

HYBRID RECOMMENDATION ALGORITHM BASED ON TWO ROLES OF SOCIAL TAGS

ZI-KE ZHANG

*Web Sciences Center, University of Electronic Science
and Technology of China, Chengdu 610054, P. R. China*

*Department of Physics University of Fribourg,
Chemin du Muse, Fribourg 1700, Switzerland
zhangzike@gmail.com*

CHUANG LIU

*School of Business, East China University
of Science and Technology, Shanghai 200237, P. R. China
Engineering Research Center of Process Systems Engineering
(Ministry of Education), East China University
of Science and Technology, Shanghai 200237, P. R. China
liuchuang@mail.ecust.edu.cn*

The past few years have witnessed the great success of a new family of paradigms, social tagging networks, which allows users to freely associate social tags to items and efficiently manage them. Thus it provides us a promising way to effectively find useful and interesting information. In this paper, we consider two typical roles of social tags: (i) an accessorial tool helping users organize items; (ii) a bridge that connects users and items. We then propose a hybrid algorithm to integrate the two different roles to obtain better recommendation performance. Experimental results on a real-world data set, *Del.icio.us*, shows that it can significantly enhance both the algorithmic accuracy and diversity.

Keywords: Personalized recommendation; social tagging networks; hybrid.

1. Introduction

The past few years have witnessed an explosion era of information: we face too much data to be able to filter out irrelevant information [Zhang *et al.*, 2008a]. Recently, *Personalized Recommendation*, which provides a promising way to solve this dilemma, has attracted increasing attention [Huang *et al.*, 2004; Herlocker *et al.*, 2004; Zhang *et al.*, 2007a; Zhang *et al.*, 2007b; Zhou *et al.*, 2007; Zhou *et al.*, 2008; Zhou *et al.*, 2010], among which the most successful one is *Collaborative Filtering* (CF) [Balabanovic & Shoham, 1997; Sarwar *et al.*, 2001]. CF tends to produce recommendations by evaluating most similar users or items to the target

user based on his/her historical activities. Despite its great success, CF has also some drawbacks: (i) a huge number of items that is far beyond users' ability to measure even a small fraction of them; (ii) many users are not willing to give explicit rate to items. Therefore, a variety of researches have been devoted to use accessorial information to obtain better performance of recommendation under the framework of CF. Examples include user profiles [Kazienko & Adamski, 2007], item attributes [Tso & Schmidt-Thieme, 2005] and content descriptions [Pazzani & Billsus, 2007]. However, profile-based methods are restricted by user privacy rules, attribute-aware algorithms are limited

by vocabulary which are defined by domain experts, and various applications (e.g. videos, images) are lack of contents.

Recently, the advent of Web2.0 and its affiliated applications bring a new form of paradigms, *Social Tagging Systems*, introducing a novel platform for users' participation. A social tagging system allows users to freely assign tags to annotate their collections into baskets, requires no specific skills for users' participation, broadens the semantic relations among users and items, and thus has attracted much attention from the scientific community. A considerable number of researches have been done to study its usage patterns [Golder & Huberman, 2006], structure [Zhang *et al.*, 2008b], evolution and dynamics [Cattuto *et al.*, 2007; Zhang & Liu, 2010]. Besides, social tagging systems have also found wide applications in recommender systems. Szomszor *et al.* [2007] considered the tag frequency as weight in movie recommender systems, and the result showed that with the consideration of weights the recommendation accuracy can be enhanced. Schenkel *et al.* [2008] proposed an incremental threshold algorithm taking into account both the social ties among users and semantic relations of different tags, which performed remarkably better than the algorithm without tag expansion. In addition, Zhang *et al.* [2010] and Shang *et al.* [2010] proposed an item-based and user-based hybrid tag algorithm, respectively, harnessing diffusion-based method [Zhou *et al.*, 2007] to obtain better recommendations. Furthermore, Shang *et al.* [Shang & Zhang, 2009] considered the tag usage frequency as edge weight in bipartite networks and improved the accuracy of recommendation.

