Research Hotspots of Data Mining Technology in the Field of Library and Information

Hui Wang*

University Library, Jilin Agricultural University, No.2888 Xincheng Street, 130118, Changchun, Jilin, China
huiwang558@163.com
**Corresponding author*

Abstract: *With the rapid development of computer network information technology, data mining technology has begun to be widely used in all walks of life. In the field of Library and information technology, data mining technology can efficiently analyze and screen a large amount of information. This paper mainly studies the useful content, readers' interests and reading habits in the database. Firstly, it theoretically expounds the literature review of text classification methods and association rule-based algorithms by scholars at home and abroad. Then the data mining technology is used to establish the relevant models in the field of Library and information. Finally, the experimental results show that economics, geography, environmental resources, language, industrial technology and literature are the types of books that readers often consult. These five categories are a collection of common books, with a confidence level of more than 50%. Although the support of public project sets decreases with the increase of the number of public project sets, the average support of the five public project sets is 30.2%, which is higher than the minimum support of public project sets. Therefore, there is a strong correlation between the categories of books in the five project collections in the library.*

Keywords: *Data mining, Library and Information, Information Field, Research Hotspot*

1. Introduction

With the rapid development of computer network technology, data mining has been more and more widely used in various fields. As a new information processing mode, data mining gives a decision-making scheme after analyzing and classifying various relationships and laws in a large number of complex and massive original databases [1-2].

Many scholars have explored library and information. In the application of data mining in the field of Library and information, the United States and Japan have made in-depth exploration. As early as around 1970, the United States, Canada and other countries began to use computer network technology in Library and information work. So far, there are many methods to solve the problem of information processing and have achieved good results, such as using big data to establish a database to collect readers' interesting things or useful knowledge information. There are also a series of measures and means to improve decision-making efficiency through various algorithms, which have made outstanding contributions to the research of similar applications in the field of Library and information in China [3-4]. There have been some research results on book classification in China, such as book reading prediction based on content analysis and Book Word Segmentation and retrieval based on clustering technology. Some scholars believe that using the method based on text association model in the field of Library and information can make more effective use of resources and improve work efficiency and data mining ability. The above research lays the foundation for the of this paper.

This paper mainly introduces the research process of Library and information material objects and their types based on data mining analysis method. Firstly, the principle of classification and clustering technology and related theories are briefly described, and then the data sets contained in the case and the similarities between them are described by mining sample examples with similar characteristics.

2. Discussion on Research Hotspots in the Field of Library and Information Based on Data Mining Technology

2.1 Library and Information Field

The field of Library and information refers to the process of effectively screening a large amount of information through data mining technology to serve readers. With the rapid development of computer and network communication industry, the growth trend of people's demand for Internet applications and the arrival of the big data era, opportunities and challenges coexist and accelerate [5-6]. In this case, we need to improve our professional ability and quality to adapt to this rapidly changing world. At the same time, we should also pay attention to the impact of data mining technology on the research direction in the field of Library and information, because it is the emergence of a new thing and concept. In the field of Library and information, data mining technology has been widely used. From the early simple processing mode to now, data mining, machine learning and other methods can provide people with very effective, accurate and reliable information.

2.2 Application of Data Mining Technology in Library and Information Field

The use of data mining technology is very extensive and important in the field of Library and information. Using data mining in library has important practical significance. Now the library is an electronic information library, which can store a large amount of information in various forms. Users can easily access these information through the network, and information storage and user access are not limited by Region [7-8]. This can help readers get the news content they are interested in faster and better. For example, when searching for books, first input the relevant text, pictures and other image data into the computer, and then select the text and reference materials according to the required file type. Then send the query to the system and return to the user, and finally complete the data mining task. The library integrates the digitization, storage management, search and transmission of various information including multimedia, so that these information can be transmitted through the Internet and used to the greatest extent. At the same time, due to the development of computer technology, the amount of information that the library needs to preserve and transmit is becoming larger and larger, and the quantity and form of information are becoming more and more extensive. Therefore, it has become an urgent problem to use these information more reasonably and bring better personalized services to users. Due to the increasingly diverse requirements and forms of users' access to information, and the increasingly wide variety of documents, personalized information service has formed a new development trend. Personalized service also requires the support of important information such as user interest and association with books. This part of information can also be obtained through the classification and query of library daily operation data. In addition, enterprise data mining can also provide decision-making assistance for libraries, reasonably allocate collections, accurately capture users' personality and needs, and provide personalized services. The application of data mining technology in the field of Library and information is mainly to improve readers' utilization of information resources and content in the process of reading, and better meet people's needs for books and articles. Data mining technology is very good at dealing with abnormal laws that cannot be directly analyzed and understood by conventional methods [9-10].

