Published in "The European Physical Journal B - Condensed
Matter and Complex Systems 85(8): 286, 2012"
which should be cited to refer to this work.

Heat conduction information filtering via local information of bipartite networks

Q. Guo[1,2], R. Leng[1], K. Shi[1], and J.G. Liu[1,3,a]

[1] Research Center of Complex Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, P.R. China
[2] Department of Physics, University of Fribourg, 1700 Fribourg, Switzerland
[3] CABDyN Complexity Center, Saïd Business School, University of Oxford, Park End Street, Oxford, OX1 1HP, UK

Abstract. Information filtering based on structure properties of user-object bipartite networks is of both theoretical interest and practical significance in modern science. In this paper, we empirically investigate the framework of heat-conduction-based (HC) information filtering [Y.-C. Zhang et al., Phys. Rev. Lett. **99**, 154301 (2007)] in terms of the local node similarity. We compare nine well-known local similarity measures on four real networks. The results indicate that the local-heat-conduction-based similarity has the best accuracy and diversity simultaneously. Embedding the object degree effect into the heat conduction process, we present a new user similarity measure. Experimental results on four real networks demonstrate that the improved similarity measure could generate remarkably higher diversity and novelty results than the state-of-the-art HC information filtering algorithms based on local information, and the accuracy is also increased greatly or approximately unchanged. Since the improved similarity index only need the local information of user-object bipartite networks, it is therefore a strong candidate for potential application in information filtering of large-scale bipartite networks.

1 Introduction

Information filtering based on statistical properties of user-object bipartite networks has been an important issue for both academic and commercial interests in the last decades [1,2]. Comparing with traditional users activities, the development of information systems nowadays not only provide a platform for users to directly collaborate with each other through reviews and ratings, but also provide us a practical way to analyze users' collective behaviors [3]. The increasing popularity of e-commerce and online social networks brings massive amount of accessible information, more than every individual's ability to process. Another impact of Internet and social networks is that any person with access to the Internet can become an author and a publisher. As a consequence, the fast development of Internet and social networks renders the quality of the information extremely diverse and the quantity of information available is enormous [3–5]. Meanwhile, information that is highly important for one individual has no meaning for many others. By predicting users' interests and habits based on their historical records, online recommender systems play an increasingly important role in information filtering, which could increase the sales and enhance users loyalties to web sites [5]. Zhang et al. [6] proposed a new information framework based on the heat conduction process, namely heat-conduction-based (HC) recommendation model. HC model supposes that the objects one user has collected have the recommendation power to help the target user find potential relevant objects. It firstly constructs a propagator matrix $\mathbf{W}^h$, where the element $w_{\alpha\beta}$ denotes the conduction rate from object o_β to o_α. Denote H as the temperature vector of m components: the source components are of temperature one, while the remaining components are of temperature zero. The task of HC model is to find the temperatures of the remain nodes through thermal equilibrium [6] by solving the equation $\mathbf{W}^h H = f$, where f is the flux vector. This is the discrete analog of the function $-\kappa\nabla^2 T(\boldsymbol{r}) = \boldsymbol{\nabla}\cdot\boldsymbol{J}(\boldsymbol{r})$, with the discrete matrix $\mathbf{W}^H$ analog of ∇^2, $H(i)$ playing the role of $-\kappa T(\boldsymbol{r})$ and $f(i)$ playing the role of $\boldsymbol{\nabla}\cdot\boldsymbol{J}(\boldsymbol{r})$ [6]. Inspired by the HC recommendation model, random walks have been successfully embedded into the object similarity measurement [7]. In this kind of algorithm, we need to calculate the object similarity matrix firstly, then generate the recommendation lists according to object similarities. Based on the local information of the user-object bipartite network, collaborative filtering (CF) algorithm is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). If we suppose

[a] e-mail: jianguo.liu@sbs.ox.ac.uk

the users giving ratings to one specific object have the recommendation powers to predict the potential interesting users, the CF algorithm is one kind of the user-based HC model. Liu et al. [8] introduced random walks to calculate the user similarity and found that the modified algorithm has remarkably higher accuracy. By considering the high-order correlation of the users and objects, Liu et al. [9] and Zhou et al. [10] proposed the ultra accurate algorithms, in which the second-order correlation information is used to measure user or object similarities.

Since the HC model is implemented based on matrix operations, it is very time-consuming and cannot be applied for large-scale systems. Despite the random walks [7,8,11], heat conduction [6,12] and hybrid algorithm [1] have been successfully introduced to measure user or object similarities, there are lots of ways to organize the local information of user-object bipartite networks to predict users' interests and habits. It is thus interesting to understand the physics of user-object local interactions and their influence on information filtering.

