Preliminary Results in Tag Disambiguation using DBpedia

Andrés García-Silva
Ontology Engineering Group, Facultad de Informática, Universidad Politécnica de Madrid. Madrid, Spain
hgarcia@fi.upm.es

Martin Szomszor
School of Electronics and Computer Science, University of Southampton, SO16 1BJ, UK.
mns2@ecs.soton.ac.uk

Harith Alani
School of Electronics and Computer Science, University of Southampton, SO16 1BJ, UK.
h.alani@ecs.soton.ac.uk

Oscar Corcho
Ontology Engineering Group, Facultad de Informática, Universidad Politécnica de Madrid. Madrid, Spain.
ocorcho@fi.upm.es

ABSTRACT
The availability of tag-based user-generated content for a variety of Web resources (music, photos, videos, text, etc.) has largely increased in the last years. Users can assign tags freely and then use them to share and retrieve information. However, tag-based sharing and retrieval is not optimal due to the fact that tags are plain text labels without an explicit or formal meaning, and hence polysemy and synonymy should be dealt with appropriately. To ameliorate these problems, we propose a context-based tag disambiguation algorithm that selects the meaning of a tag among a set of candidate DBpedia entries, using a common information retrieval similarity measure. The most similar DBpedia entry is selected as the one representing the meaning of the tag. We describe and analyze some preliminary results, and discuss about current challenges in this area.

Keywords
Folksonomy, Ontology, Semantic, Learning, Disambiguation, Wikipedia, DBpedia.

INTRODUCTION
Folksonomies emerge in the Web 2.0 as a result of tagging processes. These are performed as a way to index and retrieve information in such an environment where information creation is not centralized in the owners of the web sites like in the traditional web, but distributed across users. Tagging has been successful mainly because users do not need special skills to perform this task and they get benefits instantaneously without too much effort [13]. Traditional information classification schemes are maintained by a closed group of people who create a taxonomy and then place resources under a particular category. In contrast, folksonomies are tailored to the needs of each user since categories, in this case tags, are freely created by users according to the context in which the user is tagging a resource.

K-CAP'09, September 1-4, 2009, Redondo Beach, California, USA.

As time goes by, some tags are used more frequently than others to annotate a resource, usually because of tag recommendation strategies relying on the most frequently preassigned tags [3]. This stabilization of the vocabulary used to annotate a resource can be seen as an agreement among users about the concept used to annotate a resource.

However, current folksonomies lack of an explicit semantics of tags, hampering the information sharing and retrieval [1; 3; 7; 12]. For instance, users can use different tags to represent the same concept when annotating a resource but the system is not aware of this relation, and thus it cannot take advantage of this information when retrieving resources.

In more detail, problems related to the lack of semantics affecting applications relying on folksonomies are caused by: 1) morphological variations to represent the same tag such as plurals, acronyms, conjugated verbs or misspelling words, 2) the use of synonyms, 3) the use of polysemous tags, and 4) the so-called basic level problem, which refers to the use of different tags to annotate a resource according to the level of expertise of the user in a particular domain.

For instance, polysemous tags are those that have more than one meaning. When a user tries to retrieve information using a polysemous tag, he can receive unintended results due to the fact that the system retrieves resources tagged with that particular tag regardless of the intended meaning of the tag. Currently, Delicious allows users to assign text descriptions about the meaning of a tag. However, these text descriptions, written in natural language, are intended to be used by users and not by machines. Thus, the system cannot easily take advantage of those textual descriptions to improve the information retrieval process.

In this paper we present an approach to automatically disambiguate polysemous tags, although it can be extended to deal with the other problems. The approach relies on DBpedia and Wikipedia information, being the former an RDF structured representation of some of the information from the latter. Each tag in a specific tagging activity is related to many DBpedia entries, which define the possible meaning of that tag. From the Wikipedia page related to each DBpedia entry we have extracted the most frequent terms and their frequency. We compare the term information of each DBpedia entry with the terms in the tag context. This comparison is carried out modeling the information about the DBpedia entries and the tag context as vec-

tors, and then calculating the similarity among those vectors using as measure the cosine of the angle they form.

The document is structured as follows. First, we describe related work. Then, in the Sense Repository section we present how we relate tags to DBpedia entries and how we extract from Wikipedia the information about the most frequent terms. After that, we present our notion of context in folksonomies. In the Disambiguation Approach section we describe the disambiguation algorithm. Next, we present some disambiguation results using the proposed algorithm. Finally, we present the conclusions.

RELATED WORK

Folksonomy-related challenges have been studied by the semantic web research community from two perspectives. First, folksonomy information is not easy to share and reuse due to the fact that they lack a standard representation. Research works as [16, 17] have proposed ontologies to model the tagging information by means of concepts as *tagger*, *resource*, *tag*, and *tagging*. The SCOT ontology [8] extends Newman ontology [17] with the *TagCloud* concept, which represents aggregated user information like all user tags, frequencies of use, and tag co-occurrence. The MOAT ontology [18] allows defining the meaning of a tag by means of its association with ontology URIs. A survey of these ontologies can be found in [6].

