Tag Recommendations in Folksonomies

Robert Jäschke[1,2], Leandro Marinho[3,4], Andreas Hotho[1],
Lars Schmidt-Thieme[3], and Gerd Stumme[1,2]

[1] Knowledge & Data Engineering Group (KDE), University of Kassel,
Wilhelmshöher Allee 73, 34121 Kassel, Germany
http://www.kde.cs.uni-kassel.de

[2] Research Center L3S, Appelstr. 9a, 30167 Hannover, Germany
http://www.l3s.de

[3] Information Systems and Machine Learning Lab (ISMLL), University of Hildesheim,
Samelsonplatz 1, 31141 Hildesheim, Germany
http://www.ismll.uni-hildesheim.de

[4] Brazilian National Council Scientific and Technological Research (CNPq) scholarship holder

Abstract. Collaborative tagging systems allow users to assign keywords—so called "tags"—to resources. Tags are used for navigation, finding resources and serendipitous browsing and thus provide an immediate benefit for users. These systems usually include tag recommendation mechanisms easing the process of finding good tags for a resource, but also consolidating the tag vocabulary across users. In practice, however, only very basic recommendation strategies are applied.

In this paper we evaluate and compare two recommendation algorithms on large-scale real life datasets: an adaptation of user-based collaborative filtering and a graph-based recommender built on top of FolkRank. We show that both provide better results than non-personalized baseline methods. Especially the graph-based recommender outperforms existing methods considerably.

1 Introduction

Folksonomies are web-based systems that allow users to upload their resources, and to label them with arbitrary words, so-called *tags*. The systems can be distinguished according to what kind of resources are supported. Flickr, for instance, allows the sharing of photos, del.icio.us the sharing of bookmarks, CiteULike[1] and Connotea[2] the sharing of bibliographic references, and Last.fm[3] the sharing of music listening habits. *BibSonomy*,[4] allows to share bookmarks and BibTeX based publication entries simultaneously.

To support users in the tagging process and to expose different facets of a resource, most of the systems offered some kind of tag recommendations already at an early stage. Del.icio.us, for instance, had a tag recommender in June 2005 at the latest,[5] and also included resource recommendations.[6] As of today, nobody has empirically shown

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

[2] http://www.connotea.org

[3] http://www.last.fm

[4] http://www.bibsonomy.org

[5] http://www.socio-kybernetics.net/saurierduval/archive/2005_06_01_archive.html

[6] http://blog.del.icio.us/blog/2005/08/people_who_like.html

J.N. Kok et al. (Eds.): PKDD 2007, LNAI 4702, pp. 506–514, 2007.

the quantitative benefits of recommender systems in such systems. In this paper, we will quantitatively evaluate a tag recommender based on collaborative filtering (introduced in Sec. 3) and a graph based recommender using our ranking algorithm FolkRank (see Sec. 4) on the two real world folksonomy datasets BibSonomy and Last.fm. We make the BibSonomy dataset publicly available for research purposes to stimulate research in the area of folksonomy systems (details in Section 5).

The results we are able to present in Sec. 6 are very encouraging as the graph based approach outperforms all other approaches significantly. As we will see later, this is caused by the ability of FolkRank to exploit the information that is pertinent to the specific user together with input from other users via the integrating structure of the underlying hypergraph.

2 Recommending Tags—Problem Definition and State of the Art

Recommending tags can serve various purposes, such as: increasing the chances of getting a resource annotated, reminding a user what a resource is about and consolidating the vocabulary across the users. In this section we formalize the notion of folksonomies, formulate the tag recommendation problem and briefly describe the state of the art on tag recommendations in folksonomies.

A Formal Model for Folksonomies. A folksonomy $\mathbb{F}$ describes the users U, resources R, and tags T, and the user-based assignment of tags to resources by the ternary relation $Y \subseteq U \times T \times R$. We depict the set of all posts by P. The model of a folksonomy we use here is based on the definition in [9].

Tag Recommender Systems. Recommender systems (RS) in general recommend interesting or personalized information objects to users based on explicit or implicit ratings. Usually RS predict ratings of objects or suggest a list of new objects that the user hopefully will like the most. In tag recommender systems the recommendations are, for a given user $u \in U$ and a given resource $r \in R$, a set $\tilde{T}(u, r) \subseteq T$ of tags. In many cases, $\tilde{T}(u, r)$ is computed by first generating a ranking on the set of tags according to some quality or relevance criterion, from which then the top n elements are selected.

