International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307, Volume-2, Issue-3, July 2012

Web Mining using Semantic Data Mining Techniques

K.Ganapathi Babu, A.Komali, V.Mythry, A.S.K.Ratnam

Abstract— The purpose of Web mining is to develop methods and systems for discovering models of objects and processes on the World Wide Web and for web-based systems that show adaptive performance. Web Mining integrates three parent areas: Data Mining, Internet technology and World Wide Web, and for the more recent Semantic Web. Semantic Web Mining is the outcome of two new and fast developing domains: Semantic Web and Data Mining. The 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 cooperation. Data Mining is the nontrivial process of identifying valid, previously unknown, potentially useful patterns in data. Semantic Web Mining refers to the application of data mining techniques to extract knowledge from World Wide Web or the area of data mining that refers to the use of algorithms for extracting patterns from resources distributed over in the web. The aim of Semantic Web Mining is to discover and retrieve useful and interesting patterns from a huge set of web data. This web data consists of different kind of information, including web structure data, web log data and user profiles data. Semantic Web Mining is a relatively new area, broadly interdisciplinary, attracting researchers from: computer science, information retrieval specialists and experts from business studies fields. Web data mining includes web content mining, web structure mining and web usage mining. All of these approaches attempt to extract knowledge from the web, produce some useful results from the knowledge extracted and apply these results to the real world problems. This paper gives an overview of how the semantic web is used for mining the World Wide Web.

Index Terms—Data mining, Semantic web, Web mining, World Wide Web.

I. INTRODUCTION

The research area of Semantic Web Mining is aimed at combining two fast developing fields of research: the Semantic Web and Web Mining. These two fields address the current challenges of the World Wide Web (WWW): turning unstructured data into machine-understandable data using Semantic Web tools, and (semi-)automatically extract knowledge hidden in the vast amounts of Web data using Web Mining tools [1]. Semantic Web Mining is a Semantic Web tools, and (semi-)automatically extract knowledge hidden in the vast amounts of Web data using Web Mining tools [1]. Semantic Web Mining is a convergence of these two fields, where the tools of the Semantic Web can be used to improve Web Mining and vice versa. For example, in the vast quantities of data, Web Mining can discover semantic structures to build semantics for the Semantic Web. Similarly, semantic structures can improve the task of mining by allowing the algorithms to operate on certain semantic levels or choose appropriate levels of abstraction. Semantic Web Mining can then be thought of as Semantic (Web Mining) or (Semantic Web) Mining to cover the spectrum of topics [1].

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. 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 [2]. The main tools that are currently being used in the Semantic Web are ontologies based on OWL (Web Ontology Language) and its associated reasoners.

"Web Mining is the application of data mining techniques to the content, structure and usage of Web resources." [1] The three main areas of Web Mining are:

- Content Mining - Analyses the content of Web resources. Mainly based on text mining techniques, but extensions to multimedia content is beginning to emerge in the research.
- Structure Mining - Analyses the hyperlink structure between Web pages.
- Usage Mining - Analyses the user's clicks from Web server logs.

Semantic Web Mining aims at combining the two areas Semantic Web and Web Mining. This vision follows our observation that trends converge in both areas: Increasing numbers of researchers work on improving the results of Web Mining by exploiting (the new) semantic structures in the Web, and make use of Web Mining techniques for building the Semantic Web. Last but not least, these techniques can be used for mining the Semantic Web itself. The wording *Semantic Web Mining* emphasizes this spectrum of possible interaction between both research areas: It can be read both as *Semantic* (*Web Mining*) and as (*Semantic Web*) *mining*.

The aim of Semantic Web Mining is to discover and retrieve useful and interesting patterns from a huge set of web data. This web data consists of different kind of information, including web structure data, web log data, and user profiles data. This paper gives an overview of how the semantic web is used for mining the world wide web.

Manuscript received on July, 2012.

K.Ganapathi Babu Pursuing M.Tech in Computer Science Engineering at Vignan's LARA Institute Of Technology and Sceince, Vadlamudi, Guntur Dist., A.P., India.

A.Komali Pursuing M.Tech in Computer Science Engineering at Vignan's LARA Institute Of Technology and Sceince, Vadlamudi, Guntur Dist., A.P., India.

