Application of Data Mining in the Field of University Libraries for Book Borrowing Services

Zhihui Liu
Hanjiang Normal University Library, Shiyan, Hubei 442000, China
E-mail: zhihlzhui@outlook.com

Keywords: data mining, book borrowing, cluster analysis

Received: April 19, 2024

University libraries are knowledge resource repositories open to all teachers and students. Using data mining techniques to explore the hidden relationships between readers and books in the borrowing data can maximize the utilization of information and better apply it to the library field. This article uses cluster analysis to implement personalized recommendation algorithms for books. Two recommendation algorithms based on book clustering and reader clustering were proposed, and a hybrid recommendation algorithm was formed by combining the two algorithms. A case analysis was conducted. The results showed that the hybrid recommendation algorithm provided more personalized and specific recommendations and had a higher accuracy than the other two algorithms regardless of how many books the target user borrowed.

Povzetek: Tehnike rudarjenja podatkov so uporabljene za izboljšanje knjižničnih storitev pri izposoji knjig. Predlagana je hibridna priporočilna rešitev, ki natančneje personalizira knjižne priporočilne sezname.

1 Introduction

The library is a place where university students can access a large amount of information [1]. The collection of a university library is extensive and covers a wide range of disciplines. Almost every day, students or teachers borrow books from the library's collection, resulting in a large amount of borrowing data. How to use the borrowing data to obtain hidden information about readers and provide better services to readers of different types and disciplines and to manage the library more efficiently has become a focus of attention for librarians. Relevant studies are reviewed in Table 1.

Table 1: Relevant studies.

Author	Main content
Wang [2]	The author proposed an integrated method for electronic book borrowing history data in libraries based on big data technology. It was found that this method had the advantages of short delay and low integration error rate and the maximum integration throughput reached 97%.
Wang et al. [3]	They used the back-propagation neural network (BPNN) algorithm to model and analyze students' book borrowing. The results of their study suggested that students' scores decreased as their reservation and borrowing frequency decreased and students who often borrowed books had a strong motivation to learn.
Silwattananusarn and Kulkanjanapiban [4]	They used data mining techniques such as association rules and clustering to analyze the relationship between university library usage and student grades. The research results revealed the patterns of customer borrowing behavior and the relationship between book usage patterns and student grades.
Iqbal et al. [5]	They constructed new prediction models based on a deep neural network (DNN), support vector regression, and random forests. They found that the performance of the DNN model was significantly better than that of the other two models and the DNN model could help library management departments effectively plan and manage library resources.
Xie [6]	The author proposed an artificial intelligence-based library classification method and found that books could be marked and classified with invisible colors using an algorithm for detecting book image edges.

2 Data mining algorithm design

2.1 Cluster analysis

Data mining is to discover hidden information through integration and analysis of a large amount of real and

effective data. In universities, both physical and online libraries are available [7]. By leveraging data from offline library transactions, such as book borrowing records, and electronic activities like browsing and downloading e-literature in the online library, data mining techniques can be employed to uncover hidden information. According to relevant literature, the methods for information analysis mainly include cluster analysis [8], association analysis [9], and time series analysis [10]. This paper chooses cluster analysis as the main way to mine book borrowing data. Cluster analysis effectively categorizes similar samples into distinct groups, which has the advantages of simple principles and easy implementation [11]. The cluster analysis process used in this paper is as follows.

(1) The borrowing data of books in the library is collected and preprocessed, which includes removing irrelevant textual information, employing interpolation to fill in missing numerical values, and normalizing the data.

(2) n data are selected from X data as the initial cluster centers.

(3) The distance (similarity) between every object and these n centers is calculated using Euclidean distance, and the formula is:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n} (x_j - x_{i,j})^2} \quad , (1)$$

where x stands for a sample, x_i is a clustering center, x_j stands for the j-th eigen value of x, and $x_{i,j}$ is the j-th eigen value of x_i. Each sample has n features.

(4) Based on the calculated distance, the object is grouped into clustering center c_i with the shortest distance.

(5) Steps (3) and (4) are repeated until the computed values no longer significantly change; otherwise, clustering continues [12].

