Characterizing Semantic Relatedness of Search Query Terms

Dominik Benz, Beate Krause, G. Praveen Kumar,
Andreas Hotho, and Gerd Stumme

Knowledge & Data Engineering Group, University of Kassel,
34121 Kassel, Germany
{benz,krause,praveen,hotho,stumme}@cs.uni-kassel.de

Abstract. Mining for semantic information in search engine query logs bears great potential for both the optimization of search engines and bootstrapping Semantic Web applications. The interaction of a user with a search engine (more specifically clicklog information) has recently been viewed as *implicit tagging* of resources by query terms. The resulting structure – previously called a *logsonomy* – exhibits structural similarities to folksonomies, which evolve during the *expclicit* process of annotating resources with freely chosen keywords in social bookmarking systems. For the folksonomy case, appropriate measures of relatedness have shown to be capable to harvest the emerging semantics inherent in the tripartite graph of users, tags and resources. Motivated by the reported structural similarities, in this work we extend this methodology to logsonomies. More specifically, we apply several measures of query term relatedness to the logsonomy graph and provide a semantic characterization for each measure by grounding it against user-validated relatedness measures based on WordNet. Comparing the outcome with prior results of analyzing folksonomy data we find that the formalization of log data in logsonomies retains the semantic information. Some relatedness measures we applied prove to be able to capture these emergent semantics similarly to the folksonomy case, while others exhibit different characteristics. In this way we provide a novel and systematic approach to compare the emergent semantics of user interactions with search engines and social bookmarking systems. We conclude that the *type* of semantic information inherent in both emerging structures is similar, and inform the choice of an appropriate measure of query term relatedness for a given task.

1 Introduction

Folksonomies are complex systems consisting of user-defined labels added to web content such as bookmarks, videos or photographs by different users. In contrast to classical search engines, which index the web and offer a simple user interface to search in this index, a folksonomy can be explored in different dimensions taking users, tags and resources into account. With logfiles containing queries and clicks of search engine users, a similar relation between users, query terms and a resource can be found: a user submits a query and clicks on a specific URL. The resulting structure of this process, previously called *logsonomy* [1], is a tripartite graph of a set of users, queries and clicked URLs with hyperedges, each connecting one query, one clicked URL and one specific user.

While folksonomies aggregate *explicit* lightweight metadata annotations (namely tags), logsonomies can hence be seen as *implicit* annotations by user clicks. Previous work [1] revealed that both show similar structural characteristics, e. g., small world properties, a power law distribution of tags and users, and a similar co–occurrence behaviour of tags. These first insights indicate the possibility that logsonomies contain – similar to the folksonomy graph – inherent semantics emerging from the "collaborative" process of searching similar information and being interested in the same resources.

In prior work, we found that measures of semantic relatedness based on the folksonomy graph are able to extract these *emerging semantics* [2]. Motivated by the structural analogies mentioned above, we explore in this paper the potential of logsonomies to extract semantic relations between different queries or query parts. In [2], different relatedness measures considering statistical, distributional and structural characteristics of a folksonomy have been applied to learn about related tags on a large-scale snapshot of the social bookmarking system del.icio.us. These measures are the co-occurrence count, three distributional measures which use the cosine similarity in the vector spaces spanned by users, tags and resources, respectively, and FolkRank [3], a graph-based measure which is an adaptation of the well-known PageRank [4] to folksonomies.

We will apply these measures in the same manner to a logsonomy built from an AOL click dataset. This allows for a direct comparison of the findings of tag relatedness in folksonomies to the ones in logsonomies. Especially, the semantic grounding based on a comparison of related tags / query terms to semantic relations of terms in the lexical database WordNet helps to characterize the major differences of relatedness measures between folk- and logsonomies. We follow the choice of [2] and measure the semantic term relatedness within WordNet by using both the taxonomic path length and a similarity measure by Jiang and Conrath [5]. The latter resembles most closely to what humans perceive as semantically related [6], while the first allows the inspection of the edge composition of paths leading from one tag to the corresponding related tags, which has proven to be especially insightful.

Learning about the hidden structure of a search engine's query vocabulary will be interesting for a variety of applications: similar or related tags can be used to refine and expand search queries, to correct spelling errors or to improve a search engine's ranking. For example, first results show that the tag context relatedness is often able to extract synonyms of a given query term. Other measures (like FolkRank) seem to point to more general terms, which can be useful for broadening a search. Another application of our work is harvesting the usage-driven semantics of search query logs by ontology learning procedures based on logsonomies. This would directly tackle the knowledge acquisition problem of many Semantic Web applications. Our work establishes hereby the connection between prior work (like [7] on mining for semantics in search query logs) and recent approaches of bridging the gap between the "Web 2.0" and the Semantic Web.

