CHAPTER 1

INTRODUCTION

Due to the enormous amount of data in various application domains, the requirements of database systems have changed. The convergence of computing and communication has produced a society that feeds on information. Yet most of the information is in its raw form: data. If data is characterized as recorded facts, then information is the set of patterns, or expectations, that underlie the data. The data is overwhelmed. The amount of data in the world, in our lives, seems to go on and on increasing and there’s no end in sight. The personal computers that are omnipresent make it too easy to save things that previously trashed. Inexpensive multi gigabyte disks make it too easy to postpone decisions about what to do with all this stuff.

Moore’s law, states that processing speed of a computer doubles about every 18 months. Computer storage capacity doubles about every nine months [GG, 99]. Computer databases expand to fill available storage capacity. This, results in large amounts of data in databases representing an untapped resource. Information can be extracted from these data. As data continue to amass and yield more information, another level of distillation called knowledge is produced. Techniques to analyze the given information and find so far hidden knowledge are mandatory to draw maximum benefit from the collected data.

Knowledge is the essence of information. A highly successful and widely popular process to extract knowledge from very large mountains of data is data
mining. Data Mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner [DHP, 01]. The idea is to build computer programs that sift through databases automatically, seeking regularities or in the process of discovering patterns in data. Many current KDD methods and tools are not truly interactive and cannot easily incorporate prior knowledge about a problem except in simple ways. The use of domain knowledge is important. The process must be automatic or (more usually) semiautomatic. Data mining is a difficult and laborious activity that requires a great deal of expertise for obtaining quality results.

1.1 Data Mining

In the last few years the application domain of Data Mining (DM) and its related techniques and technologies have been greatly expanded. Progress in digital data acquisition and storage technology has resulted in the growth of huge databases. This has occurred in all areas of human endeavor, from the mundane (such as supermarket transaction data, credit card usage records, telephone call details, and government statistics) to the more exotic (such as images of astronomical bodies, molecular databases, and medical records). The interest has grown in the possibility of tapping these data, extracting information that might be valuable to the owner of the database. The discipline concerned with this task has become known as data mining. Data mining is a new discipline that seeks to do just that: by sifting through these databases, summarizing
them, and finding patterns. Huge data collections may be analyzed and examined in an unlimited number of ways.

Data mining is often set in the broader context of knowledge discovery in databases, or KDD. This term originated in the artificial intelligence (AI) research field. The KDD process involves several stages: selecting the target data, preprocessing the data, transforming them if necessary, performing data mining to extract patterns and relationships, and then interpreting and assessing the discovered structures. This research work does not concentrate on data preprocessing issues such as data cleaning, data verification, and defining variables. Instead this research work concentrates on automatically analyze and extract knowledge from data contained within a database.

The development of automated data collection tools and the ensuing tremendous data explosion have fueled the imperative need for better interpretation and exploitation of massive data volumes. The continuous improvement of hardware along with the existence of supporting algorithms has enabled the development and flourishing of sophisticated DM methodologies. Richard and Michael [RM,03] define data mining as the process of employing one or more computer learning techniques to automatically analyze and extract knowledge from data contained within a database. The knowledge gained from a data mining session is given as a model or generalization of the data.

Data Mining is the process of nontrivial extraction of implicit, previously unknown, and potentially useful information such as knowledge rules,
constraints, and regularities from data stored in repositories using pattern recognition technologies as well as statistical and mathematical techniques [SF, 00]. Data mining has, for good reason, recently attracted a lot of attention: it is a new technology, tackling new problems, with great potential for valuable commercial and scientific discoveries. Data Mining is predicted to be "one of the most revolutionary developments of the next decade," according to the online technology magazine ZDNET News (February 8, 2001). Data Mining is one of the 10 emerging technologies that will change the world says the MIT Technology Review.