Bipartite networks has been studied extensively using different approaches ranging from statistics [13] to projection methodology [7]. In this class of networks, there are two different kinds node sets U and O, and only the connection between two kinds of nodes is allowed. As an important class of bipartite networks, user-object networks play the central role in e-commerce systems [13–16]. In addition to the empirical analysis on the user-object bipartite networks [17–19], great efforts have been made to project bipartite networks into monopartite networks [7], how to model bipartite networks [20–23], and how to characterize bipartite networks [24,25]. However thus far a comprehensive picture of the dependence of algorithmic performance of the HC model on network topology is lacking. The reason is twofold: (i) the current works have not linked the the statistical properties of bipartite networks and the current state of development together, while (ii) the statistical physics community has not paid enough attention to the information filtering and information overload problems. Accordingly, dozens of important issues are still insufficiently explored. For example, one may be concerned with how to choose a suitable algorithm given some structural descriptions of a network, such as the degree heterogeneity [18], mixing pattern [26], assortative property [27,28], community structure [25], small-world effect [29,30] and so on. Meanwhile, comparison of the performances of some recommendation algorithms may reveal some of the structural information of the bipartite networks. It is just like the synchronizing process can be used to reveal the underlying community structure [31]. The algorithms based only on local information are generally fast but of lower accuracy, while the ones making use of knowledge of global topology are of higher accuracy yet higher computational complexity.

In this paper, we empirically investigate the framework of the HC information filtering model on the basis of the user similarity, namely user-based HC model. Although the framework is simple, it opens a rich space for exploration since the design of similarity measures is challenging and can be related to very complicated physical dynamics and mathematical theory, such as random walks [7,8] and heat conduction [6,12]. Here we concentrate on local-information-based similarities. We compare nine well-known local measures on four real networks, and the results indicate that the local-heat-conduction-based index has the best overall performance. Motivated by the object degree effect on user similarity measurement [8], we present an improved HC index, namely IHC index. By using the square function to depress popular object effects, the IHC index could further enhance the accuracy, diversity and novelty greatly. Since the objects of four data sets are different, this outstanding performance of HC and IHC indices indicates that enhancing small-degree users' recommendation powers and depressing large-degree object effects could improve accuracy and diversity simultaneously.

2 Heat-conduction-based information filtering model

A recommendation system consists of a set of user nodes, object nodes and connections between two nodes corresponding to an object voted or collected by a user, represented by a bipartite network $G(U, O, E)$. We denote the object set as $O = \{o_1, o_2, \ldots, o_m\}$, the user set as $U = \{u_1, u_2, \ldots, u_n\}$ and the connection set as $E = \{e_1, e_2, \cdots, e_p\}$. The bipartite network can then be represented by an adjacent matrix $A = \{a_{\alpha j}\} \in R^{m,n}$, where $a_{\alpha j} = 1$ if object o_α is collected by user u_j, and $a_{\alpha j} = 0$ otherwise.

The HC model could be used in two ways: item-based and user-based models. The item-based HC model supposes that the objects one user has collected have the power to find the potential relevant objects to the target user. The general framework of the item-based HC model is as follows: (i) construct the weighted object network (i.e. determine the matrix W) from the known user-object relations; (ii) determine the initial resource vector $\mathbf{f}$ for each user; (iii) get the final resource distribution via

$$\mathbf{f}' = W\mathbf{f}; \tag{1}$$

(iv) recommend those uncollected objects with highest final resource. Note that the initial configuration $\mathbf{f}$ is determined by the user's personal information, thus for different users, the initial configuration is different. To the user-based HC model, for a given object o_α, the ith element of $\mathbf{f}^\alpha$ should be zero if $a_{\alpha i} = 0$. That is to say, one should not put any recommendation power (i.e. resource) onto an unrated user. The simplest case is to set a uniform initial configuration as $f_i^\alpha = a_{\alpha i}$. Under this configuration, all users rated object o_α have the same recommendation power.

Based on the user-object matrix A, the object similarity network has been calculated by the random walks [7,8,11,32], heat conduction [6,12], and hybrid [1] measures. However, several similarity measures have been successfully used in the CF model [33–35] and link predictions [36], such as common neighbor index, Jaccard index [37], Searenson index [38], hybrid index [39], second-order index [9] and so on. The CF model supposes that the users who have similar tastes or interests prefer to have similar interests in the future [40–44], which is similar with the user-based HC model. Some of these measures are not directly proposed to bipartite networks, for example, Salton index [45] is proposed by Salton in 1983 and has a long history of study in citation networks. However, this idea could be directly introduced to measure the node similarities of bipartite networks. The measure of user similarity is very important for both user-based and item-based HC models. Since most of previous works focus on item-based HC model [6–8], the understanding of the similarity effect on user-based HC model is lacking. In this paper, we extensively investigate nine well-known similarity measures based on the local information of bipartite networks and compare their performances on four different data sets. The detail definitions would be introduced in the next section.

3 Nine similarity measures based on the local information

In this section, we introduce nine similarity measures. All these measures are based on the local structural information of user-object bipartite networks contained in the testing set. We firstly give a brief introduction of each measure as follows.

