Volume 4, Issue 8, August 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Revised Two Phase Prediction Model for Better Web Page Prediction

Alpa Sharma[1], Sarbjit kaur[2], Himanshu[3]
[1]Master of Technology In Computer Science And Engineering, Modern Institute Of Engineering And Technology, Mohri, Kurukshetra, India
[2]Assistant 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- World Wide Web is growing rapidly. It has become the world's largest repository of knowledge. As more and more number of user's are accessing web as a source of information there is a great opportunity to learn from the logs to learn about the user's probable actions in the future. Web prediction is a classification problem in which we have to predict the next set of web pages that user may visit based on the knowledge of previously visited pages. This paper introduces a revised two phase prediction model which make use of Markov Model. This two stage prediction model enables web miners to identify and analyze web user navigation patterns. To validate the proposed model experiments were conducted and results were proved.

Keywords- Web Usage Mining, User's Browsing Pattern, Markov Model, Clustering etc

I. INTRODUCTION

Web has become the World's largest repository and the popularity of Web has been increasing day by day. Web has now became an important source of information retrieval now a days. The users who are accessing web are from different backgrounds. Extracting knowledge from the Web efficiently and effectively is becoming a tedious process. The exponential growth of the Web has greatly increased the amount of usage data in server logs. Web mining can be categorized into three categories as shown in Fig.1, web content mining, web structure mining and web usage mining. Web content mining focuses on useful knowledge which is extracted from web pages.

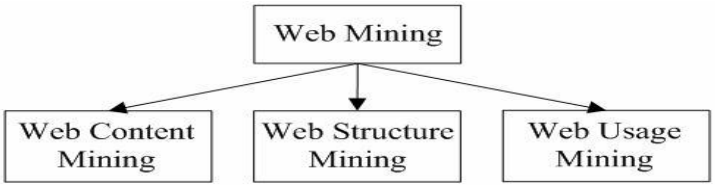

Fig. 1 Web Mining Categorization

Web structure mining is used to analyze the links between web pages through the web structure to infer the knowledge. Web usage mining is extracting the information from web log file which is accessed by users.

Web Usage Mining Procedure: The general process of web usage mining includes (i) Pre processing: Process of cleaning, Integrating and Transforming of the result of resource collection, (ii) Pattern discovery: Process of uncovered general patterns in the pre process data and (iii) Pattern analysis: Process of validating the discovered patterns. The process of web usage mining procedure.

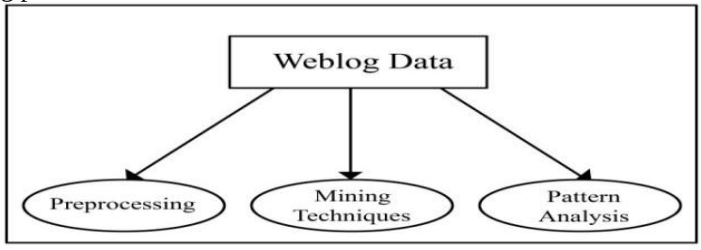

Fig. 2 Web Usage Mining

Web Log File: Web usage mining is used to find out the interrelated information from web log file which is involved. All of users' browsing behavior is clearly recorded in the web log file with users' name, IP address, date, and request time etc. The format of web log file can be divided into two types, common log files and extended log. Common log file is constructed by access log and error log, referrer log and agent log are appended to it, and form extended log. After web log file is acquired, the procedure of web usage mining must be executed.

This paper introduces a Revised Two Stage Prediction model for representing and analyzing the Web User navigation data. This model enables the identification of user navigation patterns and can also be used to foresee the next link choice of a user. It will also help to reduce the operation scope in Stage two just in some specific categories instead of all categories. After that, the web pages in suitable categories are predicted. It is expected that the two Stages of prediction model can reduce the operation scope and increase the accuracy precision.

The rest of this paper is organized as follows: Section 2 is the related work. The two Stages of prediction model is proposed in Section 3. Section 4 consists of the experimental analysis. The conclusion and future work will be discussed in Section 5.

II. RELATED WORK

Many of the previous authors are expressing the criticality and importance of identifying the user's browsing behavior available visiting data available in web log. Most of the works in the literature concentrates on pattern discovery to identify the browsing behavior of the user. Several models in the literatures proposed for identifying the association between the pages without considering the category.

