

Information filtering via weighted heat conduction algorithm

Jian-Guo Liu^{a,b,*}, Qiang Guo^a, Yi-Cheng Zhang^{a,c}

^a Research Center of Complex Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, PR China

^b CABDyN Complexity Centre, Säid Business School, University of Oxford, Park End Street, Oxford OX1 1HP, United Kingdom

^c Department of Physics, University of Fribourg, Chemin du Musée 3, CH-1700 Fribourg, Switzerland

In this paper, by taking into account effects of the user and object correlations on a heat conduction (HC) algorithm, a weighted heat conduction (WHC) algorithm is presented. We argue that the edge weight of the user-object bipartite network should be embedded into the HC algorithm to measure the object similarity. The numerical results indicate that both the accuracy and diversity could be improved greatly compared with the standard HC algorithm and the optimal values reached simultaneously. On the Movielens and Netflix datasets, the algorithmic accuracy, measured by the average ranking score, can be improved by 39.7% and 56.1% in the optimal case, respectively, and the diversity could reach 0.9587 and 0.9317 when the recommendation list equals to 5. Further statistical analysis indicates that, in the optimal case, the distributions of the edge weight are changed to the Poisson form, which may be the reason why HC algorithm performance could be improved. This work highlights the effect of edge weight on a personalized recommendation study, which may be an important factor affecting personalized recommendation performance.

1. Introduction

The amount of information in the world is increasing far more quickly than our ability to process it. All of us have known the feeling of being overwhelmed by the number of new books, journal articles, and conference proceedings coming out each year [1–5]. Technology has dramatically reduced the barriers to publishing and distributing information. Being an effective tool to address this problem, the recommender system has caught an increasing amount of attention from researchers to engineers, and it has become an essential issue in Internet applications such as e-commerce systems and digital library systems [6]. Recommender systems are one of the fastest growing segments of the Internet economy today. They help reduce information overload and provide customized information access for targeted domains. It is being widely used in many application settings to suggest products, services, and information items to potential consumers. A personalized recommender system includes three parts: data collection, model analysis and the recommendation algorithm, among which the algorithm is the core part. Motivated by the significance in economy and society, various kinds of algorithms have been proposed, including collaborative filtering (CF) approaches [7–15], content-based analyses [16,17], tag-based algorithms [18–20], link prediction approach [21], hybrid algorithms [22–24], and so on. For a review of current progress, see Refs. [4,5] and the references therein.

Very recently some physical dynamics, such as the heat conduction (HC) process [25] and probability diffusion method [26] have been successfully applied in personalized recommendation. In the standard HC algorithm, the objects one target

* Corresponding address: Business School, University of Shanghai for Science and Technology, No. 516 Jungong Road, Yangpu District, 200093 Shanghai, PR China.

E-mail address: liujg004@ustc.edu.cn (J.-G. Liu).

