A note on Pearson Correlation Coefficient as a metric of similarity in recommender system

Leily Sheugh
Faculty of Computer and IT
Islamic Azad University, Qazvin branch
Qazvin, Iran
leily.sheugh@gmail.com

Sasan H. Alizadeh
Faculty of Computer and IT
Islamic Azad University, Qazvin branch
Qazvin, Iran
Sasan.H.Alizadeh@qiau.ac.ir

Abstract— **Recommender systems help users to find information that best fits their preferences and needs in an overloaded search space. Most recommender systems researches have been focused on the accuracy improvement of recommendation algorithms. Choosing appropriate similarity measure is a key to the recommender system success for this target. Pearson Correlation Coefficient (PCC) is one of the most popular similarity measures for Collaborative filtering recommender system, to evaluate how much two users are correlated. While Correlation-based prediction schemes were shown to perform well, they suffer from some limitations. In This paper we present an extension toward Pearson Correlation Coefficient measure for cases which does not exist similarity between users by using it. Experimental result on the film trust data set demonstrate via our proposed measure and PCC we can achieve better result for similarity measure than traditional PCC.**

Keywords—recommender system, Collaborative Filtering, similarity measure, Pearson Correlation Coefficient

I. INTRODUCTION

Recommender systems have become a mainstream research field in information technologies. They use the different sources of information (Collaborative, social, demographic, content, implicit and explicit data acquisition, etc) .Information can be acquired explicitly or implicitly by monitoring users' behavior. Thereupon they enable to collect information on the priority of its users for a set of items or elements. [1-3].

One of the most well-know and successful recommender system is the Collaborative Filtering[1] that commonly used techniques to generate recommendations. It analyzes relationships between users (as users like mined neighbors) and mutual-dependencies between products to identify new user-item associations [4 -6].

The Collaborative Filtering includes item-based, user-based and model-based [7]. In case of Item based Collaborative Filtering, predicts the similarity among items by adopt pairwise item similarities which are more reliable than user similarities [8]. the user-based approaches, a similarity matrix is adopted to store the rating of each user for every item as it's based on the ratings given by the users nearest neighbors has been find [8]. The model-based construct a model to describe the behavior of user and then predict the rating of items via take advantage of the sparsity of data in the similarity matrix [9]. The problem of all collaborative filtering system is estimated how well a user will like an item that he/she has not rated so the similarity computation phase for any Collaborative Filtering plays an important role for its success [10].

The current most common similarity measure for collaborative filtering recommender system is Pearson Correlation Coefficient[2] measure [6, 11, 12].

However traditional PCC can be used as similarity measure for CF. the ultimate goal is to improve the accuracy of the Pearson Correlation Coefficient measure; it will be happen by different techniques including: extension [13-16] or changed the PCC[17].

Traditional PCC does not consider the size of the set of common users. To solve this problem, Pearson Correlation Coefficient based on weight has been proposed [13].SPCC is scalable Pearson Correlation Coefficient algorithm that uses the cluster for neighbor pre-selection [14].CPCC is confidence-aware Pearson Correlation Coefficient that consider the rating confidence; confidential weight of an item rated by the active user[15]. Most of the traditional measure such as PCC are symmetric which means that they always assign equal similarity to each user even when one user' behavior is quite

[1] Collaborative Filtering(CF)

[2] Pearson correlation coefficient

978-1-4799-8733-7/15/$31.00 ©2015 IEEE

similar to the other but not conversely. Incorporating a weighting schema in PCC can improve that's result [16].

Proximity-Impact-Popularity (PIP) is a similarity measure that analyzed the drawback of Pearson correlation coefficient. This new similarity considered three aspects: proximity, impact and popularity of the user ratings. But, this similarity considers only the local information of the ratings and does not consider the global preference of user ratings [17].

