# Application of Data Mining Technology in Digital Library

Mei Zhang
Library, Linyi University, Linyi, Shandong, China
Email: zhangmei7596@163.com

***Abstract*—With the rapid development of computer technology and network technology, it ushers in a new Internet era characterized by information and knowledge. There is an urgent need for a new generation of technologies and methods to help exploit the treasures of information, and to be refined, so as to become useful knowledge. Data mining is the core part of knowledge discovery. Its techniques and methods have great application space and value in the digital library. Data mining technology can help people to develop vast amounts of information in depth, extract the inherent link of the heterogeneous information to promote the digital library. This paper describes technologies relating with data mining, introduces the process of data mining, illustrates the main features, explores applications of three aspects in the digital library, indicates application meaning, analyzes the key problems of implementation in digital libraries, and finally take a study case on analysis of low utilization readers in Linyi university library by decision tree algorithm.**

***Index Terms*—data mining, digital library, knowledge discovery, information service**

## I. INTRODUCTION

In recent years, as the rapid development of information technology, communication technology and computer technology, digital has become the main direction of library. However, the digital library with a wealth of information easily comes into the situation of rich data but poor information. Therefore, libraries need to strengthen the capacity of information processing and resource organization. Data mining technology can help people to develop vast amounts of information in depth, extract the inherent link of the heterogeneous information to promote the digital library.

## II. DATA MINING AND RELATED TECHNOLOGIES

### A. Baisc concepts of data mining

Data mining is a new information technology as the development of database technology and artificial intelligence technology. Data mining is the process of extracting the hidden information and knowledge that people do not know in advance but potentially useful[1]. The aim is to discover unknown relationships and summarize data in the innovative way of understanding by the data owner and value, and predict possible future behavior, so as to provide stronger support for decision making. According to the different forms of the main data structure, data mining can be divided into three categories in general, that is data mining, Web data mining and text data mining.

*Data mining.* The data mining is for structured data, such as SQL, Server, Qracle, Informix and other data or data warehouse. At present, the following software DB2 Intelligent Miner for Data SAS Enterprise Miner of IBM can be used.

*Web mining.* The object of data mining is a traditional database or data warehousing, and Web data mining is the various Web data including Web pages, structure between pages, the user access to information, and business transaction information, which is to discover useful knowledge to help people extract knowledge from the World Wide Web, improving site design and e-commerce to better or improve services. Web data mining can be divided into Web content mining, Web usage mining and Web structure mining.

*Text data mining.* When the objects of data mining is composed entirely of text type, the process of automated information processing and analysis massive text information is called text data mining, which is in the way of data mining algorithms and information retrieval algorithms[2]. It includes feature extraction, text summarization, text classification and clustering, concept operations and exploratory data analysis. The techniques of text data mining contains the vector of word frequency, the word string representation, Bayesian classifier, Bag of word, text clustering algorithm based on the concept and classification algorithm of K - the nearest neighbor.

Data mining technology and its application is a hot topic in the international arena currently, and it has been applied in many industries, which demonstrates its advantages and development potential. In the area of information management, it is the only way to achieve development of knowledge retrieval and knowledge management, as long as we integrate data mining technology with artificial intelligence technology, access to user knowledge, literature and other kinds of knowledge.

Data mining is to discover and extract hidden information from large databases and vast network information space in the digital library, and the purpose is to help information workers search for the potential association between the data and find the neglected


doi:10.4304/jcp.6.4.761-768

elements, which is very useful to predict trends and make decision.

### B. Principal means of data mining

*Inductive learning method.* Inductive learning methods include information methods or decision tree and set theory methods. Decision tree uses attribute structure to represent the decision-making set, and these decision set products rules through classification of data sets. Typical methods of decision tree have classification and regression tree and mining used in classification rules. Decision tree classification method is faster and more easily converted into simple understandable classification rules, etc., especially in problem areas of high dimension, it can get a very good classification results[3]. Set theory method is a new mathematical tool dealing with vague and uncertain problems. It has some advantages of a strong mathematical foundation, simple method, highly targeted and small computation. Using this method it can handle the following problems, including data simplification, data association discovery, data meaning evaluation and approximate analysis of data.

