Improving Business Processes Management Systems by introducing Soft Computing Elements

Francisco P. Romero¹, Arturo Peralta², Jose A. Olivas¹, Jesus Serrano-Guerrero¹ and Sergio Moreno²
¹Dept. of Information Technologies and Systems, University of Castilla La Mancha, Ciudad Real, Spain
²Docpath Document Solutions SL, Madrid, Spain

Abstract—Business document management systems may provide a lot of useful features in a wide range of document storage and retrieval situations. However, the provision of search systems only based on static indexes will not meet the requirements of the users. Regarding to this background, smarter business management systems are needed with adaptive capabilities for flexible information retrieval, semantic annotation, etc.

In this work, different soft computing techniques such as fuzzy logic have been applied in order to enhance information retrieval capabilities in business document management systems. The core of the proposed system is based on the application of natural language processing, machine learning algorithms and fuzzy linguistic modeling to semantic annotate business documents and to improve search engine results.

A prototype was implemented and successfully tested in several study cases (invoices, reports, medical records, etc.)

Keywords: document management systems, soft computing, natural language processing, fuzzy logic, document categorization

1. Introduction

Business documents range from brief accounting documents (invoices, bills of entry, transport documents, etc.), to complex legal agreements, circular letters, statistical reports, and so on. These kind of documents are used extensively by professionals in their execution of their own work and to share information with others [1]. Moreover, a company uses documents to communicate, transact business and analyze its productivity, thus, ineffective information retrieval leads to a lack of productivity because it contributes lost work time and duplication of effort [2].

Moreover, in today’s business environment, the volume of information is increasing at an unstoppable and constant rate [3]. The storage of an increasing number of business documents and their proper management has become a real challenge for businesses.

A document management system (DMS) can be defined as a computer system that is used to store, track, and retrieve electronic documents [4]. Aware of the above mentioned challenges, the current document management systems tools allows storage of documents and web-based queries.

These document management systems can be used for both individual paper management and business process management [5]. For example, an official document management system is implemented in [6], the proposed system is more oriented to a framework to model business process management than an information retrieval system.

Important causes of the business documents retrieval problem are an incorrect automatic understanding of the information need and of the content of business document texts, and a poor strategy for matching query and document content [7]. In order to solve this problem, a semantic information retrieval mechanism to enable the user to obtain required information with semantics-based query is proposed in [8].

On the other hand, current technological possibilities allow management systems to exploit these texts by using some well known text analysis software frameworks like Unstructured Information Management Architecture (UMIA)¹ and General Architecture for Text Engineering (GATE)².

In the work of [9] the goal is to identify and annotate structural parts of regulation documents by using the UMIA framework.

The goal of this work is to enhance information retrieval performance of a business DMS according to the user’s view. The proposed approach incorporates in a DMS an array of advanced features that go far beyond mere storage, resulting in a semantic annotation of the documents and a deep analysis of the information contents extracted from the documents.

For this purpose, it required the integration of innovative soft computing techniques and well established natural language processing algorithms to enable an efficient content-based retrieval. These tools allow us a highly advanced document management software for optimal document management.

The rest of the paper is organized as follows: The proposed model based on Natural Language Processing and Soft Computing techniques is explained in section 2. Section 3 contains the description of the preliminary experiments carried out. Finally, section 4 presents some conclusions derived from the experienced.

¹http://uima.apache.org/
²http://gate.ac.uk
2. Proposal

A DMS provide an initial basis for collaboration by organizing your documents and implementing different tools to optimize the flow of documents between employees. This kind of tools has some key functionalities such as the registry and classification of documents with versioning, the forwarding of documents with alarms and the procedural organization of documents [10].

A DMS includes two main components: the Data Storage System, a fundamental element in the infrastructure layer, integrated and used by other company applications[11]; and the Content Management System, that helps to mask the complexity of underlying data management tasks and provides an efficient platform for users to contribute, share and retrieve particular content [12].

2.1 DMS Workflow

One of the main problems that DMS have to deal in an electronic context is the issue of workflow. Our tool users to access organization documents and interact in a simple way. This means that users are able to retrieve and organize documents. The basic workflow can be seen in figure 1.

Different business applications can produce documents according to the business context in a standardized form (invoices, contracts, ...), often in different languages. In this case the users cannot decide whether or not to save a document into the DMS but a batch process loaded these documents into the repository. In this process, the documents are annotated using the most relevant concepts extracted from their content by another batch process (the semantic annotation process 2.2). On the other hand, a DMS user can load individually documents into the repository. The content is annotated on-line in the same procedure and, moreover, the user receive a recommendation (see Section 2.5) about how to classify the document in the DMS System (folder, type of document, etc). This recommendation is very important in order to keep the repository structure consistent. Moreover, this system includes semantic and fuzzy search capabilities to effectively offset existing defects of the traditional keyword-based search.

