A Study of Sense-disambiguated Networks Induced from Folksonomies

Hans-Peter Zorn and Iryna Gurevych

Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt
Hochschulstraße 10
64289 Darmstadt, Germany

Abstract. Lexical-semantic resources are fundamental building blocks in natural language processing (NLP). Frequently, they fail to cover the informal vocabulary of web users as represented in user-generated content. This paper aims at exploring folksonomies as a novel source of lexical-semantic information. It analyzes two prototypical examples of folksonomies, namely BibSonomy and Delicious, and utilizes NLP and word sense induction techniques to turn the folksonomies into word sense–disambiguated networks representing the vocabulary and the word senses found in folksonomies. The main contribution of the paper is an in-depth analysis of the resulting resources, which can be combined with conventional wordnets to achieve broad coverage of user-generated content.

1 Introduction

Lexical-semantic resources are fundamental building blocks of natural language processing systems. For a long time, WordNet has been widely deployed in a great variety of tasks. A few years ago, the attention of researchers has turned to the so-called collaboratively created lexical-semantic resources such as Wikipedia[1] and Wiktionary[2] (Gurevych and Wolf, 2010). They have been found to perform well as sources of background knowledge in multiple NLP tasks.

While collaboratively created resources represent an excellent addition to conventional lexical-semantic resources, their coverage is insufficient in terms of the represented domains. Also, there is often lack of contextual information about individual sense descriptions. This is why, in this paper, we turn to folksonomies as a promising source of lexical-semantic information. *Folksonomy* is a term used first by Vander Wal (2004) and defined as "tagging is ontology that works", in reference to social bookmarking systems like Delicious. Such applications allow users to assign tags to web sites or other resources. By doing so, they create a structure consisting of the three entities, "tag, the object being tagged and [the user's] identity" which is referred to as a folksonomy.

Our goal is to derive a resource that can effectively complement conventional resources, *e.g.*, WordNet, as folksonomies are known to contain special vocabulary reflecting users' interests, current trends, or neologisms typically lacking in conventional resources. To achieve this, we investigate two widely deployed folksonomies, Delicious and BibSonomy. We constructed three types of tag-based graphs utilizing the tag co-occurrence with resources, users, and other tags. Then, we apply graph clustering techniques to perform word sense induction and obtain a word sense–disambiguated lexical-semantic network. Based on this graph, we present a detailed analysis of the resulting networks in terms of their graph-theoretic properties and compare them with the properties of conventional resources. We also compare their coverage and find that a folksonomy-derived resource includes web-specific vocabulary that is lacking in other resources. Furthermore,

[1] `http://wikipedia.org`

[2] `http://wiktionary.org`

we investigate the nature of the resulting word sense distributions for individual lexemes and relate them to the word senses encoded in WordNet and Wiktionary.

There has been work on inducing hierarchies (Heymann and Garcia-Molina, 2006) and ontologies (Schmitz, 2006) from folksonomies based on co-occurrence patterns. Recently, work has been published on clustering and disambiguating tag similarity graphs. Yeung *et al.* (2009) employ the Girvan–Newman algorithm (Girvan and Newman, 2002) to optimize the betweenness of nodes by incrementally removing edges and recalculating betweenness. They only look at a small data set of ten exemplary tags. Our goal is to go beyond simple clustering and do actual disambiguation, meaning to instantiate the different readings of a tag as individual nodes within the network. Jurgens (2011) presents a way to use community detection to induce word senses from word collocations. He builds a graph from word co-occurrences and then applies a community detection algorithm. In contrast to the approach in this work, they do not split the nodes into different senses but use a clustering algorithm that deals with nodes belonging to multiple communities. The algorithm is a hierarchical agglomerative clustering operating on a similarity measure defined for edges, not vertices.

The paper is structured as follows: First, we describe how to derive a tag similarity graph from folksonomies and introduce our approach for sense disambiguation of individual tags (Section 2) . We analyze the resulting graphs in terms of their properties (Section 3) and coverage (Section 4). With this analysis we are laying ground to combine traditional LSRs and folksonomies in applications.