Alexandras Nanopoulos,Dimitris Katsaros and Yannis Manolopoulos[1] focused on web pre-fetching because of its importance in reducing user perceived latency present in every web based application.This research presents important factors which affect on web pre-fetching algorithm like order of dependencies between web document access and the interleaving of requests belonging to patterns with random ones within user transactions. 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, derived from the access log of proxy server. Siriporn Chimphlee, Naomie Salim, Mohd Salihin, Bin Ngadiman , Witcha Chimphlee [3] 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. Christos Makris, Yannis Panagis and Athanasios Tsakalidis [4] Proposed a technique for predicting web page usage patterns by modeling users' navigation history using string processing technique and validated experimentally the superiority of proposed technique and weighted suffix tree is used for modeling user navigation history. Vincent S. Tseng, Kawuu Weicheng Lin, Jeng-Chuan Chang in [5] Propose a novel data mining algorithm named Temporal N-Gram (TN Gram) for constructing prediction models of Web user navigation by considering the temporality property in Web usage evolution. Chu-Hui Lee , Yu-lung Lo, Yu-Hsiang Fu [6] propose an efficient prediction model, two-level prediction model (TLPM), using a novel aspect of natural hierarchical property from web log data. TLPM can decrease the size of candidate set of web pages and increase the speed of predicting with adequate accuracy.

To extract useful browsing patterns one has to follow an approach of pre processing and discovery of the hidden patterns from possible server logs which are non scalable and impractical. Hence to reduce the operation scope there is a need of a model, which can identify the category and can reduce the operation scope

III. PROPOSED WORK

3.1 A revised Prediction Model

In this paper, it focuses on the preprocessing step and modifies the Two Levels of Prediction Model framework further. Because users' browsing features are diverse, the hierarchical agglomerative clustering is used to group users' browsing behaviors and acquire many different user clusters. The information of clusters can be seen as cluster view (that is each cluster has its own relevant matrix) for replacing of the global view (that is the only one relevance matrix for all users in the previous model). Therefore, we proposed a modified Prediction Model. In this model, the view selection will be used by which user's browsing feature is matched and used for predicting and improving the accuracy hopefully.

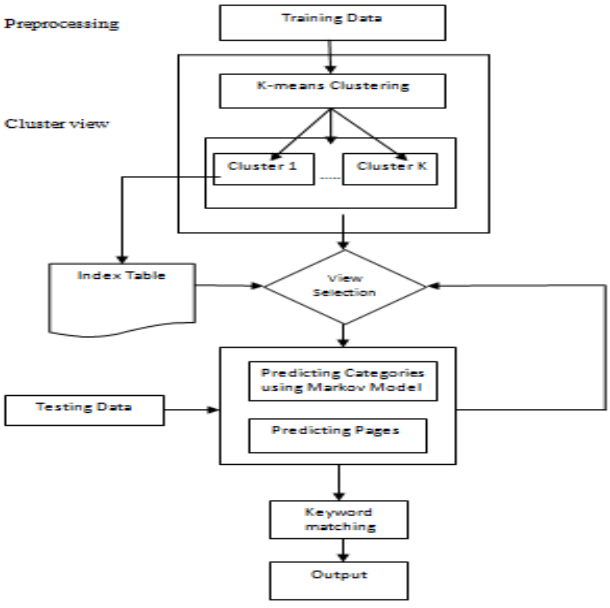

Figure 3. A Revised Two Phase Prediction Model

3.1.1 Framework

The steps of framework (as Fig. 3) are described as follows. The user database will be divided into training data and testing data. In step one of preprocessing, training data is processed by hierarchical agglomerative clustering. The k number of cluster view will be obtained which include k similarity matrices S , k first-order transition matrices P and k second-order transition matrices P2 between categories. Therefore, we get k relevant matrices R to represent k cluster views.

In step two, the centric vector of clusters will be released for creating an index table. The index table is used for view selection based on user's browsing behavior in time.

In step three, after view selection, testing data will be fed into the prediction model. The prediction result will be released as output.

3.2. TWO STAGE - PRE PROCESSING

The steps of Two Stage Pre Processing are described as follows. At first, the similarity matrix S of category is established. The establishment of similarity matrix is to gather statistics and to analyze the users' behavior browsing which can be acquired from web log data. In step two, it is to establish the first-order transition matrix P and second-order transition matrix

p 2 of Markov model. The transition matrix of Markov is established by the same approach, statistical method, from web log file.

