IJITKMSpecial Issue (ICFTEM-2014) May 2014 pp. 33-37 (ISSN 0973-4414)

A Review Paper on Web Usage Mining and future request prediction

Priyanka Bhart[1], Dr.SonaMalhotra[2]
[1]M.Tech. , CSE Department, U.I.E.T. Kurukshetra University, Kurukshetra, India
[2]HOD , CSE Department, U.I.E.T. Kurukshetra University, Kurukshetra, India
[1]priyankabhart@gmail.com , [2]Sona_malhotra@yahoo.com

Abstract:-Web usage mining is the application of data mining techniques to web log files in order to extract the useful patterns. The Web usage mining includes the data from the web server logs, poxy server logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, user profiles, registration data and any other data as the results of interactions.With the continued growth and proliferation of Web services and Web based information systems, the volumes of user data have reached astronomical proportions. Analyzing such data using Web Usage Mining can help to determine the visiting interests or needs of the web user. Lots of research has been done in this field but this paper deals with user future request prediction using web log record or user information. This paper gives the overview of various methods of future request prediction.
Keywords: Web usage mining, future request prediction, proxy server logs, browser logs, user sessions.

1. Introduction

The web has become the world's largest knowledge repository. The popularity of WWW is rapidly developing and is a golden mount with a lot of valuable information. Extracting the knowledge from the web efficiently and effectively is becoming a tedious process.It is an important type of web mining which deals with log files for extracting information about user how to use website inorder to understand user browsing behaviour. As due to the growing demand of web more and more organizations rely on the internet and the World Wide Web to conduct the business. The use of this type of web mining helps to gather the important information from customers visiting the site which can be applied for the complete analysis of a company's productivity, target the most effective web server for the promotion of their product and services..In this paper literature survey on user future request prediction in web usage mining is done. 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 review on users next request prediction, and Section 4 gives the conclusion.

2. Motivation

Popularity of the World Wide Web is growing day by day which led to the increase in the number of users accessing websites in all over the world. Whenever any user access a website a large amount of data related to that user such as its IP address, URL requested are gathered automatically by servers and collected in access log files which is very important because many times user repeatedly access the same type of web page and the record is maintained in log files. Due to the incremental nature of web log, the conventional data mining techniques were proved to be inefficient. Web access pattern which is the series of accessed web pages helps to find out the user behaviour information. Though this behaviour information helps to predict accurately the user next request prediction which can reduce the browsing time of the user and decrease the server load. The main motivation of this study is to know what research has been done on web usage mining in future request prediction.

3. Literature Review

The focus of the literature survey is to study and collecting information about web usage mining which will be used to understand user navigation behaviour which helps in predicting user next request.AlexandrasNanopoulos et al. [3] focused on "web pre-fetching" because of its importance in reducing user perceived latency present in every web based application. Due to the increasing web popularity, there is heavy traffic in the internet which result in the delay of response. The reasons of delay are the web servers are 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 as it is noteconomical. So a technique was proposed for reducing the delay of client future requests for web objects by getting those objects into the cache in the background before an explicit request is made for them. Thearchitecture shows that the prediction engine is implemented by exchange of messages between the server and clients, having the server piggybacking information about the predicted resources onto regular response messages, avoiding establishment of any new TCP connections .They assumed that there is a system implementing a

server-based predictive prefetcher, which piggybacks its predictions as hints to its clients. 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.

