ATLANTIS PRESS
Advances in Engineering Research (AER), volume 61
International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2017)

A Collaborative Filtering Algorithm based on Improved Similarity

Yan Zhou

School of Electrical and Information Engineering, Nanchang Institute of science & Technology, Nanchang, 330108, China

zhouxiazju@163.com

Keywords: Collaborative Filtering, Movie Recommendation, Pearson Similarity.

Abstract. The collaborative filtering recommendation implements recommendation by using his neighbor user's preference. And the similarity calculation is the key. The traditional similarity calculation neglects the impact of co-rating item number and user average rating on similarity calculation. This causes the poor similarity calculation of users in case of sparse data. This paper introduces the two improved factors to the improved algorithm, so as to improve the traditional similarity. Meanwhile, the improved recommendation algorithm has been applied to film recommendation system. The simulation experiment shows that the improved recommendation algorithm can get a lower MAE value than traditional recommendation algorithm. In addition, the improved algorithm can improve the quality of film recommendation system.

1. Introduction

In contrast with the filtering recommendation based on content, the collaborative filtering recommendation firstly analyses the user's interest by finding the similarity users of target users in the consumer space, then evaluates one product in accordance with these similarity users, so as to form recommendation predict [1]. The algorithm is divided mainly into the collaborative filtering based on item and the collaborative filtering based on user, they are called as the collaborative filtering model based on domain relations. This paper will discuss the improved recommend algorithm based on user similarity. This paper based on the existing collaborative filtering recommendation algorithm, improves the similarity calculation. The algorithm improves rating factors by the user co-rating items and average rating, integrates it into traditional similarity calculation method, and form a collaborative filtering recommendation algorithm based on the improved similarity.

2. The Collaborative Filtering Algorithm of Similarity

2.1 The Traditional Collaborative Filtering Algorithm.

In the traditional collaborative filtering algorithm it needs search the neighbor of the target user in accordance with user rated matrix and interest rating of neighbor user in order to produces the recommendation for the target user. The operation processing mainly includes three steps: similarity calculation, selection of the neighbor user and rating prediction [2]. It implements recommendation in accordance with the similarity of the user rating of film in the film recommendation. Among them, the similarity calculation is the core part of collaborative filtering recommendation algorithm based on user. The general method of the similarity calculation mainly includes COS similarity, Pearson similarity and ACOS similarity, etc. These similarity calculation methods are mainly based on vector [3]. It takes rating of all users as a vector, so as to calculate the similarity of users. These similarity calculations can be accurately calculated at an early stage. With the rapid development of E-ecommerce, the number of users and items will be dramatically increased, which cause the rating matrix of item-rating become fairly sparse. Therefore in recent years many studies focus on improving the accuracy of similarity calculation based on the sparse data. MS Shang proposed a similarity calculation method based on graphic [4]. These algorithms improve accuracy of the similarity calculation to some extent, but there are still some defects.

In the traditional similarity calculation, because the Pearson similarity is easy to understand and calculate, it is used widely. The specific expression is shown as formula (1):

$$sim(u,v)=\frac{\sum_{i\in I_{uv}}(r_{u,i}-\overline{r_u})(r_{v,i}-\overline{r_v})}{\sqrt{\sum_{i\in I_{uv}}(r_{u,i}-\overline{r_u})^2}\times\sqrt{\sum_{i\in I_{uv}}(r_{v,i}-\overline{r_v})^2}} \tag{1}$$

Where $r_{u,v}$ represents user *u* ratings for the item i, $\overline{r}$ represents average ratings for all evaluated items, I_{uv} represents co-ratings items of user *u* and user *v*.

The Pearson similarity is generally only used for calculating compactness of two Interval variables, which is [-1, 1]. Where "1" value represents two users have consistent evaluation of each items.

2.2 Drawback of the Traditional Similarity Calculating method.

With the development of industry, the numbers of users and items are increased geometrically, which cause the rating matrix of item-rating become fairly sparse. Although the traditional similarity calculations are widely applied in the recommendation, it cannot get real the nearest neighbor sets. The main reason of the accident has the following several aspects: the effect of co-rating items similarity and the average rating.

3. Collaborative Filtering Recommendation Algorithm Based on Improved Similarity

The traditional similarity calculation only focus on rating value of the co-rating items, and it has much impact on recommendation quality in the sparse data. According to analysis of the traditional similarity calculation methods, this paper adapts the similarity calculation methods. The method is mainly implemented in the following three aspects.

(1) It needs consider the impact of co-rating items to the user similarity when improve similarity, this paper calculates the impact factor in accordance with proportion of co-rating items.