In this paper, we propose a hybrid recommendation algorithm which considers two basic roles of social tags: (i) an accessorial tool helping users organize items; (ii) a bridge that connects users and items. Experimental results show that the present hybrid algorithm can significantly improve both the recommendation accuracy and diversity.

2. Hybrid Algorithm and Performance Metrics

A social tagging network considered in this paper consists of three sets, respectively of users $U = \{U_1, U_2, \dots, U_n\}$, items $I = \{I_1, I_2, \dots, I_m\}$, and tags $T = \{T_1, T_2, \dots, T_r\}$. A tripartite graph representation can be described by three adjacent matrices: A ,

A' and A'' for user-item, item-tag and user-tag relations. If user U_i has collected item I_j , we set $a_{ij} = 1$, otherwise $a_{ij} = 0$. Analogously, we set $a'_{jk} = 1$ if I_j has been assigned by the tag T_k , and $a'_{jk} = 0$ otherwise. Furthermore, we consider that users have personal preferences on tags, that is, a weighted adjacent matrix A'' is constructed, of which each element, a''_{ik} , is denoted as how many times that U_i has used tag T_k .

Subsequently, we introduce the diffusion-based hybrid algorithm. It can be easily understood that, in a real social tagging system, a user might have two different tagging behaviors [Zhang & Liu, 2010]: (i) one might be aware of an item via random web surfing and save it as his/her favorites by adding some related words (tags) to describe them; (ii) s/he might firstly input one (or several compound) tag(s), then pick up some items s/he is interested in among all the possible retrieval results related to the required tag(s). The former one indicates social tags play the role of an organization tool for users, and the latter suggests tags as an information retrieval bridge. Given a target user U_i , the above three algorithms will generate final score of each item, f_j , that are pushed into recommendation list for him/her, are described as following:

Algorithm I. Supposing that a kind of resource, a_{is} for user U_i and item I_s , is initially located on items, each item will averagely distribute its resource to all neighboring tags, and then each tag will redistribute the received resource to all the items associated with it. The final resource vector, \mathbf{f} , after the two-step diffusion is:

$$f_j = \sum_{l=1}^r \frac{a'_{jl}}{k(T_l)} \sum_{s=1}^m \frac{a_{is}}{k'(I_s)}, \quad j = 1, 2, \dots, m, \quad (1)$$

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

Algorithm II. Different from Algorithm I, the initial resource, a''_{il} for user U_i and tag T_l , are located on tags according to their frequencies used by the given user U_i . Then each tag will distribute the initial resource directly to all its neighboring items. Thus, the final resource vector, \mathbf{f}' , reads:

$$f'_j = \sum_{l=1}^r \frac{a'_{jl} a''_{il}}{k(T_l)}, \quad j = 1, 2, \dots, m. \quad (2)$$

We then adopt a linear superposition to combine the above two algorithms, the final resource can be written as:

$$\mathbf{f}^* = (1 - \lambda)\mathbf{f} + \lambda\mathbf{f}', \quad (3)$$

where $\lambda \in [0, 1]$ is a freely tunable parameter, \mathbf{f} is the vector obtained from Eq. (1), \mathbf{f}' is the vector derived from Eq. (2), and both the vectors are attained for the same target user. In the extremal cases $\lambda = 0$ and $\lambda = 1$, the hybrid algorithm degenerates to Algorithms I and II, respectively. After generating the final resource vector by Eq. (3), all the items that s/he has not collected are ranked in a descending order, and the top L items will be recommended to the target user.

To measure the performance of the proposed hybrid algorithm, we employ two metrics, including accuracy and diversity:

- (1) *Ranking Score (RS)* [Zhou *et al.*, 2007, 2010]. In the present case, for a particular user, RS is calculated as the rank of the deleted item in testing set divided by the number of all uncollected items for this user. Apparently, the lesser RS is obtained, the higher is the accuracy of the algorithm. Then an overall $\langle RS \rangle$ is generated by averaging over all the pairs in the testing set.
- (2) *Diversity (D)* [Zhou *et al.*, 2008; Zhou *et al.*, 2009, 2010]. D measures the differences of different users' recommendation lists, thus can be understood as the inter-user diversity. Denote I_R^i as the set of recommended items for user U_i , then

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

where $L = |I_R^i|$ for any i is the length of the recommendation list. Greater or lesser values of D mean respectively greater or lesser personalization of users' recommendation lists.

