Data Mining: Next Generation Challenges and Future Directions

Dipti Verma and Rakesh Nashine

Abstract—**Data mining is the process of discovering actionable information from large sets of data. Data mining, or knowledge discovery, has become an indispensable technology for businesses and researchers in many fields. The data mining methods such as clustering, association rules, sequential pattern, statistics analysis, characteristics rules and so on can be used to find out the useful knowledge, enabling such data to become the real fortune for decisions and development. This paper introduces the significance of the application of data mining in different areas, challenges its future directions. Finally, it is pointed out that the data mining technology is becoming more and more powerful.**

Index Terms—**Data mining concept, applications, challenges, future direction.**

I. Introduction

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

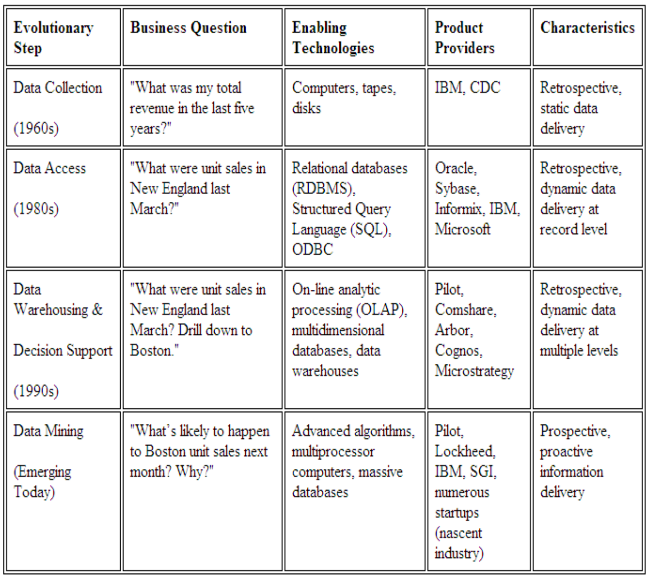
Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?"

II. Foundation of Data Mining

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Commercial databases are growing at unprecedented rates. A recent META Group survey of data warehouse projects found that 19% of respondents are beyond the 50 gigabyte level, while 59% expect to be there by second quarter of 1996.1 in some industries, such as retail, these numbers can be much larger. The accompanying need for improved computational engines can now be met in a cost-effective manner with parallel multiprocessor computer technology. Data mining algorithms embody techniques that have existed for at least 10 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods.



Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

In the evolution from business data to business information, each new step has built upon the previous one. For example, dynamic data access is critical for drill-through in data navigation applications, and the ability to store large databases is critical to data mining. From the user's point of

Manuscript received July 14, 2012; revised August 18, 2012.

Dipti Verma is with the G. H. Raisoni College of Engineering, Digdoh hills, Nagpur, Maharashtra (e-mail:diptidverma@ rediffmail.com).

Rakesh Nashine is with Citi Bank Mumbai, Maharashtra(e-mail: nashine.rakesh@gmail.com).

view, the four steps listed in Table 1 were revolutionary because they allowed new business questions to be answered accurately and quickly.

III. Data Mining Concept

Data mining uses a relatively large amount of computing power operating on a large set of data to determine regularities and connections between data points. Algorithms that employ techniques from statistics, machine learning and pattern recognition are used to search large databases automatically. Data mining is also known as Knowledge-Discovery in Databases (KDD).

Like the term artificial intelligence, data mining is an umbrella term that can be applied to a number of varying activities. In the corporate world, data mining is used most frequently to determine the direction of trends and predict the future.[1] It is employed to build models and decision support systems that give people information they can use.

Data Mining can be done through various types of data mining software. These can be simple data mining software or highly specific for detailed and extensive tasks that will be sifting through more information to pick out finer bits of information. For example, if a company is looking for information on doctors including their emails, fax, telephone, location, etc., this information can be mined through one of these data mining software programs. This information collection through data mining has allowed companies to make thousands and thousands of dollars in revenues by being able to better use the internet to gain business intelligence that helps companies make vital business decisions.