(2) It needs consider the impact of the two users' average ratings to the user similarity when improve similarity, this paper calculates impact factor in accordance with the average ratings factor.

(3) After the improved similarity calculation is used, the mean absolute error (MAE) should be lower than the traditional method.

3.1 The Co-rated Items.

In the user rating matrix, if the preferences of the two user *u1* and *u2* are similar, the rating items of the *u1* and *u2* are N1and N2 respectively, the co-rating items are n, and the value of the n/N1 and n/N2 should be larger. Therefore, it can effectively improve the defect in accordance with the proportion of co-rating items. For the two users, if $N1 \geq N2$, the percentage of the co-rating item n and

N1 is more important. But above all, this paper introduces the proportion *R(u,v)* to improve the factor

The proportion of item-ratings is shown as formula (2):

$$R(u,v)=\frac{n}{\max(N_u,N_v)} \tag{2}$$

Where *n* represents the co-rating items of the user *u* and *v*, N_u and N_v represents the rating items of *u* and *v* respectively. If the larger R, the more similar the whole similarity of *u* and *v*, while the smaller *R*, the lower similar the whole similarity of *u* and *v*.

3.2 The Average Rated Items.

The similarity calculation is not as accurate as possible in the sparse data. For the problems, many experts and scholars proposed some methods. J. Herlocker proposes an algorithm [5], which improves the problem of the users co-rating items by setting a threshold. For example, when threshold is 50, if the co-rating value is less than 50, the original similarity will be multiplied by the improved factor n/50. If the co-rating value is more than 50, the original similarity will be kept. But the method also has some defects, while is only suitable to special cases. In addition, this method didn't focus on the problem of the user average rating. Therefore, this paper introduces *d (u, v)* to balance the average rating difference of the

user *u* and *v* in accordance with the improved factor of user rating difference. The calculation method is shown in formula (3):

$$d(u,v)=\frac{1}{n}\sum_{i\in I_{uv}}\left|r_{u,i}-r_{v,i}\right| \tag{3}$$

Where $r_{u,i}$ represents rating of the user *u* to item *i*, I_{uv} represents the co-rating of the user u and v, n represents the co-rating numbers of the user u and v. The large the *d (u, v)* means the larger the user average rating difference, and the whole similarity is low similarity. The improved average rating factor *p(u, v)* is shown in formula (4):

$$p(u,v)=\frac{1}{1+d(u,v)} \tag{4}$$

Where the larger the *p*, the more similar the whole similarity of *u* and *v*, while the smaller *p*, the lower similar the whole similarity of *u* and *v*.

The improved factors *R (u, v)* and *p(u, v)* in this paper comprehensively considers Herlocker, Tanimoto similarity calculation, it can improve comprehensively problems of similarity calculation.

3.3 Collaborative Filtering Recommendation Algorithm Based on Improved Similarity.

The traditional collaborative filtering recommendation method is based on co-rating items. For sufficient data case, the traditional similarity calculation method can calculate effectively the similarity of users. With the development of film and TV industry, the rating data is rather sparse, the traditional similarity calculation method is unable to accurately calculate the similarity of users, and the recommendation efficiency of the recommendation algorithm will be decreased. The improved factors *R(u, v)* and *p(u, v)* in this paper comprehensively considers the problem of proportion of co-ratings items and the average rating difference. Therefore, it can relieve effectively the inaccurate problems of the traditional similarity calculation in accordance with the sparse data. The similarity calculation expression is shown in formula (5):

$$NSim(u,v)=sim(u,v)\times R(u,v)\times p(u,v) \tag{5}$$

Where the *sim (u, v)* represents the traditional similarity calculation method.

It integrates improved similarity calculation method into traditional item recommendation based on the collaborative filtering, and can get the collaborative filtering recommendation algorithm based on the improved similarity. The specific steps of the algorithm are as follows.

(1) Calculating the similarity of users by the formula (5).

(2) Selecting the neighbor user sets of the target users in accordance with the result of the step 1, so as to predict the rating.

(3) Predicting the target user rating to items, the calculation method is shown in formula (6):

$$p_{u,i}=\bar{r}_u+\frac{\sum_{v\in N_u}sim(u,v)\times\left|r_{v,i}-\bar{r}\right|}{\sum_{v\in N_u}sim(u,v)} \tag{6}$$

Where $p_{u,i}$ represents rating of user *u* prediction to item *i*, $\bar{r}_u$ represents the average rating of the user, N_u represents the neighbor sets of the user *u*, *sim(u, v)* represents the similarity of *u* and *v*, $r_{v,i}$ represents the rating of user *v* to item *i*.

