**COMP9417**

**ASSIGNMENT 2**

**TOPIC:9**

**Recommender System using collaborative filtering**

**Team member: z5044541 Wangyujie**

**Team member: z5044423 Chenhaoyu**

**1.Introduction**

Recommender system, as one of the technology which is applied to personal recommendation as well as Web data information prediction service, has been commonly equipped by most of companies. Great progress has been achieved, however, due to the enormous data growth, accuracy, high quality recommendation and efficiency are the most common challenges which are required to improve. One of those recommending technology is Collaborative filtering.

In this report, an item-based collaborative system are explored to predict the rating. The item-based method creates an user-item matrix in order to analyse the relationship between users and items, then only calculate the similarity items and provide recommendation to users.

**2: Method:**

**2.1 : Item similarity calculation**:

**2.1.1 cosine model**

In cosine model , we measures two items as vectors , the similarity in cosine model are the angel of these two vectors in this function. The scheme of the cosine model:

<> means the dot product of two vectors

|||| means the norm of two vectors

**2.1.1 Person correlation coefficient :**

In pearson model, Only inner items of two vectors need to be calculated, Assume be the items x, y both rated by user A . In each calculation step, average rate of two items must be reduced to prevent rate inflation which means users follows others rate ,and the rate of that item always stay high or stay low. The scheme as following:

Here, means the rate of x in S, avg r means the average rating of two items which means the mean(sum(column values)) in item matrix.

**2.1.2 Adjusted cosine similarity:**

In this case, the adjusted cosine similarity method measures that difference in rating scale by users are not taken into account. Adjusted similarity also measures if the data set is sparse, the method would preform better ,The scheme of adjusted cosine similarity is:

Here, means the rate of x in S, avg means the average rating of two users.

**2.2 Prediction:**

In this project, we use weighted Sum method after calculating all the similarities between two items, then given scheme could be use to predict the value for different user

Here sim(i, N) means the similarity between item i and item N , R(u,N) means the rate of N by users u

P(u,i) is an adjusted rated distributed between -1 and 1, in this case, p = -1 means rate 1 and p =1 means rate 5. We need to adjust R(u, N) to satisfy the demand by following scheme

means the rate of N by user u , Max rate means 5 and Min rate Means 1

**3: Result:**

**3.1.1 data size:**

In this project, we download data from “MovieLens/ml-100k.zip”, totally 100000 ratings from 943 users and 1682 movies, we choose 80% data as training set and 20% data as testing set to ensure the data size and there are totally 5 training set and 5 testing set to execute. we use 5-fold validation method to calculate the mean value of the 5 result.

**3.1.2 : prediction results:**

In this project, cousin, cosine Pearson and adjust-cosine are selected to execute, we calculate the 5 testing set and compare them with three methods and we use MAE(mean absolute errors) to describe the accuracy of three prediction method.

Here , means the value of predicted rate , and means the value of actual rate.

The following picture is the result of three models, prediction. The y label is the MAE value, the x-label means the 5-fold validation.



Figure(1)

**4: discussion:**

**4.1: the explanation of total number of prediction elements**

There are some situations that can not be predicted by pearson and adjusted-cousine model. (1): If we only find 1 inner element between two items which means the denominator in formula sim(x,y) is 0. We tag this element as unpredictable element. (2): If two items are rated by all users are 1, due to the same denominator reason as (1), we also tag these element as unpredictable element. After isolation these elements, the total number is between 19950 and 20000

**4.2: result analysis:**

the totally performance with three methods are in Figure(1) so that in this test during the 5–fold validation, each method could predict the rate with in reasonable interval, however, cosine method preforms the best among the three methods. It actually means these 5 training test do not have strong rate inflation as well as sparsification problems. In these 5 testing sets, cosine method is the best way to predict the result.

**5: extra function : recommend movie to users:**

In this project, we use three methods to predict the rating and conclude the max n number of rating movies to the user.

**自己贴一下。图**

**6：conclusion**

In this project, we use three methods to predict the rating of movies, Those three methods preformed good prediction of item based collaborative filtering predictions with training datasets. In 5-fold test, cosine preformed the best among three and could help to impove the accuracy of rate prediction. With the reasonable interval, this method could be an alternative choice for us to choose.

7.Reference:

1. Badrul Sarwar, George K, Joseph K, John R. Item-based collaborative filtering recommendation algorithms[R]. International Conference on World Wide Web, 2001. 285-295