), (Addr, 1 Oak St)}

f_3= {(First, Mary), (Last, Jones), (Addr, 2 9th St)}

If the ER process were to determine that these three input fragments refer to same identity, then an identity capture process would build an EIS for them. There are many in which these could be integrated into a single structure. Two of the most common are record-based EIS and attribute-based EIS. A record-based EIS is simply the union of the record fragments, i.e.

λ_R= $f_1 \cup f_2 \cup f_3$
= {{(First, Mary), (Last, Smith), (Addr, 3 Pine St)},
{(First, Mary), (Last, Smith), (Addr, 1 Oak St)},
{(First, Mary), (Last, Jones), (Addr, 2 9th St)}}

On the other hand in an attribute-based EIS, all attribute values of the same attribute are collected into the same fragment, i.e.

λ_A = {{(First, Mary)}, {(Last, Smith), (Last, Jones)},
{(Addr, 3 Pine St), (Addr, 1 Oak St), (Addr, 2 9th St)}}

Even though the attribute-based EIS is more compact, it is not necessarily the best choice for all situations. The choice of EIS structure can have an impact on the ER outcome.

5 Proposed Forms of Assertion

This section describes five forms of asserted resolution that are needed to effectively perform EIIM. These are reference-to-reference assertion, reference-to-structure assertion, structure integration assertion, structure split assertion, and negative structure-to-structure assertion.

5.1 Reference-to-Reference Assertion

Reference-to-reference assertion often occurs in an ER context where some pairs of equivalent references will have insufficient evidence available within the references sources themselves to make the determination by inferred resolution methods, thereby leaving them as false negatives. For example, in Table 1 there are three references Mary Smith at 3 Pine St, Mary Smith at 1 Oak St, and Mary Jones at 2 9th St which represent a same entity in real-world. Because they have different names and addresses these references would typically not be determined equivalent by direct matching with pre-defined identity equivalent rules (PIER). Unless there were other references in the same input file that provided a chain of equivalence or a pattern of relationships between these references, they would be considered references to distinct entities.

Table 1: references to be resolved

ID	First Name	Last Name	Address
1	Mary	Smith	3 Pine St
2	Mary	Smith	1 Oak St
3	Mary	Jones	2 9th St

However if this same ER process had access to information from a magazine subscription service that asserted that a person of this name had reported a change of address notice where the reported addresses agreed with the two addresses on the customer records, then the system could decide that the two records actually are equivalent references. An asserted equivalence often is represented by a special attribute that the references to an entity are assigned the same values, as shown in Table 2. Attribute Assert is used to tag the same entities. References 1, 2, and 3 refer to one entity by Assert value 1. References 4 and 5 refer to another entity by Assert value 2.

Table 2: references with asserted information

ID	First Name	Last Name	Address	Assert
1	Mary	Smith	3 Pine St	1
2	Mary	Smith	1 Oak St	1
3	Mary	Jones	2 9th St	1
4	Rose	Smith	3 Pine St	2
5	Rose	Smith	1 Oak St	2

Reference-to-reference assertion is primarily intended as a way to initialize an identity knowledgebase for identity resolution operations or to add new identities to an existing knowledgebase when PIER might be problematic. References in Table 2 will two EIS with different identifiers.

5.2 Reference-to-Structure Assertion

Reference-to-structure assertion can be used when an EIS has been built in a previous ER process and one or more new input references are asserted as equivalent to this EIS. Reference-to-structure assertion bypasses PIER to force references to a specific EIS.

For example, consider the attribute-based EIS λ_A described in Section 4. If in the input fragment

f_4 = {(First, Marife), (Last, Smith), (Addr, 3 Pine St)}

were to be asserted to λ_A, then the resulting integration structure would be

$\lambda_{A'}$ = {{(First, Mary), (First, Marife)}, {(Last, Smith), (Last, Jones)}, {(Addr, 3 Pine St), (Addr, 1 Oak St), (Addr, 2 9th St)}}

5.3 Structure Integration Assertion

Structure integration assertion occurs in the ER context that references are under consolidated. The existing EISs are asserted to be one structure. For example, there are structure λ_1 "ABCDEFGH" and structure λ_2 "XYZWOPQ",

λ1 = {{(First, Mary)}, {(Last, Smith), (Last, Jones)}, {(Addr, 3 Pine St), (Addr, 1 Oak St), (Addr, 2 9th St)}}

λ2 = {{(First, Mary)}, {(Last, Jones)}, {(Addr, 4 119th St)}}

λ1 and λ2 are asserted to be one entity and integrated to λ1. After integration, λ1 is updated to {{(First, Mary)}, {(Last, Smith), (Last, Jones)}, {(Addr, 3 Pine St), (Addr, 1 Oak St), (Addr, 2 9th St), (Addr, 4 119th St)}}.

Structure integration assertion is primarily intended to force structures to combine when there is credible external knowledge that they are equivalent.

5.4 Structure Split Assertion

Structure split assertion occurs in the ER context that references have been over consolidated, i.e. non-equivalent references have been integrated into the same EIS. PIER will almost never be 100% accurate in their resolution. The structure split assertion is an aid in fine-tuning the resolution process by correcting a small number of false positive resolutions made by the PIER rather than changes the PIER.

Splitting an EIS is much easier to manage for record-based EIS because record integrity (provenance) has been preserved. For example consider EIS λ_R from Section 4. If it were determined that the first reference to Mary Smith at Pine Street was not equivalent to the other two references, then a split assertion could be used to split λ_R into two EIS λ_1 and λ_2 of the form

λ_1= {{(First, Mary), (Last, Smith), (Addr, 3 Pine St)}}
λ_2= {{(First, Mary), (Last, Smith), (Addr, 1 Oak St)}, {(First, Mary), (Last, Jones), (Addr, 2 9th St)}}

5.5 Negative Structure-to-Structure Assertion

Negative structure-to-structure assertion is used when it is known that PIER would consolidate two EIS that are known to be non-equivalent. In essence negative structure-to-structure assertion is a preventative measure for the structure split assertion that is applied after the fact. Again the value of type of assertion is fine-tuning the PIER by addressing certain exceptions through negative assertion rather than a change in the PIER.

An example is the problem of resolving twins in student records. Twins often attend the same school, have the same teachers, the same date-of-birth, the same last name, and often very similar first names. The granularity of PIER rules necessary to discriminate between twins can potentially create a disproportionally large number of false negative resolutions. In these cases negatively asserting known twins will allow more latitude in the PIER resolution.

Negative structure-to-structure assertion can be easily implemented by cross-referencing the EIS in the metadata section of the EIS. An example of this can be seen in the next section describing asserted resolution in the open source system OYSTER.

6 Asserted Resolution in OYSTER

OYSTER (Open sYSTem Entity Resolution) is an open source project from the Center for Advanced Research in Entity Resolution and Information Quality (ERIQ) at the University of Arkansas at little Rock. The OYSTER source code and documentation are freely available from the Center's website at ualr.edu/eriq. Originally designed to support education and research in entity resolution, OYSTER can be configured to run in several modes of operation including merge-purge/record linking, identity capture, identity resolution, and identity update. The resolution engine supports probabilistic direct matching, transitive resolution, and asserted resolution. A key feature of the system is that entity and reference-specific information including identity rules is interpreted at run-time through user-defined XML scripts which gives the user the flexibility to configure and manage its operation.

An OYSTER run requires a run script in XML, an attribute script in XML, source descriptor(s) in XML, input identities in text file(s) (this is only needed when running in identity resolution mode or identity update mode), and the input references in text file(s) or from database(s). The run script stores the paths to other scripts and files. Each reference source has a corresponding XML source descriptor that describes the file or database path, record delimiter, record layout, and other features of the input source. The run script also gives the path to the attribute script that defines the identity attributes and identity rules (PIER) that apply to all input sources.

OYSTER is unique in that it has built-in logic for processing assertion sources. The latest version 3.1 supports reference-to-reference assertion logic and support for the other types of assertion described in the previous section are in currently being implemented and tested.

For input references with asserted information as shown in Table 2, items labeled with the attribute "@RefRefAssert" are recognized by OYSTER as assertion links. Figure 2 shows part of the source descriptor for the assertion input in Table 2. It defines the fifth column (in ordinal position) as the asserted attribute where "@RefID" is the OYSTER keyword for reference ID.

```
<OysterSourceDescriptor Name="Assertion">
 …
<ReferenceItems>
 <Item Name="RefID" Attribute="@RefID" Pos="0"/>
 <Item Name="First Name" Attribute="FN" Pos="1"/>
 <Item Name="Last Name" Attribute="LN" Pos="2"/>
 <Item Name="Address" Attribute="Addr" Pos="3"/>
 <Item Name="Asserted" Attribute="@RefRefAssert" Pos="4"/>
 …
</ReferenceItems>
</OysterSourceDescriptor>
```

Figure 2: Part of a Source Descriptor for Asserted Input

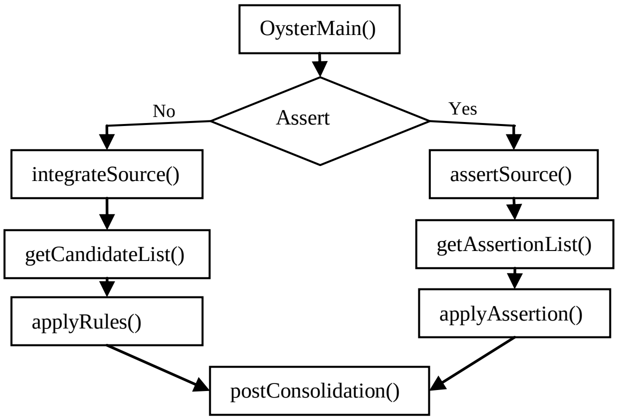


Figure 3: Logic Flow for Reference-to-Reference Assertion

When processing an input source, OYSTER will first check if the input file is a reference-to-reference assertion source. If it is not an assertion input, OYSTER will perform normal resolution on the input using direct matching (according to the PIER provided) and transitive resolution. If it is a reference-to-reference assertion source, OYSTER will only build and update entity identities based on the asserted identifiers. Other attribute values such as name values are passed to the identity structure and index, but they do not participate in resolution or consolidation decisions. Figure 3 shows the work flow of main methods in OYSTER that handle reference-to-reference assertion.

The EIS in OYSTER are described in XML. As an output file, the identity created from the References 1, 2, and 3 in Table 2 is shown in Figure 4.

```
<root>
<Metadata>
…
        <Attributes>
            <Attribute Name="@RefID" Tag="A"/>
            <Attribute Name="FN" Tag="B"/>
            <Attribute Name="LN" Tag="C"/>
            <Attribute Name="Addr" Tag="D"/>
        </Attributes>
</Metadata>
    <Identity Identifier="UOS0NA7ANSVM4WBG">
        <References>
            <Reference Value="A^Source.1|B^Mary|C^Smith|D^3 Pine St"/>
            <Reference Value="A^Source.2|B^Mary|C^Smith|D^1 Oak St"/>
            <Reference Value="A^Source.3|B^Mary|C^Jones|D^2 9th St"/>
        </References>
    </Identity>
    …
</root>
```

Figure 4: An EIS Create from Table 2

The element <Metadata> contains gives a mapping of attributes to alias tags that allow for compression of the XML representation of the EIS. Each EIS is assigned a unique 16 character identifier given by the element <Identity> called the OYSTER Identifier or OID. For example in Figure 4, the EIS labeled with the OID value "UOS0NA7ANSVM4WBG" has been created from three references. Input reference fragment

{(ID, 1), (First, Mary), (Last, Smith), (Addr, 3 Pine St)}

Is represented in the EIS as

"A^Source.1|B^Mary|C^Smith|D^3 Pine St".

For reference-to-structure assertion, OYSTER will use the keyword "@RefStrAssert" as the attribute label that will indicate to which EIS the reference should be joined. For structure integration assertion, the input assertion file should contain two fields that contain the identifiers of two existing EIS that should be integrated (merged)..

For structure split assertion, the input assertion file should contain three items, a reference ID, an OID, and an assertion group label. The references in the designated EIS will be extracted from the EIS and used to create a new EIS with a new label. The references with same assertion group values will all be go into the same newly created EIS.

For negative-structure-to-structure assertion, the input assertion file contains two OID values indicated which two EIS should not be integrated in future ER processes. To achieve this, a new element <NegStrStr> will be added to the <Metadata> of the EIS. For example, if two structures labeled "ABCDEFGH" and "XYZWOPQR" are negatively asserted, they will be cross linked. The <NegStrStr> in the two EIS will be updated. In EIS "ABCDEFGH", it will appear as

<NegStrStr><OID name="XYZWOPQR"></NegStrStr>.

In EIS "XYZWOPQR", it will appear as

<NegStrStr><OID name= "ABCDEFGH"></NegStrStr>.

7 Conclusion

Asserted resolution can add significant capability to ER systems that typically rely only upon direct matching and transitive resolution for their operation. Asserted resolution provides a way to supplement knowledge inferred from the input sources with external knowledge from trusted sources of information. By working in concert, inferred and asserted resolution techniques have the capability of provided more accuracy ER results.

Asserted resolution also plays a key role in support of EIIM. The reference-to-reference assertion implemented in OYSTER provides a simple method for users to describe and load known identities into an identity resolution ER systems.

The reference-to-structure, structure integration, structure split, and negative structure-to-structure assertions described in this paper have the potential to provide important new capabilities in the long-term management of identities in EIIM-capable ER systems. This family of assertions enhances the capability PIER-only resolution systems by allowing users to fine-tune PIER results by adjusting or overriding inferred resolutions thereby providing for more effective management of entity identities over time.

8 Acknowledgement

The work reported in this system was partially supported by grants from Arkansas Department of Education and SAS Software.

9 References

[1] John Talburt. “Entity Resolution and Information Quality”. Morgan Kaufmann, Burlington, MA, 2011.

[2] Greg Holland and John Talburt. “An entity-based integration framework for modeling and evaluating data enhancement products”; Journal of Computing Sciences in Colleges, 24, 5, 65-73, 2009

[3] David Loshin. “Master Data Management”. Morgan Kaufmann, 2008

[4] Key findings of the eHealth Initiative. (n.d.). Retrieved August 9, 2010, from eHealth: eHealth (2010) Key Findings. http://www.ehealthinitiative.org/key-findings.html

[5] Ee-Peng Lim, Satya Prabhakar, Jaideep Srivastava, and James Richardson. “Entity identification in database integration”; Ninth International Conference on Data Engineering, 294-301, 1993

[6] Thomas Herzog, Fritz Scheuren, and William Winkler. “Data Quality and Record Linkage Techniques”. Springer, 2007

[7] Gonzalo Navarro. “A Guided Tour to Approximate String Matching”; ACM Computing Surveys, 33, 1, 31-88, 2008

[8] Omar Benjelloun, Hector Garcia-Molina, David Menestrina, Qi Su, Steven E. Whang, and Jennifer Widom. “Swoosh: A Generic Approach to Entity Resolution”. The VLDB Journal, 18, 1, 255-276, 2009

[9] Xiaowei Xu, Nurcan Yuruk, Zhidan Feng, and Thomas Schweiger. “SCAN: A Structural Clustering Algorithm for Networks”; The Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007

[10] John Talburt, Yinle Zhou, and Savitha Shivaiah. “SOG: A Synthetic Occupancy Generator to Support Entity Resolution Instruction and Research”; Proceedings of the 14th International Conference on Information Quality. 91-105. Potsdam, Germany, 2009

[11] Jeff Jones. “Non-obvious Relationship Awareness (NORA)”. IBM Entity Analytics Solutions Presentation. Las Vegas, NV, 2005.

[12] Acxiom. From http://www.acxiom.com/products_and_services/cdi/Recognition/AbiliTecandAbiliTecDigital/Pages/AbiliTecandAbiliTecDigital.aspx. Retrieved on March 11, 2011.

[13] John Talburt, Richard Wang, Kimberly Hess, and Emily Kuo. “An Algebraic Approach to Data Quality Metrics for Entity Resolution over Large Datasets.” In Information Quality Management: Theory and Applications, 1-22. Hershey, PA: Idea Group Publishing, 2007.

[14] John Talburt and Ray Hashemi. “A Formal Framework for Defining Entity-Based, Data Source Integration.” 2008 International conference on Information and Knowledge Engineering. Las Vegas, NV, 2008. 394-398.

Restructuring medical package leaflets to improve knowledge transfer

Fabian Merges, Madjid Fathi
Department of Computer Science and Electrical Engineering,
University of Siegen
Hölderlinstrasse 3, D-57068 Siegen, Germany
fabian.merges@uni-siegen.de

Abstract — **This document presents a readability test designed to optimise the readability of package leaflets for medicinal products. Patients often find leaflets incomprehensible, which in part explains the high percentage of medications disposed of before being taken. So patient information must be formulated that an averagely educated user can understand its content and follow its instructions. The paper discusses the existing methods and the additional properties that the new readability method offers. A newly developed catalogue of criteria forms the foundation for the new method. In addition, the problems associated with current package leaflets are discussed. A medical thesaurus translate technical terms into their colloquial equivalents. To verify the new method there will be an evaluation with 400 test subjects. The paper shows also shortly the readability project in the context of previous projects of the Knowledge Management and Medical Engineering group (KMME) and how these projects can be combined to improve knowledge transfer.**

Keywords: readability test, readability indices, knowledge transfer, readability criteria, medical thesaurus, method evaluation

I. Introduction

In Germany, according to the Rote Liste, there are about 8500 product entries [1]. During 2008, medicines with a total value of about 40.7 billion euros or 1.5 billion packs were dispensed to patients by pharmacies. Given the constantly rising age of the population as a whole, this figure is likely to increase. Ten per cent of the medications sold in 2008, with a value of about 4 billion euros, were disposed of by patients even before taking them because they found the patient information incomprehensible. [2] After medical practitioners and pharmacists, the most important source of information for patients is the patient information [3]. In view of its significance, patient leaflets must not be written in such a way that a patient feels at a loss, fails to take a medication properly or neglects to take it altogether. The patient information must be so formulated that an averagely educated user can understand its content and follow its instructions. As investigations show, however, often this is not the case, which in part explains the high percentage of medications disposed of before being taken.

As a result of the many regulations issued since the German Medicines Act (AMG) came into effect on 1 January 1978 [4] patient information on medicines has become ever more extensive and complex, such that those for whom the information is written, namely the patients, often no longer understand it.

The scientific institute of the health insurer AOK showed the package leaflets of the 100 most commonly prescribed medicines to 70 test subjects. They were asked to assess the leaflets for readability and comprehensibility. In addition, 1900 subjects were surveyed on the topic of package leaflets. [5]

According to the results, the overwhelming majority of consumers do read the patient information before taking the medicine, but the information contained in it is most patients regard the information as not very helpful. 28% of those surveyed had at least once in the past discontinued a medication or not even started taking it because of the package leaflet. Almost every other person surveyed thinks that package leaflets are too long, and one in five finds them incomprehensible.

In September 2005, by way of the 14th Amendment to AMG, the German Bundestag stipulated in § 22 (7) that patient information on medicines must be comprehensible, and to demonstrate this comprehensibility a readability test must be performed with the participation of a patient target group. [6]

The situation abroad is not that much different. Here too, patients are mostly overtaxed by package leaflets, so that for example the US Food and Drug Administration (FDA) provides patients with further information via its web portal. [7] As shown in an FDA press report dated January 2006, the FDA also provides support for the optimisation of patient information, for example a guideline for the labelling of prescription-only medications. It does point out, however, that these are only suggestions and not mandatory regulations. [8]

Since then, efforts have been made on the basis of these statutory provisions to optimise patient information. But there is no system in place for well-founded, scientific examination of the texts on the basis of linguistic criteria. Instead, a patient test (pilot test) is performed in the first step, in order to identify the problems associated with a package leaflet. The results of this test, the statutory provisions, the guidelines of the BfArM (German Federal Institute for Drugs and Medical Devices) and the tester's experience and assessments form the basis for a subsequent redesign. On completion of the revision phase, the amended package leaflet is tested again and if necessary revised once more. Instead of working with a standardised method, the process is one of trial and error. [9]

The revision itself is always carried out manually. In other words, a reviser has to adapt patient information without computer-supported aids. The revision is based on the first systematic readability tests, which were performed in Australia. The procedure recommended there, which serves as a standard for the testing of package leaflets, was described in detail by Sless and Wiseman in their book "Writing about medicines for people". [9]

Over the next two years, the Institute for Knowledge-based Systems at the University of Siegen will develop an electronic readability test with a view to significantly improving the readability of package leaflets. The development will be carried out by the Knowledge Management and Medical Engineering group (KMME) at the Institute for Knowledge-based Systems. In the past, with the help of knowledge-based methods such as content-based filter methods and recommender systems, it has set up medical portals for patient information and expert systems for medical practitioners. [10]

Below, we describe briefly the readability project of the Institute for Knowledge-based Systems at the University of Siegen, and provide an overview of the project's phases.

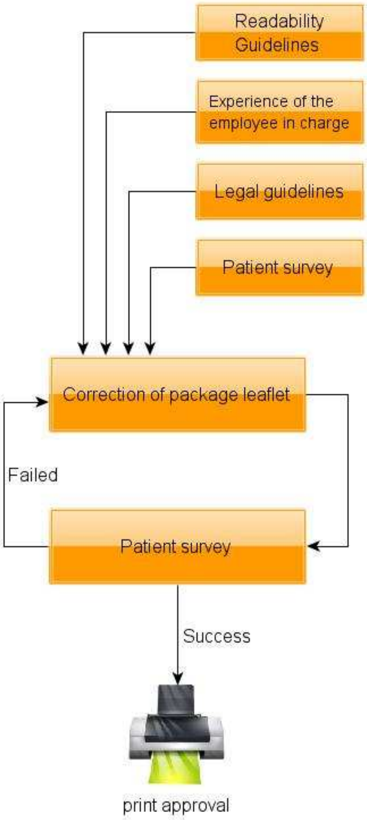


Figure 1. Trial and error

II. Current Readability Methods

Current methods compute the indices of a text. The index is meant to provide data about the readability of a text. Existing methods from linguistic sciences are relied upon. These include the Flesch Reading Ease method [11, 12], the Flesch-Kincaid Grade Level [11], the Gunning Fog Index [13] and the Wiener Sachtextformel (Vienna non-fiction formula) [13]. These are also combined, as in Phillips *et al.* [15] Thus for example, Comlab in its TextLab product combines the various indices in order to draw conclusions about the readability of a text. [16] When computing the existing readability indices, static parameters such as e.g. the number of syllables are collected, placed in a relation to each other and a readability index calculated. However, this gives a reviser only a rough indication of the extent to which the text requires revision. The current methods do not offer concrete help in respect of individual sentences or paragraphs, e.g. those of a package leaflet. As a rule, revisers will fall back on the readability method of Sless and Wiseman [9], but this does not exist in electronic form. The readability test presented here is designed to show the user specific errors, sentence by sentence, thus offering revisers far more support in their work.

The use of a readability index would seem to be especially useful where a large number of texts are to be sorted by incomprehensibility. This allows a reviser to search out from the mass of texts those in most urgent need of revision.

III. Readability Criteria

The preparation of a package leaflet is based on readability criteria. They offer a means of measuring and assessing the readability of an existing leaflet. Such a catalogue of criteria would have to be drawn up during the first phase of the project. In linguistic sciences, investigations have been conducted into various sub-fields such as e.g. font styles, in order to establish the degree to which they affect readability. These findings need to be collated so that an informed catalogue of criteria can be compile from them. A first draft of criteria structuring has been drawn up, in conjunction with an application for funding, which is currently being reviewed. According to this approach, the first level distinguishes between formal, typographical, visual and linguistic criteria. The following levels break down the criteria further, e.g. into sub-criteria that influence the macrostructure of the text. This includes, amongst others, font size and style. Often, a compromise needs to be reached between the requirements of various criteria. For example, the font size must remain legible for older patients and those with mildly impaired vision. On the other hand, the entire text needs to be printed on a single package leaflet, the size of which is predetermined and limited.

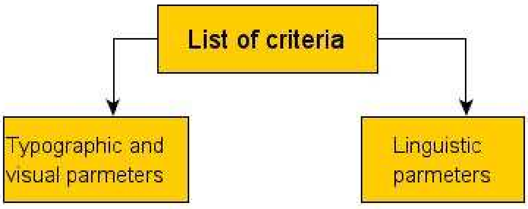


Figure 2. Macrostructure of the readability test

IV. Readability Method

The objective of the planned readability method is to prepare a text in such a way that it can be optimised easily by a reviser. It is not meant to replace revisers, but to support them in identifying, rapidly and systematically, at which points in the text the information is not being conveyed to the recipient. This is done by running the text through several phases. In the first phase the text is prepared for the subsequent phases. With the help of a dedicated parser, it is broken down into paragraphs, sentences, lists and tables. Additional metadata need to be collected, so that the individual parts can be related to the overall text. Further metadata, such as the number of words per sentence or paragraph, facilitate the execution of subsequent analytical steps. These analytical steps are performed in the second phase and rely on different partial aspects of computer linguistics. In-depth comparative analysis is needed in order to show which aspects these will be in each particular case.

V. Medical Thesaurus

For the users of package leaflets, foreign words are the greatest obstacle to understanding. The scientific institute of the German health insurer AOK has shown the package leaflets of the 100 most commonly prescribed medicines to 70 test subjects. They were asked to assess the texts for readability and comprehensibility. In addition, 1900 subjects were surveyed on the subject of package leaflets. [5]

The study shows unequivocally that it is impossible to improve readability without replacing foreign words. Thus, slightly more than one quarter of those surveyed had discontinued a medicine or not taken it at all, because the enclosed patient information left them feeling at a loss. The Institute for Knowledge-based Systems intends to solve this problem of technical terminology in medical texts in two ways.

The first step involves cooperation with MTW Healthcare GmbH. The company has produced a medical thesaurus containing more than 2500 technical terms and published it in electronic form. [17] It is shown in Fig. 3. These terms can be called via interfaces and made available for a readability test. The technical terms themselves were translated by experts into their colloquial equivalents.

In the second step, the present capabilities of the thesaurus will be expanded to pursue the basic idea of identifying for the reviser expressions that are difficult to understand. Building on a corpus of newspaper articles from German-language daily papers, content-based filter methods will be used to decide whether a word in a patient leaflet text is comprehensible to an average user. Methods such as TF-IDF [18, 19] could be used here as a basis. The results of such analysis could be used to provide a reviser with a first impression of which passages in a text require revision most urgently.

VI. Evaluation of Readability Method

The development of a text method for readability and its availability in the form of text processing software will be followed by an evaluation phase. The patient information optimised with the aid of the text processing software should be tested for readability in patient interviews based on a questionnaire, in order to demonstrate the validity of the method developed. The requirements and criteria for success laid down in the EU readability guideline [20] are to serve as target parameters. The concept of readability has two aspects that will be reviewed in the evaluation: findability of information in the text and comprehensibility of wording.

In the practical test, 400 test subjects will be surveyed (20 subjects each on a total of 20 patient leaflets). Both aspects "findability" and "comprehensibility" of information will be checked with reference to 12-15 questions on a patient leaflet. The questions will focus on treatment-relevant topics, such as indications or side effects of a medicine.

The outcome is likewise to be measured based on the requirements laid down in the EU readability guideline. This states that "a satisfactory test outcome for the method outlined above is when the information requested within the package leaflet can be found by 90% of test participants, of whom 90% can show that they understand it. That means to have 16 out of 20 participants able to find the information and answer each question correctly and act appropriately". [20]

In her position paper of February 2011, Co-ordination Group for Mutual Recognition and Decentralised Procedures - Human set up by the European Medicines Agency (EMA) supports this procedure. [21] Described as the Australian method, this procedure is attributed to Sless and Wiseman. [9] Even if this method is virtually the standard in patient tests, the position paper clearly states that other types of patient survey would be possible. Instead of having patients answer questions verbally, they could also write down their answers. [21]

As things stand at present, the patient survey in in the evaluation phase is carried out using the Australian method. This has the advantage that, if necessary, any tests carried out without an improvement in readability as a result of the new method can be compared with tests after the new method has been applied.

If the tests were not successful, they would have to be used as a starting point for a subsequent adjustment of the readability tests. A repeat test would then be necessary to show the effectiveness of the measures.

VII. Readability Test in the Context of Previous Projects

In the past, the KMME group was occupied amongst other things with the development of portal systems for patients and doctors. In these cases, there was always a strict separation between these two groups. For example, doctors often work with sensitive data that have to be kept safely for data protection reasons. In portal systems, these data also have to be de-identified. [10] Figure 4 shows the structure of the stroke portal 'StroPoS' as an example.

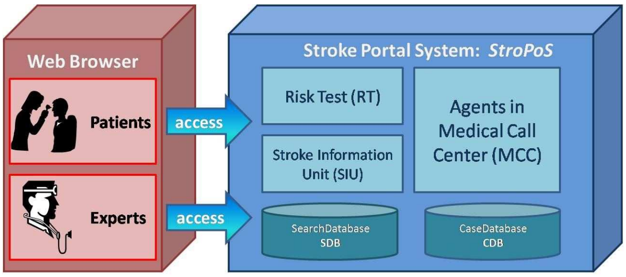

Web Browser
Patients
Experts
access
access
Stroke Portal System: StroPoS
Risk Test (RT)
Stroke Information Unit (SIU)
Agents in Medical Call Center (MCC)
SearchDatabase SDB
CaseDatabase CDB

Figure 3. Architecture of the stroke portal StroPoS

Knowledge is communicated in all fields, except the Medical Call Centre (MCC), via websites. This is common to all portal solutions. In the interplay with the readability method the aim should be to review these websites retrospectively for their readability. With the medical partners they have so far only been checked for the correctness of their content.

Two technical tasks must be resolved for this. The readability test must be able to parse not only Word documents correctly, but also websites. Only then can there be a retrospective analysis of the existing portals. Yet the aim should not only be to improve existing texts, but also to check new texts directly for their legibility. An extension of the readability test to include web services would be conceivable, so that present and future portal solutions can access the readability analysis.

VIII. Conclusions and Outlook

An automated system of generating and optimising patient information in particular and medical texts for lay persons in general means having a rapid, economical and effective method in place that can improve the reading compliance of end-users with regard to a medical text, for example a package leaflet, and make a substantial contribution to drug safety. Reducing patients' anxieties will help encourage them to take their medication correctly. Patients often avoid a medicine because they are afraid of side effects or interactions. The resulting losses to the economy are huge.

In 2008, for instance, about 10% of medications were disposed of before being taken because patients could not understand the information in the package leaflet. [2] Based on the 2008 figures, if it proved possible to optimise patient information through the project presented here in such a way that only 0.1% of patients took their prescribed medication instead of discarding it, then 40 million euros could be saved annually and thus relieve the burden on payers of health insurance premium.

The readability test of the Institute for Knowledge-based Systems is still in its early stages. Nevertheless, even now the possibilities and the way to the completion of the project are clear. The benefits not only to patients, but also to the healthcare system that can be derived from the savings highlighted here speak for themselves.

The readability test can also be extended to other fields without changing the criteria, simply by changing the range of values, which allows application to the texts of other products. An improvement in the readability of user instructions for consumer electronics, such as video recorders, would also be conceivable.

In an era of distributed applications, the readability test could also be offered as a cloud application service. Simple integration in existing application programs would be possible via web services.

References

[1] Rote Liste Service GmbH, "Rote Liste 2010"; Frankfurt am Main, p 9, 2010

[2] J. Bublies , "Angst vor Beipackzetteln", NRZ (Neue Ruhr / Neue Rhein Zeitung), 25.07.2009

[3] W.-D. Ludwig, "Arzneimittelsicherheit - Patientenrechte – Schutz vor Schaden"; Patientenforum Medizinethik, p 18, Tutzing 2007

[4] Gesetz über den Verkehr mit Arzneimitteln (Arzneimittelgesetz) in der Fassung der Bekanntmachung vom 12. Dezember 2005; §2 Arzneimittelgesetz

[5] K. Nink; H. Schröder, "Zu Risiken und Nebenwirkungen: Lesen Sie die Packungsbeilage?"; Scientific institute of the german health insurance company 'AOK', Bonn 2005

[6] BMGS: Vierzehntes Gesetz zur Änderung des Arzneimittelgesetzes vom 29. September 2005 § 22 Absatz 7

[7] U.S. Food and Drug Administration (FDA), "Index to Drug-Specific Information", Rockville (Maryland) 2011

[8] U.S. Food and Drug Administration (FDA)," Guidance for Industry and Review Staff - Labeling for Human Prescription Drug and Biological Products - Determining Established Pharmacologic Class for Use in the Highlights of Prescribing Information", Rockville (Maryland) 2009

[9] D. Sless, R. Wiseman, "Writing about medicines for people - Usability Guidelines for Consumer Medicine Information", 2nd edition, Canberra 1997

[10] M. Bohlouli, P. Uhr, S. M. Hassani, F. Merges, M. Fathi, "Practical Approach of Knowledge Management in Medical Science", IKE'10 - 9th International Conference on Information and Knowledge Engineering, Las Vegas 2010

[11] R. Baud; M. Fieschi; P. Le Beux, "The new navigators: from professionals to patients : proceedings of MIE2003"; p 657; Amsterdam 2003

[12] C. Finkbeiner, "Interessen und Strategien beim fremdsprachlichen Lesen"; pp 293-294, Tübingen 2005

[13] D. Graham, J. Graham, "Can Do Writing: The Proven Ten-Step System for Fast and Effective Business"; p 162; New Jersey 2009

[14] J. Schweizer; "Subprime, Hypotheken und faule Kredite – Alles klar?!: Wie verständlich ...", pp 38-39; Norderstedt 2009.

[15] T.J. Phillips, C.M. Daily, M.S. Luechlfing, "A note on the readability of professional materials for management accountants", Advances in Management Accounting, pp 311-318; Amsterdam 2007

[16] M. Sturmer, T. Holzinger; "Die Online-Redaktion: Praxisbuch für den Internet-Journalismus"; p 113; Berlin/Heidelberg 2010

[17] K. Menges, R. Schraitle, F. Merges, "Thesaurus medizinischer Fachbegriffe und deren umgangssprachliche Entsprechungen", MTW Verlag, Kleve 2009

[18] M. A. Russell; "Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites", pp 209-216, Sebastopol 2011

[19] Ali Khodaei, Cyrus Shahabi, Chen Li, "Hybrid Indexing and Seamless Ranking of Spatial and Textual Features of Web Documents", Database and Expert Systems Applications: 21st International Conference, pp 454-455, Berlin/Heidelberg 2010

[20] European Commission: "A Guideline on the Readability of the Label and the package Leaflet of Medicinal Products for Human Use", Revision 1, Brussels 2009

[21] Co-ordination Group for Mutual Recognition and Decentralised Procedures - Human, "Position paper on user testing of package leaflet - Consultation with target patient groups (Compliance with article 59(3) of Council Directive 2001/83/EC)", Doc. Ref: CMDh/234/2011; February 2011

Assessment of Hemoglobin Level of Pregnant Women Before and After Iron Deficiency Treatment Using Nonparametric Statistics

M. Abdollahian*, S. Ahmad*, S. Nuryani#, +D. Anggraini
*School of Mathematical and Geospatial Sciences, RMIT University, Melbourne, Australia
#Ulin Hospital (RSUD Ulin) Banjarmasin Indinesia
+Department of Mathematics Lambung Mangkurat University, Banjarbaru, Indonesia

Abstract
Iron Supplementation is generally recommended during pregnancy to meet the iron requirement of both mother and fetus. When detected early in pregnancy, iron deficiency anemia (IDA) is associated with an increase in the risk of preterm delivery. Deficiency of Vitamin C also causes poor iron absorption leading to IDA. This paper deploys non-parametric statistics to analyse and compare the hemoglobin level of pregnant women before and after consumption of Vitamin C and Sulfas Ferroses. The objective of this study is to compare the hemoglobin level of pregnant women before and after the IDA treatment and assess the association between maternal hemoglobin (Hb) level and pregnancy outcome. Johnson transformation is used to estimate the proportions of women with hemoglobin level out side the required specification limits before and after the treatment. Wilcoxon Rank Sum Test is deployed to compare the hemoglobin level before and after the treatment. The result indicates a significant difference in the hemoglobin level before and after the treatment.

Key words: Anemia, Hemoglobin, Johnson transformation, Performance analysis, Proportion of non-conformance, Wilcoxon Rank Sum Test.

1. Introduction

Due to the increased iron requirements during pregnancy, pregnant women are recognized as the group most vulnerable to iron deficiency anemia (IDA). Anemia, as determined by low hemoglobin or hematocrit, is most common among women in their reproductive years. Until recently, it was assumed that anemia during pregnancy had few untoward sequelaes. In recent years it has been suggested that a proportional relationship between anemia and preterm delivery now exists [1]. Infants born to women with a low Hemoglobin **level** have subsequently resulted in a lower birth weight, height and Apgar scores. In Asia, the prevalence of anemia was estimated to be 44% in non-pregnant and 60% in pregnant women [10]. At least 50% of the anemia cases have been attributed to iron deficiency ([2]; [3]; [12]). Maternal iron deficiency anemia increases the risk of premature delivery and subsequent low birth weight, and may contribute to low iron status and poor health of infants ([1]; [11]; [6]).

It was found that pregnant women with anemia are at a greater risk of perinatal mortality and morbidity ([5]; [9]; [8]). The results which have been gathered and analysed are due to a of trial experiment conducted in Indonesia relating to pregnant women. Specifically candidate mothers were observed over a nine month period and asked to consume 90 tablets of Vitamin C 100 mg and Sulfas Ferroses (SF) 350 mg after their third semester of pregnancy. In this paper we assess the performance of non-normal hemoglobin level data collected before and after treatment. The proportion of hemoglobin measurements falling below the WHO recommended level for either case is estimated using Johnson transformation. We will then use Wilcoxon Rank Sum Test to assess the effectiveness of the Vitamin C and Sulfas Ferroses (SF) treatment.

2. Subjects

The study was conducted among 125 women enrolled in a maternity clinic in Banjarmasin Indonesia. We are using 125 pairs of data collected during February 2007-September 2010. The pregnant women who deliver their baby in the clinic had been monitored during their nine-month pregnancy and each consumed 90 tablets of Vitamin C 100 mg and Sulfas Ferroses (SF) 350 mg after their third semester of pregnancy. The hemoglobin level of the individual patient was measured before and after the treatment. One of the objectives in this research is to analyse the hemoglobin level of these women before and after treatment. We also assess the distribution of these data and take necessary steps to transform the data (before and after treatment) to estimate the proportion of women with hemoglobin level outside WHO recommended lower and upper specification limits {LSL , USL] of 11 and 14.

In section 3.2 we explain the reason for transforming data and propose the transformation model to be adopted. At the start of the study the authors have used statistical package Minitab to fit the commonly used normal distribution function to each set of the hemoglobin data.

For both sets the p-value of the fit was less than 0.01 indicating the measurements of the hemoglobin level do not follow normal distribution. The histograms of the data are presented in Figures 1-2. The summary statistics including skewness, Upper and lower quartiles (Q3 and Q1), mean, standard deviation, minimum and maximum are given in table 1 below.

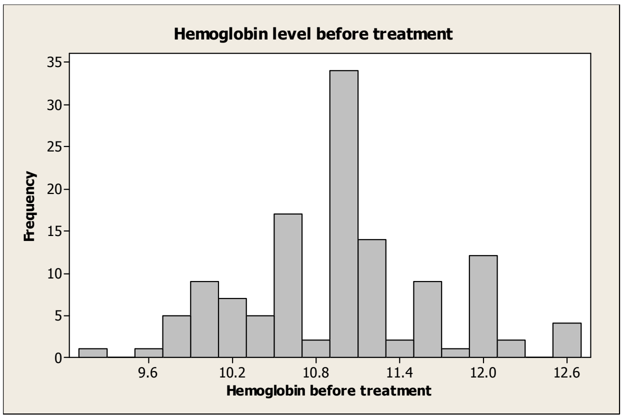


Figure 1: Distribution of the Hemoglobin level measurements before treatment.

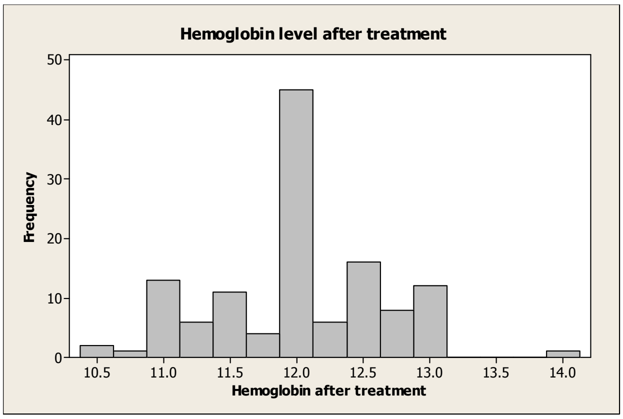


Figure 2: Distribution of the Hemoglobin level measurements after treatment.

Table 1-a: Descriptive Statistics for Hemoglobin level before and after treatment

		N	Mean	St. Dev	Min	Q1
Hb	before	125	10.94	0.7	9.2	10.5
Hb	after	125	12	0.63	10.5	11.5

Table 1-b: Descriptive Statistics for Hemoglobin level before and after treatment

		N	Median	Q3	Max	Skewness
Hb	before	125	11	11.2	12.5	0.13
Hb	after	125	12	12.5	14	-o.o7

3. Method

One of the main objectives of this paper is to estimate the proportion of women with the hemoglobin level outside the WHO lower and upper specification limits (LSL and USL) [13]. This requires the knowledge of the data distribution both prior to the treatment and after the treatment. Figures 1-2 show that none of the data set follows normal distribution. Furthermore, the summary statistics confirm a significant difference between the mean hemoglobin level before and after the treatment with positive skewness coefficient before the treatment and negative after the treatment. It is well documented that the low hemoglobin levels are a growing concern to the medical profession [3], [5]; therefore it is only logical that the aim should be to reduce the proportion of women with Hb level below the LSL. Since the data does not follow normal distribution we have used the Johnson transformation to estimate the proportion of non conforming data. This proportion is estimated using the performance analysis in the statistical package Minitab. We have employed the Wilcoxon Rank sum test which is a nonparametric alternative to paired test to assess the effectiveness of the treatment process undertaken by the clinic.

3.1 Johnson Transformation

Many analyses require an assumption of normality. In cases when data is not normal, one can apply a function to make the data approximately normal and complete the required analysis. In this research we desire to perform a capability analysis on the non-normal hemoglobin data, therefore, we will first transform the data and then apply the performance analysis on the amended data. Depending on the nature of the data, there are many different functions such as square root, logarithm, power, reciprocal or arcsine, that one could apply to transform the data.

The Johnson transformation function is selected from three types of functions in the Johnson system [4]. Because the functions cover a wide variety of distributions by changing the parameters, the statistical package usually finds an acceptable transformation. The Johnson transformation function is complicated, but is well suited for finding an appropriate transformation for our purposes.

3.2 Performance Analysis of Non-normal Data

Process capability analysis is used to ensure that the outcomes of a process are capable to fulfil certain requirements or specifications. Application of process capability analysis is an essential part of the overall quality improvement process. The concept of process capability was introduced by Juran et al.[7]. The two most popular indices used in performance analysis are Cp, Cpk and defined as follows:

$$C_p = \frac{USL - LSL}{6\sigma} \qquad (1)$$

$$C_{pk} = \min\left(\frac{USL-\mu}{3\sigma}, \frac{\mu-LSL}{3\sigma}\right)$$

$$C_{pk} = \min\left(C_{pu}, C_{pl}\right) \qquad (2)$$

μ and σ are the population mean and standard deviation respectively.

In the capability calculations, we are mainly interested in three points within the process distribution: the upper tail, the point of central tendency and the lower tail. In terms of quantiles, these points for the normal distribution correspond to $X_{0.99865}=\mu+3\sigma$, $X_{0.50}=\mu$ $X_{0.00135}=\mu-3\sigma$ In case of normally distributed data, it is easy to estimate quantile points. But for the non-normal data, it is quite cumbersome to estimate them. In fact when the population is not normal, these quantiles do not necessarily corresponds to $\mu+3\sigma$, μ, $\mu-3\sigma$, respectively. To deal with non-normality; one approach is to transform the non-normal output data to approximately normal data using mathematical functions. In this paper we have used the Johnson transformation to transform the data and used the statistical package Minitab to estimate the capability indices and the proportion of data falling out side the specification limits.

3.3 Non-parametric Wilcoxon Rank Sum Test

The Wilcoxon test is a nonparametric analog of the sample t-test because it does not require the data to come from a normally distributed population, as the t-test does. It is a nonparametric hypothesis test for the median of a single population. The procedure uses the null hypothesis that the population median (η) is equal to a hypothesized value ($H0$: $\eta = \eta_0$), and tests it against an alternative hypothesis, which can be either left-tailed $\eta < \eta_0$), right-tailed ($\eta > \eta_0$), or two-tailed ($\eta \neq \eta_0$). In this paper we apply the Wilcoxon test to the population of the hemoglobin difference of the study group. Since the differences would also be non-normal. The differences are defined as:

Difference = Hb level after treatment – Hb level before treatment;

If the treatment is significantly increasing the hemoglobin level, then we expect the median of the population of differences to be greater than zero. However if the treatment had no significant effect then the population of differences would have a median of zero, i.e., we test the hypothesis:

$H0 : \eta = 0.0$ vs $Ha : \eta > 0.0$

4. Case Study and Discussion

As stated earlier this study has used the hemoglobin measurements of 125 women who were monitored in a maternity clinic in Banjarmasin between January 2005 and December 2010. Each patient was given 90 tablets of Vitamin C 100 mg and Sulfas Ferroses (SF) 350 mg after their third semester of pregnancy. The hemoglobin level of individual patient was measured before and after the treatment. The authors have examined the distribution of each data set (before and after treatment) and presented the results in Figures 1-2.

→The results clearly demonstrate the data does not follow normal distribution. The Johnson transformation was employed to convert the skewed data to approximately normal data. The capability analysis was performed on the data to estimate the proportion of patients whose globins level fall out side the WHO recommended specification limits defined by LSL=11 and USL=14. The WHO anemia classifications based on the hemoglobin level are:

- Hb: 11 gr% : normal
- Hb: 9-10 gr% : mild anemia
- Hb: 9-10 gr% : mild anemia
- Hb: 7 – 8 gr%: medium anemia
- Hb: < 7 gr% : severe anemia

The results of the performance analysis are presented in Figures 3-4. The outputs indicate that if patients had not received the treatment then, approximately 416000 per million would have had anemia (PPM < LSL = 416000). Figure 4 shows the significant reduction of this proportion after the treatment, i.e. PPM < LSL = 32000 (under observed performance).

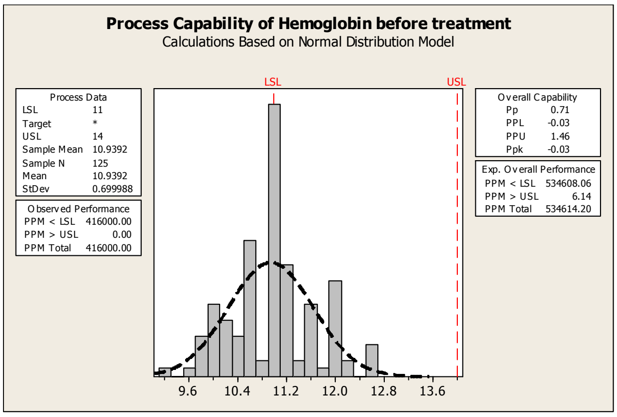


Figure 3: Performance analysis of Hemoglobin measurements before the treatment.

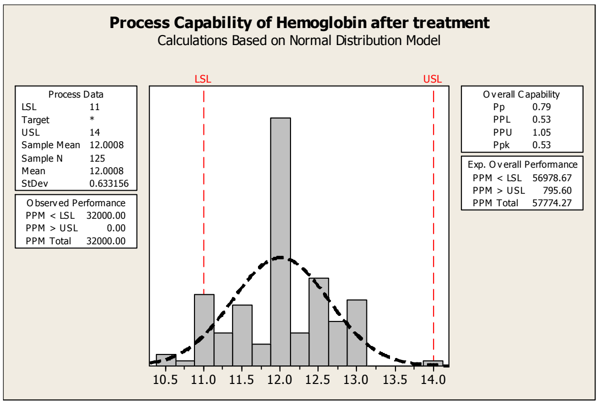


Figure 4: Performance analysis of Hemoglobin measurements after the treatment

We have also employed the non-parametric Wilcoxon Rank Sum test to investigate the effectiveness of the treatment in increasing the hemoglobin level of the women in the study group. Figure 5 presents the box plot comparison of the two data sets. The results of the Wilcoxon Rank Sum test are presented in Table 2 and show that the median of the differences is 1.05 and the 95% confidence interval for the median difference is [1.05 1.10]. The results also indicate that the null hypotheses.

H0: median of differences = 0.0 should significantly be rejected in favour of the alternative;

Ha: median of differences > 0.0 with the p-value of 0.0. Therefore, we can conclude that the treatment has significantly increased the median of the hemoglobin level by 1.05 units. Furthermore, on average we would expect the median to increase between 1.05 to 1.10.

Table 2: Out put for Wilcoxon Signed Rank Test where difference = Hemoglobin after treatment - Hemoglobin before treatment.

```
H0: median of differences = 0.0 versus
Ha: median of differences > 0.0
```

	N	N for Test	Wilcoxon Statistic
Diff after-before	15625	14587	101319492

	N	P	Estimated Median
Diff after-before	15625	0.000	1.05000

```
95% confidence interval for the median of
the differences = [1.05    1.10]
```

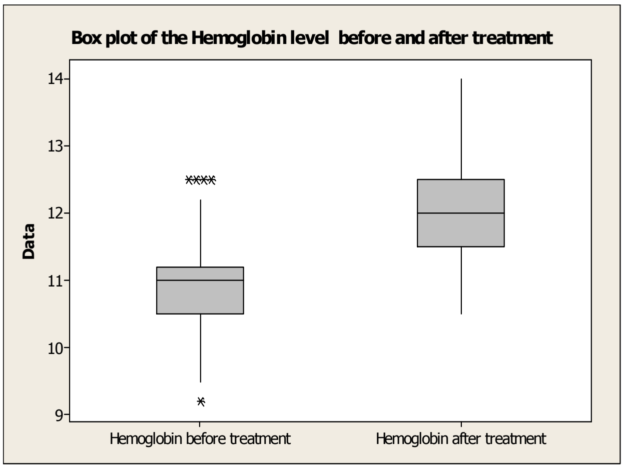


Figure 5: Box plot of Hemoglobin level measurements before and after the treatment.

5. Conclusion

Anemia is an increasingly prevalent factor in the Maternal Mortality Rate (MMR). It has been well established in the research literature that consumption of Vitamin C and Sulfas Ferroses during the pregnancy term reduces the incidence of hemoglobin level <11 which leads to anemia. The paper has investigated the distribution of the hemoglobin measurements before and after the treatment for 125 patients who have been monitored in the clinic. Johnson transformation is used to transfer the non-normal data. Performance analysis is employed to estimate the proportion of anemia patients before and after the treatments. The results show a significant reduction of 416000/1000000 to 32000/1000000 in the proportion after the treatment. The Wilcoxon Signed Rank test is used to assess the effectiveness of the treatment in increasing the hemoglobin level of pregnant women and reducing the maternal mortality rate. The results have showed an increase of 1.05 to 1.10 in the median of the hemoglobin level after the treatment. Therefore, we can conclude that the treatment has been extremely effective in reducing the anemia rate in pregnant women.

Acknowledgement

The authors would like to thank the medical practitioners and staff in the maternity clinic in Banjarmasin Indonesia for their effort in collecting this data and providing them to us for this research.

References

[1] Allen LH (2000). Anemia and iron deficiency: effects on pregnancy outcome. Am J Clin Nutr 71, S1280–S1284.

[2] Allen LH (2001). Biological mechanisms that might underlie iron's effects on fetal growth and preterm birth. J Nutr 131, S581–S589

[3] DeMaeyer E, Adiels-Tegman M (1985). The prevalence of anaemia in the world. World Health Stat Quart 38, 302–316.

[4] Johnson NL (1949) System of frequency curves generated by methods of translation. Biometrika 36:149–176

[5]. Haas JD, Brownlie 4th T (2001). Iron deficiency and reduced work capacity: a critical review of the research to determine a causal relationship. J Nutr 131, S676–S688.

[6] Kaiser LL, Allen LH (2002). Position of the American dietetic association: nutrition and lifestyle for a healthy pregnancy outcome. J Am Diet Assoc 102, 1479–1490.

[7] Juran, J 1974, *Juran's Quality Control Handbook*, 3rd edn, McGraw-Hill, New York.

[8] Ramakrishnan U, Yip R (2002). Experiences and challenges in industrialized countries: control of iron deficiency in industrialized countries. J Nutr 132, S820–S824

[9] Rasmussen KM (2001). Is there a causal relationship between iron deficiency or iron deficiency anemia and weight at birth, length of gestation and perinatal mortality? J Nutr 131, S590–S603.

[10] Rush D (2000). Nutrition and maternal mortality in the developing world. Am J Clin Nutr 72 (Suppl), S212–S240.

[11] Scholl TO, Reilly T (2000). Anemia, iron and pregnancy outcome. J Nutr 130, S443–S447

[12] Singh K, Fong YF, Arulkumaran S (1998). Anaemia in pregnancy – a cross-sectional study in Singapore. Eur J Clin Nutr 52, 65–70.

[13] World Health Organization The world health report 2002: reducing risks, promoting healthy life. Geneva: World Health Organization; 2002

Modelling Non-Normal Neonatal Weight Data to Estimate Mortality Rate of Newborn Babies

S. Ahmad[*], M. Abdollahian[*], S. Nuryani[#], [+]D. Anggraini
[*]*School of Mathematical and Geospatial Sciences, RMIT University, Melbourne, Australia*
[#]*Ulin Hospital (RSUD Ulin) Banjarmasin Indinesia*
[+]*Department of Mathematics Lambung Mangkurat University, Banjarbaru, Indonesia*

Abstract
Neonatal Mortality Rate (NMR) can be defined as the number of newborn deaths' aged 0 to 28 days. The most significant factor influenced NMR is low birth weight of newborn babies approximately 48%, followed by asphyxia 37% and sepsis 32.5%. This paper deploys a Burr XII distribution to model the Non-Normal Neonatal Weight data. The parameters of the fitted Burr XII distribution are estimated using Simulated Annealing (SA) method. The mortality rate defined as the proportion of newborn with weight out side the accepted weight specification limits is then estimated using the fitted Burr XII distribution. The efficacy of the model is assessed by comparing the estimated mortality rate based on the fitted distribution with the actual mortality rate obtained from the data. The results indicate that the Bur XII distribution provides an estimate of Mortality rate which is very close to the actual mortality rate.

.

Key words: *Burr XII distribution, Quantile based capability indices, Simulated Annealing (SA), Process Capability Indices (PCIs), Mortality rate, Proportion of non-conformances.*

1. Introduction

Neonatal Weight (Newborn weight) is an important indicator of infant survival and childhood morbidity [10] and appears to be related to the subsequent risk of Type 2 Diabetes, Hypertension, Cardiovascular Disease, and other disorders [11, 6]. Therefore, many studies have attempted to identify sources of variation in newborn size [6 , 7]. Total maternal weight gain in pregnancy is a well-established, modifiable influence on newborn size [13]. The estimate of the risk of neonatal death can provide important information to paediatricians, especially to neonatal intensive care physicians, with respect to the attention a newborn requires. However, less is known about the true distribution of the Neonatal Weight.

The purpose of the current study is to identify the most appropriate distribution that can describe Neonatal Weight data. The fitted distribution would then be used to estimate the mortality rate defined as the proportion of newborn with weight out side the World Health Organization (WHO) accepted specification limits. The study uses a secondary data of 420 neonatal weights obtained during February 2007- September 2010 in a maternity clinic in Indonesia.

The babies are delivered by mothers who have been monitored in the clinic during their pregnancy. In this study several commonly used statistical distributions have been fitted to the Neonatal Weight data. It has been observed that none of these distributions are capable of describing the true distribution of the Neonatal Weight. Based on the latter observation we have decided to fit Burr XII distribution to the Neonatal Weight data. We will also deploy commonly used Johnson and Box-Cox transformation techniques to transfer the non-normal weight data to normal data.

This paper is organized in the following manner. A capability analysis for the non-normal data and Neonatal Mortality Rate (NMR) estimation are discussed in section 2. A review of the Burr distribution and Johnson and Box & Cox transformation techniques are presented in section 3. The modified quantile based calculations with Burr XII distribution are then used to obtain relevant capability measures in section 4. Application example with real clinical data is presented in section 5.

2. Non-Normal Process Capability Index

The main objective of this capability study is to determine whether the process of monitoring women during the pregnancy period is capable of producing Neonatal Weight within the World Health Organization WHO specifications limits. The most common indices being applied by performance analysts are Cp and Cpk.

Cp is the process capability ratio and is defined as follows:

$$C_p = \frac{Specification\ spread}{process\ spread}$$

which can be also expressed as follows:

$$C_p = \frac{USL - LSL}{6\sigma} \qquad (1)$$

The other important indices Cpk is the process capability ratio for off-centre processes and is defined as the minimum value of one sided upper or lower process capability ratios.

$$C_{pk} = Min\{C_{pu}, C_{pl}\} \qquad (2)$$

These one sided capability measures are defined as follows:

$$C_{pl} = \frac{LSL - \mu}{3\sigma} \quad , \quad C_{pl} = \frac{\mu - LSL}{3\sigma} \qquad (3)$$

where USL and LSL are the upper and lower specification limits, μ and σ are the population mean and standard deviation respectively. Since the population mean μ and process variance σ^2 are unknown, they are often estimated using mean and standard deviation of the collected data.

These capability indices are statistical measures and their calculations heavily depend on the validity of certain assumptions. For instance, the population under examination must be under control and stable. The output data must be independent, identically distributed and follows normal distribution. However, these basic assumptions of traditional process capability indices are not usually fulfilled in practice. Most of the processes that have occurred in the real practical world produce non-normal data. Researchers and practitioners need to examine these basic assumptions before deploying a conventional process capability indices technique. Clements [1] used non-normal percentiles to modify the classical capability indices. He proposed the method of non-normal percentiles to calculate Cp and Cpk indices for a distribution of any shape, using the Pearson family of curves. In this paper we use Burr XII distribution instead of Pearson family of curves in the Clements percentile method to estimate capability indices. The Maximum Likelihood method is used to estimate the parameters of the fitted Burr XII.

The morality rate NRM (proportion of data falling out side the specification limits) is defined [5] by;

$$NMR = 1 - \int_{lsl}^{usl} f(x)dx \qquad (4)$$

Where f(x) is the fitted Burr distribution function. This estimated value of the morality rate based on Burr distribution will be compared with the actual proportion of Neonatal Weights that falls out side the WHO recommended specification limits of [2500 , 4000].

One can also use capability indices to estimate NRM[13] by;

$$\text{NRM} = \Phi(-3\ \text{Cp}) \qquad (5)$$

where $\Phi(\text{y})$ is the cumulative distribution function of standard normal at point y.

Another approach to deal with non-normality is to transform the non-normal output data to approximately normal data using mathematical functions. In this paper we have used Johnson and Box-Cox techniques to transform the data and furthermore used the statistical package Minitab to estimate the capability indices and the proportion of data falling out side the specification limits. A brief discussion of these two techniques is presented in the following section.

3. Review of Burr Distribution

Burr [3] developed a number of useful cumulative frequency functions which can describe various non-normal distributions. One of them is the Burr XII distribution. This distribution is widely used in reliability and quality literature. The mathematical expression of the probability density function of the Burr XII distribution is defined as follows:

$$f(y) = \begin{cases} \dfrac{cky^{c-1}}{(1+y^c)^{k+1}} & if \;\; y \geq 0; c \geq 1; k \geq 1 \\ 0 & if \;\; y < 0 \end{cases} \qquad (6)$$

The mathematical expression of the cumulative distribution function of the Burr XII distribution is therefore given by

$$F(y) = \begin{cases} 1 - \dfrac{1}{(1+y^c)^k} & if \;\; y \geq 0 \\ 0 & if \;\; y < 0 \end{cases}$$

(7)

In the above equations, c and k represent the skewness and kurtosis coefficients of the Burr distribution. Burr [2] showed that a wide range of probability density functions can be fitted by an appropriate Burr XII distribution. For example, the normal density function can be estimated by a Burr distribution with c =4.85437 and k =6.22665 and a Gamma distribution with shape parameter 16 can be approximated by a Burr XII distribution with c = 3 and k = 6, and log-logistic distribution is a special case of Burr XII distribution. Rodriguez [15] demonstrated that the Weibull distribution is a limiting distribution of the Burr XII distribution. In practice, it has been observed that the majority of the quality characteristics follow Weibull distribution. Hence, the two-parameter Burr XII distribution can be used to describe real data.

Burr [2] has tabulated the mean and standard deviation as well as skewness and kurtosis coefficients for the family of Burr distribution. These tables enable users to make a standardized transformation between a Burr variate (say *Q*) and another random variate (say *X*). The mathematical expression of the transformation is defined as:

$$\frac{X - \overline{X}}{S_x} = \frac{Q - \mu}{\sigma} \tag{8}$$

where $\overline{X}$ and S_x are the sample average and standard deviation of the original sample data, and. μ and σ are the mean and standard deviation for the fitted Burr distribution. Zimmer and Burr [16] developed a method for sampling variables from non-normal populations using the Burr XII distribution. Castagliola [5] used Burr's approach to compute the proportion of nonconforming items.

3.1 Non- Normal Transformation Techniques

Data transformation refers to the application of a known deterministic mathematical function to each point in the data set, i.e. each data point X_i is replaced with the transformed value $Y_i = f(X_i)$, where the function $f(.)$ is an appropriate mathematical function. The main objective of the data transform technique is to transform the non-normal data to normally distributed data so that it can closely meet the assumptions of a statistical inference procedure that need to be applied to improve the interpretability of the data set. In this paper we deploy Johnson and Box-Cox transformation techniques (two of the most commonly used transformation techniques) that are available in most statistical packages to estimate Cp and NRM for the non-normal Neonatal Weight Data.

3.1.1 Johnson Transformation technique

Johnson [8] proposed a system of distributions based on the moment method to transform the non-normal data to normally distributed data. The Johnson transformation function is selected from three types of functions in the Johnson system. Because the functions cover a wide variety of distributions by changing the parameters, the statistical package usually finds an acceptable transformation. The Johnson transformation function is complicated, but is very suitable for finding an appropriate transformation. This transformation method is available in most statistical software packages as a standard feature.

3.1.2 Box-Cox power Transformation technique

Power transformation is a variant of transformations that map non-normal data from one space to another using power functions. This is a useful data transformation technique employed to reduce data variation and to ensure that the data is normally distributed. The Box-Cox power transformation is the most commonly used technique in the medical industry. This technique was proposed by Box and Cox [4]. The Box-Cox power transformation on necessarily positive response variable *X* is expressed by

$$X(\lambda) = \begin{cases} \dfrac{(X^{\lambda} - 1)}{\lambda} & for \quad \lambda \neq 0 \\ Ln\,(X) & for \quad \lambda = 0 \end{cases}$$

where $-5 \le \lambda \le +5$ (9)

This transformation depends upon a single parameter λ that can be estimated by Maximum Likelihood Estimation (MLE) method [4]. Using the optimal $\overset{*}{\lambda}$ value, data values for each individual X are transformed to a normal variate using equation (9). Box-Cox transformation can be applied to non-zero, positively skewed data. The transformation method is available in most statistical software packages as a standard feature. Consequently, the user can deploy this technique directly and with ease to evaluate process capability indices for non-normal data first transforming.

4. Fitting Burr Distribution to Data

In this paper we fit Burr distribution function $f(x)$ to neonatal data. To fit the appropriate Burr distribution, we need to estimate the parameters c and k. We will use the method of Maximum Likelihood Estimation (MLE) to estimate these parameters. For a random sample of size n, $x_1, x_2, \ldots x_n$ from the population , the likelihood function of Burr XII distribution is defined by

$$L(c, k; x_1, \ldots., x_n) = \frac{c^n k^n \prod_{i=1}^{n} (x_i)^{c-1}}{\prod_{i=1}^{n} (1 + x_i{}^c)^{k+1}} \tag{10}$$

The corresponding log-likelihood function is defined by

$$\log\ L = n\log(\ c\ ) + \log(\ k\ ) - (1 + k\ )$$
$$* \sum_{i=1}^{n} \log(1 + x_i^{\ c}) + (c-1)\sum_{i=1}^{n} \log\ x_i \qquad (11)$$

The log-likelihood function (equation 11) is used as an objective function to find the Burr fitted MLE estimators of c and k. To determine the MLE estimators of c and k the Simulated Annealing (SA) method is adopted in this paper.

4.1 Simulated Annealing Search Method

The Simulated Annealing (SA) method is based upon that of Metropolis et al. [8], which was originally proposed as a mean of finding the equilibrium configuration of a collection of atoms at a given temperature. The connection between this method and mathematical minimization was first observed by Pincus [14], but it was Kirkpatrick et al. [9] who subsequently proposed it as an optimization technique for combinatorial and other optimization problems. Ease of use and the provision of good solutions to real-world problems makes this method one of the most powerful and popular meta-heuristics method to solve many optimization problems.

5. Case Study

The study was conducted among women enrolled in a maternity clinic in Banjarmasin Indonesia. It is assumed that if pregnant women have been monitored regularly in the clinic during their nine-month of pregnancy they will deliver babies with normal weight. The study is using 420 data sets collected during February 2007- September 2010. One of the objectives in this research experiment is to investigate the performance of the clinical monitoring program and assess its capability in reducing the proportion of Neonatal Weights that fall outside the WHO specification limits (mortality rate). The numerical values that measure the process capability are capability indices. The definition and properties of these indices are explained in Section 2. To estimate theses indices one needs to know the distribution of the data.

In the initial step of the study the authors have used a statistical package to fit commonly used continuous probability distribution functions such as, Normal, lognormal, 3-parameter Lognormal, Exponential, 2-parameter Exponential, Weibull, 3-parametrs Weibull, Logistic, Log logistic, 3-parameter Log logistic, Gamma and 3- parameter Gamma distributions to the Neonatal Weight data. In all cases the p-value for the fit was less than 0.01 indicating that none of the fits is appropriate. The histogram of the data is presented in Figure1.

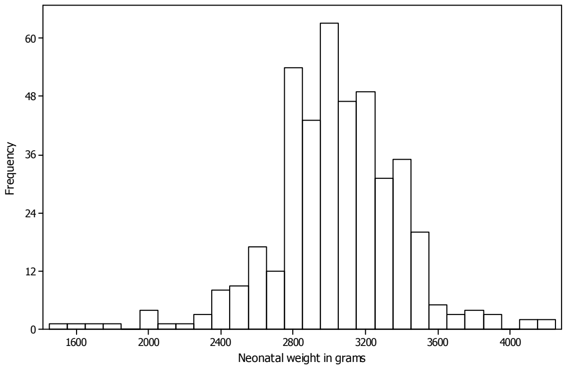


Figure 1: distribution of the Neonatal Weight data

One of the main objectives of this paper is to examine the possibility of fitting the Burr XII distribution to this non-normal data set. This will then be followed by estimating the proportion of non-conforming Neonatal Weight using non-normal capability index (PCR).

One criterion, proposed by many researchers [1] for assessing the efficacy of their proposed method is to estimate the proportion of non-conformance (Neonatal Mortality Ratio (NMR)) based on their proposed method and compare it with the actual proportion of non-conformance obtained from the actual data.

In this paper the estimated parameters of the fitted Burr distribution are obtained using Simulation Annealing method and are displayed in Table 1. We have also included the estimates of NRM and cp values based on Johnson and Box-Cox transformations in table 2.

The proportions of nonconforming data NRMs and their corresponding capability index are obtained using equations (4), (5) and (6).

Table 1: Burr distribution parameters (c, k) estimated using Simulated Annealing method.

Burr parameters estimation methods

Neonatal weight specification limits	Simulated Annealing (SA)	
	c	k
4000	12.0462	0.0508
2500	0.0985	0.0504

The proportion of nonconforming data NMR for the fitted Burr distribution is obtained using equation (4) and replacing $f\ (\ x\ )$ by the fitted Burr distribution based on SA method. It is worth mentioning that

$$\int_{lsl}^{usl} f(x)dx = F(usl) - F(lsl)$$

where usl = 4000, lsl = 2500 and $F(x)$ is the cumulative distribution function for x and is defined by equation (7).

Table2: Proportion of non-conforming NRMs and process capability indexes using fitted Burr distributions, Johnson and Box-Cox transformations.

	Burr based on (SA)	Box Cox	Johnson
Cp	0.5114	0.56	0.49
NRM	0.0625	0.0465	0.0708
TRUE NRM		0.0595	

Cp: capability index

NMR: Neonatal Mortality Ratio

Using the NMR criterion, table 2 shows that the NMR obtained using the fitted Burr distribution is closer to the actual NMR (0.0595) in comparison with NMR obtained based on Box-Cox and Johnson transformations. The actual NMR represents the actual proportion of data that fall outside their respective specification limits given by the WHO.

6. Conclusion

In this paper we have proposed to model the non-normal Neonatal Weight data by Burr XII probability distribution. The parameters of the fitted Burr distribution are estimated using the Simulation Annealing method. The mortality rate defined by the proportion of newborns with Neonatal Weights out side the WHO specification limits is then estimated using the proposed fitted Burr distributions. We have also deployed the commonly used Johnson and Box-Cox transformation techniques to estimate the mortality rate. The results show that the estimated mortality rate based on the proposed Burr model is very close to the actual mortality rate.

The paper also estimated the capability index of the process of monitoring mothers during the nine month of their pregnancy. The estimated capability index indicates that the monitoring process should be improved to increase the proportion of babies within the WHO specified weight limits (reducing the mortality rate caused by the low birth weight). The process of monitoring pregnant women during their pregnancy can help to improve the quality of newborn care, especially in the case of premature babies. Clinicians who correctly estimate neonatal survival probabilities of premature newborns tend to provide more appropriate care than those who underestimate survival probability. It has also been observed that neonatologists with the correct estimation of neonatal survival intervene more often with appropriate invasive therapies, such as mechanical ventilation, cardiopulmonary resuscitation, inotropes, and intravenous fluids.[7,12].

Acknowledgement

The authors would like to thank the medical practitioners and staff in the maternity clinic in Banjarmasin Indonesia for their effort in collecting these data sets and provide them to us for this research paper.

References

[1] Ahmad, S, Abdollahian, M, Zeephongsekul, P & Abbasi, B 2008, 'Measuring Process Performance for Non-Normal Quality Characteristics Data', Ubiquitous Computing and Communication Journal, vol. 3, pp. 8-12.

[2] Burr IW (1973) Parameters for a general system of distributions to match a grid of á3 and á4. Commun Stat 2:1–21

[3] Burr IW (1942) Cumulative frequency distribution. Ann Math Stat 13:215–232.

[4] Box, G & Cox, D 1964, 'An analysis of transformation', Journal of royal statistics Society, vol. 26, pp. 211-243

[5] Castagliola P (1996) Evaluation of non-normal process capability indices using Burr's distributions. Qual Eng 8(4):587–593

[6] Godfrey KM. Maternal regulation of fetal development and health in adult life. Eur J Obstet Gynecol Reprod Biol 1998;78:141–50.

[7] Haywood JL, Morse SB, Goldenberg RL, Bronstein J, Nelson KG, Carlo WA. Estimation of outcome and restriction of interventions in neonates. Pediatrics 1998;102:e20

[8] Johnson NL (1949) System of frequency curves generated by methods of translation. Biometrika 36:149–176

[9] Kirkpatrick, S., Gerlatt, C. D. Jr., and Vecchi, M.P. (1983), Optimization by Simulated Annealing, Science 220, 671-680

[10] McCormick MC. The contribution of low birth weight to infant mortality and childhood morbidity. N Engl J Med 1985;312:82–90

[11] Mi J, Law C, Zhang KL, Osmond C, Stein C, Barker D. Effects of infant birth weight and maternal body mass index in pregnancy on components of the insulin resistance syndrome in China. Ann Intern Med 2000;132:253–60

[12] Morse SB, Haywood JL, Goldenberg RL, Bronstein J, Nelson KG, Carlo WA. Estimation of neonatal outcome and perinatal therapy use. Pediatrics 2000;105:1046-50

[13] National Academy of Sciences. Nutrition during pregnancy. I. Weight gain. II. Nutrient supplements. Washington, DC: National Academy Press, 1990:176–221.

[14] Pincus, M. (1970), A Monte Carlo Method for the Approximate Solution of Certain Types of Constrained Optimization Problems, Oper. Res. 18, 1225-1228.

[15] Rodriguez RN (1977) A guide to the Burr type XII distributions. Biometricka, 64:129–134

[16] Zimmer WJ, Burr IW (1963) Variables sampling plans based on non normal populations. Ind Qual. Control July:18–36

Identification of Human Skin Regions Using Color and Texture

Mark Smith
University of Central Arkansas
Conway, Arkansas 72035

Ray Hashemi
Armstrong Atlantic University
Savannah, Georgia 31419

Leslie Sears
Armstrong Atlantic University
Savannah, Georgia 31419

Abstract

A novel approach identifying and segmenting skin regions within images is presented in this paper. The identification and recognition of facial regions are a central focus of this work. A set of standard images containing facial/skin objects is first manually segmented into the interested regions. These regions are utilized in the training the system. Dominant color features (i.e., the most frequently occurring quantized colors) along with texture features generated from the co-occurrence matrix are extracted from the training regions. An example image is then presented to the system. The image undergoes a standard image segmentation algorithm that splits the image into consistent objects. The same color/texture features are extracted from the example regions. A similarity measurement is computed and the regions of the example image are subsequently classified as skin/non skin regions. Results are shown for several standard mpeg such as Foreman, Salesman, Miss America, and others.

1. Introduction

The convenience in capturing and encoding of digital images has caused a massive amount of visual information to be produced and processed rapidly. Hence, efficient tools and systems for searching and retrieving visual information are needed. There is especially a need to identify people and human features as they appear in images and video sequences. The identification of humans within images allows for a more intelligent and efficient retrieval system. Other applications consist of classifying video scenes as active or background and allows for filtering large regions of the video. This system focuses on visual information pertaining to skin, especially facial regions, which even increases the complexity of the matching and retrieval algorithm significantly.

Skin identification and segmentaion has received much interest in recent years as a research topic[1], [3]. There are currently several general-purpose systems available for skin and facial identification: QBIC [2], PhotoBook [5], Virage [4], and VisualSEEk [1]. Our system focuses only on the analysis and identification of human skin objects [6-7]. The following steps of our system are summarized below:

- A set of training images undergo manual image segmentation. All skin regions are identified.
- The Dominant Color feature is extracted from each skin region.
- Texture features consisting of the co-occurrence matrix are extracted from each skin region.
- Example images are automatically segmented using a standard region segmentation algorithm.
- Dominant color and co-occurrence texture features are extracted from each region.
- A similarity measurement is computed between the training objects and each example region.
- A sufficiently small similarity measurement results in the example region classified as a skin region.

2. Training Set

The primary objective of this step is to provide a representative set of color and texture features corresponding to true skin data. The steps involved with this portion consist of the following:

1. Manually segment skin textures within a set of images
2. Extract Dominant Color Features
3. Extract texture features from Co-occurrence Matrix

The training set utilized by this system was extracted from the *Foreman* video. Examples of the manual segmentation of this video sequence is shown below:

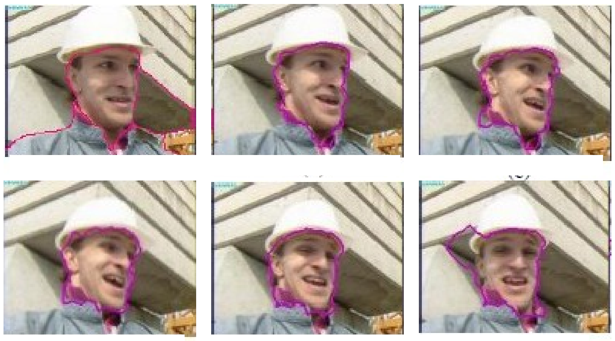


Fig. 1. Foreman test sequence between 126-146.

The next step extracts the Dominant Colors/Texture features from these manually segmented images.

3. Dominant Colors

The color feature utilized in this measurement consists of all quantized RGB colorshaving a concentration greater than 5% extracted from a given object. These colors arealso referred to as the dominant colors of the object. The Dominant Colors are one of the MPEG-7 features. A color quantization step is performed that quantizes the total number of colors to only 16. The color processing to detect those 16 with 5% or more concentration is illustrated below in Fig. 2:

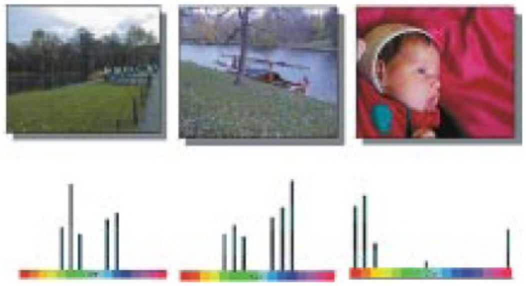


Fig. 2. Detection of Dominant Colors in Images.

4. Texture Features

The Gray-Level Co-occurrence Matrix (GLCM) is one of the most popular statistical texture measurements [2] and has been used as the primary component in a wide range of image segmentation applications [4]. The GLCM is a second-order statistical measurement; second-order statistics take into account the relationship between groups of two (usually neighboring) pixels in the original image. In contrast, first-order stats, (e.g., mean and variance), do not consider any neighborhood associations.

4.1 Computing Co-Occurrence Matrix

The process by which the GLCM is computed is outlined as follows:

1. The GLCM computation utilizes the relation between two pixels at a time; one is called the reference and the other the neighbor pixel.
2. A displacement vector d, distance in horizontal direction or distance in vertical direction.
3. A displacement vector is selected and determines the relationship between the pixels in the image. Utilizing only neighboring pixels (d = 1) is the most commonly used distance
4. There are 8 possible relationships (i.e., displacement vectors) that can be formed between neighboring pixels (directions between neighboring pixels are shown in parenthesis –the first component refers to the horizontal displacement, whereas the second parameter refers to the vertical displacement.

Positive horizontal values represent right neighboring pixels while negative values correspond to left neighbors. Positive vertical values represent a neighboring pixel above the reference pixel, while negative values correspond to a neighboring pixel below the reference pixel.

The values extracted from the co-occurrence matrix represent an excellent texture measurement and is stored in a vector as defined as follows. The co-occurrence matrix is best represented as a probability density as shown in Equation (1) below:

$$P_i = \frac{V_{ij}}{\sum_{i=0}^{N} \sum_{j}^{M} C_{ij}} \qquad (1)$$

The co-occurrence matrix is computed over each facial object and maintained for matching with example objects. A seven dimensional vector, consisting of:

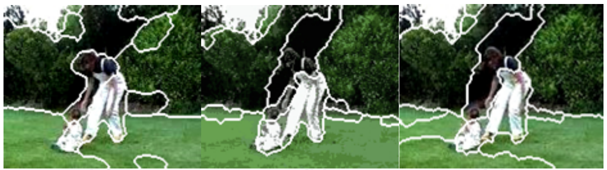
- entropy
- homogeneity
- energy
- mean
- standard deviation
- coefficient of correlation
- contrast

provides the texture features utilized by this system.

5. Segmenting Example Images

There are several image segmentation algorithms available in the literature, thus providing many different choices for this step in our system. The image segmentation process used in this system is fully described in an earlier work of the authors [4] and is applied to each frame of the video sequence.

The results of our image segmentation algorithm applied to 3 frames of the *HappyGranny* sequence are shown in Fig. 3. Note the inconsistencies between segmented regions.



(a) (b) (c)

Fig. 3. Examples of initial image segmentation for *HappyGranny* sequences for (a) Frame 90 (b) Frame 91 (c) Frame 92.

6. Similarity Measurement

After the example image has been segmented into meaning objects, the dominant color and co-occurrence texture features are extracted from each object. Each object is then compared with the features extracted from the training objects. The similarity function is defined as the norm difference between the vector features as shown below:

$$Diff = \frac{|T_t - T_{obj}|}{|T_t|} + \frac{|D_{ct} - D_{cobj}|}{|D_{ct}|} \quad (2)$$

The resulting difference is then compared with an empirically computed threshold *T* as

$$T > Diff \quad (4)$$

If Diff is less than T as shown in equation (4), the object is classified as skin region.

7. Testing and Results

The system is tested on the following images/videos - *Car, Tennis, Claire, and Susie.* All regions identified as skin objects are outlined in purple. The facial objects are automatically segmented using the algorithm described in section 5 and the color/texture features are extracted as described in sections 3 and 4. The training objects/features have already been processed and the similarity measurement is applied as in section 6.

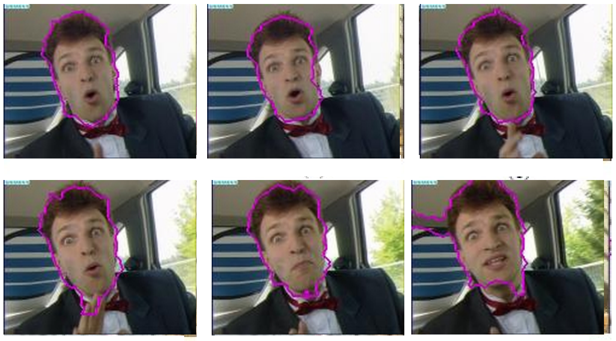


Fig. 4. Carphone test sequence between 168-182.

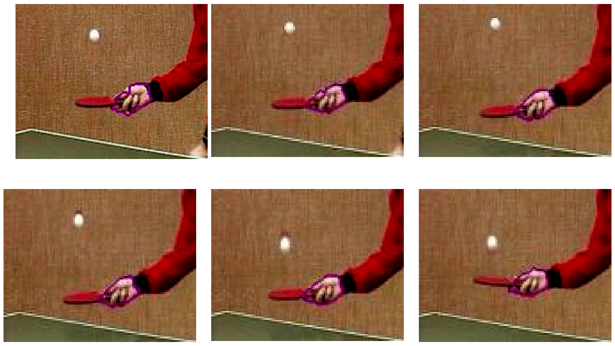


Fig. 5. Tennis test sequence between 0-20.

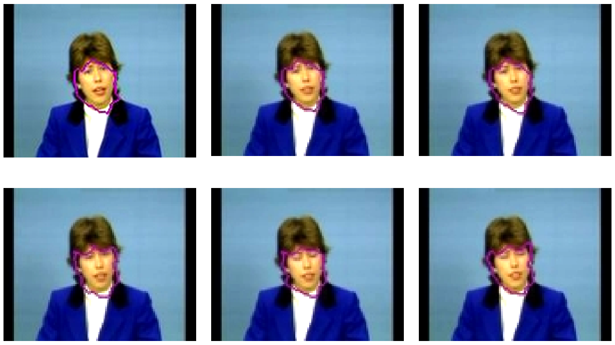


Fig. 6. Claire test sequence between 71-88.

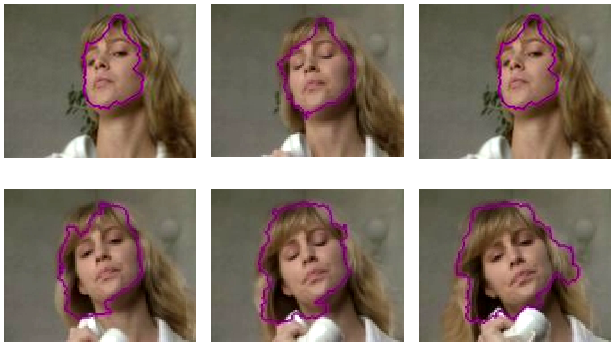


Fig. 7. Susie test sequence between 53-73.

The following video clips where segmented and then compared using the proposed system. The following results pertaining to video clip comparison are shown below:

Table 1 Results

Video	#Objects	#Skin	Correct	Percent Correct
Let It Be	35	9	8	97.1
Say the Word	23	12	10	91.3
Taxman	9	6	5	88.9
Love me do	29	11	10	96.5

Overall, the results are very positive and the system does a good job classifying the skin regions based on the specified criteria.

.

8. Conclusion and Future Work

The algorithm described in this work accurately classifies human skin regions based on test set criteria provided by the user The system is successfully tested on a variety of standard mpeg-4 videos, consisting of many human skin regions - totaling nearly 100 skin objects in all. The algorithm performs well under a variety of different conditions and circumstances. The results and the technology derived from this work has proven to be very exciting and we look forward to developing a series of additional applications from this work.

9. References

[1] Y. Deng and B. S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 6, pp. 939-954, 2001.

[2] T. Aach, A Kaup, and R. Mester, "Statistical model-based change detection in moving video," *IEEE Trans. on Signal Processing*, vol. 31, no 2, pp. 165-180, March 1993.

[3] A. Nagasak and Y. Tanka, "Automatic video indexing and full video search for object appearances," in *Visual Database System II*, Elsevier, 1992, pp. 113-127.

[4] M. Smith and A. Khotanzad, "Unsupervised object-based video segmentation using color and texture features," *IEEE Southwest Symposium on Image Analysis*, March, 2006.

[5] J. Goldberger and H. Greenspan," Context-based segmentation of image sequences," *IEEE Transactions On Pattern Analysis And Machine Intelligence*, vol. 28, no. 3, pp. 463-468, March 2006.

[6] H. Tao, "Object tracking with Bayesian estimation of dynamic layer representations," *IEEE Trans. On Pattern Anal. And Mach Intelli.*, vol. 24, no. 1, pp. 75- 89, January 2002.

[7] F. Porikli, "Real-time video object segmentation for MPEG encoded video sequences,"*TR-2004-011*, pp. 178-189, March 2004.

Wireless Power Transmission

Stephanie Shreck, Shahram Latifi
University of Nevada, Las Vegas
Department of Electrical and Computer Engineering
University of Nevada, Las Vegas
4505 Maryland Parkway
Las Vegas, NV 89154-4026

Abstract

Wireless Power Transmission (WPT) has been an ongoing research area since 1873 when James Maxwell first theorized transferring power could be achieved through electromagnetic radiation. While microwave systems were first to be developed, in recent years the viability and demand for optical systems has grown significantly. In the space elevator contest sponsored by NASA in 2009, LaserMotive successfully demonstrated the potential of a laser based WPT system. The defense and space communities have started to recognize the impact that this technology could have given the multitude of battery-reliant unmanned systems that exist today. The current research is a study analyzing the background and benefits of laser based WPT systems, the system components, demonstrated/proposed WPT systems, and proposed research area needs within laser based WPT systems.

Keywords: Laser, Microwave, UAV, Photovoltaic, Concentrator lens.

1. Introduction

Military and Homeland Security personnel are becoming more reliant on the advancing technology in autonomous vehicles. Unmanned Aerial Vehicles (UAVs), Unmanned Combat Air Vehicles (UCAV), Unmanned Surface Vehicles (USV), Unmanned Underwater Vehicles (UUV), and Unmanned Ground Vehicles (UGV) are all being developed. Whether the need is to locate people after a natural disaster or to keep troops out of harm's way, the job of these systems becomes imperative.

UAVs are probably the most well known of the unmanned vehicles and commonplace in places like Afghanistan and Iraq. The troops feel a sense of security knowing that a UAV has already sent pictures and data about what type of obstacles lie ahead. Unfortunately, as more functionality is added to these surveillance systems, the endurance of the vehicles is greatly reduced. Many of the small battery powered vehicles have the ability to last an hour or two, depending on how much data they are collecting and transmitting. Larger gas powered vehicles are able to increase endurance to a few days.

Endurance is not limited to UAV's though; any of the unmanned vehicles run into the same problem. To get the most information possible, functionality is added to these systems, increasing weight and power demand. This results in a need to refuel periodically during a mission. In a natural disaster or war, time wasted to refuel can result in lives lost. The need to keep systems running 24/7 is a crucial one.

Efforts have been made to increase the endurance of these vehicles, by combining solar powered vehicles with small battery capacity. For QinetiQ's high altitude UAV, Zephyr, this has significantly increased the availability of the vehicle to two weeks [8]. While this solar powered system moves towards the ultimate goal, not all of the unmanned systems have access to solar power regularly, or the surface area to provide the needed power per cell. To accommodate these cases, another method that provides power transfer on demand is required. Wireless Power Transmission provides a viable solution for these needs.

2. Wireless Power Transmission

While Wireless Power Transmission (WPT) is not a new concept, it is one that is underdeveloped for its age. Transferring power through electromagnetic radiation was first theorized by James Maxwell in 1873. Thirty years later Nikola Tesla proved Maxwell's theory by transporting energy using electromagnetic waves through vacuum. Soon after, in 1918, Heinrich Hertz also validated the findings in principle. [6]

It was not until the Klystrom and Magnetron were developed that the next big step in WPT took place. William Brown used microwave energy to power a small tethered helicopter. In 1968, Peter Glaser proposed the first solar powered satellite systems based on work done by Tsickovski, Oberth, and Brown. Following in 1979, a study was done by the Department of Energy (DoE) and NASA about the potential of creating a solar power plant in space and beaming the energy down to earth. The "Solar Power System" report created by NASA and the DoE concluded that while the technology was feasible, the size and cost of such a system was too high to pursue [6]. Since the 1980's, several demonstrations and concepts have been proposed.

Wireless Power Transmission is typically considered using one of several methods: inductive coupling, microwaves, and lasers. Inductive coupling works well for two objects that can come into contact (or be relatively close) to each other, such as charging toothbrushes or medical equipment. For power systems over longer distances, microwave or lasers are the two options.

Since microwave systems have advanced more over the years than lasers, many more demonstration systems and designs have been created using microwaves. The first experiment was William Brown's helicopter following his invention of the rectenna in 1964. The rectenna, which is short for rectifying antenna, is the receiver that is needed to capture the microwaves and convert them into electrical energy. In 1975 the Jet Propulsion Lab (JPL) developed a system that transmitted 30 kW to a 26-meter diameter rectenna over a distance of 1.54 km. 85% efficiency was achieved. In 1985, N. Kaya was able to use microwaves to transmit power in space. A ground to plane transfer of power via microwaves was completed by the Canadians in 1987. In the early 90's Japan joined the WPT work by transmitting microwave power to a small airplane in 1993 and a balloon system in 1995. Radio Frequency Identification (RFID) systems would also be a smaller scale of proven wireless power technology using microwaves. [6]

Laser demonstrations, while not as prolific, are still another method that has been investigated over the years. In the 1980's there was some presumably classified work done in this area. Between 2002 and 2003, Steinsiek and Schäfer demonstrated a ground to ground transmission of laser power to a rover. This system used a green, frequency doubled Nd:YAG laser sending a couple of Watts to a rover 280-meters away.

The space elevator contest sponsored by NASA was a recent driver to get laser WPT developed. Contestants were provided a vertically suspended ribbon that teams had to create a laser powered mover to climb 1 km. This competition focused mostly on power level optimization and was not concerned with beam control or steering aspects of a WPT system. The competition ended in 2009 with LaserMotive winning. Since then LaserMotive has published a white paper specifically looking at the use of laser WPT to power UAV's [5].

3. Laser versus Microwave WPT

Both laser and microwave power transmission has been demonstrated and studied over the last 30 years offering different advantages for different applications. Table 1 provides a quick look at some of the positive and negative aspects of each approach. For applications where size and weight are limiting factors, such as small unmanned systems or space applications, laser systems provide a big advantage compared to microwave.

The systems differ in wavelength for operation to minimize atmosphere attenuation which drives the properties of the overall system. "Microwave frequencies of either 2.45 or 5.8 GHz (0.12-0.05 m; both in the industrial, scientific and medical (ISM) frequency band), laser energy transmission takes advantage of the atmospheric transparency window in the visible or near infrared frequency spectrum" [6]. Figure 1 highlights that atmosphere opacity by wavelength. As can be seen, laser systems in the visible range are attenuated slightly, while microwave systems are transparent to the atmosphere. These wavelength differences determine the size of the receiver and transmitter needed to operate. The laser system at 1μm versus the 0.12-0.05 m microwave system can be achieved using significantly smaller transmit and receive components.

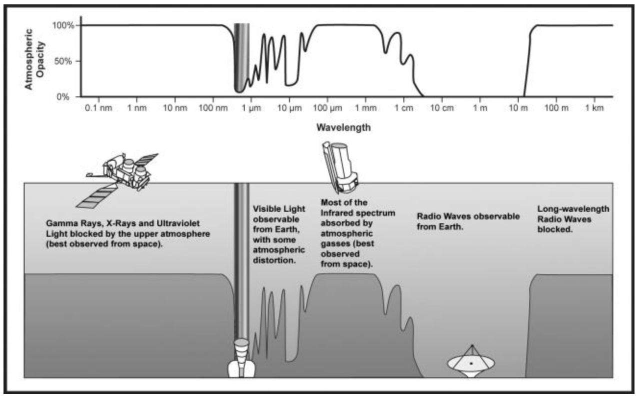


Figure 1: Transmission and absorption in Earth atmosphere (source NASA) [6]

Table 1: Laser versus Microwave Wireless Power System Advantages and Disadvantages (compilation of findings)

Type:	Microwave	Laser
Positives	- Little attenuation due to atmosphere - High efficiency rates - Electronic Beam Steering (well developed and implemented) - Ideal for terrestrial applications	- High Energy Density - Narrow, Focused Beam - Small receiver and transmitter system - Electronic Beam Steering (not very developed) - Ideal for space applications
Negatives	- Filtering needed to deal with side and grating lobes - Large transmitter and receiver (size and weight) - Low energy density - Safety systems necessary	- Atmosphere Attenuation - Low efficiency rates - High power systems require large cooling systems - Safety systems necessary

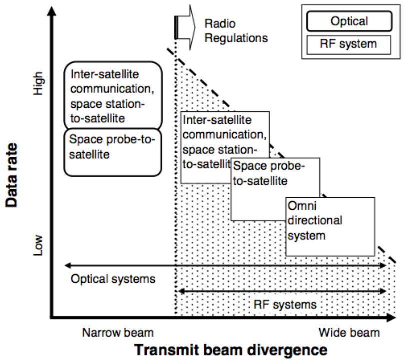


Figure 2: Classification of Satellite Communication Systems by Beam Divergence and Data Rate [6]

The energy density of the system also creates an advantage for laser systems. "Similar to the higher data rate achievable with optical data links (Figure 2), laser energy transmission allows much higher energy densities, a narrower focus of the beam" [6].

Electronic beam steering allows for minimal moving parts to be used in the system which increases overall system reliability. This was first developed for microwave antenna systems which was a clear advantage over laser based systems. A study done by "Schafer and Kaya demonstrated that a similar system is, in principle, also possible for laser based systems, by presenting a new concept for a retrodirective tracking system" [6]. This work, done in 2007, allows laser systems to have the same electronic steering capability as microwave systems.

The big drawback with laser systems is the reduced efficiency that comes from power conversions within the system. Where microwave systems can be created with an 80%-90% efficiency, laser systems are in the 10%-20% range. This low efficiency is attributed to many factors. Some factors can be improved with system design; others such as the atmosphere opacity of laser wavelengths cannot.

4. WPT System Components

A typical wireless power transmission system consists of several key components: transmitter, receiver, safety system, cooling, conversion electronics, power source, and pointing control system. Figure 3 shows these parts within a simple system. The two most important components being the transmitter and receiver that will be discussed more in this section.

The transmitter can be microwave or laser based. From the earlier discussion in Section 3, going forward it will be assumed that we are working with a laser system. A receiver in this case could either be, the traditional photovoltaic (PV) cells (convert the light into electric energy), or thermovoltaic cells (produce energy based on thermal gradient that is produced by the transmitter). Both receivers come with their own drawbacks when designing a system. Neither PV cells, nor thermovoltaics provide the highest efficiencies. To get better efficiencies, PV cells need to be optimized for the laser that is being used. The angle of incidence also needs to be taken into account in order to provide the maximum efficiency. Since thermovoltaic cells depend on the thermal gradient, its efficiency will be a factor of the environment. For lab settings this can be controlled, however applications like those discussed in Section 1 adds much uncertainty to the system if thermovoltaics are used.

As with any power system, safety needs to be integrated into the system. Some proposed designs desire to create a beam that is safe enough to walk through [4], while other designers prefer to have the beam turn off if an obstruction moves in the way of the beam [3]. Depending on the use of the system the power transmitted could range from a couple of watts to power a small vehicle to Megawatts to transmit power from a solar power plant in space to Earth. These safety concerns will introduce constraints on the overall system that need to be accounted for at all steps in the design process.

The power supply for the transmitter is an important design parameter within the system design. This power could come straight from the grid or from another power source (solar, battery, etc). An example of this would be a direct or indirect solar pumped laser. A laser system for a satellite would most likely be an indirect solar pumped laser. This indicates that the initial energy for the system is received via solar arrays, converted into electricity, and then converted into a laser beam. For many military applications or natural disaster scenarios, where electricity is not readily available, this might be the best choice for system design. Direct solar pumped lasers, on the other hand, is where the conversion to electricity is removed, and the collected solar energy converts directly to a laser beam. [6]

Cooling, pointing control, and miscellaneous electronics (data transmission, handshaking, etc) will be dependent on the specific design systems. The complexity of the cooling system will depend on the amount of power that is being transmitted through the system. A system of less than 100W can use something like Peltier cooling units [7]. As the output exceeds a couple of hundred Watts, a more complex and larger cooling system is necessary.

Pointing control is an important part of the system to increase the overall efficiency and safety of the system. Whether mechanically or electronically performed, the pointing accuracy will allow the laser to hit the receiver at the correct angle and avoid interference with other items in the environment around the receiver. As mentioned in Section 2 this is an area that has been neglected in recent work associated with the Space elevator contests.

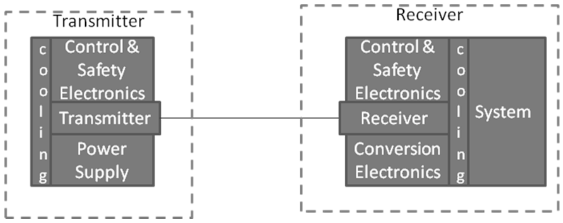


Figure 3: Simple Wireless Power Transmission System Using Laser Technology

4.1. Transmitter - Laser Systems

There are several types of lasers that exist for various applications. Trade-offs between the laser power (both average and peak), wavelength, propagation, and compatibility with the PV receiver need to be evaluated against system size, cost, and needed efficiency [1]. For example, the most common PV cells are Silicon (Si) and Gallium Arsenide (GaAs) with peak conversion efficiency at 900 nm and 840 nm respectively. Unfortunately, no laser has a high average power at either of these wavelengths [1].

