Personalized Recommendation Algorithm for books and its implementation

Personalized Recommendation Algorithm for books and its implementation

Li Chun-mei[1], Ma yi-han[1], Pi Wei[1], Qi Yan[2], Jiang Jie-teng[1], Dong Shuo[1]

[1]Department of Computer Technology and Application, Qinghai University, Xining 810016, China

[2]Institute of Information Technology, SuZhou Top Institute Of Information Technology, KunShan, 215311, china

Abstract: In recent years, with the completion of the new library of Qinghai University, the collection of books in the library has greatly increased. The library has a total collection of 880,000 volumes, covering a dozen disciplines such as science, engineering, agriculture, literature, history, economics, philosophy, law, education, management and medicine. It is difficult for users to find the books they are interested in among the numerous materials. Based on the actual situation of the library of Qinghai University, the differences of different professional users and their personal interests, this paper chooses the item-based collaborative filtering algorithm to realize personalized recommendation. First of all, in the calculation of book similarity, the traditional user score data is not chosen to calculate the similarity, but to calculate the similarity between books and books according to the feature vector of book name. Secondly, in order to avoid the problem of cold start, the system recommends the users who have no borrowing record, but the most borrowed books in their department. The combination of the two realized the personalized recommendation of books. By comparing with other traditional recommendation algorithms, it is found that the algorithm adopted in this paper has better recommendation effect.

1. The introduction

With the rapid development of today's Internet technology, the amount of data is also increasing. People increasingly feel helpless in the face of massive data. In order to counter the problem of information overload, recommendation systems (sometimes called recommendation engines) are proposed and search engines are created at the same time. The two share the same goal of addressing information overload, but the specifics vary from person to person. Recommendation engines are more inclined to people without a clear purpose, or their purpose is fuzzy. The purpose of the recommendation system through the history of user behavior or the user interest preferences or demographic characteristics of the user to give recommendation algorithm, then the recommendation system using the recommended algorithm to generate the user might be interested in a list of items.

The research and application of collaborative filtering and other recommendation algorithms are reported as follows. Sarwar et al.[1] proposed an item-based collaborative filtering recommendation algorithm based on the deficiencies of traditional recommendation algorithms, and introduced the method to measure Item similarity. They concluded that the item-based collaborative filtering algorithm had better performance and recommendation effect than the user-based collaborative filtering algorithm. Zhang Weiguang et al.[2] wrote a review of collaborative filtering recommendation algorithm, summarized the research status of collaborative filtering recommendation algorithm at home and abroad,

and analyzed the key technologies and existing problems in detail. In this paper, the sparse problem, cold start problem and recommended speed of user-project matrix are discussed. This paper is of great help to the study and research of collaborative filtering algorithm. The first system that applied collaborative filtering algorithm was Grundy[3], which could establish users' interest model and recommend relevant books to users according to the model. Collaborative filtering algorithm is now one of the most successful application in the world is the amazon online website. G Linden[4] and others according to the actual situation of the amazon website, useing the collaborative filtering algorithm based on item, by comparing the similar items rather than the more similar to the user which is better applicable to the item number and much greater than the actual situation of the number of users, to produce high quality recommendations.

With the establishment of the new Qinghai University Library, the collection of books in the library has been increasing. The total collection of the university library has reached 880,000 volumes, covering more than ten disciplines such as science, engineering, agriculture, literature, history, economics, philosophy, law, education, management and medicine. Although the library staff organized the books systematically and arranged them in an orderly manner, it took a lot of time for users to find the books which they were interested in among the numerous materials due to the large number and variety of books. Especially with the completion of the online lending platform of Qinghai University Library, the design of the personalized recommendation system matching it is extremely urgent. At present, the method of finding books is basically that users input the subjects they are interested in, such as mathematics, searching for some books related to this subject through keywords, and then finding the books they are interested in from the search results for reading. Or go directly to the library bookshelf, find the corresponding classification number of bookcase, manually search book by book. Find books that interest you. This manual method is time-consuming, especially when the volume of books is increasing, and can be even less efficient.

