Information filtering based on transferring similarity

Duo Sun,[1] Tao Zhou,[1,2,*] Jian-Guo Liu,[1,2] Run-Ran Liu,[1] Chun-Xiao Jia,[1] and Bing-Hong Wang[1,3]
[1]*Department of Modern Physics and Nonlinear Science Center, University of Science and Technology of China, Hefei Anhui 230026, People's Republic of China*
[2]*Department of Physics, University of Fribourg, Chemin du Musee 3, CH-1700 Fribourg, Switzerland*
[3]*Research Center for Complex System Science, University of Shanghai for Science and Technology, Shanghai 200093, People's Republic of China*

In this Brief Report, we propose an index of user similarity, namely, the transferring similarity, which involves all high-order similarities between users. Accordingly, we design a modified collaborative filtering algorithm, which provides remarkably higher accurate predictions than the standard collaborative filtering. More interestingly, we find that the algorithmic performance will approach its optimal value when the parameter, contained in the definition of transferring similarity, gets close to its critical value, before which the series expansion of transferring similarity is convergent and after which it is divergent. Our study is complementary to the one reported in [E. A. Leicht, P. Holme, and M. E. J. Newman, Phys. Rev. E **73**, 026120 (2006)], and is relevant to the missing link prediction problem.

With the exponential growth of the internet [1] and the world-wide-web [2], a prominent challenge for modern society is the information overload. Since there are enormous data and sources, people never have time and vigor to find out those most relevant for them. A landmark for solving this problem is the use of search engine [3,4]. However, a search engine could only find the relevant web pages according to the input keywords without taking into account the personalization, and thus returns the same results regardless of users' habits and tastes. Thus far, with the help of *Web2.0* techniques, personalized recommendations become the most promising way to efficiently filter out the information overload [5]. Motivated by the significance in economy and society, devising efficient and accurate recommendation algorithms becomes a joint focus from theoretical studies [5] to e-commerce applications [6]. Various kinds of algorithms have been proposed, such as collaborative filtering (CF) [7,8], content-based methods [9,10], spectral analysis [11,12], iterative refinement [13], principle component analysis [14], network-based inference [15–18], and so on.

A recommender system consists of users and objects, and each user has rated some objects. Denoting the user set as $U=\{u_1,u_2,\ldots,u_N\}$ and the object set as $O=\{o_1,o_2,\ldots,o_M\}$, the system can be fully described by an $N\times M$ rating matrix $\mathbf{V}$, with $v_{i\alpha}\neq 0$ denoting the rating user u_i gives to object o_α. If u_i has not yet evaluated o_α, $v_{i\alpha}$ is set as zero. CF system has been one of the most successfully and widest used recommender systems since its appearance in the mid-1990s [7,8]. Its basic idea is that the user will be recommended objects based on the weighted combination of similar users' opinions. In the standard CF, the predicted rating $v'_{i\alpha}$ from user u_i to object o_α is set as

$$v'_{i\alpha}=\bar{v}_i+I\sum_j s_{ij}(v_{j\alpha}-\bar{v}_j), \tag{1}$$

where s_{ij} is the similarity between u_i and u_j, $\bar{v}_i$ means the average rating of u_i and $I=(\Sigma_j s_{ij})^{-1}$ serves as the normalization factor. Here, j runs over all users having rated object o_α excluding u_i himself. The similarity, s_{ij}, plays a crucial role in determining the algorithmic accuracy. In the implementation, the similarity between every pair of users is calculated first, and then the predict ratings by Eq. (1). Various similarity measures have been proposed, among which the Pearson correlation coefficient is the widest used [7], as

$$s_{ij}=\frac{\sum_c (v_{ic}-\bar{v}_i)(v_{jc}-\bar{v}_j)}{\sqrt{\sum_\alpha (v_{i\alpha}-\bar{v}_i)^2}\sqrt{\sum_\beta (v_{j\beta}-\bar{v}_j)^2}}, \tag{2}$$

where c, α, and β run over all the objects commonly selected by user i and j. All diagonal elements in the similarity matrix are set to be zero, which has no effect on the predicted ratings by Eq. (1). We make this small modification of the standard Pearson coefficient to make sure the transferred similarity between two nodes [see Eq. (3)] is contributed only by the medi-users.

