

Library Collection Document Lending System Based on The Statistical Equation of Probability Theory

Hui Wang¹, Jing Sheng², Guosheng Chen³, Shipeng Yang^{4,*}, Xueleli Liu⁵

¹ Library, Hunan Institute of Technology, Hengyang, China

² School of Computer Information Engineering, Yongzhou Vocational and Technical College, Yongzhou, China

³ School of Business, Hunan Institute of Technology, Hengyang, China

⁴ School of Public Administration, Xiangtan University, Xiangtan, China

⁵ School of Economics and Management, Hunan Institute of Traffic Engineering, Hengyang, China

* Corresponding Author Email: 201821100500@smail.xtu.edu.cn

Abstract. This paper establishes a library-collected document lending system based on data mining technology. This paper uses the statistical method of probability theory of association rules between multi-dimensional attributes to mine the regularity of document usage. Through this method, corresponding decisions can be made to optimize the layout of the library's collections. This paper uses data mining methods such as cluster analysis and association analysis to establish a data mining processing model. This method records the retrieval behavior of teachers and students. In this way, the interest points of teachers and students' demand books can be summarized. Finally, according to the above scheme, a simple experimental system is implemented, and the experimental results are verified and tested. The research finds that this paper's bibliography recommended by the book recommendation system using data mining technology is consistent with the borrowed bibliography. Compared with the traditional method, the judgment accuracy of the algorithm is greatly improved.

Keywords: Probability theory, Statistical method, Library literature, Book lending system, Data mining.

1. Introduction

The continuous development of colleges and universities and the continuous increase in the number of students have led to a geometric increase in library resources. At present, solving the problem of information overload faced by readers has become a necessary standard to measure the quality of personalized service in university libraries. The traditional book viewing efficiency is low, and the utilization rate of library resources is not high [1]. The importance of accurately capturing the needs of different readers and providing them with high-quality, targeted, and personalized book recommendation services is self-evident. The customized book recommendation is an essential means to optimize the service level of university libraries in the new era. This method is beneficial to stimulate the reading interest of teachers and students in colleges and universities and can improve the utilization rate of book resources [2]. This paper is based on the four-year book borrowing log of college students. We conduct data mining on borrowing logs through association rules and Bayesian network algorithms. In this way, personalized book recommendations can be realized, and the service level of the library can be improved.

At present, most of the hot research is through the improvement of various recommendation algorithms. These methods aim to improve the accuracy of the recommendation and improve the quality of the request. However, the personalized semantics for different users cannot be fully reflected, and the research on mining the category relationship between model data attributes is not sufficient. This paper uses borrowing time as the primary dimension to conduct data mining through the association rule algorithm [3]. We use the Bayesian network to perform semantic verification on association rule mining results. This can improve the accuracy and reliability of the recommendation

results. This method conducts an in-depth analysis of linked data and recommendation results. Finally, this paper proposes the corresponding personalized recommendation strategy. This can improve the level of library recommendation services.

2. Research Process

2.1. Data acquisition and processing

This paper selects 1,000 readers from the 2018 undergraduates (with a four-year schooling period) through stratified sampling. We extracted the borrowing log data of the sampled readers from September 2018 to June 2021 through the book borrowing management system. This sample of readers covers 17 majors, including medical testing technology, medical imaging technology, rehabilitation therapy, pharmacy, nursing, and social work. A total of 29,211 borrowing log data were obtained this time [4]. We select the following fields for the sampled sample: reader barcode, call number, title, reader unit, borrowing grade (operation date), reader major, etc. We extract the required fields as shown in Table 1.