This paper aims at the shortcomings of the traditional method of finding books. According to the basic information of users, it can calculate the books that users may be interested in and recommend them to users, greatly reducing the unnecessary time to find books. It can also calculate the books with high similarity according to the user's reading habits and reading history, and recommend them to the user. In addition, users can also give feedback on the recommendation results. If users are not satisfied with the recommended books, the system will optimize the recommendation results according to the user feedback, so as to better meet the needs of users.

2. Data Processing

2.1 Data cleaning

2.1.1 The initial data

In order to ensure the authenticity and integrity of the data, the data source of this study was directly exported from the background of the library. After analysis and collation, 5 CSV type files were generated, namely, reader library, classified subject table, bibliography of books, circulation library 1, and circulation Library 2. Among them, the information stored in the reader library file is the reader's information, and we chose the student information. There is a total of 90,445 pieces of student information, each piece of information contains at least 8 attributes, and more than 10 attributes. Among them more important attribute is student number, name, grade, major.

The classified thesaurus files contain the classified information of books, which is a large bibliographic indexing reference book compiled on the basis of Chinese Library Method and Chinese Thesaurus to realize the integrated indexing of classified subjects and improve the retrieval efficiency. There are two main attributes in this table: the category number and the corresponding subject word. This number can be found in the index information of the book, from which we can get the subject of the book.

A bibliography is a document that holds information about books in a library. This file contains a lot

of data, a total of more than 680,000 book data, each book data has dozens of attributes, and the more important attributes are the primary key code, title and call number.

Circulation Library 1 and Circulation Library 2 store users' borrowing data, which together contain more than 12,500 data. Each piece of data also has 13 attributes, among which the more important two attributes are reader barcode and primary key code. Combine them with bibliographies and reader information to find out which book was borrowed by which reader.

2.1.2 Data cleaning processing

Although there are only five data files, the amount of data is very large and complex. Firstly, several important attributes are selected according to the requirements of the algorithm, which are student number, name, major, primary key code of the borrowed book, name and call number of the borrowed book.

Since the algorithm in this study is mainly implemented in Python, the process of data cleaning is also implemented in Python language.

First read the information in the reader's library and store it in a dictionary. The primary key of the dictionary is the reader's student number, and the value is the reader's information.

Next, read the bibliography. Because of the large size of the data files in the bibliographies, it is time-consuming to process. Since there is Tibetan language in the catalogue of the books in the collection, the results of direct copying and reading cannot identify those Tibetan languages, which will affect the subsequent results. Therefore, we choose to deal with the bibliography directly, that is, we put the primary key of the dictionary in the catalog and the primary key code of the book in the same dictionary.

2.1.3 Data after cleaning

Through the operation of the previous step, we have obtained the relatively clean user borrowing data and book name data required by the algorithm. User borrowing data are shown in Figure 1 below.

To illustrate its meaning with the first line of data: YKT11172 represents the reader code, unique. Zhang Qinwen represents the name of the reader. The College of Agriculture and Animal Husbandry represents the department of the reader. Veterinary pathology represents the name of the book borrowed by the reader, and s852.3/162 represents the call number of the book borrowed. So that's what the data represents when it's finally cleaned out.

The data needed for the subsequent algorithm implementation and the data needed for the test are based on these two data.

YKT11172, Zhang Qinwen, The College of Agriculture and Animal Husbandry, Veterinary pathology, S852.3/162
2018990009, Li Yanggui, The Department of Computer Science, Parallel algorithm practice, TP301.6/283
YKT20959, Yang Mingwei, Library, Huangpu's Godfather Sun Yat-sen, K827/532
YKT20959, Yang Mingwei, Library, Ten generals of the Republic of China, K825.2/225
1610405112, Pang Xun, The College of Chemical Engineering, An introduction to mathematical thought,01-0/176

Figure1 User borrowing data

3. Algorithm Selection

At present, due to the larger and larger scale of online library, there are many different methods about the algorithm of personalized book recommendation. Some scholars use association rule analysis to make personalized book recommendation[5] . Some scholars use machine learning to make personalized recommendations for books; Some scholars use a variety of algorithms together to do personalized book recommendation. In short, the current academic research on the personalized recommendation of books has been more mature. In this paper, a popular and mature collaborative filtering algorithm is selected to carry out personalized book recommendation.

