CHAPTER 2

REVIEW OF LITERATURE

This Chapter discusses the related work relevant to the thesis. In order to carry out the research work specified in the problem statement given in Chapter 1.8, the review of appropriate literature was carried out under the following domains.

- Automated Data Mining
- Vertical partitioning in Object Oriented Databases
- Ranking of Attributes
- Automatic Clustering

2.1 Automated Data Mining

In the age of digital information, the problem of data overload looms ominously ahead. Large databases of digital information are ubiquitous. Data mining automates the process of finding predictive information in large databases. Multi gigabyte databases with millions of records and large number of fields (attributes and variables) are common place. Although, it is a challenge to provide a high- performance, rapid- response environment that also assists users in the proper selection and matching of appropriate algorithms to achieve their goals. There needs to be more emphasis on human- computer interaction with the aim of supporting both expert and novice users. Automated data mining and modeling software gives managers a tool to perform analyses that otherwise would need to be handled by a highly trained researcher. Automated data mining
methodologies is not to provide more accurate results but strives to empower non-expert users to achieve reasonable results with minimum effort. By automating the data mining tools one can sweep through databases and discover previously unknown patterns. Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed.

Berry and Linoff compare automated Data mining to taking a picture with an automatic camera. The quality of results is not always as good as what can be achieved by an expert; however the ease of use empowers non-expert users to achieve reasonable results with minimum effort [BL, 99]. The camera can relieve the photographer from having to set the shutter speed, aperture and other settings every time a picture is taken. This makes the process easier for expert photographers and makes better photography accessible to people who are not experts. But this is still automating only a small part of the process of producing a photograph. Choosing the subject, perspective and lighting, getting to the right place at the right time, printing and mounting, and many other aspects are all important in producing a good photograph. This research work tries to develop highly automated, scalable, integrated, and reliable automated data mining system. This framework requires data mining systems to perform mining on data automatically, without asking the user too many questions.
Automating a data mining system is not an easy task. David mentioned that there are also many parts of the process that cannot be automated, including choosing a methodology to match a business problem, selecting a data set, quality checking and preparing the data for analysis, choosing among the available options within the analysis process, and interpreting and presenting the results [DSC, 02]. The author also suggested that an analysis could be automated when: The data being used is from a familiar source, the analysis has been used before in the same context, the variables included have been used before in the same type of analysis, and the results of the analysis will be interpreted and used in an established manner.

Many current KDD methods and tools are not truly interactive and do not easily incorporate prior knowledge about a problem except in simple ways. So there is a need for new generations of tools for automated data mining and knowledge discovery. Automated approach has developed in many areas of data mining.

Given below are the few related work done in the area of automated data mining during the recent past years.

Marcos, Peter and Boriana proposed a new approach to the design of data mining applications targeted business intelligence communities [MPB, 05]. Process automation refers to the specification of potentially complex data mining methodologies that can deliver the desired results with minimum intervention. This approach uses a data-centric focus and automated methodologies to make
data mining accessible to non-experts. In this approach automation of the data mining process requires many different processing stages in the application design. The following steps are usually required for implementing automated data mining methodologies: computation of statistics, sampling, attribute data type identification, attribute selection, algorithm selection, data transformation, model selection and quality assessment and output generation. This approach uses a data-centric focus and automated methodologies to make data mining accessible to non-experts. The primary benefits of this approach are good usability and performance. The research work is implemented as Oracle Predictive Analytics (OPA) in the Oracle Database 10g Release 2 Enterprise Edition with the data mining option in the form of PL/SQL package named DBMS_PREDICTIVE_ANALYTICS.

This approach targets only database users. The implementation described in this research work utilizes PL/SQL stored procedures and addresses the complete beginning to end process from input data to producing results containing predictions or explanations for specific attributes. This research work automatically determines problem type (classification or regression) only and not clustering.

Saravanan and Vivekanandan proposed an automated data mining system which compasses familiar data mining algorithms [SV, 04]. This research work deals with developing an automated data mining system which encompasses the familiar data mining algorithms. According to that the system will automatically
select the appropriate data mining technique and select the necessary field needed from the database at the appropriate time without expecting the users to specify the specific techniques and the parameters. A new framework is proposed for incorporating intelligent agents with automated data mining. One of the major goals of this system is to give the control to the computer for learning automatically by using intelligent agents for the exploratory data mining. Although this research paper discusses about automated data mining system it emphasis is mainly on association and classification algorithms. This research paper does not deal with automatic clustering of object oriented data.