Table.1. Association rule mining results

Program number	Program
1	R47, H31, D92, B84
2	TP3, R47, R71, D92
3	TP3, R47, R71, B84
4	TP3, R71, B84, D92
5	H31, R47, R71, D92
6	R47, R71, D92, R192
7	TP3, H31, R71, D92
8	TP3, R47, B84, D92
9	H31, R47, R71, B84
10	H31, R91, B84, D92
11	TP3, R71, D92, R192
12	TP3, H31, R47, R71
13	TP3, H31, R71, B84
14	TP3, R47, R71, R192
15	R47, R71, B84, R192
16	R71, B84, D92, R192
17	TP3, H31, R47, D92

2.2. Apriori algorithm and Bayesian network

The Apriori algorithm is an algorithm for mining data association rules. We extract frequently occurring data from data values to form a dataset and further analyze it. Multi-dimensional association rules need to examine the number of readers, book types, majors, and grades in the reader's loan log. The article mines the association rules between these data. The goal of the Apriori algorithm is to find the most extensive frequent set of K items. The algorithm can be divided into 2 steps. First, we iteratively retrieve the recurring settings that meet the support criteria in the database. The goal is an item set whose support is not lower than a set threshold. Then we construct a rule that satisfies the minimum trust degree of users through frequent itemsets. The specific steps of the Apriori algorithm are as follows: We first retrieve a candidate item set and record it through the iterative idea. Next, we find its corresponding support [5].

We prune and remove the 1 item set lower than the support degree to get the frequent 1 item set and record it as C1. Then we concatenate the remaining frequent itemsets to get the candidate's regular two-item sets. At this point, we call it L2. We filter out the candidate frequent two itemsets lower than the support degree to get the true frequent two itemsets. Let's call it C2. And so on and so on

until no frequent $k+1$ itemsets can be found. The set of corresponding regular k itemsets is the output result of the algorithm. The location of the results we get satisfies the minimum confidence. This is also the result of definite association rules. We can mine helpful information and directions from historical data.

The ideological basis of the Bayesian network is as follows: we solve the probability of each category appearing under the condition that the item to be classified seems [6]. The formal definition of Bayesian network classification is as follows:

(1) Suppose $x = \{a_1, a_2, a_3, \dots, a_m\}$ is an item to be classified, and each a is a feature attribute of x .

(2) There is a definite set denoted as $C = \{y_1, y_2, \dots, y_n\}$.

(3) We find a set of known classification items to be classified (training set). At this point, we statistically obtain the conditional probability estimates of each feature attribute under each category.

$$p(a_1 | y_1), p(a_2 | y_1), \dots, p(a_m | y_1);$$

$$p(a_1 | y_2), p(a_2 | y_2), \dots, p(a_m | y_2);$$

$$p(a_1 | y_n), p(a_2 | y_n), \dots, p(a_m | y_n);$$

If each feature attribute is conditionally independent, the following derivation can be obtained according to Bayes' theorem:

$$p(y_i | x) = \frac{p(x | y_i)p(y_i)}{p(x)} \quad (1)$$

(4) Each category is calculated as follows:

$$p(x | y_i)p(y_i) = p(a_1 | y_i)p(a_2 | y_i) \cdots p(a_m | y_i)p(y_i) = p(y_i) \prod_{j=1}^m p(a_j | y_i) \quad (2)$$

(5) We take the most crucial term of $P(x | y_i)P(y_i)$ as the category to which x belongs.

2.3. Personalized Recommendation Model Based on Association Rules and Bayesian Networks

We build a personalized recommendation model through association rule mining algorithms, historical data pruning, and Bayesian network correction (Figure.1.). The model makes the original database based on the reader's 4-year borrowing log data. We preprocess the data and input it into the Apriori algorithm. This also acts as an association rule mining module. We combine the association rule mining results into an association rule set. We prune it with historical data. We will prune data below the average. At this point, an optimized association rule set is obtained. We use the Bayesian network standard library to perform semantic verification through the Bayes algorithm [7]. We sort the associated itemsets by probability. We analyze the books borrowed by readers based on association rule mining. This paper obtains the output results through Bayesian network optimization, and a personalized recommendation strategy is formulated.