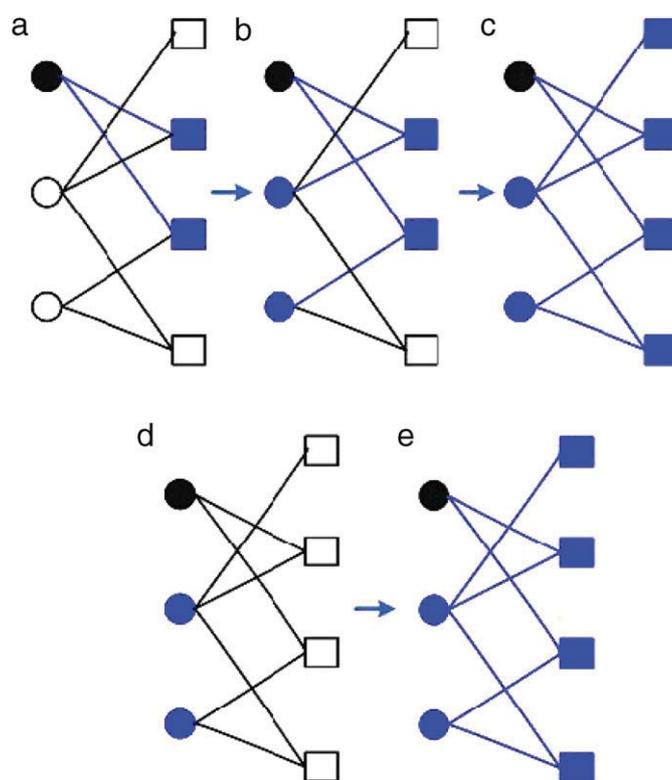


Fig. 1. (Color online) Illustration of network-based recommendation (a–c) and collaborative filtering (d, e) algorithms on the user–object bipartite network. Users and objects are shown as circles and squares, with the target user indicated by the black circle. The network-based algorithm supposes the objects one user has collected are able to recommend new objects (a–c), while the collaborative filtering algorithm is implemented based on the neighbor users’ opinions (d, e).

user has selected were set as the heat resource with constant temperature 1, all other unselected object temperatures are set as 0. The heat is diffused from the object set to user set, then it flows back again. To one heat resource with temperature 1, the unselected objects’ final temperature is defined as the direct similarity to the resource node. Since the HC algorithm is implemented based on user–object bipartite networks, it’s also called a *network-based algorithm*. Fig. 1 demonstrated the main idea of the network-based (a–c) and CF (d, e) algorithms. Combining the similar neighbors’ opinions, CF algorithms recommend new objects to the target user. Unlike the standard CF algorithm, the network-based algorithms suppose that the objects one user has collected/selected have the power to recommend new objects. The numerical results have indicated that these physical approaches have both high efficiency and low computational complexity [9–12,25–29]. However, all of the above algorithms are implemented in the topological user–object bipartite networks, in which the edge between a user and an object indicates that the user is fond of this object. In other words, all of the objects and users with far different degrees have been treated equally. By introducing the collaborative similarity, Shang et al. [30] found the correlation between users’ degree and taste diversity, which indicates that the edge contains unequal information in a bipartite network. Liu et al. [14] investigated the edge weight effect on CF algorithm and found that, in the optimal case, the edge weight obeys the power-law distribution. Inspired by the above works, we argue that the edge weight should be taken into account to analyze the effects on HC algorithm performance. Since each edge contains the target user’s special interest or habit information, the edge diversity could approximately reflect the user’s tastes.

In this paper, we suppose that the preference level endowed in each edge relies on the degrees of the two connected nodes, which should be taken into account to improve HC algorithm performance. Implementing the HC algorithm on the weighted user–object bipartite network, a modified algorithm named the weighted HC algorithm (WHC) is present. The numerical results indicate, when the edge weight obeys the Poisson distribution, both the accuracy and diversity could be improved greatly and reach the optimal values approximately simultaneously.

2. Heat conduction recommendation algorithm

A recommender system consists of users and objects, and each user has collected some objects. Denote the object-set by $O = \{o_1, o_2, \dots, o_m\}$ and user-set by $U = \{u_1, u_2, \dots, u_n\}$, the recommender system can be fully described by a bipartite network with $m + n$ nodes, where an object is connected with a user if and only if the object has been collected by this user. Connections between any two users or two objects are not allowed. Based on the bipartite user–object network,

an object–object network can be constructed, where two objects are connected if and only if they have been collected simultaneously by at least one user. Unlike the CF algorithm, the network-based algorithms suppose that the objects one user has collected have the ability to recommend other new objects for him/her.

The standard HC processes have been successfully introduced to the personalized recommendation [25]. It could produce higher diversity and lower accuracy. To a general user–object network, the weighted projection onto object–object network reads in the following ways. The HC algorithm is akin to a heat diffusion process on the bipartite user–object network [25]. Assigning objects in the network an initial level of resource denoted by the vector \mathbf{f} , we then redistribute it via the transformation $\hat{\mathbf{f}} = \mathbf{W}\mathbf{f}$, where

$$w_{\alpha\beta} = \frac{1}{k_\alpha} \sum_{i=1}^n \frac{a_{\alpha i} a_{\beta i}}{k_i} \quad (1)$$

is a row-normalized $m \times m$ probability matrix representing the diffusion process. The resulting recommendation list of uncollected objects is then sorted according to \hat{f}_i in descending order and those objects with highest values of final temperature are recommended.