Related work. General overviews on the rather young area of folksonomy systems and their strengths and weaknesses are given in [7,11,12]. In [13], Mika defines a model of semantic-social networks for extracting lightweight ontologies from del.icio.us. Recently, work on more specialized topics such as structure mining on folksonomies—e. g. to visualize trends [5] and patterns [16] in users' tagging behavior—as well as ranking of folksonomy contents [9], analyzing the semiotic dynamics of the tagging vocabulary [3], or the dynamics and semantics [6] have been presented.

The literature concerning the problem of tag recommendations in folksonomies is still sparse. The existent approaches [2,10,14] usually adapt methods from collaborative filtering or information retrieval. The standard tag recommenders, in practice, are services that provide the most-popular tags used for a particular resource by means of tag clouds, i.e., the most frequent used tags are depicted in a larger font or otherwise emphasized. These approaches address important aspects of the problem, but they still diverge on the experimental protocol, notion of tag relevance and metrics used, what makes further comparisons difficult.

3 Collaborative Filtering

Due to its simplicity and promising results, collaborative filtering (CF) has been one of the most dominant methods used in recommender systems. In the next section we recall the basic principles and then present the details of the adaptation to folksonomies.

Basic CF principle. The idea is to suggest new objects or to predict the utility of a certain object based on the opinion of like-minded users [15]. In CF, for m users and n objects, the user profiles are represented in a user-object matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$. The matrix can be decomposed into row vectors:

$$\mathbf{X} := [\vec{x}_1, ..., \vec{x}_m]^\top \text{ with } \vec{x}_u := [x_{u,1}, ..., x_{u,n}], \text{ for } u := 1, \dots, m,$$

where $x_{u,o}$ indicates that user u rated object o by $x_{u,o} \in \mathbb{R}$. Each row vector $\vec{x}_u$ corresponds thus to a user profile representing the object ratings of a particular user. This decomposition leads to user-based CF (see [4] for item-based algorithms).

Now, one can compute, for a given user u, the recommendation as follows. First, based on matrix $\mathbf{X}$ and for a given k, the set N_u^k of the k users that are most similar to user $u \in U$ are computed: $N_u^k := \arg\max^k_{v \in U} \operatorname{sim}(\vec{x}_u, \vec{x}_v)$ where the superscript in the $\arg\max$ function indicates the number k of neighbors to be returned, and sim is regarded (in our setting) as the cosine similarity measure. Then, for a given $n \in \mathbb{N}$, the top n recommendations consist of a list of objects ranked by decreasing frequency of occurrence in the ratings of the neighbors (see Eq. 1 below for the folksonomy case).

CF for Tag Recommendations in Folksonomies. Because of the ternary relational nature of folksonomies, traditional CF cannot be applied directly, unless we reduce the ternary relation Y to a lower dimensional space. To this end we consider as matrix $\mathbf{X}$ alternatively the two 2-dimensional projections $\pi_{UR}Y \in \{0,1\}^{|U| \times |R|}$ with $(\pi_{UR}Y)_{u,r} := 1$ if there exists $t \in T$ s. t. $(u,t,r) \in Y$ and 0 else and $\pi_{UT}Y \in \{0,1\}^{|U| \times |T|}$ with $(\pi_{UT}Y)_{u,t} := 1$ if there exists $r \in R$ s. t. $(u,t,r) \in Y$ and 0 else. The projections preserve the user information, and lead to log-based like recommender systems based on occurrence or non-occurrence of resources or tags, resp., with the users. Notice that now we have two possible setups in which the k-neighborhood N_u^k of a user u can be formed, by considering either the resources or the tags as objects.

Having defined matrix $\mathbf{X}$, and having decided whether to use $\pi_{UR}Y$ or $\pi_{UT}Y$ for computing user neighborhoods, we have the required setup to apply collaborative filtering. For determining, for a given user u, a given resource r, and some $n \in \mathbb{N}$, the set $\tilde{T}(u,r)$ of n recommended tags, we compute first N_u^k as described above, followed by:

$$\tilde{T}(u,r) := \underset{t \in T}{\overset{n}{\arg\max}} \sum_{v \in N_u^k} \operatorname{sim}(\vec{x}_u, \vec{x}_v) \delta(v,t,r) \qquad (1)$$

where $\delta(v,t,r) := 1$ if $(v,t,r) \in Y$ and 0 else.

