A SEMANTIC WEB APPROACH TO ENRICH INFORMATION RETRIEVAL ANSWERS

R. Carolina Medina-Ramírez and Víctor M. Ramos R.
UAM-Iztapalapa, San Rafael Atlixco 186, 09340 Iztapalapa, Mexico

Keywords: Semantic web, document retrieval.

Abstract: In previous works, we have presented the advantages of using a domain ontology and annotations for information retrieval (Medina-Ramírez et al., 2002b) as well as the translation problems between languages with different expression semantic levels (Medina-Ramírez et al., 2002a). In this paper we focus on the view point of the end-user. In fact, we explore the impact and helpfulness of a domain ontology, semantic annotations relying on this ontology and semantic resource descriptions so as to enrich end-user answers extracted from an information retrieval system. A system that embodies this approach is presented. We argue that it is necessary to improve the format of end-user answer in order to share and re-use knowledge.

1 INTRODUCTION

Huge amounts of heterogeneous data (structured data, semi-structured data, textual data, multimedia data), on the present Web is mainly addressed to human users of the Web. The Semantic Web (Berners-Lee et al., 2001; Shadbolt et al., 2006) aim to enable machines to understand, process and reason about resources in order to provide better and more comfortable support for humans. The semantic contents description of Web resources, also called semantic annotations, relying on ontologies contribute to reach this goal. They could be processed by automated tools.

In the last few years, a new generation of ontology-guided Information Retrieval systems have been proposed –SHOE (Luke et al., 1997), OntoBroker (Fensel et al., 1999), OntoSeek (Guarino et al., 1999), WebKB (Martin and Eklund, 2000), Corese (Corby et al., 2004), RDFQ (Gaspari and Guidi, 2003), Ontoweb (Kim, 2005)–. They focus on ontology knowledge representation languages and propose an ontology-guided retrieval of annotated documents.

The Web Community is invested in developing new semantic search techniques, but the question of improving the interaction with web content is at hand. Web users aim at retrieving resources or services satisfying specific criteria or constraints. They want to watch the retrieved resources in a personalized format. Particularly, results from desktop search engines are still limited.

On current widely deployed search engines, the result format is simple consisting of a set of lines describing the documents found which match a submitted query. This description is based on the keywords submitted in the query and appears in the retrieved document. Important information fields such as: document type (journals, proceedings or informal notes), publication date, author names, journal and conferences name is missing in a typical answer from the web. The presence of such information in the returned results is of relevant importance to select the pertinent document from a specific user query. A much richer expressiveness than simple keywords for describing resources is considered by the semantic web approach.

Languages like XML let us on one hand represent more explicitly the structure of electronic documents, and on the other hand they let us manage such information in an easy way. Specialized document bases such as PubMed (Database, 2002), use this approach.

We claim that an effort has to be made in the displaying of results to submitted queries for a better comprehension and transfer of knowledge.

Carolina Medina-Ramírez R. and M. Ramos R. V. (2007).
A SEMANTIC WEB APPROACH TO ENRICH INFORMATION RETRIEVAL ANSWERS.
In *Proceedings of the Ninth International Conference on Enterprise Information Systems - SAIC*, pages 299-302
DOI: 10.5220/0002409402990302

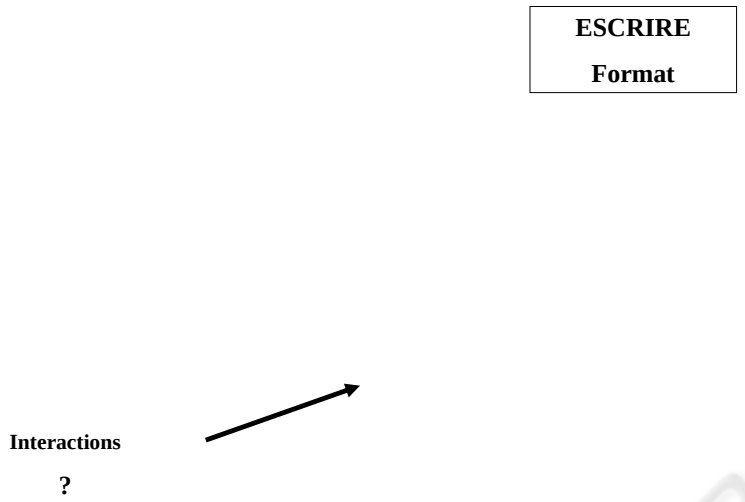

Figure 1: Escrire format.

