Two-level Prediction Model using k-means clustering for Web Browsers Usage

Nisha Yadav[1]
[1]Computer Science and Applications Department
[1]Kurukshetra University, Kurukshetra

Abstract— Due to the popularity of World Wide Web, many organizations have changed the way of doing business, which supplements the quick development of E-commerce directly and makes the development of web usage mining expertise crucial. Predicting of user's visiting behavior is an important technology of E- commerce application. The prediction results can be used for personalization, building proper websites, improving marketing strategy, promotion, product supply, getting marketing information, predicting market trends, and increasing the competitive strength of organizations etc. Markov model has been used for studying and understanding stochastic processes, and well suited for modeling and predicting a user's browsing behavior on a web at category level. Bayesian theorem can be used to infer users' browsing behaviors at webpage level. In this paper, we use the k-means clustering to cluster users' browsing behaviors. The prediction results by Two Levels of Prediction Model framework work well in general cases. However, Two Levels of Prediction Model suffer from the differences in user's behavior. The experiments will show that our model has higher hit ratio for prediction.

Keywords: Bayesian theorem, Markov model, Prediction, Web Usage Mining.

I. INTRODUCTION

The World Wide Web has become the world's largest knowledge vault. The popularity of World Wide Web is continuously increasing and is a golden mount of useful information [1]. Extracting knowledge and valuable patterns from the web efficiently and effectively is becoming a time consuming process. The rapid growth of the web has greatly increased the amount of valuable data in web server logs. Web mining can be divided into three categories as shown in fig. 1, web content mining, web structure mining, and web usage mining [2][3]. Web content mining focuses on automatic search and retrieval of useful information from web pages. Web Structure mining is used to analyze the links between web pages through the web structure to infer the knowledge presented in them. Web Usage Mining can be described as the discovery and analysis of user access patterns, through the mining of log files and associated data from a particular Web site which is accessed by the users.

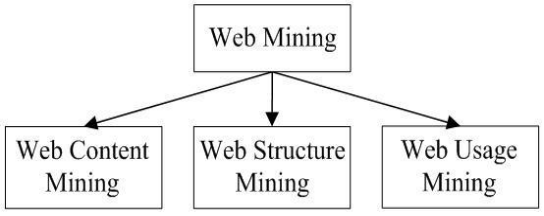

Fig. 1: Taxonomy of Web Mining

Due to the popularity of World Wide Web, many organizations have changed the way of doing business, which supplements the quick development of E-commerce directly and makes the development of web usage mining expertise crucial. Web usage mining is a technique of web mining and applies the data mining techniques on the Internet and web documents. It is used to find out and extract the pattern and the information in the web usage automatically, for example, association rule, clustering algorithm, and sequential pattern analysis etc [3].

The goal of web usage mining is to find out the useful information from web data or web log files. The other goals are to enhance the usability of the web information and to apply the technology on the web services, for example, pre-fetching and caching, personalization etc. For decision management, the results of web usage mining can be used for target advertisement, building proper websites, improving web design, improving satisfaction of customer, guiding the strategy decision of the enterprise, and forecasting marketing trends etc [4][5][6][7].

Forecasting the users' browsing behaviors is one of web usage mining applications. In order to achieve the purpose, it is necessary to understand the users' browsing behaviors by analyzing the information present in web data or web log files. Predicting the 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, increasing the competitive strength of organizations.

Markov model has been used for studying and understanding stochastic processes, and well suited for analyzing and predicting a user's browsing behavior on web at category level. The input to this problem is the sequence of web pages browsed by the user. In First order Markov model each action that can be executed by a user corresponds to a state in the model. A more complicated model computes the prediction by looking at the last two actions performed by the user. This is second order Markov model, and its states correspond to all possible pairs of actions that can be completed in sequence. This approach is generalized to the n^{th} order Markov model, which achieves the prediction by looking at the last N actions performed by the user [1].

In most of the applications, first order Markov model has low accuracy in performing accurate predictions which is why augmentation to higher order Markov models is necessary. All higher order Markov models has the advantage of higher prediction accuracy and more coverage but at the overhead of considerable increase in state space complexity. This led us to design models to combine different order Markov models so that the new model hold the advantage of both the low order Markov model and high order Markov model.

