The Knowledge Engineering Review, Vol. 27:1, 57–85. © Cambridge University Press, 2012
doi:10.1017/S026988891100018X

Review of the state of the art: discovering and associating semantics to tags in folksonomies

ANDRÉS GARCÍA-SILVA[1], OSCAR CORCHO[1], HARITH ALANI[2] and ASUNCIÓN GÓMEZ-PÉREZ[1]

[1]*Ontology Engineering Group, Facultad de Informática, Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Spain;*
e-mail: hgarcia@fi.upm.es, ocorcho@fi.upm.es, asun@fi.upm.es;
[2]*Knowledge Media Institute, The Open University, Milton Keynes MK7 6AA, UK;*
e-mail: h.alani@open.ac.uk

Abstract

This paper describes and compares the most relevant approaches for associating tags with semantics in order to make explicit the meaning of those tags. We identify a common set of steps that are usually considered across all these approaches and frame our descriptions according to them, providing a unified view of how each approach tackles the different problems that appear during the semantic association process. Furthermore, we provide some recommendations on (a) how and when to use each of the approaches according to the characteristics of the data source, and (b) how to improve results by leveraging the strengths of the different approaches.

1 Introduction

In recent years we have witnessed the transition from a Web where the content is generated mainly by the owners of websites to a more open and social Web where users are not only information consumers but also producers (*prosumers*—Tapscott & Williams, 2006). This new age of the Web, also known as Web 2.0[1], has brought a diversity of new social applications like *wikis*, *blogs*, *social networks*, *social bookmarks*, and *photo*, *music* and *video sharing sites*. These applications made it possible for all Web users to contribute and share huge amounts of multimedia content, and to *tag* these content resources with free-form keywords.

Tags serve multiple purposes, such as content organisation, description, and searching. In 2003, Delicious[2] was released as a social bookmarking tool where users are able to assign tags to Uniform Resource Locators (*URLs*) in a collaborative manner. One year later, Flickr[3] was presented as a social network for photo sharing where users can assign tags to their own photos or to other photos from their colleagues. Nowadays, tagging is part of many popular applications such as *Amazon*, *YouTube* and *Last.Fm*, to name a few, where users can assign tags to products, videos and songs, respectively.

In 2004, Vander Wal[4] coined the term Folksonomy to describe the new structure of users, tags, and objects. Folksonomy is defined as *the result of personal free tagging of information and objects (anything with a URL) for one's own retrieval. The tagging is done in a social environment (usually*

[1] http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html
[2] http://delicious.com/
[3] http://www.flickr.com/
[4] http://www.vanderwal.net/folksonomy.html

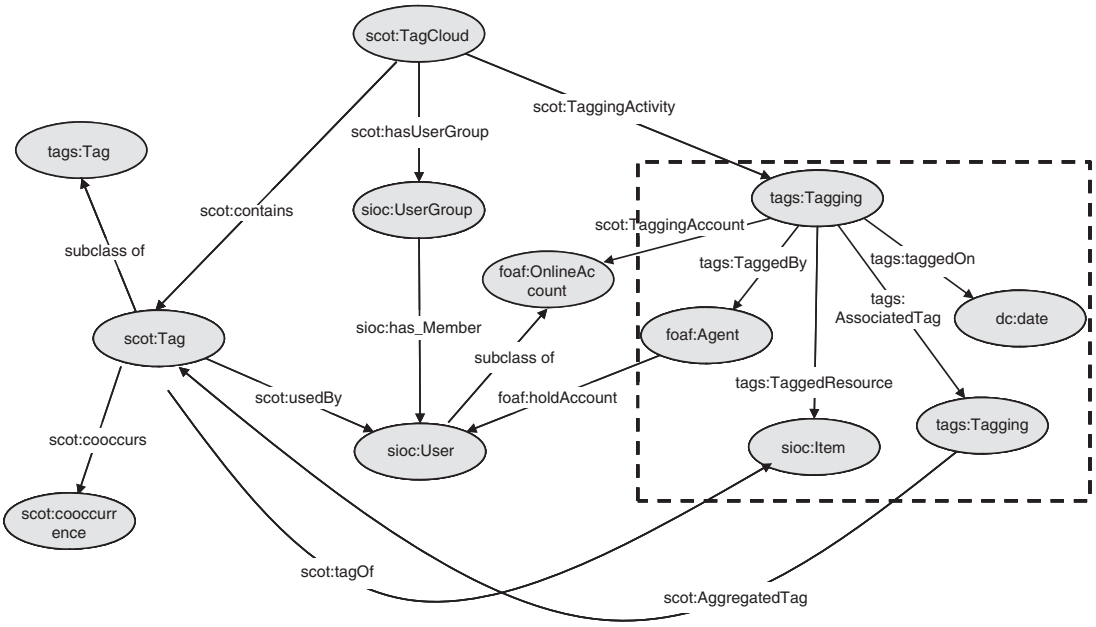

Figure 1 Graphic representation of the Social Semantic Cloud of Tags (SCOT) ontology (Kim *et al.*, 2008a)

shared and open to others). Folksonomies are good sources of terminology frequently updated by large communities of users. This contrasts with other classification schemes, such as thesauri or taxonomies, which are generally created and maintained by controlled user groups. Hence, one of the advantages of folksonomies is their ability to rapidly adapt to new changes in terminologies and domains. Furthermore, as time goes by, users tend to stabilize the vocabulary used to tag a resource (Golder & Huberman, 2006). This stabilization is the result of several user iterations and of a tag recommendation strategy based on previously assigned tags. These vocabularies can be seen as shared conceptualizations by groups of users with respect to groups of resources.

The success of tagging is attributed to two main factors: (a) they are very easy to create, where users do not need any special skills or experience to tag, and (b) the benefits of tagging are immediate (Hotho *et al.*, 2006). However, current tagging technology suffers from two main problems. First problem is that *folksonomies lack a uniform representation to facilitate their sharing and reuse*. Some Web 2.0 applications provide Application Programming Interfaces (APIs) to export their folksonomies. However, they do it in proprietary formats. To overcome this problem, ontologies have been proposed to model the tagging activities in folksonomies, with semantic concepts to represent users, tags, resources, etc. (Gruber, 2005; Newman, 2005; Knerr, 2006; Echarte *et al.*, 2007; Scerri *et al.*, 2007; Kim *et al.*, 2008a; Passant & Laublet, 2008). One example of these ontologies is the Social Semantic Cloud of Tags (SCOT) ontology (Kim *et al.*, 2008a), which is depicted in Figure 1. This ontology models tagging information, and includes concepts such as *User*, *Item*, *Tag*, and *Tag Cloud* as well as the relationships among these concepts. SCOT reuses existing vocabularies such as FOAF (Friend of a Friend)[5] and SIOC (Semantically Interlinked Online Communities)[6], the former being a set of classes and properties describing people and their interests, and the latter a popular ontology for interlinking online communities (Breslin *et al.*, 2006). Some surveys in this respect have been published such as Kim *et al.* (2008b), where authors review most of the current ontologies for folksonomy information representation. Thus, please note that this issue is out of the scope of this survey.

[5] http://www.foaf-project.org/

[6] http://sioc-project.org/

The second and more relevant problem is the *lack of formal and explicit semantic of tags*. This problem has been widely reported in Golder and Huberman (2006), Angeletou *et al.* (2008), Lee and Yong (2007), and Szomszor *et al.* (2008). Users can use different morphological variations of a tag to represent the same label such as plurals, acronyms, conjugated verbs, or misspelling words (e.g. different users can annotate a picture of a celebration with tags such as *party*, *parties*, *partying*, *partyign*). Furthermore, a user can use a tag to annotate a resource while another user can use a synonym of that tag to annotate another resource (e.g. synonyms as *party* and *celebration*). Moreover, some tags can be polysemous, where the same word has more than one meaning, such as *party* as a celebration as opposed to *party* as a political organization. Most current tagging applications do not allow users to define the intended meaning for their tags. For example, although Flickr tackles the ambiguity problem by providing clusters of related tags, the meaning of these tags, the meaning of their relationships, and the meaning of the cluster itself are not defined. Finally, different levels of granularity may be found in tags provided by users: some users issue more generic tags while others issue more specific tags. Such difference in granularity could be related to the level of user interest, or depth of expertise in the subject (e.g. a general tag as *party* in contrast to a specific tag as *banquet*).

Thus, when a user looks for resources tagged with a particular tag, current systems ignore the resources tagged with morphological variations or synonyms of that tag, as well as the resources tagged with more generic or more specific tags. If a user uses a polysemous tag when searching, all the resources tagged with that tag are retrieved without taking into account the tag sense the user was looking for. For example, the noun *bank* has at least 10 senses[7], hence if we query Flickr using *bank*, we get photos about financial institutions, fog banks, and sand banks among others.

Ontologies have been proposed as a solution to the above problems. Tags, their morphological variations and the synonym relation among tags can be modeled using ontological primitives such as classes, data properties, and object properties. Furthermore, a polysemous tag can be identified according to the context in which the tag appears in the ontology, where the context is defined as the set of concepts related to the tag. Systems can recognize ambiguous tags from the number of different ontological concepts associated with the tag to represent its possible meaning. Thus, the system could ask users to specify the sense to be applied to the given tag, or try to disambiguate the tag automatically to select the correct meaning. In addition, with ontologies, relations such as *synonymy*, *sibling*, *subclass of*, *type of*, etc., can be represented and used in query expansion processes (Qiu & Frei, 1993) where other highly relevant tags can be added to the original query to widen the search when necessary.

In this paper we describe the most relevant approaches described in the literature whose main objective is either to extract ontologies from tags in folksonomies or to associate tags to external semantic entities in order to make explicit the meaning of those tags. Ontology learning approaches (Maedche & Staab, 2001; Buitelaar *et al.*, 2005) are knowledge acquisition processes aiming at the creation of ontologies from unstructured (e.g. text files or Web pages) or structured data (e.g. Extensible Markup Language files or relational databases). In this context, we can refer to folksonomies as a new data source for ontology learning that can be analyzed using techniques already used in this area such as clustering, natural language processing, and formal concept analysis.

This survey is structured as follows. First, in Section 2 we propose a unified process and categories to describe each of the approaches in a uniform way. Then we describe a simple folksonomy, which we use to exemplify the results obtained from each one of the analyzed approaches. In Section 4 we describe each of the approaches in detail, according to our unified process, and then we present a summary of the review. After having reviewed the approaches, we will discuss in Section 5 how to improve these approaches and how their results can be used in Web applications. Finally, we present our conclusions.

[7] As described in WordNet (`http://wordnet.princeton.edu/`)

2 A unified process for the association of semantics to tags

In this section we propose a unified process that can be used to understand, evaluate, and categorize the different approaches for the association of semantics to tags. This process consists of a set of common activities identified in most of the analyzed proposals. The objective is to provide a uniform way to describe the different proposals and therefore to facilitate their assessment.

We designed this unified process following a bottom-up approach. First, we analyzed each process individually, identifying the activities carried out and main objectives. Then, we highlighted the commonalities between the activities among different processes and the result was a set of activities that most of the analyzed processes carry out with different degrees of detail. The activities are presented in a logical sequence, which does not necessarily hold for all the processes. The general process is depicted in Figure 2.

Most of the processes start with defining their data sources, and some of them explicitly describe how they gather information from these data sources. For instance, some research teams designed and developed specialized programs to crawl folksonomies when APIs are not available, or available but with limited coverage and capabilities. For our analysis, the details of how to get the data are not so important. What is relevant are the filters they implement to select and clean the data if they exist. Thus, the first activity identified in our unified process is called *data selection and cleaning*. This activity may include filters that take into account tag use frequency, lexical characteristics like tag length or allowed characters, morphological characteristics, or even the language of the tags. With regard to tag frequency, this can be measured based on the number of times the tag has been used to annotate a resource, or the number of times the tag was used by different users.

