Using WordNet to turn a folksonomy into a hierarchy of concepts

David Laniado, Davide Eynard, and Marco Colombetti

Politecnico di Milano
Dipartimento di Elettronica e Informazione
Via Ponzio 34/5, 20133 Milano, Italy
{david.laniado,eynard,colombet}@elet.polimi.it

Abstract. As the volume of information in the *read-write Web* increases rapidly, folksonomies are becoming a widely used tool to organize and categorize resources in a bottom up, flat and inclusive way. However, due to their very structure, they show some drawbacks; in particular the lack of hierarchy bears some limitations in the possibilities of searching and browsing. In this paper we investigate a new approach, based on the idea of integrating an ontology in the navigation interface of a folksonomy, and we describe an application that filters del.icio.us keywords through the WordNet hierarchy of concepts, to enrich the possibilities of navigation.

1 Introduction

As the amount of information available in the Web grows every day faster, the task of classification is getting harder, the traditional top down approach is getting inadequate [1], and the new collaborative approach of *folksonomies* is emerging [2].

In folksonomies users can associate freely chosen tags to resources and in this way they produce knowledge for the entire community. Beside their dynamism and low cost, folksonomies present many disadvantages: in particular, their lack of hierarchy limits the possibility of searching and browsing related information.

Our purpose is to enrich the possibilities of navigation in a folksonomy by adding some explicit semantics, provided by a static hierarchy of concepts, to help users orient themselves among keywords. We chose to start with del.icio.us[1], one of the most popular folksonomies for social bookmarking, and to develop an alternative tool for the suggestion of related tags, based on the WordNet hierarchy of concepts.

In this paper, after a brief description of the current related work (Section 2), we describe both the design and the implementation of our project (Section 3). In Section 4 we show some results of our tests and an evaluation of the application, then in Section 5 we conclude with a summary and a discussion of future work.

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

2 Current Work

Joshua Schachter, founder of del.icio.us, defined it as *"a way to remember in public"*; in folksonomies each user can generally explore two spaces, the one of his bookmarks and the one of everyone's bookmarks; tags can be used to filter items.

As the work of categorization is performed by users, folksonomies are democratic, scalable, current, inclusive and have a very low cost. On the other hand, the absence of an authority and of a unique coherent point of view on the domain bears several limitations: the lack of hierarchy, the absence of synonym control, the lack of both precision and recall, the possibility of *gaming* [3] [4]. While the traditional classification schemes, based on taxonomies, favor searching and browsing, folksonomies encourage another paradigm of navigation, based on *finding* and *serendipity* [5].

Despite their strong limitations, folksonomies are rapidly gaining momentum: according to Clay Shirky, *"The mass amateurization of publishing means the mass amateurization of cataloging is a forced move"*[2].

As tags are just text strings, with no explicit semantics associated, it is not trivial to organize them for presentation to the user. The most common way to show a set of tags are *tag clouds*, visual representations where each tag is displayed with a font size which is proportional to its popularity. Tag clouds, however, don't keep into account relationships among tags or their meaning.

To allow the discovery of interesting and related items many applications have introduced links to *related tags*, where relatedness is generally measured with metrics based on co-occurence data. For example in del.icio.us, when a user visits the page containing all the bookmarks tagged with a certain tag, a list of tags related to that one is shown inside a sidebar.

Flickr[3], a popular folksonomy for photo sharing, introduced *clustering* as an interesting feature to help navigation in the space of a tag. The system is able to find clusters of related keywords, so items corresponding to different contexts for that tag are grouped together.

These features are very useful but often insufficient, for different reasons. First of all, they leave the lack of hierarchy problem unsolved: they build flat spaces of tags, so there is no criterion to organize them and only a small set of items can be displayed. Furthermore, there is no explicit connection with the meaning of keywords or semantic relationships among them, that might help users to orient themselves in the tag space.

An interesting study to integrate a *top down* classification paradigm with folksonomies is presented in [6]. Some investigations about the challenge to derive ontologies from folksonomies are presented in [7] and [8].

