A Collaborative Filtering Recommendation Algorithm Based On Item Classification

HengSong Tan
Zhejiang Business Technology Institute,
Ningbo 315012, P. R. China
e-mail: tanhengsongzjbti@163.com

HongWu Ye
Zhejiang Textile & Fashion College,
Ningbo 315211, P. R. China
e-mail: yehongwuzjbti@163.com

***Abstract*—Collaborative filtering systems represent services of personalized that aim at predicting a user's interest on some items available in the application systems. With the development of electronic commerce, the number of users and items grows rapidly, resulted in the sparsity of the user-item rating dataset. Poor quality is one major challenge in collaborative filtering recommender systems. Sparsity of users' ratings is the major reason causing the poor quality and the traditional similarity measure methods make poor in this situation. To address this issue, this paper proposes a collaborative filtering recommendation algorithm based on the item classification to pre-produce the ratings. This approach classifies the items to predict the ratings of the vacant values where necessary, and then uses the item-based collaborative filtering to produce the recommendations. The collaborative filtering recommendation method based on item classification prediction can alleviate the sparsity problem of the user-item rating dataset, and can provide better recommendation than traditional collaborative filtering.**

Keywords-recommender system; collaborative filtering; item classification rating; sparsity

I. INTRODUCTION

Nowadays, the explosive growth of the Internet has brought us a vast amount of information that we can hardly digest. To deal with the flood of information, various recommender systems have been created to assist and augment this natural social process [1,2,3]. Recommender systems have been developed to automate the recommendation process. These systems recommend various types of web resources, online news, movies. Large scale commercial applications of the recommender systems can be found at many electronic commerce sites, such as Amazon, CDNow, DangDang, and Sinforyou. These commercial systems recommend products to potential consumers based on previous transactions and feedback. They are becoming part of the standard electronic business technology that can enhance electronic commerce sales by converting browsers to buyers, increasing cross-selling, and building customer loyalty [4,5,6,7].

Collaborative filtering (CF) has been the most successful recommendation system approach to date and has been widely applied in various applications. CF is based on the assumption that similar users have similar preferences. In other words, by finding users that are similar to the active user and by examining their preferences, the recommender system can predict the active user's preferences for certain items and provide a ranked list of items which the active user will most probably like. Collaborative filtering generally ignores the form and the content of the items and can therefore also be applied to non-textual items [8,9,10]. Despite its wide spread adoption, collaborative filtering suffers from several major limitations including sparsity, system scalability, and synonymy.

However, in most cases in real-world applications, the ratio of rated items to the total of available items is very low. The absence of a sufficient amount of available ratings significantly affects CF methods reducing the accuracy of prediction. The sparsity of ratings problem is particularly important in domains with large or continuously updated list of items as well as a large number of users. The sparsity problem may occur when either none or few ratings are available for the target user, or for the target item that prediction refers to, or for the entire database in average [11,12,13]. Different treatments are required and different prediction techniques must be employed depending on the sparsity conditions, making the selection of an appropriate approach a cumbersome task. Current CF approaches are limited in the sense that they address specific aspects of the above problem.

In this paper we propose a collaborative filtering recommendation algorithm based on the item classification to pre-produce the ratings. This approach classifies the items to predict the ratings of the vacant values where necessary, and then uses the item-based collaborative filtering to produce the recommendations. The collaborative filtering recommendation method based on item classification prediction can alleviate the sparsity problem of the user-item rating dataset, and can provide better recommendation than traditional collaborative filtering.

II. USING ITEM CLASSIFICATION TO PRE-PRODUCE PREDICTTION

Firstly, we classify the items using the item attribute content, and then in the sub-matrix, using user-based collaborative filtering to fill the vacant ratings.

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

978-0-7695-3614-9/09 $25.00 © 2009 IEEE
DOI 10.1109/PACCS.2009.68

attributes of the items [14,15,16]. 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. Table 1 shows examples of the descriptive information of items.

Table1 item-item attribute table

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

Where, rij 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. Pre-preducing prediction where necessary

According to the item attribute content, we cluster the items in some classifications. Just as C1(Ic11,Ic12,…,Ic1k), C2(Ic21,Ic22,…,Ic2k), …,Ci(Ici1,Ici2,…,Icik),… , Ck(Ick1, Ick2,…,Ickk). Then, in the sub-matrix, we employ user-based collaborative filtering algorithm to fill the vacant rating where necessary in the next step.

III. USING ITEM-BASED CF TO PRODUCE PREDICTION

After using the item classification to pre-produce the prediction where necessary, we have the dense user-item rating matrix. In this section, we use the item based collaborative filtering algorithm to produce recommendation.

A. Measuring the item rating similarity

There are several similarity algorithms that have been used [7,13,17]: Pearson correlation, cosine vector similarity, adjusted cosine vector similarity, mean-squared difference and Spearman correlation.

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.

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

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.

The cosine measure, as following formula, looks at the angle between two vectors of ratings as the target item t and the remaining item r.

$$sim(t,r) = \frac{\sum_{i=1}^{m} R_{it} R_{ir}}{\sqrt{\sum_{i=1}^{m} R_{it}^2 \sum_{i=1}^{m} R_{ir}^2}} \quad (2)$$

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, and m is the number of all rating users to the item t and item r.

The adjusted cosine, as following formula, is used for similarity among items where the difference in each user's use of the rating scale is taken into account.

