Semantic Disambiguation and Contextualisation of Social Tags[1]

Ignacio Fernández-Tobías, Iván Cantador, Alejandro Bellogín

Departamento de Ingeniería Informática, Universidad Autónoma de Madrid
28049 Madrid, Spain
ign.fernandez01@estudiante.uam.es, {ivan.cantador, alejandro.bellogin}@uam.es

Abstract. We present an algorithmic framework to accurately and efficiently identify the semantic meanings and contexts of social tags within a particular folksonomy. The framework is used for building contextualised tag-based user and item profiles. We also present its implementation in a system called cTag, with which we preliminary analyse semantic meanings and contexts of tags belonging to Delicious and MovieLens folksonomies. The analysis includes a comparison between semantic similarities obtained for pairs of tags in Delicious folksonomy, and their semantic distances in the whole Web, according to co-occurrence based metrics computed with results of a Web search engine.

Keywords: social tagging, folksonomy, ambiguity, semantic contextualisation, clustering, user modelling.

1 Introduction

Social tagging has become a popular practice as a lightweight mean to classify and exchange information. Users create or upload content (items), annotate it with freely chosen words (tags), and share these annotations with others. In a social tagging system, the whole set of tags constitutes an unstructured collaborative knowledge classification scheme that is commonly known as *folksonomy*. This implicit classification serves various purposes, such as for item organisation, promotions, and sharing with friends or with the public. Studies have shown, however, that tags are generally chosen by users to reflect their interests [13]. These findings lend support to the idea of using tags to derive precise user preferences, and bring with new research opportunities on personalised search and recommendation [16,18,17].

Despite the above advantages, social tags are free text, and thus suffer from various vocabulary problems. Ambiguity (polysemy) of the tags arises as users apply the same tag in different domains (e.g. `bridge`, the architectonical structure vs. the card game). At the opposite end, the lack of synonym control can lead to different tags being used for the same concept, precluding collocation (e.g. `biscuit` and `cookie`). Synonym relations can also be found in the form of acronyms (e.g. `nyc` for `new york`

[1] This manuscript is an extended version of the paper 'cTag: Semantic Contextualisation of Social Tags', presented at the 6th International Workshop on Semantic Adaptive Social Web (SASWeb 2011).

city), and morphological deviations (e.g. blog, blogs, blogging). Moreover, there are tags that have single meanings, but are used in different semantic contexts that should be distinguished (e.g. web may be used to annotate items about distinct topics such as Web development, Web browsers, and Web 2.0).

Aiming to address the above problems, we present herein a system called cTag, which consists of an algorithmic framework that allows identifying semantic meanings and contexts of social tags within a particular folksonomy, and exploits them to build contextualised tag-based user and item profiles. The system is used to preliminary analyse semantic meanings and contexts of tags belonging to Delicious and MovieLens folksonomies, and to compare semantic similarities obtained for pairs of tags in Delicious, and their semantic distances in the whole Web, according to co-occurrence based metrics computed with results of a Web search engine.

The remaining of the paper is organised as follows. Section 2 introduces our notion of semantic context in social tagging, and presents our approach to identify the semantic contexts of social tags in a particular folksonomy. Section 3 describes how our approach can be used to contextualise and disambiguate social tags within user and item profiles, and Section 4 provides a preliminary analysis of contextualisation and disambiguation results. Finally, Sections 5 and 6 end with a brief summary of related work, conclusions and potential future research lines.

2 Semantic Contexts of Social Tags

Current folksonomy-based content retrieval systems have a common limitation: they do not deal with semantic ambiguities of tags. For instance, given a tag such as sf, existing content retrieval strategies do not discern between the two main meanings of that tag: *San Francisco* (the Californian city) and *Science Fiction* (the literary genre).

Semantic ambiguity of social tags, on the other hand, is being investigated in the literature. There are approaches that attempt to identify the actual meaning of a tag by linking it with structured knowledge bases [19,1,11,7]. These approaches rely on the availability of external knowledge bases, and so far are preliminary, and have not been applied to personalised search and recommendation.

Other works are based on the concept of tag co-occurrence, and aim at extracting tag semantic meanings and contexts within a particular folksonomy by applying probabilistic models and clustering techniques on the tag space according to the tag co-occurrences in item annotation profiles [24,23,18,2,22,9]. For example, for the tag sf, often co-occurring tags such as sanfrancisco, california and bayarea may be used to define the context "San Francisco, the Californian city", while co-occurring tags like sciencefiction, scifi and fiction may be used to define the context "Science Fiction, the literary genre."

In this paper, we follow a clustering strategy as well, but in contrast to previous approaches, ours provides the following benefits:

- Instead of using simple tag co-occurrences, we propose to use more sophisticated tag similarities, which were presented by Markines et al. in [14], and are derived from established information theoretic and statistical measures.
- Instead of using standard hierarchical or partitional clustering strategies, which require defining a stop criterion for the clustering processes, we propose to

apply the graph clustering technique presented by Newman and Girvan [15], which automatically establishes an optimal number of clusters. Moreover, to obtain the contexts of a particular tag, we propose not to cluster the whole folksonomy tag set, but a subset of it.

In the following, we briefly describe the above tag similarities and clustering technique. In Section 3, we shall explain how obtained tag similarities and clusters are exploited to contextualise tag-based profiles.

2.1 Tag Similarities

A folksonomy $\mathcal{F}$ can be defined as a tuple $\mathcal{F} = \{\mathcal{T}, \mathcal{U}, \mathcal{I}, \mathcal{A}\}$, where $\mathcal{T}$ is the set of tags that comprise the vocabulary expressed by the folksonomy, $\mathcal{U}$ and $\mathcal{I}$ are respectively the sets of users and items that annotate and are annotated with the tags of $\mathcal{T}$, and $\mathcal{A} = \{(u, t, i)\} \in \mathcal{U} \times \mathcal{T} \times \mathcal{I}$ is the set of assignments (annotations) of each tag t to an item i by a user u.

