Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 89 (2016) 450 – 456

Procedia Computer Science

Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016)

A New Similarity Measure Based on Mean Measure of Divergence for Collaborative Filtering in Sparse Environment

Suryakant* and Tripti Mahara

IIT Roorkee, Roorkee 247 667, India

Abstract

Memory based algorithms, often referred to as similarity based Collaborative Filtering (CF) is one of the most popular and successful approaches to provide service recommendations. It provides automated and personalized suggestions to consumers to select variety of products. Typically, the core of similarity based CF which greatly affect the performance of recommendation system is to finding similar users to a target user. Conventional similarity measures like Cosine, Pearson correlation coefficient, Jaccard similarity suffer from accuracy problem under sparse environment. Hence in this paper, we propose a new similarity approach based on Mean Measure of Divergence that takes rating habits of a user into account. The quality of recommendation of proposed approach is analyzed on benchmark datasets: ML 100 K, ML-1 M and Each Movie for various sparsity levels. The results depict that the proposed similarity measure outperforms existing measures in terms of prediction accuracy.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016

Keywords: Collaborative Filtering; Information Filtering; Recommendation System; Recommendation Accuracy; Similarity Measure.

1. Introduction

A Recommendation system (RS) is a necessity and a popular technology to handle information explosion. It serves as an information filtering tool and is commonly used to assist the target user to filter through a large pool of products and present only those products that are of user's interest. It gathers a large amount of data on the activities, inclinations, interest or taste of its client for a set of things i.e. movies, hardware item, garments and so forth and makes utilization of this gathered data to give suggestions to different clients.

In day to day life, we regularly depend on the opinion of like-minded individuals, individuals with comparable taste and preference or other trusted sources about the nature and quality of assets to get the suggestions for a product or item. A RS automates this word-of-mouth phenomenon and is generally utilized by the online shopping sites like Amazon to prescribe products of interest and by sites like Netflix to prescribe films to the users by giving personalization suggestions. The significance of RS increases inconceivably with the presence of long tail phenomenon (Anderson 2006). A physical retail store is characterized by shortage of resources. For instance, a fabric shop has constrained rack space and can show restricted assortment of items to a client. In the physical world it is impractical to

*Corresponding author. Tel: +918894969853; Fax: +91 1792 245362.
E-mail address: surya.dpt2015@iitr.ac.in

1877-0509 © 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016
doi:10.1016/j.procs.2016.06.099

customize a physical retail store for every individual client, as it is either represented by deals figures (most prevalent products) or it relies on upon master judgment. Despite what might be expected, an online retail store can make everything accessible to the client that exists. This peculiarity to makes things accessible between the offline and on-line world has been termed as the long tail[6] phenomenon. As per the long tail, offline retail stores just give the most popular things though on-line stores offer least popular items in addition most famous things. Henceforth[19], Park and Tuzhilin[19] proposed the need of a RS to prescribe things to individual clients online as it is unrealistic to present every accessible thing to the client or things that match the essence of clients.

Till date, Collaborative filtering (CF) is the most successful and widely employed approach in RS. Bobadilla *et al.*[3,9] is based on similar taste of users. It works on the fact that if a user had similar taste in past for a set of items, then they will share common taste in future. The information preferences for some resources are far more complex and hard to determine. It is due to the fact that sometimes preferences can't be just defined by using a set of keywords or by quality and taste. It can be obtained by observing the preference, behaviour or taste of other users. Tapestry[11] the first RS recommended documents, collected from newsgroup to a set of users by building database of contents and comments.

CF is one of the most widely used approaches to design a RS. However, sparsity is one of the major weaknesses of this prosperous approach. This problem inherently occurs in the system and is attributed to ever increasing number of users and items. Because of this numerous users may have evaluated or bought just couple of items from the total accessible items. Indeed, even extremely well known items may have been bought or evaluated by very few users. Hence it is difficult to compute similarity between users that leads to high sparsity in user-item ratings matrix. This affects the performance of a RS. In case, the system manages to evaluate similarity, their might exist a possibility that this similarity may be not reliable because of insufficient information processed. According to the density of matrix is lower than 1%.

In literature diverse similarity measures have been proposed but their performance is not very satisfactory for sparse matrix. Hence, the main aim of this paper is to build a new approach based on mean measure of divergence to find similarity between users to address the sparsity issue.

