2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

A collaborative filtering framework based on local and global similarities with similarity tie-breaking criteria

André R. S. Lopes Ricardo B. C. Prudêncio Byron L. D. Bezerra

***Abstract*—Collaborative Filtering is the most commonly used technique in Recommender Systems, based on the users ratings in order to identify similar profiles and suggest them items. However, because it depends essentially on direct similarity measures between users or items, it usually suffers from the sparsity problem. Upon this situation, a good alternative is using global similarities to enrich the users neighborhood by transitively connecting them together, even when they do not share any common ratings. In this paper, we investigated the use of both local and global similarity measures with the maximin distance algorithm, along with tie-breaking criteria for neighbors with equal similarity. Our experiments showed that the maximin distance algorithm in fact produces many equally similar global neighbors, and that the criteria set for deciding between them severely improved the results of the recommendation process.**

I. Introduction

THE volume of information is growing much faster than the human capacity to process it. As a consequence, we seem to live in an information overload that creates a real difficulty in acquiring useful knowledge and making decisions. Upon this scenario, we naturally tend to search for recommendations and advice in order to help the decision-making process. Recommender systems are a promising technology that simply automates the process of learning from the collective experience in order to discover the most valuable content for us [1]. Due to its great potential, this kind of system can be seen almost everywhere. The e-commerce world, for instance, is making good use of recommender systems to decrease the gap between company seller and costumer.

Due to its efficiency and simplicity, collaborative filtering techniques are the most used ones for the development of recommender systems [1]. In this approach, items are recommended to a target-user based on the preferences of similar users. Similarity between two users in this context is commonly defined according to ratings given to items they evaluated in common. Despite being widely used and studied, collaborative filtering is very sensitive to the problem of data sparsity [2]. Since it works based on similar preferences, a low number of common ratings makes similarity between users rather useless.

In recent years, different solutions have been proposed to deal with the sparsity problem [3]. The current paper will focus on the use of global similarity measures originally proposed by Luo [4] and Anand [5]. In contrast to the traditional similarity measures based on common ratings (named as *local similarity*), global similarity measures identify transitive relationships between users. Two users are globally similar if they can be connected by a transitive chain of locally similar users. The experiments performed in previous work have revealed good results when global and local measures were combined. However, due to the nature of the global measures proposed in previous work, a very high number of neighbors with equal similarity value can be produced. Hence, in many cases the global similarity measures have a lower capacity to discriminate users, which can harm the recommendation process. In our work, we investigated open issues in the combination of global and local measures by proposing tie-breaking criteria to favor the best neighbors in these situations. In the experiments performed on different datasets, a gain in accuracy was observed in the recommendation process when the proposed criteria were adopted.

This paper is organized as follows: Section II introduces a brief background and the related work. In Section III, we discuss the proposed similarity tie-breaking criteria and the collaborative filtering framework. Section IV presents the performed experiments and obtained results compared with the ones achieved in [5]. Finally, the conclusions and future research directions are pointed out in Section V.

II. Background and related work

In spite of all its popularity, collaborative filtering still has difficulties in dealing with data sparsity, which as a significant weakness. In this section, we discuss different strategies of the literature to treat this issue, and then present the use of global and local similarity measures, which is main focus of the present work.

A. Data Sparsity

Data sparsity is characterized by the limited amount of ratings and, consequently, neighbors of a target-user considering a database of items to recommend. As real-world systems usually have a really big number of users and items, even the most active users consume just a few items of the whole database [6]. Similarly, even the most popular items are rated by just a small portion of the users in the system. Hence, a lot of real-world recommender systems based on collaborative filtering are challenged by the sparsity problem.

In order to overcome data sparsity, some previous work have tried to effectively reduce the dimension of the user-item data matrix in order to increase the likelihood that different users rate common items [7]. The main representative techniques of this approach are the Singular Value

André Lopes and Ricardo Prudêncio are with the Center of Informatics, Federal University of Pernambuco, Cidade Universitária - CEP 50740-560 - Recife (PE) - Brazil (email:{arsl, rbcp}@cin.ufpe.br and Byron Bezerra is with Polytechnic School of Pernambuco, University of Pernambuco - CEP 50100-010 - Recife (PE) - Brazil (email:byronleite@ecomp.poli.br).

