International Journal of Computational Engineering Research||Vol, 03||Issue, 4||

A Study on User Future Request Prediction Methods Using Web Usage Mining

Dilpreet kaur[1], Sukhpreet Kaur[2]

[1,]*Master of Technology in Computer Science & Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India.*

[2,]*Assistant Professor, Department Of Computer Science & Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib,Punjab, India.*

Abstract

Web usage mining is an important type of web mining which deals with log files for extracting the information about users how to use website. It is the process of finding out what users are looking for on internet. Some users are looking at only textual data, where others might be interested multimedia data. Web log file is a log file automatically created and manipulated by the web server. The lots of research has done in this field but this paper deals with user future request prediction using web log record or user information. The main aim of this paper is to provide an overview of past and current evaluation in user future request prediction using web usage mining.

Keywords: Future request prediction, log file, Web usage mining

I. INTRODUCTION

The web is an important source of information retrieval now-a days, and the users accessing the web are from different backgrounds. The usage information about users are recorded in web logs. Analyzing web log files to extract useful patterns is called web usage mining. Web usage mining approaches include clustering, association rule mining, sequential pattern mining etc., To facilitate web page access by users, web recommendation model is needed.[10]

Web usage mining is valuable in many applications like online marketing, E- businesses etc. The use of this type of web mining helps to gather the important information from customers visiting the site. This enables an in-depth log to complete analysis of a company's productivity flow. E-businesses depend on this information to direct the company to the most effective Web server for promotion of their product or service.[12]In this paper we did literature survey on user future request prediction in web usage mining. The paper gives the overview of various methods of user future request prediction. The advantages and disadvantages of these methods have also been discussed. The rest of the paper is organized as below. Section 2 presents the motivation of paper, Section 3 presents Literature Survey on users next request prediction, and Section 4 gives the conclusion and Future Work.

II. MOTIVATION

With the growing popularity of the World Wide Web, A large number of users access web sites in all over the world. When user access a websites, a large volumes of data such as addresses of users or URLs requested are gathered automatically by Web servers and collected in access log which is very important because many times user repeatedly access the same type of web pages and the record is maintained in log files. These series of accessed web pages can be considered as a web access pattern which is helpful to find out the user behavior. Through this behavior information, we can find out the accurate user next request prediction that can reduce the browsing time of web page thus save the time of the user and decrease the server load.In recent years, there has been a lot of research work done in the field of web usage mining „ Future request prediction". The main motivation of this study is to know the what research has been done on Web usage mining in future request prediction.

III. LITERATURE SURVEY

The focus of the literature survey is to study or collecting information about web usage mining which is used to find out web navigation behavior of user and collecting the information about "User Future Request Prediction" approach which is used to predict the next request of the user. Alexandras Nanopoulos, Dimitris Katsaros and Yannis Manolopoulos[1] focused on „web pre-fetching" because of its importance in reducing user perceived latency present in every web based application. From the web popularity, there is heavy traffic in the internet and the result is that there is delay in response.

The reasons of delay are the web servers under heavy load, Network congestion, Low bandwidth, Bandwidth underutilization and propagation delay. The solution is to increase the bandwidth but this is not proper solution because of economic cost. For that propose, this technique proposed in which reducing the delay of client future requests for web objects and getting that objects into the cache in the background before an explicit request is made for them. Architecture shows how web server could cooperate with a pre-fetch engine to disseminate hints every time a client request to a document of the server. In this paper author presented important factors which affects on web pre-fetching algorithm like order to dependencies between web document accesses and the interleaving of requests belonging to patterns with random ones within user transactions and the ordering of requests. WM0 (Ordered Web Mining) algorithm it compares with previously proposed algorithm like PPM, DG and existing approaches from the application of web log mining to web pre-fetching and the author got a result WM0 achieved large accuracy in prediction with quite low overhead in network traffic.[1]