Fayyad et al in his research work reviews the automated search focused mainly on automated methods for extracting patterns or models from data [FSS, 96]. Although this approach is consistent with the definition, it does not necessarily represent what other communities might refer to as data mining. For example, some use the term to designate any manual search of the data or search assisted by queries to a database management system or to refer to humans visualizing patterns in data. In other communities, it was used to refer the automated correlation of data from transactions or the automated generation of transaction reports.

In the research paper there are no established criteria for deciding which methods to use in which circumstances, and many of the approaches are based on crude heuristic approximations to avoid the expensive search required to find optimal, or even good, solutions. Hence, the reader should be careful when
confronted with overstated claims about the great ability of a system to mine useful information from large (or even small) databases. This approach chooses to focus only on methods that contain certain degrees of search autonomy.

James Malone, Ken McGarry, Chris Bowerman present an intelligent data mining architecture that incorporates both data-driven and goal-driven strategies and is able to accommodate the spatial and temporal elements of the dataset under analysis [JKC,06]. An Intelligent Data Mining Architecture was proposed which is able to automatically classify interesting proteins with a low number of false positives and false negatives. The architecture uses a combination of goal-driven (expert heuristics) and data-driven (data mining) elements to perform the proteomics data analysis. Following the data mining stage, the results of this process are then used to train and test a neural network classifier. This offers the advantage of being able to classify new, previously unseen data following successful training and hence introduce a level of automation to the process.

The technique selected is that of the Back Propagation, Multi-Layer Perceptron (MLP) neural network. This neural network is a supervised learning approach which involves training the network using both the inputs and the required outputs. Results from the data mining process technique can also be visualized using a simple line graph plot, which also aids our interpretation. This research goes some way to addressing the processing bottleneck that exists within post-experimental 2-DE gel data analysis by providing a technique that automatically extracts potentially interesting proteins from within the datasets.
Using a data mining technique to detect variance within the data before classification offers performance advantages over other statistical variance techniques in the order of between 16 and 46%. In this research an Intelligent Data Mining Architecture was proposed which is able to automatically classify interesting proteins. Since the technique involves the use of a supervised neural network, normal considerations of suitability apply, i.e. that empirical data must be available in order to train and test the network’s ability and that little human effort is needed to learn and classify correctly. This research paper addresses only classification technique in data mining but not deal with clustering of Object Oriented data. Also it needs little human effort to learn and classify the interesting proteins.

Themis et al in their research work proposed a methodology for the automated creation of fuzzy expert systems, applied in ischaemic and arrhythmic beat classification [TMC⁺,07]. The methodology automatically creates a fuzzy expert system from an initial training dataset. The approach consists of three stages: (a) extraction of a crisp set of rules from a decision tree induced from the training dataset, (b) transformation of the crisp set of rules into a fuzzy model and (c) optimization of the fuzzy model's parameters using global optimization. Specifically, a set of rules is extracted from a decision tree, developed from a training set. In the second stage, the set of crisp rules is fuzzified, resulting into a fuzzy model. Finally, all the parameters entering the
fuzzy model are tuned with respect to the classification accuracy of the fuzzy model, using global optimization.

The fuzzy model with the optimized parameters composes the final Fuzzy Expert System (FES). The generated FESs are able to provide interpretation for their decisions since they are based on sets of rules. The methodology automatically generates a FES, using an initial annotated dataset. In order to extract an initial set of rules from an annotated dataset, a rule mining technique must be employed. Here, decision trees are used, however any rule mining technique could be employed. The methodology has been evaluated in the detection of ischaemic cardiac beats in ECG recordings using data from the ESC ST-T database. Also, it has been evaluated in arrhythmic beat classification, using data from the MIT-BIH arrhythmia database. In both cases high classification results were obtained; the accuracy (Acc) is 92% and 96% for the ischaemic and arrhythmic FES, respectively. It is generic and thus it can be applied to any classification domain; given an initial annotated dataset, it can automatically generate a FES. This FES is based on a set of fuzzy rules and thus it is able to provide interpretation for its decisions. This research work automatically creates a fuzzy expert system from an initial training dataset.

This methodology is innovative since it combines data mining techniques with fuzzy modeling and introduces several novelties. The limitation is the requirement of a representative training set in order to extract reliable rules and
thus create a reliable fuzzy model. This work focuses only on decision rules for classification and not clustering techniques.

