ISSN:0975-9646

Web Usage Mining Approaches for User's Request Prediction: A Survey

[1]Avneet Saluja, [2]Dr. Bhupesh Gour, [3]Lokesh Singh

[1]M.Tech Scholar, [2] H.O.D. ,[3] Assistant Professor
Department of Comp. Sci. & Engg.
***Technocrats Institute of Technology Bhopal (M.P).India.*

***Abstract*— Web Usage Mining is an important application of data mining technique used to discover interesting usage patterns from the Web logs to understand and serve the requirements of Web-based applications. A Web log along with the identity of the user captures their browsing behaviour on a web site. Web prediction is a classification problem which attempts to predict the most likely web pages that a user may visit depending on the information of the previously visited web pages. In this paper emphasizes is given on the user future request prediction using web log record, click streams record and user information.**

The objective of the work is to provide a benchmark for evaluating various method used in past, at present and which can be used in future to minimize the search time of a user on the network. The usefulness of this work can be further enhanced when techniques with different methodologies are used in accordance with each other with the aim to increase the efficacy and probability of the user's next request prediction.

***Keywords*-Web Usage Mining, Web Log Mining, Future Request Prediction.**

I. INTRODUCTION

WEB PREDICTION is a classification problem which attempts to predict next set of pages user can access. This prediction classification concept can be applied for various applications such as search engines, caching systems and wireless applications. A huge amount of research has been done on Web Usage Mining (WUM) which gives the information about the user search behaviour. When the user browses the web pages, user leaves certain valuable information stored in the Web log. This content is very helpful in determining the web navigational pattern of user and the kind of information user wants from the specific sites.

WUM comprises of basically three major processes namely data pre-treatment, data mining and pattern analysis. Firstly, Pre-treatment of data is done on a series on Web logs to obtain logs with minimized redundancies, user, session, transaction identification and information on path completion. Secondly, mining algorithms are applies to extract user navigation patterns which represents relationship among Web pages in a particular Web site. Lastly, pattern analysing algorithm is applied to extract data for data mining applications.WUM automatically discovers knowledge from the data collected in Web logs. The collected log files, pattern analysis and click stream knowledge is helpful for recommending a set of objects(pages) to the active user which along with the actual data consists of links, ads, text or products applicable according to the user perceived preferences.

This paper presents a literature survey on Web Usage Mining approaches for user's next Request Prediction which includes various methods proposed for the prediction problem including its advantages and limitations.

The organization is as follows. In Section II, related work is presented. In Section III, introduction of various methods for user's request prediction has been explained. Section IV, gives a comparative analysis of the literature survey. In Section V, the paper is concluded emphasizing on the use Web Usage Mining for User's Request Prediction.

II. RELATED WORK

Internet nowadays is the most democratic of all the mass media. Millions of users access different Websites all around the world. When they access the network, a large amount of data is generated and is stored in Web log files which can be used efficiently as many times user repeatedly searched the same type of Web pages recorded in the log files. These series can be considered as a web access pattern, helpful to find the user behaviour. Through this personalized information, it's quite easy to predict the next set of pages user might visit based on the previously searched patterns, thereby reducing the browsing time of an user. This survey too focuses on how to improve the prediction time without compromising prediction accuracy.

In yesteryears, a huge amount of research has been done in consideration of Web Usage Mining for Web-Browsing Behaviour. However, the objective of this survey is to analyse the involvement of Web Usage Mining in user's future request prediction.

A proposed algorithm automatically discovers pages in a website for which the location is different from where the visitors expect to find them. For this purpose "Backtrack" as a key is used for the algorithm from the point where the user will backtrack. The point where the user starts to backtrack is considered as the expected location for the page[11]. One more algorithm is devised that selects the set of navigational links to optimize the prediction time and accuracy as in [4], works on a "Real Time Management Engine" that uses historical data and online visitation pattern of e-commerce site visitors. When the historical

data is huge the prediction accuracy is compensated with respect to the prediction time as bulk of data is to evaluated in order to provide the correct prediction. However, sometimes this makes the situation alleviated as branches for searching procedure multiplies which can affect the result.

III. Literature Survey

The focus of this section of the paper is to study and contrast different available techniques to predict user's future movement.

1.) Effective Prediction of Web-user Accesses using Pre-fetching:

Web Pre-fetching can be used in reducing user perceived latency problem present in every web based application. As web popularity is increasing day by day in huge folds, there is heavy traffic on the internet which results in delay in response due to congestion. The various reasons of delay on the web can be servers under heavy load, network congestion, low bandwidth, bandwidth underutilization and propagation delay. To overcome or to reduce these delays the only possible solution is to increase the bandwidth however, because of its high economic cost it cannot be considered as an optimal solution. For this purpose, a technique is proposed which aims at reducing the delay of client future request process on the web by getting the objects (pages) into the cache in a background environment before an explicit request is made for the page[2].

