OFFPRINT

Solving the cold-start problem in recommender systems with social tags

Zi-Ke Zhang, Chuang Liu, Yi-Cheng Zhang and Tao Zhou

Please visit the new website
www.epljournal.org

Solving the cold-start problem in recommender systems with social tags

Zi-Ke Zhang[1,2], Chuang Liu[3,4], Yi-Cheng Zhang[1,2(a)] and Tao Zhou[1,5]

[1] *Web Sciences Center, University of Electronic Science and Technology of China - Chengdu 610054, PRC*
[2] *Department of Physics, University of Fribourg - Chemin du Musée 3, 1700 Fribourg, Switzerland*
[3] *School of Business, East China University of Science and Technology - Shanghai 200237, PRC*
[4] *Engineering Research Center of Process Systems Engineering (Ministry of Education), East China University of Science and Technology - Shanghai 200237, PRC*
[5] *Department of Modern Physics, University of Science and Technology of China - Hefei 230026, PRC*

Abstract – Based on the user-tag-object tripartite graphs, we propose a recommendation algorithm that makes use of social tags. Besides its low cost of computational time, the experimental results on two real-world data sets, *Del.icio.us* and *MovieLens*, show that it can enhance the algorithmic accuracy and diversity. Especially, it provides more personalized recommendation when the assigned tags belong to more diverse topics. The proposed algorithm is particularly effective for small-degree objects, which reminds us of the well-known *cold-start problem* in recommender systems. Further empirical study shows that the proposed algorithm can significantly solve this problem in social tagging systems with heterogeneous object degree.

Introduction. – Many complex systems can be well described by networks where nodes represent individuals, and edges denote the relations among them [1–5]. Recently, the design of recommender systems has attracted increasing attention from physical communities [6–22], which aims at finding objects (*e.g.* books, webpages, music, etc.) that are most likely to be collected by users. For example, classical information retrieval can be viewed as recommending documents with given words [23], and the process of link prediction can be considered as a recommendation problem in unipartite networks [24–27]. The core techniques embedded in most recommender systems are twofold: estimating taste similarity based on the historical records of user activities [28,29] and utilizing accessorial information (*e.g.*, object attributes) to efficiently filter out irrelevant objects. However, the accurate descriptions of objects for the latter task is largely limited by the attribute vocabulary and they are usually simply classified into a few system-designed categories that are less helpful to dig out personalized preferences.

Recently, the advent of Web2.0 techniques brings a new paradigm named *social tagging systems* (or *collaborative tagging systems*) for users' participations. A social tagging system allows users to freely assign tags to annotate their collections, requires no specific skills for users to participate in, broadens the semantic relations among users and objects, and thus has attracted much attention from the scientific community. Golder *et al.* studied its usage patterns and classified seven kinds of tag functions [30]. Similar to the tagging functions, the statistics of keywords and PACS numbers are utilized to help characterizing the structure of co-authorship and citation networks [31,32]. Furthermore, many efforts have been done to explain the emergent properties of social tagging systems. Cattuto *et al.* [33] proposed a memory-based Yule-Simon model to describe the aging effects and occurrence frequencies of tags. Zhang and Liu [34] proposed an evolutionary hypergraph model, where users assign tags to relevant objects and retrieve objects via relevant tags.

Besides, social tagging systems have already found wide applications in *Recommender Systems*. By considering the tag frequency as weight, Szomszor *et al.* [35] proposed an improved movie recommendation algorithm.

(a)E-mail: yi-cheng.zhang@unifr.ch

Schenkel *et al.* [36] proposed an incremental threshold algorithm taking into account both the social ties among users and semantic relations of different tags, which performs remarkably better than the algorithm without tag expansion. Zhang *et al.* [37] and Shang *et al.* [38] proposed tag-aware diffusion-based methods to obtain better recommendations. Shang and Zhang [39] considered the tag usage frequency as edge weight in a user-object bipartite network and accordingly designed an improved recommendation algorithm.

