2014

ISSN : 0974 - 7435 Volume 10 Issue 20

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(20), 2014 [12526-12531]

Exploration on the data mining system construction based on massive database

Yunfeng Zhang[1], Huan Wang[2], Jie Zhu[1]
[1]Computer Science & Engineering Department, North China Institute of Aerospace Engineering, Langfang, 065000, (CHINA)
[2]Science Department, North China Institute of Aerospace Engineering, Langfang, 065000, (CHINA)

ABSTRACT

After years of rapid development, database and information technology have already evolved from the original file processing into a complicated and powerful database system. Nowadays, the primary task is how to further explore the database functions. This paper explores the data mining system construction based on a massive database. Constructing the data mining system can further explore the database functions based on a massive database, which is beneficial to the decision-makers to find useful data fast and accurately. So they can make the most reasonable and effective decisions on the basis of these data. According to the authors’ experiences, this research provides a few constructing ways of the mining system, which can be shared to other peers.

KEYWORDS

A massive database; The data mining system; Construction.

INTRODUCTION

Because of the construction of the database, it is possible to store and extract massive electronic information. But the digital information resource database is very huge. Therefore, one of the vitally important research topics in the field of a massive database is how to research and extract the useful information faster and better. The data mining technology means the unique means and processes to search the useful information in the database. It is essential for the users to master a way of data mining so that they can more rapidly and accurately get the information. Hence, this paper explores the data mining system construction based on a massive database. According to the authors' experiences, this research provides a few constructing ways of the mining system, which can be shared to the other peers.

THE DATA MINING TECHNOLOGY

The content and essence of the data mining

With the deepening of the research on the data mining technology, it is clearly recognized that there are three main technical supports for the research of the data mining, including database, artificial intelligence and mathematical statistics.

The database

In 1980s, the database was very popular. It was applied in all walks of life and it had become a trend and culture. On one hand, the culture of database spread fast, and the database as the foundation of the knowledge source was very solid. On the other hand, the starting point of knowledge acquisition was increased greatly owing to the fact that the database formalized and organized a special field. So the knowledge would be searched and extracted belonging to the database in the future. Therefore, under this circumstance, the database scientists have invested more energy to study the data mining.

Artificial intelligence

The experts in Artificial Intelligence have embarked on inference based on cases. In particular, the scientists in machine learning are not satisfied with the ivory tower which they constructed in small sample learning mode. They have paid attention to the big data sample which is partial, noisy, numerous, random and ambiguous in the reality. Gradually, they have also devoted themselves to the data mining research.

Mathematical statistics

Mathematical statistics is a very active and vital subject in applied mathematics. It was born earlier than the invention of the computer and there have been a history of hundreds of years. The foundation of modern information consulting industry is the powerful and effective mathematical statistics means and instruments. At the current information era, consulting industry has been well- developed, but the combination of the database and mathematical statistics has still been at the first stage and demanded a long time to develop. The knowledge can be discovered by the data mining as shown in TABLE 1.

TABLE 1 : The knowledge types discovered by the data mining

Types	Definition
Generalization Knowledge	The knowledge reflects the common characters of the same type stuff.
Association Knowledge	The knowledge reflects the dependence and association between stuffs.
Deviation Knowledge	The knowledge reveals the abnormal phenomenon of the stuff deviating from the norm.
Difference Knowledge	The knowledge reflects the attribute differences between different stuffs
Prediction Knowledge	Predict the future data on the basis of the historical and current data.
Characteristic Knowledge	The knowledge reflects all aspects of characteristics of the stuffs.

All knowledge can be found at different levels of conception. With the increase in the conception tree through microcosm, mesocosm and macrocosm, the demands to make decisions of different clients from different levels can be satisfied. For instance, in the database of a supermarket, a classic association rule can be easily concluded. The clients who buy bread and butter stand a good chance to buy milk too. Besides, there is a great possibility that the clients who buy food use the credit card. The rule concluded from the database is very helpful for the business to make the sales plan and the strategy.

The procedures of the database mining and the composition of the mining system

The database mining is the process to discover the knowledge. The specific procedures are shown in Figure 1. There are seven steps.

