OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

User Navigation Pattern Prediction from Web Log Data: A Survey

Vaibhav S. Dhande[1], Dinesh D. Puri[2]
[1]*Department of Computer Engineering SSBT's COET, Bambhori, Jalgaon (M.S.), India*
[2]*Department of Computer Engineering SSBT's COET, Bambhori, Jalgaon (M.S.), India*

***Abstract**: This paper proposes a survey of Web Page Prediction Techniques. Prefetching of Web page has been widely used to reduce the access latency problem of the Web users. However, if Prefetching of Web page is not accurate and Prefetched web pages are not visited by the users in their accesses, the limited bandwidth of network and services of server will not be used efficiently and may face the problem of access delay. Therefore, it is critical that we need an effective prediction method during prefetching. The Markov models have been widely used to predict and analyze users navigational behavior. All the activities of web users have been saved in web log files. The stored users session is used to extract popular web navigation paths and predict current users next web page visit.*

***Keywords:** Clustering, Markov Model, N-Grams, User Sessions, Web Usage Mining*

I. INTRODUCTION

World Wide Web (WWW) is collection of data which can be accessed by Web Browser. The World Wide Web is just a subset of Internet. The WWW is conceptual but the Internet is physical aspect like cable, router, switch etc. The Internet is actual network of networks where all the information presents. The Hyper-Text Transfer Protocol (HTTP) and File Transfer Protocol (FTP) are the methods used to transfer Web pages over the Network. Hypertext is a text which contains an address of another file or data.

Web mining research works with many areas such as data mining, text mining, Web retrieval and information retrieval. The classification depends on the aspects like the purpose and the data sources. Mining research concentrates on finding new information or knowledge in the data. On the basis of above information, Web mining can be divided into three different categories web usage mining, web content mining web structure mining [1] is the process of extracting useful information from Web documents. Content data is nothing but the collection of facts a Web page was designed to convey to the users. Web content mining is not only related but also different from data mining and text mining. It is also related to data mining because many data mining techniques can be applied in web content mining.

Web usage mining [1] is the application of data mining techniques to discover usage patterns from Web data in order to understand and better serve needs of Web based applications. It consists of three phases preprocessing, pattern discovery and pattern analysis. Web servers, client applications and proxies can easily capture data about Web usage.

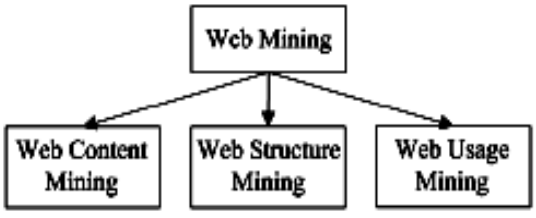

Figure 1. types of web mining

The goal of web structure mining [2] is to generate structural summary about the web page and website. A hyperlink is a structural component that connects the web page to a different location. The first type of web structure mining is extracting patterns from hyperlinks in the web.

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, 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 information of user behavior, 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 User Navigation Pattern Prediction in web usage mining. The motivation of this study is to know what research has been done on User Navigation Pattern Prediction from Web log data.

III. LITERATURE SURVEY

Recommendation systems are one of the early applications of Web prediction. Joa chims et al. [3] proposed the Web Watcher which is a path-based recommendation model based on ANN and reinforcement learning. The system contains some properties like (a) WebWatcher provides several types of assistance but most importantly highlights the interesting hyperlinks as it accompanies the user. (b) It learns from experience. (c) WebWatcher runs as a centralized server so that it can assist any Web user running any type of Web browser as well as combine training data from thousands of different users.

Su et al. [4] have proposed the N-gram prediction model and applied the all-N-gram prediction model in which several N-grams are built and used in prediction. Basically N-gram is a collection of N visited web pages by user. Their work is aimed at showing that using simple n-gram models for n greater than two will result in significant gain in prediction accuracy while maintaining reasonable applicability. They proposed path-based model for web page prediction. Their path-based model is built on a web-server log file L. They consider L to be reprocessed into a collection of user sessions, in a way that each session is indexed by a unique user id and starting time. Each session is nothing but a sequence of requests where each request corresponds to a visit to a web page (an URL).

The log L then consists of a set of sessions. Their algorithm builds an n-gram prediction model based on the occurrence frequency. Each sub-string of length n is n-gram. These sub-strings serve as the indices of a count table T. During its operation, algorithm scans through all sub-strings exactly once, recording occurrence frequencies of the next click immediately after the substring in all sessions. The request which is maximum occurred is used as the prediction for the sub-string.

