Data Mining and Knowledge Discovery, 10, 5–38, 2005

Mining Web Log Sequential Patterns with Position Coded Pre-Order Linked WAP-Tree*

C.I. EZEIFE cezeife@uwindsor.ca;(http://www.cs.uwindsor.ca/∼cezeife)
YI LU
School of Computer Science, University of Windsor, Windsor, Ontario, Canada, N9B 3P4

Editor: Fayyad

Received July 2002; Revised June 2004

Abstract. Sequential mining is the process of applying data mining techniques to a sequential database for the purposes of discovering the correlation relationships that exist among an ordered list of events. An important application of sequential mining techniques is web usage mining, for mining web log accesses, where the sequences of web page accesses made by different web users over a period of time, through a server, are recorded. Web access pattern tree (WAP-tree) mining is a sequential pattern mining technique for web log access sequences, which first stores the original web access sequence database on a prefix tree, similar to the frequent pattern tree (FP-tree) for storing non-sequential data. WAP-tree algorithm then, mines the frequent sequences from the WAP-tree by recursively re-constructing intermediate trees, starting with suffix sequences and ending with prefix sequences.

This paper proposes a more efficient approach for using the WAP-tree to mine frequent sequences, which totally eliminates the need to engage in numerous re-construction of intermediate WAP-trees during mining. The proposed algorithm 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. It then, finds each frequent sequential pattern, through progressive prefix sequence search, starting with its first prefix subsequence event. Experiments show huge performance gain over the WAP-tree technique.

Keywords: sequential patterns, Web usage mining, WAP-tree mining, pre-order linkage, position codes, apriori techniques

1. Introduction

Association rule mining is a data mining technique which discovers strong associations or correlation relationships among data. Given a set of transactions (similar to database records in this context), where each transaction consists of items (or attributes), an association rule is an implication of the form $X \rightarrow Y$, where X and Y are sets of items and $X \cap Y = \emptyset$. The support of this rule is defined as the percentage of transactions that contain the set $X \cup Y$, while its confidence is the percentage of these "X" transactions that also contain items in "Y". In association rule mining, all items with support higher than a specified

*This research was supported by the Natural Science and Engineering Research Council (NSERC) of Canada under an operating grant (OGP-0194134) and a University of Windsor grant.

Table 1. The example database transaction table with items.

TID	Items bought
100	f, a, c, d, g, i, m, p
200	a, b, c, f, l, m, o
300	b, f, h, j, o
400	b, c, k, s, p
500	a, f, c, e, l, p, m, n

minimum support are called large or frequent itemsets. An itemset X is called an i-itemset if it contains i items. Agrawal and Srikant (1994) presents the concept of association rule mining and an example of a simple rule is "80% of customers who purchase milk and bread also buy eggs." Table 1 shows an example database transaction table, which can be mined. This table has the set of items $I = \{a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p\}$, where these items could stand for retail store items like bread, butter, cheese, egg, milk, sugar or web pages if the domain of interest is web mining. Using this database, which has only five transactions for simplicity, the support of the itemset, "a, f, c" is 3 or 3/5 (60%) since the itemset "a, f, c" is present in only three of the five transactions (100, 200 and 500). Thus, if the minimum support given by the user for deciding frequent itemsets is 60% or lower, then, the itemset "a, f, c" is a frequent itemset. However, if the minimum support is 70%, then, itemset "a, f, c" is not a frequent itemset since it has a support that is lower than the minimum support threshold. The confidence of a rule, $a, f \rightarrow c$, that can be generated from the frequent itemset, "a, f, c" is 100% because all three transactions that contain the antecedent items, "a, f" also contain the consequent item "c".

Since discovering all such rules may help market baskets or cross-sales analysis, decision making, and business management, algorithms presented in this research area include (Agrawal and Srikant, 1994; Park et al., 1997; Mannila et al., 1995; Han and Kamber, 2000). These algorithms mainly focus on how to efficiently generate frequent patterns from a non-sequential (non-ordered) lists of items, and how to discover the most interesting rules from the generated frequent patterns.

While traditional association rule mining finds intra-transaction patterns, sequential pattern mining finds inter-transaction patterns, to detect the presence of a set of items in a time-ordered sequence of transactions. In basic association rule mining, the items occurring in one transaction have no order, but in sequential pattern mining, an order exists between the items (events) and an item may re-occur in the same sequence. Example of a sequential pattern is: in a video rental store, 80% of customers typically rent "star wars", then "Empire strikes back", and then "Return of Jedi". The measures of support and confidence, used in association rule mining for deciding frequent itemsets, are still used in sequential pattern mining to determine frequent sequences and strong rules that can be generated from them.

Web usage mining is used for automatic discovery of user access patterns from web servers. Interesting user access patterns can be extracted from web access logs, recorded