In step three, the relevance matrix R is computed from first-order and second-order (or n-order) transition matrix of Markov model and similarity matrix. In the proposed method, the relevance is an important factor of prediction. Relevance can be used to infer the users' browsing behavior between web categories.

It is assumed that D denotes a database, which contains m users' usage record. It means that the users' session is recorded and D = {session1, session2 ,...,sessionm} is obtained. Each user's session can also be recorded as a sequential pattern of n web pages which is browsed by time order, and sessionp={page 1, page2,…., pagen}, where page i represents the user's visiting page at time j, is obtained. If a web site has k categories, then the user's session can be reorganized by sessionc={c1,c2,…., ck}, where ci = 0. After giving the definitions, more details for the prediction model are described in the following sections.

a) Similarity matrix of web categories

Proposed Research framework of step one is to create the similarity matrix from web log file. At first, the situation of categories in each user's session has to be understood. The vectori = <v1,i,…,vh,i,…,vm,i> for each category i is gather the i th element of session c from all m user sessions, vh,i = 1 means user h visited web page of category i otherwise vh,i = 0 . Two categories can be calculated the Set similarity and Euclidean distance. Euclidean distance is further normalized . The results are computed by similarity and Euclidean distance. They are combined into a weight total similarity equation.

SetSim(A,B)= A∩B/ AUB

After the similarity is calculated, the similarity matrix S is a k x k matrix of categories similarity, where Sij is the similarity between Ci and Cj that is established by above steps.

$$S = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_k \end{array} \\ \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1k} \\ S_{21} & S_{22} & \cdots & S_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ S_{k1} & S_{k2} & \cdots & S_{kk} \end{bmatrix} \end{array}$$

b) Transition matrix of Markov model

The proposed Research framework of step two is to create the transition matrix of Markov model P, which is based on web log file as well as similarity matrix. The P matrix is first-order transition matrix of Markov model and it is presented as follows:

$$P = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_k \end{array} \\ \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1k} \\ P_{21} & P_{22} & \cdots & P_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ P_{k1} & P_{k2} & \cdots & P_{kk} \end{bmatrix} \end{array}$$

Each element in the P matrix presents a transition probability between any two categories. Pij presents a transition probability which is calculated between category i and category j. The numerator is the number of transition times between category i and category j, and the denominator is the total number of transition times between category i and every category k. The transition matrix of P2 ,..., Pn can be calculated

c) Relevance matrix

Proposed Research framework of step three is to create the relevance matrix. The element Rij of relevance matrix is equal to product of Sij and Pij , which are acquired from similarity matrix and transition matrix of Markov model respectively. The relevance is an important factor of prediction between any two categories. The relevance can be used to infer the users' browsing behavior between categories. The relevance matrix is presented as follows:

Presents a relevance, which is calculated between category i at time t − n and category j at time t. More high the value of means more relevance between category i and category j.

$$R^n = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_3 \end{array} \\ \begin{bmatrix} R^n_{11} & R^n_{12} & \cdots & R^n_{1k} \\ R^n_{12} & R^n_{22} & \cdots & R^n_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ R^n_{k1} & R^n_{k2} & \cdots & R^n_{kk} \end{bmatrix} \end{array}$$

3.3 Clustering

Clustering is the unsupervised classification of patterns (observations, feature vectors and data items) into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis. Clustering is a difficult task combinatorial and differences of assumptions and contexts in different communities has made the transfer of useful generic concepts and methodologies slow to occur.

Clustering is useful in several exploratory pattern-analysis, grouping, machine-learning situations and decision-making including image segmentation , document retrieval, data mining and pattern classification. K-means clustering: It is the most intuitive and popular clustering algorithm, iteratively is partitioning a dataset into K groups in the vicinity of its initialization such that an objective function defined in terms of the total within-group sum-of-squares is minimized The simple definition of k-means clustering is to classify data to groups of objects based on attributes/features into K number of groups. K is positive integer number. K-means is Prototype-based (center-based) clustering technique which is one of the algorithms that solve the well-known clustering issues. It makes a one-level partitioning of data objects.

3.4 Two Levels of Prediction Model

Lee and Fu proposed a Two Levels of Prediction Model in 2008 (as Fig. 4) [14]. The model decreases the prediction scope using the two levels framework in a better manner. The two Prediction Levels of Model are designed by combining Markov model and Bayesian theorem. At level one the Markov model filters the most possible of categories which user will be browsing. At level two of this model, Bayesian theorem is used to infer precisely the highest probability of web page.

