2009 International Conference on Machine Learning and Computing
IPCSIT vol.3 (2011) © (2011) IACSIT Press, Singapore

College information system research based on data mining

An-yi Lan [1], Jie Li [2]

[1] Hebei North University Hebei, China
[2] Agricultural University of Hebei Hebei, China

Abstract. In this paper, the existing data mining algorithms are improved through the research on them. And through the mentioned data mining algorithm, the student information management module in college information management system was implemented. According to the data of students'course and pre-course selection in course selection management database, using the data mining algorithm presented in this paper, it mines the association rules of the courses, identifies the relationship of students' selected courses and provide basis for planning and curriculum classification.

The experiment shows that the data mining algorithm presented in this paper can not only achieve the function of data mining, the mining speed is also improved for a certain degree, and finally this method is simple and easy to implement.

Keywords: Data mining; association rules; information system

1. INTRODUCTION

In recent years, with the development of database and information technology to be technical support, the rapid popularization of network technology to be development channel, the volume supply of computer hardware, data collection equipment and the storage medium to be material basis, people's data-collection capabilities have been significantly increased, all sectors of society have stored a lot of information such as the types of production, management and research, and the global data storage capacity is being increased dramatically. However, in sharp contrast with this is that the people's understanding of large-scale data capacity has not been effectively increased, and only relying on the traditional data search and statistical analysis methods is such as far from enough to meet the needs of others, resulting in a "data rich, but information poor "situation [3] .

In order to extract mode from the mass data storage, find the rule of the data changes and the relationship between data, fully explore the potential of data to guide decision-making and scientific discovery works, the needs of people for data analysis and even translating it into the knowledge which can be understood easily is more and more imminent. The knowledge discovery technology of data mining met the needs of people, provided a powerful mean to automatically and intelligently translate the mass data into useful information knowledge, which spanned a convenient bridge for the gulf between data and knowledge. As a new technology which can automatically and efficiently extract the valuable information and knowledge from the mass data to effectively support the decision-making, data mining has a very important theoretical and practical significance and wide application prospects. The data mining theory has aroused the interest of researchers in many fields, so data mining is not only the leading issues of information science, but also a theoretical and practical issue combining multi research areas such as machine learning, pattern recognition, statistics, intelligent databases, knowledge acquisition and expert system. At present, data mining technology has been initially used in business, finance, agriculture and other areas,and also has obtained good results [1] .

2. DATA MINING

Data mining is a process to extract the implicit information and knowledge which is potentially useful and people do not know in advance, and this extraction is from the mass, incomplete, noisy, fuzzy and random data [2] .

The essential difference between the data mining and the traditional data analysis (such as query, reporting and on-line application of analysis) is that the data mining is to mine information and discover knowledge on the premise of no clear assumption [4] . The gained information from data mining should has unknown, effective and available characteristics. The previously unknown information is that the information had not been anticipated in advance, that is, data mining is to find the information or knowledge which can not be found relying on intuition, even the information or knowledge which is contrary to intuition, and the more unexpected the mined information is, the more valuable it may be. The validity of the information requests that the data to be mined should be checked carefully, only the validity of the information (or data) to be mined is ensured the validity of the extracted information will be ensured. Most importantly, the information received must be practical, that is, the information or knowledge is valid, practical and achievable for the discussed business or research area.

2.1. Data mining process

Generally, data mining process is composed by data preparation, data mining, information expression and analysis decision-making phases, the specific process as shown in fig.1[5] .

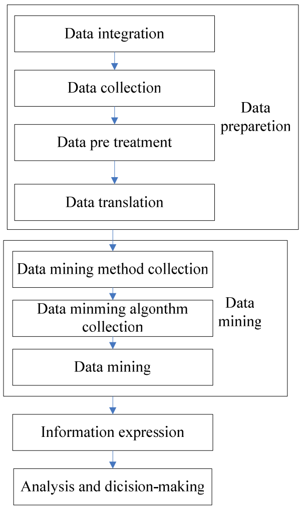

Fig.1:data mining general process

1) Data preparation

Data preparation generally consists of two processes: data collection and data collation. Data collection is the first step of data mining, and the data can come from the existing transaction processing systems, also can be obtained from the data warehouse; data collation is to eliminate noise or inconsistent data, it is the necessary link of data mining. The data obtained from the phase of the data collection may have a certain degree of "pollution", which refers to that in the data may be its own inconsistency, or some missing data, so the collation of the data is essential [9] . At the same time, through data collation the data can be done on a simple generalization processing, thus on the basis of the original data more rich data information will be obtained, which will facilitate the next data mining step.

2) Data mining

Data mining is the core stage of the entire process, it mainly uses the collected mining tools and techniques to deal with the data, thus the rules, patterns and trends will be found.

3) Information expression

Information expression is to use visualization and knowledge information expression technology to provide the mined knowledge information for users, is an important means to show the data mining results. Clear and effective mining result information expression will greatly facilitate the accuracy and efficiency of the decision-making.

