A Collaborative Filtering Recommendation Algorithm with Improved Similarity Calculation

Yang Ju[a], Liu Bailin[b] and Zhao Zhixiang
School of Computer Science and Engineering, Xi'an Technological University
Xi'an, 710021, Shaanxi
State and Provincial Joint Engineering Lab. of Advanced Network and Monitoring Control
Xi'an, 710021, China
e-mail: [a]yangju90@163.com, [b]498194312@qq.com

***Abstract*—In order to improve the accuracy of the proposed algorithm in collaborative filtering recommendation system, an Improved Pearson collaborative filtering (IP-CF) algorithm is proposed in this paper. The algorithm uses the user portrait, item characteristics and data of user behavior to compute the baseline predictors model. Instead of the traditional algorithm's similarity calculation, the prediction model is used to improve the accuracy of the recommendation algorithm. Experimental results on Moivelens dataset show that the IP-CF algorithm significantly improves the accuracy of the recommended results, and the RMSE and MAE evaluation results are better than the traditional algorithms.**

Keyword-Recommendation Algorithm; Collaborative Filtering; Similarity Calculation; Baseline Predictors Model

I. INTRODUCTION

With the rapid development of computer technology and network technology, the number of network information services and applications is growing rapidly. China Internet Network Information Center reported statistics, as of June 2016, the size of China's Internet users reached 710 million, a total of 21.32 million new netizens in half a year [1]. With the increase in the number of people on the Internet, Internet information has also seen explosive growth. How to find interesting and effective information in this vast data is a very difficult thing. In order to solve this problem, academia and industry put forward personalized recommendation system[2]. According to the user's personal information and historical habits, it can discover the potential interest of the user and recommend the resources of interest to the user actively.

Personalized recommendation system is a special form of information filtering system[3]. The recommendation system can be divided into the following categories: collaborative filtering recommendation system, content-based recommendation system and hybrid recommendation system. Because of its wide applicability, strong interpretability and good stability, the collaborative filtering recommendation system based on neighborhood model is widely used in various fields. Therefore, this paper focuses on the collaborative filtering recommendation system based on neighborhood model.

The accuracy of recommendation results in the recommendation system is the main index to measure the recommendation effect. Sarwar et al.[4] proposed an item-based collaborative filtering recommendation algorithm that looked into cosine-based similarity to compute the similarity between products. This method provides dramatically better performance than traditional recommendation algorithm, while at the same time providing better accuracy. Chen and Cheng[5] use the rating data to compute the similarity between users, and use the ranking data as the weight of similarity calculation. Yang and Gu[6] propose to use user behavior information to construct the user's interest points and use the interest points to compute the similarity between users. Experiments show that these methods are better than the classic collaborative filtering algorithm. However, these methods only consider the user-item behavioral data, and neglect the user portrait and item features, which causes deviations in the accuracy of the recommendation result. This paper improves similarity calculation method in collaborative filtering recommendation algorithm based on neighborhood model, and uses the user portrait, item characteristics and user-item behavior data to compute similarity. We experimentally evaluate our results and compare them to the classic collaborative filtering algorithm. Experiments suggest that the improved similarity calculation method can improve the accuracy of the recommended results.

II. COLLABORATIVE FILTERING RECOMMENDATION ALGORITHM BASED ON NEIGHBORHOOD MODEL

Collaborative filtering recommendation algorithm is to select the same custom hobby user groups, use other people's experience to meet their own needs, in order to achieve the purpose of reducing overhead. Typically, the workflow of a collaborative filtering system is:

(1) Compute the similarity between users.

(2) Determine the neighbor set. Find the k users whose user interest is the most similar through the similarity size, and set these users as the user sets.

(3) According to the user sets prediction rating. The system recommends items that the users have rated highly but not yet being rated by this user.

The most important thing in collaborative filtering algorithm is similarity calculation. For the calculation of

similarity, the researchers put forward a variety of similarity calculation methods.

