Improved course recommendation algorithm based on collaborative filtering

Zheng Chen[1,2] ,Xueyue Liu[1,*] ,Li Shang[1]
[1]Collaborative Innovation Center, Communication University of China Beijing, China
[2]Information Office, Communication University of China Beijing, China
zhchen@cuc.edu.cn, moonlight197@163.com*, shangli_cuc@163.com

***Abstract*—As multidisciplinary educational interest increases, it is more and more important to support students course decision. This paper proposes a new novel recommended algorithm based on collaborative filtering for the course recommender to help student's decision. In this algorithm, the improved cosine similarity is used, according to the history of students' course selection records, and the better accuracy is obtained in the recommendation task, which meets the needs of users. In addition, both text vector and user behavior record are used to improve the calculation of course similarity. This paper evaluates 2022 students' 18457 records and 309 courses' real data. The experimental results show that the algorithm has a good effect on accuracy, recall rate and F1-score index.**

Keywords—collaborative filtering; recommendation system; course recommender

I. INTRODUCTION

Colleges and universities have widely adopted elective courses for a long time. At present, most universities in China still use the traditional way of searching and audition to let students choose elective courses with cross disciplinary academic interests. The elective course itself is also designed to cultivate diversified and personalized talents. At present, the college's elective system stores a large number of multi-faceted elective courses. For students, it is difficult to obtain valuable course information through course information query based on network platform [1]. Most students choose courses blindly, which makes some courses not selected because of the order of courses or students' inadequate understanding, which cannot meet the needs of students' interest and ability improvement. At the same time, the technical application of personalized recommendation systems in e-commerce, movie platforms, and music platforms is becoming more and more mature [2].

Course recommendation can be also regard as educational data mining and recommender system (RS) task [3]. In traditional RS task, websites can recommend similar searches based on the user 's historical browsing records, user purchase (viewing) records, and user's favorite records, reducing the difficulty for users to find and select, such as Amazon, Cloud Music, etc. [4]. In one of the first recommender system, Tapestry [5] use collaborative filtering (CF) to calculate the relationship between the user and help target users to filter retrieved similar items. Major news, streaming media and e-commerce websites have started to take collaborative filtering as one of the main technical methods to increase user stickiness [6][7]. Konstan et al. [8] propose a model to calculated the similarity between users by using the user project scoring matrix, searched for the nearest neighbor user set of the target users, and obtained the recommendation result. Breese et al., respectively, put forward and confirmed that popular items have an important impact on recommendation results, which has also been integrated with many improvement ideas [9]. At present in the field of recommender system research, the relevant domestic and foreign scholars and experts have a lot of research results, the online education platform because its content is the vast majority of online courses, also can be in online mode efficient collection and user, project related data, so also is the main course recommendation algorithm using the platform. There has been some related work in personalized recommendation system of educational administration system of higher education course. Pardos et al. integrated text information and neural network model [10], Polyzou et al. used the collaborative filtering method based on Markov chain when recommending courses [11], Ma et al. used association rules to recommend college elective courses [1]. Esteban et al. proposed a multi criteria hybrid recommendation method based on genetic algorithm [12]. Gulzar et al. proposed a hybrid method based personalized course recommendation system, which includes text content and N-gram model [13]. Khosravi and others proposed a peer learning recommendation algorithm based on knowledge gap [14]. Wang et al. put forward three different course recommendation algorithms which are all related to the course order [15].

In this paper, we apply a recommendation algorithm to the course selection system. Our algorithm uses the improved cosine similarity get better accuracy in recommendation task and meets user needs based on the collaborative filtering recommendation using the history of students' course selection records. Furthermore, the text vector of the course introduction and user behavior are both used to improved calculate course similarity.

The rest of this paper is organized as follows. In Section II, we detail the item-based collaborative filtering algorithm process and improved algorithm content. In Section III we introduce the comparative analysis of the data sets and experimental results used in this paper. Finally, Section IV concludes the paper.

