TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

Data Mining [3] is a interdisciplinary study which integrates the results of the newest technologies such as Database technology, Artificial Intelligence、Machine Learning、Statistics、 Knowledge Engineering、Object-Oriented Method、Information Retrieval、High-Performance Computing and Data Visualization. With more than ten years' research, there are many new concepts and methods of data mining. Especially in recent years, some basic concepts and methods become clearer. And their research is further developing into a deeper direction. The reason that data mining is called one of the backbone technologies in the future information handling is that it, with a completely new concept, changes the manner with which people uses data. In the twentieth century, the database technology has got decisive harvest and been applied widely[4]. However, database technology, as a kind of basic form of information storage and management, still takes OLTP[5] (On-Line Transaction Processing) as core application, it lacks of supporting mechanism of higher function such as decision-making, analysis, and forecasting, and so on. As is known to everyone, with the expansion of the database capacity, especially the increasing popularity of the new type of data source such as Data Warehouse and Web, etc, the complex application such as OLAP （On-Line Analytic Processing）、Decision Support and Classification、Clustering is inevitable. Facing this challenge, data mining and knowledge discovery emerge and show powerful vitality. Data mining and knowledge discovery help data handling technology into a higher stage, which can not only query the past data, but also can find the potential associations among the past data. Then, it is possible to do higher level analysis in order to make ideal decision and forecast future developing trend.

1.1 topic background

With the development of the distributed database, the centralized data mining cannot meet the demand of the distributed data mining. Therefore, the research on the distributed data mining system has become the hot researching topic in the field of data mining[6].

Why the research on the data mining facing the distributed database has been paid so much attention? The main reasons are as follows[7]:

①The object of data mining is large-size data set. However, in the actual environment, most large-size database exists in distributed form. Therefore, it is necessary to put forward the structure of new distributed data mining system.

② Data mining system often needs data from database from different sites, which require the data mining system have the ability of distributed mining, and meanwhile, we should design new distributed data mining algorithm according to the features of distributed data mining.

The main requirements of the users to the data mining technology are as follows[7]:

① The enlarging of the enterprise size. Most of them have many branches and each branch forms its own body.

②Security and privacy are more and more important. In the market competition, the protection of the security and privacy of the data is more and more important to all enterprises.

③ Data quantity is larger and larger. The efficiency requirement of the association rule mining in the enterprises is higher and higher.

In the light of the above points, the traditional mining technology based on centralized database cannot meet the requirements, and gradually exposures its own disadvantages, mainly including the following three points[7]:

① To the application of many branch nodes, if concentrating all the data on one centre server to do data mining, network will have to be loaded too much transmission.

② Compared with mining respectively in each branch node, the efficiency of centralized mining is obviously low.

③ The centralized mining in which all concentrate on one place is disadvantageous for the protection of data privacy.

A set of good mining algorithms of the distributed association rule has the following features[8]:

—Accuracy: this is the basic requirement. That means the result of the algorithm must be accurate. The present main algorithms can meet this requirement in this aspect.

—High efficiency and privacy-preserving: it mainly refers to the executive efficiency. Because the algorithm is based on distributed database, the quantity of mutual transmitting data among networks determines the efficiency of the algorithm to a great extent. The less the quantity of the mutual transmitting data is, the higher the efficiency will be, and the security, privacy and the secrecy will increase greatly. The present main algorithms are trying to increase in this aspect, but not enough.

This topic is just based on such background, and it is focused on the shortages of the centralized algorithms and the main disadvantages to improve. First, it puts forward structure of the distributed data mining system based on multi-agent, and then the mining algorithm of the distributed association rule to make the data mining in the distributed environment increase in the aspects of both data privacy protection and operation efficiency.

1.2 The research contents

The distribution of data saving method bring challenges to traditional centralized data mining, and distributed data mining is required by circumstance.. the distribution of data and isomerism is the difficult part in distributed data mining. Centralized data mining algorithm can not suit the requirement of distributed data mining. How to design new algorithm or improve existing centralized data mining algorithm is also an urgent problem. This thesis will focus on these two points to do some research. The researching methods in this thesis are mainly in the following aspects:

□introduce in details about research on data mining, the structure and algorithm of distributed data mining in chapter two.

□introduce the concept of "agent" in chapter four, and bring about an architecture which is used in distributed data mining. This architecture is based on mass distributed saving transaction database in physics. Multi-agent technique is used here to enable the intelligence of the architecture.

□in chapter three, the theory of distributed mining association rules are discussed. And in chapter five, a new distributed association rules algorithm—RK-tree is introduce. This algorithm will make the system safer and more effective.

□the last chapter is the conclusion of the whole thesis. Also, further thinking and suggestions of distributed data mining system are raised by the author.

1.3 Organization form of the thesis

This thesis dissertates and analyzes the ideas of mining technology and the algorithms systematically, and focuses on the distributed data mining and the association rule mining algorithm. On the basis of these, the structure of the distributed data mining and a kind of new algorithm of distributed data mining are put forward. The contents of each chapter in this thesis are arranged as follows:

Chapter One: introduction

This chapter introduces purpose and meaning of this research, and put forward the core contents of this research. Meanwhile, the task of his thesis and the organization of the thesis are also stated.

Chapter two: Theory of Data Mining

① Introduce present research situation of data mining;

② Briefly introduce the distributed data mining technology;

Chapter three : General introduction of Association Rules

Focus on the basic concepts of the association rule mining algorithm, the basic ideas of the algorithm, and the general condition of the development. And then the ideas of Apriori algorithm are emphasized, and finally discusse several problems in the classification and mining of the association rule.

Chapter four :the distributed data mining system based on multi-agent

① introduce the distributed data mining system and agent technology;

② put forward a model of the distributed data mining system based on multi-agent ;

Chapter five: the mining algorithm of the distributing association rule

This chapter puts forward a kind of algorithm of the distributing association rule based on multi-agent, and analyze and conpae this algorithm. And examples to prove the advantage of the algorithm are also presented.

Chapter six: conclusion

this part will make a conclusion of the research in this thesis. And offer author's thinkings and suggestions on further research on distributed data mining system.

CHAPTER 2 Theory of Data Mining

2.1Data Mining

With quick development of database technology and the wide application of management system, more and more data are accumulated, behind which there is much hidden important information. People hope to analyze the information in higher level in order to better make use of these data. At present, the database system can realize high-efficient functions such as recording, correcting, statistics and inquiry, etc. However, it cannot discover existing association and rules among the data, and cannot forecast future developing trend according to present data. Lacking means of mining the knowledge hidden behind data leads to the phenomena of "data explosion with lack of knowledge". We use database management system to store data, the method of machine learning to analyze data, and mine knowledge hidden behind a great deal of data. The combination of these two promotes the appearance of knowledge discovery of database (KDD) [9].

Although data mining itself is a discipline, its origination can trace back to the early development of artificial intelligence in 1950s. During this period, mode identification and the development based on rule reasoning provide basic module, and data mining is based on these concepts. From then on, although we had not give a name to data mining, many technologies used today are extend from that time, and they are mainly used in science application. Figure 2.1 shows the development in the past forty years[10].

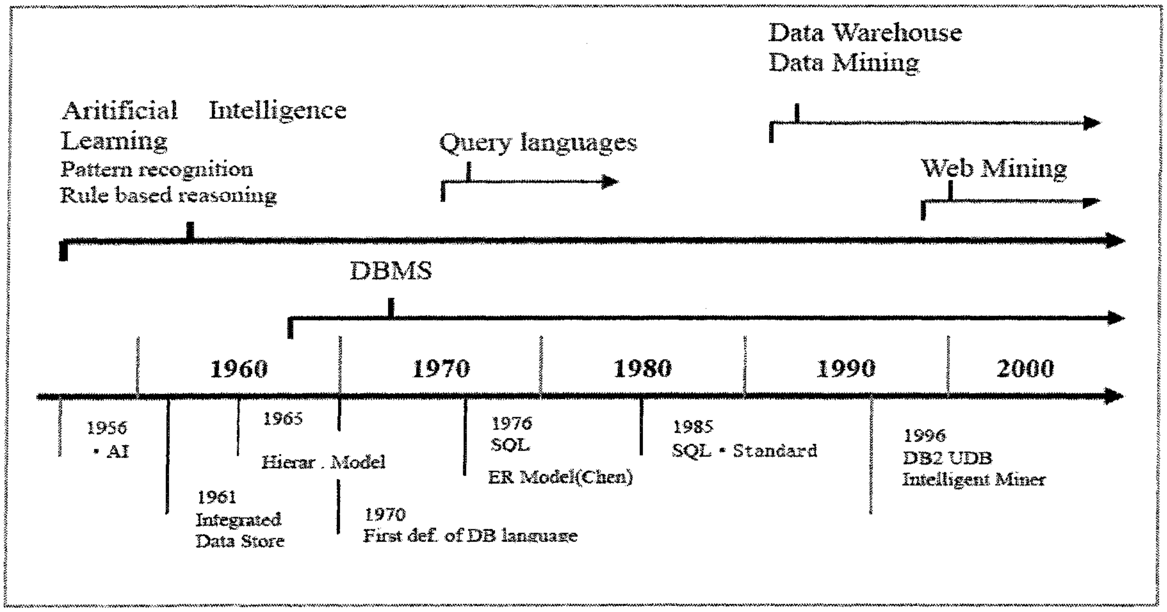

Figure2.1 History of DataMining Development

2.1.2 The structure of data mining system

A practical data mining system should have the following characteristics:

(1)the ability of quick respond. The system can send information back to customers even the operation time of system is very long;

(2)the ability to deal with large amount of data, that is, having better time complexity in realization method;

(3)friendly and alternating interface; data output can be realized in many ways.

(4)the ability to choose self-adaptation and suggest better parameters and models.

(5) Because customers are usually not experts, they don not know which model is suitable for the existing data.

Figure 2.2 is a typical structure of data mining system, the followings are discussions on the composition and function of each part[3]:

(1) Database, data warehouse and other information base: it is a information base of a database or a group of database, data warehouse and electronic table or other types. It can clear and integrate data on data.

(2) Server of database or data warehouse: according to users' requirement of data mining, server of database or data warehouse are responsible for extracting relevant data.

(3) Knowledge base: knowledge base is used to save knowledge needed in data mining. This knowledge will be used to guide the searching process of data mining, or to help evaluate the data mining results. The threshold value defined by

customers in mining algorithm is the simplest area knowledge.

(4) Data mining engine: it is the basic part of data mining system, which is composed of a group of functional modules, used for the analysis of characterization, relevance, classification and clustering, and evolution and variance analysis.

(5) The module of mode evaluation: usually, this component uses the degree of interest to measure, and interacts with data mining module in order to focus the search on the interesting modules. It may use the threshold of interest degree to filter the discovered modes. The module of mode evaluation can also integrate with mining module, which depends on the realization of all the used data mining methods.

(6) The interface of graphs and users: this module communicates between users and data mining system, allowing users to interact with system, pointing the inquiry or task of data mining, providing information, helping search focus, doing data mining with searching method according to the middle results of data mining. Besides, this component also allows users to browse database and data warehouse or data structure, to evaluate mining mode, and to visualize the mode with different forms.

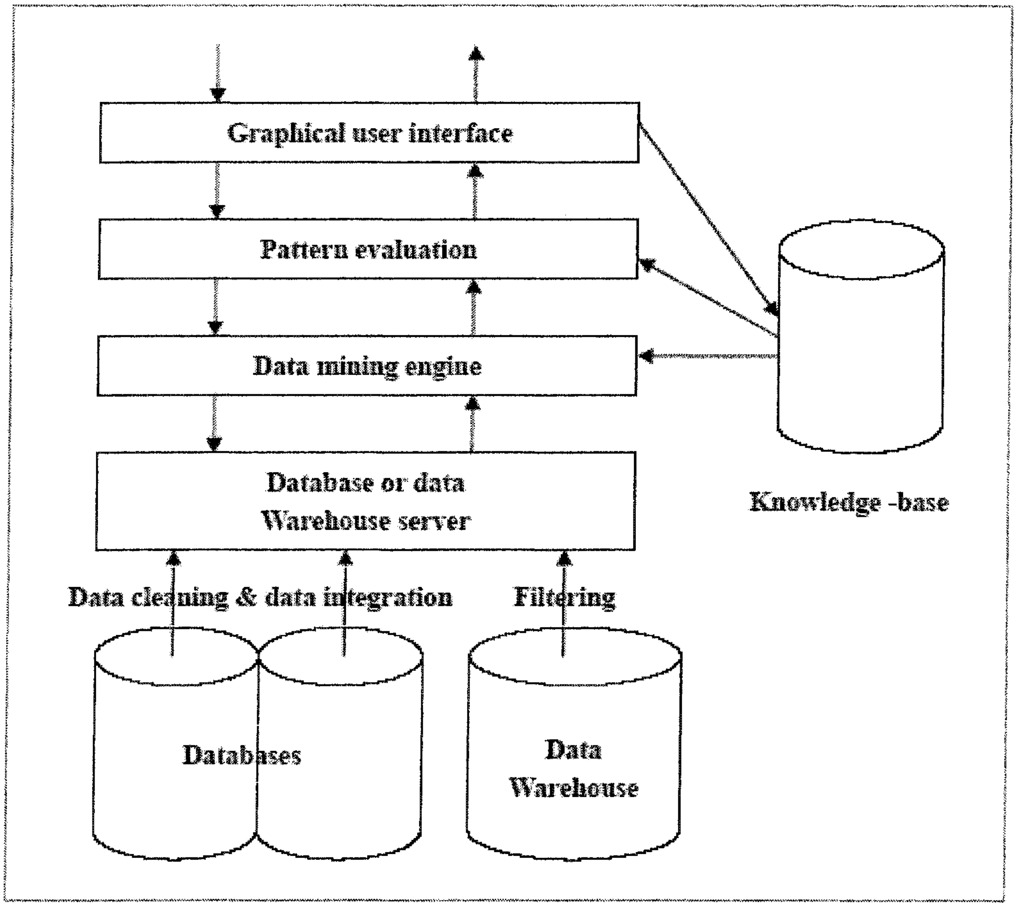

Figure2.2 The structure of data mining system

2.1.2 Functions of data mining

Data mining makes the proactive decision based on knowledge through forecasting future trend. The aim of data mining is to discover the hidden and meaningful knowledge from database. It mainly has the following five kinds of functions[11]:

①forecasting trend and behaviors automatically

Data mining looks for forecasting information automatically in the large-size database. Nowadays, the problems that need more manual analysis in former times can get conclusion more quickly and directly from data themselves. A typical example is market forecasting problem. Data mining uses the data about promotion in the past to look for the users who will give the largest investment in future. The other problems that can be forecasted include forecasting bankruptcy and identify the groups that will probably respond to given events.