Other lasers offer decent transition power levels with other parameters that are unacceptable. Chemical Oxygen Iodine Lasers (COIL) “demonstrated high power with moderately good beam quality (necessary for efficient propagation) but has high consumable costs and requires large infrastructures” [1].

The most common lasers for wireless power applications are Solid-State or Semiconductor (diode) lasers. The most popular solid-state laser is the Nd:YAG laser. The Nd:YAG laser uses neodymium-doped (Nd) yttrium aluminium garnet (YAG) crystal as the lasing medium and has a typical wavelength of 1064 nm. Studies by Steinsiek, Summerer, and Kawashima all use an Nd-YAG laser for theoretical and demonstration purposes [2, 6, 7]. Diode lasers, which are a subset of semiconductor lasers, operate using an active semiconductor medium that is similar to that of a light emitting diode (LED). Their frequency range most significant to WPT is the 795-830 nm range [6].

In addition to selecting the lasing medium necessary to meet requirements, the pumping process that the laser uses must be evaluated. Laser pumping is a process that raises atoms from a lower energy level to an upper energy level. Lasers can be pumped several different ways. Optically pumped lasers use light to perform the population inversion. A solar pumped, as mentioned before, can be done using sunlight directly or indirectly. Pumping can also be used to change certain attributes about the lasing medium. A Diode Pumped Solid State (DPSS) ND: YAG Laser can emit at 808 nm rather than the typical 1064 nm.

4.2. Receiver - Photovoltaic Cells

On top of optimizing the photovoltaic (PV) cells for the laser wavelength, the PV’s overall conversion efficiency also has to be considered when developing a system. While cell material changes offer small advantages in efficiencies, adding an optical concentrator element to the PV cells can make a more significant impact.

The concentrator is a glass lens that allows the light to be more intense on each individual cell. “Higher light intensities enable higher efficiencies in converting sunlight to electricity, and greatly reduces the size of the PV cell required” [9]. At University of California, Merced it has been found that a Fresnel lens can be added as a concentrator on top of either, a silicon PV cell, or a multi-junction PV cell. The Fresnel Lens, by their experiments, have proven to be the best lens type for this application. They note, however, to keep the concentrator photovoltaic (CPV) cost effective the requirement of alignment needs to become less stringent, and a lens with a wider field of view needs to be used.

A Stretched Lens Area (SLA) architecture using Fresnel lenses was developed for space applications under NASA and Auburn University [11]. The Fresnel lens was used in combination with multi-junction PV cell receivers. Their design allows for 85% less PV material to be used per Watt of power produced. The “slope error tolerance of the symmetrical (Fresnel) refraction lens is more than 100 times better” than a conventional flat or reflective concentrator. The unique arch shape of this architecture is what produces the increased slope error tolerance. This shape would also provide an excellent aerodynamic surface if placed on a UAV.

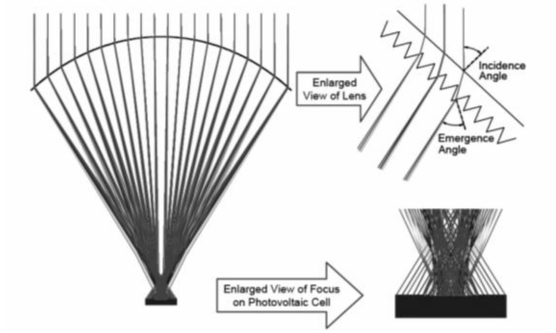


Figure 4: Symmetrical Refraction Lens with False-Color Rays Showing Wavelengths in the Photovoltaic Cell Response Range (0.36 µm to 1.80 µm for All Three Junctions of a Triple-Junction GaInP/GaAs/Ge Cell) [11]

The popular method of solar concentrators prior to O’Neill’s work was mirror solar concentrators [11]. This design deploys mirrors at 60° to focus the solar rays as they hit onto the cell area. The problem that arises with this design is slope errors with the mirrors. The rays no longer reflect evenly on the cell area. Figure 5 shows the issues associated with these errors. The vibration environment on a small UAV would make it difficult for precise placement of mirrors like this.

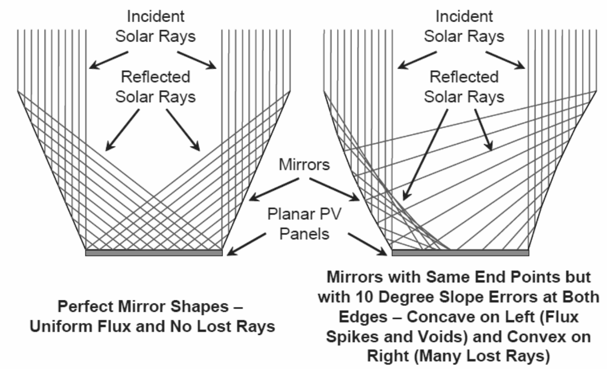

Incident Solar Rays
Reflected Solar Rays
Incident Solar Rays
Reflected Solar Rays
Mirrors
Planar PV Panels
Perfect Mirror Shapes – Uniform Flux and No Lost Rays
Mirrors with Same End Points but with 10 Degree Slope Errors at Both Edges – Concave on Left (Flux Spikes and Voids) and Convex on Right (Many Lost Rays)

Figure 5: Ray Traces for 60° Tilted Mirrors for Perfect Mirrors (Left) and for Mirrors with Shape Errors (Right) [11]

In the early 90's the Photovoltaic Array Space Power (PASP) Plus Mission was the first deployment of this type of concentrator technology. The lenses were used on multi-junction cells (GaAs over GaSb). The design performed well and was able to withstand cell voltage excursions to 500V.

In 1998, Solar Concentrator Array with Refractive Linear Element Technology (SCARLET) launched with silicone Fresnel lens to focus sunlight with 8 times concentration onto triple-junction cells. The cells were able to produce 200 W/m^2, which was the best metric performed to date. The SCARLET was the basis for the O'Neill's SLA architecture development.

Figure 6 shows the deployment mechanism and the lens once deployed. The receiving cells in this design are triple-junction (GaInP/GaAs/Ge) cells. The efficiency found was approximately 27.5% when tested as NASA's Glenn facility. The lens efficiency for this design was near 90%.

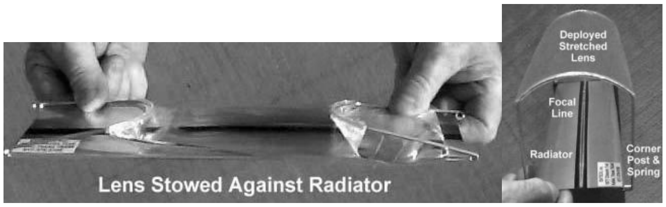


Figure 6: Model Showing Basic Stretched Lens Approach [11]

Current products that exist in concentrator technology include panels by LightPath and 3M. LightPath has developed Gradium® glass solar concentrator lens. These cells were developed for space applications through a contract with AFRL. The product is said to offer a "wide field of view" to minimize the need for re-pointing [10].

5. Demonstrated or Proposed WPT systems

While demonstrated and proposed laser WPT systems are not as well refined as their microwave counterparts, several studies have been done advancing the technology and capabilities of such systems. Most of the studies focus on the overall system design and proof of concept activities, saving system performance improvements for follow-up activities.

5.1. PowerSphere (PS)

In Ortabasi's work, a device called a Powersphere (PS) is developed. The PS is "a high efficiency Photovoltaic Cavity Converter (PVCC) that is under development for Wireless Power Transmission (WPT)" [1]. The PS system, pictured in Figure 7, uses a "lamp pumped Nd:YAG laser operated in the CW mode, with a simple flat-flat cavity resonator" [1]. This design uses a near lambertian beam-diffuser to scatter the light onto the solar panels that reside around the cavity (Figure 8 shows half of the sphere). The properties of the lambertian diffuser allow the light to scatter such that the reflected light is of identical brightness, regardless of the viewing angle.

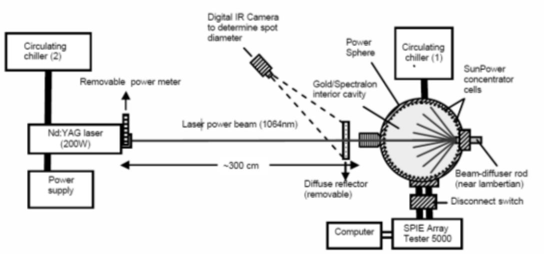

Digital IR Camera to determine spot diameter
Circulating chiller (2)
Removable power meter
Nd:YAG laser (200W)
Laser power beam (1064nm)
Power supply
~300 cm
Gold/Spectralon interior cavity
Power Sphere
Circulating chiller (1)
SunPower concentrator cells
Diffuse reflector (removable)
Beam-diffuser rod (near lambertian)
Disconnect switch
Computer
SPIE Array Tester 5000

Figure 7: Experimental PowerSphere Setup at United Innovations in San Marcos, CA, USA [1]

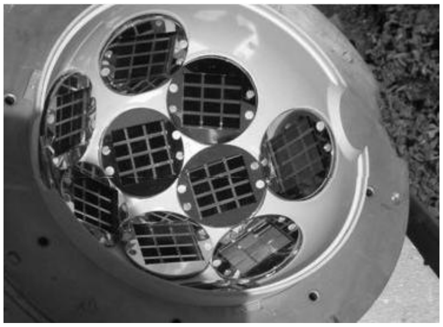


Figure 8: The interior of the PowerSphere (one of the two hemispheres) [1]

The test system operated at about 14% efficiency using SunPower Silicon (Si) concentrator cells (HEDA312) [1]. The low efficiency was attributed to the fact that: "a) Si cells are not a good match for the 1064nm wavelength, b) the flux density inside the sphere is 30% less than one sun, though the cells are optimized for 500 suns, c) the standard AR coating for the test cells inside the PowerSphere have a reflectance of ~15% at 1064nm and, d) the cell population inside the cavity is only 24%" [1]. With corrections, Ortabasi believes 40% efficiency can be achieved, working towards 60% if better matched PV cell are used.

The PS design shows how optical methods and geometrical techniques can increase the efficiency of the PV cells. If pointing accuracy is a problem with the system, a reflective dish that concentrates the beam on a center disk containing PV cells (much like a satellite dish) could also be a technique worth exploring.

5.2. Powering of Remote Rover for Space Applications

In 2005 the European Aeronautic Defense And Space (EADS) Company developed a Space Power Infrastructure (SPI) project "aim(ing) at a commercial application of Power from Space in the long term, embedded in an international economical, political and legal network" [2]. EADS designed a system to power a rover from a distance of 30-200 m away. The application for this technology was looking at a rover that would be exploring a dark part of a planet, and a ground relay station that had access to sun. The ground relay station would capture the energy and transmit it to the rover. The paper investigates the development of a small demonstration system and analysis on how to build to a full scale model.

The test system was developed using a 5 Watt, Nd:YAG Solid State Laser. A 532 nm wavelength was used in testing for visibility purposes. The final system would be optimized to use a 1064 nm IR laser. Foil was placed around the cells to reflect light around the PV cells. This reflection provided feedback to the transmitter, allowing it to adjust pointing as needed. Gallium Indium Phosphide (GaInP) PV cells were used. GaInP has a 1.85 eV band gap and was "optimized with respect to the Gaussian laser beam profile" [2]. Based on the measurements made with the test system, it was determined that a 40% efficiency could be achieved.

Building on this initial test system, a second test system and third demonstration system were proposed. The second system would have a stationary laser transmitter with a relay airship, and a mobile rover. Its design would increase the power level, increase the transmission distance, try to improve the pointing accuracy, and potentially include data transmission capability. The demonstrator system could be run off the International Space Station and evaluate the ability to beam power to earth. This work was proposed to discuss the possibilities of WPT for remote powering of rovers on exploratory missions in space, and lay foundation for eventual Solar Power Plants.

5.3. Powering of Unmanned Systems

In Kawashima and Takeda's work, students developed a laser WPT system to power small autonomous vehicles [7]. Three systems were designed and tested at an indoor facility. The first was laser powering a rover. A 60 W laser diode was driven through a 400 µm fiber, transmitting to a 70 cm diameter solar panel receiver. The transmission was done at a distance of 1 km and produced a 20% overall efficiency.

The second setup used a small kiteplane. This system used a 200 W, 808 nm laser diodes driven through a 400 µm fiber to a 30 cm diameter receiving GaAs solar panel. GaAs provides 40% conversion efficiency per cell, to result in an overall panel efficiency of 25%. The average power measured on board was about 40 W. Figure 9 shows the system configuration for the kiteplane test. This system required a Peltier cooling system to be used due to the power levels.

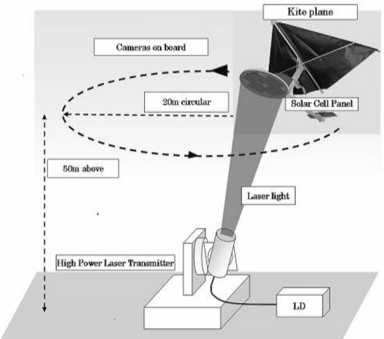


Figure 9: System configuration of a laser energy driven kiteplane [7]

A third test system was developed using a small helicopter shown in Figure 10. This system used a 530W laser and the same GaAs receiving solar array as the kiteplane. The array was placed under the propellers to provide the cooling needed for the system.

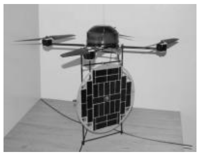


Figure 10: Helicopter with a vertical solar panel underneath [7]

5.4. Current Product Development in WPT

PowerBeam is a small company that has already started to explore the commercial options of WPT [3]. Using lasers, PowerBeam has developed small, safe transmitters and receivers to power things like speakers, computers, and televisions. Their system consists of a number of IR lasers on the transmitter side, and a PV cell detector on the receiving end. The laser operates at 1400+ nm with a collimated beam. This allows the beam to traverse relatively large distances (1 m to 100 m) with minimal loss in power or efficiency.

LaserMotive is another privately owned company, that is developing systems in WPT. While most of LaserMotives work was concentrated on winning the space elevator competition over the past few years, they recognize the importance of what they developed. LaserMotive released a white paper in March of 2010 about how WPT laser systems could be used to power or recharge small UAV's [5]. Figure 11 shows LaserMotives proposed systems.

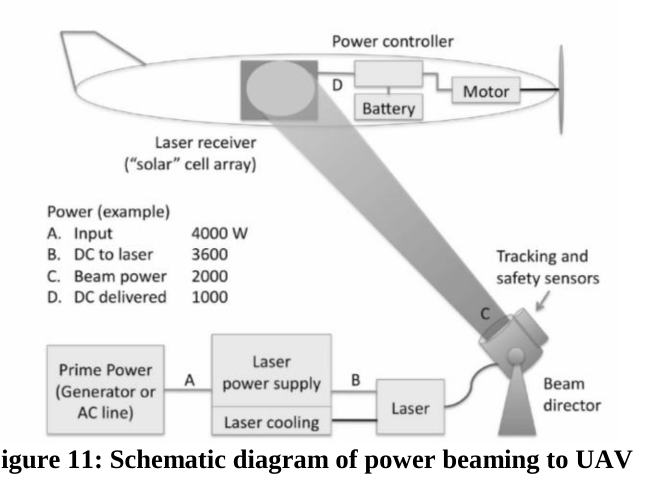


Figure 11: Schematic diagram of power beaming to UAV [5]

Initial concepts would look to use a near-infrared laser diode with >50% DC power into light efficiencies. For longer distances, a diode pumped fiber laser would need to be used. This change would create a lower divergence beam, smaller transmitter, but would increase the cost of the system.

PV cells would be the optimal receiver due to how developed the technology is currently. The cells for such a system would need to be matched to the laser wavelength and beam intensity to provide the best efficiency possible. [5]

LaserMotive highlights UAV applications of interest to be: station keeping, extended/multi-mission operations, and unlimited patrol. A safety system is also discussed that would shut the beam off if the path was blocked for any reason. The white paper provides the first real step towards laser recharging and identifies future steps that need to be taken.

6. Conclusion and Research Area Needs

While this study shows substantial research and test systems that have been completed in the area of WPT, all the authors' list similar future work opportunities.

The top areas of need include:

- Tests in increased power levels
- Tests with increased transmission distances
- Increase in Efficiency
  - On the front end, the electric to lasing conversion
  - On the back end, the beam to electric conversion using PV cells
- Improved Pointing Accuracy
- Development in efficient lightweight cooling systems
- PV cell development to better match lasing wavelength

While the first two areas are more suited for industry, the last four provide room for academic development. As mentioned, increasing the efficiencies of the PV cells can come from using optical techniques like concentrators on the cells, or methods such as the work being done for the PowerSphere. Pointing accuracy through electronic or mechanical steering seems to be an area not thoroughly explored by this community. This area might provide the most opportunity for discovery. Cooling techniques and PV cell matching are areas that will most likely need an element of academic and industry to get the most accomplished.

Wireless Power Transmission has felt a resurge in the last several years for a number of reasons. The military surge in investment for autonomous vehicles, will be a large driver for innovation in the area of WPT. Socially there is also a need for clean renewable energy. NASA continues to invest heavily in robotic missions, where power is always an important and limiting factor. Laser WPT provides an answer for all these needs, and has the proven potential to be extremely successful.

7. Acknowledgment

This work was supported in part by the Nevada EPSCoR Programs, and funded by NSF "Grant # NSF-NV 0814372", and in part by University of Nevada, Las Vegas Graduate College.

8. References

[1] Ortabasi, Ugur; Friedman, Herbert; , "Powersphere: A Photovoltaic Cavity Converter for Wireless Power Transmission using High Power Lasers," Photovoltaic Energy Conversion, Conference Record of the 2006 IEEE 4th World Conference on , vol.1, no., pp.126-129, May 2006.

[2] Steinsiek, F.; Weber, K.H.; Foth, W.P.; Foth, H.J.; Schafer, C.; , "Wireless power transmission technology development and demonstrations," Recent Advances in Space Technologies, 2005. RAST 2005. Proceedings of 2nd International Conference on , vol., no., pp. 140- 149, 9-11 June 2005.

[3] "Technology – How PowerBeam Works." PowerBeam Wireless Electricity Inc. Web. 15 Sept. 2010.

[4] Gray, Richard, "Lasers to Beam Energy to Earth from Space," The Telegraph. 23 January 2010.

[5] Nugent, T.J.; Kare J.T.;, "Laser Power for UAVs: A White Paper," LaserMotive, LLC. March 2010.

[6] Summerer, Leopold; Purcell, Oisin;, "Concepts for Wireless Energy tyransmission via Laser," Europeans Space Agency (ESA) - Advanced Concepts Team, 2009.

[7] Nobuki Kawashima and Kazuya Takeda (2008). "Laser Energy Transmission for a Wireless Energy Supply to Robots," Robotics and Automation in Construction, Carlos Balaguer and Mohamed Abderrahim (Ed.), ISBN: 978-953-7619-13-8, InTech.

[8] "High Altitude Long Endurance UAV – Zephyr." QinetiQ. Web. 10 Oct. 2010.

[9] "Efficient Fresnel Lens Concentrator for Solar Cells," University of California, Merced. Web. 15 September 2010.

[10] "GRADIUM® Lenses." LightPath Technologies. Web. 10 Dec. 2010.

[11] O'Neill, M.J, "Ultralight stretched Fresnel lens solar concentrator for space power applications," Optical Materials and Structures Technologies. Edited by Goodman, William A. Proceedings of the SPIE, Volume 5179, pp. 116-126, 2003.

[12] Schafer, Christian; Matoba, Osamu; Kaya, Nobuyuki, "Optical Retrodirective Tracking System Approach Using an Array of Phase Conjugators for Communication and Power Transmission," Applied Optics, 46(21):4633, July 2007.

[13] Fikes, J.C.; Howell, J.; Mankins, J.C.;, "Space Solar Power Technology Demonstration for Lunar Polar Applications: Laser-Photovoltaic Wireless Power Transmission," Pratt and Whitney Rocketdyne (PWR) Engineering.

[14] Paschotta, Rüdiger, "High-Power Lasers," Encyclopedia of Laser Physics and Technology Ed 1, October 2008.

Optical Character Recognition of Non-flat Small Documents Using Android: A Case Study

Leslie Sears[1], Ray Hashemi[1], and Mark Smith[2]

[1]Department of Computer Science
Armstrong Atlantic University
Savannah, GA 31419, USA

[2]Department of Computer Science
University of Central Arkansas
Conway, AR 72035, USA

Abstract

Optical Character Recognition (OCR) using Android has been received a great attention. In all these efforts a flat document is selected for OCRing and non-flat documents are ignored. Labels mounted on cylindrical surfaces such as wine bottle, pill box, cans, etc, are examples of non-flat documents. The goal of this research effort is to perform optical character recognition on a non-flat document. To be more specific, we investigate OCRing of a cylindrical pill box label. Two pictures of the label are needed for the OCRing. The methodology was applied on 30 synthesized non-flat labels and on average, the system was able to accurately recognize 92.4% of the characters.

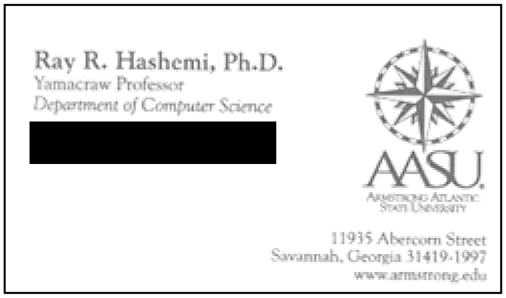
Key Words: *Optical Character Recognition, Non-flat documents, Android, Mobile Device, and Segmentation.*

1. Introduction

Optical character recognition using Android has been received a great attention [1, 2, 3]. In all these efforts a flat small document such as a business card, Figure 1, is selected for OCRing. However, there are a large number of applications that demand OCRing of small documents that are not flat. Labels mounted on cylindrical surfaces such as wine bottle, pill box, cans, etc, are examples of non-flat documents, Figure 2.

The goal of this research effort is to perform optical character recognition on a non-flat surface. To be more specific, we investigate OCRing of a cylindrical pill box label.

The rest of the paper is organized as follow: The imaging is the subject of section 2. Sharpening is covered in section 3. Methodology is presented in section 4. Experimental results are discussed in section 5. Conclusion and future research are the subjects of section 6.

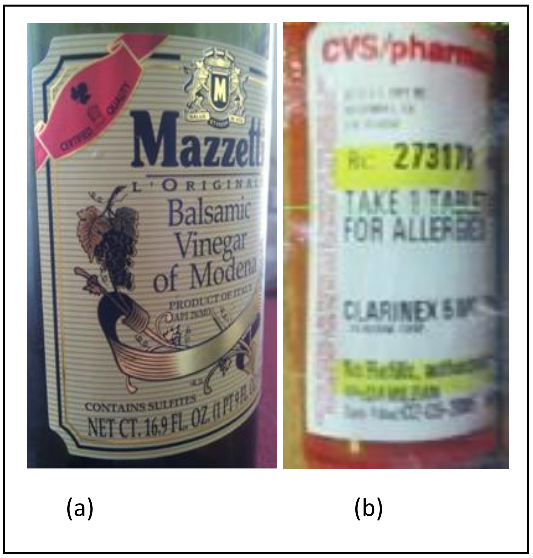


Ray R. Hashemi, Ph.D.
Yamacraw Professor
Department of Computer Science
[redacted]

AASU
Armstrong Atlantic State University

11935 Abercorn Street
Savannah, Georgia 31419-1997
www.armstrong.edu

Figure 1: An example of a flat small document.



(a) (b)

Figure 2: Examples of non-flat small documents: (a) label of a vinegar bottle and (b) label on a pill box.

2. Imaging

When taking a picture of a label on a non-flat surface the text cannot be adequately captured in a single image. To handle this situation, the system processes the first image and then gives the user the option of taking another picture to capture the remaining text. For the second picture, user rotates the bottle to bring the other text into view while providing some overlap with the first image.

Each image is captured into memory as a compressed byte array stored internally in the JPEG format as an android.graphics.Bitmap object. The mobile device has a 5 Megapixel camera which produces an image with a width of 2048 pixels and a height of 1536 pixels. To reduce the number of computations and thereby improve interactivity, the image is scaled to a width of 1280 pixels and a height of 960 pixels, preserving the original image aspect ratio. Since the text characters of interest are black on a white background, the color information is not necessary and the image is converted into a 256 level grayscale image. Additionally, converting the image to grayscale reduces the image's size, conserves memory and increases computational efficiency in latter steps by reducing the amount of image data.

When converting color images to grayscale for the purpose of human consumption, it is important to preserve the luminance values so that the relative brightness of the grayscale image matches that of the original color image. This is typically done by using a weighted average method in which the weights are assigned based on the human eye's sensitivity to each of the primary colors. The NTSC values for the weights are .30 for red, .59 for green and .11 for blue [4]. Given that the pill bottle label images are not intended to be viewed by the user and the black text tends to have red, green and blue values that are nearly identical, it is not necessary to use the weighted average in this case. Instead, the three colors are averaged together un-weighted to reduce computational complexity and, thereby, processing time. This is done using Equation 1 by taking each of the color bands of Red (R), Green (G) and Blue (B) and finding the average of the three colors.

$$Gray = \frac{(Red+Green+Blue)}{3} \qquad (1)$$

3. Sharpening

The final step of preprocessing enhances the sharpness of the image so the individual characters in the image have pixel values that are closer to one another. Due to the nature of capturing the image from a camera, many of the pixels around the edges of the characters have varying degrees of blackness. Enhancing the sharpness of the image will aid in segmenting the image accurately. To sharpen the image an Unsharp Mask (USM) filter is employed [5]. The USM filter uses a Gaussian kernel, formula (2), to create a blurred image of the original image.

$$G(i) = \frac{1}{\sqrt{2\pi r^2}} e^{-\frac{i^2}{2\sigma^2}} \qquad (2)$$

The radius r determines the width of the Gaussian kernel which in turn determines how many of the surrounding pixels are used to influence the value of the pixel under consideration. The value i is the distance from the center of the kernel. To create the blurred image *B*, the one dimensional Gaussian kernel *G* is applied both vertically and horizontally to the pixels of the two dimensional image's matrix of pixels *I*. This is done in two steps by first applying Equation 3 to the pixels in row-column order and then a second time in column-row order. In each step, the value of the blurred pixel *B(x)* is obtained from the sum of the surrounding pixels of pixel I*(x)* of the original image's matrix multiplied by the respective values of the Gaussian kernel *G(i)*, formula (3).

$$B(x) = \sum_{i=-r}^{r} I(x+i)G(i) \quad (3)$$

The sharpened image *S* is produced using formula 4, in which the blurred image is subtracted from the original image *I* taking into consideration the *amount, A*, and *threshold, T*. The value *A* determines the magnitude of the effect the kernel has on the pixel value and the *T* determines if the filter will be applied to an individual pixel based on the minimum difference between the blurred and original pixel values. During application of the USM the *A* and *T* are given.

$$S(x) = \begin{cases} |I(x) - B(x)| < T, \quad I(x) \\ \\ |I(x) - B(x)| > T, \\ \quad A * \big(I(x) - B(x)\big) + B(x) \end{cases} \qquad (4)$$

The pixel values of the sharpened image *S* are then normalized to be in the 0 – 255 range by setting values less than 0 to 0 and values greater than 255 to 255. Figure 3 demonstrates the effect sharpening has on a single character in the image.

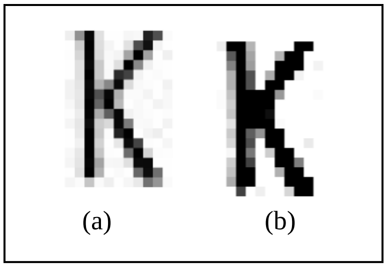


Figure 3: Effects of Unsharp Mask (USM) on a single Character: (a) Original Image and (b) After applying USM

4. Methodology

To capture a non-flat small document, two images are taken. The first logical step is to stitch the two images into one and then the OCR process be applied on the stitched image. However, we decided to postpone the stitching process for a later stage. To explain it further, in recent years, there has been considerable research into how to properly stitch multiple images into a single image for human viewing [6, 7]. The most successful approach uses scale and rotation invariant features that are common in the images [8]; frequently corner detection is employed. Since the images in this project are not intended for human consumption, there's simply no need to combine the images. Finding the common text among the images is a more efficient and reliable method. In addition, the Java heap limit of 17MB would require a combined image to be processed at a much lower resolution, as there is simply not enough memory available to process a single large image in memory. The lower resolution would affect the number of pixels that comprise each character and reduce the accuracy of the character decoding process.

As a result the stitching takes place after the individual images of the same label are gone through the process of character recognition. This process is composed of the following steps: segmentation of the image and filtering the segments, character decoding, region filtering by ratio, construction of lines of characters in each processed image, and stitching the processed images to create full Lines of text. Each step is described in the following subsections.

4.1 Segmentation and Filtering

Segmentation is the process of grouping pixels into regions in which all pixels in the region share a common characteristic and the pixels are spatially related to one another. Two pixels are defined as *connected* to another if one pixel is in the 8 pixel neighborhood of the other. Two pixels are *similar* if the absolute value of the difference of the two pixel's grayscale values are less than the given threshold. All the pixels that are connected and similar make a region. The seed of each region is selected randomly seed selection is continued after each region is grown

An *unassigned* pixel is a pixel that has not yet been identified as belonging to a region. Conversely, an *assigned* pixel is a pixel that has been identified as belonging to a specific region. An *ungrown* pixel is a pixel that has been assigned to a region but has not yet examined its neighboring pixels. Again, conversely, a *grown* pixel is a pixel that has been *assigned* and has examined its neighbors.

The region growing algorithm, Figure 4, used for this project is the succession of selecting an *unassigned* pixel and/or an *ungrown* pixel. Next, examining connected pixels, assigning the similar connected pixels to the current pixel's region. And then marking that connected pixel as *assigned* but *ungrown* and finally marking the current pixel as *grown*.

```
Algorithm: GrowRegions
Given:  A 2D matrix of pixels as an image
        (imageMatrix).
Objective: Return a list of regions.
 1: Set done ← false;
 2: Set regionIndex ← 0;
 3: Set lastUnassigned ← 0;
 4: Set unAssignedList ← null;
 5: Set ungrownList ← null;
 6: Set regionList ← null;
 7: Repeat until not done
      Set firstUnassigned ← -1;
      Set firstUngrown ← -1;
      If ungrownList is not empty
      Then
         Set ungrown ← ungrownList.next;
         ungrownList remove ungrown;
         check neighbors of ungrown;
         If firstUngrown = -1
         Then
            If unAssignedList is empty
            Then
                 Set ← done = true;
            SetfirstUnassigned ←unAssignLixt.next;
            Set region ← regionIndex++;
            regionList add region;
            ungrownList add firstUnassigned;
            check neighbors of firstUnassigned;
 8: Return regionList;
 9: End;
```

Figure 4: Region Growing Algorithm.

Considering the fact that characters on the label are darker than the background, it is logical that each region is a character, logo, two connected characters, portion of a character, or a blub courtesy of droopy corners and poor lighting. To make a distinction among the resulting regions, each region is boxed by its Minimum Bounding Rectangle (MBR). The MBR is computed by defining the top left and bottom right corners of the region. The coordinates for the top left corner are found by taking the minimum x and y values from the set of pixels in the region and the bottom right corner is found by taking the maximum x and y values from the set region pixels.

Given that the domain of the task is limited to finding characters on a pill bottle label, there are some reasonable assumptions that can be made regarding the grayscale value, size and shape of boxed regions that are likely to be characters. First, since prescription labels contain black text on a white background, all regions whose mean pixel value is above a certain threshold (t_{mean}) are removed. These regions are too white to be text characters. Second, thresholds are defined for minimum size(t_{min}) and maximum size (t_{max}) of a boxed region based on the size range of characters on the label. Boxed regions that do not meet the size characteristics are removed from the list of regions. The remaining regions are subject to character decoding process.

4.2 Character Decoding

For decoding each eligible region into the encompassed characters, two assumptions are made. First, though not required by regulation [9], for every label examined, the dosing instruction was a plain sans serif font of only capital letters. As such, the decoding algorithm is designed to decode only the capital English characters A through Z and the numerals 0 through 9. Second, the character decoding is not designed to handle rotated characters. It is expected that the user is able to take a properly oriented image of the pill bottle label with the text upright and within a few degrees of horizontal. Even if the label had been misplaced on the bottle, the user should orient the text in a upright position. While there are numerous methods to decode a set of pixels into a character, this project relies on identifying certain characteristics of the region and comparing those characteristics to the set of known characteristics for each of the characters.

This method of decoding by characteristics is scale invariant which is useful as it eliminates the need to scale the characters to a known size before decoding. A *cut* is a transition from an edge of the boxed region to black pixels or any transition of white pixels to black pixels in the boxed region. The characteristics determined for each boxed region are (1) the presence of straight vertical or horizontal lines at the edges of the region and (2) the number of *cuts*.

To determine if a boxed region contains a straight line of pixels near the edges, a process counts the longest line of pixels near each of the edges of the MBR of the region. Figure 8 shows the areas of the boxed region that are examined. If the boxed region contains a vertical or horizontal line of pixels that is greater than 75% of the MBR in the edge of the region, a "1" is placed in an encoded bit string, otherwise a "0". For example, the boxed region of Figure 5.a has been evaluated for straight lines in the shaded areas (Figure 5.b) and three lines at the left, top, and bottom are detected. As a result, the boxed region is encoded with the bit string of "1011".

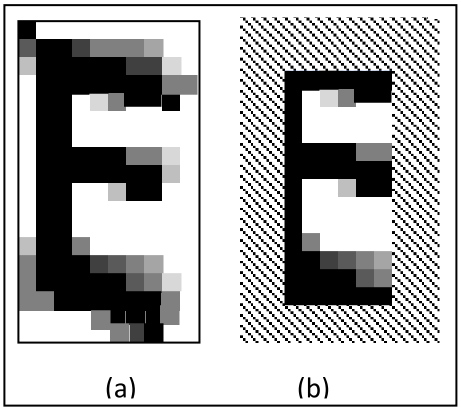


Figure 5: Edge Line Examples: (a) a boxed region and (b) the areas (shaded) on the region checked for lines of pixels.

The encoded bit string is four bits long and represents the four edges of the boxed regions in order of left, right, top and bottom. The encoded bit string is then used to create subsets of the potential characters. Due to irregularities in the image, certain characters belong to multiple subsets. For instance, for the character "B" the top edge of the character may or not meet the 75% threshold. For this reason, "B" belongs to the two character subsets defined by the encoded bit strings "1011", "1010".

In the second, the number of *cuts* is determined and is encoded into a value string. Five horizontal and vertical lines are drawn on the boxed region. The horizontal and vertical lines are drawn at 10%, 25%, 50%, 75% and 90% of the height and width of the boxed region, respectively. For each line the number of cuts is counted. The frequencies collectively make the value string representing the boxed region. For example, In Figure 6, numbers of cuts for the five horizontal lines from top to bottom are 1, 1, 1, 1, and 1". The numbers of cuts for five vertical lines from left to right are 1, 3, 3, 3, and 2. As a result, the boxed region is encoded as “1111113332” which is called *cut string*. In addition, for a well defined character, the number of *cuts* was determined using the same five horizontal and vertical lines and encoded as *expected string.*

Figure 7 lists the expected encoding of the characters. Each cut string is compared to expected strings of the reduced character set using Equations 5 and 6. The boxed region is decoded to the character in which the difference in the value strings is minimized.

$$d_i = \sum_{j=1}^{n} \left| u_j - k_j^i \right| \text{ (for i= 1 to m)} \qquad (5)$$

$$d_{min} = min\{d_1, d_2, d_3, \ldots, d_i\} \qquad (6)$$

Where, u is the unknown encoded string, k is the known encoded string, n is the number of bits in u, and m is the number of encoded characters in Figure 7.

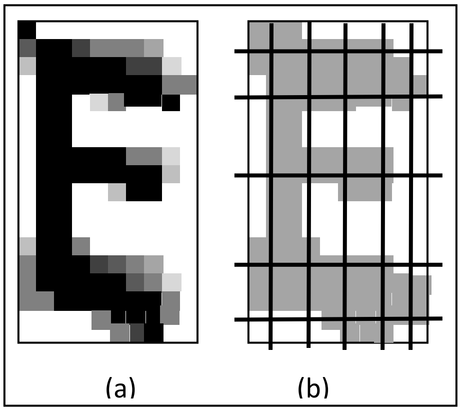


Figure 6: Determining the number of cuts: (a) a boxed region and (b) Horizontal and vertical lines drawn on the boxed region.

Character Decoding				
No Edge Lines - 0000			**Left and Right Edges Only - 1100**	
A	1222211211		H	2212211111
C	1111112222		M	2244311111
G	1112112222		M	2343211111
I	1111122122		N	2333211111
I	1111111111		N	2232211111
O	1222112221		U	2222111111
Q	1222112222		**Right and Bottom Edges Only - 0101**	
S	1111122322		J	1111112211
V	2221111111		**Left and Bottom Edges Only - 1001**	
W	3322211111		D	1222112221
X	2212212121		L	1111111111
Y	2211111111		**Top and Bottom Edges Only - 0011**	
6	1122111332		Z	1111123322
G	1122112232		S	1111122322
8	1212114341		3	1111123332
9	1221122331		5	1111123322
4	1221112211		I	1111122122
Left Edge Only - 1000			**Left, Top and Bottom Edges Only - 1011**	
D	1222112221		B	1222213331
K	2212211221		D	1222112221
R	2212212221		E	1111111332
Right Edge Only - 0100			**Left and Top Edges Only - 1010**	
J	1111112211		F	1111112221
4	1121111211		P	1211112211
W	2332211111		R	1212212221
Top Edge Only - 0010			B	1222113342
T	1111111111		**Left, Right and Bottom Edges Only - 0111**	
7	1111122211		I	1111122122
5	1111123322		J	1111112211
Bottom Edge Only - 0001				
2	1111112322			
S	1111122331			
1	1111122111			

Figure 7: Encoded Characters.

4.6 Region Filtering by Ratio

Since all English characters and Arabic numerals are roughly square, with the exception of “1” and “I”, Thresholds are defined for the minimum and maximum ratio of width to height of the regions and because we have the two exceptions to the ratio requirements, boxed “1” and “I”, the regions are not filtered by ratio until they have been decoded. Filtering by ratio removes characters that have been compressed horizontally due to the curvature of the pill bottle. Those characters, at the visible edge of the rounded pill bottle, will be captured in the subsequent overlapping images.

4.4 Lines Construction

The decoded characters must are organized properly into lines of text. This task is done by sorting regions based on their position in the image. The first level of sorting places the image in order from top to bottom. The second level sorts the regions that are considered the same distance from the top of the image within a given threshold from left to right. The vertical threshold is employed to place characters on the same line that may not be perfectly aligned due to the curvature of the label on the bottle or because of a slight rotation in the image. As stated previously, significant rotation of the image is not accounted for or necessary. After the regions are sorted, lines of text are constructed. Starting with the first region in the sorted list, the decoded character is used to build a string. Moving through the sorted list, the decoded characters from each region are appended to the string until either of the following two conditions is encountered. The first condition verifies the horizontal difference in position of the previous region to the current region. If greater than a given threshold, a space character is appended to the current string before the decoded character. The second condition verifies the vertical difference in position from the previous region and the current region. If the distance is greater than the given threshold, a new line string is created to represent the new line of text.

4.5 Stitch Images to Create Full Lines of Text

The lines of text from the individual images of the non-flat document need to be combined to create the full lines of text. Algorithm Stitch, Figure 8, accomplishes this process by taking the first line of text from the two successive images and searching for the longest common substring of characters. The longest common substring represents the area in which the images overlap. The common characters are removed from the line of text in the second image and the remaining characters are appended to the line of text from the first image. This process continues until the extracted text from each of the images has been appended. The algorithm for finding the longest common substring uses a traditional “brute-force” method. Although there are more efficient algorithms based on dynamic programming methods [10], they require more memory and are not necessary since the document is small.

```
Algorithm: StitchLines
Given: Text line extracted from image one
    (firstStr) and image two (secondStr).
Objective: firstStr and secondStr merged
    together at longest overlap.
 1: Set subLongestLen ← 0;
 2: Set subLongestInd ← 0;
 3: Repeat for each character, a, in firstStr
      Set subIndex ← 0;
      Repeat for each character, b, in secondStr
          subIndex ← subIndex + 1;
          If a = b
          Then  Set subLen ← 0;
                Let c ← 1;
                Repeat do while a + c = b + c
                     subLen ← subLen + 1;
                If subLen > subLongestLen
                Then
                   subLongestLen←subLen;
                   subLongestInd←subIndex;
 4: Set stitchedStr ←
    firstStr.SUBSTR(0,subLongestInd) +
    secondStr;
 5: End:
```

Figure 8: Line Stitch Algorithm.

5. Experimental Results

Android based mobile devices present many unique challenges. Among them are the device’s limited processing power and available memory. The particular device used for this project was the Motorola Droid. It has an ARM Cortex A8 processor running at 550Mhz, 256 MB of internal RAM and a 5 megapixel camera. The Android OS provides a Java Software Development Kit for creating custom applications which further constrains the application environment by limiting the Java heap size to a maximum of 17MB.

The methodology was applied on 30 images of synthesized non-flat documents. The synthesized labels were contained only the relevant information. For example, the pharmacy logo, barcode, prescriber name, was not included in the label.

Due to the cylindrical nature of the documents two pictures of each document were required to capture the full text of the document. The pictures are processed separately and then resulting characters for each line segment were stitched to make a complete line. The pill bottle were located on a flat surface. The mobile device was not mounted on a stand for taking the picture of the non-flat label and the second picture was taken roughly from the same distance. The system was able to recognize 92.4% of the characters.

We also applied the methodology on the same synthesized labels as flat small documents. In this case only one picture was taken from the label and 94.4% of the characters were recognized correctly.

During the experimental evaluation it was noted that proper and even illumination was required to obtain consistent results. In addition, the camera on the mobile device is better suited for taking snapshots and is difficult to hold steady when taking close-ups of the labels. These issues could be remedied by providing a stand to hold the camera and the prescription bottle in the proper position.

6. Conclusion and Future Research

The results revealed that application of the presented methodology for optical character recognition of the non-flat small documents is an effective one and potentially it is a powerful one.

As future research extracting the dosing instructions and other information from actual prescription pill bottle labels that includes other text has been planned.

7. References

1. Kevin Jackson, :CamCard Makes Collecting Business Cards Easier", posted in "Android Software", http://www.androidthoughts.com/tags/ocr, 2010.

2. Arnold Zafra, Best Business Card Reader App for Android, Simon Hall (Editor and Publisher), http://www.brighthub.com/mobile/google-android/articles/93647.aspx, Oct, 2010.

3. Jenny Curtis, "DocuScan Plus Offers OCR for iPhone and Android", http://www.ocrworld.com/software/4-news/362-docuscan-plus-offers-ocr-for-iphone-and-android.html, December 2010.

4. Pratt, W. K. (2001). *Digital Image Processing* (3rd ed.). John Wiley and Sons.

5. Gonzalez, R. C. and R. E. Woods (2001). *Digital Image Processing*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

6. Nguyen, H. (2005). "Panorama Imaging For Cell Phones (Java Image Processing Development)." MS Project, Armstrong Atlantic State University.

7. Szeliski, R. (2006). Image alignment and stitching: a tutorial. *Found. Trends. Comput. Graph. Vis. 2* (1), 1-104.

8. Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision 60* (2), 91-110.

9. "General Labeling Provisions." Code of Federal Regulations Title 21, Pt. 1.B.201, 2008. http://ecfr.gpoaccess.gov.

10. Cormen, T. H., C. Stein, C. E. Leiserson, and R. L. Rivest (2001). Introduction to Algorithms (2nd Revised edition ed.). B&T. pp. 350-356.

SESSION

APPLICATIONS AND RELATED DISCUSSIONS

Chair(s)

TBA

XML Schema for Aircraft Conceptual Model Representation

Shubhangi G. Deshpande, Layne T. Watson,
and Robert A. Canfield
Departments of Computer Science, Mathematics,
and Aerospace and Ocean Engineering
Virginia Polytechnic Institute
and State University
Blacksburg, VA 24061

Maxwell Blair
and
Philip S. Beran
Air Force Research Laboratory
Wright Patterson AFB
Dayton, OH 45433

Abstract. *Today's modern conceptual aircraft designs exhibit a strong, nonlinear, interdisciplinary coupling and require a multidisciplinary, collaborative approach. Efficient transfer, sharing, and manipulation of aircraft design and analysis data in such a collaborative environment demands a formal structured representation of data. XML, a W3C recommendation, is one such standard concomitant with a number of powerful capabilities, thus alleviating interoperability issues involved in a collaborative environment. A generic XML schema for an aircraft design markup language is proposed to represent aircraft data, mathematical models, and geometry. The purpose of this unified data format is to provide a common language for discourse and data communication, and to improve efficiency and productivity within a multidisciplinary, collaborative aricraft design environment. The ultimate goal is to develop a generic, comprehensive, and compact XML schema for representing conceptual aircraft design data and models including, but not limited to aircraft geometry, structural layout, and materials.*

Keywords. conceptual aircraft design; multidisciplinary, collaborative environment; interoperability issues; XML schema

1. Introduction

Conceptual aircraft design is characterized by a large number of design alternatives and trade studies, and a continuous, evolutionary change to the aircraft concepts under consideration [13]. Numerous design alternatives are studied during the conceptual design phase. Traditional conceptual aircraft design systems were developed as isolated disciplines with a minimal interdisciplinary coupling and mostly linear interactions between disciplines. However, today's modern aerospace systems exhibit strong, nonlinear, interdisciplinary coupling and require a multidisciplinary, collaborative approach [15]. It has become more of a collaborative endeavor that involves many individuals from diverse groups around the world working together in an extended enterprise environment to achieve a common goal. This multidisciplinary coupling in an aircraft design poses additional challenges beyond those encountered in a single disciplinary design and analysis of aircraft. As a result, multidisciplinary aircraft conceptual design and analysis need to manage a large amount of data including, but not confined to analysis inputs and outputs. Efficient transfer, sharing, and manipulation of aircraft design and analysis data across different platforms, systems, and users demands a formal structured representation of the data in a well structured data sharing and validation environment. In general, if a uniform representation is not used, the same information is stored multiple times, each time in a different format specific to the underlying implementation. In order to exchange this information between different disciplines, an import/export tool needs to be designed that converts one format to/from another at every facility and for each format. Thus, the lack of a standard, uniform representation results in redundancy in codes and duplication of information and efforts. Another common problem with this kind of data exchange is data inconsistency. All these factors greatly hinder sharing and exchanging of interdisciplinary data and make the design process less efficient. To alleviate this burden, a unified system is sought that provides certain capabilities, for modeling the massive amount of multidisciplinary data, such as portability, maintainability, reusability, platform independence, integrity, (syntactic) correctness, and system recovery. With a platform and language independent data exchange standard, like XML (extensible markup language), information can flow seamlessly in a heterogeneous environment with diverse computing platforms, programming languages, and hardware systems. The authors of this paper propose some first steps for the conceptual aircraft design and analysis community to move in this same direction.

The organization of this paper is as follows, Section 2 motivates the need for a data exchange standard. Section 3 provides a rationale for using an XML based markup language for exchanging aircraft design and analysis data. Section 4 provides an overview of the proposed XML schema, and a use case for the application and implementation of the proposed XML schema based markup language ADML is presented in Section 5. Section 6 offers concluding remarks.

2. Motivation

As discussed in [9], the multidisciplinary optimization branch at NASA Langley Research Center (LaRC) is promoting a multidisciplinary, collaborative approach and sharing of information among disciplines for aircraft design and analysis systems. The advanced engineering environments (AEE) study committee sponsored by NASA has identified a number of technical, management, cultural, and educational barriers that need to be overcome in order to realize a multidisciplinary, collaborative environment [15]. Several design requirements related to information management and integration of tools, systems, and data need to be addressed first in order to realize a unified system. Readers are referred to [15] for a detailed discussion about the interoperability issues involved in a multidisciplinary, collaborative, conceptual aircraft design system.

A platform and language independent format to represent aircraft design and analysis data is a desirable way to meet all these requirements and support a multidisciplinary system distributed across a network of heterogeneous computing environments. Over the past two decades or so, several data modeling languages and technologies have emerged for representation and exchange of product manufacturing information. IGES (initial graphics exchange specification) [8] is a language neutral data format that allows exchange of product data among computer aided design (CAD) systems. Reference [7] describes CAPRIS, a vendor-neutral system for accessing a variety of CAD systems through a unified and simple programming interface. CAPRIS maintains a BRep (boundary representation) data structure that is common to all participating CAD systems. CAPRIS relies on SOAP (simple object access protocol) for exchanging structured information and relies on XML for messaging. However, the geometry schema for CAPRIS is not publicly available. STEP (standard for exchange of product model data), a successor of IGES, is a comprehensive ISO standard (ISO 10303) that describes a mechanism to represent and exchange product data, and has been widely used in the aerospace, automobile, electrical, electronic, and other industries [4]. STEP uses a data modeling language called EXPRESS ([10], [17]) to describe and exchange product data between CAD, CAM (computer aided manufacturing), CAE (computer aided engineering), and other CA* systems.

As discussed in [11], STEP has a proven record of success in modeling aircraft geometry. STEP's EXPRESS language provides rich facilities for data modeling at the semantic level. However, unfamiliarity of today's application programmers with the traditional STEP based data modeling techniques impedes its widespread usage. Furthermore, XML has become a de facto standard for representing and exchanging digital data for several domains, including domains that are within the scope of STEP. Moreover, STEP can semantically model the high fidelity information required by many XML applications. Thus, the STEP data modeling standard and XML are complementary technologies. It is a logical next step to merge the traditional STEP technology within XML. With the integration of the two, the best of both worlds can be achieved. Therefore, the ADML developers are motivated to explore and test EXPRESS' semantic capabilities in the context of geometry and geometric sensitivities, ultimately leading to an XML schema that supports a multidisciplinary design optimization (MDO) environment.

The next section provides the rationale for proposing an XML based markup language for representing aircraft design and analysis data.

3. Rationale for using XML

XML, a W3C (World-Wide Web Consortium) recommendation, is a standard concomitant with a number of powerful capabilities (extensibility, flexibility, reusability, maintainability, and so on) and a generic, robust syntax for developing specialized markup languages. The platform, language, and vendor neutral format of XML makes it well suited to the task of satisfying multidisciplinary aircraft data requirements. In addition, the inherent hierarchical nature of XML provides a way to define structural relationships that exist in the data and facilitates employment of object oriented principles into conceptual aircraft design data. Name, attributes, and content model of an XML element are closely related to class name, properties, and composition associations in an object oriented aircraft design. Thus, with the use of an XML based markup language, it is possible to faithfully model aircraft design and analysis data as well as structural and functional relationships among different data elements.

There are several XML based languages developed for various application domains. There are compelling examples of success from various disciplines, e.g., a systems biology markup language (SBML) [6] developed for systems biology models and data; MathML [16], an XML based language developed for mathematical notations; Office Open XML, a Microsoft file format (commercial application) for storage of electronic data, and many more.

Although the aerospace industry is no exception for developing XML based standards for exchanging aircraft data and models, there are only a handful of successful examples. The JSBSim flight dynamics model software library [3] is a batch simulation application aimed at modeling flight dynamics and control for aircraft. JSBSim is an XML based model description specification where input files are supplied in XML format. These XML files contain descriptions of aerospace vehicles, engines, scripts, etc. DAVE-ML is a markup language for a draft AIAA standard [1], inspired by JSBSim, for the interchange of flight dynamics modeling data between facilities. However, there is no work known to the authors developing a generic schema for aircraft conceptual design and analysis. Both JSBSim and DAVE-ML are intended to provide a platform and language neutral format for exchanging flight dynamics modeling, verification, and documentation data where the major XML elements are mathematical objects. However, JSBSim provides its own XML tags for representing mathematical constructs (e.g., product, sum, quotient, etc.), whereas DAVE-ML uses the verbose MathML format for representing mathematical constructs.

The XML schema presented in this paper is a novel approach to provide a uniform language for discourse and data communication.

4. The proposed XML Schema

Although an XML based markup language is well suited for addressing interoperability issues involved in a multidisciplinary, collaborative environment, the actual development is not as easy as it first appears. Developing a generic, comprehensive, and compact XML schema for each and every discipline involved in aircraft conceptual design phase is a very challenging task. Every discipline has its own set of modeling requirements and constraints that adds up to the overall complexity of the final design. However, the inherent hierarchical nature of an XML schema plays a significant role in structuring various components of conceptual aircraft design. Figure 1 presents a simplified version of the taxonomy of the proposed XML schema. The specification for the ADML would need to include the capability to define aircraft data specific to each and every discipline involved in the conceptual design phase. An overview of three XML schema modules (data, functions, and aircraft geometry) and the existing and the future modeling capabilities of the proposed XML schema follows.

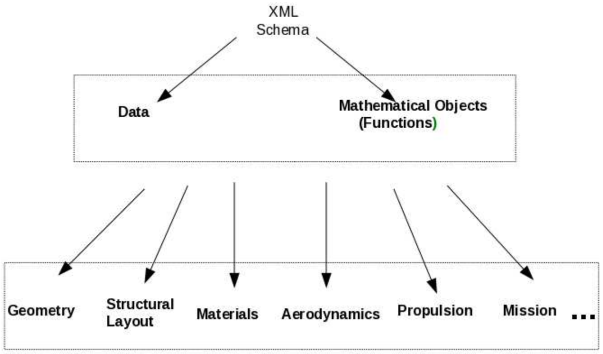


Figure 1. XML Schema Taxonomy.

4.1. Modular Schema Development

Modular schema development facilitates logical decomposition of XML elements into subsets where each individual subset focuses on specific functional capabilities thereby enabling reusability. Each small subset or module that results from this exercise can work as a building block for other more complex modules thereby enabling extensibility. The inherent modular or hierarchical structure of multidisciplinary aircraft design elicits modular schema development. The top level modules in the XML taxonomy, data and mathematical objects, serve as the foundation for developing more complex aricraft design constructs that appear at a lower level in the inheritance hierarchy. Every discipline involved in an aircraft design phase can be viewed as a separate module in the XML schema development process and can be used either as a single, isolated entity or as a part of a hierarchical structure built by combining several disciplines together. The following subsections elaborate on the fundamental modules of the proposed schema (data and functions) as well as modeling the aircraft geometry.

4.2. Data Schema

At a very high level, everything is data. However, the rationale for dividing the XML schema in different sections (data, functions, and geometry) is to exploit the functional and logical distinction among different aircraft model objects and to maintain their inherent hierarchy. In a top down view, the data schema is at the lowest level of the hierarchy, representing the simplest form of data. Elements of the data schema are used as the building blocks for all other elements. The XML schemata for more complex, high level aircraft model objects such as mathematical functions and geometry are presented in the following subsections.

The major element of the data schema is the *variable* element. Variables are used to define inputs and/or outputs to/from a design or an analysis. A variable element has a human readable *name*, a *description*, a *unit*, a *min*, a *max*, some *flags*, and other *scalar* parameters associated with it, and a machine readable variable identifiler, *vid*. A variable defined in an XML document can be referenced at a later point in a mathematical expression. The value space of a variable element consists of a scalar (an atomic value) or a tensor (a multidimensional array). A tensor is an XML element defined recursively to represent an array of arbitrary dimensions. A higher rank tensor is defined in terms of a lower rank tensor. A vector ($vtype$ element) is a rank 1 tensor and a matrix is a rank 2 tensor. In general, a k-dimensional array can be defined as a rank k tensor, e.g., a $2 \times 3 \times 4$ tensor is defined as a sequence of two 3×4 matrices that are defined in terms of three 4-dimensional vectors each. A typical use of a tensor element could be to define relational data (function tables).

4.3. Representing Math

Mathematical objects such as functions, expressions, arbitrary dimensional lists, and operators constitute a significant part of aircraft design and analysis data. Therefore, communicating mathematical objects among different disciplines plays a crucial role in exchanging data in a multidisciplinary, collaborative conceptual aircraft design and analysis environment. Careful thought has been given to a format for representing mathematical objects while developing the proposed XML schema. Three possible candidates are the XML based markup language MathML and two widely used computational software tools, Mathematica and Matlab. The most significant advantage of using a MathML format to represent mathematics is that MathML itself is an XML based markup language and can be parsed and validated easily using available XML parsers; however, MathML is an extremely verbose and unreadable format. Editing mathematical expressions in MathML requires a special editor because the markup is very complex and bandwidth intensive. This makes it impractical to edit by hand. Furthermore, the conceptual aircraft design and analysis community is more interested in communicating content rather than representing mathematical objects. Moreover, Matlab and Mathematica are among the most popular tools used to evaluate mathematical expressions in the aerospace community. Therefore, this paper proposes the use of Mathematica or Matlab syntax over the verbose MathML format for representing mathematics. However, if an application intends to parse the mathematical data being exchanged at the other end, then parsing subroutines need to be written specific to the underlying implementation. The supported format is more useful when the mathematical objects being exchanged are meant to be passed to either Matlab or Mathematica tools for evaluation. For sharing mathematical data (mathematical lists or arrays) that are meant to be parsed, an application should make use of the more relevant and easy to parse XML element, tensor, defined in the data schema.

The major schema elements for representing mathematical objects include *operator*, *relation*, *mlist*, and *expression*. These elements can be represented in either Mathematica or Matlab format using the *format* attribute associated with them.

An *operator* is a generalization of the familiar notion of a function. Typically, an operator is used to represent the operations performed on functions to produce other functions.

Another type of element is the *relation* element. A relation might be defined by an expression that involves logical or relational operations. A relation can also be viewed as a subset of the Cartesian product of k sets. Thus, the first $k - 1$ values in a k-tuple correspond to the arguments or inputs to the relation, and the kth value corresponds to the output. The corresponding relation table can be defined using the *mlist* element.

Although an mlist element somewhat resembles a tensor element from the data schema, its intended usage is quite different. A tensor element is primarily used to transfer a multidimensional array across different systems (platforms or users) and not for manipulating the array. However, the intended use of an mlist element is to define and manipulate an arbitrary list structure (where every list element can have a different cardinality). A tensor, being a recursive XML element, facilitates an easy parsing process at the other end, whereas an mlist element has the advantage of a compact representation using either Matlab or Mathematica format.

The rationale for having two different elements (expression and relation) to represent mathematical relations is that a relation is a special type of an expression involving only relational operations. The intended use of an expression element is to represent intermediate computations or evaluations in an analysis or a design process.

4.4. Modeling Aircraft Geometry

Rapid development of computer technology over the past decade has changed the conduct of conceptual aircraft design. Aircraft analysis methods that were considered feasible only for advanced and detailed designs are now available and even practical at an early stage of the aircraft

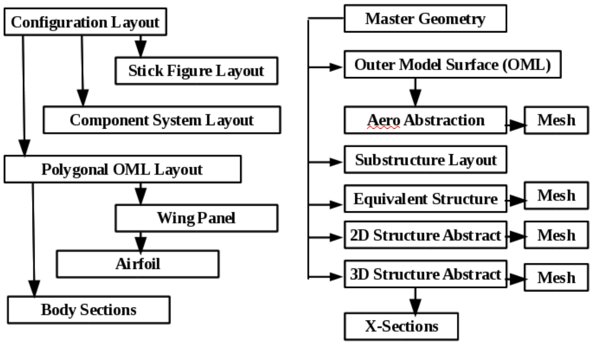


Figure 2. A high level class diagram for aircraft geometry (adapted from Ref. [2]).

design process. To fully exploit the available computing resources and analysis methods, the geometric model of aircraft must be generated rapidly and easily so as not to inhibit the conceptual aircraft design process. However, aircraft geometry is one of the most complex constructs among various conceptual aircraft design components, and likewise the representation of the geometry model is complex.

Figure 2 depicts a high level class diagram for a particular aircraft (SensorCraft) geometry. A detailed discussion of each component of the class diagram can be found in [2]. Providing a capability to represent structural and functional specifications of all the geometry components presented in Figure 2 is one of the major goals in the development of ADML geometry schema.

The ADML schema for aircraft geometry starts with low level, common geometry elements such as *point*, *line*, *plane*, *nurbs*, and *frame*, and builds on these more complex aircraft geometry elements such as *liftingSurface*, *body*, *airfoil*, *wing*, *tail*, *enginePod*, *landingGear*, *payload*, *fuelZone*, etc. The *frame* element of the geometry schema facilitates the boundary representation (discussed in detail in the next subsection) for aircraft geometry elements. Each frame element is defined in terms of an origin and three angles, and two different frame elements can be referenced through a *parentID-frameID* relationship.

4.5. Object Oriented Features

W3C XML schema, with a hierarchical type system, closely resembles an object oriented programming paradigm. Amongst the significant features of an XML schema are extensions (and restrictions), element references, and an object like behavior of an element (that carries attributes and other elements).

The modular schema development of the proposed schema, as discussed in the previous subsection, facilitates reusability and extensibility. The geometry schema follows an object oriented programming approach. Figure 3 presents a simplified view of the inheritance hierarchy for a subset (only two elements - airfoil and wingPanel) of all the components of an aircraft geometry model. Owing to the complexity of the geometry schema and a large number of XML elements to represent geometry components, it is not feasible to list and discuss each and every element in this paper.

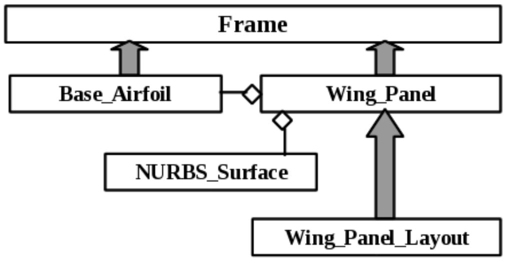


Figure 3. A simplified view of inheritance hierarchy for the geometry schema.

4.6. Next Steps

Multidisciplinary analysis and design requires a single geometric representation for a configuration that is shared amongst the various disciplines involved. For conceptual aircraft designs, boundary representation based geometry models are more suitable than CAD (computer aided design) based models.

The software behind most commonly used CAD systems is extensive and tailored to serve its community of mechanical system designers. In contrast, computational design optimization is an extension of conceptual design merged with high fidelity computational models; the geometric requirements are significantly specialized compared to general industrial CAD systems. The system described in [14] is tailored to support the generation of lofted geometry and the creation of computational meshes and computational fluid dynamaics (CFD) and computational structural mechanics (CSM) design models. It identifies the use of an XML schema for the complete aeroelastic model of a concept aircraft, including the geometric attributes. However, the schema is not publicly accessible as far as the present authors are aware.

The Air Force Research Laboratory Multidisciplinary Sciences and Technology Center (AFRL MSTC) is currently focused on adapting an existing BRep (boundary representation) geometry engine, GGTK [5], developed and maintained by the ETLab, University of Alabama at Birmingham, for use in a computational design optimization environment. The proposed ADML geometry schema must ultimately support BRep geometry data

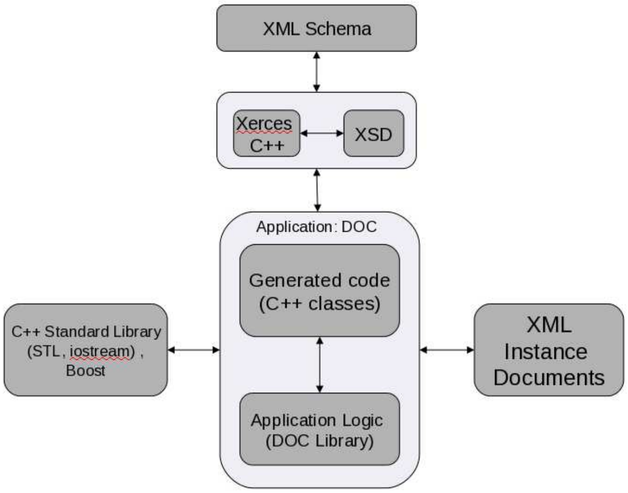

XML Schema
Xerces C++
XSD
Application: DOC
Generated code (C++ classes)
Application Logic (DOC Library)
C++ Standard Library (STL, iostream) , Boost
XML Instance Documents

Figure 4. Integration of geometry schema with DOC project.

structures. The BRep form [5] contains a set of geometry entities (e.g., NURBS) collected into various topologies, another set of data. ADML must address both sets of geometric data and the interrelationship between them.

For more complex constructs in aircraft geometry, ADML geometry schema developers intend to start with an exploration of existing standard forms as discussed in Section 2. ADML will specifically include XML schema to support BRep data with emphasis on NURBS curves and surfaces.

5. An Example Application: Airfoil Geometry

A C++ project, design optimization in C++ (DOC), that computes design sensitivities for conceptual aircraft design applications is used as a pilot project to demonstrate an application of the proposed geometry schema. An open source, cross platform W3C schema to C++ data binding compiler, Codesynthesis XSD, is used to convert the XML schema to C++ classes. Once the C++ classes are generated from the XML schema, the data stored in XML instance documents can be accessed through the C++ objects (member variables and functions) rather than dealing with the intricacies of reading and writing XML. The software architecture of the application in Figure 4 depicts the geometry integration and related tools/packages. The XSD software uses Xerces-C++ as the underlying XML parser. Xerces-C++ is a validating XML parser written in a portable subset of C++ and is available under Apache Software License.

The geometry schema presented in this paper supports a NURBS (nonuniform rational B-spline) based geometry model to represent curves and surfaces, as for the airfoil shown in Figure 5. Below is the code listing for a C++ NURBS structure and the corresponding XML schema definition. Each *nurbs* element is defined in terms of a set of control points (the *controlPoints* element), a knot vector (the *knotVector* element), a weight vector (the *weightVector* element), and the NURBS order (the *order* element).

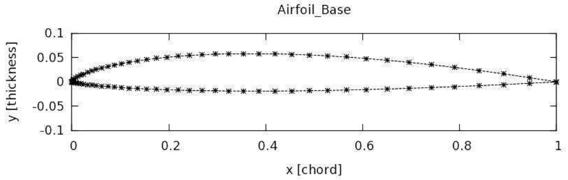


Figure 5. Airfoil geometry.

```
typedef struct{
 int numControlPoints;
 int order;
 double * knotVec;
 Point3d * controlPoints;
 double * weights;
 } NURBSCurve;
```

```
<xs:element name="nurbs">
 <xs:complexType><xs:extension base="nd">
  <xs:attribute name="nurbsID" type="xs:ID"
  maxOccurs="1"/>
  <xs:attribute name="ntype" type="nurbstype"
  maxOccurs="1"/>
  <xs:attribute name="ncp" type="xs:integer"
  maxOccurs="1">
  <xs:element name="controlPoints" type="mlist"
  maxOccurs="1"/>
  <xs:element name="knotVector" type="vtype"
  maxOccurs="1"/>
  <xs:element name="weightVector" type="vtype"
  maxOccurs="1"/>
  <xs:element name="order" type="xs:int"
  minOccurs="1"/>
 </xs:extension></xs:complexType></xs:element>
```

A sample code listing for the XML schema definition for a NURBS based airfoil object follows.

```
<xs:element name="airfoilBase">
 <xs:complexType><xs:sequence>
  <xs:element ref="frame" maxOccurs="1"/>
  <xs:element name="LEOffset" type="xs:double"/>
  <xs:element name="chord" type="xs:double"/>
  <xs:element> name="thickness" type="xs:double"/>
  <xs:element name="curveTop" type="nurb"
  maxOccurs="1"/>
  <xs:element name="curveBot" type="nurb"
  maxOccurs="1"/>
 </xs:sequence></xs:complexType>
</xs:element>
```

The geometry model for the airfoil shown in Figure 5, defined by two NURBS curves (top and bottom) with four control points associated with four weights each, follows.

```
<airfoilBase><curveTop ncp="4">
 <controlPoints format="Mathematica"><definition>
  { { 0,0,0},{ 0,0.020,0 {,
 { 0.25,0.12,0 },{ 1,0,0 } }
  </definition></controlPoints>
 <weights format="Mathematica">
  <definition>{ 1,1,1,1 }</definition>
 </weights></curveTop><curveBot ncp="4">
<controlPoints format="Mathematica">
 <definition>{ { 0,0,0 },{ 0,-0.005,0 },
 { 0.25,-0.04,0 },{ 1,0,0 } }
</definition>></controlPoints>
<weights format="Mathematica">
 <definition>{ 1,1,1,1 } </definition>>
 </weights></curveBot></airfoilBase>
```

6. Conclusion and Future Work

An XML schema is proposed for an aircraft design markup language (ADML) to represent aircraft design models (geometry) and analysis data (raw data and mathematical objects). ADML addresses data exchange and interoperability issues involved in a multidisciplinary, collaborative, conceptual aircraft design environment by providing a common language for discourse and data communication. The XML schema discussed in this paper follows a modular schema development and takes a bottom up approach by starting the schema development from the simplest form of data and building on that more complex constructs occurring in a conceptual aircraft design process. Thus, the XML elements from the data and function schema serve as the building blocks for other more complex elements. An airfoil geometry example presented in Section 5 illustrates the modeling capabilities of the proposed geometry schema. The geometry schema presented in this paper is not comprehensive, however, it provides an infrastructure with all basic geometry elements to extend the schema to other aircraft geometry components, and to expand the schema to integrate further high level geometry constructs such as topology (BRep) and grid. As discussed in previous sections, the ADML developers are motivated to explore and test EXPRESS as its schema medium in the context of geometry and geometric sensitivities, ultimately leading to a schema to support a multidisciplinary design optimization (MDO) environment.

The ultimate goal is to develop a generic, comprehensive, and compact XML schema for representing design and analysis data for all the disciplines involved in a multidisciplinary, collaboarative conceptual aircraft design and analysis process (Figure 1). Ultimately, all disciplines can natively understand the ADML standard and can communicate with each other through a common, language and platform neutral data format. The schema described in this paper is organized for the design and analysis of fixed wing aircraft, but it is readily extensible to flapping wing MAVs (micro air vehicles) and morphing vehicles, whose shapes change in time.

References

[1] American Institute of Aeronautics and Astronautics: Flight dynamics model exchange standard (draft), “BSR/AIAA S-119-201X,” AIAA, 2010

[2] M. Blair, “Air Vehicle Environment in C++: A Computational Design Environment for Conceptual Innovations,” in *Journal of Aerospace Computing, Information, and Communication*, Vol. 7, 85-117, 2010

[3] J. S. Berndt, “JSBSim, an open source platform independent flight dynamics model in C++,” JSBSim Reference Manual v1.0

[4] M. P. Bhandarkar and R. Nagi, “STEP-based feature extraction from STEP geometry for Agile Manufacturing,” in *Elsevier, Computers in Industry*, Mar 2000

[5] S. Gopalsamy and T. Yu, “A Geometry Engine for CAD/GRID Integration,” in *AIAA 2003-800, 41st Aerospace Sciences Meeting and Exhibit*, Jan 2003, Reno, Nevada

[6] M. Hucka, F. Bergmann, S. Hoops, S. M. Keating, S. Sahle, and D. J. Wilkinson, “The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 1 Core,” Dec 2009

[7] R. Haimes and J. F. Dannenhoffer, III, “Control of Boundary Representation Topology in Multidisciplinary Analysis and Design,” in *AIAA 2010-1504, 48th AIAA Aerospace Sciences Meeting*, Jan 2010, Orlando, Florida

[8] Initial Graphics Exchange Specifications, “A Century of Excellence in Measurements, Standards, and Technology—A Chronicle of Selected NBS/NIST Publications, 1901 - 2000, David L. Lide, Editor,” NIST Special Publication 958, January 2001

[9] R. Lin and A. Afjeh, “An extensible, interchangeable, and sharable database model for improving multidisciplinary aircraft design,” in *AIAA 2002-5613*, 2002, Atlanta, GA

[10] M. Pratt, “Extension of ISO 10303: The Step Standard, for the Exchange of Procedural Shape Models,” in *Proc. Int’l Conf Shape Modeling and Applications (SMI)* , 2004

[11] R. Peak, J. Lubell, V. Srinivasan, and S. C. Waterbury, “STEP, XML, and UML: Complementary Technologies,” in *ASME, Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2004,Salt Lake City, USA

[12] A. Rappoport, “An Architecture for Universal CAD Data Exchange,” in *Proceedings, Solid Modeling â^03*, Jun 2003, Seattle, WA, SCM Press

[13] D. P. Raymer, *Aircraft Design: A Conceptual Approach*, AIAA Education Series, New York, NY, 2006

[14] A. Rizzi , J. Oppelstrup, M. Zhang and M. Tomac, “Coupling Parametric Aircraft Lofting to CFD and CSM Grid Generation for Conceptual Design,” in *AIAA 2011-160, 49th AIAA Aerospace Sciences Meeting*, Jan 2011, Orlando, Florida

[15] G. L. Roth, J. W. Livingston, M. Blair, and R. Kolonay, “CREATE-AV DaVinci: Computationally based engineering for conceptual design,” in *AIAA 2010-1232*, Jan 2010, Orlando, FL

[16] R. Sandhu, *The MathML Handbook*, Charles River Media, 2003

[17] D. Schenck and P. Wilson, *Information Modeling the EXPRESS Way*, Oxford University Press,1994

New pattern for development Decision Support System(DSS)

Marjan Abdeyazdan [1], Vahid Arjmand [2], Hamid Raeis Ghanavati [3], Atefeh Parsa[4]

[1] Department of Computer Engineering, Mahshahr branch, Islamic Azad Universiy, mahshahr, Iran.
e-mail: m.abdeyazdan@mahshahriau.ac.ir
[2] Department of Computer Engineering, Mahshahr branch, Islamic Azad Universiy, mahshahr, Iran.
e-mail: vahid.arjmand@gmail.com
[3]Mahshahr, Iran.
e-mail: Hamid_raeis40@yahoo.com
[4] Department of Computer Engineering, Mahshahr branch, Islamic Azad Universiy, mahshahr, Iran.
e-mail: parsa_atefeh89@yahoo.com

Abstract - *DSS(Decision Support System) are the family of information system that developed by the aim of use models, data, information. knowledge collection to help the managers for solving unstructured problems and semi-structure problems.From breaking out till now it was introduced different methods for developing this type of system (D.S.S).However it was introduced different methods, but the plurality of methods and the weakness that related to them made the selection really difficult to choose one of them for using to development.On the other hand the process pattern subject in software developing methods, is a good solution with the aim of extracting the commons and repeatation from the family of methods of software creation to reach to public process creation.*

The process pattern is the result of abstracting sight to the action of the family of methods software creation the show the apparition ruling on them.How ever the process pattern defined in different area (yenge).Despite they didn't research about some methods to develop D.S.S till right now. in the research we will define a public life cycle for developing these type of system by extracting process pattern of DSS.The propose of process pattern collection due to pructure of public and compare with the D.S.S methodology is gathered.

Keywords: DSS(Decision Support System), development methodologies-process, patterns methodologies, engineering-meta modeling.

1 Introduction

The D.S.S is kind of systems that include tools and technique for supporting from-high-level-managing Decision (HLMD) is developed we can mention to the advantages of (DSS) system as improving easy and quick access to information-quick calculation easy uses, simple users interface for more connection with management. ability of giving complex report and keeping plenty of information. In fact DSS systems are part of daily management for making better management Decisions[1,2,5,7,10,11].

Researchers who work on the methodology of DSS believe that the process of developing these type of system has special attribution and action that make them different from the process of developing commercial information system. For example Discovering or understanding the actual requirement may discover at the process of developing system. However we attend to in introduce Different models for developing D.S.S but the explanation of methodologies is about life cycle and we don't attend to do detail for developing system . that is why they are not useless[3,4,15,16].

So methodologies of developing D.S.S systems are with special observations. all the attempt and research have been done for defining suitable methodologies in D.S.S range however it was whit up and down but it was the subject of research in this recent years.Mean while the different research has been done related to compare and analysis study and understanding the weak points and strong points these methodologies. the majority of methodologies were designed from 1978 till 1991 during some years break, during 1966 till 2005 some methodologies were defined base on pervious methodologies. however more than 30 methodologies of developing D.S.S system be given till now. nevertheless we need standard and integrate methodology[6,8,9,12]. the main problem of the all developing D.S.S systems methodologies are project-.specific and unquality for using to develop the systems in different scale[13,14].

In better way these days developing process of DSS for an international organ is different from developing for small organ this problem means that the process cause of the subject properties is specific. That is has been in the center of attention for engineering society. the process to make software, are software themselves. so every organ required and defined its specific methodology for doing its project. so the definition of a customize process adequate with available project's properties, before doing a project is necessary. These days of one solution to access to developing felexible process and consistent with project properties is using the process pattern[20,21]. The process pattern of extraction: on commons and successful repeatetive solution in developing software systems. With extant these commons we can reach to the collection of successful pattern to make a DSS system[17,18,19,22].

Our main target (aim) for doing this research:

_access to public process to developing DSS systems with pattern process set.

_access to parts of DSS methodology and plant them in parts methodology repository for doing engineering methodology.

_Enrichment avaible methodologies with process pattern.

_the possibility of comparing and evaluating methodologies base on meta model created from process patterns.

2 proposal pattern

2.1 architecture creation step pattern

In this step, organization's software and hardware sub manufacture are analized, that make system's final architecture.

Sub manufacture is like a base, that, final DSS is deployed on it. So, according to readiness of excited base, comprehensiveness is studied.

Organizational process and dominant rules on them are inspected and development team would know from their effect on DSS. Quality of hardware process capabilities are inspected and if there was any cast, organization's technical sub manufacture would be improved, in order to be accepted from DSS. Organization's meta data, which are information about business regulation and their maintenance place, are identified. Existed limitations on organization's sources are studied. Inheritance systems are identified to determine quantity of their deployment capability in new system. Based on performed analysis on existed architecture, architecture introduction and improvement plan have provided and auording to that, necessary improvements are performed.

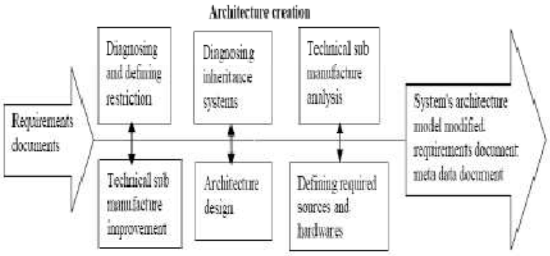

Architecture creation
Requirements documents
Diagnosing and defining restriction
Diagnosing inheritance systems
Technical sub manufacture analysis
Technical sub manufacture improvement
Architecture design
Defining required sources and hardwares
System's architecture model modified. requirements document meta data document

Figure 1: step pattern of Architecture creation

2.2 modeling step patterns

in this step required models are provided for problem. Based on req uirements and problem's difficulties, a solution and model are defined. Bank customers or works are processed in line. It is maybe that various models are provided as alternative and are deployed for similirization and analysis(figure 1). Being arbitrary of modeling step pattern means that all methodologies are typely based on dss development depended on data sources.

And they don't consider similizer systems development so we suppose it arbitrary. However, modeling step pattern is required for process pattern deployment and for similizer systems development.

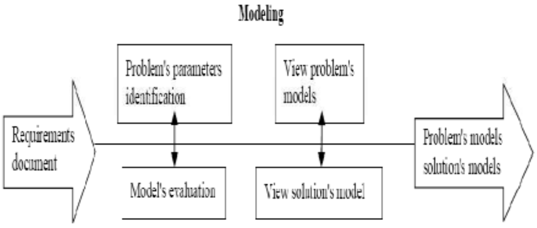


Figure 2: step pattern of Modeling

2.3 data analysis step pattern

Generally, main goal of work pattern set in this step, in identification of all organization's data sources and their quality level.Here, at first, storing parameters and other important information in related to data sources are identified, as data base and organization's data sources(work pattern of data base and data source identification)analysis of stored data quality in data source is one of the very important works in DSS development. Data quality analysis in data base and data source are very essential for production of authentic reports and appropriate statistical analysis. So, it is necessary to inspect generally data quality and quantity. Data quality means existence unreliable , faulted and incompatible data in various data sources.

Finally all organization’s data source relation are produced in a semi relation model. This model indicate data source relation model. In this step, ejected product are a report of data source relation state and stored data’s quality in data sources. (figure 2)

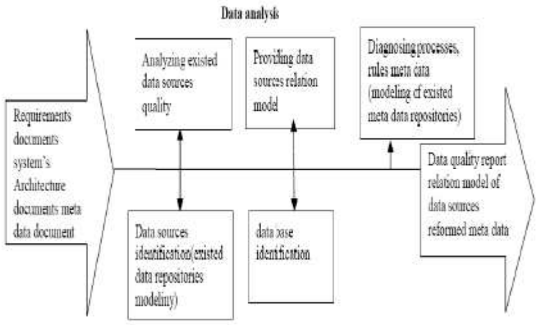

Data analysis
Requirements documents system's Architecture documents meta data document
Analyzing existed data sources quality
Providing data sources relation model
Diagnosing processes, rules meta data (modeling of existed meta data repositories)
Data sources identification(existed data repositories modeliny)
data base identification
Data quality report relation model of data sources reformed meta data

Figure 3 : step pattern of data analysis

2.4 step pattern of data architecture

The main aim, in this step, is new sources data definition depended on necessity. Here designing of data repositories or new data base are performed. There are various procedures for designing data base and data source[22,23]. After designing this data sources, contain organization's old information.which will contain clean data, during implementing work patterns of data cleaning. This step's ejected product is a set of models related to new data sources or their modifaction. (figure 3)

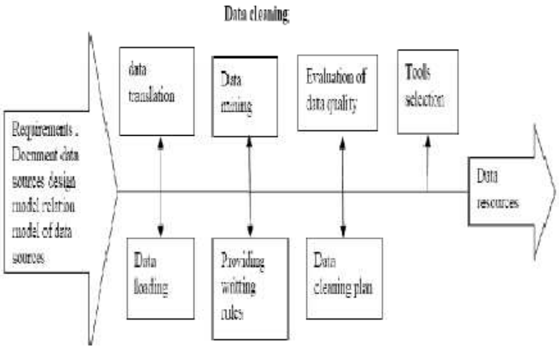


Figure 4: step pattern of data cleaning

2.5 step pattern of application development

The main aim, here is developing the required user mediator and reports, in order to work with system.

The required tools are selected for system's coding. There are popular report tools for OLAP implementation and multi dimentional analysis on data base and data resources. A set of tools selection criterias have been brought in [23,24] after tools selection , probably, some of their predeterminedLibrary following will customized in order to use in the current project. So this work pattern has been supposed arbitrary. Work entered and ejected products to this step pattern have been come in figure 4.

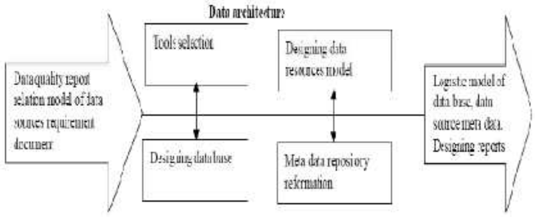

Data architecture
Dataquality report relation model of data sources requirement document
Tools selection
Designing data resources model
Desiging data base
Meta data reposiory reformation
Logistic model of data base, data source meta data. Designing reports

Figure 5 : step pattern of data Architecture

2.6 step pattern of data cleaning

One of the main step in DSS development in access to suitable quality of data , in order to take managerial decisions , based on them. During , at first existed data's quality should be evaluated in terms of authenticity level(work pattern of data quality evaluation). Typically , in this step , existed data are mined from various data resources by tools(mining work pattern). Then based on writing rules,(work pattern of providing writing rules), required information are selected, incompatibilities are removed and incorrect data are eliminated(work pattern of translation)and finally are stored in designed data resource (in step pattern of data architecture)(work pattern of loading). In data cleaning bus , suitable tools selection in essential.(work pattern of tools selection). In this direction, Solomon[25] and weir [26] have suggested the general principles for successful implementation of data cleaning process which can be used for duty patterns in this step enterened and ejected work product to this step pattern have been brought in figure 6.

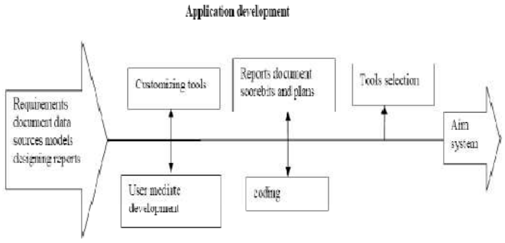

Application development
Requirements document data sources models designing reports
Customizing tools
Reports document scorebits and plans
Tools selection
User mediate development
coding
Aim system

Figure 6: step pattern of Application development

2.7 test step pattern

In this step, manufactured system is tested during passing from applications development step. Although, every product test is an important and obvious matter, but in dss, most important work pattern in this step are effectiveness and pressure tests implementation [23] and [site]. Clean data's quality is also examined. If there was any fault, it would be removed. (figure 7)

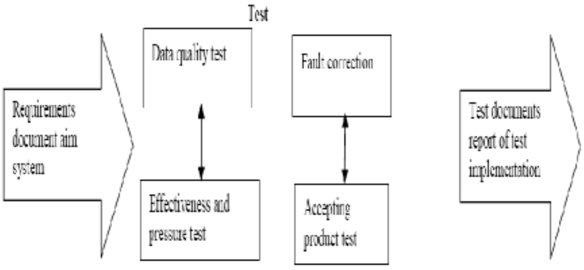

Test
Requirements document aim system
Data quality test
Effectiveness and pressure test
Fault correction
Accepting product test
Test documents report of test implementation

Figure 7: step pattern of test

2.8 Deployment step pattern

In this step, the final product which is a decision support system will be set in user environment. Users, who are often high ranking managements are though necessary trainings. most important works of this step are work patterns of obtained experiences review and version evaluation. In work pattern of obtained experiences review, development test has obtained experiences through the problems in system's development bus, and so it will not face with that problems in the next projects or in repeatation. Version evaluation is in order to diagnose probable faults and errors in system and remove them. (figure 8)

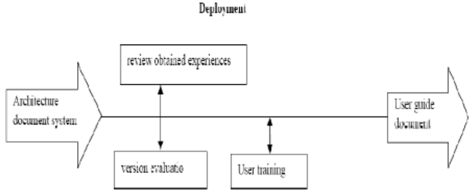


Figure 8: step pattern of deployment

3 proposal process patterns evaluation

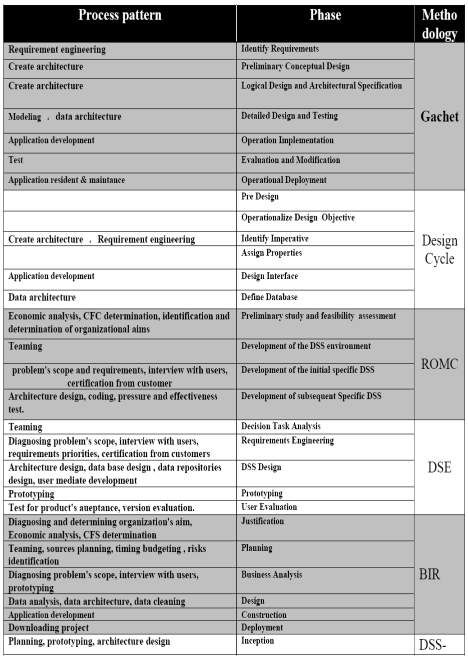
After offering public model of DSS development, it turns to evaluate it's correctness. In the present, there isn't a documented and valid way for process patterns have compared themselves by Inspecting the obtained meta model coverage level on existed methodologies and through this, they have evaluated their mined meta model correctness.

In more exact words, obtained process patterns should be able to actualize the presented methodologies and be complete. It means that through process patterns, it must be able to actualize everything is observed in DSS development methodologies. It makes assurance that proposal process patterns contains enough correctness and completeness otherwise, it should attempt to remove faults of process patterns set.

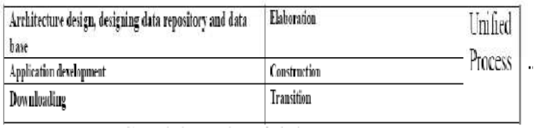
Then in this section, we consider the offered public model, which is in real a set of development process patterns, in term of coverage level and capability of creating DSS methodologies.

Success evaluation indicates that produced process patterns provide an appropriate implementation of methodology's engineering procedures. It is obvious that it can be added new process patterns in repository. Table 1 illustrate that our process pattern, that make public model of DSS development, are able to cover on DSS methodologies and actualize them.

Table1.evaluationn proposal pattern for development DSS



Process pattern	Phase	Metho dology
Requirement engineering	Identify Requirements	Gachet
Create architecture	Preliminary Conceptual Design	
Create architecture	Logical Design and Architectural Specification	
Modeling ، data architecture	Detailed Design and Testing	
Application development	Operation Implementation	
Test	Evaluation and Modification	
Application resident & maintance	Operational Deployment	
	Pre Design	Design Cycle
	Operationalize Design Objective	
Create architecture ، Requirement engineering	Identify Imperative	
	Assign Properties	
Application development	Design Interface	
Data architecture	Define Database	
Economic analysis, CFC determination, identification and determination of organizational aims	Preliminary study and feasibility assessment	ROMC
Teaming	Development of the DSS environment	
problem's scope and requirements, interview with users, certification from customer	Development of the initial specific DSS	
Architecture design, coding, pressure and effectiveness test.	Development of subsequent Specific DSS	
Teaming	Decision Task Analysis	DSE
Diagnosing problem's scope, interview with users, requirements priorities, certification from customers	Requirements Engineering	
Architecture design, data base design , data repositories design, user mediate development	DSS Design	
Prototyping	Prototyping	
Test for product's aueptance, version evaluation.	User Evaluation	
Diagnosing and determining organization's aim, Economic analysis, CFS determination	Justification	BIR
Teaming, sources planning, timing budgeting , risks identification	Planning	
Diagnosing problem's scope, interview with users, prototyping	Business Analysis	
Data analysis, data architecture, data cleaning	Design	
Application development	Construction	
Downloading project	Deployment	
Planning, prototyping, architecture design	Inception	DSS-



Architecture design, designing data repository and data base	Elaboration	Unified Process
Application development	Construction	
Downloading	Transition	

4 conclusion and future works

Today, a specific methodology definition is necessary for every project implementation. This problem is also true about DSS development. For specific methodology definition, it should be referenced to process patterns until it, would be produced an entiched collection of different methodology's parts. In this research through DSS methodology inspection, we try to mine commons and obtain a set of process patterns and organize them in a public life cycle. In providing path of process patterns we observe that DSS development methodologies are not so imatured and must of them have remained in life in cycle definition level. So, we used matured scopes of DSS development for our pattern completion. Produced patterns set can be used as a repository of method parts in methodology engineering way. We aim to develop patterns more and try to complete them in future. Obtained patterns set, are a collection of first process patterns of decision support system development.

References

[1] R.R.Veronica, Decision Support Systems Developmentavailable at:http://steconomice.uoradea.ro/anale/volume/2007/v2-statistics-and- economic-informatics/36.pdf.

[2]A.Gachet,P.Haettenschwiler,Development Processes of Intelligent Decision-making Support Systems: Review and Perspective.

[3] B.Arinze,A contingency model of DSS development methodology. Journal of Management Information Systems 8(1): 149-166,1991.

[4] DR.Arnott, A framework for understanding decision support systems evolution. 9th Australasian Conference on Information Systems, Sydney, Australia: University of New South Wales, 1998.

[5] G.M.Marakas, Decision support systems in the 21st century. Upper Saddle River, NJ, Prentice Hall, 2003.

[6] A.Gachet , R.Sprague, A context-based approach to the development of decision support systems, International workshop on Context Modeling and Decision Support, Paris, France, 2005..

[7] T.Moss.Larissa, Atre Shaku, Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications, Addison Wesley, 2003.

[8] R.R.Veronica, Decision Support Systems Development.

[9] RW.Blanning, The functions of a decision support system, Information and Management 2, Page 71-96, 1979.

[10] Martin MP, Determining information requirements for DSS, Journal of Systems Management, Page14-21, 1982.

[11] CB. Stabell, A Decision-Oriented Approach to Building DSS. Building Decision Support Systems. J. L. Bennett. Reading, MA, Addison-Wesley, Page 221-260,1983.

[12] Sage AP (1991) Decision support systems engineering. New York, Wiley.

[13] JC.Courbon, J.Drageof, J.Tomasi, L'approche évolutive, Informatique et Gestion,1979.

[14] JC.Courbon, J.Grajew , J.Tolovi, Design and Implementation of Decision Supporting Systems by an Evolutive Approach, Unpublished working paper, 1980.

[15] M.Alavi, IR.Weiss, Managing the risks Associated with end-user computing, Journal of Management Information Systems 2(3), Page 5-20, 1985.

[16] S. W. Ambler, Process Patterns: Building Large-Scale System Using Object Technology, Cambridge University Press, 1998.

[17] J. O. Coplien, A Generative Development Process Pattern Language, In Pattern Languages of Program Design, ACM Press/Addison-Wesley, 1995, pp. 187-196.

[18] J.O.Coplien, A Development Process Generative Pattern Language. In Proceedings of the First Annual Conference on Pattern Languages of Programming (PLoP), 1994.

[19] Harmsen, A. F., Situational Method Engineering, Moret Ernst & Young, 1997.

Conference (CSICC'08), Kish Island, Persian Gulf, Iran, March 2008.

[20] S.Tasharofi, R.Ramsin, "Process patterns for Agile methodologies", In Situational Method Engineering: Fundamentals and Experiences, J. Ralyté, S. Brinkkemper, B. Henderson-Sellers (Eds.), Springer, 2007, pp. 222-237.

[21] E.Kouroshfar, H.Yaghoubi Shahir, R.Ramsin, "Process patterns for component-based software development", In Proceedings of the 12th International Symposium on Component-Based Software Engineering (CBSE’09), 2009, pp. 54-68.

[22] D. F. D’Souza, A. C. Wills, Objects, Components and Framewroks with UML: The Catalysis Approach,Addison-Wesley, 1998.

[23] L.T. Moss, S.Atre, Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications, Addison Wesley, 2003.

[24] E.Turban,Decision Support Systems And Intelligent Systems, Seventh Edition,Printic-Hell, 2006.

[25] Solomon, Ensuring a successful data warehouse initiative, information systems management, Vol. 22 Iss. 1, p26, 2005.

[26] R.Weir, T.Peng, J.M. Kerridge, Best Practice for Implementing a Data Warehouse: A Review for Strategic Alignment, Proceedings of the 5th Intl. Workshop DMDW'2003, Berlin, Germany, September 8, 2003.

A Comprehensive Evaluation Model of Algal Water Bloom in Rive and Lakes

Zaiwen LIU[1],Xiaoyi WANG[1],Wei WEI[1], and Qiaomei WU[2]
[1] College of Computer and Information Engineering, Beijing Technology and Business University, Beijing,China
[2] Guangxi University, Nanning, Guangxi, China

Abstract - *In order to evaluate the phenomenon of algal water bloom in lakes efficiently and properly, an integrated evaluative function F of algal water bloom is proposed in this paper. The study of the function involves three aspects: algal average activation energy of photosynthesis（E）, integrated nutritional status index （TLI（∑））, and transparency （SD）,which are considered from the microcosmic level., the macroscopic level and the intuitionistic level respectively. The values of the function are classified properly. At the meantime, the weight value of each evaluative parameter is determined objectively, via the theory of multiple criteria decision making,. By analyzing and calculating the experimental data, the obtained values of the function and the classification results can be verified using the data of the samples. Good agreement is obtained between the results and the fact. In this way, the evaluative function of algal water bloom offers a significant theoretical basis for the algal water bloom prediction of lakes.*

Keywords: Modeling, Forecasting, Water bloom, Integrated nutritional index, MAMD, Evaluation method

1 Introduction

As the aggravation of global water eutrophication, the outbreak of algal water bloom is becoming increasingly frequent. The environmental and economic problems caused by algal water bloom are paid more and more attention by the public. Most part of the lakes in China is also facing the outbreak of algal bloom[1-2]. Algal water bloom damages the biodiversity in bodies of water, seriously hindering the economic construction and social development. In recent years, in order to control algal bloom caused by eutrophication in lakes, a large sum of money has been given by the Chinese central and local government, among which, 4 billion RMB are cost by DianChi, 10 billion RMB are cost by TaiHu, and millions upon millions of RMB are also easily cost by some small urban lakes[3]. Therefore, further study of growth principle of algal water bloom should be made, rational evaluation function of algal bloom should be found, and efficient mathematical model should also be provided to accurately predict the outbreak of water bloom so as to predict and control or remove the algal water bloom efficiently, which, in turn, will benefit the economic construction and ecological protection.

At present, many studies about eutrophication in inland lakes exist at home as well as at abroad, and all of these studies are relatively mature with great achievement. However, the occurrence of algal bloom and its evaluation system is rarely studied. Some scholars have made a research about the phenomenon of algal bloom on the Three Gorges Area and have established exploratory algal bloom outbreak evaluative function[4]and algal bloom state function[5]. However, geographical differences of water quality and algal growth must be drew proper attention. Moreover, the weight of each evaluative factor in the mathematic model mentioned above is analyzed experiences and calculated on the basis of the original data which were obtained in the Three Gorges Area. As a result, no mathematical model of algal bloom evaluation for lakes of Beijing has been reported by far.

In this paper which adopts the characteristics of the lake in Beijing, it could determine the algal average activation energy of photosynthesis（*E*）, status index of nutritional （*TLI*（∑））, and transparency（*SD*）are the parameters of evaluation function for water bloom, according to the analysis of microcosmic feature, macroscopic feature and intuitionistic feature. And the model for evaluation function of water bloom F is established utilizing the weights of those parameters determined objectively by multiple attribute decision making.. A sunlight laboratory is utilized to simulate the algal bloom incubator. The obtained experimental data is used to calculate the evaluative function value of algal water bloom and the function values are properly classified. The verification results of the samples are in line with the true fact. In this way, the evaluative function of algal water bloom offers a significant theoretical basis for the algal water bloom intelligent prediction of lakes in Beijing.

