CHAPTER 7

CONCLUSION

7.1 Discussion of the Research Results of the Thesis

With a view to improve the performance of mining, hiding, and post mining of association rules a number of new techniques have been proposed in this thesis.

The process of producing association rules involves the task of finding the set of all the frequent itemsets and generating promising rules. The first technique described in Chapter 2 proposes a single scan procedure for determining the frequent itemsets. Experiments show that this reduces the time by a factor of above 30, on an average. 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 total computation time.

Maximal Frequentset (MFS) generation is identified as NP-Complete problem. Different techniques exist to generate MFS. The second proposed technique is for mining MFS more efficiently. It requires merely two passes over the transactional database. It generates MFS with the help of simple data structures. A partitioning approach is used to guide the search for MFS. Because of all these, appreciable reduction in time has been achieved, while generating the MFS correctly. The reduction time achieved has been experimentally found to be about 10 to 20% in comparison with the existing Pincer method [19, 20].

Data mining has a few demerits apart from its usefulness. Sensitive information contained in the database may be brought out by the data mining tools. Various approaches are being used to hide the sensitive information. The
third technique proposed in this thesis has addressed the problem of reducing the
time taken to hide association rules. This method has adapted Frequent Pattern
Tree (FP Tree) approach to access the generating transactions from the
transactional database quickly. An data structure called Generating Transaction
Count (GTC) array has been devised to facilitate rapid access. Theoretical
investigation shows that the time reduction can be as much as 250% to 750%.
This predicted performance has been verified by the experimental results of
Chapter 4, while correctly hiding the rules as specified by the user.

Another technique proposed in this thesis has applied to hide essential
association rules even more rapidly. The generating transactions of the rules to
be hidden can be located from the transactional database with minimum effort.
Another partitioning approach is applied to determine the generating transactions
quickly. Hence, it reduces the scan of the transactional database to merely two.
Experimentation of Chapter 5 has shown drastic reduction in time—by as much
as 5 to 15 times.

Discovering valid rules from different-sized data sources without any
artificial manipulation of the data sources is a problem. This has been addressed
in Chapter 6. Two methods have been proposed in this thesis there for
discovering valid high-frequency rule from different-sized data sources without
splitting or merging of data sources to make them equal-sized. The first of these
methods introduces weighted data source selection. The second method equalizes
the supports of the data sources in order to eliminate the low-frequency rules and
thereby exclude them from further processing. Experimentation conducted with a
number of different-sized data sources with different characteristics show that
correct discovery of valid overall rules could be achieved by the proposed
techniques.
7.2 Scope for Future Work

The various methods proposed in this thesis have achieved good improvements in solving the problem addressed by them. However, there are certain points which need further investigations.

The technique proposed in Chapter 2 for single scan mining of association rules achieves reduction of time by cutting-down disk access. The space complexity aspect needs to be further investigated with a view to controlling the memory space required.

Similarly, in Chapter 3 the time complexity has been the main pursuit. The space complexity needs to be further investigated here also.

Again, Chapters 4 and 5 reduce the time taken for hiding the specified rules. But the side effects such as the number of new rules introduced and the number of legitimate rules lost are important. There is good scope for further investigation on these aspects.