CHAPTER 7
CONCLUSION

The growth of Information Technology is overwhelming in many corporate, governmental, and scientific communities with an influx of data that is routinely stored in on-line databases. This trend creates an opportunity for a new research area, Data Mining or Knowledge Discovery in Databases (KDD) that promises to provide high payoffs in many business and scientific domains.

Recently, many companies have disseminated more and more information using Intranets or the Internet. However, most researches have concentrated on mining in centralized data warehouse. Thus, applying knowledge discovery techniques on distributed data warehouse in a parallel environment becomes increasingly important. Over the years, data set sizes have been growing rapidly with the advances in technology, the ever-increasing computing power and computer storage capacity, and the increasingly automated business, manufacturing and scientific processes. To mine large and distributed data sets, it is important to investigate efficient parallel algorithm for distributed environment to reduce the communication overhead, central storage requirements, and computation cost. This thesis covers the issues of selecting dataset from a centralized warehouse and allocating them to a distributed environment in a parallel approach using parallel data mining algorithms and dynamic remote
memory utilization to obtain an optimized response time with optimal mathematical models.

7.1 Broad Conclusion

The distributed workstations are identified and the dataset is allocated based on the CPU utilization time and the task allocation schedule. This task allocation is carried out using a simulated annealing algorithm explained in Chapter 3. Association rule mining algorithms are used for the mining process in the distributed workstation. Association rule mining algorithms are widely used in various areas such as Telecommunication networks, Market and Risk management, Inventory control etc., Association rule mining is to find out association rules that satisfies the predefined minimum support and confidence for a given database. The association rule mining is usually decomposed into two sub problems. One is to find those itemsets whose occurrences exceed a predefined threshold in the database; those itemsets are called frequent or large itemsets. The second is to generate association rules from those large itemsets with the constraints of minimal confidence.

The computational cost of association rules mining can be reduced in four ways:

- by reducing the number of passes over the database
- by sampling the database
- by adding extra constraints on the structure of patterns
- through parallelization.
7.2 Research Contributions

This section presents the research contributions of this thesis. The thesis has contributed towards the mining process in geographically distributed environment with the result of an optimized response time. The specific contributions made by this research are as follows:

i. Development of a novel framework for performing parallel and distributed data mining with dynamic remote memory utilization.

ii. Designing a model for the cost calculation for obtaining the optimized response time with suitable mathematical formulation.

iii. Development of a new protocol in Microsoft C#.NET for simulating the new framework and to produce the required results.

iv. A novel model for knowledge integration is developed from the geographically distributed workstations with optimal mathematical models.

v. Monitoring the mining process with mobile intelligent agents.

vi. XML dataset is used for the mining process and multiple nesting is handled properly.

vii. An optimized response time is obtained with mathematical cost computations.

viii. The results are validated and presented in two journals and four international conferences.
7.3 Major Strength of the Approach

Today many organizations, companies, and scientific centers produce and manage large amounts of complex data and information. Climate data, astronomic data and company transaction data are just some examples of massive amounts of digital data repositories that today must be stored and analyzed to find useful knowledge in them. This process is both computationally intensive and collaborative and distributed in nature. The major strengths of this research works are listed as follows


The data mining is carried out in distributed workstations with parallel algorithms so that the mining process could be carried out in parallel as soon as the data allocation is completed.

b. Dynamic remote memory utilization for utilizing the memory effectively.

The utilization of memory for the distributed workstations is carried out effectively with remote memory utilization. The intelligent agents can monitor the memory utilization during the mining process. If there is any fault in one workstation then they could be allocated to the next available workstation.

c. Knowledge Integration Process with relevant models.

The knowledge obtained from the distributed locations is integrated with global exceptional identification. The Bayesian network is also used for the knowledge integration purpose which provides more accurate results.
d. Optimized Response Time.

The response time obtained after performing the mining, knowledge integration and the communication process with an optimizer using relevant mathematical models, an optimized solution is obtained.

e. XML Rules.

The XML dataset is used so that the general XQuery optimization techniques are carried.

f. Intelligent Agents

Intelligent agents are used in parallel and distributed data mining for performing the monitoring process during mining, task allocation and in dynamic remote memory utilization.

7.4 Future Research Directions

- Extensible to Neural Network Applications

In a parallel and distributed data mining environment the response time could be modeled with a neural network approach. Bivens et al., [BE, 00] used neural network structures which reflected the physical network topology to determine and avoid network problems. In these neural net-work structures, the hidden layer had as many neurons as the input layer and the hidden layer neurons aggregated data from input neurons of the corresponding network workstation as well as the input neurons corresponding to workstations connected to this workstation in the physical network.
This method of network topology inclusion is discussed in [BE, 00]. This could easily be extended to application workflow information and used to build a specialized NN model. The neural network application could be used for identifying the distributed workstations from different locations and representing different graphical models based on the geographically distributed site.

