Datenbank Spektrum (2010) 10: 15–24
DOI 10.1007/s13222-010-0004-8

SCHWERPUNKTBEITRAG

Query Logs as Folksonomies

Dominik Benz · Andreas Hotho · Robert Jäschke · Beate Krause · Gerd Stumme

Received: 2 February 2010 / Accepted: 4 March 2010 / Published online: 6 May 2010

Abstract Query logs provide a valuable resource for preference information in search. A user clicking on a specific resource after submitting a query indicates that the resource has some relevance with respect to the query. To leverage the information of query logs, one can relate submitted queries from specific users to their clicked resources and build a tripartite graph of users, resources and queries. This graph resembles the folksonomy structure of social bookmarking systems, where users add tags to resources. In this article, we summarize our work on building folksonomies from query log files. The focus is on three comparative studies of the system's content, structure and semantics. Our results show that query logs incorporate typical folksonomy properties and that approaches to leverage the inherent semantics of folksonomies can be applied to query logs as well.

D. Benz (✉) · R. Jäschke · B. Krause · G. Stumme
Hertie-Stiftungslehrstuhl Wissensverarbeitung, Universität Kassel, Kassel, Germany
e-mail: benz@cs.uni-kassel.de

R. Jäschke
e-mail: jaeschke@cs.uni-kassel.de

B. Krause
e-mail: krause@cs.uni-kassel.de

G. Stumme
e-mail: stumme@cs.uni-kassel.de

A. Hotho
Data-Mining- und Information-Retrieval-Gruppe, Universität Würzburg, Würzburg, Germany
e-mail: hotho@informatik.uni-wuerzburg.de

1 Introduction

Collaborative tagging systems such as Delicious,[1] BibSonomy,[2] or Flickr[3] have become popular among Internet users in the last years. Taggers actively index and describe web resources by adding keywords to interesting content and sharing their entries with other web users. The data structure evolving from these activities is called a *folksonomy*. Over the last years, a significant number of resources has been collected, offering a new form of searching and exploring the web. To folksonomy users, this personalized, community driven search has become an alternative to traditional web search engines.

The major differences between a folksonomy and a search engine can be found in the way content is created and information presented. Folksonomies allow users to explore their content in different dimensions taking users, tags and resources into account. In contrast, classical search engines offer a simple user interface, where users enter their information need and view a result list sorted according to relevancy. In folksonomies, users themselves—not an algorithm—decide about relevance by explicitly tagging the contents. Search engines create their indexes and results without human input by means of automatic crawlers and intelligent ranking mechanisms.

User relevance feedback can be integrated into ranking algorithms as well. The feedback is extracted from log files which track a user's click history. As the evolvement of social bookmarking systems has shown, many web searchers are not only interested in a ranked list of search results.

[1] http://www.delicious.com/.
[2] http://www.bibsonomy.org/.
[3] http://www.flickr.com/.

The discovery of interesting, new information, based on navigation through similar users or tags can also help to fulfill an information need. One possibility to realize such "search communities" within search engines is the building of an anonymized folksonomy similar to the Delicious social bookmarking system from search engine logdata. As logdata contains queries, clicks and session IDs, the classical dimensions of a folksonomy can be reflected: Queries or query words represent tags, session IDs correspond to users, and the URLs clicked by users can be considered as the resources that they tagged with the query words. Search engine users can then browse this data along the well known folksonomy dimensions of tags, users, and resources. We refer to this structure as *logsonomy*.

Folksonomies and logsonomies raise a variety of questions. Are user interactions with search engines and folksonomies similar? How does the structure of logsonomies differ from the one of folksonomies? Do logsonomies adhere to inherent semantics as has been shown for folksonomies [5]?

Understanding the similarities and differences between folk- and logsonomies can help to apply approaches used with folksonomies to query logs. One application can be the realization of a folksonomy-alike navigation through search histories. Another application is the filtering of inherent semantics from logsonomies and the enrichment of common IR methods with it. For example, one can detect synonyms or spelling variants of queries. Another possibility is to identify related words to an original query and use those words for query expansion [1, 5]. Further, the computation of rankings can be enhanced by introducing frequently clicked URLs or related queries as features into the ranking or by computing a graph based ranking using the triparted hypergraph built from logsonomies instead of using the web graph.

In this article, we summarize our findings from different comparative studies between folk- and logsonomies to provide a deeper understanding of both paradigms. We focus on three areas. A content-based analysis of the vocabulary used by searchers and taggers gives initial insights into the usage of both systems [11]. In order to find out, if the logsonomy graph offers similar navigation structures as folksonomies, we look at small world properties [10, 12]. Finally, we analyze the inherent semantics of queries by computing similar query terms with the help of relatedness measures established for folksonomies [2].

2 Background

The term *folksonomy* refers to a lightweight classification structure which is built from shared tag annotations added by different users to their resources. Following [8], we model a folksonomy as a quadruple $\mathbb{F} := (U, T, R, Y)$ where U, T, and R are finite sets, whose elements are called *users*, *tags* and *resources*, resp., Y is a ternary relation between them, $Y \subseteq U \times T \times R$, whose elements are called tag assignments (*tas* for short). In social bookmarking systems, the quadruple consists of users (U), adding keywords (T) to bookmarks (R), which results in tag assignments (Y).