(AA) Adamic-Adar Index. For a user u_i, let $\Gamma(u_i)$ denotes the object set user u_i has selected or rated. This index refines the simple counting of common neighbours by assigning less connected neighbours more weight, and is defined as [46]

$$w_{ij} = \sum_{z \in \Gamma(u_i) \bigcap \Gamma(u_j)} \frac{1}{\log k_z}, \tag{2}$$

where k_z denotes the degree of object z.

(CN) Common Neighbours. By common sense, two users, u_i and u_j, are more likely to have similar interests if they have many commonly rated objects. The simplest measure of the neighbourhood overlap is the directed count, namely

$$w_{ij} = |\Gamma(u_i) \bigcap \Gamma(u_j)|. \tag{3}$$

(HPI) Hub Promoted Index. This index is proposed for quantifying the topological overlap of pairs of substrates in metabolic networks [47], and is defined as

$$w_{ij} = \frac{|\Gamma(u_i) \bigcap \Gamma(u_j)|}{\min\{k_{u_i}, k_{u_j}\}}. \tag{4}$$

According to the definition, the similarities to large-degree users are likely to be assigned high scores since the denominator is determined by the lower user degree only.

(JAC) Jaccard Index. This index was proposed by Jaccard [37] over a hundred years ago, and is defined as

$$w_{ij} = \frac{|\Gamma(u_i) \bigcap \Gamma(u_j)|}{|\Gamma(u_i) \bigcup \Gamma(u_j)|}. \tag{5}$$

(LHN) Leicht-Holme-Newman Index. This index assigns high similarity to node pairs that have many common neighbours not comparing with the possible maximum, but with the expected number of such neighbours [48]. It is defined as

$$w_{ij} = \frac{|\Gamma(u_i) \bigcap \Gamma(u_j)|}{k_{u_i} \times k_{u_j}}, \tag{6}$$

where the denominator, $k_{u_i} \times k_{u_j}$, is proportional to the expected number of common neighbours of user u_i and u_j in the configuration model.

(SEA) Searensen Index. This index is used mainly for ecological community data [38], and is defined as

$$w_{ij} = \frac{2|\Gamma(u_i) \bigcap \Gamma(u_j)|}{k_{u_i} + k_{u_j}}. \tag{7}$$

(SAL) Salton Index. The Salton index [45] is defined as

$$w_{ij} = \frac{|\Gamma(u_i) \bigcap \Gamma(u_j)|}{\sqrt{k_{u_i} \times k_{u_j}}}, \tag{8}$$

where $k_{u_i} = |\Gamma(u_i)|$ denotes the degree of user u_i. The Salton index is also called the cosine similarity in the literature.

(RW) Random-walk-based index. To the similarity from u_j to u_i, a certain amount of resource is associated with user u_j, and the weight w_{ij} represents the proportion of the resource u_j would like to distribute to u_i. The measure of This directed measure is defined as follows [7,8]

$$w_{ij} = \frac{1}{k_{u_j}} \sum_{l=1}^{m} \frac{a_{li} a_{lj}}{k_{o_l}}. \tag{9}$$

(HC) Local-Heat-conduction-based index. When calculating the directed similarity from user u_j to u_i, user u_j is set as a heat resource and the temperature is set as 1, then the heat will diffuse from u_j to his collected objects, and diffused back to user u_i [6,12]. The final temperature user u_i received is defined as follows

$$w_{ij} = \frac{1}{k_{u_i}} \sum_{l=1}^{m} \frac{a_{li} a_{lj}}{k_{o_l}}. \tag{10}$$

4 Experimental results

4.1 Data

In this paper, we consider four representative networks drawn from disparate fields: (1) *MovieLens* data set consists of 6040 users on 3592 movies(objects) and rating scale from *one* (i.e., worst) to *five* (i.e., best). (2) *Netflix* data

Table 1. Basic statistics of four data sets, namely MovieLens, Netflix, Delicious and Amazon. m and n represent the number of users and objects nodes, respectively. E is the number of the actual user-object entries in each data set. $E/(m \times n)$ is an equation used to calculate the sparsity of the data set.

Data sets	m	n	E	$E/(m \times n)$
MovieLens	6040	3592	750 000	0.0346
Netflix	10 000	6000	701 947	0.0117
Delicious	10 000	232 657	1 233 997	5.3×10^{-4}
Amazon	34 808	80 774	855 857	3.04×10^{-4}

set is a random sample of the whole records of user activities in Netflix.com, which consists of 6000 movies and 10 000 users and 824 802 ratings. We apply a coarse graining method: a movie is considered to be collected by a user only if the rating is larger than *two*. In this way, the Movielens data has 750 000 edges, and the Netflix data has 701 947 edges. (3) *Delicious* data set is a random sample of the whole records of user selections in Del.icio.us. (4) *Amazon* data set is also a random sample of the whole records. The statistical properties of the four data sets are shown in Table 1. To test the performance of these nine user similarity measures, the data set E is randomly divided into two parts $E = E^T \cup E^P$, where the training set E^T is treated as the known information, contains p percent of the data, and the remaining $1-p$ part is set as the probe set E^P, whose information is not allowed to be used for prediction.