②association analysis

Association analysis is a kind of important knowledge that can be discovered in database. If there is some regularity between values of two or more than two variables, it is called association. Association can be divided into simple association, time-sequence association, and cause-result association. The aim of association analysis is to find out the association web hidden in database. Sometimes we do not know the association function of data in database, even we know it, it is uncertain. So the rule of association analysis is reliable.

③clustering

The record in database can be divided into a series of meaningful itemsets, that is, clustering. Clustering strengthens people's understanding to the objective reality and it is the precondition of concept description and derivation analysis. Clustering technology mainly includes the method of traditional mode identification and mathematics taxonomy in the early 1980s, Michalski[12] put forward the concept of clustering technology, whose points are that when dividing objects it not only needs to consider the distance among objects, but also the types divided have the discretion of some kind of connotation, so that some sidedness of traditional technology can be avoided.

④concept description

Concept description is to describe the connotation of some type of objection, and summerize its relevant features. Concept discretion is divided into feature description and distinction description. The former one describes common features of some type of object while the latter one describes the distinction among different types of objects. The feature description generating one type only evolves the general feature of all objects in this type. There are many methods of generating distinction description, such as decision-making tree method, and genetic method, and so on.

⑤derivation testing

There are often some abnormal records of data in database. It is meaningful to test these derivations in database. Derivation includes much potential knowledge, such as the unusual examples, the special cases not meeting the rules, the derivation between the observing results and model forecasting value, the change of value with time, and so on. The basic method of derivation testing is to look for meaningful differences between observing results and the reference value.

2.1.3 Methods used in data mining

①rule summing[13]

It is to sum and extract valuable if-then rules through statistics method, such as association rule mining.

②the method of decision-making tree

It uses tree shape to stand for decision set. These decision sets will generate rules through the classification of data collection. The method of decision-making tree first uses information to look for the field with the greatest information in database and then establishes a node of decision-making tree. And then according to different values of fields, branches of the tree will be established; then in each itemset of branch, establish repeatedly the lower-layer node of the tree and branches, i.e. establishing decision-making tree. The most influential international decision-making method is the method of ID3 (Interaction Detection[14]) developed by Quinlan[15], whose typical application is classification data mining.

③artificial neural network[16]

This method mainly imitates the neuron structure of human being, and it is also a kind of non-linear forecasting model to study through training. It can finish many kinds of data mining tasks such as classification, clustering, and feature rules, and so on. Meanwhile, it is based on the MP[17] (McCulloch Pitts) model and HEBB[18] (Hebbian) study rule to establish three kinds of network models : forward feedback network, backward feedback network, and self-organization network .

④genetic algorithm

This is a kind of algorithm imitating the process of biology evolution, which is put forward by Holland [19] in the 1970s for the first time. It is an iterative process based on groups, and with the features of random and directional search. These processes include four kinds of typical operators: gene portfolio, crossover and mutation, and natural selection. Genetic algorithm acts on a group composed of many potential solutions of the problems, and each individual in the group is expressed by a code. Meanwhile, each individual needs to be given a suitable value according to the objective function of the problem in order to perform the ability of advantageous search of genetic algorithm.

⑤fuzzy technology[20]

Fuzzy technology is making use of the theory of fuzzy set to make fuzzy assessment, fuzzy decision-making, fuzzy mode identification and fuzzy clustering analysis. This kind of fuzziness exists objectively, and the higher the complexity of the system is, the stronger the fuzziness will be. On the basis of the traditional fuzzy theory and probability statistics, the cloud model shaped through putting forward the qualitative and quantitative and uncertain transferring model combines the fuzziness and randomness of the concepts together, which provides a kind of new method of concept and knowledge expression, qualitative and quantitative transferring and comprehension and decomposition for data mining.

⑥Rough Set method[21]

It is a kind of completely new data analysis method put forward by Poland logician Pawlak[22] . In recent years, it has been paid great attention to and widely applied in the fields of machine learning and KDD, and so on. This kind of rough set method is an effective researching method in uncertain and imprecise problems in information system, whose basic principle is based on the ideas of equal-price type. However, the elements of this kind of equal-price type are regarded as indistinctive. The basic method is first using the method of rough set approximation to discrete the attribute value of information system; and then dividing each attribute into equal-price type, and then making use of the equivalence to simplify the information system; and finally getting a minimum decision association in order to get rules conveniently.

⑦visualized technology[23]

It uses the form of visual graphics to display the information mode, the association or the trend of data to the decision-makers. In this way the decision-makers can analyze the data association interactively through the visualized technology. The visualized technology mainly includes three aspects of visualization of data, model, and process. Among them, data visualization mainly has histogram, box map and fall apart point chart; the concrete method of model visualization has something to do with the algorithm of data mining. For example, the algorithm of decision-making tree is expressed by tree shape while process visualization uses data flow to describe the knowledge discovery process.

Although data mining technologies mentioned above have their own features and applicable scope, the types of knowledge they discover are not the same. Among them, induction is often suitable for association rule, feature rule, sequence mode and the mining of discrete data; the method of decision-making and genetic algorithm and rough set method are often suitable for classification mode and structure; but neural network method can be used to realize many kinds of data mining such as classifying, clustering, feature rules; fuzzy technology is often used in mining association, fuzzy classification and fuzzy clustering rule.

2.1.4 the main application and the developing trend

The research of data mining is driven by application. Since it was born, it is featured with application function. Due to features of data mining itself, it has great application foreground in any fields such as finance, insurance, retail trade, medicine, manufacturing, transportation business, science and engineering research, and so on[24][25]. In finance, data mining can be used to analyze users' credit condition, as well as forecast the repayment condition of loan; in the field of biomedicine, data mining can be used to research DNA sequence; in the production manufacturing, it can be used in aspects of failure diagnosing, storage optimizing, manufacturing attempering, and so on. Due to the diversity of data form, the task of data mining, there are many challenging topics in the field of data mining. The high –effective and useful method of data mining, the design of data mining language, the establishment of interactive integrated data mining environment, all these problems are research and development personnel of data mining facing. The future focus and developing trend of data mining field can be expressed in the following aspects[2][26][27]:

(1) The standardization of data mining platform: the standardization of data mining language will provide convenience for the development of the systemization of the mining project, and is helpful for the mutual operation between each data

mining system and function module; it is also convenient for the training and using in the enterprises. Maybe it will be standardized and easy to use like SQL in the future.

(2) The visualization method in the process of data mining: the research on this aspect will make the process of knowledge discovery understood by the users visually, and it is convenient for intercommunion between people and machine.

(3) The stretchable method of data mining: most of the traditional data analyzing methods are based on memory. What data mining faces is large quantity of data. Therefore, it becomes a research direction that how to effectively and intersecting handle these large amounts of data. The complexity of a good algorithm of data mining should increase linearly with the growth of the number of data record and attribute.

(4) Web mining: at present, Internet has been the most immense and global information service centre. There is a great deal of information on Internet. The mining of Web contents, Web log and Web structure has been one of the most important hot topics in present and future data mining field.

(5) The new method of complex data type mining: at present, the mining with complex data type, such as the mining of geography space, multimedia, time series, and so on, has made some progress. But there is a long way to go for actual

application. Therefore, further research on this field is very important.

(6) Privacy protection and information security in the process of data mining: with the development of data mining, how to guarantee the privacy security and information security is a very important problem. So, it needs to do research in this field.

(7) The algorithm of distributed data mining and the platform of distributed data mining: with the development of network technology and the globalization of enterprises and organizations, data information of enterprises and organizations may distribute in different physical positions. In order to excavate useful knowledge from these data, it is challenging to do some distributed improvement in the original centralized algorithms and construct high-effective distributed data mining platform.

2. 2 Distributed Data Mining

The global distribution of the enterprises and various kinds of organizations causes that a great deal of data or information is stored in different geographic positions. quick development of network technology, internet technology and the increase of the computer performance makes it possible to analyze and handle these data, and excavate the valuable knowledge from them. And this has been the urgent need to these enterprises and organizations. But most traditional data mining systems are centralized. And at this time the distributed data mining system emerges

When data mining is operating in such environment: users, data, hardware resource and software resource needed in mining are distributed physically. We call it Distributed Data Mining[28][29] (DDM for short). It is a process that makes use of distributed computer technology and discovers knowledge from distributed database. Typically, this kind of environment has the features of heterogeneity data, many users, and large-size data quantity.

2. 2. 1 the basic principle of Distributed Data Mining (DDM) [28]

Distributed data mining is a new research field put forward recent years. Because it has tempting foreground, at present, there are considerable research personal devoting to the research on this field and having made some results. The two basic steps of the typical distributed data mining algorithm are :① partial data analyzing, and producing partial data model(partial knowledge).② combining partial data model in different data points and then getting the overall data model(overall knowledge) as the Fig.2.3 shows:

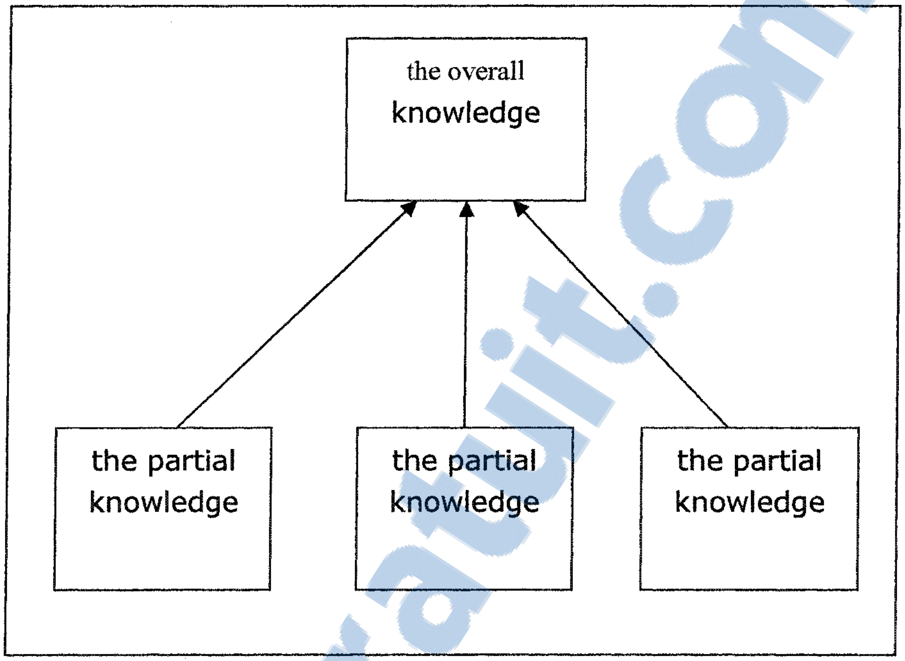

Figure 2.3 the generalization of the overall knowledge of the distributed data

2. 2. 2 the necessity of the distributed data mining[7]

The present data mining algorithm and model are mainly concentrative. Even under the condition of data distributed storing, it also requires to recollect these data into a concentrative place (like data warehouse), which demands high-speed network of data communications; also, responding time will become longer and the privacy and security will be damaged, Especially when the distributed data are not in the same structure. Although the band network is increasing, it cannot catch up with the speed of data increasing. As a result, it needs to use limited band network to move large capacity of data. What is more, the centralized concentrative data mining algorithm is not suitable for future analysis application of large capacity and distributed data. Therefore, because of privacy and secrecy of data and the incompatibility of system, it is not realistic to put all the data into a concentrative platform.

2.2.3 the key technological problems in distributed data mining

In the distributed data mining, there are the following four aspects of key technologies needing to be noticed:

(1) Data Consistency[30]

The first stage of data mining is to collect data from the data source of logical distribution or physical distribution. The traditional method is to extract data table first from association database, and then put it into a concentrative data warehouse or data set. Therefore, to the distributed data mining system, it is very important to provide a consistent storing structure for all the data mining processes. Besides, it is also critical to minimize the data movement of the whole data mining period as much as possible in the distributed environment. And also, an important topic is to develop an inquiry interface which is compatible with SQL for data mining algorithm to visit the information of the distributed database directly.

(2) Parallel Data Mining

On the side of servers, running data mining under large scales of data set are very time consuming, because data mining algorithm has very high complexity .A better way is parallel data mining algorithm. Many data mining algorithm have been developed, such as Association Rules [31], Neural Network [32], Genetic Algorithms [33], Decision Tree [34], and so on. But traditional algorithms used only think about the using of single data base, commonly they are series algorithms. Along with parallel and distributed technique development, more and more mining algorithms based on the parallel and distributed emerged. For example , recently parallel mining association algorithms have CD (Count Distribution)、CaD (Candidate Distribution)、DD (Data Distribution) [35] by Agrawal , PDM [36] by Park . Based on distributed data based mining algorithms have DMA [37] and FDM [38] by Chueng. Parallel classifier analysis algorithms have SPRINT [39] by Shafer. Usually, parallel mining algorithms and distributed data mining algorithms are universal.

(3) Knowledge Assimilation[40]

In data distributed and function distributed environment, knowledge assimilation is very important. Its basic ideas are using data mining algorithm to assimilate knowledge from several data sets (generally not disjoint), and then using the knowledge fragment produced in the data mining process to compose complete knowledge.

(4) Distributed Software Engineering[41]

In recent years, internet has become the super-structure of Client/Server computer mode in the worldwide. In new environment, application development is mainly to develop software parts, and then combine them. Software parts have encapsulation. Its compatibility with outside is finished through the applied procedure interface (API) which is defined beforehand. The biggest advantage of software parts is that they support software diplex. In this way system designing stuff can use existing software parts. The most popular distributed models nowadays are CORBA[42], ActiveX/DCOM and vTava Beans.

