ISSN: 0067-2904

A Modified Similarity Measure for Improving Accuracy of User-Based Collaborative Filtering

Nadia F. AL-Bakri[1*], Soukaena H. Hashim[2]

[1]Department of Computer Science, AL Nahrain University, Baghdad, Iraq.
[2]Department of Computer Science, University of Technology, Baghdad, Iraq.

Abstract

Production sites suffer from idle in marketing of their products because of the lack in the efficient systems that analyze and track the evaluation of customers to products; therefore some products remain untargeted despite their good quality. This research aims to build a modest model intended to take two aspects into considerations. The first aspect is diagnosing dependable users on the site depending on the number of products evaluated and the user's positive impact on rating. The second aspect is diagnosing products with low weights (unknown) to be generated and recommended to users depending on logarithm equation and the number of co-rated users. Collaborative filtering is one of the most knowledge discovery techniques used positively in recommendation system. Similarity measures are the core operations in collaborative filtering; however, there is a certain deviance through using traditional similarity measures, which decreases the recommendation accuracy. Thus, the proposed model consists of a combination of measures: constraint Pearson correlation, jaccard distance measure and inverse user frequency (IUF).

The experimental results implemented on movielens data set using MATLAB show a comparison between the results of the proposed model and some of the traditional similarity measures. The outcome results of the comparison show that the proposed model can be used as a parameter in the prediction process to achieve accurate prediction results during recommendation process.

Keywords: Recommendation system, Collaborative filtering, Similarity measurement, inverse user frequency, Jaccard.

مقياس التشابه المعدل لتحسين دقة التصفية التعاونية القائمة على المستخدم

ناديه فاضل البكري*[1]، سكينه حسن هاشم[2]

[1]قسم علوم الحاسبات، جامعه النهرين، بغداد،العراق.
[2]قسم علوم الحاسبات، الجامعه التكنولوجيه، بغداد، العراق.

الخلاصه

تعاني مواقع الإنتاج من الكساد في تسويق منتجاتها بسبب الافتقار إلى النظم الفعالة التي تحلل وتتبع تقييم العملاء للمنتجات؛ وبالتالي تبقى بعض المنتجات غير مسوقه على الرغم من نوعيتها الجيدة. يهدف هذا البحث إلى بناء نموذج متواضع يهدف إلى أخذ جانبين في الاعتبار. الجانب الأول هو تشخيص المستخدمين يمكن الاعتماد عليهم من خلال المواقع الالكترونيه بناءا على عدد المنتجات التي تم تقييمها والأثر الإيجابي

*Email: Nadiaf_1966@yahoo.com

للمستخدم على التصنيف. أما الجانب الثاني فيتمثل في تشخيص المنتجات ذات الأوزان المنخفضة (غير المعروفة) التي سيتم إنشاؤها وتوصيتها للمستخدمين اعتمادا على تحويل ومعكوس تكرار المستخدم وعدد المستخدمين الذين يتم تصنيفهم بشكل مشترك. التصفية التعاونية هي واحدة من معظم تقنيات اكتشاف المعرفة المستخدمة بشكل إيجابي في نظام التوصية. مقاييس التشابه هي من العمليات الأساسية في التصفية التعاونية. ومع ذلك، هناك اختلاف معين من خلال استخدام مقاييس التشابه التقليدية، مما يقلل من دقة التوصية. وهكذا، فإن النموذج المقترح يتكون من مجموعة من المقاييس: مقياس ارتباط بيرسون ، قياس المسافة جاكارد وتكرار المستخدم معكوس تظهر النتائج التجريبية المطبقة على مجموعة بيانات موبيلنس مقارنة بين نتائج النموذج المقترح ونتائج مقاييس التشابه المذكورة انفا". نتائج المقارنة بينت أن نتائج النموذج المقترح يمكن ان تستخدم كمعاملات في حساب التنبؤ لتحقيق نتائج تنبؤ دقيقة خلال عملية التوصية.

وتظهر النتائج التجريبية التي أجريت على مجموعة بيانات أن نتائج النموذج المقترح يمكن استخدامها بكفاءة. في عملية التنبؤ لتحقيق نتائج دقيقة.

1. Introduction

Recommender Systems (RSs) are software tools and methods that come up with recommendations for things that are probably of interest to a particular person. The recommendations relate to numerous decision-making methods, such as what products to buy, what song to listen, or what movie to watch. “Item” is the universal word used to designate what the system recommends to users [1]. The basic tasks of Recommender systems are [2]:

The First task

Is providing a personalized recommendation to users by using user preferences to predict the rating of a particular item, A selection of interested items is done first by the user, which is done frequently when surfing an online websites.

The Second task

Is recommending a list of items to the target users Recommender systems have been established to be effective in e-commerce, such as recommending movies by movielens, recommending books/products by Amazon.com and DVD's by Netflix recommenders. Technology has intensely reduced the obstacles to reproducing and distributing information. One of the most encouraging technologies is collaborative filtering which exploit information from neighborhoods to predict which item the current user will most probably like or be interested in. These systems are well-known industrial usage today to promote additional items and increase sales [1].

