A General Framework for Personalized Text Classification and Annotation*

Andrea Baruzzo, Antonina Dattolo, Nirmala Pudota, and Carlo Tasso

University of Udine, Via delle Scienze 206, 33100 Udine, Italy
{andrea.baruzzo, antonina.dattolo, nirmala.pudota, carlo.tasso}@dimi.uniud.it

Abstract. The tremendous volume of digital contents available today on the Web and the rapid spread of Web 2.0 sites, blogs and forums have exacerbated the classical information overload problem. Moreover, they have made even worse the challenge of finding new content appropriate to individual needs. In order to alleviate these issues, new approaches and tools are needed to provide *personalized* content recommendations and classification schemata.
This paper presents the PIRATES framework: a Personalized Intelligent Recommender and Annotator TEStbed for text-based content retrieval and classification. Using an integrated set of tools, this framework lets the users experiment, customize, and personalize the way they retrieve, filter, and organize the large amount of information available on the Web. Furthermore, the PIRATES framework undertakes a novel approach that automates typical manual tasks such as content annotation and tagging, by means of personalized tags recommendations and other forms of textual annotations (e.g. key-phrases).

1 Introduction

In the context of Semantic Web and Web 2.0 environments, finding an appropriate content is regarded not only as a problem of information overload but also as a problem of Web personalization [1], which deals with personalizing content retrieval and access with respect to a specific user model. Moreover, this large volume of data makes impractical or even impossible several manual activities such as extracting *small* portions of relevant information from available contents, or classifying contents according to a specific model of user interests [2]. As a consequence, the gap between the performance of traditional information retrieval tools (e.g. search engines) and the user satisfaction in their use continues to grow. In order to alleviate this issue [3], more sophisticated approaches and tools become necessary for providing *personalized* content recommendations and classification. Furthermore, in a world of collaborative publishing we have to take into account e-Learning, knowledge management and Web 2.0 as typical

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

application environments. Indeed, we can discover new relevant information by looking the *community* of people that, for example, share a common set of documents or use the same tags to label them. In this wider setting, automatic text classification remains a significant research field with several challenges such as:

- *Associating rich and precise semantics to information contents*. For describing an object, people tend to assign to it a very small number of tags, based on their knowledge background; of consequence, same tags, used by different users, do not share a common semantics [4, 5].
- *Adapting information retrieval strategies to an evolving user model*, providing run-time malleability to end-users [6]. Certainly, continuously updating a user profile is more difficult than building a single static representation, and requires the availability of some forms of user feedback to keep synchronized the model.
- *Finding relationships between contents and using a uniform method to share and reuse tagging data* amongst users or communities [7]. The topicality criteria alone may not be sufficient to relate contents when there is no shared semantics for a tag.

Our main goal in building the PIRATES framework is to empower social bookmarking tools, allowing users to easily add new contents in their personal collection of links, automatically supporting them when categorizing by means of keywords (tags) in a personalized and adaptive way. This work is a first step towards the generation and sharing of personal information spaces described in [8]. We have designed PIRATES keeping in mind several applications where it can provide innovative adaptive tools enhancing user capabilities: 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 digital libraries [9]) 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.

The paper is organized as follows: Section 2 illustrates the overall architecture and operation of PIRATES; Section 3 describes a typical interaction session and Section 4 concludes the paper.

2 The PIRATES framework

PIRATES (Personalized Intelligent Recommender and Annotator TEStbed) is a general framework for text-based content retrieval and categorization and exploits social tagging, user modeling, and information extraction techniques. Rather than proposing a rigid classification toolset, we have developed a testbed platform for integrating (and experimenting with) various tools and techniques, providing an interactive environment where users can customize the way they