*Imitation biotechnology method.* Imitation biotechnology method includes neural networks and genetic algorithms. Artificial neural networks imitating biological neural network in structure, is a nonlinear prediction model through training, and can be used for classification, clustering, feature extraction and other operations. Genetic algorithm is an optimization technique, it uses a series of concepts of biological evolution to search issues, and ultimately to optimize. It calculates the fitness of individuals after genetic, then carries out chromosome replication, exchange and mutation, etc., produces new individuals, repeats this operation until obtains the best or better individual. In data mining, it is often expressed as a search problem, and uses powerful search capabilities of genetic algorithms to find the optimal solution.

*Formula discovery.* Formula discovery includes discovery system BACON of physical laws and discovery system FDD of empirical formula. It identifies the new record through combination of K similar historical records, and this technique can be used as data mining tasks of clustering and deviation analysis.

*Statistical analysis.* Statistical analysis is to extract unknown mathematical model from samples analysis. In data mining, it often involves a certain statistical procedures, judgment hypothesis and errors control.

*Fuzzy mathematics.* Fuzzy logic is fusion of fuzzy set and Boolean Logic. The true value of a formula can be any value in [0,1]. In data mining and knowledge discovery, it often uses fuzzy logic for data query, sort, and the evidence combination and confidence calculation.

*Visualization.* Visualization shows association trends of information mode and the data in the intuitive graphical way, and decision makers interactively analyze data.

In general, there is not a universally applicable method of data mining. A method or algorithm is very effective in a particular field, but may not be suitable in another area. Therefore, we need carefully choose effective data mining models and algorithm against specific areas in practical application.

## III. Steps of Data Mining

Data mining is a completed process which mines unknown, effective and practical information from large databases. Data mining process includes four steps, identification area objects, data preparation, mining process and results expression and analysis, which is shown in Fig. 1. The above four steps is not a linear, and in the actual operation only continuously repeat we can obtain good results. The explanation of the four steps is as follows.

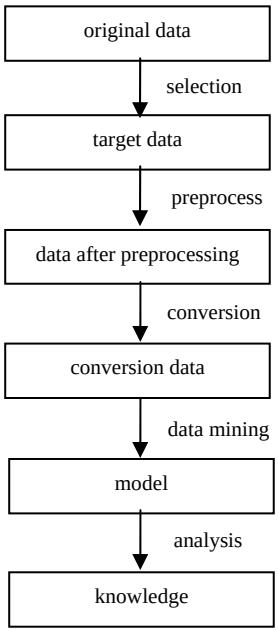


Figure 1. Data mining process.

### A. Determine area object

Firstly, we must understand a variety of knowledge of application areas and analyzes the application target, to determine the purpose and requirements of data mining. Only understand and familiar with the background knowledge in the field, understand the needs of users, we can clearly define the problem to be solved, and to prepare high-quality data for the mining, which can correctly analyze mining results, and provide useful information for mining areas[4]. Therefore, it is an important step to determine the business object and clearly demand for mining purposes.

### B. Data preparation

This phase includes two aspects. On the one hand, the required data will be integrated from multiple data sources; on the other hand, the required indicators of the existing data will be determined according to the experience of miner and easy use of mining tools. This phase can be further divided into the following three sub-steps:

*Data selection.* In this step, it only simply removes some of the redundant or irrelevant data, and select out all data information relating with application areas and suitable for mining application from internal and external data sources.

*Data preprocessing*. It further analyzes the selected data to improve the quality of the data. In particular, the incomplete and inconsistent data containing noise must be preprocessed to improve the quality of the knowledge gained from data mining.

*Data conversion*. The data will be normalized, even if the data converts to applicable form.

### C. Data Mining

In the field of data mining, no method or tool is a panacea, which is applicable to all data. Under normal circumstances, it needs to build different models, parameter or algorithm, to choose the most appropriate one by comparison.

### D. Results expression and analysis

According to mining object, it extracts the most valuable information, expresses in the understandable way, analyzes and evaluates the results. In this step, it is not only to express the results, but also filters the information, to obtain the information consistent with the purpose of mining[5]. If the obtained knowledge does not meet the requirements, it needs to repeat the above steps. In addition, we should optimize specific stages of knowledge discovery according to the actual implementation, until meet user requirements.

It is very challenging how to get information through effectively applying data mining algorithms. Algorithm research and testing is often used in standard data sets. However, in reality, data sources of users are often various, which have a lot of missing data, noise data and non-standardized data. So in the data mining process, it is very important for data preprocessing and transformation, and it is an important guarantee for data mining to improve the efficiency and get better result.

