

A hybrid approach to item recommendation in folksonomies

Robert Wetzker
DAI Labor
Technische Universität Berlin
robert.wetzker@dai-labor.de

Winfried Umbrath
DAI Labor
Technische Universität Berlin
winfried.umbrath@dai-labor.de

Alan Said
DAI Labor
Technische Universität Berlin
alan.said@dai-labor.de

ABSTRACT

In this paper we consider the problem of item recommendation in collaborative tagging communities, so called folksonomies, where users annotate interesting items with tags. Rather than following a collaborative filtering or annotation-based approach to recommendation, we extend the probabilistic latent semantic analysis (PLSA) approach and present a unified recommendation model which evolves from item user and item tag co-occurrences in parallel. The inclusion of tags reduces known collaborative filtering problems related to overfitting and allows for higher quality recommendations. Experimental results on a large snapshot of the *delicious* bookmarking service show the scalability of our approach and an improved recommendation quality compared to two-mode collaborative or annotation based methods.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

Keywords

folksonomies, tagging, recommendation, PLSA, delicious

1. INTRODUCTION

Collaborative tagging has become the most common content categorization technique of the Web 2.0 age, allowing the creators or consumers of content to assign freely chosen keywords (tags) in order to simplify later retrieval. The concept of tagging has been proven successful in multiple areas, enabling the success of resource sharing services such as *delicious*, *last.fm* or *flickr*. These social tagging communities have become known as folksonomies. The distributed, user-centric annotation of (web-) content was shown to provide relevant meta-data and is expected to boost the semantic quality of labels[7].

In this paper we consider the problem of item recommenda-

tion in large folksonomies. Our intent is to help folksonomy users discover new interesting items based on their item history. To this mean, we exploit the semantic contribution of tags and extend the classical collaborative filtering approach by user-generated annotations. This allows us to improve recommendations by calculating item similarities not only based on the user item distribution, but also in the tag space. Our approach is thus algorithmically related to previously presented work on hybrid recommender systems that combine collaborative and content-based models for better recommendation quality. However, instead of relying on the often complex, hard to extract and possibly heterogeneous content of items, we only consider an item's annotations.

The probabilistic latent semantic analysis (PLSA), as introduced by Hofmann [9], has been shown to improve recommendation quality for various settings by assuming a latent lower dimensional topic model as origin of observed co-occurrence distributions. Our approach extends the PLSA algorithm such that the topic model is estimated from the item user as well as the item tag observations in parallel. This allows us to benefit from user annotations during the recommender training and to combine collaborative and annotation based models into a unified representation.

The evaluation of our recommender system is performed on a large snapshot of 109 million bookmarks of the *delicious* on-line bookmarking service¹. The service allows its users to centrally collect and manage their bookmarks by assigning tags. Being one of the first and most researched real world folksonomies, *delicious* represents a congruous evaluation object. Due to its large size, our dataset not only reflects the structure and size of a real world social bookmarking corpus but also allows us to demonstrate the scalability of our approach. Even though we limit our evaluation to social bookmarking, we believe our method is generally applicable to the task of item recommendation in collaborative tagging communities.

1.1 Related Work

Historically, recommender systems are categorized into collaborative filtering, content-based or hybrid systems, where the latter combine, or unify, user and content oriented approaches and have shown to outperform their two-mode counterparts in many scenarios [3]. Even though we do not consider the actual content of items rather item annotations generated by users, our scenario is algorithmically similar to

¹<http://delicious.com>

the fusion of collaborative and content-based models. The authors of [2] present a multi-dimensional technique which incorporates contextual information for an optimized recommender training. The fusion of co-occurrence relationships among multiple types of objects is also proposed in [15], where the authors present a multi-type extension of the latent semantic analysis (LSA) algorithm that outperforms standard LSA in a variety of domains, such as collaborative filtering or text categorization. A more general overview of recommender systems is given in [3, 6].

Probabilistic latent semantic analysis (PLSA) has been shown to improve the quality of collaborative filtering based recommenders [9] by assuming an underlying lower dimensional latent topic model. Similar to our approach, the authors of [5] consider the problem of document clustering and extend the PLSA algorithm to combine content-based and hyperlink-based similarities into a unified model. Model fusion using PLSA was also successfully applied to the discovery of navigational patterns on the Web [12], in music recommendation combining multiple similarity measures [4] and for the cross-domain knowledge transfer [17].

Until recently, research on recommender systems and folksonomies mainly focused on tag recommendation [8, 11, 13]. The authors of [14] enrich a collaborative movie recommender by incorporating tags that were assigned to each movie in external folksonomies. Finally, [1] proposes to smooth tag item distributions based on usage patterns in order to improve resource retrieval.