In this paper, we bring together two research branches: First, work targeted on harvesting semantic information from social annotations (i. e., by appropriate measures of semantic relatedness), and second work on structural similarities between interactions of users with social bookmarking systems and search engines. We expect synergies for both directions by posing the following research questions:

- Is there evidence for emergent semantics in logsonomies as it is in folksonomies?
- Can we infer further structural similarities or differences between folksonomies and logsonomies by comparing the output of relatedness measures on both structures?
- Does a given measure of relatedness exhibit the same semantic characteristics when applied to a folksonomy and to a logsonomy graph?

The rest of the paper is organised as follows. Section 3 briefly defines folk- and logsonomies, describes the construction of logsonomies and the datasets used for computing tag and query term relatedness. In Section 4, we introduce the applied relatedness measures. Some qualitative insights are presented in Section 5, while a thorough analysis of query term relatedness is conducted in Section 6. Finally, implications of this work to other fields are discussed and an outlook on future work given.

2 Related Work

In this section, we discuss related work which considers the analysis of semantics in query click logs. To the best of our knowledge, a comprehensive comparison of the semantics extracted from folksonomies and search engine logs has not been conducted before, but each data structure has been considered individually.

Besides a variety of analytical studies about the nature of clickdata, extensive reseach has been conducted in the area of information retrieval where click data was used to expand queries and to improve the retrieval performance. For a detailed discussion of folksonomies, social bookmarking systems and the extraction of semantics the reader may refer to [2].

The transformation of clickdata to a tripartite hypergraph as well as a comparison to folksonomy properties has been carried out in [1]. The work could show similar structural properties of logsonomies and folksonomies. Given these results, the analysis of query term similarity seems to be very promising.

A further consideration of the tripartite structure of query logs has been presented in [8], where an algorithm to rank resources based on the relationships among users, queries and resources was proposed. In [7], Baeza-Yates and Tiberi proposed to present query-logs as an implicit folksonomy where queries can be seen as tags associated to documents clicked by people making those queries. The authors extracted semantic relations between queries from a query-click bipartite graph where nodes are queries and an edge between nodes exists when at least one equal URL has been clicked after submitting the queries. As an extension of the above work, Francisco et al. [9] cluster these bipartite graphs using clique percolation and priori induced cliques and consequently extract semantic relations between queries. However, the semantic grounding of the relations is not in the core of this work. By constructing a folksonomy-alike structure, we can build on systematic investigations of various topological characteristics of the well-established folksonomy model. Our tag-tag–co-occurrence analysis is closely related to the graph analysis of [9], but operates on a different kind of dataset created by splitting the original search queries into single search terms. This dataset is compared to results of the del.icio.us folksonomy. Overall, our contribution is a comparison between hypergraphs constructed of real-world folksonomy and logsonomy datasets. We are not aware of other work which examines differences and commonalities of user interactions with folksonomy and search engine systems as we do in this paper.

3 Folksonomies and Logsonomies

As mentioned above, social bookmarking systems contain *explicit* annotations while search engine clicklogs provide *implicit* annotations of resources. In this section, we provide a formal model of folksonomies and show how search engine clicklogs can be adapted to this model, resulting in *logsonomies.* We consequently detail on the datasets used to evaluate our approach.

3.1 Formal Model of a Folksonomy

The central data structure of a social bookmarking system is called *folksonomy.* It can be seen as a lightweight classification structure which is built from *tag* annotations (i. e., freely chosen keywords) added by different users to their resources. A folksonomy consists thus of a set of users, a set of tags, and a set of resources, together with a ternary relation between them.

Following [3], we formally define a *folksonomy* as 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).

For convenience we also define, for all $u \in U$ and $r \in R$, $\textsc{tags}(u, r) := \{t \in T \mid (u, t, r) \in Y\}$, i. e., $\textsc{tags}(u, r)$ is the set of all tags that user u has assigned to resource r. The set of all *posts* of the folksonomy is $P := \{(u, S, r) \mid u \in U, r \in R, S = \textsc{tags}(u, r), S \neq \emptyset\}$. Thus, each *post* consists of a user, a resource and all tags that the user has assigned to the resource.

3.2 Adaptation to Search Engine Query Logs: Logsonomies

In order to apply our established folksonomy analysis techniques [2] to search engine logs, we need similar structures on both sides. We adhere to the approach described in [1] and transform a search engine log into a folksonomy alike structure, called *logsonomy.* User IDs represent the users of a folksonomy,[1] and the *clicked* URLs represent resources. The latter are implicitly annotated by the given

[1] In datasets other than the AOL data at hand, one may need to switch to session IDs, if there are no explicit user IDs in the log files.

query; in order to mimic most closely social annotations, we split composed queries into single words. These *query terms* are thus all substrings of a query that are separated by whitespaces. They correspond to the tags in a folksonomy. For sake of simplicity, we will also use the term "tag" when addressing "query words" in the remainder of the paper. This means when we talk about relatedness of "tags" in a logsonomy, we do talk about relatedness of query words. The decision to use the splitted queries instead of complete ones was motivated by the findings of [1] that the resulting network structure comes closer to an actual folksonomy.

