Recommendation Systems Using Hybrid Collaborative Filtering

T.Aravindharaj, T.Rohit, P.Richard Antony Raj, M.Romuald Belwin
TRP Engineering College ,Trichy.

Abstract: Personalized Recommendation on any product or entity is a primary requirement for any commercial website. Collaborative filtering is one of the most successful techniques available. However the biggest issue which still remains is data sparsity. Data sparsity causes problems such as low prediction rates and poor recommendation quality. To resolve this problem we use content based filtering recommendation for new users until we obtain an index of the user's taste. Based on nearest neighbour's the weighted slope algorithm is used to predict the user's likings and present it as a recommendation to the user. This method can greatly increase the accuracy of the prediction.

Keywords: Personalized Recommendation , Collaborative filtering, Data Sparsity, Weighted Slope one algorithm.

1 Introduction

With the development of internet and e-commerce shopping online has become a popular way of consumption since more and more people rely on a series of online services. However, the number of commodities is growing increasingly. In order to help people find their desired commodities recommendation systems were introduced. Recommendation systems can provide items which each user is interested in. The core issue of recommendation is to find out the item which may interest the user. Personalized recommendation algorithms have good prospects but these systems are facing problems such as data sparsity , cold start, scalability, and other challenges. With further research these questions have been resolved to certain extent and have created satisfactory recommendations in movie, music and other fields. The present recommendation algorithms are mainly divided into four kinds : content based ,Knowledge based algorithms and collaborative filtering algorithms. Among these filtering algorithms collaborative filtering algorithm is the most widely used and effective method. Slope one is a simple collaborative filtering algorithm, it is based on linear regression model and utilizes rating deviations to predict ungraded items. However Slope one algorithm is not better than traditional collaborative filtering algorithm, when the data is sparse , so this algorithm cannot resolve data sparsity problem. Many improved slope one algorithms have been proposed to improve accuracy in recommendation. So to resolve this conflict we use content based filtering algorithm until the data sparsity is resolved and then we revert to the original collaborative filtering algorithm. We suggest that this would give an insight to new users and also increase the accuracy of recommendation.
2 Similarity between users

In order to find user’s neighbors, this study uses the collaborative filtering method to calculate the similarity between users and to find the nearest neighbor set. Collaborative filtering is mainly classified into two types : model based collaborative filtering algorithm, memory based collaborative filtering algorithm. The latter algorithm includes user based and content based collaborative filtering algorithm. So the first step of collaborative filtering algorithm is to find the users who are similar to each other, then do the recommendations to the user. The aim of similarity calculation is to find similar users and to get the neighbor set based on the value of similarity. The similarity is particularly important for that it affects the quality of the recommendation. Finding the similarity follows the following method: First the rating matrix \( R(m,Q) \) of users is defined, where \( m \) denotes the number of users and \( n \) represents the number of items . \( R_{i,j} \) is the rating of the item rated by the user \( I_i \).

1) Cosine Correlation

The user’s rating are seen as a vector of \( n \)-dimension spaces on the project, user \( I_i \) and user \( I_j \) where \( I_i \cap I_j \). If the user has no ratings on the items , the ratings are defined as zero, soothe similarity calculation between user \( I_i \) and user \( I_j \) is as follows:

\[
\cos(d1, d2) = \frac{(d1 \cdot d2)}{||d1|| ||d2||}
\]

Here the numerator is the inner product of two user’s rating vectors , the denominator is the product of two user’s rating norm of vectors

2) Adjusted cosine vector similarity

\( I_i \) is the item set which is user \( I_i \), \( I_j \) is the item set rated by the user, so the formula is as follows:

\[
sim(i,j) = \frac{\sum_{u \in U}(R_{u,i} - \bar{R}_u)(R_{u,j} - \bar{R}_u)}{\sqrt{\sum_{u \in U}(R_{u,i} - \bar{R}_u)^2} \sqrt{\sum_{u \in U}(R_{u,j} - \bar{R}_u)^2}}
\]

Where \( R_i \) is the average rating of user \( I_i \) for all co-rated items, \( R_j \) is the average rating of the user \( I_j \) for all the co-rated items.

3) Pearson correlation

\[
r = \frac{n(\Sigmaxy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}
\]

Where \( R_i \) is the average rating of the user on the users set \( I_{i,j} \), \( R_j \) is the average rating of users set \( I_{i,j} \). The calculation of cosine correlation is simple but when the data is sparse it is not a good way to find the nearest neighbors. Pearson correlation can achieve better results compared with other methods , it was used by many recommendation systems . In this research, the similarities among users are calculated by Pearson correlation method.

3 Slope one algorithm

Slope one is a typical item-based collaborative filtering algorithm. The algorithm takes into account both the information from other users who rate the same items and the information of the other items rated by the same user, and it is based on the rating difference between items.

<table>
<thead>
<tr>
<th>Table1. Ratings for five items</th>
</tr>
</thead>
<tbody>
<tr>
<td>item</td>
</tr>
<tr>
<td>item1</td>
</tr>
<tr>
<td>item2</td>
</tr>
<tr>
<td>item3</td>
</tr>
<tr>
<td>item4</td>
</tr>
</tbody>
</table>
The basic Slope one algorithm [5][6][7] is derived from a simple linear model: $y = f(x) = x + b$, where $x$ is a variable presented the difference of ratings and $b$ is a constant, for example, the ratings of five items which are rated by user1, user2, user3, user4 are taken and are calculated as follows: $((3-1)+(2-2)) / 2 = 1$, $((3-1)+(2-3)) / 2 = 0.5$. The computational result of table 1 is: $((3+1)+(2+0.5)) / 2 = 3.25$, thus this result fill the vacant rating of user 4 rated to item i and item j could be defined as $dev_{ij}$.

