Towards Social Semantic Suggestive Tagging in a Digital Repository of Bookmarks

Fabio Calefato, Domenico Gendarmi, Filippo Lanubile, Giovanni Semeraro
University of Bari, Dipartimento di Informatica
{calefato,gendarmi,lanubile,semeraro}@di.uniba.it

Abstract

The organization of the knowledge on the web is increasingly becoming a social task performed by online communities whose members share a common interest in classifying different types of information for a later retrieval. Collaborative tagging systems allow people to organize a set of resources of interest through unconstrained annotations based on free keywords commonly named tags. Suggestive tagging techniques support users in this organization process and have shown to be helpful also in fostering a quick convergence to a shared tag vocabulary. In this paper, we propose a tag recommender which relies on the content analysis of the resource to be tagged, as well as on the personal and collective tagging history. The main contribution of this work is a model which combines semantic content analysis methods with existing suggestive tagging techniques. The expected benefit is the improvement of the user experience in social bookmarking systems, and more generally in collaborative tagging systems.

Categories and Subject Descriptors

H.3.1 Content Analysis and Indexing; H.3.7 Digital Libraries;

General Terms

Design

Keywords

collaborative tagging, folksonomy, recommender system, semantic web, content analysis, suggestive tagging, social bookmaking.

1 Introduction

The phenomenon of Web 2.0 [10] has led to the development of many tools, which have succeeded in making the task of knowledge organization more attractive to a broader audience. Tools for accomplishing this activity, such as collaborative tagging systems, harness the power of virtual communities and have been shown effective in gathering quickly large amounts of information directly generated by users.

Collaborative tagging systems allow people to organize a set of resources, by annotating them with tags through a browser. Tags can be regarded as free keywords used by people to label resources of interest. The activity of labelling is called tagging, as it consists of attaching one or more tags to the resource. Although this tagging activity is accomplished individually, while using the system, everyone can see who else is participating by observing others' tagging behaviours. This tight feedback loop makes these systems social and the result is a collection of annotations, also called folksonomy [15]. Unlike top-down centralized classification approaches, folksonomies have revealed a noteworthy ability in adhering to the personal way of thinking [4]. The opportunity of using free tags with no restrictions allows users to express their own perspective on the annotated resource. Therefore, these annotations can become a reliable indicator of interests and preferences of the active participants in such systems.

To date, most collaborative tagging systems provide a limited support to users in the annotation process, as they typically recommend tags by arranging suggested tags in a tag cloud that

emphasizes tags on the basis of their popularity: the bigger the font, the more used the tag. By suggesting to the user his/her most used tags, as well as the most popular tags in the whole community, this form of tag recommendation takes into account both the personal and social dimension of folksonomies. Nevertheless, this approach falls short of considering the semantic dimension for the content of the resource that is going to be annotated.

We acknowledge that suggesting meaningful tags to a user, according to personal and social interests, can enhance the user experience and augment the number of active participants in the annotation process. However, we argue that, the content analysis of the resources can significantly improve the accuracy of suggestive tagging, thus, fostering a quick convergence to a shared tag vocabulary and limiting the tag synonymy issue [5].

In this paper, we propose a tag recommender which relies on the semantic analysis of the resource content which is going to be annotated, as well as on the personal and collective tagging history. Such an approach is able to address the typical cold-start problem affecting recommender systems [12]. In fact, our recommender system will be able of suggesting tags to users who have not yet tagged any resource, by putting forward tags which are popular in the community. Further, when there are resources not yet tagged by anyone in the community, our recommender will suggest tags which have been gathered through a semantic analysis of the resource content.

The main contribution of this work is a model which combines semantic content analysis methods with existing suggestive tagging techniques. The expected benefit is the improvement of the user experience in social tagging bookmarking systems, and more generally in collaborative tagging systems.

The remainder of the paper is structured as follows. Section 2 describes how the content analyzer works and how it is going to be integrated in the proposed tag recommender system. In Section 3 we present our model through four typical scenarios which can take advantage from a mix of semantic content analysis and traditional suggestive tagging. Section 4 surveys novel related work concerning suggestive tagging in folksonomies. Finally, Section 5 draws conclusions and points out some challenges we are going to address in the near future.

