Trend Detection in Folksonomies

Andreas Hotho[1], Robert Jäschke[1,2], Christoph Schmitz[1], Gerd Stumme[1,2]

[1] 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
[2] Research Center L3S, Expo Plaza 1, D–30539 Hannover, Germany
http://www.l3s.de

Abstract. As the number of resources on the web exceeds by far the number of documents one can track, it becomes increasingly difficult to remain up to date on ones own areas of interest. The problem becomes more severe with the increasing fraction of multimedia data, from which it is difficult to extract some conceptual description of their contents.
One way to overcome this problem are social bookmark tools, which are rapidly emerging on the web. In such systems, users are setting up lightweight conceptual structures called folksonomies, and overcome thus the knowledge acquisition bottleneck. As more and more people participate in the effort, the use of a common vocabulary becomes more and more stable. We present an approach for discovering topic-specific trends within folksonomies. It is based on a differential adaptation of the PageRank algorithm to the triadic hypergraph structure of a folksonomy. The approach allows for any kind of data, as it does not rely on the internal structure of the documents. In particular, this allows to consider different data types in the same analysis step. We run experiments on a large-scale real-world snapshot of a social bookmarking system.

1 Social Resource Sharing and Folksonomies

With the growth of the web, both the number and the heterogeneity of types of available resources have increased dramatically. The management of such a collection of resources includes many subtasks like search, retrieval, clustering, reasoning, and knowledge discovery. For all these tasks, some sort of conceptual description of the documents is essential. While there are many approaches that have been applied successfully for years for extracting such descriptions from text documents — ranging from the bag-of-words model for information retrieval to ontology learning — there are fewer solutions for images, videos, audio tracks and music data up to now. The way from the features of the different resources to a conceptual description is generally far more difficult for multimedia data. Furthermore, these techniques have to be developed separately for each kind of data. For applications like the detection of trends from a collection of resources consisting of several types of (multimedia) data — which is the topic of this paper — first a common format for the representation of the conceptual model plus extraction techniques for each of the data types would have to be defined.

Complementing the extraction of conceptual descriptions from the documents themselves, social resource sharing tools are currently emerging on the web, as a part of what is called “social software” or “Web 2.0”. In these user-centric publishing and knowledge

management platforms, a conceptual description is provided to each document by the user in the form of a collection of 'tags', i. e., of arbitrary, user-defined catchwords. As this description is independent of the format of the resource, the social tagging approach provides a unified model for all kinds of resources, including in particular multimedia formats.

Social resource sharing tools, such as Flickr[3] or del.icio.us[4] (see Fig. 1), have acquired large numbers of users within less than two years. The social photo gallery Flickr, for instance, is estimated to have over a million users. The reason for the immediate success of these systems 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, and to label them. All these systems share the same core functionality. Once a user is logged in, he can add a resource to the system, and assign arbitrary labels, so-called *tags*, to it. Resources can be almost anything. In systems such as our *BibSonomy*,[5] for instance, resources are bookmarks and bibliographic references, in *Flickr* they are photos, in *last.fm*[6] music files, in *YouTube*[7] videos, and in *43Things*[8] even goals in private life.

The collection of all assignments of a user is called his *personomy*, the collection of all personomies is called *folksonomy*. The user can also explore the personomies 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; 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 (see Fig. 1).

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* [17, 18] which result from the converging use of the same vocabulary.

In this paper, we will analyze this emergence of common semantics by exploring trends in the folksonomy. Since the structure of a folksonomy is symmetric with respect to the dimensions 'user', 'tag', and 'resource', we can apply the same approach to study upcoming users, upcoming tags, and upcoming resources. We present a technique for analyzing the evolution of topic-specific trends. Our approach is based on our *FolkRank* algorithm [10], a differential adaptation of the PageRank algorithm [3] to the tri-partite hypergraph structure of a folksonomy. Compared to pure co-occurrence counting, FolkRank takes also into account elements that are related to the focus of interest with respect to the underlying graph/folksonomy. In particular, FolkRank ranks synonyms higher, which usually do not occur in the same bookmark posting together.

[3] http://www.flickr.com/ [4] http://del.icio.us [5] http://www.bibsonomy.org
[6] http://www.last.fm [7] http://www.youtube.com/ [8] http://www.43things.com/

