Research on Book Personalized Recommendation Method Based on Collaborative Filtering Algorithm

Research on Book Personalized Recommendation Method Based on Collaborative Filtering Algorithm

Yuhe Gao, Can Huang, Mengqi Hu, Jianan Feng, Xiaoxue Yang

Kunming, China Yunnan University of Finance and Economics, China

Abstract. Collaborative filtering recommendation algorithm is one of the most used in information filtering and information systems. This paper proposes a recommendation algorithm based on user collaborative filtering technology, and uses decision tree algorithm to predict the recommended books. The experimental results show that the algorithm has a more accurate prediction recommendation function and has a good reference value for book recommendation work.

1. The book recommendation system and its related algorithms

The book recommendation system provides the targeted personalized book recommendation service by using the borrowing data of the reader's history to predict the potential interests and hobbies of the readers.

The recommended algorithms commonly used in book recommendation systems can be divided into three categories:

recommendation algorithms based on association rules;

content-based recommendation algorithms;

Collaborative filtering recommendation algorithms.

The advantages and disadvantages of these algorithms are compared as shown in the following table.

Table 1. The advantages and disadvantages of three algorithms

algorithms	advantages	disadvantages
recommendation algorithms based on association rules	The algorithm is complex and the recommended quality is high	Difficult to mine, low computational efficiency, difficult to personalize
content-based recommendation algorithms	Simple analytical method	Complex algorithm, complicated processing
collaborative filtering recommendation algorithm	Simple algorithm for handling complex objects	Will sparse problems and cold start problems

The library management systems what we use are based on the existing borrowing information. And the recommendation algorithms based on association rules can analyze the hidden association rules, but the rule extraction is difficult, time consuming and computationally inefficient; Due to the

large amount of data and complex data types in the library management system, it is easy to cause the content-based recommendation algorithm to comprehensively represent the characteristics of the book during data modeling which will result in low quality of recommendation results; And the collaborative filtering recommendation algorithm does not require in-depth analysis of the knowledge content of the book resources. It only needs to analyze the characteristics and borrowing records of the readers, and analyzes their interests and their book needs according to the characteristics of the readers. The algorithm could deal with more complex processing objects. Therefore, the collaborative filtering recommendation algorithm is widely adopted by the mainstream recommendation systems. For example, Amazon and Dangdang adopt the collaborative filtering recommendation algorithm for personalized recommendation service. However, since the user evaluation information collected is relatively limited during the initial stage of the recommendation system which leads the nearest neighbor generated by the generated sparse evaluation matrix may not be accurate enough. At the same time, the algorithm ignores the user's interests in the recommended objects and only considers the user's evaluation data, which means, the algorithm only focuses on the users and the objects two-dimensional, ignoring other latitudes.

2. The collaborative filtering recommendation algorithm

The collaborative filtering recommendation algorithm mainly generates users' evaluation matrix by collecting users' information and calculates the similarity degree of the users through the evaluation matrix, then generates a nearest neighbor set of the recommended users. It provides a recommendation service according to the nearest neighbor evaluation information. The collaborative filtering recommendation algorithms are mainly divided into two categories: one is based on the user's collaborative filtering algorithm which is using similar statistical methods to obtain neighbor users' sets with similar interests; the second is model-based collaborative filtering algorithm and it uses historical data to get a recommendation model at first, and then improve model after the recommendation model is evaluated. The algorithm in this paper is based on the improved collaborative filtering algorithm.

2.1. The collaborative filtering recommendation algorithm process

The process of collaborative filtering recommendation algorithm is mainly divided into three steps: firstly establish a user model, then find the nearest neighbor user, and finally generate a recommendation list.

