The Development of Data Mining

Fang Weiping
Wang Yuming
Shanghai Engineering Technology University İnstitute of Management
Shanghai, 201620

Abstract

Mining is the current hot spots, the most promising research areas has broad one, through data mining research status, algorithms and applications of analysis to explore data mining problems and trends, which is the development of data mining has certain reference value.

Key Words: Database; evolution; algorithm

1. Introduction

In our daily life, we often encounter such a situation: message tell you XXX store or e-commerce platform will present new products or promotions recently, or emails notify you about mall anniversary celebration recently. This is because retailers at the store checkout the customer data and collect their data, retailers can make use of this information, combined with electronic commerce log , shopping record and so on, better understanding of the needs of the consumers, to make a correlation analysis. In the movement of the era of big data, data analysis tools in the information management system have not been extracted information from huge complex data. For example, information systems can not analysis of these data which are collected and observe from satellite surface, Marine atmospheric, because of the size and characteristics of time and space. Also, there are vast amounts of genetic data, etc. In this way, we need to develop a new method to mine the data.

2. Concept of data mining

Data Mining is the advanced process which extracts the potential and effective and comprehensive mode from the vast amounts of Data in accordance with the established business goals. [1] many people consider data mining as commonly term of knowledge discovery , while others simply put data mining as a basic steps in the process of knowledge discovery [2]. It appeared in the late 1980s, which is new areas with a great researching value in the study of the database, and overlapping subject, combined with artificial intelligence, database technology, pattern recognition, machine learning, statistics, data visualization, and other fields of theory and technology. As a kind of technology, the life cycle of data mining is in unclear stage, experts need takes time and energy to research, develop and propel to be mature gradually, eventually been accepted [3-4]. Data mining is a kind of technology, which combines the traditional data processing methods with different algorithms, to analyze new data types and extract knowledge from huge amounts of data. Found in huge amounts of data, there are two kinds of knowledge, one is on-line analytical processing (OLAP), the other is a data mining (DM). Both are analysis tools based on data warehouse, but on-line analytical processing appears earlier than the data mining, based on a multidimensional view, emphasizing the execution efficiency and quick response for user commands; And data mining is pay attention to the useful model to people hiding in the depths of data, and it is done by automation, without participation of customers.

2.1 On what data mining

Data mining can examine any type of data and information flow, its difficulty is relative with database type.
(1) Mining for relational database. A relational database is the set of tables; table is composed of attributes group, depositing large number of tuples. Usually use ER model to represent the connection between the database and the real. Excavating from a relational database, the trends and data model can be obtained.

For example, the customer's income, age, education level, and other information can be obtained and by commercial relational database for mining, then making targeted marketing for customer and avoiding fraud, and shape a company's strategy.

(2) Mining for data warehouse. Data warehouse is a subject oriented, integrated, non-volatile and time-variant collection of data, which contains consistent data used in enterprise decision support [5-6]. The data warehouse is the data environment that can be used as the single integrated source of data for processing information. Data warehouses deposit aggregated data, which are processed to find hidden patterns and relation to structure analytical model to classify and forecast.

（3）Mining for new database. The new database includes spatial database, time database and text database and multimedia database. These data include spatial data, text, data, image and audio data and streaming data and web data. The data structure is more complex and dynamic change, more difficult to handle. For example, through the data mining technology can find the evolution of the object characteristics and trends; streaming data are clustered and compared to find interesting patterns.

2.2 The evolution process of data mining

1960s, database technology, and information technology has gradually developed from the basic document processing system to more complex and more powerful database system, such as hierarchical and network database are typical representative of this era with little data independence and abstraction. 1970s, relational databases appear, allowing users access to a flexible data access language and interface, OLTP technology make the relational database technology application gained popularity; Mid-1980s, the rise of a powerful database system, and put forward many advanced data models. for example expanding the relational model, object-oriented model, the interpretation of the model, etc. by the end of 80 advanced data model and application oriented database were developed.After 2000, the ability to store large amounts of data is over the capacity of human analysis and understanding; there is no suitable tool to help extracting information and knowledge from the data. The existence of specific patterns and rules can be found by data mining tools in a large amount of data, which can provide the necessary information for commercial activity, scientific exploration and medical research and many other areas.Business intelligence (BI) based on data mining has become the new darling of the IT industry. Currently the data mining has been successfully applied in retail goods basket data analysis, financial risk prediction, product quality, molecular biology, genetic engineering, discovery of Internet site access patterns, the information search and classification and many other fields.

