
DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

# SEMANTICS DISAMBIGUATION IN FOLKSONOMY: A CASE STUDY

Pierre Andrews, Juan Pane
and Ilya Zaihrayeu

December 2010

Technical Report # DISI-10-063

# Semantics Disambiguation in Folksonomy: a Case Study

Pierre Andrews, Juan Pane, and Ilya Zaihrayeu

Dipartimento di Ingegneria e Scienza Dell'Informazione,
Universit degli studi di Trento, Italy
{andrews,pane,ilya}@disi.unitn.it

**Abstract.** Social annotation systems such as del.icio.us, Flickr and others have gained tremendous popularity among Web 2.0 users. One of the factors of success was the simplicity of the underlying model, which consists of a resource (e.g., a web page), a tag (e.g., a text string), and a user who annotates the resource with the tag. However, due to the syntactic nature of the underlying model, these systems have been criticised for not being able to take into account the explicit semantics implicitly encoded by the users in each tag. In this article we: a) provide a formalisation of an annotation model in which tags are based on concepts instead of being free text strings; b) describe how an existing annotation system can be converted to the proposed model; c) report on the results of such a conversion on the example of a del.icio.us dataset; and d) show how the quality of search can be improved by the semantic in the converted dataset.

## 1 Introduction

One of the cornerstones of what we now call the "Web 2.0" is unconstrained user collaboration and creation of content. Some of the first sites to allow such features were del.icio.us and Flickr where users could share resources – bookmarks and photos respectively – and freely annotate them. Both websites allowed the creation of so called Folksonomies: social classification of resources created by the community that have shown to be very important for organising the large amount of content online, but also for, later on, studying the collaborative creation of shared vocabularies and taxonomies.

These folksonomies are now widely studied, in particular with the model of tripartite graphs of *tags-users-resources*. However, in this model, *tags* are free-form terms with no explicit semantic, therefore a number of issues arise from their use, such as:

- the loss in precision due to the ambiguity of tags – for example, the tag "java" can refer to the "Indonesian island", the "programming language", and a "beverage".
- the loss of recall due to the synonymy of terms – for instance, if you search for the tag "travel", you might be interested by the results for the tag "journey".

The use of different forms of the same word also exacerbate these issues as some users would, for example, use the tag "running", others would use instead "run", "runs", "torun", etc.

A number of approaches try to disambiguate tags in folksonomies and to create organised formal vocabularies automatically from them [1]. This has shown to be a difficult task and has not yet been fully characterised and evaluated. In this article, we propose a case study of a sample of the del.icio.us tripartite graph that was manually annotated with senses from a controlled vocabulary (Wordnet). We show a number of properties of the vocabulary shared by the users of this folksonomy and identify important features that have been overlooked by previous studies on disambiguation and sense extraction from such folksonomies. Moreover, we provide a quantitative analysis of the impact of the introduction of formal semantics in folksonomies on the construction of digital libraries on top of them where the data can be more easily accessed and processed by computers.

This article is organised as follows. First, in Section 2, we introduce the issue of folksonomy modelling and how we believe it can be extended to formalise the semantic of tags; we then discuss in Section 3 our case study and the methodology that was used to construct the dataset we examine. In Section 5 we then introduce and analyse a number of features of the vocabulary used, in particular on: *a*) how preprocessing of different forms of the same term can reduce the vocabulary size by ca. 60%, *b*) how a general controlled vocabulary is too static to encode the vocabulary of the folksonomy users as only around 50% of terms can be mapped to a controlled sense. In Section 6, we extend this analysis by showing quantitatively how the disambiguation of tags to senses can improve search. Finally, in Section 7 we discuss the related work and how it compares to the results we have obtained.

## 2 Semantic Folksonomy Model

### 2.1 Syntactic Folksonomy

The term folksonomy was coined in 2004 by T. Vander Wal [2] who characterised the new social tagging web sites that were appearing at the time. He defined a folksonomy as "*the result of personal free tagging of information and objects (anything with a URL) for one's own retrieval*". This "result" is one of the simplest form of annotation of resources with metadata that can serve to help the indexing, categorisation or sharing of such resources: a tag annotation.

Mika[3] introduced a formalisation of this results to ease its processing in multimodal graph analysis. Doing so, the author enables the formal representation of the social network resulting from the folksonomy building activity. Mika represents a folksonomy as a tripartite graph composed of three disjoint types of vertices, the *actors* $A$ (the user creating the tag annotation), the *concepts* $C$ (tags, keywords) used as metadata and the *objects* $O$ or resources being annotated. A tag annotation is thus a triple combining three vertices from each type:

$$T = \langle u, t, r \rangle \text{ where } u \in A, t \in C \text{ and } r \in O$$

According to Mika, such tripartite graph can be used to describe an ontology representing the knowledge of the community that created this folksonomy. This model has been used since to exploit different social networking analysis tools and distributional semantic models to extract a more formal representation of the semantic knowledge encoded in these tripartite graphs.

### 2.2 Semantic Folksonomy

An important point in Mika's [3] description of the folksonomy model is that "tags" or "keywords" are considered to be mapped one-to-one to the *concepts* of the ontology and that these are the semantic units of the language used in the community that created the folksonomy. However, we believe that a more granular model has to be used to represent the conceptual part of folksonomies. This will enable a better understanding of its underlying semantic and of the overlap of vocabularies between the users of the folksonomy.

In fact, tags and keywords, while they represent a specific concept and have a known semantic for the agent that creates them, are just stored and shared in the folksonomy as purely free-form natural language text. Because of the ambiguous nature of natural language [4], a number of issues arise when sharing only the textual version of the annotations:

**Base form variation** This problem is related to natural language input issues where the annotation is based on different forms of the same word (e.g., plurals vs. singular forms, conjugations, misspellings) [4].