Several algorithms [19–21] have recently been proposed to improve the accuracy of the standard CF via modifying the definition of user-user similarity. However, all those algorithms have not fully addressed the similarity induced by indirect relationship, say, the high-order correlations. Note that, the Pearson correlation coefficient, s_{ij}, considers only the direct correlation. We argue that to appropriately measure the similarities between users, the indirect correlations should also be taken into consideration. To make our idea clearer, we draw an illustration in Fig. 1. Suppose there are three users, labeled as A, B, and C. Although the similarity between user A and C is quite small, A and C are both very similar with B. Actually, A, B, and C may share very similar tastes, and the very small similarity between A and C may be

*zhutou@ustc.edu

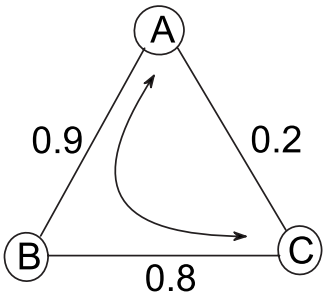

FIG. 1. Illustration for transferring similarity.

caused by the sparsity of the data. That is to say, A and C has a very few commonly selected objects. The sparsity of data set makes the direct similarity less accurate, and thus we expect a new measure of similarity properly integrating high-order correlations may perform better.

Denoting ε a decay factor of similarity transferred by a medi-user, a self-consistent definition of *transferring similarity* can be written as

$$t_{ij} = \varepsilon \sum_{v} s_{iv} t_{vj} + s_{ij}, \tag{3}$$

where s_{ij} is the direct similarity as shown in Eq. (2). The parameter ε can be considered as the rate of information aging by transferring one step further [22]. Clearly, the transferring similarity will degenerate to the traditional Pearson correlation coefficient when $\varepsilon=0$. Denoting $\mathbf{S}=\{s_{ij}\}_{N\times N}$ and $\mathbf{T}=\{t_{ij}\}_{N\times N}$ the direct similarity matrix and the transferring similarity matrix, Eq. (3) can be rewritten in a matrix form, as

$$\mathbf{T} = \varepsilon \mathbf{S}\mathbf{T} + \mathbf{S}, \tag{4}$$

whose solution is

$$\mathbf{T} = (1 - \varepsilon \mathbf{S})^{-1}\mathbf{S}. \tag{5}$$

Accordingly, the prediction score reads

$$v'_{i\alpha} = \bar{v}_i + I' \sum_{j} t_{ij}(v_{j\alpha} - \bar{v}_j), \tag{6}$$

where multiplier $I' = (\Sigma_j t_{ij})^{-1}$ serves as the normalizing factor and j runs over all users having rated object o_α excluding u_i himself.

To test the algorithmic accuracy, we use a benchmark data set, namely, *MovieLens*, which consists of $N=943$ users, $M=1682$ objects, and 10^5 discrete ratings from 1 to 5. The sparsity of the rating matrix $\mathbf{V}$ is about 6%. We first randomly divide this data set into two parts: one is the training set, treated as known information, and the other is the probe, whose information is not allowed to be used for prediction. Then we make a prediction for every entry contained in the probe (resetting $v'_{i\alpha}=5$ and $v'_{i\alpha}=1$ in the case of $v'_{i\alpha}>5$ and $v'_{i\alpha}<1$, respectively), and measure the difference between the predicted rating $v'_{i\alpha}$ and the actual rating $v_{i\alpha}$. For evaluating the accuracy of recommendations, many different metrics have been proposed [7]. We choose two commonly used measures: *root-mean-square error* (RMSE) and *mean absolute error* (MAE). They are defined as