3. Edge weight definition on bipartite networks

In the user–object bipartite network, if both o_i and o_j have been selected by the same user u_i , they probably have similar features. In the standard HC algorithm, the heat could diffuse between object and user set twice, therefore, statistically speaking, one target object’s neighbors are $k_o k_u$, where k_o and k_u denote the average object and user degree. Fig. 3 investigates the distribution of the degree time product distribution, from which a clear power-law distribution could be found. In other words, in the object similarity network, lots of objects’ neighbor sets have few nodes. We also noticed that, in the final step of HC process, the temperature would be obtained by dividing the termination node’s degree. Therefore the influence of the objects with few neighbors would be enhanced. In order to give more opportunities to the objects with a larger neighbor set, we introduce a tunable parameter λ to measure the degree time product effect on the HC algorithm. According to the above analysis, the edge weight is given in the following way

$$e_{\alpha i} = (k_\alpha k_i)^\lambda. \quad (2)$$

Based on the weighted bipartite network, the object–object similarity could be given by

$$w_{\alpha\beta} = \frac{1}{k_\alpha} \sum_{i=1}^n \frac{a_{\alpha i} e_{\alpha i} a_{\beta i} e_{\beta i}}{k_i}. \quad (3)$$

Based on the new object similarity measurement, the unselected objects could be recommended according to a network-based algorithm framework.

4. Simulation results

4.1. Data sets

Two benchmark datasets, namely MovieLens¹ and Netflix², were used to test the above algorithms. The MovieLens data is a randomly-selected subset of the huge data, which consists of 1682 movies (objects) and 943 users. We applied a coarse-graining method: A movie is set to be collected by a user only if the giving rating is larger than 2. The original data contains 10^5 ratings, 82.52% of which are larger than 2, that is, the user–object (user–movie) bipartite network after the coarse gaining contains 1574 objects and 82,520 edges³. The Netflix data has 10,000 users and 6000 objects, and there are 701,947 edges in the bipartite network.

4.2. Algorithmic performance metrics

To test a recommendation method on a dataset we remove at random 10% of the links and apply the algorithm to the remainder to produce a recommendation list for each user. We then employ three different metrics, one to measure accuracy in the recovery of deleted links and two to measure recommendation popularity and diversity.

The average ranking score is adopted to measure the accuracy, which is defined as follows. For an arbitrary user u_i , if the entry u_i-o_j is in the probe set (according to the training set, o_j is an uncollected object for u_i), we measure the position of o_j

¹ <http://www.grouplens.org>.

² <http://www.netflix.com>.

³ In order to eliminate the effect of coarse gaining method on object set, the quick sort algorithm is implemented in the ranking process, which makes the average ranking score measurement reasonable.

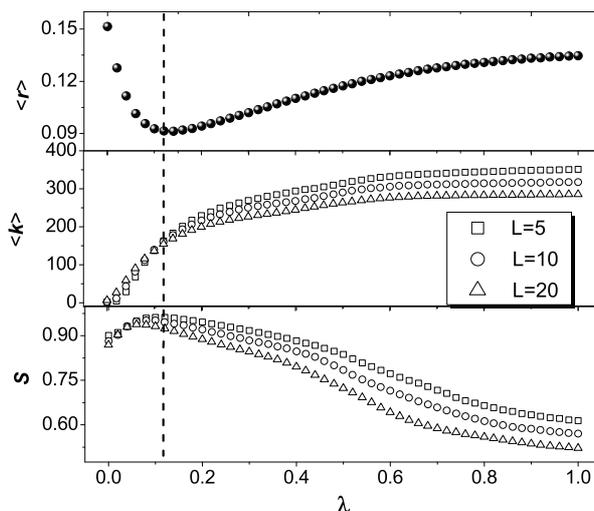


Fig. 2. The average ranking score $\langle r \rangle$, average object degree $\langle k \rangle$ and diversity S vs. λ on Movielens dataset. The accuracy is improved from 0.1516 (corresponding to the routine case $\lambda = 0$) to 0.0914 (corresponding to the optimal case $\lambda_{opt} = 0.14$). When $L = 5$, the diversity is improved from 0.9009 to 0.9587. All the data points are averaged over five independent runs with different data-set divisions.

in the ordered list. For example, if there are 10 uncollected objects for u_i , and o_j is the 3rd from the top, we say the position of o_j is 3/10, denoted by $r_{ij} = 0.3$. Since the probe entries are actually collected by users, a good algorithm is expected to give high recommendations to them, leading to small r_{ij} . Therefore, the mean value of the position r_{ij} , $\langle r \rangle$, averaged over all the entries in the probe, can be used to evaluate the algorithmic accuracy: the smaller the average ranking score, the higher the algorithmic accuracy, and vice versa.

