

# A Collaborative Recommender Based on User Information and Item Information

SongJie Gong

Zhejiang Business Technology Institute, Ningbo 315012, China

Email: songjie\_gong@sina.com

**Abstract**—Collaborative recommender is the most popular recommendation technique nowadays and it mainly employs the user item rating data set. Traditional collaborative filtering approaches compute a similarity value between the target user and each other user by computing the relativity of their ratings, and they only consider the ratings information. User attribute information associated with a user's personality and item attribute information associated with an item's inside are rarely considered in the collaborative filtering recommendation process. In this paper, a hybrid collaborative filtering recommender is proposed which employs the user attribute information and the item attribute information. This approach combines the user rating similarity and the user attribute similarity in the user based collaborative filtering process and then it combines the item rating similarity and the item attribute similarity in the item based collaborative filtering process to produce recommendations. The collaborative filtering recommender employs the user attribute and item attribute can alleviate the sparsity issue in the recommender systems.

**Index Terms**—recommender system, collaborative filtering, rating similarity, user attribute similarity, item attribute similarity

## I. INTRODUCTION

With the development of Network, the problem of information overload is becoming increasing serious and we all have experienced the feeling of being overwhelmed. Many researchers and practitioners pay more attention on building a proper tool which can help users obtain resources and services which wanted. Personalized recommendation systems are used to help users obtain recommendations for unseen items based on their preferences, which are able to distinguish one user from another to provide information [1, 2]. The famous electronic commerce website Amazon and CD-Now have employed recommendation technique to recommend products to customers and it has improved quality and efficiency of their services.

The most techniques used in today's recommendation systems fall into two distinct categories: content-based methods and collaborative filtering methods [3, 4]. And collaborative filtering has been known to be the most successful recommendation techniques. Collaborative methods recommend items based on aggregated user ratings of those items and these techniques do not depend on the availability of textual descriptions. They share the common goal of assisting in the user's search for items of interest, and thus attempt to address one of the key research problems of the information age: locating

needles in a haystack that is growing exponentially. Collaborative filtering systems can deal with large numbers of people and with many different items. However there is a problem that the set of ratings is sparse, such that any two users will most likely have only a few co-rated items. The high dimensional sparsity of the user-item rating matrix and the problem of scalability result in low quality recommendations.

Traditional collaborative filtering approaches compute a similarity value between the target user and each other user by computing the relativity of their ratings, which is the set of ratings given on the same items. Based on the ratings of the most similar users, commonly referred to as neighbors, the algorithms compute recommendations for the target user. They only consider the ratings information. User attribute information associated with a user's personality and item attribute information associated with an item's inside are rarely considered in the collaborative filtering recommendation process. In this paper, a hybrid collaborative filtering recommender is proposed which employs the user attribute information and the item attribute information. This approach combines the user rating similarity and the user attribute similarity in the user based collaborative filtering process and then it combines the item rating similarity and the item attribute similarity in the item based collaborative filtering process to produce recommendations. The collaborative filtering recommender employs the user attribute and item attribute can alleviate the sparsity issue in the recommender systems.

## II. EMPLOYING USER ATTRIBUTE IN FOR THE USER BASED COLLABORATIVE FILTERING

### A. Analyzing the problem

In real world, user has owner demography not relation to the ratings and the information is important to the personalized recommendation system. For example, all users are required to register and to provide demographic information including sex, age, profession, department, specialty, etc. The demographic information of each user can be used to classify users that like similar categories or subjects of items. The specialty information is very useful for generating recommendations. The specialty field refers to the research field of the user. So if two users have the same specialty, they will have the same interest in some items. But, it is not reflecting in the user-item ratings. Therefore, if we know the specialty to which a user belongs, we can partially know which items the user

will be interested in. This relationship can be used to initialize the user preferences of a new user [5,6,7].

### B. User Attribute

We use MovieLens collaborative filtering data set. MovieLens data sets were collected by the GroupLens Research Project at the University of Minnesota.

Except for ratings awarded by users on items, the MovieLens data set includes information regarding specifically the users. The included data consists of a sequential list, with 943 vectors of the following form [3]:

user id | age | gender | occupation | zip code

The user ids are the ones also used in the main data file. The gender can be either ‘M’, for male, or ‘F’, for female. The occupation takes a value from a list of 21 distinct possibilities.

1. Age: The user demand is different according to the user age, and so the user interest is the same. Children would like to watch animation and children’s films, young people would like to watch romance film, the middle age people would like to watch life film, and the old people would like to watch documentary film.