2.2.4 The Research Result of Distributed Data Mining

2.2.4.1 The existing distributed mining algorithms

□ Distributed Classifier Learning[43]

This algorithm mainly uses the technology of Meta-Leaming. According to different distributed forms of each site, it is divided into the algorithm for the same structure data and the algortihm for the different structure algorthm.

□ Collective Data Mining[44]

The ideas of this algorithm is that, at first, each site calculates automatically an approximate orthonormal basis coefficient, and then select some special samples from the data set of each station to move them into one site. Calculate the approximate basis coefficient targeting at non-linear crossterms according to the collective data set. And finally, according to the basis coefficient, unite the partial model into the overall model and submit it to the users.

□ Distributed Association Rule Mining[45]

This algorithm is divided into two kinds: Count Distributed and Data Distributed. The Count Distributed form is mainly used in the system in which data are distributed in the same structure. The statitics of frequent item sets is that each site counts respectively and then collects them to one site for the final arbitration. Data distributed form needs to exchange data banding in the process of mining. Therefore, it is mainly used in the system with high-performance network joining. These algorithms are all the distributed versions of Apriori algorithm and its changed form.

④ Distributed Clustering [46]

This algorithm is often used in the environment of the same structure data. It usually adopts the form of collective control. Some scholars apply the Agent technology to this algorithm, and have achieved good effect.

2.2.4.2 the architecture of the existing distributed data mining

In the aspect of the structure of the distributed data mining, many structures based on different technologies have appeared, such as Jiangchun Song,and Junyi Shen[47] ,They developed a n kind of distributed web-mining system based on CORBA (DWMBC) ; GUO Li-ming and ZHANG Yan-zhen[48] , have done research on the Distributed Data Mining System Based on Multi-agent Technology ; Jiang Wu-shan and Yu Ji-hui[49], put forward a service-oriented architecture of DDM on the grid , which can realize the data mining of large quantity under the distributed environment of heterogeneous; Krishnaswamy[50] has researched a kind of federated data mining system based on different structures and distributed environment; this system is used for electronic business application; J . Omer Rana, etc. put forward a kind of distributed data mining system frame with good expansibility based on groupware technology. This frame can conveniently integrate the third-side plug-in and the groupware defined by users themselves. Different from the centralized data mining system, the present distributed data mining systems are mainly in the research stage. There are still not mature commercial products. The present research hotspot of the distributed data mining mainly focuses on the handling of the extra-large-size data collection and increasing the overall performance of the distributed mining system. Grossman[51], etc. put forward a kind of integrated system called PDS(photonic data services). This frame has first integrated data service supporting teledata analysis and distributed data mining. It has designed the network agreement used to transmit data high-effectively in the network of high performance, and it has also designed chain-road service used for light-fibre network. This frame can do the distributed data mining of Gigabyte large-size data quantity.

CHAPTER 3
General introduction of Association Rules

The association rule mining is an important research topic in the filed of data mining, and the important content of KDD data research[52]. It is put forward in the environment of supermarket data with the motive of discovering how each kind of commodity is evolved. In the trade database of supermarket, there is association between different commodity items that users buy. From this association we can find the mode of users' buying behavior through association rule algorithm, such as the effect of buying one kind of commodity on other commodities. The discovery of such rules is useful for the design of commodity shelves, inventory arrangement and the classification of users according to their buying mode.

In 1993, Agrawal, etc. first put forward the problem of the association rule of the items in mining users' trade database[53]. From then on, many researchers have done research on the mining problem of association rules. Their work includes the optimization of the original algorithms, such as introducing random sampling, distributed, parallel ideas, and so on, to increase the efficiency of the algorithm of mining rule and promote the application of the association rule. At present, the association rule mining has been applied into many fields such as business decisions, market analysis in the enterprises,, testing the financial fraud, biopharmaceuticals and mode identification, and so on. With the deepening of the research and application promotion of the association rule mining, its importance and practicability will show out prominently.

3.1 Basic conception and problem describing

Set I= $\{i_1, i_2... i_m\}$ are itemset, in which the element is item, marked D as the aggregate of (transaction) T, here the transaction T is the itemset, and $T \subseteq I$. Every transaction has its unique identity, such as transaction number, marked TID. Set X is a aggregate of item in I, if $X \subseteq T$, then transaction T included the X.

A association rule is a containing formula like $X \Rightarrow Y$, here $X \subset I$, $Y \subset I$, and $X \cap Y = \Phi$. The support of the rule $X \Rightarrow Y$ in transaction data base D is the ratio of X and Y data and all the transaction data, it is called support($X \Rightarrow Y$):

Support($X \Rightarrow Y$) $= | \{T \mid X \cup Y \subseteq T, T \in D\} | / | D |$ (2.1)

If to item X, there is support(X) which is bigger than min-support (min-supp) presented by user, it is called X frequent itemsets or big itemsets.

The confidence of rule $X \Rightarrow Y$ in transaction aggregate means the ratio of included X and Y transaction data and data of all the transactions , which is called confidence(X Y) :

$\text{confidence}(X \Rightarrow Y) = |\{T \mid X \cup Y \subseteq T,\ T \in D\}| \,/\, |\{T : X \subseteq T\ T \square D\}|$ □ (2.2)

Presented a transaction aggregate D, mining association rules problem are association rules which generate min-support (min-supp) and min-confidence (min-conf) whose support and confidence are bigger than users'.

3.2 The variety of association rules

We classify the association rule according to different situations[54]:

(1) Based on the types of handling variables in the rules, the association rule can be divided into Boolean type and number type.

The values in Boolean type are all discrete, and are classified by type. It shows associations among these variables ; while number type association rule can be combined with multi-dimensional cooassociation or multi-layer association rule to handle the fields of number type, and do dynamic division, or handle the original data directly. Of course type variables can be included in the number type association rule.

For example : sex = "female"=> carrer= "secretary" is the Boolean type relate rule. sex = "female"=> avg income =1300, the involved income is number type. Therefore, it is a number type association rule.

(2) Based on the abstract level of data in rule, it can be divided into single layer rule and multi-layer association rule.

In the single-layer association rule, all variables do not consider that realistic data have different levels; but in the multi-layer association rule, the multi-layer of data has been fully considered.

For example, IBM desktop=>Sony printer , it is a single-layer association rule in detailed data; desktop=>Sony printer , is a multi-layer association rule in a relatively higher lever and is more detailed.

(3) Based on the involved dimension of data in the rule, the association rules can be divided into single dimension and multi-dimension.

In the association rule of single dimension, we only deal with one dimension of data, such as the goods that user buys; while in the association rule of multi-dimension, the data to be handled will involve many dimensions. In other words, the single association rule handles the association of single attribute; multi-dimension association rule handles the associations among each attribute.

For example, beer=>diaper, this rule only involves the goods users purchase; gender= "female"=>career= "secretary", this rule involves the information of two fields, and an association rule in two-dimension.

After giving the classification of the association rule, in the actual application of the association rule, we can consider which concrete method is fit for which kind of mining of rule, and how many different methods can be used to handle some kind of rule.

3.3 The steps and classic algorithm of mining association rule

3.3.1 The steps of mining association rule

The mining of the association rule is generally divided into the following two steps[55]:

□Discover all the item sets that the supporting degree of all transactions is bigger than the minimum supporting degree. The supporting degree of one item refers to the transaction number including this item. The item with the minimum supporting degree is called frequent item, and the others are all non-frequent items.

□Construct the rule in the frequent items that the confidence degree is not lower than the minimum confidence degree. If ABCD and AB are frequent items, and we can calculate the percentage conf≥min-conf, then the rule is right. (This rule is certain to have the minimum supporting degree because ABCD is frequent item.)

In the two steps mentioned above, the fist step is the critical step in the mining association rule. The overall performance of the association rule mining is decided by the performance of this step. Therefore, present researches all focus on the first step, i.e. the mining handling of the frequent items. Relatively speaking, the second step is easier to realize, because it only needs to list all the possible association rules in the frequent items that have been mined, and then weigh these association rules with the threshold of the minimum supporting degree, and finally find the interesting association rules.

3.3.2 Classical Association Rules Algorithm —— Apriori Algorithm[56]

Single-dimensional, single-level, Boolean association rules are the simplest association rules for data mining. Apriori is an influential algorithm for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties, as we shall see below. Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k+1)-itemsets. First, the set of frequent 1-itemsets is found.This set is denoted L_1. L_1 is used to find L_2, the set of frequent 2-itemsets, and so on, until no more frequent k-itemsets can be found. The finding of each L_k requires one full scan of the database.

(1) Apriori property

To improve the efficiency of the level-wise generation of frequent itemsets, an important property called the Apriori property, presented below, is used to reduce the search space.

In order to use the Apriori property, all nonempty itemsets of a frequent itemset must also be frequent. This property is based on the following observation. By definition , if an itemset *I* does not satisfy the minimum support threshold , min-sup , then *I* is not frequent , that is , $P(I) <$ min-sup .If an item *A* is added to the itemset *I*, then the resulting itemset (i.e., *I*□*A) cannot* occur more frequently than *I*. Therefore , *I*□*A* is not frequent either ,that is , *P(I*□*A)* < min-sup

So we can use the ob-axiom : If a set cannot pass a test , all of its supersets will fail the same test as well。 So it is easy to make sure the Apriori property stand. "How is the Apriori property used in the algorithm?" To understand this ,let us look at how L_{K-1} is used to find L_K.A two-step process is followed ,consisting of join and prune actions.

□.The join step

To find L_k, a set of candidate k-itemsets is generated by joining L_{k-1} with itself. This set of candidates is denoted C_K. Let l_1 and l_2 be itemsets in L_{k-1}..The notation $l_i[j]$ refers to the *j*th item in l_i (e.g., l_i [k-2] refers to the second to the last item in l_i). By convention, Apriori assumes that items within a transaction or itemset are sorted in lexicographic order. The join, $L_{k-1} \infty L_{k-1}$, *is* performed, where members of L_{k-1} are joinable if their first (k-2) items are in common. That is, members l_1 and l_2 of L_{k-1} are join if (l_1 [1] = l_2 [1]) (□ l_1 [2]= l_2 [2]) ... (□ □ l_1 [k-2] = l_2 [k-2]) (□ l_1 [k-1] = l_2 [k-1]). The condition l_1[k-1] < l_2 [k-1] simply ensures that no duplicates are generated . The resulting itemset formed by joining l_1 and l_2 is l_1 [1] l_1 [2]...l_1 [k-1] l_2 [k-1].

□.The prune step

C_K is a superset of L_k, that is, its members may or may not be frequent,but all of the frequent k-itemsets are included in C_K. A scan of the database to determine the count of each candidate in C_K would result in the determination of L_k (i.e., all candidates having a count no less than the minimum support count are frequent by definition, and therefore belong to L_k). C_K, however, can be huge, and so this could involve heavy computation. To reduce the size of C_K, the Apriori property is used as follows. Any (k-1)-itemset that is not frequent cannot be a itemset of a frequent k-itemset.Hence, if any (k-1)-itemset of a candidate k-itemset is not in L_{k-1}, then the candidate cannot be frequent either and so can be removed from C_K. This itemset testing can be done quickly by maintaining a hash tree of all frequent itemsets.

(2) Apriori algorithm describing[15]

Apriori algorithm describing as algorithm 3.1 showed.

Algorithm: 3.1 (Apriori) Find frequent itemsets using an iterative level-wise approach based on candidate generation.

Input : Database D , of transactions ; Mini-support threshold, min-sup.

Output : L , frequent itemsets in D.

```
L1 = find_frequent_1_itemset ( D );
For ( k = 2 ; $L_{k-1} \neq \Phi$ ; k+ + ){
    $C_k$ = apriori_gen ( $L_{k-1}$ , min_sup );
        //Based on frequent ( k - 1 ) -itemset candidate k-itemsets
        for each t□D{ //scan D for counts
            $C_t$ = itemset ( $C_t$ , t ); //get the itemsets of t that are candidates
            for each c□$C_t$
            c.count++;
}
$L_k$ = { c□$C_k$|c.count>min_sup }
}
Return L = $U_k L_k$
Procedure apriori_gen ( $L_{k-1}$ , min_sup )
for each $L_1$□$L_{k-1}$
```

```
        ( 2 )   for each $L_2$□$L_{k-1}$
        ( 3 )   if(( $L_1$[1]= $L_2$[1] ) □…□ ( $L_1$[k-2]= $L_2$[k-2] ) □ ( $L_1$[k-1]< $L_2$[k-1] ))
{
        ( 4 )   c= $L_1$ □$L_2$ ; //join step : generate candidates
        ( 5 )   if has_infrequent_itemset ( c , $L_{k-1}$ )
        ( 6 )   delete c ; //prune step : remove unfruitful candidate
        ( 7 )   else $C_k$ = $C_k$ U {c}
        ( 8 )   }
        ( 9 )   return $C_k$
        Procedure has_infrequent_itemset ( c , $L_{k-1}$ )
        ( 1 )   for each ( k-1 ) itemset s of c
        ( 2 )   if s≇ Lk-1 return TRUE ;
        ( 3 )   else return FALSE
```

The first step of Apriori algorithm is finding frequent 1-itemsets L1 ; From (2) to (8) step , used Lk-1 to created Ck for getting Lk . Apriori_gen process produced relevant candidate itemsets ; Then used Apriori Property to deleted those itemsets which are candidate itemsets of the non-frequent itemsets(the 3ird step)。Till found all of the candidates , then searched the data base (the 4th step) , To every transaction of the data base used itemset function to helping find all the itemsets of the candidate itemsets in that transaction records (the 5th step) , Added up the support frequent of every candidate itemset (the 6th step) . Finally the satisfied the min-supports candidate itemset made up of the frequent itemset L. So we can used this process to helping get all the association rules in frequent itemsets.

Apriori process completed two operations, the one is connection, the two is deleted operation. Such as the above introduced , in connected process , Lk-1 and Lk-1 connected to create latency candidate itemsets (from the 1st to the 4th step of the algorithm) ; In deleted operation (from the 5th to the 6th step of the algorithm) used Apriori property to deleted itemsets of the non-frequent itemsets in the candidate itemsets . (from the 1st to the 4th step of the algorithm) ; In deleted operation (from the 5th to the 6th step of the algorithm) used Apriori property to deleted itemsets of the non-frequent itemsets in the candidate itemsets .Has_infrequent_itemset process finished to check the non-frequent itemsets.

