Consiglio Nazionale delle Ricerche

Semantic Web gets into collaborative tagging

M. Tesconi, F. Ronzano, S. Minutoli, A. Marchetti, M. Rosella

IIT TR-06/2007

Technical report

Maggio 2007

Istituto di Informatica e Telematica

Semantic Web gets into collaborative tagging

Tesconi Maurizio[1], Ronzano Francesco, Minutoli Salvatore[1]
Marchetti Andrea[1], Rosella Marco[1]

[1]CNR, IIT Department, Via Moruzzi 1, I-56124, Pisa, Italy
(maurizio.tesconi, salvatore.minutoli, andrea.marchetti, marco.rosella)@iit.cnr.it,
fr.ronzano@virgilio.it

Abstract. Collaborative tagging is a new content sharing and organization trend, mainly diffused over the Web, which has attracted growing attention during the last years. It refers to the process by which many users add metadata in the form of keywords to shared content. Today many different collaborative tagging systems are available on the Web, enabling users to add descriptive keywords to different types of Internet resources (web pages, photos, videos, etc.). The great number of advantages offered by the availability of collaboratively tagged resources in terms of their organization and shared information is underlined by their growing adoption, also in non-technical communities of users. In spite of this, analyzing the current structure and usage patterns of collaborative tagging systems, we can discover many important aspects which still need to be improved so as to bring tagging systems to their full potential. In particular, problems related to synonymy, polysemy, different lexical forms, different spellings and misspelling errors, but also the lack of accurancy caused by different levels of precision and distinct kinds of tag-to-resource association represent a great limit, causing inconsistencies among the terms used in the tagging process and thus reducing the efficiency of content search and the effectiveness of the tag space structuring and organization. This kind of problems is mainly caused by the lack of semantic information inclusion in the tagging process. Considering the increasing attention focused on the Semantic Web, we propose a new model of tagging system, based on semantic keywords. We let the users easily define the meaning of their tags, referencing some sort of social ontology. As social ontology we explore the adequacy of the support offered by the entries of Wikipedia and WordNet. Finally we present SemKey, a tool that allows users to tag in a semantic context, providing an evaluation of the system proposed in comparison with classical tagging tools.

Table of Contents

List of Tables

List of Figures

1 Introduction

During the last few years, the Web has experienced the growing diffusion of many kinds of collaborative tagging systems and the related increase of communities of taggers [CM06] [Wei05]; they are actively involved in the process of pointing out and cataloguing resources of interest, exploiting the growing amount of information collected to improve their searches and content discovery process. Some of the most used and representative collaborative tagging services are [THS05]:

- **Del.icio.us** (http://del.icio.us): it allows user to assign a free set of tags to a Web resource identified by its URL; this kind of tagging schema is also known as 'social bookmaring', because users can create and share resources' annotations in a way similar to locale bookmarking systems integrated in existing browsers;
- **Flickr** (http://www.flickr.com): this is a photo sharing system; each user can share and tag his personal photo and access and tag photos of other users;
- **Technorati** (http://www.technorati.com): it allows authors to tag their blog's posts, aggregating information contained in weblogs and facilitating their search.

All these tagging systems just listed are usually adopted by particular communities of users; del.icio.us by Computer Science experts, Flickr mainly by amateur photographers and Technorati by bloggers.

As we can argue also reading this short description of significant examples, tagging represent a collaborative social effort of a community of users constituted around a tagging service; with his tagging action, every user, mainly on the basis of its interests, directly contributes to the creation of a shared metadata collection, progressively augmenting its relevance and its richness of useful and shared data. The three main components of collaborative tagging systems are: users, resources and tags [LA05]. Users may be connected in groups with common interests; resources may be related by the different kinds of links which constitute the basis of current Web; tags provide the connection between a single user and a particular resource. When every user can assign a freely defined set of tags to a resource, the tag collection will reflect the social attitudes of the community of users and a shared social organization and structuring of the tag-space will emerge: this phenomenon is referred to as emergent semantics. It continuously adapts the tag space to the way users choose to describe resources, reflecting their tagging behavior.

The result of this process of adaptive social structuring of the tag-space in a collaborative tagging system has recently been defined as folksonomy [Smi04]. A folksonomy is the outcome of the fusion of two words: 'folk' and 'taxonomy'. 'Folk' is used to indicate the social collaborative component of the process of tags definition; 'taxonomy' instead refers to the method of organizing concepts in predefined and sometimes rigid structures, in order to better define their semantics and relations. When we speak about folksonomy, we refer to the collaborative and progressive definition of a relaxed categorization and organization of content, not based on a rigid hierarchical structure, and the related emergent semantic specification of concepts, or better of the meaning of tags. In this way the user has the freedom of choose autonomously his tags.

Many formal (research articles references [Bec06]) and informal (blogs references [She05] [Kro]) analysis of collaborative tagging system have also identified the low user learning curve and the relatively low bootstrapping cost of this kind of services as two relevant factors influencing their spread and rapid diffusion.

Analyzing in more depth the current structure and usage patterns of collaborative tagging systems, we can discover many important aspects which still need to be improved so as to really exploit their real potential.

2 Weak points of current collaborative tagging systems

When we analyse existing collaborative tagging systems, we can point out some relevant weak features; in particular, many of them can be related to the absence of any semantic information in the process of assigning descriptive keywords to a resource ([GH05], [ZXS06], [CM06], [GT06], [Mat04]). As a consequence, we can identify the following main causes of weakness:

- **Polysemy** (2.1)
- **Synonymy** (2.2)
- **Different lexical forms** (2.3)
- **Misspelling errors or other spelling** (2.4)
- **Different levels of precision** (2.5)
- **Different kinds of tag-to-resource association** (2.6)

Through the following example, we give a summary of the most important weak points of existing collaborative tagging systems just listed above. Let suppose that there are four different Web users: John, Monica, Bill and Anne. John, Monica and Bill are browsing the same Web resource speaking about a new model of Jaguar, a British luxury car manufacturer and decide to tag it. They have in mind to state that the Web resource is about cars, intended as a concept. Anne is also browsing and is searching information about the jaguar, the large spotted feline; after a lot of Web searching activity, she decides to tag the jaguar section of an interesting web site about jaguars, intended as a concept. All those situations are represented in Figure 1. Every user may freely choose one or more tags (characters' strings) to describe a Web resource. John, Monica and Bill refer to the concept of car choosing only one tag; John uses the word 'automobile' (one of the many synonyms associated to the concept of car), Monica the plural form of the word 'car' (different lexical form), Bill better specifies the concept with an increased level of precision using the word 'jaguar' (different level of precision) that has also another meaning, that one intended by Anne who adopts the same tag string to refer to the large spotted feline (polysemy). Moreover, 'motor-car' and 'motor_car' are possible different spelling of the same word and 'autmobile' represents a possible user spelling error during the tagging of a resource. Finally Anne links two tags to the visited Web resource, 'jaguar' and 'interesting' with

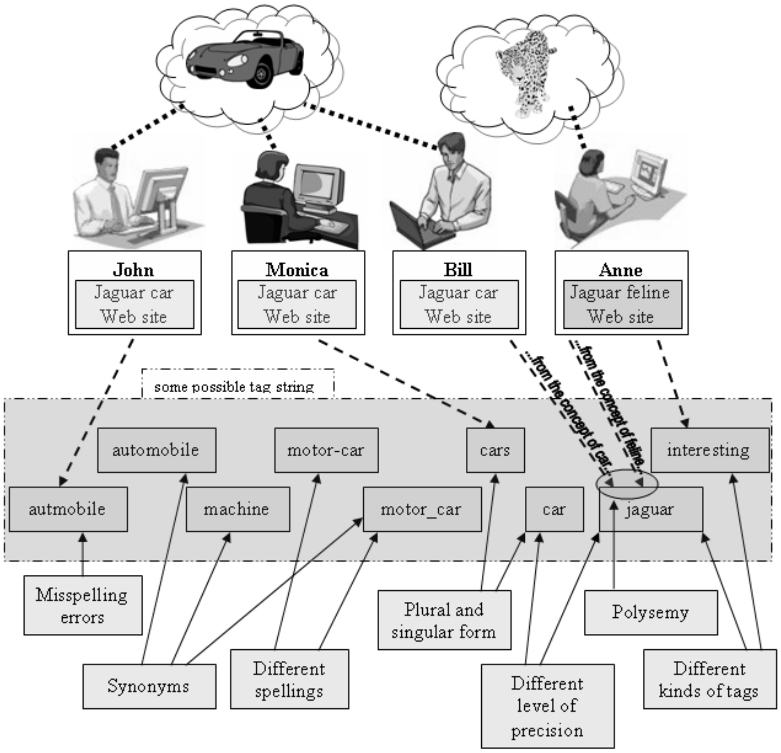


Fig. 1. The tag choice problems

a different purpose: the first one to describe the topic of the resource and the second one to express his personal opinion about the resource (different kinds of tag-to-resource association).

We have considered two main parameters to evaluate tag based searches in terms of their retrieval effectiveness; these parameters are usually adopted to measure how well an information-retrieval system, in this case a tag based search system, is able to execute a specific search [SS02]:

- **Precision** : the percentage of all retrieved resources that are actually relevant to the query;
- **Recall** : the percentage of all relevant resources present in the system that are retourned by the search.

