An improved recommendation algorithm via weakening indirect linkage effect*

Chen Guang(陈 光), Qiu Tian(邱 天)†, and Shen Xiao-Quan(沈小泉)

School of Information Engineering, Nanchang Hangkong University, Nanchang 330063, *China*

(Received 24 October 2014; revised manuscript received 2 February 2015; published online 29 May 2015)

We propose an indirect-link-weakened mass diffusion method (IMD), by considering the indirect linkage and the source object heterogeneity effect in the mass diffusion (MD) recommendation method. Experimental results on the MovieLens, Netflix, and RYM datasets show that, the IMD method greatly improves both the recommendation accuracy and diversity, compared with a heterogeneity-weakened MD method (HMD), which only considers the source object heterogeneity. Moreover, the recommendation accuracy of the cold objects is also better elevated in the IMD than the HMD method. It suggests that eliminating the redundancy induced by the indirect linkages could have a prominent effect on the recommendation efficiency in the MD method.

Keywords: bipartite network, mass diffusion, recommender system, indirect linkage effect

PACS: 89.75.Hc, 87.23.Ge, 05.70.Ln **DOI:** 10.1088/1674-1056/24/7/078901

1. Introduction

Internet has brought a great convenience to people for providing sufficient information, which on the other hand makes one in the dilemma of choosing what one wants when facing the enormous information. For example, how to quickly find out a favorable movie from the huge number of online movies. As a powerful tool, recommender engine emerges to make predictions for users according to their past activities, which therefore greatly helps people filter out the redundant information.[1]

Various recommendation algorithms have been proposed. The most widely applied algorithms include the so-called collaborative filtering algorithm and its improved ones.[2–6] Another line of algorithms is the content-based algorithm, and its extensive versions.[7–10] To enhance the recommendation efficiency, different accessorial information is also considered in some algorithms, such as the social tags.[11–16] Recently, the network-based recommendation algorithms have attracted great interest of physicists,[16–18] for the development of the complex network theory.[19,20] Different physical processes have been introduced to the information filtering, such as the heat conduction process. Inspired by a heat conduction algorithm[21] and a mass diffusion (MD) algorithm,[22] numerous network-based recommendation algorithms have been proposed. Liu *et al.* proposed a biased heat conduction method by considering the heterogeneity of the target objects,[23] which greatly improves the recommendation accuracy, compared with the original heat conduction method. The heterogeneity of the source objects is further considered,[24] which well alleviates the recommendation bias between the cold and hot objects, and further improves both the recommendation accuracy and diversity. A hybrid method of heat conduction and mass diffusion (HHP) optimizes the diffusion process by introducing a hybridization parameter, which well resolves the dilemma of recommendation accuracy and diversity.[25] Based on the HHP, an improved hybrid version is proposed by considering the object heterogeneity, and it greatly enhances the recommendation accuracy of the cold objects.[26] Scaling behavior is observed in the recommender systems, and the scaling-based algorithm shows better performance not only on the recommendation accuracy of the cold objects, but also on the recommendation diversity.[27] Heterogeneity effect of the initial configuration is investigated using different methods,[28,29] and the higher-order correlation is eliminated in the mass diffusion method,[30] and in the HHP method.[31] These improved methods are found to be beneficial for the recommendation efficiency in different aspects.

In this article, we propose an indirect-link-weakened MD method (IMD), by weakening both the higher-order redundancy induced by the indirect linkages and the object heterogeneity effect in the MD method. We try to understand which factor could be important to the recommendation efficiency for a particular algorithm. By comparing the recommendation results of four algorithms, the indirect linkage effect is found to have a prominent impact on the recommendation performance in the mass diffusion method.

2. Algorithms

The network-based recommender system can be described by the bipartite graph,[32,33] which is composed of

*Project supported by the National Natural Science Foundation of China (Grant No. 11175079) and the Young Scientist Training Project of Jiangxi Province, China (Grant No. 20133BCB23017).