2.1.1. User borrowing model construction process. When we represent the data entered in the general collaborative filtering recommendation algorithm as the number of users is $m \times n$ and the evaluation matrix is $R\langle m,n\rangle$ where the row represents the number of readers and the column represents the number of books. And R_{ij} is expressed as the i th reader's evaluation of the j th book. , the evaluation matrix R can be expressed as:

$$R = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1n} \\ R_{21} & R_{21} & \cdots & R_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ R_{m1} & R_{m2} & \cdots & R_{mn} \end{bmatrix} \tag{1}$$

2.1.2. Generating the nearest neighbor set. The nearest neighbor set refers to a set of users that are similar to the target users. The user scoring matrix is used to calculate the similarity between users, and the nearest neighbor set with the highest similarity of the target user is generated. The process of

generating the nearest matrix is actually using the matrix R to calculate a similarity of the target users U from large to small.

2.1.3. Generate a list of recommended books. The summary recommendation results are stored in a database table, and the reader's borrowing card number, name, and the like are recorded in the table. After entering the recommendation system, the readers can recommend relevant information according to the readers' information, then it will realize the personalized service of the book recommendation based on the difference of the reader data.

3. The experimental process

The research background of this paper is based on the university library. The experimental data comes from a university library. The data includes three parts: the reader information table, the book information table and the reader borrowing table. This part uses the collaborative filtering recommendation algorithm to generate a wide range of recommendation results based on reader similarity. Since the data set is based on the reader and the book-based perspective, there are two kinds of calculation methods: cosine similarity and related similarity algorithm and four algorithm combinations are generated, namely:

User-based collaborative filtering uses cosine similarity to calculate similarity;

User-based collaborative filtering uses similarity to calculate similarity;

Object-based collaborative filtering uses cosine similarity to calculate similarity;

Object-based collaborative filtering uses similarity to calculate similarity;

The experiments in this section are based on the implementation of the collaborative filtering algorithm based on the Mahout open source project, and combined with the above four algorithms. Mahout converts the input data into a scoring matrix for calculation and processes the original borrowing information data table into the input format of the experimental data, as shown in the following figure:

Table 2. The experimental data

28400	21055
6704	36874
9894	39065
3027	14742
4713	32336
10533	55993
29068	55993
4509	56004
3747	17042
2223	14383

The data sample contains two data items, the first column is id, the second column is bookid, and the input data represents the borrowing relationship between the user and the book.

In the experiment, the average absolute error MAE value is calculated by the similarity algorithm to compare the quality of the algorithm. This experiment starts from 20, grows to 40 in units of 5, observes the change of MAE, and organizes the output of the program as follows:

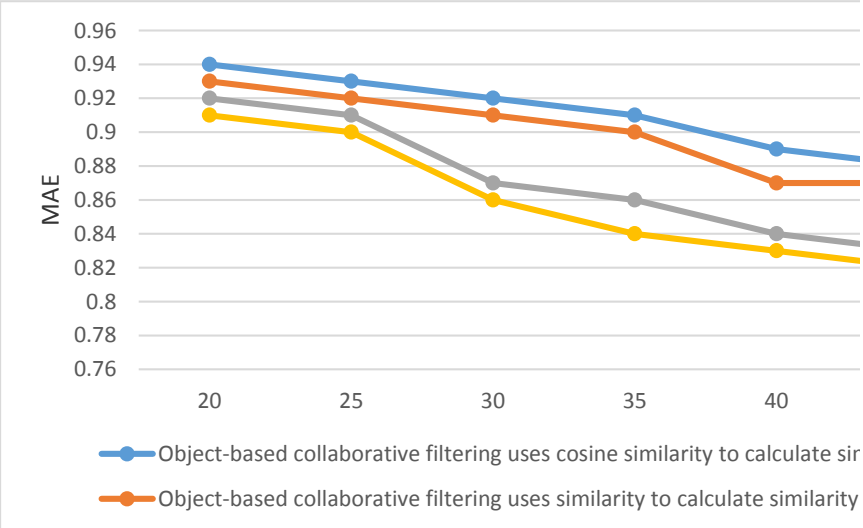

Figure 1. Algorithm comparison result

The smaller the MAE value is, the higher the similarity is. Therefore, it can be seen from the figure that the algorithm combining the user-based collaborative filtering algorithm and using the similarity similarity calculation algorithm has the smallest MAE value.

