Mining History of Changes to Web Access Patterns
(Extended Abstract)

Qiankun Zhao and Sourav S. Bhowmick

Nanyang Technological University Singapore, 639798
{pg04327224,assourav}@ntu.edu.sg

1 Introduction

Recently, a lot of work has been done in web usage mining [2]. Among them, mining of frequent Web Access Pattern (WAP) is the most well researched issue[1]. The idea is to transform web logs into sequences of events with user identifications and timestamps, and then extract association and sequential patterns from the events data with certain *metrics*. The frequent WAPs have been applied to a wide range of applications such as personalization, system improvement, site modification, business intelligence, and usage characterization [2]. However, most of the existing techniques focus only on mining frequent WAP from snapshot web usage data, while web usage data is dynamic in real life. While the frequent WAPs are useful in many applications, knowledge hidden behind the historical changes of web usage data, which reflects how WAPs change, is also critical to many applications such as adaptive web, web site maintenance, business intelligence, etc.

In this paper, we propose a novel approach to discover hidden knowledge from historical changes to WAPs. Rather than focusing on the occurrence of the WAPs, we focus on the frequently changing web access patterns. We define a novel type of knowledge, *Frequent Mutating WAP (FM-WAP)*, based on the historical changes of WAPs. The FM-WAP mining process consists of three phases. Firstly, web usage data is represented as a set of WAP trees and partitioned into a sequence of *WAP groups* (subsets of the WAP trees) according to a user-defined *calendar pattern*, where each WAP group is represented as a *WAP forest*. Consequently, the log data is represented by a sequence of WAP forests called WAP history. Then, changes among the WAP history are detected and stored in the *global forest*. Finally, the FM-WAP is extracted by a traversal of the *global forest*. Extensive experiments show that our proposed approach can produce novel knowledge of web access patterns efficiently with good scalability.

2 FM-WAP Mining

Given a WAP forests sequence, to measure the significance and frequency of the changes to the support of the WAPs, two metrics, *S-value* and *F-value*, are proposed. Formally, they are defined as follows.

J.-F. Boulicaut et al. (Eds.): PKDD 2004, LNAI 3202, pp. 521–523, 2004.

Definition (S-Value) Let F_{wi} and $F_{w(i+1)}$ be two WAP forests in the WAP history. Let T_{wk} be a subtree of F_{wi}, $T_{w(k+1)}$ be the new version of T_{wk} in $F_{w(i+1)}$. Then the *S-value* of T_{wi} is defined as

$$S_i(T_{wi}) = \frac{|support(T_{w(k+1)}) - support(T_{wk})|}{max(support(T_{wk}),\ support(T_{w(k+1)}))}$$

Given a threshold α for *S-value*, a WAP tree T_{wi} **changed significantly** from F_{wi} to $F_{w(i+1)}$ if $S_i \geq \alpha$.

Here $support(T_{w(k+1)})$ denotes the support values of $T_{w(k+1)}$ in WAP forest $F_{w(i+1)}$. *S-value* is defined to represent the significance of changes to the support values of a WAP in two consecutive WAP forests. Given a WAP forest sequence, there will be a sequence of *S-value*s for a specific WAP tree. It can be observed that *S-value* is between 0 and 1. A larger *S-value* implies a more significant change.

Definition (F-Value) Let H be a WAP history and $\langle F_{w1}, F_{w2}, \cdots, F_{wn} \rangle$. Let $\langle S_1(T_{wk}), S_2(T_{wk}), \cdots, S_{n-1}(T_{wk}) \rangle$ be the sequence of S-values of $T_{wk} \in F_{wj}$, for $1 \leq j \leq n$. Let α be the threshold for the S-value. Then the F-value for T_{wk} is defined as: $\mathrm{F}(T_{wk}, \alpha) = \frac{\sum_{i=1}^{n} f_i}{n}$, where $f_i = 1$, if $S_i(T_{wk}) \geq \alpha$; else $f_i = 0$.

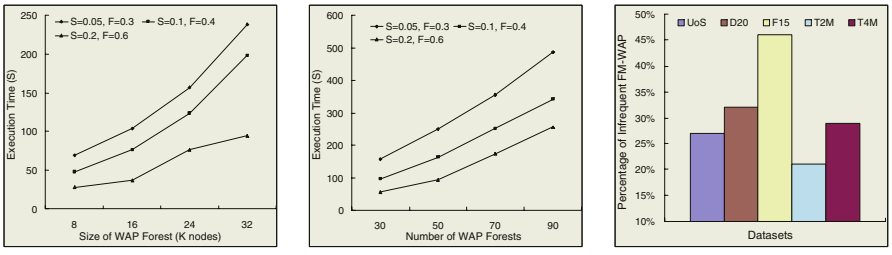
The *F-value* for a tree is defined to represent the percentage of times this tree changed significantly against the total number of WAP forests in the history. The *F-value* is based on the threshold α for *S-value*. It can also be observed that *F-value* is between 0 and 1. Based on the *S-value* and *F-value* metrics, *frequently mutating web access pattern*(FM-WAP) is defined as following.

Definition (FM-WAP) Let H be a WAP history and $\langle F_{w1}, F_{w2}, \cdots, F_{wn} \rangle$. Let α and β be the thresholds for the *S-value* and *F-value* respectively. Then a WAP tree $T_{wj} \in F_{wm}$, where $1 \leq m \leq n$, is a **FM-WAP** if $F(T_{wj}, \alpha) \geq \beta$.

FM-WAPs represent the access patterns that change significantly and frequently (specified by α and β). FM-WAPs can be frequent or infrequent according to existing WAP mining definitions [1]. The difference between frequent WAPs in [1] and the FM-WAPs is that frequent WAPs are defined based on the occurrence of WAPs in the web log data; while FM-WAPs are defined based on the change patterns of web log data. The problem of FM-WAP mining is defined as follows. Given a collection of web usage data, FM-WAP mining is to extract all the FM-WAPs with the user-defined calendar pattern and thresholds of *S-value* and *F-value*.

3 Experimental Results

In the experiments, four synthetic datasets generated by using a tree generator program and a real dataset downloaded from *http://ita.ee.lbl.gov /html/contrib/ Sask-HTTP.html* are used. Figures 1 (a) and (b) show the scalability of our algorithm by varying the size and number of WAP forests. Figure 1 (c) shows that FM-WAP is novel knowledge by comparing with the frequent WAP mining results using WAP-mine[1]. It has been observed and verified that not all FM-WAPs are frequent WAPs and not all frequent WAPs are FM-WAPs.



(a) Size of WAP forests (b) No. of WAP forests (c) Infrequent FM-WAP

Fig. 1. Experiment Results

4 Conclusion

In this paper, we present a novel approach to extract hidden knowledge from the history of changes to WAPs. Using our proposed data structure, our algorithm can discover the FM-WAPs efficiently with good scalability as verified by the experimental results.

References

1. J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu. Mining access patterns efficiently from web logs. In *PAKDD*, pages 396–407, 2000.
2. J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations*, 1(2):12–23, 2000.