Generalizing, most problems of current collaborative tagging systems can be traced back to the existence of $n : m$ relations between concepts and tags used to identify an intended concept. When a single tag is used to point out different concepts ($Tag(1) \longrightarrow Concept(n)$), polysemy issue occurs. When we adopt that tag to find all resources related to a specific intended concept, *precision decreases* because of the noise generated by the other retrieved resources dealing with different concepts but tagged using the same tag. In Figure 2 we show an example of result noise generated when we search for all the resources tagged with the keyword 'machine' meaning the concept of four wheels vehicle.

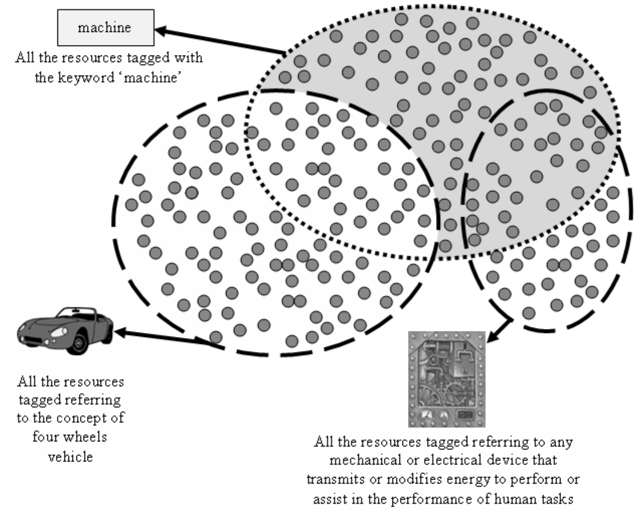

machine
All the resources tagged with the keyword 'machine'
All the resources tagged referring to the concept of four wheels vehicle
All the resources tagged referring to any mechanical or electrical device that transmits or modifies energy to perform or assist in the performance of human tasks

Fig. 2. A single tag used to point out different concepts

On the other side, multiple tags can be used to refer to the same concept ($Tag(n) \longrightarrow Concept(1)$); this can be caused by synonymy, different lexical forms, misspelling errors or other spelling. In this case, using one of the different tags to refer to a concept and find all related resources, *recall decreases* because of the presence of other relevant resources that are not retrieved since they are tagged using distinct words that point to the same concept. In Figure 3 we show an example of retrieval rate lowering that occurs when we search for all the resources tagged with the keyword 'machine', but we are not able to retrieve all the resources tagged thinking about the concept of four wheels vehicle using different tags ('cars', 'auto', 'automobile' and 'car').

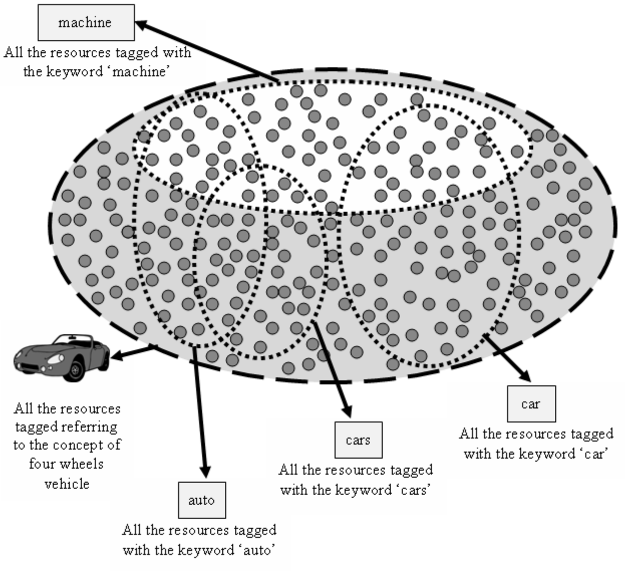

machine
All the resources tagged with the keyword 'machine'
All the resources tagged referring to the concept of four wheels vehicle
auto
All the resources tagged with the keyword 'auto'
cars
All the resources tagged with the keyword 'cars'
car
All the resources tagged with the keyword 'car'

Fig. 3. The same concept referred by different tag

2.1 Polysemy

When the user performs a tag based search, he needs to properly modify the search tag set so as to increase precision and recall. Usually a set of tags is used to specify a particular meaning; in fact we can notice that, during a tag-based search, one or more new tags are often added to the search tag set to disambiguate the meaning of tags already present.

If the user of a social tagging system wants to find information about the Jaguar car manufacturer, he could type only the word 'jaguar' to form the search tag set, obtaining as result every Web resource tagged with this word. Obviously only part of this result is of real interest to the user, in fact the word ‘jaguar’ may show a lot of other meanings (polysemy [GH05], [CM06]): a large felid animal, the codename of an Apple operating system, a video game console made by Atari, a guitar built by Fender, etc. In such a situation the user usually adds other tags to the search tag set, in order to better define the intended meaning; for instance, he could choose the tag ‘car’ to find all resources tagged with ‘jaguar’ and ‘car’ as an attempt to increase the precision of his search. Considering del.icio.us[1], the most popular tags used to refer to the ‘automobile’ word are: ‘car’, ‘cars’, ‘auto’ and ‘automotive’. As a consequence we can assume that when a user wants to specify the meaning of the tag ‘jaguar’, he will usually add one of the four tags just mentioned.

When we search for all Web resources tagged using ‘jaguar’, we obtain 1450 results. In Table 1 we can see the four subsets of this resources’ group produced adding to the tag ‘jaguar’ one of the most popular tags used to refer to the ‘automobile’ word, mentioned before and called disambiguation tags.

Search tag set	**Number of Web resources found**
jaguar car	217
jaguar cars	183
jaguar auto	75
jaguar automotive	37

Table 1. Single disambiguation tag

In this way, we obtain a first refinement of the search results and an increase in precision. Now we better analyze the structure of these four sets of resources so as to understand their level of overlapping and the size of their intersections. To do this, we examine the number of Web resources tagged with all possible combinations of the four disambiguation tags ‘car’, ‘cars’, ‘auto’ and ‘automotive’.

In what follows, we suppose that the total number of resources relevant to our search (all resources dealing with Jaguar cars) is equal to the number of different resources identified by the four search results showed in Table 1. Thus, we can determine the supposed total number of relevant resources present in the tagging system applying the ‘Inclusion-exclusion principle’ [Mat]. It is used to compute the cardinality of a set composed by the union of other finite sets, through the cardinality of their intersections. We suppose to have n finite sets: A_1, A_2, A_3, …, A_n, we can compute the cardinality of the n sets’ union using the following formula:

[1] All the numeric data reported in the examples included in this chapter are obtained querying del.icio.us (http://del.icio.us) in data 3.1.2007.

$$\sum_{i=1}^{n}((-1)^{i+1}\sum_{0\leq j_1\leq..\leq j_n}\bigcap_{k=1}^{i}(A_{j_k}))$$

Applying the ‘Inclusion-exclusion principle’ to compute the total number of relevant resources we obtain the result of 324.

Starting from the previous analysis of the tag space, we can notice that when we use only one disambiguation tag among the most popular tags used to refer to the automobile word, we obtain the results represented in Table 2 in terms of recall.

Search tag set	**Recall**
jaguar car	217/324 = 0.6697 - **67%**
jaguar cars	183/324 = 0.5648 - **56%**
jaguar auto	75/324 = 0.2315 - **23%**
jaguar automotive	37/324 = 0.1142 - **11%**

Table 2. Recall with one disambiguation tag

Only a relatively small part of all relevant resources present in the system is selected and showed as search result to the user; this fraction ranges between 11% and 67%, depending on the popularity of the disambiguation tag added to the tag ‘jaguar’ to form the search tag set.

2.2 Synonymy

Besides polysemy, another search limitation in the existing tagging systems is due to synonymy ([GH05], [Mat04]) that is the presence of different tags/words having the same meaning. For example, we can refer to a computer using tags like ‘computer’ or ‘pc’ or to automobiles using tags like ‘car’, ‘auto’, ‘automobile’, etc. When we search all Web resources dealing with computer, we chose a tag so as to identify this concept, excluding from the result all the relevant resources tagged using its synonyms; as a consequence we must face the following situation graphically represented by a Venn diagram in Figure 4.

Only 4188 Web resources, compared to a total number equal to 343126 (1.2%), have been tagged using both tags, maybe by expert users in order to relieve the search limitations caused by synonymy. But we can’t rely on the users’ tagging behaviour, wishing that the user will add all possible synonyms of every word used when he tags a resource; it could also represent a great limit to the directness of the process of tagging.

Many times the limitations caused by synonymy are strictly related to those caused by polysemy. In fact, it is possible that a word or tag has more than one meaning, but it presents also many synonyms. In such a situation the search

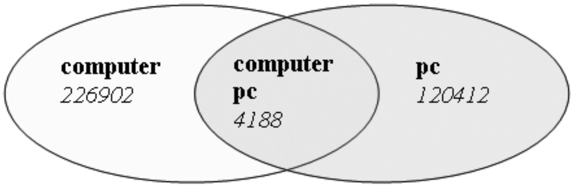


Fig. 4. The relevant Web resource partition due to different synonyms of the word computer.

is complicated by the presence of both polysemy and synonymy. Due to polysemy, the user will need to add other tags in order to disambiguate those already chosen together with all problems related in terms of recall and precision. Furthermore, all resources tagged with tags synonym to the one chosen and therefore potentially relevant, will not be included among the search results.

2.3 Different lexical forms

Problems similar to those previously described may often arise as a consequence of the use of different lexical forms ([GT06]) to refer to the same concept.