The second aspect is that tags in folksonomies lack of formal and explicit semantics. In this respect, [1, 2, 11] have associated tags with ontologies. The approach proposed in [12] assigns Wikipedia URIs to tags. However, tags related to more than one Wikipedia page are discarded. A manual approach where users assign ontology concepts to tags is suggested in [10]. On the other hand, other research works like [8,5] have focused on the tag co-occurrence to form groups of related tags. These groups define implicitly the meaning of the tags. Mika [8] proposes to use set operations and some social networks metrics to identify the semantic of the tags. In [5] authors suggest creating a concept lattice from a particular tag set. Then, the graphical representation of the concept lattice can be analyzed by an ontology engineer to create manually an ontology.

The disambiguation of tags has been addressed in [1, 4, 11]. In [11] authors create groups of related tags based solely on their co-occurrence. However, when a tag is ambiguous it can have more than one pattern of co-occurrence. Thus, groups can contain co-occurring tags related to more than one meaning. Authors propose to create subgroups of tags based on high co-occurrence. Those subgroups are supposed to cluster tags according to one meaning. Nevertheless, the association of an ontology concept to the tag is carried out manually.

The approach proposed in [1] analyzes a tag set. If a tag has more than one sense in WordNet, then its hierarchy of senses extracted from WordNet is used to calculate the similarity with the senses of all tags in the tag set. The most similar sense to the senses of all tags in the tag set is selected as the meaning of the analyzed tag. Then, the sense information of the tag is used to associate automatically this tag to an ontology concept.

Hamasaky *et al*. [4], propose an algorithm for disambiguation, based on the idea that if a tag is used to annotate different instances by different groups of users (neighbors), the tag may have different meanings. Otherwise, the tag has the same meaning. The proposed algorithm treats each user tag as a pre-concept, and then these pre-concepts are merged if they have the same labels and share the same users/resources or neighboring users. However, this approach does not define explicitly the meaning of the tags.

There are several differences of our approach with respect to the ones presented in this section. First, our disambiguation approach is completely automatic, in contrast to [11]. We define the meaning of tags by associating them to DBpedia concepts, unlike [4] that does not make any association to existing vocabularies but lets a vocabulary emerge independently. Finally, we use DBPedia, instead of WordNet, to disambiguate the tag meaning, in contrast to [1]. Tags reflect the changes in the user vocabulary. The main difference between them is that WordNet is maintained by a close group of people and it does not evolve at the same speed as the folksonomy vocabulary does. In contrast, Wikipedia (hence DBPedia) is collaborative maintained and continuously evolving.

SENSE REPOSITORY

The proposed disambiguation approach uses as dictionary the TAGora sense repository (TSR) where tags are related to DBpedia concepts and Wikipedia pages. TSR is a linked data enabled service endpoint that provides extensive metadata about tags and their possible senses. When the TSR is queried with a particular tag string, by forming a URI that contains the tag in a REST style (e.g. http://tagora.ecs.soton.ac.uk/tag/apple/rdf), the tag is processed, grounded to a set of DBpedia.org resources[1], and an RDF document is returned containing the results. For the purposes of this experiment, we also provide a SPARQL end-point.

Creating The Resource Index The first stage in building the TSR was to process the XML dump of all Wikipedia pages to index all titles, mine redirection and disambiguation links, and extract term frequencies for each of the pages. For the current version we use a dump available from http://download.wikimedia.org, created on the 08/10/2008. For each Wikipedia page in the dump, we extract and index the page title, a lower case version of the title, and a concatenated version of the title (i.e. the title Second_life be-

[1] According to wiki.dbpedia.org/About DBpedia currently describes 2.6 million things, including at least 213,000 people, 328,000 places, 57,000 music albums, 36,000 films, 20,000 companies