3.3.3 The existing improvement of Apriori algorithm[56]

(1) Hash-based technique (hashing itemset counts) [36]

A hash-based technique can be used to reduce the size of the candidate k-itemsets , C_K ,for k > 1 . For example, when scanning each transaction in the database to generate the frequent 1-itemsets, L_1, from the candidate 1-itemsets in

C_1, we can generate all of the 2-itemsets for each transaction, hash (i.e., map) them into the different buckets of a hash table structure, and increase the corresponding bucket count. A 2-itemset whose corresponding bucket count in the hash table is below the support threshold cannot be frequent and thus should be removed from the candidate set. Such a hash-based technique may substantially reduce the number of the candidate k-itemsets examined (especially when k=2).

(2) Partitioning data[57]

A partitioning technique can be used to find candidate itemsets in just two database scans. It consists of two phases. In Phase □, the algorithm subdivides the transactions of D into nonoverlapping partitions. If the minimum support count for a partition is min-sup ,the frequency of minimum support for each part is : min_sup×number_of_transaction_of_partition. For each partition, all frequent itemsets within the partition should be found. They are referred to as local frequent itemsets. The procedure employs a special data structure that, for each itemsets, records the TIDs of the transactions containing the items in the itemset. In this way, it can find all of the local frequent k-itemsets, for k=1,2,....

To the entire database D, A local frequent itemset may not be a overall frequent itemset. Any itemset that is potentially frequent with respect to D must occur as a frequent itemset in at least one of the partitions. Therefore, all local frequent itemsets are candidate itemsets with respect to D. The collection of frequent itemsets from all partitions forms the global candidate itemsets with respect to D .In hash; a second scan of D is conducted in phase□ which the actual support of each candidate is assessed in order to determine the global frequent itemsets. Partition size and the number of partitions are set so that each partition can fit into main memory and therefore be read only once in each phase.

(3) Sampling (mining on a itemset of the given data) [58]

The basic idea of the sampling approach is to pick a random sample S in the given data D, and then search for frequent itemsets in S instead of D. In this way, we trade off some degree of accuracy against efficiency. The sample size of S is such that the search for frequent itemsets in S can be done in main memory, and so only one scan of the transactions in S is required overall. Because we are searching for frequent itemsets in S rather than in D. it is possible that we will miss some of the global frequent itemsets. To lessen this possibility, we use a lower support threshold than minimum support to find the frequent itemsets local to S (denoted Ls). The rest of the database is then used to compute the actual frequencies of each itemset in Ls. A mechanism is used to determine whether all of the global frequent itemsets are included in Ls. If Ls actually contains all of the frequent itemsets in D, then only one scan of D is required. Otherwise, a second pass can be done in order to find the frequent itemsets that were missed in the first pass. The sampling approach is especially beneficial when efficiency is of utmost importance, such as in computationally intensive applications that must be run on a very frequent basis.

(4) Dynamic Item Counting[59]

Dynamic item counting is to add candidate items at different moment of scanning. Dynamic item counting is put forward in the process of dividing and mining database. Each divided data module is marked with the sign of "beginning". During this change, at any beginning point, it can add new candidate item; before each database scanning, the candidate items have been decided. This kind of technology is dynamic because it needs to estimate the supporting degree of all items that have been counted up to now; if all the sub-sets of an item are estimated as frequent, and then add a new candidate item. The algorithm obtained through this method needs scanning twice.

(5) Cyclical Market Shopping Analysis

Cyclical market shopping analysis is to discover the corresponding frequent items in the period defined by the users. Cyclical market shopping analysis makes use of the trade record with time mark to determine the set in the trade database and mark it as period. The so-called period is a group like "the first day of a month" and so on. From the item of each day in period, we extract corresponding association rule. In this way, an item that doesn't meet the minimum supporting threshold can be regarded as frequent in a data sub-set that meets the constraint of period.

(6) Sequence Pattern

Sequence pattern is to discover the trade sequence (pattern) with the change of time. The aim of sequence pattern analysis is to mine the items according to time. This kind of research is scarce in reality.

(7) Other methods include the association rule mining with multi-layer and multi-dimension and the mining of time-sequence data, and so on.

3.4 The Generalization of Association Rule

After mining all the frequent items in database D, it is easier to obtain corresponding association rule. That is to say that to generate the strong association rule meeting the minimum supporting degree and minimum confident degree, the forum introduced in 2.3.1 can be used to calculate the confident degree of association rule[53].

The detailed operation introduction on generating association rule is as following:

(1) For each frequent itemset l , generate all nonempty itemsets of l ;

(2) For each non-empty itemset of 1, if $\frac{\text{support\_count(l)}}{\text{support\_count(s)}} \geq \text{min\_conf}$, then a association rule "s=〉(l - s) " will be generalized; thereinto, min_conf is the

minimum confident degree threshold.

Since the rules are generated from frequent itemsets, each one automatically satisfies minimum support. Frequent itemsets can be stored ahead of time in hash tables along with their counts so that they can be accessed quickly.

3.5 The analysis of some existing association rule mining algorithms

The mining of the association rule has made remarkable achievements, and already put forward many good association rule mining algorithms. According to the application environment, we classify them into three types:

One is the association rule mining algorithm used in centralized database system [60-64]. This kind of algorithm includes AIS, Apriori, AprioriTid, AprioriHybrid[65] put forward by Agrawal, etc. and DHP[66] put forward by Park, etc., and the dividing algorithm PARTITION[67] put forward by Savasere, etc., and the sampling[68] algorithm put forward by Tovionen, etc., and some updating algorithms of the association rule such as FUP, IUA and NEWIUA, and so on.

Among them, the basic idea of Apriori algorithm is scanning database repeatedly. The length of scanning in the kth time is the big itemset L_k ; when scanning in the k+1th time, on the basis of k itemset in L_k, the candidate collection

C_{k-1}will be produced; DHP algorithm uses Hashing technology to improve the producing process of the standby itemset C_k; the algorithm PARTITION is to divide the database, reducing the I/O times in the mining process; the algorithm Sampling is first to sample the database, and then mine the sampling data to increase mining efficiency.

The second one is the algorithms solving the problems of the association rule mining in the parallel environment: the CD (Count Distribution) put forward by Agrawal, etc., DD (Data Distribution), CaD(Candidate Distribution) [69] and PDM(efficient Parallel Data Mining for association rules)[70] put forward by Park, etc., and so on. All these algorithms are based on the algorithm Apriori. The precondition is that the processor contains special memory and disk and there is no region that can be shared in structure. The processor is joined by communication network, and the information transmitting is used for communicating; and data are allotted evenly to the special disk of each processor.

The basic idea of CD algorithm is that: on each processor, there stores the overall candidate itemsets and frequent itemsets. In each step of calculation, we use Apriori algorithm to calculate the number that the candidate itemsets supports in the local data. And then do synchronization once, and each handling itemset exchanges the supporting number of the local candidate item collection to make all the candidate item collections of each processor get the overall supporting number, so that the overall frequent item collection L_k can be got. CD algorithm allows parallel redundant calculation and redundant storing in other processors, avoiding transmitting large quantity information as much as possible.

DD algorithm makes use of the overall storing space more effectively. It stores different candidate item collection in each processor. Each processor, in order to calculate the overall supporting number of local candidate item collection C^i_k, must both calculate the supporting number of the local candidate item collection and the supporting number in all the other processors. Therefore, it must broadcast the local data, and receive the data transmitted from the other processors. Because the communication load is very large, the machine should have higher communication speed.

The algorithm CaD combines CD and DD algorithm. When generating one item collection, we use CD algorithm or DD algorithm. But when generating the following k (k>1) item collection, the algorithm allots the frequent item collection L_{i-1}, and also re-allots the transaction database; when generating C^i_k, it should separate it from other processor, not waiting other processors to transmit the complete pruning information, but only when other processors transmits lag pruning information, leaving them to be handled in the next pruning time. Although CaD algorithm avoids large quantity information transmitting, its efficiency is not so ideal due to the re-allocation of the transaction database.

The third one is the algorithm solving the problems of the association rule mining in the distributed environment, such as DMA[71] , FDM, etc. [72-78]. The design of algorithm DMA is based on the principle that "if the item collection X in DB is big item collection, then it must also be big item collection in some DB". The algorithm uses local pruning technology to generate the candidate big item collection which is smaller than that of the algorithm CD. When each sites exchange supporting number, the algorithm uses polling site technology to make the communication cost of each tem collection X degrade to 0 (n) from 0 (n^2)of the algorithm CD, and n is the number of the site. Although the algorithm DMA overcomes some weakness of the algorithm CD, it needs the supporting number of all the other sites when generating K frequent big itemset, having more synchronization times with other sites. The algorithms FDM and DMA are almost the same. The difference is only that FDM adds the overall pruning technology.

Because each site in the distributed environment is joined through tele-network. Constrained by the speed and the reliability of the network joining, the time cost of transmitting the data and information among the sites is high, and the reliability is not so good. Therefore, in the distributed environment, it should reduce the unnecessary communication and data transmitting as much as possible.

CHAPTER 4
The distributed data mining system based on multi-agent

4. 1 the distributed data mining system (DDMS)

Specifically speaking, data mining can be regarded as a forecasting model or rule set got from one or more (distributed) data collections applying corresponding data mining algorithm. Here different strategies can be used mainly according to the data themselves, the distribution of the data, the software and hardware resources that can be used, and the required precision. Accordingly, the centralized distributed data mining systems have some differences in the following strategies [79.80]:

(1)Data Strategy

The distributed data mining can choose the final result of moving data, or moving middle result, or providing forecasting model, or moving data mining algorithm. We can use the distributed data mining system of Local Learning to establish models in each distributed places, and then carry these models to a centre region. We can also use the data mining system of Centralized Learning to carry the data to the centre region and then establsih models. Besides, some data mining systems use Hybrid Learning, i.e. the strategy combining partial leaning and the centralized leaning.

(2)Task Strategy

The distributed data mining system can choose to co-ordinately use one kind of data mining algorithm in several data stations, and can also choose to use different data mining algorithms independently in each data station. In the mode of Independent Learning, each kind of data mining algorithm is respectively applied in each distributed data station; in the mode of Coordinated Learning, one (or more) data station use one kind of data mining algorithm to coordinate mining task in several data stations .

(3)Model Strategy

There are many methods of combining the forecasting models established in different places. Among these methods, the simple and the most often used one is making use of voting, which is to combine the output of the models of each type according to the majority voting. But the method of Knowledge Probing is to establish a comprehensive model according to the input and output of all kinds of models and the expected output.

A distributed data mining system should have good performance in Scalability、 Efficiency、Portability、Adaptivity、and Extensibility.

The extensibility of a distributed data mining system is such a kind of ability of the system: when the number of the data sites is increasing, the performance of the system has no substantive and obvious declining. The effectiveness means to make use of the centralized system resources effectively and get the correct mining results.

The Portability refers tothat a distributed data mining system should normally operate in the multi-environment with software and hardware equipments, and can combine multi-model with different expressions.

Almost all the environment of most data mining systems will change. The adaptivity of the distributed data mining system refers to the ability hoe to evolve and adjust according to the changed environment.

Not only data and mode will change with time, but algorithms and tools will have some change with the progress of machine leaning and data mining. The extensibility of the distributed data mining system means that ii must have enough flexibility to adapt the present and future data mining technology; or else, it will be inapplicable quickly and close to obsolescence.

4.2 Agent introduction

The direct background of agent and its relevant concepts is distributed artificial intellengence[81], its basic theory was brought about by John.McCarthy[82] in 1950s. And now it is widely used in areas such as process control, production-manufacturing, information management, intelligent database, data mining, network management and e-business. Agent technology is a new algorithm model, which is highly intelligent, easy to construct distributed system and having strong reusability.

4.2.1 the characteristics and definition of Agent

The two characteristics of Agent is intelligent and acting ability. Intelligent means the ability to use reasoning, learning, and other skills to analyze and explain various information and knowledge which it meets or receives. Generally speaking, Agent should have the following four basic characteristics[83]:

(1)autonomy: agent can be operated without the intervening of people or other agents. Also, agent can control its own behavior and inner situation; the behavior of agent should be active and voluntary; Agent should has its own objective and intention; Agent should make plans for its own behavior according to objective and environment.

(2) reactivity: Agent can sense and understand its environment, and respond in

time to the changes of environment.

(3)pro-activeness: can not only respond to environment, it can also adopt behavior to face the objective through receiving some starting information.

(4)socialibility: Agent with sociability is very friendly. It has good social relationship diffuse skills. Agents can communicate with each other by agent language. They can communicate, share out the work and help one another, and constitute a society or group with many agents.

The concept of agent and technology has appeared in the development of distributed applied system and shown its remarkable effectiveness. From some research about agent and developing work in the aspect of distributed application, we can see the meaning of the concept and techonology of agent.

□. Agent techonology can improve the application of internet such as the agent which develops "finding person with information". The agent, according to the information, can initiatively notice information provider that who needs the provided information at present ;

□. Agent techonology can improve the application of parellel projects, such as the manager of agent technology developing work. It can make the workflow and programming known to each workstation, and initiatively guide each workstation to promoe the work according to the workflow and programming, handle and estimate the reports of work condition of each workstation, and manage centrally all kinds of data,and so on.

□. Agent techonology can be used to develop the distributed interactive simulation system. For example, it can connet the simulator of flight training and several workstations in the computer network, and realize many agents imitating airplanes in workstations to form interactive aviation simulation system together with simulator. For this kind of simulator operation, the trained stuff can not only experience all kinds of skills of operating planes, but also realzie various kinds of air actions through the interaction with the intelligent autonomy imitating airplane

4.2.2 Multi-agent system

A multi-agent system (MAS) [84] is a system composed of multiple interacting intelligent agents. Multi-agent systems can be used to solve problems which are difficult or impossible for an individual agent or monolithic system to solve. Examples of problems which are appropriate to multi-agent systems research include online trading[85],disaster response[86], and modelling social structures。 [87]. At present, people have begun to apply multi-agent system into the research of distributed data mining system. For example, Centralized data mining system--BODHI[44]; expandable distributed data mining system—(Parallel Data Mining Agents)[88]; middle-learning distributed data mining system—JAM[89], etc..

