CORRELATION AMONG SIMILARITY MEASUREMENTS FOR COLLABORATIVE FILTERING TECHNIQUES: AN IMPROVED SIMILARITY METRIC

Swaleha Zubair*, MuaadhAbdo Al Sabri, Afreen Khan

Department of Computer Science, Aligarh Muslim University, Aligarh, India

ABSTRACT

Information is rising exponentially over the Internet. The World Wide Web has emerged as a treasure trove of knowledge and provide relevant information pertaining to any exclusive topic as per the individual's performance or demand. Frequently, the user gets confused while seeing such a large number of over the Internet to choose which one to buy. In this situation, it is essential to filter the available information so as to recommend user about items and determine what different users prescribe. To avoid information overload, we can employ the recommended system that helps us to effectively filter, prioritize and deliver tremendously vital information. A recommendation system refers to a system that can personalize or filter preference in a set of items. Various similarity measures are the key to the analysis. The collaborative filtering recommendation system is supported to be the best approach for personalized user or service recommendations. User-based collaborative filtering approach possesses certain shortcomings, thus item-based collaborative filtering method is taken into consideration. To fill this gap, we compared correlation similarity measures and the distance similarity measures to study the performance of various existing similarity calculation models in order to enhance the recommendation performance. The results of the study were then used to develop an improved method by employing statistical accuracy metrics to give the most accurate recommendation. Therefore, similarity measures can be assessed and contrasted with the outcomes of the similarity measures as discussed in this study. The objective of this paper is to identify appropriate distance measures for datasets and furthermore to facilitate comparison and assessment of the proposed similarity measures with that of conventional ones.

Keywords- collaborative filtering, recommendation system, similarity measurement, correlation, distance

[1] INTRODUCTION

The recommendation system (RS) relies on statistical methods and knowledge discovery techniques to recommend filtering category. A collaborative filtering (CF) approach assesses the preferences of other neighbourhood customer based on their past preference and priorities [1,2]. There are two main ways to recommend items in the collaborative filtering category: the neighbourhood-based CF and the model-based CF [1,2,3]. Neighbourhood-based CF is a simple method which responds to any user request promptly as it does not involve any learning stage [4]. An advantage of a neighbour-based approach is that it works with a single parameter (K-number for the neighbourhood) while a model-based approach works with more than one parameter (learning
parameter $g$, regulation parameters, etc.). In general, collaborative filtering exploits a similarity scale to find active user neighbours and common elements of a candidate [1,5].

The CF algorithm process first gathers information of users to create a user profile or sample of forecasting tasks, including the user attribute, behaviour, or resource contents etc [6]. Next step involves the calculation of similarities among items and users. In this paper, the chief focus is on the similarity measurement and the common properties of similarity metrics, by making use of the following similarity measurements:

1. Sim $(U, V) = 1$ only if $U = V$ (maximum Similarity).
2. Sim $(U, V) = 0$ only if $U \neq V$
3. Sim $(U, V) = Sim (V, U)$ for all $U$ and $V$
   where, Sim $(U, V)$ is the similarity between data objects, $U$ and $V$.

After identifying the similarities, next we analyzed the drawback of the similarity measures, and then provided the hypothesis and motivation for the suggested similarity measure approach. It involves a mathematical model of the proposed similarity measure approach.

Let us assume that:

$U = \{ u_1, u_2, u_3, \ldots, u_N \}$ and
$V = \{ v_1, v_2, v_3, \ldots, v_M \}$ are the set of users and items respectively.

The matrix user-item rating is denoted as:

$R = ( r_{i,j} )_{N \times M} \cdot i = 1,2, \ldots, N, j = 1,2,3, \ldots, M$

In this paper, we illustrate different methods to expand the capabilities of recommendation systems. We compared the correlation similarity measures and the distance similarity measures in order to study the performance of different similarity calculation models that exist. Furthermore, the findings of the study were then used to build a method by using statistical accuracy metrics to give the most precise recommendation. However, before doing this, we first present related work and theoretical background of similarity models in Section 2 and Section 3. Then, in Section 4 we discuss the experimental results, followed by future study and conclusion in Section 5.

