08391 – Working Group Summary

Analyzing Tag Semantics Across Collaborative Tagging Systems

— Dagstuhl Seminar —

Dominik Benz[1], Marko Grobelnik[2], Andreas Hotho[1], Robert Jäschke[1,3], Dunja Mladenic[2], Vito D. P. Servedio[4], Sergej Sizov[5], Martin Szomszor[6]

[1] Knowledge & Data Engineering Group, University of Kassel, Germany
[2] Department of Knowledge Technologies, J. Stefan Institute, Ljubljana, Slovenia
[3] Research Center L3S, Hannover, Germany
[4] "La Sapienza" University of Rome, Italy
[5] Information Systems and Semantic Web, University of Koblenz-Landau, Germany
[6] University of Southampton, England

Abstract. The objective of our group was to exploit state-of-the-art Information Retrieval methods for finding associations and dependencies between tags, capturing and representing differences in tagging behavior and vocabulary of various folksonomies, with the overall aim to better understand the semantics of tags and the tagging process. Therefore we analyze the semantic content of tags in the Flickr and Delicious folksonomies. We find that: tag context similarity leads to meaningful results in Flickr, despite its narrow folksonomy character; the comparison of tags across Flickr and Delicious shows little semantic overlap, being tags in Flickr associated more to visual aspects rather than technological as it seems to be in Delicious; there are regions in the tag-tag space, provided with the cosine similarity metric, that are characterized by high density; the order of tags inside a post has a semantic relevance.

Keywords. Social Web Communities, Folksonomy, Tag, Semantics

1 Introduction and Motivation

Popularity and data volume of modern Web 2.0 content sharing applications originate in their ease of operating for even unexperienced users, suitable mechanisms for supporting collaboration, and attractiveness of shared annotated material. Understanding of annotation properties is crucial for constructing accurate and efficient navigation and browsing mechanisms, including content recommendations (favorites), ranked retrieval of relevant items for user queries, or user assistance in annotating new contents (tag recommendation). For this reason, the discussion in our focused group was centered around the semantic grounding of tagging. Our objective was to exploit state-of-the-art Information

Dagstuhl Seminar Proceedings 08391
Social Web Communities
http://drops.dagstuhl.de/opus/volltexte/2008/1785

Retrieval methods for finding associations and dependencies between tags, capturing and representing differences in tagging behavior and vocabulary of various folksonomies, with the overall aim to better understand tags and the tagging process.

2 Description of Flickr and Delicious

Since few years, social bookmarking has become of common use in the world of internet. The vast spread of personal computers, laptops, high performance mobile telephones, has lead to the necessity of users to upload their bookmarks, documents, photos or general resources, onto a remote web server easily accessible from any electronic device with an internet connection. But the very revolution brought by social bookmarking is the open possibility of users to label their uploaded resources with one or more semantically meaningful words, called "tags" by the internet community. The procedure of tagging is therefore a selfish process that users willingly carry on for their personal use in order for them to render the information retrieval process (e.g., the visualization of a particular photo) the easiest as possible. Tags carry semantic informations that characterize the resource tagged, but that, from another point of view, characterize tags themselves by defining a precise semantic context of use. Tags are exposed to all users, and as a result, may be used for purposes of searching for resources in a way more oriented towards the semantic aspect rather than the brutal analysis of resource content. The possibility to access to the semantic aspect of resources is clearly much more crucial in the particular case of picture retrieval since, nowadays, there exists no software capable of extracting the semantic informations embedded in a picture in a unsupervised way.

In the present work we shall analyze two major folksonomy systems:

Delicious,[1] in which we used data collected in November 2006. In total, data from 667,128 users of the Delicious 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 Delicious, and to the resources/users that have been associated with at least one of those tags. 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.

Flickr,[2] a large-scale reference data set obtained by systematically crawling the Flickr portal during 2006 and 2007. The target of the crawling activity were the core elements of a folksonomy: the users, tags, resources and tag assignments. We also gathered additional information about the interests of the users. The additional information included the contact list of the users, their comments to photos, their favorite photos and memberships in user groups. The full dataset consists of $|U| = 3,074,947$ users, $|T| = 5,556,568$ tags, $|R| = 41,278,715$ resources, and $|Y| = 187,168,654$ tag assignments.

In addition, informations about 29,842,973 publicly visible user contacts, 50,058,103 personal favorites, and 132,816 groups (13,243,481 group members in total) were gathered.

Delicious is considered as a broad folksonomy, where each user may tag any resource (URL), while *Flickr* as a narrow folksonomy, where only the user who owns the resource (a picture) may tag it, or allow a list of friends to do so.