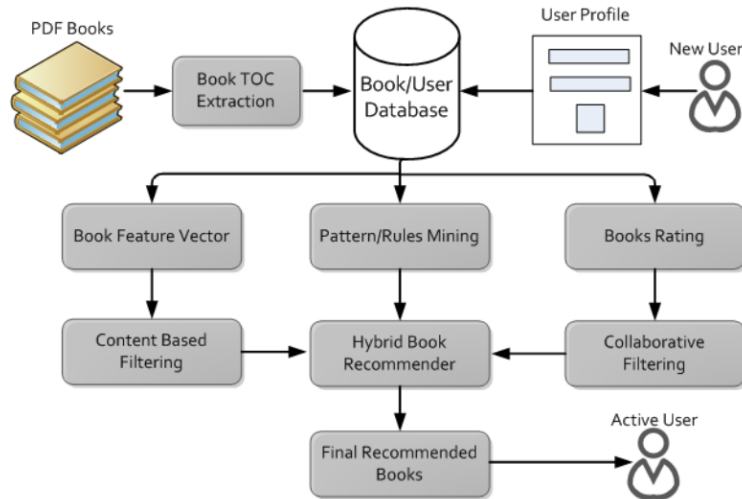


Figure.1. Personalized model recommendation

2.4. Book Personalized Association Recommendation Algorithm

D_1 is the borrowed data for the preprocessed readers. The algorithm mines association rules from D_2 . We prune the frequent itemsets in the mined association rule set through historical data. In this paper, the Bayesian network verifies the pruned association rule set. At this point, we get the optimized set of association rules. In this way, a personalized book recommendation scheme is developed [8]. Where Algorithm 1 will call Algorithm 2.

Algorithm 1: Bayesian personalized association rule recommendation algorithm

Input: D_1

output: S_b

Begin:

- 1) This article obtains the borrowing data of readers after preprocessing and denote it as D_1 .
 - 2) According to the *Apriori* algorithm, we set the support degree *Support_degree* and generate the k itemset. We obtain frequent itemsets $S_i, S_x \in S_j, k \in [1, n]$ to be recommended to readers through *Apriori* algorithm, n is the number of items.
 - 3) User *user* historical data is recorded as *Huser*. We combine *Huser* and call *Prune*($S_j, Huser$). We perform personalized semantic pruning on S_j and output $S_R = \{S_r | S_r \in S_i\}$.
 - 4) If S_R satisfies the set k item set requirements, then go to step.
 - 5) Otherwise, go to step 2) until the k itemset requirements are met.
 - 6) We construct a borrowing record database D_2 that reflects readers' interests.
 - 7) We take S_R as the input to the *Bayes* algorithm. We run the *Bayes* algorithm to output $S_b, S_b = \{S_y | S_y \in S_i\}$ sorted by reader interest.
 - 8) Return (S_b); S_b is the recommended solution.
- End
- Among them N_4 is the selected threshold.

2.5. Evaluation of Recommendation Algorithms

Generally, we can evaluate the recommendation results according to the precision rate and recall rate. Recall rate refers to the proportion of recommended books that meet the reader's interest in the reader's interest set. The accuracy rate refers to the ratio of recommended books that meet readers' interests in the total recommended book collection. The calculation formulas of recall rate and precision rate are shown in formula (3), formula (4), and formula (5).

Recall rate:

$$R_e = \sum_{i=1}^u \frac{L_i}{M \times P_i} \quad (3)$$

Accuracy:

$$P_r = \frac{\sum_{i=1}^M L_i}{M \times N} \quad (4)$$

The harmonic means of the two:

$$F = \frac{2 \times R_e \times P_r}{R_e + P_r} \quad (5)$$

Among them, P_i is the total number of books that represent the reader's interest. L_i indicates recommended books that match the reader's interests. M is the total number of readers. N means the total number of recommended books. When the F value is higher, the recommendation effect is better.

3. Analysis and discussion of experimental results

We use 29211 items as the original borrowing data for data mining. (1) We perform association rule mining on the borrowing database. We select a reader and prune it with his personal history of borrowing books [10]. We analyzed borrowing records from a single study subject. This paper uses the Apriori algorithm. Set the support to 0.05 to get frequent item sets (Table.1.), and we use the Bayesian network to perform personalized semantic verification on the itemsets to get the results (Table. 2.). By comparing Table 1 and Table 2, we can find that the results mined by the association rules in Table 1 do not meet personalization requirements. Based on Table 1, this paper uses the Bayesian network to perform personalized semantic verification to obtain the probability ranking of priority recommendations. In this way, a personalized book recommendation for different readers is realized. To get customized book recommendation results, we selected nursing readers with student number 3130608019.