Let us consider the query log of a search engine. To map it to the three dimensions of a folksonomy, we set U to be the set of users of the search engine. Depending on how users in logs are tracked, a user is represented either by an anonymized user ID, or by a session ID. Let T be the set of queries the users submitted to the search engine (where one query either results in one tag, or will be split at white spaces into several tags). Let R be the set of URLs which have been clicked by the search engine users. In a logsonomy, we assume an association between t, u and r when a user u clicked on a resource r of a result set after having submitted a query t.

Please note that queries from a query log can be handled in two ways. The query t can either consist of a full query or can be split into single terms $t_1, \ldots, t_k$. For example, the query "semantic web tutorials" can either be handled as one tag, or be split into the single terms "semantic", "web", "tutorials". This results in either one or three triples of the relation $Y \subseteq U \times T \times R$, which correspond to the tag assignments in a folksonomy. We call the resulting structure a *logsonomy*, since it resembles the formal model of a folksonomy. The process of creating such model is similar to the one of folksonomies: Users describe an information need in terms of a query. They then restrict the result set of the search engine by clicking on those URLs whose snippets indicate that the website has some relation to the query. In some important points, however, one can note differences:

- While tagging a specific resource can be seen as an indicator for relevance, users may click on a resource to check if the result is important and then disappointedly return to the initial search list.
- Users experience a bias towards clicking the top results of a result list.
- In search engines, queries are submitted first; afterwards a result list of different URLs is shown. In logsonomies, we interpret the query as the description of the underlying, clicked resource. Splitting these descriptions in single words may destroy or change the intended meaning.
- Queries are processed by search engines leaving open to which extent the terms influence the search results. They may be ignored or enhanced with similar query terms.
- When a resource never comes up in a search result, it cannot be tagged as such.
- If no user IDs are available, logsonomies can only be build using session IDs to represent a user. The IDs are

Table 1 Statistics of Delicious and MSN in May, 2006

	MSN	Delicious	MSN ∩ Delicious
Terms	31,999,521	8,995,085	–
Distinct terms	2,224,550	377,515	97,626
Average	14,38	23,83	–
Frequent terms	127,509	39,043	18,616
Frequent terms containing "_"	95	1,840	0
Frequent terms containing "-"	1,821	1,613	142
Frequent terms containing "www.", ".com", ".net" or ".org"	19,664	145	36

probably more coherent, as users tend to view similar topics in one session; however, they cannot reflect a typical user who has several interests or changes interests over time as can be done with folksonomies.

As we focus on a comparative study using methods which have already been applied to folksonomies, we will refrain from explicitly analyzing the influence of these issues to structural or semantic differences between log- and folksonomies. Some of our findings, however, can be attributed to those distinctions. For example, the appearance of URLs as queries in a logsonomy, which contrasts to folksonomies, is due to the fact, that people express an information need, not a description of a resource when searching (see Sect. 4).

3 Datasets

We used three different data sources for our experiments. In November 2006 we crawled Delicious to obtain a comprehensive social bookmarking set with tag assignments from the beginning of the system to October 2006. Based on the time stamps of the tag assignments, we were able to produce different snapshots, for example all users, resources and tags of May 2006.

We obtained a click data set from Microsoft for the period of May 2006.[4] The MSN dataset consists of about 15 million queries submitted in 7,470,915 different sessions which were tracked from the MSN search engine users in the United States in May 2006.

A second data set is from AOL [13]. The data was collected from March, 1st to May, 31st 2006. The dataset consists of 657,426 unique user IDs, 10,154,742 unique queries, and 19,442,629 click through events.

For each of the experiments we constructed appropriate snapshots of the data. The sizes and specifics are provided with the different experiments.

[4] http://research.microsoft.com/ur/us/fundingopps/RFPs/Search_2006_RFP.aspx.

4 Query Word and Tag Usage Analysis

To begin with an analysis of logsonomies, we compare the usage of both systems by analyzing their contents. The overlap of the set of query words with the set of tags is an indicator of the similarity of the usage of both systems. We use all tags of May 2006 of Delicious to represent social bookmarking systems, and all queries of the MSN dataset. We normalized both systems by removing terms containing only one single character.

Table 1 shows statistics about the usage of query words in MSN and tags in Delicious. The first row reflects the total number of queried words, and the total number of used tags in Delicious. The following row shows the number of distinct terms in all systems. As can be seen, both the total number of words and the number of distinct words is significantly larger in MSN compared to the total number of tags and the number of distinct tags in Delicious. Interestingly, the average frequency of a term is quite similar in both systems (see third row). These numbers indicate that Delicious users focus on fewer topics than search engine users, but that each topic is, on average, equally often addressed.

The relative overlap of the MSN query words with the Delicious terms (2nd row/3rd column of Table 1) is rather low. This observation can be explained by the numerical distribution of terms (i.e., tags and query words) in both systems. Figure 1 shows the distribution of terms in both systems on a log-log scale. The right-most box, for instance, comes from the query word 'yahoo' which is the only query word that is occurring 179,651 times in MSN. The top-most box comes from 1,439,604 different query words that occur once only.