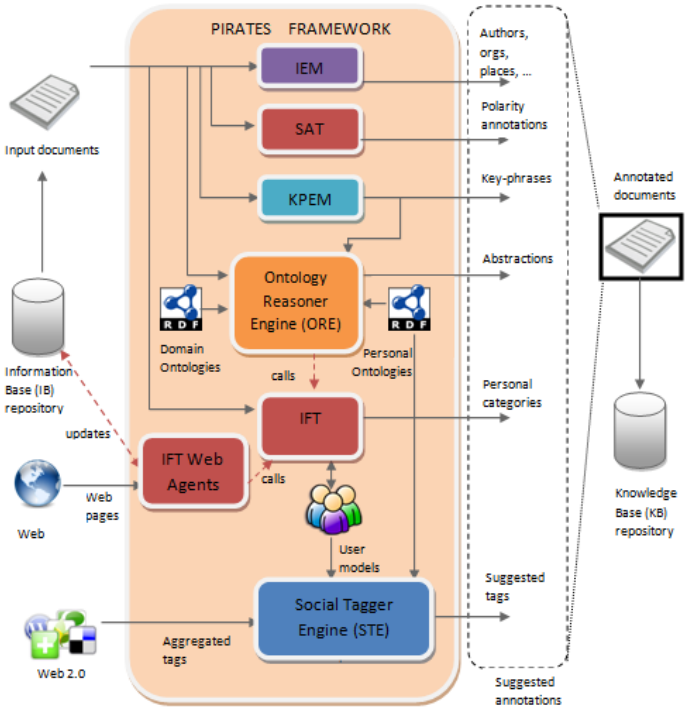


Figure 1. Overall architecture of PIRATES.

retrieve and classify information on the Web. 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). The framework operates on a set of input documents stored in the Information Base (IB) repository and suggests for these some personalized tags and other forms of textual annotations (e.g. key-phrases) in order to classify them. The original documents are then annotated with these tags, forming the Knowledge Base (KB) repository. Personalization is achieved exploiting user profiles (which represent the user interests), personal ontologies, personal tags, etc., as discussed in Section 3. Furthermore, PIRATES provides several mechanisms of user feedback that helps to provide personalized adaptive information.

The PIRATES architecture is illustrated in Figure 1. On the left-hand side, all the possible input sources are shown: single textual documents, specific IB repositories which can be contained within an e-learning knowledge management environment, and the Web, with specific (but not exclusive) focus on Web 2.0 portals, social networks, etc.. The right-hand side shows the suggested annotations and the resulting KB repository. The main modules of PIRATES are:

- `IEM` (*Information Extraction Module*), which is based on the GATE platform [10] to extract named entities, adjectives, proper names, etc. from input documents, contained in the IB.
- `SAT` (*Sentiment Analysis Tool*), which is a specific plug-in for personalized sentiment analysis (typically to be activated for online marketing applications), that is capable of mining consumer opinions in the blogosphere and classify them according to their polarity (positive, negative, or neutral) [11].
- `KPEM` (*Key-Phrases Extraction Module*), 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 each input document. This information is provided to the user, and is also given as input to the subsequent modules.
- `ORE` (*Ontology Reasoner Engine*), which suggests new *abstract concepts* by navigating through ontologies, classification schemata, thesauri, lexicon (such as WordNet), etc. An abstract concept is identified by looking for a match between the annotations found by the other modules (IEM, KPEM, IFT, and STE) and the concepts stored in ontologies. When a match is found, ORE navigates through the ontology, looking for the common parent node which represents the more abstract term to suggest as annotation. ORE also assists users in creating personal ontologies with techniques similar to those described in [13].
- `IFT` (*Information Filtering Tool*), which evaluates the relevance (in the sense of topicality) of a document according to a specific model of user interests represented with semantic (co-occurrence) networks [14].
- `IFT Web Agents`, which continuously monitor the Web (and the blogosphere) looking for new information, cooperates with IFT to filter contents according to the user model, and updates the IB repository. IFT and its Web agents form together the Cognitive Filtering module discussed in [8].
- `STE` (*Social Tagger Engine*), which suggests new annotations for a document relying on *aggregated tags*, i.e. the user's personal tags (tags previously exploited) and the more popular tags used by the community of people that classify the same document in social bookmarking sites such as Del.icio.us[1], Faviki[2] or Bibsonomy[3]. This social information is integrated with content-based analysis techniques as discussed in [15].

3 A typical usage scenario

In this section we provide a typical scenario that illustrates a use case for our framework. Consider a user interested to read scientific publications in the area of software engineering. He trains the IFT tool providing the training data (e.g. 2-3 relevant papers in the field, some keywords and a short textual description for the argument) in order to setup the user model. After training, the IFT

[1] http://delicious.com
[2] http://www.faviki.com/pages/welcome/
[3] http://bibsonomy.org

