Global Journal of Computer Science and Technology
Volume 11 Issue 17 Version 1.0 October 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Knowledge Discovery from Web Logs - A Survey

By S. Chitra, Dr. B. Kalpana
Avinashilingam University, Coimbatore

Abstract - Web usage mining is obtaining the interesting and constructive knowledge and implicit information from activities related to the WWW. Web servers trace and gather information about user interactions every time the user requests for particular resources. Evaluating the Web access logs would assist in predicting the user behavior and also assists in formulating the web structure. Based on the applications point of view, information extracted from the Web usage patterns possibly directly applied to competently manage activities related to e-business, e-services, e-education, on-line communities and so on. On the other hand, since the size and density of the data grows rapidly, the information provided by existing Web log file analysis tools may possibly provide insufficient information and hence more intelligent mining techniques are needed. There are several approaches previously available for web usage mining. The approaches available in the literature have their own merits and demerits. This paper focuses on the study and analysis of various existing web usage mining techniques.

Keywords : Web Usage Mining, Personalization, Pre -processing, Web Log, Navigation Patterns.

GJCST-C Classification : H.2.8

Strictly as per the compliance and regulations of:

Knowledge Discovery from Web Logs – A Survey

S. Chitra[α] , Dr. B. Kalpana[Ω]

Abstract - Web usage mining is obtaining the interesting and constructive knowledge and implicit information from activities related to the WWW. Web servers trace and gather information about user interactions every time the user requests for particular resources. Evaluating the Web access logs would assist in predicting the user behavior and also assists in formulating the web structure. Based on the applications point of view, information extracted from the Web usage patterns possibly directly applied to competently manage activities related to e-business, e-services, e-education, on-line communities and so on. On the other hand, since the size and density of the data grows rapidly, the information provided by existing Web log file analysis tools may possibly provide insufficient information and hence more intelligent mining techniques are needed. There are several approaches previously available for web usage mining. The approaches available in the literature have their own merits and demerits. This paper focuses on the study and analysis of various existing web usage mining techniques.

Keywords : Web Usage Mining, Personalization, Pre - processing, Web Log, Navigation Patterns.

I. Introduction

THE World Wide Web (WWW), is the current era of information explosion, has become the large source of online data, which includes text, graphics, videos, sound, etc. WWW is a comprehensive information medium in which the users can read, write and communicate through the use of computers connected to the Internet. Recent studies have estimated that the Web have more than one billion pages. It has become the powerful technology for sharing new ideas and content exchange. The impact of the Internet on everyday life is tremendous and it has changed the way of doing business, providing and receiving education, organization management, etc. The manner of information collection and sharing has changed with the advancement of hardware and communication software.

The growth has motivated the web service providers to predict the user's web usage behaviors so that, they can

- Personalize the information provided to them
- Make the websites more user friendly
- Reduce the traffic load
- Create or modify their website to suit different group of people.

The current requirement focuses on some tools which will help system analysts and business persons to learn user/consumer's needs, so that user requirements or demand can be solved immediately.

Web mining is the application of data mining techniques to web-based data for the purpose of learning or extracting knowledge. The techniques in web mining focus on providing solutions to content provider, web designer and programmers to improve their website and also to the web users with navigation assistance tools. It is a part of data mining where knowledge is gained from WWW.

Web servers trace and gather information about user interactions every time the user requests for particular resources. Evaluating the Web access logs would assist in predicting the user behavior and also assists in formulating the web structure.

Web usage mining, popularly also known as web log mining, works on the secondary data like web log file, click streams to extract knowledge with regard to web usage. It is the process which uses data mining techniques abundantly, the result of which can be used for several uses like personalization, system improvement and site modification.