4.2 Metrics

Accuracy. An accurate method will put preferable objects in higher places. Here we use *average ranking score* [7] to measure the ability of the measure to produce a good uncollected object ranking list that matches the target user's preference. For an arbitrary user u_i, if the object o_α is not collected by user u_i, while the entry u_i-o_α is in the probe set, we use the rank of o_α in the recommendation list to evaluate the accuracy. For example, if there are 8 uncollected objects for user u_i, and object o_α is ordered at the 3rd place, we say the position of o_α is 3/8, denoted by $r_{i\alpha} = 0.375$. Since the probe entries are actually collected by users, a good algorithm is expected to give high recommendations to them, leading to a small $r_{i\alpha}$. Therefore, the mean value of the positions, averaged over all the entries in the probe set, can be used to evaluate the algorithmic accuracy

$$\langle r\rangle = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\sum_{(u_i,o_\alpha)\in E^p} r_{i\alpha}}{m-k_{u_i}}\right), \tag{11}$$

where E^p is the edge set existing in the probe set, and m is the number of objects in the system. The smaller the average ranking score, the higher the algorithmic accuracy, and vice verse.

Diversity. The analysis results on Facebook data set shows that, besides the common interests, users of online social networks also have their specific tastes and interests [3], leading to diverse selection behaviors. Liu et al. [12] found that users' tastes on Movielens and Netflix data could also be divided into two categories: common interests and specific interests [11]. Therefore, besides accuracy, the diversity of the recommendation list is taken into account to evaluate the algorithmic performance. In general, most of the users would not show negative altitude to popular objects, therefore, ranking popular objects at the top part of the recommendation lists would generate higher accuracy. However, personalized recommendation algorithms should not only present accurate prediction but also generate different recommendations to different users according to their specific tastes or habits. The diversity of user u_i and u_j's recommendation lists can be quantified by the Hamming distance,

$$S_{ij} = 1 - \langle Q_{ij}(L)\rangle / L, \tag{12}$$

where L is the length of the recommendation list and Q_{ij} is the number of overlapped objects in user u_i and u_j's recommendation lists. The average Hamming distance S could be used to measure the algorithmic diversity. The largest $S = 1$ indicates recommendations to all users are completely different, in other words, the system has highest diversity. While the smallest $S = 0$ means all of recommendations are exactly the same.

Popularity. A accurate and diverse recommender system is expected to help users find the niche or unpopular objects which is hard for them to identify yet match their preferences. The metric *popularity* is introduced to quantify the capacity of an algorithm to generate unexpected recommendation lists, which is defined as the average collected times over all recommended objects

$$\langle k\rangle = \frac{1}{n}\sum_{i}\left(\frac{1}{L}\sum_{o_\alpha\in O_i^L} k_{o_\alpha}\right), \tag{13}$$

where O_i^L is user u_i's the recommendation list with length L. A smaller average degree $\langle k\rangle$, corresponding to the less popular objects, are preferred since those small-degree objects are hard to be found by users themselves.

F-measure. F-measure has been extensively used to in the information retrieval and natural language processing communities [49,50]. In fact, the F-measure is a harmonic mean of *recall* (R) and *precision* (P). Since all of accuracy $\langle r\rangle$, diversity S and popularity $\langle k\rangle$ are very important to measure information filtering algorithm'c performance, we apply the similar idea to construct F-measure. Because the smaller value of $\langle r\rangle$ indicates higher accuracy, while larger hamming distance S means higher diversity, meanwhile, small popularity $\langle k\rangle$ indicates that the algorithm could present fresh new information to the user. Therefore, F-measure could be given in a harmonic mean way

$$F = \frac{3}{\frac{1}{1-\langle r\rangle}+\frac{1}{S}+\frac{1}{(k_{\max}-\langle k\rangle)/k_{\max}}}, \tag{14}$$

where $k_{\max}$ denotes the largest object degree in the system. When the accuracy $\langle r\rangle$ equals to 0, diversity S equals

Table 2. Algorithmic performances for MovieLens, Netflix, Delicious and Amazon data sets when $p = 0.9$, including the accuracy $\langle r\rangle$, diversity S and popularity $\langle k\rangle$ corresponding to the length of recommendation list $L = 10$. The abbreviations, AA, CN, HPI, JAC, LHN, SEA, SAL, RW, HC and IHC stand for Adamic-Adar index, Common Neighbours, Hub Promoted Index, Jaccard Index, Leicht-Holme-Newman Index, Searensen Index, Salton Index, Random-walks-based Index, Local-Hear-Conduction Index and improved HC index respectively. The entries corresponding to the highest accuracies among these nine measures are emphasized in black. Each number is obtained by averaging over five runs with independently random division of training set and probe set.

