Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/110137

Modified Heuristic Similarity Measure for Personalization using Collaborative Filtering Technique

SARANYA K. G. and G. SUDHA SADASIVAM*

Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, Tamil Nadu, India

Received: 10 Jul. 2016, Revised: 8 Nov. 2016, Accepted: 11 Nov. 2016
Published online: 1 Jan. 2017

Abstract: Collaborative filtering is one of the most widely used techniques for personalized recommendation services to users, since it can assist users to specify their interest on available items. The key feature of this technique is to find similar users by applying similarity measures on user-item rating matrix. Personalized system can thus provide recommendations for users based on the interest of the active user as well as a likeminded users. The success of the recommendation process depends upon the similarity metric used to find the most similar users. Similarity measures like cosine, Pearson Correlation Coefficient, Jaccard Uniform Operator Distance etc are not much effective when user-item rating matrix is sparse. This paper presents a new similarity model to calculate the similarities between each user, when only few ratings are available in the user profile. The proposed model considers both global preference as well as the local context of the user behavior. Experiments are conducted on two different datasets and compared with many existing similarity measures. The results of the experiments show that the proposed similarity measure improves the performance of the personalized recommendation process.

Keywords: Neighborhood Similarity Measure; Personalization; Recommendation Systems; Top-K Similarity; Collaborative Filtering; Similar users

1 Introduction

With rapidly increasing amount of information in the World Wide Web, it becomes difficult to locate relevant information from a large volume of data. Personalized recommendation services are used to help users to find the information of interest to them. Personalized recommendation services are of great importance in e-commerce web sites such as Amazon, Flip kart, Snap deal, digital library and online news portals.

The collaborative filtering techniques are the most widely used techniques in the field of personalized recommendation services to recommend an item to the users [17]. It gives recommendation based on similar users with the active user or the similar items with the items which is rated by the active user. Collaborative filtering recommendations depend upon the preferences of a set of users. The preferences of users can be tracked explicitly or implicitly. The explicit feedback which can be described as rating value given by the particular user on particular item. The implicit feedback which can be described by a user behavior, like clicking on a particular item. Memory-based CF [16,21] and Model based CF [18] are the two different types of collaborative filtering techniques. Model based methods constructs a model that reflects the behavior of the users and then it predicts the rating for the item or it recommends the item to the user. Memory based method calculates the similarity between the active user and all other users in the database and then it selects the most similar users as the neighbors, it then makes the recommendation according to the active user as well as the neighbors preferences.

This paper focuses on the recommendation system performance based on memory-based CF algorithms. The data set of any recommendation systems is very large since it has U*I matrix. Plenty of items have been rated by only a few of the total number of users present in the database. so, even active users can see just a few of items present in the database. This problem is called as a sparsity problem. It has a negative impact on the collaborative filtering based recommendation process. If the user item rating matrix is very sparse, then the similarity value calculated between two users will not be reliable and hence has a negative impact on CF based recommendation process.

* Corresponding author e-mail: saranyaa87@gmail.com

Collaborative filtering needs a lot of computations (permutations & combinations) to calculate a similarity between two users. They grow non linearly with increasing number of users. For example, if the database have U number of users and I number of items then the time complexity will be O(U*I). This problem is called as a scalability problem. P. Resnick et al [11] has suggested a solution to this problem. One of the existing solutions to this problem is to run the time-consuming training step in an offline mode and then produce a prediction with in a shorter time period in an online mode.

The core success of CF algorithms is either to compute a similarity between active user and all other users in the database or compute a similarity among items. Many researchers in the field of recommendation systems proposed lot of similarity measures such as Cosine, Jaccard, Pearson Correlation Coefficient, PIP, Adjusted Cosine, Euclidian Distance etc.. But all these similarity measures suffer from lower coverage and lower accuracy due to data sparsity and scalability problem in user-item matrix.

This paper presents an improved similarity measure which combines proportion of common ratings, global preferences, and ratings on non co-rated items of each user ratings. Experiments were conducted to analyze the performance of the proposed approach and it is compared with the existing similarity measures. Experimental results show that the proposed approach outperforms when the user-item rating matrix is sparse.

The remainder of this paper is organized as follows: section 2 describes the advantages and shortcomings of traditional similarity measures related to user based CF. Section3 describes the need and working principle of the proposed similarity measure. Experimental results are discussed in section4 and the conclusion of this paper is discussed in section5.

2 Related works

This section details about the working principle of neighborhood based CF approach in detail and the different existing similarity measures.

2.1 Neighborhood based approach

The memory based approach also called as Neighborhood based approach is introduced in Group lens systems, and this approach has been used in wide variety of recommender systems. It will use user profile (which contains user interest on particular item) to generate a list of items to recommend for an active user. Generally the user interest on a particular item is specified as a categorical rating from 1 to 5. An entry 0 in the user item rating matrix indicates that the particular user u has not rated the particular item. The prediction task of the neighborhood based approach is to formulate a k nearest neighbors (like minded people) based on the ratings given by the users on different items. The prediction task of the neighborhood based approach is to predict the rating of a particular item based on the neighborhood information. Hence this approach formulates k nearest neighbors to predict a interest of a particular user on particular item. For this purpose, this method computes a similarity between the active user and all other users in the user-item dataset, and then it selects k closest users to form the nearest neighbors of the active user. Finally, based on the neighborhood cluster, this method will predict a rating interest of a particular user on a particular item, or recommend set of items to the user based on the neighbors interest. Neighborhood based method can be categorized as User based CF methods and Item based CF methods. The user based CF methods [3] predict the rating of a particular item or recommend set of items based on ratings of ith item made by the neighbors of the active user.