Collaborative filtering algorithm is the core idea based on user history data to recommend[6]. For example student A which is majored in Computer science likes to read *the C programming language, Java from entry to master* this kind of books, the system will guess computer student B will also like to

read books of this type, so he will be recommended some books which Student A has read. It is now often said as the collaborative filtering algorithm based on the user.

Later, after extensive application and research, it was found that this user-based algorithm was not very accurate in some cases, so some people proposed the item-based collaborative filtering algorithm. The main idea is to make recommendations based on similarity between items and users' historical data. For example, if User A is interested in books like *Introduction to Algorithmic Competition Classics* and *The Art of Computer Programming*, the system will assume that he is also interested in the same type of *Challenge Programming Competition* and *Fundamentals of Algorithm Design and Analysis*, and will recommend these books to User A.

This item-based collaborative filtering algorithm is effective for systems where the number of items far exceeds the number of users. This paper analyzes the obtained data and finds that the number of books is more than 300,000, and the number of users is only tens of thousands. Moreover, the number of users with record of borrowing books is less than 10,000, and the number of books is ten times that of users. Therefore, this paper finally chooses the item-based collaborative filtering algorithm as the algorithm of this study.

4. Calculate book similarity based on content

4.1 The book feature is converted to a vector space

In order to improve the accuracy of personalized recommendation, the first step is to choose a good method to calculate the similarity between books.

At present, many scholars calculate the similarity between books according to the user's borrowing records. After obtaining the user-book matrix, they calculate the similarity between books through a certain algorithm. Although this method can obtain relatively high accuracy in many cases, it is not very suitable for the situation where the user borrows less data, and the user's choice will affect the similarity between the books. For example, if two users like *Gone with the Wind* and *Python from Beginner to Proficient*, this algorithm will result in a high similarity between the two books, which is obviously not quite right.

After the analysis of the name, content and author of the book, it can be found that the book has some characteristics of its own: it will have a series of books, such as *Queen Xiao Xuan (1)* and *Queen Xiao Xuan (2)* is the same series of books; As can be seen from the names, there are also some correlations between books. For example, *the Analects of Confucius* and *the Analects of Confucius*. Just from the names, we can see that they have strong correlations.

These features show that even if there is no user choice, there are objectively some relevant relationships between books, which will not change because of the differences between users. Therefore, in order to get more accurate similarity between books, this paper will calculate the similarity from the book's own attributes.

There are many properties that can be used to describe book information, such as book name, book catalog, abstract, publisher, author, and so on. Some scholars have done relevant studies[7] using these attributes to calculate the similarity between books, and observing which attributes are more accurate by comparing the experimental methods. Their results showed that the most accurate method was to calculate the similarity of the abstract, followed by the name of the book. Due to the limitation of data, there is no summary of each book in the data obtained in this study. Finally, the attribute of book name is chosen to calculate the similarity.

For the convenience of calculation, the feature of book name needs to be converted into vector space first. This paper used is more common method of TF, it refers to a certain word in the sentence the number of occurrences[8]. For example there are two books such as "physics foundation" and "basic college physics", first for word segmentation, the result is: physics/foundation, university /foundation/physics, there are three different words are: physics, foundation and the university, but this a few words in the two books the number of occurrences of respectively is: 1, 1, 0 and 1, 1, 1, finally get the two vector space are: [1,1,0] and [1,1,1], then similarity calculations can be made.

4.2 Cosine similarity

When calculating the similarity between texts, cosine similarity algorithm is used more. Its main idea is to evaluate the similarity between two vectors according to the cosine of the Angle between them[9].

To get the similarity between two vectors, you first have to get the cosine between the vectors. Conventional methods cannot directly obtain the cosine value between vectors, so the Euclidean dot product formula is needed, and its expression is as follows:

$$a \cdot b = \| a \| \| b \| \cos\theta.$$

Where a and b represent two vectors in the space, a and b represent the magnitude of these two vectors, represents the Angle between them[10].