2 From Folksonomies to Lexical-Semantic Resources

2.1 Datasets

We explore two datasets for our investigations, Delicious and BibSonomy, which are available for research purposes.

The first dataset was extracted by TU Berlin from the Delicious social bookmarking site (Wetzker *et al.*, 2008). Delicious is a general-purpose social bookmarking system founded in 2003. TU Berlin crawled about 142 million bookmarks resulting in more than 450 million tag assignments. We used the first 30 million tag assignments for our experiments, a number where the vector representation fits in about 12 GB of RAM. The Delicious system is generally open to everyone, but the users seem to be mostly technical and IT people, resulting in a higher coverage in these domains. We choose Delicious because of its size and its broad coverage.

BibSonomy (Benz *et al.*, 2010) is a social bookmarking system for both publications and web sites. It was created by the Knowledge and Data Engineering research group at the University of Kassel to provide a useful service to the research community and to collect data for their social bookmarking research. Therefore the dataset[3] is available for research purposes and has been used in various studies on tag recommendation. We choose BibSonomy as a reference dataset in the field of folksonomies. We perform only light preprocessing by removing special characters[4] because we see concatenations by hyphens, underscores, and colons as possibly emerging neologisms or Internet jargon that we are actually interested in.

2.2 Graph Representation of Lexical-Semantic Resources

Previously, folksonomies have been principally considered a tool to recommend tags to users, and have been studied with regard to their semantic properties. The contribution of our work is the conversion of a folksonomy to a lexical-semantic resource (LSR) that could be used alongside other LSRs. LSRs can be represented as graph structures that contain words and their relations to

[3] Knowledge and Data Engineering Group, University of Kassel: Benchmark Folksonomy Data from BibSonomy, version of January 1, 2009

[4] `[!#?%/&+,;]`

each other. By interpreting the LSR as a graph, we can identify three types of LSRs that differ in what the vertices represent. In a **lexical graph**, each vertex denotes exactly one *lexical item*. For example, there is exactly one vertex for the lexeme "order" even though this word has multiple meanings. **Disambiguated lexical graphs** represent each *reading* of every word as a separate vertex. In this case, there is a vertex for "order" as a command, and for "order" as a "ranking". Finally, a **semantic graph** represents each *meaning* as a vertex. In such a structure, there is exactly one vertex combining (for example) the common meaning of the lexemes "order" and "command".

We will describe and apply a multi-stage process to transform folksonomies to an LSR. We first create a *lexical graph* based on the tags in a folksonomy. Then, we split the lexical nodes with the help of community detection techniques to create a *disambiguated lexical graph*. As future work, we plan to apply graph clustering methods to transform the resulting graph to a *semantic graph*.

2.3 Creating a Tag Similarity Graph

As a first step, we create a lexical graph from tags, where each vertex represents a tag. Next, we calculate similarities between tags to create edges between similar tags. Cattuto *et al.* (2008) and Markines *et al.* (2009) summarize the state of the art in similarity measures on folksonomies. We choose one specific category from those distributional measures that represent tags in a vector space. These measures work with the disambiguation algorithm proposed in Section 2.4.

Resource Similarity. We create a vector space over all resources (*e.g.*, web sites) and represent each tag as the number of resources it has been assigned to by users. Then

$$\mathrm{sim}_{\mathrm{res}}(t_1, t_2) = \frac{v_{r1} \cdot v_{r2}}{|v_{r1}| \cdot |v_{r2}|}, \tag{1}$$

where $v_{rn} \in \{R\}$ is the vector space representation of all resources t_n has been assigned to.