$$\overline{r}_{u,i} = \overline{r}_i + \frac{\sum_{n=1}^{N} S(I_i, I_n) * (r_{u,n} - \overline{r_n})}{\sum_{n=1}^{N} |S(I_i, I_n)|} \quad (1)$$

$\overline{r_i}$= average rating of item I_i .
$S(I_i, I_n)$= similarity between the target item I_i and the n^{th} similar item of I.
$r_{u,n}$= is the rating made by the active user u on the nth similar item of I. Jamali et al [M. Jamali] introduced the item based CF methods. This method computes similarity between target item and all other items to find k- most similar items. Finally unknown rating is predicted based on the ratings of k items made by the active user.

$$\overline{r}_{u,i} = \overline{r_u} + \frac{\sum_{n=1}^{N} S(U_u, U_n) * (r_{ni} - \overline{r_{nu}})}{\sum_{n=1}^{N} |S(U_u, U_n)|} \quad (2)$$

Where,

$\overline{r_u}$ = average rating of the user U_u .
$S(U_u, U_n)$ = similarity value between and its n^{th} neighbor
$\overline{r}_{nu}$ = average rating of n^{th} neighbor with respect to active user U.
r_{ni} = rating value made by n^{th} neighbor on i^{th} item.

Hence the similarity measurement model plays a vital role in any recommendation systems. Similarity measures in Memory based Collaborative Filtering: Traditional measures such as cosine similarity, Euclidean distance, Pearson correlation coefficient are frequently used similarity measures in personalized recommendation systems. A list of existing similarity measures in user based CF algorithm is given in table 1. Salton et al [17] proposed a cosine similarity measure in information retrieval domain.

Table 1. Advantages and disadvantages of existing similarity Measures.