2.3 Data Mining Technology

2.3.1 Concept

Data mining is a method of using information visualization technology to comprehensively analyze and process a large number of complex, unknown and uncertain things in many aspects, such as meaningful definition and implicit process, so as to obtain useful knowledge. In a broad sense, it is to find potential through various algorithms, which exist in the research hotspots in related fields in the real world. In the narrow sense, it is also called "nonparametric" mining. Data mining is based on big data, based on the principle of information theory, using a large number of useful data in the database, through the process of screening, classifying and processing massive and complex information, so as to obtain valuable or semi-structural key rules from a large number to individual users with different quality and the same characteristics, Extract valuable and useful knowledge [11-12].

2.3.2 Tasks

Data mining usually has the following six main tasks:

(1) Classification. Classification is one of the most common data mining tasks. Business issues such as risk management and advertising positioning often involve event classification. Classification refers to the classification of business cases into the following types according to a predictable attribute. Each event includes an attribute, where there is a predictable attribute or a Category attribute. In the classification task, we need to find a model that defines the classification attribute as a function of the input attribute. Classical classification calculation includes decision tree calculation, neural network calculation and Bayesian algorithm.

(2) Cluster analysis. System clustering is also called classification. It classifies events according to combination attributes. In a cluster analysis method, all cases have more or less the same attribute value. Because cluster analysis is an unsupervised data mining task, no attribute can be used to guide the whole model establishment process and treat all input attributes fairly. Therefore, most clustering algorithms use repeated iteration to establish the model, and the calculation stops when the model converges, that is, the calculation stops when the subdivision boundary tends to be stable.

(3) Correlation analysis. Association analysis is to find interesting associations or correlations between element sets in a large amount of data. Correlation can be divided into simple correlation, temporal correlation and causal correlation. The purpose of correlation analysis is to find out the correlation network hidden in the database. The commonly used correlation algorithm is a priori algorithm.

(4) Return. The regression problem is similar to the classification problem. The biggest difference is that all estimable attributes in the return task are continuous. The most common regression analysis techniques include linear regression and logistic regression. Other regression analysis techniques include regression tree and neural network.

(5) Forecast. Prediction is an important data mining task. The data set used in prediction technology is time series data set. Prediction technology can carry out overall trend analysis, periodic analysis and noise filtering. The most commonly used time series analysis method is ARIMA model.

(6) Deviation analysis. Gap analysis involves looking for special situations whose behavior is significantly different from other situations. Deviation analysis, also known as outlier detection, is used to identify behaviors that show significant changes compared with previously observed behaviors.

2.3.3 Association Rules

Although the common item crawling algorithm is better for some non dense databases, for dense databases, or when the support threshold is small, the number of records is relatively small. Common items are growing exponentially, so all common items cannot be found. Possible tasks. However, in fact, there is more redundancy in the common element set, so people use different methods to try to reduce the redundancy in the common element set. At present, frequent closed itemset (FCI) or maximum frequency itemset (MFI) are mainly used to reduce the result set. The concept of frequent closed itemset comes from the conventional concept analysis in mathematics. Let I be the set of items, t be the set of TIDs, and DB be the transaction database. Two mappings are defined:

$$X \in I, f(x) = \{y \in T \mid \forall X \in X, (x, y) \in DB\} \tag{1}$$

$$Y \in T, g(x) = \{x \in I \mid \forall Y \in Y, (x, y) \in DB\} \tag{2}$$

If G (f (x)) = x, X is called the closed set of elements. If f (g (y)) = y, y is called a closed label set. If x is the closed set of elements and Y is the closed set of labels, the binary relationship (x, y) is called a concept. For any group of common elements x, the common closed element set of X can be obtained by the operation of F and G. Public closed itemsets are several orders of magnitude smaller than public itemsets, but do not lose the frequency information of all public itemsets, so association rules can be generated from them.

The size of the largest common element set is the smallest of all common element sets. FCI and FI can be derived from MFI. However, MFI loses the frequency information of its subset, so it cannot generate any allocation rules. If you want to generate association rules, you also need to calculate the frequency of some subsets. However, for databases with long biological information models, MFI mining still has great practical value.