To compute semantic similarities between tags, we follow a two-step process. First, we transform the tripartite space of a folksonomy, represented by the triples $\{(u, t, i)\} \in \mathcal{A}$, into a set of tag-item relations $\{(t, i, w_{t,i})\} \in \mathcal{T} \times \mathcal{I} \times \mathbb{R}$ (or tag-user relations $\{(t, u, w_{t,u})\} \in \mathcal{T} \times \mathcal{U} \times \mathbb{R}$), where $w_{t,i}$ (or $w_{t,u}$) is a real number that expresses the relevance (importance, strength) of tag t when describing item profile i (or user profile u). In [14], Markines et al. call this transformation as tag assignment "aggregation", and present and evaluate a number of different aggregation methods. We focus on two of these methods, *projection* and *distributional* aggregation, which are described with a simple example in Figure 1. Projection aggregation is based on the Boolean use of a tag for annotating a particular item, while distributional aggregation is based on the popularity (within the community of users) of the tag for annotating such item.

Tag assignments [user, tag, item]

Alice	conference	recommender	research
www.umap2011.org	1	1	
www.delicious.com		1	
ir.ii.uam.es		1	1

Bob	conference	recommender	research
www.umap2011.org	1	1	1
www.delicious.com		1	
ir.ii.uam.es			

⇩

Tag assignment aggregation [tag, item]

Projection	conference	recommender	research
www.umap2011.org	1	1	1
www.delicious.com		1	
ir.ii.uam.es		1	1

Distributional	conference	recommender	research
www.umap2011.org	2	2	1
www.delicious.com		2	
ir.ii.uam.es		1	1

Figure 1. An example of projection and distributional tag assignment aggregations. 2 users, Alice and Bob, annotate 3 Web pages with 3 tags: `conference`, `recommender` and `research`.

Second, in the obtained bipartite tag-item (or tag-user) space, we compute similarities between tags based on co-occurrences of the tags in item (or user) profiles. In [14], the authors compile a number of similarity metrics derived from established information theoretic and statistical measures. The cTag system computes some of these metrics, whose definitions are given in Table 1.

2.2 Tag Clustering

We create a graph G, in which nodes represent the social tags of a folksonomy, and edges have weights that correspond to semantic similarities between tags. By using the similarity metrics presented in Section 2.1, G captures global co-occurrences of tags within item annotations, which in general, are related to *synonym* and *polysemy* relations between tags.

Table 1. Tested tag similarity metrics. $I_1, I_2 \subseteq I$ are the sets of items annotated with $t_1, t_2 \in \mathcal{T}$.

Similarity	*Projection aggregation*	*Distributional aggregation*
Matching	$sim(t_1,t_2) = \lvert I_1 \cap I_2 \rvert$	$sim(t_1,t_2) = -\sum_{t \in I_1 \cap I_2} \log \mathrm{p}(t)$
Overlap	$sim(t_1,t_2) = \frac{\lvert I_1 \cap I_2 \rvert}{\min(I_1, I_2)}$	$sim(t_1,t_2) = \frac{\sum_{t \in I_1 \cap I_2} \log \mathrm{p}(t)}{\max(\sum_{t \in I_1} \log \mathrm{p}(t), \sum_{t \in I_2} \log \mathrm{p}(t))}$
Jaccard	$sim(t_1,t_2) = \frac{\lvert I_1 \cap I_2 \rvert}{\lvert I_1 \cup I_2 \rvert}$	$sim(t_1,t_2) = \frac{\sum_{t \in I_1 \cap I_2} \log \mathrm{p}(t)}{\sum_{t \in I_1 \cup I_2} \log \mathrm{p}(t)}$
Dice	$sim(t_1,t_2) = \frac{2\lvert I_1 \cap I_2 \rvert}{\lvert I_1 \rvert + \lvert I_2 \rvert}$	$sim(t_1,t_2) = \frac{2\sum_{t \in I_1 \cap I_2} \log \mathrm{p}(t)}{\sum_{t \in I_1} \log \mathrm{p}(t) + \sum_{t \in I_2} \log \mathrm{p}(t)}$
Cosine	$sim(t_1,t_2) = \frac{I_1}{\sqrt{\lvert I_1 \rvert}} \cdot \frac{I_2}{\sqrt{\lvert I_2 \rvert}} = \frac{\lvert I_1 \cap I_2 \rvert}{\sqrt{\lvert I_1 \rvert \cdot \lvert I_2 \rvert}}$	$sim(t_1,t_2) = \frac{I_1}{\lVert I_1 \rVert} \cdot \frac{I_2}{\lVert I_2 \rVert}$

Once G is built, we apply the graph clustering technique presented by Newman and Girvan in [15], which automatically establishes an optimal number of clusters. However, we do not cluster G, but subgraphs of it. Specifically, for each tag $t_l \in \mathcal{T}$, we select its T_1 most similar tags and then, for each of these new tags, we select its T_2 most similar tags[2] to allow better distinguishing semantic meanings and contexts of t_l within the set of T_1 most similar tags. With all the obtained tags (at most $1 + T_1 T_2$), we create a new graph G_l, whose edges are extracted from the global graph G.

Tables 2 and 3 show examples of semantic meanings and contexts retrieved by our approach for Delicious[3] and MovieLens[4] tags. Delicious is an online system where users bookmark and tag Web pages. Since bookmarks can be related with any topic, a wide range of domains are covered by Delicious tags, and semantic meanings are easily distinguished in many cases. It can be seen, for instance, that most of the Web pages tagged with `sf` are about *San Francisco* and *Science Fiction*. Moreover, for a particular meaning, several contexts can be found. Web pages about San Francisco may belong to *restaurants* or announce *events* in that city.

MovieLens, on the other hand, is a recommender system where users rate and tag movies. We may expect that the number of contexts for a particular tag in MovieLens folksonomy is much lower than in Delicious' since the scope of the former (movies belonging to a limited number of genres) is smaller than the latter (Web pages related to any domain and topic). Moreover, we may also expect that distinct meanings and

[2] In preliminary experiments, we have tested $T_1 = 20, 25, 30$ and $T_2 = 3, 5$

[3] Delicious - Social bookmarking, http://www.delicious.com

[4] MovieLens - Movie recommendations, http://www.movielens.org

contexts of a particular tag are hardly differentiated in MovieLens since the number of tags and tag assignments per user and item is lower than in Delicious. Examples in Table 3, however, show that is not necessarily the case: there are `animation` movies produced by different studios (e.g. Disney and Pixar), movies interpreted by `Will Smith`, the American actor, with different genres (e.g. comedy, action, and science fiction), and movies with characters that can be described based on different facets, e.g. `James Bond`, as a spy, as a killer, or as a hero.

