ISSN : 0974 - 7435 Volume 10 Issue 13

2014

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(13), 2014 [7602-7609]

XML data mining research based on multi-level technology

Jiangrong Cheng

Longnan Teachers College, Gansu Chengxian, 742500, (CHINA)

ABSTRACT

Data mining technology is a new hot-spot in database research, which combines traditional data analysis technique with relatively sophisticated algorithms for dealing with mass data, exploring and analyzing new data type and processing old data type with new method. After the emergency of XML technology, it has soon become to the standard about indicating and exchanging of information, also providing a new method for data mining. This article explains the background of data mining technology, decision tree, XML technique and discusses the basic process of data mining based on XML, proposing a decision tree analysis model in XML database mining, and finally pointing out the application field and future trend of data mining technology.

KEYWORDS

Data mining; Multi-level technology; XML technique.

INTRODUCTION

With the rapid development of the Internet and database technology, the huge mount of data was stored in the database, which makes difficult to find the needed information. The data mining technology is the exact method to solve this problem and to look for the needed data in mass data, thus can help company dig the valid data and reduce unnecessary expense. Recently, many companies have already adopted the data mining technology to identify the most valuable customer groups, and then revise their promotion strategy which aims to get higher return with lower investment. The data mining is also called data exploration, mining of data, which generally refers to the process of extracting the hidden, unknown, but potentially valuable information among large number of noisy, ambiguous, incomplete and random data. This article will firstly talk about the data mining analysis technique.

COMMONLY USED ANALYSIS TECHNOLOGY OF DATA MINING

Generally speaking, according to the specific type of data and model, the more technology of data mining system has been used, the more reliable conclusion will be get. In terms of function, the analysis technology of data mining can be classified into following techniques:

Association analysis technique

The purpose of association analysis is to dig out the correlation among the large amounts of data. Given a set of Item and a record collection, the association analysis is to know the correlation among Item through analyzing the characteristics of record collection. For example, “30% is the record including Item A, Item B and Item C, which also includes Item E and Item F”, the 30% is referred to the credibility of rule for “the record including Item A, Item B and Item C, which also includes Item E and Item F”, and Item A, Item B and Item C are called the antithesis to Item E and Item F.

This article will use a typical example of goods sale for easy discussion. The goods retailer keeps all the transaction record which includes transaction number, customer number and item number, etc. Now it has association analysis on customer’s buying behavior, and the object of association analysis is a group of Item of different goods type. The goods retailer saves the transaction record and forms a record collection for association analysis, each record, as shown in TABLE 1, is composed of transaction number, customer number, item number and date of every transaction.

TABLE 1 : Climate training set

Number	property				category
	Weather	Temperature	humidity	Wind	
1	sunny	hot	hot	windless	N
2	sunny	hot	hot	windless	N
3	cloudy	hot	hot	windless	P
4	rainy	moderate	hot	windless	P
5	rainy	cold	normal	windless	P
6	rainy	cold	normal	windy	N
7	cloudy	cold	normal	windy	P
8	sunny	hot	hot	windless	N
9	sunny	cold	normal	windless	P
10	rainy	moderate	normal	windless	P
11	sunny	moderate	normal	windy	P
12	cloudy	moderate	hot	windy	P
13	cloudy	hot	normal	windless	P
14	rainy	moderate	hot	windy	N

Before doing association analysis, it needs to preset two parameters: 1) the minimum degree of confidence, whose role is to filter out the possibility of smaller rules, and in this case, you can set a minimum confidence level of 0.4; 2) the minimum support, whose role is to represent the probability of occurrence of a specific rule, and the credibility value will be 0.4 based on the minimum confidence level.

According to the above description, it can set rule as ”customer who buys product X will also buys product Y”, and setting the confidence level as C for this rule, so doing support level as S. Using formula to obtain: C= (the transaction number of both purchase product X and product Y)/ (the transaction number of only purchase product X), S= (the transaction number of both purchase product X and product Y)/ (total transaction number).