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

Decomposition (SVD) [8], the Latent Semantic Indexing (LSI) [9] and the Principal Component Analysis (PCA) [10]. Although such techniques can be successfully employed to improve the accuracy of collaborative filtering systems facing the sparsity problem, the process of eliminating users and items may cause the lost of potential valuable information.

Another alternative that has been well explored is to make use of a hybrid approach, combining collaborative filtering and content-based filtering techniques [11] [1]. In such combination, the hybrid recommender system tries to benefit from the strengths of each technique, overcoming their weaknesses [12].

Prediction propagation strategies have also been a good alternative to to deal with data sparsity [13] [14] by adopting an iterative data completion procedure. In this approach, strategic missing values in the user-item rating matrix are predicted and assumed to be known. Following, such values are propagated in order to iteratively produce a dense matrix.

Trust is another interesting concept that addresses which neighbors participate in the recommendation process [15] [16], and thus can help to treat the data sparsity problem. This approach is adopted to complement the use of common similarity measures, and is able to increase the target-user neighborhood set by enabling the contribution of users that do not share any common ratings with him or her, but are trustful instead [17] [18].

B. Local Similarities

The similarity measurement between users is a crucial step for collaborative filtering recommender systems. In order to compute what we call the local similarity of two users, the presence of common ratings is important. The main similarity metrics in the literature are the Pearson Correlation Coefficient (PCC) and the Vector Cosine Similarity (VS). The former aims to measure how much two users vary together from their common evaluations, that is, the difference from each of their ratings to their average rating, which is given by the formula:

$$sim_{x,y} = \frac{\sum_{i \in S_{xy}} (r_{x,i} - \bar{r}_x)(r_{y,i} - \bar{r}_y)}{\sqrt{\sum_{i \in S_{xy}} (r_{x,i} - \bar{r}_x)^2 \sum_{i \in S_{xy}} (r_{y,i} - \bar{r}_y)^2}} \quad (1)$$

where $sim_{x,y}$ is the similarity between users x and y, $r_{x,i}$ and $r_{y,i}$ correspond to their ratings for the item i, $\bar{r}_x$ and $\bar{r}_y$ are their respective average rating and S_{xy} denotes the common ratings set between these users.

Other popular similarity metric, the VS, treats each user as a vector with dimensionality equal to the number of items in the database, and values corresponding to their ratings. The similarity of two users is then given by the cosine of the angle between their vectors. The idea behind this is that vectors with same direction reveal a high similarity, and is given by the formula:

$$sim_{x,y} = \frac{\sum_{i \in S_{xy}} (r_{x,i})(r_{y,i})}{\sqrt{\sum_{i \in S_{xy}} (r_{x,i})^2 \sum_{i \in S_{xy}} (r_{y,i})^2}} \quad (2)$$

In Luo [4], we are presented a novel similarity metric based on the principle that if two users rated an item similarly, but very differently from the other users, we then have a strong evidence of similarity that should be taken into consideration. Thus, to find the similarity of two users the authors first measured what they called *surprisal*, a measure that provides information about how far a rating was from the database average for it. The rating of each item was modeled as a Laplacian random variable and its surprisal was computed as:

$$I(r_{x,i}) = -ln(f(r = r_{x,i} | \hat{\mu_i}, \hat{b_i})) = ln(2\hat{b_i}) + \frac{|r_{x,i} - \hat{u_i}|}{\hat{b_i}} \quad (3)$$

where $r_{x,i}$ is the rating of the item i given by the user x, $\hat{\mu i}$ and $\hat{bi}$ are the maximum likelihood estimative for the location and scale parameters, in that order. Instead of using the original ratings, the authors adopted the surprisals with direction given by the function:

$$S_{x,i} = sgn(r_{x,i} - \hat{\mu_i}) I(r_{x,i}) \quad (4)$$

where sgn is the signal function that represents whether the user x rated the item i higher or lower in comparison to $\hat{\mu_i}$, the average rating for that item. The similarity between the two users was achieved using the VS formula with the surprisal vectors.