1.2 Vertical Partitioning

The current trend in information management is towards a object-oriented data base environment [BOS, 95]. Next generation database applications require extra modeling supports, higher performance requirements and new design techniques. Object oriented databases (OODBs) have gained a considerable attention mainly because they reduce the gap between real world concepts and data representation models. Data volumes are increasing at an astonishing rate in the commercial world due to increase in number and complexity of transactions. The world surrounding us generates various types of data in abundance. Object oriented database provides rich modeling features like, encapsulation of methods, inheritance, object identity, arbitrary data types and complex objects.
The partitioning of related objects should be performed before clustering for an efficient access in object-oriented databases. Class partitioning is the process of clustering relevant data accessed by an application into a class. This reduces the amount of irrelevant data accessed, thus reducing the number of disk accesses. In databases, the clustering of data is needed to store and retrieve related data together. Generally, the clustering of data can be divided into two phases. In the first phase, partitioning of related data is done. The second phase is to rearrange data in the partition block so that data which are more likely to be accessed together are located closely to increase the performance.

Partitioning in database design is the process of assigning a logical object (relation) from the logical schema of the database to several physical objects (files) in a stored database. Vertical partitioning subdivides the attributes into groups and assigns each group to a physical object. In other words, vertical partitioning refers to the dissection of a relation into a set of relations, each containing a subset of attributes of the original relation. Horizontal partitioning subdivides object instances (tuples) into groups, all having the same attributes of the original object. In other words, horizontal partitioning refers to the dissection of a relation so that each smaller relation contains the same number of attributes as the original relation but with different tuples in each partition. Here the physical objects are a result of vertical or horizontal partitioning as horizontal or vertical fragments. Vertical partitioning is the process of subdividing the
attributes of a relation or a record type, creating fragments. Vertical partitioning is a process of generating the fragments, each of which is composed of attributes with high affinity. The concept of vertical partitioning are applied to many research areas, especially databases and distributed systems, in order to improve the performance of query execution and system throughput. Vertical partitioning is a technique for facilitating efficient execution of next generation database applications by reducing irrelevant instance variable (attribute) access. In this research work an object oriented database is vertically partitioned using vertical partitioning algorithm.

1.3 Ranking of Attributes

The ubiquitous usage of databases for managing structured data, compounded with the expanded reach of the Internet to end users, has brought forward new data retrieval scenarios. Ranking and returning the most relevant results of a query is a popular paradigm in Information Retrieval. Ranking elements have been widely discussed in a variety of settings. In new applications, such as E-commerce or multimedia, users want to find best matching tuples over a database. The success of many data mining algorithms relies on their abilities to identify a small subset of relevant attributes. Attributes play a crucial role in all data mining algorithms. Their qualities, quantities, relevancy, and redundancy have different levels of impact on the performance of data mining algorithms [BL, 97]. The motivation for ranking has increased along with the size of databases. Users often need to optimize the selection of objects
By appropriately weighing the importance of multiple object attributes. The attributes in an object oriented database are ranked in decreasing order of their relevance to a predictive data mining task, i.e., the attribute most relevant for the prediction of the target attribute's value is at the top of the ranked list are more useful while clustering. The higher the ranking of attributes the higher the importance is given during clustering the object oriented database.

1.4 Clustering

There is an enormous wealth of information embedded in huge databases belonging to enterprises and this has spurred tremendous interest in the areas of knowledge discovery and data mining. Large data sets are now commonly used in most business organizations. So much data has been gathered that asking even a simple question about the data has become a challenge. Interesting patterns can be discovered from the large volume of data. There are several data mining techniques fulfilling these objectives. These are categorized as association, classifications, sequential patterns, and clustering. Clustering is a useful technique for discovery of data distribution and patterns in the underlying data. The goal of clustering is to discover dense and sparse regions in a data set. An important class of problems in the areas of decision support and reporting are clustering, in which, one is interested in finding groupings (clusters) in the data.

