Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol 2, No.8, 2011

Data Mining Tools and Techniques: a review

Karimella Vikram
Associate Professor in CSE Dept.
Medak College of Engineering and Technology
Siddipet , Medak (D)
Tel: 9030231455 Email : vikramkariella@yahoo.com

Niraj Upadhayaya
Professor & HOD CSE Dept
JBIET Moinabad , hyderbad
Tel: 9906213890 Email: drnirajup@gmail.com

Received: 2011-10-27
Accepted: 2011-10-29
Published:2011-11-04

Abstract

Data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. Because data mining tools predict future trends and behaviors by reading through databases for hidden patterns, they allow organizations to make proactive, knowledge-driven decisions and answer questions that were previously too time-consuming to resolve. 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 of logistics companies and support their decisions and development. This paper introduces the significance use of data mining tools and techniques in logistics management system, and its implications. Finally, it is pointed out that the data mining technology is becoming more and more powerful in logistics management.

Keywords— Logistics management, Data Mining concepts, application areas, Tools and Techniques

I. INTRODUCTION TO LOGISTICS MANAGEMENT

Through the use of automated statistical analysis (or "data mining") techniques, businesses are discovering new trends and patterns of behavior that previously went unnoticed. Once they've uncovered this vital intelligence, it can be used in a predictive manner for a variety of applications. Logistics is the management of the flow of goods, information and other resources in a repair cycle between the point of origin and the point of consumption in order to meet the requirements of customers. Logistics involves the integration of information, transportation, inventory, warehousing, material handling, and packaging, and occasionally security. Logistics is a channel of the supply chain which adds the value of time and place utility. Today the complexity of production logistics can be modeled, analyzed, visualized and optimized by plant simulation software.

Logistics management is that part of the supply chain which plans, implements and controls the efficient, effective forward and reverse flow and storage of goods, services and related information between the point of origin and the point of consumption in order to meet customer and legal requirements. A professional working in the field of logistics management is called a logistician. Logistics management is the governance of supply chain functions. Logistics management activities typically include inbound and outbound transportation management, fleet management, warehousing, materials handling, order fulfillment, logistics network design, inventory management, supply/demand planning, and management of third party logistics services providers. To varying degrees, the logistics function also includes customer service, sourcing and procurement, production planning and scheduling, packaging and assembly. Logistics management is part of all levels of planning and execution -- strategic, operational and tactical. It is an integrating function, which coordinates all logistics activities, as well as integrates logistics activities with other functions including marketing, sales manufacturing, finance, and information

2. DATA MINING CONCEPS

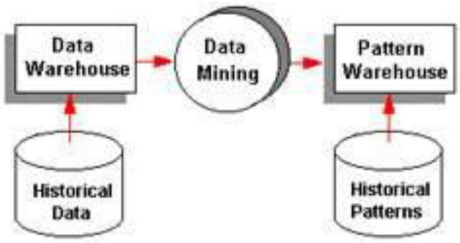
Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining fondly called patterns analysis on large sets of data uses tools like association, clustering, segmentation and classification for helping better manipulation of the data. Simply put, data mining is a continuous, iterative process that is the very core of business intelligence. It involves the use of data mining software, sound methodology and human creativity to achieve new insight through the exploration of data to uncover patterns, relationships, anomalies and dependencies. We have achieved our reputation as the data mining industry's leading innovator by developing powerful, user friendly and affordable data mining technology, and by delivering comprehensive knowledge transfer to customers to enable them to take advantage of the business benefits data mining technology makes possible. For almost a decade we has taken the leadership role in broadening user understanding and acceptance of this technology as a highly value decision support system for a wide range of business applications in many different industries. Our data mining customers -- one of the largest installed base of active users of this technology in the world -- have been increasing their revenues, lowering their costs, and enhancing their competitive positions because they have openly embraced and actively explored the possibilities data mining technology offers to them. Data mining derives its name from the similarities between searching for valuable business information in a large database 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. In some cases the data are consolidated in a data warehouse and data marts; in others they are kept on the Internet and intranet servers. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing the following capabilities:

2.1. 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 and quickly from the data. A typical example of a predictive problem is targeted marketing. Data mining can use data on past promotional mailings to identify the targets most likely to respond favorably to future mailings. Other predictive examples include forecasting bankruptcy and other forms of default and identifying segments of a population likely to respond similarly to given events.