†Corresponding author. E-mail: tianqiu.edu@gmail.com

© 2015 Chinese Physical Society and IOP Publishing Ltd

the user set $U = \{u_1, u_2, \ldots, u_i, \ldots, u_m\}$ and the object set $O = \{o_1, o_2, \ldots, o_\alpha, \ldots, o_n\}$. If an object o_α is collected by a user u_i, then add a link between them. Therefore, the recommender system can be characterized by an adjacent matrix $\boldsymbol{A} = a_{i\alpha}$, with $a_{i\alpha}$ being 1 for the linked user–object pairs; otherwise, being 0.

Assume each object has an initial resource, which could spread from one object to another along the link between the user and object. Here we name the object disseminating resource as the source object, and the object receiving resource as the target object. The resource reallocation process can be formulated as

$$f' = Wf, \tag{1}$$

where W is the transformation matrix, characterizing the resource diffusion process; f is the initial resource of the object, which can be assigned to 1 or 0 for simplicity, depending on whether it is collected by the user or not, i.e., $\boldsymbol{f}_0^i(o_\alpha) = a_{i\alpha}$; f' is the final resource that the object obtains. For each user, rank the resources of his/her uncollected objects according to the descending order of the resources, the objects with the top L resources would then be recommended to the user.

The transformation matrix $\boldsymbol{W}$ is therefore very essential to the resource reallocation. In the MD algorithm,[22] an equal probability spreading process is introduced, as illustrated in Fig. 1. At first, each object distributes the resource to its neighboring users with an equal probability. For example, the first object β who has an initial resource of 1 as the source object would disseminate the resource to its two neighboring users, i.e., the first and second users, with an equal probability. Therefore, the two users would obtain $1/2$ resource from the object β, respectively. Summing up the resources from all the linked objects, the user then redistributes the total resource to his/her neighboring objects, also with the equal probability. For example, the user i who has a total resource of 1 would feedback the resource to i's two neighboring objects with an equal probability as well. The final resource of the object is the total resource obtained from all its neighboring users. The transformation matrix of the MD algorithm is formulated by

$$W_{\alpha\beta}^{\mathrm{MD}} = \frac{1}{k_\beta} \sum_{i=1}^{m} \frac{a_{\alpha i} a_{\beta i}}{k_i}, \tag{2}$$

where k_β is the degree of the source object o_β, and k_i is the degree of user u_i. Compared with the user-similarity based collaborative filtering algorithm, the MD method greatly improves the recommendation accuracy.

Previous studies have shown that the object heterogeneity has a great impact on the recommendation accuracy.[23,24] It has been reported that a more homogeneous object distribution may result in a better recommendation accuracy.[28] Based on the MD method, a heterogeneity-weakened MD algorithm (HMD) is introduced in Ref. [28], by relieving the source object heterogeneity with a tunable parameter η. The transformation matrix of the HMD can be formulated by

$$W_{\alpha\beta}^{\mathrm{HMD}} = \frac{1}{k_\beta^\eta} \sum_{i=1}^{m} \frac{a_{\alpha i} a_{\beta i}}{k_i}. \tag{3}$$

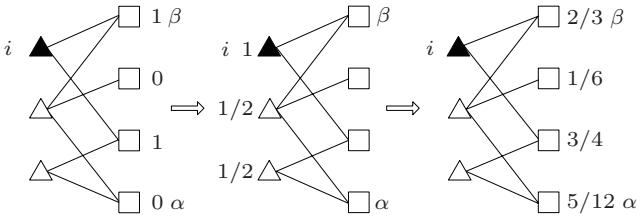


Fig. 1. An illustration of the resource transformation of the MD method.

Furthermore, the redundancy induced by the indirect linkage is considered in the MD method. The redundant correlation effect has been investigated in different algorithms.[30,31] Properly eliminating the redundant correlations can well elevate the algorithm performance. To depict how the indirect linkage works in the bipartite network based recommender system, we present the simplest example in Fig. 2. As shown in Fig. 2(a), for the particular objects α and β, which are both collected by the same user i, the similarity between α and β can be described as L_1 along with the direct linkage $\alpha - i - \beta$. However, the correlations between objects could be much complex in real systems. Except the direct linkage, there are numerous indirect linkages. The simplest example is shown in Fig. 2(b). Besides the direct correlation via the user i, the objects α and β could also be correlated via some other media objects, such as the media object γ in Fig. 2(b). That is, the object α is collected by user i, and the object β is collected by user j, while users i and j both collect the object γ. Therefore, there is a second-order correlation L_2 between object α and β along the indirect linkage $\alpha - i - \gamma - j - \beta$, which would lead to the linkage redundancy of the system. In order to eliminate the redundancy induced by the indirect linkage, we propose the indirect-link-weakened MD method (IMD), with its transformation matrix formulated by