With this formula, the cosine similarity of the eigenvector of the book name can be solved. The first step is to get the property vector A and B of the two books, with the Angle between them being. Then, the following formula can be used to get the cosine of the Angle between them, which is their similarity.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\| A \| \| B \|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}.$$

In this formula, $A_i$and $B_i$represent the components of A and B, and n represents the length of the vector space.

From the above formula, the similarity between the two books can be obtained. By calculating all books in the database in pairs, a similarity table between books can be obtained, which is also the basic data for my recommendation in this study.

5. Recommendation Process

Using the above method, you can get a table of the similarity between books, and with this table, you can make recommendations. Here, the system will recommend two kinds of people with different methods, one is the record of borrowing, one is no record. Its recommendation process is shown in Figure 2 below:

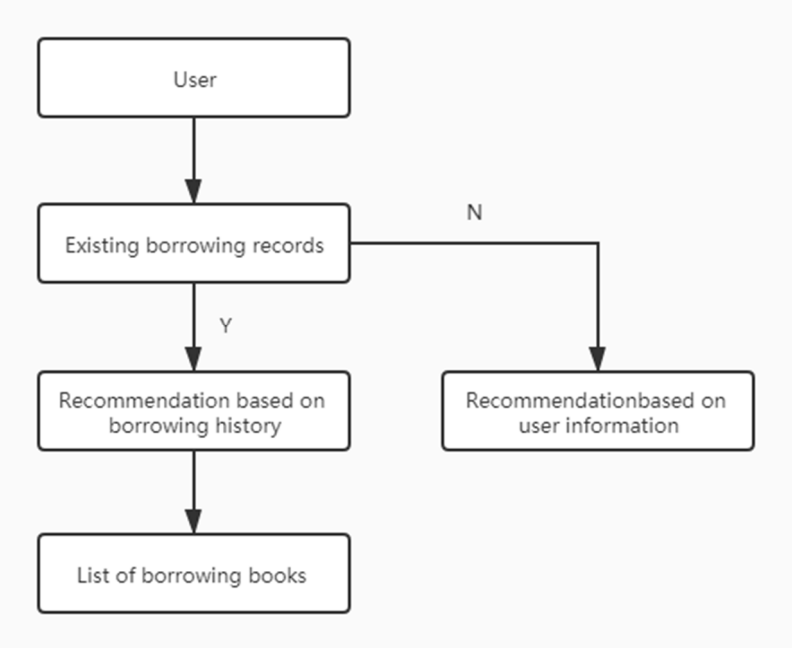

Figure2 Recommended flow chart

For users who have borrowed books, the system will find the books with the highest similarity according to the books they have borrowed. For the convenience of display, only the top 10 books with similarity are selected for recommendation.

For users who have no record of borrowing books or who are newly registered, the most frequently borrowed books of their departments will be recommended according to the information of their

registered departments.

In addition to these two recommendation methods, the system will also show the most borrowed books to readers. That is, there will be a borrowing list which contains books borrowed more, and most of the people will be interested in the books. It will have a certain recommendation effect.

The above is the main recommendation process of this system. One of its important advantages is that it solves the cold start problem.

Cold start is a very common problem in the recommendation system. Because some systems do not have too much user data for you to train in the early stage, so the recommendation effect of the system is often not satisfactory[11]. For this personalized recommendation system of university of Qinghai books, it takes advantage of the particularity of student users. Even if a student has not borrowed books from the library, he or she can recommend books that students of the university are interested in according to the borrowing records of other students of his or her department, so as to complete personalized recommendation. In addition, set a borrowing list this function, can also achieve a certain recommendation effect. In addition, relevant studies have shown that newly registered users are more likely to use the ranking list as a basis for borrowing books. Cold start is a very common problem in the recommendation system. Because some systems do not have too much user data for you to train in the early stage, the recommendation effect of the system is often not satisfactory[11] . For this personalized recommendation system of university of Qinghai books, it takes advantage of the particularity of student users. Even if a student has not borrowed books from the library, he or she can recommend books that students of the university are interested in according to the borrowing records of other students of his or her department, so as to complete personalized recommendation. In addition, also set a borrowing list this function, can also achieve a certain recommendation effect. In addition, relevant studies have shown that newly registered users are more likely to use the ranking list as a basis for borrowing books.

