Volume 2, No. 6, June 2011
Journal of Global Research in Computer Science
ISSN-2229-371X
RESEARCH PAPER
Available Online at www.jgrcs.info

SEMANTIC WEB MINING FOR INTELLIGENT WEB PERSONALIZATION

Anil Sharma[*1], Suresh Kumar[2], Manjeet Singh[3]

[1]M.Tech (IS) Scholar, CSE, A.I.T. - G, New Delhi, India
sharma.anil19@yahoo.com
[2]Assistant Professor, CSE, A.I.T. - G, New Delhi, India
sureshpoonia@yahoo.com
[3]Assistant Professor, CSE, YMCA US&T- F, India
mstomar2000@yahoo.com

Abstract- 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. 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. In this paper, various techniques for Semantic Web mining like web content mining, web usage mining and web structure mining are discussed. Our main focus is on web usage mining and its application in web personalization. Study shows that the accuracy of recommendation system has improved significantly with the use of semantic web mining in web personalization.

Keywords: Navigation Pattern, Pattern Analysis, Semantic Web, Web Personalization, Web Usage Mining.

INTRODUCTION

With the increase of larger and larger collection of various data resources on the World Wide Web (WWW), Semantic Web Mining has become one of the most important requirements for the web users. Web servers will store various formats of data including text, image, audio, video etc., but servers cannot identify the contents of the data. These search techniques can be improved by adding some special techniques including semantic web mining and probabilistic analysis to get more accurate results. Semantic web mining technique can provide meaningful search of data resources by eliminating useless information with mining process. 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 [1]. Data Mining is the nontrivial process of identifying valid, previously unknown, potentially useful patterns in data [2]. Semantic Web Mining refers to the application of data mining techniques to extract knowledge from World Wide Web [3] 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 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 [4].

SEMANTIC WEB MINING TAXONOMY

Web mining is divided into three mining categories according to the different sources of data analyzed [5] as shown in Fig.1:

A. 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.

B. 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.

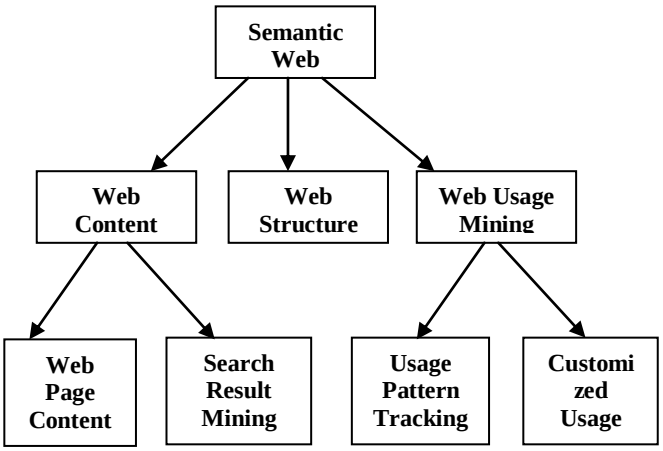

Figure 1. Semantic Web Mining Taxonomy

C. 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.

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 discov ery 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" [1], 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.

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.

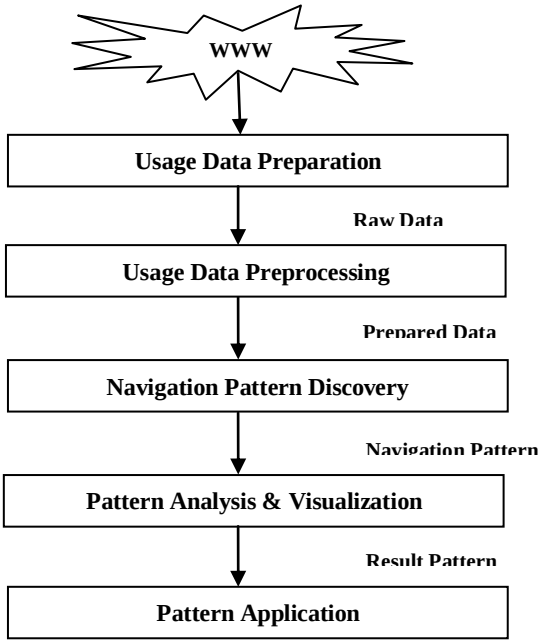

Figure 2. Web Usage Mining Architecture (Level-1)