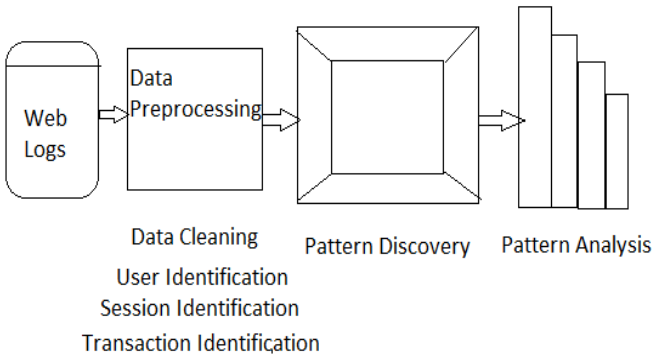
It is essential to investigate what kind of features a WUM system is estimated to have with the intention of performing effective and efficient Web usage mining, and what kind of challenges may be faced in the process of developing new Web usage mining techniques. A Web usage mining system should be able to:

- Gather useful usage data thoroughly
- Filter out irrelevant usage data
- Establish the actual usage data
- Discover interesting navigation patterns
- Display the navigation patterns clearly
- Analyze and interpret the navigation patterns correctly and
- Apply the mining results effectively.

Web usage mining comprises of three phases, specifically, preprocessing, pattern discovery and pattern analysis. Different phases of Web Usage Mining are depicted in Figure-1. There are several approaches previously available for web usage mining. This paper focuses on the study and analysis of various existing web usage mining techniques.

Author[α]: Assistant Professor, Government Arts College (Autonomous) Coimbatore - 641018.
Email : chitra.sivakumar@ymail.com
Author[Ω]: Associate Professor, Avinashilingam University for Women Coimbatore - 641043. Email : kalpanabsekar@yahoo.com

Figure 1: Phases of Web Usage Mining



II. Literature Survey

Web usage mining is a major application of data mining technology to extract the data of the Web server log file. It can find out the browsing behavior of user and a certain type of correlations among the web pages. Web usage mining offers the support for the Web site design, and also offers personalization server and further business decision making, etc. Web mining applies the data mining, the artificial intelligence and the chart technology and so on to the Web data and tracks the users browsing characteristics, and then obtains the users using pattern. Qingtian Han et al., [1] investigated on Web Mining Algorithm based on Usage Mining. And it also provides the design approach of the electronic commerce website application algorithm. This approach is easy, efficient and easy to understand, it is appropriate to the Web usage mining demand of building a low cost B2C website.

In order to enhance the Web site, it is necessary to estimate current usage of the particular website. Web usage mining and statistical analysis are two approaches to estimate usage of Web site. The integration of Web usage mining and statistical analysis provides more exact information about Web usage. With the help of Web usage mining approaches, graph mining focuses complex Web browsing patterns like parallel browsing. With the use of statistical analysis approaches, investigating the page browsing time provides precious information about Web site and its user's behavior. Heydari et al., [2] presented a Web usage mining technique which integrates Web usage mining and statistical analysis by taking into consideration of client side data. In additional way, it integrates graph based Web usage mining and browsing time examination with the consideration of client side data. It assists in rebuilding user session precisely as it has been and in accordance with these data, this Web usage patterns offers better accuracy.

Web usage mining has turn out to be very popular in different business fields associated with Web site development. In Web usage mining, frequently browsed navigational paths are obtained with the assistance of Web page addresses from the Web server visit logs, and the patterns are utilized in different applications together with recommendation. Normally the semantic information of the Web page contents is not considered in Web usage mining. Salin et al., [3] provided a structure for combining semantic information with Web usage mining. The common navigational patterns are obtained as the form of ontology instances rather than Web page addresses and the outcome is used for generating Web page recommendations to the visitor. Additionally, an assessment method is implemented with the intention of testing the accomplishment of the recommendation. Test result confirms that precise recommendations can be achieved by including semantic information in the Web usage mining.

Nasraoui at al., [4] presented a comprehensive structure and findings in mining Web usage patterns by using Web log files of a real Web site that has all the demanding aspects of real-life Web usage mining, together with evolving user profiles and external data describing an ontology of the Web content. Therefore, the authors present a technique for determining and tracing the mounting user profiles. The authors also discuss how the obtained users profiles can be improved with clear information obtained from search queries of Web log data. Profiles are also enhanced with additional domain-specific information aspects that provide a panoramic view of the discovered mass usage modes. Many experiments have been done by the author to assess the excellence of the mined profiles, especially their adaptability in the face of developing user behavior.