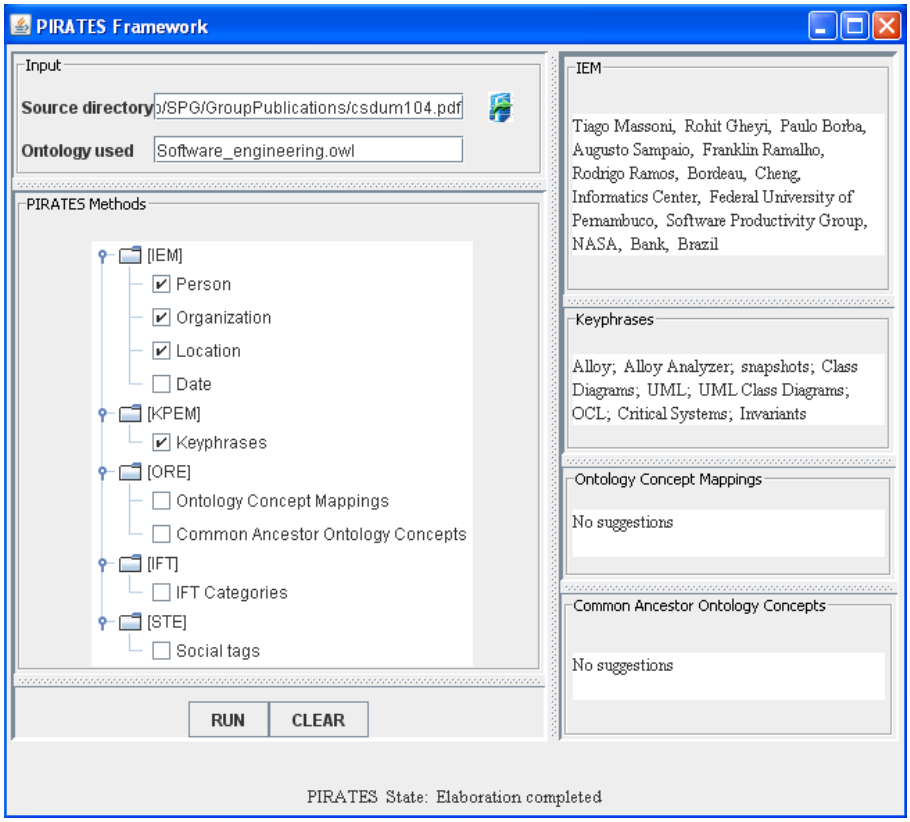


Figure 2. The PIRATES user interface running our example

agents periodically monitor the Web (in our case especially Web 2.0 sites such as Del.icio.us, Bibsonomy, CiteseerX[4], etc.), download new content and scrap selected data from them to filter out irrelevant information (e.g. ads and navigational links). When a relevant content (with respect to the user model) is retrieved, the agents add it to the IB repository and informs the user with a notification (e.g. an e-mail message). This information retrieval workflow has been already discussed in [14, 16], so in the rest of the section we concentrate on the classification features added by the PIRATES framework. Indeed, PIRATES aims expressly to support the user in *organizing* the IB repository, easing the work of classifying new contents by means of personalized tag suggestions.

Suppose now that an IFT agent notifies (among the others) the paper "A UML Class Diagram Analyzer"[5]. In order to classify this new content, the user can enable some PIRATES annotator modules, as illustrated in the left side of

[4] http://citeseerx.ist.psu.edu/
[5] http://twiki.cin.ufpe.br/twiki/pub/SPG/GroupPublications/csduml04.pdf.

