Research on Data Mining Methods

Ying Chen and Cheng Luo

Nanchang Institute of Science & Technology

Keywords: Data mining method; Definition; Current situation; Association analysis

Abstract. Data mining technology has a strong technical foundation, which is the result of a long-term research and development of database technology. At first, all kinds of business data is stored in the database of computers, and then the database can be queried and accessed, then the database can be real-time traversed. Data mining makes the database technology get into a more advanced stage where the database cannot only query and traverse the past data, but also can identify the potential links between the data in the past, so as to promote the transfer of information. This paper expounds the contradiction between the rapid expansion and high efficiency of the data and people's not getting the effective information and knowledge which are needed in scientific decision-making, puts forward the development and evolution of data mining, points out the prospect of data mining and finally describes all the work done in this paper.

Introduction

At present, the application of database has touched every aspect of human life. Banking, business, industry, agriculture, science and technology, military and other industries are in the application of database. According to statistics, in 1989, the total number of the database of the world was 5 million, and the number increased double each 20 months, in which the stored data was showing exponential growth. The millions of genetic gene in biology, the regular census of population of each country, geographic information of land and resources, the railway dynamic scheduling control, the cases of public security and judicial departments are all massive data. The main use of the data is mainly search and query and the efficiency is very low. What's more, a considerable number of data has a strong timeliness, and a lot of data has not been analyzed before it is outdated. The value of the data has not been fully used.

Facing the rapid expansion and the height limitation of data, on the one hand, people cannot timely get the reliable knowledge needed in scientific decision-making; on the other hand, a large number of valuable data resources are outdated before being used, which leads to a new problem, that is the so-called "rich data and poor knowledge". There is an urgent need to study a new generation of data processing technology, in order to improve the utilization rate of data. Data mining technology is produced under such a background. Its purpose is to analyze and process the massive data, to discover useful knowledge and to provide users with needed answer.

The Definition of Date Mining

In 1989, on the 11th International Conference on Artificial Intelligence in Detroit, America, Usama M.Fayyad and other people put forward the original descriptive definition of KDD which was a non trivial process of data to identify valid, novel, potentially useful and ultimately understandable patterns. But on the conference, the data mining definition did not given, which in fact led that many scholars have confusion of these two terms. Now data mining sector generally believes that data mining is the process to extract the information and knowledge which are implied in a large number of uncompleted, noisy, fuzzy, random and practical data that are not known by people in advance but is potentially useful.

This definition includes several layers of meaning: the data source must be true, massive, noisy; the discovered knowledge must be interested in and can be accepted, understood and used by users; it is not required to find the knowledge which is universally applicable or is the new natural science and pure mathematics formula or mechanical theorem proving.

In fact, all the discovered knowledge is relative with specific conditions and constraints. The discovered knowledge is for a specific area, but also is easy to be understood by users and it is best to use natural language to express the results found. The knowledge which is obtained by data mining shall have the character of previously unknown. Previously unknown knowledge refers to the knowledge or information which is unexpected, even is contrary to the intuition. The more unexpected the information which is dig out, the more valuable will it be.

Research Status of Data Mining Methods

1. R.Agrawal and other people integrated the machine learning and database technology, dealt with three kinds of data mining which are classification, association and sequence as a rule which is unified and reserves huge amounts of data, gave a unified model and several basic operations in rule discovery process and put forward how the data mining problem map the model and how to solve the problems found by the proposed basic operations. They proposed the classifier algorithm CDP which used the basic operating structure cannot only be effective in mining classification rules, but also has the accuracy of ID3 (ID3 is one of the best classifiers currently).

2. S.Anand and other people proposed general framework for data mining EDM based on Evident Theory. The algorithm developed in EDM framework is parallel, which has good efficiency in mining the distributed and heterogeneous data sets. The uses' prior knowledge and previously identified can be coupled to discovery process; also in (to meet the excavations on the minimum support and minimum confidence of association rules) of the excavation and spatial database, were used to test the proposed method.

3. For a wide variety of data mining techniques, Wei-Ming Shen and BingLeng proposed an automatic mining integration method based on meta model (meta query), which was different from the integration method which is like tool box. The concept of meta patterns they proposed is facilitated for the interdependence of the automatic use of induction, deduction and the human guidance

4. Visualization has become a trend in the entire computer industry and it is one of the important research directions of data mining. In the field of data mining, there are many similarities in automatic knowledge discovery and visualization. The data mining of visualization is also helpful in the interpretation of data analysis, and it is also has a great potential in large data sets mining.