Clustering has been studied in statistics, pattern recognition, machine learning, and many other fields of science and engineering for decades. Each cluster includes records that are more similar to members of the same cluster
than they are similar to rest of the data. Clustering is a useful technique for grouping data points such that points within a single group/cluster have similar characteristics (or are close to each other), while points in different groups are dissimilar. Clustering is one of the most important tasks performed in Data Mining Applications. Clustering algorithms partition data into certain number of clusters groups, subsets, or categories. The goal of clustering, or unsupervised learning, is to discover "natural" groupings in a set of patterns, points, or objects, without prior knowledge of any class labels.

Cluster Analysis finds groupings in the data, and identifies homogenous ones of the groupings as clusters. Clustering has modern applications in numerous domains, such as biomedical data, software engineering, psychology, marketing, image processing, geographical information systems, statistics, biology, and others. Datasets emerging in such domains are often too large and too complex for human analysis. Such domains often need automated clustering solutions to support long-term strategic planning and decision making [AK, 06], [AA, 00], [UA, 99], [ZK, 05]. Surveys regarding clustering algorithms are found in [RD, 05], [ARS, 01].

1.5 Software Agents

Data Mining systems encompass a number of discrete, nevertheless dependent tasks, they can be viewed as networks of autonomous, yet collaborating units, which regulate, control and organize all the potentially distributed, activities involved in the knowledge discovery process. Agent
technology has introduced a windfall of novel computer-based services that promise to dramatically affect the way humans interact with computers. The use of agents may transform computers into personal collaborators that can provide active assistance and even take the initiative in decision-making processes on behalf of their masters.

Although the word agent is used in a multitude of contexts and is part of our everyday vocabulary, there is no single all encompassing meaning for it. Perhaps, the most widely accepted definitions for this term is that “An agent acts on behalf of someone else, after having been authorized”. Software agents are instantiated and act instead of a user or a software program that controls them.

Agents are defined as software entities that perform some set of tasks on behalf of users with some degree of autonomy [RN, 95]. An agent is a virtual entity, which runs approximately as follows: [Fer, 99]

- Which is capable of acting in an environment,
- Which can communicate directly with other agents,
- Which is driven by a set of tendencies,
- Which possesses resources of its own,
- Which is capable of perceiving its environment,
- Which possesses skills and can offer services,
- Which may be able to reproduce itself,
- Which has only a partial representation of this environment,
- Whose behavior tends towards satisfying its objectives, taking account of the resources and skills available to it and depending on its perception, and the communications it receives.

Agents are not directed by commands which come from a user or another agent, but by a set of tendencies, which can take the form of individual goals to be achieved or satisfaction or survival functions which the agent attempt to optimize. There are many terms used in describing the characteristics of software agents, such as autonomy, intelligence, reactivity, proactivity, personality, adaptivity, mobility, continuity, rationality, collaboration, and so on. Actually no agents have all these characteristics at the same time. Agents are endowed with autonomy. Autonomy is the most essential factor that differentiates the agent from other simple program, and has been the most widely accepted.

As mentioned in Jennings and Wooldridge autonomy is a difficult concept to pin down precisely, but it means that the system should be able to act without the direct intervention of humans (or other agents), and should have control over its own actions and internal state [JW, 98]. This can be illustrated in the following figure 1a.
As shown in Figure 1a, an agent encapsulates so-called “control” and “resource,” which means that an agent knows how to function by itself. This is the difference between an agent and a general software entity.

Agents also must have intelligence. Intelligence is the degree of reasoning and learned behavior: the agent’s ability to accept the user’s statement of goals and carry out the task delegated to it.

According to Arnon and Onn the following are the general concepts according to an agent. [AO, 03]

a. Autonomy: is the ability of an agent to operate without supervision.

b. Reactiveness: is the ability of an agent to respond in a timely manner to changes in the environment.

c. Proactiveness: is the ability of an agent to pursue new goals.

d. Sociality: is the ability of an agent to interact with other agents by sending and receiving messages, routing these messages, and understanding them.
Software agents are becoming a mainstream technology used widely in many areas including intelligent user interface, industry, Internet and WWW, electronic commerce, business process management, digital library, electronic meeting, entertainment, network management, virtual reality, and artificial society and so on. Software agents are becoming more and more an important software development technology. Software agents can reduce the amount of interaction between humans and the data mining process, which are expected to operate in a more efficient and intelligent way.