More formally, this transformation of a search engine log to a logsonomy can be described as follows:

- Let U be the set of *users* of the search engine.
- T be the set of *query terms* contained in the queries the users gave to the search engine,
- R be the set of *URLs* which have been clicked by the search engine users.

We add a tuple (u, t, r) to Y whenever user u clicked on resource r of a result set after having submitted the query term t (eventually with other terms). The resulting relation $Y \subseteq U \times T \times R$ corresponds to the tag assignments in a folksonomy.

The process of creating a logsonomy shows similarities to the creation of a folksonomy. Users describe an information need by means of a query. They then restrict the result set of the search engine by clicking on those URLs whose snippets indicate that the web page has some relation to the query. These query/click combinations result in the logsonomy. However, one needs to keep some major differences in mind, when applying folksonomy techniques to logsonomies:

- Users have a bias towards clicking the top URLs of a result list. In query log analysis, these clicks are usually discounted.
- 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. We nevertheless assume that the act of clicking already indicates an association between query and resource in the logsonomy, since the log data under study did not contain any explicit user feedback (which could have been used for further differentiation).
- 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 which do not publish the techniques applied. One does not know to which extent the query terms serve as a description of 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.

3.3 Datasets

In order to make our results comparable to prior work on semantic relatedness measures on folksonomies, we detail here on the social bookmarking data which was

Table 1. Folksonomy and Logsonomy Datasets

dataset	$\|T\|$	$\|U\|$	$\|R\|$	$\|Y\|$
Del.icio.us	10,000	476,378	12,660,470	101,491,722
AOL split queries	10,000	463,380	1,284,724	26,227,550

the basis of [2] before describing the click data we used to build the logsonomies. Table 1 summarizes some statistics about both datasets.

Social Bookmarking Data. For our experiments, we used data from the social bookmarking system del.icio.us, collected in November 2006. In total, data from $667,128$ users of the del.icio.us community were collected, comprising $2,454,546$ tags, $18,782,132$ resources, and $140,333,714$ tag assignments. As one main focus of this work is to characterize tags by their properties of co–occurrence with other tags, we restricted our dataset to the $10,000$ most frequent tags of del.icio.us, and to the resources/users that have been associated with at least one of those tags. One could argue that tags with low frequency have a higher information content in principle — but their inherent sparseness makes them less useful for the study of both co-occurrence and distributional measures. The size of the restricted folksonomy is shown in Table 1.

Click Data. We used a click dataset from the AOL search engine. The data was collected from March, 1st to May, 31st 2006. The original dataset consists of 657,426 unique user IDs, 10,154,742 unique queries, and 19,442,629 click-through events [10]. We constructed the logsonomy as described in Section 3.2. Again, we apply the restriction of only using the 10,000 most frequent tags (i. e., query words) to the dataset. The resulting sizes are shown in Table 1. Since the AOL data was only available with truncated URLs, we reduced the URLs to host-only URLs, i. e., we removed the path of each URL leaving only the host name.

4 Measures of Relatedness

The underlying structure of a logsonomy can — analogously to a folksonomy — also be regarded as an undirected tri-partite hyper-graph (see section 3.2). As measures of similarity and relatedness are not well-developed for this kind of data yet, we follow the approach of [2] and stick to two- and one-mode views on the data. These views are complemented by a graph-based approach for discovering related tags (FolkRank), which makes direct use of the three-mode structure. Please note that the remaining paragraphs of this section summarize the measures of relatedness used in our prior work [2]; we include their description in order to explain their adaption to logsonomies. For a detailed description of the computational complexity of each measure, we refer to [2].

Co-Occurrence. Given a logsonomy (U, T, R, Y), we define the *query word co-occurrence graph* as a weighted undirected graph whose set of vertices is the set T of query words. 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.,

$$w(t_1, t_2) := \text{card}\{(u, r) \in U \times R \mid t_1, t_2 \in T_{ur}\} \ . \tag{1}$$

Co-occurrence relatedness 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 will denote this co-occurrence relatedness by *co-occ*.

Distributional Measures. 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 – and thus between the measures – is the feature space used to describe the tags, which varies over the possible 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 $\boldsymbol{v}_t \in \mathbb{R}^X$, as described below.

Tag Context Similarity. The Tag Context Similarity (TagCont) is computed in the vector space $\mathbb{R}^T$, where, for tag t, the entries of the vector $\boldsymbol{v}_t \in \mathbb{R}^T$ 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 weight zero 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.

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

User Context Similarity. 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 $\boldsymbol{v}_t \in \mathbb{R}^U$ 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 [11]: If two tags t_1 and t_2 are represented by $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^X$, their cosine similarity is defined as: $\text{cossim}(t_1, t_2) := \cos\measuredangle(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{\boldsymbol{v}_1 \cdot \boldsymbol{v}_2}{||\boldsymbol{v}_1||_2 \cdot ||\boldsymbol{v}_2||_2}$.