2 Content Analysis for Semantic Tag Suggestion

The idea behind applying content analysis for semantic suggestive tagging is to provide a user who wants to tag a resource, not only with relevant words extracted from a resource, but also with a set of synonyms. In this way, other than fostering tag convergence, we also increase the probability of suggesting tags that fit better to users' personal way of thinking, without affecting the meaning.

To make this possible, a word sense disambiguation (WSD) algorithm is needed, which can assing a word w occurring in a given resource (e.g., a web document), to the appropriate sense, according to the context (i.e., the set of words that precede and follow w). Then, once the appropriate sense of a word w is identified, a dictionary or a lexical ontology can be used to find its synonyms, and to provide the user with a set of recommended tags alternative to w.

META (MultilanguagE Text Analyzer) [3] is a tool developed at the University of Bari, which implements an algorithm to perform WSD on text documents in a variety of formats (e.g., pdf, doc). The tool has also been used by the Item Recommender system (ITR) to learn sense-based user profiles [13]. In addition to performing the basic content analysis tasks (e.g., stop-words elimination, stemming), META is also able to analyze different parts of text documents, called slots. For instance, when processing papers from a conference proceedings, META performs content analysis on the title, abstract, and body, separately. Furthermore, META relies on WordNet for obtaining a sense inventory. Thus, after performing the WSD, META returns the unique id in WordNet (called offset) of the correct sense identified, for each word extracted. Our idea is to use offsets to retrieve from the lexical ontology the whole set of synonyms (SYNSET, in short) for each relevant word extracted that will be suggested as a tag.

3 The Suggestive Tagging Model

In [1, 9] a generic collaborative tagging system is defined as a tripartite 3-uniform hypergraph $F=(N,E)$ where $N= U \cup T \cup R$ is the union of three disjoint sets of entities, namely a set of registered users (U), a set of applied tags (T), and a set of annotated resources (R). Furthermore, we define $E= \{(u,t,r) \mid u \in U, t \in T, r \in R\}$ as the set of all the annotations that compose the folksonomy. Given the above definitions, we can define a typical social bookmarking system as a folksonomy where R is replaced by a set of bookmarks B, pointing to resources in R.

For each entity within such system, we can also discern among different kinds of tags, users and bookmarks. Given a user u and a selected bookmark b we can identify three sets of tags:

- *Personal Tags(u),* all the tags assigned by u to all bookmarks.
- *Social Tags(b),* all the tags assigned by all users to b.
- *Semantic Tags*(b), all the tags extracted by analyzing the content of the resource pointed by b.

Depending on the amount of tags adopted by a single user, we can also discriminate a user as *novice*, if he/she has no tags, or as *expert*, when he/she has started to annotate bookmarks using tags. The definition of novice includes hence both new users just registered to the system and users registered for a while but loath in using tags to save bookmarks. Finally, a bookmark can be categorized as *tagged* if it has been annotated with at least one tag, or *untagged* if there are no associated tags.

According to the above definitions, we illustrate four scenarios which depict how a user can be supported by tag recommendations in his/her task of saving a bookmark (Table 1). Because of the huge number of registered users and the availability of suggestive tagging features, we use del.icio.us[1] as the reference system for the proposed approach. In the following, we consider four users, namely John and Dexter (as novices), and Alice and Bea (as experts). We also assume the Collaborative Development Group Research page[2] as an untagged bookmark and DELOS home page[3] as a tagged one.

	Untagged Bookmark	Tagged Bookmark
Novice	Scenario 1	Scenario 3
Expert	Scenario 2	Scenario 4

Table 1. Four exemplary scenarios

Scenario 1: A Novice user saving an Untagged Bookmark
John is going to save Collaborative Development Group Research page as a bookmark. Being a novice user, he has no tags yet. In this scenario del.icio.us cannot provide any suggestion because John has no personal tags and nobody else has saved this bookmark yet (Figure 1).