2.3 Data Mining Tasks

2.3.1 Forecasts

With assignment to the line attribute or object attribute unknown category data, we use the obtained learning model to forecast. Classification and regression are two major kinds of prediction model. The former is used to predict discrete or symbolic value, and the regression is used to predict continuous values. The answer is Y or N for the question of purchase of goods online. While to forecast future stock prices and trends are regression-based tasks. Predictive models can determine the benefits and risks of future market, also can predict the earth's resources consumption.

2.3.2 Description

Summed up the relationship that exists potential data model is a role of authentication and interpretation. Usually correlation analysis use to describe model with strong corelational characteristics, to extract interesting pattern to find the correlation in the data. Extract characteristic formulas expressed the general characteristics of the data set from the data warehouse, Or find other features to distinguish the characteristics of style.[5] For example, feature extraction and distinction from other cases. While they are commonsense, association rule mining can find many other interesting interactions, such as 'beer and diaper'. This is the famous market basket analysis, which was once a secret weapon in Wal-Mart, market basket analysis can help us find stores selling process with affiliated merchandise.

2.4 Classification of Data Mining

Data mining can be divided into two categories, direct and indirect data mining. The goal of direct data mining is to use the available data to create model with a description of variables. The goal of indirect data mining is to no choice of a specific variable, but to establish a relationship of all the variables [5].Classification, estimation and prediction are direct data mining; Association rule, clustering, description, and visualization are indirect data mining.

Association rules is not known in advance what would be the knowledge obtained, what get after the analysis of data. Such as customers buy A product with B product. Clustering is grouping on similar records and put the similar records in a gathering.The difference between clustering and classification is that clustering does not rely on pre-defined classes, no training set. Description and visualization is a representation of the data mining results.

3. Common method of data mining

3.1 Correlation analysis

Discover useful patient or associated knowledge useful from a large data set. Basic idea express as: W - > B, W represents the set of attributes, B represent attributes individual, rules simply interpreted if W with true value, individual B obtains the possibility and trend of true value in the database list[6]. After buying a commodity, how likely does continue to buy B commodity?

3.2 Decision Tree.

A decision tree composed by a series of nodes and branches, and the nodes link sub-nodes by branches .Nodes represents attributes considered in the decision-making process, and different attribute values form different branches [7]. Using the decision tree model to make decisions, you can search from the root to the leaves; leaf nodes containing decision the results of each scheme.

3.3 Genetic Algorithm

Genetic algorithms are a probabilistic search to find the optimal process. It is generated from random or specific groups, in accordance with certain rules of operation to continue iterative calculation, such as selection, reproduction, crossover and mutation, etc., and it is the process of retaining fine varieties, eliminating inferior varieties, and guiding the search to approaching optimal solution according to each individual's fitness. Genetic algorithm execution requires two data conversion, the encoding and decoding. Coding is to convert parameters of search space to chromosome or individuals of genetic space; Decoding is to convert chromosome or individuals of genetic space to parameters of search space. Genetic algorithm is developed on the basis of simulation of genetics, to operate directly on the structure objects. It has very good robustness without restriction of derivative and function [8-9].

3.4 Bayesian Networks

Bayesian network is based on the mathematical model of probabilistic inference, a probabilistic inference is through some information to obtain the probability of other variables, Bayesian network based on probabilistic inference is proposed to solve the problem of uncertainty and incompleteness, it has the very advantage of solving fault caused by complex uncertainty and correlation, widely used in many fields. With N nodes in bayesian networks can be expressed by $BN_N = \langle \langle V, E \rangle, P \rangle, \langle V, E \rangle$ is N nodes with a directed acyclic graph, node $V_i \in V$ is abstraction of unit status, observation, personnel operation. Directed edges $(V_i, V_j) \in E$ represents there is a direct influence or causal relationship between nodes V_i and V_j. V_i is V_j's father node. P represents associated with each node condition probability distribution, it expresses the nodes associated with the parent node. Using bayesian network structure and the conditional probability tables, calculate the probability of values of certain node after giving evidence

3.5 Rough Set Approach

Rough set theory is a mathematical methods to deal with the ambiguity and uncertainty , using the method of rough set can analyze the decision table, can evaluate the importance of a specific attribute, build property set reduction, nuclear, and get rid of redundant attributes from the decision table, classification rules arise from the table of reduction to make decision. The main idea of rough set based on the existing knowledge on the given problem, through classifying to manage for the actual data, divide the domain of the problem,reduct data under the premise of retaining critical information, get reduction and nuclear of knowledge, assess the dependencies between data ,derived classification rules of concept [10].