Plural nouns, different verb conjugation and name-adjective couples. Referring to the example previously analyzed when speaking about polysemy, the tags 'car' and 'cars', both used to indicate the concept of automobile, represent the *singular and the plural form* of the same word, but are managed as different entities. In fact, if we consider a search tag set composed by the tags 'jaguar' and 'car', the system will select 217 resources as search results, omitting those tagged with 'jaguar' and 'cars' but without the tag 'car' and thus preventing the user to access to other 65 relevant Web resources.

We can observe those problems also in the e-commerce Web sites' tagging. E-commerce web sites are often tagged with 'buy' or 'buying' (*the gerundive form of buy*). When the user wants to find all the Web sites that sell scooters, depending on the search tag set used, 'scooter buy' or 'scooter buying', he will retrieve different sets of results (without considering the problems caused by the presence of polysemy and synonymy regarding the tags considered).

A similar case of search precision loss happens when a document describing different kinds of sources of energy is tagged by different users respectively with the keywords 'energy' and 'energetic'; both tags express very similar meaning but *the former is a noun and the latter the respective adjective*. When a user asks for all the resources speaking about 'energy', only the first set of the two ones will be showed as search result (obviously including their intersection). The same problem may arise with couples of very similar keywords like: 'pollute' - 'pollution', 'dance' - 'dancing', etc.

Multi-word tags. In many tagging system there are a lot of tags composed by more than a single word: *multi-word tags* ('semantic web', 'personal computer', 'web design', etc.). When a user tags a Web resource, if he divides the words that constitute the tag using blank characters, the system will consider all those words as different tags and not as a single lexical form referring to the same concept.

Moreover, to overcome this problem, the users of tagging systems usually adopt different solutions and thus different lexical form to refer to the same concept, causing a search space partitioning. For instance, we want to retrieve all resources speaking about the Semantic Web. Some of these are tagged with two separate tags: 'semantic' and 'web'. They will be included in every search even if only one of these two tags constitutes the search tag set. Other alternative tags which refer to the same concept are: 'semWeb', 'semanticWeb', etc. It is also possible that a single user or a community of users defines a new tag to refer to the Semantic Web, for example 'sWeb'. When we search for Semantic Web related resources, if we type the tag 'semWeb', we will identify 5387 resources, omitting other 9435 results tagged with 'semanticWeb' and not with 'semWeb' (represented in the Venn diagram shown in Figure 5).

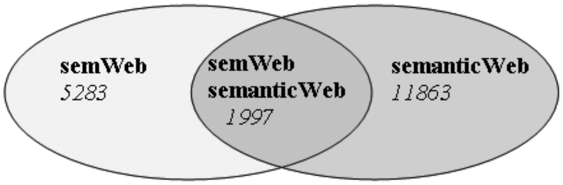


Fig. 5. The relevant Web resource partition due to different multiword tags referring to the Semantic Web.

One aspect of current tagging system, related to multi-words tags, is represented by *the different notations* which could be adopted by users. For instance, when a user chooses to collapse a tag composed by different words he could use the CamelCase notation (removing all blank spaces and starting every word besides the first one using an upper-case character) or he could replace the blank characters with other characters like underscore, slash, dot, etc. Also in this case their effect is a bad tag space partitioning.

2.4 Misspelling errors or other spelling

Also misspelling errors or other spelling ([GT06]) represent a possible source of search imprecision; indeed, when a user makes a mistake typing a tag, he isolates the selected resource decreasing the possibility of a future retrieval, especially if the resource considered isn't so popular. He can *misspell a tag* related to a document which describes Condoleezza Rice using the keyword 'Condoleeza' (only one 'z'), making these resources isolated from the others. Moreover, a user could write the word/tag 'colour' in a slightly different way, adopting *a different spelling*: 'color'. They both identify the same concept. As shown in Table 3 in del.icio.us there are 72949 web resources tagged with at least one tag among 'color' and 'colour'; only 2698 of them belong to their intersection and thus are tagged with 'color' and 'colour'.

Search tag set	**Number of Web resources found**
color	61446
colour	14101
color colour	2698

Table 3. Different spelling of the same word

Different forms of spelling could be present especially when we refer to *proper names*, *acronyms* or *word punctuation*, for example 'Al-Jazeera' and 'AlJazeera' represent different tags which refer to the same concept as a consequence of different spelling behaviours. We could also include in this broad category those problems created by *different rules of capitalisation*.

2.5 Different levels of precision

Another problem which may occur during the tagging process is related to the different level of precision that could be adopted while a user chooses a keyword to describe and characterize a resource; this question is also referred as "the basic level" problem ([GH05]). Let suppose that a user wants to tag a Web page about a new musical record review. Depending on the different level of experience and musical knowledge and on the aim and the accuracy of his tagging behaviour, he could choose a general keyword like 'music' or a more specific one, e.g. 'jazz'. A tag-space based search regarding all resources tagged with the tag 'music' will not find those tagged with 'jazz', lowering the recall. Similarly a user could tag a document which describes Java programming language with 'programming' (general) or 'Java' (more specific). Different users could be characterized by different levels of precision and every user usually uses a personal basic level of precision while tagging; the level of specificity of keywords is influenced by the aim of tagging but also by the knowledge and the expertise of the user.

2.6 Different kinds of tag-to-resource association

Analysing the keywords used in existing tagging systems and the tagging behaviour of users, we can define different implicit kinds of relations that links a tag to a specific resource. Indeed the association of a keyword or tag to a resource is made without specifying the relation. Many times the tag represent the topic of the resource, other times it is a sort of rating about the resource or a simple and often personal memo information.

Through a tag analysis, it is possible to group tags in distinct sets on the basis of the kind of relation which associates the tag to a particular resource. Several research papers have proposed possible categorizations of the kinds of tags. In [GH05] seven different groups of tags are identified analysing del.icio.us; in [CM06] those seven sets are grouped together in three main categories: tags which identifies properties of the resource referred, tags which describe the content of a resource in terms of its relation to the tagger and tags which collect information about a particular task in which the tagger is involved. Also [ZXS06] proposes five groups of tags' types: *content based tags* ('computer', 'AMD', 'programming', etc.); *context based tags* (for example those tags which describe location and time to specify the context in which the resource was created or saved, e.g. 'Rome', '10-10-2001'); *attribute tags* (which describe a resource but could not be directly derived from its content, e.g. 'Jimmy's blog', 'post' etc.); *subjective tags* ('interesting', 'funny', etc.); *organizational tags* ('mywork', 'toread', etc.).

Another relevant problem briefly mentioned before is represented by *the presence of tags, or better of lexical forms expressed in different languages*. A tagging system is used by communities of people from different countries and even if English represents the mainly adopted language over the Web, the multilanguage support is an important issue in a global Web system like this one. From the analysis made in [GT06], we observe that about 64% of the set of tags present in del.icio.us are valid English language dictionary words and about 32% of the same set is unrelated to a particular language (proper names, acronyms, etc.). Therefore, more than 95% of the total tag set of del.icio.us is expressed using the English language or at least consistent to it. Besides the English, on del.icio.us other used languages are the Spanish, the German, the French and the Portuguese.

3 Semantic collaborative tagging

Examining the different causes of inconstancies and loss of precision in tag-space based searches, we can infer that all of them may be solved or substantially reduced bringing semantics in collaborative tagging systems. Each tag should not represent just a simple sequence of characters, but should be defined specifying its meaning. When a user decides to tag a resource, describing it through one or more keywords, he must be able to disambiguate each of them, defining their semantics or better pointing out their contextualized meaning. Moreover, we added properties to links concepts to a specific resources; this process will be referred to as *semantic collaborative tagging*. In this way, the outcome of semantic tagging activity consists of *a set of unambiguous assertions on resources*. Each of them could represent statements about topic or kind of resource or concern the personal opinion of the user about the resources.

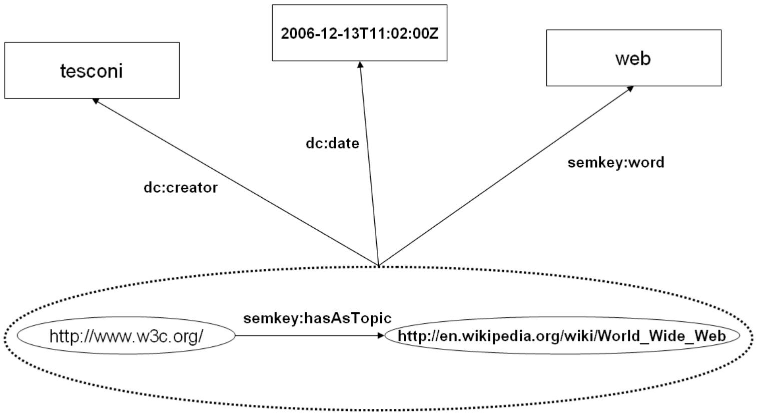


Fig. 6. Example of RDF triple

These assertions represent the classical RDF triples that are composed by the following parts:

- **Subject:** the URL of the web resource.
- **Property:** the URI that identifies the relationships between a resource and a concept.
- **Object:** the URI that identifies of the concept associated to the web resource.

Other data need to be added to those just mentioned in order to fully describe the association of a concept to a particular Web resource:

- the *lexical form* (string) employed by the user to identify the particular concept referred in the moment of the generation of the semantic tag;
- the *username* adopted in our system in order to uniquely identify the user, author of the semantic annotation;
- the *date* and the *time* of generation of the semantic annotation.