Through these methods, the system studied in this paper solves the problem of cold startup, so that users can recommend some books that they may be interested in even if they have no borrowing data, which has a good recommendation effect. The most important two parts of the whole system are book similarity calculation and recommendation. Its working process is shown in Figure 3 below:

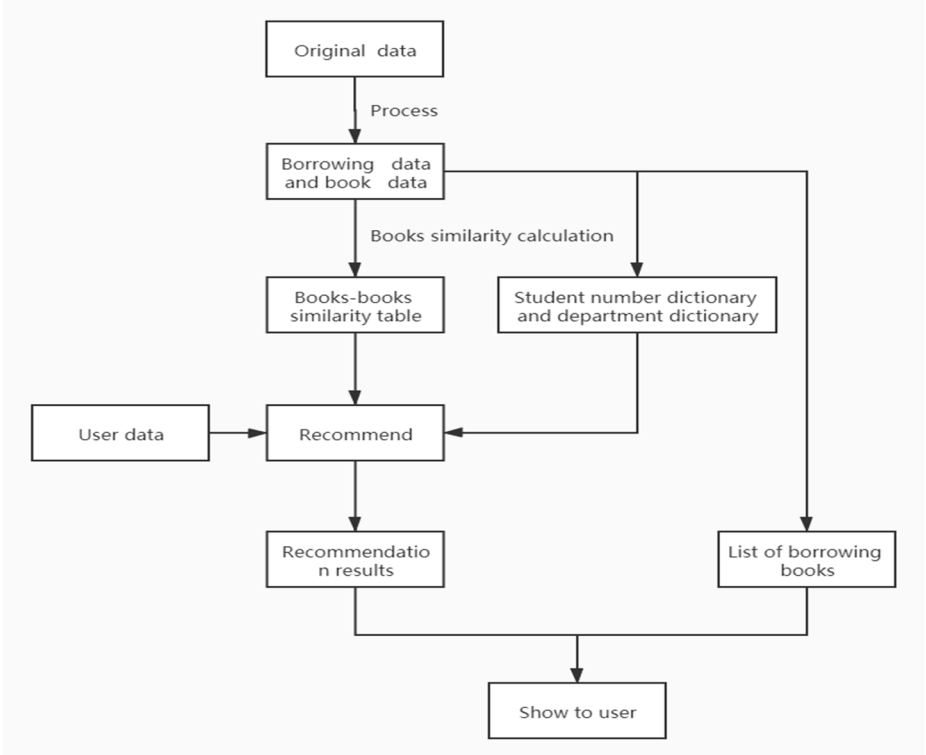

Figure3 System flow chart

Firstly, the book name information is read, and the similarity between two pairs of books is calculated through two loops. The first step in the similarity calculation is to start with the word segmentation of the book name, using jieba, the Python Chinese word segmentation library. At the beginning, jieba.cut() was used for word segmentation. It was found that there was an error in the result, which would separate out some symbols, Spaces and other things, which would affect the similarity calculation to some extent. The JieBA library has functions specifically designed to calculate TF-IDF keywords:

