Semantic Clickstream Mining

Mehrdad Jalali[1], **and Norwati Mustapha**[2]
[1] Department of Software Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran
[2] Department of Computer Science, Universiti Putra Malaysia, Malaysia

Abstract - *Nowadays Web users are drowned in all kind of available information. However, only a tiny part of it is usually relevant to their preferences. Web usage mining which extracts knowledge for usage and clickstream data has become the subject of exhaustive research, as its potential for Web-based personalized services, prediction of user near future intentions, adaptive Web sites, and customer profiling are recognized. Moreover, semantic web aims to enrich the WWW by machine processable information which supports the user in his tasks. Semantic clickstream mining which integrates semantic in Web usage mining processes aims to improve the quality of the Web usage mining systems. Given the primarily syntactical nature of data Web usage mining operates on, the discovery of meaning is impossible based on these data only. Therefore, formalizations of the semantics of Web resources and navigation behavior are increasingly being used. In this paper, we discuss the interplay of the Semantic Web with Web usage mining and also we give an overview of where the two research areas meet today. Moreover a proposed framework to integrate semantic Web and Web usage mining discuss in the rest of the paper..*

Keywords: Web usage mining, Clickstream Mining, Semantic Web, Semantic Web usage mining

1 Introduction

Semantic Web Usage Mining (SWUM) or Semantic Clickstream Mining aims to integrate two research areas Semantic Web and Web Usage Mining for obtaining more fine-grained and meaningful user behaviours in the Web environment. To better understand user's next intentions from observing him while navigating on a Website, all semantically interaction data needs to be tracked as well as tracking clickstream data. The most important issue facing in the classical Web Usage Mining system is quality of the results. The aim of this paper is to give an overview of where the two areas meet today, and what we can do to improve the results of integrating semantic Web and Web Usage Mining.

The remainder of this paper is organized as follows: Section 2 covers a brief overview of the areas Semantic Web and Web Usage Mining. Section 3 describes some related research about semantic Web usage mining and introduces proposed framework for integrating semantic Web and Web usage mining. Finally, section 4 concludes the current study and sheds light on some directions in the future works.

2 Web Usage Mining and Semantic Web

In the first part of this section, we cover some backgrounds in the WUM systems. In the second part, we recall our understanding of semantic Web. A brief discussion about integrating these two areas will be illustrated in the end of this section.

2.1 Web Usage (Clickstream) Mining

In general, Web mining can be characterized as the application of data mining to the content, structure, and usage of Web resources [1, 2]. The goal of Web mining is to automatically discover local as well as global models and patterns within and between Web pages or other Web resources. However, Web mining tools aim to extract knowledge from the Web, rather than retrieving information. Research on Web mining is classified into three categories, which are Web structure mining that identifies authoritative Web pages, Web content mining that classifies Web documents automatically or constructs a multilayered Web information base, and Web usage mining that discover user access patterns in navigating Web pages [3]. The goal of Web usage mining, in particular, is to capture and model Web user behavioral patterns. The discovery of such patterns from the enormous amount of data generated by Web and application servers has found a number of important applications. Among these applications are systems to evaluate the effectiveness of a site in meeting user expectations [4], techniques for dynamic load balancing and optimization of Web servers for better and more efficient user access [5], and applications for dynamically restructuring or customizing a site based on users' predicted needs and interests .

From the data-source perspective, both Web structure and Web content mining target the Web content, while Web usage mining targets the Web access logs. Web usage mining (WUM) comprises three major processes: data pretreatment, data mining, and pattern analysis [3]. Pretreatment performs a series of processing on Web log files, which are data conversion, data cleaning, user identification, session identification, path completion, and transaction identification. Next, mining algorithms are applied to extract user navigation patterns. A navigation pattern represents the relationships among Web pages in a particular Web site. Some pattern analyzing algorithm is applied to extract data from data mining part for the recommendation system. Recently, a number of Web usage mining (WUM) systems have been proposed to predict user's preferences and their navigation behaviors.

More recently, Web usage mining techniques have been proposed as another user-based approach to personalization which alleviates some of the problems associated with collaborative filtering. In particular, Web usage mining has been used to improve the scalability of personalization systems based on traditional CF-based techniques. In [6] we advance an architecture for online predicting in Web usage mining recommendation system and propose a novel approach to classifying user navigation patterns for predicting users' future requests. The approach is based on using the longest common subsequence (LCS) algorithm in classification part of the system. All of these works attempt to find architecture and algorithm to improve accuracy of personalized recommendation, but the accuracy still does not meet satisfaction especially in large-scale Websites. Fig. 1 illustrates the state of the proposed system.