$$\begin{aligned} W_{\alpha\beta}^{\mathrm{IMD}} &= W_{\alpha\beta}^{\mathrm{HMD}} - \lambda \sum_{\gamma=1}^{n} W_{\alpha\gamma}^{\mathrm{HMD}} W_{\gamma\beta}^{\mathrm{HMD}} \\ &= \frac{1}{k_\beta^\eta} \sum_{i=1}^{m} \frac{a_{\alpha i} a_{\beta i}}{k_i} - \frac{\lambda}{k_\beta^\eta} \sum_{\gamma=1}^{n} \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{a_{\alpha i} a_{\gamma i} a_{\gamma j} a_{\beta j}}{k_i k_j k_\gamma^\eta}, \end{aligned} \tag{4}$$

where λ and η are two tunable parameters. Therefore, the IMD method can be taken as a combination of weakening both the indirect linkage[30] and the object heterogeneity effect.[28] When tuning λ and η to the proper values, the IMD method would achieve the optimal recommendation results.

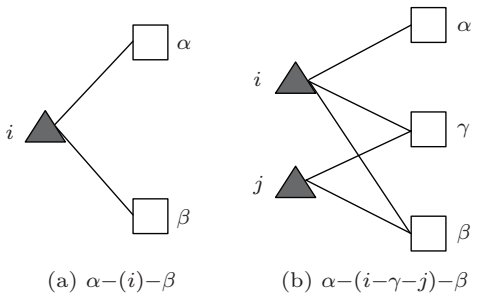


Fig. 2. Illustration of the indirect linkage effect.

To better evaluate the performance of the proposed methods, the traditional user-similarity based collaborative filtering algorithm (CF) is also investigated for comparison, by calculating the similarity between users i and j via a cosine similarity,

$$s_{ij} = \frac{\sum_{\alpha=1}^{n} a_{i\alpha} a_{j\alpha}}{\sqrt{k_i k_j}}. \tag{5}$$

The score of the user i to object α is $f_{\alpha i} = \sum_{j \neq i} s_{ij} a_{\alpha j}$. Ranking the scores of the user's uncollected objects as the descending order, then the objects with the top L scores would be recommended to the user i.

3. Datasets and evaluators

Three empirical datasets, i.e., the MovieLens, the Netflix, and the RYM, are used to test the performance of the recommendation algorithms. The MovieLens and the Netflix datasets are both five-level rating movie systems, and the RYM is a ten-level music rating system. The MovieLens, downloaded from the website of GroupLens Research (http://grouplens.org), contains 943 users and 1682 objects. The Netflix, randomly selected from the huge dataset of the Netflix Prize, contains 9999 users and 5870 objects. The RYM, downloaded from the music rating website (RateYourMusic.com), contains 10159 users and 5250 objects. If the rating of the user to the object is no less than three for the MovieLens and the Netflix, and no less than six for the RYM, it can be regarded that the user likes the object, and therefore a link is added between the user and the object. For the MovieLens, it has 100000 links, for the Netflix, it has 815917 links, and for the RYM, it has 559634 links. To test the algorithm performance, the links of the network are split into the training set and the test set. The test set is composed of 10% links, which are the links randomly deleted from the total links. The rest 90% links are used as the training set to make predictions for users.