In this letter, we propose a diffusion-based recommendation algorithm which treats tags as a bridge connecting users and objects, namely users can efficiently find relevant objects via tags. In particular, we consider the usage frequencies of tags as users' personal preference, while the semantic relations between tags and objects as global information. Experimental results show that the present algorithm can considerably improve the recommendation accuracy, especially for the objects collected by few users, which reminds us of the well-known *cold-start* problem [40,41]. Since tags can build up relations between existent objects and the new ones, the incorporating of tags can remarkably help users in finding the new (or less popular) yet interesting objects, and thus enhance the overall accuracy. In addition, we employ entropy-based and Hamming-distance–based methods to measure the *inner-* and *inter-diversity* of tag usage patterns, respectively. Experimental results show that there are different tag usage patterns in the two datasets: users assign more diverse tags in *Del.icio.us* than *MovieLens*, and it might shed lights on the understanding of why the improvement of algorithmic performance for *Del.icio.us* is remarkably higher than for *MovieLens*.

Data. – The empirical data used in this paper include: i) *Del.icio.us* —one of the most popular social bookmarking web sites, which allows users not only to store and organize personal bookmarks (URLs), but also to look into other users' collections and find what they might be interested in by simply keeping track of the baskets with social tags; ii) *MovieLens* —a movie rating system, where each user votes movies in five discrete ratings 1–5. A tagging function is added in from January 2006. In both data sets, we remove the isolated nodes and guarantee that each user has collected at least one object, each object has been collected by at least two users, assigned by at least two tags, and each tag is used by at least two users, and each tag is used at least twice by every adjacent user. Table 1 summarizes the basic statistics of the purified data sets.

Every data set consists of many entries, and each follows the form $\mathbb{F} = \{\text{user}, \text{object}, \text{tag}_1, \text{tag}_2, \ldots, \text{tag}_t\}$, where t is the number of tags assigned to this object by this user. Each data set is randomly divided into two parts: the training set is treated as known information, while the testing set is used for testing. In this letter, the training set always contains 90% of entries and the remaining 10% of entries constitute the testing set.

Table 1: Basic statistics of the two data sets, *Del.icio.us* (*Del.*) and *MovieLens* (*Mov.*). n, m, r are the total numbers of users, objects and tags, respectively. $\langle k \rangle$, $\langle k' \rangle$ and $\langle k'' \rangle$ denote the average number of objects collected by a user, tags assigned by an object and tags adopted by a user, respectively.

Data	n	m	r	$\langle k \rangle$	$\langle k' \rangle$	$\langle k'' \rangle$
Del.	4902	36224	10584	43.85	38.82	286.86
Mov.	648	1590	1382	15.04	19.89	22.89

Algorithms. – A recommender system considered here consists of three sets, respectively of users $U = \{U_1, U_2, \ldots, U_n\}$, objects $O = \{O_1, O_2, \ldots, O_m\}$, and tags $T = \{T_1, T_2, \ldots, T_r\}$. The tripartite graph representation can be described by three matrices, A, A' and A'' for user-object, object-tag and user-tag relations. If U_i has collected O_j, we set $a_{ij} = 1$, otherwise $a_{ij} = 0$. Analogously, we set $a'_{jk} = 1$, if O_j has been assigned by the tag T_k, and $a'_{jk} = 0$, otherwise. Furthermore, the users' preferences on tags can be represented by a weighted matrix A'', where a''_{ik} is the number of times that U_i has adopted T_k.

We firstly introduce two baseline algorithms: user-object diffusion [11] (I); user-object-tag diffusion [37] (II), and then propose a new algorithm: user-tag-object diffusion (III). Given a target user U_i, the above three algorithms will generate final score of each object, f_j, for her/him according to following rules:

(I) Supposing that a kind of resource is initially located on objects. Each object averagely distributes its resource to all neighboring users, and then each user redistributes the received resource to all his/her collected objects. The final resource vector for the target user U_i, $\vec{f}$, after the two-step diffusion is

$$f_j = \sum_{l=1}^{n} \sum_{s=1}^{m} \frac{a_{lj} a_{ls} a_{is}}{k(U_l) k(O_s)}, \quad j = 1, 2, \ldots, m, \quad (1)$$

where $k(U_l) = \sum_{j=1}^{m} a_{lj}$ is the number of collected objects for user U_l, and $k(O_s) = \sum_{i=1}^{n} a_{is}$ is the number of neighboring users for object O_s.