4 A Graph Based Approach

The seminal PageRank algorithm reflects the idea that a web page is important if there are many pages linking to it, and if those pages are important themselves. In [9], we employed the same underlying principle for Google-like search and ranking in folksonomies. The key idea of our FolkRank algorithm is that a resource which is tagged with important tags by important users becomes important itself. The same holds, symmetrically, for tags and users, thus we have a graph of vertices which are mutually reinforcing each other by spreading their weights.

For generating a tag recommendation for a given user/resource pair (u, r), we compute the ranking as described in [9], and then restrict the result set $\tilde{T}(u, r)$ to the top n tag nodes.

5 Evaluation

In this section we first describe the datasets we used, how we prepared the data, the methodology deployed to measure the performance, and which algorithms we used, together with their specific settings.

Datasets. To evaluate the proposed recommendation techniques we have chosen datasets from two different folksonomy systems: *BibSonomy* and *Last.fm*. Table 1 gives an overview on the datasets. For both datasets we disregarded if the tags had lower or upper case.

BibSonomy. Since three of the authors have participated in the development of BibSonomy, [7] we were able to create a complete snapshot of all users, resources (both publication references and bookmarks) and tags publicly available at April 30, 2007, 23:59:59 CEST.[8] From the snapshot we excluded the posts from the DBLP computer science bibliography[9] since they are automatically inserted and all owned by one user and all tagged with the same tag (*dblp*). Therefore they do not provide meaningful information for the analysis.

Last.fm. The data for Last.fm[10] was gathered during July 2006, partly through the web services API (collecting user nicknames), partly crawling the Last.fm site. Here the resources are artist names, which are already normalized by the system.

Core computation. Many recommendation algorithms suffer from sparse data or the "long tail" of items which were used by only few users. Hence, to increase the chances of good results for all algorithms (with exception of the most popular tags recommender) we will restrict the evaluation to the "dense" part of the folksonomy, for which

[7] http://www.bibsonomy.org

[8] On request to bibsonomy@cs.uni-kassel.de a snapshot of BibSonomy is available for research purposes.

[9] http://www.informatik.uni-trier.de/~ley/db/

[10] http://www.last.fm

Table 1. Characteristics of the used datasets

dataset	$\|U\|$	$\|T\|$	$\|R\|$	$\|Y\|$	$\|P\|$	date	$k_{\max}$
BibSonomy	1,037	28,648	86,563	341,183	96,972	2007-04-30	7
Last.fm	3,746	10,848	5,197	299,520	100,101	2006-07-01	20

Table 2. Characteristics of the p-cores at level k

dataset	k	$\|U\|$	$\|T\|$	$\|R\|$	$\|Y\|$	$\|P\|$
BibSonomy	5	116	412	361	10,148	2,522
Last.fm	10	2,917	2,045	1,853	219,702	75,565

we adapt the notion of a p-core [1] to tri-partite hypergraphs. The p-core of level k has the property, that each user, tag and resource has/occurs in at least k posts.

An overview on the p-cores we used for our datasets is given in Table 2. For BibSonomy, we used $k = 5$ instead of 10 because of its smaller size. The largest k for which a p-core exists is listed, for each dataset, in the last column of Table 1.

Evaluation methodology. To evaluate the recommenders we used a variant of the leave-one-out hold-out estimation [8] which we call *LeavePostOut*. In all datasets, we picked, for each user, one of his posts p randomly. The task of the different recommenders was then to predict the tags of this post, based on the folksonomy $\mathbb{F} \setminus \{p\}$.

As performance measures we use precision and recall which are standard in such scenarios [8]. With r being the resource from the randomly picked post of user u and $\tilde{T}(u,r)$ the set of recommended tags, recall and precision are defined as

$$\operatorname{recall}(\tilde{T}(u,r)) = \frac{1}{|U|} \sum_{u \in U} \frac{|\operatorname{tags}(u,r) \cap \tilde{T}(u,r)|}{|\operatorname{tags}(u,r)|} \tag{2}$$

$$\operatorname{precision}(\tilde{T}(u,r)) = \frac{1}{|U|} \sum_{u \in U} \frac{|\operatorname{tags}(u,r) \cap \tilde{T}(u,r)|}{|\tilde{T}(u,r)|}. \tag{3}$$

For each of the algorithms of our evaluation we will now describe briefly the specific settings used to run them.

Most popular tags. For each tag we counted in how many posts it occurs globally and used the top tags (ranked by occurence count) as recommendations.

Most popular tags by resource. For a given resource we counted for all tags in how many posts they occur together with that resource. We then used the tags that occured most often together with that resource as recommendation.