2.2 Personalized book recommendation algorithm

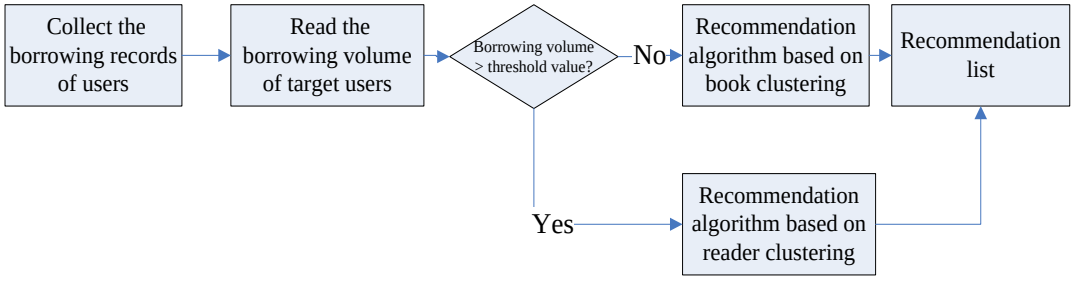

Figure 1: The hybrid algorithm based on the clustering algorithm for personalized book recommendation.

As mentioned above, using clustering algorithms can uncover hidden patterns in borrowing records within the library and provide personalized recommendation services to users based on these discovered patterns. For example, readers can be clustered based on their borrowing records in the library, and when recommending books to readers, relevant book titles can be obtained from the borrowing records of readers in the same cluster. Additionally, books can be clustered based on their borrowing records, and when recommending books to readers, recommendations can be made according to the classification based on the popularity of the books. In the above two recommendation algorithms, the former can take into account the reader's interests and hobbies, but it requires a sufficient number of borrowing records that reflect the reader's interest preferences. The latter recommends books based on their popularity without considering individual reader's borrowing records. However, this method only recommends relatively popular books, which may not be of interest to the reader. Therefore, this article combined both recommendation algorithms, and the flowchart of the combined algorithm is shown in Figure 1.

(1) The borrowing records of user groups in the library are collected.

(2) The borrowing volume of the target user to be recommended is read.

(3) It is determined whether the borrowing volume of the target users exceeds a threshold value. If not, it indicates that they do not borrow books frequently. Then, a recommendation algorithm based on book clustering is used. Firstly, the total number and annual number of borrowings for each type of book are calculated as feature attributes for book samples. Then, equation (1) is used to cluster and classify books into three types: popular, general, and unpopular. Lastly, a recommendation list ordered by popularity is provided to readers.

(4) If the borrowing volume of the target user exceeds the threshold, it indicates that the user frequently borrows books and has formed an interest range. Therefore, a recommendation algorithm based on reader clustering is adopted. First, the total number and annual number of borrowings for books are calculated as feature attributes for book samples. Then, according to equation (1), readers are clustered into active, general, and passive types. Finally, books from the borrowing records of the same cluster as the target reader are recommended.

3 Case analysis

3.1 Data source and processing

In this study, the circulation data of the library of Hanjiang Normal University were selected as the experimental data subject. The circulation data of paper books in the library for the whole year of 2022 were extracted to construct the data set. Due to the disorder of the data, the data were processed. The borrowing records and reader information were combined together, obtaining a total of 27,882 borrowing records. Useless information in the borrowing data was deleted, such as reader ID, class information, overdue days of unreturned books, borrowing dates, etc. Useful information, such as reader major information, book borrowing times, book information, etc., was screened out. The processed borrowing information is shown in Tables 2 and 3.

Table 2: The overview table of book information.

Serial number	Category number	Book category name	Book name	Number of borrowings	Renewal times
1	F	Economy	*The Economic Way of Thinking*	42	7
2	H	Language, writing	*Jane Eyre*	27	0
3	J	Art	*Power of Color Matching*	31	3
4	A	Marxism-Leninism and Mao Deng Thought	*Red Start Over China*	10	1
5	B	Philosophy	*Eastern and Western Cultures and Their Philosophies*	37	3
6	X	Environmental science, safety science	*Introduction to Environment Protection*	53	5
7	S	Agricultural science	*Agricultural Mechanics*	14	7
8	T	Industrial technology	*Hands-On Programming with R*	50	10
9	C	General social science	*Introduction to Social Work*	33	2
10	D	Politics, law	*Private School of Criminal Law*	12	8
.....	.....	.....	.....	.....	.....

Table 3: The overview table of reader information.