2.2. Automated discovery of previously unknown patterns.

Data mining tools 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, such as baby diapers and beer. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying invalid (anomalous) data that may represent data entry keying errors. When data mining tools are implemented on high-performance, parallel-processing systems, they can analyze massive databases in minutes. Often, these databases will contain several years' worth of data. Faster processing means that users can 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. Data mining also can be conducted by nonprogrammers. The "miner" is often an end user, empowered by "data drills" and other power query tools to ask ad-hoc questions and get answers quickly, with little or no programming skill. Data mining tools can be combined with spreadsheets and other end-user software development tools, making it relatively easy to analyze and process the mined data. Data mining appears under different names, such as knowledge extraction, data dipping, data archeology, data exploration, data pattern processing, data dredging, and information harvesting. "Striking it rich" in data mining often involves finding unexpected, valuable results.



Data mining yields five types of information*:*

1. Association. Relationships between events that occur at one time (e.g., the contents of a shopping cart, such as orange juice and cough medicine) 2. Sequences. Relationships that exist over a period of time (e.g., repeat visits to a supermarket) 3. Classifications. The defining characteristics of a certain group (e.g., customers who have been lost to competitors) 4. Clusters. Groups of items that share a particular characteristic that was not known in advance of the data mining. 5. Forecasting. Future values based on patterns within large sets of data (e.g., demand forecasting) Data miners use several tools and techniques: case-based reasoning (using historical cases to recognize patterns); neural computing (a machine-learning approach by which historical data can be examined for patterns through massive parallel processing); association analysis (using a specialized set of algorithms to sort through data sets and express statistical rules among items); and intelligent agents (expert or knowledge-based software embedded in information systems).

3. DATA MINING PROCESSS

The data mining process consists of four basic steps as is shown in Figure 1: question definition; data preparation and pre-processing; data mining; result interpretation and validation.

- Question definition: - The methods of data mining are selected according to actual requirement. So the related knowledge of the object must be well learned before data mining and the goal of data mining must be definite.

- *Data* selection and pre-processing:- 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.

- 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.

3.1 Result interpretation and validation*:* - T

. *Figure2.Stepwise treatment model of typical Data Mining process*

3.2 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.

3.3. ADVANTAGES AND APPLICATION AREAS OF DATA MINING

Retailing and sales distribution. Predicting sales, determining correct inventory levels and distribution schedules among outlets

- Banking*:* Forecasting levels of bad loans and fraudulent credit card use, predicting credit card spending by new customers, predicting which kinds of customers will best respond to (and qualify for) new loan offers data mining The process of searching a large database to discover previously unknown patterns; automates the process of finding predictive information.

- Manufacturing and production:Predicting machinery failures, finding key factors that control optimization of manufacturing capacity

- Brokerage and securities trading*:* Predicting when bond prices will change, forecasting the range of stock fluctuations for particular issues and the overall market, determining when to buy or sell stocks

- Insurance: Forecasting claim amounts and medical coverage costs, classifying the most important elements that affect medical coverage, predicting which customers will buy new policies

- Computer hardware and software: Predicting disk-drive failures, forecasting how long it will take to create new chips, predicting potential security violations

- Police work.*:*Tracking crime patterns, locations, and criminal behavior; identifying attributes to assist in solving criminal cases

- Government and defense*:* Forecasting the cost of moving military equipment, testing strategies for potential military engagements, predicting resource consumption; improving homeland security by mining data from many sources