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

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_i is the average rating of user i for all the co-rated items, and m is the number of all rating users to the item t and item r.

B. Prediction using item-based CF

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:

$$P_{ut} = \frac{\sum_{i=1}^{c} R_{ui} \times sim(t,i)}{\sum_{i=1}^{c} sim(t,i)} \quad (4)$$

Where Rui 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 item i, and c is the number of the neighbours.

IV. DATASET AND MEASUREMENT

A. Data set

We use MovieLens collaborative filtering data set to execute of our subsequent experiments and evaluate the performance of proposed algorithm. MovieLens data sets were collected by the GroupLens Research Project at the University of Minnesota. The historical dataset consists of 100,000 ratings from 943 users on 1682 movies with every user having at least 20 ratings. [18]

The complete data set includes in random order 100,000 vectors of the following form:

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.

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.

B. Performance measurement

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 [19].

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 (MAE) 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 p1, p2, p3, ..., pn is the prediction of users' ratings, and the corresponding real ratings data set of users is q1, q2, q3, ..., qn. See the MAE definition as following:

$$MAE = \frac{\sum_{i=1}^{n} | p_i - q_i |}{n} \quad (5)$$

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 with the traditional collaborative filtering

We compare the proposed collaborative filtering algorithm based on item classification with the traditional collaborative filtering. The performance of our proposed collaborative filtering is better than the traditional collaborative filtering in terms of the MAE measure.

V. CONCLUSIONS

Personalized recommender systems represent services of personalized that aim at predicting a user's interest on some items available in the application systems. With the development of electronic commerce, the number of users and items grows rapidly, resulted in the sparsity of the user-item rating dataset. Poor quality is one major challenge in collaborative filtering recommender systems. Sparsity of users' ratings is the major reason causing the poor quality and the traditional similarity measure methods make poor in this situation. In this paper, we propose a collaborative filtering recommendation algorithm based on the item classification to pre-produce the ratings. This approach classifies the items to predict the ratings of the vacant values where necessary, and then uses the item based collaborative filtering to produce the recommendations. The collaborative filtering recommendation method based on item classification prediction can alleviate the sparsity problem of the user-item rating dataset, and can provide better recommendation than traditional collaborative filtering.

REFERENCES

[1] S. Maneeroj, H. Kanai, K. Hakozaki, "Combining Dynamic Agents and Collaborative Filtering without Sparsity Rating Problem for Better Recommendation Quality", Proceedings of the Second DELOS Network of Excellence Workshop, 2001, pp.33-38

[2] Resnick, P., and Varian, H. R. (1997). Recommender Systems. Special issue of Communications of the ACM. 40(3).

[3] Schafer, J. B., Konstan, J., and Riedl, J. (1999). Recommender Systems in E-Commerce. In Proceedings of ACM E-Commerce 1999 conference.

[4] Huang, Z., Chen, H. and Zeng, D. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. ACM Transactions on Information Systems, 22, 1 (2004), 116-142.

[5] Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J. (2000). Analysis of Recommendation Algorithms for E-Commerce. In Proceedings of the ACM EC'00 Conference. Minneapolis,MN. pp. 158-167

[6] Schafer J.B., Konstan J., Riedl J. (2000). Electronic Commerce Recommender Applications. Journal of Data Mining and Knowledge Discovery, Vol 5 (1/2), 115-152.

[7] Herlocker, J. (2000). Understanding and Improving Automated Collaborative Filtering Systems. Ph.D. Thesis, Computer Science Dept., University of Minnesota.

[8] Miha Grcar, Dunja Mladenic, Blaz Fortuna and Marko Grobelnik, Data Sparsity Issues in the Collaborative Filtering Framework ,Lecture Notes in Computer Science,Volume 4198 2006,pp58-76

[9] 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.

[10] Billsus, D., and Pazzani, M. J. (1998). Learning Collaborative Information Filters. In Proceedings of ICML '98. pp. 46-53.

[11] Chong-Ben Huang, Song-Jie Gong, Employing rough set theory to alleviate the sparsity issue in recommender system, In: Proceeding of the Seventh International Conference on Machine Learning and Cybernetics (ICMLC2008), IEEE Press, 2008, pp.1610-1614.

[12] Songjie Gong, Chongben Huang, Employing Fuzzy Clustering to Alleviate the Sparsity Issue in Collaborative Filtering Recommendation Algorithms, In: Proceeding of 2008 International

Pre-Olympic Congress on Computer Science, World Academic Press, 2008, pp.449-454.

[13] Sarwar B, Karypis G, Konstan J, Riedl J. Item-Based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International World Wide Web Conference. 2001. 285-295.

[14] 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.

[15] SongJie Gong, GuangHua Cheng, Mining User Interest Change for Improving Collaborative Filtering, In:Second International Symposium on Intelligent Information Technology Application(IITA2008), IEEE Computer Society Press, 2008, Volume3, pp.24-27.

[16] 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.

[17] George Lekakos, George M. Giaglis, A hybrid approach for improving predictive accuracy of collaborative filtering algorithms, User Model User-Adap Inter (2007) 17:5–40

[18] M.G. Vozalis, K.G. Margaritis, Using SVD and demographic data for the enhancement of generalized Collaborative Filtering, Information Sciences 177 (2007) 3017–3037.

[19] Huang qin-hua, Ouyang wei-min, Fuzzy collaborative filtering with multiple agents, Journal of Shanghai University (English Edition), 2007,11(3):290-295.