Semantic Grounding of Tag Relatedness in Social Bookmarking Systems

Ciro Cattuto[1], Dominik Benz[2], Andreas Hotho[2], and Gerd Stumme[2]

[1] Complex Networks Lagrange Laboratory, Institute for Scientific Interchange (ISI) Foundation, 10133 Torino, Italy
ciro.cattuto@isi.it
[2] Knowledge & Data Engineering Group, University of Kassel, 34121 Kassel, Germany
{benz,hotho,stumme}@cs.uni-kassel.de

Abstract. Collaborative tagging systems have nowadays become important data sources for populating semantic web applications. For tasks like synonym detection and discovery of concept hierarchies, many researchers introduced measures of tag similarity. Even though most of these measures appear very natural, their design often seems to be rather ad hoc, and the underlying assumptions on the notion of similarity are not made explicit. A more systematic characterization and validation of tag similarity in terms of formal representations of knowledge is still lacking. Here we address this issue and analyze several measures of tag similarity: Each measure is computed on data from the social bookmarking system del.icio.us and a semantic grounding is provided by mapping pairs of similar tags in the folksonomy to pairs of synsets in Wordnet, where we use validated measures of semantic distance to characterize the semantic relation between the mapped tags. This exposes important features of the investigated measures and indicates which measures are better suited in the context of a given semantic application.

1 Introduction

Social bookmarking systems have become extremely popular in recent years. Their underlying data structures, known as *folksonomies*, consist of a set of users, a set of free-form keywords (called *tags*), a set of resources, and a set of tag assignments, i. e., a set of user/tag/resource triples. As folksonomies are large-scale bodies of lightweight annotations provided by humans, they are becoming more and more interesting for research communities that focus on extracting machine-processable semantic structures from them. The structure of folksonomies, however, differs fundamentally from that of e.g., natural text or web resources, and sets new challenges for the fields of knowledge discovery and ontology learning. Central to these tasks are the concepts of similarity and relatedness. In this paper, we focus on similarity and relatedness of tags, because they carry the semantic information within a folksonomy, and provide thus the link to ontologies. Additionally, this focus allows for an evaluation with well-established measures of similarity in existing lexical databases.

Budanitsky and Hirst pointed out that similarity can be considered as a special case of relatedness [1]. As both similarity and relatedness are semantic notions, one way of defining them for a folksonomy is to map the tags to a thesaurus or lexicon like Roget's thesaurus[3]

[3] http://www.gutenberg.org/etext/22

or WordNet [2], and to measure the relatedness there by means of well-known metrics. The other option is to define measures of relatedness directly on the network structure of the folksonomy. One important reason for using measures grounded in the folksonomy, instead of mapping tags to a thesaurus, is the observation that the vocabulary of folksonomies includes many community-specific terms which did not make it yet into any lexical resource. Measures of tag relatedness in a folksonomy can be defined in several ways. Most of these definitions use statistical information about different types of *co-occurrence* between tags, resources and users. Other approaches adopt the *distributional hypothesis* [3, 4], which states that words found in similar contexts tend to be semantically similar. From a linguistic point of view, these two families of measures focus on orthogonal aspects of structural semiotics [5, 6]. The co-occurrence measures address the so-called syntagmatic relation, where words are considered related if they occur in the same part of text. The contextual measures address the paradigmatic relation (originally called associative relation by Saussure), where words are considered related if they can replace one another without affecting the structure of the sentence.

In most studies, the selected measures of relatedness seem to have been chosen in a rather ad-hoc fashion. We believe that a deeper insight into the semantic properties of relatedness measures is an important prerequisite for the design of ontology learning procedures that are capable of harvesting the emergent semantics of a folksonomy.

In this paper we analyse five measures of tag relatedness: the *co-occurrence count*, *three distributional measures* which use the cosine similarity [7] in the vector spaces spanned by users, tags, and resources, respectively, and *FolkRank* [8], a graph-based measure that is an adaptation of PageRank [9] to folksonomies. Our analysis is based on data from a large-scale snapshot of the popular social bookmarking system del.icio.us.[4] To provide a semantic grounding of our folksonomy-based measures, we map the tags of del.icio.us to synsets of WordNet and use the semantic relations of WordNet to infer corresponding semantic relations in the folksonomy. In WordNet, we measure the similarity by using both the taxonomic path length and a similarity measure by Jiang and Conrath [10] that has been validated through user studies and applications [1]. The use of taxonomic path lengths, in particular, allows us to inspect the edge composition of paths leading from one tag to the corresponding related tags, and such a characterization proves to be especially insightful.

The paper is organized as follows: In Section 2 we discuss related work. In Section 3 we provide a formal definition of a folksonomy and describe the del.icio.us data on which our experiments are based. Section 4 describes the measures of relatedness that we will analyze. Section 5 provides examples and qualitative insights. The semantic grounding of the measures in WordNet is described in Section 6. We discuss our results in the context of ontology learning and related tasks in Section 7, where we also point to future work.

2 Related Work

One of the first studies about folksonomies is Ref. [11], where several concepts of bottom-up social annotation are introduced. Ref. [12, 13, 11] provide overviews of the strengths and weaknesses of such systems. Ref. [14, 15] introduce a tri-partite graph representation for folksonomies, where nodes are users, tags and resources. Ref. [16] provides a first quantitative analysis of del.icio.us. We investigated the distribution of tag co-occurrence frequencies

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

in Ref. [17] and the network structure of folksonomies in Ref. [18]. Tag-based metrics for resource distance have been introduced in Ref. [19]. To the best of our knowledge, no systematic characterization of tag relatedness in folksonomies is available in the literature.

Ref. [20] generalizes standard tree-based measures of semantic similarity to the case where documents are classified in the nodes of an ontology with non-hierarchical components. The measures introduced there were validated by means of a user study. Ref. [21] analyses distributional measures of word relatedness and compares them with measures of semantic relatedness in thesauri like WordNet. They concluded that "even though ontological measures are likely to perform better as they rely on a much richer knowledge source, distributional measures have certain advantages. For example, they can easily provide domain-specific similarity measures for a large number of domains, their ability to determine similarity of contextually associated word pairs more appropriately [...]."