[2] http://many.corante.com/archives/2005/01/22/folksonomies_are_a_forced_move_a_response_to_liz.php

[3] http://flickr.com/

3 Our Project

The goal of our work is to investigate the possibility of integrating an ontology in the navigation interface of a folksonomy, filtering tags through a predefined semantic hierarchy to improve the possibilities of searching and browsing. In particular we chose to improve the *related tags* panel in del.icio.us; filtering a set of related tags through WordNet noun hierarchy it is possible to display a much higher number of them, organized according to a semantic criterion. As WordNet is a semantic lexicon of English, developed to reflect the semantics of natural language and the way in which humans classify objects, the relations and categories that it contains are likely to be immediately understood by most people [9].

The first problem when trying to map tags to WordNet is the one of tags that are not recognizable as words in the lexicon, even after a stemming process, and therefore cannot be mapped. To evaluate the relevance of the excluded data we have collected a large dataset, relative to about 30,000 del.icio.us users and containing about 480,000 different tags. Studying these data we found that only about 8% of the different tags used are contained in the lexicon, but we also observed that the most popular tags have a much higher probability of belonging to WordNet. This distribution in particular follows a power-law curve, very common in the field of collaborative systems, as showed in Figure 1. Of the 20 million total tagging relations present in our dataset, about 68.1% involve words contained in WordNet. We think this data might be much increased by using local wordnets in other languages and domain ontologies to cover more specific terms.

There is then the problem of words that are recognized as belonging to the lexicon, but not as nouns: these tags too can't be mapped, as the hierarchy of WordNet is only defined on nouns. According to the distinction formulated in [10] among *factual*, *subjective* and *personal* tags, we can argue that factual tags tend to correspond to nouns, as nouns fit better to describe factual knowledge, while adjectives tend to correspond to subjective tags. Further studies about this issue can be found in [11]. From a quantitative point of view, our dataset confirms the intuition that most of the tags, and especially most of the most popular tags, are nouns. Indeed the 85% of the different tags recognized by WordNet are nouns, while of the over 20 million total tagging relations, about 64.9% involve WordNet nouns, and just about 3% involve words belonging to the lexicon without being nouns; in other words this data tells that, in our dataset, about 95% of the times that a tag belonging to WordNet is used it has almost one meaning as a noun: the power law distribution is accentuated for nouns.

The application we have developed is based on a client-server paradigm, where all the tasks relative to the processing and storing of information are left to the server and the client has only to manage the visualization of results. The system architecture is shown in Figure 2.

The server is composed of a *scraper*, that extracts the data from del.icio.us HTML pages and stores them on a database, a module for *tag disambiguation* and a core module that builds the *semantic tree* of tags related to a given one,

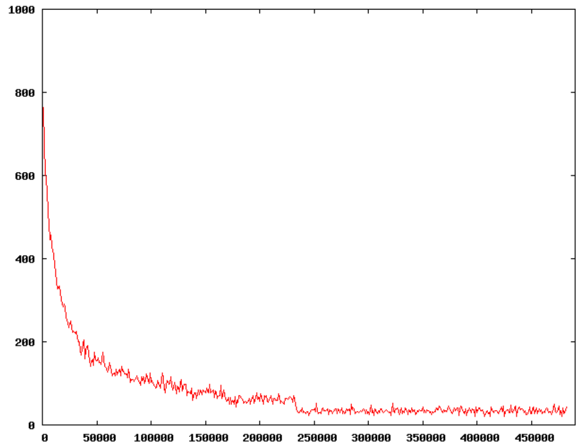


Fig. 1. The image shows the probability that a tag belongs to WordNet, in (inverse) function of its popularity. Along the X axis are represented tags from our dataset, grouped by 1000 and ordered by decreasing popularity; the Y axis shows the number of tags belonging to WordNet for each group of tags. The most popular tags are much more likely to belong to WordNet, following a power law distribution.

based on the hierarchy of concepts of WordNet. On the client side, according to the principle of *active navigation*, a JavaScript script executed inside the browser dynamically modifies the pages visualized by the user, integrating the additional information provided by the server.

3.1 Tag disambiguation

One problem when trying to map tags on an ontology is polisemy: as no explicit semantics is associated to tags by the users, the same tag can have different meanings according to different acceptation of the word, and consequently different positions in the ontology. For example the word "turkey" may refer to the country or to the animal, and in the second meaning you could want to distinguish between biological and gastronomic meaning, according to the context. In WordNet semantic relationships are not defined among words, but among *synsets*, groups of synonyms that represent units of meaning; each word can belong to different synsets according to its different acceptations. The word "turkey", for example, belongs to five synsets, where the first one is "turkey, Meleagris gallopavo" and the second is "Turkey, Republic of Turkey" .

