



College Student Reader Model and Its Reading Recommendation Algorithm Based on Data Mining

Xin Peng^(✉) and Lixin Nie

Xi'an Eurasia University, Xi'an 710000, Shaanxi, China
pengxin@eurasia.edu

Abstract. This paper expounds the necessity of applying data mining technology to the recommendation service of university library, and introduces the present situation of applying data mining technology to the literature resource recommendation service, literature resource retrieval service and literature resource management service of library. The informatization construction of colleges and universities is one of the important fields of social informatization construction in China, and it is an important measure to comprehensively improve teaching quality and scientific research ability. The reading behavior of university readers is biased, utilitarian, the reading content tends to popular culture, the classical reading consciousness is relatively weak, and the reading style is digital. University library is an indispensable part of cultivating high-quality talents, and its informatization construction degree affects the cultivation level of college students' overall quality to a certain extent. As a service institution of higher education, university library should fully understand the reading trend of readers and grasp their reading interests, which directly affects the service quality of university library. The existing university library management system has accumulated a large amount of reader information and borrowing history information, which provides a data basis for recommendation service. This paper analyzes the three influencing factors of the construction of precision recommendation service model based on College Students' reading data mining technology, namely database construction, precision recommendation calculation method and reader oriented operation process.

Keywords: Data mining · College students · Readers · Model · Reading recommendation · Algorithm research

1 Introduction

As an important supporting body of scientific research services, the university library provides a platform and necessary conditions for the quality education of university readers, and plays a leading role in reading for the whole people with the advantages of collection, personnel, technology and equipment. With the development of readers' reading diversification and individuation, the library pays more and more attention to

the application of data mining technology in accurate recommendation service, which requires the library to analyze and model from massive data resources, screen out the target resources that match readers' needs to the maximum extent, and recommend them to readers. In recent 10 years, China has been advocating nationwide reading activities. As an important force in the industry, university library is also an academic institution for teaching and research services in colleges and universities, and it is the second classroom for college students. Reading promotion has become a routine activity and duty requirement of university library. However, there are some urgent problems to be solved in the construction of library digital service, such as how to improve the management level and service efficiency of the library, how to promote college students to make more efficient use of library resources and so on. University libraries have accumulated a large amount of data in the process of automation and informatization. Because the resources are updated every day, it is difficult for readers to obtain the latest resource information at the first time in most cases. Moreover, because the readers of the library have different backgrounds such as their age, major and identity, the demand for books is also different [1]. Therefore, in order to solve these problems, we need to carry out personalized recommendation through data mining, so that it can make a scientific, objective and reliable judgment on the needs of users.

As a relatively novel technology, data mining technology has been widely used in commercial fields such as banking, telecommunications and other industries, which provides a new perspective for the development of university libraries. University library, as the biggest reading source for readers, is also an important position of university education, and it is more important to understand the change of readers' reading tendency, which is of great theoretical and practical significance for library collection construction, readers' reading guidance, readers' education, etc., and also of great significance for giving full play to the educational function of university library [2]. In a broad sense, digital library is a computer processable and orderly information collection, which can also be regarded as the storage of digital information. It uses digital technology to organize and manage information resources. Users can query and retrieve information efficiently and conveniently through the network to obtain information services, and its information storage and user access are not limited by time and region. As the main reading group of society and colleges, College Readers' reading behavior has become an important research object. In practical work, this kind of accurate recommendation is the concrete expression of the library's active service concept. Accurate recommendation can provide readers with highly targeted and personalized information services, which fully reflects the library's reader centered service idea. In order to improve the sustainability and effectiveness of reading promotion in university libraries, university libraries need to make extensive analysis of readers' borrowing, resource browsing and downloading, and use data mining technology to find out the relevance, understand the knowledge needs of college students at different stages of school, and actively carry out reading promotion.