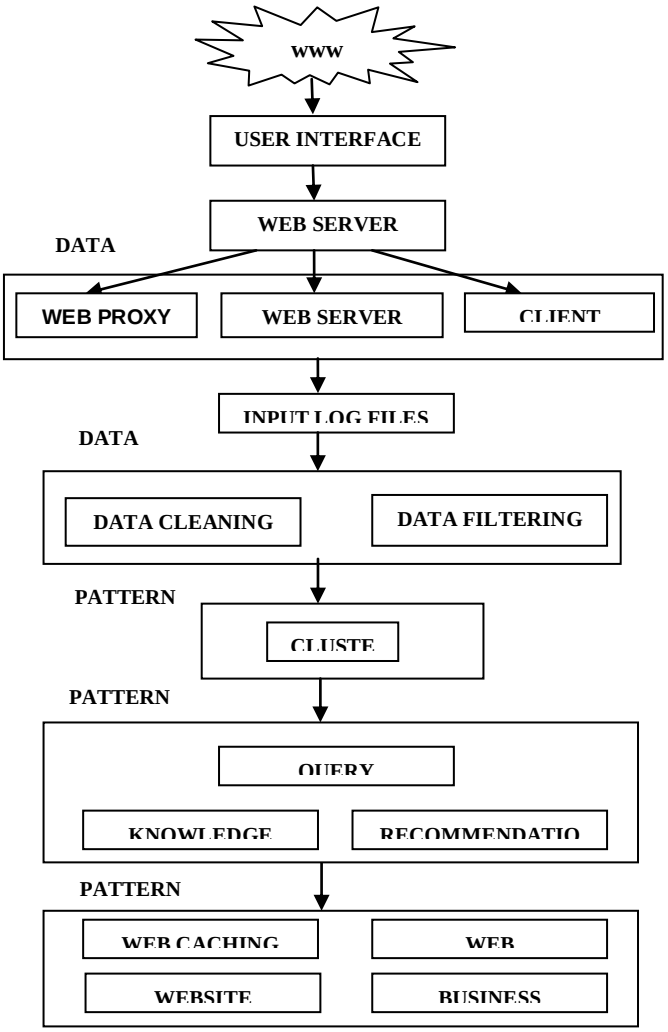

Figure 3. Web Usage Mining Architecture

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 [6, 7, 8]. 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.

A variety of implementations and realizations are employed by Web usage mining systems. **Figure-2** gives a generalized structure of the systems, each of which carries out five major tasks:

a. Usage data gathering: Web logs, which record user activities on Web sites, provide the most comprehensive, detailed Web usage data.
b. Usage data preparation: Log data are normally too raw to be used by mining algorithms. This task re-stores the users' activities that are recorded in the Web server logs in a reliable and consistent way.
c. Navigation pattern discovery: This part of a usage mining system looks for interesting usage patterns contained in the log data. Most algorithms use the
d. method of sequential pattern generation, while the remaining methods tend to be rather ad hoc.

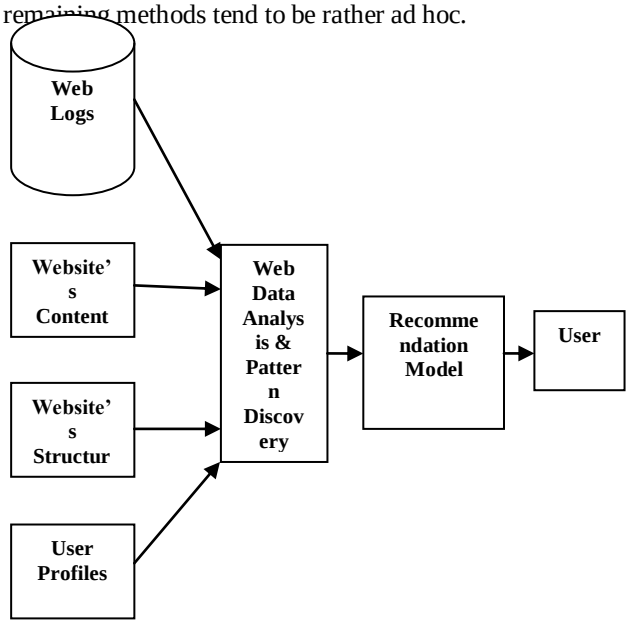

Figure 4. Web Personalization Process

e. Pattern analysis and visualization: Navigation patterns show the facts of Web usage, but these re-quire further interpretation and analysis before they can be applied to obtain useful results.
f. Pattern applications: The navigation patterns discovered can be applied to the following major areas, among others: i) improving the page/site design, ii) making additional product or topic recommendations, iii) Web personalization, and iv) learning the user or customer behavior.

