ISSN: 2472-3703 DOI: 10.6911/WSRJ.202310_9(10).0001

Clustering Methods of University Library Readers Based on Borrowing Behaviors and Their Applications

Jiajia Liu*

Library of Huaiyin Normal University, Huai'an, Jiangsu, 223001, China

*Corresponding Author

Abstract

Based on the Chinese Library Classification, this paper proposes a similarity measurement method for bibliographic content, and further constructs a similarity measurement method for readers' borrowing behaviors. The PAM algorithm of K-center clustering is used to cluster readers' borrowing data, and analysis and utilization methods for clustering results are proposed. The experimental results show that the above bibliographic similarity measurement method and reader borrowing behavior similarity measurement method are feasible and effective. By analyzing the clustering results, it is easy to explore the borrowing characteristics of reader groups and achieve accurate and personalized bibliographic recommendation services.

Keywords

Reader clustering; Personalized recommendation; Similarity measurement.

1. INTRODUCTION

With the continuous advancement of the modernization process of university libraries, the demand for personalized services is increasing, and the requirements are becoming more and more sophisticated. College students' needs for reading are diversified and personalized, and there have also been great changes in the way they read [1]. Different groups of readers have different borrowing behaviors and reading preferences [2]. Facing a large amount of reader borrowing data, librarians urgently need to find useful information from it to understand readers' borrowing needs and provide them with better quality personalized services. Personalized recommendation refers to recommending the most likely books needed for readers. To do this, we urgently need to tap the most likely needs of readers from known borrowing data. From the perspective of personalized recommendation algorithms, collaborative filtering is currently the most successful. The concept of collaborative filtering was proposed by Goldberg, Nichols, Oki and Terry in 1992 [3]. Collaborative filtering is a pure personalized recommendation method [4], which has limitations in targeted recommendations with group characteristics. The success rate of recommended bibliographies being accepted needs to be further improved.

This paper discusses the division method of reader groups based on readers' borrowing data, and uses clustering techniques to obtain natural groupings (groups) of readers. The borrowing behaviors of members within the group have a high degree of similarity. For individuals in the group, these borrowing behaviors have a group-oriented role. The natural boundaries between groups mean that there are certain differences in borrowing behaviors between different groups. The group divisions obtained by natural clustering can reflect readers' real borrowing needs better than the administrative divisions.

2. MEASUREMENT METHODS FOR READER SIMILARITY

When clustering reader borrowing data, it is necessary to measure the distance or similarity between two data. According to the requirements of the clustering algorithm, the similarity measurement method between any two readers must be determined first. The more similar the borrowing behaviors of two readers are, the higher the degree of similarity should be. Borrowing behaviors are reflected through borrowed bibliographies. Therefore, the similarity measurement of readers can be measured through the similarity between bibliographies.

2.1. Definition of bibliographic similarity based on Chinese Library Classification

In order to measure the similarity between the borrowing behaviors of two readers, the similarity between any two books must be determined first. The measurement of bibliographic similarity is essentially based on the similarity of book content. The content of the bibliography itself cannot be quantified, so it is difficult to directly calculate the similarity between the contents of two books. However, by classifying bibliographies, the approximate similarity between different types of bibliographies can be obtained. For this purpose, this paper designs a similarity measurement method to classify bibliographies using the Chinese Library Classification (hereinafter referred to as the CLC). By analyzing the similarity between different classes, the similarity between bibliographies can be determined. If two books are in the same bottom class, the similarity between the two bibliographies needs to be further determined. The inter-class bibliographic similarity and intra-class bibliographic similarity depend on the depth of the classification number directory in the CLC and the definition of inter-class similarity by experts [5].