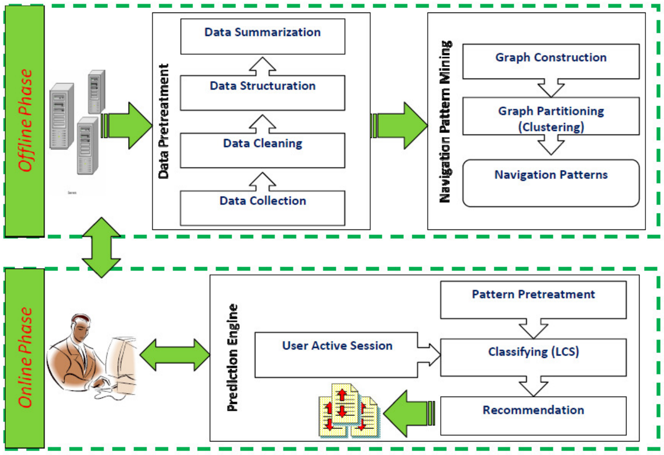

Fig.1 A WUM system framework [6]

To improve the quality of the results of the WUM systems, domain knowledge about a Web site can be integrated to the WUM process. Domain knowledge can be integrated into the Web usage mining process in many ways. This includes leveraging explicit domain ontologies or implicit domain semantics extracted from the content or the structure of documents or Web site. In general, however, this process may involve one or more of three critical activities [7]:

2.1.1 Domain ontology acquisition

The process of acquiring, maintaining and enriching the domain ontologies is referred to as "ontology engineering". For small Web sites with only static Web pages, it is feasible to construct a domain knowledge base manually or semi-manually. The outcome of this phase is a set of formally defined domain ontologies that precisely represent the Web site. Good representation should provide machine understandability, the power of reasoning, and computation efficiency.

2.1.2 Knowledge base construction

While the first phase generates the formal representation of concepts and relations among them, the second phase, knowledge base construction, can be viewed as building mappings between concepts or relations on the one hand, and objects on the Web. The goal of this phase is to find the instances of the concepts and relations from the Web site's domain, so that they can be exploited to perform further data mining tasks. Information extraction methods play an important role in this phase.

2.1.3 Knowledge-enhanced pattern discovery

Domain knowledge enables analysts to perform more powerful Web data mining tasks. For example, semantic knowledge may help in interpreting, analyzing, and reasoning about usage patterns discovered in the mining phase.

In the following we introduce semantic Web which can be integrated to the Web usage mining process.

2.2 Semantic Web

The Semantic Web aims to obtain machine-understandable information from WWW which is based on a vision of Tim Berners-Lee, the inventor of the WWW. The great success of the current WWW leads to a new challenge: a huge amount of data is interpretable by humans only; machine support is limited. He suggested to enrich the Web by machine-processable information which supports the user in his tasks. For instance, today's search engines are already quite powerful, but still frequently return overly large or inadequate lists of hits. Machine-processable information can point the search engine to the relevant pages and improve the quality of the results. Fig. 2 shows the layers of the Semantic Web as suggested by Berners-Lee. This architecture is discussed in detail for instance in [8] and [9].

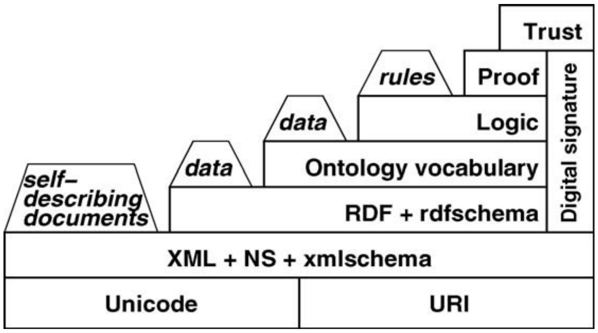

Fig. 2 the layers of the Semantic Web

Today, it is almost impossible to retrieve information with a keyword search when the information is spread over several pages. Consider the query for Web mining experts in a company intranet, where the only explicit information stored are the relationships between people and the courses they attended on one hand, and between courses and the topics they cover on the other hand. In that case, the use of a rule stating that people who attended a course which was about a certain topic have knowledge about that topic might improve the results.

The process of building the Semantic Web is today still under way. Its structure has to be defined, and this structure should be brought to life. Given the primarily syntactical nature of data Web usage mining operates on, the discovery of meaning is impossible based on these data only. Therefore, semantic knowledge of Web documents and navigation behaviors can be utilized in recommendation systems for predicting different types of complex Web document and object based on underlying properties and attributes especially in large-scale and dynamic Web sites. Mapping between user navigation transactions in Web usage mining to semantic transaction based on concepts and objects can improve the accuracy of the Web usage mining personalization system.

