CONCLUSION
CHAPTER 9

CONCLUSION

This chapter projects the research contributions in this thesis as well as gives the directions for further research work.

9.1 Thesis Summary

The main contributions of this thesis are:

A set of fast algorithms for the manipulation of generalized association rules and sequential patterns and a set of algorithms for incremental updating techniques for mining generalized association rules and sequential patterns have been proposed. The algorithms were implemented on an experimental basis and their performances were analyzed. The experiments showed that the algorithms performed well on large databases. The major contributions of this research work are summarized below:

a) In the second chapter, two new incremental updated algorithms have been presented for mining generalized association rules for solving business and scientific problems that are fundamentally different from the previously known incremental algorithms. Experiments with synthetic as well as flat data showed that the performances on the proposed algorithms are more efficient than the performance of earlier algorithms.

b) In chapter 3, two fast and efficient algorithms have been proposed for mining generalized association rules, which are fundamentally different from Cumulate algorithm. The first algorithm reads the database utmost 2 times to generate all the association rules. The proposed method can also effectively reduce the I/O and CPU overheads for all cases. In the second method, a
Cumulate-PD algorithm has been proposed for mining generalized association rules that quickly reduce the size of the database on each pass making it more efficient to mine all generalized frequent patterns in a large database. The proposed algorithm avoids the costly process of candidate set generation, and saves a great amount of counting time with reduced databases. The study shows that the given method has an excellent performance over the best existing algorithms.

c) In chapter 4, two new fast algorithms have been proposed for mining generalized association rules, given database and the taxonomy of the attributes. All sequential algorithms for mining association rule are multiple passes over the data. An interesting feature of our algorithms is that the database D is not used for counting support after the first pass. The proposed algorithms named BasicTID and CumulateTID have such an efficient feature that it replaces a pass over the original database by a pass over the temporary database. Hence, algorithms BasicTID and CumulateTID are very effective in later passes when the size of the temporary database becomes small compared to the size of the original database. In the second part of this chapter, an improved fast greedy algorithm for mining generalized association rules in large databases was discussed. An interesting feature of the proposed algorithm is that the database DB is not used for counting support after the first pass. Another significant feature of this method is that it replaces a pass over the original database by the temporary database. Hence, the improved algorithm is very effective in later passes when the size of the temporary database becomes small when compared to the size of the database. Performance evaluation shows that our algorithms are superior to Cumulate algorithm. The proposed algorithms were tested on an experimental basis and its performance has been presented. The performance study shows that the given algorithms have an excellent performance over the Cumulate algorithm. Hence, the two algorithms are very much suitable for massive database.
d) In the fifth chapter, a new and efficient algorithm called Fast Distributed Generalized Mining algorithm (FDGM) was proposed for mining generalized association rules on a distributed environment to improve its performance. The algorithm was implemented on an experimental test bed and its performance has been studied. The performance study shows that FDGM has an excellent performance over the direct application of a typical sequential algorithm called Cumulate.

e) In the sixth chapter, the problem of incremental mining of sequential patterns when the sequences in the database are fixed with changing the minimum support has been considered. The first part deals with the proposed incremental updating algorithm IUSP-SD to achieve the fast response time demand. The second part of this chapter presents an efficient algorithm called IUSP-DD for updating frequent sequences when new customer sequences are updated to the set of old frequent customer sequences in a transaction database. IUSP-DD works a dynamic look-ahead strategy in adding the existing transaction database that will not remain frequent patterns after the contribution of the new set of customer sequences. The proposed algorithms work in an iterative manner, but it differs from the other algorithm by scanning the existing old customer transaction database once at the maximum and the new customer database exactly once. Also, it generates and counts the minimum number of candidate sub-sequences in the new database. The experimental test shows that the proposed algorithms perform significantly faster than the AprioriAll algorithm. Hence, the proposed incremental updating algorithms for mining sequential pattern are very much suitable for large static database with dynamic minimum thresholds.

f) In chapter 7, two efficient algorithms called SP-Partition (one is for sequential algorithm and another one for parallel algorithm) have been proposed for mining sequential patterns that are fundamentally different from the known algorithms like AprioriAll. Compared with the previous algorithms, the
proposed algorithms not only reduced the I/O cost significantly but also have lower CPU overhead for most cases. The performance study shows that the SP-Partition algorithms are efficient and scales linearly for mining both long and short patterns. Hence the proposed algorithms are especially suitable for large databases.

g) The eighth chapter of this thesis presents fast and efficient algorithms called AprioriAllSID and GSPSID for mining sequential patterns that are fundamentally different from the known algorithms like AprioriAll and Generalized Sequential Patterns (GSP). The algorithms were implemented on an experimental basis and its performance was studied. The performance study shows that the proposed algorithms have an excellent performance over even the best existing algorithms. The AprioriAllHybrid algorithm which is combination of AprioriAllSID algorithm with AprioriAll algorithm scales linearly with the number of sequences. It has excellent properties with respect to the sequence size and the number of items in the database sequence.

9.2 Future Extension

Data mining or KDD refers to the overall process of discovering new, useful, and understandable patterns in data. Developing fast algorithms is just one of the steps in the KDD process. The other steps include data selection, cleaning and preprocessing, transformation, data-mining task and algorithm selection, and finally post-processing. This KDD process tends to be highly interactive and iterative. Future research needs to target the interaction of the overall KDD process, and sequential, parallel and distributed computing, in addition to developing faster sequential, parallel, incremental and distributed solutions for the core data mining tasks. Some of the problems that need to be addressed are:
a) **Scalability:** Databases will continue to increase in size, in both the number of records and in the number of attributes. Hash based sequential, parallel and distributed processing is ideally suited for addressing issues of scalability. The ultimate goal is to handle giga/tera-bytes of data efficiently.

b) **Incremental Techniques:** In many domains, the data change over time. Old extracted knowledge may need to be updated or may even become invalid. Parallel incremental algorithms can provide fast modification, deletion or augmentation of the discovered knowledge.

c) **Applications:** KDD is ultimately motivated by the need to analyze data from a variety of practical applications, be they in business domains such as finance, markets, telecommunications and manufacturing, or in scientific fields such as biology, bio-informatics, chemist-informatics, astronomy and medicine. Future research must identify new application domains that can benefit from data mining. This will lead to the refinement of the existing techniques and also to the development of new methods where current tools are inadequate.