ABSTRACT

The advancements in the field of database technology have made it possible to store a huge amount of data. Data mining techniques have been widely used for extracting non-trivial information from such massive amounts of data. It is useful in many applications like strategic decision-making, financial forecast, medical diagnosis etc. Data mining can be applied either as a descriptive or as a predictive tool. Association rule mining is one of the functionalities of data mining. This thesis proposes a few methods for improving association rule mining, association rule hiding, and post mining.

The process of producing association rules involves the task of finding the set of all the frequent itemsets and generating promising rules. This thesis proposes a technique for determining the frequent itemsets with a single scan of the transactions database in the disk. During this single database scan, information about itemsets and their occurrences are captured in a table kept in the main memory. While determining the frequent itemsets, this table is scanned instead of the disk. This results in great reduction in the total computation time.

In some cases, the number of frequent itemsets to be generated is large. All these itemsets are difficult to handle and manage. This problem can be solved by mining Maximal Frequent Sets (MFS) alone. An MFS of length $m$, implies the presence of $2^m-2$ frequent itemsets. Hence, all the frequent itemsets can be readily inferred from the MFS. Therefore, the generation of MFS reduces the time taken for obtaining all the frequent itemsets. A new faster technique is proposed in this thesis for mining the MFS, which reduces the number of database scans to the maximum of two and also avoids the generation of candidates.

Though data mining has a lot of merits, it has a few demerits also. Sensitive information contained in the database may be brought out by the data
mining tools. Different approaches are being used to hide the sensitive information. It is observed that most of the hiding algorithms in the existing literature, work at the transactions level to hide some sensitive information. This is a time consuming step in the hiding process. In this thesis, two new techniques have been proposed to reduce the time complexity of the hiding process. One technique achieves the reduction of time complexity by modifying the structure of the Frequent Pattern Growth Tree. The other technique achieves the reduction by appropriate partitioning of the database. Theoretical analysis of the algorithms shows that the hiding of data minimizes the time and space complexities to a great extent.

A big organization may have multiple branches spread across different locations. Processing of data from these branches becomes a huge task when innumerable transactions take place. Also, branches may be reluctant to forward their data for centralized processing but ready to pass their association rules. Local mining may also generate a large amount of rules. Further, it is not practically possible for all local data sources to be of the same size. A model is proposed in this thesis for discovering valid rules from data sources of different sizes where the valid rules are high-weighted rules. These rules can be obtained from the high frequency rules generated from each of the local data sources. Support Equalization is another method proposed which focuses on eliminating the low frequency rules at the local sites themselves, thus reducing the rules by a significant amount.

The performance of the improved techniques discussed in the thesis has been examined thoroughly by extensive simulation.