Before this data mining software came into being, different businesses used to collect information from recorded data sources. But the bulk of this information is too much too daunting and time consuming to gather by going through all the records, therefore the approach of computer based data mining came into being and has gained huge popularity to now become a necessity for the survival of most businesses. This collected information is used to gain more knowledge and based on the findings and analysis of the information make predictions as to what would be the best choice and the right approach to move toward on a particular issue.

The data mining also empowers companies to keep a record of fraudulent payments which can all be researched and studied through data mining. This information can help develop more advanced and protective methods that can be undertaken to prevent such events from happening. Buying trends shown through web data mining can help you to make forecast on your inventories as well[2]. This is a direct analysis, which will empower the organization to fill in their stocks appropriately for each month depending on the predictions they have laid out through this analysis of buying trends.

The data mining technology is going through a huge evolution and new and better techniques are made available all the time to gather whatever information is required. Web data mining technology is opening avenues on not just gathering data but it is also raising a lot of concerns related to data security. There is loads of personal information available on the internet and web data mining had helped to keep the idea of the need to secure that information at the forefront.

A. Benefits of Data Mining in Business

Data mining technology delivers two key business intelligence benefits:

1) It enables enterprises, regardless of industry or size, in the context of defined business objectives, to automatically explore, visualize and understand their data, and to identify patterns, relationships and dependencies that impact on business outcomes (such as revenue growth, profit improvement, cost containment, and risk management) - a descriptive function.
2) It enables relationships uncovered and identified through the data mining process to be expressed as business rules, or predictive models. These outputs can be communicated in traditional reporting formats (presentations, briefs, electronic information sharing) to guide business planning and strategy. Also these outputs, expressed as programming code, can be deployed or "hard wired" into business operating systems to generate predictions of future outcomes, based on newly generated data, with higher accuracy and certainty - a predictive function.

B. Need for Data Mining

Data mining is very important for the logistics management decision, which helps improve the efficiency of decisions, make right sale decisions, reduce inventory costs and analyse the market and trend[3].

1) Efficiency of decisions can be improved

Data mining can help mine the hidden valuable information from the large amount of data and make timely and accurate decisions.

2) Limiting the cost

Data mining system can integrate transport data with inventory data, and analyze data, which decide to carry out first shipment of goods to ensure the appropriate inventory. It can also send information of goods forecast and inventory directly to customers through the electronic data interchange system (EDI), so as to increase or reduce inventory regularly, thereby reducing the burden of their own.

3) Making right sale decision

Data mining helps identify the purchase model of target customers, predict the purchase trend, and mine the purchase power of customers, so as to provide real-time, dynamic and accurate decision support for the promotion.

4) Market and trend can be analyzed

Data mining tools and statistical models help analyze information about seasons of goods, transport amount, stocks, and varieties of goods in order to forecast the risk of goods, and then make decisions about the logistics operational management.

5) Psychology of customer better understood

Data mining helps in identifying the types of customer interested in making purchase and the psychology of the customer while making the purchase.

IV. The Process of Data Mining

The data mining process consists of four basic steps. question definition; data preparation and pre-processing; data mining; result interpretation and validation.

1) *Deciding Business Objective for data mining:* - The entire data mining process is driven by objectives, which are to be decided in advance. The business objective may depend upon the conclusions that will be drawn by data mining process.

The precise definition of the business needs is the first step in this direction. So the related knowledge of the object must be well learned before data mining and the goal of data mining must be definite.

2) *Data Preparation:* - It comprises of data selection, pre-processing and data transformation. Data selection is very important to data mining. The efficiency and the validity of data mining are directly affected by the data preparation. The preparation and transformation of the data sets are an important step in the process of data mining and it costs about 60% total mining time. The data preparation includes two key tasks. One is to choose appropriate input and output attributes according. The other is to identify the data and define the goal of the data mining.