1.6 Automated Data Mining

Advances in data collection methods, storage and processing technology provide a unique challenge and opportunity for automated data exploration techniques. Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools. Automated data mining and modeling software gives marketing managers a tool to perform analyses that otherwise would need to be handled by a highly trained researcher. There is a need for new generations of tools for automated data mining and knowledge discovery.
1.7 Motivation of the thesis

Data mining is a technique that discovers previously unknown relationships in data. Although data mining is a valuable technology for many application domains, the full potential of data mining is still under exploited. One of the reasons data mining is under-exploited is because of the existing framework of current data-mining systems [HAK+, 02]. Most current systems rely heavily on user’s interaction for inputs. The user must specify which data-mining algorithm to use and all the parameters associated with the algorithm. The user requires a great level of expertise in setting these parameters. Sometimes, the user might even need to define the goal, or what the user is looking for in the data. After the system applies a data-mining algorithm on the data of interest and returns the results, the user must do more work to decide which results are useful or interesting. For most of the organizations the required human resource is too expensive. Even if the required human resource is available, the person might still not be able to efficiently tackle the task on hand, due to the number of unknowns in the data to be mined. Therefore, the need arises to develop highly automated, scalable, integrated, and reliable data-mining systems and tools. A data mining framework is required to perform discovery on data automatically, without asking the user too many questions. So that, any non-expert user with little or no knowledge about the domain will be able to mine the data.
1.8 Problem Statement

To investigate and propose an approach for an automated data mining in object oriented data with the help of software agents. The goal is to empower non-expert users to achieve reasonable results with minimum effort. The proposed approach provides an interface to the non-experts and also hides the mining concepts away from the users thus helping to bridge the conceptual gap usually associated with data mining.

Automated data mining in object oriented data requires many different processing stages in the research work. The following steps are usually required for implementing automated data mining methodologies for an object oriented data.

i. Identifying the properties of object oriented data.

ii. Vertically partitioning the object oriented data.

iii. Identification, selection, and ranking of the attributes.

iv. Selection of appropriate data mining clustering algorithms.

The proposed research work uses the software agents in the following steps:

- Used to identify the properties of object oriented data and vertically partition the object oriented database.

- Ranking of attributes are automated.

- Used to select the appropriate Clustering Techniques to minimize the user interaction.
1.9 Need for the Proposed System

Data mining is a difficult and laborious activity that requires a great deal of expertise for obtaining quality results. Object oriented database systems are popular and influential in advanced database applications, it is important to investigate the mechanisms for knowledge discovery in object oriented data. The concept of domain knowledge is very important in data mining. The user needs to have enough prior domain knowledge in order to define the objective and specify the correct mining algorithm and its parameters. Then only the user gets the correct knowledge from the data mining system. It will take more time for the end-users to choose the correct mining algorithm and the user must have enough domain knowledge. Therefore, the need arises to empower non-expert users to achieve reasonable results with minimum effort. The approach provides an interface to the non-experts and also hides the mining concepts away from the users thus helping to bridge the conceptual gap usually associated with data mining. The new proposed framework deals with developing an automated data mining system for an Object Oriented data using software agents.
1.10 Thesis Organization

The remaining chapters of the thesis are organized as follows:

Chapter 2 gives a brief review on the literature available in automated data mining. The literature review shows a picture of how the vertical partitioning of object oriented databases, ranking and automated clustering looked prior to the work proposed in this thesis.

Chapter 3 introduces the new framework for an automated data mining system and the role of software agents adopted in the system. This chapter explains the different agents used in the research work.

Chapter 4 provides the methodology and implementation of the automated data mining system. This chapter also presents the various algorithms used and the database used in the research work.

Chapter 5 presents the results and discussions in the given context of an automated data mining system for an Object Oriented Data using Software agents. The results are analyzed from the student data.

Chapter 6 sums up the main ideas of this thesis and gives the conclusion and main directions for future research.