Bayesian theorem can be used to predict the most possible users' next request at the page level. Naïve Bayesian techniques try to find the probability that a condition will be satisfied, given only that it is known how many times it was satisfied in the past, and how many times it was not satisfied. In Bayesian theorem, some of the

information is used to revise the prior probability and obtain the posterior probability.

In this paper, we focus on the preprocessing step for improving the Two Levels of Prediction Model framework. Due to the diversity in users' browsing behavior, the k-means clustering algorithm is used to class users' browsing features [8][9][10][11]. Many different user clusters will be obtained and appear as cluster view for superseding the global view. The experiments will show that our prediction model will achieve higher hit ratio than Two Levels of Prediction Model.

The rest of this paper is organized as follows: Section 2 is the related work. The prediction model is proposed in Section 3. Section 4 shows the experiment results. Section 4 concludes the work and discusses the future work.

II. Related Work

The term web mining was first proposed by Etzioni in 1996. Many of the previous authors have manifested the urgency and importance of recognizing the users' visiting pattern available in web usage logs obtained

From different sources. Most of the work reported in the literature focus on pattern discovery to recognize the browsing behavior of the user. Several models have been proposed in the literature for recognizing the relationship between the pages without considering the category.

Myra Spiliopoulou [12] suggests employing web usage mining to website assessment to determine the needed improvements, basically to the site's design of page content and link structure between them.R. Walpole, R. Myres and S. Myres [13] proposed that Bayesian theorem can be utilized to forecast the most possible users' next request.Lee and Fu proposed a Two levels of Prediction Model in 2008[14]. The model reduces the prediction scope using two levels framework. The Two Levels of Prediction Model framework is designed by combining Markov Model and Bayesian Theorem. In The Two Levels of Prediction Model framework, the transition matrix is formed by examining all the users' browsing patterns and Bayesian theorem is used for presumption. The experimental results of the model are entirely well for predicting the class of web pages at the first level and predicting the web pages at the second level. However, the preprocessing step the model needs to improve further: the most users' properties are examined for making the transition matrixes. The transition matrix provided the global view for all users' behavior.

III. Two Levels of Prediction Model

Lee and Fu proposed a Two levels of Prediction Model in 2008(as Fig. 2) [14]. The model reduces the prediction scope using two levels framework. The Two Levels of Prediction Model framework is designed by combining Markov Model and Bayesian Theorem. At level one, Markov Model is used to filter the most possible class which will be browsed by user. At the second level, Bayesian theorem is used to determine exactly the highest possibility of web page.

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 is used to infer 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.

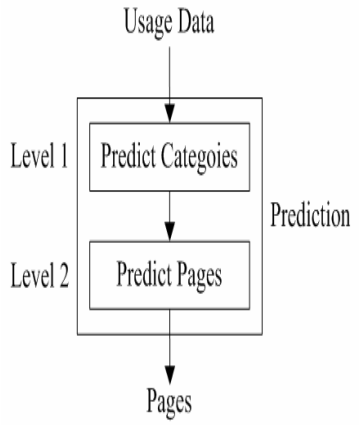

Fig. 2: Two Levels of Prediction Model [14]

In the Two Levels of Prediction Model framework (as Fig. 3), in step one, the similarity matrix *S* of category is established. The approach of establishing similarity matrix is to gather statistics and to analyze the users' behavior browsing which can be obtained from web log data.

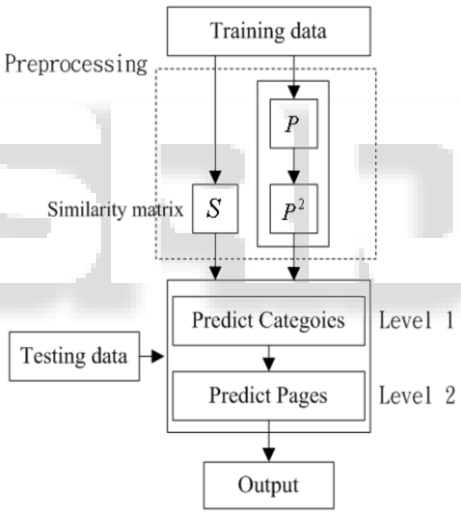

Fig. 3: Two Levels of Prediction Model Framework [14]

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.

IV. Preprocessing: K-means Clustering

In this research work, we use the k-means clustering for grouping users' browsing behaviors. The similarity measure is Euclidean distance function. In the initialization, every user is seen to be a cluster. The similar users' browsing feature will be found out and merged into a cluster until final condition is met. Finally, the user clusters and index table will be presented as output.

