Research on Readers' Behavior of College Libraries Based on Data Mining Technology

Jianfei Hu, Jinling Tang, Bin Zhang
Library, University of Jinan, Jinan, China

Abstract. College libraries produce a large amount of book circulation data every day. In addition to recording readers' personal information, these data are generally used for routine business statistics. Their potential use value is far from being fully tapped and utilized. A careful study of readers' long-term borrowing history data reveals that there is a certain correlation between readers' borrowing behavior and books, between different disciplines and between different types of readers and borrowing books. Digging out the correlation between these data is conducive to the rational allocation of resources and improving the utilization of resources. It can also improve the library's service level. Based on the author's learning and practical experience, this paper first analyzed the necessity of applying data mining technology to library and then discussed the basic techniques of data mining. The article proposed the basic steps of applying data mining to digital library systems and finally studied The application of data mining technology in the analysis of library readers' borrowing behavior.

1. Introduction
In the library's existing digital system, a large number of statistical data and forms are generated every moment. After these statistical data and forms are analyzed and processed, the internal information volume can be used for various services of the library. In particular, it will play a strong guiding role in procurement, collection, and consulting services. With the continuous development of the digitalization of libraries, the characteristics of the data that the library has to deal with and provide are more complex and presenting a multiple, new, and broad trend. This makes higher requirements for processing and providing data for libraries. If there is no deep-level development of these massive amounts of data, it will fall into the dilemma of “rich data but poor information”. With the improvement of the information level and requirements of library readers, it also objectively requires that the information services provided by the library to the readers can be more proactive and personalized. Therefore, the library should further strengthen its processing and analysis capabilities of readers' borrowing behavior information and the ability to further organize and integrate readers' borrowing information resources to extract valuable internal links from a large number of seemingly unscrupulous readers' borrowing behavior information for library, so as to satisfy readers' new and more complex borrowing requirements. In terms of the processing of statistical data and forms generated in the library's digital system, if it only stays in the breadth of data collection, these rapidly growing and excessive data may become "data tombs" and cannot provide valuable and instructive information to the library. For readers' data, with the popularization of digital library technology, a large number of readers' borrowing behavior statistics and forms are generated in the library's digitization system, which raises the requirements for libraries to handle and analyze readers' borrowing information. In recent years, data mining has become a very active field in applied research. It has a wide range of applications in various industries in China and also provides technical support for library information analysis. The use of various data mining
techniques in the digital library system can analyze various indicators of readers' borrowing information resources, such as the utilization of library resources, the efficiency of library resources and the denial rate of library resources. This will help the library to strengthen services, such as procurement, collections, and consulting service to improve service quality.

2. Analysis of necessity of applying data mining technology to library

2.1. The rapid development of information technology
With the advent of the age of information and networking, the library's existing automation system generates a large number of statistical data and forms every year, every month and every day. They play a guiding role on the procurement, retrieval, consulting and circulation of libraries. However, the rapid growth of excessive data collection tends to become "data tombs" and loses its guiding significance. It is difficult for libraries to digest excessive information, to distinguish the true and false information, to guarantee information security, to ensure consistent information forms and to handle the information in a unified manner. This urgently requires an effective data collection, processing and processing technology to provide technical support and decision management support for library work. Data mining is such an emerging technology.

2.2. A lot of data is not being used
With the rapid development of database technology and the extensive use of database management system in libraries, a large number readers' data of accessing and borrowing resources and books from the library is accumulated. There is a lot of important information hidden in these data and people want to be able to analyze it to use this data better to serve the reader. The current library management system cannot discover the relationships and rules that exist in these data and cannot predict readers' information needs. It lacks the means to mine the hidden knowledge in the data. In the ocean of knowledge in a library, readers find it is difficult to find the information resources they need. Using data mining techniques, libraries can explore and understand readers' interest and hobbies, and provide technical support and decision management support for library readers' needs analysis. This will make libraries more accessible and active to provide readers with the required information.