FolkRank. FolkRank employs the principle of the PageRank algorithm [12] for folksonomies [3]: a resource which is tagged with important tags by important users becomes important itself. The same holds, symmetrically, for tags and users. By modifying the weights for a given tag in the random surfer vector, FolkRank can compute a ranked list of relevant tags.

To apply the FolkRank to a 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. Refer to [2] for a more detailed description of the experimental procedure and to [3] for a detailed description of the FolkRank algorithm.

Table 2. Examples of most related tags for each of the presented measures.

rank	tag	measure	1	2	3	4	5
5	lyrics	*co-occurrence*	song	love	music	day	songs
		folkrank	song	love	music	myspace	songs
		tag context	titles	listen	lyric	called	theme
		resource context	wanna	lyric	gonna	ya	goodbye
		user context	song	music	songs	school	center
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
4764	vest	*co-occurrence*	herb	life	patterns	hd	pattern
		folkrank	herb	motorcycle	vests	patterns	shooting
		tag context	vests	sweaters	jacket	sweater	knit
		resource context	jacket	set	bag	shorts	stainless
		user context	herb	hd	vests	sec	lawsuits

5 Qualitative Insights

A first natural question that arises when trying to compare tag relatedness measures on both logsonomies and folksonomies is to which extent both vocabularies overlap. We found that 4,451 out 10,000 tags[2] were present in both datasets (i. e., roughly 44%). Looking up these tags in an English dictionary showed that 92% of them are proper English words; this confirms the intuition that the vocabulary used for tagging and searching is substantially different, but has an overlap of "generic" terms. Figure 1 plots the tag rank for each overlapping tag in the folksonomy (del.icio.us) against its rank in the logsonomy (AOL). Please note that a low tag rank corresponds to a high usage frequency. One can see that high-frequency (i. e., low-rank) tags in the folksonomy tend to be frequently used in a logsonomy as well (roughly the top 1,000 tags); apart from this, there seems to be no special correlation between the usage frequency of tags in both datasets.

Following the methodology of [2], our next step was to compute, for each of the 10,000 most frequent tags of the AOL log, its most closely related tags using each of the measures described above.

Table 2 provides a few examples of the related tags returned by the measures under study. A first observation is that the cooccurrence relatedness seems to

[2] Please recollect that we use for sake of simplicity the term "tag" to subsume tags in a folksonomy and query words in a logsonomy.

Table 3. Overlap between the 10 most closely related tags.

	co-occurrence	*FolkRank*	*tag context*	*resource context*
user context	*2.28*	*2.16*	0.71	1.11
resource context	1.93	*2.25*	1.5	
tag context	0.88	1.1		
FolkRank	**5.91**			

often "restore" compound expressions like *news channel, guitar tabs, brain tumor*. This can be naturally attributed to the way how the logsonomy was constructed, namely by splitting queries (and consequently also compound expressions) using whitespace as delimiter. We could observe this behaviour also partially in our last study on folksonomy data (see [2]), however to a much lesser extent. Another observation which is identical to the folksonomy data is that cooccurrence 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 related by tag context relatedness is often a synonym (e. g., *gun* – *guns*, *vest* – *vests*), 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.

Our next systematic step is to check whether the most closely related tags are shared across the measures of relatedness. We consider the 10,000 most popular tags in AOL, and for each of them we compute the ten most related tags according to each of the relatedness measures. Table 3 reports the average number of shared tags for the relatedness measure we investigate. The results are again very close to our folksonomy analysis – in general, there is a rather small overlap between the lists of the 10 most related tags (between 0.71 and 2.28) – with an exception of almost six shared tags in average between cooccurrence and FolkRank. This supports our assumption that the computation of FolkRank on a logsonomy also tends to be dominated by the tag-tag cooccurrence network.

To better investigate this point, for each of the 10,000 most frequent tags in the AOL log, we computed the average rank (according to global frequency) of its ten most closely related tags, according to each of the relatedness measures under study. The results are shown in Figure 2, along with the folksonomy results

