

# Information Retrieval in Folksonomies: Search and Ranking

Andreas Hotho,<sup>1</sup> Robert Jäschke,<sup>1,2</sup> Christoph Schmitz,<sup>1</sup> Gerd Stumme<sup>1,2</sup>

<sup>1</sup> Knowledge & Data Engineering Group, Department of Mathematics and Computer Science,  
University of Kassel, Wilhelmshöher Allee 73, D-34121 Kassel, Germany

<http://www.kde.cs.uni-kassel.de>

<sup>2</sup> Research Center L3S, Expo Plaza 1, D-30539 Hannover, Germany

<http://www.l3s.de>

**Abstract.** Social bookmark tools are rapidly emerging on the Web. In such systems users are setting up lightweight conceptual structures called folksonomies. The reason for their immediate success is the fact that no specific skills are needed for participating. At the moment, however, the information retrieval support is limited. We present a formal model and a new search algorithm for folksonomies, called *FolkRank*, that exploits the structure of the folksonomy. The proposed algorithm is also applied to find communities within the folksonomy and is used to structure search results. All findings are demonstrated on a large scale dataset.

## 1 Introduction

Complementing the Semantic Web effort, a new breed of so-called “Web 2.0” applications is currently emerging on the Web. These include user-centric publishing and knowledge management platforms like Wikis, Blogs, and social resource sharing tools.

These tools, such as Flickr<sup>3</sup> or del.icio.us,<sup>4</sup> have acquired large numbers of users within less than two years.<sup>5</sup> The reason for their immediate success is the fact that no specific skills are needed for participating, and that these tools yield immediate benefit for each individual user (e.g. organizing ones bookmarks in a browser-independent, persistent fashion) without too much overhead. Large numbers of users have created huge amounts of information within a very short period of time. The frequent use of these systems shows clearly that web- and folksonomy-based approaches are able to overcome the knowledge acquisition bottleneck, which was a serious handicap for many knowledge-based systems in the past.

Social resource sharing systems all use the same kind of lightweight knowledge representation, called *folksonomy*. The word ‘folksonomy’ is a blend of the words ‘taxonomy’ and ‘folk’, and stands for conceptual structures created by the people. Folksonomies are thus a bottom-up complement to more formalized Semantic Web technologies, as they rely on *emergent semantics* [11, 12] which result from the converging use of the same vocabulary. The main difference to ‘classical’ ontology engineering approaches is their aim to respect to the largest possible extent the request of non-expert users not

<sup>3</sup> <http://www.flickr.com/> <sup>4</sup> <http://del.icio.us> <sup>5</sup> From discussions on the del.icio.us mailing list, one can approximate the number of users on del.icio.us to be more than three hundred thousand.

to be bothered with any formal modeling overhead. Intelligent techniques may well be inside the system, but should be hidden from the user.

A first step to searching folksonomy based systems – complementing the browsing interface usually provided as of today – is to employ standard techniques used in information retrieval or, more recently, in web search engines. Since users are used to web search engines, they likely will accept a similar interface for search in folksonomy-based systems. The research question is how to provide suitable ranking mechanisms, similar to those based on the web graph structure, but now exploiting the structure of folksonomies instead. To this end, we propose a formal model for folksonomies, and present a new algorithm, called *FolkRank*, that takes into account the folksonomy structure for ranking search requests in folksonomy based systems. The algorithm will be used for two purposes: determining an overall ranking, and specific topic-related rankings.

This paper is organized as follows. Section 2 reviews recent developments in the area of social bookmark systems, and presents a formal model. Section 3 recalls the basics of the PageRank algorithm, describes our adaptation to folksonomies, and discusses experimental results. These results indicate the need for a more sophisticated algorithm for topic-specific search. Such an algorithm, FolkRank, is presented in Section 4. This section includes also an empirical evaluation, as well as a discussion of its use for generating personal recommendations in folksonomies. Section 5 concludes the paper with a discussion of further research topics on the intersection between folksonomies and ontologies.

## 2 Social Resource Sharing and Folksonomies

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<sup>6</sup> and Connotea<sup>7</sup> the sharing of bibliographic references, and 43Things<sup>8</sup> even the sharing of goals in private life. Our own system, *BibSonomy*,<sup>9</sup> allows to share simultaneously bookmarks and bibtex entries (see Fig. 1).

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 had uploaded, together with the tags he had assigned to them (see Fig. 1); 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.

The systems allow for additional functionality. For instance, one can copy a resource from another user, and label it with one's own tags. Overall, these systems provide a very intuitive navigation through the data. However, the resources that are displayed are usually ordered by date, i. e., the resources entered last show up at the top. A more sophisticated notion of 'relevance' – which could be used for ranking – is still missing.

<sup>6</sup> <http://www.citeulike.org/>

<sup>7</sup> <http://www.connotea.org/>

<sup>8</sup> <http://www.43things.com/>

<sup>9</sup> <http://www.bibsonomy.org>

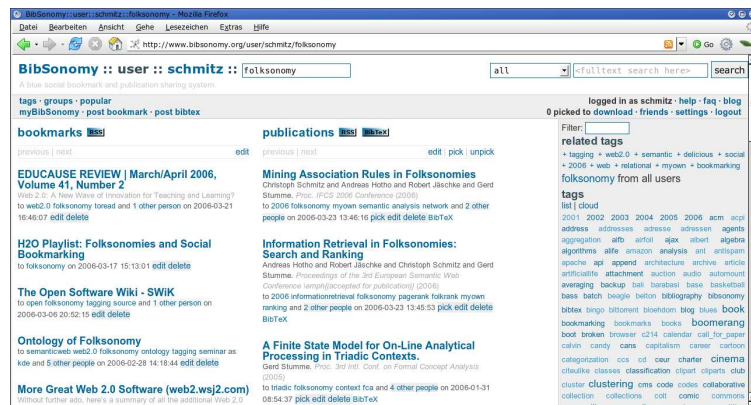


Fig. 1. Bibsonomy displays bookmarks and BibTeX based bibliographic references simultaneously.