To properly map a tag to the corresponding position in the ontology you need first to disambiguate it, in relation with the context in which it has been used. A fair solution naturally offered by a folksonomy is to use the other tags associated by some users to the same resource as the context for disambiguation.

Our algorithm for tag disambiguation acts for each tagged resource in the following way: the C most used tags for the resource are compared among them,

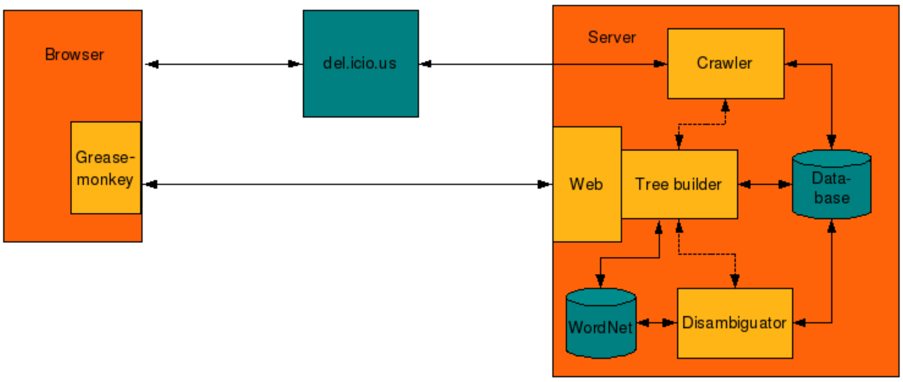


Fig. 2. The system architecture

and for each of them the meaning that is more strictly related to the other tags is selected; semantic relatedness among tags is calculated according to a choice of metrics based on WordNet [12] (adapted lesk, Hirst and St. Onge) and disambiguation is performed using the Perl library SenseRelate [13]. In the same way the remaining tags are disambiguated using the first C as a context. This solution is effective, as it reduces the sensitivity to less used tags, and efficient, as it avoids the exponential growth of the algorithm complexity with the number of different tags associated with a resource.

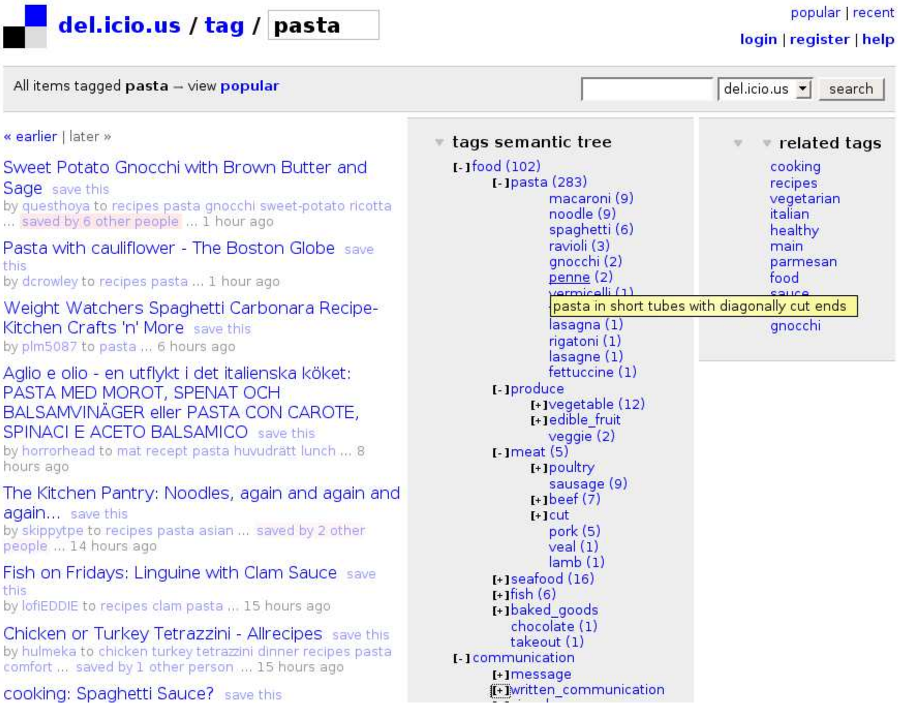
3.2 Building the tag semantic tree

The core module, for the construction of the tree of related tags, acts in four steps: *tree building*, *compression*, *branch sorting* and result output. All the algorithms developed have linear complexity with the number of input tags.