Cecil chua, Roger H.L. Chiang and Ee- Peng Lim present a research paper which helps the analysts to extract and validate statistical results to facilitate data mining [CCE, 00]. It describes an integrated data mining system called the Linear Correlation Discovery System (LCDS) which is used on a sample data set, and extracts relevant statistical output such as include R2 (regression), η2 (ANOVA) and Goodman and Kruskal’s λ (chi-square) for effective mining of data. LCDS implemented in Visual Basic using Microsoft Access 97 and SPSS v.8.0. Runs on any PC supporting Windows 95/NT. A new validation algorithm based on measuring the consistency of mining results was described. In this research work an integrated data mining system was developed that can select appropriate measures of association for a data set, execute the association measurement function, extract results from the statistical output and validate them. This research work addresses only measures of association, appropriate association rules only when attributes are ordinal, categorical, and interval respectively.

Moustafa et al., in their approach describes an approach which is based on developing and using an improved automatic feature selection method in conjunction with traditional classifiers [MYH, 01]. The feature selection method used is based on capturing frequently occurring keyword combinations (or motifs) within short segments of the text of a document and has proved to
produce more accurate classification results than approaches relying solely on using keyword-based features. The task addressed in this paper is that of developing a system to automatically curate a database of scientific papers by analyzing a training data set of past human curation decisions. Sub-tasks 1 and 2 of this task (providing a ranking of the relevance of the papers and deciding whether a paper should be curated or not) can be handled directly using a document categorization framework.

With a little preprocessing, sub-task 3 (identifying whether a particular item mentioned in a paper is related to a given concept) can be easily converted into a categorization question. The first attempt was to experiment with a simple traditional automatic information retrieval approach. The second and third attempts were based on using domain knowledge to choose only relevant keywords as a basis for constructing the feature vectors we decided to use an approach that captures the association between words appearing in each document. This is based on automatically identifying frequently co-occurring localized word patterns or motifs. By restricting the search for these patterns to localized parts of each document (e.g. a sentence or neighboring sentences) this approach models the associations between these words and generates classifiers based on these associations. The output patterns are defined using regular expressions on words automatically extracted from the documents.

Marcos, Boriana and Peter paper on, "Data–Centric Automated Data Mining" developed a data mining methodology for business intelligence
communities [MPB, 05]. This approach uses a data centric focus and automated methodologies to make data mining accessible to non-experts. Automation of the data mining process requires including many different processing stages in the application design. The following steps are used in automated data mining methodologies: Computation of statistics, Sampling, Attribute data type identification, Attribute selection, Algorithm selection, Data transformation, Model selection and quality assessment and output generation. The approach uses Two Applications:

- Oracle Predictive Analytics feature of Oracle10g Release 2
- Oracle Spreadsheet Add-in for Predictive Analytics

The approach described in Section 2 is currently implemented as Oracle Predictive Analytics (OPA) in the Oracle Database 10g Release 2 Enterprise Edition with the Data Mining option in the form of a PL/SQL package named DBMS_PREDICTIVE_ANALYTICS. This research work targets database users. The implementation described here utilizes PL/SQL stored procedures.

The following concepts are considered for this research work from the paper discussed above. They are a ranking procedure for attribute selection and algorithm selection automatically which helps the user to achieve reasonable results with minimum effort. Automated methodologies are used which helps to produce more accurate results with less human intervention. In this research work, software agents are used in automating the data mining system from user input till the knowledge is mined from the object oriented data.
Ranjit Bose, Vijayan Sugumaran on their paper titled "IDM: An Intelligent Software Agent Based Data Mining Environment", designed and developed an intelligent software agent based data mining environment called IDM [RV, 98]. In their research IDM data mining environment is designed using the following agents: a) user interface agent b) IDM coordinator agent c) data mining agent d) data-set agent and e) report/visualization agent. IDM agents are autonomous and independent and may reside in a single machine or in a distributed environment where they can communicate and cooperate with each other. Managers from different business functions and domains who have access to IDM will be able to make fact-based and timely decisions. IDM’s design is based on object-orientation and is implemented using “JATLite” (Java Agent Template, Lite) from Stanford University. These systems help make critical business decisions faster or with a greater degree of confidence. To remain competitive, managers are increasingly depending on intelligent systems for accessing, analyzing, summarizing, and interpreting information from large and multiple data sources.