## 2.1 State of the Art

There are currently virtually no scientific publications about folksonomy-based web collaboration systems. The main discussion on folksonomies and related topics is currently taking place on mailing lists only, e.g. [3]. Among the rare exceptions are [5] and [8] who provide good overviews of social bookmarking tools with special emphasis on folksonomies, and [9] who discusses strengths and limitations of folksonomies. In [10], Mika defines a model of semantic-social networks for extracting lightweight ontologies from del.icio.us. Besides calculating measures like the clustering coefficient, (local) betweenness centrality or the network constraint on the extracted one-mode network, Mika uses co-occurrence techniques for clustering the folksonomy.

There are several systems working on top of del.icio.us to explore the underlying folksonomy. CollaborativeRank<sup>10</sup> provides ranked search results on top of del.icio.us bookmarks. The ranking takes into account how early someone bookmarked an URL and how many people followed him or her. Other systems show popular sites (Populicious<sup>11</sup>) or focus on graphical representations (Cloudalicious<sup>12</sup>, Grafolicious<sup>13</sup>) of statistics about del.icio.us.

Confoto,<sup>14</sup> the winner of the 2005 Semantic Web Challenge, is a service to annotate and browse conference photos and offers besides rich semantics also tagging facilities for annotation. Due to the representation of this rich metadata in RDF it has limitations in both size and performance.

Ranking techniques have also been applied in traditional ontology engineering. The tool Ontocopi [1] performs what is called Ontology Network Analysis for initially populating an organizational memory. Several network analysis methods are applied to an already populated ontology to extract important objects. In particular, a PageRank-like [2] algorithm is used to find communities of practice within sets of individuals represented in the ontology. The algorithm used in Ontocopi to find nodes related to an individual removes the respective individual from the graph and measures the difference of the resulting Perron eigenvectors of the adjacency matrices as the influence of that individual.

<sup>10</sup> <http://collabrank.org/>

<sup>11</sup> <http://populicious.us/>

<sup>12</sup> <http://cloudalicio.us/>

<sup>13</sup> <http://www.neuroticweb.com/recursos/del.icio.us-graphs/>

<sup>14</sup> <http://www.confoto.org/>

This approach differs insofar from our proposed method, as it tracks which nodes benefit from the removal of the individual, instead of actually preferring the individual and measuring which related nodes are more influenced than others.

## 2.2 A Formal Model for Folksonomies

A folksonomy describes the users, resources, and tags, and the user-based assignment of tags to resources. We present here a formal definition of folksonomies, which is also underlying our BibSonomy system.

**Definition 1.** A folksonomy is a tuple  $\mathbb{F} := (U, T, R, Y, \prec)$  where

- $U$ ,  $T$ , and  $R$  are finite sets, whose elements are called users, tags and resources, resp.,
- $Y$  is a ternary relation between them, i. e.,  $Y \subseteq U \times T \times R$ , called tag assignments (TAS for short), and
- $\prec$  is a user-specific subtag/supertag-relation, i. e.,  $\prec \subseteq U \times T \times T$ , called subtag/supertag relation.

The personomy  $\mathbb{P}_u$  of a given user  $u \in U$  is the restriction of  $\mathbb{F}$  to  $u$ , i. e.,  $\mathbb{P}_u := (T_u, R_u, I_u, \prec_u)$  with  $I_u := \{(t, r) \in T \times R \mid (u, t, r) \in Y\}$ ,  $T_u := \pi_1(I_u)$ ,  $R_u := \pi_2(I_u)$ , and  $\prec_u := \{(t_1, t_2) \in T \times T \mid (u, t_1, t_2) \in \prec\}$ , where  $\pi_i$  denotes the projection on the  $i$ th dimension.

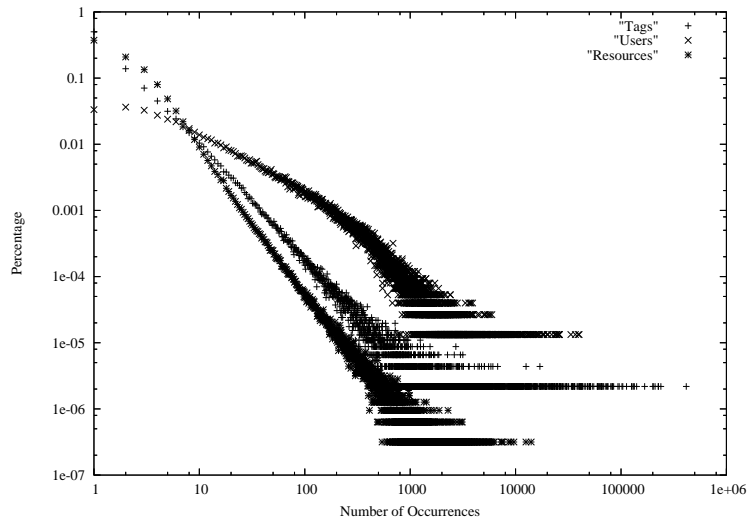
Users are typically described by their user ID, and tags may be arbitrary strings. What is considered as a resource depends on the type of system. For instance, in del.icio.us, the resources are URLs, and in flickr, the resources are pictures. From an implementation point of view, resources are internally represented by some ID.

In this paper, we do not make use of the subtag/supertag relation for sake of simplicity. I. e.,  $\prec = \emptyset$ , and we will simply note a folksonomy as a quadruple  $\mathbb{F} := (U, T, R, Y)$ . This structure is known in Formal Concept Analysis [14, 4] as a *triadic context* [7, 13]. An equivalent view on folksonomy data is that of 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.

## 2.3 Del.icio.us — A Folksonomy-Based Social Bookmark System

In order to evaluate our retrieval technique detailed in the next section, we have analyzed the popular social bookmarking system del.icio.us, which is a server-based system with a simple-to-use interface that allows users to organize and share bookmarks on the internet. It is able to store in addition to the URL a description, an extended description, and tags (i. e., arbitrary labels). We chose del.icio.us rather than our own system, BibSonomy, as the latter went online only after the time of writing of this article.