**Homography** Annotation elements may have ambiguous interpretation. For instance, the tag "Java" may be used to describe a resource about the *Java island* or a resource about the *Java programming language*; thus, users looking for resources related to the programming language may also get some irrelevant resources related to the Island (therefore, reducing the precision);

**Synonymy** Syntactically different annotation elements may have the same meaning. For example, the tags "image" and "picture" may be used interchangeably by users but will be treated by the system as two different tags because of their different spelling; thus, retrieving resources using only one of these tags may yield incomplete results as the computer is not aware of the synonymy link;

**Specificity gap** This problem comes from a difference in the specificity of terms used in annotation and searching. For example, the user searching with the tag "cheese" will not find resources tagged with "cheddar[1]" if no link connecting these two terms exists in the system.

Indeed, as we show in our case study of del.icio.us (see Section 5.2), such issues can be found in a real application of the folksonomy model. We thus propose to replace the simple "Concept"↔"tag" mapping to one that will allow for an explicit formalisation of the intended semantic of the tag. The intuition behind this new formalisation is two-fold:

[1] which is a kind of cheese

– different tags could represent different forms of the same concept – for instance, "folksonomy" and "folksonomies" or "image" and "picture",
– a tag could represent a composed concept relying on two atomic concepts – for instance "sunny italy".

One suitable formalism for the representation of concepts is the one defined by Description Logics (DL) [5]. Briefly, the semantics (or, the extension) of a concept in DL is defined as a set of elements (or, instances). For example, the extension of the concept `Person` is the set of people existing in some model (e.g., in the model of the world). Because they are defined under a set-theoretic semantics, operators from the set theory can be applied on concepts, e.g., one could state that concept `Organism` *subsumes* (or, is more general than) the concept `Person` because the extension of the former concept is a superset for the extension of the latter concept. Among other things, the subsumption relation can be used for building taxonomies of concepts. These properties lead to a number of useful reasoning capabilities such as computing the instances of concepts through the concept subsumption, computing more specific or general concepts – these capabilities can be used for building services for the end users such as semantic search, as discussed in Section 6. A more complete introduction to DL is out of the scope of this article; interested readers are referred to [5] for details.

We thus introduce two new formalisations in the model to create a quadripartite graph representing the user-resource-tag-concept link:

– A *controlled tag* $ct$ is a tuple $ct = \langle t, \{lc\}\rangle$, where $t$ is a tag, i.e., a non-empty finite sequence of characters normally representing natural language words or phrases such as "bird", "sunnydays" or "sea"; and $\{lc\}$ is an ordered list of linguistic concepts, defined as follows:
– A *linguistic concept* $lc$ is a tuple $lc = \langle c, ct\rangle$, where $c$ is a concept as defined in DL (see above); and $ct$ is a term in natural language that denotes the concept $c$.

Consider an example of a controlled tag: $ct = \langle \text{“}sunnydays'', \{lc_1, lc_2\}\rangle$, with $lc_1 = \langle$`Sunny`, “$sunny''\rangle$ and $lc_1 = \langle$`Day`, “$day''\rangle$. Note that there can be more than one term that represents the same concept as in $lc_3 = \langle$`Sunny`, “$bright''\rangle$.

Recall the syntactic folksonomy model definition (see Section 2.1) that we now extend to the definition of a controlled tag annotation, $T^C$:

$$T^C = \langle u, ct, r\rangle \text{ where } u \in A, ct \textit{ is a controlled tag}, \text{ and } r \in O$$

In the following section we discuss how controlled tag annotations can be used to "semantify" a social annotation system such as del.icio.us.

## 3 Semantifying del.icio.us

To study our model and a set of natural language technologies that can be used to help the users in specifying the semantic of tags at the time of their creation, we study the widely used del.icio.us[2] folksonomy as a case study.

[2] `http://del.icio.us`

del.icio.us is a simple folksonomy as was defined by [2] and formalised by [3] in that it links resources to users and tags in a tripartite graph. However, these tags are totally uncontrolled and their semantic is not explicit. In the current datasets, for instance the ones provided by Tagora[3] or listed in [6], no-one has yet, to the best of our knowledge, provided a golden standard with such semantics. In that, the del.icio.us dataset is not perfectly what we are looking for, the Faviki[4] website could provide such dataset, however it does not contain so many users and annotations as del.icio.us and the quality of the disambiguations is not guaranteed. To make the del.icio.us dataset fit our problem statement, we have thus decided to extend a subset of a del.icio.us dump with disambiguated tags by manual validation. We used WordNet 2.0 [7] as the underlying controlled vocabulary for finding and assigning senses for tag tokens.

### 3.1 del.icio.us Sample

We obtained the initial data from the authors of [8] who crawled del.icio.us between December 2007 and April 2008. After some initial cleaning the dataset contains *5 431 804* unique tags (where the uniqueness criteria is the exact string match) of *947 729* anonimized users, over *45 600 619* unique URLs on *8 213 547* different website domains. This data can be considered to follow the syntactic folksonomy model $\langle t, r, u \rangle$ where the resource $r$ is the URL being annotated, containing a total of *401 970 328* tag annotations.

To study the semantic used in these tags, we have thus decided to extend a subset of the data with disambiguated tags; i.e., convert $t \rightarrow ct$. This means that for each tag $t$ in this subset, we have explicitly split it in its component tokens and marked it with the Wordnet synset (its sense) it refers to and thus get to the semantic folksonomy model described in Section 2.2.

The golden standard dataset we have built includes annotations from users which have less than 1 000 tags and have used at least ten different tags in five different website domains. This upper bound was decided considering that del.icio.us is also subject to spamming, and users with more than one thousand tags could potentially be spammers as the original authors of the crawled data assumed [8]. Furthermore, only $\langle r, u \rangle$ pairs that have at least three tags (to provide diversity in the golden standard), no more than ten tags (to avoid timely manual validation) and coming from users who have tags in at least five website domains (to further reduce the probability of spam tags) are selected. Only URLs that have been used by at least twenty users are considered in the golden standard in order to provide enough overlap between users. After retrieving all the $\langle r, u \rangle$ pairs that comply with the previously mentioned constraints, we randomly selected 500 pairs. We thus obtained 4 707 tag annotations with 871 unique tags on 299 URLs in 172 different web domains.