Tag Context Similarity. The tag context similarity is a global co-occurrence similarity. Each tag is represented as the frequency of tags it appears with. This can be seen as a global co-occurrence measure. Then

$$\mathrm{sim}_{\mathrm{tag}}(t_1, t_2) = \frac{v_{t1} \cdot v_{t2}}{|v_{t1}| \cdot |v_{t2}|}, \tag{2}$$

where $v_{tn} \in \{T\}$ is the vector-space representation of all tags that have been assigned to resources where t_n has also been assigned.

Because each resource usually has multiple tags assigned by different users, feature vectors in this space are more dense than in the resource space. The tag context feature space can be limited by looking at only the N most popular tags. N has been chosen empirically as $N = 10\,000$ to achieve a good balance between performance and coverage.

User Similarity. This measure is similar to resource similarity but represents each tag by the users it has been used by:

$$\mathrm{sim}_{\mathrm{user}}(t_1, t_2) = \frac{v_{u1} \cdot v_{u2}}{|v_{u1}| \cdot |v_{u2}|}, \tag{3}$$

where $v_{un} \in \{U\}$ is the vector-space representation of all users that have been tagging the resources with t_n.

To create the tag similarity graph, we compute the neighborhood for each tag and create edges between the tag and its neighbor. The neighborhood is defined by a similarity threshold δ, after the application of which we can compute the average number of edges per vertex. Based on this we set the value of δ to 0.4 for the following experiments.

Table 1 shows the influence of the threshold on the average degree of nodes to the resulting graph for BibSonomy.

Table 1: Influence of δ on average degree of the nodes in the graph

δ	tag context similarity	resource similarity	user-similarity
0.4	99.51	6.20	96.06
0.5	74.53	3.70	85.51
0.6	42.84	2.38	72.46
0.7	18.34	1.60	58.47

2.4 Identifying Word Senses of Tags: Disambiguated Lexical Graphs

The second step is to transform the tag similarity graph to a disambiguated lexical graph. To do so, we first need to determine how many different senses a tag has and then to split each tag into new vertices representing those. Lastly, we need to modify the vector representation of each tag in a way that it only reflects its specific sense. Each ambiguous tag has a neighborhood of similar tags that have been calculated by one of the similarity measures discussed above. We assume that each neighboring tag belongs to one of the senses of the tag to be disambiguated. As discussed, each tag is represented by a feature vector. In the following procedure, for each sense of a tag, new feature vectors are created which contain only those features contributing to the specific sense.

Assuming we want to split tag t_i into a set of disambiguated tags, we define its *neighborhood* as

$$\text{neighborhood}(t_i) = \{t_j \in T | \, \text{sim}(t_i, t_j) > \delta\}$$

given a similarity measure sim() and a similarity threshold δ.

The threshold δ is chosen in such way that it produces a large neighborhood. In this case, the neighbors are not relations in a resource but are used as contexts for disambiguation. We then represent the tags in the neighborhood by their respective resource vectors and apply single-link hierarchical agglomerative clustering (HAC) (Manning *et al.*, 2008) on this neighborhood. We choose this algorithm because it is straightforward to implement, and because it does not require previous knowledge about the number of clusters like k-means does. The number of clusters determines the number of senses which is what we want to induce. This algorithm starts with each tag in its own cluster and iteratively merges the most similar clusters until a similarity threshold δ_{hac} is reached. For sparse feature spaces, such as in the resource domain, we have to choose $\delta_{hac} = 0$, as the neighbors are often only similar to t_0 but have only very low similarity to each other. By choosing $\delta_{hac} = 0$, we group all tags that have at least one feature in common. For tags t_1 and t_2 that are located in different clusters, the similarity $\text{sim}(t_1, t_2)$ is zero.

