The Applied Research of Data Mining in University Libraries

Yongtao ZHANG
Business School, Sichuan University, Chengdu, China

ABSTRACT: The data mining is the effective tool of information processing. The paper, starting with the basic conception of data mining, probes into the functions of data mining and points that the main applications of data mining technology in university libraries are academic literature recommendation, literature retrieval and intelligence consultation.
KEYWORD: Data Mining; Library; Applied research

In the network age, a large amount of information has brought us great convenience, and meanwhile, the problem of information redundancy, authenticity and security. University libraries are important collecting and distributing centers for literature retrieval and intelligence consultation. The problem to be solved now is how to find out the useful information in massive data and how to improve the information utilization of libraries. The appearance of data mining technology provides us effective way and method to solve these problems.

1 BASIC CONCEPTION, CONTENT AND ESSENCE OF DATA MINING

1.1 *Basic Conception of Data Mining*

Data mining is an interdisciplinary subfield of computer science, which involves the following disciplines: computer technology (including database technology and machine learning), information technology, statistics, data visualization and pattern recognition. Data mining is defined as the hidden information of database. And it is also called exploratory data analysis, data-driven found or inductive learning which means analyzing on the massive, incomplete, noisy, obscure and random real applied data from database. It aims to find out the unknown relations and patterns of data, and recombines the hidden potential data by new methods and techniques into useful information and knowledge.

1.2 *Content and Essence of Data Mining*

The content of data mining has a wide range, including basic theory, data warehouse, visualization technology, qualitative and quantitative mutual model, knowledge presentation method, discovered knowledge maintenance and reuse, discovered knowledge in semi-structured data and unstructured data and Internet data mining. Mining data is a process of deduction by finding a pattern through the data. The essence is to search the valuable data from the massive data and to find out the relevant information.

2 MAIN FUNCTIONS OF DATA MINING

The main purpose of data mining is to find out the hidden valuable relation in the unknown data. The two main discovery modes are predictive mode and descriptive mode. The predictive mode is to predict the data values for some purposes by utilizing the known results from different data, including classification, regression, time series analysis, and prediction. However, the descriptive mode aims at identifying the models and relations of data and providing a method to analyze the data character, rather than predicting the new character. It includes clustering, collection, association rule and sequence found. In the data mining process, these functions interconnect and interact, rather than work independently.

2.1 *Classification*

Classification: Classification is to differentiate the data type. The main function of classification is to learn a classified function or classified pattern (also called classifier), which can distribute data to different group according to data's characters, namely, analysis data's different characters and finding data's character pattern to determine which group they belong to. Therefore, it can be used to analyze the known data and predict which group the new data belong to. Before predicting the data, the experts should ensure the classification and create a classified function and then make the featured data reflect the defined classification.

2.2 *Regression*

In classification, the predicted variable is prototypical category which dispersing forecasting value, such as class label. In regression, the predicted variable is quantitative mode, which adopts the continuous predicted value. Therefore, in data mining technology, it is widely believed that classification adopts predicted class label, while prediction adopts continuous predicted value. Regression, firstly, has to assume that some known types of functions should match data, and then make use of error analysis to determine the best matching function with the target data. So it can be seen that some nonlinear problems can be solved by switching to linear problems via variables.

2.3 *Time Series Analysis*

The main function is to use time series plots to visualize the time series data. The basic three functions involve: using the distance measure to determine the similarity of different time series; testing line structure of time series plots to determine the behavior of time series; employing historical time series plots to predict the data's future numerical value.

2.4 *Prediction*

Prediction is to classify or assess the mode, which can be used to predict the variable. The purpose of prediction is to predict the future unknown variable and this kind of prediction needs time to test and verify. That is to say, after a certain period of time, we could know the accuracy of the prediction. As one of the functions of data mining, rather than model prediction, the main difference is to predict the future data status not the current status. Its applications are flood forecast, speech recognition and machine learning.

2.5 *Clustering*

When lack of descriptive information in data analysis, or when some modes cannot be classified, you could use the clustering analysis. Clustering is to classify one item or data into different classes according to some similar measuring methods. By determining the certain character's similarity, we can finish clustering task. The most similar data are clustered into a cluster, which is not defined in advance but is explained by experts in this field. The differences between clustering and classification is that the clustering does not depend on the defined class and does not need training set with the purpose of shortening the distance of items of the same class and make the items of different classes much longer.