However Pearson Correlation Coefficient is a good measure but the former issue refers to the difficulty in finding sufficient and reliable similarity measure Therefore in this paper we propose a novel approach for Pearson Correlation Coefficient to solve some shortage of traditional PCC.

The rest of this paper is organized as follows. Section 2 gives a brief overview of related research on traditional the similarity measure for Collaborative Filtering. Section 3 describes the Pearson Correlation Coefficient and its limitation. Section 4 proposed new similarity measure. Section 5 considers experimental result. Finally, in Section 6 we conclude our work.

II. SIMILARITY MEASURE

The core of Collaborative Filtering is to calculate similarity among items or users. A metric or a Similarity Measure[3] determines the similarity between pairs of users (user to user CF) or the similarity between pairs of items (item to item CF).

Collaborative Filtering system using neighborhood-based algorithm for providing personalized prediction. That's used Pearson Correlation to weight user similarity, used all correlated neighbors, then computed a final prediction by performing a weighted average of deviations from the neighbor's mean[18,19]:

$$p_{a,i} = \bar{r}_a + \frac{\sum_{u=1}^{n} \left(r_{u,i} - \bar{r}_u\right) * w_{a,u}}{\sum_{u=1}^{n} w_{a,u}} \qquad (1)$$

$p_{a,i}$ Represent the prediction for the active user a for item i . n Is the number of neighbors and $w_{a,u}$ is the similarity weight between the active user and neighbor u as defined by pcc .

[3] Similarity Measure(SM)

Many Statistical Coefficients can be used as similarity measures for Collaborative Filtering recommender system. The final goal in Collaborative Filtering is get a set of neighbors that are closed as possible as given active user by using similarity measure that enhanced the accuracy.

Traditionally, the series of statistical metrics have been used in Collaborative Filtering such as the Pearson Correlation Coefficient, cosine[4], constraint Correlation[5] , and mean squared differences[6] and Euclidean[7]; the relatedness concept was introduced to provide the importance of the relationship between users and items [19-22].

Table1 shows a classification of the memory-based CF similarity measure .as Heuristic Similarity Measure that named (PIP) and outperforms the traditional statistical Similarity Measure (Pearson Correlation, cosine, etc.)[23].Predicts first actual ratings and subsequently identifies prediction errors for each user named (UERROR), and a metric based on neural learning (model-based CF) and adapted for new user cold-start situations, called (NCS) [6].

TABLE I. TESTED COLLABORATIVE FILTERING SIMILARITY MEASURES

	Not based on models		**Model-based**
	No trust extraction	***Trust extraction***	
Traditional (only the ratings of both users or both items) Not tailored to cold-start users	JMSD,CORR,CCORR, COS,ACOS,MSD,EUC		GEN[8]
Tailored to cold-start users	PIP[9]	UERROR	NCS
Extended to all the ratings	SING[10]	TRUST	

However there are several similarity metrics to calculate similarity, but three commonly used similarity metrics are: PCC, COS and MSD [19, 20]. The formulas COS and MSD are defined as follow:

$$\cos(u,u') = \frac{\sum_{i \in I} r_{ui} r_{u'i}}{\sqrt{\sum_{i \in I} \left(r_{ui}\right)^2} \sqrt{\sum_{i \in I} \left(r_{u'i}\right)^2}} \qquad (2)$$

[4] Cosine(COS)
[5] constraint Correlation(CCORR)
[6] mean squared differences(MSD)
[7] Euclidean(EUC)
[8] genetic-based(GEN)
[9] Proximity-Impact-Popularity(PIP)
[10] Singularities(SING)

In Equation 2 cosine similarity between two users is $\cos(u,u')$. I Represents the set of all items rated by both user u and u'. r_{ui} denotes the rating of Item i by user u and $r_{u'i}$ denotes the rating of Item i by user u'.

$$MSD(u,u') = \frac{\sum_{i \in I} \left(r_{u,i} - r_{u',i}\right)^2}{I} \qquad (3)$$

In Equation 3 Mean square difference between two users is $MSD(u,u')$, I represents the set of all items rated by both user u and u'. r_{ui} denotes the rating of Item i by user u and $r_{u'i}$ denotes the rating of Item i by user u'.

