An Overview of Web Usage Mining

R. Suguna
Assistant Professor
Department of Computer Applications
Bannari Amman Institute of Technology
Sathyamangalam- 638 401

D. Sharmila
Professor and Head
Department of Electronics and Instrumentation Engineering
Bannari Amman Institute of Technology
Sathyamangalam- 638 401

ABSTRACT
Web Usage Mining make use of Association Rule Mining to discover the interesting pattern, identify web user behavior, predict web user expectation and improve the business strategy. Association Rule Mining is a technique of Data Mining which is used to find the relationship between the data items. In Web Usage Mining, data are stored in the web server in the form of web log files. Numerous amounts of website visitors visit the web sites. So, it is not easy to access the web log files and find the relationship among them because of the rapid growth of web log files. Some preprocessing works are needed to reduce the noisy data of web log files before applying the association rules to find the relationship between the log files. Many researchers done the variety of works on web content mining and web usage mining to improve the efficiency of the websites by providing novel methods and this paper gives an overview about the existing works done by the researchers on web usage mining.

Keywords
Web Usage Mining, Web Content Mining, Association Rule Mining

1. INTRODUCTION
Data Mining techniques are usually aimed to discover the interesting patterns or knowledge from the large set of data items. Association Rule Mining (ARM) [1] is the very important techniques in data mining used to identify the relationships in a set of transactional data item. Association Rule Mining is normally applied in a market-basket analysis to identify the frequently purchased items in a specified period of time and arrange those items together to facilitate the customers. This technique helps the organizations to maintain the existing customers and attracting the new customers in an easy way. The same method can be applied to Web Mining [2] to understand the website visitor's behavior, attracting the website visitors, personalizing the websites and enhancing the websites for business point of view.

Web Mining has divided into three categories: 1. Web Content Mining 2. Web Usage Mining (WUM) 3. Web Structure Mining. This survey paper mainly focused on Association Rule Mining in Web Usage Mining [2] [3] Web Usage Mining has sequence of steps such as data preprocessing, knowledge extraction and result analysis. Some researchers only focused on data preprocessing and some of the researchers focused on knowledge extraction and result analysis. In Web Usage Mining, user interactions with the websites are maintained in the form of web log files [4]. The web log files may reside in proxy servers, servers or in the form of cookies in browser machine. The best way of analyzing the relationships between the web logs files are in the server log files. The server contains numerous amounts of log files; it is very difficult to apply the association rules without preprocessing the log files. The result of analyzing the web log files in the server is to identify the frequently accessed web pages from the user side and maintain those web pages in a cache memory to speed up the process as well as to enhance the efficiency of the web site.

This paper is organized as follows: Section 2 is about the existing work in Data Preprocessing in Web Log files, Section 3 is about the limitations of the existing work and Section 4 is about the scope of the future work.

2. AN OVERVIEW OF THE EXISTING WORK

The author [4] has done the research on data preprocessing in web usage mining. Data preprocessing is the important process in Web Usage Mining before applying the association rule mining in the subsequent steps such as transaction identification, path analysis, association rule mining and sequential pattern mining in the web log files. The author [5] had done the work on data preprocessing in web usage mining. They presented a new algorithm called "USIA (User and Session Identification). It finds the user and session identification details. The same user is identified with the help of IP address and User ID. If the request is from the same IP address, then the algorithm concluded that the request is from same user. The session is identified based on the time in and time out period. The page request time is exceeding 30 minutes, and then the algorithm assumes that the user started the new session. In data cleaning phase unwanted data are removed to reduce the processing time of the algorithm. This data cleaning process mainly used to improve the efficiency of the algorithm. For reference, the log file which has suffix .gif, .jpeg, .jpg is removed. So, this paper mainly focused on user identification for the particular session and series of web pages viewed by the user.

