Collaborative semantic structuring of folksonomies

Freddy Limpens, Fabien Gandon
Edelweiss, INRIA
F-06902 Sophia Antipolis
{freddy.limpens, fabien.gandon}@inria.fr

Michel Buffa
KEWI, I3S - UNSA CNRS
F-06903 Sophia Antipolis
buffa@unice.fr

Abstract

The advent of tagging and folksonomies for organizing shared resources on the social Web brought promising opportunities to help communities of users capture their knowledge. However, the lack of semantics, or the spelling variations between tags lowers the potentials for browsing and exploring these data. To overcome these limitations, we propose exploiting the interactions between the users and the systems to validate or correct semantic analysis automatically applied to the tags. This process is based upon our model of the assistance of folksonomies enrichment which supports conflictual points of view. Several strategies can then be applied to propose novel browsing facilities to users.

1. Introduction

Social tagging has recently become an affordable and powerful means to categorize and organize shared resources within the social and collaborative Web, mostly thanks to its simplicity of use. The full exploitation of this type of knowledge representation is however problematic. The spelling variation of equivalent tags, such as "newyork" and "new_york", or the lack of semantic relations between related tags lower the potentials for navigating the tag space.

In this article we propose a methodology and some tools to tackle the limitations of folksonomies by building "lightweight ontologies" by integrating the users of a folksonomy-based system into the process of ontology maturing. These semantically richer structures can then be exploited to suggest semantically related terms, or to include spelling variants when retrieving resources associated with a tag. To achieve this goal, we propose associating the power of automatic handling of folksonomies and the expertise of users by integrating simple semantic functionalities within the interface of the system. Users will then be able to validate or correct the automatic inferences. This system is based on our model of semantic enrichment of folksonomies. According to this model, all the assertions that can be made on tags are first captured, even if contradictory. Then, the exploitation and application of these assertions is postponed pending further processing steps, for instance while sorting the results of a request.

This work is currently being tested in collaboration with the ADEME[1] agency. In this agency, we strive to promote the use of social bookmarking systems and social tagging among this community of users, as well as the unobtrusive embedding of the semantic enrichment of folksonomies within their every day tasks.

Our article is organized as follows. In section two we present our model of semantically enrichment of folksonomies. In section three, we explain how we implemented our method in a social bookmarking system to augment the navigation functionalities. In section four we discuss our position and conclude.

2. Model of folksonomy enrichment

The goal of our model is to describe the semantic relations that may exist between the tags of a folksonomy, and, at the same time, to support conflictual views between the users. For example, if a user says that "CO2" is narrower than "pollution", and another user says that "CO2" is narrower than "green-house gas", the model will record both assertions, even if they may contradict each other, temporarily, leaving it up to the designer of the systems to decide how to treat this conflict between several options (with a voting system for instance, or by showing explicitly the different points of view).

Our model is an extension of the RDF model of the reification of assertions[2] in the case of tags, and also includes already existing ontologies such as SIOC [1] or SCOT [2]. We propose an RDFS schema (see figure 1) in which an assertion on the semantics between two tags of a folksonomy is represented as a RDFS class (`TagSemantic-Statement`). Moreover, a user (`sioc:User`[3]), who may also be an automatic agent, may have proposed a semantic assertion (property `hasProposed`), or approved it (`hasApproved`), or rejected it (`hasRejected`). The semantic relationships between tags are specified by

[1] ADEME is the French for Environment and Energy Management Agency, see http://www.ademe.fr

[2] see http://www.w3.org/TR/rdf-mt/#Reif

[3] see http://rdfs.org/sioc/spec/

the subclasses of the class TagSemanticStatement which describes semantic relations between concepts : HasNarrower, HasBroader, HasRelated, and HasSpellingVariant. These semantic relations are those encountered within the SKOS schema, except that these relations are now classes instead of properties, and that each of these classes can be specified thanks to additional properties. The semantic relationships may also be specified, with the property rdf:predicate, by properties having the same meaning, such that SKOS subproperties of skos:semanticRelation, or such that scot:spellingVariant..