3 Semantic Clickstream Mining

Our goal in this section is to provide a road map for the integration of semantic and ontological knowledge into the process of Web usage mining. Semantics can be utilized for Web Usage Mining for different purposes which are introduced in this section.

To better integrate semantic Web and WUM it would be desirable to have a rich semantic model of content and structure of a site. This model should capture the complexity of the manifold relationships between the concepts covered in a site, and should be ”built into” the site in the sense that the pages requested by visitors are directly associated with the concepts and relations treated by it. This leads to semantic Web usage mining. Semantic Web usage mining involves the integration of domain knowledge into Web usage mining [7] . Utilizing semantic knowledge can lead to deeper interaction of the Website's user with the site. Integration of domain knowledge allows such systems to infer additional useful recommendations for users based on more fine grained characteristics of the objects being recommended, and provides the capability to explain and reason about user actions.

The interpreting, analysing, and reasoning about usage patterns discovered in the mining phase can be done by using semantic knowledge. Moreover, it can improve the quality of the recommendations in the usage-based system.

Several studies have considered various approaches to integrate content-based semantic knowledge into traditional usage-based recommender systems. An overview of the existing approaches as well as a some framework for integrating domain ontologies with the personalization process based on Web usage mining is given in the following.

Our main research question is; Can usage patterns reveal further relations to help build the Semantic Web? This field is still rather new, so we will only describe an illustrative selection of research approaches.

Ypma and Heskes propose a method for learning content categories from usage [10]. They model navigation in terms of hidden Markov models, with the hidden states being page categories, and the observed request events being instances of them. Their main aim is to show that a meaningful page categorization may be learned simultaneously with the user labeling and intercategory transitions; semantic labels (such as “sports pages”) must be assigned to a state manually. The resulting taxonomy and page classification can be used as a conceptual model for the site, or used to improve an existing conceptual model.

Chi et al.[7] identify frequent paths through a site. Based on the keywords extracted from the pages along the path, they compute the likely “information scent” followed, i.e. the intended goal of the path. The information scent is a set of weighted keywords, which can be inspected and labeled more concisely by using an interactive tool. Thus, usage creates a set of information goals users expect the site to satisfy. These goals may be used to modify or extend the content categories shown to the users, employed to structure the site's information architecture, or employed in the site's conceptual model.

Stojanovic, Maedche, Motik, and Stojanovic [11] propose to measure user interest in a site's concepts by the frequency of accesses to pages that deal with these concepts. They use these data for ontology evolution: Extending the site's coverage of high-interest concepts, and deleting low-interest concepts, or merging them with others.

The combination of implicit user input (usage) and explicit user input (search engine queries) can contribute further to conceptual structure. User navigation has been employed to infer topical relatedness, i.e. the relatedness of a set of pages to a topic as given by the terms of a query to a search engine. A classification of pages into “satisfying the user defined predicate” and “not satisfying the predicate” is thus learned from usage, structure, and content information. An obvious application is to mine user navigation to improve search engine ranking [12].

Many approaches use a combination of content and usage mining to generate recommendations. For example, in contentbased collaborative filtering, textual categorization of documents is used for generating pseudo-rankings for every userdocument pair [8]. In [9], ontologies, IE techniques for analyzing single pages, and a user's search history together serve to generate recommendations for query improvement in a search engine.

In [13], Authors have presented a general framework for using domain ontologies to automatically characterize usage profiles containing a set of structured Web objects. Their motivation has been to use this framework in the context of Web personalization, going beyond page-level or item-level constructs, and using the full semantic power of the underlying ontology. They considered a Web site as a collection of objects belonging to certain classes (resulting in a concept

Hierarchy of Genre's a portion of which is shown in Fig. 3. Given a collection of similar user sessions (e.g., obtained through clustering) each containing a set of objects, they have shown how to create an aggregate representation of for the whole collection based on the attributes of each object as defined in the domain ontology (Fig. 4). This aggregate representation is a set of pseudo objects each characterizing objects of different types commonly occurring across the user sessions. They have also presented a framework for Web personalization based on domain-level aggregate profiles.