The remainder of this paper is structured as follows. In section 2 we extend the PLSA model to a recommendation model which unifies annotations and usage patterns. We then present our experimental settings and the results obtained from our experiments in sections 3 and 4 and summarize our conclusions and ideas for future directions in section 5.

2. MODEL FUSION USING PLSA

According to [10], a folksonomy can be described as a tripartite graph whose vertex set is partitioned into three disjoint sets of users $U = \{u_1, \dots, u_l\}$, tags $T = \{t_1, \dots, t_n\}$ and items $I = \{i_1, \dots, i_m\}$. We simplify this model to two bipartite models where the collaborative filtering model IU is built from the item user co-occurrence counts $f(i, u)$ and the annotation-based model IT derives from the co-occurrence counts between items and tags $f(i, t)$. In the case of social bookmarking IU becomes a binary matrix ($f(i, u) \in \{0, 1\}$), as users can bookmark a given web resource only once. Given our model, we want to recommend the most interesting new items from I to a user u_l given the user's item history.

The aspect model of PLSA associates the co-occurrence of observations with a hidden topic variable $Z = \{z_1, \dots, z_k\}$. In the context of collaborative filtering an observation corresponds to the bookmarking of an item by a user and all observations are given by the co-occurrence matrix IU . Users and items are assumed independent given the topic variable Z . Applying the aspect model, the probability that an item was bookmarked by a given user can be computed by sum-

ming over all latent variables Z :

$$P(i_m|u_l) = \sum_k P(i_m|z_k)P(z_k|u_l), \quad (1)$$

For the annotation-based scenario we assume the same hidden topics as origin of the item tag co-occurrence observations given by IT . Analog to (1), the conditional probability between tags and items can be written as:

$$P(i_m|t_n) = \sum_k P(i_m|z_k)P(z_k|t_n). \quad (2)$$

Following the procedure in [5], we can now combine both models based on the common factor $P(i_m|z_k)$ by maximizing the log-likelihood function

$$L = \sum_m \left[\alpha \sum_l f(i_m, u_l) \log P(i_m|u_l) + (1 - \alpha) \sum_n f(i_m, t_n) \log P(i_m|t_n) \right], \quad (3)$$

where α is a predefined weight for the influence of each two-mode model.

Using the Expectation-Maximization (EM) algorithm [5] we then perform maximum likelihood parameter estimation for the aspect model. During the expectation (E) step we first calculate the posterior probabilities:

$$P(z_k|u_l, i_m) = \frac{P(i_m|z_k)P(z_k|u_l)}{P(i_m|u_l)}$$

$$P(z_k|t_n, i_m) = \frac{P(i_m|z_k)P(z_k|t_n)}{P(i_m|t_n)},$$

and then re-estimate parameters in the maximization (M) step as follows:

$$P(z_k|u_l) \propto \sum_m f(u_l, i_m)P(z_k|u_l, i_m) \quad (4)$$

$$P(z_k|t_n) \propto \sum_m f(t_n, i_m)P(z_k|t_n, i_m) \quad (5)$$

$$p(i_m|z_k) \propto \alpha \sum_l f(u_l, i_m)P(z_k|u_l, i_m) + (1 - \alpha) \sum_n f(t_n, i_m)P(z_k|t_n, i_m) \quad (6)$$

Based on the iterative computation of the above E and M steps, the EM algorithm monotonically increases the likelihood of the combined model on the observed data. Using the α parameter, our new model can be easily reduced to a collaborative filtering or annotation-based model by setting α to 1.0 or 0.0 respectively.

We can now recommend items to a user u_l weighted by the probability $P(i_m|u_l)$ from equation (1)². For items already bookmarked by the user in the training data we set this weight to 0, thus they are appended to the end of the recommended item list.

²It is also possible to recommend items with respect to a given tag t_n based on equation (2).

3. EXPERIMENTS

3.1 Dataset

We evaluate our approach on a corpus of originally 142 million bookmarks from the *delicious* bookmarking service. These bookmarks were collected between September 19, 2007 and January 22, 2008. This is the same corpus as described in [16]. A previous analysis unveiled that the original corpus was highly polluted by spam [16]. In order to get meaningful results, we limit the impact of spam users on the initial corpus as their anomalous behavior would strongly interfere with our analysis. To identify spam users we employ a common spam usage pattern. As was shown in [16], many spam users try to heighten the visibility of their web domains and consequently post a very high number of URLs to very few domains. To reduce the spam ratio within the data, we excluded the top 10 percent of users with the highest URLs per domain rate from our analysis. The filtered data set consists of 109 million bookmarks.

