Published in "Chinese Physics Letters 26(11): 118903, 2009"
which should be cited to refer to this work.

Diffusion-Based Recommendation in Collaborative Tagging Systems *

SHANG Ming-Sheng(尚明生)[1**], ZHANG Zi-Ke(张子柯)[2***]

[1]*School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054*

[2]*Department of Physics, University of Fribourg, CH-1700 Fribourg, Switzerland*

Recently, collaborative tagging systems have attracted more and more attention and have been widely applied in web systems. Tags provide highly abstracted information about personal preferences and item content, and therefore have the potential to help in improving better personalized recommendations. We propose a diffusion-based recommendation algorithm considering the personal vocabulary and evaluate it in a real-world dataset: Del.icio.us. Experimental results demonstrate that the usage of tag information can significantly improve the accuracy of personalized recommendations.

PACS: 89.75.−k, 89.20.−a, 89.20.Ff

The exponential growth of web information has brought us into an information overload era: We face too much data and sources to be able to find out those most relevant and interesting for us. Evaluating all these alternatives by ourselves is impossible. As a consequence, an urgent problem is how to automatically find out the relevant items for us. Internet search engine[1] provide us with a useful tool to find out this information and they have achieved great success over the last decade. However, it does not take into account personalized information and returns the same results for people with far different habits. Comparatively, recommender system,[2] adopting knowledge discovery techniques to provide personalized recommendations, is now considered to be the most promising way to efficiently gather the useful information.

One of the most prominent techniques of recommender systems is collaborative filtering (CF), where a user is recommended items that people with similar tastes and preferences liked in the past. Despite its success, the performance of CF is strongly limited by the sparsity data. Thus, a number of researchers devoted to integrate additional information, such as user profiles,[3] item content[4] and attributes,[5] to filter out possibly irrelevant recommendations. However, these applications are usually strongly restricted to respect personal privacy, or limited due to the lack of available content information.

Collaborative tagging systems (CTSs), allowing users to freely assign tags to their collections, provide promising possibility to better address the above issues. CTSs require no specific skills for user participating, thus can overcome the limitation of vocabulary domains and size, widen the semantic relations among items and eventually facilitate the emergence of folksonomy.[6] In addition, tags can be treated as abstracted content of items. Especially, tags are given by users themselves and thus in somehow represent the personal vocabulary and preferences. Recently, many efforts have been addressed in understanding the structure, evolution[7] and usage patterns[8] of CTSs. A considerable number of algorithms have been designed to recommend tags to users, which may be helpful for better organizing, discovering and retrieving items.[6,9,10] The current work focuses on a relevant yet different application of CTSs, that is, to provide personalized item recommendations with the help of tag information. Schenkel *et al.*[11] proposed an incremental threshold algorithm taking into account both the social ties among users and semantic relatedness of different tags, which performs remarkably better than the algorithm without tag expansion. Nakamoto *et al.*[12] created a tag-based contextual collaborative filtering model, where the tag information is treated as the users' profiles. Tso-Sutter *et al.*[13] proposed a generic method that allows tags to be incorporated to the standard collaborative filtering, via reducing the ternary correlations to three binary correlations and then applying a fusion method to reassociate these correlations. Chi *et al.*[14] presented a model considering probabilistic polyadic factorization for personalized recommendation. Shepitsen *et al.*[15] proposed a tag clustering-based method to improve the algorithmic accuracy. Zhang *et al.*[16] and Shang *et al.*[17] proposed two hybrid algorithms for personalized recommendation making use of CTSs.

In this Letter, we propose a diffusion-based recommendation algorithm for collaborative tagging

*Supported by China Postdoctoral Science Foundation (20080431273), the National Natural Science Foundation of China under Grant Nos 60973069 and 90924011, and the Swiss National Science Foundation (200020-121848).

