Recommending new tags using domain-ontologies

Andrea Baruzzo, Antonina Dattolo, Nirmal Pudota, Carlo Tasso
Department of Mathematics and Computer Science
University of Udine
Udine, Italy
{*andrea.baruzzo,antonina.dattolo,nirmala.pudota, carlo.tasso*}*@dimi.uniud.it*

Abstract

Tagging is a representative activity of social Web, useful for organizing information into knowledge. This activity presents some open issues, due in the majority to the manual insertion of tags. On the other hand, domain ontology is a specification of the conceptualization of a domain in terms of concepts, attributes and relations. Domain ontologies have a good potential to improve information organization, management and understanding. In this paper, we propose an automated approach for recommending new tags for Web resources by using domain ontologies and key-phrases. The proposed approach is implemented in the PIRATES framework, a prototype system for personalized content retrieval, annotation, and classification. Our approach is then explained with a simple use-case scenario.

1. Introduction

We live today in a world of collaborative publishing which emphasizes (and sometimes leads to the creation of) social networks, folksonomies, e-learning communities. Considering the amount of user-generated contents available on the Web and its steady growth rate, however, Web 2.0 is actually leading to an exacerbation of information overload. In this context, we propose to handle the classical problem of retrieve, recommend, or classify new content by exploiting some typical characteristics explicitly introduced by Web 2.0 (i.e. tags and ontologies). In particular, in this paper, we concentrate on the *tagging* phenomena: a *tag* is a keyword users use to annotate the content, in order to describe, organize, and correlate it with other contents, or simply to retrieve it easily in future searches. Numerous social tagging systems, such as *del.icio.us* (http://delicious.com/) for Web pages, *Bibsonomy* (http://www.bibsonomy.org/) for scientific publications, *Flickr* (http://www.flickr.com/) for images, have become popular thanks to the tagging feature.

In order to exploit tagging to recommend new content appropriate to a specific user need, we have to better understand the nature of tagging, for which purpose it is used, and what are the typical limitations that affect it. Tagging is a textual annotation technique based on *meta-data information* (i.e. keywords); this activity may be *manual* if it is generated by a human user, or *automatic* if it is generated by a dedicated software. Users can employ tags differently because they can be guided by different tasks. Typically, tagging is used with the explicit intent of:

1) *classifying a content* by means of a corpus of concepts which are familiar to the user (e.g. taxonomies, thesauri, or any bag of keywords representing meaningful categories for him/her);
2) *summarizing a resource content* by means of a short list of keywords representing the user-generated content description;
3) *expressing a polarity judgment* about a content by means of proper adjectives provided as tags (e.g. *"sad"*, *"wonderful"*);
4) *correlating tagged resources with people and their skills* such as the level of expertise, the reputation, or the importance of a person mentioned in the resource content (e.g. *"guru"*, *"geek"*, *"vip"*, *"bill-gates"*);
5) *creating dichotomic classification criteria* in order to describe resources as belonging or not to a particular category (e.g. *"clinical"/"not-clinical"*, *"statistical"/"not-statistical'*, *"accepted"/"rejected"*);
6) *providing a temporal information* to a resource (e.g. dates of correlated events).

To some extent, all these forms of tagging express a *classification intent* targeted to establish effective schemata for organizing knowledge in the Web space and to facilitate later retrievals. In our approach, we consider content-based filtering in terms of a classification process; thus we concentrate on tagging as a classification technique.

Tagging allows users to freely determine suitable labels for their resources without relying on any predetermined vocabulary or hierarchy. Moreover, tags can be very effective for serendipitous browsing a digital archive of documents (or bookmarks) in order to find relevant information. Hence people tag the content with their own vocabulary and ultimately their mental models in order to facilitate the process

The authors acknowledge the financial support of the Italian Ministry of Education, University and Research (MIUR) within the FIRB project number RBIN04M8S8.

978-0-7695-3801-3/09 $26.00 © 2009 IEEE
DOI 10.1109/WI-IAT.2009.313

of recall. Besides the potential benefits, resulted tags suffer with some of the notable limitations [1]:

- *Ambiguity*: with an uncontrolled vocabulary, many tags can be ambiguous. Indeed in tags we can find the same ambiguity that we find in natural language (e.g., homonymy, polysemy, synonymy, spelling mistakes, disambiguation).
- *Undistinguished concerns*: social tagging systems does not enforce, or even propose, a schema for distinguishing the purpose of a meta-data value. Tags might be, indifferently, proper names, subject descriptors, genres, self-reminders; tangential remarks (such as colors or years for pictures).
- *Independence of terms*: social tagging does not provide relations to connect and relate different terms: each tag is independent of the others, and no inference is possible (the structure of a tag system is "flat").
- *Effort*: systematically (and consistently) tagging Web resources is tedious, error prone, and rather wearying.