2. Related Work

One of the most critical factors that greatly affect the performance of CF is the computation of similarities between users. Cosine (COS), Pearson correlation coefficient (PCC), adjusted cosine measure (ACOS) and Spearman's rank correlation (SRC) are the generic traditional measures for similarity computation. These similarity measures defined in Table 1 Patra *et al.*[20].

Here the users are represented as vectors of objects based on their taste or preference history. The similarity between two users is characterized as the similarity of corresponding vectors. However literature has shown that the traditional measures do not properly utilize the user preferences (ratings data); especially when the available user item rating matrix is sparse or ratings data are not sufficient. Researchers have proposed some new similarity measures to improve the performance of a RS. Luo *et al.*[17] introduced local and global user similarity measure.

Bobadilla *et al.*[3] proposed a new similarity measure called JMSD by joining Jaccard *et al.*[14] and Mean squared-difference (MSD). Mean–Jaccard–Difference (MJD)[5] combination of six similarity measures. At that point, the neural network is utilized to tune the weights of each similarity measure. Choi & Suh[8] proposed another similarity measure that chooses neighbours dynamically for each different target item Pirasteh *et al.*[2] presented new weighted plans for customary similarity measures, which changed the symmetric similarity to asymmetric similarity. Jeong *et al.*[13] proposes to utilize an iterative message passing procedure for similarity updating. Gan and Jiang[10] utilize a power function to adjust user similarity scores.

The impact factor indicates about the preference of target user about an item. The penalty is imposed if ratings are on opposite side of the median. The popularity factor captures global information of the target item. The drawbacks of PIP[1] is addressed by Liu, Hu, Mian, Tian & Zhu, (2014) and motivated them to propose a new heuristic similarity model (NHSM). Patra *et al.*[20] proposed a similarity measure based on Bhattacharyya coefficient, it utilized all ratings given by a set of users. Chen *et al.*[7] propose new similarity measure using artificial immune network. In addition, a modified PCC formula is also proposed. To alleviated the sparsity problem of Collaborative recommender systems Leng *et al.*[15] proposed a novel similarity measure based on potential field.

Table 1. Traditional Similarity Measures.

Similarity Measures	**Computational formulae's**	Major Drawbacks
Pearson Correlation(PCC)	$\mathrm{sim}(u,u')^{PCC} = \frac{\sum_{i \in I}\left(r_{u,i}-\bar{r}_u\right)-\left(r_{u',i}-\bar{r}_{u'}\right)}{\sqrt{\sum_{i \in I}\left(r_{u,i}-\bar{r}_u\right)^2}\cdot\sqrt{\sum_{i \in I}\left(r_{u',i}-\bar{r}_{u'}\right)^2}}$ Where I is the set of items, $r_{u,i}$ rating of given to item i by user u, $\bar{r}_u$ average rating of user u	It suffers from few co-rated item problem. It outputs high (low) similarity even if there exists significant difference in ratings.
Cosine (COS)	$\mathrm{sim}(u,u')^{COS} = \frac{\sum_{i \in I}\left(r_{u,i}\right)\cdot\left(r_{u',i}\right)}{\sqrt{\sum_{i\in I}\left(r_{u,i}\right)^2}\cdot\sqrt{\sum_{i\in I}\left(r_{u',i}\right)^2}}$	It suffers from few co-rated item Problem. It gives high similarity It outputs high similarity even if there exists significant difference in ratings
Adjusted Cosine (ACOS)	$\mathrm{sim}(i,i')^{ACOS} = \frac{\sum_{u \in U}\left(r_{u,i}-\bar{r}_i\right)\cdot\left(r_{u,i'}-\bar{r}_{i'}\right)}{\sqrt{\left(r_{u,i}-\bar{r}_i\right)^2}\cdot\sqrt{\left(r_{u,i'}-\bar{r}_{i'}\right)^2}}$ Where U is the set of users rated both items i.e. i and i'	Unable to compute similarity if user U have rated only few items. It outputs high (low) similarity even if there exists significant difference in ratings.
Spearman's Rank Correlation similarity	$\mathrm{sim}(u,u')^{SRCC} = 1-\frac{6\sum_{i\in I}\mathrm{rank}\left(r_{u,i}\right)^2-\mathrm{rank}\left(r_{u',i}\right)^2}{|I|\cdot\left(|I|^2-1\right)}$ Where $\lvert I\rvert$ is the cardinality of co-rated items.	High similarity even if the ratings are similar.

