e-ISSN: 2349-9745
p-ISSN: 2393-8161

Scientific Journal Impact Factor (SJIF): 1.711

International Journal of Modern Trends in Engineering and Research

www.ijmter.com

A Survey on: Utilizing of Different Features in Web Behavior Prediction

Neetu Sahu[1], Pragyesh Kumar Agrawal[2]

[1]*Reaserch Scholar, Atal Bihari Vajpai Hindi Vishwavidyalaya, Bhopal, India* (M.P.)

[2]*Professor of Physics, Sarojini Naidu Govt. Girls P.G. College, Bhopal, India* (M.P.)

Abstract— As the web user increases day by day, there are many websites which have a large number of visitors at the same instant. So handing of these user required different technique. Out of these requirements one emerging field is next page prediction, where as per the user navigation pattern different features has been studied and predict the next page for the user. By this overall web server response time is reduce. In this paper a detailed study of the different researcher paper has shown, there techniques outcomes and list of features utilization such as web structure, web log, web content.

Keywords- Page Prediction, Web log, Web content, User behavior.

I. INTRODUCTION

With the large increase in the social networking websites, e-marketing websites in these days, many researchers are working in the web mining field in order to cover the various aspects of the fields. Out of different issues one of the major issues of web mining is to reduce the web server time as the numbers of hits get increased. For this some kind of prior preparation is required for handing such kind of hard work, which is full of possibilities.

This problem of reducing the web server response is term as web page prediction, some of researcher use next page prediction, etc. Thus increase for the server response is compulsory for all kind of website whose hit ratio is much above. As the word prediction is use in the solution it requires information related to the website, user, search engine can also provide the cache in information of the user behavior which might be helpful for the server or prediction algorithm in prediction [1].

Now the next problem of this prediction is gather related information from different features of the website and then pre-process it for the better usefulness of the algorithm. There are different web features of the web that is web logs, web structure and web content each has its separate preprocessing steps, and then proper data structure is required for the obtained features to process all.

In order to predict web page access pattern of the user it is required to develop the web recommendation model. For the increase of the perfection of the page, it is necessary to study the web user behavior. By the implementation of the prediction model web site visitors will gain quick response. It was estimate in 2004 that the numbers of user raise up to 945 million by (Computer Industry Almanac) [8]. As most of the user are untrained and they learned from the daily practice, while at the same time increase the use of the web make different information available online. In various survey it has been obtained that more than one million web page are inserted in the network every day and about 600 GB of page per month [11].

From the above explanation it can notice that day by day the numbers of users are increasing for their different work proportionally the quantity of data also increases. In this fashion it is very necessary to provide the prediction model for the user behavior analysis.

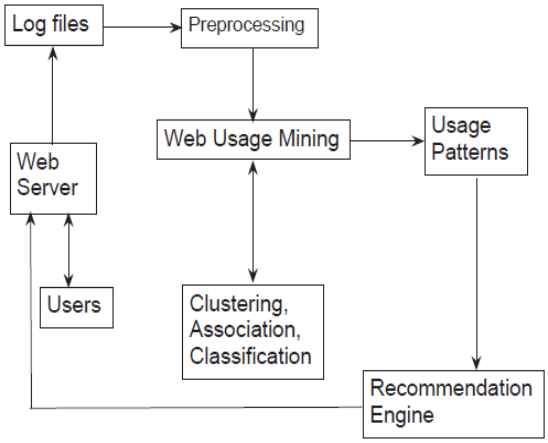

Fig.1 Web usage mining architecture.

Personalizing the Web users' content and recommending appropriate Web page simply that we are able to supply users with what they require based on their previous interactions within the same Web site. This task is viewed as a prediction task for we are trying to predict the users' level of interest in specific pages and rank these pages according to their predicted values [9].

When users access a website, a large volume 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 users 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 that what research has been done on Web usage mining in future request prediction [10].

In Web prediction, main challenges are in both preprocessing and prediction. Preprocessing challenges include handling large amount of data that cannot fit in the computer memory, choosing optimum sliding window size, identifying sessions, and seeking/extracting domain knowledge. Prediction challenges include long training/prediction time, low prediction accuracy, and memory limitation.