## IV. The Function of Data Mining

In general, the function and the target data type of data mining is relevant. Some features can only be used in a particular data type, while some functions can be applied to many different types of databases. For the determination of data mining, we should consider integratively the function of data mining, data type and user interest. It can be divided into the following categories according to their functions.

### A. Predicting trends and behavior automatically

Predicting trends and behavior automatically refers that data mining can classify and find predictive information from the past and current data in the large databases, automatically present models describing the important data, obtain results forecasting rapidly and directly from the data itself, and predict future data state.

### B. Association analysis

Association analysis is to find interesting association, relationship or causal structure between the sets from a large number of items and item set patterns. Data association is a class of important knowledge that can be found in the database. It is called association if the values of two or more variables exist some regularity. Association can be divided into simple association, temporal association and causal association. Association analysis is to identify the links among data set properties, and form association rules[6]. Association rules have two parameters, support level and confidence coefficient. Support level shows the established proportion the rule in all instances, that is representative. Confidence coefficient shows the established proportion of the consequent case when the antecedent established, that is the credibility of rules.

### C. Clustering

Cluster analysis is a metnod when do not know the pre-designated classes, we gather the information according to the principle of information similarity. The purpose of clustering is to divide a data set into different classes according to a particular standard, so in the same class there is a high similarity between objects, but in different classes they vary greatly[7]. Clustering enhances people's understanding of objective reality, and it is a prerequisite of concept description and deviation analysis. Clustering techniques mainly include the traditional method of pattern recognition and mathematical classification.

### D. Concept description

Concept description is to describe and summarize the relevant characteristics of the content of such objects through summarizing, analysizing and comparing certain types of associated data. Concept description is divided into characteristic description and discrimination description. The former describes the common features of certain objects, while the latter describes the differences between different types of objects.

### E. Time-series pattern

Time-series pattern is to search out the higher probability of repeated mode through time series. As regression, it uses the known data to forecast future values, but the differences between the data are the variables in different time. In time-series pattern, it need to find the rules whose appearance ratio in a certain minimum time is always higher than a certain minimum percentage, that is minimum support threshold. These rules will be appropriately adjusted as the situation changes[8]. In time-series pattern, an important and influential method is similar sequence. Using the method of similar sequence, it will view the time event database according to time order, to find out one or more similar events.

### F. Deviation detection

Deviation detection is to detect and analyze the deviation data in the database. The data in the database

often have a number of anomalious records, so it is meaningful to detect these deviations from the database. Deviation includes much potential knowledge, such as the abnormal instance of the classification, the special case that does not meet the rules, deviation between observation results and model predictions, and value changes over time.

## V. Applications of Data Mining in Digital Library

### A. Digital library

Digital library is a digital information resource system supported by many high technologies. It will interconnect the digital information resource spreading over the different carriers, different regions in the way of network, provide access and achieve resource sharing. Digital library is the knowledge set of computer handle and orderly organization. It uses digital technology to organize and manage information resources, can store vast amounts of information, the user can query through the network efficiently and easily retrieve information to access to information services, and its information storage and user access do not restrict by time and region. It is not simply the library home page on the Internet, but rather a set of object-oriented, distributed, platform-independent set of digital resources[9]. Digital libraries can break through the limitations of the literature unit and mine and discover information on the basis of knowledge unit, so as to find the law.

### B. Applications of data mining in digital library

With more and more online digital libraries, data mining and knowledge discovery of digital library has great application value. According to the different processing objects, data mining of digital library can be divided into three categories, structure mining, content mining and user log mining. The main elements of data mining in the digital library are shown in Fig. 2.

*Mining of digital library structure.* Web page of digital library is to use hypertext markup language to compile and use hypertext links to establish an information organization. Links are universal phenomenon of pages, and it can exchange information and expand the use value only through linking with other pages and its own content. Structure mining based on digital library is to discover knowledge from the organizational structure and links of pages, mine hyperlinks of digital library pages, documents internal structure and the directory path structure of document URL, and derive knowledge from organizational structure and links. Digital library can provide useful information in addition to document content because of the interconnection between documents. It aims to discover the structure and mode of the page, and reveal the useful model containing in these document structure, and classify and cluster, or analyze the relevant pages, so as to evaluate the quality of web pages, optimize retrieval methods, guide the site construction and master the subject development[10]. For example, professor RayLarson, school of Information Management and Systems of California University in Berkeley, collected the data on earth scientific literature by the search engine AltaVista, and analyzed the relationship of the earth sciences, GIS, satellite remote sensing disciplines and their development.