In order to alleviate some of these limitations, we propose an automated approach which assists the user when (s)he tags a Web resource: a software system analyzes the textual content of such resource, and provides new tag suggestions/recommendations by exploiting a domain ontology. Using this approach, we try to achieve two different goals:

- *use a controlled, ontology-based vocabulary*, not necessarily present in the original Web resource, in order to classify it as a result of the automatic tagging process;
- *reduce the manual effort* required to tag a Web resource.

This paper is organized as follows: Section 2 discusses the state-of-art in the field of filtering and classification by exploiting tagging and/or ontologies; Section 3 introduces the PIRATES framework, and a specific module, ORE (Ontology Reasoner Engine), which automatically provides new tag recommendations for a generic textual content exploiting a domain ontology; Section 4 provides a use case scenario which illustrates an interaction between the user and the PIRATES system. Section 5 concludes the paper with a brief overview of possible applications of our approach and future research lines.

2. Related Work

Many Web 2.0 systems, such as *Del.icio.us* or *Bibsonomy*, require minimal efforts to annotate resources with unrestrained keywords, but suffers with the aforementioned limitations, that could be alleviated by using ontologies. Ontologies have a huge potential to improve information organization, management and understanding, but their use in supporting activities like tagging and classification represents yet an open challenge. Different methodologies and approaches, used in literature, have been analyzed in [1]: some works simply extend ontologies in a folksonomy-like approach; other works add multiple labels to ontology nodes. Another line of research is concerned with extracting basic semantic relations from folksonomies or adding more ontology-like features to social tagging. For example, *Folk2onto* [2] maps social tags (taken from *Del.icio.us*) to ontological categories (using a Dublin Core-based ontology) in order to classify and give a proper structure to the tagged resources. Another system, *ePaper* [3], uses a hierarchical news ontology, based on the *IPTC* (www.iptc.org) Subject Codes taxonomy, as a common language for content based filtering in order to classify news items and to deliver personalized newspaper services on a mobile reading device. In [4] the authors propose an ontological approach in Personalised E-Learning Scenarios; in [5] the authors presents a new ontology-based model for resource inventory by integrating semantic Web technologies and agents paradigm.

In literature there are many examples of tag recommender systems, but the major part of them do not use ontologies: *Autotag* [6] recommends tags to weblog posts based on the tags assigned to similar weblog posts in a given collection; it uses information retrieval measures to find similar weblog posts. Other systems such as [7] suggest tags for new bookmarks, using textual content associated with bookmarks to model documents and users: in this case, the authors exploit the *Bibsonomy* dataset which contain Web pages and publications.

3. The PIRATES framework

PIRATES (Personalized Intelligent Recommender and Annotator TEStbed) is a framework for text-based content retrieval and categorization which exploits social tagging, user modeling, and information extraction techniques. The main feature of PIRATES concerns a novel approach that automates in a personalized way some typical manual tasks (e.g. content annotation and tagging). In particular, we proposed an automated method to assist a user interested in tagging a Web resource. Our approach analyzes the textual content of resources, and provides new tag recommendations by exploiting an existing domain ontology. We used the ontology to examine how domain knowledge can help in the tasks of classification and tagging.

3.1. PIRATES Architecture

PIRATES operates on a set of input documents stored in the Information Base (IB) repository. In order to classify them, it suggests some personalized tags and other forms of textual annotations (e.g. key-phrases). The input documents are then annotated with these tags, forming the Knowledge Base (KB) repository.

The PIRATES architecture, shown in Figure 1, is formed by three major components:

- The *Cognitive Filtering Tools* module implements IFT (Information Filtering Tool), a system based on an

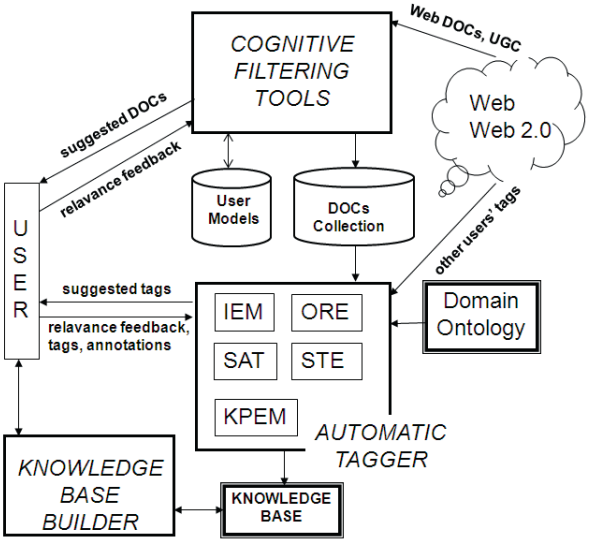

Figure 1. PIRATES architecture

algorithm [8] designed to build representations of user interests (*IFT user models*). Exploiting these models, IFT provides mechanisms of relevance feedback used to tune the classification of a document which belongs to an incoming stream of input documents (for example the results of a spidering process over the Web). The classification process produces evaluations of the relevance (in the sense of topicality) of a document according to a specific user model represented with semantic (co-occurrence) networks.

