Data Mining and Knowledge Discovery in Database

Nainja Rikhi[1]

1Research Scholar & Department of CSE & Mewar University, Chittorgarh, Rajasthan, India

***Abstract*— Knowledge discovery and data mining have become areas of growing significance because of the recent increasing demand for KDD techniques, including those used in machine learning, databases, statistics, knowledge acquisition, data visualization, and high performance computing. The motive of mining is to find a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. This article provides real-world applications, specific data-mining techniques, challenges involved knowledge discovery. This paper also discusses relation between Knowledge and Data Mining, and Knowledge Discovery in Database.**

***Keywords*— Knowledge discovery in databases, Data mining, Analysis, Information.**

I. Introduction

The rapid evolution of huge amount of data has lead to need for automated extraction of useful knowledge from such a huge dramatic pace. To extract this useful information some technique is required to focus on aspects of finding understandable patterns that can be interpreted as useful or interesting knowledge. So, as shown in Fig. 1, Data Mining is iterative and interactive process of discovering valid, novel, useful, and understandable knowledge (patterns, models, rules etc.) in Massive databases.

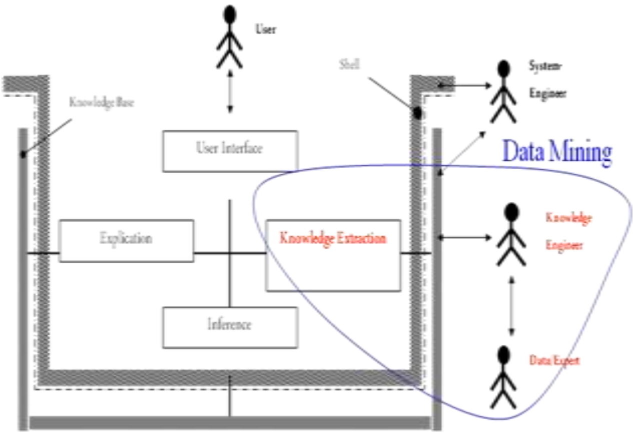

Fig. 1: Data Mining

These techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD) to discover strategic information hidden in very large databases.

The aim of this study is fourfold:

1) Provide an overview of existing techniques that can be used for extracting of useful information from databases.
2) Provide a feature classification scheme that identifies important features to study knowledge discovery and data mining software tools.
3) Investigate existing knowledge discovery and data mining software tools using the proposed feature classification scheme. These tools may be either commercial packages available for purchasing, or research prototypes developed at various universities.
4) Practical application issues of KDD and enumerate challenges for future research and development.

A. *Data, Information, Knowledge*

Data information knowledge are related to each other in such a way that one is depend on another, but then also data is rich and knowledge is poor, we understand this by following discussion:

We often see data as a string of bits, or numbers and symbols, or “objects” which we collect daily.

Information is data stripped of redundancy, and reduced to the minimum necessary to characterize the data.

Knowledge is integrated information, including facts and their relations, which have been perceived, discovered, or learned as our “mental pictures”.

Knowledge can be considered data at a high level of abstraction and generalization.

B. Impractical Manual Data Analysis

People gathered and stored so much data because they think some valuable assets are implicitly coded within it. Raw data is rarely of direct benefit. Its true value depends on the ability to extract information useful for decision support.

So, how to acquire knowledge for knowledge-based systems remains as the main difficult and crucial problem as shown in Fig. 2.

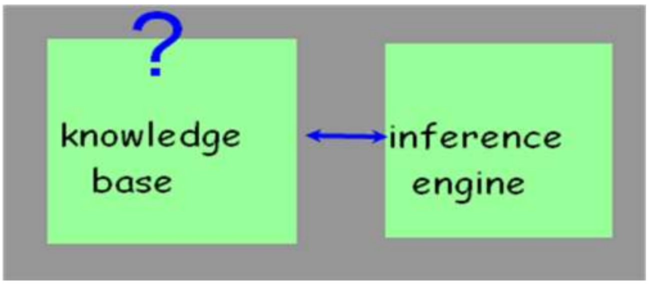

Fig. 2: Data Mining

Tradition: via knowledge engineers
New trend: via automatic programs

So, here need some intelligent system that has less consumption of time and very fast response time.