- **Incorporation of Cluster Analysis**

  Parallel Clustering algorithms can be used to perform the mining process in distributed environment. Clustering is to group data objects into classes of similar objects based on their attributes. Each class, called a cluster, consists of objects that are similar between themselves and dissimilar to objects in other classes. The dissimilarity or distance between objects is measured by the given attributes that describe each of the objects. The distance between two clusters can be determined by single link, average link, complete link or centroid-based metrics. At present this research work concentrates on parallel association rule mining algorithms, Clustering algorithms can also be used as an extension to the existing work. Different approaches of clustering techniques can be employed.

- **Privacy Preserving in Distributed Data Mining**

  Privacy preserving in parallel and distributed data mining can be a secure means of mining process. Data mining technology has emerged as a means of identifying patterns and trends from large quantities of data. Data mining has used a data warehousing model
for gathering all data into a central site, then running an algorithm against that data. Privacy considerations may prevent this approach. For example, the Centers for Disease Control (CDC) may want to use data mining to identify trends and patterns in disease outbreaks, such as understanding and predicting the progression of a flu epidemic. Privacy-preserving data mining has emerged to address this issue. One approach is to alter the data before delivering it to the data miner. The second approach assumes that the data is distributed between two or more sites, and these sites cooperate to learn the global data mining results without revealing the data at their individual sites. This approach was first introduced to the data mining community by Lindell and Pinkas[LP, 00], with a method that enabled two parties to build a decision tree without either party learning anything about the other party’s data, except what might be revealed through the final decision tree. This could also be an emerging field for the further enhancement of this research work as this research work is mainly focused on geographically distributed workstations where the data has to be handled very safely. Various levels of security aspects could be handled.

- **Incorporation of Incremental Mining Algorithm**

  Incremental algorithms can be used for online updating in a parallel and distributed environment. In a distributed incremental approach, individual agents have access only to a limited number of transactions. Therefore, by employing the Apriori algorithm, they only
have a partial view of the association rules. However, they can memorize the rules with a lower support and gradually update them, as they access more databases or communicate with other agents. In the related work the Incremental mining algorithm [Tho, 97], [Thu, 00], [AIS, 93] is used for finding new frequent itemsets with minimal re-computation when new transactions are added to or deleted from the transaction database. Incremental algorithms could be used for the real time data warehouse because the online updates would be reflected automatically. The mining need not be done for the whole data warehouse again instead the incremental algorithms could be used so that only the newly updated part will be mined.

- **Mining of Heterogeneous Datasets**

  This research work focuses on the homogeneous dataset. The ability to interact with heterogeneous data source is critical. Nowadays a significant amount of time and effort is spent in accessing/querying multiple remote or local heterogeneous data sources and integrating the results of these searches either manually, or with the aid of data integration tools. There should be a good knowledge about the content, structure, and representation of the data in such databases, and their implementation, and query capabilities. Also when users have a complex query or information from different sources needs to be combined they need to construct a query plan on their own. In the heterogeneous case the table is vertically partitioned, each site contains a collection of columns (sites
do not have the same attributes). However, each tuple at each site is assumed to contain a unique identifier to facilitate matching. It is important to stress that the global table viewpoint is strictly conceptual. It is not necessarily assumed that such a table was physically realized and partitioned to form the tables at each site. The work can be further expanded with these types of heterogeneous datasets where a new framework could be created for various datatypes from distributed locations.

- **Knowledge Integration**

  In the emerging network knowledge environment, the relevant data for many computations may reside on a number of geographically distributed databases that are connected by communication networks. A common constraint in these situations is that the data cannot be moved to other network sites due to security, size, privacy, and data ownership considerations. An example of such a situation is that there may be a need to compute decision trees, association rules, or some complex statistical quantities using data from a census database, a diseases database, a labor statistics database, and a few pollution databases located in ten different cities across the country. It is impossible to bring these databases together and join them for performing some computations. Additionally, a new instance of some computation may require data from a different set of participating workstations and databases. The knowledge discovery process takes the raw results from data mining (the process of
extracting trends or patterns from data) and carefully and accurately transforms them into useful and understandable information. The term Knowledge Data Discovery (KDD) is increasingly being used as a synonym for data mining. It is a more descriptive term and can be applied to all the activities and processes related to the discovering of useful knowledge from the data. Using a combination of techniques - including statistic analysis, neuronal logic, diffuse logic, multidimensional analysis, data visualization and intelligent agents - the KDD can discover useful patterns to develop models that can predict behaviors or consequences, in a large variety of knowledge spheres. Knowledge integration could be an emerging and latest trend for data mining researchers. The Knowledge integration part can be used with other mathematical models such as Rough Sets, and other Operations Research techniques so that it could be an optimized solution.

