Tag Recommendations in Folksonomies[*]

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

1: Knowledge & Data Engineering Group (KDE), University of Kassel,
Wilhelmshöher Allee 73, 34121 Kassel, Germany
http://www.kde.cs.uni-kassel.de
2: Information Systems and Machine Learning Lab (ISMLL), University of Hildesheim,
Samelsonplatz 1, 31141 Hildesheim, Germany
http://www.ismll.uni-hildesheim.de

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 present two tag recommendation algorithms: an adaptation of user-based collaborative filtering and a graph-based recommender built on top of FolkRank, an adaptation of the well-known PageRank algorithm that can cope with undirected triadic hyperedges. We evaluate and compare both algorithms on large-scale real life datasets and show that both provide better results than non-personalized baseline methods. Especially the graph-based recommender outperforms existing methods considerably.

1 Introduction

Social resource sharing systems 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. Our own system, *BibSonomy*,[4] allows to share bookmarks and BibTeX based publication entries simultaneously.

In their core, these systems are all very similar. Once a user is logged in, he can add a resource to the system, and assign arbitrary tags to it. The collection of all his assignments is his *personomy*, the collection of all personomies constitutes the *folksonomy*. The user can explore his personomy, as well as the personomies of the other users, in all dimensions: for a given user one can see all resources he has uploaded, together with the tags he has assigned to them; when clicking on a resource one sees which other users have uploaded this resource and how they tagged it; and when clicking on a tag one sees who assigned it to which resources. Based on the tags that are assigned to a resource, users are able to search and find her own or other users resources within such systems.

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] However, no algorithmic details were published. We assume that these recommendations basically rely on tag–tag-co-occurrences. As of today, nobody has empirically shown 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 three real world folksonomy datasets. 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 describes the users, resources, and tags, and the user-based assignment of tags to resources. Formally, a *folksonomy* is a tuple $\mathbb{F} := (U, T, R, Y)$ where U, T, and R are finite sets, whose elements are called *users*, *tags* and *resources*, resp., and Y is a ternary relation between them, i. e., $Y \subseteq U \times T \times R$, whose elements are called tag as-

[*] A shortened version of this work has been published at PKDD 2007. [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

signments (*tas* for short).[7] Users are typically described by their user ID, and tags may be arbitrary strings. What is considered a resource depends on the type of system. For instance, in del.icio.us, the resources are URLs, in BibSonomy URLs or publication references, and in last.fm, the resources are artists.

In this paper, we will use an equivalent view on the folksonomy structure. We will consider it as a tripartite (undirected) hypergraph $G = (V, E)$, where $V = U \dot{\cup} T \dot{\cup} R$ is the set of nodes, and $E = \{\{u, t, r\} \mid (u, t, r) \in Y\}$ is the set of hyperedges.

For convenience we also define, for all $u \in U$ and $r \in R$, $\mathrm{tags}(u, r) := \{t \in T \mid (u, t, r) \in Y\}$, i. e., $\mathrm{tags}(u, r)$ is the set of all tags that user u has assigned to resource r. The set of all *posts* of the folksonomy is then $P := \{(u, S, r) \mid u \in U, r \in R, S = \mathrm{tags}(u, r)\}$. Thus, each *post* consists of a user, a resource and all tags that this user has assigned to that resource.

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 [Hammond *et al.*, 2005; Lund *et al.*, 2005; Mathes, 2004]. In [Mika, 2005], 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 [Dubinko *et al.*, 2006] and patterns [Schmitz *et al.*, 2006] in users' tagging behavior—as well as ranking of folksonomy contents [Hotho *et al.*, 2006a], analyzing the semiotic dynamics of the tagging vocabulary [Cattuto *et al.*, 2006], or the dynamics and semantics [Halpin *et al.*, 2006] have been presented.