2 Integrated evaluative function of water bloom

2.1 The construction of integrated evaluative function of water bloom

Not all water with eutrophication will have water bloom occurrence, though eutrophication provide nutrition for the water bloom. As a result, in this paper, algal average activation energy of photosynthesis *E* (microcosmic level), integrated nutritional status index which serves as basic parameter(internal factors causing water bloom), and transparency(intuitionistic feature) are introduced to construct water bloom evaluative function, whose mathematical model is as follows:

$$F = \sum_{i=1}^{n} W_i T_i \qquad (1)$$

Where F is the evaluative function of the water bloom and *Wi* is the weight coefficient of each evaluative parameter. Since the unit of each evaluative factor might be different, every evaluative factor should be normalized and represented by T. The normalization formula is as follows:

$$r_{ij} = \frac{R_{ij}}{\sqrt{\sum_{j=1}^{m} R_{ij}^2}} \qquad （2）$$

2.2 Water bloom prediction model based on BP neural network

2.2.1 Lalgal average activation energy of photosynthesis *E*

Within the range of temperature suitable for the growth of algae, assimilation rate are found to increase with the increase of the temperature. While $\theta_{\min} < \theta < \theta_{\max}$ [6] $\theta_{\min}$ is the minimum value algae can withstand, and $\theta_{\max}$ is the maximum value algae can withstand ,

$$\frac{d\ln v}{dT} = \frac{E}{RT^2} \qquad (3)$$

Where v is photosynthesis rate, *T* is thermodynamic temperature and R is the gas constant. According to the literature[4], photosynthesis is defined as:

$$v = \frac{dc_A}{dt} \qquad (4)$$

Where dc_A is algal biomass concentration. The value of phytoplankton biomass can be calculated by the following formula, according to the research results found by Chinese Academy of Sciences at Wuhan East Lake[7]:

$$c_A = 7.577 + 0.328c_a \qquad (5)$$

Where C_A is the value of phytoplankton biomass(mg/L) and *Ca* is chlorophyll a concentration ($\mu g / L$)。According to the features of lakes in Beijing, the equation of chlorophyll a can be represented as follows, which is the mathematic model of algal growth mentioned in the reference[8]:

$$\frac{dc_a}{dt} = [G_{\max} \times 1.066^{(T-293)} \frac{TP}{TP+K_p} \times \frac{I}{I+K_I} - D_{\max} \times 1.08^{(T-293)} - m_p] \times c_a \qquad （6）$$

Via（2），（3），（4），（5），algal average activation energy of photosynthesis *E* can be expressed as:

$$E = \frac{R\int d\ln^{0.328\times[G_{\max}.1.066^{(T-293)}\frac{TP}{TP+K_p}.\frac{I}{I+K_I} - D_{\max}.1.08^{(T-293)} - m_p].c_a}}{\int dT^{-1}} \qquad （7）$$

Since water bloom usually breaks out during a period when water temperature is relatively stable and normally, temperature difference is a constant value. To make calculation easy, we assume *T2-T1*=1, so

$$E = RT^2 \ln^{0.328\times[G_{\max}.1.066^{(T-20)}\frac{TP}{TP+K_p}.\frac{I}{I+K_I} - D_{\max}.1.08^{(T-20)} - m_p].c_a} \qquad （8）.$$

2.2.2 Mathematic model of improved nutritional status index

The mathematic model of nutritional status index is a method which combines many limitations and factors that influence the eutrophication process in the reservoir and expresses all these factors with exponential form. The nutritional state of water is continuously divided into different levels by using this model. Based on the development of the research of this method, this method generally includes Carson Index, modified Carson Index and relative weighed nutritional status index. However, the relative weighed nutritional index is pretty complicated when it is used to calculate the relationship between two factors and its accuracy will decline because of lack of the number of samples. In this paper, an enhanced method of the weight calculation of the relative weighed nutritional status index is proposed so as to overcome these disadvantages, and the new mathematic model is called the enhanced relative weighed nutritional status index model. Weights of all the factors can be directly calculated by universal index formula and applied to nutritional status index model. Nutritional status index formula is generally expressed as follows:

$$TSI_m(i) = a_i + b_i \ln c_i \qquad (9)$$

Where C_i is measured value of each factor and a_i, b_i are undetermined coefficients changing with different factors.

By modifing the above formula with genetic algorithm, a more widely applicable universal index formula is obtained:

$$TSI_G(i)=1+10.6\ln(\frac{c_i}{c_{i0}}) \qquad （10）$$

Where C_{i0} is the base value of the factor being analyzed.The concrete situation is shown in Table1.

Table1.The base values of the eutrophication evaluative factors in lakes and reservoirs

	Chl-a	*TN*	*TP*	*COD$_{mn}$*	*SD*	*NH$_3$-N*	*BOD$_5$*
C$_{i0}$	0.1	0.01	0.001	0.1	0.02	0.002	0.05

The weight of each evaluative factor is calculated by compromise activation efficacy function. (strengthen the effect of smaller nutrient state indices, and weaken the effect of larger nutrient state indices):

$$w_i^k=\begin{cases}(\frac{u_i}{2})^{1/2}, 0\le u_{i\le 0.5}\\ 1-[\frac{1-u_i}{2}]^{1/2}, 0.5\le u_i\le 1\end{cases} \qquad （11）$$

In the actual calculations, make

$$u_i=TSI_G(i)/100 \qquad （12）$$

So as to simplify the calculations. After the value of w_i^k being worked out as well as different researched factors being normalized, it could get the weights of various factors.

2.3 Multiple attribute decision making accessing to the weight of each parameter of water bloom evaluation

Multiple Attribute Decision Making (referred to as for MADM), known as multiple objective decision making of limited scenarios as well, is the decision making via which selecting the most suitable scenario or ordering the limited scenarios, in the condition of considering a number of attributes (or indicators). In the problems of Multiple Attribute Decision Making, a great number of objective methods in terms of attribute evaluation exist, and this paper utilizes the method for ensuring attribute weights proposed in the literature [10] to get access the weight of each factor in the water bloom evaluation, the model is as follow:

$$MinF(W)=\sum_{i=1}^{n}D_i(W)=\sum_{i=1}^{n}\sum_{j=1}^{m}d^2(r_{ij},r_i')W_j^2$$

$$s.t.\quad W_j\ge 0,\ j=1,2,\cdots m \quad \sum_{j=1}^{m}W_j=1 \qquad (13)$$

In the model, r_{ij} is the value of each attribute in the matrix of standardization, r_i' is the ideal value of each attribute, $d(r_{ij},r_i')$ is the norm between the value and ideal value of each attribute, known as the proximity. Specific calculation steps are as follows:

Determining the matrix of attribute:

$$A=[a_{ij}]_{n\times m},\ i=1,2,\cdots,n,\ j=1,2,3$$

◆ The standardization for the decision-making matrix

◆ Ensuring the ideal value of each attribute

◆ Resolving the model (13) to obtain the optimal weight vector of attributes: $W_j,(j=1,2,3)$;

2.4 Calculation of water bloom evaluation function

It is normally between 25℃ - 30℃, the temperature of water, in summer of Beijing, which is the most ideal one for the growth of algae and thus it is too sensitive to cause the outbreak of water bloom in the conditions of abundant nutrition and light. The researchers in chemical lab of Beijing Technology and Business University utilize the raw water in Chang He river of Beijing as samples, putting the water in the algal incubator of Sun Lab to have the experiment to simulate the growth of water bloom. According to the observation, after 5 to 7 days' cultivation, the water bloom could break out. The steps to calculate the comprehensive evaluation of water bloom are as follows:

Calculating the average activation energy for photosynthesis of algae according to (8) formula;

Calculating the integrated nutritional status index according to (10) formula;

Utilizing Multiple Attribute Decision Making to ensure the attribute weight of parameters;

Utilizing formula (2) to standardize the calculation of each evaluation parameter

Computing the value of water comprehensive evaluation function according to formula (1)

Via research, classify the calculated value of water bloom comprehensive evaluation into different levels as follow:

$$F=\begin{cases}[0\ \ 0.25] & \text{safe level}\\(0.25\quad 0.34] & \text{mild}\\(0.34\ \ 0.45] & \text{moderate level}\\>0.45 & \text{severe level}\end{cases}\qquad(14)$$

3 Confirmation of comprehensive evaluation function for water bloom

3.1 Calculation of comprehensive evaluation function for water bloom

Calculate the data selected from No 2, 4 and 5 pools of second group, and select a value respectively in morning and afternoon everyday as the data to be calculated. The data of *TP* in each pool is 0.1 mg/L, 0.2mg/L, 0.2 mg/L. The experiment started from the afternoon of April 10, 2007, the chlorophyll value came to the first culmination in No 2, 4 and 5 pools on April 14, 15 and 16. No 5 pool in the fourth group (*TP* 0.2 mg/L) started cultivating on August 3, and come to first culmination on August 11. Parts of data and calculated values of each pool are indicated in table 2 as follow.

Table2 Calculated Data of Water Bloom Evaluate Function

Sample Point	Light(W. m2）	Chal($\mu g/L$)	Transparency	*E*(KJ)	TSIG	*F*
Second group No2 pool	30.06	1.20	5.00	23.48	30.36	0.15
	28.51	10.10	1.67	41.61	51.98	0.25
	12.17	28.30	2.50	43.46	57.28	0.26
	12.73	32.70	1.50	49.08	56.75	0.30
	8.64	52.40	0.8	45.42	61.78	0.35
	20.73	56.50	0.50	45.76	70.89	0.42
Second group No4 pool	29.29	1.10	5.00	21.68	31.16	0.12
	28.49	19.6	1.10	39.54	50.66	0.24
	29.54	38.90	0.91	48.59	56.04	0.28
	31.31	59.70	0.32	58.53	64.07	0.39
	28.81	81.60	0.17	50.44	70.22	0.49
Second group No5 pool	30.06	1.10	5.00	17.38	27.24	0.15
	28.51	7.10	3.50	43.31	35.44	0.24
	12.17	13.80	1.00	36.59	38.94	0.24
	12.73	18.70	3.33	40.96	39.90	0.25
	20.73	30.50	2.50	40.82	43.58	0.27
	24.47	46.00	0.9	43.89	50.67	0.40

3.2 Calculation result analysis of water bloom evaluation function

1) Via the analysis of experiment data to each pool, it could indicate that the quality of water was at a good state and all of the average activation energy of algae, integrated nutritional status index and the function value of water bloom evaluation water relatively low. Nevertheless, with the growth of algae each index was increasing simultaneously and the index of activation energy for algae photosynthesis was more than 40KJ at the moment approaching to the outbreak of water bloom which accorded with the conclusion of literature [9].

2) Comparing between the value of comprehensive evaluation and integrated nutritional status index: In No.2 pool of the second group, water quality, according to the comprehensive evaluation of water bloom, was at moderate level as the state of eutrophication at moderate level and severe level on April 14 and 15 2010. Meanwhile, in No.5 pool, water quality was at moderate level and the integrated nutritional status index indicated that it merely reached the state of eutrophication.

However, in fact, certain signs had appeared, such as the clear green color of water, oil-like object on the surface and slight smell, and the value of chlorophyll reached 46 mg/L, which indicated the phenomenon of algal bloom. In No5 pool of the fourth group, the water hadn't reached the state of eutrophication till April 11, however, the signs, dark-green color and slight smell of water, had appeared, and the value of chlorophyll reached 41.80mg/L, indicated the algae bloom had already occurred. Actually, the value of chlorophyll had reached 60mg/L measured in the evening of April 11, demonstrating the appearance of water bloom, which indicated that the results of the evaluation for eutrophication and comprehensive evaluation of water bloom were not entirely corresponded, and the eutrophication was the condition for the appearance of water bloom but it could be unilateral to estimate directly whether the water bloom having occurred or the extent of its level state, merely via integrated nutritional status index. However, the evaluation function of water bloom established in this paper could reflect each phrase in terms of the growth of algae much better.

Initially, the data of test Second group, as the training data of network, is trained by BP neural network function which is provided by *MATLAB* and its error is controlled in the range of 0.0001. Then, *SIM* emulational function is used for interpolation emulational output. Comparing the diagrams of prediction result with real measurement result until proper interpolation value is generated. Neural network method is used to interpolate so as to gain network training data. Interpolation graphs of some partial parameters are as follow[10].

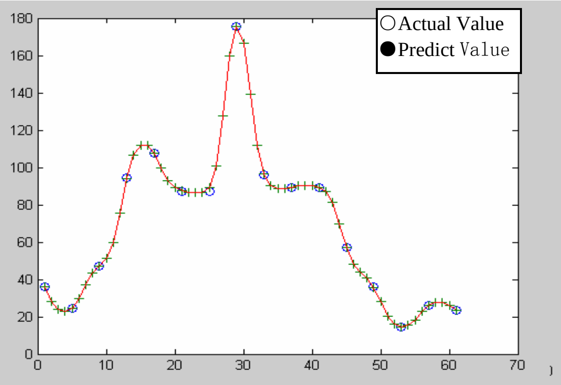


Figure 1. Prediction diagram for *Chi_a* interpolation data fitting curve

4 Conclusions

Study comprehensively the synthesis effects of the photosynthesis of algae for average activation energy, comprehensive integrated nutritional status index, and the transparency, establish the model of water bloom evaluation function, and utilize the Multiple Attribute Decision Making theory to ensure the attribute weights for all evaluated parameters impersonally. The results, concluded by the analysis and calculation of the experiment data, indicate that this evaluation model could be utilized to describe the growth of water bloom at every phrase, and its results are in line with facts, for that mater, this model could provide the references for prediction and management of water bloom.

Nevertheless, simulated in the incubator of the Sun Lab at the maximum extent to approach to the real situation, the growth of algae is affected by various kinds of uncertain factors in actual lake. Thus when applied into the water bloom evaluation with the lake of actuality, the validity of this model should be discussed and verified further. However, with the development of this research and the promotion of its methods, progress and improvement could be made for evaluation function of water bloom.

5 Acknowledgements

This supported by the Beijing Natural Science Foundation （8101003）, and the Beijing Municipal Commission of Education (PHR201007123、PHR201008238) and its Science and Technology Foundation Project. Those supports are gratefully acknowledged.

6 References

[1] Jin Xiangcan, Li zhaochun, Zheng Sufang et. Growth characteristics of microcystis aeruginosa [J]. *Environmental Science Research*, 17,2004,52-61.

[2] Jin Xiangcan,Liu Shushen,Zhang zongshe et. China Lake Environment [M]. *Beijing: China Ocean Press*,1995.

[3] Kong Fanxiang, Gao Guang, Thinking of large shallow lake eutrophication cyanobacteria bloom formation mechanism [J]. *Eco-Journal. 25*(3), 2005,589-594

[4] Liu Xinan, Zhan Min, Xie Zhaoming. Three Gorges reservoir area algal outbreaks typical watershed function model evaluation studies[J].*Environmental Science, 27*（4）,2006, 669-674

[5] Liu Xinan, Zhang Mifang. Gathered parameters for advantage algae of the Three Gorges reservoir area and algae bloom state function [J]. *Journal of Environmental Science. 28*（9）,2008， 1916-1921

[6] Wang Xian, Li Wenquan. Experimental study of the relationship between four single synechocystis assimilation rate and temperature. [J]. *Taiwan Strait, 9*(3), 1990,287~230.

[7] Yu Jingshan,Cui Guoqing. Beijing Zoo water ecology water bloom occurrence mechanism [J]. *Environmental Engineering,22*(4) ,2004,62-65

[8] Liu Zaiwen, Wu Qiaomei, Wang Xiayi, Cui Lifeng, Lian XiaoFeng. Algal growth based on the optimization theory model and the application of the water bloom prediction [J].. *Chemical Journal, 59*(7), 2008, 1869-1873

[9] Li Qiyong, Deng Xinmin, Zhao Xiaohong, Xu Nanni. Nutritional status lakes universal index formula evaluation and the effect of testing [J]. *Environmental Engineering, 20*(1), 2002,:70-72

[10] Liu Zaiwen, Wu Qiaomei, Wang Xiayi, Cui Lifeng,Intelligent technology for predicting water bloom engendering. *Proceedings - 34th Annual Conference of the IEEE Industrial Electronics Society*, IECON 2008, Orlando, FL, United states, 1896-1900

Intelligent Integration of Product Design to Production Devices

M.B. Raza, R. Harrison

mmbrm@lboro.ac.uk , R.Harrison@lboro.ac.uk

Abstract. The integration of manufacturing design knowledge with live production data in composite services improves the ability for manufacturing lines to reconfigure and recover from errors. By using manufacturing design knowledge captured as ontologies planning level faults on assembly lines and conflicts in between product design and line can be automatically identified. In addition by using embedded Web Services (WSs) production process errors and/or faults can be fed from the line into more complex and knowledge based services to aid error recovery and line reconfiguration. This approach allows more targeted alerts and reports of failures, empowering the production operative and allowing more problems to be solved at the source of origination, thus improving efficiency.

Keywords: Ontology, SOA, Semantic Web Service (SWS), knowledge based system.

I. INTRODUCTION

Service Oriented Architecture (SOA) has emerged as a reliable distributed computing method. WSs are considered the best implementation method of SOA as they are loosely coupled and platform independent. Nevertheless to construct machine accessible 'XML' in a complex product manufacturing enterprise, a higher level of semantics is required which is provided by ontology. An ontology is commonly defined as: "a formal, explicit specification of a shared conceptualization"[1]. More specifically, an ontology is an engineering artifact composed (i) of a vocabulary specific to a domain of discourse, and (ii) of a set of explicit assumptions regarding the intended meaning of the terms in the vocabulary for that domain.

The research exploits the advantages of ontologies and SWSs in real world industrial problems. The relationship between the design phase of a product and its creation on a production line is vital for manufacturing efficiency. Errors in the design or failure to create a line that suits the product design can create delays in the process as re-design and re-configuration occur. These revisions have impacts on both the supply chain and overall manufacturing output.

In the manufacturing domain, the relationship between the design phase of a product and its creation on a production line is vital for manufacturing efficiency. Errors in the design or failure to create an assembly line that suits the product design can create delays in the process as re-design or re-configuration occurs. These revisions have impacts on both the supply chain and overall manufacturing/assembly output.

The Loughborough University team in collaboration with Ford Motor Company, UK have been investigating how ontologies and SWSs could be used to improve production output through Business Driven Automation (BDA) project [2]. These investigations lead to the development of a software framework, which will facilitate the integration of product design and production line configuration. This framework builds on existing WSs developed during earlier projects such as SOCRADES [3]. Rest of the paper is organised as follows: Second section discusses the related work, third section explains device level services, the fourth sections presents the industrial case study, the fifth sections discusses the general system overview while the sixth section gives a perspective on the next steps for enhanced knowledge based systems.

II. RELATED WORK

Research work into the use of ontologies and the semantic web in manufacturing has started to be applied onto real industrial case studies. SOA and its implementation using WSs has raised significant interest as a technology facilitator for encapsulating industrial devices as loosely coupled and interoperable units. The use of Semantic Web Services (SWSs) is based upon ontologies that provide semantics and reasoning support for intelligent retrieval and discovery and use of manufacturing resources.

The use of ontologies in the manufacturing domain to form intelligent SWSs to improve productivity is emerging [4]. These ontologies are applied to a variety of points in the manufacturing lifecycle ranging from design and product production phases. These innovations are significant for companies such as the Ford which is currently facing challenges in maintaining competitive advantage when faced with competitors from less developed economies who can produce products on a large scale and lower cost. To counter this, western manufacturing has looked towards innovation in manufacturing process embracing movements such as agile manufacturing. The enablement of high quality and customized production through agile manufacturing requires changes in production process. Central to the production line is the ability for it to support change and reconfigure [5]. Time saved during this process has a direct impact on the efficiency of an organization and is a key focus of the research.

To date, re-configuration and assembly line flexibility has been managed in a variety of ways. Ford's Powertrain assembly follows this pattern with assembly lines consisting of a variety of vendor specific machinery and control software. Thus integration of machines and lines is a challenge while integration with enterprise management tools such as Enterprise Resource Planning (ERP) software is rarely achieved.

III. Device Level Services

The first phase of the work is to link to the existing data in legacy software applications, the project has developed services to automatically extract the appropriate designs, compare them against the product and produce reports for key areas of focus for the new line configuration. In addition to this services are being developed to link this data into the enterprise systems at Ford to aid scheduling of the implementation of the line and order of appropriate parts from suppliers. By encapsulating the functionality of devices as web services, composite systems can be created by adding services into Service Orientated Architectures (SOA). WSs present interfaces using open standards that can be accessed via internet protocols such as HTTP. A large amount of business systems are currently emerging to support the SOA approach in terms of enterprise integration.

SWSs are built on the SOA approach but also present interfaces to ontologies describing the relationship between the functionality of the services. This description allows SOA's to be formed in a more dynamic and automated way as service functionality can be expressed via common ontology definition and interfaced via standard interfaces. In terms of manufacturing, SWSs can foster the integration of heterogeneous production devices and of a mix of architectures in systems which would be chaotic from an ICT perspective [6].

In order to embrace SOA at production line level embedded WSs are needed to interface with simple lower level devices such as conveyors to robot arms. Presently the common approach for control at this level has been via encapsulated control languages and techniques such as PLC. However innovation in the development of embedded processing has increased the computing power that is available at line level. Thus simple devices can be controlled in real time by the use of embedded WSs.

Technical innovation in this level of control is behind the development of embedded WS toolkits such as the GSOAP and the Devices Profile for Web Services (DPWS) [7]. These WSs toolkits support standards to enable 'eventing', subscription and notification of events enabling a more efficient lower layer of device based communication.

The presence of WSs at manufacturing device level allows the use of SWSs in real time production level. This semantic management will impact the management of the line by injecting the ability to link the data from the line along with its behaviour into knowledge based services. This will enable a greater level of management and knowledge based reasoning on the line and will enhance wider production processes and automation. For example, device level service behaviour can be factored into ERP and supply chain management activities to help plan production output.

The EU Framework 6 research project SERENA [8] has provided the initial impetus for the development of device level services in manufacturing. The project has taken the application of DPWS in the home electronics field and applied it to the industrial automation field. This application is via the SODA [9] and SOCRADES [10] projects created devices that can present and be controlled by web services.
These ontologies represent how the Powertrain line is constructed, and example ontology can be seen in Figure 1.

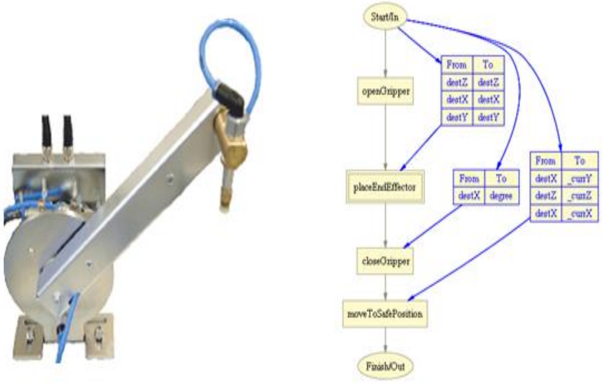


Figure1 Pick and transfer equipment, Pick service defined in ontology

IV. Case Study: Ford Powertrain Assembly Line

Downtime is a major cost at all large scale manufacturers such as Ford. Automation at Ford is delivered via the Powertrain production line. In essence the line is constantly working and is only stopped for re-configuration to support new products or in the case of mechanical or human error. In a bid to increase productivity, WSs at device level were deployed at Loughborough University on a Powertrain test rig.

These services were created using 'Arm 9 embedded processes' and the DPWS toolkit. To enhance the management of services, ontologies were introduced to enable SWSs use around the management of the line. These semantic services are linked to ontologies that describe the elements on the line. Thus increasing the computerised knowledge of the line in the live execution and monitoring process.

The machine, line, component and product data were captured into ontologies using Protégé editor and OWL language. These ontologies represent how the Powertrain assembly line is constructed and assists the mapping of data from a variety of vendors and sources that constitute a production line. This helped cross organisational usage of services hence enabling a layer of knowledge to exist over the individual elements, the knowledge is then used to support and run the production process / assembly line. With the help of ontologies full accessibility to diverse sources of data as well as automatic processing of the distributed and heterogeneous data is also a reality. For example, a breakdown on the line can be followed up by a detailed report on faulty components based on line data and distributed component design records from various vendors via the ontologies. This information reduces the recovery and repair time for the line.

V. System Overview

In terms of the wider system, the line is supported by external WSs. These services monitor the live production data from the device level web services on the line. At Ford the machines on the line are provided by a large number of vendors and are often specifically modified for different variants of products. The creation and modification of these machines creates a large level of design data which is in varying standards and formats depending on the vendor. This data is needed to analyse any mechanical faults on the line and to aid reconfiguration for the development of new products. To date this analysis has relied heavily on individual engineers and their knowledge of the machines and processes involved. Every time there is a change in the engine design, the process engineers have to manually check all the stations to determine the potential changes to be made in the assembly line. To automate this process ontologies are built to establish relational knowledge among products, processes and resources (PPR) as shown in Figure2.

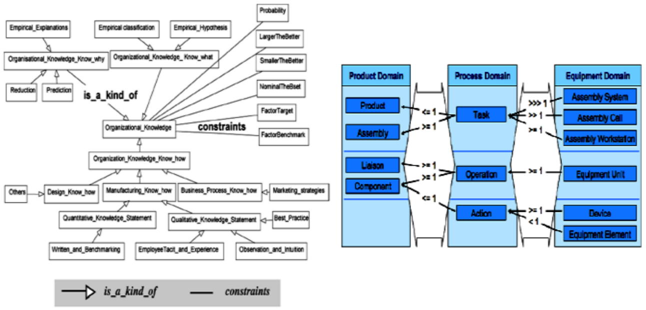


Figure 3. The UML class-based structure of organization_knowledge ontology.

Figure2: Relational information among PPR established into the ontology

By establishing relational knowledge in ontologies to describe the components in the machines, it is the aim to establish commonality between vendor machine designs at Ford. As a result, ontologies created a centralised relational knowledge base of Bill of Material (BOM) with its Bill of Process (BOP) and machine Bill of Resource (BOR), thus ontological based connections and mapping among BOM, BOP and BOR is formally established and efficiently exploited. The system then by using WSs interrogates the appropriate ontology to get the right information to aid the repair / re-configuration of lines as shown in Figure 3, where machine designers link via the use of the web to a common ontology for the line as deployed at the Ford factory.

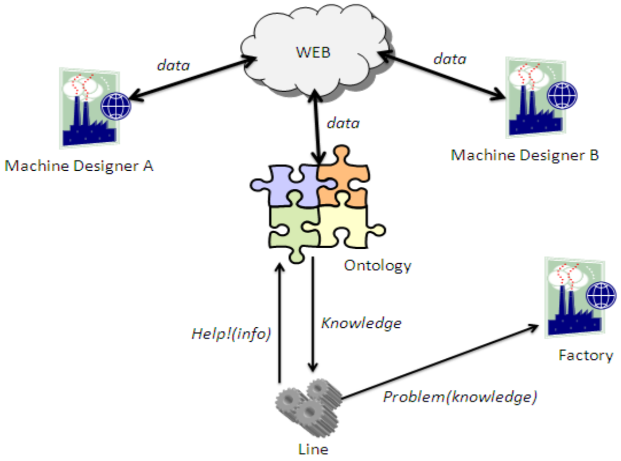


Figure 3: Knowledge based service use in line management.

Therefore using ontologies separate and distributed sources of vital line information can be both searched and linked in order to troubleshoot lines consisting of multiple vendors' machinery. In order to establish the system further it will be the aim for machine designers to provide ontologies along with the Computer Aided Design information about their machine when it is delivered to Ford. This new level of information therefore would vastly improve the amount of automated knowledge present in the Ford Powertrain manufacturing process.

VI. Configuration Management

In order to support the execution of services on the line supporting services need to be in place before, during and after device level execution. In order to make this process dynamic and transferable, knowledge is needed of service design and functionality. In the BDA project ontologies were developed to aid this activity. Machine, line, component and product data has been captured into ontologies using Protégé and OWL [11, 12]. A snapshot of created ontology for the powertrain assembly line is shown in Figure4:

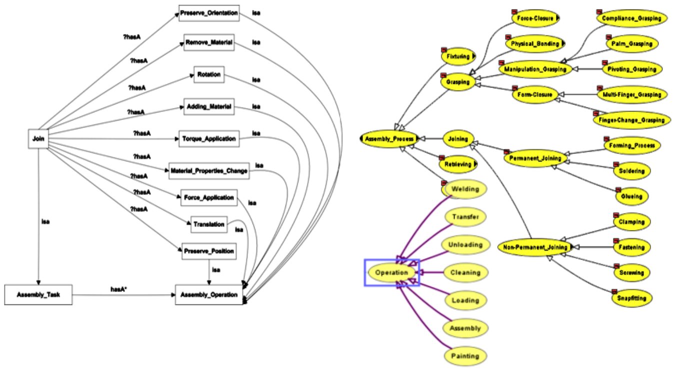


Figure 4: Ontology for a production line task and snapshot of assembly line in Protégé

Key concepts introduced into the ontology of the line are Product, Assembly Process, Resource‟ Product Characteristics, Product Parts, Process Steps etc. The development of ontologies aid the mapping of data from the variety of vendors and sources that constitute a production line. This enables a layer of knowledge to exist over the distributed data sources that are needed to help configure, support and run the line.

This layer of knowledge is wrapped and exposed as semantic web services. These services allow the knowledge bridge to the data to be represented in web service processes. Thus aiding reconfigurations of lines. For example when a new product specification is created it can be accessed using web service calls and interrogated as an OWL ontology. This will enable quick comparison of new products against existing line data also expressed as ontologies. In addition to this the ontology directs services and engineers to the appropriate sources for various elements of design data or line faults.

VII. Results

As the system develops, improvements will be added on the build and commission phase. This will be achieved by reusing knowledge from the SWS in the build and commissioning phase. This should further reduce this phase and save costs. Initial results reveal that the new approach will reduce configuration times. This will be a significant reduction that can be seen in the context of the BDA project in Figure 6.

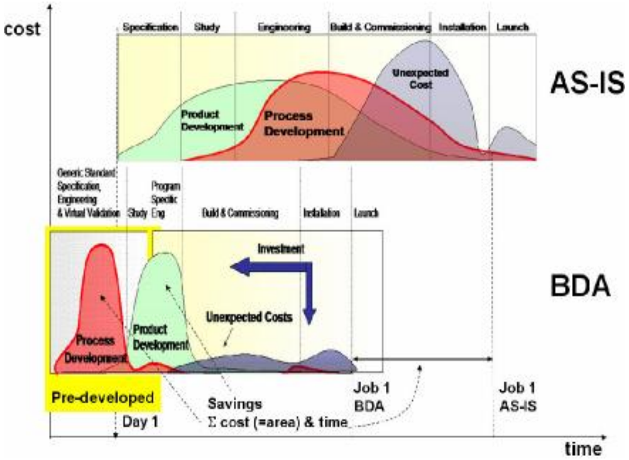


Figure 5: BDA approach compared to Ford AS-IS system

VIII. Conclusion and Future Work

In this paper we explored a methodology that incorporates and improves distributed intelligence at the shop floor level. Currently the ontologies are used to aid re-configuration of the assembly lines for new products. Live use of the ontologies has been limited to a few basic error conditions. As the development and use of the knowledge based services evolve they will be used as support in the entire production lifecycle. However, to make it a reality the data has to be captured or represented into ontologies from a wide variety of proprietary and legacy data sources throughout Ford, UK.

Using knowledge based services a new layer of manufacturing management can be envisaged that will help the entire production lifecycle. This layer is enhanced by the use of device level WSs which will enable the live use of knowledge in automated decision making on assembly lines. To date, the use of knowledge based services has been limited to the design phase of machines. By using the technology in SOCRADES and BDA projects this approach can be widened out to encompass the whole process. This will standardise production management and responses to errors that will reduce cost and improve manufacturing efficiency.

IX. Acknowledgement

The authors gratefully acknowledge the support of ‘UK EPSRC’ and Loughborough University’s ‘IMCRC’ through Business Driven Automation (BDA) project. We would like to thank all the project participants and engineers who have contributed in this research from Loughborough University and participant companies especially Ford Motor Company, UK.

X. References

[1] R. Studer, R. Benjamin, and D. Fensel, Knowledge Engineering: Principle and Methods, In Data & Knowledge Engineering, 25(1):161-197, 1998.

[2] BDA project at Loughborough University: http://www.lboro.ac.uk/departments/mm/research/manufacturing-systems/dsg/bda/index.htm

[3] T. Kirkham, D. Savio, H. Smit, R. Harrison, R. P. Monfared, and P. Phaithoonbuathong, "SOA Middleware and Automation: Services, Applications and Architectures," in 6th International Conference on Industrial Informatics (IEEE INDIN 2008): IEEE Comp. Soc., 2008.

[4] O. Lukibanov, Use of Ontologies to Support Design Activites at DaimlerChrysler, 8th Intl. Protégé Conference, 2005.

[5] Harrison, R., A.W. Colombo, A.A. West and S.M. Lee, Reconfigurable modular automation systems for automotive powertrain manufacture. International Journal of Flexible Manufacturing Systems, 2006. 18(3): p. 175-190.

[6] J. L. Martinez Lastra, "On Future Self-Orchestrating Manufacturing Systems", Position paper for the FP7 Workshop on The Agile, Wireless Manufacturing Plant, Brussels, February 10th, 2005.

[7] F. Jammes, H. Smit, Service-Oriented Paradigms in Industrial Automation, IEEE Transactions on Industrial Informatics, Vol 1(1), pp. 62-70, Feb 2005.

[8] H. Bohn, A. Bobek, and F. Golatowski, “SIRENA – Service Infrastructure for Real-time Embedded Networked Devices: A service oriented framework in different domains,” Proc of ICN,
p. 43, 2006

[9] SODA project homepage: www.soda-itea.org

[10] SOCRADES project homepage: www.socrades.eu/

[11] H. Knublauch, R. Fergerson, N. Noy, and M. Musen. The protégé OWL plugin: An open development environment for semantic web applications. In Proc. of ISWC 2004, number
3298 in LNCS, pages 229–243, 2004.

[12] D. Martin, M. Burstein, J. Hobbs, O. Lassila, D. McDemott, S. McIlraith, S. Narayanana, M. Paolucci, B. Parsia, T. Payne, E. Sirin, N. Srinivasan and K. Sycara. OWL-S: Semantic markup for web services. Member Submission 22, W3C, November 2004. http://www.w3.org/Submission/2004/SUBM-OWL-S-20041122/

Participatory Action Research: An Exploration of Electronic Banking Adoption in Saudi Arabia

Khaled AlAjmi
Maxwell School of Citizenship and Public Policy
Syracuse University
Syracuse, NY, USA
kalajmi@syr.edu

Abstract - **Banking leaders in Saudi Arabia have invested significantly on introducing and implementing new banking technologies, hoping that such technologies will lead to increase in return on investments [1]. The purpose of this qualitative participatory action research is to understand further the contextual factors needed for electronic banking leaders in Saudi Arabia to improve the adoption of electronic banking products. The focus of the study is developing and examining a model for electronic banking business using the service-oriented architecture framework. The implementation of service-oriented architecture in electronic banking enhances the customer experience and expectations in countries other than Saudi Arabia [2]. This participatory action research study indicates that banking leaders in Saudi Arabia lack a comprehensive electronic banking services representation that treats customer interaction services. The model entails organizational transformations to promote the broader public adoption of electronic banking services in Saudi Arabia [3].**

Keywords-component; service-oriented architecture; electronic banking; technology adoption model; Saudi Arabia

Introduction

The effect of information technology (IT) on business has been the subject of many studies in developed countries such as the United States [4]. Researchers have also analyzed managerial, infrastructural, organizational, and economic factors of diffusing IT with commerce in developed countries [5], but have not studied the same factors in developing countries such as Middle Eastern countries, especially in the Kingdom of Saudi Arabia. Only a limited number of researchers have explored how IT and business have diffused and evolved together to form new means of conducting business [6,7].

An example of diffusing IT with business is electronic banking. Electronic banking is a transformation of traditional financial processes [8]. The use of electronic banking has led to minimizing the traditional barriers of time and space, such as face-to-face interactions between customers and bank officers and clerks. Electronic banking channels have been gradually introduced as substitutes to traditional banking [9]. Increasingly, leaders of financial institutions are implementing electronic banking as a method of reducing operational costs and improving efficiency [10].

Although the Arabian Gulf region has undergone many technological and financial developments during the last year [11], researchers have not adequately addressed barriers limiting the large-scale use of electronic banking in the literature. Researchers have examined the public adoption of electronic banking products and services in countries neighboring Saudi Arabia, such as Oman and the United Arab Emirates [6,12,13,14,15]. The literature contains little evidence of a previous study or analysis on the adoption of electronic banking products by the public within the Saudi Arabia [16,17,18].

Problem Statement

From 2002 to 2007, Saudi gross domestic product income grew 85% from US$190 billion to US$345 billion, leading to benefits of growth in educational opportunities and an expanded technological infrastructure [11]. Saudi government leaders adopted a plan of deploying advanced technologies in the country, including creating a dedicated ministry for information and telecommunication technology [19]. The general problem is that the Saudi population has been reluctant to adopt technology despite increasing governmental expenditure on technology promotion [19]. In 2007, only 2.5% of the Saudi population subscribe to broadband services. Countries with similar characteristics to Saudi Arabia had higher rates of broadband subscriptions (Information and Communication Technology Commission, 2007). The developed country of South Africa had a rate of 64% and Estonia, which is a developing country, had a rate of 17% [20,21].

The expanding technological infrastructure in Saudi Arabia contributed to transforming the delivery channels banks use to provide better service to their customers [17]. From 2002 to 2007, banking leaders in Saudi Arabia invested in introducing Internet banking [1], but only 12.6% of banking clients used Internet banking in Saudi Arabia [11]. In 2006, the percentage of Internet banking customers of total banking clients was 20% in the United Arab Emirates, 40% in Singapore, and 60% in South Korea [22].

A number of researchers indicated how banking leaders in the Arabian Gulf region lacked an understanding that could lead to expanding the adoption of electronic banking by the public [6,17,23]. The specific problem is that banking leaders in Saudi Arabia, who represented the general population in the current study, lack an effective technology adoption methodology to strategically identify and classify the managerial, cultural, and environmental factors responsible for the lack of wide public acceptance of electronic banking products and services. Performing the qualitative action research study led to a greater understanding of the contextual factors that could improve the adoption of electronic banking products and services in Saudi Arabia. Understanding the contextual factors could enable electronic banking leaders in Saudi Arabia, who represented the specific population in the study, to identify, explore, plan, and implement the necessary organizational and managerial actions to promote a broader use of electronic banking products and services in Saudi Arabia.

Purpose Statement

The purpose of the current qualitative action research was to understand further the contextual factors needed for a service-oriented architecture model for electronic banking leaders in Saudi Arabia to improve the adoption of electronic banking products. A greater understanding of how electronic banking leaders can evolve and implement the service-oriented architecture model in Saudi Arabia might enable the participating electronic banking leaders to identify, explore, plan, and implement the necessary strategies to promote a broader use of electronic banking products and services in Saudi Arabia. Understanding the contextually relevant factors of service-oriented architecture implementation could also support electronic banking leaders playing an active role in the economic growth of Saudi Arabia by more effectively promoting electronic banking products and services [1].

The qualitative research methodology was appropriate for understanding the perceived barriers of the public toward the adoption of electronic banking practices and allowed for an exploration of the changes necessary to transform the perceived barriers. The action research design was appropriate because it helps researchers to develop iteratively the knowledge necessary to investigate and examine service-oriented architecture organizational change. The action research design was also appropriate for use in the research because an anticipated outcome of the study was to provide electronic banking leaders in Saudi Arabia with a better understanding and a better method of implementation to promote the broader adoption of electronic banking.

Research Questions

The fundamental question for the research was as follows: *What contextual and organizational knowledge might assist electronic banking leaders in Saudi Arabia to make more informed decisions concerning the design and implementation of electronic banking processes?* The strategic research outcome was to support an increased public acceptance of electronic banking products and services offered at banks in Saudi Arabia. The fundamental research question included the potential for implementing and adopting service-oriented architecture in organizations in Saudi Arabia following the successful implementation in several other countries [2,15,24,25,26]. The development of the service-oriented architecture framework occurred to increase the efficiency of organizations through the enhancement of organizational agility and to reduce the overhead costs of operations [27].

The research study involved an attempt to answer the fundamental research question through the following additional research questions:

1. *What are the results of observing how current workplace practices and processes hinder a wider public acceptance of electronic banking products and services in Saudi Arabia?*
2. *What actions, including the acquisition of local contextual knowledge, could bank leaders in Saudi Arabia take to expand public adoption of electronic banking products and services offered at banks located in Saudi Arabia?*
3. *What do banking leaders in Saudi Arabia learn when engaging with each other to derive a service-oriented architecture model to expand the public adoption of electronic banking products and services offered at banks located in Saudi Arabia?*

Engaging with the research questions required an understanding of the service-oriented architecture framework and terminology, qualitative action research, and modeling of business and service processes. Conducting the research provided additional contextual and organizational knowledge that might assist banking leaders in Saudi Arabia in making more informed decisions concerning the design and implementation of electronic banking processes. Attempting to answer the research questions also provided the electronic banking leaders in Saudi Arabia with a plan of action regarding a contextually relevant service-oriented-architecture-based model for enhancing the adoption of electronic banking products and services by the public. The results of the qualitative action research could encourage banking leadership in Saudi Arabia and along the Arabian Gulf to focus on the service-oriented architecture framework as a methodology for enhancing organizational agility and reducing operational expenditures.

Literature Review

Scholarly literature in which researchers describe how service-oriented architecture enhances the efficiency of electronic banking in Saudi Arabia is rare. The focus of a majority of the research material retrieved for this study was on the financial aspects of electronic banking and the demographic, cultural, and environmental factors limiting public adoption of electronic banking services both globally and in countries located along the Arabian Gulf [5,6,12,16,17,23,28,29,30]. A description of the public adoption of electronic banking services has been the subject of a number of master’s theses and doctoral dissertations [31,32,33]. Researchers of these studies provided the fundamental analysis for the demographic, cultural, and environmental barriers limiting a broader public adoption of electronic banking many countries like in the United States, Malaysia, and Iran.

Research with a focus on how financial organizations could use the above research publications as part of an overall organizational strategy is not common [1,2,15,34]. Leaders of a large number of organizations, including financial institutions in North America, Europe, and countries in the Asia-Pacific region, have introduced and globally accepted the emerging framework of service-oriented architecture [23,35,36]. Researchers have scarcely studied the introduction of service-oriented architecture as a framework promoting the transformation of banking and financial organizations in Arabian Gulf countries [15].

The literature did not include a thorough analysis of customers’ attitude toward electronic banking services in Saudi Arabia or how to identify and deploy the core capabilities of banks and financial institutions in Saudi Arabia. In addition to the exclusion of females in [16], which was the most relevant study to the proposed study, the results did not establish a meaningful mechanism for bank leaders to follow to gain more market share. The researchers in [17] did not elaborate on the organizational behavior aspects of implementing electronic banking in the country and did not mention the effect of reconfiguring the organizational structures to exploit the implementation of electronic banking. Researchers in [16] and [17] did not discuss how external environments affect supply chains in banks whose leaders implemented electronic banking in Saudi Arabia in terms of partnerships with other enterprises, outsourcing, and alliances.

Research Method

This study involved the use of qualitative research with an action research design. It was noted that performing qualitative action research can help provide business leaders with a greater understanding of how to enhance effectiveness in their workplaces [37]. Researchers use qualitative research to understand how social and sociocultural aspects are experienced and understood in a particular context both spatially and temporally [38]. Employing qualitative research also contributes to the analysis of sociotechnical and ergonomic aspects of adopting technology to understand how organizations, environments, and cultures react to introducing new technologies [39].

Action research includes three main stages [37,40]. The first stage involves identifying a process of societal or organizational contexts. The second stage involves analyzing underlying practical aspects of the process, including, in the present study, the current state of electronic banking practices and processes. This study involved using the first two stages to provide the initial knowledge needed to develop the service-oriented-architecture-based model recursively [2,41]. To examine the adequacy of the model developed, the researcher and the selected electronic banking leaders in Saudi Arabia iteratively studied and analyzed the service-oriented-architecture-based model to explore its suitability against the acceptance criteria established by the research participants. Interviewing electronic banking leaders have yielded information about the internal processes of a bank. The third stage of the action research involved defining the type of action needed for research participants to implement the developed service-oriented architecture model in their respective organizations.

A review of the literature indicated that academic research on applying action research to design banking processes is rare, although a number of researchers addressed the use of participatory action research to design banking services [42,43,44]. The focus of action research is the collaborative efforts of members within the same organization or from different organizations who share similar interests in addressing a specific problem [44]. This focus, coupled with the recursive feature of action research, adds a substantial amount of effort that bank leaders would not invest in unless the return on investment is evident. Researchers in [45] conducted an action research study in which 80 major banks in Finland collaborated on the implementation of deregulation changes launched in the early 1980s. Action research was an appropriate design because the design included the ability to generate plans of implementing changes required for the deregulation of banking processes.

Little research that involved developing and testing service-oriented-architecture-based models using action research exists in the literature [46]. Service-oriented architecture is a framework developed to support organizational transformations. The spiral and recursive nature of look, think, and act associated with action research requires time and effort to experiment with and reflect on the developed model [37]. A collaborative effort becomes necessary for action research to yield a tangible outcome. The researchers in [47] described the successful use of action research to develop service-oriented architecture models for a large public hospital in Ireland. The participation of the hospital's IT manager in the research represented the collaborative effort required for the success of action research.

Employing qualitative action research provides the management of local and international banks located in Saudi Arabia with the knowledge necessary to make informed organizational decisions. The outcomes of the decisions lead to the broad public adoption of electronic banking services in the country. Bank leaders in Europe, Asia, and the United States have started to strategically adopt service-oriented architecture and redesign their major businesses accordingly, resulting in greater customer satisfaction and better workplace efficiency [3]. Electronic banking leaders who provide service-oriented architecture solutions comprised the specific population of the current research study.

Selected electronic banking leaders wanted to participate in the research because they have experience in designing and managing electronic banking services in Saudi Arabia. Electronic banking leaders who provide service-oriented architecture solutions received invitations to participate in the study due to their involvement in promoting and implementing electronic banking solutions. The list of the target population is available to the public by contacting the organizations with which prospective participants have affiliations [11]. The participant sampling process was purposeful. The basis for participant selection was the maximal variation sampling strategy because business processes, structures, and operational complexities are different between banking institutions.

The research instrument was a semistructured interview process. The design included components of electronic banking and ways to implement service-oriented-architecture-based models. No expectation existed that the study would involve other instruments identified in the literature, such as surveys, because of the need to have a fully interactive and collaborative study. Interviews are a powerful research instrument that permits interviewees to explore their experiences in-depth and share their expertise to address a research problem [48].

Using semistructured interviews was appropriate for learning and understanding the terminology used in the electronic banking industry in Saudi Arabia [49]. The interviews proved to be a valuable source of information on the implementation of service-oriented architecture in other industries and on the possible implementation of service-oriented architecture in electronic banking [2,41]. The interviews included a number of levels of open-ended questions for the participants to answer. Research participants rely on open-ended questions to generate responses that are contextual and cultural in nature [48].

The first set of questions gave participants the opportunity to describe their experience and qualifications. The focus of the second set of questions was to describe current practices, operations, services, and products associated with the organization with which the participant has an affiliation. The second set of questions was the most detailed part of the interview, as it provided the foundation for the service-oriented-architecture-based model. The third set of questions involved identifying the potential limitations of current practices that hinder customer satisfaction related to electronic banking services. The decision to implement service-oriented architecture as a framework to harness the limitations was the subject of the fourth set of questions. In the final set of questions, the participants had an opportunity to extend the collaborative effort by suggesting potential participants who might have an interest in the implementation of service-oriented architecture.

The interviews of the research participants were iterative. The first round of interviews revealed a set of data to develop the initial service-oriented-architecture-based models used in [2,41]. After developing the initial model, a second iteration of interviews and discussions was necessary to examine the adequacy of the developed model, analyze its outputs, and indicate which aspects of the model needed modification. The iterations continued until the development of a satisfactory model occurred. The developed model was generic, and parameterization tasks were necessary to make the model suitable for the specific banks. The interviews also revealed a notation list of parameters and configured values necessary for the adjustment of the model behavior. The basis of the initial notation list was the analysis presented in [1,6,12,16,17,23] as well as factors affecting the adoption of electronic banking from the customers' perspective. Recursive modifications of the list occurred throughout the interviews.

The developed model was generic, and parameterization tasks were necessary to make the model suitable for the specific banking context. The interviews also revealed a notation list of parameters and configured values necessary for adjusting the model behavior. The basis of the initial notation list was the analysis presented by [12,16,17] and on factors affecting the adoption of electronic banking from the customers' perspective. Recursive modifications of the list occurred throughout the participatory action research iterations. In an action research study, the participants establish and agree upon the method of data collection, the method of analysis, and the criteria for terminating the research. The research facilitator and the members of the panel in a given action research study define the research parameters at the time of the initial interview [37].

In participatory action research, the generation of knowledge is tightly coupled with actions and leadership [50]. Participants are interested in understanding the problem as much as they want to develop a plan of action collectively that results in improving the situation that caused the problem [51]. The degree of the consequent changes and the composition of the research panel serve to measure the success of participatory action research. Panel participants iteratively work together to collect facts about the research problem and to analyze the facts to provide a better way of reacting with them [52]. The reactions in the current study centered on removing the obstacles that hinder the growth of using electronic banking in the Saudi Arabia. During the course of conducting the participatory action research, an increase in understanding on the research problem resulted in the generation of deeper knowledge and a focused list of actions proposed by the researchers [53].

The first round of interviews revealed the information needed to develop a Saudi electronic banking business services landscape, shown in Figure 1. After developing the landscape, interviews and discussions ensued to examine the adequacy and completeness of the developed landscape, analyze its components, and indicate which aspects of the model needed modification. The developed model represented the status quo of the electronic banking services at the time of conducting the participatory action research.

The interviews also revealed a notation list of the parameters and configured values necessary for the adjustment of the model behavior. Elaborated discussions took place to identify how customers' perceptions and use of electronic banking could be part of the developed model. The third discussion revealed a merging of technology adoption models and a service-oriented architecture representation of the electronic banking process. The third iteration also included an example that illustrated the design actions banking leaders need to carry out when designing electronic brokerage products in Saudi Arabia, as depicted in Figure 2.

The research panel performed a fourth and a final research iteration. In this iteration, the panel devised a questionnaire and distributed it to a number of banking leaders. The purpose of using this questionnaire was to seek the opinion of the participating banking leaders on the research outcomes developed during the first three iterations.

The research participants thought using the technology adoption model as an input to the different business services would result in maximizing the chances of adoption. Marketing and banking relationship managers have to get the customer to log in and use the service. The role of marketing is to capture the customers' profiles in the database, and the process of getting customers to log in might include incentives for the first log-in or for the first couple of log-ins. The incentives could include providing customers a gift or a voucher if the customer conducts a transaction. The use of incentives will not actually indicate adoption, but might help to get the customers to try something new. The developed model could determine whether the customer will stay with the channel or leave it and return to the branch.

In the long term, the banking leaders need to keep the customers logging in without incentives. The panel members thought that this could be achieved by tailoring customers' experience with the online channel to maximize the customers' chances of adopting the online channel. Such an experience could occur in two ways: the use of adaptable interfaces based on the profile and the use of the technology adoption model within the process of the customer interaction. The analysis and feedback from the channel itself is to be used to recall behaviors such as when the customer dropped, when the customer stopped, how far the customer went, and what kind of actions the customer likes. Such analysis indicates one of two actions could happen next. The interaction with the customer can be passive, in which customer's behavior is not observed, or active, which involves collecting and improving behavioral data over time.

When banking leaders have access to their customers' profiles and understand more about customers' behavior, the banking leaders can then change the customer interactions, displays, campaigns, and offers over time to predict customers' behavior better. Effective ongoing quantitative research activities to measure bank customers' adoption of electronic banking products are a prerequisite to the developed technology-adoption-model-based service-oriented architecture model. The activities would include online customer surveys, continuous data analysis and mining tasks, and implementing business intelligence tools, which are standard functions in many customer relationship management applications.

Research Summary

The implementation of service-oriented architecture in electronic banking has enhanced the customer experience and expectations in countries other than Saudi Arabia [2]. The purpose of the current qualitative action research was to understand further the contextual factors needed for a service-oriented architecture model for electronic banking leaders in Saudi Arabia to improve the adoption of electronic banking products. The focus of the study was developing and examining a model for electronic banking business using the service-oriented architecture framework. Participants were experienced electronic banking leaders who have implemented electronic banking services in Saudi Arabia.

Presenting electronic banking leaders in Saudi Arabia with a service-oriented-architecture-based model to perform necessary organizational transformations to improve customers' experience and expectations required employing action research [2,15]. Developing the action research required in the current study depended on building a service-oriented architecture model iteratively for the electronic banking industry in Saudi Arabia. Once developed, the intention for the study was to examine the model against the established business acceptance criteria to ensure the acceptance of the model. The model might entail organizational transformations to promote the broader public adoption of electronic banking services in Saudi Arabia [3].

The research data revealed in the current participatory action research study indicated that banking leaders in Saudi Arabia lack a comprehensive electronic banking services landscape composed of customer interaction services. Banking leaders have invested significantly in introducing and implementing new banking technologies, hoping that such technologies will result in increased market share. The research participants observed, thought about, and acted on generating the knowledge needed by Saudi banking leaders to increase customers' adoption of the introduced banking technologies.

The members of the research panel uncovered an important theme in the current study by realizing that banking leaders introduced electronic banking services without considering how the customers would react to and interact with the services [1,17,54,55]. The services were presented in a uniform manner to all customers, without taking into consideration the culture and context of designing and marketing such services. The research panel members found a number of research activities that indicated customers do not adopt electronic banking services, but negligible effort was made in the literature to recommend organizational and leadership actions to increase customer adoption. The members of the panel developed a service-oriented-architecture-based model for electronic banking services that includes the technology adoption model components as services. Including the technology adoption model when designing and marketing electronic banking services could result in a better understanding of the customers' needs, preferences, and behaviors.

The outcome of the current participatory action research could be extended, through additional participatory action research iterations, to other business units in the banking industry, such as retail banking and investment services. The research might also extend to exploring the implementation of service-oriented architecture and to broaden the public use of electronic banking in other countries. The design of the qualitative research method was appropriate for electronic banking leaders transforming their businesses to yield business and operational agility and efficiency [27,37]. The study included discussions on the implications of the research on both the banking leaders and the researchers in the area of electronic banking design. The study also contained a number of recommendations for future extensions of the research to cover other banking processes and other industries that lack customer interaction with electronic services.

References

[1] Sohail, M. S., & Shaikh, N. M. (2008). Internet banking and quality of service. Online Information Review, 32, 58-72.

[2] Shan, T. C., & Hua, W. W. (2006). Service-oriented solution framework for Internet banking. International Journal of Web Services Research, 3, 29-48.

[3] Bielski, L. (2005). Breakout systems and applications give bankers new options. ABA Banking Journal, 97(6), 61-65.

[4] Uzoka, F., Shemi, A., & Seleka, G. (2007). Behaivoral influences on e-commerce adoption in a developing country context. Electronic Journal of Information Systems in Developing Countries, 31(4), 1-15.

[5] Tarafdar, M., & Vaidya, S. D. (2006). Challenges in the adoption of e-commerce technologies in India: The role of organizational factors. International Journal of Information Management, 26, 428-441.

[6] Khalfan, A., & Alshawaf, A. (2004). Adoption and implementation problems of e-banking: A study of the managerial perspective of the banking industry in Oman. Journal of Global Information Technology Management, 7, 47-64.

[7] Zwass, V. (2003). Electronic commerce and organizational innovation: Aspects and opportunities. International Journal of Electronic Commerce, 7(3), 7-37.

[8] Harkness, M. D. (2005). Electronic banking and information assurance. Internal Auditing, 20(2), 4-20.

[9] Sarkar, P., Yadav, S. S., & Banwet, D. K. (2001). Emergence of flexible distribution channels for financial products: Electronic banking as competitive strategy for banks in India. Global Journal of Flexible Systems Management, 2(3), 29-38.

[10] Lin, J.-C., Hu, J.-L., & Sung, K.-L. (2005). The effect of electronic banking on the cost efficiency of commercial banks: An empirical study. International Journal of Management, 22, 605-611.

[11] Saudi Arabian Monetary Agency. (2007). The 43rd annual report. Riyadh, Saudi Arabia: Author.

[12] Al-Hajri, S., & Tatnall, A. (2008). Adoption of Internet technology by the banking industry in Oman: A study informed by the Australian experience. Journal of Electronic Commerce in Organizations, 6(3), 20-36.

[13] Budd, B., & Budd, D. (2007). A preliminary empirical investigation of 'brick-to-click' banking presence in the United Arab Emirates. Journal of Internet Banking and Commerce, 12(2), 1-7.

[14] Hashmi, M. A. (2007). An analysis of the United Arab Emirates banking sector. International Business & Economics Research Journal, 6, 77-88.

[15] Rabhi, F. A., Yu, H., Dabous, F. T., & Wu, S. Y. (2007). A service-oriented architecture for financial business processes. Information Systems and eBusiness Management, 5, 185-200.

[16] Al-Ashban, A. A. (2001). Customer adoption of tele-banking technology: The case of Saudi Arabia. The International Journal of Bank Marketing, 19(4/5), 191-200.

[17] Al-Somali, S. A., Gholami, R., & Clegg, B. (2009). An investigation into the acceptance of online banking in Saudi Arabia. Technovation, 29, 130-141.

[18] Dwivedi, Y. K., & Weerakkody, V. (2007). Examining the factors affecting the adoption of broadband in the Kingdom of Saudi Arabia. Electronic Government, 4, 43-58.

[19] Al-Gahtani, S. S. (2006, March 26-29). Information technology adoption, the roadmap to sustainable development: Examining three models. Paper presented at the 18th Saudi National Computer Conference 2006, Riyadh, Saudi Arabia.

[20] Dwivedi, Y. K., Papazafeiropoulou, A., & Choudrie, J. (2008). Handbook of research on global diffusion of broadband data transmission. Hershey, PA: Information Science Reference.

[21] Information and Communication Technology Commission. (2007). Annual report. Riyadh, Saudi Arabia: Author.

[22] Awamleh, R., & Fernandes, C. (2006). Diffusion of Internet banking amongst educated consumers in a high income non-OECD country. Journal of Internet Banking and Commerce, 11(3), 2.

[23] Kamel, S., & Hassan, A. (2003). Assessing the introduction of electronic banking in Egypt using the technology acceptance model. Annals of Cases on Information Technology, 5, 1-25.

[24] Lewis, G. A., & Smith, D. B. (2008). Proceedings of the International Workshop on the Foundations of Service-Oriented Architecture (FSOA 2007). Pittsburgh, PA: Software Engineering Institute.

[25] One-third of CIOs prioritize service architecture plans. (2006). Bank Technology News, 19(10), 1.

[26] Veit, D. J., & Weinhardt, C. (2007). Enterprise, applications and services in the finance industry. Information Systems & e-Business Management, 5, 139-141.

[27] Erl, T. (2008). SOA: Principles of service design. Upper Saddle River, NJ: Prentice Hall.

[28] Al-Gahtani, S. S., Hubona, G. S., & Wang, J. (2007). Information technology (IT) in Saudi Arabia: Culture and the acceptance and use of IT. Information & Management, 44, 681.

[29] Gan, C., Clemes, M., Limsombunchai, V., & Weng, A. (2006). A logit analysis of electronic banking in New Zealand. The International Journal of Bank Marketing, 24(6), 360-383.

[30] Kolodinsky, J. M., Hogarth, J. M., & Hilgert, M. A. (2004). The adoption of electronic banking technologies by US consumers. The International Journal of Bank Marketing, 22(4/5), 238-259.

[31] Abou-Robieh, M. (2005). A study of e-banking security perceptions and customer satisfaction issues. Dissertations Abstracts International, 66, 238.

[32] Adham, K. A. (2000). The adoption and implementation of information technology in Malaysian commercial banks: Phone banking and electronic terminal banking systems. Dissertations Abstracts International, 61, 4073.

[33] Alagheband, P. (2006). Adoption of electronic banking services by Iranian customers (Unpublished master's thesis). Lulea University of Technology, Lulea, Sweden.

[34] Heinonen, K. (2007). Conceptualising online banking service value. Journal of Financial Services Marketing, 12, 39-52.

[35] Glaser, J. P., Halvorson, G. C., Ford, M., Heffner, R., & Kastor, J. A. (2007). Too far ahead of the IT curve? Harvard Business Review, 7/8, 29-39.

[36] Granebring, A., & Lindh, C. (2008). Measuring the effects of SOA on business: Initial results of the study. ICFAI Journal of Management Research, 7(3), 7-24.

[37] Stringer, E. T. (2007). Action research (3rd ed.). Thousand Oaks, CA: Sage.

[38] Bloomberg, L. D., & Volpe, M. (2008). Completing your qualitative dissertation: A roadmap from beginning to end. Thousand Oaks, CA: Sage.

[39] Ahasan, R., & Imbeau, D. (2003). Socio-technical and ergonomic aspects of industrial technologies. Work Study, 52(2/3), 68-75.

[40] Park, P. (1999). People, knowledge, and change in participatory research. Management Learning, 30, 141-157.

[41] Kohlmann, F., & Alt, R. (2007). Business-driven service modeling: A methodological approach from the finance industry. Paper presented at the Business Process and Services Computing: 1st International Working Conference on Business Process and Services Computing, Leipzig, Germany. Retrieved from http://www.alexandria.unisg.ch/EXPORT/PDF/Publication/38226.pdf

[42] Cooke, B. (2003). A new continuity with colonial administration: Participation in development management. Third World Quarterly, 24, 47-61.

[43] Goggin, N. (2006). Ideology, mission or just a technique? Finance & the Common Good/Bien Commun., 25, 30-36.

[44] Rayman, P., Bailyn, L., Dickert, J., Carre, F., Harvey, M., Krim, R., & Read, R. (1999). Designing organizational solutions to integrate work and life. Women in Management Review, 14(5), 164-177.

[45] Santalainen, T. J., & Hunt, J. G. (1988). Change differences from an action research, results-oriented OD program in high- and low-performing Finnish banks. Group & Organization Studies, 13, 413-440.

[46] Josuttis, N. M. (2007). SOA in practice. Sebastopol, CA: O'Reilly.

[47] Fitzgerald, B., & Kenny, T. (2003). Open source software in the trenches: Lessons from a large-scale OSS implementation. Paper presented at the Proceedings of 24th International Conference on Information Systems, Seattle, WA. Retrieved from http://www.b4step.ul.ie/db/dir/content/brian/79-A.pdf

[48] Creswell, J. W. (2002). Educational research: Planning, conducting, and evaluating quantitative and qualitative research. Upper Saddle River, NJ: Merrill.

[49] Shafi, I. M. (2002). Assessment of the impact of Internet technology use among Saudi business organizations. Dissertations Abstracts International, 63, 3457A.

[50] Walji, N. (2009). Leadership: An action research approach. AI & Society, 23, 69.

[51] Kidd, S. A., & Kral, M. J. (2005). Practicing participatory action research. Journal of Counseling Psychology, 52, 187-195.

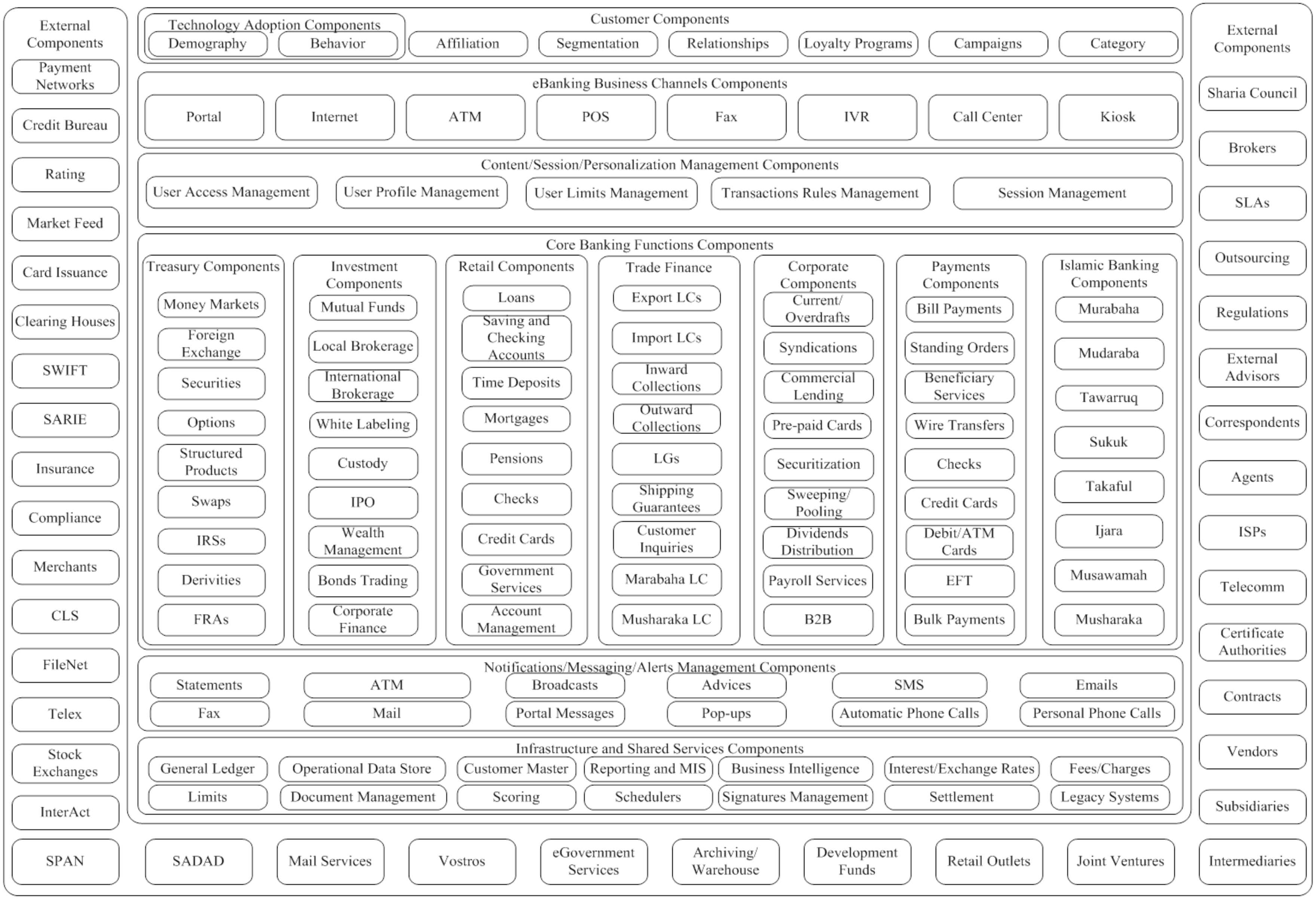
[52] Prilleltensky, I., Prilleltensky, O., & Voorhees, C. (2008). Psychopolitical validity in the helping professions: Applications to research, interventions, case conceptualization, and therapy. In C. I. Cohen & S. Timimi (Eds.), Liberatory psychiatry: Philosophy, politics, and mental health (pp. 105-130). Cambridge, England: Cambridge University Press.

[53] Dick, B., Stringer, E., & Huxham, C. (2009). Theory in action research. Action Research, 7, 5.

[54] Ahmed, A. M., Zairi, M., & Alwabel, S. A. (2006). Global benchmarking for Internet and e-commerce applications. Benchmarking, 13, 68-80.

[55] Sait, S. M., Al-Tawil, K. M., & Hussain, S. A. (2004). E-commerce in Saudi Arabia: Adoption and perspectives. Australian Journal of Information Systems, 12, 54-74.

Figure 1. Saudi Electronic Banking Business Services with TAM Components



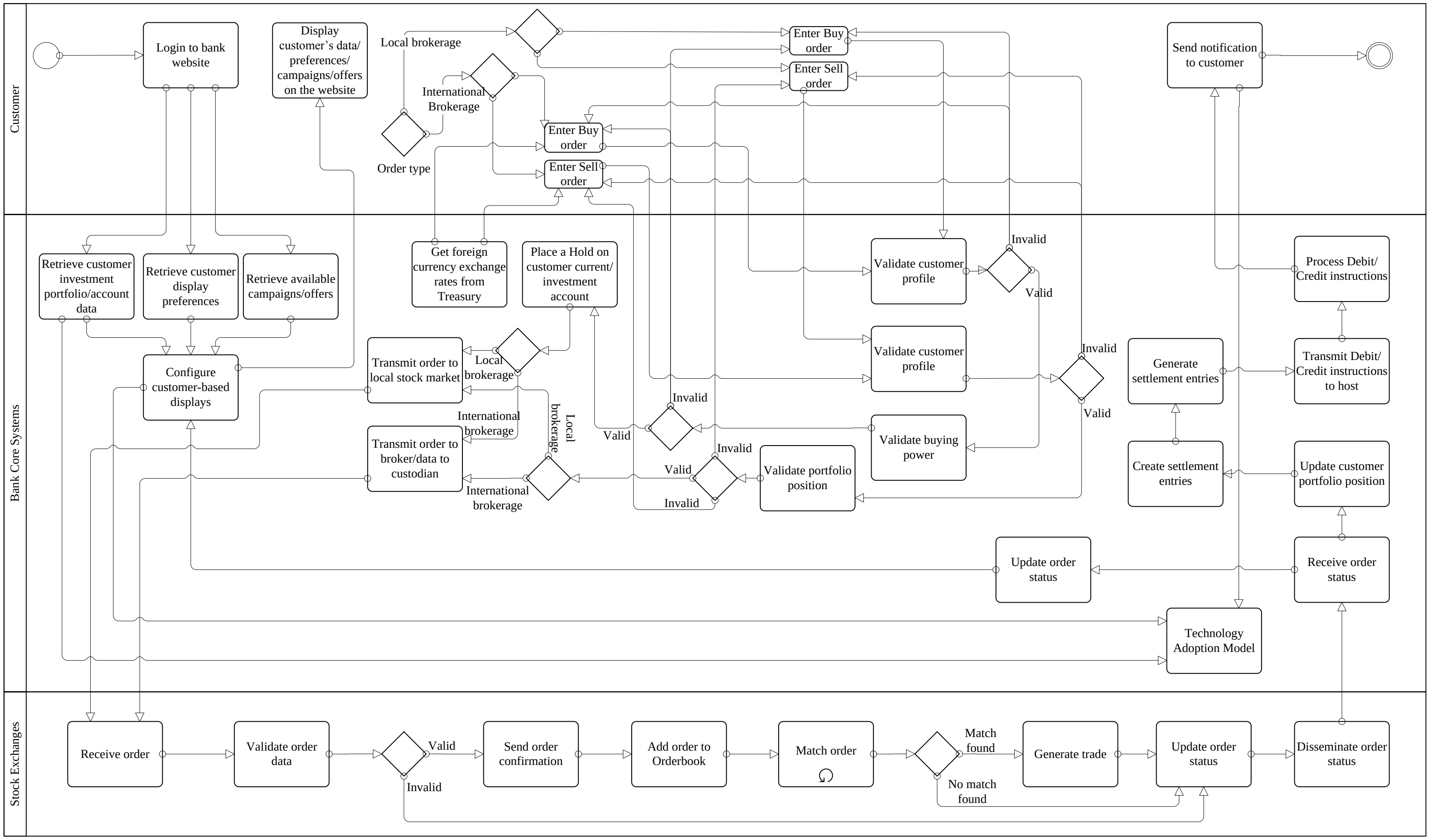

Customer
Login to bank website
Display customer's data/ preferences/ campaigns/offers on the website
Local brokerage
International Brokerage
Order type
Enter Buy order
Enter Sell order
Enter Buy order
Enter Sell order
Send notification to customer
Bank Core Systems
Retrieve customer investment portfolio/account data
Retrieve customer display preferences
Retrieve available campaigns/offers
Configure customer-based displays
Get foreign currency exchange rates from Treasury
Place a Hold on customer current/ investment account
Transmit order to local stock market
Local brokerage
International brokerage
Local brokerage
Transmit order to broker/data to custodian
International brokerage
Valid
Invalid
Valid
Invalid
Invalid
Validate portfolio position
Validate customer profile
Invalid
Valid
Validate customer profile
Invalid
Valid
Validate buying power
Generate settlement entries
Create settlement entries
Process Debit/ Credit instructions
Transmit Debit/ Credit instructions to host
Update customer portfolio position
Update order status
Receive order status
Technology Adoption Model
Stock Exchanges
Receive order
Validate order data
Valid
Invalid
Send order confirmation
Add order to Orderbook
Match order
Match found
No match found
Generate trade
Update order status
Disseminate order status

Figure 2. A Service-oriented Architecture Representation of a Standard Brokerage Business Process Including the Technology Adoption Model Service.

An Innovative and Flexible Model for Change Management

Mutlaq Al Utaibi[1], Basil Al Kasasbeh[1], and Rafe Al Asem2
[1]Department of Information Systems, Al-Imam Mohammed Ibn Saud Islamic University, Riyadh, KSA
[2]Department of Computer Science, Al-Imam Mohammed Ibn Saud Islamic University, Riyadh, KSA

Abstract *- Currently, the changes of the whole world require some changes in different fields of our life, political, economical and environmental. Economical changes have to start from the change management. The flexible model for change management is developed taking into account all positive and/or negative effects that may occur in any time and any place, the Innovative and Flexible Model for Change Management can efficiently find the solution then corrects the current situation to avoid any failure of organization management. This paper attempts to show how to implement the proposed model in change management in an efficient, flexible, and cost/time-effective manner.*

Keywords: Change management, Decision support system, External environment, Internal environment, Processing system.

1 Introduction

Chang Management (CM) is becoming increasingly important to both national and international firms. The purpose of CM is to develop a set of processes that are employed to ensure that significant changes are implemented in a corporation or with a large organization, you might have heard the phrase "change management" used from time to time. Change management has been around for a while, but has become extremely popular with organizations or corporations that would like to initiate significant change to processes that can include both work tasks and culture. Change usually involves three aspects; people, processes and culture Change management is aimed at helping system users to adopt the new system and use it productively. The role of the change management analyst includes ensuring that adequate documentation and support are available to the users. [1]

Chang Management means to plan, initiate, realize, control, and finally stabilize change processes on both, corporate and personal level. Change may cover such diverse problems as for example strategic direction or personal development programs for staff. [2]

Change is the continuous adoption of corporate strategies and structures to changing external conditions. Today, change is not the exception but a steady ongoing process. On contrast ‘business as usual’ will become the exception from phases of turbulence. Change management comprises both, revolutionary one-off projects and evolutionary transformations. [3,4]

In other world, Change Management is a management for plans the change by using the data and analysis from the previous steps to understand the degree of risk and build customized Communication, Learning and Reward plans across all target groups. These three components form an integrated change management strategy that mitigates risk, builds ownership and addresses the changes required in an organization's people and culture to ensure long-term success. [5]

Further more Change Management can be used to promote eco efficiency more broadly including not only pollution prevention but also more efficient use all inputs (fertilizer, water resource, energy, row materials, air pollution, pesticide residues, etc).

In this paper, we propose a new approach in which we use the flexible model for change management. In this way, the change of management is carried out in effective way immediately responding any changes of external and/or internal environments. By this approach always the correct selection is guaranteed.

The rest of this paper is organized as follows: Section 1 is an introduction. Section 2 introduces a short overview of the decision support systems while section 3 briefly describes the business strategy and the environment. In section 4 we describe the details of the proposed approach, and finally, section 5 concludes the paper.

2 Decision support system (DSS)

DSS is defined as an information system that supports organizational decision-making activities of the world of business. DSSs carries out function of serving the management, operations, and planning of organizations and help to make right decisions, which may be rapidly changing and not easily specified in advance. DSS includes knowledge-based systems such as databases and web-based resources. A suitable developed DSS is a software-based system intended to help decision makers find useful information from a combination of raw data, documents, personal knowledge, or business models to identify and solve problems and make decisions [6]. There are several ways to classify DSS applications. Not every DSS fits exactly into one category, but may be a mix of two or more architectures, the DSS classified into six frameworks: Text-oriented DSS, Database-oriented DSS, Spreadsheet-oriented DSS, Solver-oriented DSS, Rule-oriented DSS, and Compound DSS [7]. Any adaptive change management model must take into account the information-based DSSs to perform a right change.

3 Business Strategy and the Environment

Specialists have attempted to improve understanding of environmental management by classifying companies' environmental behavior, and evaluating their performance. Driven by both research and societal interest, this has resulted in a wave of stage or phase models, and a range of different typologies. Kolk et M. [8] gave an overview of the development of such environmental management models, analyzing their characteristics, strengths and weaknesses. An evolution can be noted in the direction of typologies and non-linear models to deal with organizational and strategic complexities. Models are starting to pay more attention to the management side. To overcome problems of operationalization and limited company and sector specificity, environmental performance evaluation systems have emerged more recently. Although comprehensive performance assessments are still unavailable, the tenets of such a system can already be delineated [8].

4 Proposed model

The proposed model “flexible model for change management “consists of many components that work together to give waited results for which the given model was designed, the components are:The external environment effects, the internal environment effects, decision support systems and change management process. Each one of them will be discussed below. (see figure 1 and figure 2)

4.1 External and internal environments effects and decision support systems (DSS)

An organization resides in environments from which it draw resources and to which it supplies goods and services. Organization and environments have a reciprocal relationship. On the one hand, organization is open to, and

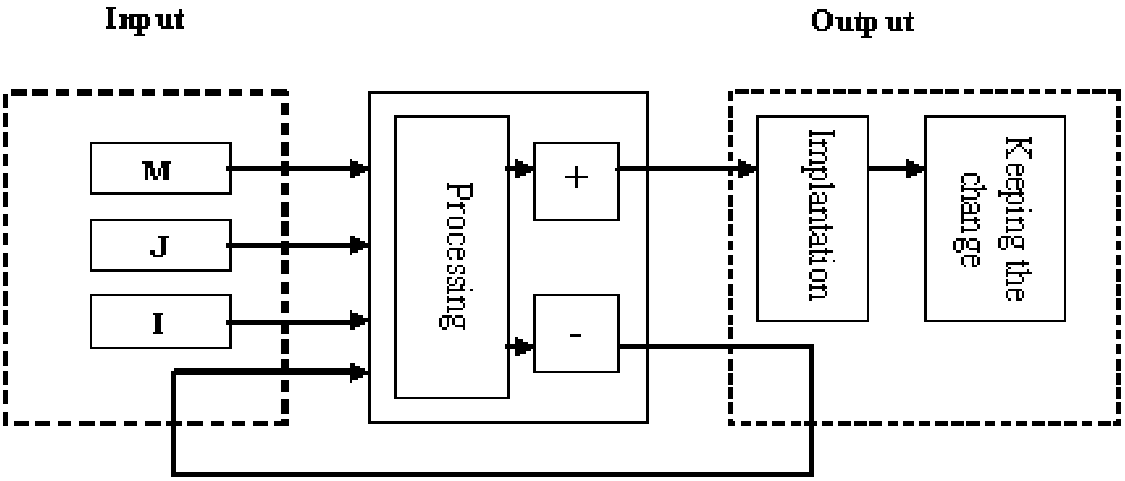


Fig.1 The proposed change management system.

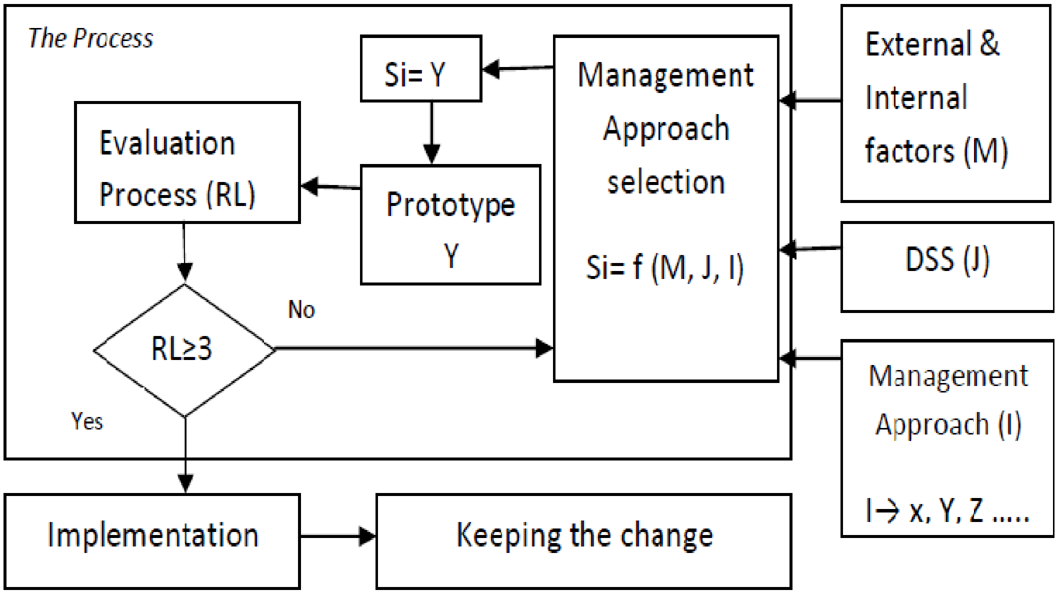


Figure 2. The proposed flexible model for change management.

dependent on, the social and physical environment that surrounds it. Without financial and human recourses organization could not exist. An organization has to obey and follows legalizations and other requirements imposed by government, as well as following the actions of market and competitors, also organization must be respond to any political change within the country or outside (see figure 2).

Decision support systems in any modern organization must contain the following: information data bases, information resources such as World Wide Web (the Internet), planning and marketing studies. [5, 9]

Information systems play a very important role to help organizations perceive changes in their environments along with helping organizations to act in their environments. Information systems are the tool for environmental survey, helping managers identifies external and internal changes that might require an organizational response. Environments typically change much faster than organizations. The organizational failure occurs due to an inability to adapt to rapidly changing environment and a lack of resources to sustain even short periods of troubled times. New coming technologies, new products and changing of public tastes and values put strains on any organization's culture, politics, and people. Many organizations do not cope well with large environmental changes. The inertia built into an organization's standard operating procedures, the political conflict raised by changes to the existing order, and the threat to closely held cultural values typically inhabit organizations from making significant changes.

4.2 Description of the model

The given model contains many parts as explained above, now we will discuss how the system works and how the parties interact with each other to give the correct and positive change that can put the organization in the right path and avoiding any failure of business.

The proposed flexible model for change management is described as a system that likes any system which must be consisted of input, processing and output (with loop back in some cases) (see figure 1). According to figure, variable M is external and/or internal environment(s) factor(s), M may be one or more of the following: external (political factor, Social factor, Legalizations factor, Economical factor, Competition factor, Investment requirements factor, and Environment factor) and internal(Human factor, Capital, Budget Econ. situation and products factor). Variable J refers to decision support system(s) (DSS); J also may be one or more of the following: Planning studies, Information source (WEB), marketing studies and data bases. Variable I refer to method(s) of management (for example centralized, distributed….).

The processing stage of the system includes a management method (approach) choice, intermediate result, prototype (concept) and evaluation process.

The management approach (method) selection step is based on an artificial intelligence that can deal with inputs (M, J, I) to select the best method that will be suitable for a current situation of an organization taking into account inside and/ or outside effects. An intermediate result is the output of the previous step that will require a prototype. What is the meaning of a prototype? Prototype is an experiment to carry out a selected method of management within an organization for few days or weeks (be sure that weeks is always better than days for this case) parallel with the current method of the given organization.

Certainly, after the experimental time in which the selected method of management temporary was carried out, the next stage of the change management process that is called evaluation step should be done.

The evaluation process is based on four factors (see figure 3), the first of them is a human resources factor, who will accept the new introduced method or reject it (resistance), the second factor is the quality of the new approach, does the new approach suitable and can respond the quality assurance issues within the organization? The percentage of quality value must be equal or more than 80%. The fourth factor is the budget of the organization, is the budget can support this new method of management? The answer of this question is important part of evaluation. Evolution process can be described by the following mathematical expression:

$$RL = Qi + Ai + Rres + S$$

Where:

RL is a positive or negative result.

Qi is the percentage of quality;

Ai is the percentage of acceptance;

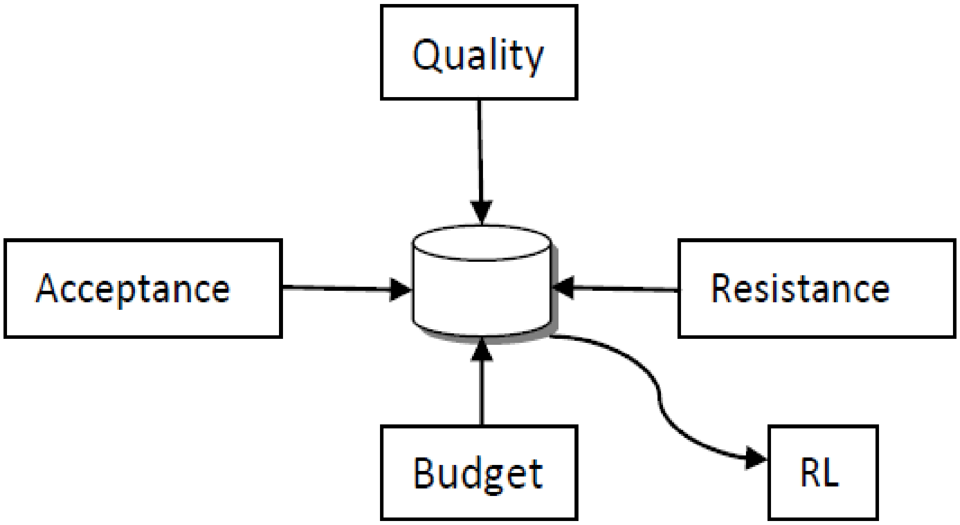


Figure 3. The Evaluation process of change management model.

Ai = Na/Nt ; Na: the number of employees who accept the new approach,

Nt: the total number of employees.

Rres : is the percentage of resistance;

Ares = Nr/ Nt; Nr is the number of employees who reject the new approach.

S is a constant that used for weighted result: S = 0.69.

Now, if the RL ≥ 3 then the result is positive, the method will be implemented in real life of a given organization, after the process of keeping the change will start. If the RL value is less than 3 then the result is negative and the loop back process will start to select another method of management that may respond the requirements of a current situation.

5 Conclusions

The work in this paper attempts to develop a new model which is called An Innovative and flexible model for change management, the given model can be used to perform the change of management based on an intelligent approach to avoid an organization failures and to help it to be sustainable. All outside and/or inside factors took into account. In future, some cases will be experimentally implemented to show the efficiency of the model, an enhanced algorithm that explains the work of the proposed will be developed.

6 References

[1] http://www.the manager.org/strategy

[2] http://www.themanager.org/strategy/change_phases.htm

[3] LaMarsh & Associations (2008).

[4] http://en.wikipedia.org/wiki/Change_management

[5] K. laudon, J. laudon, Management information systems, Prentice hall, 2004.

[6] Efraim Turban, Jay E. Aronson, Ting-Peng Liang (2008). Decision Support Systems and Intelligent Systems.

[7] Holsapple, C.W., and A. B. Whinston. (1996). Decision Support Systems: A Knowledge- Based Approach. St. Paul: West Publishing.

[8] Ans Kolk, Anniek Mauser, The evolution of environmental management: from stage models to performance evaluation, Business Strategy and the Environment, Volume 11, Issue 1, pages 14–31, January/February 2002.

[9] http://en.wikipedia.org/wiki/Management_information_ system

SESSION

NOVEL APPLICATIONS + KNOWLEDGE ENGINEERING AND MANAGEMENT + DATA AND TEXT MINING AND ANALYSIS

Chair(s)

Prof. Hamid R. Arabnia

Knowledge Realization Momentum

The Subtle Trap of Unidirectional Innovation

Gideon Samid
Department of Elect. Eng. and Computer Science, Case Western Reserve University, Cleveland, Ohio, USA

Abstract – *When a certain innovation effort proves productive, it generates momentum in its established direction. Innovators look for a repeat performance. This 'directional commitment' is likely to suppress 'sideways' innovation without which the project will not be accomplished. That 'sideways' may hide an intractable innovation challenge, which remains unattended because of the enthusiasm and success in the directional innovation. When the intractable challenge finally bursts on the scene, it may generate a daunting cost and time outlook, well beyond the prevailing timeline and budget. This article helps identify and quantify the knowledge realization momentum, (KRM), and it suggests effective ways to escape its fate. The risk of premature termination looms strong for goal-inflexible innovation projects. This methodology is therefore especially helpful for industrial R&D. KRM also generates insight with respect to open-ended scientific efforts, and with respect to the unceasing quest to understand reality. This article proposes a regressive methodology for in-depth research, and suggests that KRM may lead to accidental pattern perception of random data.*

Keywords: Knowledge Realization Momentum, Learning, Innovation Management, Innovation Productivity

1 Introduction

While knowledge has been a topic of discussion since the beginning of science, we still don't have a clear definition, not to speak of a clear metric, for its nature and quantity. And hence, we find it hard to rationally gauge an R&D or innovation effort. These efforts are essentially a knowledge realization process. Since we can't measure knowledge, we can't ascertain whether a knowledge realization session was productive or not. This author has circumvented the challenge of quantifying knowledge per se, by defining '*useful knowledge'* as the knowledge needed for an innovator to achieve his or her well specified innovation goal. To the extent that the innovator lacks a given measure of needed (useful) knowledge, his or her estimate of cost-to-complete, and time-to-finish is less credible. Samid then used the credibility metric of an estimate as the measurement of the missing useful knowledge. As the innovator realizes more useful knowledge, the credibility of his or her estimate increases. By measuring this credibility increase Samid measures the intrinsic innovation progress. This allowed him to spot research and development projects that spend a lot of time and money but don't realize much new knowledge. [15,16].

This new metric also allowed Samid to spot the phenomenon of knowledge realization momentum: the tendency to invest effort in the direction that proved productive before. This investment may deprive another innovation challenge from its due attention. If the project as a whole needs the two innovation challenges to be completed, then by not paying due attention to the 'other issue' one risks a total project failure. This is because a poor credibility estimate of cost-to-complete means that the actual cost may be much higher than what is presently computed as the most probable cost. In other words, every innovation issue that is estimated with poor credibility may become a 'cost mine': eventually 'explode' with a very high price tag in terms of dollars or duration. The price tag may render the entire project infeasible. If one succumbs to innovation momentum then the project will look very promising to begin with, because the first innovation topic is moving along very nicely while the 'cost mine' of the other innovation issue is left latent. This means that the innovator and the people who fund him celebrate a false sense of progress, and only find out about the 'cost explosion' at a much later date.

The above description fits the majority of creative innovative projects that end up doomed, and the methodology offered here is a means to avoid this pitfall.

1.1 Related Prior Work

Knowledge has been a target of human investigation for several millennia. The Stoics developed a 'theory of knowledge' [Watson 96, Russell 72], Kant founded the modern version of the same [Kant 29], and 20th century quantum mechanics totally convoluted any sense of general agreement as to the essential nature of human knowledge. Einstein in his famous thought experiment (EPR), argued that the strangeness of quantum mechanics suggests deeper knowledge, not yet revealed. Niels Bohr, and most physicists today (The Copenhagen School), argued that Einstein's box of hidden knowledge is essentially empty - there is no knowledge there to be discovered. This philosophical dimension of knowledge remains unresolved until today [Wheeler 83, Penrose 04]. It was only in the late decades of the 19th century that scientists and industrialists in Germany reframed the generic notion of knowledge [Rosenblum 96, Cornwell 03]. The association of knowledge with the ultimate truth, and non-subjective reality was downplayed.

Knowledge was increasingly regarded as means to resolve intractable technological challenges, [Rosenblum 96]. Organizational entities dedicated to the accumulation of intractability-resolution knowledge came into being. This reframing of knowledge was quickly picked up in the United States which took the lead, and throughout the 20th century became the powerhouse for intractability-resolution knowledge generation.

When artificial intelligence rose to prominence it inspired a fertile thrust of knowledge taxonomy. Notions like rehmatic knowledge, dicent knowledge, designative knowledge, appraisive knowledge, prescriptive knowledge, argumentative knowledge were introduced [Gudwin 1989]. Various knowledge generation operators were developed for the purpose of processing rich databases and extracting rules from facts, discerning similarities and dissimilarities between bodies of data, extracting equations, even conceptual classifications [Kauffman 1991].

The difficulties facing artificial intelligence bear testimony to the intractability of the step that crosses from data to knowledge. While the former was rigorously analyzed by Claude Shannon, Kolmogorov and their followers, making a fully compressed, random look-alike bit string the metric for quantity of data, corresponding metrics for knowledge remained elusive. In a recent work knowledge was quantified in terms of what one needs to know in order to resolve a well defined intractable challenge [Samid 09]. Alas, it is exactly this vagueness with respect to the process of realizing more knowledge that makes one apprehensive on whether the very process of knowledge realization has an impact on the goal of resolving the challenge at hand.

1.2 Unbounded Innovation.

The definition of useful knowledge applies to knowledge needed to accomplish a well specified goal. It does not apply to open-ended research, or to what we shall call ‘Unbounded Innovation’. When one aims to ‘understand an issue’ then at any point of his or her understanding it is hard to say whether all the significant knowledge is known. This is different from a research designed to accomplish a testable goal. Much of the academic research or the research conducted by long-range institutions (e.g. DARPA) may be categorized as Unbounded Innovation where the knowledge metric discussed above does not apply. And hence the notion of knowledge realization momentum is reduced to a conjecture only. It says: human researchers will develop a directional commitment (momentum) to pursue their research in the direction that proved productive before. This practice leaves sideways knowledge untouched, and its benefit unrealized. It further skews the view of reality because so much of reality is left unresearched. Today, when cutting edge research requires considerable funding, this blinding phenomenon only becomes worse. Fund managers tend to increase their return by betting on established directions.

To illustrate, consider two people approaching a library with the intent to describe what is in it. One person is a professional librarian. She will first catalogue all the books on the shelves, familiarize herself with their subjects, and their main message. The other person is a typical reader. He will browse the books until he would hit a book that holds a great deal of personal interest for him. He will consume this book cover to cover, and then read likewise books. When later on these two people will describe what the library is about, their descriptions will be very different from each other. Intuitively we will say that the librarian has a more accurate take on what the library is about, and she will be more useful when a need arises to find knowledge for some unexpected topic.

The implications for science in general are that our fast moving innovation train in bioinformatics, digital technology, and green energy may deprive us from some wonderful revelations in areas where no momentum has yet been established. The only way to rectify this is by being aware of the situation, which is another reason for publishing this work.

2 The KRM Model

In 2002 Samid introduced an objective metric to gauge the progress of a research and development project. Accordingly, the validities of the quantitative estimates of the project resources measure the unknown that still needs to be discovered before the project is declared a success. The KRM model builds on this metric to measure the balance among the various estimates: are they all becoming more valid, or is one resource, one aspect of the project receiving the lion share of innovative attention. The latter is what happens when knowledge realization momentum drives the innovation efforts. And so here we propose a mathematical expression that tracks the knowledge realization momentum. This model will also help countering its deficiencies. We then take a leap of faith towards open-ended innovation projects where the validity metrics don’t apply.

2.1 Measuring Knowledge Momentum

Consider an innovation project, namely an a project where a measure of innovation is needed to accomplish its goal. This means that at present one cannot specify a well defined procedure, or algorithm to achieve the stated goal. And hence one is somewhat ambiguous as to the quantities of resources that would be required to accomplish the same. Once the exact procedure, or algorithm, are specified, so would be the measure of resources. Let it be known that some r resources participate in the solution procedure. At any given point we may estimate the required quantity of resource *i* with a credibility or validity marked as V_i such that Vi=0 indicates zero validity, and Vi=1 indicates zero uncertainty. Once innovation is done, all is clear and the validity of the estimate for each required resource *i* is Vi=1 for i=*1,2,…,*r. At any point *p*' before that, the validities will be V'_1, V'_2,….V'_r ≤ 1.00, and hence the project history could be marked as a path on an *r*-dimensional metric space where the path starts at *p'* and ends at [1,1,1,…1], where each dimension reflects a validity measure of resource *i (i=1,2,…r).*

One could use the Samid validity metrics for the *r* resources, or any other metric, to construct this *r*-dimensional construct.

We may now rigorously define the Knowledge Realization Space (KRS) as follows: Let the accomplishment of a certain innovation goal be carried out through a procedure that refers to *r* resources, each needed in a definite measure, R_i, *i=1,2,…r*. At any given state of progress the best estimate for resource *i* is S_i. S_i is estimated with credibility, or validity V_i where 0 ≤ V_i ≤ 1.00, such that V_i=0 is zero validity (zero credibility, a wild baseless guess), and V_i=1 indicates zero uncertainty, a maximum validity estimate. This innovation project will be associated with a Knowledge Realization Space (KRS) – a metric space of *r* dimensions where each dimension *i (i=1,2,…r)* is associated with one project resource *i*, and such that $V_i=0$ will be at the origin, and $V_i=1$ will be at the end stretch of the dimension. The *r* dimensions will be mutually scaled to reflect the best estimate S_i values for the various *i* values. So the length of dimension *i* will be S_i.

Since the state of V_i=1 for *i=1,2,…r* represents the state of project complete (no more uncertainty with respect to any resource), then the most efficient pathway from any given point *q* to the final point [1,1,..1],will be the straight line between these two points on the knowledge realization space. This straight line can be described using vector notation as (Eq-1):

$$\vec{O} = \sum_{i=1}^{i=r} S_i(1 - V_i)\overrightarrow{U_\iota}$$

where $\vec{O}$ is the vector that starts at point *q* on the KRS, and ends at the end-point [1,1,….1] there, and $\overrightarrow{U_\iota}$ represents the unit vector in direction *i*. The actual move from point *q* to point *q'* on the KRS can be expressed using vector noation as follows (Eq-2):

$$\overrightarrow{qq'} = \sum_{i=1}^{i=r} \left[\!\!\left[\int_q^{q'} \left(\frac{\partial V_i}{\partial e}\right) de \right]\!\!\right] \overrightarrow{U_\iota}$$

where *e* is the innovation effort leading from state *q* to state *q'*. e can be measured in dollars, time or otherwise. The knowledge momentum effect will be measured by the scalar product of these two vectors (Eq-3):

$$KRM = \frac{|\vec{O}|\,\overline{|qq'|}}{\vec{O} \bullet \overrightarrow{qq'}} - 1$$

such that a well balanced innovation path will compute to KRM=0 (no momentum distraction), and the greater the knowledge realization momentum effect the higher the value of KRM.