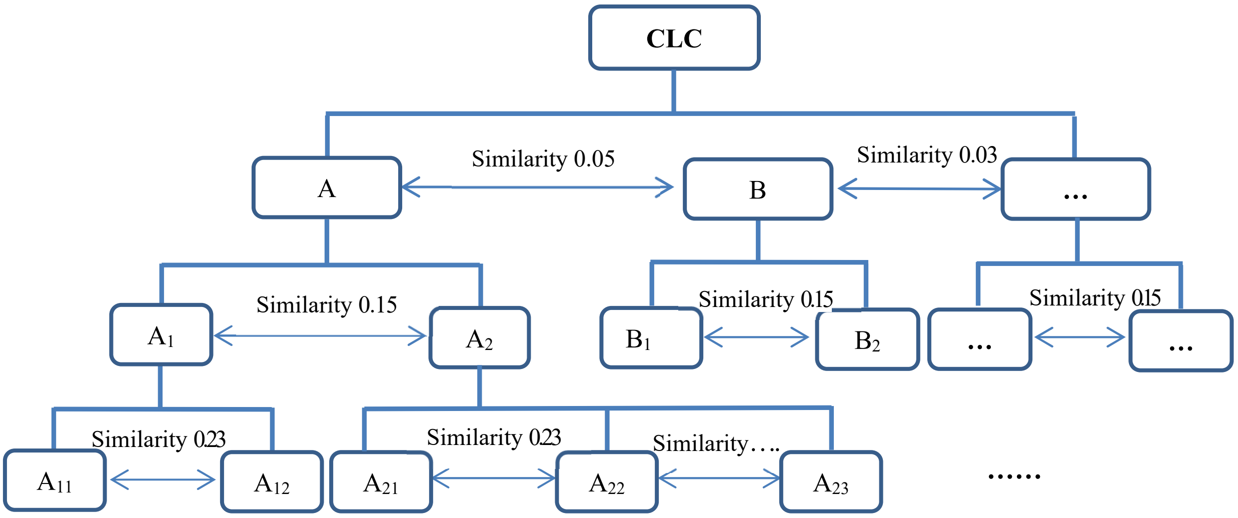

Figure 1. Example of bibliographic similarity definition based on Chinese Library Classification

The measurement of bibliographic similarity depends on the definition of bibliographic similarity. Bibliographic similarity is mainly determined from the following three aspects: First, if two bibliographies belong to two different classifications, the inter-class similarity between the two classes to which they belong, that is, the class interval similarity, must be determined first. The condition for the existence of similarity between two classes is that they have a common parent class, that is, they are divided from the same class. It can be seen from this that subclasses divided from different parent classes have no directly defined similarity. As shown in Figure 1, class A1 and class A2 have a common parent class A, so the similarity between A1 and A2 can be directly defined, with a value of 0.15. Classes A1 and B1 do not have a common parent class, so there is no directly defined similarity between them; Second, the defined inter-class similarity value is always greater than the inter-class similarity at the parent class level, that is, the inter-class similarity at the lower level is always higher than that at the higher level; Third, the incomplete similarity between different bibliographies or classes is always less than 1.

2.2. Measurement of bibliographic similarity

For convenience of description, b_n represents the bibliography, n represents the bibliography number, L_p represents the class divided by the Chinese Library Classification, DB represents the bibliographic similarity function, and DL represents the inter-class similarity function. When the parameter is 2, it represents the similarity between two classes. When the parameter is 1, it represents the intra-class similarity of the class. Using the above definition of bibliographic similarity, the similarity between any two bibliographies b_m and b_n can be determined according to the following three situations:

(1) b_m and b_n belong to the same bottom class L_p, and $b_m \neq b_n$, that is, b_m and b_n are not the same book. This can be expressed as $b_m \in L_p$ and $b_n \in L_p$ and $b_m \neq b_n$. Then the similarity between the two bibliographies is $DB(b_m, b_n) = DL(L_p)$;

(2) b_m and b_n are the same book, that is, $b_m = b_n$. In this case, it is not necessary to consider the similarity within the class bibliography. That is, $DB(b_m, b_n) = 1$;

(3) b_m and b_n belong to two different bottom classes. Their similarity is the similarity between the two classes they belong to. If the two belonging classes have no directly defined similarity, go up one level to find the inter-class similarity definition value between the parent classes of the belonging classes. Loop like this until the classes L_p and L_q with similarity definition values are found. The similarity between b_m and b_n can then be determined, that is, $DB(b_m, b_n) = DL(L_p,\ L_\mathrm{q})$.