Web usage mining utilizes data mining methods to examine the user access of Web sites. As with any KDD (knowledge discovery and data mining) process, WUM comprises of three main phases: preprocessing, knowledge extraction, and results analysis. Tanasa at al., [5] concentrates on data preprocessing, a difficult and complicated phase. Analysts intend to find out the accurate list of users who browsed the Web site and to reconstitute user sessions-the order of actions every user carried out on the Web site. Inter-sites WUM focuses on the Web server logs from numerous Web sites, commonly belonging to the similar organization. Therefore, analysts must reconstruct the users' path through all the different Web servers that they browsed. This solution is to integrate all the log files and reconstitute the visit. Traditional data preprocessing comprises of three phases: data fusion, data cleaning, and data structurization. The author calls this solution as advanced data preprocessing. This technique comprises of a data summarization phase, which will permit the analyst to choose only the information of importance. The authors have effectively tested this technique in an experimentation with log files from INRIA Web sites.

Jianxi et al., [6] provides a Web usage mining method based on fuzzy clustering in identifying target group. Data mining is a procedure of non-trivial mining of inherent, formerly unidentified, and extremely useful data from very large quantity of data. Web mining can be defined mainly as the utilization of data mining techniques to Web data. Web usage mining is a noteworthy and fast developing area of Web mining where several researches has been carried our earlier. The author utilized the fuzzy clustering method for identifying groups that allocate comparable interests and behaviors by investigating the data gathered in Web servers.

Internet and Web technologies are extensively available, enabling it simpler for organizations to carry out business and transfer data to customers. Furthermore, they accelerate financial transactions competently by decreasing the transaction costs of commercial actions that businesses would generally incur. As a result, Internet business has generated aggressive surroundings, a flourishing organizations wanted to survive and increase a competitive advantage must offer a satisfactory package of customized services that convince customers' needs. Regardless of the Internet's apparent benefits as a novel communication medium its advertising provides the same advertising information to all customers and so has experienced from poor reactions. To increase a Web ad's usefulness, Sung Min Bae et al., [7] developed a Web ad selector with the intention of personalizing advertising information for customers according to their preferences and interests. The Web ad selection method segregates the Web site customers with comparable preferences into numerous segments through Web usage mining. It makes use of fuzzy rules that conveys customer segments' surfing patterns on the basis of specialist recommendation, and recommends suitable ads by fuzzy inference.

Wu et al., [8] recommended a Web Usage Mining technique according to the sequences of clicking patterns in a grid computing environment. Predicting user's browsing behavior is an important process of web usage mining. It can support the web designers to improve the web structure or enhance the performance of the web servers. Mining on the sequences of such clicking patterns (MSCP) can be considered as a data mining operation. MSCP is normally an expensive process because of its considerable quantity of time for computation and storage for archiving a huge quantity of information. Executing MSCP turns out to be unsuccessful or even not realistic on a computer with limited resources. The author finds out the handling of MSCP in a distributed grid computing environment and expresses its efficiency by experimental cases.

Web usage mining is a part of data mining technology to extract the data of the Web server log files. It can find out the session patterns of user and certain kinds of correlations among these Web pages. Web usage mining offers the support for the Web site design, by offering personalization server and additional business making decision. There are several session patterns accumulated in Web server log files, page attribute of the same is in Boolean quantity. With the purpose of enhancing the effectiveness of presented algorithms and decrease the time of scanning database, and consequently focusing to these aspects, Gang Fang at al., [9] proposes a double algorithm of Web usage mining in accordance with the sequence number that is appropriate for mining several session patterns. The algorithm transforms session pattern of particular user into binary, and subsequently uses up and down search approach to double generate candidate frequent itemsets. The algorithm works out support by sequence number dimension with the intention of scanning once session pattern of a particular user, which is dissimilar from conventional double search mining algorithm. in addition to this, the effectiveness of Web usage mining is competently enhanced because of this approach. The experiment result confirms that the efficiency is more rapid and more competent than the similar algorithms.

A lot of models are available and practices that analyze user behavior according to their user navigation data and use clustering algorithms to differentiate their access patterns. The navigation patterns recognized are predicted to satisfy the user's interests. Raghavendra et al., [10] modeled user behavior as a vector of the time the particular user spends at each URL, and additionally categorize a new user access pattern. The clustering and classification methods of k-means with non-Euclidean similarity measure, and artificial neural networks with consistent inputs were implemented and evaluated. Despite recognizing user behavior, this model can also be utilized as a prediction system in which it can be used to identify deviational behavior.