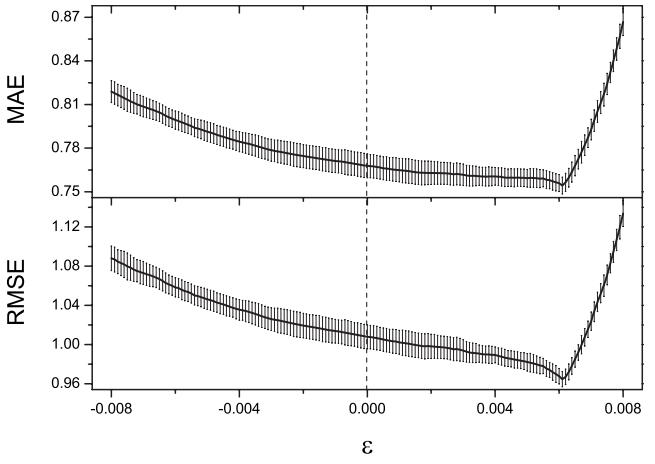

FIG. 2. Prediction accuracy of the present algorithm, measured by MAE and RSME, as functions of ε. The transferring similarities are directly obtained by Eq. (5). The numerical results are averaged over 20 independent runs, each corresponds to a random division with training set containing about 90% of data while the probe consisted of the remain 10%. The error bars denote the standard deviations of the 20 samples.

$$\mathrm{RMSE} = \sqrt{\sum_{(i,\alpha)} (v'_{i\alpha} - v_{i\alpha})^2 / E}, \tag{7a}$$

$$\mathrm{MAE} = \frac{1}{E} \sum_{(i,\alpha)} |v'_{i\alpha} - v_{i\alpha}|, \tag{7b}$$

where the subscript (i,α) runs over all the elements in the probe, and E is the number of those elements.

In Figs. 2–4, we report the numerical results about the algorithmic accuracy, where the divisions of training set and probe are 90% vs 10%, 50% vs 50%, and 10% vs 90%, respectively. In every case, there exists an optimal value of ε, denoted by $\varepsilon_{\mathrm{opt}}$, corresponding to both the lowest MAE and the lowest RMSE. Around the optimal value, $\varepsilon_{\mathrm{opt}}$, the present algorithm obviously outperforms the standard CF. The

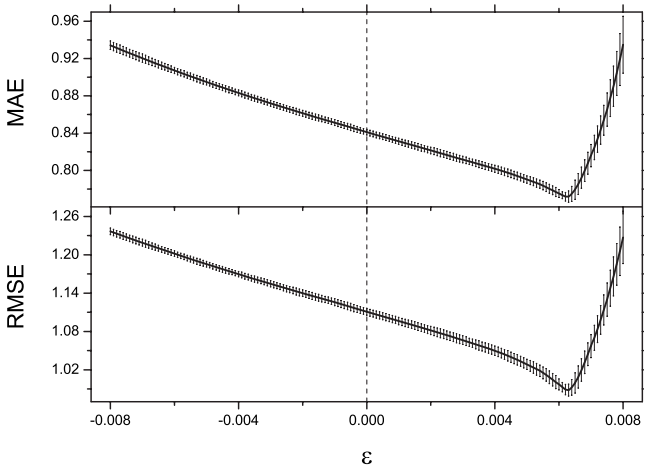

FIG. 3. Prediction accuracy of the present algorithm, where the division of training set and probe is 50% vs 50%. Other conditions are the same as what presented in Fig. 2.

