A Collaborative Filtering Algorithm Based on Mixed Similarity

Yang ZOU¹ and Ying-ding ZHAO²,*

¹Department of Computer and Information Engineering, Jiangxi Agricultural University, China
²Department of Software, Jiangxi Agricultural University, China
*Corresponding author

Keywords: User interest, Item popularity factor, Data sparsity.

Abstract. The traditional collaborative filtering algorithm ignores the influence of user interest and item popularity in the calculation of similarity, which will lead to inaccurate calculation of similarity in terms of sparse data. This paper proposes a collaborative filtering algorithm which is based on mixed similarity algorithm. Both user interest similarity and item popularity similarity are taken into consideration. Results show that the proposed algorithm significantly reduces the calculation error comparing with the traditional algorithm. At the same time, the problem of data sparsity has been alleviated.

Introduction

At the age of information explosion, recommendation systems have been widely employed to filter undesired information for users. Collaborative filtering algorithm is one of the most widely employed algorithms in the recommendation system. However, there are always some problems including sparsity, accuracy and cold start problems in traditional collaborative filtering algorithms. In order to solve the problems mentioned above, domestic and foreign scholars have done a lot of meaningful research. Li Ma et al proposed Pearson correlation coefficient to smooth and cluster the data. The data sparsity problem has been apparently improved [1]. Besides, Haihui Huang et al integrated the score difference and user interest into the similarity calculation. Results show that the proposed algorithm not only reduces deviation but also improves accuracy [2].

In this work, a mixed similarity is proposed to improve the collaborative filtering algorithm. Both user interest similarity and item popularity similarity are taken into consideration. The accuracy of the calculation results can be improved while the problem of data sparsity can be alleviated by mixed similarity.

Theory of the Algorithm

User Interest Similarity and Item Popularity Similarity

The user interest similarity is an important factor when rating a certain item. It will directly affect the results of rating. Therefore, they should be considered when calculating the similarity. Expressions of user interest similarity are as follows:

\[ H_{u,f} = \frac{N_{u,f}}{N_u} \] (1)

\[ SimI(U_x, U_y) = \frac{\sum_{f \in \Omega} H_{x,f} H_{y,f}}{\sqrt{\sum_{f \in \Omega} H_{x,f}^2} \sqrt{\sum_{f \in \Omega} H_{y,f}^2}} \] (2)
where \( Nu \) refers to the total number of the rating for user ‘\( u \)’. \( Nu,t \) denotes the number of rating in terms of category ‘\( t \)’ for user ‘\( u \)’. \( Ux \) and \( Uy \) refer to user ‘\( x \)’ and ‘\( y \)’. \( Hx,t \) and \( Hy,t \) represent the rating of user ‘\( x \)’ and ‘\( y \)’ in terms of category ‘\( t \)’.

Traditional collaborative filtering algorithm will lead to higher similarity due to item popularity. This is because the popularity of a certain item will affect users’ rating. As a result, item popularity should be taken into consideration when calculating the similarity. This paper introduces the item popularity factor into the calculation in order to characterize the influence brought by item popularity. Expressions of item popularity similarity as follows:

\[
SimH(U_x, U_y) = p \times SimP
\]  

(3)

\[
p = \frac{1}{\log(1 + Nu(i))}
\]  

(4)

where \( SimP \) refers to the Pearson similarity [3]. \( Nu(i) \) represents the number of users who rated item \( i \).

### Mixed Similarity

In the case when the rating data set is sufficient, traditional similarity calculation method can accurately calculate the similarity between users. However, without sufficient data set, the data sparse will become observable. As a result, the traditional similarity calculation method is difficult to accurately characterize the similarity between users. The accuracy of prediction will be reduced. In order to solve this problem, this paper proposes a mixed similarity that comprehensively considers the influence of user interest and item popularity. A weighted factor was employed in this work to combine the influence of them. The expression of the mixed similarity is shown in Eq.(5).

\[
HSim = \theta \times SimH(U_x, U_y) + (1 - \theta) \times SimI(U_x, U_y)
\]  

(5)

where \( \theta \) is the weighted factor. \( SimH \) and \( SimI \) refer to the user interest similarity and item popularity similarity.