3) *Data mining:* - The typical methods of data mining are as follows, classification analysis, clustering analysis, association analysis, and sequence analysis and time sequence, outlier analysis and so on. The effective data mining algorithm should be selected according to the specified research field. The results of data mining can be descriptive knowledge or predictive knowledge.

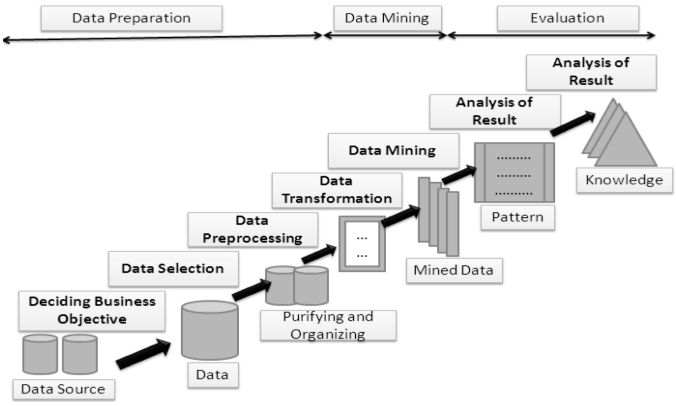


Fig. 1. Stepwise treatment model of typical Data Mining process

4) *Analysis of Result and its interpretation and validation:* - This step is for understanding the meaning of data mining results and its validity. The valuable knowledge and rules are achieved and expressed in an understandable mode. The extracted information is also assessed. Visualization can transform the derived useful knowledge into a format that is easy to understand, such as images or graphs[5].

5) *Assimilation of Knowledge:* - This step is the action to be taken to implement the knowledge gained by data mining. Knowledge discovered in the data mining is to be implemented in the business cycle [9].

V. The Scope of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database — for example, finding linked products in gigabytes of store scanner data — and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

Automated prediction of trends and behaviors. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events[10].

Automated discovery of previously unknown patterns. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high performance parallel processing systems, they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions.

Databases can be larger in both depth and breadth:

More columns: Analysts must often limit the number of variables they examine when doing hands-on analysis due to time constraints. Yet variables that are discarded because they seem unimportant may carry information about unknown patterns. High performance data mining allows users to explore the full depth of a database, without preselecting a subset of variables.

More rows: Larger samples yield lower estimation errors and variance, and allow users to make inferences about small but important segments of a population.

A recent Gartner Group Advanced Technology Research Note listed data mining and artificial intelligence at the top of the five key technology areas that "will clearly have a major impact across a wide range of industries within the next 3 to 5 years."2 Gartner also listed parallel architectures and data mining as two of the top 10 new technologies in which companies will invest during the next 5 years. According to a recent Gartner HPC Research Note, "With the rapid advance in data capture, transmission and storage, large-systems users will increasingly need to implement new and innovative ways to mine the after-market value of their vast stores of detail data, employing MPP [massively parallel processing] systems to create new sources of business advantage

The most commonly used techniques in data mining are:

1) *Artificial neural networks:* Non-linear predictive models that learn through training and resemble biological neural networks in structure.

2) *Decision trees:* Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).

3) *Genetic algorithms:* Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

4) *Nearest neighbor method:* A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called the k-nearest neighbor technique.

5) *Rule induction:* The extraction of useful if-then rules from data based on statistical significance.

Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms. The appendix to this white paper provides a glossary of data mining terms.

VI. Components Of The Data Mining

Data mining progress usually consists of three major components of the data preparation, data acquisition, and expression and interpretation of results.

A. Data Preparation

Data preparation critical to data mining includes processing the data in logistics management system, checking the integrity and consistency of the data and dealing with incorrect and useless data. Its process consists of data selection, data integration, and data conversion.

1) Data selection: All internal and external data about business are searched, and the data suitable for data mining application are selected in order to narrow the scope of data mining so as to avoid blind search and enhance the speed and quality of data mining.

