A Study of Semantic Web Mining: Integrating Domain Knowledge into Web Mining

V. Sitha Ramulu, Ch. N. Santhosh Kumar, K. Sudheer Reddy

Abstract*— The Semantic Web and Web Mining are two fast-developing research areas, which have many points in contact. Web mining applies data mining techniques on Web content, usage, and structure. Methods of Web content mining can, on the one hand, be used to create semantic annotations from Web page content; on the other hand, content mining can profit from content that is already structured in XML, RDF, or ontological format. Methods of Web usage mining can profit from semantically enriched descriptions of the Web pages visited; this will provide for the identification of more meaningful patterns within site visits, and better site improvements, recommendations, and personalization options based on these patterns. Usage patterns can in turn serve to improve the semantic annotations of pages. The third form of Web Mining, Web structure mining, utilizes the hyperlink structure. Crawlers that take into account structure as well as semantic content can significantly improve search engine results. The rapid development of the field means that Semantic Web Mining now plays a wide range of different roles. Semantic web mining is the combination of the semantic web and web mining. This paper presents an overview of semantic web mining.

Index Terms*—Domain knowledge, Ontologies, Semantic web, Web mining.

I. INTRODUCTION

The World Wide Web [1], has grown in the past few years from a small research community to the biggest and most popular way of communication and information dissemination. Every day, the WWW grows by roughly a million electronic pages, adding to the hundreds of millions already on-line. WWW serves as a platform for exchanging various kinds of information, ranging from research papers, and educational content, to multimedia content and software. The continuous growth in the size and the use of the WWW imposes new methods for processing these huge amounts of data. Because of its rapid and chaotic growth, the resulting network of information lacks of organization and structure. Moreover, the content is published in various diverse formats. Due to this fact, users are feeling sometimes disoriented, lost in that information overload that continues to expand.

Web mining [2] is a very broad research area emerging to solve the issues that arise due to the WWW phenomenon. The Web mining research is a converging research area from several research communities, such as Databases, IR and AI. There are three axes of web mining namely:

- Web structure mining
- Web content mining
- Web usage mining

The goal of Web structure mining is to categorize the Web pages and generate information such as the similarity and relationship between them, taking advantage of their hyperlink topology. In the latter years, the area of Web structure mining focuses on the identification of authorities, i.e. pages that are considered as important sources of information from many people in the Web community.

Web content mining has to do with the retrieval of information (content) available on the Web into more structured forms as well as its indexing for easy tracking information locations. Web content may be unstructured (plain text), semi- structured (HTML documents), or structured (extracted from databases into dynamic Web pages). Such dynamic data cannot be indexed and consist what is called "the hidden Web". A research area closely related to content mining is text mining.

Web usage mining is the process of identifying browsing patterns by analyzing the user's navigational behavior. This information takes as input the usage data, i.e. the data residing in the Web server logs, recording the visits of the users to a Web site. Extensive research in the area of Web usage mining led to the appearance of a related research area, that of Web personalization. Web personalization utilizes the results produced after performing Web usage mining, in order to dynamically provide recommendations to each user.

It is evident that the distinctions between the three axes of Web mining (especially of Web content and Web structure mining) are in many cases ambiguous. Most of the current research efforts concentrate on combining methods proposed in either one of them, in order to enhance the results of their studies. This combination implies the advancing of Web mining to a more abstract level. To achieve this abstraction, Web data (usage, content, structure) are represented using another emerging model of representation, ontologies. This representation, closes the gap between Semantic Web and Web Mining areas, to create a fast-emerging research area, that of Semantic Web Mining[3]. This paper presents an overview of the semantic web mining- Integration of domain knowledge in to web mining to form semantic web mining, the concepts of semantic web mining.

II. SEMANTIC WEB MINING: INTEGRATING DOMAIN KNOWLEDGE INTO WEB MINING

Web mining is the process of discovering and extracting useful knowledge from the content, usage, and structure of one or more Web sites. Semantic Web mining involves the integration of domain knowledge into the Web mining process. Domain knowledge can be integrated into the Web

Manuscript received on July, 2012.

V. Sitha Ramulu , Assoc. Professor, Dept. of IT, Swarna Bharathi Inst. of Sc. & Tech, Khammam, AP, India

Ch. N. Santhosh Kumar , Assoc. Professor, Dept. of CSE, Swarna BharathiInst. of Sc. & Tech, Khammam, AP, India,

K. Sudheer Reddy, Research Scholar, Dept. of CSE, Acharya Nagarjuna University, Guntur, AP, India.

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 Website. In general, however, this process may involve one or more of three critical activities: domain ontology acquisition, knowledge base construction, and knowledge-enhanced pattern discovery.