(1) Data cleaning: clean noise or inconsistent data;

(2) Data integration: combine data sources;

(3) Data selection: search and analyze the related data in the database;
(4) Data conversion: convert or unify the data to be the form of a suitable for mining, for example, aggregate operation, summary and so on;
(5) Data mining: the basic step, extract the data model by using the intelligent method;
(6) Data evaluation: measure the data according to some interesting degree and recognize the really interesting model represented the knowledge;
(7) Knowledge representation: provide the mined knowledge for the users by the means of knowledge representation and visualization technology.

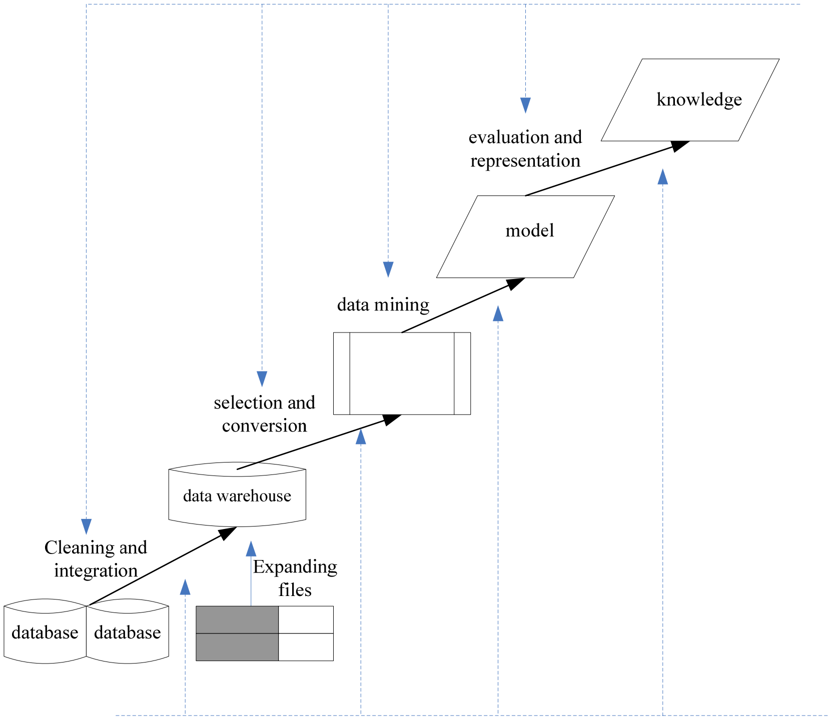

Figure 1 : The procedures of the data mining

The database mining procedures can realize the interaction with the knowledge database or the users, providing the interesting models for the users and storing the new knowledge in the database. From this point of view, the data mining just one step in the whole process but the most important step. However, in the database industry and the database research community, the knowledge discovery in the database is a normal term in the data mining. Thus, the generalized point of the database mining is that data mining means the process to mine the interesting knowledge from the numerous data stored in the database. From this point of view, the database mining system includes the following parts, whose system construction is as shown in Figure 2.

(1) Database: a database can restore information and it is possible to integrate or clean data in the database;
(2) Data warehouse server: according to the users' demands of the data mining, extract the related data in the data warehouse server;
(3) Knowledge base: it is a kind of domain knowledge used to evaluate the interesting degree of the result pattern or guide the research. The concept of this kind of knowledge base is hierarchical, including the knowledge about the users' confirmation;
(4) Data mining engine: it is the essential part of the data mining, which is combined by a group of function module, and it is used to characterize, classify, convert and deviation analyze;
(5) Data evaluation module: it is generally measured by the interesting degree, and it can interact with the data mining module, which makes it possible for the search to focus on the interesting model;
(6) Graphical user interface: the communication between the users and the data mining system is realized in this module, which allows the system to interact with the users, assigns the data mining, provides information, helps search focus, and does the exploratory data mining.