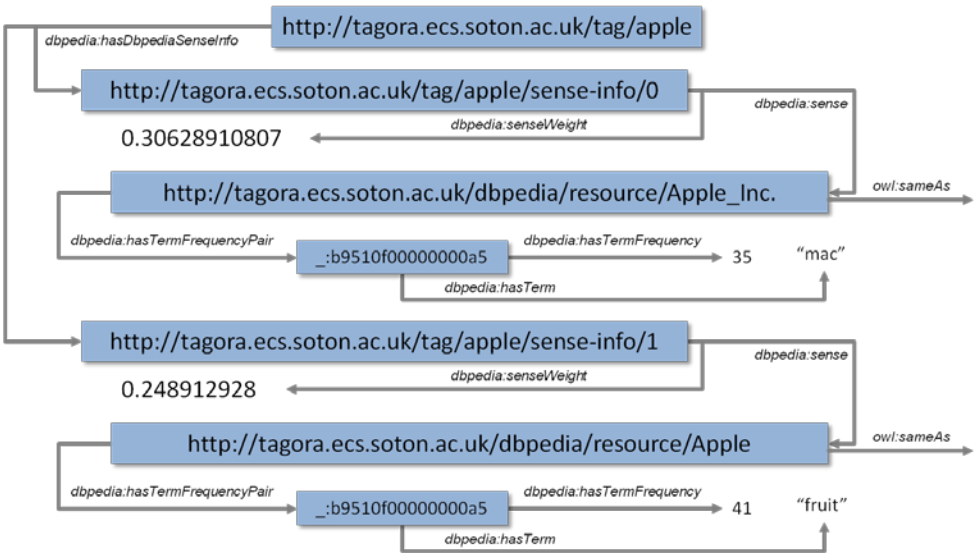

Figure 1. Linked data representation of tag senses.

comes secondlife). This multiple title indexing enables us to match more easily tags that are made up of compound terms. We also extract redirection links, disambiguation links, as well as the terms contained in the page and their frequencies. During this indexing process, we also store a list (and total) of all incoming links to each page. Since the dump is large, we only store terms with a frequency greater than the mean frequency of all terms in that page. This data is stored in a Triple Store using our own extended DBpedia ontology[2] since we are providing more detailed metadata about the entries than DBpedia.org such as the term frequencies. Each Wikipedia page in the TSR is also linked to DBpedia via the owl:sameAs property.

Searching For Senses When the TSR is queried with a tag, the first step is to find a list of candidate DBpedia resources that represent possible senses of the tag. We begin by normalizing the tag string (i.e. removing non-alphanumeric characters as described in [12]). The Triple Store is then queried for all entries with the same lowercase title or concatenated title as the tag. During this process, we are likely to encounter redirection links and /or disambiguation links, both of which are followed. When a set of candidate senses has been created, we calculate the total number of incoming links for each resource (including the sum of incoming links for any pages that redirect to it). Finally, a weight is associated with each possible sense as the fraction of incoming links associated with that sense / the total number of incoming links for all senses associated with the tag. This basic page rank inspired measure means senses that have very specific meanings receive much lower weights than general those associated with general concepts. Note that TSR associates tags to concepts or instances since DBpedia resources identify instances or concepts such as London and City.

Figure 1 provides a visual example of the linked data associated with the tag 'apple' – a common tag that could refer the computer company (*Apple_Inc.*), or the of fruit (*Apple*). In this example, the URI for *apple* (center, top) is linked to a number of sense-info instances (only two of which are shown here) via the *dbpedia:hasdbpediaSenseInfo*. Each sense-info pair gives the weight (0.306 for *Apple_Inc*. and 0.249 for *Apple*) and corresponding DBpedia resource. Each resource is linked to a set of blank nodes (of type *termFrequencyPair)* that states the frequencies of terms within the Wikipedia page of that resource.

CONTEXT IN FOLKSONOMIES

Disambiguation processes are strongly dependent on the notion of context. In this section we describe the notion of context in disambiguation approaches based on dictionaries. Then, we present our notion of context in folksonomies exploiting the different levels of information provided by folksonomies.

In [19] the author proposed a word sense disambiguation algorithm relying on a dictionary. This approach retrieves from the dictionary the sense definitions of the word to disambiguate. Then, for each word in the context, except for the stop words, the corresponding sense definitions are retrieved from the dictionary. In this case, the context are words in the sentence where the word to disambiguate appears. Finally, the definitions of the words in the context are compared against the sense definitions of the word to disambiguate. The most similar sense according to the terms shared among the context and the sense is selected as the meaning of the analyzed word.

Tags in folksonomies do not appear in sentences as it occurs in natural language paragraphs. However, tags usually

[2] http://tagora.ecs.soton.ac.uk/schemas/dbpedia

co-occur with other tags, and this co-occurrence information can be used to build a context for specific tags. For instance, when a user is tagging a resource with a set of tags, those co-occurring tags can be considered as the context in which the user is using them. Nevertheless, in disambiguation processes based on dictionary definitions there are problems when the definitions are short because it is likely that none of the definition words overlap with the words in the context [20]. In the same way, the co-occurring tags when a user is annotating a resource might not give enough information to disambiguate the meaning of ambiguous tags. Thus, we can take into account for instance, all the tags co-occurring in that resource regardless of the user. This new information can provide more evidence about the meaning of a tag.

Following this idea, we propose to exploit different definitions of tag co-occurrences in a folksonomy to define the context where the tag was used by a user when annotating a resource:

- User tags co-occurring in the same resource.
- Co-occurring tags in the resource regardless of the user. In broad folksonomies[3] it is likely that other users have tagged the resource.
- User tags regardless of the resource. We suppose that the whole user vocabulary can provide some clues when disambiguating one of his tags. Moreover, we can take advantage of the user vocabulary in the period of time when he tagged the resource with the ambiguous tag. This period of time could be measured from minutes to years.
- Tags co-occurring in the user social network. In [4] authors suggest that the tags of the user contacts can be used as a complement when the user tagging information is scarce.
- Co-occurring tags in the whole folksonomy when annotating any resource regardless of the users. This information is useful to find the most used meaning of tag in the whole folksonomy.