Figure 1. No tags suggested by del.icio.us

[1] http://del.icio.us
[2] http://cdg.di.uniba.it/index.php?n=Research.HomePage
[3] http://www.delos.info/

However, even if no suggestions are available from either personal or social tags, according to our view, it is still possible to support John with tag recommendations by providing *Semantic Tags* as the output of the content analysis (Figure 2).
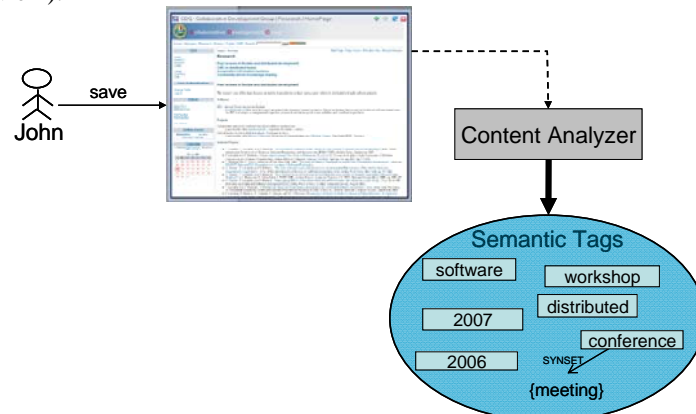

Figure 2. Semantic Tags as output of the content analysis

In particular, META extracts the following words from the Collaborative Development Group Research page (for the sake of space, we limit to six the number of words extracted):

- software (occurring 27 times)
- distributed (21)
- 2007 (16)
- 2006 (10)
- conference (7)
- workshop (5)

In addition, META identifies the correct sense for the words extracted. For instance, for the word "conference" the sense extracted from the inventory is "prearranged meeting, especially with a formal agenda", while the other offset (i.e., "association of sport teams") is just skipped. Finally, the whole set of semantic tags is obtained by also retrieving from WordNet the SYNSET for "conference" (i.e., {meeting}).

Scenario 2: An Expert user saving an Untagged Bookmark
Alice is an expert user and thus, she has already used some tags to annotate bookmarks in del.icio.us. Now she wants to save a bookmark never tagged before in the system. In such a case, del.icio.us can suggest only those tags which have been already adopted by Alice, even though most of them might be inappropriate (Figure 3).

Figure 3. Personal tags suggested by del.icio.us

Instead, in this scenario, we argue that a hopefully more useful set of tags can be suggested to Alice by intersecting both Alice's *Personal Tags* (i.e., collaborative, Web2.0, 2007, conference, …) and the *Semantic Tags* extracted from the bookmark (i.e., software, distributed, 2007, conference).

We define the intersection of the two sets as the *Personal Semantic Tags(u, b)*, i.e., all the tags from a user which have also been obtained from the content analysis of the resource pointed by a bookmark. If the set of Personal Semantic Tags is not empty, these tags will be suggested to Alice as recommended tags, in addition to the remaining *Personal Tags* and *Semantic Tags* (Figure 4).
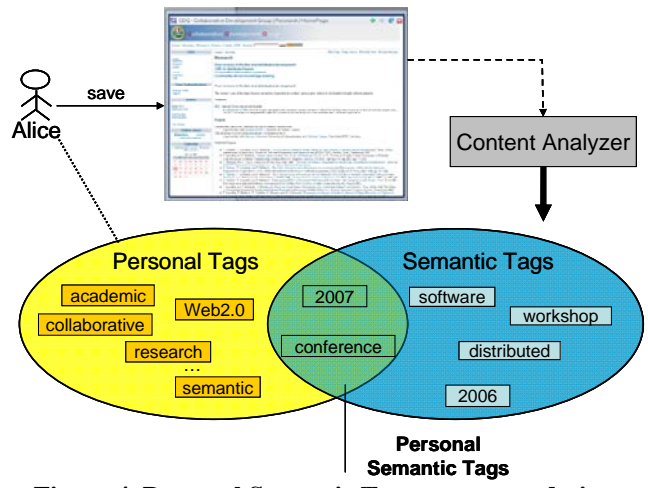
Figure 4. Personal Semantic Tags recommendation

Scenario 3: A Novice user saving a Tagged Bookmark

Dexter is a novice user who has been registered to del.icio.us for two months, but he has never used tags. Now he is going to save the DELOS home page that has already been tagged by Alice and other users. Typically, del.icio.us suggests only popular tags if the selected bookmark has been annotated by more than one user (Figure 5).