From the above paper the concept occurred in developing and implementing a platform for data mining system for an object oriented data using Software Agents. In this research work multiple agents are used that interact and negotiate with each other. Four types of agents namely user interface agent, partitioning agent, ranking agent and data mining agent. These
agents are detailed in chapter 3. Each agent is intelligent in its own field and interacts with each other or behaves autonomously.

Saravanan in his dissertation titled "Design and Development of a Unified Data Mining System using Intelligent Agents: An Automated Approach", designed and developed a unified data mining system using intelligent agents. According to the data mining system the user specifies the objectives for a given data warehouse and the developed data mining system will take care of choosing the necessary parameters using intelligent agents. The system uses a relational data. If the user is using the system frequently the previous results also will be prompted if user enters into the system next time. The user can perform different types of data in much shorter time because the new system selects the algorithm and the parameter automatically. Because of the use of intelligent agents, selecting the appropriate parameter, for the given problem domain will be done by the system itself.

The following are the concepts considered from the above discussion. An agent is used to perform the mining process using the automated approach for a relational database. In this research work software agents are used to create an automated data mining system for an object oriented data.

Ayse Yasemin Seydilm in his research paper "Intelligent Agents: A Data Mining Perspective", discusses the agent paradigm along with the main applications and the use of this technology in data mining [Ays, 99]. In his paper he said the inherent parallelism and complexity of the classification and
discovering patterns from large amounts of data can be delegated to intelligent software agents. In the research paper an intelligent agent can use domain knowledge with embedded simple rules and using the training data it can learn and reduce the need for domain experts. In the interpretation of what is learned, an agent can go through the rules and facts generated and identify items that can possibly contain valuable information.

The following concepts were from the above research paper the idea of using the intelligent agents to automate the individual tasks. Agent can be especially used in learning parameters and also agents are implemented for classification, clustering, summarization and generalization which have learning nature and rule generation.

Although many researches were carried out related to automated data mining, there is no discussion on automated data mining system for an object oriented data using software agents. All the above related work although it is automated, it needs human assistance and enough domain knowledge to run the data mining system.

In this research paper software agents are used in automating the individual tasks such as analyzing, partitioning, ranking the attributes and in the selection of appropriate clustering algorithms for mining the object oriented data.

In the proposed research work, an automated data mining system is developed for an object oriented data. Also with the help of the software agents a new user will be able to interact with the data mining system without much
data mining technical knowledge. The system will automatically select the appropriate clustering algorithm and select the necessary field needed from the object oriented data with less human intervention based on the user query.

### 2.2 Vertical partitioning in Object Oriented Databases

Object oriented database systems are popular and influential in advanced database applications. It provides rich modeling features like, encapsulation of methods, inheritance, object identity, arbitrary data types and complex objects. 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.

Vertical Partitioning is a technique for facilitating efficient way of executing next generation database applications by reducing irrelevant instance variable (attribute) access. The aim of vertical partitioning is to partition a relation into fragments. Although this problem has been addressed in relational database systems, there is very little work on class partitioning for object oriented database (OODB) systems. Related works of vertical partitioning in object oriented database are addressed below.

Hoffer and Severance in their early work developed the concept of attribute affinity [HS, 75]. This paper presents a means for overcoming this difficulty in the design of data base records. A metric with which to measure the similarity of usage among data items is developed and used by a clustering algorithm to reduce the space of alternative designs to a point where solution is
economically feasible. They defined an algorithm in which attributes of an object were permuted in such a way that attributes with "high affinity" were clustered together. Affinity among attributes expresses the extent to which they are used together in processing. Attribute affinity means the degree to which pairs of attributes are accessed together, and hence, grouping attributes with high affinity will reduce access costs. In this paper clustering of attributes on the basis of their affinity was achieved by applying the Bond Energy Algorithm (BEA) developed by McCormick.

McCormick et al developed the Bond Energy Algorithm (BEA). In which clustering of attributes is carried out on the basis of their affinity achieved by applying the BEA algorithm [MST, 72]. It operates upon a raw input object-object or object-attribute data array by permuting its rows and columns in order to find informative variable groups and their interrelations. The Bond Energy Algorithm is a top-down approach. It begins with a group composed of all attributes and splits them. Their attribute similarity function uses the notion of an attribute being "required for processing", or "required only if the record is selected", or finally "not required". This paper describes the algorithm and illustrates by several examples its use for both problem decomposition and data reorganization.