Factors which affects web pre-fetching algorithm such as order of dependencies between web document accesses and interleaving of requests belonging to patterns within user transactions and the ordering of requests.

In addition to this, a new algorithm Ordered Web Mining(WMO) is generated and is compared against previously proposed and existing algorithms on web-log mining for web pre-fetching and the result showed that WMO algorithm achieved higher accuracy levels in prediction with quite low overhead in network traffic[2]. Fig. 1 shows architecture how a web server will cooperate with a pre-fetch engine to disseminate hints every time a client request for a document in the server as in [2].

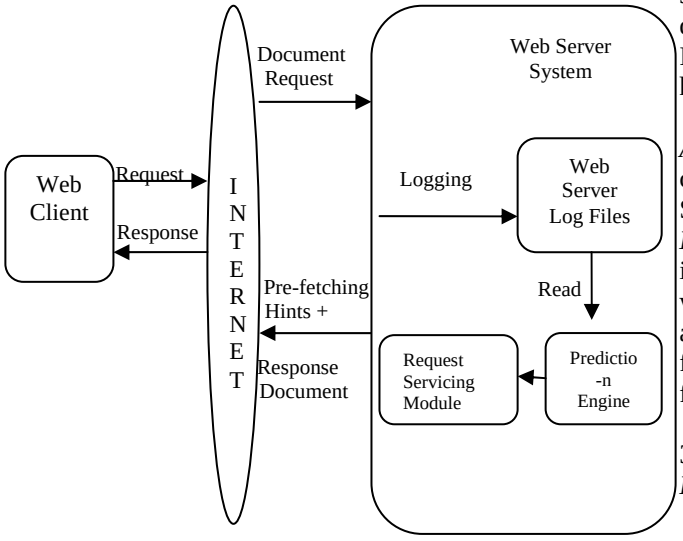

Fig. 1 Proposed Architecture of a Prediction Enabled Web Server

Advantage: Pre-fetching prevents bandwidth underutilization and hides part of the latency outcome by assuming that there is a system implementing a server-based predictive pre-fetcher, which piggybacks its predictions as hints to its clients. This technique also allows interleaving of requests belonging to patterns with random ones within user transactions. Lastly, idle time for searching can be reduced considerably as sources with various spatial differentiality can be considered in advance and given as the output whenever needed. However, none of the existing approaches has considered the aforementioned factors altogether [10] .

Disadvantage: The main disadvantage of this approach can be considered when the resources to be considered are highly dynamic that is it changes very frequently. When caching is used pages with high temporal data cannot be considered and may lead to unnecessary computational cost, due to maintenance of a large number of rules.

2.) *Evaluation of Web Usage Mining Approaches for User's Next Request Prediction:*

In this method three distinguished web mining approaches that are capable of exploring web logs were developed and experiments were also conducted with real web log data sets:

a) Association Rule (AR): In data mining Association Rule learning is a popular research method for discovering interesting relations between variables in large database as in [10].It describes, analyse and present strong rules for discovery in databases using different measures of interestingness as in .The problem of finding web pages that are visited together in a particular sequence is similar to finding an association among item sets in transaction databases. Once transactions have been identified each of them could represent a basket and each page in the basket could represent an item [6].

b)Frequent Sequences (FS): This approach is used to discover time ordered sequences of URL's that have been accessed by past user's.

c)Frequent Generalized Sequences (FGS): A generalized sequence is a sequence allowing wildcards in order to determine the user's navigational pattern in a flexible way. In order to extract frequent generalized subsequence's they have used generalized algorithm proposed by Gaul [6].

Advantage: The result of various experiments on a collection o f web log datasets evaluated that Frequent Sequences (FS) gives better accuracy than AR and FGS [6].

Disadvantage: A disadvantage of association mining is that it does not inherently use the notion of temporal distance which we believe is crucial for deciding which rules to apply for a given Web transaction. In addition to this frequent sequences too cannot predict navigational patterns for data sets that have not been used before [6].

3.) *A Customizable Behaviour Model for Temporal Prediction of Web User Sequences:*

Clustering and Association Rule, capture the sequential and the temporal nature of web pages as they will be visited

and are termed as Sequential Rules. To achieve sequentiality the rules are deployed in a manner such that they can store the click stream of the antecedent and the consequent. For Temporality the distance between the antecedent and the consequent is measured by the number of clicks required to go from one page to another. The prediction system with uses the concept of antecedent and consequent is reliable as they gives the information about when the pages are going to be accessed along with the information of what pages are going to be accessed. They also proposed a customizable prediction system i.e.; the prediction system is adaptable to various environment depending on the characteristics of the server which can include number of pages, architecture of server, number of links per page etc ,in order to capture more accurately the behaviour of its user's[5].