[3] `http://www.tagora-project.eu/data/`
[4] `http://faviki.com/`

### 3.2 Manual Validation

Selecting the right tag split and the right disambiguation for each token in such split is a tedious task for the human annotators and we try to make this task as straightforward as possible. Thus we use some supporting tools to simplify the work of the validators and streamline the annotation process. A team of three annotators have already annotated a sample of one thousand bookmarks from a del.icio.us crawl in less than a week. To enable such streamlined annotation, some pre-annotation is performed automatically so that the most probable splits are already available to the validators and the most probable disambiguation is also proposed. These supporting tools are described in the following sections.

### 3.3 Preprocessing

The goal of the preprocessing step is to recognise a word sequence in a tag that may consist of several concatenated tokens that might have been written with syntactic variations (e.g,. plurals, exceptional forms). This step is composed of the following sub-steps:

1. **Tag split**: split the tags into the component tokens. This step is needed considering the fact that many annotation systems such as del.icio.us do not allow spaces as word separators and, therefore, users just concatenate multi-words (javaisland) or concatenate them using the Camel case (javaIsland), slashes (java-island), underscores (java_island) or other separator they deem useful. The tag split preprocessing runs a search in WordNet and tries to place *splits* when it recognises valid tokens. This preprocessing can generate different splits for the same tag, for instance, the tag "javaisland" can be split into {"java", "island"} or {"java", "is", "land"}. The output of this step is ranked to present the most plausible split to the annotator first. The ranking prefers proposals with fewer number of splits and with the maximum number of tokens linked to the controlled vocabulary.
2. **Lemmatization**: in order to reduce different forms of the word into a single form (that can later be found in a vocabulary such as Wordnet), a number of standard lemmatization heuristics are applied. For example, "banks" would be preprocessed as "bank".

### 3.4 Disambiguation

In this step we run an automatic disambiguation algorithm in order to suggest to the validator the possibly correct sense of the word (as preprocessed in the previous step). The algorithm is an extension of the one reported in [9] and based on the idea that collocated tags provide context for disambiguating (as generalised in the survey by Garcia-Silva [1]). In our approach, given a token within a tag split, we consider three levels of context: 1. the other tokens in the tag split provide the first level of context, 2. the tokens in the tag splits for the other tags used for the annotation of the same resource by the *same* user

provide the second level, 3. the tokens in the tag splits for the tags used for the annotation of the same resource by *other* users, provide the third level of context.

The possible semantic relations between the senses of the given token and the senses of the tokens from its contexts are then mined to find a disambiguation. When a relation is found, the score of the corresponding word sense is boosted by a predefined value. The relations used are as follows (in decreasing order of their boost value):

1. synonymy (e.g., "image" and "picture");
2. specificity, measured as the length of the is-a path between two senses (e.g., "dog (Canis familiaris)" is more specific than "animal (a living organism)"); and
3. relatedness, measured as the sum of the lengths of the paths from the two given senses to the nearest common parent sense (e.g., "table (a piece of furniture)" is related to "chair" (a seat for one person) through the common parent sense "furniture (furnishings that make a room)").

For the specificity and relatedness relations, the scores are adjusted according to the length of the path (the shorter the length, the higher the score). The scores for all the relations are also boosted according to the level of the used context (level one leads to higher scores, whereas level three leads to lower scores). The algorithm then uses two other heuristics to boost the scores of word senses, namely: 1. we boost the sense of a word if the part-of-speech (POS) of that sense is the same as the one returned by a POS tagger[5]; and 2. we boost the sense of a word according to the frequency of usage of the sense[6].

The sense with the highest score is then proposed to the validator as the suggested meaning of the token. If more than one sense has the highest score we applied an heuristic were the POS is preferred in the following order: nouns, verbs, adjectives and adverbs – as this follows the distribution of the tag tokens by POS in annotation systems such as del.icio.us as reported in [10] and confirmed in our own analysis (see Figure 5a)). Finally, if more than one candidate remains, then the sense with the highest frequency of usage is selected.

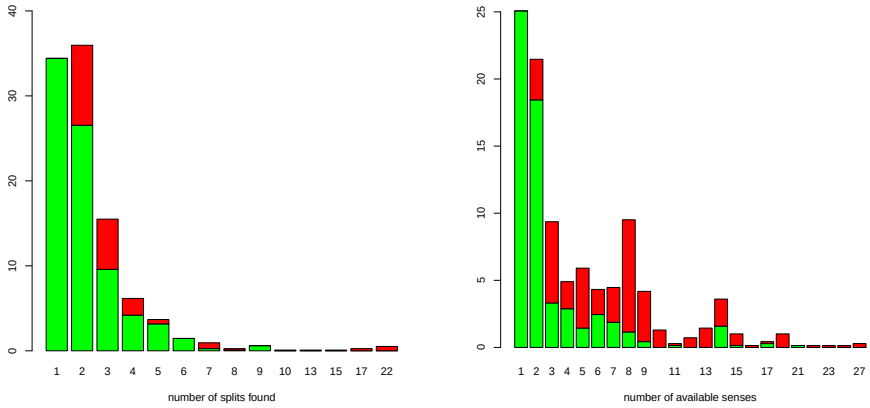
## 4 Results

In the following paragraphs we describe a first evaluation of the validity of the algorithms we described in the previous section, based on the annotated sample from del.icio.us.

### 4.1 Preprocessing

The accuracy of the preprocessing step (see Section 3.3) in this validation task reached 80.31%. In Figure 1a) we provide a detailed analysis of the accuracy of

[5] which can reach more than 97% in accuracy on metadata labels as shown in [9]
[6] this data is available in linguistic resources such as WordNet [7]




a) Preprocessing – by Number of Splits b) WSD – by Number of Senses

**Fig. 1.** Accuracy of the Preprocessing and WSD Algorithms

the algorithm for different numbers of possible splits. The Y axis corresponds to the distribution of tags per number of possible splits proposed to the validator, the top box is the amount of wrong splits ranked as best split while the bottom one represents the amount of accurate splits that were ranked top by the preprocessing algorithm. The plot should be read as follows: ∼35% of all the tags have two possible splits and the accuracy of the algorithm for these tags is ∼80% (see the second bar from the left).