The distributional hypothesis is also at the basis of a number of approaches to synonym acquisition from text corpora [22]. As in other ontology learning scenarios, clustering techniques are often applied to group similar terms extracted from a corpus, and a core building block of such procedure is the metric used to judge term similarity. In order to adapt these approaches to folksonomies, several distributional measures of tag relatedness have been introduced in theoretical studies or implemented in applications [23, 24]. However, the choice of a specific measure of relatedness is often made without justification and often it appears to be rather ad hoc.

A task which depends heavily on quantifying tag relatedness is that of tag recommendation in folksonomies. Scientific publications in this domain are still sparse. Existing work can be broadly divided in approaches that analyze the content of the tagged resources with information retrieval techniques [25, 26] and approaches that use collaborative filtering methods based on the folksonomy structure [27]. An example of the latter class of approaches is Ref. [28], where we used our FolkRank algorithm [8] for tag recommendation. FolkRank-based measures will be also covered in this paper.

Relatedness measures also play a role in assisting users who browse the contents of a folksonomy. Ref. [29] shows that navigation in a folksonomy can be enhanced by suggesting tag relations grounded in content-based features.

A considerable number of investigations are motivated by the vision of "bridging the gap" between the Semantic Web and Web 2.0 by means of ontology-learning procedures based on folksonomy annotations. Ref. [15] provides a model of semantic-social networks for extracting lightweight ontologies from del.icio.us. Other approaches for learning taxonomic relations from tags are provided by Ref. [23, 24]. Ref. [30] presents a generative model for folksonomies and also addresses the learning of taxonomic relations. Ref. [31] applies statistical methods to infer global semantics from a folksonomy. The results of our paper are especially relevant to inform the design of such learning methods.

3 Folksonomy Definition and Data

In the following we will use the definition of folksonomy provided in Ref. [8]: [5]

Definition 1. *A* folksonomy *is a tuple* $\mathbb{F} := (U, T, R, Y)$ *where* U*,* T*, and* R *are finite sets, whose elements are called* users*,* tags *and* resources*, respectively., and* Y *is a ternary relation*

[5] Ref. [8] additionally introduces a user-specific sub-tag/super-tag relation, which we will ignore here as it is not relevant for del.icio.us.

between them, i. e., $Y \subseteq U \times T \times R$*. A* post *is a triple* (u, T_{ur}, r) *with* $u \in U$*,* $r \in R$*, and a non-empty set* $T_{ur} := \{t \in T \mid (u, t, r) \in Y\}$.

Users are typically represented by their user ID, tags may be arbitrary strings, and resources depend on the system and are usually represented by a unique ID. For instance, in del.icio.us the resources are URLs, while in YouTube the resources are videos.

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 restricted folksonomy consists of $|U| = 476,378$ users, $|T| = 10,000$ tags, $|R| = 12,660,470$ resources, and $|Y| = 101,491,722$ tag assignments.

4 Measures of Relatedness

A folksonomy can be also regarded as an undirected tri-partite hyper-graph $G = (V, E)$, where $V = U \cup T \cup R$ is the set of nodes, and $E = \{\{u, t, r\} \mid (u, t, r) \in Y\}$ is the set of hyper-edges. Alternatively, the folksonomy hyper-graph can be represented as a three-dimensional (binary) adjacency matrix. In Formal Concept Analysis [32] this structure is known as a *triadic context* [33]. All these equivalent notions make explicit that folksonomies are special cases of three-mode data. Since measures of similarity and relatedness are not well developed for three-mode data yet, we will consider two- and one-mode views on the data. These views will be complemented by a graph-based approach for discovering related tags (FolkRank) which makes direct use of the three-mode structure.

4.1 Co-Occurrence

Given a folksonomy (U, T, R, Y), we define the *tag-tag co-occurrence graph* as a weighted undirected graph whose set of vertices is the set T of tags. Two tags t_1 and t_2 are connected by an edge, iff there is at least one post (u, T_{ur}, r) with $t_1, t_2 \in T_{ur}$. The *weight* of this edge is given by the number of posts that contain both t_1 and t_2, i. e.,

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

Co-occurrence relatedness between tags is given directly by the edge weights. For a given tag $t \in T$, 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*. For its computation, we first create a sorted list of all tag pairs which occur together in a post. The complexity of this can be estimated as $O(\frac{|Y|^2}{2|P|} \log(\frac{|Y|^2}{2|P|}))$. Then, we group this list by each tag and sort by count, which corresponds to an additional complexity of $O(|T|^2 \log(|T|^2))$. Y, P, T denote the set of tag assignments, posts and tags, respectively (see Section 3).

4.2 Distributional Measures

We introduce three distributional measures of tag relatedness that are based on three different vector space representations of tags. 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 folksonomy. Specifically, for $X \in \{U, T, R\}$ we consider the vector space $\mathbb{R}^X$, where each tag 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. The complexity of this measure comprises the cost of computing co-occurrence (see above), i.e., $O(\frac{|Y|^2}{2|P|}\log(\frac{|Y|^2}{2|P|}) + |T|^2\log(|T|^2))$, plus the cost of comparing each tag pair, which is $O(|T|^2 2|X|), X \subseteq T$. In our case $|X| = 10,000$.

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 $r \in R$: $v_{tr} := \mathrm{card}\{u \in U \mid (u, t, r) \in Y\}$. In terms of complexity, the tag-resource counts amount for $O(|Y|\log(|Y|))$, plus the pairwise comparison cost of $O(|T|^2 2|R|)$.

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 $v_{tu} := \mathrm{card}\{r \in R \mid (u, t, r) \in Y\}$. In this case, the complexity is $O(|Y|\log(|Y|) + |T|^2 2|U|))$.

