A Personalized Intelligent Recommender and Annotator TEStbed for Text-Based Content Retrieval and Classification: The PIRATES Project

Felice Ferrara and Carlo Tasso

Department of Mathematics and Computer Science, University of Udine
{felice.ferrara,carlo.tasso}@uniud.it

Abstract. This paper presents the PIRATES (Personalized Intelligent Recommender and Annotator TEStbed for text-based content retrieval and classification) Project. This project faces the information overload problem by taking into account semantic and social issues: an integrated set of tools allow the users to customize and personalize the way they retrieve, filter, and organize Web resources.

1 Introduction and Motivation

The tremendous volume of digital contents available on the Web generates the information overload problem: the task of filtering new contents appropriate to individual needs is hard. Moreover, the growing amount of user generated contents published by means of Web 2.0 environments (such as forums, blogs, and social networks) makes this problem even harder. Obviously, an effective classification of the Web resources can tackle this problem: such classification can be used to filter the resources which match the user interests. For this reason, a core function in a framework aimed at facing information overload is the classification module: more accurate is the classification of resources and more precise is the description of the user interests and better filtering performance can be obtained. However, such classification cannot be exploited by means of a manual activity (such as extracting small portions of relevant information from available contents, or classifying contents according to a specific model of user interests) due to the large amount of resources to be considered.

In order to overcome this limitation, Semantic Web and adaptive personalization technologies have been proposed: the classification and matching processes do not involve human intervention since a semantic layer is automatically added in order to classify resources. Ontologies are one of the main Semantic Web tools able to associate a clear semantic to Web resources. However, ontologies are domain dependent and, for this reason, it is quite difficult to integrate these technologies in domain independent frameworks. On the other hand, in social tagging systems the classification task is performed by users: each user of a Web 2.0 site can freely choose a set of terms, called tags, in order to classify Web

M. Agosti et al. (Eds.): IRCDL 2011, CCIS 249, pp. 136–139, 2011.

resources. Obviously, such a domain independent approach cannot be rigorous since users do not have to respect specific rules and consequently tags often do not have a clear semantic meaning.

So shortcomings of Semantic Web technologies can be faced by using socially defined classifications, and vice versa, user generated classifications can be supported by Semantic Web tools in order to produce more meaningful classifications. Following this vision, we have proposed an experimental testbed (called PIRATES) to merge social and semantic technologies in order to enhance the access to the available knowledge available on the Web. More specifically the PIRATES integrates several tools in order to support the users in the following tasks:

- **Classifying resources**. Given an input document, several strategies are used to capture its meaning: extracting keyphrases from the specific document, browsing ontologies in order to find more abstract relevant concepts, and exploiting the collective intelligence provided by Web 2.0 users.
- **Finding relevant contents and people**. Web resources are crawled (by a set of software agents) and classified by adaptively taking into account the specific user interests. Moreover, in order to identify relevant resources for a specific information need, both content-based approaches and collaborative filtering strategies are exploited.

PIRATES is a general framework aimed at providing support in many different scenarios: in PIM (Personal Information Management), for supporting the identification of relevant Web contents in a personalized way; in E-Learning for supporting the tutor and teacher activities for monitoring (in a personalized fashion) student performance, behavior, and participation; in knowledge management contexts (including for example scholarly publication repositories and, more in general, digital libraries) for supporting document filtering and classification and for alerting users in a personalized way about new posts or document uploads relevant to their individual interests; in online marketing for monitoring and analyzing the blogosphere where word-of-mouth and viral marketing are nowadays more and more expanding and where consumer opinions can be listen.

2 The PIRATES Framework

Figure 1 shows the general architecture of the PIRATES framework.

A set of software agents are used for crawling Web resources as well as other meaningful information provided by Web 2.0 users. Web resources are then classified/labeled by means of a set of tools:

- the IEM (Information Extraction Module), based on the GATE platform, extracts named entities, adjectives, proper names, etc. from input documents;
- the KPEM (Key-Phrases Extraction Module) [1] extracts meaningful keyphrases which summarize each input document;