Now we apply the actual sense induction algorithm:

- For each cluster, we create a new tag t_n and assign it the feature-vector $\overrightarrow{f}$ of the original tag.
- For each cluster, we add its tags' feature-vectors together to create a masking vector $\overrightarrow{m}$.
- We set

$$f_i = \begin{cases} f_i & \text{if } m_i > 0 \\ 0 & \text{if } m_i = 0 \end{cases} \tag{4}$$

where f_i is the ith component of vector $\overrightarrow{f}$, and m_i is the ith component of vector $\overrightarrow{m}$. This way, the newly created tag is represented by only those features that are also present in the feature vectors of all the tags in the same cluster. Now we reconstruct the graph by applying the process described in Section 2 again to create a new, much larger tag similarity graph in which all senses of a tag are represented as a separate vertex.

Table 2: Graph-theoretic properties of the tag similarity graph based on BibSonomy

	$\|V\|$	$\|E\|$	$\#CC$	$\|V_{LCC_1}\|/\|V\|$	$\|V_{LCC_2}\|/\|V\|$	L	$\bar{k}$	C_{LCC}
bib-res	13,275	107,101	4941	0.63	< 0.01	3.23	12.85	0.19
bib-tag	13,275	91,518	157	0.99	< 0.01	2.41	40.00	0.40
bib-user	13,274	466,740	4902	0.63	< 0.01	3.31	56.55	0.31
bib-res-da	62,461	1,181,465	1572	0.96	< 0.01	4.29	19.53	0.57
bib-tag-da	21,160	91,101	2	1.00	< 0.01	3.04	58.80	0.42
bib-user-da	15,721	846,421	81	0.99	< 0.01	1.00	54.12	0.53
del-res	31,166	122,090	15,298	0.48	< 0.01	6.57	7.93	0.16
del-tag	32,724	2,327,591	111	0.99	< 0.01	3.45	40.00	0.36
del-user	15,766	416,688	7,253	0.54	< 0.01	0.00	48.99	0.25
del-res-da	133,497	1,567,909	17,003	0.82	< 0.01	6.02	14.15	0.39
del-tag-da	133,505	1,567,898	16,999	0.82	< 0.01	6.02	14.16	0.39
del-user-da	33,281	1,457,235	362	0.97	< 0.01	8.87	44.96	0.64
WiktG	20,011	33,650	8,214	0.57	< 0.01	5.03	5.80	0.0822
GNG	42,129	99,130	355	0.67	0.21	8,77	2.82	0.0155

3 Analyzing Network Properties of Tag Similarity Graphs

To investigate the potential value value of the generated resources, we now compare graph-theoretic properties to traditional LSRs. Topological analysis can help to understand the structure of the resource. Furthermore, it helps to understand if the tag similarity graph is structurally similar to conventional LSRs. We understand that manually crafted LSRs such as WordNet aim to be linguistically and psychologically sound, while the bottom-up characteristics of folksonomies are much more domain-specific and unbalanced. However, analyzing similarities and differences between those resources will help to understand if and how they could be used in combination in certain application scenarios. Last but not least, many LSR-based applications and algorithms operate on the graph, and they are likely to work with a resource based on a folksonomy if they work on a conventional, topologically similar resource.

Since the construction of tag similarity graphs is dependent on a set of empirically derived parameters as described in Section 2, it should be noted that many graph properties can be related to those parameters. If parameters are changed at this step, the graph-theoretic properties will change. In this work, we focus on an analysis of the automatically induced resources constructed with fixed parameters. An in-depth study of how the parameters are correlated with the resulting resources is left to future work.

Table 2 shows the results for tag similarity graphs induced from BibSonomy and Delicious. The table cover the graphs created by using three similarity measures as described in Section 2.3: resource (res), tag context (tag), and user (user) for both the tag similarity graph and the disambiguated version (-da). For comparison we include the values reported by Garoufi *et al.* (2008) for the German Wiktionary (WiktG) and GermaNet (GNG), the German wordnet.