In this paper, a modified similarity model for user-based collaborative filtering recommendation system is suggested after conducting the problems in general similarity measures. In section 2, the related works on this field is subjected. In section 3, the collaborative filtering definitions and traditional similarity measure methods are summed up. In section 4, the general concepts of the modified similarity model of user-based collaborative filtering recommendation is presented. In section 5, experiments on the suggested similarity model and comparison with other measures is conducted. Last section is the conclusion of this work.

2. Related Work

In what follows, some of the previous research literatures related to the techniques using similarity measures in user-based collaborative filtering are presented.

a- KG, S., & Sadasivam, G. S. [3], in this paper, A new modified heuristic similarity measure is suggested that combines Proximity-Significance-Singularity (PSS), Jaccard and Modified Bhattacharya coefficient to calculate a similarity between two users on sparse datasets. In PSS model, Proximity is the distance between two users on a particular item. The Significance is the distance between the median value of the rating scale and the rating values of two distinct users on a specific item. Singularity is the measure of how far each rating made by a distinct user from the mean rating of a particular item. Bhattacharya Coefficient similarity is the measure of overlapping between two probability distributions. The suggested heuristic model considers both global preference and the local context of the user behavior. The model was tested on two datasets. The results shown that the proposed similarity measure improves the performance of the personalized recommendation process and it is outperformed when the user-item rating matrix is sparse

b- Aygün S., Okyay, S., [4], in this paper, a new mathematical equation enhancing Pearson similarity measure called age parametrized Pearson similarity equation is proposed and used during the similarity calculation between users. The proposed similarity measure utilizes time information taken from users' ages used in the recommender systems. Time generation gap between users can make sense positively or negatively in terms of the amount of gap. 10-year interval is considered as a transition between generations, a 3-year long difference can be a boundary in the same wavelength. It was tested with a real valued Movielens dataset. Results have shown that improvements are obtained when the user ages are scientifically taken into consideration and the generation gap between users are processed to increase the efficiency of the recommender system

c- Huang, B. H., & Dai B. R., [5], in this paper, a similarity model called a Weighted Distance (WD) is proposed that perform item-based similarity. The model (WD) is combined with jaccard similarity coefficient to take the number of common ratings into account. Results conducting movielens dataset shown that the proposed similarity function significantly improves accuracy of recommender system and performs much better than traditional similarity measures in the cold-start problem

d- Liang, S., et al. [6], in this paper, a singularity-based model is presented to get accurate prediction results for collaborative filtering in recommender system. Singularity method describes the relation between two ratings made on a particular item by two different users with respect to the mean rating of that item. A Pearson correlation coefficient was improved by incorporating the number of common items rated by users. The jaccard measure was improved to consider the rating values given on items. Then a combination of the two methods formulated in two ways to produce three singularity aspects: the positive, negative and empty singularity for each item in the data base. The model was tested on two different datasets, the results shown a good performance in prediction accuracy in recommender system.

e- Liu, H., et al.[7],in this paper, a New Heuristic Similarity Model (NHSM) is Proposed to improve the recommendation performance when only few ratings are available. The model utilized a non-linear function concept called sigmoid function. The Proximity, Impact, Popularity (PIP)measure was improved as PSS (Proximity-Significance-Singularity) and combined with the improved jaccard formula to produce a new similarity measure called Jaccard Proximity-Significance-Singularity (JPSS). The mean and variance of the rating to model the user reference was adopted and combined with JPSS to produce the NHSM scheme. The model considered the global preference and local information of user ratings. The model was tested on three datasets and the results were shown better recommendation performance and better utilization of user ratings in cold start user conditions.

f- Candillier, L., et al. [8], in this paper, a modified similarity measure is proposed that combine jaccard with traditional similarity measures, in order to benefit from their integration and to overcome the shorting existed in these measures due to sparsity problem. Jaccard used as a weighted factor to the similarity measure. wpearson, wcosine and wmanhattan were adopted where they stands for weighted Pearson, weighted cosine and weighted manhattan respectively. The adopted weighting scheme was tested on two movie datasets: MovieLens and Netflix used in collaborative filtering. The results superiorly improved for both the item-based and user-based in two prospective the prediction accuracy and scalability problem.

g- Hyung, J., & Ahn, J., [9], in this paper, a scheme (PIP) is proposed which conducted three subjects: Proximity-Impact-Popularity of the user behavior for the collaborative filtering. Proximity factor reflects the agreement or disagreement of two ratings, giving punishment to the disagreement based on the difference between two ratings. The Impact factor considers the high rating given to an item by users reflecting the likeness. Popularity factor consider the distance between the user sand their average rating. Greater distance value means higher value to a similarity. The model was developed exploiting explicit description of user preference on items in order to overwhelm the flaw of traditional similarity measures like Pearson correlation coefficient and cosine measures. It was shown through experiments that the proposed model (PIP) considered the local information of the ratings and did not consider the global behavior of user ratings. Specific formula for each aspect was formulated and applied on three available datasets.The results shown that a great performance is achieved in new user cold-start problem.

h- Weng, L.T., et al.[10],in this paper, a similarity measure Statistical Attribute Distance (SAD) used in memory based collaborative filtering is proposed which includes a modification to Inverse User Frequency(IUF) transformation and under the assumption that the item ratings is more appropriate in considering the item importance among users. The SAD combines statistical information of the item features with item correlation method (such as Euclidean distance measure) to generate accurate prediction results especially in case for few rating from users.

