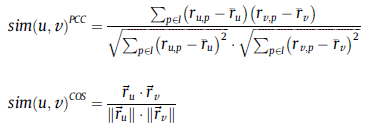
4.1 Disadvantages of existing similarity measures

Pearson correlation and cosine coefficient is as follows



where I represents the set of common rating items by user u and v.

 and  is the average rating value of user u and v respectively.

r and r denotes the rating of item p by user u and v respectively.

and  is the vector of the user u and v rated respectively.

The magnitude of vector is represented as ||.||.

For considering the impact of positive and negative ratings, the constrained Pearson correlation coefficient (CPCC) has been presented. The CPCC is defined as follows:



where r is the median value in the rating scale.

The traditional cosine similarity does not account for the preference of the user’s rating. For considering the preference of the user’s rating,the adjusted cosine measure (ACOS) has been introduced.The ACOS is defined as follows:



where P is the set of all items. If user u has not rated the item p  P ,the rating r is zero.

Jaccard [17] and mean squared difference (MSD) [6] are another two widely used measures. The formulas areas follows.Jaccard only considers the number of common ratings between two users. The basic idea is that users are more similar if they have more common ratings. The drawback is that it does not consider the absolute ratings. For example, user1 rates 5 and 4 on item1 and item2, user2 rates 1 and 2, user3 rates 4 and 5.Obviously, user1 and user 3 are more similar. MSD only considers the absolute ratings, but not consider the number of common ratings.The drawback is that it ignores the credibility of the similarity. A like the previous example, assume that user1, user2 and user3 have rated 5, 8 and 100 items respectively. Obviously, the similarity between user1 and user2 is more credible than the similarity between user1 and user3. Jaccard and MSD can be combined to form a new metric. Jaccard-



MSD-



JMSD-



where r and r represents the set of items by user u and v rated respectively.

Table 1An example of the user-item rating matrix. The missing ratings are represented by the

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| User\item | I | I | I | I |
| U | 4 | 3 | 5 | 4 |
| U | 2 | 1 | - | - |
| U | 4 | 3 | 3 | 4 |
| U | 3 | - | - | - |
| U | 1 | - | - | - |

symbol –.

Table 1.gives an example of a user-item rating matrix. We assume that there are four items and five users in the systems. The missing ratings of the rating matrix are represented by the symbol -.Then we calculate the similarities of users in the table according to those similarity measures described the above. Fig. 1 shows the results of the user similarities in Table 1. Since user similarity matrix is symmetric, we only show partial values.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | U | | U | | U | | U | |  | |  | |  | | U | | U | | U | | U | |  | |  |  | U | U | U | U |
| U | | 0.707 | | 0 | | 0 | | 0 | |  | |  | | U | | 0.606 | | 1 | | 0.492 | | 0.492 | |  | |  | U | -0.448 | 0.579 | 0 | -1 |
| U | | 1 | | 1 | | 0 | | 0 | |  | |  | | U | | 1 | | 0.695 | | 0.896 | | 0.896 | |  | |  | U | 1 | -0.448 | 0 | -1 |
| U | |  | | 1 | | 0 | | 0 | |  | |  | | U | |  | | 1 | | 0.565 | | 0.565 | |  | |  | U |  | 1 | 0 | 0 |
| U | |  | |  | | 1 | | 0 | |  | |  | | U | |  | |  | | 1 | | 1 | |  | |  | U |  |  | 1 | -1 |
| a)Pearson | | | | | | | | | |  |  | | b)cosine | | | | | | | | | |  | |  | | c)CPCC | | | | |
|  |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  | |  |  |  |  |  |
|  | U | | U | | U | | U | |  | |  | |  | | U | | U | | U | | U | |  | |  | |  | U | U | U | U |
| U | -0.287 | | 0 | | 0 | | 0 | |  | |  | | U | | 0.5 | | 1 | | 0.25 | | 0.25 | |  | |  | | U | -3 | 0 | 0 | -8 |
| U | 1 | | 0 | | 0.907 | | 0.906 | |  | |  | | U | | 1 | | 0.5 | | 0.5 | | 0.5 | |  | |  | | U | 1 | -3 | 0 | 0 |
| U |  | | 1 | | 0.867 | | 0.867 | |  | |  | | U | |  | | 1 | | 0.25 | | 0.25 | |  | |  | | U |  | 1 | 0 | 0 |
| U |  | |  | | 1 | | 1 | |  | |  | | U | |  | |  | | 1 | | 1 | |  | |  | | U |  |  | 1 | -3 |
| d)ACOS | | | | | | | | |  | |  | | e)Jaccard | | | | | | | | | |  | |  | | f)MSD | | | | |

Fig 1. user similarity values for table1 using different coefficients

The new similarity formula can be stated as follows:

sim(u,u)=

where u,u-two different users

N-total number of common products between u and u

-ratings given by user u and uto common product n

-average for users u and uof those products which are common between   
 them

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | U | U | U | U |
| U | 0.833 | 1 | 0.5 | 1 |
| U | 1 | 0.833 | 0.5 | 0.5 |
| U |  | 1 | 0.5 | 1 |
| U |  |  | 1 | 1 |

Using above formula we get user similarities as follows