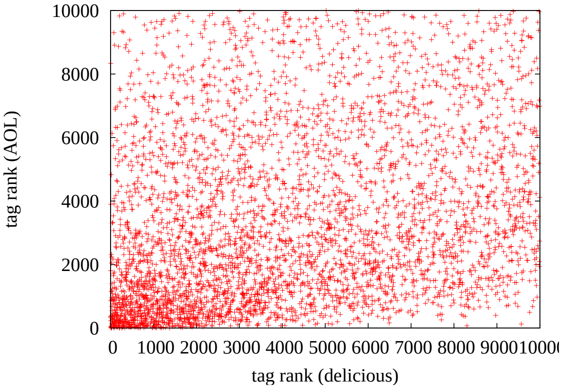
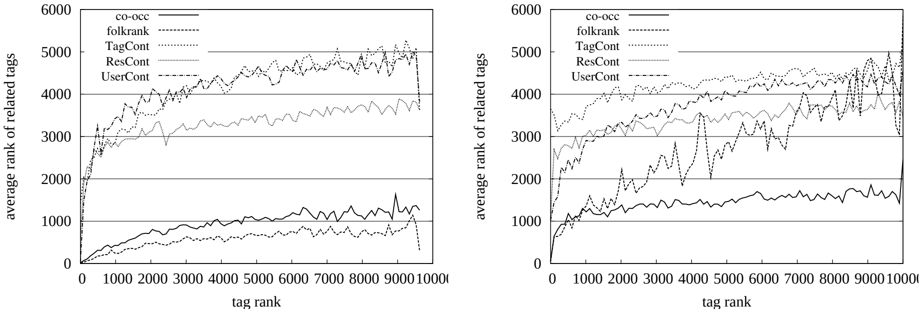


Fig. 1. Correlation between the tag ranks in the AOL logsonomy and the delicious folksonomy.

for comparison. One can see that co-occurrence, resource and user context relatedness show almost identical behaviour compared to folksonomies: Especially the co-occurrence relatedness does have the same strong bias towards high-frequency (i. e., low-rank) tags, independently of the frequency of the original tag. This effect is not so strong for the context measures; however, the distribution for tag context relatedness shows a significant difference: For very popular tags (roughly the top 2000 tags), its most related tags are comparatively rarely used ones (with a tag rank around ∼3500). We hypothesize that this could reflect the topical diversity of both datasets: Because the folksonomy is dominated by technophile topics, its most popular tags are probable to fall into that category. This implies especially that "sibling" tags of a "technical" tag are probably also used frequently. If the topical diversity among the popular tags in the logsonomy is higher, then the sibling tags are more probable to point towards less frequently used tags – which is what we observe here. Another remarkable difference is the behaviour of FolkRank; despite its high overlap with the co-occurrence relatedness reported in Table 3, their profiles differ significantly. The strongly peaked plot of folkrank suggests that there might exist some "outliers" – i. e., very infrequent tags – besides the overlap with the co-occurrence relatedness.

The last question we asked ourselves in this first step is to which extent a given measure returns the same tags when applied to a logsonomy and a folksonomy. To this end we restricted the examination to the overlapping 4451 tags (i. e., for each tag in the overlap, we computed its ten most closely related tags by all measures, whereby the most related tags had to be in the overlap again). Interestingly, we did not find a significant overlap between any two measures (the overlap values ranged from 0.5 to 2.3). For this reason we skip the inclusion of the complete overlap table, as it does not provide additional information. We take this as an indicator that – despite the similarities reported above – the actual semantics contained in folksonomies and logsonomies differ.



(a) folksonomy (b) logsonomy

Fig. 2. Average rank of the related tags as a function of the rank of the original tag.

Table 4. WordNet coverage of logsonomy tags.

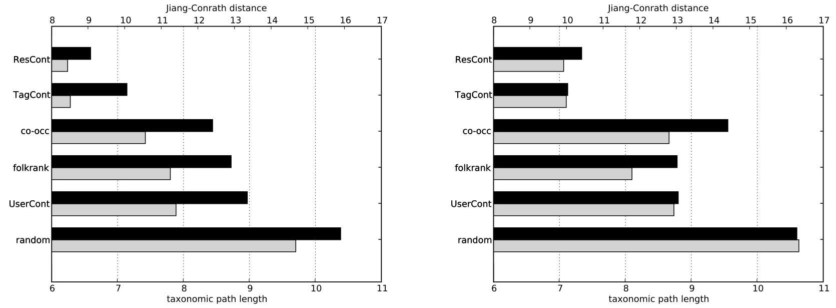
# top-frequency tags	100	500	1,000	5,000	10,000
fraction in WordNet (AOL split query logsonomy)	95 %	96 %	94 %	88 %	81 %
fraction in WordNet (del.icio.us folksonomy)	82 %	80 %	79 %	69 %	61 %

6 Semantic Analysis

Following the qualitative analysis of the previous section, we now go one step further to a more formally grounded characterization of the measures under consideration. In our previous work [2], we introduced the notion of *Semantic Grounding*: The basic idea hereby is to ground the relations between the original and the related tags 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 follow this approach and use WordNet [13], 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. 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 Jiang and Conrath [5] which combines the taxonomic path length with an information-theoretic similarity measure by Resnik [14]. The choice of this measure was guided by a work of Budanitsky and Hirst [6], who showed by means of a user study that the Jiang-Conrath distance comes most closely to what humans perceive as semantically related. We use the implementation of those measures available in the `WordNet::Similarity` library [15].