[2] RELATED WORK

The use and existence of the recommendation system started back in the 1990s and became an essential component for online users in the procurement of items and browsing. Recommendation system guides the user to find the right information or product [1-5]. The best method in RS is collaborative filtering, making automatic predictions about the interests of a user by analysis preferred information from the nearest users. To find the nearest user there are many similarity measurements. Earlier and recent studies focused on cosine similarity and person similarity measurement [3,9,11]. Al Hassanieh, Demerjian [2], discussed the drawback of PC and settled this issue by employing a weighted Pearson correlation coefficient. Suryakant and T. Mahara [8], observed that the Euclidean Distance Measure consistently performs well and produces better quality results compared to other similarity measures. H. Liu, Z. Hu, A. Mian [4] et. al. provided enhancement and solved the drawback of cosine similarity measure by adjusted cosine similarity. F. O. Isinkaye [21] concluded that Pearson’s Correlation Coefficient by providing a better outcome. S. Al-anazi [9] proposed approaches such as Squared difference, cosine, and Person correlation coefficient. However, the result was not remarkable, especially in the cold user problem. One more similarity metric based on the rating to enhance the accuracy of the recommendation system was proposed in the later years [6].

The item-based recommendation system [11] was proposed because it is better and faster as compared to the user-based recommendation system. A series of recent studies have demonstrated that
still, the system have many disadvantages regarding the amount of data, cold star problem, scalability, and data sparsity, etc.

[3] SIMILARITY MODELS

Collaborative filtering’s main challenge is to identify a matrix of distance or similarities (item-item similarity or user-user similarity) among a set of items from the dataset [10]. Similarity can be measured either as a correlation or distance (dissimilarity) and majority of similarities measure can convert it into the distance and vice-versa [12,13]. An outline of the similarity model has been illustrated in Figure 1.

![Figure 1. Similarity Model](image_url)

Following the above diagram, we elucidate each similarity type by focusing on the disadvantages, as to why it did not work and why the other similarity metrics were needed.

3.1 Correlation Similarity Measurement

Correlation measures the linear relationship between variables [14,16,17]. It is a type of numerical measurement, which measures as to how similar two data objects are; higher when objects are more alike often falls in the range (0,1) [14,17]. There are many correlation measurement models. The most effective ones are described below:

1. **Pearson correlation similarity (PCS):** It is a popular measure in recommendation systems where the result shows a value between [-1 and +1]. Where -1 is a negative correlation while +1 is a high positive correlation. 0 value shows no relation and
therefore sometimes it is also called a zero-order correlation [19]. The PCS formula is given as [26]:

\[
\text{PCS}(u, v) = \frac{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)^2 (r_{vi} - \bar{r}_v)^2}}
\]

(1)

where \(u\) is the user \(u\), \(v\) is the user \(v\), \(I_{uv}\) is the list of items rated by both users \(u\) and \(v\), \(r_{ui}\) is the rating of user \(u\) on item \(i\), and \(\bar{r}_v\) is the average of the ratings provided by user \(v\).

The main drawback of this measure is that it does not consider the personal impact of web services on the calculation of similarity. It also does not show an accurate result when a particular user has rated only one variable or when the two users have one rating in common [2].

2. **Constrained Pearson correlation coefficient (CPCS):** This is a variant of PCC called a Constrained Pearson correlation coefficient [19]. It helps in overcoming the drawbacks found in PCC and consider the impact rating whether positive or negative.

\[
\text{CPCS}(u, v) = \frac{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)^2 (r_{vi} - \bar{r}_v)^2}}
\]

(2)

where \(\bar{r}_v\) the median value in the rating scale.

For example, if we have from 1 to 20 the median value is 10.5.