S.No	Similarity Measure	Formula	Major Drawbacks
1.	Cosine Similarity (COS) [12,20]	$sim(u,v) = \frac{\sum_{i\epsilon I'}(r_{u,i})(r_{v,i})}{\sqrt{\sum_{i\epsilon I'} r_{u,i}^2 \sum_{i\epsilon I'} r_{v,i}^2}}$ $r_{u,i} = rating\ made\ by\ user\ u\ on\ item\ i.$ $r_{v,i} = rating\ made\ by\ user\ v\ on\ item\ i.$ $I' = set\ of\ co\ rated\ items\ by\ both\ users\ u\ and\ v.$	1) It provides high similarity regardless of the significant difference in ratings made by two users. 2) This method suffers from the problem of few co-rated items by both users.
2.	Adjusted Cosine Similarity (ACOS) [5]	$sim(u,v) = \frac{\sum_{i\epsilon I'}(r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{(r_{u,i} - \bar{r}_u)^2}\sqrt{(r_{v,i} - \bar{r}_v)^2}}$ $r_{u,i} = rating\ made\ by\ user\ u\ on\ item\ i.$ $\bar{r}_u = average\ rating\ of\ the\ user\ u.$ $r_{v,i} = rating\ made\ by\ user\ v\ on\ item\ i.$ $\bar{r}_v = average\ rating\ of\ the\ user\ v.$ $I' = is\ the\ set\ of\ corated\ items\ by\ both\ users$	1) It provides low similarity regardless of the similar ratings made by two users. 2) If set of co rated items by both users u and v is very small then the similarity value produced by this model will not be reliable.
3.	Pearson Correlation Coefficient (PCC) [11]	$sim(u,v) = \frac{\sum_{i\epsilon I'}(r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i\epsilon I'}(r_{u,i} - \bar{r}_u)^2}\sqrt{\sum_{i\epsilon I'}(r_{v,i} - \bar{r}_v)^2}}$ $r_{u,i} = rating\ of\ user\ u\ on\ item\ i.$ $r_{v,i} = rating\ of\ user\ v\ on\ item\ i.$ $I' = set\ of\ corated\ items\ by\ both\ users$	1) it provides low similarity value regardless of the similar ratings made by two users on items. 2) If the co- rated items present in the user-item rating matrix is very few, then it will not provide a reliable similarity value.
4.	Constrained Pearson Correlation Coefficient (CPCC) [4]	$sim(u,v) = \frac{\sum_{i\epsilon I'}(r_{u,i} - r_{med})(r_{v,i} - r_{med})}{\sqrt{\sum_{i\epsilon I'}(r_{u,i} - r_{med})^2}\sqrt{\sum_{i\epsilon I'}(r_{v,i} - r_{med})^2}}$ $I' = set\ of\ corated\ items\ by\ both\ users$ $r_{med} = median\ value\ in\ rating\ scale$ $r_{u,i} = rating\ of\ user\ u\ on\ item\ i$ $r_{v,i} = rating\ of\ user\ v\ on\ item\ i.$	1) It suffers from few co - rated items problem.
5.	Sigmoid Pearson Correlation Coefficient (SPCC) [2]	$sim(u,v) = sim(u,v)^{PCC} . \frac{1}{1 + exp^{(-\frac{\lvert I\rvert}{2})}}$	1) It provides high similarity value regardless of the difference between the two user ratings.
6.	Mean Square Difference (MSD) [7]	$sim(u,v) = 1 - \frac{\sum_{i\epsilon I'}(r_{u,i} - r_{v,i})^2}{\lvert I'\rvert}$	1) It ignores the proportion of common ratings. This may lead to low accuracy.
7.	Jaccard Coefficient [6]	$sim(u,v) = \frac{\lvert I_u\rvert \cap \lvert I_v\rvert}{\lvert I_u\rvert \cup \lvert I_v\rvert}$	1) This approach does not consider absolute rating value of the two user's while calculating a similarity.
8.	Jaccard Mean Square Difference (JMSD) [8]	$sim(u,v)^{JMSD} = sim(u,v)^{Jaccard} . sim(u,v)^{MSD}$	1) It particularly addresses the drawbacks of jaccard and MSD. This measure utilizes all ratings provided by two users u and v. but this approach suffers from cold user problem.
9.	PIP (Proximity, Impact, Popularity)[9]	$sim(u,v)^{PIP} = \sum_{i\epsilon I'} PIP(r_{u,i}, r_{v,i})$ $PIP(r_{u,i}, r_{v,i}) = PIP\ value\ for\ the\ two\ ratings\ r_{u,i}\ and\ r_{v,i}\ on\ item\ i \in I'\ by\ user\ u\ and\ v\ respectively.$ $PIP(r_{u,i}, r_{v,i}) = Proximity(r_{u,i}, r_{v,i}).Impact(r_{u,i}, r_{v,i}).Popularity(r_{u,i}, r_{v,i})$	1) This approach does not consider the proportion of common ratings made by two users. Hence this will lead to low accuracy. (misleading of similarity exist in this approach. 2) It is not normalized. 3) Global preferences of the user behavior is not addressed.
10.	NHSM (New Heuristic Similarity Measure) [13]	$sim(u,v)^{NHSM} = sim(u,v)^{JPSS} . sim(u,v)^{URP}$ $sim(u,v)^{PSS} = \sum_{i\epsilon I'} PSS(r_{u,i}, r_{v,i})$ $sim(u,v)^{JPSS} = sim(u,v)^{PSS} . sim(u,v)^{Jaccard}$ $sim(u,v)^{URP} = 1 - \frac{1}{1 + \exp(-\lvert\mu_u - \mu_v\rvert . \lvert\sigma_u - \sigma_v\rvert}$	1) Ratings on non co-rated items are neglected in this approach. 2) Similarity computation is very complex.
11.	Bhattachary a Coefficient	$BC(i,j) = BC(\bar{P}_i, \bar{P}_j) = \sum_{h=1}^{m} \sqrt{(\bar{P}_{i_h})(\bar{P}_{j_h})}$	1) This approach cannot be used to find a similarity between pair of users if they rate on few or no similar items.
12.	Jaccard Uniform Operator Distance (JacUOD)[15]	$sim(u,v) = \begin{cases} \frac{\lvert S_{u,v}\rvert}{\lvert S_u \cup S_v\rvert} * \frac{\sqrt{m(Vmax - Vmin)^2}}{\sqrt{\sum_{s\epsilon S_{uv}}(r_{u,s} - r_{v,s})^2}} & if\ \ni s \in S_{u,v}, r_{u,s} \neq r_{v,s} \\ \frac{\lvert S_{u,v}\rvert}{\lvert S_u \cup S_v\rvert} * \frac{\sqrt{m(Vmax - Vmin)^2}}{0.9 + \sqrt{\sum_{s\epsilon S_{uv}}(r_{u,s} - r_{v,s})^2}} & if \forall s \in S_{u,v}, r_{u,s} = r_{v,s} \end{cases}$	1) This approach suffers from few or no co rated items.
13.	A new similarity measure using Bhattachary a coefficient for CF. (BCF) [14]	$sim(u,v) = Jacc(u,v) + \sum_{i\epsilon I_u}\sum_{j\epsilon I_v} BC(i,j) loc(r_{u,i}, r_{v,j})$ $loc_{cor}(r_{u,i}, r_{v,j}) = \frac{(r_{u,i} - \bar{r}_u)(r_{v,j} - \bar{r}_v)}{\sigma_u \sigma_v}$ $loc_{med}(r_{u,i}, r_{v,i}) = \frac{(r_{u,i} - r_{med})(r_{v,i} - r_{med})}{\sqrt{\sum_{k\epsilon I_u}(r_{u,k} - r_{med})^2} . \sqrt{\sum_{k\epsilon I_v}(r_{v,k} - r_{med})^2}}$	1) This approach is not scalable and similarity computation is very complex.

3 Modified heuristic similarity measurement model

The main motivation of the proposed similarity model is to combine the local context as well as the global preferences of the user behavior in order to improve the prediction of nearest neighbors and performance of the recommender systems.

The similarity measurement model plays a vital role in neighborhood based CF approach during the formulation of nearest neighbors of an active user. From the literature survey, it is interpreted that the traditional similarity measures are not suitable for sparse rating dataset. The proposed similarity measure uses Jaccard similarity which is used to compute a proportion of common ratings made by two users, Modified Bhattacharya coefficient measure, which is introduced to compute divergence between the ratings made by two users and PSS which is introduced to utilize the absolute ratings of the two users during the similarity computation.