4. About decision tree

4.1. Clustering and classification

Classification is classified by type, rank or nature, while clustering is the process of dividing a collection of physical or abstract objects into multiple classes of similar objects. Simply, put together similar things and classify them. And clustering is not only that the classification is a target-driven classification, but clustering is a target-driven classification.

4.2. Decision tree

Decision trees are a typical classification method. First, the data is processed, the inductive algorithm is used to generate readable rules and decision trees, and then the new data is analyzed using the decision. Essentially, a decision tree is the process of classifying data through a series of rules.

The ID3 algorithm is a classical decision tree learning algorithm proposed by Quinlan in 1979. It is the most influential and typical algorithm in the decision tree classification method. The basic idea of the ID3 algorithm is that based on the information theory proposed by Shannon in 1948, the information entropy is a measure, which is used for the attribute selection of decision tree nodes. Each time the attribute with the most information is selected first, the entropy value can be minimized. Attributes to construct a decision tree with the fastest drop in entropy, and the entropy value at the leaf node is zero. At this time, the instances in the instance set corresponding to each leaf node belong to the same class.

$$\begin{aligned} Info_s(D) &= -\sum_{i=1}^{v} p_i \log_2 p_i \\ &= -(p_1 \log_2 p_1 + p_2 \log_2 p_2 + \ldots + p_n \log_2 p_n) \\ &= -(\frac{|C_{1,D}|}{|D|}\log_2 \frac{|C_{1,D}|}{|D|} + \frac{|C_{2,D}|}{|D|}\log_2 \frac{|C_{2,D}|}{|D|} + \ldots + \frac{|C_{m,D}|}{|D|}\log_2 \frac{|C_{m,D}|}{|D|}) \end{aligned} \quad (2)$$

Where D is a set of tuples with a set of attributes; S is a special attribute which refers to the attribute as a division criterion; m is the number of values of S, that is, the value of S has m kinds; The value $C_{i,D}$ is D divided into m subsets, which $C_{i,D}$ is the i-th subset; $Info_s(D)$ is the weighted sum of the information entropies required to obtain D according to the m classifications divided by the special attribute S; p_i is any element in the set D and the probability of occurrence in the i-th subset, ie $\frac{|C_{i,D}|}{|D|}$.

$$\begin{aligned} Info_A(D) &= \sum_{i=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \\ &= \frac{|D_1|}{|D|} \times Info(D_1) + \frac{|D_2|}{|D|} \times Info(D_2) + ... + \frac{|D_v|}{|D|} \times Info(D_v) \end{aligned} \tag{3}$$

We will expand the $Info_s(D)$ and we will get the following formula:

$$\begin{aligned} Info(D_j) &= -\sum_{i=1}^{m} p_i \log_2 p_i \\ &= -(p_1 \log_2 p_1 + p_2 \log_2 p_2 + ... + p_m \log_2 p_m) \\ &= -(\frac{|C_{1,D_1}|}{|D_1|} \log_2 \frac{|C_{1,D_1}|}{|D_1|} + \frac{|C_{2,D_2}|}{|D_2|} \log_2 \frac{|C_{2,D_2}|}{|D_2|} + ... + \frac{|C_{m,D_v}|}{|D_v|} \log_2 \frac{|C_{m,D_v}|}{|D_v|}) \end{aligned} \tag{4}$$

Therefore, it is not difficult to see that formula 3 is a call to formula 2. In formula 2, A is an attribute other than the special attribute S in the tuple set; v is the number of values of A; D_j is the number of values A according to divide D into v subsets, which is the jth subset; $Info_A(D)$ is the information entropy required to obtain D according to the v classifications divided by A; $Info_s(D_j)$ is based on the m classifications of special attributes S, and obtains the D_j required information entropy ; p_i is the probability that any element in the D_j collection is in the jth subset.

$$Gain(D) = Info_s(D) - Info_A(D) \tag{5}$$

$Gain(A)$ can determine the information gain with A, and the larger $Gain(A)$, the more information the selection test attribute provides for the classification. Therefore, the attribute A with the highest information gain is selected as the split attribute of the root node, so that the information required to complete the tuple classification is minimized.