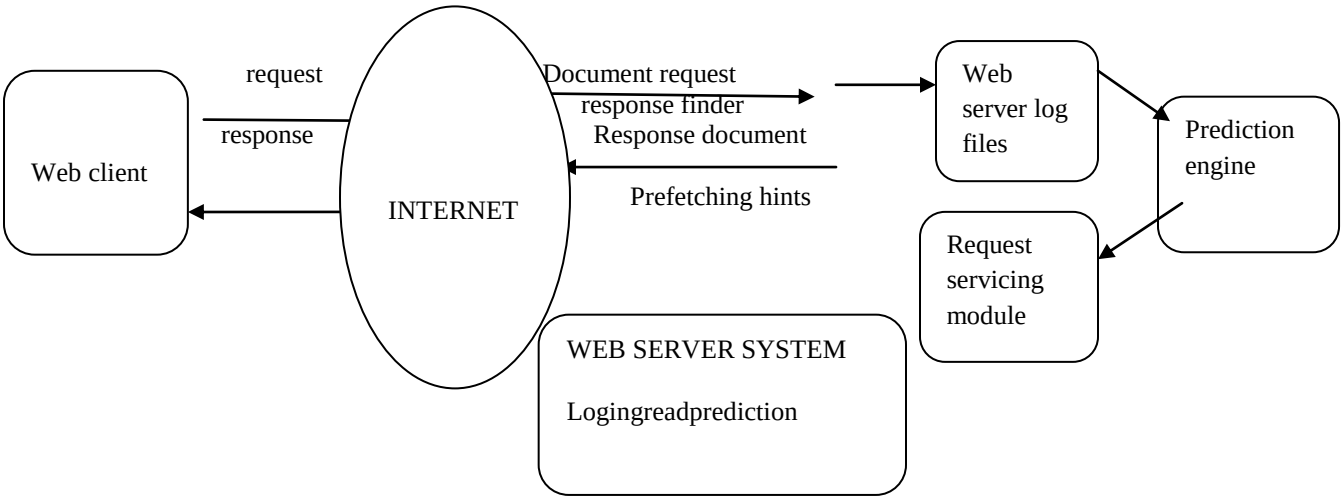

Fig.1 Proposed architecture of a prediction enabled Web server.

Yi-Hung Wu et al. [4] defined user behaviors by sequences of consecutive web page accessed by him and derived from the access log of a proxy server.The frequent sequences are identified and organized as an index. Based on these indexes, they propose a scheme for predicting user requests and a proxy based framework for prefetching web pages.

According to Mathis Gery&HatemHuddad,[5] Author compares three web mining approaches that exploit web logs: Association Rules (AR), Frequent Sequences (FS) and Frequent Generalized Sequences (FGS). Algorithms for three approaches were developed and experiments have been done with real web log data. In this paper, they have presented and took experiments on a framework for a recommender system that predicts the user's next requests based on their behaviour discovered from Web Logs data. They had combined three Web Usage Mining approaches (association rules, frequent sequence rules and frequent generalized sequence rules) with two prediction strategies, and we have evaluated these combinations using three collections of real Web usage dataAuthor 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.

Vincent S. Tseng et al. [6] Proposed a novel data mining algorithm named Temporal N-Gram (TNGram) for constructing prediction models of Web user navigation by considering the temporality property in Web usage evolution. In this three kinds of new measures Support-based Fundamental Rule Changes, Confidence-based Fundamental Rule Changes, and Changes of Prediction Rules are proposed for evaluating the temporal evolution of navigation patterns under different time periods. As the user's Web usage patterns may change with time, i.e., a visitors may have different behavior on the same Web at different interval of time. It is important to consider the temporal evolution of Web usage patterns feature in order to construct an effective prediction models for user navigation patterns. The process of prediction consist of Three steps , first step is data preprocessing for cleaning the dataset and converting the Web log into a session file that contains a click-stream of page-views for each visitor. The second step is pattern discovery. The third step is pattern analysis that predicts the user's next request. Christos Makris, YannisPanagis, EvangelosTheodoridis,andAthanasiosTsakalidis in [6] Proposed a technique by modeling users' navigation history for predicting web page usage patterns using string processing techniques, and the superiority of proposed technique was validated experimentally. In this paper for modeling user navigation history weighted suffix tree is used. The method proposed has the benefit that it demands a constant amount of computational effort per user action and it consumes a relatively small amount of extra memory space.