II. KDD, DATA MINING, AND RELATION TO OTHER FIELDS

This part provides an introduction into the area of knowledge discovery, data mining and serves as background explanation for our feature classification scheme.

KDD is the automatic extraction of non-obvious, hidden

knowledge from large volumes of data while data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data. The distinction between the KDD process and the data mining step is a central point of this paper. The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporating appropriate prior knowledge, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data.

The unifying goal is extracting high-level knowledge from low-level data in the context of large data sets. KDD overlaps with machine learning and pattern recognition in the study of particular data mining theories and algorithms: means for modelling data and extracting patterns. Hence data mining is just one step in the overall KDD process.

The above implies that we can define quantitative measures for evaluating extracted patterns. In many cases, it is possible to define measures of certainty or utility.

III. BASIC DEFINITIONS

We define KDD as Knowledge Discovery in Databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

Data Mining is a step in the KDD process consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data.

KDD Process is the process of using the database along with any required selection, pre-processing, sub sampling, and transformations of it; to apply data mining methods (algorithms) to enumerate patterns from it; and to evaluate the products of data mining.

The data mining component of the KDD process is concerned with the algorithmic means by which patterns are extracted and enumerated from data. The overall KDD process includes the evaluation and possible interpretation of the "mined" patterns to determine which patterns may be considered new "knowledge." The KDD process also includes all of the additional steps described in Section 4.

IV. THE KDD PROCESS

Knowledge discovery in databases (KDD) is the process of discovering useful knowledge from a collection of data. This widely used data mining technique is a process that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results.

It does this by using data mining methods (algorithms) to extract (identify) what is deemed knowledge, according to the specifications of measures and thresholds, using a database along with any required preprocessing, subsampling, and transformations of that database.

A. An Outline of the Steps of the KDD Process

The KDD process is interactive and iterative, involving numerous steps with many decisions being made by the user. The overall process of finding and interpreting patterns from data involves the repeated application of iterate steps as shown in Fig. 3.

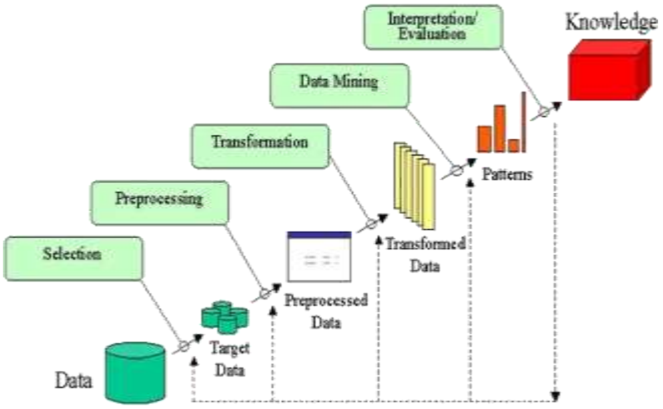

Fig. 3: An overview of the steps comprising the KDD process

Steps involved in the entire KDD process are:

1. Developing an understanding of the application domain, the relevant knowledge, and the goals of the end-user.
2. Creating a target data set on which discovery is to be performed.
3. Data cleaning and pre-processing includes removal of noise or outliers, collecting necessary information, Strategies for handling missing data fields.
4. Data reduction and projection includes finding useful features to represent the data and using dimensionality reduction or transformation methods.
5. Choosing the data mining task.
6. Choosing the data mining algorithm(s).
7. Data mining step includes searching for patterns or a set of such representations as classification rules or trees, regression, clustering, and so forth.
8. Interpreting mined patterns, possibly return to any of steps 1-7 for further iteration. This step can also involve visualization of the extracted patterns/models or data.
9. Consolidating discovered knowledge incorporating this knowledge into another system for further action to interested parties.