3. Data and Experimental Results

The empirical data used in this paper is collected from *Del.icio.us*,¹ one of the most popular social bookmarking web sites, which allows users not only to store and organize personal bookmarks (URLs), but to also look into other users' collections and find what they might be interested in by simply

keeping track of the baskets with social tags. In addition, we remove the isolated nodes and guarantee that each user has collected at least one item, each item has been collected by at least two users, assigned by at least two tags, and each tag is used at least twice by its respective owners and at least used by two different users. The final purified dataset contains 4902 users, 36 224 items and 10 584 tags. On average, each user collects 43.85 items, each item has been assigned by 38.82 tags, and each user has 286.86 tags. Then the data set is randomly divided into two parts: the training set which constitutes 90% of the entries, and the remaining 10% entries is treated as the testing set. Each entry in social tagging networks refers to $\mathbb{F} = \{\text{user}, \text{item}, \text{tag}_1, \text{tag}_2, \dots, \text{tag}_t\}$, where t is the number of tags assigned to the relevant item by the very user.

Figures 1 and 2 show the experimental results according to Eq. (3). In Fig. 1, it can be seen that: (i) Algorithm II ($\lambda = 1$) performs better than Algorithm I ($\lambda = 0$); (ii) hybrid algorithm can provide more accurate recommendations than the pure cases. In Table 1, we compare the RS for the two pure algorithms and for the optimum by hybrid algorithm. Comparing with Algorithms I and II, at the optimal values, the improvements are 8.7%, 1.7% respectively. It indicates that the second role of social tags can more effectively help to find relevant items than the first one, and the hybrid of the two roles can further improve the

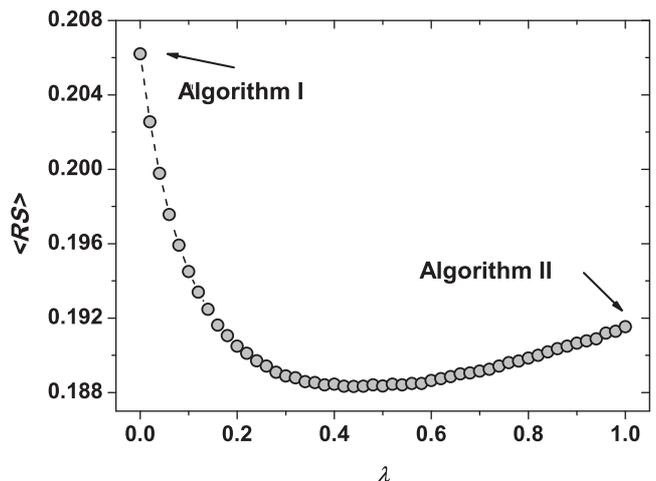


Fig. 1. RS as the function of λ . The results reported here are averaged over 50 independent runs, each of which corresponds to a random division of training and testing sets.

¹<http://del.icio.us/>

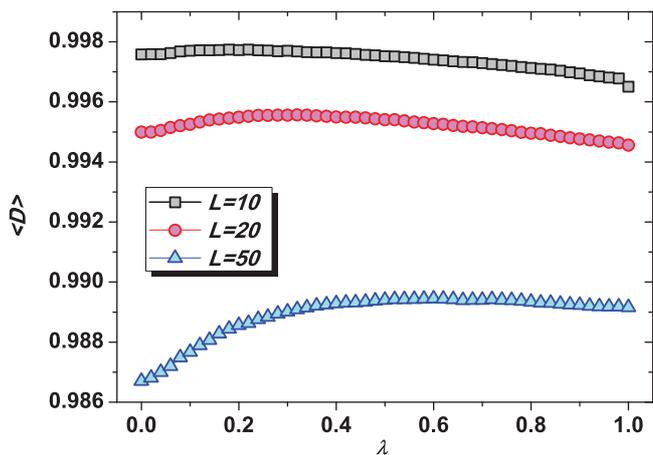


Fig. 2. D as the function of λ for three typical lengths of recommendation list: $L = 10$, $L = 20$ and $L = 50$. The results reported here are averaged over 50 independent runs, each of which corresponds to a random division of training and testing sets.