- Airlines: Capturing data on where customers are flying and the ultimate destination of passengers who change carriers in hub cities so that airlines can identify popular locations that they do not service, checking the feasibility of adding routes to capture lost business

- Health care: Correlating demographics of patients with critical illnesses, developing better insights on symptoms and their causes, learning how to provide proper treatments

- Broadcasting*:* Predicting the most popular programming to air during prime time, predicting how to maximize returns by interjecting advertisements

- Marketing*:* Classifying customer demographics that can be used to predict which customers will respond to a mailing or buy a particular product

DISADVANTAGES OF DATA MINING

- Privacy Issues *:* Personal privacy has always been a major concern in this country. In recent years, with the widespread use of Internet, the concerns about privacy have increase tremendously. Because of the privacy issues, some people do not shop on Internet. They are afraid that somebody may have access to their personal information and then use that information in an unethical way; thus causing them harm. Although it is against the law to sell or trade personal information between different organizations, selling personal information have occurred. For example, according to Washing Post, in 1998, CVS had sold their patient's prescription purchases to a different company.7 In addition, American Express also sold their customers' credit care purchases to another company.8 What CVS and American Express did clearly violate privacy law because they were selling personal information without the consent of their customers. The selling of personal information may also bring harm to these customers because you do not know what the other companies are planning to do with the personal information that they have purchased.

- Security issues*:* Although companies have a lot of personal information about us available online, they do not have sufficient security systems in place to protect that information. For example, recently the Ford Motor credit company had to inform 13,000 of the consumers that their personal information including Social Security number, address, account number and payment history were accessed by hackers who broke into a database belonging to the Experian credit reporting agency.9 This incidence illustrated that companies are willing to disclose and share your personal information, but they are not taking care of the information properly. With so much personal

information available, identity theft could become a real problem. *Misuse of* information/inaccurate information:Trends obtain through data mining intended to be used for marketing purpose or for some other ethical purposes, may be misused. Unethical businesses or people may used the information obtained through data mining to take advantage of vulnerable people or discriminated against a certain group of people. In addition, data mining technique is not a 100 percent accurate; thus mistakes do happen which can have serious consequence.

4.-. TECHNIQUES FOR DATA MINING

Just as a carpenter uses many tools to build a sturdy house, a good analyst employs more than one technique to transform data into information. Most data miners go beyond the basics of reporting and OLAP (On-Line Analytical Processing, also known as multi-dimensional reporting) to take a multi-method approach that includes a variety of advanced techniques. Some of these are statistical techniques while others are based on artificial intelligence (AI).

- Cluster Analysis:-Cluster analysis is a data reduction technique that groups together either variables or cases based on similar data characteristics. This technique is useful for finding customer segments based on characteristics such as demographic and financial information or purchase behavior. For example, suppose a bank wants to find segments of customers based on the types of accounts they open. A cluster analysis may result in several groups of customers. The bank might then look for differences in types of accounts opened and behavior, especially attrition, between the segments. They might then treat the segments differently based on these characteristics.

- Linear Regression:-Linear regression is a method that fits a straight line through data. If the line is upward sloping, it means that an independent variable such as the size of a sales force has a positive effect on a dependent variable such as revenue. If the line is downward sloping, there is a negative effect. The steeper the slope, the more effect the independent variable has on the dependent variable.

- Correlation:-Correlation is a measure of the relationship between two variables. For example, a high correlation between purchases of certain products such as cheese and crackers indicates that these products are likely to be purchased together. Correlations may be either positive or negative. A positive correlation indicates that a high level of one variable will be accompanied by a high value of the correlated variable. A negative correlation indicates that a high level of one variable will be accompanied by a low value of the correlated variable. Positive correlations are useful for finding products that tend to be purchased together. Negative correlations can be useful for diversifying across markets in a company's strategic portfolio. For example, an energy company might have interest in both natural gas and fuel oil since price changes and the degree of substitutability might have an impact on demand for one resource over the other. Correlation analysis can help a company develop a portfolio of markets in order to absorb such environmental changes in individual markets.

