ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 133

2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE2016)

A User Interest Recommendation Based on Collaborative Filtering

Wenqi Wu, Jianfang Wang, Randong Liu, Zhenpeng Gu and Yongli Liu

Henan Polytechnic University of Computer Science and Technology, Jiaozuo 454000, China

***Abstract*—The traditional collaborative filtering algorithm cannot response user interest with time and is lack of time effectiveness. These problems lead to poor recommendation quality. On the basis of the neighbor-based collaborative filtering, a fused method of improved similarity and user interest is proposed. To begin with, we compute similarity from global perspectives obtained with Jaccard similarity, local perspectives obtained with Bhattacharyya Coefficient. Furthermore, we adopt the forgetting curve to represent the user interest preference, adding the interest weight to the new similarity method to update user interest. Finally, we make recommendation prediction by calculating similarity using the method. Experimental results on the Movielens datasets demonstrate that our approach has advantages over state-of-the-art methods in terms of both the discovery of user interest preference and providing highly accuracy recommendations.**

Keywords-collaborative filtering; Bhattacharyya Coefficient; forgetting curve; interest weight; similarity

I. INTRODUCTION

Personalized recommendation is a type of information filter to overcome the problem of information overload. The Collaborative Filtering (CF) [1][2], one of the most prevalent recommendation methods, learns the user interest preferences and behavior patterns by collecting and analyzing the data [3]. Then that can recommend information or item that users need.

CF has been applied to a wide variety of fields: movies, music, e-commerce, social news and other commercial domain. CF can be classified as: *k* nearest neighbor (KNN) based CF, model based CF [4]. KNN based CF computers the recommendations using the *k* most similar users to the target user in terms of ratings. The *k* most similar users are obtained by calculating similarity, so the results of similarity directly affect the quality of recommendation. The traditional CF doesn't take into account the issue that user's preferences change continuously with time, which is also widely known as interest drift [5]. The traditional approaches are carried out on the presupposition that the user interest is stable, which cannot represent the changes of user interest can produce low efficiency and precisions.

To sidestep the shortcomings of the algorithm above, we propose a user interest recommendation based on collaborative filtering. The paper takes advantage of user's historical ratings and single rating discrepancies between users to improve the similarity method, taking the user interest into account while computing the similarity. Therefore, the user's preference of each genre in the item can be defined, which can reflect the needs of user's behavior.

In the remainder of the paper is structured as follows. Section 2 describes the related works, Section 3 presents a user interest recommendation based on collaborative filtering, Section 4 shows experimental results and analysis, and Section 5 contains conclusions and further work.

II. RELATED WORK

Personalized recommendation has become core competitiveness between e-commerce sites and social media, according to the user's surfing and purchasing behavior, predicted user's preferences at different time points. There are some different ways to solve the interest drift. Chen et al. [6] propose a temporal recommender system-TencentRec [7], and deploy the TencentRec in a series of production applications. Zhu et al. [8] propose a dynamic user-interest-model. User's long term interest and short term interest can be embodied clearly in this model. Liu [9] proposes algorithm of collaborative filtering based on user interest, by building interest intensity model and discovering interest correlation among different items through that model. Patra et al. [10] propose a new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data. Cheng et al. [11] propose a new collaborative filtering recommendation method based on users' interest sequences (ISCF). These updated similarities, transition characteristics and dynamic evolution patterns of users' preferences are considered. Xiao et al. [12] propose a time-ordered collaborative filtering recommendation algorithm, which takes the time sequence characteristic of user behaviors into account. The above methods consider either time factor or similarity, a user interest recommendation based on collaborative filtering is proposed. It considers the user interest in improving the similarity calculation method, which can improve the accuracy of the recommendations.

Time is one of the most important context information, which has a profound influence on user preferences. In this paper, three different genres of movie whose the proportional of audience changes over time are shown in Figure I, that is analyzed with Movielens dataset as an example (from September 1997 to April 1998). The genre of Item 1 is comedy animation, the genre of Item 50 is a war sci-fi action, and the genre of Item 181 is romantic comedy in Figure I. As can be seen in Figure I the popularity of the film changes over time for the different genres of movie. The proportion of audience is gradually declining over time no matter what genre of movie. In general its law is similar to the forgetting curve in psychology. When the movie just releases, audiences pay more attention, the audiences slowly decline over time until been

forgotten. These human behaviors can be explained by Ebbinghaus forgetting curve.