3. **Weighted Pearson correlation similarity (WPCS):** PCS does not consider the size of the group of subscribed users [19, 20]. To solve this problem, weighted Pearson’s correlation coefficient was suggested and the equation could be written as:

\[
WPCS(u, v)^{wpcc} = \begin{cases} 
\text{sim}((u, v)^{pcc}) \frac{|I|}{|I| \leq H} & \text{if } |I| \leq H \\
\text{sim}((u, v)^{pcc}), \text{otherwise} & \text{otherwise}
\end{cases}
\]

(3)

where \(H\) is an experimental value.

4. **Sigmoid function-based Pearson correlation similarity (SFBPCS):** Since the drawbacks in all previous models are there and all of them reinforce by using this common function by increasing the logistic curve, which usually describes restricted growth or cumulative quantity.

\[
SFBPCS(u, v)^{spcc} = \text{sim}(u, v)^{pcc} \cdot \frac{1}{1 + \exp\left(-\frac{|I|}{2}\right)}
\]

(4)

5. **Spearman Rank Correlation similarity (SRCS):** A calculating similarity in Spearman Rank Correlation depends on ranks instead of ratings and this procedure avoids the problem of the rating of normalization. It does not work well for partial orderings [3]. There are high similarities even if the ratings are similar. Equation of SRCS is:
SRCS(u, v) = \frac{6 - \sum_{h_0}^{n_i} d_h^2}{n_i (n_i^2 - n)} \quad (5)

where \(d_h\) is the difference in ranks of the item, h is co-rated by both users, \(n_i\) is a number of items co-rated by both users.

The chief disadvantage with SRCS is, that it does not operate well for partial orderings.

6. Jaccard Similarity (JACS): The Jaccard similarity coefficient is used to compare the members of two groups to identify common and distinct members.

\[ JACS(U, V) = \frac{|I_u| \cap |I_v|}{|I_u| \cup |I_v|} \quad (6) \]

Like others, the Jaccard similarity approach possesses its own disadvantages. Likewise, it does not calculate the absolute value into account. It lays emphasis on rating-based CF methods and also undergoes from limited or no overlapping items [25].

7. Mean squared difference (MSD): Mean square difference assesses the similarity between two users- u and v as the inverse of the average square difference between ratings given by two users on the same items [7]. MSD does not take into account the number of common ratings instead it includes absolute ratings.

\[ MSDS(U, V) = 1 - \frac{\sum_{p \in I} (r_{u,p} - r_{v,p})^2}{|I|} \quad (7) \]

Mean square difference’ disadvantage is limited in comparison with PCS because it does not capture negative correlations between users’ preferences. With these negative correlations, the accuracy of the rating prediction can be improved [7].

8. Jaccard and MSD Similarity (JASMDS): Jaccard and mean squared difference is combined to form a new metric and to solve the drawback of two similarity models.

\[ JASMDS(u, v) = sim(u, v)^{jaccard} \cdot sim(u, v)^{MSD} \quad (8) \]

where \(r_u\) and \(r_v\) represents the set of items by user u and v rated respectively.

3.2 Distance Similarity Measurement

The familiar definition of distance is to find the length between two variables or two points in two dimensions case or many dimensions case and written into a new matrix as a distance matrix of size: \(u_i \times v_i\) [18]. There are many distance measures. Selection of the most commonly used and most effective measures are described below:

1. Cosine similarity (CS): In calculating the similarity between two or more items of users, CS is most widely used with CF to measure the distance between two-
dimensional using cosine angle [20]. Here we represent the first dimensional by user1 and the second dimensional by user2. Equation of CS is (5):

$$\cos(u, v) = \frac{\mathbf{r}_u \cdot \mathbf{r}_v}{||\mathbf{r}_u|| \cdot ||\mathbf{r}_v||}$$  \hspace{1cm} (9)

\(r_u\) is the vector of user u and \(r_v\) is the vector of user v rating respectively. \(||\cdot||\) represents a vector.

The main disadvantage of cosine-based similarity is that the difference in the ratings given to the items between the various users is not taken for the calculation.