Figure 4. Two Levels of Prediction Model

In level one, it is to predict the most possible user's current state (web page) of category at time t, which depends on user's category at time t-1 and time t-2. Bayesian theorem predicts the most possible web pages at a time t according to user's states at a time t-1. Finally, the prediction result of two levels of prediction model is released.

In the Two Levels of Prediction Model framework (as Fig. 4), in step one, the similarity matrix S of category is established. The approach establishes similarity matrix to gather statistics and to analyze the users' behavior browsing which can be acquired from web log data.

In step two, it is to establish the first-order transition matrix P and second-order transition matrix P2 of Markov model. The transition matrix of Markov is established by the same approach, statistical method, from web log file.

3.5 View Selection and Prediction

The suitable view will be selected when predicting a user's browsing behavior. That is to choose the suitable relevant matrix for user. The view selection is executed by considering the distance between user session vector and the vectors of index table. Predicting the current user's browsing behavior through the Two Levels of Prediction Model after the suitable view is selected.

IV. EXPERIMENT ANALYSIS

In the experiment result of level one (as Fig. 5), there are five cases in level one which are Topr-1 to Topr-5. The Hit Ratio varies to 100% on Topr-5.

In the experiment result of level two (as Fig. 6). The Hit Ratio is from 51% on Topr-1 probability to 100% on Topr-2 and Topr-3 probability. The experiment results prove our revised prediction model has better predicting ability than Two Levels of Prediction Model.

