Collaborative Filtering based on demographic attribute vector

Tian Chen  
East China Normal University  
Shanghai, China  
tchen@ ica.stc.sh.cn

Liang He  
East China Normal University  
Shanghai, China  
lhe@ ica.stc.sh.cn

Abstract—In present recommender systems, users receive items recommended on basis of their purchase records. New user experiences the cold start problem: as there records is very poorly. This paper proposed an NCT/TF(number of common terms / term frequency) collaborate filtering Algorithm Based on demographic vector. First, generates user demographic vector base on the user information (age, occupation, gender), then calculate two users similarity base on previous result, and generate new similar by combine it with cosine or PCC similar, And then predict item rates by top N similar neighbors. The experiments show that the quality of recommendations improved, while the new user effort is smaller as no initial rating are asked.

Keywords: recommendation system; Collaborative Filtering; Cold start problem

I. INTRODUCTION

The amount of information in the world is increasing far more quickly than our ability to access it. Now it is time to create the technologies that can help us sift through all the available to find that which is most valuable to us. Fortunately, recommendation systems technology offers a potential solution to this problem. Nowadays, recommendation systems have become an important research area and there has been much work done both in industry [3][11] and academia[4] on developing new approaches to recommender systems over the last decade.

At present, recommendation systems technology usually classified according to their information filtering approach as below three types:

1. Content-based filtering recommends items similar to those the user preferred in the past. However, content-based filtering has some limitations[4]: It is hard for content-based filtering to pro-vide serendipitous recommendations, be-cause all the information is selected and recommended based on the content.

2. Collaborative filtering recommends[4] items that uses with similar preferences have liked in the past. Collaborative filtering system overcomes some limitations of content-based filtering. The system can suggest items (the things to be recommended, such as books, music etc.) to users and recommendations are based on the ratings of items, instead of the contents of the items, which can improve the quality of recommendations. Although collaborative filtering can improve the quality of recommendations based on the user ratings, it completely denies any information that can be extracted from contents.

3. Hybrid filtering approaches combine content-base and collaborative filtering methods in several different ways [4]. It is obvious that the content-based filtering does not suffer the above problems. So it is a natural way to combine them in order to achieve a better performance of filtering, and take the advantages of each.

In present recommender system, users receive items on the basis of the ratings they have already provided. When a new user connects to the system, his/her list of ratings is empty, which makes it difficult for the system to provide recommendations. This problem is one of the well-know cold start problems, called new-user problem [12].

This paper proposed a method to cold start problem. And this algorithm also improved the recommendation performance when user's rating sparsely. This method focus on the calculation of new users' similarity. We calculates similarity between users base on there demographic attribute and ratings of items separately. And new similarity is generated by combine the previous results. At last, we find Top N neighbors base on new similarities and suggests items. The experiment shows our approach achieve good performance while new users have no rating or ratings very sparsely.

The reminder of this paper is organized as follows. In section2, we provide an overview of several major approaches for calculates similarity. Section 3 shows the method of similarity computation. Simulation of our approach is presented in section 4, followed by a conclusion in section 5.

II. RELATED WORK

In this section, we review collaborative filtering and several major approaches for calculates similarity between users. Three methods of similarity calculations are widely used in collaborative filtering: Pearson, cosine and adjusted cosine.

A. Collaborative filtering

Give a recommendation system consists of M users and N items, the relationship between users and items is denoted by an M×N matrix, called user-item matrix. Every entry in this matrix $r_{ij} = 0$ represents the score value, $r$, the user i rates an item j. When user i does not rates the item j, then $r_{ij} = 0$. Table 1 is the user-item rating matrix.
Measure the similarity between user i and user j, first of all, we should get the items which user i and user j were rating. Then measure similarity computation using different method. Sim(i,j) denotes the similarity between user i and user j.

