ISSN:0975-9646

Semantic Web Mining - A Critical Review

S.S. Dhenakaran[#1], S.Yasodha[*2]

 # Asst. Prof. in Computer Science
Alagappa University
Karaikudi, Tamilnadu, India
** Asst. Prof. in Computer Science*
Govt. Arts College(W),
Pudukkottai, Tamilnadu, India

Abstract **: Over the last decade, there is an explosive growth in the information available on the World Wide Web (WWW). Today, web browsers provide easy access to myriad sources of text and multimedia data. More than one billion pages are indexed by search engines, and finding the desired information is not an easy task. This profusion of resources has prompted the need for developing automatic mining techniques on the WWW, thereby giving rise to the term "Web Mining". The "Semantic Web" aims to address this problem by providing machine interpretable semantics to provide greater machine support for the user. These two areas pave way for the extraction of relevant and meaningful information from the web, thereby giving rise to the term "Semantic Web Mining".**

The objective of this paper is to provide an outline of web mining, its various classifications, its subtasks, and to give a perspective to the research community about the potential of applying techniques to extract meaningful patterns. This paper also gives a survey of the recent works in the area of semantic web mining and a comparison of traditional web applications and semantic web applications thereby providing guidelines for future research in the area of semantic web mining.

Keywords: **Information retrieval, selection, extraction, preprocessing, content mining, structure mining, usage mining.**

I. INTRODUCTION

A. Web Mining

Web Mining is the application of data mining techniques to the content, structure and usage of Web resources[1]. Web mining is growing rapidly since its inception in or around 1996, and new methodologies are being developed both using classical and soft computing approaches concurrently. Web mining, when looked upon in data mining terms, can be said to have three operations of interests – clustering (eg. Finding natural groupings of users, pages, etc.), associations (eg. which URLs tend to be requested together) and sequential analysis (eg. the order in which URLs tend to be accessed).

The three main areas of web mining are:

- *Content Mining* - Analyses the content of Web resources. It describes the discovery of useful information from the web contents. Mainly based on text mining techniques, but extensions to multimedia content is beginning to emerge in the research. The web content consists of several types of data such as textual, image, audio, video, metadata, as well as hyperlinks.

  Most of the efforts on web content mining are concentrated on the text or hypertext contents. The textual parts of web content data consist of unstructured data such as text documents, semi-structured data such as HTML documents and more-structured data such as data in tables or database-generated HTML pages.
- *Structure Mining* - Analyses the hyperlink structure between Web pages. Web Structure Mining is concerned with discovering the model underlying the link structures of the web. It is used to study the topology of the hyperlinks. This model can be used to categorize web pages and is useful to generate information such as the similarity and relationship between different websites.

  While Web Content Mining attempts to explore the structure within a document (intra-document structure), Web Structure Mining studies the structures of documents within the web itself (inter-document structure).
- *Usage Mining* - Analyses the user's clicks from Web server logs. Web Usage Mining deals with studying the data generated by the web surfer's sessions or behaviours. Web Usage Mining mines the secondary data derived from web server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls.

Web mining can be viewed as consisting of four tasks [2] as shown below:

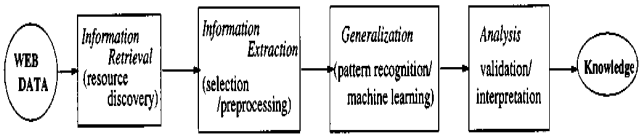

Fig. 2. Web mining subtasks.

1) Information Retrieval (IR) (Resource Discovery):

Resource discovery or IR deals with automatic retrieval of all relevant documents, while at the same time ensuring that the irrelevant ones are fetched as few as possible. The IR process mainly includes document representation, indexing, and searching for documents.

2) Information Selection/Extraction and Preprocessing:

Once the documents have been retrieved the challenge is to automatically extract knowledge and other required information without human interaction. Information extraction (IE) is the task of identifying specific fragments

of a single document that constitute its core semantic content.