2.3. Data mining improves information acquisition speed
The library generates a large amount of book circulation data every day. In addition to recording the account information of readers, these data are generally used for routine business data statistics. The potential use value of these data is still far from being fully tapped and utilized. By using data mining, the library can develop targeted plans based on the mining results and take appropriate actions to improve the competitiveness of the library. The data, such as borrowing and retrieval information in the library automation system, is the reader's active satisfaction of personal information needs, and it is also the best evidence for readers to use library resources. The data mining and analysis of the library's borrowing records is to change the passive services of the library and the library will take the initiative to improve its overall image in readers' minds.

3. The basic overview of data mining technology

3.1. Selection of data mining technology
The traditional collection of library information is determined by a dedicated interviewer alone or with a small number of experts. It is unavoidably one-sided, subjective and personally pleasing. They cannot discover the relationships and rules existing in these data, and predict readers' information needs. Moreover, the rapidly growing and excessive data collection will often lead to the situation of "rich data but poor information" and the data loses its guiding value. This requires to find a powerful data collection and processing tool to solve this problem and data mining is exactly what is needed. The readers of college libraries mainly include a large number of undergraduates, postgraduates, employees of all
departments in the college, teachers and researchers who are responsible for different teaching tasks. Using data mining methods to analyze their historical behavior patterns and actively explore needs and preferences of different groups of readers will provide effective technical support and decision-making reference for the library's resource organization and management, budget allocation, service quality improvement and service scope expansion. Besides, it has important practical significance for improving the ability of the library to serve the scientific research.

3.2. Introduction to data mining technology
Data Mining (DM) is a process to extract unknowing and potentially useful information and knowledge from a large number of incomplete, noisy, fuzzy and random practical application data. It incorporates statistical methods, computer techniques and artificial intelligence techniques to extract patterns, changes, associations, anomalies, regularizations and statistics of important structures and events from large datasets semi-automatically. Compared with the previous structured data forms, the current data representation forms are more diverse. Therefore, according to the data structure of its main research objects, people divide the data mining in a broad sense into narrow data mining, web data mining and text data mining. Narrow data mining is aimed at structured data, which is common database or data warehouse, such as SQL Server and Oracle. Its application is mainly reflected in prediction, regression analysis, data classification, cluster analysis, correlation analysis, exploratory data analysis (EDA) and anomaly detection. Corresponding data mining technology and algorithms include statistical methods, case-based reasoning, rule-based reasoning, correlation analysis, sequence pattern algorithms, neural networks, decision-making trees, genetic algorithms, Bayesian belief networks, fuzzy sets, rough sets and visualization technology. It can be seen that on one hand, using data mining technology can analyze borrowing data to discover the needs of readers and find the collections that the library lacks, so that the library can provide targeted supplements and enrich resources. On the other hand, with the correlation analysis of the categories of borrowed documents by users, the library can find the characteristics of relation between readers' borrowing and associating behavior and various types of documents, so as to realize the optimization of library resources and improve the library's service quality.

4. Basic steps for applying data mining to the library's digital system
Data mining is a new technology after upgrading the traditional demand analysis technology. When applied in a library digital system, it can analyze the needs of readers from the massive historical data generated in the library digital system. This is a kind of decision support. This process will need artificial intelligence, machine learning and statistics to achieve highly automated analysis of massive amounts of historical data. Therefore, inductive reasoning is derived from the massive historical data to unearth the potential readers' borrowing behavior model to predict their borrowing behavior in the future. This will help library staff adjust book procurement strategies, further improve collection construction and strengthen consulting services of book information.

4.1. Establishing a data warehouse
The primary task of applying data mining technology to the library is to establish a data warehouse. There are many business data in the digital library system, and there are complex horizontal relationships between data (Figure 1).
Generally, information contained in the data warehouse can be divided into three main categories: reader information, book information and readers’ borrowing and circulation information.