Extracted usage and user behavior patterns may be used in targeting specific groups of users; in various recommendation systems; and in evaluation and reconstruction of the web site to meet design efficiency issues and user satisfaction requirements.

A. 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 [9]. Google search engine, for instance, is based on Page Rank algorithm [10], 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 [11] 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 [12, 13]

WEB USAGE MINING ARCHITECTURE

Figure-3 depicts the detailed web usage mining architecture (level-2), with various phases as follows:

a. Data Preparation: - The quality of patterns discovered in the web usage mining process highly depends upon the quality of data used in the mining process]. When the browser traces the web pages its information is stored in the server log file. Any user browser hits a URL in a domain; the information related to that operation is recorded in an access log file.
b. Data Preprocessing: - Data cleaning is also a customized step [14, 15], which includes integrating different usage logs and parsing data from these usage logs. Filtering is the most important task in web usage mining, since the quality of mined patterns depends upon this directly.
c. Pattern Discovery: - Pattern discovery is the main issue with both web usage mining & data mining. The search space increases exponentially as the lengths of patterns to be discovered increases [16, 17]. The nature of data to be clustered plays a key role when choosing the right algorithm for clustering.
d. Pattern Analysis: - This involves the interpretation of navigation patterns and other useful information extracted from web usage mining before its application to real world problems.
e. Pattern Application: - The last step in the web usage mining is the application of result to deliver Recommendations to the user. The discovered knowledge can be used for various applications such as web caching, web personalization, website improvements and business intelligence.

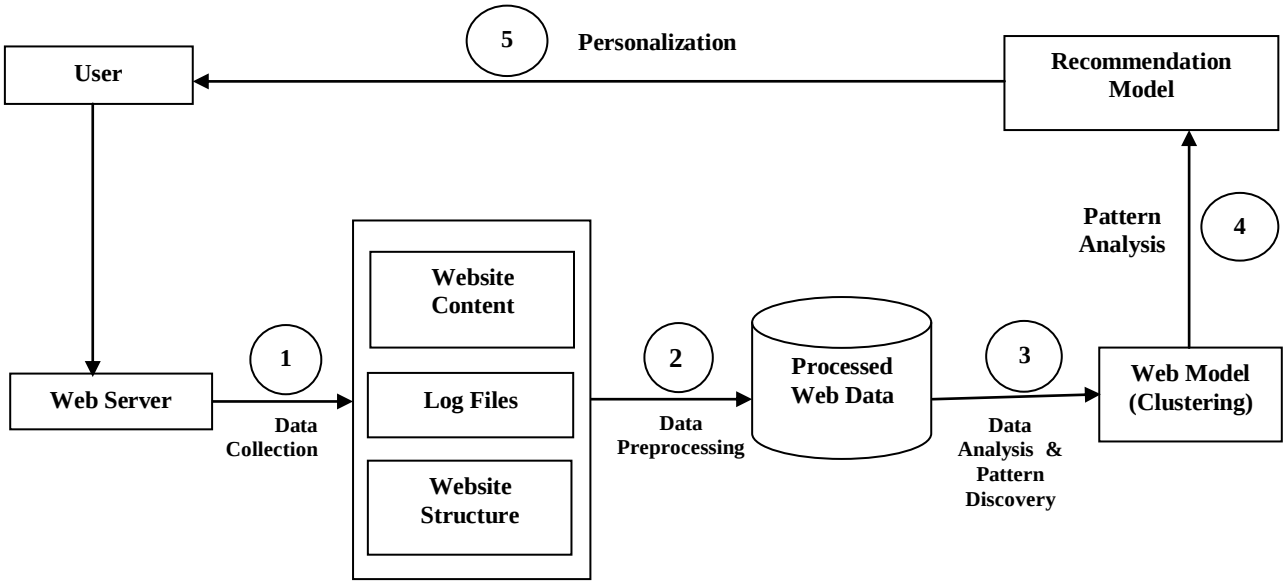

Figure 5. Web Personalization System Architecture

WEB PERSONALIZATION

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 [18]. 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).

Content-based filtering is the most common method for web personalization from server log files and has attracted considerable attention from researchers [19, 20, 21] for constructing user models that represent the behavior of users. A generalize web personalization process is depicted from figure-4. Web personalization (re)organizes Web sites/pages based on the Web experience to fit individual users' needs [22, 23]. It is a broad area that includes adaptive Web sites and recommender systems as special cases.