As one can notice, for each similarity measure discussed, only the set of items commonly evaluated was taken into consideration. Therefore, it could be possible that some users would be considered perfectly similar even if they shared just one single common rating. As this would be bad for the recommendation process, Luo [4] applied Mas proposal [19] to penalize the similarity weights that were based on a small number of co-rated items. His final formula received the name of Surprisal-Based Vector Similarity with Significance Weighting (SVSS):

$$sim'_{x,y} = \frac{min(|Iu_x \cap Iu_y|, \gamma)}{\gamma} sim_{x,y} \quad (5)$$

where $|Iu_x \cap Iu_y|$ is the number of co-rated items between the users x and y, and γ is the threshold for the significance weighting.

C. Global Similarities

When the database is sparse, local similarities just do not produce good user neighbor relationships, thus affecting the quality of the recommendation process [5]. In these cases, global similarities can be used to enrich the users neighborhood, since they define two users as being similar whether it is possible to connect them through their local neighbors [4].

Following this principle, some proposals can be found in the literature. Fouss [20] introduced a random walk on a bipartite graph of both users and items. By connecting the users to the items they have experienced, the values of average commute time and average first passage time, achieved by the random walk were employed as similarity measures for two users. Average commute time is the mean number of steps a random walker takes to go back and forth

from one user node to another. Average first passage time, in turn, represents the mean number of steps necessary to reach the other user node for the first time. The authors demonstrated that these similarity measures tend to grow when the number of paths connecting two user nodes grows, and also when the distance of the paths decreases [20]. Nevertheless, this kind of recommendation process lacks easily understandable explanations for the final users of the system.

Desrosiers [21] presented an approach that computes global similarities both between the pairs of users and the pairs of items, based on the solution of a system of linear equations that relates the user similarities to the item similarities. This strategy helps the contention of the data sparsity problem, since it allows calculating the similarities of users that do not have any common ratings but whose items experienced have some similarity between each other.

D. Combining Local and Global Similarities

Luo [4] proposed a framework that combines both local and global similarities in the recommendation process. For that, the users were initially represented in a weighted graph, in which each node corresponds to a user and each edge weight between two connected nodes corresponds to the local similarity. The similarity measure applied to define the edge weights was the SVSS discussed in Section II. After the graph construction, the global similarity between each pair of users could then be calculated through indirect associations using the maximin distance algorithm:

1) Find all paths that connect two users;
2) For each path, identify its minhop, that is, the edge with the lowest weight;
3) Among all path minhops, identify the one with the highest value. This represents the maximin distance between the two users, and is used as their proper global similarity.

It is worth pointing out that despite the word "distance" in the maximin distance algorithm, the higher the result achieved the better. That is because the graph edges actually represent the users local similarities. That way, two users can become globally more similar because they can be connected through some locally more similar neighbors [4].

In order to compute all maximin distance pairs, the Floyd-Warshall algorithm can be adopted with an algorithm complexity of $O(N^3)$ [22]. In practice, an efficient algorithm based on message passing can query the global similarity of a specific user and the rest with a time complexity of $O(N^2)$ [23].

The rating prediction of a given item for the target-user was then expressed as a linear combination of the predictions obtained by individually adopting the local and the global neighborhood:

$$predR = (1 - \alpha) * predR_L + \alpha * predR_G \qquad (6)$$

where $predR_L$ is the prediction obtained with the target-user local neighborhood (i.e., the set of neighbors identified according to the local similarities), $predR_G$ is the analogous prediction using the global neighborhood (i.e., the set of neighbors identified using only the global similarities) and α is a significance weight to balance local and global predictions.

In [4], the authors demonstrated that for sparse databases, the linear combination in favor of global similarities achieves better results (i.e., greater values of α), while the opposite is true for dense databases. Fixing the significance weight at 0.5, Luo [4] observed the superiority of their technique compared to different prediction algorithms proposed in the literature, such as user-based collaborative filtering using PCC (UPCC), item-based collaborative filtering using PCC (IPCC), Effective Missing Data Prediction (EMDP) and Similarity Fusion Algorithm (SF).

