Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 147 (2019) 416–421

Procedia Computer Science

www.elsevier.com/locate/procedia

2018 International Conference on Identification, Information and Knowledge in the Internet of Things, IIKI 2018

An Efficient Similarity Measure for Collaborative Filtering

Yi Mu[a], Nianhao Xiao[a], Ruichun Tang[a,*], Liang Luo[a], Xiaohan Yin[a]

[a]*Ocean University of China, No.238 Songling Road, Qingdao 266100, China*

Abstract

In the field of recommendation system, the memory-based Collaborative filtering has been proven to be useful in lots of practices. Similarity measures like Pearson correlation coefficient tend to only focus on improving as much as possible the accuracy. Handling datasets with different features, exiting measures cannot apply to different types of data simultaneously. In this paper, an improved similarity measure Common Pearson Correlation Coefficient (COPC) was proposed. Unlike existing measures, it strongly depends on chosen distance function, which adhere to the natural property of monotonicity and utilize consensus evaluation measure to capture an optimal value to improve PCC measure. To mitigate sparse problem, we also introduce the Hellinger Distance (Hg) as global similarity to lower the impact of lacking co-rated items. Experimental results on real-world datasets demonstrates that our measure outperformed the existing schemes of predicting ratings.

Keywords: recommendation system; User-based Collaborative filtering; similarity measure; versatility problem; sparse data

1. Introduction

In recent years, as the most notable application of handling the problem of "information overload", Memory-based Collaborative filtering (CF) successfully refine the data between users and items whether in the field of data mining or e-commerce [1, 2]. By providing users with personalized suggestions for products or projects suiting their interests, useful information from the massive data could be automatically obtained or manually extracted by individual users.

* Corresponding author. Tel.:+8613153206505.
E-mail address: tangruichun@mail.ouc.edu.cn

10.1016/j.procs.2019.01.258

Despite the advantage of CF in recommendation system, several problems, including poor versatility and sparse data problem are still existing. Previous research shows that the measure of similarity influenced the prediction accuracy of the algorithm directly. To alleviate the aforementioned problems, various algorithms have been incorporated in different similarity measures. Existing similarity measures mainly depend on the common items rated by users. While, measures like Pearson correlation coefficient (PCC) and its variant cannot fully exploit all the available information. Regularly, the ratings are identified with numbers to assign the arithmetic mean or median as the metric by users [3]. Otherwise, the global similarity of the items should also be considered into calculation.

As the purpose of defining a more effective similarity metric to apply to different types of datasets, an attempt has been made in this paper to address this problem of "versatility". A similarity calculation measure COPC was proposed. In contrast with widely practiced ongoing measures, COPC makes CF more compatible by improving the traditional Pearson correlation coefficient (PCC) calculation measure. To overcome the sparse problem, we estimate the global users' similarity by calculating the Hellinger Distance and Jaccard value of all ratings.

The main contribution of this paper is proposing a similarity model COPC-Hg， which solved the problem that low versatility of existing measures and negative effect caused by the sparsity data in the classical model.

2. Related work

This section briefly reviews some related work about advanced similarity calculation approaches for solving existing problems. Readers may refer to relevant publications herein for more details.

Commonly used measures of calculating similarity include cosine, Jaccard, and Pearson correlation coefficient, etc. Among them, PCC and its variants are prevalently employed all the time because of its practicability. However, measures like PCC are always limited by a single user-item. Faced with sparse data, the accuracy of this measure will be greatly reduced. What' more, most calculation measures are not suitable for applying multiple types of data simultaneously. The impact of these two problems is not negligible.

Bobadilla et al. proposed a singularity measure by using hidden attributes in the CF processes to obtain higher predictive accuracy [4]. However, the performance of the singularity measure is limited when there are fewer singularities of co-rated items. Lee, Yang, and Park proposed measure to capture similarity between users [5], that is, to discover hidden similarity (DHS). This measure is based on the connection on human network (such as the friendship network with friends' friends) and then increased similarity matrix density and predicted coverage. It solves the problem of sparsity problem to a certain extent, while, the connection generated by the transfer relationship can easily cause similar phenomena. Owning to the limitation of improved local similarity, the other way is gradually being applied, that is using global similarity calculation to balance the rare of co-rated items. Desrosiers, Karypi proposed a model that calculates the global similarity between the project and the user by calculating the linear equations related to user similarity and project similarity [6]. M. SAEED and EG MANSOORI et al. proposed a measure based on fuzzy collaborative filtering to calculate the similarity [7].