3. Recommender System Aspects

The general recommendation techniques can be classified into [11]:

1. Content-based recommendation: Recommendation process finds items that have similar content and features to the items that the target user interested in the past. A various similarity measures can be used to calculate the similarity between items.

2. Collaborative filtering recommendations: Recommendation process is based on a consideration that community of users that share past behavior can also share future behavior. Rating matrix is a table used to store user ratings to be further used in Collaborative filtering works.

Two types of collaborative filtering algorithms [2, 12]:

(a) Memory-based collaborative filtering (neighbor-based): in this type, the items chosen by users who have similar behavior with the target user will be recommended. Similarity measures are used to find users that are similar to the target user (neighbors). Then, prediction methods are applied on the ratings of these neighbors. Memory-based algorithms can be categorized into user-based algorithm and item-based algorithm depending on the neighbors of similar users or items.

(b) Model-based collaborative filtering: in this type, models are created and trained offline to predict missing rating for the target user.

3. Hybrid approaches: a combination of the previous mentioned methods is combined in different manner to utilize the advantageous of each method [11].

The fundamental stage of collaborative filtering methods is similarity computation between users or items. The common similarity measures, such as cosine, Pearson correlation coefficient and jaccard are not sufficient to find the real similar users [12].

This paper focuses on using user-based method and applying a modified similarity measure in finding similar users to get accurate prediction.

3.1 User-Based method (neighborhood-based)

It is one of the standard methods of collaborative filtering. The rating matrix is held in memory and that is why it is said memory-based. User-based method follows the following steps:

(1) Similarity Calculation between the target user and other users in the rating matrix.

(2) A neighborhood Selection according to the similarity with the target user.

(3) Prediction Computation using the selected neighborhood [12].

3.2 Computation of Similarity between Users

In recommendation systems, the rating matrix is made up of of *n* users $U = \{u_1, u_2, \ldots, u_n\}$ and *m* items $\{I_1, I_2, ..., I_m\}$. The user-item rating matrix can be denoted as $R\ (n \times m)$, $R_{a,}i$ be the rating of user u_a on item I_i. $\overline{R_u}$ $\overline{R_V}$ are average rating for users *u*, *v* respectively . Let *SIM* (u, v) be the similarity between user *u* and user *v*. All items that are rated by *u* and *v* will be under similarity calculation using different similarity measures [2].

A-Similarity Weight Computation

The computation of the similarity weights can have a substantial influence on both its accuracy and performance [1].The similarity value is a non-negative numeric value that measures the degree of likeness or dislike. They are often between 0 (not similar) and 1(complete similar).

If SIM (u, v) is the similarity between u and v, then:

1. SIM (u, v) =1 only if u=v. ($0 \leq SIM \leq 1$).

2. SIM (u, v) = SIM (v, u) for all u and v. (Symmetry)

The similarity weight value has a double role in neighborhood-based computation, they are:

1) Trusted neighbors are selected during the similarity calculation.

2) Give more or less importance to the neighbors during the prediction computation [13].

B- Similarity Measurements

In this subsection, the most commonly used traditional similarity methods are explained.

1- Pearson Correlation Similarity

It is a measure of the linear relationship between the ratings of the two objects (users or items). Pearson's correlation coefficient between two users u and v is defined by the following similarity equation [7]:

$$Sim(u,v) = \frac{\sum_{i\in I}\left(R_{u,i} - \overline{R_u}\right)\left(R_{v,i} - \overline{R_V}\right)}{\sqrt{\sum_{i\in I}\left(R_{u,i} - \overline{R_u}\right)^2}\sqrt{\sum_{i\in I}\left(R_{v,i} - \overline{R_V}\right)^2}} \tag{1}$$

The Pearson correlation coefficient takes values from +1 (strong positive correlation) to −1 (strong negative correlation). The Pearson algorithm makes use of negative correlations as well as positive correlations to make predictions [12].

2- Constrained Pearson correlation

This measure is derived from the Pearson correlation coefficient which does not make use of negative "correlations" as the Pearson algorithm does. It uses median value instead of average rating.so it takes into consideration the influence of positive and negative ratings. The constrained Pearson correlation coefficient (CPCC) [14] is defined as follows [7]:

$$Sim(u,v) = \frac{\sum_{i\in I}\left(R_{u,i} - R_{Median}\right)\left(R_{v,i} - R_{Median}\right)}{\sqrt{\sum_{i\in I}\left(R_{u,i} - R_{Median}\right)^2}\sqrt{\sum_{i\in I}\left(R_{v,i} - R_{Median}\right)^2}} \tag{2}$$

R_{Median} is the median value in the rating scale. The median value for rating scale [1-5] is 3.

3. Cosine similarity measure

Is a measure of the angle between u and v vectors. If the angle between u and v is 0, then the cosine similarity is 1and they are similar. But if the angle between u and v is 90, then the cosine similarity's value is 0 and they are not similar.