The set of tags to be considered is selected by collecting, for each of the latest N sites associated with the given tag, the M most frequent tags for that site; M and N are parameters that can be specified in the HTTP request. The construction of the tree is performed by an iterative algorithm; for each different tag present in the set of interest in a particular acceptation, the chain of the hypernyms is created as a path till the unique root of the noun hierarchy of WordNet and then merged with the existing tree. At the end of this process the tree is a subpart of WordNet noun hierarchy, chosen to contain all the tags of the set of interest.

As WordNet is very fine-grained, it can take more than 10 steps to descend from the root to a word; the tree has to be compressed to be useful for navigation, eliminating the useless nodes. The compression algorithm performs a breadth-first visit of the tree, in which all nodes considered unnecessary are deleted and replaced by their children. On one hand, all the nodes corresponding to high level categories in WordNet, contained in a black list, are deleted; the information content of these nodes is generally too low to be useful for navigation. On the other hand all the nodes that do not correspond to any tag and have a branching factor lower than K or have no siblings are replaced by their children. The default value for K is 2; in this way the structure of the hierarchy is preserved and at the same time the most specific terms can ascend in the tree.

The branches are ordered by weight, where the weight of a node is calculated as the number of resources in the set of interest that have been tagged with the corresponding word in that acceptation. This guarantees that the branches of the hierarchy that are most strictly related to the given tag are shown first to the user. As a last step, the tree is output by the server in HTML or XML format.



All items tagged **pasta** – view popular

« earlier | later »

Sweet Potato Gnocchi with Brown Butter and Sage save this
by questhoya to recipes pasta gnocchi sweet-potato ricotta ... saved by 6 other people ... 1 hour ago

Pasta with cauliflower - The Boston Globe save this
by dcrowley to recipes pasta ... 1 hour ago

Weight Watchers Spaghetti Carbonara Recipe- Kitchen Crafts 'n' More save this
by plm5087 to pasta ... 6 hours ago

Aglio e olio - en utflykt i det italienska köket: PASTA MED MOROT, SPENAT OCH BALSAMVINÄGER eller PASTA CON CAROTE, SPINACI E ACETO BALSAMICO save this
by horrorhead to mat recept pasta huvudrätt lunch ... 8 hours ago

The Kitchen Pantry: Noodles, again and again and again... save this
by skippytpe to recipes pasta asian ... saved by 2 other people ... 14 hours ago

Fish on Fridays: Linguine with Clam Sauce save this
by lofiEDDIE to recipes clam pasta ... 15 hours ago

Chicken or Turkey Tetrazzini - Allrecipes save this
by hulmeka to chicken turkey tetrazzini dinner recipes pasta comfort ... saved by 1 other person ... 15 hours ago

cooking: Spaghetti Sauce? save this

tags semantic tree

```
[-]food (102)
    [-]pasta (283)
            macaroni (9)
            noodle (9)
            spaghetti (6)
            ravioli (3)
            gnocchi (2)
            penne (2)
            vermicelli (1)
            lasagna (1)
            rigatoni (1)
            lasagne (1)
            fettuccine (1)
    [-]produce
        [+]vegetable (12)
        [+]edible_fruit
            veggie (2)
    [-]meat (5)
        [+]poultry
            sausage (9)
        [+]beef (7)
        [+]cut
            pork (5)
            veal (1)
            lamb (1)
    [+]seafood (16)
    [+]fish (6)
    [+]baked_goods
        chocolate (1)
        takeout (1)
[-]communication
    [+]message
    [+]written_communication
```

pasta in short tubes with diagonally cut ends

related tags

cooking
recipes
vegetarian
italian
healthy
main
parmesan
food
sauce
gnocchi

Fig. 3. A screenshot from the del.icio.us page for tag "pasta", where the inner sidebar shows an expandable hierarchy of related tags, provided by our application.