However, the above measures are still not suitable in the requirement of satisfying the accuracy of different types of data at the same time. For the compatibility, we now know that for different types of datasets, the corresponding calculation measures and recommendation models are often selected according to their characteristics. For large-scale e-commerce companies, multiple sets of different characteristics need to be handled simultaneously. For this consideration, we propose a similarity measure COPC that can more effectively satisfy multiple types of data for recommendation. What's more, to balance spare data problem, we introduced Hellinger Distance as global similarity calculation. The introduction of dynamic weight threshold α, a value obtained by the dynamic calculation of the dataset helped our measure more generalized.

3. Novel scheme for similarity calculation

3.1. Proposed model of COPC-Hg

In this section，a model combining predictions from local and global neighbors was discussed, and the model of COPC-Hg was shown in Fig.1.

In our model, we propose a user-based label distance calculation to improve PCC measure. To obtain the similarity between user u and user v, we first calculate the local similarity and local similarity and global similarity by using

COPC measure and Hellinger distance simultaneously. Secondly, we calculate the weight coefficient α to weigh the local neighbors. Finally, we make a prediction for recommendation by using our model.

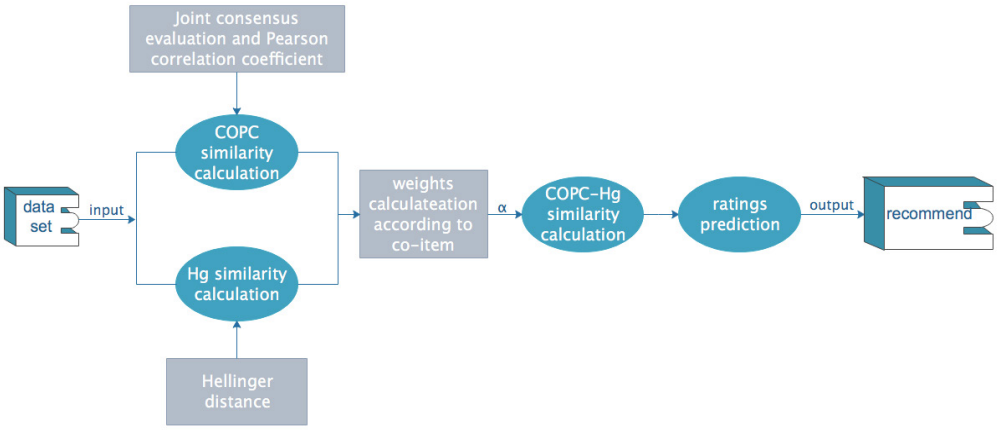

Fig. 1. The model of COPC-Hg measure.

3.2. *Local* similarity

In the field of sensory evaluation, there is a problem of obtaining the consensus evaluation of multiple objects. In order to solve this problem [8,9], Raúl Pérez-Fernández et al. proposed the consensus assessment theory to reasonably estimate the optimal solution so that the total cost of converting the full set to the optimal solution is the minimum. when an object has multiple evaluations from multiple users, the evaluation results can be obtained through reasonable mathematical calculations. The traditional way to solve this problem is to start from the data collection step and obtain the evaluation results of different experts on different objects. The goal is to reach a consensus assessment based on these experts' assessments.

In order to measure the cost of changing a label from the ordinal scale to another one, we considered cost function M on label L and proposed a user-based evaluation model that can obtain the user's consensus grading problem. The matrix of labels for the user is as follows:

$$\mathbb{R}^u = \begin{bmatrix} f_1(U_1) & f_1(U_2) & \cdots & f_1(U_r) \\ f_2(U_1) & f_2(U_2) & \cdots & f_2(U_r) \\ \vdots & \vdots & & \vdots \\ f_n(U_1) & f_n(U_2) & \cdots & f_n(U_r) \end{bmatrix} \tag{1}$$

