# A Combinative Similarity Computing Measure for Collaborative Filtering

Lin Guo
Systems Engineering Institute
Xi’an Jiaotong University
Xi’an, China
guolin.xjtu@stu.xjtu.edu.cn

Qinke Peng
Systems Engineering Institute
Xi’an Jiaotong University
Xi’an, China
qkpeng@mail.xjtu.edu.cn

***Abstract*—Similarity method is the key of the user-based collaborative filtering recommend algorithm. The traditional similarity measures, which cosine similarity, adjusted cosine similarity and Pearson correlation similarity are included, have some advantages such as simple, easy and fast, but with the sparse dataset they may lead to bad recommendation quality. In this article, we first research how the recommendation qualities using the three similarity methods respectively change with the different sparse datasets, and then propose a combinative similarity measure considering the account of items users co-rated. Compared with the three algorithms, our method shows its satisfactory performance with the same computation complexity.**

***Keywords-user-based collaborative filtering; similarity method; item’s account users co-rated;***

## I. INTRODUCTION

With the large amount of information on the Internet, it is hard for Netizens to select what they need or like. The recommender system has emerged in such a background and it can actively recommend items to users according to their interests and behaviors. As one kind of recommender system, Collaborative Filtering system has been very successful in both research and applications such as GroupLens [1], Web Watcher [2] and Let’s Browse [3]. User-based CF algorithm is one of the most popular techniques in CF system and it utilizes the similarity among profiles of users to recommend interesting items. The *k*-Nearest Neighbor (KNN) method is a popular way for the realization of user-based CF system. Its key technique is to calculate the similarity between target user and the others, and then find the *k* nearest neighbors to predict the target user’s interest [4]. To calculate users’ similarity, there are three basic methods, and they are cosine similarity, adjusted cosine similarity and Pearson’s correlation similarity respectively [5]. The three methods are simple, intuitive and easy to implement. However, because of the sparse data they may bring about unsatisfied recommendation qualities. To solve the sparsity problem, Chun Zeng et al [4] present a matrix conversion method for similarity measure to improve the accuracy of the collaborative filtering algorithm. Xiangwei Mu et al [6] propose stability degree to improve the accuracy of collaborative filtering both based on item and user. Balabanovic et al [7] and Claypool et al [8] put forward a content-based Collaborative Filtering method, which utilizes the contents browsed by users to compute the similarity among users. Sarwar et al [9] uses Latent Semantic Indexing (LSI) to capture the similarity among users and items in a reduced dimensional space. Yu et al [10] uses a feature-weighting method to improve the accuracy of Collaborative Filtering algorithms. Lei Shen and Yiming Zhou [11] apply a basic fractional function and an exponential function to calculate the similarity between users by taking both common features and different features into consideration.

In this paper, we first compare the basic similarity methods with different sparse datasets, and acquire the change trend of recommendation qualities using the three algorithms with the different sparse data. Then according to the comparison experimental results we present a new combinative similarity measure. It is a combination of adjusted cosine similarity and Pearson’s correlation similarity and takes the amount of items users co-rated into consideration. The experimental results of our method show that it enhances the precision of recommendation compared with the three basic similarity algorithms. Meanwhile, since our method is the combination of basic algorithms, the computational expense is equal to the basic methods. Therefore, our method outperforms the three basic algorithms with the same costs.

The rest of the paper is organized as follows. The next section provides an analysis to the three traditional similarity methods. Section 3 shows the process and result of comparing the three methods under ten groups of different sparse datasets. Section 4 describes our combinative method for similarity measure. In section 5 we describe the experimental work for our method and discuss the results. The final section provides some conclusion and directions for future research.

## II. BASIC SIMILARITY METHODS

User-based CF is also called nearest-neighbor based Collaborative Filtering [6]. It first finds target user’s nearest-neighbors, and then combines the preferences of neighbors to produce a prediction or top-N recommendation for the target users. Similarity computing which measures the similarity between two users is the most important part of user-based CF. Choosing a proper similarity method can obviously improve the performance of user-based CF. The three basic similarity methods are as follows:

- **Cosine Similarity** In this case, two users are regarded as two vectors in the *n* dimensional item space. The similarity between them is measured by computing the cosine of the angle between these two vectors. Formally, similarity between users *i* and *j* is given by

Published by Atlantis Press, Paris, France.

$$sim(i,j)=\frac{I\cdot J}{\|I\|\|J\|}=\frac{\sum\limits_{c\in Item} R_{ic}R_{jc}}{\sqrt{\sum\limits_{c\in Item} R_{ic}{}^{2}}\sqrt{\sum\limits_{c\in Item} R_{jc}{}^{2}}} \tag{1}$$

where $I,J$ represent the $n$ dimensional vectors that users $i$ and $j$ rated on the $n$ items; $Item$ represents the whole items; $R_{i,c},R_{j,c}$ denote the ratings user $i$ and $j$ on the item $c$.