When building a data warehouse, the library can do the following steps: First, data modeling of various business data in the library. This is a process of determining the subject of library data warehouses. The second step is to design database of the data warehouse and determine the data conversion procedures. The third step is to manage the metadata. It needs to define the metadata of the library data warehouse, signify and define various business data of the library with metadata, determine its specific meaning, and establish the relationship between the components of the database system. The fourth step is to determine analysis tools in the data mining. Based on the implementation and use of data warehouses, the library needs to establish a structured decision-making support queries to help libraries make decisions and meet library decision support needs.

4.2. Construction of data mining structure
After building the library data warehouse, it is time to start building the data mining structure. This process can be divided into the following aspects: First, the library needs to conduct data integration, filtering and conversion of a variety of business data in the library, including readers’ basic information, book and circulation information, and then store the data in the database or data warehouse server. Then, the library needs to establish a tool engine of data mining, call data analysis tools of data mining (such as Excel data analysis tool and OLAP data analysis tool), and store data mining algorithms in the knowledge base (such as association rules, cluster analysis and algorithms). This is the core process of data mining. Finally, the library obtains the data mining results of various business data and displays them to users. Various business data in the library calls back-end data packages for data mining through front-end development tools. The forms can be tables, cubes or other models, which are intuitive and concise.

4.3. Data mining analysis of the practical use of books
The library needs to use association rules, cluster analysis and other algorithms to analyze actual utilization of books through data mining. Massive data in the data warehouse, such as readers’ information, book information, and readers’ borrowing and circulation information, is structured data.
There are many algorithms for data mining. The two most commonly used methods are association rule algorithms and cluster analysis algorithm.

1) The use of association rule algorithm is mainly focused on the analysis of borrowing history data. The library analyzes the association rule algorithm of the borrowing history of information resources. For example, it is found that a large number of readers who read the A book also read the B book. In the service work of the library, a service for recommending B book to readers who read A Book needs to be added. The discovery of these rules is helpful for improving the utilization of library resources and providing more user-friendly services for the readers.

2) The use of cluster analysis algorithm is mainly based on the cluster analysis of recent books and historical books to determine the borrowing value of recent books for readers. The data warehouse built on the large number of statistical data and forms generated in the library's existing digital system is real and noisy. For libraries, readers' borrowing information and knowledge from these large amounts of statistical data and forms is potential. The understandable and applicable knowledge which libraries are interested in is hidden in these large amounts of readers' information, books information and circulation information.

4.4. Decision after data mining analysis

The basis for decision making after analyzing readers' borrowing behavior through data mining includes two dimensions:

1) Comparison of reader borrowing quantity information of books in readers' borrowing behavior data exhibition can be achieved by taking the time of duration of targeted group readers as the measure value and taking readers' category and total holding time to conduct comparison. The analysis of the actual holding time of the reader according to the reader category can be obtained.

2) Borrowing quantity dimension comparison of book information in the library is based on the "Chinese Library Classification Law". In this way, the library can make horizontal comparisons according to the holding time of certain books borrowed by all readers. The analysis shows that which type of book is more popular among readers and which type of book meets readers' needs better. The author once carried out statistics on the library and classified the book information into several levels, such as first-level category, second-level category, third-level category and book bar code level. The author found the popular and unwelcome book types to ensure that the library's collection resources can be used reasonably and fully. The author takes his library as an example to show decision making after analyzing the actual use of books through data mining. The decision-making tree is shown in Figure 2.
5. Application of data mining technology to the analysis of readers' borrowing behavior in the library

5.1. Establishment of a data warehouse

(1) Data source: historical borrowing and returning record of the author's library; (2) Owner: The author's library; (3) Number of records: 10,242 borrowing records and 9673 returning records (4) The format of the original data is mainly the Excel table, including 13 fields of borrowing record and 16 fields of returning record. See Table 1.

| number | name | unit | date of borrowing | book title | demand number | console operator |
|--------|------|------|-------------------|------------|---------------|------------------|

Table 2 Original data format of returning record

| number | name | unit | date of borrowing | date of return | demand number | console operator |
|--------|------|------|-------------------|----------------|---------------|------------------|