Let $U = \{U_1, U_2 \cdots, U_r\}$ denotes the set of experts. Let $I = \{I_1, I_2, \cdots, I_n\}$ denotes the set of items. The range of labels is represented as $L = \{L_1, L_2, \cdots, L_l\}$. $f_i(U_u)$ represents the ratings of I_i from I given by U_u from U. The consensus evaluation rating of user U_u is formulated as below:

$$f(U_u) = \arg\min_{L_t \in L} \sum_{i=1}^{O} M(f_i(U_u), L_t) \tag{2}$$

where L_t is in a range from 1 to L_t, and $M(f_i(U_u), L_t)$ denotes the cost of changing rating into label. Then, we use the consensus evaluation result $f(U_u)$ and $f(U_v)$ to represent the rating personalization of u and v. The similarity of COPC is formulated as below:

$$Sim_{COPC}(u,v)=\frac{\sum_{i\in I(u,v)}(r_{ui}-f(U_u))(r_{vi}-f(V_v))}{\sqrt{\sum_{i\in I(u,v)}(r_{ui}-f(U_u))^2}\sqrt{i\in I(u,v)}(r_{vi}-f(V_v))^2} \tag{3}$$

3.3. Global similarity

The ‘‘sparse problem” in CF refers to inability to find a sufficient quantity of good quality neighbors to aid in the prediction process due to insufficient overlap of ratings. This can happen when the ratings matrix is sparse, or the number of users participating is not large. Even when the data is dense enough to allow quality predictions for most users, some users may not have rated enough items, with the result that such users get poor quality predictions.

Therefore, we propose a novel global similarity calculation measure, which utilize the probability estimate measure given by Hellinger distance. Hellinger distance is an approach of measuring two discrete probability distributions. It has been widely used in statistics and is mainly used to measure the separability between classes.

Thus, users whose local neighborhood set is sparse can be balanced by using predictions from the global similarity calculation. The similarity of Hellinger distance is formulated as below：

$$Sim_{Hg}(u,v)=1-\sqrt{\sum_{h=1}^{m}\sqrt{(\hat{P}_{uh})(\hat{P}_{vh})}} \tag{4}$$

where $\hat{P}_{uh}$ represents the percentage, which equals h label's number divides the total number of ratings rated by user u. In global similarity calculation, we also introduced the Jaccard measure to provide more importance to the number of common items, which balanced the effect of Hg measure on the accuracy of global similarity when the volume of data is insufficient.

3.4. Defining and calculating the similarity with weight coefficient

The parameter α serves to adjust the weight that we give to the similarity of COPC with regards to co-rated neighbors, and $(1-\alpha)$ was defined as the weight coefficient of Hellinger Distance similarity. The progress of calculation is counting the proportion of the number of common-item to the total number of ratings to determine the sparsity of the datasets. The function of COPC-Hg is formulated as below:

$$Sim_{COPC-Hg}=\alpha*Sim_{COPC}+(1-\alpha)*(Sim_{Hg}+Sim_{Jaccard}) \tag{5}$$

4. Experimental evaluation

4.1. Data preparation and experimental design

In this section, to evaluate the performance of the proposed measure COPC-Hg, we compared it with other classic algorithms on real world datasets (MovieLens, Jester, Anime, BookCrossing), which is shown on Table 1. Among the entire users in each dataset, 50 users who had watched movies more than 20 times via datasets were randomly selected as target users. Seventy percent of each target user's data were used as training data, and the remaining data were used as test data.

Table 1. Real-world datasets description.

Dataset	Users	Items	Ratings
MovieLens 100k	943	1682	100000
Anime	6900	9927	6337241
Jester	50691	150	1728830
Bookcrossing	105281	340541	1149763

4.1. Experimental results and analysis

In this section, we carry out recommendation experiments based on four empirical datasets and demonstrated the performance of different measures separately.

First of all, we conduct experiments on the MovieLens 100k datasets and Anime. As shown in Fig. 2-3.