This author [6] focused on grouping the customer transactions by using the clustering technique. The set of transactions in a group has some similarities, so we can easily identified the customer behaviour and the web site analyst can able to understand the customer expectation and make the website customer friendly. In other point of view, make the website is more personalized and more user friendly. The researcher used the pattern based clustering approach to group the similar type of transactions. Some measure is followed to group the transactions, for example {starting_time = morning, avg_time_page < 2, category = 3, total_time < 10 min} may

be the behavioural pattenr for grouping the transactions. The result may be the webpages of news, finance, share or email.

The author [7] dealt with two types of groups one is Web Clustering Groups which groups the relative pages from the web server log files, the second is User Clustering Groups which groups the user who refers the same type of web pages. Divisive Hierarchical Clustering Algorithm is used to group the Web Log files and User of similar type. Then the association rule mining with support and confidence measure is applied to each group to fine the relationship among them.

This author [8] focused on the first phase of Web Usage Mining called Data Pre-processing and they suggested a novel approach for feature selection based on Rough set Theory for Web Usage Mining. The problem in web Log Files is their size and unwanted data. This paper used two algorithms Quick reduct and Variable Precision Rough Set Algorithm to identify the necessary data from the web log files, the actual process of feature selection. The k-means clustering algorithm is used to segment the similar patterns before applying the above two algorithms. So the algorithms are applied only to the group of similar items to identify the feature selection. So, this technique given the optimal solution for eliminating the unwanted data in the web log files.

The author [9] mainly focused on the data pre processing step to remove the unnecessary data such as images, extra click events. Pattern discovery algorithms are used to eliminate the unwanted data from the web server log files. They taken the data from NASA website server log files and remove the unwanted data to improve the efficiency of the web log data analysing process. No specific data mining techniques are applied to web log files after pre processing. That work is open for future research workers.

The author [10] compared the algorithm Apriori and RCS (Reduced Candidate Set) to generate the frequent item set from the given set of candidate item set. They applied the algorithm only in the large set of database not in the web log files. From the comparison result the Apriori algorithm requires multiple passes of the data set to find out the support and confident count values, whereas in the case of RCS algorithm the number of passes equal to the number of items in the item set. In future the same research may applied for web log files to generate the frequent item set to improve the efficiency of the website, website personalization and enhance the business strategy of the website.

The author [11] has done the research on applying the association rule algorithms in the web log files to find the frequent item set. Rom the result the website is personalized, enhance the searching behaviour and finding the user behaviour and also improve the business strategy. They used Apriori algorithm with hash tree, Apriori Hash Tree with Fuzzy, Modified Apriori Hash Tree and Modified Apriori Hash Tree with Fuzzy and the performance and efficiency of each algorithms are monitored. From the evolution result the later one given the better efficiency to find the frequent item set in the web log files than the former things.

This author [12] focused on finding the relationship among the website link structure to facilitate the website analyst. In their former research [13] they used the freely available data mining tool Weka to find the relationship among the website link structure. But the tool not supports the file format of web log files. Some complex pre-processing is needed to convert the weblog file format into weka tool supported format .arff. So this paper, the object oriented programming language C# is used to write an algorithm to find the interesting measure and this given a best result and better efficiency.

The author [14] suggested a hybrid prediction model to improve the accuracy of the web prediction by using the classification techniques ANN, ARM, Markov and ALL- Kth model. Markov has the limitation, that only predict the trained data, whereas Kth and ANN model can predict the data beyond the training set. So, the author combined Kth model and ANN model using Dempster's rule to enhance the web page prediction. They followed ANN model for web navigation and eliminate irrelevant classes to improve prediction time and accuracy. Then ANN model is combined with Markov model using Dempster's rule to improve prediction accuracy. Finally each new model is combined with individual models to check the accuracy for web prediction.

The author [15] has done the comparative study on various sequential association rule mining algorithms with the various sequence and temporal constraints to predict the next request from the user. The result is affected based on the set of constraints. So, choosing the correct constraint given the better predictions result.