4) Analysis and decision-making

The ultimate goal of data mining is to assist the decisionmaking.Decision-makers can analyze the results of data mining and adjust the decision-making strategies combining with the actual situation.

2.2. Data mining classification

There are many classification methods of data mining technology. According to the mining tasks it can be divided into association rule mining, data classification rules mining, clustering rules mining, model-dependent analysis and discovery, as well as the concept description, deviation analysis, trend analysis and pattern analysis; According to the mined database it can be divided into relational database, object-oriented database, spatial database, time-based databases, multimedia databases and heterogeneous databases; According to the technology used it can be divided into artificial neural networks, decision trees, genetic algorithm, the neighborhood principle and vision, and so on [3] .

2.3. Association rule mining

In the knowledge modes data mining discovered, association rule mode is a very important and also the most active branch. Association rule refers to the rules of certain association relationship between groups of objects in the database. It can be used to find the contact among the different commodities (terms) in the transaction database, and so that the behavior patterns of customer purchases will be found. Fining that rules can be applied for merchandise shelf design, inventory arrangements and users classification according to the purchase patterns.

Association rule mining can be described as following: assuming $\mathrm{I}=\left\{i_1, i_2, \cdots, i_n\right\}$ is n aggregates with different terms, then for a transaction database D , each element T inD is a set composed by some terms in I, $T \subseteq I$ .The association rule is expressed as $X \Rightarrow Y$,hereinto, $X \subset I$, $Y \subset I$ and $X \cap Y = \Phi$ .The association rule mining is to discover all condition implicative expression meeting the minimum degree of confidence and support users given, that is, association rules. The confidence and support degree of these rules are all greater than or equal to the minimum degree of confidence and support [7] .

The confidence and support degree of the association rules respectively reflects the correct degree and the support rate of them. Analyzing from semantic perspective, the degree of confidence refers to the trustworthiness of the rules; support degree refers to the possibility of the rule mode appearance, reflects the importance of the antecedent to the consequent.

In general, the user can define two thresholds which is respectively set as minimum support threshold and minimum confidence threshold, the support and confidence degree the data mining system generated are required to be not less than the two given thresholds, then we can say that this rule is valid, otherwise it is null and void. Thus a specific association rules can be uniquely identified by using a implication expression and two thresholds.

3. ASSOCIATION RULE MINING ALGORITHM IMPROVEMENT

After analyzing the characteristics of the association rules in relational database and the two basic ideas of mining association rules in the current relational database, in order to help the "seamless" connectivity between the mining algorithm and the existing relational database management system and improve the mining efficiency, in this paper the technical idea of aggregate operation based on SQL language is adopted, and through the process of generating the search table with smaller data scale than the mining source table and the key technology which uses generating attribute combinations to support the connection query between the aggregation operations and the frequent relation item tables in the search replace, a high-performance association rule mining algorithm in relational database is presented.

The whole algorithm is divided into frequent item set search and the generation of strong association rules two steps. The search of frequent relationship item set is still the core of the whole mining algorithm, and its efficiency has a great influence to the overall efficiency of the algorithm.

Through collecting the support count of all the relationship item sets in the search replace table, the algorithm can adopt the minimum support threshold to select out all the frequent relationship item sets. The generation method of the strong association rules has the same principle with Boolean association rules, which adopted the packet processing methods to improve the overall operation efficiency, and therefore in this paper the search process of the frequent item set will be focused on.

3.1. Algorithm idea

The relation database contains complex multi-valued, multi-dimensional association rules, if analyzed from Boolean-based mining idea, the mining process is bound to be complex and cumbersome than the one-dimensional Boolean rule in affair database; but if analyzed from the view of SQL-based operation technology, the mining algorithm of the association rules in relational database is more easily understood and realized. SQL language only needs to use its nine verbs to meet users' operation request on the database.

One of the main reasons is that its commands contain some simple and efficient operate clauses and functions. And using it the discovery process of the frequent relational item set will become simple and intuitive. Therefore, using all the nonempty sub sets which have associated attributes with the mining task in the relational database to be the grouping attributes of SELECT statement, all the unions of the implementation results are the implied all frequent relational item sets which can meet the minimum support users appointed, the specific algorithm process as shown in Fig.2.

