Making Item Predictions through Tag Recommendations

Jing Peng[1] Daniel Zeng[1, 2]

[1] Institute of Automation, Chinese Academy of Sciences
[2] Department of Management Information Systems, University of Arizona
Email: jing.peng@ia.ac.cn; zeng@email.arizona.edu

Abstract

As opposed to the search engine, social tagging can be considered an alternative technique tapping into the wisdom of the crowd for organizing and discovering information on the Web. Effective tag-based recommendation of information items is a critical aspect of this social information discovery mechanism. While most existing work in the tagging domain makes item recommendations directly after constructing or learning the user profiles, items are not particularly recommendable indeed due to the limiting descriptive ability of the binary values they were assigned on interacting with users. In response to this problem, we propose to recommend the more recommendable tags, which have numerical interactions with users, to refine users' tag preference first, and then deliver quality item recommendations based on the global relationship between tags and items. Experiments on three real-world social tagging datasets demonstrate the effectiveness of our approach.

Keywords: social tagging, tag-based recommendation, tag recommendation, item recommendation, collaborative filtering

1. Introduction

In the past few years, social tagging has been gaining wide-spread popularity in a variety of applications, from social bookmarking sites (e.g., Delicious and CiteULike), movie rating sites (e.g., MovieLens), to E-commerce sites (e.g., Amazon). Social tagging systems encourage users to save and annotate Web resources of interest with tags. These tags not only allow users to conveniently revisit and retrieve previously-visited Web resources, but also enable them to search and explore what other users are interested in.

Social tagging can be considered a crowd-wisdom-based approach to information organization and discovery, an alternative to the traditional Web search engine approach. Enabling automated recommendation of various kinds in social tagging systems can further enhance this important social information discovery mechanism. In E-commerce applications, such recommendations can be a direct marketing tool. From the point of view of collaborative filtering research, tagging data generated by social tagging systems offer the potential to deliver substantially improved recommendation results as tags constitute a novel source of data complementing standard user-item interaction/rating information.

However, research on how to improve item recommendation leveraging tagging information is just emerging. Several methods that have been proposed, including the tag-aware fusion method [1] which profiles users/items with extended item/user-plus-tag vectors, and the topic-

based method [2] which views each tag as a distinct topic and computes the probability of a user saving an item by summing the transition probability through all tags. Most existing research in the tagging domain makes item recommendations directly after constructing or learning the active user's profile. Whereas, as we will point out in this paper, items are not so recommendable as tags because of the limiting descriptive ability of the binary values they were assigned on interacting with users. In response to this problem, we propose to recommend the more recommendable tags first, and then make item recommendations based on items' correlation with the recommended tags. In particular, to make effective use of both the user-item and user-tag interaction information, we propose to project the item profile of each user to a subspace while preserving users' similarities in the tag space using Locality Preserving Projections (LPP) [3].

The rest of this paper is organized as follows. We first briefly review the literature on tag-based recommendation in section 2. We then present the proposed approach to make item predictions through tag recommendations in section 3. Next, in section 4, we report on empirical evaluation using three real-world datasets. Finally, we conclude this paper and point out future research directions in section 5.

2. Related Work

Collaborative filtering (CF) is the most widely-used and commercially successful approach to recommendation. A few methods have been proposed for tag-based collaborative filtering. A straightforward method is to use tags for computing user or item similarity. The user (item) similarities in standard user-based (item-based) CF methods are derived from the similarity of items (user) the users (items) interacted with. Zeng and Li [4] introduced two variants of the standard user- and item-based methods by calculating user and item similarities based on TF-IDF weighted tag vectors. Further, Zhao et al. [5] proposed to compute the similarity of two users based on the semantic distance of their tag sets on common items they have bookmarked. Tso-Sutter et al. [1] extended the item vectors for user profiles and user vectors for item profiles with tags and then constructed the user/item neighborhoods based on the extended user/item profiles. In addition, several other alternatives have been proposed to facilitate similarity computation using tags [6-8].

There are also a number of recent studies aiming at further use of tagging information for tag-based recommendation. The topic-based method [2] exploits tag information in a probabilistic framework, viewing each tag as an indicator of a topic and then estimating the probability of a user bookmarking an item by summing the transition probabilities through all tags. Zhen et al. [9] used users' tag vectors to regularize the user-item matrix factorization results by making sure that the similarity between two user's latent feature vectors are correlated with the tag sets of the two users. The subject-based method [10] tries to extract informative tagging patterns (subjects) from the user-tag and item-tag co-occurrence matrices using Consistent Nonnegative Matrix Factorization to explain why a user saved (or might save) an item. Recently, a diffusion method [11] was proposed to generate recommendations based on fusion of information diffusion on user-item and item-tag bipartite graphs.