Using the data in TABLE 1, getting the confidence level C and support level S of each rule through the above formula, the result is shown in TABLE 2. By the second row of data in TABLE 2, it can be seen that the confidence level C of the customer who buys product B will also buys product A is 0.33, and the support level is 0.33. Thus the market staff will know that the customer who buys electromagnetic oven will has 33% chances to buy the matching tableware. This usage is relatively simple for the ordinary market staff who will master it after some training, the results have a certain reference value. This method does not involve complex algorithm, but from which can understand the basic method and reference value of association analysis.

TABLE 2 : Mining rule number and time for rstbor and godriis(s)

data collection	GODRIIS		RSTBORD	
	rule number	time	rule number	time
Mushroom	9	28.80	11	2.00
Anneal	16	5.90	16	1.20
Balance-scale	5	0.40	8	0.05
Restricted	15	40.80	15	0.20

Sequential pattern analysis technique

The Sequential pattern analysis technique is also to dig out the relevance of large amounts of data, the purpose is somewhat similar to association analysis technique, but sequential pattern analysis focuses on the causal relationship of data mining. With the above examples of data and uses the customer number, date, number and quantity of goods to rebuild the table (see TABLE 3).

TABLE 3 : Classification by timetable

Customer Number	date	product number	quantity
1	3/4/95	A	14
		B	3
	4/4/95	C	11
		C	2
2	5/6/95	B	3
		D	13
	8/6/95	B	10
		D	12

Similarly, it needs to preset the minimum confidence level C and minimum support level S before doing the sequential pattern analysis. As in this example, presetting the minimum confidence level as 0.4, and so doing the minimum support level as 0.4. And the degree of confidence of rule "customer who first buys the product X will then buy goods Y" is C, and the support level is S. Using

the formula to Obtain: C = (the group number of customer who first buys the product X will then buy goods Y) / (the group number of first buying product X); S = (the group number of customer who first buys the product X will then buy goods Y) / (total number of groups).

TABLE 4 : Record form of sequence analysis

Item 1	Item 2	confidence level C	support level S
A	B	1.0	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,DB	D	0.5	0.5

As shown in the TABLE 4, choosing the group as a reference object, from the data of second row in TABLE 4 can know that the confidence level of customer who buys product B and then buys Product C is 0.5, and the support level of it is 0.5. Using sequential model to analyze the above data, the product retailer will find the potential way for shopping of the customers, for example, which product does the customer usually buy before purchasing the electromagnetic oven, and then has the sales strategy for the product. The sequential model analysis technique can also be used in stock analysis, for example, it can find some regular pattern through sequential model analysis technique, in medical insurance industry it also has a good effect, the insurance company can use this new technique to predict the most common medical approach and then adjust the insurance policies.

Classification analysis technique

The classification analysis technique needs to assume the record collection has a set of mark which needs to have the characteristics to distinguish from other categories. In preparation of classification analysis, it needs to define a mark for every record firstly, which means to classify the records according to mark when doing the record, then check out those records with marks, finally, describing the characteristics of those records with marks. The description method can be different, it can be explicit way, such ac directly define a set of rule or characteristic: and it can be implicit, such as the definition of a mathematical formula of a mapping method. Now, many classification model has started to adopt classification analysis technique to analyze and predict.