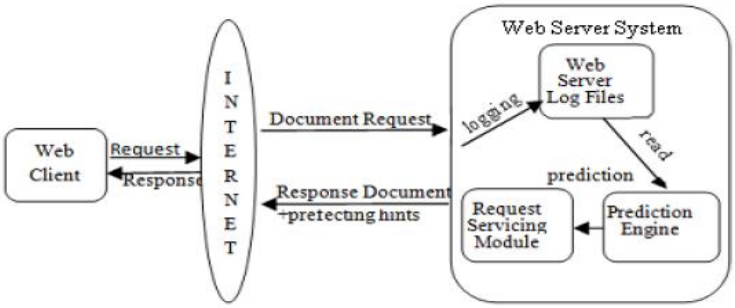

Figure 1 Proposed architecture of a prediction enabled Web server[1]

Yi-Hung Wu and Arbee L. P. Chen,[2] proposed user behaviors by sequences of consecutive web page accesses, derived from the access log of a proxy server. Moreover, the frequent sequences are discovered and organized as an index. Based on the index, they propose a scheme for predicting user requests and a proxy based framework for prefetching web pages. They perform experiments on real data. The results show that their approach makes the predictions with a high degree of accuracy with little overhead. In the experiments, the best hit ratio of the prediction achieves 75.69%, while the longest time to make a prediction only requires 1.9ms.The disadvantage of this experiment is that the average service rate is very low. The other problem is the setting of the three thresholds used in the mining stage. Thesethresholds have great impacts on the construction of the pattern trees. The use of minimum support and minimum confidence is to prune the useless paths. Obviously, some information may be lost if the pruning effects are overestimated. On the other hand, the grouping confidence is only useful for the strongly related web pages due to some editorial techniques, such as the embedded images and the frames.

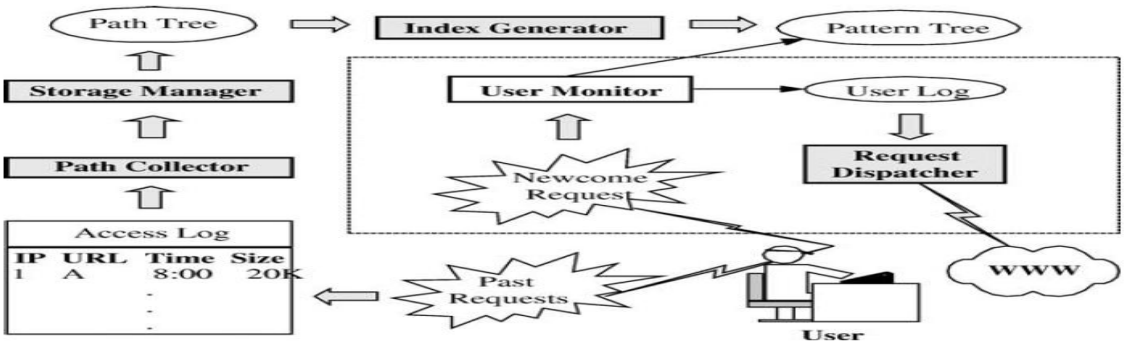

Figure 2 The flowchart of prediction system using proxy server log.[2]

According to Mathis Gery & Hatem Huddad,[3] Author distinguished three web mining approaches that exploit web logs: Association Rules (AR), Frequent Sequences (FS) and Frequent Generalized Sequences (FGS). Algorithm for three approaches were developed and experiments have been done with real web log data. Association Rule: In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large database. Describes analyze and present strong rules discovered in database using different measures of interestingness.

In [3] The problem of finding web pages visited together is similar to finding associations among item sets in transaction databases. Once transaction have been identified each of them could represent a basket and each research an item. Frequent Sequences: The attempt of this technique is to discover time ordered sequences of URLs that have been followed by past users. Frequent Generalized Sequences (FGS): a generalized sequence is a sequence allowing wildcards in order to reflect the users navigation in a flexible way. In order to extract frequent generalized subsequences they have used the generalized algorithm proposed by Gaul.