II. A COLLABORATIVE FILTERING RECOMMENDATION ALGORITHM WITH IMPROVED SIMILARITY CALCULATION

In this section, we first present traditional collaborative filtering algorithm. Then we introduce our course recommendation algorithm in detail.

978-1-7281-5900-3/20/$31.00 ©2020 IEEE
DOI 10.1109/ICBDIE50010.2020.00115

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on August 02,2020 at 08:03:04 UTC from IEEE Xplore. Restrictions apply.

A. *Item based collaborative filtering algorithm*

Item-based Collaborative filtering recommendation system is based on the idea that similar users and similar projects can be given priority [4], and also based on neighborhood method. Its basic principle is shown in Fig. 1.

The ItemCF algorithm mainly consists of two steps:

a) *step 1:* according to the user project scoring matrix, the projects with high score value of the target user are obtained, and the sets of similar projects of these projects are calculated.

b) *step 2:* calculates the predicted score value of target users for each project in the similar project set, and the predicted score value will be higher than a certain threshold, or generate a recommendation list for target users based on the top-N of the predicted score value. Also, be careful to filter out items where the target user has already generated behavior.

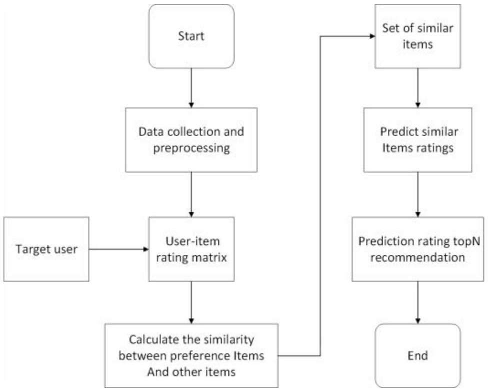
Start
Data collection and preprocessing
Target user
User-item rating matrix
Calculate the similarity between preference Items And other items
Set of similar items
Predict similar Items ratings
Prediction rating topN recommendation
End

Figure. 1. The basic principle of item-based collaboration filtering algorithm.

The key of step 1 is to use the scoring matrix of all users' projects in the system to calculate the project similarity between all projects in the matrix. In the recommendation stage, it is necessary to obtain the respective scores according to the historical scoring records of each project in the project set. Meanwhile, it is necessary to filter out the projects that target user has interacted with. The most commonly used calculation method is shown in (1):

$$\hat{r}_{targetuser,j} = \frac{\sum_{j \in Neighbors(i)}^{k} sim(i,j) \times rating(targetuser, i)}{\sum_{j \in Neighbors(i)}^{k} sim(i,j)} \quad (1)$$

Among them, $\hat{r}_{targetuser,j}$ is the target user's prediction score for item *j*, *sim(i, j)* is the similarity between item *i* and item *j*, item *i* ∈ *Neighbors (targetuser)* is the item preferred by the target user, item *j* ∈ *Neighbors (i)* is the assemble of similar items of *i*, *rating (target user, i)* represents target user's known score for item *i* from user behavior record, calculate target user's rating of item *j* by weighted average method. Finally, from all the potential item assemble of which the target user has no behavior, the top-N item assemble are recommended to target user according to the ranking of scoring values.

1) *Cosine similarity (ItemCF-cosine):*

The similarity between items is calculated according to user behavior records. For the set of items that the target user likes, find the K items that are most similar to each item in the set. By sorting the similarity calculation results, the items of the generated behaviors of the target users are filtered to recommend items to the target users. *N(i)* refers to the number of users who like item *i*, *N(j)* refers to the number of users who like item *j*, and the similarity between item *i* and *j* can be expressed in (2) through cosine similarity calculation.