The plot confirms a finding of [7], that the distribution of the counts of tags (and of resources) in a folksonomy follows a power law distribution, $P(k) \sim k^{-\gamma}$, where k is the count and γ the exponent of the distribution. In a log-log-scaled plot, as in Fig. 1, a power law means that the data points are all on a straight line (which has $-\gamma$ as gradient). A power law distribution implies in particular, that a very high number of terms occurs very rarely, and that only very few terms have very high counts.

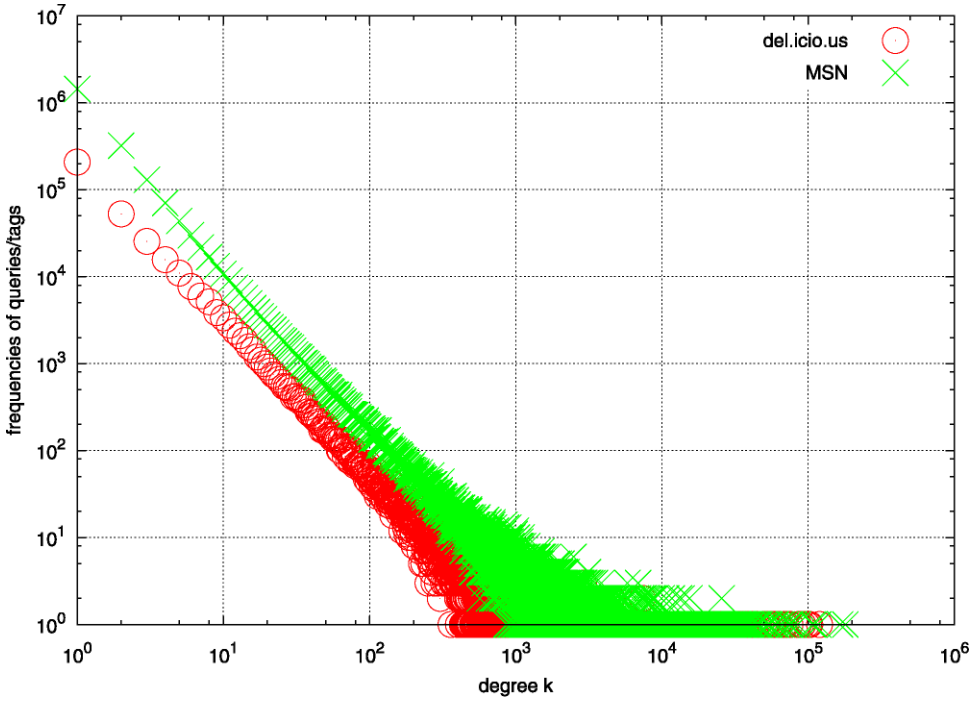


Fig. 1 Term distribution in MSN and Delicious. The x-axis denotes the count of terms in the data set, and the y-axis shows the number of terms that have the given count

The fact that both Delicious tags and MSN query words follow a power law explains the above-mentioned small overlap of the MSN query words with the Delicious terms: Most of the tags and query words are in the "long tail" of their distribution, i.e., show up very rarely only. It is therefore very unlikely that they consist of common English words (or words of some other natural language). Hence their potential to show up also in the other system is very low.

In order to analyze the overlap for the more central terms, we restricted both sets to query words/tags that showed up in the respective system at least ten times. The resulting frequencies are given in the first line of the second part of Table 1. It shows that the sizes of the reduced MSN and Delicious datasets become more equal, and that the relative overlap increases.

When browsing both reduced data sets, we observed that the non-overlapping parts result very much from the different usages of both systems. In social bookmarking systems, for instance, people frequently encode multi-word lexemes by connecting the words with either underscores, hyphens, dots, or no symbol at all. (For instance, all of the terms *artificial_intelligence*, *artificial-intelligence*, *artificial.intelligence* and *artificialintelligence* show up at least ten times in Delicious). This behavior is reflected by the second and third last rows in Table 1. Underscores are basically used for such multi-word lexemes only, whereas hyphens occur also in expressions like *e-learning* or *t-shirt*. Only in the latter form they show up in the MSN data.

A large part of the query words in MSN that are not Delicious tags are URLs or part of URLs, see the last row of Table 1. This indicates that users of social bookmarking systems prefer tags that are closer to natural language, and thus easier to remember, while users of search engines (have to) anticipate the syntactic appearance of what they are looking for.

5 Structural Properties

The observations of the last section have shown that the usage of both systems is similar for common terms. Due to the different processes of searching and tagging, however, the vocabulary differs in the long tail. In this section, we will turn away from a direct comparison of contents to explore structural properties.

Folksonomies exhibit small world characteristics: a graph topology for which the degree of clustering is almost as high as that of a regular graph with a same degree distribution for each node; but the average shortest path length is almost as small as that of a random graph [14]. These characteristics are one explanation for the popularity of social bookmarking systems: on the one hand, resources fulfilling a specific information need are clustered together in the folksonomy, on the other hand, users can reach most of the contents within a few clicks.