3.6 Neural Network

Neural network is a dynamic system with topology structure of directed the graph; it deals with information through response for the continuous or intermittent input state. Neural network system is made up of large and simple processing units, through widely connected to each other and formed a complex network of systems. Although the structure and function of each neuron is very simple, but the behavior of the network system consisting of a large number of neurons is really colorful and very complicated. The algorithm is suitable for data clustering, which can make a lot of complex information data is systematic and ordinary, so as to find the inner relationship between the data through the analogy of time and space.

3.7 Statistical Analysis

Statistical analysis is accurate method of data mining based on statistics and probability theory. For example, regression analysis and factor analysis, through the modeling of objects, find a conclusion. Usually divided into the following steps: description to the nature of the analytical data, researching group of data relationship, building of a model, summary of data and relationship of basis group, explanation the validity of the model, and finally prediction for the future development. SPSS and SAS are widely used as application software of statistical.

4. Data Mining Applications

Data mining technology is application-oriented. In many areas, the data mining have played a major role, especially in the banking, insurance, and transportation and retailing, data mining can solve a lot of business issues, increase business profits and make wise decisions. Retail is not only the first application of data mining techniques but also important areas. Because the retail industry has accumulated a lot of sales data, such as customer purchase records, consumer information and service information, etc. Enterprises can use the data to classify customers, from basic customer groups found gold customer characteristics and consumer intent, and as far as possible to provide satisfactory products and services for them. Application of data mining software, choose the appropriate algorithm, knowledge hidden behind the data would be found. The information obtained by data mining applies in making marketing strategy to guide enterprise decision-making. Retail data Mining main functions are: market positioning, consumer analysis, forecast sales trends, marketing strategy optimization, inventory requirements analysis, customer buying pattern recognition, shelf placement assistance, the time to develop promotional activities, promotional merchandise combination as well as slow-moving and selling goods of the situation of other commercial activities. Goods basket analysis is the most widely used in data mining, with product sales, product pricing, promotion activities to reduce business costs and increase profits.

Now, many of popular e-commerce are the application of data mining, product portfolio marketing and product recommendations, but will provide customers with personalized marketing at the same time, optimization of website design and product promotion. Secondly in the financial industry, data mining also played a big role. Insurance industry is a serious information asymmetry, the analysis of the policy-holder can reduce the risk, and it is only through the mining customs' history to understand customer behavior effectively, prevent the happening of insurance fraud. Bank loans, securities heavily depend on data mining. When banks lend, it need find customers' wage income, education level, credit history and other factors that affect credit, and then decide whether to loan, set up a model of credit fraud to predict the risks and benefits. Medicine and biotechnology, the genetic analysis of these data through data mining technology is helpful to research and understand [11]. Data mining decision inference which has been widely applied in the field of medicine, build medical model aimed at the study of difficult problems to find the essence of connection and phenomenon, reasoning a new treatment of the disease.

BES are committed maybe the loss of bank customers through SPSS software to identify the characters of customers who are likely to leave the bank. Jorge Portugal and his team analyze these dynamic relationships and build models to develop appropriate adjustment strategy to improve customer satisfaction. Results show that customer churn rate was reduced by 15% - 15%, profits rose by 10% to 10%. HSBC Bank may have a number of banks with branches in the same area, resulting in a sustainable competitive to attract and retain potential customers nearby. In order to maintain a high level of customer acquisition, maintenance of profitability, banks must expand existing customer relationships, control marketing costs in order to maintain profits and rapid transfer market, HSBC dig on growing customer data , build predictive modeling to discover opportunities for cross-selling and tumbling.