3.2 Experimental setup

Our experiments are performed on a 6 month section, July–December 2007, of the spam filtered corpus. We remove items, users and tags occurring less than 10 times within these 6 months, thus generating the p-core 10 of the initial tripartite graph. We then split the remaining data into 6 monthly snapshots, each containing approximately 1.6 million bookmarks corresponding to more than 5.6 million tag assignments. For each month, the numbers of elements in each dimension, I , T , U , roughly sum up to 200,000, 95,000 and 200,000 respectively. The corresponding co-occurrence matrices IU and IT are very sparse and only contain a percentage of around 0.004 and 0.012 non-zero entries. All results presented in this paper are averaged over all 6 months.

For each month we randomly select 80% of all bookmarks for training and the remaining bookmarks are saved for testing. This split is done on a per user basis. The bookmarks from the training period are then used to create the co-occurrence matrices IU and IT on which the recommenders are trained. After training we select a random set of 1000 users, with at least 10 test items each. For every user we recommend all items sorted by $P(i_m|u_i)$ where items bookmarked by the user during the training or before the evaluated month are weighted with $P(i_m|u_i) = 0$. The quality of the recommended item list is evaluated using performance measures commonly found in relevant literature [6], such as the area under curve (AUC) value of the receiver operating characteristic (ROC) curves or the precision measure. Results are averaged over all test users.

Multiple variables have to be taken into consideration when evaluating recommender systems. Among these is the question whether items that do not appear in the training data should be included into the evaluation. As we are only interested in the relative improvement of our approach, we remove all previously unseen items. For the same reasons, we also exclude items which appear in the training but not in the test data.

All obtained results are compared with the performance of a baseline recommender (*most-popular*) that weights items by how often they were bookmarked during the training period. These item weights are global and do not take into

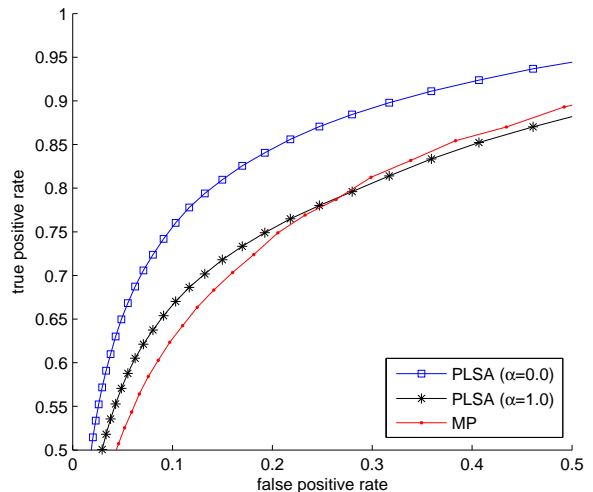


Figure 1: Magnified ROC curves for the item recommendation task on the *delicious* dataset. The number of latent topics (k) is set to 80 for the annotation-based PLSA recommender ($\alpha = 0.0$) and to 5 for the collaborative version ($\alpha = 1.0$). The MP line represents the performance of a *most-popular* baseline classifier.

account user preferences. However, as for the PLSA recommender, we set the weight of previously bookmarked items to 0. Most-popular recommenders have become a standard feature of Web 2.0 resource sharing communities.

4. RESULTS

Figure 1 presents a section of the ROC curves for the collaborative filtering ($\alpha = 1.0$) and the annotation-based ($\alpha = 0.0$) PLSA recommenders with the number of latent topics k set to 5 and 80 respectively. All values are averaged over the 6 evaluation months. The figure shows a significant boost in recommendation quality when using an annotation-based PLSA recommender ($\alpha = 0$) reaching AUC values of 0.9022 compared to 0.8425 for the *most-popular* recommender. For the collaborative method ($\alpha = 1$) with an optimal k set to 5 we obtain an AUC result only slightly above the baseline performance (0.8467). However, the collaborative recommender performs better for small numbers of recommended items.

Table 1: Area under curve (AUC) for different parameter settings. Bold entries indicate the best AUC value for a given number of latent topics k .

