ISSN : 0974 - 7435 Volume 10 Issue 14

2014

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(14), 2014 [7849-7855]

Study on data mining technology based on MySQL database

Fen Liu
Institute of Computer Science, Yan'an University, Yan'an 716000, Shaanxi, (CHINA)

ABSTRACT

Data mining technology is a new hotspot in database research field. It combines the traditional data analysis technology with the complex arithmetic for dealing with big data, which helps to explore and analyze the new data type as well as deal with old data type by applying new method. The historical background for data mining technology is elaborated in this paper to make discussion by the classification on the main analysis technology involved in data mining technology and describe the arithmetic commonly used at present as well as point out the application areas and future trend of data mining technology.

KEYWORDS

Data mining; Association analysis; Classification analysis.

INTRODUCTION

With the rapid development of the internet and database technology, large amount of information is stored in the database, and it’s difficult to mine out the information we need from the database. However the data mining technology is the very method to help to solve the problem that the data we need has to be sought from the mass data, accordingly helping the enterprise to mine the effective data, to reduce the unnecessary expenditure. Now many companies have already adopted the data mining technology to help identify the group of the most valuable customers, and reenact publicity and promotion strategy for this reason, getting superior returns by smaller inputs. Data mining technology is also known as data exploration and data mining, generally refers to the process of the information which is hidden among them, previously unknown but with potential value extracted from the large amount of data with noise, which is ambiguous and incomplete and provided with randomness. This paper discusses the commonly used analysis technology for data mining at first.

ANALYSIS TECHNOLOGY FOR COMMONLY-USED DATA MINING

Generally speaking, according to the concrete type of data and model, the more technologies adopted in data mining system, the conclusion reached is more reliable. Classification shall be made from the functions, and the analysis technology for data mining can be divided into as the following.

Association analysis technology

The purpose of the association analysis is to mine out the interconnectedness between the large amounts of data. Now, a group of Items and a record set are given, association analysis is to infer the correlation between the items by analyzing the features of the set of records. For example, “30% is the record containing Item A, Item B and Item C, and the records contain Item E and Item F at the same time”, the 30% here is called credibility for rules of “the record containing Item A, Item B and Item C, contains Item E and Item F at the same time”, but Item A, Item B and Item C are called the opposite of Item E and Item F.

To facilitate the discussion, this paper takes a typical example of sale of goods of the market. The commodity retailer keeps all the detailed records of transaction, among them transaction number, customer number, article number, etc. are included. Now association analysis is made to the behavior of commodity s purchase of the customers, but the object of the association analysis is to put different kinds of commodity s into a group of Item. However commodity retailer keeps the transaction records and forms a record set for association analysis, among them each record is composed of transaction number, customer number, article number, quantity, date and other data of each transaction shown in TABLE 1.

TABLE 1 : Transaction table of a customer

Transaction number	Customer number	Article number	Quantity	Date
	A	A	14	3/4/95
	A	B	3	3/4/95
2	B	C	2	5/6/95
	B	B	3	5/6/95
	B	D	13	5/6/95
3	B	B	10	8/6/95
	B	D	12	8/6/95

Before carrying out the association analysis, two parameter values are required to be preset: 1) minimum confidence degree, its purpose is to filter out the rules with small possibility, and the minimum confidence degree can be set in this example as 0.4; 2) minimum support degree, its purpose is to express the probability that the specific rules may occur, the credibility calculated according to the minimum confidence degree is 0.4.

According to the above description, the rule can be set as “the customer purchasing commodity X may also purchase the commodity Y”, the confidence degree of the rule can be set as C, support degree is S. C= (trading quantity of not only the commodity X but also purchasing the commodity Y) / (trading quantity of purchasing the commodity X), S= (trading quantity of not only the commodity X but also purchasing the commodity Y) / (total trading quantity) can be gotten by the formula.

According to the data shown in TABLE 1, confidence degree C and support degree S of each rule can be gotten through above-mentioned formula, the data gotten is as shown in TABLE 2. The data in lines 2 of TABLE 2 shows that the confidence degree C of which the customer purchasing commodity B may also purchase commodity A at the same time is 0.33, and support degree S is 0.33. Market staff can accordingly learn that: there is 33% probability for the customer who has purchased induction cooker to buy the supporting tableware. This method is relatively simple, and can be mastered by common market personnel after some training, and the results are of certain reference value. The method doesn’t involve complex arithmetic, but from which, the basic approach and reference value of association analysis can be understood briefly.