Anand [5] extended Luo's work [4] by proposing the use of sparsity measures to automatically define the significance weight α. The idea behind their work was to measure the sparsity level of the input databases and use this information to favor the local or the global similarities properly. More specifically, the higher is the level of sparsity, the greater were the suggested values of α. Five different sparsity measures were introduced, taking into consideration not only the database as a whole, but also the target-user and item, aiming to capture and benefit from the various sparsity aspects of the database. The authors [5] also proposed a framework to combine all these sparsity measures by using genetic algorithms, which they called as Unified Measure of Sparsity (UMS). Their experiments showed that the individual use of any of the proposed sparsity measures improved the system accuracy when compared to Luo's [4] fixed-α scheme, and also that the UMS obtained the best results from all techniques investigated.

III. Proposed recommender system framework

Although Luo [4] and Anand [5] achieved outstanding results using global similarity, we identified a good opportunity of improvement in the proposed framework regarding the application of the maximin distance algorithm. The algorithm is capable of indirectly associating local neighbors in order to deal with the data sparsity problem, however, as a side-effect not yet reported, a very high percentage of the global neighbors have the same similarity value.

Figure 1 illustrates this situation, where there are not any similarity coincidence in the target-user using local similarity, but after measuring the global similarities with the maximin distance algorithm, three users (number 2, 4 and 5) are considered equally similar to the target-user. This happens mainly because the maximin distance algorithm can only reproduce an existing local similarity value, and because it just uses the best paths that connect the users, which tend to vary little.

This consequence is bad in the sense that, depending on the neighborhood size, possible best neighbors could end up out of the target-user global neighborhood, losing place to others considered equally similar by the maximin distance algorithm, but who may happen to be worse for the recommendation process.

As stated before, in [4] [5] did not discuss such issue deeply and no criterion for decision-making in such cases was suggested, even though the neighborhood formation is

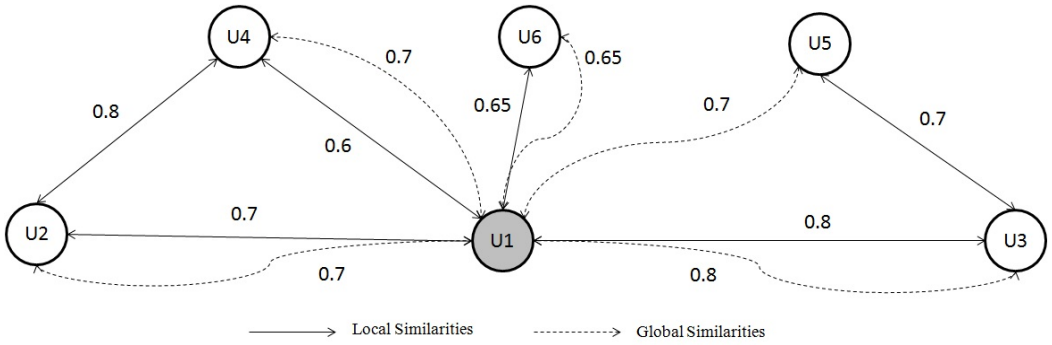

Fig. 1. Local and Global Similarities with the maximin distance algorithm.

one of the most important steps of collaborative filtering recommender systems. Therefore, it is extremely important to have criteria for tie-breaking the global users who will form the target-user neighborhood.

This work proposes a set of rules for that matter. Given a target-user, when facing neighbors' similarity ties, we will then opt for a certain neighbor among the others when:

1) His local similarity is higher than all others. Neighbors without a local similarity with the target-user are automatically deferred in favor of the ones that do. The motivation of this rule is that local similarities (when available) represent direct information of two users and then are more reliable than indirect and inferred associations;
2) In the case of a draw at the local similarity comparison, we choose the one with the highest number of co-ratings with the target-user. If there is still a draw, then we pick the one that has the higher Jaccard Coefficient [24] with the target-user, that is, the result of the intersection of their ratings divided by their union. The explanation for this rule is simple: explicit ratings might not always reflect the real user preferences, since people do not know themselves very well and happen to give different ratings for different items motivated by unclear reasons, or even evaluate the same item differently at distinct moments. The voluntary choice of consuming an item, on the other hand, reflects a true implicit preference;
3) In the scenario where neighbors don't have local similarities with the target-user, or where the other rules could not favor one single neighbor, we will compare what would be the next maximin distance, namely, the next highest minhop. This step is repeated until a single neighbor becomes chosen.