The cosine similarity equation between user u and user v is expressed below [13] [15]:

$$sim(u,v) = cosine(\overline{R_u}, \overline{R_v}) = \frac{\overline{R_u} \cdot \overline{R_v}}{\|R_u\|\|R_v\|} = \frac{\sum_{i=1}^{N} R_{u,i} \times R_{v,i}}{\sqrt{\sum_{i=1}^{N}\left(R_{u,i}\right)^2}\sqrt{\sum_{i=1}^{N}\left(R_{v,i}\right)^2}} \tag{3}$$

4-Jaccard Measure

It is a measure of the closeness of two vectors of values. Jaccard distance measure is 1 minus the Jaccard similarity. The concept behind this measure is that users are similar according to the number of common ratings. The formula for jaccard similarity is defined as follows [13]:

$$Sim\ (u,v) = \frac{|I_U| \cap| I_V|}{|I_U|\cup|I_V|} \tag{4}$$

Where $|I_U|$ | $\mathrm{I_V}$| is the total number of items rated by u and v respectively.

5- Inversed User Frequency (IUF)

It is one of the most transformations used in memory-based collaborative filtering. Based on Breese's research [16], the IUF has significantly improved the recommendation accuracy for the correlation coefficient method about 11% averagely [10]. The idea behinds Inversed User Frequency (IUF) was from the information retrieval method Inverse Document Frequency (IDF) [17]. The IUF transformation usage in similarity measure equation decreases the weight on common items, because these items are less beneficial in recommendation process to target users.

IUF can be formularized into [16]:

$$f_i = log\frac{n}{n_i} \tag{5}$$

f_i The significance of i^th item during the similarity calculation.

n users' total number in the user-item matrix.

n_i the number of users rated item i.

The similarity between user u and another user v using the IUF with the Pearson correlation equation can be defined as [10]:

$$sim(u,v) = \frac{\sum_{i=1}^{N} f_i^2\ (R_{u,i} - \overline{R_u})(R_{v,i} - \overline{R_v})}{\sqrt{\sum_{i=1}^{N} f_i^2\ (R_{u,i} - \overline{R_u})^2}\sqrt{\sum_{i=1}^{N} f_i^2\ (R_{v,i} - \overline{R_v})^2}}$$

C- Similarity Methods Analysis and its Limitations

1- Pearson correlation measurement not consider the fact of finding similar users for common items have less influence in recommendation process than finding similar users on uncommon items[12].

2- When the number of common items is 1 between two users, the Pearson correlation result will be 0 and the cosine correlation result will be 1 regardless of differences in individual ratings. Also during computation, Pearson correlation coefficient does not consider the number of co-rated items between two users [9].

3- Pearson correlation and cosine correlation may be confusing, where similar users may seem to be different to each other by using these similarity measures [9].

4- The rated items and its amount in the rating matrix do not well reflect the correlations between users. The sparsity problem of the dataset has a strong influence on the correlation [10].

5- Cosine similarity does not account for the preference of the user's rating [18].

6- Jaccard coefficient does not consider the absolute ratings. Discarding the absolute value of rating will become difficult to distinguish different users [7].

7- Ignoring the proportion of common ratings will lead low accuracy [7].

4. The Proposed Method

In this section, a description of the user-based collaborative filtering recommender system based on a modified similarity model is conducted. First, the framework of the proposed model is introduced. Next, the algorithm of the modified similarity model is provided.

A. Framework

The user-movie rating matrix is constructed using the movielens dataset. After matrix formulation from the dataset, then, the proposed similarity model is applied between the target user and the rest of users of the matrix. The proposed model is depicted in Figure-1.

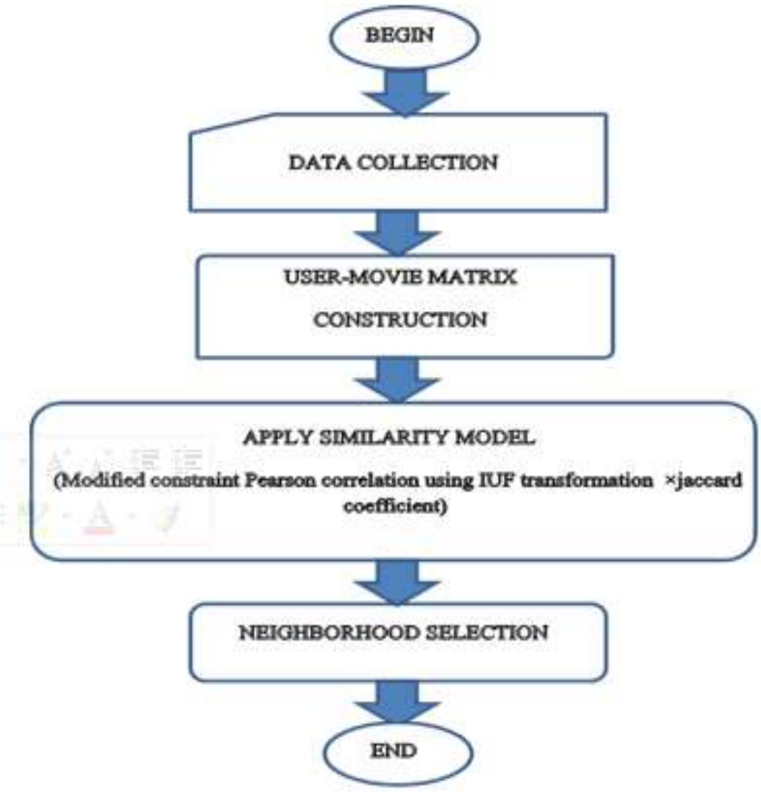
BEGIN
DATA COLLECTION
USER-MOVIE MATRIX CONSTRUCTION
APPLY SIMILARITY MODEL
(Modified constraint Pearson correlation using IUF transformation ×jaccard coefficient)
NEIGHBORHOOD SELECTION
END

Figure 1-the block diagram of the user-based collaborative filtering with Modified similarity measure

B- The similarity model

An approach based on an offline precomputation of the dataset by constructing the user similarity matrix that describes the pairwise similarity of all users.