3 Cross-Folksonomy Analysis

With the growth of Web 2.0, it is becoming increasingly common for users to maintain a presence in more than one folksonomy site. For example, one could bookmark pages in Delicious, publish images with Flickr, record music listening habits with Last.fm, blog in Technorati, etc. In a recent survey, Ofcom found that 39 % of UK adults with at least one social networking profile have indeed two or more profiles [1]. It has even been predicted that by 2010, each of us will have between 12 and 24 online identities [2]. Hence, the issue of understanding the similarities and differences in tagging behaviour across such folksonomies will become increasingly significant. In recent work, cross-folksonomy analysis has been used to automatically illicit and model user interests based on a user's interaction with various folksonomy sites [3]. Tag filtering has also shown to be an important factor when considering the consolidation of resources tagged in different domains [4].

In this section, we descibe experiments undertaken using a subset of Delicious and Flickr where users' profiles in each domain have been correlated using Google's OpenSocial API[3] (as described in [3]). The data set consists of $|U_d| = 2,045$ Delicious users, $|U_f| = 2,045$ Flickr users, $|T_d| = 185,400$ Delicious tags, $|T_f| = 341,908$ Flickr tags, $|R_d| = 341,908$ Delicious resources, $|R_f| = 2,534,467$ Flickr resources.

The first observation to be made is that 50,264 of the tags occur both in Delicious and Flickr (27.1 % and 14.7 % of the respective totals). Figure 1 contains a histogram of the tags found in Delicious that also appear in Flickr vs. those that do not. Tags are grouped on the horizontal axis according to their frequency (the most frequent tags on the left, tags that appear only once on the right). The group containing tags with a frequency between 10,000 and 100,000 (i.e., the most common terms) is almost entirely represented in Flickr (except for one tag). If a tag is used less frequently in Delicious, it is less likely to appear in Flickr. Following this observation, we continued to investigate the similarity in the way tags that intersect both domains are used. Previous work [5] has defined *tag context similarity* (in which a tag is described by its co-occurrence-vector against other tags) as a good method for establishing semantically similar terms. We analyzed all tags in this data set appearing in both Flickr and Delicious, recording

[3] `http://code.google.com/apis/socialgraph/`

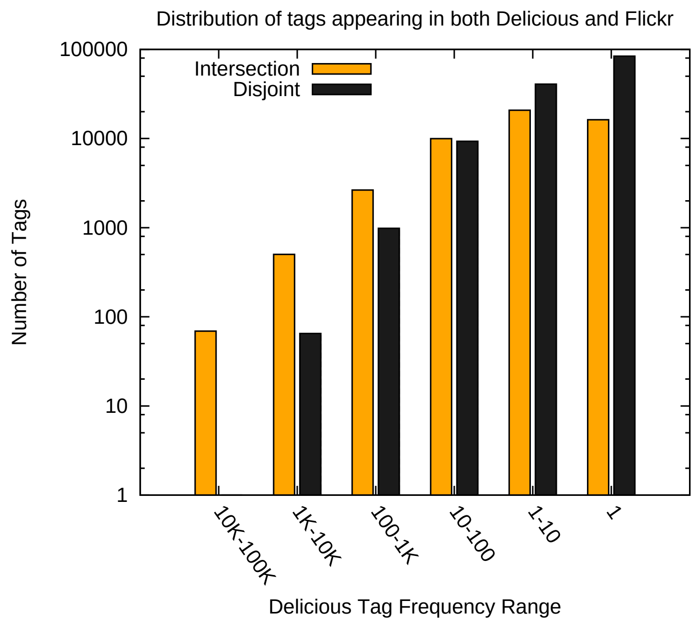

Fig. 1. A histogram showing the distribution of tags that appear in both Delicious and Flickr (intersection), and those that appear only in Delicious (disjoint).

the similiarity between co-occurrence-vectors. Table 1 summarises the 40 most similar tags.

Most of these tags have a low similarity, *macintosh* has the highest with 0.7718. There is no discernible pattern in the distribution: one might expect more general terms or those corresponding to unambiguous concepts would have the highest similarity. A unique insight that can be gained from this particular dataset is how individuals use particular tags in Delicious and Flickr. Tag context similarity can be applied at a user level by comparing the user's co-occurrence-vector. To establish whether high tag context similarity can be used as a meaningful semantic measure between tags in Delicious and Flickr, we examined use of the tag *apple*. Due to the technology bias, the tag *apple* is likely to be used in Delicious to represent the concept of the computer company. In Flickr, both the computer company is featured (i.e. with pictures of Apple computers, Apple OS screenshots, or Apple stores), and the fruit. The table below shows the 4 most frequently co-occurring tags for the user with the highest similarity for *apple* (0.9316).

Table 1. The 40 most similar tags.