Author performed some experiments for this purpose they used three collections of web log datasets. One weblog dataset for small web site, another for large website and the third weblog dataset for intranet website. By using above three web mining approaches they evaluate the three different types of real web log data and they found Frequent Sequence (FS) gives better accuracy than AR and FGS.[3]Siriporn Chimphlee, Naomie Salim, Mohd Salihin, Bin Ngadiman , Witcha Chimphlee [4] proposed a method for constructing first-order and second-order Markov models of Web site access prediction based on past visitor behavior and compare it association rules technique. In these approaches, sequences of user requests are collected by the session identification technique, which distinguishes the requests for the same web page in different browses. In this experiment, the three algorithms first-order Markov model, second-order Markov and Association rules are used. These algorithms are not successful in correctly predicting the next request to be generated. The first-order Markov Model is best than other because it can extracted the sequence rules and choose the best rule for prediction and at the same time second-order decrease the coverage too. This is due to the fact that these models do not look far into the past to discriminate correctly the difference modes of the generative process.

Christos Makris, Yannis Panagis, Evangelos Theodoridis,and Athanasios Tsakalidis in [5] Proposed a technique for predicting web page usage patterns by modeling users' navigation history using string processing techniques, and validated experimentally the superiority of proposed technique. In this paper weighted suffix tree is used for modeling user navigation history. The method proposed has the advantage that it demands a constant amount of computational effort per user action and consumes a relatively small amount of extra memory space.

Vincent S. Tseng,Kawuu Weicheng Lin, Jeng-Chuan Chang in [6]Propose a novel data mining algorithm named *Temporal N-Gram* (*TNGram*) for constructing prediction models of Web user navigationby considering the temporality property in Web usage evolution. In this three kinds of new measuresSupport-based Fundamental Rule Changes, Confidence-based Fundamental Rule Changes, and Changes of Prediction Rules are proposed for evaluating the temporal evolution of navigationpatterns under different time periods. Through experimental evaluation on both of real-life and simulated datasets, the proposed *TN-Gram* model is shown to outperform otherapproaches like N-gram modeling in terms of prediction precision, in particular when the web user's navigating behavior changes significantly with temporal evolution.

Mehrdad Jalali, Norwati Mustapha, Md. Nasir Sulaiman, Ali Mamat in [7] Proposed a recommendation system called WebPUM, an online prediction using Web usage mining system for effectively provide online prediction and propose a novel approach for classifying user navigation patterns to predict users' future intentions. The approach is based on the new graph partitioning algorithm to model user navigation patterns for the navigation patterns mining phase. LCS algorithm is used for classifying current user activities to predict user next movement. The architecture of WEBPUM is divided into two parts:-

[1] Offline phase This phase consists two main modules, which are data pretreatment and navigation pattern mining. Data pretreatment module is designed to extract user navigation sessions from the original Web user log files. A new clustering algorithm based on graph partitioning is introduced for navigation patterns mining.
[2] Online phase The main objective of this phase is to Classifying the user current activities based on navigation patterns in a particular Web site, creating a list of recommended Web pages as prediction of user future movement. The main online component is the prediction engine.

