Information Retrieval in Folksonomies: Search and Ranking

Andreas Hotho, Robert Jäschke, Christoph Schmitz, Gerd Stumme

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

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, there exists no foundational research for these systems. 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[1] or del.icio.us,[2], have acquired large numbers of users (from discussions on the del.icio.us mailing list, one can approximate the number of users on del.icio.us to be more than one hundred thousand) within less than two years. 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

[1] http://www.flickr.com/

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

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 to be bothered with any formal modeling overhead. Intelligent techniques may well be inside the system, but should be hidden from the user and have to be designed to set up the needed semantic structure in the background.

A first step to searching these 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.

Hybrid approaches to ranking search results, augmenting content-based measures with rankings based on the hyperlink structure of the documents, are successfully employed by the major web search engines today, providing search results which are to a large extent influenced by people’s opinions of web pages (expressed by the tendency to put links to pages one likes).

Applying these ranking techniques in intranets, however, is more difficult. Corporate intranets will consist of large collections of documents, which typically do not link to each other and are often stored in formats such as PDF or MS Office not having the idea of hypertext in mind. The hyperlink structure of intranets is often purely navigational and does not express any kind of recommendation or semantic links between contents, but will rather be engineered from scratch by a knowledge engineer or even the person who is in charge of the technical infrastructure of the intranet. This lead to two motivating observations: (a) folksonomies can augment the rigid structure of corporate knowledge management, adding individual statements about resources which can be used for ranking search results, and (b) from this additional structure, recommendations for intranet users can be extracted.

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 have developed a new algorithm, called *FolkRank,* that takes into account the folksonomy structure for ranking search requests in internet and intranet based folksonomy systems.

This paper is organized as follows.

Section 2 reviews recent developments in the are of social bookmark systems, and presents a formal model. Section 3 introduces and evaluates the FolkRank algorithm for ranking search results and generating personal recommendations in folksonomies. Section 4.1 concludes the paper with a discussion of further research topics on the intersection between folksonomies and ontologies.

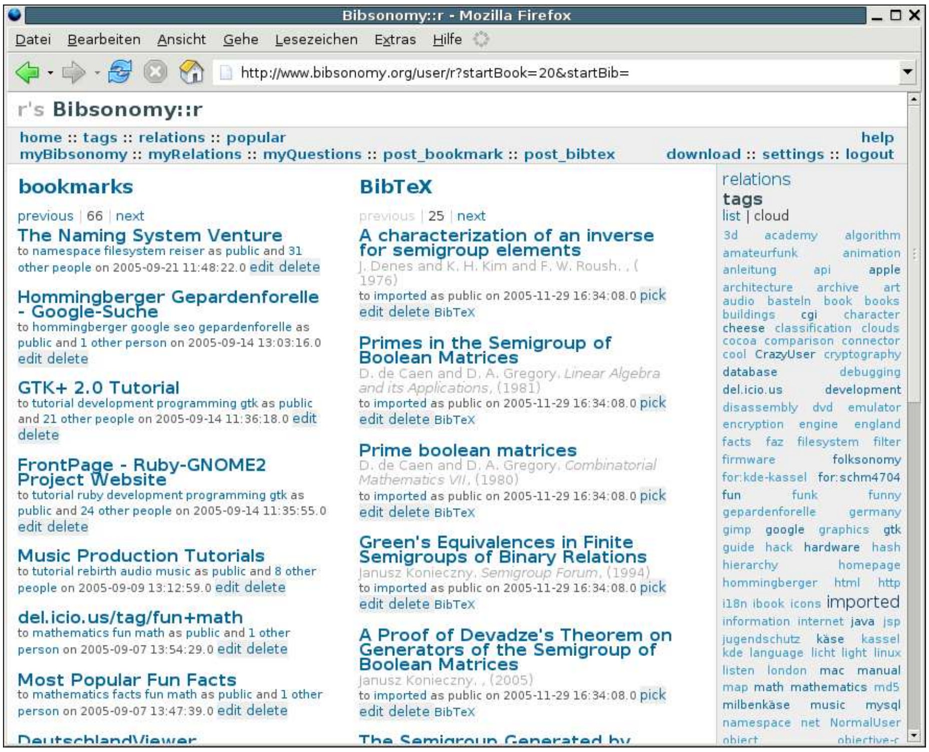
2 Social Resource Sharing and Folksonomies

Social resource sharing systems are web-based systems that allow users to upload their resources, and to label them with names. 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[3] and Connotea[4] the sharing of bibli-

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

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

ographic references, and 43Things [5] even the sharing of goals in private life. Our own upcoming system, called *BibSonomy*,[6] will allow to share simultaneously bookmarks and bibtex entries (see Fig. 1).