(II) The initial resources are set as same as I), but each object equally distributes its resource to all neighboring tags, and then each tag redistributes the received resource to all its neighboring objects. Thus, the final resource vector, $\vec{f'}$, is

$$f'_j = \sum_{l=1}^{r} \sum_{s=1}^{m} \frac{a'_{jl} a'_{ls} a_{is}}{k'(T_l) k'(O_s)} \quad j = 1, 2, \ldots, m, \quad (2)$$

where $k'(T_l) = \sum_{j=1}^{m} a'_{jl}$ is the number of neighboring objects for tag T_l, $k'(O_s) = \sum_{l=1}^{r} a'_{sl}$ is the number of neighboring tags for object O_s.

(III) Different from (I) and (II), here, the initial resources are located on tags according to their frequencies used by the target user U_i. Then each tag distributes the initial resource directly to all its neighboring objects. Thus, the final resource vector, $\vec{f''}$, reads

$$f''_j = \sum_{l=1}^{r} \frac{a'_{jl} a''_{il}}{k'(T_l)}. \qquad (3)$$

After we obtain the final score of objects, all the objects having not been collected by the target user U_i are ranked in a descending order, and the top L objects will be recommended to U_i.

Comparing with algorithms I and II, the advantages of algorithm III are threefold. Firstly, since social tags highly reflect users' personal preferences, algorithm III is promisingly expected to generate more personalized recommendation. Secondly, the one-step diffusion can clearly save computational time especially for large-scale data. Thirdly, algorithm III reveals the essential role of tags: helping users retrieve and organize collections without the limit of hierarchial structure and vocabulary of words.

Metrics. $-$ To give solid and comprehensive evaluation of the proposed algorithm, we employ three different metrics that characterize the accuracy and diversity of recommendations.

1) *Ranking Score* (RS) [11]. In the present case, for each entry in the testing set (*i.e.* a user-object pair), RS is defined as the rank of the object, divided by the number of all uncollected objects for the corresponding user. Apparently, the less the RS, the higher accuracy the algorithm is. $\langle RS \rangle$ is given by averaging over all entries in the testing set.

2) *Inter Diversity* ($InterD$) [11,42]. $InterD$ measures the differences of different users' recommendation lists, thus can be understood as the inter-user diversity. Denote O_R^i the set of recommended objects for user U_i, then

$$InterD = \frac{2}{n(n-1)} \sum_{i \neq j} \left(1 - \frac{|O_R^i \cap O_R^j|}{L} \right), \qquad (4)$$

where $L = |O_R^i|$ is the length of recommendation list. In average, greater or smaller $InterD$ mean, respectively, greater or smaller personalization of users' recommendation lists.

3) *Inner Diversity* ($InnerD$) [42]. $InnerD$ measures the differences of objects within a user's recommendation list, thus can be considered as the inner-user diversity. It reads,

$$InnerD = 1 - \frac{2}{nL(L-1)} \sum_{i=1}^{n} \sum_{j \neq l, j, l \in O_R^i} S_{jl}, \qquad (5)$$

Table 2: Algorithmic accuracy for *Del.icio.us*. $\langle RS \rangle_{k_o \leqslant 10}$ is the average ranking score over objects with degree equal to or less than 10, and $\langle RS \rangle_{k_o > 10}$ is the average ranking scores over objects with degree greater than 10. Each value is obtained by averaging over 50 realizations, each of which corresponds to an independent division of training set and testing set.