### B. Traditional Similarity Computation

Similarity between two users is computed by the items which have been rate and then applying a similarity computation technique. As a rule, there are three common techniques:

1. **Cosine-based similarity**
   - The similarity between user i and user j is measured by computing the cosine of angle between different vectors, which denoted by Sim(i,j) is given by:
   \[
   \text{sim}(i, j) = \frac{\sum s_{i,j} (r_{i,s} - \bar{r}_i)(r_{j,s} - \bar{r}_j)}{\sqrt{\sum s_{i,j} (r_{i,s} - \bar{r}_i)^2} \sqrt{\sum s_{i,j} (r_{j,s} - \bar{r}_j)^2}}
   \]

2. **Pearson Correlation Coefficient**
   - Collaborative filtering engaging PCC was used in a number of recommendation system [4], since it can be easily implemented and can achieve high accuracy when comparing with other similarity computation methods. PCC[4][5][6] is employed to define the similarity between two, user i and user j based on the items they rated in common:
   \[
   \text{sim}(i, j) = \cos(\hat{\vec{r}}_i, \hat{\vec{r}}_j) = \frac{\hat{\vec{r}}_i \cdot \hat{\vec{r}}_j}{\|\hat{\vec{r}}_i\| \cdot \|\hat{\vec{r}}_j\|}
   \]

3. **Adjusted Cosine Similarity**
   - Computing similarity using basic cosine measure has one important drawback—the differences in rating scale between different users are not taken into account. The Adjusted cosine similarity offsets this drawback by subtracting the corresponding user average from each co-rated pair. Formally, the similarity between user i and user j using this scheme is given by:
   \[
   \text{sim}(i, j) = \frac{\sum s_{i,j} (r_{i,s} - \bar{r}_i)(r_{j,s} - \bar{r}_j)}{\sqrt{\sum s_{i,j} (r_{i,s} - \bar{r}_i)^2} \sqrt{\sum s_{i,j} (r_{j,s} - \bar{r}_j)^2}}
   \]
   Here $\bar{r}_i$ is the average of the i-th user’s ratings.

### C. Existing Problems and Our Contributions

In traditional similarity method, the user has to rate a number of items. Therefore, a new user, having no rating, would not be able to calculate similarities. Besides, if a user's rating records sparsely, the recommendation performance is very poor. [1][4].

Several techniques have been proposed to address this problem. Vozalis and Margaritis[7]propose an approach enhancing the collaborative filtering, which compute the user similarity based on demographic data. This paper bring forward the character which use SVD technical to decompose demographic data, at the same time, it enhance the collaborative filtering with the method computing the cosine similarity. However , SVD-based ethnologies will lead to information loss[8],moreover, it is not appropriate to compute the similarity among the m dimension vector the which consisted with discrete values using cosine similarity. For example, it can not calculate the similarity with cosine similarity or PCC method on the data collected by the MovieLens, which numerical value are all keywords character except the age.

[1][9] proposed a method to improve the collaborative filtering on the ways of increasing the number of projects users score together when the user-score are sparse using the score predictable manner. However, this method can not apply to the situation of the user without score.

### III. Collaborative Filtering based on demographic attribute vector

In order to calculate this user's similarity with other users without their rating records, this paper proposed a method base on demographic attribute vector. This approach is based on an assumption: If two people are similar in the age, occupation, income, gender and other physical attributes, then they may both have a common interest in the same item. The algorithm is described as follows:

1. The information users filed out the registration, including age, sex, occupation, self-description, can be collected to create a user profile.
2. Establish user's attributes keyword set, then according to the users’ registered information and through a series of rules, users can be described by the keyword in the set.
3. Uses our method to calculate the user's similarity on demographic between two users by the common keywords and its weights.
4. Combining the above calculated similarity and traditional approaches (e.g. PCC or cosine), a new similarity can be generated.

The third step is the focus of this paper.

**A. Collect users’ information and generate a users’ attributes keywords set.**

This step is completed in user registrations. User information can be classified into two categories: specific information and vague information. Different methods have been taken to deal with these two types of information.

1. **Specific Information**
   - Specific information includes user’s occupation, age, income level, gender and so on. This information we can describe them by keywords easily. For instance, a user may

<table>
<thead>
<tr>
<th>User1</th>
<th>Item1</th>
<th>Item2</th>
<th>Item3</th>
<th>Item4</th>
<th>Item5</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1,1</td>
<td>×</td>
<td>r1,3</td>
<td>r1,4</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>User2</td>
<td>×</td>
<td>r2,2</td>
<td>r2,3</td>
<td>×</td>
<td>r2,5</td>
</tr>
<tr>
<td>User3</td>
<td>r3,1</td>
<td>r3,2</td>
<td>×</td>
<td>r3,4</td>
<td>×</td>
</tr>
</tbody>
</table>

Figure 1. user-item rating matrix
have the following set of keywords attributes: (young man, teacher, master and high-income).

