Customizing information access according to domain and task knowledge : the OntoExplo System

Nathalie Hernandez
IRIT
118 Route de Narbonne
31062 Toulouse Cedex 4 France
0561556899, +33
hernandez@irit.fr

Josiane Mothe
IRIT
118 Route de Narbonne
31062 Toulouse Cedex 4 France
0561556444, +33
mothe@irit.fr

Sandra Poulain
IRIT
118 Route de Narbonne
31062 Toulouse Cedex 4 France
0561557436, +33
poulain@irit.fr

ABSTRACT
In this paper we present a system that allows a user to explore or mine a document collection. This system is based on domain and task knowledge modelled in the form of ontologies and allows direct access both to information as it is stored and to information that is built from it. The system has been developed in Java.

Categories and Subject Descriptors
H3.3 [Information Storage And Retrieval]: Information Search and Retrieval; H.3.1 [Content Analysis and Indexing] – *Search process*, Retrieval models.

General Terms
Design

Keywords
content domain ontology, task domain ontology, interface, corpus exploration, information retrieval, information mining

1. INTRODUCTION
More and more, the answer to an information need cannot be found simply in a document or a set of documents, but has to be extracted or even created from a variety of resources. In this case sophisticated information management systems are necessary to provide users with global views of the available information and the structure of the domain (knowledge discovery and information mining systems). These tools should help them find the nuggets and knowledge that are hidden in these masses of data they potentially have access to. The system we present aims at providing such types of information from scientific publications.

2. INFORMATION REPRESENTATION : DOMAIN ONTOLOGIES
The common representation of documents is the so-called "bag of words" representation in which the content of the document only is considered. However, it is now acknowledged that the document content is not the only important component for a user. Many other document facets can be of interest depending on the user's task [1]. Although meta-data play a key role in document description, current Information Retrieval Systems do not use them intensively. We promote an approach in which documents are represented according to two types of domain knowledge: content and task, both modelled in the form of ontologies [3].

SIGIR'05, August 15–19, 2005, Salvador, Brasil.

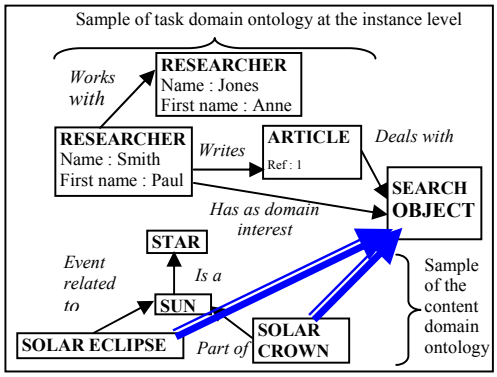

Figure 1 : Link between the content and task ontologies for Astronomy

Content is represented according to a domain-ontology representing the domain treated in the documents. Contrary to ontologies or thesauri that represent the entire world (e.g. WordNet); we promote an approach in which an ontology is devoted to a single domain (e.g. Mesh). In our application in which the domain is Astronomy, this ontology is built from the IAU thesaurus[1] (see lower part of Figure 1). To each concept is associated a set of terms (or labels) that can be used to represent this concept. Different relationships are also represented such as 'is a', 'part of', and 'event linked to'.

Task is represented by a second ontology. Its content depends on the type of information users want to extract or discover. It organises the meta-data associated to documents according to the role they will play in the task the user is carrying out. In the application that is presented in this poster, the task ontology corresponds to the extraction of the domain structure and to scientific monitoring. (see upper part of Figure 1).

[1] http://msowww.anu.edu.au/library/thesaurus

The two ontologies are linked to each other according to the corpus knowledge. 'Search object' is the concept that makes the link in our example. Instances of 'search object' of the task ontology are found in the content ontology by identifing the most representative concepts of the document set concerned (books, author's publications) according to a measure presented in [2].

The system developed on this model is implemented to assign documents to both ontologies. Regarding document and content ontology mapping, traditional document indexing is used. Documents in which a term corresponding to a concept occurs are associated to the concept using a weighted link (based on tf.idf weighting). Document and task ontology mapping is based on information extraction mechanisms. For example, IE technology is used to extract the author's names and affiliations from the publication. Once done, corresponding instances are created within the task ontology.

3. QUERYING THE COLLECTION

Query language is based on the concepts of the two ontologies and exploring the collection is on ontology browsing. The interface is presented in Figures 2 and 3 .Two scenarios are used: querying the task ontology or querying the content ontology.