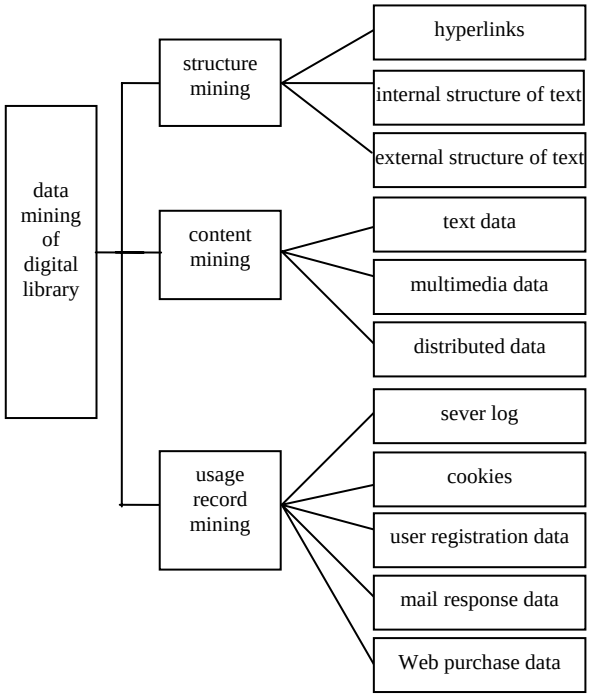


Figure 2. Main elements of data mining in digital library.

*Mining of digital library content.* Content mining refers to extract knowledge from the information of Web document. Content mining is divided into text documents, including text, html, pdf and other formats, multimedia files, including image, sound, audio, video and other media types, and distributed data mining. The content mining based on digital library is to discover meaningful knowledge through pattern recognition and analysis of the information in the digital library. Content mining includes the following five aspects:

(1) Organize the literature data

Using machine learning techniques, it studies the data to form hierarchical classification structure, and do not need pre-defined subject categories, but rather divide the document into several clusters, which requires the similarity as large as possible within a cluster of document, and the similarity as small as possible among different clusters, and divide the search results of search engine into several clusters using text clustering technology. Users only need those related cluster, which greatly reduces the number of results.

(2) Automatic extraction and description of feature

There is a huge amount of digital information in the digital library, at the same time, it will provide search services for network users, so digital libraries must make use of more advanced technology to describe data reasonably and effectively. In order to describe the data, it will discribe the digital information in accordance with certain language[11]. Feature selection is used to identify individual characteristics of the storing objects, and it can

retrieve effectively for the objectives of selected features. Therefore feature selection is conducive to efficient indexing and retrieval. Automatic segmentation of characteristics is to implement the appropriate analysis according to the level of understanding through a content analyzer, from which it extracts the relative content feature to label and organize. Users make this as the basis of search, complete matches of information stored in the database to realize the direct location and find information.

(3) Text summary or abstract

Text summary is to extract key information from the document, summarize and explain the document in the concise way. So the users can understand the overall content of the document or document collection, but need not browse the full story. Text summary usually gives a summary of the document when the search engine returns query results to the user.

(4) Automatic classification of documents

Document classification is a process to make a large number of documents into one or more categories according to the contents or property of documents. The key point is to construct a classification model, and map the unknown document to a given type of space. Construction of classification commonly uses machine learning methods and neural network methods.

(5) Collecting and organizing thematic information automatically

It mines knowledge which can reflect the law available to users from the large amount of raw data. According to a given area of information needs, it automatically captures, collects and collates the required information, and then filters the information sources. After determining the information sources, according to the model algorithm, it calculates and determines the search path, automatically gives priority to the best search path, organizes search keywords according to the logic, which can correspond to more specific areas of information capture[12]. Its main function is to filter redundant information, intelligent concept extraction and generate information summary.