For our experiments, we collected data from the del.icio.us system in the following way. Initially we used `wget` starting from the top page of del.icio.us to obtain nearly 6900 users and 700 tags as a starting set. Out of this dataset we extracted all users and resources (i. e., del.icio.us' MD5-hashed urls). From July 27 to 30, 2005, we downloaded in a recursive manner user pages to get new resources, and resource pages to get new



**Fig. 2.** Number of TAS occurrences for tags, users, resources in del.icio.us

users. Furthermore we monitored the del.icio.us start page to gather additional users and resources. This way we collected a list of several thousand usernames which we used for accessing the first 10000 resources each user had tagged. From the collected data we finally took the user files to extract resources, tags, dates, descriptions, extended descriptions, and the corresponding username.

We obtained a core folksonomy with  $|U| = 75,242$  users,  $|T| = 533,191$  tags and  $|R| = 3,158,297$  resources, related by in total  $|Y| = 17,362,212$  TAS.<sup>15</sup> After inserting this dataset into a MySQL database, we were able to perform our evaluations, as described in the following sections.

As expected, the tagging behavior in del.icio.us shows a power law distribution, see Figure 2. This figure presents the percentage of tags, users, and resources, respectively, which occur in a given number of TAS. For instance, the rightmost ‘+’ indicates that a fraction of  $2.19 \cdot 10^{-6}$  of all tags (i. e. one tag) occurs 415950 times – in this case it is the empty tag. The next ‘+’ shows that one tag (“web”) occurs 238891 times, and so on. One observes that while the tags follow a power law distribution very strictly, the plot for users and resources levels off for small numbers of occurrences. Based on this observation, we estimate to have crawled most of the tags, while many users and resources are still missing from the dataset. A probable reason is that many users only try posting a single resource, often without entering any tags (the empty tag is the most frequent one in the dataset), before they decide not to use the system anymore. These users and resources are very unlikely to be connected with others at all (and they only appear for a short period on the del.icio.us start page), so that they are not included in our crawl.

<sup>15</sup> 4,313 users additionally organised 113,562 of the tags with 6,527 so-called *bundles*. The bundles will not be discussed in this paper; they can be interpreted as one level of the  $\prec$  relation.

### 3 Ranking in Folksonomies using Adapted PageRank

Current folksonomy tools such as del.icio.us provide only very limited search support in addition to their browsing interface. Searching can be performed over the text of tags and resource descriptions, but no ranking is done apart from ordering the hits in reverse chronological order. Using traditional information retrieval, folksonomy contents can be searched textually. However, as the documents consist of short text snippets only (usually a description, e. g. the web page title, and the tags themselves), ordinary ranking schemes such as TF/IDF are not feasible.

As shown in Section 2.2, a folksonomy induces a graph structure which we will exploit for ranking in this section. Our *FolkRank* algorithm is inspired by the seminal PageRank algorithm [2]. The PageRank weight-spreading approach cannot be applied directly on folksonomies because of the different nature of folksonomies compared to the web graph (undirected triadic hyperedges instead of directed binary edges). In the following we discuss how to overcome this problem.

#### 3.1 Adaptation of PageRank

We implement the weight-spreading ranking scheme on folksonomies in two steps. First, we transform the hypergraph between the sets of users, tags, and resources into an undirected, weighted, tripartite graph. On this graph, we apply a version of PageRank that takes into account the edge weights.

**Converting the Folksonomy into an Undirected Graph.** First we convert the folksonomy  $\mathbb{F} = (U, T, R, Y)$  into an *undirected* tripartite graph  $\mathbb{G}_{\mathbb{F}} = (V, E)$  as follows.

1. The set  $V$  of nodes of the graph consists of the disjoint union of the sets of tags, users and resources:  $V = U \dot{\cup} T \dot{\cup} R$ . (The tripartite structure of the graph can be exploited later for an efficient storage of the – sparse – adjacency matrix and the implementation of the weight-spreading iteration in the FolkRank algorithm.)
2. All co-occurrences of tags and users, users and resources, tags and resources become undirected, weighted edges between the respective nodes:  $E = \{\{u, t\}, \{t, r\}, \{u, r\} \mid (u, t, r) \in Y\}$ , with each edge  $\{u, t\}$  being weighted with  $|\{r \in R : (u, t, r) \in Y\}|$ , each edge  $\{t, r\}$  with  $|\{u \in U : (u, t, r) \in Y\}|$ , and each edge  $\{u, r\}$  with  $|\{t \in T : (u, t, r) \in Y\}|$ .

**Folksonomy-Adapted Pagerank.** The original formulation of PageRank [2] reflects the idea that a page is important if there many pages linking to it, and if those pages are important themselves. The distribution of weights can thus be described as the fixed point of a weight passing scheme on the web graph. This idea was extended in a similar fashion to bipartite subgraphs of the web in HITS [6] and to n-ary directed graphs in [15]). We employ the same underlying principle for our ranking scheme in folksonomies. The basic notion 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.

Like PageRank, we employ the random surfer model, a notion of importance for web pages that is based on the idea that an idealized random web surfer normally follows hyperlinks, but from time to time randomly jumps to a new webpage without following a link. This results in the following definition of the rank of the vertices of the graph the entries in the fixed point  $\vec{w}$  of the weight spreading computation  $\vec{w} \leftarrow dA\vec{w} + (1-d)\vec{p}$ , where  $\vec{w}$  is a weight vector with one entry for each web page,  $A$  is the row-stochastic<sup>16</sup> version of the adjacency matrix of the graph  $G_{\mathbb{F}}$  defined above,  $\vec{p}$  is the random surfer component, and  $d \in [0, 1]$  is determining the influence of  $\vec{p}$ . In the original PageRank,  $\vec{p}$  is used to outweigh the loss of weight on web pages without outgoing links. Usually, one will choose  $\vec{p} = \mathbf{1}$ , i. e., the vector composed by 1's. In order to compute personalized PageRanks, however,  $\vec{p}$  can be used to express user preferences by giving a higher weight to the components which represent the user's preferred web pages.