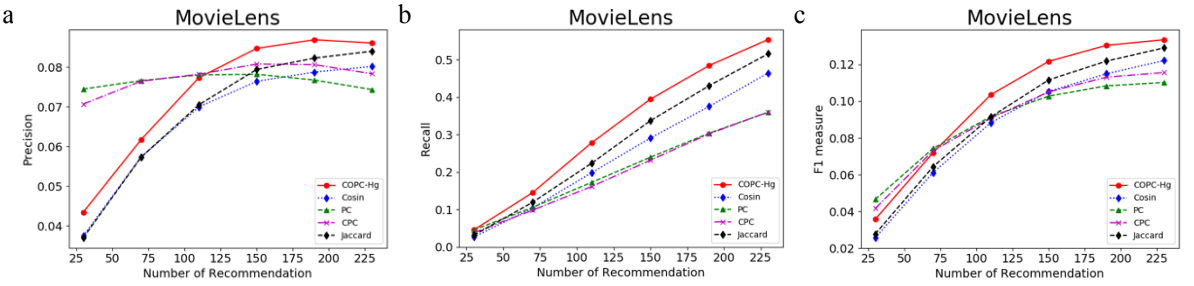

Fig. 2. (a) Precision on MovieLens; (b) Recall on MovieLens; (c) F1 measure on MovieLens ;

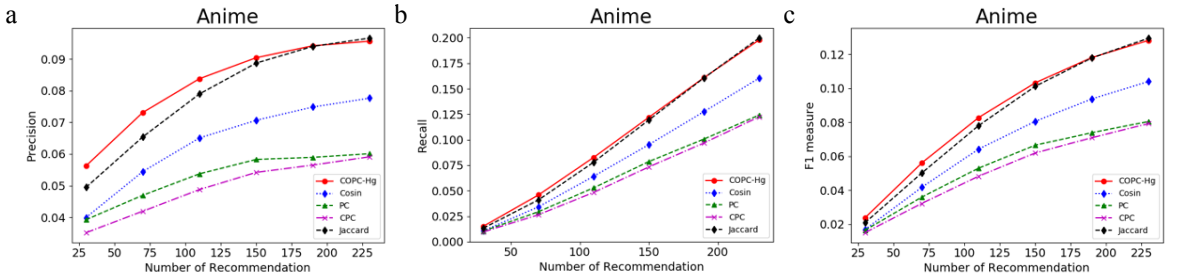

Fig. 3. (a) Precision on Anime; (b) Recall on Anime; (c) F1 measure on Anime;

With the increase of recommendation number, precision, recall and F1-measure are increased gradually. Notably, in terms of accuracy, PC and PCC measures on the MovieLens dataset are in a lower-increased state, but high value. Still, the COPC-Hg measure exceeds the above two measures when the recommended bibliography is around 110, and its recall rate behaves better than other measures. From the F1 value, we can see that with the recommended number gradually increases, COPC-Hg performs better than other measures.

Secondly, we observed the performance of different algorithms on the Jester dataset, as shown in Fig. 4.

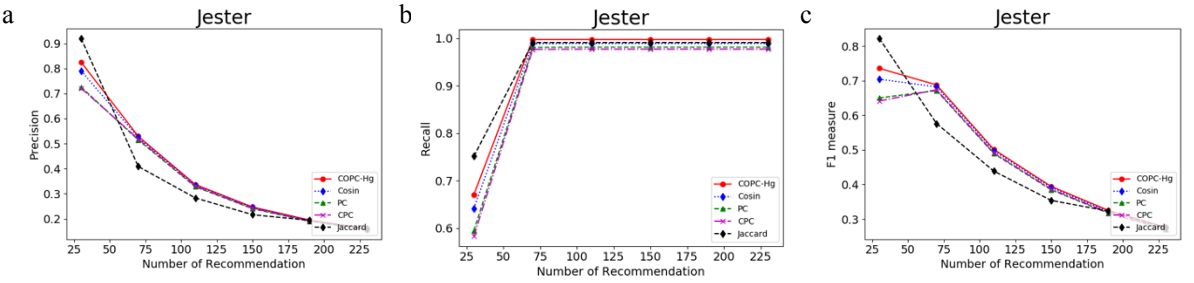

Fig. 4. (a) Precision on Jester; (b) Recall on Jester; (c) F1 measure on Jester;

Different from the above two datasets, the accuracy and F1 values of all measures show a downward trend with the increase of the recommended number, and the curves presents a dense distribution. The Recall rate is on the rise and eventually approaches saturation as the number of recommendation increases. In terms of accuracy, recall and F1, performance of COPC-Hg measure is optimal. Finally, we conducted experiments on BookCroosing dataset. As

shown in Fig. 5, the accuracy fluctuated with the recommended number, and the performance of COPC-Hg measure is also superior to the other groups. The results of our experiments show that our measure COPC-Hg proved to be efficient with good versatility and also improved the quality of CF-based recommendation system in precision, recall, F1 measure.