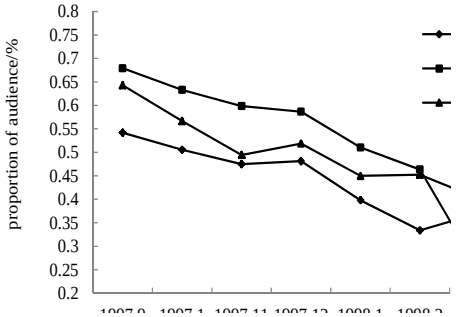

FIGURE I. THE RATIO CHART OF AUDIENCES ON THREE DIFFERENT GENRES OF MOVIES

KNN based CF exploits all the available rating information in the training dataset to predict the preference (vote or rating) of an active user on target item. It makes recommendations based on similarity methods of the users or items. The nearest neighbor set can be achieved by calculating the similarity method in KNN based CF. Traditional similarity measures such as cosine similarity, adjust cosine similarity and Pearson correlation coefficient are frequently used. KNN based CF can be divided into user-based CF and item-based CF. Take the user-based CF as an example, computing similarity between user *u* and user *v*:

The cosine similarity function:

$$sim(u,v)=\frac{\sum_{i\in I_{uv}} r_{u,i}r_{v,i}}{\sqrt{\sum_{i\in I_{uv}} r_{u,i}^{\,2}}\sqrt{\sum_{i\in I_{uv}} r_{v,i}^{\,2}}} \tag{1}$$

The adjusted cosine similarity function:

$$sim(u,v)=\frac{\sum_{i\in I_{uv}}\left(r_{u,i}-\overline{r_u}\right)\left(r_{v,i}-\overline{r_v}\right)}{\sqrt{\sum_{i\in I_u}\left(r_{u,i}-\overline{r_u}\right)^2}\sqrt{\sum_{i\in I_v}\left(r_{v,i}-\overline{r_v}\right)^2}} \tag{2}$$

The Pearson correlation coefficient function:

$$sim\left(u,v\right)=\frac{\sum_{i\in I_{uv}}\left(r_{u,i}-\overline{r_u}\right)\left(r_{v,i}-\overline{r_v}\right)}{\sqrt{\sum_{i\in I_{uv}}\left(r_{u,i}-\overline{r_u}\right)^2}\sqrt{\sum_{i\in I_{uv}}\left(r_{v,i}-\overline{r_v}\right)^2}} \tag{3}$$

Equations (1)-(3):where $r_{u,i}$ and $r_{v,i}$ are ratings given by user *u* and *v* to item *i* respectively, $\overline{r_u}$ is the average of the ratings made by the user *u*, and $\overline{r_v}$ is the average of the ratings made by the user *v*, I_{uv} indicates the items that users *u* and *v* co-evaluated. I_u is item set that user *u* rated. I_v is item set that user *v* rated.

Although traditional similarity measures is commonly used metric in the process of user-based CF, these choice is not always backed by the nature and distribution of the data in the recommender systems.

III. AN USER INTEREST RECOMMENDATION BASED ON COLLABORATIVE FILTERING

A. Bhattacharyya Coefficient

The Bhattacharyya Coefficient (BC) is calculated as an overlap between the two rating vectors, which can be used to measure the correlation between the ratings of two users. The Bhattacharyya measure has been widely used in signal processing, image processing and pattern recognition research community. In the continuous domain the BC between two density distributions $p_1(x)$ and $p_2(x)$ is defined as follows:

$$BC(P_1,P_2)=\int\sqrt{P_1(x)P_2(x)}dx \tag{4}$$

The BC is defined over a discrete domain X as follows:

$$BC(P_1,P_2)=\sum_{x\in X}\sqrt{P_1(x)P_2(x)} \tag{5}$$

Densities of $p_1(x)$ and $p_2(x)$ are estimated from the given rating data. Let p_u and p_v be the estimated discrete densities of the two users *u* and *v* obtained from rating data. Then, BC similarity between user *u* and user *v* is computed as:

$$BC(u,v)=BC(\hat{p}_u,\hat{p}_v)=\sum_{h=1}^{m}\sqrt{(\hat{p}_{uh})(\hat{p}_{vh})} \tag{6}$$

where *m* is the number of bins; $\hat{p}_{uh}=\frac{\#h}{\#u}$, where #*u* is the number of items by user *u* rated; #*h* is the number of items by user *u* rated with rating value '*h*'.