In the following, we investigate to which extent these characteristics hold for logsonomies. We created six folk- and logsonomy datasets from the available Delicious, MSN and AOL data as follows (Table 2): In the dataset *MSN complete queries*, the set of tags is the set of complete queries, the set of users is the set of sessions and the set of resources

Table 2 Sizes, average shortest path lengths, cliquishness and connectedness for the structural properties experiments

Dataset	Sizes				ASPL			Cliquishness			Connectedness		
	$\lvert T\rvert$	$\lvert U\rvert$	$\lvert R\rvert$	$\lvert Y\rvert$	Raw	Shuffled	Binomial	Raw	Shuffled	Binomial	Raw	Shuffled	Binomial
Delicious, complete URLs	430,526	81,992	934,575	14,730,683	3.59	3.08	3.99	0.86	0.55	0.20	0.85	0.37	0.00
Delicious, host only URLs	430,526	81,992	2,913,354	16,217,222	3.48	3.06	3.67	0.75	0.51	0.05	0.83	0.32	0.00
AOL, complete queries	4,811,436	519,250	1,620,034	14,427,759	4.11	3.81	5.76	0.85	0.66	0.32	0.33	0.03	0.00
AOL, split queries	1,074,640	519,203	1,619,871	34,500,590	3.62	3.20	3.90	0.70	0.43	0.04	0.66	0.10	0.00
MSN, complete queries	3,545,310	5,680,615	1,861,010	10,880,140	5.43	4.10	8.78	0.87	0.75	0.47	0.42	0.03	0.00
MSN, split queries	902,210	5,679,240	1,860,728	24,204,125	3.94	3.42	5.48	0.85	0.50	0.23	0.70	0.11	0.00

is the set of clicked URLs. For the second dataset, *MSN split queries*, we decomposed each query t at whitespace positions into single terms $(t_1, \ldots, t_k)$ and collected the triples (u, t_i, r) (for $i \in 1, \ldots, k$) in Y instead of (u, t, r). This splitting shall better resemble the tags added to resources in folksonomies which typically are single words. To make both datasets comparable to the AOL dataset, which consists of host-only URLs, we reduced the MSN full URLs to host-only URLs.

The AOL data was transformed into the two datasets *AOL complete queries* and *AOL split queries* analogously to the MSN datasets. We used unique user IDs for the AOL dataset, because session IDs were not included in the AOL dataset.

To compare the logsonomy structure to a folksonomy, we used a dataset from Delicious containing posts from 81,992 users up to July, 31st 2005. Again, we have two datasets: one consisting of full URLs to be comparable to prior work on folksonomies, and one reduced to the host part of the URL only to be comparable to the logsonomy datasets.

We followed the experiments of [4] in order to be comparable to former findings regarding folksonomy properties. In these experiments, binomial and shuffled (hyper-)graphs of the same size as the original folksonomy were selected to compare the original graph to random graphs. For a given folksonomy (U, T, R, Y), a *binomial* random graph is a logsonomy $(U, T, R, \hat{Y})$ where $\hat{Y}$ consists of $|Y|$ randomly drawn tuples from $U \times T \times R$. A *shuffled* random graph is then a folksonomy $(U, T, R, \check{Y})$ where $\check{Y}$ is derived from Y by randomly shuffling all occurrences of tags in Y, followed by shuffling all occurrences of the resources. (For a complete shuffling, it is sufficient to shuffle any two of the three dimensions.) The binomial graph has thus the same number of tag assignments as the original graph, while the shuffled graph has additionally the same degree distribution.

5.1 Average Shortest Path Length

The average shortest path length (ASPL) denotes the mean distance between any two nodes in the graph. In a tripartite hypergraph, a path between any two nodes is a sequence of hyperedges that lie between them. The shortest path is a path with the minimum number of hyperedges connecting the two nodes.

For complexity reasons, we approximated the average shortest path length as follows. For each of the datasets, we randomly selected 4,000 nodes and calculated the shortest path length of each of those nodes to all other nodes in its connected component.

Table 2 shows the average shortest path length of each dataset together with the values for the corresponding random graphs. Comparing the two Delicious datasets, the average shortest path length does not vary to a large extent when considering host only URLs (3.48 for the host-only-graph versus 3.59 for the graph with complete URLs). The

average shortest path length of the AOL and MSN datasets with split queries are smaller than those of the datasets with complete queries. This can be explained by the higher overlap, which is produced by the splitting of queries. As a side effect, this also leads to a mixing of contents, e.g., the word *java* in *java programming language* and *java island* will link to different topics. However, such wording issues also exist in folksonomies.

Compared to Delicious, all four datasets from MSN and AOL provide larger path lengths. Capturing the intuition of serendipitous browsing, it takes longer to reach other queries, users, or URLs within a logsonomy than it takes to jump between tags, users and resources in a folksonomy. In particular, the high values for MSN are likely to result from the fact that a user cannot bridge between different topics if he searched for them in different sessions.

Small world properties are still confirmed by the shortest path length: Comparing each logsonomy to the corresponding binomial and random graphs, the path lengths differ only slightly.

5.2 Clustering Coefficient

The clustering coefficient characterizes the density of connections in the environment of a node. It describes the cliquishness, (i.e., *are neighbor nodes of a node also connected among each other*) and the connectedness of a node, (i.e., *would neighbor nodes stay acquainted if the node was removed*). In a tripartite graph, one needs to consider these two characteristics separately. In [4], two measures were proposed which are summarized in the following.