*User mining of digital library.* Usage mining of digital library is using data mining technology to analyze log files after access to the digital library, mine access pattern, and provide decision support for Web site management, structural adjustment and personalized service. Usage mining mainly refers to server log, Cookie, the user registration data, e-mail query response data and Web purchase data. Currently, usage mining can be divided into two categories:

(1) Access pattern tracking

General access mode is to understand the user's access patterns and trends through analyzing usage records to improve the organizational structure of the site. From the large amount of access information, it can mine user access patterns, predict the user's access interest from document hyperlink, find different user groups by means of association rule and clustering method, and provide information customization services for different groups to help members search and process knowledge.

(2) Tracking records of individual use

Track of personal usage records tends to analyze individual user preferences, and its purpose is to provide the corresponding customized services according to access patterns of different users. Through mining access information and usage information, it matches between the digital object and user, the object classification and subject[13]. It uses different mining techniques, such as clustering based on the business and association rules, to automatically extract knowledge, determine personalized services and improve the automation level of services for users. Professor Michael Cooper, had mined and analyzed record data of catalog use of the digital library in California University, and found several types of users, users of real use the directory, the network robbert only collect data for retrieval services, general visitors just visit the website but not find the directory, their duration of stay is different. Cooper also designed a model, for user's time and the process, it adopted cluster analysis and time series analysis, and found that the query number, time, results number, result number of different users had different characteristics[14]. Through data analysis, it understood and mastered the characteristics of digital library users, predicted the future trend to study the behavior law of digital library users.

### *C. The application meaning of data mining*

As the future development trend of library, digital library encounters all kinds of difficulties in the research process of effective restructuring and discovering knowledge. In academia, it is generally considered that data mining can provide key technology for digital library construction, in view that data mining has enormous potential in data organization, analysis, data mining and knowledge discovery.

*Optimize resource construction.* It mines library borrowing, circulation condition, retrieval request and collection bibliographic database, counts literature refusal collection and frequent borrowing collection according to classification, to provide targeted decision-making support for supplement and enrich information resource. And it can analyze utilization ratio of the document, promptly obsolete outdated information, or reduce the interview of some literature information. It carries out association analysis of borrowing literature, discovers association rules and proportional relation between various types of literature, to optimize information construction and collection layout.

It collects and organizes the data of the online message, converts into a structured database, and discovers user interest of using data mining. In particular, using appropriate mining algorithm, academic leaders and experts can find gaps of information resources timely and adjust the direction of collection, make good collection and order of literature information[15]. Using a variety of data mining techniques and methods, we judge the utilization and efficiency of digital library information resources, and guide collection construction of library. We also can develop characteristic collection and construct in-depth special resources in accordance with library resources and personnel structure.

*Provide personalized service.* Digital library is a data information system, which not only has rich content and a variety of digital information resources, but also depends on the support of modern high-tech, and efficiently meets the needs of users. Currently, Digital library includes a large number of digital collections, a wide variety of databases, Web resources links of full text and a wealth of information on the Internet. It is the user actually needs only identify the truly valuable knowledge and information behind the data through organization, analysis and data mining[16]. We can provide more optimized information services to meet the personalized needs of users if data mining is used in the whole process of information discover and provision. Personalized service is a key part in digital library, so we should turn passive service to active service and shift information presentation to generation. It is mainly divided into two levels, the first level is to customize information based on user request, and the second level is that the digital library mines user interest model and initiatively provides services to make digital library be an intelligent and proactive information provider.

*Improve information access speed.* The digital library has a huge amount of information, from which there are much useful knowledge to be extracted. Users are more concerned about their own needs than the total amount of information[17]. We must have a good search mechanism to provide users with faster and more efficient services. Data mining technology provides advanced information retrieval tools for digital libraries. The design system will have greater intelligence if we use the related theory and method of data mining in retrieval.

*Expand service form.* Data mining can realize improvement of service quality and business expansion. Digital library uses modern information technology, which not only refers to changes of media and services, but to improve service structure and levels by means of data mining technology, such as information retrieval services, selective dissemination of information service, novelty search and information analysis services.

### D. The key problems of data mining application

*Computer network support.* The data mining of information resource construction in the digital library need to be based on good computer network. Data transfer and storage is convenient, constructs large-scale data warehousing, which provides effective protection for data mining to carry out.

*Using a variety of data mining.* Single data mining methods have been unable to meet the needs of optimum retrieval. It will greatly improve the efficiency of data mining if using a variety of mining methods and different types of methods as far as possible.

*Deeply mining.* It will mine more potential knowledge if it mines deeply from mass data information. Mathematical statistics in the regression analysis, association analysis and artificial intelligence can solve some of depth mining tasks.