For example, the database of major banks save the records of various customers, and have classification according to personal credit degree (namely the mark) which can be divided into: good, fair, poor, very poor, etc. This classification method is classification analysis technique. Banks can use classification analysis technique to give an explicit description of the credit rating as:”the fair credit consumers generally refers to ones who have monthly income of 5000RMB, the age between 25 to 35, and mostly living in certain area,”

Cluster analysis technique

Clustering is the process of classifying the collection of physical or abstract objects and compose the multiple categories of similar objects. The input set of cluster analysis technique is a group of marked records, which is not classified, and the purpose is to divide the record collection reasonably according to specified rules, then adopting explicit or implicit rule to classify. Generally, cluster analysis technique can adopt various algorithms, so the same record collection may has different classification result, and the above classification analysis substantially applies to cluster analysis. In fact, the classification analysis technique and the cluster analysis technique is complementary in many aspects, such as, during the early data analysis, the analyst can label and classify the data according to experience or general rule, then analyze the data with classification analysis technique, getting a general description

of each category, then using the description as a new classification rule to re-divide the data, thus can get a better division effect, and analyst can cycle use the two analysis techniques to obtain a satisfactory result.

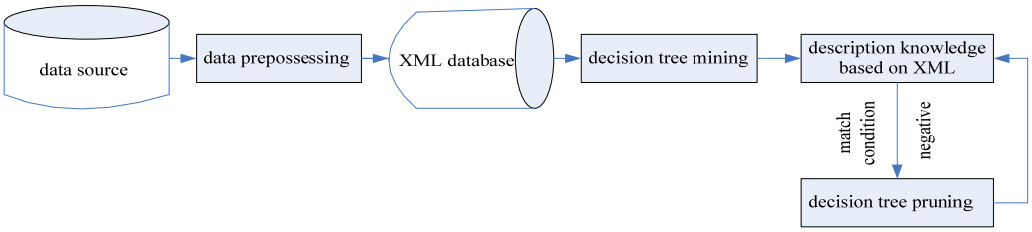

Figure 1 : Data mining model based on XML

In terms of data mining system, sometimes it needs integrated utilization of four methods that described above. To take product sales as example, now it needs to have market position for a certain product, the data mining system may comprehensively use four methods: 1) using association analysis technique to sum up the product which is usually bought with the certain product; 2) using sequential analysis technique to find the customer group of purchasing the certain product, and summarize the sequence of shopping; 3) using classification analysis technique to sum up the shopping mode of the certain product according to the result of sequential analysis technique; 4) using the above shopping mode as the rule of cluster analysis technique, and adopting cluster analysis method to find the potential customer with the same shopping mode and having a promotion for the potential.

COMMONLY USED TECHNOLOGY OF DATA MINING

Neural network

Neural network is also known as connection model, which is the algorithm mathematical model of an intimation of animal nervous feature, and having distribution parallel information processing which is usually used to solve the classification and regression problems. Neural network is considered as a set of connected input or output unit, each of which is connected to a power. In the learning phase of neural network, people can adjust the weights of the neural network so that it can accurately predict the category number of samples for learning.

Decision tree

Decision tree has an extensive usage, now it can be used to judge the condition under which the value can be get with certain rule method. For example, when bank staff dealing with the loan, it needs to judge the risk of the loan applicant. The ID3 decision Tree is shown as Figure 2.

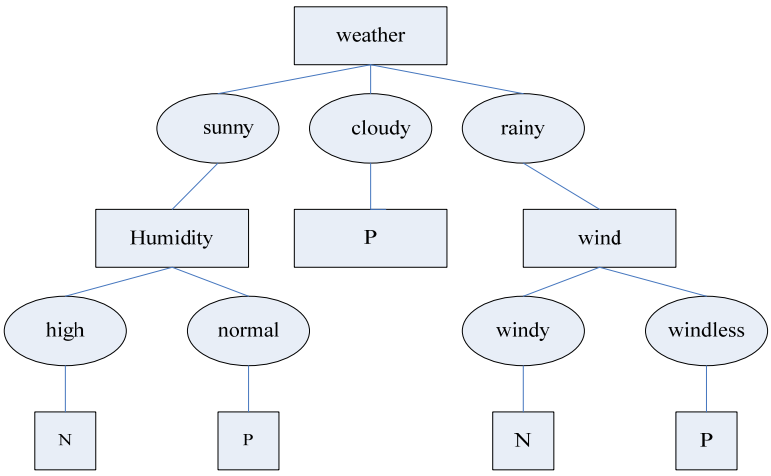

Figure 2 : ID3 decision tree

For example, bank employees can use the decision tree as shown in Figure 1 to decide whether granting or refusing the loan applicant, and using the size of loan to determine the loan risk. Such as “income > 40,000” and “ high liability” consumer will be treated as “high risk”, otherwise, "income <40,000" and "working time> 5 years," user will be considered as "low risk "and get the loan.

