International Journal of Computer Science and Business Informatics

IJCSBI.ORG

Web Page Access Prediction based on an Integrated Approach

Phyu Thwe
Faculty of Information and Communication Technology,
University of Technology (Yatanarpon Cyber City),
Pyin Oo Lwin, Myanmar

ABSTRACT

Predicting the user's web page access is a challenging task that is continuing to gain importance as the web. Understanding users' next page access helps in formulating guidelines for web site personalization. Server side log files provide information that enables to build the user sessions within the web site, where a user view a session consists of a sequence of web pages within a given time. A web navigation behavior is helpful in understanding what information of online users demand. In this paper, we present the system that focuses on the improvements of predicting web page access. We proposed to use clustering techniques to cluster the web log data sets. As a result, a more accurate Markov model is built based on each group rather than the whole data sets. Markov models are commonly used in the prediction of the next page access based on the previously accessed pages. Then, we use popularity and similarity based-page rank algorithm to make prediction when the ambiguous results are found. Page Rank represents how important a page is on the web. When one page links to another page, it is a vote for the other page. The more votes for a page, the more important the page must be.

Keywords

Web Log Mining, Web Page Access Prediction, K-means Clustering, Markov Model, Page Rank Algorithm.

1. INTRODUCTION

As Internet is becoming an important part of our life, the quality of the information is more considered and how it is displayed to the user. The research area of this work is web data analysis and methods how to process this data. This knowledge can be extracted by collecting web servers' data log files, where all users' navigational patterns about browsing are recorded. Server side log files provide information that enables to rebuild the user sessions within the particular web site, where a user view a session consists of a sequence of web pages within a given time. A web navigation behavior is helpful in understanding what information of online users demand. Following that, the analyzed results can be seen as knowledge to be used in intelligent online applications, refining web site maps, web based personalization system and improving searching accuracy when seeking information. However, an online navigation behavior grows each passing day, and thus extracting information intelligently from it is a difficult issue. Web Usage Mining (WUM) is the process of extracting knowledge from Web users' access

data by using Data Mining technologies. It can be used for different purposes such as personalization, business intelligence, system improvement and site modification.

In this paper, we present the system that focuses on the improvements of predicting web page access. Data preprocessing is the process to convert the raw data into the data abstraction necessary for further applying the data mining algorithm. We proposed to use clustering techniques to cluster the data sets so that homogenous sessions are grouped together. As a result, a more accurate Markov model is built based on each group rather than the whole data sets. The proposed Markov model is low order Markov model so that the state space complexity is kept to a minimum. 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.

The rest of this paper is organized as follows: Section 2 describes the theory background about preprocessing technique, Markov Model and Page Rank Algorithm. In section 3, we review some researches that advance in web page access prediction. Section 4 describes the proposed method for the predicting of web page access in web log file. Results of an experimental evaluation are reported in section 5. Finally, section 6 summarizes the paper.

2. BACKGROUND STUDY

2.1 Preprocessing Technique

2.1.1 Data Cleaning

This step is to remove all the data useless for data analyzing and mining e.g. requests for graphical page content (e.g., jpg and gif images); requests for any other file which might be included into a web page; or even navigation sessions performed by robots and web spiders. The quality of the final results strongly depends on cleaning process. Appropriate cleaning of the data set has profound effects on the performance of web usage mining. The discovered associations or reported statistics are only useful if the data represented in the server log gives an accurate picture of the user accesses to the Web site.

The procedures of general data cleaning are as follows:

- Firstly, it should remove entries that have status of “error” or “failure”. It’s to remove the noisy data from the data set. To accomplish it is quite easy.
- Secondly, some access records generated by automatic search engine agent should be identified and removed from the access log. Primarily, it should identify log entries created by so-called crawlers or spiders that are used widely in Web search engine tools. Such data offer nothing to the analyzing of user navigation behaviors [6].

2.1.2 User Identification

There are some heuristics for user identification [6]. The first heuristic states two accesses having the same IP but different browser or operation system, which are both recorded in agent field, are originated from two different users. This heuristic

is that a user, when navigating the web site, rarely employs more than one browser, much more than one OS. But this method will render confusion when a visitor actually does like that. The second heuristic states that when a web page requested is not reachable by a hyperlink from any previously visited pages, there is another user with the same IP address. But such method introduces the similar confusion when user types URL directly or uses bookmark to reach pages not connected via links.

2.1.3 Session Identification

To define user session, two criteria are usually considered [6]:

1. Upper limit of the session duration as a whole;
2. Upper limit on the time spent visiting a page.

Generally the second method is more practical and has been used in Web Usage. It is achieved through a timeout, where if the time between page requests exceeds a certain limit, it is assumed that the user is starting a new session. Many commercial products use 30 minutes as a default timeout. Once a site log has been analyzed and usage statistics obtained, a timeout that is appropriate for the specific Web site can be fed back into the session identification algorithm.

