Semantic Analysis of Tag Similarity Measures in Collaborative Tagging Systems

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

Abstract. Social bookmarking systems allow users to organise collections of resources on the Web in a collaborative fashion. The increasing popularity of these systems as well as first insights into their emergent semantics have made them relevant to disciplines like knowledge extraction and ontology learning. The problem of devising methods to measure the semantic relatedness between tags and characterizing it semantically is still largely open. Here we analyze three measures of tag relatedness: tag co-occurrence, cosine similarity of co-occurrence distributions, and FolkRank, an adaptation of the PageRank algorithm to folksonomies. Each measure is computed on tags from a large-scale dataset crawled from the social bookmarking system del.icio.us. To provide a semantic grounding of our findings, a connection to WordNet (a semantic lexicon for the English language) is established by mapping tags into synonym sets of WordNet, and applying there well-known metrics of semantic similarity. Our results clearly expose different characteristics of the selected measures of relatedness, making them applicable to different subtasks of knowledge extraction such as synonym detection or discovery of concept hierarchies.

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. Crucial hereby are the concepts of similarity and relatedness. Here we will focus on similarity and relatedness of tags, because this affords comparison with well-established measures of similarity in existing lexical databases.

Ref. [2] points out that similarity can be considered as a special case of relatedness. 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] or WordNet [6], 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. There are several obvious possibilities and most of them use statistical information about different types of co-occurrence between tags, resources and users. Another possibility is to adopt the *distributional hypothesis* [7, 11], which states that words found in similar contexts tend to be semantically similar. One important reason for using distributional measures in folksonomies 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.

The distributional hypothesis is also at the basis of a number of approaches to synonym acquisition from text corpora [5]. 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 used in theory or implemented in applications [12, 24]. In most studies, however, 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 successfully harvesting the emergent semantics of a folksonomy.

In this paper, we consider the three following measures for the relatedness of tags: the *co-occurrence count*, the *cosine similarity* [23] of co-occurrence distributions, and *FolkRank* [13], a graph-based measure that is an adaptation of PageRank [20] 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 [14] that has been validated through user studies and applications [2]. 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 the next section, we discuss related work. In Section 3 we provide a definition of folksonomy and describe the del.icio.us data on which our experiments are based. Section 4 describes the three measures of relatedness that we will analyze. Section 5 provides first 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 in Section 7, where we also point to future work.

[1] Complex Networks Lagrange Laboratory (CNLL), ISI Foundation, 10133 Torino, Italy, email:cattuto@isi.it

[2] Knowledge & Data Engineering Group, University of Kassel, 34121 Kassel, Germany, email: {benz,hotho,stumme}@cs.uni-kassel.de

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

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

2 Related Work

One of the first scientific publications about folksonomies is [17], where several concept of bottom-up social annotation are introduced. Ref. [15, 18] introduce a tri-partite graph representation for folksonomies, where nodes are users, tags and resources. Ref. [9] provides a first quantitative analysis of del.icio.us.

A considerable number of investigations is 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. [18] provides a model of semantic-social networks for extracting lightweight ontologies from del.icio.us. Other approaches for learning taxonomic relations from tags are [12, 24]. Ref. [10] presents a generative model for folksonomies and also addresses the learning of taxonomic relations. Ref. [25] applies statistical methods to infer global semantics from a folksonomy. The distribution of tag co-occurrence frequencies has been investigated in [3] and the network structure of folksonomies was investigated in [4].

After comparing distributional measures on natural text with measures for semantic relatedness in thesauri like WordNet, [19] concluded that "distributional measures [. . .] can easily provide domain-specific similarity measures for a large number of domains [. . .]." Our work presented in this paper indicates that these findings can be transferred to folksonomies.

3 Folksonomy Definition and Data

In the followin we will use the definition of folksonomy provided in [13]:

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

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 our experiments we used data from the social bookmarking system del.icio.us, collected in November 2006. As one main focus of this work is to characterize tags by their distribution of co-occurrence with other tags, we restricted our data 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 of co-occurrence makes them less useful for the study of 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 [8] this structure is known as a *triadic context* [16]. All these equivalent notions make explicit that folksonomies are special cases of three-mode data. Since measures for similarity and relatedness are not well developed for three-mode data yet, we will consider two- and one-mode views on the data. These two views will be complemented by a graph-based approach for discovering related tags (FolkRank) which makes direct use of the three-mode structure.