Table 2. Examples of semantic contexts identified for different Delicious tags.

tag	context centroid	context popularity	context tags
sf	fiction	0.498	fiction, scifi, sciencefiction, sci-fi, stores, fantasy, literature
	sanfrancisco	0.325	sanfrancisco, california, bayarea, losangeles, la
	restaurants	0.082	restaurants, restaurant, dining, food, eating
	events	0.016	events, event, conferences, conference, calendar
web	webdesign	0.434	webdesign, webdev, web_design, web-design, css, html
	web2.0	0.116	web2.0, socialnetworks, social, socialmedia
	javascript	0.077	javascript, js, ajax, jquery
	browser	0.038	browser, browsers, webbrowser, ie, firefox
holiday	christmas	0.336	christmas, xmas
	travel	0.274	travel, trip, vacation, tourism, turismo, planner
	airlines	0.104	airlines, arline, flights, flight, cheap
	rental	0.019	rental, apartment, housing, realestate

Table 3. Examples of semantic contexts identified for different MovieLens tags.

tag	context centroid	context popularity	context tags
animation	animals	0.354	animals, children, fun, kids, talking animals
	pixar	0.147	cartoon, inventive, pixar, toys come to life, vivid characters
	disney	0.127	classic, disney, disney studios, family, fantasy
	anime	0.032	anime, hayao miyazaki, japanese, studio ghibli, zibri studio
will smith	fantasy	0.226	fantasy, seen more than once, adventure, action, exciting
	funny	0.032	funny, comedy, jim carrey, claymation, very funny
	conspiracy	0.020	conspiracy, michael moore, twist ending, politics
	comic	0.016	comic, adapted from comic, superhero, based on a comic
james bond	murder	0.427	murder, bond, 007, assassin, killer as protagonist, serial killer
	action	0.079	action, scifi, adventure, superhero
	espionage	0.074	espionage, matt damon, robert ludlum, tom cruise, spies
	england	0.041	england, british, uk, based on a book

3 Semantically Contextualised Tag-based Profiles

We define the profile of user u as a vector $\mathbf{u} = (u_1, \ldots, u_T)$, where u_t is a weight (real number) that measures the "informativeness" of tag t to characterise contents annotated by u. Similarly, we define the profile of item i as a vector $\mathbf{i} = (i_1, \ldots, i_T)$, where i_t is a weight that measures the relevance of tag t to describe i. There exist different schemes to weight the components of tag-based user and item profiles. Some

of them are based on the information available in individual profiles, while others draw information from the whole folksonomy. We have implemented several forms of weighting strategies based on the well-known TF, TF-IDF, and BM25 information retrieval models [5].

In each of the built profile, a tag t is transformed into a semantically contextualised tag t^u (or t^i), which is formed by the union of t and the semantic context $c_{t,u}$ (or $c_{t,i}$) of t within the corresponding user profile u (or item profile i). For instance, tag `sf` in a user profile with tags like `city`, `california` and `bayarea` may be transformed into a new tag `sf|sanfrancisco`, since in that profile, "sf" clearly refers to San Francisco, the Californian city. With this new tag, matchings with item profiles containing contextualised tags such as `sf|fiction`, `sf|restaurants` or `sf|events` would be discarded by a personalised search or recommendation algorithm because they may annotate items related to Science Fiction, or more specific topics of San Francisco like restaurants and events in the city. More formally, the context (centroid) $c_{t,u}$ (or $c_{t,i}$) of tag t within the user profile u (or item profile i), and the corresponding contextualised tag t^u (or t^i) are defined as follows:

$$\forall (u,t,i) \in \mathcal{A}, \quad \begin{aligned} c_{t,u} &= \mathrm{c}(t,u) = \arg\max_{c_t} cos(\mathbf{c}_t, \mathbf{u}) \Rightarrow t^u = t \cup c_{t,u} \\ c_{t,i} &= \mathrm{c}(t,i) = \arg\max_{c_t} cos(\mathbf{c}_t, \mathbf{i}) \Rightarrow t^i = t \cup c_{t,i} \end{aligned}$$

where $\mathbf{c}_t = (c_1, \ldots, c_T)$ is the weighted list of tags that define each of the contexts c_t of tag t_l within the folksonomy (see Tables 2 and 3).

Tables 4 and 5 show several examples of contextualised tag-based Delicious and MovieLens profiles generated by our approach. Each table shows four item profiles in which two of them contain a certain tag, but used in two different contexts: `sf` as *San Francisco* and *Science Fiction*, `web` in the contexts of *Web development* and *Web 2.0*, *Disney* or *Anime* `animation` movies, `will smith` featuring *fantasy* or *funny* movies.

Table 4. Four semantically contextualised tag-based item profiles of Delicious dataset. Each original *tag* is transformed into a *tag*|*context* pair.

bayarea|sf	california|sf	city|sustainability	conservation|green	eco|green
environment|recycle	government|activism	green|environment	home|green	local|sanfrancisco
recycle|environment	recycling|environment	sanfrancisco|sf	**sf|sanfrancisco**	solar|environment
sustainability|recycling	sustainable|green	trash|green	urban|sustainability	volunteer|environmental
culture|philosophy	essay|interesting	fiction|sf	future|scifi	futurism|philosophy
god|science	interesting|science	literature|scifi	mind|philosophy	read|philosophy
religion|philosophy	research|science	sci-fi|sf	sciencefiction|sf	scifi|writing
sf|fiction	storytelling|fiction	toread|philosophy	universe|philosophy	writing|fiction
ajax|javascript	css|javascript	design|web	embed|webdesign	framework|javascript
gallery|jquery	html|javascript	icons|web	javascript|ajax	jquery|webdev
js|javascript	library|javascript	plugin|webdev	programming|javascript	site|webdev
toolkit|webdev	tutorials|webdev	**web|javascript**	web2.0|web	webdev|javascript
articles|web	blogs|web2.0	idea|community	internet|tools	library|opensource
network|tools	podcasts|education	rdf|web	reading|education	school|educational
semantic|semanticweb	semanticweb|web	semweb|semanticweb	software|utilities	technology|web2.0
tim|web	trends|technology	**web|web2.0**	web2.0|social	wiki|web2.0

Table 5. Four semantically contextualised tag-based item profiles of MovieLens dataset. Each original *tag* is transformed into a *tag*|*context* pair.