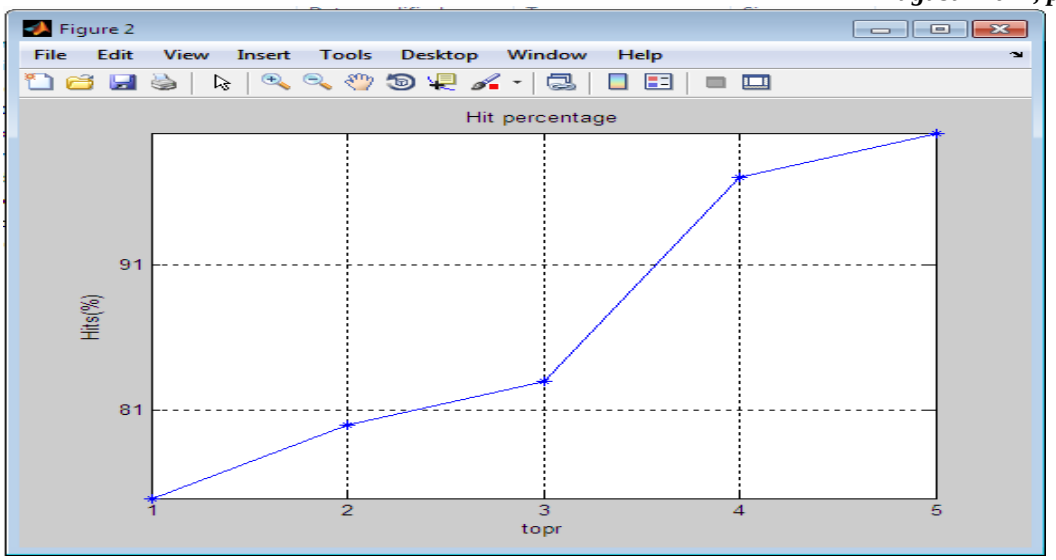

Figure 5: Showing Category Prediction

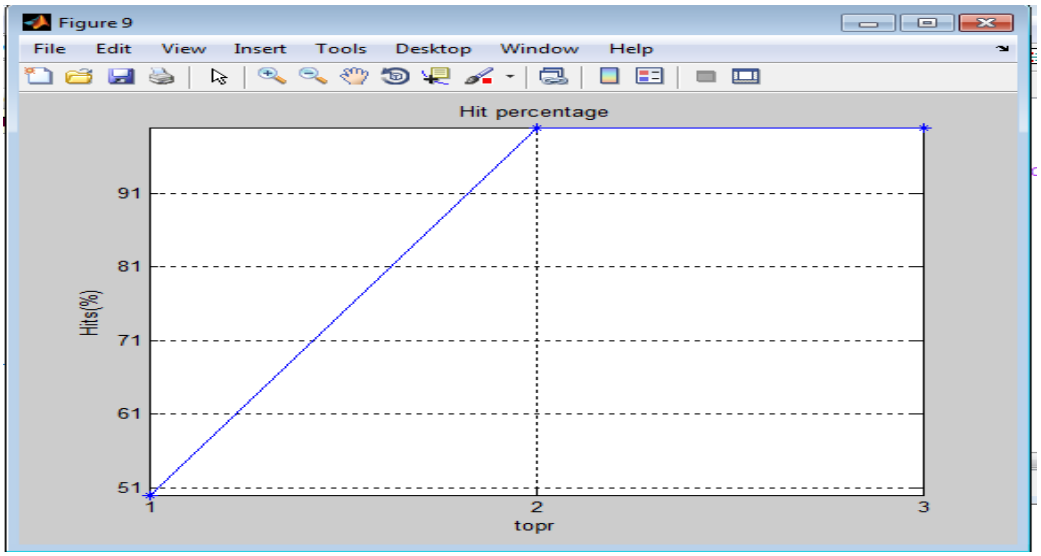

Figure 6: Showing Page Prediction

V. CONCLUSION

In this paper, it focuses on the preprocessing step and modifies the Two Levels of Prediction Model framework further. The information of clusters can be seen as cluster view for replacing of the global view. Therefore, we proposed a revised prediction model. The experiment results prove the Hit Ratio is better than before model.

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] Siriporn Chimphlee, Naomie Salim, Mohd Salihin Bin Ngadiman, Witcha Chimphlee, Surat Srinoy , "Rough Sets Clustering and Markov model for Web Access Prediction", in Proceedings of the Postgraduate Annual Research Seminar 2006.

[4] Christos Makris, Yannis Panagis, Evangelos Theodoridis,and Athanasios Tsakalidis "A Web-Page Usage Prediction Scheme Using Weighted Suffix Trees" © Springer-Verlag Berlin Heidelberg 2007.

[5] Vincent S. Tseng, Kawuu Weicheng Lin, Jeng-Chuan Chang "Prediction of user navigation patterns by mining the temporal web usage evolution" ©

[6] Chu-Hui Lee ,Yu-Hsiang Fu, "Web Usage Mining based on Clustering of Browsing Features" in the proceedings of Eighth International Conference on Intelligent Systems Design and Applications.

[7] S. Araya, M. Silva and R. Weber, "A Methodology for Web Usage Mining and Its Application to Target Group Identification," *Fuzzy Sets and Systems 148*, 2004, pp. 139-152.

[8] W. Bin and L. Zhijing, "Web Mining Research," *ICCIMA'03 IEEE*, 2003, pp. 84-89.

[9] S.K. De and P.R. Krishna, "Clustering web transactions using rough approximation," *Fuzzy Sets and Systems 148*, 2004, pp.131-138.

[10] F. M. Facca and P. Luca Lanzi, "Mining Interesting Knowledge from Weblogs: A Survey," *Data and Knowledge Engineering 53*, 2005, pp. 225-241.

[11] A.K. Jain, M.N. Murty, and P.J. Flynin, "Data Clustering: A Review," *ACM Computing Surveys*, Vol. 31, No 3, September 1999, pp. 265-323.

[12] P. Kumar, P.R. Krishna, R.S. Bapi and S.K. De, "Rough clustering of sequential data," *Data and Knowledge Engineering*, 2007.

[13] R. Kosala, H. Blockeel, "Web Mining Research: A Survey," *SIGKDD Explorations*, Volume 2, Issue 2, pp. 115.

[14] C.H. Lee and Y.H. Fu, "Two Levels of Prediction Model for Users' Browsing Behavior, *The 2008 IAENG International Conference on Internet Computing and Web Services (IMECS'08)*, 2008, pp.751-756.

[15] J. Srivastava, R. Cooley, M. Deshpande and P. N. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns form Web Data," *SIGKDD Explorations*, Vol. 1, Issue 2, Jan 2000, pp. 12-23.

[16] M. Spiliopoulou, "Web Usage Mining for Site Evaluation," Comm. ACM, vol. 43, no. 8, 2000, 127–134.

[17] 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), pp. 2-9, 2005.

[18] X. Chen and X. Zhang, "A Popularity-Based reduction Model for Web Pre fetching," Computer, pp. 63-70, 2003.

[19] R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the World Wide Web. In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), 1997.

[20] V.V.R. Maheswara Rao,Dr. V. Valli Kumari,, "An Efficient Hybrid Predictive Model to Analyze the Visiting Characteristics of Web User using Web Usage Mining",2010 International Conference on Advances in Recent Technologies in Communication and Computing.