We employ a similar motivation for our ranking scheme in folksonomies. The basic notion 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 tripartite graph in which the vertices are mutually reinforcing each other by spreading their weights. Formally, we spread the weight as follows:

$$\vec{w} \leftarrow \alpha\vec{w} + \beta A\vec{w} + \gamma\vec{p} \quad (1)$$

where  $A$  is the row-stochastic version of the adjacency matrix of  $G_{\mathbb{F}}$ ,  $\vec{p}$  is a preference vector,  $\alpha, \beta, \gamma \in [0, 1]$  are constants with  $\alpha + \beta + \gamma = 1$ . The constant  $\alpha$  is intended to regulate the speed of convergence, while the proportion between  $\beta$  and  $\gamma$  controls the influence of the preference vector.

We call the iteration according to Equation 1 – until convergence is achieved – the *Adapted PageRank* algorithm. Note that, if  $\|\vec{w}\|_1 = \|\vec{p}\|_1$  holds,<sup>17</sup> the sum of the weights in the system will remain constant. The influence of different settings of the parameters  $\alpha$ ,  $\beta$ , and  $\gamma$  is discussed below.

As the graph  $G_{\mathbb{F}}$  is undirected, part of the weight that went through an edge at moment  $t$  will flow back at  $t + 1$ . The results are thus rather similar (but not identical) to a ranking that is simply based on edge degrees, as we will see now. The reason for applying the more expensive PageRank approach nonetheless is that its random surfer vector allows for topic-specific ranking, as we will discuss in the next section.

### 3.2 Results for Adapted PageRank

We have evaluated the Adapted PageRank on the del.icio.us dataset described in Section 2.3. As there exists no 'gold standard ranking' on these data, we evaluate our results empirically.

First, we studied the speed of convergence. We let  $\vec{p} := \mathbf{1}$  (the vector having 1 in all components), and varied the parameter settings. In all settings, we discovered that  $\alpha \neq 0$  slows down the convergence rate. For instance, for  $\alpha = 0.35, \beta = 0.65, \gamma = 0$ , 411 iterations were needed, while  $\alpha = 0, \beta = 1, \gamma = 0$  returned the same result in only 320 iterations. It turns out that using  $\gamma$  as a damping factor by spreading equal weight

<sup>16</sup> I. e., each row of the matrix is normalized to 1 in the 1-norm. <sup>17</sup> ... and if there are no rank sinks – but this holds trivially in our graph  $G_{\mathbb{F}}$ .

**Table 1.** Folksonomy Adapted PageRank applied without preferences (called *baseline*)

Tag	ad. PageRank	User	ad. PageRank
system:unfiled	0,0078404	shankar	0,0007389
web	0,0044031	notmuch	0,0007379
blog	0,0042003	fritz	0,0006796
design	0,0041828	ubi.quito.us	0,0006171
software	0,0038904	weev	0,0005044
music	0,0037273	kof2002	0,0004885
programming	0,0037100	ukquake	0,0004844
css	0,0030766	gearhead	0,0004820
reference	0,0026019	angusf	0,0004797
linux	0,0024779	johncollins	0,0004668
tools	0,0024147	mshook	0,0004556
news	0,0023611	frizzlebiscuit	0,0004543
art	0,0023358	rafapol	0,0004535
blogs	0,0021035	xiombarg	0,0004520
politics	0,0019371	tidesonar02	0,0004355
java	0,0018757	cyrusnews	0,0003829
javascript	0,0017610	bidurling	0,0003727
mac	0,0017252	onpause_tv_anytime	0,0003600
games	0,0015801	cataracte	0,0003462
photography	0,0015469	triple_entendre	0,0003419
fun	0,0015296	kayodeok	0,0003407

URL	ad. PageRank
http://slashdot.org/	0,0002613
http://pchere.blogspot.com/2005/02/absolutely-delicious-complete-tool.html	0,0002320
http://script.aculo.us/	0,0001770
http://www.adaptivepath.com/publications/essays/archives/000385.php	0,0001654
http://johnvey.com/features/deliciousdirector/	0,0001593
http://en.wikipedia.org/wiki/Main_Page	0,0001407
http://www.flickr.com/	0,0001376
http://www.goodfonts.org/	0,0001349
http://www.43folders.com/	0,0001160
http://www.csszengarden.com/	0,0001149
http://wellstyled.com/tools/colorscheme2/index-en.html	0,0001108
http://pro.html.it/esempio/nifty/	0,0001070
http://www.alistapart.com/	0,0001059
http://postsecret.blogspot.com/	0,0001058
http://www.beelerspace.com/index.php?p=890	0,0001035
http://www.techsupportalert.com/best.46.free.utilities.htm	0,0001034
http://www.alvit.de/web-dev/	0,0001020
http://www.technorati.com/	0,0001015
http://www.lifehacker.com/	0,0001009
http://www.lucazappa.com/brilliantMaker/buttonImage.php	0,0000992
http://www.engadget.com/	0,0000984

to each node in each iteration speeds up the convergence considerably by a factory of approximately 10 (e. g., 39 iterations for  $\alpha = 0, \beta = 0.85, \gamma = 0.15$ ).

Table 1 shows the result of the adapted PageRank algorithm for the 20 most important tags, users and resources computed with the parameters  $\alpha = 0.35, \beta = 0.65, \gamma = 0$  (which equals the result for  $\alpha = 0, \beta = 1, \gamma = 0$ ). Tags get the highest ranks, followed by the users, and the resources. Therefore, we present their rankings in separate lists.

As we can see from the tag table, the most important tag is “system:unfiled” which is used to indicate that a user did not assign any tag to a resource. It is followed by “web”, “blog”, “design” etc. This corresponds more or less to the rank of the tags given by the overall tag count in the dataset. The reason is that the graph  $G_{\mathbb{F}}$  is undirected. We face thus the problem that, in the Adapted PageRank algorithm, weights that flow in one direction of an edge will basically ‘swash back’ along the same edge in the next

iteration. Therefore the resulting is very similar (although not equal!) to a ranking based on counting edge degrees.

The resource ranking shows that Web 2.0 web sites like Slashdot, Wikipedia, Flickr, and a del.icio.us related blog appear in top positions. This is not surprising, as early users of del.icio.us are likely to be interested in Web 2.0 in general. This ranking correlates also strongly with a ranking based on edge counts.

