FolkRank: A Ranking Algorithm for Folksonomies

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

In social bookmark tools users are setting up lightweight conceptual structures called folksonomies. Currently, 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. A long version of this paper has been published at the European Semantic Web Conference 2006 [3].

1 Introduction

Social resource sharing tools, such as Flickr,[1] del.icio.us,[2], or our own system *BibSonomy*[3] (see Fig. 1) have acquired large numbers of users 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 are web-based systems that allow users to upload their resources (e. g., bookmarks, publications, photos; depending on the system), and to label them with arbitrary words, so-called *tags*. For an overview over the state of the art of folksonomy research, we refer to [3].

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.

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

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

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

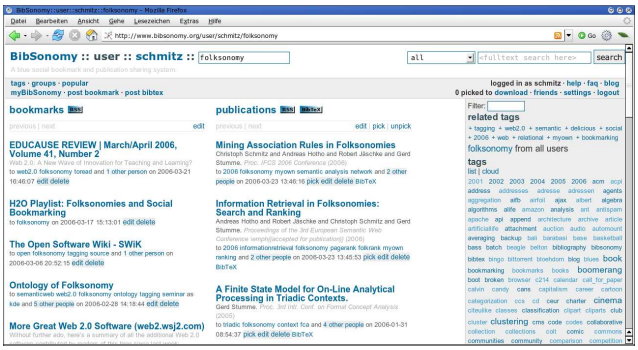

Figure 1: Bibsonomy displays bookmarks and BibTeX based bibliographic references simultaneously.

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.

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.

2 Folksonomies

A folksonomy describes the users, resources, and tags, and the user-based assignment of tags to resources. The following definition of folksonomies is also underlying our BibSonomy system.

A *folksonomy* is a tuple $\mathbb{F} := (U, T, R, Y)$ where U, T, and R are finite sets, whose elements are called *users*, *tags* and *resources*, resp., and Y is a ternary relation between them, i. e., $Y \subseteq U \times T \times R$, whose elements are called tag assignments (TAS for short). [4]

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.

This structure is known in Formal Concept Analysis [2] as a *triadic context* [5; 8]. An equivalent view on folksonomy data is that of a tripartite (undirected) hypergraph

[4] In the long version of this paper, we introduce additionally a user-specific tag hierarchy $\prec$.

$G = (V, E)$, 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.

In order to evaluate our retrieval technique detailed in the next section, we have analyzed the popular social bookmarking sytem 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, from July 27 to 30, 2005, data from the del.ico.us system, and obtained a core folksonomy with $|U| = 75,242$ users, $|T| = 533,191$ tags and $|R| = 3,158,297$ resources, related by in total $|Y| = 17,362,212$ tag assignments.

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 discussed above, 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 [1]. 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.

First we convert the folksonomy $\mathbb{F} = (U, T, R, Y)$ into an *un*directed 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 $|\{t \in T : (u, t, r) \in Y\}|$.

The original formulation of PageRank [1] reflects the idea that a page is important if there are 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. 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. Formally, we spread the weight as follows: $\vec{w} \leftarrow dA\vec{w} + (1 - d)\vec{p}$ where A is the row-stochastic[5] version of the adjacency matrix of $\mathbb{G}_{\mathbb{F}}$, $\vec{p}$ is the random surfer component, and $d \in [0, 1]$ is a constant which controls the influence of the random surfer.

Usually, one will set $\vec{p}$ as the vector where all values equal 1. 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 call the iteration according to the assignment above– until convergence is achieved – the *Adapted PageRank* algorithm. Note that, if $||\vec{w}||_1 = ||\vec{p}||_1$ holds,[6] the sum of the weights in the system will remain constant.

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 identical only if $d = 1$) to a ranking that is simply based on edge degrees. 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.ico.us dataset. As there exists no 'gold standard ranking' on these data, we evaluated our results empirically.

First we ran the algorithm with $d = 1$. We obtained the highest ranks for the tags, followed by the users, and the resources. This ranking provides an overview over the content of del.ico.us. 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. 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.ico.us are likely to be interested in Web 2.0 in general.