Three widely adopted evaluators are applied to quantify the recommendation accuracy performance, i.e., the ranking score $\langle \mathrm{RS} \rangle$, precision P, and recall R.[22,34] The ranking score $\langle \mathrm{RS} \rangle_{\alpha i}$ is defined as $\mathrm{RS}_{\alpha i} = p_\alpha/(n-k_i)$, where n is the number of all objects, k_i is the degree of the user u_i, and p_α is the position of the recommended object o_α located in all the uncollected objects of the user u_i. The average ranking score $\langle \mathrm{RS} \rangle$ is taken an average of $\langle \mathrm{RS} \rangle_{\alpha i}$ over all the deleted links. The smaller the $\langle \mathrm{RS} \rangle$, the higher rank of the deleted link, and the more accurate the algorithm. The precision P is defined as $P = (1/m)\sum_{i=1}^{m} q_{iL}/L$, where q_{iL} is the number of the user i's deleted links contained in the top L recommended object list. The higher the precision, the more accurate the recommendation, and *vice versa*. The recall R is defined as $R = (1/m) \sum_{i=1}^{m} (q_{iL}/l_i)$, where l_i is the number of the user i's deleted links in the test set. The higher the recall, the more accurate the recommendation, and *vice versa*.

Two evaluators are used to quantify the personalized recommendation, i.e., the novelty NL and Hamming distance H.[28,35] The novelty NL is defined as $\mathrm{NL} = (1/mL) \sum_{i=1}^{m} \sum_{o_\alpha^i \in O_\mathrm{R}^i} k_{o_\alpha^i}$, where O_R^i is the object set of the user i's recommendation list. The smaller the NL, the more novel the algorithm, and *vice versa*. The recommendation diversity indicated by the Hamming distance H is defined as

$$H = \frac{2}{m(m-1)} \sum_{i=1}^{m} \sum_{j=i+1}^{m} \left(1 - \frac{\phi_i \cap \phi_j}{L}\right),$$

where $\phi_i \cap \phi_j$ is the number of the common objects recommended for the user u_i and u_j in the top L recommendation list, and therefore quantifies the difference between two users' recommendation lists. The higher the H, the more diverse the algorithm, and *vice versa*.

4. Results

To investigate how the heterogeneity of source objects and the indirect linkage effect on the recommendation efficiency, we compare the HMD and the IMD with the original MD method, and also the traditional user-similarity based CF method. For the HMD and IMD methods, we firstly try to find out the optimal value of the tunable parameter. Since the optimal value of parameter may differ in the evaluators, here we use the widely adopted ranking score as the evaluator to estimate the optimal value. As shown in Fig. 3, for the HMD method, the optimal value is obtained at the minimal value of the ranking score $\langle \mathrm{RS} \rangle$, with $\eta = 1.87$ for the MovieLens, 1.58 for the Netflix, and 1.78 for the RYM, respectively. For the IMD method, as shown in Fig. 4, the minimal value of the ranking score $\langle \mathrm{RS} \rangle$ is obtained at $(\lambda, \eta) = (0.58, 0.94)$ for the MovieLens, $(0.46, 0.92)$ for the Netflix, and $(0.78, 1.02)$ for the RYM, respectively. All the following results are based on the optimal value of the parameter, and the results are averaged over six independent runs.

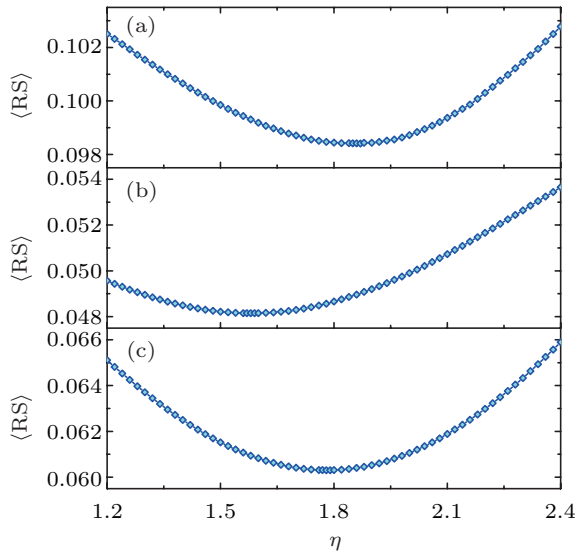


Fig. 3. (color online) The ranking score ⟨RS⟩ on the tunable parameter η of the HMD algorithm for (a) the MovieLens, (b) the Netflix, and (c) for the RYM.