Multi-agent technique's applying to distributed data mining system has the following advantages:

①Agent's autonomy: agent's autonomy is correspondent to the autonomy od data source. Agent can visit local data according to local visit limitation and safety strategy, cooperate with information on different data sources. In this way, the protection of private information can be strengthened.

②Agent's go-aheadism: it can limit customer's supervision and the intervening to data mining process. Customers can set objective and method for agents at the early stage. And during the operating process, agents can adjust the task exercising process.

③Agent's self-adaptation: this means agents can choose data source and collect data independently. One important problem of distributed or real environment is the changing of environment, which will result in the changing of data source. In this circumstance, agents can search and select data according to standard set beforehand, such as expected amount, type and quality of data, etc.. in this way, agents can be basic tools to search for data source for static data mining method in dynamic environment, thus static method can be used to analyze dynamic data.

④Agent's coordination: this means it can expand traditional data mining method, which will enable them to suit mass data in distributed environment; realize data mining with multi-techniques; integrate results from each local site point, and achieve results in overall situation.

4.3 the distributed data mining system based on multi-agent

4.3.1 Structure

In this part, the thesis also brings about a distributed data mining system which is based on multi-agent. This system can not only mine local data information, but can also do distributed data mining in different data site point. It is composed of users' interface agent, users' information base, knowledge managemen agent, task management agent, the overall knowledge base, coordianting machine agent, and data minign agent.

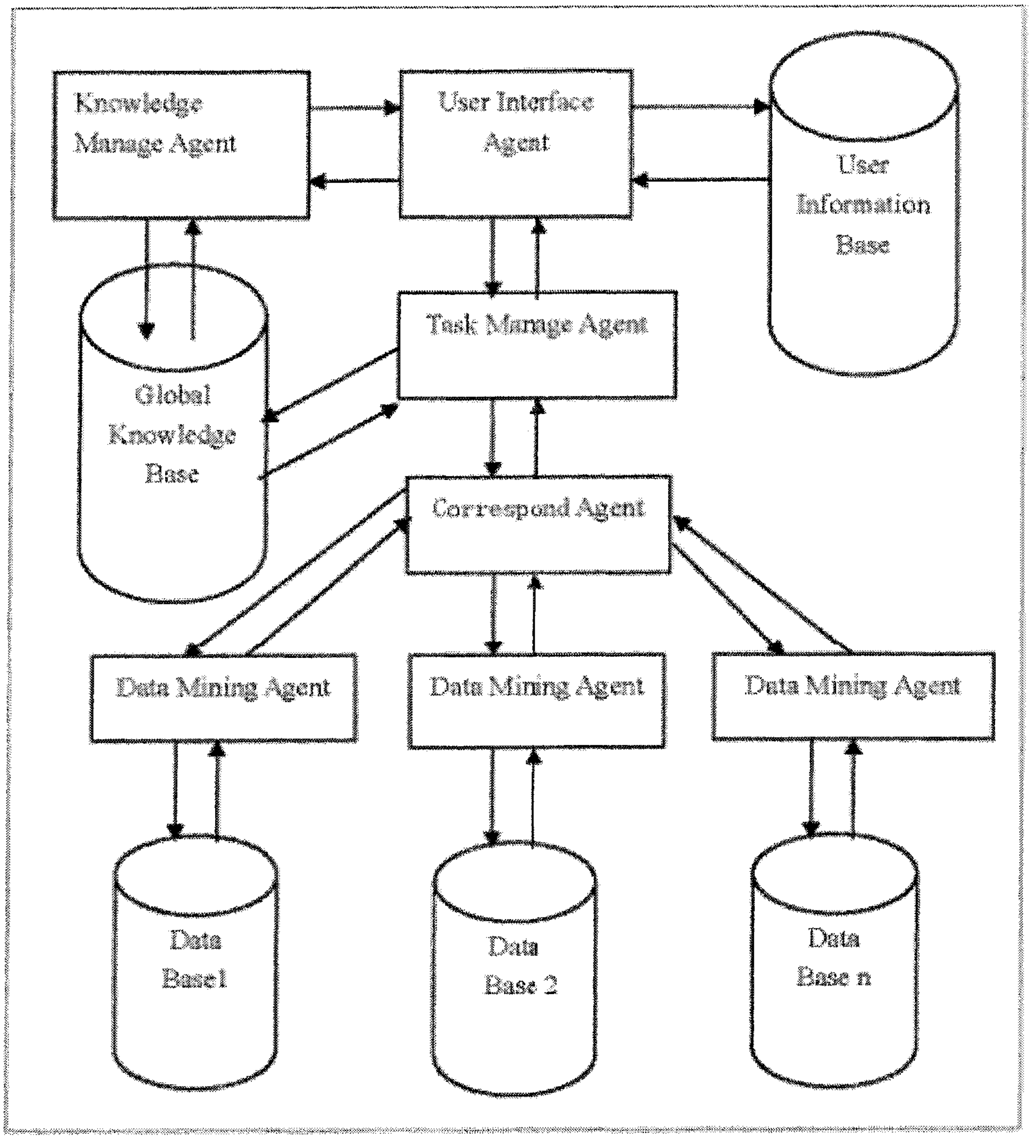

Figure 4.1 structure of the distributed data mining system based on multi-agent

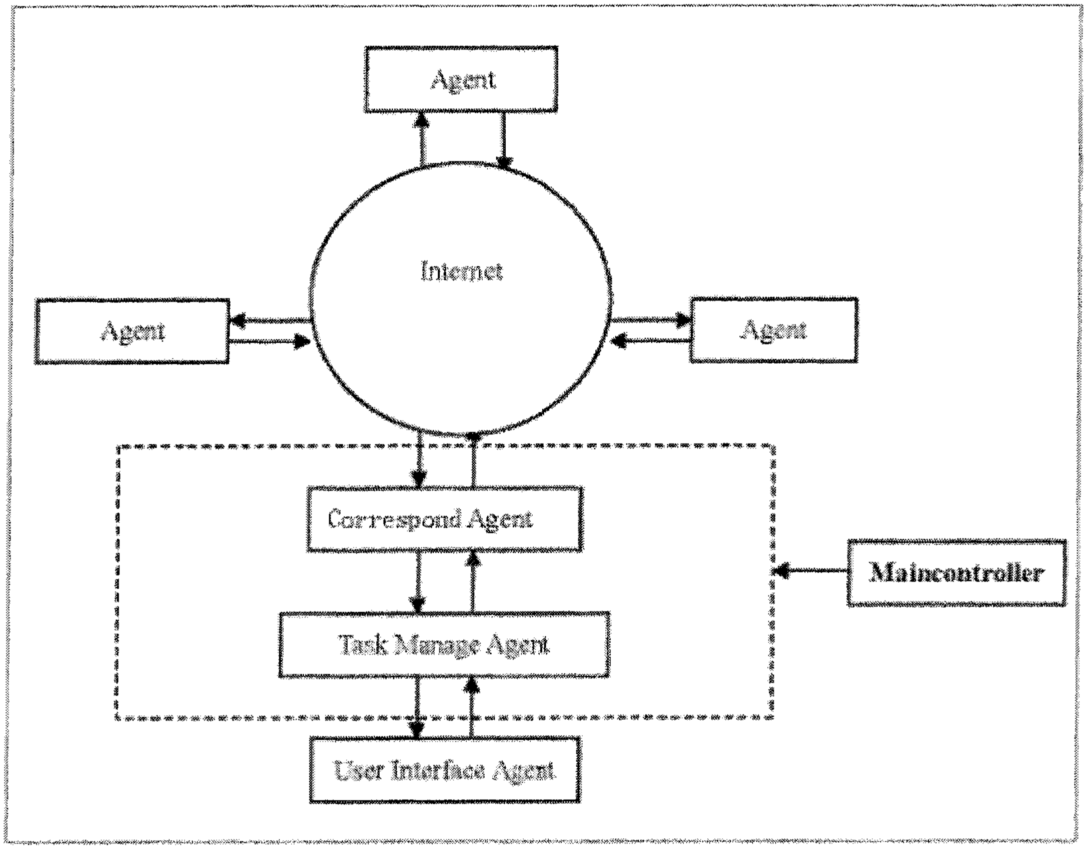

Figure 4.2 system topology structure

4.3.2 Module function

Each function module of this system is composed of multi-agent. The agents coordinate with each other to realize functions of the system. Therefore, this system is the distributed data mining system based on multi-agent. In the following, we will introduce the main functions of each module :

4.3.2.1 Users' interface agent

□It is used to realize the alternation between users and the computers. The communicatin between users and the system is finished through the users'interface agent. The users need not to communicate with other agents.

□It can provides users nicer interface. The users' interfaces all use WEB form. There are three advntages :

□. It is convenient for the users to deposit and take out.

□. Separated from the platform: the interface has nothing to do with the system platform. Users need not to master professional knowledge. They only need to input the mining demand precisely and then the system will automatically transform the demand into mining task language and allot task to mining agent to mine the data; after mining, the mining result will be displayed to users in the proper form.

□.low costs in establishment and management

□Gathering users' demand, and according to the demand to analyze the information such as users' interest, and storing the analyzing result into the information base f users;

□Taking the system security into consideration, in the process of the alternation between users' interface agent and users, it will, according to the information in information base, validate the identification of users.

4.3.2.2 Task management agent

Task management agent is used to establish, manage, start-up and execute data mining task. Making use of task management agent can pack the handling steps needed by the data mining task including data picking-up manner, data pre-handling manner, data dispersing method, data mining algorithm, and so on, into a data mining task. Task management agent can manage this mining task, and through task management agent, start-up the mining task, and record the state information of mining execution. This is in favors of the establishment, management, repeat execution of mining task, and so on. Specifically speaking, task management agent has the following functions:

□After receiving the mining demand of users from the users' interface agent, it disassembles the users' demand into each mining sub-task;

□Determining which kind of algorithm will be used (in this system, the mining algorithm applied in data mining sub-agent is not the only. We set many kinds of effective mining algorithm suitable for different conditions in it), and making the task known to the corresponding agent;

□Meanwhile, establishing mining log, and recording the information such as the data, task, result of mining every time for the future comprehensive analysis of the mining knowledge.

4.3.2.3 Correspond agent

It is used to realize the communication between agents.

□After receiving the mining task given by task management agent, according to the specific situation, it makes the task known to the corresponding data mining agent; after mining, it receives the mining results transmitted from mining agent, and then transmits them to the task management agent;

□Transferring the metadata, and providing the overall sharing;

□In the process of mining, it coordinates the information transferring among the data mining gents. When it needs to make the mining task known to some data mining agent, it should first test whether this mining agent is busy or not. If this mining agent is not busy, then make the mining task known immediately; or else, wait until the present mining task is over, then make the new mining task known.

□Logging on the mining log

4.3.2.4 data mining agent (DMA)

DMA is the core of the system, which can realize the data mining analysis of the local database. It mainly includes three functions: data picking-up, data pre-handling, and data mining. They can be expressed by different sub-agents. Data picking-up and data pre-handling are data preparing process. Data preparing is a very important link in the data mining process [8]. If the task of data preparing is done well, the data quality is high, then the process of data mining will be faster and more convenient, and the modes and rules that are mined finally will be more effective and applicable, and the results will be more successful.

□Data picking-up sub-agent: Picking-up data from one or more tables in the appointed database to a new table, and managing the picking-up situation; displaying the data in the table in the graphics mode and providing the attribute that users choose to extract.

□Data pre-handling sub-agent: transforming data into the form easy to mine, and establishing and managing the concept number of the attribute; transforming according to the concept number to form the mining data table.

Data pre-handling is an important step in the process of data mining (knowledge discovery), especially when mining data that contain yawp or are incomplete, even conflicting, it needs data pre-handling even more in order to increase the quality of the data mining objects, and finally achieve the aim of increasing the quality of the mode knowledge obtained by data mining. The so called yawp data means that there are wrong or abnormal (deviating the expected value) data; incomplete data refers that the attribute of the interesting data has no value; while the conflicting data means that there is conflicting situation in the data connotation (for example, the coding as the key word in one department have different value).

Data pre-handling includes data cleaning, data integrating and data dispersing. Data cleaning is to handle the pretermission in data and clean dirty data. Data integrating is to combine and handle the data of multi-data sourece to solve semantic fuzziness and conform into coherent data storing. Data dispersing will discern the data muster that needs to mine, and narrow handling scope.

□Data mining sub-agent : using mining algorithm to mine the data information that has been extracted and handled.

4.3.2.5 Knowledge management agent

It provides the managing function of the data mining result knowledge, stores mining result into the overall knowledge base, and displays mining results, and provides the interface that the experts estimate the mining results and according to the expert analyzing result, deletes some useless rules.

The results of data mining have many forms such as mode model, concept rules, report graph, and so on. Estimating and testing the results are the indispensable part in the whole process, including the mode explaining discovery, and displaying the discovered knowledge in the form that is easy to understand to the End-User; it needs to estimate the reliability and practicability of the model, and on the basis of understanding the model, adjust and improve the model. After the comprehensive estimation and repeated testing, the redundancy and conflicting in the results will be solved, and the aim of eliminating the false and retaining the true has been realized to make it achieve the expected aim, and invest it into the actual application finally. If the mining results are not satisfying, it can repeat the previous steps and links until the more ideal concept rules or mode models are generalized.

4.3.2.6 Users' information base

there are two kinds of information storing in the users; information base: one is users' managing information that is used to enrol, maintain and manage the users' information, and authenticate the users' logging, and setup the purview and PRI. The other is the information about users' interest and hobbies that is regarded as the reasoning rules used for the alternation between the users' interface and users.

4.3.2.7 The overall knowledge base

The results of data mining not only can be provided to the users through person-computer interface, but also can be stored into the overall knowledge base for the future further analysis. Because the model diversity of the mining, the representation form of the knowledge will be different, having no uniform forms. Therefore, it can setup a table for each kind of mining algorithm to store the knowledge got through this algorithm.

