CS54701 Information Retrieval

Project 2 Collaborative Recommendation Algorithm

Penghao Wang

wang1128@purdue.edu

March 22, 2016

1 Introduction

In this project, I developed different algorithms to make recommendations for movies. In the first method, memory-based collaborative filtering algorithm based on vector similarity method is develop. The related file is Project2.py. In the second part, a model-based collaborative filtering algorithm based on correlation-based Similarity is installed. The related file is Project2_method2.py. In the third part, I modified the memory-based model. The Project2_method3.py is using this method.

2 Functions of Program

All of the explanations of the functions is in the readme.md.

3 Accuracy

3.1 Method

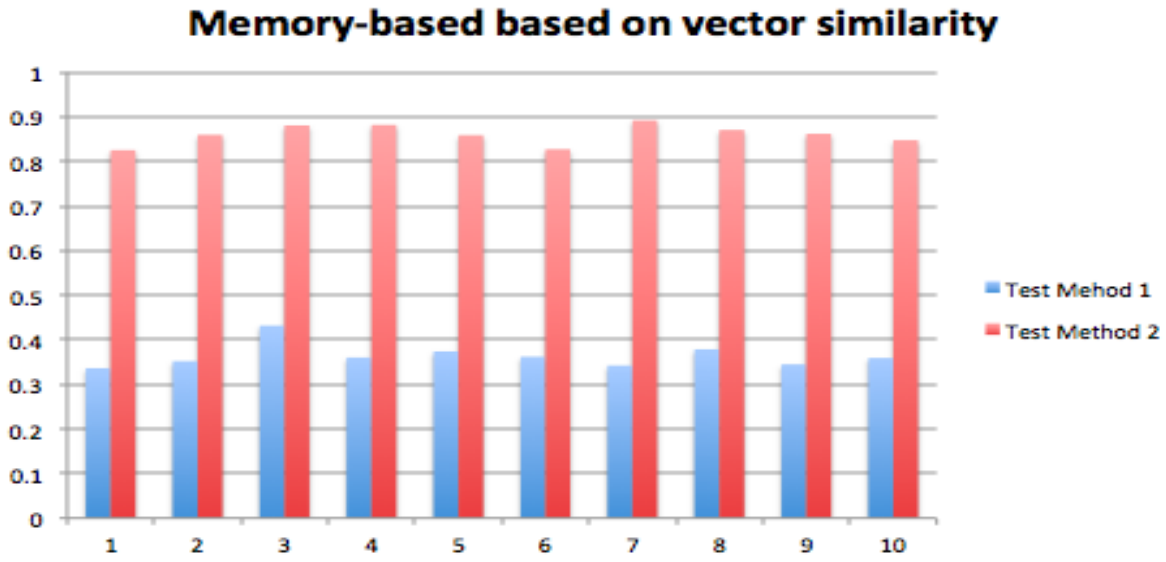
In this project, two measurement of accuracy are implement. The first is the strict standard. The second is the soft standard. In the beginning, predict every user rating about every movie. The way to test accuracy is to compare the existing rate and the prediction. In the first method, if the round of prediction is same as existing rate, then this prediction is correct. In the second method, if the round of prediction is equal to existing rating plus or minus one, then the prediction is count as right one. For instance, if the prediction is 3.2 and the original rate is 4. In the first method, the prediction is wrong. In the second method, the prediction is right.

3.2 10-fold cross validation

In the project, I use 10-fold validation to test the accuracy. I randomly divide the original data to 10 parts. Using the 9 parts for training and the other one part for testing. Then, I do ten times. Each time it will return a average of accuracy. The accuracy is the testing of one part. In three methods, I use one random seed. In that way, I could compare them.