Tile KDD process can involve significant iteration and may contain loops between any two steps. Most previous work on KDD has focused on step 7 the data mining. We now focus on the data mining component, which has by far received the most attention in the literature.

V. THE DATA MINING STEP OF THE KDD PROCESS

Data mining process is used to extract information from a data set and transform it into an understandable structure for further use as shown in Fig. 4. The data mining component of the KDD process often involves repeated iterative application of particular data mining methods. It is achieved by using application domain like prior knowledge, user goals etc. to create target dataset that will be used in data mining algorithms that interpret, evaluate and visualize patterns and manage discovered knowledge.

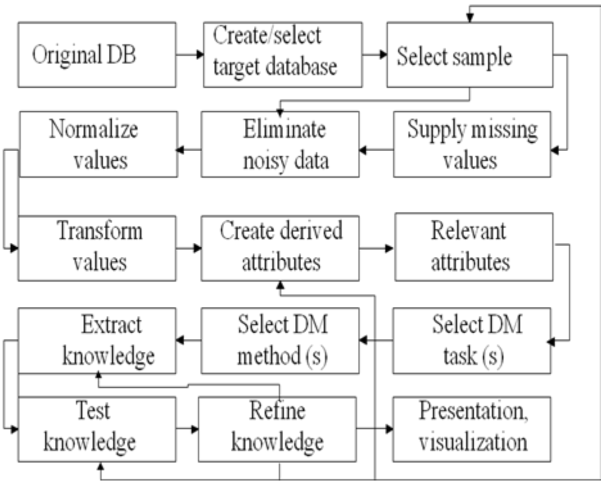

Fig. 4: Data mining process

The knowledge discovery goals are defined by the intended use of the system. We can distinguish two types of goals: Verification, where the system is limited to verifying the user's hypothesis, and Discovery, where the system autonomously finds new patterns. In this paper we are primarily concerned with discovery-oriented data mining. The goals of prediction and description are achieved via the following primary data mining methods:

1) *Classification:* learning a function that maps (classifies) a data item into one of several predefined classes.

2) *Regression:* learning a function which maps a data item to a real-valued prediction variable and the discovery of functional relationships between variables.

3) *Clustering:* identifying a finite set of categories or clusters to describe the data. Closely related to clustering is the method of probability density estimation which consists of techniques for estimating from data the joint multi-variant probability density function of all of the variables/fields in the database.

4) *Summarization:* finding a compact description for a subset of data, e.g., the derivation of summary or association rules and the use of multivariate visualization techniques.

5) *Dependency Modeling:* finding a model which describes significant dependencies between variables (e.g., learning of belief networks).

6) *Change and Deviation Detection:* discovering the most significant changes in the data from previously measured or normative values.

VI. DATA MINING TECHNIQUES

Data mining adopt its technique from many research areas, including statics machine learning, database systems, rough sets, visualization and neural networks.

A. Statistical Approach

Statistical models are built from a set of training data. Many statistical tools have been used for data mining including, Bayesian network, correlation analysis, regression analysis and cluster analysis. For example simple Bayesian network for traffic jam problem. In the Bayesian network nodes represents states or variable while edges represents dependencies between nodes. From Fig. 5 we can see that rush hour, bad weather or accident affect the traffic which in turn causes traffic jam.

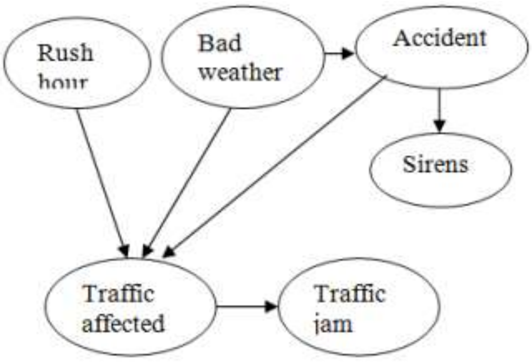

Fig. 5: Example of an unacceptable low-resolution image

B. Machine Learning Approach

The most common machine learning methods used for data mining include conceptual learning, inductive concept learning and decision tree induction. By following the path from root to leaf node an objects class can be determine by decision tree. Decision trees are induced from the training set and decision trees give classification rules. A simple decision tree is given in Fig. 6; it determines the car’s mileage from its size, transmission type and weight. The leaf nodes are in square boxes.

