Volume 4, Issue 7, July 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Study on Web Page Prediction Methods Using Web Usage Mining

Alpa Sharma[1], Sarbjit kaur[2], Miss Himanshu Tayal[3]
1Master of Technology In Computer Science And Engineering, Modern Institute Of Engineering And Technology, Mohri, Kurukshetra, India
2Assistant Professor ,Department of Computer Science And Engineering, Modern Institute Of Engineering And Technology, Mohri, Kurukshetra, India
3.Master of Technology In Computer Science And Engineering, Modern Institute Of Engineering And Technology, Mohri, Kurukshetra, India

Abstract— The web has become the world's largest repository of knowledge. Web usage mining is the process of discovering knowledge from the interactions generated by the user in the form of access logs, cookies, user sessions data. There is an exponential growth of web log due to which conventional methods were proved to be inefficient. Web log is incremental and heterogeneous in nature, it becomes very crucial to predict the user's browsing behavior. For the web miners it has become very necessary to use efficient predictive techniques so as to know the exact user's browsing behavior. Moreover the web log is heterogeneous and non scalable. So there is a need to reduce the operation scope which in turn increases the accuracy precision significantly. Many researches has already been done in this context. The main aim of this paper is to give an overview of past and current evaluation in user's future request prediction using web usage mining.

Keywords— Future request prediction, User's Browsing Behaviour,Web Usage Mining, Markov Model, Web log.

I. INTRODUCTION

The web is huge, diverse and dynamic. The web has become the world's largest knowledge repository. The popularity of WWW is rapidly developing. Extracting the useful knowledge from web has become a tedious process as web logs are heterogeneous and non scalable in nature. Web mining has been categorized in three categories : Web content mining, Web Structuring mining and Web usage mining. Web content mining is used to extract useful knowledge which can be extracted from web pages.

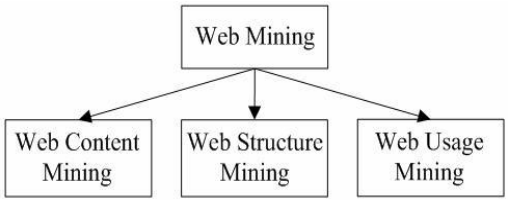

Figure 1: Taxonomy of Web Mining

Web structuring mining focuses on analyzing the links between web pages through web structure. The goal of Web Usage Mining is to find out the useful information from web data. The other goals are to enhance the usability of the web information and to apply the technology on the web applications, for instance, pre-fetching and caching. Forecasting the user's browsing behavior is one of web usage mining issue. In order to achieve the purpose, it is necessary to understand the customer's browsing behavior through analyzing the web data or web log files. Predicting the most possible user's next requirement is based on the previous similar behavior. There are many advantages to implement the prediction, for example personalization, building proper web site, improving marketing strategy. The rest of the paper is organized as follows: Section 2 presents the motivation, Section 3 presents the literature review and Section 4 gives the conclusion and future work.

II. MOTIVATION

With the massive usage of World Wide Web, large number of users access web sites from all over the world. When the user access a web sites, a large volumes of data such as IP Address of users , access log, agent log etc. information are collected and the log is maintained in log files as the user may access a series of web pages repeatedly. These series of accessed web pages can be considered as the browsing pattern of user which can be of great use for predicting next web

© 2014, IJARCSSE All Rights Reserved

pages. With the help of user future request prediction the browsing time can be reduced and server load can be decreased as well. In recent years there has been lots of research in this field. The main motivation of this study is to know how much research has been done on "Web Usage Mining-Web Page Prediction".

III. LITERATURE SURVEY

Literature survey which is the study or collection of information about web usage mining is used to find the web navigation behavior of user and collecting the information about "User Future Request Prediction" approach that is used to predict the next request of the user. Alexandras Nanopoulos, Dimitris Katsaros and Yannis Manolopoulos[1] focused on web pre-fetching as it has great significance in reducing user perceived latency present in every web based application. Due to web popularity and heavy traffic over web, it results in response delay of requests by clients which in turn leads to time wastage and increases the load on server too.

To name a few reasons of the delays, the web servers under heavy load, Network congestion, Low bandwidth, Bandwidth under-utilization and propagation delay. The solution lies in increasing the bandwidth but this is not the best solution economically. For these reasons a technique was proposed in which the delay of client future requests for web objects was reduced by transferring those objects into cache in the background before an explicit request is made for them. The important factors which affect on web pre-fetching algorithm like order of dependencies between web documents access and interleaving of requests belonging to patterns with random ones within user transactions were discussed in their paper. The architecture of prediction enabled Web server [1] Yi-Hung Wu and Arbee L. P. Chen [2] of user behaviors generates sequences of consecutive web page accesses that is derived from the access log of proxy server. The frequent sequences thus organized, discovered and then used as an index. Based on this index, a new scheme for predicting user requests was made and a proxy based framework for pre-fetching web pages. Thus experiments on real data were performed and predictions convey with a high degree of accuracy with very little overhead. In these experiments results show the best hit ratio of the prediction achieves 75.69%, where the longest time to make a prediction requires only 1.9ms.The disadvantages are the low average service rate, the setting of the three thresholds used in the mining stage. These thresholds has a great impact in the construction of the pattern trees. The usage of minimum support and minimum confidence lies in pruning or omitting the useless paths which leads to over estimation of pruning effects. On the other hand, the grouping confidence is useful for the strongly related web pages due to use of editorial techniques, such as the embedded frames and the images.