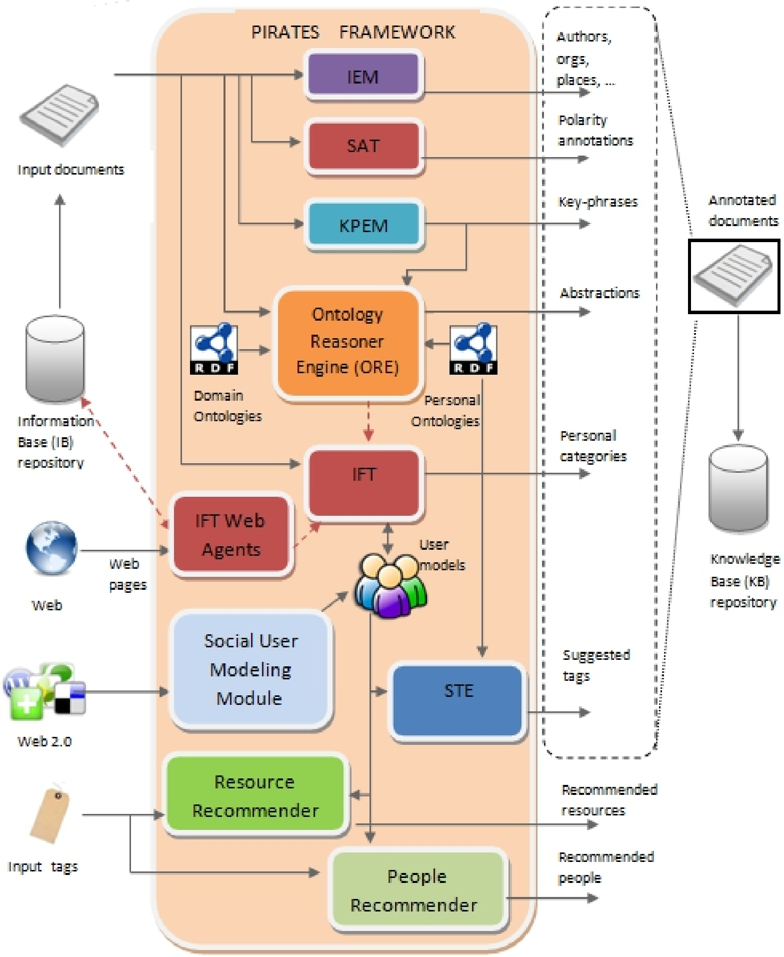

Fig. 1. The general architecture of PIRATES

- the IFT (Information Filtering Tool) [2] evaluates the relevance (in the sense of topicality) of a document according to a specific personalized model of user interests represented with semantic (co-occurrence) networks;
- the STE (Social Tagger Engine) suggests new annotations for a document relying on the tags generated by Web 2.0 users: social applications (such as delicious, BibSonomy, etc.) are also monitored in order to model the behavior of Web 2.0 users. The personal interests of each user are inferred by taking into account the set of resources that he/she tagged [3].
- the ORE (Ontology Reasoner Engine) [1] suggests more abstract concepts by browsing through ontologies, classification schemata, thesauri, lexicon (such as WordNet) and by using information extracted by the IEM, KPEM, IFT, and STE modules.
- the SAT (Sentiment Analysis Tool) [4] is a specific plug-in for personalized sentiment analysis that is capable of mining consumer opinions in the blogosphere;

PIRATES is also capable to recommend new potentially relevant contents and to identify people with interests similar to the user. For this purpose, PIRATES includes:

- the Resource Recommender module which filters resources according to an analysis of tags and resources considered by the users;
- the People Recommender module, which identifies people which share specific interests with the user.

3 Conclusions

The development of PIRATES is ongoing and has been planned in an incremental fashion, interleaved with several experimental evaluation phases. In fact, several modules has been already developed, integrated, and tested and the first experiments show that the proposed framework is a promising approach to automatic, personalized classification of Web contents.

References

1. Pudota, N., Dattolo, A., Baruzzo, A., Ferrara, F., Tasso, C.: Automatic keyphrase extraction and ontology mining for content-based tag recommendation. International Journal of Intelligent Systems, Special Issue: New Trends for Ontology-Based Knowledge Discovery 25, 1158–1186 (2010)
2. Minio, M., Tasso, C.: User modeling for information filtering on internet services: Exploiting an extended version of the umt shell. In: 5th UM Inter. Conf. UM for Information Filtering on the WWW (1996)
3. Ferrara, F., Tasso, C.: Extracting and Exploiting Topics of Interests from Social Tagging Systems. In: Bouchachia, A. (ed.) ICAIS 2011. LNCS, vol. 6943, pp. 285–296. Springer, Heidelberg (2011)
4. Casoto, P., Dattolo, A., Tasso, C.: Sentiment classification for the italian language: A case study on movie reviews. Journal of Internet Technology 9, 365–373 (2008)