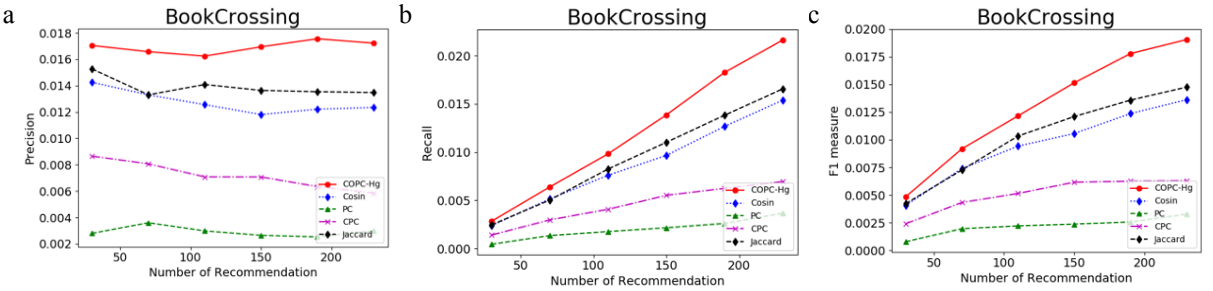

Fig. 5. (a) Precision on BookCrossing; (b) Recall on BookCrossing; (c) F1 measure on BookCrossing;

5. Conclusions and future work

Most measures for calculating similarity in CF area mainly focus on the precision instead of compatibility. Especially in large-scale e-commerce companies, multiple sets of different characteristics need to be handled simultaneously. Therefore, with the consideration of getting the PCC measure better versatility and compatibility, we introduced the consensus evaluation measure into it. We also utilize the probability estimate measure given by Hellinger distance to solve sparse data problem. Results shows that our measure outperformed other measures on the situation facing datasets with different characters.

In the future work, we will extend this research in two directions. Firstly, some important factors like diversity, the popular degree of items and viewing time would be considered into recommender system to further improve the performance. Secondly, We will be committed to making breakthroughs in computational efficiency and scalability. The deep learning method would be introduced into our algorithm to make algorithm work better.

Acknowledgements

This work was supported by the National Key R&D Program of China (No. 2017YFC0806205).

References

[1] Adomavicius G, Tuzhilin A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions[M]// Multimedia Services in Intelligent Environments. Springer International Publishing, 2013:734-749.

[2] Bobadilla J, Ortega F, Hernando A. Recommender systems survey[J]. Knowledge-Based Systems, 2013, 46(1):109-132.

[3] Xia N, Desrosiers C, Karypis G. A Comprehensive Survey of Neighborhood-Based Recommendation Methods[M]// Recommender Systems Handbook. Springer US, 2015:107-144.

[4] J. Bobadilla, F. Ortega, and A. Hernando. A collaborative filtering similarity measure based on singularities[J]. Inf. Process. Manage, 2012: 204–217,

[5] Lee S, Yang J, Park S. Discovery of hidden similarity on collaborative filtering to overcome sparsity problem[J]. In Proceedings of discovery science: Seventh international conference, 2004:396–402.

[6] Desrosiers C, Karypis G. Solving the sparsity problem: Collaborative filtering via indirect similarities. Technical Report, University of Minnesota, 2008:08-044.

[7] Saeed M, Mansoori E G. A Novel Fuzzy-Based Similarity Measure For Collaborative Filtering To Alleviate The Sparsity Problem[J]. Iranian Journal of Fuzzy Systems, 2017, 14(5).

[8] Pérez-Fernández R, Sader M, Baets B D, et al. Joint consensus evaluation of multiple objects on an ordinal scale: an approach driven by monotonicity[J]. Information Fusion, 2017, 42.

[9] Pérez-Fernández R, Rademaker M, Baets B D. Monometrics and their role in the rationalisation of ranking rules[J]. Information Fusion, 2017, 34:16-27.