Chapter - I

Introduction

1.1 Preamble

Data mining is a subfield of computer science that comprises of techniques for determining interesting, potentially useful and previously unknown patterns from large datasets. It involves many techniques to accomplish different mining tasks. These techniques can be described into two categories: Predictive data mining and Descriptive data mining. Predictive model is used to make predictions on the basis of historic data. It includes classification, regression, time series analysis and prediction. Descriptive model is used to identify patterns or relationships in data. It serves a way to explore the properties of examined data. Descriptive mining includes clustering, summarization, association rules and sequence mining.

With growing size of organizations, size of data repositories is also growing at an exponential rate. Conventional mining techniques fail to mine from such large data. Distributed sources of voluminous data, reduced IT infrastructure cost and advancements in computing has opened up the opportunity of mining in distributed computing environment. Data mining algorithms are implemented and evaluated on parallel multi core low cost GPUs which has proved to be a cost effective solution for large scale mining. A cost effective solution to mine data in distributed computing environment is developed. Data mining when implemented on high performance parallel processing and distributed systems, can help to analyze massive datasets.

1.2 Distributed Systems

A distributed system is a collection of autonomous computers that are connected through a network. Here, a middleware helps the computers to coordinate the activities and to share the resources. The components are located on networked computers that communicate and coordinate their actions by passing messages. Users of distributed system perceive the system as a single, integrated computing facility.
1.3 Distributed Computing

Distributed computing is a field of computer science where computations are executed in distributed environments. Distributed computing also refers to the use of distributed systems to solve computational problems. In distributed computing, a problem is divided into many tasks, each of which is solved by one or more computers. It is a combination of hardware and software having more than one processing or storing element. It is a type of parallel computing where the system may be loosely or tightly coupled.

1.4 Distributed Computing vs. Parallel Computing

Distributed computing is basically a type of parallel computing, but parallel computing term is generally used to refer to processing where different parts of a program run simultaneously on two or more processors that are part of the same computer. Distributed computing also requires the division of the executing program or the data but different sections runs on different autonomous computers.

1.5 Distributed Data Mining

Distributed Data Mining explores techniques of how to apply Data Mining in a non-centralized way. Distributed computing plays an important role in the Data Mining. There are mainly two reasons that necessitate the use of distributed mining. First, data mining generally requires large amount of storage resources and computation time. Second, there are applications where data is generated in different storage locations. Transferring of data to a centralized location may be time consuming and makes the system inefficient. To make data mining process efficient and scalable, it is important to develop new techniques in the direction of distributed data mining.

1.6 Motivation

These days, the use of Information and Communication Technology has spanned across a large number of applications. Data mining techniques [1] evolved as a
requirement when enormous data started accumulating in digital format. With the availability of large data in application areas like telecommunication, banking, insurance, retail, health care, scientific areas, medical sciences etc., better data mining solutions are in real demand.

Data mining techniques [2] are helpful in digging out hidden previously unknown information from raw data. Data mining offers a wide range of algorithms that are used for analysis, pattern discovery and prediction. It can be helpful to find relationships and categories within data. Decisions based on better analysis might help to increase profitability. Thus, management needs to provide budget and resources for these services.

Conventional data mining techniques works well on structured data which is clean and properly stored at a central location. Mining data from distributed sources is challenging. Distributed data mining [3] is a new paradigm that helps in knowledge extraction in distributed scenarios.

1.7 Rationale

The field of data mining has emerged as a fundamental research area with important applications to science, engineering, medicine, business, and education. While studying the application of data mining in different domains like WHO database and ATM machine maintenance, it was felt that there exists a need to develop superior data mining solutions for distributed environment.

Challenges that call for improvement in existing mining techniques are listed below [4]:

- Mining information from distributed and heterogeneous databases
- Mining from Complex Data types
- Mining from very large Data
- Improving efficiency of existing data mining algorithms
- Parallel, distributed, and incremental mining
- Security, privacy and data integrity in Data Mining
- Handling Data Mining process-related problems
In view of the challenges, a number of research papers are studied. Centralized approaches may result in delayed outcomes. Therefore, parallel and distributed approaches are required for cost effective and time efficient data mining systems.

1.8 Current Approaches

In last two decades, lot of research is carried out in improving performance of data mining techniques. Three important data mining approaches are discussed next.

1.8.1 Centralized Data Mining

The basic central data mining techniques included in this study are [5]:

<table>
<thead>
<tr>
<th>Data mining Technique</th>
<th>Centralised Data Mining Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Association Rule</td>
<td>-A Priori, Pincer Search, FP Tree Growth</td>
</tr>
<tr>
<td>Clustering</td>
<td>-K Means, K Mode, EM Algorithm</td>
</tr>
<tr>
<td>Classification</td>
<td>-C4.5, CART, SVM, kNN, NaiveBayes,</td>
</tr>
<tr>
<td>Prediction</td>
<td>-AdaBoost</td>
</tr>
<tr>
<td>Others</td>
<td>-PageRank</td>
</tr>
</tbody>
</table>

1.8.2 Parallel Data Mining

Various data mining algorithms for finding association rules, clustering and classification are modified for parallel processing architectures [6, 7, 8, 9]. Two main approaches were used for parallelization: Task parallelism and Data parallelism [10]. Initially, parallel mining was restricted to multi core CPUs and shared memory multi computers [11]. Parallel computing techniques took a boost with the advent of multi core CPUs and cheaper GPUs [12]. A combination of CPU and GPU resulted in multi fold performance benefit.
Parallel programming is incomplete without discussing on the recent approach called Map Reduce [13] that was introduced by Google in 2004. It offers a parallel programming environment. Map Reduce is a popular platform for mining large-scale datasets in parallel as well as distributed environment [14].