Besides accuracy, the personalized recommendation algorithm should present different recommendations to different users according to their tastes and habits. The diversity can be quantified by the average Hamming distance, $S = \langle H_{ij} \rangle$, where $H_{ij} = 1 - Q_{ij}(L)/L$, L is the length of recommendation list and $Q_{ij}(L)$ is the overlapping number of objects in u_i and u_j 's recommendation lists. The largest $S = 1$ indicates that the recommendations to all of the users are totally different, in other words, the system has the highest diversity. While the smallest $S = 0$ means all of the recommendations are exactly the same. In addition, the average degree of all recommended objects, $\langle k \rangle$ is taken into account to measure the algorithmic popularity. The smaller average degree, corresponding to the less popular objects, are preferred since those lower-degree objects are hard to be found by users themselves.

4.3. Numerical results

The results of algorithmic accuracy, popularity and diversity on Movielens are demonstrated in Fig. 2. From which we can find that the curve of the average ranking scores $\langle r \rangle$ have clear minimums around $\lambda = 0.14$ for the WHC algorithm, which strongly support the above discussion that the algorithmic accuracy could be improved by taking into account the edge weight. Compared with the routine case ($\lambda = 0$), the optimal average ranking score can be reduced by 39.7% for the standard HC algorithm, which is indeed a great improvement. More importantly, we find that when the parameter increases from the standard case $\lambda = 0$ the diversity is also increased from 0.9009 to 0.941. We also notice that the optimal parameter λ_{opt} is positive and smaller than 1, which indicates that, in the optimal case, the edge weights are sorted in the same order as the case $\lambda = 1$, while the relative difference between any two edges is minimized. For example, Fig. 3 shows that the difference between the largest and smallest degree time product changes from 2×10^5 (the case $\lambda = 1$) to $(200\,000)^{0.14} = 5.5226$ (the optimal case $\lambda = 1$). We also find that the mean value of the Poisson is 3.8, which indicates that the edges whose weights are around 1500 ($15\,000^{0.14} = 3.8$) should be enhanced to improve WHC algorithm's performance. Fig. 2 reports the average degrees of all recommended movies as a function of λ , from which one can find that $\langle k \rangle$ is positively correlated with λ . Thus increasing the parameter λ from 0 to 0.14 would enhance edge effects with higher weights, which could give more opportunities to the user interested objects. We also notice that, in the optimal case $\lambda_{opt} = 0.14$, the average object degrees of different recommendation length L are approximately equal. More importantly, the diversity S reaches its highest value when $\lambda = 0.14$. When $L = 10$, the diversity S of the WHC algorithm is increased to 0.941, which is the best diversity value for all of the recommendation algorithms. Table 1 has demonstrated the results obtained by several algorithms. We can find that, besides the hybrid algorithm [24] and the algorithms considering the second-order similarities [29,10], the presented algorithm could generate the best accuracy and diversity simultaneously. Furthermore, the edge weight distributions are investigated in Fig. 3. One can see that, unlike the exponential distribution form when $\lambda = 1$, the edge weight obeys the Poisson distribution, which may be the reason why the HC algorithm on the weighted bipartite network has excellent performance. Similar results are also found on the Netflix data (see Fig. 4), from which we can

Table 1

Algorithmic performance for MovieLens data. The precision, popularity and diversity are corresponding to $L = 10$. GRM, CF and NBI are abbreviations of the global ranking method, collaborative filtering and the network-based inference [27]. Heter-NBI, Hybrid, HC and HO-CF are abbreviations of NBI with heterogeneous initial resource distribution proposed in Ref. [28], hybrid algorithm proposed in Ref. [24], standard heat conduction algorithm proposed in Ref. [25], and the high-order CF algorithm proposed in Ref. [10]. The parameters in Heter-NBI, Hybrid, HO-CF are set as the ones corresponding to the lowest ranking scores (for Heter-NBI $\lambda_{opt} = -0.80$, Hybrid $\lambda_{opt} = 0.2$, and HO-CF $\lambda_{opt} = -0.82$). Each number presented in this table is obtained by averaging over five runs, each of which has an independently random division of training set and probe.