Among the aforementioned methods, most makes item recommendations in a user-based manner. However, a critical problem that may undermine the performance of a user-based algorithm on tagging data (applicable to general binary

datasets as well) is that the user-item interactions are binary. In tagging data, all the saved items are treated equally, whereas intuitively users tend to like different items to different extents. As a consequence, the user-based methods are prone to perform poor on tagging data in that the recommended items from each neighbor (similar user) are weighted equally. In this sense, tags should be more recommendable than items as they have numerical interactions with users that can precisely describe how much a user likes a tag. This intuition inspires us to attempt tag recommendations before item recommendation. A method highly relevant to ours is the joint item-tag recommendation approach [12], which first makes joint item-tag recommendations and then projects them to the item space for final item recommendation. Nevertheless, no light has been shed on the recommendablity of items and tags in this approach, though it also involves a tag recommendation step.

3. The Algorithm

In the proposed approach, we first synthesize users' item and tag profiles using a subspace learning method and then compute the similarities between users based on their lower dimensional representation. After that, we recommend the more recommendable tags to users in a user-based manner to refine users' tag profile. Finally, we make item recommendations based on the refined tag profiles under the topic-based recommendation framework [2].

Notation: In this paper, matrices are denoted by boldface capital letters, e.g., $\mathbf{A}$. Vectors are denoted by italic boldface lowercase letters, e.g. $\boldsymbol{z}_i$. Scalars are denoted by italic lowercase letters, e.g., α, k. Let $\mathcal{U} = \{u_1, u_2, \dots, u_m\}s$ be a set of users, $\mathcal{I} = \{i_1, i_2, \dots, i_n\}$ be a set of items, and $\mathcal{T} = \{t_1, t_2, \dots, t_l\}$ be a set of tags. We use a n dimensional column vector $\boldsymbol{x}_i$ to represent the item profile of user u_i, and a l dimensional column vector $\boldsymbol{y}_i$ to represent the tag profile of the same user. Let $p(i|u)$ denote the probability of an arbitrary user u saving an arbitrary item i, $p(t|u)$ the probability of user u using tag t, and $p(i|t)$ the probability of saving item i if tag t is used for annotation.

3.1 Subspace Learning

To make effective use of both the item and tag profiles of users, as well as to guarantee the scalability of our approach, we propose to project users' item profiles to a lower dimensional subspace while preserving users' similarities in the tag space, taking advantage of the LPP method [3]. LPP is a manifold learning approach that approximates an transformation matrix to project the high dimensional data into a lower dimensional subspace, kind of like the well-known Principle Component Analysis (PCA) [13]. The superiority of LPP over PCA is that it holds the capability to preserve certain affinities during the projection process. Generally, the affinities to be kept are data points' similarities in the original space. However, in our specific application, we propose to preserve users' similarities in the tag space instead, which enables us to integrate the user-tag interaction information into the projection process in a principled manner.

Formally, the subspace learning process of our approach aims to minimizing the objective function

$$f(\mathbf{A}) = \sum_{i,j} w_{ij} \left\| \mathbf{A}^T \boldsymbol{x}_i - \mathbf{A}^T \boldsymbol{x}_j \right\|^2 \quad (1)$$

under certain constraints. Herein, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are n dimensional column vectors representing the item profiles of user u_i and u_j, respectively. $\mathbf{A}$ represents the $n \times k$ $(k < n)$ dimensional transformation matrix projecting item profiles to the

lower dimensional subspace. w_{ij} indicates the similarity between user u_i and u_j in the tag space, given by

$$w_{ij} = \frac{\boldsymbol{y}_i^T \boldsymbol{y}_j}{\|\boldsymbol{y}_i\|\|\boldsymbol{y}_j\|} \quad (2)$$

where $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ are column vectors representing the tag profiles of user u_i and u_j respectively, with each entry being the corresponding co-occurrence counts of the given user with the given tag in the training data.

Let vector a represent an arbitrary column of $\mathbf{A}$, and $\mathbf{X}$ represent the $n \times m$ item-user interaction matrix. If we specify the constraint of the objective function to be $\boldsymbol{a}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\boldsymbol{a} = 1$, where $\mathbf{D}$ is a diagonal matrix given by $d_{ii} = \sum_j w_{ij}$. By some simple math in spectral graph theory [3], we can prove that the vectors $\boldsymbol{a}$ that minimize the above objective function are given by the minimum eigenvalue solutions of the following generalized eigenvalue problem:

$$\mathbf{X}(\mathbf{D} - \mathbf{W})\mathbf{X}^T\boldsymbol{a} = \lambda\mathbf{X}\mathbf{D}\mathbf{X}^T\boldsymbol{a} \quad (3)$$

One advantage of the proposed approach over existing tag-based recommendation methods that involve subspace learning [10, 12] is that an explicit transformation matrix is learned for projection, so our approach can be applied to new users easily without any additional effort.