3. Experiment

3.1 Application Process of Data Mining Technology in Library and Information Field

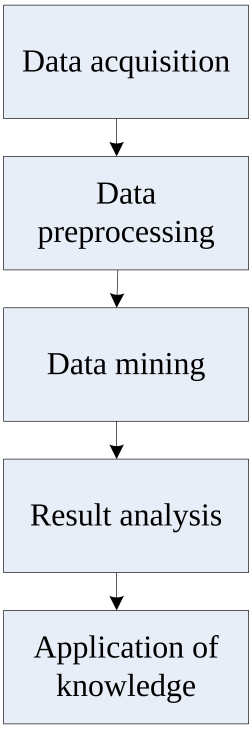

Figure1: Data mining process

The application process of data mining technology in the field of Library and information can be divided into three stages (as shown in Figure 1): the first stage is information collection, acquisition and analysis. It is mainly through sorting out a large number of original materials and transforming them into available and useful resources. This process needs the help of some new media, such as microblog, to realize the dissemination of these contents. The second stage (preprocessing) Data mining and prediction decision-making process and model establishment and improvement, as well as follow-up work. It is mainly to solve how to find effective information in massive library and information, and provide accurate and real-time analysis results and corresponding solutions. At this stage, a variety of models can be used for modeling. The third stage is to sort out and retrieve the relevant literature and knowledge base contents, and transform them into highly practical and quantifiable research objects or products, which contain useful information or useless value materials. After mining and processing the problems existing in the field of Library and information, form conclusions for decision-makers' reference.

3.2 Data Mining Technology Performance Test Steps

The purpose of data mining performance test is to analyze the errors in the processing, operation and maintenance of the system, and then judge whether the system can meet the business requirements according to the analysis results, so its function must be determined first. (1) Select the appropriate application environment. Different users have their own unique preferences for computer hardware and software. Therefore, it is a very important step to select suitable for their own development and use habits as the principle to evaluate the performance test method of data mining. (2) After the experimental scheme is designed and implemented, the whole system is analyzed and evaluated. A large number of invalid and redundant information are screened out and integrated to form new knowledge, and then these useless or redundant information are sorted together as training sets or statistical models to simulate possible problems in the actual situation, so as to improve the test efficiency, reduce the cost investment and reduce some unnecessary losses.

4. Discussion

4.1 Performance Test and Analysis of Data Mining Technology

Table 1 is the performance test data of data mining technology.

Table1: Data mining test

Test book type	Support(%)	Confidence level (%)
Economic	30.2	51.3
Geography and environmental resources	35.4	53.2
Language, text	34.8	54.2
Industrial technology	41.2	58.2
Literature	43.4	51.7