V.Mythry, Pursuing M.Tech in Computer Science Engineering at Vignan's LARA Institute Of Technology and Sceince, Vadlamudi, Guntur Dist., A.P., India.

A.S.K.Ratnam Head, Department of CSE, Vignan's LARA Institute Of Technology & Sceince,Vadlamudi Guntur Dist., A.P., India.

II. SEMANTIC WEB MINING

The Semantic Web offers to add structure to the Web, while Web Mining can learn implicit structures. The combine use of the Semantic Web and Web Mining were briefly discussed in the Introduction. There are other ways of building the Semantic Web by data mining and mining the Semantic Web as detailed in Stumme et al [1]. From these combined use of tools from both fields, a feedback loop can be setup between the two fields that automates the transformation of the Web to the Semantic Web. This is an interesting way for Semantic Web Mining to create itself as the dependence between the Semantic Web and Web Mining increases. The resulting research benefits many areas of industry such as "e-activities", health care, privacy and security, and knowledge management and information retrieval [2].

Semantic Web Mining is a relatively new area, broadly inter-disciplinary, attracting researchers from: computer science fields like artificial intelligence, machine learning, databases and information retrieval specialists and from business studies fields like marketing, administrative and e-commerce specialist and from social and communication studies such as social network analyzers etc. Web data mining includes web content mining, web structure mining, web usage mining. All of these approaches attempt to extract knowledge from the web, produce some useful results from the knowledge extracted and apply these results to the real world problems. To improve the internet service quality and increase the user click rate on a specific website, it is necessary for a web developer to know what the user really want to do, predict which pages the user is potentially interested in, and present the customized web pages to the user by learning user navigation pattern knowledge [3].

A. Taxonomy of Semantic Web Mining

Web mining is divided into three mining categories according to the different sources of data analyzed [4] as shown in Fig.1:

(1) Web content mining focus on the discovery of knowledge from the content of web pages and therefore the target data consist of multivariate type of data contained in a web page as text, images, multimedia etc.

(2) Web usage mining focus on the discovery of knowledge from user navigation data when visiting a website. The target data are requests from users recorded in special files stored in the website's servers called log files.

(3) Web structure mining deals with the connectivity of websites and the extraction of knowledge from hyperlinks of the web. It is the process of using graph theory to analyze the node and connection structure of a web site.

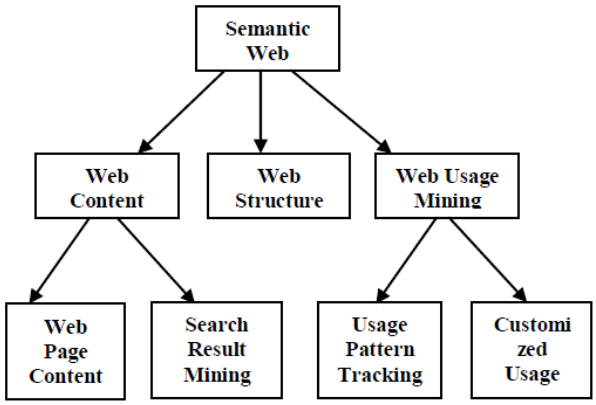

Figure 1. Semantic Web Mining Taxonomy

B. Layers of the Semantic Web

The Semantic Web is based on a vision of Tim Berners- Lee, the inventor of the WWW. Berners-Lee suggested a layer structure for the Semantic Web. This structure is shown in fig 2.

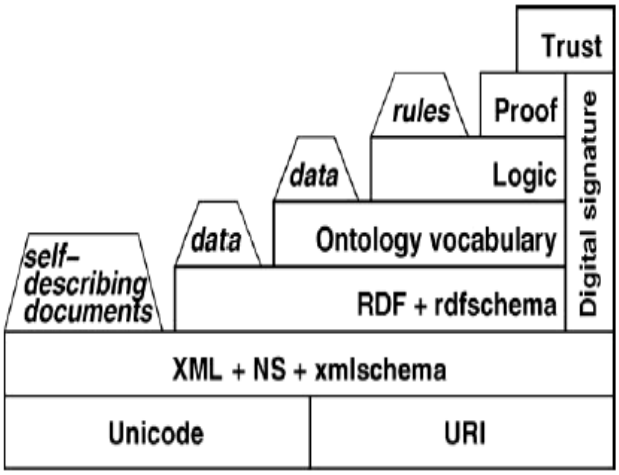

Figure. 2. The layers of the Semantic Web.