3.2 Tag Recommendation

When the lower dimensional representation of user is obtained, we can compute the cosine similarities between users in the subspace as follows:

$$s_{ij} = \frac{\boldsymbol{z}_i^T \boldsymbol{z}_j}{\|\boldsymbol{z}_i\|\|\boldsymbol{z}_j\|} \quad (4)$$

where $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ are k dimensional column vectors representing the profiles of user u_i and u_j in the subspace, given by $\boldsymbol{z} = \mathbf{A}^T\boldsymbol{x}$. Note that other similarity metrics, such as Correlation and Euclidean distance, are also applicable here.

After the user similarities are computed, we can recommend items to each user directly following the traditional user-based methods [14, 15]. However, as we argued earlier, the binary-valued item recommendations from each neighbor to the active user cannot precisely capture how much the neighbor likes the recommended items. On the contrary, the co-occurrence frequencies of users with tags can take value on any non-negative integral numbers, thus the tag profiles of users are generally much more informative than the item profiles. In this sense, tags should be more recommendable than items because of their fine-grained representation. Consequently, we propose to recommend tags rather than items to the active user in a user-based manner as follows:

$$\boldsymbol{y}_i^R = \alpha \frac{\sum_{j \neq i} s_{ij} * \boldsymbol{y}_j}{\sum_{j \neq i} s_{ij}} + (1 - \alpha)\boldsymbol{y}_i \quad (5)$$

where $\boldsymbol{y}_i^R$ indicates the recommended tag vector to the active user u_i. $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ are normalized (unit sum) l dimensional column vectors representing the tag preference of user u_i and u_j, respectively. Considering that we may not want to ignore the active user's original tag preference completely, we weight the active user's own tag profile separately, with a weighting parameter α $(0 \leq \alpha \leq 1)$, so as to facilitate the tuning of the relative importance of the original tag profile in the recommended tag vectors. Similar ideas on tag recommendation can be found in [12].

It should be noted that the term "tag recommendation" generally refers to

recommending tags to the active user on saving some item in the literature, whereas the "tag recommendation" mentioned herein means recommending tags to a user without specifying any item. The former type of recommendation aims at easing the tagging process of users' bookmarking activities, while the latter one aims at refining the tag preference of users.

3.3. Item Recommendation

When the refined tag preference of users is obtained, we can recommend items to users based on items' correlation with users' preferred tags. More specifically, we can compute the probability of a user saving an item following the topic-based recommendation framework [2] as follows:

$$p(i|u) = \sum_t p(t|u)p(i|t) \qquad (6)$$

where $p(t|u)$ can be found in users' recommended tag profiles directly and $p(i|t)$ can be easily estimated from the overall item-tag co-occurrence matrix in the training data.

Finally, we can recommend the top-N items with the highest probabilities to the active user if they haven't been saved before. Figure 1 summarizes the major steps of the proposed approach to make item predictions through tag recommendations.

Algorithm: Making Item Predictions through Tag Recommendations

Input: user set, item set and the training records

Parameters: dimension of subspace – k, weighting parameter – α, number of item recommendations – N

Output: N item recommendations for each user

1. Construct user-item, user-tag and item-tag co-occurrence matrices from the training data.
2. Learn a lower dimensional representation of users using LPP.
3. Perform user-based tag recommendation to refine users' preference on tags.
4. Estimate the probability of a user saving an item by inspecting the correlation of the target item with the active user's preferred tags.
5. Choose the top-N items of each user that haven't been saved yet to recommend.

Fig. 1: Recommendation procedure.

4. Empirical Analysis

4.1 Dataset

We have evaluated our proposed approach on three datasets. The first dataset was crawled from Delicious, the largest social bookmarking site. The collected dataset consists of bookmarking data of 5000 users dated from 12/1/2008 to 12/31/2008. The second dataset is a snapshot of the CiteULike database [1] downloaded on 1/21/2010. We collected transactions that took place in year 2009. The last dataset is the Bibsonomy dataset[2] widely used in the tagging domain, and what we used is the 2009-07-01 snapshot. The Bibsonomy dataset contains bookmarks for both bibliographies and general Web resources, of which only the part for general Web resources was used in our experiment.