Tag	Similarity	Delicious Freq	Tag	Similarity	Delicious Freq
macintosh	0.7718	1,460	drawing	0.6045	1,649
vista	0.7459	1,524	furniture	0.6041	2,080
osx	0.7329	18,976	ubuntu	0.6013	4,803
mozilla	0.7262	2,911	inspiration	0.6009	16,511
xhtml	0.7242	4,348	cinema	0.5971	1,659
bands	0.7200	1,209	iphone	0.5968	3,509
nintendo	0.7046	2,008	sanfrancisco	0.5964	1,438
recipe	0.6945	1,924	band	0.5917	1,054
painting	0.6924	1,030	ipod	0.5836	6,136
portfolio	0.6782	8,309	live	0.5819	1,253
guitar	0.6693	1,119	recipes	0.5735	4,254
wii	0.6632	1,228	indie	0.5722	2,058
rock	0.6630	1,090	html	0.5609	9,615
cooking	0.6456	3,045	atom	0.5600	1,119
newyork	0.6287	1,283	feeds	0.5597	1,787
restaurants	0.6274	1,063	gnu	0.5538	1,158
rubyonrails	0.6214	3,927	satire	0.5425	1,096
cellphone	0.6207	1,253	feed	0.5383	1,525
apple	0.6171	17,277	javascript	0.5377	23,429
logo	0.6152	1,343	playlist	0.5373	1,159

Co-occurring Tag	Delicious Weight	Co-occurring Tag	Flickr Weight
apple	174	apple	74
osx	45	screenshot	17
mac	36	osx	17
ipod	30	mac	14

In this case, it is clear that *apple* is used in both Delicious and Flickr to refer to the same concept. However, choosing the individual with the lowest similarity (0.3838) gives a different picture:

Co-occurring Tag	Delicious Weight	Co-occurring Tag	Flickr Weight
apple	48	goldendelicious	40
mac	20	fruits	36
ipod	21	food	20
osx	16	apples	20

This particular user is using *apple* in Delicious to refer to the computer company, but in Flickr, it is used to refer to the fruit.

4 Cosine Similarity Between Tags

Prior work on analyzing collaborative tagging systems has given evidence for emergent semantics [6,7]. Cattuto et al. [5] characterized several measures of tag relatedness. Tag context similarity (whereby each tag was described by its co-occurrence-vector v_i with other tags) provided the most precise semantics hereby. For computing the similarity between two tag feature vectors v_1 and v_2 we use the common notion of IR-style cosine measure, involving the euclidean scalar product:

$$\sim (v_1, v_2) = \frac{v_1 \cdot v_2}{||v_1|| \, ||v_2||} \tag{1}$$

Their analysis was based on a dataset containing the 10,000 most popular tags from Delicious (crawled in 2006), along with all users and resources connected to at least one of those tags in the folksonomy graph.

Having a dataset of the social photo sharing platform Flickr at hand, these results inspired us to the following research questions:

1. Does tag context similarity also convey meaningful results in a narrow folksonomy like Flickr?
2. To which extent does the semantics of Delicious and Flickr differ?
3. Is it possible to refine the semantics of certain tag relations by restricting the dataset to a thematically focused user group?

To answer the first two questions, we computed the tag context similarity as defined in [5], based on the 10,000 most popular tags in Flickr and once again all users and resources associated with at least one of these tags. Table 2 summarizes the 10 most similar tags based on the full tag context for Delicious (*del_full*) and Flickr (*flickr_full*).

The last research question was led by the intuition that a thematically restricted user group might yield even more precise tag similarities for tags which belong semantically to this group. Flickr offers the possibility for users to create groups;[4] we selected a user group concerned with Scottish castles.[5] Our expectation was that tag context similarity restricted to this group (i.e., only retaining tags and resources that belong to at least one group member) would yield rather precise semantic relations for tags like *castle*, *scotland*, *loch* and comparable ones. The restricted dataset contained all postings from 360 group participants, using in total 87,383 tags. The 10 most similar tags are again summarized in Table 2 (*flickr_restricted*).

The first impression is that the tag context similarity also seems to yield meaningful results for a narrow folksonomy like Flickr. For the tag *bug*, as an example, different types of bugs (*wasp, hoverfly, grasshopper, ...*) are deemed similar. The next obvious thing is that the two folksonomies (Delicious and Flickr) nicely disambiguate tags like *bug*, *windows* and *net*. Delicious is known to have a strong focus towards technophile users, hence this is not too suprising. But it gives evidence that the tag context measure is able to help disambiguating terms when applied to appropriate folksonomies.