Adapted PageRank. With the parameter $d = 0.7$ we stopped computation after 10 iterations or when the distance between two consecutive weight vectors was less than 10^{-6}. In $\vec{p}$, we gave higher weights to the user and the resource from the post which was chosen. While each user, tag and resource got a preference weight of 1, the user and resource from that particular post got a preference weight of $1 + |U|$ and $1 + |R|$, resp.

FolkRank. The same parameter and preference weights were used as in the adapted PageRank.

Collaborative Filtering UT. For this collaborative filtering algorithm the neighborhood is computed based on the user-tag matrix $\pi_{UT}Y$. The only parameter to be tuned in the CF based algorithms is the number k of best neighbors. For that, multiple runs where performed where k was successively incremented until a point where no more improvements in the results were observed. For this approach the best values for k were 20 for the BibSonomy and 60 for the Last.fm dataset.

Collaborative Filtering UR. Here the neighborhood is computed based on the user-resource matrix $\pi_{UR}Y$. For this approach the best values for k were 30 for the BibSonomy and 100 for the Last.fm dataset.

6 Results

In this section we present and describe the results of the evaluation. We will see that both datasets show the same overall behavior: ‘most popular tags’ is outperformed by all other approaches; the CF-UT algorithm performs slightly better than and the CF-UR approach approx. as good as the ‘most popular tag by resource’, and FolkRank uniformly provides significantly better results.

The diagrams 1 and 2 show precision-recall plots as usual. A datapoint on a curve stands for the number of tags used for recommendation (starting with the highest ranked tag on the left of the curve and ending with ten tags on the right). Hence, the steady decay of all curves in both plots means that the more tags of the recommendation are regarded, the better the recall and the worse the precision will be.

BibSonomy. Figure 1 shows the precision and recall of the chosen algorithms. The top-rightmost curve depicts the performance of FolkRank and it can clearly be seen that the graph based algorithm outperforms the other methods in both precision and recall. With ten recommended tags the recall reaches up to 80%, while the second best results only reach around 65% with a comparable precision. While CF-UT, CF-UR and the ‘most popular tags by resource’ algorithms have a quite similiar performance, the adapted PageRank is significantly worse, especially with its dropdown of precision already after the third recommended tag. Finally, using the most popular tags as recommendation gives very poor results in both precision and recall.

Let us now look at Table 3. We will focus here on a phenomenon which is unique for this dataset. With an increasing number of suggested tags, the precision decrease is steeper for FolkRank than for the collaborative filtering and the ‘most popular tags by resource’ algorithm such that the latter two approaches for ten suggested tags finally overtake FolkRank. The reason is that the average number of tags in a post is around 4 for this dataset and while FolkRank can always recommend the maximum number of

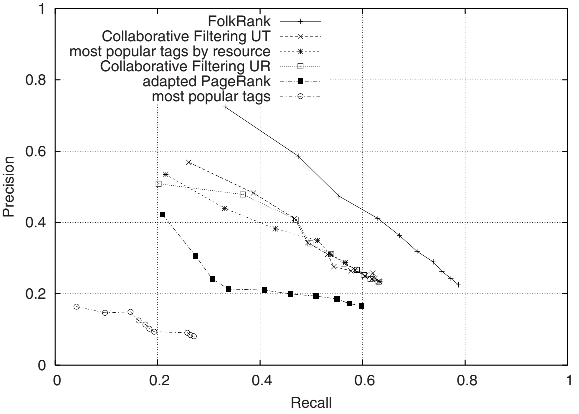

Fig. 1. Recall and Precision for BibSonomy p-core at level 5

Table 3. Precision for BibSonomy p-core at level 5

Number of recommended tags	1	2	3	4	5	6	7	8	9	10
FolkRank	0.724	0.586	0.474	0.412	0.364	0.319	0.289	0.263	0.243	0.225
Collaborative Filtering UT	0.569	0.483	0.411	0.343	0.311	0.276	0.265	0.257	0.243	0.235
most popular tags by resource	0.534	0.440	0.382	0.350	0.311	0.288	0.267	0.250	0.241	0.234
Collaborative Filtering UR	0.509	0.478	0.408	0.341	0.311	0.285	0.267	0.252	0.241	0.234