These added data are all descriptive information that is added to the core RDF-triple previously mentioned. To represent it we need to exploit another RDF expressive conventionalism: the reification [Fut06]. It is used to make RDF statements that describe an entire RDF triple; to do this we need to univocally refer to the RDF-triple that must be described assigning it an identifier. It could consist of an URI, which is unambiguous over the Web or of a blank node identifier which is unambiguous inside the local RDF document that contains it. This ID is formally assigned to the RDF-triple using other four additional RDF-triples in order to respectively specify its subject, its predicate, its object (*rdfs:subject*, *rdfs:predicate*, *rdfs:object*) and the class of belonging (*rdfs:statement*). Once determined this ID, we can define other properties referred to the entire RDF-triple, in particular:

- **semkey:word** : the lexical form used to refer the concept during the semantic tagging process;
- **dc:date** : date and time of generation of the semantic annotation;
- **dc:creator** : username of the user which has generated semantic annotated data.

In the properties just described, the namespace 'semkey' refers to the local RDF Schema namespace of our semantic tagging reference and the namespace 'dc' refers to the Dublin Core Metadata RDF Schema namespace [Dub]. As a consequence the set of information related to the association of a disambiguated tag to a Web resource made by a particular user in a precise moment is represented as showed in Figure 6. The main RDF-triple is represented by the information contained in the dotted circle; through the RDF reification conventionalism three other descriptive data are added to the main RDF-triple. If we want to represent those data using the XML/RDF serialization, we obtain the following schema:

```
<?xml version='1.0'?>
<rdf:RDF xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
xmlns:semkey='http://www.semkey.org/schema/'
xmlns:dc='http://purl.org/dc/elements/1.1/'>

<!-- Triple 1 -->
<rdf:Description rdf:about='http://www.w3.org/'>
<semkey:hasAsTopic rdf:nodeID='id00001'
rdf:resource='http://it.wikipedia.org/wiki/World_Wide_Web'/>
</rdf:Description>

<!-- Descriptive information added to triple 1 exploiting its
reification -->
<rdf:Description rdf:nodeID='id00001'>
<semkey:word>web</semkey:word>
<dc:date>2006-12-13T11:02:00Z</dc:date>
<dc:creator rdf:resource='http://www.semkey.org/users/tesconi'/>
</rdf:Description>

</rdf:RDF>
```

Adding semantic information to tagging process, search efficiency and effectiveness will be considerably improved and new important information access and organization patterns will be exploitable. In order to make it possible a sort of shared ontology should be available and concepts expressed should be easily and univocally referenced. In Figure 7 we graphically schematize a possible scenario in which a shared ontology is used to semantically tag three resources referring to specific concepts.

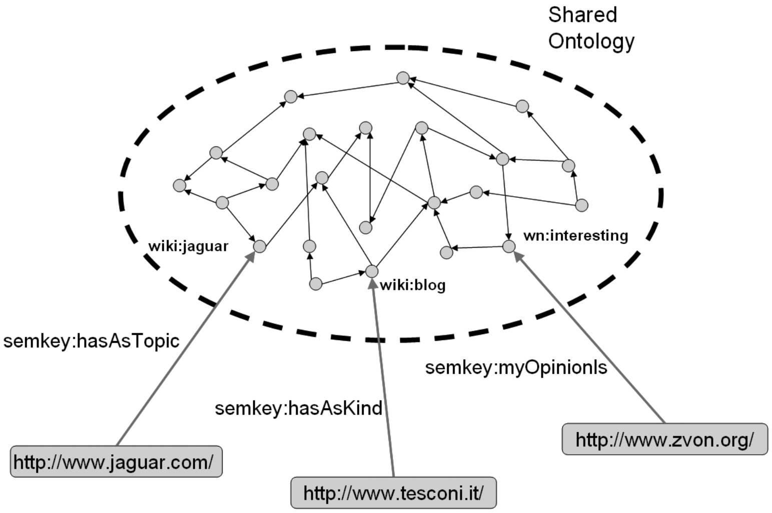


Fig. 7. Example of semantic tagging: resources that references a shared ontology

Such a kind of global generic-domain organization of concepts (the shared ontology) should be provided in a way which doesn't represent an obstacle to the usability of the system and to its adoption by a great number of users. At present, especially in specific and limited domain of interest (e.g. academic research, corporate knowledge management, medical classification etc.), tagging systems are supported by *controlled vocabularies* [MM06]; they consist of a set of terms structured and interconnected by relations so as to organize them in order to consider also their semantic relations. Using controlled vocabularies while tagging, we can manage more easily synonymy, polysemy, misspelling, plural words and other different lexical forms of a concept. They are adopted as a reference to define tags for a resource, mainly in a library or text cataloguing context. They are structured and brought up to date by experts of a specific domain. They also usually arrange their content in a hierarchical or taxonomical manner from more general to more specific concepts. Nevertheless controlled vocabularies present a rigid structure and are too strongly domain dependent to represent a valid support to the definition of the semantics of a generic tagging system [Ros01]. Usually the language used by a community of taggers changes continuously reflecting its social behaviour and it shouldn't be forced to adapt itself to the rigid constraints of structure of a controlled vocabulary or also to its limited set of terms. Moreover keeping up to date a controlled vocabulary represents an expensive task because involves domain experts and knowledge engineers; in this process few people define a structure of information used by many more users, which cannot directly take part to its definition [MH06].

Considering existing collaborative tagging systems, all these problems are absent because they leave the users complete freedom when choosing resources' keywords. This results in *the social emergence of a defined structuring of the tag space, called folksonomy*, which continuously adapts itself to the way the community of users tag resources and which is not limited by structural and organizational constraint. But when we leave such a freedom of tagging to the users all the problems and inconsistencies described in the previous chapter arise.

In order to define a semantic keyword disambiguating the meaning of a particular lexical form, we need to exploit some resource that should support the following tasks:

- starting from a particular lexical form it should identify all its possible meanings (or concepts), providing for example a short textual description in order to express each one;
- it should allow univocally referencing every single concept.

Considering these fundamental requirements, we have identified two different and may be complementary kinds of resource currently available over the Web:

- **WordNet**: a lexical database which is based on the concept of set of synonym words, called synset, which define a particular meaning; it is sufficiently structured and includes a lot of lexical and semantic relations between words and synsets. At present, WordNet version 3.0 [Wne] is available; it includes 117597 concepts (or distinct synsets).
  Wordnet [Wne] is updated by a group of lexicon explerts and presents quite a complex net of internal relations, in fact it has been developed in order to support text mining and information extraction. WordNet has a broad coverage of all parts of speech (names, verbs, adverbs and adjectives).
- **Wikipedia**: the famous collaboratively-edited free encyclopaedia, which represents the result of the efforts of many editors worldwide, directly involved in this project; it is rich of extensively described and easily referenced definitions of concepts and it is continuously increasing its dimension and completeness.

Wordnet could be used for disambiguating personal opinion (with adjectives) of user about resources. Wikipedia does not cover all parts of speech like WordNet, but it is extremely rich and constantly updated. It provides descriptions of many specific proper-named concepts that are not present in WordNet and could be useful for creating topics assertions. Wikipedia is obviously less strongly structured than WordNet, but thanks to the possibility to collaboratively edit its data, it is constantly enriched with new updated contents. It supports the disambiguation of polysemous words through the introduction of disambiguation pages which allows users choosing a specific meaning among those available. If there are many synonyms of a single word in Wikipedia is possible to use the redirect mechanism in order to redirect all of them to the same article. Moreover since May 2004 Wikipedia includes also a sort of relaxed classification system of its documents: the Wikipedia categories. Every description included in the encyclopedia can be assigned to one or more categories in order to provide a new way of accessing and cataloguing it. Users can create new categories arranging them in a hierarchical-like structure. Every document is also related to many other documents through simple links usually used to point to extended descriptions of terms. In Table 4 we mention some important numerical data [Wikb] regarding the English version of Wikipedia in order to better quantify the great amount of information collected. To obtain additional information see [Wika].

Number of articles included	1,4 Millions
Number of active editors (who edited at least 10 times since they arrived)	150.000
Number of links between Wikipedia articles	32,1 Millions
Number of redirects	1,4 Millions
Number of categories	176.000
Percentage of categorized articles	86%

Table 4. Wikipedia statistics - English - October, 2006

The exploitation of Wikipedia contents and relations so as to build a collaboratively edited collection of concepts and relations between them represents an important opportunity [DMW06] [Vos06] [MH06]. It is possible to extract from Wikipedia a relaxed controlled vocabulary or thesaurus-like structure which can be used to support the semantic tagging activity. The flexible and less strict set of semantic connections that characterizes thesauri, usually represented by equivalence, hierarchy and related-concepts relations, compared with rigid classification systems, can provide the right degree of structuring for such a collaboratively maintained resource and Wikipedia constitutes a considerable collection of data and relations that may be exploited to bootstrap it. It is also relevant, but may be initially too difficult to organize and maintain, the possibility to extend Wikipedia with an increased sets of user defined semantic relations; this is an attempt to fully import Semantic Web vision in this socially edited resource [MK06] [Pla].

At present, an interesting project that is attempting to build a sort of social ontology which may be useful to support the disambuguation of concepts and their unambiguous reference during semantic tagging activity is represented by OmegaWiki [Ome]. It is a free, multilingual resource with lexicological, terminological and thesaurus information. It is substantially a collection of concepts; each of them is characterized by a short description and one or more strings that refers to it. All these data associated to a concept can be provided adopting different languages and some simple type of relation between concepts can be estabilished. OmegaWiki is still in an intial phase of development; it aims to be collaboratively-edited: it relies on social editing efforts as its fundamental growing factor. Currently users still has read-only access and only a group of tester has the possibility to modify OmegaWiki contents; anyway its collection of data is considerably growing.