3. Implementation and results

In this section we present our implementation of a folksonomy-based system of bookmarks management, and how we combine automatic processing and semantic functionalities in order to assist the users in contributing to the semantic enrichment of the folksonomy.

3.1. Automatic treatments on tags

One of the widely known limitations of folksonomies is the spelling variations between supposedly equivalent tags such as "déchet" and "dechets". A simple solution to this problem consists in measuring the editing distance between these tags (such as the Levenshtein distance [3]), and to identify equivalent tags above a given threshold value. The experimentations we made so far gave good results with a value of 0.84, however further investigations will be made to understand the behavior of this threshold with other data sets

Another type of analysis consists in measuring the "similarity distance" between all the tags thanks to an analysis of the links between the tags, the users, and the tagged resources in a folksonomy. To this regard, [4] made a distinction between different ways of measuring such similarities: the measures based on simple cooccurrence of two tags for the same resource, and the distributional measures, which take into account three ways of associating tags: (1) *via* their usage for a single user (user-tag context), or (2) *via* their usage for a single resource (resource-tag context), or (3) *via* their common associations with other tags (tag-tag context). In our implementation, we have used the distributional measures of similarity based on the tag-tag context. This distributional measure of similarity between two tags t_1 and t_2 consists, first, in computing their associated vectors v_1 and v_2, whose components v_{ik} are equal to the value of the frequency of cooccurrence of the tags t_i and t_k which is incremented each time t_i and t_k are used for the same resource. The similarity measure is then computed as the cosine distance between the vectors:

$$\cos(v_1, v_2) = \frac{v_1.v_2}{\|v_1\|_2.\|v_2\|_2}$$

Table 1 shows a series of tags having a similarity value in the tag-tag context above 0.7. To obtain these measures, we have picked up delicious.com bookmarks which have the tag "ademe" (or its spelling variants)[4]. The results show relevant associations of related tags regarding the topic of ecology and sustainable development, which is what we could be expecting since the use of the tag "ademe" suggests a connection with these topics.

3.2. Integration in a bookmarks navigation system

The system we propose is a bookmarks navigator which is able to automatically include spelling variants within the results of a query, and to suggest related tags. Our system is composed of: (1) automatic agents applying semantic treatments on folksonomies, and (2) a user interface to browse the bookmarks database, and at the same time, to validate or correct the automatically suggested tags and semantic relationships. Figure 2 shows this interface displaying bookmarks tagged with "environment". One of the suggested functions consists in rejecting included spelling variants by clicking on a red cross. The second type of functionality proposes the users to reject (with the same symbol) or choose other types of semantic relationships between the original tag and the suggested related tags, such as "is narrower" (symbolized by arrows pointing the center of a circle) or "is broader" (symbolized by arrows pointing outside a circle). The actual use of these functionalities remains completely optional and is non intrusive to the regular use of the system.

In our model every assertion is recorded and added to the database, even when it is contradictory with other assertions (for example the assertion "pollution" is related to "car" has been approved by John, and rejected by Paul). The administrators of the system may then decide: (1) to make visible the contradictions by organizing them through different points of view, explicitly shown in the user interface (e.g. the point of view of the "car's opponents", and the point of view of the "car's defenders"); or (2) to show the results of an assertion according to the community to which the current user belongs (e.g. John and Paul belong to different communities, so we won't take Paul's assertion into account when displaying results to John); or (3) to rely on approval or rejection of the users to keep the assertions which collect the higher number of implicit votes.

[4] this excerpt is made of the 100 bookmarks of the 75 users who associated 221 distinct tags to 107 URLs

