Efficient Web Usage Mining Process for Sequential Patterns

Sang T.T. Nguyen
Decision Systems & e-Service Intelligence (DeSI) Lab
Centre for Quantum Computation & Intelligent Systems (QCIS)
Faculty of Engineering and Information Technology
University of Technology, Sydney
P.O. Box 123, Broadway, NSW 2007, Australia
+61401446501
tsang@it.uts.edu.au

ABSTRACT

The tremendous growth in volume of web usage data results in the boost of web mining research with focus on discovering potentially useful knowledge from web usage data.

This paper presents a new web usage mining process for finding sequential patterns in web usage data which can be used for predicting the possible next move in browsing sessions for web personalization. This process consists of three main stages: preprocessing web access sequences from the web server log, mining preprocessed web log access sequences by a tree-based algorithm, and predicting web access sequences by using a dynamic clustering-based model. It is designed based on the integration of the dynamic clustering-based Markov model with the Pre-Order Linked WAP-Tree Mining (PLWAP) algorithm to enhance mining performance. The proposed mining process is verified by experiments with promising results.

Keywords

Markov Model, Sequential Patterns, Web Access Patterns (WAP), Pre-Order Linked WAP-Tree (PLWAP-tree), Web Usage Mining (WUM)

1. INTRODUCTION

Web logs contain not only simple web usage sessions, but also useful information which can be used to trace web usage patterns in relation to browsing behavior and recommending more relevant web pages to users. By mining the web logs using more advanced data mining techniques, the web usage patterns of users can be discovered. This process is called Web Usage Mining (WUM) which aims to discover potential knowledge hidden in the web browsing behavior of users [1].

Web Usage Mining algorithms can be classified into many categories, such as clustering, classification, association rules, and sequential pattern discovery. There are two major methods of sequential pattern discovery: deterministic techniques (recording the navigational behavior of the user) and stochastic methods (using “the sequence of web pages that have been visited in order to predict subsequent visits”) [1]. This paper focuses on the advanced model-based techniques of the stochastic methods for predicting hidden sequential patterns.

Some applications of WUM are clustering web users based on interesting patterns [2], mining conceptual link hierarchies from web log files for adaptive website navigation [3], building the frequent web access sequences using a tree algorithm [4], predicting web navigations using the Markov model or association rules [5], [6]. Predicting web access patterns using the Markov model is very interesting in web personalization because of its special features, such as modeling a collection of navigation records, modeling user web navigation behavior, or classifying browsing sessions into different categories. According to J. Borges and M. Levene [7], the Markov model is a powerful and probabilistic model to estimate the probability of visiting web pages. Each web page is referred to as a state in the Markov model. The N-order Markov model enables us to predict the next visited page affected by the previous N-1 visited pages. Although the predictive probability of the N-order Markov model is higher than the lower-order model, the number of its states exponentially increases more. Because the model complexity is measured by the number of states, the complexity of a higher-order Markov model excessively increases when using it to model a huge number of web pages. Some good news for this problem is that there are usually a significant number of unessential web pages which are included in the Markov model in WUM. Effective filtering of such unessential web pages in the web usage data can resolve the aforementioned complexity problem. Some clustering methods have been applied to filter web pages, such as EM and k-means algorithms in [6].

Hybrid probabilistic predictive models based on the Markov model, such as the dynamic clustering-based Markov model of J. Borges and M. Levene [7], have shown improved prediction accuracy over the traditional Markov model. However, the complexity problem of the model has not really been resolved as it still mines the web usage data with a large amount of redundant data.

In order to resolve this complexity problem, this paper proposes a new mining algorithm for WUM based on the Markov model from the stochastic approach. The new feature of this algorithm is that it predicts users’ web navigation patterns using the frequent web access sequences extracted from the web logs rather than all web pages. As a result, the complexity of the Markov model can significantly decrease because only frequently accessed web pages are used, resulting in a small number of states. The frequent web access sequences fed into the model can be discovered from web log data by using a tree algorithm. Consequently, we have a set of possible cases of interesting web navigation patterns. The frequent web navigation patterns will then be modeled by a higher-order Markov model to predict the next page accessed by a user.