3.1 Working principle of the Proposed similarity measure

The modified heuristic similarity measure combines PSS, Jaccard and Modified Bhattacharya coefficient to calculate a similarity between two users. The Proximity Significance Singularity (PSS) similarity between two users is calculated as follows.

$$sim(u,v)^{PSS} = \sum_{i \in I'}^{PSS(r_{u,i},\ r_{v,i})} \tag{3}$$

$$\begin{aligned} PSS(r_{u,i}\ ,\ r_{v,i}) = \ & Proximity\ (r_{u,i}\ ,r_{v,i}) * \\ & Significance(r_{u,i}\ ,\ r_{v,i}) * Singularity(r_{u,i},\ r_{v,i}) \end{aligned} \tag{4}$$

Where, Proximity is defined as the distance between rating made on a particular item by two different users.

$$Proximity(r_{u,i}\ ,r_{v,i}) = 1 - \frac{1}{1+exp^{(-|r_{u,i}-r_{v,i}|)}} \tag{5}$$

Where, $r_{u,I}$= rating value made by user u on item i.
$r_{v,I}$ = rating value made by user v on item i.

Significance is defined as the distance between the median value of the rating scale and the rating value made on a particular item by two different users.

$$Significance(r_{u,i}\ ,r_{v,i}) = 1 - \frac{1}{1+exp^{(-|r_{u,i}-r_{med}|\ .|r_{v,i}-r_{med}|)}} \tag{6}$$

Singularity is describes how two ratings made on a particular item by two different users with respect to the mean rating of that item

$$Singularity(r_{u,i}\ ,r_{v,i}) = 1 - \frac{1}{1+exp^{(-|\frac{r_{u,i}+r_{v,i}}{2}-\mu_i|)}} \tag{7}$$

Where,μ_i= average rating of item i.

The proportion of common ratings made by two users should also be considered to increase the accuracy of the similarity.

$$sim(u,v) = \frac{|I_u| \cap |I_v|}{|I_u| \cup |I_v|} \tag{8}$$

Bhattacharya coefficient:

Bhattacharya Coefficient similarity measure provides similarity between two probability distributions. It provides a measure of the amount of overlap between two statistical samples or population. If p and q are a discrete probability distributions over the same domain X, then the Bhattacharya distance between p and q is defined as,

$$BC(p,q) = \sum_{x \in X} \sqrt{p(x)\,q(x)} \tag{9}$$

If p and q are a continuous probability distributions over the same domain x means, then the Bhattacharya distance between p and q is defined as,

$$BC(p,q) = \int \sqrt{p(x)\,q(x)}\ dx \tag{10}$$

The similarity between two users u and v are calculated using rating made by those two users on available items and it is computed as follows.

$$BC(U,V) = \sum_{h=1}^{m} \sqrt{(\hat{X}_{Uh})\,(\hat{X}_{Vh})} \tag{11}$$

Where, $\hat{X}_{uh}$ and $\hat{X}_{vh}$ are the users rating value on different items under the domain X and it is calculated as follows,

$$\hat{X}_{Uh} = \frac{\#h}{\#u} \tag{12}$$

Where, h=number of items rated with rating value h. u= number of items rated by user u. This will be illustrated with the following example. Let us consider the rating scale lies between 1 to 4. i.e1,2,3,4. U1 and U2 are the two users made a rating on four different items.

Table 2. Example User-Item Rating Matrix

Users/Item	I1	I2	I3	I4
U1	2	-	4	2
U2	2	3	1	2

$$BC(U1,U2) = \sqrt{\left(\frac{0}{3}\right)\left(\frac{1}{4}\right) + \left(\frac{2}{3}\right)\left(\frac{2}{4}\right) + \left(\frac{0}{3}\right)\left(\frac{1}{4}\right) + \left(\frac{1}{3}\right)\left(\frac{0}{4}\right)} \tag{13}$$

BC(U1,U2)= 0.333

The disadvantage of existing Bhattacharya Coefficient Measure is that, it does not give any importance to local similarity, namely when a pair of users made dissimilar rating value on similar items. It will return 0 even if there exist a number of co-rated items by two users with dissimilar rating value. It does not give any importance to the number of common items rated by two users. Hence the similarity value provided by BC will not be reliable for all kind of situations. To avoid this problem, the proposed similarity measure modifies the Bhattacharya Coefficient as follows,

$$sim(u,v)^{MBC} = \frac{1}{1+exp^{-|sim(u,v)^{BC}|}} \tag{14}$$

$$sim(u,v)^{BC} = \sum_{h=1}^{m} \sqrt{\left(\frac{h_{u,i}}{|I_u|}\right)\left(\frac{h_{v,i}}{|I_v|}\right)} \tag{15}$$

Where, $h_{u,i}$= number of items rated with rating value h by user u.
$h_{v,i}$= number of items rated with rating value h by user v.
I_u= total number of items rated by user u.
I_v = total number of items rated by user v. The formalization of the modified heuristic similarity measure is defined as follows:

$$sim(u,v) = w_1 * sim(u,v)^{Jaccard} + [w_2 * (sim(u,v)^{PSS} * Sim(u,v)^{MBC})] \tag{16}$$

Where w_1 and w_2 value is taken as 0.5. i.e equal weight has been assigned to both proportion of common ratings and the absolute rating of the user.