The **size of the resource** in terms of the number of vertices ($|V|$) and edges ($|E|$) directly influences the coverage, and the ratio between vertices and edges is a measure for the density of lexical-semantic relations. While the Delicious-derived graphs are larger than BibSonomy in terms of vertices, they have more edges for each node. The number of **connected components ($\#CC$)** is the number of distinct subgraphs in which all vertices are connected to each other. A high number of CCs means a lot of nodes are not reachable from each other. If the largest connected component ($|LCC_1|$) is magnitudes higher than the second largest component ($|LCC_2|$), then LCC_1 can probably be used as a resource and all the small outliers may be simply ignored. If the CCs have all similar sizes, then the graph is fragmented, which hinders traversal-based algorithms. In our case, all graphs are highly connected and feature one, large set of nodes covering almost all nodes

aside from some outliers. This means that random-walk or graph-traversal algorithms should work properly on the folksonomy-induced resource proposed in this paper.

The following properties refer to LCC_1 of the respective graph only. The **average path length** L denotes the average distance between all vertices. It is an indication of whether the graph is a small-world network. The **average degree** $\bar{k}$ is the average number of all edges entering and leaving a node. The **network clustering coefficient** C_{LCC} is the average of all local clustering coefficients C_i. The local clustering coefficient is calculated as the proportion of all actual existing edges between the neighbor vertices of a vertex and the possible edges. BibSonomy is a much smaller resource than Delicious, even though we only use a fraction of the data available from Delicious. However, the structure in terms of connected components is about the same and depends only on the used similarity measure. The ratio between largest CC and second-largest is also similar. BibSonomy has a smaller average path length. The average degree is slightly larger in BibSonomy, the clustering coefficient is about the same. This means in terms of graph properties there is no large difference between the both datasets even though they are of such different size - the influence of the similarity measure is much larger. Tag and user similarity produce graphs with much higher average degrees and clustering coefficients. The disambiguated graphs are much larger, they have a smaller average degree and higher clustering coefficients. The smaller average degree comes from the fact that relations are now distributed between the newly created senses.

4 Analyzing the Vocabulary of Automatically-induced Tag Similarity Graphs

4.1 Vocabulary Coverage

Since the folksonomies under investigation can be freely edited by users without any editorial control, we expect to find large differences from the conventional resources like WordNet on the one hand and from collaboratively constructed dictionaries like Wiktionary on the other hand. Our expectations regarding the vocabulary found in folksonomies can be summarized as follows:

Mixed natural languages. Users can specify tags in any language they are comfortable with.

Free-form. Tags do not need to be constrained to natural language words; they can be any symbols which the users find useful.

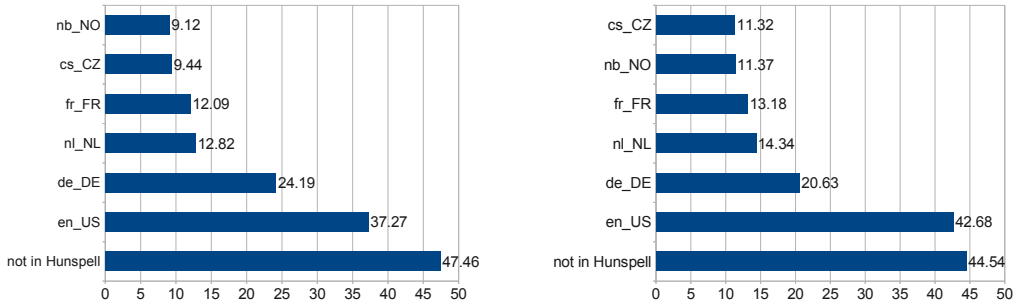
Creative. People can use neologisms such as novel portmanteau words.

To analyze the multilingual vocabulary of the folksonomies, we calculated the overlap with the Hunspell spelling dictionaries provided by OpenOffice.org[5] covering 36 languages. Figures 1a and 1b show that the most of the tags could not be assigned to any language. The most common language is English, followed by German, Dutch, French and Czech. Note that the coverage is overlapping and in some languages loanwords are more common.