- Querying the task ontology

As the task in our application is monitoring a scientific domain, a typical query is based on getting knowledge on the main authors or laboratories and their relationships. Such a query is based on browsing concepts from the task ontology. Figure 2 provides the results obtained once the user has selected the author "Stephano Cecchini". The interface is then automatically customized so that the user visualizes only the instances that are related to the selected one. In this example, co-authors of the selected author are displayed as well as the list of documents of which he is the author, and the organization to which he is affiliated. Instances of the search object concept are displayed in the left side of the windows. Colored concepts are the concepts about which the author has published. Concepts are displayed within their context (related concepts).

- Querying the content ontology

Alternatively, the user can decide to start with the content ontology. He can select a concept (different levels of generality can be chosen); related authors and publications are then displayed (See Figure 3).

4. CONCLUSION

Contrary to other systems dealing with document content and meta-data, the system proposed aims at separating the management of both, providing better reusability [4]. The ontologies proposed are based on richer semantic relations than concept hierarchies traditionally used [1]. The system we implemented is now being evaluated by astronomers in order to quantify the contribution of such semantic features for specific tasks. Improvements are being made to expand the meta-data extraction process (currently based on manually defined policies for each meta-data) and the indexing process (by taking into consideration the semantic relations).

5. REFERENCES

[1] Aussenac-Gilles N., Mothe J., Ontologies as Background Knowledge to Explore Document Collections, RIAO, pp 129-142, 2004.

[2] Hernandez N., Mothe J., An approach to evaluate existing ontologies for indexing a document corpus, AIMSA, Semantic Web Challenges, pp 11-21, 2004.

[3] Fensel D., Ontologies: a silver bullet for Knowledge Management and Electronic Commerce, Berlin, Springer Verlag, ISBN 3-540-00302-9, 2001.

[4] Stuckenschmidt, H.; van Harmelen, F.; de Waard, A.; Scerri, T.; Bhogal, R.; van Buel, J.; Crowlesmith, I.; Fluit, C.; Kampman, A.; Broekstra, J.; van Mulligen, E., Exploring large document repositories with RDF technology: the DOPE project, Intelligent system, Vol. 19, No. 3, pp 34- 40, 2004

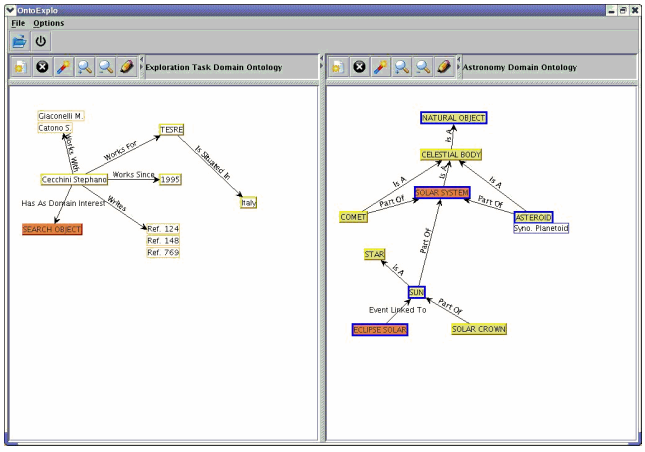

Figure 2: Result upon the receipt of the selection of author 'Stephano Cecchini'.

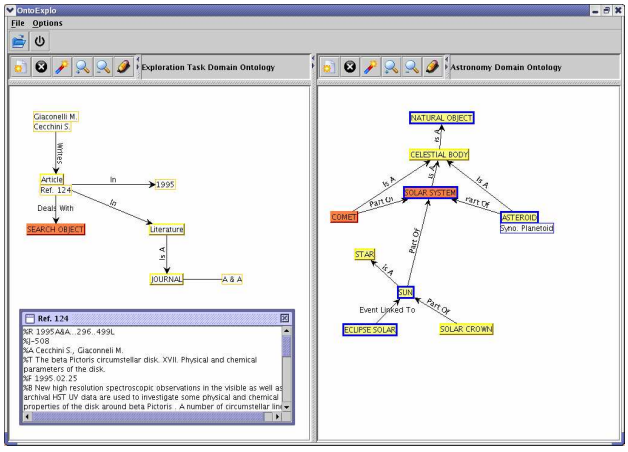

Figure 3: Result upon the receipt of the selection of a document dealing with 'Solar System'.