2.6 *Association Rules*

The data in database have some relevant relations. There is some regularity between dereferencing of two or more variables. The main relations are simple relation and time sequence relation. All of these two relations explain the relevance of the same thing or different items in the process and they are recognizable patterns. The main target of association rules is object database. And its main purpose is to find out the hidden relevant net in database and describe closeness or relation of data in one item, so as to make sure what things could happen. The most famous mining association rule, putting forward by Agrawal in 1994, is called Apriori Algorithm. The main idea of Apriori Algorithm is to count the frequency of many items in a shopping and then switch it into association rules. The core is to use one scanning result to get the candidate sets of item of this scanning, so as to improve searching efficiency.

2.7 *Sequence found*

Sequence found is used to determine the sequence mode between data and time. Sequence found is similar to association analysis, but it pays more attention to analyze the sequence relations. There are three factors that affect the results: the lasting time of sequence, the foldout window of time and the interval of found mode.

3 APPLICATION OF DATA MINING TECHNOLOGY IN UNIVERSITY LIBRARIES

3.1 *Application of Recommendation of books and references*

Recommendation of books and references are the major components of personalized services in university libraries. Through the association analysis of readers' borrowed record, we could find the association rules between the readers' borrowing behaviors and books and references. Calculating the associa-

tion rules' support degree and credit degree could create the borrowing mode, by which we could recommend the available books and references to the readers. For example, by association rules, we find that the reader A often borrows the books on computer visual programming, and meanwhile he also borrows the books on database. Therefore, we could give some advice about the books when reader A borrows books. This kind of technology could save the time and improve the service quality and efficiency.

3.2 *Optimization of Shelves Management*

The shelves of university libraries are the intensive place of books and references, including various professional fields. Some may offered for teaching and research activities for all teachers and students. Shelves management is much more important in library management. The author believes that the most important part in shelves management is to predict the books and references changing trend and reserve the shelves. It is data mining that could tackle the above problems. Firstly, employing the sequence analysis and regression to analyze all borrowed records and find out the periodical changes of these books. Secondly, doing the classification analysis on the books on circulation in order to find out the catalogs of high-frequency books and highly increasing books. Thirdly, updating the books. Classification analysis should be used on data and do the statistics on the numbers and positions of the books, which are old, or with many copies, or damaged. In the end, making a list of all these books and feedback to the relevant departments.

3.3 *Application on Information Retrieval*

Information retrieval is an important part in personalized services in university libraries. The traditional information retrieval can only offer simple information for readers, rather than personalized services. It can be easily achieved by data mining. Firstly, preparing the data. It is the basic step in data mining. Without data, data mining technology is meaningless. Preparing means collecting the borrowing information including borrow records, Internet notes, reservation information and renewal information. Secondly, filtering and process the data. We have to remove the noise data and repeating data to preprocess and transfer the data. Based on the filtering, we have to build the dynamic structure database to do the data mining. Thirdly, doing the data mining. Doing the association and classification analysis on data mining database and find out the collection of books and hobbies of different readers. At last, combining the data mining and visualization to provide personalized retrieval service. When doing the retrieval by data mining, the retrieval results can automatically provide the relevant books to readers and automatically extract the valuable information.

3.4 *Application on Intelligence Consultation*

The timely knowledge and information providing service is a standard measure of intelligence consultation. In modern time, all the university libraries have great relation with Internet, which provides a great convenience for intelligence agents to get information. It is hard to get the right information from the Internet. However, Web mining can tackled the problem. Firstly, using the web mining technology to mine the web on the Internet on mining server according to the needs of teaching, research and development of the university. Web mining can do the deep analysis on the data according to the requirement and also guarantee the completeness and security of data. Secondly, employing clustering and classification to analyze the retrieval results and classify the relevant information by discipline construction and research direction. At last, combined with visualization and artificial intellegence to build the retrieval interfaces. Therefore, users can do the retrieval via keywords, subjects or other information by proxy servers.

4 SUMMARY

Nowadays, data mining technology in university libraries stands in an initial stage, for it shows great potential in processing, analysis and organizing. It is believed that in the near future, with the development of database expansion and network technique, data mining can yield unusually brilliant influences. Furthermore, it can also have great impact on the transition process from traditional libraries to digital libraries.

REFERENCES

[1] Zhe Zhang & Guiran Chang. 2003. The Application of Data Mining Technology in CRM. Chinese Journal of Management Science: 11(1), 53-59.

[2] Jiejun Huang & Heping Pan. 2003. Application Research on the Technology of Data Mining. Computer Engineering and Applications: 2, 45-48.

[3] Weimin Zheng & Gang Huang. 1999. The Tool and Choice of Data Mining Technology. Beijing: Tsinghua University Press.

[4] Wenke Liu. 2006. Application of Data Mining on University Library Reader Management. Sci-Tech Information Development & Economy: 16(8), 67-68.

[5] Yan Wang. 2003. The Application of Data Mining in Digital Library. Information Science: 21(2), 211-214.