PEARSON CORRELATION COEFFICIENT AND LIMITATION

In Collaborative Filtering predictions for a user can be based on the similarity between the interest profile of that user and other users. Suppose that a database of user ratings items to exist, where users indicate their interest an item on a numeric scale. It is now possible to define similarity measures between two user profiles, as u and u' by Pearson Correlation Coefficient, $pcc(u,u')$. Once the similarity between profiles has been quantified, it can be used to compute personalized recommendations for users[18,21].

Pearson Correlation Coefficient is a statistical measurement of linear Correlation between two variables. Equation 4 gives the pcc formula of two users u and u' [19].

$$pcc(u,u') = \frac{\sum_{i \in I} \left(r_{u,i} - \bar{r}_u\right)\left(r_{u',i} - \bar{r}_{u'}\right)}{\sqrt{\sum_{i \in I} \left(r_{u,i} - \bar{r}_u\right)^2} \sqrt{\sum_{i \in I} \left(r_{u',i} - \bar{r}_{u'}\right)^2}} \qquad (4)$$

Here let $r_{u,i}$ and $r_{u',i}$ be the rating scores from two users, and $\bar{r}_u$ and $\bar{r}_{u'}$ denote the average ratings by the two users and $I = I_u \cap I_{u'}$ denotes the set of items rated by both users u and u'. Where $pcc(u,u') \in$ [-1, 1] is the similarity between two users u and u'. $pcc(u,u')$=1 is for the purpose of consistency. In particular, $pcc(u,u')$>0 indicate positive Correlation between users u and u'. $pcc(u,u')$<0 mean opposite Correlation and $pcc(u,u')$=0 implies no Correlation. Therefore, pcc value can capture the rating similarity between the two users.

So, by pcc measure, we can filter out some users' pairs that have more or less similarity to rating scores on the same items. Based on the assumption users with similar tastes on different types of products have higher probability to form a community and are likely to make associates to each other even if they don't know each other before.

Similarity predictions based on Pearson Correlation Coefficient suffer from several limitations:

First, Correlation between two user profiles can only be computed based on items that both users have rated, i.e. If users can select among thousands of items to rate, it is likely that overlap of rated items between two users will be small in many cases. Therefore, computed Correlation cannot be regarded as a trustworthy measure of similarity because computed Correlation Coefficients are based on just a few observations[24].

the Second, the Correlation approach induces one global model of similarities between users, rather than separate models for classes of ratings (e.g. positive rating vs. Negative rating)[24].

Third, two users can only be similar if there is overlap among the rated items, i.e. if users did not rate any common items, their user profiles cannot be correlated. Consider the following example: Users A and B are highly correlated, as are users B and C. This relationship provides information about the similarity between users A and C as well. However, in case users A and C did not rate any common items, a Correlation-based similarity measure could not detect any relation between the two users [24].

Forth, and perhaps most importantly, when the rating data for Collaborative Filtering are extremely sparse, it will be difficult to present accurate predictions using the Pearson correlation-based Collaborative Filtering [25].

In addition of listed above the other problems of using Pearson Correlation Coefficient don't get any number[11] of results for similarity measure between two users. It happen by following reasons (NAN is a measure that used in matlab software):

1. If the users don't have common items with the other users, similarity measure can be considered NAN results.
2. Be zero the variance of the problems that they gives NAN similarity and occur for two bellow reasons:

[11] not a number(NAN)

2.1. If the user has one rating in this condition mean rating is equal rating; therefore the result of this problem, it gives as zero variance and *NAN* similarity.

2.2. If all of the items have the same ratings, the mean of ratings will be equal with common rating, which it gives as zero variance and *NAN* similarity again.