4 SemKey: s semantic collaborative tagging system

We present SemKey, a tool that allows users to tag in a semantic context, providing an evaluation of the system proposed in comparison with classical tagging tools. SemKey provides all the facilities needed to experiment semantic tagging and its possible advantages over current social tagging systems.

4.1 Requirements and global architecture

In this section we describe the main architectural chooses faced when structuring the proposed social semantic tagging system: SemKey. First of all we identify and specify in more detail its requirements, its desired features. Then we examine the global architecture and the more relevant organizational issues describing and justifying the decisions made, but also mentioning ideas about possible future relevant improvements.

General requirements The main idea that supports our system, which is also its main requirement, is the following: *giving the user the possibility to express semantic assertion about a Web resource.*

Usability. One of the most important features that have supported the wide diffusion of current collaborative tagging systems is the directness of the process of tagging; every user can immediately tag a Web resource using one or more keywords. We have considered the importance of this aspect trying *not to excessively increase the cognitive weight of the semantic tagging process.* It is obvious that we need a greater amount of information to be provided by a user in order to specify the intended meaning of a tag, but we have paid attention to organize and graphically arrange the interactions in order to make the semantic tagging process as fast as possible.

Motivation. Moreover we must consider *user's motivation to produce semantica assertions.* A critical aspect related to the diffusion of our system is represented by the possibility of experimenting concrete advantages in information organization and accessibility, when adopting this new way of tagging. For this reason we have laid great emphasis also on the completeness and the availability of added value information organization and search features.

Main user interaction patterns Starting from the system requirements just analyzed, we can define the structure of the principal interactions between our semantic tagging system and its typical users.

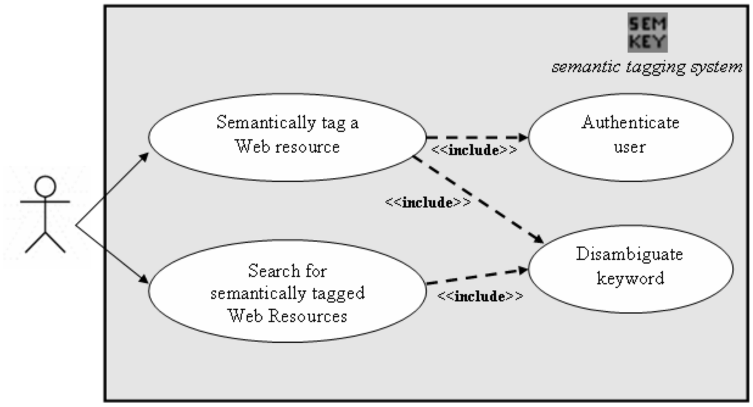


Fig. 8. System UML use case diagram.

First of all a generic user can access our system from two fundamental different perspectives. He may use the system only *as a search engine performing a semantic assertion search* in order to find relevant Web resources' references; this is a passive exploitation of the semantic information collected by the system, meaning that the user accesses to the contents already present in the system without giving any contribution to its enrichment. In this way the user takes advantage of only a part of all the possibilities offered by our system, not exploiting one of its fundamental aspects: the social component. On the other end, the user, after having completed a registration phase, may *authenticate himself and access to his personal area.* Thus he can exploit all the possibilities of SemKey. He can semantically tag Web resources of interest producing semantic assertions, manage his collection of resources and semantic assertions and organize them. Moreover, through his tagging activity, every user gives his contribution to the enrichment of the informative data collected by the system, thus making searches possibly more effective. We can graphically schematize the described user-system typical interactions through the UML use case diagram in Figure 8.

Global architecture of the system Our system architecture is based on *three main modules*: two server-side resident components and a client side one. The main functionalities provided by each module are:

- **Semantic tagging manager** (client side): this module is intended to be strictly integrated in user browsers so as to allow a fast process of semantic tagging in order not to alter the usual Web browsing activity of a common user. In this way while using our semantic tagging system, we aim not to introduce any change in the diffused browsing interaction patterns;
- **Sense disambiguation module** (server side): this module provides access to all information and services needed during the lexical form disambiguation process; it mainly supports the client in the choice of the intended concept described by a lexical form collecting the different meanings and thus allowing the definition of a semantic assertion;
- **Metadata store and access module** (server side): this is the principal module of our system. It mainly stores and provides Web access to all collected semantic tagging information. It is also responsible of the users' management.

In Figure 9 we represent SemKey high-level modules just described.

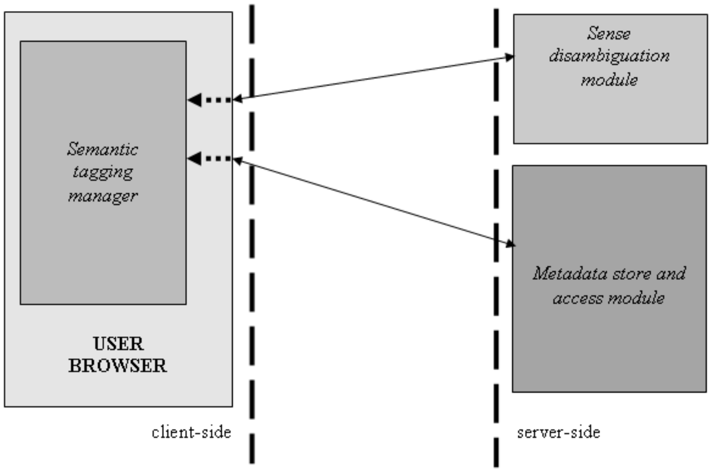


Fig. 9. SemKey high-level modules.

When a user browses the Web and visits a resource of interest, he can decide to semantically tag it. He activates the 'Semantic tagging manager' that retrieves the URL of the resource and eventually a tag (or lexical form) selected by the same user inside the browsed resource (1). If the user isn't still logged in SemKey, logging credentials are required in order to identify him; they are validated interacting with the 'Metadata store and access module' (2). After the authentication phase is successfully completed, the user will be driven in the choice of the intended meaning of the selected tag. Interacting with the tagging Web APIs of del.icio.us [del] and Yahoo My Web 2.0 [MyW] the 'Semantic tagging manager' retrieves and shows the user the most popular tags concerning the selected resource, in order to provide possible suggestions (3). Once the user has chosen a tag, it will be sent to the 'Sense disambiguation module' in order to receive a list of all possible concepts that can be referred using that tag (4). The user selects the intended meaning of his tag and the specific property of the Web resources to describe: thus he formulates a semantic assertion. It is sent to the 'Metadata store and access modile' to be stored (5). Theen the 'Semantic tagging manager' ends its execution and the user can prosecute his browsing (6).

Main organizational issues In this section we analyze the basic organizational issues faced when structuring the modules which constitute our system. We discuss and motivate every adopted solution, but we also describe possible improvements and future scenarios.

The structure of Sense disambiguation module: WordNet and Wikipedia exploitation The 'Sense disambiguation module' represents the core of our system; it is devoted to support the client-side 'Semantic tagging manger' during the process of semantic tagging and it is responsible for the collection of available meanings of user tags. As stated before, in this initial version of SemKey we have decided to explore the semantic content of WordNet [Wne], and Wikipedia [Wika]. During the disambiguation process, the 'Sense disambiguation module' accesses the Web interfaces of WordNet and Wikipedia to collect the meanings of the typed tag. In particular, given a tag we consider:

- **in WordNet**, all synsets which the tag belongs to;
- **in Wikipedia**, the description of the meaning of the tag or the different meanings associated to a polysemyc tag through its disambiguation page.

The 'Sense disambiguation module' selects for every concept two information:

- an **URI** which identify the concept;
- a short textual **description** or gloss of the concept.

Compared to WordNet, Wikipedia contents is often more difficult to manage to disambiguate a tag because of its relaxed organizational structure that doesn't provide many facilities to support this task.

Semantic assertion model Another relevant issue is represented by the organization of the set of data stored during the semantic tagging of a Web resource. In a generic collaborative tagging system as del.icio.us an user can associate a tag to a resource without specifying the relation type. Normally the tag represents the topic of the resource but this is not always true and this semantic information