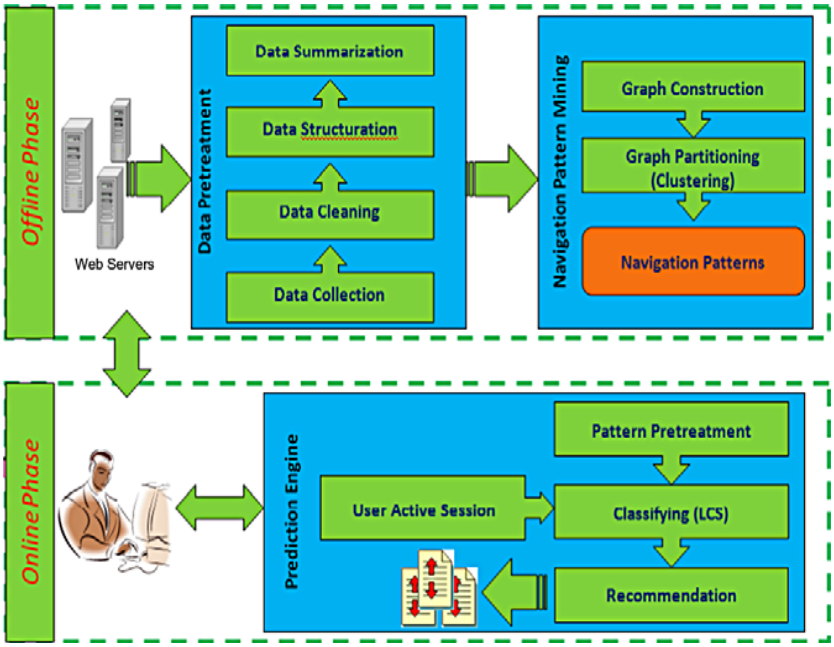

Figure 3 Architecture of WebPUM.[7]

Chu-HuiLee , Yu-lung Lo, Yu-Hsiang Fu [8] propose an efficient prediction model, two-level prediction model (TLPM), using a novel aspect of natural hierarchical property from web log data. TLPM can decrease the size of candidate set of web pages and increase the speed of predicting with adequate accuracy. The experiment results prove that TLPM can highly enhance the performance of prediction when the number of web pages is increasing.

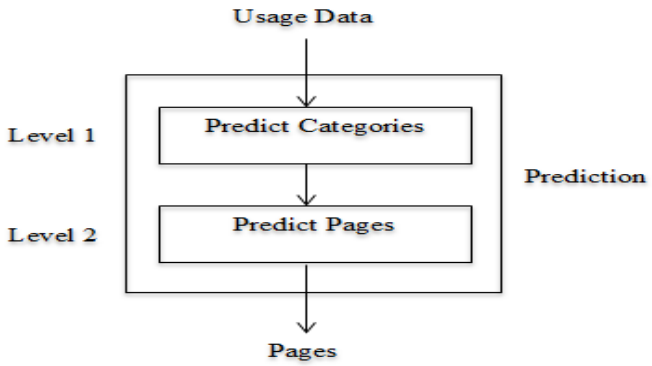

Figure 4 Two level prediction model(TLPM).[8]

In the TLPM [8] , in level one, Markov model is used to predict the next possible category which will be browsed by the user . Inlevel two, Bayesian theorem is used to predict the next possible page which belongs to the predicted category of level one to archive the goal ofreducing prediction scope more efficiently through the two-level framework. The experiment result proves that TLPM can archive the purpose and improve the efficiency of prediction by the way of finding out the important category in level one and decreasing the candidate page set in level two. Finally, the prediction result of TLPM can be applied in pre-fetching and caching on web site, personalization, target sales, improving web site design,etc.

V. Sujatha, Punithavalli [9] proposed the Prediction of User navigation patterns using Clustering and Classification (PUCC) from web log data. In the first stage PUCC focuses on separating the potential users in web log data, and in the second t stage clustering process is used to group the potential users with similar interest and in the third stage the results of classification and clustering is used to predict the user future requests.

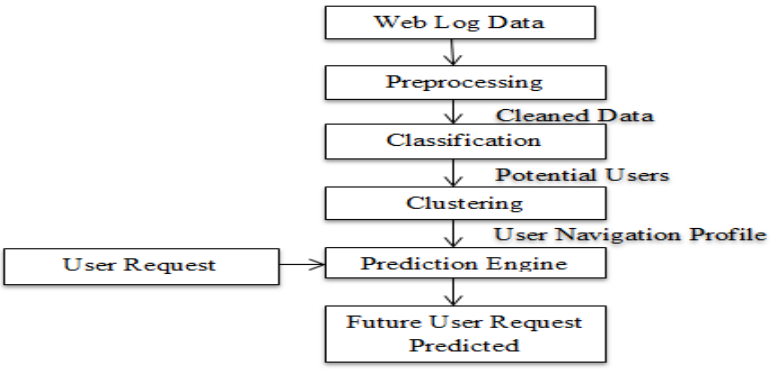

Figure 5 PUCC Model.[9]