2.2 Semantic Annotation Process

The proposal presented in this section includes several phases of data processing to automatically extract relevant concepts from the selected document set, as can be seen in figure 2.

The Semantic Annotation Process is a batch process in order to transform and analyze the documents stored in the DMS. This process can process the new documents, the document which previously fail, and the documents under certain conditions established by the system administrator. The aim is to extract a set of relevant concepts that represent a brief summary of the document content.

2.2.1 Linguistic Preprocessing

In this step, the document set is preprocessed in order to characterize texts by their topically significant words. The goal is to extract textual information in the form of individual words from the selected documents. The process begins with a document transformation to deal with the different format types (.doc, .txt, .pdf, xml, etc.). Then, all non-textual information like digits, dates and punctuation marks, is removed from the documents (lexical analysis). Plus, some relevant structural information is extracted (titles, bold and italic words, etc.). Next, collocations were extracted according to the method described in [13]. After that, the system carried out a part-of-speech (POS) tagging process using the Apache Open NLP tool\(^3\). The POS tagging is the task of labeling each word or token in a sentence with its appropriate syntactic category called part of speech (e.g., adjective, noun, verb, etc.) [14].

Finally, three techniques are used to reduce the vocabulary and make the representation of texts more meaningful: stop words [15], stemming [16] and zipf law, therefore, words that provide low semantic meaning for the domain context are eliminated. Language detection and spelling correction processes are also included in this step.

\(^3\)https://opennlp.apache.org/
2.2.2 Top Concepts Extraction and Storage

The main purpose in this step is to extract a set of chosen concepts which will help to generate a semantic representation of the document content.

In this research work, we used a concept weighting measures to quantify the value or usefulness of a concept in a document and, thus, determine the so-called relevance weight of a concept that was used to select it as “top concept”. The concept weighting measure used is an extension of the classic tf-idf, which we call absolute weight \( (aw) \):

\[
aw_i = \frac{k * w_{ij}}{|j| * idf^D}
\] (1)

The absolute weight of a concept \( i \) \( (aw_i) \) is calculated mainly using its weight on the document \( (w_{ij}) \). Each concept has a weight on each document according to FIS-CRM Model [17]. The fundamental basis of FIS-CRM is to “share” the occurrences of a contained word among the fuzzy synonyms that represent the same concept, and to “give” a fuzzy weight to the words that represent a more general concept that the contained one. In this way, a word may have a fuzzy weight in the new vector even if it is not contained in it, as long as the referenced conceptunderlies the document. As the total number of concepts in a document \( (|j|) \) is generally much bigger than a single concept weight, the formula gives a very low-value result. To avoid this problem, a factor \( k \) is incorporated to make the result more significant. The value of this factor depends of three parameters: if the concept occurs in a title, if the concept is a collocation, and if the selected word is bold or italic or not. On the other hand, in order to consider the concept discriminative power within a document, it is employed the \( idf^D \) factor (Inverse Document Frequency in a Domain), which can be calculated based on the number of times the concept occurs within a domain corpus as a whole. The domain corpus is build from the number of times the concept occurs within a domain (Frequency in a Domain), which can be calculated based on the documents (described in the previous section).

Query techniques such as query expansion can be very interesting in order to retrieve relevant documents even if they do not contain the words used in the original. This is because the ambiguity of natural language and also the difficulty in using a single term to represent an information concept [18].

The proposed procedure of query expansion in the DMS contains the following steps:

1) **Selection of new terms**: Once the user sends a query to the system, the first step is to build a set of additional queries that complement the original one. In these queries new terms semantically related to those of the original one are included. This process is guided through the set of top concepts previously extracted from the documents (described in the previous section).

2) **Word sense disambiguation**: The main handicap of the search process is managing weak words (words with several meanings) [19]. So, sense disambiguation of this kind of words is fundamental for carrying out this process successfully. The disambiguation of a weak term consists on identifying its right semantic area. Automatic disambiguation methods are suggested by papers such as [20]. In this work, the disambiguation method proposes in [21] is used. This method can be applied when some semantic relation between the terms of the query can be determined. To be able to apply it, it is necessary to have an ontology or thesaurus that allows discovering that relation. An ontology automatically built from the domain templates previously obtained in the semantic annotation process is used for this purpose.

3) **Weighting of new terms**: All terms in the expanded query should have equal weighting or whether the new terms should have a higher/lower weighting. Voorhees [22] found that assigning lower weights to added concepts enhances retrieval accuracy. She used a factor between 0 and 1 for weighting added terms. In this step the approach described in [23] is used with the aim of generating a set of queries/concepts whose meaning is very similar to the original query and which can help to maximize the search process. Each new query can be submitted as an individual request and the final results retrieved by the queries could be merged in order to build a final list of results.