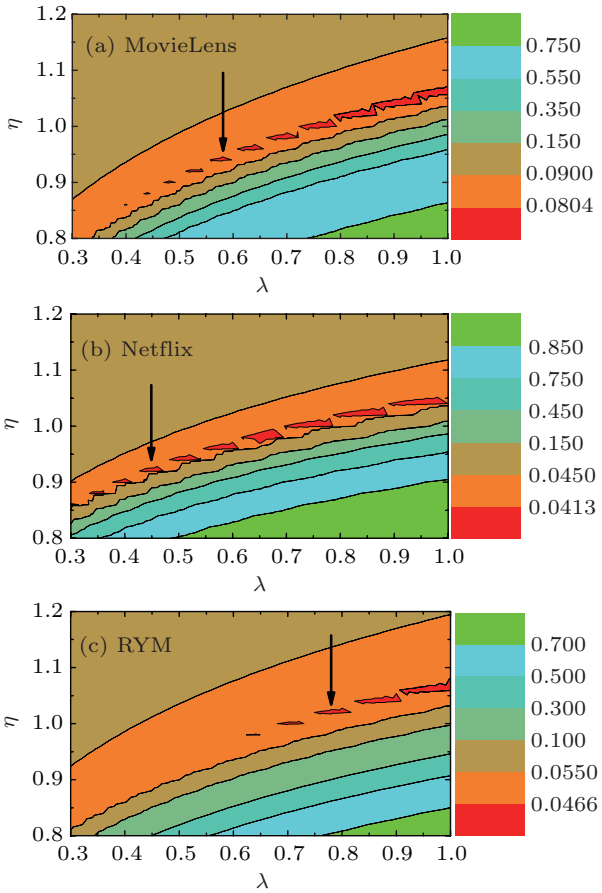


Fig. 4. (color online) The ranking score ⟨RS⟩ on the tunable parameter λ and η of the IMD algorithm. The arrow points to the optimal value of the parameter. (a) MovieLens, (b) Netflix, (c) RYM.

The results of the CF, MD, HMD, and IMD are shown in Table 1. It is observed, for the MovieLens, Netflix, and RYM, the HMD and IMD methods outperform the original MD, as well as the CF, in all the evaluators of accuracy and personalization. To quantify the improvement, we show the improvement percentage of the IMD against the CF, MD, and HMD in Table 2. Taking the MovieLens as an example, the improvement percentage of the ranking score ⟨RS⟩ of the IMD against the MD is 22.9%, and further against the HMD is 18.2%. In addition, the improvement percentage of the diversity H of the IMD against the MD is 28.2%, and against the HMD is 15.0%.

From the above results, the recommendation performance can be improved by weakening both perspectives of the object heterogeneity and the indirect linkage effect. However, compared with the HMD method, the improvement of the IMD method is much greater. That is to say, the redundancy induced by the indirect linkage could have an important impact on the recommendation efficiency for the MD method. The possible reason for the relatively smaller improvement of the HMD algorithm could be that the HMD optimizes the source object resource, which may have less effect on the recommendation list obtained based on the order of the target object resource.

To investigate the robustness of the results, the novelty NL and Hamming distance H on the recommendation list length L is studied for the CF, MD, HMD and IMD methods. As shown in Fig. 5, for the MovieLens and Netflix, the novelty NL is found to be improved for the HMD method to a certain degree, and greatly improved for the IMD method for all the investigated range of the recommendation list length, compared with that of the CF and MD methods. Similarly, as shown in Fig. 6, the Hamming distance H is observed to be improved for the HMD method to a certain degree, and greatly improved for the IMD method for all the investigated range of L for the MovieLens and Netflix, compared with that of both the CF and MD methods. The results suggest that the IMD algorithm indeed greatly improves the personalized recommendation.