The method used measures the distance between the antecedent (first page) and the consequent (last page) in terms of number of clicks required to go from one page to other. As with the fast growing internet a customizable prediction system is required. The author has presented such a system that represents order of the set of URL's between antecedents and consequent along with the distance between them which makes it possible to determine when the pages are growing to be visited [5].

Advantage: The necessity of a customizable model comes from the fact that Internet servers have very different characteristics. The Customizable Model allows a trade-off between the number of rules and the prediction accuracy. Also, the model is able to capture the inherent sequentiality of web visits.

Disadvantage: Prediction accuracy depends on the parameter n, the distance in clicks between the antecedent and the consequent. If n is large the prediction accuracy suffers [5].

4.) *Web User Behaviours Prediction System Using Trend Similarity:*

A trend based application system is used to analyse user's behaviour and predict the future path of user based on trend similarity. It is not viable to predict the browsing behaviour of current user according to the similar past behaviour of other user's. Hence, a trend based prediction model is proposed to predict the future travelling path by generating ordered browsing sequence. The system proposed in fig. 2 works in two phases. One is the Construction phase and another is the Prediction phase. The construction phase helps to discover useful common browsing patterns for the experts and then they use it to predict the further browsing sequences. In predicting phase, the browsing behaviour of a new user is fed into the system to be compared within the prediction system so as to generate pages that can be pre-fetched to improve the browsing performance. Application of replacement algorithm on proxy servers and experimental results shows that the performance of the proposed model is useful to pre-fetch candidate's pages in advance thereby reducing search time.

Advantage: The browsing behaviours of new visitors in the website can be predicted according to the prediction model. It can be extracted to compare with all matching patterns in prediction model for predicting the future travelling path.

Disadvantage: Historical data may not give a true picture of an underlying trend hence, we will not be able to predict the behaviour.

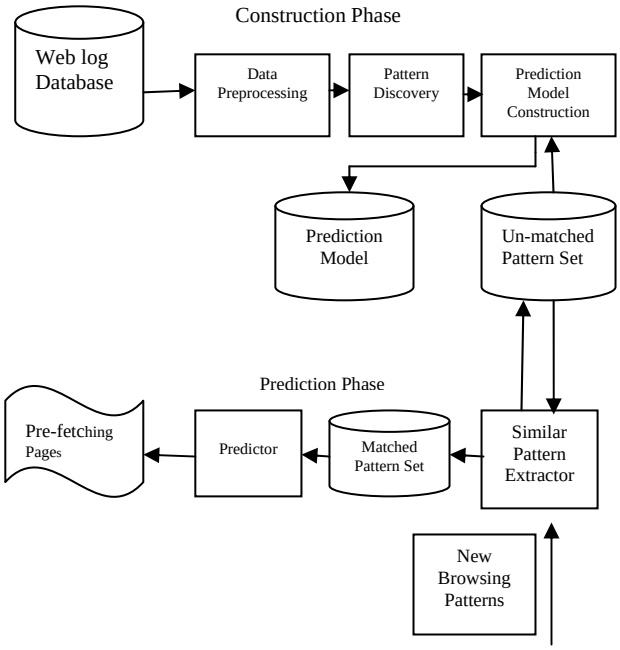

Fig. 2 Prediction System Architecture

5.) *A new classification model for online predicting user's future movement:*

Web Usage Mining is usually implemented for two components online and offline. The offline structure extracts knowledge from the historical log files and then this knowledge is used by the online component. Author proposed advance architecture for improving accuracy of classification in online phase. In the architecture fig. 3 classification is done using Longest Common Subsequence (LCS) algorithm. The architecture is partitioned into online and offline phases which work simultaneously. In offline phase data pre-treatment module process the web logs and reformat it to identify all web access session. The navigational pattern mining module cluster the group sessions according to certain common properties. In online phase, according to the URL requested and session identifier the user belongs to that session, the underlying knowledge base is updated and the list of suggestion is appended in the prediction list which finds the cluster based on LCS algorithm .Pre-treatment makes the data refined so that redundancy and anomalies can be removed in the earlier stage itself [7].

Advantage: The semantic knowledge about underlying domain can be used to improve the quality of the recommendations.

Disadvantage: Web sites made up from dynamically generated pages cannot be easily maintained [7].