The rest of this paper is organized as follows. Section 2 reviews some existing research related to sequential pattern discovery. Section 3 explains the new web usage mining algorithm. Section 4 analyzes the performance of the new mining algorithm. Section 5 presents the experimental results of the new algorithm against two sets of web logs taken from two real websites. Section 6 concludes the paper and points out the further work.

2. RELATED WORKS

Sequential pattern mining is considered as an efficient method in web usage mining because of its tolerance of crawling attacks which create fake profiles [8]. In the last decade, many sequence mining algorithms have been proposed in the literature. Ezeife and Lu [4] proposed the PLWAP (tree) algorithm which "builds the frequent header node links of the original WAP-tree in a pre-order fashion and uses the position code of each node to identify the ancestor/descendant relationships between nodes of the tree". This algorithm was successful in extracting the set of all frequent patterns from the web access sequences. Their experiment results show a huge performance boost by predicting frequent navigation patterns of web users.

Regarding the modeling of user web navigation behavior, José Borges [7] proposed a dynamic clustering-based method to increase accuracy of a Markov model in representing a collection of user web navigation sessions. The novelties of this approach are to use the state cloning concept to duplicate states in a way that separates the in-links whose corresponding second-order probabilities diverge, and to use a clustering technique which determines an efficient way to assign in-links with similar second-order probabilities to the same clone. The algorithm only clones inaccurate states whose first and second-order probabilities diverge or the difference between these probabilities is greater than a certain threshold. The results show that the number of additional states induced by the dynamic clustering method can be controlled through a threshold parameter, so the state space does not increase significantly, and the performance of the method has linear time regarding the size of the model.

Hybrid methods might be more robust than standard data mining algorithms such as the Markov models, clustering algorithms, and association rules in current trustworthy recommender systems [9]. Specially, some research has shown that the adaptive mixture of Markov models achieves higher performance for web personalization.

Some approaches to adaptive mixture of Markov models are the combination of the Markov model with clustering techniques or association rules. Liu et al. [5] proposed a clustering method based on a mixture of Markov models to cluster users and capture the sequential relationships hidden in user web navigation histories. The performance of this method is higher than the traditional Markov models, the association rules, or clustering methods. Jalali et al. [10] proposed a clustering approach based on "the graph partitioning for modeling user navigation patterns". Their experimental results could improve the quality of clustering user navigation patterns in WUM systems. Khalil [6] combined three techniques of clustering, association rules and a low-order Markov model. This combination obtains better accuracy of web page access prediction, less state space complexity and fewer generated rules than the single methods.

All of these works attempt to build a more accurate and efficient model for mining the users' web navigation patterns from the web usage data. However, most of them are not concerned with the datasets in the mining process. This may lead to mining unessential or uninteresting data. Considering the methods proposed in [4] and [7], while mined patterns seem to be separated and have not yet presented relationships between access points in the PLWAP-tree [4], the Markov model could fail to mine a huge dataset including necessary and unnecessary information [7].

Compared with the existing work, our work presents an innovative idea of combining the tree algorithm and the Markov model to build a novel mining process. This process can predict users' web navigation patterns more effectively by only using the interesting web access patterns and can resolve the drawbacks of the existing methods, i.e., the complexity problem.

3. NEW MINING PROCESS

The new mining process is designed based on the combination of the PLWAP algorithm [4] and the dynamic clustering-based Markov Model [7] to achieve higher performance in predicting frequent web navigation patterns. The block diagram of this mining process is depicted in Figure 1.