Figure 5. Popular tags suggested by del.icio.us

This time, a recommender that implements our approach might benefit from both *Semantic Tags(b)* and *Social Tags(b)*. Assuming that the *Social Tags(b)* and the *Semantic Tags(b)* sets are not disjoint, we define the intersection between these two sets as *Social Semantic Tags(b)* (i.e., delos, library, research, digital, …). As in the previous scenario, our social/semantic tag recommender would suggest tags belonging to the intersection as *Recommended Tags* and also the remaining *Social Tags* and *Semantic Tags* (Figure 6).
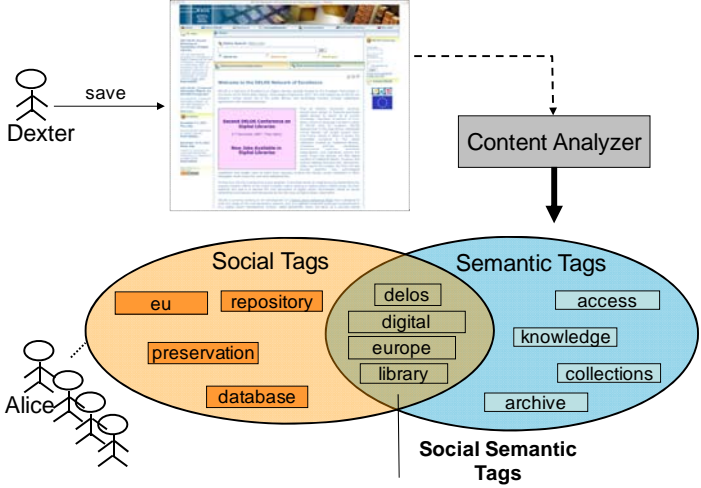

Figure 6. Social Semantic Tags recommendation

Scenario 4: An Expert user is saving a Tagged Bookmark

Bea is an expert user who uses del.icio.us to save her bookmarks on a daily basis, and thus she has a large set of personal tags. She is now going to save and tag the DELOS home page, which has been also tagged by both Alice and Dexter, among the others. In such a scenario, del.icio.us would provide Bea with *Her tags*, i.e., Bea's personal tags used for other bookmarks, *Popular Tags*, i.e., the most used tags for that bookmark by other users, and *Recommended Tags*, i.e., the personal tags that have been also used by others for that bookmark (Figure 7).

Figure 7. Personal, Popular and Recommended tags suggested by del.icio.us

Other than suggesting these three sets of tags, according to our approach, it is also possible to exploit the *Semantic Tags(b)* obtained through the content analysis of the DELOS home page. In this scenario, Bea is supported with four different kinds of tags recommendation:

- *Personal Semantic Tags(u, b)*
- *Social Semantic Tags(b)*,
- *Shared Tags(b, u):* those tags belonging to the intersection between *Personal Tags(u)* and *Social Tags(b)*