(a) folksonomy (b) logsonomy

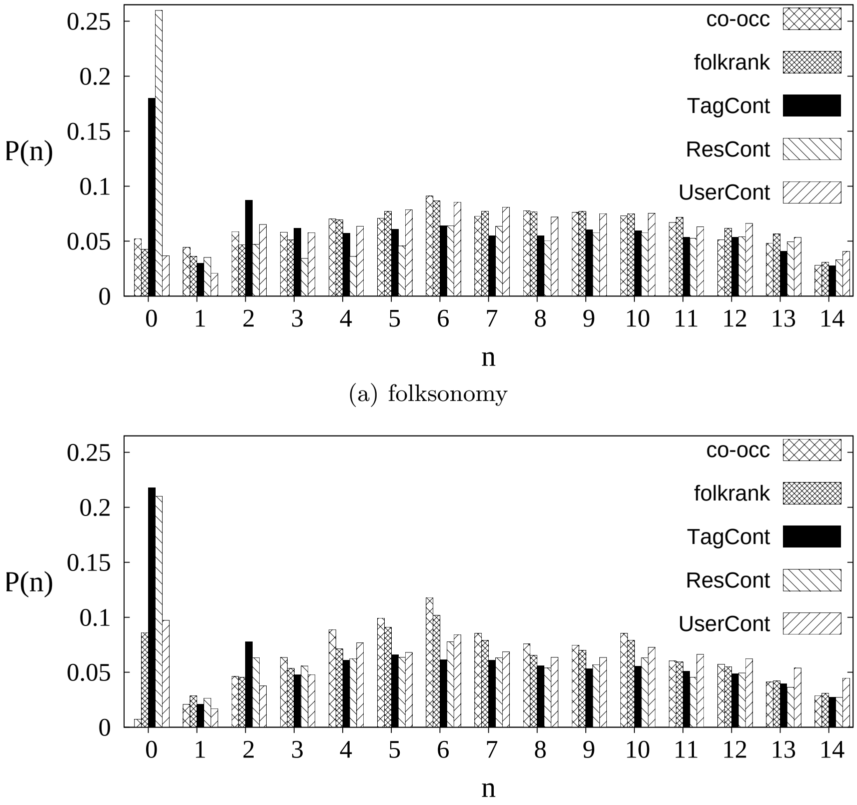
Fig. 3. Average semantic distance, measured in WordNet, from the original tag to the most closely related one. The distance is reported for each of the measures of tag similarity discussed in the main text (labels on the left). Grey bars (bottom) show the taxonomic path length in WordNet. Black bars (top) show the Jiang-Conrath measure of semantic distance.

A natural prerequisite for the semantic grounding described above is that a significant fraction of the most popular tags in the logsonomy (in other words, the most popular search query terms) is present in WordNet. Despite some limiting factors (different languages, misspellings, queries for names of persons or things, . . .), Table 4 shows a relatively high overlap – 81 % of the 10,000 most popular search terms are in fact proper English words. This is significantly more than in our previous work with a snapshot of the del.icio.us folksonomy, which is probably due to the more idiosyncratic nature of folksonomy tags. The higher overlap puts the following grounding process on an even more solid basis.

Following the pattern proposed in [2], 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 containg 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 3 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 Figure 3a), for the logsonomy the tag and resource context relatedness measures yield the semantically most closely related tags. However, in the logsonomy case, the context resource relatedness could not repeat the superior performance it showed for the folksonomy. We attribute this to the way how the logsonomy is built: When users



(a) folksonomy

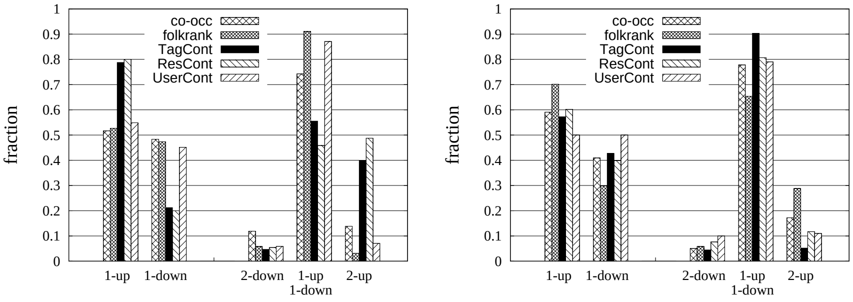
(b) logsonomy

Fig. 4. Probability distribution for the lengths of the shortest path leading from the original tag to the most closely related one. Path lengths are computed using the subsumption hierarchy in WordNet.

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. A further examination (see section 5) revealed that co-occurrence often "reconstructs" compound expressions; e. g., the most related tag to *power* according to co-occurrence relatedness is *point*. This is a natural consequence of splitting queries and consequently splitting compound expressions as we did; so our results confirm the intuitive assumption that the semantics of isolated parts of a compound expression usually are semantically complementary.