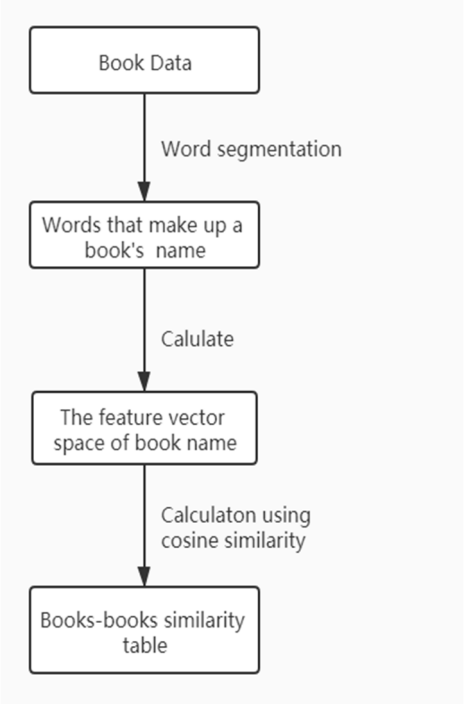

Figure4 algorithm flow chart

Through the above step, you can get the books-books similarity table. It is worth mentioning here that, since it is expected in this article that no more than ten books are recommended to users, only the top ten data of the similarity of each book is kept for convenience of searching and saving storage space. In addition, to avoid double-counting, the table needs to be placed in a document so that it can be called directly when recommended. The contents of the document are shown in Figure 5 below:

```
Selected Chinese Mini Novels in 2007    300 Mini Novels    0.5163977794943222
Psychological tricks in sales    Psychology    0.5773502691896258
Psychological tricks in sales    General psychology    0.40824829046386296
Psychological tricks in sales    Psychology and life    0.40824829046386296
Psychological tricks in sales    Introduction to psychology    0.40824829046386296
```

Figure5 Book similarity table

Since two different types of users need to be recommended, two different dictionaries need to be calculated based on the borrowing data. A dictionary takes the user as the key and the list of books borrowed as the value; The other dictionary has the department as the key value, and the list of books that the student has borrowed from the department as the value.

The algorithm flow is shown in Figure 6 below:

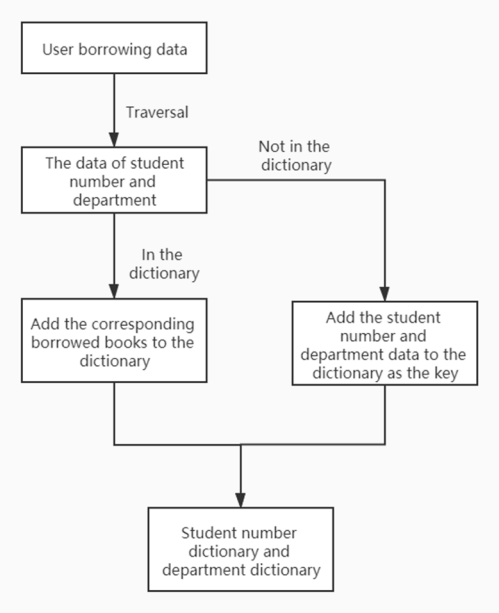

Figure6 Algorithm flowchart

6. System Implementation

Interaction is a very important part of a website. A good website must have user-friendly interaction design, which can make users find the information they want easily and quickly.

In view of users' demand for the personalized recommendation system of Books in Qinghai University, the interaction designed in this paper mainly consists of the following login, homepage and search:

6.1 The Login

The login interface is an essential part of the website. Users can log in with a student number and the password is also a student number. Its page effect is shown in Figure 7 below:

Figure7 System login interface

6.2 The main interface

Home page: After logging in, the user will enter the main interface of the system, which is mainly divided into several modules. One module is the borrowing record module. The user can see his borrowing record here. One module is the introduction module, which gives a general introduction to the library of Qinghai University. One module is book recommendation module, which will recommend

books to users according to user data. The final module is the Popular books module, which shows the most basic books that users have borrowed. Its page effect is shown in Figure 8below:

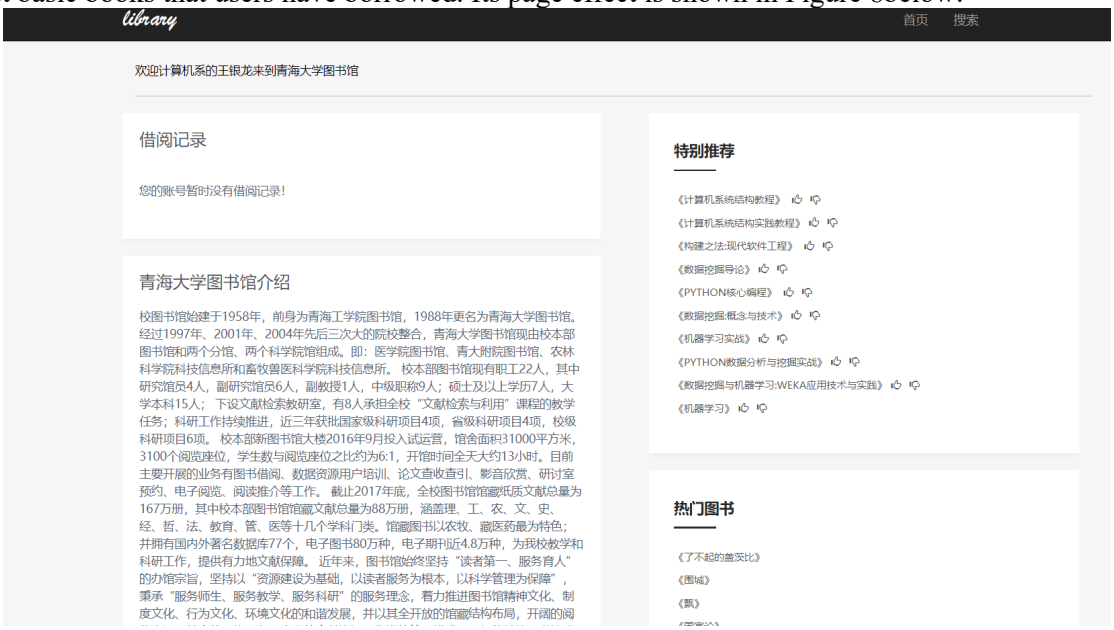

Figure8 Main page

6.3 The interface of search

Search: In the search interface, after the user enters the title, the book will be blurred with the database, books with the same words will be displayed in the search results page. And in the search box below you often search some of the book titles are displayed for user reference. The effect is shown in Figure 9 below:

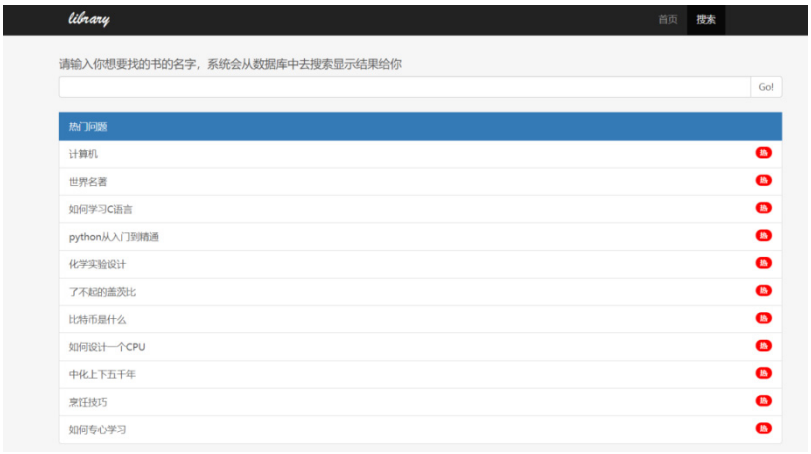

Figure9 Recommendation page

7. System Test

After passing the test, the accuracy of this algorithm can be intuitively felt. In the personalized recommendation system, there is an index to measure whether the recommendation is accurate, called recall rate[12], which indicates how much data is predicted correctly. In the book recommendation system[13][14], it means how many book recommendations are successful and what users are interested in.

In order to reflect whether the algorithm selected in this paper is superior, the recommendation results of other algorithms will be compared to observe which algorithm has the highest recall rate and the most accurate recommendation results.

The other algorithms selected here are user-based collaborative filtering algorithm and traditional

item-based collaborative filtering algorithm, which calculates the similarity between items based on the user's borrowing data.

In order to ensure the accuracy of the experiment, 10 sets of data were randomly selected, and each set of data contained 10 users. Here, the method of result evaluation is marked manually. Several students are invited to mark the final recommendation results to see whether the recommended books may be of interest to users, so as to calculate the recall rate.