TABLE 2 : An example for association analysis

Item 1	Item 2	Confidence Degree C	Support Degree S
A	B	1.00	0.33
B	A	0.33	0.33
B	C	0.33	0.33
B	D	0.66	0.66
C	B	1.00	0.33
C	D	1.00	0.33
D	B	1.00	0.66
D	C	0.50	0.66

Sequence pattern analysis technology

Sequence pattern analysis technology is for the purpose to mine out the correlation of large amounts of data, its purpose is more or less close to the association analysis technology, but the sequence pattern analysis technology focuses on mining the causal relationship between the data. The table shall be remade by the data from the above example and customer number, date, article number and quantities (referring to TABLE 3).

TABLE 3 : Table classified by time

Customer No.	Date	Article number	Quantity
A	3/4/95	A	14
	4/4/95	B	3
		C	11
		C	2
B	5/6/95	B	3
		D	13
	8/6/95	B	10
		D	12

Similarly, before conducting the sequence pattern analysis, minimum confidence degree C and minimum support degree S are also required to be preset. For example, in this example, the minimum confidence degree preset is 0.4, minimum support degree is 0.4. But the confidence degree of the rule of “the customer who purchases commodity X at first may purchase the commodity Y next” is C, support degree is S. C = (the number of group for the customer who purchases commodity X at first may purchase the commodity Y next) / (the number of group for purchasing commodity X at first); S = (the number of groups of the customer who purchases commodity X at first may purchase the commodity Y next) / (total number of group) can be gotten by the formula.

TABLE 4 : Sequence analysis record table

Item 1	Item 2	Confidence Degree C	Support Degree S
A	B	1.00	0.5
B	C	0.5	0.5
A,B	C	0.5	0.5
B	B	0.5	0.5
B	D	0.5	0.5
B	A,D	0.5	0.5
B,C,D	D	0.5	0.5

The data in TABLE 4 shows that by group as the reference object, the confidence degree for the customer may purchase commodity C after purchasing commodity B which can be obtained from the data in lines 2 of the table 4 is 0.5, support degree is 0.5. By analyzing the above data by sequence pattern, the commodity retailer can discover customer's potential shopping mode, for example, what commodity is most frequently bought by the custom before purchasing induction cooker, and the sales strategy of goods can be specified accordingly. In fact, sequence pattern analysis technology can also be adopted to stock analysis, for example, some law of stock can be found by sequence pattern analysis technology, and there are also good result in medical and insurance industries, insurance company can also use sequence pattern analysis technology to forecast the medical way the most commonly used by users buying insurance, to adjust insurance policy.

Classification analysis technology

Classification analysis technology is required to assume that the record set is provided with a group of marks, but the marks are required to be provided with characteristics differing from other categories. When preparing for classification analysis, each record shall be defined with a mark first, that is the classification shall be made by the marks when recording, and then the records with marks shall be inspected, and the characteristics of the records with marks shall be described finally. However the description methods can be multiple methods, and can be explicit way, for instance defining directly a group of rules or characteristics; and can also be implicit, for instance defining a mathematical formula or a mapping mode. Nowadays many classification models are beginning to use classification analysis technology to analyze and forecast.

As an analogy, the records of various customers are kept in the database of all major banks, and may be classified according to personal credit degree (marks), and can be divided into several kinds: good, general, poor, very bad, etc. The way of classification is classification analysis technology. Now the bank can give a explicit description for credit scores handily by classification analysis technology: "the users with general credit have about RMB 5000 monthly income generally, with age between 25 and 35, and most of them live in a certain areas".

Clustering analysis technology

The process forming by classification the set of the physical or abstract objects into multiple categories made up of similar objects is called clustering. But the input set of clustering analysis technology is a group of records with marks, that is the existing input set is not classified yet, its purpose is to divide the record set reasonably according to specific rules, and then use explicit or implicit rules to classify. In general, clustering analysis technology can be adopted with various different algorithms, so the same record set may be provided with different dividing results, but the above classification analysis method is basically suitable for clustering analysis method in the same way. Actually, classification analysis technology is perfect mutually with clustering analysis technology in many aspects, for example: during the early data analysis, analyst can mark, divide categories to the data to be analyzed according to experience or general rules, and then analyze the date by classification analysis technology, obtaining general description for each category, next re-divide the set by using these description as a new classification rule, getting better dividing result accordingly, analyst can use the two analysis technologies circularly, to get satisfactory results.