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, and where 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.,

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

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 tags $t' \in T$ with $t' \neq t$ such that $w(t, t')$ is maximal. In the sequel, we will denote the co-occurrence relatedness also by *freq*.

Cosine Similarity

We introduce a distributional measure of tag relatedness by computing the cosine similarity of tag-tag co-occurrence distributions. Specifically, we compute the cosine similarity [23] in the vector space $\mathbb{R}^T$, where each tag t is represented by a vector $\vec{v}_t \in \mathbb{R}^T$ with $v_{tt'} := w(t, t')$ for $t \neq t' \in T$ 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.

If two tags t_1 and t_2 are represented by $\vec{v}_1, \vec{v}_2 \in \mathbb{R}^n$, then their cosine similarity is defined as:

$$\operatorname{cossim}(t_1, t_2) := \arccos \measuredangle(\vec{v}_1, \vec{v}_2) = \frac{\vec{v}_1 \cdot \vec{v}_2}{||\vec{v}_1||_2 \cdot ||\vec{v}_2||_2} \qquad (2)$$

FolkRank

The PageRank algorithm [1] reflects the idea that a web page is important if there are many pages linking to it, and if those pages are important themselves. The same principle was employed for folksonomies in [13]: 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. Ref. [13] provides a detailed description.

5 Qualitative insights

Using each of the three measures introduced above, we computed, for each of the 10,000 most frequent tags of del.icio.us, its most closely related tags. Tables 1 – 3 show a few selected examples. We observe that in many cases the cosine similarity provides more synonyms than the other measures. For instance, for tag *web2.0* is returns some of its other commonly used spellings.[5] For tag *games*, the cosine similarity also provides tags that one could consider as semantically *similar* (like the singular form *game* or its German and French

[5] The tag *"web* at the fourth position is likely to stem from some user who typed in *"web 2.0"* which in the earlier del.icio.us was interpreted as two separate tags *"web* and *2.0"*.

Table 1. Examples of most related tags measured by co-occurrence

rank	tag	1	2	3	4	5
13	web2.0	ajax	web	tools	blog	webdesign
15	howto	tutorial	reference	tips	linux	programming
28	games	fun	flash	game	free	software
30	java	programming	development	opensource	software	web
39	opensource	software	linux	programming	tools	free
1152	tobuy	shopping	books	book	design	toread

Table 2. Examples of most related tags measured by cosine similarity

rank	tag	1	2	3	4	5
13	web2.0	web2	web-2.0	webapp	“web	web_2.0
15	howto	how-to	guide	tutorials	help	how_to
28	games	game	timewaster	spiel	jeu	bored
30	java	python	perl	code	c++	delphi
39	opensource	open_source	open-source	open.source	oss	foss
1152	tobuy	wishlist	to_buy	buyme	wish-list	iwant

Table 3. Examples of most related tags measured by Folkrank

rank	tag	1	2	3	4	5
13	web2.0	web	ajax	tools	design	blog
15	howto	reference	linux	tutorial	programming	software
28	games	game	fun	flash	software	programming
30	java	programming	development	software	ajax	web
39	opensource	software	linux	programming	tools	web
1152	tobuy	toread	shopping	design	books	music

translations *spiel* and *jeu*), while the other two measures provide *related* tags like *fun* or *software*. The same observation is also made for the “functional” tag *tobuy* (see [9]), where the cosine similarity provides tags with equivalent functional value, whereas the other measures provide rather categories of items one could buy. An interesting observation is also that *java* and *python* could be considered as siblings in some suitable concept hierarchy. A possible justification for these different behaviors is that the cosine measure is measuring the frequency of co-occurrence with other words *in the global contexts*, whereas the co-occurrence measure and — to a lesser extent — FolkRank measure the frequency of co-occurrence with other words *in the same posts*. We will substantiate this assumption later in the paper on a more general level.