The influence of the restricted dataset is not clearly visible. This is probably due to the way how we extracted the dataset. In fact, we included all tags

[4] see `http://flickr.com/groups/`
[5] see `http://flickr.com/groups/scotlands_castles/`

Table 2. Examples of most related tags by tag context similarity (see [5]) based on different folksonomies and restricted to a user group. *del_full* is based on the same dataset as [5], *flickr_full* is based on a comparable dataset from Flickr, and *flickr_restricted* contains only data from a user group on Scottish castles.

tag	**sim measure**	**1**	**2**	**3**	**4**	**5**
bug	*del_full*	bugs	msie	ie6	ie7	internetexplorer
	flickr_full	wasp	hoverfly	grasshopper	dragonfly	insecte
	flickr_restricted	grasshopper	wasp	dragonfly	bugs	beetle
castle	*del_full*	division	duck	forest	battle	kent
	flickr_full	geotagged	church	europe	village	palace
	flickr_restricted	0506	skye	highlands	glasgow	highland
scotland	*del_full*	wales	england	cornwall	london	britain
	flickr_full	meadows	hill	view	hills	geotagged
	flickr_restricted	tree	grass	light	silhouette	water
windows	*del_full*	utilities	utility	opensource	open_source	freeware
	flickr_full	structure	roof	facade	window	balcony
	flickr_restricted	window	roof	wall	architecture	old
net	*del_full*	internet	sites	services	www	service
	flickr_full	rope	fisherman	fishermen	sunny	wind
	flickr_restricted	trawler	nets	macduff	boat	harbour
low	*del_full*	affordable	cost	reduce	costs	fast
	flickr_full	shadow	lamp	dark	lamp	shadows
	flickr_restricted	stonefaction	morayshire	faved	fife	personalfave

and pictures of the group members, so that content not related to Scotland or castles may have tainted the analysis. Further restricting the content (e.g., only to pictures explicitely assigned to this group) might lead to a less blurred picture.

5 Analysis of Tag-Tag Space

In this section we used the same tag vector representation as in Section 4 which is based on its co-occurring tags on a post level. Now we are interested in a deeper analysis of the resulting tag-tag space. We based this analysis on the Delicious dataset. Figure 2 shows the visualization of the tag-tag space for Delicious, where tag similarity is reflected in the relative distance between the tags as proposed in [8]. Lighter parts in Figure 2 show regions of higher density, containing tags such as, *webdevelopment*, *resources*, *downloads* (on the left hand-side), *audio*, *music*, *hobby* (in the middle) or *tourism*, *transport*, *fishing*, *bike* (in the right hand-side). Somewhat less frequent are items tagged as *accounting*, *mortgage*, *cash* (at the top darker region) or *handmade*, *clustering* (at the bottom).

Figure 3 zoomed-in on the right part with tags related to tourism and transportation, where we can see tags for different touristic destinations such as, italy, taiwan, asia, spain, disney, tokyo, newyork, as well as tags on globalization, politics or tags on cancer.

The analysis of the Delicious tag-tag space was performed using the OntoGen tool for semi-automatic data driven ontology construction that enable interactive browsing through the constructed ontology and several visualizations of the un-

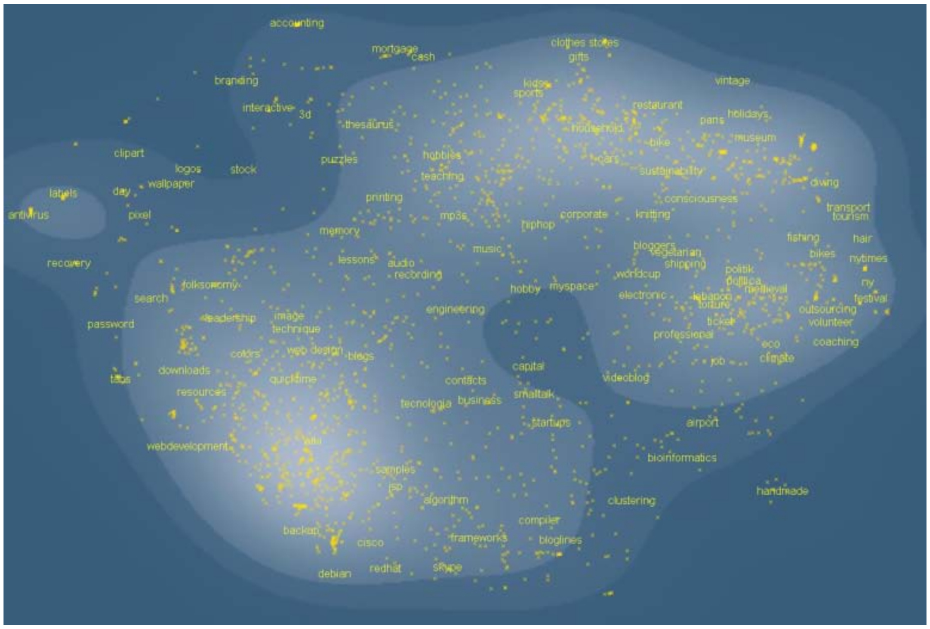

Fig. 2. Visualization of Delicious tag-tag space on 10, 000 most frequent tags.

derlying instances. OntoGen [9,10] is a semi-automatic and data-driven ontology editor focusing on editing of topic ontologies. The system combines text-mining techniques with an efficient user interface in order to optimize both computation time and user friendliness. In this way it bridges the gap between complex ontology editing tools and domain experts, who deal with ontology construction and not necessarily have the skills of ontology engineering. The tool combines several knowledge discovery approaches including handling textual data and representing them as vectors, automatic discovery of concepts, adding topics to the ontology, concept naming, incorporating domain knowledge, data visualization, addition of new instances, ontology contextualization and overview of the ontology through concept browsing and various kind of visualization.