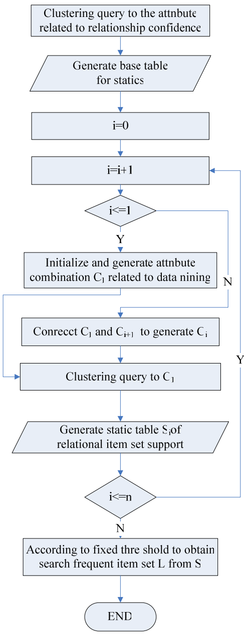
Clustering query to the attnbute related to relationship confidence
Generate base table for statics
i=0
i=i+1
i<=1
Y
N
Initialize and generate attnbute combination C1 related to data nining
Conrecct C1 and Ci+1 to generate Ci
Y
Clustering query to C1
Generate static table Si of relational item set support
i<=n
N
According to fixed thre shold to obtain search frequent item set L from S
END

Fig.2:algorithm process figure

4. DATA MINING ALGORITHM APPLICATION IN COLLEGE INFORMATION MANAGEMENT

Due to space limitations, in this paper only the application of data mining algorithm in student information management module will be introduced. In student information management module, the goal should be achieved is to identify the relationship of students' selected courses and provide basis for planning and curriculum classification according to the data of students' course and pre-course selection in course selection management database.

4.1. Application of Data mining algorithm in college information management

Assuming that macrocosm is the course aggregate the distance education provider can provide, and then each course has a Boolean variable to show the availability of the course. Each elective sequence can be expressed by using a Boolean vector. The Boolean vectors can be analyzed to obtain the elective mode reflected the course frequent association. These modes can be expressed in the form of between the curriculums students selected can be found through mining the association rules of the elective course in student elective management database. Because that only the one-dimensional data of the courses the students selected, in this paper the data mining algorithm introduced above should be adopted to generate association rules from the frequent item sets, the realized process as shown in figure 3.

4.2. Mining test results

1) Experiment circumstance

This experiment uses HP Pavilion g3628cx PC machine with storage of 2G, the database circumstance is Visual FoxPro 6.0, SPSS Clementine 10.0.

2) Experiment data

Assuming the course macrocosm is: {I1=software engineering, I2=data structure, I3=database principle,I4=object-oriented program design, I5=C language,I6=college English, I7=higher mathematics, I8=computer principle, I9=computer network}. From the student course management database and through relationship query the related group set will be achieved, and the specific content as shown in table 1.

TABLE 1. STUDENT COURSE DATA GROUP SET

Group ide ntifier	Data item list
1	14,15,16
2	13,14,19
3	13,18,19
4	17,18
5	12,13,19
6	11,13,14
7	12,13,14
8	11,12

Assuming that the minimum affair support is 2, the output rule is: I3> I4, I9> I3. That is, the students selected the course of "database principle" are also apt to select the course of "object-oriented program design" ; the students selected the course of "computer network" are also apt to select the course of "database principle".

5. CONCLUSION

Data mining is a hot topic of the computer science research in recent years, and it has a extensive applications in various fields. Data mining technology is an applicationoriented technology. It not only is a simple search, query and transfer on the particular database, but also analyzes, integrates and reasons these data to guide the solution of practical problems and find the relation between events, and even to predict future activities through using the existing data [8] . In this paper, through the mentioned data mining algorithm, the student information management module in college information management system was implemented.

According to the data of students' course and pre-course selection in course selection management database, using the data mining algorithm presented in this paper, it mines the association rules of the courses, identifies the relationship of students' selected courses and provide basis for planning and curriculum classification. The experiment shows that the data mining algorithm presented in this paper can not only achieve the function of data mining, the mining speed is also improved for a certain degree, and finally this method is simple and easy to implement.

6. REFERENCES

[1] Ming-Syan Chen, Jiawei Han, Philip S yu. Data Mining: An Overview from a Database Perspective[J]. IEEE Transactions on Knowledge and Data Engineering, l996, 8(6):866-883.

[2] R Agrawal ,T 1 mielinski, A Swami. Database Mining: A Performance Perspective[J]· IEEE Transactions on Knowledge and Data Engineering, 1993,12:914-925.

[3] Vasant Dhat. Data Mining In Finance: Using Counterfactuals to generate knowledge from organizational information

[4] system[J].Information System,1998,23(7):423-437.

[5] Rakesh Agrawal, Sakti Ghosh, Tomasz Imielinski, Bala lyer, Aran Swami. An Interval Classifier for Database Mining Applications[M]. CLDB92. Vancouver, British Coumbia, Vanada, 1992:560～573.

[6] J Han, Y Cai, N Cercone. Knowledge Discovery in Database: An Attribute-Oriented Approach [M]. VLDB-92, Vancouver, British Columbia,Canada,1992:547-559.

[7] J.Hipp, U.Guntzer, G.Nakhaeizadeh. Algorithms for association generation. In Proc ACM-SIGMOD. Dallas, TX, May 2000. 1-12.

[8] Agrawal R, Srikant R. Privacy-Preserving Data Mining. ACM SIGMOD Conference on Management of Data.Dallas,Texas,2000.439-450.

[9] Hand D, Mannila H, Smyth P. Principle of Date Mining, Cambridge, CA. MIT Press, 2001:1-2.

[10] Han J W, Kamber M. Data Mining: Concepts and Techniques. San Francisco: CA. Morgan Kaufmann Publishers. 2001:2-4.