Levene and Loizou [5] computed the information gain from the navigation trail to construct a Markov chain model to analyse the user navigation pattern through the Web. Navigation through the web, colloquially known as "surfing", is one of the main activities of users during interaction with web. When users follow a navigation trail they often tend to get disoriented in terms of the goals of their original query and thus the discovery of typical user trails could be useful in providing navigation assistance. Herein they give a theoretical underpinning of user navigation in terms of the entropy of an underlying Markov chain modelling the web topology. They present a novel method for online incremental computation of the entropy and a large deviation result regarding the length of a trail to realise they said entropy. They provide an error analysis for our estimation of the entropy in terms of the divergence between the empirical and actual probabilities. They also provide an extension of our technique to higher order Markov chains by a suitable reduction of a higher-order Markov chain model to a first-order.

M. Deshpande, G. Karypis [6], presented a class of Markov model-based prediction algorithms that are obtained by selectively eliminating a large fraction of the states of the All-Kth-Order Markov model. Their experiments on a variety of datasets have shown that the resulting Markov models have a very low state-space complexity and at the same time achieve substantially better accuracies than those obtained by the traditional algorithms.

M. Awad and L. Khan [7] have successfully combined several effective prediction models along with domain knowledge exploitation to improve the prediction accuracy. However, the module endures expensive training and prediction overheads because of the large number of labels/classes involved in the WPP.

M. T. Hassan, K. N. Junejo, and A. Karim [8] presented Bayesian models for two things like learning and predicting key Web navigation patterns. Instead of modeling the general problem of Web navigation they focus on key navigation patterns that have practical value. Furthermore, instead of developing complex models they present intuitive probabilistic models for learning and prediction. The patterns that they consider are: short and long visit sessions, page categories which visited in first N positions, rank of page categories in first

N positions, and the range of page views per page category.They learn and they predict these patterns under four settings corresponding to what is known about the visit sessions (user ID and/or timestamp).

F.Khalil, J. Li, H. Wang [9] improved the Web page access prediction accuracy by integrating all three prediction models: Clustering, Markov model, and association rules according to certain constraints. Their model, IMAC, integrates the three models using lower order Markov model. Clustering is used to group homogeneous user sessions. Low order Markov models are built on clustered sessions. Association rules are used when Markov models could not make clear predictions. The integrated model has been demonstrated to be more accurate than all three models implemented individually, as well as, other integrated models. The integrated model has less state space complexity and is more accurate than a higher order Markov model.

Bhawna Nigamand and Dr. Suresh Jain [10] proposed three different Prefetching and Caching schemes i.e. Prefetching only, Prefetching with Caching and Prefetching from caching. Dynamic Nested Markov model is used for predicting next accessed web page. The Experimental result shows that the Prefetching with caching scheme will give good results. By applying these schemes, users' web access latency can be minimized and quality of service can be provided to the web user.

Mamoun A. Awad and Issa Khalil [11] analysed and studied Markov model and all- Kth Markov model in Web prediction. They proposed a new modified Markov model to alleviate the issue of scalability in the number of paths. They have used standard benchmark data sets to analyse, compare, and demonstrate the effectiveness of our techniques using variations of Markov models and association rule mining.Their experiments show the effectiveness of modified Markov model in reducing the number of paths without compromising accuracy. Additionally, the results support their analysis conclusions that accuracy improves with higher orders of all Kth model.

Poornalatha G, Prakash S Raghavendra [12]presented a paper to solve the problem of predicting the next page to be accessed by the user based on the mining of web server logs that maintains the information of users who accessed the web site. Prediction of next page to be visited by the user may be pre fetched by the browser which in turn reduces the latency for user. Thus analysing user's past behavior to predict the future web pages to be navigated by the user is of great importance.

Li Yue et al. [13] propose a DOM-Based Block Text Identification method to detect navigation page. This method should extract block-text segments from a web page. If the number of segments is too small or too large, then that web page is classified as a navigation page. This method is based on this observation: a common content page contains main content, and this main content is not divided into a lot blocks. So if a web page contains no block text or contains too many block texts, it is rather a navigation page than a content page.

A.Anitha [14] proposed to integrate Markov model based sequential pattern mining and clustering. With the help of proposed approach approximately 12% of prediction accuracy increases compared to traditional Markov model. The main advantage of proposed hierarchical clustering approach is that every object must be candidate of only one cluster. The traditional Markov models have serious limitation which is, the low order Markov models have good coverage but they lack accuracy due to poor history and high order Markov models suffers from high state space complexity, because they use long browsing history, but high order. Markov models provides good prediction accuracy for that purpose in proposed approach combined the advantages of both Markov models and in order to improve the accuracy of prediction process, sequential mining used. We have summarized various method of web usage mining in next session.