Table 4. Overlap between the ten most closely related tags.

freq–folkrank	cosine–freq	cosine–folkrank
6.7	1.7	1.1

The first natural aspect to investigate is whether the most closely related tags are shared across relatedness measures. 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 4 reports the average number of shared tags for the three relatedness measures. We observe that relatedness by co-occurrence (freq) and by FolkRank share a large fraction of the 10 most closely related tags, while the cosine relatedness displays little overlap with both of them. To better investigate this point, we plot in Figure 1 the average rank (according to global frequency) of the 10 most closely related tags as a function of the rank of the original tag. The average rank of the tags obtained by co-occurrence relatedness (black) and by FolkRank (green) is low and increases slowly with the rank of the original tag: this points out that most of the related tags are among the high-frequency tags, independently of the original tag. On the contrary, the cosine relatedness (red curve) displays a different behavior: the rank of related tags increases much faster with that of the original tag. That is, the tags obtained from cosine-similarity relatedness belong to a broader class of tags, not strongly correlated with rank (frequency).[6]

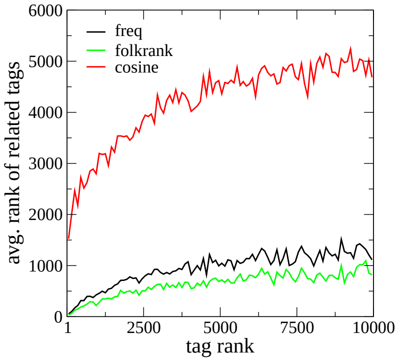

Figure 1. Average rank of the related tags as a function of the rank of the original tag.

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 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 (obtained by using any of the relatedness measures we study).

In the following we ground our measures of tag relatedness by using WordNet [6], 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.

For nouns and verbs it is possible to restrict the links in the network to (directed) *is-a* relationships only, so that a subsumption hierarchy can be defined. The *is-a* relation connects a *hyponym* (more specific synset) to a *hypernym* (more general synset). Since the *is-a* WordNet network for nouns and verbs consists of several disconnected hierarchies, it is useful to add a (fake) global root 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 measures to ground our tag-based measures of relatedness in folksonomies.

We measure the similarity in WordNet using both the taxonomic shortest-path length and a distance measure introduced by Jiang and Conrath [14] that combines the taxonomic path length with an information-theoretic similarity measure by Resnik [22]. We use the implementation of those measures available in the WordNet::Similarity library [21]. We remark that [2] provides a pragmatic grounding of the Jiang-Conrath measure by means of user studies and by its superior performance in the correction of spelling errors. This way, our semantic grounding in WordNet of the folksonomy similarity measures is extended to a pragmatic grounding in the experiments of [2].

The program outlined above is only viable if a significant fraction of the popular tags in del.icio.us is also present 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 and is an open-ended system. This is not a big problem in practice because, to date, the vast majority of del.icio.us tags are grounded in the English language. 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 [9] 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: 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.

Table 5. Average semantic distance, measured in WordNet, from the original tag to the most closely related one.

similarity metric	freq	folkrank	cosine
shortest path	7.4	7.8	6.3
Jiang-Conrath	13.1	13.6	10.8

[6] Notice that the curve for the cosine-similarity relatedness (red) approaches a value of 5 000 for high ranks: this is the value one would expect if tag relatedness was independent from tag rank.

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 of relatedness to find the corresponding most related tag. If the most related tag is also in WordNet, we measure the semantic distance between the synsets that contain the original tag and the most closely related tag, respectively. In the case of the shortest-path distance, if any of the tags occurs in more than one synset, we select synsets which minimizes the path length. Table 5 reports the average semantic distance, computed in WordNet by using both the (edge) shortest-path length and the Jiang-Conrath distance. The cosine relatedness points 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 [2], so that Table 5 actually deals with distances cognitively perceived by human subjects. The closer semantic proximity of tags obtained by cosine relatedness was intuitively apparent from the comparison of Table 2 with Table 1 and Table 3, but now we are able to ground this statement through user-validated measures based on the subsumption hierarchy of WordNet.

As noted in Section 5, the tags obtained via the cosine-similarity relatedness measure appear to be "synonyms" or "siblings" of the original tag, while the two other measures of relatedness seem to provide "more general" tags. The possibility of locating 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 similarity measures), and characterize the length and edge composition (hypernym/hyponym) of such paths.