3.3 User interface

The system rests on Firefox Browser and Greasemonkey extension to execute some JavaScript code inside the browser . When the user is visiting the del.icio.us page for a certain tag, the script connects to our server to get the semantic tree of related keywords for that tag; as soon as the information is ready, a new sidebar is dynamically integrated in the page, showing an expandable tree. For each node of the hierarchy there are two links, one directed to the del.icio.us page for that tag and one to the page of the resources tagged both with that tag and with the given one; the size of each tag's intersection with the current keyword is shown in parenthesis and represents an indicative measure of relatedness for the users. Tooltips guide users showing WordNet definitions of the concepts corresponding to each node and indicating the destinations of links.

Figure 3 shows the result obtained for tag "pasta", where all the tags associated to the latest 300 sites tagged with "pasta" are displayed; in the picture you can see the first branches (i.e. the most related ones, in this case those about "food"), that have been expanded.

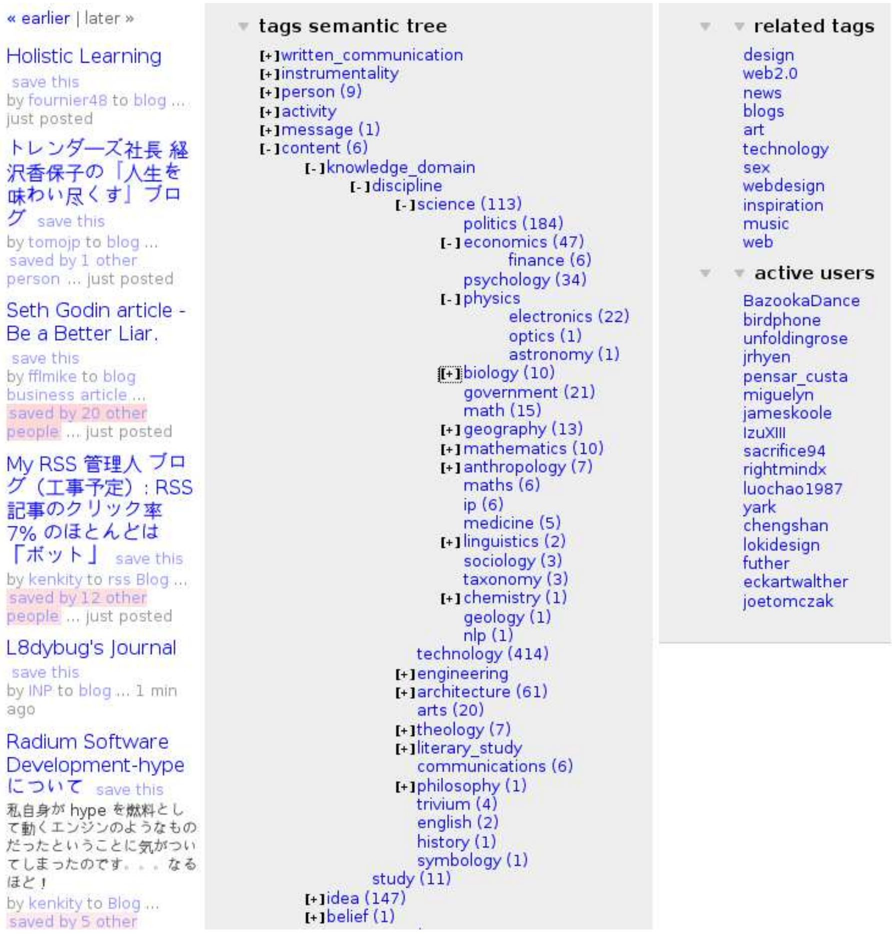
4 Tests and evaluation

We tested the system with different kinds of tags, according to different dimensions. The first dimension is the specificity of the tag from which the exploration starts; it's very different to display the space of a keyword situated in a specific domain or in a generic one. In the first case the resulting tree tends to be compact and to allow easier navigation, while in the second case it tends to have a high branching factor and a high number of first level nodes; anyway, as the branches are always ordered by weight, the most interesting concepts in relation to the given one are reachable exploring the first branches, also in case of very general keywords. The second dimension is given by the popularity of a tag, while the third one is given by the semantic field; each semantic field has its specificity and some of them rest on more conventional and ordered sets of words, such as the *food* context, visible in Figure 3, while some others are more prone to slang and neologisms, such as the one of *software*.