A. 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. In [4] a semi-manual approach is adopted for defining each domain concept as a vector of terms with the help of existing vocabulary and natural language processing tools. However, manual construction and maintenance of domain ontologies requires a great deal of effort on the part of knowledge engineers, particularly for large-scale Websites or Websites with dynamically generated content. In dynamically generated Websites, page templates are usually populated based on structured queries performed against back-end databases. In such cases, the database schema can be used directly to acquire ontological information. Some Web servers send structured data files (e.g., XML files) to users and let client-side formatting mechanisms (e.g., CSS files) work out the final Web representation on client agents. In this case, it is generally possible to infer the schema from the structured data files.

When there is no direct source for acquiring domain ontologies, machine learning and text mining techniques must be employed to extract domain knowledge from the content or hyperlink structure of the Web pages. The outcome of this phase is a set of formally defined domain ontologies that precisely represent the Website. A good representation should provide machine understandability, the power of reasoning, and computation efficiency.

B. Knowledge Base Construction

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 Website's domain, so that they can be exploited to perform further data mining tasks. Learning algorithms plays an important role in this phase.

In [5] a text classifier is learned for each "semantic feature" (somewhat equivalent to the notion of a concept) based on a small manually labeled data set. First, Web pages are extracted from different Websites that belong to a similar domain, and then the semantic features are manually labeled. This small labeled data set is fed into a learning algorithm as the training data to learn the mappings between Web objects and the concept labels. In fact, this approach treats the process of assigning concept labels as filling "missing" data.

C. Knowledge-Enhanced Web Data Mining

Domain knowledge enables analysts to perform more powerful Web data mining tasks. The applications include content mining, information retrieval and extraction, Web usage mining, and personalization. On the other hand, data mining tasks can also help to enhance the process of domain knowledge discovery. Domain knowledge can improve the accuracy of document clustering and classification and induce more powerful content patterns. For example, in [6] domain ontologies are employed in selecting textual features. The selection is based on lexical analysis tools that map terms into concepts within the ontology. The approach also aggregates concepts by merging the concepts that have low support in the documents. After preprocessing, only necessary concepts are selected for the content clustering step.

Traditional approaches to content mining and information retrieval treat every document as a set or a bag of terms. Without domain semantics, we would treat "human" and "mankind" as different terms, or, "brake" and "car" as unrelated terms. In [4] a concept is defined as a group of terms that are semantically relevant, for example, as synonyms. With such concept definitions, concept distribution among documents is analyzed to find interesting concept patterns. For example, one can discover dominant themes in a document collection or in a single document; or find associations among concepts.

III. SEMANTIC WEB MINING CONCEPTS

The Semantic Web can be defined as an extension of the current web. Here the information is presented in a well-defined manner, better enabling computers and people to work in cooperation. Data in the Semantic Web is defined and linked in a way that can be used for more effective discovery, automation, integration and reuse across applications. This data can be shared and processed by automated tools as well as people. The Semantic Web will provide an infrastructure that enables not just web pages, but databases, services, programs, sensors, personal devices, and even household appliances to both consume and produce data on the web.Semantic web mining is essentially mining the information pertaining to the semantic Web. This means mining Web pages so that the machine can better understand the information. It also means mining the data sources to develop an effective semantic Web.

Figure 1 illustrates semantic Web mining. It is essentially mining various XML and RDF documents as well as mining ontologies and metadata. It also includes mining the data sources on the Web and mining the information relating to the information management technologies. The semantic Web is essentially an evolution of the Web. We will not end with the semantic Web. It will evolve as far as we can see. In the same way, semantic Web mining will evolve from Web mining. The ultimate goal is to make the Web easier to use. In addition, we want to mine the large quantities of information on the Web so that humans can better perform their tasks.

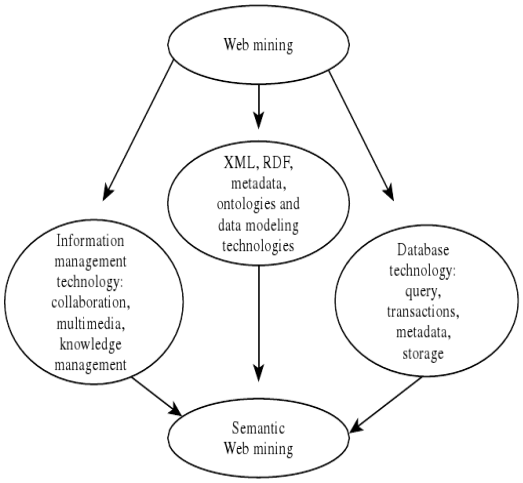


Figure 1 : Semantic Web Mining Technologies

Figure 2 illustrates a semantic Web mining concept of operation. With semantic Web mining, the brokering services [7] use Web mining to determine the best publishers for the subscribers and to advise them. That is, it is not just a matching service; it mines the information and data sources, finds the best match, and provides advice for future services and enhancements.