2. Gender: In many aspects, users choose different items as the different genders. The females would like to watch fantasy film, and the males would like to watch war film.

3. Occupation: Many users can divide into one category according to their occupation. The level of the artists is higher than the educations, so they have different interest in the films.

4. Zip code: Users in the same region may have the same interest in same ways.

### C. Combing the rating similarity and user attribute similarity

We propose a hybrid method that groups users by integrating the user rating similarity and user attribute information similarity. The relative weighting is adopted to adjust the importance of rating similarity and attribute similarity. We initially establish a user-item rating matrix and a user-attribute matrix. Then, users rating similarity and the user attribute similarity are computed. The integrated measurement of similarity is then derived as following formula.

$$sim(i, j) = \omega sim_1(i, j) + (1 - \omega) sim_2(i, j) \quad (1)$$

Where,  $\omega$  and  $1 - \omega$  represent the relative importance of the user rating similarity and user attribute similarity, respectively. If  $\omega = 0$ , then the method becomes user information-based method. If  $\omega = 1$ , then the method becomes traditional CF method.

### D. Fill the Vacant Ratings

Since we have got the membership of user, we can calculate the weighted average of neighbors’ ratings, weighted by their similarity to the target user. When count the object user U ratings for not graded items, produce the prediction according to the nearest neighbor for user ratings.

The rating of the target user u to the target item t is as following:

$$P_{ut} = A_u + \frac{\sum_{i=1}^c (R_{it} - A_i) * sim(u, i)}{\sum_{i=1}^c sim(u, i)} \quad (2)$$

Where  $A_u$  is the average rating of the target user u to the items,  $R_{it}$  is the rating of the neighbour user i to the target item t,  $A_i$  is the average rating of the neighbour user i to the items,  $sim(u, i)$  is the combining similarity of the target user u and the neighbour user i, and c is the number of the neighbours.

## III. USING ITEM ATTRIBUTE TO PRODUCE RECOMMENDATIONS

### A. Item attribute content

The content of many items such as books, videos, or CDs is difficult to analyze automatically by a computer, but the items may be categorized or clustered based on the attributes of the items. For example, in the context of movies, every movie can be classified according to the “genre” attribute of each item. Other item descriptions such as title, category, subject, authors, and published time also reflect the interests of a user when a user reads or downloads items [8,9]. Table 1 shows examples of the descriptive information of items.

TABLE I  
ITEM-ITEM ATTRIBUTE TABLE

Attribute Item	A1	A2	... ..	At
Item1	r11	r12	... ..	r1t
Item2	r21	r22	... ..	r2t
... ..	... ..	... ..	... ..	... ..
Itemn	rn1	rn2	... ..	rnt

Where,  $r_{ij}$  denotes the express value of the item to its attribute. The symbol n denotes the total number of items, and t denotes the total number of item attributes.

### B. Data set including item ratings and attributes

We use MovieLens collaborative filtering data set. The complete data set includes in random order 100,000 vectors of the following form [3]:

user id | item id | rating | time stamp

Obviously, users are enumerated from 1 to 943, items from 1 to 1682, while ratings take values between 1 and 5. The time stamps are unix seconds since 1/1/1970 UTC.

Except for ratings awarded by users on items, the MovieLens data set includes information regarding specifically the items. The items, which in the case of the MovieLens data set correspond to movies, there is another sequential list, with 1682 vectors of the following form:

movie id | movie title | release date | video release date | IMDb URL | unknown | Action | Adventure | Animation |

Children's | Comedy | Crime | Documentary | Drama | Fantasy | Film-Noir | Horror | Musical | Mystery | Romance | Sci-Fi | Thriller | War | Western

The movie ids are the ones used in the main data set. The movie title is a string with the title of the movie. The release dates are of the form dd-mmm-yyyy, e.g. 14-Jan-1967. The IMDb URL is a web link leading to the Internet Movie Database page of the corresponding movie. The last 19 fields are the film genres. Items can belong to more than one genres at the same time.

### C. Measuring the item rating similarity

There are several similarity algorithms that have been used in the item based collaborative filtering: Pearson correlation, cosine vector similarity, adjusted cosine vector similarity, mean-squared difference and Spearman correlation.

In this paper, we will use the Pearson correlation measurement.