2) Data integration: Data are selected and integrated from heterogeneous databases, files or missing systems, and the ambiguity and redundancy of data are eliminated. Different data formats become the same, at the same time, the missing and abnormal data are processed.

3) Data conversion: Data will be converted into an analysis model suitable for mining algorithms, which is the key to success of data mining.

B. Data Acquisition

Data collection is the core technology of data mining, mainly including the following four parts.

1) Determining the type of the task: At first, functions of the system and the type of the task are determined[6].

2) Choosing mining technology: The appropriate data mining technique is selected on the basis of the task, such as the classification model often using the neural network o inductive technology, clustering by clustering analysis, correlation analysis by use of the technology finding relations and sequences.

3) Selecting algorithms: In accordance with selected technique, a kind of specific algorithm is selected, which will determine the way of searching the hidden pattern, with the specific algorithm matching with the overall goal of data mining.

4) Mining data: The selected algorithm is used for the repeated search in the model space, collecting hidden and new models from the set of data.

C. Interpretation and Evaluation of Results

The results of the data mining do not necessarily satisfy users, and the analysis method is generally determined by the data mining process. According to the decision purpose, extracted information is analyzed, and the most valuable information is distinguished and sent to decision-makers by the decision support system. Interpretation and evaluation filtering the information determine whether or not to send the found rules into knowledge base. If the results have no significance for decision-makers, data mining process will be repeated.

D. User Interface and Knowledge Base

The knowledge gotten by the analysis is integrated into the business information system. With the visualization technology and appropriate visualization tools, the users confirm the reliability of knowledge. The knowledge in the knowledge base and found by data mining method and other scientific methods has a variety of forms such as charts and rules which provide decision-makers with the strong support. The visual interface helps managers of logistics companies understand the knowledge and make decisions.

VII. How Data Mining Works

The technique that is used to perform these feats in data mining is called modeling. Modeling is simply the act of building a model in one situation where you know the answer and then applying it to another situation that you don't. For instance, if you were looking for a sunken Spanish galleon on the high seas the first thing you might do is to research the times when Spanish treasure had been found by others in the past. You might note that these ships often tend to be found off the coast of Bermuda and that there are certain characteristics to the ocean currents, and certain routes that have likely been taken by the ship’s captains in that era. You note these similarities and build a model that includes the characteristics that are common to the locations of these sunken treasures. With these models in hand you sail off looking for treasure where your model indicates it most likely might be given a similar situation in the past. Hopefully, if you've got a good model, you find your treasure.

This act of model building is thus something that people have been doing for a long time, certainly before the advent of computers or data mining technology. What happens on computers, however, is not much different than the way people build models. Computers are loaded up with lots of information about a variety of situations where an answer is known and then the data mining software on the computer must run through that data and distill the characteristics of the data that should go into the model. Once the model is built it can then be used in similar situations where you don't know the answer [7]. For example, say that you are the director of marketing for a telecommunications company and you'd like to acquire some new long distance phone customers. You could just randomly go out and mail coupons to the general population - just as you could randomly sail the seas looking for sunken treasure. In neither case would you achieve the results you desired and of course you have the opportunity to do much better than random - you could use your business experience stored in your database to build a model.

VIII. Aspects Paid Attention To When Applying Data Mining Technology

As the business data are numerous and jumbled, the following aspects should be paid attention to

A. Data Collection and Preparation

From a variety of data sources, data needed by data mining are integrated to ensure data quality. Choosing the right data source is critical to the entire data mining project. The abnormal data input into the database, irrelevant or conflicting fields, and out of date are repaired, and the redundant data are removed.

B. Reducing Costs, and Avoiding the Extra Investment

Logistics companies should pay attention to updating the database available continuously and the integration of data mining system and other systems in order to reduce the costs of construction and use of the database, improve economic efficiency, and avoid the repeated investment.

C. Choice of Data Mining Tools

Logistics enterprises should select the data mining tools reflecting the specific operating conditions based on the specific situation and the actual needs of enterprises, rather than blindly pursuing the mature and developed tools.