The top 5 book sets recommended by this algorithm are R47, R71, B84, R192; R71, B84, D92, R192; R47, R71, D92, R192; R47, H31, D92, B84; TP3, H31, R71, D92. Therefore, the priority order of personalized recommended book types is {R47, R71, B84, R192, D92, TP3}. We demonstrate the effectiveness and authenticity of customized recommendations for a single reader by analyzing the reader's visit [11]. The types of books that the reader is interested in are R47 (Nursing), R71 (Obstetrics and Gynecology), B84 (Psychology), R192 (Nurse Vocational Qualification Examination), D92 (Medicine and Law), TP3 (Computer Science) grade exams). The results are in line with the borrowing rules of the reader's nursing major.

The algorithm of this paper is used to mine and analyze 29211 loan log data. The thresholds of association rules are respectively set as Rule 1: 0.0345; Rule 2: 0.0462; Rule 3: 0.0511; Rule 4: 0.2876; Rule 5: 0.0786; Rule 6: 0.0899; This paper uses book recommendation accuracy (Pr), recall (Re) and harmonic mean F as the test indicators. The larger the F value, the better the recommendation effect [12]. We mine and analyze the borrowing log data through different recommendation algorithms. The experimental results are shown in Figure.2. and Table.3.

Table.2. Bayesian personalized semantic verification results

New recommended serial number	Original item set serial number	Program	Priority
1	15	R47, R71, B84, R192	0.02151 1
2	16	R71, B84, D92, R192	0.02154 5
3	6	R47, R71, D92, R192	0.02016 8
4	1	R47, H31, D92, B84	0.01960 2
5	7	TP3, H31, R71, D92	0.01861 1
6	5	H31, R47, R71, D92	0.01816 8
7	11	TP3, R71, D92, R192	0.01802 1
8	14	TP3, R47, R71, R192	0.01890 4
9	12	TP3, H31, R47, R71	0.01865 8
10	17	TP3, H31, R47, D92	0.01852 1
11	10	H31, R91, B84, D92	0.01816 6
12	9	H31, R47, R71, B84	0.01659 1
13	3	TP3, R47, R71, B84	0.01611 1
14	13	TP3, H31, R71, B84	0.01408 8
15	8	TP3, R47, B84, D92	0.01268 8
16	2	TP3, R47, R71, D92	0.00885 6
17	4	TP3, R71, B84, D92	0.00685 5

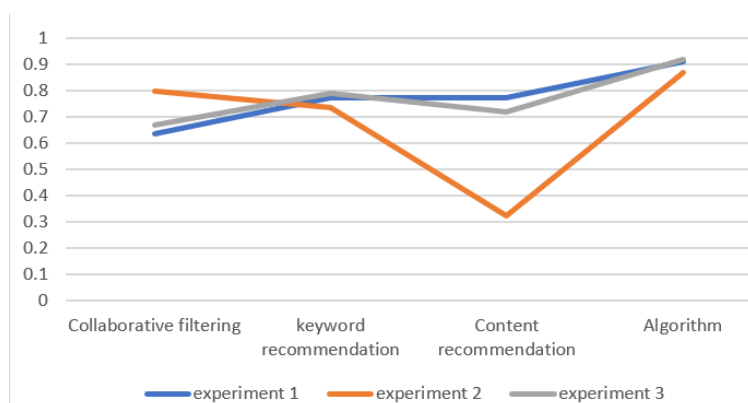


Figure. 2. Test comparison results of recommendation algorithms

It can be seen from Figure 2 that the book recommendation accuracy and recall rate of the algorithm in this paper are higher than other recommendation algorithms. This shows that the personalized recommendation of the algorithm in this paper is highly reliable and accurate.