We believe that the current accuracy of the preprocessing algorithm can be increased by some simple improvements on the lemmatization heuristics as well as by using a lexicon of existing words in the English language.

### 4.2 Word Sense Disambiguation

The average homography of the tag tokens in the dataset is 4.68, i.e., each tag token has 4.68 possible senses on average. The proposed WSD algorithm performed at 59.37% in accuracy. In Figure 1b) we provide a detailed analysis of the accuracy of the algorithm for different levels of homography. The Y axis corresponds to the distribution of tokens per number of possible homographs, the top box is the amount of wrong disambiguations ranked as best while the top one represents the amount of accurate disambiguations that were ranked top by the WSD algorithm. The figure should be read as follows: the number of cases with two possible senses in the controlled vocabulary is ∼22% and the accuracy of the algorithm for these cases is ∼90% (see the second bar from the left).

It is worth noting that, on Figure 1b), we can see that the WSD algorithm has an accuracy lower than 50% for the tokens with many available senses, however, the biggest amount of tokens only have two senses available and in this case, the WSD algorithm performs at an accuracy close to 90%.

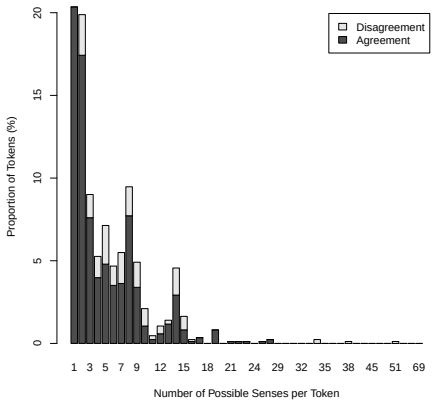
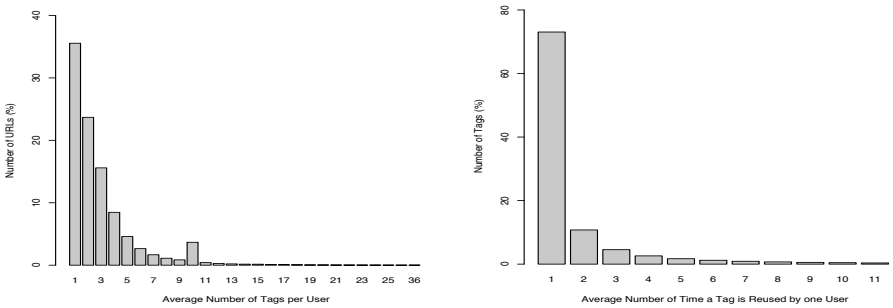



**Fig. 2.** Agreement between Annotators on Sense Validation, per Number of Available Senses

From the result we conclude that the WSD problem can be harder in its application in the domain of tag annotations than in its application in the domain of web directory labels, which are closer to tags in their structure than well formed sentences but still provide a more specific context for disambiguation. In fact, as reported in [9], the WSD algorithm proposed by the authors reaches 66.51% in accuracy which is only 2.61% higher than the baseline, when the most frequently sense is used. This suggests that the annotators should not fully rely on the result of the WSD algorithm and that they may need to check and/or provide the input manually at this annotation phase.

### 4.3 Validation

In order to guarantee the correctness of the assignment of tag splits and tag token senses, two different validators validated each $\langle URL, u \rangle$ pair. The "agreement without chance correction" [11] between users in the task of disambiguation of tokens is of 0.76. As mentioned in [11], there is not yet any accepted best measure for the agreement in the task of sense annotation and thus we currently report only the raw agreement. It is intuitive to think that the agreement will fall when there are more available senses for one token as the annotators will have more chance to choose a different sense. This could also happen because, as we show in Figure 5b), sometimes the annotators cannot decide between too fine grained senses in the controlled vocabulary. Figure 2 shows a more detailled view of the effect of number of available senses on the annotators' agreement.




a) Tags per URL per User b) Tag Reuse across Bookmarks

**Fig. 3.** Use of Tags in del.icio.us

# 5 Analysis

## 5.1 Considerations on the Dataset Uncontrolled Vocabulary

del.icio.us is used in many research groups that work on folksonomies as a large dataset showing how users use tags to organise and share their resources. We have thus started by a basic analysis of how users used tags in the dataset and what we could observe from this. In the following paragraphs, we discuss the analysis that we performed on the whole dataset of 45 600 619 URLs, with all the users and tags available. The analyis and first conclusion on the manual disambiguation batch of 500 $\langle URL, u\rangle$ pairs is discussed in the next section.

While the annotation task on del.icio.us is quite simple as it does not require the specification of semantics, we can already see that the users are not motivated to provide a large amount of annotations. Note that we cannot make any conclusions on why this might be the case as this would require a direct users study, however, as illustrated by Figure 3a), we can see that in 35.5% of the cases, users use only one tag per bookmark and only in 12.1% of the cases they would add more than five tags per bookmark.

This might be because each user only uses very specific tags to classify/categorize the bookmark and thus does not require many indexing terms to find the resource in the future. This assumption would be a "dream" scenario as it would mean that the users are already ready to provide very specific descriptors for their resources and if these descriptors are linked to the underlying controlled vocabulary, we can retrieve them using synonymous and/or more general terms very easily. However, it might just be that the users are not bothered to add more tags as they do not see the value of adding many indexing terms for future retrieval.

An interesting point is that there is an out-of-the-norm peak at ten tags per bookmark that seems too strong to be coincidental. We have not yet studied in details why this happens but hypothesise that it might be created by spambots providing a lot of bookmarks with exactly ten tags.

