ABSTRACT

Data mining is an emerging field in the database technology. The goal of data mining is knowledge discovery, that is, to excavate information from historical organizational databases that can be used to guide business strategies and decision making. In this dissertation, fast and efficient algorithms for mining generalized association rules and sequential patterns in massive databases are presented. An association rule, for example, could be " *98% of the customers who buy bread and butter also buy jam* ". The problem is to find out all such rules whose frequency is greater than some user-defined minimum support.

This thesis deals with algorithmic and systems aspects of scalable data mining algorithms applied to massive databases. The algorithmic aspects focus on the design of efficient and scalable algorithms for two-key rule discovery techniques – generalized association rules and generalized sequential patterns. The systems aspects deal with the scalable implementation of these methods on sequential machines.

Two incremental updated techniques for mining generalized association rules and sequential patterns to generate rules and patterns in massive datasets are presented. The first one is the database, which is fixed with changing minimum support. The second one is the given original database, when new increment db (transactional database) is added to the original database DB with fixed minimum support and minimum confidence.

Using partition method association rules and sequential patterns have been generated. The major advantage of the partition method is scanning the database exactly two times to compute the large itemsets by means of constructing a transaction list for each large itemset. In sequential pattern, large maximal sequences are generated using parallel partition method. The speed-up and size-up properties show that the proposed parallel partition method is better than sequential partition method. The method of pattern decomposition can avoid the costly process of candidate set generation and save a great amount of computing time with reduced database size.

The problem of mining generalized association rules and sequential patterns using TID method has been analyzed. By using this method, the cost of execution time has been reduced and linearly scalable.

Another most important problem of mining generalized association rules in the distributed environment has also been presented here. The computing time of the fast distributed algorithm is in the O (n), whereas the parallel based algorithm for mining generalized association rule is in the O (n^2). Hence, the proposed fast-distributed algorithm is more reliable than the previous parallel and sequential algorithm.

Extensive experiments have been conducted for solving the above two problems (generalized association rules and generalized sequential patterns), showing immense improvement over the previous approaches, with linear scalability in database size.