Once we have the data set we want to work with, the next activity is *context identification* of the tagging activity. By context we mean the set of things we take into account to figure out the tag meaning. This will help to identify groups of related tags or to associate a semantic concept to formally define the tag meaning. Context in linguistics is defined as the surrounding words that appear near the word in question in a sentence. This notion of context can be applied to the tags of a folksonomy in two ways, the first one uses tags that occur together when they are used to annotate the same resource or group of resources, the second one uses tags that are used together by the same user or group of users. Furthermore, the notion of context can include linguistic information, such as synonyms or hyperonyms, or other known morphological variations of the tags. On the other hand, context can also include tag or resource metadata information such as location coordinates and timestamps (Kennedy *et al.*, 2007).

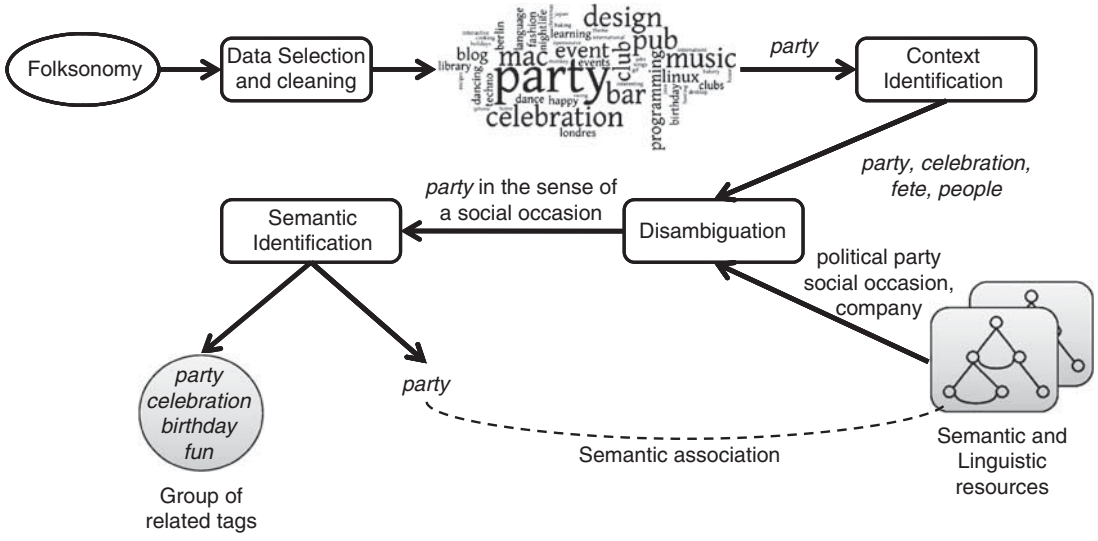

Figure 2 A unified process for the association of semantic to tags

As mentioned earlier, one of the main problems with folksonomies is that most tags have more than one meaning. This ambiguity yields inaccurate and irrelevant results when these tags are used to search and retrieve information. Therefore, a disambiguation activity is an important step in the semantic association process. The *disambiguation* activity can be carried out using external semantic resources, such as WordNet, and all the tag context information. Some tag disambiguation approaches, such as those presented by Au Yeung *et al.* (2007) and Specia and Motta (2007), use clustering techniques in order to group tags according to the resources they annotate or to the users who authored the tags. In either case, according to these methods, if a tag is used to annotate different resource groups or if a tag is used by different user groups, then the tag is considered to have more than one meaning. In general, this type of analysis is more focused on identifying the existence of ambiguity, rather than on identifying the true meaning of a tag.

Finally, the last activity is *semantic identification* in which tag semantic is made formal and explicit. This activity consists of matching between tags and semantic entities, or identifying relations between tags or semantic entities. The matching between tags and semantic entities like classes or instances is carried out using predefined ontologies or ontologies retrieved at runtime by means of Semantic Web search engines. This matching may result in several semantic entities for a tag (Angeletou *et al.*, 2008), hence aggregation of these entities is required in order to identify which of them refer to the same topic and which does not. In the case that the semantic entities refer to more than one topic, a disambiguation task might be carried out. Furthermore, this activity could use clustering techniques to identify groups of synonyms, or social network measures, such as clustering coefficient and local centrality, to identify groups of narrower terms and broader terms similar to the relations found in a thesaurus (Mika, 2007).

In addition to defining this unified process, we will also categorize existing approaches according to the main technique they use. One method for distinguishing between these approaches is proposed by Angeletou *et al.* (2008), which is based on whether these approaches use statistical clustering techniques to implicitly describe their meaning, or ontology-based techniques to align tags with existing semantic resources. Besides these two categories, we introduce a hybrid category for those approaches mixing clustering and ontology-based techniques. This categorization is useful since most of the proposals based on statistical techniques do not state explicitly the meaning of the tags or the relationships between them, while the ontology-based proposals usually do. On the other hand, hybrid approaches exploit the benefits of statistical and ontology-based techniques to associate semantics to tags or to find groups of related tags.

3 An illustrative example

In this section we present a simple folksonomy that we will use to illustrate how each of the approaches presented in the following section works. Let us assume that we have the folksonomy shown in Figure 3, which consists of four users. User A has no explicit relation with any other user, while user B is explicitly related to users C and D. These relationships are symmetric, as is usually the case in most social networking sites. Although not all tagging systems allow for social relations to be established among users, we will include those relations in the example folksonomy because of the fact that nowadays more and more tagging systems, such as Delicious and Flickr, are supporting social relations.

There are five tags in the example folksonomy: *Coffee*, *Java*, *Language*, *Program*, and *Code*, which are assumed to have been used by our users to tag three resources R1, R2, and R3.

The tagging carried out by these users is presented in Table 1. User A tagged R1 with *Coffee* and *Java*. User B tagged R1 with *Java*, and R2 with *Java* and *Language*. User C tagged R2 with *Language* and *Program*. Finally, user D tagged R2 with *Language* and *Program*, and also tagged R3 with *Program* and *Code*.

In the following section we will describe the results that can be obtained for this simple example from each of the approaches presented next, so that the similarities and differences between them can be better understood.

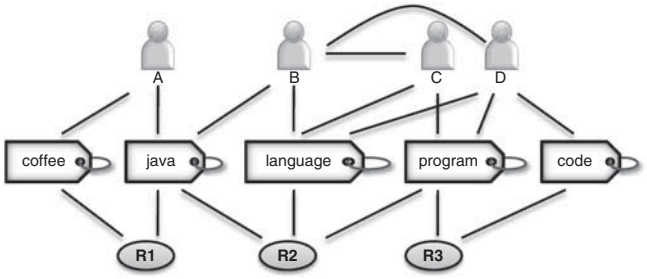

Figure 3 Graphical representation of the folksonomy example

Table 1 Tagging details of the Folksonomy example

Resources	R1		R2			R3	
Users/tags	*Coffee*	*Java*	*Java*	*Language*	*Program*	*Program*	*Code*
A	X	X					
B		X	X	X			
C				X	X		
D				X	X	X	X

4 Review of approaches about discovering and associating semantics to tags

In this section, we review the most relevant approaches that aim to enrich folksonomies with semantics. We identify three groups of approaches according to if they are based on (1) *clustering techniques*, (2) *ontologies*, or (3) on a *hybrid approach* mixing clustering techniques and ontologies. In general, clustering-based approaches goal is to group related tags in the hope that such grouping will indirectly expose a meaning for their tags. Hence, these approaches do not formally define the meaning of tags or their relations. Ontology-based approaches aim at stating the meaning of the tags and their relations by means of associating semantic entities to tags. Hybrid approaches objective can be either (1) to group tags using semantic information, or (2) to associate semantic entities to tags using as context groups of tags.

4.1 Clustering-based approaches

Several approaches exist, whose goal is to identify the semantics of tags, that propose to cluster tags according to some relations among them (Begelman *et al.*, 2006; Hamasaki *et al.*, 2007; Kennedy *et al.*, 2007; Mika, 2007; Jäschke *et al.*, 2008). Tag co-occurrence is a well-known measure of tag relatedness that can be measured when two tags are used to annotate the same resource regardless of the annotator, or when two tags are used by the same user regardless of the resource. In this section, we present research works that exploit these tag relatedness measures to find groups of tags as the those depicted in Figure 4.

4.1.1 Mika's approach

Mika describes an approach to generate two lightweight ontologies from folksonomies using statistical techniques (Mika, 2007): an ontology of concepts based on the overlapping set of user communities (O_{ac}), and an ontology of concepts based on the overlapping set of resources (O_{ci}). The approach is tested with two data sources; a set of users, terms, and Web pages from the Semantic Web research community, and a folksonomy extracted from the Delicious website. We use the latter case to describe this approach following our uniform process detailed in Section 2.

Data selection and cleaning: The data selection and cleaning activity of this approach is limited to filtering out from the folksonomy those tags with less than 10 items classified under them and

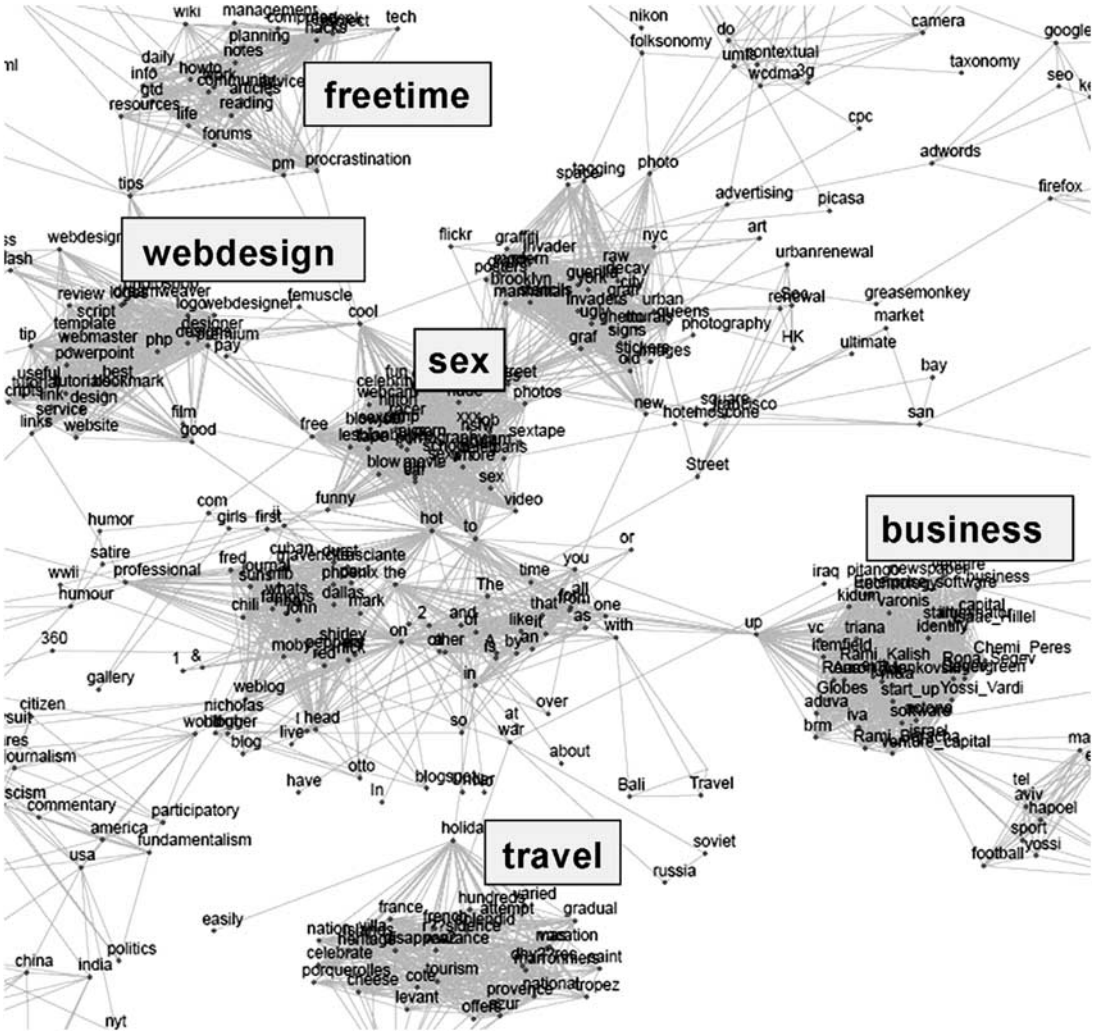

Figure 4 Tag clustering based on tag co-occurrence when annotating resources (Mika, 2007)

those persons who have used less than five tags. The Delicious data set used to test the approach consists of 51 852 anotations of 30 790 resources by 10 198 users using 29 476 distinct tags.