Table 1. Comparison of algorithmic accuracy, measured by the RS . I and II denote the Algorithms I and II, respectively corresponding to $\lambda = 0$ and $\lambda = 1$. The optimal value of λ as well as the corresponding optima of RS are presented for comparison.

I	II	Optimum	λ_{opt}
0.20621	0.19154	0.18833	0.44

recommendation performance by fusing them. Figure 2 reports the values of diversity versus the parameter λ . It shows that the pure Algorithm I performs the best for small recommendation list $L = 10$, the hybrid method can outperform the pure cases for median length $L = 20$, and Algorithm II and hybrid algorithm can obtain almost the same performance for large length $L = 50$. This result suggests that the two roles of tags contribute to different scales of importance in tagging networks: (i) the first role is effective to help users to manage their own items, which is a small fraction compared with the total number of items in the whole system; (ii) comparatively, the second role can help users find interesting items from a large scale of candidates, which makes it similar to finding targets by inputting keywords in a search engine.

4. Conclusions and Discussion

In this paper, we have proposed a hybrid recommendation algorithm based on two roles of social tags. (i) an accessorial tool helping users

organize items; (ii) a bridge that connects users and items. Experimental results demonstrate that the hybrid algorithm can significantly improve accuracy and diversity of recommendations. In addition, the result of diversity shows Algorithm I is efficient for small L , Algorithm II is effective for large L , and the hybrid method alleviates this contradiction and gives more diverse recommendation results for median length of recommendation list. Furthermore, this result might shed some light on the fact that the two different roles of social tags correspond to different purposes of tagging: personal organization and information retrieval, respectively.

This paper only provides a simple start point in understanding the roles of social tags, as well as making use of them to design recommendation algorithms. A couple of open issues remain for future study. First, we lack quantitative understanding of the structure and evolution of social tagging networks to well understand the observed roles. In addition, the two pure algorithms might reduce some information of social tagging networks. The hypergraph theory [Ghoshal *et al.*, 2009; Zhang & Liu, 2010] is expected to harness the whole network structure without losing any information and thus provide a promising way to obtain better recommendation performance.

Acknowledgments

This work is partially supported by the Swiss National Science Foundation (Project 200020-121848). Z.-K. Zhang acknowledges the National Natural Science Foundation of China under the grant no. 60973069. Both authors also acknowledge the Scholarship Program supported by China Scholarship Council (CSC Program).

References

- Balabanović, M. & Shoham, Y. [1997] “Fab: Content-based, collaborative recommendation,” *Commun. ACM* **40**, 66–72.
- Cattuto, C., Loreto, V. & Pietronero, L. [2007] “Semi-otic dynamics and collaborative tagging,” *Proc. Natl. Acad. Sci. USA* **104**, 1461.
- Ghoshal, G., Zlatić, V., Caldarelli, G. & Newman, M. E. J. [2009] “Random hypergraphs and their applications,” *Phys. Rev. E* **79**, 066118.
- Golder, S. A. & Huberman, B. A. [2006] “Usage patterns of collaborative tagging systems,” *J. Info. Sci.* **32**, 198–208.