**Email: msshang@uestc.edu.cn

***Email: zike.zhang@unifr.ch

systems. Diffusion process on complex networks have been systematically investigated,[18] yet mostly for unipartite simple networks[19–21] or unipartite weighted networks.[22] Recently, Zhang *et al.*[23,24] and Zhou *et al.*[25,26] have studied the diffusion process on bipartite networks, and applied it to personalized recommendations. Shi *et al.* and Fu *et al.* studied statistical properties of bipartite graphs,including node strength connectivity correlation[30] and node weight distribution and disparity.[31] We consider a hybrid diffusion process on bipartite networks, involving users, items and tags. Experimental results based on a large-size real data demonstrate that the usage of tag information can significantly improve the accuracy of personalized recommendations.

We adopt a weighted variant of diffusion-based method, where the weights are given according to personal vocabulary in CTSs. A CTS consists of three sets for users $U = \{U_1, U_2, \ldots, U_n\}$, items $I = \{I_1, I_2, \ldots, I_m\}$, and tags $T = \{T_1, T_2, \ldots, T_s\}$, respectively. Figure 1 gives a simple CTSs using a bipartite graph, where three users U_1, U_2 and U_3 use four tags T_1, T_2, T_3 and T_4 to label four items I_1, I_2, I_3 and I_4. Actually, it is easy to understand that different users may consider differently for the same item, and such difference can be characterized to some extent by looking into the different usage patterns of tags. Although those tags are freely given, people are supposed to give their most favorite words to describe their best collections. A latent assumption is that the more frequently a user uses a tag, the more likely the user likes this tag as well as the items labeled with it. On the other hand, users are not willing to give too many tags for a single item.

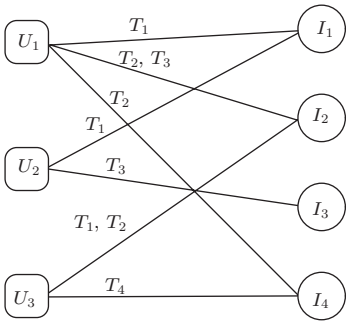

Fig. 1. Illustration of a collaborative tagging system.

We introduce a simple way that utilizes the tag information to provide better recommendations. As mentioned above, we will consider two factors: (i) the frequency of each tag used by each user; (ii) the number of tags assigned with a single item. Since our aim is to find the most relevant items for a particular user, so-called personalized recommendation, we will describe our algorithm for a target user U_i. The algorithm can be expressed in the following steps:

Step 1. Define the initial value vector $\boldsymbol{f}$ for all the items, whose element reads

$$f_j = \frac{1}{\sum_{j=1}^{m} \sum_{s'=1}^{|T_{ij}|} K(t_{is'})} \sum_{s=1}^{|T_{ij}|} K(t_{is}), \qquad (1)$$

where $|T_{ij}|$ denotes the number of tags that U_i has assigned to item I_j, and $K(t_{is})$ is the number of times tag t_s has been used by U_i.

Step 2. Distribute the value of each item evenly to the users who collect it, then the value a user U_l will receive reads

$$r_l = \sum_{j \in \Gamma(U_l)} \frac{f_j}{d(I_j)}, \qquad (2)$$

where $\Gamma(U_l)$ denotes the set of items collected by U_l, and $d(I_j)$ is the degree of I_j in the user-item bipartite graph.

Step 3. Redistribute the value of each user U_i to his/her collections according to the weight defined in Step 1. Then the final value vector $\boldsymbol{f'}$ of items will be summarized as

$$f'_j = \sum_{k=1}^{|U_{I_j}|} \frac{r_k}{\sum_{j=1}^{m} \sum_{s'=1}^{|T_{kj}|} K(t_{ks'})} \sum_{s=1}^{|T_{kj}|} K(t_{ks}), \qquad (3)$$

where $|U_{I_j}|$ is the number of users collected item I_j.

The above procedure constitutes of a mutual reinforcement process that allows the values transferred between users and items. At the first step, we highlight the items selected by U_i and assign each of them with an initial value according to U_i's tagging activities. Step 2 transfers values from items to users. In step 3, we consider the personal vocabulary again and distribute the values to items, which generates final score for each item. Finally, we sort these scores in a descending order, and the top items in the list having not been collected by U_i will be recommended to U_i.