4.3.3The Work Process of System

In this system, data mining agent DMA is responsible to store-etract data and mine higher-level users' information from data. DMA works in the parallel form. The coordinator is used to communicate and share information between DMA. Coordinator collaborative agent provides information to users, and feedbacks the users' information to the agent. The basic work principle of the system is as follows:

□The users (who have passed the identification test of users' mining agent) give out the mining requirements;

□The task managing agent accepts mining requirements, and packages the mining requirements according to the scheduled format and then transmits it to the coordinator;

□The coordinator analyzes the mining requirements and fixes the involved DMA;

□The coordinator checks up the state of DMA and if DMA is not operating, establish DMA;

□The coordinator broadcasts the mining requirements to DMA;

□DMA mines automatically the corresponding information according to the mining requirements;

□The coordinator collects the corresponding information from each DMA, and then analyzes it comprehensively, and gets the final result information.

□Task managing agent submits the result information to the users through the users' interface agency

4.4 Summery

This chapter put forward a distributed data mining system based on multi-agent. The structure features of this system are mainly as follows:

① Making use of multi-agent technology, and the features of multi-agent coordinating system such as sociality, automatism, and collaborative to make the system more intelligent, and meet the users' demand.

②Introducing task management agent. and making use of it can pack each handling step needed by a data mining task such as data picking-up mode, data discrete method, data mining algorithm, and so on, into one data mining task, which is convenient for system automatically to control the automatic execution of multi-step of a mining task, and also convenient to manage mining task, and convenient for one mining task to be executed many times according to the requirements.

③The system structure is complete, which provides complete supporting for each handling step of knowledge discovery.

CHAPTER5 The distributed association rule mining algorithm based on multi-agent --RK-tree algorithm

We have put forward a distributed data mining system based on multi-agent. The realization of a better distributed mining system will depend on a high-effective algorithm. The association rule mining is a kind of data mining algorithm that is often used. Chapter two has stated that there is the problem of high communication cost in the present distributed association rule mining algorithms that have been put forward. To solve this problem, this chapter will put forward a distributed association rule mining algorithm whose communication cost is lower--RK-tree algorithm.

5. 1 the basic concept and theory

Definition 4. 1 supposing a distributed database system S that is composed of n sites s^1、s^2、s^3、......s^n. DB is the distributed database of S. the database in the station s^i is DB. DB=DB^1 U DB^2U ......UDB^n . D and D^i respectively stand for the size of the database in DB and DB^i, D= D^1+D^2 +... ...+ D^n. DB is called overall database, and

DB^i is called local database.

5. 2 the basic principle of RK -tree

RK –tree algorithm has following steps :

(1) each site point adopts local-mining, and then gets local rules set R(i)(i=1 , 2 , ... , n) , in which local-mining use Apriori algorithm to realize association mining.

(2) local rules set R (i) resulted from each site point is sent to the main controlling site point as results. The main site point builds an overall rules knowledge database, which is used to collect all rules sent by sub-site points, and reflect them onto an association tree. Then an association rules tree—RK-tree is generated.

RK-tree includes all association rules database. And final association rules will be mined in this overall database.

The constructing method of RK -tree :

□first, create a root node of tree, and mark it with "null";

□start scanning rules knowledge base D, D is an overall rules base which is formed by collections from each sub-site point. It creates a branch in tree with each rule mined from site point 1 according to the order of first component (P) —> consequent (B) . The rule's consequent is the leaf node of branch, and record this rule's appearing times as 1 at the leaf node;

□and then, compare rules mining from site point 2 with rules in the tree. If it is same with a certain rule in the tree, them new branch with not be created. Then record this rule's appearing times as 1, and record it into next rule. If not, then create a new branch, and record appearing times as 1;

□with this method, reflect all ruled mined from all site points which are saved in ruled knowledge base D onto the tree.

(3) with the given value N— the least appearing times rule, scan through the rule tree formed by overall rules base. And then compare the appearing times of each rule recorded in rule tree, delete branch whose times is smaller than value N, and delete correspondent rule from rule knowledge base.

Because what we want is overall association rule, apparently this rule should exist commonly in all sub-site points database. If some rule is generated just from certain site points, them it is definitely not overall rule. Therefore, the selection of rule times value N will influence directly this algorithm's speed and rate of convergence. For different value N, this thesis offering some experimental results.

(4) after deleting all branches smaller than value N, scan all sub-site points' database again, and obtain information of left branch rules in each sub-site point, such as supporting rate, confidence coefficient, and event number of supports.

(5) count supporting rate and confidence coefficient of each branch rule left in

RK-tree with branch rule information obtained in step four. And determine the overall rule according to the given smallest supporting rate and confidence coefficient.

Here it needs to say that the concentrative association rule mining algorithm selects association rules according to the given supporting degree and confidence degree. In distributed association rule mining, according to the ideas in this chapter, the obtaind overall rules not only need to meet the given overall supporting degree and confidence degree, but also need to meet the supporting number of sites, which means that the overall rules must at least be supported by n (n<the number of distributed sites) sites, i.e. overall rules must be the local/partial association rules in n sub-sites. This algorithm is based on this kind of thought to first select possible overall association rules from RK -tree.

5.3 the description of RK –tree algorithm

algorithm : RK -tree algorithm :

Input : the distributed database DB{DB_1 , DB_2 , ... , DB_n} ; the threshold of the minimum supporting degree min-sup ; the threshold of the minimum confidence degree min-confi; the number of the minimum sites : N

Output : the mined overall rule

```
( 1 ) for all sites do
{
  (2) each sub-point adopts Aprior algorithm local-mining to generate local rule set R(i);
  (3) site point i sends each rule's attribute information generated from rule set R(i) and R(i) to the center;
}
( 4 ) establishing rule knowledge base D; storing all the rule set R(i) transferred from each site in D ;
( 5 ) for i=1 to n do
  {
  (6) mapping all the rules in D  the rule tree, generating RK –tree in which the appearing times n ( i ) of each rule are recorded ;
}
( 7 ) for ( reading each branch of the tree one by one until the whole tree is read )
( 8 ) if  n ( i ) < N then delete this branch
( 9 ) else scanning the database again, getting the information of each sub-sites in the left branch rule in the RK –tree; through the information calculating the
```

```
supporting degree RK -tree and the believing degree $conf_i$ ;
    {
    ( 10 ) if  $sup_i < min-sup$  then delete this branch
    ( 11 ) else
    {
    ( 12 )      if  $conf_i < min-conf$  then delete this branch
    ( 13 )      else  this rule is the overall rule, output
          }
    }
    ( 14 ) end
```

5.4 The example of RK-tree algorithm

In order to make readers understand more deeply, we explain a simple example with RK -tree algorithm specifically.

Eg. 3.1 in table 3.1, we give the mining rules of three sub-sites storing in the rule database D. We use RK -tree algorithm to these rules to find the overall rule. (suppsoig the transaction numebr that each site transfers is : site 1 : 25000 ; site 2 : 27500 ; site 3 : 30000)

rule of station 1

Num	rule
1	a=5 □b=3□ c=2 □d=1
2	b=5□c=3□d=4
3	C=4□d=5□ a=2□b=3
4	b=2□a=5□d=1□c=4
5	a=3□c=6□d=5□b=4

rule of station 2

Num	rule
1	b=2□ c=4 □a=3
2	A=5□b=3□c=2□d=3
3	B=5□c=3□d=4
4	C=4□d=5□a=2□b=3
5	B=1□a=3□d=2□c=1

rule of station 3

Num	rule
1	C=4□d=5□a=3□b=3

2	B=2□a=5□d=1□c=3
3	A=5□b=3□c=2□d=1
4	B=5□c=3□d=4
5	D=2□c=4□b=5

Tab.5.1 the rule of the sub-site in rule knowledge base D

(1) first, we construct the rule of the site 1 in rule knowledge base D into the following RK –tree according to the constructing method mentioned above. establish the root node of the "null"' of the tree, and each rule constitutes a branch according to the order that the former piece - > the latter piece. The latter one is the leaf node of the branch, and record the appearing times of this rule at the place of leaf node.

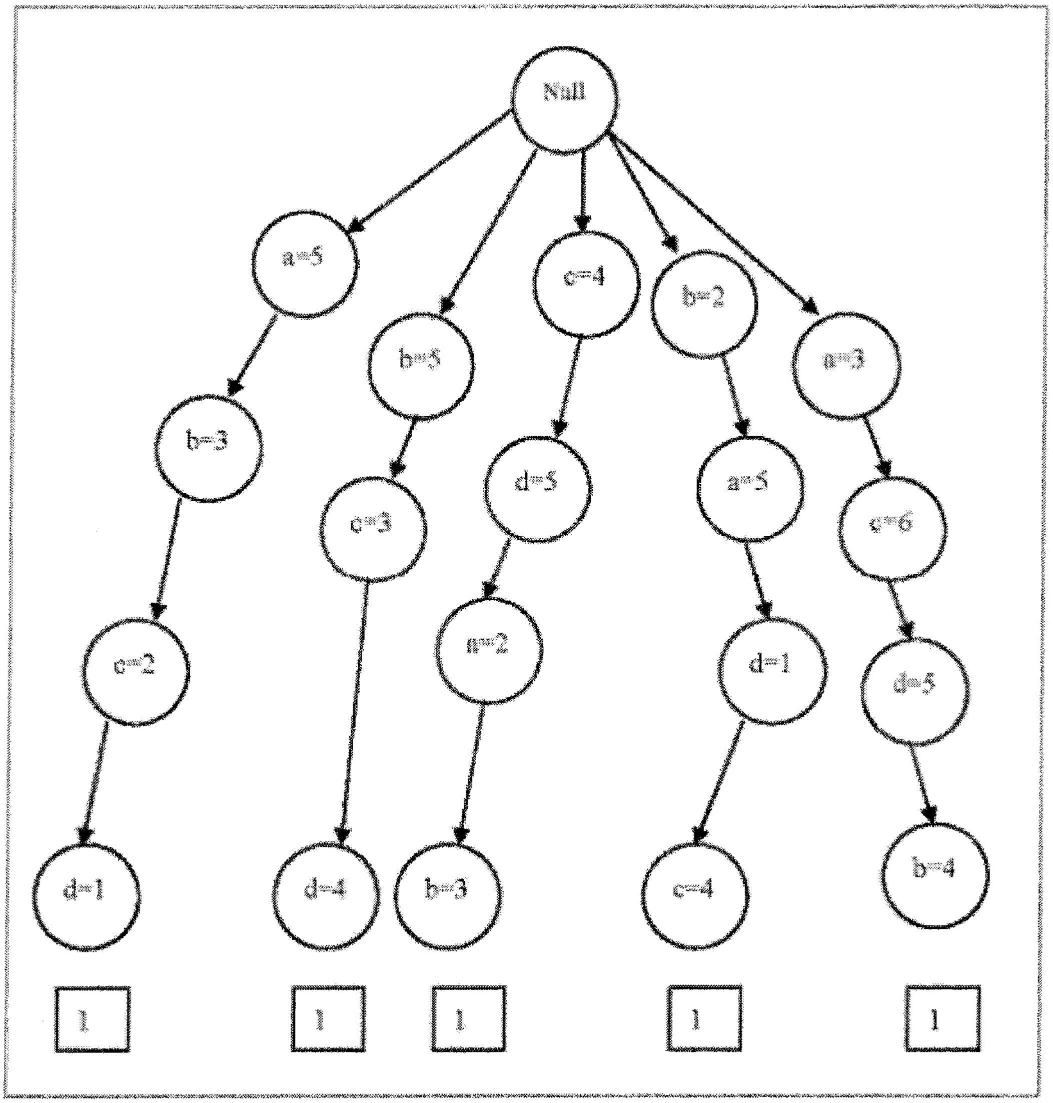

Figure 5.1 RK –tree formed by rule of the site 1

(2) Then, begin to read each rule from the site 2. if it is the same as the existing rules in the tree, then add one to the appearing times of this rule; or else, establish another branch. According to this method read all rules. Fig. 4.2 is the constructed RK -tree

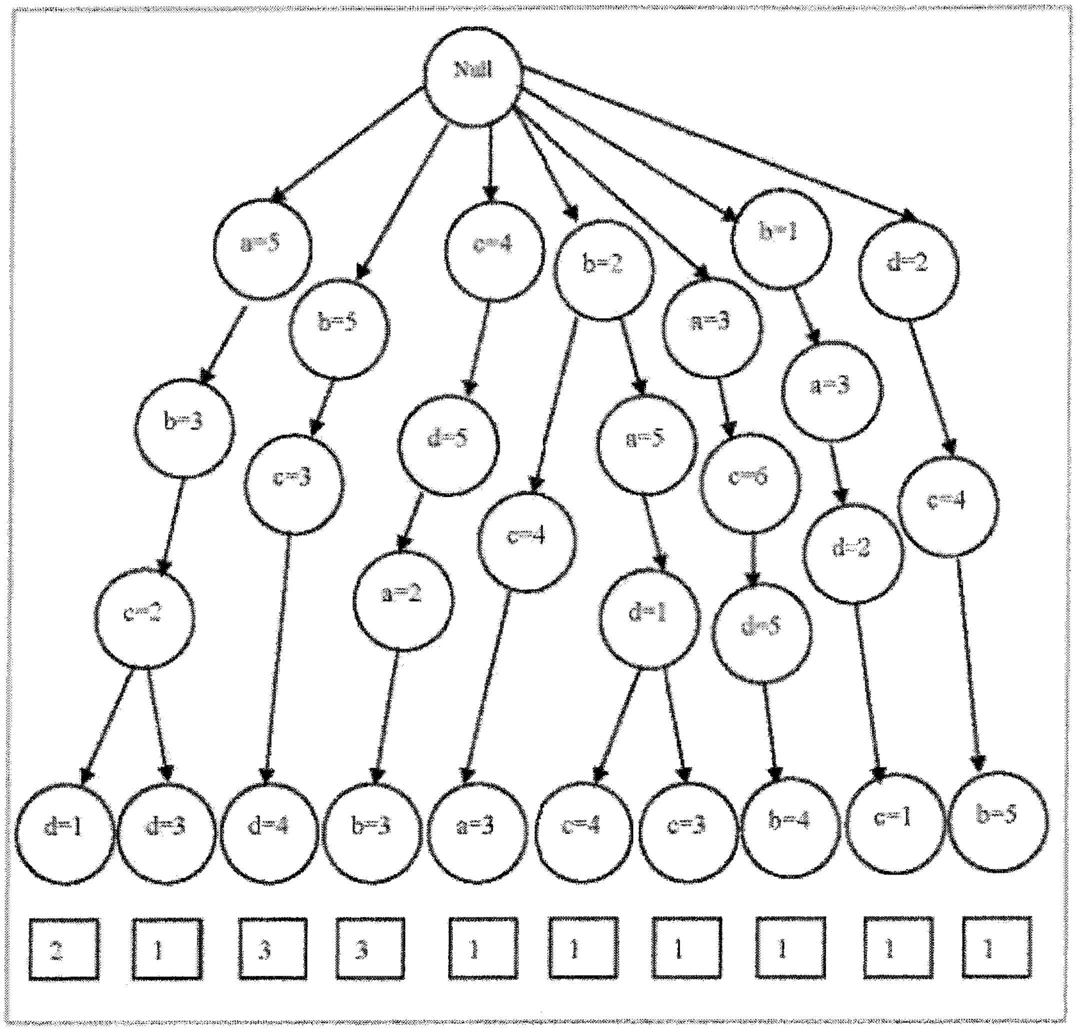

Figure5.2 the constructed RK -tree

(3) here, we regulate N=2 , minsup=0.1,minconf=0.7 then

□{a=5□b=3□c=2 }=>d=1 中 , n=2=N , keep this branch ;