	MO			NE			DE			AM		
	$\langle r\rangle$	$\langle k\rangle$	S	$\langle r\rangle$	$\langle k\rangle$	S	$\langle r\rangle$	$\langle k\rangle$	S	$\langle r\rangle$	$\langle k\rangle$	S
AA	0.1336	**1964**	0.5816	0.0923	**3007**	0.6053	0.3821	**328**	0.8378	0.2854	403	0.8636
CN	0.1217	1995	0.5649	0.0590	3124	0.5760	0.2373	567	0.5564	0.1356	573	0.8283
HPI	0.1209	1999	0.5667	0.0590	3147	0.5892	0.2323	576	0.5589	0.1343	576	0.8326
JAC	0.1176	1991	0.5881	0.0573	3117	0.6397	0.2206	547	0.6323	0.1318	539	0.8664
LHN	0.1173	1992	0.5864	0.0576	3102	0.6281	0.2212	522	0.6665	0.1324	528	0.8751
SEA	0.1186	1994	0.5833	0.0577	3124	0.6338	0.2207	550	0.6278	0.1319	543	0.8641
SAL	0.1193	1997	0.5764	0.0580	3140	0.6158	0.2219	568	0.5877	0.1314	555	0.8606
RW	0.1143	1975	0.6003	0.0525	3074	0.5876	0.2178	497	0.7033	0.1351	443	0.9175
HC	**0.1080**	1968	**0.6242**	**0.0499**	3089	**0.6564**	**0.2029**	401	**0.8399**	**0.1313**	**324**	**0.9620**
IHC	0.0999	1833	0.7306	0.0487	2658	0.8099	0.2126	237	0.9634	0.1356	140	0.9959

to 1 and $\langle k\rangle = 1$, F-measure would equal to 1. On the contrary, F-measure equals to 0, when $\langle r\rangle = 0$, $S = 0$ or $\langle k\rangle = k_{\max}$. The larger the F-measure, the better the algorithmic performance, and vice versa.

4.3 Simulation results

Table 2 shows the effects of local similarity measures on user-based HC information filtering model for MovieLens, Netflix, Delicious and Amazon data sets, from which we find that both accuracy and diversity of HC index are the best one in all nine measures. For MovieLens, Netflix and Delicious data sets, the AA index could generate the lowest popularity, while for Amazon, the lowest $\langle k\rangle$ is obtained by the HC index.

According to the information which have been used in the measures, including users or objects' degree and the number of common neighbors, these nine measures could be divided into two categories: directed and undirected similarities. Although CN, SAL, JAC, SEA, HPI, LHN indices are designed in different ways, they didn't take into account the similarity direction, which are only determined by the number of common neighbors and two neighbor users' degrees, who have at least one common rated object. It should be noticed that although the AA index is undirected, the popular object effects are depressed by using the log function, which means that if two uses commonly rated small-degree objects, their similarity would be larger than other user pairs rated large-degree objects. In reality, there are lots of channels for all users to obtain the popular objects' information, while it is hard to identify users' specific interests or tastes, therefore, popular objects' effects should be depressed. Besides these seven measures, RW and HC indices not only take into account the similarity direction but also depress popular object effects. Extensively experimental results [1] on the item-based HC model indicates that the RW index could generate high accuracy and low diversity recommendation by emphasizing large-degree objects' recommendation powers. On the contrary, by highlighting small-degree object effects, the HC model is of high diversity and low accuracy. The experimental results in Table 2 show that, different with the item-based HC model, the HC index of the user-based HC model could simultaneously generate best accuracy and diversity recommendations for different data sets.

5 Improved HC measure and experimental results

The extensively experimental results on four different data sets indicate that the HC index outperforms all undirected measures, as well as the RW index in accuracy and diversity simultaneously. The common property of RW and HC indices indicates that depressing popular object effects could enhance information filtering performance, however, we don't know whether we could improve the recommendation performance by further depressing large-degree objects' effects. Inspired by object degree effects on the user-based HC model [8], we argue that it deserves the further investigation about the influence of popular objects on item-based HC. As a result, we rewrite the HC index as

$$w_{ij}^{t} = \frac{1}{k_{u_i}} \sum_{l=1}^{m} \frac{a_{li}a_{lj}}{k_{o_l}^{\alpha}}, \qquad (15)$$

where α is a free parameter to investigate when the suppression of popular objects starts to be effective or ineffective respectively in four data sets. As shown in Figure 1, the average ranking score $\langle r\rangle$ as a function of α, one can find that the average ranking score in each subplot has a clear minimum. According to the benchmark MovieLens data set, the parameter $\alpha_{opt} = 2.0$ when the average ranking score $\langle r\rangle$ reaches the minimal value, thus we set the