Algorithms	$\langle RS \rangle$	$\langle RS \rangle_{k_o \leqslant 10}$	$\langle RS \rangle_{k_o > 10}$
I	0.276	0.369	0.054
II	0.209	0.275	0.049
III	0.196	0.249	0.068

Table 3: Algorithmic accuracy for *MovieLens*.

Algorithms	$\langle RS \rangle$	$\langle RS \rangle_{k_o \leqslant 10}$	$\langle RS \rangle_{k_o > 10}$
I	0.207	0.307	0.039
II	0.130	0.168	0.055
III	0.123	0.146	0.070

where $S_{jl} = \frac{|\Gamma_{O_j} \cap \Gamma_{O_l}|}{\sqrt{|\Gamma_{O_j}| \times |\Gamma_{O_l}|}}$ is the cosine similarity between objects O_j and O_l, where Γ_{O_j} denotes the set of users having collected object O_j. In average, greater or smaller $InnerD$ suggests, respectively, greater or smaller topic diversification of users' recommendation lists.

Results. $-$ To make the role of tags clear, a microscopic picture of algorithmic accuracy is very helpful. Especially, since tags are used to describe the objects, we would like to see the dependence of accuracy on object degree, namely the number of users collecting it. Given an object degree k_o, the degree-dependent average ranking score, denoted by $\langle RS \rangle_{k_o}$, is defined as the mean positions averaged over all the entries in the testing set with object degree equal to k_o. In table 2 and table 3 we give the overall $\langle RS \rangle$ of the three algorithms for the observed data sets, showing that $\langle RS \rangle$ is significantly enhanced by the present algorithm. Figure 1 reports the correlation between accuracy and object degree. The ranking score decays with the increasing k_o for all the three algorithms. In addition, the three curves intersect around $k_o = 10$, which is a relatively small value considering the heterogeneous object-degree distribution shown in fig. 2, yet the majority of objects are of degree $k_o \leqslant 10$ (90.04% and 69.35% in *Del.icio.us* and *MovieLens*, respectively). From fig. 1, it is seen that the algorithmic accuracy of algorithm III is better than that of algorithms I or II for $k_o \leqslant 10$, but worse when $k_o > 10$ (see also table 2 and table 3), which reminds us of the well-known cold-start problem in recommender systems: how to recommend the unpopular and/or new objects to users? It is very difficult for a user to be aware of these *cold objects* by random surfing since they are not hot items, and for a system to recommend them to right places since there are usually
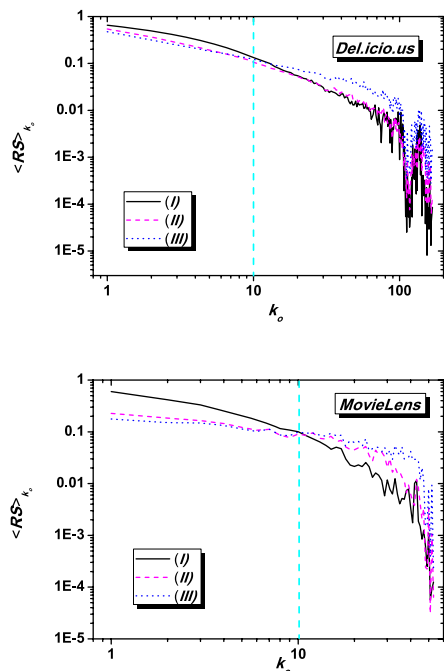
Fig. 1: (Color online) Object-degree–dependent ranking score for the three algorithms in *Del.icio.us* and *MovieLens*. Each data point is obtained by averaging over 50 realizations, each of which corresponds to an independent division of training set and testing set.
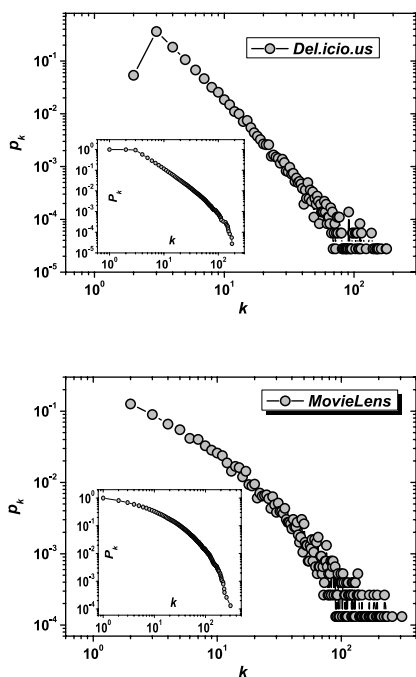