Context identification: The context identification activity is carried out in a distinct way for each of the resulting ontologies. In the case of the ontology O_{ci}, context is defined as the tags that co-occur with a particular tag when they are used to annotate a resource. In the case of the ontology O_{ac}, context is defined as the tags that co-occur with a particular tag when they are used by a user or group of users.

Disambiguation: This approach does not explicitly deal with disambiguation problems.

Semantic identification: Finally, the semantic identification activity is also carried out differently for each of the ontologies that this approach generates. A graph of concepts is built for O_{ci} where the edges indicate that two concepts (tags) were used together when annotating one or more resources. These links are weighted by the number of resources annotated using both tags. Similarly, a graph of concepts is also built for O_{ac} where the edges indicate that two concepts were used together by one or more users. These links are weighted by the number of people who have used both tags. After this graph generation process, clustering is performed to identify specific words inside each cluster and general words bridging different clusters. Set theory principles are then applied to define relations between concepts as broader and narrower terms. The output is a hierarchy of concepts O_{ac} based on subcommunity relationships and a classification hierarchy O_{ci} based on resource overlap.

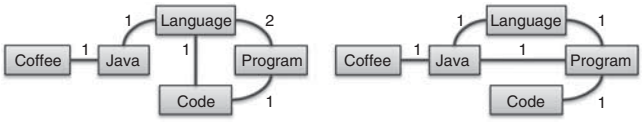

Figure 5 Tags' relations found by Mika's approach; (a) left side: O_{ac} tags are related if used by same user and (b) right side: O_{ci} tags are related if used to tag the same resource

Using our folksonomy example from Section 3, the ontologies obtained with this approach are presented in Figure 5. In O_{ac} the *Language* and *Code* tags are related because user D used them to tag R2 and R3, respectively, while in O_{ci} these tags are not related because they were not used to tag the same resource. On the other hand in O_{ci} tags *Java* and *Program* are related because these tags were used to tag the same resource R2, while in O_{ac} these tags are not related because they were not used together by any user. Then, if we want to discover in more detail the relations between these tags, we have to analyze the tag set using set operations. For instance, let us assume that we are interested in discovering the relation between *Java* and *Language* in the case of O_{ci}. If all the resources annotated with the *Java* tag are included in the set of resources annotated with the *Language* tag, and this last set is large enough, then according to this approach *Language* is broader than *Java*, meaning that all resources classified under *Java* are also classified under *Language*.

This approach allows reflecting the ongoing behavior of folksonomies for a set of user communities. However, there are some limitations to this approach. Tag ambiguity is one of the main problems present in folksonomies, as described in the introduction, and it is not clear in this approach how ambiguous tags can affect or be reflected in the generated ontologies. Furthermore, the identified relations mostly represent co-ocurrence, rather than any ontological relation such as *subclass of* or *part of*. Finally, there is no explicit catering for misspelled tags, or for tags in morphological variations.

4.1.2 Hamasaki et al.*'s approach*

Hamasaki *et al.* (2007) extended Mika's work with the notion of user neighborhood, which can be described as the direct contacts of a user. Particularly, the O_{ac} ontology is modified by taking into account tagging information of the user neighbors in the folksonomy. The data source used by Hamasaki *et al.* to evaluate this approach is a folksonomy adapted from a community support system for academic conferences, where users can bookmark documents of interest.

Data selection and cleaning: No general rules for data selection and filtering are described for this approach. The data set used to test the approach comprises 314 tags, 297 resources, and 75 users who have bookmarked at least one resource, and a total of 323 users in the social network.

Context identification: Context is shaped from the user tags along with the tags used by his neighbors. Tagging information of neighbors could help to overcome any lack of tagging information for a particular user.

Disambiguation: Unlike Mika's approach, this approach proposes an algorithm for disambiguation. The algorithm is based on the idea that if a tag is used to annotate different resources by different groups of users (neighbors), the tag may have different meanings. Otherwise, the tag has only one (or very similar) 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.

Semantic identification: There is no description in this approach of what concerns the semantic identification activity. As pointed out above, each pre-concept previously identified and possibly merged with others is converted into a concept. However, the relations between those concepts are not explicitly defined.

Figure 6 shows the O'_{ac} ontology that is obtained when applying this approach to our sample folksonomy. The O_{ci} ontology remains the same as the one described in Mika's approach. If we compare this O'_{ac} ontology with the O_{ac} ontology described in Mika's approach, we can see that

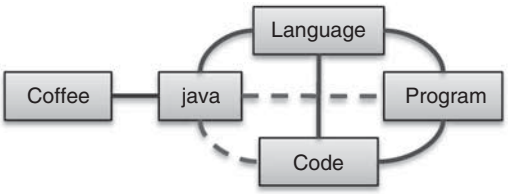

Figure 6 O_{ac} tags are related if used by the user or its social network

two new relations appear. First, the *Java* and *Program* tags are related even though they are not used by the same user. In this case, user B is related to user C, and following this approach, the tags from user C are considered to be indirect tags for user B. Therefore, user B uses *Java* and *Program* tags, being the *Program* tag taken from user C. Second, the relation between *Java* and *Code* is established from the relation between users B and D.

This approach presents some advances over Mika's approach. However, the data source used to evaluate the approach, as described in Hamasaki *et al*. (2007), has some characteristics that are not present in real folksonomies. In this data source owners of resources assign tags to them. However, if a user, other than the resource owner, bookmarks these resources, then the tags assigned previously by the owner are considered to have also been assigned by this user. Thus, all users who bookmark a document share, in a mandatory way, the same vocabulary. In contrast, users in major existing folksonomies assign tags freely, and the convergence of a vocabulary for a resource occurs when several users tag the same resource using a tag recommendation strategy that includes the most popular tags used by other users to tag the same resource. We want to note that besides recommendations based on tag popularity, other tag recommendation strategies have been developed based on collaborative filtering (Jäschke *et al*., 2007), and association rules (Schmitz *et al*., 2006), among others.

The proposed disambiguation strategy is strongly influenced by the group of users selected and how they tag. However, as explained above, the data source used by the authors of this approach is biased because all users who bookmark a document are assumed to share the same tags, and hence it is unclear whether this strategy would be adequate for open environments. Finally, the approach does not propose any method to semantically define the meaning of tags and their relationships.

4.1.3 Jäschke et al.*'s approach*

Jäschke *et al*. (2008) proposed a statistical approach for discovering subsets of users who implicitly agree on common tags for a set of resources in folksonomies. The approach is tested with three data sources; Delicious and BibSonomy[8] folksonomies, and a non-folksonomy application called *IT baseline security manual*.

Data selection and cleaning: The data selected for the experiments is a snapshot of the chosen folksonomies. For instance, in the case of Delicious, authors used as their data set all tagging information entered into the system before June 16, 2004. This data set contained over 3.3K users, around 30.5K unique tags, over 220K resources, with close to 617K links. With respect to Bibsonomy, the snapshot included all data up to November 23, 2006, excluding any automatic insertions (such as DBLP (Digital Bibliography and Library Project) publications) as well as any automated default tags (such as *imported*). The Bibsonomy testbed contained almost 45K tag assignments, 262 users, and over 11K resources (publications) tagged with close to 6K distinct tags. As for the IT security manual data set, this was set up by a closed group of experts and not by an open folksonomy, and used by the authors as an ontology for analyzing their approach, rather than for discovering any tag semantics.

[8] BibSonomy is a social bookmarking and publication sharing system. `http://www.bibsonomy.org/`

Context identification: Context identification in this approach consisted of mining all frequent tri-concepts over the selected information in order to obtain a set of triples, where each triple contains a set of users, a set of tags, and a set of resources. Each user in the set of users has tagged each resource in the set of resources with all the tags in the set of tags.
Disambiguation: Disambiguation of the tags used in these triples is not addressed in this approach.
Semantic identification: The semantic identification of tags is carried out by selecting those tag sets that we are interested in from the triples found in the previous activities. Then for each tag set a concept lattice is created. A concept lattice is a hierarchical conceptual clustering of tags. The formal context of the concept lattice is composed of resources tagged with at least one of the tags in the tag set by a particular number of users, and of tags which were used by the majority of users in a resource. The graphic representation of this concept lattice could then be used by ontology engineers to manually build a concept hierarchy that corresponds to the original folksonomy.

With respect to our running example folksonomy, let us extend the tagging of users C and D with the *Java* tag to annotate R2. Table 2 shows the shared conceptualizations that can be found in this folksonomy. Users B, C, and D agree on the use of *Java* and *Language* as tags for R2. Users C and D agree on the use of *Java*, *Language*, and *Program* as tags for R2. Users A and B agree on the use of *Java* to tag R1. According to the authors the sets of users, tags, and resources of each shared conceptualization have the property that none of them can be extended without shrinking one of the other two sets. That is, if we want to expand the tag set in which users B, C, and D agree to annotate R2 with the tag *Program*, users B will be eliminated from the user set.

As an example, from these triples we can create a concept lattice to analyze the tags *Java*, *Language*, *Program*. Let us suppose that the tag *Java* has been used only together with the *Language* tag, while the *Language* tag has been used independently in other cases. Thus, in the graphical representation of the concept lattice the *Language* tag will be above the *Java* tag, meaning that all the resources tagged with *Java* are also tagged with *Language*, and that *Language* has been used in a more general sense to annotate another group of resources. This hierarchical relation can be then analyzed by an ontology engineer to find the appropriate semantic relation between *Java* and *Language* tags.

Triples and the corresponding concept lattices that are generated by applying this approach could provide useful information for ontology construction. However, on the one hand, in this approach two different triples can be generated including the same tag set, but for different sets of users and resources. This will be the case if some people use the tag set to annotate a resource set, while some other people use the same tag set to annotate another resource set. Nevertheless, the resource sets could be related, for instance if they are about the same topic, and thus the shared conceptualization could be extended to cover all users and resources. On the other hand, the output of the process is a hierarchical representation of tags, but the relationships between tags in different hierarchical levels are not defined semantically, and this task is left to an ontology engineer. Finally, in this approach there is no strategy to deal with ambiguous tags.

4.1.4 Other clustering-based approaches

Other clustering algorithms have been proposed for identifying groups of related tags. Here, we describe briefly two aproaches: (1) Begelman *et al.* (2006) cluster tags according to tag co-occurrence

Table 2 Groups of users, tags, and resources

	Shared conceptualizations		
Users	B, C, D	C, D	A, B
Tags	*Java, Language*	*Java, Language, Program*	*Java*
Resources	R2	R2	R1

when annotating resources and (2) Kennedy *et al.* (2007) cluster tags based on time and location metadata.

Begelman *et al.*'s approach proposes to create a graph where the vertices are tags, and edges between two tags exist if they co-occur in the annotations of one or more resources. These edges are weighted by their co-ocurrence frequency. A technique called Spectral bisection is used to split this graph in two clusters, and a modularity function is used to compare the quality of the new clusters against the previous one to evaluate whether the new clusters are accepted or not. This technique is executed recursively on the new clusters.

On the other hand, Kennedy *et al.*'s approach presents a clustering-based technique to identify tags related to locations and events. The approach relies on the latitude and longitude of the geotagged resources as well as on the timestamp. Each tag has an associated spatial and temporal distribution. The spatial distribution is a list of the geographical coordinates of the pictures annotated with that tag, and the temporal distribution is a list of the timestamps where the picture was taken or when it was uploaded to the system. To identify location and event tags, a clustering algorithm was applied to the spatial distribution to find groups of tags sharing spacial patterns, and on the temporal distribution to find tags sharing temporal patterns.