In all three representations, we measure vector similarity by using the cosine measure, as is customary in Information Retrieval [7]: 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: $\mathrm{cossim}(t_1, t_2) := \cos \measuredangle(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{\boldsymbol{v}_1 \cdot \boldsymbol{v}_2}{||\boldsymbol{v}_1||_2 \cdot ||\boldsymbol{v}_2||_2}$. The cosine similarity is thus independent of the length of the vectors. Its value ranges from 0 (for totally orthogonal vectors) to 1 (for vectors pointing into the same direction).

4.3 FolkRank

The PageRank algorithm [34] reflects the idea that a web page is important if there are many pages linking to it, and if those pages are important themselves. We employed the same principle for folksonomies [8]: 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.

More specifically, FolkRank considers a folksonomy (U, T, R, Y) as an undirected graph $(U \cup T \cup R, E)$ with $E := \{\{u, t\}, \{u, r\}, \{t, r\} \mid (u, t, r) \in Y\}$. For a given tag t, it computes in this graph the usual PageRank [34] with a high weight for t in the random surfer vector.[6] Then, the resulting vector is compared to the case of PageRank without random

[6] In this paper, we have set the weights in the random surfer vector as follows: Initially, each tag is assigned weight 1. Then, the weight of the given tag t is increased according to $w(t) = w(t) + |T|$. Afterwards, the vector is normalized. The random surfer has an influence of 15 % in each iteration.

Table 1. Examples of most related tags for each of the presented measures.

rank	tag	measure	1	2	3	4	5
13	web2.0	*co-occurrence*	ajax	web	tools	blog	webdesign
		folkrank	web	ajax	tools	design	blog
		tag context	web2	web-2.0	webapp	"web	web_2.0
		resource context	web2	web20	2.0	web_2.0	web-2.0
		user context	ajax	aggregator	rss	google	collaboration
15	howto	*co-occurrence*	tutorial	reference	tips	linux	programming
		folkrank	reference	linux	tutorial	programming	software
		tag context	how-to	guide	tutorials	help	how_to
		resource context	how-to	tutorial	tutorials	tips	diy
		user context	reference	tutorial	tips	hacks	tools
28	games	*co-occurrence*	fun	flash	game	free	software
		folkrank	game	fun	flash	software	programming
		tag context	game	timewaster	spiel	jeu	bored
		resource context	game	gaming	juegos	videogames	fun
		user context	video	reference	fun	books	science
30	java	*co-occurrence*	programming	development	opensource	software	web
		folkrank	programming	development	software	ajax	web
		tag context	python	perl	code	c++	delphi
		resource context	j2ee	j2se	javadoc	development	programming
		user context	eclipse	j2ee	junit	spring	xml
39	opensource	*co-occurrence*	software	linux	programming	tools	free
		folkrank	software	linux	programming	tools	web
		tag context	open_source	open-source	open.source	oss	foss
		resource context	open-source	open	open_source	oss	software
		user context	programming	linux	framework	ajax	windows
1152	tobuy	*co-occurrence*	shopping	books	book	design	toread
		folkrank	toread	shopping	design	books	music
		tag context	wishlist	to_buy	buyme	wish-list	iwant
		resource context	wishlist	shopping	clothing	tshirts	t-shirts
		user context	toread	cdm	todownload	todo	magnet

surfer (which equals the simple edge count, as the graph is undirected). This way we compute the winners (and losers) that arise when giving preference to a specific tag in the random surfer vector. The tags that, for a given tag t, obtain the highest FolkRank are considered to be the most relevant in relation to t. Ref. [8] provides a detailed description of the algorithm. The complexity of FolkRank can be estimated as $O(i|Y|)$, where i is the number of iterations (the typical values used in this study were 30-35).

5 Qualitative insights

Using each of the measures introduced above, we computed, for each of the $10,000$ most frequent tags of del.icio.us, its most closely related tags. As we used different (partially existing) implementations for the measures we investigate, runtimes do not provide meaningful information on the computational cost of the different measures. We refer the reader to the discussion of Section 4 on computational complexity.

Table 1 provides a few examples of the related tags returned by the measures under study. A first observation is that in many cases the tag and resource context similarity provide more synonyms than the other measures. For instance, for the tag *web2.0* they return some of its alternative spellings.[7] For the tag *games*, the tag and resource similarity also provide tags that could be regarded as semantically *similar*. For instance, the morphological variations *game* and *gaming*, or corresponding words in other languages, like *spiel* (German), *jeu* (French)

[7] The tag *"web* at the fourth position (tag context) is likely to stem from users who typed *"web 2.0"*, which the early del.icio.us interpreted as two separate tags, *"web* and *2.0"*.

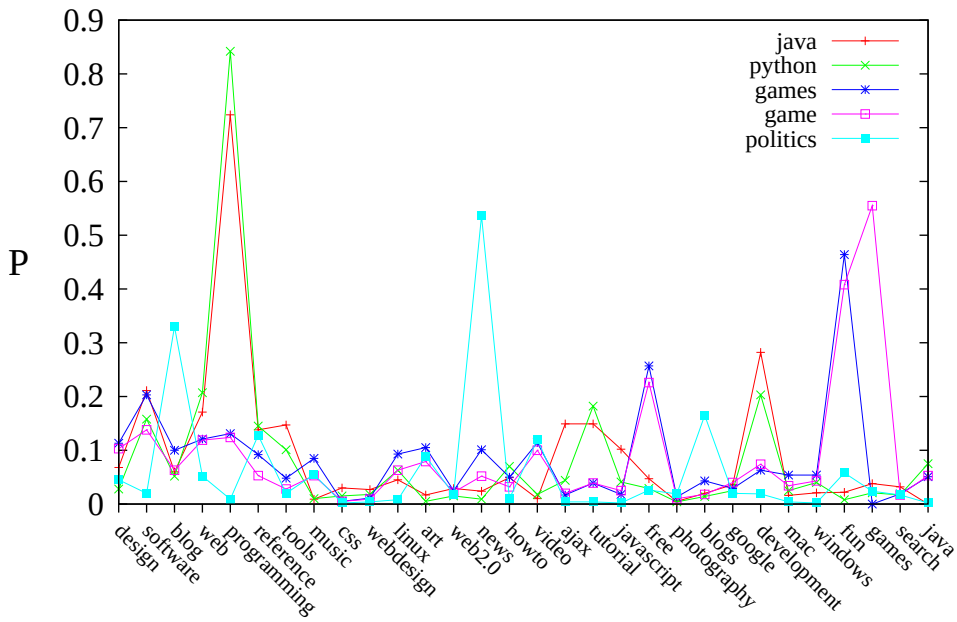

Fig. 1. Tag co-occurrence fingerprint of five selected tags in the first 30 dimensions of the tag vector space.