2.2 Markov Model

The 1st-order Markov models (Markov Chains) provide a simple way to capture sequential dependence [14, 15, 16], 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.

2.3 Page Rank Algorithm

Page Rank is used to determine the importance of the page on the web. Surgey Brin and Larry Page [13] 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 behaviour 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 $\varepsilon = 0.85$.

RELATED WORKS

In recent years, there has been an increasing number of research works done with regard to web usage mining. They [1] describe a prediction system to predict the future occurrence of an event that is a prediction system based on fuzzy logic. A subtractive clustering based fuzzy system identification method is used to successfully model a general prediction system that can predict future events by taking samples of past events. Historical data is obtained and is used to train the prediction system. Recent data are given as input to the prediction system. All data are specific to the application at hand. The system, that is developed using Java, is tested in one of the many areas where prediction plays an important role, the stock market. Prices of previous sessions of the market are taken as the potential inputs. When recent data are given to the trained system, it predicts the possibility of a rise or a fall along with the next possible value of data.

The prediction models [2] that are based on web log data that corresponds with users' behavior. They are used to make prediction for the general user and are not based on the data for a particular client. This prediction requires the discovery of a web users' sequential access patterns and using these patterns to make predictions of users' future access. They will then incorporate these predictions into the web prefetching system in an attempt to enhance the performance.

In [3], it is proposed to integrate Markov model based sequential pattern mining with clustering. A variant of Markov model called dynamic support pruned all k^{th} order Markov model is proposed in order to reduce state space complexity. Mining the web access log of users of similar interest provides good recommendation accuracy. Hence, the proposed model provides accurate recommendations with reduced state space complexity.

An Efficient Hybrid Successive Markov Prediction Model (HSMP) is introduced in [4]. The HSMP model is initially predicts the possible wanted categories using Relevance factor, which can be used to infer the users' browsing behavior between web categories. Then predict the pages in predicted categories using techniques for intelligently combining different order Markov models so that the resulting model

has low state complexity, improved prediction accuracy and retains the coverage of the all higher order Markov model.

In [5], they propose the use of CRFs(Conditional Random Fields) in the field of Web page prediction. They treat the previous Web users' access sessions as observation sequences and label these observation sequences to get the corresponding label sequences, then they use CRFs to train a prediction model based on these observation and label sequences and predict the probable subsequent Web pages for the current users.

3. PROPOSED 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.

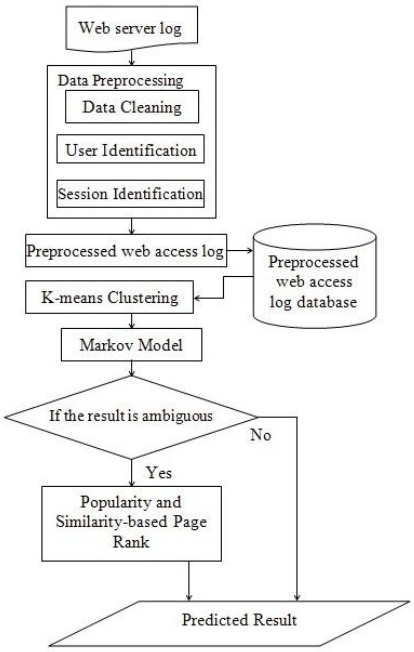

Figure 1. Proposed System Architecture

3.1 Web Server Log

A Web log file [12] records activity information when a Web user submits a request to a Web Server. The main source of raw data is the web access log which we shall refer to as log file. As log files are originally meant for debugging purposes. A log file can be located in three different places: i) Web Servers, ii) Web proxy Servers, and iii) Client browsers. Server-side logs: These logs generally

supply the most complete and accurate usage data, but their two major drawbacks are:

- These logs contain sensitive, personal information, therefore the server owners usually keep them closed.
- The logs do not record cached pages visited. The cached pages are summoned from local storage of browsers or proxy servers, not from web servers.

NASA web server log file is considered for the purpose of analysis. Web Server logs are plain text (ASCII) files, that is independent from the server platform. There are some distinctions between server software, but traditionally there are four types of server logs: Transfer Log, Agent Log, Error Log and Referrer Log. The first two types of log files are standard. The Referrer and Agent Logs may or may not be "turned on" at the server or may be added to the Transfer log file to create an "Extended" Log File format. A Web log is a file to which the Web server writes information each time a user requests a resource from that particular site. Most logs use the format of the common log format.

3.2 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 [19]. The popularity value of page rank was discussed in [17]. 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

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

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, the popularity and similarity-based page rank (PSPR) for every page can be calculated. In order to make rank calculations faster, the required steps of our calculations are stored in the 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.