Cliquishness. Consider a resource r. Then the following sets of tags T_r and users U_r are said to be connected to r: $T_r = \{t \in T \mid \exists u \in U : (t, u, r) \in Y\}$, $U_r = \{u \in U \mid \exists t \in T : (t, u, r) \in Y\}$. Furthermore, let $tu_r := \{(t, u) \in T \times U \mid (t, u, r) \in Y\}$, i.e., the (tag, user) pairs occurring with r.

If the neighborhood of r was maximally cliquish, all of the pairs from $T_r \times U_r$ would occur in tu_r. So we define the cliquishness coefficient $\gamma_{cl}(r)$ as:

$$\gamma_{cl}(r) = \frac{|tu_r|}{|T_r| \cdot |U_r|} \in [0, 1]. \tag{1}$$

The cliquishness is defined likewise for tags and users.

Connectedness. Consider a resource r. Let $\widetilde{tu_r} := \{(t, u) \in tu_r \mid \exists \tilde{r} \neq r : (t, u, \tilde{r}) \in Y\}$, i.e., the (tag, user) pairs from tu_r that also occur with some other resource than r. Then we define:

$$\gamma_{co}(r) := \frac{|\widetilde{tu_r}|}{|tu_r|} \in [0, 1]. \tag{2}$$

γ_{co} is thus the fraction of r's neighbor pairs that would remain connected if r were deleted. It indicates to what extent the surroundings of the resource r contain "singleton" combinations *(tag, user)* that only occur once. The connectedness is defined likewise for tags and users.

The results in Table 2 show that the average cliquishness and connectedness coefficients of the original AOL, MSN and Delicious graphs are in general higher than the ones of the corresponding random graphs. This indicates that there is some systematic aspect in the search behavior which is destroyed in the randomized versions. Comparing the two logsonomies to the folksonomy, however, one can conclude that the clustering coefficients of the folksonomy exceeds those of logsonomies. This is probably due to the higher variety of topics in the logsonomy datasets—whereas Delicious is very focused on computer related terms. Additionally, users in folksonomies tend to add similar tags to a resource, while resources in web search engines will be retrieved by many different queries. This relates to the issues in Sect. 2: The process in which tags are created in logsonomies and folksonomies is different.

6 Semantic Properties

The previous section revealed that folksonomies and logsonomies show similar structural characteristics, e.g., small world properties. These findings support the idea of enabling some kind of browsing facilities in search engines. Another exciting property of folksonomies is the inherent semantic which occurs from the process of tagging. We now present some of the experiments of [2], which investigate to which extend a logsonomy allows the extraction of semantics emerging from the "collaborative" process of searching similar information and being interested in the same resources.

For our folksonomy experiments, we used the Delicious data from 2006. We restricted the dataset to the 10,000 most frequent tags, and to the resources/users that have been associated with at least one of those tags. For the logsonomy representation, we used the click dataset from the AOL search log. Again, we constructed a logsonomy, this time with the restriction of only using the 10,000 most frequent query words to the dataset. The resulting sizes of the datasets are shown in Table 3.

6.1 Relatedness Measures

Different relatedness measures can be used to extract semantic similarities between query parts from logsonomies.

Given a logsonomy (U, T, R, Y), we define the *query word co-occurrence graph* as a weighted undirected graph

Table 3 Folksonomy and logsonomy datasets

Dataset	$\lvert T\rvert$	$\lvert U\rvert$	$\lvert R\rvert$	$\lvert Y\rvert$
Delicious	10,000	476,378	12,660,470	101,491,722
AOL split queries	10,000	463,380	1,284,724	26,227,550

whose set of vertices is the set T of query words. For all users $u \in U$ and resources $r \in R$ let T_{ur} be the set of query words within one query, i.e., $T_{ur} := \{t \in T \mid (u,t,r) \in Y\}$. Two query words t_1 and t_2 are connected by an edge, iff there is at least one query (u, T_{ur}, r) with $t_1, t_2 \in T_{ur}$. The *weight* of this edge is given by the number of queries that contain both t_1 and t_2, i.e.,

$$w(t_1,t_2) := \text{card}\{(u,r) \in U \times R \mid t_1, t_2 \in T_{ur}\}. \tag{3}$$

Co-occurrence relatedness (Co-Occ) between query words is given directly by the edge weights. For a given query word $t \in T$, the tags that are most related to it are thus all the tags $t' \in T$ with $t' \neq t$ such that $w(t,t')$ is maximal.

We introduce three distributional measures of query word relatedness that are based on three different vector space representations of query words. The difference between the representations is the feature space used to describe the tags, which varies over the three dimensions of the logsonomy.

Specifically, for $X \in \{U, T, R\}$ we consider the vector space $\mathbb{R}^X$, where each query word t is represented by a vector $\mathbf{v}_t \in \mathbb{R}^X$, as described below.

The *Tag Context Similarity* (TagCont) is computed in the vector space $\mathbb{R}^T$, where, for tag t, the entries of the vector $\mathbf{v}_t \in \mathbb{R}^T$ are defined by $v_{tt'} := w(t,t')$ for $t \neq t' \in T$, where w is the co-occurrence weight defined above, and $v_{tt} = 0$. The reason for giving a zero weight between a node and itself is that we want two tags to be considered related when they occur in a similar context, and not when they occur together.

