Using Markov Model and Popularity and Similarity-based Page Rank Algorithm for Web Page Access Prediction

Phyu Thwe

***Abstract*—**With the rapid growth of the World Wide Web, the Internet has become a huge repository of data and serves as an important platform for the dissemination of information. The users' accesses to Web sites are stored in Web server log file. These series can be considered as a web access pattern which is helpful to find out the user behaviour. Through this behaviour information, we can find out the accurate user next request prediction that can reduce the browsing time of web pages. In this paper, we proposed to use clustering techniques to cluster the data sets from the results of the preprocessing phase. As a result, a more accurate Markov model is built based on each group rather than the whole data sets. The accuracy of low order Markov model is normally not satisfactory. Therefore, we use popularity and similarity-based page rank algorithm to make prediction when the ambiguous results are found.

***Keywords*—** Page Rank Algorithm, Web Log Mining, Web Page Prediction, Web Usage Mining.

I. Introduction

THE study of web mining The study of Web mining techniques to discover useful knowledge has become increasingly important because more organizations rely on the Internet to conduct daily business. However, with the magnitude and diversity of available information from the Internet, it is important to locate the relevant information to satisfy the requirements of people with different backgrounds. To assist Web surfers in browsing the Internet more efficiently, one of the topics that have attracted much attention is modelling the Web user's browsing patterns and making prediction.

Many researchers have applied data mining to extract user navigation patterns from Web logs. The discovered patterns can be used to predict which pages are likely to be clicked by a user given a sequence of pages that the user already visited. With the prediction, the server can automatically send the predicted pages to the client cache before the pages are actually requested to reduce the network latency. Alternatively, the predicted pages can be recommended to the user by dynamically generating the links to the pages on the current browsing window of the user to help the user find relevant information more efficiently. The techniques that have been used to find such patterns include association rule mining, sequential pattern mining and clustering.

The rest of the paper is organized as follows: We overview some related work in Section 2. We present the required theoretical background concerning the Markov Models and the Page Rank Algorithm in Section 3. We provide the architecture of the proposed system in which PSPR can be applied in Section 4. We prove that this PSPR algorithm can be applied to any web site's log data. Finally, we conclude with our system in Section 5.

Phyu Thwe is with the University of Technology (Yatanarpon Cyber City, Pyin Oo Lwin, Myanmar (e-mail: pthwe19@gmail.com).

II. Related Work

In [4], users' browsing behavior will be predicted at two levels to meet the nature of the navigation. One is category stage and the other is web page stage. In stage one is to predict category. The unnecessary categories can be excluded. The scope of calculation is massively reduced. A dynamic clustering-based method is used to increase a Markov model's accuracy in representing a collection of user web navigation sessions [5]. The method makes use of the state cloning concept to duplicate states in a way that separates in-links whose corresponding second-order probabilities diverge. In [6], a rough set clustering is presented to cluster web transactions from web access logs and Markov model is used for next access prediction. Using this approach, users can effectively mine web log records to discover and predict access patterns.

In [7], this paper investigates into using Markov chains to make link prediction and the transition matrix derived from Markov chains to acquire structural knowledge about Web sites. In this paper, a method for predicting the ranking position of a Web page is proposed [8]. Assuming a set of successive past top-*k* rankings, the evolution of Web pages in terms of ranking trend sequences used for Markov Models training is studied, which are in turn used to predict future rankings. A hybrid probabilistic predictive model extending the properties of Markov models by incorporating link analysis methods is presented in [9]. More specifically, the use of a Page Rank-style algorithm is proposed for assigning prior probabilities to the web pages based on their importance in the web site's graph.

In [11] Usage Based Page Rank (UPR) is introduced. UPR is a variation of the Page Rank algorithm, based on the visit

frequency data obtained from previous users' sessions. In [12] Access time-length and frequency-based Page Rank (TFPR) is introduced and is based on the length of time spent on visiting a page and the frequency that a page was visited. In [13] Duration based Page Rank (DPR) and Popularity based Page Rank (PPR) is introduced. Duration Based Rank (DPR), which focuses on page duration with size proportion. Popularity Based Page Rank (PPR) ranking model, which focuses on both page duration with size proportion and frequency value of page visits.