The literature concerning the problem of tag recommendations in folksonomies is still sparse. The existent approaches usually lay in the collaborative filtering and information retrieval areas. AutoTag [Mishne, 2006], e.g., is a tool that suggests tags for weblog posts using information retrieval techniques. Xu et al. [Xu *et al.*, 2006] introduce a collaborative tag suggestion approach based on the HITS algorithm [Kleinberg, 1999]. A goodness measure for tags, derived from collective user authorities, is iteratively adjusted by a reward-penalty algorithm. Benz et al. [Benz *et al.*, 2006] introduce a collaborative approach for bookmark classification based on a combination of nearest-neighbor-classifiers. There, a keyword recommender plays the role of a collaborative tag recommender, but it is just a component of the overall algorithm, and therefore there is no information about its effectiveness alone. The standard tag recommenders, in practice, are services that provide the most-popular tags used for a particular resource. This is usually done by means of tag clouds where the most frequent used tags are depicted in a larger font or otherwise emphasized.

The approaches described above address important aspects of the problem, but they still diverge on the notion of tag relevance and evaluation protocol used. Xu et al. [Xu *et al.*, 2006], e.g., present no quantitative evaluation, while in [Mishne, 2006], the notion of tag relevance in not entirely defined by the users but partially by experts.

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 [Sarwar *et al.*, 2001]. 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:

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

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. (The matrix can alternatively be represented by its column vectors leading to item-based recommendation 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: $N_u^k := \arg\max^k_{v \in U} \mathrm{sim}(\vec{x}_u, \vec{x}_v)$ 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).

This brief discussion refers only to the user-based CF case, moreover, we consider only the recommendation task since in collaborative tagging systems there are usually no ratings and therefore no prediction. For a detailed description about the item-based CF algorithm see [Deshpande and Karypis, 2004].

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.

[7] In the original definition [Hotho *et al.*, 2006a], we introduced additionally a subtag/supertag relation, which we omit here.

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:

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

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

4 A Graph Based approach

The seminal PageRank algorithm [Brin and Page, 1998] reflects the idea that a web page is important if there are many pages linking to it, and if those pages are important themselves.[8] In [Hotho *et al.*, 2006a], 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. In this section we briefly recall the principles of the FolkRank algorithm, and explain how we use it for generating tag recommendations. More details can be found in [Hotho *et al.*, 2006a].

Because of the different nature of folksonomies compared to the web graph (undirected triadic hyperedges instead of directed binary edges), PageRank cannot be applied directly on folksonomies. In order to employ a weight-spreading ranking scheme on folksonomies, we will overcome this problem in two steps. First, we transform the hypergraph into an undirected graph. Then we apply a differential ranking approach that deals with the skewed structure of the network and the undirectedness of folksonomies, and which allows for topic-specific rankings.

Folksonomy-Adapted Pagerank.

First we convert the folksonomy $\mathbb{F} = (U, T, R, Y)$ into an *un*directed tri-partite graph $G_{\mathbb{F}} = (V, E)$. The set V of nodes of the graph consists of the disjoint union of the sets of tags, users and resources. All co-occurrences of tags and users, users and resources, tags and resources become edges between the respective nodes (more details in [Hotho *et al.*, 2006a]).

The rank of the vertices of the graph are the entries in the fixed point $\vec{w}$ of the weight spreading computation

$$\vec{w} \leftarrow dA\vec{w} + (1-d)\vec{p} \ , \qquad (2)$$

where $\vec{w}$ is a weight vector with one entry for each node, A is the row-stochastic version of the adjacency matrix of the graph $G_{\mathbb{F}}$ defined above, $\vec{p}$ is the preference vector, and $d \in [0,1]$ is determining the influence of $\vec{p}$.

For a global ranking, one will choose $\vec{p} = \mathbf{1}$, i. e., the vector composed by 1's. In order to generate recommendations, however, $\vec{p}$ can be tuned by giving a higher weight to the user and to the resource for which one currently wants to generate a recommendation. The recommendation $\tilde{T}(u,r)$ is then the set of the top n nodes in the ranking, restricted to tags. In the experiments presented below, we will see that this version performs reasonable, but not exceptional. This is in line with our observation in [Hotho *et al.*, 2006a] which showed that the topic-specific rankings are biased by the global graph structure. As a consequence, we developed the following differential approach.

FolkRank—Topic-Specific Ranking.

