A hybrid genre-based personalized recommendation algorithm

Yajie Hu, Yi Yang*,Caihong Li,Yachen Wang
School of Information Science & Engineering
Lanzhou University
Corresponding Author: yy@lzu.edu.cn

Lian Li
School of Computer and Information
Hefi University of Technology
Hefei, China

***Abstract*—Because of the serious information overload problem on the internet, the recommender system as one of the most important solutions has been widely used to help users find more valuable information. However, the traditional collaborative filtering method is seriously affected by the rating sparseness and cold start to obtain the precise recommendation. In this paper, a hybrid collaborative filtering method (CGCF) based on the traditional user-based collaborative filtering, as well as, a new approach named genre-based collaborative filtering (GCF) is proposed. GCF uses term frequency–inverse document frequency(TF-IDF) to combine users' former ratings with item genres to quantize individual genre preference. Combining GCF and User-bsed collaborative filtering with dynamic weight, we proposed the CGCF. According to the experiment on Movielens dataset, when comparing with Item-based collaborative filtering, CGCF has reduced MAE by 2.2% and improved Coverage by 16.9%. When comparing with User-based collaborative filtering, CGCF has reduced MAE by 2.5% and improve Coverage by 6.2%. The results demonstrate that the proposed method improves the precision and coverage of recommendation obviously comparing with the traditional ones.**

Keywords—component; collaborative filtering; TF-IDF; data spareness; user genre profile; dynamic weight

I. INTRODUCTION

In recent years, the rapid development of e-commerce systems have brought a great convenience to people. We can browse, purchase, and enjoy all kinds of servives indoors. However, countless goods make us dizzy. A rapid and efficient way of finding the goods becomes our urgent desire. With the capacity of dealing with the serious information overload and provide more personalized services for individual users, personalized recommender systems are regarded as the most useful solutions. Collaborative filtering is the most polular approach of personalized recommender systems.

Collaborative filtering recommender systems [1,2] collect users' preferences from the historic behaviors, and then recommend specific items to them depending on their interests. User-based and Item-based are the most popular techniques in the current recommendation systems. The former technique simulates the process your like-minded friends who introduce a new product to you. It will find user's potential interest. The other one will recommend items that have the similar feedback informations to the target users consumed before. It is just like a user spoke highly of one movie named Interstellar, a similar rating movie called Marvel's The Avengers is recommended. Obviously, these two techniques rely on the ranking matrix totally, however, the quantity of goods is enormous, nevertheless, some users may not comment on goods they have consumed, which makes the matrix is insufficient for collaborative filtering which brings out two apparent problems [3,4] called “data sparseness” and “cold start”.

The way of reducing the data sparseness and solving the cold start problem effectively is the main directions of current research. The most common method considers demographic information to fill the missing scores. Dai et al. [5] assumed that users have the similar demographic information may have a similar preference. However, the demographic information is not detailed enough to ensure that users have similar preferences. In addition, predict filling algorithm based on resource characteristics [6] adopt weighted value of neighbor resource ratings to fill the missing rating. Whereas the characteristics we need are hard to obtain, and the representative of the characteristics are difficult to judge. Many clustering approaches have been applied in collaborative filtering, such as k-means [7] and co-clustering method [8,9]. They can reduce influence caused by the sparsity of the matrix and improve predicting accuracy. However, these methods will consume unavoidable time that is impossible to offer online services. Jesse vig.et al. [10] presented that tags marked by users are able to strongly reflect the users’ real preferences, however, if tag technology is applied to recommend items for users, the process of getting labels will burden users.

Aiming to solve the data spareness and cold start problems without much participation of users. In this paper, we propose a hybrid collaborative filtering method named CGCF, which is composed of two parts: a genre-based collaborative filtering (GCF) for handling new item cold start and making recommendations for users with few ratings; and the traditional user-based collaborative filtering algorithm. Combining both parts by dynamic weight, the hybrid method is presented.

The rest of the paper is organized as follows: Section 2 reviews basic flowchart of recommender system and relevant similarity calculation techniques. In section 3, CGCF hybrid collaborative filtering method and its formal filtering process are presented. Section 4 evaluates the performance of the new method on Movielens Dataset and campares the results with the classical methods. Section 5 concludes our method and puts forward of our future research direction.

978-1-4673-8644-9/16/$31.00

II. RELATED TECHNIQUES

Based on rating matrix, the traditional collaborative filtering aims to solve the information overload problem and helps users find their interests. Section 2.1 presents a classic recommendation process. Section 2.2 reviews several most useful similarity calculation techniques. A simple sample of the rating matrix is given in TABLE I.