Cosine-based similarity: For user u and user v, $N(u)$ denotes the set of positive feedback items for user u, and $N(v)$ denotes the set of positive feedback items for user v. And similarity between items u and v, denoted by $sim(u, v)$ is given by:

$$sim(u, v) = \frac{|N(u) \cap N(v)|}{\sqrt{|N(u)||N(v)|}} \quad (1)$$

Pearson correlation coefficient[7-8]: In this case, similarity is computing based on the vector of the rating. Among them, $\bar{r}_u$ in the formula has two forms in the traditional recommendations. One is the average rating of user u, and the other is the average rating of item i scored by all users. The Pearson correlation is given by:

$$sim(\mathrm{u}, \mathrm{v}) = \frac{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{i \in I_{uv}} (r_{vi} - \bar{r}_v)^2}} \quad (2)$$

The similarity between users can be calculated by the above formula, and the similarity ranking of each user with other users can be obtained to obtain the nearest neighbor user set. After getting the user set, the next step is interest prediction computation. We can denote the prediction $p(u, i)$ as:

$$p(u, i) = \sum_{v \in S(u,K) \cap N(i)} sim(u, v) r_{vi} \quad (3)$$

III. Improved Similarity Measures

This paper considers the user rating data from the overall situation, introduces the characteristics of personal habits, item quality and category to improve the similarity computation formula. Thus the approximated correlation coefficient is given by:

$$S_{m,n} = \frac{\sum_{i \in P_{m,n}} (r_{mi} - b_{mi})(r_{ni} - b_{ni})}{\sqrt{\sum_{v \in P_{m,n}} (r_{mi} - b_{mi})^2} \sqrt{\sum_{v \in P_{m,n}} (r_{ni} - b_{ni})^2}} \quad (4)$$

Here, $S_{m,n}$ denotes the similarity between user m and n. r_{mi} is the raw rating of the user m for the item i. $P_{m,n}$ represent the user m, n common rating set. b_{mi} is a baseline predictors for rating r_{mi}. The baseline predictors model is as follows:

$$b_{mi} = \mu + b_m + b_i + \sum_{g \in G_i} c_g \quad (5)$$

μ denotes the intercept of the baseline predictors model. The parameters b_m and b_i indicate the deviations of user m and item i, respectively, from the average. The last term of the formula denotes the preference of the user on the item category. G_i represents the set of categories to which the item belongs. Here we use an example to illustrate the baseline predictors model. Create a baseline predictors model for the rating of movie i for user m. Assume that the mean rating of all the movie scores is 3.5. m is a critical user, who tends to rate 0.3 stars lower than the average. i is a movie with a relatively high standard, so its rating is 0.5 stars higher than the average rating. In addition, the movie i belong to c_x, c_y, c_z, with a bias relative to the average of -0.05, 0.08, 0.12. Therefore, the prediction for the movie i rating by user m is 3.5-0.3+0.5-0.15+0.18+0.12=3.85. For this formula, the purpose is to find b_m, b_i and c_g. This paper solves the problem by solving the least-squares problem[9-11]. The cost function formula is as follows:

$$e_{mi} = r_{mi} - \mu - b_u - b_i - \sum_{g \in G_i} c_g \quad (6)$$

$$\min \sum_{(m,i) \in \kappa} (e_{mi})^2 + \lambda \left(\sum_{m \in U} b_m^2 + \sum_{i \in I} b_i^2 + \sum_{i \in I} \sum_{g \in G_i} c_g^2 \right) \quad (7)$$

In the above (6) formula, r_{mi} is the true value of user for item. In formula (7), κ represents the set of user ratings for items, U represents user set, and I represents the set of item. The solution process is to obtain the best fitting b_m, b_i and c_g by minimizing the first term $\sum_{(m,i) \in \kappa} (e_{mi})^2$ in equation (7). The second item is the L1 regular, that is added to prevent overfitting. The size of λ indicates the degree of intervention to fit, and the larger the general λ is, the smoother the fitting curve is.

Because the proposed model of this paper is different from the traditional one, the data matrix cannot be directly applied to the training of the model. So the matrix of the training data is restructured, adding personal habits, item quality and category bias. For example, when a movie i was released, it was called a masterpiece of elements such as comics, entertainment, suspense, etc. These classified data were useful for the model but could not be used. Through the transformation of the data format, useful information is used, and the information is vectorized according to the classification categories. Each row of data through the transformation training matrix can be expressed as: $\{(u, i), u_1 \ldots u_m, i_1 \ldots i_n, c_1 \ldots c_j, r_{ui} \mid c \in G\}$.