As the graph $G_{\mathbb{F}}$ that we created in the previous step is undirected, we face the problem that an application of the original PageRank would result in weights that flow in one direction of an edge and then 'swash back' along the same edge in the next iteration, so that one would basically rank the nodes in the folksonomy by their degree distribution. This makes it very difficult for other nodes than those with high edge degree to become highly ranked, no matter what the preference vector is.

This problem is solved by the *differential* approach in FolkRank, which computes a topic-specific ranking of the elements in a folksonomy. Let $\vec{w}_0$ be the fixed point from Equation (2) without preference vector and $\vec{w}_1$ be the fixed point with preference vector $\vec{p}$ and in this case $d = 0.7$. Then $\vec{w} := \vec{w}_1 - \vec{w}_0$ is the final weight vector. Thus, we compute the winners and losers of the mutual reinforcement of nodes when a user/resource pair is given, compared to the baseline without a preference vector. We call the resulting weight $\vec{w}[x]$ of an element x of the folksonomy the *FolkRank* of x.[9] For generating a tag recommendation for a given user/resource pair, we compute the ranking as described above, 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 three different folksonomy systems: *del.icio.us*, *Last.fm* and *BibSonomy*. They have different sizes, different resources to annotate and are probably used by different people. Therefore they form a good basis to test our tag recommendation scenario in a general setting. Table 1 gives an overview on the datasets. For all datasets we disregarded if the tags had lower or upper case since this is the behaviour of most systems when querying them for posts tagged with a certain tag (although often they store the tags as entered by the user).

Del.icio.us. One of the first and most popular folksonomy systems is del.icio.us [10] which exists since the end of 2003. It allows users to tag bookmarks (URLs) and had according to its blog around 1.5 Mio. users in February 2007. We used a dataset from del.icio.us we obtained from July 27 to 30, 2005 [Hotho *et al.*, 2006a]. Since del.icio.us allows its users to *not* tag resources at all (they can be accessed by

[8] This idea was extended in a similar fashion to bipartite subgraphs of the web in HITS [Kleinberg, 1999] and to n-ary directed graphs in [Xi *et al.*, 2004].

[9] In [Hotho *et al.*, 2006a] we showed that $\vec{w}$ provides indeed valuable results on a large-scale real-world dataset while $\vec{w}_1$ provides an unstructured mix of topic-relevant elements with elements having high edge degree. In [Hotho *et al.*, 2006b], we applied this approach for detecting trends over time in folksonomies.

[10] http://del.icio.us

the tag “system:unfiled”) we added those posts with the tag “system:unfiled” to the dataset.

Last.fm. Audioscrobbler[11] is a music engine based on a collection of music profiles. These profiles are built through the use of the company’s flagship product, Last.fm,[12] a system that provides personalized radio stations for its users and updates their profiles using the music they listen to. Audioscrobbler exposes large portions of data through their web services API. The data 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.

BibSonomy. This system allows users to manage and annotate bookmarks and publication references simultanously. Since three of the authors have participated in the development of BibSonomy, [13] 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.[14] From the snapshot we excluded the posts from the DBLP computer science bibliography[15] 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.

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 we adapt the notion of a p-core [Batagelj and Zaversnik, 2002] 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.

To construct the p-core, recall that a folksonomy (U, T, R, Y) can be formalized equivalently as tri-partite hypergraph $G = (V, E)$ with $V = U \dot\cup T \dot\cup R$. First we define, for a subset $\tilde{V}$ of V (with $\tilde{V} = \tilde{U} \dot\cup \tilde{T} \dot\cup \tilde{R}$ and $\tilde{U} \subseteq U, \tilde{T} \subseteq T, \tilde{R} \subseteq R$), the function

$$\operatorname{posts}(v, \tilde{V}) = \begin{cases} \{(v, S, r) \mid r \in \tilde{R}, S = \operatorname{tags}_{\tilde{V}}(v, r)\} & \text{if} \quad v \in \tilde{U} \\ \{(u, v, r) \mid u \in \tilde{U}, r \in \tilde{R}\} & \text{if} \quad v \in \tilde{T} \\ \{(u, S, v) \mid u \in \tilde{U}, S = \operatorname{tags}_{\tilde{V}}(u, v)\} & \text{if} \quad v \in \tilde{R} \end{cases} \tag{3}$$