III. NEW SIMILARITY MEASURE

One of the limitations of using Pearson Correlation Coefficient for similarity measure is given a set of *NAN* result that these reasons to mention in the previous section.

In this section, we will present the proposed method the basic principle of which is reduced *NAN* result in similarity measure for all of the items have the same ratings, and the mean of ratings is equal with common rating. Equation 5 gives the *newpcc* formula of two users u and u' for addresses this problem.

$$newpcc(u,u') = \frac{\left(-\left|\mu_u - \mu_{u'}\right| + \max rate\right) - meanrate}{meanrate} \qquad (5)$$

μ_u Is the average rating of user u, $\mu_{u'}$ is the average rating of user u' and Mean rate is the middle of all ratings.

4.1. An example:

In this subsection we intend to exemplify step by step the use of *newpcc* to decrease a *NAN* similarity for a given users. Suppose there are five user and five items, denoted by u_k and i_k where $k, j \in [1,5]$. Each user may rate a few items by giving an integer rating rated in [1, 5] as shows in table 2.

The first step of *newpcc* is to identify the users that rate the common items. Second, peruse the proviso equal items rating with mean of all ratings. Third calculates the similarity measure for users with common rating for items and mean rating by Eq. (5) and other similarity user can be inferred by Eq. (4), as shows in table 3.

TABLE II. THE DATA SET CONSISTING OF USER- ITEM RATING MATRIX

	I_1	I_2	I_3	I_4	I_5	I_6
U_1		1	5			
U_2	3	5	1			
U_3				5	5	5
U_4				3	3	3
U_5						

TABLE III. USER-USER SIMILARITY MATRIX

	U1	U2	U3	U4	U5
U1	1	-1	0	0	0
U2	-1	1	0	0	0
U3	0	0	1	-0.6	0
U4	0	0	-0.6	1	0
U5	0	0	0	0	1

IV. CASE STUDY AND EXPERIMENTAL RESULT

In order to verify the effectiveness of the new proposed method, we conduct experiments on Real word data set which namely film trust that contains user-item rating. Film trust is a trust-based social site in which users can rate and review movies. Its contain 1896 user, 2071 movies and 35497 rating.

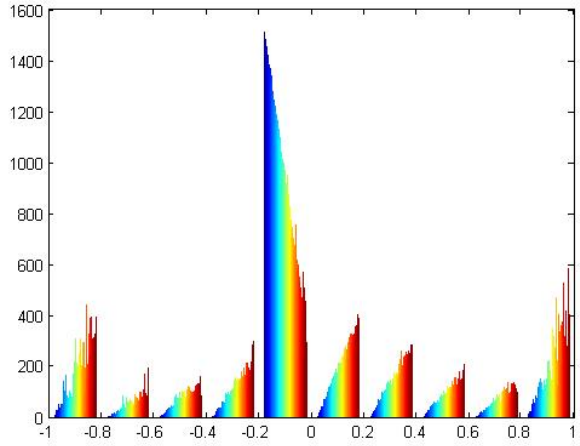

Fig.1. similarity of film trust data set based on *pcc*

Fig. 1 show the histogram plot of *NAN* similarity measure based on traditional PCC for Film Trust. It is the extent to which two user linearly relate with each other as well as represents a lot of users no Correlation.

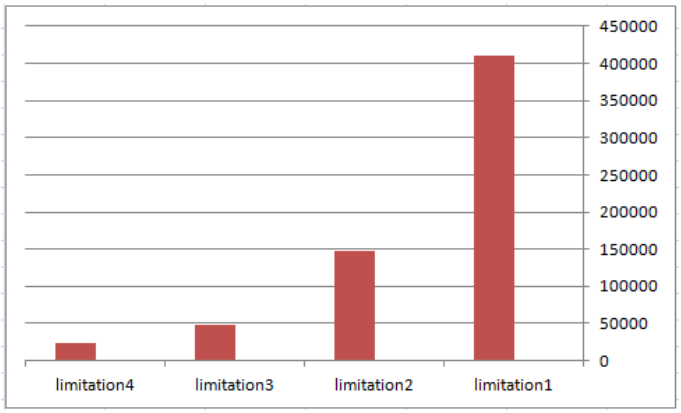

Fig.2. the histogram limitation for traditional PCC

Fig.2. shows the reason that it gives the similarity of on the Film Trust data set by traditional PCC. The figure has seen the most of user don't have intersected item with others. It is the first and main limitation of recommender system based on PCC to accrue especially when data are sparse. The second