In fact, for data mining system, sometimes the above four methods shall be used comprehensively. Take sale of goods as an example, now market positioning shall be conducted to a certain commodity, data mining system may be adopted with the four analysis method comprehensively: 1) generalize the commodity commonly bought together with this commodity by association analysis technology; 2) find out the user group purchasing this commodity by sequence analysis technology, and summarize the order of their shopping; 3) summarize the shopping mode of this commodity by classification analysis technology according to the result of sequence analysis technology; 4) take the above summarized shopping mode as the rule for clustering analysis technology, apply clustering analysis method to find the user with this shopping mode but not buying this commodity, and focus on propaganda to it.

GENERAL TECHNIQUES FOR DATA MINING

Neural network

Neural network is also called as connection model, which is an algorithm mathematics model by simulating animal's neural characteristics to treat information by distributing and paralleling, which is often used to solve classification and regression problems. Neural network can be specifically deemed as a group of connected output or input unit, among which each connection may be connected with a weight. During the learning stage of the neural network, the weight of the neural network can be adjusted to enable it can forecast the accurate category number of samples to learn.

Decision tree

The purpose of decision tree is very wide, now decision tree can be used to discriminate that the method of rules of a certain value can be gotten under what conditions. For example when processing bank loan, the bank is required to estimate the size of the risk resulted from loan of the applicant.

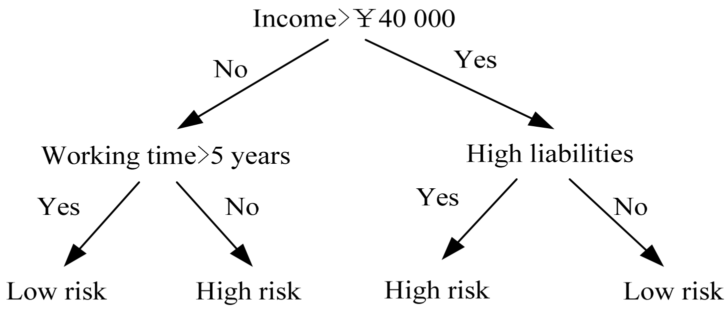

Figure 1 : A simple decision tree

For example, bank employees can use the decision tree in Figure 1 to decide whether gives the loans or rejects the loans, and judge the size of loan risk by amount of loans. For example, the user with "income>40 000" and "high liabilities" may be deemed as "high risk", but the user with "income<40 000" and "Working time>5 years" may deemed as "low risk", and will be allowed for loan.

The decision tree used in data mining technology can be used to analyze data and also can be predicted, the decision tree shown in Figure 1 can be forecasted. The algorithms commonly used for forecasting are: CHAID, Quest, CART and C5.0. Provided data in the database is scanned, situations with variable can be treated easily by setting up decision tree then. But the decision tree set up shall not be too large, otherwise the forecasting effect may be affected, at this moment, the method of the maximum number of layers of decision tree can be set to restrict the scale of the decision tree, the other restricting method is to preset node to include the smallest record count, when the record count included in the node is less than the minimum record count, the partition will be stopped. Sometimes the decision tree can be allowed to grow as much as possible, and then the decision tree can be clipped to the size as required, attention shall be paid to keep the accuracy of the decision tree during the process of clipping.

When treating the decision tree to divide the node, greedy algorithm can be adopted, the algorithm thought is: all the division shall be carried out in order, a node will not be considered the rationality after being divided that is each division depends on the rationality of the former division. The greedy algorithm is least used due to the boundedness.

Relatively speaking, the decision tree is good at dealing with the nonnumeric data, such as bank information; but neural network is good at treating numeric data.

Genetic algorithm

Genetic algorithm is a kind of algorithm with randomization evolved by learning from evolutionary law (such as, survival of the fittest, selective genetic mechanism) in the biological world, to search optimal solution. The characteristics are can be operated for the object directly, without limit of derivation and continuous function, being provided with internal implicit parallelism and global convergence ability; but use of the convergence method with randomization can obtain and optimize the region of search automatically and can adjust the search direction adaptively. Genetic algorithm consists of three basic operators: 1) reproduction ion (selection), is the process of selecting individual with strong vitality from parent (former population), accordingly generating a new population; 2) intersecting (recombination), is the process of selecting two different individuals for intersecting and recombination, thereby generating a new individual; 3) variation (mutation), is the process of making mutation treatment to certain parts of certain individual, thereby generating a new individual.