Begelman *et al.*'s approach finds groups of tags sharing co-occurring patterns over resources. However, the meaning of the tags in those groups as well as the relations among the tags in each group or in different groups are not stated. Kennedy *et al.*'s approach is a step forward to the identification of tag semantics using clustering techniques based on the information in tag metadata. Nevertheless, this approach does not represent the meaning of the grouped tags by means of a formal language such as Research Description Framework (RDF). Furthermore, relations between the tags belonging to the same or different groups are not established.

4.2 Ontology-based approaches

The approaches presented earlier mainly focus on applying statistical techniques to group tags or to show relatedness between them. In addition to these approaches, there exists a number of approaches aiming at associating semantic entities to tags as a way to formally define their meaning (Maala *et al.*, 2007; Passant, 2007; Angeletou *et al.*, 2008; Cantador *et al.*, 2008; Tesconi *et al.*, 2008; García-Silva *et al.*, 2009). In this section, we review these works.

4.2.1 Angeletou et al.*'s approach*

Angeletou *et al.* (2008) proposed an automatic approach to enrich folksonomy tags with formal semantics by associating them with relevant concepts defined in online ontologies. The data source used to test the approach is a Flickr data set.

Data selection and cleaning: The initial Flickr data set comprised 250 resources and 2819 tags. As in the previous approach, during the data selection and cleaning activity, some tags are filtered out, including numbers, special characters, and non-English tags. The main reason to eliminate these tags is that this approach relies on WordNet, a lexical database for the English language, and usually its entries do not include numbers or special characters.

Context identification: The filtered tags then go through context identification. For each tag, all of its possible lexical representations, such as singular, plurals, or various delimited types of compound tags, are generated. The context is defined as the whole filtered tag set along with their lexical representations.

Disambiguation: Tags and their contexts are taken as input to the disambiguation activity. In this activity, if a tag has more than one sense in WordNet, then the hierarchy of its senses as extracted from WordNet is used to calculate the similarity with the senses of all tags in the tag set and thus disambiguating them. The similarity between senses is calculated using the Wu and Palmer (1994) similarity measure.

Unlike other approaches described so far, in this approach there is a phase of context expansion after disambiguation. For each tag, synonyms and hypernyms are extracted from WordNet using

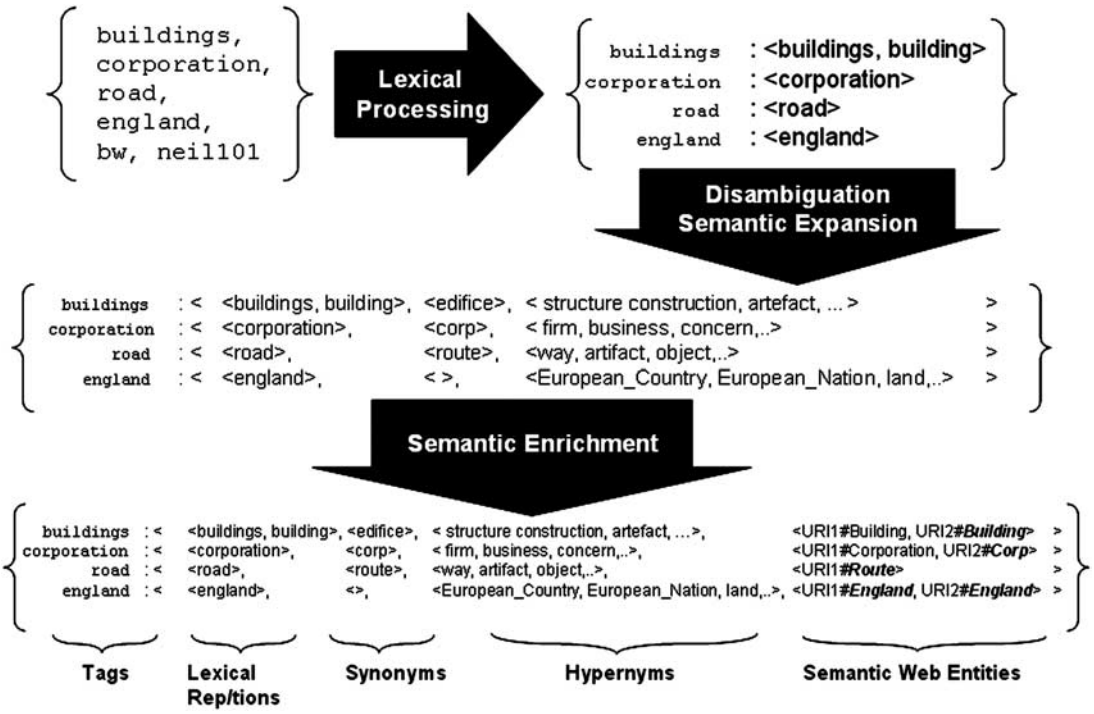

Figure 7 Semantic enrichment process (Angeletou *et al.*, 2008)

Table 3 Tags and related semantic entities

Tags	Semantic entity
Java	`http://dbpedia.org/resource/Java` *typeOf* `http://dbpedia.org/resource/Java_(programming_language)`
Code	`http://www.lt4el.eu/CSnCS#ComputerCode` *subClassOf* `http://www.loa-cnr.it/ontologies/IOLite.owl#DigitalResource`
Program	`http://www.lt4el.eu/CSnCS#Program` *subClassOf* `http://www.lt4el.eu/CSnCS#ProgrammingSoftware`

the sense assigned in the disambiguation phase. This information is used later to find the right ontology entity that will be associated with each tag.

Semantic identification: The semantic identification activity in this approach focuses on relating the expanded set of tags to ontological entities, using the Watson[9] semantic search engine. For each tag, several ontological entities may be retrieved, which are then integrated in order to group similar ontological entities. The similarity measure used for this integration process compares the entity labels and the semantic neighborhood information including superclasses and subclasses. Finally, tags are associated with one or more ontological entities, comparing the ontological parents of the merged entities with the tag hypernyms.

The whole process proposed is depicted in Figure 7. Table 3 presents the result of applying this approach to some of the tags in our sample folksonomy. In this case, the most probable sense of the *Java* tag is Java as a programming language because most of the tags are related with this sense. Thus, the *Java* tag is associated with an instance of the class *Java_(programming_language)*. The *Code* tag is

[9] `http://watson.kmi.open.ac.uk/`

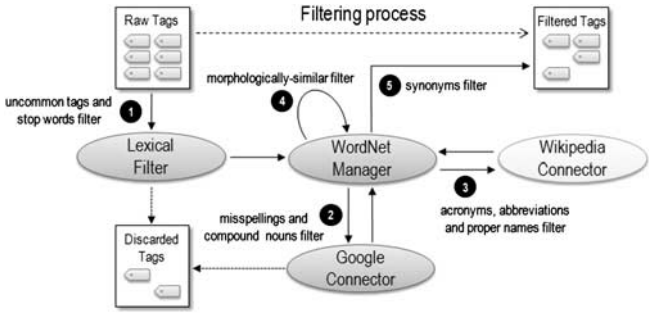

Figure 8 Tag filtering process (Cantador *et al.*, 2008)

associated with the *ComputerCode* class which is a *subclass of* the *DigitalResource* class. The *Program* tag is associated with the class *Program* which is a *subclass of* the *ProgrammingSoftware* class.

Angeletou *et al.* (2008) evaluated their approach, achieving a high precision rate of 93%, with an initial low recall rate of 24.5%, which then was raised to 49% by estimating how many of the tags actually could be related with ontological entities using the information provided by Watson. This evaluation was carried out using 226 photos whose tags were enriched semantically using this approach, and the associations between tags and ontological entities were manually checked.

These results are promising, since they show that these types of automatic techniques can achieve good results. However, there are also some limitations. For instance, this approach tries to find one sense for a tag. Although this could be valid for a particular tag set, there is the possibility that an ambiguous tag is used in more than one sense. The authors also mention that some tags and their context were not found in the WordNet hierarchy of senses, and thus the disambiguation activity failed. WordNet is limited in terms of its terminological coverage, and hence seriously limits the scope of this approach. Finally, the semantic identification activity tries to match hypernyms of WordNet with ontological parents. WordNet is a lexical database and its hierarchy of hypernyms does not necessarily correspond to the hierarchy of concepts found in ontologies.

4.2.2 Cantador et al.*'s approach*

Cantador *et al.* (2008) proposed an automatic approach to associate folksonomy tags with domain ontology concepts using Wikipedia[10] categories as an intermediate shared representation between tags and ontology classes. The data source used to test the approach is a tag set extracted from Delicious and Flickr.

Data selection and cleaning: The initial tag set contains 28 550 tags, gathered from Delicious and Flickr. The data selection and cleaning activity, depicted in Figure 8, focuses on filtering out tags which are too small (one letter tags) or too large (tags with more than 25 characters). Moreover, special characters are converted to their base form (e.g. ü is converted to u), and tags with a low frequency or that are common stop-words are removed. Then, each tag is searched in WordNet. If a tag does not exist in WordNet, then the Google *did you mean* mechanism is used to correct any possible tag misspellings, or to break up any compound tags (tags made up of concatenated words). Otherwise, the tags are assumed to be acronyms, abbreviations, or proper names. In the latter case, those tags are searched in Wikipedia for an agreed representation. Furthermore, morphologically similar tags are grouped in a single tag using a singularization algorithm and a stemming function, and the shortest term in WordNet is used as the representative tag. Finally, tags which are non-ambiguous synonyms are merged. The synonym information of each tag is retrieved from WordNet.

Context identification: The context identification activity retrieves information from Wikipedia for each tag, including the Wikipedia page URL, and the Wikipedia category list for that page.

[10] `http://en.wikipedia.org/`

Table 4 Semantic association between tags and ontology concepts

Tags	Semantic entity
Coffee	`http://en.wikipedia.org/wiki/Coffee`
	type of `http://mydomain.org/userPreferencesOntology/Coffee`
Java	`http://en.wikipedia.org/wiki/Java`
	type of `http://mydomain.org/ProgrammingLangOntology/Java`
Language	`http://en.wikipedia.org/wiki/Language`
	type of `http://mydomain.org/ProgrammingLangOntology/FormalLanguage`

Disambiguation: The disambiguation activity is not specified in this approach, although Wikipedia disambiguation pages are pointed out as a possible source of information to disambiguate tags (see the next approach).

Semantic identification: During semantic identification each tag is given a concept Uniform Resource Identifier (URI) using the tags context information, which includes the Wikipedia page name and the Wikipedia category previously associated with each tag. To this end, the terms in the context, that is the terms in the category names, are compared against the domain ontology classes, and the most appropriate ontology classes among the matching ones are selected. Finally, an instance of each of the ontology classes is created. The URI of those instances is the Wikipedia page name, and the categories are assigned as instance labels.

In our sample folksonomy, all the tags are ambiguous according to WordNet and Wikipedia. However, Wikipedia only displays a disambiguation page for the *Program* tag, while for the other tags the most probable page is displayed and in this page there is a link to the disambiguation page. Therefore, the *Program* tag is not processed for this approach because of its ambiguity. The *Coffee* tag is related to the Coffee Wikipedia page, which refers to the coffee beverage. This page has the categories: Coffee, Arabic culture, Arabic loanwords, and Crops. In this case, the Coffee Wikipedia category will be selected as it matches exactly the *Coffee* tag. Therefore an instance of this class will be created using the Wikipedia page name as its URI (as shown in Table 4). However, for the other tags the Wikipedia pages displayed are not in the context of programming languages. For instance, for the *Java* tag the Wikipedia page refers to Java as the island in Indonesia.

According to the authors of this approach, the advantage of using Wikipedia as a shared representation for tags is that Wikipedia is maintained collaboratively by a large user community. Thus, Wikipedia incorporates new terminology faster than linguistic resources like WordNet. However, this approach fails when the Wikipedia page is not directly related with the intended meaning of the tag according to its context, mainly because the approach lacks a disambiguation process.

4.2.3 *García-Silva* et al.*'s approach*

García-Silva *et al.* (2009) proposed an approach to link tags to DBpedia[11] resources by means of the selection of the Wikipedia page that best represents the tag intended meaning according to the context where the tag was used. This approach was exemplified using some Flickr pictures.