and *juegos* (Spanish). This effect is not obvious for the other measures, which tend to provide rather *related* tags instead (*video*, *software*). The same observation holds for the "functional" tag *tobuy* (see Ref. [16]), for which the tag context similarity provides tags with equivalent functional value (*to_buy*, *buyme*), whereas the FolkRank and co-occurrence measures provide categories of items one could buy. The user context similarity also yields a remarkable amount of functional tags, but with different target actions (*toread*, *todownload*, *todo*).

An interesting observation about the tag *java* is that *python*, *perl* and *c++* (provided by tag context similarity) could all be considered as siblings in some suitable concept hierarchy, presumably under a common parent concept like *programming languages*. An approach to explain this behavior is that the tag context is measuring the frequency of co-occurrence with other tags *in the global context* of the folksonomy, whereas the co-occurrence measure and — to a lesser extent — FolkRank measure the frequency of co-occurrence with other tags *in the same posts*.

Another insight offered by this first visual inspection is that context similarities for tags and resources seem to yield equivalent results, especially in terms of synonym identification. The tag context measure, however, seems to be the only one capable of identifying sibling tags, as it is visible for the case of *java* in Table 1. This is also visible in Fig. 1, which displays the tag co-occurrence vectors of 5 selected tags. The vectors are restricted to co-occurrence with the 30 most frequent tags of the folksonomy, i.e., to only 30 dimensions of the vector space $\mathbb{R}^T$ introduced in Section 4.[8] The figures shows that both *java* and *python* appear frequently together with *programming*, and (to a lesser degree) with *development*. These two common peaks alone contribute approx. 0.68 to the total cosine similarity of the

[8] The length of all the vectors was normalized to 1 in the 2-norm.

Table 2. Overlap between the 10 most closely related tags.

	co-occurrence	*FolkRank*	*tag context*	*resource context*
user context	1.77	1.81	1.35	1.55
resource context	*3.35*	*2.65*	*2.66*	
tag context	1.69	1.28		
FolkRank	**6.81**			

two tags *java* and *python* of 0.85. A similar behavior can be seen for *game* and *games* both displaying peaks at *fun* and (to a lesser degree) *free*. Here we also see the effect of imposing $v_{tt} = 0$ in the definition of the cosine measure: while the tag *game* has a very high peak at *games*, the tag *games* has by definition a zero component there.

The high value for tag *game* in the dimension *games* shows that these two tags are frequently assigned together to resources (probably because users anticipate that they will not remember a specific form at the time of retrieval).

In the case of *python*, on the other hand, we observe that it seldom co-occurs with *java* in the same posts (probably because few web pages deal with both java and python). Hence — even though *python* and *java* are "most related" according to the tag context similarity, they are less so according to the other measures. In fact, in the lists of tags most closely related to *java*, *python* is at position 21 according to FolkRank, 34 according to co-occurrence, 97 according to user context similarity, and 476 according to resource context similarity.

Our next step is to substantiate these first insights with a more systematic analysis. We start by using simple observables that provide qualitative insights into their behavior.

The first natural aspect to investigate is whether the most closely related tags are shared across measures of relatedness. We consider the $10,000$ most popular tags in del.icio.us, and for each of them we compute the 10 most related tags according to each of the relatedness measures. Table 2 reports the average number of shared tags for the relatedness measures we investigate. We first observe that the user context measure does not exhibit a strong similarity to any of the other measures. The same holds for the tag context measure, with a slightly higher overlap of 2.65 tags with the resource context measure. Based on the visual inspection above, this can be attributed to shared synonym tags. A comparable overlap also exists between resource context and FolkRank / co-occurrence similarity, respectively.

Based on the current analysis, it is hard to learn much on the nature of these overlapping tags. A remarkable fact, however, is that relatedness by co-occurrence and by FolkRank share a large fraction (6.81) of the 10 most closely related tags. That is, given a tag t, its related tags according to FolkRank are – to a large extent – tags with a high frequency of co-occurrence with t. In the case of the context relatedness measures, instead, the suggested tags seem to bear no special bias towards high-frequency tags. This is due to the normalization of the vectors that is implicit in the cosine similarity (see Section 4), which disregards information about global tag frequency.

To better investigate this point, for each of the $10,000$ most frequent tags in del.icious we computed the average rank (according to global frequency) of its 10 most closely related tags, according to each of the relatedness measures under study. Fig. 2 shows the average rank of the related tags as a function of the original tag's rank. The average rank of the tags obtained by co-occurrence relatedness and by FolkRank is low and increases slowly with the rank of the original tag: this points out that most of the related tags are high-frequency tags, independently of the original tag. On the contrary, the context (distributional) measures display a different behavior: the rank of related tags increases much faster with that of the

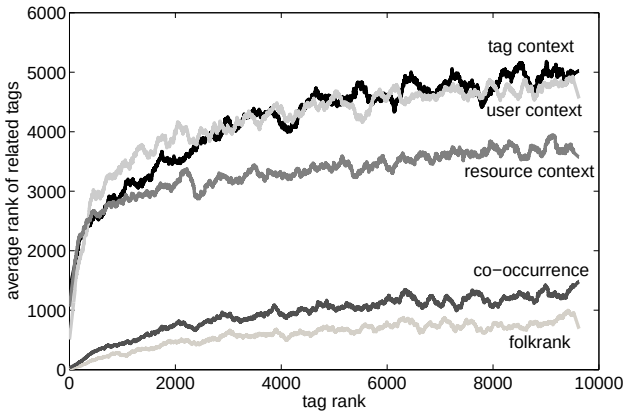

Fig. 2. Average rank of the related tags as a function of the rank of the original tag.

original tag. That is, the tags obtained from context relatedness span a broader range of ranks.[9]

6 Semantic Grounding

In this section we shift perspective and move from the qualitative discussion of Section 5 to a more formal validation. Our strategy is to ground the relations between the original and the related tags by looking up the tags in a formal representation of word meanings. As structured representations of knowledge afford the definition of well-defined metrics of semantic similarity, one can investigate the type of *semantic* relations that hold between the original tags and their related tags, defined according to any of the relatedness measures under study.