The results for the top users are of more interest as different kinds of users appear. As all top users have more than 6000 bookmarks; “notmuch” has a large amount of tags, while the tag count of “fritz” is considerably smaller.

To see how good the topic-specific ranking by Adapted PageRank works, we combined it with term frequency, a standard information retrieval weighting scheme. To this end, we downloaded all 3 million web pages referred to by a URL in our dataset. From these, we considered all plain text and html web pages, which left 2.834.801 documents. We converted all web pages into ASCII and computed an inverted index. To search for a term as in a search engine, we retrieved all pages containing the search term and ranked them by  $tf(t) \cdot \vec{w}[v]$  where  $tf(t)$  is the term frequency of search term  $t$  in page  $v$ , and  $\vec{w}[v]$  is the Adapted PageRank weight of  $v$ .

Although this is a rather straightforward combination of two successful retrieval techniques, our experiments with different topic-specific queries indicate that this adaptation of PageRank does not work very well. For instance, for the search term “football”, the del.icio.us homepage showed up as the first result. Indeed, most of the highly ranked pages have nothing to do with football.

Other search terms provided similar results. Apparently, the overall structure of the – undirected – graph overrules the influence of the preference vector. In the next section, we discuss how to overcome this problem.

## 4 FolkRank – Topic-Specific Ranking in Folksonomies

In order to reasonably focus the ranking around the topics defined in the preference vector, we have developed a differential approach, which compares the resulting rankings with and without preference vector. This resulted in our new *FolkRank* algorithm.

### 4.1 The FolkRank Algorithm

The FolkRank algorithm computes a topic-specific ranking in a folksonomy as follows:

1. The preference vector  $\vec{p}$  is used to determine the topic. It may have any distribution of weights, as long as  $\|\vec{w}\|_1 = \|\vec{p}\|_1$  holds. Typically a single entry or a small set of entries is set to a high value, and the remaining weight is equally distributed over the other entries. Since the structure of folksonomies is symmetric, we can define a topic by assigning a high value to either one or more tags and/or one or more users and/or one or more resources.
2. Let  $\vec{w}_0$  be the fixed point from Equation (1) with  $\beta = 1$ .
3. Let  $\vec{w}_1$  be the fixed point from Equation (1) with  $\beta < 1$ .
4.  $\vec{w} := \vec{w}_1 - \vec{w}_0$  is the final weight vector.

Thus, we compute the winners and losers of the mutual reinforcement of resources when a user preference is given, compared to the baseline without a preference vector. We call the resulting weight  $\bar{w}[x]$  of an element  $x$  of the folksonomy the *FolkRank* of  $x$ .

Whereas the Adapted PageRank provides one global ranking, independent of any preferences, FolkRank provides one topic-specific ranking for each given preference vector. Note that a topic can be defined in the preference vector not only by assigning higher weights to specific tags, but also to specific resources and users. These three dimensions can even be combined in a mixed vector. Similarly, the ranking is not restricted to resources, it may as well be applied to tags and to users. We will show below that indeed the rankings on all three dimensions provide interesting insights.

## 4.2 Comparing FolkRank with Adapted PageRank

To analyse the proposed FolkRank algorithm, we generated rankings for several topics, and compared them with the ones obtained from Adapted PageRank. We will here discuss two sets of search results, one for the tag “boomerang”, and one for the URL <http://www.semanticweb.org>. Our other experiments all provided similar results.

The leftmost part of Table 2 contains the ranked list of tags according to their weights from the Adapted PageRank by using the parameters  $\alpha = 0.2, \beta = 0.5, \gamma = 0.3$ , and 5 as a weight for the tag “boomerang” in the preference vector  $\vec{p}$ , while the other elements were given a weight of 0. As expected, the tag “boomerang” holds the first position while tags like “shop” or “wood” which are related are also under the Top 20. The tags “software”, “java”, “programming” or “web”, however, are on positions 4 to 7, but have nothing to do with “boomerang”. The only reason for their showing up is that they are frequently used in del.icio.us (cf. Table 1). The second column from the left in Table 2 contains the results of our FolkRank algorithm, again for the tag “boomerang”. Intuitively, this ranking is better, as the globally frequent words disappear and related words like “wood” and “construction” are ranked higher.

A closer look reveals that this ranking still contains some unexpected tags; “kassel” or “rdf” are for instance not obviously related to “boomerang”. An analysis of the user ranking (not displayed) explains this fact. The top-ranked user is “schm4704”, and he has indeed many bookmarks about boomerangs. A FolkRank run with preference weight 5 for user “schm4704” shows his different interests, see the rightmost column in Table 2. His main interest apparently is in boomerangs, but other topics show up as well. In particular, he has a strong relationship to the tags “kassel” and “rdf”. When a community in del.icio.us is small (such as the boomerang community), already a single user can thus provide a strong bridge to other communities, a phenomenon that is equally observed in small social communities.

A comparison of the FolkRank ranking for user “schm4704” with the Adapted PageRank result for him (2nd ranking from left) confirms the initial finding from above, that the Adapted PageRank ranking contains many globally frequent tags, while the FolkRank ranking provides more personal tags. While the differential nature of the FolkRank algorithm usually pushes down the globally frequent tags such as “web”, though, this happens in a differentiated manner: FolkRank will keep them in the top positions, *if* they are indeed relevant to the user under consideration. This can be seen for example for the tags “web” and “java”. While the tag “web” appears in schm4704’s tag list – but not very of-

**Table 2.** Ranking results for the tag “boomerang” (two left at top: Adapted PageRank and FolkRank for tags, middle: FolkRank for URLs) and for the user “schm4704” (two right at top: Adapted PageRank and FolkRank for tags, bottom: FolkRank for URLs)