One can use Eq-3 both in order to appraise a planned innovation move, and as a way to measure how focused on the project goal (and not distracted by knowledge momentum) have we been.

Projects that show chronic distraction (persistent high KRM values) warrant special attention. Perhaps a leadership change, or even defunding.

2.2 Choice of Resources

It is practically impossible to list and manage all the various resources used in accomplishing an innovation goal. One must pick a good representation of the required resources. A good representation will be such that when (i) the required measures of all the listed resources is well known, the innovation goal can be readily accomplished, and (ii) when the innovation progress cannot be properly managed using a subset of the listed resources. The latter requirement prevents one from amassing secondary resources of no much meaning for the project. For instance, the number of test tubes needed to synthesize a desired chemical. The first requirement is designed to insure that all the resources that are essential for the project are accounted for. For instance: innovation efforts to make oil from coal may need to measure the required coal per a barrel of oil, the required quantities of the non-coal ingredients, and the required energy for this transformation. As long as either one of these quantities is not known with high validity – the innovation load is still looming.

2.3 Capabilities as Resources

Any component of an innovation project can be measured by its required resources, or by the estimate of the effort needed to make it work as required. In the first option the credibility of the estimate of the required resources will be tracked, and in the second case the credibility of the estimate of the effort to complete will be tracked. For example, to innovate a cancer killing drug one needs to have a high credibility estimate of the effort to engineer a drug that will kill the cancer cell once there, and also to have a high credibility estimate of the required effort to innovate an effective way to lead the drug to the site of the cancer. In this case the resources can be viewed as a service for the goal. Clearly when one has a good estimate as to what will it take to synthesize the cancer killing drug, and what will it take to haul it there – then the original project can also be estimated in very good credibility. One can also mix a project component with some nominal resources to build the project management resource list.

Using a project component in appraising and managing an innovation project may lead to a cascade in which the progress of the original project is estimated from the credibility of the estimates of some derived projects. The derived projects are not necessarily components of the original project, but rather associated projects such that when they are accomplished they make the accomplishment of the deriving goal easier. Such relationship is described in Samid 06. Accordingly, if the effort to accomplish an innovation goal is *P*, then one could search for an associated innovation challenge for which the accomplishment effort is *P'* and such that the following inequality will hold, (Eq-4):

$$P > P' + P|P'$$

where *P|P'* is the effort to resolve the original challenge after having resolved the associated challenge *P*'. Eq-4 can be cascaded indefinitely (Eq-5):

$$P > P|P' + P'|P"+\mathrm{P"|P'"} + \ldots$$

and where the associated projects are either a breakdown project, an extension project, or an abstraction project as described in Samid-06.

3 KRM v. The Gantt Chart

Many innovation projects are based on nominal projects and hence are designed as a Gantt chart where a series of project components follow each other based on their logical sequence. The innovator who would follow the Gantt sequence will show a poor showing on his KRM score. The question arises: *should one adhere to the logical sequence expressed in Gantt, or vie for the KRM strategy*? The answer depends on the innovation content of the project at hand. If the innovation is low to moderate, then the Gantt sequence offers order and manageability, but if the innovation content is high, it would make sense to give priority to KRM consideration. It is not always possible, many projects are so structured that task B must follow task A, but to the extent possible KRM should have priority.

3.1 Illustration: New Drug Administration

A certain pharmaceutical company experimented with a new molecule for its target disorder. It was clear upfront that the new drug would have to be administered as a skin patch. Alas, due to the large size of the molecule it was not clear whether the prevailing patch will be effective. The Gantt chart logic called to first develop the drug, be sure of exactly what molecular structure it is, and only then worry about how to get it into the body. The KRM approach called for dividing the research attention between the two parts because it does not make sense to resolve the uncertainty in only one part, allowing the other part to doom the project. By shifting to the KRM strategy the company researched the drug administration with a like-size molecule. It discovered some insurmountable difficulties and changed the project.

3.2 Illustration: Using Car Traffic to Clean the Air

An academic research group has developed a concept whereby an ingenious powerful adsorber will be installed in front of a typical car radiator fan, and the drawn air will be scrubbed from air pollutants. Once the adsorption contraption becomes mandatory, so the idea went, the urban air will be spot clean! The group worked on the adsorption chemistry with a lot of enthusiasm, neglecting the uncertainty with respect to the expected impact on the air pollution. It was late into the research when one computed the figures to show that even if all the cars in America would have been fitted with this device, and even if it would have functioned in top performance – the impact on the urban air would have been minimal. A level headed KRM approach would have attended to this computation much earlier.

4 KRM and Result Flexibility

The KRM methodology is critical for innovation projects with poor result flexibility. Namely projects where the goal is dictated from above, and is not subject to change. This is the case when the R&D effort is an attempt to solve a problem that showed up, and needs a solution. The researcher in that case does not have the freedom to say, I will solve a different problem. In most industrial and military contexts, the innovation goal is fixed and nonnegotiable, and in such cases, it is crucial to implement KRM. It is a different case when the R&D is academic. In that case the research team could say, *yes we set out to discover A, but, guess what, we have discovered B – so be it, we will report the unexpected discovery of B*. If the goal, the result, is flexible the innovator could say *I will develop something else*. In fact a large majority of start-up companies set out to develop one product but end up developing quite another.

5 KRM and Funding Philosophy

Fund managers are naturally apprehensive of research directors reporting high percentage of progress along the way. They suspect that bad news are suppressed to the last moment, and this apprehension halts their funding. By implementing KRM, a research director is communicating to the funding authority that the residual project uncertainty will be reduced as fast as possible, and if indeed the project hides a latent 'cost mine' – it will be flashed out much earlier than otherwise. This attitude will increase the fund manager inclination to invest in that innovation project.

6 KRM v. Knowledge Per Se

The KRM methodology as defined here is applicable only to situations where a fixed innovation goal guides the action and measures its efficiency. This is not the case in a 'general learning' environment where one tries to learn what there is, as opposed to achieving a well stated goal. However, we may define our research goal as learning all that there is to learn, regardless of utility. The difficulty with such a goal is, of course, that at any point we are not sure whether there is more that is still hidden, so we can't use this goal the way we use a regular R&D objective. Albeit, being aware of the knowledge Realization Momentum, we may reasonably suspect that our learning path is skewed in the direction of our learning momentum, and as a result large knowledge zones are left untouched, and unrecognized. This conjecture brings forth the question, how can we reach out to this unchartered territory. In this open case we don't have the estimate credibility metric to guide us. One possible approach is herewith outlined:

The process of science may be described as a series of conclusions drawn from a body of data. So we have body of data D_1 leading to conclusion C_1, body of data D_2 along with C_1 leading to conclusion C_2, etc. Given our KRM tendency we should suspect that every conclusion we have drawn from a body of data is not exhaustive, namely there are more conclusions that may be drawn from the same data. We missed them originally because we acted under the influence of the knowledge realization momentum. Yet, upon revisiting the same body of data we may discover formerly overlooked conclusions. Once we have found such a new conclusion, we may regard it as the starting point of another conclusion sequence. In other words, we are calling here for a regressive approach: revisiting past conclusions, searching for additional conclusions we missed before. To be productive we may focus on bodies of data that led to conclusions that were part of an enthusiastic discovery drive. These bodies of data, in particular, might hide some important latent conclusions that were passed over as the knowledge realization momentum directed the research attention into its unidirectional way.

This mechanism is consistent with the anecdotal finding that big leaps of science are discovered and built by unlikely explorers or builders. These innovators, because they are not well learned in the subject, don't suffer from the knowledge realization momentum. They have a fresh look. Thomas Edison said that had he been schooled in electrical engineering he probably would not have experimented so stubbornly with the electrical bulb. The Wright brothers were not highly educated mechanical engineers but rather bicycle repairmen. Paracelsus (1493-1541) revolutionized 16^{th} century medicine despite lacking any formal medical education, concocting specific prescriptions to specific ailments, and ignoring the search for a single 'philosopher stone' as a catch-all generic remedy. He also discredited the vague 'four humors' theory that prevailed among the well educated physicians. In World War II, the German suspected that their Enigma cipher was compromised, but they did not redesign a new cipher – they simply added a fourth wheel to the three wheels cipher they used before – a classic KRM case that shortened the war and saved many lives... Windows – the quintessential operating system for personal computers was developed by two college dropouts, not by learned computer scientists, and ever since it is being patched and repatched for security holes, rather than redesigning a solution from the ground up. Venture capitalists today invest millions in very young innovators because they have not been tainted by KRM.

Because KRM may be rather productive for quite a long time, it effectively hides the fact that more astonishing unknowns are waiting untouched. It's only when an established direction erodes in its productivity that off-shoot directions are being sought.

Modern physics drowns in complexity while it has revolutionized the role of the scientist from a neutral observer to a reality player. This conclusion suggests that we should have a better grasp of physics if we advance our understanding of the human brain and its psychology. Alas, such an interdisciplinary research is unwelcome by most modern physicists, so it is left largely unpracticed. This highlights another aspect of KRM: *expertise*. If an effective research plans calls for work to be done in a discipline which is alien to the researching team, then this conclusion will be de-emphasized, and KRM will reign.

Regarding the off-shoots of conclusion sequences from a given body of data one wonders whether a major scientific conclusion may be drawn through more than one logical path. One may conjecture that much as Feynman's '*sum over histories'* states that moving particles take all possible routes from A to B, so it may be that major conclusions of science can be reached through various logical pathways, and the way to distinguish between theories is not just to what extent they correctly predict the results of a future observation, but to the extent that they suggest creative and un-thought of new experiments and observations.

7 KRM and a Theory of Patterns

Knowledge Realization Momentum can be employed as an explanation for perceived patterns by a generic data reader. Let's consider a random source of data: spewing random bits. These bits are met by a bit reader who has an *a-priori* probability $p_0 < 1$ to read any bit coming its way. The KRM phenomenon is expressed mathematically as follows: the probability of the bit reader to read a given oncoming bit increases with the number of bits of same identity that have been read by the reader before. These two terms would lead the reader to de-randomize the incoming data, and establish a stable pattern based solely on the chance reading of the first bits. A simple exercise using $p_0 = 0.1$, and bit reading probability: $p_0 + (1- p_0)r_b$ where r_b is the bit fraction of same identity bit, will transform a pseudo-random bit series into a clear-pattern series (where one bit appears 60% to 70% of the times). Such pattern-perception may be cascaded upwards, and also may be applied to 2, 3 or more bit sequences. The conclusion here is that if our biological sensors of reality are subjugated to the KRM phenomenon (as Darwinian evolution suggests) then the patterns that we observe in reality are accidental to the history of our reality reading before.

8 Conclusions

Knowledge Realization Momentum (KRM) is identified as a natural human way to conduct research, to innovate, to acquire new knowledge. While KRM may appear very productive to begin with, it may hinder the fulfillment of goal-oriented projects that also require innovation in a different direction. KRM may also distort an open-research aimed at knowledge per-se.

For goal oriented, results-inflexible R&D, we have developed a methodology to measure and correct for KRM and thereby boost the productivity of the innovative effort. For open-ended research we propose a regressive methodology to revisit bodies of data from which earlier conclusions have been drawn under the driving knowledge realization momentum. Such data, upon re-examination, may yield new insight that lay latent as a KRM victim.

This work is consistent with the emerging trend in science: to improve productivity by accounting for common psychological tendencies.

9 References

[1] Albus, J.S., "Outline for a Theory of Intelligence", IEEE Transactions on System Man and Cybernetics, Vol. 21, No. 3, May/June 1991

[2] Bess J. 1995 "Creative R&D Leadership" Quorum Books

[3] Carlisle 2004 "Innovations and Discoveries" John Wiley and Sons

[4] Carlson C, Wilmot W. 2006 "Innovation" crown Business NYC

[5] Cornwell, J 2003 "Hitler's Scientists" Viking Press

[6]Doignon Falmagne 1999 "Knowledge Spaces" Springer

[7] Ernst Newell 1969 "GPS A Case Study in Generality and Problem Solving" Academic Press

[8] Fagerberg et al 2005 "The Oxford Handbook on Innovation" Oxford Press

[9] Gudwin, R. 1998 "On the generalized deduction, induction and abduction as the elementary reasoning operators within computational semiotics" Intelligent Control (ISIC), 1998.

[10] Jelinek, M 1979 "Institutionalizing Innovation" Praeger Publishers

[11] Kant, I. 1929, "Critique of Pure Reason" trans. Norman Kemp Smith. New York: St. Martin's Press

[12] Kauffman, K. 1991 "knowledge extraction from databases" 6th international symposium on methodologies for intelligent systems ISMIS91

[13] Rosenbloom Spencer 1996 "Engines of Innovation" Harvard Business School

[14] Samid, G. 2006 “The Innovation Turing Machine” DGS Vitco

[15] Samid, G. 2002 “R&D Cost Estimation” PhD Dissertation Technion – Israel Institute of Technology

[16] Schmitt, S 1969 "Measuring Uncertainty" Addison Wesley

Parallel All Pairs Similarity Search

Amit Awekar, and Nagiza F. Samatova[1]
acawekar@ncsu.edu, samatovan@ornl.gov
North Carolina State University, Raleigh, NC
Oak Ridge National Laboratory, Oak Ridge, TN
[1]Corresponding Author

Abstract— **This paper presents the first scalable parallel solution for the All Pairs Similarity Search ($APSS$) problem, which involves finding all pairs of data records that have a similarity score above the specified threshold. With exponentially growing datasets and modern multi-processor/multi-core system architectures, serial nature of all existing $APSS$ solutions is the major rate limiting factor for applicability of $APSS$ to large-scale real-world problems and calls for parallelization. Our proposed $index\ sharing$ technique divides the $APSS$ computation into independent searches over the central inverted index shared across all processors as a read-only data structure and achieves linear speed-up over the fastest serial $APSS$ algorithm in shared memory environment. We demonstrate effectiveness of our solution over four large-scale real world million record datasets.**

I. Introduction

All Pairs Similarity Search *(APSS)* is the problem of finding all pairs of data records having similirty score above the specified threhsold. Similarity between two records is defined via well known similarity measures, such as the cosine similarity or the Tanimoto coefficient. Many Business and scientic applications like search engines, and systems biology frequently solve the $APSS$ problem over high diemnsional datasets having several millions or billions of records.

The nature of the existing $APSS$ solutions is compute- as well as data-intensive. Given a dataset with n data records in a d dimensional space, where $n << d$, existing $APSS$ algorithms compute $O(n^2)$ similarity scores, while searching through $O(n*d)$ size inverted index that maps each dimension to a list of data records having a non-zero projection along that dimension.

Existing solutions for $APSS$ are all limited to serial algorithms [1], [2], [3], [4], [5], [6]. The compute- and data-intensive nature of $APSS$ is a rate limiting factor for the $APSS$ applicability to large-scale real-world problems and calls for alternative approaches. Processor clock rates are not expected to increase dramatically in the near future [7]. With the emergence of shared memory multi-processor, multi-core architectures, parallel algorithms that take advantage of such emerging architectures are a promising strategy. Throughout this paper, we will use the term *processor* to refer to a single processor or a processing core within a multi-core processor, unless stated otherwise. Inspired by the success of parallel computing in dealing with large-scale problems [8], [9], [10], we explore parallelization to further speed-up $APSS$ computation.

Parallel algorithms for $APSS$ should enable processing of large datasets in a reasonable amount of time. Web-based applications like search engines, online social networks, and digital libraries are increasingly dealing with more massive datasets [11], [12]. Without scalable parallel $APSS$ algorithms, it will likely not be practical to run $APSS$ over some of these applications' datasets, which are growing at an exponential rate [13], [14].

A scalable, parallel solution for $APSS$ will effectively help design scalable, parallel solutions for important data mining tasks like clustering and collaborative filtering that use $APSS$ as their underlying operator. Middlewares like pR [15] can use parallel solution for $APSS$ to speed-up statistical analysis algorithms.

The compute- and data-intensive nature of the $APSS$ problem poses the following technical challenges for its parallel solution:

- The inverted index is shared across all processors and updated incrementally.
- The huge size of the dataset and of the inverted index makes data transfers between processors prohibitively expensive.

Our proposed *index sharing* technique addresses these challenges by parallelizing the $APSS$ computation into independent searches over the central inverted index, which is shared as a read-only data structure across all processors. Index sharing builds the whole central inverted index before starting the search for *matching pairs*, unlike existing $APSS$ algorithms that incrementally build the inverted index while searching for *matching pairs*. Each processor keeps and updates its own copy of index metadata of reasonably small size, resulting in slightly larger memory footprint. A subset of data records is assigned to each processor to find the corresponding matching pairs. We explore various static and dynamic strategies for distributing data records among processors. We empirically evaluate the performance of the proposed index sharing technique using four real-world million record datasets described in Section VI.

To the best of our knowledge, this is the first work that explores parallelization for $APSS$. We propose the following contributions:

- We develope the index sharing technique to parallelize the APT algorithm [6], which is the fastest serial $APSS$ algorithm (Please, refer to Figures 1 and 2).

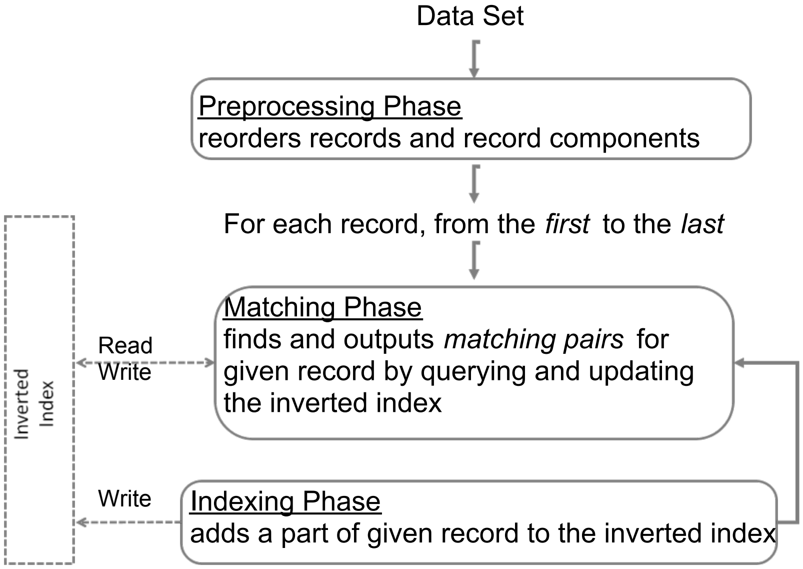

Data Set
Preprocessing Phase
reorders records and record components
For each record, from the *first* to the *last*
Matching Phase
finds and outputs *matching pairs* for given record by querying and updating the inverted index
Indexing Phase
adds a part of given record to the inverted index
Inverted Index
Read
Write
Write

Fig. 1: Unifying Framework for Recent Exact $APSS$ Algorithms

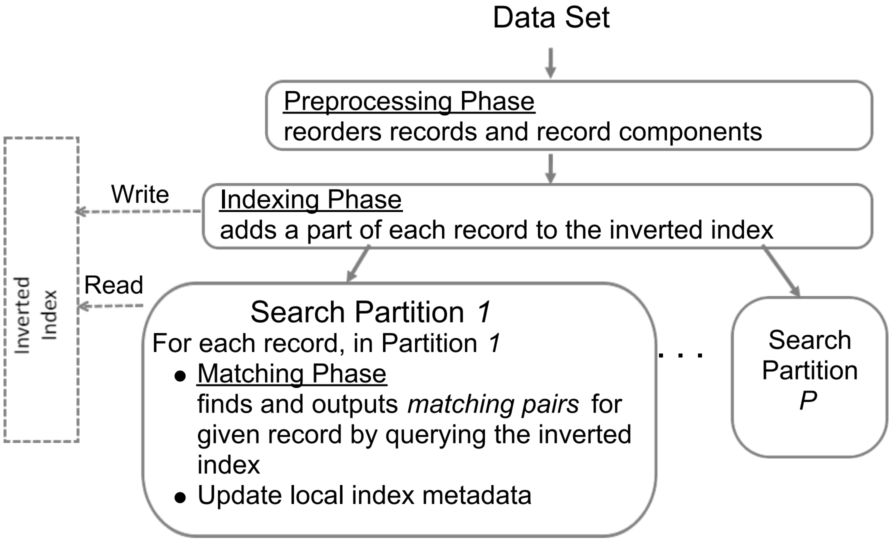

Data Set
Preprocessing Phase
reorders records and record components
Indexing Phase
adds a part of each record to the inverted index
Write
Read
Inverted Index
Search Partition *1*
For each record, in Partition *1*
Matching Phase
finds and outputs *matching pairs* for given record by querying the inverted index
Update local index metadata
. . .
Search Partition *P*

Fig. 2: Proposed Parallel $APSS$ Solution

- Our index sharing based parallel $APSS$ algorithm achieves linear speed-up over the APT algorithm in a shared memory environment.
- We provide a scalable solution to perform $APSS$ over large datasets in a reasonable time.

II. Definitions and Notations

In this section, we define the $APSS$ problem and other important terms referred throughout the paper.

Definition 1 (*All Pairs Similarity Search*): The all pairs similarity search ($APSS$) problem is to find all pairs (x, y) and their exact value of similarity $sim(x, y)$ such that $x, y \in V$ and $sim(x, y) \ \geq t$, where

- V is a set of n real valued, non-negative, sparse vectors over a finite set of dimensions D; $|D| = d$;
- $sim(x, y) : V \times V \rightarrow [0, 1]$ is a symmetric similarity function; and
- t, $t \in [0, 1]$, is the similarity threshold.

Definition 2 (*Inverted Index*): The inverted index maps each dimension to a list of vectors with non-zero projection along that dimension. A set of all d lists $I = \{I_1, I_2, \ldots, I_d\}$, i.e., one for each dimension, represents the inverted index for V. Each entry in the list has a pair of values (x, w) such that if $(x, w) \in I_k$, then $x[k] = w$. The inverse of this statement is not necessarily true, because some algorithms index only a part of each vector.

Definition 3 (*Candidate Vector* and *Candidate Pair*): Given a vector $x \in V$, any vector y in the inverted index is a candidate vector for x, if $\exists$ j such that $x[j] > 0$ and $(y, y[j]) \in I_j$. The corresponding pair (x, y) is a candidate pair.

Definition 4 (*Matching Vector* and *Matching Pair*): Given a vector $x \in V$ and the similarity threshold t, a candidate vector $y \in V$ is a matching vector for x, if $sim(x, y) \geq t$. We say that y matches with x, and vice versa. The corresponding pair (x, y) is a matching pair.

During subsequent discussions we assume that all vectors are of unit length ($||x|| = ||y|| = 1$), and the similarity function is the cosine similarity. In this case, the cosine similarity equals the dot product, namely:

$$sim(x, y) = cos(x, y) = dot(x, y).$$

Our algorithm can be extended to the Tanimoto coefficient and other similarity measures using simple conversions derived by Bayardo *et al.* [3].

III. Previous Work

All existing algorithms for $APSS$ are serial in nature. Previous work on $APSS$ can be divided into two main categories: heuristic and exact.

Main techniques employed by heuristic algorithms are hashing, shingling, and dimensionality reduction. Charikar [16] defines a hashing scheme as a distribution on a family of hash functions operating on a collection of vectors. For any two vectors, the probability that their hash values will be equal is proportional to their similarity. Fagin *et al.* [17] combined similarity scores from various voters, where each voter computes similarity using the projection of each vector on a random line. Broder *et al.* [18] use shingles and discard the most frequent features.

Recent inverted index based exact algorithms for $APSS$ have outperformed the heuristic algorithms. These exact algorithms share a common three-phase framework of:

- data preprocessing: sorting data records and computing summary statistics;
- pairs matching: computing similarity between selective record pairs; and
- record indexing: adding a part of data record to an indexing data structure.

The preprocessing phase reorders data records and record components using various attributes such as: the number of components in a data record or the maximum component value within a data record. Then, the matching phase identifies, for a given record, corresponding pairs with the similarity above the specified threshold by querying the inverted index. The matching phase also removes redundant entries from the inverted index. The indexing phase then adds a part of the given data record to the inverted index. The matching phase dominates the computing time of $APSS$, and the time spent during preprocessing and indexing is negligible.

The APT algorithm [6] is also based on the common framework described above and is the fastest serial algorithm for $APSS$. Please, refer to Figure 1 for an overview of the APT algorithm.

IV. Index Sharing

The index sharing technique is based on parallelizing the matching phase of $APSS$ into independent searches over *the central inverted index which is shared across all processors* as a read-only data structure. In contrast, *the data records are partitioned among processors* to perform the matching phase. We explore both static and dynamic partitioning strategies. Each processor performs the matching phase independently of other processors using the central inverted index. While finding the matching pairs for a given record using the central inverted index, the procedure used by the index sharing technique is the same as the procedure used by the APT algorithm. Please, refer to Figure 3 for an overview of the index sharing technique.

Read and write access to the inverted index during the matching phase is the major bottleneck in parallelizing the APT algorithm (please, refer to Figure 1). This bottleneck arises because of the following two reasons:

1) APT algorithm adds a given data record to the inverted index only after performing the matching phase for that record. Thus, new entries are added to the inverted index during the matching phase.
2) The matching phase in the APT algorithm updates the inverted index during each search by discarding redundant entries from the inverted index.

If the matching phase in a parallel $APSS$ algorithm requires read and write access to the central inverted index, then each processor will require exclusive access to the inverted index, resulting in *synchronization overheads*. Index sharing overcomes this bottleneck by building the inverted index before starting the matching phase and by replicating the reasonably small size index metadata across all processors.

To perform the matching phase for a given data record, all the data records with the ids prior to the given record's id must be present in the inverted index. Hence, the APT algorithm cannot be adopted directly to perform search for multiple data records simultaneously. Index sharing overcomes this limitation by building the whole inverted index before starting the matching phase (please refer to Figure 2).

A. Index Metadata Replication

Index metadata is the set of start offset values maintained, one for each list of data record ids in the inverted index, indicating the front of the list. Index sharing replicates the index metadata across all processors to eliminate the need

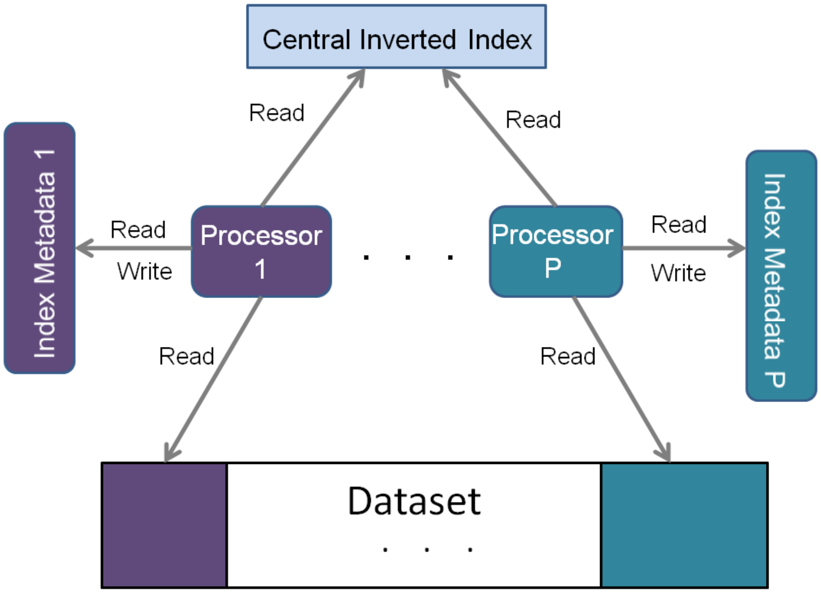


Fig. 3: Overview of the Index Sharing Technique

for any synchronization among processors while performing the matching phase. For a dataset with n data records in a d dimensional space, the inverted index contains $O(n * d)$ entries, while the size of the metadata is only $O(d)$. Compared to the size of the inverted index, the size of the metadata is reasonably small and grows linearly with the number of dimensions.

The matching phase of the APT algorithm requires write access to the inverted index to discard redundant entries from the inverted index. The preprocessing phase in the APT algorithm sorts data records in the decreasing order of the maximum value of any component within the record. Using this sort order, the APT algorithm derives a lower bound on the size of data records in the inverted index to match with any of the remaining data records. While performing the matching phase for a given record, the entries that correspond to data records not satisfying the lower bound on their size are discarded from the inverted index.

For time efficiency purposes, the APT algorithm does not actually remove the redundant entries from the inverted index, but only ignores them using the index metadata. The APT algorithm uses arrays for representing lists in the inverted index. Deleting an element from the beginning of a list will have linear time overhead. Instead of actually deleting such entries, the algorithm simply ignores these entries by removing them from the front of the list. The start offset corresponding to an inverted list array is incrementally advanced as entries are removed from the front.

Index sharing replicates the index metadata across all processors to eliminate the need for synchronization between processors while performing the matching phase. Each processor updates its local index metadata after performing the matching phase for every data record assigned to it.

V. Load Balancing Strategies

The goal of the index sharing technique is to divide the computation workload of the matching phase roughly equally across all processors. We consider two static partitioning strategies (Block and Round-Robin) and a dynamic load balancing strategy.

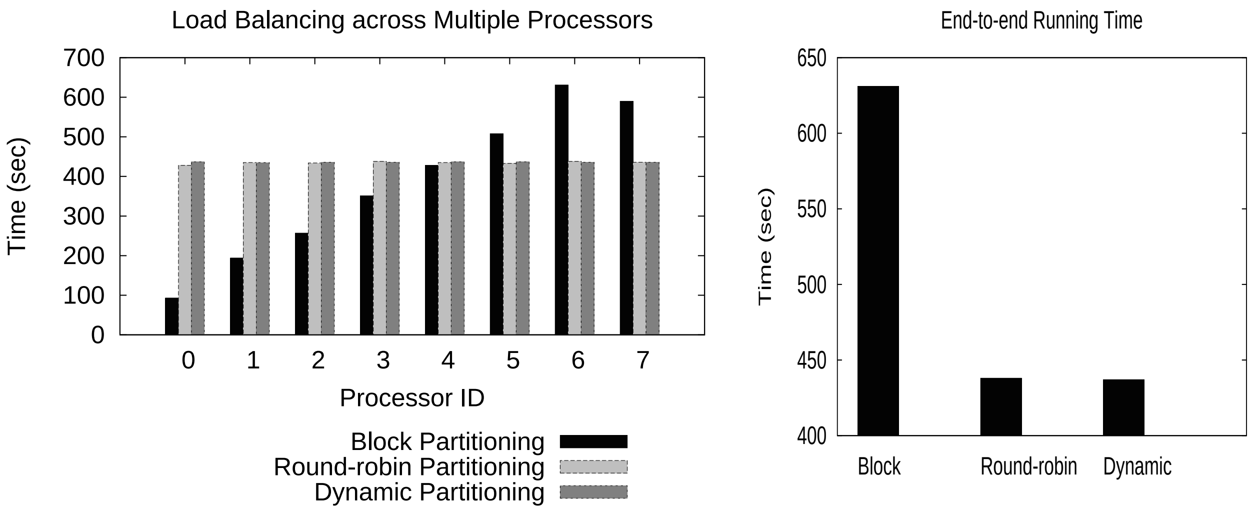
A. Block Partitioning

The block load balancing strategy assigns a contiguous block of data records to each processor. The time required to perform the matching phase for a given data record increases as $APSS$ proceeds from the beginning of the dataset to the end. This variation arises because the preprocessing phase puts short data records, i.e. records with fewer number of non-zero components at the beginning of the dataset. Compared to short data records, longer data records require more time to perform the matching phase because they generate comparatively more candidate pairs for evaluation.

Due to the variation in the time required for the matching phase of data records, assigning equal number of contiguous data records to each processor will likely create severe work imbalance among the processors. The block load balancing strategy tries to compensate this imbalance by assigning an equal number of components to each processor. If short data records are assigned to a processor, then that processor will have more number of data records assigned than a processor with longer data records assigned.

B. Round-Robin Partitioning

The block load balancing strategy assigns all short data records to initial processors and all longer data records to later processors, resulting in a severe imbalance in the distribution of the computation workload of the matching phase across



(a) Different Processors' Execution Times

(b) End-to-end Running Time

Fig. 4: Comparison of Various Load Balancing Strategies

various processors. Please, refer to Figure 4 for an example of this imbalance. This example was generated while running an index sharing based parallel $APSS$ algorithm using block load balancing strategy for the Orkut dataset.

Work imbalance induced by the block load balancing strategy can be reduced by assigning data records to each processor in a Round-Robin fashion. If there are P processors, then any consecutive P data records in the dataset are assigned to a different processor by the Round-Robin load balancing strategy. Please, refer to Figure 4 for an example of the performance improvement achieved in evenly distributing the computation workload of the matching phase by the Round-Robin strategy over the block strategy.

C. Dynamic Partitioning

Dynamic partitioning strategy aims at maximizing the processor utilization efficiency by dynamically assigning a small batch of data records to a processor as soon as the corresponding processor finishes the previous batch. As a result, all processors are expected to finish their computation almost the same time.

Contrary to the general experience in parallel computing, dynamic load balancing strategy performs only marginally better than the Round-Robin load balancing strategy which is a static strategy. The specific sort order of data records is the reason for the exceptionally good performance of Round-Robin strategy.

VI. Index Sharing Performance Evaluation

We performed experiments on four real-world million record datasets: Medline, Flickr, LiveJournal, and Orkut. Medline dataset comes from the scientific literature collaboration information in Medline indexed papers, while the rest come from popular online social networks: Flickr, LiveJournal, and Orkut. These datasets represent a variety of large-scale web-based applications like digital libraries and online social networks that we are primarily interested in. More detailed description these datasets is available in [19].

The distribution of the vector sizes in these datasets is the power law distribution [20], [6], [3]. These datasets are high dimensional and sparse (please, refer to Table I). The ratio of the average vector size to the total number of dimensions is less than 10^{-4}. All these characteristics are common across datasets generated and used by many large-scale web based applications [4], [3]. Therefore, we expect our index sharing technique to be relevant to other similar datasets as well.

We performed experiments for both the cosine similarity and the Tanimoto coefficient measures. Results for both similarity measures are quite similar. We present results only for cosine similarity for the sake of brevity. The results presented here are an aggregate of experiments preformed by varying the similarity threshold value from 1.0 to 0.5 in decrements of 0.1. The time spent for preprocessing and indexing is negligible as compared to the time spent for the matching phase. In all our experiments, we consider the time required only for the matching phase.

All of our implementations are for shared memory environment and coded in C++. We implement parallelization using the POSIX Pthreads library [21]. We used the standard template library for most of the data structures. We used the dense hash map class from Google^{TM} for the hash-based partial score accumulation [22]. The code was compiled using the GNU gcc 4.2.4 compiler with $-O3$ option for optimization. The experiments were performed on multi-processor as well as multi-core shared memory computers, each with eight processing elements. The code, the datasets, and additional experimental results are available for download on the Web [23].

TABLE I: Datasets Used

Dataset	Number of Records	Total Non-zero Components	Average Size
Medline	1565145	18722422	11.96
Flickr	1441433	22613976	15.68
LiveJournal	4598703	77402652	16.83
Orkut	2997376	223534153	74.57

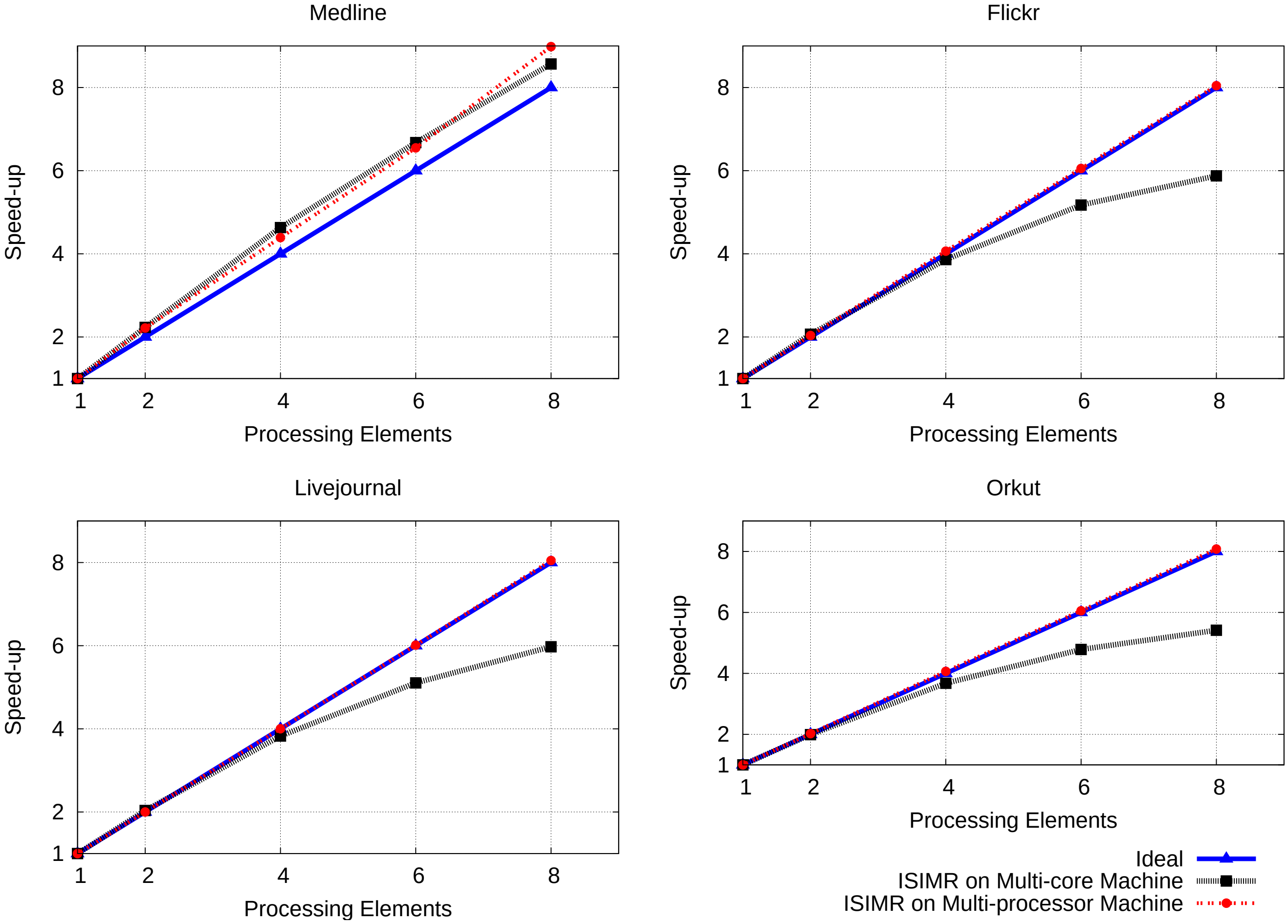


Fig. 5: Speed-up Over APT Algorithm vs. Number of Processing Elements for Index Sharing Technique

As described in Section I, motivation for parallelizing $APSS$ is to create a solution that scales with the number of processing elements as well as with the size of datasets. Therefore, we evaluate the performance of the index sharing technique based on scalability with respect to the number of processing elements and to the number of data records. The index sharing technique achieves ideal performance for both metrics.

For multi-processor environment, index sharing achieves ideal strong scaling behavior, i.e. linear speed-up over the APT algorithm (please, refer to Figures 5. The performance of the index sharing technique degrades in the multi-core environment due to *cache thrashing* and *memory bandwidth* limitation. In our experiments, the size of the datasets and of the inverted index range from few hundred megabytes to multiple gigabytes. $APSS$ algorithms access the inverted index and the dataset randomly, resulting in cache thrashing. Multi-core environment has multiple processing cores, but they still share the bus connection to the shared memory. When more cores start competing for memory access, the index sharing technique performance degrades. Thrashing effect is more visible for larger datasets like Orkut, while linear speed-up is maintained for smaller datasets, like Medline.

Scalability of index sharing with respect to variations in dataset sizes is plotted in Figure 6. The performance of the index sharing technique remains consistent. This result suggests that the index sharing technique will likely scale well with other large datasets.

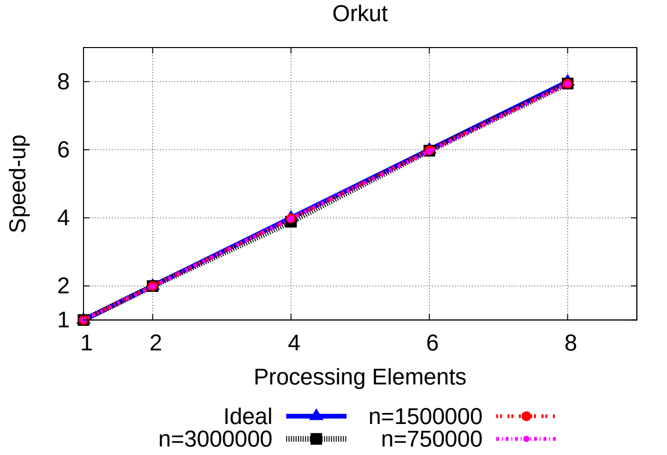


Fig. 6: Comparison of Speed-up of Index Sharing Over APT Algorithm for Different Dataset Sizes (n) in Multi-processor Environment

VII. Conclusion and Future Work

We presented a scalable, parallel solution for the $APSS$ problem based on the index sharing technique. The index sharing technique based parallel $APSS$ algorithm achieves ideal strong scaling performance and this performance remains consistent with variations in the dataset sizes. The work presented in this paper demonstrates that $APSS$ can be performed over large datasets in a reasonable time using parallelization.

Acknowledgment

This work is performed as part of the Scientific Data Management Center (http://sdmcenter.lbl.gov) under the Department of Energy's Scientific Discovery through Advanced Computing program (http://www.scidac.org). Oak Ridge National Laboratory is managed by UT-Battelle for the LLC U.S. D.O.E. under contract no. DEAC05-00OR22725.

References

[1] Sarawagi, S., and Kirpal, A. Efficient set joins on similarity predicates. In *SIGMOD '04: Proceedings of the 2004 ACM SIGMOD international conference on Management of data, Paris, France*, pp. 743–754.

[2] Arasu, A., Ganti, V., and Kaushik, R. Efficient exact set-similarity joins. In *VLDB '06: Proceedings of the 32nd international conference on Very large data bases, Seoul, Korea*, VLDB Endowment, pp. 918–929.

[3] Bayardo, R. J., Ma, Y., and Srikant, R. Scaling up all pairs similarity search. In *WWW '07: Proceeding of the 16th international conference on World Wide Web, Banff, Alberta, Canada*, pp. 131–140.

[4] Xiao, C., Wang, W., Lin, X., and Yu, J. X. Efficient similarity joins for near duplicate detection. In *WWW '08: Proceeding of the 17th international conference on World Wide Web, Beijing, China*, pp. 131–140.

[5] Xiao, C., Wang, W., and Lin, X. Ed-join: an efficient algorithm for similarity joins with edit distance constraints. In *Proc. VLDB Endow.*, vol. 1, VLDB Endowment, pp. 933–944.

[6] Awekar, A., and Samatova, N. F. Fast matching for all pairs similarity search. In *WI-IAT '09: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Milan, Italy.*, pp. 295–300.

[7] Geer, D. Industry trends: Chip makers turn to multicore processors. *Computer 38*, 5 (2005), pp. 11–13.

[8] Agrawal, R., and Shafer, J. C. Parallel mining of association rules. *IEEE Trans. on Knowl. and Data Eng. 8*, 6 (1996), pp. 962–969.

[9] Olman, V., Mao, F., Wu, H., and Xu, Y. Parallel clustering algorithm for large data sets with applications in bioinformatics. *IEEE/ACM Trans. Comput. Biol. Bioinformatics 6*, 2 (2009), pp. 344–352.

[10] Barua, S., and Alhajj, R. High performance computing for spatial outliers detection using parallel wavelet transform. *Intell. Data Anal. 11*, 6 (2007), pp. 707–730.

[11] Dean, J. Challenges in building large-scale information retrieval systems: invited talk. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining, New York, NY, USA*, pp. 1–1.

[12] Kumar, R., Novak, J., and Tomkins, A. Structure and evolution of online social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, Philadelphia, PA, USA*, pp. 611–617.

[13] Mislove, A., Koppula, H. S., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. Growth of the flickr social network. In *WOSP '08: Proceedings of the first workshop on Online social networks* (New York, NY, USA, 2008), ACM, pp. 25–30.

[14] Cheng, X., Dale, C., and Liu, J. Statistics and social network of youtube videos. In *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*, pp. 229–238.

[15] Breimyer, P., Kora, G., Hendrix, W., and Samatova, N. F. pr: Lightweight, easy-to-use middleware to plugin parallel analytical computing with r. In *IKE '09: Proceedings of the International Conference on Information and Knowledge Engineering, Las Vegas, Nevada, USA* (2009), pp. 667-673.

[16] Charikar, M. S. Similarity estimation techniques from rounding algorithms. In *STOC '02: Proceedings of the thiry-fourth annual ACM symposium on Theory of computing, Montreal, Quebec, Canada* (2002), pp. 380–388.

[17] Fagin, R., Kumar, R., and Sivakumar, D. Efficient similarity search and classification via rank aggregation. In *SIGMOD '03: Proceedings of the ACM SIGMOD international conference on Management of data, San Diego, California* (2003), pp. 301–312.

[18] Broder, A. Z., Glassman, S. C., Manasse, M. S., and Zweig, G. Syntactic clustering of the web. *Comput. Netw. ISDN Syst. 29*, 8-13 (1997), 1157–1166.

[19] Awekar, A., Samatova, N. F., and Breimyer, P. Incremental all pairs similarity search for varying similarity thresholds. In *SNAKDD '09: Workshop on Social Network Mining and Analysis Held in Conjunction with KDD '09, Paris, France* (2009), ACM.

[20] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. Measurement and analysis of online social networks. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, San Diego, California, USA* (2007), pp. 29–42.

[21] Lewis, B., and Berg, D. J. *Multithreaded programming with Pthreads*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1998.

[22] Google dense hash map library : `code.google.com/p/google-sparsehash/`.

[23] Code and data sets for our algorithms : `www4.ncsu.edu/~acawekar/ike10/`.

Embedded DMAIC Methodology into Financial Knowledge Management System

Yi-Chuan Lu[1], Hilary Cheng[2], Calvin Sheu[3]
[1]Department of Information Management, Yuan Ze University, Chung-Li 320, Taiwan
[2]College of Management, Yuan Ze University, Chung-Li 320, Taiwan
[3]Department of Information Management, Tungnan University, New Taipei City 222, Taiwan

Abstract - *We present a "define, measure, analyze, improve, and control" (DMAIC) methodology to model the knowledge discovery process for BI applications, specifically in statistical analysis and data mining. We implemented the DMAIC methodology embedded into the existed Financial Knowledge Management System (FKMS), which can be used as an effective means to dramatically improve knowledge product and process quality. It is meant for the achievement of total financial industry power users' satisfaction by producing near defect-free financial knowledge. We have emphasized the knowledge engineering perspective for developing knowledge sets into historic reference decision with regard to formed knowledge management and sharing architecture in the knowledge discovery process in addition to the importance of leveraging structures and DMAIC methodology in design as well as implementation. The resulting knowledge from each experiment defined as a knowledge set consisting of strings of data, model, parameters, and reports are stored, shared, disseminated, and thus made helpful to support decision making. We finally illustrate the above claims with a process of applying data mining techniques to support corporate bonds classification.*

Keywords: DMAIC, financial knowledge management system

1 Introduction

One of the biggest challenges that most security investment institutions experienced was the lack of an intelligent data mining system to support investment researches decisions. The problems their system encountered included the following (Cheng, Lu, Wang, & Sheu, 2004): Though data for financial applications are simple data, the data typically include time series information, and the relationships among the financial instruments are complex. For example, consider a derivative security objects. The derivative security object often shares underlying securities with other derivatives. Underlying securities can come from many classes of instruments, from a simple currency to an interest rate swap to a hedge. As the securities become more complex, the data management and knowledge discovery problems become more difficult. Consider a security portfolio. The portfolio construction is a process of quantitative analysis over massive amounts of data, and the Data Cube and ad-hoc analysis techniques are invisible solutions to support this process. We present the concept of FKMS, which is a prototype of a KM environment specifically for financial research purposes. The environment generates groups of knowledge sets with strings of data, models, parameters, and reports and tracking of the lifecycle of knowledge set creation among build, use, maintenance, and desertion. The ontology design of knowledge management and knowledge sharing is presented. Finally, a realization of decision support and knowledge sharing processes is illustrated. With FKMS, knowledge workers can freely extract sets of financial and economic data, analyze data with different decision support modules, rerun experiments with different sets of parameters, and finally disseminate value-added information (knowledge) through middleware or the Internet to remote clients. Not to mention that the knowledge generated is being collected, classified, and shared with colleagues, and thus it is well archived into corporate business intelligence databank. The remainder of this paper proceeds as follows. Section 2 introduces a perspective of knowledge engineering in the knowledge discovery process. Section 3 presents the DMAIC methodology. Finally, in Section 4 we conclude this paper.

2 The Perspective of Knowledge Engineering in the Knowledge Discovery Process

The development of knowledge representation is considered complex, and both the domain knowledge defined and the implemented feasible are subject to considerable uncertainty. The tendency exists therefore to focus knowledge discovery process modeling development from the knowledge management and sharing architecture to the domain knowledge form of enterprise production, which is logically described as the knowledge factory. As mentioned in Lu and Cheng (2003), a successful knowledge management system enhances the way people work together and enables knowledge workers and partners to share information easily so they can build on each other's ideas and work more effectively and efficiently. The goal is to gather company proprietary knowledge to come up with the best decision making or to quickly seize the initiative with innovative ideas.

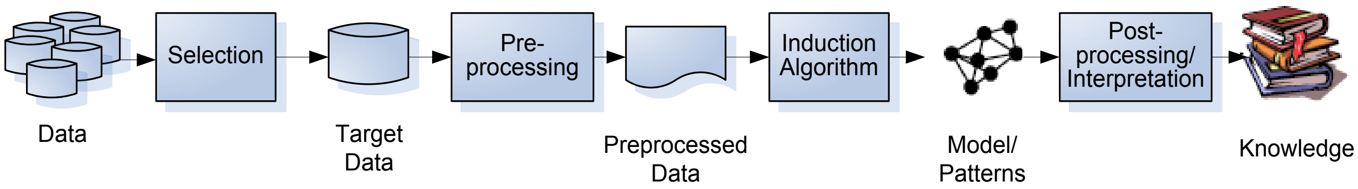


Fig. 1 The knowledge discovery process (adapted from Fayyad et. al. 1996)

To optimize the flow of information and interaction among knowledge workers so that the company can always make better trading or investment decisions, the specific group of data and modeling results should be administered and properly shared. Future knowledge can be generated by capturing existing (shared) knowledge via filtering, storing, retrieving, and disseminating explicit knowledge and by creating and testing new knowledge (Nemati, Steiger, Iyer, & Herschel, 2002).

According to the knowledge discovery process defined by Fayyad et al. (1996), as shown Fig. 1, the flow of data and shared information proceeds as follows: users select data with different criterion to perform quant analysis. The data retrieved will automatically be saved as data cubes. Each data cube records the time this file has been created, the screening rules, the data item names and types, etc., so that colleagues can easily share and exchange their knowledge and value-added information with others. In the meantime, the company will keep good management of each individual worker's knowledge and so will be ready for good customer relationship management. However, this traditional knowledge discovery process is based on the prospect of problem solving (McDermott, 1988; Decker, Daniel, Erdmann, & Studer, 1997) in database. These knowledge engineering activities are quite unique in contrast to those from other types of processes, such as manufacturing. From the perspective of knowledge engineering, the knowledge discovery process is a transfer process (Hayes-Roth, Waterman, & Lenat, 1983) and a modeling process (Clancey, 1989; Morik, 1990). So, unlike a manufacturing process, a completed knowledge discovery process is non-repetitive. Each knowledge discovery instance invokes a different set of inputs and outputs. The higher levels of cognition increase the process variance. To further explain this, consider the following example (see Fig. 2)

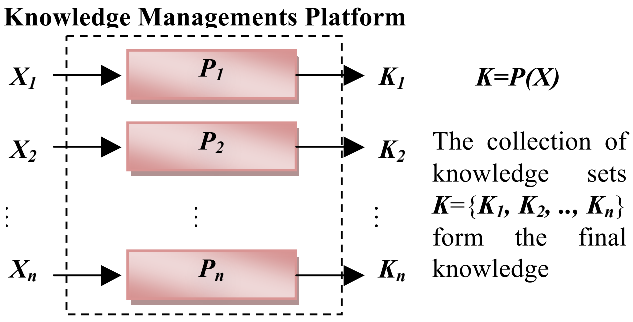

Knowledge Managements Platform
X_1
P_1
K_1
$K=P(X)$
X_2
P_2
K_2
The collection of knowledge sets $K=\{K_1, K_2, .., K_n\}$ form the final knowledge
X_n
P_n
K_n

Fig. 2 Each Knowledge Discovery Instance Invokes a Different Set of I/O.

Based on the input set X={X1, X2, .., Xn}, a set of knowledge discovery processes is P ={P1, P2, ..., Pn} that was invoked by knowledge workers {A, B, C} and then produced knowledge sets K ={K1, K2, .., Kn}. So, the formula of knowledge sets is K=P(X). Much of the statistical analysis was done based on the assumption of a stable process. Due to the uniqueness of the knowledge discovery process, the collection of individual instances of the knowledge discovery processes is no longer stable. Since process capability represents the full range of normal process variations for a given characteristic, it makes little sense to speak of process capabilities in the knowledge discovery process.

3 DMAIC Methodology

We know the business process design and modeling technologies are being increasingly used by both traditional and newly formed ecommerce's enterprises in order to improve the quality and efficiency of their administrative and production processes. From a process modeling and automation perspective, this has several implications; for example, the business processes should be correctly designed, their execution should be supported by a system platform that can fit the knowledge worker's requirements, and the process resources should be able to carry out their work in a timely change. We applying DMAIC methodology to optimize knowledge management and sharing architecture that supports business and IT users in managing process implementation quality. FKMS can support full modeling tool suite in the business process, since it is based on the application of business intelligence techniques to business processes. In fact, FKMSTM store many types of incident that occur during process executions, including the start and completion time of each stage, its input and output data, the resource that executed it, and any failure that occurred during stages or process execution. By process records into a warehouse and by analyzing them with business intelligence technologies, we can extract knowledge about the circumstances in which high- or low-quality executions occurred in the past, and use this information to explain why they occurred as well as predict potential problems in running processes. The architecture of the FKMSTM is a layered structure as shown in Fig. 3.

The DMAIC methodology provides the financial industry with the opportunity to achieve quality and cost-effective development measures that will not only save companies substantial financial resources in rework and waste, but will also ultimately ensure financial knowledge continuity.

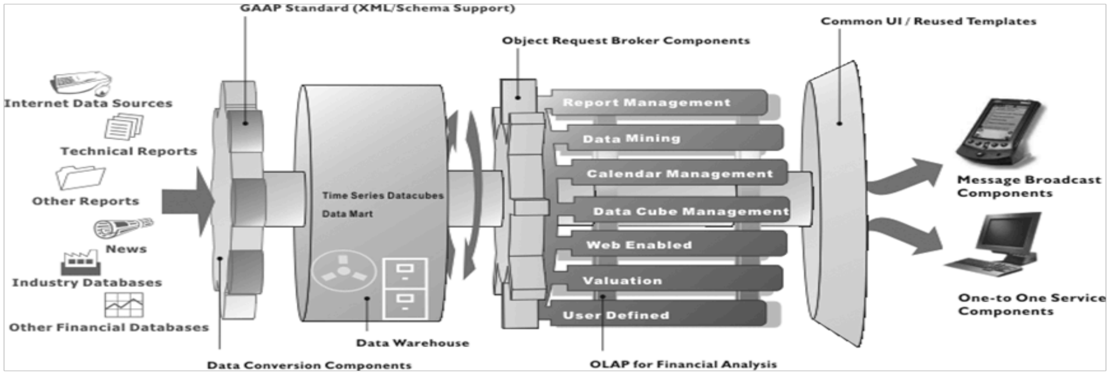

GAAP Standard (XML/Schema Support)
Object Request Broker Components
Common UI / Reused Templates
Internet Data Sources
Technical Reports
Other Reports
News
Industry Databases
Other Financial Databases
Time Series Datacubes
Data Mart
Report Management
Data Mining
Calendar Management
Data Cube Management
Web Enabled
Valuation
User Defined
Message Broadcast Components
One-to One Service Components
Data Warehouse
Data Conversion Components
OLAP for Financial Analysis

Fig. 3 The System Architecture of FKMS

Applying DMAIC methodology in the development of the knowledge discovery process requires a fundamental cultural change as well as management commitment.

We refined the full lifecycle of a knowledge set of knowledge discovery results consisting of simple four basic phases: building, use, maintenance/update, and desertion, as depicted in Fig 4. During the first phase, the knowledge set must be designed and built with the discipline of DMAIC domain expertise based on FKMS to ensure that the knowledge that is delivered to production meets the domain expert's needs. Once in production, the knowledge set must be maintained and improved over time to ensure a long and opulent lifecycle of the business's financial investment.

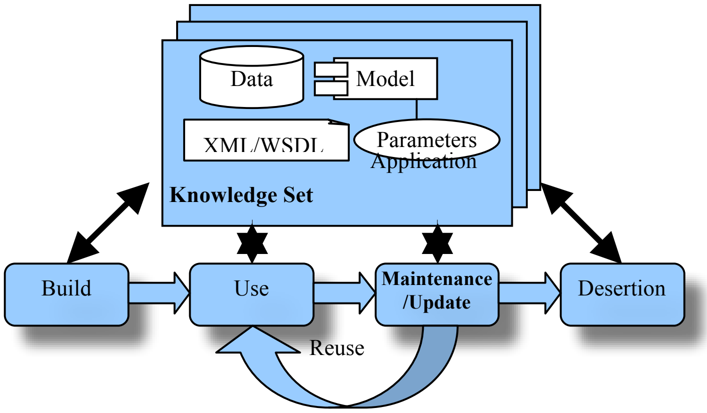


Fig. 4 The Lifecycle of Knowledge Set Creation

Since DMAIC was developed specifically to improve the existing knowledge discovery processes, it seems only natural to apply the knowledge engineering perspective roadmaps into an extended roadmap (shown in Fig 5.) that can be used to ensure this DMAIC methodology quality throughout the lifecycle of the knowledge set creation. No matter which modeling framework is chosen, the DMAIC methodology can always be incorporated, as outlined below.

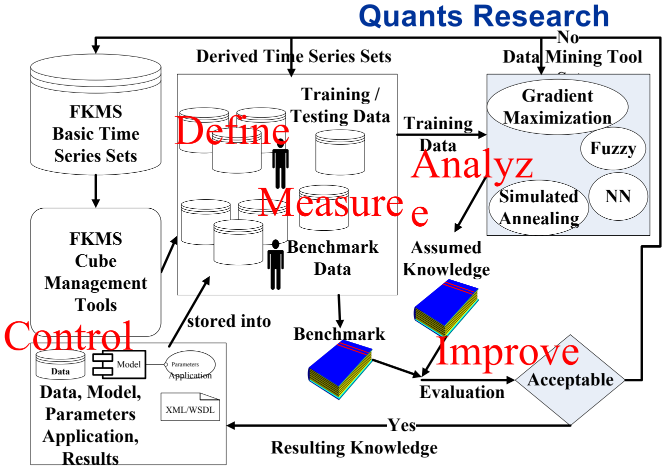


Fig. 5 Embedded DMAIC Methodology into FKMS

Define. The program of knowledge discovery should be defined as early as the knowledge engineering project-planning phase. The DMAIC first asks knowledge workers to define their core processes following two steps: 1) Define knowledge project goals by domain expert. 2) Build knowledge sets, recognition of data cube's factors, data cube gathering for training and testing, conceptual application for domain customers, model selection about mathematical models like neural networks (NNs) and fuzzy logic (FL), and designing model. Finally, the result will be created. To design the knowledge sets, we have to clearly conceptualize recognition of the data cube's factors. To further specify the problem, a set of measurable requirements was gathered. For knowledge sets, these requirements are usually of the following types:

Data cube's factors. For example, 1) relation to the operation of the actual data cube on corporate bond clusters, 2) requirements defining lower bounds on the data cube.

Model's parameters. 1) Integration with certain models' data mining techniques, like neural networks (NNs) and fuzzy logic (FL). 2) Set parameters from reliability metrics.

This is the phase in which one gathers the data and model of resources on the importance of knowledge quality measures and expectations. The progress of the knowledge discovery program should be tracked as part of the project schedule tracking activities. The milestones to be achieved by knowledge discovery program should be reflected as project milestones. A high level process map should also be outlined at this stage.

Measure. In this phase, the process is measured to determine current performance and to quantify the problem. One must answer the following questions: 1) What should be monitored? 2) How should the benchmark be monitored? 3) What is the plan to describe signals? Who will own the plan? 4) How will the monitoring system be maintained? Who will maintain it? This process defines what constitutes a mathematical model weight, the training time period, and the measurement unit for benchmark. Then a data collection plan should be drawn and people assigned to be responsible for identifying the needed data. This plan should be carried out during the whole knowledge development lifecycle for measuring key process deliverables.

Analyze. In the analyze phase, the system provides a signal when there is suggestion by the domain expert or other statistical proof that there has been a significant change in one of the leading indicator values. Then data collected are analyzed in addition to the model's parameters that must be tuned, monitored, and maintained to ensure the classifier's optimal performance and that the root causes of benchmark are determined. By doing so, mathematical model weights for improvement are identified. In this step, each statistical tool can be used to validate the root causes of the benchmark to identify data sources of process variations and to determine the most promising alternatives.

Improve. During the improve step of the DMAIC methodology, ideas and solutions are put to work. In this phase, the process provided maintenance and updating to improve knowledge sets' quality by optimization of mathematical model weights, use of design of experiments, and identification of possible solutions. The target process is improved by designing creative solutions to rectify and prevent mathematical model weights. There must be checks to ensure that the desired results are being achieved. Some experiments and trials may be required in order to find the best model weights.

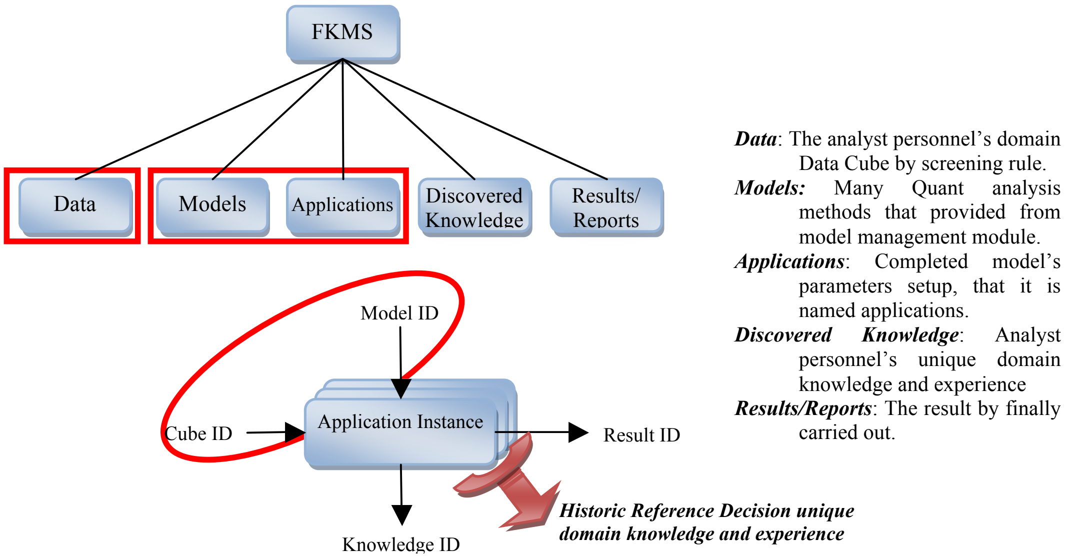
Control. During this phase, future process performance is controlled. This is done through performance tracking mechanisms and measurements based on FKMS platform in order to ensure, at a minimum, that the gains made in the knowledge engineering project are not lost over a period of time. The improvements are also institutionalized through the modification of systems and structures. This effort includes defining and validating in knowledge sets. Control mechanisms, development of standards and procedures, and verification of benefit and cost savings, etc., are determined. With this, the DMAIC methodology really starts to create returns; ideas and projects in one part of the organization are translated in a very rapid fashion into implementation in another part of the organization.

3.1 Ontology design based on DMAIC methodology.

From novel knowledge sets to formed historic reference decision. In this research, a multiplicity of methods has been conceived to support the knowledge discovery process modeling. The methods can be distinguished according to those aspects of the knowledge discovery processes modeling, which they aim to support, according to their chosen perspective on the problem (e.g., data cube-oriented, application-oriented, model-oriented, and result-oriented). As a result, attempts are being made to create a methodology (theory of methods) for the development methods. In the aim of categorizing and evaluating the methods, there is another set of reasons for considering knowledge management system design methodologies. These reasons result from the fact that, in general complex development, projects involve several partners who may implement diverse development methods and whose working results overlap. In this situation, only a conceptual framework which allows the categorization, of the various methods, and consequently their conformities and discrepancies, can generate mutual understanding. The fact that such a conceptual framework can, and even must, also lead to uniformity in the use of methods is, of course, also relevant. Architecture is generally understood as the art of construction. In practice, as one of the key problems of computer environment for knowledge schemes, the effective establishment and management of knowledge base systems is still troubled by the sharing and reuse of domain knowledge bodies in computational form at present.

So, what is domain knowledge in quant analysis? The answer to this question should begin with a description of knowledge itself. In this research, for supported DMAIC methodology that knowledge can be described as a set of knowledge that describes various properties and behaviors within a domain. The relation within knowledge management function is established through the knowledge set, which consists of five key parts: cube ID, model ID, results ID, application ID, and knowledge ID. We aim to define the knowledge scheme that decomposes complex knowledge sets into several simple IDs and the retrieval of knowledge of related records in FKMS by means of establishing relations within IDs. As a result, the issues of more correspondence relations existing in FKMS architecture are finally solved, as shown in Fig. 6. In the framework of this research, an extreme position on domain knowledge would be, "Any knowledge set that shapes the knowledge discovery process and experiences recorded."

So, that which we defined with the domain experts, called upon from a committee of experienced financial staffs to conduct a clear review of a data cube of cases from our database and arrive at a unanimity decision for each case, is referred to as the standard of historic reference decision. The resulting knowledge sets containing the entire standard of historic reference decisions was used in the measurement system analysis of the original data cube. The knowledge sets



Data: The analyst personnel's domain Data Cube by screening rule.
Models: Many Quant analysis methods that provided from model management module.
Applications: Completed model's parameters setup, that it is named applications.
Discovered Knowledge: Analyst personnel's unique domain knowledge and experience
Results/Reports: The result by finally carried out.

Fig.6. The Correspondence Relations of Knowledge Sets.

of standard of historic reference decisions also served as the benchmark decisions against which the decisions of both the knowledge sets and the current process of DMAIC could be independently compared. Without the knowledge sets of standard of historic reference decisions, there would have been scarcely able to measure the accuracy of the final knowledge set. The knowledge sets of standard of historic reference decisions were used during the building of the knowledge sets as well.

As started above, the DMAIC methodology can be a new perspective of knowledge engineering to modeling knowledge discovery process. In this analytic conceptual of the knowledge scheme, its domain knowledge base serves as a powerful tool to 1) manage data in enterprise, 2) formalize data cubes about behaviors and rules, 3) supply model management for separate application and data, 4) classify applications more effectively, 5) generate knowledge report to supply aid for users to make decisions, and 6) standardize historic reference decision to provide practical experience in the next round of the knowledge discovery process.

The rationality of the knowledge set structure directly affects creativity of knowledge schemes as well as its effective in enterprise intelligence. In addition, the modeling process can solve the efficiency issue and play a decisive role in deciding whether schemes which are finally generated could realize computer-aided innovative design in the real domain knowledge. We then proposed a basic idea in FKMS architecture that will support this DMAIC process approach and then advance domain knowledge of reusable ontology in a standard formalism, which each knowledge worker was supposed to adopt.

Meta taxonomy carried out through knowledge sets. To define the scopes of Quant research within domain knowledge that includes the following:

Data sets management. Data sets are registered and classified according to Meta Taxonomy that is named "Meta data about data sets." We have to define user defined data sets and derived data sets.

Applications management. Applications, models, and model associated parameters are registered and classified according to Meta Taxonomy. We call this "Meta data about applications," "Meta data about models," and "Meta data about model associated parameters." After defining (and constructing) a data set for back testing, data mining models (and applications) are selected from the mining tools set. Templates/reports management. The data mining results (discovered knowledge) and reports are also registered and classified according to Meta Taxonomy.

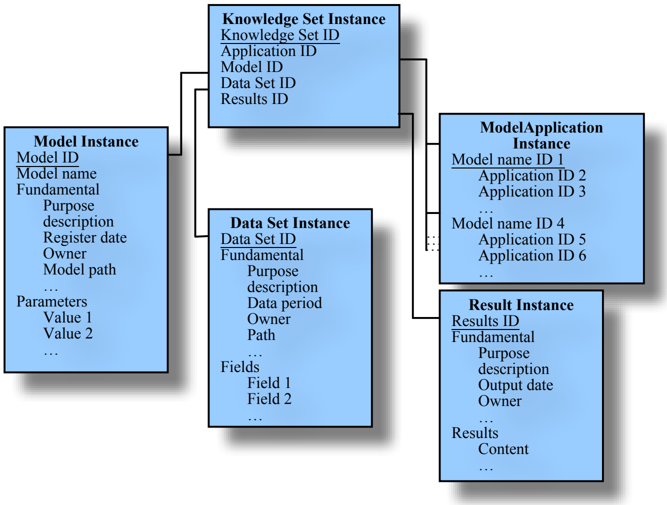
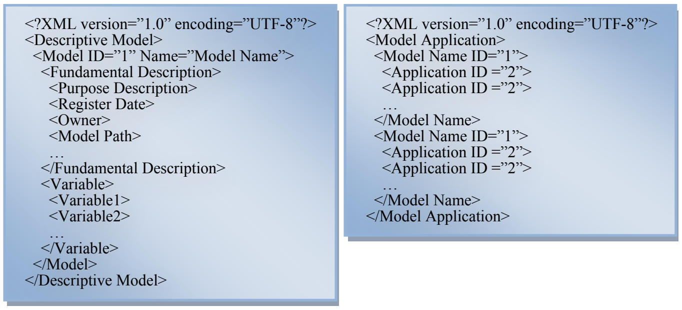


Fig. 7. The Scheme of Knowledge Sets.

Fig. 7 shows the overall scheme of the knowledge sets and their relationships. Underlined attributes denote primary keys, while links among tables denote foreign key constraints. For clarity of presentation, in the figure we have slightly simplified the structure of the tables. In addition to instance execution tables, authorized users can access definition tables that describe the processes, nodes, services, and resources also defined in the FKMS.

We know that XML possessed characters of Metadata that TagName and TagValue have used to define Metadata about data. Users can use different elements and attributes to define the elaboration of data that users and applications can access easily. So, the XML document structure is a description result of Metadata. We also regard XML as the storage tools of data in addition to regarding the XML data model as the Metadata form of models. We also illustrated our knowledge sets scheme design with the description of two mainly XML documentations.

The content of DescriptiveModel.xml filled in while registering in the model. The content is divided into two parts to describe mainly just a single model (see left in Fig. 8). The content of ModelApplication.xml is a relevant description that records between mode and application. Every tag will be named with the model or application name and also have another ID attribute added as identification. The level 2 recording model and level 3 recording derivational application of model are demonstrated (see right in Fig. 8).



```
<?XML version="1.0" encoding="UTF-8"?>
<Descriptive Model>
  <Model ID="1" Name="Model Name">
    <Fundamental Description>
      <Purpose Description>
      <Register Date>
      <Owner>
      <Model Path>
      …
    </Fundamental Description>
    <Variable>
      <Variable1>
      <Variable2>
      …
    </Variable>
  </Model>
</Descriptive Model>
```

```
<?XML version="1.0" encoding="UTF-8"?>
<Model Application>
  <Model Name ID="1">
    <Application ID ="2">
    <Application ID ="2">
    …
  </Model Name>
  <Model Name ID="1">
    <Application ID ="2">
    <Application ID ="2">
    …
  </Model Name>
</Model Application>
```

Fig. 8. DescriptiveModel.xml and ModelApplication.xml.

Ontology of knowledge abstraction. This research addresses the concept of FKMS, which is a prototype of KM environment specifically used for financial research purposes. The environment generates groups of knowledge sets with strings of data, models, applications, and reports. Based on the above taxonomy of enterprise metadata management, the data administrator can define enterprise metadata to align with corporation structure and operations based on the business requirements. Enterprise metadata management is the brain of the data conversion layer. For the purpose of being expandable, scalable, and portable, the ontology of the knowledge abstraction was represented with XBRL schema. Also, the design of enterprise metadata management is flexible to accommodate corporate structure and business changes if a corporate re-engineering is needed.

4 Conclusions

Domain experts need a flexible system environment where they can freely select data and models and running different settings of parameters for decision support purposes. Knowledge sets of each research experiment containing data, models, parameters, and results essentially provide great value for business intelligence generation. In this study, we propose a DMAIC methodology to optimize knowledge management and sharing architecture in financial service institutions. Then the integration of decision support and knowledge management processes is crucial for enterprises to create its niche business intelligence and to maintain global competitive advantages.

We implement the DMAIC methodology embedded into the FKMS, which can convert data from various sources into the data warehouse and retrieve data cubes in response to different knowledge workers' request for report generating or for running decision support applications.

We have emphasized the knowledge engineering perspective for developing knowledge sets into the historic reference decision to form knowledge management and sharing architecture in the knowledge discovery process and the importance of leveraging structures and DMAIC methodology in design as well as implementation. We have demonstrated how this can be achieved by describing the development of knowledge sets embedded in the FKMS. In conclusion, we would like to highlight the key elements of the DMAIC roadmap as they apply to the development and implementation of any knowledge set for decision support.

5 References

[1] Bergholz, A. (2000). Extending your markup: An XML tutorial. IEEE Internet Computing , 4 (4), pp. 74-79.

[2] Bertino, E., & Catania, B. (2001). Integrating XML and databases. IEEE Computer , 5 (4), pp. 84-88.

[3] Bolloju, N., Khalifa, M., & Turban, E. (2002). Integrating Knowledge Management into Enterprise Environments for the Next Generation Decision Support. Decision Support Systems , 33 (2), pp. 163-176.

[4] Bourret, R., Bornhovd, C., & Bornhovd, A. (2000). A generic load/extract utility for data transfer between XML documents and relational databases. WECWIS-2000: Proc. the Second International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems, (pp. 134-143).

[5] Cheng, H., Lu, Y., Wang, W., & Sheu, C. (2004). An Ontology-based Data Mining Approach in a Financial Knowledge Management System. The Third Workshop on e-Business (WEB 2004), (p. 166). Washington DC, USA.

[6] Clancey, W. (1989). The Knowledge Level Reinterpreted: Modeling How Systems Interact. Machine Learning , 4, pp. 285-291.

[7] Decker, S., Daniel, M., Erdmann, M., & Studer, R. (1997). An Enterprise Reference Scheme for Integrating Model-based Knowledge Engineering and Enterprise Modeling. In E. Plaza, & R. Benjamins (Ed.), Knowledge Acquisition, Modeling, and Management, 10th European Workshop (EKAW'97), Lecture Notes in Artificial Intelligence 1319.

[8] Dutta, S., & Shekhar, S. (1988). Bond Rating: A Non-Conservative Application of Neural Networks. Proceedings of ICNN-88, 2, pp. 443-450.

[9] Fayyad, U., Shapiro, G., & Smyth, P. (1996). From data Mining to Knowledge Discovery: An Overview. In Advances in Knowledge Discovery and Data Mining (pp. 1-34). MA AAAI/ MIT Press.

[10] Hayes-Roth, F., Waterman, D., & Lenat, D. (1983). Building Expert Systems. New York: Addison-Wesley.

[11] Lu, Y., & Cheng, H. (2003). Towards Automated Optimal Equity Portfolios Discovery in a Financial Knowledge Management System. In Computational Intelligence in Economics and Finance (pp. 387-402). Springer-Verlag.

[12] McDermott, J. (1988). Preliminary Steps toward a Taxonomy of Problem-solving Methods. In S. Marcus, Automating Knowledge Acquisition for Experts Systems. Boston: Kluwer Academic Publisher.

[13] Morik, K. (1990). Underlying Assumptions of Knowledge Acquisition as a Process of Model Refinement. Knowledge Acquisition , 2 (1), pp. 21-49.

[14] Nemati, H., Steiger, D., Iyer, L., & Herschel, R. (2002). Knowledge Warehouse: an Architectural Integration of Knowledge Management, Decision Support, Artificial Intelligence and Data Warehousing. Decision Support Systems , 33 (2), pp. 143-161.

[15] Sammon, J. (1969). A Nonlinear Mapping for Data Structure Analysis. IEEE Trans. Computing , 18 (5), pp. 401-409.

DEA of Assurance Region Malmquist Index: An Illustration with International Tourist Hotels in Taiwan

Hilary Cheng[1], Yi-Chuan Lu[2], Jen-Tsung Chung[3]
[1]College of Management, Yuan Ze University, Chung-Li 320, Taiwan
[2]Department of Information Management, Yuan Ze University, Chung-Li 320, Taiwan
[3]Dept. of Computer Science and Information Engineering, Asia University, Taichung 41354, Taiwan

Abstract - *This study supplements Malmquist Index calculation of basic DEA model, and includes weights of input and output factors so as to more precisely measure the vertical productivity change of industries. The model we proposed revises the Assurance Region Malmquist Index (AR-MI) by combining importance of weights with Malmquist Index. Analytic Hierarchy Process is used to acquire the weights of input and output items, and empirical analysis is conducted on operational performance of international tourist hotels in 1998 - 2007 in Taiwan by the modified model. The results shows that: (1) long-term productivity of overall industry grows and the main reason is due to technical transformation instead of growth of technical efficiency; (2) returns to scale of hotel industry declines, which indicates the severe competition in the industry; (3) growth or decline of international tourist hotels is closely related to government promotion for industry development.*

Keywords: Assurance Region, Malmquist Index, DEA, Analytical Hierarchy Analysis, international tourist hotels

1 Introduction

In the tourism industry, tourist hotels are multi-functional places for lodging, shopping, and various social activities of tourists. According to statistics of the Tourism Bureau [1], Ministry of Transportation and Communications, in 2006, the average expenditure of each tourist in Taiwan in tourist hotels was 44.74% of daily consumption/per person. Therefore, service quality and operational performance of tourist hotels are the keys of tourism industry. Hotel industry in Taiwan can be divided into ordinary tourist hotels and international tourist hotels. In 1992-2006, the numbers of ordinary tourist hotels and rooms reduce from 42 hotels and 4,706 rooms to 29 hotels and 3,265 rooms by 31%; numbers of international tourist hotels and rooms increase from 47 hotels and15, 018 rooms to 60 hotels and 17,830 rooms by 18.72%. It indicates that larger scale international tourist hotels are the mainstream in the market. Thus, performance management of international tourist hotels is an important researcher topic.

Past studies have indicated the importance of weights of input and output. Sun and Lu [2] included influence of weights in calculation of model, but the weights of input and output variables were fixed and possible relative relations among weight ratios might be neglected. Therefore, this study attempts to integrate the importance weights of input and output items in research of Färe et al. [3] on Malmquist Index of year-to-year productivity change in order to complete research method.

The subjects of this study were 34 international tourist hotels in Taiwan. Data were sourced from the "Operational analysis and report of international tourist hotels in Taiwan" of 1997-2006 edited by Tourism Bureau, Ministry of Transportation and Communications. The research purposes are below: (1) to expand the research field of DEA to serve as a reference to future studies; (2) to analyze the overall operational performance of international tourist hotels and long-term productivity change; (3) to provide references for hotel industry to gradually enhance performance.

The remainder of this paper is organized as follows: Section 2 describes the research method and modified model, data and variable selection; Section 3 discusses the empirical analysis. Finally, Section 4 gives conclusions and suggestions.

2 Research method

Due to limitation of the space, this study does not indicate the basic model of DEA and only suggest the related parts. For introduction of basic theory of DEA, please see Cooper, Seiford and Tone [30].

2.1 Assurance Region Malmquist Index

In original CCR model, weights of input and output items are acquired by the model regarding the optimized efficiency value of different DMUs. However, the weights are not based on importance of the variables. Thus, there is no proper and reasonable explanation in terms of economics or managerial implication.

Malmquist Index model proposed by Färe et al. [3] is only based on CCR and does not consider importance of input and output items. In practice, input and output items reveal different percentages of importance. By assessing the weights of input or output items, real efficiency measurement can be used.

Assurance Region (AR) is also based on CCR; however, this model includes upper and lower limit of weights of input and output variables. Thus, it can separate inefficient DMU from efficient DMU judged by CCR. Therefore, Assurance Region will more effectively indicate different performance degrees of DMU. In order to more precisely measure vertical productivity change of industry, based on Malmquist Index model, this study combines specific model concept and includes weight importance in Malmquist Index model to propose Assurance Region Malmquist Index (AR-MI).

2.1.1 Weight setting of input and output items

It is assumed that weights of input and output items are vi = (v1,…,vm) and ur = (u1,…,us), importance of input and output is shown in the following conditions:

$$l_{1,2} \le \frac{v_2}{v_1} \le u_{1,2} \qquad (1)$$

$$L_{1,2} \le \frac{u_2}{u_1} \le U_{1,2} \qquad (2)$$

In Eq. (1), l1,2 and u1,2 represent upper and lower limits of weight ratio v2/v1 of input items. L1,2 and U1,2 in Eq.(2) indicate upper and lower limits of weight ratio u2/u1 of output items

2.1.2 Malmquist Index calculation by restricted distance function

Färe et al. [3] combined Malmquist Index, modified Total Factor Productivity (TFP), defined by Caves, Christensen and Diewert [32] and calculated distance function proposed by Shepherd [33] through nonparametric techniques and geometric mean. They turned Total Factor Productivity change into product of Technical Efficiency Change (TECH) and Technology Change (TCH). Ray and Desli [34] further replaced fixed Returns to Scale by changed Returns to Scale. They not only calculated Technology Change of changed Returns to Scale, but also divided Technical Efficiency Change into Pure Technical Efficiency Change (PTECH) and Scale Efficiency Change (SECH). Meanings of the terms of input-oriented Malmquist Index are skipped in this paper due to conference presentation limit.

2.2 Selection of decision-making units

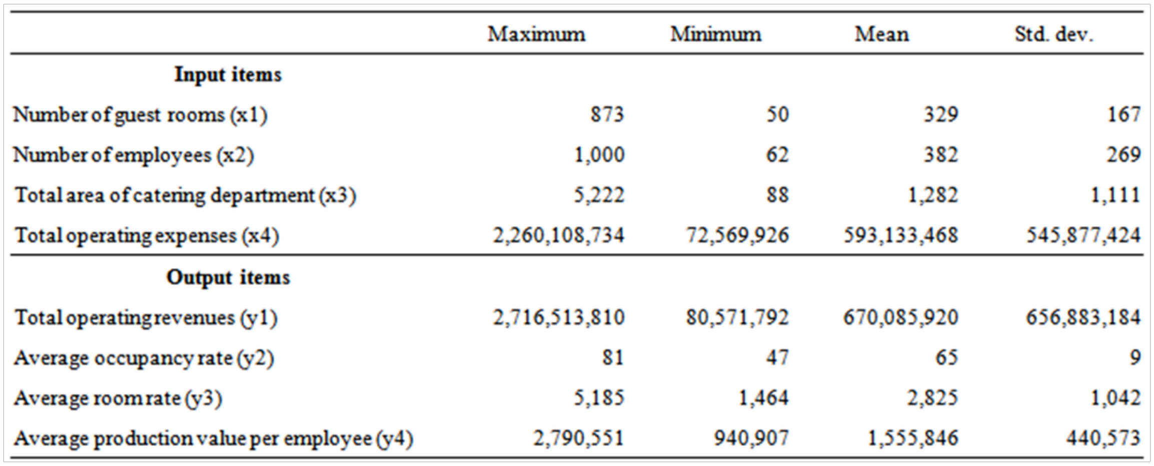
Source of research data is the “Analysis and report of operation in international tourist hotels of Taiwan (1997-2006)” edited by Tourism Bureau [1], Ministry of Transportation and Communications. In the report, the hotels which close out and are reorganized of the years are eliminated. Thus, data are complete and continuous. The subjects in study are 34 international tourist hotels which are listed in the report in consecutive ten years.

2.3 Variables selection

We summarized the variables used in the literature cited in Section 2.2 and obtained a total of 5 inputs (number of guest rooms, number of employees, total area of catering department, total operating expenses and catering expenses) and 6 outputs (total operating revenues, average occupancy rate, average room rate, average production value per employee, occupancy revenues, catering revenues). This research topic explores the overall operational performance; therefore, the input item “catering expenses” is part of the total operating expenses and can be eliminated. The output items “occupancy revenues” and “catering revenues” are both part of the total operating revenues, so these two items are also eliminated.

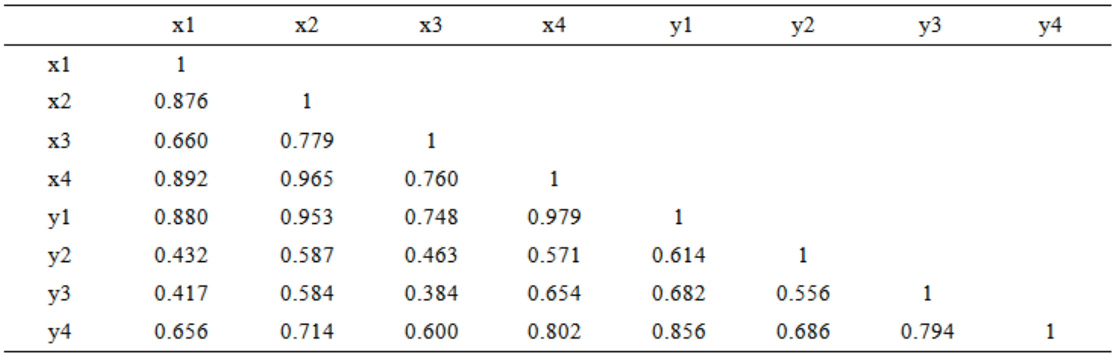
We selected a total of 34 DMUs and 8 input/output items, in line with Golany and Roll [35], who suggested that the number of DMUs should be at least twice the aggregation of input and output items. Table 1 presents descriptive statistics for our data set. To ensure the accurate representation of statistics, we calculated the arithmetic mean of all numbers of these hotels over the past 10 years. Table 2 shows the correlations that were obtained. A positive

Table 1 Descriptive statistics for the 34 hotels



	Maximum	Minimum	Mean	Std. dev.
Input items				
Number of guest rooms (x1)	873	50	329	167
Number of employees (x2)	1,000	62	382	269
Total area of catering department (x3)	5,222	88	1,282	1,111
Total operating expenses (x4)	2,260,108,734	72,569,926	593,133,468	545,877,424
Output items				
Total operating revenues (y1)	2,716,513,810	80,571,792	670,085,920	656,883,184
Average occupancy rate (y2)	81	47	65	9
Average room rate (y3)	5,185	1,464	2,825	1,042
Average production value per employee (y4)	2,790,551	940,907	1,555,846	440,573

Table 2 Correlation coefficients among inputs and outputs



	x1	x2	x3	x4	y1	y2	y3	y4
x1	1							
x2	0.876	1						
x3	0.660	0.779	1					
x4	0.892	0.965	0.760	1				
y1	0.880	0.953	0.748	0.979	1			
y2	0.432	0.587	0.463	0.571	0.614	1		
y3	0.417	0.584	0.384	0.654	0.682	0.556	1	
y4	0.656	0.714	0.600	0.802	0.856	0.686	0.794	1

correlation is observed between the input and output variables. In other words, an increase in some inputs will lead to an increase in some outputs. This is consistent with the hypothesis of constant returns to scale that has been used in this research

2.4 Weight

Weights in this study are acquired upon Analytic Hierarchy Process (AHP) proposed by Satty [36]. Integration of simplicity and flexibility of AHP and DEA will effectively result in relative importance weights among variables Cooper, Seiford and Tone [30]. Inference of relative weights can be upon experts' opinions in AHP questionnaire. Questionnaire survey is based on convenience sampling and field interview.

3 Empirical analysis

In this section, consistency test of AHP questionnaire and acquisition of weights of input and output items are conducted by Microsoft Office Excel 2007. The distance function of Assurance Region Malmquist Index is calculated by self-designed Marco program and "planning and solution" of Microsoft Office Excel 2007.

3.1 Weight calculation of input and output items

A total of 12 questionnaires were returned, including five hotels (Landis Hotel, Evergreen Hotel, Howard Hotel, United Hotel and Hotel Holiday Garden). Two questionnaires that did not pass the consistency ratio test were eliminated. In 10 valid questionnaires, post distribution is shown below: 2 assistant general managers, 2 managers of front office, 2 directors of human resource department, 2 directors of marketing planning department, 1 financial director and 1 director of counter department; as to distribution of educational level, there are 1 doctors and 9 bachelors; regarding ages, 5 subjects are 31—40 years old and 5 subjects are 41—50 years old; as to seniority in international tourist hotels, 2 subjects are 1—5 years, 2 subjects are 5—10 years, 4 subjects are 10—20 years and 2 subjects are above 20 years. Weight ratio of marginal input is shown in Table 3.

Table 3 Weight ratio of marginal input

Weight ratio	Lower bound	Upper bound
v2/v1	0.543	7.796
v3/v1	0.177	4.241
v4/v1	0.742	7.987
v3/v2	0.116	1.894
v4/v2	0.289	5.412
v4/v3	1.297	10.638

3.2 Analysis on year-to-year productivity change

In this section, four distance functions are obtained by Equations (1)-(4), empirical analysis is conducted on nine year-to-year productivity changes according to the operation of international tourist hotels in Taiwan in 1998-2007 by Malmquist Index deconstructed by Ray and Delsi [34].

3.2.1 Analysis on Technology Change

Regarding hotel industry, Technology Change is efficiency frontier change caused by increase of asset input or result input operation of research innovation. Technology Change values of hotels in different periods increase or decrease. Far Eastern Plaza Hotel (H16) grows for eight times in 9 periods. Sheraton (H10) and Royal Hotel (H11)grows for 7 times; Hotel Riverview Taipei (H04), Santos Hotel (H07), Howard Hotel (H12), Grand Formosa Regent (H14) and Evergreen Hotel (H24) grow for six times; growth of Hotel National (H22) and The Grand Hotel Kaohsiung (H30) is lower for three times in 9 periods; in 1998-1999, 1999-2000, 2001-2002, 2003-2004, 2004-2005 and 2006-2007, over half of hotels have Technology Change increased. In 2000-2001, 2002-2003 and 2005-2006, most of hotels have Technology Change decreased.

3.2.2 Analysis on technical efficiency change

Technical Efficiency consists of Pure Technical Efficiency and Scale Efficiency, including technical capacity and business scale of enhancement of hotel image and service quality which result in relative position change of DMU in production set. The changes are described below.

i. Technical efficiency change

As to Technical Efficiency Change of hotels in different periods, United Hotel (H09) and Ambassador Hotel Kaohsiung (H19) grow for six times in 9 periods. Royal Hotel (H11), Howard Hotel (H12), Sherwood Hotel (H15) and Far Eastern Plaza Hotel (H16) only have lower growth twice in 9 periods; in 1999-2000, 2000-2001, 2002-2003 and 2005-2006, over half of hotels have Technical Efficiency increase and in 1998-1999, 2001-2002, 2003-2004, 2004-2005 and 2006-2007, most of hotels have Technical Efficiency decreased.

ii. Pure technical efficiency change

Factors of Pure Technical Efficiency Change can include overall marketing of hotels, managerial process, personnel quality, resource fit and broadening sources of income and reducing expenditure, etc. As to Pure Technical Efficiency Change of hotels in different periods, Hotel Kingdom (H17), Hotel Holiday Garden (H18), Ambassador Hotel Kaohsiung (H19), Howard Hotel Kaohsiung (H21) and Hotel National (H22) grow for six times in 9 periods. Grand Hyatt Taipei (H13), Grand Formosa Regent (H14), Far Eastern Plaza Hotel (H16) and Hotel China (H29) do not have Pure Technical Efficiency changed in 9 periods; in 1999-2000, 2002-2003 and 2005-2006, over half of hotels have growth of Pure Technical Efficiency and in 1998-1999, 2003-2004 and 2006-2007, most of hotels have Pure Technical Efficiency decreased.

iii. Scale efficiency change

Growth of Scale Efficiency means hotels are approximate to MPSS. Regarding Scale Efficiency Change of hotels in different periods, Ambassador Hotel Kaohsiung (H19) grow for six times in nine periods; Royal Hotel (H11), Sherwood Hotel (H15), Far Eastern Plaza Hotel (H16) and Hotel Kingdom (H17) only have lower growth for two times in nine periods. In 2000-2001 and 2002-2003, over half of hotels have growth of Scale Efficiency and in 1998-1999, 1999-2000, 2001-2002, 2003-2004, 2004-2005, 2005-2006 and 2006-2007, most of hotels have Scale Efficiency decreased.

3.2.3 Analysis on total factor productivity change

Total Factor Productivity is product of Technology Change and Technical Efficiency Change and it can be general index of operational performance measurement by tangible input and output. As to Total Factor Productivity change of hotels in different periods, Hotel Riverview Taipei (H04) grow for eight times in nine periods; Ambassador Hotel (H02), Grand Formosa Regent (H14), Ambassador Hotel Kaohsiung (H19) and Howard Hotel Taichung (H25) grow for seven times in nine periods; Howard Hotel (H12), Sherwood Hotel (H15), Grand Han-Lai Hotel (H20), Marshal Hotel (H26), Parkview Hotel (H28) and Tainan Hotel (H33) only have three times of lower growth in nine periods; in 1998-1999, 1999-2000, 2001-2002, 2003-2004 and 2004-2005, over half of hotels have Total Factor Productivity increased and in 2000-2001, 2002-2003, 2005-2006 and 2006-2007, over half of hotels have Total Factor Productivity decreased.

4 Conclusions and suggestions

This study proposes expansion model of DEA and analyzes operation of international tourist hotels in Taiwan in 1998 - 2007. The findings not only extend research field of DEA, but also allow international tourist hotels in Taiwan to recognize their industrial position and competitive environment and function as criterion for resource or strategy adjustment. Conclusions are summarized as follows:

i. Proposal of expanded model of DEA:

Malmquist Index can indicate year-to-year change of long-term industrial development. In order to acquire more precise efficiency measurement, this study includes weights of input and output variables in the original model and proposes input-oriented Assurance Region Malmquist Index model.

ii. Productivity of hotel industry has been growing for long term and the main reason is Technology Change:

In 1998-2007, geometric mean of Total Factor Productivity of overall industry is 1.010. It demonstrates that in the past ten years, productivity of hotel industry is increasing. Among 21 hotels with increasing Total Factor Productivity, most of them have increased Technology Change and decreased Technical Efficiency. Among 12 hotels with decreasing Total Factor Productivity, most of them have increased Technology Change and declined Technical Efficiency; mean of Technology Change in overall industry is 1.061 which is increased by 0.061%, Technical Efficiency Change is 0.951 which is reduced by 0.049%. It is inferred that in the past ten years, the main reason of growth of overall hotel industry productivity is Technology Change. It implies that hotels value enhancement of asset input or R&D and innovative capacity. Therefore, for most of hotels, at present, they should enhance Technical Efficiency such as marketing or service quality in order to increase productivity.

iii. Returns to Scale of overall industry decrease, and it indicates that the hotel industry is highly competitive:

Geometric mean of Scale Efficiency Change is 0.955 which decreases. 8 hotels have increased Scale Efficiency Change and 25 of them have decreased change. It demonstrates that in the past ten years, average production scale of hotel industry deviates from MPSS. It implies that growth of tourism industry in Taiwan does not catch up with increase of international tourist hotels. Therefore, Scale Efficiency of overall industry decreases in highly competitive environment. The government should propose more effective measure in order to enhance development of tourism industry in Taiwan.

iv. Growth and decline of international tourist hotels are closely related to governmental measures to enhance industrial development:

Regarding long-term change of Total Factor Productivity in 1998-2007, Total Factor Productivity of five periods grows and Total Factor Productivity of four periods decreases. It shows that operational performance of international tourist hotels was affected by foreign and domestic political and economic events.

In diverse research themes, there are various expanded models of DEA. Besides the propriety of research topics, these studies mostly modify previously established models and compare with them to revise and improve the modified model. Malmquist Index which indicates long-term industrial change is only based on radial orientation as scale. Although this study takes into account of weight limit, it can still develop Assurance Region Malmquist Index model upon slack analysis.

5 References

[1] Taiwan Tourism Bureau. The Operating Report of International Tourist Hotels in Taiwan, Taiwan Tourism Bureau, Taipei, 1997-2006.

[2] Sun, Shinn and Lu Wen-Min, Evaluating the Performance of the Taiwanese Hotel Industry Using a Weight Slacks-based Measure, Asia-Pacific Journal of Operational Research 2005, 22(4), 487-512.

[3] Färe, Rolf, Grosskopf Shawna, Norris Mary and Zhang Zhongyang, Productivity Growth, Technical Progress and Efficiency Change in Industrialized Countries, The American Economic Review 1994, 84(1), 66-83.

[4] Lee, C., Lee, W. R. and Hsu, H. W., An Empirical Study on the Relationship between Strategic Groups and Performance in Taiwan's International Tourist Hotel Industry, Journal of Business Administration 2000, 48 , pp.89-120.

[5] Kimes, Shery E., The Basics of Yield Management, Cornell Hotel and Restaurant Administration Quarterly 1989, 30(3), 14-19.

[6] Wassenaar, Dirk J. and Stafford Edwin R., The lodging index: an economic indicator for the hotel/model industry, Journal of Travel Research 1991, 30(1), 18-21.

[7] Wijeysinghe, B. S., Breakeven occupancy for a hotel operation, Management Accounting 1993, 71(2), 32-33.

[8] Baker, Michael and Riley Michael, New perspectives on productivity in hotels：some advances and new directions, International Journal of Hospitality Management 1994, 13(4), 297-311.