2. Adjusted cosine similarity (ACOS): It enhances the performance of cosine similarity; however, it did not take into account user rating preferences [20].

The equation can be written as

$$\text{ACOS}(u, v) = \frac{\sum_{i\in P}(d_{1,p} - d_{1i})(d_{2,p} - d_{2i})}{\sqrt{\sum_{p\in P}(d_{1,p} - d_{1i}) \cdot \sqrt{\sum_{p\in P}(d_{2,p} - d_{2i})}}$$  \hspace{1cm} (10)

where \(P\) represents the set of all items. \(p \in P\) item if user u has not rated, the rating of d is zero.

ACOS measure is an improved form of vector-based similarity where various users have distinct ratings i.e. few users might give a higher rating and others might give lower ratings to the items [21]. To eliminate this disadvantage from vector-based similarity, each user average rating is subtracted from each user's rating for the pair of items being referred to [22].

3. Euclidean Distance similarity (EDS): Euclidean Distance Metric is a selection of the most effective and most commonly used measure of many distance measures [23].

$$EDS(u, v) = \sqrt{\sum_{k=1}^{n} (u_k - v_k)^2}$$  \hspace{1cm} (11)

Here \(n\) is a number of dimensions (attributes) and \(u_k\) and \(v_k\) are the attributes (components) of data user u and v.

The major drawback of this approach is, if no attribute values are common between two data vectors, then they may have a smaller distance than the other pair of data vectors comprising the identical attribute values [23].

4. Manhattan Distance similarity (MDS): The distance between Manhattan is the measure in which the distance between two points is the sum of absolute differences of Cartesian coordinates. In a simple way to say that, it is the sum of the difference between \(u_i\) and \(v_i\). It is also known as City-block similarity.
\[ MDS = \sum_{k=1}^{n} |u_{i,k} - v_{i,k}| \]  \hspace{1cm} (12)

where \( k \) is the ratings given by the user \( u_{i,k} \) and \( v_{j,k} \).

This approach may not work well for image data and Document Classification.

**[4] RESULT**

In this paper, similarity amongst users has been calculated over a proposed rating matrix as shown in Table 1.

<table>
<thead>
<tr>
<th>Users/Items</th>
<th>It1</th>
<th>It2</th>
<th>It3</th>
<th>It4</th>
<th>It5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Us1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Us2</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Us3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Us4</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Us5</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

From Table 1, it can be observed that the user preferences for items are represented as a matrix of user-item \( P_{r,s} \) where \( R \) is numbers of users and \( S \) is a number of items. A sample matrix displaying user ratings of five users for five items is given in Table 1 where set of users, \( U = \{ U_{s1}, U_{s2}, U_{s3}, U_{s4}, U_{s5} \} \) and set of items, \( I = \{ I_{t1}, I_{t2}, I_{t3}, I_{t4}, I_{t5} \} \).

We calculated and implemented the similarity measurement and compared the results using the similarity measuring methods mentioned above in Section 3. This section provides the result of experiments and analysis that we did.

**4.1 Experiment with sample rating**

Part of the above problems can be clearly observed through simple experiments using a sample dataset as seen in Table 1. The overall measurement results are summarized in Table 2 and 3.

In this sample rating, we considered 5 users and 5 items and we assumed each item rating from 1 to 5 scale. We tried to take different values for rating value of items to get the best result, also the absolute value. Here, we implemented both similarity measurement methods (correlation and distance) as discussed in Section 2 above and furthermore, the results of our analysis are illustrated in Table 2 and 3.