3) Genealization:

In this phase, pattern recognition and machine learning techniques are usually used on the extracted information.

4) Analsis:

Analysis is a data-driven problem which presumes that there is sufficient data available so that potentially useful information can be extracted and analyzed. Humans play an important role in the information or knowledge discovery process on the web since the web is an interactive medium. This is especially important for validation and or interpretation of the mined patterns which take place in this phase.

B. Semantic Web

The current WWW has a huge amount of data that is often unstructured and usually only human understandable. The Semantic Web aims to address this problem by providing machine interpretable semantics to provide greater machine support for the user.

Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in co-operation. The semantic web will provide intelligent access to heterogeneous, distributed information enabling software products to mediate between user needs and the information source available.

Semantic Web is

1. Providing a common syntax for machine understandable statements.
2. Establishing common vocabularies.
3. Agreeing on a logical language.
4. Using the language for exchanging proofs.

The Semantic Web has a layer structure that defines the levels of abstraction applied to the Web. At the lowest level is the familiar World Wide Web, then progressing to XML, RDF, Ontology, Logic, Proof and Trust. The main tools that are currently being used in the Semantic Web are ontologies based on OWL (Web Ontology Language) and its associated reasoners. Semantic Web is the advanced technique used by the web to fulfill the client's requirements. In Semantic Web, lot of information can be stored in RDF in an XML file format. This information can be extracted by the users depending upon their needs. This can be done by accepting the user request and providing response to the user by extracting the information from the RDF. Thus the information can be easily extracted by the Web.

C. Semantic web mining

The human ability for information processing is limited on the one hand, whilst otherwise the amount of available information of the Web increases exponentially, which leads to increasing information saturation[3]. In this context, it becomes more and more important to detect useful patterns in the Web, thus use it as a rich source for data mining [4].

The research area of Semantic Web Mining is aimed at combining two fast developing fields of research: the Semantic Web and Web Mining. The idea is to improve, on the one hand, the results of Web Mining by exploiting the new semantic structures in the Web; and to make use of Web Mining, on the other hand, for building up the Semantic Web. These two fields address the current challenges of the World Wide Web (WWW): turning unstructured data into machine-understandable data using Semantic Web tools.

As the Semantic Web enhances the first generation of the WWW with formal semantics, it offers a good basis to enrich Web Mining: The types of (hyper)links are now described explicitly, allowing the knowledge engineer to gain deeper insights in Web structure mining; and the contents of the pages come along with a formal semantics, to apply mining techniques which require more structured input.

- *Semantic Web Content and Structure Mining*

In the Semantic Web, content and structure are strongly intertwined. Therefore, the distinction between content and structure mining vanishes. However, the distribution of the semantic annotations may provide additional implicit knowledge. An important group of techniques which can easily be adapted to semantic Web content / structure mining are the approaches discussed as *Relational Data Mining* (formerly called *Inductive Logic Programming* (*ILP*)). Relational Data Mining looks for patterns that involve multiple relations in a relational database. It comprises techniques for Semantic Web Mining like classification, regression, clustering, and association analysis. It is quite straightforward to transform the algorithms so that they are able to deal with data described in RDF or by ontologies.

There are two big scientific challenges in this attempt. The first is the size of the data to be processed (i.e. the scalability of the algorithms), and the second is the fact that the data are distributed over the Semantic Web, as there is no central database server. Scalability has always been a major concern for ILP algorithms. With the expected growth of the Semantic Web, this problem increases as well. Therefore, the performance of the mining algorithms has to be improved, e.g. by sampling.

As for the problem of distributed data, it is a challenging research topic to develop algorithms which can perform the mining in a distributed manner, so that only (intermediate) results have to be transmitted, and not whole datasets.