5.2. Data mining system construction technology

The technology of building data mining system in the author's library mainly includes Analysis services and SQL Server. Analysis services is characterized by the integration of OLAP functions and its OLAP interface is extensible and based on COM; SQL Server2000 can support all kinds of complex database calling functions, including support for terabytes data capacity, data transmission services (providing data input/output function), all intermediate data, such as metadata repository specification (Microsoft Repository), support for multi-dimensional O-LAP, relational OLAP, and mixed OLAP. It can be used to configure and fill in the data warehouse, can provide client-side OLAP data access function (Pivot Table Service) and Core Management Service (MMC) for scheduling and storage management.

5.3. Data Analysis - a case study of the analysis of borrowing record

In this analysis, readers borrowed a total of 1468 kinds of books, and the borrowing status is shown in Figure 3.

![Figure 3 borrowing status analysis](image-url)
There are 22 species with more than 100 borrowing times, accounting for 1.50% of the total category. 257 species between 10 borrowing times accounts for 17.51% of the total category. Categories with less than 10 borrowing times are 1,319, accounting 80.99% of all books. It can be seen that a large number of documents were read by readers during this period. For example, the A1 Marx and Engels books were only read by the readers once; a few documents were frequently read by readers, such as the H310.42 English proficiency test books which were borrowed 820 times. Why are so many books borrowed by readers a few times? This requires the library staff to analyze the in-depth reasons and maximize the utilization of this part of the literature. In addition, the procurement of this part of the book should be restricted, because there are enough collections that can meet the current readers' borrowing needs.

6. Conclusion
After analyzing and processing the statistical data and forms generated in the existing digital system of the library, its internal information can play a strong guiding role in various services of the library. The continuous development of the library's digitization process puts forward higher requirements for the processing of the data that the library has to handle and provide. It also objectively requires that the library can provide readers with more active and personalized information services. Data mining technology provides technical support for the library's information analysis work. It can analyze the indicators of readers' borrowing information resources and help the library to strengthen its procurement, collection, and consulting services. At the same time, with the improvement of library informationization, the library will accumulate a large amount of circulation data, which can provide rich data resources for further data mining.

References
[1] Ma Guiying, Li Huimin. A Research on the Role of Big Data Technology in the Construction of Digital Library of Application-oriented College [J]. Journal of North China Institute of Aeronautical Engineering, 2017,(04):55-57+60.
[2] Zhang Wei. A Research on Digital Library Service of Colleges Based on Data Mining Technology [J]. Agricultural Network Information, 2016, (11): 140-143.
[3] Li Renling, Wang Jianling. A Study on the Readers' Behavior and Research of Data Mining Technology in TCM College Libraries [J]. Contemporary Educational Practice and Teaching Research, 2016, (07):35-36.
[4] Yan Junxia. A Research on Readers' Borrowing Behavior of College Libraries Based on Data Mining Technology [J]. Science & Technology Square, 2014, (07): 110-113.
[5] Yang Yang, Jian Dazhao. An Application Research of Data Mining in College Libraries [J]. Lantai World, 2013, (26): 36-37.
[6] Tan Yunjiang, Ling Yun, Yan Haifeng, Dong Ruojian. A Research on Readers' Behavior of College Libraries Based on Data Mining Technology [J]. Library and Information Service, 2010,(S2): 359-362.
[7] Yao Yanyao. A Research on the Integration of Readers' Loyalty and Data Mining in College Libraries [J]. Information Science, 2010,(11):1646-1651.
[8] Zhang Wei, Hong Xia. An Analysis of Key Technology of Personalized Service System Based on OPAC Reader Behavior Mining [J]. Library Tribune, 2010, (01): 62-64.
[9] Zhang Wenhua. A Case Study of Data Mining Based on Readers' Behavior [J]. Library Magazine, 2009, (12):55-58.
[10] Jiang Changbin, Chen Li. A Research on Active Service of College Libraries Based on Readers' Behavior Analysis [J]. Journal of Wuhan University of Technology, 2009,(05):132-136.