Figure 4 shows the result obtained for tag "blog"; as "blog" often refers to a kind of site more than to the content it can be considered a particular case, and a very general tag as there are blogs almost about everything. "Blog" is also one of the most popular tags in del.icio.us, so it is an extreme case also according to the second dimension. We obtained this result considering the latest 2000 del.icio.us bookmarks tagged "blog", and only the 15 more used tags for each of them, to cut the *long tail* of less used tags. In the picture you can see the hierarchy of scientific disciplines expanded.

According to this and other tests, the main problem for scalability seems to be the high number of nodes in the first level of the tree; some improvements could be obtained by making the tree compression algorithm more dynamic.

Comparing the related tags suggested by del.icio.us with the results we obtained, we observed that they are always somewhere in the first branches in the new sidebar. An exception must obviously be done for the words that don't belong to WordNet, that are absent in the new sidebar. Experimenting, for example, with the "Greasemonkey" tag (the experiment is possible even though



« earlier | later »

Holistic Learning
save this
by fournier48 to blog ... just posted

トレンダーズ社長 経沢香保子の「人生を味わい尽くす」ブログ save this
by tomojp to blog ... saved by 1 other person ... just posted

Seth Godin article - Be a Better Liar.
save this
by fflmike to blog business article ... saved by 20 other people ... just posted

My RSS 管理人 ブログ（工事予定）: RSS 記事のクリック率 7% のほとんどは「ボット」 save this
by kenkity to rss Blog ... saved by 12 other people ... just posted

L8dybug's Journal
save this
by INP to blog ... 1 min ago

Radium Software Development-hype について save this
私自身が hype を燃料として動くエンジンのようなものだったということに気がついてしまったのです。。。なるほど！
by kenkity to Blog ... saved by 5 other

tags semantic tree

```
[+]written_communication
[+]instrumentality
[+]person (9)
[+]activity
[+]message (1)
[-]content (6)
    [-]knowledge_domain
        [-]discipline
            [-]science (113)
                politics (184)
                [-]economics (47)
                    finance (6)
                psychology (34)
                [-]physics
                    electronics (22)
                    optics (1)
                    astronomy (1)
                [+]biology (10)
                government (21)
                math (15)
                [+]geography (13)
                [+]mathematics (10)
                [+]anthropology (7)
                maths (6)
                ip (6)
                medicine (5)
                [+]linguistics (2)
                sociology (3)
                taxonomy (3)
                [+]chemistry (1)
                geology (1)
                nlp (1)
            technology (414)
            [+]engineering
            [+]architecture (61)
            arts (20)
            [+]theology (7)
            [+]literary_study
            communications (6)
            [+]philosophy (1)
            trivium (4)
            english (2)
            history (1)
            symbology (1)
        study (11)
    [+]idea (147)
    [+]belief (1)
```

related tags

design
web2.0
news
blogs
art
technology
sex
webdesign
inspiration
music
web

active users

BazookaDance
birdphone
unfoldingrose
jrhyen
pensar_custa
miguelyn
jameskoole
IzuXIII
sacrifice94
rightmindx
luochao1987
yark
chengshan
lokidesign
futher
eckartwalther
joetomczak

Fig. 4. A screenshot from the del.icio.us page for tag "blog", where the inner sidebar shows an expandable hierarchy of related tags, provided by our application.

the word itself is not contained in the lexicon) we found that many important related tags, like "JavaScript", are not recognized, while other important words, such as "extension", are interpreted in a wrong way as WordNet doesn't contain the acceptation related to software; all the tags for which there is in WordNet an acceptation related to software have instead been correctly interpreted by the system. These limitations could be addressed by resting on some domain ontologies to integrate WordNet and on Wikipedia for reconducting slang forms to more conventional ones (for example, Wikipedia recognizes "nyc" as an alternative form for "New York City", while WordNet does not).