II. RELATED WORK

In [13] prediction of next page is base on the soft computing algorithm that is Ant colony where just like the ant behavior user behavior is learned then response is done for the same as the pages of the visited website. Here Main theme is utilization of the content feature in form of the MIK (Most Important Keywords) vector from each page. Then the user is represent as a Ant and move in the web structure of the website for finding the pattern as per the initial terms select by the Ant. Here web log is use in the hieratical form for the analysis of the behavior or to make a decision for jump for the Ant from one page to other. In this work whole training and testing will generate large number of sequences which is matched from the user generated sequence which not fulfill the prediction requirements .Although it is good for the estimation of the website pattern sequence.

In [14] prediction of next page is done by the implementation of Markov model with different order for the page. Here web log are use for the development of the Markov model Modification in the

Markov model is that at any position page sequence is change although the pages are done in manner that same pages are analyzed in different combination. Now in this way most of the prediction done by the system is as per the user requirement. One more study done in this work is the Markov model order for the evaluation of prediction. Results shows that 3rd order Markov model prediction is much accurate as compare to other Markov model

In [15] extension of the work has been done where most of the work again focuses on increasing the prediction of the web user. It is obtained that web content feature has been included in the paper which is not new for web mining but increase the prediction capacity of the work. In similar fashion of the [14] a comparison study of the Markov order is done where third order prediction works well as compare to other orders.

In [12] all the features web content, web log and web structure of the web mining are utilize. Here different models are prepare with the help of features such as TemNavNet and DomainOntology, where base on the prior knowledge of the whole web features keywords are collect from Web-pages. Then extract term sequences from the Web-page contents, and build the semantic network – TermNetWP. Now builds DomainOntoWP by web usage. Generates FWAP by Markov model Builds FVTP (frequently viewed term patterns). Then in testing phase**.** Identifies a set of currently viewed terms using query DomainOntoWP. Find next viewed terms on the 1st-order TermNavNet. Finally Predict next pages mapped to next term on DomainOntoWP. Here whole paper predicts the pages which have quite good accuracy level. Although complexity of the work is high.

In [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 pre-fetching 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. These thresholds 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.

In [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 was 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 has 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.

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.

In [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.

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.

In [6] propose 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. Three kind 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. Through experimental evaluation on both of real-life and simulated datasets, the proposed *TN-Gram* model is shown to outperform other approaches like N-gram modeling in terms of prediction precision, in particular when the web user's navigating behavior changes significantly with temporal evolution.

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:-
a. 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.
b. Online phase the main objective of this phase is to classify 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.

III. PROBLEM IDENTIFICATION

Many of the researches of previous works, work for the different combination of those features for prediction. But in [12] it has utilized all the features of the web and creates the ontology for the same. Such as with the help of the web page title it collect important keywords of the website then links those words and webpage in a structure for the betterment of the prediction quality. Here in the

module of the work propose work make changes that it use whole web page content instead of the web title only as it make only the use of the limited words in the website which is true for some websites only which has web page title in prior while there are many other website that has no such kind of title then system get helpless. So here whole content of the website each page are pre-process for the prediction then make the work easy or suitable for all kind of websites.

One more feature of the web is use that is web log which need preprocessing as web log contain many information like IP address, time, date, webpage sequence. So removing of that useless information from the web log is term as preprocessing this is one of the measure issues as the required information is only the weblog sequence.

Next problem is the pattern generate in the sequence as there are many clustering algorithm use for the same but these are not effective and thing are not fruitful for searching the best pattern. In few works where web logs are use in the form of web page are base on the Markov model which is quit big and level of the Markov model is second. This is one of the most fruitful methods for the weblog arrangement and many researchers are working with these regularly.

Now combination of the feature then prediction algorithm is another problem of the work because utilization of the feature at right time and right place is very necessary. In previous work user query is utilize to find the web keywords then related web pages are search in the weblog Markov model then finally most frequent page is term as the predicted page. But here user interest which is obtained from the query is related to many pages and then search of the frequent page is some kind of misleading work.