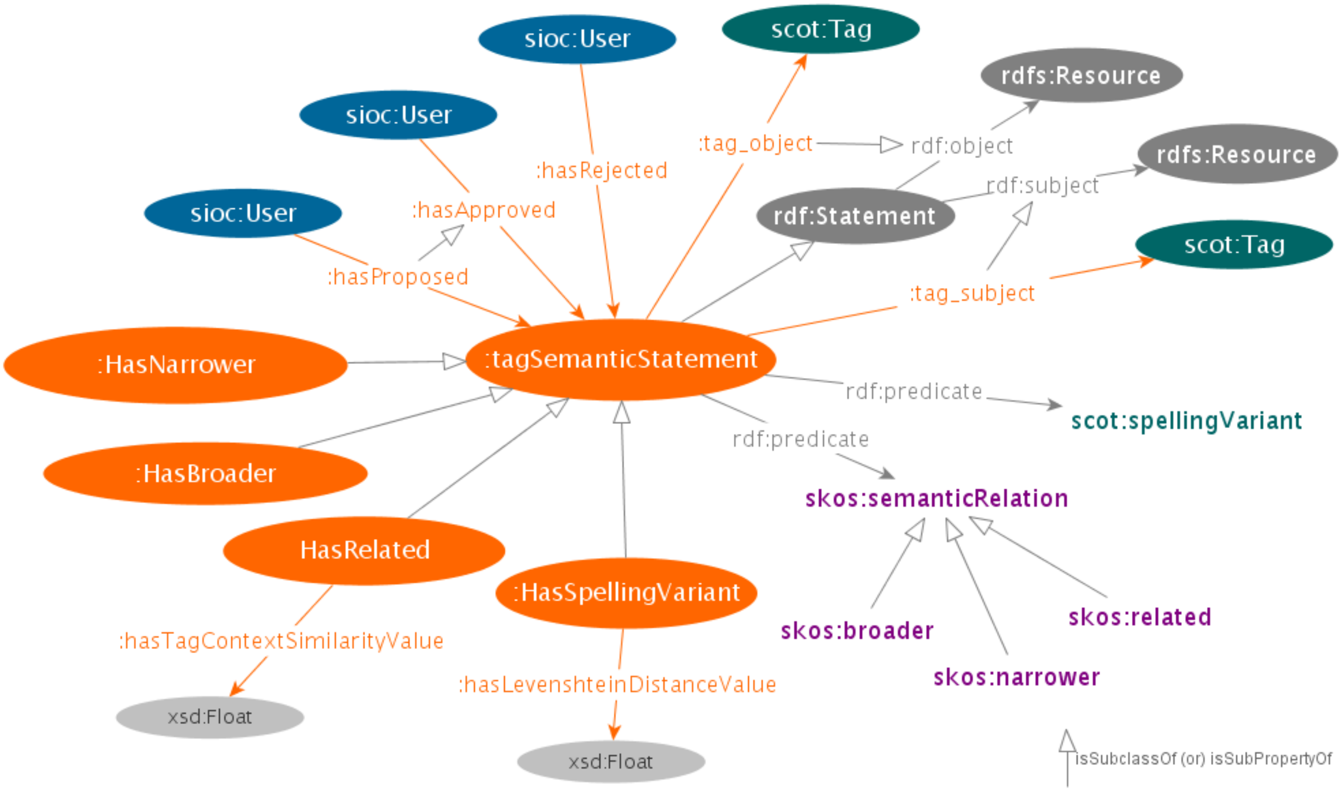
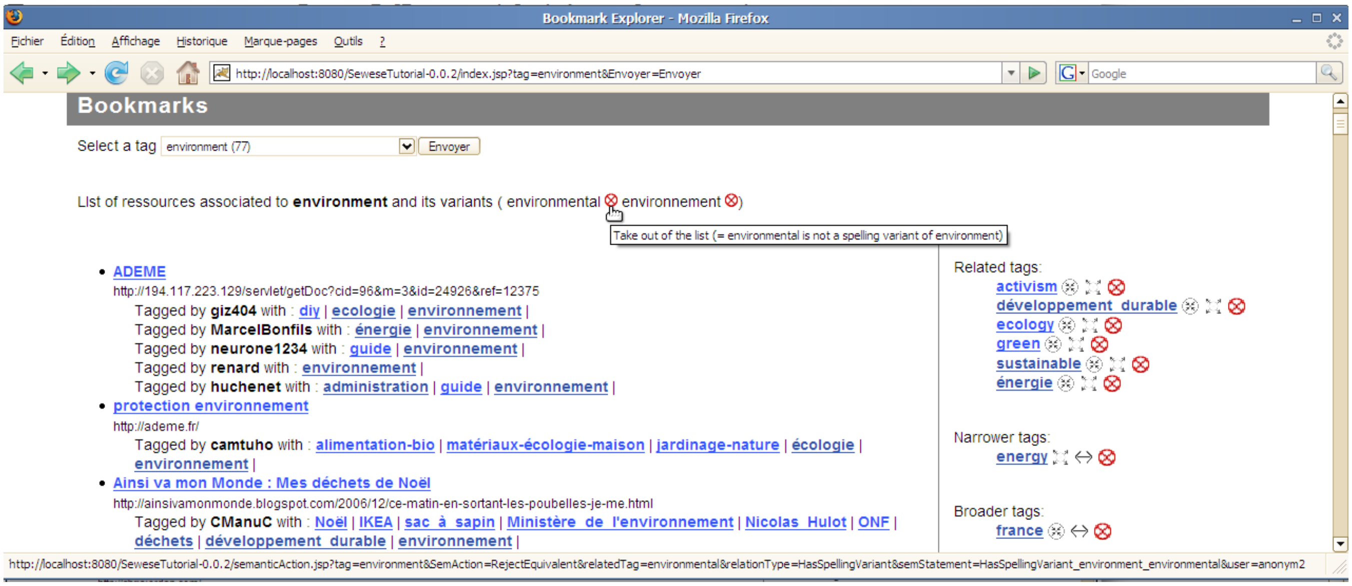