In this paper we only use the first type of context: user tags co-occurring in the same resource. We are planning to test the other context definitions in an iterative approach where we start with an initial context to disambiguate tags. If context information is not useful to disambiguate the tag meaning (e.g., none of the tags in the context are related with the tag definitions), we can extend the current context adding information of one of the other defined contexts.

DISAMBIGUATION APPROACH

We are interested in the definition of the intended meaning of a tag when it is used by a user to annotate a resource. Our disambiguation approach relies on the TSR where tags are associated with DBpedia entries, and in the notion of tag context defined in the previous section. Inspired by information retrieval techniques [13], the sense and the tag context information are represented by means of vectors, which then can be compared by measuring the angle between them.

Vector Representation

Folksonomies are formalized in [8] as a tuple F= <T, R, U, Y> where T is the set of tags, R is the set of resources, U is the set of user, and $Y \subseteq U \times T \times R$ is the relation denoting a tagging activity, that is, a user annotates a resource with a tag. We extended this basic model to include information about senses and their terms and how those senses are related to tags. Thus, our model can be defined as $F' = <T,R,U,Y,S,W,X,Z>$, where S is the set of senses, and W is the set of terms related to those sense. $X \subseteq S \times W$ is the relation between senses and terms, and $Z \subseteq S \times T$ is the relation between tags and senses.

The disambiguation process objective is to choose the most likely sense among all the senses that have been associated to a tag, according to the context in which the tag has been used. This disambiguation is carried out taking into account the terms of each sense. Thus, we need to define the following data sets related to a tag.

- $Senses(t \in T) = \{s_j : (t,s_j) \in Z\}$. The set of the senses associated to a tag.
- $Terms(s \in S) = \{w_k : (s,w_k) \in W\}$. The set of terms associated to a sense.
- $Voc(t \in T) = \cup Terms(s_j)$: $s_j \in Senses(t)$. The set of terms of all the senses associated to a tag. Voc stands for vocabulary.

On the other hand, we need to define the context of a tag.

- $Context(u \in U, t \in T, r \in R) = \{t_l \in T : (u, t_l, r) \in Y\}$. The set of all co-occurring tags used by the user u to annotate the resource r.

Now we can define a tag, its context and the senses associated to that tag using vectors. Those vectors are in $\Re^{|Voc(t)|}$ and each position in the vector corresponds to a term in Voc(t).

- A tag and its context can be represented by the vector $V_{context} = (v_i)$, where $1 \leq i \leq |Voc(t)|$ and $v_i = 1$ if the corresponding term w_i in Voc(t) appears in Context(u,t,r), otherwise $v_i = 0$.
- The sense associated to a tag can be represented by the vector $V_{sense} = (v_i)$ where $1 \leq i \leq |Voc(t)|$ and v_i is the frequency of the corresponding term $w_i \in Voc(t)$ in the sense.

In this way, when we want to analyze a tag we can create a $V_{context}$ for the specific tag and then a V_{sense} for each sense related to that tag. Then, we can compare $V_{context}$ with each

[3] In http://www.vanderwal.net/ broad folksonomies are defined as those where many people tag the same object and every person can tag the object with their own tags in their own vocabulary.

V_{sense} using a similarity measure. A well known similarity measure among vectors is the cosine function (1). The cosine of the angle between two vectors is a value between 0 and 1. When the angle is small the cosine value tends to 1, when the angle is big the cosine value tends to 0.

$$Sim(V_{context}, V_{sense}) = \cos\theta = \frac{V_{context} \cdot V_{sense}}{|V_{context}||V_{sense}|} \quad (1)$$

For example, we want to disambiguate the tag *nature* that was used along with the tags *news*, and *science* by a user *u* to annotate the resource *r = http://www.nature.com*, the website of the scientific journal Nature. Thus, context(*u*,*nature*,*r*)= {*nature, news, science*}.

Let us suppose that in the sense repository we have:

- Senses(*Nature*) = {*dbpedia:Nature, dbpedia: Nature_(journal)*}
- Terms(*dbpedia:Nature*) = {*life, nature, earth*} being the respective frequencies of those terms in the sense (62, 46, 32)
- Terms(*dbpedia:Nature_(journal)*) = {*nature, science, scientific*} being the respective frequencies of those terms in the sense (77, 29, 25)

With this information we can gather the terms of all the senses to create the set Voc(*nature*) = { *life, nature, earth, science, scientific*}. The next step is to create the vectors in $\Re^5$ to represent the tag context and the senses.