Another important evaluation of the recommendation algorithm is how it works on the cold objects, i.e., the so-called cold start problem.[11,13,36] From our datasets, the cold objects occupy a big proportion, and the object with its degree no more than 10 is 41.26%, 49.59%, and 21.73% for the MovieLens, Netflix, and RYM, respectively. For lack of the historical records, it remains a great challenge to make recommendation for them. Here we employ an object-dependent ranking score $\langle RS\rangle_k$[28] to evaluate the algorithm performance for the cold objects. The $\langle RS\rangle_k$ is defined as the average ranking score over objects with the same value of degrees. As shown in Fig. 7, the $\langle RS\rangle_k$ of the HMD is observed to be a little more advantageous than that of the CF and MD again, while that of the IMD is found to be much more advantageous than that of the CF and MD, for all the three datasets. It implies that, for the recommendation of the cold objects, the improvement of the IMD method is also much greater than that of the HMD, which further suggests that the indirect linkage effect has an important impact on the recommendation efficiency of the MD method.

Table 1. The ranking score $\langle RS \rangle$, precision P, recall R, novelty NL, Hamming distance H of the CF, MD, HMD, and IMD algorithms are shown for the MovieLens, Netflix, and RYM, with $L = 50$.

		Optimal parameter	$\langle RS \rangle$	P	R	NL	H
MovieLens	CF	–	0.116	0.068	0.424	246	0.551
	MD	–	0.105	0.074	0.477	233	0.616
	HMD	$\eta = 1.87$	0.099	0.076	0.491	219	0.687
	IMD	$(\lambda, \eta = 0.58, 0.94)$	**0.081**	**0.088**	**0.564**	**188**	**0.790**
Netflix	CF	–	0.059	0.053	0.405	2351	0.403
	MD	–	0.051	0.054	0.420	2336	0.423
	HMD	$\eta = 1.58$	0.048	0.056	0.422	2217	0.514
	IMD	$(\lambda, \eta = 0.46, 0.92)$	**0.040**	**0.067**	**0.497**	**2016**	**0.641**
RYM	CF	–	0.089	0.037	0.439	490	0.848
	MD	–	0.069	0.042	0.497	466	0.874
	HMD	$\eta = 1.78$	0.061	0.042	0.493	426	0.905
	IMD	$(\lambda, \eta = 0.78, 1.02)$	**0.047**	**0.050**	**0.563**	**415**	**0.914**

Table 2. Improvement percentage of the IMD against the CF, MD, and HMD is shown for the MovieLens, Netflix, and RYM, which is computed based on the results of Table 1.

		$\langle RS \rangle$	P	R	NL	H
MovieLens	δ_{CF}	30.2%	29.4%	33.0%	23.6%	43.4%
	δ_{MD}	22.9%	18.9%	18.2%	19.3%	28.2%
	δ_{HMD}	18.2%	15.8%	14.9%	14.2%	15.0%
Netflix	δ_{CF}	32.2%	26.4%	22.7%	14.2%	59.1%
	δ_{MD}	21.6%	24.1%	18.3%	13.7%	51.5%
	δ_{HMD}	16.7%	19.6%	17.8%	9.1%	24.7%
RYM	δ_{CF}	47.2%	35.1%	28.2%	15.3%	7.8%
	δ_{MD}	31.9%	19.1%	13.3%	10.9%	4.6%
	δ_{HMD}	23.0%	19.1%	14.2%	2.6%	1.0%