$$w_{ij} = \frac{|N(i) \cap N(j)|}{\sqrt{|N(j)||N(j)|}} \quad (2)$$

2) *Improved cosine similarity (ItemCF-IUF):*

The cosine similarity calculation method of 1) shows that there is a problem, that is, each user's contribution to the interest list will have a similar term, but obviously the interest of inactive users is very concentrated, so the contribution of active users to project similarity is less than that of inactive users. Therefore, IUF (Inverse User Frequency), which is the inverse of the logarithm of user activity, is similar to cosine degree of improvement, as (3) shown:

$$w_{ij} = \frac{\sum_{u \in N(i) \cap N(j)} \frac{1}{lag(1+|N(u)|)}}{\sqrt{|N(j)||N(j)|}} \quad (3)$$

After the similarity between the items is calculated, the user's interest preference for the unselected item *i* can be obtained by the user's rating of the historical item and the similarity between the historical item and the unselected item [16], as shown in (1).

B. *Improved collaborative filtering algorithm based on TF-IDF*

The item data in the recommendation system is usually stored in the form of text. Taking the data in this paper as an example, the project text stored in the university educational administration system database contains the text description of each course. In this paper, TF-IDF is used to extract the text features of the project content, and vector space model is used to represent the text features. TF-IDF (term frequency invested document frequency) is a commonly used weighting technology for information retrieval and text mining, which is a statistical method used to evaluate the importance of a word to a file set or one of the documents in a corpus [17][18]. In this paper, in order to increase the impact of the course curriculum features of similarity, using TF-IDF theory extracted from the course description text vector text, and then calculate the course similarity, the core idea of TF-IDF model is , for the text part considered as a word, the TF-IDF model considers a document that makes up corpus as a set of words and assigns a weight value to each word. Finally, the original text is represented as a vector, which translates the problem of calculating text similarity into computing vector similarity. The model mainly contains two factors [19]:

1) *TF (Term Frequency):*

i.e. the frequency of a word appearing in the text. the frequency of the larger mean greater contribution to the text of the word, by the (4) calculation

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on August 02,2020 at 08:03:04 UTC from IEEE Xplore. Restrictions apply.

$$TF_{i,d} = \frac{N_{t_i}}{\sum_k N_{t_i}} \tag{4}$$

Where N_{t_i} is the number of occurrences of the word t_i in the document d, and the denominator represents the total number of occurrences of all words in the document d .

2) IDF (Inverse Document Frequency): the frequency of a word appearing in other texts, the greater the frequency, the more widely used and the less representative the word is, the more difficult it is to represent the text features, calculated by (5)

$$IDF_{id} = log\frac{N}{(1+n_i)} \tag{5}$$

Where N is the total number of texts in the corpus, i.e. a collection of text, n_i represents a text assemble comprising word t_i amount of text, if t_i is not a collection of text, it will result in zero denominator, therefore, it is generally used $(1 + n_i)$.

Finally, the TF-IDF value of the word t_i in a certain text d is obtained through (6)

$$TFIDF_{id} = TF_{id} \times IDF_{id} \tag{6}$$

Through the word segmentation, each text in the text set gets its own series of word strings, solves TF-IDF value of each word in the string, and then obtains the text vector of the text, for example, the text vector of the text p may be represented as (7).

$$T_p = \{(t_{p1}, w_{p1}), (t_{p2}, w_{p2}), \ldots, (t_{pn}, w_{pn})\} \tag{7}$$

Taking into account the characteristics of the application scenario of course recommendation, it is determined that there will not be a large change in the course set and the number of students is large. Therefore, this article uses ItemCF as the basis for improving the course recommendation algorithm. Using the cosine similarity of (8) calculates a set of text similarity of each text as a supplement course similarity.

$$sim(i,j)_{i,j\in|D|} = \frac{T_j \times T_j}{\sqrt{T_i^2} \times \sqrt{T_j^2}} \tag{8}$$

Based on course collaborative filtering is improved on the similarity as (9) shown. Where w_{ij} uses the improved cosine similarity shown in (3).

$$sim(i,j)'_{i,j\in|D|} = \frac{sim(i,j)_{i,j\in|D|} + w_{ij}}{2} \tag{9}$$

III. Experimental Results and Analysis

This section evaluates the proposed improved collaborative filtering course recommendation algorithm. The experimental platform is configured as follows: operating system: Win10 x64-bit; CPU is i7 processor; development language and platform: Python + Microsoft VScode. The experimental data are 18457 experimental data records and 309 course descriptions from 2022 students in Communication University of China.