TABLE I. AN EXAMPLE OF USER-GENRE RATING MATRIX

	$item_1$	$item_2$	$item_3$	$item_4$	$item_5$	$item_6$
$user_1$	0	5	0	0	1	4
$user_2$	1	3	0	2	0	0
$user_3$	0	0	0	0	2	0
$user_4$	0	0	0	5	0	5
$user_5$	2	5	0	0	3	1

A. classic recommendation process

As we all know, the classic recommendation process of collaborative filtering can be divided into the following four steps .

- Finding similar users or item neighbours in terms of user-item rating matrix.
- Using the weighted summation of neighbours' ratings to predict the new item.
- Sorting the prediction ratings of new items by descending order.
- Recommending the top-N items to the target users.

B. Relevant similar calculating techniques

As known, the most important part of the whole method is to gain the appropriate user or item neighbours. How can we get the appropriate neighbors? The traditional similarity measures are shown in TABLE Ⅱ.

TABLE II. AN EXAMPLE OF USER-GENRE RATING MATRIX

Technique	Formula	Descriptions
Cosine	$sim(i,j)=\frac{\vec{i}\cdot\vec{j}}{\|\vec{i}\|\times\|\vec{j}\|}=\frac{\sum_{k\in I_{ij}}R_{i,k}R_{j,k}}{\sqrt{\sum_{k\in I_i}R_{i,k}^2}\sqrt{\sum_{k\in I_j}R_{j,k}^2}}$	$sim(i,j)$ the similarity of i and j $\vec{i}$ the rating vector of user i. $\vec{j}$ the rating vector of user j. $R_{i,k}$ the rating on item k by user i. $R_{j,k}$ the rating on item k by user j. $\overline{R_i}$ the average rating by user i. $\overline{R_j}$ the average rating by user j. I_i the set of items rated by user i. I_j the set of items rated by user j. I_{ij} the set of items rated by user i and j $I_{ij}=I_i\cap I_j$.
Adjusted Cosine	$sim(i,j)=\frac{\sum_{k\in I_{ij}}(R_{i,k}-\overline{R_i})(R_{j,k}-\overline{R_j})}{\sqrt{\sum_{k\in I_i}(R_{i,k}-\overline{R_i})^2}\sqrt{\sum_{k\in I_j}(R_{j,k}-\overline{R_j})^2}}$	
Person Correlation	$sim(i,j)=\frac{\sum_{k\in I_{ij}}(R_{i,k}-\overline{R_i})(R_{j,k}-\overline{R_j})}{\sqrt{\sum_{k\in I_{ij}}(R_{i,k}-\overline{R_i})^2}\sqrt{\sum_{k\in I_{ij}}(R_{j,k}-\overline{R_j})^2}}$	
Jaccard	$sim(i,j)=\frac{\|I_i\cap I_j\|}{\|I_i\cup I_j\|}$	

Cosine is the classic approach in item-based collaborate filtering. It treats the target user A and B as two rating vectors and calculates the angle of the vectors as the similarity. Due to the Cosine similarity does not take personal rating scales into account, some people prefer to mark high ratings but the others not. So the Adjusted Cosine uses the mean rating of users to correct the predicting result. Person Correlation Coefficient that measures the similarity of two users based on the linear relation, the result ranges from -1 to +1. Jaccard Coefficient used to evaluate the similarity of two users based on the probability of rating concentration and dispersion.

Based on the common items, all those techniques depend on the ratings marked by users. Faced with the high degree of sparsity of the rating databases just like TABLE I, the results from above calculations are not accurate enough to the recommendation. For example, if we evaluate the similarity of $user_1$ and $user_2$ in TABLE I with the techniques talked above respectively. A variety of results got by Person Correlation Coefficient, Cosine, Adjusted Cosine and Jaccard are so different as 100%, 61.9%, 40.9%, 20%, respectively.

For the above measures, the example indicates the missing of capacity to compute the similarities with spare rating data. In order to improve the recommendation performance, the algorithm must address the issue of data sparseness.

III. PROPOSED METHOD CGCF

The method uses item genre informations to modify the rating matrix. User-genre profile is regarded as the criteria of choosing new items and evaluating the comsumed items. The filtering process for prediction rating is shown in section 3.1. Some other important parts of the new method are stated in the following four parts.

A. Overview of the framework

The novel filtering model predicts the rating by user-item profile and user-genre profile. The flowchart is shown in Fig. 1. The details of some parts are described as follows.