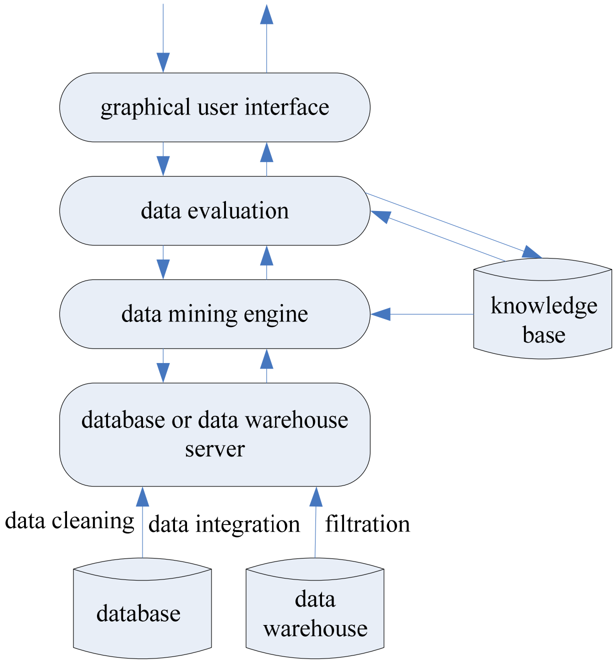

Figure 2 : The classic data mining system

The functions of the data mining system

Data cleaning and generalization

The data mining system can generalize the existing data to a higher level. Applying the generalized integration algorithm of GDBR, the complexity of the space and time can be related conditionally. Then, the method of adopting N-Gram can search effectively and accurately for the repeated records which are similar in the system, and then sort and test them. Intelligent operations such as normative insertion, deletion, exchange and replacement can cope with the common spelling mistake, cleaning the data. But there are a few deviations in the precision detection by adopting the normal elimination basic algorithm, so this system improves the elimination basic algorithm, applying the principles of statistics reasonably and combining with the direct and converse repeat matrixes, which is able to increase the detection rate of spelling mistake and the correct rate of modification.

The mining function of the data

According to the related association rule and the time sequence rule, the system classifies and aggregates the data by the means of the data mining, achieving the expected application aim of the data mining system. Realize Apriori algorithm by searching and integrating the frequent item sets among the data. Then, form the association rule by the frequent item sets. The methods are as following. Suppose record the frequent item set as *I*, and record all the nonvoid subsets in *I* as *a*. If the value of support (*I*)/support (*a*) is greater than minconf, the rule a=> (1-a) will be output directly. If the nonvoid subsets in *I* do not correspond to the condition, the related rule will not be output. In other words, the association rule is not formed as *a*. Similar with the association rule, but the time sequence rule tends to the time association of the item sets in the system. The time sequence rule in this system is formed by AprionAII. In a broad sense, the association rule includes strong rule, exception rule and random rule. The rule a small amount of data obeyed is represented by exception rule. Although the number is small, its confidence is high. It is the rule formed for the unknown information at this stage and unpredictable information. The system setting of the minimum confidence is realized by the exception association rule. So the system can form CAR and ECAR and delete SCAR.

For those well defined data and classified data, the data category which is representative can be formed, and the according classification standard of the data belonging to the unknown category can be formed, which is the classifier. In this system, applying the interval classifier can increase the level of the correct rate and classification accuracy and decrease the over deep tree extension of the decision tree classifier.

Clustering algorithm means to merge high-density clusters and adopt CURE algorithm to mark different clusters by many representative points. So a certain cluster distributed framework is formed. Then recognize effectively the special shape, which enlarges the data processing and enhances processing capacity. The main clustering means which this system adopts is the hierarchical cluster procedure. Before utilizing the means, the data mining system will divide space distribution automatically, which makes the information object form many data units. Then according to the characteristic of the unit, computer the clusters' distribution. Another unique clustering algorithm is the density-based clustering algorithm. Through improving Dbscan algorithm, the data division can be realized by the small division cluster, and the acceleration of the algorithm speed is realized by selecting the representative expanding seed points of the neighboring objects, and the whole database clustering is realized by the sample data clustering. It makes the clustering algorithm of the system more effective.