□{a=5□b=3□c=2}=>d=3 中 , n=1<N , delete this branch ;

□{b=5□c=3}=> d=4 中 , n=3>N , keep this branch ;

□{c=4□d=5□a=2 }=> b=3 中 , n=3>N , keep this branch ;

□{b=2□c=4}=> a=3 中 , n=1<N , delete this branch ;

□{b=2□a=5□d=1}=> c=4 中 n=1<N , delete this branch ;,

□{b=2□a=5□d=1}=> c=3 中 , n=1<N , delete this branch ;

□{a=3□c=6□d=5 }=> b=4 中 , n=1<N , delete this branch ;

□{b=1□a=3□d=2 }=> c=1 中 , n=1<N , delete this branch ;

□{d=2□c=4 }=>b=5 中 , n=1<N , delete this branch ;

(4) scanning the database of these three stations again to find the information of the left branch in the three sub-sites, the findings are as follows :

rule □{a=5□b=3□c=2 }=>d=1 information				
	sup_i	$conf_i$	transaction supporting number	transaction number including the former piece
station 1	0.17	0.8	4250	5312
station 2	0.03	0.45	825	1833
station 3	0.15	0.67	4500	6716

rule□{b=5□c=3}=> d=4 information				
	sup_i	$conf_i$	transaction supporting number	transaction number including the former piece

station 1	0.15	0.63	3750	5952
station 2	0.08	0.6	2200	3667
station 3	0.11	0.73	3300	4520

rule□{c=4□d=5□a=2 }=> b=3 information				
	sup$_i$	conf$_i$	transaction supporting number	transaction number including the former piece
station 1	0.13	0.75	3250	4333
station 2	0.09	0.64	2475	3867
station 3	0.18	0.85	5400	6353

Tab.5.2 the basic information of rules

According to Tab.4.2, we can easily calculate the supporting degree and the confidence degree of the left rules, as follows ：

First , according to the definition 4.1 , D= $D^1+D^2 +\ldots \ldots+D^n$, there is

Adding the transaction number of three sites :

25000+27500+30000=82500

minsup=0.1,minconf=0.7 则 then

□{a=5□b=3□c=2 }=>d=1 ：

Support= (4250+825+4500) ÷82500=0.116 > minsup

confidence= (4250+825+4500) ÷ (5312+1833+6716) =0.69<minconf

delete this branch ;

□{b=5□c=3}= > d=4 :

Support= (3750+2200+3300) ÷82500=0.11 > minsup

confidence= (3750+2200+3300) ÷ (5952+3667+4520) =0.654<minconf

delete this branch ;

□{c=4□d=5□a=2 }= > b=3 :

Support= (3250+2475+5400) ÷82500=0.13 > minsup

confidence= (3250+2475+5400) ÷ (4333+3867+6353) =0.764 > minconf

this rule is the overall rule

through calculating, we know the rule {c=4□d=5□a=2 }= > b=3 is the overall rule.

5.5 comparison between RK –tree algorithm and other algorithms

(1)The amount of communication. In each iteration step, in order to mine overall K frequency maxterm set, FDM algorithm needs to exchange candidate supporting numbers of K frequency maxterm set and local maxterm set. The times and amount of communication are all large. And complicated communication agreement is needed. On the other hand, RK-tree just needs to come and go twice between the main site point and each local site point. The amount of communication is very small.

(2)High autonomy of sub-site point, which is good for reducing the complexity of distributed mining controlling. In each iteration step's end, each site point of FDM algorithm should be synchronous. They should wait to collect the supporting number of candidate frequency maxterm set in other site points. On the other hand, RK-tree algorithm is independently in each site point's mining. Each site point can do off-line mining. And at the same time the amount and times of communication is smaller. Therefore, the efficiency of RK-tree is better than FDM.

(3) Reduce the workload of algorithm research and development. The first step of RK-tree algorithm—each site point found out that the task of local frequency itemset is an undistributed frequency itemset mining task, therefore, it can be realized by frequency itemset mining algorithm at present. And RK-tree can supplement on this basis, which can realize the combination of algorithm knowledge, overall knowledge test and distributed mining control, etc.. in this way, the workload of research is reduced. What's more, currently, many effective frequency mining algorithms have been raised. With these excellent algorithm, RK-tree algorithm's local frequency

mining task is good to improve the efficiency of the whole algorithm.

(4)This algorithm is simple. It just need twice knowledge combination

5.6 The comparison of optimized experimental results

Experimental parameters:

Experimental test data: database of customers of Da Hua supermarket

Item Entry = 77120 ;

PC : Dual 2 Core Processor 1.6 GHz , memory 2G

Agent number : 3

Center number : 1

Network connections: six cable, kilomega connection

The structure of distributed network topology is as follows:

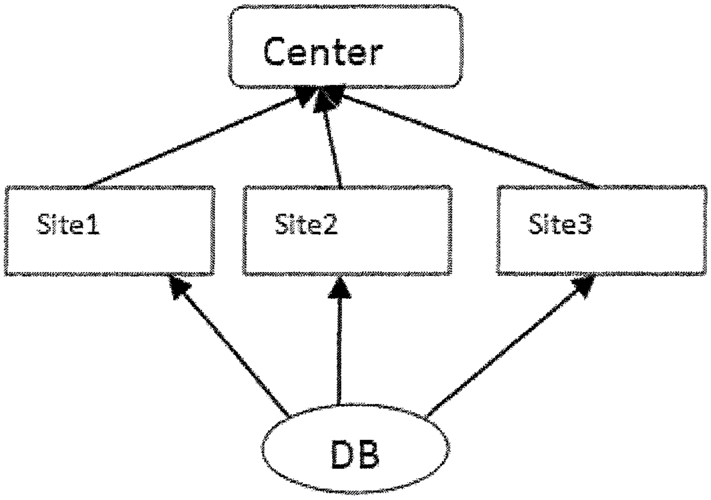

Figure 5.3 The structure of distributed network topology

Time :

FDM : 47 minutes 32 seconds

RK - Tree : 32 minutes 12 seconds

Optimized performance : 31.9%

In order to further analyze the accountability of this algorithm, we test a group of experimental data according to the objective entry number in different database. The data objectives are as follows:

Object DB Entry Number :

512	1024	2048	4096	8192	10240	12240	15480	19900	24320

The comparison results of the experiments:

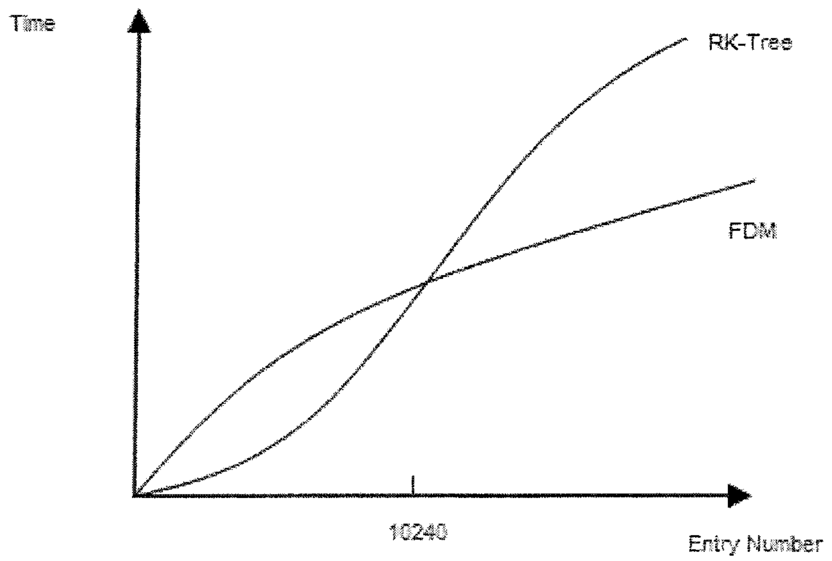

Figure 5.4 The comparison results of the experiments

By comparing the above experimental results, we discover that RK-tree is better than FDM when Entry is bigger.

On 10240 entry node, there is one point of intersection of RK-tree algorithm and FDM algorithm. Of course, according the different complexity of different data, as well as the difference of network hardware and optimization realization, there may be some differences in points of intersection.

Reasons:

As for DB with less data objective, RK-tree needs to divide overall database into local database, and send concrete information on internet. In this way, large amount of time is wasted in sending information and rearrange them. However, as the scale of data is increasing, RK-tree's efficiency in communication and mining which is concluded in 5.5 is more obvious.

5.7 Summery

This chapter puts forward a distributed association rule mining algorithm based on multi-agent——RK-tree algorithm. Through the above example, we can see that it greatly reduces the network communication cost to use such algorithm to do distributed association rule mining. It can calculate the supporting degree and the believing degree of each rule and get high-effective and reliable mining results that users need only by mapping the mining results of the sub-sites to a association tree and through the basic information of each rule. In the whole mining process, it only needs communication two times in sub-stations and the main station, and data transferring is less(only the mining results of the sub-sites needs transferring), and the requirement to the network bandwidth is low; the mining efficiency is high; the security and privacy of data have been guaranteed. As a distributed association rule mining algorithm, RK-tree has good practicability.

CHAPTER6
CONCLUSION

The development of information technology and network technology promotes the information process of enterprises and society. Most kinds of mass and useful data are stored in the distributed places of physical position, which brings challenge to the traditional centralized data mining. Therefore, the distributed data mining comes into being as time requires. The inhomogeneous and the diversity of data are one of the different problems in distributed data mining. It is another different problem in distributed data mining that how to design and reconstruct centralized mining algorithm to adapt the requirement of distributed data mining.

This thesis has done some research in these two aspects, and put forward some novel thoughts and good ideas. The main work includes the following aspects :

(1) On the basis of analyzing the model of the centralized distributed data mining system, we apply multi-agent technology to the distributed data mining system. The self-adaptability and intelligence of multi-agent provide a novel and effective solving scheme.

(2) The thesis puts forward a distributed association rule algorithm based on multi-agent, and analyzes, compares this algorithm. Meanwhile we use concrete example to prove the advantage of this algorithm. the main characteristics of RK-tree are as follows:

□few amount and times of communication is good for improving the mining efficiency. Each sub-site point mine local frequency itemset from local database, and then sends local frequency itemset to the main site point. The main site point combines local frequency itemset sent by sub-site points, and send the overall frequency itemset to each sub-site point. After this, each sub-site point will make supporting counting statistics of these overall frequency itemsets and send it back to the main site point. At last, the main site point will combine results sent by each sub-site point and get overall frequency itemset and overall association rules. This algorithm just needs three times communication between the main site point and sub-site points.

□ the sub-site point has high autonomy, which is convenient for reduce the controlling complexity of distributed mining. RK-tree algorithm is independent in mining in each sub-site point. Each sub-site point can do off-line mining, which will improve the efficiency of data mining.

□ reduce the research on algorithm and workload of development. The first step of RK-tree algorithm—each sub-site point discover task of local frequency itemset, is a non-distributed frequency itemset mining task. Therefore, it can be realized by frequency itemset mining algorithm. And RK-tree algoeithm can be further expanded on this basis, which will realize the comibination of algorithms, overall knoledge identification and distributed mining control, etc.. In this way, the development workload of research on algorithem will be decreased. What's more, currently, many effective frequency module mining algorithems have been rasied. These excellent methods can be used to realize RK-tree algorithm's local frequency itemset mining task.

□ this algorithm is simple, which requires only twice knowledge combination.

Although our research work has got some valuable achievements, there are still some shortages, and this is the research work we are going to do next step:

(1) It is a complex work with much workload to construct a distributed data mining system with complete functions. We just put forward the structure of mining system with main functions. There are many function modules needing to be perfected, such as humanity users' interface, the secure measure of multi-agent, and so on.

(2) The research and realization of distributed data mining algorithms of other types. In this thesis, we only research the association rule distributed mining algorithm. In the future work, we will discuss about other distributed mining algorithms like classification, and so on.

REFERENCE

[1] Benjamin A. Lieberman. *Information architecture essentials*, Part 6: Distributed data mining

[2] J.Han and M.Kamber .*Data Mining:cConcept and Techniques* ,Morgan Kaufmann 2001 . p649---p685

[3] J.Han and M.Kamber .*Data Mining:cConcept and Techniques* , Morgan Kaufmann 2001 . p1---p9

[4] *Cases on Database Technologies And Applications* , Idea Group Publishing (March 20, 2006) | ISBN: 1599043998

[5] J.Han and M.Kamber .*Data Mining:cConcept and Techniques* , Morgan Kaufmann 2001 . p105---p156

[6] *Distributed and Ubiquitous Data Mining* , From DDMWiki , http://www.umbc.edu/ddm/wiki/Distributed_and_Ubiquitous_Data_Mining.

[7] *Distributed and Ubiquitous Data Mining* , From DDMWiki , http://www.umbc.edu/ddm/wiki/Distributed_and_Ubiquitous_Data_Mining

[8] William K. Cheung, Xiao-Feng Zhang, Ho-Fai Wong,et al. *Service-Oriented Distributed Data Mining,* IEEE Internet Computing, Volume 10 , Issue 4 (July 2006), Pages: 44 - 54,2006.