Table.3. Harmonic mean of different recommendation models in three groups of experiments

Recommendation algorithm	Experiment 1	Experiment 2	Experiment 3
Collaborative filtering	0.633	0.798	0.668
Keyword recommendation	0.773	0.737	0.789
Content recommendation	0.773	0.321	0.717
Algorithm	0.912	0.869	0.919

The higher the harmonic mean F, the higher the recommendation accuracy. In this paper, the F value of book recommendations using this method is higher than other algorithms. Book recommendation accuracy is high, and reliability is good. This shows the superiority of this method.

4. Conclusion

This paper is based on the library book borrowing log data. This paper obtains a book personalized recommendation model by combining association rule mining and Bayesian network verification. The model can filter out the books that readers are interested in. Compared with other recommendation algorithms, this algorithm has high accuracy, high reliability, and a good recommendation effect. We can grasp and understand the interests and needs of readers of different grades and different majors in the data mining and analysis of borrowing logs. Moreover, this paper can provide book recommendation services according to the school's discipline construction and professional structure. This is more in line with the requirements of personalized book recommendations.

Acknowledgements

The authors gratefully acknowledge the financial support from General scientific research project of Hunan Education Department "Influence Comprehensive Evaluation of Economic management academic journals based on TOPSIS-Weight Method in the Context of New Media" (No. 19C0535); General project of Education and Teaching reform in Hunan Province, "Research on the Design and Practice of Ideological and Political Teaching of < management > Course under the Background of" Golden Course "Construction" (No. HNJC-2021-1280); Projects generally supported by Hunan Provincial Social Science Achievement Appraisal Committee "Research on the Construction of Evaluation Index System of Innovation and Entrepreneurship Education in Higher Vocational Colleges in Hunan Province Based on AHP Model" (No. XSP22YBZ150).

References

- [1] Husain T, Hidayati N. The Optimize of Association Rule Method for The Best Book Placement Patterns in Library: A Monthly Trial [J]. TEKNOKOM, 2021, 4(2): 53-59.
- [2] Karthikeyan R, Satheesbabu S, Gokulakrishnan P. Machine learning based student performance analysis system [J]. Information Technology in Industry, 2021, 9(1): 1174-1181.
- [3] Wang Z, Jiang C, Zhao H, et al. Mining semantic soft factors for credit risk evaluation in peer-to-peer lending[J]. Journal of Management Information Systems, 2020, 37(1): 282-308.
- [4] Hamad F, Al-Aamr R, Jabbar S A, et al. Business intelligence in academic libraries in Jordan: Opportunities and challenges [J]. IFLA journal, 2021, 47(1): 37-50.
- [5] Law K S, Chung F L. Knowledge-driven decision analytics for commercial banking [J]. Journal of Management Analytics, 2020, 7(2): 209-230.

- [6] Nikitha G N, Chandana C, Neelashree N, et al. Bank customer complaints analysis using natural language processing and data mining [J]. International Journal of Progressive Research in Science and Engineering, 2020, 1(3): 22-25.
- [7] Daugherty A L. Improving the licensing workflow at a major research university library [J]. Technical Services Quarterly, 2020, 37(3): 239-255.
- [8] Yang Y T, Shieh J C. Is there the Pareto principle in public library circulation? A case study of one public library in Taiwan[J]. Malaysian Journal of Library & Information Science, 2019, 24(2): 97-113.
- [9] Gençoğlu M T, Agarwal P. Use of quantum differential equations in sonic processes [J]. Applied Mathematics and Nonlinear Sciences, 2021, 6(1): 21-28.
- [10] Şahin R, Yağcı O. Fractional calculus of the extended hypergeometric function [J]. Applied Mathematics and Nonlinear Sciences, 2020, 5(1): 369-384.
- [11] Li S, Jiao F, Zhang Y, et al. Problems and changes in digital libraries in the age of big data from the perspective of user services [J]. The Journal of Academic Librarianship, 2019, 45(1): 22-30.
- [12] Wicaksono E A, Pakereng M A I. Implementation of Laravel Framework in The Development of Library Information System (Study Case: Smk Pgri 2 Salatiga) [J]. Jurnal Pilar Nusa Mandiri, 2020, 16(2): 261-270.