Figure 1. Reification of the notion of semantic relation

voiture (car)	auto (0.81), automobile (0.83), co2 (0.85), pollution (0.83)
développement (development)	durable (sustainable) (0.88), ecologie (ecology) (0.8)
solaire (solar)	photovoltaïque (photovoltaic) (0.74)

Table 1. Series of tags sharing a similarity value computed in the tag-tag context and above 0.7 (English translations between parentheses)


Bookmark Explorer - Mozilla Firefox
Fichier Édition Affichage Historique Marque-pages Outils ?
http://localhost:8080/SeweseTutorial-0.0.2/index.jsp?tag=environment&Envoyer=Envoyer
Google

Bookmarks

Select a tag environment (77) Envoyer

List of ressources associated to **environment** and its variants (environmental ⊗ environnement ⊗)

Take out of the list (= environmental is not a spelling variant of environment)

- ADEME
  http://194.117.223.129/servlet/getDoc?cid=96&m=3&id=24926&ref=12375
  Tagged by **giz404** with : diy | ecologie | environnement |
  Tagged by **MarcelBonfils** with : énergie | environnement |
  Tagged by **neurone1234** with : guide | environnement |
  Tagged by **renard** with : environnement |
  Tagged by **huchenet** with : administration | guide | environnement |
- protection environnement
  http://ademe.fr/
  Tagged by **camtuho** with : alimentation-bio | matériaux-écologie-maison | jardinage-nature | écologie | environnement |
- Ainsi va mon Monde : Mes déchets de Noël
  http://ainsivamonmonde.blogspot.com/2006/12/ce-matin-en-sortant-les-poubelles-je-me.html
  Tagged by **CManuC** with : Noël | IKEA | sac à sapin | Ministère de l'environnement | Nicolas Hulot | ONF | déchets | développement durable | environnement |

Related tags:
activism ⊗
développement durable ⊗
ecology ⊗
green ⊗
sustainable ⊗
énergie ⊗

Narrower tags:
energy ↔ ⊗

Broader tags:
france ↔ ⊗

http://localhost:8080/SeweseTutorial-0.0.2/semanticAction.jsp?tag=environment&SemAction=RejectEquivalent&relatedTag=environmental&relationType=HasSpellingVariant&semStatement=HasSpellingVariant_environment_environmental&user=anonym2