Data selection and cleaning: The data selection and cleaning activity consists of the same steps presented in Cantador *et al.* (2008) since these works share the tagging database used to test the approaches. However, authors do not provide statistics regarding the data set used to test the approach.

Context identification: This approach tries to disambiguate the meaning of each tag in each user post. Consequently, context has been defined as the co-ocurring tags in the user post. However, authors propose that in cases where co-occurring tags in a user post might not give enough information to disambiguate the meaning of a tag, the context could be improved using more folksonomy information such as (1) co-occurring tags in the resource regardless of the user,

[11] DBpedia is an RDF representation of part of Wikipedia. `http://dbpedia.org`

Table 5 Semantic association between tags and DBpedia resources

User	Tags	Semantic entity	Resource
A	*Coffee*	`dbpedia/resource/Coffee` *rdf:type* `dbpedia/ontology/Beverage`	R1
B	*Java*	`dbpedia/resource/Java_coffee` *skos:subject* `dbpedia/resource/Category:Coffee`	R1

B	*Java*	`dbpedia/resource/Java_(prog_lang)`
		skos:subject
		`dbpedia/resource/Category:Object-oriented_prog_lang`
			R2

C	*Language*	`dbpedia/resources/prog_lang`
		skos:subject
		`dbpedia/resource/Category:Computer_languages`
			R2

(2) user tags regardless of the resource, (3) co-occurring tags in the user social network, and (4) co-occurring tags in the whole folksonomy.

Disambiguation: The disambiguation activity starts by retrieving a set of candidate Wikipedia pages related to the ambiguous tag using the Tagora Sense Repository[12]. In addition, from each Wikipedia page the most frequent terms are retrieved. Thus, the tag and its context can be compared against each one of the candidate Wikipedia pages measuring the overlapping of the terms in the context with the terms in each Wikipedia page. The tag, along with its context, and the Wikipedia pages are represented as vectors which then are compared by means of the cosine function. The most similar Wikipedia page vector to the tag and its context is selected as the most probable meaning for that tag.

Semantic identification: Semantic identification is carried out by selecting the corresponding DBpedia concept to the selected Wikipedia page in the previous phase. In DBpedia each concept has a *foaf:page* object property which links the DBpedia concept with the corresponding Wikipedia page. Using this property the DBpedia concept can be easily identified from the Wikipedia page URL.

The result of applying this approach to our folksonomy example is shown in Table 5. García-silva *et al.*'s approach is able to disambiguate the meaning of tags used in each post according to the context. Thus, each tag is assigned a DBpedia resource describing the user intended meaning when tagging a particular resource. In this example, we can see that the ambiguous tag *Java* has been used in two different contexts by the user B when annotating resources R1 and R2, respectively. We want to note that some DBpedia resources such as `http://dbpedia.org/resource/Coffee` are instances of an ontology. However, some other like `http://dbpedia.org/resource/Java_(programming_language)` are not related to an ontology, they are just subjects of RDF triples.

In this approach authors propose several context definitions as well as a tag disambiguation technique; yet, the approach has not been evaluated properly in terms of precision and recall. In addition, authors proposed to use as term weight the frequency of the term in the corresponding Wikipedia page. However, most of the current information retrieval approaches use more sophisticated term weights using measures such as the inverse document frequency, or the keyword density value (Baeza-Yates & Ribeiro-Neto, 1999). Authors might explore these term weights to evaluate if the current results improve.

4.2.4 Tesconi et al.*'s approach*

Tesconi *et al.*'s (2008) approach is based on mapping tags to Wikipedia pages and then associating those tags with other semantic resources. The approach has been tested using tagging information retrieved from Delicious.

[12] `http://tagora.ecs.soton.ac.uk`

Data selection and cleaning: Tagpedia[13] is used as a sense repository to find the set of candidate Wikipedia pages related to a particular tag. Tagpedia associates terms to Wikipedia pages by gathering information from Wikipedia disambiguation and redirection pages. Thus, morphological variations as well as synonyms are implicitly managed. The data set consists of the tagging information of nine Delicious user comprising 3520 tags used to annotate 3926 resources.
Context identification: The context of a tag consists of the user tags co-occurring with the tag when annotating any resource tagged by the user, plus the Delicious most popular tags for the set of resources annotated by the user.
Disambiguation: In this approach, the disambiguation activity calculates for each relevant Wikipedia page associated to an ambiguous tag a sense-rank value and selects the one with the highest value. The sense-rank value is calculated by taking into account co-occurrence or popularity frequency of each tag in the context. Co-occurrence is used when the tag in the context co-occurrs with the ambigous tag. On the other hand, popularity is used when the tag in the context was extracted from Delicious popular tags. In addition, the sense-rank value includes the number of occurrences of each one of the tags in the context in the analized Wikipedia pages.
Semantic identification: In the semantic identification activity the selected Wikipedia page is used to find the Wikipedia categories containing that page, and the corresponding DBpedia resource. From DBpedia resources authors extract references to Yet Another Great Ontology (YAGO) ontology[14] classes and WordNet Synsets.

The results of applying this approach to our folksonomy example are similar to the results presented previously in Table 5, extending them with the YAGO concepts and WordNet synsets. For each DBpedia resource in Table 5 we looked for the corresponding YAGO concept. We did not find any YAGO concept related to the DBpedia resources *Coffee*, *Java_coffee*, and *Programming_language*. Nevertheless, we found a YAGO relation for the DBpedia resource *Java_(programming_language)* stating that it has an *owl#sameAs* relation with the concept *yago:Java_(programming_language)*.

Authors of this approach produced in their evaluation of the disambiguation process a 89.15% of correct disambiguations of distinct polysemous tags. 11.71% of the tags have not been associated to any Wikipedia page. In addition, they evaluated the coverage of Wikipedia categories, YAGO classes and WordNet synsets, and produced 95%, 58%, and 18% accuracy, respectively. The accuracy of the disambiguation process seems very promising. However, this approach assumes that users always use ambiguous tags with just one meaning, which might not be true in some cases.

4.2.5 Passant's approach

Passant describes a collaborative approach, where users can manually perform all the tasks in our unified process. These users are assumed to be the taggers in the folksonomy, and will share the results of associating semantics to tags. Hence, unlike the approaches above, Passant's (2007) approach aims to generate tag-semantics at tag-creation time. The data source selected by Passant for the evaluation of this approach is a folksonomy from a corporate Web blog platform, where blog posts are annotated with tags.
Data selection and cleaning: In this approach, the data selection and cleaning activity is carried out by each user of the system who can annotate posts. Those tags used in a post that do not have a semantic association are displayed in a different color, so that the contributing user can enrich them semantically. The author does not provide statistics about the data set used to test *de* approach.
Context identification, disambiguation, and semantic identification: These activities are also carried out by users. The assumption is that users know the context because they know what the post content is, and thus if the tag is ambiguous they are able to choose the right meaning, which is

[13] `http://www.tagpedia.org/`

[14] A semantic knowledge base created from Wikipedia information `http://www.mpi-inf.mpg.de/yago-naga/yago/`

defined by concepts or instances of a predefined domain ontology. Polysemic tags can be associated to more than one ontology concept. In this case, users need to associate blog posts with tags, and with the concept they represent to avoid ambiguity. Users are provided with a list of URIs to select from, which are in turn selected from existing ontologies based on how similar concept names are to the given tags.

For our sample folksonomy, let us suppose that a user has two blog posts R1 and R2 in the system. In R1 the user has used two tags *Coffee* and *Java*. Then, he associates each tag with an appropriate class in the preference domain ontology. The system internally associates both tags with the post and their meaning in that post, that is, the ontology classes. With respect to R2, the user uses the *Java* tag, among others. Then, he associates this tag with an ontology class in the programming languages domain ontology. In this case, the system associates the *Java* tag, with the post and with its meaning in this post. Therefore, the system is able to differentiate between the two meanings of the *Java* tag, and also the system knows which meaning has been used in which post.

Involving users in the process is a straightforward approach to get rid of ambiguity. However, tagging proved highly successful because of its simplicity of creation. Passant's approach has been tested in a controlled environment of a corporate blog platform, but it has not been evaluated in an open environment. It is unclear how taggers would react to this approach, which controls and restricts their tagging activities. This approach will not scale very well because of its dependence on users to do most of the work themselves.

4.2.6 Maala et al.*'s approach*

The approach of Maala *et al.* (2007) uses semantic resources to automatically convert tags of photos into RDF semantic descriptions. The data source they used in their approach consists of a set of photos and their tags from Flickr.

Data selection and cleaning: In this approach, photos with tags that include at least a verb were selected. These tags are then transformed into their non-inflectional form using a stemmer. This selection is carried out because the authors needed to semantically describe photos with actions, and in consequence at least a verb is needed. Authors do not provide statistics about the data set used to test the approaches.

Context identification: Tag context is taken here to be the set of tags used to annotate a particular photo, and thus each photo is processed one at a time. Furthermore, each tag is classified according to one of the following categories: location, time, event, people, camera, and activity. This classification is carried out using some semantic resources including domain ontologies that have been created from external resources such as WordNet and existing Web sites.

Disambiguation: There are no disambiguation activities defined in this approach.

Semantic identification: The semantic identification activity in this approach creates RDF descriptions for each photo as follows. First, location tags are ordered according to an inclusion relation, and then RDF triples are created stating that the photo is *in* the smallest location, and that this location is *in* a broader location, and so on. Second, all 'time' tags are ordered according to an inclusion relation, and then RDF triples are created stating that the photo is *at* the smallest time tag, and that this time tag is *at* a broader time tag, and so on. Third, for each event tag an RDF triple is created stating that the photo *event* is the current event tag. Fourth, for each camera tag an RDF triple is created stating that the photo was *shot by* the current camera tag. And finally, for each activity tag an RDF triple is created stating that the photo *describes* an activity. Furthermore, WordNet is used to find the arguments of the verb related to the activity including the subject type of who performs the activity. If any of the photo tags correspond to the subject type, then a triple is created stating that the activity *agent* is that tag.

In the context of our sample folksonomy, let us suppose that we have a photo R1, and we have tagged it with all the tags, except for the *Coffee* tag. In addition, we have annotated this picture with more tags including *Madrid*, *Spain*, *January*, *2009*, and *John*. According to this approach, the tags *Madrid* and *Spain* are identified as locations. Hence, we can assert that the photo is *in Madrid*,

and that *Madrid* is *in Spain*. Furthermore, the tags *January* and *2009* are identified as time tags, and thus, we can assert that the photo is *at January*, and that *January* is *at 2009*.

On the other hand, the *Program* tag is the unique tag that can be considered as a verb (although it could also be a noun). Therefore, we can assert that the photo *describes* a *Program*. Furthermore, we extract from WordNet the sentence frames of the *Program* verb: (1) *somebody programs something* and (2) *somebody programs*. From these two sentence frames we could identify the arguments of the verb consisting of a mandatory subject of type *somebody*, and of an optional object of type *something*. In this case, the tag *John* could be identified as an instance of *somebody* so that we can assert that *John* is an *agent* of *Program*.

Although the authors of this approach show some examples of use, they do not provide any evaluation metric about the generated RDF descriptions. In a study about photo tags in Flickr the authors estimate that about 53% percent of photos include a tag representing an activity, and thus, the approach leave out of the process the remaining 47% of photos.

In the context activity where tags are placed in some predefined categories, some of these tags can be misclassified because of tag ambiguity and the approach does not provide any technique to fix this problem.

4.3 Hybrid approaches

So far we have described approaches aiming to group related tags using statistical techniques and some others aiming to associate tags to ontologies. In this section, we present some approaches relying on ontologies and clustering techniques whose goal is either to group related tags (Giannakidou *et al.*, 2008) or to associate semantic entities to tags (Specia & Motta, 2007).

4.3.1 Giannakidou et al.*'s approach*

Giannakidou *et al.* (2008) proposed a statistical approach for discovering the semantic of tags by clustering tags and resources, being resources represented by their annotations. This approach is based on a similarity measure that mixes tag co-occurrence with semantic similarity. The approach was tested with a set of Flickr photos depicting cityscape, seaside, mountain, roadside, landscape, sport-scenes, and locations.