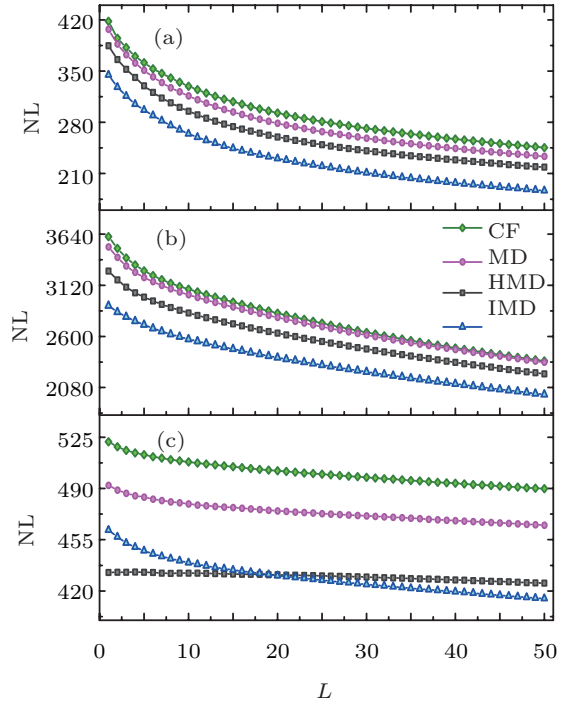


Fig. 5. (color online) The novelty NL on the recommendation list length L for (a) the MovieLens, (b) the Netflix, and (c) the RYM.

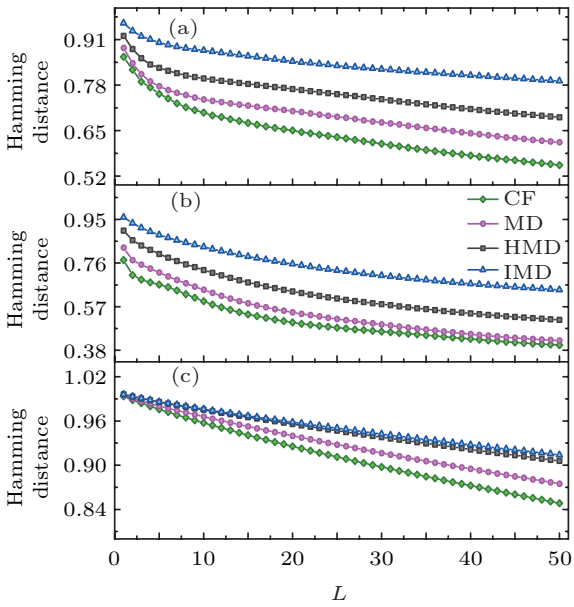


Fig. 6. (color online) The Hamming distance H on the recommendation list length L for (a) the MovieLens, (b) the Netflix, and (c) the RYM.

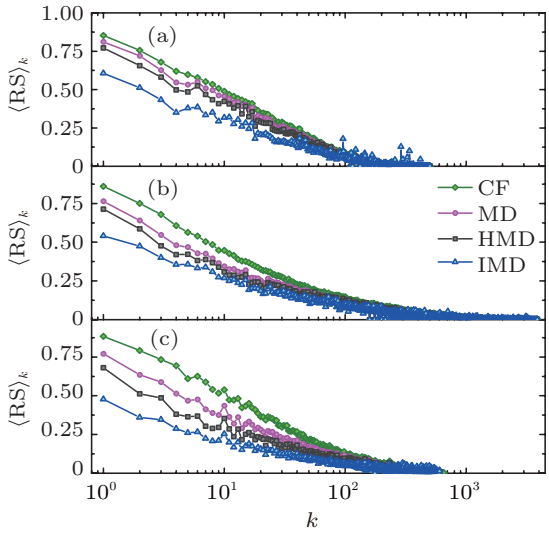


Fig. 7. (color online) The object-dependent ranking score $\langle RS \rangle_k$ on the object degree k for (a) the MovieLens, (b) the Netflix, and (c) the RYM.

5. Conclusion

We propose an improved network based recommendation method, the IMD method, by considering the indirect linkage and the object heterogeneity effect, based on the mass diffusion (MD) method. Tested on three empirical datasets, the IMD method is found to greatly improve both the recommendation accuracy and diversity of the MD method, compared with the HMD method which merely considers the source object heterogeneity. Moreover, the IMD method also greatly improves the recommendation accuracy of the cold objects. The obtained results indicate that the indirect linkage effect could have an important impact on the recommendation efficiency in the MD method.