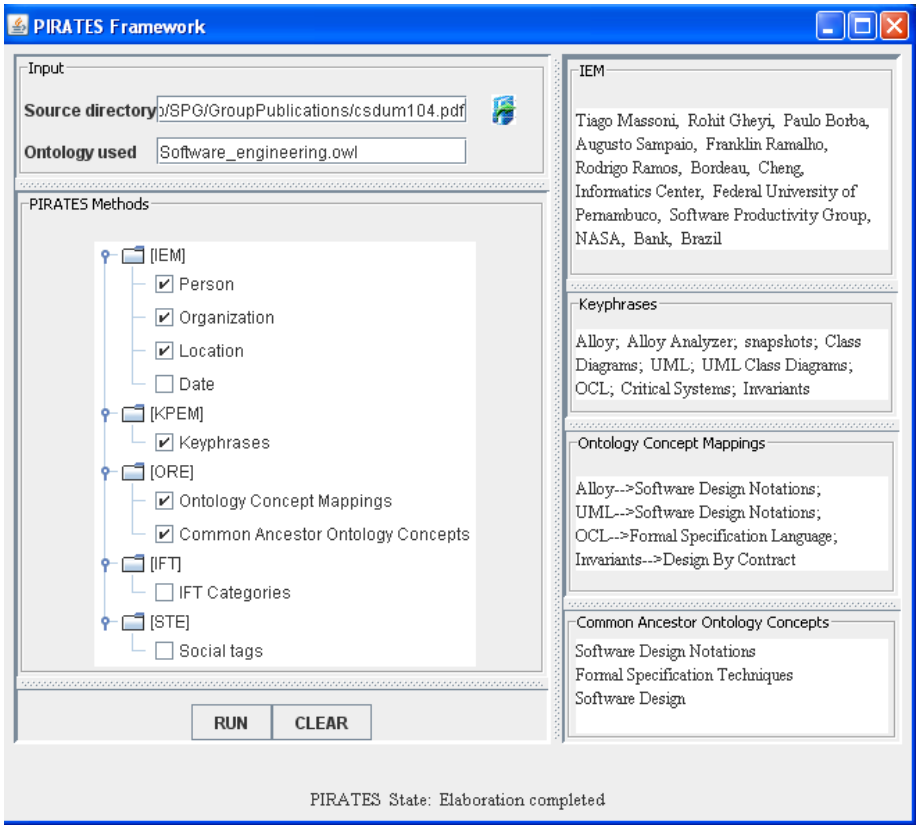


Figure 3. The PIRATES user interface running our example

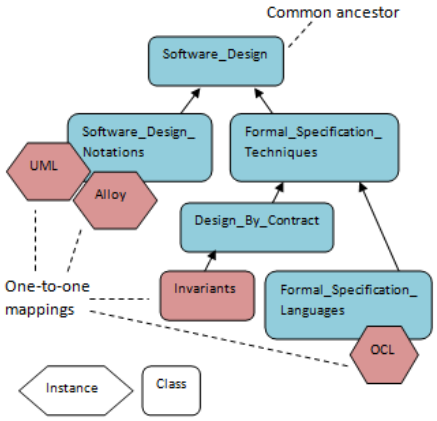
Figure 2. Let us assume that he enables only IEM and KPEM modules in order to extract, respectively:

- person's names, organizations, and places (using IEM);
- keyphrases, i.e. n-grams long three terms at maximum (using KPEM).

With these settings, the framework produces the tag recommendations showed in the right side of Figure 2. In particular, the suggested tags concern persons such as the authors (`Tiago Massoni`, `Rohit Gheyi`, and `Paulo Borba`) and the people acknowledged in the paper (`Bordeau`, `Chang`, `Augusto Sampaio`, `Franklin Ramalho` and `Rodrigo Ramos`), locations (`Brazil`), and organizations cited in the text (the `Informatics Center` of the `Federal University of Pernambuco`, the `Software Productivity Group`, and the `NASA`). As keyphrases, KPEM provides many terms related to Alloy specification language (`Alloy`, `Alloy Analyzer`, `snapshots`), to UML (`UML`, `UML Class Diagrams`, `OCL`) and to the specification of dependable systems (`Critical Systems`, `Invariants`).

Tags suggested by IEM + KPEM + ORE modules	
IEM	***Orgs***: Federal University of Pernambuco, Informatics Center, Software Productivity Group, NASA ***Places***: Brazil ***People***: Paulo Borba, Tiago Massoni, Rohit Gheyi, Franklin Ramahlo, Rodrigo Ramos, Augusto Sampaio
KPEM	***Keyphrases***: Alloy Alloy Analyzer UML OCL Critical Systems UML Class Diagrams Invariants Snapshots
ORE	***Ontology tags***: Software Design Notation, Formal Specification Language, Design by Contract, Formal Specification Techniques, Software Design

(a) IEM, KPEM, and ORE outputs



(b) Ontology reasoning

Figure 4. Personalized annotations proposed by PIRATES

The tag suggestions provided so far are extracted by the text present in the input document: no personalization is present at all. Suppose now that the user enables also the ORE module which exploits (in our example) a *personal* ontology[6] in the field of software engineering (see left side of Figure 3).

ORE implements a navigation strategy, taking in input the key-phrases extracted by other annotators (KPEM in this case). 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 (see Figure 4(b)). Starting from these nodes, ORE uses a spreading activation algorithm to find common ancestors representing more abstract subjects. Then both one-to-one ontology mappings and common ancestors are provided by PIRATES as potential tag recommendations, as summarized in Figure 4(a). The ontology navigation process highlighted by the spreading activation algorithm is depicted in Figure 4(b). In conclusion, the ORE module recommends five new tags which are not present in the text (i.e. `Software Design Notation`, `Formal Specification Language`, `Design by Contract`, `Formal Specification Techniques`, and `Software Design`)[7].