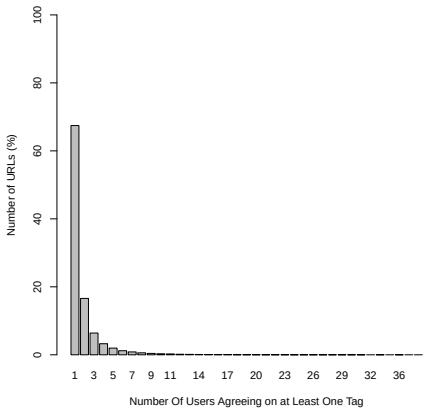



**Fig. 4.** Average Agreement on Tags for the same Resource

In Figure 3b), we consider another interesting feature of the tagging behaviour of users on del.icio.us. While an often used assumptions in folksonomy study algorithms is that we can learn a lot from tag collocations on different resources, we can see that users do not often reuse the same tag more than once. In fact, from our analysis, in 73% of the cases, a tag is used only once on the whole set of bookmarks by a single user. This means that in a majority of the cases, a tag will not be found located on different resources, at least not by the same user. Only in 7.3% of the cases a tag is reused on more than seven resources.

This might support our previous assumption that the users use very specific tags when they annotate resources and thus they do not use them on multiple documents. However, this might create difficulties when sharing knowledge between users as they might not use the same vocabulary (as they use very specific/personal terms). It might also impair the ontology learning algorithms [1] that are based on the measure of collocation of tags.

When annotating shared goods such as web pages, if there is no agreement between the users on what the resource means, it is difficult to reuse these annotations to improve search and ranking of resources. It is also difficult to learn the meaning of the resource or of the annotations attached to it. We have thus done a preliminary analysis of the general agreement of the users in the del.icio.us dataset when they tag a resource. Here we are interested to see how many tags are used by more than one user on the same resource.

To do this, we have adopted a naïve measure of agreement where we count how many users have used the same tag on the same resource. For instance, if there is user $U_1$ who tagged a resource $R_1$ with $T_1$ and $T_2$ while user $U_2$ tagged this resource with $T_3$ and $T_4$, then there is only one user using any of the four tags. If $U_3$ tagged $R_2$ with $T_5$ and $T_6$, $U_4$ tagged it with $T_6$ and $T_7$ and $U_5$ with $T_8$ and $T_9$, then there are two users agreeing on at least one tag for that resource.

Note that we only consider URLs in the dataset bookmarked by at least two users. Figure 4 shows the results of this measure. In 67.5% of the cases, there is only one user "agreeing" on at least one tag, which means different users used different tags on the same resources. In only 9.3% of the cases more than three users agreed on at least one tag.

In a sense this is a good result in that users do provide very diverse tags for the same resource and thus we can learn more about the resource itself. However, if there is no agreement between the users, it is difficult to consider that tags are valid as they might be very personal or subjective.

It is interesting to note that these percentages apply on millions of tags, resources and users and in this, a small percentage still represent a large mass of resources and users on which automatic semantic extraction algorithms can be applied. Also, these figures were computed without any preprocessing of the different forms of tags, or without their disambiguation. As we show in the next section, this might be an important factor for the lack of overlap of tags between resources and users that we are seeing.

However, seeing these results, it is clear that there is a need to create better incentives for the users to provide annotations. In particular, they should be motivated to provide diverse annotation, but also annotations that create a consensus on the meaning of the resources as both these factors are important for leveraging the power of semantic search, navigation and knowledge learning.

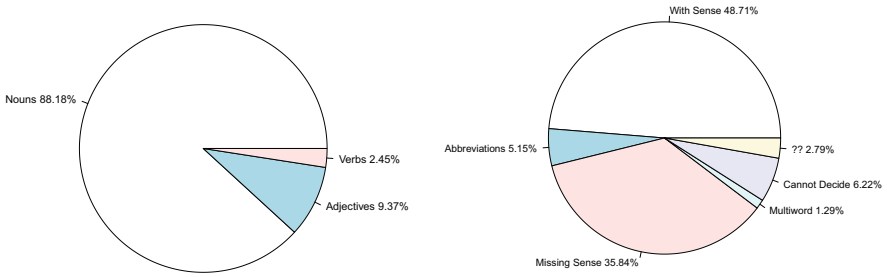
### 5.2 Consideration on the Dataset Controlled Vocabulary

As discussed earlier, we have obtained a quality, disambiguated, sample of the del.icio.us folksonomy for which we know the sense of each tag. In this section, we analyse this subset to see the tagging behaviour when tags are disambiguated to the terms in a controlled vocabulary. In the following paragraphs we present some first conclusions on the use of a controlled vocabulary and how it maps to the users' vocabulary. In the following analysis, we only consider entries that were validated and agreed upon by two validators.

**Use of Nouns, Verbs and Adjectives** In a previous study Dutta et al. [10] point out that the users of del.icio.us tend to use mainly Nouns as descriptors of the urls. In the current dataset we have a validated sense (with all its metadata provided by Wordnet) for each term and thus we can easily reproduce such observation.

Figure 5a) shows that we can come to the same conclusions as [10]. In fact, Nouns are used most of the times (88.18%) while Verbs and Adjectives, even if they are used sometimes cannot be found in great numbers in the annotations. Note that Adverbs seem to be never used, at least in the sample of del.icio.us that we are studying.

**Controlled Vocabulary vs. the Users' Vocabulary** While disambiguating the tags to a sense in Wordnet, the manual annotators could decide that no




a) Part of Speech b) Valid vs. Missing Senses

**Fig. 5.** Properties of Validated Tokens

sense provided by the controlled vocabulary was adequate to express the sense meant by the user. For example, the tag "ajax" was found in the dataset and usually referred to the ajax technology used in web applications[7]. However, the only sense present in Wordnet for this tag is "a mythical Greek hero".

As shown in Figure 5b), the case of the missing sense happened in 35.8% of the cases. However, the validators were able to find a matching sense in Wordnet for 48.7% of the terms used in the validated batch. For diverse reasons (the users use abbreviations, there is no sense in wordnet, etc.) less than half of the vocabulary used by the users can be mapped to the WordNet controlled vocabulary.

This is an important observation as it shows the inadequacy of fully automatic folksonomy processing systems based on fixed controlled vocabularies such as Wordnet. For instance, if we consider the issue of Word Sense Disambiguation, the state-of-the-art tools cannot often achieve more that 60% accuracy. However, given the fact that only half of the terms from our dataset can be found in a vocabulary such as WordNet, from the end user perspective, it means that the user will be suggested the right sense for a given tag token in much less than 60% of cases.

