Name: Lalit Mohan Goyal

Supervisor: Prof. Tanvir Ahmad

Co-supervisor: Prof. M. M. Sufyan Beg

Department: Computer Engineering

Title of Thesis: Efficient Mining of Data in Distributed Databases

Abstract

A geographical expansion of business leads the databases to be distributed at many sites (computers). While mining distributed databases, many challenges have been faced by the researchers. One of the challenges is to generate frequent itemsets. To resolve this challenge, distributed association rule mining, one of the data mining techniques, is used. It investigates distributed databases for generating frequent itemsets. In this research work, detailed survey has been steered to trace the research gaps and to identify the challenges not only for distributed but also for centralized databases. This survey motivated us to setup the objectives of this research work and directed us to provide some solutions in this regard.

FDM algorithm is one of the most popular algorithms in the field of distributed association rule mining. In this research work, a novel framework and a corresponding algorithm have been proposed to generate frequent itemsets efficiently. Before presenting the proposed framework and the proposed algorithm, new definitions and a lemma have been explained. Proposed algorithm generates optimal number of candidate itemsets and exchanges optimal number of messages among the sites. A numerical example has been illustrated to compare the behavior of both, proposed and *FDM*, algorithms. To evaluate the performance of proposed algorithm, twelve databases have been synthetically generated as described by the previous researchers. These databases have been equally distributed among sites horizontally to carry out the experiments. From the experiments, it has been observed that proposed algorithm works more efficiently than the existing *FDM* algorithm. This research work also emphasized on mining the centralized databases. For this, a probabilistic *Apriori* algorithm has been proposed to generate frequent itemsets. It provides an approximate solution to generate frequent itemsets fast with a compromise on the

number of frequent itemsets. In this algorithm, a new technical term named as "*Probability of Co-existence*" of a potential candidate itemset has been defined, explained and computed. For each potential candidate itemset, the value of *Probability of Co-existence* has been compared against the user specified probability threshold value. This comparison accepts or rejects the scanning of potential candidate itemset in the database. Proposed probabilistic *Apriori* algorithm has been evaluated on synthetic databases and it has been observed that there is great saving in the execution time when compared to standard *Apriori* algorithm.

Standard *Apriori* algorithm always prunes the candidate itemset by assuring that all subsets of the itemset are frequent. In this research work, a new pruning approach has been given which prunes the candidate itemsets with the help of infrequent itemsets. This new approach has been named as filtration approach. The performance of this algorithm has been evaluated on synthetic databases and it has been observed from the experiments that new pruning approach can be used as an alternate to the existing approach.

The work presented in this thesis can be extended further by applying the proposed concept of "*Probability of Co-existence*" in the distributed environment. Approximate solution may be discovered for the other association rule mining algorithms like *FP-Growth, Eclat etc.* Some other concepts about the *Probability of Co-existence* can be evolved.