Data selection and cleaning: This approach performs tag spelling normalization where the different spellings of a tag are mapped to a normalized version of that tag. Infrequent tags are filtered out, along with those that do not have a corresponding concept in WordNet, which is the terminological resource they use for calculating semantic similarity. The data set used to test the approach contains 3000 resources. From these resources the 30 most frequent tags were extracted to be analyzed.

Context identification: Giannakidou *et al.* (2008) consider the context for a tag to be the set of tags that co-occur with the given tag when anotating resources. Furthermore, the context of a resource is defined as the tags the users have assigned to it.

Disambiguation: This approach does not explicitly deal with disambiguation problems, however, authors claim that the grouping of resources and tags found in the next activity helps to disambiguate the meaning of tags.

Semantic identification: The semantic identification activity creates a graph where the resources, and the most frequent tags in the folksonomy, are represented as vertices. The graph edges associate resources with tags. An edge between a resource and a tag exists if their similarity value is above a certain threshold. In this approach each resource is represented by the set of tags used to annotate it so that the similarity between a tag and a resource is calculated as the maximum similarity value of the tag with each one of the tags used to annotate the resource. The similarity between two tags is a weighted sum of their social similarity and their semantic similarity. Social similarity is based on the co-ocurrence of both tags when annotating resources. For the semantic similarity, authors propose to map tags to concepts in a semantic resource. Then, the semantic similarity is calculated proportionally to the path distance between those concepts in the semantic

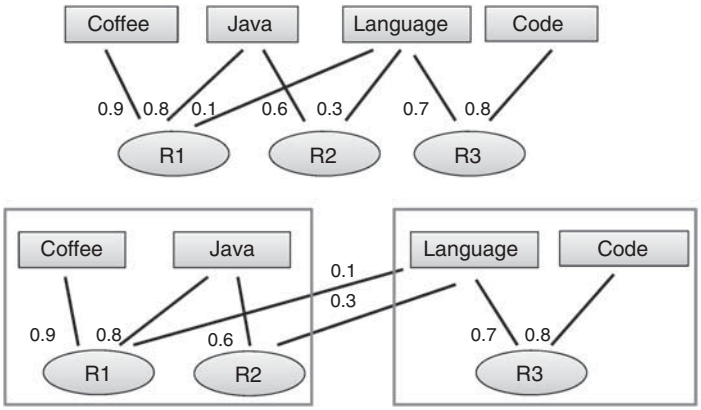

Figure 9 Top: sample bipartite graph relating resources and tags using social and semantic similarity. Bottom: clustered graph

resource. The bipartite graph relating resources and tags is then clustered using a spectral graph clustering algorithm whose goal is to create disjoint clusters so that the elements in the same cluster have high similarity and elements in different clusters have low similarity.

Let us suppose that the *language* tag has been also assigned to R1 and R3 in the example folksonomy. We have extracted from our folksonomy example the bipartite graph shown in Figure 9. This graph relates tags and resources by means of the similiarity value calculated relying on social and semantic similarity. Besides in the bottom of Figure 9 two groups found by the clustering algorithm are shown where the sum of similarities between elements of the same cluster is maximized while the sum of similarities of different clusters is minimized.

Giannakidou *et al.* (2008) tested their approach with varying weights assigned to the social and semantic similarity measures to see how each one of these measures affect the clusters found. Authors concluded that social similarity helps to disambiguate ambiguous tags since the context (i.e. co-occurring tags) helps to highlight the meaning of tags. Authors also stated that semantic similarity allows to find groups of synonyms, but fails to handle ambiguous tags. However, this approach clusters tags into disjoint groups. This means that a tag can belong to just one group and therefore if a tag has several meanings the approach will only identify the most frequent meaning for that tag according to the tag co-occurrence pattern. Moreover, the tags are grouped according to an abstract relation found by the clustering algorithm, but this relation is not defined semantically in terms of *subclass*, *part of*, *synonym* or any other relation. Similarly, the meaning of the tags is not defined explicitly.

4.3.2 Specia and Motta's approach

Specia and Motta (2007) proposed a semi-automatic approach using a mix of clustering and ontology-based techniques, focusing on two data sources (Flickr and Delicious), although the approach could be extended to any other folksonomy data source.

Data selection and cleaning: The data selection and cleaning activity starts by filtering out unusual tags. For instance, it filters out tags that do not start with a letter. Then, morphologically similar tags are grouped using the Levenshtein similarity metric. Finally, infrequent and isolated tags are filtered out. The data set used to test the approach comprised data from Delicious and Flickr. The Delicious data contains 7164 users, 14 211 resources, and 11 960 tags. On the other hand, the Flickr data consists of 6140 users, 49 087 resources, and 17 956 tags.

Context identification: The goal of the context identification activity here is to build clusters of related tags. First, the context of a tag is defined as the set of tags that co-occur with the current tag when annotating a resource or when they are used by the same user. To represent the context of a tag the authors use a vector whose number of elements is equal to the number of distinct tags in the folksonomy, and the values of each position corresponds to the number of times the tag co-occurs with the tag corresponding to the current position. In the case where the element of the

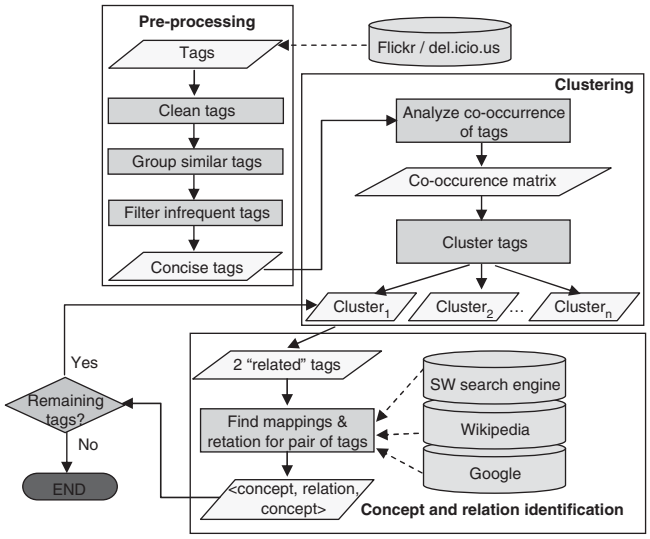

Figure 10 Process of associating semantics to tags (Specia & Motta, 2007)

Table 6 Tag and context vector representation

	Coffee	*Java*	*Language*	*Program*	*Code*
Coffee	1	1	0	0	0
Java	1	5	3	2	0
Language	0	3	3	2	0

vector correspond to the tag that is identifying the vector, the value for that element is the frequency of use of that tag in the folksonomy. Then, each tag is compared with other tags using their context vectors in order to find similar tags.

Disambiguation: When a tag is ambiguous it can have more than one pattern of co-occurrence. Thus, the set of similar tags found in the context identification may include tags with different meanings. The disambiguation activity analyzes each group of similar tags in order to find clusters of related tags based on high co-occurrence.

Semantic identification: Finally, for each cluster of related tags the semantic identification activity is carried out manually. A user uses a Semantic Web search engine (e.g. Swoogle[15]) to look for ontologies containing pairs of tags in the cluster. If an ontology is found that contains a pair of tags, then the semantic information about the tags (type, parents, domain, range) is used to establish relations between them.

The proposed approach is depicted in Figure 10. Let us apply this approach to our folksonomy example. First, we need to create the vectors to represent each tag and its context. Some examples of these vectors are shown in Table 6. Then, over this matrix we have to apply the clustering algorithm proposed in the disambiguation activity. As a result of this activity we can get two two groups of tags. A group with tags *Coffee* and *Java*, and the other group with tags *Language*, *Java*, *Program*, and *Code*. Then, the user looks in Swoogle for each pair of tags in each group and tries to establish manually the relations among tags. The Cyc Ontology[16] has a direct relation among the tags *coffee* as a beverage and *Java*, the latter being an *english alias* of the former. On the other

[15] http://swoogle.umbc.edu/

[16] http://sw.opencyc.org/

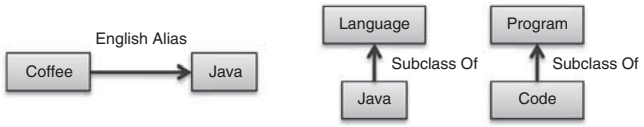

Figure 11 Tags' relations identified within group of tags

hand, in the LT4eL ontology[17] we found that *Java* is a *subclass of Language* as a programming language. In addition, we found that *Code* as source code is *subclass of Program* as a computer program. The final ontology is depicted in Figure 11.

One of the main advantages of this approach is that it combines clustering and ontology-based techniques. However, the approach has also some limitations. For example, the semantic identification activity requires users to analyze manually the ontologies retrieved from a Semantic Web search engine like Swoogle. However, an approach to automate this process has been introduced in Angeletou *et al.* (2008). Specia and Motta's approach is highly dependent on finding relations between tags in existing ontologies. It is therefore natural to expect that many tag pairs found in folksonomies will not be found in any ontology libraries, thus limiting the output of this approach.

With respect to evaluation, the authors do not evaluate how well the clusters of highly co-occurring tags in the similar tag sets help in the disambiguation of tag senses.

4.4 Consolidated overview of approaches

In this section we will provide a summary and comparison of the approaches presented in this paper. This overview is summarized in Table 7, which uses the activities identified in our unified process and some other characteristics that we have considered in our descriptions. In this table, the first column contains the reviewed approaches. The following columns are the characteristics evaluated using the unified process. The approach type can take the values of *Clu* for clustering-based approaches, *Ont* for ontology-based approaches, and *Hyb* for hybrid-based approaches. The auto column describes if the approach is automatic or manual, and it can take the values of *Yes* for automatic, *No* for manual, or *Semi* for semi-automatic. The data source column shows the folksonomies used to test the approaches in the original publications. Then, there are columns to specify if the approaches include or not some of the activities of the proposed general process.

Three of the approaches use clustering techniques to identify the hidden semantics of tags in folksonomies, while six more use ontologies to associate semantics to tags, and just two use a hybrid approach. Clustering techniques are used most of the time to find groups of related tags, whereas ontology-based approaches are used to associate semantic entities to individual tags.

Most of the approaches are automatic, except for Specia and Motta's, which is semi-automatic, and Passant's which is completely manual and focuses on user-generated semantic enrichment. The most studied data sources are Delicious and Flickr. In Hamasaki *et al.* (2007), the folksonomy was adapted from an academic conference support system, and in Passant a folksonomy of an enterprise blogging platform was used. Almost all the approaches implement a data selection and cleaning activity, defining the initial tag set and filtering out the tags they do not want to deal with, except for Hamasaki *et al.* (2007) where this activity was not described.

In all approaches some kind of context identification is included. The objective of this activity is usually for tag disambiguation or for semantic identification. Table 8 presents the different context definitions found in the reviewed approaches. Most of the approaches rely on tag co-occurrence when annotating resources regardless of the user. However, as these approaches ignore the user in the context definition, they are mixing the different meanings of tags given by the different users. Other approaches, such as Maala *et al.* (2007) and García-Silva *et al.* (2009) analyze the tags in the

[17] `http://www.lt4el.eu/index.php?content=tools#ontology`

Table 7 State of the Art consolidated overview

Approach	Type	Auto	Data Source	Data selection and cleaning	Context identification	Disambiguation	Semantic identification
Mika (2007)	Clu	Yes	Delicious, Other	Yes	Yes	No	Yes
Hamasaki *et al.* (2007)	Clu	Yes	Polyphonet	No	Yes	Yes	No$^{\mp}$
Giannakidou *et al.* (2008)	Hyb	Yes	Flickr	Yes	Yes	Yes	No$^{\mp}$
Jäschke *et al.* (2008)	Clu	Yes	Delicious, BibSonomy, Other	Yes	Yes	No	No$^{\mp}$
Specia and Motta (2007)	Hyb	Semi	Delicious, Flickr	Yes	Yes	Yes	Yes
Angeletou *et al.* (2008)	Ont	Yes	Flickr	Yes	Yes	Yes	Yes^{+}
Cantador *et al.* (2008)	Ont	Yes	Union of Flickr and Delicious	Yes	Yes	No	Yes^{+}
García-Silva *et al.* (2009)	Ont	Yes	Flickr	Yes	Yes	Yes	Yes^{+}
Tesconi *et al.* (2008)	Ont	Yes	Delicious	Yes	Yes	Yes	Yes^{+}
Passant (2007)	Ont	No	Enterprise Folksonomy	Yes	Yes	Yes	Yes^{+}
Maala *et al.* (2007)	Ont	Yes	Flickr	Yes	Yes	No	Yes

+Tags are related to semantic resources.