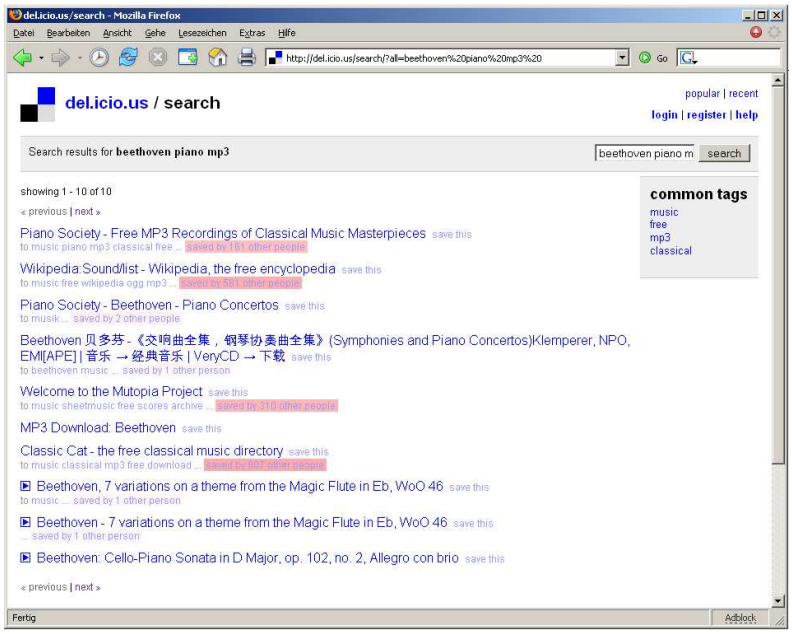

Fig. 1. Del.icio.us, a popular social bookmarking system.

With FolkRank, we compute topic-specific rankings on users, tags, and resources. In a second step, we can then compare these rankings for snapshots of the system at different points in time. We can discover both the absolute rankings (who is in the Top Ten?) and winners and losers (who rose/fell most?).

The contributions of this work are:

Ranking in folksonomies. We describe a general ranking scheme for folksonomy data. The scheme allows in particular for topic-specific ranking.

Trend detection. We introduce a trend detection measure which allows to determine which tags, users, or resources have been gaining or losing in popularity in a given time interval. Again, this measure allows to focus on specific topics.

Application to arbitrary folksonomy data. As the ranking is solely based on the graph structure of the folksonomy – which is resource-independent – we can also apply it to any kind of resources, including in particular multimedia objects, but also office documents which typically do not have a hyperlink structure per se. It can even be applied to an arbitrary mixture of these content types. Actually, the content of the tagged resources will not have to be accessible in order to manage them in a folksonomy system.

Evaluation. We have applied our method to a large-scale dataset from an actual folksonomy system.

The paper is organized as follows. In the next section, we describe our ranking and trend detection approach. In Section 3, we apply the approach to a large-scale dataset,

a one-year snapshot of the del.icio.us system. Section 4 discusses related work, and Section 5 concludes with an outlook on future topics in this field.

2 Trend Detection in Folksonomies

For discovering trends in a social resource sharing system, we will need snapshots of its folksonomy at different points of time. For each snapshot, we will need a ranking, such that we can compare the rankings of consecutive snapshots. As we also want to discover topic-specific trends, we will additionally need a ranking method that allows to focus on the specific topic. We will make use of our search and ranking algorithm *FolkRank* [10] which we summarize below.

2.1 Basic Notions

A folksonomy basically describes the users, resources, tags, and allows users to assign (arbitrary) tags to resources. We will make use of the following notions. 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., $Y \subseteq U \times T \times R$, whose elements are 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 $\prec_u := \{(t_1, t_2) \in T \times T \mid (u, t_1, t_2) \in \prec\}$ (whereas π stands for the projection).

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 [20, 7] as a *triadic context* [13, 19]. 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 ($\dot\cup$ is the disjunctive union).

2.2 Ranking

In this section we recall the principles of the *FolkRank* algorithm that we developed for supporting Google-like search in folksonomy-based systems. It is inspired by the seminal PageRank algorithm [3].

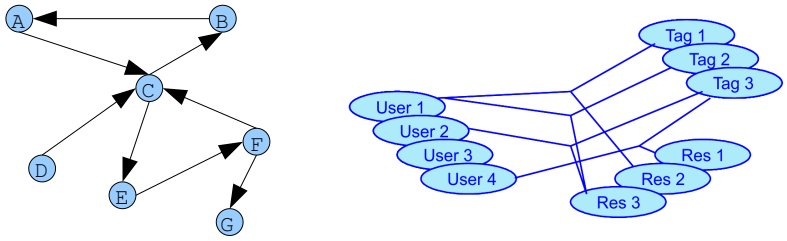

Fig. 2. Webgraph vs. Folksonomy Hypergraph

Because of the different nature of folksonomies compared to the web graph (undirected triadic hyperedges instead of directed binary edges, see Figure 2), PageRank cannot be applied directly on folksonomies. In order to employ a weight-spreading ranking scheme on folksonomies, we will overcome this problem in two steps. First, we transform the hypergraph into an undirected graph. Then we apply a differential ranking approach that deals with the skewed structure of the network and the undirectedness of folksonomies.