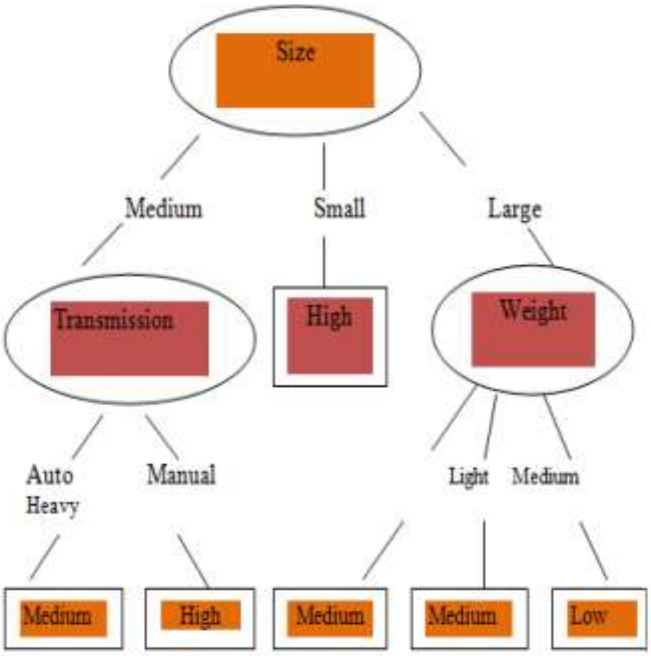

Fig. 6: A simple decision tree

From decision tree we can conclude, for example, large size; heavy weight car will have low mileage. Nodes represent three classes of mileage.

VII. DATA MINING TOOLS

Data mining is not all about the tools or database software that you are using. You can perform data mining with comparatively modest database systems and simple tools, including creating and writing your own, or using off the shelf software packages. Complex data mining benefits from the past experience and algorithms defined with existing software and packages, with certain tools gaining a greater affinity or reputation with different techniques.

For example, IBM SPSS®, which has its roots in statistical and survey analysis, can build effective predictive models by looking at past trends and building accurate forecasts. IBM InfoSphere® Warehouse provides data sourcing, preprocessing, mining, and analysis information in a single package, which allows you to take information from the source database straight to the final report output.

Now an entirely new range of tools and systems available, including combined data storage and processing systems.

You can mine data with a various different data sets, including, traditional SQL databases, raw text data, key/value stores, and document databases. Clustered databases, such as Hadoop, Cassandra, CouchDB, and Couchbase Server, store and provide access to data.

In particular, the more flexible storage format of the document database causes a different focus and complexity in terms of processing the information. Document databases that have a standard such as JSON enforcing structure, or files that have some machine-readable structure are also easier to process, although they might add complexities because of the differing and variable structure. For example, with Hadoop's entirely raw data processing it can be complex to identify and extract the content before you start to process and correlate it.

VIII. DATA MINING APPLICATIONS

Data mining application uses unstructured textual information and examines it in attempt to discover structure and implicit meanings hidden within the text. Compared with the kind of data stored in databases, text is unstructured, amorphous, and difficult to deal with. Through text mining, we can uncover hidden patterns, relationships, and trends in text. It is argued that the benefits of using text mining are to get to decision points more quickly.

Various fields use data mining technologies because of fast access of data and valuable information from vast amount of data. Data mining technologies have been applied successfully in many areas like marketing, telecommunication, fraud detection, and finance, medical and so on. Some of the application is listed below.

A. Financial Data Analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates the systematic data analysis and data mining. Here are the few typical cases: Design and construction of data warehouses for multidimensional data analysis and data mining. Loan payment prediction and customer credit policy analysis. Classification and clustering of customers for targeted marketing. Detection of money laundering and other financial crimes.