- $V_{context}$= (0,1,0,1,0)
- V_{nature}= (62,46,32,0,0)
- $V_{nature(journal)}$= (0,77,0,29,25)

Now we can calculate the similarity between the vectors. Sim($V_{context}$, V_{nature}) = 0,389 and Sim($V_{context}$, $V_{nature(journal)}$) = 0,872. Therefore, we can assert that the most probable meaning of the tag *nature* according to its context is given by the *dbpedia.org:Nature_(journal)* entry.

Disambiguation Algorithm

The disambiguation process (See Figure 1) takes as input a tagging activity described by a user u, a tag t and a resource r. First, $V_{context}$ is created using Voc(t) and Context(u,t,r), the vocabulary associated to the tag and the tags in the context respectively. Second, for each sense associated to the tag, V_{sense} is created using Voc(t) and Terms(s), the vocabulary associated to the tag and the terms of the sense respectively. Then, the similarity value is calculated by means of the cosine of the angle between the two vectors. The variables *maxSimilarityValue* and *mostSimilarSense* are used to keep track of the sense with higher similiarity value. Finally, when all the senses have been processed, the most similar sense is returned.

This algorithm allows us to use several definition of the context of a tag such as the ones presented in the Context in Folksonomies section. We plan to test the other definitions of context in order to find the best context definition to achieve the best disambiguation results.

```
Disambiguation(User u, Tag t, Resource r)
  V_context= createContextVector(Voc(t),Context(u,t,r))
  Sense mostSimilarSense = null
  maxSimilarityValue = 0
  For each s in Senses(t) do
     V_sense=createSenseVector(Voc(t),Terms(sense))
     similarityValue = cosine(V_context,V_sense)
     If similarityValue > maxSimilarityValue then
        mostSimilarSense = sense
        maxSimilarityValue = similarityValue
     end if
  End for
  Return mostSimilarSense
```

Figure 1. Disambiguation algorithm.

PRELIMINARY RESULTS

Tagging data, including user, tags, resources, and tagging activities, are stored in an RDF triple store, which provides a SPARQL endpoint to query the data. The TSR provides also a SPARQL endpoint to query the information about the DBpedia entries related to a tag. Hence we use ARQ (a SPARQL processor for Jena) to query the triple store. The input of the program is a tuple describing a user, a tag and a resource, and the output is the DBpedia page URL chosen by the algorithm as the tag intended meaning.

In this section, we want to discuss some real examples we have used to test our disambiguation algorithm.

Running in London

The first example is about a tagging activity where the resource is a picture of a group of amateur runners running in the streets of a city. The user has annotated the photo with the tags *london*, *londonmarathon*, and *running*. *londonmarathon* and *running* are associated in the TSR to the entries *dbpedia/resource/London_Marathon* and *dbpedia/resource/Running* respectively. On the other hand, the *london* tag has been associated to 91 DBpedia entries.

In Table 1 we present just 15 of these entries and their similarity value. The disambiguation algorithm chose the *dbpedia/resource/London* entry as the most similar entry to the tag and its context. This entry refers to London the capital of the UK, where term *London* appears 324 times. However, the tags *running* and *marathon* appear 3 and 2 times in this entry. In this case, the disambiguation was successful because the high frequency of the tag *London* and in a lower scale because of the tags in the context.

Table 1. Disambiguation results for the *london* tag

london	
dbpedia/resource/London	**0,905**
dbpedia/resource/London,_Arkansas	0,222
dbpedia/resource/London,_California	0,179
dbpedia/resource/London,_Kentucky	0,224
dbpedia/resource/London,_Ohio	0,362
dbpedia/resource/London,_Ontario	0,710
dbpedia/resource/London_and_Croydon_Railway	0,590
dbpedia/resource/London_and_North_Eastern_Railway	0,203
dbpedia/resource/London_and_North_Western_Railway	0,435
dbpedia/resource/London_and_Port_Stanley_Railway	0,358
dbpedia/resource/London_Broil	0,380
dbpedia/resource/London_Labour_and_the_London_Poor	0,594
dbpedia/resource/London_Majors	0,595
dbpedia/resource/London_Marathon	0,522
dbpedia/resource/The_London_Magazine	0,244

Ice Skating

Let us analyze a more complex tagging activity. Some user x has tagged a picture r with the tags *ice*, *iceskating*, *nottingham*, and *skating*. The picture is about a group of people in an ice rink wearing winter apparel and ice-skates.

According to the TSR the tag *iceskating*, which has been preprocessed and split in *ice skating*, is associated to a unique entry: *dbpedia/resource/Ice_skating*. The rest of the tags have associated more than one entry. The tags, their DBpedia entries and the similarity measure are shown on Table 2. Our disambiguation algorithm assigned to the tags *ice*, *skating* and *nottingham* the entries *dbpedia/resource/Ice*, *dbpedia/resource/Ice_skating* and *dbpedia/resource/Nottingham* respectively.