So measure goal are prediction model is to identify the subsequent requests of a user, given the current request that a user has made. This way the server can pre-fetch it and cache these pages or it can pre-send this information to the client. The idea is to control the load on the server and thus reduce the access time. Careful implementation of this technique can reduce access time and latency, making optimal usage of the server's computing power and the network bandwidth.

IV. EVALUATION PARAMETER

In order to evaluate this work there are different parameter present for the different techniques. The best parameter which suit this work is the precision where it give the value which is a measure of the prediction which is correctly identify by proposed model to the all the logs pass in the experiment. The other important measure is the Satisfaction which is include new in this era. This is the ration of the prediction page which are satisfactory but not the actual or the correct prediction to the total number of logs pass in the model. This can be understand part of the web log pass for the test will predict one page then that page is compare with the next to next page of the log if it same then consider it as the satisfactory result.

Precision, Recall, F-Measure: In this evaluation parameter let us consider a Weblog = {L1,L2, L3……..Ln}. Here L is the particular web page sequence such L1 = (P1, P2, P4, P5 , P6 , P9). In order to find the prediction the part of the log is pass in the system such as (P1, P2, P4) then for this pass correct prediction is P5 if the system generate the P5 value then consider it as the correct prediction otherwise consider it as the incorrect one.

$$\Pr ecision = \frac{True\_Positive}{True\_Positive + False\_Positive}$$

$$\text{Recall} = \frac{True\_Positive}{True\_Positive + False\_Negative}$$

$$F\_Score = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

In above true positive value is obtained by the system when the prediction by the user is same as the system. While if prediction by system is say page one and user open page two then True negative.

Satisfaction: In this evaluation parameter let us consider a Weblog = {L1,L2, L3……..Ln}. Here L is the particular web page sequence such L1 = (P1, P2, P4, P5, P6 , P9). In order to find the prediction the part of the log is pass in the system such as (P1, P2, P4) then for this pass satisfactory prediction is P6 if the system generate the P6 value then consider it as the satisfactory prediction otherwise consider it as the unsatisfactory one.

So Satisfaction = Rs/ R

Where Rs is the number of correct prediction
R is the total number of logs

Execution Time: Total Time for the execution of the algorithm in the prediction of the page is based on the different size of dataset. This can be understood with the use of different number of web logs for initial training the total time of execution gets differed. So if a dataset for training have more number of web log session then it will take more time, here time is calculated in seconds.

V. CONCLUSIONS

World Wide Web has necessitated the users to make use of automated tools to locate desired information resources to follow and asses their usage pattern. Web page pre fetching has been widely used to reduce the user access latency problem of the internet; its success mainly relies on the accuracy of web page prediction. Markov model is the most commonly used prediction model because of its high accuracy. Low order Markov models have higher accuracy and lower coverage. Clustering by Association rule is one of the best solutions for resolving the problem of worse prediction accuracy of Markov model. It is a powerful method for arranging users' session into clusters according to their similarity. Finally a powerful method still required to develop a high prediction which utilize different features and reduce time complexity.

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] Thi Thanh Sang Nguyen, Hai Yan Lu, Jie Lu " Web-page Recommendation based on Web Usage and Domain Knowledge", 1041-4347/13/$31.00 © 2014 IEEE.

[13] Pablo Loyola∗, Pablo E. Rom´an∗ and Juan D. Vel´asquez. "Clustering-Based Learning Approach for Ant Colony Optimization Model to Simulate Web User Behavior", 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology.

[14] Mamoun A. Awad and Issa Khalil. "Prediction of User's Web-Browsing Behavior: Application of Markov Model". IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 42, NO. 4, AUGUST 2012 1131.

[15] Sonal Vishwakarma, Shrikant Lade, Manish Kumar Suman, Deepak Patel. "WEB USER PREDICTION BY: INTEGRATING MARKOV MODEL WITH DIFFERENT FEATURES" . *Vol. 2, No. 4, November 2013 IJERST.*