3d|animated	**animation|disney**	pixar animation|animation	comedy|animation	fun|adventure
disney|family	kids|toys come to life	animated|pixar animation	funny|animation	bright|toys come to life
computer|animation	disney animation|pixar	favorite|toys come to life	fantasy|animation	family|disney
toys|toys come to life	pixar|toys come to life	toys come to life|animated	classic|comedy	funny|animation
fantasy|zibri studio	dragon|anime movie	mythical creatures|anime	secret door|anime	japan|zibri studio
animation|anime	miyazaki|zibri studio	hayao miyazaki|myazaki	zibri strudio|anime	myazaki|zibri studio
fun|adventure	adventure|zibri studio	environment|mythical creatures	animated|animation	strange|foreign
foreign|japan	great anime film|anime	anime movie|mythical creatures	fanciful|zibri studio	anime|zibri studio
oscar winner|scifi	aliens|scifi	**will smith|fantasy**	frantic|scifi	end of the world|scifi
adventure|scifi	want|scifi	seen more than once|scifi	sf|scifi	action|fantasy
alien invasion|action	scifi|fantasy	seen at the cinema|scifi	war|action	disaster|scifi
dvd|space	watchfully|action	patriotic|scifi	invasion|scifi	et|scifi
comedy|funny	humor|comedy	end of the world|scifi	stupid|comedy	aliens|stupid
funny|comedy	amazing|fantasy	formulaic|will smith	action|fantasy	very funny|funny
predictable|scifi	fight|funny	seen more than once|comedy	futurism|scifi	cool|comedy
will smith|funny	cool but freaky|funny	violently stupid|comedy	dvd|space	space|alien invasion

4 Preliminary Analysis

We have implemented the algorithmic framework for tag and profile contextualisation described in Sections 2 and 3 in a system called cTag[5], which consists of a Web application and a Web service.

Figure 2 shows a screenshot of cTag Web application. The user selects a dataset – Delicious or MovieLens–, an aggregation method, and a tag similarity metric. Then, she queries for a tag available in the dataset, and is presented with the semantic contexts associated to that tag, which are obtained by our approach with the selected aggregation method and similarity metric. The contexts are shown as an ordered list of tag sets. Each context is assigned a name, which corresponds to the centroid tag of the context cluster, and a colour. The tags within a context are shown as a tag cloud, and have different sizes based on their weights in the semantic context. On the right side of the screen, the system also shows the graph associated to the input tag, and colour the different semantic clusters. The user can also introduce a profile manually or automatically by introducing a Delicious user name. The given profile is used to contextualise the input tag. In this case, only the contexts (clusters) that are related to the profile are shown.

The cTag system has allowed us to use and test our approach on various folksonomy datasets. In this section, we describe these datasets, and present a preliminary analysis of obtained semantic meanings, contexts, and similarities between tags. This analysis includes a comparison of semantic similarities for pairs of tags, and their semantic distances in the Web, by means of co-occurrence based metrics computed with results of a Web search engine.

[5] cTag Web application and Web service, http://ir.ii.uam.es/reshet/results.html

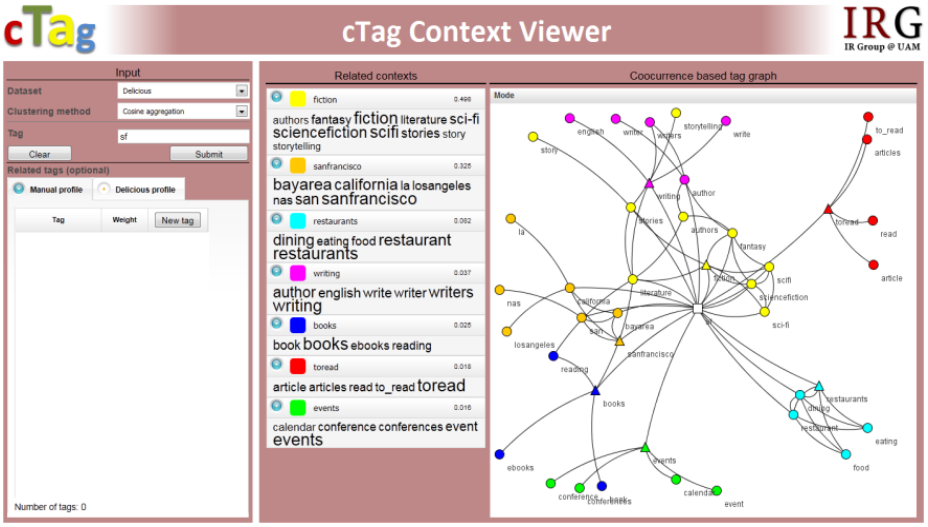

Figure 2. Screenshot of cTag Web application.