Some researchers have also utilizes the metaheuristic algorithms to find the similarity. For example Bobadilla *et al.*[4] have adapted Genetic algorithm to compute similarity between a pair of users. It combines values and weights to obtain this similarity. The similarity between each pair of users is obtained by first computing the values and then genetic algorithm is adapted for weights adjustment. The weights in the framework are computed just once.

3. Proposed Similarity Measure: CjacMD

The conventional similarity measures can't be used in sparse environment, as they don't utilize ratings data efficiently. Therefore, we propose our similarity measure: CjacMD (Cosine-Jaccard-Mean Measure of Divergence) suitable in sparse dataset. This measure combines Cosine, Jaccard and Mean Measure of Divergence to compute overall similarity.

The cosine similarity measure is a popular similarity measure. It treats each user as a vector in the space of items and uses the cosine between these vectors as a measure of similarity. However, it doesn't compute the individual's personal habits to express its preference. Hence, we propose to take into account the personal habits of a user. To achieve this Mean measure of divergence (MMD) is used. MMD[12] is the most common and most popular distance measures used for the computation of bio-distances based on no-metric traits.

Each individual has personal habits to express his preferences. Some user tends to give high or low ratings as compared to others. This biasness in ratings influences the relationship between the users. The traditional similarity neglects this factor. Mean measure of divergence is used to calculate likenesses in users ratings based upon rating habits. In this way, the θ_u represents a vector based on user u ratings, $|I_u|$ represents a total number of ratings made

by u and r represents number of co-rated item between u and u'. For example, let ratings vector of u and u' are $I_u = (1, 5, 2, 0, 1, 5, 2, 0, 3, 0, 1, 4, 4)$ and $I_{u'} = (2, 4, 2, 1, 0, 0, 5, 2, 2, 0, 0, 5, 5)$ respectively, where $m = 1$ and $M = 5$ (that is to say that ratings lie in $\{1, \ldots, 5\}$). Then θ_U can be represents as $\theta_u = (\theta^{\#m}, \ldots, \theta^{\#M})$ where each component $\theta^{\#m}$ represents the number of times the rating value m occurs in rating vector (I_u) of user u. Therefore θ_u and $\theta_{u'}$ can be computed as $\theta_u = (3, 2, 1, 2, 2)$ and $\theta_{u'} = (1, 4, 0, 1, 3)$. Therefore, *MMD* between a pair of users can be computed as Eqs. 1:

$$sim(u, u')^{MMD} = \frac{1}{1 + \left(\frac{1}{r} \sum_{i=1}^{r} \left\{ (\theta_u - \theta_{u'})^2 - \frac{1}{|I_u|} - \frac{1}{|I_{u'}|} \right\} \right)} \tag{1}$$

To give more significance to the co-rated items, we have added Jaccard similarity in our model. The Jaccard similarity between a two users is defined as Eqs. 2:

$$sim(u, u')^{Jaccard} = \frac{|I_u| \cap |I_{u'}|}{|I_u| \cup |I_{u'}|} \tag{2}$$

where $|I_u|$is the cardinality of items rated by user u.

Taking cosine, jaccard and Mean measure of divergence based similarity into account we combine these similarities to calculate the final similarity between users. The hybrid user similarity between two users is given by Eqs. 3:

$$sim(u, u')^{CJacMD} = sim(u, u')^{COS} + sim(u, u')^{Jaccard} + sim(u, u')^{MMD} \tag{3}$$

After the similarities between users have been calculated the next step is to find the top K users with the highest similarity to active user. These users are called neighbour of an active user. In order to predict the ratings of the active user for the unrated items, the following function (Eqs. 4) is used:

$$Pred_{u,i} = \overline{r_u} + \frac{\sum_{u' \in N(u)} sim(u, u') \times (r_{u',i} - \overline{r_{u'}})}{\sum_{u' \in N(u)} sim(u, u')} \tag{4}$$

where, $sim(u, u')$ indicates the similarity between two users, $\bar{r}_u$ is the mean rating of user u, $r_{u,i}$ is the rating given by user u to item i and $N(u)$ is the neighbors of user u.

The Algorithm 1 sketches the general scheme of proposed approach that is used make prediction for unknown ratings in user-item rating matrix.