In many cases synonyms or just different ways of spelling a word happen to be close to each other and easily recognizable in the tree provided by the new sidebar: the semantic hierarchy helps to face the problem of the synonym control to which a folksonomy is naturally prone.

As a last consideration we want to mention the problem of gaming. It's not unusual in del.icio.us to see the related tags sidebar entirely mucked up by spam, as we found in some of our examples. Gamers can trick del.icio.us to gain a good position for the tags they want to show and, as there are just a dozen tags suggested, the whole sidebar can easily be compromised. In the new sidebar the problem is embanked: as a much higher number of tags is shown, the presence of some spam tags doesn't make the whole suggestion system unuseful;however, the order of branches could be gamed .

5 Conclusions and Future Work

We have proposed a new approach to integrate the navigation interface of a folksonomy adding explicit semantics provided by an ontology; we have developed a tool that uses WordNet to build a semantic hierarchy that helps users navigate and find related resources in del.icio.us.

We have shown that in this way it is possible to combine some advantages of the traditional top down approach to classification with the ones of the collaborative paradigm that is emerging on the Web, providing richer possibilities of searching and browsing, and dealing with some of the limitations to which folksonomies are prone, such as lack of recall, synonym control and gaming.

Our application is actually just a prototype and can be improved in several directions. The algorithm for the tree compression is one of the most delicate issues and could be improved by making it dynamic also for higher levels of the hierarchy, instead of just eliminating words contained in a black list.

Many improvements might be reached in tag recognition by using local wordnets in different languages and domain ontologies for specific terms.

As future work, it would be also interesting to use the results of tag disambiguation, performed by our application, to filter resources and not only tags; in this way it might be possible, for example, to show, among the del.icio.us bookmarks tagged as "turkey", only the ones that have been individuated as related to the geographical acceptation.

References

1. Clay Shirky. Shirky: Ontology is overrated – categories, links, and tags, 2005. http://shirky.com/writings/ontology_overrated.html.
2. Emanuele Quintarelli. Folksonomies: power to the people. June 2005. http://www-dimat.unipv.it/biblio/isko/doc/folksonomies.htm.
3. Ellyssa Kroski. The hive mind: Folksonomies and user-based tagging, December 2005. http://infotangle.blogsome.com/2005/12/07/the-hive-mind-folksonomies-and-user-based-tagging/.
4. Harry Halpin, Valentin Robu, and Hana Shepard. The dynamics and semantics of collaborative tagging. In *Proceedings of the 1st Semantic Authoring and Annotation Workshop (SAAW'06)*, 2006.
5. Adam Mathes. Folksonomies – cooperative classification and communication through shared metadata, December 2004. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.
6. E. Quintarelli, L. Rosati, and A. Resmini. Facetag: Integrating bottom-up and top-down classification in a social tagging system. In *IA Summit 2007*, 2007.
7. Christoph Schmitz, Andreas Hotho, Robert Jschke, and Gerd Stumme. Mining association rules in folksonomies. In V. Batagelj, H.-H. Bock, A. Ferligoj, and A. iberna, editors, *Data Science and Classification. Proceedings of the 10th IFCS Conf.*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 261–270, Heidelberg, July 2006. Springer.
8. Celine Van Damme, Martin Hepp, and Katharina Siorpaes. Folksontology: An integrated approach for turning folksonomies into ontologies. In *Bridging the Gap between Semantic Web and Web 2.0 (SemNet 2007)*, pages 57–70, 2007.
9. C. Fellbaum. *WordNet – An Electronic Lexical Database.* MIT Press, 1998.
10. Scott Golder and Bernardo A. Huberman. The structure of collaborative tagging systems, Aug 2005. http://arxiv.org/abs/cs.DL/0508082.
11. Hend S. Al-Khalifa and Hugh C. Davis. Towards better understanding of folksonomic patterns. In *HT '07: Proceedings of the 18th conference on Hypertext and hypermedia*, pages 163–166, New York, NY, USA, 2007. ACM Press.
12. Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. Wordnet: : Similarity - measuring the relatedness of concepts. In *AAAI*, pages 1024–1025, 2004.
13. S. Patwardhan, T. Pedersen, and S. Banerjee. SenseRelate::TargetWord - A Generalized Framework for Word Sense Disambiguation. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 73–76, Ann Arbor, MI, June 2005.