Fig. 2: Object degree distributions of the two data sets. The insets show the accumulative distributions.

insufficient information about them. Comparing with the algorithms I and II, the present one can effectively help users find those cold objects via tags.
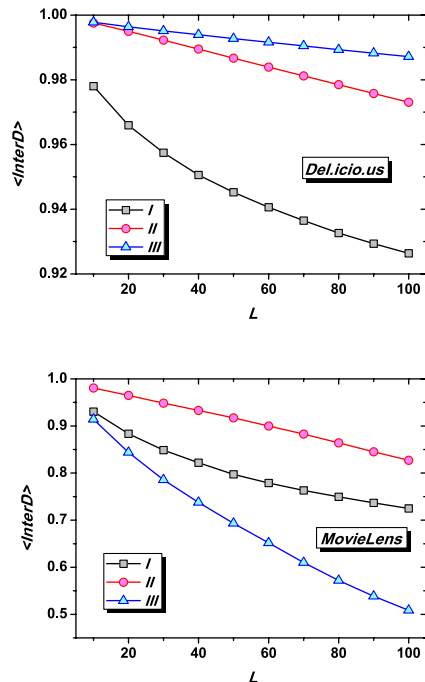

Fig. 3: (Color online) $\langle InterD \rangle$ as a function of the length of the recommendation list for the three algorithms in *Del.icio.us* and *MovieLens*.
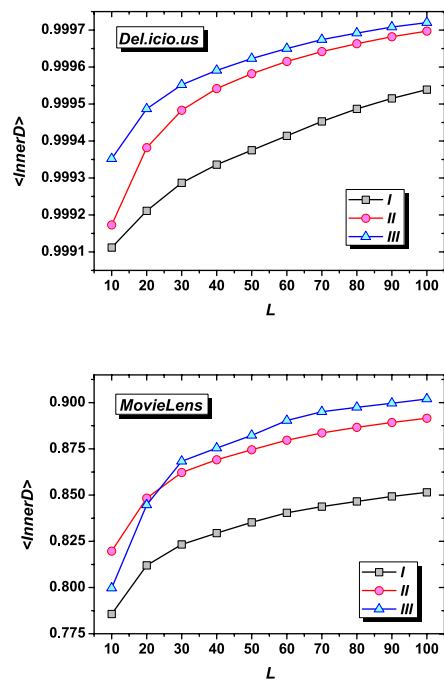

Fig. 4: (Color online) $\langle InnerD \rangle$ as a function of the length of the recommendation list for the three algorithms in *Del.icio.us* and *MovieLens*.

Figures 3 and 4 show the experimental results of $\langle InterD \rangle$ and $\langle InnerD \rangle$, respectively. In fig. 3, $\langle InterD \rangle$ is enhanced only for *Del.icio.us*. The reason for small $\langle InterD \rangle$ of algorithm III in *MovieLens* is that there are only movies in that data set, and thus a comparatively
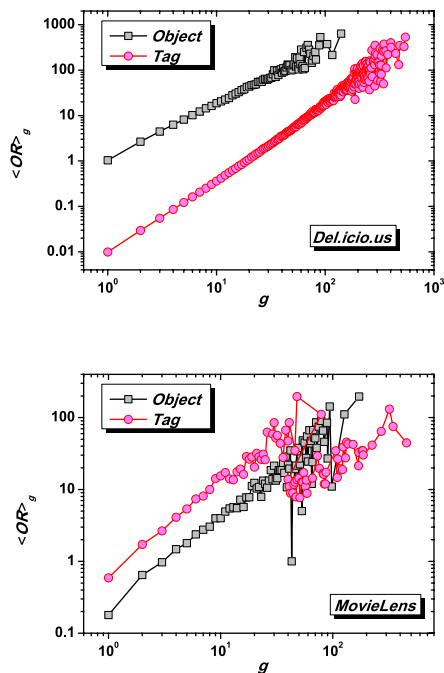
Fig. 5: (Color online) $\langle OR \rangle$ as a function of g for the two data sets. The black squares and red circles represent $\langle OR \rangle$ for objects and tags, respectively.