Algorithms	$\langle r \rangle$	S	$\langle k \rangle$
GRM	0.1390	0.398	259
CF	0.1168	0.549	246
NBI	0.1060	0.617	233
Heter-NBI	0.1010	0.682	220
Hybrid	0.0840	0.9173	216
HC	0.1516	0.7880	3.09
HO-CF	0.0826	0.9127	237
WHC	0.0914	0.941	179

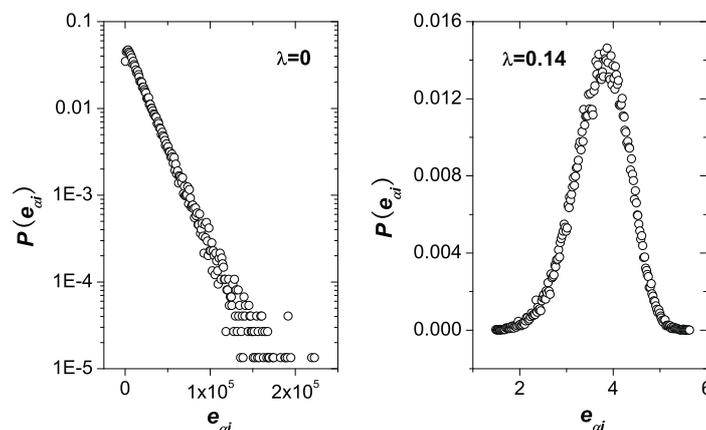


Fig. 3. The distributions of the degree times e_{cd}^λ at the case $\lambda = 1$ and optimal case $\lambda_{opt} = 0.14$ of the MovieLens dataset. The edge weight obeys the power-law distribution when $\lambda = 1$, and approximately obeys the Poisson distribution when $\lambda_{opt} = 0.14$.

find that both the accuracy and diversity reach their optimal values simultaneously when $\lambda_{opt} = 0.10$, which is coincident with the results on MovieLens. The edge weight approximately obeys the Poisson distribution in the optimal case $\lambda_{opt} = 0.10$ (see Fig. 5), which is coincident with the MovieLens data result.

5. Conclusions and discussions

In this paper, the edge weight of the bipartite network is introduced and be embedded into the standard HC algorithm. We argue that the edge weight could reflect the preference level of one target user to the collected objects. By introducing a tunable parameter, the correlations between the parameter and WHC algorithmic performances are investigated. The numerical results on MovieLens and Netflix show that the accuracy, measured by the average ranking score, could be improved from 0.1516 to 0.0914 in the optimal case $\lambda = 0.14$ and the result on Netflix could be improved from 0.1063 to 0.0466 when $\lambda = 0.10$, which has improved the accuracy 39.7% and 56.1%, respectively. Surprisingly, the diversity is also improved and could reach its optimal value in the optimal case λ_{opt} . The diversities of MovieLens and Netflix could be improved to 0.9587 and 0.9317 when the recommendation list $L = 5$. In addition, we find that the edge weight distribution has been approximately changed to the Poisson form, which may be the reason why the algorithm performance is improved. The present algorithm could be used to find the relevant reviewers for the scientific papers or funding applications [31,32], and the link prediction in social and biological networks [33]. We believe the current work can enlighten readers in this promising direction.

Although the higher accuracy, measured by the average ranking score, has been found in several algorithms [24,29,10], the computational complexities of these algorithms are much higher than the presented algorithm and it's hard to reach the optimal accuracy and diversity at the same time. How to automatically find out relevant information for diverse users is