[9] Anderson, Randy I., Fish Mary, Xia Yi and Michello Frank, Measuring Efficiency in the Hotel Industry: A Stochastic Frontier Approach, International Journal of Hospitality Management 1999, 18,45-57.

[10] Banker, Rajiv D., Charnes Abraham, Cooper William W. and Clarke R., Constrained game formulations and interpretations for data envelopment analysis, European Journal of Operational Research 1989, 40(3), 299-308.

[11] Anderson, Randy I., Lewis Danielle, and Parker Mike E., Another Look at the Efficiency of Corporate Travel Management Departments, Journal of Travel Research 1999, 37, 267-272.

[12] Wang, F. C., Hung, W. T. and Shang, J. K. Measuring the Efficiency of China and Non-chain International Tourist Hotel in Taiwan, Asia-Pacific Economic and Management Review (7:1), 2004, pp.109-123.

[13] Wang, Y. H., Lee, W. F. and Wong, C. C. Productivity and Efficiency Analysis of International Tourist Hotels in Taiwan: An Application of the Stochastic Frontier Approach, Taiwan Economic Review (35:1), 2007, pp.55-86.

[14] Morey, Richard C. and Dittman David A., Evaluating a Hotel GM's Performance-A Case Study in Benchmarking, Cornell Hotel and Restaurant Administration Quarterly 1995, 36(5), 30-35.

[15] Charnes, Abraham, Cooper William W. and Rhodes E., Measuring the Efficiency of Decision Making Units, European Journal of Operations Research 1978, 2(6), 429-444.

[16] Anderson, Randy I., Fok Robert, and Scott John, Hotel Industry Efficiency：An Advanced Linear Programming Examination, American Business Review 2000, 18(1), 40-48.

[17] Tsaur, Sheng-Hshiung, The Operating Efficiency of International Tourist Hotels in Taiwan, Asia Pacific Journal of Tourism Research 2001, 6(1), 73-87.

[18] Hwang, Shiuh-Nan and Chang Te-Yi, Using Data Envelopment Analysis to Measure Hotel Managerial Efficiency Change in Taiwan, Tourism Management 2003, 24, 357-369.

[19] Barros, Carlos Pestana, Evaluating the efficiency of a small hotel chain with a Malmquist Productivity Index, International Journal of Tourism Research 2005, 7, 173-184.

[20] Cooper, William W., Deng Honghui, Gu Bisheng, Li Shanling and Thrall R. M., Using DEA to improve the Management of congestion in Chinese industries (1981-1997), Socio-Economic Planning Science 2001, 35, 227-242.

[21] Yang, Chyan and Lu Wen-Min, Performance benchmarking for Taiwan's International Tourist Hotels, INFOR 2006, 44(3), 229-245.

[22] Morita, Hirohi, Hirokawa Koichiro and Zhu Joe, A slack-based measure of efficiency in context-dependent data envelopment analysis, The International Journal of Management Science 2005, 33(4), 357-362.

[23] Cheng Hilary, Lu Yi-Chuan and Chung Jen-Tsung, Improved slack-based context-dependent DEA – A study of international tourist hotels in Taiwan, Expert Systems with Applications 2010, 37, 6452–6458.

[24] Cheng Hilary, Lu Yi-Chuan and Chung Jen-Tsung, Assurance Region context-dependent DEA with an application to Taiwanese hotel industry, Int. J .Operational Research 2010 , 7, 1-20

[25] Charnes, Abraham, William W. Cooper and E. Rhodes, Short Communication: Measuring the Efficiency of Decision-making Units, European Journal of Operational Research 1979, 3(4), 339.

[26] Cherchye, L., W. Moesen, N. Rogge, T. Van Puyenbroeck, M. Saisana, A. Saltelli, R. Liska and S. Tarantola, Creating Composite Indicators with DEA and Robustness Analysis: The Case of the Technology Achievement Index, Journal of the Operational Research Society 2008, 59(2), 239-251.

[27] Thompson, Russell G., Jr., F. D. Singleton, Robert M. Thrall and Barton A. Smith, Comparative Site Evaluations for Locating a High-Energy Physics Lab in Texas, Interfaces 1986, 16(6), 35-49.

[28] Dyson, R. G. and E. Thanassoulis, Reducing Weight Flexibility in Data Envelopment Analysis, Journal of the Operational Research Society 1988, 39(6), 563-576.

[29] Allen, R., A. Athanassopoulos, R. G. Dyson and E. Thanassoulis, Weights Restrictions and Value Judgments in Data Envelopment Analysis: Evolution, Development and Future Directions, Annuals of Operations Research 1997, 73, 13-34.

[30] Cooper, William W., Seiford Lawrence M. and Tone Kaoru, Data Envelopment Analysis – A Comprehensive Text with Models, Applications, References and DEA – Solver Software, Boston: Kluwer Academic Publishers 2000.

[31] Dyson, R. G., R. Allen, A. S. Camanho, V. V. Podinovski, C. S. Sarrico and E. A. Shale, Pitfalls and Protocols in DEA, European Journal of Operational Research 2001, 132(2), 245-259.

[32] Cave, Douglas W., Laurits R. Christensen and W. Erwin Diewert, The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity, The Journal of the Econometric Society 1982, 50(6), 1392-1414.

[33] Shepherd, Ronald William, Theory of Cost and Production Function, Princeton University Press, Princeton, 1970.

[34] Ray, Subhash C. and Evangelia Desli, Productivity Growth, Technical Progress, and Efficiency Change in Industrialized Countries: Comment, The American Economic Review 1997, 87(5), 1033-1039.

[35] Golany, Boaz and Roll Y., An Application Procedure for DEA, The International Journal of Management Science 1989, 17(3), 237-250.

[36] Satty, Thomas L., The Analytic Hierarchy Process, McGraw Hill, New York, 1980.

A Mutually Supervised Ensemble Approach for Clustering Heterogeneous Datasets

M. Hossain[1], **S. Bridges**[2], **Y. Wang**[2], **and J. Hodges**[2]
[1]Department of Computer Science, Fairmont State University, WV, USA
[2]Department of Computer Science and Engineering, Mississippi State University, MS, USA

Abstract— *We present an algorithm to address the problem of clustering two contextually related heterogeneous datasets that use different feature sets, but consist of non-disjoint sets of objects. The method is based on clustering the datasets individually and then combining the resulting clusters. The algorithm iteratively refines the two sets of clusters using a mutually supervised approach to maximize their mutual entropy and finally computes a single set of clusters. We applied our algorithm on a document collection using multiple feature sets that were extracted by natural language preprocessing methods. Empirical results demonstrate that our method outperforms clustering based on individual feature sets, clustering based on unified feature sets, and clustering based on a well-studied ensemble method.*

Keywords: Clustering, Ensemble, Mutual Entropy

1. Introduction

The traditional clustering paradigm pertains to a single dataset. In some problem domains, multiple contextually related heterogeneous datasets may be available that can provide complementary information. Clustering these datasets may reveal interesting relationships between objects that cannot be obtained by clustering a single dataset. The increasing availability of this type of heterogeneous data in different problem domains, e.g., information retrieval, molecular biology, etc. has motivated research into developing algorithms for clustering contextually related heterogeneous datasets [1], [2], [3], [4].

Heterogeneous datasets can be clustered by constructing a unified feature space [1], [2] or by clustering the datasets individually and then combining the resulting cluster sets [3], [4]. In some domains, several feature sets may be available to represent the same set of objects, but it may not be easy to compute a useful and effective unified feature space because of structural and semantic heterogeneity between the feature sets. We were motivated to pursue an ensemble based approach for clustering heterogeneous datasets.

Ensemble methods [5], [6], [7], [8] have been developed that typically focus on combining different clustering results from a single data set. These algorithms do not work well on clustering results generated from different datasets. In [9], an iterative clustering method was proposed that uses two independent disjoint subsets of the original feature set, but clustering is performed on the same set of objects. In [10], disjoint subsets of objects are clustered independently and the resulting sets of clusters are combined by matching the cluster centroids, but all the objects are represented by the same features. In [4], a correspondence approach was used to extract the correspondence between the different sets of clusters, but a final set of clusters is not produced. In this paper, we address the problem of clustering datasets that are represented by heterogeneous feature sets and that may not necessarily contain the same set of objects.

Even though clustering is traditionally perceived to be an unsupervised task, in semi-supervised clustering [11], the performance of an unsupervised clustering algorithm can be improved with some supervision in the form of some labeled data or constraints. In the context of clustering two heterogeneous datasets, once individual datasets are clustered, each set of clusters can be used to supervise the refinement of the other set of clusters in order to provide a combined and improved clustering of the two datasets. In this paper, we present a mutually supervised clustering algorithm for clustering two heterogeneous datasets. It starts with a set of clusters generated from each dataset, then iteratively refines each set of clusters using the other, and finally generates a single set of clusters.

The remainder of the paper is organized as follows. In section 2, we present a formal description of our clustering problem. In section 3, we present our algorithm. In section 4, we describe our datasets and present experimental results. Finally, in section 5, we provide concluding remarks.

2. Mutually Supervised Ensemble Clustering

Before giving a formal description of the problem of clustering heterogeneous datasets, we first give an intuitive explanation of our clustering approach. We are interested in extracting a single set of partitional clusters from two related heterogeneous datasets. Our method will use a mutually supervised approach to iteratively recompute each set of clusters computed from the individual datasets. The purpose is to allow each individual clustering to refine the other clustering so that at the end of each iteration, the similarity between the two sets of clusters increases. We view this problem as maximizing the *mutual entropy* [12] between the individual clusterings. Mutual entropy can be used to represent the information content shared by two random variables. In our previous work [13], we presented an ensemble algorithm that computes a set of *seed clusters* and then recomputes each set of clusters around the seed clusters iteratively so that the mutual entropy between the two sets of clusters converges. Even though this algorithm was shown to provide an effective solution for clustering heterogeneous datasets, a more effective approach can be to use each set of clusters to iteratively refine the other.

2.1 Problem Statement

Let D_1 and D_2 be two datasets having feature sets F_1 and F_2 respectively. Heterogeneity between two datasets can occur at the feature level, at the object level, or both. The datasets may have same features, some common features, or no common feature. The same situation can occur with the objects. Ensemble algorithms of [5], [6], [7], [8] primarily deal with the situation where $F_1 = F_2$ and $D_1 = D_2$. In this paper, we deal with a situation where $\emptyset \subseteq (F_1 \cap F_2) \subseteq (F_1 \cup F_2)$ and $(D_1 \cap D_2) \supset \emptyset$, i.e., the datasets have disjoint or non-disjoint feature sets and consist of non-disjoint object sets.

Definition 1: Let $D_1 = \{o_1^{(1)}, o_2^{(1)}, \ldots, o_{n_1}^{(1)}\}$ and $D_2 = \{o_1^{(2)}, o_2^{(2)}, \ldots, o_{n_2}^{(2)}\}$, where n_x is the number of objects in D_x. D_1 and D_2 are said to be *congruent* if they represent the same set of objects and *semi-congruent* if some of the objects they represent (not all) are contained in both.

Definition 2: Let $D = D_1 \cup D_2 = \{o_1, o_2, \ldots, o_n\}$, where $max(n_1, n_2) \leq n \leq (n_1 + n_2)$. We define the *object representation vector* of D_x as $V_x = \{v_1^{(x)}, v_2^{(x)}, \ldots, v_n^{(x)}\}$, where $v_i^{(x)}$ is one if $o_i \in D_x$ and $v_i^{(x)}$ is zero if $o_i \notin D_x$.

Definition 3: Let $\mathcal{C}_x = \{C_1^{(x)}, C_2^{(x)}, \ldots, C_k^{(x)}\}$ be the set of clusters computed from D_x. Given $D = \{o_1, o_2, \ldots, o_n\}$, the *cluster assignment vector* of $C_j^{(x)}$ is defined as $A_j^{(x)} = \{a_{j1}^{(x)}, a_{j2}^{(x)}, \ldots, a_{jn}^{(x)}\}$, where $a_{ji}^{(x)}$ is one if $o_i \in C_j^{(x)}$ and $a_{ji}^{(x)}$ is zero if $o_i \notin C_j^{(x)}$. A cluster $C_j^{(x)}$ is assumed to be represented by a cluster centroid $\mu_j^{(x)}$, computed using the feature space of D_x.

Given two datasets D_1 and D_2 where $(D_1 \cap D_2) \supset \emptyset$, two sets of clusters $\mathcal{C}_1$ and $\mathcal{C}_2$ computed from D_1 and D_2 respectively, the goal is to compute a set of k disjoint clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ so that $\forall o_i \in (D_1 \cup D_2)\ \exists C_j \ni o_i$.

2.2 Mutual Entropy

The notion of mutual entropy is based on Shannon's information theory and quantifies the amount of information two random variables share. If two random variables are independent, their mutual entropy is zero, i.e., none contains any information about the other. If they are identical, then all the information contained in one is shared by the other. The mutual entropy of two random variables X and Y can be defined using their individual and joint entropies [12]:

$$\sum_{i,j} P(X = x_i, Y = y_j) \log \frac{P(X = x_i, Y = y_j)}{P(X = x_i) P(Y = y_j)}$$

In our case, each random variable is a given clustering. $P(X = x_i)$ is estimated as the fraction of the objects that are in a cluster in one of the clusterings and $P(X = x_i, Y = y_j)$ as the fraction of the objects that are common in two clusters from two different clusterings.

Definition 4: Let $m_i^{(x)}$ be the number of objects in the i^{th} cluster of $\mathcal{C}_x$, i.e., $m_i^{(x)} = |C_i^{(x)}|$. Let $m_{ij}^{(xy)}$ be the number of objects shared by the i^{th} cluster of $\mathcal{C}_x$ and j^{th} cluster of $\mathcal{C}_y$, i.e., $m_{ij}^{(xy)} = |C_i^{(x)} \cap C_j^{(y)}|$. The *mutual entropy* between two sets of clusters $\mathcal{C}_x$ and $\mathcal{C}_y$ is then defined as:

$$\begin{aligned} \gamma_{xy} &= \sum_{i,j} \frac{m_{ij}^{(xy)}}{n} \log \frac{m_{ij}^{(xy)}/n}{\left(m_i^{(x)}/n\right)\left(m_j^{(y)}/n\right)} \\ &= \frac{1}{n} \sum_{i,j} m_{ij}^{(xy)} \log \frac{m_{ij}^{(xy)} n}{m_i^{(x)} m_j^{(y)}} \end{aligned}$$

The motivation for considering the mutual entropy stems from its ability to measure a general dependence among random variables. In our case, when the mutual entropy between two sets of clusters is large, it means they are more similar. The goal of our algorithm is to maximize this mutual entropy based on complementary information contained in the two datasets, in order to obtain more meaningful clustering than obtained by either dataset individually.

2.3 Seed Clusters

Seed clusters play an important role in the final step of our clustering approach. Once the mutual entropy between two sets of clusters reaches a maximum value, we use the seed clusters to compute the final set of clusters.

Definition 5: Given two datasets D_1 and D_2, a *seed cluster* S_j is defined as a set of objects that are in both D_1 and D_2, i.e., $S_j \subset (D_1 \cap D_2)$. An object is called a *seed object* if it belongs to a seed cluster and is called a *non-seed object* if it does not belong to any seed cluster.

3. Algorithm Overview

In this section, we present an overview of our clustering algorithm. For convenience, we will refer to it as CLUST$_H$. The algorithm is based on computing a set of clusters from each dataset and then using each to refine the other iteratively. When the mutual entropy between the two sets of clusters is maximized, a conglomeration step is carried out to generate a single set of clusters.

3.1 Iterative Refinement

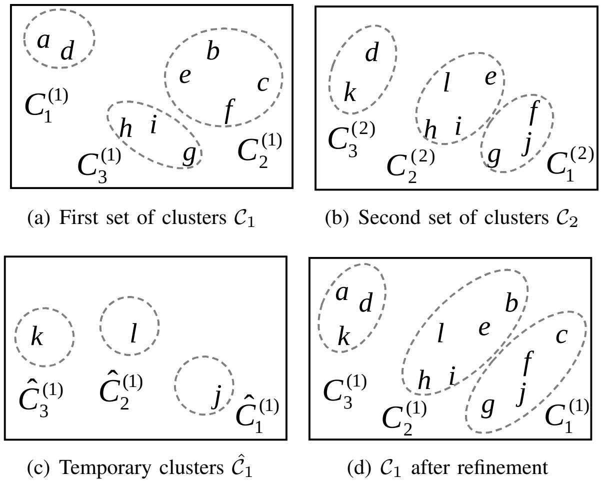
Once the individual clustering has been performed on each dataset, the algorithm progresses iteratively through some refinement steps. In these steps, the algorithm will refine the first set of clusters using the second set of clusters and vice-versa. At the beginning of each refinement step, a set of *temporary clusters* is created for each clustering. Each set of temporary clusters will be initially empty and will correspond to the clusters from the other clustering.

When the first set of clusters $\mathcal{C}_1$ is refined, a set of temporary clusters $\hat{\mathcal{C}}_1 = \{\hat{C}_1^{(1)}, \hat{C}_2^{(1)}, \ldots, \hat{C}_k^{(1)}\}$ will be created, each of which will be initially empty. Note that, each $\hat{C}_j^{(1)}$ will correspond to $C_j^{(2)}$ and will have a precomputed centroid $\hat{\mu}_j^{(1)}$ associated with it. Each $\hat{\mu}_j^{(1)}$ will be computed using the objects that are in both datasets, i.e., the objects of $C_j^{(2)}$ that are also in the first dataset D_1. But, the centroid computation will be based on the feature space of D_1 only.

After the centroids have been computed for $\hat{\mathcal{C}}_1$, the objects that are exclusively in the second dataset will be added to the respective clusters, i.e., each $\hat{C}_j^{(1)}$ will consist of objects of $C_j^{(2)}$ that are only in D_2. Then, each object that appears in D_1 will be added to the closest temporary cluster. At the end of this step, each temporary cluster will consist of objects from both datasets and the set of temporary clusters $\hat{\mathcal{C}}_1$ will replace the existing clusters in the first set of clusters $\mathcal{C}_1$.

The same process will be carried out for the second set of clusters, i.e., the clusters in the second set of clusters $\mathcal{C}_2$ will be recomputed using the first set of clusters $\mathcal{C}_1$. This is a mutually supervised process in the sense that one set of clusters guides the computation of the other set of clusters. At the end of each refinement step, the mutual entropy between the two sets of clusters will be computed. These iterations will be repeated until the mutual entropy between the two sets of clusters converges. We will use the average mutual entropy in the most recent iterations to determine convergence.

Let us explain the refinement step with the two sets of clusters shown in Figs. 1(a) and 1(b). There are two datasets, $D_1 = \{a, b, c, d, e, f, g, h, i\}$ and $D_2 = \{d, e, f, g, h, i, j, k, l\}$. The first set of clusters $\mathcal{C}_1$ contains the clusters $\{a, d\}$, $\{b, c, e, f\}$, and $\{g, h, i\}$. The second set of clusters $\mathcal{C}_2$ contains the clusters $\{f, g, j\}$, $\{e, h, i, l\}$, and $\{d, k\}$. The process starts by creating $\hat{\mathcal{C}}_1$, a set of three temporary clusters that are initially empty. The centroids will be pre-computed for these clusters using $\{f, g\}$, $\{e, h, i\}$, and $\{d\}$ respectively, since each of these sets of objects is contained in the corresponding cluster of $\mathcal{C}_2$ and belong to both D_1 and D_2. Note that the centroid computation will be done based on the feature space of D_1 only. Then, the temporary clusters will be populated to contain $\{j\}$, $\{l\}$, and $\{k\}$ respectively as shown in Fig. 1(c), since these are the objects that exclusively belong to D_2. Then, each object of D_1 will be assigned to one of the temporary clusters of $\hat{\mathcal{C}}_1$ based on the pre-computed cluster centroids. Let us assume that this reassignment results in the clustering shown if Fig. 1(d). This is essentially the refinement of $\mathcal{C}_1$ using $\mathcal{C}_2$. The same process, as explained above, will be carried out to refine $\mathcal{C}_2$ using the pre-refinement state of $\mathcal{C}_1$. These refinement steps will be repeated until the mutual entropy converges and the conglomeration step will be carried out as explained next.



(a) First set of clusters $\mathcal{C}_1$

(b) Second set of clusters $\mathcal{C}_2$

(c) Temporary clusters $\hat{\mathcal{C}}_1$

(d) $\mathcal{C}_1$ after refinement

Fig. 1: Refinement of one set of clusters using another.

3.2 Conglomeration Step

The conglomeration step is based on the computation of *seed clusters*. To generate seed clusters, we will identify k pairs of similar clusters across the two clusterings. One straightforward approach for computing the similarity between two clusters is to count the number of common objects shared between the clusters. However, this may cause a tie between two pairs of clusters and will often cause large diverse clusters to be selected over smaller purer clusters. Instead, we will compute the similarity between $C_i^{(1)}$ and $C_j^{(2)}$ using the information theoretic measure $2m_{ij}^{(12)}/(m_i^{(1)} + m_j^{(2)})$. This measure will give us a continuous value between 0 and 1. If there is no common object between $C_i^{(1)}$ and $C_j^{(2)}$, it will evaluate to 0, and if the clusters are exactly the same, it will evaluate to 1. We need to compute the similarity between all pairs of clusters across the two clusterings.

We considered two methods for pairing similar clusters across the two clusterings. One approach is to find the "most similar" pairs of clusters. However, this is a combinatorial optimization problem in itself, and can be computationally intensive if there are many clusters. Instead, we adopted the second approach where a greedy algorithm is used for selecting similar cluster pairs. In this approach, for each of the k clusters from one clustering, we select the most similar cluster from the other clustering that has not yet been paired with a cluster from the first set of clusters. Once the k pairs of similar clusters are identified, the seed clusters will be computed from the intersection of each pair of clusters.

Let us explain the seed cluster computation for the two sets of clusters shown in Figs. 1(a) and 1(b). For $C_1^{(1)}$, the first cluster in Fig. 1(a), the most similar one in Fig. 1(b) is $C_3^{(2)}$. These two clusters will be used to generate the first set of seed cluster, i.e., $S_1 = C_1^{(1)} \cap C_3^{(2)} = \{d\}$. For $C_2^{(1)}$, the second cluster in Fig. 1(a), the most similar one in Fig.

1(b) is $C_1^{(2)}$. These two clusters will be used to generate the second seed cluster, i.e., $S_2 = C_2^{(1)} \cap C_1^{(2)} = \{f\}$. Finally, the only match for $C_3^{(1)}$ is $C_2^{(2)}$ and $S_3 = C_3^{(1)} \cap C_2^{(2)} = \{h, i\}$. The seed clusters are shown in Fig. 2.

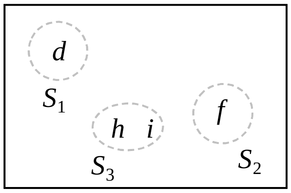


Fig. 2: Seed clusters computed from the two sets of clusters shown in Figs. 1(a) and 1(b).

Fig. 3 shows the algorithm for generating seed clusters. Once the seed clusters are computed, two centroids will be computed for each seed cluster, each based on one of the datasets. There will still be objects in both the datasets that will not be in any of the seed clusters. We will assign each of these non-seed objects to the closest seed cluster. If the non-seed object is contained in both the datasets, then seed cluster centroids computed from both datasets will be used to find the closest seed cluster. If the non-seed object is contained in one dataset but not the other, then seed cluster centroids computed from the respective dataset will be used to find the closest seed cluster.

```
Algorithm: $FindSeedClusters(\mathcal{C}_1, \mathcal{C}_2)$
input : Two sets of clusters $\mathcal{C}_1$ and $\mathcal{C}_2$
output: A set of seed clusters $\{S_1, S_2, \ldots, S_k\}$

for $j \leftarrow 1$ to $k$ do
    $paired_j \leftarrow 0$
for $i \leftarrow 1$ to $k$ do
    $q \leftarrow \mathrm{argmax}_j (1 - paired_j) \cdot 2m_{ij}^{(12)} / (m_i^{(1)} + m_j^{(2)})$
    $paired_q \leftarrow 1$
    $S_i \leftarrow C_i^{(1)} \cap C_q^{(2)}$
return $\{S_1, S_2, \ldots, S_k\}$
```

Fig. 3: Algorithm for generating seed clusters.

3.3 The CLUST$_H$ Algorithm

Fig. 4 describes the CLUST$_H$ algorithm. The inputs to the algorithm are two datasets D_1 and D_2 where $(D_1 \cap D_2) \supset \emptyset$, and two sets of clusters $\mathcal{C}_1 = \{C_1^{(1)}, C_2^{(1)}, \ldots, C_k^{(1)}\}$ and $\mathcal{C}_2 = \{C_1^{(2)}, C_2^{(2)}, \ldots, C_k^{(2)}\}$ computed from D_1 and D_2 respectively. The output of the algorithm is a single set of partitional clusters, $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$. CLUST$_H$ uses two sets of temporary clusters $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$. For describing the algorithm, we use ˆ to represent the parameters corresponding to the temporary clusters. The similarity between the i^{th} object o_i and the centroid of cluster $C_q^{(x)}$ is represented by $sim(o_i, \mu_q^{(x)})$.

```
Algorithm: $CLUST_H(D_1, D_2, \mathcal{C}_1, \mathcal{C}_2)$
input : Dataset $D_1$ and the set of clusters $\mathcal{C}_1$
        Dataset $D_2$ and the set of clusters $\mathcal{C}_2$
output: A single set of clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$

//Refinement Step
repeat
    for $x \leftarrow 1$ to 2 do
        for $j \leftarrow 1$ to $k$ do
            $\hat{C}_j^{(x)} \leftarrow \left\{o_i : a_{ji}^{(x\%2+1)} = v_i^{(x)} = v_i^{(x\%2+1)} = 1\right\}$
            compute $\hat{\mu}_j^{(x)}$
    for $x \leftarrow 1$ to 2 do
        for $j \leftarrow 1$ to $k$ do
            $\hat{C}_j^{(x)} \leftarrow \left\{o_i : a_{ji}^{(x\%2+1)} = v_i^{(x\%2+1)} = 1, v_i^{(x)} = 0\right\}$
    for $x \leftarrow 1$ to 2 do
        foreach $o_i \in D_x$ do
            $j \leftarrow \mathrm{argmax}_q \, sim\left(o_i, \hat{\mu}_q^{(x)}\right)$
            $\hat{C}_j^x \leftarrow \hat{C}_j^x \cup \{o_i\}$
    for $x \leftarrow 1$ to 2 do
        for $j \leftarrow 1$ to $k$ do
            $C_j^{(x)} \leftarrow \hat{C}_j^{(x)}$
until $\gamma_{xy}$ converges

//Conglomeration Step
$\mathcal{S} \leftarrow FindSeedClusters(\mathcal{C}_1, \mathcal{C}_2)$        // $\mathcal{S} = \{S_1, S_2, \ldots, S_k\}$
for $j \leftarrow 1$ to $k$ do
    $C_j \leftarrow C_j^{(1)} \leftarrow C_j^{(2)} \leftarrow S_j$
    compute $\mu_j^{(1)}$ and $\mu_j^{(2)}$
foreach $o_i \notin \{S_1 \cup S_2 \cup \ldots \cup S_k\}$ do
    $j \leftarrow \mathrm{argmax}_q \max\left(v_i^{(1)} sim\left(o_i, \mu_q^{(1)}\right), v_i^{(2)} sim\left(o_i, \mu_q^{(2)}\right)\right)$
    $C_j \leftarrow C_j \cup \{o_i\}$
```

Fig. 4: The CLUST$_H$ algorithm.

3.4 Complexity of CLUST$_H$

In the refinement step, CLUST$_H$ needs to scan all the objects for each cluster. If n is the number of total objects from both the datasets, k is the number of clusters, and t is the number of iterations required for convergence, then the overall complexity of these steps is $O(2nkt)$. The conglomeration step computes the seed clusters. There are k^2 pairs of clusters and the computation of the similarity between each pair of clusters requires $O((n/k)^2)$ operations. Hence, the computation of seed clusters requires $O(n^2)$ operations. Finally, assigning each non-seed object to a cluster requires $O(nk)$ operations. Thus the overall complexity is $O(2nkt + n^2 + nk)$. In general, $n \gg k$ and $n \gg t$, and the complexity can be formalized as $O(n^2)$.

4. Algorithm Evaluation

We evaluated the effectiveness of CLUST$_H$ in two situations: (1) two datasets are clustered that represent the same sets of objects - we call these *congruent* datasets, and (2) two datasets are clustered that share some but not all objects - we call these *semi-congruent* datasets.

4.1 Datasets

To evaluate the effectiveness of our approach, we applied CLUST$_H$ and several baseline methods to the task of clustering a document collection consisting of 10,000 journal abstracts from ten different Library of Congress categories. The collection was divided into five non-overlapping subsets where each subset contained 2,000 abstracts (200 abstracts from each category). Different NLP methods were used to preprocess each subset in order to capture different aspects of the documents in different feature spaces. We generated four feature vectors from each abstract; one using syntactic preprocessing and three using semantic preprocessing. The feature vectors extracted using each preprocessing method were used to build a dataset for each of the five document subsets. This resulted in a total of 20 data subsets each containing 2,000 documents.

The syntactic preprocessing was based on constructing a "bag of keywords" from each abstract using the Collins parser[1] and the semantic preprocessing was based on constructing a "bag of senses" from each abstract using the WordNet semantic network[2] and the WordNet::Similarity package[3]. We used three semantic preprocessing methods: node-based, edge-based, and combined node-and-edge-based. We used the cosine coefficient method to calculate the pairwise similarity between feature vectors. A detailed discussion of the data preprocessing can be found in [14].

4.2 Experimental Design

To test the effectiveness of CLUST$_H$, we used several baselines. Ten clusters were generated during each clustering because the datasets were known to have ten categories. The first set of baselines was the partitional clusterings based on individual feature sets. For these baselines, we performed graph-based partitional clustering using METIS[4]. A similarity matrix was computed from each set of syntactic feature vectors and from each set of semantic feature vectors. The similarity matrix generated from a dataset was converted into an adjacency matrix for a graph before applying the graph-partitioning. Each object was treated as a vertex and each similarity value between a pair of objects was treated as the weight of the edge between the two corresponding vertices. Another baseline was partitional clustering based on an unified feature set obtained from the two respective feature sets and we used the same graph-based clustering approach. We also compared the performance of CLUST$_H$ against the well-known ensemble algorithm CSPA [5].

We also wanted to observe how CLUST$_H$ performs with semi-congruent datasets. To generate data for these experiments, we randomly excluded 50% of the objects from each of the 20 datasets with the constraint that each category of documents remained equal sized. On average, this resulted in an overlap of approximately 50% between two datasets constructed from the same document subset using different feature sets, i.e., each pair of combined datasets had approximately 1,500 objects and each individual dataset had exactly 1,000 objects.

When constructing an unified feature set using two semi-congruent datasets, we had to deal with missing values for objects that appear only in one of the datasets. There are two general methods for handling missing feature values: *imputation* and *marginalization* [15]. Imputation requires inferring values from neighbors or averaging and marginalization requires missing values to be ignored. In our document clustering domain, the feature space is extremely sparse and therefore the imputation method is not applicable. We decided to employ marginalization with pairwise deletion when dealing with the unified feature sets, i.e., when computing the similarity of two objects, only the features with observed values in both objects were used.

4.3 Evaluation Results

We used *F-Measure* [16] to compare cluster qualities. The F-measure quantifies how the clustering fits the actual classification of data where the most desirable value is 1 and the least desirable value is 0. Fig. 5(a) shows the graphical representation of the F-measure values for CLUST$_H$ and baseline clustering schemes with congruent datasets. For comparison with the clustering based on individual feature sets, we only show the individual clustering yielding the higher F-measure value ($\mathcal{C}_{ind}$). For each feature set or a combination of feature sets, five data subsets were used. Each graph shows the F-measure values for a particular combination of two heterogeneous feature sets and six different combinations were used. For example, the first graph shows the F-measure values when syntactically preprocessed and semantically (node-based) preprocessed feature sets were used for clustering. Each curve represents the F-measure for one of the clustering schemes. It is evident that CLUST$_H$ produces high-quality clusters (F-measure $\approx$ 0.9) and outperforms the baselines in all instances.

If we compare the performance of the clustering performed on the unified feature sets with clustering based on individual feature sets, we can see that the unified feature sets do not always produce better quality clusters. This demonstrates that feature set integration does not necessarily take advantage of the mutual information in the individual feature sets to yield improved clustering. It is evident that CLUST$_H$ consistently yields higher quality clusters than clustering based on individual feature sets and unified feature sets. CLUST$_H$ also outperforms the clustering based on the ensemble method CSPA. Note that CSPA was tested using identical experimental setup as CLUST$_H$, i.e., clustering was performed on individual feature sets and then the resulting individual clusterings were combined.

[1] http://www.cs.columbia.edu/ mcollins/code.html

[2] http://wordnet.princeton.edu

[3] http://search.cpan.org/dist/WordNet-Similarity

[4] http://www-users.cs.umn.edu/ karypis/metis/metis/index.html

Fig. 5(b) presents the comparison of F-measure values for CLUST$_H$ and baseline clustering schemes when we used semi-congruent datasets by randomly removing 50% of the objects from each data subset. As before, we obtained higher quality clusters with CLUST$_H$ compared to the baseline schemes. It is interesting to note that clustering with unified feature sets always yields lower quality clusters with semi-congruent datasets when compared to other clustering approaches.

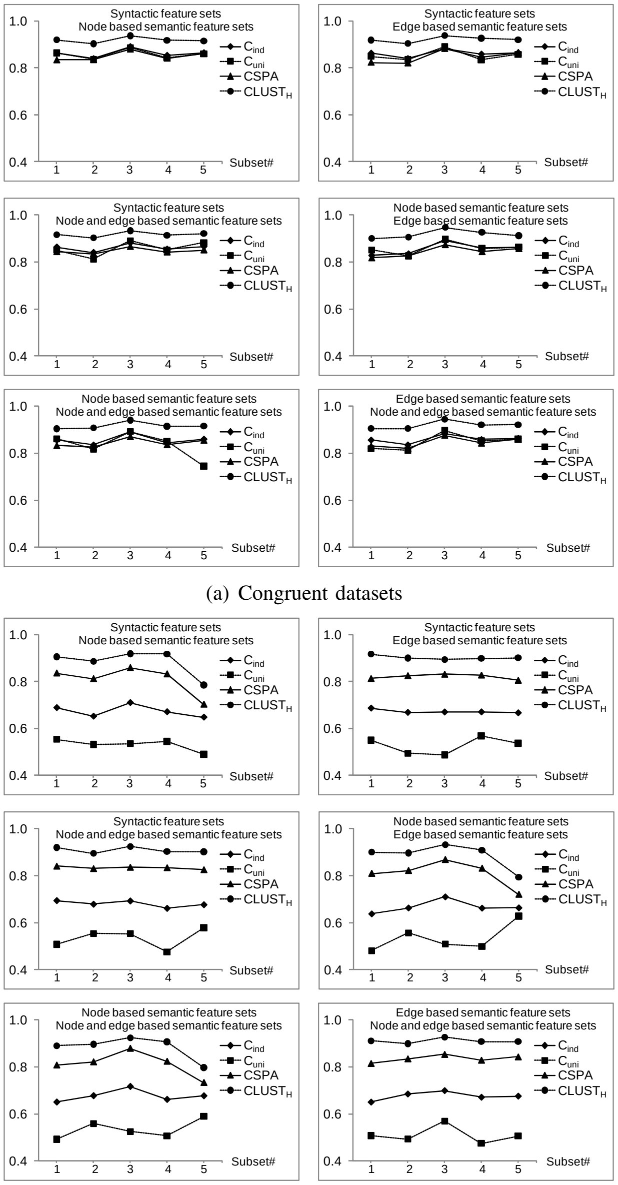
The results presented in Figs. 5(a) and 5(b) are summarized graphically in Figs. 6(a) and 6(b) respectively. The results shown are averages over all thirty observations along with the standard deviation. In addition to higher average F-measure values for CLUST$_H$, lower standard deviations of F-measure values are also observed for CLUST$_H$. Note that for congruent datasets, neither clustering with unified features sets nor CSPA produces higher quality clusters compared to clustering on individual feature sets. When the datasets only share 50% of the same objects (semi-congruent), CSPA and CLUST$_H$ clearly outperform clustering based on individual feature sets and clustering based on unified feature sets with CLUST$_H$ outperforming CSPA.

To test the statistical significance of our results, we performed one-tailed paired T-tests. Table 1 shows the corresponding results. We used an α value of .05 where $T_{critical} = 1.699$ for one-tail tests. The hypotheses tested are CLUST$_H \succ \mathcal{C}_{ind}$, CLUST$_H \succ \mathcal{C}_{uni}$, and CLUST$_H \succ$ CSPA, where $\mathcal{C}_x \succ \mathcal{C}_y$ means that clusters generated by $\mathcal{C}_x$ are of significantly higher quality than clusters generated by $\mathcal{C}_y$. For each hypothesis, the table shows T-value (T_{stat}) and p-value ($P_{T \leq t}$). A T-value of 1.699 or higher and a p-value of 0.05 or lower indicate evidence that the hypothesis is true. The T-test results demonstrate that CLUST$_H$ is significantly better than the baseline clustering schemes.

Table 1: The one-tailed T-test results for comparing CLUST$_H$ with other clustering schemes where $\alpha = .05$ and $T_{critical} = 1.699$.

Hypothesis	Congruent Datasets		Semi-congruent Datasets	
	T_{stat}	$P_{T<t}$	T_{stat}	$P_{T<t}$
CLUST$_H \succ \mathcal{C}_{ind}$	41.09	<0.0001	35.78	<0.0001
CLUST$_H \succ \mathcal{C}_{uni}$	14.59	<0.0001	33.34	<0.0001
CLUST$_H \succ$ CSPA	37.37	<0.0001	33.56	<0.0001

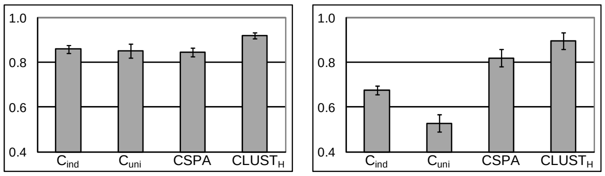
Finally, we wanted to observe how CLUST$_H$ performs when the initial clusterings of the individual datasets differ. Figure 7 shows the % improvement of F-measure values that CLUST$_H$ achieves over CSPA with different initial mutual entropy between two individual clusterings (for congruent datasets). It is evident that CLUST$_H$ has a tendency to perform better when two individual clusterings have less commonality, i.e., smaller mutual entropy.



(a) Congruent datasets

(b) Semi-congruent datasets

Fig. 5: F-measure for CLUST$_H$ and baseline clustering methods. C_{ind}: individual clustering with the higher F-measure, C_{uni}: clustering with unified feature sets, CSPA: clustering based on the ensemble method of [5].



(a) Congruent Datasets

(b) Semi-congruent Datasets

Fig. 6: Comparison of F-measure averages for CLUST$_H$ and baseline clustering methods.

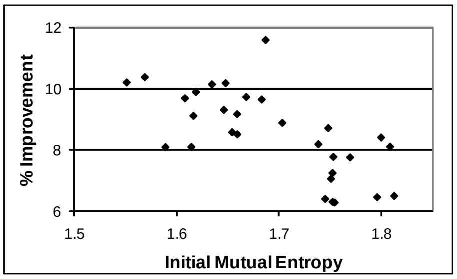


Fig. 7: Improvement in F-measure for CLUST$_H$ over CSPA with different initial mutual entropy between individual clusterings.

5. Conclusion

In this paper, we addressed the problem of clustering two related heterogeneous datasets using an ensemble approach. We presented a mutually supervised algorithm for combining partitional clusterings generated from two heterogeneous datasets. The algorithm uses the cluster membership of objects in each clustering to refine the other clustering iteratively and this mutual refinement process continues until the mutual entropy between the two clusterings converges. Finally, a single set of clusters is computed using the two refined sets of clusters. We applied our algorithm for clustering a document collection using heterogeneous feature sets extracted with natural language preprocessing methods. It was shown to yield higher quality document clusters than clustering based on individual feature sets, unified feature sets, and the well-studied ensemble algorithm CSPA. Our algorithm was observed to perform better than the baseline clustering schemes even when all the objects are not represented by both datasets.

References

[1] P. Pavlidis, J. Weston, J. Cai, and W. Grundy, “Gene functional classification from heterogeneous data,” in *Proc. of 5th Intl. Conf. on Computational Biology (RECOMB 2001), Montreal, Canada, Apr 22-25*, 2001, pp. 249–255.

[2] I. Dagan, Z. Marx, and E. Shamir, “Cross-dataset clustering: Revealing corresponding themes across multiple corpora,” in *Proc. of 6th Conf. on Computational Natural Language Learning (CoNLL 2002), Taipei, Taiwan, Aug 31-Sep 1*, 2002, pp. 15–21.

[3] N. Iam-on, S. Garrett, C. Price, and T. Boongoen, “Link-based cluster ensembles for heterogeneous biological data analysis,” in *Proc. of Intl. Conf. on Bioinformatics and Biomedicine (BIBM 2010), Hong Kong, Dec 18-21*, 2010, pp. 573–578.

[4] V. Rinsurongkawong and C. F. Eick, “Correspondence clustering: An approach to cluster multiple related spatial datasets,” in *Proc. of 14th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD 2010), Hyderabad, India, Jun 21-24*, 2010, pp. 216–227.

[5] A. Strehl and J. Ghosh, “Cluster ensembles - A knowledge reuse framework for combining multiple partitions,” *Journal of Machine Learning Research*, vol. 3, no. Dec, pp. 583–617, 2002.

[6] N. Nguyen and R. Caruana, “Consensus clusterings,” in *Proc. of 7th Intl. Conf. on Data Mining (ICDM 2007), Omaha, Nebraska, Oct 28-31*, 2007, pp. 607–612.

[7] H. G. Ayad and M. S. Kamel, “On voting-based consensus of cluster ensembles,” *Pattern Recognition*, vol. 43, no. 5, pp. 1943–1953, 2010.

[8] G. Forestier, P. Gançarski, and C. Wemmert, “Collaborative clustering with background knowledge,” *Data & Knowledge Engineering*, vol. 69, no. 2, pp. 211–228, 2010.

[9] S. Bickel and T. Scheffer, “Multi-view clustering,” in *Proc. of 4th Intl. Conf. on Data Mining (ICDM 2004), Brighton, UK, Nov 1-4*, 2004, pp. 19–26.

[10] P. Hore, L. O. Hall, and D. B. Goldgof, “A scalable framework for cluster ensembles,” *Pattern Recognition*, vol. 42, no. 5, pp. 676–688, 2009.

[11] M. Bilenko, S. Basu, and R. J. Mooney, “Integrating constraints and metric learning in semi-supervised clustering,” in *Proc. of 21st Intl. Conf. on Machine Learning (ICML 2004), Banff, Alberta, Jul 4-8*, 2004, pp. 81–88.

[12] X. Zhou, X. Wang, E. R. Dougherty, D. Russ, and E. Suh, “Gene clustering based on clusterwide mutual information,” *Journal of Computational Biology*, vol. 11, no. 1, pp. 147–161, 2004.

[13] M. Hossain, S. Bridges, Y. Wang, and J. Hodges, “Extracting partitional clusters from heterogeneous datasets using mutual entropy,” in *Proc. of IEEE Intl. Conf. on Information Reuse and Integration (IRI 2007), Las Vegas, NV, Aug 13-15, 2007*, pp. 447–454.

[14] M. Hossain, S. Bridges, Y. Wang, and J.Hodges, “An ensemble approach for generating partitional clusters from multiple cluster hierarchies,” in *Proc. of IEEE Intl. Conf. on Granular Computing (GrC 2006), Atlanta, GA, May 10-12, 2006*, pp. 666–670.

[15] P. D. Green, J. Barker, M. P. Cooke, and L. Josifovski, “Handling missing and unreliable information in speech recognition,” in *Proc. of 8th Intl. Workshop on Artificial Intelligence and Statistics (AISTATS 2001), Key West, FL, Jan 4-7*, 2001.

[16] B. Larsen and C. Aone, “Fast and effective text mining using linear-time document clustering,” in *Proc. of 5th Intl. Conf. on Knowledge Discovery and Data Mining (KDD 1999), San Diego, CA, Aug 15-18*, 1999, pp. 16–22.

Combined Application of Game Theory and Data Envelopment Analysis as a Methodological Approach for National Defense and Economy

Hilary Cheng[1], Yi-Chuan Lu[2], Ming-Shan Niu[3]
[1]College of Management, Yuan Ze University, Chung-Li 320, Taiwan
[2]Department of Information Management, Yuan Ze University, Chung-Li 320, Taiwan
[3]Graduate School of Management, Ph.D. Program, Yuan Ze University, Chung-Li 320, Taiwan

Abstract - In this paper, instead of conforming to the limitations of traditional research frameworks, the authors have established a game theoretic model between the government and the Ministry of National Defense by focusing on the interactive equilibrium configuration between national security and economic growth to analyze Taiwan and South Korea's national defense expenditures and the scheme of growth that would allow the configuration of national defense expenditure (from the injection of state financial resources) and economic constructions to reach equilibrium. In order to avoid this paper from becoming a purely theoretical discussion (a prevalent trend in previous researches) while verifying the discrepancies between the theoretical values and true values, the DEA method has been adopted to examine the equilibrium solutions to Taiwan and South Korea's national defense expenditure games on top of relative input and output performances of Taiwan, South Korea, Japan, India and Israel to complement and fortify the contents of this study. Findings from this research will serve as a useful reference for competent authorities in their decision-making processes for the effective allotment of state financial resources to achieve a configuration of equilibrium between national defense and economic construction.

Keyword: Equilibrium; Defense expenditure; Data Envelopment Analysis; Malmquist Index

1. Introduction

National defense expenditure also refers to the grand total of resources that a country commits to the construction of national defense in a specific period of time. The most fundamental issue that most researches on national defense expenditure tend to focus on is the proportion of resources from a country's total social economic resources (i.e. GDP) allotted to national defense construction and the ideal configuration for two reasons(1).national defense expenditure has always been one of the government's largest expenditure items and (2)it is useful to the understanding of the correlation between national defense expenditure and economic growth for the facilitation of a balanced development of both the economy and national defense construction.

In this study, Taiwan, South Korea, Japan, India and Israel have been chosen as the subjects of the research with the following objectives: (1) to construct a game theoretic model to reveal the interactive relationship between national defense expenditure and economy for Taiwan and South Korea. The model should shed light on solutions that would allow DMUs to minimize the impact of national defense expenditure on economies while accommodating to the demands for national defense and security within reasonable limits. This approach would make up for the lack of national defense demand analysis (the flaw with the first school of thought as previously mentioned). 2. To further discuss the relative national defense expenditure utilization efficiency for 7 DMUs to make up for the lack of efficiency analysis (the flaw with the second school of thought).

2. Data acquisition

The data used in the DEA estimation comprised information for 7 DMUs during 48 years for a total of 336 observations. First, data availability is an empirical criterion. Second, the literature survey is another criterion of ensuring the validity of the research.

2.1 Data and Input-output variables descriptions

In this research, game theoretic model has been adopted to derive the formula for national defense expenditure in order to formulate the national defense expenditure equilibrium solution for Taiwan and South Korea from 1961-2008. Next, the Malmquist Index from the DEA model will be used to examine the relative efficiency between national defense budget (input) and military capability (output) and compare the relative efficiencies among Taiwan, South Korea, Japan, India and Israel from 1961 through 2008.

3. Research Framework

The research framework is described in Figure 1 as follows.

4. Model derivation

4.1Analysis of national defense capital model

From an economic perspective, the effective military deterrence capability (S_t) of an army comes from two primary components: the capital deposit of the national defense (M_t) (including weaponry, equipment and human resources), and military management and operational knowledge (KM_t) (including military tactic theories and organizational system), and it can be represented as the function: St=F(M_t , KM_t). The formula for national defense capital may be represented as:

$$M_t = M_{t-1} + Y_t R_t (1-r_t) - \lambda \sum Y_{t-i} R_{t-i} (1-r_{t-i})$$
$$(i = 1,2,\cdots)$$

λ represents the rate of depreciation. We can represent the national defense security production function as:

$$S_t = F\left[M_t = M_{t-1} + Y_t R_t (1-r_t) - \lambda \sum Y_{t-i} R_{t-i} (1-r_{t-i}), KM_t\right]$$

(1)

Using Cobb-Douglas Production Function $Y_t = AK^{\alpha} L_t^{\beta}$ as the basis, we can derive the economic growth rate:

$$\Delta Y / Y_t = \Delta A / A_t + \alpha \Delta K / K_t + \beta \Delta L / L_t \qquad (2)$$

Then $y_t = \mathrm{a}_t + \alpha k_t + \beta l_t$

$$y(D_t) = \mathrm{a}(D_t) + \alpha k(D_t) \qquad (3)$$
$$y(R_t, r_t) = \mathrm{a}(r_t) + \alpha k(R_t, r_t)$$

Where a(r_t) is obtained from the production function $Y_t = AK^{\alpha} L_t^{\beta}$ and the TFP for technical advancement refers to the productivity of a set of traditional input and it can be computed using the following formula:

$$TFP = Y / (K^{\alpha} L^{\beta}) \quad (4)$$
$$g_{TFP} = y - ak - \beta l \quad (5)$$
$$g_{TFP} = \Omega(RD/Y) \quad (6)$$

National defense expenditure capital deposit K_t is derived from prior capital deposit K_{t-1} and the net investment function I_t.

$$K_t = K_{t-1} + I_t - \lambda \sum I_{t-i} \qquad (i = 1,2,\cdots)$$

(7)

And thus,

$$\alpha k(R_t, r_t) = \left[\mathrm{a} Y_t R_t (1-r_t) + \mathrm{a}(1-\lambda) \sum_{i=1,2,\cdots} Y_{t-i} R_{t-i} (1-r_{t-1}) / k_t\right]$$

4.2 Model derivation

$$R_t = \mu r_t + R_0 \quad (8)$$

In order to derive the maximum national defense security output, the optimized strategy function may be represented as:

$$Max\{S_t = F\left[M_t = M_{t-1} + Y_t R_t (1-r_t) - \lambda \sum Y_{t-i} R_{t-i} (1-r_{t-i}), K\right.$$

$$\begin{aligned} s.t. \quad & R_t = \mu r_t + R_0 \\ & R_0 < R_t < \pi - \pi_0 \\ & 0 < r_t < 1 - r_0 \qquad (9) \end{aligned}$$

In order for the government to facilitate economic growth, the optimized strategy function may be represented as:

$$s.t. \quad R_t = \mu r_t + R_0 \qquad (10)$$

$$R_0 < R_t < \pi - \pi_0$$

$$0 < r_t < 1 - r_0$$

$$\partial S / \partial r_t = (\partial F / \partial M_t)(\partial M_t / \partial r_t) + (\partial F / \partial M_t)(\partial KM_t / \partial r_t) = (\partial F / \partial M_t) Y_t (\mu + R_0 + 2\mu r_t) = 0$$

We get: $r_t = (\mu - R_0) / 2\mu$;

By plugging $r_t = (\mu - R_0) / 2\mu$ into $R_t = \mu r_t + R_0$

We arrive at $R_t = (\mu + R_0) / 2$

By plugging in $b_t = \Omega_1 / y_t$; $c_t = \alpha Y_t$; $d_t = \alpha\{(1-\lambda)\sum Y_{t-i} R_{t-i} (1 - r_{t-i})\} / K_t y_t \quad (i = 1, 2, \cdots)$,

$$\varphi(R_t, r_t) = (b_t + c_t) R_t r_t - c_t R_t - d_t \qquad (11)$$

$$\varphi(\mu) = 0.25(b_t + c_t)(\mu - R^2 / \mu) - 0.5 c_t \mu - 0.5 c_t R_0 - d_t$$

(12)

$$\partial\varphi(\mu) / \partial\mu = 0.25(b_t + c_t)(\mu - R^2 / \mu) - 0.5 c_t = 0$$

We can obtain the solution $\mu = [(c_t + b_t)/(c_t - b_t)]^{1/2} R_0$

$$r_t^* = 1/2 - 1/[4(c_t + b_t / c_t - b_t)]^{1/2} \qquad (13)$$

$$R_t^* = \{[(c_t + b_t)/(c_t - b_t)]^{1/2} + 1\} R_0 / 2$$

$$\upsilon_t = Y_t R_t^* / Y_{t-1} R_{t-1}^* - 1$$

By substituting the equilibrium solution $r_t^* = 1/2 - 1/[4(c_t + b_t / c_t - b_t)]^{1/2}$ and $R_t^* = \{[(c_t + b_t)/(c_t - b_t)]^{1/2} + 1\} R_0 / 2$ back into the original equation

We will be able to derive the growth rate for national defense budget as:

$$\upsilon_t = (1 + y_t)\{[(c_t + b_t)/(c_t - b_t)]^{1/2} + 1\} / \{[(c_{t-1} + b_{t-1})/(c_{t-1} - b_{t-1})]^{1/2} + 1\} - 1$$

(14)

5. Empirical analysis for the model

Now that we have established the model $Y_t = AK^{\alpha} L_t^{\beta}$, we may apply it to calculate Taiwan and South Korea's capital output elasticity α and labor output elasticity β before computing the results for $TFP = Y / (K^{\alpha} L^{\beta})$.

$LnY = b + aLnK + \beta LnL + \mu$ for regression analysis, we can derive both α and β for Taiwan:

$$LnY = \underset{2.398}{2.420} + \underset{5.601}{0.439} LnK + \underset{1.917}{0.346} LnL + 0.350 \underset{1.136}{D}$$

$$R^2 = 0.899$$

Significance-F=148.840

Results of South Korea's capital output elasticity α and labor output elasticity β are as follows:

$$LnY = \underset{-14.27}{-18.1} + \underset{1.687}{0.211} LnK + \underset{14.901}{4.726} LnL + 0.268 \underset{1.623}{D}$$

$$R^2 = 0.979$$

Significance-F=650.670

We can therefore establish that during the period between 1956 and 2009, Taiwan's capital output elasticity α = 0.439 and her labor output elasticity β = 0.346, whereas South Korea's α = 0.138 and β = 5.127 during the period between 1961 and 2009. we will be able to calculate the three countries' TFP growth rate from 1956 (1961) through 2009 as shown in Table 2.

5.1 Data correction and estimation

Let M_t be Taiwan's annual national defense R&D budget, and $M_t = R_t r_t + (R_2 / Y)_t$. We can modify equation (9) into:

$$g_{TFP} = a + b_i M_{t-i} + cD_t + \mu \qquad (15)$$

$$g_{TFP} = \underset{16.250}{0.029} - \underset{-4.788}{1.083} R_t r_t + \underset{1.215}{0.005} D$$

In 1982, due to the petroleum energy crisis, South Korea's former president Chun Doo-Hwan saw the need to develop the country's hi-tech sectors. As such, 1982 will be chosen as the dividing point for South Korea.

$$g_{TFP} = a + b_i M_{t-i} + cD_t + \mu \qquad (16)$$

$$g_{TFP} = \underset{8.612}{0.00001677} - \underset{-1.228}{0.00062} R_t r_t + \underset{2.837}{0.000003089} D$$

6. Empirical results

During the span of 52 years from 1957 through 2008, Taiwan's true national defense expenditure growth rate was greater than that of the equilibrium solution for 23 years (see Table 3). This suggests that Taiwan's actual national defense expenditure did not exceed the reasonable limit for social economic development.

South Korea went through 10 years where her actual national defense expenditure was higher than that of the equilibrium solution (see Table 3). This shows that South Korea's DMU has been more effective than Taiwan's in terms of national defense resource utilization.

6.1 Analysis on changes in productivity

From Table 4, we see that only Israel and Japan showed improvement in terms of TFP; TFP for the remaining 5 DMUs in the study are all smaller than 1 (indicating decline). The geometric mean for TFP change came to 0.9716364 for all 7 DMUs, indicating that TFP for the subject DMUs in the past 47 years have regressed slightly.

6. CONCLUSIONS

1. Inspirations from the equilibrium solutions from Taiwan and South Korea's game

(1)On Taiwan: During the span of 52 years from 1957 through 2008, Taiwan's true national defense expenditure growth rate was greater than that of the equilibrium solution for 23 years (see Table 3).

(2)On South Korea: South Korea went through 10 years where her actual national defense expenditure was higher than that of the equilibrium solution (see Table 3).

2. Discussion on the overall productivity of national defense expenditure

With regards to TFP: Only Israel (1.0912) and Japan (1.0191) showed improvement in their average TFP and the remaining 5 DMUs showed decline. In terms of equilibrium solution, the average TFP for South Korea's equilibrium solution was higher than the true value while the situation turned out to be different for Taiwan.

6.2 Policy Implication

The contribution from this paper stems from the fact that instead of conforming to the limitations of traditional research frameworks, the authors have established a game theoretic model between the government and the Ministry of National Defense by focusing on the interactive relationship between national defense expenditure, national defense security and economic growth to analyze the scheme of growth that would allow the configuration of national defense expenditure and economic construction to reach equilibrium under limited state financial resources for Taiwan and South Korea. In order to prevent this paper from becoming a purely theoretical discussion (a prevalent trend in previous researches) while verifying the discrepancies between the theoretical values and true values, the DEA method has been adopted to examine the DMUs performances. Findings from this research will serve as a useful reference for competent authorities in their decision-making processes for effective allotment of state financial resources to achieve configuration of equilibrium between national defense and economic construction.

References

Barro, R. J. (1990)." Government spending in a simple model of endogenous growth. ",*Journal of Political Economy,* 98(5), S103–S125.

Benoit, E. (1973), Defense and Economic Growth in Developing Countries, Lexington, MA: Lexington Books.

Benoit, E. (1978), "Growth and Defense in Developing Countries," *Economic Development and Cultural Change,* 26(2), 271–280.

Deger, S. and S. Sen (1983), "Military Expenditure, Spin-Off and Economic Development," Journal of Development Economics, 13, 67–83.

Deger, S. and S. Sen (1992), "Military Expenditure, Aid and Economic Development," Proceedings of the World Bank Annual Conference on Development Economics, 159–189,

Deger, S. and R. Smith (1983), "Military Expenditure and Growth in Less Developed Countries," Journal of Conflict Resolution, 27, 335–353.

Deger, S. (1986)," Economic development and defence expenditure. Economic .",Development and Cultural Change, 35,179–196.

Deger, S. and S. Sen (1995), "Military Expenditure and Developing Countries," in K. Hartley and T. Sandler (eds), Handbook of Defense Economics,1, 275–307, Amsterdam: Elsevier.

Mintz, A. and C. Huang (1990), "Defense Expenditures, Economic Growth and the Peace Dividend," American Political Science Review, 84, 1283–1293.

Mintz, A. and R. Stevenson (1995), "Defense Expenditures, Economic Growth and the Peace Dividend: A Longitudinal Analysis of 103 Countries," Journal of Conflict Resolution, 39,283–305.

Sezgin, S. (2001)," An empirical analysis of Turkey's defence-growth relationships with a multi-equation model (1956-1994).", Defence and Peace Economics, 12, 69–86.

Shieh, J. Y., Lai, C. C. and Chang, W. Y. (2002),” The impact of military burden on long-run growth and welfare,” Journal of Development Economics, vol. 68, issue 2, pages 443-454.

Definition of relevant variables

e	able	ator	nition	ce
Game model	Taiwan and South Korea’s actual GDP	GDP	GDP is the basic measure of a country’s overall economic output. It is an important indicator of a country (region)’s economic status.	OECD Database
Game model	Taiwan and South Korea’s labor force	LABOUR	The population of people above 15 years of age and ready for employment, including the employed and unemployed.	OECD Database
Game model	Taiwan and South Korea’s actual capital deposit	CAPITAL	Capital deposit or deposit capital. From the perspective of corporate capital operation, it refers to all existing capital resources in a corporation’s possession.	OECD Database
Game model	Taiwan and South Korea’s national defense technology R&D budget	DEFENSE BUDGET FOR R&D	Annual national defense budget that goes to technology R&D	OECD Database
Game model	Taiwan and South Korea’s national defense expenditure	Military Budgets	True value of national defense expenditure from each fiscal year	Military Balance
DEA Input	Military Budgets	Military Budgets	The true value of each nation’s annual national defense expenditure Equilibrium solution to Taiwan and South Korea’s national defense expenditure game	Military Balance
DEA Output	Military Capability	Defensive military strength	No. of active troops / Area of territory	Military Balance
		Organizational structure ratio	No. of troops in the navy and air force / No. of active troops	

Source: Compiled by the author

TABLE 2 Growth rate of Taiwan and South Korea’s TFP over the years

YEAR	Taiwan TFP	Korea TFP	YEAR	Taiwan TFP	Korea TFP	YEAR	Taiwan TFP	Korea TFP
1957	0.009896	N/A	1974	-0.008400	0.001876	1991	0.005441	0.012674
1958	0.013687	N/A	1975	-0.006950	0.001989	1992	-0.001860	0.013552
1959	0.043149	N/A	1976	0.017222	0.002576	1993	0.003498	0.023820
1960	0.309356	N/A	1977	-0.199310	0.003119	1994	0.001768	0.027191
1961	-0.299710	N/A	1978	0.138126	0.004244	1995	0.006231	0.028615
1962	0.006221	0.000455	1979	0.002583	0.005232	1996	0.007995	0.029561
1963	-0.000530	0.000721	1980	-0.000170	0.005501	1997	0.000349	0.029903
1964	0.015206	0.000787	1981	0.006737	0.006450	1998	7.57E-05	0.030825
1965	-0.000290	0.000779	1982	0.005539	0.006732	1999	0.004597	0.029822
1966	-0.001640	0.000727	1983	0.012291	0.007061	2000	0.003263	0.031987
1967	-0.001260	0.000778	1984	0.013269	0.007394	2001	0.011468	0.032786
1968	-0.001480	0.000814	1985	0.010656	0.008055	2002	0.004547	0.041107
1969	0.002606	0.000987	1986	0.007527	0.008787	2003	0.005446	0.043515
1970	0.008217	0.001022	1987	0.006985	0.007861	2004	-0.004080	0.048496
1971	0.003073	0.001213	1988	0.000839	0.009297	2005	0.000846	0.055175
1972	0.008428	0.001231	1989	0.003201	0.010994	2006	0.008457	0.033637
1973	0.011758	0.001466	1990	0.001867	0.011099	2007	0.006893	0.036223

TABLE 3 TWN and KOREA’s economic growth, national defense budget growth and national defense budget in equilibrium

YEAR	TWN economic growth	TWN defense budget growth	TWN defense budget in equilibrium growth	YEAR	KOREA economic growth	KOREA defense budget growth	KOREA defense budget in equilibrium growth
1957	0.072545917	0.290863891	N/A	1957	N/A	N/A	N/A
1958	0.082989412	**0.272026962 >**	0.152587780	1958	N/A	N/A	N/A
1959	0.250030176	0.021953066 <	0.401285100	1959	N/A	N/A	N/A
1960	0.062033576	0.085555556 <	0.448308010	1960	N/A	N/A	N/A
1961	0.063161450	**0.141589901 >**	-0.189424230	1961	N/A	N/A	N/A
1962	0.079037003	0.038254632 <	0.095756520	1962	0.990185676	0.1818181820	N/A
1963	0.093539337	0.078583765 <	0.113319234	**1963**	0.180601093	**0.407692308 >**	0.3054606894
1964	0.121989402	0.083800374 <	0.148879618	1964	0.031437655	-0.185792350 <	0.1405177040
1965	0.111350000	**0.148731839 >**	0.130432695	1965	0.041095890	-0.020134228 <	0.1511965774
1966	0.089132160	**0.163879957 >**	0.103818020	1966	0.282051282	0.232876712 <	0.4176342859
1967	0.107112471	0.059397735 <	0.126187210	1967	0.014000000	0.1111111111 <	0.1212359763
1968	0.091707273	**0.155163421 >**	0.104627261	1968	0.432692308	0.170000000 <	0.5842059706
1969	0.089483757	**0.156369930 >**	0.097133438	**1969**	0.117449664	**0.273504274 >**	0.2356260458
1970	0.113708893	**0.155658229 >**	0.130563942	1970	0.156505343	0.117449664 <	0.2788111432
1971	0.128951059	0.092234923 <	0.145911640	**1971**	0.004025765	**0.243243243 >**	0.1102055473
1972	0.133171674	**0.274836315 >**	0.150889331	1972	0.295825771	0.050724638 <	0.4328658957
1973	0.128327204	**0.175226464 >**	0.143780371	1973	0.377564979	0.094252874 <	0.5232490649
1974	0.011620503	**0.147482881 >**	0.010176734	**1974**	0.071534274	**0.558823529 >**	0.1848541082
1975	0.049283886	**0.271868534 >**	0.053163510	1975	0.308452776	0.270889488 <	0.4468274023
1976	0.138606148	0.143870983 <	0.152885964	**1976**	0.292779487	**0.590668081 >**	0.4294967868
1977	0.101896859	**0.261094507 >**	0.111211641	1977	0.476442274	0.355333333 <	0.6325826647
1978	0.135939160	**0.222974077 >**	-0.130110624	1978	0.267411938	0.272011805 <	0.4014464418
1979	0.081739434	**0.136440180 >**	0.100867603	**1979**	0.040450614	**0.244779582 >**	0.1504827962
1980	0.073012254	0.058777181 <	0.089224417	1980	0.116941735	0.078285182 <	0.2350630751
1981	0.061626927	**0.380613344 >**	0.076521832	**1981**	0.065927655	**0.234514549 >**	0.1786540662
1982	0.035512312	**0.216423434 >**	0.035240087	1982	0.066424682	0.015169195 <	0.1792036303
1983	0.084469087	**0.145530063 >**	0.099291801	1983	0.037831228	0.013103448 <	0.1475863001
1984	0.105996571	-0.056096026 <	0.137146334	1984	0.107484599	-0.016791468 <	0.2246058036
1985	0.049525101	**0.103005132 >**	0.056707013	1985	0.105172041	0.066466651 <	0.2220486850
1986	0.116370175	0.129620079 <	0.137396495	**1986**	-0.067712359	**0.105172041 >**	0.0308810222

1987	0.127448658	0.044422953 <	0.157953231	1987	0.208852271	0.021930683 <	0.3366934722
1988	0.078404261	0.040418331 <	0.093708566	1988	0.249736401	0.208852271 <	0.3819013746
1989	0.082324736	**0.152849121 >**	0.100388592	1989	0.090473333	0.008559201 <	0.2057955453
1990	0.053949764	-0.091632639 <	0.062451923	1990	0.185944118	0.043061449 <	0.3113627310
1991	0.075539582	0.082084047 <	0.086313290	1991	0.057551487	-0.029682085 <	0.1693920492
1992	0.074874594	0.045174743 <	0.084275142	1992	0.761070153	0.116304348 <	0.9473107612
1993	0.070138327	0.033468494 <	0.082005017	**1993**	0.160944108	**0.668382251 >**	0.2837189261
1994	0.071080235	-0.046310684 <	0.076838609	1994	0.078186351	0.064198766 <	0.1922091069
1995	0.064240461	-0.024095919 <	0.073199733	1995	0.048082147	0.110858665 <	0.1589212469
1996	0.061022509	0.024107901 <	0.070445851	**1996**	0.002363874	**0.140559983 >**	0.1083680717
1997	0.063667142	0.040602237 <	0.067732993	1997	-0.047236375	-0.051817957 <	0.0535223833
1998	0.043294441	0.022551187 <	0.048508201	1998	-0.034703761	-0.156123647 <	0.0673803737
1999	0.053191402	0.035529984 <	0.057850582	1999	0.107591945	-0.065842349 <	0.2247244474
2000	0.057811474	**0.416180218 >**	0.063874417	2000	0.010723431	0.033752482 <	0.1176116931
2001	-0.022240002	**-0.330831061 >**	-0.025733376	2001	0.260766291	0.010723431 <	0.3940976289
2002	0.039444946	-0.032263687 <	0.043448314	2002	0.071970566	0.035629454 <	0.1853359783
2003	0.033349860	-0.008641403 <	0.035702656	2003	0.120770966	0.118577982 <	0.2392972118
2004	0.056985754	0.020398302 <	0.061819889	2004	0.143460272	0.120770966 <	0.2643860083
2005	0.036580777	-0.018024434 <	0.039277152	2005	0.12660084	0.238748628 <	0.2457436203
2006	0.071874117	-0.026340865 <	0.077778525	2006	0.102463744	0.213262443 <	0.2190539232
2007	0.060261393	**0.210544041 >**	0.064996197	2007	0.020234692	0.078839521 <	0.1281288010
2008	-0.038697133	**0.092862193 >**	-0.043577266				

TABLE 4 TFP changes for 7 DMU'S defense expenditure from(1961- 2008)

Malmquist	**TWN**	**KOR**	**TWN-E**	**KOR-E**	**JAPAN**	**INDIA**	**ISRAEL**	**Average**
1961=>1962	0.82851697	0.83364356	0.55768673	0.81570016	5.35684164*	0.86685863	0.78773778	1.43528364
1962=>1963	0.87258885	0.71322494	0.89222174	1.26732620*	1.00136166*	1.02959228*	0.78560235	0.93741686
1963=>1964	1.01766082*	1.34182083*	0.99104932	0.79305335	0.83367332	0.80670675	0.77138722	0.93647880
1964=>1965	0.76240512	1.07421354*	0.82108978	1.04819579*	0.90871915	0.93593622	0.79346945	0.90628986
1965=>1966	0.89198885	0.91958945	0.92544024	0.63997835	0.88862779	2.90116962*	0.51100404	1.09682833
1966=>1967	0.94913806	0.86179165	0.84683570	1.15783483*	0.88576779	0.87206111	0.96544276	0.93412456
1967=>1968	0.83560911	0.86903750	0.92880246	0.72683807	0.91473661	0.90714666	1.47282929*	0.95071424
1968=>1969	0.90899662	0.71898084	0.91605832	1.08138539*	0.92939380	1.00892805*	0.68641844	0.89288021
1969=>1970	0.81463656	1.01370804*	0.79019744	0.58913857	0.80991231	1.70682323*	0.40004413	0.87492290
1970=>1971	0.94621844	0.75557976	0.88230194	0.88722351	0.87166812	0.79543161	0.72439353	0.83754527
1971=>1972	0.72630966	0.86151909	0.84407311	1.18162806*	0.66298598	0.90690002	5.28534594*	1.49553741
1972=>1973	0.85600523	1.21364226*	0.79398993	0.52898583	0.75851892	0.78066984	0.96971213	0.84307488
1973=>1974	0.85068210	0.57279704	0.94016439	0.79229119	1.02373032*	1.02954825*	0.26515596	0.78205275
1974=>1975	0.79104872	0.94074973	0.84109671	0.55046339	0.86576816	0.91842105	1.23819338*	0.87796302
1975=>1976	0.83175202	0.48668907	0.68346250	0.87948446	0.95352217	0.93984182	0.28523224	0.72285490
1976=>1977	0.77609045	0.63816733	0.88763931	0.97386363	0.90959675	0.80499343	1.04683472*	0.86245509
1977=>1978	0.84256470	0.83455558	1.03589548*	0.82378833	0.62270011	0.96638655	1.64348120*	0.96705314
1978=>1979	1.00060771*	1.06166457*	0.74074791	1.28424807*	0.85607679	0.95967742	3.43569020*	1.33410181
1979=>1980	0.76785468	0.97476032	0.72290772	0.85100638	1.13382232*	0.91976925	0.20770801	0.79683267
1980=>1981	0.74547331	0.82914205	0.98346409	1.02232167*	0.86871041	0.85937500	0.68917711	0.85680909
1981=>1982	0.84577854	0.97198366	0.77450244	0.92498619	1.01727388*	0.97338403	1.38263564*	0.98436348
1982=>1983	0.87295832	0.95937267	0.77455478	0.99061136	0.85343073	0.93319963	0.87076928	0.89355668
1983=>1984	1.10509488*	1.01707824*	0.88063296	1.11155942*	1.00999066*	0.87828011	1.42097020*	1.06051521
1984=>1985	0.83168731	0.90226676	1.04514538*	0.67147379	0.90089338	0.91790909	1.62242264*	0.98454262
1985=>1986	0.84537711	0.92990155	0.80461547	0.97274767	0.62842261	0.88067855	0.58219399	0.80627671
1986=>1987	0.95746651	1.16543480*	0.86968511	1.12911860*	0.83169182	0.78304994	1.11805864*	0.97921506
1987=>1988	1.12413962*	0.91764390	1.18519386*	1.01133923*	0.85184468	0.94051386	0.86404851	0.98496052
1988=>1989	0.70929072	1.26554779*	0.78120192	1.37224846*	0.99037977	1.10332967*	0.91146436	1.01906610
1989=>1990	1.00449862*	0.70305861	0.81957019	0.45384991	1.09146143*	0.96153109	0.88499169	0.84556593
1990=>1991	0.92414263	1.03059006*	1.07673745*	1.35888500*	0.84077686	1.02063923*	1.02367377*	1.03934929
1991=>1992	0.94154575	1.17177441*	0.91111073	0.84420627	0.95726043	1.32740741*	1.18742000*	1.04867500
1992=>1993	1.15475218*	0.59938302	1.14426735*	0.96666448	0.87207164	1.06635071*	1.27268516*	1.01088208
1993=>1994	1.00823030*	0.93967408	0.93474803	0.94126327	0.94000752	0.86712329	0.90993911	0.93442651
1994=>1995	1.12930285*	0.90020453	1.15959863*	0.92240598	0.76111213	0.99323467	1.12345197*	0.99847297
1995=>1996	0.96469504	0.80433411	1.01495715*	0.95496498	1.22699060*	0.96666667	1.05859528*	0.99874340
1996=>1997	0.97270127	1.07505225*	0.99077508	1.05274529*	1.03960427*	0.84848485	1.05405762*	1.00477438
1997=>1998	0.91205782	1.18500773*	0.91269090	1.17307904*	1.23324496*	1.15164410*	0.84102373	1.05839261
1998=>1999	0.99664071	1.07048312*	1.00036499*	0.77549889	0.87534732	0.90047415	1.05859139*	0.95391437
1999=>2000	0.69163903	1.02969708*	0.94054711	1.23864095*	0.87755059	0.64911508	0.93145209	0.90837742
2000=>2001	1.49439094*	0.98794175	0.77092534	0.92220033	1.15291466*	0.84976347	0.67872804	0.97955207
2001=>2002	1.03333934*	0.96559633	1.39552080*	0.87625301	0.94600939	0.99358974	0.95339788	1.02338664
2002=>2003	1.12670735*	0.89399221	1.16280496*	0.90559525	1.03314836*	1.26704480*	1.51710498*	1.12948541
2003=>2004	0.60730267	0.89224296	0.60974510	0.82494039	0.91692321	0.83505155	0.98214707	0.80976471
2004=>2005	1.23184350*	0.73567268	1.21112436*	0.81166588	1.01008949*	0.88181818	1.10052140*	0.99753364
2005=>2006	1.02705348*	0.82542250	0.98203342	1.12440193*	1.01240114*	0.93617021	1.26318381*	1.02438093
2006=>2007	0.82607486	0.92786638	1.03935855*	0.89156244	1.01389354*	0.67859867	0.70741015	0.86925208
2007=>2008	0.91502845	0.94622584	0.87968840	0.98591256	0.95904219	1.15794292*	1.00402253*	0.97826613
Average	0.92063591	0.92188779	0.91696224	0.93837395	1.01914705*	0.99317516	1.09127278*	0.97163641

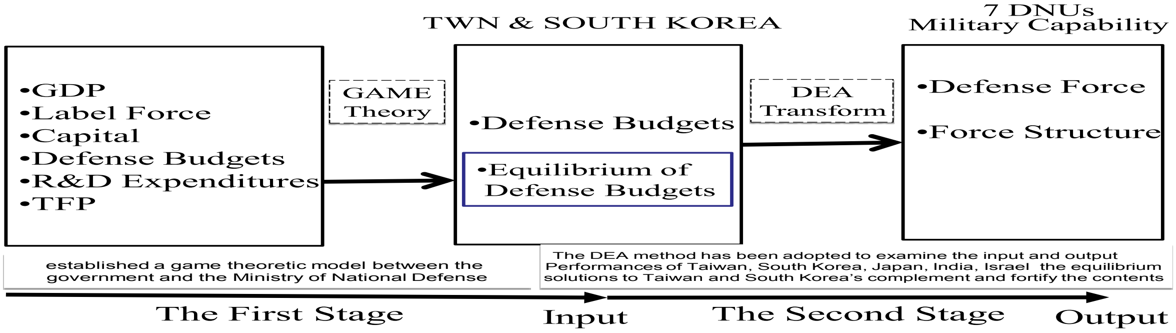

TWN & SOUTH KOREA
7 DNUs
Military Capability
•GDP
•Label Force
•Capital
•Defense Budgets
•R&D Expenditures
•TFP
GAME Theory
•Defense Budgets
•Equilibrium of Defense Budgets
DEA Transform
•Defense Force
•Force Structure
established a game theoretic model between the government and the Ministry of National Defense
The DEA method has been adopted to examine the input and output Performances of Taiwan, South Korea, Japan, India, Israel the equilibrium solutions to Taiwan and South Korea's complement and fortify the contents
The First Stage
Input
The Second Stage
Output

Figure 1 The Research Framework

Layered Multi-Modal Network Analysis of Textual Data for Improved Situation Awareness

Peter M. LaMonica and Todd V. Waskiewicz
Air Force Research Laboratory, AFRL/RIED
525 Brooks Road, Rome, NY 13441-4505

Abstract

A team of researchers at the Air Force Research Laboratory's Information Directorate investigated combining multiple textual databases that consist of relational data to determine from different data networks if there is an improvement in a user's situation awareness. Currently, users manually process heterogeneous data types and are required to make mental correlations to merge databases and derive patterns in space and time across the different network types. This is because the vast majority of this data does not exist independently, as much of it is related and connected across networks. The researchers experimented with combining heterogeneous relational data types to determine the impact on situation awareness. The findings of this effort support future research to establish a layered multi-modal network analysis approach that would connect numerous data types for analysis purposes.

Keywords: Layered multi-modal network analysis, social network analysis, link analysis, situation awareness, and pattern matching.

1. Introduction

This initial research effort sought to combine social networks and communication networks into one common operating picture. A social network consists of nodes (e.g. people) and edges (relationships), and is the study of social relationships, mainly people and how they are related. Social Network Analysis (SNA) consists of social network metrics that determine the level of importance of a particular node, or person based on relationships in a network. The goal of the research was to apply social network analysis metrics to determine high value targets (HVTs) within a multi-layer network derived from the combination of social and communication networks. The objective was to develop the software infrastructure for a SNA system to include real-time analysis of audio sources. This effort represents an initial step towards the analysis of multi-layered and multi-modal networks. Multi-modal networks contain numerous node types. An example would be a network that contains nodes that are people, locations, and organizations.

2. Approach

The SNA approach leverages two key data inputs: an existing social network, and metadata that describes the communications from the audio sources. The social network was created with existing relational data and contained individuals common to the audio dataset. This effort did not analyze raw audio, rather it leveraged the descriptions of the audio. The metadata included speaker identification, duration of the communication, frequency, and location. It was used to create a communications network that was integrated with the existing social network. The audio metadata also served as attributes for those relationships. Several issues needed to be resolved in order to integrate these diverse data types. The database that was created to store the various types of data needed to standardize and normalize the existing data. Data standardization created a common entity and attribute definition, making all data conform. Data normalization was used to reduce redundancies in the data. For example, each person should only have one record in the database. In addition, the database needed to be dynamic in order to allow real-time updates as new information was published.

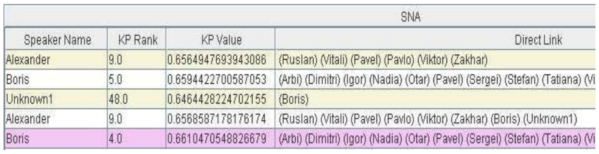
The focus of this research was to apply social networking techniques to a new domain where these methods are not currently used in order to determine if the application of SNA methodologies provided beneficial information by improving situation awareness for end users [1]. Based on end user needs, the researchers determined which SNA techniques would provide optimal results. Those needs included determining groups, identifying the most influential nodes, and conducting link analysis techniques. Thus, the researchers decided to use a group detection algorithm based on 21st Century Technologies' Best Friends Group Detection Algorithm [3] and the Key Player algorithm [4] as part of their approach. As a result, the SNA architecture can analyze data to determine how individual nodes are connected in a network of interest. This SNA methodology is able to discover the key players in the given network, determine what group(s) a node belongs to, and also what other nodes a particular node is directly connected to. Given this information, the end user would gain valuable insight and awareness of a situation.

3. Experiment

The research team conducted an experiment to test and evaluate the SNA methodologies that were being proposed. The team developed a scenario and dataset based on a smuggling event that consisted of people, events, materials, and locations. Since the core focus of this research was on the social aspect of the battlespace, the team concentrated on the persons that were contained in this data. Relationships were to be derived from previously known associations, as well as new communication between nodes. The team used this as a foundation for future data and this represented a historical context for this experiment. During the experiments, all components of the SNA implementation were tested – direct connections, key player values and ranking, and group detection. The experiment was designed to first test functionality and utility, and then performance.

4. Results

The results from the experiment clearly demonstrated the benefits of utilizing social network analysis. In terms of functionality and utility, the methodologies were able to return direct connections, key player values and rankings, and any related groups. The results were published to a main interface as new activity occurred (see Figure 1 below). As a result, the end user does not need to search or provide any input. Information is provided in a near real-time fashion (as new calls are made). Users are able to see the history of calls that have been made providing an opportunity for forensically analyzing the data. Collectively, this approach made the SNA techniques very efficient and effective for users to quickly obtain information about the network.



SNA

Speaker Name	KP Rank	KP Value	Direct Link
Alexander	9.0	0.6564947693943086	(Ruslan) (Vitali) (Pavel) (Pavlo) (Viktor) (Zakhar)
Boris	5.0	0.6594422700587053	(Arbi) (Dimitri) (Igor) (Nadia) (Otar) (Pavel) (Sergei) (Stefan) (Tatiana) (Vi
Unknown1	48.0	0.6464428224702155	(Boris)
Alexander	9.0	0.6568587178176174	(Ruslan) (Vitali) (Pavel) (Pavlo) (Viktor) (Zakhar) (Boris) (Unknown1)
Boris	4.0	0.6610470548826679	(Arbi) (Dimitri) (Igor) (Nadia) (Otar) (Pavel) (Sergei) (Stefan) (Tatiana) (Vi

Figure 1. A portion of the SNA interface that displays speakers and their respective key player rank, value and direct links

From a performance standpoint, all of the audio messages were correctly processed and accessible via the interface. All new people were published to the database as well as the accurate computation of direct connections and key player values and ranks. This all occurred in near real-time, which provided results to the end user within roughly a second of a call being placed in the field.

The experiment also proved to be a successful integration of two heterogeneous network data types. An existing social network of relational data was augmented with information from communication networks. Records of communications were used to add individuals and relationships to the network. This provided a more accurate representation of the network and results of calculated metrics and algorithms. Key player algorithm values and ranks, and groups were more precisely reported to the interface than using the standard relational network alone.

Collectively, this equates to improved situation awareness as these SNA methods identified previously unknown aspects of the underlying social network. Prior to implementing the SNA components, users had no insight into these aspects of either network (e.g. key players, groups, and important links). Now, users can quickly learn a person’s influence, how they are connected, and what groups they belong to. This helps the user to perceive and comprehend data elements within their scenario, which is imperative to improving situation awareness [5]. For example, if a person has connections to a known insurgent, then the user can infer that this person has access to their resources. Ultimately, this provides the end user with greater comprehension of the network and allows them to gain this insight immediately. This information helps users by directing them where to look and by providing them information on the network (entities and relationships).

5. The Next Step – Layered Multi-Modal Network Analysis

These results provided the researchers with positive insight into combining heterogeneous network types to gain a better understanding of the battlespace. Despite this effort being limited to communication and social network data, the results demonstrate that there is potential for combining other data sources and types. On a much grander scale, this concept could be applied to all networked data (per a given scenario), so that this data can be collectively analyzed to discover how different data types correlate and potentially make inferences across networks.

This has generated research ideas for experimental and theoretical development of technologies to fill requirements for the development of a layered multi-modal network analysis (LMMNA) approach. This is especially important since understanding the structure and dynamics of networks are of vital importance to winning the global war on terror [6]. Current analysis of network data occurs independently within one or two particular domains. However, to fully understand the network environment, users must be able to simultaneously investigate interconnected relationships of many diverse network types as they evolve spatially and temporally. It is important to realize that no single network exists independently of others. This creates what is being defined here as a layered multi-modal network approach. Each layer represents a diverse network data type and these layers can be connected through nodal and edge similarities as is depicted in Figure 2 below.

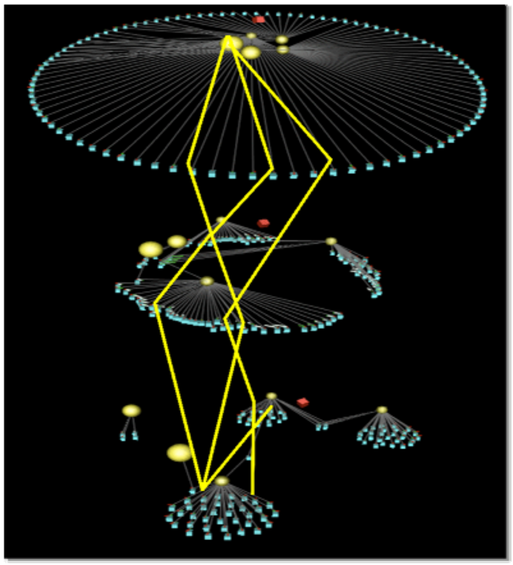


Figure 2. Layered Multi-Modal Network Analysis

The focus of this future research would be to develop a layered multi-modal networking approach that the user can exploit and shape for their mission. A layered multi-modal network analysis (LMMNA) approach will be able to assemble previously disconnected networks (e.g. social networks, financial transactions, computer networks, etc.) into a common battlespace picture. This will provide the user with timely situation awareness, understanding and anticipation of threats, and support for effective decision-making in diverse environments. Combining numerous networks will require various data transformation techniques such as standardization, normalization, as well as entity resolution.

The concept behind LMMNA is derived from social network analysis and link analysis techniques. The collective network is layered because it would consist of many interconnected networks that were previously distinct. It is also multi-modal because the network will incorporate many different entity types (e.g. people, organizations, locations, objects). Having this data in a collective network will allow users to apply techniques derived from traditional SNA and link analysis measures. In addition, this may introduce new areas of research for metrics, entity resolution and pattern learning as the user will ultimately have to analyze data of multiple types.

There are many objectives to proposing this research. First, an underlying unified network model is necessary that will allow multimodal and multidimensional layered network analysis to be possible [7]. This framework would serve as the connecting language and translation mechanism between the heterogeneous data networks. This would allow any network in any domain to be mapped and connected. Ultimately, users would analyze and visualize these collective networks.

Once the data is in a common network, new network analysis metrics will need to be designed in order to measure the importance of entities, groups, and their relationships across networks. Currently, social network analysis metrics are designed to analyze only social relations. Also, current SNA metrics do not analyze attribute data, rather they measure connectivity (how well each node is connected to other nodes in the network). Thus, it is unclear how these metrics will be applicable. In addition, new threat patterns will need to be developed per given scenarios that can operate across networks, time, and entity types. The end user has previously done this manually.

Initially, the research team is focusing on developing an approach to apply social network analysis metrics based on specific measures (e.g. eigenvector centrality, group detection) and visualization technology to the collective network. At this initial stage in the research, the team has identified two risks with this approach – scalability and visualization. However, the researchers feel that these two risks can be mitigated by providing only the data that user needs for their scenario. This can be accomplished either through selecting which data sources the user needs or through group detection algorithms to determine subgraphs related to the user’s scenario.

5.1 LMMNA – An Example

An example of a layered multi-modal network graph is shown in Figure 3. There are many important aspects of this graph to note. First, one can see that this graph consists of four distinct networks or subgraphs, lettered *a, b, c*, and *d*. These subgraphs would represent the diverse network data types that a user has available. Examples could include social networks, computer networks, and financial transactions. Next, notice that nodes *x* and *y* belong to multiple subgraphs. These nodes demonstrate how specific nodes can be identified in several different subgraphs. Entity resolution is an important aspect of LMMNA for this reason – when two similar nodes exist in distinct networks, there must be a determination if these nodes are the same or if they are separate entities [8]. In this case, it was determined that nodes *x* and *y* were the same across the subgraphs and connect their collective networks.

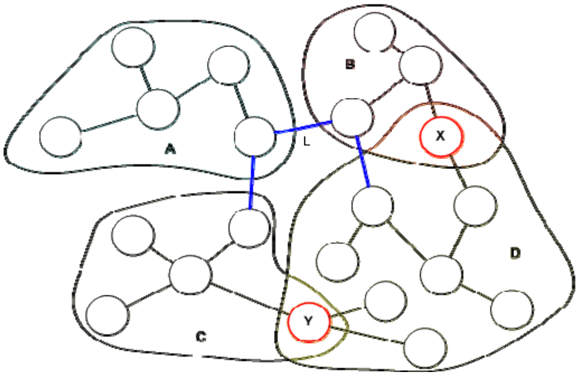


Figure 3. A layered multi-modal network graph

In addition to nodes connecting subgraphs, relationships (edges) can also connect these diverse networks as an edge can connect one or more data types to another. For example, see edge *l* in Figure 3 and note how this relationship connects subgraphs *a* and *b* together. In a given scenario, this edge could connect a social network (consisting of people) to a computer network (consisting of computer devices on a local area network) and could represent that a specific person has used a particular computer device.

From the way that nodes and edges can connect these diverse network datasets (subgraphs) this can immediately allow users to learn and understand relationships in and across databases that they never knew existed. This can also help to infer new relationships, help uncover unknowns, and identify new groups and patterns across networks.

5.2 Significance of LMMNA

Developing a layered network approach to network analysis will eliminate the current stovepipes and manual association that exist today. This approach would create an environment where users can visualize and analyze how numerous networks (entities and their relationships) are interconnected. This will allow the user to have all network data in a layered network application where they can correlate and analyze activity and time occurrence in related networks to better understand the battlespace. From this information, the user would be able to derive the most influential nodes and relationships, discover groups, and learn and predict the behavior of how these different network types interact.

LMMNA will allow users to identify and detect emerging threats from large collections of data spatially and temporally. The result will be a dramatic improvement in situational awareness across multiple domains and significantly advance the current state-of-the-art network analysis capabilities by having an accurate and complete multi-dimensional understanding of the layered, dynamic network environment. This work will provide the user with the ability to identify high value targets in multiple network types and anticipate their role and activity.

6. Conclusion

This preliminary research effort proved very successful, as it was able to develop a SNA methodology that could combine different types of network data and ultimately analyze this data over the collective network. These results demonstrate that this same approach can be applied on a larger scale to include numerous diverse relational datasets. The researchers proposed a novel technique known as layered multi-modal network analysis (LMMNA) and its potential impact on end users can be prodigious.

Acknowledgements

Work described in this technical report was funded as in-house research by Air Force Research Laboratory’s Information Directorate (AFRL/RI) at Rome, NY.

Reference

[1] Peter LaMonica and Todd Waskiewicz. Developing an intelligence analysis process through social network analysis. *Proceedings of SPIE, Evolutionary and Bio-Inspired Computation: Theory and Applications II*, 6964, May 2008.

[2] Metzler, J. M., Linderman, M. H., Seversky, L. M. (2009, October). "N-CET: Network-centric exploitation and tracking," *Military Communications Conference, MILCOM 2009,* IEEE, pp.1-7.

[3] Wan, J., Moy, M., Darr, T., Coffman, T., Snyder, J., Hollinger, M., and Thomason, B. (2006, Fall). Key Elements of an evidence detection system. *AAAI Fall Symposium on Capturing and Using Patterns for Evidence Detection,* pp. 62-67.

[4] Borgatti, S.P. (2006). Identifying sets of key players in a network. *Computational, Mathematical and Organizational Theory, 12*(1): 21-34.

[5] Jones, D.G., Bolstad, C.A., Riley, J.M., Endsley, M.R., & Shattuck, L. (2003). Situation awareness requirements for the future objective force**.** *Proceedings of the ARL Collaborative Technology Alliances Conference,* Adelphi, MD: ARL.

[6] Hoffman, F.G. (2006). Complex irregular warfare: The next revolution in military affairs. *Orbis, 50*(3), 395-411.

[7] Hall, C.M., McMullen, S.A., Hall, D.L., McMullen, M.J., & Pursel, B.K. (2008). Perspectives on visualization and virtual work technologies for multi-sensor data fusion. *Information Fusion, 2008 11th International Conference.*

[8] Knoblock, C.A., Ambite, J.L., Ganesan, K., Muslea, M., Minton, S., Barish, G., Gamble, E., Nanjo, C., See, K., Shahabi, C. & Chen C.C. (2007). EntityBases: Compiling, organizing and querying massive entity repositories. *Proceedings of the International Conference on Artificial Intelligence (ICAI'07).*

Enhancing the Delivery and Reception of Information Objects Over Deprived Channels and Computational Devices – Part III

Gerard T. Capraro[1], Christopher T. Capraro[1], Ivan Bradaric[1], Wayne Perrigo[1], Daniel Klockowski[1], and John Spina[2]
[1]Capraro Technologies, Inc., 2118 Beechgrove Place, Utica, NY 13501
[2]US Air Force Research Laboratory, Information Directorate, Rome, NY 13441

Abstract: *We wish to extend the basic concepts of information theory and modify information objects and yet maintain their inherent meaning. Our approach is to reduce the quality of the original messages or information objects without severely degrading a message's intent before shrinking the resultant messages for transmission. Our motivation is to enhance and improve the efficiency in the delivery and reception of information objects from computational devices with limited bandwidth and processing capability. Three instantiations are described dealing with Microsoft Office and portable document format (PDF) files, email systems, and gathering information using smartphones.*

1. Introduction

The science and technology vision of the USAF is to "Anticipate, Find, and Fix, Track, Target, Engage, and Assess— Anything, Anytime, Anywhere" (AF2T2EA4). The AF2T2EA4 vision will require that the military manage information delivery both dynamically and securely, getting the right information to the right people who use multiple devices in multiple locations as well as gathering information from them. Managing information delivery includes gathering, storing, and distributing information objects (any collection of digital multimedia data) throughout the Department of Defense (DoD). The US military must provide the right information to its warfighters in a timely manner. This must be done securely, to a large variety of mobile devices across different networks, in different formats, and using multiple protocols. Mobile devices should be used to receive information from the network, gather data, and send it to the network.

In two previous papers [1- 2] we provided a brief review of information theory followed by a method for obtaining a user's perception of the amount of information contained in images. We also discussed how this information can be used to modify images before they are encoded (using information theory) based upon the terminal devices and the bandwidth available. We also demonstrated two instantiations of this software. One instantiation dynamically changed the size of images for the real estate market based upon the bandwidth connection and the properties of the terminal smartphone. Another instantiation demonstrated the capability of shrinking Microsoft Office and PDF files and compared our results with the best commercial software product currently on the market. In section two we will discuss our latest instantiation of our work called PocoDoc and how users may use this capability free of charge. We will also provide an example of its capability. The third section presents how PocoDoc can be used with an email server to shrink attachments to be sent to users with smartphones thereby retrieving information and minimizing the use of their bandwidth. Section four will discuss how a user can gather information for first responders using their smartphone, shrink its images and send the results to the cloud for further processing. Section five will present a summary, conclusions and future work.

2. PocoDoc.Com

Visiting PocoDoc.Com will allow one to send us a Microsoft Office and/or a PDF file for shrinkage. The user uploads the file they wish to have shrunken, their email address, and the degree of shrinkage they wish on a scale of 1 to 19. Once the file is shrunk the system will email the user with the shrunken file. The default shrunken value is 10 and this will usually reduce the file between 0 to 50% while maintaining the information contained in the file. If the file contains text only then the reduction is minimal.

If the file contains many images then the shrinkage can be up to 50% with a shrunken value of 10. Please visit our web site and test it's capability. We will be releasing a new version, which will also be free, called PocoDoc Lite in which a portion will run on your computer and will eliminate the step of having to enter your email address. Eventually we will be offering a standalone desktop version so one does not have to send their files to our web site.

An illustration of PocoDoc is shown in Figure 1. The original PowerPoint document was 4.4 MB and the shrunken PocoDoc version was 671 KB. Comparing two of the images shows that the amount of information lost is not directly proportional to the size in reduction, i.e. approximately 85%. It's not the size of the document that counts but the amount of information.

3. PocoBox an Email Instantiation

We have developed a software instantiation of PocoDoc for the real estate industry. The software is web based and modifies the original information object (IO) based upon the type of IO, the bandwidth available, the user's device attributes (e.g. screen size, color depth, stylus), the user's quality of experience and the requested time they are willing to wait for the IO to be delivered. A demonstration of a system with text and images can be tried at http://mobile.PrudentialCarucci.com/.

We are currently extending PocoDoc as an appliance to work with one's email server called PocoBox. See Figure 2 for a view of this instantiation. When an email is sent with an attachment a service oriented architecture (SOA) call will be made to PocoDoc which will strip out all the images, video, and audio files and shrink them. They will then be put back into the original IO and sent back to the email server which will then pass them onto its ultimate destination. The user can then download the shrunken version if they receive it on their smartphone or the original document if they are connected with their desktop computer and are not bandwidth restricted. We are currently investigating different instantiations of PocoBox depending upon the email server and the different requirements that an organization may have.

4. PocoDoc and First Responders

The last section dealt with sending information to a smartphone with limited bandwidth. We have also developed an instantiation of PocoDoc that will allow a smartphone user to gather information and send it to the Cloud. One can envision the military need for this technology such as for a forward observer, or for a soldier gathering data for battle damage assessment, or for enemy interrogation and for a medic or first responders tending to the needs of injured comrades. We have elected to demonstrate this instantiation by addressing the need of a first responder for a natural or manmade disaster.