This can be illustrated with an example. Let $U=(1,0,2,0,1,0,2,0,3,0)^T$, $V=(0,1,0,2,0,1,0,2,0,3)^T$ be the rating vectors of users *u* and *v*, respectively. The ratings lie in {1,2,3,4,5}. Then, BC coefficient between users *u* and *v* can be obtained as:

$$BC(u,v)=\sum_{h=1}^{5}\sqrt{\hat{p}_{uh}\hat{p}_{vh}}=\sqrt{\left(\frac{2}{5}\right)*\left(\frac{2}{5}\right)}+\sqrt{\left(\frac{2}{5}\right)*\left(\frac{2}{5}\right)}+\sqrt{\left(\frac{1}{5}\right)*\left(\frac{1}{5}\right)}+0+0=1 \tag{7}$$

B. Proposed Similarity Measure

Traditional similarity measures have some drawbacks. The cosine similarity focuses on the angle between the vectors of user's ratings and gets high similarity in spite of significant

difference in ratings. The adjust cosine similarity is unable to recognize its positive and negative correlation. The Pearson correlation coefficient shows low (high) similarity regardless of similar (difference) in the ratings. One effective way to solve the problems mentioned above is the new metric JBC, innovated on the basis of BC coefficient and Jaccard similarity. Jaccard similarity measures the probability of having common neighbors between user *u* and user *v* to the number of unions of *u* and *v*'s neighbor nodes. Jaccard is calculated as follows:

$$sim_{Jac}(u,v)=\left|\frac{I_u \cap I_v}{I_u \cup I_v}\right| \tag{8}$$

where I_u is the item by user *u* rated, I_v is the item by user *v* rated.

The proposed a new similarity measure is termed as JBC that combines local and global similarity to obtain final similarity value. Jaccard can be used to describe the global similarity. By using BC coefficient we can calculate similarity of the rating distribution, which can be used to describe the local similarity. The local similarity plays an important role and it provides local information, the higher the BC value is, the more similar the rating distribution of two users are, the closer the user interest is. If ratings are made on a same rating distribution between users *u* and *v* as the BC value is 1 ($BC(u,v)=1$). It does not give any importance to local similarity if the ratings distribution completely different ($BC(u,v)=0$). To provide importance to the number of common users, $sim_{Jac}(u,v)$ is added to $sim_{JBC}(u,v)$, the final similarity can be written as:

$$sim_{JBC}(u,v)=sim_{Jac}(u,v)BC(u,v) \tag{9}$$

C. Interest Weight

Assuming that a user may have a relatively stable interest within a certain short period, the interval time of the user's behaviors play important roles in determining the similarities between users, the smaller interval time, the higher interest weight to ratings. A time span is set to divide linear time into time sequences. This paper employs fitted Ebbinghaus forgetting curve as the decay parameter, when a timeframe goes by every function value is multiplied by a decay factor:

$$w_t=e^{-\frac{t_{ui}-t_0}{T}} \tag{10}$$

where t_{ui} is the rating time by user *u* to item *i*, t_0 is the sampling time by active user, *T* is the time span of the dataset.

This paper adopts JBC similarity measure and adds interest weight to the similarity calculation formula to enhance the recommendation performance. The JBC similarity calculation method based on the interest weights (TJBC) is modified as:

$$sim_{JBC}(u,v)=sim_{Jac}(u,v)BC(u,v)\cdot w_t \tag{11}$$

D. Weighted Prediction Rating

After calculating user's similarities, we rank all the other users that have rated the target item according to their similarities with the active user, and then select the top *k* users as the active user's neighbors for the target item. Considering the user's present behavior and recent behavior should be more relations with time. The interest weight w_t is added to the prediction rating. The predicted rating for active user *u* for item *i* is rewritten as follows:

$$P_{u,i}=\overline{r_u}+\frac{\sum_{v=1}^{n}\left[sim(u,v)\left(r_{v,i}\times w_t-\overline{r_v}\right)\right]}{\sum_{v=1}^{n}sim(u,v)} \tag{12}$$

where $sim(u,v)$ is the similarity between active user *u* and the nearest neighbor user *v*. $r_{v,i}$ is the rating for user *v* to item *i*, $\overline{r_u}$ means the average rating of the user *u*, $\overline{r_v}$ means the average rating of the user *v*.

