cTag: Semantic Contextualisation of Social Tags

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. In this paper, we present 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. 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.

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 (resources, 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 [8]. 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 [11,12,13].

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 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 such 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.

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 [1,6]. These approaches, however, 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, that is, on extracting tag semantic meanings and contexts within a particular folksonomy by clustering the tags according to their co-occurrences in item annotation profiles [2,7,14]. 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 [9], 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 [10], 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 $\mathcal{F} = \{\mathcal{T}, \mathcal{U}, I, \mathcal{A}\}$, where $\mathcal{T} = \{t_1, \dots, t_L\}$ is the set of tags that comprise the vocabulary expressed by the folksonomy, $\mathcal{U} = \{u_1, \dots, u_M\}$ and $I = \{i_1, \dots, i_N\}$ are respectively the sets of users and items that annotate and are annotated with the tags of $\mathcal{T}$, and $\mathcal{A} = \{(u_m, t_l, i_n)\} \in \mathcal{U} \times \mathcal{T} \times I$ is the set of assignments (annotations) of each tag t_l to an item i_n by a user u_m.

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 $\{(u_m, t_l, i_n)\} \in \mathcal{A}$, into a set of tag-item relations $\{(t_l, i_n, w_{l,n})\} \in \mathcal{T} \times I \times \mathbb{R}$ (or tag-user relations $\{(t_l, u_m, w_{l,m})\} \in \mathcal{T} \times \mathcal{U} \times \mathbb{R}$), where $w_{l,n}$ (or $w_{l,m}$) is a real number that expresses the relevance (importance, strength) of tag t_l when describing item profile i_n (or user profile u_m). In [9], 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. 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 [9], the authors compile a number of similarity metrics derived from established information theoretic and statistical measures. cTag computes some of these metrics, whose definitions are given in Table 1.

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`.

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

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

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.

Once G is built, we apply the graph clustering technique presented by Newman and Girvan in [10], 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[1] 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[2] and MovieLens[3] 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.

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, schi-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, airline, flights, flight, cheap
	rental	0.019	rental, apartment, housing, realestate

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 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.

[1] In preliminary experiments, we have tested $T_1 = 20, 25, 30$ and $T_2 = 3, 5$
[2] Delicious - Social bookmarking, http://www.delicious.com
[3] MovieLens - Movie recommendations, http://www.movielens.org

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_m as a vector $\mathbf{u}_m = (u_{m,1}, \ldots, u_{m,L})$, where $u_{m,l}$ is a weight (real number) that measures the "informativeness" of tag t_l to characterise contents annotated by u_m. Similarly, we define the profile of item i_n as a vector $\mathbf{i}_n = (i_{n,1}, \ldots, i_{n,L})$, where $i_{n,l}$ is a weight that measures the relevance of tag t_l to describe i_n. 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 [3].

In each of the built profile, a tag t_l is transformed into a semantically contextualised tag t_l^m (or t_l^n), which is formed by the union of t_l and the semantic context $c_{l,m}$ (or $c_{l,n}$) of t_l within the corresponding user profile u_m (or item profile i_n). 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) $c_{l,m}$ (or $c_{l,n}$) of tag t_l within the user profile u_m (or item profile i_n), and the corresponding contextualised tag t_l^m (or t_l^n) are defined as follows:

$$\forall (u_m, t_l, i_n) \in \mathcal{A}, \quad \begin{array}{l} c_{l,m} = \mathrm{c}(t_l, u_m) = \arg\max_{c_l} cos(\mathbf{c}_l, \mathbf{u}_m) \Rightarrow t_l^m = t_l \cup c_{l,m} \\ c_{l,n} = \mathrm{c}(t_l, i_n) = \arg\max_{c_l} cos(\mathbf{c}_l, \mathbf{i}_n) \Rightarrow t_l^n = t_l \cup c_{l,n} \end{array}$$

where $\mathbf{c}_l = (c_{l,1}, \ldots, c_{l,L})$ is the weighted list of tags that define each of the contexts c_l 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 cTag

cTag[4] is a system with the implementation of the algorithmic framework for tag and profile contextualisation presented in Sections 2 and 3, and allows using and testing it through a Web application and a Web service. Figure 2 shows a screenshot of cTag Web application. The user selects a dataset –Delicious or MovieLens– and a tag similarity, queries for a social tag available in the dataset, and obtains the semantic contexts associated to that tag. The user can also set a profile (manually or automatically via Delicious API) to contextualise. The retrieved contexts (clusters) are shown in the form of weighted lists of tag clouds, and a coloured clustered graph.