Table 3 shows the term overlap between the folksonomy-derived resources and two common LSRs, WordNet and Wiktionary. A node is considered as overlapping if the strings are exactly the same. Because folksonomies are multilingual, and we make no attempt to separate languages, these statistics only indicate which part of the LSR's vocabulary is found in a folksonomy-induced resource and not the other way round. However, both the English Wiktionary and WordNet include some foreign-language vocabulary, which is why the coverage is higher than the coverage for US English detected by Hunspell. Table 4 shows the categories of folksonomy-specific tags we discovered in our analysis of the folksonomy vocabulary not found in other resources.

This lexical-semantic information is complementary to the knowledge in common LSRs and can be directly utilized to analyze user texts in order to achieve better coverage in automatic language analysis and induce semantic knowledge about user-specific vocabulary.

[5] `http://wiki.services.openoffice.org/wiki/Dictionaries`



(a) BibSonomy

(b) Delicious

Figure 1: Percentage of tags matched by Hunspell

Table 3: Percentage of folksonomy vocabulary that is also contained in WordNet or Wiktionary.

	BibSonomy	Delicious
WordNet	40.68	53.65
Wiktionary	58.13	61.59

4.2 Word Sense Distribution

Meyer and Gurevych (2010) study the distribution of senses per lexeme in Wiktionary and WordNet. We follow their approach to analyze the distribution of senses per lexeme and calculate $\omega_r(\ell)$, the number of word senses per lexeme (Table 5). The sense distribution $d_r(k) = |\{\ell|\omega_r(\ell) = k\}|$ specifies the number of occurrences of lexemes with a certain number of senses. While the majority of lexemes in WordNet 3.0 and in Wiktionary[6] (70% and 83%, respectively) have only a single sense, the sense distribution in the folksonomy-induced resources strongly depends on the employed similarity measure and other parameters, as described in Section 2.3. For example, user similarity on the BibSonomy dataset has a sense distribution that resembles that of WordNet and Wiktionary. If used with resource similarity, the average number of senses is much higher. Due to the sparseness of the resource space, many of a tag's neighboring tags feature no similarity to each other and will therefore end up in separate clusters yielding more senses.

Next, we turn to a qualitative analysis of polysemous lexemes found in the both generated disambiguated graphs from Section 2.3. Table 6 shows the number of senses for six lexemes

[6] The English edition from June 5, 2011, accessed using JWKTL 0.14.1 (Zesch *et al.*, 2008)

Table 4: Examples of words found in the folksonomy but not in Wiktionary

Category	**Examples**
abbrevations	cvs, pda, tdd, dhtml, nyc
concatenations	real_estate, bayarea, zipcode
meta-tags	mek:untagged, firefox:import
multiword expressions	milkandcookies, home_improvement,
spelling errors	acadamia
neologisms	hometheater, pilates, howto, etext, bioinformatics, portscan
named entities	junit
other	hamradio

Table 5: Sense distribution for different tag similarity graphs and other resources in percent

	bib-res	bib-user	del-res	del-user	WKT	WN
$d_r(1)$	15.45	83.75	44.81	62.99	70.16	82.68
$d_r(2)$	15.62	13.74	22.16	8.75	16.68	10.70
$d_r(3)$	14.68	2.12	12.74	26.83	6.49	3.33
$d_r(4)$	12.05	0.32	7.25	0.92	2.97	1.40
$d_r(\geq 5)$	42.20	0.08	13.04	0.50	3.71	1.88

Table 6: Number of senses in tag similarity graphs compared to WordNet and Wiktionary

	del-res	del-user	bib-res	bib-user	synsets in WN	senses in WKT
bank	5	1	6	1	18	12
life	5	3	5	1	14	14
order	8	2	8	1	24	16
sf	2	1	4	1	0	2
table	3	1	10	1	8	8
tree	9	2	9	1	7	10

(*table*, *bank*, *sf*, *order*, *tree* and *life*) in the Delicious and BibSonomy graphs constructed with the user and resource similarity measures, along with those for WordNet and Wiktionary.