Bibsonomy::r - Mozilla Firefox
Datei Bearbeiten Ansicht Gehe Lesezeichen Extras Hilfe
http://www.bibsonomy.org/user/r?startBook=20&startBib=
r's Bibsonomy::r
home :: tags :: relations :: popular
help
myBibsonomy :: myRelations :: myQuestions :: post_bookmark :: post_bibtex
download :: settings :: logout

bookmarks

previous | 66 | next

The Naming System Venture
to namespace filesystem reiser as public and 31 other people on 2005-09-21 11:48:22.0 edit delete

Hommingberger Gepardenforelle - Google-Suche
to hommingberger google seo gepardenforelle as public and 1 other person on 2005-09-14 13:03:16.0 edit delete

GTK+ 2.0 Tutorial
to tutorial development programming gtk as public and 21 other people on 2005-09-14 11:36:18.0 edit delete

FrontPage - Ruby-GNOME2 Project Website
to tutorial ruby development programming gtk as public and 24 other people on 2005-09-14 11:35:55.0 edit delete

Music Production Tutorials
to tutorial rebirth audio music as public and 8 other people on 2005-09-09 13:12:59.0 edit delete

del.icio.us/tag/fun+math
to mathematics fun math as public and 1 other person on 2005-09-07 13:54:29.0 edit delete

Most Popular Fun Facts
to mathematics facts fun math as public and 1 other person on 2005-09-07 13:47:39.0 edit delete

DeutschlandViewer

BibTeX

previous | 25 | next

A characterization of an inverse for semigroup elements
J. Denes and K. H. Kim and F. W. Roush. , (1976)
to imported as public on 2005-11-29 16:34:08.0 pick edit delete BibTeX

Primes in the Semigroup of Boolean Matrices
D. de Caen and D. A. Gregory. *Linear Algebra and its Applications*, (1981)
to imported as public on 2005-11-29 16:34:08.0 pick edit delete BibTeX

Prime boolean matrices
D. de Caen and D. A. Gregory. *Combinatorial Mathematics VII*, (1980)
to imported as public on 2005-11-29 16:34:08.0 pick edit delete BibTeX

Green's Equivalences in Finite Semigroups of Binary Relations
Janusz Konieczny. *Semigroup Forum*, (1994)
to imported as public on 2005-11-29 16:34:08.0 pick edit delete BibTeX

A Proof of Devadze's Theorem on Generators of the Semigroup of Boolean Matrices
Janusz Konieczny. , (2005)
to imported as public on 2005-11-29 16:34:08.0 pick edit delete BibTeX

The Semigroup Generated by

relations
tags
list | cloud

3d academy algorithm amateurfunk animation anleitung api apple architecture archive art audio basteln book books buildings cgi character cheese classification clouds cocoa comparison connector cool CrazyUser cryptography database debugging del.icio.us development disassembly dvd emulator encryption engine england facts faz filesystem filter firmware folksonomy for:kde-kassel for:schm4704 fun funk funny gepardenforelle germany gimp google graphics gtk guide hack hardware hash hierarchy homepage hommingberger html http i18n ibook icons imported information internet java jsp jugendschutz käse kassel kde language licht light linux listen london mac manual map math mathematics md5 milbenkäse music mysql namespace net NormalUser object objective-c

Fig. 1. Bibsonomy displays bookmarks and BibTeX based bibliographic references 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 labels, so-called *tags*, to it. The collection of all his assignments is called his *personomy*, the collection of all personomies is called *folksonomy*. The user can also explore the folksonomies of the other users in all dimensions: for a given user he can see the resources that user had uploaded, together with the tags he had assigned to them (see Fig. 1); when clicking on a resource he sees which other users have uploaded this resource and how they tagged it; and when clicking on a tag he 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 ones own tags. Overall, these systems provide a

[5] http://www.43things.com/

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

very intuitive navigation through the data. However, the resources that are displayed are usually ordered by date, i. e., the lastly entered resources show up at the top. A more sophisticated notion of ‘relevance’ – which could be used for ranking – is still missing.