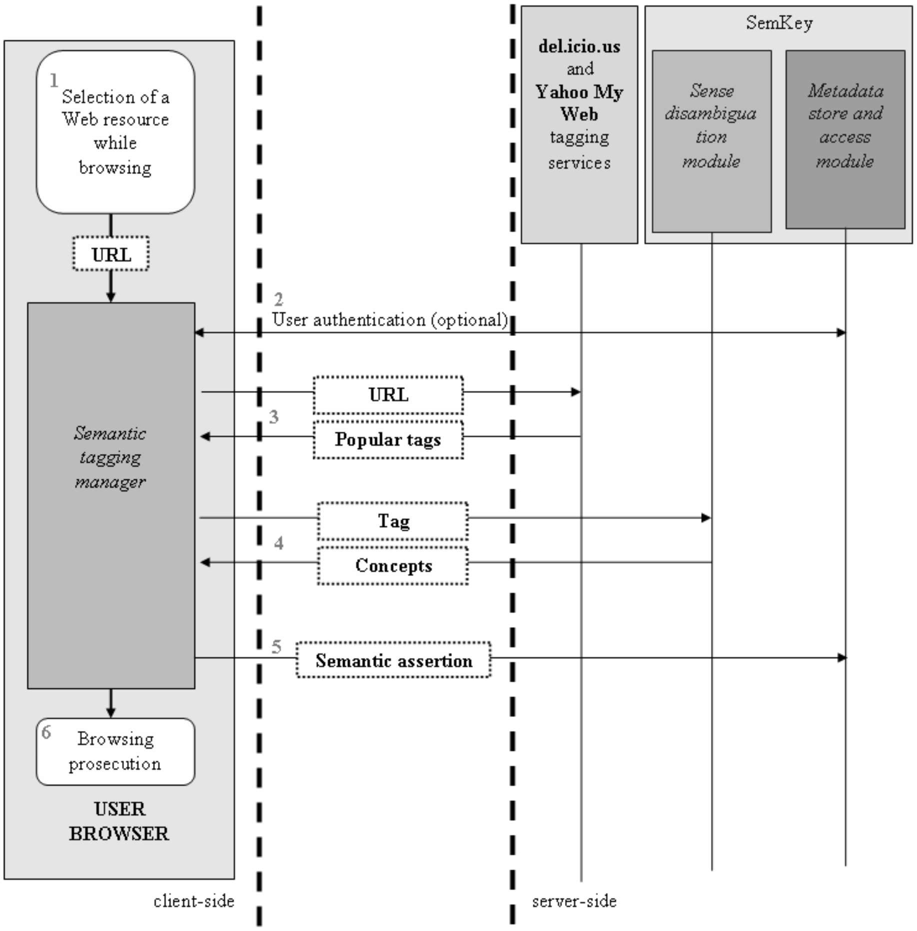

1
Selection of a Web resource while browsing
URL
Semantic tagging manager
6
Browsing prosecution
USER BROWSER
client-side
del.icio.us and Yahoo My Web tagging services
SemKey
Sense disambigua tion module
Metadata store and access module
2
User authentication (optional)
URL
3
Popular tags
Tag
4
Concepts
5
Semantic assertion
server-side

Fig. 10. Modules interaction to produce a semantic assetion.

remain in the head of the user. The solution could be to force the user to explicate the kind of relation for each tag, but this can make boring the semantic tagging activity, so we adopted an intermediary solution.

Starting from the analysis of the different kinds of tags managed by existing collaborative tagging systems, we have decided to manage only *three different relations*:

1. **hasAsTopic** : this relation will be used to describe the topic of the resource such as book, Web design, sport, politics, cars, animal, medicine, etc.;
2. **hasAsKind** : this relation will be used to characterize the kind of informative content of the resource such as blog, application, mashup, podcast, official Web site, streaming, video, e-commerce, Web API, etc.;
3. **myOpinionIs** : this relation concerns all subjective opinions such as cool, funny, interesting, boring, amazing, expensive, boring, etc..

The choice of the right relation to connect a concept to a particular resource is left to the user. In this way the model of a semantic assertion is a particular type of RDF triple (see 6)

Semantic search patterns When a user searches for relevant resources, he must specify the structure of one or more *generic semantic assertions*; they are semantic assertions defined without referring to a particular resource. Each of them specify a concept that describes a particular characteristic (or property) of the resource to find. All the resources that are described by the set of generic semantic assertions specified by the user are considered to form search results.

For instance, the user could ask the system to find all 'blogs' (property: kind of resource) which deal with 'Web design' (property: topic of resource) and are reputed to be 'interesting' (property: personal opinion); 'blog', 'Web design' and 'interesting' are disambiguated lexical forms, referring to specific concepts. The search parameters just described are composed by three generic semantic assertions. The user could specify none, one or more semantic assertions.

Exploitation of WordNet and Wikipedia net of relations. Those just described represent only the basic search capabilities and content structuring possibilities that such a kind of system offers. A possible relevant improvement could be obtained considering all the nets of relations that could connect the concepts used to support the disambiguation of lexical forms or could relate two or more different tag lexical forms. This further part of informative content is usually strongly present in lexical resources.

In this first development phase of our system, we have decided to exploit the sense disambiguation information provided by the lexical resource *WordNet* [Wne]. In WordNet, the meanings and the lexical forms used to refer to a particular concept are connected by a set of 18 different kinds of relations. Some of these are very specific and have been introduced in order to exploit the semantic information available with an original distinct purpose: text mining and information extraction. However, other relations could be exploited to further enrich search possibilities and structured exploration of contents. For example, *the hyponymy/hypernymy relations* that represent the hierarchical specialization / generalization of concepts may be used to suggest, during the disambiguation of lexical forms, all their hyponyms or hypernyms in order to better define the level of precision adopted by the user; in this way we can solve or at least reduce the basic level of precision problem, mentioned before. Moreover we can allow users to extend the coverage of their search including all the hyponym concepts of those related to particular chosen concept; in a similar way we can suggest users to choose one of the hyponyms of a disambiguated tag to better specify the search parameters. He can also substitute a disambiguated tag with one of its hypernyms in order to eventually increase search coverage.

For instance, if a user wants to find all the resources tagged with 'automobile', after the choice of the intended meaning for this word, he could the examine all the concepts hyponyms of this concept: 'jeep', 'coupe', 'station wagon', etc. so as to extend the search coverage including all resources tagged with a least one of the hyponyms or to further refine his search replacing, for example, 'car' with 'jeep' and increasing the level of precision adopted. Part of the considered WordNet subsumption hierarchy of concepts is schematized in Figure 11.

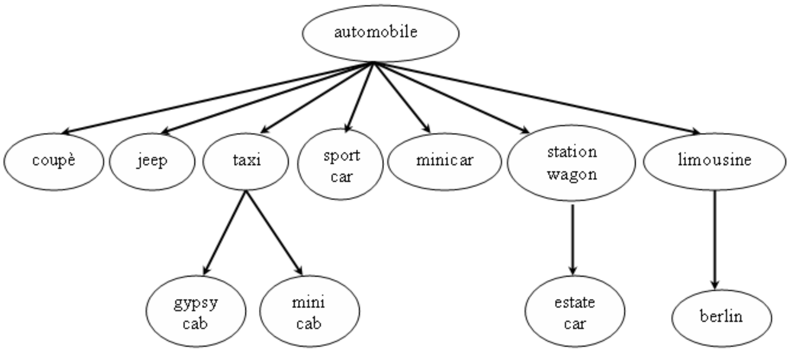


Fig. 11. Part of WordNet hierarchy of concepts referred to the automobile world.

Another exploitable WordNet's relation is *the meronym or 'part of' relation.* It connects a concept with other concepts which constitute its parts. For example, the disambiguated tag 'automobile' has the following parts or meronyms: 'accelerator', 'air bag', 'auto engine', etc. It could be useful to show all meronyms of a concept in order to help users to better structure and organize their search tag set.

When we analyze *Wikipedia* [Wika] and the semantic concept references extracted by this other resource, we should consider that it is not a coherent lexical resource, but a collaboratively edited encyclopedia. Also Wikipedia presents a sort of content categorization system: *the Wikipedia categories.* They are collaboratively edited and managed and don't constitute a hierarchical structure; they form a direct acyclic graph. Every category could be included in one or more general ones, and sometimes there are also cyclic inclusions, even if editors are explicitly advertised to avoid such a situation. All those categories constitute a sort of specialization / generalization structure similar to that previously described speaking about WordNet, with more relaxed constraints. We can consequently exploit this added informative content in a way similar to that described above, in order to improve search completeness. Besides the category structure, in Wikipedia there is a highly dense net of simple inter-document references and every concept description or encyclopaedia entry presents a collection of related Web resources which could be exploited to provide the user with useful links' suggestions in order to deeply examine a concept.

4.2 Detailed system modules' architecture

Considering the main high-level modules of our system, their interactions and the fundamental organizational issues faced when specifying their architectural structure, in Figure 12 we detail the internal organization of each of them.

The 'Semantic tagging manager', realized as a browser extension, can directly interact with del.icio.us [del] and Yahoo My Web 2.0 [MyW] Web APIs to retrieve poular tags' suggestions.

The 'Sense disambiguation module' can be accessed directly from the Web browser when the user must single out a specific concept considering a particular tag, in order to support SemKey search functionalities; this module can be also queried by the 'Semantic tagging manager' during the formultion of a semantic assertion in order to point out a particular concept.

The 'Metadata store and access module' can be accessed by the 'Semantic tagging manager' in order to save one or more semantic assertions, by a request to SemKey Web APIs or by the user browser in order to execute semantic searches or to manage the personal data of every user of our system.

4.3 Implementation and functioning

We describe some implementation details with some examples of SemKey in action, considering each one of its three main modules. For try a first version of this tool see the site http://www.semkey.org.

The Semantic Tagging Manager (STM). STM is the client-side module of our system: it must provide support to make the semantic tagging process (choice of the concept starting from a lexical form and of the relation) easy to the user and to *maintain high the global usability of our system.* We have decided to realize it as a Mozilla Firefox extension [Moz].