In the following we ground our measures of tag relatedness by using WordNet [2], a semantic lexicon of the English language. In WordNet words are grouped into *synsets*, sets of synonyms that represent one concept. Synsets are nodes in a network and links between synsets represent semantic relations. WordNet provides a distinct network structure for each syntactic category (nouns, verbs, adjectives and adverbs). For nouns and verbs it is possible to restrict the links in the network to (directed) *is-a* relationships only, therefore a subsumption hierarchy can be defined. The *is-a* relation connects a *hyponym* (more specific synset) to a *hypernym* (more general synset). A synset can have multiple hypernyms, so that the graph is not a tree, but a directed acyclic graph. Since the *is-a* WordNet network for nouns and verbs consists of several disconnected hierarchies, it is useful to add a fake top-level node subsuming all the roots of those hierarchies, making the graph fully connected and allowing the definition of several graph-based similarity metrics between pairs of nouns and pairs of verbs. We will use such metrics to ground and characterize our measures of tag relatedness in folksonomies.

In WordNet, we will measure the semantic similarity by using both the taxonomic shortest-path length and a measure of semantic distance introduced by Jiang and Conrath [10] that

[9] Notice that the curves for the tag and user context relatedness approach a value of $5\,000$ for high ranks: this is the value one would expect if the rank of the related tags was independent from the rank of the original tags.

Table 3. WordNet coverage of del.icio.us tags.

# top-frequency tags	100	500	1,000	5,000	10,000
fraction in WordNet	82 %	80 %	79 %	69 %	61 %

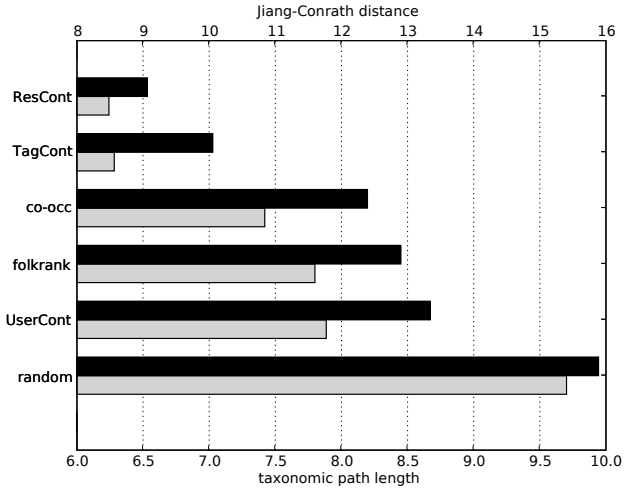

Fig. 3. Average semantic distance, measured in WordNet, from the original tag to the most closely related one. The distance is reported for each of the measures of tag similarity discussed in the main text (labels on the left). Grey bars (bottom) show the taxonomic path length in WordNet. Black bars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

combines the taxonomic path length with an information-theoretic similarity measure by Resnik [35]. We use the implementation of those measures available in the `WordNet::Similarity` library [36]. It is important to remark that [1] provides a pragmatic grounding of the Jiang-Conrath measure by means of user studies and by its superior performance in the context of a spell-checking application. Thus, our semantic grounding in WordNet of the similarity measures is extended to the pragmatic grounding in the experiments of [1].

The program outlined above is only viable if a significant fraction of the popular tags in del.icio.us is also present in WordNet. Several factors limit the WordNet coverage of del.icio.us tags: WordNet only covers the English language and contains a static body of words, while del.icio.us contains tags from different languages, tags that are not words at all, and is an open-ended system. Another limiting factor is the structure of WordNet itself, where the measures described above can only be implemented for nouns and verbs, separately. Many tags are actually adjectives [16] and although their grounding is possible, no distance based on the subsumption hierarchy can be computed in the adjective partition of WordNet. Nevertheless, the nominal form of the adjective is often covered by the noun partition. Despite this, if we consider the popular tags in del.icio.us, a significant fraction of them is actually covered by WordNet: as shown in Table 3, roughly 61% of the $10\,000$ most frequent tags in del.icio.us can be found in WordNet. In the following, to make contact with the previous sections, we will focus on these tags only.

A first assessment of the measures of relatedness can be carried out by measuring – in WordNet – the average semantic distance between a tag and the corresponding most closely related tag according to each one of the relatedness measures we consider. Given a measure

of relatedness, we loop over the tags that are both in del.icio.us and WordNet, and for each of those tags we use the chosen measure to find the corresponding most related tag. If the most related tag is also in WordNet, we measure the semantic distance between the synset that contains the original tag and the synset that contains the most closely related tag. When measuring the shortest-path distance, if either 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 (edge) shortest-path length and the Jiang-Conrath distance. The tag and resource context relatedness point to tags that are semantically closer according to both measures. We remark once more that the Jiang-Conrath measure has been validated in user studies [1], and because of this the semantic distances reported in Fig. 3 correspond to distances cognitively perceived by human subjects.