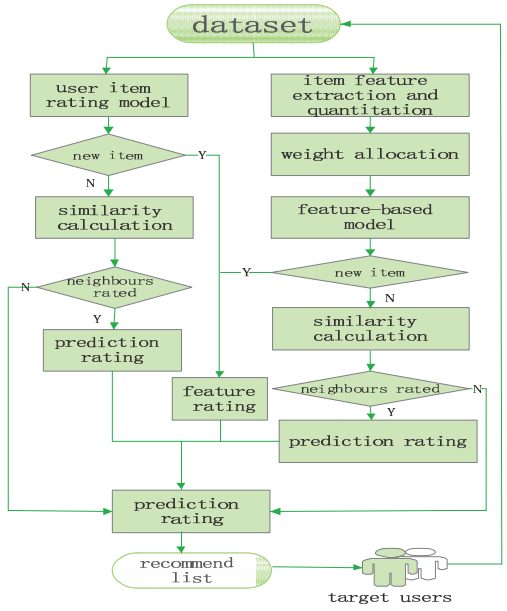

Fig. 1. Filtering process for recommendations

B. Personalized item genres extraction and quantization

Before that, some assumptions are defined here.

- User ratings for items can reflect users' preferences on item genres.
- User profile for item genres has a filterable function to select a new item.
- Multiple genres codetermine the user profile for items.

Obviously, an item can have one or more genres attributes. For example, a movie can be labelled adventure, comedy, action and western. So every item can be indicated as below.

$$I = \{G_i \mid i \in 1...m\} \tag{1}$$

G_i can be 0 or 1, m is the number of features selected by the algorithm.

C. Personalized genres weight allocation

Because users have different preference levels on different genres, TF-IDF principle is applied to quantify the importance of item genres for individuals. Using $user_t$ as an example, the details as follows.

The algorithm uses the average rating of $user_t$ as the boundary to divide R into L_t and H_t two parts. The more items in L_t own G_a and the less items in H_t do not own G_a can suggest that $user_t$ likes the items own G_a. And G_a can be more likely to predict the profile of $user_t$. On the contrary, the less items in L_t own G_a and the more items in H_t do not own G_a can suggest that $user_t$ do not like items own G_a. And G_a can be weakly to predict the positive profile of $user_t$.

Based on this principle, every personalized genre weight for users can be calculated as follows.

$$w_{ta} = n_{tal} \times \log(\frac{N_a}{n_a}) \tag{2}$$

n_{tal} is the count of items which have G_a in L_t, n_a is the count of items which have G_a in R, N_a is the count of items of R.

D. Rating matrix transformation

According to the item rating and the feature weight, every user' genre quantitative rating can be obtained with formula (3). The user-genre rating matrix can be got as TABLE III.

$$s_{ta} = \frac{1}{n}\sum_{i=1}^{n} r \times \frac{w_{ta}}{\sum_{j=1}^{m} w_{tj}} \quad a \in \{j \mid j \in I_i \wedge j \neq 0\} \tag{3}$$

w_{ta} stands the rating on from $user_t$, n is the number of items which is liked by $user_t$, r stands the rating on $item_i$ from $user_t$.

TABLE III. AN EXAMPLE OF USER-GENRE RATING MATRIX

	G_1	G_2	G_3	.....	G_m
$user_1$	1.0186	0. 5204	2.0586	.....	0.1225
$user_2$	0.5235	1.0245	0.9650	.....	0.5897
$user_3$	0.0125	2.0530	1.0251	.....	2.0147
.....	.....	.....	.....	.....	.....
$user_5$	0.9602	0.0065	2.0500	.....	3.0141

E. Rating prediction

Our hybrid method uses similar neighbors to predict the missing rating with formula (4). If the target item has not been rated by anyone, it will be predicted with formula (5). All these ratings are sorted in descending order, the top-N items that are higher than the threshold will be recommended to target user. The threshold is considered as the average rating of the target user.

$$p_{t,i}=\overline{R_t}+\lambda\times\frac{\sum_{v\in s_1}sim(t,v)\times(R_{v,i}-\overline{R_v})}{\sum_{v\in s_1}sim(t,v)}+(1-\lambda)\times\frac{\sum_{v\in s_2}sim(t,v)\times(R_{v,i}-\overline{R_v})}{\sum_{v\in s_2}sim(t,v)} \quad (4)$$

$\overline{R_t}$, $\overline{R_v}$ are the average ratings of $user_t$ and $user_v$ separately , $sim(u,v)$ is the similarity of $user_t$ and $user_v$, $R_{v,i}$ stands the rating on $item_i$ ·from $user_v$, λ is the ratio of the number of users who rated this item in similar set to the number in s_1 and s_2.

$$p_{t,i}=\frac{1}{n}\sum_{k=1}^{n}s_{tk} \qquad k\in\{j\mid j\in I_i\wedge j\neq 0\} \quad (5)$$

s_{tk} stands the rating on genre from $user_t$, n is the number of genres contained by $item_i$.