1. Split dataset $R_{m \times n}$ into disjoint training $\{Tr_{m \times n}\}$ and testing $\{Tr_{m \times n}\}$ sets
2. **for** all users u **in** $Tr_{m \times n}$ **do**
3. $sim(u, u')^{Jaccard}$ ← Compute the *Jaccard* similarities between user u and u' i.e target user and rest of the user respectively.
4. $sim(u, u')^{COS}$ ← Compute the *COS* similarities between u and u' .
5. $sim(u, u')^{MMD}$ ← Compute the *MMD* similarities between u and u' .
6. $sim(u, u')^{CJacMD} \leftarrow sim(u, u')^{COS} + sim(u, u')^{Jaccard} + sim(u, u')^{MMD}$
7. $K \leftarrow u' \in N(u)$ % find the nearest neighbors of the active user
8. $Pred_{u,i}$ ← Compute UserBased Prediction
9. **end for**
10. *Evaluate RMSE and MAE of predicted ratings for* Te.

Algorithm 1. *CjacMD* Algorithm

4. Experiments

4.1 Data sets

To evaluate the RS approach, a large number of benchmark datasets are publically available. We evaluate the proposed model with three well known real-life datasets (i) MovieLens 100 K (ML-100 K). The ML-100K dataset, consists of 100,000 ratings on 1682 items (movies) made by 943 users. In this dataset, the sparsity rating is 93.69%, which means that 4.25% of the movie has been rated by users. (ii) MovieLens 1 M (ML-1 M), consists of approximately one million ratings from 6,040 users who reviewed 3,952 movies. The sparsity rating of ML-1 M is about 95.75%. In both the MovieLens dataset, each movie is rated on a scale from 1 to 5 and (iii) the each movie dataset where we extracted a subset of 1,004 users who reviewed 1,091 movies with rating scale from 1 to 6.

4.2 Evaluation measures

The effectiveness and accuracy of rating predictions of our proposed similarity measure and other traditional similarity measures was evaluated using most popular evaluation metrics namely Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

The Mean Absolute Error (MAE) is defined as Eqs. 5:

$$MAE = \frac{1}{|S|}\sum_{i=1}^{S}|Pred_i - r_i| \tag{5}$$

where $Pred_i$ is the predicted rating for a movie, r_i the actual rating, and $|S|$ is the cardinality of the test ratings,

The RMSE is defined as Eqs. 6:

$$RMSE = \sqrt{\frac{1}{|S|}\sum_{i=1}^{S}(Pred_i - r_i)^2} \tag{6}$$

4.3 Experimental result and analysis

Assume K denotes the number of nearest neighbours. In CF, the performance of recommendation can be greatly affected by the number of nearest neighbours (K). In order to examine the sensitivity of the neighbourhood size, we perform our experiment with different number of nearest neighbours. It is observed that the different number of neighbours (K) will give different predictions and accuracy. To prove the effectiveness of the proposed CJacMMD method, it is compared with four different similarity measures i.e. COS, PCC, ACOS and SRC.

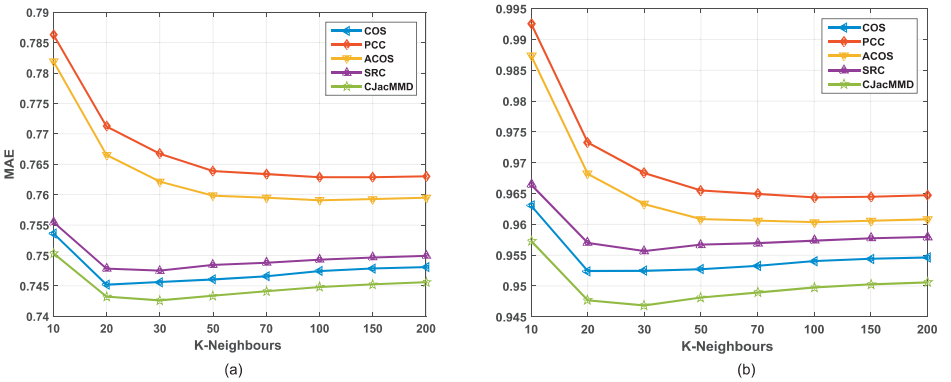

Fig. 1. Performance Comparison of Different Approaches on ML-100 K Dataset: (a) MAE; and (b) RMSE.