- **Adjusted Cosine Similarity** Basic cosine measure has one important drawback that the differences in rating scale between different users are not taken into account. The adjusted cosine similarity offsets this drawback by subtracting the corresponding user average rating from each co-rated pair. Formally, the similarity between user $i$ and $j$ is given by

$$sim(i,j)=\frac{\sum\limits_{c\in I_{ij}}\left(R_{ic}-\overline{R_i}\right)\left(R_{jc}-\overline{R_j}\right)}{\sqrt{\sum\limits_{c\in Item}\left(R_{ic}-\overline{R_i}\right)^2}\sqrt{\sum\limits_{c\in Item}\left(R_{jc}-\overline{R_j}\right)^2}} \tag{2}$$

where $I_{ij}$ represents the items that user $i$ and $j$ co-rated; $\overline{R_i}$ , $\overline{R_j}$ denote the average rating of user $i$ and $j$ .

- **Pearson's Collection Similarity** In this case, similarity between users $i$ and $j$ is measured by computing the Pearson correlation. To make the correlation computation accurate we isolate the co-rated cases. The correlation similarity is given by

$$sim(i,j)=\frac{\sum\limits_{c\in I_{ij}}\left(R_{ic}-\overline{R_i}\right)\left(R_{jc}-\overline{R_j}\right)}{\sqrt{\sum\limits_{c\in I_{ij}}\left(R_{ic}-\overline{R_i}\right)^2}\sqrt{\sum\limits_{c\in I_{ij}}\left(R_{jc}-\overline{R_j}\right)^2}} \tag{3}$$

## III. COMPARISON of BASIC SIMILARITY METHODS

### A. *Experimental Datasets*

Our experimental datasets are from the DouBan website (http://movie.douban.com/) which supplies us with plenty of evaluations from users to movies they have watched. The experimental datasets include ten groups of different sparse data which includes the same 733 users and 687 items. In order to describe the sparse degree of a dataset, we introduce the *closeness* concept as follows:

$$closeness=\frac{num}{m\times n} \tag{4}$$

where $num$ represents the number of rating samples; $m,n$ denote the number of users and items respectively.

According to the definition of the *closeness* concept, the higher *closeness* is, the less sparse the data set is. The experimental data sets are shown in table 1.

TABLE 1 Different Sparse Experimental Datasets

| *num* | *m* | *n* | *closeness* |
|---|---|---|---|
| 24536 | 733 | 687 | 0.0487 |
| 48718 | 733 | 687 | 0.0967 |
| 72999 | 733 | 687 | 0.1450 |
| 97139 | 733 | 687 | 0.1929 |
| 121233 | 733 | 687 | 0.2407 |
| 145580 | 733 | 687 | 0.2891 |
| 169794 | 733 | 687 | 0.3372 |
| 194001 | 733 | 687 | 0.3852 |
| 218237 | 733 | 687 | 0.4334 |
| 242124 | 733 | 687 | 0.4808 |

### B. *Experimntal Procedure*

In our experiment, we divide 80% of the dataset into training set and 20% into test set. And the experiment procedure is as follows:

First, we construct the $m\times n$ user-item rating matrix based on our experiment dataset.

Second, we compute the similarity between users using cosine similarity, adjusted cosine similarity and Pearson's correlation similarity.

Third, we choose $k$ nearest neighbors for every user. In the experiment, we let $k$ take 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50.

Fourth, we compute a prediction of the target user's rating to an item from a combination of the selected neighbors' ratings. The prediction formula is as follows:

$$P_{uc}=\overline{R_u}+\frac{\sum\limits_{v\in KNB} sim(u,v)\left(R_{vc}-\overline{R_v}\right)}{\sum\limits_{v\in KNB}\left|sim(u,v)\right|} \tag{5}$$

where $P_{uc}$ denotes the prediction of user $u$'s rating to item $c$ ; $KNB$ represents the neighbors of user $u$ ; $sim(u,v)$ denotes the similarity of user $u$ , $v$ .