Mehrdad Jalali, Narwati Mustapha, Md. Nasir Sulaiman, Ali mamat [15] advanced their previous work and renamed there architecture as WebPUM. In WebPUM they proposed a novel formula for assigning weights of edges of undirected graph to classify the current user activity. They used Longest Common Subsequence algorithm to predict user near future movements and they conducted two main experiments for navigation pattern mining and in second experiment, prediction of the user next request has been performed and they found quality of clustering for user navigation pattern and the quality of recommendation for both CTI and MSNBC datasets improved.

V. Sujatha, Punithavalli [16] 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 while in third stage the results of classification and clustering is used to predict the user future requests. Figure (2) shows PUCC model. The first stage is the cleaning stage, in which the unwanted log entries were removed. In the second stage, the cookies were identified and removed. The result was then used 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.

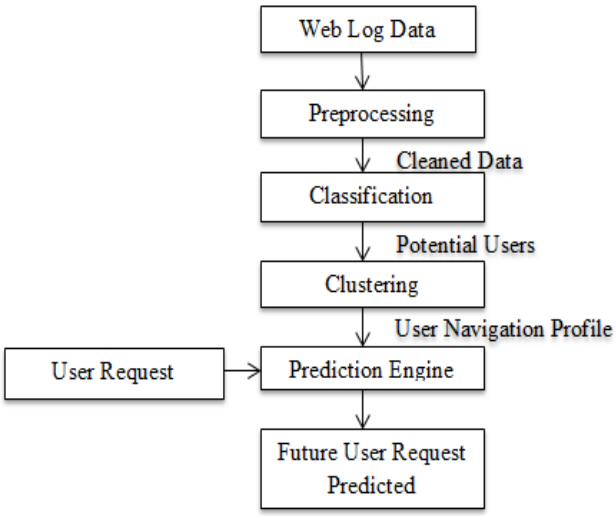

Figure 2 PUCC Model

Trilok Nath Pandey, Ranjita Kumari Dash , Alaka Nanda Tripathy ,Barnali Sahu [17] proposed IMC(Integrating Markov Model with Clustering) approach for user User Navigation Pattern 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 meaningful. After this the 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.

Mathis Gery & Hatem Huddad,[18] 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, the association rule learning is very 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 [18] 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.They have used the generalized algorithm In order to extract frequent generalized subsequences 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.

Yi-Hung Wu and Arbee L. P. Chen,[19] 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.

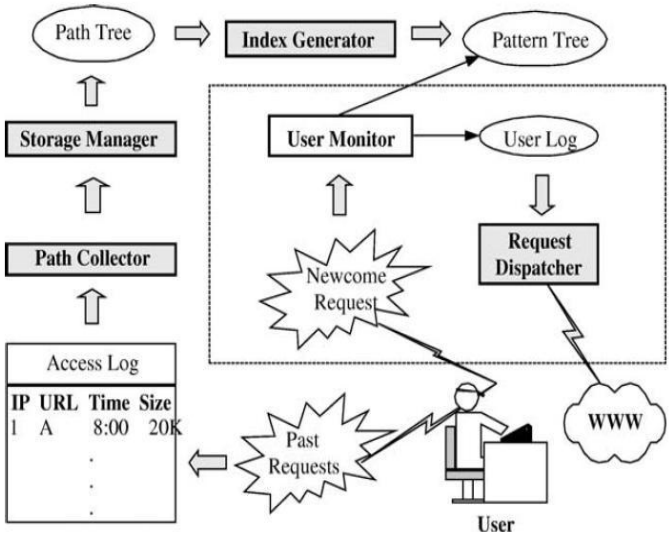

Figure 3. The flowchart of prediction system using proxy server log.

Siriporn Chimphlee, Naomie Salim, Mohd Salihin, Bin Ngadiman , Witcha Chimphlee [20] 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 by using association rules technique. In these approaches, the session identification technique collects the sequences of user requests, which distinguishes the requests for the same web page in different browses. In this experiment, the three algorithms Association rules, first-order Markov model and second-order Markov model are used. These algorithms are not successful in correctly predicting the next request to be generated. but first-order Markov Model is best than other because it can extracted the sequence rules and also choose the best rule for prediction and at the same time second-order decrease the coverage. This is only due to the fact that these models do not look far into the past to discriminate correctly the difference modes of the generative process.

IV. CONCLUSION

The conclusion based on the literature survey is that various researches had done on User Navigation Pattern 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 navigation Pattern Prediction and many types of models are developed for prediction.

World Wide Web has necessitated the users to make use of automated tools to locate desired information resources and to follow and assess their usage pattern. Web page prefetching 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.

