CFUI: COLLABORATIVE FILTERING WITH UNLABELED ITEMS

Jing Peng[1] **Daniel Zeng**[2, 1] **Bing Liu**[3] **Huimin Zhao**[4]
[1]Institute of Automation, Chinese Academy of Sciences
[2]Department of Management Information Systems, The University of Arizona
[3]Department of Computer Science, University of Illinois at Chicago
[4]Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee
jing.peng@ia.ac.cn zeng@email.arizona.edu liub@cs.uic.edu hzhao@uwm.edu

Abstract

As opposed to Web search, 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 critical to the success of this social information discovery mechanism. Over the past few years, there have been a growing number of studies aiming at improving the item recommendation quality of collaborative filtering (CF) methods by leveraging tagging information. However, a critical problem that often severely undermines the performance of tag-based CF methods, i.e., sparsity of user-item and user-tag interactions, is still yet to be adequately addressed. In this paper, we propose a novel learning framework, which deals with this data sparsity problem by making effective use of unlabeled items and propagating users' preference information between the item space and the tag space. Empirical evaluation using real-world tagging data demonstrates the utility of the proposed framework.

Keywords: Social tagging, sparsity, tag-based recommendation, unlabeled items

1. Introduction

In recent 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. Social tagging can be considered a crowd-wisdom-based approach to information organization and discovery, as opposed to the traditional Web search 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 recommendation can be a direct marketing tool. From the point of view of collaborative filtering (CF) 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.

There have been a growing number of studies aiming to improve the item recommendation quality of CF methods by leveraging tags. While previous CF research on tagging data mainly focused on how to make full use of tagging information, the sparsity of user-item/tag interaction data, which might severely limit the performance of tag-based recommendation methods, has not been adequately addressed yet. In this paper, we propose a novel learning framework named CFUI, built on the special nature of tagging data, to deal with the sparsity problem by making effective use of unlabeled items. In this framework, the user-item and user-tag interactions are iteratively smoothed by inspecting the complementary interplay of users' preferences in the item and tag spaces.

2. Related Work

A number of methods have been proposed for tag-based CF. Tso-Sutter et al. (2008) extended the item vectors for user profiles and user vectors for item profiles with tags and then constructed the user/item neighborhoods for prediction based on the extended user/item profiles. Peng and Zeng (2009b) viewed each tag as an indicator of a topic and then estimated the probability of a user bookmarking an item by aggregating the transition probabilities through all tags. Zhen et al.

(2009) 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. Zhang et al. (2010) proposed a diffusion method, which generates recommendations based on fusion of information diffusions on user-item and item-tag bipartite graphs. Recently, Peng et al. (2010) presented a joint item-tag recommendation framework, which explicitly pointed out the topical interests of users in the recommended items. An advantage of this approach is that it is able to make full use of all available interactions among users, items, and tags. However, no method has been developed particularly to deal with the sparsity problem of tagging data yet.

Besides tag-based CF, another line of work closely related to ours is learning classifiers using only positive and unlabeled examples (PU learning) (Liu et al. 2002). There are generally two approaches to dealing with the imbalance of PU data. One is to identify a portion of unlabeled examples as reliable negative examples using some heuristics before applying a standard binary classification method that works on positive and negative examples (Liu et al. 2002; Yu et al. 2004). The other is to assign a certain amount of weight to unlabeled examples, treating them as weighted negative examples for training (Lee et al. 2003; Liu et al. 2003). A comprehensive evaluation (Liu et al. 2003) shows that the latter approach is generally superior. Since the literature of PU learning demonstrates that unlabeled examples carry useful information for classification, we posit that unlabeled items may also contribute helpful information to CF. A difference is that we try to find some potentially positive items (items that are likely to be saved in the future), rather than negative items, to alleviate the data sparsity problem.

3. Collaborative Filtering with Unlabeled Items

As tags hold semantic information annotating each user-item interaction in the tagging data, a natural choice to take full advantage of tagging information is to replace the latent variables in traditional probabilistic latent semantic analysis (PLSA) methods (Hofmann 2003; Wetzker et al. 2009) with tags and perform tag-based semantic analysis (TSA), as shown in Figure 1 (see only the solid arrows).