4 Weighted slope one and user based collaborative filtering
Although Slope one algorithm performs well, it does not take the difference between users into account. Weighted Slope one algorithm and user-based collaborative filtering algorithm are adopted in this paper to solve this problem. An important step of this algorithm is to calculate the similarities of users. Pearson correlation is used to calculate the similarities between users and to select the most nearest neighbors of the target user. After that, the ratings of the unrated items are predicted and the recommendation results are given, based on the weighted Slope one algorithm. The number of neighbors is represented by the letter $k$, the value of $k$ gets from experiments, the nearest neighbors set is represented by $S(k)$, then the prediction formula is as

$$P(u_j) = \frac{\sum_{i \in S(k)} (dev_{ji} + u_i)c_{ji}}{\sum_{i \in S(k)} c_{ji}}$$

To get higher accuracy, the algorithm predict the blank ratings based on the nearest neighbor set, and the neighbor set is chosen according to the similarity between users, the analysis of this algorithm is as follows:
Input: the training data set, rating matrix $R(m \cdot n)$, the target user $u$, the target item $j$.
Output: the prediction rating:
1) First, the users who have rated the same items as the target user are found. Then the formula (3) is used to calculate the similarities between users. If the target user does not have the same rated items with others, the similarity is set to zero.
2) According to the similarities, the top $k$ users were added to the target user's nearest neighbor set $S(k)$.
3) The formula (7) is used to calculate the ratings $P(u_j)$ which are not rated by the target user. If the items were not rated by the neighbors, the user's rating will be replaced by the average value of the rated items of the user.
4) The top $N$ ($N=10$) items are recommended according to the ratings calculated by the above procedures.

5 Experimental evaluation
5.1 Data Set
The Movielens data set was used in the present experiment. The data set[8] has been widely used, and it consists of 100,000 ratings from 943 users on 1682 movies. In the data set, each user rated at least 20 movies. The ratings ranged 1-5 expressed the fondness of users to movies, 4 and 5 represents positive ratings, 1 and 2 represents negative ratings. The density of the data set is:

$$\frac{100000}{943 \times 1682} \times 100\% = 6.3\%$$

the lowest level of sparsity for the tests is defined as: $1 - 6.3\% = 93.7\%$. Hence one can see that the data set is very sparse. This experiment divided the data set into training
set and test set, and used 5-fold cross-validation. The data set was split up into five disjoint subsets, of which training set and test set ratio is 0.8/0.2. Training set contains history rating records, and recommendation is conducted based on it to the users of the testing set.

5.2 Metrics
The metrics for evaluating the accuracy of prediction algorithms can be divided into two main categories [9][10][11][12]: decision-support metrics and statistical accuracy metrics. Statistical accuracy metrics, Mean Absolute Error (MAE), was used to measure prediction performance of the proposed algorithm in the present study. The MAE is calculated by the difference between user's predictive score and the user's actual score, which is defined as:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} | P_i - R_i |$$

Where, $iP$ is the prediction rating, $iR$ is the real rating of the testing set, $n$ is the total number of ratings in the testing set, the smaller the value of MAE is, the better the prediction accuracy would be.

5.3 Get the Number of Neighbors
In the proposed algorithm, the chosen of neighbors' scale [13] affects recommendation quality. The number of the neighbors was represented by $k$, in order to select the appropriate value of $k$, different values of $k$ ( $k=2, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50$ ) were tested to predict the ratings accuracy results. Each $k$ value was tested five times and the average value of the 5 values was used as the final results. The results are shown in

![Comparison of MAE of different neighbors](image1.png)

From the Figure 1, as the number of neighbors increased, the MAE values decreased first and then increased when $k$ was greater than 10, which means that the accuracy of prediction was increased first and then decreased when $k$ was greater than 10. Therefore, value 10 was selected for the best $k$ value.

5.4 Predicting Results
In the end, the experimental results of MAE from the improved weighted Slope one algorithm were compared with MAE from weighted Slope one algorithm and user-based collaborative filtering algorithm, as shown in Figure 2

![MAE of different algorithms](image2.png)

As it can be seen, the MAE from the improved algorithm is lower than the basic weighted Slope one when $k=10$. The proposed algorithm alleviates the data sparsity problem. The experimental results show that the improved Slope one algorithm is more accurate than weighted Slope one algorithm.

6 CONCLUSIONS
The result of traditional Slope one algorithm is less accurate, because it does not take the
similarities between users into consideration. An improved weighted Slope one algorithm, which is based on the similarity between users, is proposed in the present study to improve the recommendation accuracy. Specifically, the most similar users are chosen as neighbors, then the weighted Slope one algorithm is used to fill the blank ratings. If the user does not have neighbors (k=0), the ratings of the unrated items are replaced by the average ratings. The experimental results show that the improved weighted Slope one algorithm is more accurate than weighted Slope one algorithm, and satisfactory results can be achieved under data sparse condition.

REFERENCES