The best performance is achieved by similarity according to resource context. This is not surprising as this measure makes use of a large amount of contextual information (the large vectors of resources associated with tags). While similarity by resource context is computationally very expensive to compute, it can be used as a reference for comparing the performance of other measures. To this end, we also computed the distances for the worst case scenario of a measure (marked as *random* in Figure 3) that associates every tag with a randomly chosen one. All the other measures of relatedness fall between the above extreme cases. Overall, the taxonomic path length and the Jiang-Conrath distance appear strongly correlated, and they induce the same ranking by performance of the similarity measures. Remarkably, the notion of similarity by tag context (*TagCont*) has an almost optimal performance. This is interesting because is it computationally lighter that the similarity by resource context, as it involves tag co-occurrence with a fixed number ($10\,000$) of popular tags, only. The closer semantic proximity of tags obtained by tag and resource context relatedness was intuitively apparent from direct inspection of Table 1, but now we are able to ground this statement through user-validated measures of semantic similarity based on the subsumption hierarchy of WordNet.

As already noted in Section 5, the related tags obtained via tag context or resource context appear to be "synonyms" or "siblings" of the original tag, while other measures of relatedness (co-occurrence and FolkRank) seem to provide "more general" tags. The possibility of looking up tags in the WordNet hierarchy allows us to be more precise about the nature of these relations. In the rest of this section we will focus on the shortest paths in WordNet that lead from an initial tag to its most closely related tag (according to the different measures of relatedness), and characterize the length and edge composition (hypernym/hyponym) of such paths.

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. All similarity measures share the same overall behavior for $n > 3$, with a broad maximum around $n \simeq 6$, while significant differences are visible for small values of n. Specifically, similarity by tag context and resource context display a strong peak at $n = 0$. Tag context similarity also displays a weaker peak at $n = 2$ and a comparatively depleted number of paths with $n = 1$. For higher values of n, the histogram for resource context and tag context has the same shape as the others, but is systematically lower due to the abundance of very short paths and the normalization of $P(n)$. The peak at $n = 0$ is due to the detection of actual synonyms in WordNet. As nodes in WordNet are synsets, a path to a synonym appears as an edge connecting a node to itself (i. e., a path of length 0). Similarity by tag context points to a synonym in about $18\,\%$ of

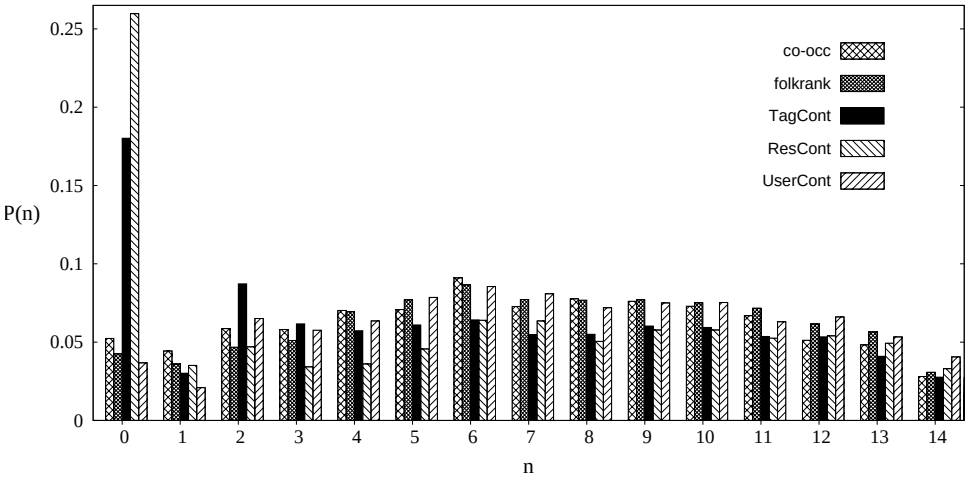

Fig. 4. Probability distribution for the lengths of the shortest path leading from the original tag to the most closely related one. Path lengths are computed using the subsumption hierarchy in WordNet.

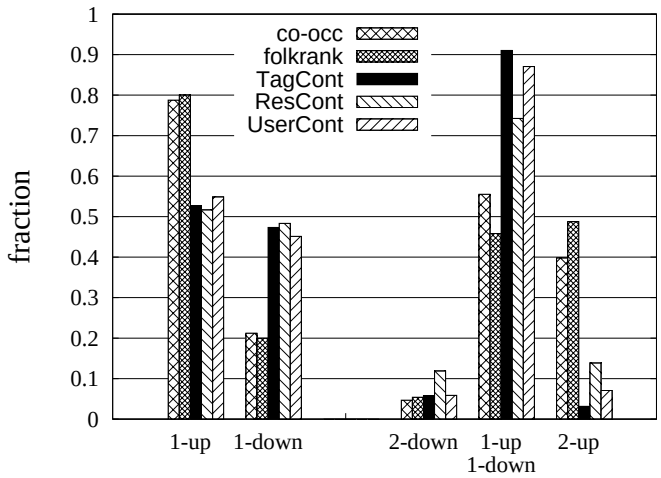

Fig. 5. Edge composition of the shortest paths of length 1 (left) and 2 (right). An "up" edge leads to a hypernym, while a "down" edge leads to a hyponym.

the cases, while using resource context this figure raises to about $25\,\%$. In the above cases, the most related tag is a tag belonging to the same synset of the original tag. In the case of tag context, the smaller number of paths with $n = 1$ (compared with $n = 0$ and $n = 2$) is consistent with the idea that the similarity of tag context favors siblings/synonymous tags: moving by a single edge, instead, leads to either a hypernym or a hyponym in the WordNet hierarchy, never to a sibling. The higher value at $n = 2$ (paths with two edges in WordNet) for tag context may be compatible with the sibling relation, but in order to ascertain this we have to characterize the typical edge composition of these paths.