Serial number	School	Major	Grade	Number of borrowings
1	School of Economic and Management	Economics and finance	Freshman	8
2	School of Foreign Languages	English	Freshman	15
3	School of Physics and Electrical Engineering	Physics	Junior	17
4	School of Economic and Management	Auditing	Sophomore	21
5	School of Foreign Languages	Business English	Senior	16
6	School of Chemical and Environmental Engineering	Applied chemistry	Sophomore	26
7	School of Physics and Electrical Engineering	Electrical engineering and automation	Freshman	29
8	School of Mathematics and Computer Science	Network engineering	Freshman	30

9	School of Chemical and Environmental Engineering	Environmental engineering	Senior	15
10	School of Art	Musicology	Junior	10
.....	.....	.....	.....	.....

3.2 Experimental setup

In the hybrid recommendation algorithm used in this article, the K values for both the book clustering-based algorithm and the reader clustering-based algorithm were set to 3. To validate the proposed hybrid algorithm, a comparison was made with other book recommendation algorithms, namely association rule-based algorithm and collaborative filtering-based algorithm. The support threshold for the association rule-based algorithm was set to 0.1, and the confidence threshold was set to 0.5. For the collaborative filtering-based algorithm, non-repetitive book items were selected from the borrowing records of the top ten users with highest similarity as recommended results.

In addition, due to the use of clustering algorithms in the hybrid recommendation algorithm adopted in this article, i.e., a book-based clustering algorithm and a reader-based clustering algorithm, when recommending books, the appropriate clustering algorithm will be selected based on demand. This means that the quality of the clustering algorithm directly affects the quality of the recommendation algorithm. Therefore, this article set different values for K (2, 3, 4, 5) in order to compare and evaluate the classification performance of different clustering algorithms under different K values. The evaluation indicators are:

$$\begin{cases} s_i = \dfrac{1}{N}\sum_{j=1}^{N}\dfrac{b_{ij}-a_{ij}}{\max(a_{ij},b_{ij})} \\ s = \dfrac{1}{M}\sum_{i=1}^{M} s_i \\ DBI = \dfrac{1}{M}\sum_{i=1}^{M}\max_{i\neq j}\dfrac{s_i+s_j}{d_{ij}} \end{cases}, \quad (2)$$

where s refers to the silhouette coefficient of the clustering algorithm, s_i is the silhouette coefficient of the i-th kind of cluster in clustering results, DBI denotes Davies-Bouldin index, a_{ij} is the average distance between sample j in the i-th kind of cluster and the other samples in the cluster, b_{ij} is the average distance between sample j in the i-th kind of cluster and samples in the nearest cluster, N is the total number of samples in the i-th kind of cluster, M is the total number of clusters, and d_{ij} is the distance of the clustering center between clusters i and j.

3.3 Experimental results

The silhouette coefficient and DBI were employed to measure the classification effect of the clustering algorithm. The larger and closer to 1 the silhouette coefficient was, the better the classification effect. The closer to 0 the DBI, the better the classification effect. The classification performance of clustering algorithms based on books and readers under different K values is shown in Table 4. It can be seen that as the K value increased, the classification effect of both clustering algorithms first improved and then deteriorated. When the K value was 3, the clustering algorithm had the best classification effect.

Table 4: The performance of clustering algorithms under different K values.

Algorithm	Evaluation indicator	2	3	4	5
The book-based clustering algorithm	*S*	0.75	0.88	0.64	0.42
	DBI	0.31	0.11	0.34	0.59
The reader-based clustering algorithm	*S*	0.73	0.87	0.65	0.41
	DBI	0.32	0.12	0.35	0.61

Table 5 presents partial recommendation results obtained by applying three different algorithms to various target users. Due to the large number of books, it is challenging to list their specific names individually. To provide a clear overview of the tendency in the recommendation results, numerical codes were used to represent book types. Firstly, when comparing the recommendation results of three different algorithms for the same individual, it can be observed that there were differences in the recommended results among all three algorithms. The collaborative filtering algorithm suggested a diverse range of book genres, while the association rules algorithm only showed variations in one or two book categories. On the other hand, the hybrid recommendation algorithm consistently yielded recommendations with similar book genres. When comparing the recommendation results of different users utilizing the same collaborative filtering algorithm, it can be observed that a diverse range of book genres were recommended. In contrast, association rule-based recommendations exhibited some overlap in the recommended books among various users. Regarding hybrid recommendations, the recommended book genres varied across different users while maintaining consistency within each individual user.

Table 5: Partial recommendation results of three algorithms.