Three aspects are considered in the proposed similarity model:

First, the model considers the impact of positive and negative ratings by using the constraint Pearson correlation measure.

Second, reducing the weight on commonly watched movies when calculating the user preferences to movies using the $f_j$factor mentioned in equation (5) as a weight in the constraint Pearson correlation equation.

Third, considering the number of co-rated movies by using the jaccard measure.

As a final stage, the mathematical formalization of the proposed similarity model is formed from the combination of the three aspects. The proposed similarity model is presented in the following algorithm.

Algorithm: the proposed similarity model.

Input : user-movie rating matrix.
Output: the modified similarity measure.

```
Begin
U the number of all users in the dataset.
M the number of all movies in the dataset.
$U_m$ the number of users rated movie m.
Step1: For all M in user-movie rating matrix do
 Compute the inverse user frequency (IUF) for each movie m∈M using the following formula :
        $F_m = log\frac{U}{U_m}$    // $F_m$ Indicates the weight of the movie m in the similarity computation.
        End For
Step2: For all U in user-movie rating matrix do
Compute the similarity between each pair of users using the constraint Pearson correlation coefficient with its modification using $F_m$. the modified constraint Pearson correlation coefficient ($Sim^{CPCC\ modified}$) formula is:
  $Sim^{CPCC\ modified}(user\ u, user\ v) = \frac{\sum_{movie\ m\in M} F_m^2(R_{u,m} - R_{Median})(R_{v,m} - R_{Median})}{\sqrt{\sum_{movie\ m\in M} F_m^2(R_{u,m} - R_{Median})2}\sqrt{\sum_{movie\ m\in M} F_m^2(R_{v,m} - R_{Median})2}}$
      End For
Step3:For all U in user-movie rating matrix do
 Compute the similarity between each pair of users using the jaccard similarity distance to consider the proportion of common movies using the formula:
      $Sim^{Jaccard}(user\ u, user\ v) = 1 - \frac{sum(m_u\ \&\ m_v)}{sum(m_u\ |\ m_v)}$    // $m_u$ , $m_v$ arethe no. of movies rated by users u and v respectively.
End For
Step4: For all U in user-movie rating matrix do
        Multiply step2 and step 3 for user u and user v to gain the proposed similarity model:
                $Sim^{modified}(user\ u, user\ v) = Sim^{CPCC\ modified}(u,v) \times Sim^{jaccard}(u,v)$
        End For
End
```

5. Experimental Results and Analysis

In this section, an empirical methodology and analysis for each similarity measure mentioned in subsection B of section 3.2 is presented. The experiments are implemented using MATLAB. Then a comparison is done between the proposed similarity model and traditional similarity measures to verify the accuracy of the suggested model. A movielens dataset is used as a input data in the calculation. This dataset has 943 users and 1862 movies. triple fields were extracted from the Movielens data set, the user ID, Movie ID and rating. This dataset was chosen because it has different scaling values for users and different number of ratings for each user. Results will be presented in the following tables and a discussion will be shown for each table. Experiments were implemented over all the users and the movies of movielens dataset; however, the following tables will show the similarity matrix for 10 users as an example. The user-user 10×10 adjacency matrix is used for representation and the values above the diagonal is considered only. Table-1 shows the number of co-rated movies between users. These numbers are needed during the similarity calculation.

Table 1-The Number of co-Rated Movies between 10 Users

	User1	**User2**	**User3**	**User4**	**User5**	**User6**	**User7**	**User8**	**User9**	**User10**
User1	262	15	7	4	73	87	140	29	4	70
User2	15	52	8	3	3	26	16	4	3	12
User3	7	8	44	6	1	8	12	5	0	5
User4	4	3	6	14	1	0	5	5	0	1
User5	73	3	1	1	165	39	97	18	4	32
User6	87	26	8	0	39	201	123	14	6	88
User7	140	16	12	5	97	123	393	35	8	111
User8	29	4	5	5	18	14	35	49	1	16
User9	4	3	0	0	4	6	8	1	12	7
User10	70	12	5	1	32	88	111	16	7	174

Table-2 presents the similarity results after MATLAB implementation using PCC formula on 10 users of Movielens data set. As shown in the Table, the similarity values cannot distinguish between users that have positive or negative impact and all the similarity values are positive. So effective users will not be recognized by this measure. Another concern about Pearson, the similarity calculations gives no indication between users that rate few movies from users with large number of rated movies. Table-3 presents the similarity values after applying cosine measure formula. It was shown from the results, that this measure cannot be used to generate the neighborhood because the values are far from the average rating of each user. Using this measure will lead to poor neighborhood.