Revisiting Figure 1, if one has to choose between the neighbors 2, 4 and 5, the result of applying these criteria would be the tie-break in favor of the user number 2. That is because the user 5 does not have a local similarity with the target-user, and the user 2 presents a higher local similarity in comparison to the number 4. It is important to clarify that these criteria do not exclude any neighbor candidate, but just help rank them for the neighborhood formation process. Thus, in a scenario where there was room for only two of these users, users number 2 and 4 would make it.

These comparative steps should be enough to prioritize the best candidate to participate in the global neighborhood set of the target-user. It is possible, though extremely unlikely, that we still do not get to decide between the neighbors by the end of these comparisons. For real-world recommender systems, then, other interesting criteria can be added to the proposed ones, like opting for the neighbor that most actively participates in the system, the oldest user, or the one that has more connections to other users.

IV. Experiments

In order to verify the presence of global neighbors with same similarity value and the improvement in the accuracy of the recommender system framework achieved by the application of the proposed tie-breaking criteria, several experiments were conducted on the vastly popular MovieLens [25] and Jester [26] databases.

A. Database

The databases chosen for the experiments were the same ones adopted in Anand's work [5]: MovieLens, a movie recommendation dataset, and Jester, a dataset composed of judgments of jokes. These databases are good choices for our purpose, not only because they have been adopted in many previous works, but also because they differ a lot from each other. This is convenient to the analysis of the proposed techniques under diverse data environments.

The MovieLens dataset comprises 100.000 item ratings provided by 943 users over 1682 different movies. Each user evaluated at least 20 items, with ratings ranging in the discrete interval [1,5]. In turn, Jester consists of 4.1 million ratings by 73.421 users concerning 100 jokes. The ratings are continuous and lie in the range -10 to 10. In terms of data sparsity, the two databases differ a lot from each other, since the overall sparsity of MovieLens is high whereas the Jester dataset is very dense.

B. Experimental Setup

In the experiments, both datasets received the same treatment as in Anand [5]: 300 users from each dataset were randomly chosen and then divided into a training group consisting of 200 users, and a test group with the remaining 100. For each user on the test group, the Given-X protocol [4] was applied for X equal to 5, 10 and 20.

In other words, their number of available evaluations were randomly reduced to 5, 10 and 20, thus originating 6 different configurations: ML300G5, ML300G10, ML300G20, Jester300G5, Jester300G10 and Jester300G20. The removed items were then kept as test items, that is, items whose ratings were not available to the recommender system and should be predicted. Finally, the ratings of the Jester dataset were discretized by rounding the rating value to the nearest integer.

The recommender system framework was the one proposed at Anand [5], using both local and global similarities and weighting their significance with the UMS achieved with genetic algorithms. The local similarity metric was always the SVSS presented at Luo [4]. The γ threshold parameter that penalizes the similarities based on a small number of co-rated items was set to 30, and the k corresponding to the number of nearest neighbors used for prediction was also kept at 30. We chose the same settings adopted in Anand [5], for a better comparison of the results. The MAE and RMSE [3] were used as evaluation measures for the quality of the recommendations. Each experiment was repeated 30 times and the average results were collected to perform parametric hypothesis testing.

C. Experiment 1

Our first experiment aimed to analyse the ratio of similarity ties in the target-user global neighborhood produced by the maximin distance algorithm. Therefore, we counted the number of target-user neighbor candidates available for each prediction, that is, users that have a similarity with the target-user, as well as how many similarity ties were present, for both the local and the global neighborhoods. The ratio of similarity ties was then the quotient of the latter by the former. The results can be seen at the Tables I, II, III and IV.