- *Semantic Web Usage Mining*

Usage mining can also be enhanced further if the semantics are contained explicitly in the pages by referring to concepts of an ontology. Semantic Web usage mining can for instance be performed on log files which register the user behavior in terms of an ontology. A system for creating such semantic log files from a knowledge portal has been developed at the AIFB. These log files can then be mined, for instance to cluster users with similar interests in order to provide personalized views on the ontology.

II. APPLICATIONS

The web is not only a publishing infrastructure but also an application platform for web applications. But existing applications use the web primarily as a means to access their application, generating HTML pages from their database content and serving these pages over HTTP. These database-driven applications result in a "deep" or

"hidden web", whose dynamically-generated pages do not conform to traditional web principles such as hyperlinks and are thus hard to crawl and index.

TABLE I
TRADITIONAL VS. SEMANTIC WEB APPLICATIONS

Web Application	**Semantic Web Application**
centralized	Decentralized
one fixed schema	semi-structured
one fixed vocabulary	arbitrary vocabulary
centralized publishing	publish anywhere
one data source	many distributed data sources
closed systems	open systems

More importantly, these database-driven applications are closed systems that rely on a single centralized data source; due to the inherent limitations of their relational databases, these Web applications operate on fixed data structures and schema, use one fixed vocabulary, and do not interlink their data. In contrast, semantic web applications are more aligned with the principles of the web, such as interoperability, universality, and decentralization. Semantic Web applications are decentralized and open, operate on distributed data that can be published anywhere, may conform to arbitrary vocabulary and follow semi-structured schemas [5].

Semantic web mining benefits many areas such as e-activities, health care, privacy and security, and knowledge management and information retrieval. It offers great opportunities and challenges in many areas, including business, commerce, marketing, finance, publishing, education, research and development [6].

Current web mining research on E-learning is based on web usage mining as the focus has been on how the student performs. Now-a-days digital libraries are also accessible form the web. Many commercial institutions are transforming their businesses and services electronically. The challenge of the Semantic Web Mining technologies in the e-Learning domain can relate to the provision of personalized experiences for the users. Particularly, these applications can take into consideration the individual needs and requirements of learners [7].

The utilization of data mining on semantic web information for business intelligence has got not much attention in the research community in comparison to the overall research investments in this field. There are existing application scenarios of relational data mining where semantic web can be used in enterprises (risk management, competitive analysis etc.) [8].

Semantic web mining is applied in the E-Services areas of societal interest like E-Government, E-Politics and E-Democracy. Only recently web mining applications have been related to these three fields. Much of the government information are gradually being placed on the web in recent years. Current web mining research focus on E-Politics is based on web structure mining to identify political groups. It seems that the fields of E-services and web mining have recently met each other leading to more benefits to society [9].

Semantic web mining is also applied in genetics, molecules, social network analysis, as well as natural language processing [10].

III. REVIEW OF RELATED WORK

Diana Cerbu presented two new and fast-developing domains: Semantic Web and Data Mining. The author suggested how these areas can be combined and present three different approaches to semantic web mining: an approach to recurring pattern mining, a text classification algorithm called AdaBoost and a framework for generating better customized content on the web by using web mining combined with embedded ontologies [11].

Mahindra Pratap Singh Dohare et al. presented Sindice, an indexing and lookup service for Semantic Web data sources. Sindice allows application developers to easily discover relevant data sources for their application. Lookups for data sources can be performed directly using resource URIs, indirectly through uniquely-identifying inverse functional properties, and through a full-text search over the literals [5].

Thomas Fischer et al. have motivated the utilization of relational data mining algorithms in context of the semantic web. They have outlined important differences to the traditional knowledge discovery process. Especially the selection, modeling and transformation steps are different to the standard approach. The knowledge discovery process suggested by them selects parts of semantic data to fully utilize information and background information derived from a variety of sources [8].

Nizar R. Mabroukeh et al. introduced a comprehensive generic framework named SemAware that integrates semantic information into all phases of web usage mining. Semantic information can be integrated into the pattern discovery phase, such that a semantic distance matrix is used in the adopted sequential pattern mining algorithm to prune the search space and partially relieve the algorithm from support counting. A 1*st*-order Markov model is also built during the mining process and enriched with semantic information, to be used for next page prediction [12].