2 Data Mining Related Theory and Technology

2.1 Data Mining Concepts

With the rapid development of information technology and Internet technology, data shows an explosive growth trend. As the construction and exchange center of document resources, library pays more and more attention to the construction and improvement of database. Data mining and knowledge discovery are rising with the development of information technology such as database, data warehouse, machine learning and so on. Using the algorithms of association rules, clustering and decision tree in data mining technology, this paper models different situations respectively, and excavates the information and knowledge about students' borrowing habits, so as to guide the reading promotion of college students. As a new technology, data mining has been applied in some fields and made some achievements. The most typical case is the case of beer and diapers. The data mining technology such as association analysis is applied to the mining database to obtain the readers' borrowing habits, interest patterns and reading trends, so as to establish the readers' borrowing behavior model. Data mining can find the potential laws hidden behind the data through certain algorithms. Therefore, data mining provides a new way to solve this problem. It uses data mining technology to analyze the circulation data of the library, and then master the reading characteristics and reading trend of readers [3]. The data mining system consists of various databases (see Fig. 1), pre-processing module, mining operation module, pattern evaluation module and knowledge output module, which constitute the architecture of the data mining system. Data mining, as the core technology of knowledge discovery, is a process of extracting or mining information or knowledge from large-scale data sets, which is hidden and unknown to people in advance and valuable for decision-making. Simply put, data mining is a process of "mining" or extracting knowledge from specific data sets such as databases, data warehouses or other information bases.



Fig. 1. Database management module

2.2 Research Status of Data Mining

Compared with foreign countries, domestic research on data mining started a little late and immature, and is currently in the stage of development. Latest development: in the research of classification technology, we try to establish its set theory system to realize massive data processing; Constructing fuzzy system identification method and fuzzy system knowledge model; Construct intelligent expert system. Data mining system is an application software system integrating information retrieval, management, expert system, analysis and evaluation, data warehouse and so on. Scholars use the key technology

of data mining to mine the knowledge hidden in the book circulation data, and master the characteristics, behavior and trend of readers' reading, so as to improve the service quality of the library [4]. Pre-process the data and determine the type of data mining to be carried out. The basic steps of converting data into an analysis model CRISP-DM model include: business understanding-data understanding-data preparation-model building, model evaluation and model implementation. Figure 2 describes these steps and their relationship. In data mining, cluster analysis, association analysis and decision tree are very important methods, and scholars generally only use one of them to study. In industry, media and research, data mining usually refers to the whole process of knowledge discovery, that is, data mining in the broad sense above, so data mining can be defined as: data mining is the process of mining interesting patterns and knowledge from a large amount of data.

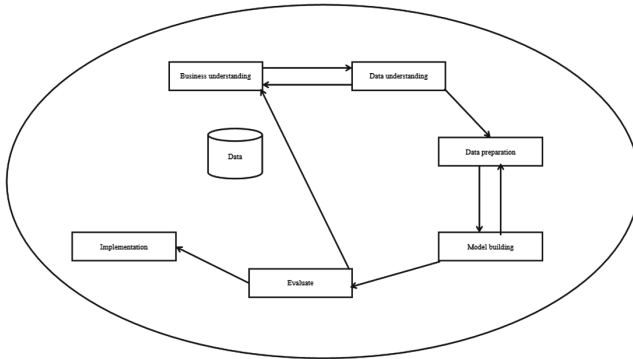


Fig. 2. CRISP-DM data mining process

3 Algorithm Selection of Library Circulation Data Mining

3.1 College Students' Reading Data Collection

In the face of a large amount of library lending information and students' personal information, data suitable for data mining applications are screened out. Because of the differences in readers' reading purposes, ways and habits, the reading behavior varies from person to person, but there are common manifestations. According to their own knowledge discovery, knowledge construction, readers gradually establish their own reading behavior. The purpose of constructing accurate recommendation service mode based on reader reading data mining in library is to provide personalized and distinctive push experience for readers. Data mining and association analysis based on the characteristics of reader reading data can highlight this individuality and characteristics. Through the research of classification technology and mining algorithm in data mining, this paper classifies the data type characteristics such as borrowing type, author type and college major in the existing lending records of libraries, and establishes data sources. Through

the analysis of reader reading data mining technology and library accurate recommendation service, this paper puts forward a specific service mode implementation scheme from four aspects: front-end application, algorithm implementation, role modeling and data collection (see Fig. 3) [5].

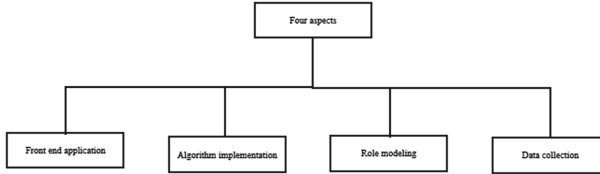


Fig. 3. Four aspects

Find out the hidden association network in the database through association analysis: on the one hand, use association rules to find the bibliography with strong association; On the other hand, the use of classification technology to determine the formation of multiple reader groups with different styles and interests, and recommend the bibliographies that have been consulted more frequently in the reader group, so as to establish a library personalized recommendation model. At the level of data collection, the library should fully consider the situation of readers' historical browsing records, readers' behavior and professional reserves, so as to recommend information resources with higher value for readers and improve readers' experience satisfaction. Classify data types and establish data sources [6]. At the same time, according to the characteristics of library data, through association analysis, find out the hidden association network in the database and establish the model.