**Sense Disambiguation** One of the issues presented in the raw tags analysis we discussed in Section 5.1 is that there is not a great agreement between users in the tags they use and there is not a great overlap in their personal vocabularies. One of the hypothesis for this is that there are many lexical variations of the same term that cannot be matched without preprocessing the tags (for example, "javaisland", "java_island", "java" and "island", etc.) and as we have already discussed earlier, there are different terms that can be used for the same concept (for example, "trip" and "journey").

In the validation process for the batch, we have actually cleaned all these issues by collapsing different lexical variations and linking them to their relevant

[7] `http://en.wikipedia.org/wiki/Ajax_(programming)`


**Fig. 6.** Decrease in the Amount of ambiguities after pre-processing and after sense disambiguation

concepts. We can thus evaluate the amount of ambiguity that is added by these different type of variations.

Figure 6 shows a summary of this decrease in ambiguity when going from *tags* – that can represent the same word in different forms – to *tokens* – that are preprocessed tags collapsed to the normal form of the world – and then to *synsets* – that disambiguate the meaning of the tag. The top bar represents the number of tags we started from (742), the middle bar represents the number of tokens to which they collapse (265) and the bottom bar represent the number of synsets from wordnet to which these tokens can be mapped.

We can thus see that by preprocessing alone (splitting and lemmatazing tags), the vocabulary size shrinks by 64.7%, thus reducing the ambiguity of the annotations significantly without the need to disambiguate them to the terms in a controlled vocabulary (e.g., a user searching for "blog" will be able to find bookmarks tagged with "blogs", "coolblog", "my_blog", etc.).

The disambiguation provided by the linking to the controlled vocabulary, in the current batch, does not actually provide a great amount of reduction in the vocabulary size. In fact, in the current batch, only seven tokens can be mapped to a smaller set of synsets. This means that there is not a great amount of synonymy in the tags that we have studied.

We believe that this is not a general feature of the full del.icio.us folksonomy and that synonyms and homograph tags will happen in a bigger number in different domains. We are now extending the size of our study batch to observe this hypothesis. In fact, in the current batch, the main topic was focused on computer and web technologies that use a very restricted vocabulary where words do not often have synonyms. We believe that this phenomenon might appear more often in less technical domains and are thus extending our study to the domains of cooking, education and travel.

## 6 Evaluating Semantic Quality of Service

It is often argued that the quality of search would improve if the explicit semantic of the resources were known by the search engine [12]. In order to evaluate this

improvement in the Quality of Service (QoS) of search in annotation systems such as del.icio.us, we implemented and evaluated the performance of a semantic search algorithm in the golden standard. The key difference from keywords-based search algorithms is that instead of using strings as query terms, the algorithm uses concepts from the controlled vocabulary and searches results in the semantically annotated dataset of del.icio.us discussed in Section 3.

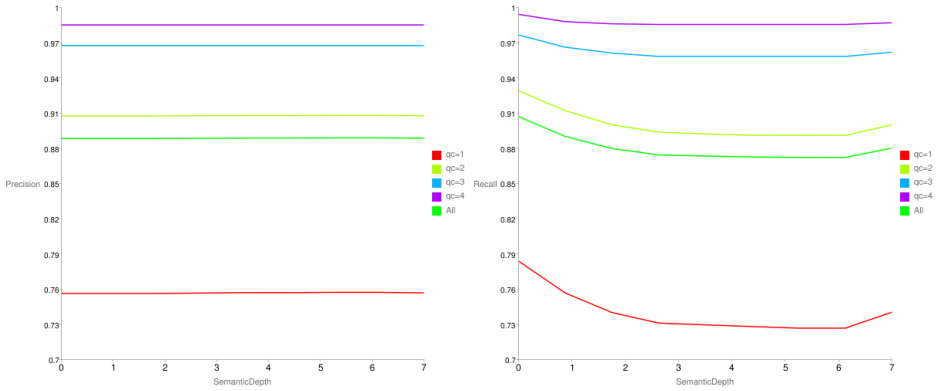
We built queries from validated tag tokens, i.e., tokens for which an agreement on their meaning was reached amongst the validators. The key intuition here was that if the users used these tags to annotate web resources, then they are likely to use the same tags and in the same meaning to find these and other resources.

In order to implement search, we built two indexes: a *keyword index* and a *concept index*. The keyword index contains mappings from tag tokens (e.g., "java") to all the resources annotated with this tag token (e.g., pages about the Java island but also about the programming language, the coffee beverage, etc). The concept index contains mappings from the concepts of the validated tag tokens to all the resources annotated with this tag token in the meaning represented by the concept (e.g., given the token "java" in the meaning of the Java island, the index would point to all resources about the java island but *not* about the programming language or about the coffee beverage). From the golden standard, we generate 377 entries in the concept index, 369 entries in the keyword index, which both point to 262 resources.

Given a number of tag tokens (which corresponds to the desired number of query terms) we built two queries: a *keyword-based query* and a *concept-based query*. The keywords-based query is the conjunction of the token strings, whereas the concepts-based query is the conjunction of the corresponding validated concepts of the tokens. The results of the keyword-based queries might be incorrect and incomplete due to, among other things, the issues discussed in Section 2.2 such as base form variation, homography, synonymy and specificity gap.

The results of the concept-based queries were computed by matching concepts in the query to those in the index. Thus, a query with a particular concept would return all and only resources that have this concept amongst its tag tokens independently of any linguistic variation used to denote this concept in the tag token (e.g., synonymy, homography, as from above). Therefore, the results of concepts-based queries are *correct* and *complete* as long as the meaning of tag tokens in the resource annotations and of the terms in the concepts-based queries is properly disambiguated, which is the case for the analysed dataset due to its manual disambiguation as described in Section 3

In order to address the specificity gap problem, the concept-based search described above was extended to support searching of more specific terms. In this we followed the approach described in [13]. In short, we introduced a variable *"semantic depth"* parameter that indicated the maximum distance between a query concept and a concept according to the `is-a` hierarchy of concepts in the underlying taxonomy in order for a resource annotated with such a concept to be considered as a query result for this query concept. For example, given the following path in the taxonomy: `transport` → `vehicle` → `car` and the query




a) b)

**Fig. 7.** Precision and Recall vs. Semantic Depth

`transport`, then if the semantic depth parameter is set to 1, then resources annotated with the concept `transport` and/or with the concept `vehicle` will be returned as results; if this parameter is set to 2, then the resources annotated with the concept `car` will also be returned.