[9] Piatetaky-Shapior G.*Dataming and Knowledge Discovery in Bussiness Database.* ISMIS'96,56—67

[10] Fan Jianhua and Li Deyi, *An Overview of Data Mining and Knowledge Discovery* , Department of Computer Science,Communication Engineering Institute Nanjing 210016,P.R.China,1998.4

[11] J.Han and M.Kamber. *Data Mining:cConcept and Techniques* , Morgan Kaufmann 2001 . p21---p27

[12] (America)Ryszard S.Michalski,Ivan Bratko,Miroslav Kubat, etc., Zhu Ming, etc. Tra. *Machine Study and Data Mining: Method and Application* [M]. Beijing: Electronic Press, 2004.

[13] Clark Glymour , David Madigan , Daryl Pregibon , Padhraic Smyth , *Statistical inference and data mining. Communications of the ACM* , Volume 39 , Issue 11 (November 1996) , Pages: 35 - 41 , 1996.

[14] Tom M.Mitchell. *Machine Study*. Machinery Industry Press , 2003.1

[15] Quinlan J R. *Induction of decision tree*. Machine Learning,1986

[16] Hongjun Lu, R. Setiono, and Huan Liu, "*Effective data mining using neural networks,*" Knowledge and Data Engineering, IEEE Transactions on, Volume: 8 Issue: 6, pp.957 -961.1996.

[17] McCulloch W S,Pitts W. *A logical calculus of the ideas immanent in nervous activity* .1943

[18] Donald O. Hebb's, *The Organization of Behavior*.1949

[19] J . H . Holland , *Adaptation in natural and artificial systems*[M] , Ann arbor: University of Michigen press,1975

[20] McNeill, Daniel, and Freiberger, Paul, "*Fuzzy Logic: The Discovery of a Revolutionary Computer Technology*", Simon and Schuster,1992. ISBN 0-671-73843-7.

[21] Pawlak Z. *Rough Set Theory and its Applications to Data Analysis* 1998(7)

[22] Pawlak,Z. *Rough Sets.*International Journal of Information and Computer Science,1982,11

[23] Kurt Thearling, Barry Becker, Dennis DeCoste, Bill Mawby, Michel Pilote, and Dan Sommerfield.*Visualizing Data Mining Models*. Information Visualization in Data Mining and Knowledge Discovery.2001.

[24] *Data mining*, From Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Data_mining

[25] Groth R. *Data Mining: Building Competitive Advantage* [M]. New Jersey: Prentice Hall.1999.1-10.

[26] Park J S, Ashak Savasere. *An Effective Hash-based Algorithm for Mining Association Ulles* [M]. Heraklion. Proceedings of the ACM SIGMOD. Greece: Ctete

Press. 1995,3:175-186.

[27] Han J.*Mining Knowledge at Multiple Concept Levels* [M]. In: Proceeding of the International Conference on Information and Knowledge Management(CIKM'95).Maryland:Baltimore Press,1995,8:19-24.

[28] YongJian Fu.*Distributed Data Mining : An Overview* [B].IEEE TCDP newsletter,Spring,2001

[29] Park, B. & Kargupta, H. *Distributed Data Mining: Algorithms, Systems, and Applications.* In N. Ye (Ed.), The Handbook of Data Mining. (pp. 341-358). Lawrence Erlbaum Associates. 2003.

[30] *Data Consistency Explained*, by Recovery Specialties

[31] H. Toivonen, *Discovery of Frequent Patterns in Large Data Collections*,Tech. Report A-1996-5, Dept. Computer Science, Univ. Of Helsinki, 1996

[32] C. Bishop, *Neural Networks for Pattern Recongition*, Oxford Univ. Press,Oxford, England, 1995

[33] D.E.Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning,* Addison-Wesley, Reading, Mass. 1989

[34] J.R.Quinlan, *C4.5: Programs for Machine Learning, MorgenKaufmann,*
San Francisco, 1993

[35] R.Agrawal, et al. *Parallel mining of association rules.* IEEE Transactions on knowledge and data engineering, 1996, 8(6), 962~969

[36] J.S.Park, et al. *Efficient parallel data mining for association rules*.Proc .Fourth int' l conf .Information and Knowledge management, Baltimore,Nov.1995.

[37] D.W.Cheung, et al. *Efficient mining of association rules in distributed databases.* IEEE Transactions on knowledge and data engineering, 1996.8(6), 910~921

[38] D.W.Cheung, et al. *A fast distributed algorithm for mining association rules.* Proc. Of 1996 Int'l Conf. On Parallel and Distributed Information Systems (PDIS'96), Miami Beach, Florida, USA, Dec, 1996

[39] John Shafer, Rakesh Agrawal, and Manish Mehta: *SPRINT: A Scalable*
Parallel Classifier for Data Mining. Proceedings of the 22nd VLDB Conference Mumbai (bombay), India, 1996

[40] Decker, H. *Abduction for knowledge assimilation in deductive databases.* Computer Science Society, Proceedings., XVII International Conference of the Chilean. 1997.

[41] Jeff Kramer,*Distributed software engineering.* IEEE Computer Society Press Los Alamitos, CA, USA .1994.

[42] *Common Object Request Broker Architecture*, From Wikipedia, the free encyclopedia

[43] A.Lazarevic and Z.Obradovic.Boosting *Algorithms for Parallel and Distributed Learning.* Distributed and Parallel Databases: An International Journal, Special Issue

on Parallel and Distributed Data Mining, 2:203–229, 2002.

[44] Kargupta H ,Park B ,Hershberger D , et al. *Collective Data Mining: A New Perspective Toward Distributed Data Mining.* In:Advances in Distributed and Parallel Knowledge Discovery ,AAAI/ MIT Press,2000.131～178

[45] Schuster and R. Wolff. *Communication-Effcient Distributed Mining of Association Rules.* Data Mining and Knowledge Discovery, 8(2), March 2004.

[46] T. Li, S. Zhu, and M. Ogihara. *Algorithms for Clustering High Dimensional and Distributed Data.* Intelligent Data Analysis Journal, 7(4), 2003.

[47] Jiangchun Song, Junyi Shen , *Research on the Architecture of Distributed Web Mining System* , Journal of Communication and Computer , 2006/01.

[48] GUO Li-ming , ZHANG Yan-zhen。*A Distributed Data Mining System Based on Multi-agent Technology* ,JOURNAL OF DONGHUA UNIVERSITY(ENGLISH EDITION) , 2006 23 (6)

[49] Jiang Wu-shan and Yu Ji-hui, *Distributed data mining on the grid* . Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on Volume 4, Issue , 18-21 Aug. 2005 Page(s):2010 - 2014 Vol. 4

[50] Krishnaswamy A. *Program slicing: an application of object-oriented program dependency graphs.* Technical Report, TR94-108, Department of Computer Science, Clemson University.

[51] Grossman R, Kasif S, Moore R, et al. *Report of three NSF Workshops on Mining Large, Massive, and Distributed Data*[M]. New York: AAAI Press, 1999.

[52] C Zhang, S Zhang , *Association rule mining: models and algorithms* - Lecture Notes In Artificial Intelligence; Vol. 2307, 2002

[53] Rakesh Agrawal, Tomasz imielinski, and Arun Swami. *Mining association rules between sets of items in large databases.* In Proceedings of the ACM SIGMOD Conference on Management of Data, Washington, D. C., May 1993.

[54] Jiawei Han , Micheline Kamber : *Data Mining Concepts and Techniques*.232~234.2001.

[55] Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; & Verkamo, A. *Fast discovery of associationrules.* In Advances in KDD. MIT Press. 1996.

[56] Jiawei Han , Micheline Kamber : *Data Mining Concepts and*

Techniques.227~240. 2001.

[57] A.Savasere , E.Omieeinski , andS.Navathe.*An efficient algorithm for mining assoeiation rules in large databases.*proceedings of the 21st Intemational Conference on Very Large Database , P.P.432 — 443 , SeP.1995.

[58] H.Mannila , H.Tivnoen , andA.Vekramo.*Efficient algorithm for discovering assoeiation rules*.AAAI Workshop on knowledge Discovery in databases , P.P.181 — 192 , Jul.1994.

[59] S.Brin , R.Motwani , J.D.Ullman , andS.Tsur *Dynamic itemset counting and implication rules for makret basket data*.In ACM SIGMOD Intenrational Conefrence on the Management of Data , P.p.255 — 264 , May1997.

[60] Fong J, Wong, H.K, Huang, S.M, *Continuous and incremental data mining association rules using frame metadata model Knowledge-Based Systems*,March,2003,16(2) p91 — 100

[61] Tung, Anthony K.H. Lu Hong jun, Han Jiawei, Feng Ling. *Efficient mining of intertransaction association rules.* IEEE Transactions on Knowledge and Data Engineering. 15(1),2003,p 43-56

[62] Wang De-Xing, Hu Xue-Gang,Liu Xiao-Ping, Wang Hao, Guo Jun. *Research on model of association rules mining with added-newly measure criteria.* Proceedings of 2004 International Conference on Machine Learning and Cybernetics.2004.vo1 2:1187-1190

[63] Aggarwal, Charu, C. *Fast algorithms for online generation of profile association rules.*2002,14(5):1017-1028

[64] Tsai. Cheng-Fa,Lin Yi-Chau,Chen Chi-Pin.*A new fast algorithms for mining association rules in large databases.* Proceedings of the IEEE International Conference on Systems,Man and Cybernetics.2002,vo1(7):251-256

[65] R.Agrawal, R.Srikant. *Fast algorithms for mining association rules.* Pros. 20thint' 1 conf. very large databases, Santiago, Chile, Sept .1994,487-499.

[66] J. S. Park , et al .*Using a hash-based method with transaction trimming for mining association rules* .IEEE Transactions on knowledge and data engineering,1997,9(5),813-825.

[67] A. Savasere, E, Omiecinski and S. Navathe .*An efficient algorithm for mining associateion rules.* Proceedings of the 21st international conference on very large databases, Zurich, Switzerland , Sept .1995, 432-444.

[68] Hannu Toivonen. *Sampling large databases for association rules*. Porccedings of the 22nd international conference on very large databases, Bombay, India,1996,134-145.

[69] R .Agrawal ,et al. *Parallel mining of association rules* .IEEE Transactions on knowledge and data engineering ,1996,8(6),962-969

[70] J. S. Park,et al. *Efficient parallel data mining for association rules* .Proc .Fourth int'1 conf. information and Knowledge management, Baltimore, Nov, 1995

[71] D. W. Cheng, et al. *A fast distributed algorithm for mining association rules in distributed databases.* IEEE Transactions on knowledge and data engineering.1996,8(6),910-921

[72] D. W. Cheng, et al. *A fast distributed algorithm for mining association rules* Proc .of 1996 Int'1 Conf.on Parallel and Distributed Information Systems (PDIS'96), Miami Beach, Florida, USA,Dec,1996

[73] S. Nestorov. *Mining Qualified Association Rules in Distributed Databases.* In Work-shop on Data Mining and Exploration Middleware for Distributed and Grid Computing,Minneapolis, MN, September 2003.

[74] A. Schuster and R. Wolff. *Communication-Effcient Distributed Mining of Association Rules.* Data Mining and Knowledge Discovery, 8(2), March 2004.

[75] Assaf Schuster, Ran Wolff, and Bobi Gilburd. *Privacy-Preserving Association Rule Mining in Large-Scale Distributed Systems*. In Proceedings of Cluster Computing and the Grid (CCGrid), 2004.

[76] M. Zaki. *Parallel and Distributed Association Mining: A Survey*. IEEE Concurrency,1999.

[77] R. Wolff, A. Schuster, and D. Trock. *A High-Performance Distributed Algorithm for Mining Associatio nRules.* InThe Third IEEE International Conference on Data Mining(ICDM’03), November 2003.

[78] V. S. Ananthanarayana, D. K. Subramanian, and M. N. Murty. Scalable, *Distributed and Dynamic Mining of Association Rules*. In Proceedings of HIPC’00, pages 559–566,Bangalore, India, 2000.

[79] Botia J,A., Garijo , J,R, and Skarmeta, A,F, (1998), "*A Generic Data Mining System: Basic Design and Implementation Guidelines*", in Workshop on Distributed Data Mining at the 4th Int. Conf. on Data Mining and Knowledge Discovery(KDD-98), New York, USA, AAAI Press

[80] Chatratichat, J , Darlington, J, Guo, Y, Hedvall, S, Kohler, M, and Syed, J,(1999), "*An Architecture for Distributed Enterprise Data Mining*", in Proc. of the 7th Int Conf. on High Performance Computing and Networking(HPCN Europe'99),Amsterdam, The Netherlands, Springer- Verlag LNCS 1593.

[81] Shoham Y , *Agent-oriented programming ,Artificial Intelligence* ,vol 60 ,51-92 , No.1 ,(1993) .

[82] McCarthy, John; Minsky, Marvin; Rochester, Nathan; Shannon, Claude (1955), *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, http://www.formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

[83] *Software agent*, From Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Software_agent

[84] Michael Wooldridge, *An Introduction to MultiAgent Systems*, John Wiley & Sons Ltd, 2002

[85] Alex Rogers and E. David and J.Schiff and N.R. Jennings. *The Effects of Proxy Bidding and Minimum Bid Increments within eBay Auctions,* ACM Transactions on the Web, 2007

[86] Nathan Schurr and Janusz Marecki and Milind Tambe and Paul Scerri et al. *The Future of Disaster Response:Humans Working with Multiagent Teams using DEFACTO,* 2005.

[87] Ron Sun and Isaac Naveh.*Simulating Organizational Decision-Making Using a Cognitively Realistic Agent Model*, Journal of Artificial Societies and Social Simulation

[88] Rana O F, Walker D W, Li Mao zhen , et al. *PaDDMAS: Parallel and distributed data mining application suite* [A]. In: Proc.14th Int Conf. on Parallel and Distributed Processing Symposium[C]. Cancun Mexico: IEEE, 2000. 387~392

[89] Stolfo SJ , Prodromidis A L , Tselepis S, et al. *JAM: Java agents for meta learning over distributed databases*. In: Hecherman D ,Mannila H ,eds. Proc. Third International Conference on Knowledge Discovery and Data Mining(KDD 97) , Newport Beach , California , USA , AAAI Press ,1997. 74~81