To some extent, the smaller the information entropy of A, the greater the information gain, so Equation 5 is not needed in ID3.

5. The recommendation system based decision tree

The recommendation algorithm proposed in this paper consists of two parts. The first part is to use the collaborative filtering algorithm to make a rough recall of the recommendation results. This part only uses the borrowing information to generate the borrowing relationship matrix between the user and the

book to generate the recommendation result, but it is recommended. The precision is not enough, so this section will build the model by using the decision tree algorithm in the Python language.

We constructed the model by selecting the previously selected features, and used 80% of the previously processed borrowed information data as the training set, and 20% as the test set, and compared with the results of the previous experiments. The results are as follows:

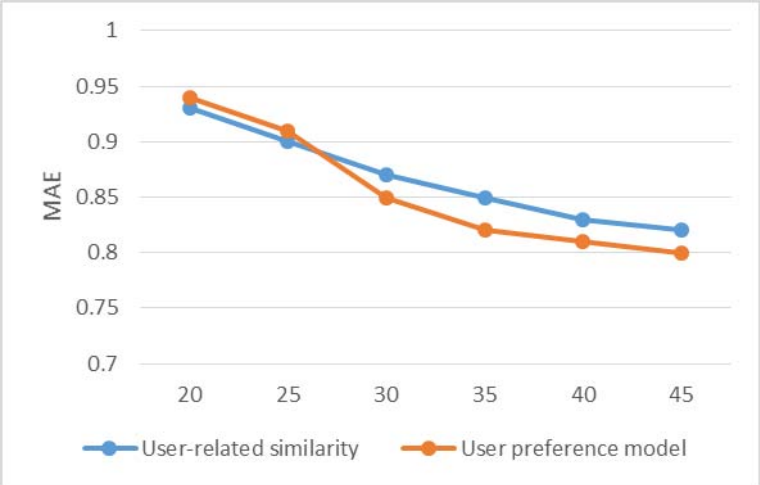

Figure 2. Algorithm comparison result

6. Conclusion

This paper based on the model generated by the feature selected by the collaborative filtering recommendation algorithm, the generated results are brought into the decision tree preference model, and the user interest preference model is constructed based on the rough recall. The saved book has a lower MAE value than the simple use of the collaborative filtering algorithm, so this recommendation algorithm can better realize the personalized recommendation of the book.

Acknowledgments

First of all, I would like to extend my sincere gratitude to National College Students Innovation and Entrepreneurship Training Program Project Funding and the project number is 201720689006.

Secondly, I would like to express my heartfelt gratitude to Ms.Huang Can, who led me into the word of algorithm. I am also greatly indebted to my partners: Hu Mengqi and Feng Jianan, who have instructed and helped me a lot.

Last my thanks would go to my beloved family for their loving considerations and great confidence in me all through these years.

References

[1] Li Yueyang. Research and Application of Personalized Recommendation Method of Books Based on Borrowing Record [D]. North China University of Technology, 2017.

[2] Liu Kan. Books Borrowing Information Analysis and Mining [A]. . Proceedings of the 25th China Database Academic Conference of the Chinese Computer Society Database Professional Committee (1) [C] China Computer Society Database Professional Committee: . , 2008: 3.

[3] Jing Minchang, Yu Yinghui. CF Recommending Model Based on Borrowing-time Scores and Its Application [J]. Library and Information Service, 2012, 56 (03): 117-120.

[4] Luo Lixia.Design and realization of book recommendation system based on the client collaborative filtering [J]. Journal of Xinyu University, 2014, 19 (06): 23-25.