When the plug-in is installed, a multi-coloured button is added to the user interface of the browser. It is used to add one or more semantic assertion to

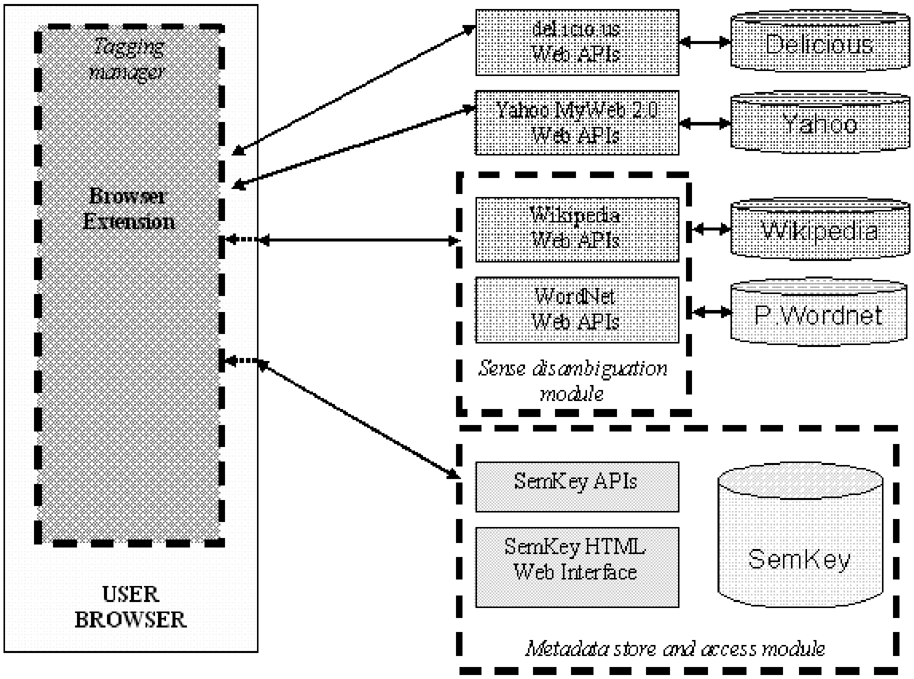


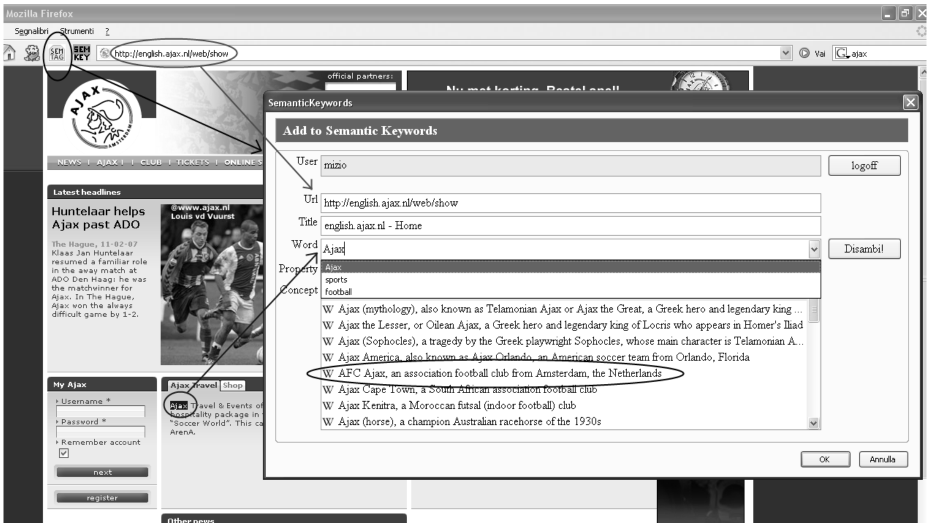
Fig. 12. Detailed system modules' architecture.

the current resource displayed on the browser by activating a dialog window as shown in Figure 13.

If the user is not still logged in the system, it requires typing logging credentials (username and password) in order to identify him (2), interacting with the 'Metadata store and access module' so as to validate them. After the log-in phase is successfully completed, the user will be driven in the composition of the semantic assertion.

In fact the STM proposes initially some tags corresponding to the most popular ones used by delicious users to annotate the current resource. The user can select one of these tags or insert a new one and the relative relation (by default it is selected the 'hasAsTopic' relation).

Immediately the STM answers with a list of available meanings. Once selected the intended meaning of the considered tag, the semantic assertion is completed and STM will save it sending all data to the "metadata store and access module".



http://english.ajax.nl/web/show

Huntelaar helps Ajax past ADO

The Hague, 11-02-07
Klaas Jan Huntelaar resumed a familiar role in the away match at ADO Den Haag: he was the matchwinner for Ajax. In The Hague, Ajax won the always difficult game by 1-2.

@www.ajax.nl
Louis vd Vuurst

Ajax Travel & Events of hospitality package in "Soccer World". This ca ArenA.

Add to Semantic Keywords

User mizio
Url http://english.ajax.nl/web/show
Title english.ajax.nl - Home
Word Ajax
Property
Concept

Ajax
sports
football

W Ajax (mythology), also known as Telamonian Ajax or Ajax the Great, a Greek hero and legendary king ...
W Ajax the Lesser, or Oilean Ajax, a Greek hero and legendary king of Locris who appears in Homer's Iliad
W Ajax (Sophocles), a tragedy by the Greek playwright Sophocles, whose main character is Telamonian A...
W Ajax America, also known as Ajax Orlando, an American soccer team from Orlando, Florida
W AFC Ajax, an association football club from Amsterdam, the Netherlands
W Ajax Cape Town, a South African association football club
W Ajax Kenitra, a Moroccan futsal (indoor football) club
W Ajax (horse), a champion Australian racehorse of the 1930s

Fig. 13. Semantic tagging manager dialog window.

The Sense Disambiguation Module (SDM) This module has to support the process of disambiguation of the tag chosen by the users. SDM gathers the different meanings of a particular lexical form by exploiting the available concepts of WordNet and Wikipedia. To carry out this goal, the SDM filters the web pages of these two lexical resources producing a list of the collected meaning associated to the lexical form. This list is serialized in order to compose a JSON array [JSO] with all collected couples of concept URLs and respective short concept descriptions; this array is sent back as the reply to the STM.

If Wikipedia and Wordnet would provide some suitable Web APIs to access to their content, we could bypass the SDM.

The Metadata Store and Access Module (MSAM) This module provides storage functionalities to save and retrieve all semantic annotations. It is also responsible for the users' management. All these features are available through a Web based HTML interface.

User Management. In our system each user must be registered in order to be identifiable; in this way we can manage his personal data and his tagging metadata and we can support him with additional system functionalities.

Semantic Annotation. The main goal of our system is the collection of the semantic assertions produced by the semantic tagging activity. Every semantic assertion has been generated by a particular user in a precise *moment.* All these data are stored by the 'Metadata store and access module' as the outcome of semantic tagging activity.

Users oriented views. When we speak about user oriented views, we mean all the available ways that a registered user, after his authentication, can exploit to interact with the system and visualize his personal profile data and his tagging metadata. Here is a list of all those implemented in our system:

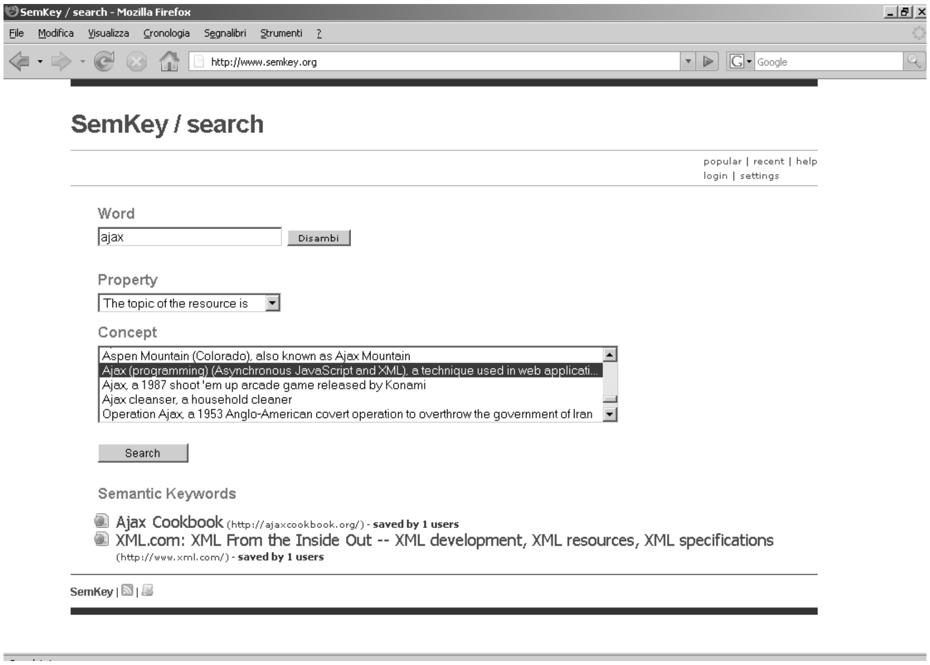
- **Visualization of user semantic tagging metadata** :
  - *my Web resources* view : all the Web resources semantically tagged by the user ordered by date, each one with all the associated semantic assertions;
  - *my semantic assertions* view : all the semantic assertions made by the user ordered by referred concept and property (every semantic keyword is a link to the next view);
  - *my Web resources tagged with* view : all the Web resources semantically tagged with one particular concept (is a list of Web resources links ordered by date).
- **Deletion of a semantic assertion from a web resource** (*delete a tag* view).
- **Visualization / partial modification of user personal profile data** (*my profile* view);

Generic search-oriented view. This section includes all available ways that a generic user, authenticated in the system or not, can exploit to execute a semantic search among the metadata collected:

- *basic search* view : search of all Web resources semantically tagged with one or more generic semantic assertion:
  - the user chooses one or more lexical forms;
  - every lexical form is disambiguated interacting with 'The sense disambiguation module' previously described and retrieving all the meanings used to refer to it; in this way the user can define the intended concept, choosing between the multiple meanings presented;

- onece a concept has been chosen, the user can select a particular property that links the Web resource he wants to find to the concept, thus defining a generic semantic assertion;
- SemKey, interacting with the 'Metadata store and access module', will retrieve all resources matching the set of generic semantic assertions previously defined.