Chu-Hui Lee et al. [7] proposed an efficient prediction model, two-level prediction model (TLPM), using natural hierarchical property from web log data. TLPM decreases the size of candidate set of web pages and increases the speed of predicting with adequate accuracy. The experiment results proves that TLPM can highly enhance the performance of prediction when the number of web pages is increasing .In the TLPM, works in levels, in level one,

the next possible category which will be browsed by the user is predicted by using markov model . In level two, the next possible page which belongs to the predicted category of level one is predicted using Bayesian theorem .The experiment result proves that TLPM can archive the goal 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 resultof Bayesian theorem is used to prediction of TLPM can be applied in pre-fetching and caching on web site, personalization, target sales, improving web site design, etc.

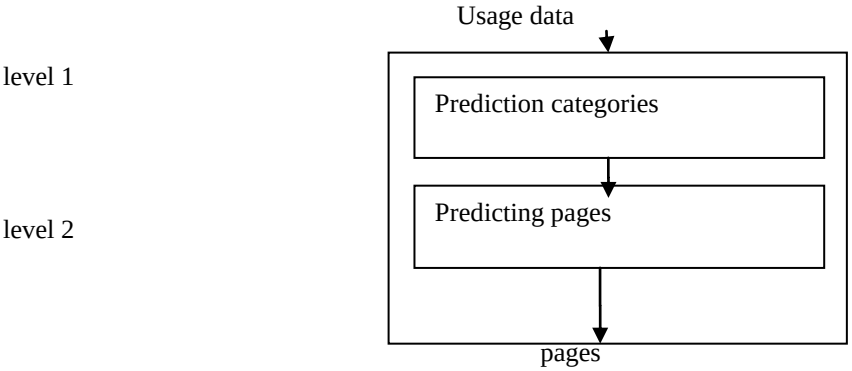

Fig.2 two level prediction model

Lee et al. [8] in their paper used the hierarchical agglomerative clustering to cluster user browsing behaviors due to the heterogeneity of user browsing features. The prediction results by two levels of prediction model framework work well in general cases. However, two levels of prediction model suffer from the heterogeneity user's behavior. So they have proposed a prediction model which decreases the prediction scope using two levels of framework, this prediction model is designed by combining Markov model and Bayesian theorem. Here categories of web pages are predicted in level one and prediction of web page is done in level two. To improve the hit ratio for prediction, it is worth to classify the different user behavior.

To effectively provide online prediction MehrdadJalali et a. [9] Proposed a recommendation system called WebPUM, an online prediction using Web usage mining system for effectively provide online prediction and proposed a approach to predict online future intentions of users by classifying user navigation patterns . The approach is based on the new graph partitioning algorithm to model user navigation patterns for the navigation patterns mining phase and longest common subsequences algorithm is used to classify current user activities to predict user next movement.
The architecture of WEBPUM is divided into two phases:-

A. Offline phase: This phase consist of two main modules. In first module is Data pretreatment module where user navigation sessions from the original Web user log files is done. Second is navigation pattern mining, where a new clustering algorithm based on graph partitioning is introduced for navigation patterns mining.

B. Online phase: Goal 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.
TrilokNathPandey et al. [10] proposed IMC(Integrating Markov Model with Clustering) approach for user future request prediction. In this paper author improvemedthemarkov model accuracy by grouping web sessions into clusters. Firstly the categorization of web pages in the user sessions is done according to the functionality of web services and usefulness .Then clustering is done using k-means clustering algorithm. Lastly Markov model techniques are applied to each cluster as well as to the whole data set. Advantage of this approach is, accuracy of lower order markov model is improved and disadvantage of this method is that the state space complexity of higher order markov model is it reduced.
V.V.R. MaheswaraRao et al.[11] in their paper "An efficient hybrid predictive model to analyze the visiting characteristics of web user using web usage mining**"** introduced an efficient hybrid predictive model, which is a combination of Markov model and Bayesian theorem. This two stage predictive model to enables the web miner to identify and analyze web user navigation patterns. In this model, the Markov model helps to reduce the operations scope by filtering possible categories which is a compact way of representing a collection of sessions and Bayesian

theorem improves accuracy in predicting the web pages in identified category. This model is used to enable the identification of user navigation patterns and also used to foresee the next link choice of a user. It will help to reduce the operation scope in Stage two just in some specific categories instead of all. After that, the web pages in suitable categories are predicted by Bayesian theorem in the Stage two of prediction model. It is expected that the two Stages of prediction model can reduce the operation scope and increase the accuracy precision.