Figure 3b shows – as in the folksonomy case – that the analysis of the semantic measures for the logsonomy data yields basically the same results with the path length as with the Jiang-Conrath measure. Therefore, we will stick to the simpler-to-understand path length in the sequel.



(a) folksonomy

(b) logsonomy

Fig. 5. Edge composition of the shortest paths of length 1 (left) and 2 (right). An "up" edge leads to a hypernym, while a "down" edge leads to a hyponym.

A more fine-grained analysis of the nature of related tags can be obtained by analyzing the shortest paths that lead from a given tag to its most closely related tag (according to our measures under consideration). Figure 4 displays the normalized distribution $P(n)$ of shortest-path lengths n (number of edges) connecting a tag to its closest related tag in WordNet. The most obvious analogy to [2] is the strong peak of both tag and resource context relatedness at a path length of 0. Paths of length 0 reveal a synonym relation in WordNet, which means that the two query words appear in the same synset in WordNet. Interestingly, the co-occurrence measure shows very low probability of finding synonyms, but is highest when it comes to longer path lengths (for example $n = 6$). This again is in line with our assumption, that the co-occurrence relatedness finds compound expressions which appear to have longer path lengths within WordNet.

For both the logsonomy and the folksonomy and for all measures under study, a path length of 1 occurs very infrequently. This indicates that none of the measures frequently returns direct hypernyms or direct hyponyms. The contrast between high and low probabilities for the path lengths of 0 and 1 can be attributed to the fact, that a path length of 1 leads to either a hypernym or a hyponym in the WordNet hierarchy, never to a sibling. None of the applied measures reveal such a hierarchical relation.

Next, we focus on the shortest path lengths of $n = 1$ and $n = 2$ in the two datasets, e. g., the potential hypernym/hyponym and sibling relations. For $n = 2$ (right-hand side of the subfigures in Figure 5), all measures show – both for folksonomies and logsonomies – a prevalent peak for siblings (1-up/1-down, corresponding to a hypernym edge (up) and a hyponym edge (down)). This observation especially holds for the tag, resource and user context measures. Surprisingly, in the logsonomy, this also holds (though with a lower probability) for the co-occurrence relatedness – in contrast to the folksonomy case. Considering the process of search, some users probably tend to describe their information need with "sibling" query terms like *microwave oven* or *black white*. When interpreting these results, one also has to keep in mind that the absolute number of 1-up-1-down pairs is much larger for tag and resource context relatedness (424 / 367) compared to the other three

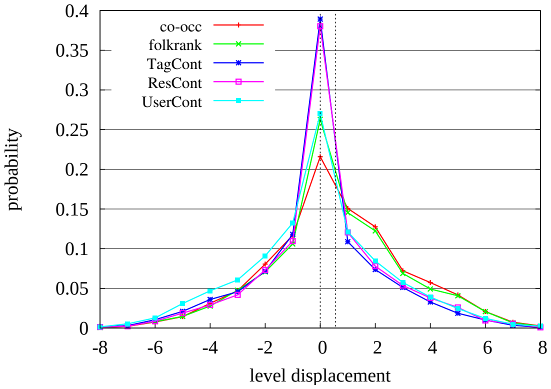


Fig. 6. Probability distribution of the level displacement Δl in the WordNet hierarchy.

measures (279 / 257 / 200 for co-occurrence, FolkRank and user context, respectively). For paths with $n = 1$, we observe – except of the user context relatedness – a slight preference towards hypernym edges (e. g., one level up in the WordNet taxonomy). This finding is the strongest for the FolkRank, but also the other three measures show a slight tendency to reveal hypernyms rather than hyponyms. Especially for the co-occurrence relatedness this behaviour is rather different from our results on folksonomies; again we think that one can see this as another indicator that co-occurrence relatedness restores compound terms in the logsonomy case.

When we generalize the analysis of Figure 5 to paths of arbitrary length, however, the slight tendency towards hypernym edges for the context relatedness measure vanishes. Figure 6 displays the *hierarchical displacement* Δl, i. e., the difference in hierarchical depth between the synset where the path starts and the synset where the path ends. Δl is the difference between the number of edges towards a hypernym (up) and the number of edges towards a hyponym (down). We do not include the folksonomy data for comparison here because it is nearly identical (see [2]). In both cases, we observe a strong peak at $\Delta l = 0$ for all context relatedness measures, which means that the measures do not imply a systematic bias towards more general or more specific terms. The average value of Δl for all the contextual measures is $\overline{\Delta l} \simeq 0$ (dotted line at $\Delta l = 0$). The probability distributions for both co-occurrence and folkrank relatedness are less symmetric and have both an average of $\overline{\Delta l} \simeq 0.55$ (right-hand dotted line). This means that for these measures – as we have already observed – the related tags lie preferentially higher in the WordNet hierarchy.