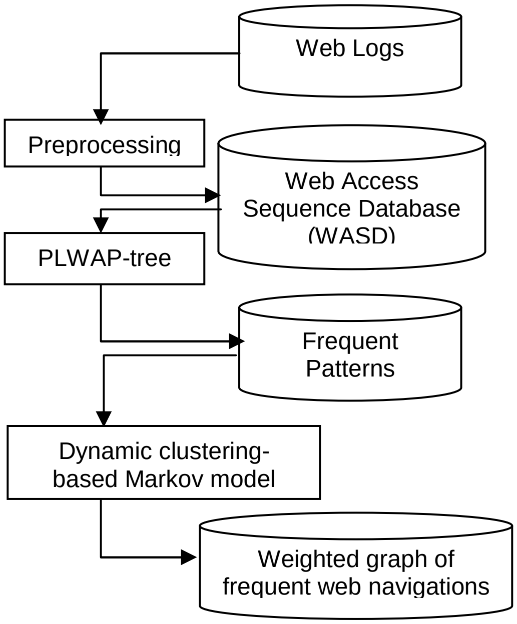
Web Logs
Preprocessing
Web Access Sequence Database (WASD)
PLWAP-tree
Frequent Patterns
Dynamic clustering-based Markov model
Weighted graph of frequent web navigations

Figure 1. WUM process

There are three steps in this process.

Step 1: Preprocess the web log data.

In this step, sequential patterns or web access sequence database (WASD) are obtained after cleansing the web log, i.e., removing erroneous and invalid pages from the web log and eliminating multi-media files such as gif, jpg and script files such as js and cgi.

Step 2: Extract the frequent web page sequences from WASD using the PLWAP-tree algorithm [4].

Given a transaction set T and a set S = {S_1, S_2… S_n} of frequent (contiguous) sequential patterns over T, the support of each S_i is defined as follows [11]:

$$\sigma(S_i)=\frac{|\{t\in T: S_i\ is\ (contiguous)\ subsequence\ of\ t\}|}{|T|}$$

The support is a parameter commonly used to estimate the frequent sequences of web pages navigated by users. It is used in the PLWAP algorithm to select web access patterns that have the support values greater than the predefined minimum support (MinSup). In this step, a complete set of frequent patterns in WASD is discovered.

Step 3: Predict frequent web navigations based on the frequent patterns

The dynamic clustering-based Markov model [7] is used to discover a weighted graph of the frequent web navigations with second-order probability accuracy. Compared with the second-order Markov model, this used model is more efficient and flexible because of its dynamic prediction of the next pages based on the previous pages.

This new mining process allows the prediction of user web navigation based on web access history. It predicts the next web pages that the user likely to visit based on his or her previously visited web pages, and its predictive accuracy is higher than the first-order Markov model which predicts the next visiting web pages only based on the current state. More importantly, the complexity of the Markov model in this new mining process can be significantly decreased because the input of the Markov model is only the useful and essential information extracted from the web log using the PLWAP-tree (all the unessential web pages are filtered out). Consequently, the final mined result of this new mining process allows accurate prediction or recommendation of interesting web navigation patterns. It will be very useful for web personalization using recommender systems.

4. EXPERIMENTAL EVALUATION

In order to validate the proposed mining process, two web logs are used for the experiments. They are taken from the following two websites: http://www.cs.kent.edu/ (KENT) and http://science.ksc.nasa.gov (NASA). The KENT dataset was collected from 18/Sep/2002:12:05:25 through 23/Sep/2002:12:05:11, with a total of 6 days. The NASA dataset was collected from 00:00:00 August 1, 1995 through 10:46:43 August 13, 1995, with a total of 13 days. The web log cleaning tool[1] is used to clean the web log files and extract the data sets in the preprocessing step. In this step, we can optionally remove invalid pages and select correct web pages. The data set is a set of web access sequences which is the input data of the next mining steps. Table 1 shows the data sets after the preprocessing step (Step 1).

Table 1. Datasets after preprocessing

Dataset	Number of users	Size (Mbytes)	Number of sessions
KENT	4472	2.14	8412
NASA	26037	10.20	49406

From the sessions listed in table 1, web pages were extracted and were processed in the next mining steps. Table 2 shows the number of web pages (states) of the Markov model of frequent web navigations after applying the PLWAP algorithm (Step 2).