III. Background Theory

A. Markov Model

The 1st-order Markov models (Markov Chains) provide a simple way to capture sequential dependence [2, 3, 9], but they do not take into consideration the long-term memory aspects of web surfing behaviour since they are based on the assumption that the next state to be visited is only a function of the current one. Higher-order Markov models are more accurate for predicting navigational paths. But, there exists a trade-off between improved coverage and exponential increase in state space complexity as the order increases. Moreover, such complex models often require inordinate amounts of training data, and the increase in the number of states may even have worse prediction accuracy and can significantly limit their applicability for applications requiring fast predictions, such as web personalization. There have also been proposed some mixture models that combine Markov models of different orders. However, such models require much more resources in terms of preprocessing and training. Therefore, it is evident that the final choice that should be made concerning the kind of model that is to be used, depends on the trade-off between the required prediction accuracy and model's complexity/size.

B. Page Rank Algorithm

Page Rank is used to determine the importance of the page on the web. Surgey Brin and Larry Page [1] proposed a ranking algorithm named Page Rank (PR) that uses the link structure of the web to determine the importance of web pages. According to this algorithm, if a page has important links to it, then its links to other pages also become important. Therefore, it takes back links into account and propagates the ranking through links. In Page Rank, the rank score of a page is equally divided among its outgoing links and that values of outgoing links are in turn used to calculate the ranks of pages pointed by that page.

Page Rank [11] is the most popular link analysis algorithm, used broadly for assigning numerical weightings to web documents and utilized from web search engines in order to rank the retrieved results. The algorithm models the behavior of a random surfer, who either chooses an outgoing link from the page he's currently at, or "jumps" to a random page after a few clicks.The Web is treated as a directed graph G = (V, E), where V is the set of vertices or nodes, i.e., the set of all pages, and E is the set of directed edges in the graph, i.e., hyperlinks. In page rank calculation, especially for larger systems, iterative calculation method is used. In this method, the calculation is implemented with cycles. In the first cycle all rank values may be assigned to a constant value such as 1, and with each iteration of calculation, the rank value become normalized within approximately 50 iterations under $\varepsilon = 0.85$.

IV. System Architecture

The processing steps of the system have three main phases. Preprocessing is performed in the first phase. The second phase is clustering web sessions using K-means clustering. In the final phase, Markov model is used to predict next page access based on resulting web sessions. The popularity and similarity-based page rank algorithm is used to decide the most relevant answer if the ambiguous result is found in Markov model prediction.

The input of the proposed system is a web log file. A web log is a file to which the web server writes information each time a user requests a resource from that particular site.

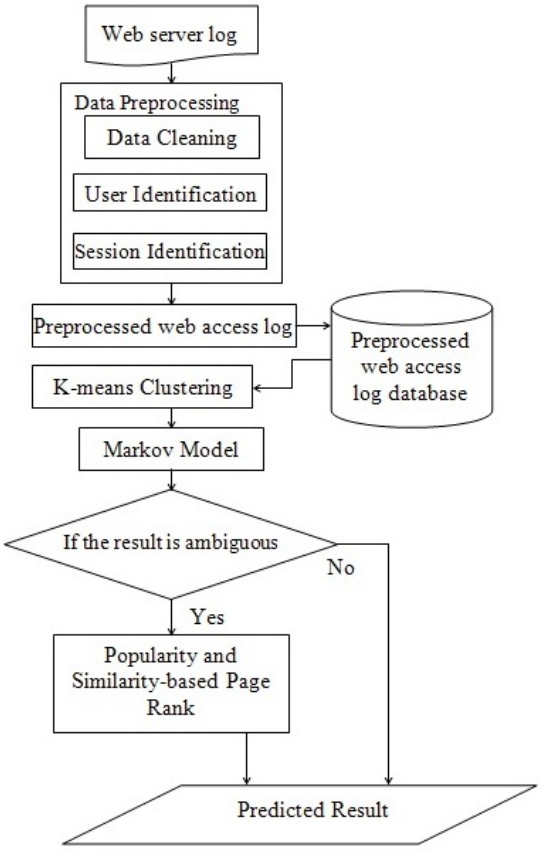

Fig. 1 Processing step of the proposed system

A. Data Preprocessing

The data preprocessing includes three basic steps [10]. They are:

- Data Cleaning
- User Identification
- Session Identification

The first issue in the preprocessing phase is data cleaning. Web log data may need to be cleaned from entries involving pages that returned an error or graphics file accesses. In some cases such information might be useful, but in others such data should be eliminated from a log file.