4) **Ranking and filtering**: The relevancy of each object is calculated to determine the order in which they will appear. A ranking of the new items is obtained according the similarity between the expanded query and the original query, the position of the item in the results, and the sum of relevance in all queries (chorus effect).

---

4http://office.microsoft.com/es-es/templates
2.4 Fuzzy Search

Our purpose is to include fuzzy search capabilities in the DMS without modifying its technical architecture. The use of the proposed fuzzy techniques allow the user to write flexible conditions over numeric attributes based on linguistic labels instead of precise (crisp) values.

In our system, an attribute is capable of fuzzy treatment if it is in an ordered underlined fuzzy domain[24], in this case, linguistic labels can be defined on it. Therefore, an attribute can be seen as a linguistic variable. A linguistic variable is a variable whose values are linguistic words or sentences in a natural language[25] and can be divided into various linguistic values (for example: high, low, medium).

Triangular and trapezoidal membership functions are used for each linguistic value defined in each quantitative attribute for simplicity. Hence, each linguistic value is a fuzzy number, which is a fuzzy subset in the universe of discourse that is both convex and normal [26].

The number of fuzzy sets refers to the number of partitions that should be done in each ordered continuous domain, i.e., the number of labels that should be used to describe the continuous numeric variable. This number is manually established by the system administrator when the document type is introduced in the DMS. There are an automatic mapping between this number and the labels, for example, if three fuzzy sets are defined, the values are mapped to “high”, “medium” and “low” labels.

The system allows to automatically define the fuzzy sets on each continuous domain based on trapezoidal functions and the maximum and the minimum. In a first approach, the attribute \( A \) and its membership functions \( \mu^A_{K,i_m}(x) \) can be represented as follows [24]:

\[
\mu^A_{K,i_m}(x) = \max\{1 - \left| \frac{x - a^K}{b^K} \right|, 0\} \tag{2}
\]

where:

\[
a^K = m + (m - m) \times \frac{i_m - 1}{K - 1} \tag{3}
\]

\[
b^K = \frac{m - m}{K - 1} \tag{4}
\]

where \( K \) is the predefined number of linguistic values in a linguistic variable, \( i_m \) (\( 1 \leq i_m \leq K \)) is the \( i_m \)th linguistic value of \( K \) various linguistic value defined in the attribute, \( m \) and \( m \) are the maximum and minimum values, respectively, of the attribute’s domain.

For example, suppose that 3 fuzzy sets should be defined, the fuzzy sets are defined as can be seen in figure 3:

![Fig. 3: Fuzzy Partition.](image)

2.5 Document Categorization

Text categorization (also known as text classification) is the task of automatically sorting a set of documents into categories from a predefined set. In automatic text categorization, the decision criterion of the text classifier is usually learned from a set of training documents, labeled for each class.

Each category is established using two different ways, manually or automatically by the application of a clustering algorithm. Once the categories are established, a set of chosen concepts is extracted to semantically defined them. Considering each concept has a weight on each document according to FIS-CRM Model (\( \text{FIS – CRM}(c_i, d) \)); we distinguished three levels of concept relevance (relevant, sub-relevant and other). The weight \( kmod^c_{i} \) of a concept \( c_i \) within a category \( \zeta \) can be calculated as proposed in the equation 5.

\[
kmod^c_{i} = \sum_{j \in c} w_{ij} \times \left( 1 + \frac{\text{docs}(c_i, \zeta)}{|D|} \right) \times \text{Ln} \left( \frac{1 + \text{docs}(c_i, \zeta)}{|\zeta(c_i)| + 1} \right) \tag{5}
\]

where \( w_{ij} \) represents the relevance degree of the concept \( c_i \) in the document \( d_j \) using the FIS-CRM model, \( \text{docs}(c_i, \zeta) \) is the number of documents in the category \( \zeta \) in which the concept \( c_i \) occurs, \( |D^\zeta| \) is the total number of documents considered in the category, \( |\zeta| \) is the total number of categories in the context and \( \zeta(c_i) \) represents the number of categories in which the concept \( c_i \) has a positive membership degree.

According to this formula a concept is regarded as relevant in a category if it occurs more frequently than other concepts in a certain user document set, but occasionally elsewhere. A concept that occurs frequently in several document set could be a relevant domain concept but it is not useful to represent the category.

Once the concepts weights are calculated, it is possible to identify the relevance distribution of all the concepts and use them to construct the groups according to the following statements: Relevant concepts (concepts with higher relevance degrees), Sub-Relevant concepts (concepts with relevance degrees higher than the average) and Other concepts concepts with relevance degree lower than the average.
The category is defined by the most relevant concepts from the Relevant Concepts Set (75% to 100%). These concepts have an associated weight proportional to their importance.