In Figure 14 is represented an example of the *basic search* view system interface; the user has chosen the word 'ajax' and, among the list of oncepts retrieved disambiguating it, has pointed out the concept of 'AJAX (programming) (Asynchronous JavaScript and XML), a thechnique used in Web applicatins...'. Then selecting the property 'The topic of the resource is' has formed a generic semantic assertion, requesting to find all Web resources that speak about the Asynchronous JavaScript and XML Web programming technique. SemKey shows a list of all the resources semantically tagged that match the parameters previously specified.


SemKey / search - Mozilla Firefox
File Modifica Visualizza Cronologia Segnalibri Strumenti ?
Google
popular | recent | help
login | settings
Disambi
Search
SemKey |

http://www.semkey.org

SemKey / search

Word: ajax

Property: The topic of the resource is

Concept:

Aspen Mountain (Colorado), also known as Ajax Mountain
Ajax (programming) (Asynchronous JavaScript and XML), a technique used in web applicati...
Ajax, a 1987 shoot 'em up arcade game released by Konami
Ajax cleanser, a household cleaner
Operation Ajax, a 1953 Anglo-American covert operation to overthrow the government of Iran

Semantic Keywords

Ajax Cookbook (http://ajaxcookbook.org/) - **saved by 1 users**
XML.com: XML From the Inside Out -- XML development, XML resources, XML specifications (http://www.xml.com/) - **saved by 1 users**

Fig. 14. Example of semantic search.

5 Conclusions

In this work we have described the architecture and the main implementation choices of SemKey a semantic collaborative tagging system.

Starting from the fundamental weak points of existing social tagging systems, we have deduced that most of them are referable to the absence of any semantic support in tagging activity. We have proposed a nwe model of tagging referred to as **semantic collaborative tagging**. The outcome of semantic tagging activity consists of producing a set of unambiguous assertions on resources. We allows users easily defining the meaning of their tags sa as to produce semantic assertions, referencing some sort of social ontology. As social ontology we have explored the adequacy of the support offered by the entries of Wikipedia and WordNet. Finally we have present SemKey, a tool that allows users to tag in a semantic context, describing his architecture and implementation.

At the basis of our idea of semantic tagging stands the availability and completeness of a global collection of concepts and lexical forms in order to point out and univocally reference semantic keywords; as said, both WordNet and Wikipedia have been used in order to test their possible support to this tasks. We have explored their main organizational features: WordNet presents a rich set of parts of speech and a strongly structured net of relations between them, but it lacks many data useful to support proper names disambiguation and it is not collaboratively edited; Wikipedia is an encyclopaedia so its content is composed mainly by a very rich set of names along with their extended descriptions. Wikipedia has strong proper names coverage; it is also continuously updated, but lacks a structured set of relations between the concepts described, even if its documents are interconnected by a huge number of links: at present, only the system of Wikipedia categories is available as an attempt to provide some sort of relaxed structure to its informative content. Besides Wikipedia and WordNet we must mention an early project OmegaWiki [Ome]; it is attempting to build a free socially-edited multilingual thesaurus; it organizes concepts and terms adopting a structure that seems adapt to support lexical forms' disambiguation and concept referencability as needed by semantic tagging. In parallel with the growing of OmegaWiki informative content, future works should be oriented to better explore its possibilities of cooperation with semantic tagging systems.

Since now we have analysed semantic tagging activity concerning a global domain. Recently, the advantages provided by tagging activity has been introduced also in enterprise networks; IBM has announced its version of an internal social bookmarking system: Dogear [DMK05]. The exploitation of SemKey in specific knowledge domains represents another important potential field of application. Indeed, our semantic tagging system can be used referring to defined collections of concepts in order to describe a particular domain of interest. We think that future works could concern the possibility to exploit our semantic tagging tool as a corporate knowledge management and organization support; it can support the organization and improvement of the accessibility to shared information like internal collections of documents or, in general, any huge amount of data which needs to be collaboratively organized. Many domain specific concepts' collections are actually available: for instance MeSH [MeS], the National Library of Medicine's controlled vocabulary thesaurus is a terminological medical reference actually widely used and that could be adopted as a specific tagging reference. The analysis of this possibilities constitutes an interesting new semantic tagging application scenario.

Summarizing, in this work we have suggested a new semantic tagging system in order to combine semantic technologies with the collaborative tagging paradigm in a way that can be highly beneficial to both areas.

References

[Bec06] Dave Beckett. Semantics through the tag. Conference presentation XTech 2006: Building Web 2.0, Amsterdam, The Netherlands, May 2006. Yahoo! Inc.

[CM06] Danah Boyd Marc Davis Cameron Marlow, Mor Naaman. Position paper, tagging, taxonomy, flickr, article, toread. Technical report, Yahoo! Research Berkeley 1950 University Avenue, Suite 200 Berkeley, CA 94704-1024 - UC Berkeley School of Information 102 South Hall Berkeley, CA 94720-4600, 2006.

[del] del.icio.us tagging system web site. http://del.icio.us/.

[DMK05] Jonathan Feinberg David Millen and Bernard Kerr. Social bookmarking in the enterprise - ibm. *ACM Queue*, vol. 3, no. 9, 2005.

[DMW06] Olena Medelyan David Milne and Ian H. Witten. Mining domain-specific thesauri from wikipedia: A case study. Technical report, Department of Computer Science, University of Waikato, 2006.

[Dub] Dublin core metadata initiative web site. http://dublincore.org/.

[Fut06] Joe Futrelle. Harvesting rdf triples. Technical report, Natioanl Center for Supercomputing Applications 1205 W. Clark St., Urbana IL 61801, US, 2006.

[GH05] Scott A. Golder and Bernardo A. Huberman. The structure of collaborative tagging systems. Technical report, Information Dynamics Lab, HP Labs, 2005.

[GT06] Marieke Guy and Emma Tonkin. Folksonomies: Tidying up tags? *D-Lib Magazine*, Volume 12 Number 1, January 2006.

[JSO] Json (javascript object notation) - web site. http://json.org/.

[Kro] Ellyssa Kroski. The hive mind: Folksonomies and user-based tagging. Blogsite.

[LA05] R. Lambiotte and M. Ausloos. Collaborative tagging as a tripartite network. Technical report, SUPRATECS, Universit de Lige,B5 Sart-Tilman, B-4000 Li'ege, Belgium, 2005.

[Mat] Wolfram mathworld - the web's most extensive mathematics resource. http://mathworld.wolfram.com/Inclusion-ExclusionPrinciple.html.

[Mat04] Adam Mathes. Folksonomies - cooperative classification and communication through shared metadata. Technical report, Computer Mediated Communication - Graduate School of Library and Information Science - University of Illinois Urbana - Champaign, 2004.

[MeS] Medical subject headings - official web site. http://www.nlm.nih.gov/mesh/meshhome.html.

[MH06] Katharina Siorpaes Martin Hepp, Daniel Bachlechner. Harvesting wiki consensus - using wikipedia entries as ontology elements. Technical report, Digital Enterprise Research Institute (DERI), University of Innsbruck - Florida Gulf Coast University, Fort Myers, FL, USA, 2006.

[MK06] Max Volkel Markus Krotzsch, Denny Vrandecic. Wikipedia and the semantic web. Technical report, Institute AIFB, Unviersity of Karlshrue, Germany, 2006.

[MM06] George Macgregor and Emma McCulloch. Collaborative tagging as a knowledge organisation and resource discovery tool. Technical report, Centre for Digital Library Research, Department of Computer & Information Sciences, University of Strathclyde, 2006.

[Moz] Mozilla developer center - plugin. http://developer.mozilla.org/en/docs/Plugins.

[MyW] Yahoo my web 2.0 apis reference. http://developer.yahoo.com/search/myweb/.

[Ome] Omegawiki web site. http://www.omegawiki.org/Main_Page.

[Pla] Platypus wiki - the semantic wiki wiki web. http://platypuswiki.sourceforge.net/.

[Ros01] Lou Rosenfeld. Folksonomies? how about metadata ecologies? Blog article, January 2001.

[She05] Rashmi Shena. A cognitive analysis of tagging (or how the lower cognitive cost of tagging makes it popular). Blogsite, September 2005.

[Smi04] Gene Smith. Folksonomy: social classification. Blog article, August 2004.

[SS02] Korth Silberschatz and Sudarshan. *Database system concepts*, chapter 22, pages 852–853. Mc Graw Hill, 2002.

[THS05] Ben Lund Tony Hammond, Timo Hannay and Joanna Scott. Social bookmarking tools (i) - a general review. *D-Lib Magazine*, Volume 11 Number 4, 2005.

[Vos06] Jacob Voss. Collaborative thesaurus tagging the wikipedia way. Technical report, Wikimedia Detushland e.V., 2006.

[Wei05] David Weinberger. Tagging and why it matters. Technical report, Harvard Berkman Center for the Internet and Society, 2005.

[Wika] English wikipedia web site. http://en.wikipedia.org/wiki/.

[Wikb] Wikipedia statistics english. http://stats.wikimedia.org/EN/TablesWikipediaEN.htm.

[Wne] Princeton wordnet web site. http://wordnet.princeton.edu/.

[ZXS06] Jianchang Mao Zhichen Xu, Yun Fu and Difu Su. Towards the semantic web: Collaborative tag suggestions. Technical report, Yahoo! Inc 2821 Mission College Blvd., Santa Clara, CA 95054, 2006.