Recommending Scientific Literatures in a Collaborative Tagging Environment*

Ping Yin, Ming Zhang, and Xiaoming Li

School of Electronics Engineering and Computer Science
Peking University, Beijing, China
pkufranky@gmail.com, mzhang@net.pku.edu.cn, lxm@pku.edu.cn

Abstract. Recently, collaborative tagging has become popular in the web2.0 world. Tags can be helpful if used for the recommendation since they reflect characteristic content features of the resources. However, there are few researches which introduce tags into the recommendation. This paper proposes a tag-based recommendation framework for scientific literatures which models the user interests with tags and literature keywords. A hybrid recommendation algorithm is then applied which is similar to the user-user collaborative filtering algorithm except that the user similarity is measured based on the vector model of user keywords other than the rating matrix, and that the rating is not from the user but represented as user-item similarity computed with the dot-product-based similarity instead of the cosine-based similarity. Experiments show that our tag-based algorithm is better than the baseline algorithm and the extension of user model and dot-product-based similarity computation are also helpful.

1 Introduction

Collaborative recommendation and content-based recommendation are widely used in recommendation systems. Due to advantages and flaws of both technologies, it's a hot research to combine them to achieve better results [1, 2].

In recent years, collaborative tagging [3] becomes more and more popular. Tags can reflect both user's opinion and content features of resources. The utilization of the tag content for recommendation is worthy of a further research.

This paper focuses on scientific literature recommendation in a collaborative environment, considering both collaborative tags and content information.

There is much work related to ours. Digital libraries such as ACM[1] list similar papers in the form of text search. CiteSeer[2] provides content-based and citation-based recommendations. McNee etc. generate recommendations by mapping the web of citations between papers into the CF user-item rating matrix [4, 5].

The remainder of the paper is organized as follows. Section 2 describes the key steps for scientific literature recommendation in a collaborative tagging environment. Section 3 experimentally evaluates the algorithm. Section 4 summarizes this paper.

* This work is supported by the National Natural Science Foundation of China under Grant No. 90412010, HP Labs China under "On line course organization".

[1] http://www.acm.org/dl

[2] http://citeseer.ist.psu.edu

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 478–481, 2007.

2 Recommending Scientific Literatures in a Collaborative Tagging Environment

This paper proposes a hybrid recommendation algorithm similar to the user-user CF algorithm for a collaborative tagging environment. The difference lies in the user interest modeling, the user similarity computation and the user rating simulation.

2.1 The Representation of User Interest and Literature

The user's interest keywords have three sources: the user tags of literatures, keywords of the tagged literatures and their citations. To distinguish the importance of these three sources, different weights are assigned to them respectively. Then, the keywords frequencies are used to form an m-dimension user interest vector as follows.

$$U =< u_1, ..., u_m > \tag{1}$$

Here u_i denotes the weighted word frequency of the *i*th keyword.

Similarly, the model of a single literature consists of its keywords, keywords of its citations and all users' tags on it.

$$D =< d_1, ..., d_m > \tag{2}$$

Here d_i denotes the relative weighted frequency of the *i*th keyword summed to one.

2.2 The Computation of User Interest Degree

The user rating is simulated by user interest degree which is not directly from the user, but measured by similarity between vectors of the user interest and the literature.

The formula for interest degree is as follows, where dot-product-based similarity is used instead of cosine similarity since the length of user interest vector is meaningful.

$$R(U, D) = \sum_{i=1}^{m} u_i d_i \tag{3}$$

2.3 The Computation of User Similarity and Prediction

Once the set of most similar users is isolated with the correlation-based similarity [6], the adjusted weighted sum approach is used to obtain prediction [7].

Formally, we can denote the prediction P_{ui} as

$$P_{ui} = \overline{R_u} + \frac{\sum_{n \in NSet} sim(u,n) \times (R_{ni} - \overline{R_n})}{\sum_{n \in NSet} |sim(u,n)|} \tag{4}$$

Here NSet denotes the nearest neighbor set, $sim(u,n)$ denotes similarity between user *u* and *n*. R_{ni} denotes user *n*'s rate on item *i*, $\overline{R_u}$ denote the average rating of user *u*.

3 Experimental Evaluation

The dataset is adapted from citeulike[3] including ten thousands literatures with tags.