3. LIMITATIONS OF THE EXISTING WORK

From the literature survey I found that, many researchers done the work on part of the web usage mining in data mining. Majority of the researchers concentrated on data pre-processing work which is part of the Web Usage Mining. Only some of the authors concentrated on applying the association rule algorithms to find the relationship among the candidate item set. Apriori algorithm with Hash Tree and Fazzy, FP-Growth algotirhm, Reduced Candidate Set Algorithm are used to find the interesting measure of the web log files.

4. CONCLUSION AND FUTURE WORK

World Wide Web is the hot area in Computer field. Billions of people daily accessing the internet for searching some kinds of information. So, web server contains huge amount of data every day. Web Usage Mining facilitates to analyze the Web Log Files. Still some research is needed to improve the efficiency of the algorithms to facilitate the website visitors, website analyst and website personalization. The proposed work will concentrate on the above said measures in a better manner.

5. REFERENCES

[1] Agrawal R., Imielinski T., and Swami A. (1993). Mining Associations between sets of items in Massive Databases. In Proceeding of the ACM-SIGMOD International Conference on Management of Data, pp. 207-216, Washingtom D.c USA.

[2] Kosala R., Blockeel H., (2000). Web mining research: a survey. SIGKDD explorations: newsletter of the special interest group (SIG) on knowledge discovery & data mining, ACM 2(1), pp. 1–15.

[3] Cooley R., Mobasher B., & Srivastava J. (1997). Web mining: Information and pattern discovery on the World Wide Web. In Proceeding of the IEEE International Conference on Tools with AI. pp. 558-567.

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

[5] Zhang Huiying, Liang Wei.An (2004). Intelligent Algorithm of Data Pre-processing in Web Usage Mining. In Proceeding of the 5th World Congress on Intelligent Control and Automation. pp. 15-19. Hangzhou, P.R. China.

[6] Yinghui Yang and Balaji Padmanabhan. (2005). GHIC: A Hierarchical Pattern-Based Clustering Algorithm for Grouping Web Transactions. IEEE Transactions on Knowledge and Data Engineering, Vol 17, No. 9.

[7] Yi Dong, Huiying Zhang and Linnan Jiao. (2006). Research on Application of User Navigation Pattern Mining Recommentation. In Proceeding. of the 6th World Cogress on Intelligent Control and Automation. Dalian, China.

[8] Hannah Inbarani H., Thangavel K., and Pethalakshmi A. (2007). Rough Set based Feature Selection for Web Usage Mining. International Conference on Computational Intelligence and Multimedia Applications.

[9] Suneetha K. R., and Krishnamoorthi R. (2009). Identifying User Behavior by Analysizing Web Server Access Log File. IJCSNS International Journal of Computer Science and Network Security, Vol 9, No.4.

[10] Manoj Bahel and Chhay Dule. (2010). Analysis of Frequent Itemset generation process in Apriori and RCS (Reduced Candidate Set) Algorithm. International Journal of Advanced Networking and algorithms. Vol 02, Issue 02. pp. 539-543.

[11] Veeramalai S., Jaisankar S., and Kannan A. (2010). Efficient Web Log Mining Using Enhanced Apriori Algorithm with Hash Tree and Fuzzy. International. Journal of Computer Science and Information Technology (IJCSIT) Vol.2. No.4.

[12] Maja Dimitrijevic and Zita Bosnjak. (2011). Web Usage Association Rule Mining System. Interdiciplinary Journal of Information, Knowledge and Management, Vol 6.

[13] Maja Dimitrijevic and Zita Bosnjak. (2010). Discovering interesting association rules in the web log usage data. Interdisciplinary Journal of Information, Knowledge, and Management, 5,pp. 191-207.

[14] Mamoun A. Awad and Latifur R. Khan. (2007). Web Navigation Prediction Using Multiple Evidence Combination and Domain Knowledge. IEEE Transactions on Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 37, No. 6.

[15] Wang Yong Li and Zhanhuai Zhang Yang. (2005). Mining Sequential Association-Rule for Improving WEB Document Prediction. In Proceedings of the Sixth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'05).