- **Intelligent Distributed Data Mining**

  Advances in computing and communication over wired and wireless networks have resulted in many pervasive distributed computing environments. The internet, intranets, local area networks are some examples. Many of these environments have different distributed sources of voluminous data and multiple compute nodes. Analyzing and monitoring these distributed data sources require data mining technology designed for distributed applications. A scalable solution for distributed applications calls for distributed processing of
data controlled by the available resources and human factors. For example consider an ad-hoc wireless sensor network where the different sensor nodes are monitoring some time critical events. Central collection of data from every sensor node may create heavy traffic over the limited bandwidth wireless channels and this may also drain a lot of power from the devices. A distributed architecture for data mining is likely to reduce the communication load and also reduce the battery power more evenly across the different nodes in the sensor network. There are various similar needs in ad-hoc wireless networks of mobile devices like PDAs, cell phones, and wearable computers. Intelligent Distributed data mining could be used for these types where the algorithms can be selected automatically depending upon the time and need of the application. Automated Intelligence mining may be applied with less human intervention so that the time complexity could be reduced.

- **Ant Colony Optimization**

  Ant colony optimization (ACO) metaheuristic, a novel population-based approach was proposed in 1992 by Marco Dorigo et al [MT, 04], to solve several discrete optimization problems. The ACO mimics the way real ants find the shortest route between a food source and their nest. The ants communicate with one another by means of pheromone trails and exchange information about which path should be followed. The more the number of ants traces a given path, the more attractive this path (trail) becomes and is followed by
other ants by depositing their own pheromone. This auto catalytic and collective behavior results in the establishment of the shortest route.

Ants find the shortest path from their nest to the food source with the help of pheromone trail. This characteristic of ants is adapted on ant colony optimization algorithms to solve real problems with using exactly some characteristics of ants and some new addition. The method improved by modeling real ants using exactly the same specifications taken from real ants is below:

- the communication established with ants through pheromone trail.
- Paths deposited with more pheromone preferred previously.
- Pheromone trail on short paths increase more rapidly.

Addition of new specifications to this new technique is below:

- They live in an environment where time is discrete.
- They will not be completely blind, they will reach the details about the problem.
- They will keep information formed for the solution of the problem with has some memory.

Ant colony optimization algorithm aiming can be used to solve the data-clustering problem and the technique used is given below.

Two important directions for future research are as follows. First, it would be interesting to extend Ant-Miner to cope with
continuous attributes, rather than requiring that this kind of attribute be discretized in a preprocessing step. Second, it would be interesting to investigate the performance of other kinds of heuristic function and pheromone updating strategy. This method could be well applied for optimizing the model which is applied for performing the mining process.

- **Extensible with other types of data**

Data mining is a difficult task. It requires complex methodologies, including problem definition, data preparation, model selection, and model evaluation. This has limited the adoption of data mining at large and in the database and business intelligence (BI) communities more specifically.

A data-centric design refers to analytical activities where:

- No concepts of a model or complex methodologies are exposed to the user.

- The only information required for a computation is a data source. All supporting objects are either deleted at the end of the computation or are linked to a data source.

Based on these a new data centric and a common approach for all data types may be provided for performing an easier mining process. A novel framework can be developed so that depending on the type of the data it would be accommodating the values.
• **Parallel Algorithms for Distributed data Mining**

There are many parallel algorithms for performing the distributed data mining process. The different parallel algorithms are Count Distribution, Candidate Distribution, Hybrid Count, Hash Partitioned Apriori, etc. A novel and new parallel algorithm can be designed based on the dynamicity of the application to be suited in a distributed environment. The scalability, time complexity and the performance could be improved.

**7.5 Research Summary**

In this research work, parallelization is used for reducing the computational cost. The two parallel algorithms in Apriori namely the Count Distribution and Hash Partitioned Apriori are used with different sizes of datasets as described in Chapter 4.

The Intelligent Agents play a vital role in this research. They are used for monitoring the entire distributed environment before the mining and also during the mining process. A mobile agent can navigate through a network and perform tasks on remote machines as per the instructions given. The different types of intelligent agents and their roles with relevant pseudo code are dealt with elaborately in Chapter 3. The dynamic remote memory utilization is also taken care of efficiently in this research work so that the memory will be utilized properly without any heavy overload. This is also dealt with in Chapter 3.

After obtaining the mined results, the knowledge from the distributed workstations are integrated to form the result which is to be communicated
to the server. This knowledge integration uses a specific algorithm to identify exceptional patterns dealt with in chapter 5. The response time is an important primary concern of this research work. The response time is calculated with relevant mathematical models and the optimized solution obtained is discussed in chapter 5. XML is used as the dataset as it is becoming a standard to represent semi-structured data. XML data becomes an important data mining domain, and it is of interest to develop association rule mining methods for such data. The results and discussions are given in chapter 6.