Pearson's correlation, as following formula, measures the linear correlation between two vectors of ratings as the target item  $t$  and the remaining item  $r$ .

$$sim_1(t, r) = \frac{\sum_{i=1}^m (R_{it} - A_t)(R_{ir} - A_r)}{\sqrt{\sum_{i=1}^m (R_{it} - A_t)^2 \sum_{i=1}^m (R_{ir} - A_r)^2}} \quad (3)$$

Where  $R_{it}$  is the rating of the target item  $t$  by user  $i$ ,  $R_{ir}$  is the rating of the remaining item  $r$  by user  $i$ ,  $A_t$  is the average rating of the target item  $t$  for all the co-rated users,  $A_r$  is the average rating of the remaining item  $r$  for all the co-rated users, and  $m$  is the number of all rating users to the item  $t$  and item  $r$ .

### D. Measuring the item attribute similarity

We also use the Pearson correlation measurement to compute the item attribute similarity, as following formula.

$$sim_2(t, r) = \frac{\sum_{a=1}^m (R_{ta} - A_t)(R_{ra} - A_r)}{\sqrt{\sum_{a=1}^m (R_{ta} - A_t)^2 \sum_{a=1}^m (R_{ra} - A_r)^2}} \quad (4)$$

Where  $R_{ta}$  is the express value of the target item  $t$  to its attribute  $a$ ,  $R_{ra}$  is the express value of the remaining item  $r$  to the attribute  $a$ ,  $A_t$  is the average value of the target item  $t$  for all the co-rated attributes,  $A_r$  is the average rating of the remaining item  $r$  for all the co-rated attributes, and  $m$  is the number of all rating attribute to the item  $t$  and item  $r$ .

### E. Combining the two similarities

We propose a hybrid method that clusters items by combining the item rating similarity and item attribute similarity. The relative weighting is adopted to adjust the importance of rating similarity and attribute similarity. The integrated measurement of similarity is then derived as following formula.

$$sim(i, j) = wsim_1(i, j) + (1-w)sim_2(i, j) \quad (5)$$

Where,  $w$  and  $1-w$  represent the relative importance of the item rating similarity and item attribute similarity, respectively. If  $w=0$ , then the method becomes item attribute-based method. If  $w=1$ , then the method becomes traditional item-based CF method.

### F. Producing Recommendations

Since we have got the membership of item, we can calculate the weighted average of neighbors' ratings, weighted by their similarity to the target item.

The rating of the target user  $u$  to the target item  $t$  is as following:

$$P_{ut} = \frac{\sum_{i=1}^c R_{ui} \times sim(t, i)}{\sum_{i=1}^c sim(t, i)} \quad (6)$$

Where  $R_{ui}$  is the rating of the target user  $u$  to the neighbour item  $i$ ,  $sim(t, i)$  is the similarity of the target item  $t$  and the neighbour it user  $i$  for all the co-rated items, and  $m$  is the number of all rating users to the item  $t$  and item  $r$ .

## IV. EXPERIMENT RESULTS

In this section, we describe the dataset, metrics and methodology for the comparison between traditional and proposed collaborative filtering algorithm, and present the results of our experiments.

### A. Data Set

We use MovieLens collaborative filtering data set to evaluate the performance of proposed algorithm. MovieLens data sets were collected by the GroupLens Research Project at the University of Minnesota and MovieLens is a web-based research recommender system that debuted in Fall 1997. Each week hundreds of users visit MovieLens to rate and receive recommendations for movies [3].

### B. Performance Measurement For the Collaborative Filtering

Several metrics have been proposed for assessing the accuracy of collaborative filtering methods. They are divided into two main categories: statistical accuracy metrics and decision-support accuracy metrics. In this paper, we use the statistical accuracy metrics [10,11].

Statistical accuracy metrics evaluate the accuracy of a prediction algorithm by comparing the numerical deviation of the predicted ratings from the respective actual user ratings. Some of them frequently used are mean absolute error (MAE), root mean squared error (RMSE) and correlation between ratings and predictions. All of the above metrics were computed on result data and generally provided the same conclusions. As statistical accuracy measure, mean absolute error is employed.

Formally, if  $n$  is the number of actual ratings in an item set, then MAE is defined as the average absolute difference between the  $n$  pairs. Assume that  $p_1, p_2, p_3, \dots, p_n$  is the prediction of users' ratings, and the corresponding real ratings data set of users is  $q_1, q_2, q_3, \dots, q_n$ . See the MAE definition as following:

$$MAE = \frac{\sum_{i=1}^n |p_i - q_i|}{n} \quad (7)$$

The lower the MAE, the more accurate the predictions would be, allowing for better recommendations to be formulated. MAE has been computed for different prediction algorithms and for different levels of sparsity.