The first stage is the cleaning stage, where unwanted log entries were removed. In the second stage, cookies were identified and removed. The result was then segmented to identify potential users. From the potential user, a graph partitioned clustering algorithm was used to discover the navigation pattern. An LCS classification algorithm was then used to predict future requests.[9]

Trilok Nath Pandey, Ranjita Kumari Dash , Alaka Nanda Tripathy ,Barnali Sahu [11] proposed IMC(Integrating Markov Model with Clustering) approach for user future request prediction. In this paper author presented the improvement of markov model accuracy by grouping web sessions into clusters. The web pages in the user sessions are first allocated into categories according to web services that are functionally meaning full. Then k-means clustering algorithm is implemented using the most appropriate number of clusters and distance measure. Markov model techniques are applied to each cluster as well as to the whole data set. The advantage of this approach is that it improves the accuracy of lower order markov model and disadvantage of this method is that it reduce the state space complexity of higher order markov model.

IV. CONCLUSION AND FUTURE WORK

The conclusion based on the literature survey is that Various research had done on future request prediction approach. In existing research various algorithms of pattern discovery techniques like graph partition techniques of clustering, LCS and Naive Bayesian techniques of classification etc are used for user future request prediction and many types of models are developed for prediction. In future prediction can be improved by using different techniques of data mining pattern discovery like classification, clustering, association rule mining etc.

REFERENCES

[1] Alexandros Nanopoulos, Dimitris Katsaros and Yannis Manolopoulos "Effective prediction of web-user accesses: A data mining approach," in Proc. Of the Workshop WEBKDD, 2001.

[2] Yi-Hung Wu and Arbee L. P. Chen, "Prediction of Web Page Accesses by Proxy Server Log" World Wide Web: Internet and Web Information Systems, 5, 67–88, 2002.

[3] Mathias Gery, Hatem Haddad "Evaluation of Web Usage Mining Approaches for User"s Next Request Prediction" WIDM"03 Proceedings of the 5th ACM international workshop on web information and data management p.74-81, November 7-8,2003.

[4] Siriporn Chimphlee, Naomie Salim, Mohd Salihin, Bin Ngadiman , Witcha Chimphlee "Using Association Rules and Markov Model for Predict Next Access on Web Usage Mining" © 2006 Springer.

[5] Christos Makris, Yannis Panagis, Evangelos Theodoridis,and Athanasios Tsakalidis "A Web-Page Usage Prediction Scheme Using Weighted Suffix Trees" © Springer-Verlag Berlin Heidelberg 2007.

[6] Vincent S. Tseng,Kawuu Weicheng Lin, Jeng-Chuan Chang "Prediction of user navigation patterns by mining the temporal web usage evolution" © Springer-Verlag 2007.

[7] Mehrdad Jalali, Norwati Mustapha, Md. Nasir Sulaiman, Ali Mamat, "WebPUM: A Web-based recommendation system to predict user future movements" Expert Systems with Applications 37 , 2010.

[8] Chu-Hui Lee , Yu-lung Lo, Yu-Hsiang Fu, "A novel prediction model based on hierarchical characteristic of web site", Expert Systems with Applications 38 , 2011.

[9] V. Sujatha, Punithavalli, "Improved User Navigation Pattern Prediction Technique From Web Log Data", Procedia Engineering 30 ,2012.

[10] A. Anitha, "A New Web Usage Mining Approach for Next Page Access Prediction", International Journal of Computer Applications, Volume 8– No.11, October 2010

[11] TrilokNathPandey, Ranjita Kumari Dash , Alaka Nanda Tripathy ,Barnali Sahu, "Merging Data Mining Techniques for Web Page Access Prediction: Integrating Markov Model with Clustering", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 1, November 2012.

[12] http://www.webdatamining.net/usage/