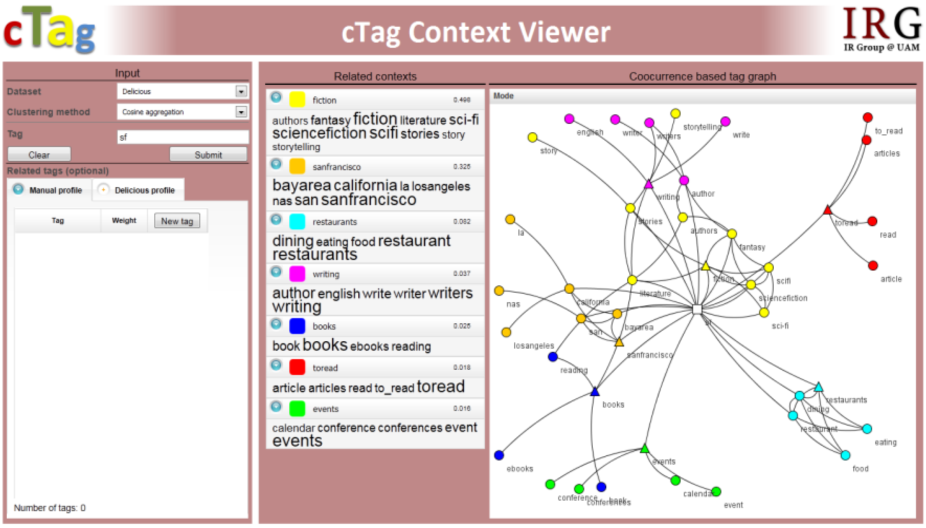

Figure 2. Screenshot of cTag Web application.

Figure 3 shows the XML response from cTag Web service for the input tag `sf` and profile {(`books`, 0.7), (`sci-fi`, 0.3)}, 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.

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

```
<tag_contextualization_results method="cosine_aggregation_20_5" dataset="delicious">
   <tag value="sf">
      <profile>
         <profile_tag weight="0.7">books</profile_tag>
         <profile_tag weight="0.3">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="0.099144">stories</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 XML response from cTag Web service.

As shown in Table 6, in addition to the differences in the number and nature of their domains, cTag datasets[5] obtained from Delicious and MovieLens systems present distinct characteristics that may affect the contextualisation process (Table 7), 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

[5] 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)

5 Conclusions and Future Work

In this paper, we have presented cTag, a system which consists of 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 cTag approach is that it utilises a clustering technique that exploits sophisticated co-occurrence based similarities between tags, 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.

As shown in previous works [1,7,11,13], semantic disambiguation and contextualisation of social tags can be used to improve folksonomy-based personalised search and recommendation strategies. Recently, in [3], we have preliminary evaluated cTag with a number of state of the art recommenders [4] 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 that 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), does have on personalization and recommendation strategies, 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.

The distinction of the users' tagging purposes –describing content and context, making subjective opinions, and providing self-references– may be also taken into consideration to enhance the tag disambiguation/contextualization process [5].

Acknowledgements

This work was supported by the Spanish Ministry of Science and Innovation (TIN2008-06566-C04-02), and the Regional Government of Madrid (S2009TIC-1542).

References

1. Angeletou, S., Sabou, M., Motta, E.: Improving Folksonomies Using Formal Knowledge: A Case Study on Search. In: 4th Asian Semantic Web Conference, 276--290. Springer-Verlag (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. ACM Press (2009)
3. 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. Springer-Verlag (2011)
4. Cantador, I., Bellogín, A., Vallet, D.: Content-based Recommendation in Social Tagging Systems. In: 4th ACM Conference on Recommender Systems, pp. 237--240. ACM Press (2010)
5. 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)
6. 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)
7. 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. ACM Press (2009)
8. Golder, S. A., Huberman, B. A.: Usage Patterns of Collaborative Tagging Systems. Journal of Information Science, 32(2), 198--208 (2006)
9. 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. ACM Press (2009)
10. Newman, M. E. J., Girvan, M.: Finding and Evaluating Community Structure in Networks. Physical Review, E 69, 026113 (2004)
11. 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. IEEE Press (2006)
12. Sen, S., Vig, J., Riedl, J.: Tagommenders: Connecting Users to Items through Tags. In: 18th International Conference on World Wide Web, pp. 671--680. ACM Press (2009)
13. 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. ACM Press (2008)
14. Weinberger, K. Q., Slaney, M., Van Zwol, R.: Resolving Tag Ambiguity. In: 16th ACM Conference on Multimedia, pp. 111--120. ACM Press (2008)