∓The approach finds groups of related tags without identifying the relations among tags nor their meaning.

Table 8 Context definitions

Approach	Context
Maala *et al.* (2007) García-Silva *et al.* (2009)	Co-occurring tags in the user post
Tesconi *et al.* (2008)	Co-occurring tags when annotating any resource tagged by the user
García-Silva *et al.* (2009)	User tags regardless of the resource
Hamasaki *et al.* (2007) García-Silva *et al.* (2009)	Co-occurring tags in the user social network
Mika (2007) Giannakidou *et al.* (2008) Specia and Motta (2007) García-Silva *et al.* (2009)	Co-occurring tags when they are used to annotate a resource
Mika (2007) Specia and Motta (2007)	Co-occurring tags when they are used by an annotator
Angeletou *et al.* (2008)	All analyzed tags
Kennedy *et al.* (2007)	Latitude and longitude of the geo-tagged resources and timestamp
Cantador *et al.* (2008)	Wikipedia category list associate to each tag

user post getting rid of the use of tags in different meanings. Tesconi *et al.* (2008) propose to use the co-occurring tags in the resources annotated by the user, and in the same respect, García-Silva *et al.* (2009) proposed to use all the user vocabulary. If a user has used the tag in more than one sense, then the analysis of this context will result in the most frequent sense for that tag. The idea behind using the co-occurring tags in the user social network first introduced in Hamasaki *et al.* (2007), is that groups of people sharing some interest, for example, scientist and practitioners, tend to use the same vocabulary in a particular field. Finally, only Kennedy *et al.* (2007) use tagging metadata information as part of context definition.

Three of the research works ignore the disambiguation problem, while the other four suggest some technique to disambiguate tag meanings. Regarding the semantic identification activity only Hamasaki *et al.* (2007) and Jäschke *et al.* (2008) do not describe how to define explicitly or formally the tag semantics.

Mika (2007) uses social network metrics and set theory to define the tag semantics, while Specia and Motta (2007), Angeletou *et al.* (2008), García-Silva *et al.* (2009), Tesconi *et al.* (2008), and Passant (2007) use external semantic resources. Maala *et al.* (2007) also use external semantic resources to define in which category (e.g. location, time, event, etc.) each tag fits better, however, in the end the RDF triples generated are not related to the semantic resources, losing the advantage of using the obtained intermediate results.

Passant's approach differs from all others in which all the activities are carried out manually by users. For instance, the user is the one who decides which tags are to be enriched semantically. Also, when the user has to define which ontology concept he wants to associate with the tag, he has to understand the context in which the tag is used and if the tag is ambiguous then he has to define the right meaning in order to associate the best ontology concept. Furthermore, the user will need to understand the meaning of the ontology concepts the system suggests for the given tags to be able to select the correct URIs.

Finally, in spite of the fact that folksonomies as collaborative Web applications have a worldwide scope reaching a wide range of users, multilinguality issues have not been addressed explicitly by any of the reviewed approaches. Nevertheless, clustering-based approaches are able of processing multilingual tags due to the fact that they rely on relatedness measures among tags and to calculate them the tag language is not relevant. On the other hand, ontology-based approaches are highly dependent of the language of the semantic resource and thus multilingual tags cannot be processed by these approaches.

5 Recommendations

After analyzing the most relevant research works on association of semantics to tags, we provide some recommendations on how to improve the current approaches giving suggestions on each one of the identified activities in the unified process. Besides, we present future applications based on the results of the process of associating semantics to tags.

5.1 Semantic association process

In Marlow *et al.* (2006) folksonomies have been surveyed to identify their main characteristics. Authors claim that folksonomies differ in functionalities such as user tagging rights, tag recommendation strategies, social networks, etc. Another classification is proposed by Vander Wal[18], who differentiate among broad and narrow folksonomies. A broad folksonomy such as Delicious has many people tagging the same object and every person can tag the objects with their own tags in their own vocabulary. In this kind of folksonomies, as time goes by, more and more users, often because of a recommendation strategy, use the most frequent tags, and thus a vocabulary emerges for a group of users to refer to a resource. On the other hand, narrow folksonomies like Flickr are those where one or few people provide tags for one resource.

In summary, tags in broad folksonomies tend to have a frequency of use with respect to the annotated resources (i.e. used by many to tag the same resource), while this frequency does not exist in narrow folksonomies (i.e. resources are not usually tagged by many people). We suggest that *the data source must be categorized according to whether it is a narrow or broad folksonomy, and each of the approaches can be better applied to a different type of data source*. Thus, the approaches that do not take into account the tag frequency by resource, such as the approach of Mika, are suitable for both types of folksonomies, while the approaches that do, such as Specia and Motta approach, should only be used for broad folksonomies.

Moreover, we have identified some recommendations in each one of the activities of the uniform process for the associaton of semantics to tags presented in Section 2.

Data selection and cleaning: Ontology-based and clustering-based approaches can benefit from tag preprocessing. We suggest that the following preprocessing tasks should be carried out:

- Common stop-words such as pronouns, articles, etc. should be removed (Cantador *et al.*, 2008).
- Misspelling tags can be processed easly using mechanism such as *Google did you mean*[19] as suggested by Specia and Motta (2007), or the *Yahoo spelling suggestion service*[20]. However, those services have a maximum number of calls per day per IP (Internet Protocol) address limiting their use.
- Acronyms, abbrevations, and proper names can be managed using Wikipedia (Cantador *et al.*, 2008).
- Meaningless tags should be filtered out (e.g. tags mixing numbers and letters, or including special characters; Specia & Motta, 2007; Angeletou *et al.*, 2008; Cantador *et al.*, 2008; Giannakidou *et al.*, 2008).
- By identifying the stem of a tag as proposed in Giannakidou *et al.* (2008)several morphological variations of a tag will be grouped in a unique term.
- Compound words used as tags should be splitted, in case those words represent different concepts, or represented in an standard way as propossed in Cantador *et al.* (2008).

In addition, ontology-based approaches can increase their coverage, by generating many representations of the compound word, that is, using different characters to concatenate the words, as proposed in Angeletou *et al.* (2008). However, generating and processing these alternative representations will increase the overall processing time.

[18] `http://www.vanderwal.net/random/entrysel.php?blog=1635`

[19] `http://code.google.com/intl/en/apis/ajaxsearch/`

[20] `http://developer.yahoo.com/search/boss/boss_guide/Spelling_Suggest.html`

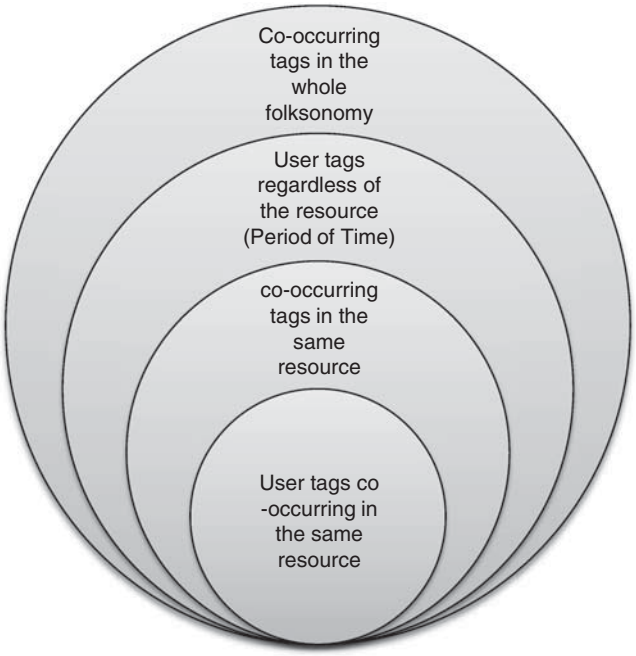

Figure 12 Incremental context definition

Context identification: The objective of a semantic association process can be to identify the general semantics of a part or of the whole folksonomy as in Mika (2007) and Specia and Motta (2007), or to identify the semantics of tags for each user as in Maala *et al*. (2007), Passant (2007), and García-Silva *et al*. (2009). According to the objective of the semantic association process, we suggest that the *contextualization of tags can be improved by taking advantage of the different levels of information contained in a folksonomy*.

We propose an incremental definition of the context of a tag (see Figure 12). Initially, the context can be defined using the co-occurring tags when the user is tagging a resource as Maala *et al*. (2007) suggested. If this context is not enough to achieve the tag disambiguation or to choose the right semantic entity, we can go a step further and add new information to the context using tags co-occurring for the same resource used by other users as Mika (2007), Giannakidou *et al*. (2008), and Specia and Motta (2007) suggested. Following with this idea, the context of a tag can be increased using all the user tags (García-Silva *et al*., 2009), all the tags of the user contacts (Hamasaki *et al*., 2007), the most frequent co-occurring tags in the whole folksonomy (García-Silva *et al*., 2009), or taking into account the time dimension, for instance, with the tags used by the user in a particular period of time (García-Silva *et al*., 2009).

Besides, output of statistical approaches are clusters or hierarchies of tags, but the meaning of these tags or groups is not defined explicitly. We suggest that *statistical approaches should be used in combination with ontology-based approaches*, which actually define the meaning of tags, as a means of improving the contextualization of tags and of actually proceeding to the semantic enrichment of these clusters. For instance, the clusters of tags found in Mika's approach or the triples of users, tags, and resources found by Jäschke *et al*. can be used to extend the contextualization of Angeletou *et al*. approach, which could then lead to better disambiguation results and also to better associations of ontology entities with tags.

Disambiguation: Clustering-based approaches addressing tag ambiguity (Hamasaki *et al*., 2007; Specia & Motta, 2007; Giannakidou *et al*., 2008) find groups of tags that implicitly define the meaning of each tag. In contrast, ontology-based approaches (Angeletou *et al*., 2008; Tesconi *et al*., 2008; García-Silva *et al*., 2009) explicitly states the meaning of each tag by means of its association to a semantic entity. WordNet has been used initially as sense repository (Angeletou *et al*., 2008;

Giannakidou *et al.*, 2008). However, due to its low coverage, researches have started to use as sense repository collaborative-created knowledge bases such as Wikipedia and DBpedia (Cantador *et al.*, 2008; Tesconi *et al.*, 2008; García-Silva *et al.*, 2009).

When the disambiguation or the semantic association activities cannot be carried out automatically because of insufficient information, user participation could be requested, as proposed by Passant. We can go a step further and propose a semi-automatic approach, where the system calculates the list of probable meanings for a tag and select one as the intended meaning. However, users have the possibility to manually assert the meaning for a tag either by confirming the system decision or by picking another concept as the tag meaning.

Semantic identification: So far most of the approaches associating semantic entities to tags rely on string matching techniques to find candidate ontology concepts and then use the tag context to choose the one that better describes the meaning of a tag. However, this activity implies the transition from a flat space, that is, without hierarchies, in the folksonomy side, to a hierarchical space in the ontology side. Some research works as Angeletou *et al.* (2008) tackle this problem by associating tags initially to WordNet synsets, and then the synset hierarchical structure is compared against ontologies.

We suggest, as an alternative to current approaches, to create tag hierarchies based on the folksonomy information which then can be compared against ontology hierarchies. In Mika (2007) the creation of a tag hierarchy is proposed. A concept A is a super-concept of concept B if the set of resources classified under B is a subset of the entities classified under A, and that A should be significantly larger than B.

Heymann and García-Molina (2006) presents a greedy algorithm to create tag hierarchies. In this approach tags are represented using vectors, where each position represents the times the tag was used to annotate a resource. Those vectors are compared using as similarity measure the cosine of the angle they form. With this information a similarity graph is built, where the vertices are tags, and edges are weighted by the similarity measure calculated previously. Next, for each tag the network centrality is calculated in the similarity graph. Finally, the tags are ordered in a list according to this centrality value. This list is processed, starting with the most central tag, to create the tag hierarchy.

