Project 3: Content Recommendation

1. Movielens Data Set

In this project, we use Movielens 1M data set (http://grouplens.org/datasets/movielens). The data set contains 1 million ratings from 6040 users on 4000 movies. The **README.txt** file describes the detailed information and format of the three files (**users.dat**, **movies.dat** and **ratings.dat**). All of such files are downloaded from the url http://files.grouplens.org/datasets/movielens/ml-1m.zip.

Note that in the **ratings.data** file, the ratings are sorted only by the UserID. For the ratings of each user, we further sort the ratings by the **Timestamp**. Based on the sorted ratings of each user, the 90% most earliest ratings are used as the training data set and the remaining 10% ratings as the test data set. For example, the 53 ratings of UserID 1 are as follow:

1::1193::5::978300760
1::661::3:: 978302109
1::914::3:: 978301968
.....
1::608::4::978301398
1::1246::4::978302091

After sorting the above ratings, we then have the following result:

1::3186::4::978300019
1::1270::5::978300055
1::1721::4::978300055
.....
1::1907::4::978824330
1::48::5::978824351

With the above 53 sorted ratings, the top 47 (=53*90%) most earliest ratings (ie., those with timestamp within the range between 978300019 and 978824268) are as the training data set, and the remaining 6 ratings with timestamp between 978824291 and 978824351 are as the test data set. With the above approach to choose the training and test data set, we utilize the training data set to predict the ratings in terms of the test data set. For example, we need to predict the ratings of the UserID 1 over the 6 MovieIDs 2355, 2294, 783, 1566, 1907 and 48 (we assume that the ratings of the UserID 1 and 6 MovieIDs are missed). Based on the predicted ratings and real ratings given by the test data set, we then validate the accuracy of the prediction by computing the root mean squared error (RMSE).

In the following tasks, we use the specific approaches over the training data set to predict the rating scores of those test data set, and then to compute the RMSE value.

2. Task Description

Use the following approaches to predict the rating scores of a specific user:

(a) The user-user based collaborative filtering
(b) The item-item based collaborative filter
(c) The approach in page 31 of Lec7-recommendation to enhance the above two approaches

Each student will use their student ID (say S) to select the specific user ID. Specifically, given the student ID S, the target user ID is selected by the result of (**S mod 6040**).

Once the user ID is selected, you then need to predict the rating scores of those movies appearing in the test data with respect to such a user ID.

During the above approaches, use the following metrics to measure the similarity.

(a) the Pearson correlation coefficient
(b) the cosine similarity

In addition, use three different numbers ||k|| to select the neighbors

(a) ||k||=2
(b) ||k||=10
(c) ||k||=50

3. What to submit

3.1 A text files to list the predicted rating score. The format follows the similar format as the **ratings.dat** (without the timestamp column bu with an extra column M):
userID::movieID::rating_2_p::rating_10_p::rating_50_p::rating_2_c::rating_10_c:: rating_50_c::M

- The above rating_[2|10|50]_[p|c] is the predicted rating score depending upon the different ||k|| =2, 10, 50 and similarity metric (=Pearson, Cosine).
- The above last column ***M*** indicates the method to compute the predicted score. The value of ***M*** is u, i, U, I. The four items of ***M*** indicate the user-user, item-item, enhanced User-user, and enhanced Item-item CF, respectively.
- Each user and movie pair is associated with four methods (or lines). Together with the different number of ||k|| and different similarity metrics, each user and movie pair is associated with 4*2*3=24 predicted scores with 4 lines and 6 scores per line.

3.2 A pdf document file to discuss and compare the RMSE value given the different similarity metrics and ||k|| numbers. The file should contain a bar figure with such metrics and numbers.
3.3 the r source code, the output file and saved R workspace file.

4. When and where to submit

4.1 Submission Deadline: 2014-12-31. If n (>=1) days are late to submit the file, the final score of this project is reduced by n*15 %.

4.2 All files are zipped with the naming conversion like proj3-xxx-xxx-xxx.zip, where each xxx indicates the student num. of a group member.