The application is built for the Android smartphone. It requres a minimum set of software to be loaded on the smartphone and allows a user to take multiple images of the victim, gather their vtal signs, and place a 2D barcode on the victim. Once the 2D barcode is scanned with our built in software scanner, multiple medical personnel can access the demographics and vital data on the victim. Figure 3 depicts the first screen seen by the smartphone user and the ability to save an image to the victim's file. Figure 4 shows one of the screens illustrating the victim's demographics and the tabs to view the treatment performed, information about the medical provider, and the images taken to date. The first responder has a number of unused 2D barcode wrist bands to attache to each of the victims. By scanning these barcodes a smarphone can automatically retrieve the stored information on the victim. The desktop version allows a user to retrieve records based on the unique identifier placed on the barcode. In the example shown it is mrn_21.

Once the first responder gathers the information for the victim and wishes to send it to the Cloud the software automatically sends parameters on all the images and sends it to a centeral server where PocoDoc is resident. PocoDoc retrieves the parameters about the images and computes a new set of parameters for the software on the Android to shrink the images. This process is very fast and requires a minimum amount of bandwidth. Once the Android shrinks the images then the total record is sent to the Cloud while minimizing its use of available bandwith.

In this instantiation we have built-in security and maintenance. First all transmissions between the smartphone and the Cloud are all sent using the https protocol. Second, note that if you use a standard barcode reader on the barcode shown in Figure 4 you will not be able to access the victim's information. Third, even if one obtains a copy of this software for their Android, it will not work. At the central server PocoDoc accesses a database that has the International Mobile Equipment Identity (IMEI) of each smartphone that has our registered loaded software. When the barcode reader software directs the browser to our web site the smartphone will not match one of the unique identifiers in the IMEI database, and access will not be allowed. Fourth, if one loses their smartphone a call to our host site will direct an administrator to remove the IMEI from the database and the smartphone will be rendered useless since none of the data are stored on the device and access will be restricted. Lasty, our system checks the software running on the smartphone and if necessary it will update its software when the user accesses PocoDoc.

5. Summary, Conclusions and Future Work

The major thrust of our work is to deliver and receive information from computational devices with limited bandwidth and processing capability. We have described three different instantiations of PocoDoc. One being accessable via a web browser, one being developed to work with an email server, and one built on an Android smartphone for gathering of information for first responders. We feel that this technology is necessary as our military becomes more mobile and will always require more information. In the future we need to extend our information objects to include audio and video formats and integrate these results with quantitative measures to dynamically change an IO based upon an individual's quality of experience, i.e what is acceptable for them given their time constraints and the information they require.

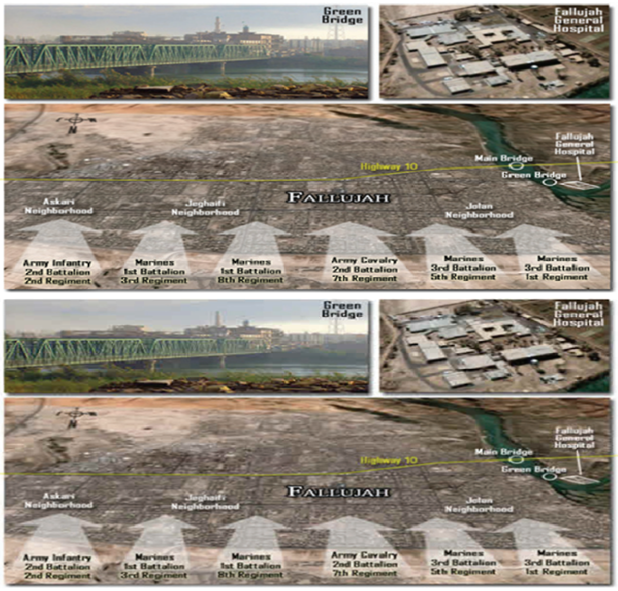
Acknowledgements

The authors would like to thank the US Air Force for supporting this work under contract FA8750-09-C-0008.

References

1. G. T. Capraro, C. T. Capraro, I. Bradaric, J. M. Scherzi, and J. Spina, "Enhancing the Delivery and Reception of Information Objects Over Deprived Channels and Computational Devices", Proceedings of the International Conference on Artificial Intelligence, Volume I, pp 429 – 434, July 13-16, 2009.

2. G. T. Capraro, C. T. Capraro, I. Bradaric, J. M. Scherzi, D. Klockowski, and J. Spina, "Enhancing the Delivery and Reception of Information Objects Over Deprived Channels and Computational Devices – Part II", Proceedings of the International Conference on Artificial Intelligence, Volume I, pp 281-285, July 12-15, 2010.


Green Bridge
Fallujah General Hospital
Main Bridge
Green Bridge
Highway 10
Askari Neighborhood
Jeghaifi Neighborhood
FALLUJAH
Jolan Neighborhood
Army Infantry 2nd Battalion 2nd Regiment
Marines 1st Battalion 3rd Regiment
Marines 1st Battalion 8th Regiment
Army Cavalry 2nd Battalion 7th Regiment
Marines 3rd Battalion 5th Regiment
Marines 3rd Battalion 1st Regiment

Original

Original Power Point
Document size is
4400 KB
PocoDoc size is
671 KB

PocoDoc

Figure 1. PowerPoint Optimization Image Comparison

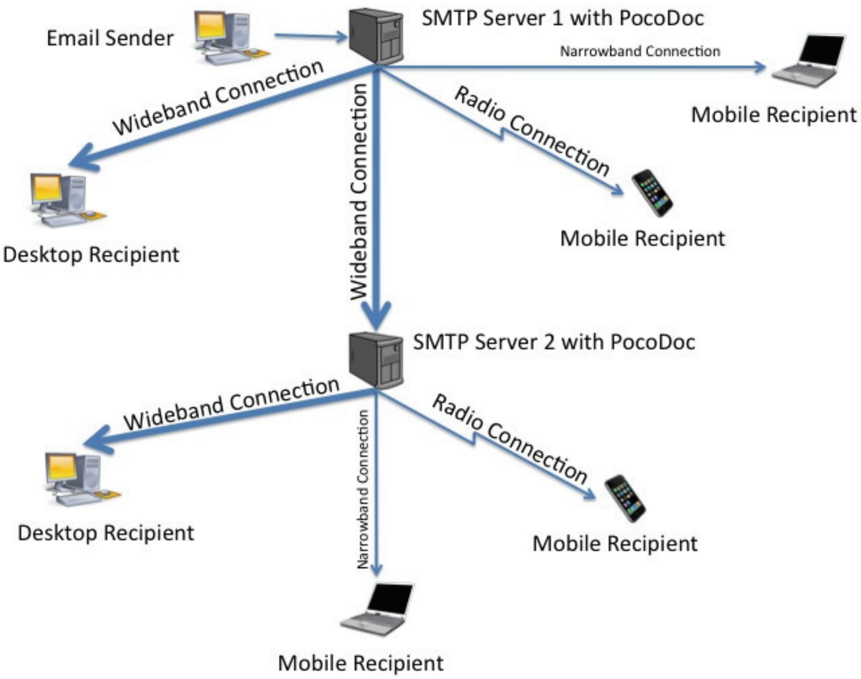


Figure 2. PocoBox an Email Appliance

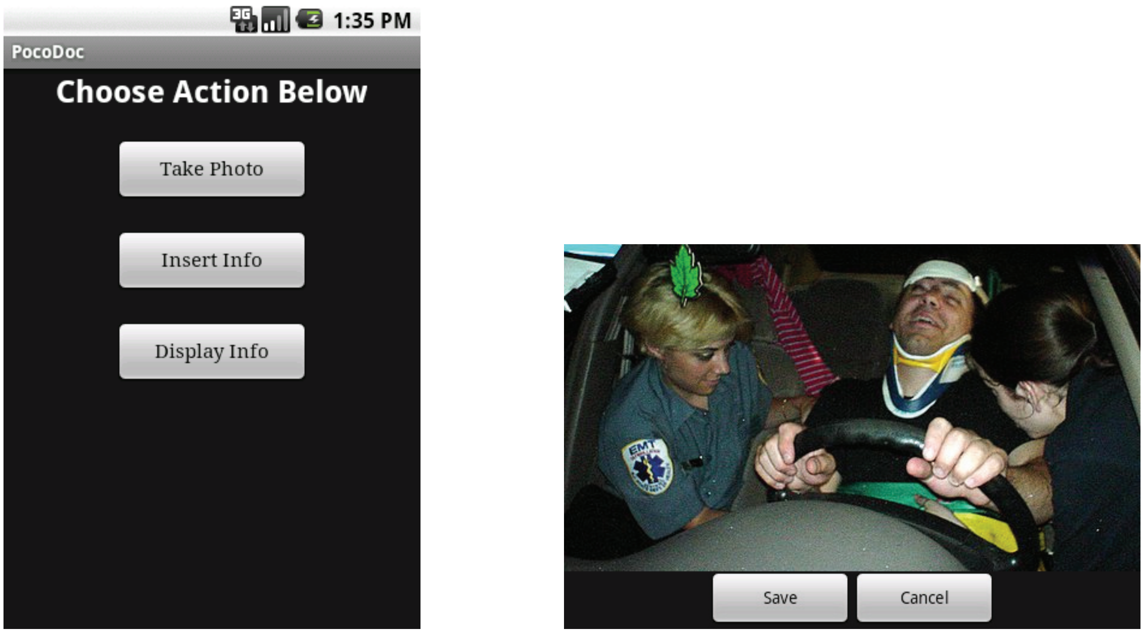
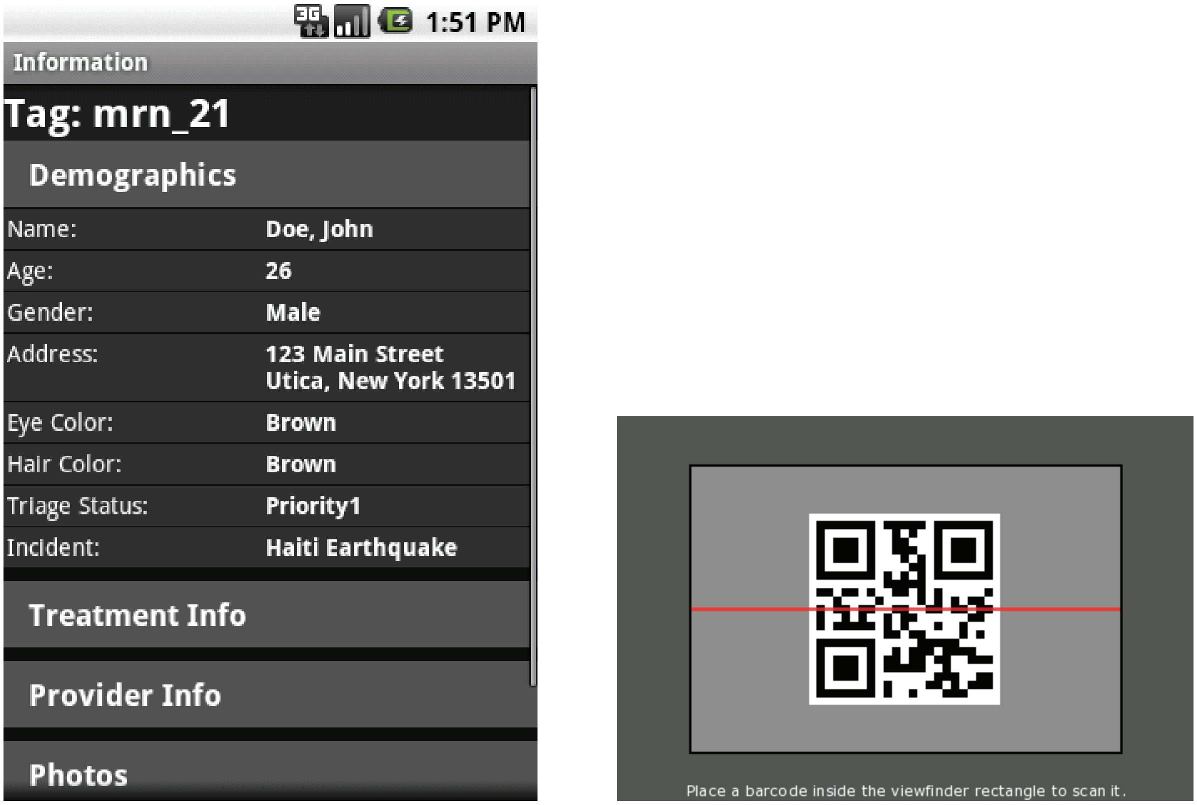


Figure 3. First Screen on Smartphone and Saving Images



Tag: mrn_21

Demographics

Name:	Doe, John
Age:	26
Gender:	Male
Address:	123 Main Street Utica, New York 13501
Eye Color:	Brown
Hair Color:	Brown
Triage Status:	Priority1
Incident:	Haiti Earthquake

Place a barcode inside the viewfinder rectangle to scan it.

Figure 4. Smartphone Demographics Screen and 2D Barcode

Efficient personalized e-learning material recommender systems based on incremental frequent pattern mining

Mohammad H. Nadimi-Shahraki

Faculty of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Isfahan, Iran.

E-mail: nadimi@ieee.org

Abstract – *Personalized e-learning material recommenders are known for discovering associations between learner's requirements and learning materials. They usually use association rule mining in which the most time-consuming part is frequent pattern mining from log files. Since the content of log files and learner profiles are frequently changed, frequent patterns must be updated to discover valid association rules. Obviously, updating frequent patterns by rerunning the mining process from scratch can be very time-consuming. In this paper, firstly we propose a general architecture for developing efficient personalized learning recommender systems using incremental association rule mining. Consequently, a new method is proposed for incremental frequent pattern mining from log files, which is the most computationally-intensive process in the proposed architecture. The content of log file is captured by using a well-organized tree in one database scan. While the log files are changed the tree can be incrementally updated. The experimental results show that using the proposed method enhances the efficiency of personalized e-learning material recommenders.*

Keywords: Personalized e-learning material recommender; Incremental frequent pattern mining; Web-based learning; E-learning.

1. Introduction

Over the past decade, applications of data mining techniques in web-base learning and e-learning systems have been increased. Meanwhile, personalized e-leaning material recommenders (PEMR) have been proposed based on discovering associations between learner requirements and learning materials. Usually, they assist learning management system (LMS) which is a main part of web-based learning and e-learning systems. LMSs are mainly considered to distribute the learning materials among learners in a course. Obviously, well-distributed learning materials can enhance the effectiveness of web-based learning and e-learning systems. There have been introduced some commercial LMSs such as WebCT [1] and Top-Class [2] while some examples of free LMSs are Ilias [3], Moodle [4] and Claroline [5]. Usually they can create a variety of several courses using the different learning material repositories and web-based learning materials. Web-based learning and e-learning systems record various tasks such as reading, writing, taking tests, and even communicating with peers performed by active learners in log files. Systematically, they provide a database consisted of personal information about the learners which is called learner profile. Since a huge amount of data can be daily generated by these systems, it is very difficult to analysis these data manually to use in LMSs. Therefore, using an efficient PEMR in LMS can enhance the effectiveness of the learning material distribution. Some efficient PEMRs have been proposed based on association rule mining in which the most time-consuming part is frequent pattern mining from log files.

When the association rules are used in PEMRs most of the research efforts have been devoted for two following issues: 1) improving the algorithmic performance [6] and 2) reducing the output set. One of the important problems involved in the first issue is incremental updating of the association rules. In other words, once the association rules have been discovered, they are valid until the content of log files and learner profiles are changed. Therefore, frequent patterns must be frequently updated to discover valid association rules. Obviously, updating frequent patterns by rerunning the mining process from scratch can be very time-consuming. Although many efficient association rule mining methods has been applied to web-based learning and e-learning systems, thus far a few methods have been introduced for incremental updating of association rules mined from log files.

In this paper, firstly we propose a general architecture for developing efficient personalized learning recommender systems using incremental association rule mining. Consequently, a new method is proposed to incremental frequent pattern mining from log files, which is the most computationally-intensive process in the proposed architecture. The content of log file is captured by using a well-organized tree structure called Log ordering-based FP-tree or LFP-tree in short. LFP-tree is an extension of FP-tree [7]. It has been considered to capture the whole records of log file by one scan in same ordering used in the log file.

Therefore, LFP-tree is constructed independent of the frequency of items. In other words, when log file is updated and frequency of patterns is changed, LFP can be updated without tree reconstruction and scanning the log file. Once LFP-tree is constructed, frequent patterns can be mined from the tree by using the original FP-growth [7].The experimental results verify the proposed method is able to mines frequent patterns from log files incrementally which is result to enhance the efficiency of personalized e-learning material recommenders based on association rules mining.

The rest of the paper is organized as follows. Section 2 introduces the problem and reviews some related works. The proposed architecture is introduced in Section 3. Consequently, a new method for incremental frequent pattern mining is proposed in Section 4. The experimental evaluation and results are presented in Section 5 and Section 6 contains some conclusions and future works.

2. Problem and Related Work

2.1. Problem Description

The main objective of using a personalized e-learning material recommender is to make learning recommendations to the learners with respect to their personalized activities. Although several data mining tasks and techniques have been used in personalized e-learning material recommender to make these recommendations, association rule mining has been the most common task [8].

An association rule is an implication of the form $X \rightarrow Y$, where X and Y are patterns and $X \cap Y = \varnothing$. Usually the rules are measured by two metrics called support and confidence. The support of pattern X or Sup(X) is the percentage of transactions that contain pattern X. Consistently, the support of the rule is the joint probability that transactions containing both X and Y. It is given as:

$$Sup(X \rightarrow Y) = Sup(X \cup Y)$$

Confidence of the rules $X \rightarrow Y$ is the conditional probability that a transaction contains Y, given that it contains X. It is given as:

$$Con(X \rightarrow Y) = \frac{Sup(X \cup Y)}{Sup(X)}$$

In fact, support shows how often a rule is applicable to the data set, while value of confidence shows how frequently items in Y appear with items in X in same transactions. The set of all strong association rules (SAR) are referred to those association rules whose support is greater than or equal the minimum support threshold *minsup* and confidence is greater than or equal the minimum confidence threshold *mincon*. It is denoted as:

$$SAR = \{r \in AR \mid Sup(r) \geq minsup \wedge Con(r) \geq mincon\}$$

Therefore, association rule mining problem is to find all rules having Sup ≥ *minsup* and Con ≥ *mincon*, where *minsup* and *mincon* are the corresponding support and confidence thresholds. Based on the above definition, the association rule mining problem can be broken down into two sub-problems: frequent pattern mining and rule generation [9]. The first sub-problem is to find all frequent patterns and the second sub-problem is to inference the interesting rules from frequent patterns.

The first sub-problem is the more computationally-intensive and has received the lion share of data mining community's attention [9, 10]. Given *i* distinct items within the database DB, there are 2^i-1 candidate itemsets to explore. Obviously, naive mining methods are often intractable especially when *i* is large.

Frequent patterns are itemsets or substructures that exist in a dataset with frequency no less than a user specified threshold. Let I= $\{i_1, i_2 \ldots i_n\}$ be a set of items and DB be a transaction database and T= $\{t_1, t_2 \ldots t_N\}$ be the set of all transactions such that each transaction t_i is a subset of items from I and N=|DB| is the number of transactions in DB. Given X= $\{i_j \ldots i_k\}$ be a subset of I ($1 \leq j \leq k$ and $k \leq n$) which is called a pattern or itemset. If a pattern contains k items, it is termed a k-pattern or k-itemset. An important property of a pattern X is its support count that is the number of transactions in DB that contain the particular pattern X. It can be formulated mathematically as follow:

$$Support\ count(X) = |\{ t_i \mid X \subseteq t_i, t_i \in T \}|$$

Consequently, the support of pattern X or Sup(X) is the percentage of transactions that contain pattern X. The pattern X will be called frequent if its support is no less than a user specified threshold called minimum support threshold or minsup denoted by σ ($0\ \% < \sigma \leq 100\ \%$). The problem of frequent pattern mining is to find all frequent patterns from transaction database DB with respect to σ.

In general, when a database is updated, some existing transactions are deleted from the database or some new transactions are inserted into database to form an updated database. Hence, some existing frequent patterns are no longer important and new frequent patterns may be introduced. In particular, in web-based learning and e-learning systems, when a log file is updated, some associations between learner's requirements and learning materials are no longer valid and new associations may be introduced.

Although many efficient association rule mining methods has been applied to web-based learning and e-learning systems [11-14], thus far a few methods have been introduced for incremental updating of association rules mined from log files [8]. Since generating association rules from updated frequent patterns is not a time-consuming process, therefore, the efficiency of updating association rules is mainly determined by the efficiency of method used for updating frequent patterns.

2.2. Related Work

Since introducing association analysis by AIS and Apriori algorithms [9], association rule mining has become a mature field of data mining research. Association rule mining has been applied traditionally for finding associations between items in a dataset and particularly for discovering associations between learner's requirements and learning materials in e-learning systems.

Yu at el. [15] have used association rule mining to find out the relationship between each learning-behavior pattern by which the teacher can promote collaborative learning behavior on the web. Minaei-Bidgoli at el. [16] have discovered interesting associations to improve online education systems for both teachers and students. In [17], association rule mining has been used to develop a personalized e-learning material recommender system. It assists students to find learning materials properly. Kuo et al. [18] proposed a real-time leaning algorithm to improve traditional sequential mining method. It mines the learner's behavior log incrementally. Moreover, the sequence log files are systematically divided into three sets to dynamically generate association rules efficiently. Consequently, the association rules are used by learning management system to make learning decision.

As reviewed above, there have been addressed several applications of association rule mining to make recommendations to learners. But a few methods have been proposed to incremental update the associations discovered from log files in web-based learning and e-learning systems.

To develop an incremental association rule mining, reviewing incremental frequent pattern mining methods is necessary. As mentioned above, the association rule mining problem can be broken down into two sub-problems: frequent pattern mining and rule generation. The first sub-problem is the more computationally-intensive and has received the lion share of data mining community's attention [10]. Although many efficient algorithms have been proposed for frequent pattern mining, thus far a few research efforts have been devoted for incremental frequent pattern mining [19-21] especially for using in web-based learning and e-learning systems [8].

Koh and Shieh [19] proposed the AFPIM algorithm for incremental mining. Specifically, it was designed to produce an FP-tree for the updated database, in some cases, by adjusting the old FP-tree via the bubble sort. However, in many other cases, it requires rescanning the entire updated database in order to build the corresponding FP-tree. The AFPIM algorithm suffers from several problems/weaknesses when handling incremental updates. A problem is the amount of computation spent on swapping, merging, and splitting tree nodes. Swapping is required because items are arranged according to a frequency dependent ordering. Another problem of the AFPIM algorithm is its requirement for an additional mining parameter *preMinsup*.

Chang et al. [20] proposed NFUP (New Fast Update Method). The NFUP is an Apriori-based algorithm same the FUP. To mine new frequent patterns in updated database, NFUP partitions the incremental database logically according to unit time interval (month, quarter or year, for example). For each item, assume that the ending time of exhibition period is identical. The NFUP scans each partition backward and it does not require the rescanning of the original database and can determine new frequent itemsets at the latest time intervals. The NFUP requires only the incremental database to be scanned. Their experimental results showed that the NFUP is suitable for frequently updated databases. However, the NFUP suffers from Apriori weaknesses inherently.

Leung et al. [21] proposed a canonical-order tree or CanTree in short, which is an efficient extension of the FP-tree. It captures the content of the transaction database by one database scan in canonical order, specifically; items can be consistently arranged in lexicographic order. Hence, there is no need to search and find merge-able paths like those in the CATS tree, and swapping of tree nodes affected by the frequency ordering. Once the CanTree is constructed, frequent patterns can be mined from the tree by using the original FP-growth [7]. The experimental results show that CanTree are very suitable for incremental frequent pattern mining.

3. The Proposed Architecture

This section introduces the proposed architecture for developing efficient personalized learning recommender systems using incremental association rule mining. Consequently, in the next section the proposed method for incremental frequent pattern mining is introduced. In fact, proposing a general architecture is the first objective of this research. It aims to introduce a general architecture for developing personalized e-learning material recommender systems that may efficiently assist the learning management systems.

As reviewed in Section 2, there have been introduced many e-learning recommender using association rules. They use association rules discovered from previous log files created by previous learners' session and profile. Obviously, while new learners are connected to web-based learning and e-learning systems, their log file is created. Therefore, some associations between learner's requirements and learning materials are no longer valid and new associations may be introduced. Obviously, rerunning the mining algorithm from scratch to update the associations is very time-consuming.

As previously mentioned in Section 2, an efficient solution to avoid rerunning the mining algorithm is incremental updating the association rules. As shown in Figure 1 , in the proposed architecture, association rules are updated by incremental association rule mining from incremental logs. Therefore, the recommender can use valid associations even the log files are frequently changed which result in precious recommendations to learners.

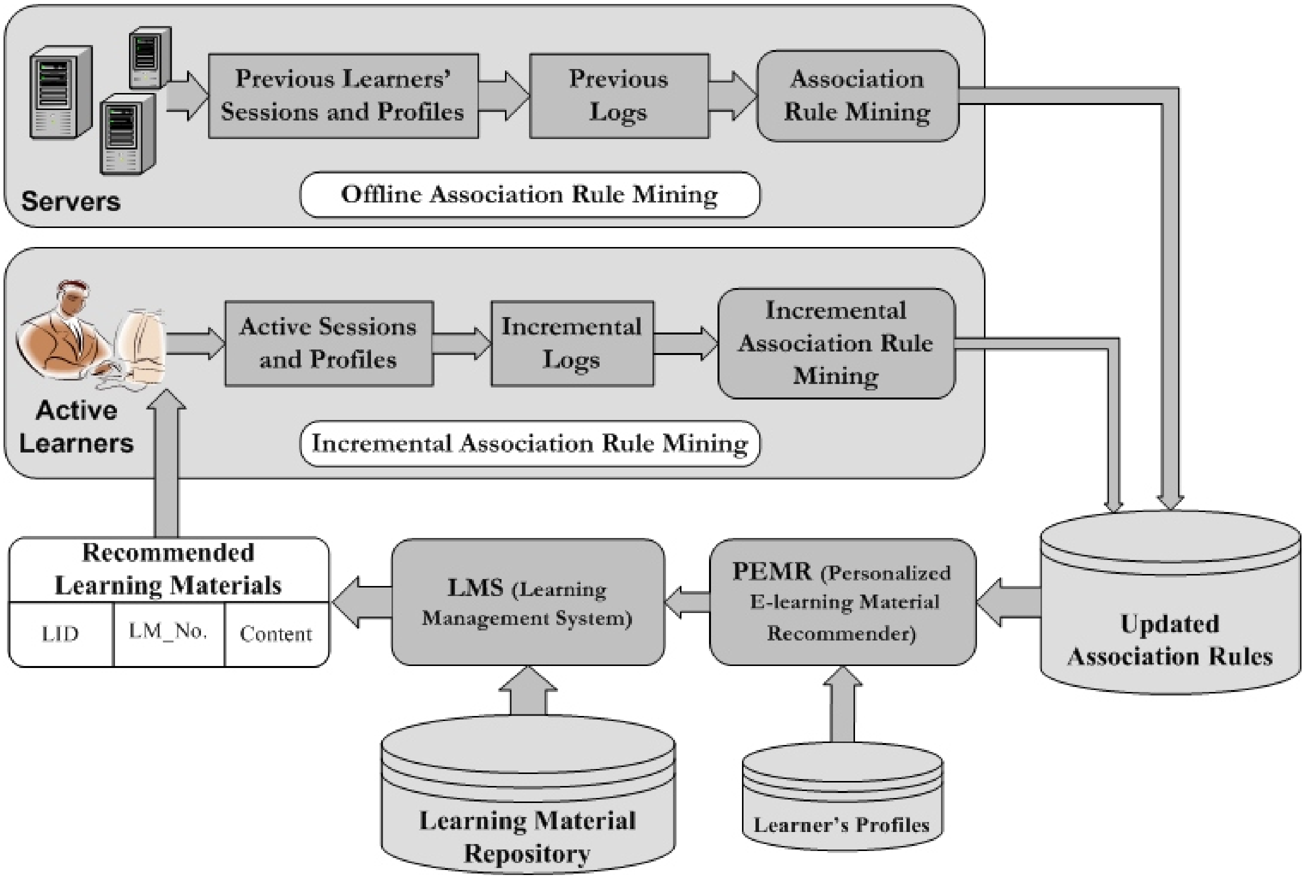

Servers
Previous Learners' Sessions and Profiles
Previous Logs
Association Rule Mining
Offline Association Rule Mining
Active Learners
Active Sessions and Profiles
Incremental Logs
Incremental Association Rule Mining
Incremental Association Rule Mining
Recommended Learning Materials
LID
LM_No.
Content
LMS (Learning Management System)
PEMR (Personalized E-learning Material Recommender)
Updated Association Rules
Learning Material Repository
Learner's Profiles

Figure 1. General architecture for personalized e-learning materials recommender systems

4. Incremental Frequent Pattern Mining

As previously discussed in Section 2, in web-based learning and e-learning systems, when a log file is updated, some associations between learner's requirements and learning materials are no longer valid and new associations may be introduced.

Although many efficient association rule mining methods has been applied to web-based learning and e-learning systems, thus far a few methods have been introduced for incremental updating of association rules mined from log files [8]. Since generating association rules from updated frequent patterns is not a time-consuming process, therefore, the efficiency of updating association rules is mainly determined by the efficiency of method used for updating frequent patterns. Moreover, previously it was discussed that rerunning the mining algorithm to update frequent patterns is unacceptable. Therefore, in this section a new method for incremental frequent pattern mining from log files is proposed.

In the proposed method, the content of log file is captured by using a well-organized tree structure called Log ordering-based FP-tree or LFP-tree in short. LFP-tree is an extension of FP-tree. Although FP-tree is pioneer of frequent pattern mining without candidate generation, it keeps only the frequent patterns by two database scans. In other words, when the log file is updated (i.e., any record is inserted, deleted, and/or modified), FP-tree must be reconstructed by two scans of the entire log file. To solve this weakness, LFP-tree captures the whole records of log file by one scan in same ordering used in the log file. Therefore, it is constructed independent of the frequency of items. In other words, when log file is updated and frequency of patterns is changed, LFP can be updated without tree reconstruction and scanning the log file. To illustrate, let us examine it through an example. As shown in Figure 2, given M as learning material repository consisted of different content materials (learning materials). Consistently, many courses can be produced by combining the content materials.

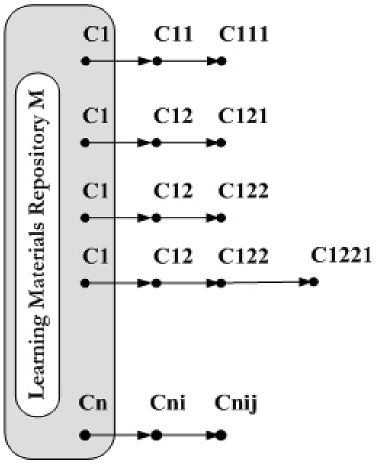


Figure 2. Learning materials repository M

For example the first course can be produced by three different content materials C1, C11 and C111 denoted {C1, C11, C111}. If the number of content materials (items) is denoted by *N*, then, 2^N -1 different course can be produced from these content materials.

Since in web-based learning and e-learning systems there are a big number of content materials (i.e. N is a big number), therefore, there are a huge number of possible courses. Particularly, it means there are 2*N* -1 candidate patterns in the log files created by learners which use content materials of M in different sequences. Let the first six rows shown in Table 1 be the logs recorded for the six current learners who are using the content (learning) materials. In this table, the first column shows the learner's identifier (LID) and the second column shows their course sequences. For example, the third row means that, the learner with identifier of 3 has used content materials C1, C11, and C1112 respectively.

As explained above, LFP-tree captures the whole records of log file by one scan in same ordering used in the log file. Figure 3 shows the LFP-tree constructed by the first six logs. Unlike FP-tree, which keeps only the frequent items found in its first database scan, the first six logs are scanned, and then added to the LFP-tree using the same ordering used in log files. While the logs are scanned, if the same content materials exist in the new log and the tree path, then the frequency of the nodes in the path is incremented by 1. Otherwise, a new branch will be created associated with the path. By following above, LFP-tree is constructed independent of the frequency of items.

Table 1. Log File *LF*

LID	Content Sequence
1	C1,C11,C111,C1111
2	C1,C11,C111
3	C1,C11,C1112
4	C1,C11,C111,C1112
5	C2,C21,C211
6	C2,C21,C212
7	C1,C11,C1112
8	C1,C11,C1113

Given two new learners are connected to the e-learning system and the log file is incrementally updated by inserting their logs shown by the last two rows in Table 1. By updating the log file, the frequency of candidate patterns and the number of records are incremented which result in introducing new frequent patterns.

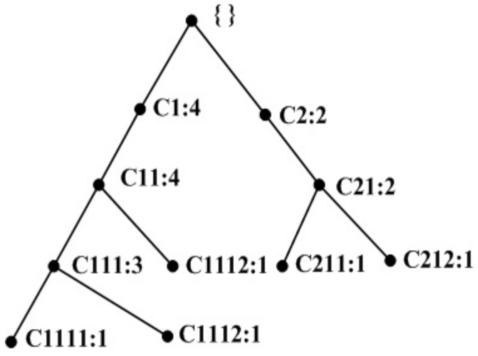


Figure 3. LFP-tree constructed by the first six logs

Consequently, the current frequent patterns and association rules are no longer valid. Therefore, frequent patterns must be updated. If FP-tree is used, then whole log file must be scanned twice. Since LFP-tree is not dependent on the frequency, it can be updated only by scanning the incremented logs of the log file. Therefore, the seventh and eighth rows of log file LF shown in Table 1 are added to LFP-tree in the same fashion used for the previous logs. Figure 4 shows the LFP-tree updated by the incremented logs shown in seventh and eighth rows in Table 1. Once LFP-tree is constructed, frequent patterns can be mined from the tree by using the original FP-growth [7].

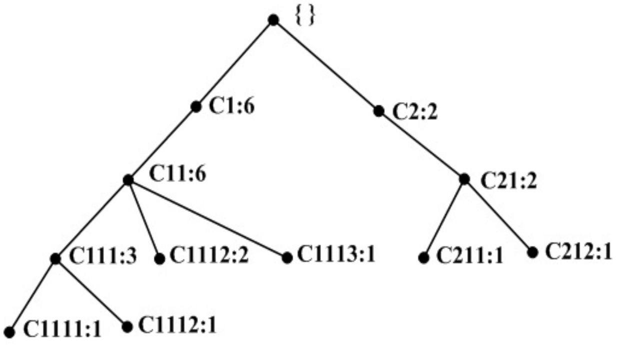


Figure 4. LFP-tree updated by incremented logs

5. Experimental Results

In this section, we evaluate the performance of our method. All experiments were performed in a time-sharing environment in a 2.4 GHz PC. All the algorithms are implemented using Microsoft Visual C++ 6.0. We performed comprehensive experimental analysis on the performance of LFP-tree on several synthetic datasets. According to the space limitation and the problem specifications, we only show the result of efficiency evaluation on synthetic dataset T10I6D100k by two following experiments. The dataset T10I6D100k is generated by the program developed at IBM Almaden Research Center [22]. The number of transactions, the average transaction length and the average frequent pattern length of T10I6D100k are set to 100k, 10 and 6 respectively.

In each experiment, LFP-tree and FP-tree are separately run in the same experimental environment. Once they are constructed, frequent patterns are mined by using original FP-growth. The results reported in figures have been computed by the average of multiple runs.

The first experiment aims to evaluate the efficiency of the proposed method in term of runtime for frequent pattern mining. Similar to related works, for doing this, the effect of change in the *minsup* value on the runtime must be examined. The graphs in Figure 5 indicate the runtime of our method denoted LFP-tree and FP-tree running for different values of minsup (i.e. minsup versus runtime). To evaluate the efficiency for frequent pattern mining, for each value of minsup, our method and FP-tree are run from scratch. It is expected that, if the minsup were decreased, the runtime would be increased. Moreover, FP-tree must be faster than LFP-tree, since it is constructed in frequency-depending ordering by which tree become smaller. The graphs agree with the expectation; the runtime has been increased when the *minsup* was decreased and FP-tree is faster than LFP-tree for individual or static mining.

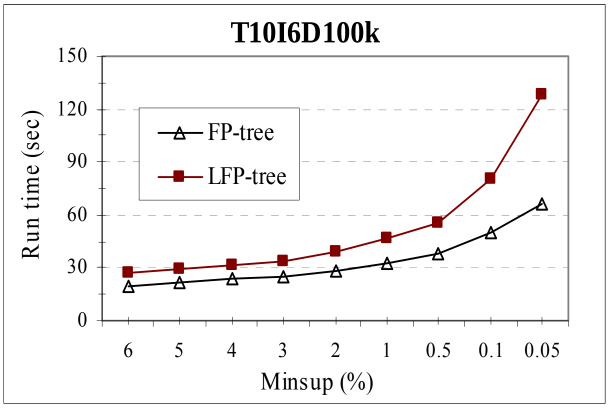


Figure 5. Efficiency of runtime vs. minsup

The second experiment is to evaluate incremental frequent pattern mining using LFP-tree and FP-tree. For doing this, we consider p% of the dataset where p will be increased from 10 to 100. The graphs in Figure 6 indicate the runtime of our method denoted LFP-tree and FP-tree running for different portion (i.e. size of incremented portions versus runtime). It is expected that, incremental frequent pattern mining using LFP-tree is the faster than rerunning mining algorithm from scratch using FP-tree. The graphs agree with the expectation.

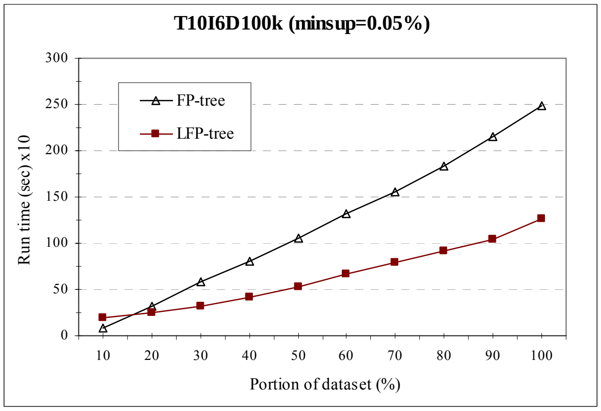


Figure 6. Incremental frequent pattern mining

6. Conclusions and Future Works

In this paper, we proposed a general architecture for developing efficient personalized e-leaning material recommenders (PEMR) systems using incremental association rule mining. Then, we proposed a new method for incremental frequent pattern mining from log files, which is the most computationally-intensive process in the proposed architecture. A well-organized tree structure called LFP-tree was proposed to capture whole log file by one scan. It is an extension of FP-tree, but independent of item-frequency. While the log files are changed by new loge, the incremental logs can be added to LFP-tree efficiently. Consistently, the frequent patterns can be updated by using this LFP-tree. The experimental results verified the efficiency of the proposed method over synthetic datasets. However, in future studies it can be also examined by real data collected from web-based learning and e-learning systems.

References

[1] "WebCT, available at http://www.webct.com/ , 2011."

[2] "TopClass, available at http://www.topclass.nl/ , 2011."

[3] "Ilias, available at http://www.ilias.de/, 2011."

[4] "Moodle, available at http://moodle.org/, 2011."

[5] "Claroline, available at http://www.claroline.net/, 2011."

[6] A. Ceglar and J. F. Roddick, "Association mining," *ACM Computing Surveys (CSUR),* vol. 38, pp. 5-es, 2006.

[7] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," *Data mining and knowledge discovery,* vol. 8, pp. 53-87, 2004.

[8] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews,* vol. 40, pp. 601-618.

[9] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," *International Conference on Very Large Data Bases (VLDB'94),* p. 487499, 1994.

[10] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: current status and future directions," *Data Mining and Knowledge Discovery,* vol. 15, pp. 55-86, 2007.

[11] C. M. Chen, Y. L. Hsieh, and S. H. Hsu, "Mining learner profile utilizing association rule for web-based learning diagnosis," *Expert Systems with Applications,* vol. 33, pp. 6-22, 2007.

[12] E. García, C. Romero, S. Ventura, and T. Calders, "Drawbacks and solutions of applying association rule mining in learning management systems," in *the International Workshop on Applying Data Mining in e-Learning*, 2007, p. 13.

[13] E. García, C. Romero, S. Ventura, and C. Castro, "An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering," *User Modeling and User-Adapted Interaction,* vol. 19, pp. 99-132, 2009.

[14] L. Yuemin and Z. Shenghui, "An association rule mining approach for intelligent tutoring system," in *Computer Engineering and Technology (ICCET), 2010 2nd International Conference on*, pp. V6-460-V6-464.

[15] P. Yu, C. Own, and L. Lin, "On learning behavior analysis of web based interactive environment," 2001, pp. 1-10.

[16] B. Minaei-Bidgoli, P. N. Tan, and W. F. Punch, "Mining interesting contrast rules for a web-based educational system," in *International Conference on Machine Learning Applications*, Los Angeles, USA, 2004.

[17] L. Jie, "A personalized E-Learning material recommender system," in *the 2nd International Conference on Information Technology for Application (ICITA 2004)*, 2004.

[18] Y. H. Kuo, J. N. Chen, Y. L. Jeng, and Y. M. Huang, "Real-Time Learning Behavior Mining for e-Learning," 2005, pp. 653-656.

[19] J. L. Koh and S. F. Shieh, "An efficient approach for maintaining association rules based on adjusting FP-tree structures1," 2004, pp. 221-227.

[20] C. C. Chang, Y. C. Li, and J. S. Lee, "An Efficient Algorithm for Incremental Mining of Association Rules," in *15th IEEE International Workshop on Research Issues in Data Engineering: Stream Data Mining and Applications (RIDESDMA'05)*, 2005.

[21] C. K. S. Leung, Q. I. Khan, Z. Li, and T. Hoque, "CanTree: a canonical-order tree for incremental frequent-pattern mining," *Knowledge and Information Systems,* vol. 11, pp. 287-311, 2007.

[22] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," *Proc. 20th Int. Conf. Very Large Data Bases, VLDB,* vol. 1215, p. 487499, 1994.

Towards Emerging Green Information and Communication Technologies: A Review

G.Bekaroo† & C.Bokhoree†

†School of Sustainable Development and Tourism,
University of Technology, Mauritius

gbekaroo@umail.utm.ac.mu sbokhoree@umail.utm.ac.mu

Abstract

Though the ICT industry has been one of the fastest growing industries in many countries over the past years, its relationship with the natural environment has not been given much consideration. This growing industry has as main negative impact on the environment, climate change in the form of global warming, caused by the emissions of carbon in the air. In order to minimize the harms caused by the growing IT industry on the environment, Green ICT is an emerging solution which is increasingly being adopted by businesses and computer users. Green ICT refers to environmentally sustainable computing and is a discipline that studies, develops and promotes environmentally friendly and resource-efficient ICT products over their entire life cycle. This paper discusses the effects of the growing IT industry on the environment along with a review of emerging green information and communications technologies and best practices which will shape the future of computing at home and in businesses.

Keywords

Green ICT, Sustainable Development, Emerging Technologies, Energy Consumption, Energy Cost, Environment

Introduction

Down the past years, the Information and Communications Technology (ICT) has had significant impact on businesses and the society due to the various benefits provided by its adoption. Computer technologies and the Internet are being increasingly used today, be it at home, in businesses, schools and even in public places via wireless hotspots. As such, an increasing number of computing and electronic devices are being used in an unsustainable manner which negatively impacts the environment in the main form of climate change which several countries around the world are facing.

Studies have estimated that IT accounts for two per cent of worldwide carbon emissions which is the same level of CO_2 emissions as the airline industry (Gartner 2007). This figure is also expected to double over the next decade since the number of Internet and mobile phone subscribers is constantly increasing. On the other hand, energy costs kept on rising during the past years which has had an impact on businesses and society. Taking cognizance of the potential threats, Green ICT is being promoted around the world and different technologies and best practices are being developed so as to reduce energy costs while at the same time favouring the environment. However, to bring about proper best practices and adoption of green technologies, thorough investigation in consumption pattern, policy analysis and e-waste management practices are required.

Unsustainable Energy Consumption and its Adverse Environmental Impacts

In UK, the proportion of household computer owners increased from 72% to 75% between 2008 and 2009 (Office for National Statistics 2010), similar to several countries around the world (Internet World Stats 2010). Studies have shown that personal computers are not being actively used during most of the time they are switched on (Miller, 2008). At home, much power is consumed when computers are left on especially during period of inactivity or during downloading from the Internet at night, which is a common practice by youngsters. A desktop computer normally has a 200-watt power supply and if 100 million of these machines are turned on at once worldwide, it means that together 20,000 megawatts of electricity are being used, which is the total output of 20 average-sized nuclear power plants (Tanenbaum, 2001).

Several businesses around the world including Google and Microsoft are building several large data centres consisting of thousands of electronic and computing devices needed to support the services being provided including search, email, etc (Barroso, 2007). This expansion of ICT infrastructure unsurprisingly leads to an increase in the company's energy costs due to increased power consumption in four main categories namely: from critical computational systems, cooling, during conversion and hostelling (Saul 2008). Other studies have shown that roughly 40-50% of corporate energy consumption and computing centre power costs have doubled over the last 5 years (Cluster Resources 2009) and that an average-sized company with approximately 10,000 PCs wastes at least $165,000 in electricity costs each year due to computers being left on overnight (PC Energy Report 2007). As an attempt to minimise costs, companies are attempting to relocate to places where power and human resources are cheaper (Katz 2007).

Though ICT is being increasingly used within business and society, its relationship with the environment has often been overlooked in the past. This progressive adoption of ICT adversely impacts the environment in the form of global warming, caused by the emissions of green house gases in the air that eventually leads to climate change (International Socialist Group 2006). In addition to GHG emissions, the lifecycle environmental impact of ICT also concerns the extraction and disposal of harmful materials (Plepys 2002).The fast growing IT industry is contributing to the production of large quantities of electrical and electronic waste (e.g. computers, telephones, printers, etc) due to the life span of computers and some electronic devices including mobile phones ranging between 2 and 3 years (Jaragh and Boushahri 2009).

THE EMERGING GREEN ICT

With entities using ICT resources looking for new ways to save money especially during the global economic recession, energy costs are an important part of the picture. Power factor correction is an efficient way that businesses or computer owners can put forth the effort to save energy which in turn will save money while at the same time reduce CO_2 emissions. Most companies are running using electricity with efficiency between 70-80% approximately (PRWeb 2011). With a power factor correction after applying energy efficiency techniques, 5-25% of the energy bills can be saved if the efficiency is brought closer to 100%.

Green ICT refers to environmentally sustainable computing and is a discipline that studies, develops and promotes designing environmentally friendly and resource-efficient ICT products over their entire life cycle starting with the conception and design of the product up to usage and recycling. Adopting Green ICT techniques and best practices does not only contributes to cost reduction but also favours the environment in the long term. The sustainability concept is commonly associated with three dimensions: social, environmental, and economic as it is used in connection with human and natural systems. Social sustainability refers to maintaining organisational social conditions that do not undermine people's ability or jeopardize their potential to meet their needs in the future. Economic sustainability, in turn, means the amount of consumption that can be sustained indefinitely and whether organisations will be able to respond to adverse changes in financial conditions (Costanza and Wainger 1991) and ecological/environmental sustainability can be defined as improving the quality of human life within organisations while living within the carrying capacity of supporting ecosystems (IUCN 1991).While taking these sustainability dimensions into practice, some best practices are being applied in order to promote Green ICT and the most common ones include:

- Using laptops instead of desktop computers where possible,
- Turn off equipments when not in use,
- Using computer power saving modes,
- Buying/Using efficient equipments,
- Minimise printing,
- Using virtualization technologies wherever possible,
- Green Education.

EMERGING GREEN TECHNOLOGIES IN ICT

Before appropriate measures can be taken to reduce consumption during computer usage at home and in businesses, energy consumption measurement metrics are essential. Metrics help to draw conclusions on whether adopted processes really contribute to a reduction in energy consumption, based on comparison of measured values. Discussion will begin with emerging technologies facilitating energy measurement from ICT resources, followed by technologies and strategies on how to reduce energy consumption at different levels of hardware, software, building and awareness.

Emerging Green Energy Efficiency Metrics

An organisation needs good environmental measurement solutions and metrics for measuring energy consumption. Due to the complexity of ICT businesses which consist of different components including hardware, software, people, building, among others; saving energy in each and every components is beneficial.

For effective energy consumption measurement in data centres, different metrics have been proposed and the most common ones include PUE, DCiE, DCeP and SpecPower_ssj Benchmark. The Power Usage Effectiveness (PUE) and its reciprocal Data Centre infrastructure Efficiency (DCiE), proposed by the Green Grid (2011), are benchmarking standards to help to determine the energy efficiency of data centres and to monitor the impact of their efficiency efforts. PUE is a measure of how efficiently a data centre uses its power in terms of computing equipment. It computes the ratio of the total power used by a computer data centre facility to the power delivered to computing equipment. Under facility power usage, non-computing devices like cooling and lighting are considered. DCiE in turn gives the percentage efficiency and is the reciprocal of PUE. Brought about after the DCiE and PUE by the Green Grid, the Data Centre energy Productivity (DceP) incorporates both infrastructure and IT equipment unlike PUE and DCiE which focus primarily on infrastructure (mechanical and electrical systems) while evaluating energy efficiency within IT organisations. The SPECpower_ssj benchmark, in turn, is a Standard Performance Evaluation Corporation (SPEC) benchmark to evaluate power and performance characteristics of servers. The benchmark helps to compare the energy efficiency of servers whereby determining the amount of power servers require at different levels of utilisation by using the the Server Side Java Operations per Watt metric (Ou 2008).

To measure energy consumed by hardware/electric components, different power meters are available on the market that can be used for this purpose. These power meters can tell users how much energy is being used for a particular instant or period of time by appliances and electronics plugged to the device. Examples of power meters and their features are discussed in the emerging green hardware section. Similarly, some studies have attempted to measure energy consumed by software, when in use. For example, Seo et al (2008) have developed an application that can be used to estimate energy consumption of pervasive Java based software systems. Likewise, Amsel and Tomlinson (2010) have produced a tool, named Green Tracker, which can be used to estimate the energy consumption of currently installed software systems by making use of CPU usage of the software being monitored.

Emerging Green Building Design Considerations

During the construction, expansion or innovation of buildings hosting ICT companies today, much energy can be saved if proper green design techniques are applied. Green building involves a comprehensive process of design and construction which also employs techniques to minimize adverse environmental impacts and reduce the energy consumption of a building, while at the same time contributing to the health and productivity of its occupants. Green buildings might need a higher initial investment cost but in the long run they turn out to be minimising both costs and energy (Eco-Business, 2011).

Roodman and Lenssen (1995) showed that buildings account for one-sixth of the world's fresh water withdrawals, one-quarter of its wood harvest, and two-fifths of its material and energy flows. Common practices being applied during design and construction of green buildings include optimised orientation in order to minimise solar heat gains, with minimal direct West-facing facades and architectural designs that maximise day-lighting. Extensive overhangs and planters are being used in green buildings in order to block direct solar exposure and facade, and roof greening are being introduced to mitigate urban heat effect and solar heat gain. Rainwater drainage systems can also be hidden within the building cladding and collected water from rain can be re-used in toilets and for irrigation. Today, some projects are also coming with the use of extensive photovoltaic panels and some are even opting for eco-friendly materials such as 'green concrete' comprised of copper slag, recycled concrete aggregates and ground granulated blast furnace slag (Tay 2011).

Green building techniques are better defined as best practices in standards like BRE Environmental Assessment Method (BREEAM) and Leadership in Energy and Environmental Design (LEED) that can be adopted by businesses. BREEAM is a popular environmental assessment method for buildings as it sets the standard for best practice in sustainable design and has become the de facto measure used to describe a building's environmental performance (BREEAM, 2009). LEED is basically a third-party certification program for design, operation and construction of high performance green buildings and ensures the buildings are environmentally compatible, provide a healthy work environment and are profitable (LEED, 2009) especially if the IT organisation rents office space. This is because certified buildings are commanding higher rental rates and great occupancy than non-green buildings (LEED, 2009).

Emerging Green Software Systems

Software with the aim of optimising energy usage is often referred to as a Green Software System. There are many emerging green software systems coming on the market that do not only monitor and better utilise system resources, but also go beyond and nullify the need to add extra hardware to the existing infrastructure. For example, Microsoft Joulemeter (Microsoft Research 2010) is a software based mechanism to measure the energy usage of virtual machines (VMs), servers, desktops, laptops, and even individual pieces of software running on a computer. Recently released on Windows 7, the Joulemeter also provides additional insight into just how much energy Windows computers manage to swallow, be them servers, desktops or laptops. The eMeter (2011) is another green software system that crunches the data from electrical meters for the benefit of utilities. It can monitor power consumption, maintenance issues, trends, outrages, billing management, among others. Likewise, Microsoft released a beta web-based application, called Microsoft Hohm (2011), which helps household owners to better understand home energy usage while at the same time helping to save energy and money from recommendations given based on data fed in the application.

Many similar applications are emerging on the market and focus on different areas. In server farms, Niyato et al (2009) proposed an Optimal Power Management (OPM) solution based on the constrained Markov Decision Process developed by Feinberg and Shwartz (2001), which observes the state of a server farm and makes decision to switch the operation mode of a server in order to minimise the power consumption in the server farm. Similarly, Verma et al (2008) presented an application placement controller, called pMapper, which dynamically places applications in server farms so as to minimise power. To help in green building construction and repairs, Sustainable Spaces (Recurve 2011) recently came up with an application that allows contractors to determine the optimal repairs for a building and Autodesk

(2011) is focusing on improving energy efficiency in buildings, reducing the amount of raw materials in manufactured goods, and replacing fossil fuels. In agriculture, GeoMation (Hitachi 2011) crunches satellite data to determine the optimum time to harvest wheat and rice.

Emerging Green Computer Hardware

Much work has been done by computer designers and vendors to reduce power consumption in personal computers and servers, mostly concentrated on improving battery life in portable computers. Different power meters and home energy monitoring kits are now available on the market which help better to measure and monitor energy consumed in households or businesses. The best way to reduce power consumption in a house is by adopting a two-pronged approach (PowerMeterStore 2011) which involves firstly to measure and monitor the total power consumption in the house or building, and secondly to measure and monitor power consumption of individual electronics (including computers and associated peripheral devices) and appliances. Common devices for power measurement and monitoring power of individual electronics include Kill A Watt (P3 International Corporation 2008) and Watts Up Pro (Watts-Up 2011) and these devices can tell users how much energy is being used for a particular instant or period of time by appliances and electronics plugged to the device. For measuring and monitoring total power consumption, common home energy monitoring kits including Cent-a-Meter (2011) and Power Cost Monitor (2011) which are increasingly being used today.

Recently, Agarwal et al (2008) came up with an architecture named Somniloquy meaning "the act or habit of talking in one's sleep" that aims to augment network interfaces to allow computers in sleep mode to be responsive to network traffic. This enables computers using the interface to appear to be "sleep talking" (Science Daily 2009) and following this realization, the team built a small USB-connected hardware and software plug-in system that allows a PC to remain in sleep mode while continuing to maintain network presence and run well-defined application functions (e.g. VoIP, web downloads, file sharing, etc). In the future, Somniloquy could potentially be incorporated into the network interface card of new PCs, which would eliminate the need for the prototype's external USB plug-in hardware (Science Daily 2009).

Emerging Green Education

Increasing people awareness is one of the key steps towards promoting Green ICT. Users of the ICT need to understand the current climate change problems being faced everywhere in the world and also how one can provide help in cutting down energy usage to help against this major problem facing the environment today. Attitudes and behaviours of human is complex and takes time for positive change to happen. Continuous and capacity building trainings helps for change of mind sets.

As such, many businesses around the world have already started to provide courses to their staff in order to increase their awareness in Green ICT, which can help the companies to cut down electricity bills in the long term. Similarly, many training institutions are now providing courses on going green in different areas thus making the participants 'Green Literate'.

Even at the tertiary education level, universities are considering to include environmental modules in Computer Science programmes (Talebi 2009) thus promoting green education. In the same educational institutions, green teaching techniques are being employed where interactive whiteboards are increasingly being used instead of the overhead projector using plastic slides. Also, soft copies of lecture materials are being given to students instead of printed handouts. Different books on promoting Green ICT are now available in the libraries and conferences and workshops are now being held more often in different countries in the world for the same main reason, that is, to increase people's awareness in Green ICT.

CONCLUSION

Adoption of emerging technologies within Green ICT is promoting environmental responsibility between computer users at home and in organisations. This paper has discussed the current problems being faced by the environment due to the growing ICT industry and has provided a roadmap for research and development on emerging green information and communication technologies being adopted to cut down energy costs while at the same time promoting environmental friendliness.

REFERENCES

AGARWAL, Y., HODGES, S., SCOTT, J., CHANDRA, R., BAHL, V.; GUPTA, R. (2008), *Somniloquy: Maintaining network connectivity while your computer sleeps* [online]. Microsoft Research. Accessed on: 01 Nov 2010, Available at: http://research.microsoft.com/research/pubs/view.aspx?type=Technical%20Report&id=1458

AMSEL, N.; TOMLINSON, B. (2010), *Green tracker: a tool for estimating the energy consumption of software*, Conference on Human Factors in Computing Systems, Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems, p.3337-3342, ISBN:978-1-60558-930-5

Autodesk (2011), Autodesk, Accessed on: 9 Feb 2011, Available at: http://usa.autodesk.com/

BARROSO, L.A. (2007), *Warehouse –scale computers*, In USENIX Annual Technical Conference

BREEAM (2009), What is BREEAM? [online], BRE Environmental Assessment Method, Accessed on: 29 Dec 2010, Available at: http://www.breeam.org/page.jsp?id=66

Cent-a-meter (2011), *Think Green* [online], Accessed on: 21 Jan 2011, Available at: http://www.centameter.com.au/

Cluster Resources (2009), Moab Energy-Saving and Green Computing Solutions [online], Accessed on: 20 Oct 09, Available at: http://www.clusterresources.com/docs/220

COSTANZA, R.; WAINGER, L. (1991), *Ecological economics: mending the Earth*, North Atlantic Books, Berkeley

CURTIS, L. (2008), *Environmentally Sustainable Infrastructure Design*, Green Computing, The Architecture Journal #18, Microsoft, p.2-8

Eco-Business (2011), *Green buildings in Singapore: Adding the green touch with technology* [online], Accessed on: 10 Feb 2011, Available at: http://www.eco-business.com/news/green-buildings-in-singapore-adding-the-green-touch-with-technology/

EIA (2011), *Short-Term Energy Outlook* [online], Accessed on: 13 Mar 2011, Available at: http://www.eia.doe.gov/emeu/steo/pub/contents.html

eMeter (2011), *eMeter - Energy Information You Can Act On* [online], Accessed on: 12 Feb 2011, Available at: http://www.emeter.com/

FEINBERG, E.A.; SHWARTZ, A. (2001), *Handbook of Markov Decision Processes: Methods and Applications*, 1 edition, Published by: Springer, ISBN: 0792374592

Gartner (2007), PETTEY, C., *Gartner Estimates ICT Industry Accounts for 2 Percent of Global CO2 Emissions* [online], Press Releases, Accessed on: 12 Dec 2010, Available at: http://www.gartner.com/it/page.jsp?id=503867

Green Grid (2011), *The Green Grid* [online], Accessed on: 18 Mar 2011, Available at: http://www.thegreengrid.org/

Hitachi (2011), "GeoMation Farm" Agriculture Information Management System [online], Accessed on: 11 Apr 2011, Available at: http://www.hitachi.com/environment/showcase/solution/it/geomation.html

International Socialist Group (2006), *Climate Change- The biggest challenge facing humanity* [online], Accessed on: 02 Oct 09, Available at: http://www.isg-fi.org.uk/spip.php?article303

Internet World Stats (2010), *Internet Usage Statistics- the Internet Big Picture* [online], Accessed on: 07 May 2011, Available at: http://www.internetworldstats.com/stats.htm

IUCN; UNEP; WWF (1991), *Caring for the Earth*, IUCN Gland, Switzerland.

JARAGH, M; BOUSHAHRI, J. (2009), *The e-waste impact*, Proceedings of the First Kuwait Conference on e-Services, ISBN: 978-1-60558-797-4

KATZ, R. (2007), *Research directions in internet-scale computing*, In 3rd International Week on Management of Networks and Services

LEED (2008), Promoting LEED Certification and Green Building Technologies [online], Accessed on: 29 Dec 2010, Available at: http://leed.net/

Microsoft Hohm (2011), How energy efficient is your home? [online], Accessed on: 12 Mar 2011, Available at: http://www.microsoft-hohm.com/

MILLER (2008), *Computer power management* [online], Information Technology, MILLER School of Medicine University of Miami, Accessed on: 5 Jan 2011, Available at: http://it.med.miami.edu/x1159.xml

Microsoft Research (2010), KANSAL, A; GORACZKO, M., *Microsoft Joulemeter User Manual* [online], Accessed on: 13 Nov 2010, Available at: http://research.microsoft.com/en-us/projects/joulemeter/UserManual.pdf

Office for National Statistics (2010), Consumer Durables - Ownership increases [online], Accessed on: 07 May 2011, Available at: http://www.statistics.gov.uk/cci/nugget.asp?id=868

Ou, G. (2008), Measure energy efficiency for Server Side Java [online], ZD Net, Accessed on: 19 Mar 2011, Available at: http://www.zdnetasia.com/measure-energy-efficiency-for-server-side-java-62037771.htm

NIYATO, D.; CHAISIRI, S.; SUNG, L.B (2009), *Optimal Power Management for Server Farm to Support Green Computing*, Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, p. 84-91, ISBN:978-0-7695-3622-4

P3 International Corporation (2008), *Kill A Watt* [online], Accessed on: 04 Dec 2010, Available at: http://www.p3international.com/products/special/P4400/P4400-CE.html

PowerCostMonitor.com (2011), *PowerCost Monitor Home Energy Meter* [online], Accessed on: 21 Jan 2011, Available at: http://www.powercostmonitor.com/

PC Energy Report (2007), *Climate Savers Computing* [online], Accessed on 01 Mar 2010, Available at: http://www.climatesaverscomputing.org/docs/Energy_Report_US.pdf

PLEPYS, A. (2002), *The grey side of ICT*, Environmental Impact Assessment Review, Vol.22, no. 5, p.509- 523

PowerMeterStore (2011), *Home Energy Monitor Kits* [online], Optimum Energy Products Ltd, Accessed on: 20 Jan 2011, Available at: http://www.powermeterstore.com/c628/home_energy_monitor_kits.php

PRWeb (2011), *As Obama States the Need for Green Businesses, U.S. Power Services Company Launches a New Website to Help Companies Save Electricity* [online], Accessed on: 10 Mar 2011, Available at: http://www.prweb.com/releases/2011/01/prweb5002754.htm

Recurve (2011), Powering Home Energy Professionals, Accessed on: 9 Feb 2011, Available at: http://www.recurve.com/

ROODMAN, D.M; LESSEN, N. (1995), *A Building Revolution: How Ecology and Health Concerns are Transforming Construction*, Worldwatch Paper 124, Worldwatch Institute, Washington, DC, p. 5.

SEO, C.; MALEK, S.; MEDVIDOVIC, N. (2008), *Estimating the Energy Consumption in Pervasive Java-Based Systems*, In Proceedings of Sixth Annual IEEE international Conference on Pervasive Computing and Communications (Mar 2008), PERCOM, IEE Computer Society, Washington, DC, 243-247.

Science Daily (2009), *'Sleep Talking' PCs Save Energy and Money*, University of California - San Diego, Accessed on: 03 Nov 09, Available at: http://www.sciencedaily.com/releases/2009/04/090424114216.htm

TALEBI, M.; WAY, T. (2009), *Methods, Metrics and Motivation for a Green Computer Science Program*, Proceedings of the 40th ACM technical symposium on Computer science education, Publisher: ACM (2009), ISBN:978-1-60558-183-5, pp. 362-366.

TANENBAUM, A.S. (2001), *Modern Operating Systems*, 2nd Edition, Published by: Prentice Hall, ISBN: 0130313580

TAY, E. (2011), *Adding the green touch with technology* [online], GreenBusinessTimes.com, Accessed on: 19 Mar 2011, Available at: http://www.greenbusinesstimes.com/2011/04/26/adding-the-green-touch-with-technology-news/

VERMA, A.; AHUJA, P.; NEOGI, A. (2008), *pMapper: Power and Migration Cost Aware Application Placement in Virtualized Systems*, Proceedings of the 9th ACM/IFIP/USENIX International Conference on Middleware, p.243-264, ISSN:0302-9743

Watts-Up (2011), *Watts-Up* [online], Accessed on: 10 Apr 2011, Available at: https://www.wattsupmeters.com/secure/index.php

An Efficient Technique for Clustering High Dimensional Data Set

Dharmveer Singh Rajput dharmveer@iiitm.ac.in

P. K. Singh pksingh@iiitm.ac.in

M. Bhattacharya mb@iiitm.ac.in

ABV – Indian Institute of Information Technology and Management, Morena Link Road, Gwalior, 474010, India

Abstract: - In the modern world, advance technologies produce huge amount of data with many objects and dimensions. Traditional clustering algorithms do not perform well in the high dimensional data sets as similarity measures are no more meaningful, hence the data objects are equidistant from each other in high dimensions. Some traditional algorithms produce local optimum results as they start with random initial clusters centers. Relevant feature selection and selection of optimal initial clusters centers are two major issues of many partitioning clustering algorithms. In this paper, we propose a technique for selecting most relevant dimensions of data set and efficient initialization of clusters centers. When, we compare the results of proposed technique with existing techniques then our proposed technique produce clustering of better quality.

Keywords: - Clustering, feature selection, high dimensional data, initial centers, k-means.

I. INTRODUCTION

Clustering is an unsupervised data mining process, which creates the groups (clusters) of objects where objects having similar property belong to one group (cluster) and the objects having dissimilar property belong to another cluster. Some distance function, e.g., Euclidean distance is used to compute the similarity between two objects. The objects are more similar if they have less Euclidean distance.

Conventional clustering methods, such as partitioning algorithms and hierarchical algorithms use prior knowledge of the number of clusters. Partitioning algorithms start with the random selection of *k* objects, from the data set, as initial clusters centers and assign other objects to the nearest cluster center. These initial clusters are refined to minimize the sum of squared error among the clusters; it is done iteratively by updating the cluster centers using mean value of objects in the cluster and reassigning the objects to the clusters.

Hierarchical clustering algorithms create a tree structure, known as dendrogram of the objects in the data set. Initially every object belongs to its own cluster then successively two closest clusters are merged until all the objects belong to a single cluster. These methods create k clusters by cutting the dendrogram at the specific level.

All such traditional clustering algorithms create clusters based on the similarity or distance, hence fail to work on high dimensional data set as all objects become equidistant in high dimensional data set; there is no difference between closest and farthest data objects.

As partitioning clustering algorithms choose the initial clusters centers randomly, they suffer with the problem of local optimum clusters; in different runs they produce different results.

Dimensionality reduction is a major issue in the supervised as well as unsupervised data mining techniques. It reduces the size of the representation of the data objects which also results in less computational cost in further steps. However, a care should be taken that the reduced data set truly represents the original data set, i.e., there should be minimal descriptive inaccuracy.

Dimensionality reduction techniques may be categorized into two distinct groups: feature transformation (FT) and feature selection (FS). Feature transformation methods such as Principal Components Analysis (PCA) or Singular Value Decomposition (SVD) perform some linear transformation on high dimensional data set and produce low dimensional data set for further processing [1, 2, 4]. On the other hand, feature selection methods select the most relevant dimensions of the high dimensional data set based on ranking of the dimensions and leave the irrelevant dimension of the data set [5, 6].

Feature transformation methods produce a better reduced data set of the original data set as compared to feature selection methods, but the relationship between the original data set and the transformed low dimensional data set is very difficult to interpret.

Feature selection is a pre-processing step that uses ranking or weighting strategy for removing the irrelevant or noisy dimension of the data set for both supervised and unsupervised data mining task. For classification, it selects the dimensions which give the highest accuracy for a class, whereas in clustering, it chooses the dimensions which produce better clusters.

There are two well-known methods for unsupervised feature selection: filter methods and wrapper methods. Earlier methods compute the rank of every dimension of data set then select highest ranked dimensions [10], whereas later methods determine every subset of dimension of data set then select the subset which has better clustering tendency. These feature selection techniques provide better clusters but at a higher computational cost [8, 9].

One of the general solutions for initialization of clusters centers is to run the algorithm for different initial centers, and then select the clusters which give higher accuracy based on the objective function criteria.

In this paper, we propose a technique which selects discriminate dimensions of data set as well as finds efficient initial clusters centers to obtain good quality clusters. Our proposed technique consists of three phases; feature selection phase, centers initialization phase and refinement phase. In feature selection phase, it uses the ranking method for selection of relevant dimensions of data set, then uses trimmed mean of selected dimensions to obtain the efficient initial clusters centers in the centers initialization phase. Finally, these initial centers are used in the k-means algorithm to obtain final clusters.

Rest of the paper is organized as follows. Section II summarizes the previous relevant work and Section III presents the proposed technique. We analyze the performance of the proposed technique and compare the results with other existing clustering algorithms in Section IV. Finally, Section V gives the conclusion and further scope of the work.

II. LITERATURE REVIEW

Dimensionality reduction or projection techniques can transform large data of multiple dimensions into a smaller, more manageable set. Thus, we can uncover hidden structure that aids in the understanding and visualization of the data. Dimensionality reduction techniques such as Principal Component Analysis [1] and Multidimensional Scaling [2, 3, 4] are capable of handling the linear data, but nonlinear techniques for dimensionality reduction such as Locally Linear Embedding (LLE) [5] and Isometric Feature Mapping (Isomap) [6] can handle nonlinear data with some type of topological manifold, e.g., The Swiss Roll. However, LLE and Isomap both fail in some other type of data, e.g., sphere or torus data set [7].

Feature selection techniques such as filter approach and wrapper approach differ in the way that they evaluate a given feature subset. The filter method evaluates the rank of every dimension and then it selects the highest ranked dimension [8, 9]. On the other hand, wrapper method evaluates every feature subset of data set then selects one of the feature subset by estimating the generalization performance of the feature subset of the original data [10].

Gheyas et al. [11] introduce a hybrid technique based on the simulated annealing and genetic algorithm (SAGA), which uses GRNN classifiers for examining the candidate feature subset solutions, where higher accuracy solution represent the better solution. If two solutions have equal accuracy, then the solution having smaller number of features wins.

Apolloni et al. [12], present a method called Boolean Independence Component Analysis (BICA), which fetch Boolean bits with minimal joint entropy and consistence assignment from full dimensions, then it produces a vector of Boolean variable with unique assignment. This assignment is directly proportional to relevance of the feature and same assignment does not assign to another classification level.

Hu et al. [13], propose feature evaluation and selection technique based on the concept of soft-margin support vector machines and neighbourhood rough sets. This technique combines the concept of classification loss and neighbourhood margin which determine the classification accuracy of feature subsets.

Liu et al. [14], present an efficient non-linear dimension reduction technique called multi-layer isometric feature mapping (ML - Isomap), which automatically produce clusters by using rushes editing.

Sugiyama et al. [15], proposed Direct Density-ratio estimation with Dimensionality reduction (D3) which improve density-ratio estimation accuracy in high-dimensional data sets.

Kabir et al. [16] propose a technique which is based on the concept of the wrapper approach and sequential search strategy called Constructive Approach for Feature Selection (CAFS). It uses three layered feed-forward neural network that combines the feature selection with the neural network architecture by using correlation information.

Arai et al. [20] utilize all the clustering results of K-means in certain times even though some of them reach the local optima. To determine the initial centers for K - means, it transforms the result by combining with hierarchical algorithm.

Khan et al. [19], present an algorithm called CCIA, which first compute the mean and standard deviation of each dimension of data set and then partition the data by normal curve into some clusters to determine the similarity of data objects.

An algorithm by Barakbah et al. [17] first compute the grand mean of all data objects and then gradually selects the farthest object from all previously selected objects as initial clusters centers. This algorithm is based on the pillar designates approach.

Celebi [18], evaluate the efficiency of k-means algorithm as a colour quantization as well as different clusters centers initialization scheme.

To overcome the problems of algorithms in both the groups and meet the demand for feature selection for high dimensional data and efficient initialization of clusters centers, we develop a novel algorithm which can effectively identify the relevant dimensions and initial clusters centers of data set.

III. PROPOSE TECHNIQUE

Our proposed technique solve two problems of existing clustering algorithms, first it removes the irrelevant dimensions of high dimensional data set and then search the efficient initial clusters centers. The proposed technique uses the Fisher's criterion score to find out the relevant dimensions of data set and then uses the trimmed mean to select the efficient initial clusters centers. The results indicate that it produces better quality clusters in comparison to the existing clustering techniques.

Input: *N* x *D* data set, where *N* is number of objects in data set and *D* represents the dimensions of data set. And *K* is the number of clusters in data set.

Output: *K* optimal clusters of *d* dimension, where *d* is less than *D*.

Feature selection phase

1. Sort all dimensions of data set in ascending order, and divide every dimension in *K* parts.

2. Calculate the mean and standard deviation of the each part of all dimensions.

$$\mu_X = \frac{\sum_{i=1}^{n} X_i}{n}, \sigma_X = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{(n-1)}}$$

3. Calculate the *F* Score of each dimension and select highest score dimensions.

$$F_t = \frac{(\mu_1 - \sum_{i=2}^{k} \mu_i)^2 + \ldots.. + (\mu_j - \sum_{\substack{i=1 \\ i \neq j}}^{k} \mu_i)^2 + \ldots\ldots... + (\mu_k - \sum_{i=1}^{k-1} \mu_i)^2}{\sum_{i=1}^{k} \sigma_i^2}$$

Centers initialization phase

4. Sort highest score dimension and move the elements of other selected dimensions accordingly.

5. Divide selected dimensions into *K* partitions and calculate *l* percent trimmed mean of each group. Those trimmed mean are referred as initial clusters centers.

Refinement phase

6. Assign each object to its closest centroid based on minimum Euclidean distance.

7. Update the centroid by calculating mean value of objects in the cluster.

8. Repeat steps 6-7 until no change occur or no object move to other cluster.

First three steps of the algorithm selects relevant dimension of data set, steps 4 –5 find the efficient initial clusters centers and reduce the effect of outliers. Finally, clustering is done by steps 6 – 8.

IV. EXPERIMENTAL RESULTS

We take two standard dataset for our experiment: milk data set [21] and breast cancer data set [22]. We apply some existing clustering algorithm such as k-means algorithm, principal component analysis based k-means [28], singular value decomposition methods based k-means [29] and our proposed technique and compare the results using some well known quality measures. We summarize the results of all above mentioned clustering techniques with quality measures in table 1 and table 2.

A. Milk dataset

Milk data set contain 25 mammals and 5 ingredients of their milk, where mammals are represented objects *(O_1,O_2.........O_{25})* and ingredients of their milk are considered as the dimensions *(d_1,d_2......d_5)* of the data set. Here number of clusters is an input parameter which is equals to 4 (K = 4). So we have to cluster these mammals into 4 clusters based on the similarity in their ingredients of milk as shown in figure 1.

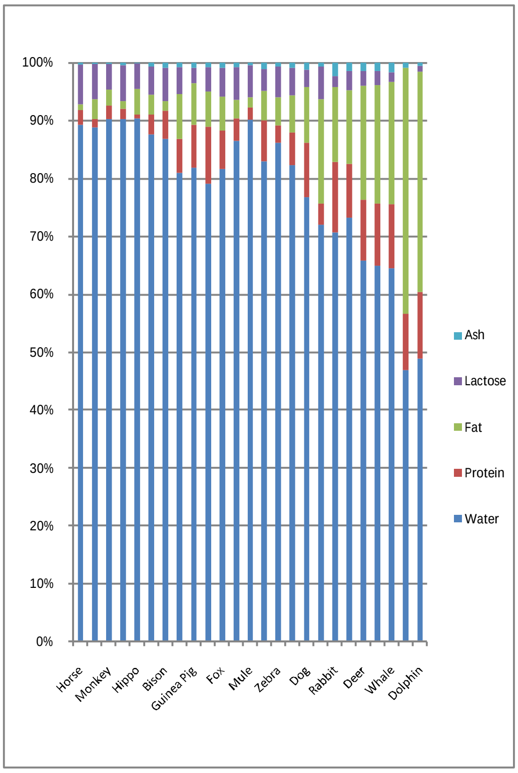


Figure 1. Milk Data Set

When we apply our propose technique on this data set, it first sort all dimensions in ascending order then divide each dimension in 4 (K = 4) classes. After that calculate mean and standard deviation of each class and find F score, then select highest score dimensions. Here, F score of dimensions d_1, d_2, d_3, d_4 and d_5 are 650.24, 190.26, 23.58, 166.57 and 74.89 respectively. So selected dimensions are $d_1$and d_2 (water, protein), because it's has highest F score. In second phase, sort the selected dimensions in ascending order according to highest score dimension and divide dimensions in K groups then take trimmed mean of each group. Here, the trimmed mean of 4 groups are (59.58, 9.35), (77.53, 9.15), (84.41, 5.05) and (89.34, 2.00).

These objects are referred as initial centers of clusters for C_1, C_2, C_3 and C_4 respectively. These objects also reduce the effect of outliers present in data set. Then in the third phase, initial cluster centers obtained in second phase are used in the k-means to find optimal clustering of data set. Then it takes only six iterations for convergence. The centers of the final clusters are as follows: C_1 = (45.65, 10.15), C_2 = (68.33, 9.55), C_3 = (81.18, 7.42) and C_4 = (88.50, 2.57). The objects contained in these clusters are *(O_{24}, O_{25}), (O_{18}, O_{19}, $O_{20,}$ O_{21}, O_{22}, O_{23}), (O_8, O_9, O_{10}, O_{11}, O_{14}, $O_{16,}$ O_{17})* and *(O_1, O_2, O_3, O_4, O_5, O_6, O_7, $O_{12,}$ $O_{13,}$ O_{15})* respectively, Clusters are shown in figure 2.

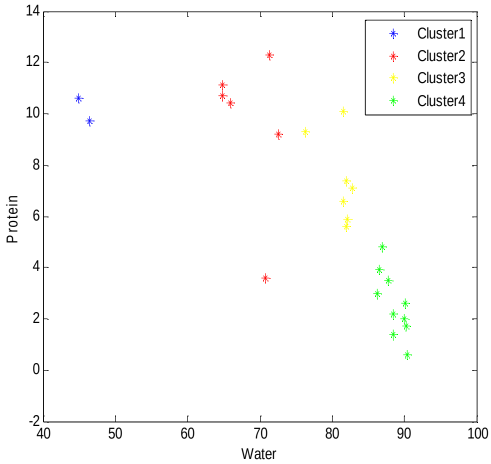


Figure 2. Clustering of Milk Data Set using Propose Technique

B. *Breast Cancer Dataset*

We also perform the experiment on standard data set of Wisconsin Diagnostic Breast Cancer (WDBC), which contain 198 objects and 35 dimensions. Every attribute of the data set has been determined by the digitized image of a fine needle aspirate (FNA) of a breast mass and it represents the cell nuclei involve in the image. Here, the number of clusters equals to 10 (K = 10), when we apply our proposed approach step by step on this data set then results are as follows: first calculate the F score of all dimensions and then select the dimensions which has maximum value of F score, here maximum F scores are 71251, 65162 of dimensions d_6 and d_{11}. After that we apply the next phase of approach for finding the initial centers. The trimmed mean of the selected dimensions are (0.0828, 0.0540), (0.0899, 0.0574), (0.0937, 0.0597), (0.0976, 0.0593), (0.1007, 0.0602), (0.1037, 0.0643), (0.1070, 0.0656), (0.1114, 0.0652), (0.1163, 0.0682) and (0.1256, 0.0727) of clusters C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9 and C_{10} respectively. Finally we apply k-means algorithm to cluster the data set according to these initial centers then the algorithm is converse is in 16 iterations. Final clusters centers are (0.0824, 0.0538), (0.0910, 0.0563), (0.0940, 0.0637), (0.0991, 0.0590), (0.1066, 0.0589), (0.1046, 0.0666), (0.1098, 0.0734), (0.1152, 0.0639), (0.1201, 0.0739) and (0.1386, 0.0787). And final clusters C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9 and C_{10} are contain 19, 28, 16, 35, 17, 25, 12, 24, 17 and 5 objects as shown in figure 3.

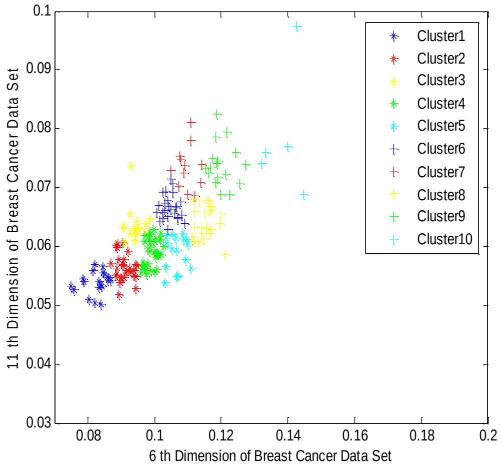


Figure 3. Clustering of Breast Cancer Data Set Using Propose Technique

C. Quality Measures

We apply five quality measures on the results of existing techniques as well as our proposed technique then compare the results as shown in table 1 and table 2.

Dunn index (D) [23]: Dunn Index is the ratio of d_{min} (which represent the smallest distance between two objects from different clusters) and d_{max} (denotes the largest distance of two objects from the same cluster).
The value of Dunn Index always lies between [0, 1] and maximum value of Dunn Index represents the better clustering.

$$D = \frac{d_{\min}}{d_{\max}} \qquad (1)$$

Davies - Bouldin index (DB) [24]: *Davies - Bouldin index* determined by the equation no. 2 as given below. Where, n, σ_i, σ_j and $d(c_i, c_j)$ represent the number of clusters, the average distance of all objects in cluster i to their cluster center c_i, the average distance of all objects in cluster j to their cluster center c_j, and the distance between clusters centers c_i and c_j respectively.
Here, small values of *DB* Index indicate that the clusters are compact, and there centers are far away from each other.

$$DB = \frac{1}{n}\sum_{i=1, i\neq j} \max\left(\frac{(\sigma_i + \sigma_j)}{d(c_i, c_j)}\right) \qquad (2)$$

Jagota index (Q) [25]: Jagota Index calculates the average distance of all objects in the cluster from its centroid. It determines the compactness and homogeneity of the objects in the clusters. Here, $|C_i|$ represents the number of data objects in cluster i, k denotes the total number of clusters, μ_i, x and $d(x, \mu_i)$ represents the centroid of i^{th} cluster, object in the cluster and distance between object x and the cluster centroid μ_i respectively.
Q will be small if the data objects in each cluster are close.

$$Q = \sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{x \in C_i} d(x, \mu_i) \qquad (3)$$

Siddheswar Index (S) [26]: This validity measure calculates the ratio of intra cluster distances and inters cluster distance.
Good clustering produces minimum value for this validity index.

$$S = \frac{intra}{inter} \qquad (4)$$

Sum of Squared Error [27]: Sum of Squared Error criteria are most popular criteria for calculating the homogeneity of the objects in the clusters.
This criterion should be minimizing for good clustering.

$$SSE = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n(s_i)} \left\| m_{ij} - \overline{s_i} \right\|^2}{N} \qquad (5)$$

Here, we show the comparison of results of existing clustering techniques and proposed technique on both milk data as well as Breast cancer data using five above mentioned quality measures. The results of milk data set and breast cancer data set are shown in table 1, Figure 4 and table 2, Figure 5 respectively. Here, it is clearly shown that our proposed technique produce better clustering.

Methods	Dunn Index	DB Index	Jagota Index	Siddh. Index	SSE
K–Means	0.4828	0.2665	13.56	0.0746	507.88
PCA + K-Means	0.4677	0.1794	12.11	0.0727	472.57
SVD + K-Means	0.3696	0.1847	15.14	0.0714	525.93
Proposed Technique	0.4984	0.1761	8.9695	0.0619	196.08

TABLE I. COMPARISION OF RESULTS ON MILK DATA

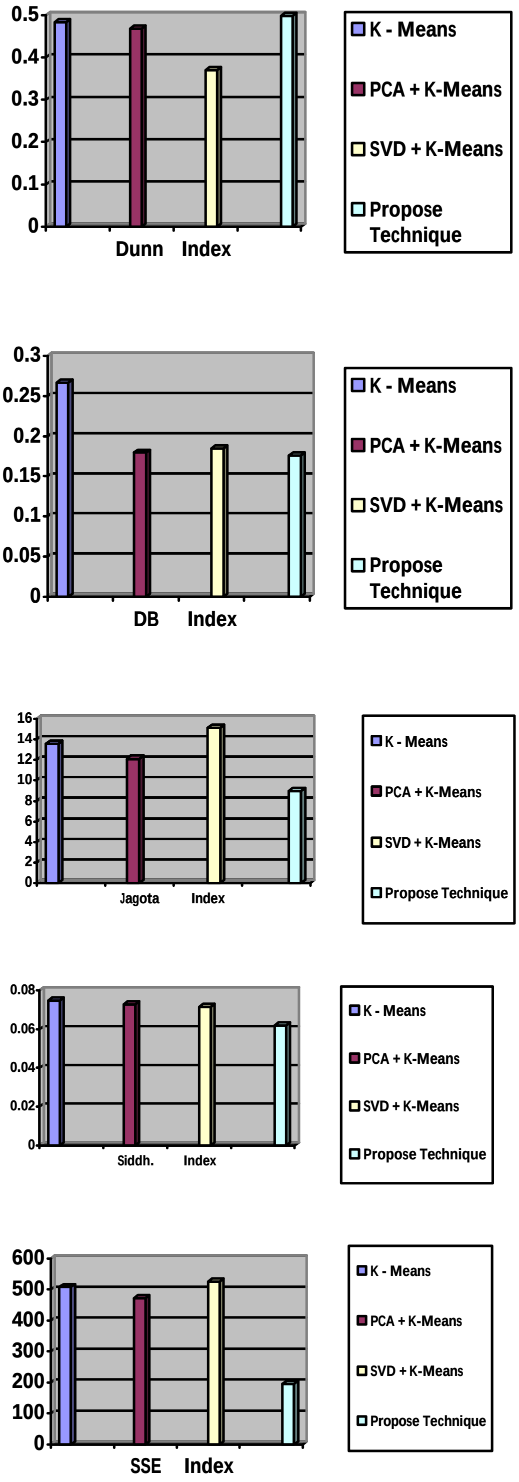


Figure 4. Comparison of Results on Milk Data Set.

Methods	Dunn Index	DB Index	Jagota Index	Siddh. Index	SSE
K–Means	0.0236	0.2343	1474.48	0.0364	5362660
PCA + K - Means	0.0253	0.2263	1368.84	0.0382	4981100
SVD + K -Means	0.0324	0.2257	1355.96	0.0326	4930040
Proposed Technique	0.0374	0.1585	911.65	0.0259	1626850

TABLE II. COMPARISION OF RESULTS ON BREAST CANCER DATA SET

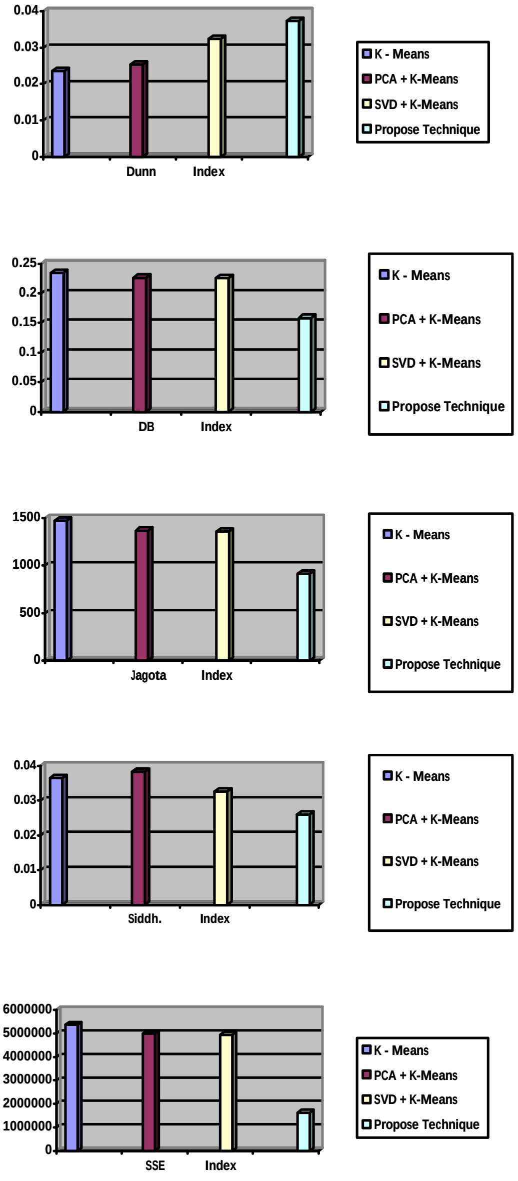


Figure 5. Comparison of results on Breast Cancer Data Set.

V. CONCLUSIONS

High dimensional data contain many noisy irrelevant dimensions, so the resulting clusters are hiding in sea of noise. And random selection of initial clusters centers gives the local optimal clustering. In this paper, we present a clustering technique, which provide the solution for both of the problems. It determines the relevant dimension and finds out the efficient clusters centers of the selected dimensions. The quality measures indicate the superior performance (good quality results) of the proposed algorithm in comparison to the existing algorithms.

REFERENCES

[1] Pearson, K. "On lines and planes of closest fit to systems of objects in space", Philosophical Magazine, Series B, 2(11), 559–572, 1901.

[2] Cox, T. F., and Cox, M. A. A., "Multidimensional scaling (2nd ed.). /CRC", New York: Chapman & Hall, 2000.

[3] Davison, M., "Multidimensional scaling", Florida: Krieger Publishing Company, 2000.

[4] Kruskal J. and Wish M., "Multidimensional scaling", London: Sage Publications, 1978.

[5] Roweis S. T. and Saul L. K., "Nonlinear dimensionality reduction by locally linear embedding", Science, 290(5500), 2323–2326, 2000.

[6] Tenenbaum J., de Silva V. and Langford J., "A global geometric framework for nonlinear dimensionality reduction", Science, 290(5500), 2319–2323, 2000.

[7] Saul L. K. and Roweis S. T., "Think globally, fit locally: unsupervised learning of low-dimensional manifolds", Journal of Machine Learning Research, 4, 119–155, 2003.

[8] Richard O. Duda and Peter E. Hart, "Pattern classification and scene analysis", Wiley, 1973.

[9] I. Kononenko and S. J. Hong,"Attribute Selection for Modeling", Future Generation Computer Systems, ISSN 0167 - 739X, 13 (2 - 3), pp. 181 - 195, 1997.

[10] I. Guyon, J. Weston and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines", Machine Learning, Vol. 46, S. 389 – 422, 2002.

[11] Iffat A. Gheyas and Leslie S. Smith, "Feature subset selection in large dimensionality domains", Pattern Recognition 43, 5 – 13, 2010.

[12] Bruno Apolloni, Simone Bassis and Andrea Brega, "Feature selection via Boolean independent component analysis", Information Sciences 179, 3815–3831, 2009.

[13] Qinghua Hu, Xunjian Che, Lei Zhang and Daren Yu, "Feature evaluation and selection based on neighbourhood soft margin", Neurocomputing 73, 2114–2124, 2010.

[14] Yang Liu, Yan Liu and Keith C.C. Chan, "Dimensionality reduction for heterogeneous dataset in rushes editing", Pattern Recognition 42, 229 – 242, 2009.

[15] Masashi Sugiyama, Motoaki Kawanabec and Pui Ling Chui," Dimensionality reduction for density ratio estimation in high-dimensional spaces", Neural Networks 23, 44-59, 2010.

[16] Md. Monirul Kabir, Md. Monirul Islam and Kazuyuki Murase, "A new wrapper feature selection approach using neural network", Neurocomputing 73, 3273–3283, 2010.

[17] Ali Ridho Barakbah and Yasushi Kiyoki, "A pillar algorithm for kmeans optimization by distance maximization for initial centroid designation", IEEE, 2009.

[18] M. Emre Celebi, "Effective Initialization of k-means for colour quantization", IEEE, ICIP, 2009.

[19] Khan, S.S. and Ahmad, A., "Cluster center initialization algorithm for K-means clustering", Pattern Recognition Letter, 2004.

[20] Kohei Arai and Ali Ridho Barakbah, "Hierarchical K-means: an algorithm for centers initialization for K-means", Reports of the Faculty of Science and Engineering, Saga University, Vol. 36, No.1, 2007.

[21] http://www.unikoeln.de/themen/statistik/data/cluster/milk.dat.

[22] http://mlearn.ics.uci.edu/databases/breast-cancer/wisconsin/wdbc.names.

[23] Dunn, "well separated clusters and optimal fuzzy partitions", Journal of Cybernetics, 4, 95-104, 1974.

[24] Davies D.L. and Bouldin D.W., "A cluster separation measure", IEEE Trans. Pattern Anal. Machine Intell, 1(4), 224-227, 2000.

[25] Arun Jagota "Novelty detection on a very large number of memories stored in a Hopfield-style network", IJCNN, 1991.

[26] Siddheswar Ray and Rose H. Turi, "Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation", ICAPRDT, pp. 27-29, 1999.

[27] C.J. Veenman, M.J.T. Reinders, E. Backer, "A maximum variance cluster algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1273-1280, 2002.

[28] Ding, C. And He, X. "k-means clustering via principal component analysis", In 21 international conference on machine learning, New York, ACM, 2004.

[29] Andrews, H. And Patterson, C. "Singular value decompositions and digital image processing", IEEE Trans. On Acoustics, Speech and Signal processing, 24:26-53, 1976.

FinFuzRelat: A new Fuzzy Relational Algorithm for Financial Data Analysis

Dr. Nabil EL KADHI
Computer Engineering Department Chairperson
AHLIA University Manama, Bahrein
nelkadhi@ahliauniversity.edu.bh

Dr. Nahla EL ZANT MIS Department
AHLIA University Manama, Bahrein
nahla@ahliauniversity.edu.bh

Dr. Naser NAJJAR
Banking and Finance Department
AHLIA University Manama, Bahrein
najjarnaser@hotmail.com

Abstract— This paper suggests an add-in solution for financial data and Depeche analysis. The idea is to adopt a former relational data analysis algorithm in order to apply it to financial situation analysis. This algorithm is used to filter and classify financial documents in order to help decision making with partial data and ratios.

Keys words: Data Mining, Research Algorithm, Financial Rate, Fuzzy Relation

1. *Introduction*

Decision making process are nowadays one of the most crucial and important tasks in any company. In fact with a large globalization and a complex interaction between related events it becomes more difficult to go through a precise event analysis in order to take suitable decision. As we can notice it [6] ERP systems and Business Intelligence tools [9] are used more and more as an integrated and mandatory component of any efficient Information Systems. ERP [15] comes in general with a set of modules helping in having a coherent Information System able to handle heterogeneous data and business processes. Beside the complexity and diversity of such tools [7, 12, 10], we need to generate boarding tables in order to assist manager during any decision making process. Business intelligence tools that could be or not an integrated part to ERP solutions are offering a set of functionalities for gathering and reporting data. Unfortunately, those tools does not, in general, include any decisional support system helping in transforming boarding tables to knowledge and decision. In fact, interpretation and related analysis with similar previous cases remains one of the hard decision making process. In this paper, we start by presenting in section 2 the Business Intelligence tools and their use for financial data analysis in order to easily identifier their advantages and weaknesses. This leads us to introduce a specific algorithm (see section 3) named KMRCRelat [1] that has been initially used for fuzzy redundancy detection in bioogical sequences. Section 4 introduces the concept of ‘fuzzy event’ or fuzzy codification as we intend to apply it to financial data. Though a running example we will illustrate the importance of considering any resemblance with previous situations in a decision making process. This will allow us to go through a specific codification/representation of financial rates that are mostly used in depeche. Our suggested algorithm *FinFuzRelat* is presented in section 5. A general description of the redundancy search algorithm will be given in section 6 as well as the suggested relational subgroups in the financial context. The conclusion argues the need for a training step before going through a complete fuzzy codification and we suggest a possibly efficient way to automate partially the codification step as future work.

2 *ERP/ B.I tools and financial data analysis*

Enterprise resource planning (ERP) [13, 14]) is a term used to describe a software system providing multiple application modules to run a business in the areas of financial management, logistics, manufacturing, Human Resources (HR) and extended supply chain operations. The term Enterprise Resource Management came from the concept that inventory, time, and people are all resources of the company and an integrated software solution should be a tool to manage those resources. In another definition: ERP systems integrate primary business applications. All the applications in an ERP suite share a common set of data that is stored in a central database. A typical ERP system provides applications for accounting and controlling, production and materials management, quality management, plant maintenance, sales and distribution, human resources, and project management. ERP can apply their organizational/integration role major business processes, such as employee benefits or financial reporting as shown figure A.

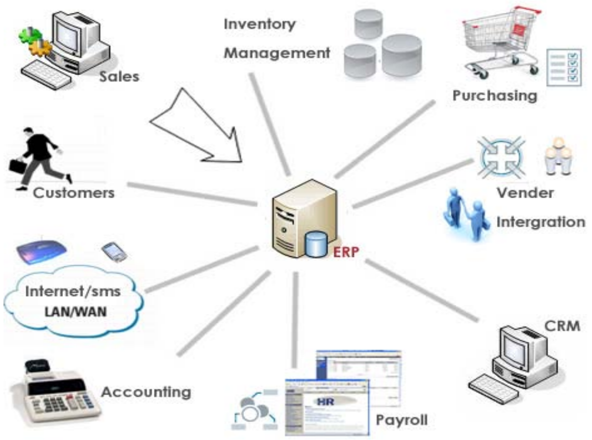


Figure A: ERP Solution Overview

ERP softwares are in general organized as a set of integrated software modules (Figure 2). Each ERP software module mimics a major functional area of an organization. Common ERP modules include modules for product planning, parts and material purchasing, inventory control, product distribution, order tracking, finance, accounting, marketing, and HR. The financial module is the core of many ERP software systems. It can gather financial data from various functional departments, and generate valuable financial reports such balance sheet, general ledger, trail balance, and quarterly financial statements. Financial reports can be used to generate a set of specific ratios classified in different categories. As mentioned in [16]. Five major ratios categories could be cited:

- Leverage financial ratios.
- Liquidity financial ratios.
- Operating financial ratios.
- Profitability financial ratios.
- Solvency financial ratios.