### Evaluation of the Proposed Algorithm

#### Experiment Dataset Description

The experimental data is based on the Movie Lens dataset provided by the Group Lens project team of the University of Minnesota, USA, which is widely used in the simulation experiments to evaluate the recommendation system. It contains 943 users and 1682 movies, each user has at least 20 ratings, the rating is ranging from 1 to 5. This dataset consists of total 100004 ratings and the data set sparsity is 93.7%.

#### Experimental Procedure and Evaluation Criterion

In this experiment, for reducing data contingency, 80% of the data in the data set were randomly selected as the training set. The training set was then divided into five parts randomly. Each part is performed separately in order to improve the reliability of the data. Then, the final result was achieved by the statistical mean. In terms of the other 20% of the data, they were chosen as the test set.

Based on the training set mentioned above, three tables including movie-sorted table, user-movie table and movie rating table were established. They contain the information of the user, users’ rating and movie item. With the calculation method in Eq.(1-2) and (3-4), the user interest similarity and item popularity similarity were calculated separately. By repeating iteration, the weighted factor \( \theta \) can be extracted. Then, the mixed similarity can be achieved. Built on the calculated similarity, the
neighbor set can be extracted. Finally, the prediction of the item rating can be achieved with the neighbor set. The flowchart of the experimental procedure is presented in Figure 1.

The evaluation of the algorithm mentioned above is based on the metric absolute mean error (MAE) and standard error (RSME) [4]. MAE is also called the average absolute dispersion, which can intuitively reflect the actual prediction error. RSME is also known as root mean square error, which can well reflect the degree of dispersion of a data set. The expressions of them are shown as follows.

\[
MAE = \frac{1}{N} \sum_{i=1}^{N} |R_{x,i} - R'_{x,i}|
\]  

(6)

\[
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (R_{x,i} - R'_{x,i})^2}
\]  

(7)

where \(R_{x,i}\) is the predicted score of user \(x\) for item \(i\) in the recommendation algorithm, \(R'_{x,i}\) is the true score of user \(x\) for item \(i\), and \(N\) is the number of neighbor sets.

**Results and Discussion**

The value of parameter \(\theta\) is significant for mixed similarity. The extraction of \(\theta\) can be performed by experimental iteration method. The value for \(\theta\) was ranging from 0.1 to 1.0 stepped by 0.2 in this work. The experiment result with different values of \(\theta\) are presented in Figure 2.

![Figure 2. MAE and RSME calculated with different \(\theta\).](image)

It can be seen from Figure 2. The MAE and RSME reach the minimum value when \(\theta\) is 0.1 and 0.3. The calculation of mixed similarity was performed separately when \(\theta\) is equal to 0.1 and 0.3. Results of the mixed similarity are presented in Figure 3. Pearson in Figure 3 refers to the traditional algorithm.
Figure 3. Comparison between the traditional and improved algorithm: (a) MAE and (b) RSME.

It’s clear that the mixed similarity calculation method in this work is superior to the traditional one in both MAE and RSME. This means that the problem of data sparsity has been alleviated while the accuracy of the algorithm has been improved at the same time. Besides, it seems that the lower value of $\theta$ is, the higher accuracy the improved algorithm can achieve.

**Summary**

This paper proposes a collaborative filtering algorithm which is based on mixed similarity algorithm. Both user interest similarity and item popularity similarity are taken into consideration to solve the problem the traditional algorithm which ignores the influence of user interest and item popularity in the calculation of similarity. Results show that the proposed algorithm significantly reduces the calculation error comparing with the traditional algorithm. At the same time, the problem of data sparsity has been alleviated.

**Acknowledge**

The research reported in this paper is supported by the Jiangxi Province Graduate Innovation Special Fund Project (YC2017-S190).

**Reference**