Figure 3 shows the XML response from cTag Web service for the input tag `sf` and profile {(`books`, 0.3), (`sci-fi`, 0.7)}, by using the cosine aggregation method with T_1=20 and T_2=5, on Delicious dataset. It can be seen that two semantic contexts are retrieved: *books* and *fiction*. Both of them are related to *Science Fiction* genre, but the former takes a higher weight since it focuses on books and readings, which is the main topic of the input profile.

```
<tag_contextualization_results method="cosine_aggregation_20_5" dataset="delicious">
   <tag value="sf">
      <profile>
         <profile_tag weight="0.3">books</profile_tag>
         <profile_tag weight="0.7">sci-fi</profile_tag>
      </profile>
      <contexts>
         <context name="books" similarity="0.107571">
            <context_tag weight="0.35857">books</context_tag>
            <context_tag weight="0.229219">book</context_tag>
            <context_tag weight="0.207827">ebooks</context_tag>
            <context_tag weight="0.204383">reading</context_tag>
         </context>
         <context name="fiction" similarity="0.0806848">
            <context_tag weight="0.145413">fiction</context_tag>
            <context_tag weight="0.144174">scifi</context_tag>
            <context_tag weight="0.12935">sciencefiction</context_tag>
            <context_tag weight="0.115264">sci-fi</context_tag>
            <context_tag weight="0.0890222">fantasy</context_tag>
            <context_tag weight="0.0834318">literature</context_tag>
            <context_tag weight="0.0683994">authors</context_tag>
            <context_tag weight="0.0661398">story</context_tag>
            <context_tag weight="0.0596612">storytelling</context_tag>
         </context>
      </contexts>
   </tag>
</tag_contextualization_results>
```

Figure 3. Example of XML response from cTag Web service.

4.1 Datasets

As shown in Table 6, in addition to the differences in the number and nature of their domains, cTag datasets[6] obtained from Delicious and MovieLens systems present distinct characteristics that may affect the contextualisation process, and its further application to folksonomy-based personalisation and recommendation strategies. Although the number of users is quite similar (~2K) for both datasets, the number of tagged items (and tag assignments) is much different; the purpose of Delicious is bookmarking and tagging Web pages, and MovieLens's is rating movies. Moreover, in Delicious dataset, a significant amount of tags was not contextualised because they are expressions that are not commonly shared by the community.

Table 6. Description of cTag datasets.

	Delicious	**MovieLens**
#users	1867	2113
#items	69226	5909
#tags	53388	5291
Avg. #tags/user	123.697 (99.870)	10.093 (52.193)
Avg. #tags/item	7.085 (3.397)	6.353 (8.141)
#TAS	437593	47958
Avg. #TAS/user	234.383 (192.395)	22.697 (169.948)
Avg. #TAS/item	6.321 (6.356)	8.116 (12.638)
#contextualised tags	14295	5291

4.2 Tag Meanings and Contexts

Table 7 shows some statistics about the clusters (semantic contexts) generated by the different tag aggregation and similarity strategies for Delicious and MovieLens datasets. From these results we can derive a number of conclusions. First, *Matching similarity*, which corresponds to the basic definition of co-occurrence, provides fewer clusters per tag in both Delicious and MovieLens datasets. It seems thus that simple co-occurrence is not enough to distinguish all the meanings and contexts of tags within a folksonomy. The other similarities, on the other hand, help to increase the differences between tag semantic distances, allowing a better clustering of tag graphs. Second, for all the strategies, there are not significant differences between the sizes of the generated clusters, and the average size of the clusters is quite similar for each strategy in the two datasets. These sizes (numbers of tags) represent the detail with which semantic meanings and contexts are described by our approach, and can be adjusted by the number of most similar tags used to build the semantic subgraph of each particular tag t, i.e. they can be adjusted by changing the values of parameters T_1 and T_2. Since the number of clusters and the cluster sizes are quite similar for the strategies, it seems that the tested similarities are able to capture the semantics underlying the tag subgraphs. This has also been observed in experiments we have conducted to evaluate the impact of the different contextualisation strategies on several tag-powered item recommenders [5].

[6] cTag datasets, published at HetRec'11 workshop: http://ir.ii.uam.es/hetrec2011

Table 7. Description of obtained clusters for each dataset and tag similarity.

		Delicious		MovieLens	
		Avg. #clusters/tag	**Avg. cluster size**	**Avg. #clusters/tag**	**Avg. cluster size**
Projection aggregation	**Matching**	**4.870** (1.517)	8.698 (3.897)	**6.165** (1.743)	7.875 (4.433)
	Overlap	**9.687** (3.022)	7.310 (3.270)	**10.154** (2.721)	7.305 (3.547)
	Jaccard	8.397 (2.848)	6.630 (2.674)	8.616 (2.902)	6.768 (3.501)
	Dice	8.407 (2.846)	6.622 (2.678)	8.633 (2.909)	6.754 (3.497)
	Cosine	8.579 (2.878)	6.538 (2.678)	8.719 (2.967)	6.689 (3.477)
Distributional aggregation	**Matching**	**4.875** (1.502)	8.687 (3.885)	**6.036** (1.745)	7.995 (4.382)
	Overlap	**9.767** (3.031)	7.244 (3.213)	**10.443** (2.796)	7.019 (3.402)
	Jaccard	8.403 (2.844)	6.640 (2.686)	8.868 (2.823)	6.808 (3.328)
	Dice	8.413 (2.845)	6.631 (2.682)	8.887 (2.832)	6.793 (3.326)
	Cosine	9.019 (2.858)	6.511 (2.576)	8.874 (3.135)	6.182 (3.169)

4.3 Tag Similarities

In this section, we compare the semantic similarities obtained for pairs of tags in the Delicious folksonomy against their semantic distances in the whole Web. As an approximation of tag distance in the Web, we make use of the Google Similarity Distance described in [8], but using Bing as search engine instead of Google, because of restrictions on the API access. In this way, we approximate the (semantic) similarity between two tags by using the Bing page counts of each tag along with the number of pages containing both tags. The distance is defined as follows:

$$distance(t_1, t_2) = \frac{f(t_1, t_2) - \min(f(t_1), f(t_2))}{\max(f(t_1), f(t_2))}$$

where $f(t)$ is the number of pages retrieved by Bing for the query composed by tag t, and $f(t_1, t_2)$ is the number of pages retrieved by Bing for the query "t_1 AND t_2".

Figure 4 presents a comparison between the tag frequency (page counts) distributions obtained for Delicious and Bing. We can observe that, even though both distributions are not equivalent, they present a power law distribution and (as expected) satisfy the Zipf's law.