Folksonomy-Adapted Pagerank. First we convert the folksonomy $\mathbb{F} = (U, T, R, Y)$ into an *un*directed tri-partite graph $G_{\mathbb{F}} = (V, E)$ 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: $E := \{\{u, t\} \mid \exists r \in R : (u, t, r) \in Y\} \cup \{\{t, r\} \mid \exists u \in U : (u, t, r) \in Y\} \cup \{\{u, r\} \mid \exists t \in T : (u, t, r) \in Y\}$.

The original formulation of PageRank [3] 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 [12] and to n-ary directed graphs in [21]. 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 jumps to a new webpage without following a link. This results in the following definition. The rank of the vertices of the graph are the entries in the fixed point $\boldsymbol{w}$ of the weight spreading computation

$$\boldsymbol{w} \leftarrow dA\boldsymbol{w} + (1 - d)\boldsymbol{p} \ , \tag{1}$$

where $\boldsymbol{w}$ is a weight vector with one entry for each web page, A is a row-stochastic version of the adjacency matrix of the graph $G_{\mathbb{F}}$ defined above, $\boldsymbol{p}$ is the random surfer component that outweighs the loss of weight in dangling links, and $d \in [0, 1]$ is determining the influence of $\boldsymbol{p}$. Usually, one will choose $\boldsymbol{p} = \mathbf{1}$, i. e., the vector composed by 1's, to achieve uniform damping. In order to compute personalized PageRanks, however, $\boldsymbol{p}$ can be used to express user preferences by giving a higher weight to the components which represent the user's preferred web pages. If $||\boldsymbol{w}||_1 = ||\boldsymbol{p}||_1$,[9] the weight in the system will remain constant.

As the graph $G_{\mathbb{F}}$ is undirected, most 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, due to the damping) 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.

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

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

1. The preference vector $\boldsymbol{p}$ is used to determine the topic. It may have any distribution of weights, as long as $||\boldsymbol{w}||_1 = ||\boldsymbol{p}||_1$ holds. Typically a single entry or a small set of entries is set to a higher 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 giving a higher value to either one or more tags and/or one or more users and/or one or more resources.
2. Let $\boldsymbol{w_0}$ be the fixed point from Equation (1) with $d = 1$.
3. Let $\boldsymbol{w_1}$ be the fixed point from Equation (1) with $d < 1$. In our experiments, we set $d = 0.85$.
4. $\boldsymbol{w} := \boldsymbol{w_1} - \boldsymbol{w_0}$ is the final weight vector.

Thus, we compute the winners and losers of the mutual reinforcement of nodes when a user preference is given, compared to the baseline without a preference vector. We call the resulting weight $\boldsymbol{w}[x]$ of an element x of the folksonomy the *FolkRank* of x. In [10] we showed that $\boldsymbol{w}$ provides indeed valuable results on a large-scale real-world dataset while $\boldsymbol{w_1}$ provides an unstructured mix of topic-relevant elements with elements having high edge degree.

2.3 Trend Detection

In order to analyze the trends around a specific topic, we first have to describe the topic by defining the preference vector $\boldsymbol{p}$. Then we compute, for each point in time

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

$t \in \{0, \ldots, n\}$, the rank vector $\boldsymbol{w_t}$ within the folksonomy $\mathbb{F}_t$ which consists of all tag assignments performed before t.[10]

We select then from the resulting rank vectors those entries which are assigned to one of the three dimensions 'tags', 'users', and 'resources' — depending on where we want to see rising and falling elements. Else an analysis would be difficult, since users have higher weights than tags, which in their turn have higher weights than resources, due to the different sizes of the sets U, T, and R.

As the total weight in the system will differ at different points of time because of new tags, users, and resources, we normalize at last each rank vector such that its largest value equals 1. This allows to compare rankings from different points in time. If the preference vector has only one distinguished element, then this element is the one with the highest value in the resulting weight vector. The closer another entry is to this value, the more important is its associated element to the topic. By plotting the values of the Top 10 or Top 20 over time, one can thus discover the rise and fall of the most popular elements. Figure 3 shows such a plot for the del.icio.us users which are most important for the topic 'music', while Figure 4 shows the tags which are most important for the topic 'politics'. How these diagrams are to be read, and what the most important findings are, will be described in detail in the next section.

Going a step further, we may not only be interested in the most important elements, but also in those where the increase or decrease of rank is the steepest. To this end, we have developed the following *popularity change* measure, which allows for detecting topic-specific trends.

Assume x is a tag, user or resource of the folksonomy $\mathbb{F}$, i. e. $x \in U \cup T \cup R$. (In the following, we assume it is a resource; the same methods apply symmetrically for tags and users.) Similar to the relative change used for word occurrences in [11], we define the *popularity change* $pc_{t_0 \to t_1}(x)$ of x from t_0 to t_1 as follows.