Semantic Shared Tags(u, b): those tags belonging to the intersection of all the above available sets of tags, namely *Personal Tags(u), Social Tags(b),* and *Semantic Tags(b)* (Figure 8). We argue that the quality of tag recommendations provided to Bea could be significantly improved by presenting intersections which limit the information overload.
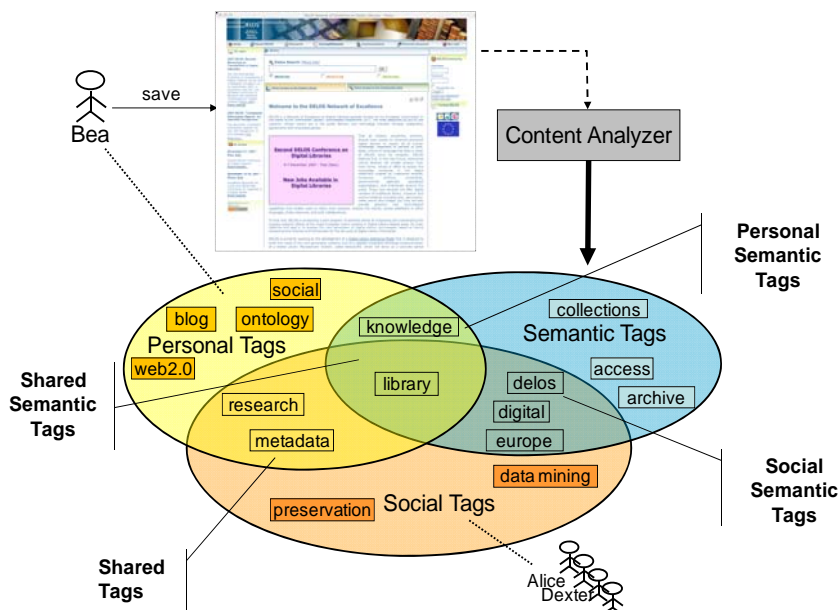

Figure 8. Four different kinds of tags recommendation

Finally, in the following table, we summarize all the recommended tags, according to each presented scenario. In particular, strongly recommended tags, i.e., those tags that, in our view, are hopefully more useful, are shown in bold (Table 2).

	Untagged Bookmark	*Tagged Bookmark*
Novice	Semantic Tags	**Social Semantic Tags** Semantic Tags Social Tags
Expert	**Personal Semantic Tags** Semantic Tags Personal Tags	**Semantic Shared Tags** **Shared Tags** **Social Semantic Tags** **Personal Semantic Tags** Semantic Tags Social Tags Personal Tags

Table 2. Recommended tags for each scenario

4 Related Work

Suggestive tagging within folksonomies is a rather novel field of research [8]. The evidence of such a novelty is the quite sparse literature related to the state of the art on tag recommendations.

One existing approach to tag suggestions is referred to as *selection of tags,* which indicates that systems select a small number of tags to display, among the sheer size of terms already associated to an item. With respect to this approach, Sen et al. [14] investigated how different algorithms for selecting tags to display, influence users' personal vocabularies while annotating movies in a movie recommendation system.

A similar approach was also proposed by Xu et al. [16], who defined a set of general criteria for a good tag suggestion algorithm, in order to identify the most appropriate tags, while eliminating noise and spam. These criteria, identified through a study of tag usage by real users in My Web 2.0, include high coverage of multiple facets to ensure good recall, least effort to reduce the cost involved in browsing, and high popularity to ensure tag quality.

Based merely on the social dimension of tagging systems is the work of Jaschke et al. [7], who presented two different algorithms for recommending tags. The first algorithm is based on collaborative filtering [11], whereas the second is based on the FolkRank algorithm, defined in [6], and exploits the graph structure of folksonomies. The comparison, performed using two datasets from real–life folksonomies, namely Last.fm and Bibsonomy, showed that the graph-based FolkRank algorithm outperforms collaborative filtering approaches.

Finally, following a similar approach to the one proposed in this paper, Byde et al. [2] described a tag recommender system based on two different resource similarity metrics, which take into account the tag used by one user to annotate resources and their content, respectively. Although this work was the first to introduce content-based methods for recommending tags, it failed to take into account the social dimension of folksonomies (i.e., the community tags) to compute the resource similarity, considering only the personal resources and tags.

5 Conclusions and Future Work

Suggestive tagging fulfils several needs: it helps users in the annotation process, fostering a quick tag vocabulary convergence, and enhances the likelihood of a resource to get tagged. However, current systems suggest tags only on the basis of personal recent use or because of their popularity among the community.