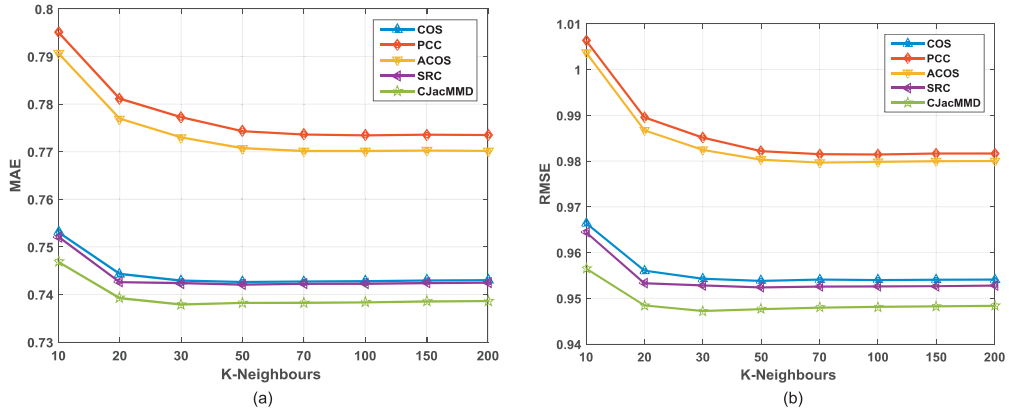

Fig. 2. Performance Comparison of Different Approaches on ML1M Dataset: (a) MAE; and (b) RMSE.

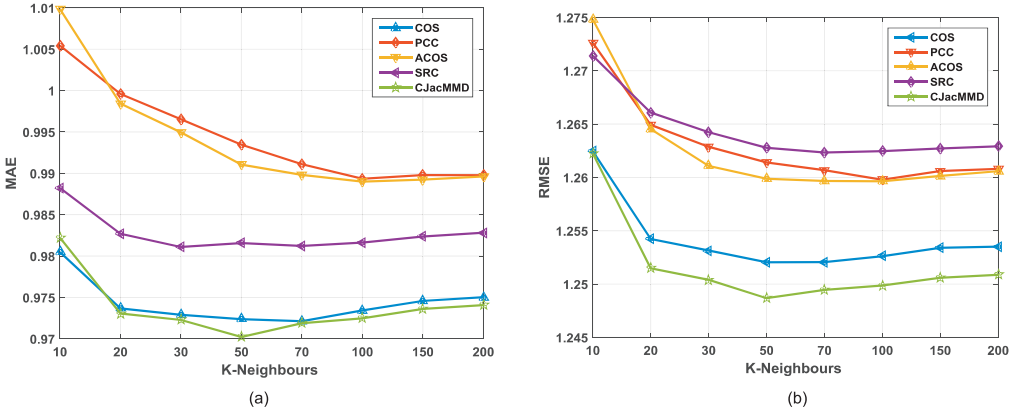

Fig. 3. Performance Comparison of Different Approaches on EachMovie Dataset: (a) MAE; and (b) RMSE.

Figure 1 illustrates the performance of different similarity method with varying the number of nearest neighbours on the ML 100 k dataset. The evaluated MAE and RMSE of all the similarity measures decrease as the number of nearest neighbors increases.

Figure 2 elaborate the same conclusion hold using ML-1 M dataset except change in the performance of different similarity measure.

Figure 3 shows the performance comparison of existing similarity measures on EachMovie dataset. CJacMMD outperforms the other approaches consistently, regardless of neighbourhood size. It is observed that the range of MAE and RMSE values in all the three dataset decreases as the value of K increases. That is because some similarity measures are sensitive to false neighbours, which results in the high similarity between two users, but in fact, their preference is not similar this often lead by the data sparsity.

5. Conclusions

CF has become an attractive subject for researchers because its capability to handle information overload efficiently. It provides personalized and automated suggestions to users based on their taste and those of similar interest. In order to overcome the problem of data sparsity, we proposed a new similarity measure based on Mean measure of divergence that considers the behaviour habit effects of rated items to measure the similarity between users. The detailed analysis

on three benchmark datasets indicates that the results of proposed approach are comparable with the traditional approach. The proposed technique provides better prediction accuracy. We plan to extend the proposed algorithm in a real-world recommender system.

References

[1] Ahn, Hyung, A New Similarity Measure for Collaborative Filtering to Alleviate the New User Cold-Starting Problem, *Information Sciences*, vol. 178(1), pp. 37–51, June (2008).

[2] J. Bobadilla, F. Ortega, A. Hernando and A. Gutiérrez, Recommender Systems Survey, *Knowledge-Based Systems*, vol. 46, pp. 109–32, (2013).

[3] J. Bobadilla, F. Serradilla and J. Bernal, A New Collaborative Filtering Metric That Improves the Behavior of Recommender Systems, *Knowledge-Based Systems*, vol. 23(6), pp. 520–28, (2010).