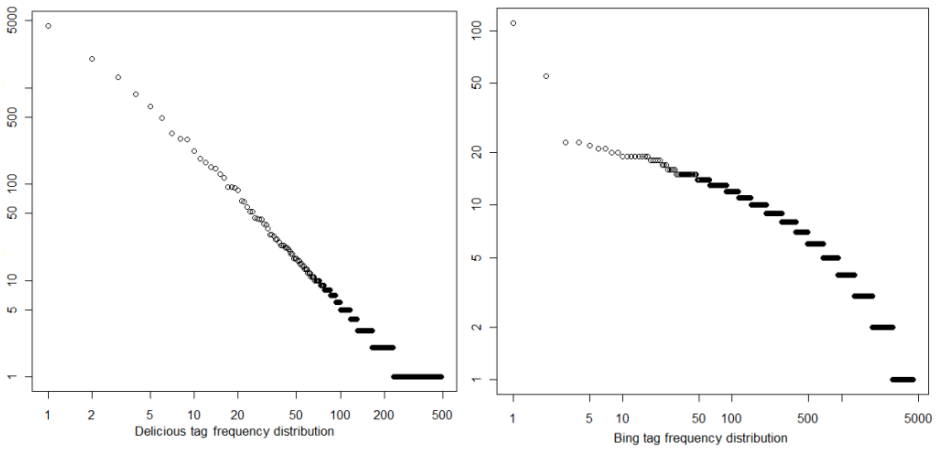

Figure 4. Distribution of tag/keyword frequencies in Delicious and Bing datasets (log-log plot)

There are a number of differences between the similarities found by using Delicious data and the ones obtained with Bing (i.e., the whole Web). First, the number of tags present in the Delicious dataset is different, more limited, to the potentially unlimited vocabulary available in the Web. In fact, it is very difficult (if possible) for Delicious' folksonomy to describe / cover all the semantics available in the Web. Furthermore, it is very likely that our collected data presents a biased representation of the whole knowledge available in the entire Delicious site. Figure 5 supports this claim, since there is no clear correlation between the tag frequencies found in each dataset. Moreover, in the figure, we can also observe how the values in the X axis (Delicious tag frequencies) have a more compact range than the Y axis; in particular, most of the tags appear less than 1000 times. The frequencies obtained by Bing, on the other hand, distribute themselves more uniformly in the space. This is likely due to the fact that we are making use of a more complete knowledge dataset, and thus, more accurate, in this situation.

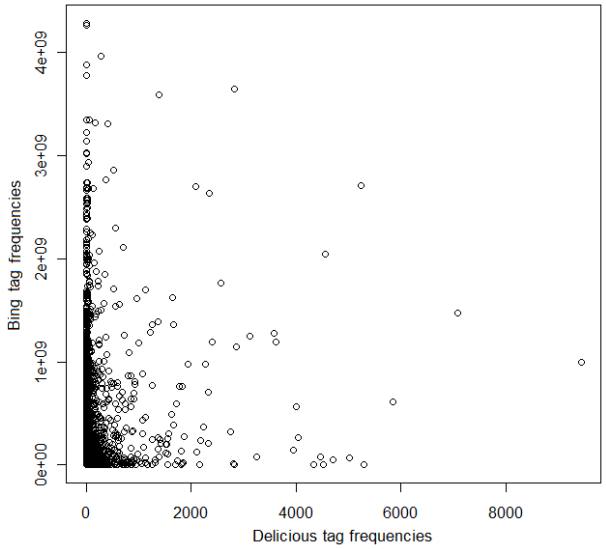

Figure 5. Scatterplot of tag/keyword frequencies in Delicious and Bing datasets.

Despite the fact that there is not a clear correlation between tag/keyword frequencies in Delicious folksonomy and the Web, we found out significant and diverse nonlinear correlations between the semantic similarities obtained with our approach in Delicious folksonomy, and the semantic distances computed with the Google Similarity Distance with Web search results. As shown in Table 8, *Jaccard* and *Dice* similarities do have a Spearman correlation value of -0.35[7]. This result supports the motivation of exploiting our semantic contextualisation approach to enhance folksonomy-based personalised web search [21] and recommendation [6]. Nonetheless, other strategies, such as *Matching* and *Overlap* similarities, do not have such correlation, and thus may not be good candidates for the above purposes. As

[7] Negative correlation values are obtained because the comparison is done between tag similarities and distances: the higher the similarity between two tags, the lower their distance.

already explained in Section 4.2, these metrics, which are based on simple tag co-occurrences, do not capture all the semantic meanings and contexts of the tags.

Analysing the correlations between semantic similarities (Table 8), it seems that in general, there are not significant differences between projection and distributional aggregation strategies for each particular semantic similarity, since their correlation values are in the range [0.84, 1.00], except for *Cosine* similarity, which has a correlation of 0.49 for its projection and distributional versions. On the contrary, differences appear when comparing semantic similarity strategies. Again, *Matching* and *Overlap* similarities strongly differ from the rest of similarities, with which they have positive and negative correlation values. *Jaccard* and *Dice* similarities, on the other hand, highly correlate, and *Cosine* similarity maintains moderate correlations.

Table 8. Spearman correlation values between tag semantic similarities obtained by our approach in Delicious folksonomy, and semantic distances computed with Bing Web search results.

		Projection aggregation					**Distributional aggregation**					**Web distance**
		Matching	**Overlap**	**Jaccard**	**Dice**	**Cosine**	**Matching**	**Overlap**	**Jaccard**	**Dice**	**Cosine**	
Projection aggregation	**Matching**	1.00	0.19	-0.10	0.12	0.17	0.84	0.18	0.12	0.12	0.39	-0.06
	Overlap		1.00	-0.10	-0.10	0.22	0.24	0.99	-0.09	-0.09	-0.15	0.08
	Jaccard			1.00	**1.00**	0.94	0.17	-0.08	**1.00**	**1.00**	0.58	**-0.35**
	Dice				1.00	0.94	0.16	-0.08	**1.00**	**1.00**	0.58	**-0.35**
	Cosine					1.00	0.23	0.23	0.94	0.94	0.49	**-0.31**
Distributional aggregation	**Matching**						1.00	0.26	0.17	0.17	0.24	-0.07
	Overlap							1.00	-0.08	-0.08	-015	0.07
	Jaccard								1.00	**1.00**	0.58	**-0.35**
	Dice									1.00	0.58	**-0.35**
	Cosine										1.00	**-0.25**
Web distance		-0.06	0.08	**-0.35**	**-0.35**	**-0.31**	-0.06	0.07	**-0.35**	**-0.35**	**-0.25**	1.00

5 Related Work

Semantic ambiguity is a phenomenon that occurs too frequently to be ignored by a social tagging system. As representative examples, as for October 2011, Wikipedia contains over 200K disambiguation entries[8], and Gemmell et al. [12] demonstrate that ambiguity and redundancy impede the evaluation and performance of tag-based recommender systems, especially in folksonomies that include broad domains.