We see that, for some of the selected words, there have been more senses created in BibSonomy, and for others more in Delicious. Thus, the number of senses does not seem to depend much on the size of the resource. Sense splitting using user similarity does not work at all in BibSonomy while it does produce some senses in Delicious. In comparison to WordNet and Wiktionary, the folksonomies produce quite different number of senses. For *table* and *tree*, for example, the folksonomy-derived LSRs have more senses than WordNet, but for *bank*, *life*, and *order* there are fewer. The abbreviation *sf* for "science fiction" or "San Francisco" is not contained in WordNet. To better understand the nature and granularity of word senses encoded in the analyzed resources, we look closer at the clusters for *sf* and *table* in the two resources derived from Delicious and BibSonomy. We choose resource similarity because besed on the resulting sense distribution this one looks most similar to the conventional resources.

The word *table* has three senses in Delicious, one of them refers to HTML tables, one of them refers to the periodic table, and one of them is an artifact which probably should have been merged into another cluster. The sense of *table* as a piece of furniture does not appear at all in the folksonomy. We see from Figures 2a-2d that the folksonomies are very unbalanced. Some topics are covered extensively, some meanings are not contained at all. If course, this depends on the way people use the Internet. The absence of the "furniture" meaning of *table* may come from the fact that furniture is one of the things people do not yet buy on-line so much and as a result do not annotate web sites about this topic.

5 Conclusions

In this paper, we presented experiments on inducing sense-disambiguated lexical-semantic networks based on two popular folksonomies, BibSonomy and Delicious. We apply a multi-step process to analyze the tags assigned by users to resources and construct a tag similarity graph which is further processed to induce word senses. We compare the resources resulting from a set of different system configurations with conventional lexical-semantic resources such as WordNet and Wiktionary. Regarding the graph-theoretic properties, we find that the induced resources have similar properties as traditional resources in terms of connectivity. Furthermore, we analyze the coverage of different resources and observe that folksonomies overlap in large part with existing

novel, via:boingboing, reality, charles, etext, culture, worldbuilding, webzine, utopia, ebook, starwars, science.fiction, popularmedia, writing, lovecraft, zine, read, writing, short, horror, california, transhumanism, future, science-fiction, author, singularity, sanfrancisco, fantasy, science-fiction, science, sci-fi, sciencefiction, fiction, scifi
francisco, bay, san, california, bayarea, sanfrancisco

(a) *sf* in Delicious

chemistry, elements
tables
sequence, sequences
robot
convert, latex, converter, openoffice,
data, presentation, liste, table, elements, year2006, tree, graphs, informationdesign, visualization, übersicht, phawk, methoden, knowledgevisualization

(b) *table* in Bibsonomy

writing, storytelling, fiction, tricks, plot, scifi, stories, fiction, sciencefiction
frameworks, codegeneration, sourceforge

(c) *sf* in Bibsonomy

webauthoring, webdevelopment, sorted, programming/javascript, data, web, example, cssgalleries, cssgalleries, web, dev, web.design, voorbeeld, css.example, xhtml,js
table, elements, periodictable, periodic, chemistry
absolute

(d) *table* in Delicious

Figure 2: Induced senses of the tags *sf* and *table*

resources but also contribute vocabulary that is not included in them. Finally, we compare the word senses encoded in the individual resources and find that the proposed sense induction algorithm detects many of the senses that are encoded in the folksonomy, although due to the folksonomies' nature, they differ substantially from the senses in traditional LSRs.

In summary, the folksonomy-induced resources show good potential to complement the knowledge in conventional resources because they directly encode information regarding non-standard terms frequently found in user-generated content and lacking in conventional resources. The resulting lexical-semantic knowledge can be directly utilized for example in query expansion, or for personalized information management in the context of Web applications. For such applications we plan to look into how to couple folksonomies with traditional LSRs to extend their domain-specific vocabulary.