Offline experiment the use of 4 embodiment of the fold cross-validation, i.e. the total test set of training set ratio of 25% taken 4 times the mean of each set of parameters as the test the final results, respectively, in the neighborhood K is 5, 10, At 20, 40, 80 and 160, compare ItemCF, ItemIUF and the similarity correction recommendation algorithm in this article.

The quality of the recommendation system is measured by its prediction results. In this paper, three indicators of accuracy, recall rate and F1-score are used to evaluate the accuracy of the recommendation results.

*1) Precision:*which is used to measure the accuracy of the recommendation result obtained by the algorithm on the training set on the test set, is defined as the ratio of the user's realistic interest in the recommendation list to the length of the entire recommendation list. The calculation of recommendation accuracy for a single user $u \in U$ is shown in (10) .

$$Precision_u = \frac{|L_u \cap B_u|}{|L_u|} \tag{10}$$

In (10), L_u represents the recommendation result obtained by the algorithm for user u on the training set, and B_u represents the item of realistic interest of user u on the test set. The value range of accuracy is from 0 to 1, from small to large, indicating more accurate recommendation result. The calculation of recommendation accuracy of the whole system is shown in (11).

$$Precision_{all} = \frac{\sum_{u\in U} |L_u \cap B_u|}{\sum_{u\in U} |L_u|} \tag{11}$$

2) Recall: which is used to measure the degree of users' real interest in the recommendation list, and to define the ratio between the recommended correct items and the real interest items. The recommended recall rate of a single user u is shown in (12) .

$$Recall_u = \frac{|L_u \cap B_u|}{|B_u|} \tag{12}$$

It can be seen that the numerator in the calculation of recall rate is consistent with the numerator of accuracy, but the denominator is transformed into a list of test sets. Similarly, the recall rate of the whole system is shown in (13):

$$Recall_{all} = \frac{\sum_{u\in U} |L_u \cap B_u|}{\sum_{u\in U} |B_u|} \tag{13}$$

*3) F1 – score:*usually adopts the accuracy and recall rate from two dimensions to evaluate the accuracy of the algorithm to recommend the results, but in some cases, the accuracy of algorithm A is higher than that of algorithm B, and the recall rate of algorithm B is higher than that of algorithm A, when taking into account the need to use the F1 - score accuracy rate and recall rate of performance evaluation, the calculation method is shown in (14) .

$$F_1 = \frac{2 \times Precision_{all} \times Recall_{all}}{Precision_{all} + Recall_{all}} \tag{14}$$

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on August 02,2020 at 08:03:04 UTC from IEEE Xplore. Restrictions apply.

By substituting the system's recommendation accuracy and recall rate into the, the f1-score of the whole recommendation algorithm can be calculated.

As said in Section II, ItemCF and ItemIUF algorithms are commonly used in the recommendation system, the algorithm is to calculate the difference between the two different items of similarity, herein incorporated ItemCF text similarity on the basis of, and through the conventional ItemCF Improve the comparison of ItemIUF algorithm with similarity. The experimental results of accuracy and recall are shown in Fig. 2 respectively.

The experimental results of F1-score are shown in Fig. 3.

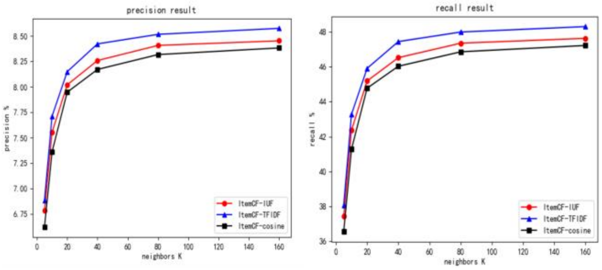

Figure. 2. precision and recall results in different neighborhoods K.