Association Analysis of Data Mining

Association analysis is the most mature part of the research in data mining. Association analysis is to find interesting association or association relationship between sets in a large number of data. With a large amount of data constantly being collected and stored, many business people are becoming increasingly interested in finding interesting association from their databases. Finding interesting association from a large number of business affairs can help to develop many business decisions, such as the classification of design, the cross shopping and discount analysis. A typical example is the shopping basket analysis. This process analyzes the buying habits of customers by analyzing the association of the different goods which customers buy at the same time. To understand those goods which are bought by customers frequently can help retailers to develop marketing strategies, which further stimulates customers buy these goods at the same time.

For example, set $I = \{i_1, i_2, \cdots i_n\}$ as the collection of items. The goods collection which is mentioned is a collection of items. Set D as the collection of database transaction, in which each transaction T is the collection of items. $T \subseteq I$. Each transaction has an identifier which is called TID. Set A as a collection of items, transaction T contains A when and only when $A \subseteq T$. The contain relationships of the technology which is got by association analysis is as follows: $A \rightarrow B$, in which $A \subset I$, $B \subset A$ and $A \cap B = \varphi$, which is called association rules. An association rule can be described by Support, Confidence, Expected Confidence and Life.

Table 1 The calculation formula of the four association rules

Names	Description	Formula
Support	The probability of item sets A and B appear at the same time	$P(A \cup B)$
Confidence	When item set A appears, the probability of item set B appears	$P(B\|A)$
Expected Confidence	The probability of item set B appears	$P(B)$
Life	Ratio of confidence to expected confidence	$P(B\|A)/P(B)$

Association analysis is to find out the association rules of MinSup and MinConf which are given by users in the database, which can be decomposed into two problems:

(1) All frequent sets exist in the database. Support(X) of item set X is no less than MinSup which is given by users, than X can be called frequent set.

(2) Use frequent sets to generate association rules. For each frequent set A, if B A, $B \neq \varphi$ and Confidence ($B \rightarrow (A- B)) \geq$ MinConf, the association rule $B \rightarrow (A- B)$ is generated.

Conclusion

Data mining has attracted a great attention in information industry. The main reason is that it can be widely used because there are large amounts of data which is in urgent need to transfer these data into useful information and knowledge. The information and knowledge acquired can be widely used in a variety of applications, including business management, production control, market analysis, engineering design and scientific exploration. According to the existing theory of data mining, this paper discusses the association analysis of data mining, and the classical method of association analysis and the negative association analysis method based on interests. Through an example, this paper points out that the classical association analysis method has the wrong association rules in the "support degree confidence" framework; And in view of this situation, this paper puts forward the negative association analysis method based on interests. Analyzing the cited example, this paper shows that this method can mine the negative association rules which will be more practical and be interested in by users. It is believed that the association analysis can be more used in data mining in the modern society.

References

[1] Chen M S, Han J, Yu P S. Data Mining: An Overview from a Database Perspective [J]. IEEE Transactions on Knowledge & Data Engineering, 1997, 8(6):866-883.

[2] Azuaje F. Witten IH, Frank E: Data Mining: Practical Machine Learning Tools and Techniques [J]. Biomedical Engineering Online, 2006, 5(1):1-2.

[3] Pereziratxeta C, Bork P, Andrade M A. Association of genes to genetically inherited diseases using data mining. [J]. Nature Genetics, 2002, 31(3):316-9.

[4] Lu Z Q J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction [M]// The elements of statistical learning: Springer, 2001:192-192.

[5] Fayyad U, Piatetsky-Shapiro G, Smyth P. Knowledge Discovery and Data Mining: Towards a Unifying Framework [J]. Shapiro, 2010:82--88.

[6] Pudi V. Data Mining: Concepts and Techniques[C]// Oxford University Press, 2009.

[7] Piel W H, Sanderson M J, Donoghue M J. The Small-world dynamics of tree networks and data mining in phyloinformatics [J]. Bioinformatics, 2003, 19(9):1162-8.

[8] Lee W, Stolfo S J, Mok K W. A data mining framework for building intrusion detection models [J]. Proceedings of the IEEE Symposium on Security & Privacy, 1999:120-132.

[9] Fayyad U M, Piatetsky-Shapiro G G, Smyth P, et al. Advances in Knowledge Discovery & Data Mining [J]. Technometrics, 1996, 40(1).

[10] Tiffin N, Kelso J F, Powell a R, et al. Integration of text- and data-mining using ontologies successfully selects disease gene candidates [J]. Nucleic Acids Research, 2005, 33(5):1544-52.

[11] Apté C, Weiss S. Data mining with decision trees and decision rules [J]. Future Generation Computer Systems, 1997, 13(2):197-210.

[12] Larose D. Data mining methods and models [M]. Wiley-Interscience, 2006.