7 Conclusions and Outlook

In this paper, we investigated emergent semantics in search engine logs by means of term relatedness measures that have been shown to reveal semantic information inherent in the folksonomy graph [2]. We built our analysis on a folksonomy-like representation of the logdata, namely on *logsonomies*, because prior work showed promising structural similarities between log- and folksonomies [1]. This approach allowed us to directly compare log- and folksonomies in respect to the kinds of semantics captured by the different relatedness measures. In order to provide a semantic grounding of the measures under study, we used WordNet and well-established measures of semantic relatedness.

Resuming the research questions stated in the introduction of this paper, we can summarize our contributions as follows:

Emergent semantics in Logsonomies: The presence of inherent semantics in query log data has been reported before. Our contribution is to show that – despite some differences – the formalization of log data into logsonomies retains the semantic information and facilitates the application of established folksonomy analysis techniques to capture the semantics. However, the process of logsonomy construction seems to play an important role: In our case (i. e., when tags are created by splitting the original queries), the co-occurrence relatedness tends to restore compound expressions contained in the original queries.

Comparison of semantics in logsonomies and folksonomies: Our presented approach allows for a direct comparison of the semantics emerging from *explicit* tagging in social bookmarking systems and *implicit* tagging by clicking on search engine results. The results demonstrate that the *type* of inherent semantic information is similar in both cases, but the actual *instances* seem to vary. In other words: Both structures allow mining for synonym and sibling terms, but the *actual* synonyms and siblings retrieved *for a given term* differ.

Characteristics of relatedness measures: Interestingly, applying the resource context relatedness to logsonomies is much less precise in discovering semantically close terms, compared to a folksonomy. We attribute this mainly to incomplete user knowledge about the content of a result page they click on, leading e. g., to "erroneous" clicks. The behaviour of the tag context measure is more similar to the folksonomy case, which recommends it as a candidate for synonym and "sibling" term identification. Additionally, the semantics of the co-occurrence relatedness is strongly influenced by the process of constructing the logsonomy.

In general, we think that our work can help to model the semantic implications of user interactions with search engines. Ultimately, a deeper understanding of this will facilitate the improvement of search engines (e. g., via query expansion) on the one hand, and the harvesting of ontologies for Semantic Web applications on the other hand. We are currently working on voting approaches for combining several measures of relatedness in order to separate even more clearly e. g., synonym and sibling terms. Another promising research direction is to further characterize and understand the structural differences between logsonomies and folksonomies which are responsible for the different behaviour of some relatedness measures.

Acknowledgement. This research was partially funded by the European Commission in the project "Tagora – Emergent Semiotics in Social Online Communities" and by DFG in the project "Info 2.0 – Informationelle Selbstbestimmung im Web 2.0".

References

1. Krause, B., Jäschke, R., Hotho, A., Stumme, G.: Logsonomy - social information retrieval with logdata. In: HT '08: Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia, New York, NY, USA, ACM (2008) 157–166
2. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic grounding of tag relatedness in social bookmarking systems. The Semantic Web - ISWC 2008 (2008) 615–631
3. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: Search and ranking. In Sure, Y., Domingue, J., eds.: The Semantic Web: Research and Applications. Volume 4011 of LNAI., Heidelberg, Springer (2006) 411–426
4. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web. In: WWW'98, Brisbane, Australia (1998) 161–172
5. Jiang, J.J., Conrath, D.W.: Semantic Similarity based on Corpus Statistics and Lexical Taxonomy. In: Proceedings of the International Conference on Research in Computational Linguistics (ROCLING), Taiwan (1997)
6. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of lexical semantic relatedness. Computational Linguistics **32**(1) (2006) 13–47
7. Baeza-Yates, R., Tiberi, A.: Extracting semantic relations from query logs. In: KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, ACM (2007) 76–85
8. Zhang, D., Dong, Y.: A novel web usage mining approach for search engines. Computer Networks **39**(3) (june 2002) 303–310
9. Francisco, A.P., Baeza-Yates, R.A., Oliveira, A.L.: Clique analysis of query log graphs. In Amir, A., Turpin, A., Moffat, A., eds.: SPIRE. Volume 5280 of Lecture Notes in Computer Science., Springer (2008) 188–199
10. Pass, G., Chowdhury, A., Torgeson, C.: A picture of search. In: Proc. 1st Intl. Conf. on Scalable Information Systems, ACM Press New York, NY, USA (2006)
11. Salton, G.: Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1989)
12. Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems **30**(1-7) (April 1998) 107–117
13. Fellbaum, C., ed.: WordNet: an electronic lexical database. MIT Press (1998)
14. Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: Proceedings of the XI International Joint Conferences on Artificial. (1995) 448–453
15. Pedersen, T., Patwardhan, S., Michelizzi, J.: Wordnet::similarity - measuring the relatedness of concepts (2004) http://citeseer.ist.psu.edu/665035.html.