When analyzing financial situations, and in order to take appropriate decisions, a set of computation are done to fulfill an overall analysis. Unfortunately, in some situations some data are missing making the overall analysis almost impossible or partially incomplete. Because of such feature and because of the high cost of ERP, Business intelligence tools are often used as a high level integration tool for producing valuable reports even they still suffer from being unable to handle incomplete data.

Business intelligence (BI) refers to skills, technologies, applications and practices used to help a decision maker to acquire a better understanding of its commercial context. Business Intelligence may also refers to the collected information itself. A BI application provides historical, current, and predictive views of business operations. Common functions of BI applications are reporting, analytics, data mining, business performance management, benchmarks, text mining and predictive analytics. BI software tools (fig 3) are traditionally associated with in-depth analysis of historical transaction data, supplied by either a data warehouse or an online analytical processing server linked to a database system.

There is a very fine line distinguishes between BI tools and data mining techniques. In this paper, we suggest a new algorithm for data analysis that could be applied as an additional filter or decision support system beyond any data/report produced by BI tools. In this paper we suggest an *add-in* solution or process that fits with specific financial data. We argue that previous situation and similar events/decisions are of a high importance in financial analysis. In fact, what makes an analyst more accurate than other is his instinct that was build based on experience and extra ability of

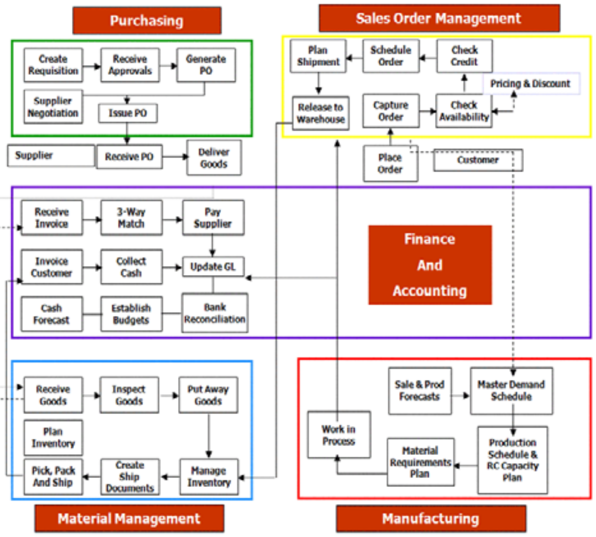


Figure B: ERP Modules architecture sample

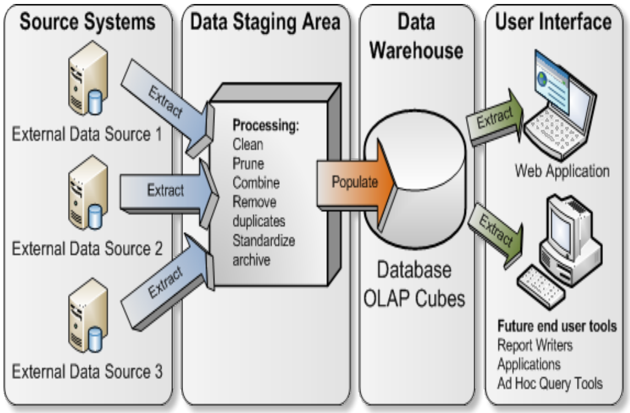


Figure C: Business intelligence process

build based on experience and extra ability of linking events and situations even if they didn't seem to be directly related. To have accurate analysis, the analyst should take into consideration previous linked events and situations even if they didn't seem to be directly related. In crises [8, 14] or critical situations, experts relay usually on '*previous cases'* similarity analysis and '*already* seen' scenarios to justify and give more validity to their decisions and suggestions. Basically, they go through a huge number of Depeche's and try to find out similar ratio values in order to link different situation. Our concern is with some conflicted situations where the available data didn't allow decision makers to compute the same financial ratio. For example, let try to make decision about two situations where in one only the ROC (Return Of Capital employed ratio) can be computed and in another one just the Rate Of Interest can be computed. Any regular decision making system or data mining system will not consider those situations as similar or linked since the computed ratio are different. A deep analysis can help us noticing that the two ratios belong to the same group or that they are correlated in some way. *FinFuzRelat* aims to take advantages from such consideration by extending KMRCRelat a relational redundancy search algorithm to handle financial concepts. Such kind of situation leads us to suggest a particular data mining /data analysis algorithm for fuzzy redundancy. As a starting point we consider KMRCRelat [1] algorithm and its specific application for microsatellite detection.

3 *KMRCRelat Algorithm review*

Finding regularities in sequences is an important problem in many areas. In most cases entities forming sequences are described as symbols. Sequences can represent, after some representational changes, any kind of sequential objects. KMRCRelat deals with cases in which items in the sequence are not only described as symbols but also through what relates them to the other items in the sequence. KMRCRelat introduces the concept of relational patterns, namely flexible relational words that extend words and flexible words. KMRCRelat consider what relates the items at positions i and j in a sequence S defined on an alphabet ∑. In fact, each pair (i, j) of items (with i < j) in a sequence, has a value $r_{i,\,j}$ (called a relational value) belonging to a relational alphabet noted Σ_R. We represent the relational values of a sequence as a set of delay vectors Rd where for each vector R_d, $R_d[i]$ contains the relational value $r_{i,\,i+d}$. $R_0[i]$ is simply the non relational value s_i . ***Example*** Let Σ_R= {rb, rs, ra}.The relational information about the 3-length sequence S is represented for instance as:

R_2 = *rs*
R_1 = *rb* - *ra*
R_0 = *S* = *a* - *a* - *b*

This means that the relation $r_{1,2}$ between position 1 and 2 has the value *rb*, that $r_{1,3}$= *rs*, and that $r_{2,3}$= *ra*. Let items of the sequence be events that each have some duration and let us suppose that $r_{i,j}$= *rb* means that event at position *i* finishes *before* event at position *j*, that $r_{i,j}$= *rs* means that the event at position *i* finishes at the *same* time as the event at position *j*, and that $r_{i,j}$= *ra* means that the event at position *i* finishes after the event at position *j*. Non relational values *a, b* simply are labels of the events. Then the previous sequence S is such that: the first event has label *a* and finishes before the second event, and finishes at the same time as the third one. The second event has label *a* and finishes after the third one. The third event has label *b*. *S* corresponds then to the following configuration of the events:

- *a----------------------------.-end*
- *a-------------------------------------.-end*
- *b--------------------.-end*

Note that this information can be inferred from the ending time of the events. So here the relational values can be computed whenever necessary and do not need to be stored. In what follows we suppose that when considering a sequence *S* we either have access or we can compute all the corresponding delay vectors R_d.

3.1 KMRCRelat main idea

The main idea of ***KMRCRelat*** is to find the repeated relational flexible words in an n-Length sequence S. Searching repeated relational flexible words consist of finding all occurrences of this word in S. The search of repeated words in KMRCRelat is based on recurrent waves. In fact in order to find relational flexible words of length k, ***KMRCRelat*** relay on finding occurrences of relational flexible words of length k' with k'= k-1. Let M be a k-length word of a sequence S, the relation between the first and the second element of M will be presented by $M_{0,1}$. The relations between the elements at position i and i+1 are also presented by $M_{i,\,i+1}$ for all i ∈ [0..|M|].

3.2 Theoretical bases of KMRCRelat algorithm

A relational flexible word is repeated q time in a sequence (q-repeated word) if we can found q occurrences of this word in the sequence. ***KMRCRelat*** apply a basic lemma in k loops in order to find k-length repeated words.

3.3 KMRCRelat STEPS

KMRCRelat is a recurrent algorithm. In this section we present different steps of the algorithm.

Step 1: the goal of this step is to build a vector V[n] where n is the length of the word. To construct 1-length repeated words we consider n as s equal to one. In fact, any component of the alphabet is an element of one or many alphabet cover. For each alphabet cover, we will affect a number. Each element i of the vector V[1] presents the number of alphabet covers that contain s_i. In other word, each position i of V[n] presents a stack that contains all alphabet cover that contain the element at position i of the sequence. Each element i of V[n>1] presents a stack that contains all number of groups containing the occurrence i.

For example: let Σ = {a, b, c}, q= 2 and
S= a a b a a b a c
C= { C1={a, b}, C2={b, c}}: there are two groups of words of length 1:C1= 1, C2= 2.V[1] will be presented by figure 2.

a	a	b	a	a	b	a	c
1	2	3	4	5	6	7	8

Figure 2: vector V[1].

Step 2: in the second step, P-stack is used to gather the occurrences of n-length ***repeated flexible words*** (at the beginning n=1) found in the sequence. P-stack will contain all occurrences groups of repeated words. The size of P-stack is equal to the number of the existing alphabet cover. P-stack for 1-length repeated words for this example is presented in figure 3.

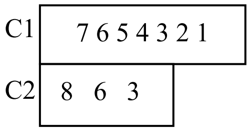


Figure 3: P-stack.

Step 3*: the third step consists to find the occurrences of different (n+1)-length ***repeated flexible words in the sequence. These occurrences will be presented in Q-stack. Q-stack will contain the same number of groups as P-stack. We keep only the groups verifying the repetition constraint. Thus means that we keep only groups that contains q occurrences. For finding all occurrences of repeated words the algorithm use P-stack and the vector V[1]. For all occurrence i of P-stack, we take the number of alphabet covers at position i+1 of V[1]. Then we put the occurrence i in the group of i+1. Q-stack of our example is presented in figure 4.

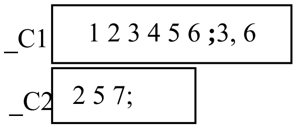


Figure 4: Q-stack.

The first group of Q-stack that contains {1, 2,…} presents the occurrences of the 2-length repeated word M= C1 C1, the second group that contains {3, 6} presents the occurrences of the 2-length repeated word M= C2 C1 and so on. Groups will be then filtered in order to eliminate all groups that does not respect the quorum q. For example, if q= 3, the second group (3,6 in C1[1 2 3 4 5 6**; 3 6**]) will be deleted.

***Step 4*:** a vector W[n], where n represents the length of the words, will be constructed in this step. W[n] represents the relations between the occurrence i and i+n (n=1 for searching 2-length flexible repeated relational words).

We affect a number for each relational cover. Each element of W is a stack containing all number of relation covers that contain the relation between the components i and i+n.

Let $G(\Sigma_R)$= {Cr1={r1, r2}, Cr2={r2, r3}}, the relations between elements, in such case, is presented by figure5.

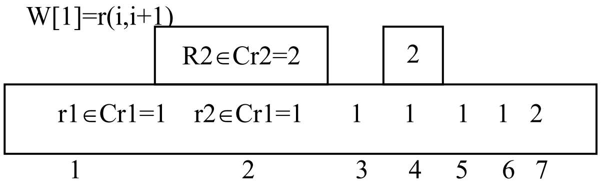


Figure 5: vector W[1].

For all W[n] the position i presents a stack of relational groups that contains the relation between the elements at position i and i+n.

Step 5*:** the fifth step of the algorithm find the ***repeated relational flexible words in the sequence. The occurrences of different repeated relational flexible words will be presented in Qrelat-stack. The algorithm use W and Q-stack in order to construct Qrelat-stack,. In fact, for all occurrence i of Q-stack, we take all numbers of relational covers at position i of W. Then we insert the occurrence i in the corresponding relational group. Qrelat-stack of our example is presented in figure 6.

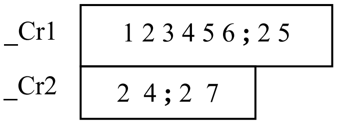
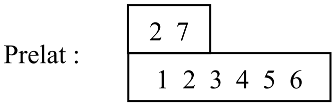


Figure 6: Qrelat-stack.

The group that contain {2, 7}, for example, presents two occurrences of a relational flexible word M presented by:

Cr1
M= C1 C1

***Step 6*:** in this step, previously founded groups will be filtered. This step delete all groups that are included in other ones. The sequence will be represented by an indexed vector based on the retained groups. In fact, we will attribute a number at each group and then we will use them to generate the indexed sequence representation. The algorithm stops when we have found repeated longest flexible relational words.



This step will remove any group that contains less than q occurrences. For our example, the groups {2, 5} and {2, 4} will be filtered because they are included in the group {1, 2, 3, 4, 5, 6}. We obtain, at the end of this step, two groups in Prelat-stack:
The second group represents the relational word :

Cr2

M'= C1 C2

4. Using Kmrcrelat Word Train Searches For Financial Analysis

The term ***financial crisis*** [11] is applied broadly to a variety of situations in which some financial institutions or assets suddenly lose a large part of their value. In the 19th and early 20th centuries, many financial crises were associated with banking panics, and many recessions coincided with these panics. Other situations that are often called financial crises include stock market crashes and the bursting of other financial bubbles, currency crises, and sovereign defaults. Many economists have offered theories about how financial crises develop and how they could be prevented. Data Mining, Artificial Intelligence and Genetic Algorithms apply the concepts of evolution to solve mathematical problems. Those technologies has been used to solve different problems in different biological, financial, and business areas. Data mining is used today by companies, with a strong consumer focus, retail, financial, communication, and marketing organizations. Data Mining enables companies to determine relationships among internal factors such as price, product positioning, or staff skills, and external factors such as economic indicators, competition, and customer demographics. It enables companies to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables us to "drill down" into summary information to view detail transactional data. KMRCRelat can be modified and used to help companies to detect future events and avoid future crisis.

4.1 Fuzzy codification for financial event

As mentioned in section 1, financial ratios are classified in 5 groups according to the ratio nature or category. This will be our first grouping criteria. KMRCRelat allows users to define a set of groups or equivalences. *FinFuzRelat* the derived algorithm, will follow the same approach by grouping financial data and defining the potential existing relations between them. The ratios will be grouped into different similarity groups.

The first algorithm parameter will specify that the set of used relations R1, R2, R3 (for example) belong to the group G1, R5 R6, R7 to the group G2 and so on. For example, we will explicitly mention through the parameter sets that the ratios CR and QR are in the same group. The financial considered groups are codified as represented table 1

The first step of the algorithm will consider as linked or related any depeche's where ratios from the same group are mentioned. **This will be the first filter to decided which documents should be kept or analyzed.** This will reduce or classify documents according to financial easy known criteria. From a set of non classified and non structured documents we will have a grouped or classified sets this will reduce the research space and will improve our algorithm efficiency and accuracy. Even this grouping criteria is not new or exceptional, it is important to include it as an initial grouping criteria. Let us now introduce our fuzzy decision making criteria. The idea is to analyze the previous ratio according to the used computational parameters and to define 'new' fuzzy groups. A fuzzy group includes all the ratios that have at least a shared parameter. The Fuzzy Groups (FG) are defined as following:

FG1: LVFR LFR
FG2: PFR, PrfFR
FG3: PrFFR, SFR

The mean of fuzzy groups is that when analyzing depeches and situation, an equivalent value of any ratios from LVR group will be considered as similar to any ration on LFR group. In fact in case that computing the same ratio between two different situations is not possible, we can replace the non computable ratio (decision making level) by any other ration that belongs to the same fuzzy group FGi. The fuzzy group can be detailed and made more accurate by defining sub fuzzy groups between ratios belonging to different groups in a more tiny way. This step is the most difficult and crucial one among the efficiency of our algorithm. In fact, this step is equivalent to a knowledge extraction and codification step in which expert will describe their non deterministic criteria and transform them to a set of fuzzy groups.

After defining and specifying the fuzzy groups, we have to clearly identify our ‘similarity’ ratios. KMRCRELAT uses a similarity factor to identify word-trains that represent set of similar repeated groups. In the case of financial Depeche’s, similarity will be defined based on first the previously FG defined group as well as a fuzzy rate (between 0 and 1). The rate is initialized by 1 and will be decreased by 0.0x each time a fuzzy group relation is used. Practically, any time we are in a situation, where we are not able to compute or compare the same ratio between two Depeche’s or situation, and that we are using an equivalent rate from our previously defined group, we will decrease the similarity rate (initialized to one) by a specific value (0.0x). We assume that financial expert will also provide a set of fuzzy penalty rates (0.0x) for each grouping set. The rate is in fact measuring the disparity or distance that should be considered when dealing with the two ‘fuzzy similar’ rates.

It is clear from the described FunFuzRelat process that document classification accuracy is related to the accuracy of fuzzy group definition and penalty rate establishment. When discussing and experimenting the process with some financial expert, we face the problem that experts claim being ‘accurate’ and almost ‘sure’ from their defined rates even though we notice that the same ratios, and when linked to the same fuzzy group, may not be assigned the same penalty rate by different experts.

5. *Conclusion*

In this paper, we intend to introduce a new classification/decision making filter for financial Depeche’s analysis. The main goal of our work is to adapt a relational redundancy detection algorithm to data mining in financial area.

After a brief overview of BI solutions and their major use in financial information system, we argue the need of non deterministic tools for financial data analysis. We recall a relational algorithm for redundancy detection step and we present a fuzzy based extension to it.

Our FinFuzzRelat algorithm introduces the concept of fuzzy similarity between financial ratios in order to allow decision making in case of partial ratio availability.

A key success of our approach relay on an accurate and long training step. In fact, after a set of experimentations we came out to conclude that such fuzzy groups are hard to be automatically defined and may lead, in case of weak definition, to wrong decision or classification.

Actually, FunFuzzRelat implementation has been realized in C++ language and we are focusing on financial fuzzy group definition. We do believe that expert should be assisted during this step by using some training techniques such as neural networks [2, 3, 4, 5] or symbolic training. In conjunction with the Business college at Ahlia University, we are selecting a set of documents and Depeche’s including a high rate of computed ratios in order to be analyzed by different researchers and to extract an initial sub set of fuzzy groups. When done, the obtained results will be used to train our classification tool in order to tune our fuzzy parameters.

Grp	Ratio	Formula
LVFR	*Curent Ratio (CR)*	$\frac{\text{Total Current Assets}}{\text{Ttal Current Lialability}}$
	Quick Ratio (QR)	$\frac{\text{Cash} + \text{ Accounts Receivable } (+\text{any other Quick Assets})}{\text{Current Lialability}}$
LFR	*Debt to Equity*	$\frac{\text{Total Liabilities (Debt)}}{\text{Net Worth (Total Equity)}}$
	EBIT/Interest	$\frac{\text{Earnings Before Interest Taxes}}{\text{Interesr Charges}}$
	Cash Flow to Current Maturity of Long-Term Debt	$\frac{\text{Net Profit} + \text{Non Cash Expenses}}{\text{Current Portion of Long} -\text{term Debt}}$
PFR	*Gross Profit Margin*	$\frac{\text{Gross Profit}}{\text{Total Sales}}$
	Net Profit Margin	$\frac{\text{Net Profit}}{\text{Total Sales}}$
PrfFR	*Return on Assets*	$\frac{\text{Net Profit Before Taxes}}{\text{Total Assets}}$
	Return on Equity	$\frac{\text{Net Profit Before Taxes}}{\text{Net Worth}}$
SFRs	*Accounts Receivable Turnover*	$\frac{\text{Total Net Sales}}{\text{Accounts Receivable}}$
	Accounts Receivable Collection Period	$\frac{365\text{ Days}}{\text{Accounts Receivable Turnover}}$
	Accounts Payable Turnover	$\frac{\text{Cost of Goods Sold}}{\text{Accounts Payable}}$
	Days Payable	$\frac{365\text{ Days}}{\text{Accounts Payable Turnover}}$
	Inventory Turnover	$\frac{\text{Cost of Goods Sold}}{\text{Inventory}}$
	Days Inventory	$\frac{365\text{ Days}}{\text{Inventory Turnover}}$
	Sales to Net Worth	$\frac{\text{Total Sales}}{\text{Net Worth}}$
	Sales to Total Assets	$\frac{\text{Total Sales}}{\text{Total Assets}}$
	Debt Coverage Ratio	$\frac{\text{Net Profit} + \text{Any Non} -\text{Cash Expenses}}{\text{Principal on Debt}}$

T
T
a
b

Table 1: Financial Ratios in and Groups

References

[1] Extension And Use Of Kmrcrelat Algorithm For Biological Problems', Nahla El Zant El Kadhi, Nabil El Kadhi, and Pierre-Antoine Gourraud, extended version, JCMSE, Volume 6, PP 157-170, 2006.

[2] Application of fuzzy neural networks for predicting seismic subsidence coefficient of loess subgrade, Tian-Feng Gu; Jia-Ding Wang; Natural Computation (ICNC), volume 8, pp1556 - 1559 2010.

[3] Artificial Neural Networks for Forecasting of Fuzzy Time Series, U. Reuter, B. Möller, Computer-Aided Civil and Infrastructure Engineering, Volume 25, Issue 5, pages 363–374, July 2010.

[4]http://beginnersinvest.about.com/od/ financialratio/a/ratiocategories.htm

[5] Boudreau, M.C., & Robey," Organizational transition to Enterprise Resource Planning Systems: Theoretical choices for process

research" Paper presented at the International Conference on Information Systems (ICIS) (1999)

[6] Business-Software ,"Top 12 ERP Software Vendors – 2009" , Business-Software Research Center (2009)

[7] Business-Software ,"Top 12 ERP Software Vendors – 2009" , Business-Software Research Center (2009)

[8] Charles P. Kindleberger and Robert Aliber (2005), *Manias, Panics, and Crashes: A History of Financial Crises*, 5th ed. Wiley, ISBN 0471467146.

[9] Galit Shmuli, Nitin R. Patel, Pete C. Bruce, “Data Mining for Business Intelligence” , Wiley 2008.

[10] Jim Mazzullo , "SAP R/3 for Everyone: Step-by-Step Instructions, Practical Advice, and Other Tips and Tricks for Working with SAP" ,Prentice Hall PTR; illustrated edition edition (July 25, 2005)

[11] Luc Laeven and Fabian Valencia (2008), 'Systemic banking crises: a new database'. International Monetary Fund Working Paper 08/224

[12] Marianne Bradford, "Modern ERP: Select, Implement & Use Today's Advanced Business Systems" , Lulu.com (September 2008)

[13] Marzi, H.; Turnbull, M.; Marzi, E.; , "Use of neural networks in forecasting financial market," Soft Computing in Industrial Applications, 2008. SMCia '08. IEEE Conference on , vol., no., pp.240-245, 25-27 June 2008

[14] Piszczalski, Margin, “ERP: Consider Connectivity; Enterprise Resource Planning Systems”, Automotive Manufacturing & Production, April 1997

[15] Robert Jacobs , David Clay Whybark,Clay Whybark,Robert Jacobs,"Why ERP? A Primer on SAP Implementation" ,McGraw-Hill/Irwin; 1 edition (January 6, 2000).

[16] Zhou Yixin, Jie Zhang, "Stock Data Analysis Based on BP Neural Network," iccsn, pp.396-399, 2010 Second International Conference on Communication Software and Networks, 2010.

Frequent Patterns Mining over Data Stream Using an Efficient Tree Structure

M. Deypir[1], **M. H. Sadreddini**[2]
[1,2] Department of Computer Science and Engineering, Shiraz University, Shiraz, Fars, Iran

Abstract - *Mining frequent patterns over data streams is an interesting problem due to its wide application area. In this study, a novel method for sliding window frequent patterns mining over data streams is proposed. This method utilizes a compressed and memory efficient tree data structure to store and to maintain sliding window transactions. The method dynamically reconstructs and compresses tree data structure to control the amount of memory usage. Moreover, the mining task is efficiently performed using the data structure when a user issues a mining request. The mining process reuses the tree structure to extract frequent patterns and does not need additional memory requirement. Experimental evaluations on real datasets show that our proposed method outperforms recently proposed sliding window based algorithms.*

Keywords: Data Stream Mining; Frequent Patterns; Sliding Window; Patricia Tree

1 Introduction

The Apriori [1] is a well-known algorithm for solving frequent itemset mining problem in static databases. An itemset (set of items) is frequent in a database if the number of its occurrences in the database is greater than a user given threshold. The frequent patterns mining in data streams is more challenging than static databases. This is due to dynamic nature of data streams and unbounded amount of incoming data. Commonly used approaches to handle and model data streams for frequent itemsets mining is sliding window model [2, 3, 4, 5]. In this paper, we have proposed a new data structure namely PatDS (Patricia tree for Data Stream mining) for frequent patterns mining which adapts a special type of prefix tree for data stream processing. Moreover, an adapted version of FP-growth [6] algorithm is also proposed to extract frequent patterns from the most recent transactions stored in PatDS when the user requests. Our proposed method provides better results in terms of memory usage and run time with respect to recently proposed algorithms. The reminder of paper is as follows. The next section presents a review on recently proposed frequent patterns mining algorithms over data stream. Problem statement is described in Section 3. In Section 4 the new method is introduced. The experimental results are presented in Section 5. Finally Section 6 concludes the paper.

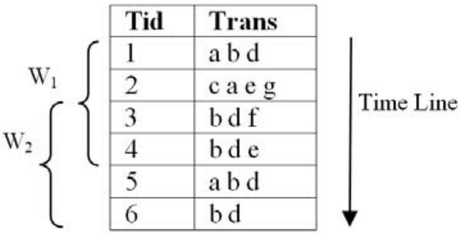
2 Related works

Han et al. introduced a novel algorithm, known as the *FP-growth* [6] method, for mining frequent itemsets in static databases. The *FP-Growth* (Frequent Pattern Growth) method is a depth-first tree based search algorithm for mining patterns without candidate itemset generation. In this method, a data structure termed the *FP-tree* is used for storing frequency counts of items belong to the input database. In [7] a new tree based algorithm for frequent itemset mining called *PatriciaMine* was introduced. It uses *Patricia Tree* [8] instead of the *FP-Tree* to store frequent items of transactions. This data structure is used since it needs smaller memory with respect to *FP-Tree* due to have capability to store more than one item in a node of the tree. Additionally, applying pattern growth method on this tree structure leads to better run time due to smaller tree structure.

The first algorithm for frequent patterns mining over data streams are proposed by Manku et al [9] where the authors start with frequent items mining and then extend their idea to frequent itemsets mining. Lin et al. [4] proposed a method for mining frequent patterns over time sensitive sliding window. In their method the window is divided into a number of batches for which itemsets mining is performed separately. The *estWin* [2] finds recent frequent patterns adaptively over transactional data streams using sliding window model. This algorithm requires the significant support in addition to the minimum support threshold to adaptively maintain the approximate frequent patterns. There are a number of algorithms in the literature in which the *FP-Tree* structure is adapted to store sliding window information. For the mining phases these method use the same *FP-Growth* algorithm to extract frequent patterns. One of these methods is the *DSTree* [3]. In this algorithm, transactions within a window are divided into a number of batches (or panes) and the information about every batch is maintained in a prefix tree. Nodes of the prefix tree show items in the transactions which are sorted in a canonical order. Transactions are inserted to and removed from the tree in a batch by batch manner. Frequent itemsets are mined using the *FP-Growth* method when the user requests. Another recently proposed prefix tree based algorithm is the *CPS-Tree* [5] which is more efficient than the *DSTree*. This method is similar to the *DSTree* algorithm but reconstructs the prefix tree to reduce its memory requirements as the incoming data stream is changed. In the *CPS-Tree*, nodes are sorted in descending order of their support. For controlling the main memory usage, the tree is dynamically reconstructed to maintain support descending order of nodes. In this study, a new method to extract set of all frequent itemsets over sliding window is proposed. Our algorithm adapts *Patricia Tree* for frequent itemsets mining over sliding window on transactional data streams.

3 Problem statement

Let $I=\{i_1,i_2,\ldots,i_m\}$ be a set of items. Let S be a stream of transactions in sequential order, wherein each transaction is a subset of I. For an itemset X, which is a subset of I, a transaction T in S is said to contain the itemset X if $X \subseteq T$. The support of X is defined as the percentage of transactions in S that contain X. For a given support threshold s, X is frequent if the support of X is greater than or equal to s%, i.e., if at least s% of transactions in S contain X. Transaction sensitive sliding window over data stream S contain $|W|$ recent transactions in the stream, where $|W|$ is the size of the window. The window slides forward by inserting a new transaction and deleting the oldest transaction from the window. Due to efficiency issues, instead of a single transaction, the unit of insertion and deletion can be a pane (or batch) of transactions. The current transactional window over the data stream S is defined as $TW_{n-w+1} = \{T_{n-w+1}, T_{n-w+2}, \ldots, T_n\}$ where n-$w+1$ and T_i are the window's identifier and i th transaction in the S, respectively. In fact the window contains the n most recent transactions of the data stream. An itemset X is said to be frequent in TW if $Sup(X) \geq s$, where $Sup(X)$ and s are support of X in TW and the minimum support threshold, respectively. Thus, having a transactional sliding window TW and a minimum support threshold specified by the user, the problem is defined as finding all the frequent itemsets that exists in the recent TW.



Tid	Trans
1	a b d
2	c a e g
3	b d f
4	b d e
5	a b d
6	b d

W1
W2
Time Line

Figure 1. A transactional data stream and sliding window model

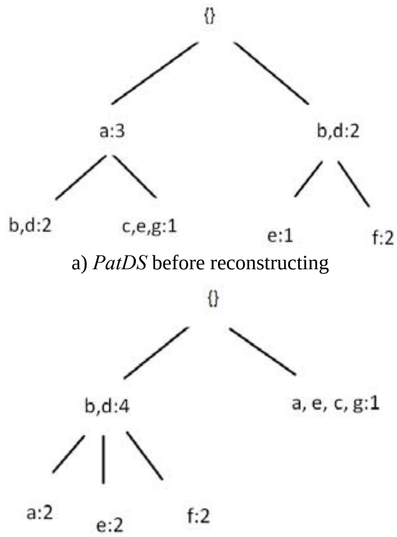
Example 1. Figure 1 shows a data stream of transactions, where the left column shows the transaction id, i.e., *Tid* of incoming transactions and the right column shows the items within each transaction. Two consecutive transactional sliding windows W_1 and W_2 are defined over the so far received transactions. The pane size is 2 transactions and a window contains 2 panes. Given a minimum support threshold, the aim is to mine all frequent itemsets that exist within the recent window after the user makes a mining request.

4 The proposed method

As mentioned previously, with respect to *DSTree* the *CPS-Tree* is more memory efficient. This efficiency is due to its prefix tree which is maintained in the support descending order. Additionally, the *CPS-Tree* has better runtime due to its effective pane extraction mechanism and applying the *FP-Growth* to a support descending ordered prefix tree. However, this method suffers from a number of shortcomings which affect its runtime and memory usage during data stream processing and the mining phase. Since, in a data stream of transaction we have to store all items (frequent or infrequent), the number of nodes in the *CPS-Tree* becomes prohibitively large. Moreover, in some paths of *CPS-Tree*, there exist a number of items with the same support value. Therefore, long paths may be formed which consume significant amount of memory. Additionally, in the *CPS-Tree* the *FP-Growth* algorithm is used to find frequent patterns after a user submits a request. We believe that this type of mining is not suitable for data streams since it generates large number of conditional trees during its processing. These conditional trees require more memory in addition to the original tree. Moreover, constructing these trees is a time consuming task. To overcome these shortcomings we have utilized Patricia tree [7] for sliding window frequent patterns mining over a data stream. Based on this data structure, we have developed a new data structure namely *PatDS* for storing and updating sliding window transactions. Additionally, an efficient in place mining method has been exploited to extract frequent patterns from the *PatDS*

Patricia tree can be considered as a special type of a prefix tree. In a Patricia tree items in all branches are sorted according to a predefined order, e.g., lexicographical or support descending. In this structure, each sequence of items existing consecutively in the same path with the same count, resides in a single node. Therefore, any node in Patricia tree can store a list of items if they have the same support value and appear in the same branch of tree consecutively. However, the Patricia tree allocates a node for an item if its adjacent nodes have different support values or if there is not any adjacent node. The *same support items* in a path of a prefix tree, is an ordered sequence of items having the same support value in that path.

The *same support items* in a path of prefix tree could reside in a single node with single support value. The *same support items* in a path could reside in a single node without any lose of information. It is important to note that, *same support items* must appear in the same path. A set of items could have the same support value but do not co occur with each other in the same set of transactions and thus does not form any the *same support items* in a specific path. A prefix tree which benefits from the *same support items* to compress its structure is called a *Patricia* tree. By using smaller number of nodes in a *Patricia* tree, the tree structure exhibits lower space requirements, especially in the sparse set of transactions [7]. In this study, we have adapted *Patricia* tree for space efficient storing sliding window information and proposed a new data structure namely *PatDS*. In our method, the structure of *PatDS* is dynamically reconstructed using a previously proposed restructuring algorithm. Similar to *CPS-Tree*, *PatDS* is used to store sliding window transactions. However, *PatDS* ensures smaller memory requirement and thus better mining time. In data stream mining we have to maintain both frequent and infrequent items since an infrequent item may become frequent in the near future, thus requiring its support value. Therefore a number of items become large in the tree data structure. The larger the number of items, the further probability of the *same support items* in every path of the tree. Therefore, *PatDS* memory footprint can be far less than *CPS-Tree*. We need to maintain a *PatDS* in support descending order and manage the *same support items* in all branches dynamically as the content of the window is changed. A restructuring method is applied on the *PatDS* to obtain support descending ordered tree. In [5], a restructuring method namely *BSM* (Branch Sorting Method) is proposed. This method uses the merge sort to sort every path of the prefix tree. In the BMS, unsorted paths are removed and then are sorted and reinserted to the tree structure. We use *BMS* to restructure the *PatDS* so that it has support descending order of items in every path. After tree restructuring, a compression process is started to identify the *same support items* in each path and merge them to a single node.



a) *PatDS* before reconstructing

b) *PatDS* after reconstructing and compression

Figure 2. *PatDS* before and after restructuring

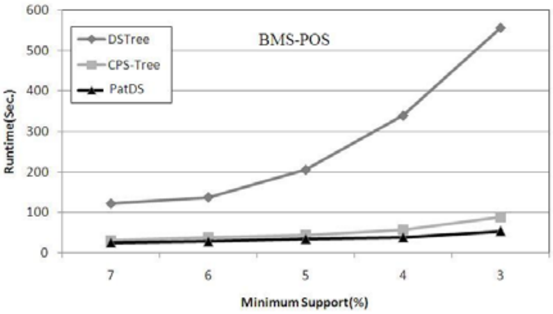
Example 2: Again consider the data stream shown in Figure 1. Suppose that the window size and pane size are set to 3 and 2 respectively. Therefore the sliding window contains all transactions. Suppose that the processing is started with lexicographical order and tree restructuring is performed after the first window. Figure 2.a shows *PatDS* structure where a lexicographical order of items is used and Figure 2.b shows its corresponding tree after restructuring and compression.

Therefore, using *PatDS* instead of CSP-Tree needs small number of nodes before and after restructuring and compression due to residing the *same support items* in a single node. However, *PatDS* restructuring and compressing are time consuming process and we must perform them when required. We call the process of tree restructuring and tree compressing when significant change is observed in the order of items and distinct support values in the header table. Therefore, we test two criterions after every sliding. Change in the order of items reflects the required change in the order of items in the paths of *PatDS*. Moreover, the number of distinct values appeared consecutively in the header table with respect to the number of items is a good approximation for the number of the *same support items* in different paths. As the number of distinct values appeared in the header table reduced, the number of the *same support items* in different branches is increased. We use an adaptation of FP-Growth algorithm for the mining phase. *FP-Growth* is a bottom up approach since it extracts conditional pattern bases by starting from bottom of the header table. This method constructs large number of conditional *FP-Tree*s during the mining process in addition to the original *FP-Tree*. These conditional *FP-Tree*s have further memory requirement beside original *FP-Tree*. On the other hand, *Top-Down FP-Growth* [10] for static databases, process nodes at upper level first to extract conditional pattern bases. This method reuses the paths in the original *FP-Tree* to form conditional FP-Trees. Therefore, since it is an in place approach, it does not waste memory for conditional trees and requires smaller memory during the mining. Only additional memory is header tables of conditional trees, i.e., conditional header tables. Since in data stream processing we have the limitation of main memory, here, we utilize this method for mining frequent patterns from *PatDS*.

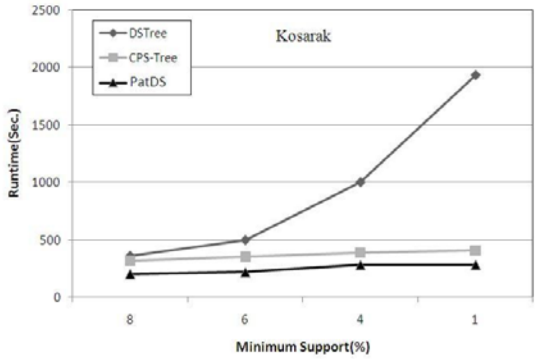
5 Experimental evaluations

We empirically evaluate the performance of *PatDS* method. To fair comparison, two similar algorithms are selected which operate in the same model of *PatDS*. We have implemented the *PatDS* and two recently proposed algorithms of *DSTree* and *CPS-Tree*. All programs were written in C++ and executed in Windows XP on a 2.66 GHz CPU with 1 GB memory. To show the applicability of our proposed method two real life datasets of BMS-POS and Kosarak are used in the experiments. The first experiment compares the run time of *PatDS*, *CPS-Tree* and *DSTree*. In this experiment, the average runtime of all active windows are computed for the three algorithms on all datasets. After every window, the mining is performed on the current window. For the BMS-POS, every window contains three panes for which the pane size is 50K transactions. In the Kosarak dataset, the pane size and window size are set to 50K transactions and 4 panes respectively. Figure 3 shows the result of this experiment for different minimum support thresholds. In this Figure, for both charts, horizontal axis show minimum support values and vertical axis shows the run time in second. As shown in

Figure 3, the *PatDS* is more efficient than *DSTree* and *CPS-Tree* algorithms for both of the datasets. As the minimum support decreased, the efficiency of the *PatDS* is more revealed and the performance gap becomes more significant. Figure 3 shows that the *PatDS* is faster than other algorithms even for low support thresholds where the number of generated frequent itemsets is high. The efficiency of the *PatDS* is due to its efficient storing sliding window information using smaller number of nodes and in place mining of frequent patterns which does not needs any additional tree creation time.



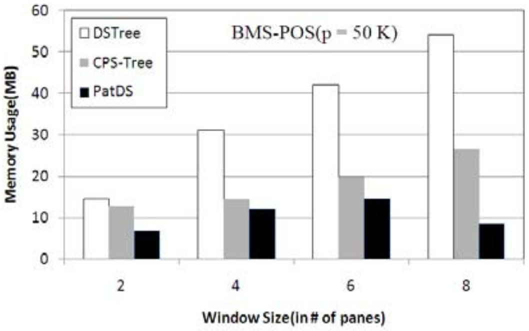
a)BMS-POS



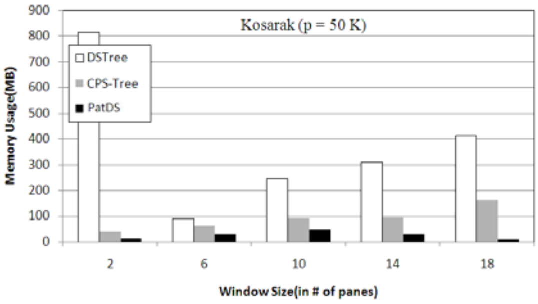
b)Kosarak

Figure 3. Runtime comparison of *DSTree*, CPS-Tree and *PatDS*

The second experiment is about the memory usage. In this experiment, the memory usages of all algorithms are measured with respect to different window sizes on the datasets. Again, for both datasets the pane size is set to 50K. Sizes of window for each dataset are varied by using different number of panes. The results are shown in Figure 4. In this figure, memory usage (in mega byte) is plotted with respect to each window size for all the algorithms. As shown in Figure 4, memory usage of the *PatDS* is significantly lower than other methods. This is true for all datasets and all window sizes. As can be inferred from Figure 4, the amounts of memory enhancement with respect to other algorithms in both datasets are almost fixed for different window sizes. This is due to the more compact data structure of *PatDS*. The *PatDS*, has far smaller number of nodes and thus has better memory usage. Therefore, the *PatDS* outperforms *CPS-Tree* and *DSTree* in terms of memory usage.



a) BMS-POS



b) Kosarak

Figure 4. Memory usage comparison of *DSTree*, *CPS-Tree* and *PatDS*

6 Conclusions

In this paper, a new method namely *PatDS* is proposed for frequent patterns mining in data streams. This method uses a new data structure which is an adaptation of Patricia tree for data stream mining. The *PatDS* mines the complete set of frequent patterns in recent sliding window faster than recently proposed algorithms. Moreover due to use of a compact data structure and efficient maintenance of this structure, the proposed algorithm requires less memory with respect to other algorithms. Frequent patterns are extracted using an in place mining algorithm which does not need additional tree structures during the mining.

7 References

[1] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules”, Proc. VLDB Int. Conf. Very Large Databases, 1994, pp. 487–499.

[2] J. H. Chang and W.S. Lee, “estWin: Online data stream mining of recent frequent itemsets by sliding window method”, Journal of Information Science, vol. 31(2), 2005, pp. 76–90.

[3] C. K. S. Leung and Q. I. Khan, “DSTree: a tree structure for the mining of frequent sets from data streams”, Proc. ICDM, 2006, pp. 928–932.

[4] C. H. Lin, D. Y. Chiu, Y. H. Wu, and A. L. P. Chen, “Mining frequent itemsets from data streams with a time-sensitive sliding window”, Proc. SIAM Int. Conf. Data Mining, 2005.

[5] S. K. Tanbeer, C. F. Ahmed, B. S. Jeong, and Y. K. Lee, “Sliding window-based frequent pattern mining over data streams”, Information Sciences, vol. 179(22), 2009, pp. 3843-3865.

[6] J. Han, J. Pei, Y. Yin, and R. Mao, “Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach”, Data Mining and Knowledge Discovery, vol. 8(1), 2004, pp. 53-87.

[7] A. Pietracaprina and D. Zandolin, “Mining Frequent Itemsets Using Patricia Tries,” Proc. IEEE ICD Workshop Frequent Itemset Mining Implementations, CEUR Workshop Proc., vol. 80, Nov. 2003.

[8] D. Knuth, “Sorting and Searching”. Reading, Mass.: Addison Wesley, 1973.

[9] G. S. Manku and R. Motwani, “Approximate frequency counts over data streams”, Proc. VLDB Int. Conf. Very Large Databases, 2002, pp. 346–357.

[10] K. Wang, L. Tang, J. Han, and J. Liu, “Top Down FP-Growth for Association Rule Mining”, Proc. 6th Pacific-Asia Conf. on Advances in Knowledge Discovery and Data Mining, Lecture Notes In Computer Science; Vol. 2336, 2002, pp. 334 – 340.

A Framework for Web Host Quality Detection

A. Aycan Atak, and Şule Gündüz Ögüdücü

Abstract— **With the rapid growth of World Wide Web, finding useful and desired information in a short amount of time becomes an important issue for Web users. Search engines and focused crawlers help people to navigate the internet. A user expresses her information need in the form of a query and there is huge number of Web pages returning to this query. However, the majority of users view only a single page (the top 10 Web pages as ranked by the search engine) returned by a search engine. Even if the returned Web pages do not provide the exact information they need, the users also do not refine their query based on the returning results of their initial query. Thus, not only finding relevant Web pages but also ranking them plays an important role for the search engines. For this reason, determining the quality of Web pages is one of the main priorities of search engines, since low quality Web pages cause search engines results to be extremely vague and flooded with irrelevant Web pages. In this paper, we propose a novel method for determining the quality of Web pages. The proposed method first identifies the genre of Web pages and then it determines the quality of Web pages based on their genre. Our experimental results show that our proposed method is very effective and efficient.**

I. Introduction

The number of Web pages grows rapidly every day. The advent of Web has caused a dramatic increase of the use of Internet as a huge, widely distributed, global information service for every kind of information. Since there is no central system to control the Web, it is impossible to estimate the precise number of Web sites and Web pages on Internet. Monthly surveys by sites like Netcraft[1] have shown that in September 2010 there are nearly 227,225,642 sites on the Web. In this environment, search engines help people to locate information relevant to their search needs expressed in the form of a query. However, the lack of central control has also increased the number of Web pages consisting of highly noisy, contradictory and unreliable information. Due to this fact, even the search results in the initial results pages are being heavily spammed. Since Web searchers usually examine the top ten results, it becomes an important issue for search engines, to list really relevant and high quality Web pages at top of search results.

Web spam can significantly decrease the quality of search engine results. Search engines work on efficient algorithms to determine and block spam Web pages. Without using such algorithms the search engine results may be unreliable and Web searchers lose their trust and confidence in search engine. Common approaches for spam detection are based on extracting a set of content-based and link-based features from Web pages. From the machine learning point of view, Web page spam detection is considered as a binary classification problem of Web site content as spam or non-spam. In this problem, the Web pages are represented with feature vectors with dimensions corresponding to the terms appeared on them. However, this technique is vulnerable to Web sites faking high relevance with respect to some topics. This is called Search Engine Persuasion (SEP) in [9]. Link analysis is one of the solutions to overcome this problem. Thus, using PageRank scores for eliminating spam Web pages from search results has become a standard for years.

But, applying methods for spam filtering may still not guarantee that search engines list relevant and high quality Web pages first. The nature of result ranking task based on the relevancy and quality of results is different to that on more traditional spam detection. One of the differences is that the measurement of relevance and quality of search results is "subjective" since it is highly dependent on peoples' perceptions of the relevance and quality of information. Low quality is not simply equivalent to Web spam. To promote the research and practice of new strategies for determining the overall rank, quality and importance of a Web site and estimating Web content quality, the ECML/PKDD Discovery Challenge is organized in 2010.

Different from the traditional Web spam detection problem, the aim in this challenge is to develop site-level classification for the genre of the Web sites (editorial, news, commercial, educational, "deep Web", or Web spam and more) as well as their readability, authoritativeness, trustworthiness and neutrality. The motivation behind this labeling procedure is to help organizations, such as search engines and Web archives, in their efforts to prioritize their procedures to gather, store and organize their collection of Web pages.

Although link-based features are commonly used for determining relevant and trusted Web hosts today, term vectors obtained from Web content are still considerably important components of this kind of quality determination problems. Using only link-based features such as PageRank scores, it is difficult to determine the factuality or genre of a Web site that effect its quality score. However, when considering the size of the Web, it is not feasible to extract the content of Web hosts in order to determine a quality score for them. Besides, for organizations such as search engines or Web archives, it is important to determine the quality score of a Web host without downloading the content of the Web page.

In this paper, we focus on determining the quality score of Web pages. For this task, we used the data set provided by the

A. Aycan Atak is with the Department of Computer Engineering, Istanbul Technical University, Istanbul, 34469, Turkey (phone: +90-212-285-3682; fax: +90-212-285-3689; email: ataka@itu.edu.tr).

Şule Gündüz Ögüdücü is with the Department of Computer Engineering, Istanbul Technical University, Istanbul, 34469, Turkey (phone: +90-212-285-3597; fax: +90-212-285-3597;email: sgunduz@itu.edu.tr).

[1] http://news.netcraft.com/archives/category/web-server-survey/, Accessed on 27 Sept. 2010.

Discovery Challenge 2010. The Discovery Challenge tasks included the prediction of the quality score predefined based on genre, trust, factuality and bias and spamicity. It is found that, content based features are useful when predicting the genre of Web pages. However, determining more subjective characteristics of Web pages, such as trustiness, neutrality and bias, predicted genre labels of Web pages yields better results.

The rest of this paper is organized as follows; in section 2, detailed description and analysis of available dataset is given. In section 3, classifiers and feature selection methods used in this study are mentioned. Experimental results are shared and discussed in section 4. And finally, conclusion and future work plans are given in section 5.

II. Related Work

Web spam detection has been pointed as a serious problem for search engines and Web archives. However, the studies on this problem have been slow down since the problem of determination spam in social networks have become more attractive for researchers. Another important research area is to determine the utility of Web page in relation to an information need represented as a query. It has been shown that a range of factors affect human judgments of relevance. However, these studies have been conducted on textual documents which structures are different from Web documents with a wide range of formally and informally produced multimedia content and hyperlinks. However, studies on users' perceptions of the relevance of information need on the Web are few. Recently, several studies are performed within ECML/PKDD 2010 Discovery Challenge (DC2010)[2].

Geng et. al. used multi-scale attributes composed of attribute groups including content features, page and host level link features and TFIDF features [1]. With the fusion of different sets of features, bagging is applied to C4.5 decision tree to classify Web sites according to the categories given in the DC2010 data sets. In that study, it has been found that the host level link features are robust for classifying tasks and that feature fusion is necessary for statistical Web content quality assessment.

Sokolov et. al. used RankBoost algorithm in their instance based model and propagation schemes in their graph based model separately [2]. They showed that iterative algorithm for learning propagation scheme is comparatively more efficient on revealing correlations between different quality levels.

Nikulin reduced dimensionality of host attributes using Wilcoxon-based feature selection [3]. He also reduced multi-class decision problem to corresponding number of two-class decision problem using one class against all method. In that study, Nikulin took the final decision using minimal and maximal values from the result set of two-class decision problems. For predicting host quality, Lex et. al. used voting with three classifiers; J48 decision tree, class-feature-centroid classifier (CFC) and support vector machine (SVM)[4]. In that study, each classifier applied on different types of attributes and oversampling method is used to deal with imbalanced dataset problem.

III. Data Set

The data set, DC2010, used in this study is provided by European Archive Foundation as the material of the ECML/PKDD Discovery Challenge 2010 on Web Quality [5]. It is created through crawling the Web sites in the .EU domain in three languages: English, German and French. Thus, the data set is separated into three parts where each part contains Web hosts from a different language with the same group of attributes. In this data set, four sets of attributes are provided: link-based attribute set, content-based attribute set, natural language processing (NLP) attribute set and term frequency vectors of hosts. Also URL and host graph of crawled piece of the Web are provided with the data set. The details of these sets of attributes are described in detail in the next paragraphs[3]:

- The provided link-based attributes are obtained from the Web graph. Among the 178 link-based attributes, the most salient attributes of this group seem to be: PageRank, out-degree, in-degree and TrustRank values. These attributes give information about the graph properties of hosts.
- Content-based attributes are obtained using full content of hosts. For a host, the number of words in the homepage or compression rate of the homepage can be given as examples of this set of attributes. The total number of attributes in this set is 98.
- NLP attributes are computed per URL using text content of Web hosts. This attribute set includes features such as the number of tokens in a URL or counts of token types such as adverb or pronoun in a URL.
- Term frequency vector of each host is computed by counting the number of times words appear within each host. The term frequency vector consists of the most frequent 50,000 terms after eliminating the stop words.

One of the tasks in this challenge is the quality determination of web sites whereas the quality of a Web site is measured as an aggregate function of its genre, neutrality, bias and trustiness. Therefore, predicting class attributes of Web sites is the main problem which contains the answers of quality prediction.

The first class attribute to be predicted is genre. There are 6 possible values of this attribute in the provided data set. Thus a Web site is labeled with one of the following categories; spam, news-editorial, commercial, educational, discussion or personal-leisure. As can be seen from Table If genres of Web hosts are considered, as seen from Table I, the data set is highly imbalanced with more Web site labeled as commercial or educational. There is very few information available for some genre labels in the training set which will affect the learning results negatively. For example, for French data set, there is only one host labeled as discussion.

[2] http://www.ecmlpkdd2010.org/index.php?md=articles&id=2041&lg=eng

[3] http://datamining.sztaki.hu/?q=en/DiscoveryChallenge/

TABLE I
GENRE DISTRIBUTION FOR EACH DATASET

Dataset	Spam	News-Editorial	Commercial	Educational	Discussion	Personal-Leisure
English	2.7%	3.0%	36.2%	34.0%	4.3%	19.9%
German	3.7%	11.2%	44.8%	12.0%	10.4%	18.0%
French	4.8%	3.7%	33.3%	13.2%	0.5%	44.4%

TABLE II
TRUSTINESS (A), NEUTRALITY (B) AND BIAS (C) DISTRIBUTION FOR EACH PART OF DATASET

Dataset	1	2	3
English	0.3%	1.1%	98.5%
German	5.7%	70.0%	24.1%
French	0.0%	2.0%	98.0%

(A)

Dataset	1	2	3
English	0.3%	2.3%	97.4%
German	3.4%	61.0%	35.6%
French	1.0%	5.2%	93.8%

(B)

Dataset	1	2
English	1.2%	98.8%
German	4.6%	95.4%
French	0.0%	100.0%

(C)

The second class attribute, trustiness, can take three values; 1, 2 and 3. A value of 1 means that the host is not reliable. Values of 2 and 3 mean that the host is reliable. The third class attribute, neutrality indicates factuality of a host and it also has three values; 1, 2 and 3. A value of 1 means that the host is problematic. The fourth class attribute, bias, can take two values; 1 and 2. A value of 1 indicates that there are significant bias problems in the host. As can be seen in Table II, the training data set is imbalanced. For example, the training set of the French data contains any example with a trustiness or bias value of 1.

IV. METHODOLOGY

In this study, we only utilize the term vector attribute set. Remaining 3 sets of attributes, namely link-based, content-based and NLP attributes, are determined irrelevant in predicting class attributes based on the results of feature selection methods. It is observed that these 3 sets of attributes do not contain sufficient information to determine the class labels. At this point it must be noted that, there are some instances in the data set that do not have term frequency vector. These instances are removed from both training and validation sets.

It is observed that genre values of hosts are more informative than the term frequency vectors in determining the hosts' trustiness, neutrality and bias values. For this reason, in this study first the genre of Web hosts' is determined based on the term frequency vectors which is used in turn to identify the trustiness, neutrality, bias values of hosts'. The obtained values are used to compute the quality score of the hosts'. An overview of our methodology is given in Figure 1.

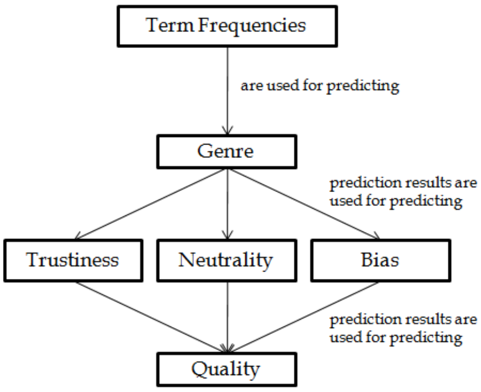

Term Frequencies
are used for predicting
Genre
prediction results are used for predicting
Trustiness
Neutrality
Bias
prediction results are used for predicting
Quality

Fig. 1. Proposed methodology

For machine learning algorithms including attribute selection, classification and oversampling methods, implementations in WEKA[4] data mining tool are used. However, it is observed that oversampling methods are not successful to handle the imbalanced data set problem and they reduce the performance of classification algorithms. For this reason, oversampling methods and the classification results with oversampling methods are not reported in this study.

A. Genre Prediction

If genre prediction is considered, we are facing several problems; multi-class decision problem, text-categorization problem and high dimensional vector space.

Yang and Pedersen compared several attribute selection methods for text categorization including information gain, chi-square statistics, document frequency, mutual information and term strength [6]. According to that study, information gain and chi-square statistics are among the best performing attribute selection methods for text categorization. In this study, for all datasets, chi-square statistics also appears to be performing better. It decreases classification execution times while increasing the accuracy of the classifier. For each data set, number of selected terms is determined with hill-climbing method. Attribute selection method and number of selected terms which minimize the training error are selected for the final classification.

For classification, different classifiers are applied on the same data set. Support vector machines, (SVM) which is

[4] http://www.cs.waikato.ac.nz/ml/weka/

considered as the best performing classifier for many text categorization problems [7], appears to be performing better in this study. Similarly as in attribute selection phase, the classifier and its parameters are also selected based on the results of hill-climbing method.

From original term vectors, new term vectors are generated with binary weighting. According to binary weighting, for host h and term t, if h contains t then the weight of t in term vector of h is set to 1, otherwise it is set to 0. Each experiment to find the best classifier and attribute selection method is applied to both data sets to illustrate the effect of different weighting schemes on classification results. Subsequently, term vectors with binary weighting have better results in terms of classification accuracy. Table III presents the applied classifiers and feature selection methods to each data set and the resulting number of terms after feature selection.

TABLE III
FINAL DECISIONS FOR ATTRIBUTE SELECTION METHOD AND CLASSIFIER

Dataset	Feature Selection Method	# Of Selected Terms	Weighting	Classifier
English	Chi-square Stat.	11000	Binary	SVM
German	Chi-square Stat.	100	Binary	SVM
French	Chi-square Stat.	8500	Binary	SVM

B. Trustiness, Neutrality and Bias Prediction

Trustiness, neutrality and bias are predicted based on the results of genre prediction. For host h, genre prediction process produces 6 probability values. Each of these probability values indicates the probability of belonging to a genre. A host h_i can be then represented with 6 attributes, such as $g_{i1}, g_{i2}, ..., g_{i6}$, where the values of each attribute is obtained from genre prediction and $\sum_{j=1}^{6} g_{ij} = 1$. Linear regression is applied to this data set for predicting trustiness, neutrality and bias values of each host. Thus, linear regression produces a score for each hosts' trustiness, neutrality and bias values. Then, the hosts in the test set are sorted by each of these values individually.

C. Quality Prediction

For determining quality levels of hosts, the prediction of the utility score predefined based on genre, trustiness, neutrality and bias values are also inclusded in the DC2010 tasks. For each host h, utility value is calculated as in Algorithm 1 by combining the genre, trustiness, neutrality and bias values of h.

Algorithm 1 Quality Determination

```
$value = 0$;
**if** News-Edit OR Educational **then**
  $value = 5$;
**else if** Discussion **then**
  $value = 4$;
**else if** Commercial OR Personal-Leisure **then**
  $value = 3$;
**end if**
**if** $neutrality = 3$ **then**
  $value+ = 2$;
**end if**
**if** $bias = 1$ **then**
  $value- = 2$;
**end if**
**if** $trustiness = 3$ **then**
  $value+ = 2$;
**end if**
```

Based on Algorithm 1, the utility value of a host can range between -2 and 9. The categories News and Educational have the highest quality. Also, trusted, unbiased and neutral contents have a high quality score. By default, Web Spam hosts have the lowest quality. The utility value may be utilized to predict the quality of a host. Similarly, linear regression method can be applied to obtain quality scores.

V. EXPERIMENTAL RESULTS

The results of the experiments are conducted in terms of the evaluation metrics used at the DC2010. Depending on tasks determined by this challenge, quality prediction is evaluated for English, German and French datasets while genre, trustiness, neutrality and bias prediction are evaluated for only English dataset.

A. Evaluation

For evaluation, normalized discounted cumulative gain (nDCG) is used at DC2010. nDCG is obtained with normalization of discounted cumulative gain (DCG) with ideal DCG value. DCG and nDCG can be computed with Eq. 1 and Eq. 2 respectively.

$$DCG = \sum_{rank=1}^{N} utility\,(rank) \times \left(1 - \frac{rank}{N}\right) \quad (1)$$

$$nDCG = \frac{DCG}{IdealDCG} \quad (2)$$

In Eq. 1, N is the number of test instances in the test set and ideal DCG is the DCG value obtained by the ideal ordering of the hosts. According to $nDCG$ formula, it is important to rank hosts with higher utility values at top. Please note that, in a perfect ranking algorithm, the nDCG values will be the same as the Ideal DCG producing an nDCG of 1.0. Thus, all nDCG calculations are then relative values on the interval 0.0 to 1.0. For genre prediction problem, when evaluating results for genre g, utility value is 1 for host h, if h belongs to genre g. Otherwise utility value is 0.

TABLE IV
GENRE PREDICTION RESULTS FOR ENGLISH DATASET

Genre	nDCG
Spam	0.88
News-Editorial	0.73
Commercial	0.84
Educational	0.87
Discussion	0.76
Personal-Leisure	0.81
Overall	0.82

B. Genre Prediction

Genre prediction results are given in Table IV. For each genre, nDCG value is computed separately. Overall result of genre prediction can be computed with arithmetic average of all nDCG values obtained from genre prediction. This experiment is applied on English dataset only.

As can be seen from Table IV, the nDCG results are high compared to the results from previous publications. These results indicate that our proposed method yields higher nDCG values for spam genre prediction. The nDCG values for spam and educational genres are higher. This may be due to the fact that spam and educational hosts have more discriminative words.

According to Table IV, it can be said that spam and educational genres are more word-specific than other genres. There are more discriminative words that mostly exist in spam and educational hosts. However, discussion and news-editorial hosts contain more words in common.

C. Trustiness, Neutrality and Bias Prediction

Results of the trustiness, neutrality and bias predictions are given in Table V. These class attributes are predicted with term vectors and genre prediction results separately. As in genre prediction, this experiment is also applied on English dataset only.

TABLE V
TRUSTINESS, NEUTRALITY AND BIAS PREDICTION RESULTS FOR ENGLISH DATASET

Class Attribute	nDCG Using Term Vectors	nDCG Using Genre Predictions
Trustiness	0.35	0.62
Neutrality	0.44	0.46
Bias	0.40	0.53

As can be seen from Table V, except for neutrality, using genre prediction results instead of term vectors causes a significant improvement in the nDCG values. Also it can be concluded that for Web hosts, neutrality is less depended on genre values than trustiness and bias. However, also for neutrality, using genre prediction results still gives better results than using term vectors.

If other studies are considered, these results are quite satisfactory. For trustiness, the best nDCG value is obtained so far. Also bias prediction results are at the same level with top results.

TABLE VI
QUALITY PREDICTION RESULTS

Dataset	nDCG
English	0.85
German	0.81
French	0.82

D. Quality Prediction

Quality prediction results are given in Table VI. For each dataset, quality is computed separately using trustiness, neutrality and bias prediction results.

These results are also compatible with the results obtained in other studies focused on the same dataset. Using prediction results from other classification processes reduces the size of the attribute set which in turn reduces the execution time of the classifier. That's why, this results are more satisfactory when execution time and accuracy are considered together.

VI. CONCLUSION AND FUTURE WORK

In this study, a method to measure the quality of Web hosts is proposed. It is first hypothesized and then observed that terms of hosts contain more information about Web host genre than link-based metrics. For predicting class attributes related with Web hosts, a framework is presented. In order to evaluate the proposed framework, a data set from a conference challenge is used. The experimental results show that our proposed method is superior to the previous proposed methods in terms nDCG and execution time.

We are extending the classification method in several ways. More accurate predictions can be made using hybrid datasets including category prediction results and term vectors. However, after quality prediction with two-class decision problem [8] and this multi-class decision problem studies, detecting quality using propagation on graph structures is the main subject for our future work. Due to existence of Web host graph structure in the DC2010 dataset, possible future studies can focus on the same dataset with this study.

VII. ACKNOWLEDGEMENT

Author Gunduz-Oguducu was partially supported by the Scientific and Technological Research Council of Turkey (TUBITAK) EEEAG project 110E027.

REFERENCES

[1] G. Geng, X. Jin, X. Zhang and D. Zhang, "Evaluating Web content quality via multi-scale features", *ECML/PKDD 2010 Discovery Challenge Workshop*, 2010.

[2] A. Sokolov, T. Urvoy, L. Denoyer and O. Richard, "Madspam consortium at the ECML/PKDD Discovery Challenge 2010", *ECML/PKDD 2010 Discovery Challenge Workshop*, 2010.

[3] V. Nikulin, "Web-mining with Wilcoxon-based feauture selection, ensembling and multiple binary classifiers", *ECML/PKDD 2010 Discovery Challenge Workshop*, 2010.

[4] E. Lex, I. Khan, H. Bischof, M. Granitzer, "Assessing the quality of Web content", *ECML/PKDD 2010 Discovery Challenge Workshop*, 2010.

[5] A. A. Benczur, C. Castillo, M. Erdelyi, Z. Gyongi, J. Masanes and M. Matthews, ”ECML/PKDD 2010 Discovery Challenge Data Set”, *Crawled by the European Archive Foundation*.

[6] Y. Yang and J. O. Pedersen, ”A comparative study on feature selection in text categorization”, *In proceedings of the fourteenth international conference on machine learnin (ICML ’97)*, Douglas H.Fisher (Ed.), Morgan Kaufmann Publishers Inc., San Fransisco, CA, USA, pp.412-420, 1997.

[7] J. Thorsten, ”Text categorization with support vector machines: Learning with many relevant features”, *The 10th European Conference on Machine Learning*, 1998.

[8] A. A. Atak, S. G. Oguducu, ”A framework for social spam detection based on relational bayesian classifier”, *In proceedings of the 6th International Conference on Data Mining (DMIN10)*, pp.71-77, 2010.

[9] M. Marchiori, ”The quest for correct information on the Web: hyper search engines”, In Selected papers from *the sixth international conference on World Wide Web*, Phillip H. Enslow, Jr., Mike Genesereth, and Anna Patterson (Eds.). Elsevier Science Publishers Ltd., Essex, UK, pp. 1225-1235, 1997.

Effective Parameters on Response Time of Data Stream Management Systems

Shirin Mohammadi[1], Ali A. Safaei[1], Mostafa S. Hagjhoo[1] and Fatemeh Abdi[2]

[1]**Department of Computer Engineering, Iran University of Science and Technology**
Tehran, Iran
sh_mohammadi@comp.iust.ac.ir, safaeei@iust.ac.ir, haghjoom@iust.ac.ir

[2]**Department of Science, Babol-Branch, Islamic Azad University, Babol, Iran**
sulmaz_abdi@yahoo.com

***Abstract* – considering rapid and time variant (bursty) nature of data streams, data would rapidly lost its value while time is going on. So, the results with high response time are not reliable in Data Stream Management Systems (DSMSs). In other words, one of the most important factors in data stream management systems is the response time (i.e., the amount of time which a data stream tuple arrives into the system until it exits as the output while processed by a query).**

In this paper, the parameters which are more effective on DSMSs' response time are considered, categorized and analyzed. Static and dynamic system properties, input and output data streams' properties, and also properties of queries and query processing algorithms are factors which influence on DSMS's response time. Experimental results are shown to illustrate the impact of each parameter on the response time metric.

Keywords: DSMS, Query processing, Response time, Effective parameters

IKE'11 - 10th Int'l Conference on Information and Knowledge Engineering

I. INTRODUCTION

Traditional Database management systems as finite set of stored data are able to respond ad-hoc queries in best cases. But most of new applications need data stream processes which are infinite continuous streams of data [1, 2]. These applications need a new series of systems called as data stream management systems (DSMS). DSMSs provide requirements of mentions applications. Continuous data streams are infinite and rapid and varying time. DSMSs are able to discuss queries on data streams. These queries are executed in long time processes since being received continuously and are called as continuous queries [2]. Such queries with long lasting execution time need to be evaluated by the system until they finish [3]. One of the important factors of evaluating DSMSs is the response time of the system.

Response time or tuple latency is defined as the average time which a tuple needs to be processed by a query. Of course, it includes all waiting times in buffers [3]. In other words, response time for an output tuple is the time period since providing all required information for concluding the output tuple until generation of the output tuple in real [4].

Generally, figure 1 represents DSMS architecture. Incoming data streams on the left produce data indefinitely and drive query processing. In many applications stream data also may be copied to an archive, for preservation and possible offline processing of expensive analysis or mining queries which we primarily concerned with the online processing of continuous queries. Finally, processing of such queries requires intermediate state, which denoted as Scratch Store in the figure and could be stored and accessed on disk or in memory. Applications or users register their Continuous Queries (CQ), which they remain active until explicit deregistration. Results are generally transmitted as output streams of data, although they could be relational results being updated over time [1].

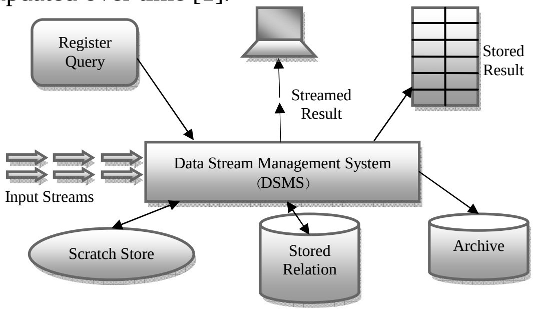


Figure 1- DSMS Architecture [1]

The remained parts of this paper are structured as follows: related work is studied in section 2. Effective parameters on DSMS's response time are categorized and analyzed in section 3. In section 4, experimental results are shown to illustrate the impact of these parameters on the response time. Finally, we conclude the paper in section 5.

II. RELATED WORKS

Lots of researches on DSMSs are done [13]. Several primary samples of DSMSs like the STREAM [1,2], Aurora [5] and TelegraphCQ [14] are provided too. scheduling strategies of operators to process continuous queries on data streams varies from simple ones like the Round Robin [6], chain [15] and Greedy [6] to more complex ones [16,17]. Some of them are provided for optimizing a performance factor [18, 19], while some others try to optimize multiple factors or a compound one [20, 21]. Totally most of these methods are provided to minimize tuple latency or the response time factor. Determination and analysis of effective parameters on response time of DSMSs are explicitly studied in few references which [22] is one of them.

III. EFFECTIVE PARAMETERS ON RESPONSE TIME

A. *Categorizing the parameters based on DSMS architecture*

Response time of a query in a Data stream management system depends on several parameters, some with less influence and some with more importance. Considering the figure 2 effective parameters on response time of DSMSs are represented in total categories of: *Data stream properties*, *Query properties*, *Query execution properties*, *Output properties*, *System properties (static conditions)* and *System condition (Dynamic).*

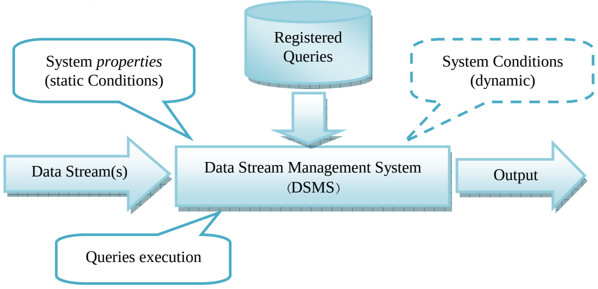

System properties (static Conditions)
Registered Queries
System Conditions (dynamic)
Data Stream(s)
Data Stream Management System (DSMS)
Output
Queries execution

Figure 2 - categorizing the effective parameters on response time

B. *Data stream* properties

A data stream includes data elements generated in an infinite, continuous and rapid manner which varies in time. In other words Data stream of *S* is a set of *s* elements with time stamp of τ which the elements arrive to the system in time stamp order. The time stamp specifies the logical entrance time of a tuple into the data stream. By using discrete and regular domain of T a data stream can be defined as below [1]:

$$S = \{\langle s, \tau \rangle \mid \tau \in T\} \quad (1)$$

1) Type of elements in a data stream: Data stream *S* includes data elements *s* which are divided into three categories of well-structured data, semi-structured data and unstructured data [3].

2) Domain of attributes: Domain or element type of tuples, belongs to attribute set of *Att* which its members are *m* types of a_1, a_2, a_3, …, a_m as:

$$s = (e_1, e_2, ..., e_n) \quad , \quad \forall e_i \in Att \quad , \quad Att = \{a_1, a_2, ..., a_m\} \quad (2)$$

3) Number of attributes: each tuple *s* is represented by an ordered list of elements like (e_1,e_2, … , e_n) which the *n* represents the number of attributes of a tuple.

4) Data stream distribution: The DSMSs usually don't have any control on order, rate and distribution of input streams [2]. Data stream distribution parameter represents the manner of distribution in stream arrival into the DSMS. Distribution of data arrival can be Uniform or bursty distribution such as the Pareto distribution.

5) Arrival rate into the system: A data stream includes an ordered set of tuples which arrives to the system continuously. Time rate of data which arrive to the system is called as arrival rate. Arrival information of a data stream usually could not be controlled or predicted. The produced tuples often has fluctuate and high arrival rate [3].

C. *Query* properties

Continuous queries are those which have several processes on new data to generate new results. They are executed in a long lasting mode and generate the results continuously [3].

1) Type of query: Registered queries of a DSMS can be categorized as below [2]:

- One-time query which is evaluated in a specific moment of time.
- Continuous query which is continuously evaluated until arriving tuples to the system. The output is generated continuously. It could be saved as a relation or updated by processing new tuples or the output could be sent as a data stream.

2) Number of operators in a query: Each query includes a set of operators. If Q represents a query and Op_i represents the i^{th} operator, then *n* represents the number of operators in a query, as we have:

$$Q = (Op_1, Op_2, ..., Op_n) \quad (3)$$

3) Arrangement of operators in query plan: Arrangement of operators represents the query plan and its execution procedure. By changing the Arrangement and execution order of the operators, which occurs in query optimization they can lead to decrease cost of query execution and tuple latency.

4) Type of operators: A query includes a set of operators which is shown by symbol O. If we consider o_j as the j^{th} operator and *m* as number of operators in a system, then we have:

$$Q = (Op_1, Op_2, ..., Op_n) \quad , \quad \forall Op_i \in O \quad , \quad O = \{o_1, o_2, ..., o_m\} \; (4)$$

Each Op_i gets one or more streams as input and generates an output. The output stream of each operator may provide the input for one or more other operators.

D. *Execution of Queries*

To execute multiple queries concurrently in a DSMS, first the query selection has to be executed. In second step the manner of accurate execution of queries has to be specified. The first and second steps are discussed as scheduling queries and scheduling query operators.

1) Scheduling queries: To schedule queries in non-continuous systems algorithms like the FIFO can be used but they could not be applied in DSMSs considering the continuous nature of queries and streams. So we have to use scheduling algorithms with circular manner, like the RR and the WRR (weighted RR). The Quantum length in such scheduling algorithms is an important factor of response time of the system.

2) Scheduling query operators: A DSMS includes several query plans which are connected as a DAG of operators which are connected by connector queues and transforms the input stream into the output stream. Considering the high rate and the input variable of data stream, also the limitations on resources such as processor and memory, scheduling algorithms of operators are important factors on responding the queries [5]. Scheduling algorithms like *Round-Robin, FIFO, Greedy, Chain* and *two-Phase* are proposed algorithms in scheduling the query operators [6, 7].

E. Output properties

1) Type of output: Four types of output are considered for a DSMS: *output as a data stream (continuous), output as a relation, output as an announcement* and *output at once*.

2) Number of outputs: number of outputs affects the DSMS's response time since for example, preparing a stream as the output result takes less time rather than preparing a relation in addition to the output stream.

F. System propertie*s (static conditions)*

Depending on total or the static conditions of the DSMS, even in case of number of queries, amount of available memory, number of processors and processing capacity of the them, effective parameters on response time in this category can be considered as below:

1) Number of registered queries: Continuous queries are stored in DSMS and are used permanently to process the queries. In a DSMS the achieved data of various query executions from resources is processed.

2) query registration time : Queries are separated into two categories of below about the time of registration as below:

-Predefined queries: before starting to receive data stream(s) the query is registered. This kind of query is often a continuous type of queries.

-Ad-hoc queries: after starting data streams arrival, it can be continuous or one time use and may need to process previous data.

3) Number of processes (logical machines): In parallel processing of a query, we have number of N_p same processes (logical machines) which are associated. These machines can be physical machines (such as processes of a multi-processor system or nodes of a clustered computer) or virtual machines (such as threads which are executed on each cores of a multi core processor) [9].

If a proper architecture for parallel processing of queries in DSMSs is provided, an increase in number of processes can highly improve the response time.

4) Processing capacity of processor(s): Time of completing a job by the computer depends on several factors which first one is the processor speed. Obviously the processing capacity of processors has reverse relation on the response time.

5) System architecture: Several types of architectures exist for parallel machines. The leading ones are *shared memory, shared disk, shared nothing* and *hierarchical* [10].

6) The amount of available memory: If arrival rate of the input stream is greater than the output rate of the output stream or the sliding windows are mostly used in operators of query, the amount of memory consumption will increase. If the amount of required memory is greater than the amount of available memory, the system will be forced to discard the overload or to use secondary memories. If discarding the overload is prevented, it will influence the factor of validity and if the method of buffering on secondary memories is used, in cause of I/O operations, response time will increase.

G. System condition (Dynamic)

Dynamic status of the system includes changes which may occur while systems execution, like the amount of assigned memory to execute a specified query when we have several processes in processing state or occurring a deadlock.

1) Allocated processing capacity of processor(s) for query execution: Considering that the assigned processor(s) to the DSMS are simultaneously assigned to execute other processes, then always a part of the processor is assigned to execute the query.

2) Memory usage: When memory consumption exceeds the available memory, then overload problem occurs and the system is forced to discard some parts of data. As been considered in static state of the system, this influences the quality of output.

3) Overload in DSMS: Overload situation occurs when requested system resources exceeds the available capacity [8,11]. In such a situation, most part of data are accumulated in systems queues which may cause to increase response time if required main memory is available and in cases of insufficient main memory and requirement to data transfer with the secondary memory, vast delay occurs in generating results of the query.

4) Occurring deadlock: If three conditions to occur a deadlock are established in the DSMS, then failure occurs and the response time will infinitely increase.

IV. EVALUATION

As the evaluation process we developed a prototype which been implemented in the Java language with JDK 6.0 on a machine which was equipped with a Core i7 2930 Intel processor and 6GB of RAM in Linux environment. The Input data set includes data of monitoring IP packets which is located in Internet Traffic Archive (ITA) [12]. One of traces, specifically the "DEC-PKT" contains all wide-area traffic of an hour between Digital Equipment Corporation and the rest of the world. This real-world data set is used in our experiments. Two types of monitored packets, the TCP and the UDP packets, are selected as input streams. Each TCP packet contains five items of *source address, destination address, source port, destination port,* and *length*. UDP packets are the same as TCP missing the length of packets.

So elements of the stream are in type of well-structures data (tuple). Registered queries are continuous and registration time of them in system is pre-defined. To schedule queries the RR algorithm and to schedule query operators the FIFO algorithm are used. Also type of relations is assumed. The experiment is done through a 60000 milliseconds time period with 10 times of runs, considering average results of these runs as the final experimental results.

Effect of parameters of input rate, processing capacity of the processor, number of query operators, type of operators, Quantum length of scheduling between queries, number of registered queries of a system, number of processes and size of buffer memory are evaluated.

1) Effects of parameter of Stream arrival rate to the system: Considering the definition if response time based on "the time period since arriving a tuple until it exits as an output tuple" and considering the figure 3, if stream arrival rate increases as much the value that the system is unable to execute desired queries on input tuples, soon buffers of system will be filled of tuples and waiting time for tuples in queues will increase, then these cause the response time to increase. Conversely, if arrival rate of tuples of data streams is less that the response time, then the response time will decrease. Consequently input rate of data stream has direct relation with the response time.

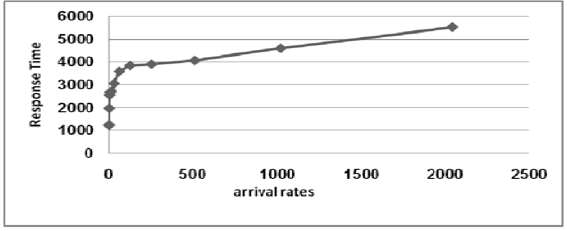


Figure 3- response time vs input rate of data stream

2) Effects of parameter of number of query operators: As shown in figure 4, as much as registered query operators in a system we have, it causes an arrived tuple to get more time to be processed. It means that response time increases while the number of operators increases. So, number of operators has direct effect on the response time.

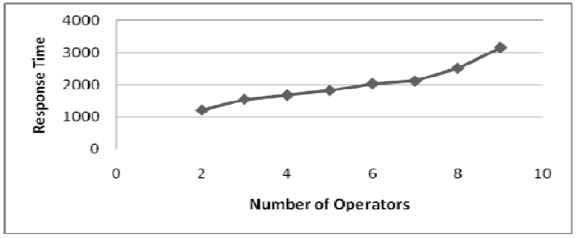


Figure 4- response time vs number of query operators

3) Effects of parameter of type of operators: Type of operator has great influence on response time. For example an operator such as the *Join* is more time consuming than the *Selection* operator. So as much as the response time of an operator been used in query increases, response time of DSMS will increase too. In figure 5 response time of six operators of *Selection, Count, Max, Min, Union* and *Join* are evaluated.

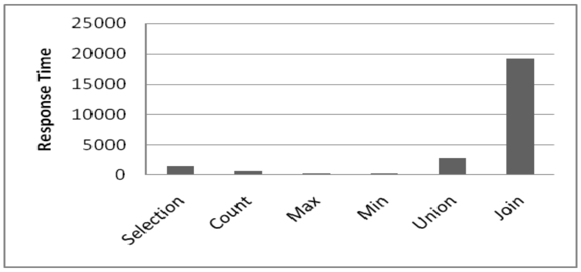


Figure 5 – response time **vs** type of operator

4) Effects of parameter of Quantum length in scheduling queries: Considering the continuous nature of queries and that the best scheduling is one such as the RR, the quantum length been used in algorithm is an important parameter of response time. As shown in figure 6, in short quantum samples the response time is high. Then in an optimized mode, by increasing the quantum length the response time increases too.

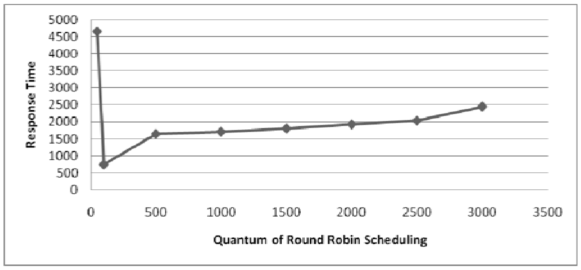


Figure 6 – response time vs quantum length of the RR algorithm

5) Effects of parameter of number of registered queries of a system: Number of queries which are executed on input tuples is competing to achieve system resources such as the processor and memory and when resources are allocated to a query which is under execution, on execution tuples of other queries will be kept on a waiting queue and the response time increases. As shown in figure 7, the number of queries has direct effect on the response time.

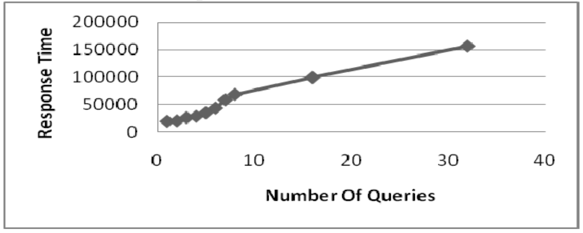


Figure 7- response time vs number of registered queries

6) Effects of parameter of processing capacity of the processor(s): In figure 8, effects of number of queries parameter on a Dual core and a Quad core processor is studied. It is obvious that processing capacity of the processors has reverse relation with the response time. As much as the capacity increases the response time of each operator and consequently the total response time of the system will decrease.

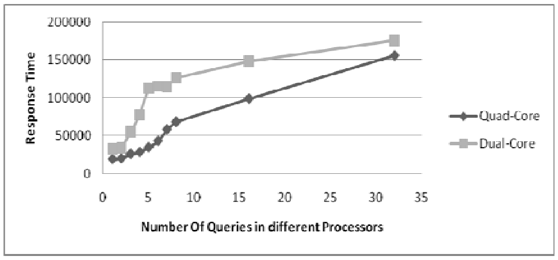


Figure8 – response time vs processing capacity of processors

7) Effects of parameter of number of queries: As we can see in figure 9, if a proper architecture is used for parallel processing of queries in a DSMS, increasing the number of processes can great influence on improving the response time. So the parameter of number of processes or N_p has direct effect on the response time.

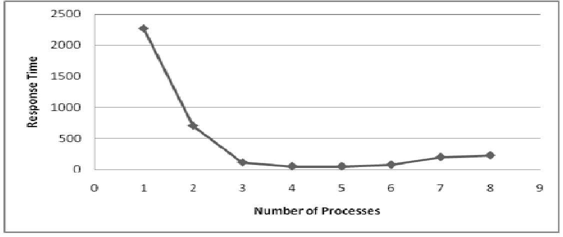


Figure 9 – response time vs number of processes

8) Effects of parameter of memory buffer size: Considering the evaluation result of figure 10, as much as the buffer size increases, tuples in waiting queues will wait more and the response time of the system increases, on the other hand when the buffer size is small, the system is forced to discard some tuples, so accuracy of result decreases.

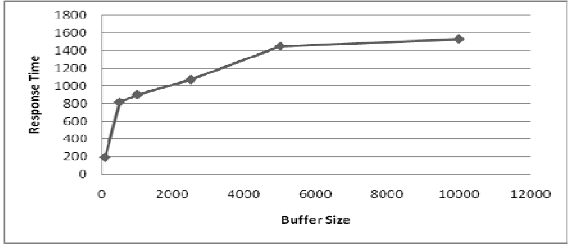


Figure 10 – response time vs memory buffer size

V. Conclusion and Future Works

Rapid and bursty arrival of input data streams raises this fact that, to be fast is a major challenge for a DSMS. We considered the time period from arriving a tuple to the DSMS, until it exits as an output tuple as the response time of the system. Many different factors affect on response time of the system. In this paper, these factors were studied in six categories: *data stream properties, query properties, query execution properties, output properties, static properties* and also *dynamic status of the system*. Experimental results are shown to illustrate the impact of each parameter on the response time.

As the further works, dynamically setting of the changeable parameters based on the of machine learning techniques, determining and providing a function to compute the response time w.r.t the values of effective parameters and also using proper mechanisms to estimate the response time in a DSMS can be followed.

References

[1] A. Arasu, et. al., "*STREAM: The Stanford Stream Data Manager*". In: Proc. of ACM SIGMOD, USA, 2003.

[2] B. Shivnath, "*Adaptive Query Processing in Data Stream Management Systems*", Ph.D. thesis, Department of Computer Science, Stanford University, USA, September 2005.

[3] S. Chakravarthy, et. al., "*Stream data processing: a quality of service perspective: modeling, scheduling, load shedding, and complex event processing*", book, springer, USA, 2009.

[4] Y. Bai, et. al., "Minimizing Latency and Memory in DSMS-a Unified Approach to Quasi-Optimal Scheduling", Proceedings of the 2nd international workshop on Scalable stream processing system, University of California, Los Angeles, 2008.

[5] D. Abadi, et. al., "*Aurora: A New Model and Architecture for Data Stream Management*", In VLDB Journal (12)2: 120-139, August 2003.

[6] B. Babcock, et. al., "*Operator Scheduling in Data Stream Systems*", VLDB Journal, 13(4):333–353, 2004.

[7] D. Carney, et al., "*Operator Scheduling in a Data Stream Manager*", in Proceedings of the 29th international conference on Very large data bases, Germany, pp. 838-849, 2003.

[8] N. Tatbul, et. al., "*Load Shedding Techniques for Data Stream Management Systems*", Ph.D. thesis, Brown University, May 2007.

[9] A.A.Safaei, et. al., "*Parallel Processing of Continuous Queries over Data Streams*", Distributed and Parallel Databases, Volume 28, Numbers 2-3, 93-118, 2010.

[10] A Silberschatz, "*Database System Concepts*", book, 5th edition, 2005.

[11] F. Reiss, et. al., "*Data Triage: An Adaptive Architecture for Load Shedding in TelegraphCQ*", U.C. Berkeley Department of Electrical Engineering and Computer Science, And Intel Research Berkeley, Conference paper, Proceedings of the 21st International Conference on Data Engineering, ICDE 2005, 5-8 April 2005, Tokyo, Japan, 2005.

[12] Internet Traffic Archive, http://www.acm.org/sigcomm/ITA/

[13] B. Babcock, et. al., "*Models and Issues in Data Stream Systems*", Invited paper in Proc. of PODS 2002, June 2002.

[14] S. Chandrasekaran, et al., "*TelegraphCQ: Continuous Dataflow Processing*", in ACM SIGMOD, 2003.

[15] B Babcock, et al., "*Chain: Operator Scheduling for Memory Minimization in Data Stream Systems*", Proceedings of the ACM SIGMOD International conference, 2003.

[16] M. A. Sharaf, "*Preemptive Rate-Based Operator Scheduling in a Data Stream Management System*", in IEEE/AICCSA, 2005.

[17] M. S. Soliman, G. Tan, "*Operator-scheduling using dynamic chain for continuous-query processing*", IEEE Int. Conference on Computer Science and Software Engineering , 2008.

[18] S. Chakravarthy, et. al., "*Scheduling Strategies and Their Evaluation in a Data Stream Management System*", Springer LNCS 4042, 2006.

[19] M. A. Sharaf, et. al., "*Scheduling Continuous Queries in Data Stream Management Systems*", in PVLDB, 2008.

[20] B. Srivastava, et. al., "*Exploiting k-Constraints to Reduce Memory Overhead in Continuous Queries over Data Streams*", Technical Report, November 2002.

[21] M. Ghalambor, et. al., "*DSMS scheduling regarding complex QoS metrics*", IEEE/ACS International Conference on Computer Systems and Applications (AICCSA), 10-13 May 2009.

[22] S. Chakravarthy, et. al., "*Stream data processing: a quality of service perspective: modeling, scheduling, load shedding, and complex event processing*", book, springer, USA, 2009.

Utilizing the Significant Spectral Bands for Image Outputs for SAV from the Water Effect Correction Module

Gibel Gaye[1], Hyunju Kim[1], and Hyun J. Cho[2]
[1]Department of Computer Science, Jackson State University, Jackson, MS, USA
[2]Department of Biology, Jackson State University, Jackson, MS, USA

Abstract - *A water effect correction algorithm was developed by our research team in order to correct/remove water effects for improved detection of SAV (Submerged Aquatic Vegetation) from remotely sensed hyperspectral data. The algorithm was implemented in IDL (Interactive Data Language) and incorporated to ENVI as user-defined module. This paper specifically explains the two different methods in presenting outputs from the module: spectral profiles of selected pixels and two-dimensional images on the significant spectral bands. These significant bands were identified and further verified with a series of PCA (Principal Component Analysis) and a subsequent clustering process, and utilized for presenting output images and spectral profiles within the correction module.*

Keywords: Remote Sensing, Hyperspectral Data, SAV (Submerged Aquatic Vegetation), Water Impact Correction, Clustering

1 Introduction

SAV (Submerged Aquatic Vegetation) is a group of vascular plants that grows underwater. SAV includes seagrass species that play a vital role in the ecological processes/dynamics/productivity of shallow coastal and marine ecosystems. Thus, information on SAV distribution and abundance is widely used as an indicator of aquatic environmental quality.

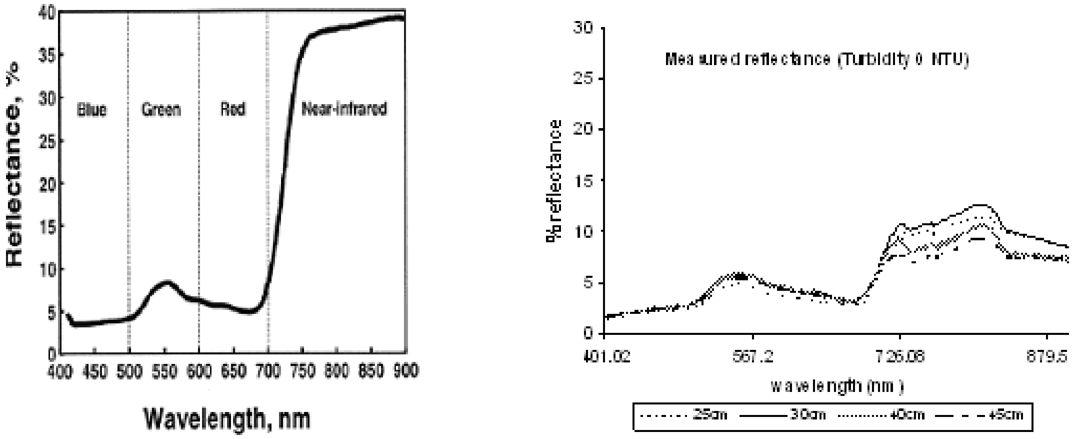
Terrestrial green plants produce distinctive spectral characteristics, such as low reflectance in the visible light and high reflectance in near-infrared (NIR). These characteristics have been used to develop multispectral indices, such as Simple Vegetation Index (SVI = NIR reflectance – red reflectance) and Normalized Difference Vegetation Index (NDVI = (NIR reflectance – red reflectance) / (NIR reflectance + red reflectance)). These indices are used to assess the distribution, abundance, and health of green vegetation from multispectral or hyperspectral data.

However, SVI or NDVI may not be effectively used for plants that grow underwater or that are temporarily flooded [1] because the water overlying the vegetation canopies reduces the vegetation effects of 'red absorption' and the 'NIR reflectance' [2][3][8]. In addition, the water turbidity further intervenes in the vegetation effects that are presented within the spectral bands. For example, Figure 1 compares the two spectral profiles of green plants on the ground and underwater. The profiles of underwater plants indicate that the water effects attenuate the spectral characteristics of vegetation.

In order to remove or correct the water effects, our research team developed the water correction algorithm in [11]. This algorithm was designed to remove water effects from the hyperspectral data taken over shallow waters that contain SAV. We recently implemented this algorithm in IDL (Interactive Data Language) and have incorporated it into ENVI [5], the software that is widely used in remotely sensed image processing and analysis. Since the water effect` correction module, our implementation of the algorithm, processes large amounts of data such as hyperspectral data and needs to present outputs in efficient and effective ways, we also studied which spectral bands were significant in detecting SAV [7]. This paper specifically reports how we designed and implemented the two different methods in order to present outputs to the users of the water correction module.

2 Related work

Mapping of SAV using remotely sensed data has focused on classifications based on signal variations in the multispectral bands, especially those in the short visible wavelengths with high water penetration, but the NIR region was often not used due to its high attenuation in water. However, according to the research in [1][8][9], SAV produces the spectral characteristics which can be used in mapping of SAV from the data: the NIR reflectance becomes two peaks near at approximately 710 – 715 nm and 810 – 815 nm due to the water absorption of the energy in between those two NIR peaks.



(a) Spectral profile of green plants on the ground (http://www.cnr.berkeley.edu)

(b) Spectral profiles of underwater plants at different depths

Figure 1. Spectral profiles of green plants on the ground and underwater

In order to study SAV effects within the composite upwelling hyperspectral signals, we conducted various controlled indoor experiments, which measured upwelling energy over water tanks with white, gray, and black bottoms. Details of the experiments conducted by our team can be found in [4]. The data from these experiments were utilized in developing the water correction algorithm [4][11]. The algorithm corrects water effects including absorption and scattering at varying water depths as well as at varying turbidity levels.

We also applied a series of PCA (Principal Component Analysis) and model-based clustering over the collected data in order to identify significant spectral bands in detecting SAV [7]. In this study, we analyzed the data in terms of upwelling energy, which was converted to reflectance, water depth, and spectral band. The result showed that the spectral bands between about 500 – 600 nm and between about 700 – 900 nm were the most significant bands that retain SAV characteristics. We also noted that as the water depth increased, more bands in the range of 500 – 600 nm and fewer bands in the range of 700 – 900 nm were identified. Figure 2 shows these significant bands with respect to the water depths.

3 Utilizing the SAV significant bands for the outputs

3.1 Verification of the significant bands

In order to further verify the significant bands from our previous study, we conducted an experiment over the hyperspectral image data that were collected in 43 bands, ranged from 400 nm to 900 nm. We sampled 3 subsets from the data, which consist of an area of shallow water with bare sediment without SAV beds, a sparsely-populated SAV area, and a densely-populated SAV area. The same method reported in [7], a series of PCA followed by model-based clustering was applied over the sample data, and successfully identified SAV-contained pixels from the bare sediment and water (no SAV)-only pixels when all of the 43 bands were used.

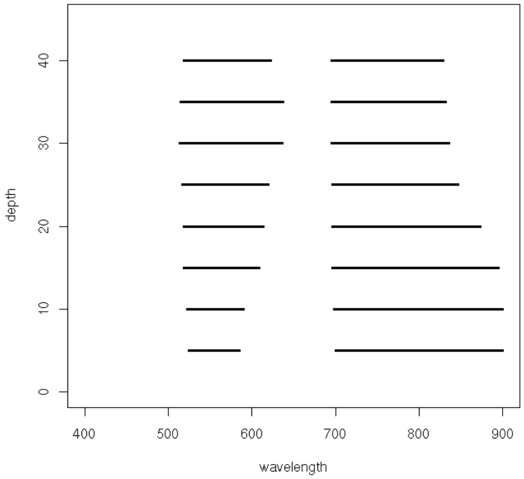
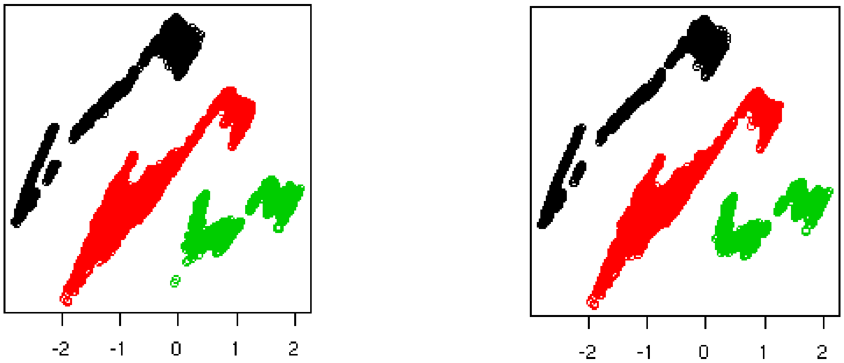


Figure 2. The most significant spectral bands for detecting SAV [7]

We then limited our data to the ones from the significant bands between 500 – 600 and 700 – 850 for each pixel from the sample data, which resulted in 28 bands out of the 43 bands. The same PCA and clustering method was applied over the reduced data, and the result showed that about 97% of the data were meaningfully clustered after the 2nd iteration. This implies that SAV components were sufficiently-embedded within the significant bands.



(a) Spectral features of the 43-band data

(b) Spectral features of the 28-band data

The components of the first group at the top indicate pixels of water; the components of the second group in the middle indicate pixels of sparsely-populated SAV area; the components of the third group at the right bottom corner indicate pixels of densely-populated SAV area.

Figure 3. Comparison of the spectral features of the original and reduced data sets

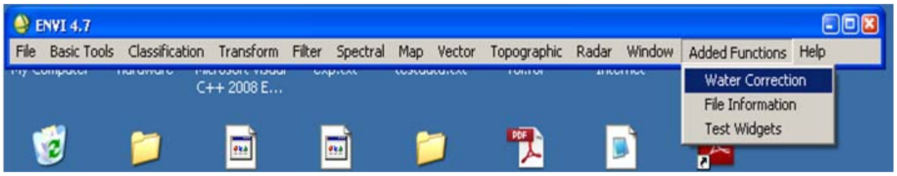


Figure 4. The water effect correction module on ENVI

Figure 3 shows the spectral features of the original data with 43 bands and the reduced data with 28 bands. As indicated in the figure, the two sets contain the spectral features that are similar to each other.