Positioned at each customer with the best value, maximize products sales, minimal marketing costs. Just three years, the bank product sales increased by 50%, the marketing cost reduced by 30%, improve the ability to establish and carry out real-time marketing strategy.

5. Data Mining Challenges

5.1 Performance of data mining

Large data increases the challenge of algorithms and computing platforms curse of dimensionality is more severe, computational overhead is also increased. Fundamentally speaking, from large data production, storage, protection, archiving to safety maintenance of various angles, this is the category of IT management maintenance, but when data quantity is beyond conventional management scale, the difficulty to maintain management appeared upward trend.

5.2 The diversification of data types

The structure of data on the web is poor, and most of them is semi-structured or unstructured, since semi-structured and unstructured information can't clearly express by the data model, therefore it is necessary to develop new data mining tools which is suitable for web data in the Web mining. In the field of remote sensing, scanning the earth via satellite, every day we can get a lot of remote sensing image of the earth's surface, In the field of remote sensing, scanning the earth via satellite, every day we can get a lot of remote sensing image of the earth's surface, spatial data mining technology must be used from a large amount of data of each image hidden [12].

5.3 data security

As the issues of cyber attacks become more prominent, the importance of network security audit is increasingly apparent. Prism that broke out in June is a typical data mining problem of safety, U.S. intelligence agencies attack and monitor on his country's Internet to obtain confidential data through IT enterprises network to collect data from countries around the world. Laws alone can not safeguard the interests of a country, how to build an effective security strategy to protect against unauthorized use of these data will be the focus of future research directions.

6. Data Mining Trends

Data mining is a new kind of intelligent information processing technology. With the rapid development of information technology, the application in the field of data mining will broaden and deepen ceaselessly, especially in the military, security, business intelligence applications. Mobile data mining will be the future direction of data. the powerful Internet capture the science data sets and the social sequence of data set, topology, geometry and other characteristics, connection graph mining and social network analysis usefully. Data mining is directly facing massive databases, so data mining algorithm must be efficient and scalable. The mostly current databases are relational database; the emergence of database of models in future need ensure the ability of processing the type of data is important. Operators with appropriate participation in data mining can accelerate data mining process. On the one hand, the interactive interface for users to provide convenience to express requirements and strategies; On the other hand, interactive interface transform the generated results which are varied to the user, namely data mining system requires strong interactivity. At the same time, that the research of data mining can lead to illegal data invasion is a problem to be solved in real life [13]. In China, data mining technology started late, compared with the United States and other developed countries, there are many shortcomings, for example, almost all our computers still use Microsoft's operating system, the U.S. intelligence agencies on the implementation of the network listening and attacks, resulting in a lot of data leakage, we need in data mining, and must continue to develop technologies to prevent data leakage. In this paper, author hope to provide people who are interested in some of the data mining work with some guidance and help

References

Zhong Xiao, Ma Shaoping, cymbals, zhang YuRuiZhao. Survey of Data Mining [J].fuzzy recognition and artificial intelligence. 2001.01 (14)

Jiawei Han,Micheline Kamber.Data Mining Concepts and Techniques,[M]Second Edition.2007.3.page3-4

Agrawal R.Data mining:Crossing the chasm[R/OL].http://www.almaden.ibm.com.2002-11-20

Wang Guanghong Jiang Ping. Survey of Data Mining [J].Journal of Tongji University. In February 2004, 32 (2)

Lin Wenyuan. Theory of data mining and analysis [J]. The 2011-09

H Mannila,H Toivonen et AL.Efficient algorithms for discovering association rules[C].In:Knowledge Discovery in Databases (KDD 94),AAAI Press,1994:181-192

John Durkin ,Cai Jingfeng, CAI zixing. Decision tree technology and its current research direction[J].control engineering.12 (1) 2005.1

Li Min. Research and application of data mining algorithm [D] the dalian institute of technology, master thesis, 2004.6

JiGenLin. Genetic algorithm (ga) review [J] computer applications and software, 2004.2

Bao Xin zhong,Xiao Ming. Securities investment decision based on rough set theory [J] journal of systems management. In October 2010

Neena Buck.Eureka.Knowledge discovery[J]Software Magazine,2000-12

Qian Feng. Domestic research review [J] intelligence data mining tools. 2008 10th

Huang Xiejun. Data mining research[J]. Computer engineering and application. 2003.2