IV. Experimental Results and Analysis

A. Datasets

In order to observe the effectiveness of system, we used two different real datasets, namely MovieLens_100K, MovieLens_1M in experiments. Brief description of these three datasets is given in Table Ⅰ.

TABLE I. DESCRIPTION OF DATASET IN THE EXPERIMENTS

Name	*Users*	*Items*	*Ratings*	*Sparsity*
Movielens_100K	943	1682	10^5	93.7%
Movielens_1M	6040	3706	10^8	95.81%

B. Evaluation Metrics

To make the experimental results comparable and reproducible, we adopt two well-known metrics, MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error), to measure the closeness of predicted ratings to the actual ones. These metrics are defined as follows:

$$MAE=\frac{\sum_{u,i\in N}\left|r_{u,i}-p_{u,i}\right|}{N} \tag{13}$$

$$RMSE=\sqrt{\frac{\sum_{u,i\in N}\left|r_{u,i}-p_{u,i}\right|^2}{N}} \tag{14}$$

where $p_{u,i}$ and $r_{u,i}$ denote the ture and predicted rating values given by user *u* to item *i*, respectively; *n* denotes the test set. It is clear that lower values of RMSE and MAE correspond to higher recommendation accuracy.

C. Results

This paper uses two different sizes of Movielens datasets in experiments. Movielens datasets are randomly split into a training set and a testing set according to a certain proportion (8:2 in this paper). To avoid overfitting problems, we conduct 5-fold cross-validation experiments, and the average value of ten crossover experiments is taken as the final result.

1) Comparison of traditional similarity on Movielens_100K dataset: To evaluate performance of our proposed based CF. The equation (9) of interest weight adds to the traditional similarity calculation formula. we implement user-based CF using the TJBC and the traditional similarity measures. The number of neighbor *k* is taken as 5, 10, 15, 20, 25 and 30 respectively in the following experiments, the experimental results are shown in Figure II.

It can be seen from Figure II that MAE values on the user-based CF by using the TJBC are lower than the tradition similarity measures, especially using cosine similarity. The results prove our argument that integrating improved similarity and user interest can describe the similarity between users more accurately, as well as alleviate the interest drift problem.

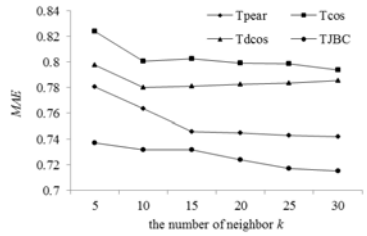

FIGURE II. MAE COMPARISONS BETWEEN PROPOSED METHOD AND TRADITIONAL SIMILARITY

2) Performance comparison between different algorithms on Movielens_100K dataset: We use the proposed method TJBC to improved the collaborative filtering algorithm, the proposed algorithm is called TJBCF. TJBCF is compared with CFBUI proposed in reference [9], ISCF proposed in reference [11] on Movielens datasets.

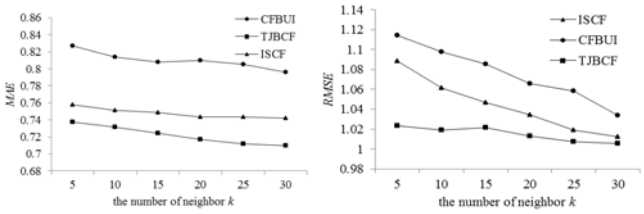

FIGURE III. COMPARION BETWEEN THE PROPOSED ALGORITHM AND THE REFERENCES OVER MAE AND RMSE RESULTS ON THE MOVIELENS_100K DATASET

It is depicted in Figure III that the MAE value of TJBCF is 3.4% lower than ISCF and 10.8% lower than CFBUI. The RMSE value is 2.7% lower than ISCF and 5.6% lower than CFBUI on the Movielens_100K dataset. The performance of the TJBCF is better than the algorithm of references [9] [11].