2.1 State of the Art

There are currently virtually no scientific publications about folksonomy-based web collaboration systems. 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. The main discussion on folksonomies and related topics is currently only going on mailing lists, e.g. [3]. To the best of our knowledge, the ideas presented in this paper have not been explored before, but there is a lot of recent work dealing with 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-occurence techniques for clustering the concept network.

There are several systems working on top of del.icio.us to explore the underlying folksonomy. CollaborativeRank[7] 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[8]) or focus on graphical representations (Grafolicious[9], Cloudalicious[10]) of statistics about del.icio.us.

Confoto[11], 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.

The tool Ontocopi described in [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 individuals represented in the ontology. The algorithm used there to find related nodes of an individual removes the respective individual from the graph and measures the difference of the resulting Perron eigenvectors of the matrices as the influence of that individual. This approach differs insofar from our proposed method, as it tracks which nodes benefit from the removal of the invidial, 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 basically describes the users, the resources, tags, and allows users to assign (arbitrary) tags to resources. We present here a formal definition of folksonomies, which is also underlying our BibSonomy system.

[7] http://collabrank.org/

[8] http://populicio.us/

[9] http://www.neuroticweb.com/recursos/del.icio.us-graphs/

[10] http://cloudalicio.us/

[11] http://www.confoto.org/

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* $\prec_u := \{(t_1, t_2) \in T \times T \mid (u, t_1, t_2) \in \prec\}$.

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.ico.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 sytem del.icio.us [12]. Del.icio.us 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 is going online only after the time of writing of this article.

For our experiments, we collected data from the del.ico.us system in the following way. Initially we used `wget` starting from the top page of del.ico.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 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.[13] After

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

[13] 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 one level of the $\prec$ relation.

inserting this dataset into a MySQL database, we were able to perform our evaluations, as described in the subsequent chapters.

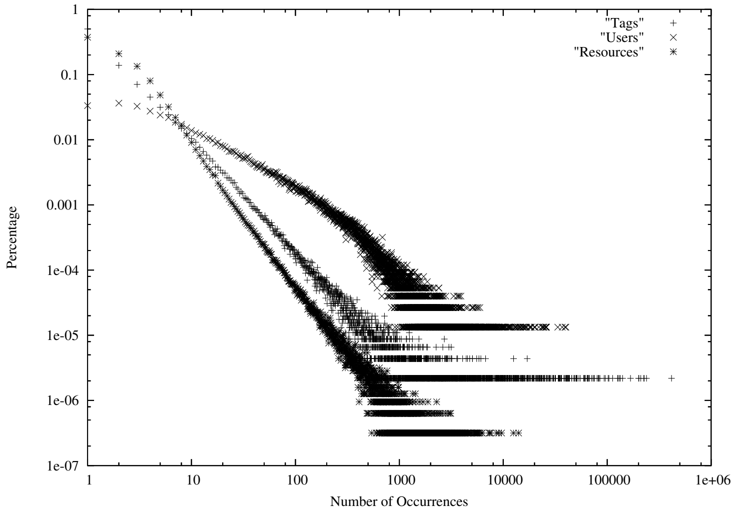


Fig. 2. Number of TAS occurrences for tags, users, resources in del.icio.us

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.

We see 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. This can be explained by the fact that many users only ever try posting one resource, often leaving out the 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 linked to at all (they only appear for a short period on the del.icio.us start page), so that they are not included in our crawl.

3 Searching in Folksonomies

Current folksonomy tools such as del.icio.us provide only very limited searching 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.

3.1 Folkrank: Ranking of Search Results

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, using an algorithm we call *FolkRank* which is inspired by the seminal PageRank algorithm [2]. 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 difference 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.

Converting the Folksonomy into an Undirected Graph. First we convert the folksonomy $\mathbb{F} = (U, T, R, Y)$ into an *un*directed tri-partite 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 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 edges between the respective nodes:

$$\begin{aligned} E = \{\{u,t\} \mid \exists r \in R : (u,t,r) \in Y\} &\cup \\ \{\{t,r\} \mid \exists u \in U : (u,t,r) \in Y\} &\cup \\ \{\{u,r\} \mid \exists t \in T : (u,t,r) \in Y\}& \end{aligned}$$

Folksonomy-Adapted Pagerank. The original formulation of PageRank [2] employed the random surfer model, a notion of importance for web pages that is based on the idea that an idealized random web surfer which follows hyperlinks ends up on any given page with a certain probability.