TABLE I. NUMBER OF NEIGHBORS AND SIMILARITY TIES OF THE LOCAL NEIGHBORHOOD ON MOVIELENS

	ML300G5	**ML300G10**	**ML300G20**
Number of neighbors	*Mean (std)*	*Mean (std)*	*Mean (std)*
	26.48 (2.40)	31.61 (1.66)	35.10 (2.36)
Similarities tied	*Mean (std)*	*Mean (std)*	*Mean (std)*
	13.08 (0.85)	7.35 (0.45)	4.22 (0.22)
Ratio of ties	*Mean*	*Mean*	*Mean*
	0.49	0.23	0.12

TABLE II. NUMBER OF NEIGHBORS AND SIMILARITY TIES OF THE GLOBAL NEIGHBORHOOD ON MOVIELENS

	ML300G5	**ML300G10**	**ML300G20**
Number of neighbors	*Mean (std)*	*Mean (std)*	*Mean (std)*
	36.21 (2.44)	37.21 (1.66)	38.35 (2.31)
Similarities tied	*Mean (std)*	*Mean (std)*	*Mean (std)*
	35.35 (2.44)	35.72 (1.64)	35.50 (2.24)
Ratio of ties	*Mean*	*Mean*	*Mean*
	0.97	0.96	0.92

When we compare the number of neighbors with similarities tied to the total number of neighbors available for the predictions in both local and global neighborhood sets, we notice that the application of the maximin distance algorithm resulted in too many global neighbors with coincident similarities. Actually, sometimes the number of ties almost reached the whole neighborhood set. Thus, it is clear that the recommendation system can benefit a lot from criteria for tie-breaking decisions on the global neighborhood.

TABLE III. NUMBER OF NEIGHBORS AND SIMILARITY TIES OF THE LOCAL NEIGHBORHOOD ON JESTER

	Jester300G5	**Jester300G10**	**Jester300G20**
Number of neighbors	*Mean (std)*	*Mean (std)*	*Mean (std)*
	131.56 (4.36)	138.46 (4.91)	151.65 (5.21)
Similarities tied	*Mean (std)*	*Mean (std)*	*Mean (std)*
	11.48 (1.14)	7.62 (0.53)	1.21 (0.29)
Ratio of ties	*Mean*	*Mean*	*Mean*
	0.08	0.05	0.01

TABLE IV. NUMBER OF NEIGHBORS AND SIMILARITY TIES OF THE GLOBAL NEIGHBORHOOD ON JESTER

	Jester300G5	**Jester300G10**	**Jester300G20**
Number of neighbors	*Mean (std)*	*Mean (std)*	*Mean (std)*
	139.44 (3.97)	142.45 (4.74)	151.79 (5.21)
Similarities tied	*Mean (std)*	*Mean (std)*	*Mean (std)*
	119.67 (3.34)	104.41 (3.27)	83.95 (3.89)
Ratio of ties	*Mean*	*Mean*	*Mean*
	0.86	0.73	0.55

D. Experiment 2

Our next experiment has the purpose of evaluation if the proposed criteria for deciding among equally similar global neighbors positively impact the accuracy of the system. We then evaluated and compared Anands framework [4] with and without the tie-breaking criteria application. The results are presented in the Tables V, VI, VII and VIII.

TABLE V. EVALUATION OF THE SYSTEM (MAE) WITH AND WITHOUT THE CRITERIA ON MOVIELENS

	ML300G5	**ML300G10**	**ML300G20**
Framework	*MAE*	*MAE*	*MAE*
Without criteria	0.8771	0.8182	0.7503
With criteria	**0.8362**	**0.7901**	**0.7312**

TABLE VI. EVALUATION OF THE SYSTEM (RMSE) WITH AND WITHOUT THE CRITERIA ON MOVIELENS

	ML300G5	**ML300G10**	**ML300G20**
Framework	*RMSE*	*RMSE*	*RMSE*
Without criteria	1.1644	1.0995	1.001
With criteria	**1.1201**	**1.0661**	**0.978**

TABLE VII. EVALUATION OF THE SYSTEM (MAE) WITH AND WITHOUT THE CRITERIA ON JESTER

	Jester300G5	**Jester300G10**	**Jester300G20**
Framework	*MAE*	*MAE*	*MAE*
Without criteria	3.7989	3.5819	**3.1511**
With criteria	**3.7916**	**3.5800**	3.1532