Tag	ad. PRank	Tag	FolkRank	Tag	ad. PRank	Tag	FolkRank
boomerang	0,4036883	boomerang	0,4036867	boomerang	0,0093549	boomerang	0,0093533
shop	0,0069058	shop	0,0066477	lang:ade	0,0068111	lang:de	0,0068028
lang:de	0,0050943	lang:de	0,0050860	shop	0,0052600	shop	0,0050019
software	0,0016797	wood	0,0012236	java	0,0052050	java	0,0033293
java	0,0016389	kassel	0,0011964	web	0,0049360	kassel	0,0032223
programming	0,0016296	construction	0,0010828	programming	0,0037894	network	0,0028990
web	0,0016043	plans	0,0010085	software	0,0035000	rdf	0,0028758
reference	0,0014713	injuries	0,0008078	network	0,0032882	wood	0,0028447
system:unfiled	0,0014199	pitching	0,0007982	kassel	0,0032228	delicious	0,0026345
wood	0,0012378	rdf	0,0006619	reference	0,0030699	semantic	0,0024736
kassel	0,0011969	semantic	0,0006533	rdf	0,0030645	database	0,0023571
linux	0,0011442	material	0,0006279	delicious	0,0030492	guitar	0,0018619
construction	0,0011023	trifly	0,0005691	system:unfiled	0,0029393	computing	0,0018404
plans	0,0010226	network	0,0005568	linux	0,0029393	cinema	0,0017537
network	0,0009460	webring	0,0005552	wood	0,0028589	lessons	0,0017273
rdf	0,0008506	sna	0,0005073	database	0,0026931	social	0,0016950
css	0,0008266	socialnetworkanalysis	0,0004822	semantic	0,0025460	documentation	0,0016182
design	0,0008248	cinema	0,0004726	css	0,0024577	scientific	0,0014686
delicious	0,0008097	erie	0,0004525	social	0,0021969	filesystem	0,0014212
injuries	0,0008087	riparian	0,0004467	webdesign	0,0020650	userspace	0,0013490
pitching	0,0007999	erosion	0,0004425	computing	0,0020143	library	0,0012398

Url	FolkRank
http://www.flight-toys.com/boomerangs.htm	0,0047322
http://www.flight-toys.com/	0,0047322
http://www.bumerangclub.de/	0,0045785
http://www.bumerangfibel.de/	0,0045781
http://www.kutek.net/trifly_mods.php	0,0032643
http://www.rediboom.de/	0,0032126
http://www.bws-buhmann.de/	0,0032126
http://www.akspiele.de/	0,0031813
http://www.medco-athletics.com/education/elbow_shoulder_injuries/	0,0031606
http://www.sportsprolo.com/sports%20prolotherapy%20newsletter%20pitching%20injuries.htm	0,0031606
http://www.boomerangpassion.com/english.php	0,0031005
http://www.kuhara.de/bumerangschule/	0,0030935
http://www.bumerangs.de/	0,0030935
http://s.webring.com/hub?ring=boomerang	0,0030895
http://www.kutek.net/boomplans/plans.php	0,0030873
http://www.geocities.com/cmorris32839/jonas_article/	0,0030871
http://www.theboomerangman.com/	0,0030868
http://www.boomerangs.com/index.html	0,0030867
http://www.lmifox.com/us/boom/index-uk.htm	0,0030867
http://www.sports-boomerangs.com/	0,0030867
http://www.rangsboomerangs.com/	0,0030867

Url	FolkRank
http://jena.sourceforge.net/	0,0019369
http://www.openrdf.org/doc/users/ch06.html	0,0017312
http://dsd.lbl.gov/hoschek/colt/api/overview-summary.html	0,0016777
http://librdf.org/	0,0014402
http://www.hpl.hp.com/semweb/jena2.htm	0,0014326
http://jakarta.apache.org/commons/collections/	0,0014203
http://www.aktors.org/technologies/ontocopi/	0,0012839
http://eventseer.idi.ntnu.no/	0,0012734
http://tangra.si.umich.edu/radev/	0,0012685
http://www.cs.umass.edu/mccallum/	0,0012091
http://www.w3.org/TR/rdf-sparql-query/	0,0011945
http://ourworld.compuserve.com/homepages/gaeme_birchall/HTM_COOK.HTM	0,0011930
http://www.emory.edu/EDUCATION/mfp/Kuhn.html	0,0011880
http://www.hpl.hp.com/semweb/rdql.htm	0,0011860
http://jena.sourceforge.net/javadoc/index.html	0,0011860
http://www.geocities.com/mailsoftware42/db/	0,0011838
http://www.quirksmode.org/	0,0011327
http://www.kde.cs.uni-kassel.de/lehre/ss2005/googlespam	0,0011110
http://www.powerpage.org/cgi-bin/WebObjects/powerpage.woa/wa/story?newsID=14732	0,0010402
http://www.vaughns-1-pagers.com/internet/google-ranking-factors.htm	0,0010329
http://www.cl.cam.ac.uk/Research/SRG/netos/xen/	0,0010326

ten, “java” is a very important tag for that user. This is reflected in the FolkRank ranking: “java” remains in the Top 5, while “web” is pushed down in the ranking.

The ranking of the resources for the tag “boomerang” given in the middle of Table 2 also provides interesting insights. As shown in the table, many boomerang related web pages show up (their topical relatedness was confirmed by a boomerang aficionado). Comparing the Top 20 web pages of “boomerang” with the Top 20 pages given by the “schm4704” ranking, there is no “boomerang” web page in the latter. This can be explained by analysing the tag distribution of this user. While “boomerang” is the most frequent tag for this user, in del.icio.us, “boomerang” appears rather infrequently. The first boomerang web page in the “schm4704” ranking is the 21st URL (i. e., just outside the listed TOP 20). Thus, while the tag “boomerang” itself dominates the tags of this user, in the whole, the semantic web related tags and resources prevail. This demonstrates that while the user “schm4704” and the tag “boomerang” are strongly correlated, we can still get an overview of the respective related items which shows several topics of interest for the user.