A. Pattern representation

In the task of pattern representation, user sessions are

created from web usage log files. User sessions can be organized as a *m* x *k* matrix as table 1, each row can be presented by $session^u = (p_{u,1}, p_{u,2}, ..., p_{u,k})$ where $p_{u,i} = 1$ denotes that user u browsed the web page otherwise $p_{u,i} = 0$. The *k* is the number of web pages. The session means the user's browsing situation which is also user's browsing feature.

	P_1	P_2	…	P_k
$session^1$	1	0	…	1
$session^2$	0	1	…	0
			…	
$session^m$	1	1	…	1

Table. 1: User Sessions

B. *Definition Similarity Measure*

The similarity between any two users can be calculated by distance measure. Euclidean distance function (1) is used for computing the similarity between user *i* and user *j*, the similarity can be presented by $Sim(user_i, user_j) = (session^i, session^j)$.Euclidean distance is further normalized by equation(2). Further, the *m* x *m* matrix of user similarity will be obtained.

Euclidean distance:

$$D(A,B) = \sqrt{\sum_{i=1}^{m} (A_i - B)_i^2}$$

Normalization:

$$N(D(A,B)) = 1 - \sqrt{\sum_{i=1}^{m} (A_i - B)_i^2}/m$$

C. *Clustering*

The simplest unsupervised learning algorithm that solve clustering problem is K- Means algorithm. It is a simple and easy way to classify a given data set through a certain number of clusters. When the documents are clustered [15] using K-Means algorithm, the cluster contains more similar documents and it increases the relevancy rate of search results. K-means clustering is the simplest and most commonly used clustering algorithm. Initially the algorithm takes the collection of documents, number of clusters (K) and centroids of each cluster as input. Algorithm finds the distance of documents from the initial centroids and documents are assigned to nearer clusters. This process continues until some stopping criterion is met. Selection of initial centroids and the number of clusters (K) is done randomly. Once the clusters are formed, the cluster labels are generated by finding the terms with high frequency inside the documents. The algorithm is composed of the following steps:

1) Initialize k cluster centers to be seed points (These centers can be randomly produced or use other ways to generate).
2) For each sample, find the nearest cluster center, put the sample in this cluster and re-compute centers of the altered cluster (Repeat n times).
3) Exam all samples again and put each one in the cluster identified with the nearest center (don't re-compute any cluster centers). If members of each cluster haven't been changed, stop. If changed, go to step 2.

D. *View Selection and Prediction*

The suitable view will be selected when predicting a user's browsing behavior. That is to choose the suitable relevance matrix for user. The view selection is performed by examining 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.

V. EXPERIMENT

The source of database in the experiment is from UCI Dataset Repository [16]. This repository contains a lot of different research fields. Therefore, this repository is very popular in the research field. The dated of the database is September, 28, 1999 on msnbc.com website and a part of news data is from msn.com website. Each sequence data is corresponding to users' page views. The time interval is 24 hours. The categories are presented in sequential data. The 17 types of categories are frontpage, news, tech, local, opinion, on-air, music, weather, health, living, business, sports, summary, bbs, travel, msn-news, and msn-sports. The number of category is represented from 1 to 17 in the sequential data.

The UCI Dataset Repository is a quite popular database, but the users' browsing behavior is just recorded by category in the database, the user's web pages visiting are simulated by Gaussian distribution. The numbers of web pages in each category are assumed equal.

The experiment environment in this paper is Dell inspiron. The hardware is performed on Intel® Core TM(2) Duo CPU, T6600 @2.20 GHZ and Microsoft Windows XP operating system. The software is performed on MATLAB R2012b.

There are five cases in level one prediction, which are Top-1, Top-2, Top-3, Top-4 and Top-5. The Hit Ratio is from 0% on top-1 relevance, 16.67% on top-2 relevance to 100% on Top-5 relevance (as Fig. 4).