Reader serial number	The recommendation result of collaborative filtering	The recommendation result of association rules	The recommendation result of the hybrid algorithm
1	D92;F923;E9;O1;I2	F0;F1;F49;A2;A8	F0;F1;F49;F4;F7
2	F5;D9;G2;F12;G1	H0;H3;H81;B5;A2	H0;H3;H81;H9;H83
3	O1;O4;P3;F12;H1	T-0;T-6;TD;C3;C8	T-0;T-6;TD;TF;TS
10	G2;F12;D9;G1;I2	D0;D2;D4;F49;F59	D0;D2;D4;D8;D9

This article used accuracy as a measure of the performance of book recommendation algorithms. The calculation method for accuracy involves first utilizing the recommendation algorithm to provide suggested results and then tallying the number of books that users find interesting among these recommendations. The proportion of these books in the recommended results represents the accuracy of the book recommendation algorithm. The accuracy of three different recommendation algorithms for users with varying borrowing volumes is depicted in Figure 2. It can be observed that as users' borrowing volume increased, the collaborative filtering algorithm exhibited a gradual improvement in accuracy, while the association rule algorithm initially experienced an increase and subsequently declined. The hybrid algorithm consistently achieved the highest level of accuracy overall. However, for users with a borrowing volume of less than 30 books, the association rule algorithm outperforms the collaborative filtering algorithm in terms of accuracy. Conversely, for users with a borrowing volume of at least 40 books, the collaborative filtering algorithm surpassed the association rule algorithm in terms of accuracy.

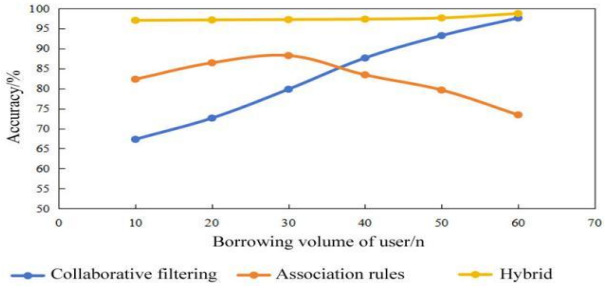

Figure 2: Accuracy of recommendation results for users with different borrowing volumes obtained using three algorithms.

4 Discussion

In college libraries, there is typically an extensive collection of books encompassing various genres. However, relying solely on individual searches within the library is inefficient and challenging for users in determining a book's relevance based solely on its cover. The emergence of computers and the internet has significantly facilitated book retrieval, enabling users to efficiently search for relevant books using keywords. However, keyword-based book retrieval is a relatively precise approach, and its efficiency can be greatly reduced if users lack a clear direction for their search. The book recommendation algorithm proposed in this article utilized data mining on user borrowing records and recommend books accordingly.

The book recommendation algorithm was enhanced by incorporating a clustering analysis algorithm in this paper, which effectively classified both users and books. Users were categorized into three groups: active, general, and passive, while books were classified as popular, average, and unpopular. The hybrid recommendation algorithm was formed by integrating the book clustering-based algorithm with the reader clustering-based algorithm. During usage, the selection of which clustering recommendation algorithm to employ was based on the user's book borrowing volume: when the borrowing volume fell below a predetermined threshold, the book clustering-based algorithm was utilized; otherwise, the reader clustering-based algorithm was employed. The final case study compared the hybrid recommendation algorithm with other recommendation algorithms based on association rules and collaborative filtering, as demonstrated in the previous section. The proposed hybrid algorithm was capable of consistently providing users with relevant recommendations tailored to their professional field. In terms of accuracy, not only did this hybrid recommendation algorithm outperformed others, but it also maintained stability even when faced with an increasing number of user borrowings. The reasons for the aforementioned results are analyzed. The collaborative filtering algorithm and association rule algorithm approach book recommendations from different perspectives. The collaborative filtering algorithm focuses on recommending books based on users with similar interests, emphasizing their interests as a key factor. However, it is important to note that having similar interests does not necessarily imply reading the same genre of books, which leads to diverse book recommendations. The association rule algorithm explores the collective borrowing records of users to uncover the association rules among books, i.e., the interconnections between different books. However, it fails to consider individual variations in user interests. Consequently, the recommended outcomes tend to be relatively limited and overlapping in terms of book categories. The hybrid algorithm integrates cluster analysis for precise clustering and incorporates borrowing volume as a threshold to flexibly adjust the utilization of two recommendation algorithms. This ensures that the recommendation algorithm is consistently employed in appropriate scenarios.