When using the random surfer vector to express user-specific preferences (e. g., by giving a considerably higher weight to a tag like 'boomerang'), we observed that although tags, users, and resources that are related to this preference are now ranked higher in the result, many of the general results mentioned above still hold the top positions. We ran this experiment with several settings of the preference vector, always with similar results. (For details see [3]). Apparently the preference vector is not strong enough to overcome the global graph structure.

[5] I. e., each row of the matrix is normalized to 1 in the 1-norm.

[6] ... and if there are no rank sinks – but this holds trivially in our graph $G_{\mathbb{F}}$.

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 the iteration with $d = 1$.
3. Let $\vec{w_1}$ be the fixed point from the iteration with $d < 1$.
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 $\vec{w}[x]$ 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 one set of search results, for the tag 'boomerang'. Our other experiments all provided similar results.

The top leftmost part of Table 1 contains the ranked list of tags according to their weights from the Adapted PageRank by using $d = 0.625$ 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. The right column in Table 1 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 right 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

Table 1: Ranking results with preference for the tag "boomerang" for the tags (left: Adapted PageRank, right: FolkRank for tags) and for the URLs (bottom: FolkRank)

Tag	ad. PRank
boomerang	0,4036883
shop	0,0069058
lang:de	0,0050943
software	0,0016797
java	0,0016389
programming	0,0016296
web	0,0016043
reference	0,0014713
system:unfiled	0,0014199
wood	0,0012378
kassel	0,0011969
linux	0,0011442
construction	0,0011023
plans	0,0010226
network	0,0009460
rdf	0,0008506
css	0,0008266
design	0,0008248
delicious	0,0008097
injuries	0,0008087
pitching	0,0007999

Tag	FolkRank
boomerang	0,4036867
shop	0,0066477
lang:de	0,0050860
wood	0,0012236
kassel	0,0011964
construction	0,0010828
plans	0,0010085
injuries	0,0008078
pitching	0,0007982
rdf	0,0006619
semantic	0,0006533
material	0,0006279
trifly	0,0005691
network	0,0005568
webring	0,0005552
sna	0,0005073
socialnetworkanalysis	0,0004822
cinema	0,0004726
erie	0,0004525
riparian	0,0004467
erosion	0,0004425

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/.../elbow_shoulder_injuries/	0,0031606
http://www.sportsprolo.com/...pitching%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

Table 2: Ranking results with preference for user "schm4704" for the tags (left: Adapted PageRank, right: FolkRank)

Tag	ad. PRank
boomerang	0,0093549
lang:ade	0,0068111
shop	0,0052600
java	0,0052050
web	0,0049360
programming	0,0037894
software	0,0035000
network	0,0032882
kassel	0,0032228
reference	0,0030699
rdf	0,0030645
delicious	0,0030492
system:unfiled	0,0029393
linux	0,0029393
wood	0,0028589
database	0,0026931
semantic	0,0025460
css	0,0024577
social	0,0021969
webdesign	0,0020650
computing	0,0020143

Tag	FolkRank
boomerang	0,0093533
lang:de	0,0068028
shop	0,0050019
java	0,0033293
kassel	0,0032223
network	0,0028990
rdf	0,0028758
wood	0,0028447
delicious	0,0026345
semantic	0,0024736
database	0,0023571
guitar	0,0018619
computing	0,0018404
cinema	0,0017537
lessons	0,0017273
social	0,0016950
documentation	0,0016182
scientific	0,0014686
filesystem	0,0014212
userspace	0,0013490
library	0,0012398

"rdf". When a community in del.ico.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 (right column) 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 often, "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 at the bottom of Table 1 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 resources for "boomerang" with the Top 20 resources given by the "schm4704" ranking (not shown here), 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.

This example– as well as the other experiments we performed (see also [3]) – 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.

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.

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. In the long version [3] of this paper, we discuss also how to use FolkRank for generating recommendations.

The FolkRank ranking scheme has been used in this paper to generate personalized rankings of the items in a folksonomy. We have seen that the top folksonomy elements which are retrieved by FolkRank tend to fall into a coherent topic area, e.g. "boomerang". 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.

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

References

[1] 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.

[2] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical foundations*. Springer, 1999.

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.