IV. The Experimental Results and Analysis

A. Data set

Movielens dataset was utilized to evaluate the validity of the hybrid method. It was collected by the GroupLens Research Project. The dataset consists of 100000 ratings from 943 users on 1682 movies. All the movies have a total of 19 different genres. Each movie was tagged at least one genre and maximum five genres. Each user has rated at least 20 movies. All the rating is between minimum value 1 and maximum value 5. The sparsity of the dataset is approximate to 93.7%. 80% dataset is randomly selected as the training set, the other 20% as the test set.

B. Measurements of prediction quality

There are multiple metrics to measure prediction and recommendation quality. Two popular metrics, mean absolute error (MAE) and coverage are applied in this paper to measure the performance of the hybrid collaborative filtering algorithm.

$$E=\{i\in I\mid p_i\neq\Phi\wedge r_i\neq\Phi\} \quad (6)$$

E stands items set where a prediction can be obtained from the algorithm.

Mean Absolute Error (MAE) which measures the prediction accuracy. It considers the mean error between the predicted ratings on the items and the actual ratings. The smaller the value is, the higher the prediction accuracy will be.

$$MAE=\frac{1}{n}\sum_{i\in E}|pi-ri| \quad (7)$$

r_i is the actual rating, p_i is the forecasted rating by the algorithm, n is the number of items in E.

Coverage [11] which measures the capacity of recommend new items of users' k-neighbours. It mainly considers the ratio of the number of items which can be predicted to the total test set. It is one of the most important indicators of practicality. The higher the value is, the better recommendation quality will be.

$$Coverage=\frac{n}{N} \quad (8)$$

n is the number of items can be predicted, N is the number of the total test set.

C. Performance evaluation

In order to assess the practicality of our CGCF, a comparison of our CGCF and GCF, User-based, Item-based by the two statistical variables (MAE and Coverage) was adopted. All those methods use AdjustedCosine as the basic method to measue the similarities of users. The results are shown in Fig. 2, Fig. 3 and TABLE Ⅳ. Judging from the experimental results, it can be seen that CGCF has the lowest value in MAE and the highest value in Coverage, which means the CGCF has the best recommending performance. By observing each value of different neighbor lists, when the neighbor number is 10, CGCF almost has the same MAE value with Item-based collaborative filtering, but the Coverage is higher 35.8%; Item-based collaborative filtering has the better MAE than User-based and GCF collaborative filtering when neighbor number is less than 20, along with the increasing number of neighbors, Item-based collaborative filtering loses the advantage on the MAE. All those algorithms will be close to optimal MAE value when neighbor number is 30. From the Fig. 3 we can see the Coverage value of CGCF has the absolute advantage than the other three algorithms. In addition, CGCF can make predictions about the 29 cold start items which cannot be predicted by User-based and Item-based collaborative filtering. It is a great breakthrough. All the experiment results are shown in TABLE Ⅳ.

The comparison experiment results obtained show that CGCF has a better MAE and Coverage. It outperforms the other three methods.