Table 6. Probabilities of 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.

shortest path length	0	1	2	≥ 3
freq	0.05	0.04	0.06	0.85
folkrank	0.04	0.04	0.05	0.87
cosine	0.18	0.03	0.09	0.70

Table 6 summarizes the probabilities of the shortest-path lengths n (number of edges) connecting a tag to its closest related tag in WordNet. The FolkRank and co-occurrence relatedness have similar probabilities. The cosine relatedness displays higher values at $n = 0$ and $n = 2$ and a comparatively depleted number of paths with $n = 1$. The higher value at $n = 0$ is due to the detection of actual synonyms; i. e., the cosine relatedness, in about 18 % of the cases, points to a tag which belongs to the same synset of the original tag. The smaller number of paths with $n = 1$ (one single edge in WordNet) is consistent with the idea that the cosine relatedness 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) may be compatible with the sibling relation, but in order to ascertain it we have to characterize the average edge composition of these paths.

Figure 2 displays the average edge type composition (hypernym/hyponym edges) for paths of length 1 and 2. For the cosine-similarity relatedness (blue), we observe that the paths with $n = 2$ (right-hand side of Figure 2) consist almost entirely (90%) of one hypernym edge (up) and one hyponym edge (down), i. e., these paths do lead to siblings. Notice how the path composition is very different for the other relatedness measures: in those cases roughly half of the paths consist of two hypernym edges in the WordNet hierarchy. We observe a similar behavior for $n = 1$, where the cosine relatedness has no statistically preferred direction, while the other measures of relatedness point preferentially to hypernyms.

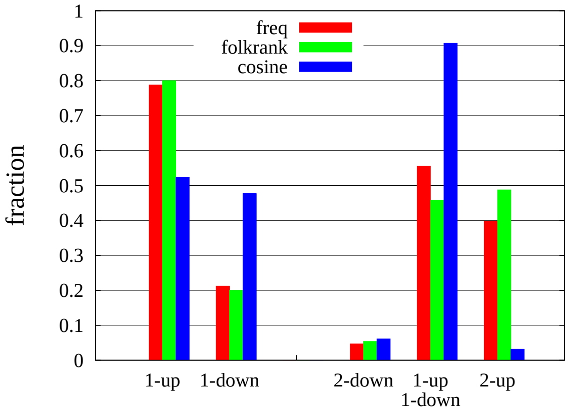

Figure 2. Edge composition of the shortest paths of length 1 (left) and 2 (right). An "up" edge leads to a hypernym, while a "down" edge leads to a hyponym.

7 Discussion and Perspectives

The main contribution of this paper is a methodological one. Several measures of relatedness have been proposed in the literature, 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 these measures is needed. Here we proposed to ground different measures of tag relatedness in a folksonomy by mapping del.icio.us tags, when possible, on WordNet synsets and using well-established measures of semantic distance in WordNet to gain insight into their respective characteristics.

Our results can be taken as indicators that the choice of an appropriate relatedness measure is able to yield valuable input for learning semantic term relationships from folksonomies. We will close by briefly discussing which of the three relatedness measures we studied is best for …

- … *synonym discovery.* The cosine similarity is clearly the measure to choose when one would like to discover synonyms. As shown in this work, cosine similarity delivers not only spelling variants but also terms that belong to the same WordNet synset.
- … *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.
- … *discovery of multi-word lexemes.* Depending on the allowed tag delimiters, it can happen that multi-word lexemes end up as several tags. Our experiment indicates that FolkRank is best to discover these cases. For the tag *open*, for instance, it is the only of the three algorithms which has *source* within the ten most related tags and vice versa.[7]

[7] *Open* is at position 6 for *source*, and *source* is at position 3 for *open*.

Future work includes the analysis of further relatedness measures, e. g., based on representations in the vector spaces spanned by the users or resources. We are furthermore currently working on adapting existing ontology learning techniques to folksonomies, including the presented measures.

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] Sergey Brin and Lawrence Page, 'The Anatomy of a Large-Scale Hypertextual Web Search Engine', *Computer Networks and ISDN Systems*, **30**(1-7), 107–117, (April 1998).

[2] Alexander Budanitsky and Graeme Hirst, 'Evaluating wordnet-based measures of lexical semantic relatedness', *Computational Linguistics*, **32**(1), 13–47, (2006).