*Coordination of information management personnel.* Data mining in digital libraries needs collaboration of information management personnel. It is not enough only rely on the computer. Information managers need to have a higher quality, master data mining technology, be able to carry out information processing proficiently by means of a variety of data mining tools[18]. Information managers must track data mining methods to improve the original tool and update the data warehouse.

*Centering on readers and timely processing feedback information.* Libraries and readers should keep an interactive relationship. Information management staff inquire readers, and timely establish the reader files and record their behavior. According to the needs of readers, it improves the existing mining technology to make data mining targeted.

## VI. A Study Case on Analysis of Low Utilization Readers

### A. Background

After registering, the readers very little or even never go to the library, which is the major factor of inefficient resource utilization. For Linyi University Library, the annual number of new readers is about 8,000, of which can really make full use of library resources is less than 30%. It is put a lot of manpower, material and financial resources for library construction, but almost nobody is interested in a lot of resources. For some of the reading room, there are less than 30 readers daily, which not only waste money, manpower and space, but also show insufficiency of the university research environment.

### B. Purpose analysis

According to the nature and behavior of readers of high borrowing rate and low borrowing rate, we mine and analyze, and establish the predication model of low utilization readers, to analyze which readers have larger inertia probability, their borrowing behavior, and the relevant factors resulting in the inert. It can provide decision-making basis for library to manage, develop appropriate strategies and attract readers, and forecast reader use in the strategy.

### C. Mining process

In this analysis, we synthetically use steps of CRISP-DM and SEMMA. The detailed description of the whole data mining process is as follows:

*Data preparation.* Firstly, we extract record data table of low borrowing rate from the existing database. We select the readers whose total borrowing number is 0 as the object to study, determine the reader identification number, name, sex, unit and occupation as data items related to the theme of the model, extract data, generate derived variables, and constantly repeat until get all the satisfied data variables. Finally, we set these derived variables in a collection document, including various types of information for each reader, and store into the data mart of lazy reader.

*Sampling.* We extract samples from the collection document of derived variables, including the number of different borrowing, and many samples of different size, and make it as a training model and test model. The

extracted samples are directly sent to the data mining servers.

*Building model.* We select decision tree technology, and train and build models by means of training set.

*Validating the model.* Validation data is a new data for the established model, but the data set needs to be handled through constant inspection, until it has a general accuracy with the established model.

*Model score and monitoring.* The model built by data mining is outputted as a C language subroutine, returns data mining host, and is called by main program of the model score.

### *D. Description of the data source*

In the information table of lazy reader, it defines lazy readers whose borrowing times is 0, and derive the needed data. The table contains the following field, reader identification number, name, grade, units, profession, total number of borrowing, reader registration date, cancellation data, reader type, borrowing rules, average borrowing number, the total sample number ` and so on.

The table of borrowing history, borrowing rules, reader types contains the field of bar code, classification number and so on.

### *E. Creating a model of low utilization readers using the decision tree method*

The reader information tables of which borrowing number is 0, contains 3358 records. We use C5.0 algorithm to construct prediction decision tree or rule set, select units as the target field and grade input field.

In this model, the grade field is divided into five node set, and in each set we get the most unit of zero borrowing. We can find that inert readers are more in junior college students, art department and P. E. department.

### *F. Result analysis*

It will generate multiple inert groups from analysis of inert model and output of decision tree. We should examine and diagnose combination with other condition of readers whether the implied meaning of each group is reasonable or not.

The model grasps limited reader information, so it severely limits the depth of research. However, as the inert reader group of the various departments, it can combine with other dimensions to analyze and predict the possible inert group in other readers.

## VII. CONCLUSION

Data Mining is a cross-discipline, and it is one of the most cutting-edge researches in the international database and information system, so its application in the field of digital libraries is still in its infancy. We should strengthen the learning and applications of data mining. Firstly, we should update the idea and establish effective mechanisms of technology innovation and application. Library management should open mind and change ideas, encourage staff to try and innovate service technologies. Mining technology innovation is the basic premise of information survival and development. It must establish a more perfect technology innovation system to survive the fierce competition and gain competitive advantages. Secondly, we should emphasize learning and promotion of data mining. We can organize young backbone personnel to learn the basic courses of the technology, take the opportunity, improve technological levels, and promote universal basic theory of data mining techniques. Thirdly, we should strength inter-museum exchanges. Because this is a new emerging technology, it is impossible to work behind closed door. Only exchange is it can lay foundation for unification and standardization of expression language. As technology advances and requirement change, the application of data mining in the library management will be paid more attention. It will promote faster development of librarianship and create good social benefits.