2.3. Reader similarity measurement method

Taking the set of bibliographies borrowed by readers as borrowing behaviors, and on the basis of the above bibliographic similarity measurement, the similarity between the borrowing behaviors of two readers can be measured. The basic idea is to measure the similarity between the borrowing behaviors of any two readers, that is, to measure the similarity between the two sets of bibliographies borrowed by the two readers. Compare one bibliography borrowed by reader A with all the bibliographies borrowed by another reader B one by one, and add up the results to obtain the similarity between the single bibliography borrowed by reader A and the set of bibliographies borrowed by reader B. Add up the similarities between all the single bibliographies of reader A and the bibliography sets of reader B to obtain the similarity between A and B. This reader similarity measurement method is a content-based measurement method, which can better reflect the similarity of readers' borrowing behaviors.

Let reader A be ($a_1,a_2,\ldots,a_n$), that is, the set of bibliographies borrowed by A is { $a_1,a_2,\ldots,a_n$ }, where n>0. Let reader B be ($b_1,b_2,\ldots,b_k$), and the set of bibliographies borrowed is { $b_1,b_2,\ldots,b_k$ }, where k>0. Then the similarity S(A,B) between reader A and reader B can be calculated according to the following formula. There are two cases:

Case 1 (There is no duplicate between the bibliographies borrowed by reader A and reader B):

$$\mathrm{S(A,B)} = (DB(a_1, b_1) + DB(a_1, b_2) + \cdots + DB(a_1, b_k) +$$

$$DB(a_2, b_1) + DB(a_2, b_2) + \cdots + DB(a_2, b_k) +$$

$$\ldots +$$

$$DB(a_n, b_1) + DB(a_n, b_2) + \cdots + DB(a_n, b_k))/(n * k)$$

Considering that the number of bibliographies borrowed by different readers is different, but the importance of each reader is the same, the total amount of bibliographies borrowed by each reader should be given the same weight. If the number of bibliographies a reader borrows is n, then the weight occupied by each bibliography borrowed should be $1/n$. The similarity between two readers is the similarity between the two sets of bibliographies borrowed, which is the sum of the similarities between the borrowed bibliography sets/(n*k).

Case 2 (There are duplicate bibliographies borrowed by reader A and reader B):

Let the duplicate bibliographies be { t_1， t_2， ……， t_q}, where q>0. Before calculating the similarity between A and B, the bibliographies of the two need to be preprocessed by temporarily hiding the duplicate bibliographies. Duplicate bibliographies are not involved in similarity calculations, only the duplicate number q is involved in the calculation. After preprocessing, the bibliographies borrowed by reader A are { a_1,a_2,...,a_n }, n>0, and the bibliographies borrowed by reader B are { b_1,b_2,...,b_k}, k>0. The similarity should be calculated as follows:

$$\mathrm{S(A,B)} = (q + Db(a_1, b_1) + Db(a_1, b_2) + \cdots + Db(a_1, b_k) +$$

$$Db(a_2, b_1) + Db(a_2, b_2) + \cdots + Db(a_2, b_k) +$$

$$\ldots +$$

$$Db(a_n, b_1) + Db(a_n, b_2) + \cdots + Db(a_n, b_k))/(n * k + q)$$

When there are duplicate bibliographies borrowed by two readers, the duplicate bibliographies do not conform to the rules of inter-class similarity definition, so the related calculations cannot be performed as in the first case. The similarity of duplicate bibliographies is 1. According to the above formula, the more duplicate bibliographies there are, the greater the similarity between A and B. This conforms to the definition of similarity.

3. READER CLUSTERING USING PAM ALGORITHM

The K-means clustering algorithm and the K-center clustering algorithm[6] are typical partitioning-based clustering algorithms. As a kind of K-center clustering algorithm, the PAM algorithm has better tolerance for "noise" data and outlier data. There are often atypical "noise" or outlier data in reader borrowing data. Therefore, this paper chooses the PAM algorithm for clustering.

The basic process of reader clustering based on PAM[7]: First, randomly select k readers as the representative objects Ri of k clusters, that is, the center points. For the remaining readers Rh, assign them to the cluster represented by the representative object with the highest similarity to it according to the similarity between it and the representative object. Then, repeatedly replace the representative object with a non-representative object R_h, and use the replacement cost function C_{jih} to evaluate the cost of replacement. If it is a good replacement, make a formal replacement, otherwise do not replace. Finally, the optimal clustering results are obtained. Adding up the replacement costs C_{jih} of all n-k R_j with R_h replacing R_i yields the total replacement cost: $TC_{ih} = \sum_j C_{jih}$.