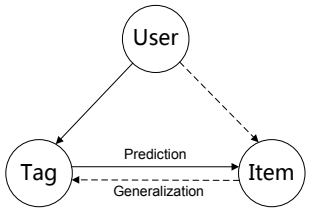


Figure 1. Tag-based semantic analysis model

An advantage of TSA over traditional PLSA models is that no special effort is needed to interpret the estimated model as a tag often carries informative semantics in its own right. In addition, the interactions of tag with the other two entities of the TSA model (i.e., user and item) are usually partially observed in the historical data. This partial observability enables us to obtain reasonable initial values for model parameters by counting co-occurrences in the training data. These initial values can also help to guide the inference procedure of TSA. Similar to the PLSA models (Hofmann 2003; Wetzker et al. 2009) well-studied in the literature, we employ the annealed EM algorithm to maximize the likelihood of TSA on the observed data, given by $L = \sum_{<u,i>} w_{ui} \, log \, p(i|u)$, as follows

E step: $p(t|u,i) = [p(t|u)p(i|t)]^{\lambda} / \sum_t [p(t|u)p(i|t)]^{\lambda} \quad (0 < \lambda \leq 1)$

M step: $p(t|u) = \alpha \sum_i w_{ui} P(t|u,i) / \sum_t \sum_i w_{ui} P(t|u,i) + (1-\alpha) p(t|u)^{init} \quad (0 \leq \alpha \leq 1)$

$p(i|t) = \beta \sum_u w_{ui} P(t|u,i) / \sum_i \sum_u w_{ui} P(t|u,i) + (1-\beta) p(i|t)^{init} \quad (0 \leq \beta \leq 1)$

where w_{ui} represents an entry of the user-item interaction matrix (**UI**), taking the value 1 if user u has saved item i and 0 otherwise. The initial values of $p(t|u)$ and $p(i|t)$ are estimated from the co-occurrence matrices of user-tag (**UT**) and item-tag (**IT**) derived from the training data, respectively. α and β are parameters reflecting the importance of observed initial values in guiding the inference procedure. λ is the annealing parameter used to alleviate the over-fitting problem. When the model is fixed, we can compute the probability of a user saving an item by

$$p(i|u) = \sum_t p(t|u)p(i|t) \quad (1)$$

Despite the earlier mentioned advantages of TSA over PLSA, a critical problem that might severely undermine the performance of TSA is the sparsity of **UT** and **UI**. The sparsity of **UT** may lead to poor initial values for $p(t|u)$ and hence hurt the prediction accuracy. The sparsity of **UI** may cause over-fitting to a small number of positive items, which might not be a comprehensive representation of all the potential positive items.

A major cause for the sparsity of **UT** is users' tendency to annotate items with only a small set of tags, even none sometimes. In addition, the personal trait of some users in using tags interpretable only to themselves (e.g., "***", "toread") also causes difficultly to share preference information of users among the community. In response to these problems, we take advantage of the tag generalization method (Peng et al. 2009a), which posits that the assigning of tags to an item should be independent of users. Specifically, we standardize tag usage across different users and smooth user-tag interaction as shown in matrix form in equation (2). Note that both **UI** and **IT** are normalized to unit row sum before the multiplication so that the resulting **UT** matrix also represents probabilities. The rationale behind this smoothing strategy is that the tag set aggregated over all users on an item is generally more complete and objective than those of individual users. The process of tag generalization is illustrated by the dashed arrows in Figure 1.

$$\mathbf{UT}_{\text{normalized}} = \mathbf{UI}_{\text{normalized}} \cdot \mathbf{IT}_{\text{normalized}} \quad (2)$$

