Traversal Pattern Mining in Web Usage Data

Jenq-Foung (J.F.) Yao
Georgia College & State University, USA

Yongqiao Xiao
SAS Co., USA

INTRODUCTION

Web usage mining is designed to discover useful patterns in Web usage data, i.e., Web logs. Web logs record the user's browsing of a Web site, and the patterns provide useful information about the user's browsing behavior. Such patterns can be used for Web design, improving Web server performance, personalization, etc.

BACKGROUND

Pattern discovery applies methods and algorithms from different fields such as statistics, data mining, machine learning, etc., to the prepared data. By applying the statistical techniques to the Web usage data, some useful statistics about the users' browsing behavior can be obtained, e.g., the average view time of a page, and the most frequently accessed pages, etc. By applying clustering (a.k.a., unsupervised learning) to the usage data, groups of users which exhibit similar browsing behavior can be found. Such knowledge is useful for market segmentation and personalization. The focus here is on discovering traversal patterns from the Web usage data.

A traversal pattern is a list of pages visited by a user in one session. Several different traversal patterns and the corresponding methods of discovering them have been proposed in the literature, namely, Association Rules, Sequential Patterns, Frequent Episodes, Maximal Frequent Forward Sequences, and Maximal Frequent Sequences. The details about each type of traversal pattern are described in the following section.

TRAVERSAL PATTERNS AND DISCOVERING METHODS

Association Rules

Association rules were originally proposed for market basket data (Agrawal et al., 1993; Agrawal and Srikant, 1994). Association rules describe the associations among items bought by customers in the same transaction, e.g., 80% of customers who bought diapers also bought beer in some store.

To mine association rules from the transactions, there are two steps: first finding the frequent item sets, and then generating association rules from the frequent item sets. Since the second step is straightforward compared to the first one, the research focus is on the first step. An item set (or itemset for short) is frequent, if the support for the itemset is not less than some predefined threshold. The support for an itemset in a database of transactions is defined as the percentage of the transactions that contain the itemset.

Sequential Patterns

Sequential patterns (Srikant and Agrawal, 1997) were also originally proposed for market basket data. For example, customers buy a digital camera, then a photo printer, and then photo papers. Such sequential patterns capture the purchasing behavior of customers over time.

Sequential patterns have also been applied to Web logs (Buchner et al., 1999; see also Spiliopoulou, 2000; Pei et al., 2000). The sessions are ordered by the user id and the access time. As for association rules, the duplicate pages are discarded. Then for each user, there is a user sequence, which consists of all sessions of the user. A sequential pattern is a maximal sequence of itemsets whose support is not less than some predefined threshold. A sequence is maximal if it is not contained in any other sequence. The support of a sequence is the percentage of user sequences that contain the sequence.

Algorithm AprioriAll was proposed in Srikant and Agrawal (1995) for finding all sequential patterns given some support threshold. AprioriAll was then improved by Generalized Sequential Patterns (GSP) (Srikant and Agrawal, 1996). Traversal patterns were generalized to allow time constraints, sliding time window, and user-defined taxonomy.

Frequent Episodes

Frequent episodes were originally proposed for telecommunication alarm analysis (Mannila et al., 1997). Episodes are collections of events, which occur together within some time window. In general, they are partially ordered sets of events. There are two special types of episodes: parallel episodes and serial episodes. They differ in whether the events in the episodes are ordered. In parallel episodes the events are not ordered, while in serial episodes the events are ordered sequential. An episode is frequent if it occurs in the event sequence not less than some predefined threshold.

Frequent episodes were also applied to Web logs (Mannila et al., 1997). The clicks (pages) correspond to events. They are ordered by the access time, and usually the users need not be identified, i.e., there are no sessions.

Maximal Frequent Forward Sequences

Maximal Frequent Forward Sequences (MFFS for short) were proposed in Chen et al. (1998) Notice that MFFS was referred to as large reference sequence in Chen et al. (1998). An MFFS describes the path traversal behavior of the user in a distributed information-providing environment like World Wide Web. There are two steps to mine MFFSs from the sessions. First each session is transformed into maximal forward sequences (i.e., the backward traversals are removed). The MFFSs are then mined using level-wise algorithms (Park et al., 1995) from the maximal forward sequences.

In the raw sessions, there are often backward traversals made by the user. A backward traversal means revisiting a previously visited page in the same user session. It is assumed that such backward traversals happen only because of the structure of the Web pages, not because the user wants to do this. When a backward traversal occurs, a forward traversal path terminates. This resulting forward traversal path is called maximal forward sequence. It then backtracks to the starting point of the next forward traversal and resumes another forward traversal path.

An MFFS is a traversal sequence (consecutive subsequence of a maximal forward sequence) that appears not less than some predefined threshold in the set of maximal forward sequences. The pages in an MFFS are required to be consecutive in the maximal forward sequences, and an MFFS is also maximal, which means that it is not a subsequence of any other frequent traversal sequence.

Maximal Frequent Sequences

Maximal Frequent Sequences (MFS) were proposed in Xiao and Dunham (2001). In contrast to maximal frequent forward sequences, MFSs do not remove backward traversals from the sessions. It was argued in Xiao and Dunham (2001) that such backward traversals are useful for discovering the structures of the Web pages. For example, if a pattern <A, B, A, C> is found frequent, it may suggest that a direct link from page B to page C is needed, while the resulting maximal forward sequences <A, B> and <A, C> lose such information.

An MFS is a traversal sequence (consecutive subsequence of a session) that appears not less than some predefined threshold. Since the backward traversals are kept in the sessions, a traversal sequence may occur in a session more than once. In order to measure the actual number of occurrences of a traversal sequence, the support of an MFS is defined as the ratio of the actual number of occurrences to the total length of all sessions. The length of a session is the number of clicks in the session. The pages in an MFS are required to be consecutive in the sessions, and an MFS is also maximal.

Summary

Table 1 compares the different types of traversal patterns by the following features:

- Ordering: the pages in a traversal pattern can be ordered or not.
- Duplicates: which indicate whether backward traversals are allowed in the traversal pattern.
- Contiguity: the page references in a traversal pattern may be contiguous or not.
- Maximality: a frequent pattern is maximal if it is not contained in any other frequent pattern. A pattern could be maximal or not.

Notice that for frequent episodes, parallel episodes are not ordered, while serial episodes are ordered and the general episodes are partially ordered. Due to the different features of the traversal patterns, the support for each type of pattern is defined quite differently, which is also shown in Table 1.

These features used by different patterns can be used for different purposes. Backward traversals capture the structure information of the Web, and therefore can be used to improve the design of Web pages by adding new links to shorten future traversals. The maximality feature can reduce the number of meaningful patterns discovered. The contiguity and the ordering features could be used to predict future references and thus for prefetching and caching purposes.

These traversal patterns uncover the associations or sequences among the Web pages browsed by the user. They can be used together with other data mining techniques, such as classification and clustering, to further