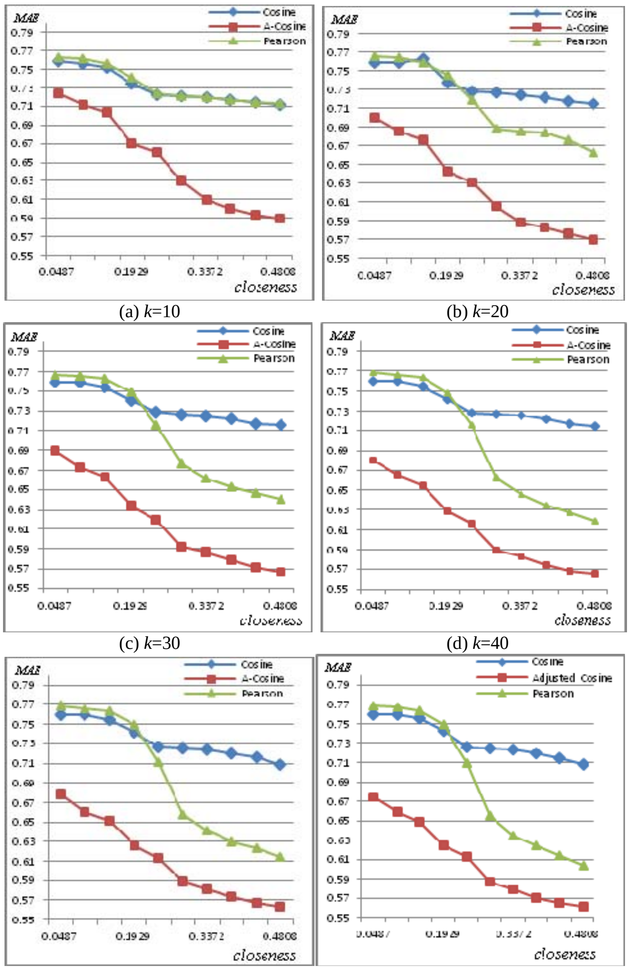
Fifth, we use $MAE$ as the evaluation metrics. The less $MAE$ is, the better recommendation performance is. The formula is as follows:

$$MAE=\frac{\sum\left|P_{uc}-R_{uc}\right|}{n} \tag{6}$$

where $n$ denotes the total number of prediction.

### C. *Experimntal Results*

In the experiment, we compare the three basic similarity methods under 10 groups of different sparse datasets, and for every dataset we also test ten groups of different $k$ . The result shows that for different $k$ the transformation trends of $MAE$ using the three algorithms with different sparse datasets are almost same. So in order to save the space, we just demonstrate the result when $k$ takes 10, 20, 30, 40, 45, 50 in Figure 1.



(a) $k$=10 (b) $k$=20

(c) $k$=30 (d) $k$=40

(e) $k$=45 (f) $k$=50

Figure 1 Comparison of recommendation quality: cosine, adjusted cosine and Pearson

From Figure 1, we can conclude that:

- For different $k$, the transformation trend of $MAE$ is first down and then almost stable with cosine, is also first down and then almost stable with adjusted cosine and is first down slowly and then rapidly with Pearson's correlation.
- Comparing the three methods and the result is that at first(when the data is relatively sparse) adjusted cosine is the best one; and as the data becomes closer, the *MAEs* of the three methods all decline, but the decline range of adjusted cosine is greater; when the data's *closeness* comes to a certain value, the decline range of Pearson's correlation becomes greater than that of adjusted cosine; in the end, the *MAEs* of cosine and adjusted cosine both approximate changeless, but the *MAE* of Pearson's correlation is still down. So we can assume that when the data's *closeness* is great enough the *MAE* of Pearson's correlation is less than that of adjusted cosine, and this assume is the key of our method.

## IV. Combination of Basic Similarity Methods

According to the results of comparing the three similarity methods, we introduce a new similarity measure which is a combination of adjusted cosine similarity and Pearson's correlation similarity considering the account of items which two users co-rated.