Mathis Gery & Hatem Huddad,[3] Author proposed three web mining approaches to exploit web logs and that are : Association Rules (AR), Frequent Sequences (FS) and Frequent Generalized Sequences (FGS). These three algorithms are developed with real web log data.

Association Rule: In data mining, association rule learning is a popular method for discovering relations between variables in large database. It describes, analyzes and then presents strong rules that discovered in database using different measures of interestingness .The problem of finding web pages that were visited together is similar with finding associations among item sets in transaction databases. These transactions correspond to a basket and each research is an item.

Frequent Sequences: This technique discovers time ordered sequences of URLs that have been followed by users in the past.

Frequent Generalized Sequences (FGS): It's a generalized sequence allowing wildcards which reflects user's navigation in a very flexible way. To extract frequent generalized subsequences they have used the generalized algorithm proposed by Gaul.

The experiments used three collections of web log datasets for small web site, then for large website and another third weblog dataset for intranet websites. Usage of 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.

Siriporn Chimphlee[3], Naomie Salim, Mohd Salihin, Bin Ngadiman , Witcha Chimphlee [4] proposed a method for constructing the first-order and second-order Markov models for Web site access prediction based on past visitor behavior and then compare it with association rules technique. These approaches, sequences of user requests are collected by the session identification technique,which further distinguishes the requests for the same web page in different browses. The three algorithms first-order Markov model, second-order Markov and Association rules are used but 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 whereas the second-order decreases the coverage too. This is because of the fact that these models do not look far into the past to correctly discriminate the difference modes of the generative process. Christos Makris, Yannis Panagis, Evangelos Theodoridis,and Athanasios Tsakalidis [5] Proposed a technique for predicting web page usage patterns by modeling user's navigation history using string processing technique and validated the superiority of proposed technique experimentally and weighted suffix tree is used for modeling user navigation history. The method proposed has the advantage that it demands a constant or fixed amount of computational effort per user action and consumes a relatively minimal amount of extra memory space. Vincent S. Tseng, Kawuu Weicheng Lin, Jeng-Chuan Chang in [6] Propose a novel data mining algorithm named Temporal N-Gram (TN Gram) for constructing a

© 2014, IJARCSSE All Rights Reserved

prediction model for 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 various time slots. Through experimental evaluation on both real-life and simulated datasets, the proposed TN-Gram model is shown to outperform the various other approaches like N-gram modeling in terms of prediction precision, particularly when the web user's navigating behavior changes significantly with temporal evolution. Mehrdad Jalali, Norwati Mustapha, Md. Nasir Sulaiman, Ali Mamat in [7] Proposed a recommendation system called Web PUM, an online prediction using Web usage mining system which effectively provides online prediction and propose a novel approach for classifying user navigation patterns to predict user's future intentions and desires. The approach is based on the new graph partitioning algorithm in order to model user navigation patterns for the navigation patterns mining phase. LCS algorithm is used for classifying current user activities to predict user next intention or movement. Chu-HuiLee , Yu-lung Lo, Yu-Hsiang Fu [8] propose an efficient prediction model, two-level prediction model (TLPM), using an aspect of natural hierarchical property from web log data. TLPM can decrease the size of candidate set of web pages and can significantly increase the speed of predicting with adequate accuracy.

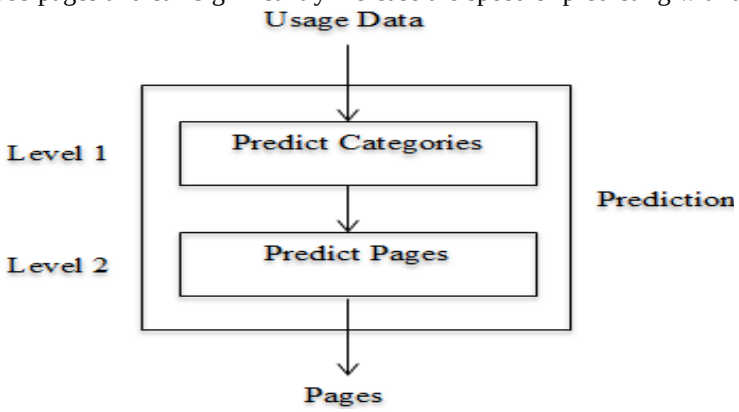

Figure 2: Two level prediction model(TLPM).