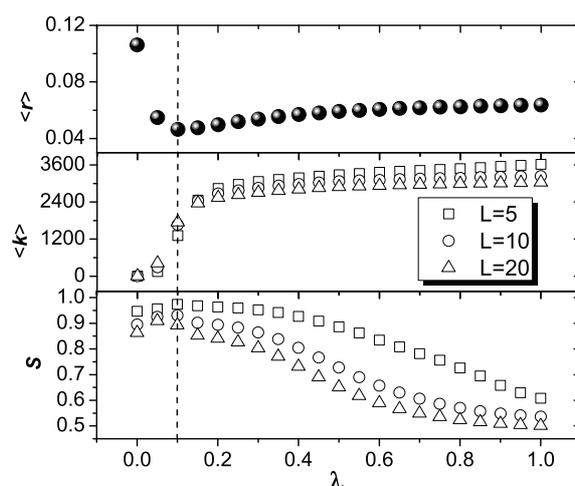


Fig. 4. The average ranking score $\langle r \rangle$, average object degree $\langle k \rangle$ and diversity vs. λ on the Netflix dataset. The accuracy is improved from 0.1063 (corresponding to the standard case $\lambda = 0$) to 0.0467 (corresponding to the optimal case $\lambda_{\text{opt}} = 0.1$). When $L = 5$, the diversity is improved from 0.8949 to 0.9317. All the data points are averaged over five independent runs with different data-set divisions.

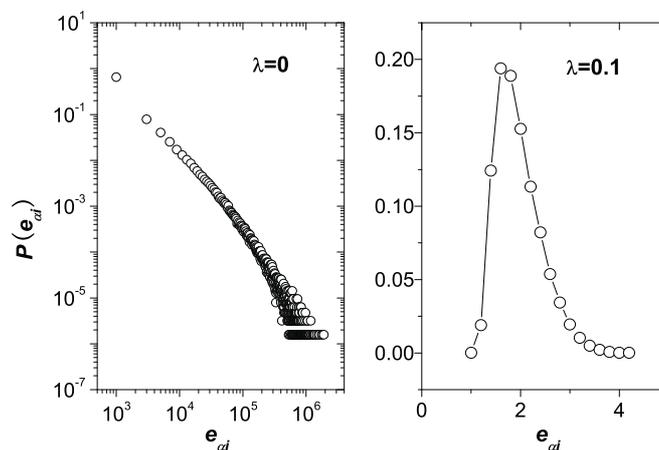


Fig. 5. The distributions of the edge weight e_{ai}^λ at the case $\lambda = 1$ and optimal case $\lambda_{\text{opt}} = 0.10$ of the Netflix dataset. The edge weight obeys the power-law distribution when $\lambda = 1$, and approximately obeys the Poisson distribution when $\lambda_{\text{opt}} = 0.10$.

a long-standing challenge in modern information science, the presented algorithm shows that the statistical properties of the data influence the algorithmic performance greatly. We could present the adaptive algorithm according to the dataset properties. Since there are fewer measurements used to describe the bipartite network, such as the node correlation between different types of nodes, the degree product is used to measure the edge weight, it also could be measured by other ways. Further work can be investigated. Firstly how to select the optimal parameter, then finding the relationship between the Poisson distribution of the edge weight and WHC algorithm.

Acknowledgements

The authors would like to thank Ning Xi for his helpful discussion. We acknowledge the *GroupLens* Research Group for providing us with *MovieLens* data and Netflix Inc. for *Netflix* data. This work is partially supported by National Natural Science Foundation of China (Grant Nos. 10905052, 70901010, 91024026 and 71071098), Processing Fund of Shanghai Key Laboratory of Intelligent Information (IPL-2010-006), and Research and Innovation Project of Shanghai Education Commission (Grant Nos. 11ZZ135, 11YZ110), the Shanghai Development of Undergraduate Education Base III-Electronic Commerce, GQ acknowledges Shanghai Leading Discipline Project (No. S30503).

References

- [1] A. Clauset, C. Moore, M.E.J. Newman, *Nature* 453 (2008) 98.
- [2] G.-Q. Zhang, G.-Q. Zhang, Q.-F. Yang, S.-Q. Cheng, T. Zhou, *New J. Phys.* 10 (2008) 12307.