3.2 Cluster Analysis Algorithm Selection

The formal description of cluster analysis is: the given data set is $V = \{V_1, V_2, \dots, V_n\}$, where v_i is called data point, and the data set is divided into K groups according to the similarity between data points, namely: $\{C_1, C_2, \dots, C_n\}$, $C_i = \{C_{i1}, C_{i2}, \dots, C_{in}\}$ ($i = 1, 2, \dots, k$). The algorithm of library circulation data mining chooses K-Means algorithm in clustering method, hierarchical method [7].

K-means algorithm is an algorithm based on centroid. It divides an object into clusters (parameters), so that the similarity within clusters is high and the similarity between clusters is the lowest. It mainly refers to dividing n objects into k categories according to their attributes, $K < n$, then locating the central position of K categories, calculating the distance from n objects to K central points, reclassifying n objects into the nearest category, recalculating the central position of each category, and pushing them in turn until the end. Similarity calculation is based on the average of all objects in a family (regarded as the center of gravity of a cluster or cluster). Generally, the criterion function is the sum of squares of distances, as shown in formula (1).

$$E = \sum_{i=1}^k \sum_{p \in C_i} |p - m_i|^2 \quad (1)$$

where p is the point in the space representing the object, and m is the central point of cluster c .

Hierarchical method decomposes a given data set object hierarchically until certain conditions are met. According to the way of decomposition and formation of hierarchy, hierarchy method can be divided into two types: condensation method and splitting method [8]. Density-based method the density-based method overcomes the disadvantage that the distance-based algorithm can only find the circle-like clustering algorithm. As long as the density of nearby areas (the number of data points or objects) exceeds a certain threshold, clustering will continue.

Nearest neighbor collaborative filtering recommendation technology is the most successful recommendation technology at present. Its basic idea is to generate recommendations to target users based on the score data of nearest neighbors with similar scores, that is, generate a recommendation list for target users according to the views of other users. The analysis of library circulation data based on association rules is actually the association mining of the bibliographic information borrowed by readers, and then find out the effective borrowing rules, and then recommend the books that may be interested to readers according to the readers' borrowing interest. Whether the nearest neighbor query of the target user is accurate or not is directly related to the recommendation quality of the whole recommendation system. In order to accurately query the target user, it is necessary to accurately calculate the similarity between different items, so how to calculate the similarity between items becomes the key to improve the recommendation accuracy.

4 Cosine Similarity

The user rating is regarded as a vector in the N -dimensional item space. If the user does not rate the item, the user rating for the item is set to 0, and the similarity between users is measured by the cosine angle between vectors. Let the scores of user U and user V in the N -dimensional project space be expressed as vectors U and V respectively, and the similarity calculation method between user U and user V is shown in formula 2.

$$\text{sim}(u, v) = \cos(u, v) \frac{u \cdot v}{|u| \times |v|} \quad (2)$$

In the formula, the numerator is the inner product of two user scoring vectors, and the denominator is the product of two user vector modules. The smaller the included angle is, the higher the similarity is.

(1) Correlation similarity

Suppose that the item set scored by user u and user V is represented by I_{uv} , then the similarity $\text{sim}(U, V)$ between user u and user V is measured by Pearson correlation coefficient, as shown in formula (3).

$$\text{sim}(u, v) = \frac{\sum_{C \in I_{UV}} R_{U,C} \cdot R_{V,C}}{\sqrt{\sum_{C \in I_{UV}} R_{U,C}^2} \sqrt{\sum_{C \in I_{UV}} R_{V,C}^2}} \quad (3)$$

where, $R_{u,c}$ represent the user U 's rating on item C , $R_{u,c}$ and $R_{v,c}$ represent the average rating of user U and user V respectively.