Let us analyze in more detail the tag *skating*. This tag has 10 possible meanings according to DBpedia. The set Voc(*skating*) has 77 terms gathered from the associated senses. The tags *ice* and *skating* that belong to the *skating* context appears in Voc(*skating*). This means that vectors $V_{context}$, and V_{sense} can be compared by means of these two shared tags. In general, the similarity value is greater than 0 when at least one of the tags in the context is present in the most frequent terms of the DBpedia entry.

The two entries with highest similarity are *dbpedia/resource/Ice_skating* and *dbpedia/resource/Tour_skating*. The tags *ice* and *skating* appear 46 and 25 times respectively in the former, and 6 and 11 times in the latter. The algorithm chose *Ice_skating* as the meaning of the tag *skating* in this context. However, *Tour_skating* also achieved a high similarity value. This entry has in the TSR just 4 terms: *skating, ice, tour, and sweeden*. Thus, the similarity between $V_{tour\_skating}$ and $V_{context}$ is high because they share the terms *ice* and *skating* and differs only in the other two terms. On the other hand, *Ice_skating* has in the TSR 12 terms. In this case the number of tags in which $V_{context}$ and $V_{ice\_skating}$ differs is in 10 terms. Nevertheless, this difference is compensated with the high frequency of the terms shared between $V_{context}$ and $V_{ice\_skating}$.

Table 2. Similarity measures: *ice, Nottingham, skating*

ice	
dbpedia/resource/Ice	**0,911**
dbpedia/resource/Ice_(comics)	0,735
skating	
dbpedia/resource/Artistic_roller_skating	0,671
dbpedia/resource/Figure_skating	0,569
dbpedia/resource/Freestyle_slalom_skating	0,000
dbpedia/resource/Ice_skating	**0,893**
dbpedia/resource/Road_skating	0,451
dbpedia/resource/Roller_skating	0,394
dbpedia/resource/Skateboarding	0,197
dbpedia/resource/Snowboarding	0,000
dbpedia/resource/Speed_skating	0,549
dbpedia/resource/Tour_skating	0,831
nottingham	
dbpedia/resource/East_Nottingham_Township,_Pennsylvania	0,000
dbpedia/resource/Elizabeth_I_of_England	0,000
dbpedia/resource/Nottingham	**0,750**
dbpedia/resource/Nottingham,_New_Hampshire	0,386
dbpedia/resource/Nottingham_Cooperative	0,524
dbpedia/resource/Nottingham_Township,_Harrison_County,_Ohio	0,000
dbpedia/resource/Nottingham_Township,_Pennsylvania	0,000
dbpedia/resource/Nottinghamshire	0,428
dbpedia/resource/Sheriff_of_Nottingham	0,640
dbpedia/resource/West_Nottingham_Township,_Pennsylvania	0,000

The tag *nottingham* is a special case where only one tag in the context, in this case the same tag, is also in Voc(*nottingham*). This means that the comparison among $V_{context}$ and V_{sense} is made basically in terms of the frequency of the *nottingham* term in the respective sense, and in the amount of terms of each sense in Voc(*Nottingham*). Let us analyze this correlation among similarity, term frequency and sense terms in Voc(*tag*).

Table 3. Detail analysis of *nottingham* tag disambiguation

DBpedia Entry	Freq (*nottingham*)	Terms in Voc(*nottingham*)
../Nottingham	181	16
../Sheriff_of_Nottingham	9	3
../Nottingham_Cooperative	12	12
../Nottinghamshire	27	15
../Nottingham,_New_Hampshire	14	17

Table 3 shows detailed information about the DBpedia entries associated to the *nottingham* tag, which are ordered from the highest similarity value to the lowest. The *dbpedia/resource/Nottingham* entry, the one with the highest similarity value, has the highest frequency of the term *nottingham* and an average number of terms in Voc(*nottingham)*.

The *dbpedia/resource/Sheriff_of_Nottingham* has the lowest frequency of *nottingham* but also has the lowest number of terms in Voc(*nottingham)*. As we mentioned previously, just one tag in the context appears in Voc(*nottingham*), thus, $V_{context}$ and $V_{Sheriff\_of\_Nottingham}$ are highly similar because they share one term and differ just in two terms.

We can see that even if a sense has a high frequency of the term n*ottingham* like in *dbpedia/resource/Nottinghamshire*, one with a lower term frequency like *dbpedia/resource/Nottingham_Cooperative* can achieve a higher similarity due to the fact that the latter has a lower number

of terms in Voc(*nottingham*). This characteristic is also present when a user has used just one tag to annotate one resource. We have to think in how to improve the algorithm to deal with DBpedia entries having a low number of terms in Voc(*tag*) since they can bias the similarity results. One possibility is to assign a weight to each DBpedia entry according to the number of terms in Voc(tag), and then we can take into account this weight to modify the similarity formula.