1.8.3 Data Mining in Distributed Computing Environments

Distributed Computing environments include PEER to PEER systems [15,16], GRIDS [17,18,19,20], CLUSTERS [21,22,23, 24] and Cloud [25].

Peer to peer system represents a group of computers connected together to combine computing and processing abilities. In P2P systems, computers can directly communicate without the help of a server. GRID is a collection of tightly coupled and geographically distributed heterogeneous computers which are combined to work together for related problems. In a CLUSTER, all machines are homogeneous and work as a single unit. The computers in the cluster are normally in a single location. CLOUD architecture refers to a distributed infrastructure that provides services and resources through internet.

A great deal of work is carried out in PEER to PEER, GRIDS, and CLUSTERS but very less literature is available related to the execution of data mining techniques in cloud environment.

1.9 Challenges

Business volumes are growing at a fast pace resulting in exponential growth of database. Mining of large sized data is a big challenge these days. The data is so voluminous that digging out hidden information becomes an exhaustive process. Fast and efficient mining solution requires high end costly servers. The setup may not be affordable for small and middle level organizations. Mining large data in a faster and economically efficient way is one of challenge that was addressed.

Conventional mining techniques are centralized and the data needs to be accumulated at some central location. But in a distributed system, where the data is distributed over the network, it is challenging to create a central data store.
Extra time is incurred in pre-processing and collecting the data. If the process of data extraction is automated, and the data mining techniques are provided as a service over the network, performance of overall knowledge discovery system can be improved.

Another challenge is to mine iterative data mining algorithms in distributed scenario. Few data mining algorithms are highly iterative and the basic parallel programming model cannot be directly exploited for them. There is a requirement of efficient programming model for such algorithms.

Looking at the challenges mentioned above, following solutions are proposed:

i. Parallel data mining using GPUs.
ii. Development of efficient framework for distributed data mining.
iii. Improving the efficiency of iterative data mining algorithms by modifying the working of Hadoop Map Reduce model.

Frequent Item set Mining and K- Means algorithm are investigated in parallel and distributed environment respectively.

1.9.1 Why these challenges?

The title of the thesis is “A Framework for Query Optimization for Data Mining in Distributed Environment”. There are some query languages in data mining. Among them, 'DMQL', 'MSQL', 'MINERULE' and 'OLEDB for Data Mining' [26, 27, 28] are worth mentioning. These data mining query based languages were mainly developed to exploit the advantages of database technology. During the beginning of our study, it was felt that existing query languages can be optimized. However, later part of the research revealed that query languages are not commercially viable. With the advancements in data mining tools, alternatives used against these query languages have resulted in better knowledge extraction. Therefore, our research is focused on optimization and efficiency of existing data mining algorithms through parallel and distributed computing. Using parallel and distributed approaches, performance of mining algorithms is investigated.
1.10 Research Objectives

- Study of existing data mining algorithms and their applications.
- Survey of data mining techniques in centralized, parallel and distributed systems.
- Optimizing performance in data mining techniques using parallel processing.
- Developing a framework for algorithm optimization for data mining in distributed environment.
- Designing efficient programming model for iterative data mining algorithms.

1.11 Our Contributions

The thesis investigates the research aspects and challenges in the direction of parallel and distributed data mining. The overall contribution from the thesis can be summarized as below: -

i) Investigation of data mining algorithms and their applications. Application of data mining techniques on different domains like World Health Organisation and ATM machine maintenance is studied.

ii) A survey of centralised, parallel and distributed data mining techniques is carried out and various possibilities of optimizing data mining in parallel and distributed environment are explored.

iii) Designed a parallel algorithm for Frequent Item Set Mining on GPU (Graphical Processing Unit). Performance improvement is achieved by applying data compaction technique so as to ensure less data scans. Instead of using only CPU cores, a combination of CPU and GPU was used. Efficiency was improved through task parallelism implemented on parallel GPU cores.

iv) Proposed a cloud based data mining solution for cost effective data mining. Distributed data mining was implemented using Hadoop Map Reduce and data
storage as a service using ownCloud. It also served the purpose of developing a cloud based academic environment for student experimentation.

v) Developed a new framework for Map Reduce that supports iterative computation efficiently. Parallel K-Means algorithm is designed and tested on the proposed framework and its performance with typical Hadoop Map Reduce is carried out.

1.12 Organization of Thesis

Background of the proposed research work is discussed in Chapter 1. Different data mining techniques are introduced followed by rationale, challenges and research objectives. Based on the rationale presented in Chapter-I, extensive research on different data mining techniques was carried out and the details of the literature reviewed is stated in Chapter-II. The chapter presents a summary of literature related to algorithms and research conducted in three areas: centralised data mining, parallel data mining and distributed data mining. As a result of extensive literature survey, different algorithmic and architectural approaches followed in various parallel and distributed data mining techniques were identified. Chapter-III presents the methodology for applying and analysing centralised, parallel and distributed data mining. Various emerging trends in these mining techniques are also proposed. After detailed study of different data mining techniques, experiments were carried out to practically execute data mining techniques on different application domains. Chapter-IV describes the application of data mining techniques the data of World Health Organisation and ATM machine maintenance. Chapter-V is about applying frequent item set mining using parallel processing. Extensive experimentation revealed multi fold improvements in the performance of FIM algorithm. In Chapter-VI, a framework for Data Mining as a Service on Private Cloud is proposed. The framework is developed using a combination of ownCloud, ETL and Hadoop MapReduce. K-Means algorithm is optimized by executing the algorithm using distributed computing approach. Chapter-VII tries to elaborate upon the bottlenecks in Hadoop Map Reduce Programming model. With few modifications in the basic model, a
framework is proposed and it is observed that new framework performs better on iterative data mining algorithms. Chapter-VIII is about conclusion and future scope.