In Web Usage Mining (WUM), web session clustering plays a significant key part to categorize web visitors according to the user click history and comparison measure. Swarm dependent web session clustering assists in several ways to handle the web resources efficiently such as web personalization, schema modification, website modification and web server performance. Hussain et al., [11] propose a structure for web session clustering at initial level of web usage mining. The structure will cover the data preprocessing phase to organize the web log data and transform the categorical web log data into numerical data. A session vector is acquired, so that suitable comparison and swarm optimization possibly will be applied to cluster the web log data. The hierarchical cluster based technique will improve the existing web session techniques for additional structured information about the user sessions.

With the huge development of World Wide Web and e-commerce the investigation of users' navigation

patterns has developed into more significant. Predicting users' navigation behavior is a challenging subject for e-commerce enterprises. Web usage mining approaches can be utilized for modeling and predicting users' navigation patterns. In actual fact, mining users' navigation pattern is the basic approach for producing recommendations. In practice, the user interests are unpredictable, and it is complicated to follow the exact user navigation pattern. Khosravi et al., [12] proposed a technique based on naïve Bayesian method for modeling and predicting users' navigation behavior. The author has used Web server logs as source data, and carried out his experiment.

Huge volumes of data are collected automatically by Web servers and accumulated in access log files. Examination of server access data can offer important and helpful information. Web Usage Mining is the method of using data mining approaches to find the usage patterns by using the Web data and is aimed towards applications. It extracts the secondary data based on the interactions of the users throughout certain amount of Web sessions. Web usage mining contains three stages, that is, preprocessing, pattern discovery, and pattern analysis. Web usage mining has seen a huge increase of interest from the research people together with practice communities. Etminani et al., [13] applied Kohonen's SOM (Self Organizing Map) to pre-processed Web logs of Web server logs and extract frequent patterns. Experimental result of this technique confirms that this technique would be more useful for Web site owner.

In order to offer the online prediction efficiently, Shinde et al., [14] formulated a architecture for online recommendation for predicting in Web Usage Mining System. This approach provides the structural design of on-line recommendation system in Web usage mining (OLRWMS) for enhancing the exactness of classification by dealing between classifications, estimation, and provides user activities and user profile in online phase of this architecture.

Nowadays, Internet has turned out to be an essential tool for each individual, in the same way Web usage mining becomes a hotspot, which uses huge amounts of data in the Web server log and other appropriate data sets for mining analysis and achieves valuable knowledge model about usage of relevant Web site. At the moment, numerous works have to be performed with the positive association rules in Web usage mining, other than negative association rules is considerably more significant, Yang Bin at al., [15] have applied negative association rules approach to Web usage mining, in the course of the research the author have proved that the negative association rules have an additional significant role on access pattern to Web visitors, provide the mining algorithms, to resolve the deficiencies in which positive association rules are referred.

The recent development in Web technology has mounted the users and web pages at an exponential speed. The evolutionary modifications in tools have made it promising to confine the users' concentrate and communications with web applications through web server log file. Web log data is stored as text (.txt) file. Because of huge quantity amount of unrelated information in the web log, the initial log data can not be straightforwardly utilized in the web usage mining (WUM) process. As a result the preprocessing of web log data turns out to be very important. The appropriate examination of web log file is valuable to control the web sites efficiently for organizational and users' potential. Web log preprocessing is preliminary required process to enhance the quality and efficiency of the future processing of web usage mining. There are number of methods existing at preprocessing level of web usage mining. Various methods are utilized at preprocessing level like data cleaning, data filtering, and data integration. Hussain at al., [16] analyzed the existing the preprocessing methods to recognize the concerns and how WUM preprocessing can be enhanced for pattern mining and examination.

III. Problems and Directions

The main objective of web usage mining is to recognize the interesting web usage patterns. In order to recognize the interesting web usage patterns, a lot of researches are needed. They researches may focus on the following:

- To provide precise page recommendation, it is necessary to understand the browsing behavior of the user and it can be effectively done using the Machine Learning algorithms. This would definitely help in providing the accurate page recommendations according to the user needs.
- Based on the statistical analysis of the user, specifically, the amount of time user spending on a particular page, the kind of links on which the user is interested, number of visit on a particular page will help to understand the behavior of an user.
- The clustering technique can be used in grouping the user access pattern plays a significant role in determining the resemblance in used browsing sessions. Therefore, the clustering technique can be improved to enhance the performance of grouping the user sessions.