Inspired by the idea of identifying negative examples from unlabeled data to deal with data imbalance in PU learning (Lee et al. 2003; Liu et al. 2003), we propose to alleviate the sparisty of **UI** by discovering some potential positive items. Considering that the task of CF is to recommend the best N items to each user, we can imagine that there is a total amount of εN weight distributed among the unlabeled items of each user ($\varepsilon \in [0,1]$, a parameter reflecting the relative importance of unlabeled potential positive items as compared to observed positive items). The item recommendation problem then becomes to re-assign this amount of weight to the subset of the best N unlabeled items of each user as much as possible. To provide a systematic principle to guide this re-assigning process, we propose to minimize the perplexity of TSA on unlabeled items while maximizing its likelihood on positive items. As the negative logarithm of a perplexity is in the form of likelihood, this criterion is equivalent to maximizing the likelihood of TSA on both positive and unlabeled items. Formally, the objective of TSA can be rewritten as:

$$\begin{aligned} &\underset{w_{ui},\ p(t|u),\ p(i|t)}{\operatorname{argmax}} \sum_{<u,i>} w_{ui}\, log\, p(i|u) \\ &s.t. \quad w_{ui}^P = 1 \\ &\qquad w_{ui}^U \le \varepsilon \\ &\qquad \textstyle\sum_i w_{ui}^U = \varepsilon N \end{aligned}$$

where w_{ui}^P indicates observed positive entries in **UI** and w_{ui}^U unlabeled entries. To solve this constrained optimization problem, we propose to optimize the objective with respect to the user-item interaction weights (w_{ui}) and model parameters ($p(t|u)$ and $p(i|t)$) alternatively, which guarantees at least a local optimal solution. Since w_{ui} can be initialized from the training data, we can first estimate $p(t|u)$ and $p(i|t)$ following the EM procedure discussed earlier and then

optimize w_{ui} while keeping the model parameters fixed. Ideally, the above likelihood will be maximized with respect to w_{ui} if the assumed weight is completely allocated to unlabeled items with the largest $p(i|u)$, which can be computed using equation (1). However, taking the constraints into consideration and to avoid w_{ui} being over-optimized with respect to the model parameters of any single iteration, we propose to update the weights on unlabeled items incrementally as follows, with a learning rate η.

$$w_{ui}^{U\ new} = (1-\eta)w_{ui}^{U\ old} + \eta\varepsilon N \frac{p(i|u)^U}{\sum_i p(i|u)^U} \quad (3)$$

If the weight of an unlabeled item exceeds ε, the superfluous weight is reallocated evenly to other unlabeled items. Apparently, if all the unlabeled items are taken into consideration during model estimation, the huge number of unlabeled items may incur extremely high computational cost. To address the efficiency problem, we propose to pre-select a moderate number of candidate items (called workset items) that are likely to be positive from the unlabeled items. Let $\mathbf{UI}^{\mathrm{P}}$ and $\mathbf{UI}^{\mathrm{W}}$ be the set of observed positive entries and discovered workset entries in **UI**, respectively. Figure 2 outlines the proposed approach to dealing with the sparsity of tagging data.

Workset Selection: Train the TSA model using only $\mathbf{UI}^{\mathrm{P}}$ and select the top-kN ($k \geq 1$) unlabeled items of each user as $\mathbf{UI}^{\mathrm{W}}$. Initially, the weights of all the workset items are the same.

Repeat: Until the likelihood stops to increase or the number of iterations exceeds a given threshold (e.g., 3~5)

- ✧ Update the weights of $\mathbf{UI}^{\mathrm{W}}$ based on the prediction result of the last trained model following equation (3).
- ✧ Initialize $p(i|t)$ based on **IT** derived from the training data and $p(t|u)$ based on the updated **UI** following equation (2).
- ✧ Train a new model based on both $\mathbf{UI}^{\mathrm{P}}$ and $\mathbf{UI}^{\mathrm{W}}$ and then predict the probabilities of entries in $\mathbf{UI}^{\mathrm{W}}$ being observed using equation (1).

Output: The final weights of $\mathbf{UI}^{\mathrm{W}}$ for item recommendation.