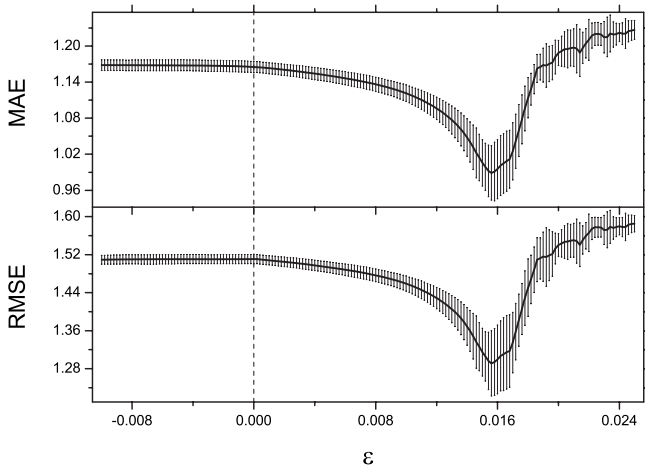

FIG. 4. Prediction accuracy of the present algorithm, where the division of training set and probe is 10% vs 90%. Other conditions are the same as what presented in Fig. 2.

present algorithm can also beat a recently proposed algorithm based on an opinion diffusion process for the same data set [16], which gives predictions with RMSE$\approx$1.00 and MAE$\approx$0.80 for the 90% vs 10% division (the corresponding errors in the optimal cases for the present algorithm are RMSE$\approx$0.96 and MAE$\approx$0.75).

The optimal values of ε are different for different cases, and the one corresponding to sparser data is larger. In addition, the improvement of accuracy is larger for sparser data. In the sparse case, the Pearson coefficient considering only local information is not distinguishable for two users generally vote only a very few overlapped objects, therefore the information from medi-users plays significant role and the improvement is great as well as the difference between T and S is remarkable. While in the dense case, two users usually have many commonly voted objects, and thus the Pearson coefficient can give accurate description on user similarity and the information contained by long-range interactions is less helpful. In addition, the specific case as shown in Fig. 1 is very unlikely to happen. Since in the real world, the data sets are usually extremely sparse (the density of *MovieLens* is about 6%, while for *Netflix.com* it is about 1%, for *RateYourMusic.com* about 0.3%, for *Del.icio.us* about 0.05%), the transferring similarity is practically useful.

Equation (5) can be expanded by a power series, as

$$\mathbf{T} = \mathbf{S} + \varepsilon\mathbf{S}^2 + \varepsilon^2\mathbf{S}^3 + \cdots . \tag{8}$$

Since to directly inverse $(\mathbf{1}-\varepsilon\mathbf{S})$ takes long time for huge-size systems ($\mathbf{1}-\varepsilon\mathbf{S}$ is generally not a sparse matrix, so the computational time scales as N^3 by *Gaussian elimination*, $N^{2.807}$ by *Strassen algorithm*, and $N^{2.376}$ by *Coppersmith-Winograd algorithm* [23]), the cutoff

$$\mathbf{T} = \mathbf{S} + \varepsilon\mathbf{S}^2 + \cdots + \varepsilon^n\mathbf{S}^{n+1}, \tag{9}$$

is usually used as an approximation in the implementation (although the matrix multiplication has the same order of computational complexity as the inversion, it takes much shorter time, and its advantage is that the multiplication of matrix can be saved and reused in searching the optimal ε while the matrix inversion has to be redone when changing ε). However, in this paper, since the system size in not too large, we always directly use Eq. (5) to obtain the transferring similarity matrix, which works out less than one second in a desktop computer with a single *Inter CoreE2160* processor (1.8 GHz) and 1 GB EMS memory. Note that, even if $(\mathbf{1}-\varepsilon\mathbf{S})$ is inversable, Eq. (8) may not be convergent. Actually, Eq. (8) is convergent if and only if all the eigenvalues of $(\mathbf{1}-\varepsilon\mathbf{S})$ are strictly smaller than 1. The mathematical proof of a very similar proposition using *Jordan matrix decomposition* can be found in Ref. [22]. Although Ref. [22] only gives the proof of the sufficient condition, the necessary condition can be proved in an analogical way. Accordingly, there exists a critical point of ε, before which the spectral radius of $\varepsilon\mathbf{S}$ is less than 1 and after which it exceeds 1. Since this critical value is also the maximal value of ε that keeps the convergence of Eq. (8), we denote it by $\varepsilon_{\max}$. The optimal and maximal values of ε for the three cases corresponding to Figs. 2–4 is presented in Table I. It is very interesting that $\varepsilon_{\rm opt}$ is always smaller yet very close to $\varepsilon_{\max}$.