TABLE I. TRAINING DATA MATRIX

(u, i)	u_1	...	u_m	i_1	...	i_n	c_1	...	c_j	r_{ui}
(1,1)	1	...	0	1	...	0	1	...	0	4
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
(1,n)	1	...	0	0	...	1	0	...	0	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
(m,n)	0	...	1	0	...	1	1	...	1	5

TABLE II. MOIVELENS DATASET INFORMATION

Version	users	movies	size	Sparsity
ML	943	1682	100K	93.695%

The training data matrix is shown in Table I.

The complete algorithm steps are described as follows:

a) Compute the baseline predictors model. According to the formula (5) combined with the rating matrix, personal habits, item quality and category, the least square method is used to solve for b_{mi}, b_m, b_i, and $\sum_{g \in G_i} c_g$. The solution formulas are (6) and (7).

b) Compute similarity. Using formula (4), compute the similarity between each two users.

c) Getting a set of nearest neighbors. According to the similarity computed in the steps (b), we sort the nearest neighbors of user m that need to be predicted, and determine the relevant user set S_m^k according to sorting order and k.

d) Rating prediction. According to formula (3), system recommends items that the users have rated highly but not yet being rated by this user.

IV. EXPERIMENTAL EVALUATION

A. Data Set And Evaluation Metrics

In order to verify the actual recommendation effect of the proposed algorithm (Improved Pearson Similarity Collaborative Filtering, IP-CF) in this paper, the MoiveLens film data set was used for verification. This data set consists of: ①100,000 ratings (1-5) from 943 users on 1682 movies. ②User data and item data have simple feature portraits. ③ Users with less complete personal portraits and fewer comments in the data have been cleaned. The dataset information is shown in Table II.

Select the root mean square error (RMSE) and mean absolute error (MAE) to evaluate the accuracy of the recommendation algorithm on the rating data[12-13]. The smaller the value, the higher the accuracy of the prediction. For a user u and item i in the test set, r_{ui} is the actual rating, $\hat{r}_{ui}$ is the predicted rating, and T is the total number

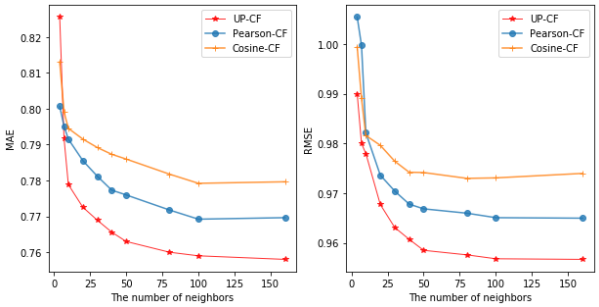

Figure 1. Comparison of precision between IPCF algorithm and traditional algorithm

of items that need to be predicted. The RMSE and MAE formulas are as follows:

$$RMSE = \sqrt{\frac{1}{|T|}\sum_{u,i \in T}\left(r_{ui} - \hat{r}_{ui}\right)^2} \tag{8}$$

$$MAE = \frac{1}{|T|}\sum_{u,i \in T}\left|r_{ui} - \hat{r}_{ui}\right| \tag{9}$$

B. Experimental Results

1) Comparison of accuracy of recommendation algorithms

The traditional cosine similar algorithm (Cosine-CF) and Pearson similar algorithm (Pearson-CF) were compared with the proposed algorithm (IP-CF). We tested them on our data sets by computing RMSE and MAE. The size k of similar user set is from 5 to 180. Figure 1 shows the experimental results.