In CTSs, different individuals have different sizes of vocabulary, and each tag may take different significance. Some tags are frequently used while some others are seldom picked. Those frequently used tags should be of higher importance in the user's viewpoint. If the user applies those frequently used tags to a specific item, it would indicate that this user prefers it to some other items assigned with infrequently used tags. Similar phenomenon also widely exists in our daily life, one can imagine that people are willing to illustrate a question using their familiar words. In addition, the number of tags assigned to an item represents how willing the user likes to describe it. By aggregating the fractions of all the tags labeling a specific item, one can estimate the importance of this item.

We use a benchmark dataset, *Del.icio.us*, to evaluate the proposed algorithm. *Del.icio.us* is 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 pools with same tags or items. The data used in this Letter is crawled from the website http://del.icio.us/ in May 2008. We guarantee that each user has collected at least one item, each item has been collected by at least two users, and assigned by at least one tag. Table 1 summarizes the basic information of the data set.

Table 1. Basic information of the data set.

Value	Description
9991	number of users
243737	number of items
102732	number of tags
1257908	number of user-item relations
4391073	accumulative number of tags

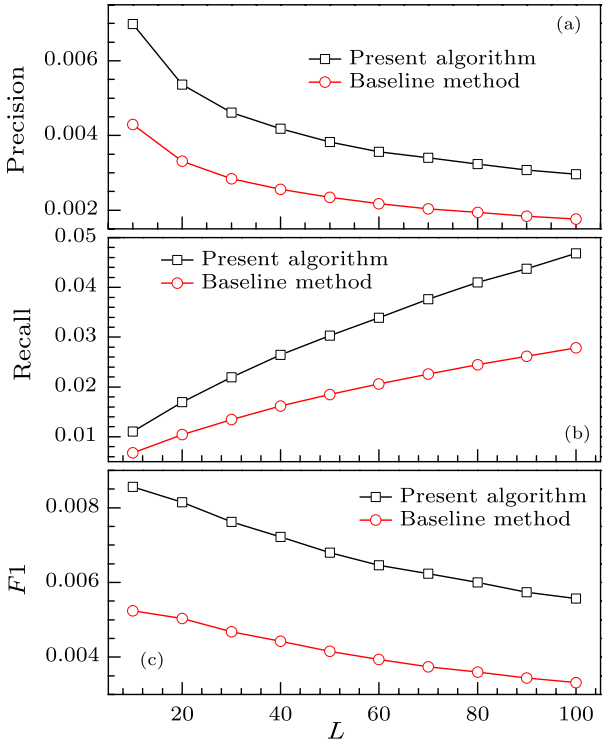

Fig. 2. Precision(a), Recall(b) and $F1$(c) versus the length of recommendation list. The results are averaged over 10 independent runs, each of which corresponds to a random division of training set and testing set.

To test the algorithmic performance, the data set is randomly divided into two parts: the training set, which is used as known information, contains 95% of entries, and the remaining 5% of entries constitute the testing set. We employ three metrics to characterize the algorithmic accuracy: Precision, Recall and $F1$, which are defined as follows:[27]

$$\text{Precision} = \frac{\sum_i N_r^i}{nL}, \tag{4}$$

where n is the number of users, L is the length of recommendation list, and N_r^i is the number of recovered items in the recommendations for user U_i.

$$\text{Recall} = \frac{\sum_i N_r^i}{\sum_i N_p^i}, \tag{5}$$

where N_p^i is the number of items collected by user U_i in the testing set.

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

Figure 2 shows the experimental results of Precision, Recall and $F1$, respectively. Since the typical length for recommendation list is tens, our experimental study focuses on the interval $L \in [10, 100]$. For comparison, we choose the method described in Ref. [25] as the baseline algorithm. It can be seen that our proposed algorithm considering the personal vocabulary significantly outperforms the baseline method in all three measurements.