B. Retail Industry

Data Mining has its great application in Retail Industry because it collects large amount data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of increasing ease, availability and popularity of web. The Data Mining in Retail Industry helps in identifying customer buying patterns and trends. That leads to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in retail industry:

- Design and Construction of data warehouses based on benefits of data mining.
- Multidimensional analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer retention.
- Product recommendation and cross-referencing of items.

C. Telecommunication Industry

Today the Telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, Internet messenger, images, email, web data transmission etc. Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business. Data Mining in Telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list examples for which data mining improve telecommunication services as:

- Multidimensional analysis of telecommunication data.
- Fraudulent pattern analysis.
- Identification of unusual patterns
- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.
- Use of visualization tools in telecommunication data analysis.

D. Biological Data Analysis

Now a days we see that there is vast growth in field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is very important part of Bioinformatics. Following are the aspects in which Data mining contribute for biological data analysis:

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.

E. Other Scientific Applications

The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy etc. There is large amount of data sets being generated because of the fast numerical simulations in various fields such as climate, and ecosystem modelling, chemical engineering, fluid dynamics etc

IX. RESEARCH AND APPLICATION CHALLENGES

The major challenging issue in data mining arise from the complexity of a natural language itself. The natural language is not free from the ambiguity problem. One phrase or sentence can be interpreted in various ways, thus various meanings can be obtained. Although a number of researches have been conducted in resolving the ambiguity problem, the work is still immature and the proposed approach has been dedicated for a specific domain.

We outline some of the current primary research and application challenges for KDD:

1) *Larger databases:* Databases with hundreds of fields and tables, millions of records, and multi gigabyte size are quite commonplace and terabyte (1019-bytes) databases are beginning to appear. Methods for dealing with large data volumes include more efficient algorithms (Agrawal et al. 1996), sampling, approximation methods, and massively parallel processing (Holsheimer et al. 1996).

2) *High dimensionality:* Not only is there often a very large number of records in the database, but there can also be a very large number of fields (attributes, variables) so that the dimensionality of the problem is high. A high dimensional data set creates problems in terms of increasing the size of the search. In addition, it increases the chances that a data mining algorithm will find spurious patterns that are not valid in general. Approaches to this problem include methods to reduce the effective dimensionality of the problem and the use of prior knowledge to identify irrelevant variables.

3) *Over fitting:* When the algorithm searches for the best parameters for one particular model using a limited set of data, it may model not only the general patterns in the data but also any noise specific to that data set, resulting in poor performance of the model on test data. Possible solutions include cross-validation, regularization, and other sophisticated statistical strategies.

4) *Assessing statistical significance:* A problem (related to over fitting) occurs when the system is searching over many possible models. For example, if a system tests N models at the 0.001 significance level, then on average, with purely random data, N/1000 of these models will be accepted as significant. This point is frequently missed by many initial attempts at KDD. One way to deal with this problem is to use methods which adjust the test statistic as a function of the search, e.g., Bonferroni adjustments for independent tests, or randomization testing.

5) *Changing data and knowledge:* Rapidly changing (non-stationary) data may make previously discovered patterns invalid. In addition, the variables measured in a given application database may be modified, deleted, or augmented with new measurements over time. Possible solutions include incremental methods for updating the patterns and treating change as an opportunity for discovery.

6) *Missing and noisy data:* This problem is especially acute in business databases. U.S. census data reportedly has error rates of up to 20%. Important attributes may be missing if the database was not designed with discovery in mind. Possible solutions include more sophisticated statistical strategies to identify hidden variables and dependencies.

7) *Complex relationships between fields:* Hierarchically structured attributes or values, relations between attributes, and more sophisticated means for representing knowledge about the contents of a database will require algorithms that can effectively utilize such information. Historically, data mining algorithms have been developed for simple attribute value records, although new
8) techniques for deriving relations between variables are being developed.