[6] We exploit an extended version of the existing domain ontology available from http://www.seontology.org/.

[7] Note also that tag `Design by Contract` was not already present nor in the input document, nor in the original ontology, but it was added to the ontology by means

These tags represent *abstractions* of the key-phrases extracted by the other annotators available in PIRATES.

4 Conclusions

We believe that the presented framework is a promising approach to automatic, personalized classification of Web contents. It is a first step in the direction of automatically organize document repositories into personal concept maps, moving from information to knowledge. The development of PIRATES has been planned in an incremental fashion, interleaved with experimental evaluation. Several modules have been already developed and integrated in a testbed environment: IEM with the sentiment analysis plug-in [16], KPEM with key-phrases extraction capabilities, and the Cognitive Filtering comprising an extended version of IFT capable to monitor Web 2.0 sources (specifically newsgoups, forums, and blogs). The integration of these modules is currently being evaluated. Prototyping and integration of ORE, SAT, and STE within PIRATES are ongoing processes, and evaluation experiments are planned. Moreover, we are working specifically on integrating the PIRATES modules in a Web-based version of the environment, which let us validate each module thoroughly. Finally, we have also planned to implement the conceptual map editor described in [8] in order to completely validate the framework.

References

1. Brusilovsky, P., Tasso, C.: Preface to special issue on user modeling for web information retrieval. User Modeling and User-Adapted Interaction **14**(2-3) (2004) 147–157
2. Gauch, S., Speretta, M., Chandramouli, A., Micarelli, A.: User profiles for personalized information access. In: The Adaptive Web. (2007) 54–89
3. Bunt, A., Carenini, G., Conati, C.: Adaptive content presentation for the web. In: The Adaptive Web. (2007) 409–432
4. Katakis, I., Tsoumakas, G., Vlahavas, I.: Multilabel text classification for automated tag suggestion. In: Proc. of the ECML/PKDD 2008 Discovery Challenge. (2008)
5. Marchetti, A., Tesconi, M., Ronzano, F.: Semkey: A semantic collaborative tagging system. (2007)
6. Lonchamp, J.: A platform for cscl practice and dissemination. In: ICALT '06: Proc. of the Sixth IEEE International Conference on Advanced Learning Technologies, IEEE Computer Society (2006) 66–70
7. Kim, H., Yang, S., Jung, J., Kim, K., Breslin, J., Decker, S., Kim, H.: Combining tags and the semanticweb for linked tagging data (2008)
8. Casoto, P., Dattolo, A., Ferrara, F., Pudota, N., Omero, P., Tasso, C.: Generating and sharing personal information spaces. In: Proc. of the Workshop on Adaptation for the Social Web, 5th ACM Int. Conf. on Adaptive Hypermedia and Adaptive Web-Based Systems. (2008) 14–23

of a user feedback mechanism provided by PIRATES. This is where personalization comes from.

9. Omero, P., Polesello, N., Tasso, C.: Personalized intelligent information services within an online digital library for medicine: the bibliomed system. In: IRCDL '07: Proc. of the Third Italian Research Conference on Digital Library Systems. (2007) 46–51
10. Cunningham, H.: Gate, a general architecture for language engineering. Computers and the Humanities **36** (2002) 223–254
11. Casoto, P., Dattolo, A., Tasso, C.: Sentiment classification for the italian language: A case study on movie reviews. Journal of Internet Technology **9**(4) (2008) 365–373
12. Frank, E., Paynter, G., Witten, I., Gutwin, C., Nevill-Manning, C.: Domain-specific keyphrase extraction. In: IJCAI '99: Proc. of the Sixteenth International Joint Conference on Artificial Intelligence, Morgan Kaufmann (1999) 668–673
13. Speretta, M., Gauch, S.: Using text mining to enrich the vocabulary of domain ontologies. Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on **1** (2008) 549–552
14. Tasso, C., Asnicar, F.A.: 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)
15. Tasso, C., Rossi, P., Virgili, C., Morandini, A.: Exploiting personalization techniques in e-learning tools. In: SW-EL'04: Proc. of the Workshop on Applications of Semantic Web Technologies for Adaptive Educational Hypermedia. (2004)
16. Pudota, N., Casoto, P., Dattolo, A., Omero, P., Tasso, C.: Towards bridging the gap between personalization and information extraction. In: IRCDL '08: Proc. of the Forth Italian Research Conference on Digital Library Systems. (2008) 33–40