limitation happens when two users have only rated one item in common and its rating is equal with the mean items ratings by users. Three and four limitation include the condition of all items that ratings and mean ratings are equal.

To address above limitation and propose a better similarity measure, we design a new equation for similarity measure that present in previous section. By using the new proposed measure the volume of useful information has been increased.

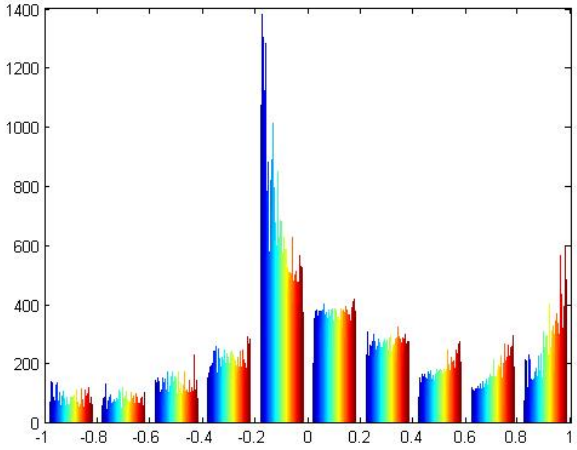

Fig.3. similarity Film Trust data set based on *pcc* and *newpcc*

Fig.3 shows the histogram plot of similarity measure based on traditional *pcc* and *newpcc* for similarity measure. Its result shows that using *newsim* to decrease *NAN* result for state all items has the same ratings with mean ratings. The result shows that it can be calculate the limitation of *pcc* and improve the correlation between users about 11.3%.

V. CONCLUSION AND FUTURE WORK

At first using the review of papers related to Pearson correlation coefficient has a number of known limitations by other researchers in this paper were collected. Then according to the variance of the ratings is needed to calculate the PCC in the denominator of fraction; PCC does not exist for cases, this measure becomes zero. By considering this limitation, we have proposed a new modified version of PCC.

The result of applying PCC measure for Film Trust dataset demonstrated that PCC measure could not be calculated as a similarity measure for almost 30% of total dataset; but almost 11.3% of it can be calculated with the new measuring now. Therefore using the new proposed measure the volume of useful information has been increased potentially and it will be create a suitable platform for rising the recommender system.

In Future work, whereas mentioned the introduction has been a very comprehensive research to create a recommender system based on PCC similarity measure, this research creates the necessary of infrastructure to do more research in order to improve and enhance accuracy these systems too. Used the result of new similarity measure for recommender system and revise social CRM by real dataset.

REFERENCES

[1] S.K. Lee, Y.H. Cho and S.H. Kim, "Collaborative filtering with ordinal scale-based implicit ratings for mobile music recommendations," Information Sciences, 180 (11), pp: 2142–2155, 2010.

[2] K. Choi, D. Yoo, G. Kim and Y. Suh, "A hybrid online-product recommendation system: combining implicit rating-based Collaborative filtering and sequential pattern analysis," Electronic Commerce Research and Applications, 11(4), pp: 309-317, 2012.

[3] E.R. Núñez-Valdéz, J.M. Cueva-Lovelle, O. Sanjuán-Martı´nez, V. Garcı´ a-Dı az, P.Ordoñez and C.E. Montenegro-Marı´ n, "Implicit feedback techniques on recommender systems applied to electronic books," Computers in Human Behavior 28 (4) ,pp:1186–1193,2012.