At times $t_0 < t_1$, let the resource x be ranked at position r_0 and r_1, respectively, in the descending weight order of the FolkRank computation. Let n_0 and n_1 be the sizes $|R|$ of the resource dimension at times t_0, t_1. The popularity change is defined as

$$pc_{t_0 \to t_1}(x) := \left(\frac{r_0}{n_0} - \frac{r_1}{n_1} \right) \log_{10} \left(\frac{n_1}{r_1} \right) \tag{2}$$

(where elements not present at time t_i are treated as being positioned at $r_i = n_i + 1$). Here, the fractions in the first term indicate the relative positions of x at the given times, $1/n_i$ being the best (i. e. having maximum FolkRank) and 1 being the worst. The second term discounts the change with respect to the relative position where the change took place: to get from a top 90 % position to a top 80 % one would be considered three times easier than to get from the top 0.09 % to the top 0.08 %.

Combined with a topic-directed FolkRank computation, we use this measure of a change in popularity to get an insight into what are the trends in a certain community in the folksonomy. We point out the winning and losing elements of the folksonomy in a given time interval.

[10] If no entries were deleted, $\mathbb{F}_{t+1}$ contains thus $\mathbb{F}_t$, for all t.

3 Experiments

3.1 Evaluation of Popularity Change in del.icio.us

In order to evaluate our approach, we have analyzed the popular social bookmarking sytem del.icio.us.[11] 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. The resources del.icio.us is pointing to cover various formats (text, audio, video, etc.). In particular, the system is not restricted to a single type (like photos in Flickr). As discussed above, our approach is specially suited for this situation. In addition to the URL, del.icio.us allows to store a description, an extended description, and tags (i. e., arbitrary labels). Del.icio.us is online for a sufficiently long time (since May 2002) to allow for extracting significant time series.

For our experiments, we collected data from the del.icio.us system between July 27 and July 30, 2005 in the following way. Initially we used `wget` starting from the start 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). 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 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. We created monthly snapshots as follows. $\mathbb{F}_0$ contains all tag assignments performed on or before June 15, 2004, together with the involved users, tags, and resources; $\mathbb{F}_1$ all tag assignments performed on or before July 15, 2004, together with the involved users, tags, and resources; and so on until $\mathbb{F}_{13}$ which contains all tag assignments performed on or before July 15, 2005, together with the involved tags, users, and resources.

Figure 3 shows the evolution of the ranking of all users tags that were among the Top 10 in at least one month for the topic 'music'. The diagram was obtained with $d = 0.85$, and the preference vector $\boldsymbol{p}$ set such that the tag 'music' gets 50% of the overall preference, the rest is spread uniformly as described above. The user names have been omitted for privacy reasons. The diagram shows three outstanding users. The first one could keep the top position for the first four months, followed by a steep fall. Another user could approach him steadily during the first four months, followed by almost the same fall. The fall of both was caused by the steep rise of a new user, which also shadowed the rankings of all other users related to 'music'. A detailed analysis of this user's data in the system revealed us that he posted more than 5500 bookmarks, 85 % of which tagged with 'music'. In total he used only about 100 tags. The 5500 bookmarks account for about 2% of *all* occurrences of 'music' in the system (with more than 70.000 users in the system at that time), and are about 3.5 times as many as those of the second user for that tag.

Figure 4 shows the evolution of all tags that were among the Top Ten in at least one month for the topic 'politics'. The line for the topic 'politics' itself can't be seen, as it

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

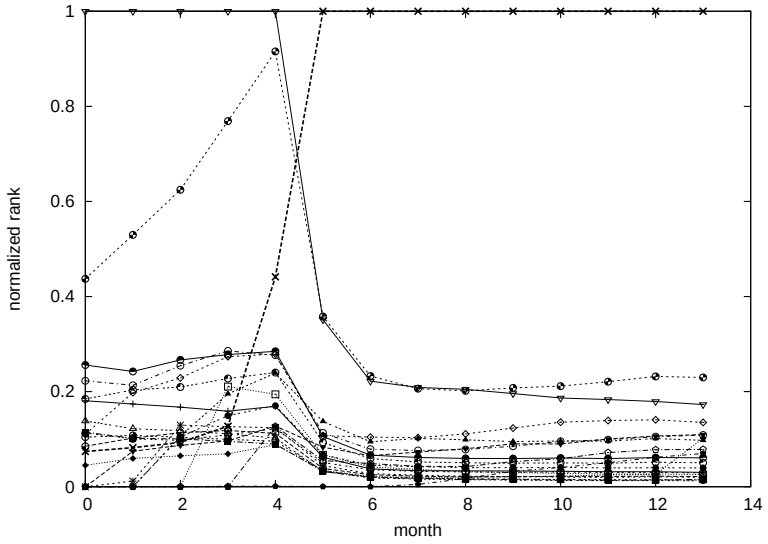

Fig. 3. Evolution of the ranking of users related to 'music'. User names are omitted for privacy reasons.