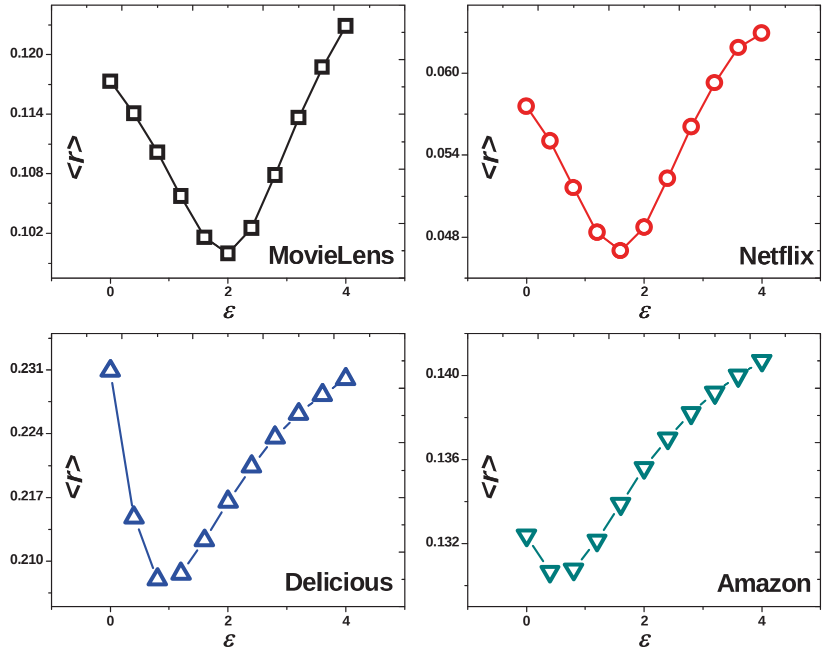

Fig. 1. (Color Online) The average ranking score $\langle r \rangle$ for MovieLens, Netflix, Delicious and Amazon data sets as a function of α. According to the results above, one can find each subplot has a clear minimum around $\alpha = 2.0$ for MovieLens, $\alpha = 1.6$ for Netflix, $\alpha = 0.8$ for Delicious and $\alpha = 0.4$ for Amazon, respectively. All the numerical results are obtained by averaging over five independent runs with random data division of training and probe set.

parameter α is equal to 2.0 for all data sets. Due to the different statistical properties of each data set, for example sparsity, the optimal parameter α is distinct actually. In spite of the above fact, we find that the optimal parameter for Netflix is close to 2.0. At the same time, the optimal parameters for Amazon and Delicious data sets have a large difference, however, the average ranking score $\langle r \rangle$ is stable relatively. In conclusion, we finally set the denominator, the object degree, as the squared term, and present an improved HC index, namely IHC index, which could be written as

$$w_{ij} = \frac{1}{k_{u_i}} \sum_{l=1}^{m} \frac{a_{li}a_{lj}}{k_{o_l}^2}. \qquad (16)$$

Experimental results show that the IHC index could greatly improve the diversity S, decrease the popularity $\langle k \rangle$ and increase the accuracy for MovieLens, Netflix and Delicious data sets. Although the accuracy for Amazon data set is not good as the one obtained by the HC index, it is also approximately unchanged.

Figure 2 shows the results of F-measure for MovieLens, Netflix, Delicious and Amazon data sets, from which one can see that F-measure of the six undirected measures are close for different data sets. Meanwhile, the results of direct measures, including IHC, HC and RW indices, are much larger than the ones obtained by undirected ones. It should be emphasized that the performances of the IHC index are much better than HC and RW indices. We also notice that the F-measure of the AA index is much better than the ones of undirected measures, which may be caused by the fact that the average object degree $\langle k \rangle$ is very large. The above results suggest that enhancing small-degree users' recommendation power and depressing large-degree object effects are two effective ways to enhance information filtering performance. Since four data sets are collected from different web sets and the characteristics of the objects are different, therefore, the above results suggest that there are somehow common user collective behavior patterns in online systems, which could be investigated in the future work.

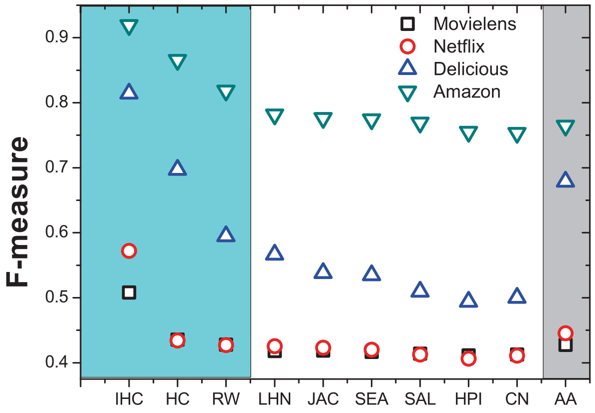

Fig. 2. (Color Online) F-measures of ten similarity measures for MovieLens, Netflix, Delicious and Amazon data sets, from which one can find that the largest one is obtained by the IHC index, which is much larger than the ones of HC and RW indices. While the results of six undirected measures, including LHN, JAC, SEA, SAL, CN, LHN and HPI indices, are close with each other. The results of the AA index are better than the above six undirected measures, but are also worse than the ones obtained by the IHC index.