Let us consider a second example. Table 3 gives the results for the web page <http://www.semanticweb.org/>. The two tables on the left show the tags and users for the adapted PageRank, resp., and the two ones on the right the FolkRank results. Again, we see that the differential ranking of FolkRank makes the right decisions: in the Adaptive PageRank, globally frequent tags such as “web”, “css”, “xml”, “programming” get high ranks. Of these, only two turn up to be of genuine interest to the members of the Semantic Web community: “web” and “xml” remain at high positions, while “css” and “programming” disappear altogether from the list of the 20 highest ranked tags. Also, several variations of tags which are used to label Semantic Web related pages appear (or get ranked higher): “semantic web” (two tags, space-separated), “semantic\_web”, “semweb”, “sem-web”. These co-occurrences of similar tags could be exploited further to consolidate the emergent semantics of a field of interest. While the discovery in this case may also be done in a simple syntactic analysis, the graph based approach allows also for detecting inter-community and inter-language relations.

The user IDs can not be checked for topical relatedness immediately, since they are not related to the users’ full names – although a former winner of the Semantic Web Challenge and the best paper award at a Semantic Web Conference seems to be among them. The web pages that appear in the top list, on the other hand, include many well-known resources from the Semantic Web area. An interesting resource on the list is PiggyBank, which has been presented in November 2005 at the ISWC conference. Considering that the dataset was crawled in July 2005, when PiggyBank was not that well known, the prominent position of PiggyBank in del.icio.us at such an early time is an interesting result. This indicates the sensibility of social bookmarking systems for upcoming topics.

These two examples – as well as the other experiments we performed – show that FolkRank provides good results when querying the folksonomy for topically related elements. Overall, our experiments indicate that topically related items can be retrieved with FolkRank for any given set of highlighted tags, users and/or resources.

Our results also show that the current size of folksonomies is still prone to being skewed by a relatively small number of perturbations – a single user, at the moment, can influence the emergent understanding of a certain topic in the case that a sufficient number of different points of view for such a topic has not been collected yet. With the

**Table 3.** Ranking for the resource <http://www.semanticweb.org> (Left two tables: Adapted PageRank for tags and users; right two tables: FolkRank for tags and users. Bottom: FolkRank for resources).

Tag	ad. PRank	User	ad. PageRank	Tag	FolkRank	User	FolkRank
semanticweb	0,0208605	up4	0,0091995	semanticweb	0,0207820	up4	0,0091828
web	0,0162033	awenger	0,0086261	semantic	0,0121305	awenger	0,0084958
semantic	0,0122028	j.deville	0,0074021	web	0,0118002	j.deville	0,0073525
system:unfiled	0,0088625	chaizzilla	0,0062570	semantic_web	0,0071933	chaizzilla	0,0062227
semantic_web	0,0072150	elektron	0,0059457	rdf	0,0044461	elektron	0,0059403
rdf	0,0046348	captsolo	0,0055671	semweb	0,0039308	captsolo	0,0055369
semweb	0,0039897	stevag	0,0049923	resources	0,0034209	dissipative	0,0049619
resources	0,0037884	dissipative	0,0049647	community	0,0033208	stevag	0,0049590
community	0,0037256	krudd	0,0047574	portal	0,0022745	krudd	0,0047005
xml	0,0031494	williamteo	0,0037204	xml	0,0022074	williamteo	0,0037181
research	0,0026720	stevecassidy	0,0035887	research	0,0020378	stevecassidy	0,0035840
programming	0,0025717	pmika	0,0035359	imported-bo...	0,0018920	pmika	0,0035358
css	0,0025290	millette	0,0033028	en	0,0018536	millette	0,0032103
portal	0,0024118	myren	0,0028117	.idate2005-04-11	0,0017555	myren	0,0027965
.imported	0,0020495	morningboat	0,0025913	newfurl	0,0017153	morningboat	0,0025875
imported-bo...	0,0019610	philip.fennell	0,0025338	tosort	0,0014486	philip.fennell	0,0025145
en	0,0018900	mote	0,0025212	cs	0,0014002	webb.	0,0024671
science	0,0018166	dnaboy76	0,0024813	academe	0,0013822	dnaboy76	0,0024659
.idate2005-04-11	0,0017779	webb.	0,0024709	rfid	0,0013456	mote	0,0024214
newfurl	0,0017578	nymetbarton	0,0023790	sem-web	0,0013316	alphajuliet	0,0023668
internet	0,0016122	alphajuliet	0,0023781	w3c	0,0012994	nymetbarton	0,0023666

URL	FolkRank
<a href="http://www.semanticweb.org/">http://www.semanticweb.org/</a>	0,3761957
<a href="http://fink.semanticweb.org/">http://fink.semanticweb.org/</a>	0,0005566
<a href="http://simile.mit.edu/piggy-bank/">http://simile.mit.edu/piggy-bank/</a>	0,0003828
<a href="http://www.w3.org/2001/sw/">http://www.w3.org/2001/sw/</a>	0,0003216
<a href="http://infomesh.net/2001/swintro/">http://infomesh.net/2001/swintro/</a>	0,0002162
<a href="http://del.icio.us/register">http://del.icio.us/register</a>	0,0001745
<a href="http://mspace.ecs.soton.ac.uk/">http://mspace.ecs.soton.ac.uk/</a>	0,0001712
<a href="http://www.adaptivepath.com/publications/essays/archives/000385.php">http://www.adaptivepath.com/publications/essays/archives/000385.php</a>	0,0001637
<a href="http://www.ontoweb.org/">http://www.ontoweb.org/</a>	0,0001617
<a href="http://www.aaai.org/AITopics/html/ontol.html">http://www.aaai.org/AITopics/html/ontol.html</a>	0,0001613
<a href="http://simile.mit.edu/">http://simile.mit.edu/</a>	0,0001395
<a href="http://itip.evcc.jp/itipwiki/">http://itip.evcc.jp/itipwiki/</a>	0,0001256
<a href="http://www.google.be/">http://www.google.be/</a>	0,0001224
<a href="http://www.letterjames.de/index.html">http://www.letterjames.de/index.html</a>	0,0001224
<a href="http://www.daml.org/">http://www.daml.org/</a>	0,0001216
<a href="http://shirky.com/writings/ontology_outrated.html">http://shirky.com/writings/ontology_outrated.html</a>	0,0001195
<a href="http://jena.sourceforge.net/">http://jena.sourceforge.net/</a>	0,0001167
<a href="http://www.alistapart.com/">http://www.alistapart.com/</a>	0,0001102
<a href="http://www.federalconciierge.com/WritingBusinessCases.html">http://www.federalconciierge.com/WritingBusinessCases.html</a>	0,0001060
<a href="http://pchere.blogspot.com/2005/02/absolutely-delicious-complete-tool.html">http://pchere.blogspot.com/2005/02/absolutely-delicious-complete-tool.html</a>	0,0001059
<a href="http://www.shirky.com/writings/semantic_syllogism.html">http://www.shirky.com/writings/semantic_syllogism.html</a>	0,0001052

growth of folksonomy-based data collections on the web, the influence of single users will fade in favor of a common understanding provided by huge numbers of users.