has a rank of 1. The diagram was obtained with $d = 0.85$, and the preference vector $\boldsymbol{p}$ set such that the tag 'politics' gets 50% of the overall preference, while the rest is spread uniformly over the other tags, users and resources. The diagram shows that the early users of del.icio.us were more critical/idealistic, as they used tags like 'activism', 'humor', 'war', and 'bushco'[12]. With increasing time, the popularity of these tags faded, and the tags turned to a more uniform distribution, as the closing lines at the right of the figure indicate. In particular one can discover the rise of the tags 'bush' and 'election', both having a peak around the election day, November 2nd, 2004, and remaining on a high level afterwards. Within the analysis of the topic 'technology' (not displayed due to space reasons), we have discovered a similar trend: The early adoptors of del.icio.us used the tag 'technology' together with tags like 'culture', 'society' or 'apple', while later tags like 'gadgets', 'news' or 'future' rise, converging towards more mainstream topics.

Both Figures 3 and 4 show that there is a change of structure in autumn 2004 (month 4 in the diagrams). This is supported by Figure 5 showing the development of the top resources. Analysing possible reasons for this change in behavior, one indicator is that the number of elements passed in month 4 the threshold of 10.000 users, 70.000 tags, and 500.000 resources. Apparently, with this number of users, one reaches a critical mass which modifies the inherent behavior of such a system. Figure 5 shows the rank of those resources which where among the top 5 at the beginning or the end. Our hypothesis that the del.icio.us community changes significantly at month 4 is supported by two observations: specific topics of the beginning, such as web design, see a decline,

[12] In del.icio.us, 'bushco' was used for tagging webpages about the interference of politics and economics in the U. S. administration.

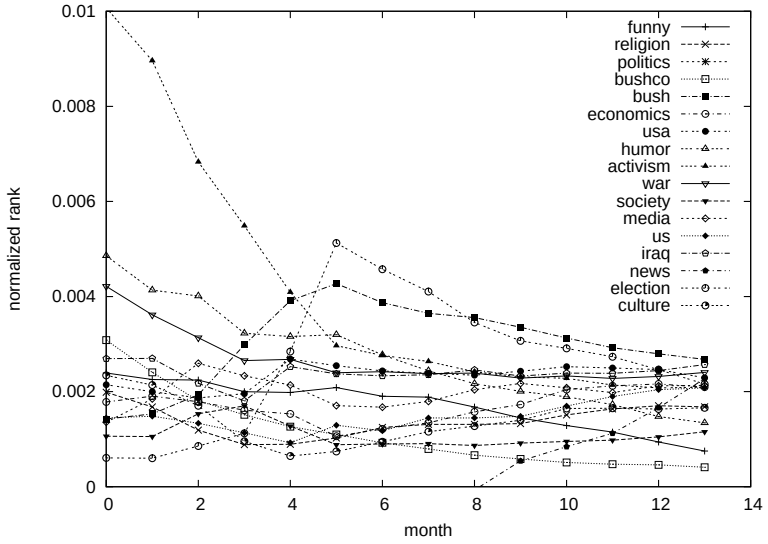

Fig. 4. Evolution of the ranking of tags related to 'politics' over time. 'Politics' has value 1.0 due to normalization and is left out for clarity of the presentation of the other values.

while on the other hand, mainstream pages gain rapidly, such as Slashdot, as well as pages concerned with folksonomies per se.

Finally, we have analyzed those resources that were the strongest winners and losers within specific topics, according to the popularity change measure defined in Section 2.3, to automatically identify trends within certain topics in del.icio.us. Our aim was to discover trends in the Semantic Web community in the month around the European Semantic Web Conference (ESWC) 2005.

For the computation, we took those resources that ended up in the Top 100 in the June 2005 ranking for the preference vector highlighting the tags 'semantic_web', 'semantic', 'web', and 'semanticweb', since the top results e. g. in searches are typically the ones attracting the most users. For these 100 URLs, we computed the popularity change coefficient from May 15 to June 15. Table 1 shows the 20 ULRs with the highest popularity change.

The top winner (#1) is a site about shallow semantic markup in XHTML, which was obviously first discovered by the community during the period under consideration and made it to the 39th position out of 2.2M resources; the corresponding line in Table 1 shows that the FolkRank value and the position in May is undefined for this resource, while the rank in June is 0.13065 and the position in this ranking is 39. Among the followers are articles about the Semantic Web and folksonomies (e. g. #2, #3, #5, #8, #9, #16), pages about new Semantic Web projects (#4, #15, #17, #19), or events such as the Scripting workshop (#7) that took place together with the ESWC conference during the period under consideration, introducing new Semantic Web projects. Note that while the #1 page leaped from nowhere to the 39th position out of 2.2 million entries, the popularity change measure still honors movements at the top of the ranking:

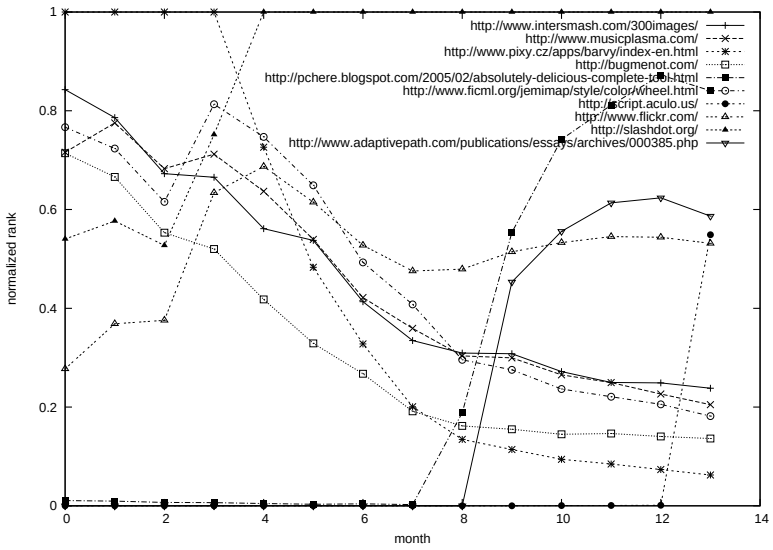

Fig. 5. Evolution of the global ranking of resources, without specific preference vector.

the Piggy Bank site (which is an important semantic web project that has been promoted at the ESWC conference), improving from 21 to 1 in the period, still gets into the top 15 winners.

Together, the results of the FolkRank computation and the popularity change measure presented in this section can thus be used to get an insight into the structure and development of communities in folksonomy systems, independent of and across different media types.

3.2 Comparison with the Interestingness of Dubinko et. al.

Closest to the approach of this paper is the visualization of Dubinko et. al. [6]. We tried to get an insight into how our FolkRank compares to the interestingness of [6]. In that paper, the authors introduce an efficient way of mining large-scale folksonomy data sets for frequent tags in given time intervals. A measure of *interestingness*, is introduced and computed for a sliding one-day window over a Flickr dataset. Similar to the TF/IDF measure from Information Retrieval, the interestingness is defined as $\text{Int}(o, I) = \sum_{i \in I} \gamma(o, i)/(C + \gamma(o))$, where $\gamma(o, i)$ is the number of occurrences of object o in time interval i out of a larger interval I, and $\gamma(o)$ is the total number of occurrences of o. As the interestingness is based on a count of occurrences of items[13] in a given interval, it does not allow for an easy integration of topic-specific rankings. Thus, one obtains a ranking of one particular tag (user, resource), which does not generalize to related elements of the folksonomy.

We computed the equivalent of Figure 5 for the interestingness measure, i. e., we show the rankings for those resources that were within the Top 5 for any of the months.

[13] In [6], only tags are evaluated. Still, the method can be applied symmetrically to users and resources.

Table 1. Popularity Change from May 15 to June 15, 2005.

#	URL	Pop.Chg.	May		June	
			Rank	Pos	Rank	Pos
1	http://mezzoblue.com/downloads/markupguide/	4.822604	undef	undef	0.13065	39
2	http://www.betaversion.org/~stefano/linotype/news/89/	4.515983	undef	undef	0.08296	79
3	http://shirky.com/writings/ontology_overrated.html	0.073704	0.00866	28598	0.39329	4
4	http://simile.mit.edu/piggy-bank/index.html	0.000805	0.05160	377	0.18740	24
5	http://www.dlib.org/dlib/april05/hammond/04hammond.html	0.000183	0.08831	142	0.09532	61
6	http://www.w3.org/2004/02/skos/	0.000175	0.08282	155	0.08369	78
7	http://www.semanticscripting.org/SFSW2005/	0.000134	0.09427	124	0.09055	67
8	http://www.scientificamerican.com/article.cfm?...	0.000133	0.08396	152	0.07208	97
9	http://jena.hpl.hp.com/~stecay/papers/xmleur...	0.000129	0.09979	111	0.09990	56
10	http://www.tantek.com/presentations/2004ete...	0.000112	0.09047	137	0.07407	92
11	http://users.bestweb.net/~sowa/peirce/ontometa.htm	0.000111	0.10273	106	0.09550	60
12	http://www.sciam.com/print_version.cfm?articleID=...	0.000101	0.09608	121	0.08178	81
13	http://www.xml.com/pub/a/2001/01/24/rdf.html	0.000089	0.09391	127	0.07314	94
14	http://developers.technorati.com/wiki/hCalendar	0.000071	0.09748	117	0.07389	93
15	http://simile.mit.edu/piggy-bank/	0.000057	0.29151	21	1.00000	1
16	http://en.wikipedia.org/wiki/Semantic_web	0.000033	0.10472	102	0.07186	98
17	http://www.semanticplanet.com/	0.000025	0.10893	91	0.07510	90
18	http://pchere.blogspot.com/2005/02/absolute...	0.000023	0.18154	41	0.13729	35
19	http://swoogle.umbc.edu/	0.000022	0.16785	48	0.12429	43
20	http://www.scientificamerican.com/print_ve...	0.000022	0.13367	68	0.09142	66

As our time window was one month, we used $C = 1500$ instead of $C = 50$ as in the original paper which used a one-day window. For lack of space and because the diagram did not yield any clear structure, we omit the diagram and summarize the findings.