[4] Bobadilla, Jesus, Fernando Ortega, Antonio Hernando and Javier Alcalá, Improving Collaborative Filtering Recommender System Results and Performance using Genetic Algorithms, *Knowledge-Based Systems*, vol. 24(8), pp. 1310–16, (2011).

[5] Bobadilla, Jesús, Fernando Ortega, Antonio Hernando and Jesús Bernal, A Collaborative Filtering Approach to Mitigate the New User Cold Start Problem, *Knowledge-Based Systems*, vol. 26, pp. 225–38, (2012).

[6] Celma and Oscar, Music Recommendation and Discovery, (2010).

[7] Chen, Meng Hui, Chin Hung Teng and Pei Chann Chang, Applying Artificial Immune Systems to Collaborative Filtering for Movie Recommendation, *Advanced Engineering Informatics*, (2014).

[8] Choi, Keunho and Yongmoo Suh, A New Similarity Function for Selecting Neighbors for Each Target Item in Collaborative Filtering, *Knowledge-Based Systems*, vol. 37, pp. 146–53, (2013).

[9] M. D. Ekstrand, J. T. Riedl and J. A. Konstan, Collaborative Filtering Recommender Systems, *Foundations and Trends®in Human-Computer Interaction*, vol. 4(2), pp. 175–243, (2011).

[10] Gan, Mingxin and Rui Jiang, Improving Accuracy and Diversity of Personalized Recommendation through Power Law Adjustments of User Similarities, *Decision Support Systems*, vol. 55(3), pp. 811–21, (2013).

[11] Goldberg, David, David Nichols, Brian M. Oki and Douglas Terry, Using Collaborative Filtering to Weave an Information Tapestry, *Communications of the ACM*, vol. 35(12), pp. 61–70, (1992).

[12] Irish and D. Joel, The Mean Measure of Divergence: Its Utility in Model-Free and Model-Bound Analyses Relative to the Mahalanobis D2 Distance for Nonmetric Traits, *American Journal of Human Biology*, vol. 22(3), pp. 378–95, (2010).

[13] Jeong, Buhwan, Jaewook Lee and Hyunbo Cho, Improving Memory-Based Collaborative Filtering via Similarity Updating and Prediction Modulation, *Information Sciences*, vol. 180(5), pp. 602–12, (2010).

[14] Koutrika, Georgia, Benjamin Bercovitz and Hector Garcia-Molina, FlexRecs: Expressing and Combining Flexible Recommendations, *Proceedings of the 35th SIGMOD International Conference on Management of Data*, pp. 745–58, (2009).

[15] Leng, Yajun, Qing Lu and Changyong Liang, A Collaborative Filtering Similarity Measure Based on Potential Field, *Kybernetes*, vol. 45(3), pp. 434–45, (2016).

[16] Liu, Haifeng, Zheng Hu, Ahmad Mian, Hui Tian and Xuzhen Zhu, A New User Similarity Model to Improve the Accuracy of Collaborative Filtering, *Knowledge-Based Systems*, vol. 56, pp. 156–66, (2014).

[17] Luo, Heng, Changyong Niu, Ruimin Shen and Carsten Ullrich, A Collaborative Filtering Framework Based on Both Local User Similarity and Global User Similarity, *Machine Learning*, vol. 72(3), pp. 231–45, (2008).

[18] Nikita and Efthymia, A Critical Review of the Mean Measure of Divergence and Mahalanobis Distances using Artificial Data and New Approaches to the Estimation of Biodistances Employing Nonmetric Traits, *American Journal of Physical Anthropology*, vol. 157(2), pp. 284–94, (2015).

[19] Park, Yoon-Joo and Alexander Tuzhilin, The Long Tail of Recommender Systems and How to Leverage It, *Proceedings of the 2008 ACM Conference on Recommender Systems RecSys 08*. vol. 11, (2008).

[20] Patra, Bidyut Kr, Raimo Launonen, Ville Ollikainen and Sukumar Nandi, A New Similarity Measure using Bhattacharyya Coefficient for Collaborative Filtering in Sparse Data, *Knowledge-Based Systems*, vol. 82, pp. 163–77, (2015).

[21] Pirasteh, Parivash, Dosam Hwang and Jai E. Jung, Weighted Similarity Schemes for High Scalability in User-Based Collaborative Filtering, *Mobile Networks and Applications*, vol. 20(4), pp. 497–507, (2014).