5. EXPERIMENTAL EVALUATION

This paper introduces a method that integrates k-means clustering, Markov model and popularity and similarity-based 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 system. 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 (D1), the web logs from 16/Aug/1995 are used. For the second testing data set (D2), the web logs from 16/Aug/1995 to 17/Aug/1995 are used. For the third testing data set (D3), the web logs from 16/Aug/1995 to 18/Aug/1995 are used. For the fourth testing data set (D4), the web logs from 16/Aug/1995 to 19/Aug/1995 are used. For the fifth testing data set (D5), the web logs from 16/Aug/1995 to 20/Aug/1995 are used. For the sixth testing data set (D6), the web logs from 16/Aug/1995 to 21/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. Table 1 show the data after processing the preprocessing phase.

Table 1. Testing data set after preprocessing

	D1	D2	D3	D4	D5	D6
Records after preprocessing	7965	17804	27400	33054	39006	50019
Sessions	346	736	1124	1376	1617	2072

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, 17, 18] 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, 17, 18] 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.

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

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. The results of the experiment for the next page prediction accuracy for popularity and similarity-based page ranking algorithm under KSim and OSim similarity are given in Table 2.

As depicted in Table 2, PSPR based on 2nd order Markov model prediction outperforms PSPR based on 1st order Markov model significantly in all OSim and KSim values in the top-3 prediction. 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.

Table 2. Top 3 prediction based on Average OSim and KSim value

	OSim Results based on 1st order Markov Model (%)	KSim Results based on 1st order Markov Model (%)	OSim Results based on 2nd order Markov Model (%)	KSim Results based on 2nd order Markov Model (%)
D1	35.59	53.04	41.46	54.49
D2	38.68	55.92	48.22	59.23
D3	38.08	55.46	49.6	61.1
D4	38.39	55.66	50.22	61.94
D5	39.5	56.6	50.84	62.44
D6	40.53	57.33	51.53	63.2

6. CONCLUSIONS

The method presented in this paper is to improve the Web page access prediction accuracy by integrating all three algorithm K-means Clustering, Markov Model and Popularity and Similarity-based Page Rank algorithm. OSim and KSim algorithm are used to calculate the similarity of our 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. We used the idea of Page Rank algorithm to improve the prediction accuracy and modified this algorithm in order to analyze the user behavior.

7. ACKNOWLEDGMENTS

My Sincere thanks to my supervisor Dr. Ei Ei Chaw, Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City), Myanmar for providing me an opportunity to do my research work. I express my thanks to my Institution namely University of Technology (Yatanarpon Cyber City) for providing me with a good environment and facilities like internet, books, computers and all that as my source to complete this research. My heart-felt thanks to my family, friends and colleagues who have helped me for the completion of this work.

REFERENCES

[1] Vaidehi .V, Monica .S, Mohamed Sheik Safeer .S, Deepika .M, Sangeetha .S, "A Prediction System Based on Fuzzy Logic", Proceedings of the World Congress on Engineering and Computer Science 2008, WCECS 2008, October 22 - 24, 2008, San Francisco, USA.

[2] Siriporn Chimphlee, Naomie Salim, Mohd Salihin Bin Ngadiman, Witcha Chimphlee, Surat Srinoy, "Rough Sets Clustering and Markov model for Web Access Prediction", Proceedings of the Postgraduate Annual Research Seminar 2006, pp. 470-475.

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

[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".

[5] Yong Zhen Guo and Kotagiri Ramamohanarao and Laurence A. F. Park, "Web Page Prediction Based on Conditional Random Fields.

[6] Ke Yiping, "A Survey on Preprocessing Techniques in Web Usage Mining", The Hong Kong University of Science and Technology, Dec 2003.

[7] S. Brin, L. Page, 1998. "The anatomy of a large-scale hypertextual Web search engine", Computer Networks, Vol. 30, No. 1-7, pp. 107-117, Proc. of WWW7 Conference.

[8] 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.

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

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

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

[12] K. R. Suneetha, Dr. R. Krishnamoorthi, "Identifying User Behavior by Analyzing Web Server Access Log File", IJCSNS International Journal of Computer Science and Network Security, Vol. 9 No. 4, April 2009.

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

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

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

[16] R. Khanchana, Dr. M. Punithavalli, "Web Page Prediction for Web Personalization: A Review", Global Journal of Computer Science and Technology, Vol. 11, No. 7, 2011.

[17] 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, pp. 687-690.

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

[19] P. Thwe, "Proposed Approach for Web Page Access Prediction Using Popularity and Similarity based Page Rank Algorithm", International Journal of Science and Technology Research, Vol. 2, No. 3, March 2013.

This paper may be cited as:

Thwe, P. 2014. Web Page Access Prediction based on an Integrated Approach. *International Journal of Computer Science and Business Informatics, Vol. 12, No. 1, pp. 55-64.*