The top resources were more volatile than in our method. I. e., in our approach, ten different resources made up the top five over all months. In the interestingness computation, there were 70 resources, i.e. each month had a new top five; Table 2 shows the top resource for each month. This indicates that the interestingness is more sensitive to momentary changes in the folksonomy than the FolkRank, and makes it harder to discover long- and medium-term trends. In the top resources, there were few general interest pages such as Slashdot or Flickr. Instead, there were more sites that seemed to be popular at one particular moment in time, but to fade soon afterwards. Figure 6 presents those four resources out of the 70 that overlap with Figure 5. It can be seen that while the interestingness shows some more jitter, the results have the same general direction for both computations.

We conclude that the interestingness, while more scalable and lending itself to a sliding-window visualization as in [6] due to its computational properties, lacks the dampening and generalizing effect of the FolkRank computation, so that it is more useful for short-term observations on particular folksonomy elements.

4 Related Work

There are currently only very few scientific publications about folksonomy-based web collaboration systems. Among the rare exceptions are [6] as discussed above, [8] and [14] who provide good overviews of social bookmarking tools with special emphasis on folksonomies, and [15] who discusses strengths and limitations of folksonomies. The main discussion on folksonomies and related topics is currently only going on mailing lists, e.g. [4]. In [16], Mika defines a model of semantic-social networks for

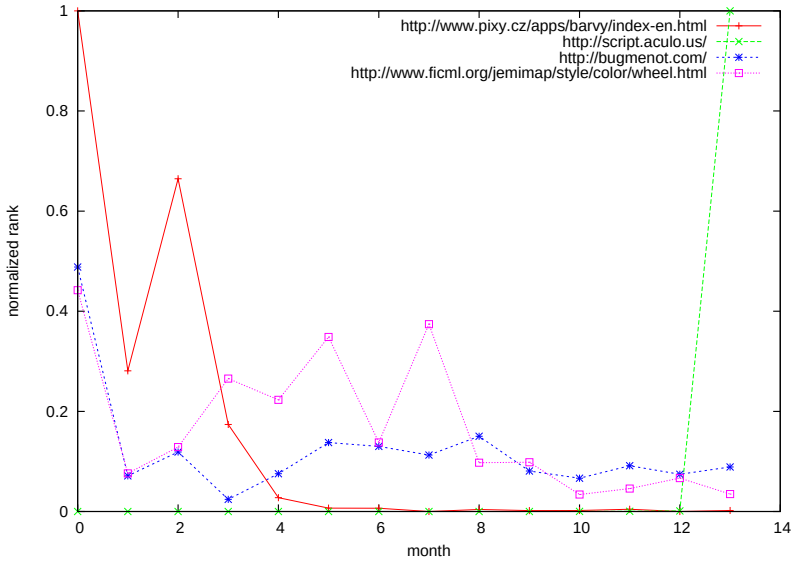

Fig. 6. Evolution of the interestingness values of those resources which overlap with Figure 5; graph plotted the same way as Figure 5.

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[14] 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[15]) or focus on graphical representations (Cloudalicious[16], Grafolicious[17]) of statistics about del.icio.us.

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 [3] algorithm is used to find communities of practice within individuals represented in the ontology. OntoRank [5] uses a PagesRank-like approach on the RDF graph to rank search results within Swoogle, a search engine for ontologies.

Along the same line, in [9], we have presented a technique for analysing ontologies that considers not only the first eigenvector (as PageRank and Ontocopi do), but the full eigensystem of the adjacency matrix of the ontology.