The Web Personalization system [24] uses a subset of Web log and session clustering techniques to derive usage profiles, which are then used to generate recommendations. Web ersonalization system architecture in figure-4 demonstrates the use of semantic web mining in web recommendation.

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 focuses the application of semantic web mining in web personalization. Web personalization system not only provides user a set of personalized pages but also gives user a list of domains the user may be interested in. Thus user can switch to different interests when surfing on the web for information. Besides this, web personalization has increased the accuracy of recommendations very significantly.

ACKNOWLEDGMENT

We want to extend our thanks to our Principal "Prof. Ashok De" and HoD "Dr. Vishal Bhatnagar" for providing us all the infrastructure and facilities to work in need and to carry out this research work.

REFERENCES

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

[2] Brachman, R.J., and Anand, T. The Process of Knowledge Discovery in Databases: A Human-Centered Approach. In Advances In Knowledge Discovery And Data Mining , eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press/The MIT Press, Menlo Park, CA. 1996, pp. 37-57.

[3] Bellaachia, A., Vommina, E., and Berrada, B. (2006). Minel: A Framework for Mining E-learning Logs, Proceedings of the 5th IASTED International Conference on Web-based Education, Puerto Vallarta, Mexico, 259-263.

[4] 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.

[5] 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.

[6] Eirinaki, M., and Vazirgiannis, M. (2003). Web Mining for Web Personalization. ACM Transactions on Internet Technology, 3(1):1-27.

[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] Wang, X., Abraham, A., and Smith, K. (2005). Intelligent Web Traffic Mining and Analysis. Journal of Network and Computer Applications, 28:147-165.

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

[10] 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.

[11] Desikan, P., Srivastava, J., Kumar, V., and Tan, P.N. (2002). Hyperlink Analysis: Techniques and Applications, Technical Report (TR 2002-0152), Army High Performance Computing Center.

[12] Park, H.W. (2003). Hyperlink Network Analysis: A New Method for the Study of Social Structure on the Web, Connections, 25(1), 49-61.

[13] Yao, Y.Y., Hamilton, H.J., and Wang, X. (2002). Page Prompter: An Intelligent Agent for Web Navigation Created Using Data Mining Techniques, Lecture Notes In Computer Science, 2475: 506-513.

[14] Junjie Chen and Wei Liu, "Research for Web Usage Mining Model", International Conference on Computational Intelligence for Modeling Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06), IEEE, vol-2, pp. 8, 2006.

[15] Junjie Chen and Wei Liu, "Research for Web Usage Mining Model", International Conference on Computational Intelligence for Modeling Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC '06): pp-8, 2006.

[16] Eirinaki, M., and Vazirgiannis, M. (2003). Web Mining for Web Personalization. ACM Transactions on Internet Technology, 3(1):1-27.

[17] Spiliopoulou, M., Pohle, C., and Faulstich, L. (1999). Improving the Effectiveness of a Web Site with Web Usage Mining, In Proceedings of WEBKDD99. San Diego, CA, USA, 142-162.

[18] Srivastava, J., Cooley, R., Deshpande, M., and Tan, P. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, 1: 12- 23.

[19] F. Masseglia, P. Poncelet, and M. Teisseire, "Using data mining techniques on web access logs to dynamically improve hypertext structure". In ACM SigWeb Letters, 8(3): 13-19, October 1999.

[20] Diamanto Oikonomopoulou, Maria Rigou, Spiros Sirmakessis, Athanasios Tsakalidis,, "Full-Coverage Web Prediction based on Web Usage Mining and Site Topology". Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04).pp. 716-719, 2004.

[21] Wolfgang, G., and Lars, S. (2000). Mining Web Navigation Path Fragments. In Workshop on Web Mining forE-Commerce (KDD2000), Boston, MA, USA, 105-110.

[22] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Automatic personalization based on Web usage mining. Communications of the ACM, 43(8):142-151, 2000.

[23] Myra Spiliopoulou. Web usage mining for site evaluation: Making a site better fit its users. Communications of ACM, 43(8):127-134, August 2000.

[24] Bamshad Mobasher, Honghua Dai, Tao Luo, and Miki Nakagawa. Discovery and evaluation of aggregate usage profiles for Web personalization. Data Mining and Knowledge Discovery, Kluwer Publishing, 6(1):61-82, Jan 2002.