which assigns to each $v \in \tilde{V}$ the set of all posts in which v occurs. Here, $\operatorname{tags}_{\tilde{V}}$ is defined as in Section 2, but restricted to the subgraph $(\tilde{V}, E_{|\tilde{V}})$. Let $p(v, \tilde{V}) := |\operatorname{posts}(v, \tilde{V})|$. The p-core at level $k \in \mathbb{N}$ is then the subgraph of (V, E) induced by $\tilde{V}$, where $\tilde{V}$ is a maximal subset of V such that, for all $v \in \tilde{V}$, $p(v, \tilde{V}) \geq k$ holds.

Since $p(v, \tilde{V})$ is, for all v, a monotone function in $\tilde{V}$, the p-core at any level k is unique [Batagelj and Zaversnik, 2002], and we can use the algorithm presented in [Batagelj and Zaversnik, 2002] for its computation. 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 [Herlocker *et al.*, 2004] 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 [Herlocker *et al.*, 2004]. 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

$$\operatorname{recall}(\tilde{T}(u, r)) = \frac{1}{|U|} \sum_{u \in U} \frac{|\operatorname{tags}(u, r) \cap \tilde{T}(u, r)|}{|\operatorname{tags}(u, r)|} \tag{4}$$

$$\operatorname{precision}(\tilde{T}(u, r)) = \frac{1}{|U|} \sum_{u \in U} \frac{|\operatorname{tags}(u, r) \cap \tilde{T}(u, r)|}{|\tilde{T}(u, r)|}. \tag{5}$$

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 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. Collaborative filtering algorithm where 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 80 for the deli.icio.us, 60 for the Last.fm, and 20 for the BibSonomy dataset.

[11] http://www.audioscrobbler.net [12] http://www.last.fm

[13] http://www.bibsonomy.org [14] On request to bibsonomy@cs.uni-kassel.de a snapshot of BibSonomy is available for research purposes.

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

Table 1: Characteristics of the used datasets.

dataset	$\lvert U\rvert$	$\lvert T\rvert$	$\lvert R\rvert$	$\lvert Y\rvert$	$\lvert P\rvert$	date	$k_{\max}$
del.icio.us	75,245	456,697	3,158,435	17,780,260	7,698,653	2005-07-30	77
Last.fm	3,746	10,848	5,197	299,520	100,101	2006-07-01	20
BibSonomy	1,037	28,648	86,563	341,183	96,972	2007-04-30	7

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

dataset	k	$\lvert U\rvert$	$\lvert T\rvert$	$\lvert R\rvert$	$\lvert Y\rvert$	$\lvert P\rvert$
del.icio.us	10	37,399	22,170	74,874	7,487,319	3,055,436
Last.fm	10	2,917	2,045	1,853	219,702	75,565
BibSonomy	5	116	412	361	10,148	2,522

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

6 Results

In this section we present and describe the results of the evaluation. We will see that all three 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.

We will now study the results in detail. There are two types of diagrams. The first type of diagram (Figure 1) shows in a straightforward manner how the recall depends on the number of recommended tags. In the other diagrams with usual precision-recall plots (Figures 2 and 3) 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 those plots means that the more tags of the recommendation are regarded, the better the recall and the worse the precision will be.

Del.icio.us. Figure 1 shows how the recall increases, when more tags of the recommendation are used. All algorithms perform significantly better than the baseline based on the most popular tags—whereas it is much harder to beat the resource specific most popular tags. The surprising result is that the graph based recommendations of FolkRank have superior recall—independent of the number of regarded tags. The second best results come from the collaborative filtering approach based on user tag similiarities. For ten recommended tags it reaches 89% of the recall of FolkRank (0.71 of 0.80)—a significant difference. The idea to suggest the top most popular tags of the resource gives a recall which is very similiar to using the CF recommender based on users resource similiarities—both perform worse than the aforementioned approaches. Between most popular tags by resource and most popular tags is the adapted PageRank which is influenced by the most popular tags, as discussed earlier.