α/k	1	5	10	20	40	80
0.0	0.8402	0.8736	0.8877	0.8936	0.9004	0.9022
0.2	0.8416	0.8491	0.8936	0.8975	0.9009	0.9023
0.4	0.8430	0.8419	0.8944	0.8986	0.8954	0.8935
0.6	0.8437	0.8423	0.8720	0.8916	0.8848	0.8722
0.8	0.8438	0.8418	0.8727	0.8678	0.8461	0.8178
1.0	0.8435	0.8467	0.8348	0.8110	0.7766	0.7466

Table 1 compares the resulting AUC values for the PLSA recommender and different choices of α and k . Once again,

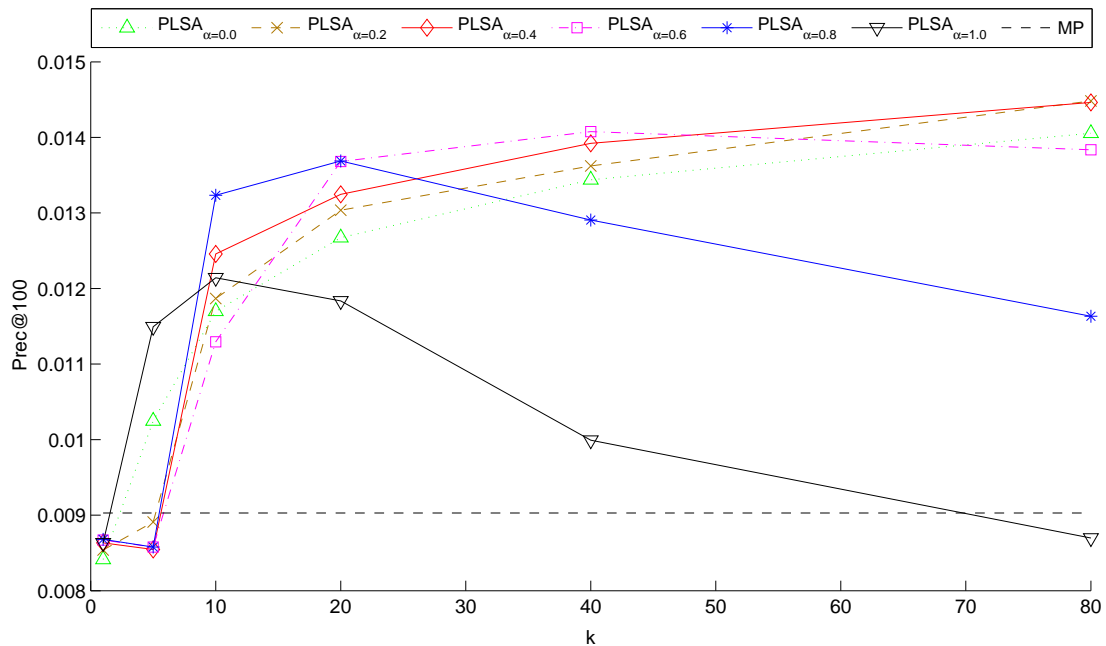


Figure 2: Prec@100 for the item recommendation task on the *delicious* dataset. The figure shows the effect of different values of α and k on the performance of a hybrid PLSA recommender. The *MP* line represents the performance of a *most-popular* baseline classifier.

we find that for an increasing number of latent topics k the collaborative filtering recommender behaves very differently from the annotation-based recommender. While the AUC values increase with k when $\alpha = 0$, quality decreases in the collaborative filtering scenario. We believe this is due to the fact that there exist more item tag than item user relations which in turn may lead to faster overfitting in the sparser user item model.

The results also indicate that recommendation quality can be improved using model fusion independent of the number of latent topics, although this effect lessens with higher k values. This observation is backed by the results plotted in Figure 2. The figure shows the precision when recommending the 100 items with the highest value $P(i_m|u_l)$ (Prec@100). We find that few latent topics cannot cope with the complexity of the unified model, and the two-mode recommenders perform better for k values below 10. However, once the number of latent classes is able to fit the model, we see a firm improvement in recommendation quality. Furthermore, we observe that for more than 10 latent classes our hybrid recommenders constantly outperform their two-mode counterparts. We also find, that models with a high α value tend to overfit earlier than more annotation oriented models. We believe that this tendency is caused by the denser nature of the item tag graph, although this has to be further investigated. The most interesting observation of our evaluations is the overall bad performance of the collaborative filtering recommender, and the drastically increased precision when considering annotations during the model building process. Consistent with the AUC results we find that any PLSA recommender with an appropriate k performs better than the *most-popular* baseline recommender.