A two-phase affinity-based vertical partitioning algorithm was developed by Navathe and Ceri [NCG+, 84]. They extended the work of Hoffer and
Severance [HS, 75]. Vertical partitioning is applied in three contexts: a database stored on devices of a single type, a database stored in different memory levels, and a distributed database. An Iterative binary partitioning method is used based on first clustering the attributes and then applying empirical objective functions to perform the partitioning. In this paper the algorithms designed for an automatic selection of vertical fragments, which substitutes for the designer's subjective judgment.

Navathe and Ra in their research paper proposed an affinity based graph theoretic algorithm for vertical partitioning [NR, 89]. This algorithm starts from the attribute affinity matrix by considering it as a complete graph. Then, forming a linearly connected spanning tree, it generates all meaningful fragments simultaneously by considering a cycle as a fragment. It shows its computational superiority. It provides a cleaner alternative without arbitrary objective functions and provides an improvement on vertical partitioning. The limitation of these affinity based methods is that they can find a feasible partition, but partitioning by attribute affinity does not imply that the partitioning scheme found can reduce the total number of disk accesses.

Chu and Leong developed a vertical partitioning method that optimizes the number of disk accesses based on clustering of attributes accessed by transactions [CL, 93]. This approach allows the optimization of the partitioning based on a selected set of important transactions. An optimal binary partitioning (OBP) algorithm based on the branch and bound method is presented, with the
The worst case complexity of $O(2^{\sup n})$, where $n$ is the number of transactions. To handle systems with a large number of transactions, an algorithm Bpi with complexity varying from $O(n)$ to $O(2^{\sup n})$ is also developed. The experimental results reveal that the performance of vertical partitioning is sensitive to the skewness of transaction accesses. Further, Bpi converges rather rapidly to OBP. Both OBP and Bpi yield results comparable with that of global optimum obtained from an exhaustive search. While partitioning, insistence is given only to transactions because of semantic meanings than the attributes.

Karlapalem and Li in their research work presented three major partitioning schemes for Object Oriented Databases, namely, vertical class partitioning, path partitioning and horizontal class partitioning [KL, 95]. The resultant class fragments can be represented and implemented as classes in an OODB. In [KL, 90] designing the partitioning schemes, a fundamental issue considered was the role of "methods", since methods are the basic primitives for accessing the objects. Based on the various types of methods, and the values returned by the methods, various combinations of the partitioning schemes can be developed resulting in method induced partitioning schemes.

Path partitioning was first introduced in OODB [KL, 95], [KLV, 96] [KL, 00]. It is required to access the complete composite objects in many applications. Path partitioning is a concept describing the clustering of all the objects forming a composite object into a partition. A path partition consists of grouping the objects of all the domain classes that correspond to all the instance
variables in the subtree rooted at the composite object. The resultant path fragments thus serve as structural indices for the composite objects, with each node of the structural index pointing to the objects of the domain class of the component object. The structural index thus contains the reference to all the component objects of a composite object, eliminating the need to traverse the class composition hierarchy. The nodes of the structural index at each level of hierarchy point to a set of component objects (which themselves can be composite objects). Therefore, the structural index can be used to directly access a composite object which itself may also be a component object of another high level composite object.

Jin and Myoung propose an adaptable vertical partitioning method that can support both best-fit and n-way vertical partitioning [JM, 04]. Adaptable vertical partitioning method supports n-way partitioning as well as best-fit vertical partitioning based on the cost model. Bottom up approach is used in the AVP method. In this research work vertical partitioning proceeds into two directions: one is to generate an overall optimal partitioning that minimizes the processing cost of data queries running on the fragments, which is called best-fit vertical partitioning and the other is to generate a specific number of fragments required by the user, which is generally called n-way vertical partitioning. The limitation is that the adaptable vertical partitioning method is used in distributed systems only but not in the centralized Object Oriented database systems.
Jiawei Han, Shojiro Nishio, Hiroyuki Kawano, Wei Wang on their paper titled "Generalization-based data mining in object-oriented databases using an object cube model" investigated in 3 aspects: Generalization of complex objects, class-based generalization, Extraction of different kinds of rules. An object cube model is proposed for class-based generalization, on-line analytical processing, and data mining. A data retrieval process and an object cube based generalization is performed to generalize and compress the set of relevant data into a compact generalized object cube [JSH, 98]. Different kinds of knowledge rules can be extracted.