Table 2. Number of states (web pages)

Dataset	MinSup	Number of web pages (or states) after running PLWAP-tree algorithm	Number of web pages (or states) before running PLWAP-tree algorithm
KENT	0.01	11	7134
	0.02	4	7134
	0.03	0	7134
NASA	0.01	50	1446
	0.02	22	1446
	0.03	12	1446

As we can see from the last two columns in Table 2, if we do not use the PLWAP-tree before the Markov model based modeling, the number of web pages fed into the Markov model will be much larger (7134 for the KENT dataset and 1446 for the NASA dataset). This will make the state space of the Markov model too large to be practical. If we apply the PLWAP-tree before, the number of web pages fed into the model can be significantly decreased. For example, the number of web pages reduced to 11 (with the MinSup = 0.01) for the KENT dataset from the original number of web pages 7134 and 50 (with the MinSup = 0.01) for the NASA dataset from 1446, because the web pages whose the supports were less than MinSup had been removed. Moreover, it is possible to control the size of the frequent web navigations by adjusting the MinSup values.

In step 3, the final result of the dynamic clustering-based Markov model is a graph presenting the weighted relationships of frequent web links. Each web link has the historical links visited before the web link and the recommended links which might be visited after the web link with the corresponding predictive probabilities, as shown in the following two examples.

Example 1: In the case of the KENT dataset choosing MinSup = 0.01, we have a graph linking 11 web pages. Figure 2 depicts a part of the graph containing the node ”*stoc.html*” with its historical links and the predicted links for the next move (the recommended links).

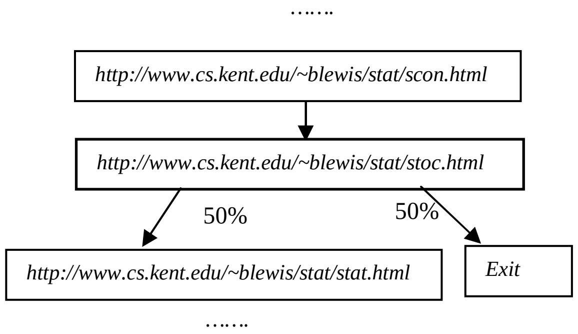
…….
http://www.cs.kent.edu/~blewis/stat/scon.html
http://www.cs.kent.edu/~blewis/stat/stoc.html
50%
50%
http://www.cs.kent.edu/~blewis/stat/stat.html
Exit
…….

Figure 2. Example of a node in the weighted graph of frequent web navigations (KENT dataset)

[1] http://sol.cs.uwindsor.ca/~cezeife/webcleaner.tar.gz

This graph shows, for the web page "*http://www.cs.kent.edu/~blewis/stat/stoc.html*", its historical link is "*http://www.cs.kent.edu/~blewis/stat/scon.html*", which is in the same parent directory "*http://www.cs.kent.edu/~blewis/stat/*". It is predicted that a user has 50% probability to visit the page "*http://www.cs.kent.edu/~blewis/stat/stat.html*", which is in the same parent directory, and 50% probability to visit an irrelevant page or end the session (we do not consider this probability).

Example 2: In the case of the NASA dataset choosing MinSup = 0.01, we have a graph linking 50 web pages. Figure 3 depicts a part of the graph containing the node "*history.html*" with its historical links and the predicted links for the next move (the recommended links).