- The *Automatic Tagger* module implements a set of modules devoted to automatically annotate an incoming stream of text (the content of a document) by means of tag recommendations: IEM (Information Extraction Module) suggests named entities, KPEM (Key-Phrases Extraction Module) provides key-phrases, SAT (Sentiment Analysis Tool) identifies polarity judgments, STE (Social Tagger Engine) assigns tags used by a community of Web 2.0 users, while ORE (Ontology Reasoner Engine) recommends tags extracted from an ontology. In this paper we focus on the description of the ORE submodule, while interested readers may find detailed descriptions of the other modules in [9], [10].
- The *Knowledge Base Builder* module organizes documents in a knowledge base repository, producing annotated documents and user conceptual maps. A more detailed description of this module is proposed in [11].

3.2. An Ontology-based Tag Recommender System

Ontology-based tag recommender system is based on the ORE module of PIRATES framework. ORE works on the result produced by KPEM which implements a variation of the KEA algorithm [12] for key-phrases extraction. KPEM identifies n-gram key-phrases (typically n between 1 and 4) that summarize the input document. Initially, for each key-phrase provided by KPEM for the given document, ORE is programmed to find the corresponding match with the terms in ontology. ORE is useful if there exists at least one match. In this case, ORE follows a special navigation strategy to find ancestor nodes and common ancestor nodes of the corresponding matches. We have followed spreading activation algorithm [13] to implement the navigation strategy composed by the following steps:

1) For each key-phrase extracted by KPEM for a given document, the algorithm looks for a corresponding match in the ontology, retrieving its immediate super class by following children-parent relationships.
2) As second step, the retrieved superclass is marked as ontology concept mapping node.
3) Then, if there are at least two ontology concept mapping nodes, it retrieves the common ancestor node for them and possibly all the nodes in the path between the ontology concept mapping nodes and the common ancestor node.

4. A Use Case Scenario

Suppose the Cognitive Filtering Tools module notifies (among the others) the paper "A UML Class Diagram Analyzer"[1]. In order to classify this new content, the user exploits two PIRATES annotators, KPEM and ORE (Figure 2). In particular, in this example, the user configured the ORE annotator in order to use an ontology in the field of software engineering, named *Software_engineering.owl*[2]. Using this ontology and starting from the key-phrases extracted by KPEM, ORE implements the navigation strategy described in Section 3.2. For four out of the suggested key-phrases (i.e. *Alloy*, *UML*, *OCL*, and *Invariants*), ORE identifies a corresponding one-to-one match in the ontology (as shown in Figure 3). Starting from these nodes, ORE uses the spreading activation algorithm to find common ancestors representing more abstract subjects. Then both one-to-one ontology mappings and common ancestors are provided to the user by PIRATES as potential tag recommendations.

In this way, for the input document, ORE recommends five new tags which are not presented in the text (i.e. *Software Design Notations*, *Formal Specification Languages*, *Design by Contract*, *Formal Specification Techniques*, and *Software Design*). These tags represent *abstractions* of the key-phrases extracted by the other annotators available in PIRATES.

1. http://twiki.cin.ufpe.br/twiki/pub/SPG/GroupPublications/csduml04.pdf.
2. This example makes use of a personalized version of the domain ontology available from http://www.seontology.org/.