In [2], the evolution of the web graph over time is analyzed. The application of the proposed method lies in the improved detection of current real-life trends in search engines. In comparison to our work, they base their approach on counting timestamped

[14] http://collabrank.org/ [15] http://populicio.us/ [16] http://cloudalicio.us/
[17] http://www.neuroticweb.com/recursos/del.icio.us-graphs/

Table 2. Top resources for each month according to the interestingness defined in [6]

Month	Resource	Int'ness
0	http://www.pixy.cz/apps/barvy/index-en.html	0.1937
1	http://craphound.com/msftdrm.txt	0.0970
2	http://extensions.roachfiend.com/howto.html	0.1339
3	http://richard.jones.name/google-hacks/gmail-filesystem/gmail-filesystem.html	0.1983
4	http://37signals.com/papers/introtopatterns/	0.2150
5	http://www.fuckthesouth.com/	0.1898
6	http://www.supermemo.com/articles/sleep.htm	0.2585
7	http://www.returnofdesign.com/spectacle/specials/colors.html	0.2958
8	http://www.hertzmann.com/articles/2005/fables/	0.4117
9	http://fontleech.com/	0.4906
10	http://pro.html.it/esempio/nifty/	0.6511
11	http://www.alvit.de/vf/en/essential-. . . -developers.html	0.5678
12	http://www.newscientist.com/channel/being-human/mg18625011.900	0.6222
13	http://script.aculo.us/	0.8478

links on pages returned by web searches on given topics, while our contribution infers communities around given users, sites, or topics from the structure of the web graph itself. The algorithm of [2] can currently not be applied to folksonomies, as there exist no folksonomy search engines yet.

Kleinberg [11] summarizes several different approaches to analyze online information streams over time. He distinguishes between three methods to detect trends: using the normalized absolute change, relative change and a probabilistic model. The popularity gradient that we introduced in Section 2.3 is related to the second approach, but differs insofar as it allows for the discovery of *topic-specific* trends, and that we honor steep rises more if they occur higher in the ranking, where the text mining scenario described in [11] requires focusing on words that are neither too frequent nor too infrequent.

5 Conclusion and Outlook

In this paper we have shown how topic-specific trends can be discovered in folksonomy-based systems. The analysis can be done regardless of the types of the underlying resources, which makes folksonomies interesting for multimedia applications.

As folksonomies are still rather young, there are many fascinating research topics left open that are related to the work presented here. They include predicting the change of structure of the folksonomy during its growth, discovering stable and volatile communities, and generating recommendations.

Acknowledgement. This work was partially supported by the European Commission in the projects "NEPOMUK — The Social Semantic Desktop" and "TAGORA - Emergent Semantics in Online Social Communities" under the 6th Framework Programme.

References

1. H. Alani, S. Dasmahapatra, K. O'Hara, and N. Shadbolt. Identifying Communities of Practice through Ontology Network Analysis. *IEEE Intelligent Systems*, 18(2):18–25, 2003.

2. E. Amitay, D. Carmel, M. Herscovici, R. Lempel, and A. Soffer. Trend detection through temporal link analysis. *J. Am. Soc. Inf. Sci. Technol.*, 55(14):1270–1281, 2004.
3. S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, April 1998.
4. Connotea Mailing List. https://lists.sourceforge.net/lists/listinfo/connotea-discuss.
5. L. Ding, R. Pan, T. W. Finin, A. Joshi, Y. Peng, and P. Kolari. Finding and ranking knowledge on the semantic web. In *International Semantic Web Conference*, pages 156–170, 2005.
6. M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *Proc. 15th Int. WWW Conference*, May 2006.
7. B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical foundations*. Springer, 1999.
8. T. Hammond, T. Hannay, B. Lund, and J. Scott. Social Bookmarking Tools (I): A General Review. *D-Lib Magazine*, 11(4), April 2005.
9. B. Hoser, A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Semantic network analysis of ontologies. In Y. Sure and J. Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *LNAI*, pages 514–529, Heidelberg, June 2006. Springer.
10. A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In Y. Sure and J. Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *LNAI*, pages 411–426, Heidelberg, June 2006. Springer.
11. J. Kleinberg. Temporal dynamics of on-line information streams. In M. Garofalakis, J. Gehrke, and R. Rastogi, editors, *Data Stream Management: Processing High-Speed Data Streams*. Springer, 2006.
12. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
13. F. Lehmann and R. Wille. A triadic approach to formal concept analysis. In *Conceptual Structures: Applications, Implementation and Theory*, volume 954 of *LNCS*. Springer, 1995.
14. B. Lund, T. Hammond, M. Flack, and T. Hannay. Social Bookmarking Tools (II): A Case Study - Connotea. *D-Lib Magazine*, 11(4), April 2005.
15. A. Mathes. Folksonomies – Cooperative Classification and Communication Through Shared Metadata, December 2004. http://www.adammathes.com/academic/computer-mediated-communication/folk%sonomies.html.
16. P. Mika. Ontologies Are Us: A Unified Model of Social Networks and Semantics. In Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen, editors, *ISWC 2005*, volume 3729 of *LNCS*, pages 522–536, Berlin Heidelberg, November 2005. Springer-Verlag.
17. 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.
18. 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.
19. G. Stumme. A finite state model for on-line analytical processing in triadic contexts. In B. Ganter and R. Godin, editors, *Proc. 3rd Intl. Conf. on Formal Concept Analysis*, volume 3403 of *Lecture Notes in Computer Science*, pages 315–328. Springer, 2005.
20. 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.
21. 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.