6 Conclusion and discussions

In this paper, based on the local information of user-object bipartite networks, we empirically investigate the effects of nine similarity measures on user-based HC information filtering model. Experimental results indicate that the undirected measure, local-heat-conduction-based index, performs the best accuracy and diversity for different data sets. Combining the HC measure and the square function, we present an improved HC (IHC) measure. Experimental results show that the accuracy $\langle r\rangle$, diversity S and popularity $\langle k\rangle$ could be greatly improved for MovieLens, Netflix and Delicious data sets. Although the accuracy of the IHC index for Amazon data set is approximately unchanged, both popularity $\langle k\rangle$ and diversity S are increased greatly. According to the IHC measure definition, small-degree users' recommendation powers are enhanced and large-degree object effects are depressed by the square function. Therefore, we could say that there are two general ways to improve information filtering performance. Firstly, depressing large-degree object effects. Secondly, increasing small-degree users' recommendation powers. We here strongly recommend the IHC measure to relevant applications and theoretical analyses, not only for its good performance, but also for its simplicity and grace. In real applications, the traditional HC model based on global matrix may be less efficient for they need long time and huge memory, while the local information based measures only exploited limited information may be less effective for their low accuracies. A properly designed algorithm can provide a good tradeoff just like the IHC index presented in this paper. Although the HC information filtering framework is very simple, it opens a rich space for investigation since in principle, all measures can be embedded into this framework. Besides these measures discussed in this paper, a number of similarities are based on the global structural information, such as the pseudoinverse of the Laplacian matrix [51], the transferring similarity [52], the PageRank index [53] and so on. These measures may give better performances than the local ones, however, the calculation of such measures, including determination of the optimal parameters for specific networks, is of high complexity, and thus infeasible for huge-size networks.

Up to now, although random walks [7,11] and the local heat conduction [12] process have been successfully embedded into the HC information filtering model, we lack systematic comparison and understanding of the performances of these measures, which is set as our future works. Empirical analysis on more known and newly proposed similarity measures as well as more real networks is very valuable for deeply understanding users' collective behavior patterns and building up knowledge and experience. A clear picture of this issue can be completed by putting together of many fragments from respective empirical studies. Besides the empirical results, an alternative way is to build artificial network models with controllable topological features. In this way, we could have a clear picture on the unknown and uncontrollable ingredients which are always mixed together in real networks.

We acknowledge *GroupLens* Research Group for providing us *MovieLens* data and the Netflix Inc. for *Netflix* data. This work is supported by NSFC (10905052, 70901010, 71071098 and 71171136), the Innovation Program of Shanghai Municipal Education Commission (11ZZ135, 11YZ110), JGL is supported by the Shanghai Leading Discipline Project (S30501), Shanghai Rising-Star Program (11QA1404500) and Key Project of Chinese Ministry of Education (211057).