Figure 5 displays the average edge type composition (hypernym/hyponym edges) for paths of length 1 and 2. The paths analyzed here correspond to $n = 1$ and $n = 2$ in Figure 4. For tag context, resource context and user context, we observe that the paths with $n = 2$

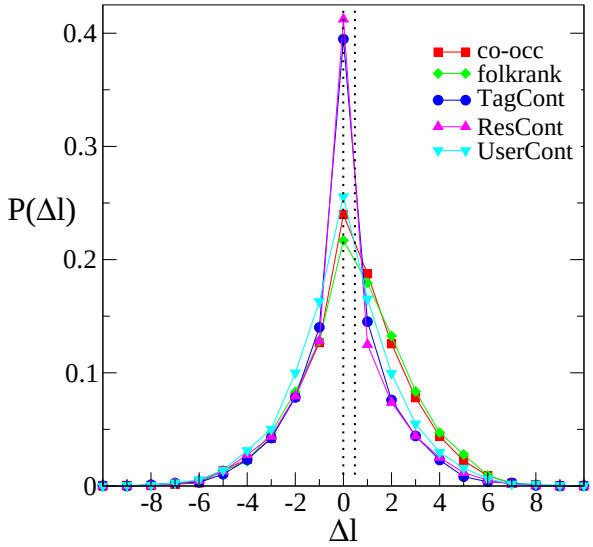

Fig. 6. Probability distribution of the level displacement Δl in the WordNet hierarchy.

(right-hand side of Figure 5) consist almost entirely of one hypernym edge (up) and one hyponym edge (down), i. e., these paths do lead to siblings. This is especially marked for the notion of similarity based on tag context, where the fraction of paths leading to a sibling is about 90% of the total. Notice how the path composition is very different for the other non-contextual measures of relatedness (co-occurrence and FolkRank): in these cases roughly half of the paths consist of two hypernym edges in the WordNet hierarchy, and the other half consists mostly of paths to siblings. We observe a similar behavior for paths with $n = 1$, where the contextual notions of similarity have no statistically preferred direction, while the other measures point preferentially to hypernyms (i. e., 1-up in the WordNet taxonomy).

We now generalize the analysis of Figure 5 to paths of arbitrary length. Specifically, we measure for every path the *hierarchical displacement* Δl in WordNet, i. e., the difference in hierarchical depth between the synset where the path ends and the synset where the path begins. Δl is the difference between the number of edges towards a hypernym (up) and the number of edges towards a hyponym (down). Figure 6 displays the probability distribution $P(\Delta l)$ measured over all tags under study, for the five measures of relatedness. We observe that the distribution for the tag context and resource context is strongly peaked at $\Delta l = 0$ and highly symmetric around it. The fraction of paths with $\Delta l = 0$ is about 40%. The average value of Δl for all the contextual measures is $\overline{\Delta l} \simeq 0$ (dotted line at $\Delta l = 0$) . This reinforces, in a more general fashion, the conclusion that the contextual measures of similarity involve no hierarchical biases and the related tags obtained by them lie at the same level of the original one, in the WordNet hierarchy. Tag context and resource context are more peaked, while the distribution for user context, which is still highly simmetric around $\Delta l = 0$, is broader. Conversely, the probability distributions $P(\Delta l)$ for the non-contextual measures (co-occurrence and FolkRank), look asymmetric and both have averages $\overline{\Delta l} \simeq 0.5$ (righ-hand dotted line). This means that those measures – as we have already observed – point to related tags that preferentially lie higher in the WordNet hierarchy.

7 Discussion and Perspectives

The contribution of this paper is twofold: First, it introduces a systematic methodology for characterizing measures of tag relatedness in a folksonomy. Several measures have been proposed and applied, but given the fluid and open-ended nature of social bookmarking systems, it is hard to characterize – from the semantic point of view – what kind of relations they establish. As these relations constitute an important building block for extracting formalized knowledge, a deeper understanding of tag relatedness is needed. In this paper, we grounded several measures of tag relatedness by mapping the tags of the folksonomy to synsets in WordNet, where we used well-established measures of semantic distance to characterize the inestigated measures of tag relatedness. As a result, we showed that distributional measures, which capture the context of a given tag in terms of resources, users, or other co-occurring tags, establish – in a statistical sense – *paradigmatic* relations between tags in a folksonomy. Strikingly, our analysis shows that the behavior of the most accurate measure of similarity (in terms of semantic distance of the indicated tags) can be matched by a computationally lighter measure (tag context similarity) which only uses co-occurrence with the popular tags of the folksonomy. In general, we showed that a semantic characterization of similarity measures computed on a folksonomy is possible and insightful in terms of the type of relations that can be extracted. We showed that despite a large degree of variability in the tags indicated by different similarity measures, it is possible to connotate *how* the indicated tags are related to the original one.

The second contribution addresses the question of emergent semantics: Our results indicate clearly that, given an appropriate measure, globally meaningful tag relations can be harvested from an aggregated and uncontrolled folksonomy vocabulary. Specifically, we showed that the measures based on tag and resource context are capable of identifying tags belonging to a common semantic concept. Admittedly, in their current status, none of the measures we studied can be seen as *the* way to instant ontology creation. However, we believe that further analysis of these and other measures, as well as research on how to combine them, will help to close the gap towards the Semantic Web.

In an application context, the semantic characterization we provided can be used to guide the choice of a relatedness measure as a function of the task at hand. We will close by briefly discussing which of the relatedness measures we investigated is best for ...

- ... *synonym discovery.* The tag or resource context similarities are clearly the first measures to choose when one would like to discover synonyms. As shown in this work, these measure delivers not only spelling variants, but also terms that belong to the same WordNet synset (see especially Fig. 4). This kind of information could be applied to suggest concepts in tagging system or to support users by cleaning up the tag cloud.
- ... *concept hierarchy.* Both FolkRank and co-occurrence relatedness seemed to yield more general tags in our analyses. This is why we think that these measures provide valuable input for algorithms to extract taxonomic relationships between tags.
- ... *tag recommendations.* The applicability of both FolkRank and co-occurrence for tag recommendations was demonstrated in Ref. [28]. Both measures allow for recommendations by straightforward modifications. Our evaluation in Ref. [28] showed that FolkRank delivered superior and more personalized results than co-occurrence. On the other hand, similar tags and spelling variants as frequently provided by the context similarity are less accepted by the user in recommendations.

– *. . . query expansion.* Our analysis suggests that resource or tag context similarity could be used to discover synonyms and – together with some string edit distance – spelling variants of the tags in a user query. The original tag query could be expanded by using the tags obtained by these measures.