Algorithm: Reader Clustering Based on PAM

Input: Expected number of reader clusters k, database containing n readers and their borrowing data

Output: k reader clusters

Step 1: Randomly select k initial representative readers;

Step 2: Repeat the following steps

Step 3: Assign the remaining n-k readers to the cluster represented by the representative reader with the highest similarity;

Step 4: For the representative reader R_i, arbitrarily select a non-representative reader R_h;

Step 5: Calculate the total replacement cost TC_{ih} to replace R_i with R_h;

Step 6: If TC_{ih}<0, replace R_i with R_h to generate new k representative readers;

Step 7: Repeat until the formed k reader clusters no longer change.

4. EXPERIMENTS AND RESULTS ANALYSIS

The experimental data uses the borrowing data of the 2022 undergraduate students of a university, with a total sample size of 5600. The natural grouping is divided into 18 colleges by college and 72 majors by major. A fragment of the statistical table is shown in Table 1.

Table 1. Fragment of reader classification statistics by college and major

College	Major	Number of People	
College of Liberal Arts	Chinese Language and Literature	173	418
	Chinese Language and Literature(Normal)	144	
	Secretarial Studies	55	
	Teaching Chinese as a Foreign Language	46	
College of Foreign Languages	English Language Studies	87	313
	English (Normal)	96	
	Translation	39	
	French Language Studies	31	
	Japanese Language Studies	60	
…			
College of Chemistry	Chemistry(Normal)	34	285
	Applied Chemistry	104	
	Environmental Science	45	
	Engineering and Technology	102	

The reader borrowing data sample is shown in Table 2. The bibliography information only lists the titles. Other related fields are not listed one by one as examples. n is the maximum borrowing volume of readers in the sample.

Due to many obvious noise in the reader borrowing data, such as reader information without borrowing behavior, pre-processing is performed as needed before input to obtain relatively neat and usable data as much as possible. Taking the borrowing data of the 2022 undergraduate students in the 2022-2023 school year after preprocessing as input, and after experimental verification and adjustment, the clustering quality is better when the initial k value is set to 21. A fragment of the clustering results is shown in Table 3.

Table 2. Reader borrowing data structure example

Name	Student ID	College	Major	Total Borrowed Books	Bibliography 1	Bibliography 2	Bibliography 3	...	Bibliography n
Yao Zhang	0122110579	College of Liberal Arts	Chinese Language and Literature	n	The Palace of Eternal Life	Dragon Totem	Countryside in August	...	The Story of a Swing
Minrui Yu	0422110590	College of History	Tourism Management	12	Global History: Civilizations of Humankind	Toast of the Office	Currency Wars	...	null
Wenjing Wang	1722110815	College of Physics and Electrical Engineering	Electronic and Information Engineering	7	Calculus	I am a Cat	Ordinary World	...	null
...									
Ting Zhang	2122170568	College of Media and Communication	Radio and Television Editing and Directing	23	Listening to Children's Scribbles	The World Art History	Tibetan Art	...	null

Table 3. Fragment of clustering results example

Cluster Number	Number of Readers
Cluster 1	175
Cluster 2	186
Cluster 3	286
...	...
Cluster 21	232

Select the representative cluster 7, which is characterized by a large number of borrowings of philosophy books and obvious borrowing tendencies. Its detailed data fragment is shown in Table 4. The maximum borrowing volume of readers in the sample is 207.