User identification means identifying each user accessing Web site, whose goal is to mine every user's access characteristic. There are several ways to identify individual visitors. The most obvious solution is to assume that each IP address (or each IP address/client agent pair) identifies a single visitor.

The next step is to perform session identification, by dividing the click stream of each user into sessions. The usual solution in this case is to set a minimum timeout and assume that consecutive accesses within it belong to the same session, or set a maximum timeout, where two consecutive accesses that exceed it belong to different sessions.

B. Popularity and Similarity-Based Page Rank Algorithm

Popularity and Similarity-based Page Rank (PSPR) calculation simply depends on the duration values of pages and transitions, the frequency value of pages and transitions, their web page file size and similarity of web page [15]. The popularity value of page rank was discussed in [13]. Popularity defines in two dimensions. They are page dimension and transition dimension. For both dimensions, popularity defines in terms of time user spends on page, size of page and visit frequency of page. Page popularity is needed for calculating random surfer jumping behaviour of the user and transition popularity is needed for calculating the normal navigating behaviour of the user.

Similarity of web page is important to predict next page access because million of users generally access the similar web page in a particular Web site. The calculation of the similarity is based on web page URL. The content of pages is not considered and this calculation does not need for making a tree structure of the Web site. For example, suppose "/shuttle/missions/sts-73/mission-sts-73.html" and "/shuttle/missions/sts-71/mission-sts-71.html" are two requested pages in web log. These two URLs are stored in string array by dividing "/" character. And then, we compute the length of the two arrays and give weight to the longer array: the last room of the array is given weight 1, the second to the last room of the array is given weight 2, the third to given weight 3 and so on and so forth, until the first room of the array is given higher length of the array. The similarity between two strings is defined as the sum of the weight of those matching substrings divided by the sum of the total weights.

This similarity measurement includes:

(1) $0 <= SURL_{j\to i} <= 1$, i.e. the similarity of any pair of web pages is between 0.0 and 1.0;

(2) $SURL_{j\to i} = 0$, when the two web pages are totally different;

(3) $SURL_{j\to i} = 1$, when the two web pages are exactly same.

$$PSPR_i = \varepsilon \times \sum_{P_j \in In(P_i)} \left[PSPR_j \times \frac{w_{j\to i}}{\sum_{P_k \in Out(P_j)} w_{j\to k}} \times \frac{(d_{j\to i}/s_i)}{\max(d_{m\to n}/s_n)} \times \frac{SURL_{j\to i}}{\sum_{P_k \in Out(P_j)} SURL_{j\to k}} \right] + (1-\varepsilon) \times \frac{w_i}{\sum_{P_j \in WS} w_j} \times \frac{(d_i/s_i)}{\max(d_m/s_m)} \quad (1)$$

In the equation 1, ε is a damping factor and usually ε = 0.85. $In(p_i)$ is the set that keeps the in-links of that page. $Out(p_j)$ is the set of pages that point to p_j. $w_{j\to i}$ is the number of times pages j and i appear consecutively in all user sessions. $d_{j\to i}$ is the duration of the transaction and s_i is the size of the transition's result page. WS is the web session. $SURL_{j\to i}$ is the similarity of web page j to page i. $\frac{w_{j\to i}}{\sum_{P_k \in Out(P_j)} w_{j\to k}} \times \frac{(d_{j\to i}/s_i)}{\max(d_{m\to n}/s_n)}$ is the transition popularity based on transition frequency and duration. $\frac{SURL_{j\to i}}{\sum_{P_k \in Out(P_j)} SURL_{j\to k}}$ is the similarity calculation between web pages. $\frac{w_i}{\sum_{P_j \in WS} w_j}$ is the frequency calculation for page i. $\frac{(d_i/s_i)}{\max(d_m/s_m)}$ is the average duration calculation for page i. The popularity of page is calculated based on page frequency and average duration of page.

By using this equation, we can calculate the popularity and similarity-based page rank (PSPR) for every page. In order to make rank calculations faster, we record required steps of our calculations to database. The step values related to rank calculations are, average duration value of pages, average duration values of transitions, page size, frequency value of pages, frequency value of transitions, the similarity value of pages. The result can be used for ambiguous result found in Markov model to make the correct decision.