The genetic algorithm can play a role in improving the offspring, which has brought pivotal effect to the aspects of optimal computation and the classified machines study.

PROCESS OF DATA MINING

The basic procedure for data mining process shall be further discussed (as shown in Figure 2), the procedures for data mining include: 1) determine the study object, and the process of the determination is the base of the whole process of data mining; 2) build up the database, thus the data can be collected, described, selected and preprocessed; 3) analyze the data, the commonly used analysis technologies introduced in this paper (such as association analysis technology, sequence pattern analysis technology and other analysis technologies) can be applied to analyze data, and summarize the features of the data; 4) prepare data, the preparing procedures include selecting variable, methods for recording and converting variable. 5) set up model, suitable mining algorithm is required to be selected to mine the converted data; 6) evaluate and explain, evaluate the conclusion of the model set up, and make necessary explanation; 7) implementation (assimilation of knowledge), the conclusion gotten through analysis shall be adopted to the system with same type.

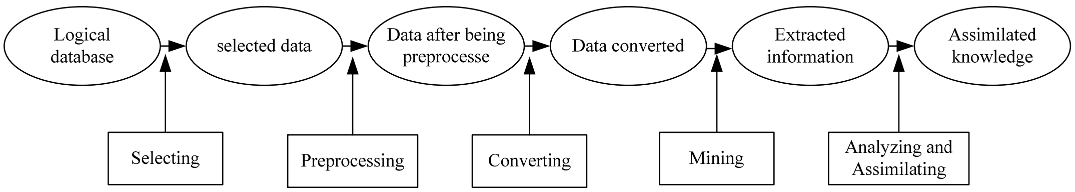

Figure 2 : Step of data mining process

APPLICATION OF DATA MINING TECHNOLOGY

Data mining technology for biomedicine and DNA data analysis

At present, a large amount of research for biomedicine is mainly focused on the data analysis of DNA sequence, causes for many diseases and disability genes have been found through DNA sequence analysis. The gene sequence appearing at the same time can be identified by association analysis technology; the morbigenous gene sequence of diseases in different stages can be found by path analysis technology; tools for visualization and genetic data analysis technology can also be used to show complex structure and sequence pattern of gene.

Data mining technology for finance

The bank and financing institution all may provide various saving business, credit business, investment and managing finances service and insurance business. For its business characteristics and data features, the following several typical data mining technologies are being given: 1. database is required to be designed and established for features of multidimensional data; 2. the data shall be analyzed by loan repayment predicting technique and customer credit policy analysis method; 3. make classification and clustering treatment for customers of the specific target market.

Application of data mining technology for retail industry

Retail industry is the main application and research field of data mining technology, retail industry has accumulated a large amount of sales data for a long term, such as, purchasing record, consuming and service records of customers and records in and out of goods. The data mining technology of retail industry is helpful for merchant to identify customers’ purchasing behaviors and shopping mode, to improve service quality and improve the sales ratio of goods, reduce the cost.

The data mining technologies involved are: design and build up database by data mining technologies; make multidimensional analysis to sales, customers, goods, time and areas; make analysis to customers’ shopping loyalty.

FUTURE TREND OF DATA MINING TECHNOLOGY

Due to multiple forms of the data, data mining task and data mining ways, many study directions have been given to data mining technology, such as application expanding of data mining technology; data mining technology with elasticity; data mining tools with visualization; new mining technology for complex data types; protection of privacy, information safety and other directions in the process of data mining.

CONCLUSION

Data mining technology is adopted with method of artificial intelligence to make statistics analysis to data in database, to mine law of them, including multiple analysis techniques, commonly used algorithm, and the application field of data mining technology is very wide, with large development potential in future.

REFERENCES

[1] Chen Jianjun, Sheng Yizhi, Chen Jinyun; Application Study of OLAP Based on Basic Data Warehouse in DSS [J], **1**, 30-31 **(2003)**.

[2] Zhu Ming; Data Mining [M], Hefei: Publishing House of University of Science and Technology of China, **(2002)**.

[3] Chen Jingmin, et al; Data Warehouse and Data Mining Technology [M], Beijing: Publishing House of Electronic Industry, **(2002)**.

[4] Mao Guojun; Concept of Data Mining [J], Computer Engineering and Design, **23(8)**,13-17 **(2002)**.

[5] Chen Wenwei, et al;. Data Mining Technology [M], Beijing: Publishing House of Beijing University of Technology, **(2002)**.