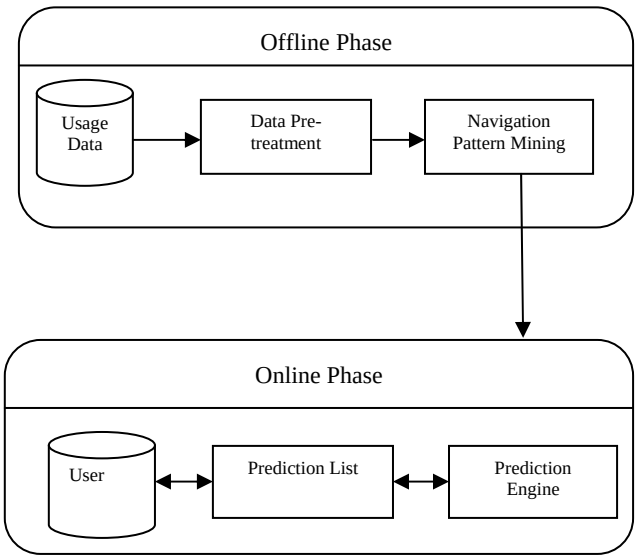

Fig. 3 Online /Offline Phase Architecture

6.) *WebPUM: A Web-based recommendation system to predict user future movements:*

Mehrdad Jalali,Narwati Mustapha,Md.Nasir Sulaiman,Ali Mamat advanced their previous work and renamed there architecture as WebPUM. In this they proposed a novel formula for assigning weights to edges of undirected graphs to classify current user activity. They used LCS algorithm to predict user near future movement and conducted two main experiments for navigation pattern mining and prediction of user's next request. In addition they found clustering patterns for user navigational behaviour and quality of the used datasets CTI and MSNBC improved [8].

Advantage: A Web usage mining architecture called WebPUM and proposed a novel approach to classify user navigation pattern for online prediction of user future intentions through mining Web server logs.

Disadvantage: The memory required to store Web server pages is quadratic in the number of pages, which will severely affect large Web sites that are made up from millions of pages [8] .

7.) A New Web Usage Mining Approach for Next Page Access Prediction:

Integration of Markov model based on sequential pattern mining with clustering showed that prediction accuracy is increased by 12% as compared to traditional Markov model. Clustering was used to identify similar access pattern from web logs using pair-wise nearest neighbour and then sequential pattern mining is performed on these identifies patterns to determine next page possible accesses. The compactness of clusters is improved by setting similarity threshold while forming clusters. When in future mining is done on these patterns, prediction accuracy will be improved as compared to the accuracy when mining is done on dissimilar access patterns. Therefore, a sequential mining technique called "Markov model" is suggested in combination with pattern discovery [1].

Advantage: The advantage of this approach is that every object will be a candidate of only one cluster. Hence by applying this approach for pattern discovery, the object that participate in prediction module are exact and have good resemblance with one another. The outcome will be a page with highest probability which the required motive of this survey. Markov Model provides good prediction accuracy if it is used in accordance with sequential mining [1] .

Disadvantage: Method does not consider loosely connected access sequences for mining process which can be considered as a limitation of this approach. Low order Markov Models have good accuracy however, they lack accuracy due to poor history or past web logs. In the same manner, high order Markov Model suffer from high state and space complexity as they use long browsing history[1] .

7.) *Prediciton of User's Search Behaviour :Application of Markov Model:*

Another variant of Markov Model for prediction of user's web browsing behaviour. They focused on a new modified Markov model to alleviate the scalability issue in the number of paths. In addition to this, a new two-tier prediction framework that creates an Example Classifier (EC),based on the training examples and the generated classifiers is developed. Experiments showed that such framework can improve the prediction time without compromising the accuracy. According to their work, the next action corresponds to predicting next set of pages to be visited and the previous actions corresponds to the pages that have already been visited. In Web prediction, the K^{th}-order Markov model gives the probability that a user will visit the k^{th} page provided that the user has visited the ordered k-1 pages. For example, in second-order Markov model, prediction of the next Web pages is computed only on the basis of the two web pages previously visited. Table I presents the All K-th Order Markov Model which gives the prediction steps to compute the next set of pages. The function(x,m_k) is considered to predict the next page visited of session x using the k^{th}-order Markov model. If m_k fails, then m_k-1 is considered using a new session x' of length k-1.This process repeats until the first-order Markov model is reached assuring that all orders fails to predict and no prediction is possible [13].