D. Competent Personnel

Logistics enterprises should attach importance to the introduction and retention of qualified personnel with extensive statistics, logistics and industry knowledge, because the technology and methods they select have the significant impact on the effectiveness of the model.

IX. Successful Application Area

A pharmaceutical company can analyze its recent sales force activity and their results to improve targeting of high-value physicians and determine which marketing activities will have the greatest impact in the next few months. The data needs to include competitor market activity as well as information about the local health care systems. The results can be distributed to the sales force via a wide-area network that enables the representatives to review the recommendations from the perspective of the key attributes in the decision process. The ongoing, dynamic analysis of the data warehouse allows best practices from throughout the organization to be applied in specific sales situations [4].

A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product. Using a small test mailing, the attributes of customers with an affinity for the product can be identified. Recent projects have indicated more than a 20-fold decrease in costs for targeted mailing campaigns over conventional approaches [8].

A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. Using data mining to analyze its own customer experience, this company can build a unique segmentation identifying the attributes of high-value prospects. Applying this segmentation to a general business database such as those provided by Dun & Bradstreet can yield a prioritized list of prospects by region.

A large consumer package goods company can apply data mining to improve its sales process to retailers. Data from consumer panels, shipments, and competitor activity can be applied to understand the reasons for brand and store switching. Through this analysis, the manufacturer can select promotional strategies that best reach their target customer segments.

X. Conclusion

Data Mining has proved to be one of the most competent tools in business. The system based on data mining technically overcomes the difficulty of forecasting and analysis of decisions, which has brought great social and economic benefits. In short, the data mining technology is becoming more and more powerful. Comprehensive data warehouses that integrate operational data with customer, supplier, and market information have resulted in an explosion of information. Competition requires timely and sophisticated analysis on an integrated view of the data. However, there is a growing gap between more powerful storage and retrieval systems and the users' ability to effectively analyze and act on the information they contain. Both relational and OLAP technologies have tremendous capabilities for navigating massive data warehouses, but brute force navigation of data is not enough. A new technological leap is needed to structure and prioritize information for specific end-user problems. The data mining tools can make this leap. Quantifiable business benefits have been proven through the integration of data mining with current information systems, and new products are on the horizon that will bring this integration to an even wider audience of users.

REFERENCES

[1] C. Westphal and T. Blaxton, *Data Mining Solutions*, John Wiley, 1998.

[2] Q. Song and M. Shepperd, "Mining web browsing patterns for Ecommerce," *Computers in Industry*, vol. 57, no. 7, pp. 623-629, 2006.

[3] P. Yingmei, "Modern logistics decision based on data mining Technology," *Logistics Technology*, vol. 27, no. 7, pp. 47-49, 2007.

[4] Liu Dejun and Zhang Guangsheng, "Application of data mining technology in modern agricultural logistics management decision," *Journal of Shenyang Normal University (Natural Science)*, vol. 26, no. 3, pp. 310-312, 2008.

[5] L. L. Weldon, "Data mining and visualization," *Database Programming and Design*, 1996, vol. 9, no. 5, pp. 21-24.

[6] C. Min, "The function of data mining in logistics management," *Computer Development & Applications*，2007, vol. 20, no. 1, pp. 47-48.

[7] D. Pyle, *Data Preparation for Data Mining*, Morgan Kaufmann, 1999.

[8] J. Q. Li, S. L. Wang , C. L. Niu , and J. Z. Liu, "Research and Application of Data Mining Technique in Power Plant," *International Symposium on Computational Intelligence and Design*, 2008

[9] C. S. R. Prabhu, *Data Mining*.

[10] A. Pujari, *Data Mining Techniques*.

Dipti Verma belongs to Nagpur, Maharashtra, India born at Tumsar on 29th July . Had done.B.Sc. (Chemistry, botany, zoology) followed by MCA from Nagpur University, Maharashtra, India in 2007. She is working as a professor at G. S. Raisoni College of Engineering in Nagpur since 3 years .