Mining Frequent Patterns from Web Log Sequences Using WAP Tree Mine Algorithm

Reshma R. Gulwani
Sudhirkumar D. Sawarkar
Department of Computer Engineering,
Mumbai University, Mumbai-Maharashtra, India
E-mails: reshmagulwani@gmail.com, sudhir_sawarkar@yahoo.com

ABSTRACT

Web usage mining is the type of web mining activity that involves discovery of user's access behaviour for the purpose of predicting and hence pre-fetching the web pages that the user is likely to visit. Sequential pattern mining discovers frequent user access patterns from web logs and is the process of applying data mining techniques to a Sequential database. The task of discovering frequent sequences is challenging, because the algorithm needs to process a combinatorially explosive number of possible sequences. As Apriori-like sequential pattern mining techniques requires expensive multiple scans of database and Web Access Pattern Tree (or WAP-tree) can also be applied efficiently to find all access patterns that satisfy user specified criteria. A WAP-tree is a tree to represent the sequence database where all the nodes of the same label are linked when the tree is built. In PLWAP, it is used to link all the nodes of the same label in the order determined by a pre-order traversal of the tree, after it is built. The central theme of these two algorithms is as follows: Scan the Web Access Database twice. However, PLWAP-tree algorithm can't denote ancestor-descendent relationship of nodes in PLWAP-tree clearly. This paper proposes a more efficient approach focusing the BFWAP-tree to mine frequent sequences, which reflects ancestor-descendant relationship of nodes in BFWAP tree directly and efficiently. The proposed algorithm builds the frequent header node links of the original WAP-tree in a Breadth-First fashion and uses the layer code of each node to identify the ancestor-descendant relationships between nodes of the tree. It then, finds each frequent sequential pattern, through progressive Breadth-First sequence search, starting with its first Breadth-First subsequence event.

Keywords: Web Usage Mining, WAP, Tree Mining, Plwaptree Mining, BFWAP-Tree Mining, Layer Codes.