Web Log Mining with Adaptive Support Thresholds

Jian-Chih Ou
Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan, ROC
alex@arbor.ee.ntu.edu.tw

Chang-Hung Lee
BenQ Corporation 18, Jihu Road, Neihu
Taipei, Taiwan, ROC
MichaelCHLee@BenQ.com

Ming-Syan Chen
Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan, ROC
mschen@cc.ee.ntu.edu.tw

ABSTRACT

With the fast increase in Web activities, Web data mining has recently become an important research topic. However, most previous studies of mining path traversal patterns are based on the model of a uniform support threshold without taking into consideration such important factors as the length of a pattern, the positions of Web pages, and the importance of a particular pattern, etc. In view of this, we study and apply the Markov chain model to provide the determination of support threshold of Web documents. Furthermore, by properly employing some techniques devised for joining reference sequences, a new mining procedure of Web traversal patterns is proposed in this paper.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Databases Applications—*Data mining.*

General Terms

Algorithms

Keywords

Web mining, path traversal pattern, Markov model

1. INTRODUCTION

With the rapid expansion of WWW, Web data mining has recently become an important research topic and is receiving an increasing amount of research interest from both academic and industrial environments. Among others, an important class of web data mining problem is mining of path traversal patterns, that can be used to decide the next likely web page requests based on significant statistical correlations. If such a sequence appears frequently enough, then this sequence indicates a frequent traversal pattern. However, most previous studies of path traversal pattern mining are based on the model of a uniform support threshold without taking into consideration such important factors as the length of the pattern, the positions of Web pages, etc. For instance, consider a Web site shown in Figure 1 and a database that contains four sequences: ABD, AB, AC, $ABFP$. Suppose the minimum support is 2. Web pages A and B at higher levels of Web site are deemed frequent in this case, which might, however, be due more to their locations than to their contents. As a result, a low support threshold will lead to lots of uninteresting patterns derived while a high support threshold may cause some interesting patterns with lower supports to be ignored. Hence, different support thresholds are deemed necessary for Web pages at different levels of Web sites.

WWW 2005, May 10–14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

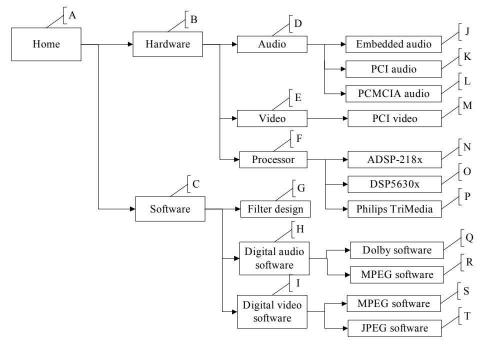

Figure 1: A site map of a real Web site

2. DETERMINATION OF ADAPTIVE THRESHOLD

Definition 1: The adaptive support threshold of the Web page p is defined as $apt\_sup(p)$. The adaptive support of a reference sequence c, denoted by $AptSup(c)$, is the lowest apt_sup value among the pages in the reference sequence c, i.e., $AptSup(c) = \min_{p\in c}\{apt\_sup(p)\}$.

The definition of adaptive threshold has been given above. However, without specific knowledge, it is very difficult and intricate for users to set adaptive threshold for every single Web page . If the support threshold is set too high, users cannot obtain enough rules. Therefore, users have to set a lower threshold and conduct the mining again, which may or may not lead to results of better quality. If the threshold is too small, there may be an excessive number of rules for the users and the runtime may be unacceptably long. To overcome this problem, we need an automatic and reasonable methodology to provide the determination of support threshold of Web documents. Therefore, we introduce the general probabilistic framework based on the Markov chains which can be used to determine the adaptive threshold of each Web page in this section.

A Markov chain is a discrete-time stochastic process defined over a set of states S in terms of a matrix P of transition probabilities. The entry P_{ij} in the transition probability matrix P is the probability that the next state will be j, given that the current state is i. Thus, for all $i, j \in S$, we have $0 \leq P_{ij} \leq 1$, and $\sum_j P_{ij} = 1$.

The surfing on a Web can be viewed as a Markov chain whose states correspond to the pages and whose transition probability matrix entry P_{ij} is defined by the probability of following a hyperlink from page i to page j. We will denote by X_t the page surfing at time t. And for all pages i, j contained in the Web, define the t-step transition probability as $P_{ij} = Prob[X_t = j \mid X_0 = i]$.

The mean recurrence time of page i is denoted as μ_i. Moreover, $1/\mu_i$ represents the proportion of times, in the long run, the surfer will be in page i. The following lemma will give a straight way of finding $1/\mu_i$.

Lemma 1: For any two pages i, j contained in the Web, the transition probability matrix entry P_{ij} is defined as follows:

$$P_{ij} = \begin{cases} \dfrac{1}{od(i)} & \text{if there exists a hyperlink from } i \text{ to } j \\ 0 & \text{otherwise} \end{cases}$$

where $od(i)$ is the out degree of page i, i.e., the number of hyperlink that is incident from page i.