References

[1] Adomavicius J and Tuzhilin A 2005 *IEEE Trans. Knowl. Data Eng.* **17** 734
[2] Goldberg D, Nichols D, Oki B M and Terry D 1992 *Commun. ACM* **35** 61
[3] Schafer J B, Frankowski D, Herlocker J and Sen S 2007 *The Adaptive Web* (Berlin Heidelberg: Springer) p. 291
[4] Breese J S, Heckerman D and Kadie C 1998 *Proc.* 14*th Conf. Uncertainity Artif. Intel* (Morgan Kaufmann Publishers Inc.) p. 43
[5] Hofmann T 2007 *Proc.* 26*th Ann Intl SIGIR Conf Research Devel* (Berlin Heidelberg: Springer) p. 259
[6] Ren J, Zhou T and Zhang Y C 2008 *EPL* **82** 58007
[7] Pazzani M J and Billsus D 2007 *The Adaptive Web* (Berlin Heidelberg: Springer) p. 325
[8] Lipczak M, Hu Y, Kollet Y and Milios E 2009 *ECML PKDD Discovery Challenge* p. 157
[9] Cantador I, Vallet D and Jose J M 2009 *TECML PKDD Discovery Challenge* p. 17
[10] Ju S and Hwang K B 2009 *Proc. the ECML/PKDD* 2009 *Discovery Challenge Workshop* p. 109
[11] Zhang Z K, Zhou T and Zhang Y C 2010 *Physica A* **389** 179
[12] Shang M S, Zhang Z K, Zhou T and Zhang Y C 2010 *Physica A* **389** 1259
[13] Zhang Z K, Liu C, Zhang Y C and Zhou T 2010 *EPL* **92** 28002
[14] Zhang Z K and Liu C 2012 *Int J. Bifurcat. Chaos* **22** 1250166
[15] Kim H N, Ji A T, Ha I and Jo G S 2010 *Electron Commerce Research Appl.* **9** 73
[16] Zhang Z K, Zhou T and Zhang Y C 2011 *J. Comput. Sci. Technol.* **26** 767
[17] Lü L Y, Medo M, Yeung C H, Zhang Y C, Zhang Z K and Zhou T 2012 *Phys. Rep.* **519** 1
[18] Bai M, Hu K and Tang Y 2011 *Chin. Phys. B* **20** 128902
[19] Watts D J and Strongatz S H 1998 *Nature* **393** 440
[20] Barabási A L and Albert R 1999 *Science* **286** 509
[21] Zhang Y C, Blattner M and Yu Y K 2007 *Phys. Rev. Lett.* **99** 154301
[22] Zhou T, Ren J, Medo M and Zhang Y C 2007 *Phys. Rev. E* **76** 046115
[23] Liu J G, Zhou T and Guo Q 2011 *Phys. Rev. E* **84** 037101
[24] Qiu T, Wang T T, Zhang Z K, Zhong L X and Chen G 2013 *EPL* **104** 48007
[25] Zhou T, Kuscsik Z, Liu J G, Medo M, Wakeling J R and Zhang Y C 2010 *Proc. Natl. Acad. Sci. USA* **107** 4511
[26] Qiu T, Chen G, Zhang Z K and Zhou T 2011 *EPL* **95** 58003
[27] Qiu T, Zhang Z K and Chen G 2013 *PLoS One* **8** e63531
[28] Zhou T, Jiang L L, Su R Q and Zhang Y C 2008 *EPL* **81** 58004
[29] Liu C and Zhou W X 2012 *Physica A* **391** 5704
[30] Zhou T, Su R Q, Liu R R, Jiang L L, Wang B H and Zhang Y C 2009 *New J. Phys.* **11** 123008
[31] Qiu T, Han T Y, Zhong L X, Zhang Z K and Chen G 2014 *Computer Phys. Commun.* **185** 489
[32] Holme P, Liljeros F, Edling C R and Kim B J 2003 *Phys. Rev. E* **68** 056107
[33] Chen H B, Fan Y, Fang J Q and Di Z R 2009 *Acta Phys. Sin.* **58** 1383 (in Chinese)
[34] Herlocker J L, Konstan J A, Terveen L G and Riedl J T 2004 *ACM T. Inform. Syst.* **22** 5
[35] Lü L Y and Liu W P 2011 *Phys. Rev. E* **83** 066119
[36] Lam X N, Vu T, Le T D and Duong A D 2008 *Proceedings of the* 2*nd International Conference on Ubiquitous Information Management and Communication* p. 208