On the first two layers, a common syntax is provided. *Uniform resource identifiers* (URIs) provide a standard way to refer to entities, while *Unicode* is a standard for exchanging symbols. The *Extensible Markup Language* (XML) fixes a notation for describing labeled trees, and XML Schema allows the definition of grammars for valid XML documents. XML documents can refer to different *namespaces* to make explicit the context (and therefore meaning) of different tags. The formalizations on these two layers are nowadays widely accepted, and the number of XMLdocuments is increasing rapidly. WhileXMLis one step in the right direction, it only formalizes the structure of a document and not its content.

The *Resource Description Framework* (RDF) can be seen as the first layer where information becomes machine *understandable*: According to the W3C recommendation, RDF "is a foundation for processing metadata; it provides interoperability between applications that exchange machine understandable information on the Web." RDF documents consist of three types of entities: Resources, properties, and statements. Resources may be Web pages, parts or collections of Web pages, or any (real-world) objects which

are not directly part of the WWW. In RDF, resources are always addressed by URIs. Properties are specific attributes, characteristics, or relations describing resources. A resource together with a property having a value for that resource form an RDF statement. A value is either a literal, a resource, or another statement. Statements can thus be considered as object-attribute value triples.

The next layer is the *ontology vocabulary*. Following [5], an ontology is "an explicit formalization of a shared understand ing of a conceptualization". This high-level definition is realized differently by different research communities. However, most of them have a certain understanding in common, as most of them include a set of *concepts*, a hierarchy on them, and *relations* between concepts.

Logic is the next layer according to Berners-Lee. Today, most research treats the ontology and the logic levels in an integrated fashion because most ontologies allow for logical axioms. By applying logical deduction, one can infer new knowledge from the information which is stated explicitly. For instance, the axiom given above allows one to logically infer that the person addressed by 'URI-AHO' cooperates with the person addressed by 'URI-GST'. The kind of inference that is possible depends heavily on the logics chosen.

Proof and *trust* are the remaining layers. They follow the understanding that it is important to be able to check the validity of statements made in the (Semantic) Web, and that trust in the Semantic Web and the way it processes information will increase in the presence of statements thus validated. Therefore, the author must provide a proof which should be verifiable by a machine. At this level, it is not required that the machine of the reader finds the proof itself, it 'just' has to check the proof provided by the author.

III. TECHNIQUES FOR SEMANTIC WEB MINING

Various techniques for Semantic Web mining are web content mining, web usage mining and web structure mining.

A. Web Content Mining

A well -known problem, related to web content mining, is experienced by any web user trying to find all and only web pages that interests him from the huge amount of available pages. Current search tools suffer from low precision due to irrelevant results. Search engines aren't able to index all pages resulting in imprecise and incomplete searches due to information overload. The overload problem is very difficult to cope as information on the web is immensely and grows dynamically raising scalability issues. Moreover, myriad of text and multimedia data are available on the web prompting the need for intelligence agents for automatic mining. Agents search the web for relevant information using domain characteristics and user profiles to organize and interpret the discovery information. Agents may be used for intelligent search, for classification of web pages, and for personalized search by learning user preferences and discovering web sources meeting these preferences.

Web content mining is more than selecting relevant documents on the web. Web content mining is related to information extraction and knowledge discovery from analyzing a collection of web documents. Related to web content mining is the effort for organizing the semi-structured web data into structured collection of resources leading to more efficient querying mechanisms and more efficient information collection or extraction. This effort is the main characteristic of the "Semantic Web" [6], which is considered as the next web generation. Semantic Web is based on "ontologies", which are meta-data related to the web page content that makes the site meaningful to search engines.

B. Web Usage Mining

Web usage mining research focuses on finding patterns of navigational behavior from users visiting a website. These Patterns of navigational behavior can be valuable when searching answers to questions like: How efficient is our website in delivering information? How the users perceive the structure of the website? Can we predict user's next visit? Can we make our site meeting user needs? Can we increase user satisfaction? Can we targeting specific groups of users and make web content personalized to them? Answer to these questions may come from the analysis of the data from log files stored in web servers.