An interesting by-product of our research is a parameterizable framework for word sense induction. Such a framework is not limited to folksonomies, but can be applied to any network based on term distributions, such as those based on n-gram statistics from the Web or other corpora or the co-occurrence of terms in a semantically structured corpus such as Wikipedia. While current methods, such as Explicit Semantic Analysis (Gabrilovich and Markovitch, 2007), operate on the term co-occurrence vectors lacking sense disambiguation, our framework can be applied to construct a sense-disambiguated resource. We plan to explore this direction in the future. Another important issue to systematically study is the relation between empirically set parameters for similarity graph construction and the properties of the resulting lexical-semantic networks. While we experiment with a set of predefined parameters in this paper, we plan to conduct an in-depth study of this issue in the future. Utilizing the resulting resources in NLP applications as mentioned above is another important avenue of research.

6 Acknowledgements

We thank Tristan Miller and Christian M. Meyer for their helpful comments. This work has been supported by the Volkswagen Foundation as part of the Lichtenberg Professorship Program under grant № I/82806.

References

Benz, Dominik, Andreas Hotho, Robert Jäschke, Beate Krause, Folke Mitzlaff, Christoph Schmitz, and Gerd Stumme. 2010. The Social Bookmark and Publication Management System BibSonomy. *The VLDB Journal*, 19(6), 849–875.

Cattuto, Ciro, Dominik Benz, Andreas Hotho, and Stumme Gerd. 2008. Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In *Proceedings of the 7th International Conference on The Semantic Web*, ISWC '08, pp. 615–631, Berlin, Heidelberg. Springer-Verlag.

Gabrilovich, Evgeniy and Shaul Markovitch. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pp. 6–12.

Garoufi, Konstantina, Torsten Zesch, and Iryna Gurevych. 2008. Graph-theoretic analysis of collaborative knowledge bases in natural language processing. In Christian Bizer and Anupam Joshi, eds., *International Semantic Web Conference (Posters & Demos)*.

Girvan, Michelle and Mark E J Newman. 2002. Community Structure in Social and Biological Networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12), 7821–7826.

Gurevych, Iryna and Elisabeth Wolf. 2010. Expert-Built and Collaboratively Constructed Lexical Semantic Resources. *Language and Linguistics Compass*, 4(11), 1074–1090.

Heymann, Paul and Hector Garcia-Molina. 2006. Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Technical Report 2006-10, Stanford InfoLab.

Jurgens, David. 2011. Word Sense Induction by Community Detection. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics Human Language Technologies*, pp. 24–28. ACL.

Manning, Christopher D, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Markines, Benjamin, Ciro Cattuto, Filippo Menczer, Dominik Benz, Andreas Hotho, and Stum Gerd. 2009. Evaluating Similarity Measures for Emergent Aemantics of Social Tagging. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pp. 641–650, New York, NY, USA. ACM.

Meyer, Christian M. and Iryna Gurevych. 2010. How web communities analyze human language: Word senses in wiktionary. In *Proceedings of the Second Web Science Conference*, Raleigh, NC, USA.

Schmitz, Patrick. 2006. Inducing Ontology from Flickr tags. In *Proceedings of the Collaborative Web Tagging Workshop at WWW '06*, Edinburgh.

Vander Wal, Thomas. 2004. Folksonomy Coinage and Definition. `http://vanderwal.net/folksonomy.html`.

Wetzker, Robert, Carsten Zimmermann, and Christian Bauckhage. 2008. Analyzing Social Bookmarking Systems: A del.icio.us Cookbook. In *Proceedings of the ECAI 2008 Mining Social Data Workshop*, pp. 26–30. IOS Press.

Yeung, Ching Man Au, Nicholas Gibbins, and Nigel Shadbolt. 2009. Contextualising Tags in Collaborative Tagging Systems. In *20th ACM Conference on Hypertext and Hypermedia*, Torino, Italy. ACM Press.

Zesch, Torsten, Christof Mueller, and Iryna Gurevych. 2008. Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation*. (electronic proceedings).