ABSTRACT

ABSTRACT

Automatic discovery and identification of frequently occurring patterns from very large database is the most desired technique in businesses and industries. Data mining and knowledge discovery provides several powerful algorithms for this purpose. Frequent pattern mining is a technique that is used to discover patterns from large transactional database.

The frequent pattern algorithms perform the mining process in two phases. In the first phase, all frequent itemsets that satisfy the user specified minimum support are generated and in the second phase uses these frequent itemsets in order to discover all the association rules that meet a confidence threshold. This research analyzes algorithms that produce compact databases for knowledge discovery from large transaction databases like market basket database and web log databases. From these compact representations, association rule mining is applied to mine frequent patterns. In this research, two variants of Apriori and FP-Growth algorithms, namely, CT-Apriori and CT-PRO are compared and their performances are analyzed.

The CT-Apriori algorithm uses a compact tree structure, called CT-tree, to compress the original transactional data. The tree representation allows the CT-Apriori algorithm, which is revised from the Apriori algorithm, to generate frequent patterns quickly by skipping the initial database scan and reducing a great amount of I/O time per database scan. The CT-PRO algorithm uses a compact tree structure called CFP-Tree, which is more compact than the FP-Tree of the FP-Growth algorithm. An algorithm called CT-PRO is used to mine frequent patterns from CFP-Tree. The CT-PRO algorithm divides the CFP-Tree into several projections represented by CFP-Trees. Then CT-PRO conquers the CFP-Tree for mining all frequent patterns in each projection.

The compression efficiency in terms of storage size required showed a gain of 7.65% over CT-Apriori. Similarly, while considering the number of transactions, CT-PRO was more efficient (7.18%) than CT-Apriori. The

execution speed results also indicated that the CT-PRO algorithm was the fastest among all the algorithms. The efficiency achieved in terms of execution speed on average by CT-PRO algorithm over CT-Apriori was 4.91% while it was 14.54% and 10.60% for Apriori and FP-Growth algorithm. Similar trend was also observed for experiments with web log data.

All these results point CT-PRO is the right candidate for generating a compact version of the original transaction database, which is small in size and which performs frequent pattern mining in a fast and efficient manner.