Figure 4 shows topic ontology constructed on the Delicious tag-tag space, showing different groups of tags such as, *Food/Travel*, *sysadmin/ip/ubuntu*, *tagging/seo/del.icio.us*, *ajax*, *pictures/kids*, *Politics/Finances*, *Car/Energy*, etc. We have split some of the groups into smaller groups, just for illustration, in general the structuring can go all the way to single tags. Furthermore, each group of tags can be visualized as a group as shown in Figure 5. From the visualization of tags about *food* and *travel* we can see that they can be split into a big group and a smaller group. The big group of tags contains two subgroups, one on *travel* (right hand-side) containing tags such as *airplane*, *vegas*, *Portugal*, *holiday*, *sanfrancisco* and the other group on *food* (left bottom hand-side) containing tags

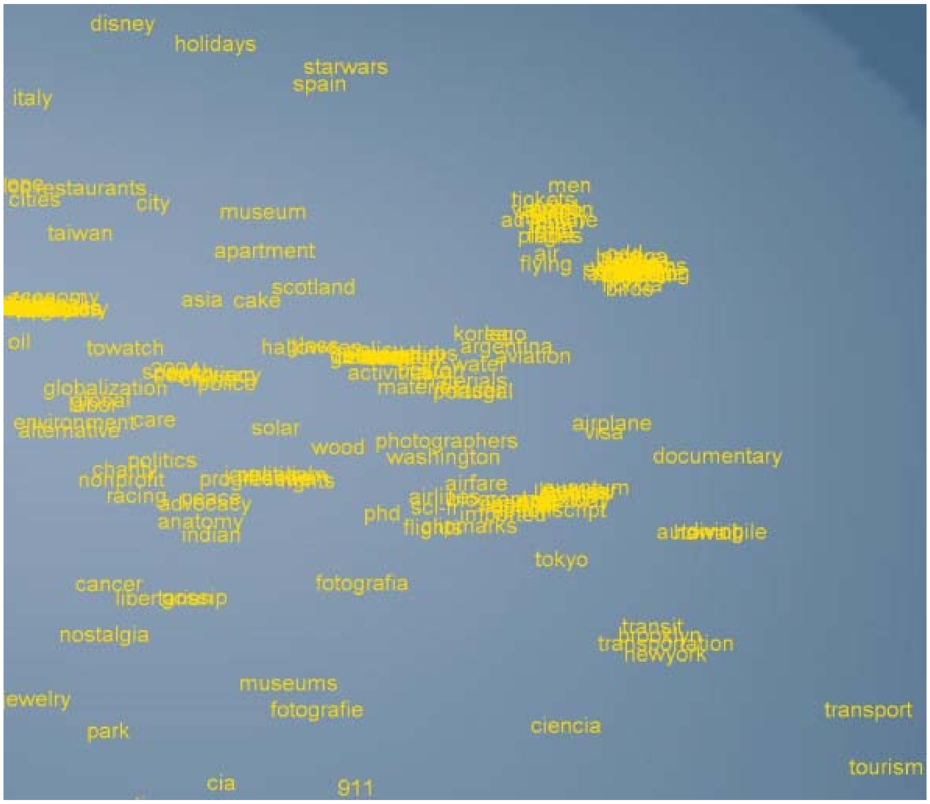

Fig. 3. Visualization of a part of Delicious tag-tag space zoomed-in on the right part of Figure 2

such as *recipes*, *alcohol*, *Indian*, *organic*, etc. The small group of tags (left top hand-side) contains tags such as *medical*, *insurance*, *yoga*, *body*, etc.

6 Directed Co-occurrence Network

Users annotate resources with posts by adding tags in a certain order. The tag-tag co-occurrence graph presented in Section 4 does not take into account tag order inside posts. In order to establish whether tag ordering has a kind of semantic relevance, we construct and analyze a weighted, *directed* co-occurrence network that fully encodes the order of tags inside posts. Figure 6 shows with an example how we construct this graph. Each tag points to its following tag, i.e., in the example, the tag *bibtex* precedes the tag *bibliography* and defines the directed edge (*bibtex*, *bibliography*). More formally, we construct the graph $G = (V, E)$ with $V = T$ and $(t_1, t_2) \in E$ iff there exists a post (u, T_{ur}, r) with