Future work includes the application of different measures of relatedness in the context of the tasks listed above. In particular, we plan to adapt existing ontology learning techniques to the case of folksonomies, building upon the semantic characterization of tag relateness that we presented here.

Acknowledgment

This research has been partly supported by the TAGora project (FP6-IST5-34721) funded by the Future and Emerging Technologies program (IST-FET) of the European Commission. We thank A. Baldassarri, V. Loreto, F. Menczer, V. D. P. Servedio and L. Steels for many stimulating discussions.

References

1. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of lexical semantic relatedness. Computational Linguistics **32**(1) (2006) 13–47
2. Fellbaum, C., ed.: WordNet: an electronic lexical database. MIT Press (1998)
3. Firth, J.R.: A synopsis of linguistic theory 1930-55. Studies in Linguistic Analysis (special volume of the Philological Society) **1952-59** (1957) 1–32
4. Harris, Z.S.: Mathematical Structures of Language. Wiley, New York (1968)
5. de Saussure, F.: Course in General Linguistics. Duckworth, London ([1916] 1983) (trans. Roy Harris).
6. Chandler, D.: Semiotics: The Basics. Second edn. Taylor & Francis (2007)
7. Salton, G.: Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1989)
8. 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
9. 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
10. 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)
11. Mathes, A.: Folksonomies – Cooperative Classification and Communication Through Shared Metadata (December 2004) http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.
12. Hammond, T., Hannay, T., Lund, B., Scott, J.: Social Bookmarking Tools (I): A General Review. D-Lib Magazine **11**(4) (April 2005)
13. Lund, B., Hammond, T., Flack, M., Hannay, T.: Social Bookmarking Tools (II): A Case Study - Connotea. D-Lib Magazine **11**(4) (April 2005)
14. Lambiotte, R., Ausloos, M.: Collaborative tagging as a tripartite network. Lecture Notes in Computer Science **3993** (Dec 2005) 1114
15. Mika, P.: Ontologies are us: A unified model of social networks and semantics. In: International Semantic Web Conference. LNCS, Springer (2005) 522–536

16. Golder, S., Huberman, B.A.: The structure of collaborative tagging systems. Journal of Information Science **32**(2) (April 2006) 198–208
17. Cattuto, C., Loreto, V., Pietronero, L.: Semiotic dynamics and collaborative tagging. Proceedings of the National Academy of Sciences (PNAS) **104** (2007) 1461–1464
18. Cattuto, C., Schmitz, C., Baldassarri, A., Servedio, V.D.P., Loreto, V., Hotho, A., Grahl, M., Stumme, G.: Network properties of folksonomies. AI Communications Journal, Special Issue on Network Analysis in Natural Sciences and Engineering **20**(4) (2007) 245–262
19. Cattuto, C., Baldassarri, A., Servedio, V.D.P., Loreto, V.: Emergent community structure in social tagging systems. Advances in Complex Physics (2007) (to appear) Proceedings of the European Confeence on Complex Systems ECCS2007.
20. Maguitman, A.G., Menczer, F., Erdinc, F., Roinestad, H., Vespignani, A.: Algorithmic computation and approximation of semantic similarity. World Wide Web **9**(4) (2006) 431–456
21. Mohammad, S., Hirst, G.: Distributional measures as proxies for semantic relatedness Submitted for publication, http://ftp.cs.toronto.edu/pub/gh/Mohammad+Hirst-2005.pdf.
22. Cimiano, P.: Ontology Learning and Population from Text — Algorithms, Evaluation and Applications. Springer, Berlin–Heidelberg, Germany (2006) Originally published as PhD Thesis, 2006, Universitt Karlsruhe (TH), Karlsruhe, Germany.
23. Heymann, P., Garcia-Molina, H.: Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Computer Science Department (April 2006)
24. Schmitz, P.: Inducing ontology from Flickr tags. In: Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland. (May 2006)
25. Mishne, G.: Autotag: a collaborative approach to automated tag assignment for weblog posts. In: WWW '06: Proceedings of the 15th international conference on World Wide Web, New York, NY, USA, ACM Press (2006) 953–954
26. Brooks, C.H., Montanez, N.: Improved annotation of the blogosphere via autotagging and hierarchical clustering. In: WWW '06: Proceedings of the 15th international conference on World Wide Web, New York, NY, USA, ACM Press (2006) 625–632
27. Xu, Z., Fu, Y., Mao, J., Su, D.: Towards the semantic web: Collaborative tag suggestions. In: Proceedings of the Collaborative Web Tagging Workshop at the WWW 2006, Edinburgh, Scotland (May 2006)
28. Jäschke, R., Marinho, L.B., Hotho, A., Schmidt-Thieme, L., Stumme, G.: Tag recommendations in folksonomies. In: Proc. PKDD 2007. Volume 4702 of Lecture Notes in Computer Science., Berlin, Heidelberg, Springer (2007) 506–514
29. Aurnhammer, M., Hanappe, P., Steels, L.: Integrating collaborative tagging and emergent semantics for image retrieval. In: Proceedings WWW2006, Collaborative Web Tagging Workshop. (May 2006)
30. Halpin, H., Robu, V., Shepard, H.: The dynamics and semantics of collaborative tagging. In: Proceedings of the 1st Semantic Authoring and Annotation Workshop (SAAW'06). (2006)
31. Zhang, L., Wu, X., Yu, Y.: Emergent semantics from folksonomies: A quantitative study. Journal on Data Semantics VI (2006)
32. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer (1999)
33. Lehmann, F., Wille, R.: A triadic approach to formal concept analysis. In Ellis, G., Levinson, R., Rich, W., Sowa, J.F., eds.: Conceptual Structures: Applications, Implementation and Theory. Volume 954 of Lecture Notes in Computer Science., Springer (1995)
34. 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
35. 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
36. Pedersen, T., Patwardhan, S., Michelizzi, J.: Wordnet::similarity - measuring the relatedness of concepts (2004) http://citeseer.ist.psu.edu/665035.html.