The *Resource Context Similarity* (ResCont) is computed in the vector space $\mathbb{R}^R$. For a tag t, the vector $\mathbf{v}_t \in \mathbb{R}^R$ is constructed by counting how often a tag t is used to annotate a certain resource $r \in R$: $v_{tr} := \text{card}\{u \in U \mid (u,t,r) \in Y\}$.

The *User Context Similarity* (UserCont) is built similarly to ResCont, by swapping the roles of the sets R and U: For a tag t, the vector $\mathbf{v}_t \in \mathbb{R}^U$ is defined as $v_{tu} := \text{card}\{r \in R \mid (u,t,r) \in Y\}$.

In all three representations, we measure vector similarity by using the cosine measure, as is customary in Information Retrieval: If two tags t_1 and t_2 are represented by $\mathbf{v}_{t_1}, \mathbf{v}_{t_2} \in \mathbb{R}^X$, their cosine similarity is defined as:

$$\text{cossim}(t_1,t_2) := \cos \measuredangle(\mathbf{v}_{t_1}, \mathbf{v}_{t_2}) = \frac{\mathbf{v}_{t_1} \cdot \mathbf{v}_{t_2}}{||\mathbf{v}_{t_1}||_2 \cdot ||\mathbf{v}_{t_2}||_2}. \tag{4}$$

The *FolkRank algorithm* transfers the principle of the PageRank algorithm to folksonomies [8]. A resource which is tagged with important tags by important users becomes important itself. The same holds, symmetrically, for tags and users. In contrast to the web graph with its binary directed edges, a folksonomy consists of undirected triadic hyperedges. In order to propagate weights along those edges, the hypergraph is transformed into an undirected graph by splitting each triple (u,t,r) in Y into three undirected edges $\{u,t\}$, $\{u,r\}$, and $\{t,r\}$ in E. The PageRank is then applied in a differential approach: By modifying the weights for a given tag (users, resources) in the random surfer vector, FolkRank computes a ranked list of relevant tags (user, resource). Due to the undirectedness of the graph, a baseline without a specific preference vector is subtracted.

To compute a tag ranking in the logsonomy, we assigned high weights to a specific query term t in the random surfer vector. The final outcome of the FolkRank is then (among others) a ranked list of tags which FolkRank judges as related to t.

6.2 First Insights

Table 4 provides a few examples of the related tags returned by the measures under study. A first observation is that the co-occurrence relatedness seems to often "restore" compound expressions like *news channel*, *guitar tabs*, *brain tumor*. This can be attributed to the way how the logsonomy was constructed, namely by splitting queries (and consequently also compound expressions) using whitespace as delimiter. Another observation which is identical to the folksonomy data is that co-occurrence and FolkRank relatedness seem to often return the same related tags.

The tag context relatedness seems to yield substantially different tags. Our experience from folksonomy data (where this measure discovered preferentially synonym or sibling tags) seems to also prove true for logsonomy data: The most similar by tag context similarity often refers to a type of synonym[5] (e.g., *gun—guns*, *news—news.com*), whereas the remaining tags can be regarded as "siblings". For example, for the tag *brain* it gives other organs of the body, whereas for the tag *guitar* it gives other music instruments. When we talk about "siblings" we mean that these tags could be subsumed under a common parent in some suitable concept hierarchy; in this case, e.g., under *organs* and *music instruments*, respectively. In our folksonomy analysis, this effect was even stronger for the resource context relatedness—a finding which does not seem to hold for logsonomy data, based on this first inspection. The resource context relatedness does exhibit some similarity to the tag context relatedness, but gives in general a mixed picture. User context relatedness is even more blurred—the latter observation is again in line with the folksonomy side. These first observations suggest that despite the reported differences, especially the tag context in a logsonomy seems to hold a similar semantic information to the one we found in folksonomy data.

[5]Please note that we do not use the term 'synonym' in a linguistically precise way; we regard two words as being synonyms when they basically refer to the same concept. This also includes e.g., singular/plural forms of a noun.

Table 4 Examples of most related tags for each of the presented measures

Rank	Tag	Measure	1	2	3	4	5
37	News	*co-occurrence*	channel	daily	fox	paper	newport
		folkrank	channel	fox	daily	newspaper	county
		tag context	news.com	newspaper	weather	obituaries	newspapers
		resource context	news.com	arrested	killed	accident	local
		user context	county	center	edging	state	city
399	Guitar	*co-occurrence*	tabs	chords	tab	free	bass
		folkrank	tabs	chords	lyrics	tab	music
		tag context	banjo	drum	piano	acoustic	bass
		resource context	tabs	tab	tablature	chords	acoustic
		user context	chords	tabs	tab	guitars	chord
474	Gun	*co-occurrence*	smoking	paintball	parts	laws	control
		folkrank	guns	rifle	paintball	parts	sale
		tag context	guns	pistol	rifles	rifle	handgun
		resource context	smoking	pistol	rifle	handgun	guns
		user context	safes	guns	pistol	holsters	pellet
910	Brain	*co-occurrence*	tumor	stem	injury	symptoms	tumors
		folkrank	cancer	symptoms	tumor	blood	disease
		tag context	pancreas	intestinal	liver	thyroid	lungs
		resource context	tumor	tumors	syndrome	damage	complications
		user context	stem	feline	tumor	acute	urinary

6.3 Semantic Grounding

We further investigate this assumption by looking up the tags in an external structured dictionary of word meanings. Within these structured knowledge representations, there exist often well-defined metrics of semantic similarity; based on these, one can infer which type of *semantic* relation holds between the original and the related tags.