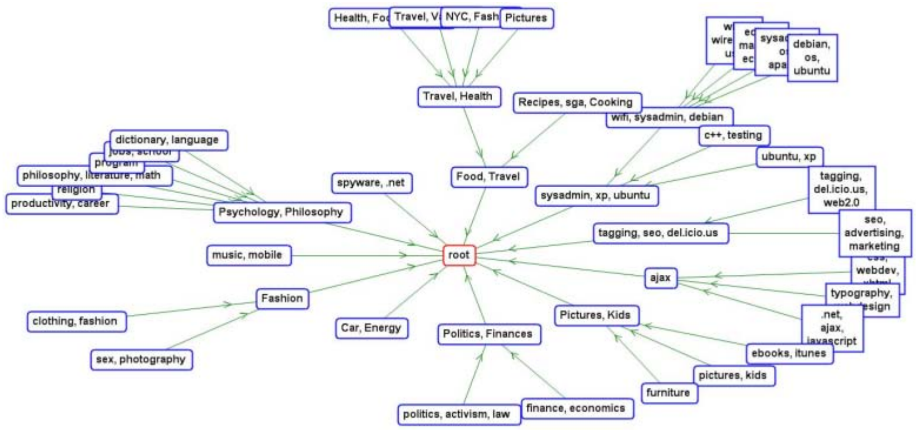

Fig. 4. Topic ontology of Delicious tag-tag space generated using OntoGen tool.

$t_1, t_2 \in T_{ur}$ and t_1 appears before t_2 in the post. The *weight* $w(t_1, t_2)$ of this edge is given by counting in how many posts t_1 appears before t_2:

$$w(t_1, t_2) := |\{(u, r) \in U \times R \mid t_1, t_2 \in T_{ur}, t_1 \text{ appears before } t_2\}|$$

In order to understand whether tag ordering in posts is a semantic relevant feature and not the result of a random uncorrelated process, we exchanged randomly the position of tags inside single posts and create a new fictitious folksonomy. This simple randomization process will result in a directed weighted tag-tag co-occurrence graph where correlations among tags in the same posts are artificially destroyed. We expect that tags appearing preferentially at the beginning or at the end of a post are most influenced by the process of shuffling.

One of the simplest measures that can be performed in directed graphs is the correlation between in-degree and out-degree of nodes. In our case the network is also weighted so that the important quantities in this case are the in-strength and out-strength of nodes, whereby strength it is meant the sum of corresponding edge weights. The scatter plot shown in Fig. 7 shows the correlation of the in-strength of tags versus their respective out-strength. We notice that the shuffling process narrows the original picture, thus showing that the order of tags in posts is important and not the result of a random process. We conjecture that the average user annotates resources by using tags with increasing (or decreasing, it is not clear at this stage) degree of generality in the posts.

In order to unravel similarities among tags and get semantic informations, we can define a cosine similarity measure following the ideas of [5]. In our case, the order of tags in a post is important and the cosine similarity measure has to be extended to directed networks where the adjacency matrix A ($A = (a_{ij})_{1 \le i,j \le |T|}$ with $a_{ij} = w(i, j)$) is no more symmetric. We may consider different quantities as explained in Table 3.

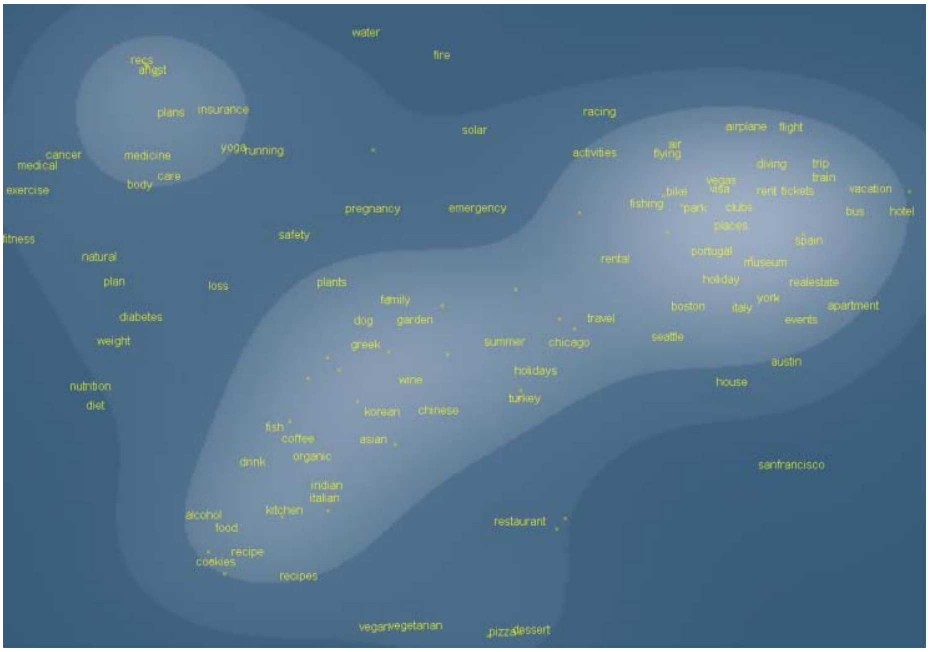

Fig. 5. Visualization of the tag group Food, Travel from Figure 4.