V. Experimental Evaluation

This paper introduces a method that integrates clustering, Markov model and page rank algorithm in order to improve the Web page prediction accuracy. In this section, we present experimental results to evaluate the performance of our algorithm. Overall our experiment has verified the effectiveness of our proposed techniques in web page access prediction based on a particular website.

For our experiments, we used NASA web server data sets. We obtained the web logs in August, 1995 and used the web logs from 01/Aug/1995 to 15/Aug/1995 as the training data

set. For the first testing data set, the web logs from 16/Aug/1995 to 17/Aug/1995 are used. We filtered the records (such as *.jpg, *.gif, *.jpeg) and only reserved the hits requesting web pages. When identifying user sessions, we set the session timeout to 30 minutes, with a minimum of 10 pageviews per session. After filtering out the web session data by preprocessing, the training data set contained 94307 records and 5574 sessions, while the testing data set contained 1271 records and 74 sessions. For the second testing data set, we used the web logs from 16/Aug/1995 to 22/Aug/1995. After filtering out the web session data, the testing data set contained 60612 records and 2482 sessions.

In comparing the predictions with the real page visits, we use two similarity algorithms that are commonly preferred for finding similarities of two sets. The first one is called Osim [11, 12, 13] algorithm, which calculates the similarity of two sets without considering the ordering of the elements in the two sets between A and B and is defined as:

$$OSim(A,B) = \frac{A \cap B}{n} \qquad (2)$$

As the second similarity metric we use Ksim similarity algorithm, which concerns Kendall Tau Distance [11, 12, 13] for measuring the similarity of next page prediction set produced by training data set and real page visit set on the test data. Kendall Tau Distance is the number of pairwise incompatibility between two sets.

$$KSim(r_1,r_2) = \frac{\left|(u,v) : r_1', r_2' \, have same ordering of \, (u,v), u \neq v\right|}{\left|A \cap B\right|(\left|A \cap B\right| - 1)} \qquad (3)$$

Where, r_1' is an extension of r_1, containing all elements included in r_2 but not r_1 at the end of the list (r_2' is defined analogously). In our experiment setup, we make experiment with top-3 comparison that are measured by Ksim and Osim methods.

A. Experimental Results

The results of the experiment for the next page prediction accuracy for popularity and similarity-based page ranking algorithm under Ksim and Osim similarity metrics are given in Figure 2.

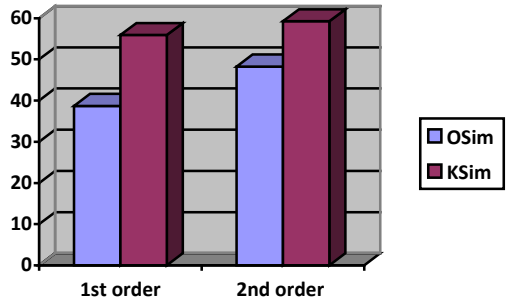

Fig. 2 Average OSim, KSim values for top-3 prediction

As depicted in Figure 2, PSPR based on 2nd order Markov model outperforms PSPR based on 1st order Markov model significantly in all OSim and KSim values. In the top-3 prediction, PSPR based on 2nd order markov model outperforms 48.24% and 59.25% for OSim and KSim respectively, while PSPR based on 1st order Markov model outperforms 38.68% and 55.92% respectively for the first testing data set. Therefore, we can confirm that popularity and similarity-based page rank depend on 2nd order Markov model can improve the accuracy of Web page prediction.

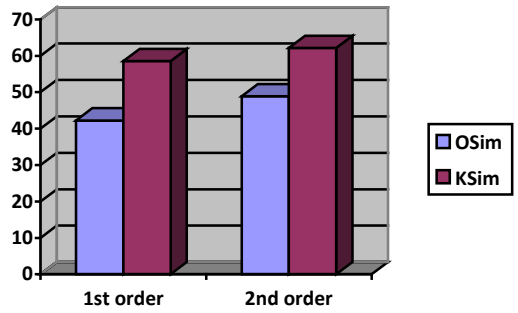

Fig. 3 Average OSim, KSim values for top-3 prediction

As depicted in Figure 3, PSPR based on 2nd order Markov model outperforms PSPR based on 1st order Markov model significantly in all OSim and KSim values. In the top-3 prediction, PSPR based on 2nd order Markov model outperforms 48.89% and 62.13% for OSim and KSim respectively, while PSPR based on 1st order Markov model outperforms 42.25% and 58.54% respectively for the second testing data set. Therefore, we can confirm that popularity and similarity-based page rank depend on 2nd order Markov model can improve the accuracy of Web page prediction.