The results of these ten experiments are shown in Figure 10below:

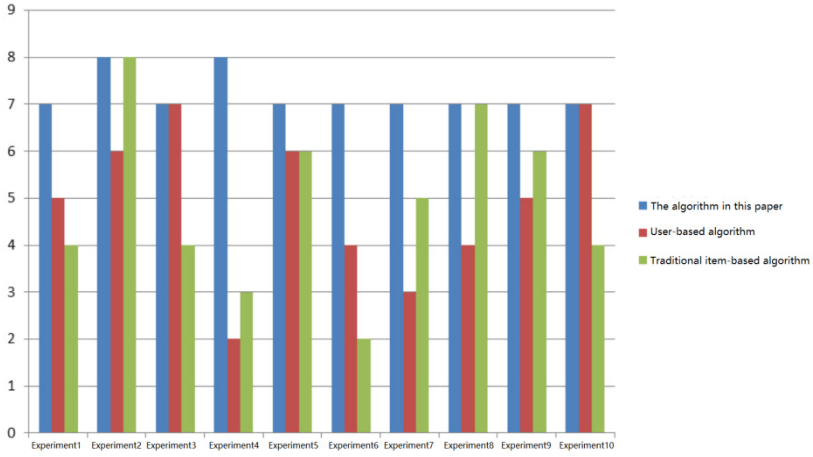

Figure10 experimental result

8. Conclusion

Starting from the actual needs of the library of Qinghai University, this paper takes the library history borrowing data as the research object, through data preprocessing, algorithm selection and implementation. In this paper, the collaborative filtering algorithm based on articles and the feature vector of book name are adopted to calculate the similarity between books and books, and the recommendation method based on the most borrowed books in the user's department is adopted as the auxiliary method to realize the realization of this system. It is proved that this method is effective in this data set by comparing 10 data sets with other algorithms.

Acknowledgements

This work is supported by the Education and teaching research project of qinghai university Grant JY201925.

Reference

[1] Sarwar B, Karypis G, Konstan J, et al.Item-based collaborative filtering recommendation algorithms.Proc10th Int'l WWW Conf, Hong Kong, 2001:1—5.
[2] Ma Hongwei, ZHANG Guangwei, LI Peng. A review of collaborative filtering recommendation algorithms [J]. Minicomputer systems,2009,30(07):1282-1288.
[3] Rich E. User modeling via stereotypes. Cognitive Science, 1979, 3 (4):329—354.
[4] Linden G, Smith B, York J. Amazon.com Recommendations: Item-to-Item Collaborative Filtering[J]. IEEE Internet Computing, 2003, 7(1):76-80.
[5] Ji Wei. Research on Algorithm and Model of University Book Recommendation System [D]. Inner Mongolia University,2017.
[6] Li Yang. Research on Recommendation Algorithm based on Collaborative filtering technology [D]. Xidian University,2015.
[7] Shang Xuejing, SUN Chengjie, Lin Lei, LIU Bingquan. Research on book Recommendation Technology based on Content Similarity [C]. Annual Conference of Digital Library High-level Forum. 2010.

[8] Wu Yong-liang, ZHAO Shu-liang, LI Chang-jing, WEI Na-di, Wang Zi-yan. Text classification method based on tf-idf and cosine similarity [J]. Chinese journal of information science,2017,31(05):138-145.
[9] Gouhanwen, Gouxiantai. Word Separation and Sentence Similarity Analysis based on Word Vector [J]. Science and Technology Innovation,2018(33):55-56.
[10] Zhang Zhenya, WANG Jin, CHENG Hongmei, WANG Xufa. Research on text Space Index Method based on cosine Similarity [J]. Computer Science,2005(09):160-163.
[11] Liu Lu. Research and Implementation of cold startup in recommended Algorithm [D]. Beijing University of Posts and Telecommunications,2019.
[12] Zhu Yu-xiao, Lu Lin-yuan. A review of recommended system evaluation indicators [J]. Journal of university of electronic science and technology of China,2012,41(02):163-175.
[13] Zhang Long, Zhang Han. Research and Application of Cross-platform responsive Front-end Framework Technology [J]. Electronic Design Engineering, 2016,26(22):6-9.
[14] Zheng Zhifang, Wei Kaile, Li Bin, Xie Yizhuang. The reason why MySQL is widely used and its embedded application [J]. The wind science and technology, 2020 (5): 114.