2. Vague Information

Vague information refers to the user’s self-described document information, such as self-introduction, self-evaluation.

For this information, a keyword extract technology is adopted: we prepare a keyword database \( Z \), extracting the keywords set in their vague information that the keyword in database \( Z \). As to the selection of keyword database \( Z \), we can only choose keywords which are relevant.

B. Calculate similarity of demographic attribute

Step 1. Calculates the common keywords set between two users.

Let \( \text{profile}_i \) denotes the keywords set of user \( i \), \( \text{profile}_j \) denotes the keywords set of user \( j \), and \( \text{profile}_{ij} \) denotes the common keywords set between user \( i \) and user \( j \). Their relationships can be defined as \( \text{profile}_i = \text{profile}_j \cap \text{profile}_{ij} \) and \( \text{profile}_{ij} = (w_1, w_2, ..., w_k) \& \{ \forall k \mid 1 \leq k \leq x \& w_k \in Z \} \)

Where \( Z \) denotes the set that contains all the keywords.

Step 2. Assume that \( x_w \) denotes the keyword in \( \text{profile}_{ij} \). Then \( \text{Vector}_{ij} \) can be defined as a vector of weight \( (TF_1, TF_2, ..., TF_x) \), where each weight \( TF_x \) the importance of keyword \( w_x \) to similarity between \( i \) and \( j \). \( TF_x \) is given by:

\[
TF_x = \log(\frac{\max_x}{f_x})
\]

Where \( f_x \) denotes the number of times the keyword \( w_x \) appears in all users profiles. \( \max_x \) denotes the maximum number of times the keyword appears in all users' attributes keywords sets, that is:

\[
\max_x = \text{Max } \{ \forall f_k \mid 1 \leq k \leq x \& w_k \in Z \}
\]

\( TF_x \) denotes the importance of \( w_x \) to similarity between \( i \) and \( j \). The larger \( TF_x \), is, the greater impact \( w_x \) has on the similarity of two users.

This formula means if a keyword \( w_x \) is very prevalent as it appears in almost every users' attributes keywords set, then its \( TF_x \) is smaller. \( TF_x \). On the contrary, if a keyword \( w_k \) in two users' profile has a low frequency of appearance in all users' profile sets, then it has a larger \( TF_x \). In extreme cases, if a certain keyword \( w_k \) appears in every user's profile set, then the value of its \( TF_x \) will be 0.

Step 3. Calculates \( Sim_{Dom} \) base on \( Vector_y = (TF_1, TF_2, ..., TF_x) \) by following formula:

\[
Sim_{New,y} = \sum_{w_x \in \text{Z}} TF_x
\]

C. Calculate similarity by combine the result of step 2 and traditional similarity.

As the sparse of users' ratings will cause low accuracy of recommendation, when there are few ratings from new users we'd generates \( Sim_y \) using the following formulation:

\[
Sim_y = \frac{\lambda \cdot Sim_{New,y} + (1-\lambda) \cdot Sim_{Col,y}}{Sim_{New,y} + Sim_{Col,y}}
\]

where \( \lambda = \frac{k}{N} \) for \( 0 \leq k \leq N \) \& \{ \lambda = 0 \mid k \geq N \}

When \( k \) denotes the total number of items which have had rated by user \( i \). \( N \) is a threshold , which determines whether to use the demographic information.

\{ \lambda = 0 \mid k \geq N \} means that if the number of ratings are more than \( N \) by user \( i \). \( Sim_y \) equals to \( Sim_{Col,y} \).

D. Calculate nearest neighbor set (NBS) and suggest prediction.

Step 1. Calculates the nearest Top N neighbor set.

According to \( Sim_y \) values, select the nearest neighbors set \( C = \{ C_1, C_2 \} \), here user between active user a have the maximum similarity, \( C_2 \) is the second, and so on.