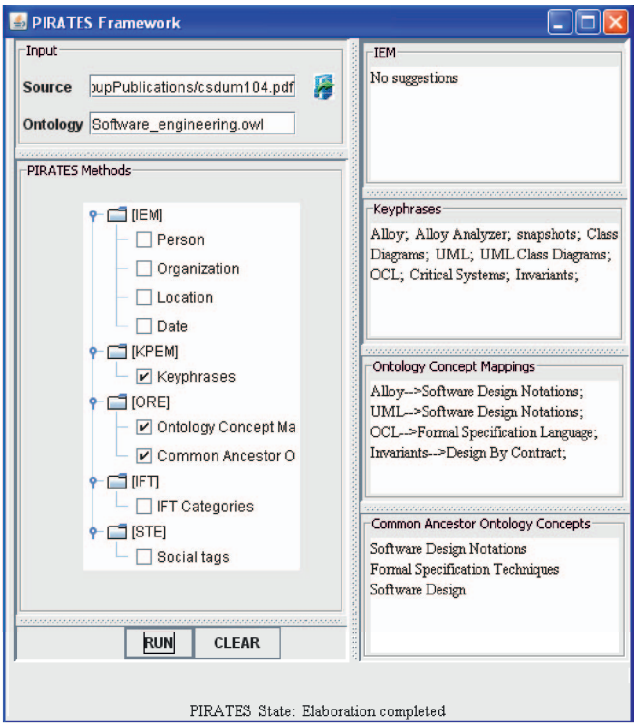

Figure 2. A screenshot of our Pirates prototype

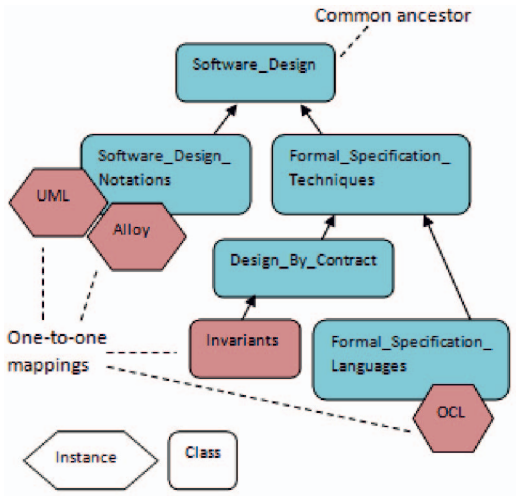

Figure 3. Ontology reasoning

5. Conclusion and Future Work

In this paper we have presented our approach for automatically suggesting new tags using key-phrases and domain ontologies. We presented an example of navigation strategy on ontology in order to identify meaningful ancestors for relevant terms and recommend them as new possible tags. Our future research will be addressed on two main objectives: (1) emphasize the aspects related to the personalization of the recommending process; (2) generalize the proposed methodology, in order to automate the extension of an ontology on the basis of user choices and preferences.

References

[1] A. Dattolo, F. Tomasi, and F. Vitali, "Towards disambiguating social tagging systems," in *Handbook of Research on Web 2.0, 3.0 and X.0: Technologies, Business and Social Applications*, S. Murugesan, Ed., vol. Chapter 27. IGI-Global, 2009.

[2] B. Sotomayor, "folk2onto: Mapping social tags into ontological categories," 2006. [Online]. Available: www.deli.deusto.es/Resources/Documents/folk2onto.pdf

[3] L. Tenenbaum, B. Shapira, and P. Shoval, "Ontology-based classification of news in an electronic newspaper," in *In Proc. of INFOS 2008*, 2008, pp. 89–97.

[4] G. Acampora, M. Gaeta, and V. Loia, "An ontological approach for memetic optimization in personalised e-learning scenarios," *Convergence Information Technology, International Conference on*, vol. 2, pp. 1204–1213, 2008.

[5] A. Adamo, L. Cafaro, V. Loia, C. Romano, and M. Veniero, "A multi-layered agent ontology system for resource inventory," in *Industrial Electronics, 2008. ISIE 2008. IEEE International Symposium on*, 30 2008-July 2 2008, pp. 2317–2322.

[6] G. Mishne, "Autotag: a collaborative approach to automated tag assignment for weblog posts," in *15th international conference on World Wide Web (WWW)*, 2006, pp. 953–954.

[7] T. Tatu, M. Srikanth, and T. D'Silva, "Rsdc'08: Tag recommendations using bookmark content," in *Proc. of ECML PKDD Discovery Challenge (RSDC08)*, 2008, pp. 96–107.

[8] C. Tasso and F. A. Asnicar, "ifweb: a prototype of user model-based intelligent agent for document filtering and navigation in the world wide web," in *Adaptive Systems and User Modeling on the WWW, 6th UM Inter. Conf.*, 1997.

[9] A. Baruzzo, P. Casoto, A. Dattolo, and C. Tasso, "Handling evolution in digital libraries," in *IRCDL '09: 5th Italian Research Conference on Digital Library Systems*, 2009.

[10] A. Baruzzo, A. Dattolo, P. Nirmala, and C. Tasso, "A general framework for personalized text classification and annotation," in *International Workshop on Adaptation and Personalization for Web 2.0 in connection with UMAP 2009, Trento, Italy, June 22-26*, 2009, pp. 31–39.

[11] N. Pudota, P. Casoto, A. Dattolo, P. Omero, and C. Tasso, "Towards bridging the gap between personalization and information extraction," in *IRCDL '08: 4th Italian Research Conference on Digital Library Systems*, 2008, pp. 33–40.

[12] E. Frank, G. Paynter, I. Witten, C. Gutwin, and C. Nevill-Manning, "Domain-specific keyphrase extraction," in *IJCAI '99: 16th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, 1999, pp. 668–673.

[13] M. Quillian, "Semantic memory," in *Semantic Information Processing*. MIT Press, 1968, pp. 227–270.