The experiment results also prove that TLPM can highly enhance the performance of prediction when the number of web pages are consistently increasing. Finally, the prediction result of TLPM can be applied for various aspects such as pre-fetching and caching on web site, personalization, target sales, improving web site design,etc. V. Sujatha, Punithavalli [9] proposed the Prediction of User navigation patterns using Clustering and Classification (PUCC) from web log data. In the first stage PUCC focuses on separating the potential users in web log data, and in the second stage clustering process is done to group the potential users having similar interest and then in the third stage the results of classification and clustering are used to predict the user future requests.

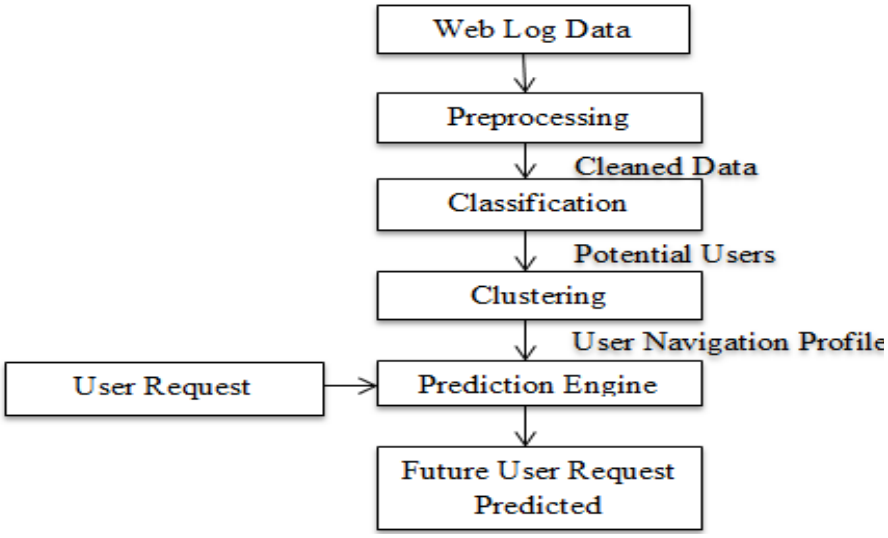

Figure 3: PUCC Model

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

© 2014, IJARCSSE All Rights Reserved

X. Chen, [19], proposed partial match forest, in this the roots signifies the popular web pages and the branches represents the non-popular web pages. It was assumed that user will start navigation with popular web pages. J. Borges, [20], has analyzed the summarization ability of the Variable Length Markov model. Nizar R, [22], proposed the Markov model with semantic information for pre-fetching the web page. They have incorporated the semantic information with transition probability of Markov model. Nizar R, [23], Proposed the sequential pattern mining algorithm with semantic in Markov model. Result shows that the semantics-aware sequential pattern mining algorithm is four times faster than the non semantic aware sequential pattern mining algorithms.

Several authors have proposed models for modeling the user web navigation sessions. J. Borges, [4], proposed a Markov model for modeling the user web navigation sessions. In this author has preprocessed the web log file then modeled it through the Markov model and finally model is used to identify the useful patterns. F. Khalil, [10], has proposed a new framework for predicting the next web page access. Authors of [10] have used the Markov model for web prediction. If the Markov model is not able to predict the next page then the association rule are used to predict the next web page. They have also proposed the solution for ambiguity in the prediction. Ambiguity will be resolved by taking the help of association rule.

J. Borges, [12, 14, 21], proposed Higher-order Markov model for web usage mining. There are various problems associated with lower-order Markov model. The low accuracy is the major limitations of lower- order Markov model. As Higher-order Markov model suffer from the state space complexity, K-mean clustering technique has been used to reduce the state space complexity. The experimental result shows that the accuracy is improved by introducing the clustering technique in Markov model. Author has proposed the Higher-order Markov model with clustering technique to improve the effectiveness of Markov model and to reduce the state space complexity.

M. Desponde, [13], proposed the new approach for reduction of the complexity of Markov model. Three approaches frequency-pruning, error-pruning and support-pruning have been used to reduce the state space complexity. With the help of chi square test the predictive power of first-order, second-order and higher-order Markov model has been tested. R. Popa, [18], proposed the Hybrid-order tree like Markov model, and it is found that the Hybrid-order tree like Markov model is predicting accurately than traditional Markov model.

M. Eirinaki, [16], proposed the combine work of Page Rank and Markov model for predicting the next web page. The prior probability and the transition probability are calculated through the Page Rank and then it is used with Markov model for prediction. J. Zhang, [15], proposed the preprocessing of web log file for mining. The web navigation sessions have been prepared for modeling. J. Borges, [18], has tested the predictive power of variable length Markov model.