The precision-recall plot in Figure 2 again reveals clearly the quality of the recommendations given by FolkRank compared to the other approaches. The top 10 tags given by FolkRank contained in average 80 % of the tags the users decided to attach to the selected resource. Nevertheless, the precision is rather poor with values below 0.2. So why do we call this a good result anyhow?

A post in del.icio.us contains only 2.45 tags in average. A precision of 100 % can therefore not be reached when recommending ten tags. However, from a subjective point of view, the additional 'wrong' tags may even be considered as highly relevant, as the following example shows, where the user *tnash* has tagged the page http://www.ariadne.ac.uk/issue43/chudnov/ with the tags *semantic, web,* and *webdesign*. Since that page discusses the interaction of publication reference management systems in the web by the OpenURL standard, the tags recommended by FolkRank (*openurl, web, webdesign, libraries, search, semantic, metadata, social-software, sfx, seo*) are adequate and capture not only the user's point of view that this is a webdesign related issue in the semantic web, but also provide him with more specific tags like *libraries* or *metadata* which the users nevertheless did not use. The CF based on user/tag similiarities recommends very similiar tags (*openurl, libraries, social-software, sfx, metadata, me/toread, software, myndsi, work, 2read*). The additional tags may thus animate users to use more tags and/or tags from a different viewpoint for describing resources, and thus lead to converging vocabularies.

The essential point in this example is, however, that FolkRank is able to predict—additionally to globally relevant tags—the exact tags of the user which CF could not. This is due to the fact that FolkRank considers, via the hypergraph structure, also the vocabulary of the user himself, which CF by definition doesn't do.

Last.fm. For this dataset, recall and precision for FolkRank are considerably higher than for the del.icio.us dataset, see Table 3. Even when just two tags are recommended, the recall is close to 60 %. Though the precision of the user-resource collaborative filtering approach is always slightly better than on the del.icio.us dataset, the recall is only better until the 7th tag where it falls below the recall reached on the del.icio.us dataset. 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. Section 4).

BibSonomy. For the BibSonomy dataset the precision for FolkRank is similiar to the Last.fm dataset (see Table 3), but the recall (omitted here because of space restrictions)

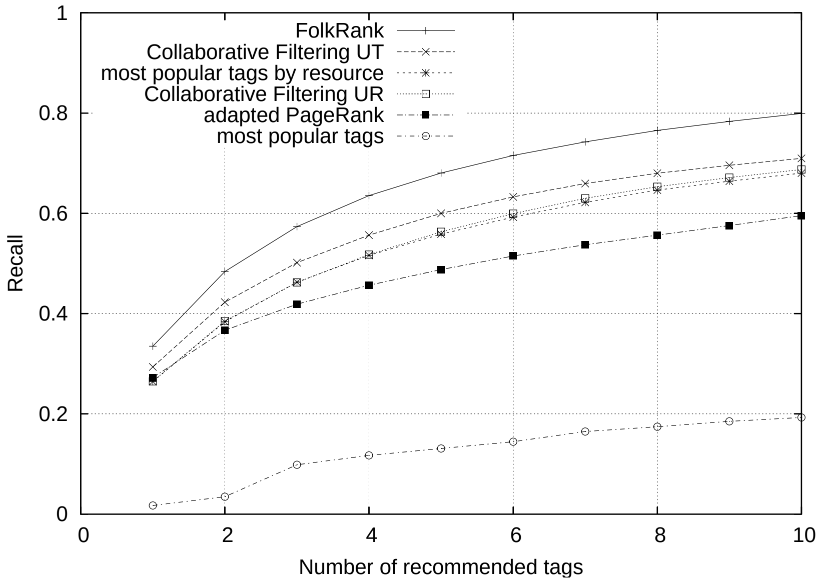

Figure 1: Recall for del.icio.us p-core at level 10

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

reaches only values comparable to the del.icio.us dataset. We will focus here on a phenomenon which is unique for that 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 tags, for the other approaches there are often not enough tags available for recommendation. 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).