This probability is reflected by a component in the fixed point R of the weight spreading computation $R \leftarrow c(A\,R + P)$, where R is a weight vector with one entry for each web page, A is a row-stochastic version of the adjacency matrix, P is a damping vector to take care of dangling links[14], and c is a normalization constant. Usually, one will choose $P = \alpha \cdot \mathbf{1}$ to achieve uniform damping. In order to compute personalized PageRanks, however, P can be used to express user preferences by giving a higher weight to the components which represent the user's preferred web pages.

These ideas were extended in a similar fashion to bipartite subgraphs of the web in HITS [6] and to n-ary directed graphs (Link Fusion, [15]).

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 graph of vertices which are mutually reinforcing each other by spreading their weights.

[14] In the original paper, the vector is called E, but that would collide with the edge set.

Formally, we spread the weight as follows:

$$R \leftarrow c(\alpha R + \beta AR + \gamma P) \tag{1}$$

where A is the row-stochastic version of the adjacency matrix of $\mathbb{G}_{\mathbb{F}}$, P is a preference vector, α, β, γ are constants and c is a normalization factor such that $\|R\|_1 = 1$. A damping factor α is used to avoid oscillation and speed up convergence, while β and γ control the influence of the preference vector.

FolkRank. As the graph $\mathbb{G}_{\mathbb{F}}$ 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.

Furthermore, the structure of our evaluation dataset favors some nodes to an extent which makes it very difficult for other nodes to become ranked high, no matter what the preference vector is.

This problem is solved by our *differential* approach, which computes a personalized ranking of the elements in a folksonomy as follows:

- A preference vector P reflecting the user's preferences or search goals is given by the user, extracted from a query, or determined from his behavior
- Let R_{AP} be the fixed point from Equation (1) with $\gamma = 0$.
- Let R_{pref} be the fixed point from Equation (1) with $\gamma > 0$.
- $R := R_{\text{pref}} - R_{AP}$ is the final weight vector.

Thus, we compute the winners and losers of the mutual reinforcement of resources when user preferences are given, compared to the baseline without a preference vector. We call the resulting weight $R(v)$ of an element v of the folksonomy the *FolkRank* of v.

3.2 Results for Adapted PageRank and FolkRank

In this section we will present the results for the different ranking methods. As described in section 3.1, we use the folksonomy adapted PageRank and the FolkRank algorithms to rank search results. There are different ways to do this: the first idea is to use the adapted PageRank to compute weights for all resources to rank retrieved web sites in a "classical" sense (cf. [2]). The second way is to compute a ranking with the adapted PageRank and compare it to our FolkRank by using the introduced preference vector to search for items (which in this case are not restricted to resources).

Ranking of Web Sites by Adapted PageRank Table 1 shows the result of the adapted PageRank algorithm for the 20 most important tags, users and resources computed with the parameter $\alpha = 0.35, \beta = 0.65, \gamma = 0$ on our del.icio.us dataset (cf. sec. 2.3). As we can see from this table the most important tag is the tag "system:unfiled" which is implicitly added to resources without any user-supplied tags, 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 results for the top users are of more interest as different kinds of user 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. Popular web sites like Slashdot, Wikipedia, Flickr, or a del.icio.us related blog appearing in top positions are not surprising. As one can see from the weights of the FolkRank, tags get the highest ranks, followed by the users, and the resources.

To see how good the ranking by adapted PageRank works, we downloaded all 3 million web pages referred to in our dataset. After that, we restrict ourselves to plain text and html web pages, which left 2.834.801 documents. We converted all web sites into ASCII and computed an inverted index. To search for a term as in a search engine, we retrieved all pages which contain the search term once and ranked the retrieved web sites by $tf \cdot R_{AP}$ where tf is the term frequency of the search term in the document and R_{AP} is the folksonomy-adapted PageRank.

Overall this method does not work very well; for the search term “football” we got the del.icio.us web site as the first result. The inspection of the next pages does not change this result and most of the pages have nothing to do with football.

In this approach we used the information of the del.icio.us users only indirectly by ranking web sites which were retrieved using a full-text search. In the next section, we will try to use the information provided by the tagged bookmarks more directly to find web sites for a given search term.