Once the tag hierarchy has been created, ontology mapping techniques (Euzenat & Shvaiko, 2007) can be used to match the tags and its hierarchy against ontology concepts.

Finally, as we mention in Section 4.4, none of the ontology-based approaches address multilingual tags. To overcome this limitation, we suggest to use mutilingual semantic resources such as EuroWordNet[21], Wikipedia, and DBpedia in the different languages in which they are available.

5.2 Applications of the results of associating semantics to tags

Another interesting characteristic to be dealt with is how to use the results obtained from the described approaches, that is, how to use the ontologies or the semantic associations obtained for tags or groups of tags. First of all, *tag suggestion strategies can rely on the ontological relations between the entities to suggest new tags to a user*, instead of only relying on syntactic relations. For instance, if the user has typed a tag that has been associated with an ontological entity, then the system can recommend as tags the tags associated with the superclasses or subclasses of that entity, or to other entities that are associated through object properties. If the tag is ambiguous, then the system can also use the superclasses or subclasses to ask the user to choose which meaning of the tag he wants to refer to.

The *search process in collaborative tagging sites can also make use of the semantic association of tags* in order to get more accurate results. If a user wants to search for a tag, the system can figure out whether the tag is ambiguous or not using the generated ontology. If the tag is ambiguous, the

[21] http://www.illc.uva.nl/EuroWordNet/

system can exploit the ontological relations to build a context for each one of the meanings of the tag. Then, for each sense the system could execute a particular contextualized search and present the results to the user in accordance to the senses found. Moreover, query expansion can be carried out using semantic relations among the ontology concepts (e.g. synonyms, siblings, subclass of, etc.) to improve the results of the retrieval processes.

User profiles built from folksonomy information, as those proposed in Szomszor *et al.* (2008), can also *benefit from the association of semantic to tags*. Traditional user profiles consist of sets of concepts related to objects including, in some cases, ratings. These profiles are used, for instance, for content-based recommendation systems to suggest new objects to users based on their past preferences. If these profiles are enriched semantically, then new recommendations can be provided to users, exploiting the relations between ontological concepts which are part of the user profile.

In Specia and Motta (2007) authors mention that *semantically enriched folksonomies are a new source of semantic information to enrich or populate existing ontologies*. According to the NeOn glossary[22], ontology enrichment refers to the activity of extending an ontology with new conceptual structures (e.g. concepts, roles, axioms, etc.). On the other hand, ontology population is defined as a knowledge acquisition activity that relies on (semi-) automatic methods to transform unstructured, semi-structured, and structured data sources into instance data. Folksonomies include quickly new vocabulary as a result of the collaborative process where many users participate in the tag supply. This makes a difference with respect to traditional semantic resources used to add new information to existing ontologies like taxonomies and thesauri (García-Silva *et al.*, 2008). When an ontology is modified there is the necessity of keeping track of those changes. In this respect ontology evolution processes as defined in Noy and Klein (2004) are in charge of managing ontology changes and their effects by creating and maintaining different variants of the ontology. However, in this case, the challenge is how to synchronize the evolution of the two structures, the folksonomy on the one hand and the ontology on the other hand.

Finally, we want to mention that currently there exists *a lack of standard testbeds to test all these approaches under the same conditions*, allowing us to compare, in a uniform way, the results of the different approaches. In this context, we have noticed also the lack of standard evaluation metrics to assess the different approaches. We believe that creation of this testbed will leverage the research in associating semantics to folksonomies to a next level, and thus, we encourage researches to pursue the creation of such database.

6 Conclusions

In recent years, there has been a growing number of research works trying to associate semantic information to tags in folksonomies. The main objective of these works is to identify the shared conceptualizations hidden in folksonomies using a wide variety of statistical clustering and ontology-based techniques. Clustering techniques use tag co-occurrence as a measure of tag relatedness, and some data mining tasks with the objective of finding groups of related tags. Ontology-based techniques aim to associate tags with ontological concepts, exploiting publicly available semantic resources.

In this paper, we have reviewed the most relevant research works in this context. We defined a unified process to characterize and compare these approaches. This unified process identifies the main activities carried out in most of the analyzed approaches and also the categories relevant to our analysis. We used this process along with an example folksonomy in order to describe and discuss the different approaches. Furthermore, we have presented a summary of the approaches and a table to compare all their activities. Finally, we have presented a set of recommendations on how to improve current approaches.

[22] `http://www.neon-project.org/web-content/images/Publications/neonglossar-yofactivities.pdf`

Acknowledgment

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. In addition, we want to thank Sofia Angeletou for the valuable information provided.

References

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

Au Yeung, C. M., Gibbins, N. & Shadbolt, N. 2007. Understanding the semantics of ambiguous tags in folksonomies. In *Proceedings of the International Workshop on Emergent Semantics and Ontology Evolution (ESOE2007)*, Busan, South Korea.

Baeza-Yates, R. & Ribeiro-Neto, B. 1999. *Modern Information Retrieval*. ACM Press.

Begelman, G., Keller, P. & Smadja, F. 2006. Automated tag clustering: improving search and exploration in the tag space. In *Proceedings of the 15th International World Wide Web Conference*, Edinburgh, Scotland.

Breslin, J., Decker, S., Harth, A. & Bojars, U. 2006. SIOC: an approach to connect web-based communities. *International Journal of Web Based Communities* **2**(2), 133–142.

Buitelaar, P., Cimiano, P. & Magnini, B. 2005. Ontology learning from text: an overview. In *Ontology Learning from Text: Methods, Evaluation and Applications Frontiers in Artificial Intelligence and Applications Series*, Buitelaar, P., Cimiano, P. & Magnini, B. (eds). 3–12.

Cantador, I., Szomszor, M., Alani, H., Fernández, M. & Castells, P. 2008. Enriching ontological user profiles with tagging history for multi-domain recommendations. In *Proceedings of the 1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008)*, Tenerife, Spain.

Echarte, F., Astrain, J. J., Córdoba, A. & Villadangos, J. 2007. Ontology of folksonomy: a new modeling method. In *Proceedings of the Semantic Authoring, Annotation and Knowledge Markup (SAAKM)*, Whistler, Canada.

Euzenat, J. & Shvaiko, P. 2007. *Ontology Matching*. Springer-Verlag.

García-Silva, A., Szomszor, M., Alani, H. & Corcho, O. 2009. Preliminary results in tag disambiguation using DBpedia. In *Proceedings of the 1st International Workshop in Collective Knowledge Capturing and Representation (CKCaR09)*, California, USA.

García-Silva, A., Gómez-Pérez, A., Suárez-Figueroa, M. C. & Villazón-Terrazas, B. 2008. A pattern based approach for re-engineering non-ontological resources into ontologies. In *Asian Semantic Web Conference (ASWC 2008)*, Bangkok, Thailand.

Giannakidou, E., Koutsonikola, V., Vakali, A. & Kompatsiaris, Y. 2008. Co-clustering tags and social data sources. In *Proceedings of the 9th International Conference on Web-Age Information Management (WAIM2008)*, Zhangjiajie, China.

Golder, S. & Huberman, B. 2006. Usage patterns of collaborative tagging systems. *Journal of Information Science* **32**(2), 198–208.

Gruber, T. 2005. Ontology of folksonomy: a mashup of apples and oranges. In *Proceedings of the 1st On-line Conference on Metadata and Semantics Research (MTSR'05)*.

Hamasaki, M., Matsuo, Y., Nisimura, T. & Takeda, H. 2007. Ontology extraction using social network. In *International Workshop on Semantic Web for Collaborative Knowledge Acquisition*, Hyderabad, India.

Heymann, P. & García-Molina, H. 2006. *Collaborative creation of communal hierarchical taxonomies in social tagging systems*. Technical Report, Stanford University.

Hotho, A., Jäschke, R., Schmitz, C. & Stumme, G. 2006. Information retrieval in folksonomies: search and ranking. In *The Semantic Web: Research and Applications*. Springer, 411–426.

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

Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L. & Stumme, G. 2007. Tag recommendations in folksonomies. In *Knowledge Discovery in Databases: PKDD 2007, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Warsaw, Poland.

Kennedy, L., Naaman, M., Ahern, S., Nair, R. & Rattenbury, T. 2007. How Flickr helps us make sense of the world: context and content in community-contributed media collections. In *Proceedings of the ACM Multimedia 2007*, Augsburg, Germany.

Kim, H. L., Breslin, J. G., Yang, S. K. & Kim, H. G. 2008a. Social semantic cloud of Tag:Semantic Model for social tagging. In *Proceedings of the 2nd KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications*, Incheon, Korea.

Kim, H. L., Scerri, S., Breslin, J. G., Decker, S. & Kim, H. G. 2008b. 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*, Berlin, Germany, 128–137.

Knerr, T. 2006. Tagging ontology—towards a common ontology for folksonomies. `http://tagont.googlecode.com/files/TagOntPaper.pdf`

Lee, S. & Yong, H. 2007. Tagplus: a retrieval system using synonym tag in folksonomy. In *Proceedings of the 2007 International Conference on Multimedia and Ubiquitous Engineering*. IEEE Computer Society, 294–298. Washington, DC.

Marlow, C., Naaman, M., Boyd, D. & Davis, M. 2006. HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Proceedings of the seventeenth conference on Hypertext and hypermedia (HYPERTEXT '06)*, Odense, Denmark.

Maala, M. Z., Delteil, A. & Azough, A. 2007. A conversion process from Flickr tags to RDF descriptions. In *Proceedings of the BIS 2007 Workshop on Social Aspects of the Web*, Poznan, Poland.

Maedche, A. & Staab, S. 2001. Ontology learning for the Semantic Web. *IEEE Intelligent Systems* **16**(2), 72–79.

Mika, P. 2007. Ontologies are us: a unified model of social networks and semantics. *Journal of Web Semantics* **5**(1), 5–15.

Newman, R. 2005. *Tag ontology*. `http://www.holygoat.co.uk/projects/tags/`

Noy, N. F. & Klein, M. 2004. Ontology evolution: not the same as schema evolution. *Knowledge and Information Systems* **6**(4), 428–440.

Passant, A. 2007. Using ontologies to strengthen folksonomies and enrich information retrieval in weblogs. In *Proceedings of the 1st International Conference on Weblogs and Social Media (ICWSM)*, Boulder, CO.

Passant, A. & Laublet, P. 2008. 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*), Beijing, china.

Qiu, Y. & Frei, H. P. 1993. Concept based query expansion. In *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval*, Pittsburgh, USA.

Scerri, S., Sintek, M., Ludger van, E. & Handschuh, S. 2007. *NEPOMUK Annotation Ontology*. `http://www.semanticdesktop.org/ontologies/nao/`

Schmitz, C., Hotho, A., Jäschke, R. & Stumme, G. 2006. Mining association rules in folksonomies. In *Data Science and Classification: Proceedings of the 10th IFCS Conference*, Ljubljana, Slovenia, 261–270.

Specia, L. & Motta, E. 2007. Integrating folksonomies with the Semantic Web. In *Proceedings of the 4th European Conference on the Semantic Web: Research and Applications*. Innsbruck, Austria.

Szomszor, M., Alani, H., Cantador, I., O'Hara, K. & Shadbolt, N. 2008. Semantic modelling of user interests based on cross-folksonomy analysis. In *Proceedings of the 7th International Semantic Web Conference (ISWC)*. Karlsruhe, Germany.

Tapscott, D. & Williams, A. D. 2006. *Wikinomics: How Mass Collaboration Changes Everything*. Penguin Books Ltd, ISBN 978-1-59184-138-8.

Tesconi, M., Ronzano, F., Marchetti, A. & Minutoli, S. 2008. Semantify del.icio.us: automatically turn your tags into senses. In *Social Data on the Web, Workshop at the 7th International Semantic Web Conference*, Karlsruhe, Germany.

Wu, Z. & Palmer, M. 1994. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, New Mexico, USA, 133–138.