- Factor Analysis:-Factor analysis is a data reduction technique. This technique detects underlying factors, also called "latent variables," and provides models for these factors based on variables in the data. For example, suppose you have a market research survey that asks the importance of nine products attributes. Also suppose that you find three underlying factors. The variables that "load" highly on these factors can offer some insight about what these factors might be. For example, if three attributes such as technical support, customer service, and availability of training courses all load highly on one factor, we might call this factor "service." This technique can be very helpful in finding important underlying characteristics that might not be easily observed but which might be found as manifestations of variables that can be observed. Another

good application of factor analysis is to group together products based on similarity of buying patterns. Factor analysis can help a business locate opportunities for cross-selling and bundling. For example, factor analysis might indicate four distinct groups of products in a company. With these product groupings, a marketer can now design packages of products or attempt to cross-sell products to customers in each group who may not currently be purchasing other products in the product group.

- Decision Trees*:*-Decision trees separate data into sets of rules that are likely to have different effects on a target variable. For example, we might want to find the characteristics of a person likely to respond to a direct mail piece. These characteristics can be translated into a set of rules. Imagine that you are responsible for a direct mail effort designed to sell a new investment service. To maximize your profits, you want to identify household segments that, based on previous promotions, are most likely to respond to a similar promotion. Typically, this is done by looking for combinations of demographic variables that best distinguish those households who responded to the previous promotion from those who did not. This process gives important clues as to who will best respond to the new promotion and allows a company to maximize its direct marketing effectiveness by mailing only to those people who are most likely to respond, increasing overall response rates and increasing sales at the same time. Decision trees are also a good tool for analyzing attrition (churn), finding cross-selling opportunities, performing promotions analysis, analyzing credit risk or bankruptcy, and detecting fraud. Decision trees are tree-shaped structures that represent decision sets. These decisions generate rules, which then are used to classify data. Decision trees are the favored technique for building understandable models. Auditors can use them to assess, for example, whether the organization is using an appropriate cost-effective marketing strategy that is based on the assigned value of the customer, such as profit.

- Neural Networks*:* - Neural networks mimic the human brain and can "learn" from examples to find patterns in data or to classify data. The advantage is that it is not necessary to have any specific model in mind when running the analysis. Also, neural networks can find interaction effects (such as effects from the combination of age and gender) which must be explicitly specified in regression. The disadvantage is that it is harder to interpret the resultant model with its layers of weights and arcane transformations. Neural networks are therefore useful in predicting a target variable when the data are highly non-linear with interactions, but they are not very useful when these relationships in the data need to be explained. They are considered good tools for such applications as forecasting, credit scoring, response model scoring, and risk analysis. Artificial neural networks are non-linear, predictive models that learn through training. Although they are powerful predictive modeling techniques, some of the power comes at the expense of ease of use and deployment. One area where auditors can easily use them is when reviewing records to identify fraud and fraud-like actions. Because of their complexity, they are better employed in situations where they can be used and reused, such as reviewing credit card transactions every month to check for anomalies.

- Association Models*:* - Association models examine the extent to which values of one field depend on, or are predicted by, values of another field. Association discovery finds rules about items that appear together in an event such as a purchase transaction. The rules have user-stipulated support, confidence, and length. The rules find things that "go together." These models are often referred to as Market Basket Analysis when they are applied to retail industries to study the buying patterns of their customers.