7 Conclusion

In this paper we presented two methods for tag recommendations in folksonomies, a straightforward collaborative filtering adaptation based on projections and an adaptation of the well-known PageRank algorithm named FolkRank. We conducted experiments in three real-life datasets and showed that FolkRank outperforms the other methods. Some conclusions of our experiment were:

- The exploitation of the hypergraph structure in FolkRank yields a significant advantage.
- Despite its simplicity and non-personalized aspect, the 'most popular tags' achieved reasonable precision and recall on the small datasets (Last.fm and BibSonomy) what indicates its adequacy for the cold start problem.
- The adapted PageRank profits also from this good performance of the 'most popular tags' on small datasets.

Currently, our approach for FolkRank always returns a fixed number of tags, often yielding low precision. Future work will include a method to determine a good cut-off point automatically.

Acknowledgement. Part of this research was funded by the European Commission in the Nepomuk[16] (FP6-027705) and the X-Media[17] (IST-FP6-026978) projects.

References

[Batagelj and Zaversnik, 2002] V. Batagelj and M. Zaversnik. Generalized cores, 2002. cs.DS/0202039, http://arxiv.org/abs/cs/0202039.

[Benz *et al.*, 2006] D. Benz, K. Tso, and L. Schmidt-Thieme. Automatic bookmark classification: A collaborative approach. In *Proceedings of the Second Workshop on Innovations in Web Infrastructure (IWI 2006)*, Edinburgh, Scotland, 2006.

[Brin and Page, 1998] Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, April 1998.

[Cattuto *et al.*, 2006] Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Collaborative tagging and semiotic dynamics, May 2006. http://arxiv.org/abs/cs/0605015.

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

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

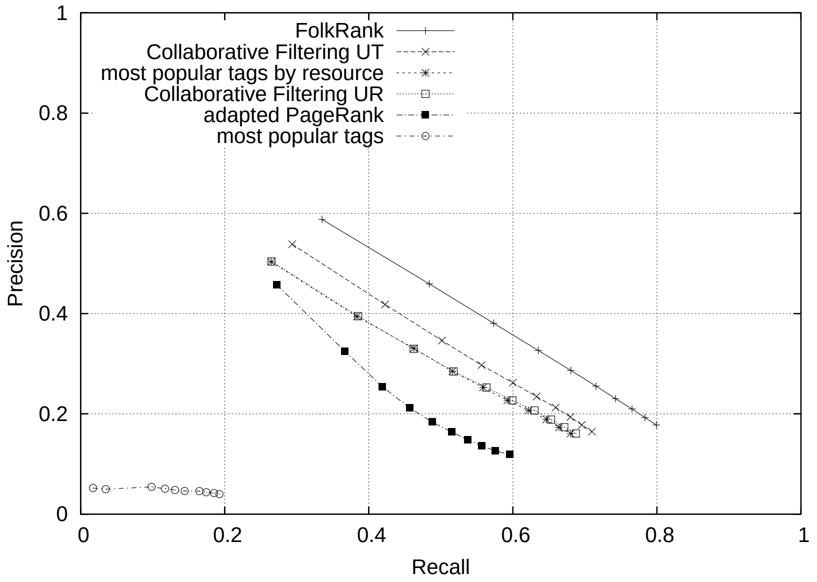

Figure 2: Recall and Precision for del.icio.us p-core at level 10

[Deshpande and Karypis, 2004] Mukund Deshpande and George Karypis. Item-based top-n recommendation algorithms. *ACM Trans. Inf. Syst.*, 22(1):143–177, 2004.

[Dubinko *et al.*, 2006] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *Proc. of the 15th International WWW Conference*, Edinburgh, Scotland, 2006.

[Halpin *et al.*, 2006] H. Halpin, V. Robu, and H. Shepard. The dynamics and semantics of collaborative tagging. In *Proceedings of the 1st Semantic Authoring and Annotation Workshop (SAAW'06)*, 2006.

[Hammond *et al.*, 2005] Tony Hammond, Timo Hannay, Ben Lund, and Joanna Scott. Social Bookmarking Tools (I): A General Review. *D-Lib Magazine*, 11(4), April 2005.

[Herlocker *et al.*, 2004] 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.

[Hotho *et al.*, 2006a] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In York Sure and John Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *Lecture Notes in Computer Science*, pages 411–426, Heidelberg, June 2006. Springer.