According to the resource annotated, in this case a picture describing a group of people in an ice rink, we can say that the tags *ice* and *skating* have been correctly disambiguated. We cannot say anything about the tag *Nottingham*. The picture does not provide any evidence of where it was taken. The other tags the user assigned to the picture are not enough information to clear state what is the intended meaning of the tag. However, as the resource is a photo, it is likely that *nottingham* refers to the geographical location where the photo was taken.

Thus, when the tag context does not provide more evidence to disambiguate the tag, the tag is a geographical location, and the resource is a photo, we can modify the similarity formula in order to assign a higher similarity value to those DBpedia entries specifying a geographical location.

Holidays in Italy

Sometimes the disambiguation algorithm fails to assign the right meaning to a tag. For instance, a user has assigned the tags *italy*, *siena*, *tuscany*, and *holiday* to a picture in which a guy is eating a sandwich in a town. The tag *siena* in our sense repository is associated to *dbpedia/resource/Siena* entry. Table 4 shows the disambiguation results for the tags *italy*, *tuscany*, and *holiday*. The tag *italy* as a country and the tag *tuscany* as a region of Italy were properly assigned to the corresponding DBpedia entries.

The tag *holiday* was assigned to *dbpedia/resource/holiday_(TV_series)*, a UK television program about travelling on holidays. We cannot say for sure that this assignment is wrong because we don´t know what the user had in mind when he tagged the photo. However, the entry *dbpedia/resource/Holiday* is more general, and can be also assigned to the *holiday* tag in this context. Again, in this case none of the tags in the context appears as frequent terms in the DBpedia entries associated to the tag *holiday*, and thus the similarity value is conditioned to the frequency of the *holiday* term, and the number of frequent terms in the entry.

CONCLUSIONS AND FUTURE WORK

In this paper we have presented a tag disambiguation algorithm, whose goal is to rank and select DBpedia entries representing the meaning of the analyzed tag. The algorithm relies on a vector representation of the tag context and of the candidate DBpedia entries using a common vocabulary, based on term frequency of the candidate DBPedia entries. These vectors are compared using a common cosine-based similarity measure, and the most similar candidate DBpedia entry is selected as the tag meaning.

Table 4. Disambiguation of tags *italy, tuscany, and holiday*

italy	
dbpedia/resource/Italy	**0,792**
dbpedia/resource/Italy,_Texas	0,364
dbpedia/resource/Italy_Again	0,577
dbpedia/resource/Italy_First	0,778
dbpedia/resource/Italy_of_Values	0,176
dbpedia/resource/Italy_Runestones	0,319
dbpedia/resource/Italy-USA_Foundation	0,625
tuscany	
dbpedia/resource/Apache_Tuscany	0,302
dbpedia/resource/Grand_Duchy_of_Tuscany	0,347
dbpedia/resource/Hugh_of_Tuscany	0,355
dbpedia/resource/Matilda_of_Tuscany	0,104
dbpedia/resource/Renaissance	0,116
dbpedia/resource/Toscana_(wine)	0,122
dbpedia/resource/Tuscany	**0,682**
dbpedia/resource/Tuscany_Lion	0,297
holiday	
dbpedia/resource/Billie_Holiday	0,637
dbpedia/resource/Christmas_and_holiday_season	0,545
dbpedia/resource/Christmas_music	0,179
dbpedia/resource/Holiday	0,459
dbpedia/resource/Holiday,_Florida	0,132
dbpedia/resource/Holiday_(TV_series)	**0,815**
dbpedia/resource/Holiday_Affair	0,000
dbpedia/resource/Holiday_Records	0,369
dbpedia/resource/The_Holiday	0,149

To improve poor results that may be due to data scarceness in folksonomies, we propose using the notion of tag context. We have provided several definitions and have tested with the simplest one, providing some preliminary results of the execution of the algorithm using real tagging examples.

All our work is based on the general assumption that the association of tags and ontology components is useful to improve tasks like search, to tagging recommendation strategies, etc. For instance, once the disambiguation algorithm selects an ontology component (e.g., a DBPedia entry) regardless if it is an instance or a concept, the search process can use its semantic relations to make more specific queries. Besides, tag recommendations strategies can benefit from those relations to suggest new tags to users, so that they are not only based on syntactic similarities or co-occurrence.

Our future work will be focused on several activities: first, we plan to **test this approach more systematically, with a larger set of tagsets and with different context definitions**. The goal is to gather enough data to analyze which context definition allows better disambiguation results, or to identify the situations where it is better to use a specific type of context according to some criteria. In cases where the tagged resource is a text document, we would like to evaluate the use of natural language processors to extract meaningful terms that then can be added to the context.

Furthermore, we want to test **more sophisticated similarity measures** that behave better with DBpedia entries with a low number of terms. An important open challenge is how to process those tags whose context terms do not overlap with DBpedia entry terms.