The decision tree which used in data mining technology can be used in analyzing data and predicting, the decision tree which is shown in Figure 1 can be used to predict. The commonly used algorithms for predicting are: CHAID, Quest, CART and C5.0, etc. As long as scanning the data in database, then creating a decision tree which can easily handle cases with variable. But, the decision tree can not be too huge to effect the predicting result, setting the maximum layer of decision tree can be used to limit the size of decision tree, another method is to preset node which contains the smallest record, when the node contains the record that is smaller than the smallest record, it is the time to stop the division. Sometimes it can let decision tree to grow freely, then prune it to the desired size, but the accuracy of the decision tree need to be maintained during the pruning,

When dealing with the node division of the decision tree, it can adopt greedy algorithm whose idea is: all division are based on sequential order, the rationality is not considered after division of each node, which means each division is dependent on the rationality of the last time division. Since the limitation of the greedy algorithm, it is seldom used.

Relatively speaking, the decision tree is good at dealing with non-numerical data, such as bank information; while neural network is good at processing numerical data.

Genetic algorithm

Genetic algorithm borrows the biological evolution law (such as survival of the fittest, survival of the fittest genetic mechanism) and evolves into a kind of algorithm which has ability of random searching for the optimal solution. Its characteristic is to operate directly on the object without the limitation of deviation and continuous function, and having the inherently implicit parallelism and overall optimization capability; the adoption of probabilistic optimizing method can automatically obtain and optimize searching field, and have self-adapt adjustment of searching direction. The genetic algorithm consists of three basic operators: 1) reproduction (selection), namely choosing a more vital individual from the parental generation (former population), thus generating a new population; 2)crossing (recombination), which means to select two different parts of individual to have genetic crossing and recombination, thus generating new individual; 3) variation (mutation), i.e. having mutation treatment on some parts of some individuals, thus generating a new individual.

Genetic algorithm can play a role in improving offspring, which has played a key role in optimizing calculation and classified machine learning

PROCESS OF DATA MINING

Now to have a further discussion on basic steps of data mining process (as shown in Figure 3), data mining steps include:1) determine the research object, this determination process is the foundation of whole mining process; 2) establish the database, through the establishment, it can collect, describe, and select the data in advance; 3) analyze the data, it can use the commonly used analysis methods which introduced in this article (such as association analysis technique, sequential mode analysis technique, etc.) to analyze the data and classify the characteristic of the data; 4) prepare the data, which includes choosing the variables, recording and transforming the variables, etc. 5) establish the model, which needs to select suitable mining algorithm to mine on the transformed data; 6) evaluate and explain, which needs to evaluate the established model and has some necessary explanations; 7) implementation (knowledge assimilation), applying the analysis result to the similar system.