Table 2-Pearson Similarity computation Matrix

	User1	**User2**	**User3**	**User4**	**User5**	**User6**	**User7**	**User8**	**User9**	**User10**
User1	1.0000	0.9545	0.8555	0.9318	0.9285	0.9527	0.9401	0.9754	0.9690	0.9677
User2	0.9545	1.0000	0.9522	0.9918	0.9829	0.9565	0.9624	0.9664	0.8907	0.9770
User3	0.8555	0.9522	1.0000	0.9484	1.0000	0.8808	0.8721	0.8785	0.0000	0.9214
User4	0.9318	0.9918	0.9484	1.0000	1.0000	0.0000	0.9058	0.9816	0.0000	1.0000
User5	0.9285	0.9829	1.0000	1.0000	1.0000	0.9355	0.9036	0.9537	0.8807	0.9340
User6	0.9527	0.9565	0.8808	0.0000	0.9355	1.0000	0.9579	0.9885	0.9583	0.9796
User7	0.9401	0.9624	0.8721	0.9058	0.9036	0.9579	1.0000	0.9645	0.9337	0.9772
User8	0.9754	0.9664	0.8785	0.9816	0.9537	0.9885	0.9645	1.0000	1.0000	0.9839
User9	0.9690	0.8907	0.0000	0.0000	0.8807	0.9583	0.9337	1.0000	1.0000	0.9931
User10	0.9677	0.9770	0.9214	1.0000	0.9340	0.9796	0.9772	0.9839	0.9931	1.0000

Table 3-Cosine Similarity computation Matrix

	User1	**User2**	**User3**	**User4**	**User5**	**User6**	**User7**	**User8**	**User9**	**User10**
User1	0.0000	0.1468	0.0507	0.0513	0.3648	0.4122	0.4380	0.2955	0.0825	0.3620
User2	0.1468	0.0000	0.1258	0.1177	0.0494	0.2236	0.1028	0.0861	0.0959	0.1227
User3	0.0507	0.1258	0.0000	0.2367	0.0234	0.0730	0.0623	0.0735	0.0000	0.0535
User4	0.0513	0.1177	0.2367	0.0000	0.0131	0.0000	0.0508	0.1548	0.0000	0.0171
User5	0.3648	0.0494	0.0234	0.0131	0.0000	0.2327	0.3613	0.2267	0.0797	0.1886
User6	0.4122	0.2236	0.0730	0.0000	0.2327	0.0000	0.4718	0.1535	0.1066	0.5174
User7	0.4380	0.1028	0.0623	0.0508	0.3613	0.4718	0.0000	0.2588	0.1156	0.4605
User8	0.2955	0.0861	0.0735	0.1548	0.2267	0.1535	0.2588	0.0000	0.0285	0.1976
User9	0.0825	0.0959	0.0000	0.0000	0.0797	0.1066	0.1156	0.0285	0.0000	0.1608
User10	0.3620	0.1227	0.0535	0.0171	0.1886	0.5174	0.4605	0.1976	0.1608	0.0000

In Table-4, the CPCC measure formula is applied; it is shown from the results that this measure distinguishes between positive and negative impacts of users according to their rating. NAN (divide by zero) values showed up during the calculation (as shown in Table-4), this is because the rating values for some users has the same median's value of the rating scale. This measure has good impact during neighborhood generation, because only positive similarity values are taken and negative values will be discarded (inverse relation between users). And by that, using this measure in the proposal model may contribute to find similar users.

Table 4-Constraint Pearson Similarity computation Matrix

	User1	**User2**	**User3**	**User4**	**User5**	**User6**	**User7**	**User8**	**User9**	**User10**
User1	1.000	0.632	-0.105	0.309	0.465	0.607	0.543	0.830	0.738	0.718
User2	0.632	1.000	-0.674	0.816	0.866	0.467	0.689	0.680	0.192	0.802
User3	-0.105	-0.674	1.000	-0.195	1.000	-0.433	-0.182	0.123	0.000	0.302
User4	0.309	0.816	-0.195	1.000	NaN	0.000	0.000	0.837	0.000	1.000
User5	0.465	0.866	1.000	NaN	1.000	0.420	0.160	0.594	-0.175	0.181
User6	0.607	0.467	-0.433	0.000	0.420	1.000	0.630	0.863	0.385	0.802
User7	0.543	0.689	-0.182	0.000	0.160	0.630	1.000	0.717	0.513	0.828
User8	0.830	0.680	0.123	0.837	0.594	0.863	0.717	1.000	NaN	0.874
User9	0.738	0.192	0.000	0.000	-0.175	0.385	0.513	NaN	1.000	0.927
User10	0.718	0.802	0.302	1.000	0.181	0.802	0.828	0.874	0.927	1.000

In Table-5, the Jaccard similarity measure formula was implemented on the dataset, it was shown from the results, users with more co-rated movies will have a relatively high similarity value and by that the proposal model can benefit from this aspect.