Discussions on the proposed approach:

1. The proposed similarity measure utilizes all the absolute ratings made by each user on available items.

2. The similarity between two users are calculated based on both local context and global preferences of the user rating. Hence misleading of similar user cluster can be avoided.

3. It assigns equal weight for number of co-rated and non co-rated items. Hence the proposed similarity measure works well even if there is no co rated items exist between two users.

4. In many existing similarity measures like cosine, PCC, Jaccard, it s not possible to compare each users, since it provides a same similarity value. But in the proposed approach, each user becomes comparable, since it provides different similarity values for each pair of users.

5. The proposed measure is a normalized similarity measure as it lies between 0 to 1.

4 Experimental Results

4.1 Dataset

News dataset and Jester datasets[19] are used in the experiments. The news articles were collected from google news website at various time interval (from 1st june2015 to 30th Aug 2015), and it forms the news database. The proposed system maintains user profile and news item database. User profile data base contains the information like, userID, Category, rating value, $news_{id}$,snippet,URL,content,time stamp, and Click Frequency. News item data base contains the informations like $news_{id}$, category, URL, Snippet, Content and published time. The data set consists of 5058 users under 5 categories from sports domain namely, cricket, football, hockey, tennis and athletics. Each user is interested in at least two categories. Jester dataset is used for online jokes recommendation system. Jester dataset includes data from 24,938 users who have rated 15 to 35 jokes. The rating matrix of dimension contains 24,938 X 101 ratings real values from -10.00 to 10.00. Hence sparseness of the dataset is more. The dataset also includes the number of jokes rated by each user. To demonstrate the performance of the proposed measure, 80% of users are used for training while 20% is used for testing.

4.2 Evaluation Metrics

Performance of the proposed similarity measure is evaluated using precision and recall. Precision is the proportion between the number of items that are actually liked by the testing users and the number of top-N items recommended. Recall is the ratio between the amount of items liked by the testing users and number of items liked by active users in the testing set. There is often a tradeoff between these two measures (Precision & Recall). For example, if the number of items increases in the top-n recommendation list then recall will increase while the precision decreases. Therefore F1-Measure which combines precision and recall is used to measure the accuracy of predicting number of nearest neighbors and performance of the recommendation system.

$$F\ Measure = 2 * \left(\frac{Precision * Recall}{Precision + Recall}\right) \tag{17}$$

4.3 Performance Comparison

In this section several experiments were conducted on the two different datasets and the proposed similarity measure is compared with many other traditional similarity measures. Number of nearest neighbors and number of recommendations are the two parameters which can impact the performance of recommendation

systems. The results are compared with different values of these two parameters.

A Performance of different similarity measures on Jester dataset:

i For K-Nearest Neighbors: Precision gets decreases as number of k-neighbors increases in all the similarity measures. But among all the traditional measures, MHSM provides better precision value. PIP provides worst precision value when k=80. JacUOD provides better precision when k¡=20.it is shown in Fig 1. Recall of ACOS increases at k¿50.Recall of MHSM is better when compared to all similarity measures and it is shown in the Fig 2. F-measure is shown in Fig 3.

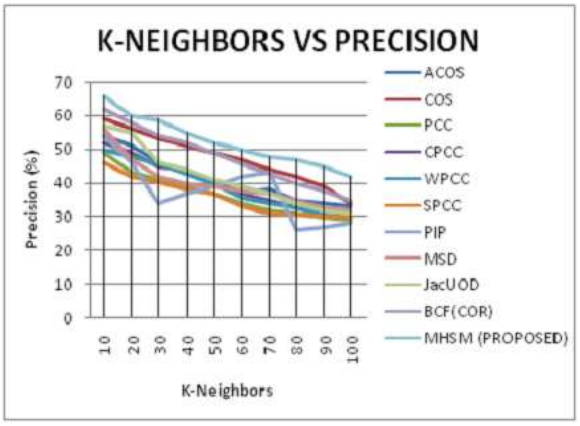

Fig. 1: Comparison of precision against K-Neighbors on jester dataset.

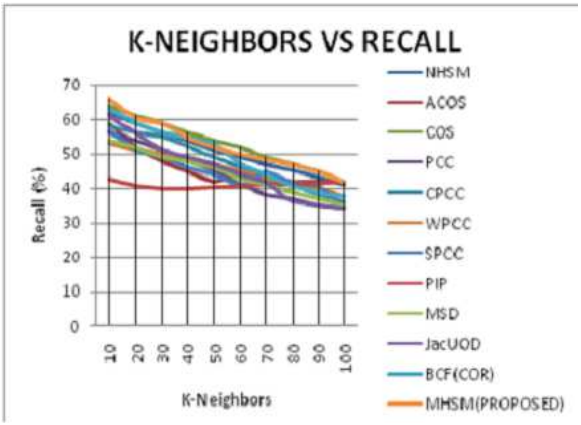

Fig. 2: Comparison of Recall against K-Neighbors on Jester dataset.

