ABSTRACT

Association rule mining is the most important and well investigated data mining technique, used by an organization's decision makers to improve the overall profit. It aids businesses to infer useful information on customer purchase patterns, shelving criterion in retail chains, stock trends and more. Association rule mining is usually carried out in two steps; they are finding frequent item sets and then using these item sets to identify the association rules. It is well known that the first step, frequent item set mining dominates the computational and I/O requirements requiring repeated passes over the entire database. As the volume of data in warehouses and on the internet is growing faster, the scalability of mining algorithms is a major concern.

Classical association rule mining algorithms that require more number of passes over the entire database, can take hours or even days to execute and in the future this problem will only become worse. Sampling approach can be used to solve this scalability problem. In the context of "standard" association-rule mining, use of samples can make mining studies feasible that were formerly impractical due to the enormous time requirements. Indeed, a number of large companies routinely run mining algorithms on a sample of their data rather than on the entire warehouse. Especially, if data comes as a stream flowing at a faster rate, sampling seems to be the only choice. Current research efforts are focused on inventing efficient ways of discovering these rules from large databases.

Likewise, most of the association rule mining algorithms presented in the literature are used for mining the static databases. Now-a-days, research community has focused their researches into the incremental database on behalf

of the real world applications and the necessity of handling the dynamically updating new records.

Applying data mining techniques to real-world applications is a challenging task because the databases are dynamic i.e., changes continuously due to addition, deletion, modification etc., of the contained data. Generally if the dataset is incremental in nature, the frequent item sets discovering problem consumes more time. Once in a while, the new records are added in an incremental dataset. Generally when compared to the entire data set, the size of the increments or the number of records added to the dataset is very small. But the assumption of the rules in the updated dataset may get distorted due to the addition of these new records. Hence a few new association rules may be created and a few old ones may become obsolete. When new transactions are inserted into the original databases, traditional batch-mining algorithms resolve this problem by reprocessing the entire new databases. But they require much computational time and ignore the available mined knowledge.

Several algorithms have been developed for association rule mining which has now become an interesting field of research in the knowledge discovery domain. Discovery of association rules from large and incremental databases is one of the toughest tasks in data mining. Researchers have been motivated to design innovative and incremental algorithm for association rule mining because the quantity of data available in the real life databases are increasing at a tremendous rate.

In this thesis, progressive sampling based algorithms have been developed for efficient association rule discovery from large databases. The core idea behind the proposed approach is that, progressive sampling, based on negative border will enable in determining an optimal sample size for effective mining of association rules. The approach selects an initial sample based on the temporal characteristics of the database and in addition with, the size of an input database. Initially, frequent itemsets are mined from the initial sample using Apriori algorithm. Subsequently, negative border is computed and the itemsets in the negative border is sorted based on their support level. The midpoint itemset is scanned on the remaining set of records in the input database to find the support level. If the support of the midpoint itemset is greater than the user specified support, the chosen sample size is progressively increased. This procedure is repeated until an optimal sample size is obtained and then, association rules are mined from the optimal sample. Finally, the support of the midpoint itemset is analyzed with the different percentage of databases. The empirical validation provides the suitable database size for conducting the midpoint itemset scan. Comparison is also made with random sampling. And the result reveals that the proposed sampling approach achieves over 95% of accuracy.

The thesis also discusses about an efficient incremental mining algorithm, called Enhanced Pre-FUFP algorithm which extends the Pre-large item set algorithm further by including the recency concept. The main aim of the proposed approach is to efficiently handle the items that are included recently in the updated database based on adaptive support threshold. At first, the FP-tree is constructed for the old database and then the transactions of incremental database are processed one at a time. After that, the association rules are mined

from the updated FP-tree by incorporating an adaptive support. Experiments are performed on extensive real life datasets to compare the performance of the proposed approach with that of the Pre-FUFP algorithm. The comparison results show the superiority of the enhanced Pre-FUFP algorithm over other existing incremental algorithms.