- The nearest-neighbor*:* - method classifies dataset records based on similar data in a historical dataset. Auditors can use this approach to define a document that is interesting to them and ask the system to search for similar items. Each of these approaches brings different advantages and

disadvantages that need to be considered prior to their use. Neural networks, which are difficult to implement, require all input and resultant output to be expressed numerically, thus needing some sort of interpretation depending on the nature of the data-mining exercise. The decision tree technique is the most commonly used methodology, because it is simple and straightforward to implement. Finally, the nearest-neighbor method relies more on linking similar items and, therefore, works better for extrapolation rather than predictive enquiries. A good way to apply advanced data mining techniques is to have a flexible and interactive data mining tool that is fully integrated with a database or data warehouse. Using a tool that operates outside of the database or data warehouse is not as efficient. Using such a tool will involve extra steps to extract, import, and analyze the data. When a data mining tool is integrated with the data warehouse, it simplifies the application and implementation of mining results. Furthermore, as the warehouse grows with new decisions and results, the organization can mine best practices continually and apply them to future decisions. Regardless of the technique used, the real value behind data mining is modeling the process of building a model based on user-specified criteria from already captured data. Once a model is built, it can be used in similar situations where an answer is not known. For example, an organization looking to acquire new customers can create a model of its ideal customer that is based on existing data captured from people who previously purchased the product. The model then is used to query data on prospective customers to see if they match the profile. Modeling also can be used in audit departments to predict the number of auditors required to undertake an audit plan based on previous attempts and similar work.

- Link Analysis*:* - This is another technique for associating like records. Not used too much, but there are some tools created just for this. As the name suggests, the technique tries to find links, either in customers, transactions, etc. and demonstrate those links.

- Visualization*:* - This technique helps users understand their data. Visualization makes the bridge from text based to graphical presentation. Such things as decision tree, rule, cluster and pattern visualization help users see data relationships rather than read about them. Many of the stronger data mining programs have made strides in improving their visual content over the past few years. This is really the vision of the future of data mining and analysis. Data volumes have grown to such huge levels; it is going to be impossible for humans to process it by any text-based method effectively, soon.

5. DATA MINING APPLICATION IN LOGISTICS MANAGEMNT

In today's fastest growing world the term logistics has become the buzz word. As per the research work carried out it has been observed that Data Mining is very essential in Logistics Management. Logistics basically deals with all type of stuff whether it is perishable goods non perishable goods. It can be easily illustrated by an example: if Company A deals with all the food materials. And it wants to transport the perishable material like grains from City X to City Y. Then the delivery could be delayed for few days and it will not harm the food material. But in case of grains if fruit (which is non perishable item) are to be supplied same from city X to city Y than the delay in delivery could cause damage of the food item. But when we deal with Logistic Management there are many items or material which are to be supplied to the companies every now and then. So it becomes difficult in Logistics to avoid that damage of goods and take care of all the perishable and non perishable goods. So to overcome this disadvantage the data mining techniques are to be applied in the Logistics Management. While mining the data from the data warehouse every small thing could be noticed and this could lead to the avoidance of wastage of goods or Logistics also deals with the flow of goods, information and other resources. Different data mining techniques could be applied in Logistics Management and their results could be analysed, and based on the observation it could be applied in Logistic Management.

6. CONCLUSION

In this paper application of data mining in terms of perishable and non perishable goods is introduced. Data mining makes it possible to analyze routine business transactions and glean a significant amount of information about individuals buying habits and preferences. Logistics industry helps enhance core

References

Song and Shepperd M. (2006), “Mining web browsing patterns for ecommerce”, Computers in Industries,57(7), pp.623-629.

Pei Yingmei, (2007), “Modern logistics decision based on data mining technology”, Logistics Technology, 27(7),pp47-49.

Liu Dejun and Zang Guangsheng (2008), “ Application of Data mining Technology in modern agricultural Logistic Management Decision” Journal of Shenyang Normal University (Natural Science), 26(3) pp 310-312.

Lay-Louise Weldon. (1996)” Data mining and visualization”,Databse Programming and Design”, 9(5),pp21-24.

Cao Min. (2007), “The function of Data mining in Logistic Management”, Computer Development & Applications, 20(1), pp-47-48.

Dorian Pyle, Morgan Kaufmann.(1999)” Data Preparation for Data Mining”, 1999.