[3] Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero, 'Semiotic dynamics and collaborative tagging', *Proceedings of the National Academy of Sciences (PNAS)*, **104**, 1461–1464, (2007).

[4] Ciro Cattuto, Christoph Schmitz, Andrea Baldassarri, Vito D. P. Servedio, Vittorio Loreto, Andreas Hotho, Miranda Grahl, and Gerd Stumme, 'Network properties of folksonomies', *AI Communications Journal, Special Issue on Network Analysis in Natural Sciences and Engineering*, **20**(4), 245–262, (2007).

[5] Philipp Cimiano, *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.

[6] *WordNet: an electronic lexical database*, ed., Christiane Fellbaum, MIT Press, 1998.

[7] J. R. Firth, 'A synopsis of linguistic theory 1930-55.', *Studies in Linguistic Analysis (special volume of the Philological Society)*, **1952-59**, 1–32, (1957).

[8] B. Ganter and R. Wille, *Formal Concept Analysis: Mathematical Foundations*, Spring-Verlag, 1999.

[9] Scott Golder and Bernardo A. Huberman, 'The structure of collaborative tagging systems', *Journal of Information Science*, **32**(2), 198–208, (April 2006).

[10] H. Halpin, V. Robu, and H. Shepard, 'The dynamics and semantics of collaborative tagging', in *Proceedings of the 1st Semantic Authoring and Annotation Workshop (SAAW'06)*, (2006).

[11] Z. S. Harris, *Mathematical Structures of Language*, Wiley, New York, 1968.

[12] Paul Heymann and Hector Garcia-Molina, 'Collaborative creation of communal hierarchical taxonomies in social tagging systems', Technical Report 2006-10, Computer Science Department, (April 2006).

[13] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme, 'Information retrieval in folksonomies: Search and ranking', in *The Semantic Web: Research and Applications*, eds., York Sure and John Domingue, volume 4011 of *LNAI*, pp. 411–426, Heidelberg, (2006). Springer.

[14] Jay J. Jiang and David W. Conrath, 'Semantic Similarity based on Corpus Statistics and Lexical Taxonomy', in *Proceedings of the International Conference on Research in Computational Linguistics (ROCLING)*. Taiwan, (1997).

[15] R. Lambiotte and M. Ausloos, 'Collaborative tagging as a tripartite network', *Lecture Notes in Computer Science*, **3993**, 1114, (Dec 2005).

[16] F. Lehmann and R. Wille, 'A triadic approach to formal concept analysis', in *Conceptual Structures: Applications, Implementation and Theory*, eds., G. Ellis, R. Levinson, W. Rich, and J. F. Sowa, volume 954 of *Lecture Notes in Computer Science*. Springer, (1995).

[17] Adam Mathes. Folksonomies – Cooperative Classification and Communication Through Shared Metadata, December 2004. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.

[18] Peter Mika, 'Ontologies are us: A unified model of social networks and semantics', in *International Semantic Web Conference*, LNCS, pp. 522–536. Springer, (2005).

[19] Saif Mohammad and Graeme Hirst. Distributional measures as proxies for semantic relatedness. Submitted for publication, http://ftp.cs.toronto.edu/pub/gh/Mohammad+Hirst-2005.pdf.

[20] L. Page, S. Brin, R. Motwani, and T. Winograd, 'The PageRank citation ranking: Bringing order to the web', in *WWW'98*, pp. 161–172, Brisbane, Australia, (1998).

[21] T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet::similarity - measuring the relatedness of concepts, 2004. http://citeseer.ist.psu.edu/665035.html.

[22] Philip Resnik, 'Using Information Content to Evaluate Semantic Similarity in a Taxonomy', in *Proceedings of the XI International Joint Conferences on Artificial*, pp. 448–453, (1995).

[23] Gerard Salton, *Automatic text processing: the transformation, analysis, and retrieval of information by computer*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[24] Patrick Schmitz, 'Inducing ontology from Flickr tags.', in *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, (May 2006).

[25] Lei Zhang, Xian Wu, and Yong Yu, 'Emergent semantics from folksonomies: A quantitative study', *Journal on Data Semantics VI*, (2006).