Fig. 6: (Color online) (a) $\langle E \rangle$ as a function of the user degree; (b) $\langle E \rangle$ as a function of the object degree.

small number of tags are used with huge overlapping. The overlapping ratio of tags, OR_j, for user pairs with g commonly collected objects is defined as

$$OR_g = \frac{1}{N_g} \sum_{i \neq j, G(i,j)=g} OR(i,j), \qquad (6)$$

where N_g is the number of user pairs (i,j) such that $i \neq j$, and $G(i,j) = g$ denotes the number of common objects collected by users i and j. $OR(i,j)$ is defined as the times the same tags are assigned to the same objects. Similar definition can also be used to quantity the overlapping ratio of objects collected by users with the same tags. Clearly, larger OR indicates smaller diversity, and vice versa. Figure 5 shows the correlation between $\langle OR \rangle_g$ and g. One can see that the $\langle OR \rangle_g$ of tags is smaller than that of objects in *Del.icio.us*, while it is not the case for *MovieLens*. In a word, social tags can help generate more diverse recommendation only if the tags are themselves used in a diverse way.

Figure 4 shows that $\langle InnerD \rangle$ is generally improved by our proposed algorithm, indicating that it can help users broaden their horizons. Recently, the *Shannon entropy* is widely used to quantify network diversity in social sharing networks [43] and social economics [44]. In this letter, we also employ it to measure individual usage pattern of tags:

$$E(U_i) = -\sum_t p_{i;t} \ln(p_{i;t}), \qquad (7)$$

where $p_{i;t}$ is the ratio of the occurrence frequency of tag t to the total occurrence times of all U_i's tags. Then the
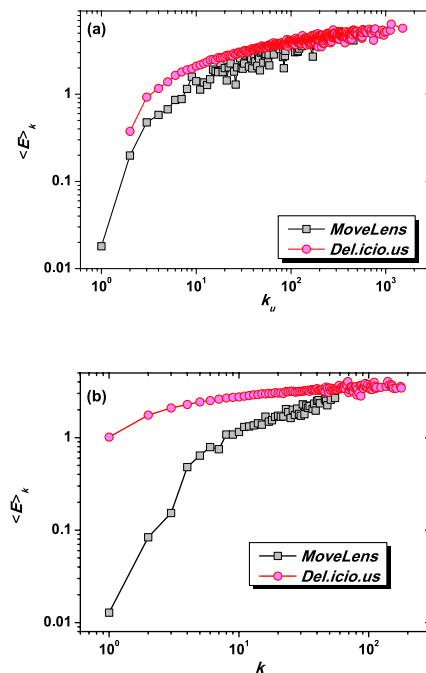
dependence of entropy on user degree, $\langle E \rangle_k$, is given by averaging all the $E(U_i)$ with $k(U_i) = k$. Similar definition can be used to quantify the dependence of entropy for objects. Clearly, larger $\langle E \rangle_k$ means that the users are more willing to use diverse topics of tags, or the objects are more likely to be assigned to more diverse tags, and vice versa. Figure 6 shows that $\langle E \rangle_k$ of *Del.icio.us* is higher than that of *MovieLens*, indicating that *Del.icio.us* is a more diverse system than *MovieLens*, and further giving a reasonable explanation why algorithm III can obtain better $InnerD$ in *Del.icio.us* than *MovieLens*.