TABLE I
ALL K^{TH}-ORDER MARKOV MODEL

Algorithm: All K^{th} Prediction
Input: user session x, of length K
Output: Next page to be visited,p
1. p ← predict(x,m_k)
2. If p is not 0 then return p
3. x←remove page ID from x
4. K ← K-1
5. if (K=0)return 'failure'
6. Go to step 1
7. Stop

Advantage: Prediction accuracy rate increased to a higher rate.

Disadvantage: Approach is not feasible for the cases when the past log files are less i.e.; 'n' is small.

III. COMPARATIVE ANALYSIS

S.No	METHOD	OUTCOME	PREDICTION RATE
1.	Pre-fetching of Web pages into cache	Prediction Enabled Web Server/WMO	0.65
2.	AR,FS and FGS	FS produce optimal outcomes	0.69
3.	Clustering , Sequential Association Rule	Behaviour Model	0.72
4.	Trend based Prediction	Prediction System Architecture	0.75
5.	Online/Offline phases of architecture Clustering	Online Prediction of user's movement	0.80
6.	LCS algorithm,Clustering	WebPUM	0.85
7.	Sequential Pattern Mining with Kth order Markov model clustering	Prediction System to determine next page access	0.86
8.	Modified Markov Model with Association Rule mining	Two-tier framework with EC	0.89

IV. CONCLUSION

In this survey, the current state-of-the-art for Web Prediction Problem (WPP) has been reviewed. The information generated in log files can be extensively used through Web Usage Mining, to efficiently increase the prediction time without compromising accuracy. It is analysed that all K^{th}- Markov model with the use of two-tier architecture achieves good prediction accuracy. However, this approach is not feasible for the cases when the past log files are less i.e.; 'n' is small. In, the near future, the research can be extended for a small number of previous log files without affecting the accuracy and an in-depth analysis can be done on other features of session's log files to further improve the prediction accuracy level.

REFERENCES

[1] A.Anitha, "A New Web Usage Mining Approach for next page access prediction", *International Journal of Computer Applications*, vol.8, no.11, Oct. 2010.

[2] Alexandros Nanopoulous, Dimitris Katsaros and Yannis Manolopoulous, "Effective prediction of web-user accesses:A data mining approach", in *Proc. Workshop WEBKDD*, 2001.

[3] Bamshad Mobasher, Robert Cooley, Jaideep Srivastava, "Automatic Personalization Based on Web Usage Mining" *Communication of the ACM, Vol.43, no. 8*, Aug.2000.

[4] Debra Vander Meer,Kaushik Dutta,Anindya Dutta, "Enabling scalable online personalization on the Web", *in Proc. of the ACM conference on Electronic Commerce*, Minneapolis, Minnesota, Oct. 17-20,2000.

[5] Enrique Frias-Martnez, Vijay Karamcheti, "A Customizable Behaviour model for Temporal Prediction of Web User Sequences", in *WEBKDD 2002,LNAI 2703*, pp 66-85,2003.

[6] Mathias Gery, Hatem Haddad, "Evaluation of Web Usage Mining approaches for User's next request prediction", *in WIDM ACM conf. New Orleans*, Louisiana, USA, Nov. 7-8,2003.

[7] Mehrdad Jalali, Norwati Mustapha, Ali Mamat, Md.Nasir B. Sulaiman, "A new classification model for online predicting user's future movement", in *International Symposium on Information Technology*,pp. 1-7, Aug. 26-28,2008.

[8] Mehrdad Jalali, Norwati Mustapha, Ali Mamat, Md. Nasir B. Sulaiman, "WebPUM : A new Web-based recommendation system to predict user's future movements", in *International Journal Expert Systems with Applications* ,pp. 6201-6212,2010.

[9] Nien-yi Jan ,Nancy P. Lin, " Web User Behaviour Prediction System using Trend Similarity", in *Proc. of the 7th WSEAs International Conference on Simulation, Modelling and Optimization*, Beijing, China, Sep. 15-17, 2007.

[10] Piatestsky-Shapiro, "Discovery, analysis and presentation of strong rules", & W.J.Frawley, "Knowledge Discovery in Databases", in *AAAI/MIT Press, Cambridge*, vol. 13, no. 3,1992.

[11] Ramakrishnan Srikant, Yinghui Yang, "Mining Web Logs to improve website organization", in *Proc. of 10th International conf. of WWW*, Hong Kong, May 1-5, 2001.

[12] Zhong Su,Qiang Yang, Hongjiang Zhang, "WhatNext: A prediction system for Web requests using N-gram Sequence Models", in *Proc. of 1st Int. Conf. Hong Kong*, pp. 200-207,Jun. 2000.

[13] Mamoun A.Awad, Issa Khalil, "Prediciton of User's Search Behaviour :Application of Markov Model", *IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics*, vol. 42, no. 4,Aug. 2012