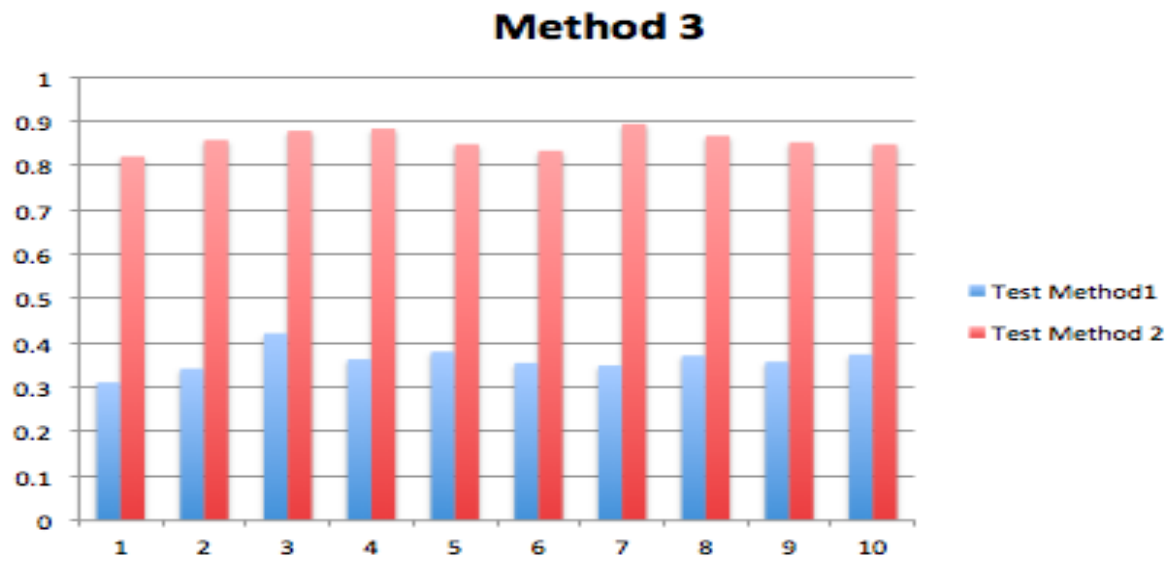
3.3 Memory-based based on vector similarity

The following picture represents the accuracy.



3.4 Method 3

From the paper, there are several extension of the memory-based model. I choose one of them. The method 3 is based on the correlation by using the memory-based model. The accuracy is similar with the Memory-based based on vector similarity. The correlation are useful if there are some users who votes everything good or bad. In this project, it do help increase some part of the accuracy. However, because there are only 200 users, there are few people how vote all things good or bad. In that case, it doesn't improve the accuracy a lot. The following picture shows the accuracy of 10 folds. The result is reasonable.

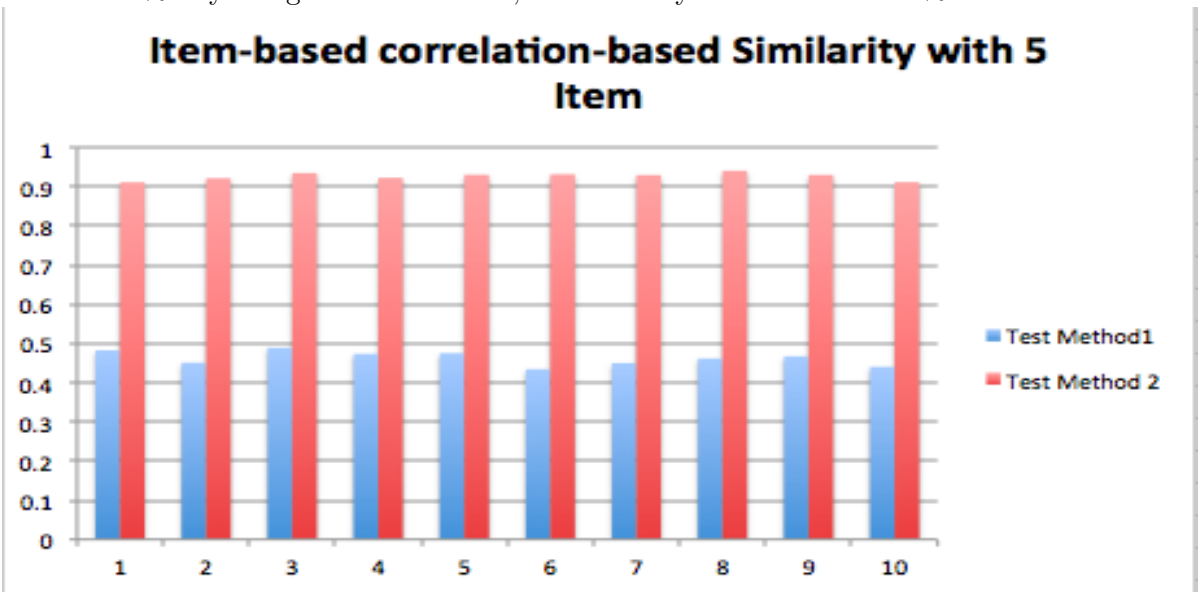


3.5 Item-based based on correlation-based Similarity

The Item-based model based on correlation-based Similarity performs better than Memory-based based on vector similarity. By using the strict method to evaluate, the average of accuracy is about 10% higher than the memory-based. By using the soft method to measure, the average of accuracy is about 6% higher than the memory-based. In the paper, the author thought 5 items are the best. In this project, the accuracy of 10-item method will be a 1% higher than 5-items.