IV. Conclusion

Web log files play a significant role in the Web Usage Mining. An important knowledge that can be obtained from web log files is the user's navigation pattern. The challenge in obtaining such knowledge is that the users are constantly shifting their focus and different users have different navigational behavior with different needs associated with them. The navigation

pattern knowledge can be used to help users by predicting their future request and it will help on the personalization of websites. This paper provides the need for Web Usage Mining and various techniques which focus on the Web Usage Mining. The directions provided in this paper will assist the researchers to perform research on Web Usage Mining.

References Références Referencias

1. Qingtian Han; Xiaoyan Gao; Wenguo Wu; "Study on Web Mining Algorithm based on Usage Mining", 9th International Conference on Computer-Aided Industrial Design and Conceptual Design (CAID/CD 2008), Pp. 1121 – 1124, 2008.
2. Heydari, M.; Helal, R.A.; Ghauth, K.I.; "A graph-based web usage mining method considering client side data", International Conference on Electrical Engineering and Informatics (ICEEI '09), Vol. 1, Pp. 147 – 153, 2009.
3. Salin, S.; Senkul, P.; "Using semantic information for web usage mining based recommendation", 24th International Symposium on Computer and Information Sciences (ISCIS 2009), Pp. 236 – 241, 2009.
4. Nasraoui, O.; Soliman, M.; Saka, E.; Badia, A.; Germain, R.; "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites", IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 2, Pp. 202 – 215, 2008.
5. Tanasa, D.; Trousse, B.; "Advanced data preprocessing for intersites Web usage mining", IEEE Intelligent Systems, Vol. 19, No. 2, Pp. 59 – 65, 2004.
6. Jianxi Zhang, Peiying Zhao, Lin Shang and Lunsheng Wang, "Web Usage Mining Based On Fuzzy Clustering in Identifying Target Group", International Colloquium on Computing, Communication, Control, and Management, Vol. 4, Pp. 209-212, 2009.
7. Sung Min Bae; Sang Chan Park; Sung Ho Ha; "Fuzzy Web ad selector based on Web usage mining", Vol. 18, No. 6, Pp. 62 – 69, 2003.
8. Chih-Hung Wu, Yen-Liang Wu, Yuan-Ming Chang and Ming-Hung Hung, "Web Usage Mining on the Sequences of Clicking Patterns in a Grid Computing Environment", International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 6, Pp. 2909-2914, 2010.
9. Gang Fang; Jia-Le Wang; Hong Ying; Jiang Xiong; "A Double Algorithm of Web Usage Mining Based on Sequence Number", International Conference on Information Engineering and Computer Science (ICIECS), Pp. 1 – 4, 2009.
10. Raghavendra, P.S.; Chowdhury, S.R.; Kameswari, S.V.; "Comparative study of neural networks and k-means classification in web usage mining", International Conference for Internet Technology and Secured Transactions (ICITST), Pp. 1-7, 2010.
11. Hussain, T.; Asghar, S.; Fong, S.; "A hierarchical cluster based preprocessing methodology for Web Usage Mining", 6th International Conference on Advanced Information Management and Service (IMS), Pp. 472 – 477, 2010.
12. Khosravi, M.; Tarokh, M.J.; "Dynamic mining of users interest navigation patterns using naive Bayesian method", IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), Pp. 119 – 122, 2010.
13. Etminani, K.; Delui, A.R.; Yanehsari, N.R.; Rouhani, M.; "Web usage mining: Discovery of the users' navigational patterns using SOM", First International Conference on Networked Digital Technologies (NDT '09), Pp. 224 – 249, 2009.
14. Shinde, S.K. and Kulkarni, U.V., "A New Approach for on Line Recommender System in Web Usage Mining", International Conference on Advanced Computer Theory and Engineering, Pp. 973- 977, 2008.
15. Yang Bin; Dong Xiangjun; Shi Fufu; "Research of WEB Usage Mining Based on Negative Association Rules", International Forum on Computer Science-Technology and Applications (IFCSTA '09), Vol. 1, Pp. 196 – 199, 2009.
16. Hussain, T.; Asghar, S.; Masood, N.; "Web usage mining: A survey on preprocessing of web log file", International Conference on Information and Emerging Technologies (ICIET), Pp. 1 – 6, 2010.

This page is intentionally left blank