5 Design of Collection Recommendation System Model in University Library

5.1 Decision Tree Algorithm Modeling

Using $C_{5.0}$ decision tree algorithm establishes the model, takes the reader's "major", "gender" and "grade" as the input variables, and "book category" as the prediction variables to establish the decision tree model. There is an obvious characteristic of College Readers' reading, that is, reading only when it is useful. What books are useful to readers will be read by readers. For example, some of the list belong to exercise guidance reference books. Readers read such books simply to assist learning, which has obvious pertinence and purpose [9]. Accurate recommendation calculation method is the technical guarantee to improve the efficiency of Library recommendation service. The library can choose the appropriate recommendation calculation method according to the actual reading situation of readers. The decision tree will recursively branch all borrowing data based on the input variable values. The first branch, also known as "root", contains all borrowing data records. Then, the root will be sorted according to the importance of input fields such as major, grade and gender, and then divided into several subsets or sub-branches. Each sub-branch can be further divided into secondary sub-branches, which can be further divided, and so on. Branches that are no longer divided are the lowest branches of the tree. Such a branch is called a blade. Finally, an analysis node is added to this model, which can determine whether the accuracy of the model can meet the special requirements.

5.2 Building Models Using Clementine Data Mining

Using the export node in Clementine data mining software, you can create a new field of "Borrowed Quantity of Books by Categories" according to the field of "Balance Quantity of Books by Categories" in the borrowed data, and make the borrowed quantity of books by categories less than 3 false and greater than 3 true, and establish a data mining model. The construction of library's accurate recommendation service mode based on reader data mining technology needs to take the dynamic data generated by readers' reading as the research ontology. The library resource recommendation system not only emphasizes the modeling of literature resources, but also attaches importance to the modeling of user data. The library should also make full use of big data mining algorithms to improve the matching degree between the target documents of readers' retrieval needs and the documents recommended by the library system, make full use of big data mining technology to analyze readers' reading rules and explore the internal relationship between data [10]. In the next step, we will mine and analyze the use of various electronic resources, provide personalized browsing and navigation of the network and personalized push services of network information resources, avoid readers' blind browsing on the Internet, and realize the filtering of network information resources,

so as to save the time of looking for resources, improve the reading experience of college students, and cultivate the character of College Students' loving knowledge, Cultivate the habit of active reading and achieve the behavioral goal of lifelong love for learning.

6 Conclusions

Reading is a necessary way for college students to grow up and become talents, and it is also the cornerstone of knowledge society. The library should find out the reasons from its own perspective, increase the publicity of the library, even if sending push information to every classmate, try its best to increase the awareness of the library in readers' minds, actively improve the service quality of the library, provide more personalized services and make readers more willing to use the library. The successful application of data mining technology in the commercial field has brought inspiration to researchers in the field of university library. This paper studies the related knowledge of data mining technology and recommendation technology, and by combining three common technologies of data mining technology with university library services, especially reader services, constructs a collection recommendation system of university library based on data mining, which mainly mines association rules of circulation data of existing systems. Regular user data can lay a good foundation for the accurate recommendation service of the library, and help the library accurately retrieve the target resources of readers in a large number of literature data, so as to achieve the service objectives of reducing resource acquisition time, simplifying resource retrieval methods, optimizing resource recommendation efficiency and improving reader satisfaction.

References

1. Huang, W.: Research on accurate recommendation service mode based on reader reading data mining. *Lib. J. Henan* **39**(6), 3 (2019)
2. Zhu, Y.: Research on personalized recommendation model of university library based on data mining. *Times Financ.* **26**, 2 (2017)
3. Wwang, Z., Hu, D.: Research on virtual college student reader model and its reading recommendation algorithm. *Inf. Theory Practice* **41**(6), 7 (2018)
4. Zhang, R., Zhang, Y.: Research on intelligent recommendation technology of electronic library based on data mining algorithm. *J. Southwest Normal Univ.: Nat. Sci. Edn.* **42**(7), 5 (2017)
5. Ping, L., Wang, Y.: Research on book recommendation system of university library based on data mining. *Comput. Prod. Circ.* **11**, 1 (2019)
6. Liu, X., Cai, N.: Research on the application of data mining technology in book recommendation service. *Archit. Eng. Technol. Des.* **000**(030), 1650 (2017)
7. Wang, Q.: Research on personalized fast recommendation algorithm of library based on data mining technology. *Modern Electron. Technol.* **42**(5), 4 (2019)
8. Sun, W., Zhang, N., Li, Y., et al.: An empirical study on the reading needs of library readers based on improved K-Means algorithm. *New Century Lib.* **5**, 7 (2020)
9. Zhu, D.: Research on personalized reading promotion for college students based on data mining technology-taking Anhui Normal University as an example. *J. Tongren Univ.* **19**(7), 7 (2017)
10. Wang, H.: Research on the application of book recommendation algorithm based on data mining technology. *Modern Inf. Technol.* **3**(23), 3 (2019)