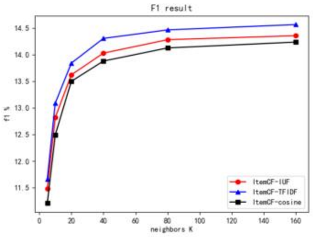

Figure. 3. F1-score results in different neighborhoods K for comparison.

It can be seen that our algorithm have achieved improved precision and recall rates than ItemCF and ItemIUF base on collaborative filtering algorithm of TF-IDF text similarity optimization. The value of F1-score also proves this conclusion.

IV. Conclusions

This paper proposes a collaborative filtering optimization algorithm based on text similarity calculation. Experimental results show the feasibility of the model. Compared with collaborative filtering recommendation based on users and items, the accuracy of the recommendation results is greatly improved. For future research, the impact of different fusion strategies on the results will be explored to maximize the effectiveness of recommendations.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61971382 and the "Fundamental Research Funds for the Central Universities".

References

[1] Ma B. (2019). Design of an Elective Course Recommendation System for University Environment. educational data mining.

[2] Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., & Riedl, J. (1997). GroupLens: applying collaborative filtering to Usenet news. Communications of the ACM, 40(3), 77-87.

[3] Romero, C., & Ventura, S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27.

[4] Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. Advances in artificial intelligence, 2009.

[5] Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. Communications of the ACM, 35(12), 61-70.

[6] Herlocker, J. L., Konstan, J. A., Terveen, L., & Riedl, J. (2004). Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems, 22(1), 5-53.

[7] Adomavicius G, Tuzhilin A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans on Knowledge and Data Engineering, 2005, 17(6): 734-749

[8] Herlocker, Jonathan L., Konstan, Joseph A., Terveen, Loren G., & Riedl, John T. (2004). Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems, 22(1), 5-53.

[9] Breese J S, Heckerman D, Kadie C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering // Proc of the 14th Conference on Uncertainty in Artificial Intelligence. San Francisco, USA, 1998: 43-52

[10] Pardos, Zachary A., Zihao Fan, and Weijie Jiang. "Connectionist recommendation in the wild: on the utility and scrutability of neural networks for personalized course guidance." User Modeling and User-adapted Interaction 29.2 (2019): 487-525.

[11] Polyzou, Agoritsa, Athanasios N. Nikolakopoulos, and George Karypis. "Scholars Walk: A Markov Chain Framework for Course Recommendation.." educational data mining (2019).

[12] Esteban, A., Gomez, A. Z., & Romero, C. (2018). A Hybrid Multi-Criteria Approach Using a Genetic Algorithm for Recommending Courses to University Students.. educational data mining.

[13] Gulzar, Z., Leema, A. A., & Deepak, G. (2018). PCRS: Personalized Course Recommender System Based on Hybrid Approach. Procedia Computer Science,, 518-524.

[14] Khosravi, Hassan, Kendra M L Cooper, and Kirsty Kitto. "RiPLE: Recommendation in Peer-Learning Environments Based on Knowledge Gaps and Interests." educational data mining (2017): 42-67.

[15] Wang, Ren, and Osmar R. Zaiane. "Sequence-Based Approaches to Course Recommender Systems." database and expert systems applications (2018): 35-50.

[16] Xiang Liang . Key technologies of dynamic recommendation system. Institute of Automation, Chinese Academy of Sciences. 2011

[17] Salton, G. and McGill, M.J. (1983.) Introduction to Modern Information Retrieval. McGraw-Hill Book Co., New York.

[18] Rajaraman A, Ullman J D. Mining of massive datasets. Cambridge University Press, 2011.

[19] Salton, G., Fox, E. A., & Wu, H. (1983). Extended Boolean information retrieval. Communications of the ACM, 26(11), 1022-1036.

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on August 02,2020 at 08:03:04 UTC from IEEE Xplore. Restrictions apply.