Table 4. Data fragment of cluster 7 example

Name	Student ID	College	Major	Total Borrowed Books	Bibliography 1	Bibliography 2	Bibliography 3	...	Bibliography 207
Wu Zhibiao	0122110761	College of Liberal Arts	Chinese Language and Literature	33	Laziness	Your Loneliness, Though Defeated, Is Glorious	Orphan of the Fog City	...	null
Gang Cao	0922110496	College of Music	Musicology	46	General Psychology	Being the Best You	Beethoven Piano Sonatas	...	null
Xu Huang	0222110335	College of Foreign Languages	English	29	Language Theory	On Beauty	The Art of Loving	...	null
Junsheng Zhao	0222110237	College of Foreign Languages	Translation	207	Proper Self-Esteem	College Students, What's Wrong with You	Jane Eyre	...	Hello, Old Times
Qing Zhou	0922110705	College of Music	Dance	26	Lolita	Philosophy and Life	Personality Psychology	...	null
Yuqian Yao	0222110630	College of Foreign Languages	English	71	The Catcher in the Rye	Psychology of Women's Success	Introduction to Psychology: The Road to Understanding Thought and Behavior	...	null
......									
Yufan Kong	1622110682	College of Mathematics and Sciences	Statistics	13	The Republic	The Diamond Sutra	Mathematical Analysis	...	null
Yihang Xu	0922130382	College of Music	Music Performance	42	Bible Stories	The Little Fisher Girl	Interpretation of Dreams	...	null

Experimental data can be screened and adjusted as appropriate to obtain more refined or macroscopic results. For example, if the number of people in a certain college is large, then all the readers of the school do not need to be used as input. Instead, the reader data of the college can be used as input, the initial center k value can be adjusted and optimized, and the above algorithm can be executed.

Comparing the experimental results with the reader data divided by college or major shows that the reader groups divided by clustering based on reader borrowing behaviors can better reflect students' real borrowing intentions. For example, in clusters with higher relevance to philosophy, literature and other books, there are students of science and engineering majors. This shows that these students have inherent needs for philosophy, literature and other aspects. Using reader clustering methods based on reader borrowing behaviors, on the one hand, decision-makers can master students' borrowing tendencies according to the clustering results, understand their real needs, and guide these needs; on the other hand, library service providers can recommend bibliographies with higher borrowing volumes in the reader clusters where these students are located to meet their borrowing needs. It has been verified that recommending bibliographies with higher ranking borrowing volumes in the cluster to students has a higher acceptance rate, which also shows that the clustering method used in this paper is feasible and effective.

5. CONCLUSION

Reader clustering based on borrowing data is a kind of reader group divided naturally based on borrowing behaviors, which is different from administrative divisions. This kind of group has more accurate reader reading preferences, there are obvious barriers between different groups, and there are obvious group borrowing characteristics. According to the group borrowing characteristics, group members can be labeled with group labels, so as to recommend more accurate bibliographies to readers with the same label. Using the methods provided in this paper, the demand analysis of reader groups and personalized bibliography recommendation can be realized, which is of great significance for improving the quality of personalized services in libraries.

6. ETHICAL APPROVAL

This article does not contain any studies with animals performed by any of the authors.

7. ETHICAL APPROVAL

This article does not contain any studies with human participants or animals performed by any of the authors.

8. FUNDING

This study is funded by the general research project of philosophy and Social Sciences foundation of Jiangsu Province(Grant No. 2018SJA1603).

9. CONFLICT OF INTEREST

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

10. DATA AVAILABILITY STATEMENT

No data were used to support this study.

REFERENCES

[1] Xiao H, Zhu H. Construction of precise reading promotion model of university library from the perspective of participatory user profile [J]. Library Work and Study, 2020(06): 122-128.

[2] Shi G, Zhang X, Yang X. Research on the influence of the differences of university readers' groups on borrowing behavior and reading preference [J]. Library, 2020(04): 59-64+78.

[3] Goldberg D, Nichols D, Oki BM, et al. Using collaborative filtering to weave an information tapestry [J]. Communications of the ACM, 1992, 35(12): 61-70.

[4] Song C. Application of an improved collaborative filtering method on recommending books in college libraries [J]. Library and Information Service, 2016, 60(24): 86-91.

[5] Zhang Y. Construction of a discipline team based on PAM clustering [J]. Library Science Research & Work, 2020(06): 48-53+71.

[6] Han J, Kamber M. Data Mining: Concepts and Techniques [M]. Beijing: Metallurgical Industry Press, 2001.

[7] Chen Z, Liu Z, Zhang J. Analysis and implementation of PAM algorithm [J]. Computer and Modernization, 2003.