THE CONSTRUCTION METHOD OF THE DATA MINING SYSTEM BASED ON A MASSIVE DATABASE

The overall frame structure settings

The system integrates all kinds of related modules closely to form the data structure being hierarchical, including the distinctive operation function of multiple outputs, multiple data sources and multiple parameters. As a result, the independence between each mining operation module can be realized, which results a more functional and more stable system. As an integrated system, there is a coordinating and unifying association among modules, which promotes to realize standard and systematic operations of the data parameters, mining results, and data sources applied by each module. On account of the data mining system based on a massive database, the data mining range of the system must be enlarged, realizing the mining objects not only existing in the database, but also in the according files. So the according file information processing method is provided by the system. To present more easily the mining results and realize the remote decision support analysis, the system also has the function of restoring the mining results automatically, expanding the application range. Because it is people who operate the computer, this system is equipped with good operation interface. It is very convenient for the users and decision makers to do the decision analysis and make a accurate decision.

Module settings

According to the above frame structure of this system, the following modules are set especially to realize the related functions of the data mining system.

(1) The mining module can achieve the mining operation function, which mines different data in the database. Each mining module is independent. The database management module can control a single module. The storage module is the site of data source. Through the mining to read in the corresponding data in the mining base, the basis of data is provided for other modules.

(2) The main function of the preprocess module is to filter, define and format the data source, further improving the operability and practicability of the whole system. The main sub-modules are data mapping, column mapping and type mapping. The data mapping is to map a source table to become ID type, and then form the relevant comparison table. Different data are mapped to form a unified module which is worth to mine. The column mapping is to extract the useful column from the data source, which is benefit to decrease the number of data and accelerate the computing speed. The type mapping is to convert the type of data source. The conversion is compulsive to unify the different types of data in the database, which is benefit for mining.

(3) Storage module is to operate uniformly the data in the whole database. However, the external files must be imported first, and then are stored and controlled. ODBC technology is adopted at lower layer interface. Memory index and buffer function are utilized to accelerate the computing power of the system.

The core module of the whole system is the mining management module. All kinds of information the users achieved from the database by the mining shall be stored in the mining base. The mining base is directly set in the system database, which is convenient for transferring and management. The mining base management includes all types of operations in the process of the data mining, data preparation and data storage. When the mining base stores the operation information, it is sequential so that it is conducive to the convenient operations. But the data mining operation is dependent in the whole process of mining. It needs a source, which is the operation result of the other data mining and results a new mining. Besides, the new result may be the data source of another mining process.

Interface settings

The main interface of this system is similar with Explorer interface, which is very user-friendly, beautiful and high operability. Different mining results are presented by different graphic technologies. The system application form presents the results of generalization and cleaning. Tree structure presents the decision tree. Two dimension and three dimension points are used to present the clustering structure. All types of rules and models are presented by text.

CONCLUSION

At present, there are a number of researches on constructing the data mining system based on a massive database. All these researches aim to mine the database deeply, so the decision makers can find the useful data resource fast and accurately from the numerous data, and make an effective decision on the basis of these useful data resource. This paper explores the data mining system construction based on a massive database. Firstly, it introduces the data mining technology, including the content and essence of the data mining, the database mining procedures and the composition of the mining system, and the functions of the data mining system. Then, it discusses the methods of constructing the database mining system, including the overall frame structure settings, module settings and interface settings. Although it is simple, it has its own characteristics. Because an increasing number of related data integration systems are issued and gained acceptance, the enterprises shall construct database mining system on the basis of their own features and demands so that provide better service for them and improve the application benefit and economic benefit.

REFERENCES

[1] Shi Yong; Construction Analysis of Knowledge Intelligent Query Data Mining System, Coal Technology, **7**, **(2012)**.

[2] Tang Hui, Li Haichen, Wang Binbin; Framework of an Automated Data Mining System Based on Intelligent Agents, Journal of Library and Information Science in Agriculture, **12**, **(2010)**.

[3] Chen Yufeng, Zhang Hongyan, Jing Song; Study on Rural Migrant Workers Employment Recommend System Based on Data Mining, Journal of Anhui Agriculture, **33 (2011)**.