[Hotho *et al.*, 2006b] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Trend detection in folksonomies. In Yannis S. Avrithis, Yiannis Kompatsiaris, Steffen Staab, and Noel E. O'Connor, editors, *Proc. First International Conference on Semantics And Digital Media Technology (SAMT)*, volume 4306 of *LNCS*, pages 56–70, Heidelberg, Dec 2006. Springer.

[Kleinberg, 1999] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[Lund *et al.*, 2005] Ben Lund, Tony Hammond, Martin Flack, and Timo Hannay. Social Bookmarking Tools (II): A Case Study - Connotea. *D-Lib Magazine*, 11(4), April 2005.

[Mathes, 2004] Adam Mathes. Folksonomies – Cooperative Classification and Communication Through Shared Metadata, December 2004. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.

[Mika, 2005] Peter Mika. Ontologies Are Us: A Unified Model of Social Networks and Semantics. In Yolanda Gil, Enrico Motta, V. Richard Benjamins, and Mark A. Musen, editors, *ISWC 2005*, volume 3729 of *LNCS*, pages 522–536, Berlin Heidelberg, November 2005. Springer-Verlag.

[Mishne, 2006] Gilad Mishne. Autotag: a collaborative approach to automated tag assignment for weblog posts. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 953–954, New York, NY, USA, 2006. ACM Press.

[Sarwar *et al.*, 2001] Badrul M. Sarwar, George Karypis, Joseph A. Konstan, and John Reidl. Item-based collaborative filtering recommendation algorithms. In *World Wide Web*, pages 285–295, 2001.

[Schmitz *et al.*, 2006] Christoph Schmitz, Andreas Hotho, Robert Jäschke, and Gerd Stumme. Mining association rules in folksonomies. In V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Žiberna, editors, *Data Science and Classification: Proc. of the 10th IFCS Conf.*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 261–270, Berlin, Heidelberg, 2006. Springer.

[Xi *et al.*, 2004] W. Xi, B. Zhang, Y. Lu, Z. Chen, S. Yan, H. Zeng, W. Ma, and E. Fox. Link fusion: A unified link analysis framework for multi-type interrelated data objects. In *Proc. 13th International World Wide Web Conference*, New York, 2004.

[Xu *et al.*, 2006] Zhichen Xu, Yun Fu, Jianchang Mao, and Difu Su. Towards the semantic web: Collaborative tag suggestions. In *Proceedings of the Collaborative*

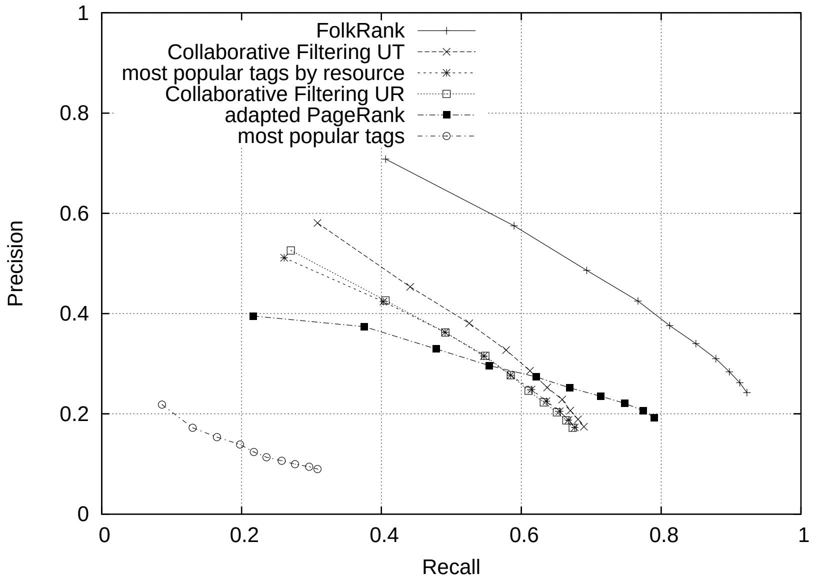

Figure 3: Recall and Precision for Last.fm p-core at level 10

Web Tagging Workshop at the WWW 2006, Edinburgh, Scotland, 2006.