5. CONCLUSIONS

In this paper we have shown that a hybrid approach to the task of item recommendation in folksonomies that includes user generated annotations produces better results than a standard collaborative filtering or annotation-based method. We presented an extension to the PLSA algorithm in order to combine usage and tagging information into a unified model. The evaluation of our approach, which was performed on a large scale corpus of a real world folksonomy, showed that the presented recommender not only outperforms two-mode methods but, because of its low dimensional data representation, also decreases recommendation time. For future work, we plan to extend our investigations to tensor-based models that fully reflect the tri-partite nature of collaborative tagging systems and were shown to improve recommendation quality in other settings [13].

6. REFERENCES

- [1] Rabeeh Abbasi and Steffen Staab, ‘Introducing triple play for improved resource retrieval in collaborative tagging systems’, in *In: Proc. of ECIR’08 Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR 2008)*, (3 2008).
- [2] Gediminas Adomavicius, Ramesh Sankaranarayanan, Shahana Sen, and Alexander Tuzhilin, ‘Incorporating contextual information in recommender systems using a multidimensional approach’, *ACM Trans. Inf. Syst.*, **23**(1), 103–145, (2005).
- [3] Gediminas Adomavicius and Alexander Tuzhilin, ‘Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions.’, *IEEE Trans. Knowl. Data Eng.*, **17**(6), 734–749, (2005).

- [4] J. Arenas-García, A. Meng, K. B. Petersen, T. L. Schiøler, L. K. Hansen, and J. Larsen, 'Unveiling music structure via PLSA similarity fusion', in *IEEE International Workshop on Machine Learning for Signal Processing*, pp. 419–424. IEEE Press, (aug 2007).
- [5] David A. Cohn and Thomas Hofmann, 'The missing link - a probabilistic model of document content and hypertext connectivity', in *NIPS*, eds., Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, pp. 430–436. MIT Press, (2000).
- [6] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl, 'Evaluating collaborative filtering recommender systems', *ACM Trans. Inf. Syst.*, **22**(1), 5–53, (2004).
- [7] Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina, 'Can social bookmarking improve web search?', in *WSDM '08: Proc. of the int. conf. on Web search and web data mining*, pp. 195–206, New York, NY, USA, (2008). ACM.
- [8] Paul Heymann, Daniel Ramage, and Hector Garcia-Molina, 'Social tag prediction', in *SIGIR '08: Proc. of the 31st ann. int. ACM SIGIR conf. on Research and development in information retrieval*, pp. 531–538, New York, NY, USA, (2008). ACM.
- [9] Thomas Hofmann, 'Probabilistic latent semantic analysis', in *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, (1999).
- [10] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme, 'Information retrieval in folksonomies: Search and ranking', in *ESWC*, eds., York Sure and John Domingue, volume 4011 of *Lecture Notes in Computer Science*, pp. 411–426. Springer, (2006).
- [11] Robert Jäschke, Leandro Marinho, Andreas Hotho, Lars Schmidt-Thieme, and Gerd Stumme, 'Tag recommendations in folksonomies', in *Workshop Proceedings of Lernen - Wissensentdeckung - Adaptivität (LWA 2007)*, ed., Alexander Hinneburg, pp. 13–20, (sep 2007).
- [12] Xin Jin, Yanzan Zhou, and Bamshad Mobasher, 'Web usage mining based on probabilistic latent semantic analysis.', in *KDD*, eds., Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel, pp. 197–205. ACM, (2004).
- [13] Panagiotis Symeonidis, Alexandros Nanopoulos, and Yannis Manolopoulos, 'Tag recommendations based on tensor dimensionality reduction', in *RecSys '08: Proc. of the 2008 ACM conf. on Recommender systems*, pp. 43–50, New York, NY, USA, (2008). ACM.
- [14] Martin Szomszor, Ciro Cattuto, Harith Alani, Kieron O'Hara, Andrea Baldassarri, Vittorio Loreto, and Vito D. P. Servedio, 'Folksonomies, the semantic web, and movie recommendation', in *Bridging the Gap between Semantic Web and Web 2.0 (SemNet 2007)*, pp. 71–84, (2007).
- [15] Xuanhui Wang, Jian-Tao Sun, Zheng Chen, and ChengXiang Zhai, 'Latent semantic analysis for multiple-type interrelated data objects', in *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 236–243, New York, NY, USA, (2006). ACM.
- [16] Robert Wetzker, Carsten Zimmermann, and Christian Bauckhage, 'Analyzing social bookmarking systems: A del.icio.us cookbook', in *Mining Social Data (MSoDa) Workshop Proceedings*, pp. 26–30. ECAI 2008, (July 2008).
- [17] Gui-Rong Xue, Wenyuan Dai, Qiang Yang, and Yong Yu, 'Topic-bridged pls for cross-domain text classification', in *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 627–634, New York, NY, USA, (2008). ACM.