We also want to **adapt our algorithm to very specific types of tags**, like those related to geographical features or locations in pictures. For instance, we can use the geographical coordinates of pictures to try to disambiguate tags related to geographical locations. And in cases where DBpedia does not cover the specific domain tags we are dealing with, it will be interesting to look for specific domain ontologies to carry out the semantic association.

Finally, evaluation of the results of these approaches for the semantic association of tags is still a challenging task. Even for an external observer it is difficult to assert if the association of tags to semantic entities is right. We have noticed the **lack of testbeds and standard evaluation metrics** that allow proper comparisons of the different research works, and we would like to work on this aspect in cooperation with producers of similar approaches.

ACKNOWLEDGMENTS

This work is supported by the Spanish Projects GeoBuddies (TSI2007-65677C02) and CENIT España Virtual (ALT0317), and the FPI grant (BES-2008-007622) of the Spanish Ministry of Science and Innovation.

REFERENCES

[1] Angeletou, S., Sabou, M., Motta, E., Semantically Enriching Folksonomies with FLOR. In *1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008)*, Tenerife, Spain (2008).

[2] Cantador, I., Szomszor, M., Alani, H., Fernández, M., Castells, P., Enriching Ontological User Profiles with Tagging History for Multi-Domain Recommendations. In *1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008)*, Tenerife, Spain, (June 2008).

[3] Golder, S., Huberman, B., Usage patterns of collaborative tagging systems. In *Journal of Information Science*, Vol. 32(2), 198-208 (2006)

[4] Hamasaki, M., Matsuo, Y., Nisimura, T., Takeda, H., Ontology Extraction using Social Network. In *International Workshop on Semantic Web for Collaborative Knowledge Acquisition*, Hyderabad, India (2007)

[5] Jäschke, R., Hotho, A., Schmitz, C., Ganter, B., Stumme, G., Discovering shared conceptualizations in folksonomies. In *Journal of Web Semantics* 6(1) 38-53 (2008)

[6] Kim, H. L., Scerri, S., Breslin, J. G., Decker, S., Kim, H. G., The state of the art in tag ontologies: a semantic model for tagging and folksonomies. In *Proceedings of the 2008 international Conference on Dublin Core and Metadata Applications*, 128-137, Berlin, Germany (2008)

[7] Lee, S., Yong, H., Tagplus: A retrieval system using synonym tag in folksonomy. In *Proceedings of the 2007 International Conference on Multimedia and Ubiquitous Engineering*, 294-298, IEEE Computer Society, Washington DC (2007)

[8] Kim, HL., Yang, SK., Song, SJ., Breslin, JG., Kim, HG., Tag mediated society with scot ontology. In *Proc. of the 5th Semantic Web Challenge at ISWC*,(2007)

[9] Mika, P., Ontologies are us: A unified model of social networks and semantics. In *Journal of Web Semantics* 5(1), 5-15 (2007)

[10] Passant, A., Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs. In *Proceedings of the First International Conference on Weblogs and Social Media (ICWSM)*, Boulder, Colorado (2007)

[11] Specia, L., Motta, E., Integrating Folksonomies with the Semantic Web. In *Proceedings of the 4th European Conference on the Semantic Web: Research and Applications*, Innsbruck, Austria (2007)

[12] Szomszor, M., Alani, H., Cantador, I., O'Hara, K., Shadbolt, N., Semantic Modelling of User Interests based on Cross-Folksonomy Analysis. In *7th International Semantic Web Conference (ISWC)*, Karlsruhe, Germany (2008)

[13] Baeza-Yates, R. A., Ribeiro-Neto, B., *Modern Information Retrieval*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, (1999)

[14] Hotho A., Jaschke R., Schmitz, C., Stumme. G., Information retrieval in folksonomies: Search and ranking. In *The Semantic Web: Research and Applications*. Springer, Heidelberg (2006), 411-426

[15] Echarte, F., Astrain, J. J., Córdoba, A. and Villadangos, J., Ontology of Folksonomy: A New Modeling Method. In *Proceedings of Semantic Authoring, Annotation and Knowledge Markup (SAAKM)* (2007)

[16] Gruber, T.: Ontology of Folksonomy: A Mashup of Apples and Oranges.In *First on-Line conference on Metadata and Semantics Research (MTSR'05)* (2005).

[17] Newman, R.: Tag ontology, available at http://www.holygoat.co.uk/projects/tags/ (2005).

[18] Passant, A. and Laublet. P., Meaning of a Tag: A collaborative approach to bridge the gap between tagging and Linked Data. In *Proceedings of Linked Data on the Web (LDOW2008)* (2008)

[19] Lesk, M., "They said true things, but called them by wrong names" – vocabulary problems in retrieval systems. in *Proc. 4th Annual Conference of the University of Waterloo Centre for the New OED* (1998)

[20] Sanderson, M., Retrieving with Good Sense. In *Information Retrieval* 2(1): 47-67 (2000)