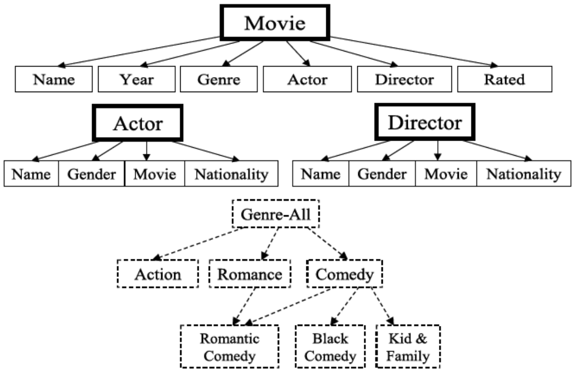

Fig. 3 The Ontology for a movie Web site

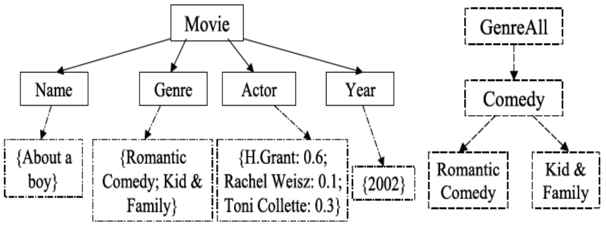
Movie
Name
Genre
Actor
Year
{About a boy}
{Romantic Comedy; Kid & Family}
{H.Grant: 0.6; Rachel Weisz: 0.1; Toni Collette: 0.3}
{2002}
GenreAll
Comedy
Romantic Comedy
Kid & Family

Fig. 4 An Example of an Object in Class Movie

In [14], they proposed an approach to track user interaction data and preserving semantic knowledge on complex and interactive Web sites. They showed that the approach comes with some major enhancements compared to some existing solutions. The usage of Microformats enables an easy integration into existing Web sites and allows then to interrelate data on these sites (Microformats are small patterns of HTML to represent commonly published things like people, events, blog posts, reviews and tags in web pages. Microformats enable the publishing of higher fidelity information on the Web; the fastest and simplest way to provide feeds and APIs for the information in your website.). This also allows them to obtain fine-grained information connected with semantic knowledge that opens new chances to personalize Web sites.

Recommender systems rely on relevance scores for individual content items; in particular, pattern-based recommendation exploits co-occurrences of items in user sessions to ground any guesses about relevancy. To enhance the discovered patterns' quality, the authors in [15] propose using metadata about the content that they assume is stored in a domain ontology.

Their approach comprises a dedicated pattern space built on top of the ontology, navigation primitives, mining methods, and recommendation techniques.

Web usage mining (WUM) approaches often use terms and frequencies to represent a Web site for the mining process. In [16], the authors show that these representations lead to poor results. Therefore, it is proposed to perform a semantic Web usage mining process to enhance quality of the mining results. In this paper it was used a concept-based Web usage mining process to generate more semantically related results.

The approach was used to enhance a real Web site and it was evaluated by comparing it with four different WUM methods. It was defined two quality measures (interest and utility) in order to evaluate the results. These measures are obtained using surveys to 100 visitors of the site. Based on interest and correlation measures, it was proved that concept-based approach allows obtaining results closer to visitors' real browsing preferences. Moreover, information produced by the proposed approach lead to the discovery of enhancements. The proposed method also finished the generalization task in a few minutes which is not too much compare with other methods.

In [10], they proposed the integration of semantic information drawn from a Web application's domain knowledge into all phases of the Web usage mining process (preprocessing, pattern discovery, and recommendation/prediction). The goal is to have an intelligent semantics-aware Web usage mining framework. This is accomplished by using semantic information in the sequential pattern mining algorithm to prune the search space and partially relieve the algorithm from support counting. In addition, semantic information is used in the prediction phase with low order Markov models, for less space complexity and accurate prediction that will help solve ambiguous predictions problem. Experimental results show that semantics-aware sequential pattern mining algorithms can perform 4 times faster than regular non-semantics-aware algorithms with only 26% of the memory requirement.

Fig. 5 illustrates the proposed architecture for a semantic Web usage mining system which can be used in a recommender system. In the offline phase of the system to perform semantic data pretreatment, Web site ontology and a knowledgebase which are created based on the content and structure of the Web site can be utilized in the process of this module. On the other hand to create semantic usage data which further will be used in semantic navigation pattern, and to understand semantic knowledge about user semantically' sessions in a particular website, this module needs to integrate with those ontology and knowledgebase.

In the next module, semantic navigation patterns will be extracted from the sessions by utilizing semantic clustering algorithm. In the online phase, the system recommends some pages and concepts which the current users intend to navigate them through the particular Web site. This phase in similar to our work which is described in [6].