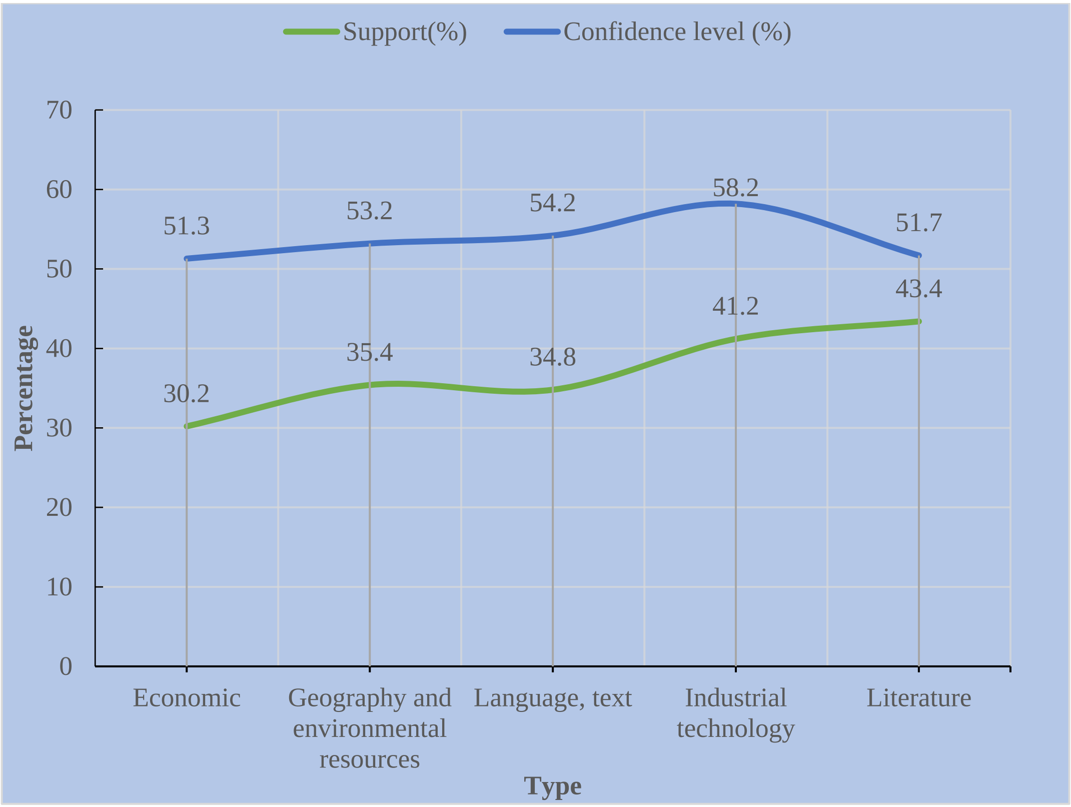

Figure 2: Data rule technical performance test

From the mining results above (Figure 2) it can be seen that economics, geography, environmental resources, language, industrial technology and literature are the types of books that readers often consult. These five categories are a collection of common books, and the confidence level is more than 50%. Although the support of public project sets decreases with the increase of the number of public project sets, the average support of the five public project sets is 30.2%, higher than that of public project sets Minimum support for the project set. Therefore, there is a strong correlation between the categories of books in the five project collections in the library. Among the five types of books, it is not difficult to deeply mine the original database. Economics, language, industrial technology and literature involve students' public basic courses, so the borrowing rate is very high. Therefore, when managing the library collection, we can try to combine the bookshelves of these five types of books, which can not only put the books on the bookshelf, but also facilitate readers to borrow, but also reduce the damage of books.

5. Conclusion

Data mining is the use of massive, complex and changeable information resources with regularity and potential significance. Its purpose is to obtain useful information without spending too much energy to find valuable information. This paper mainly studies the impact of data mining on a university library and readers in the field of Library and information. Firstly, this paper analyzes the current situation and existing problems of books and literature work in China, then summarizes the data mining technology, introduces the method based on clustering algorithm, the establishment and training process of database system and classification model, designs and tests, and obtains the results.

Reference

[1] Wang C, Qi H. Visualising the knowledge structure and evolution of wearable device research [J]. Journal of Medical Engineering & Technology, 2021, 45(3):1-16.

[2] Geest K V D, Warner K. Loss and damage in the IPCC Fifth Assessment Report (Working Group II): a text-mining analysis[J]. Climate Policy, 2019(596):1-14.

[3] Tapete D, Cigna F. InSAR data for geohazard assessment in UNESCO World Heritage sites: state-of-the-art and perspectives in the Copernicus era[J]. International Journal of Applied Earth Observation and Geoinformation, 2017, 63:24-32.

[4] Cheng Y, Huang A, Qi G, et al. Mining Customized Bus Demand Spots Based on Smart Card Data: A Case Study of the Beijing Public Transit System[J]. IEEE Access, 2019, PP(99):1-1.

[5] Okeji C C. Research output of librarians in the field of library and information science in Nigeria: a bibliometric analysis from 2000-March, 2018[J]. Collection Building, 2019, 38(3):53-60.

[6] Royal, Institute, of, et al. Education for research in library and information science: a basis for policy analysis in the Nordic countries[J]. Education for Information, 2017, 3(2):83-102.

[7] Wang Y, Zhao Y, Dang W, et al. The Evolution of Publication Hotspots in Electronic Health Records from 1957 to 2016 and Differences Among Six Countries[J]. Big Data, 2020, 8(2):89-106.

[8] Dwivedi S K, Tripathi R. Exhausting Agile Processing and Data Mining in Electronic Commerce[J]. International Journal of Scientific Research in Computer Science Engineering and Information Technology, 2019:80-84.

[9] Weng L M, Zheng Y L, Peng M S, et al. A Bibliometric Analysis of Nonspecific Low Back Pain Research[J]. Pain Research & Management, 2020, 2020:1-13.

[10] Pan X, Zhong B, Wang X , et al. TEXT MINING-BASED PATENT ANALYSIS OF BIM APPLICATION IN CONSTRUCTION[J]. Journal of Civil Engineering and Management, 2021, 27(5):303-315.

[11] Long F, Ning N, Zhang Y, et al. Mining latent academic social relationships by network fusion of multi-type data[J]. Social Network Analysis and Mining, 2020, 10(1):1-16.

[12] Ram, Vinay, Pande, et al. DaMold: A data-mining platform for variant annotation and visualization in molecular diagnostics research[J]. Human Mutation, 2017, 38(7):778-787.