### *A. Combinative Similarity Measure*

As the sparseness of the dataset implicates the amount of items users co-rated, the core of our new measure is that when the amount of items users $i, j$ co-rated is small, we use adjusted cosine to compute their similarity; otherwise use Pearson's correlation. First we set two thresholds $\alpha_1, \alpha_2$ to demarcate the small and the large amount of items two users co-rated. That is to say when $Num_{ij} < \alpha_1 * \overline{Num}$ we define $Num_{ij}$ as small amount; when $Num_{ij} > \alpha_2 * \overline{Num}$ we define it as large ($Num_{ij}$ denotes the number of items two users co-rated; $\overline{Num}$ denotes the average number of items all users rated). Then our new measure is as follows:

$$sim(i,j) = \frac{\sum_{c\in I_{ij}} \left(R_{ic}-\overline{R_i}\right)\left(R_{jc}-\overline{R_j}\right)}{\sqrt{A_1+\lambda A_2}\sqrt{B_1+\lambda B_2}} \qquad (7)$$

where $A_1 = \sum_{c\in I_{ij}} \left(R_{ic}-\overline{R_i}\right)^2 \qquad A_2 = \sum_{c\in Item \;\; but \;\; c\notin I_{ij}} \left(R_{ic}-\overline{R_i}\right)^2$

$$B_1 = \sum_{c\in I_{ij}} \left(R_{jc}-\overline{R_j}\right)^2 \qquad B_2 = \sum_{c\in Item \;\; but \;\; c\notin I_{ij}} \left(R_{jc}-\overline{R_j}\right)^2$$

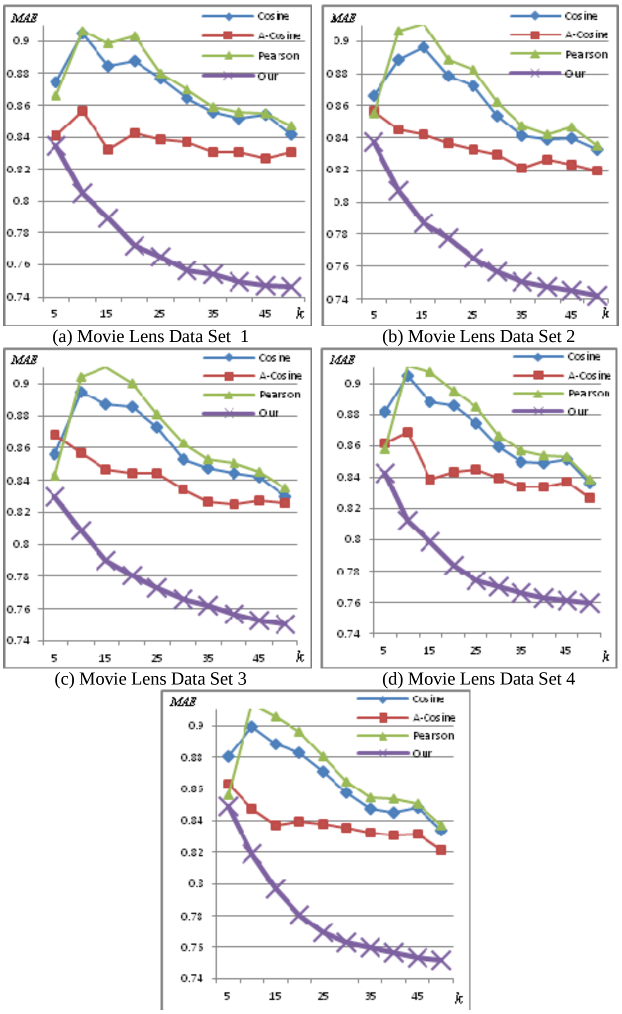
$$\lambda = \begin{cases} 1 & if \quad Num_{ij} \le \alpha_1 \overline{Num} \\ \dfrac{Num_{ij} - \alpha_2 \overline{Num}}{(\alpha_1-\alpha_2)\overline{Num}} & if \quad \alpha_1 \overline{Num} < Num_{ij} < \alpha_2 \overline{Num} \\ 0 & if \quad Num_{ij} \ge \alpha_2 \overline{Num} \end{cases}$$

From the formula, we can prove that when $Num_{ij} \le \alpha_1 * \overline{Num}$, $\lambda = 1$, then the new method is equal to adjusted cosine; when $Num_{ij} \ge \alpha_2 * \overline{Num}$, $\lambda = 0$, then the new method is equal to Pearson's correlation; when $\alpha_2 * \overline{Num} < Num_{ij} < \alpha_1 * \overline{Num}$, the new method is a combination of the basic two algorithms.