In summary, we have proposed a tag-based algorithm that takes into account the personal vocabulary. Our algorithm is based on the following hypotheses: (i) Tags assigned to a certain item by a particular user represent personal tastes of it. Even for the same item, different individuals may give different tags. (ii) Different tags play different roles for the same user. The frequency of tags might suggest the personal preferences: the higher the frequency, the more the user likes it. Experimental results demonstrate that the usage of tag information can significantly improve the accuracy of personalized recommendations.

Recently, the collaborative tagging systems have attracted more and more attention in the scientific and engineering worlds.[7,28] A great number of publications and web applications have discussed/adopted tagging functions. Our experimental results show that tags can be used to not only assist personal resource organizing, but also help to filter out mass information. We only provided a simple way to consider the use of tags, and a couple of open issues remain for future study. From the perspective of human dynamics, the rank of tags within a single collection and the time the user chooses tags could also be taken into account. In addition, the hypergraph[29] description is a promising tool to exploit a comprehensive view of CTSs and bring us an in-depth understanding of the structure and evaluation of CTSs.

References

[1] Brin S and Page L 1998 *Comput. Netw. ISDN Syst.* **30** 107
[2] Resnick P and Varian H R 1997 *Commun. ACM* **40** 56
[3] Kazienko P et al 2007 *Inf. Sci.* **177** 2269
[4] Pazzani M J and Billsus D 2007 *LNCS* **4321** 325
[5] Tso K and Schmidt-Thieme L 2005 *Proc. 29th Annual Conference of the German Classification Society* (Magdeburg,

Germany 9–11 March 2005)
[6] Hotho A et al. 2006 *LNCS* **4011** 411
[7] Cattuto C et al 2008 *PNAS* **104** 1461
[8] Golder S A and Huberman B A 2006 *J. Inf. Sci.* **32** 198
[9] Mishne G 2006 *Proc.* 15th *WWW* (Edinburgh, Scotland 23–26 May 2006)) p 953
[10] Sigurbjörnsson B and Zwol R V 2008 *Proc.* 17th *WWW* (Beijing, China 21–25 April 2008) p 327
[11] Schenkel R et al 2008 *Proc.* 31th *ACM SIGIR* (Singapore 20–24 July 2008) p 523
[12] Nakamoto R Y et al 2007 *IAENG Int. J. Comput. Sci.* **34** 214
[13] Tso-Sutter K H L et al 2008 *Proc.* 23rd *ACM SAC 2008* (Fortaleza, Ceará, Brazil 16–20 March 2008) p 1995
[14] Chi Y et al 2008 *Proc. 17th ACM Conference on Information and Knowledge Management* (California 26–30 October 2008) p 941
[15] Shepitsen A et al 2008 *Proc. ACM Conference on Recommender Systems* (Lausanne, Switzerland 23–25 October 2008) p 259
[16] Zhang Z K et al 2009 *Physica* A (accepted)
[17] Shang M S et al 2009 *Physica* A **388** 4867
[18] Zhou T et al 2006 *Prog. Nat. Sci.* **16** 452
[19] Zhou Y Z et al 2007 *Chin. Phys. Lett.* **24** 581
[20] Zhao H and Gao Z Y 2007 *Chin. Phys. Lett.* **24** 1114
[21] Zhang H F et al 2009 *Chin. Phys. Lett.* **26** 068901
[22] Yan G et al 2005 *Chin. Phys. Lett.* **22** 510
[23] Zhang Y C et al 2007 *Phys. Rev. Lett.* **99** 154301
[24] Zhang Y C et al 2007 *Eur. Phys. Lett.* **80** 68003
[25] Zhou T et al 2007 *Phys. Rev.* E **76** 046115
[26] Zhou T et al 2008 *Eur. Phys. Lett.* **81** 58004
[27] Herlocker J L et al 2004 *ACM Trans. Inf. Syst.* **22** 5
[28] Zhang Z K et al 2008 *Eur. Phys. J.* B **66** 557
[29] Ghoshal G et al 2009 *Phys. Rev.* E **79** 066118
[30] Shi J J et al 2009 *Chin. Phys. Lett.* **26** 078902
[31] Fu C H et al 2008 *Chin. Phys. Lett.* **25** 4181