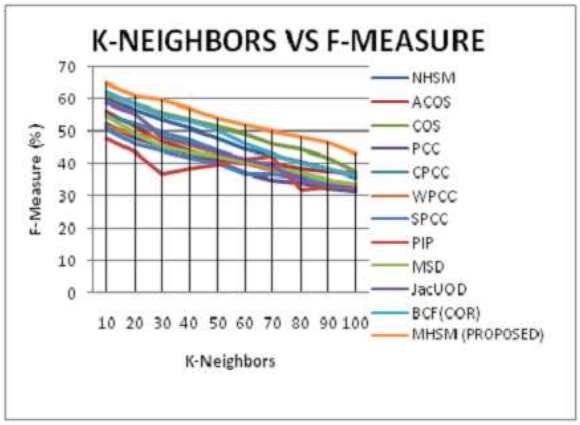

Fig. 3: Comparision of F-measure against K-Neighbors on jester data set.

ii For Number of Recommendations (k): Recalls of all the similarity measures increases with the increasing number of recommendations of jokes. MHSM gives better recall when compared to other similarities for all k values. Recall of PIP is worst than NHSM, JacUOD, ACOS, COS, MHSM, when k¿50. PCC gives low recall value when k is smaller. ACOS gives better recall when compared to cosine, but compared to MHSM, the recall value of ACOS is small. It is shown in figure 4.The precision of MHSM decreases when k value increases. But precision value is stable when k is small. The precision of WPCC is worst than all other similarity measures. The precision of MSD and ACOS similarity increases when k value increases. Precision recorded by COS is more stable when compared to MHSM. It is shown in figure 5. F-measure of the same is shown in figure 6.

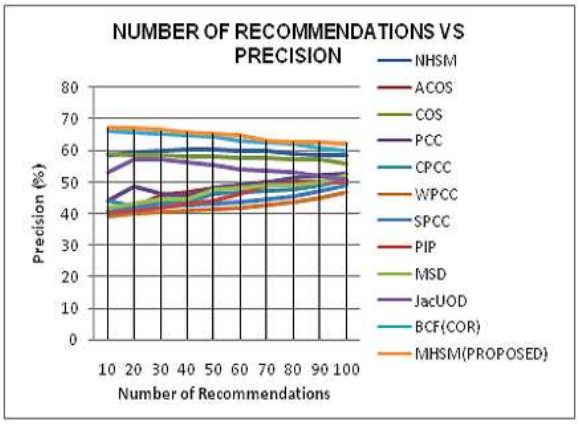

Fig. 4: The performance of different similarity measures on jester data set (Number of recommendations vs Precision)

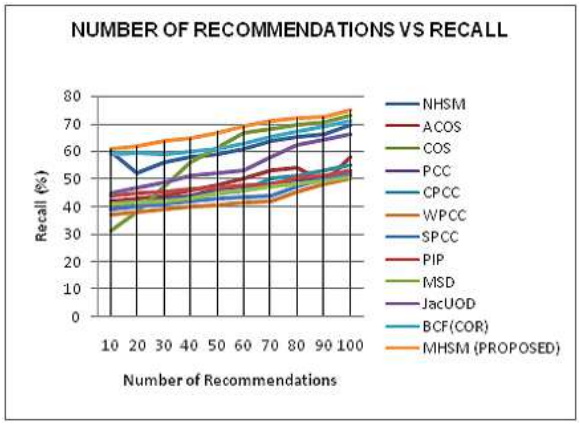

Fig. 5: Comparision of Recall against Number of Recommendations on jester data set .

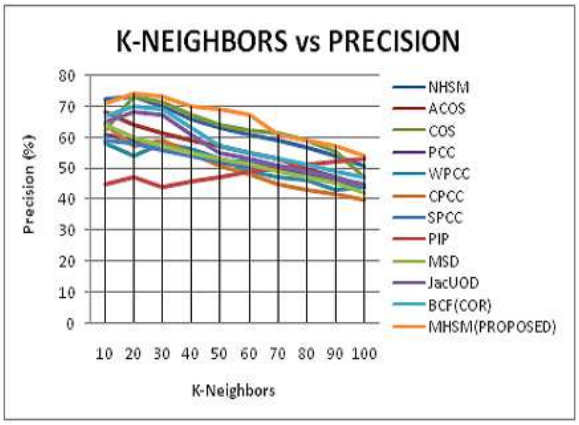

Fig. 7: Comparision of precision against K-Neighbors on News data set.

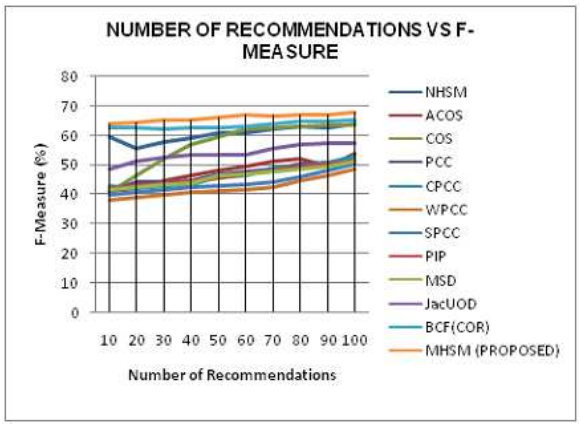

Fig. 6: Comparision of F-Measure against Number of recommendations on jester data set.