### *B. Experiment on the Combinative Similarity Measure*

In order to prove our new measure can boost the recommendation quality, we take an experiment to compare it with the traditional similarity methods using the Movie Lens datasets. The datasets include 5 groups of dataset, and each one has 100000 samples, which include 943 users and 1682 items.

The experimental procedure is the same as the former except an addition to use the new method to compute user's similarity in step 2. Besides, we try to take different values for parameters $\alpha_1, \alpha_2$, but due to limited space, we just show the results when $\alpha_1 = 0.4, \alpha_2 = 1.3$ as Figure 2.



(a) Movie Lens Data Set 1

(b) Movie Lens Data Set 2

(c) Movie Lens Data Set 3

(d) Movie Lens Data Set 4

(e) Movie Lens Data Set 5

Figure 2 Comparison of recommendation quality: cosine, adjusted cosine, Pearson and combinative method

According to the experimental results, we can conclude that for different $k$, the $MAE$ with our method is less than that of cosine, adjusted cosine and Pearson's correlation. The transformation trend of $MAE$ using our combinative method is first down and then almost changeless as k increases. By statistics, the average $MAE$ of our method is reduced by 4.8%, 3.4%, and 4.4% when compared with that of cosine, adjusted cosine and Pearson's correlation respectively. However, from the formula of our method, it is clear that the computation complexity of it is the same as the other three, except that we need to calculate the amount of items any two users both rated which can be done in the process of constructing the user-item matrix. Therefore, our method is not only better than the three in improving the recommendation performance but also the same as them in computation complexity.

## V. Conclusion

This paper first compares the three traditional similarity methods on the different sparse datasets and then proposes a combinative similarity measure. Our method is a combination of adjusted cosine similarity and Pearson's correlation similarity based on the account of items users co-rated. Experimental results show that it outperforms the three traditional similarity measures with equal computation cost.

## Acknowledgment

This work is supported by the National Natural Science Fund of China (61173111, 60774086) and the National Doctoral Fund (20090201110027).

## References

[1] JA Konstan, BN Miller, and D Maltz, GroupLens: Applying Collaborative Filtering to Usenet News, Communications of the ACM, Mar, 1997, 40(3):77-87.

[2] T Joachims, D Freitag, and T Mitchell, WebWatcher: A Tour Guide for the World Wide Web, In Proceedings of the International Joint Conference on Artificial Intelligence, Aug, 1997, 770-777.

[3] H Lieberman, N Van Dyke, and A Vivacqua, Let's browse: A Collaborative Web Browsing agent, In Proceedings of the International Conference on Intelligent User Interfaces, Jan, 1999, 65-68.

[4] Chun Zeng, Chunxiao Xing, and Lizhu Zhou, Similarity Measure and Instance Selection for Collaborative Filtering, In Proceedings of the 12th International Conference on World Wide Web, May, 2003, 652-658.

[5] Xiaobei He, Yuan Luo. Mutual Information Based Similarity Measure for Collaborative Filtering, Progress in Informatics and Computing (PIC), 2010 IEEE International Conference, Dec, 2010, 1117-1121.

[6] Xiangwei Mu, Yan Chen, and Shuyong Liu, Improvement of Similarity Algorithm in Collaborative Filtering Based on Stability Degree, Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference, Aug, 2010, 106-110.

[7] M Balabanovic, Y Shoham, Fab: Content-based Collaborative Filtering Recommendation, Communication of the ACM, Mar, 1997, 40(3):66-72.

[8] M Claypool, A Gokhale, and T Miranda, Combining Content-based and Collaborative Filters in an Online Newspaper, In ACM SIGIR Workshop on Recommender System-A Case Study, In ACM WebKDD Workshop, 2000.

[9] B Sarwar, G Karypis, and J Konstan, Application of Dimensionality Reduction in Recommender System-A Case Study, In ACM WebKDD Workshop, 2000.

[10] K Yu, Z Wen, and X Xu, Feature Weighting and Instance Selection for Collaborative Filtering, In Proceedings of the 2nd International Workshop on Management of Information on the Web-Web Data and Text Ming, 2001.

[11] Lei Shen,Yiming Zhou,A New User Similarity Measure for Collaborative Filtering Algorithm. Computer Modeling and Simulation, 2010, ICCMS'10, Second International Conference, Jan, 2010,375-379.