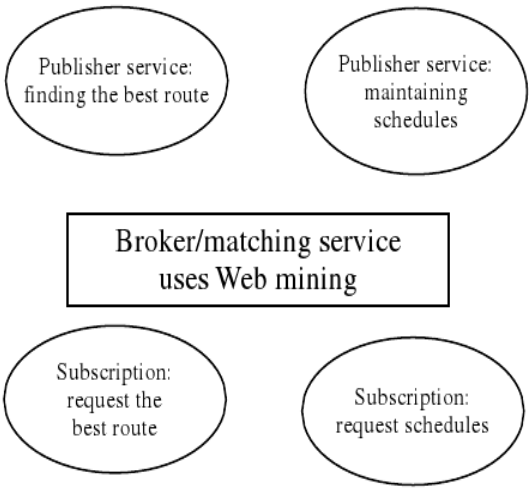
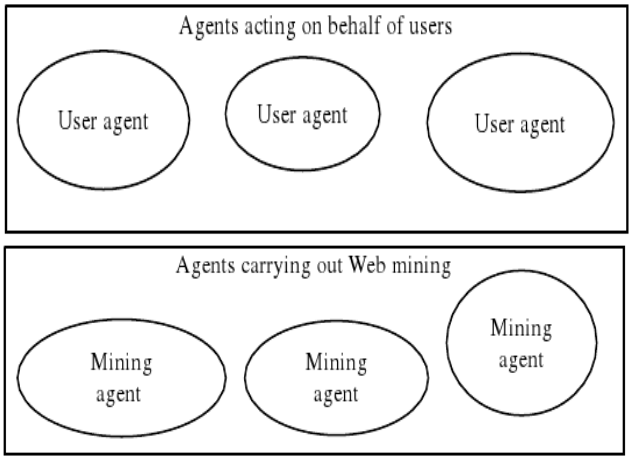


Figure 2: Concept of Operation for Semantic Web Mining

Figure 3 shows how Web mining agents can be integrated with the locator agents and user agents. That is, the Web mining agents give advice to the user agents as well as to the locator agents in terms of finding the resources.



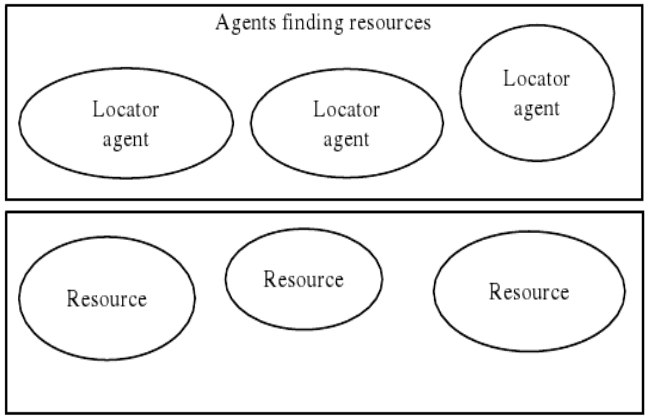


Figure 3: Agents for Semantic Web Mining

IV. CONCLUSION

The two fast developing research areas in World Wide Web are: web mining and semantic web. Web Mining is the application of data mining techniques to the content, structure and usage of Web resources. The three main areas of Web Mining are: Content mining, structure mining, usage mining. The Semantic Web can be defined as an extension of the current web. Here the information is presented in a well-defined manner, better enabling computers and people to work in cooperation. The Semantic Web offers to add structure to the Web, while Web Mining can learn implicit structures. The combined area of Semantic Web Mining offers new techniques to improve both areas. Semantic Web mining involves the integration of domain knowledge into the Web mining process. Domain knowledge can be integrated into the Web mining process in three ways- domain ontology acquisition, knowledge base construction, and knowledge-enhanced pattern discovery. The resulting research benefits many areas of industry such as "e-activities", health care, privacy and security, and knowledge management and information retrieval.

REFERENCES

[1] R. Cooley, B. Mobasher, J. Srivastava, Web Mining: Information and Pattern Discovery on the World Wide Web, in Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97).

[2] R. Kosala, H. Blockeel, Web Mining Research: A Survey, SIGKDD Explorations, 2(1):1-15, 2000.

[3] B. Berendt, A. Hotho, G. Stumme, Towards Semantic Web Mining, in Proceedings of 1st International Semantic Web Conference (ISWC 2002).

[4] Loh, S., Wives, L.K., & de Oliveira, J.P. (2000). Concept-based knowledge discovery in texts extracted from the Web. *SIGKDD Explorations*, *2*(1), 29-39.

[5] Ghani, R., & Fano, A. (2002). Building recommender systems using a knowledge base of product semantics. *Proceedings of the Workshop on Recommendation and Personalization in E-Commerce*, 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems, Malaga, Spain.

[6] Horrocks, I. (2002). DAML+OIL: A description logic for the semantic Web. *IEEE Data Engineering Bulletin*, *25*(1), 4-9.

[7] Marian Nodine, William Bohrer, Anne Hiong, "semantic Brokering over dynamic heterogeneous data sources in Infosleuth", 15th international conference on Data Engineering, pages 358- 365, 1999.