Figure 2. Screenshot of our user interface for navigating a bookmarks database

4. Conclusion

Our approach consists in integrating folksonomies into a collaborative construction of knowledge representations, aiming at providing additional functionalities to folksonomy-based systems. Several other research works tried to tackle the limitations of folksonomies by bridging them with ontologies[5]. Passant and Laublet [6] have proposed a model (MOAT) and some tools to link tags with their different meanings which are expressed within documents (Wikipedia articles) or concepts instances available on the Semantic Web. Our work differs from this by specifying the meaning of tags relatively to the other tags of the folksonomy thanks to a limited set of semantic relations (broader, narrower, etc.). But doing so does not prevent us from linking, independently, our tag-concepts to formal ontologies when this is relevant to our users (by using MOAT for instance). Other approaches propose integrating users directly in the elaboration of lightweight ontologies[7], or to semantically connect tags to each other with the help of automatic treatments and external ontological resources [8]. Our approach differs from these in that we are trying to complement and regulate automatic treatments made on tags thanks to the expertise of the users.

Our contribution is twofold. First, we proposed exploiting both the power of semantic automatic processing and the expertise of users to validate and regulate this processing. The two main functionalities we have presented in this paper are the detection of spelling variants of tags and the suggestion of related tags. These functionalities are suggested by the interface to induce users to validate, reject or correct the automatic suggestions. Second, we have also proposed a model which formalizes (1) the semantic relations between tags (to describe their meanings relatively to other tags), and (2) the semantic assertions made after automatic processing or made by the users themselves when they interact with the system. This model allows capturing and keeping track of all the semantic assertions, even when they are contradictory, and exploiting them in several ways according to the choice of the administrators of the system, who can, for instance, set up a voting system, or organize the contradictions as points of view explicitly shown to the users.

Our future work includes a testing campaign among our community of users from the "Ademe" agency, and the integration of semantic processing to detect other kinds of semantic relations (such as broader or narrower) and their corresponding functionalities within the user interface. The detection of sub-communities of interest and the semantic social network analysis [9]are also promising fields of research to us since we are seeking for different ways of personalizing the exploitation of the results of the semantic assertions. In the future, we wish to extend our research to the closer analysis of the everyday activities and working processes of our users' communities in order to identify other kinds of tasks which could be turned into opportunities for the semantic enrichment of shared knowledge.

Acknowledgment. This research was funded by the ANR agency project ISICIL ANR-08-CORD-011.

References

[1] U. Bojars, A. Passant, R. Cyganiak, and J. Breslin, Weaving SIOC into the Web of Linked Data. *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China* **(2008)**.

[2] H.-L. Kim, S.-K. Yang, S.-J. Song, J. G. Breslin, and H.-G. Kim, Tag Mediated Society with SCOT Ontology. *Semantic Web Challenge, ISWC* **(2007)**, http://iswc2007.semanticweb.org/main/default.asp.

[3] V. I. Levenshtein, Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady.* **(1966)**, Vol. 10, pp. 707–710.

[4] C. Cattuto, D. Benz, A. Hotho, and G. Stumme, Semantic grounding of tag relatedness in social bookmarking systems. *7th International Semantic Web Conference* **(2008)**.

[5] F. Limpens, F. Gandon, and M. Buffa, Bridging Ontologies and Folksonomies to Leverage Knowledge Sharing on the Social Web: a Brief Survey. *Proc. 1st International Workshop on Social Software Engineering and Applications (SoSEA)* **(2008)**.

[6] A. Passant and P. Laublet, Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China* **(2008)**.

[7] S. Braun, A. Schmidt, A. Walter, G. Nagypál, and V. Zacharias, Ontology maturing: a collaborative web 2.0 approach to ontology engineering. *CKC*, CEUR Workshop Proceedings Series, Vol. 273, CEUR-WS.org **(2007)**, http://dblp.uni-trier.de/rec/bibtex/conf/www/BraunSWNZ07.

[8] S. Angeletou, M. Sabou, and E. Motta, Semantically enriching folksonomies with flor. *CISWeb Workshop at Europ. Semantic Web Conf.* **(2008)**.

[9] G. Erétéo, M. Buffa, F. Gandon, M. Leitzelman, and F. Limpens, Leveraging social data with semantics. *W3C Workshop on the Future of Social Networking, Barcelona.* **(2009)**.