Y.Y. Yao et al. presented PagePrompter, a practical framework for designing an intelligent agent that dynamically gives the recommendations to the web site's users by learning from web usage data and users' behavior. Like a tour guide, the agent assists a user in navigating the web site. PagePrompter can also be used as a tool by a web site designer for improving the design of web sites, analyzing system performance, understanding user behavior, and generating an adaptive web site without changing the original web site. [13].

R.H.P. Engels et al. provide a technical solution which is aimed at improving the semantics. The CORPORUM tool set that is developed for this task exists for a set of programs that can fulfill a variety of tasks, either as 'stand-alone', or augmenting each other. The aim of the semantic web is not only to enhance the precision and recall of search, but also to enable the use of logical reasoning on web contents [14].

Yuefeng Li et al. developed an ontology mining technique for retrieving relevant information from the web. The discovered ontology consists of two parts: the top backbone and the base backbone. The former illustrates the linkage between compound classes of the ontology. The latter illustrates the linkage between primitive classes and compound classes. They set up a mathematical model to represent discovered knowledge on the ontology. They also

presented a novel method for capturing evolving patterns in order to refine the discovered ontology [15].

Benedicte Le Grand et al. presented how XML topic maps can be exploited to help users find relevant information in the Web. They showed how topic maps allow to characterise and "clean" web data through the definition of a profile; The analysis of a lattice generated by a classification algorithm - called Galois algorithm was used to evaluate the relevance of a web site with regard to a specific request on a traditional search engine. They finally explained how data on the web can be clustered, organised and visualised in different ways so as to enhance users' navigation and understanding of these documents [16].

Stefan Haustein implemented a system to provide an ontology based persistent blackboard communication mechanisms for connecting mining and application agents. Using ontologies and agent technologies enabled a simple extension of the system beyond the original purpose. The system can also be used to publish structured and massively linked data to the traditional “human readable” web using template based (X)HTML generation [17].

Sankar K. Pal et al. summarized the different characteristics of web data, the basic components of web mining and its different types, and their current states of the art. The limitations of some of the existing web mining methods and tools are enunciated, and the significance of soft computing (comprising fuzzy logic (FL), artificial neural networks (ANNs), genetic algorithms (GAs), and rough sets (RSs) highlighted. The prospective areas of web mining where the application of soft computing needs immediate attention are outlined [2].

P. Markellou et al. proposed a framework for personalized e-Learning based on aggregate usage profiles and domain ontology. They have distinguished two stages in the whole process, one is offline tasks that includes data preparation, ontology creation and usage mining and the other is online tasks that concerns the production of recommendations [18].

Baoyao Zhou1 et al. proposed a web usage mining approach for semantic web personalization. Providing semantic web personalization needs to tackle the technical issues on how to define web access activities, discover hierarchical relationships from web access activities, transform them into ontology automatically, and deduce personalized usage knowledge from the ontology. The approach incorporated fuzzy logic into Formal Concept Analysis to mine clientside web usage data for automatic ontology generation, and then applied fuzzy approximate reasoning to deduce personalized usage knowledge from the ontology [19].

Yuefeng Li et al. presented a model to automatically discover knowledge for a particular user or a group of users. They have used ontologies to represent user profiles and presented an abstract Web mining model. They described the details of using the abstract Web mining model for information gathering. They have also presented an efficient filtering algorithm to filter out most non-relevant inputs. The algorithm uses a numerical computational method to decrease the number of checking whether a set is a subset of other sets [20].

IV. CONCLUSION

In this paper, we have provided an outline of web mining, its various classifications and its subtasks. We have summarized the two fast developing research areas, Semantic Web and Web Mining. The combined area of Semantic Web Mining offers new techniques to improve both areas. The application of each area to the other creates a feedback loop, where the goal of Semantic Web Mining is realized.