Conclusion and discussion. – We proposed a recommendation algorithm making use of social tags, which considers the frequencies of tags as user preferences on different topics and tag-object links as semantical relations between them. Experimental results demonstrated that the proposed algorithm outperforms the two baseline algorithms in both accuracy and diversity.

Of particular importance, the present algorithm outperforms others especially for the objects with small degrees ($k_o \leqslant 10$), which constitute the majority of objects. Therefore, the incorporating of social tags could be, to some extent, helpful in solving the long-standing cold-start problem of recommender systems.

Recently, besides the accuracy, the significance of diversity has attracted more and more attention in information filtering [21]. Experimental results in this letter demonstrated that a wide-range adoption of social tags can enhance the diversity of recommendation. Therefore, we strongly encourage recommender systems to add tagging

functions and users to organize their collections by using tags. However, despite the significant role of tags, the polysemy and synonymy problems [30] might result in coarse and inaccurate performance, the tag clustering technique [45] is hopefully to provide a promising way to generate multi-scale recommendations and eventually obtain the best performance.

∗ ∗ ∗

REFERENCES

[1] ALBERT R. and BARABÁSI A.-L., *Rev. Mod. Phys.*, **74** (2002) 47.

[2] DOROGOVTSEV S. N. and MENDES J. F. F., *Adv. Phys.*, **51** (2002) 1079.

[3] NEWMAN M. E. J., *SIAM Rev.*, **45** (2003) 167.

[4] BOCCALETTI S., LATORA V., MORENO Y., CHAVEZ M. and HUANG D.-U., *Phys. Rep.*, **424** (2006) 175.

[5] COSTA L. DA F., RODRIGUES F. A., TRAVIESOR G. and BOAS P. R. U., *Adv. Phys.*, **56** (2007) 167.

[6] MASLOV S. and ZHANG Y.-C., *Phys. Rev. Lett.*, **87** (2001) 248701.

[7] CANO P., CELMA O., KOPPENBERGER M. and MASLOV S., *Chaos*, **16** (2006) 013107.

[8] BARRE J., *J. Stat. Mech.* (2007) 167.

[9] ZHANG Y.-C., BLATTNER M. and YU Y.-K., *Phys. Rev. Lett.*, **99** (2007) 154301.

[10] ZHANG Y.-C., MEDO M., REN J., ZHOU T., LI T. and YANG F., *EPL*, **80** (2007) 68003.

[11] ZHOU T., REN J., MEDO M. and ZHANG Y.-C., *Phys. Rev. E*, **76** (2007) 046115.

[12] ZHOU T., JIANG L.-L., SU R.-Q. and ZHANG Y.-C., *EPL*, **81** (2008) 58004.

[13] REN J., ZHOU T. and ZHANG Y.-C., *EPL*, **82** (2008) 58007.

[14] JIA C.-X., LIU R.-R., SUN D. and WANG B.-H., *Physica A*, **387** (2008) 5887.

[15] SHANG M.-S., LÜ L., ZENG W., ZHANG Y.-C. and ZHOU T., *EPL*, **88** (2009) 68008.

[16] SUN D., ZHOU T., LIU J.-G., LIU R.-R., JIA C.-X. and WANG B.-H., *Phys. Rev. E*, **80** (2009) 017101.

[17] LIU J. and DENG G.-S., *Physica A*, **388** (2009) 3643.

[18] LIU R.-R., JIA C.-X., ZHOU T., SUN D. and WANG B.-H., *Physica A*, **388** (2009) 462.

[19] NGUYEN V. A., KOUKOLIKOVA-NICOLA Z., BAGNOLI F. and LIO P., *J. Stat. Mech.* (2009) P01014.

[20] LIU J.-G., ZHOU T., CHE H.-A., WANG B.-H. and ZHANG Y.-C., *Physica A*, **389** (2010) 881.