TABLE I. The optimal and maximal values of ε for the three cases corresponding to Figs. 2–4. $\varepsilon_{\max}$ is obtained by averaging 20 independent runs, and we have checked that in each run $\varepsilon_{\rm opt}$ is always a little bit smaller than $\varepsilon_{\max}$. The resolution of ε is 10^{-3} since for higher resolution (e.g., 10^{-4}), the difference between two neighboring data point is very small, and the optimal value is not distinguishable with the presence of fluctuations.

Data divisions	90% vs 10%	50% vs 50%	10% vs 90%
$\varepsilon_{\rm opt}$	0.0061	0.0063	0.0156
$\varepsilon_{\max}$	0.006136	0.006311	0.015642

In summary, we designed an improved collaborative filtering algorithm based on a proposed similarity measure, namely, the transferring similarity. Different from the traditional definitions of similarity that consider the direct correlation only, the transferring similarity integrates all the high-order (i.e., indirect) correlations. The numerical testing on a benchmark data set has demonstrated the improvement of algorithmic accuracy compared with the standard CF algorithm. Very recently, Zhou *et al.* [24] and Liu *et al.* [21] proposed some modified recommendation algorithms under the frameworks of collaborative filtering [21] and random-walk-based recommendations [24], respectively. By taking into account both the direct and the second-order correlations, their algorithms can remarkably enhance the prediction accuracy. These works can be considered as a bridge connecting the nearest-neighborhood-based information filtering algorithms and the present work.

Very interestingly, we found that the optimal value of ε is always smaller yet very close to the maximal value of ε that guarantees the convergence of power-series expansion of the transferring similarity. The significance of this finding is twofold. First, Leicht, Holme, and Newman [25] have recently proposed a new index of node similarity, which is actually a variant of the well-known Katz index [26]. The

numerical tests [25] showed that their index best reproduces the known correlations between nodes when the parameter is very close to its maximal value that guarantees the convergence of power-series expansion. Although their work and the current work originate from different motivations and use different testing methods, the results are surprisingly coincident. Despite the insufficiency of empirical studies and the lack of analytical insights, this finding should be of theoretical interests. Second, $\varepsilon_{\max}$ is equal to the inverse of the maximum eigenvalue of $\mathbf{S}$, $\lambda_{\max}^{-1}$. Therefore, it is easy to determine $\varepsilon_{\max}$ since fast algorithms on calculating $\lambda_{\max}$ for a given matrix is well developed (see, for example, the *power iteration method* in Ref. [23]). When dealing with an unknown system, we can first calculate $\lambda_{\max}$, and then concentrate the search of ε_{opt} on the area around $\lambda_{\max}^{-1}$, which can save computations in real applications.

Very recently, a fresh issue is raised to physics community, that is, how to predict missing links of complex networks [27,28]. The fundamental problem is to determine the proximities, or say similarities, between node pairs [29,30]. The similarity index presented here is not only an extension of the Pearson correlation coefficient in rating systems, but also easy to be extended to quantify the structural similarity of node pair in general networks based on any locally defined similarity indices. We believe this self-consistent definition of similarity [see Eq. (3)] can successfully find its applications in link prediction problem.

We acknowledge *GroupLens Research Group* for *MovieLens* data [31]. This work is supported by the National Natural Science Foundation of China under Grants No. 60744003 and No. 10635040. T.Z. and J.-G.L. acknowledge the Swiss National Science Foundation (Grant No. 200020-121848).