## REFERENCES

[1] Siqing Zhou, and Feng Ouyang, "Analysis of the Feasibility of Data Mining Application in Academic Digital Library," *Library Work in Colleges and Universities*, vol. 27, pp. 36-38, October 2007.

[2] Jiawei Han, and Micheline Kamber, *Concepts and Techniques of Data Mining*, China Machine Press, Beijing, 2007.

[3] Jun Zhou, "Design for Personalized Service in the Digital Library Based on Data Mining," *Researches in Library Science*, no. 3, pp. 15-17, March 2007.

[4] Tian Liang, "Application of Data Mining in Information Services of Digital Library," *Academic Library and Information Service*, vol. 8, pp. 31-33, June 2009.

[5] Mo Li, "Application of web-based data mining technology in digital libraries," *Journal of Academic Library and Information Science*, vol. 25, pp. 44-46, August 2007.

[6] William A Maniatty, and Mohammed J. Zaki, "System Support for Scalable Data Mining," ACM *SIGKDD Exploration Newsletter*, vol. 2, pp. 56-65, October 2000.

[7] Myra Spilopoulou, and Carsten Pohle "Data Mining for Measuring and Improving the Success of Web Site," *Data Mining and Knowledge Discovery*, vol. 5, pp. 85-114, April, 2001.

[8] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*, China Machine Press, Beijing, 2004.

[9] Ramesh C. Agarwal, Charu C. Aggarwal, and V. V. V. Prasad, "A Tree Projection Algorithm for Generation of Frequent Item Sets," *Journal of Parallel and Distributed Computing*, vol. 61, pp. 350-371, March 2001.

[10] Bohanec M, Moyle S, and Wettschereck D, "A Software Architecture for Data Pre-processing Using Data Mining and Decision Support Models," *Workshop Integration Aspects of Data Mining, Decision Support and Meta Learning*, Freiburg, Germany, pp. 13-24, September 2001.

[11] Le Yang, Sangmun Shin, and Yongsun Choi, "A Surrogate Variable-Based Data Mining Method Using CFS and RSM," *Proceedings of the 6th WSEAS International Conference on Applied Computer Science*, Hangzhou, China, vol. 6, pp. 648-653, April 2007.

[12] Colin Shearer, "The CRISP-DM model: The New Blueprint for Data Mining," *Journal of Data Warehousing*, vol. 5, pp. 13-22, Fall 2000.

[13] George Gigli, Éloi Bossé, and George A. Lampropoulos, "An Optimized Architecture for Classification Combining

Data Fusion and Data-mining," *Information Fusion*, vol. 8, pp. 366-378, October 2007.
[14] Jochen Hollmann, Anders Ardö, and Per Stenström, "Effectiveness of Caching in a Distributed Digital Library System," *Journal of Systems Architecture*, vol. 53, pp. 403-416, July 2007.
[15] Fengrong Gao, Chunxiao Xing, Xiaoyong Du, and Shan Wang, "Personalized Service System Based on Hybrid Filtering for Digital Library," *Tsinghua Science & Technology*, vol. 12, pp. 1-8, February 2007.
[16] Olivia.P.R, and Yangyong Zhu, *Data Mining Cookbook: Modeling Data for Marketing, Risk and Customer Relationship Management*, Wiley, Hoboken, 2000.
[17] Ligang Ren, Mei Song, and Junde Song, "A Novel Data Type for the Protocol of Data Synchronization," *Proceedings of the 8th International Conference on Computer Supported Cooperative Work in Design*, Xiamen, China, vol. 1, pp. 532-535, May 2004.
[18] Tung, A. K. H., Hou, J., and Jiawei Han, "Spatial Clustering in the Presence of Obstacles," *Proceedings of the 17th ICDE*, Heidelberg, Germany, vol. 3, pp. 359-367, April 2001.

**Mei Zhang** received her B.S. degree in chemistry education from Yangzhou University in 2001, the M.A. degree from Qufu Normal University in 2009. She is a librarian of Linyi University in Shandong province, China.

Her current research interests are focused on information retrieval and instructional resources management.