The process to classify new incoming documents into different categories (folders or types) is carried out according to the following steps:

- **Linguistic preprocessing**: The document is automatically analyzed using Natural Language Processing Techniques. (see Section 2.2.1)
- **Conceptual representation**: The document is represented using the FIS-CRM model to consider all the concepts that not occurs in the content (see Section 2.2.2).
- **Comparison with different groups**: A conceptual comparison process will be done between the document and each category (folders or types) obtaining a compatibility degree among these elements ($X$). This compatibility value depends on two fundamental factors: the number of common concepts and the relevance degree of these common concepts. This way, the result of the conceptual comparison will be obtained from the following formula (Eq. 6):

$$X = \ln \left( \sum_{i \in C} \left( w_i^d \ast w_i^c \right) \right) \ast \frac{|(R \cup S) \cap C|}{|R \cup S|} \quad (6)$$

where $w_i^d$ is the weight of the concept $i$ in the document $d$, $w_i^c$ is the weight of the term with in the category, $|R \cup S|$ is the number of relevant and sub-relevant concepts that define the category, and $C$ the set of concepts with weight in the document (originals + obtained from conceptual readjustment).

3. Experiments

This section presents the results obtained from the analysis of the enhanced DMS and its comparison to the results obtained from the original DMS. Accordingly, experiments and evaluations are divided into two sections. User query development, is analyzed in the first section and the second section deals with the document categorization process. The obtained results show us that the use of soft computing and fuzzy logic techniques increases the relevant results and number of the retrieved documents which are consistent with the user’s request.

3.1 User Query Development Evaluation

In the case of DMS, it is necessary a test collection to evaluate the effectiveness of the proposed improvements. A test collection is an experimental tool to understand, compare and reproduce the results. In this research work, an Spanish Corpus has been used. This corpus contains more than one thousand files documents collected from State Commission on Access to Public Information of Sinaloa\(^5\) from the years 2008 and 2009. The complete collection includes a lot of business documents like invoices, contracts, balances, letters, etc. Each file is stored in Portable Document Format (pdf) and either represents one page from a document representing a collection of separate sheets. Each page was processed using a standard OCR software\(^6\). The resulting textual file containing the OCR recognition result is included in the corpus. For every file, a reconstruction of the original text has been prepared using a post-correction process.

Ten queries, established by a panel of experts, has been used to carried out the experiment. Each query is given to the improved DMS and the original DMS and the precision, recall and f-measure values are computed. Table 1 depicts the difference between the enhanced using and no enhanced DMS. It shows that precision per topic after the application of improvements is similar or better than the process without improvements, therefore, the new functionalities implemented in the enhanced has improved retrieval results. This improvement is more evident in the more complex queries, which are less accurate.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMS+</td>
<td>0.75</td>
<td>0.66</td>
<td>0.70</td>
</tr>
<tr>
<td>DMS</td>
<td>0.47</td>
<td>0.32</td>
<td>0.38</td>
</tr>
</tbody>
</table>

3.2 Document Categorization Evaluation

In order the evaluate the document categorization method the TREC 2011 Medical Records track has been used \([28]\). This track includes a retrieval task aiming to find EMRs that are relevant to a given natural language query. These EMRs are de-identified medical records, provided by the University of Pittsburgh BLULab NLP Repository. There is a total of more than one hundred thousand of medical reports. The categorization algorithm is tested on the types of reports existing in the corpus (radiology reports, and emergency department reports or cardiology reports).

Test collections for experiments are usually split into two parts: one part for training and a second part for testing the effectiveness of the system. To evaluate performance a 2-fold cross-validation is carried out and examined the accuracy as performance measure. The accuracy is defined the ratio of documents correctly classified (Eq. 7).

$$Ac = \frac{N_c}{|D|} \quad (7)$$

Where $|D|$ expresses the total number of documents and $N_c$ is the total number of documents correctly classified.

\(^5\)http://www.ceaipes.org.mx/index.php

\(^6\)https://code.google.com/p/tesseract-ocr/
The results on correctly classified documents are superior to 95%. Comparing this classifier with other similar classifier such as Naive Bayes Classifier, the results are similar in quality or even superior. There exist classifiers which offer in better results than the obtained in this work (SVM) but over experiments more annotated and with a minor volume of documents to deal with.

4. Conclusions

Currently, well tailored and highly personalized document business management systems are needed because there is an overwhelming need for document storage and retrieval. Our approach employs innovative techniques, such natural language processing, soft computing and fuzzy logic that have a large potential for business documents retrieval. The experiments results demonstrate the proposed innovations enhance the information retrieval capabilities of the DMS and improve the user satisfaction. Future work includes working with a larger number of users and a greater volume of documents.

References