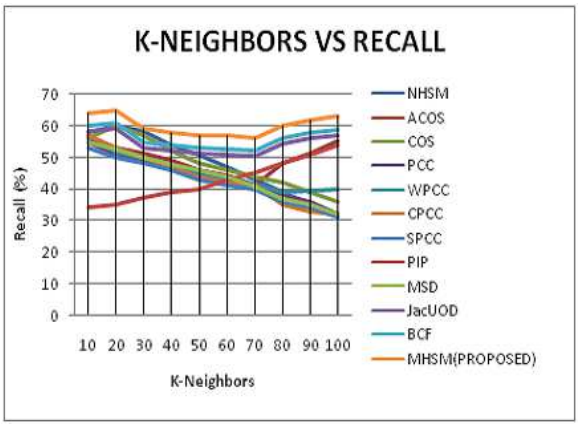

Fig. 8: Comparision of Recall against K-Neighbors on News data set.

B Performance of different similarity measures on news dataset

i For K-neighbors: Cosine similarity gives better precision value at k=70. PIP provides better precision for larger values of k. but for smaller values of k, its precision value is worst. NHSM provides better precision for smaller values of k. it is shown in Fig 7.Recall of NHSM decreases when k gets increases. MHSM provides better recall when compared to all other similarity measures. Recall of PIP increases when k gets increases, but it is smaller when compared to MHSM. Cosine similarity gives better recall for smaller values of k. it is shown in Fig 8. F-measure of the same is shown in Fig 9.

ii For Number of Recommendations(k): Cosine similarity gives better precision for smaller values of k when compared to MHSM. The precision of ACOS decreases when k value increases. MHSM provides better recall value when k is large. CPCC provides better precision when compared to PCC and MSD for all k values. It is shown in Fig 10. recall of MHSM is high

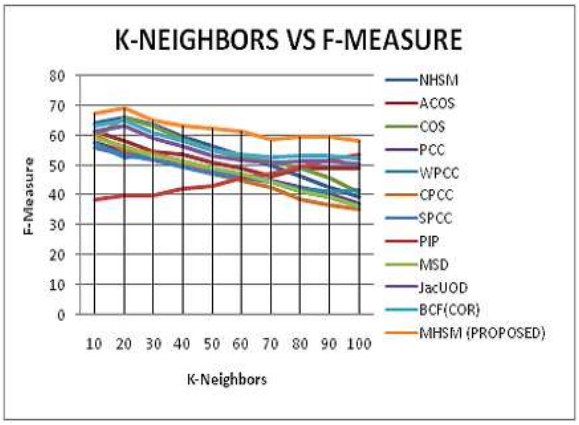

Fig. 9: Comparision of F-Measure against K-Neighbors on News data set.

when compared to all other similarity measures. It is shown in Fig 11. F-measure of the same is shown in Fig 12.

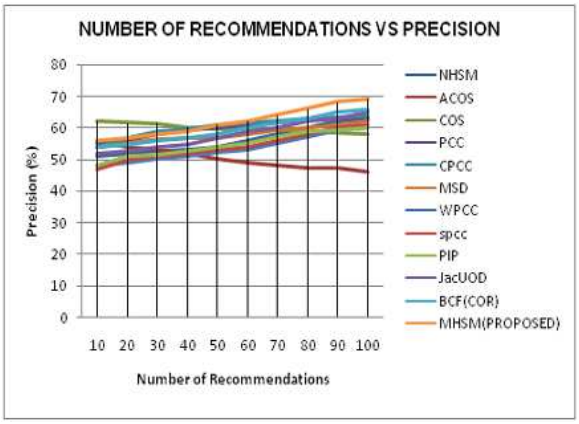

Fig. 10: Comparision of Precision against Number of Recommendations on News data set.

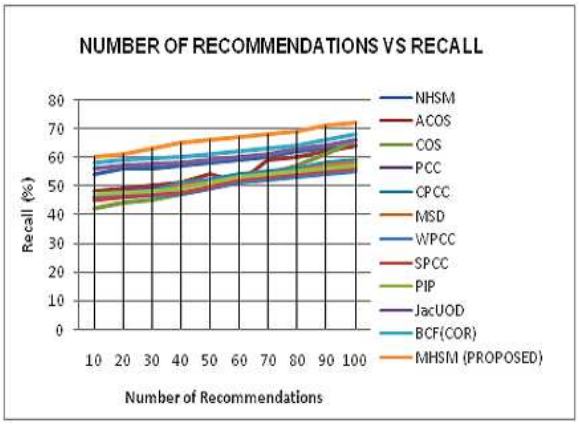

Fig. 11: Comparision of Recall against Number of Recommendations on News data set.

As some measures are sensitive to false neighbors, precision and recall obtained by some measures increases and others decreases as k value increases. This is because similarity between two users are high but actually their preferences are not same is called false neighbors. This situation often arises due to data sparsity in dataset. From figutre 7 and figure 12, the proposed similarity measure (MHSM) shows better performance than most other methods in the whole range number of recommendations.

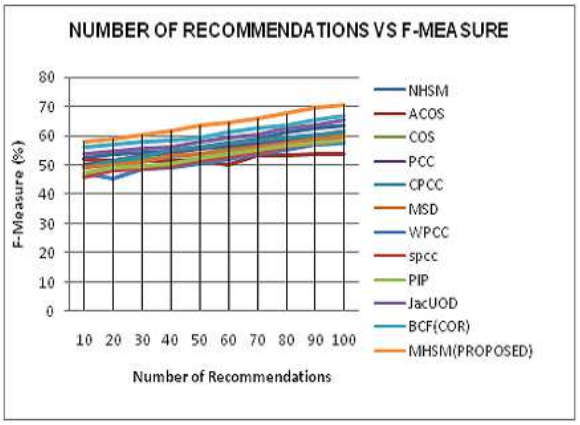

Fig. 12: Comparision of F-Measure against Number of recommendations on News data set.