Table 5-Jaccard Similarity computation Matrix

	User1	**User2**	**User3**	**User4**	**User5**	**User6**	**User7**	**User8**	**User9**	**User10**
User1	0.0000	0.9498	0.9766	0.9853	0.7938	0.7686	0.7282	0.8972	0.9852	0.8087
User2	0.9498	0.0000	0.9091	0.9524	0.9860	0.8855	0.9627	0.9588	0.9508	0.9439
User3	0.9766	0.9091	0.0000	0.8846	0.9952	0.9662	0.9718	0.9432	1.0000	0.9765
User4	0.9853	0.9524	0.8846	0.0000	0.9944	1.0000	0.9876	0.9138	1.0000	0.9947
User5	0.7938	0.9860	0.9952	0.9944	0.0000	0.8807	0.7896	0.9082	0.9769	0.8958
User6	0.7686	0.8855	0.9662	1.0000	0.8807	0.0000	0.7389	0.9407	0.9710	0.6934
User7	0.7282	0.9627	0.9718	0.9876	0.7896	0.7389	0.0000	0.9140	0.9798	0.7566
User8	0.8972	0.9588	0.9432	0.9138	0.9082	0.9407	0.9140	0.0000	0.9833	0.9227
User9	0.9852	0.9508	1.0000	1.0000	0.9769	0.9710	0.9798	0.9833	0.0000	0.9609
User10	0.8087	0.9439	0.9765	0.9947	0.8958	0.6934	0.7566	0.9227	0.9609	0.0000

In Table-6, the IUF transformation is computed for 10 movies from Movielens dataset using the logarithmic equation mentioned in section 3.2. The CPCC is modified using the obtained weight F_m value for each movie. Ten movies were taken as an example showing the weighting scheme using the IUF approach. It was shown that movies with less number of users will have high F_m weight. So movies not seen or not known to the whole (long tail case), will have a high F_m weight and so to be recommended to target users and not neglected.

Table 6-Inverse User Frequency (IUF) Transformation for 10 Movies

	Movie ID1	**Movie ID2**	**Movie ID3**	**Movie ID4**	**Movie ID5**	**Movie ID6**	**Movie ID7**	**Movie ID8**	**Movie ID9**	**Movie ID10**
Number of users to movie	392	121	85	198	79	23	346	194	268	82
Weight (IUF)	0.877	2.0533	2.4064	1.5608	2.4796	3.7136	1.0026	1.5812	1.2581	2.4423

In Table-7, the proposed model mentioned in the Algorithm of section 4 is applied. The suggested model benefited from the three measures (constraint Pearson, jaccard, IUF) to reach similar users (neighborhood) to the target user during the similarity calculation. This model considered the user's view point through the constraint and jaccard measures by considering the positive impact and number of rated movies for each user and considered the movie's influence in similarity computation through using the IUF transform by focusing on movies little being watched.

Table 7-Proposed Similarity Model Computation Matrix

	User1	**User2**	**User3**	**User4**	**User5**	**User6**	**User7**	**User8**	**User9**	**User10**
User1	0.000	0.765	-0.453	0.692	0.390	0.451	0.380	0.709	0.644	0.568
User2	0.765	0.000	-0.766	0.918	0.916	0.403	0.656	0.333	-0.306	0.761
User3	-0.453	-0.766	0.000	0.329	0.995	-0.035	-0.283	0.133	0.000	0.698
User4	0.692	0.918	0.329	0.000	NaN	0.000	-0.173	0.825	0.000	0.995
User5	0.390	0.916	0.995	NaN	0.000	0.216	0.129	0.673	-0.216	0.257

User6	0.451	0.403	-0.035	0.000	0.216	0.000	0.427	0.851	0.723	0.549
User7	0.380	0.656	-0.283	-0.173	0.129	0.427	0.000	0.730	0.562	0.640
User8	0.709	0.333	0.133	0.825	0.673	0.851	0.730	0.000	NaN	0.784
User9	0.644	-0.306	0.000	0.000	-0.216	0.723	0.562	NaN	0.000	0.897
User10	0.568	0.761	0.698	0.995	0.257	0.549	0.640	0.784	0.897	0.000

Table-8 shows a significant test as a comparison between the traditional similarity measures and the proposed model and a ranking is given for each measure. From the user-movie rating matrix, User ID 1 is considered as the target user and a similarity computation with 9 users (ID2 to ID10) is calculated using various measures. As shown in Table-8, the following clarifications are observed:

- The similarity weights using PCC differ slightly although the number of co-rated movies differs diversely and largely.
- Using PCC there is no indication for positive or negative weights as compared with CPCC and modified CPCC which have negative weights as with user ID3. The negative weight will rank the user far from the target because the movie weights which have less rated users will be increased slightly and vice versa.
- Using the cosine similarity measure will yield weights far from all calculated weights because it does not take the average rating value of the dataset scale in its formula.
- The jaccard measure gives no influence alone.
- As a resultant , the proposed model combine the three measures(CPCC, Jaccard and IUF) to enhance the similarity weights by considering the mentioned aspects.as an example, user ID 4 and user ID 9 have the same co-rated values (4),they have approximately same weights as a comparative with the other measures.