9) *Understand ability of patterns:* In many applications it is important to make the discoveries more understandable by humans. Possible solutions include graphical representations, rule structuring, natural language generation, and techniques for visualization of data and knowledge. Rule refinement strategies can be used to address a related problem: the discovered knowledge may be implicitly or explicitly redundant.

10) *User interaction and prior knowledge:* Many current KDD methods and tools are not truly interactive and cannot easily incorporate prior knowledge about a problem except in simple ways. The use of domain knowledge is important in all of the steps of the KDD process Bayesian approaches use prior probabilities over data and distributions as one form of encoding prior knowledge. Others employ deductive database capabilities to discover knowledge that is then used to guide the data mining.

11) *Integration with other systems:* A stand-alone discovery system may not be very useful. Typical integration issues include integration with a DBMS (e.g. via a query interface), integration with

spreadsheets and visualization tools, and accommodating real-time sensor readings.

X. CONCLUSIONS

In this paper, we have discussed detail study of data mining with various studies like tasks, techniques, applications and challenging issues. A primary aim is to clarify the relation between knowledge discovery and data mining. We provided an overview of the KDD process and basic data mining methods. The implementation of data mining techniques will allow users to retrieve meaningful information from virtually integrated data. These techniques provide variety of applications for industries like retail, telecommunication, Bio-medical etc. These tools predict future trends and behaviors, allowing business to make proactive and present knowledge in the form which is easily understood to human.

The focus has been given on fundamental methods for conducting data mining. The methods include natural language processing and information extraction. A brief review on application domains has been presented. The purpose of this section is to give an overview to a reader on how text mining systems can be used in real life. The paper also addressed the most challenging issue in developing data mining systems.

ACKNOWLEDGMENTS

We would like to acknowledge and extend my heartfelt gratitude to the Mr. Rudra Pratap Ojha and Ms. Santosh Upadhay for hosting this research, making available the data and valuable comments.

REFERENCES

[1] J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.

[2] Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

[3] Sholom M. Weiss and Nitin Indurkhya, "Predictive Data Mining: A Practical Guide", Morgan Kaufmann Publishers, 1998.

[4] Agrawal, Ft. and Psaila, G. 1995. Active Data Mining, In Proceedings of KDD-95: First International Conference on Knowledge Discovery and Data Mining, pp. 3-8, Menlo Park, CA: AAAI Press.

[5] Brachman, R. and Anand, T. 1996. The Process of Knowledge Discovery in Databases: A Human Centered Approach, in AKDDM, AAAI/MIT Press, 37-58.

[6] Buntine, W. 1996. Graphical Models for Discovering Knowledge, in AKDDM, AAAI/MIT Press, 59 82.

[7] Dzeroski, S. 1996. Inductive Logic Programming for Knowledge Discovery in Databases, in AKDDM, AAAI/MIT Press.

[8] Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, and Ft. Uthurusamy, 1996. Advances in Knowledge Discovery and Data Mining, (AKDDM), AAAI/MIT Press.

[9] Heckerman, D. 1996. Bayesian Networks for Knowledge Discovery, in AKDDM, AAAI/MIT Press, 273-306.

[10] Stolorz, P. et al. 1995. Fast Spatio-Temporal Data Mining of Large Geophysical Data.sets, In Proceedings of KDD-95: First International Conference on Knowledge Discovery and Data Mining, pp. 300-305, AAAI Press.

[11] Alex Freitas and Simon Lavington, "Mining Very Large Databases with Parallel Processing", Kluwer Academic Publishers, 1998.

[12] K. Jain and R. C. Dubes, "Algorithms for Clustering Data", Prentice Hall, 1988.

[13] V. Cherkassky and F. Mulier, "Learning From Data", John Wiley & Sons, 1998.

[14] Qingtian Han, Xiaoyan Gao, “Research of Distributed Algorithm Based on Usage Mining”, Knowledge Discovery and Data Mining.

[15] J. Cowie and Y. Wilks, Information extraction, New York, 2000.