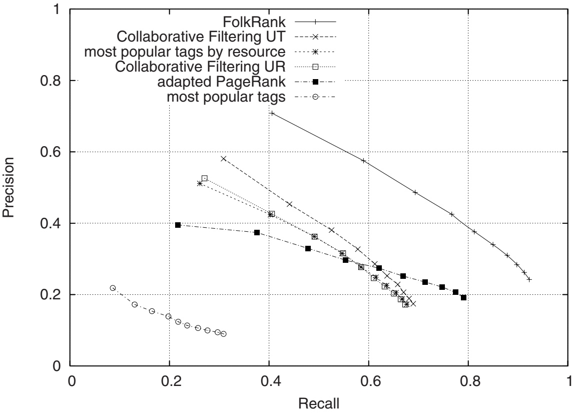

Fig. 2. Recall and Precision for Last.fm p-core at level 10

tags, for the other approaches there are often not enough tags available for recommendation. Hence, less tags are recommended. This is because in the p-core of order 5, for each post, often tags from only four other posts can be used for recommendation with these approaches. Consequently this behaviour is even more noticeable in the p-core of order 3 (which is not shown here).

Last.fm. For this dataset, the recall for FolkRank is considerably higher than for the BibSonomy dataset, see Figure 2. Even when just two tags are recommended, the recall is close to 60 %. Again, the graph based approach outperforms all other methods (CF-UT reaches at most 76 % of the recall of FolkRank). An interesting observation can be made about the adapted PageRank: its recall now is the second best after FolkRank for larger numbers of recommended tags. This shows the overall importance of general terms in this dataset—which have a high influence on the adapted PageRank (cf. Sec. 4).

The results clearly show that the graph based FolkRank algorithm outperforms base line algorithms like 'most popular tags' and collaborative filtering approaches.

Acknowledgement. Part of this research was funded by the EU in the Nepomuk[11] (FP6-027705), Tagora[12] (FP6-2005-34721), and the X-Media[13] (IST-FP6-026978) projects.

References

1. Batagelj, V., Zaversnik, M.: Generalized cores, cs.DS/0202039 (2002), http://arxiv.org/abs/cs/0202039
2. Benz, D., Tso, K., Schmidt-Thieme, L.: Automatic bookmark classification: A collaborative approach. In: Proceedings of the Second Workshop on Innovations in Web Infrastructure (IWI 2006), Edinburgh, Scotland (2006)
3. Cattuto, C., Loreto, V., Pietronero, L.: Collaborative tagging and semiotic dynamics (May 2006), http://arxiv.org/abs/cs/0605015
4. Deshpande, M., Karypis, G.: Item-based top-n recommendation algorithms. ACM Trans. Inf. Syst. 22(1), 143–177 (2004)
5. Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P., Tomkins, A.: Visualizing tags over time. In: Proc. of the 15th International WWW Conference, Edinburgh, Scotland (2006)
6. Halpin, H., Robu, V., Shepard, H.: The dynamics and semantics of collaborative tagging. In: Proceedings of the 1st Semantic Authoring and Annotation Workshop (SAAW'06) (2006)
7. Hammond, T., Hannay, T., Lund, B., Scott, J.: Social Bookmarking Tools (I): A General Review. D-Lib Magazine 11(4) (2005)
8. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. 22(1), 5–53 (2004)
9. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: Search and ranking. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 411–426. Springer, Heidelberg (2006)
10. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM 46(5), 604–632 (1999)
11. Lund, B., Hammond, T., Flack, M., Hannay, T.: Social Bookmarking Tools (II): A Case Study - Connotea. D-Lib Magazine, 11(4) (2005)
12. Mathes, A.: Folksonomies – Cooperative Classification and Communication Through Shared Metadata (December 2004), http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html
13. Mika, P.: Ontologies Are Us: A Unified Model of Social Networks and Semantics. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 522–536. Springer, Heidelberg (2005)

[11] http://nepomuk.semanticdesktop.org

[12] http://www.tagora-project.eu

[13] http://www.x-media-project.org,

14. Mishne, G.: Autotag: a collaborative approach to automated tag assignment for weblog posts. In: WWW '06: Proceedings of the 15th international conference on World Wide Web, pp. 953–954. ACM Press, New York (2006)
15. Sarwar, B.M., Karypis, G., Konstan, J.A., Reidl, J.: Item-based collaborative filtering recommendation algorithms. In: World Wide Web, pp. 285–295 (2001)
16. Schmitz, C., Hotho, A., Jäschke, R., Stumme, G.: Mining association rules in folksonomies. In: Batagelj, V., Bock, H.-H., Ferligoj, A., Žiberna, A. (eds.) Data Science and Classification: Proc. of the 10th IFCS Conf. Studies in Classification, Data Analysis, and Knowledge Organization, pp. 261–270. Springer, Berlin, Heidelberg (2006)