It can be observed from the results that the RMSE and MAE values of the improved similarity algorithm proposed in this paper decrease with the increase of the neighborhood. When the number of near-neighbor sets reaches a certain amount, it tends to a fixed value. The traditional collaborative filtering algorithm (Pearson similarity and cosine similarity) needs to find the optimal result, if the number is too large, it will affect the accuracy of the recommendation result. Overall, the RMSE and MAE of the rating prediction are 0.82% and 1.16% lower than the

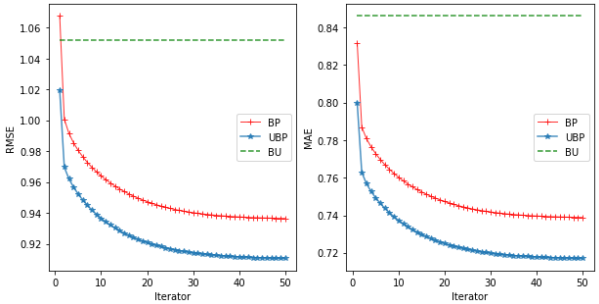

Figure 2. Comparison between the baseline predictors model and the traditional model

traditional algorithm respectively. The IP-CF algorithm has better accuracy.

2) Comparison of accuracy of baseline predictors

The following experiments verify the effectiveness of the baseline predictors model. The experiment compares the user mean model (BU), basic baseline predictors model (BP), and improved baseline predictors model (UBP). Experimental results show that the improved baseline predictors model significantly improves the accuracy of the baseline prediction.

V. CONCLUSION

In this paper, a collaborative filtering recommendation algorithm based on improved similarity computation is proposed, which takes into account user portrait, item characteristics and user behavior data in recommendation process. Experiments have shown that user portraits and item features played an important role in improving the accuracy of recommendations, and which are an important basis for analyzing potential needs. Secondly, we found that in the Top-N recommendation, the number of neighbors and the evaluation index are not a positive or negative relationship, and the size of the neighbor will affect the accuracy of the recommendation. Our further work will research the relationship between the number of neighbors and the effectiveness of recommendations, especially how to choose the best neighbor value to improve the accuracy of recommendations.

REFERENCES

[1] Statistical Report on Internet Development in China[R]. China Internet Network Information Center, 2017.

[2] Borkar V, Carey MJ, Li C. Inside Big Data management:ogres, onions, or parfaits[C]. Proceedings of the 15th International Conference on Extending Database Technology. ACM, 2012:3-14.

[3] WANG Guoxia, LIU Heping. Survey of personalized recommendation system[J]. Computer Engineering and Applications, 2012, 48(7), 66-76.

[4] Sarwar B, Karypis G, Konstan J, et al. Item based collaborative filtering recommendation algorithms[C]. Proc 10th Int'l WWW Conf, Hong Kong, 2001:1-5.

[5] Chen YL , Cheng LC. A novel collaborative filtering approach for recommending ranked items[C]. Expert Systems with Applications, 2008, 34(4): 2396 -2405.

[6] Yang MH, Gu ZM. Personalized recommendation based on partial similarity of interests[C]. Advanced Data Mining and Applications Proceedings, 2006, 4093:509-516.

[7] FU He-gang, WANG Zhu-wei. Improvement of Item-Based Collaborative Filtering Algorithm[J]. Journal of Chongqing University of Technology(Natural Science), 2010(9):71-72.

[8] GUO Lei, MA Jun. Incorporating Item Relations for Social Recommendation[J]. Chinese Journal of Computers, 2014, 37(1):220-227.

[9] X. Luo, Xia, Y. and Zhu, Q. Incremental collaborative filtering recommender based on regularized matrix factorization[J]. Knowledge-Based Systems, 2012, 27, pp.271-280.

[10] SUN Chen, XI Hongsheng, GAO Rong. A Recommendation-Support Model Using Neighborhood-Based Linear Least Squares Fitting[J]. Journal of Xi'an Jiaotong University, 2015, 49(6), 78-83.

[11] C. Rana, and Jain, S.K. A study of the dynamic features of recommender systems[J]. Artificial Intelligence Review, 2015, 43(1), pp.141-153.

[12] Item-network-based collaborative filtering: A personalized recommendation method based on a user's item network[J]. Taehyun Ha, Sangwon Lee. Information Processing and Management. 2017(5).

[13] Gai Li, Zhiqiang Zhang, et al. One-Class Collaborative Filtering Based on Rating Prediction and Ranking Prediction[J]. Knowledge-Based Systems. 2017.