5 Conclusion

This paper discusses the existing similarity measures in recommendation systems. It also discusses on the drawbacks of these measures. In order to overcome these shortages of existing similarity measure, a new similarity measure called weight based modified heuristic similarity measure is proposed. It is based on PSS, Bhattacharya Coefficient, and Jaccard similarity measures and hence it considers, the local context, global preferences and proportion of common ratings between two users while calculating similarity. Each factor in the proposed similarity measure belongs to 0 to 1 and hence it is normalized measure. Several experiments were conducted to demonstrate the effectiveness and efficiency of the proposed similarity measure. Experimental results show that the proposed similarity measure can obtain better performance when compared to other existing similarity measures. The proposed measure provides 71% of F-Measure during the recommendation process.

References

[1] F. Cacheda, V. Carneiro, D. Fernndez, V. Formoso, Comparison of collaborative filtering algorithms: limitations of current techniques and proposals for scalable, high-performance recommender system, ACM Trans. Web 5 (1) (2011) 133.

[2] M. Jamali, M. Ester, TrustWalker: a random walk model for combining trustbased and item-based recommendation, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, pp. 397406.

[3] J.L. Herlocker, J.A. Konstan, A. Borchers, J. Riedl, An algorithmic framework for performing collaborative filtering, in: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, pp. 230237.

[4] U. Shardanand, P. Maes, Social information filtering: algorithms for automating word of mouth, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1994, pp. 210217.
[5] H.J. Ahn, A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem, Inform. Sci. 178 (1) (2008) 3751.
[6] G. Koutrica, B. Bercovitz, H. Garcia, FlexRecs: expressing and combining flexible recommendations, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, 2009, pp. 745758.
[7] F. Cacheda, V. Carneiro, D. Fernndez, V. Formoso, Comparison of collaborative filtering algorithms: limitations of current techniques and proposals for scalable, high-performance recommender system, ACM Trans. Web 5 (1) (2011) 133.
[8] J. Bobadilla, F. Ortega, A. Hernando, J. Bernal, A collaborative filtering approach to mitigate the new user cold start problem, Knowledge-Based Syst. 26 (2011) 225238.
[9] H.J. Ahn, A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem, Inform. Sci. 178 (1) (2008) 3751.
[10] J. Bobadilla, F. Ortega, A. Hernando, A collaborative filtering similarity measure based on singularities, Inform. Process. Manage. 48 (2012) 204217.
[11] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, GroupLens: an open architecture for collaborative filtering of netnews, in: Proceeding of the ACM Conference on Computer Supported Cooperative Work, 1994, pp. 175186.
[12] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, IEEE Trans. Knowl. Data Eng. 17 (6) (2005) 734749.
[13] Haifeng Liu , Zheng Hu, Ahmad Mian, Hui Tian, Xuzhen Zhu, A new user similarity model to improve the accuracy of collaborative Filtering , Knowledge-Based Systems 56 (2014) 156166.
[14] Bidyut Kr. Patra , Raimo Launonen , Ville Ollikainen , Sukumar Nandi, A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data Knowledge-Based Systems 82 (2015) 163177.
[15] Hui-Feng Sun, Gang Yu, Guang Chen, JacUOD: A new similarity measurement for collaborative fltering. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 27(6): 1252:1260 Nov. 2012. DOI 10.1007/s11390-012-1301-5.
[16] Mooney R J, Roy L. Content-based book recommending using learning for text categorization. In Proc. ACM SIGIR 1999 Workshop Recommender Systems: Algorithms and Evaluation, Aug. 1999, pp.195-204.
[17] Salton, G., McGill, M, An Introduction to Modern Information Retrieval, McGraw-Hill, New York, NY (1983).
[18] Lang K, Newsweeder: Learning to filter netnews, In: Proceedings of the 12th International Conference on Machine Learning, Tahoe City, California (July 1995) 331339.
[19] http://www.ieor.berkeley.edu/ goldberg/jester-data/.
[20] Lei Li, Dingding Wang, Tao Li, Daniel Knox, Balaji Padmanabhan SCENE : A Scalable Two-Stage Personalized News Recommendation System SIGIR11, ACM 978-1-4503-0757-4/11/07 July 2428, 2011, Beijing, China.
[21] D. Anand, K.K. Bharadwaj, Utilizing various sparsity measures for enhancing accuracy of collaborative recommender systems based on local and global similarities, Expert Syst. Appl. 38 (5) (2011) 51015109.

The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

Saranya K. G. is working as an assistant professor of department of computer science and enginnering in PSG College of Technology, Coimbatore. Her research interests are in the areas of Personalized Information retrieval and semantic web technology. she has published research articles in reputed international journals.

G. Sudha Sadasivam is a professor of computer science and engineering in PSG College of Technology. her area of interest includes Distributed Computing, Grid and Cloud Computing, Software Engineering, Information Retrieval. she has published research articles in reputed international journals.