As shown, the criteria application significantly improved the accuracy of the recommender system, although it did not make much difference for the Jester dataset. We can also

TABLE VIII. EVALUATION OF THE SYSTEM (RMSE) WITH AND WITHOUT THE CRITERIA ON JESTER

	Jester300G5	**Jester300G10**	**Jester300G20**
Framework	*RMSE*	*RMSE*	*RMSE*
Without criteria	4.6473	**4.3773**	**3.8057**
With criteria	**4.6413**	4.3778	3.8089

notice that the sparser is the data, the better the accuracy enhancement, and this is probably the reason why it didnt impact the system for the Jester dataset that much.

In order to prove that the application of the proposed criteria for similarity tie-breaking decision indeed yields better results, we applied a parallel hypothesis test on the evaluation of the system with a significance level of 0.05. The null and alternative hypotheses were as follow:

- H0: mean of the error without criteria - mean of the error with criteria ≤ 0;
- H1: mean of the error without criteria - mean of the error with criteria > 0.

As we intend to reject the null hypothesis in favor of the alternative one, it asks for a right unilateral test. As dictates the z-Table [27] then, values bigger than 1.64 will serve our purpose. The test results are summarized in Table IX.

TABLE IX. HYPHOTESIS TEST FOR THE EVALUATION OF THE SYSTEM USING THE SIMILARITY TIE-BREAKING CRITERIA

Dataset	**z-value**	
	MAE	*RMSE*
ML300G5	**16.20**	**14.38**
ML300G10	**13.89**	**15.33**
ML300G20	**7.21**	**8.05**
Jester300G5	**2.95**	**2.37**
Jester300G10	0.97	-0.25
Jester300G20	-1.25	-1.99

Except for Jester300G10 and Jester300G20, all results provided strong evidences to reject our null hypothesis. Therefore, it is safe to say that in most cases the system performs better when our similarity tie-breaking criteria are applied.

V. CONCLUSION

In the present work we discussed how using the maximin distance algorithm for the global similarities measurement can generate a lot of coincident values and the problems that may arise. As a solution, we introduced a set of rules to prioritize the best neighbors laying in this situation. Our experiment results proved: (1) a big percentage of the global neighborhood end up with the same similarity value; (2) the accuracy of the recommender system based on both local and global similarities indeed improved by applying the similarity tie-breaking criteria, in most cases.

In the future, we plan to investigate other local-global prediction weighting measures than the ones existing in the literature.

REFERENCES

[1] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds., *Recommender Systems Handbook*. Springer US, 2011.

[2] J. Xu, X. Zheng, and W. Ding, "Personalized recommendation based on reviews and ratings alleviating the sparsity problem of collaborative filtering." in *ICEBE*, 2012, pp. 9–16.

[3] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Adv. Artificial Intellegence*, vol. 2009, 2009.

[4] H. Luo, C. Niu, R. Shen, and C. Ullrich, "A collaborative filtering framework based on both local user similarity and global user similarity," *Mach. Learn.*, vol. 72, no. 3, pp. 231–245, Sep. 2008.

[5] D. Anand and K. K. Bharadwaj, "Utilizing various sparsity measures for enhancing accuracy of collaborative recommender systems based on local and global similarities," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 5101–5109, May 2011.

[6] E. J. M. V. dos Santos, B. L. D. Bezerra, J. S. de Amorim, and A. I. do Nascimento, "A recommender system architecture for an inter-application environment." in *ISDA*, A. Abraham, A. Y. Zomaya, S. Ventura, R. Yager, V. Snsel, A. K. Muda, and P. Samuel, Eds. IEEE, 2012, pp. 472–477.

[7] M. Papagelis, D. Plexousakis, and T. Kutsuras, "Alleviating the sparsity problem of collaborative filtering using trust inferences," in *Proceedings of the Third international conference on Trust Management*, ser. iTrust'05, 2005, pp. 224–239.