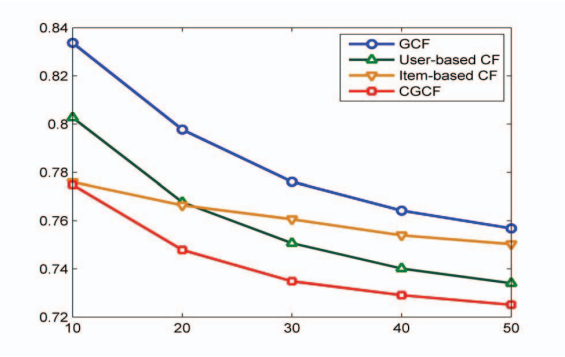

Fig. 2. Filtering MAE of four kinds of algorithms

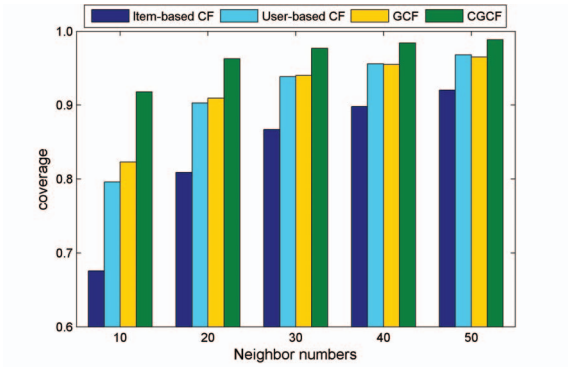

Fig. 3. Coverage of four kinds of algorithms

TABLE IV. TWO STATISTICS MEASURES OF THE FOUR

Neighbor Numbers	MAE of				Coverage of			
	User-CF	*Item-CF*	*GCF*	*CGCF*	*User-CF*	*Item-CF*	*GCF*	*CGCF*
K = 10	0.8027	0.7760	0.8336	0.7748	0.7559	0.6759	0.8231	0.9179
K = 20	0.7676	0.7663	0.7977	0.7478	0.9031	0.8089	0.9094	0.9628
K = 30	0.7507	0.7605	0.7761	0.7348	0.9388	0.8671	0.9401	0.9769
K = 40	0.7401	0.7539	0.7641	0.7290	0.9558	0.8982	0.9552	0.9840
K = 50	0.7340	0.7502	0.7568	0.7251	0.9678	0.9204	0.9640	0.9885

D. Conclusions

In this paper, we use the item genres to transform the original rating matrix. The new matrix reduces the data sparseness remarkablely. After that the personalized genre profile based on the new matrix is obtained. Combined the genre profile filtering GCF with the user-based filtering, we proposed a collaborative filtering method named CGCF. The experiment results on Movielens dataset show that it overcomes the limit of data sparseness and obtains a better performance both on recommendation precision and coverage. In future, this method will consider some other features which are useful for evaluating the user profile except item genres. In addition, the weight of genre filtering and traditional collaborative filtering could be adopted in a more reasonable way.

ACKNOWLEDGMENT

The authors would like to thank the Natural Science Foundation of P. R. of China (61300230), open fund of Guangxi Key laboratory of hybrid computation and IC design analysis and the Fundamental Research Funds for the Central Universities for supporting this research.

REFERENCES

[1] M.D. Ekstrand, J.T. Riedl, J.A. Konstan, Collaborative filtering recommender systems, Found. Trends Hum.–Comp. Interact. 4 (2010) 81-173.

[2] F. Ricci, L. Rokach, B. Shapira, P.B. Kantor, Recommender Systems Handbook,Springer Science + Business Media, 2011.

[3] H.N. Kim, A.T. Ji, H.J. Kim, G.S. Jo, Error-based collaborative filtering algorithm for top-N recommendation, Lecture Notes in Computer Science 4505 (2007)594-605.

[4] P. Massa, B. Bhattacharjee, Using trust in recommender systems: an experimental analysis, Lecture Notes in Computer Science 2995 (2004) 221-235.

[5] Dai, Yae, HongWu Ye, and SongJie Gong, "Personalized recommendation algorithm using user demography information," Knowledge Discovery and Data Mining, pp. 100-103, 2009.

[6] LIU ZB, WANGH, QuWY, et a1.Sparse matrix prediction filling in collaborative filtering[C]//IEEE International Conference on Scalable Computing and Communications: The Enighth International Conference on Embedd ed Computing. 2009:304-307.

[7] G.-R. Xue, C. Lin, Q. Yang, W. Xi, H.-J. Zeng, Y. Yu, Z. Chen, Scalable collaborative filtering using cluster-based smoothing, in: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2005, pp. 114-121.

[8] G. Chen, F. Wang, C. Zhang, Collaborative filtering using orthogonal nonnegative matrix tri-factorization, Inf. Process. Manag. 45 (3) (2009)

[9] H. Shan, A. Banerjee, Bayesian co-clustering, in: Eighth IEEE International Conference on Data Mining, ICDM'08, IEEE, 2008, pp. 530-539.

[10] Vig, Jesse, Shilad Sen, and John Riedl, "The tag genome: Encoding community knowledge to support novel interaction," ACM Transactions on Interactive Intelligent Systems, vol. 2, issue 3, no. 13, 2012.

[11] Xavier A, Neal L, Pujol JM, et al. The wisdom of the few: a collaborative filtering approach based on expert opinions from the web [C]/ /Proc of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09). New York: ACM Press,2009:532-539.