Step 2. Through the above step, we get the active user’s nearest neighbor set (NBS). Next we need to predict for active user. This method predicts the rating on an item \( k \) for user \( a \) by computing the weighted sum of the ratings given by the user similar to user \( a \). Formally, using the notion \( P_{i,k} \) denotes the prediction as:

\[
P_{i,k} = r + \frac{\sum_{j \in \text{NBS}} \text{sim}(i, j) \times (r_{j,k} - \bar{r}_j)}{\sum_{j \in \text{NBS}} (| \text{sim}(i, j) |)}
\]

\( \text{Sim}_y \) is the similarity between user \( i \) and user \( j \). \( r_{j,k} \) denotes the score user \( j \) rates on item \( k \), \( \bar{r}_j \) represents the average rate of user \( j \).

IV. Experimental Evaluation

A. data set

This study uses MovieLens data set, 100K dataset and 1 million dataset, for the analysis. 1 Million dataset and 100k dataset are composed of 1,000,209 and 100,000 of rating respectively and they are the preference values rated by 6040 and 943 users who were asked to rate their movie preference form the degree 1 to 5 over 3952 and 1682 movies. The data
set provided by the GroupLens consists of three subordinate data sets: ratings, users, and movies. The rating data set contains users, movies, ratings, and timestamps and the other two data sets contain the demographical explanation about users and the explanation of movie genres. In this research, both datasets are used. The datasets used were divided into 80% of the training set and 20% of the test set respectively.

B. Metrics

We use the Mean Absolute Error (MAE) metrics to measure the predict quality of our proposed approach with other collaborative filtering methods. MAE is defined as:

$$\text{MAE} = \frac{1}{N} \sum_{u,i} |R_{u,i} - \hat{R}_{u,i}|$$

Where $R_{u,i}$ denotes the rating that user u give to item i, and $\hat{R}_{u,i}$ denotes the rating that user u gave to item i which is predicted by our approach, and N denotes the number of tested ratings.

C. Compare

To verify the effectiveness of our approach, we compare it with traditional collaborative filtering algorithms. In traditional collaborative filtering recommendation algorithms, we use cosine and PCC similar method as similarity computation. We calculated MAE, and compare with our proposed approach, the number of neighbors are increase from 20 to 100, the experimental result as shown in Figure 2 and Figure 3:

Figure 2. Comparison of accuracy of recommendation algorithms

As shown in figure 2, all the multiples of 10 between 20 and 80 are selected as the number of nearest neighbors and then we calculate the average of their MAE. Vertical coordinates denote the value of MAE, and the abscissa indicates the magnitude of the selected N. From the simulation. It can be known that when 8 or 9 is selected as N, our approach achieves best performance. And we also find out that when N is greater than a threshold, the cosine and PCC are better than our approach. All these results illustrate that it is not best when N is selected as the largest number, while N should be an appropriate number. Here, 9 is the best.

Figure 3. The relationship of MAE and the number of Neighbors

D. Analysis

Learn from the simulation, the algorithm proposed in this paper enhances the average performance of the system in a certain way in the cold start-up phase. However, two points must be noticed when this algorithm is used:

1) N need to be a certain value as to the specific needs of the system. The suggested value is a number between 8-10 because: with the users selected points increase, the performance of the collaborative filtering gradually enhanced, but the algorithm proposed in this paper does not increase with the growing scores.

2) The experimental results show that the improvement of the algorithm accuracy is proportional to the numbers of nearest neighbors. However, taking into account the cost of computation, the value selected between 50-70 would be better. This is probably because these neighbors have been able to represent target users’ preferences to the items

V. CONCLUSIONS

As users experience the widely existed cold start problem in the present recommender systems, this paper proposed an NCT/TF(number of common terms / term frequency) collaborate filtering algorithm based on demographic vector decompose. Algorithm is based on an assumption: suppose that some users have similar attributes of the population, and then they may have a common interest in some items. First of all, collecting users’ description
keywords set, and in accordance with this we can compute users attribute similarity. Then new similarities can be produced by combining the above calculated similarity and traditional collaborative filtering as to certain rules of the PCC algorithm. Finally, selecting the nearest neighbor and recommending it. The experimental results show that this method can improve the recommendation accuracy of the system to some extent when the cold starts.

The proposed algorithm has certain dependence: the improvement of the recommendation accuracy depends on the fullness of the users’ registration information. The more information users provide, the more adequate the algorithm received the recommendation accuracy. This paper used MovieLens given the user's personal information provided only three types of users’ information: occupation, gender and age. Thus this is a major shortcoming for the improvement of the algorithm.

REFERENCES