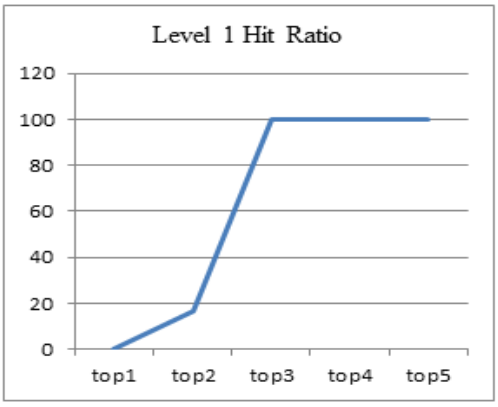

Fig. 4: Hit Ratio of Level 1

In the experiments for level two, there are five cases for the prediction result, which are Top-1, Top-2, Top-3, Top-4 and Top-5. The results are shown in the Fig. 5. For example, the Hit Ratio is from 75% (Top-1) to 98% (Top-2) and 100% on Top5.

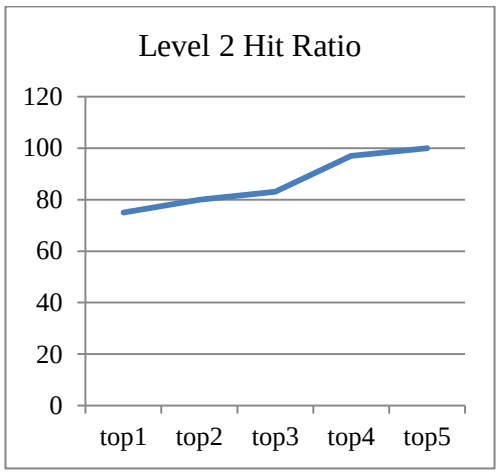

Fig. 5: Hit ratio of Level 2.

VI. CONCLUSION AND FUTUTRE WORK

Because of the huge quantity of data of web pages on many websites, for convenience, are to gather the web pages based on category. Users' browsing behavior has been observed at two levels to meet the nature of the websites. One is category level and the other is web page level. In level one is to predict category. The unnecessary categories can be excluded. The scope of calculation is greatly reduced. Next, using Bayesian theorem in the level two to predict the users' browsing page is more effective and accurate. 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 modified prediction model. The experiment results prove the Hit Ratio is better than before model.

REFERENCES

[1] V. V. R Maheswara Rao and Dr. V. Valli Kumari "An Effficient Hybrid Predictive Modelto Anayze the Visiting characteristics of Web User using Web Usage Mining" International Conference on Advances in Recent Technologies in Communication and Computing,2010, pp. 225-230.
[2] 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.
[3] F. M. Facca and P. Luca Lanzi, "Mining Interesting Knowledge from Weblogs: A Survey," Data and Knowledge Engineering 53, 2005, pp. 225-241.
[4] W. Bin and L. Zhijing, "Web Mining Research," ICCIMA'03 IEEE, pp. 84-89, 2003.
[5] D. Dhyani, W. K. Ng and , S. S. Bhowmick, "A Survey of Web Metrics," ACM Computing Surveys, Vol. 34 2002.
[6] R. Kosala, H. Blockeel, "Web Mining Research: A Survey,"SIGKDD Explorations, volume 2, Issue 2, pp. 115-120.
[7] S. Jespersen, T. B. Pedersen and J. Thorhauge, "Evaluating the Markov Assumption for Web Usage Mining," WIDM'03, pp. 82-89,2003.
[8] S.K. De and P.R. Krishna, "Clustering web transactions using rough approximation," Fuzzy Sets and Systems , 2004,pp. 131-138.
[9] K. Jain, M.N. Murty, and P.J. Flynin, "Data Clustering A Review," ACM Computing Surveys,1999.
[10] P. Kumar, P.R. Krishna, R.S. Bapi and S.K. De, "Rough clustering of sequential data," Data and Knowledge Engineering, 2007.
[11] Q. Song and M. Shepperd, "Mining web browsing pattern for E-commerce," Computers in Industry 57, 2006, pp.622-630.
[12] M. Spiliopoulou, "Web usage Mining for Site Evaluation" Comm. ACM, vol. 43, no. 8, 2000, pp. 127-134.
[13] R. Walpole, R. Myers, S. Myers and K. Ye, "Probability and Statistics for Engineers and Scientists," in Paperback, 7 ed., Pearson Education, 2002, pp. 82-87.
[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] M. Steinbach, G. Karypis, and V. Kumar "A comparison of Document Clustering Techniques," Proc KDD-2000 Workshop on Text Mining, Aug. 2000.
[16] UCI KDD archive, http://kdd.ics.uci.edu/.