We use the 4-fold cross validation and the "All but one" scheme [5]. One literature is removed randomly from the tagged literatures of each user in the test dataset, and then the modified test dataset is merged into the training dataset. Then a top-10 recommendation is run on the whole dataset.

The hit percentage [5] is used to express this expectation that the removed literature can be recommended.

$$hit - percentage = hitcount / |testset| \tag{5}$$

Here *hitcount* denotes the number of successful recommendations and |*testset*| denotes the size of the testset, that is, the number of recommendations made.

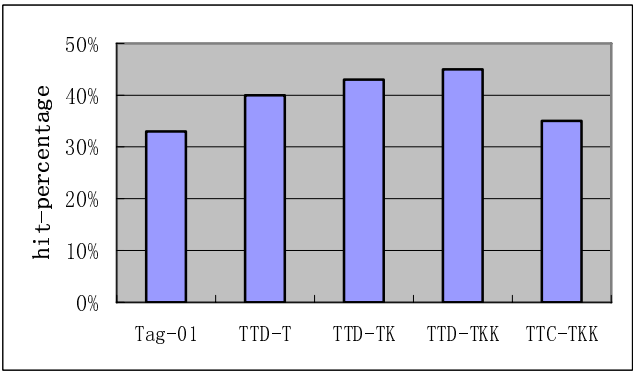

Fig. 1. Comparison of different algorithms for top-10 recommendation

All the algorithms which are used in the experiment are list below and figure 1 gives the result of all algorithms.

Tag-01: The baseline experiment which uses the user-user collaborative filtering algorithm. The rate is 0 or 1 according to whether the user has tagged the item.

Tag-text-dotproduct-T (TTD-T): User model and literature model are both represented as tag frequency vector. Dot-product-based similarity is used for the computation of user interest degree.

Tag-text-dotproduct-TK (TTD-TK): Almost the same with TTD-T except extending the user and literature model by literature keywords.

Tag-text-dotproduct-TKK (TTD-TKK): Almost the same with TTD-T except extending the user and literature model by keywords of the literature and the literature's citations.

[3] http://www.citeulike.org

Tag-text-cosine-TKK (TTC-TKK): Almost the same with TTD-TKK except that cosine-based similarity is used instead of dot-product-based similarity.

4 Conclusions and Acknowledgments

As the experiment shows, our tag-based algorithm is better than the baseline algorithm. The extension of user model with literature keywords and dot-product-based similarity computation also help to achieve better results. The prototype is now available under PKUSpace[4] [8].

This work is partially supported by NSCF Grant (60573166) as well as Network Key Lab Grant of Guang Dong Province.

References

1. Burke, R.: Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction 12(4), 331–370 (2002)
2. Balabanovic, M., Shoham, Y.: Fab: Content-Based, Collaborative Recommendation. Communications of the ACM 40(3), 66–72 (1997)
3. Golder, S.A., Huberman, B.A.: Usage Patterns of Collaborative Tagging systems. Journal of Information Science 32(2), 198–208 (2006)
4. Torres, R., McNee, S.M., Abel, M., Konstan, J.A., Riedl, J.: Enhancing Digital Libraries with TechLens+. In: Proc. of the 2004 Joint ACM/IEEE Conference on Digital Libraries, pp. 228–236 (2004)
5. McNee, S.M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S.K., Rashid, A.M., Konstan, J.A., Riedl, J.: On the Recommending of Citations for Research Papers. In: CSCW 2002: Proceedings of the 2002
6. Sarwa, B.M., Karypis, G., Konstan, J., Riedl, J.: Analysis of Recommendation Algorithms for E-commerce [R]. In: ACM Conference on Electronic Commerce, pp. 158–167 (2000)
7. Pan, H.Y., Lin, H.F., Zhao, J.: Collaborative Filtering Algorithm Based on Matrix Partition and Interest Variance. Journal of the China Society for Scientific and Technical Information 25(1), 49–54 (2006)
8. Zhang, M., Yang, D.Q., Deng, Z.H., Feng, Y., Wang, W.Q., Zhao, P.X., Wu, S., Wang, S.A., Tang, S.W.: PKUSpace: A Collaborative Platform for Scientific Researching. In: Liu, W., Shi, Y., Li, Q. (eds.) ICWL 2004. LNCS, vol. 3143, pp. 120–127. Springer, Heidelberg (2004)

[4] http://fusion.grids.cn:8080/PKUSpace/home.jsp