**Table 1.** Number of Queries

| Query Terms | Queries | Results | |
|---|---|---|---|
| | | Concept Search | Keyword Search |
| 1 | 2 062 | 9.25 | 8.87 |
| 2 | 2 349 | 4.38 | 4.20 |
| 3 | 1 653 | 2.40 | 2.32 |
| 4 | 1 000 | 1.44 | 1.44 |

Queries with different number of query terms and different values for the semantic depth parameter were generated and executed as described above (see Table 1 for details). Given that concepts-based queries, by construction, always yield correct and complete results, their results were taken as the golden standard for the evaluation of the performance of the keywords-based search. The measures of *precision* and *recall* were used for the evaluation. The results of the evaluation for these measures are presented in Figures 7 a) and b) respectively.

As can be seen in Figure 7a), the precision of the keywords-based search with one query term is about 76%, i.e., about 24% of results may not be relevant to the user query. The precision improves for keyword queries with more query terms as the combination of more keywords disambiguates implicitly each keyword (e.g. if we search for the two terms "java island", resources about the programming

language sense of "java" will rarely be returned as they will not have also been tagged with "island"). We can see that the precision of the keyword-based search is not dependent of the query depth. In fact, the number of true results, generated by the concept-based search will augment with the query depth as explained before, however, the number of results returned by the keyword-based search is constant. The precision of the search, which is computed as the number of true positives returned divided by the total number of results returned, is thus constant as the number of results returned does not change and the number of true positives is also always the same.

The recall of the keywords-based search with one query term and with the semantic depth of one is about 79%, i.e., about 21% of *correct* results are *not* returned by the query (see Figure 7b)). With the increase of the semantic depth the recall decreases. This is explained by the fact that concepts-based search is capable of retrieving resources annotated with more specific terms than those used in the query, as discussed above. Therefore, concepts-based search returns more relevant results, whereas the keywords-based search always returns the same results, which leads to a lower recall. Again, as the number of query terms increases, the recall of the keywords-based search improves as the implicit semantic of the keywords is disambiguated by the other keywords.

In practice, in the current evaluation golden standard, the different tags used on resources are very far apart in the taxonomy and thus increasing the semantic depth does not change much the number of results returned by the concept-based search (see Table 1). This is a weak point of our evaluation dataset that is yet not big enough to show a strong *specificity gap* effect. However, we do expect this effect to increase as more concepts in the taxonomy get attached to resources.

As the evaluation results show, the introduction of formal semantics for tags and query terms allows to significantly improve the precision and recall of search in annotation systems such as del.icio.us.

## 7 Related Work

Library catalogs, such as the Library of Congress [14] and Colon Classification [15], are a well known example of classification schemes where experts annotate resources for future search or navigation. The advantage of this model is that the produced classification is considered to be of good quality, which results in a good organization of the resources. On the other hand, Braun et. al. [16] point out to the cost of having dedicated experts annotating and organizing resources and building controlled vocabularies. This issue is underlined by the observation we made in Section 5.2 as these costly controlled vocabularies are not dynamic enough to follow the vocabulary of the users of the annotation system.

Several studies [17–21] analyse how the collaborative model also provides the system designers with behavioural information about the users' interests through their interaction with other users' annotations and their own annotations. However, as we have discussed in Sections 5.1 and 5.2, the problem of

semantic heterogeneity [4, 22] can hinder such analysis as matching tags might not be discovered due to homography, synonymy and morphology issues.

Some [23–25] have proposed to allow end-users to define their classes. Faceteg [26] also follows this direction by incorporating collaborative annotations and collaborative controlled vocabularies in a single system. From the analysis we discussed in Sections 5.2 and 6, such an approach is required to allow the improve the dynamic cataloging of the growing amount of resources available. We believe that the formalisation that we provide in Section 2.2 will help in the storage and reasoning over such complex collaborative annotations.

Word Sense disambiguation (WSD) is known to be a difficult problem [27, 28], especially in the domain of short metadata labels such as the categories names of a Web directories [9, 29] (e.g. DMOZ[8]). Some work also exists on the disambiguation of tags in the domain of folksonomies; According to the classification presented in a recent survey [1], our approach falls under the *ontology-based* category in which the meaning of tags is an explicit association to an ontology element (such as a class or an instance). In [30], the authors preform WSD by defining the context of a tag as the other tags that co-occur with the given tag when describing a resource, and the senses of these tags are used for the disambiguation of the sense of the tag by using the Wu and Palmer similarity measure between the senses [32]. While we use different measures for the computation of the similarities between senses, we extend them with a POS tagger and the frequency of senses to further refine the selection of the tag sense. The WSD approach presented in [31] uses Wikipedia as the source of possible meanings of a tag. To compute the sense candidate, the WSD uses a vectors distance metric between the tag's context and the frequent terms found in the Wikipedia page; note that their approach does not use relations between senses at all for disambiguation. As was pointed out by the authors of [31, 1, 6], without having any golden standards and benchmarks, it is difficult to conduct a comparative analysis with the existing approaches. Therefore, for the time being we can only describe relevant approaches pointing to the differences in algorithms with respect to our approach, however, a quantitative evaluation of our algorithm is provided in Section 4.