Figure 2. A learning framework for CFUI on tagging data

The rationale behind the way CFUI deals with sparsity is that the item and tag profiles are generally incomplete and tend to emphasize on different aspects of users' interests. Thus, propagating users' item and tag preferences iteratively between the item and tag spaces can help to complete and corroborate each other and finally lead to comprehensive representations of users' interests. For example, an MIS researcher may have tagged a lot of journal Web pages with "MIS", but have not had a chance to access the Web pages of premier journals in the MIS area, such as *MISQ*, *ISR* and *JMIS*. In this case, the tag profile of this user is complete on the "MIS" topic while her item profile is obviously not. If this incomplete user-item interaction information is directly used for model training, the user's interest in the topic "MIS" will be improperly downplayed as the most important evidence of this user's interest in MIS (i.e., saving Web pages of premier MIS journals) is missing. Under the CFUI framework, this researcher's potential interest in important MIS Web pages can be effectively identified and expressed in the item weight updating process. Likewise, the tag generalization step can help to refine a user's incomplete tag profile in some aspects with her complete item profile in the same aspects.

4. Empirical Evaluation

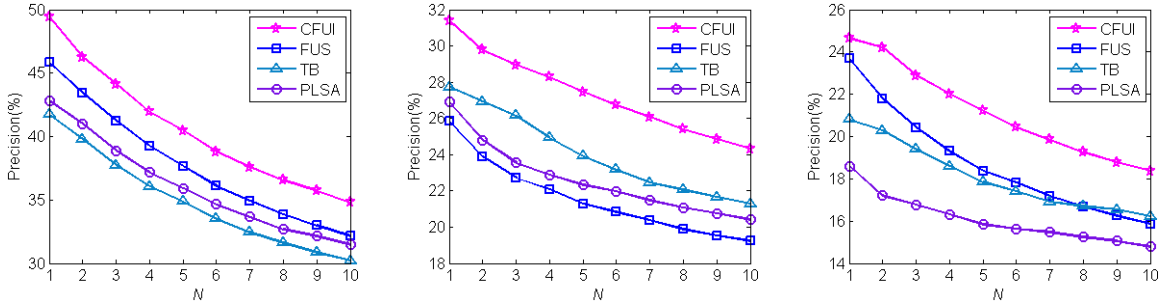
We have evaluated our proposed CFUI approach on three datasets. The first dataset was crawled from Delicious, the largest social bookmarking site. The collected dataset consists of bookmarking data of 5000 users dated from 12/1/2008 to 12/31/2008. The second dataset is a

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. 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 computational efficiency and recommendation quality, we only considered tags that had occurred more than 10 times (5 for CiteULike) in the training set. Table 1 shows the key statistics of the cleaned datasets.

Dataset	Delicious	Citeulike	Bibsonomy
Number of users *m*	177	132	125
Number of items *n*	210	225	388
Number of selected/total tags *l*	65/2251	65/1584	76/2305
Number of user-item interactions *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	39.55	13.83	27.08
Number of items per user	>=10	>=10	>=10
Number of users per item	>=10	>=10	>=8
Frequency of selected tags	>=10	>=5	>=10

Table 1. Dataset description

We compared our approach with a variety of existing tag-based recommendation methods. One of the benchmarks is the state-of-art memory-based method, the fusion (FUS) method (Tso-Sutter et al. 2008), which is a mixture of tag-aware user- and item-based methods. We also compared our approach with existing model-based algorithms in the tagging domain, including the topic-based (TB) method (Peng et al. 2009b) and the probabilistic latent semantic analysis (PLSA) method (Wetzker et al. 2009). To show the capability of the proposed approach to deal with sparse data, we randomly selected 20% of the items of each user for training and withheld the rest for prediction. In the prediction phase, we recommended the top 1, 2, …, 10 items to each user and then compare them with items in the test set. The evaluation metrics adopted in our experiment were the commonly used ones for ranked list recommendation, namely, precision, recall, F-measure, and rankscore (Peng et al. 2010). The precision curves are shown in Figure 3. Results for other metrics are similar and are omitted due to the space limit.