From the above paper the following points are considered for this research work. First the identification of the properties of object oriented data has to be identified. And the philosophy of the mining of other kinds of knowledge, interesting knowledge discovered from object-oriented databases, and the application of discovered knowledge from object oriented databases.

In this research work software agents are used which help in identifying the properties of object oriented data. The user interface agent used in this research work helps in monitoring the actions taken by the user and cooperates with the user in accomplishing the task in the application. The user interface agent used in this research work analyzes the object oriented data set and classifies them.

In all the above related works, no work has been done using software agents in vertically partitioning the attributes. In this research work vertical
partitioning is used to reduce irrelevant instance variable (attribute) access. In this proposed research work vertical partitioning is carried out based on the attributes that are closely related to each other are partitioned. Software agents are used to find the closely related attributes. A software agent called partitioning agent is used in this research work for vertical partitioning. A JACK™ Intelligent Agents (JACK) is an Agent Oriented development environment built on top of and integrated with the Java programming language is used in this research work to develop the agents which is explained in detailed in chapter 4.

2.3 Automatic Clustering

As data collection increases at an accelerating rate with the advances of computers and networking technology, analyzing the data (data mining) becomes very important. Clustering is one of the common techniques used in data mining. An ideal case is to group related data into the same cluster and unrelated data into different clusters. Even though, lot of work has been published on clustering of data in the last few decades, little has been done about automating the process. Clustering techniques can be manual or automatic. A manual clustering technique has to be started by the user. This in turn implies that the user should know all the details regarding the clustering. Rather, this research work advocate using an intelligent software agent to automate clustering that could trigger itself.

Clustering has modern applications in numerous domains, such as biomedical data, software engineering, economics, 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.

Several researchers have also worked on automating clustering. Related
works of automatic clustering are addressed below.

In Thomas Brinkhoff research work, there is a cluster manager, which re-
clusters spatial objects dynamically [Br, 01]. The reorganization is performed
using the pages kept in main memory by the buffer manager. Therefore, no
additional disk accesses are required for non-static spatial databases. For
deciding which spatial objects should be stored together, the requests for the
objects are recorded and analyzed. Because of the cluster manager no additional
human administration is required. The cluster manager cooperates with the
buffer manager. The presented cluster manager reorganizes the pages which
store the spatial objects. In this research work only simple algorithms, which are
easy to implement, have been proposed for computing the objects to be shifted.
However, more sophisticated and more efficient algorithms for computing
clusters have been proposed in literature.

Darmont et al presented three dynamic clustering techniques for Object-
Oriented Databases (OODBs) [DFR⁺, 00]. The first two, Dynamic, Statistical &
Tunable Clustering (DSTC) and StatClust, exploit both comprehensive usage
statistics and the inter-object reference graph. They are quite elaborate.
However, they are also complex to implement and induce a high overhead. The
third clustering technique, called Detection & Reclustering of Objects (DRO), is based on the same principles, but is much simpler to implement. These three clustering algorithm have been implemented in the Texas persistent object store and compared in terms of clustering efficiency (i.e., overall performance increase) and overhead using the Object Clustering Benchmark (OCB). The results obtained showed that DRO induced a lighter overhead while still achieving better overall performance. The dynamic clustering technique is perfectly viable in an OODBMS and could achieve significant gains in performance. The limitation is lack of automation in the tuning process.

Andreas Christl et al in their research paper present a new approach, in which clustering techniques are applied to support the user in the mapping activity [CKS, 07]. The result is a semi-automated mapping technique that accommodates the automatic clustering of the source model with the user's hypothesized knowledge about the system's architecture. In their research three case studies in which the semi-automated mapping technique, called HuGMe, has been applied successfully to extend a partial map of real-world software applications. In addition, the results of another case study from an earlier publication are summarized, which lead to comparable results. It also evaluated the extended versions of two automatic software clustering techniques, namely, MQAttract and CountAttract, with oracle mappings. Both clustering techniques were able to achieve a mapping quality where more than 90% of the automatic mapping decisions turned out to be correct. Moreover, the experiments indicate
that the attraction function (CountAttract') based on local coupling and cohesion is more suitable for semi-automated mapping than the approach MQAttract' based on a global assessment of coupling and cohesion. This approach is problematic when the developers of the system did not follow the principle of low coupling and high cohesion.