According to the first (second) measure, two tags are similar, if they are often predecessors (successors) of the same set of tags. The third measure is different: two tags t_i and t_j are similar, if t_i frequently is a predeccessor of tags whose successor is t_j. By construction, $M(t_i, t_j)$ is not symmetric and its symmetric form can be easily introduced by taking the arithmetic average with its transposed: $M_s(t_i, t_j) = \frac{1}{2}(M(t_i, t_j) + M(t_j, t_i))$.

For two of the measures (L and R) in Table 3, we computed the most related tags for each tag in the dataset used in [5]. Table 4 presents some example results. One can observe a substantial difference in some cases – e.g., for *java*, the L-measure yields exclusively other programming languages as similar tags, while the R-measure also contains some more general tags like *code* or *refactoring*. A first natural step to characterize is to measure the overlap of similar tags; Table 5 summarizes the results. The relatively low overlap in all cases confirms the im-

$\{(bibtex, bibliography), (bibliography, software), (software, java), (java, tools)\}$

Fig. 6. A screenshot of a post in Delicious and the corresponding set of edges as defined for the directed co-occurrence network.

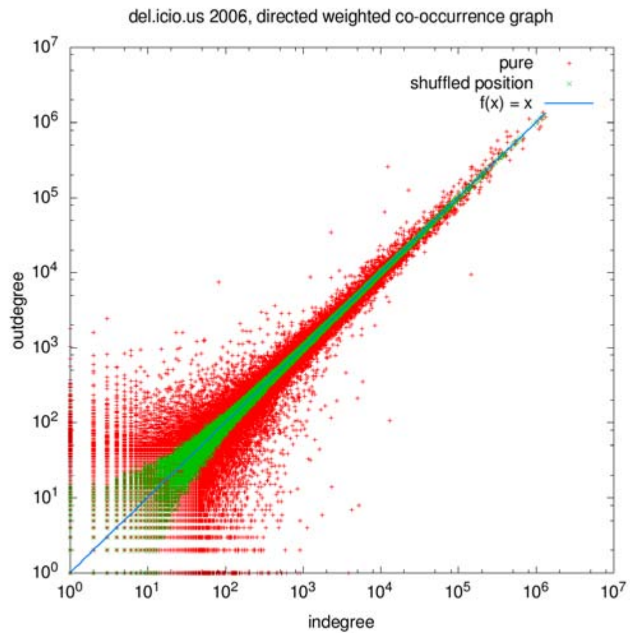

Fig. 7. Correlation of in-strength and out-strength on the directed, weighted tag-tag–co-occurrence graph of the Delicious dataset. Each + represents a tag in the original graph; each × represents a tag in the graph resulting after shuffling tags inside single posts.

pression that there is a substantial difference in the type of similarity described by the different underlying co-occurrence networks. Further work (similar to the analysis performed in [5]) is required to characterize more precisely the different characteristics.

Table 3. Similarity measures between tags in the weighted, directed co-occurrence graph. The vector $a_{i*} \in \mathbb{R}^{|T|}$ depicts the i-th row of A, and $a_{*j} \in \mathbb{R}^{|T|}$ the j-th column.

similarity measure	corresponding order of tags in post
$L(t_i, t_j) = \mathrm{sim}_{\substack{t_i \rightarrow \\ t_j \rightarrow}}(t_i, t_j) := \frac{a_{i*} \cdot a_{j*}}{\lvert\lvert a_{i*} \rvert\rvert \cdot \lvert\lvert a_{j*} \rvert\rvert}$	$\ldots, t_i, t_\alpha, \ldots$ $\ldots, t_j, t_\alpha, \ldots$
$R(t_i, t_j) = \mathrm{sim}_{\substack{\rightarrow t_i \\ \rightarrow t_j}}(t_i, t_j) := \frac{a_{*i} \cdot a_{*j}}{\lvert\lvert a_{*i} \rvert\rvert \cdot \lvert\lvert a_{*j} \rvert\rvert}$	$\ldots, t_\alpha, t_i, \ldots$ $\ldots, t_\alpha, t_j, \ldots$
$M(t_i, t_j) = \mathrm{sim}_{\substack{t_i \rightarrow \\ \rightarrow t_j}}(t_i, t_j) := \frac{a_{i*} \cdot a_{*j}}{\lvert\lvert a_{i*} \rvert\rvert \cdot \lvert\lvert a_{*j} \rvert\rvert}$	$\ldots, t_i, t_\alpha, \ldots$ $\ldots, t_\alpha, t_j, \ldots$