[21] ZHOU T., KUSCSIK Z., LIU J.-G., MEDO M., WAKELING J. R. and ZHANG Y.-C., *Proc. Natl. Acad. Sci. U.S.A.*, **107** (2010) 4511.

[22] SHANG M.-S., LÜ L., ZHANG Y.-C. and ZHOU T., *EPL*, **90** (2010) 48006.

[23] SALTON G. and MCGILL M. J., *Introduction to Model Information Retrieval* (McGraw-Hill, Auckland) 1983.

[24] ZHOU T., LÜ L. and ZHANG Y.-C., *Eur. Phys. J. B*, **71** (2009) 623.

[25] LÜ L., JIN C.-H. and ZHOU T., *Phys. Rev. E*, **80** (2009) 046122.

[26] LÜ L. and ZHOU T., *EPL*, **89** (2010) 18001.

[27] LIU W. and LÜ L., *EPL*, **89** (2010) 58007.

[28] BALABANOVIĆ M. and SHOHAM Y., *Commun. ACM*, **40** (1997) 72.

[29] SARWAR B., KARYPIS G., KONSTAN J. and RIEDL J., *Proceedings of the 10th International Conference on WWW* (ACM Press, New York) 2001, pp. 295–305.

[30] GOLDER S. A. and HUBERMAN B. A., *J. Inf. Sci.*, **32** (2006) 198.

[31] PALLA G., FARKAS I. J., POLLNER P., DERÉYI I. and VICSEK T., *New J. Phys.*, **10** (2008) 123026.

[32] ZHANG Z.-K., LÜ L., LIU J.-G. and ZHOU T., *Eur. Phys. J. B*, **66** (2008) 557.

[33] CATTUTO C., LORETO V. and PIETRONERO L., *Proc. Natl. Acad. Sci. U.S.A.*, **104** (2007) 1461.

[34] ZHANG Z.-K. and LIU C., *J. Stat. Mech.* (2010) P1005.

[35] SZOMSZOR M., CATTUTO C., ALANI H., OHARA K., BALDASSARRI A., LORETO V. and SERVEDIO V. D. P., *Proceedings of the 4th European Semantic Web Conference (Innsbruck, Austria) 2007*, pp. 71–84.

[36] SCHENKEL R., CRECELIUS T., KACIMI M., MICHEL S., NEUMANN T., PARREIRA J. X. and WEIKUM G., *Proceedings of the 31st Annual International ACM SIGIR Conference Research and Development in Information Retrieval* (ACM Press, New York) 2008, pp. 523–530.

[37] ZHANG Z.-K., ZHOU T. and ZHANG Y.-C., *Physica A*, **389** (2010) 179.

[38] SHANG M.-S., ZHANG Z.-K., ZHOU T. and ZHANG Y.-C., *Physica A*, **389** (2010) 1259.

[39] SHANG M.-S. and ZHANG Z.-K., *Chin. Phys. Lett.*, **26** (2009) 118903.

[40] SCHEIN A. I., POPESCUL A., UNGAR L. H. and PENNOCK D. M., *Proceedings of 2002 SIGIR Workshop Recommender Systems* (ACM Press, New York) 2002.

[41] SCHEIN A. I., POPESCUL A., UNGAR L. H. and PENNOCK D. M., *Proceedings of the 25th Annual International ACM SIGIR Conference Research and Development in Information Retrieval* (ACM Press, New York) 2002, pp. 253–260.

[42] ZHOU T., SU R.-Q., LIU R.-R., JIANG L.-L., WANG B.-H. and ZHANG Y.-C., *New J. Phys.*, **11** (2009) 123008.

[43] LAMBIOTTEA R. and AUSLOOSB M., *Eur. Phys. J. B*, **50** (2006) 183.

[44] EAGLE N., MACY M. and CLAXTON R., *Science*, **328** (2010) 1029.

[45] SHEPITSEN A., GEMMELL J., MOBASHER B. and BURKE R., *Proceedings of the 2008 ACM Conference Recommender Systems* (ACM Press, New York) 2008, pp. 259–266.