If all the pages in the Web satisfy the following properties:

1. All pages in the Web are *irreducible*, i.e.,there is a directed path from every page to every other page.
2. All pages in the Web are *aperiodic*, i.e., for all i, j, there are paths of all possible length, except for a finite set of path lengths that may be missing.

Then, for all pages i, the sequence $(1/\mu_i)$ will converge to the principal eigenvector of P^T, that is the transpose of the transition probability matrix P. In the interest of space, the proof of **Lemma 1** is omitted here and can be deduced from [3].

However, the above model is too ideal to match the real situation of a Web, because there are many Web pages without any outlinks. Further, directed paths incurred in real Web pages may lead into a cycle. To remedy this, a low-probability transition e [1] can be involved in the above model.

Consequently, we have the adaptive support threshold of the newly identified dynamic mining model as the following formulas:

$$\begin{aligned} apt\_sup(p) &= m * |\mathcal{D}|/u_p \text{ if } m * |\mathcal{D}|/u_p \geq S_{th} \quad (1) \\ apt\_sup(p) &= S_{th} \qquad\qquad \text{if } m * |\mathcal{D}|/u_p \leq S_{th}, \end{aligned}$$

where $|\mathcal{D}|$ is the size of the database, u_p is the mean recurrence time of page p, and $m \in [0, 1]$ is a parameter to determine the relationship between the interestingness supports of Web pages and their the mean recurrence time. S_{th} is employed for pruning some obsolete rules whose Web pages have very low expected occurrence frequencies.

3. CONSTRUCTION OF MINING PROCEDURE

The new mining procedure generalizes the FS algorithm for finding frequent reference sequences proposed in [2]. Similar to algorithm FS, frequent Web path traversal patterns are generated by using multiple passes over the database and the large reference sequences L_k found in the $(k-1)^{th}$ pass are used to generate the candidate reference sequences C_k. The candidate generation procedure of C_2 just differs from function Level2-candidate-gen()[4] in the last step, in view of the differences between sets and sequences. It is noted that the new candidate generation procedure of $C_k (k > 2)$ adopts a novel join process which is different from that proposed in [2]. In [2], two distinct sequences from L_{k-1}, say $r_1, ..., r_{k-1}$ and $s_1, ..., s_{k-1}$, are joined to form a k-reference sequence if *either* $r_1, ..., r_{k-1}$ contains $s_1, ..., s_{k-2}$ *or* $s_1, ..., s_{k-1}$ contains $r_1, ..., r_{k-2}$. To address this issue, we devise in this paper three joinable forms, namely *head_join*, *mid_join* and *tail_join* forms.

Definition 2: The minimal support page of a reference sequence r is $MSP(r) = \{p | p \in r,\ apt\_sup(p) = AptSup(r)\}$.

Definition 3: Suppose r is a k-reference sequence which contains $r_1, ..., r_k$.

(i) If $r_1 \notin MSP(r)$ and $r_k \notin MSP(r)$, then r is one reference sequence of $mid\_join\ form$, which can be generated from $r_1, ..., r_{k-1}$ and $r_2, ..., r_k$.

(ii) If $r_1 \in MSP(r)$, then r is one reference sequence of $head\_join\ form$, which can be generated from $r_1, ..., r_{k-2}, r_{k-1}$ and $r_1, ..., r_{k-2}, r_k$.

(iii) If $r_k \in MSP(r)$, then r is one reference sequence of $tail\_join\ form$, which can be generated from $r_1, r_3..., r_k$ and $r_2, r_3..., r_{k-2}, r_k$.

First, by using the union of above mentioned three join forms, we join $L_{k-1}$with L_{k-1} to obtain C_k in the join step. Then, in the prune step, we delete all sets of references $c \in C_k$ which are infrequent. Finally, it returns a superset of the set of all frequent k-references.

4. CONCLUSION

This paper broadened the horizon of frequent path traversal pattern mining by introducing a flexible model of mining Web traversal patterns with adaptive thresholds. Specifically, we apply the Markov chain model to provide the determination of support threshold of Web documents. By properly employing some techniques devised for joining reference sequences, a new mining procedure of Web traversal patterns was also proposed in this paper.

5. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[2] M.-S. Chen, J.-S. Park, and P. S. Yu. Efficient Data Mining for Path Traversal Patterns. *IEEE Transactions on Knowledge and Data Engineering*, 10(2):209–221, April 1998.

[3] G. R. Grimmett and D. R. Stirzaker. *Probability and Random processes*. Oxford Science Publications, 2nd edition, 1992.

[4] B. Liu, W. Hsu, and Y. Ma. Mining Association Rules with Multiple Minimum Supports. *Proc. of 1999 Int. Conf. on Knowledge Discovery and Data Mining*, August 1999.