3) Performance comparison between different algorithms on Movielens_1M dataset

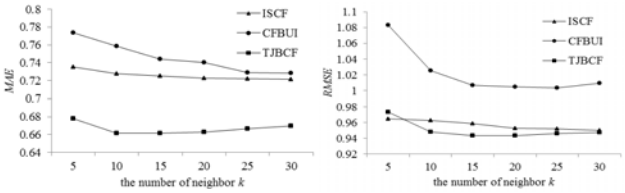

FIGURE IV. COMPARION BETWEEN THE PROPOSED ALGORITHM AND THE REFERENCES OVER MAE AND RMSE RESULTS ON THE MOVIELENS_1M DATASET

Figure IV shows the MAE and RMSE curves for each method in the Movielens_1M dataset. It can be seen that TJBCF always achieves the best performance among the improved algorithms. The reason is that TJBCF not only improves the similarity method but also takes advantage of interest weight, it can describe exactly characteristic of user behaviors. In the Figure III and IV the experimental results show that the proposed algorithm outperforms other algorithms, in terms of MAE and RMSE. With the number of neighbors' changes, the performance of the algorithm barely fluctuates. The results show that TJBCF has the character of low complexity and well stability. The obtained results over different sparsity values show that the proposed method obtained better performance compared to CFBUI specifically for higher sparsity rates.

V. CONCLUSION

This paper analyzes the law of user interest changing over time, and proposes a novel similarity metric: TJBC similarity, from the local and global to analyze similarity using Jaccard as the global similarity and BC coefficient as the local similarity, and exponential gradually forgetting curve is used to update the user interest preference. The TJBCF algorithm is designed on the user-based CF. Experiments on the real rating databases show that TJBCF can be accurately fit user interest preference and provide highly accuracy recommendations. In our future work, we plan to explore other coefficient like spearman rank to further improve recommendation accuracy.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China under Grants No. 61202286 and Foundation for University Key Teacher by Henan Province under Grant No. 2015GGJS-068.

REFERENCES

[1] M Ranjbar, P Moradi, M Azami, M Jalili. "An imputation-based matrix factorization method for improving accuracy of collaborative filtering systems," Engineering Applications of Artificial Intelligence. vol. A46, pp. 58-66, September 2015.

[2] H Koohi, K Kiani. "User Based Collaborative Filtering using Fuzzy C-Means," Measurement. Vol. A91, pp. 134-139, May 2016.

[3] Y Ar, E Bostanci. "A genetic algorithm solution to the collaborative filtering problem," Expert Systems with Applications. Vol. A61, pp. 122-128, May 2016.

[4] F Ortega, A Hernando, J Bobadilla, JH Kang. “Recommending items to group of users using Matrix Factorization based Collaborative Filtering,” Information Sciences. vol. A325, pp. 313-324, February 2016.

[5] Koren Y. “Collaborative filtering with temporal dynamics,” ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Paris, 2009, pp. 89-97.

[6] C Chen, H Yin, J Yao, B Cui. “TeRec: a temporal recommender system over tweet stream,” Proceedings of the Vldb Endowment. Trento, vol. A6, pp. 1254-1257, August 2013.

[7] Y Huang, B Cui, W Zhang, J Jiang, Y Xu. “TencentRec: Real-time Stream Recommendation in Practice,” ACM SIGMOD International Conference on Management of Data. ACM, Melbourne, 2015, pp. 227-238.

[8] M Zhu, S Yao. “A Collaborative Filtering Recommender Algorithm Based on the User Interest Model,” IEEE, International Conference on Computational Science and Engineering. IEEE, Chengdu, 2014, pp. 198-202.

[9] Z Liu. “Collaborative Filtering Recommendation Algorithm Based on User Interests,” International Journal of u- and e- Service, Science and Technology. Suining, vol. A8, pp. 311–320, April 2015.

[10] B K Patra, R Launonen, V Ollikainen, S Nandi. “A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data,” Knowledge-Based Systems. India, vol. A82, pp. 163-177, February 2015.

[11] W Cheng, G Yin, Y Dong, H Dong, W Zhang. “Collaborative Filtering Recommendation on Users’ Interest Sequences,” Plos One. Harbin, vol. A11, pp. 1–17, May 2016.

[12] Y Xiao, A I Pengqiang, C H Hsu, H Wang, X Jiao. “Time-Ordered Collaborative Filtering for News Recommendation,” Mobile Information Systems.Beijing, vol. A12, pp. 53–62, December 2015.