V Sujatha et al. [12] proposed system whose main aim is to Predicting User navigation patterns using Clustering and Classification from web log data (PUCC) is to predict user navigation patterns using knowledge from (i) Potential users from web log data are identified by classification process and (ii) a clustering process groups potential users with similar interest and (ii) results of classification and clustering, predict future user requests. The heart of the PUCC system is the web log data, which stores all the successful hit made in the Internet. The first step of PUCC is the pre-processing of web log data, where the unformatted log data is converted into a form that can be directly applied to mining process. Identification of potential user focuses on separating the potential users from others. They have used a graph partitioned clustering algorithm to group users with similar navigation pattern. An undirected graph based on the connectivity between each pair of web pages is used. Each edge in the graph is assigned a weight, which is based on the connectivity time and frequency. Connectivity Time measures the degree of visit ordering for each two pages in session. The main objective of prediction engine in this part of architecture is to classify user navigation patterns and predicts users future requests.

A.Anitha et al.[13] in the paper "A new web usge mining approach for next page access prediction" proposed a new web usage mining approach is proposed to predict next page access. It is proposed to identify similar access patterns from web log using pair-wise nearest neighbor based clustering and then sequential pattern mining is done on these patterns to determine next page accesses. The tightness of clusters is improved by setting similarity threshold while forming clusters. In traditional recommendation models, clustering by non-sequential data decreases recommendation accuracy. In this paper it is proposed to integrate Markov model based sequential pattern mining with clustering. A variant of Markov model called dynamic support pruned all kth order Markov model is proposed in order to reduce state space complexity. Mining the web access log of users of similar interest provides good recommendation accuracy. Hence, the proposed model provides accurate recommendations with reduced state space complexity.

4. Conclusion

The conclusion based on the literature survey is that variousresearches had done on future request prediction approach. In present researches various algorithms of pattern discovery techniques like Markov model, 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 for prediction are developed.

References

[1] Yan Wang "Web Mining and Knowledge Discovery of Usage Patterns", 2000.

[2] R.Cooley, B. Mobasher, and J. Srivastava"Data Preparation For Mining World Wide Web Browsing Patterns",1999.

[3] AlexandrosNanopoulos, DimitrisKatsaros and YannisManolopoulos "Effective prediction of web-user accesses: A data mining approach," in Proc. Of the Workshop WEBKDD, 2001.

[4] 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.

[5] 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 datamanagement p.74-81, November 7-8,2003.

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

[7] Christos Makris, YannisPanagis, EvangelosTheodoridis,andAthanasiosTsakalidis "A Web-Page Usage Prediction Scheme Using Weighted Suffix Trees" © Springer-Verlag Berlin Heidelberg 2007.

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

[9] Chu-Hui Lee, Yu-Hsiang Fu "Web Usage Mining based on Clustering of Browsing Features**"** Eighth International Conference on Intelligent Systems Design and Applications,IEEE, 2008.

[10] MehrdadJalali, Norwati Mustapha, Md. NasirSulaiman, Ali Mamat, "WebPUM: A Web-based recommendation system to predict user future movements" Expert Systems with Applications 37 , 2010.

[11] TrilokNathPandey, RanjitaKumariDash ,Alaka Nanda Tripathy ,BarnaliSahu, "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] V.V.R. MaheswaraRao, Dr. V. ValliKumari"An Efficient Hybrid Predictive Model to Analyze the Visiting Characteristics of Web User using Web Usage Mining" 2010 International Conference on Advances in Recent Technologies in Communication and Computing IEEE.