### C. Comparing the proposed collaborative filtering with the traditional CF

We compare the proposed method combining user attribute and item attribute collaborative filtering with the traditional collaborative filtering. The size of the neighborhood has a significant effect on the prediction quality. In our experiments, we vary the number of neighbors and compute the MAE. The obvious conclusion from Figure 1, which includes the Mean Absolute Errors for the proposed algorithm and the traditional collaborative filtering as observed in relation to the different numbers of neighbors, is that our proposed algorithm is better.

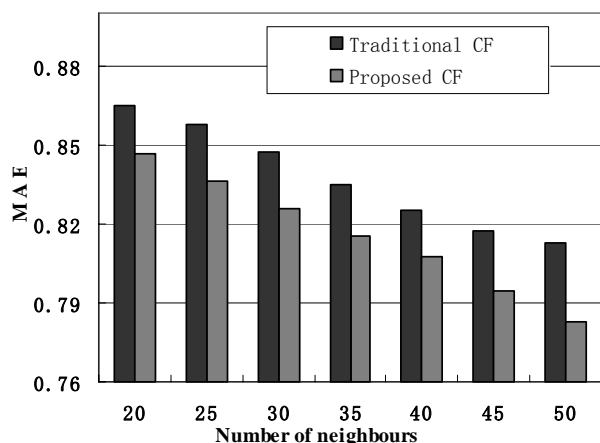


Figure1. Comparing the proposed CF algorithm with the traditional CF algorithm.

## VII. CONCLUSIONS

In this paper, we proposed a new collaborative filtering personalized recommendation algorithm which employs the user attribute information and the item attribute information. This approach combines the user rating similarity and the user attribute similarity in the user

based collaborative filtering process and then it combines the item rating similarity and the item attribute similarity in the item based collaborative filtering process to produce recommendations. The collaborative filtering recommender employs the user attribute and item attribute can alleviate the sparsity issue in the recommender systems.

## ACKNOWLEDGMENT

A Project Supported by Scientific Research Fund of Zhejiang Provincial Education Department (Grant No. Y200806038).

## REFERENCES

- [1] George Lekakos, George M. Giaglis, Improving the prediction accuracy of recommendation algorithms: Approaches anchored on human factors, *Interacting with Computers* 18 (2006) 410–431.
- [2] Gao Fengrong, Xing Chunxiao, Du Xiaoyong, Wang Shan, *Personalized Service System Based on Hybrid Filtering for Digital Library*, Tsinghua Science and Technology, Volume 12, Number 1, February 2007,1-8.
- [3] M.G. Vozalis, K.G. Margaritis, Using SVD and demographic data for the enhancement of generalized Collaborative Filtering, *Information Sciences* 177 (2007) 3017–3037.
- [4] Yu Li, Liu Lu, Li Xuefeng, A hybrid collaborative filtering method for multiple-interests and multiple-content recommendation in E-Commerce, *Expert Systems with Applications* 28 (2005) 67–77.
- [5] Breese J, Hecherman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering. In: *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI'98)*. 1998. 43–52.
- [6] Hyung Jun Ahn, A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem, *Information Sciences* 178 (2008) 37-51.
- [7] Yu Li, Liu Lu, Li Xuefeng, A hybrid collaborative filtering method for multiple-interests and multiple-content recommendation in E-Commerce, *Expert Systems with Applications* 28 (2005) 67–77.
- [8] YaE Dai, SongJie Gong, *Personalized Recommendation Algorithm using User Demography Information*, WKDD2009, IEEE Computer Society Press.
- [9] SongJie Gong, XiaoYan Shi, *A Collaborative Recommender Combining Item Rating Similarity and Item Attribute Similarity*, ISBIM2008, IEEE Computer Society Press.
- [10] Huang qin-hua, Ouyang wei-min, Fuzzy collaborative filtering with multiple agents, *Journal of Shanghai University (English Edition)*, 2007,11(3):290-295.
- [11] GuangHua Cheng, SongJie Gong, *An Efficient Collaborative Filtering Algorithm with Item Hierarchy*, In: *Second International Symposium on Intelligent Information Technology Application(IITA2008)*, IEEE Computer Society Press, 2008, Volume3, pp.28-31.