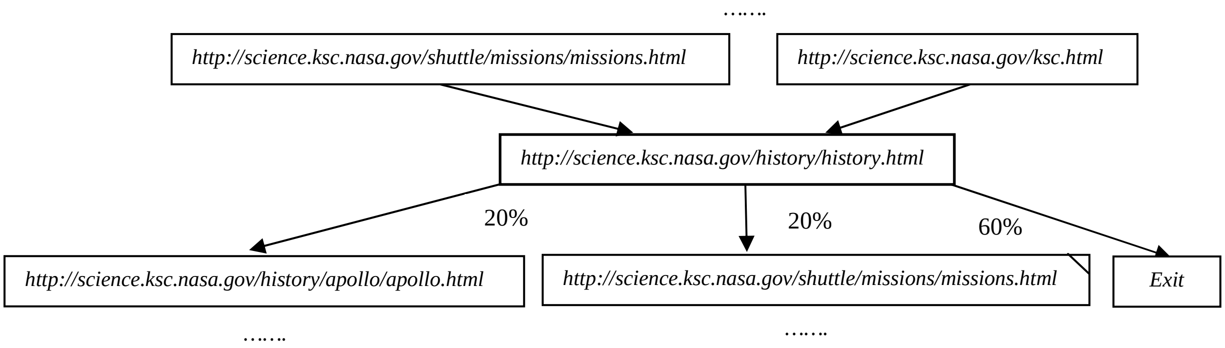
……
http://science.ksc.nasa.gov/shuttle/missions/missions.html
http://science.ksc.nasa.gov/ksc.html
http://science.ksc.nasa.gov/history/history.html
20%
20%
60%
http://science.ksc.nasa.gov/history/apollo/apollo.html
http://science.ksc.nasa.gov/shuttle/missions/missions.html
Exit
……
……

Figure 3. Example of a node in the weighted graph of frequent web navigations (NASA dataset)

This graph shows that for the web page "*http://science.ksc.nasa.gov/history/history.html*", its historical links are

"*http://science.ksc.nasa.gov/shuttle/missions/missions.html*"

and "*http://science.ksc.nasa.gov/ksc.html*",

a 20% probability is predicted that a user may either go to one page "*http://science.ksc.nasa.gov/history/apollo/apollo.html*" in the same parent directory or go back the historical page "*http://science.ksc.nasa.gov/shuttle/missions/missions.html*", and a 60% probability of going to an irrelevant page or end the session (we do not consider this probability).

Note that either the historical links or the recommended links can be in a different parent directory as the current page. This is allowed as long as the web links are listed in the related pages.

As we can see from the examples, the discovered relationships between the links are reasonable with reference to the link structure of the websites, in which the historical links are often in the same directory or the parent directory of the current link and the recommended links are often in the same directory or the subdirectory of the current link. However, web users sometimes may not follow the link structure of the website, but rather visit the interesting pages shown on the web pages. This is the reason why some discovered web navigations are different from the link structure. In such a case, the web designer may need to consider redesigning the website to reconstruct the links in order to make the website more convenient for users.

5. DISCUSSION AND CONCLUSION

In the mining step, some data mining techniques, such as the Markov model and the PLWAP algorithm are used to discover the data sets. However, the traditional Markov model might fail to model a web log containing numerous web pages with higher prediction accuracy because the number of states will increase significantly in the high-order Markov model. Even though the dynamic clustering-based Markov model tries to control the state space complexity to achieve second-order accuracy by integrating the K-means algorithm to clone the necessary number of states [7], it is a waste of time and memory to mine a considerable number of unessential web pages from a web log. It will be more efficient to collect only the essential web pages before applying the Markov model. The PLWAP algorithm is one of the promising candidates for generating the required frequent web access sequences because it not only can accomplish the task, but is also verifiable and more effective compared with Apriori-like algorithms [4]. However, the PLWAP algorithm will stop at the same level as the association rules, and cannot usually meet the demand of a user who needs the most relevant web pages to be recommended for the next navigation step in a recommender system. The probabilistic Markov model, on the other hand, can take the frequent patterns from the PLWAP algorithm and further predict the most interesting and relevant web pages to navigate for a user.