Dataset	Delicious	Citeulike	Bibsonomy
Number of users *m*	177	132	125
Number of items *n*	210	225	388
Number of selected/total tags *l*	142/2251	65/1584	161/2305
Number of transactions *p*	4093	3300	4383
Density level *p/mn* (%)	11.01	11.11	9.04
Avg. number of items per user	23.12	25.00	35.06
Avg. number of users per item	19.49	14.67	11.30
Avg. frequency of selected tags	21.55	13.83	16.20
Number of items per user	>=10	>=10	>=10
Number of users per item	>=10	>=10	>=8
Frequency of selected tags	>=5	>=5	>=5

Note: a transaction indicates a user saving an item, no matter how many tags are assigned.

Table 1: Dataset characteristic.

During data preprocessing, we iteratively removed users that had saved less than 10 items and items that had been saved by less than 10 users (8 for Bibsonomy) until the number of unqualified items was less than 20 for each dataset. In addition, we stemmed

[1] http://www.citeulike.org/faq/data.adp
[2] http://www.kde.cs.uni-kassel.de/bibsonomy/dumps

each tag to remove the effect of word variations. For computational efficiency and recommendation quality, we only considered tags that had occurred more than 5 times in the training set. Table 1 shows the key statistics of the cleaned datasets.

We randomly divided each of the datasets into a training set and a test set. The split was based on the training-testing ratio between 20%-80% and was done for each user. In the predication phase, we recommended 5, 10, …, 50 items for each user and then compared them with the items in the test set. The evaluation metrics adopted in our experiment were the commonly used ones including precision, recall, F-measure, and rankscore [12].

4.2 Experiment Results

We compared our recommendation approach with a variety of existing tag-based recommendation algorithms, including both the memory-based and model-based methods. As to the memory-based methods, we have implemented the traditional user-based (UB) algorithm [14, 15], the tagging user-based (TUB) algorithm [4], and the tag-aware fusion (FUS) method [1] as benchmarks. As to the model-based methods, we have implemented the topic-based (TB) algorithm [2] and the probabilistic latent semantic analysis (PLSA) algorithm [16] for comparison. Particularly, to show the necessity and effectiveness of making tag recommendations before item recommendation, we implemented two variants of our LPP-based subspace learning approach (LPP): the first one (LPP-I) makes item recommendations directly after obtaining user similarities; the second one (LPP-T) attempts tag recommendations before the final item recommendation.

We tuned the parameters of the implemented algorithms to their optimum according to 5 random splits (runs) and then tested their performance according to another 20 random splits. The final results are then averaged over these 20 runs. This tuning and testing strategy can help to avoid the over-fitting problem for any specific algorithm. The final results for top-5 item recommendation are shown in Table 2. Since the number of item recommendations has very minor impact on the relative strength of the implemented algorithms, we omit similar results for other numbers of item recommendation here.

Dataset	Algorithm	Precision	Recall	F-measure	Rank Score
Delicious	UB	29.81	8.23	12.89	30.21
	TUB	34.31	10.09	15.59	34.80
	FUS	37.62	11.06	17.09	38.17
	TB	33.91	10.00	15.45	34.39
	PLSA	35.28	10.25	15.88	35.77
	LPP-I	36.80	10.65	16.52	37.36
	LPP-T	**38.86**	**11.38**	**17.60**	**39.41**
CiteULike	UB	19.19	4.99	7.92	19.40
	TUB	20.63	5.50	8.68	20.93
	FUS	22.24	5.93	9.36	22.55
	TB	23.01	6.49	10.12	23.32
	PLSA	22.70	6.52	10.13	22.90
	LPP-I	17.65	4.67	7.39	17.83
	LPP-T	**24.64**	**6.85**	**10.71**	**24.92**
Bibsonomy	UB	14.68	3.02	5.00	14.81
	TUB	16.46	3.52	5.79	16.69
	FUS	19.46	4.30	7.04	19.75
	TB	19.59	4.11	6.79	19.87
	PLSA	16.16	3.33	5.52	16.24
	LPP-I	14.02	3.09	5.06	14.20
	LPP-T	**20.19**	**4.31**	**7.11**	**20.48**

Note: Except for Rankscore, all values for Precision, Recall and F-measure are showed in percentage.

Table 2: Experiment results.

It can be observed from Table 2 that the proposed approach (LPP-T) making item predictions through tag recommendations outperforms all the other implemented algorithms on all the three datasets, demonstrating the utility of the our recommendation method. More detailed findings are in order:

1) As can be seen clearly, the first variant of our approach that recommends items directly underperforms the second variant that recommends tags before recommending items significantly across all datasets, which effectively verifies our

earlier discussion that tags are more recommendable than items. The underlying reason is that the binary-valued user-item interactions lack the necessary descriptive ability to precisely capture users' preference on items.