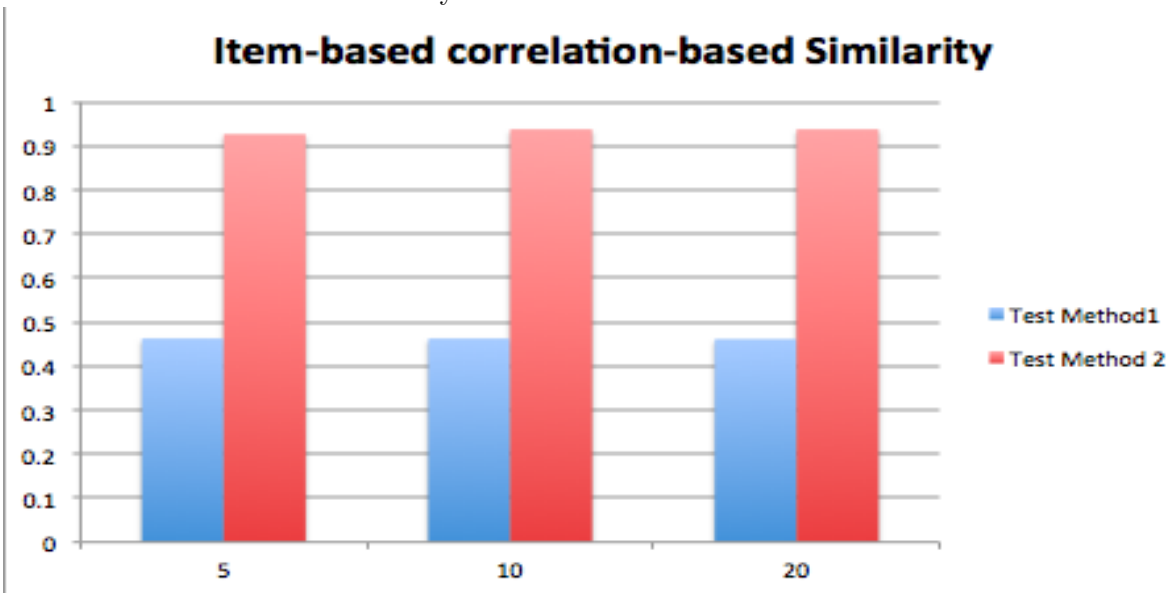
3.5.1 5-items

The following figure shows accuracy of item-based based on correlation-based similarity, where 5 items are selected to predict the rate. By using the strict method to evaluate, the average accuracy is about 46%. By using the soft method, the accuracy will be around 92%.



3.5.2 With different Items

The following figure show the accuracy based on 5-items, 10-items, and 20-items. There is not too much difference of the accuracy.



4 Efficiency

4.1 Method 1 and 3

The method 1 and method 3 is based on the memory based. The method 1 using 28 minutes and 20.6 second to test accuracy for all of the users, that is 2000 times of running the testAcc(). The method 3 is similar with method 3. It took a little bit longer because it need to calculate the correlation. It will take longer if the user increase.

4.2 Method 2

The method 2 is item based model. It takes about 15 minutes to calculate the similarity matrix and the prediction matrix I created. After I save them into a file. It takes 0.73 seconds to test all of the accuracy. Comparing with the method 1 and 3, it saves a lot of time.

5 Discussion

In this project, the item-based model is better than the memory-based model in accuracy and efficiency parts. The efficiency of method 1 and 3 will decrease when the number of users increase. However, the item-based model won't. However, it doesn't means that the item-based model is perfect. It has some disadvantages. It could not handle the cold start user. It performs bad if the user only rate several movies. If a user only rate three movies, the accuracy won't be high only based on these movies. The memory based model has the advantage on this part. The disadvantage

of the memory based model is the efficiency. The time will increase a lot if there are many users.