The higher order models have a number of limitations associated with i) Higher state complexity, ii) Reduced coverage, iii) Sometimes even worse prediction accuracy. Clustering 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. We have discussed some of the techniques to overcome the issues of web page prediction. As the web is going to expand, web usage in web databases will become more and more. The above findings will become good guide in web page prediction effectively. In this paper, we have presented a comprehensive survey of up-to-date researchers of web page prediction. Besides, a brief introduction about web mining, clustering and web page prediction have also been presented. However, research of the web page prediction is just at its beginning and much deeper understanding needs to be gained.

V. FUTURE WORK

This survey paper will help to upcoming researchers in the field of web page prediction to know the available methods. This paper will also help researcher to perform their research in right direction. In future, researcher can work on Markov model to enhance the accuracy of web page prediction. First order Markov model is based on the assumption that the next state to be visited is only a function of the current one. The first-order Markov models (Markov Chains) provide a simple way to capture sequential dependence, but do not take into consideration the long-term memory aspects of web surfing behavior. Higher-order Markov models and hidden Markov models are more accurate for predicting navigational paths. Researcher can get better result if they will do the pre-processing phase effectively. Markov model and Clustering can work together and provide better prediction results without compromise with accuracy.In future prediction can be improved by using different techniques of data mining pattern discovery like classification, clustering, association rule mining etc.

REFERENCES

[1] Raymond Kosala, Hendrik Blockeel, "Web Mining Research": A Survey, ACM SIGKDD Explorations Newsletter, Volume 2 Issue 1, June 2000.

[2] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A.Tomkins, D. Gibson, and J. Kleinberg, "Mining the Web's link structure". Computer, 32(8):60–67, 1999.

[3] T. Joachims, D. Freitag, and T. Mitchell, "WebWatcher: A tour guide for the World Wide Web", in Proceedings of IJCAI, pp. 770–777, 1999.

[4] Z. Su, Q. Yang, Y. Lu, and H. Zhang, "What Next: Aprediction system for Web requests using n-gram sequence models", in Proceedings of 1st Int. Conference Web Inf. Syst. Eng. Conference, Hong Kong pp. 200–207, Jun. 2000.

[5] M. Levene and G. Loizou, "Computing the entropy of user navigation in theWeb," Int. J. Inf. Technol. Decision Making, volume 2, no. 3, pp. 459–476, 2003.

[6] M. Deshpande, G. Karypis, "Selective Markov Models for Predicting Web Page Accesses," ACM transactions on Internet Technology, volume 4, No.2, pp.163-184, May 2004

[7] M. Awad, L. Khan, and B. Thuraisingham, "Predicting WWW surfing using multiple evidence combination," VLDB J., volume 17, no. 3, pp. 401–417, May 2008.

[10] M. Awad and L. Khan, "Web navigation prediction using multiple evidence combination and domain knowledge," IEEE Trans. Syst., Man,Cybern. A, Syst., Humans, volume 37, no. 6, pp. 1054–1062, Nov. 2007.

[8] M. T. Hassan, K. N. Junejo, and A. Karim, "Learning and predicting key Web navigation patterns using Bayesian models," in Proceedings of Int. Conf. Comput. Sci. Appl. II, Seoul, Korea, pp. 877–887, 2009.

[9] F.Khalil, J. Li, H. Wang, "An Integrated Model for Next Page Acces Prediction", Inderscience Enterprises Ltd., 2009.

[10] Bhawna Nigamand Dr. Suresh Jain, "Analysis of Markov Model on Different Web Prefetching and Caching Schemes", 978-1-4244-5967-4/10/ 2010 IEEE

[11] Mamoun A. Awad and Issa Khalil, "Prediction of User's Web-Browsing Behavior: Application of Markov Model," IEEE Trans. Syst., Man,Cybern. A, Syst., Humans, volume 42, no. 4, pp., Aug. 2012.

[12] Poornalatha G, Prakash S Raghavendra, "Web Page Prediction by Clustering and Integrate Distance Measures" IEEE/ ACM Trans. Syst., Man,Cybern. A, Syst., Humans, volume 44, no. 2, pp., Sep. 2012.

[13] Li Yue, Dong Shou-bin, Zheng Xiang, Ma Bin-Hua." Improving Navigation Page Detection by Using DOM-Based Block Text Identification" Tenth International Conference on ICT and Knowledge Engineering pp.978-1-4673-2317-8/12 IEEE-2012

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

[15] M. Jalali, N. Mustapha et al ," WebPUM: A Web-based recommendation system to predict user future movements", in international journal Expert Systems with Applications 37 (2010) 6201–6212

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

[17] Trilok Nath Pandey, 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.

[18] 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.

[19] 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.

[20] 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.