2) The proposed approach outperforms the traditional user-based method which considers only item profiles of users, the tagging user-based method which considers only tag profiles users, and the tag-aware fusion method which simply integrates users' item and tag profiles in extended item-plus-tag vectors. This observation demonstrates the necessity of making principled use of users' item and tag profiles to deliver quality item recommendations.

3) The proposed approach beats the topic-based method, which shares the same item recommendation framework with our approach but does not refine users' tag profiles through tag recommendation, indicating the effectiveness of making tag recommendations to refine users' preference on tags.

5. Conclusion

In this paper, we first discussed the recommendability of items and tags, and then proposed to recommend tags to refine a user's tag profile before making the final item recommendations following the topic-based framework. In particular, to guarantee the scalability of the proposed approach and to make sure that both item and tag profiles of users are effectively used, we proposed to project users' item profile into a lower dimensional subspace while preserving users' similarity in the tag space using Locality Preserving Projections. An empirical study on three real-world datasets demonstrates the utility of the proposed approach.

For future work, we plan to find some effective smoothing methods to remove the possible noise in users' original tag profiles, which may include spam tags or meaningless tags. In addition, we are also interested in integrating the social network information among users into our subspace learning process.

6. Acknowledgments

The authors wish to acknowledge research support from the CAS (2F07C01), NNSFC (70890084, 60875049, and 60621001), and MOST (2006AA010106).

References

[1] K.H.L. Tso-Sutter, L.B. Marinho, and L. Schmidt-Thieme. *Tag-aware recommender systems by fusion of collaborative filtering algorithms*. In *Proceedings of the ACM symposium on Applied computing*, 2008.

[2] J. Peng and D. Zeng. *Topic-based web page recommendation using tags*. In *Proceedings of the 2009 IEEE international conference on Intelligence and security informatics*, 2009.

[3] X. He and P. Niyogi. *Locality preserving projections.* Advances in neural information processing systems, **16**: 153-160, 2003.

[4] D. Zeng and H. Li. *How Useful Are Tags? -- An Empirical Analysis of Collaborative Tagging for Web Page Recommendation*. In *Proceedings of the IEEE ISI 2008 PAISI, PACCF, and SOCO international workshops on Intelligence and Security Informatics*, 2008.

[5] S. Zhao, N. Du, A. Nauerz, X. Zhang, Q. Yuan, and R. Fu. *Improved recommendation based on collaborative tagging behaviors*. In *Proceedings of the 13th international conference on Intelligent user interfaces*, 2008.

[6] S. Givon and V. Lavrenko. *Predicting social-tags for cold start book recommendations*. In *Proceedings of the third ACM conference on Recommender systems*, 2009.

[7] D. Parra and P. Brusilovsky. *Collaborative filtering for social tagging systems: an experiment with CiteULike*. In *Proceedings of the third ACM conference on Recommender systems*, 2009.

[8] S. Sen, J. Vig, and J. Riedl. *Tagommenders: connecting users to items through tags*. In *Proceedings of the 18th international conference on World wide web*, 2009.

[9] Y. Zhen, W. Li, and D. Yeung. *TagiCoFi: tag informed collaborative filtering*. In *RecSys '09: Proceedings of the third ACM conference on Recommender systems*, 2009.

[10] J. Peng and D. Zeng. *Exploring Information Hidden in Tags: A Subject-based Item Recommendation Approach*. In *Proceedings of 19th Workshop on Information Technologies and Systems*, 2009.

[11] Z.-K. Zhang, T. Zhou, and Y.-C. Zhang. *Personalized recommendation via integrated diffusion on user-item-tag tripartite graphs.* Physica A: Statistical Mechanics and its Applications, **389**(1): 179-186, 2010.

[12] J. Peng, D. Zeng, H. Zhao, and F.-Y. Wang. *Collaborative Filtering in Social Tagging Systems Based on Joint Item-Tag Recommendations*. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 2010.

[13] I.T. Jolliffe, *Principal component analysis*. 2002: Springer verlag.

[14] J.S. Breese, D. Heckerman, and C. Kadie. *Empirical Analysis of Predictive Algorithms for Collaborative Filtering*. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, 1998.

[15] R. Paul, I. Neophytos, S. Mitesh, B. Peter, and R. John. *GroupLens: an open architecture for collaborative filtering of netnews*. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, 1994.

[16] R. Wetzker, W. Umbrath, and A. Said. *A hybrid approach to item recommendation in folksonomies*. In *Proceedings of the WSDM '09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, 2009.