Web usage mining has become a necessity task in order to provide web administrators with meaningful information about users and usage patterns for improving quality of web information and service performance [7]. Successful websites may be those that are customized to meet user preferences both in the presentation of information and in relevance of the content that best fits the user.

C. Web Structure Mining

Web structure mining is closely related to analyzing hyperlinks and link structure on the web for information retrieval and knowledge discovery. Web structure mining can be used by search engines to rank the relevancy between websites classifying them according to their similarity and relationship between them [8]. Google search engine, for instance, is based on Page Rank algorithm [9], which states that the relevance of a page increases with the number of hyperlinks to it from other pages, and in particular of other relevant pages.

Personalization and recommendation systems based on hyperlinks are also studied in web structure mining. Web structure mining is used for identifying "authorities", which are web pages that are pointed to by a large set of other web pages that make them candidates of good sources of information. Web structure mining is also used for discovering community networks by extracting knowledge from similarity links. The term is closely related to "link analysis" research, which has been developed in various fields over the last decade such as computer science and mathematics for graph-theory, and social and communication sciences for social network analysis.

Web site personalization is the process of customizing the content and structure of a web site to the specific needs of each user taking advantage of user's navigational behavior. Recommendation systems support web site personalization by tracking user's behavior and recommending similar items to those liked in the past (Content-based filtering), or by inviting users to rate objects and state their preferences and interests so that recommendations can be offered to them based on other users rates with similar preferences (Collaborative filtering), or by asking questions to the user

and providing tailored to his needs services according to his answers (Rule-based filtering).

IV. CONCLUSION

World Wide Web has become one of the world's three major media, with the other two being print and television. E-commerce is one of the major forces that allow the web to flourish, but the success of electronic commerce depends upon how well the website developers understand user's behaviour and needs. Semantic Web Mining can be used to discover interesting user navigation patterns, which then can be applied to real world problems such as website improvement, additional topic/ product recommendations, customer behaviour's study etc. This paper gives an overview of how the semantic web is used for mining the world wide web.

REFERENCES

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

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

[3] Agrawal R. and Srikant R. (2000). Privacy-preserving data mining, In Proc. of the ACM SIGMOD Conference on Management of Data, Dallas, Texas, 439-450.

[4] Cooley, R., Mobasher, B. and Srivastava, J., (1997). Web Mining: Information and Pattern Discovery on the World Wide Web, 9th International Conference on Tools with Artificial Intelligence(ICTAI '97), New Port Beach, CA, USA, IEEE Computer Society, 558-567.

[5] T.R. Gruber, Towards Principles for the Design of Ontologies Used for Knowledge Sharing, in: N. Guarino, R. Poli (Eds.), Formal Ontology in Conceptual Analysis and Knowledge Representation, Kluwer, Deventer, Netherlands, 1993.

[6] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. Scientific American, 284(5):34-43.

[7] Spiliopoulou, M., and Pohle, C. (2001). Data Mining for Measuring and Improving the Success of Web Sites. Data Mining and Knowledge Discover, 5(1-2):85-114.

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

[9] Brin, S., and Page, L. (1998). The Anatomy of a Large-Scale Hyper textual Web Search Engine, Proceedings of the 7th International World Wide Web Conference, Elsevier Science, New York, 107-117.

K.Ganapathi Babu pursuing his (M.Tech in Computer Science Engineering) at Vignan's LARA Institute Of Technology and Sceince, Vadlamudi, Guntur Dist., A.P., India. His research interest includes Data Mining and Image Processing.

A.Komali pursuing her (M.Tech in Computer Science Engineering) at Vignan's LARA Institute Of Technology and Sceince, Vadlamudi, Guntur Dist., A.P., India. Her research interest includes Data Mining and Image Processing.

V. Mythry pursuing her (M.Tech in Computer Science Engineering) at Vignan's LARA Institute Of Technology and Sceince, Vadlamudi, Guntur Dist., A.P., India. Her research interest includes Data Mining and Image Processing.

A.S.K.Ratnam, Head, Department of CSE, Vignan's LARA Institute Of Technology & Sceince,Vadlamudi Guntur Dist., A.P., India. His research interest includes Data Mining and Image Processing.