<table>
<thead>
<tr>
<th>Users/Items</th>
<th>Us2</th>
<th>Us3</th>
<th>Us4</th>
<th>Us5</th>
<th>Us2</th>
<th>Us3</th>
<th>Us4</th>
<th>Us5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Us1</td>
<td>-0.5898</td>
<td>0.5898</td>
<td>0.4566</td>
<td>0.0218</td>
<td>Us1</td>
<td>-0.526</td>
<td>0.6326</td>
<td>0.5</td>
</tr>
<tr>
<td>Us2</td>
<td>-0.2188</td>
<td>-0.7741</td>
<td>-0.7741</td>
<td>Us2</td>
<td>-0.1387</td>
<td>-0.7017</td>
<td>-0.7017</td>
<td></td>
</tr>
<tr>
<td>Us3</td>
<td>0.2212</td>
<td>0.0369</td>
<td>Us3</td>
<td>0.0312</td>
<td>0.1581</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Us4</td>
<td>0.4566</td>
<td>Us4</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) Person Correlation Similarity  \hspace{1cm} (b) Constrained PCS
4.2 Experimental Results

The results of our experiment are mentioned in Table 4.

Table 4. Result grid of users’ preferences over items

|----|-------|-------|-------|-------|-------|-------|-------|-------|-------|

(c) Weighted PCS

(d) Sigmoid Function-based PCS

(e) Spearman Rank Correlation Similarity

(f) Jaccard Measurement Similarity

(g) Mean squared difference

(h) Jaccard and MSD

(i) Cosine Similarity

(j) Adjusted CS

(k) Euclidean Similarity

(l) Manhattan Similarity
The table above demonstrates the findings of the similarity measurements among various users where we gathered the outcome of similarity or disparity among data analysed. In addition, each column in Table 4, represents the comparison among users for each similarity measurement. As it can be seen that the user 1 outperforms than other users, the first 4 columns are further discussed and explained as follows:

1. **U1:U2**: In this case, the analysis was performed on the results gathered by user U1 out U2. Except for CS and JS, various other similarity measures fail to perform computation. The output computed by CS was 0.634, while it was 0.6 by JS. We can infer that similarity measures deduced deals with co-rated items that result in a negative outcome.

2. **U1:U3**: MSD rating values of U1 and U3 are (3,2,1,2,5) and (2,2,3,2,4) respectively. The rating values of U1 is nearer to the rating value of U3 i.e. small difference between them and the best value similarity measurement outcome was CPS, BCP, and CS.

3. **U1:U4**: As depicted in the table, all the rating values for U1 versus U4 are very small except for MSD where the outcome is 1.

4. **U1:U5**: In this case, we noted the main drawback of JS, MSD, and J&MSD. All the rating values of U1 are also in U5 but at a different position as the outcome comes out to be 1, which means Jaccard only takes into account the proportion of the rating value. Also, in this instance, the best measurement is CS.

Our experimentation discovered that no users have the same rating value as U1 have. It confirms that the distance correlation similarity (CS) is better than all other correlation measurement models.

We can conclude that we divided the measurement similarity to two types of figures for different ratings (correlation similarity model, Fig. 2(a)) and (distance similarity model, Fig. 2(b)).
From Fig. 2(a), we can comprehend that the PCS and WPCS obtain the negative value while Fig. 2(b) compares all famous distance similarity models. The Euclidean has the worst recall and there are variants between all of this model.

The recall of CS is better than all correlation similarity measures. Also, the recall of PC is not worthy enough but it is better than all other correlation measurements in the dataset. Although, CS has been considered as the best measure of similarity compared to all similarity measurements in previous studies. Thus, our study also confirms this fact that CS is the best model when it comes to choosing among the user based collaborative filtering approach or item-based.

[5] CONCLUSION

In the present study, we analyzed the disadvantages of the existing similarity measure. Experiments were conducted on web services data set to demonstrate the performance of the proposed similarity measure. Experimental results show the effectiveness of the new similarity measure suggests that it can overcome the drawbacks of the traditional similarity measures. The study compares all correlation similarity with the best distance similarity, the CS models performed better as compared to other similarity models. In addition, the similarity measurement is easy to implement, however not suitable for the known problems of collaborative filtering such as (cold star problem and sparse datasets, etc.). Further research is required to improve the quality of the recommendation by discovering the new measurement similarity to enhance the efficiency of the recommendation system.

REFERENCES