Comparing FolkRank with Adapted PageRank To analyze the proposed FolkRank algorithm, we present search results for the tag “boomerang”. The leftmost table in 2 contains the ranked list of tags and the weights for the adapted PageRank by using the parameter $\alpha = 0.2, \beta = 0.5, \gamma = 0.3$ and 5 as a weight for the tag “boomerang”. As expected the tag “boomerang” holds the first position while tags like “shop” or “wood” which are related are also under the top twenty. Tags like “software”, “java”, “programming” or “web” have position 4 to 7, but have nothing to do with “boomerang”. These tags are frequently used in del.icio.us (cf. table 1) and thus show up in the ranking of “boomerang”, which seems counterintuitive.

The second table from the left in Table 2 contains the results of our FolkRank for the tag “boomerang”. Intuitively, the ranking is better as the globally frequent words disappear and related words are higher ranked, like “wood” or “construction”. Nevertheless this ranking contains also unexpected tags; “kassel” or “rdf” are not obviously related tags. The analysis of the top ranked users shows the reason for this ranking. User “schm4704” is the top ranked user which has indeed a lot of bookmarks about boomerangs. This is also the reason why a ranking with weight 5 for user “schm4704” leads to the results of the two rightmost tables in 2 for adapted PageRank and FolkRank, resp. As shown in the table, the tag “boomerang” has the top position in both rankings. Comparing both rankings gives us the a similar impression as before. The adapted PageRank ranking contains a lot of the frequent tags while more personal tag are not top ranked. Coming back to the analysis of the unexpected tags in the result for “boomerang”, the reason for this result is the strong influence of that user for this tag, and the fact that this user has many resources tagged with “kassel” and “rdf”.

Tag	ad. PageRank
system:unfiled	0,0078404
web	0,0044031
blog	0,0042003
design	0,0041828
software	0,0038904
music	0,0037273
programming	0,0037100
css	0,0030766
reference	0,0026019
linux	0,0024779
tools	0,0024147
news	0,0023611
art	0,0023358
blogs	0,0021035
politics	0,0019371
java	0,0018757
javascript	0,0017610
mac	0,0017252
games	0,0015801
photography	0,0015469
fun	0,0015296

User	ad. PageRank
shankar	0,0007389
notmuch	0,0007379
fritz	0,0006796
ubi.quito.us	0,0006171
weev	0,0005044
kof2002	0,0004885
ukquake	0,0004844
gearhead	0,0004820
angusf	0,0004797
johncollins	0,0004668
mshook	0,0004556
frizzlebiscuit	0,0004543
rafaspol	0,0004535
xiombarg	0,0004520
tidesonar02	0,0004355
cyrusnews	0,0003829
bldurling	0,0003727
onpause_tv_anytime	0,0003600
cataracte	0,0003462
triple_entendre	0,0003419
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

Table 1. Folksonomy Adapted PageRank applied on 17M TAS without preferences (called *baseline*)

While the differential nature of the FolkRank algorithm usually pushes down the globally frequent tags such as “web” etc., this happens in a more differentiated manner here: the FolkRank is able to leave these 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 often – “java” is a very important tag for that user. This is reflected in the FolkRank as “java” remains in the top five, while “web” is pushed down in the ranking.

It is also interesting to regard the ranking of the resources for the tag “boomerang” given at middle of table 2. As shown in the table, a lot of boomerang related web sites show up (their topical relatedness was confirmed by a boomerang aficionado).

Comparing the top twenty web sites of “boomerang” with the top twenty sites given by the “schm4704” ranking, there is no “boomerang” web site in it.

This can be explained by analysing the tag distribution of this user. While “boomerang” is the most frequent tags for this user, in del.icio.us, “boomerang” appears rather seldomly. The first boomerang web site in “schm4704” ranking is the next URL after the 20 URLs shown.

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.

Consider another example to get an impression if the findings of the previous example still holds. Table 3 gives the results for the web site http://www.semanticweb.org/. The two tables on the left show the tags and users for the adapted PageRank, 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 are considered to be genuinely interesting by 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 highest ranked 20 tags.

Also, several variations of tags which are used to label Semantic Web related sites 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 (in this case by a simple syntactic analysis).

While the user names can not being checked for topical relatedness immediately (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 include many well-known resources from the Semantic Web area. An interesting resource on the list ist Piggy Bank, which has recently (at the time of this writing) been presented at the ISWC conference; considering that the dataset was crawled in July 2005, when Piggy Bank was not that well known, this is an interesting result.

Concluding we can see that FolkRank provides good results when querying the folksonomy for topically related elements. Our experiments – parts of the results of which we presented here – indicate that this kind of topically related items can be retrieved for many other kinds of queries as well.