Sylvain and Le in their research paper discuss an efficient algorithm for attribute clustering that dynamically and automatically generate attribute clusters based on closed item sets mined from the attributes sets found in the queries running against the database [SL, 06].

Zaman et al in their research work developed a technique for auto-indexing using clustering. [ZSG, 04] Considering the wide deployment of databases and its size, particularly in data warehouses, it is important to automate the physical design so that the task of the database administrator is minimized. An important part of physical database design is index selection. An auto-index selection tool capable of analyzing large amounts of data and suggesting a good set of indexes for a database is the goal of auto-administration. Clustering is a data mining technique with broad appeal and usefulness in exploratory data analysis. This idea provides a motivation to apply clustering techniques to obtain good indexes for a workload in the database without any automation.
William J. McIver, Jr. and Roger King in their research work presented possible architecture for performing complex object reclustering in an on-line manner that is adaptive to changing usage patterns [MR, 94]. The architecture involves the decomposition of a clustering method into concurrently operating components that each handle one of the fundamental tasks involved in reclustering, namely statistics collection, cluster analysis, and reorganization. We present results of an experiment performed to evaluate its behavior. These results show that the average miss rate for object accesses can be effectively reduced using a combination of rules that we have developed for deciding when cluster analyses and reorganizations should be performed.

All these techniques are not fully automated and require lots of parameters and users' hints. Therefore, in this research work, an automatic clustering of object oriented data is carried out by using software agents. Software agents are special types of software applications, have become a very popular paradigm in computing in recent years. Software agents are an emerging technology that is making computer systems easier to use by allowing people to delegate work back to the computer. Software agents help in choosing the appropriate algorithm without human intervention automatically. A set of clustering algorithms for different attributes numeric, categorical and for object oriented data is detailed in Chapter 4. From which the software agents chooses the appropriate clustering algorithm from the different algorithms present in the
data mining system automatically i.e. with less human intervention based on the user’s parameter (attributes) given.

2.4 Ranking

Gathering information about a dataset and learning the relationships among the attributes are important steps in data mining. Attributes play a crucial role in all data mining algorithms. Very often rankings are provided from best to worst relative to particular attributes. The attributes in a dataset are ranked in decreasing order. The attribute most relevant for the target attribute’s value is at the top of the ranked list. An attribute is selected from the top of the ranked list to be used as input into data mining clustering algorithms. Attributes in a dataset are ranked in terms of their relevance of the task. Attribute ranking was seldom discussed explicitly because it is often considered as part of attribute selection. Extracting ranking functions has been extensively investigated in areas outside database research.

Hall and Holmes discuss the algorithms used in the benchmark study of formed ranked lists on which subsequent attribute selection was based [HH, 00]. This paper presents a benchmark comparison of several attribute selection methods. All the methods produce an attribute ranking, a useful devise for isolating the individual merit of an attribute.
Attribute ranking was also used explicitly in [Hon, 97]. Se June Hong et al in the research paper present a new approach in which assigning merits to features by finding each feature's "Obligation" to the class discrimination in the context of other features. The merits are then used to rank the features, select a feature subset, and discretize the numeric variables [Tal, 99]. The limitation is more efficient refining technique will greatly improve the run time. The absolute magnitude of the merits should be understood better.

Vagelis and Yannis propose ranked queries return the top objects of a database according to a preference function [VY, 04]. A core algorithm was presented and evaluated (experimentally and theoretically) that answers ranked queries in an efficient pipelined manner using materialized ranked views. An algorithm that selects a near optimal set of views under space constraints is presented in this paper.

Sanjay et al in their research paper discuss the challenges and investigate several approaches to enable ranking in databases [SSG⁺, 03]. Ranking algorithms have been implemented on a relational DBMS. In their paper they address the ranking of numeric and categorical attributes. An attempt is made to build a generic automated ranking infrastructure for SQL databases. This is consistent with the relational database management infrastructure with functionality necessity and useful data exploration. But the modest cost to solve
database ranking to reduce the burden of an application designer or user automatically and intelligently is unanswered.

Although, there are different ranking techniques available in the literature there is no specific ranking algorithms in ranking the object oriented attributes. In this research a ranking factor is used which is discussed in chapter 4.2. Also in this research a ranking agent is used which helps in ranking the attributes automatically.