As detailed above, our ranking is based on tags only, without regarding any inherent features of the resources at hand. This allows to apply FolkRank to search for pictures (e. g., in flickr) and other multimedia content, as well as for all other items that are difficult to search in a content-based fashion. The same holds for intranet applications, where in spite of centralized knowledge management efforts, documents often remain unused because they are not hyperlinked and difficult to find. Full text retrieval may be used to find documents, but traditional IR methods for ranking without hyperlink information have difficulties finding the most relevant documents from large corpora.

### 4.3 Generating Recommendations

The original PageRank paper [2] already pointed out the possibility of using the random surfer vector  $\vec{p}$  as a personalization mechanism for PageRank computations. The results of Section 4 show that, given a user, one can find set of tags and resources of interest to him. Likewise, FolkRank yields a set of related users and resources for a given tag. Following these observations, FolkRank can be used to generate recommendations within a folksonomy system. These recommendations can be presented to the user at different points in the usage of a folksonomy system:

- Documents that are of potential interest to a user can be suggested to him. This kind of recommendation pushes potentially useful content to the user and increases the chance that a user finds useful resources that he did not even know existed by “serendipitous” browsing.
- When using a certain tag, other related tags can be suggested. This can be used, for instance, to speed up the consolidation of different terminologies and thus facilitate the emergence of a common vocabulary.
- While folksonomy tools already use simple techniques for tag recommendations, FolkRank additionally considers the tagging behavior of other users.
- Other users that work on related topics can be made explicit, improving thus the knowledge transfer within organizations and fostering the formation of communities.

## 5 Conclusion and Outlook

In this paper, we have argued that enhanced search facilities are vital for emergent semantics within folksonomy-based systems. We presented a formal model for folksonomies, the *FolkRank* ranking algorithm that takes into account the structure of folksonomies, and evaluation results on a large-scale dataset.

The FolkRank ranking scheme has been used in this paper to generate personalized rankings of the items in a folksonomy, and to recommend users, tags and resources. We have seen that the top folksonomy elements which are retrieved by FolkRank tend to fall into a coherent topic area, e.g. “Semantic Web”. This leads naturally to the idea of extracting *communities of interest* from the folksonomy, which are represented by their top tags and the most influential persons and resources. If these communities are made explicit, interested users can find them and participate, and community members can more easily get to know each other and learn of others’ resources.

Another future research issue is to combine different search and ranking paradigms. In this paper, we went a first step by focusing on the new structure of folksonomies. In the future, we will incorporate additionally the full text that is contained in the web pages addressed by the URLs, the link structure of these web pages, and the usage behavior as stored in the log file of the tagging system. The next version will also exploit the tag hierarchy.

Currently, spam is not a serious problem for social bookmarking systems. With the increasing attention they currently receive, however, we anticipate that ‘spam posts’ will show up sooner or later. As for mail spam and link farms in the web, solutions will be needed to filter out spam. We expect that a blend of graph structure analysis together with content analysis will give the best results.

When folksonomy-based systems grow larger, user support has to go beyond enhanced retrieval facilities. Therefore, the internal structure has to become better organized. An obvious approach for this are semantic web technologies. The key question remains though how to exploit its benefits without bothering untrained users with its rigidity. We believe that this will become a fruitful research area for the Semantic Web community for the next years.

*Acknowledgement.* Part of this research was funded by the EU in the Nepomuk project (FP6-027705).

## References

1. Harith Alani, Srinandan Dasmahapatra, Kieron O'Hara, and Nigel Shadbolt. Identifying Communities of Practice through Ontology Network Analysis. *IEEE Intelligent Systems*, 18(2):18–25, March/April 2003.
2. 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.
3. Connotea Mailing List. <https://lists.sourceforge.net/lists/listinfo/connotea-discuss>.
4. B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical foundations*. Springer, 1999.
5. Tony Hammond, Timo Hannay, Ben Lund, and Joanna Scott. Social Bookmarking Tools (I): A General Review. *D-Lib Magazine*, 11(4), April 2005.
6. Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
7. F. Lehmann and R. Wille. A triadic approach to formal concept analysis. In G. Ellis, R. Levinson, W. Rich, and J. F. Sowa, editors, *Conceptual Structures: Applications, Implementation and Theory*, volume 954 of *Lecture Notes in Computer Science*. Springer, 1995.
8. Ben Lund, Tony Hammond, Martin Flack, and Timo Hannay. Social Bookmarking Tools (II): A Case Study - Connotea. *D-Lib Magazine*, 11(4), April 2005.
9. Adam Mathes. Folksonomies – Cooperative Classification and Communication Through Shared Metadata, December 2004. <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>.
10. 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.
11. S. Staab, S. Santini, F. Nack, L. Steels, and A. Maedche. Emergent semantics. *Intelligent Systems, IEEE [see also IEEE Expert]*, 17(1):78–86, 2002.
12. L. Steels. The origins of ontologies and communication conventions in multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 1(2):169–194, October 1998.
13. Gerd Stumme. A finite state model for on-line analytical processing in triadic contexts. In Bernhard Ganter and Robert Godin, editors, *ICFCA*, volume 3403 of *Lecture Notes in Computer Science*, pages 315–328. Springer, 2005.
14. R. Wille. Restructuring lattice theory: An approach based on hierarchies of concepts. In I. Rival, editor, *Ordered Sets*, pages 445–470. Reidel, Dordrecht-Boston, 1982.
15. 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.