- [3] P. Resnick, H.R. Varian, *Commun. ACM* 40 (1997) 56.
- [4] G. Adomavicius, A. Tuzhilin, *IEEE Trans. Know. Data Eng.* 17 (2005) 734.
- [5] J.-G. Liu, M.Z.-Q. Chen, J. Chen, F. Deng, H.-T. Zhang, Z.-K. Zhang, T. Zhou, *Int. J. Inf. Syst. Sci.* 5 (2009) 230.
- [6] J.B. Schafer, J.A. Konstan, J. Riedl, *Data Mining Know. Disc.* 5 (2001) 115.
- [7] J.L. Herlocker, J.A. Konstan, K. Terveen, J. Riedl, *ACM Trans. Inform. Syst.* 22 (2004) 5.
- [8] J.A. Konstan, B.N. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon, J. Riedl, *Commun. ACM* 40 (1997) 77.
- [9] J.-G. Liu, B.-H. Wang, Q. Guo, *Int. J. Mod. Phys. C* 20 (2009) 285.
- [10] J.-G. Liu, T. Zhou, B.-H. Wang, Y.-C. Zhang, *Physica A* 389 (2010) 881.
- [11] R.-R. Liu, C.-X. Jia, T. Zhou, D. Sun, B.-H. Wang, *Physica A* 388 (2009) 462.
- [12] D. Sun, T. Zhou, J.-G. Liu, R.-R. Liu, C.-X. Jia, B.-H. Wang, *Phys. Rev. E* 80 (2009) 017101.
- [13] J.-G. Liu, T. Zhou, B.-H. Wang, Y.-C. Zhang, Q. Guo, *Int. J. Mod. Phys. C* 20 (2009) 1925.
- [14] J.-G. Liu, T. Zhou, B.-H. Wang, Y.-C. Zhang, Q. Guo, *Int. J. Mod. Phys. C* 21 (2010) 137.
- [15] Q. Guo, J.-G. Liu, *Int. J. Mod. Phys. C* 21 (2010) 891.
- [16] M. Balabanović, Y. Shoham, *Commun. ACM* 40 (1997) 66.
- [17] M.J. Pazzani, *Artif. Intell. Rev.* 13 (1999) 393.
- [18] M.-S. Shang, Z.-K. Zhang, *Chin. Phys. Lett.* 26 (2009) 118903.
- [19] Z.-K. Zhang, T. Zhou, Y.-C. Zhang, *Physica A* 389 (2010) 179.
- [20] M.-S. Shang, Z.-K. Zhang, T. Zhou, Y.-C. Zhang, *Physica A* 389 (2010) 1259.
- [21] L. Lü, T. Zhou, *Europhys. Lett.* 89 (2010) 18001.
- [22] M. Pazzani, D. Billsus, *Mach. Lear.* 27 (1997) 313.
- [23] N. Good, J.B. Schafer, J.A. Konstan, A.I. Borchers, B. Sarwar, J. Herlocker, J. Riedl, *Proc. Conf. Am. Assoc. Artif. Intell.* 439 (1999).
- [24] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J.R. Wakeling, Y.-C. Zhang, *Proc. Natl. Acad. Sci. USA* 107 (2010) 4511.
- [25] Y.-C. Zhang, M. Blattner, Y.-K. Yu, *Phys. Rev. Lett.* 99 (2007) 154301.
- [26] Y.-C. Zhang, M. Medo, J. Ren, T. Zhou, T. Li, F. Yang, *Europhys. Lett.* 80 (2008) 68003.
- [27] T. Zhou, J. Ren, M. Medo, Y.-C. Zhang, *Phys. Rev. E* 76 (2007) 046115.
- [28] T. Zhou, L.-L. Jiang, R.-Q. Su, Y.-C. Zhang, *Europhys. Lett.* 81 (2008) 58004.
- [29] T. Zhou, R.-Q. Su, R.-R. Liu, L.-L. Jiang, B.-H. Wang, Y.-C. Zhang, *New J. Phys.* 11 (2009) 123008.
- [30] M.-S. Shang, L. Lü, Y.-C. Zhang, T. Zhou, *Europhys. Lett.* 90 (2010) 48006.
- [31] J.-G. Liu, Y.-Z. Dang, Z.-T. Wang, *Physica A* 366 (2006) 578.
- [32] J.-G. Liu, Z.-G. Xuan, Y.-Z. Dang, Q. Guo, Z.-T. Wang, *Physica A* 377 (2007) 302.
- [33] T. Zhou, L. Lü, Y.-C. Zhang, *Eur. Phys. J. B* 71 (2009) 623.