- Herlocker, J. L., Konstan, J. A., Terveen, L. G. & Riedl, J. T. [2004] "Evaluating collaborative filtering recommender systems," *ACM Trans. Inf. Syst.* **22**, 5–53.
- Huang, Z., Cheng, H. & Zeng, D. [2004] "Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering," *ACM Trans. Inf. Syst.* **22**, 116–142.
- Kazienko, P. & Adamski, M. [2007] "AdROSA—Adaptive personalization of web advertising," *Info. Sci.* **177**, 2269–2295.
- Pazzani, M. J. & Billsus, D. [2007] "Content-based recommendation systems," *Lect. Note. Comp. Sci.* **4321**, 325–341.
- Sarwar, B., Karypis, G., Konstan, J. & Riedl, J. [2001] "Item-based collaborative filtering recommendation algorithms," *Proc. 10th Intl. Conf. WWW* (ACM Press, NY), pp. 295–305.
- Schenkel, R., Crecelius, T., Kacimi, M., Michel, S., Neumann, T., Parreira, J. X. & Weikum, G. [2008] "Efficient top-k querying over social-tagging networks," *Proc. 31st Ann. Intl. ACM SIGIR Conf. Res. Dev. Info. Retr.* (ACM Press, NY), pp. 523–530.
- Shang, M.-S. & Zhang, Z.-K. [2009] "Diffusion-based recommendation in collaborative tagging systems," *Chin. Phys. Lett.* **26**, 118903.
- Shang, M.-S., Zhang, Z.-K., Zhou, T. & Zhang, Y.-C. [2010] "Collaborative filtering with diffusion-based similarity on tripartite graphs," *Physica A* **389**, 1259–1264.
- Szomszor, M., Cattuto, C., Alani, H., Ohara, K., Baldassarri, A., Loreto, V. & Servedio, V. D. P. [2007] "Folksonomies, the semantic web, and movie recommendation," *Proc. 4th Euro. Semantic Web Conf.* (Innsbruck, Austria), pp. 71–84.
- Tso, K. & Schmidt-Thieme, K. [2005] "Tag-aware recommender systems by fusion of collaborative filtering algorithms," *Proc. 29th Ann. Conf. German Classification Society*, Magdeburg, Germany, 1995–1999.
- Zhang, Y.-C., Blattner, M. & Yu, Y.-K. [2007a] "Heat conduction process on community networks as a recommendation model," *Phys. Rev. Lett.* **99**, 154301.
- Zhang, Y.-C., Medo, M., Ren, J., Zhou, T., Li, T. & Yang, F. [2007b] "Recommendation model based on opinion diffusion," *EPL* **80**, 68003.
- Zhang, G.-Q., Zhang, G.-Q., Yang, Q.-F., Cheng, S.-Q. & Zhou, T. [2008a] "Evolution of the Internet and its cores," *New J. Phys.* **10**, 123027.
- Zhang, Z.-K., Lü, L., Liu, J.-G. & Zhou, T. [2008b] "Empirical analysis on a keyword-based semantic system," *Eur. Phys. J. B* **66**, 557–561.
- Zhang, Z.-K., Zhou, T. & Zhang, Y.-C. [2010] "Personalized recommendation via integrated diffusion on user-item-tag tripartite graphs," *Physica A* **389**, 179–186.
- Zhang, Z.-K. & Liu, C. [2010] "A hypergraph model of social tagging networks," *J. Stat. Mech.* P10005.
- Zhou, T., Ren, J., Medo, M. & Zhang, Y.-C. [2007] "Bipartite network projection and personal recommendation," *Phys. Rev. E* **76**, 046115.
- Zhou, T., Jiang, L.-L., Su, R.-Q. & Zhang, Y.-C. [2008] "Effect of initial configuration on network-based recommendation," *EPL* **81**, 58004.
- Zhou, T., Su, R.-Q., Liu, R.-R., Jiang, L.-L., Wang, B.-H. & Zhang, Y.-C. [2009] "Accurate and diverse recommendations via eliminating redundant correlations," *New J. Phys.* **11**, 123008.
- Zhou, T., Kuscsik, Z., Liu, J.-G., Medo, M., Wakeling, J. R. & Zhang, Y.-C. [2010] "Solving the apparent diversity-accuracy dilemma of recommender systems," *Proc. Natl. Acad. Sci. USA* **107**, 4511.