A survey of the recent work in the area of semantic web mining has been made. The various applications of semantic web mining are listed and a comparison has been made between the traditional web applications and semantic web applications thereby providing guidelines for future research in the area of semantic web mining.

REFERENCES

[1]. Semantic Web Mining: State of the art and future directions, Stumme.G, Hotho. A, Berendt B, Web Semantics: Science, Services and Agents on the World Wide Web 4(2) (2006) 124 – 143 Semantic Grid – The Convergence of Technologies.

[2]. Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions, Sankar K. Pal, Varun Talwar, and Pabitra Mitra, IEEE transactions on neural networks, vol. 13, no. 5, september 2002

[3]. Towards Knowledge Discovery in the Semantic Web, Krcmar H (2004), Informations management (German Edition). Springer, Berlin.

[4]. Towards Semantic Web Mining, Berendt B, Hotho A, Stumme G (2002). ISWC 2002, First International Semantic Web Conference, Sardinia, Italy, June 9-12, 2002, Springer.

[5]. Application based semantic web mining technique, Mahindra Pratap Singh Dohare*1 and Sanjaydeep Singh Lodhi, Vinod Mahor, Volume 2, No. 3, March 2011, JGRCS.

[6]. A Roadmap for Web Mining: From Web to Semantic Web, Berendt, B., Hotho, A., Mladenic, D., van Someren, M., Spiliopoulou, M., Stumme, G. Web Mining: From Web to Semantic Web Volume 3209/2004 (2004) 1–22.

[7]. Using Semantic Web Mining Technologies for Personalied E-Learning Experiences, P. Markellou, I. Mousourouli, S. Spiros, and A. Tsakalidis (Greece),Proceeding (461) Web-based Education, 2005.

[8].Towards Knowledge Discovery in the Semantic Web, Thomas Fischer, Johannes Ruhland, MKWI 2010 – Business Intelligence.

[9]. An Overview of Web Mining in Societal Benefit Areas, Georgios Lappas, IEEE E-Commerce and E-Services(CEC-EEE 2007)

[10]. Relational Data Mining, Dzeroski S and Lavrac N (2001). Springer, Berlin.

[11]. Semantic Web Mining, Diana Cerbu, Romania Konstanz, February 21, 2008.

[12]. Using Domain Ontology for Semantic Web Usage Mining and Next Page Prediction, Nizar R. Mabroukeh and Christie I. Ezeife *CIKM’09*, November 2–6, 2009, Hong Kong, China.

[13]. PagePrompter: An Intelligent Agent for Web Navigation Created Using Data Mining Techniques, Y.Y. Yao, H.J. Hamilton, and Xuewei Wang.

[14] “CORPORUM: a workbench for the Semantic Web”- R.H.P. Engels, B.A.Bremdal and R. Jones, Halden, Norway, July 31,2001.

[15]. Mining Ontology for Automatically Acquiring Web User Information Needs, Yuefeng Li and Ning Zhong, IEEE transactions on knowledge and data engineering, vol. 18, no. 4, April 2006.

[16]. XML Topic Maps and Semantic Web Mining, Benedicte Le Grand, Michel Soto, September, 2001, 12th European Conference on Machine Learning (ECML’01).

[17]. Utilising an Ontology Based Repository to Connect Web Miners and Application Agents**,** Stefan Haustein, ECML'01 Workshop on Semantic Web Mining, Sep 2001.

[18]. Using Semantic Web Mining Technologies for Personalied E-Learning Experiences, P. Markellou, I. Mousourouli, S. Spiros, and A. Tsakalidis
(Greece), Proceeding (461) Web-based Education, 2005.

[19]. Web Usage Mining for Semantic Web Personalization, Baoyao Zhou1, Siu Cheung Hui, and Alvis C. M. Fong, 2006.

[20]. Web Mining Model and its Applications for Information Gathering, Yuefeng Li and Ning Zhong, Knowledge Based Systems, 2004.