In this paper, we have described a model which combines semantic content analysis methods with existing suggestive tagging techniques. By exploiting semantics of content analysis, provided by the META tool, and social features, built-in in folksonomies, the proposed recommender can address tag recommendations even in borderline cases, such as a user which has never used tags previously or a resource with no associated tags. We also intend to extend META and adapt it for the purpose of performing the content analysis of web resources to be annotated. Web pages organize their content in the HTML <head/> and <body/> slots. The analysis of the content of the <head/> slot, in particular, can offer valuable insights for the purpose of suggesting tags to annotate a web resource. In fact, editing both the <title/>, and the *keywords* and

description <meta/> slots can be thought as an accurate form of free annotation, because their content reflect just the personal view of the web page creator/maintainer on its whole content.

Our approach to suggestive tagging has been presented in the context of del.icio.us, the most popular social bookmarking system. As future work, we plan to complete the development of the proposed recommender system and perform an explorative experimentation within del.icio.us, having the existing suggested tagging feature as a control group.

Bibliography

1. Abbattista, F., Calefato, F., Gendarmi D., Lanubile, F.: Shaping personal information spaces from collaborative tagging systems. In PB. Apolloni et al. (Eds.): KES 2007/ WIRN, Part III, LNAI 4694, pp. 728–735, 2007. Springer-Verlag Berlin Heidelberg 2007.
2. Byde, A., Wan, H., Cayzer, S.: Personalized Tag Recommendations via Social Network and Content-based Similarity Metrics. In Proceedings of the International Conference on Weblogs and Social Media (ICWSM'06), March 2006.
3. Degemmis M., Lops P., and Semeraro G.: A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation. User Modeling and User-Adapted Interaction, 17(3), pp. 217-255, July 2007.
4. Gendarmi D., Abbattista F., Lanubile F.: Fostering knowledge evolution through community-based participation. Proceedings of the 1st Workshop on Social and Collaborative Construction of Structured Knowledge at WWW'07, 2007.
5. Golder, S., Huberman, B.: Usage patterns of collaborative tagging systems. Journal of Information Science, 32(2), 198-208, 2006.
6. Hotho, A., Jaschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: Search and ranking. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 411–426. Springer, Heidelberg (2006)
7. Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., Stumme, G.: Tag Recommendations in Folksonomies. In Joost N. Kok and Jacek Koronacki and Ramon López de Mántaras and Stan Matwin and Dunja Mladenic and Andrzej Skowron, (Eds), Knowledge Discovery in Databases: PKDD 2007, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, LNCS, vol. 4702, pp. 506-514. Springer, Berlin, Heidelberg, 2007
8. Marlow, C., Naaman, M., Boyd, D., Davis, M.: HT06, tagging paper, taxonomy, Flickr, academic article, to read. Proceedings of the Seventeenth Conference on Hypertext and Hypermedia. 31–40, 2006.
9. Mika, P.: Ontologies are us: A unified model of social networks and semantics. Proceedings of the 4th International Semantic Web Conference, LNCS, Vol. 3729. Springer-Verlag, 522-536, 2005.
10. O'Reilly T.: What is Web 2.0. Design Patterns and Business Models for the Next Generation of Software.
11. Sarwar, B., Karypis, G., Konstan, J., and Reidl, J.: Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th international Conference on World Wide Web. ACM Press,2001
12. Schein, A. I., Popescul, A., Ungar, L. H., Pennock, D. M.: Methods and metrics for cold-start recommendations. In Proceedings of the 25th Annual international ACM SIGIR Conference on Research and Development in information Retrieval. ACM Press, 2002.
13. Semeraro G., Degemmis M., Lops P., Basile P. Combining Learning and Word Sense Disambiguation for Intelligent User Profiling. Twentieth International Joint Conference on Artificial Intelligence, January 6-12, 2007, Hyderabad, India, 2007.
14. Sen, S., Lam, S. K., Rashid, A., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F. M., Riedl, J.: Tagging, communities, vocabulary, evolution. Proceedings of the 20th Anniversary Conference on Computer Supported Cooperative Work, (2006), 181-190
15. Vander Wal, T. Folksonomy Definition and Wikipedia.2005
16. Xu, Z., Fu, Y., Mao, J., Su, D.: Towards the Semantic Web: Collaborative Tag Suggestions. Proceedings of Collaborative Web Tagging Workshop at 15th International World Wide Web Conference (2006).