Thus, semantic disambiguation of social tags is having increasing attention in the research literature. There are approaches that attempt to identify the actual meaning of a tag by linking it with structured knowledge bases [19,1,11,7]. These approaches rely

[8] Wikipedia disambiguation pages, http://en.wikipedia.org/wiki/Category:All_disambiguation_pages

on the availability of external knowledge resources, and so far are preliminary and have not been applied to personalisation and recommendation. Other works are based on the concept of tag co-occurrence, that is, on extracting the actual meaning of a tag by analysing the occurrence of the tag with others in describing different resources. These approaches usually involve the application of probabilistic models and clustering techniques over the co-occurrence information gathered from a folksonomy [24,23,18,2,22,9], and have been exploited by recent personalisation and recommendation approaches [16,18,17]. Their main advantage is that an external knowledge source is not required. In the following, we briefly describe and compare works on both types of approaches.

Specia and Motta [19] present a hybrid approach that exploits both tag co-occurrence information and knowledge provided by ontologies available in the Semantic Web, in order to generate groups of highly related tags that correspond to ontologies concepts. By using Wikipedia, the approach also identifies semantic relationships among subsets of grouped tags, representing thus one of the first strategies that automatically provide semantic structure to folksonomies. Extending Specia and Motta's approach, Angeletou et al. [1] exploit additional knowledge sources, such as WordNet, to identify richer semantic relationships between tags, and explore the use of obtained semantic structures to enhance search in folksonomies. As explained in this paper, our approach goes beyond grouping semantically related tags by identifying different semantic meanings and contexts of the tags. We believe Specia and Motta's, and Angeletou et al.'s approaches could be applied to our clusters providing more accurate relationships between tags.

More recent works have addressed semantic disambiguation and contextualisation of tags by exploiting Linked Open Data[9] repositories. García-Silva et al. [11] propose a framework that selects the meaning of a tag (within a particular context - set of annotations) among a number of candidate DBpedia [3] entries using information retrieval similarity functions. Cantador et al. [7] exploits YAGO ontology [20] for mapping tags to semantic concepts, and categorising them according to their purpose (describing content, context, subjective opinions, and self-organisational issues). These approaches require the existence of structured knowledge sources, and could be combined with the strategy presented herein to enrich the generated descriptions of the different semantic meanings and contexts of a tag.

Without using external semantic repositories, there exist a wide number of approaches that exploit tag co-occurrence information to address the disambiguation problem. We can categorise them in those that follow probabilistic models, and those that apply clustering strategies.

Zhang et al. [24] propose a probabilistic model that allows grouping synonymous tags together, and building global hierarchies within a folksonomy, which may be used to identify the different meanings of a tag. Weinberg et al. [23], on the other hand, address the ambiguity problem explicitly. They present a probabilistic model based on Kullback-Leibler divergence to identify the more likely meaning of a tag within an iterative tag recommendation scenario.

De Meo et al. [10] present an approach that also identifies groups of semantically

[9] Linking Open Data project, http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

similar tags based on co-occurrence metrics. The approach is capable of sorting the tags of a particular group in a "t_1 is more general that t_2" fashion, and building partial hierarchies of tags. However, it delegates the semantic disambiguation and contextualisation of tags to the users. Also exploiting co-occurrences between tags, Benz et al. [4] present an approach that automatically induces a hierarchical organisation scheme from the initially flat tag space of a folksonomy by using a strategy similar to that presented by De Meo et al. in [10], and applies a clustering approach for tag sense disambiguation. The approach, differently to ours, requires setting a number of parameters and thresholds in the different tag similarity computation, hierarchy building, and clustering and disambiguation processes.

Shepitsen et al. [18] apply hierarchical clustering to enhance personalised tag-based recommendations. The tags of a folksonomy are clustered, and obtained clusters are used to establish more accurate (less ambiguous) similarities between tag-based user profiles and item descriptions. Au Yeung et al. [2] present an alternative clustering strategy that is applied to the co-occurrence graph of a folksonomy. Similarly to our approach, Au Yeung et al.'s approach allows identifying semantic meanings and contexts of tags. However, it has a much higher computational cost since it is applied to large semantic networks. Finally, Vandic et al. [22] apply a non-hierarchical strategy for creating semantic clusters. Before the clustering process, their approach removes syntactic variations of tags by using the normalised Levenshtein distance, and the cosine similarity measure based on tag co-occurrences. The previous approaches cluster semantically related tags, but do not establish the semantic meanings and contexts explicitly. Moreover, their clustering processes are applied to the entire folksonomy graphs, which makes them having a high computational cost. Similarly to us, Datollo el al. [9] advocate for clustering subsets of semantically related tags. They propose a generic framework for finding synonym, homonym and hierarchical relationships between tags, but at the time of writing they have not fully implemented and evaluated it.

6 Conclusions and Future Work

In this paper, we have presented an algorithmic framework to identify the semantic meanings and contexts of social tags within a particular folksonomy, and exploit them for building contextualised tag-based user and item profiles. The main benefit of our approach is that it utilises a clustering technique that exploits sophisticated co-occurrence based similarities between tags, and is very efficient since it is not executed on the whole tag set of the folksonomy, and provides an automatic stop criterion to establish the optimal number of clusters.

We have preliminary analysed contextualisation results obtained by integrating state of the art tag semantic similarities into our approach, and have shown that similarities such as Jaccard and Dice seem to be better candidates than basic co-occurrence and Cosine similarity.