[1] G.-Q. Zhang, G.-Q. Zhang, Q.-F. Yang, S.-Q. Cheng, and T. Zhou, New J. Phys. **10**, 123027 (2008).
[2] A. Broder, R. Kumar, F. Moghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, Comput. Netw. **33**, 309 (2000).
[3] S. Brin and L. Page, Comput. Netw. ISDN Syst. **30**, 107 (1998).
[4] J. M. Kleinberg, J. ACM **46**, 604 (1999).
[5] G. Adomavicius and A. Tuzhilin, IEEE Trans. Knowl. Data Eng. **17**, 734 (2005).
[6] J. B. Schafer, J. A. Konstan, and J. Riedl, Data Min. Knowl. Discov. **5**, 115 (2001).
[7] J. L. Herlocker, J. A. Konstan, K. Terveen, and J. T. Riedl, ACM Trans. Inf. Syst. **22**, 5 (2004).
[8] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl, Commun. ACM **40**, 77 (1997).
[9] M. Balabanović and Y. Shoham, Commun. ACM **40**, 66 (1997).
[10] M. J. Pazzani, Artif. Intell. Rev. **13**, 393 (1999).
[11] D. Billsus and M. Pazzani, Proceedings of the International Conference in Machine Learning, 1998 (Morgan Kaufmann Publishers, San Francisco, 1998), p. 46–54.).
[12] B. Sarwar, G. Karypis, J. A. Konstan, and J. T. Riedl, Proceedings of the ACM WebKDD Workshop, 2000 (unpublished).
[13] J. Ren, T. Zhou, and Y.-C. Zhang, EPL **82**, 58007 (2008).
[14] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, Inf. Retr. **4**, 133 (2001).
[15] Y.-C. Zhang, M. Blattner, and Y.-K. Yu, Phys. Rev. Lett. **99**, 154301 (2007).
[16] Y.-C. Zhang, M. Medo, J. Ren, T. Zhou, T. Li, and F. Yang, EPL **80**, 68003 (2007).
[17] T. Zhou, J. Ren, M. Medo, and Y.-C. Zhang, Phys. Rev. E **76**, 046115 (2007).
[18] T. Zhou, L. L. Jiang, R. Q. Su, and Y.-C. Zhang, EPL **81**, 58004 (2008).
[19] J.-G. Liu, B.-H. Wang, and Q. Guo, Int. J. Mod. Phys. C **20**, 285 (2009).
[20] R.-R. Liu, C.-X. Jia, T. Zhou, D. Sun, and B.-H. Wang, Physica A **388**, 462 (2009).
[21] J.-G. Liu, T. Zhou, B.-H. Wang, and Y.-C. Zhang, e-print arXiv:0808.3726.
[22] A. Stojmirovic and Y.-K. Yu, J. Comput. Biol. **14**, 1115 (2007).
[23] G. H. Golub and C. F. Von Load, *Matrix Computation* (Johns Hopkins University Press, Baltimore, 1996).
[24] T. Zhou, R.-Q. Su, R.-R. Liu, L.-L. Jiang, B.-H. Wang, and Y.-C. Zhang, e-print arXiv:0805.4127.
[25] E. A. Leicht, P. Holme, and M. E. J. Newman, Phys. Rev. E **73**, 026120 (2006).
[26] L. Katz, Psychometrika **18**, 39 (1953).
[27] A. Clauset, C. Moore, and M. E. J. Newman, Nature (London) **453**, 98 (2008).
[28] S. Redner, Nature (London) **453**, 47 (2008).
[29] D. Liben-Nowell and J. Kleinberg, J. Am. Soc. Inf. Sci. Technol. **58**, 1019 (2007).
[30] T. Zhou, L. Lü, and Y.-C. Zhang, e-print arXiv:0901.0553, Eur. Phys. J. B (to be published).
[31] http://www.grouplens.org