5 Conclusion

The present article employed cluster analysis to implement personalized book recommendation algorithms. It introduces two recommendation algorithms based on book clustering and reader clustering, which were subsequently combined to form a hybrid recommendation algorithm. An case study was then conducted. The findings are as follows. (1) The recommendation book categories of the hybrid algorithm remained consistent within an individual user, while they varied among different users. (2) With an increase in the borrowing volume of the target user, the collaborative filtering algorithm's recommendations exhibited improved accuracy, whereas the accuracy of the association rule algorithm's recommendations initially rose and then declined. In contrast, the hybrid algorithm maintained stability and consistently achieved superior accuracy.

References

[1] Lisnawita, Devega M (2020). Implementation of ECLAT Algorithm Technology: Determining Books Borrowing Pattern in University library. *IOP Conference Series: Earth and Environmental Science*, 469(1), pp. 1-6. https://doi.org/10.1088/1755-1315/469/1/012036

[2] Wang H (2020). Research on the integration of library e-book borrowing history data based on big data technology. *Web Intelligence and Agent Systems*, 18(2), pp. 111-120. https://doi.org/10.3233/WEB-200433

[3] Wang J, Zheng L, Alsulami H, Chen J (2022). Modeling and analysis of the book borrowing of students in the library using partial differential equations. *Fractals: An Interdisciplinary Journal on the Complex Geometry of Nature*, 30(2), pp. 1-11. https://doi.org/10.1142/S0218348X22400709

[4] Silwattananusarn T, Kulkanjanapiban P (2020). Mining and Analyzing Patron's Book-Loan Data and University Data to Understand Library Use Patterns. *International Journal of Information Science and Management*, 18(2), pp. 151-172. https://doi.org/10.48550/arXiv.2008.03545

[5] Iqbal N, Jamil F, Ahmad S, Kim D (2020). Toward Effective Planning and Management using Predictive Analytics based on Rental Book Data of Academic Libraries. *IEEE Access*, 8, pp. 81978-81996. https://doi.org/10.1109/ACCESS.2020.2990765

[6] Xie C (2020). Research on classification and identification of library based on artificial intelligence. *Journal of Intelligent and Fuzzy Systems*, 40(1), pp. 1-13. https://doi.org/10.3233/JIFS-189524

[7] Xie W (2021). Discussion on E-books and Library Borrowing Service. *Journal of Electronic Research and Application: JERA*, 5(4), pp. 4-7. https://doi.org/10.26689/jera.v5i4.2505

[8] Yang H, Zhang W (2022). Data mining in college student education management information system. *International Journal of Embedded Systems*, 15(3), pp. 279-287. https://doi.org/10.1504/IJES.2022.10049644

[9] Sabna E (2019). Pemanfaatan data mining untuk penempatan buku di perpustakaan menggunakan metode association rule. *Jurnal Ilmu Komputer*, 8(2), pp. 59-63. https://doi.org/10.33060/JIK/2019/Vol8.Iss2.127

[10] Astuti T, Anggraini L (2019). Analysis of Sequential Book Loan Data Pattern Using Generalized Sequential Pattern (GSP) Algorithm. (1). https://doi.org/10.47738/ijiis.v2i1.10

[11] Liu S, Liu X (2021). Research on Density-Based K-means Clustering Algorithm. *Journal of Physics: Conference Series*, 2137(1), pp. 1-7. https://doi.org/10.1088/1742-6596/2137/1/012071

[12] Yang Y, Cai J, Yang H, Zhao X (2022). Density clustering with divergence distance and automatic center selection. *Information Sciences: An International Journal*, 596, pp. 414-438. https://doi.org/10.1016/j.ins.2022.03.027

[13] Kurniawan E (2019). Implementasi data mining dalam analisa pola peminjaman buku di perpustakaan menggunakan metode association rule. *JURTEKSI*, 5(1), pp. 89-96. https://doi.org/10.33330/jurteksi.v5i1.324

[14] Astutik F, Kharismasari A, Laksono T, Santoso I, Chusyairi A (2019). E-Library Peminjaman dan Pengembalian Buku Berbasis Web dengan Metode Prototipe. *JTIM Jurnal Teknologi Informasi dan Multimedia*, 1(3), pp. 254-260. https://doi.org/10.35746/jtim.v1i3.45

[15] Din M M, Anwar R M, Fazal F A (2021). Asset tagging for library system - does QR relevant?. *Journal of Physics: Conference Series*, 1860(1), pp. 1-11. https://doi.org/10.1088/1742-6596/1860/1/012017