Web page access prediction can be useful in many applications [14]. The improvement for accuracy can make a change in the web advertisement area. Using web page access prediction, the right advertisement will be added according to the users' browsing patterns. Also, web page access prediction helps web administrators restructure the Web sites to improve site topology and user personalization as well as market segmentation. Web page access prediction is also helpful for caching the predicted page for faster access and for improving browsing and navigation orders.

VI. Conclusion

In this paper, we used Markov model and popularity and similarity-based page rank algorithm for web page access prediction. In our experiment, we observed that in both cases PSPR based on 2nd order Markov Model are more than promising PSPR based on 1st order Markov Model in terms of accuracy (OSim and KSim). Higher order Markov model result in better prediction accuracy since they look at previous browsing history. But they are associated with higher state space complexity.

REFERENCES

[1] S. Brin, L. Page, 1998. The anatomy of a large-scale hypertextual Web search engine, Computer Networks, 30(1-7): 107-117, Proc. of WWW7 Conference.

[2] F.Khalil, J. Li and H. Wang, 2007. Integrating markov model with clustering for predicting web page accesses. Proceedings of the 13th Australasian World Wide Web Conference (AusWeb 2007), June 30-July 4, Coffs Harbor, Australia, pp: 1-26.

[3] M. Deshpande and G. Karypis. May 2004. Selective markov models for predicting web page accesses. ACM Trans. Internet Technol., 4:163-184.

http://dx.doi.org/10.1145/990301.990304

[4] V.V.R.Maheswara Rao, Dr. V. Valli Kumari, "An Efficient Hybrid Successive Markov Model for Predicting Web User Usage Behavior using Web Usage Mining", International Journal of Data Engineering (IJDE) Volume (1): Issue (5), pp 43-62

[5] J. Borges, M. Levene, "A Dynamic Clustering-Based Markov Model for Web Usage Mining", May 26, 2004

[6] S. Chimphlee, N. Salim, M. S. B. Ngadiman, W. Chimphlee, S. Srinoy, "Rough Sets Clustering and Markov model for Web Access Prediction", Proceedings of the Postgraduate Annual Research Seminar 2006

[7] J. Zhu, "Using Markov Chains for Structural Link Prediction in Adaptive Web Sites"

[8] M. Vazirgiannis, D. Drosos, P. Senellart, A. Vlachou, "Web Page Rank Prediction with Markov Models", April 21-25, 2008 · Beijing, China

[9] M. Eirinaki, M. Vazirgiannis, D. Kapogiannis, Web Path Recommendations based on Page Ranking and Markov Models, WIDM'05, November 5, 2005, Bremen, Germany

[10] R. Khanchana, Dr. M. Punithavalli, Web Page Prediction for Web Personalization:A Review, Global Journal of Computer Science and Technology Volume 11 Issue 7 Version 1.0 May 2011, ISSN: 0975-4172 & Print ISSN: 0975-4350

[11] M. Eirinaki and M. Vazirgiannis. Nov. 2005. Usage-based pagerank for web personalization. In Data Mining, Fifth IEEE International Conference on, page 8 pp.

[12] Y. Z. Guo, K. Ramamohanarao, and L. Park. Nov. 2007. Personalized pagerank for web page prediction based on access time-length and frequency. In Web Intelligence, IEEE/WIC/ACM International Conference on, pages 687-690.

[13] B. D. Gunel, P. Senkul, 2011. Investigating the Effect of Duration, Page Size and Frequency on Next Page Recommendation with Page Rank Algorithm, ACM.

[14] F. Khalil, J. Li and H. Wang, 2008. "Integrating Recommendation Models for Improved Web Page Prediction Accuracy", the Thirty-First Australasian Computer Science Conference (ACSC2008), Wollongong, Australia.

[15] P. Thwe, "Proposed Approach for Web Page Access Prediction Using Popularity and Similarity based Page Rank Algorithm", International Journal of Science and Technology Research, Volume 2 - Issue 3, March 2013 Edition [ISSN 2277-8616].