[4] B. N. Miller, I.Albert, S.K.Lam, J.A. Konstan, and J. Riedl, "Movie Lens unplugged: Experiences with an occasionally connected recommender system,"In Proceedings of the 8th international conference on intelligent user interfaces, New York, pp: 263-266, 2003.

[5] Y. Koren, "Factorization meets the neighborhood: a multifaceted Collaborative filtering model," In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp: 426–434, 2008.

[6] J. Bobadilla, F. Ortega, A. Hernando, J. Bernal," A Collaborative filtering approach to mitigate the new user cold start problem", Knowledge Based System, 26,225–238, 2012.

[7] M. Khabbaz and L. V. S Lakshmanan."Top Recs: Top-k algorithms for item-based Collaborative filtering," In Proceedings of the 14th International Conference on Extending Database Technology, pp: 213-224, 2011.

[8] J. Wang, A. P. deVries and M. J. T. Reinders, "Unifying user- based and item-based Collaborative filtering approaches by similarity fusion," 06 Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 501-508, 2006.

[9] R. Jin, L. Si and C. Zhai," A study of mixture models for Collaborative filtering. Journal of Information Retrieval, "9(3), 357–382, 2006.

[10] J. B. Schafer, D. Frankowski, J. Herlocker, and S.Sen, "Collaborative filtering recommender systems," Lecture Notes in Computer Science, 4321,pp: 291–324, 2007.

[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, pp. 175–186, 1994.

[12] Rajaraman, A., Leskovec, J., & Ullman, J. D. "Mining of massive datasets," Cambridge University Press, 2012.

[13] 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, pp. 230–237,1999.

[14] G. Xue, C. Lin, Q. Yang, W. Xi, H. Zeng, Y. Yu, Z. Chen, "Scalable collaborative filtering using cluster-based smoothing," in: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 114–121, 2005.

[15] G. Guo, J. Zhang, D. Thalmann, "Merging trust in collaborative filtering to alleviate data sparsity and cold start", Knowledge Based System. (KBS), 57 pp. 57–68, 2014.

[16] P. pirasteh, J. jung, D. Hwang," An Asymmetric Weighing Schema for Collaborative filtering," studies in computational intelligence, 572, 77-82, 2015.

[17] H.J. Ahn, "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem," Information Sciences. 178(1), 37–51, 2008.

[18] G. Adomavicius, A. Tuzhilin, "toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," IEEE Transactions on Knowledge and Data Engineering 17 (6), 734–749, 2005.

[19] J. Bobadilla, F. Ortega, A. Hernando, "A Collaborative filtering similarity measure based on singularities," Information Processing and Management 48 (2), 204–217, 2012.

[20] D. Jannach, M. Zanker, A. Felfernig, G. Friedrich, "Recommender Systems: An Introduction," Cambridge, NY, 2010.

[21] J. Wang, A.P. Vries, M.J. Reinders, "Unified relevance models for rating prediction in Collaborative filtering," ACM Transactions in Information Systems (TOIS) 26 (3), 1–42, 2008.

[22] H.J. Ahn," A new similarity measure for Collaborative filtering to alleviate the new user cold-starting problem," Information Sciences. 178 (1), 37–51, 2008.

[23] Bobadilla, J., Ortega, F., Hernando, A., Guti´errez, A. Recommender Systems Survey. Knowledge-Based Systems 46, 109–132, 2013.

[24] Billsus, D. and M. Pazzani, "Learning Collaborative information filters," In International Conference on Machine Learning, Morgan Kaufmann Publishers, 1998.

[25] X. Su, T. M. Khoshgoftaar, X. Zhu, and R. Greiner, “Imputation-boosted collaborative filtering using machine learning classifiers,” in Proceedings of the 23rd Annual ACM Symposium on Applied Computing (SAC ’08), pp. 949–950, 2008.