(a) Delicious (b) CiteULike (c) Bibsonomy

Figure 3. Experiment results on three datasets

As can be seen from Figure 3, the proposed method outperformed other algorithms remarkably on all three datasets (ANOVA test on precision according to 20 runs demonstrates that the

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

difference between CFUI and other methods are significant, with $p<0.001$ on all datasets). It is interesting that the relative performance of the three benchmark algorithms vary largely on different datasets, which demonstrates the different characteristic of the three tested datasets and the general applicability of our approach. In particular, while the PLSA method has a model and inference procedure similar to those of CFUI, its prediction accuracy is substantially inferior, demonstrating the utility of exploiting unlabeled items for CF. Moreover, we have tried to preprocess the datasets with smaller pruning thresholds and found that the relative performance of the tested algorithms kept unchanged on larger and sparser datasets.

5. Conclusions and Future Research

Sparsity is a critical problem that limits the performance of tag-based CF methods. To deal with this problem, we have proposed a novel learning framework to corroborate the user-tag and user-item interactions iteratively. Empirical evaluation on three real-world tagging datasets demonstrates the effectiveness of the proposed approach. Our work opens up prominent avenues for further research. First, the generalizability of our findings may be tested through more comprehensive experiments, especially with larger and sparser datasets. Second, as we have for the first time explored the use of unlabeled items for tag-based CF and shown its utility, more investigations may be carried out along this line. Third, the idea of leveraging unlabeled items may also be explored in other recommendation contexts.

Acknowledgements

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

References

Hofmann, T. "Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis," in: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp. 259-266.

Lee, W., and Liu, B. "Learning with Positive and Unlabeled Examples using Weighted Logistic Regression," in: *Proceedings of the 20th International Conference on Machine Learning*, 2003, pp. 448-455.

Liu, B., Dai, Y., Li, X., Lee, W.S., and Yu, P.S. "Building Text Classifiers using Positive and Unlabeled Examples," in: *Proceedings of the 3rd IEEE International Conference on Data Mining*, 2003, pp. 179-186.

Liu, B., Lee, W.S., Yu, P.S., and Li, X. "Partially Supervised Classification of Text Documents," in: *Proceedings of the 19th International Conference on Machine Learning*, 2002, pp. 387-394.

Peng, J., and Zeng, D. "Exploring Information Hidden in Tags: A Subject-based Item Recommendation Approach," in: *Proceedings of 19th Workshop on Information Technologies and Systems*, 2009a, pp. 73-78.

Peng, J., and Zeng, D. "Topic-based Web Page Recommendation Using Tags," in: *Proceedings of the 2009 IEEE international conference on Intelligence and security informatics*, 2009b, pp. 269-271.

Peng, J., Zeng, D., Zhao, H., and Wang, F.-Y. "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.

Tso-Sutter, K.H.L., Marinho, L.B., and Schmidt-Thieme, L. "Tag-aware Recommender Systems by Fusion of Collaborative Filtering Algorithms," in: *Proceedings of the ACM symposium on Applied computing*, 2008, pp. 1995-1999.

Wetzker, R., Umbrath, W., and Said, A. "A Hybrid Approach to Item Recommendation in Folksonomies," in: *Proceedings of the WSDM'09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, 2009, pp. 25-29.

Yu, H., Han, J., and Chang, K.C.-C. "PEBL: Web Page Classification without Negative Examples," *IEEE Transactions on Knowledge and Data Engineering* (16:1) 2004, pp 70-81.

Zhang, Z.-K., Zhou, T., and Zhang, Y.-C. "Personalized Recommendation via Integrated Diffusion on User-Item-Tag Tripartite Graphs," *Physica A: Statistical Mechanics and its Applications* (389:1) 2010, pp 179-186.

Zhen, Y., Li, W., and Yeung, D. "TagiCoFi: Tag Informed Collaborative Filtering," in: *Proceedings of the 3rd ACM conference on Recommender systems*, 2009, pp. 69-76.