The new web usage mining process proposed in this paper is the combination of the PLWAP algorithm and the dynamic clustering-based Markov model. It inherits the advantages of the PLWAP-tree and the dynamic clustering-based Markov model and overcomes their drawbacks by omitting uninteresting web pages. It can predict the most interesting web access patterns from the users' navigation history. It has been tested using the two web log datasets taken from real websites. The testing results show that the new mining process can considerably improve the drawbacks of Markov model and enhances the performance of PLWAP algorithm.

Another important contribution of this paper is that the resultant web page link graph not only presents all possible links of the web sites, but also predicts the frequently visited links or the frequent web navigations. This makes the new mining process be very useful for web personalization systems using recommender systems. In contrast, a web usage mining process only using the Markov model does not highlight frequent web navigation or user interests.

Our future work will apply the new process to mine the web usage data from an e-government service website to discover the user navigation patterns in order to predict users' interests and preferences in that specific service domain.

6. ACKNOWLEDGMENTS

The first author would like to thank Dr. Cuong Duc Nguyen (The University of Technology, HCMC, Vietnam), who guided her to conduct this research. She also would like to thank Prof. Jie Lu and Dr. Helen Lu (The University of Technology, Sydney, NSW, Australia) for their advices on writing this paper.

7. REFERENCES

[1] Pierrakakos, D., Paliouras, G., Papatheodorou, C., and Spyropoulos, C.D. 2003. Web Usage Mining as a Tool for Personalization: A Survey. User Modelling and User-Adapted Interaction. 13, 4, 311-372. DOI=http://dx.doi.org/10.1023/A:1026238916441.

[2] Chen, L., Bhowmick, S.S., and Li, J. 2006. COWES: Clustering Web Users Based on Historical Web Sessions. In Database Systems for Advanced Applications, Springer Berlin / Heidelberg, 541-556. DOI=10.1007/11733836

[3] Zhu, J., Hong, J., and Hughes, J.G. 2004. PageCluster: Mining Conceptual Link Hierarchies from Web Log Files for Adaptive Web Site Navigation. ACM Transactions on Internet Technology. 4, 185-208. DOI=http://doi.acm.org/10.1145/990301.990305.

[4] Ezeife, C.I., and Lu, Y. 2005. Mining Web Log Sequential Patterns with Position Coded Pre-Order Linked WAP-Tree. Data Mining and Knowledge Discovery. 10, 1, 5-38. DOI=10.1007/s10618-005-0248-3.

[5] Liu, Y., Huang, X., and An, A. 2007. Personalized Recommendation with Adaptive Mixture of Markov Models. The American Society for Information Science and Technology. 58, 12, 1851–1870. DOI=10.1002/asi.20631.

[6] Khalil, F. 2008 Combining Web Data Mining Techniques for Web Page Access Prediction. Doctoral thesis. University of Southern Queensland.

[7] Borges, J., and Levene, M. 2004 A Dynamic Clustering-Based Markov Model for Web Usage Mining. Technical Report. Available online at http://xxx.arxiv.org/abs/cs.IR/0406032.

[8] Bhaumik, R., Burke, R., and Mobasher, B. 2007. Effectiveness of Crawling Attacks Against Web-based Recommender Systems. In: Proceedings of the 5th workshop on intelligent techniques for web personalization (ITWP-07)

[9] Mobasher, B., Burke, R., Bhaumik, R., and Williams, C. 2007. Toward Trustworthy Recommender Systems: An Analysis of Attack Models and Algorithm Robustness. ACM Transactions on Internet Technology. 7, 4. DOI=10.1145/1278366.1278372.

[10] Jalali, M., Mustapha, N., Mamat, A., and Sulaiman, M.N.B. 2008. A New Clustering Approach based on Graph Partitioning for Navigation Patterns Mining. Proc. ICPR 2008. IEEE. pp. 1-4.

[11] Mobasher, B. 2007. Data Mining for Web Personalization. In The Adaptive Web, P. Brusilovsky, A.K., and W. Nejdl, Springer Berlin / Heidelberg, 90-135. DOI=10.1007/978-3-540-72079-9_3