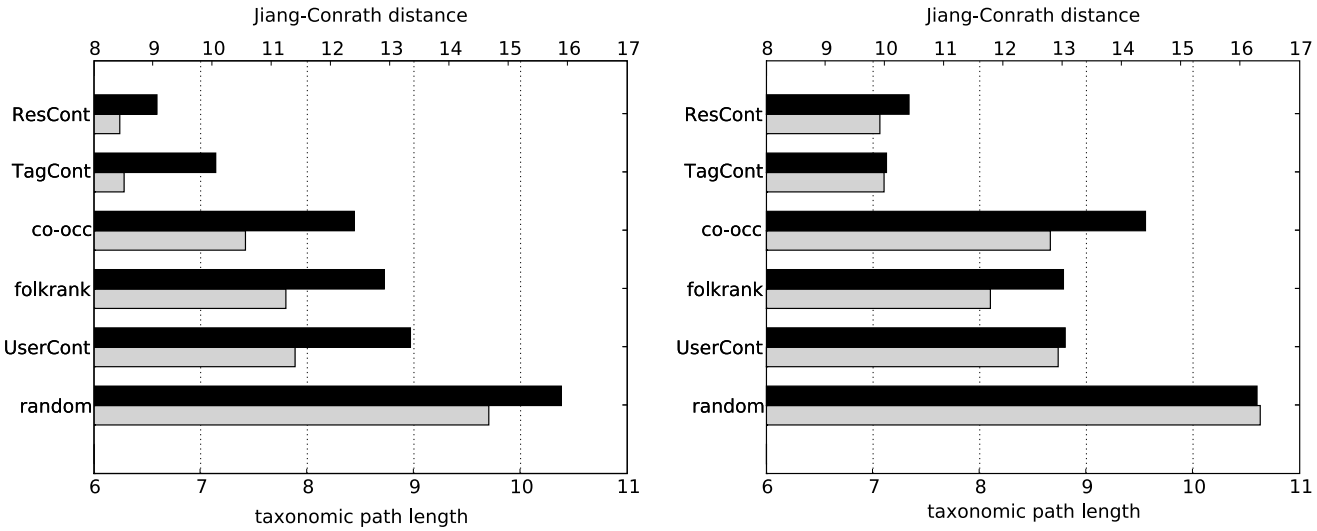
We use WordNet [6], a semantic lexicon of the English language. The core structure we exploit hereby is its built-in taxonomy of words, grouped into synsets, which represent distinct concepts. Each synset consists of one or more words, and is connected via the *is-a* relation to other synsets. The resulting directed acyclic graph connects *hyponyms* (more specific synsets) to *hypernyms* (more general synsets).

Based on this semantic graph structure, several metrics of semantic similarity have been proposed [3]. The most intuitive one is simply counting the number of nodes one has to traverse from one synset to another one. We adopted this *taxonomic shortest-path length* for our experiments. In addition, we use a measure of semantic distance introduced by [9] which combines the taxonomic path length with an information-theoretic similarity measure. The choice of this measure was guided by a work of [3], who showed by means of a user study that the Jiang-Conrath distance comes most closely to what humans perceive as semantically related.

Following the pattern proposed in [5], we carry out a first assessment of our measures of relatedness by measuring—in WordNet—the average semantic distance between a tag and the corresponding most closely related tag according to each of the relatedness measures under consideration. For each tag of our logsonomy, we find its most closely related tag using one of our measures; if we can map this pair to WordNet (i.e., if both tags are present), we measure the semantic distance between the two synsets containing these two tags. If any of the two tags occurs in more than one synset, we use the pair of synsets which minimizes the path length.

Figure 2 reports the average semantic distance between the original tag and the most related one, computed in WordNet by using both the taxonomic path length and the Jiang-Conrath distance. Overall, the diagrams are quite similar in respect to structure and scale. In both cases, the random relatedness (where we associated a given tag with a randomly chosen one) constitutes the worst case scenario.

Similar to our prior results for folksonomies (i.e., those shown in Fig. 2(a)), for the logsonomy the tag and resource context relatedness measures yield the semantically most closely related tags. In the logsonomy case, the distances of related tags for the context resource relatedness are longer



(a) Delicious folksonomy (b) AOL logsonomy

Fig. 2 Average semantic distance, measured in WordNet, from the original tag to the most closely related one. The distance is reported for each of the measures of tag similarity discussed in the main text (see Sect. 6.1). The corresponding labels are on the left. *Grey bars* (*bottom*) show the taxonomic path length in WordNet. *Black bars* (*top*) show the Jiang-Conrath measure of semantic distance

than in the folksonomy case. We attribute this to the way how the logsonomy is built: When users tag *implicitly* a certain URL by clicking on it, they are probably not as aware of the actual content of this page as a user who *explicitly* tags this URL in a social bookmarking system.