(4) Producing the recommendation result in accordance with the final rating.

4. The Experiment Results and Analyses

4.1 The Datasets.

This paper verifies the result of algorithm by the MovieLens datasets. The MovieLens datasets has three different editions, this paper select the smallest datasets. The datasets includes rating of the 943 users and 1682 movies. The datasets provides five groups of the training set and testing datasets.

4.2 The Evaluation Criterion.

This paper calculates the predictive error by MAE. Let $r_{u,i}$ to be the actual rating of the user *u* to item *i*, $\hat{r}_{u,i}$ represents the predictive rating of the recommend algorithm. The calculation expression of MAE is shown in formula (7):

$$MAE = \frac{1}{n}\sum_{i=0}^{n}\left|r_{u,i} - \hat{r}_{u,i}\right| \tag{7}$$

4.3 The Comparison of Simulation Result.

As shown in Figure 1, in the case of different neighbor users, the MAE value of the proposed similarity calculation is obviously smaller than the Pearson, COS, and Tanimoto. The MAE value of the improved algorithm is lower than the traditional Pearson algorithm.

It reflects simulation result of five groups of the training datasets and testing datasets in Figure 2. The horizontal axis represents the datasets number, and the vertical axis represents the average MAE value of datasets. As shown in Figure 2, the improved algorithm has good stability.

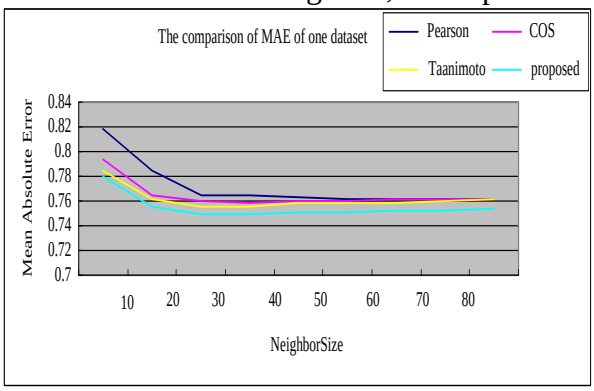

Fig.1 Comparison of MAE of same datasets

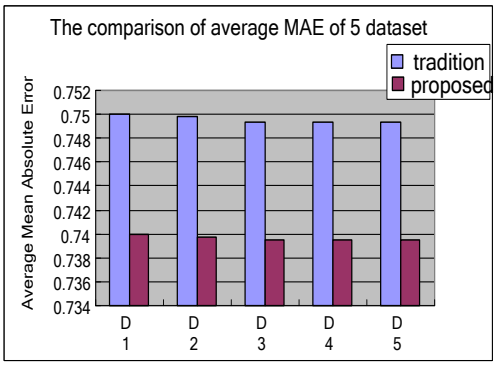

Fig.2 Comparison of average MAE of different datasets

From the simulation result, it shows that the improved algorithm have improved recommendation quality and recommendation accuracy.

5. Summary

The collaborative filtering recommendation algorithm searches the neighbor uses by similarity calculation, and implements recommendation through interest degree of the neighbor users to the target users. This paper introduces proportion of co-rating items and the average rating factors to the improved similarity calculation method, and proposes a novel collaborative filtering recommendation algorithm. The simulation experiment shows that the algorithm can effectively relieve the problem of incorrect recommendation result caused by the traditional similarity calculation. By applying the algorithm to film recommendation system, the system can improve recommendation quality.

References

[1]. Gold D. Nichols D. Using Collaborative Filtering to Weave an Information Tapestry. Communications of the ACM. Vol. 35(1992) No 12, p.61-70.
[2]. Li H, Wang G. L. A Novel Similarity Calculation for Collaborative Filtering, Proceeding of the 2013 International Conference on Wavelet Analysis and Pattern Recongnition, 2013,p.14-17.
[3]. Shang M S, Zhang Z K. Collaborative Filtering with Diffusion-based Similarity on Tripartite Graphs, Statistical Mechanics and its Application, Vol.6(2010), No 6, p.1259-1264.
[4]. Zhu X Z, Tian H. Persionalized Recommendation with Corrected Similarity, Journal of Statistical Mechanics Theory &Experiment, Vol.7(2017) No 10, p.1111-1120.
[5]. Herlocker J, Konstan J. An Empirical Analysis of Design Choices in Neighbor-based Collaborative Filtering Algorithm, Information Retrieval Journal, Vol. 5(2002) No 4,p.287-320.