Table 8-Comparison Proposed Model and Similarity Measures with their Ranks

Users ID's	Co-rated movies 0f User ID1	Pearson Similarity Measure		Cosine Similarity Measure		Constraint Similarity Measure		Modified Constraint with the IUF		Jaccard Distance Measure		Proposed Similarity Measure	
		Weight	Rank	Weight	Rank	Weight	Rank	Weight	Rank	Weight	Rank	Weight	Rank
ID2	15	0.954	4	0.146	6	0.632	4	0.804	1	0.949	6	0.765	1
ID3	7	0.855	9	0.050	9	-0.105	9	-0.465	9	0.976	7	-0.453	9
ID4	4	0.931	7	0.051	8	0.309	8	0.7023	3	0.985	9	0.692	3
ID5	73	0.928	8	0.364	3	0.465	7	0.490	8	0.793	3	0.390	7
ID6	87	0.952	5	0.412	2	0.607	5	0.586	6	0.768	2	0.451	6
ID7	140	0.940	6	0.438	1	0.543	6	0.521	7	0.728	1	0.380	8
ID8	29	0.975	1	0.295	5	0.830	1	0.790	2	0.897	5	0.709	2
ID9	4	0.969	2	0.082	7	0.738	2	0.655	5	0.985	8	0.644	4
ID10	70	0.967	3	0.362	4	0.718	3	0.7022	4	0.808	4	0.568	5

Conclusion

The proposed model utilized the specific meaning of rating rather than just calculating distances between users. Moreover in this model, less known items were focused on and treated effectively and as a result the accuracy of prediction can be improved. Many similarity measures were conducted; such as Pearson correlation, cosine, it was concluded that it was not possible to relate between users effectively, since it provides a relatively equivalent similarity values. But in the proposed approach, each user becomes comparable, since it provides different similarity values for each pair of users.it was concluded from this research, that the user can be distinguished as a dependable user in the prediction process for target users and the less known products will be more focused on.

References

1. Ricci F., Rokach L., & Shapira B. **2015.** *Recommender Systems Handbook*. 2nd ed., Springer Science+Business Media, ISBN 978-1-4899-7637-6, New York.
2. Cacheda, F., Carneiro, V.,Ferna, D. and Formoso, V. **2011.** Comparison Of Collaborative Filtering Algorithms: Limitations Of Current Techniques And Proposals For Scalable, High-Performance Recommender System. *ACM Transactions On The Web*, **5**(1).
3. KG, S. and Sadasivam, G. S. **2017.** Modified Heuristic Similarity Measure for Personalization using Collaborative Filtering Technique. *Appl. Math*, **11**(1): 307-315.
4. Aygün S. and Okyay, S. **2015.** Improving the Pearson Similarity Equation for Recommender Systems by Age Parameter. *IEEE*,: 1-6.
5. Huang, B. H. and Dai B. R. **2015.** A Weighted Distance Similarity Model to Improve the Accuracy of Collaborative Recommender System. 16th IEEE International Conference on Mobile Data Management, **2**: 104-109.
6. Liang, S., Ma, L. and Yuan, F. **2015.** A singularity-based user similarity measure for recommender systems. *International journal of innovative computing information and control*, **11**(5): 1629-1638.
7. Liu, H., Hu, Z., Mian, A., Tian, H. and Zhu, X. **2014.** A new user similarity model to improve the accuracy of collaborative Filtering. *Elsevier, Knowledge-Based Systems*, **56**: 156–166.
8. Candillier, L., Meyer, F. and Fessant, F. **2008.** Designing Specific Weighted Similarity Measures to Improve Collaborative Filtering Systems. ICDM In Industrial Conference on Data Mining, Springer, pp. 242–255.
9. Hyung, J. and Ahn, J. **2008.** A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. *Elsevier, Information Sciences*, **178**: 37–51.
10. Weng, L.T., Xu, Y., Li, Y. and Nayak, R. **2005.** An Improvement to Collaborative Filtering for Recommender Systems. IEEE, Proceedings of the 2005 International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'05), **1**: 792-795.
11. Lu L., Medo, M., Yeung, C., Zhang, Y., Zhang, Z. and Zhou, T. **2012.** Recommender systems. *Elsevier, Physics Reports*, **519**(1): 1–42.
12. Jannach, D., Zanker, M., Felfernig, A. and Friedrich G. **2011.** *Recommender Systems: An Introduction*. Cambridge University Press, USA.
13. Rajaraman J., Leskovec, A. and Ullman, J. D. **2014.** *Mining of Massive Datasets*. Cambridge University press.
14. Shardanand, U. and Maes, P. 1995. Social information filtering: Algorithms for automating word of mouth. ACM Press/Addison-Wesley Publishing Co., Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp:210–217.
15. Mao, J., Cui, Z., Zhao, P. and Li, X. 2014. An improved similarity measure method in Collaborative Filtering Recommendation Algorithm. IEEE, International Conference on Cloud Computing and Big Data, pp 297-303.
16. Breese, J. S., Heckerman, D. and Kadie, C. 1998. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. Presented at 14th Conference on Uncertainty in Artificial Intelligence, Madison.
17. Dillon, M. 1983. Introduction to modern information retrieval. McGraw-Hill, pp. xv+ 448 ISBN 0-07-054484-0, New York.
18. Adomavicius, G. and Tuzhilin, A. 2005. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Transaction, Knowledge Data Eng., 17 (6): 734–749.