References

1. T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J.R. Wakeling, Y.-C. Zhang, Proc. Natl. Acad. Sci. USA **107**, 4511 (2010)
2. G. Adomavicius, A. Tuzhilin, IEEE Trans. Knowl. Data Eng. **17**, 734 (2005)
3. J.P. Onnela, F. Reed-Tsochas, Proc. Natl. Acad. Sci. USA **107**, 18375 (2010)
4. F.R. Lynch, *The Diversity Machine: The Drive to Change the "White Male Workplace"* (Free Press, 1997)
5. G. Linden, B. Smith, J. York, IEEE Internet Computing **7**, 76 (2003)
6. Y.-C. Zhang, M. Blattner, Y.-K. Yu, Phys. Rev. Lett. **99**, 154301 (2007)
7. T. Zhou, J. Ren, M. Medo, Y.-C. Zhang, Phys. Rev. E **76**, 046115 (2007)
8. J.-G. Liu, B.-H. Wang, Q. Guo, Int. J. Mod. Phys. C **20**, 285 (2009)
9. J.-G. Liu, T. Zhou, H.-A. Che, B.-H. Wang, Y.-C. Zhang, Phys. A **389**, 881 (2010)
10. T. Zhou, R.-Q. Su, R.-R. Liu, L.-L. Jiang, B.-H. Wang, Y.-C. Zhang, New J. Phys. **11**, 123008 (2009)
11. J.-G. Liu, K. Shi, Q. Guo, Phys. Rev. E **85**, 016118 (2012)
12. J.-G. Liu, T. Zhou, Q. Guo, Phys. Rev. E **84**, 037101 (2011)
13. M. Kitsak, D. Krioukov, Phys. Rev. E **82**, 026114 (2011)
14. Z. Huang, H. Chen, D. Zeng, ACM Trans. Inf. Syst. **22**, 116 (2004)
15. M. Faloutsos, P. Faloutsos, C. Faloutsos, Comput. Commun. Rev. **29**, 251 (1999)
16. A. Broder, R. Kumar, F. Moghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Comput. Netw. **33**, 309 (2000)
17. Z. Huang, D.D. Zeng, H. Chen, Manage. Sci. **53**, 1146 (2007)
18. M.-S. Shang, L. Lü, Y.-C. Zhang, T. Zhou, Europhys. Lett. **90**, 48006 (2010)
19. Y.-L. Wang, T. Zhou, J.-J. Shi, J. Wang, D.-R. He, Phys. A **388**, 2949 (2009)
20. J. Ohkubo, K. Tanaka, T. Horiguchi, Phys. Rev. E **72**, 036120 (2005)
21. M.L. Goldstein, S.A. Morris, G.G. Yen, Phys. Rev. E **71**, 026108 (2005)
22. J.-L. Guillaume, M. Latapy, Phys. A **371**, 795 (2006)
23. E. Birmelé, Discr. Appl. Math. **157**, 2267 (2009)
24. M. Latapya, C. Magnienb, N.D. Vecchio, Social Netw. **30**, 31 (2008)
25. M.J. Barber, Phys. Rev. E **76**, 066102 (2007)
26. M.E.J. Newman, Phys. Rev. Lett. **89**, 208701 (2002)
27. M.E.J. Newman, Phys. Rev. E **66**, 016132 (2001)
28. T. Zhou, J.-G. Liu, B.-H Wang, Chin. Phys. Lett. **23**, 2327 (2006)
29. D.J. Watts, S.H. Strogatz, Nature **393**, 440 (1998)
30. A.-L. Barabási, R. Albert, Science **286**, 509 (1999)
31. A. Arenas, A. Díaz-Guilera, C.J. Pérez-Vicente, Phys. Rev. Lett. **96**, 114102 (2006)
32. J.-G. Liu, Q. Guo, Y.-C. Zhang, Phys. A **390**, 2414 (2011)

33. X. Su, T.M. Khoshgoftaar, Advances in Artificial Intelligence 421425 (2009)
34. J.L. Herlocker, J.A. Konstan, K. Terveen, J.T. Riedl, ACM Trans. Inf. Syst. **22**, 5 (2004)
35. J.A. Konstan, B.N. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon, J. Riedl, Commun. ACM **40**, 77 (1997)
36. T. Zhou, L. Lü, Y.-C. Zhang, Eur. Phys. J. B **71**, 623 (2009)
37. P. Jaccard, Bulletin de la Société Vaudoise des Sciences Naturelles **37**, 547 (1901)
38. T. Sorensen, Biol. Skr. **5**, 1 (1948)
39. X. Pan, G.-S Deng, J.-G. Liu, Chin. Phys. Lett. **27**, 068903 (2010)
40. J.L. Herlocker, J.A. Konstan, A. Borchers, J. Riedl, in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Theoretical Models* (1999), pp. 230–237
41. H. Luo, C. Niu, R. Shen, C. Ullrich, Mach. Learn. **72**, 231 (2008)
42. B. Sarwar, G. Karypis, J. Konstan, J. Riedl, in *Proceedings of the 10th International World Wide Web Conference* (2001), pp. 285–295
43. M. Deshpande, G. Karypis, ACM Trans. Inf. Syst. **22**, 143 (2004)
44. M. Gao, Z.F. Wu, F. Jiang, Inf. Proc. Lett. **111**, 440 (2011)
45. G. Salton, M.J. McGill, *Introduction to Modern Information Retrieval* (McGraw-Hill Inc., New York, 1986)
46. L.A. Adamic, E. Adar, Social Netw. **25**, 211 (2003)
47. E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.-L. Barabási, Science **297**, 1553 (2002)
48. E.A. Leicht, P. Holme, M.E.J. Newman, Phys. Rev. E **73**, 026120 (2006)
49. Y. Yang, X. Liu, in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval* (1999)
50. C.J. Rijsbergen, *Information Retireval* (Butterworths, London, 1979)
51. F. Fouss, A. Pirotte, J.-M. Renders, M. Saerens, IEEE Trans. Knowl. Data. Eng. **19**, 355 (2007)
52. D. Sun, T. Zhou, J.-G. Liu, R.-R. Liu, C.-X. Jia, B.-H. Wang, Phys. Rev. E **80**, 017101 (2009)
53. S. Brin, L. Page, Comput. Netw. ISDN Syst. **30**, 107 (1998)