## 8 Conclusion

In this article we revisited the classical social annotation model and pointed to some of its shortcomings, which mainly derive from the fact that the model is based on annotations with no formal semantics. We then proposed a model which is based on formal semantics and which can potentially overcome these shortcomings. We then described a process by which the classical model can be converted to the proposed formal model and reported on the results of such a conversion for a subset of a del.icio.us dataset. As our studies showed, the "semantified" allows for a more precise and complete search, which is one of

[8] `http://www.dmoz.org`

the key functionalities in social annotation systems. We observe that a fully automatic conversion of existing folksonomies to the formal model can hardly be possible; it should be a manual task to a significant extent, where the user can be motivated by the improved quality of services such as searching. We also observe that the use of static vocabularies such as WordNet provides ca. 50% coverage of the meaning of the tags, therefore, more dynamically evolved vocabularies need to be provisioned for semantic annotation systems.

## Acknowledgements

This work has been partly supported by the INSEMTIVES project (FP7-231181, see `http://www.insemtives.eu`). We would also like to thank Sergey Kanshin for his contribution to the development of a system used for the evaluation.

## References

1. A., G.S., O., C., H., A., A., G.P.: Review of the state of the art: Discovering and associating semantics to tags in folksonomies. The Knowledge Engineering Review (2010, (To be published))
2. Wal, T.V.: Folksonomy: Coinage and definition. http://www.vanderwal.net/folksonomy.html
3. Mika, P.: Ontologies are us: A unified model of social networks and semantics. Web Semantics: Science, Services and Agents on the WWW **5** (2007) 5–15
4. Golder, S., Huberman, B.A.: The structure of collaborative tagging systems. Journal of Information Science **32**(2) (April 2006) 198208
5. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.: The Description Logic Handbook : Theory, Implementation and Applications. Cambridge University Press (2003)
6. Körner, C., Strohmaier, M.: A call for social tagging datasets. SIGWEB Newsl. (January 2010) 2:1–2:6
7. Miller, G.: WordNet: An electronic Lexical Database. MIT Press (1998)
8. Wetzker, R., Zimmermann, C., Bauckhage, C.: Analyzing Social Bookmarking Systems: A del.icio.us Cookbook. In: Proceedings of the ECAI 2008 Mining Social Data Workshop, IOS Press (2008) 26–30
9. Zaihrayeu, I., Sun, L., Giunchiglia, F., Pan, W., Ju, Q., Chi, M., Huang, X.: From web directories to ontologies: Natural language processing challenges. In: ISWC/ASWC. (2007) 623–636
10. Dutta, B., Giunchiglia, F.: Semantics are actually used. In: International Conference on Semantic Web and Digital Libraries, Trento, Italy (2009) 62–78
11. Artstein, R., Poesio, M.: Inter-Coder Agreement for Computational Linguistics. Journal of Computational Linguistics **34**(4) (2008)
12. : Self-adaptation of ontologies to folksonomies in semantic web. Proc World Acad Sci Eng Tech **33**(September) (2008) 335–341
13. Giunchiglia, F., Kharkevich, U., Zaihrayeu, I.: Concept search. In: ESWC. (2009) 429–444
14. of Congress, L. `http://www.loc.gov/index.html` (last accessed on 01.06.2009).
15. Ranganathan, S.R.: Colon Classification. 7th edn. Asia Pub. House (1987)

16. Braun, S., Schmidt, A., Walter, A., Nagypal, G., Zacharias, V.: Ontology maturing: a collaborative web 2.0 approach to ontology engineering. In: Proceedings of (CKC 2007) at (WWW2007). (2007)
17. Mika, P.: Ontologies are us: A unified model of social networks and semantics. In: International Semantic Web Conference. (2005) 522–536
18. Schmitz, P.: Inducing ontology from flickr tags. In: Proc. of the Collaborative Web Tagging Workshop (WWW 06). (May 2006)
19. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. J. Inf. Sci. **32** (April 2006) 198–208
20. Marlow, C., Naaman, M., Boyd, D., Davis, M.: Ht06, tagging paper, taxonomy, flickr, academic article, to read. In: Proceedings of the seventeenth conference on Hypertext and hypermedia. HYPERTEXT '06, NY, USA, ACM (2006) 31–40
21. Körner, C., Benz, D., Hotho, A., Strohmaier, M., Gerd, S.: Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In: Proceedings of WWW '10, NY, USA, ACM (2010) 521–530
22. Kolbitsch, J.: WordFlickr: a solution to the vocabulary problem in social tagging systems. In: Proceedings of I-MEDIA. (2007)
23. Mathes, A.: Folksonomies - cooperative classification and communication through shared metadata. Technical report, Graduate School of Library and Information Science. University of Illinois Urbana-Champaign (December 2004)
24. Gazan, R.: Social annotations in digital library collections. D-Lib **Volume 14 Number 11/12** (December 2008)
25. Ronzano, F., Marchetti, A., Tesconi, M.: Tagpedia: a semantic reference to describe and search for web resources. In: SWKM 2008: Intl. Workshop on Social Web and Knowledge Management @ WWW. (2008)
26. Quintarelli, M., Resmini, A., Rosati, L.: Facetag: Integrating bottom-up and top-down classification in a social tagging system. In: IASummit, Las Vegas (2007)
27. Yang, C., Hung, J.C.: Word sense determination using wordnet and sense co-occurrence. In: Proceedings of the 20th International Conference on Advanced Information Networking and Applications - Vol 1, USA, IEEE (2006) 779–784
28. Agirre, E., Rigau, G.: A proposal for word sense disambiguation using conceptual distance. In: the First International Conference on Recent Advances in NLP, Tzigov Chark, Bulgaria (September 1995)
29. Autayeu, A., Giunchiglia, F., Andrews, P.: Lightweight parsing of classifications into lightweight ontologies. In: ECDL. (2010) 327–339
30. Angeletou, S., Sabou, M., Motta, E.: Semantically enriching folksonomies with flor. In: In Proc of the 5th ESWC. workshop: Collective Intelligence & the Semantic Web. (2008)
31. Garca-Silva, A., Szomszor, M., Alani, H., Corcho, O.: Preliminary results in tag disambiguation using dbpedia. In: Proc. of the First International Workshop on Collective Knowledge Capturing and Representation (KCAP), USA (2009)
32. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: Proc. of the 32nd annual meeting on Association for Computational Linguistics. (1994) 133–138