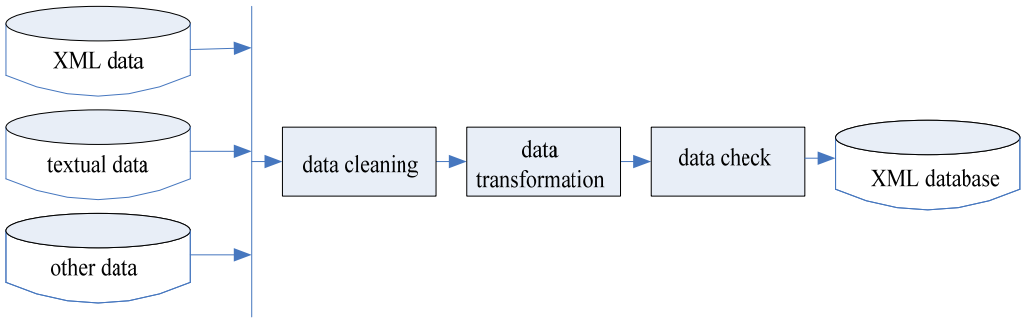

Figure 3 : Data pre-treatment process

APPLICATION OF DATA MINING TECHNOLOGY

Data mining technology aimed at bio-medical and dna data analysis

At present, many bio-medical researches mainly focus on data analysis of BNA sequence, which has found causes of many diseases and disabilities genes. Using association analysis technology to identify the gene sequence which occurs simultaneously; by means of path analysis technique to find the gene sequence of disease at different stage; and it can also use visualization tools and genetic data analysis technique to display the complex structure and sequence of gene.

Data mining technology aimed at finance

Banks and financial institutions all provides various kind of savings business, credit business, investment, financial service business and insurance business. Aimed at the characteristics of the business and data, there are several typical data mining techniques: 1. design and construct the database for the characteristic of multidimensional data; 2. Using loan repayment predicting technique and consumer credit policy analysis method to analyze data; 3. Having classification and cluster analysis on specified target market customers.

```
<? xml version=”1.0” encoding=”GB2312”?>
<Tree>
 <Weather value=”Sunny”>
 <Humidity value =”High”>
 N
 </ Humidity >
 < Humidity value=”Normal”>
 P
 < Humidity >
<Weather >
 <Weather value=”Cloudy”>
 P
<Weather >
<Weather value=”Rain”>
 <Wind value=”Windness”>
 N
 </Wind>
 <Wind value=” Windless”>
 P
 </Wind>
</Weather>
</Tree>
```

Application of data mining technology on retail business

Retail Business is the main field for the research and application of data mining, retail business has the accumulation of long-term sales data, such as the purchase record of customer, consumer, service record and goods purchase record, etc. The data mining of retail business can help sellers to identify the shopping behavior and buying mode of customer, as well as improving the service quality, increasing the goods sales ratio and reducing the cost

The related data mining techniques are: design and construct database according to data mining technology; multidimensional analysis on sales, customers, goods, time and area; analysis of customer's shopping loyalty, etc.

FUTURE TREND OF DATA MINING TECHNOLOGY

Since the variety of data, data mining task and data mining method, which gives many research direction of data mining technology, such as the application development of data mining technology; the data mining method with flexibility; data mining technology, database system and integration of Web database system; standardization of data mining language; visualized data mining tool; new mining technique for complicated data type; privacy protection of data mining process and information security, etc.

CONCLUSION

Data mining technology uses artificial intelligence and XML-related technologies to analyze the data in certain database, in order to dig out its rule, which involves various analysis techniques, commonly used algorithms, and the application field of data mining technology is quiet extensive, the development potential is huge.

ACKNOWLEDGEMENT

Research Project: High school graduate tutor research project in Gansu province - Research and development of compulsory education monitoring and evaluation of engineeringsoftware system in Gansu province (No. 1128-02).

REFERENCES

[1] Li Ze-Wen; Data Mining Technology based on Web[J], Modern Computer (Professional Edition), **1**, 30-31 **(2003)**.
[2] Zhu Ming; Data Mining [M] Hefei: China University of Technology Press, **(2002)**.
[3] Chen Qian-Xi, Wang Yong-Ping; Three Capabilities Analysis Research of Web Information System based on XML [J], Software Tribune, **06**, **(2014)**.
[4] Mao Guo-Jun; Concept of Data Mining [J] Computer Engineering and Design, **23(8)**, 13-17 **(2002)**.
[5] Chen Wen-Wei, et al; Data Mining Technology [M] Beijing: Beijing University Press, **(2002)**.