Table 4. Examples of most related tags for different co-occurrence networks. L and R are the similarity measures described in Table 3, U is the cosine similarity based on the undirected co-occurrence network (see [5])

tag	sim measure	1	2	3	4	5
web2.0	U	web2	web-2.0	webapp	“web	web_2.0
	L	webtool	webtools	2.0	webapp	webapps
	R	web2	web-2.0	application	toolkit	www
java	U	python	perl	code	c++	delphi
	L	python	c	delphi	lisp	fortran
	R	code	perl	.net	python	refactoring
metadata	U	ontology	taxonomy	classification	tags	folksonomies
	L	semantic_web	semanticweb	semweb	ontology	topicmaps
	R	classification	taxonomy	folksonomies	tag	ontology
album	U	picture	albums	foto	photograph	image
	L	abstract	fine	ascii	sculpture	deviantart
	R	photograph	picture	foto	photography	albums
germany	U	europe	berlin	france	italy	world
	L	eu	europa	austria	uk	italy
	R	deutsch	frauen	austria	heise	switzerland

Table 5. Average overlap of the 10 most similar tags according to the cosine similarity, based on two different kinds of co-occurrence networks definitions. U represents the cosine similarity calculated with the undirected co-occurrence network; L and R are cosine similarity measures pertaining to the directed co-occurrence network and are defined in Table 3.

$L-U$	$R-U$	$L-R$	$L-R-U$
3.86	4.11	3.33	2.52

7 Conclusions

This paper summarizes the work of the “Tag Semantics” Dagstuhl working group. While analyzing data from Delicious and Flickr in various ways we made the following interesting findings:

While many of the frequently occurring Delicious tags also appear in Flickr, applying the tag context similarity measures at a global scale does not give exciting insights. However, comparison of an individual’s co-occurrence network could be used to some extent to measure whether ambiguous terms are used with the same sense. Such measures are noisy and do not provide stable results. Improvement might be made by filtering the tags so morphological variations and synonyms are merged.

We performed the analysis of tag context similarity in the narrow folksonomy of Flickr and confirmed the result obtained for Delicious in a previous work. We find that tags in Flickr are obviously oriented towards their visual meaning, whereas in Delicious they are biased more towards their technical meaning. Moreover, we restricted the analysis of tag context to those users belonging to the same group of interest and found no particular variations in tag similarities with respect to the unrestricted set.

We embedded in a three dimensional space the representation of the tag-tag space with the cosine similarity metric by means of the software OntoGen. We were able to navigate in such space and find regions of high similarity density, where the cosine similarity distance between tag pairs is higher than the average.

Finally, by constructing a *directed* tag-tag co-occurrence network, in which nodes represent tags and links connect two adjacent tags inside a post from left to right, we showed that tag order in posts has a relevant semantic value.

8 Acknowledgements

This research was supported by the European Union in the Tagora project (FP6-2005-34721) and the Nepomuk project (FP6-027705).

References

1. Ofcom: Social networking: A quantitative and qualitative research report into attitudes, behaviours, and use. `http://news.bbc.co.uk/1/shared/bsp/hi/pdfs/02_04_08_ofcom.pdf` (2008)
2. Silver, D.: Smart Start-ups: How to Make a Fortune from Starting Online Communities. John Wily and Sons, Inc (2007)
3. Szomszor, M., Alani, H., Cantador, I., O'Hara, K., Shadbolt, N.: Semantic modelling of user interests based on cross-folksonomy analysis. In: ISWC 2008: Proceedings of the 7th International Semantic Web Conference, Springer (2008) 632–648
4. Szomszor, M., Cantador, I., Alani, H.: Correlating user profiles from multiple folksonomies. In: HT 2008: Proceedings of the 19th ACM conference on Hypertext and Hypermedia, ACM (2008)
5. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic grounding of tag relatedness in social bookmarking systems. In: Proc. ISWC 2008. LNCS, Karlsruhe, Germany (2008)
6. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Emergent semantics in bibsonomy. In Hochberger, C., Liskowsky, R., eds.: Informatik 2006 - Informatik für Menschen. Band 2. Volume P-94 of Lecture Notes in Informatics., Bonn, Gesellschaft für Informatik (2006) Proc. Workshop on Applications of Semantic Technologies, Informatik 2006.
7. 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)
8. Fortuna, B., Grobelnik, M., Mladenić, D.: Visualization of text document corpus. Special Issue: Hot Topics in European Agent Research I Guest Editors: Andrea Omicini **29** (2005) 497–502
9. Fortuna, B., Grobelnik, M., Mladenic', D.: Semi-automatic data-driven ontology construction system. PASCAL EPrints (2006) `http://eprints.pascal-network.org/perl/oai2`.

10. Mladenic, D., Grobelnik, M., Fortuna, B., Grcar, M.: Knowledge discovery for semantic web. In Davies, J., Grobelnik, M., Mladenic, D., eds.: Semantic Knowledge Management: Integrating Ontology Management, Knowledge Discovery, and Human Language Technologies. Springer (2008)