As shown in previous works [1,12,16,18], semantic disambiguation and contextualisation of social tags can be used to improve folksonomy-based personalised search and recommendation strategies. Recently, in [5], we have preliminary evaluated our approach with a number of state of the art recommenders [6] on a Delicious dataset, and have obtained 13% to 24% precision/recall

improvements by only contextualising 5.3% of the tags available in that dataset. In the study, we have also conducted a manual evaluation of our tag contextualisation approach. By considering as ground-truth data a set of 1,080 manual context assignments provided by 30 human evaluators for 78 distinct tags within several profiles, our approach have achieved 63.8%, 81.1% and 88.4% accuracies selecting respectively the first, second and third top contexts for each particular tag.

The effect of semantic contextualisation of tags in folksonomies describing a single domain (movies in MovieLens, music tracks in Last.fm), and in folksonomies about multiple domains (Web pages in Delicious), together with an exhaustive analysis of the proposed semantic tag similarities, and an empirical comparison of different clustering methods, are some research lines to be addressed.

Acknowledgements

This work was supported by the Spanish Ministry of Science and Innovation (TIN2008-06566-C04-02), and Universidad Autónoma de Madrid (CCG10-UAM/TIC-5877).

References

1. Angeletou, S., Sabou, M., Motta, E.: Improving Folksonomies Using Formal Knowledge: A Case Study on Search. In: 4th Asian Semantic Web Conference, pp. 276-290 (2009)
2. Au Yeung, C. M., Gibbins, N., Shadbolt, N.: Contextualising Tags in Collaborative Tagging Systems. In: 20th Conference on Hypertext and Hypermedia, pp. 251-260 (2009)
3. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A Nucleus for a Web of Open Data. In: 6th International Semantic Web Conference, pp. 722-735 (2007)
4. Benz, D., Hotho, A., Stützer, S., Stumme, G.: Semantics Made by You and Me: Self-emerging Ontologies Can Capture the Diversity of Shared Knowledge. In: 2nd Web Science Conference (2010)
5. Cantador, I., Bellogín, A., Fernández-Tobías, I., López-Hernández, S.: Semantic Contextualization of Social Tag-based Item Recommendations. In: 12th International Conference on Electronic Commerce and Web Technologies, pp. 101-113 (2011)
6. Cantador, I., Bellogín, A., Vallet, D.: Content-based Recommendation in Social Tagging Systems. In: 4th ACM Conference on Recommender Systems, pp. 237-240 (2010)
7. Cantador, I., Konstas, I., Jose, J. M.: Categorising Social Tags to Improve Folksonomy-based Recommendations. Journal of Web Semantics 9(1), pp. 1-15 (2011)
8. Cilibrasi, R. L., Vitányi, P. M. B.: The Google Similarity Distance. IEEE Transactions on Knowledge and Data Engineering 19(3), pp. 370-383 (2007)

9. Dattolo, A., Eynard, D., Mazzola, L.: An Integrated Approach to Discover Tag Semantics. In: 26th Annual ACM Symposium on Applied Computing, pp. 814-820 (2011)
10. De Meo, P., Quattrone, G., Ursino, D.: Exploitation of Semantic Relationships and Hierarchical Data Structures to Support a User in his Annotation and Browsing Activities in Folksonomies. Information Systems 34(6), pp. 511-535 (2009)
11. García-Silva, A., Szomszor, M., Alani, H., Corcho, O.: Preliminary Results in Tag Disambiguation using DBpedia. In: 1st International Workshop on Collective Knowledge Capturing and Representation (2009)
12. Gemmell, J., Ramezani, M., Schimoler, T., Christiansen, L., Mobasher, B.: The Impact of Ambiguity and Redundancy on Tag Recommendation in Folksonomies. In: 3rd ACM Conference on Recommender Systems, pp. 45-52 (2009)
13. Golder, S. A., Huberman, B. A.: Usage Patterns of Collaborative Tagging Systems. Journal of Information Science 32(2), pp. 198-208 (2006)
14. Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., Stumme, G.: Evaluating Similarity Measures for Emergent Semantics of Social Tagging. In: 18th International Conference on World Wide Web, pp. 641-650 (2009)
15. Newman, M. E. J., Girvan, M.: Finding and Evaluating Community Structure in Networks. Physical Review, E 69, 026113 (2004)
16. Niwa, S., Doi, T., Honiden, S.: Web Page Recommender System based on Folksonomy Mining for ITNG'06 Submissions. In: 3rd International Conference on Information Technology: New Generations, pp. 388-393 (2006)
17. Sen, S., Vig, J., Riedl, J.: Tagommenders: Connecting Users to Items through Tags. In: 18th International Conference on World Wide Web, pp. 671-680 (2009)
18. Shepitsen, A., Gemmell, J., Mobasher, B., Burke, R. 2008. Personalized Recommendation in Social Tagging Systems using Hierarchical Clustering. In: 2nd ACM Conference on Recommender Systems, pp. 259-266 (2008)
19. Specia, L., Motta, E.: Integrating Folksonomies with the Semantic Web. In: 4th European Semantic Web Conference, pp. 624-639 (2007)
20. Suchanek, F. M., Kasneci, G., Weikum, G.: YAGO: A Large Ontology from Wikipedia and WordNet. Journal of Web Semantics 6(3), pp. 203-217 (2008)
21. Vallet, D., Cantador, I., Jose, J. M.: Personalizing Web Search with Folksonomy-based User and Document Profiles. In: 32nd Annual European Conference on Information Retrieval, pp. 420-431 (2010)
22. Vandic, D., van Dam, J. W., Hogenboom, F., Frasincar, F.: A Semantic Clustering-based Approach for Searching and Browsing Tag Spaces. In: 26th Annual ACM Symposium on Applied Computing, pp.1693-1699 (2011)
23. Weinberger, K. Q., Slaney, M., Van Zwol, R.: Resolving Tag Ambiguity. In: 16th ACM Conference on Multimedia, pp. 111-120 (2008)
24. Zhang, L., Wu, X., Yu, Y.: Emergent Semantics from Folksonomies: A Quantitative Study. Journal on Data Semantics VI, pp. 168-186 (2006)