In summary, all of these works attempt to find reference architecture and framework to improve quality of the Web usage mining systems by integrating semantic web to WUM, In the proposed framework, we advance a framework to integrate semantic web and WUM, which by using appropriate semantic clustering algorithm and well-done ontology and knowledgebase design, the quality of the semantic Web usage mining can be enhanced as future work.

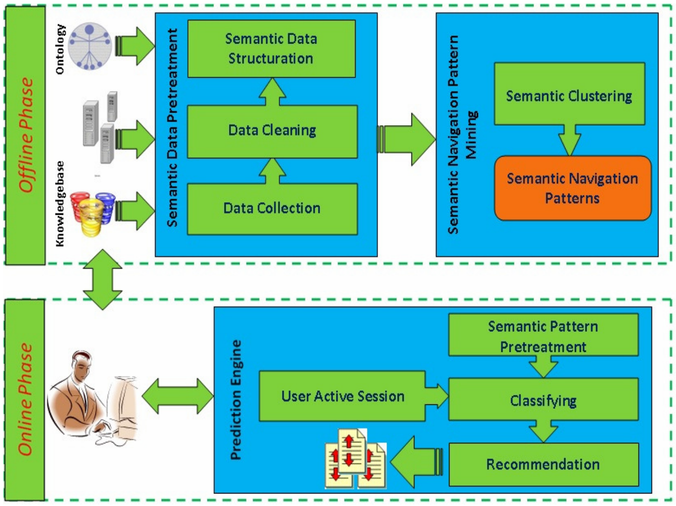

Fig. 5 The proposed framework

4 Conclusion and Future Work

In this paper, we have studied the integration of the two fast developing research areas Semantic Web and Web usage mining. We discussed how Semantic Web usage mining can improve the results of Web usage Mining systems by by exploiting the semantic structures in the process of the Web usage mining. Moreover, a proposed framework to integrate semantic Web and Web usage mining discussed in this paper. As a future direction, we plan to develop a semantic Web usage mining system which uses the proposed framework.

5 References

[1] R. Cooley, et al., "Web mining: information and pattern discovery on the World Wide Web," in Ninth IEEE International Conference on Tools with Artificial Intelligence, Newport Beach, CA, USA, 1997, pp. 558-567.

[2] J. Srivastava, et al., "Web usage mining: discovery and applications of usage patterns from Web data," ACM SIGKDD Explorations Newsletter, vol. 1, pp. 12-23, 2000.

[3] B. Mobasher, et al., "Creating adaptive Web sites through usage-based clustering of URLs," in Knowledge and Data Engineering Exchange, Chicago, IL, USA, 1999, pp. 19-25.

[4] M. Spiliopoulou, "Web usage mining for web site evaluation," Communications of the ACM, vol. 43, pp. 127-134, 2000.

[5] J. E. Pitkow, et al., WebVis: A Tool for World Wide Web Access Log Analysis: Graphics, Visualization & Usability Center, Georgia Institute of Technology, 1994.

[6] M. Jalali, et al., "WebPUM: A Web-based recommendation system to predict user future movements," Expert Systems with Applications, vol. 37, pp. 6201-6212, 2010.

[7] E. H. Chi, et al., "Using information scent to model user information needs and actions and the Web," in Proceedings of the SIGCHI, 2001, pp. 490-497.

[8] P. Melville, et al., "Content-boosted collaborative filtering for improved recommendations," in Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2002, pp. 187-192.

[9] S. Parent, et al., "An adaptive agent for web exploration based on concept hierarchies," in In Proceedings of the 9th International Conference on Human Computer Interaction, 2001.

[10] N. R. Mabroukeh and C. I. Ezeife, "Using domain ontology for semantic web usage mining and next page prediction," in CIKM '09 Proceeding of the 18th ACM conference on Information and knowledge management, 2009, pp. 1677-1680.

[11] L. Stojanovic, et al., "User-driven ontology evolution management," Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web, pp. 133-140, 2002.

[12] C. Kemp and K. Ramamohanarao, "Long-term learning for web search engines," Principles of Data Mining and Knowledge Discovery, pp. 243-311, 2002.

[13] H. Dai and B. Mobasher, "Using ontologies to discover domain-level web usage profiles," Semantic Web Mining, p. 35, 2002.

[14] T. Plumbaum, et al., "Semantic web usage mining: Using semantics to understand user intentions," User Modeling, Adaptation, and Personalization, pp. 391-396, 2009.

[15] M. Adda, et al., "Toward recommendation based on ontology-powered web-usage mining," IEEE Internet Computing, pp. 45-52, 2007.

[16] S. R os and J. Velásquez, "Semantic web usage mining by a concept-based approach for off-line web site enhancements," 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 234-241, 2008.