[8] Y. Ren and S. Gong, "A collaborative filtering recommendation algorithm based on svd smoothing," in *Proceedings of the 2009 Third International Symposium on Intelligent Information Technology Application - Volume 02*, ser. IITA '09, 2009, pp. 530–532.

[9] T. Hofmann, "Latent semantic models for collaborative filtering," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 89–115, Jan. 2004.

[10] A. I. Markos, M. G. Vozalis, and K. G. Margaritis, "An optimal scaling approach to collaborative filtering using categorical principal component analysis and neighborhood formation," in *Artificial Intelligence Applications and Innovations - 6th IFIP WG 12.5 International Conference, AIAI 2010, Larnaca, Cyprus, October 6-7, 2010. Proceedings*, H. Papadopoulos, A. S. Andreou, and M. Bramer, Eds., pp. 22–29.

[11] R. Burke, "The adaptive web," P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds., 2007, ch. Hybrid web recommender systems, pp. 377–408.

[12] B. L. D. Bezerra and F. de Assis Tenrio de Carvalho, "Symbolic data analysis tools for recommendation systems." *Knowl. Inf. Syst.*, vol. 26, no. 3, pp. 385–418, 2011.

[13] J. Zhang and P. Pu, "A recursive prediction algorithm for collaborative filtering recommender systems," in *Proceedings of the 2007 ACM conference on Recommender systems*, ser. RecSys '07, 2007, pp. 57–64.

[14] B. Piccart, J. Struyf, and H. Blockeel, "Alleviating the sparsity problem in collaborative filtering by using an adapted distance and a graph-based method." in *SDM*. SIAM, 2010, pp. 189–198.

[15] F. Lorenzi, M. Abel, S. Loh, and A. Peres, "Enhancing the quality of recommendations through expert and trusted agents," in *Proceedings of the 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, ser. ICTAI '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 329–335.

[16] D. Anand and K. K. Bharadwaj, "Pruning trustdistrust network via reliability and risk estimates for quality recommendations," *Social Network Analysis and Mining*, vol. 3, no. 1, pp. 65–84, 2013.

[17] P. Massa and P. Avesani, "Trust-aware collaborative filtering for recommender systems." in *CoopIS/DOA/ODBASE (1)*, ser. Lecture Notes in Computer Science, R. Meersman and Z. Tari, Eds., vol. 3290. Springer, 2004, pp. 492–508.

[18] M. Y. H. Al-Shamri and K. K. Bharadwaj, "Fuzzy-genetic approach to recommender systems based on a novel hybrid user model," *Expert Syst. Appl.*, vol. 35, no. 3, pp. 1386–1399, 2008.

[19] H. Ma, I. King, and M. R. Lyu, "Effective missing data prediction for collaborative filtering," in *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands,*

July 23-27, 2007, W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, Eds. ACM, 2007, pp. 39–46.

[20] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 355–369, 2007.

[21] C. Desrosiers and G. Karypis, "Solving the sparsity problem: Collaborative filtering via indirect similarities," *Technical Report, University of Minnesota*, vol. TR 08-044, no. 3, 2008.

[22] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to algorithms*. Cambridge: MIT Press, 1992.

[23] K. H. Kim and S. Choi, "Neighbor search with global geometry: a minimax message passing algorithm," in *Proceedings of the 24th international conference on machine learning*, 2007, pp. 401–408.

[24] L. Candillier, F. Meyer, and F. Fessant, "Designing specific weighted similarity measures to improve collaborative filtering systems," in *Advances in Data Mining. Medical Applications, E-Commerce, Marketing, and Theoretical Aspects, 8th Industrial Conference, ICDM 2008, Leipzig, Germany, July 16-18, 2008, Proceedings*, ser. Lecture Notes in Computer Science, P. Perner, Ed., vol. 5077. Springer, 2008, pp. 242–255.

[25] *MovieLens dataset*, 2003. [Online]. Available: http://www.grouplens.org/data/

[26] *Jester, online joke recommender system and dataset*, 2001. [Online]. Available: http://www.ieor.berkeley.edu/˜goldberg/jester-data/

[27] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*. John Wiley and Sons, 2003.