On the other hand, the results also show that the current size of folksonomies on the web 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.

We expect that similar results could be obtained analysing other folksonomy tools. Furthermore, with the growth of folksonomies on the web, the influence of single users will fade in favor of a common understanding provided by huge numbers of users.

As shown above our ranking is based on tags only, without regarding any inherent features of the resources at hand. This allows us to apply FolkRank to find e.g. pictures

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/graeme_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

Table 2. Ranking results for searching for the tag “boomerang” (two left, ad. PageRank and FolkRank) and for the user “schm4704”(two right, ad. PageRank and FolkRank)

Tag	ad. PRank
semanticweb	0,0208605
web	0,0162033
semantic	0,0122028
system:unfiled	0,0088625
semantic_web	0,0072150
rdf	0,0046348
semweb	0,0039897
resources	0,0037884
community	0,0037256
xml	0,0031494
research	0,0026720
programming	0,0025717
css	0,0025290
portal	0,0024118
.imported	0,0020495
imported-bo...	0,0019610
en	0,0018900
science	0,0018166
.idate2005-04-11	0,0017779
newfurl	0,0017578
internet	0,0016122

User	ad. PageRank
up4	0,0091995
awenger	0,0086261
j.deville	0,0074021
chaizzilla	0,0062570
elektron	0,0059457
captsolo	0,0055671
stevag	0,0049923
dissipative	0,0049647
krudd	0,0047574
williamteo	0,0037204
stevecassidy	0,0035887
pmika	0,0035359
millette	0,0033028
myren	0,0028117
morningboat	0,0025913
philip.fennell	0,0025338
mote	0,0025212
dnaboy76	0,0024813
webb.	0,0024709
nymetbarton	0,0023790
alphajuliet	0,0023781

Tag	FolkRank
semanticweb	0,0207820
semantic	0,0121305
web	0,0118002
semantic_web	0,0071933
rdf	0,0044461
semweb	0,0039308
resources	0,0034209
community	0,0033208
portal	0,0022745
xml	0,0022074
research	0,0020378
imported-bo...	0,0018920
en	0,0018536
.idate2005-04-11	0,0017555
newfurl	0,0017153
tosort	0,0014486
cs	0,0014002
academe	0,0013822
rfid	0,0013456
sem-web	0,0013316
w3c	0,0012994

User	FolkRank
up4	0,0091828
awenger	0,0084958
j.deville	0,0073525
chaizzilla	0,0062227
elektron	0,0059403
captsolo	0,0055369
dissipative	0,0049619
stevag	0,0049590
krudd	0,0047005
williamteo	0,0037181
stevecassidy	0,0035840
pmika	0,0035358
millette	0,0032103
myren	0,0027965
morningboat	0,0025875
philip.fennell	0,0025145
webb.	0,0024671
dnaboy76	0,0024659
mote	0,0024214
alphajuliet	0,0023668
nymetbarton	0,0023666

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

Table 3. Ranking for http://www.semanticweb.org, left two table adapted PageRank and the two right tables FolkRank.

in flickr or other items which are difficult to search in a content-based fashion, by using only information provided by tags of the community. 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.

3.3 Generating Recommendations

The original PageRank paper [2] already pointed out the possibility of using the damping vector E as a personalization mechanism for PageRank computations.

The results of Section 3.2 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 during the usage of a folksonomy tool.

These recommendations can be presented to the user at different points in the usage of a folksonomy system:

- Users can be presented documents which will probably be relevant to them. 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 tags which are related can be suggested. This can be used, for example, to speed up the consolidation of different vocabulary and thus facilitate the emergence of a common vocabulary.
- While folksonomy tools already use simple approaches to present tag recommendations, using FolkRank, recommendations based on other users' tagging behavior can be generated which recognizing all influences around a user.
- Other users which work on related topics can be made explicit improving the knowledge transfer within organizations.

4 Conclusion and Outlook

4.1 Conclusion

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.

A 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 webpages addressed by the URLs, the link structure of these webpages, and the usage behavior as stored in the log file of the tagging system.

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.

4.2 Future Work

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.

In Section 3.2, 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.

Furthermore, there has been a lively discussion about the usefulness of the $\prec$ relation in the folksonomy, which is partially realized as bundles in del.icio.us. We will investigate the use of ontology learning techniques to populate this relation in our folksonomy tool and augment the underlying semantic structure in the folksonomy.

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 lattices 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 Internation World Wide Web Conference*, New York, 2004.