3.2 Design and implementation of presenting outputs

The water effect correction algorithm was implemented in IDL as mentioned earlier. Since ENVI is widely used in the areas of remote sensing for image processing and analysis, we incorporated the implemented algorithm to ENVI as a user-defined function. Consequently, the user of ENVI can easily use our water effect correction module from the ENVI's main menu. This was done by editing the ENVI's menu files and porting the module to the appropriate directory of ENVI so as to make it available to the user as a call to the function. Figure 4 shows our module as an entry from the ENVI's main menu.

When the water correction option is selected, the module asks the user for information about the input file and brings a GUI (Graphical User Interface) asking for values of the correction parameters such as water depth and turbidity. As the user selects to apply the algorithm, the module executes the correction algorithm over the input data and generates an output file that contains the resulting data from the correction algorithm.

In order to present the output from the algorithm efficiently and effectively, we designed the two different methods, which are presenting spectral profiles of selected pixels and presenting two-dimensional images on a selected spectral band. Since the module needs to handle large amounts of data, it is important to present the output in a selective way so as to support the user effectively for the purpose of identifying SAV from the corrected data.

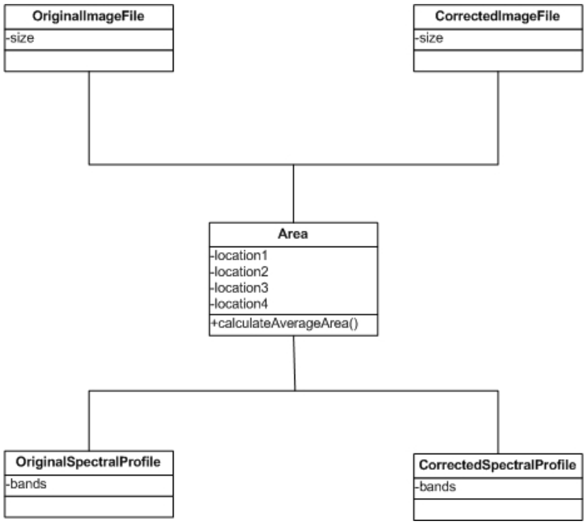


Figure 5. Class diagram for spectral profiles of selected pixels

When presenting spectral profiles, the module displays two spectral profiles, which are the one from the original data and the other from the corrected data so that the user can easily compare the two. The module first asks the user to input *x* and *y* coordinates of four pixels, which are assumed to be the four corners of a rectangular area on the two-dimensional image of the data. Currently, the user is allowed to specify rectangular areas only for the spectral profiles. In case the user inputs the same *x* and *y* coordinate for the all four pixels, the module presents two spectral profiles of a single pixel, indicated by the coordinate from the original and the corresponding corrected data. When more than one pixel is selected with different *x* and *y* coordinates, the module calculates the averaged reflectance values from the selected pixels for each spectral band, and generates the profiles for the original and corrected pixels. The class diagram in Figure 5 shows the entities that are involved in this output process.

In presenting the output in the form of two-dimensional image, the module displays the output image from the significant band first as default. Although the user can select any of the specified spectral bands from the input, with the image from one of the significant bands, the user can better examine and compare SAV characteristics.

The GUI we developed for the module contains textboxes, buttons, drop-down boxes for specifying output options and parameters. The user uses four textboxes to specify *x* and *y* coordinates and a drop-down box to select a spectral band. Along with the module itself, the GUI was also implemented in IDL with ENVI 4.7.

4 Resulting outputs and discussion

Figure 6 shows the two spectral profiles of averaged reflectance values from an input and the corresponding corrected output produced by the algorithm. As the user selected more than one pixel, the output was presented in the form of averaged spectral profiles. As seen in the figure, the profile from the area within the corrected data clearly shows the SAV characteristics that were attenuated within the original data. With these spectral profiles, the user can easily see the improvement and compare the profiles across the spectral bands.

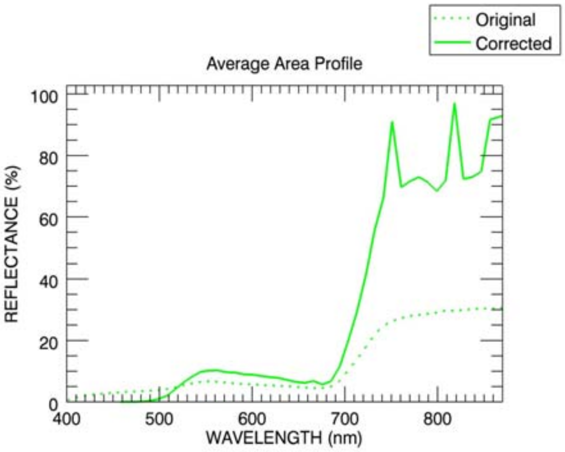
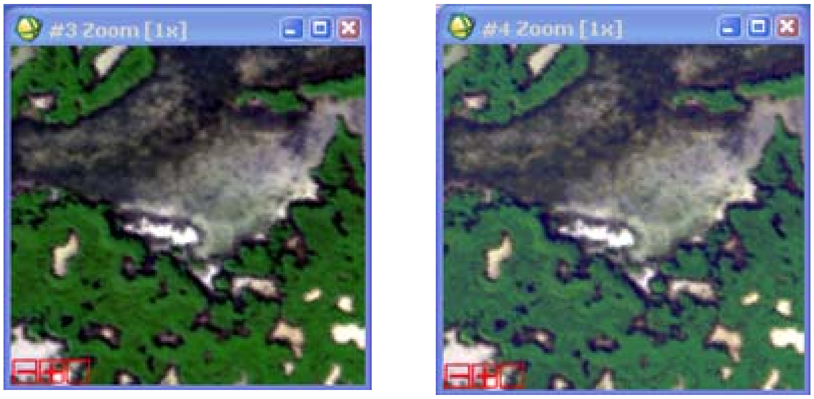
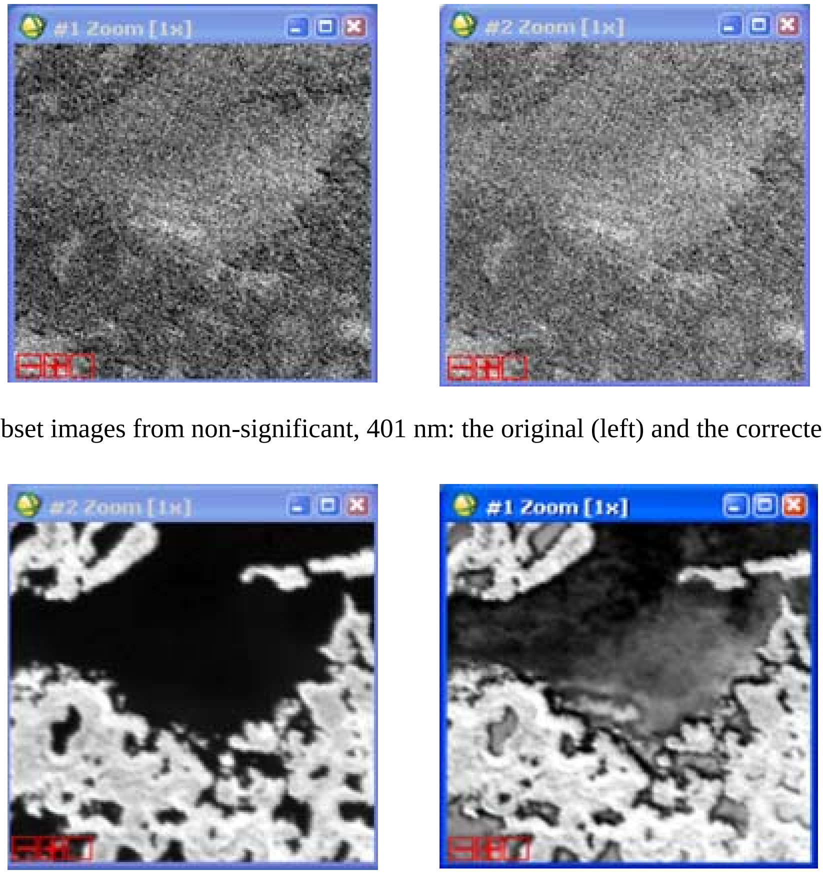


Figure 6. Spectral profiles of averaged reflectance values



(a) Subset image from the original (b) Subset image from the corrected

Figure 7. Images in RGB from the original and corrected data sets



(a) Subset images from non-significant, 401 nm: the original (left) and the corrected (right)

(b) Subset images from significant, 741 nm: the original (left) and the corrected (right)

Figure 8. Comparison of images from non-significant and significant bands

The images in Figure 7 and 8 were generated from the hyperspectral data that had been collected from Redfish Bay area in Texas. The data consist of 63 spectral bands ranging from 400 to 900 nm. Figure 7 presents the two subsets of the original and corrected data in ENVI's RGB format. Part (a) of Figure 8 shows these two subsets in the images from one of the non-significant spectral bands, namely 401 nm. The two images show little difference, which also provides little information about the existence of SAV within the area. On the other hand, the two subsets in the images from 741 nm in Part (b) provide much richer information. This indicates that the output method utilizing the significant bands is considered effective in locating and analyzing SAV components from the corrected data. Besides the current output options, in order to better support the module user, we plan to enhance these two methods by incorporating more significant bands and more options for presenting outputs.

5 Conclusion

This paper reports how we designed and implemented the two different methods for presenting outputs from the water effect correction module. The module was incorporated into ENVI for easy use, and presents the outputs in the ways of spectral profiles and two-dimensional images. The spectral bands that were identified as significant for SAV have been utilized in presenting the output for better support of the module user. We expect that the two output methods help the user analyze more features of SAV from hyperspectral data and provide a platform for comparing SAV features from different data sets.

Acknowledgement:
This research was supported by NGA (National Geospatial-Intelligence Agency) University Research Initiative Grant (HM1582-08-1-0049).

6 References

[1] Beget, M.E. and Di Bella, C.M. “Flooding: The Effect of Water Depth on the Spectral Response of Grass Canopies”, Journal of Hydrology 335: 285-294, 2007.

[2] Cho, H.J. “Depth-variant Spectral Characteristics of Submersed Aquatic Vegetation (SAV) Detected by Landsat 7 ETM+”, International Journal of Remote Sensing 28: 1455-1467, 2007.

[3] Cho, H.J., Kirui, P. and Natarajan, H. “Test of Multi-spectral Vegetation Index for Floating and Canopy-forming Submerged Vegetation”, International Journal of Environmental Research and Public Health 5: 477-483, 2008.

[4] Cho, H.J. and Lu, D. “A Water-depth Correction Algorithm for Submerged Vegetation Spectra”, Remote Sensing Letters, 1: 29-31, 2010.

[5] ENVI, http://www.ittvis.com/envi

[6] Fraley, C. and Raftery, A. “Model-based Clustering, Discriminant Analysis, and Density Estimation”, Journal of the American Statistical Analysis 97: 611-631, 2002.

[7] Gaye, G., Kim, H., and Cho, H.J. “A Study on Spectral Bands for Detecting Submerged Aquatic Vegetation from Hyperspectral Data”, In Proc. of ADMI2010 (in CD-ROM), 4 pages, 2010.

[8] Han, L. and Rundquist, D.C. “The Spectral Responses of Ceratophyllum Demersum at Varying Depths in an Experimental Tank”, International Journal of Remote Sensing 24: 859-864, 2003.

[9] Jackson, T., Chen, D., Cosh, M., Li, F., Anderson, M., Walthall, C., Doraiswamy, P., and Hunt, E. R. “Vegetation Water Content Mapping using Landsat Data Normalized Difference Water Index (NDWI) for Corn and Soybean”, Remote Sensing of Environment 92: 475-482, 2004.

[10] Jollife, I.T. “Principal Component Analysis”, Springer, 1986.

[11] Lu, D. and Cho, H.J. “An Improved Water-depth Correction Algorithm for Seagrass Mapping Using Hyperspectral Data”, Remote Sensing Letters 2: 91-97, 2011.

A Fast Algorithm Combining FP-Tree and TID-List for Frequent Pattern Mining

Lan Vu, Gita Alaghband, Senior *Member, IEEE*
Department of Computer Science and Engineering, University of Colorado Denver, Denver, CO, USA
{lan.vu, gita.alaghband}@ucdenver.edu

Abstract - *Finding frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among variables in transactional databases. The performance of a frequent pattern mining algorithm depends on many factors. One important factor is the characteristics of databases being analyzed. In this paper we propose FEM (FP-growth & Eclat Mining), a new algorithm that utilizes both FP-tree (frequent-pattern tree) and TID-list (transaction ID list) data structures to discover frequent patterns. FEM can adapt its behavior to the dataset properties to efficiently mine short and long patterns from both sparse and dense datasets. We also suggest a combination of several optimization techniques for effectively implementing FEM to speed up the mining process. The experimental results show that a significant improvement in performance is achieved.*

Keywords: knowledge mining, data mining, frequent pattern mining, association rule mining, frequent itemset.

1 Introduction

Frequent pattern mining is one of the fundamental problems in data mining. It plays an important role in finding many types of relationships among data such as associations [1], correlations [4], causality [5], sequential patterns [6], episodes [7] and partial periodicity [8]. Moreover, it helps in data indexing, classification, clustering, and other data mining tasks as well [9].

The frequent pattern mining problem aims to search for groups of itemsets, subsequences, or substructures that co-occur frequently in a dataset. In a typical transactional database, the number of distinct single items and their combinations are usually very large. For a small minimum support threshold, the number of itemsets generated can be extremely large. Hence, it is a great challenge to design algorithms for mining frequent patterns that scale with memory size and run in reasonable time [22]. Among numerous proposed methods, Apriori, FP-growth and Eclat are most popular and widely used.

The Apriori algorithm [1] utilizes the property, that a k-itemset is frequent only if all of its sub-itemsets are frequent, to reduce the search space of frequent itemsets. It is built on a recurrence relation where to find frequent k-itemsets, Apriori uses frequent (k-1)-itemsets found in the previous step. Many variants of Apriori have been proposed to improve the mining efficiency, e.g. direct hashing and pruning (DHP) [10], sampling technique [23], dynamic itemset counting (DIC) [24]. FP-growth [3] works in a divide-and-conquer fashion. It compresses the database into a FP-tree, constructs its conditional FP-trees and recursively mines on these trees to find the frequent itemsets. Some extensions of FP-growth include an array technique to reduce the FP-tree traversal time [13], the usage of FP-array data structure [16], H-mine [17] and nonordfp [18]. While Apriori and FP-growth explore the horizontal data format, Eclat [2] uses the vertical TID-lists of frequent (k-1)-itemsets to find the frequent k-itemsets by intersecting these TID-lists and computing their resulting supports. Mafia [11], AIM [14], mining using diffsets [21] are similar approaches in using the vertical data format. However, all these methods have advantages and disadvantages that make one suitable for specific databases and computing platforms. Hence, hybrid method is another approach to exploit the benefits of many mining methods.

In this paper we propose FEM (FP-growth & Eclat Mining), a new algorithm for frequent pattern mining that combines the techniques used in the FP-growth and Eclat algorithms. Our approach uses FP-tree to store the compact database in memory and recursively mine the frequent patterns from this data structure similar to the FP-growth approach. In addition, FEM will automatically switch from mining FP-trees using FP-growth to mining TID-lists using Eclat depending on the structure of the currently processed data. In order to enable the mining task using Eclat method, during the pattern growth process, the conditional pattern base [3] of a frequent item will be transformed into TID-lists [2] if its size is small enough for better mining on the vertical data structure. FEM can adapt its behavior to the database characteristics for efficiently mining both short and long patterns from sparse and dense datasets. We also suggest a combination of several optimization techniques for implementing FEM to speed up the frequent pattern mining process. Our experimental results show that a significant improvement of performance is achieved using our proposed approach.

This paper is organized as follows. Section 2 provides the essential background knowledge. Our FEM algorithm is described in section 3. Section 4 presents optimization techniques for FEM. Experiments and performance study are presented in section 5. The final section summarizes our study and points out some future research directions.

2 Background

2.1 Frequent pattern mining problem

The frequent pattern mining problem can be stated as follows: Let $I = \{i_1, i_2, . . . , i_n\}$ be the set of all distinct items in transactional database D. The *support* of an *itemset* α (a set of items) is the number of transactions containing α in D. A *k-itemset* α, which consists of k items from I, is frequent if *α's support* is no fewer than δ, where δ is a user-specified minimum support threshold. Given a database D and minimum support threshold δ, the problem statement is to find the complete set of frequent itemsets in D.

For example, given the dataset in table I and minimum support threshold δ=3, the frequent 1-itemsets include *a, b, c, d and e* while *f* and *g* are infrequent because *f* and *g* occur only 2 times. Similarly, *ab, ac, ad, ae, bc, bd* are frequent 2-itemsets and *abc* is the only frequent 3-itemset found.

TABLE I
A DATASET WITH MINIMUM SUPPORT THRESHOLD = 3

TID	Items	Sorted frequent items
1	b,d,a	a,b,d
2	c,b,d	b,c,d
3	c,d,a,e	a,c,d,e
4	d,a,e	a,d,e
5	c,b,a	a,b,c
6	c,b,a	a,b,c
7	f,g	
8	b,d,a	a,b,d
9	c,b,a,e,f,g	a,b,c,e

2.2 The FP-growth algorithm and FP-tree structure

FP-growth is a well-known algorithm proposed by Han *et al.* [3] for frequent pattern mining. It utilizes the FP-tree (frequent pattern tree) to efficiently discover the frequent patterns. FP-tree is an extended prefix-tree structure that uses the horizontal database format and stores compressed information about patterns. It consists of one root node, a set of item prefix subtrees as the children of the root, and a frequent-item header table. Each node of the FP-tree includes an *item name,* a *link* to the next node in the linked list of its appropriate frequent item and a *count* indicating the number of transactions that contains all items in the path from the root node to the current node. The header table stores the frequent items in frequency descending order. Each entry of this table includes *item name*, item *support* and a *link* to the head node in the linked list of its frequent item. The FP-tree is organized so that if two transactions share a common prefix, the shared part can be merged as long as the count properly reflects the occurrence of each itemset in the transactions.

FP-growth requires only two scans on the dataset. In the first scan, the frequency items are found to generate the header table. The dataset is re-scanned to achieve and sort the frequent items in each transaction as illustrated in the third column in Table 1. These items are inserted into FP-tree in frequency descending order. If the appropriate node of an item exists, its count is increased by one. Otherwise, a new node is inserted to the FP-tree. Figure 1 illustrates the FP-tree constructed from the dataset in Table 1.

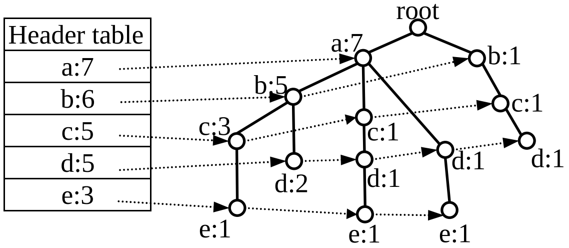


Fig. 1 FP-tree constructed from the dataset in Table I

The next step is to mine the FP-tree by constructing the conditional pattern base and then constructing the conditional FP-tree of each frequent item as described by an example in Section 3.2, and performing mining recursively on such a tree [3]. The frequent items of a resulting conditional FP-tree are combined with the suffix-pattern to generate the new frequent patterns.

2.3 The Eclat algorithm and TID-list structure

Eclat is another efficient algorithm for frequent pattern mining developed by Zaki *et al.* [2]. This algorithm utilizes the TID-list data structure (transaction ID list), a vertical format of database, for the mining task. A TID-list of an item or itemset is a list that stores the IDs of transactions containing that item or itemset. Eclat, similar to FP-growth, applies the depth-first approach to search for frequent patterns and needs only two database scans. It first scans the database to find all frequent items. In the second database scan, it generates the TID-lists of the frequent items. This algorithm organizes frequent k-itemsets into disjoint equivalence classes by common (k-1)-prefixes. The candidate (k+1)-itemsets can be generated by joining pairs of frequent k-itemsets from the same classes. The main advantage of using TID-list is that the support of a candidate itemset can be computed simply by intersecting the TID-lists of the two component subsets. A simple check on the resulting TID-list tells whether the new itemset is frequent or not. Figure 2 demonstrates the TID-lists and the Eclat mining process for the dataset in the Table I.

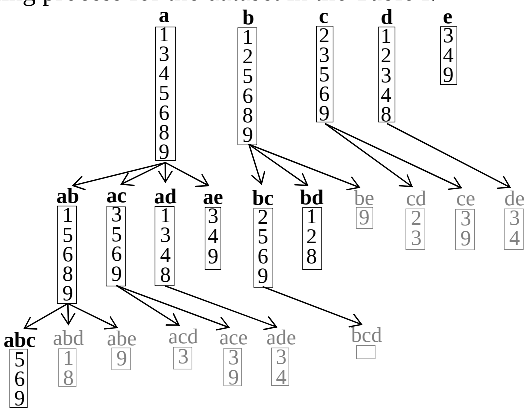


Fig. 2 Mining TID-lists using Eclat

3 The FEM algorithm

3.1 Overview of the FEM algorithm

In the FP-growth algorithm, the frequent patterns are discovered from the conditional FP-trees which are recursively constructed from the original FP-tree. The shape of FP-trees is usually wide for sparse datasets and more compact for the dense ones. In either case, the size of newly generated trees is much smaller than their original one. We found that this size will reduce to a level where mining with an alternative data structure performs better. Additionally, the conditional pattern base of a given item can be easily converted into TID-lists which are more cache-friendly than the trees with linked lists and pointers are. We therefore propose FEM, an algorithm combining mining techniques of FP-growth and Eclat, to discover frequent patterns. FEM flexibly uses both FP-tree and TID-list for its mining task. In the pattern growth process, it switches between FP-growth strategy and Eclat strategy depending on whether FP-tree or TID-list provides better performance. FEM consists of the following three main tasks:

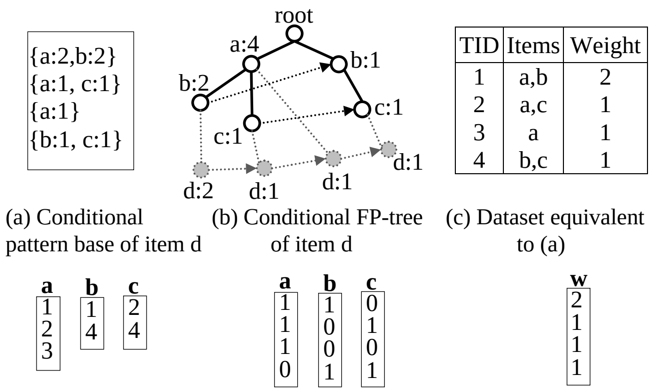
FP-tree construction: Database is scanned for the first time to find the frequent items and create the header table. A second database scan is conducted to get and sort frequent items of each transaction in the support-descending order and then build the FP-tree.

FP-tree mining: This task uses the mining solution of FP-growth. It constructs the conditional FP-trees and recursively mines these trees to find the frequent itemsets. However, before a conditional FP-tree is constructed, FEM will check the size of its conditional pattern base. If it is considerably small, FEM will transform it into TID-lists and switch to mining task using Eclat approach, described next. We represent the TID-lists in bit vector form for its efficiency in computation and memory consumption.

TID-list mining: In this task, we obtain the TID-lists using our bit-vector representation and continue searching for the frequent patterns recursively by logical ANDing these bit vectors. The new patterns are constructed by concatenating the suffix pattern of previous steps with the newly generated frequent patterns. This mining task is inspired by Eclat strategy in using vertical format of database.

3.2 Transforming a conditional pattern base into TID-lists

A conditional pattern base is a "sub-database" which consists of the sets of frequent items co-occurring with the suffix pattern [3]. Each frequent item of a FP-tree has an equivalent conditional pattern base derived from that FP-tree. For example, the conditional pattern base of item d in the FP-tree (Fig. 1) has four sets {a:2, b:2}, {a:1, c:1}, {a:1}, {b:1, c:1} (Fig. 3-a). This conditional pattern base can be used to construct the conditional FP-tree (Fig. 3-b). In the FEM algorithm, we consider these sets as transactions (Fig. 3-c) and transform them into TID-lists for mining using Eclat approach. The transformation is executed by assigning each set with an ID and grouping IDs into lists. In our example, three TID-lists (Fig. 3-d) can be generated including {1, 2, 3} of item a, {1, 4} of item b and {2, 3} of item c. To save memory and benefit bitwise operation efficiency, we represent TID-lists in bit-vector form. The advantage of this approach was shown in [11]. Figure 3-e illustrates the TID bit vectors transformed from the conditional pattern base of item d. In other side, each set in the conditional pattern base has a frequency value indicating the number of its occurrence. We combine all the frequency values into a weight vector which will be used to compute the support of the TID-list. The weight vector in the given example is {2, 1, 1, 1} (Fig. 3-f).



TID	Items	Weight
1	a,b	2
2	a,c	1
3	a	1
4	b,c	1

(a) Conditional pattern base of item d

(b) Conditional FP-tree of item d

(c) Dataset equivalent to (a)

(d) TID-lists

(e) TID bit vectors

(f) Weight vector w

Fig. 3 TID-List and Bit vectors transformed from conditional pattern base of item d

During the *FP-tree mining* stage, thousands or even millions of conditional pattern bases are processed. However, only those whose size is considerably small are transformed into TID bit vectors. We use the number of nodes in the linked list of item α in the original FP-tree to decide whether to switch from *FP-tree mining* to *TID-list mining*. This criteria is used because a small number of nodes in the linked list of α usually indicates a small size of α's conditional pattern base. If this number of nodes is less than or equal to a threshold K, FEM will use Eclat strategy. The value of K depends on the properties of the dataset. In this study, we chose a value of 128 for K. Using this value, the maximum size of a TID bit vector is 16 bytes which is usually smaller than or equal to the size of just one node of FP-tree. The memory size of all TID bit vectors is therefore not greater than the number of items in the frequent pattern base multiplied by 16. This data structure requires much less memory space than an equivalent conditional FP-tree does. Furthermore, the bitwise operations on TID bit vectors will perform faster than creating and manipulating FP-trees. In the given example, the number of node in the linked list of item d is 4 (Fig. 3-b) which is less than 128, so its conditional pattern base will be transformed into TID-lists as described above. Finding an optimal value of K for each specific database will be a subject of our future work.

3.3 The FEM algorithm

The FEM algorithm consists of three components:

a. *FEM-mining:* the mining process first constructs the FP-tree from the original database and it then calls *FP-tree-mining* as represented below.

```
FEM-mining algorithm
Input: Transactional database $D$ and min. support $\delta$
Output: Complete set of frequent patterns
Scan $D$ once to find all frequent items
Scan $D$ a second time to construct the FP-tree $T$
Call FP-tree-mining($T,\varnothing,\delta$)
```

b. *FP-tree-mining*: this component is equivalent to the *FP-tree mining* task described in Section 3.1. Lines 7-12 in the following algorithm show the switching between the two mining tasks. The value of K in our current experiments is 128.

```
FP-tree-mining algorithm
Input: Conditional FP-Tree $T$, $suffix$, min. support $\delta$
Output: Set of frequent patterns
1: If FP-tree $T$ contains a single path $P$
2: Then For each combination $x$ of the nodes in $P$
3:             Output $\beta = x \cup suffix$
4: Else For each item $\alpha$ in the header table of FP-tree $T$
5:   {  Output $\beta = \alpha \cup suffix$
6:      Construct $\alpha$'s conditional pattern base $C$
7:       If item $\alpha$ has more than $K$ nodes in its linked list
8:      Then {  Construct $\alpha$'s conditional FP-tree $T'$
9:                  Call FP-tree-mining ($T',\beta,\delta$)   }
10:      Else  {  Transform  $C$  into  TID  bit  vectors  $V$
11:                                  and weight vector $w$
12:                 Call TID-list-mining ($V,w,\beta,\delta$)  }    }
```

c. *TID-list-mining*: this component is equivalent to the *TID-list mining* task described in Section 3.1which the TID-lists are represented in the bit-vector form. This algorithm is called by FP-tree-mining and recursively mines until no new frequent pattern is found.

```
TID-list-mining algorithm
Input: Bit vectors $V$, weight vector $w$,
         suffix, min. support $\delta$
Output: Set of frequent patterns
Sort $V$ in descending other of its item support
For each vector $v_i$ in $V$
{    Output $\beta$ = item of $v_i \cup suffix$
      For each vector $v_k$ in $V$, $k < i$
      {    $u_k = v_i$ AND $v_k$
            $sup_k$ = support of $u_k$ based on $w$
            If $sup_k \geq \delta$ Then add $u_k$ into  $U$   }
       If all $u_k$ in $U$ are identical to $v_i$
       Then For each combination $x$ of the items in $U$
                    Output $\beta' = x \cup \beta$
       Else If $U$ is not empty
       Then Call TID-list-mining ($U,w,\beta,\delta$)    }
```

4 Optimization techniques for implementing FEM

The performance of a frequent pattern mining algorithm depends on many factors: data structure, database properties, CPU speed, I/O speed, memory size, minimum support threshold, etc. Table III shows the two implementations of the FP-growth algorithm (i.e. FP-growth_B and FP-growth_GZ. As the Table shows these two implementations result in varying performance on different datasets due to using different optimization approaches. We have incorporated a set of optimization techniques for implementing FEM that has effectively improved the runtime performance of our algorithm for variety of datasets. Following are the details:

FP-tree construction: In the second database scan, we pre-load the filtered transactions into a lexicographically sorted list as suggested in [12]; one copy of similar transactions is kept with its count. The transaction list size can be changed to fit the available memory. We organize this list in a binary tree and maintain its order while the list grows in size. This technique reduces the traversal and construction time of FP-tree. It also keeps the nodes most visited together to be allocated closely in the memory. Thus, it speeds up the mining stage as well.

Mining task using FP-growth: We exploit the technique proposed by [13]. An array-based implementation of prefix-tree-structure is used to improve the efficiency of the *FP-tree-mining* by reducing the need of traversal on FP-trees when constructing the conditional FP-tree.

Memory management: A chunk of memory is allocated for each FP-tree when FEM creates a new one and is discarded after all frequent itemsets from this FP-tree are generated. The chunk size is variable. This technique reduces overhead of allocating and freeing nodes [13].

Output processing: We preprocess the most frequent output values and store them in indexed tables as proposed in [15]. In addition, the similar part of two frequent itemsets outputted consecutively is processed only once. Hence, this technique improves considerably computation time on output reporting, especially when output size is huge.

I/O optimization: Data are read into a buffer before being processed into transactions. Similarly, the outputs are buffered and only written when the buffer is full. This technique reduces much I/O overhead.

5 Experiments and Performance study

5.1 Experiments

The experiments were performed on the Altus 2701 machine with dual AMD Opteron 2427, 2.2GHz, 32GB memory, running Linux OS. We used g++ for compilation. Five datasets used in our tests Connect, Mushroom, Accident, Retail and Webdocs are publicly available at the Frequent Itemset Mining Implementations Repository [19] and are reported in the Table II.

TABLE II
DATASETS AND THEIR PROPERTIES

Datasets	Type	Transactions	Items	Average length	Size
Connect	Dense	67557	129	43	8.82 MB
Mushroom	Dense	8124	119	23	557 KB
Accidents	Moderate	340183	468	33.8	33.8 MB
Retail	Sparse	88126	16470	10.3	3.79 MB
Webdocs	Sparse	1623346	52676657	177.23	1.37 GB

FEM was implemented using the optimization techniques in Section 4. For comparison, we benchmarked FEM and three state-of-the-art frequent pattern mining implementations: FP-growth_B by Borgelt [12], FP-growth_GZ by Grahne & Zhu [13] and AIM2 by Fiat & Shporer [14], [15] (i.e. an Eclat based approach) which are available at [19], [20]. The runtime with the considerably low supports for all datasets is reported in the Table III. The detailed performance comparison on Connect, Accident and Webdocs datasets with various supports are presented in Fig. 4 and Fig. 5 and Fig. 6. Table IV presents the runtime distribution between *FP-tree-mining* and *TID-list-mining* of FEM in both absolute runtime and percentage runtime.

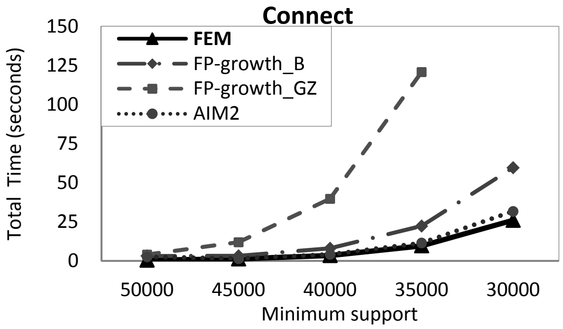


Fig. 4 Runtime on the connect dataset

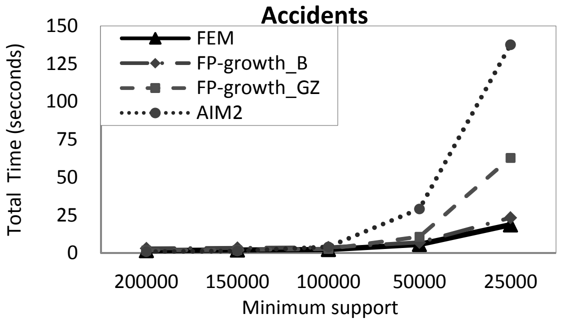


Fig. 5 Runtime on the accident dataset

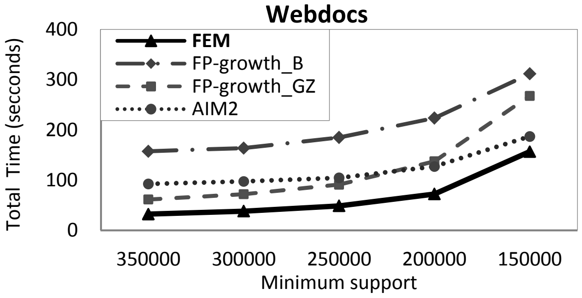


Fig. 6 Runtime on the webdocs dataset

5.2 Performance study

5.2.1 Performance comparison

All the four implementations tested on the same machine behaved differently on the dense and sparse datasets (Table III). FP-growth_B and AIM2 performed better than FP-growth_GZ on the dense datasets Connect and Mushroom. For the sparse datasets, AIM2 ran faster than both FP-growth_B and FP-growth_GZ on the Webdocs but slower on the Retail. For the Accident dataset, AIM2 did not perform as well as either the FP-growth_B or FP-growth_GZ. FEM performed quite well in every case. The performance of FEM reported in Table III and Fig. 4, 5, 6 shows that FEM outperforms the other algorithms on all test databases. Hence, FEM works better than both FP-growth and Eclat on variety of datasets.

TABLE III
RUNTIME (SECONDS) FOR FIVE DATASETS WITH SELECTED MIN SUPPORTS

Datasets	Minimum support	**FEM**	FP-Growth_B	FP-Growth_GZ	AIM2
Connect	30000	**25.55**	59.05	342.03	31.78
Mushroom	50	**21.64**	56.31	306.61	28.47
Accidents	25000	**18.61**	25.36	63.41	137.48
Retail	5	**1.62**	4.88	9.84	51.07
Webdocs	150000	**156.65**	341.56	267.09	186.49

5.2.2 The runtime distribution between two mining tasks

The runtime distribution between *FP-tree-mining* and *TID-list-mining* in Table IV and Fig. 7 shows that FEM distributes the mining workload dynamically depending on the dataset characteristics.

TABLE IV
TIME DISTRIBUTION BETWEEN FP-TREE-MINING & TID-LIST-MINING OF FEM

Datasets	Minimum support	Total time (seconds)	Mining time (seconds)	FP-tree-mining (seconds & %)	TID-list-mining (seconds & %)
Connect	30000	25.55	25.24	0.02 **0.07**%	25.22 **99.93**%
Mushroom	50	21.64	21.60	0.10 **0.5**%	21.50 **99.5**%
Accidents	25000	18.61	15.72	7.72 **49**%	8.00 **51**%
Retail	5	1.62	0.84	0.39 **47**%	0.45 **53**%
Webdocs	150000	156.65	112.51	112.25 **99.7**%	0.26 **0.3**%

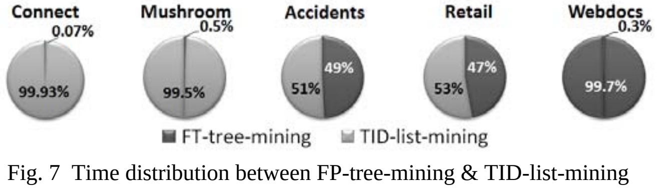


Fig. 7 Time distribution between FP-tree-mining & TID-list-mining

For the dense datasets Connect and Mushroom, *TID-list-mining* was responsible for over 99% of the mining time. The shape of FP-tree of dense datasets is usually compact, so most of the conditional pattern bases satisfy the condition to switch from *FP-tree-mining* to *TID-list-mining*. In contrast, for the very large and sparse dataset Webdocs, *FP-tree-mining* was responsible for 99.7% of the mining time because the large number of big FP-trees were generated and processed. For Accidents and Retail datasets, mining time was distributed equally for the two mining tasks which were 49% vs. 51% on Accidents and 47% vs. 53% on Retail. The percentage time was computed using the runtime of mining stage and the runtime of each mining task. It must be noted that the runtime distribution does not indicate the amount of work. In fact, *TID-list-mining* using faster bitwise operations and better memory layout will process larger amount of data than *FP-Tree-mining* does in the same amount of time. In addition, the runtime distribution will change when the minimum support varies.

6 Conclusion and future work

In this paper, we presented FEM, a new frequent pattern mining algorithm that combines the mining techniques of two famous algorithms FP-growth and Eclat. The performance merit of FEM is achieved by adapting the mining process to match the characteristics of the datasets. The combination of the optimization techniques for implementing FEM contributes to the improvement of performance as well. In future work, we plan to improve FEM further by integrating several other optimization techniques. We will investigate how to find the optimal value of K for specific databases based on their characteristics. In addition, we will study parallel approaches for implementing FEM on parallel and distributed systems as memory limitation is the largest barrier to deploy any sequential frequent pattern mining algorithm on large scale databases.

7 References

[1] R. Agrawal, R. Srikant, “Fast algorithms for mining association rules”, in *Proc. of the Int. Conf. on Very large databases*, pp. 487-499, 1994.

[2] M. Zaki, S. Parthasarathy, M. Ogihara, W. Li, “New algorithms for fast discovery of association rules”, in *Proc. of the 3rd Int. Conf. on Knowledge Discovery and Data Mining*, pp. 283-286, 1997.

[3] J. Han, J. Pei, Y. Yin, “Mining frequent patterns without candidate generation”, in *Proc. of the Int. Conf. on Management of Data*, 2000.

[4] S. Brin, R. Motwani, C. Silverstein, “Beyond market basket: generalizing association rules to correlations”, in *Proc. of the Int. Conf. on Management of Data*, 1997.

[5] C. Silverstein, S. Brin, R. Motwani, and J. Ullman, “Scalable techniques for mining causal structures”. in *Proc. of the Int. Conf. on Very Large Data Bases*, pp. 594–605, 1998.

[6] R. Agrawal and R. Srikant, “Mining sequential patterns”, in *Proc. of the Int. Conf. on Data Engineering*, pp. 3–14, 1995.

[7] H. Mannila, H. Toivonen, and A. I. Verkamo, “Discovery of frequent episodes in event sequences”, *Data Mining and Knowledge Discovery*, Vol. 1 Issue 3, pp. 259-289, Sep. 1997.

[8] J. Han, G. Dong, Y. Yin, “Efficient mining of partial periodic patterns in time series dataset”, in *Proc. of the Int. Conf. on Data Engineering*, pp. 106-115, 1999.

[9] J. Han, H. Cheng, D. Xin, X. Yan, “Frequent pattern mining: current status and future directions”, *Journal Data Mining and Knowledge Discovery*, Vol. 15 Issue 1, pp. 55-86, August 2007.

[10] JS. Park, MS. Chen, P. Yu, "An effective hash-based algorithm for mining association rules", in *Proc. of the Int. Conf. on Management of Data*, pp. 175–186, 1995.

[11] D. Burdick, M. Calimlim, J. Gehrke, “MAFIA: a maximal frequent itemset mining algorithm for transactional databases”, in *Proc. of the Int. Conf. on Data Engineering*, pp. 443–452, 2001.

[12] C. Borgelt, “An implementation of the FP-growth algorithm”, in *the 1st Int. Workshop on OSDM: Frequent Pattern Mining Implementations*, 2005.

[13] G. Grahne, J. Zhu, “Efficiently using prefix-trees in mining frequent itemsets”, in *Proc. of Workshop on FIMI*, pp 123–132, 2003

[14] A. Fiat, S. Shporer, “AIM: another itemset miner”, in *Proc. of Workshop on FIMI*, 2003.

[15] S. Shporer, “AIM2: improved implementation of AIM”, in *Proc. of Workshop on FIMI*, 2004.

[16] L. Liu, E. Li , Y. Zhang, Z. Tang, “Optimization of frequent itemset mining on multiple-core processor”, in *Proc. of the 33rd Int. Conf. on VLDB*, 2007.

[17] J. Pei, J. Han, H. Lu, S. Nishio, S. Tang, D. Yang, “Hmine : Hyper-structure mining of frequent patterns in Large Databases”, In *Proc. IEEE of Int. Conf. On Data Mining*, pp. 441–448, 2001.

[18] B. Racz. “nonordfp: An FP-growth variation without rebuilding the FP-tree”, in *Proc. of ICDM Workshop on FIMI*, 2004.

[19] Frequent Itemset Mining Implementations Repository, *Workshop on FIMI*, 2003-2004 Available: http://fimi.ua.ac.be/

[20] Christian Borgelt, “Frequent Pattern Mining Implementations”, Available: http://www.borgelt.net

[21] M. J. Zaki, K. Gouda, “Fast vertical mining using diffsets”, *Technical Report* 01-1, RPI, 2001.

[22] L. Zhou, Z. Zhong, J. Chang, J. Li, J.Z. Huang, S. Feng, “Balanced parallel FP-Growth with MapReduce”, in *Conference on Information Computing and Telecommunications, IEEE*, pp. 243 – 246, 2010.

[23] H. Toivonen, “Sampling large databases for association rules”, in *Proc. of the 1996 Int. Conf. on VLDB*, pp. 134–145, 1996.

[24] S. Brin, R. Motwani, JD. Ullman, S. Tsur, “Dynamic itemset counting and implication rules for market basket analysis”, in *Proc. of the 1997 Int. Conf. on Management of Data*, pp. 255–264, 1997.

Approach for managing ontology evolution by using Text Mining Techniques

A. Benmarouf Meriem[1], B. Tlili Yamina[2]
[1]Departement of computer science, Badji Mokhtar University, Annaba, PhD Student , Algeria
[2]Departement of computer science, Badji Mokhtar University, Annaba, Doctor in computing vision , Algeria

Abstract - *The maintenance of the domain ontology or a knowledge model after the appearance of changes in the studied domain is an essential stage. Several studies provide methodologies for the maintenance of ontology but only some of them deal with ontologies that are created from texts.*

Text mining techniques provide good results when the processing of texts is done for the purpose of modeling or classification. The objective of our work is the use of text mining techniques for the maintenance of an ontology representing a dynamic domain, according to an analysis of textual changes and their effects on the corresponding ontology.

Keywords: knowledge, text mining, ontology, Corpus, Evolution

1 Introduction

Currently there are many approaches for ontologies construction [1].However, only METHONTOLOGY [2] and ON-TO-Knowledge [3] consider the evolution process in the life cycle of ontology, but they do not provide methods or recommendations to support it. One of the problems encountered in evolving environments is the integration of new knowledge, which raises the problem of the modification of the initial ontology and its adaptation to the new needs. In the state of the art, there are no theoretical complete studies or tools able to analyze the changes effects in the semantic relationship between ontological entities, or changes in the compatibility between the text corpus containing the domain knowledge and the ontology representing this domain [4]. The analysis of the main tools has highlighted the lack of very important functionalities to ensure the identification of the ontological changes [5]. Text mining techniques offer solutions to analyze texts for multiple purposes, classification, modeling, etc. Our work consists in using text mining techniques to extend the changes made on texts onto the ontology. The paper is organized as follows: first, we present a state of the art of text mining techniques, then we present an analysis of the types of textual changes which can intervene and their effects on the ontology [6]. Finally we present the text mining algorithms we have developed: a hierarchical clustering algorithm CAH, with a similarity measure, and the algorithm APRIORI with the measure of confidence. These two algorithms were modified in order to better adapt to our research problem: how to support the ontology evolution in a dynamic environment.

2 State of the art

In this section we present an art state of the text mining techniques.

2.1 Learning lexico-syntactic patterns

The Lexico-syntactic patterns are based on the study of the syntactical regularities between two given concepts. This is an observation of the realization of a relationship in the corpus, in order to simplify the lexical and syntactical context. This simplification establishes a lexico-syntactic pattern. The advantage of this technique is to be targeted on the lexico-syntactic context. It remains effective on small-sized corpora. In [7] the experimental evaluation of a large number of patterns was made by using the CAMELEON tool.

2.2 Clustering Techniques

The classification approach proposed by [8] consists in classifying documents in collections with respect to the sense of every word by using a labeled corpus such as WordNet. Then, for each of the formed collections, the words and their respective frequencies are extracted and compared with the other collections. The approach proposed by [9], shows the construction of domain ontology from textual documents by using two hierarchical classification algorithms.

2.3 Statistical techniques and associations rules

These techniques are based on the calculation of similarity measures. Association rules describe associations between certain elements. An association rule is an implication of the form:

$$A \rightarrow B \qquad (1)$$

where $A = \{t1, t2, ..., tp\}$ and B={tp+1,tp+2,..,tq}.
A rule $A \rightarrow B$ is interpreted as: a text containing the key terms$\{t1,t2,...,tp\}$ also tends to contain the key terms $\{tp+1,$

tp+2, ..., tq}, with some probability given by the confidence of the rule. Several algorithms, e.g., Close and Pascal, implement the extracting rules process [10].The measures of support and confidence cannot alone identify the rules which make sense. Therefore, other measures are also used, such as interest, belief, dependence, novelty and satisfaction [11].

3 Ontology

Ontology evolution concerns the capacity to update an existing ontology following the emergence of new needs [12]. Ensuring the evolution in a dynamic environment where the actors are not all experts in ontology maintenance requires the ability to specify the changes in a simple and understandable way [13]. In the following, we present the types of changes according to two axes; ontology and corpus.

3.1 Analysis of ontological changes

The The structure of an ontology is represented by a tuple:

$$S := \{C, R, <, X\} \qquad (2)$$

where :
C, R : are separated sets containing the concepts and no taxonomic relationships.
< : CXC is a partial order on C, which defines the concept hierarchy.
X : R$\rightarrow$ CXC is the signature of an associative or taxonomic relationship.
The lexicon of the ontology is the tuple

$$L := \{ Lc, Lr, F, G\} \qquad (3)$$

where:
Lc, lr : are separated disjoint sets.
F, G : are two references relationships.
The hierarchy of concepts is defined by a structure:
S0 :={C,<}
The A concept is defined by

$$L0 := \{Lc,F\} \qquad (4)$$

3.2 Analysis of textual changes

Or text corpus is in French.
The corpus may be represented by the following vector :
T = (t1, t2, ..., tn) where
ti : is the included term in the corpus, with the following conditions:

- The meaningless terms are removed (le, la, les, du ,..).
- All uppercase characters become lowercase.
- All conjugations are converted to the infinitive.
- The words with the same root and synonyms are considered as equivalent terms. Let us detail these changes.

3.2.1 Adding a new text

In this case, text mining techniques provide a hierarchical classification algorithm, which is most effective for our study case with a light modification to fit our needs. The result of the algorithm is the classification of the concepts and the relationships in the ontology. This algorithm uses a similarity measure in order to group the concepts together.

3.2.2 Deleting of a text fragment

In this case of evolution (text vs. ontology), we apply some text mining techniques, such as calculation of confidence and the algorithm APRIORI in order to update the ontology while preserving its coherence. Let T be the initial corpus as: T = {t1,…,tn} where ti : the corpus terms. If T2 is the text to be removed; we work on the association rule

$$T \rightarrow T2 \qquad (5)$$

Which means: the text which contains the terms of T also contains the terms of T2.

3.2.3 Corpus update

The updating can be described as a deletion and an addition of a text fragment.

4 Algorithms

4.1 CAH Algorithm

```
Input
T={t1,…,tk} ; the initial corpus
Tn={t1,…,tn} : the added text
O=(C,R,<,K) : the global ontology
C={C1,..,Cn} : all concepts of O with their vectors
MatriSim[n+m, n+m] : matrix of similarity with n number
  of concepts of the new text , m number of existing
  concepts in the ontology O.
S : the threshold of similarity
Output
the classes Symi of the synonymic concepts
For i=1,n+m do Matrsim[i,i]= « »
for i=1, n+m do
  for j=1,n+m do
    if matsim[i,j]<> « » then matrsim[i,j]= X
    endif
   endfor
 endfor
max=0
Repeat
   for i=1 ,n+m do
      for j=1,n+m do
        if matrsim[i,j]>max then
            max:= matrsim[I,j]
        Endif
      Endfor
     Endfor
```

```
if max >seuil then mi:= mi-1 Ci , Cj - Ci, Cj
  matrsim [I,j]:= “”
Endif
Update Matrsim by taking into account the new class
Make the inference by using the new relation Until Gsim
(Ci, Cj) < seuil
Endcah
```

The algorithm continues until it ends the classification of all the concepts of C.

4.2 APRIORI Algorithm

```
Input
Tc = {t1, t2, …,tk} : The set of the extracted words from the initial corpus
Pc= {P1, …, Pp}: The set of the extracted transactions from the initial corpus (a transaction is a Sentence).
Ts= {t1, t2, …,tkk} : The set of the extracted transactions from the text to be removed ( kk <= k).
Ps={p1, p2, …, pkp} The set of extracted transaction from the texts to be removed (kp<=d p)
Output
$C = \frac{Tc \cap Ts}{Ts}$        // calculate the confidence //
if (C=1) then the total ontology is removed and we keep only the initial ontology.
else For i=1, kk do        $Ci = \frac{|\{ti/ti € Tc\}|}{|\{ti/ti€ Ts\}|}$
if Ci=1 Then remove the identifier ti from the identifiers list
else i=i+1 // go to the next ti
Repeat until i=k
Cross the ontology in order to find a concept without identifier then remove.
ENDAPRIORI
```

5 Ontology Maintenance

The ontology maintenance is usually realized by taking into account at once the correspondence between the ontology and the real environment and the ontology expression quality. For example [14] proposes three criteria to measure the
quality of ontology:

-The accuracy of the modeling itself (clarity, standardization of the vocabulary, deletion of nearly anonymous concepts).

-The reliability of the ontology (completeness, consistency, scalability).

In our case study we define two types of coherence:

-Internal Coherence: means the ontological coherence of the entities, for example: no cycle, no redundancy …

-External coherence: means maintaining the dependency of the ontology in question with other ontologies (specific or generic).

6 Conclusion

Many works in text mining deal with how to manage huge number of texts. Most of these works deal with the problem from a statistical point of view without taking into account domain knowledge, and especially when this knowledge is included in texts corpora. In this paper we called to text mining techniques by the two algorithms C.A.H and APRIORI in order to take into account the textual changes in the domain ontology. The studies of the changes which can intervene on texts as well as the formalism of representation that we adopted facilitate ontology maintenance.

7 References

[1]D.ROGOZAN: «A survey on methodologies for developing, maintaining, evaluating and reengineering’’; Ontoweb,N°. IST-2000-29243-Deliverable 1.4.

[2]M.Fernandez, M.Gomez-Perez, and Juristo.N. «Methontology : From Ontological Art Toward Ontological Engineering“;Proc. AAA1’97, Stanford, USA, 1997.

[3] Sure.Y, Staab.S, and Studer.R. «On-To-Knowledge Methodology (OTKIM)”; Springer Verlag, Handbook on ontologies (pp.117-132), 2004.

[4] Stuckenschmidt.H, and Klein.M. « Evolution management for interconnected ontologies”; Workshop on Semantic Integration, ISWC, Island, Florida 2003.

[5] Delia.C. «Gestion de l’évolution d’une ontologie : méthodes et outils pour un référencement sémantique évolutif fondé sur une analyse des changements entre versions de l’ontologie’’ ; Doctoral Research Proposal in cognitive science, (DIC 9410) 2005.

[6] Najla.N, Wassim.J. « Types de changements et leurs effets sur l’évolution de l’ontologie’’ ; Proc. JFO, Tunis, 2007.

[7] Zghal.B, Aufaure.H , Ben Mustapha.N. «Extraction of ontologies from web pages : conceptual modelling and tourism’’; JIT (Journal of Internet Technologies, 2007.

[8]Chagnoux.M, Hernandez.N, Aussenac.N. « From texts to ontologies: non taxonomical relation extraction’’; Proc JFO, Lyon 2008.

[9]Aussenac-Gilles.N, Despres.S, and Szulman.S. «The Terminae method and platform for ontology engineering from texts’’; IOS press, 2008.

[10] Bastide.Y, Taouil.R, Pasquier.N. «Un algorithme d’extraction des motifs fréquents’’ ; Techniques and computing , 21(1) : 65-95, 2002.

[11] Cherfi.H. « Etude et réalisation d’un système d’extraction de connaissances à partir de textes’’ ; PhD Thesis Henri Poincaré University Nancy 1, 2004.

[12] Maedche.A, Motic.B, and Stojanovic.L. « Managing Multiple and distributed ontologies in the semantic web’’; VLBD journal 2003.

[13] Noy.N and Klein.M . «Ontology evolution : Not the same as schema evolution’’; Knowledge and information systems, 5, 2003.

[14] Natalya.F, Deborah.L. « Développement d’une ontologie 101:guide pour la création de votre première ontologie’’ ; Proc CA, 2002.

Integrating Electronic Health Records Using Universal Patient Identifiers KSA

Ahmed Emam, Ahmed Youssef, Samir EL-Masri, Mohammed Alnuem
Dept. of Information Systems
College of Computers and Information Sciences
King Saud University, Riyadh, KSA
aemam@ksu.edu.sa, ahyoussef@ksu.edu.sa, selmasri@ksu.edu.sa, malnuem@ksu.edu.sa

Abstract— **One of the required standards of healthcare information technology (HIT) and specially Electronic Health Records (HER) is to develop a unique patient identifier (UPI) to enable physicians, hospitals, and other authorized users to share clinical and administrative records more efficiently. Till now there is no standard format of UPI which is make it hard to exchange the patient information across the countries and to integrate among heterogeneous medical information systems. This work explores and investigates the desired attributes for any developed UPI such as Unique, Non disclosing, Invariable, Canonical, Verifiable, and Ubiquitous features. A sample case study that demonstrates how much it is necessary for Saudi Arabia to adapt and develop UPI for the patients was introduced. Also, a process framework and schema for the proposed solution was proposed to give a guideline and the basic steps toward develop a solution for adapting UPI in KSA.**

Keywords- Unique Patient Identifier; Electronic Health Records; Cloud Computing; Health Information Technology.

I. INTRODUCTION

Good clinical decisions based on bad data guarantee bad clinical outcomes; it is very true statement and this is the main motivation behind this research. Nowadays, there are many definition of healthcare information technology (HIT) main goal are saving money and significantly improving the quality of health care. International Organization for Standardization (ISO) is a worldwide federation of national standards bodies' aims to setup and preparing International Standards specially ISO Technical Committee 215 for setting up standard for Health informatics. Most of development countries such as European Union, Australia, and United State of America (USA) adapt special standard, for example, USA approved Health Insurance Portability and Accountability Act (HIPAA) on 1996. To replace the paper with an electronic record while maintaining all patient's care, Electronic Medical Record (EMR) or Electronic Health Record (EHR) system become more than essential. Therefore, EHR is a computer program where patient records are created, used, exchanged, stored and retrieved. Because every healthcare provider keeps a separate paper or electronic medical record for each patient, there is no ability to integrate information between the various HER systems. When data is integrated by using EHR system, patient care improves and HIPAA compliance is ensured. HIPAA mandated setting up special requirement to improve the quality of health care and preserve the patient right. One of required standard was development of a unique patient identifier (UPI) to enable physicians, hospitals, and other authorized users to share clinical and administrative records more efficiently. UPI has too many features: reduce errors, improve interoperability, reduce the cost of marinating HER, and prevent privacy breaches. One advantage of a properly implemented UPI system is its freedom from errors through giving each person single and unique identifier that follows them throughout their lives and is used only for health records. One advantage of adapting UPI is separating between health record information and financial records information which is target for identity thieves and can improve privacy by limiting the transmission of more sensitive identifiers (names, address, and SSN). From the above advantages of adapting UPI, USA Department of Health and Human Services (DHHS) in 2005 has moved forward with steps to investigate of development of a UPI by linking patients records across different networks. The implementation of UPI is very costly and depends on several variables, including the architecture chosen to achieve connectivity between different ERH systems. To estimate the costs of implementing UPI, it would require a onetime investment and an annual maintenance cost. But before implementing UPI, there are a several assignment should be done besides money. Establishing a legal environment will be the best protection of patient privacy and encouraging the advances that interoperability would increase the health care quality and efficiency. The current situation in Kingdom Saudi Arabia (KAS), most of health care provider works as isolated island and adapt different EHR system. So, the patient can have more than on record in different EHR system, which reduce the quality of the provided service and significantly increase the risk for treatment process. The main goal of this work is to explore the adapting UPI in KSA and proposed a framework for UPI. The healthcare researcher and industry are shift violently in adapting the use of electronic health records (EHR) in medical filed. The major priority for any healthcare provider is providing a clear and high quality data to be sharable among different departments within the organization and that can be achieved though accurate Patient Identification. Because of the enormous impact that PI Integrity has on the clinical, financial, and administrative business of healthcare, it is imperative that the quality of an organization's identity integrity be addressed as a major priority within an organization and most certainly prior to sharing data externally with other stakeholders. Stakeholders should require quality data from fellow participants prior to participation in any data exchange. In development countries, health care fraud accounts for an estimated 3 to 10 percent of all health care costs, or 80 to 120 billion dollars of loss per year. Accurate identification and verification of identity is important also to reduce frauds due to medical identity theft [1].

II. RELATED WORK

Carpenter [1] mentioned that the department of Health and Human Services in 1973 reported that they are object to move forward toward "Standard Universal Identification". The proposed Universal Patient Identifier (UPI) should have the following features: uniqueness, verifiability, reliability, and tracking. The proposed UPI consists of 7 digit date code, 6 digit geographical code relate to the place of birth, 5 digit sequence code to identify born on the same date in the same geographical area, and one single check digit, which make the total size is 19 digits. For examples, 9930301^044273^00047^2 represent a person born in March 1, 1993 in Minneapolis, MN- USA. The proposed UPI can be used as Universal Provider Identification (UPI) by adding one digit refer to P(provider), or MD , or RN etc. The author assumed that the proposed UPI is reasonable and flexible and can be easily adapted using the available infrastructure. The proposed UPI will coded using base 34 digits bit base and check digit algorithm used to protect against miskey and digit inversion.

Universal Healthcare IDentifier (UHID) is the result of a 2 year standards development process by ASTM committee E31.12 on medical informatics during the summer of 1994. This research work [3] consists of selective quotes in italics of portions of the proposed standard. The author mentioned the main functions of using UHID, which are: positive identification of patients when clinical care is rendered, automated linkage of various computer based records on the same patient for the creation of lifelong electronic healthcare files; providing a mechanism to support data security for the protection of privileged clinical information, and enable the use of technology for patient records handling to keep health care operating costs at a minimum. The author mentioned the most important criteria for UHID, which are Atomic, Content-free, Cost-effective, Disidentifiable, Secured , Focused, Identifiable, Permanent, Unique, and Variable. The work proposed UHID schema structure, which starts by 16 digit Sequential Identifier (SI), a single character delimiter, 6 check digits, and 6 encryption Digits and the full identifier constitutes 29 digits (0000000123456789.012345000000). An evaluation for the proposed schema against the required standard criteria and it shows that the proposed scheme appears to adequately meet all but two of the criteria (cost and ability to "split") listed in the standard.

Kohane [3] mentioned that use of SSN is not safe and provided some article support his vision and he proposed a framework Health Information Identification and De-Identification Toolkit (HIIDIT). HIIDIT is not an identification system but a generator of identification systems and it take into consideration the following dimensions that are encompassed by HIIDIT : Directory local to determine the degree of patient consent in information (for example, 1 for Patients, 2 for Provider, 3 for Provider organization, 4 for Trusted escrow and third party, and 5 for Governmental authority), Scope of Identification to represent the geographical or organizational scope of the identification and the nature of the data linked to a particular identifier, Certifying Authority (CA) to certifies varying degrees of authority and credibility correspond to a particular patient, Scope of Identifier Secrecy to keep a patient identifier confidential and disclosed (for example, 1 for Just the patient, 2 for Patient & family, friends or guardians, 3 for Provider, 4 for Class of Providers, 5 for All providers, 6 for Healthcare institution, 7 for Insurer, 8 for Government, 9 for any combinations). The research work explained how HIIDIT system work and he claims that the HIIDIT's function matched and adequate of the required four dimensions of identification systems. Finally, he recommends using HIIDIT for sharing data between health care institutions that are competing in the market.

Integrated Advanced Information Management Systems (IAIMS) and Unified Medical Language System (UMLS) projects involved large amount of useful patient data, clinical information, and biomedical knowledge in electronic and it increased dramatically since the 1980s. Besty [2] stated that, in a 1998 the National Committee on Vital and Health Statistics (NCVHS) described three types of computer-based health records: patient, personal, and population health records are needed to facilitate coordination, research, and assessment for clinical care. Since, digital library term was introduced by National Science Foundation in 1994 and can be focus on information accessible via the Internet and encompasses. Since the digital library is not a single entity and it need technology to link different resources.

Nowadays, Identity is a key concept in the global world and the report stated that, "In 2000 the UNICEF has calculated that 50 million babies (41% of births worldwide) were not registered and thus without any identity document at all". The European Union tried to cover this gap through EURODAC system, which consists of a Central Unit equipped with a computerized central database for comparing the fingerprints of asylum applicants and a system for electronic data transmission between Member States and the database. EURODAC enables Member States to identify asylum□seekers by comparing fingerprints to determine whether an asylum□seeker or a foreign national found illegally present within a Member State [5]. In 2004 the European Commission has funded a project called Biometric Identification Technology Ethics (BITE) (www.biteproject.org) and the purpose of the BITE project was to provide a forum to initiate the public conversation on ethical and policy issues raised by the deployment and the application of biometric identification technology in various fields. BITE report defined the potential weak point of any biometric scheme, which is a liveness check (technological countermeasure to spoofing using artifacts). Latex finger, a prosthetic eye, a plaster hand, or DAT voice recordings are good examples for liveness checks.

In 2004, French government decided by law to start a national project for an electronic health record called the personal medical record (PMR). Ouantin [6] proposed this research work to establish and reassure French patients regarding the security of their medical data which will be stored at a national level through creation of a secure patient identifier. The author stated that hashing the social security number would help to meet the confidentiality of personal information contained in the PMR and provide access to patient or to public health bodies. Double hashing proposed to provide anonymity safely and a portal of the application from health professional will provide a reversible encryption coding HIN. The research proposed using of a smart card attributed to professionals in both the private sector and public hospitals. For the security of exchanges among health professionals, the author strongly recommends using of networks like virtual private network. For mobility and interoperability concern, the author suggested adapting Europeans Regulation (EC No 883/2004).

In [7], authors propose the fingerprint, iris, retina scan, and DNA (FIRD) framework that utilizes a patient's biometric characteristics to uniquely associate them to their medical data. The framework establishes an infrastructure that will distinctively identify a patient to his or her complete electronic healthcare record (EHCR) with exact precision and accuracy. The framework's inner workings collect records that are not properly assigned to the unique patient identifier (UPI), remove records that do not belong to the patient, and correct errors and omissions within the patient's EHCR. The authors suggested that creation of a standardized nationwide electronic healthcare record system in the United States would require a way to match a composite of an individual's recorded healthcare information to an identified individual patient out of approximately 300 million people to a 1:1 match, resulting in a final information compilation that provides a complete healthcare history to the healthcare provider, while reducing medical errors and lowering healthcare cost.

III. Approaches for Patient Identification

Usually, patients visiting healthcare providers identify themselves in person at the reception point and authenticate their identity by ways of picture ID, insurance card, doctor's name and/or appointment time. A patient, typically, may have many healthcare providers, including primary care physician, specialists, therapists and other medical practitioners. In addition, a patient may use multiple healthcare insurance companies for different types of insurances, such as dental, vision and so forth. Several visits for different healthcare providers result in patient's health information distributed among different healthcare providers in the form of disparate Electronic Medical Records (EMR). The above scenario raises a problem of how to integrate medical records belonging to the same patient from different healthcare providers that are disparate nationwide. What is actually needed is a national healthcare information network that allows authorized practitioners to collect and share health information about patients from different healthcare providers all over the country. One of the most challenging questions in this case is how would such system uniquely identify each patient and link him/her to composite medical records in one-to-one match.

Currently, each provider has its own centralized database of EMR and, typically, assigns unique record locators (often called medical-record numbers) to the records resulting from a patient's visits. Such record locators vary widely, from simple patient and family names to modified Social Security or insurance numbers, to provider-generated alphanumeric codes. Properly identified patients can approve the sharing of these medical records with other providers and insurers by signing an authorization form, clearly identifying the provider of record, the individual or entity to receive the record, and the boundaries or limitations on the information to be shared. The migration from traditional EMR systems to national healthcare information systems as described above involves three requirements: authenticating individuals, unambiguously linking individuals to their records, and authorizing controlled access to those records. Implementing these requirements creates new challenges, for example, face-to-face methods of identifying and authenticating patients, providers, or others logging onto a network no longer applies; methods of electronic identification and authentication are required. Likewise, knowing a patient's name or medical-record number from a single provider is not sufficient to unambiguously access that patient's records from other providers or a regional health information organization (RHIO); each entity may be using different numbering schemes or name constructions. Furthermore, demographic information either change over time such as address or are not unique such as SSN and names; the larger the network, the more likely it is that more than one person will have the same name and other demographic data. Finally, compromised data integrity, widespread unauthorized distribution, and other network security attacks are very common for the national health network, new security measures are needed [*8*]. IT proponents assure us that these challenges can be overcome, but doing so demands new solutions. This paper focused mostly on one component of these new challenges: defining the best electronic patient-identifier system for the purpose of sharing personal health information through a national health network which will improve the privacy and efficiency of the health care system and the quality of healthcare itself. There are these two approaches to accomplish this task (1) statistical matching and (2) Unique Patient Identifier (UPI) [*8*]. We will discuss each of these two approaches and the advantages and disadvantages of each one.

A) *Statistical Matching :*Statistical matching attempts to integrate enough information about an individual to form a unique key used to locate his/her electronic health record. It strings attributes such as: last name, first name, date of birth (DOB), phone number, address, zip code, and gender. It may also use medical record numbers and all or parts of social security number (SSN). The problem in such key is that some attributes, such as name, DOB, and zip code, are not unique to the individual; others, such as address, may change overtime. As the database of records gets larger, more personal attributes must be added to keep the key unique. A nearly unique and relatively stable attribute, such as SSN, patient identity, and healthcare provider name, may be used to reduce ambiguity in large databases. The difficulty to distinguish between first and last names, the usage of different format, and data entry errors, such as misspellings and number transposition, may also cause ambiguities in linking patient to their records. Searching algorithms used in this approach vary from requiring an exact match on a specific set of attributes or to more advanced probabilistic pattern matching. The development of statistical matching depends on human to clarify questions and reduce ambiguity this is called disambiguation. Advanced algorithms preprocess the health-records database to determine the frequency of every attribute and score the match according to the discriminating ability of the specific attributes of that database. For example, a match of the name Smith typically would not score nearly as well as a match of a less-common name. The scores can be used with threshold values of acceptance and rejection, as well as with regions of possible matches that can be adjudicated by humans. However, setting the acceptance and rejection limits higher or lower affects false positive, false negative, and indeterminate results. Minimizing one type of error comes at the cost of increasing other types of error.

B) Unique Patient Identifier :Unique patient identification is a method for linking patients to their electronic medical records that are exist globally in a domain (state, country, region, or world). Unique Patient Identifier (UPI) is a unique, non changing alphanumeric key for each patient that associated with every health record belonging to that patient. Finding the patient's records anywhere within the healthcare system is then a matter of verifying that the patient is the person owning the key (authentication) and asking each healthcare system or provider in the domain whether it has information associated with that key [8]. The American Society for Testing and Materials (ASTM, 2000) Standard Guide lists desirable attributes of a UPI, including that it be:

- **Unique:** Each UPI is associated with only one person; different individuals can not share the same UPI; this attribute permits the collection and aggregation of health information into one complete medical record.
- **Non disclosing:** This means the UPI should not contain any personal information such as name, address or mobile number. This attribute aims to prevent revealing patient confidential information or data inquiry. The combination of selected personal attributes used in statistical matching violates this attribute.
- **Invariable:** The UPI should not change in the person's lifetime (except in case of identity theft or similar problem). This attribute solve the main problem in statistical matching which is the changes in some of the personal attributes, such as name and address, making it difficult to find previous records.
- **Canonical:** Each individual should have only one UPI. Multiple UPIs have actually been proposed as a means of giving a patient control of disclosure, but they can also lead to fragmentation of the individual's healthcare data.
- **Verifiable:** This aims to validate of the UPI and is done generally through the use of additional check digits—numbers that must match some mathematical combination of the UPI's remaining digits without additional information. Verifiability helps to prevent input errors that exist in statistical matching method.
- **Ubiquitous:** Every patient should have one. This is difficult to achieve, particularly if participation is voluntary, but the alternative is a hybrid system, in which some patient data cannot be found using a UPI.

IV. Errors in Linking to Medical Records

There are two types of errors in statistical matching: false positives, in which there is a link to the wrong patient's records, and false negatives, in which not all of a patient's records are found. Figure 1, which is adopted from [8], shows a representation of these types of errors. The horizontal scale shows the score of a particular match. As more and more attributes match and as the match is weighted by its score, or value, the higher is the probability that the patient is correctly matched to that record. A low score indicates a low probability of match (and a high probability that it does not match). It is possible to use a threshold above which the record is assumed to match and below which it is not assumed to match, which leads to the shaded areas above and below the threshold. The area shaded to the right of the threshold is the region corresponding to false positives, or picking up the wrong patient's records. The shaded area to the left of the threshold is the region of false negatives, or the records of the patient that are not picked up because of some non matching personal attributes. Another approach illustrated in this figure is to define a region of ambiguity within which possible matches are tagged for human resolution or disambiguation [8].

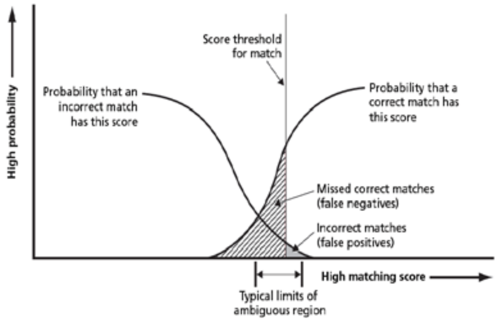


Fig. 1: False positive and false negative errors [8]

A) *False-Positive Errors*

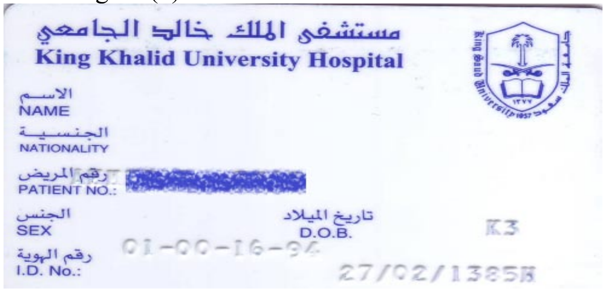
Linking to the wrong health information about a patient can cause wrong treatment based on wrong condition, perform wrong operation, serve wrong patient, mistakes on blood types, errors in lab test, or wrong medications and diagnosis. This kind of error is the result of healthcare ID theft, accidental record overlay (more than one distinct individual assigned to the same record), a threshold set too low, or a set of personal attributes used in the search that, in combination, are inadequately unique for the size and nature of the population being examined [8]. An important cause of false positives is the use of an insufficient number of attributes in a search for matches. In [8] an experiment was conducted to illustrate this problem. In this experiment, a large personal-attribute database of 80 million individuals, similar in size to a large RHIO or state-sized records database was used to evaluate false positive errors. First, a 42,000-record subset of this database is used, similar to the size of a small hospital or large clinic. For a random individual, there would be about a 2-in-3 chance (1/1.44) of finding another person's record with the same last name. However, if first name, birth year, and zip code are added, the number of possible false matches is reduced to only one in 3,500 (1/3.5E3). The use of a unique part of the SSN in the stream of keys quickly reduces the probability of a false match to near zero. This, of course, assumes that the keys for matching are entered correctly. When using the larger database of 80 million records, it is a bit more difficult to eliminate false positives. There would be a 98% chance that a false-positive match would occur with just the last name compared to roughly 66% for the small-population analysis in small database, this shows how the false-positive rate is sensitive to factors such as population size. When date of birth is added to the key, the chance of a false positive match drops to 33%. And, finally, after the last four digits of the SSN, the first name, and the zip code have been used to form the composite key, the rate of false positives drops to 1 in 39 million. In conclusion, with enough correct personal-attribute keys, the false positives can be controlled to occur with very low probability. However, eliminating the almost-unique SSN key dramatically increases the false-positive rate. If the database gets much larger, as in an NHIN, additional attributes or some, almost unique, key, such as the SSN, is certainly required to keep this error rate small. If the use of an SSN as a key is ruled out, as it increasingly appears to be in many applications, ensuring a low rate of false-positive errors becomes quite difficult in such large databases and UPI became an insisting need to reduce the error.

B) *False-Negative Errors*

False negative errors imply not finding some of the patient records. They represent a fragmentation of a patient's health history and can lead to missing or incomplete information about medical conditions, previous surgeries, medications, or allergies, which in turn lead to possible life-threatening treatment errors and potential lawsuits. Missing information can also lead to inefficiencies, such as the cost of reordering of diagnostic tests and of delays and errors in treatment. Such inefficiencies have been estimated to cost the healthcare system more than $8 billion annually [3]. It is also much more difficult to analyze patient data for research or clinical quality and process improvement when some of the patient data are not found because of such fragmentation. False negatives may be the result of changing personal attributes, such as name or address; of keying errors; and of changes in format, such as the order of first and last names. All of these situations can cause the recording of some of the patient's data as new records, effectively fragmenting potentially important health information. Another false negative problem is record duplication – records are found that falsely appear to be those of another patient when in fact they should be identified as belonging to the reference patient.

V. HEALTHCARE SYSTEM IN KSA

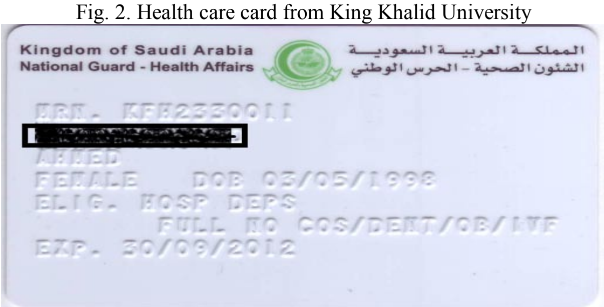
According to the Ministry of Health in Saudi Arabia, the healthcare system consist of a network of primary healthcare centers and clinics that provide basic and advanced services with some mobile clinics for remote rural areas. The Ministry of Health operates most of the hospitals and the clinics and centers. While the reset remaining facilities are operated by government agencies, including the Ministry of Defense, the National Guard, the Ministry of the Interior, and several other ministries. Some researcher classifies Saudi health care system as a mixture of the American health care system and Canadian system. On other words, there are free government hospitals like in Canada and they have private hospitals for insured and cash paying patients with instant care like America. Since health care is free for Saudi's, the Saudi government forces the companies to provide health insurance for its employees and their families. The quality of health care in Saudi generally can be classified as high and equal to that in some Europe countries, except for highly specialized treatment. In the following section, we will explore a sample case study that demonstrates how much diversity among the current health care provider in KSA. For example, King Saud University delivered health care services through two large University Hospitals; King Abdul-Aziz University Hospital (KAUH) and King Khalid University Hospital (KKUH) in conjunction with two big clinic centers. Both hospitals and clinics provide primary and secondary care services for Saudi patients from Northern Riyadh area with free of including some medications [KSU website]. The current running medical information in the both hospitals and clinics centers used a sequential assignment number for any new patient as shown in figure (2).



مستشفى الملك خالد الجامعي
King Khalid University Hospital

الاسـم
NAME
الجنسـيـة
NATIONALITY
رقم المريض
PATIENT NO.: [redacted]
الجنس
SEX
تاريخ الميلاد
D.O.B.
K3
رقم الهوية
I.D. No.: 01-00-16-94
27/02/1385H

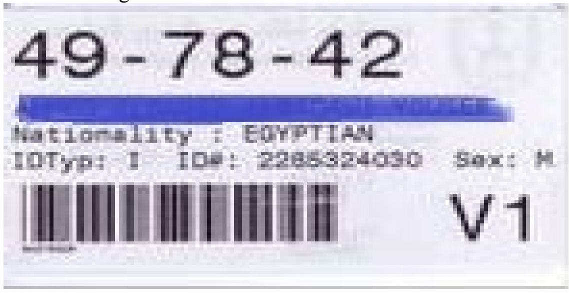
Fig. 2. Health care card from King Khalid University



Kingdom of Saudi Arabia
National Guard - Health Affairs
المملكـــة العربيـــة السعوديـــة
الشئون الصحية – الحرس الوطني

MRN. KFH2330011
[redacted]
FEMALE DOB 03/05/1998
ELIG. HOSP DEPS
FULL NO COS/DENT/OB/IVF
EXP. 30/09/2012

Fig.3. Health care card from National Guard



49-78-42
[redacted]
Nationality : EGYPTIAN
IDTyp: I ID#: 2285324030 Sex: M
V1

Fig.4. Health care card from King Saud University

As shown in figure (2), there is NO patient number but the name of the patient and identification number is given by adding the last patient ID number with 1. Another sever problem is some patient can have been treated in either/both hospital(s) and/or some clinic center, that patient can have a different ID number in all different location. Therefore, there is NO way for physician to electronically access the patient information expect through printed report carried by the patient or by his family. The same situation or close exists in the second big hospital in Riyadh, National Guard-Health affairs. Figure (2) shows the patient ID for National Guard hospital at Riyadh branch. Since National Guard has many hospitals scattered around big cities in KSA, the patient identification consist of the first three letters of hospital name and the remaining is numeric number represent the sequential number given by the medical information system as shown in figure (3). Figure (4) shows the patient ID for King Saud University and the patient identification consist of 10 digits that represent the sequence number and a bar code that contain all the patient information such as nationality, gender, incurrence class, and a file number.

VI. Framework for UPI in KSA

With the differences and specific nature of the proposed UPI system, we developed the UPI system process framework. An overview of the UPI process framework is visualized in figure (5). Our framework copes with different issues raised from related work section. The UPI process framework can be seen as three managerial levels; the strategic, tactical and operational level. For each level, processes are designed consisting of relevant activities and the relations between activities and the data produced in the activities can be achieved through SOA web service.

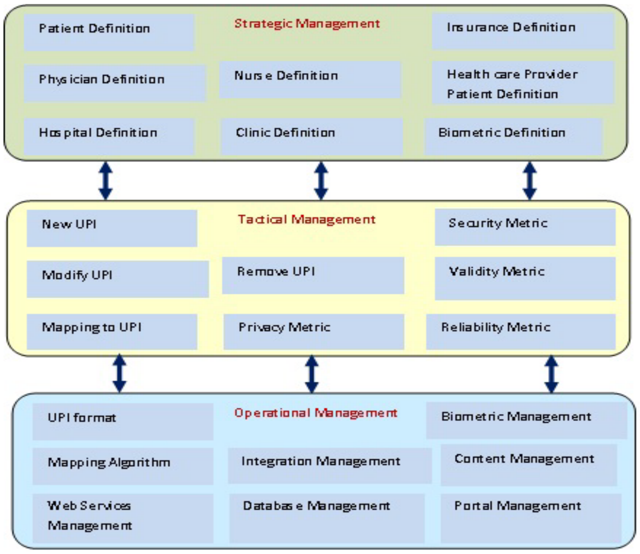


Fig. 5. UPI Process Framework

The main features of the required portal management system should contains the following features : Front End and Back End for end-user and administration management, Configuration Settings for website control, Access Rights for providing hierarchy authorities, Content for content management, Templates for providing an editable visual format of the content, Extensions for future growth and changing requirements of functionality, Multilingual front end, Simple workflow system, and Administration interface that is separated from the portal homepage. Figure (6) shows a schema for the proposed solution, the schema can seen as integration among different medical information system with portal that comply with web medical content management system. The proposed portal will store and update the UPI through a secure database system for further search and to keep track of UPI usage.

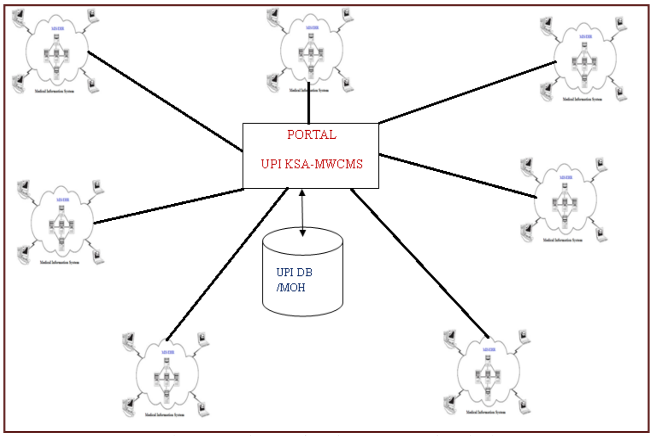

PORTAL
UPI KSA-MWCMS
UPI DB
/MOH
Medical Information System
Medical Information System
Medical Information System
Medical Information System
Medical Information System
Medical Information System

Fig. 6. Schema for the proposed Solution.

VII. Conclusions and Future Work

Faced with many challenges of existing architectures, a growing number of organizations have taken on a private cloud approach, using server virtualization to simulate on-demand services. This hybrid approach or “cloud-like” solutions can help alleviate some of this performance, security, and other challenges but at a significant cost, time, and resource expenditure. The other important aspect of this cloud computing alternative is reviewing the cultural impact of moving data and clinical applications to the cloud. Like businesses in other industries, there is a natural predisposition for physician practices and healthcare organizations wanting to “own” and have physical control over their data. Securing applications in the cloud is limited due to the difficulty in guaranteeing effective data security and integrity controls. In a traditional environment, the ability to layer stronger authentication, access control, and auditing capabilities exists because of defined network layers. By contrast, these defined network layers don’t exist in a public cloud environment. Data restoration presents another limitation as restoring data from a backup (determining what needs to be restored, from where and deposited to) can be challenging.

References

[1] Paul Carpenter and Christopher Chute, “The Universal Patient Identifier- A Discussion and Proposal”, Proc Annu Symp Comput Appl Med Care. 1993: 49–53.

[2] B. R. Hieb, “A proposal for a national health care identifier”, Proc Annu Symp Comput Appl Med Care. 1994: 469–472. PMCID: PMC2247838.

[3] . Kohane, H. Dong, and P. Szolovits, “ Health information identification and de-identification toolkit.”, Proc AMIA Symp. 1998: 356–360. PMCID: PMC2232117.

[4] Betsy L. Humphreys, “Electronic Health Record Meets Digital Library: A New Environment for Achieving an Old Goal.”, J Am Med Inform Assoc. 2000 Sep–Oct; 7(5): 444–452. PMCID: PMC79039.

[5] Emilio Mordini, MD, DPhil, “Biometric Identification Technology Ethics (BITE), FINAL SCIENTIFIC REPORT”, Centre for Science, Society and Citizenship Piazza Capo di Ferro 23 – 00186 Rome IT, February 2007.

[6] Catherine Quantin, Franc¸ois-Andr´e Allaert, Paul Avillach, Maniane Fassa, Benoˆıt Riandey, Gilles Trouessin, and Olivier Cohen, “Building Application-Related Patient Identifiers: What Solution for a European Country?”, International Journal of Telemedicine and Applications Volume 2008, Article ID 678302, 5 pages, doi:10.1155/2008/678302.

[7] D.C. Leonard, Alex P. Pons, and Shihab Asfour, “Realization of a Universal Patient Identifier for Electronic Medical Records Through Biometric Technology”, IEEE transaction on information technology in biomedicine, vol. 13, no 4, July 2009.

[8] RAND Corporation, Identity Crisis: An Examination of the Costs and Benefits of a Unique Patient Identifier for the U.S. Health care System, 2008. RAND Health, www.rand.org.

Intelligent Advising System

Ahmed Emam
Dept. of Information Systems
College of Computers and Information Sciences
King Saud University, Riyadh, KSA
aemam@ksu.edu.sa

Abstract— The primary and necessary task for any undergraduate student has to go through after his orientation is advising. Advising plays a key role in a student's success towards his degree program. Many reports were published recently on the survey of the declining college graduate students [2]. There might be several reasons behind this, but the major drawback of this decline is due to the improper advising of the students. Noticing the importance of the advising system, an Intelligent Advising System is designed which helps the students prepare their degree plan by selecting their major and minor/s. This paper mainly focuses on the importance of the advising system and how the IAS helps the students in online advising.

Key words: Advising System, Database, and Data Mining.

I. INTRODUCTION

Students are the building blocks of tomorrow's world. Every student has a wide range of ideas in which he wants to see his future career path. So, it's the responsibility of the advisor to provide with the right information in right time. Choosing a major, minor and the program that a student wishes to graduate is difficult task. Planning the courses as per their schedule for the entire period and meeting the degree and college requirements needs a high level thinking. So, Advising and tracking the success of a student throughout his entire degree program plays an important role towards the success of his program.

The National Academic Advising Association (NACADA) recommends the following goals for academic advising: development of suitable educational plans, clarification of career and life goals, selection of appropriate courses and other educational experiences, interpretation of institutional requirements, enhancement of student awareness about educational resources available such as internships and learning assistance programs, evaluation of student progress toward established goals and development of decision making skills with reinforcement of student self direction. With the high enrollment of students and ratio of the student's and advisor is increasing at a faster rate, and with the advanced technologies in the software technologies, an Online Advising System came into picture which can help the student to verify his track towards his degree program for the duration of his undergraduate program. The following are the main difference between academic ADdvisor (AD) and Intelligent Advising System (IAS): IAS is accessible most of the time while AD needs to fixes some time such as office hours or provide a means for appointments, IAS Keep in contact with your students, while AD wait for students to come to contact him, IAS gives much knowledge and information about academic requirements in different levels: university, college, and department with rationale explanation for these requirements, while AD has a limited memory capacity to memorize all details and reasons. IAS will assist students with making short-term and long-term plans using a background available knowledge and apply data mining technique while AD can create a long term plan that consistent with student interests only not consider the previous cases similar to them, finally, IAS can monitor students' progress toward educational goals by up-to-date information while AD need to access different records to keep track of the student record. The main features of IAS are : Displays the degree plan for the selected major/minor for the entire 4 years, Easy selection of the courses or the degree options from the drop down menu while preparing a degree plan, Easy selection of general education courses as per the CLEP exam selection, Emailing Advisor about the degree plan through the system, Allow students to change his major, minor, concentrations and general education requirements, Allow students/Advisor to change the courses in the degree plan at any point of time, students can update the profile with the data from backend database system such as TOPNET by click only one button, Advisors can save the degree plans of the previously graduated students and use as background data for data mining, students can select the degree plan of the previous graduated students as per the ability and interest, finally it is very easy for administrator to assign the advisors to the students or add/update the college, degree, major and minors into the system.

IAS came after limitation of iCAP such as : cannot allow students to change majors, minors, or concentrations, cannot check for pre-requisites or co-requisites, cannot automatically apply transfer coursework without an equivalency, cannot allow courses with "0" hours to fulfill requirements (Labs, etc.), show completed course work including grades and hours earned, and allow students to run "what if" audits to compare course work against other majors.

II- LITERATURE REVIEW AND EXISTING SYSTEM

Research on college students suggests that activities like advising could increase students' involvement in their college experiences. Many of the universities currently have different flavors of online advising system with different functionality. Most of the available advising system are offline system and here are a few available online advising systems chosen out of many available online.

1. Electronic Advising System (EAS): Department of Electrical and Computer Engineering, University of Colorado
2. Course Planning Consultant (CPC): Department of General Engineering, University of Illinois at Urbana Champaign
3. Student Advising System : Department of Plant Science, California State University.

Currently there is no online advising system for the undergraduates students at KSU. Every student needs to see the advisor two to three times a semester which has scheduling conflicts. The Online Advising System cannot act as a replacement for an Academic ADdvisor but it only helps the advisor and student to make the task easy upon each other.

III- IAS ARCHITECTURE

With the rapid and advanced web technologies in the software world, Microsoft .NET Framework was chosen as a platform on which the IAS was built on. SQL Server which is very well compatible with the .NET Framework was chosen as a backend data base. For the application to interact with the web, Internet Information Services (IIS) is chosen as a web server for the IAS system. IAS was built in a three tier architecture model taking into the consideration of the Web enabled advising system, which has a .NET framework as a middle tier which acts as an application server, web tier is where the GUI part of the application where the user's can view/access the application functionality. Below is the high level architecture of the IAS system. All three tiers are shown below. The firewall between the web server and the outside world prevents any un-authorized access to the network. While some pages which are at the application server level can also be accessed directly by a smart user, so in-order to avoid the direct access to application server, usually a fire wall between the application server and the web server can prevent any un-authorized access to the application server.

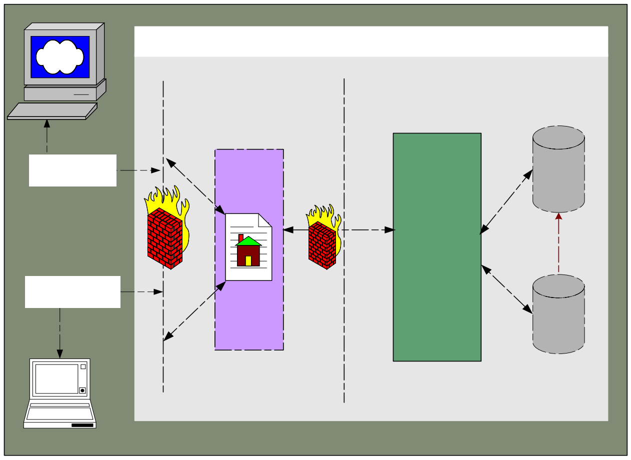


Fig. 1. *IAS system Architecture*

IV- PROPOSED FRAMEWORK

Keeping in mind the requirements for the current undergraduate students and the advisors at KSU, IAS was designed in such a way that it can be adaptable for any other department, when the requirements for a major and minor has been added to the system. Currently the data and the requirements were designed for the CS students at KSU. The Administrator of the IAS system can have a capability of changing the courses and their requirements. An administrator has full control on the system where he can add advisors to the system. An administrator can perform all the functions performed by the advisors and students. An administrator can run the reports on the advisors and students list. Role-based authentication available in the .Net frame work is used as the authentication mechanism. Once the user is authorized into the system based on the role assigned to the user, the system's functionality is divided as per the role assigned to the user. Based on the data flow the IAS system is divided into specific modules:

1) Registration module
2) Authorization module
3) Create degree plan Module
4) Modify degree plan Module
5) Update profile Module
6) Report Module
7) Course Maintenance Module

Registration module: administrator will be able to create the user id and passwords as an initial setup, and the users will be prompted with their registration page on their first time visit to the system upon proper authorization.

Authorization Module: all the users who are accessing the IAS system needs to get authorized prior to entering the system. The users will be authorized based on their assigned role (administrator, advisor, user) and are directed to the appropriate pages as per their roles.

Create a Degree Plan: when the users first create a degree plan they were asked to select a major, minor and courses for the General Education sections. Once the major and minor were selected, system will be displaying the list of core and optional courses, where the student will be selecting the optional courses list. Once all the requirements were met the system is going to display the degree plan of the student which is available to save and print the plan.

Modify degree plan: - users can pull the saved degree plan at any time and modify the existing plan and save the necessary changes. The system will be checking for all the constraints when changed before saving the plan.

Update Profile: - Users can pull their profile at any time after logging into the system and be able to update their own profile and further save the new changes.

Report Module: - Accessible by all the users of the system where users can print their degree plan. Also, administrators can print the list of users and advisors and also students grouped by the advisors. Advisors can print the list of student under his advising.

Course Maintenance Module: - Administrators can only have access to this module, where administrators can grouped by the advisors. Advisors can print the list of student under his advising.

Course Maintenance Module: - Administrators can only have access to this module, where administrators can manage the courses, manage majors, minors, and General Education requirements. Administrators can also manage at the high level by adding the colleges, departments etc.

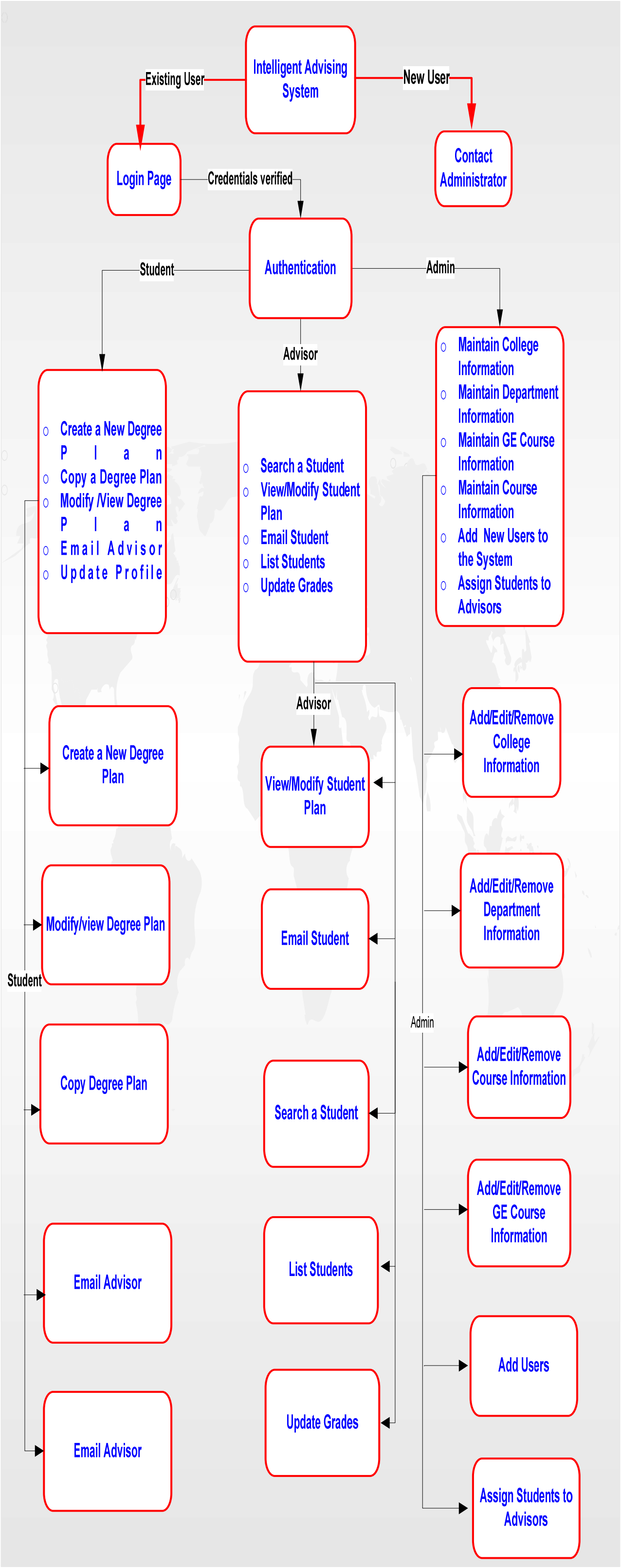


Fig. 2. IAS system Modules and functions

V- RESULTS AND DISCUSSION

Internet has become our daily part of life, where we are able to access any kind of information over WWW. Due to the high availability of the internet most of the applications were made web-enabled so that anybody can access them from anywhere in the world. So, IAS system was chosen to build as a WEB based system. Since Microsoft technologies are very easy to learn and easy to deploy, with the software being well familiar to most of the IS graduates, Microsoft Technologies were chosen for building the IAS system. Internet Information Services (IIS) Is chosen as a web server for the IAS system. This serves all the requests coming from the users, and sends the requests to the Data base server are needed. Web Server acts as a liaison between the user and the application. SQL Server 2005 is chosen as a data base server for the IAS System. .NET Framework which has a CLR (Common Language Runtime) compiles the code and acts as a liaison between the Web Server and the data base server. C# was chosen as a programming language for writing the ASP.NE web pages. The Advanced IAS system has the reports capability where the students and the advisors can print the reports of the degree plan.

VI- FUTURE WORK

The IAS system can be further extended by having the system integrated with the TOPNET where we can have a common updated database about the courses and the program requirements. The system can also be integrated with the existing ICAP system so, that we have all the functionality encapsulated into one system.

REFERENCES

[1] Enabling Technologies for Effective Student Advising. http://claymore.engineer.gvsu.edu/~steriana/Publications/ASterian.ASEE2003.pdf.

[2] http://www.detnews.com/apps/pbcs.dll/article?AID=/20060104/SCHOOLS/601040322/1026

[3] A decision support system for academic advising http://portal.acm.org/citation.cfm?id=315897

[4] Preliminary Implementation of a Web based Automated Student Advising System, http://www.actapress.com/PaperInfo.aspx?PaperID=22566

[5] Seven Principles for Good Practice in Undergraduate Education http://www.csueastbay.edu/wasc/pdfs/End%20Note.pdf

[6] https://ece.colorado.edu/~griff/cgi-bin/main.pl

[7] http://courses.webtools.uiuc.edu/cis/schedule/urbana/2006/Spring/CS/index.html

[8] http://cast.csufresno.edu/PlantSci/Advising/