Another remarkable difference compared to the folksonomy data is that the co-occurrence relatedness yields tags whose meanings are comparatively distant from the one of the original tag. This can be attributed to the fact, that co-occurrence often "reconstructs" compound expressions as already mentioned in Sect. 6.2. The finding is a natural consequence of splitting queries and consequently splitting compound expressions as we did; our results confirm the intuitive assumption that the semantics of isolated parts of a compound expression usually are semantically complementary.

7 Conclusion

In this article, we presented a comparison of the contents, structure and semantics of folk- and logsonomies. We were able to discover both similar and diverging behavior in both kinds of systems. The analysis of the tag and query contents of both systems revealed, that the overlap of MSN query words with the set of Delicious tags is only about a quarter of the size of the latter, due to a very high number of very infrequent terms in both systems. Once the sets are reduced to the frequent terms, the relative overlap is higher. The remaining differences are due to different usage, e.g., to the composition of multi-word lexemes to single terms in Delicious, and the use of (parts of) URLs as query words in MSN.

We could show that both graph structures have small world properties in that they exhibit relatively small shortest path length and high clustering coefficients. Minor differences are triggered by the session IDs which do not have the same thematic overlap as user IDs have.

Further, logsonomies retain semantic information as has been demonstrated by the application of several relatedness measures. For example, the tag context measure yields semantically related tags, which recommends it as a candidate for synonym and sibling term identification. The resource context measure is much less concise in finding related terms compared to a folksonomy. This can be contributed to incomplete knowledge about the content of a result page users click on.

Overall, these findings support the idea that social algorithms can also be leveraged for search. As query logs are a natural product of search, no specific efforts are necessary to gather the required information, which makes the data collection time- and cost-efficient. Privacy concerns, however, should be considered when accumulating and processing (to some extend personal) query data.

In future work, we want to enhance IR methods with logsonomy data and evaluate their performance in comparison to state-of-the-art approaches. Further, we want to merge folk- and logsonomies to benefit from the strength and weaknesses of both worlds.

Acknowledgements This research was funded by the European Commission in the project "Tagora—Emergent Semiotics in Social

Online Communities", by the Microsoft Grant "Social Search" and by DFG in the project "Info 2.0—Informationelle Selbstbestimmung im Web 2.0".

References

1. Baeza-Yates R, Tiberi A (2007) Extracting semantic relations from query logs. In: KDD '07: proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 76–85. ISBN 9781595936097. doi: 10.1145/1281192.1281204
2. Benz D, Krause B, Praveen Kumar G, Hotho A, Stumme G (2009) Characterizing semantic relatedness of search query terms. In: Proceedings of the 1st workshop on explorative analytics of information networks (EIN2009), Bled, Slovenia, September 2009
3. Budanitsky A, Hirst G (2006) Evaluating wordnet-based measures of lexical semantic relatedness. Comput Linguist 32(1):13–47
4. Cattuto C, Schmitz C, Baldassarri A, Servedio VDP, Loreto V, Hotho A, Grahl M, Stumme G (2007) Network properties of folksonomies. AI Commun 20(4):245–262
5. Cattuto C, Benz D, Hotho A, Stumme G (2008) Semantic grounding of tag relatedness in social bookmarking systems. In: The semantic web—ISWC 2008. Springer, Berlin, pp 615–631. ISSN 0302-9743
6. Fellbaum C (ed) (1998) WordNet: an electronic lexical database. MIT, Cambridge
7. Halpin H, Robu V, Shepard H (2006) The dynamics and semantics of collaborative tagging. In: Proceedings of the 1st semantic authoring and annotation workshop (SAAW'06), pp 211–220
8. Hotho A, Jäschke R, Schmitz C, Stumme G (2006) Information retrieval in folksonomies: search and ranking. In: Sure Y, Domingue J (eds) The semantic web: research and applications. Lecture notes in computer science, vol 4011. Springer, Heidelberg, pp 411–426
9. Jiang JJ, Conrath DW (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the international conference on research in computational linguistics (ROCLING), pp 19–33. Taiwan
10. Jäschke R, Krause B, Hotho A, Stumme G (2008) Logsonomy—a search engine folksonomy. In: Proceedings of the second international conference on weblogs and social media (ICWSM 2008). AAAI Press, Menlo Park, pp 192–193
11. Krause B, Hotho A, Stumme G (2008) A comparison of social bookmarking with traditional search. In: Macdonald C, Ounis I, Plachouras V, Ruthven I., White RW (eds) Advances in information retrieval, 30th European conference on IR research, ECIR 2008, pp 101–113
12. Krause B, Jäschke R, Hotho A, Stumme G (2008) Logsonomy—social information retrieval with logdata. In: HT '08: proceedings of the nineteenth ACM conference on hypertext and hypermedia. ACM, New York, pp 157–166. ISBN 978-1-59593-985-2. doi: 10.1145/1379092.1379123
13. Pass G, Chowdhury A, Torgeson C (2006) A picture of search. In: Proceedings of the 1st international conference on scalable information systems. ACM, New York, p 1
14. Watts DJ, Strogatz S (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442