Based on the literature survey done some of the basis are there for the comparison of different Markov Models.

Table 1. Comparison on some Markov Models

BASIS/TECHNIQUES	**TMM**	**DNMM**	**CMM**	**KTHORDER MM**	**HYBRID MM**
SEARCH SPACE	Increased	Decreased	Decreased	Increased	Decreased
COVERAGE	Decreased	Increased	Increased	Decreased	Increased
TRANSITION MATRICES FORMED	Yes	No	Yes	Yes	Yes
UPDATION	Tedious	Easy	Difficult but less than TMM	Tedious	Tedious
MEMORY WASTAGE	Yes	No	Improved from TMM	Improved from TMM	Improved from TMM
ACCESS TIME	Slow	Fast	Improved from TMM	Improved from TMM	Improved from TMM

Where TMM stands for Traditional Markov Model, DNMM for Dynamic Nested Markov Model, CMM for Clustering with Markov Model using Apriori algorithm, Kth order MM for Kth order Markov Model and Hybrid MM stands for Hybrid Markov Model.

IV. CONCLUSION AND FUTURE WORK

The conclusion based on the literature survey is that various research work had been done to predict user browsing behavior. For pattern discovery different techniques such as graph partition techniques of clustering, Bayesian techniques of classification etc are used for user's future request prediction. Many types of models are developed for better prediction. In future, the prediction can be improved by using different types of techniques of data mining pattern discovery like clustering, association rules and use of higher order Markov model etc so as to improve the hit ratio without increasing the access time.

© 2014, IJARCSSE All Rights Reserved

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] J. Borges. "A data mining model to capture user web navigation. Ph. D. thesis", University College London, London University, 2000.

[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] F. Khalil, J. Li, and H. Wang "A framework of combining Markov model with association rules for predicting web page accesses", Proc. Fifth Australasian Data Mining Conference (AusDM2006), vol. 61, 2006, pp 177–184.

[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] J. Borges, and M. Levene, "A clustering-based approach for modelling user navigation with increased accuracy", Proc. Second Int'l Workshop Knowledge Discovery from Data Streams, Oct. 2005,

[13] M. Desponde, and G. Karpis "Selective Markov models for predicting web page accesses" ACM Transactions on Internet Technology, vol. 4, no. 2, May 2004, pp.163–184.

[14] J. Borges, and M. Levene, "Generating dynamic higher-order Markov models in web usage mining," Proc. Ninth European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD), eds. A. Jorge, L. Torgo, P. Brazdil, R. Camacho, and J. Gama, Oct. 2005, pp. 34–45.

[15] J. Zhang, and A. A. Ghorbani, "The reconstruction of user sessions from a server log using improved timeoriented heuristics." in CNSR. IEEE Computer Society, 2004, pp. 315–322.

[16] M. Eirinaki, M. Vazirgiannis, and D. Kapogiannis, "Web path recommendations based on page ranking and Markov models," Proc. Seventh Ann. ACM Int'l Workshop Web Information and Data Management (WIDM '05) , 2005, pp. 2–9.

[17] R. Popa, and T. Levendovszky "Marcov models for web access prediction" 8th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics, Nov 2007.

[18] J. Borges, and M. Levene, "Testing the predictive power of variable history web usage," J. Soft Computing, special issue on Web intelligence, 2006.

[19] X. Chen, and X. Zhang, "A Popularity-Based Prediction Model for Web Prefetching," Computer, 2003, pp. 63–70.

[20] J. Borges, and M. Levene, "Evaluating variable-length Markov chain models for analysis of user web navigation sessions", IEEE Trans. Knowl. Data Eng., Vol. 19, No. 4, 2007, pp. 441–452.

[21] J. Borges, and M. Levene, "Data Mining of User Navigation Patterns," Web Usage Analysis and User Profiling, eds. B. Masand and M. Spiliopoulou, LNAI 1836, pp. 92–111, Springer, 2000.

[22] Nizar R. Mabroukeh, and C. I. Ezeife, " Semantic-rich Markov Models for Web Prefetching, in the proceedings of the 2009 IEEE International Conference on Data Mining (ICDM) Workshops (Workshop on Semantic Aspects in Data Mining (SADM 09)), Miami Florida, December 6-9, 2009, pp. 465–470.

[23] Ni Mabroukeh, and C.I. Ezeife, "Using domain ontology for semantic web usage mining and next page prediction", in the proceedings of the18th ACM Conference on Information and Knowledge Management (CIKM), Hong Kong, November 2-6, 2009.

[24] UCI KDD archive, http://kdd.ics.uci.edu/.

© 2014, IJARCSSE All Rights Reserved