2 ESCORSERVER ARCHITECTURE

The process of representing the meaning of documents in a better suited form to processing by computers (called semantic annotations) is of key importance in the semantic web approach.

The information retrieval needs on the Web are present in different scales in scientific communities, also called corporate Semantic Webs. The framework of the semantic Web can be applied on these communities in order to gain benefit from that approach. In particular, among the heterogeneous resources belonging, for example, to a scientific community or a company, documents (in electronic or paper form) constitute a significant source of knowledge needing to be represented, handled, queried and diffused.

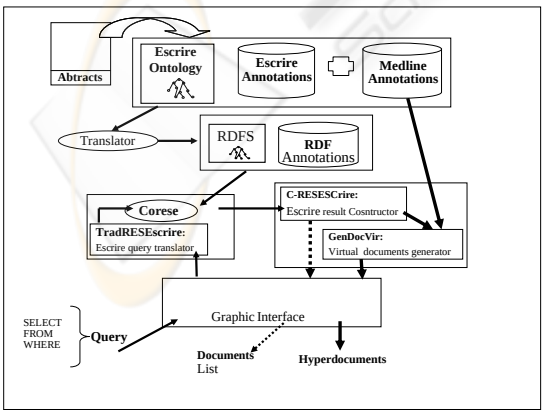

Figure 2: EsCorServerArchitecture.

The ESCRIRE project (R. et al., 2000) is focused on the context of corporate semantic webs. The aim of the ESCRIRE project is the representation and handling of document contents for document retrieval. In ESCRIRE, a test base composed of a set of 4500 abstracts of articles in biology from PubMed with semantic annotations on their contents has been used. The result format used to display the answers retrieved is a list of pertinent documents (similar to the major search engines) including also the query submitted by the user. This query is included in order to refresh the user's goals or information needs so as to help him/her to select from the list the best document match. Such format is presented in Figure 1.

In order to capitalize and diffuse the knowledge on genetic interactions in the documentary memory, we proposed EsCorServer. EsCorServer is a document server to handle, share and capitalize explicit knowledge (document content and data) from a specific domain (Drosophila melanogaster's gene interactions) for information retrieval. EsCorServer is based on an ontology-guided information retrieval, semantic annotations of domain articles abstracts, PubMed descriptions and adaptive hypermedia techniques. The heterogeneous aspects of this documentary memory reside on the nature of its resources and on its document content representation format. Figure 2 shows

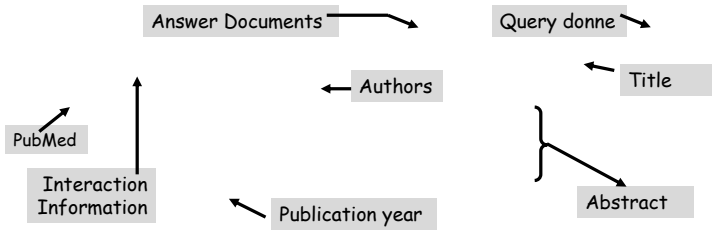

Figure 3: Enriched end-user approach.

the EsCorServer architecture.

Typical information retrieval systems present problems in the adequacy of queries like representations of informations needs. In order to improve these problems, an interface is provided in EsCorServer to help users to get from goals or information needs to structured queries. This interface charged of introducing, translating and displaying results from a query. The translation and retrieving mechanism are described in (Medina-Ramírez et al., 2002a; Medina-Ramírez et al., 2002b). The main element in EsCorServer is the *GenDocVir module* which generates on-demand the virtual interaction information document. We describe this document in Section 3.

Figure 4: Interaction information generated on-demand.

3 ENRICHED END-USER ANSWER APPROACH

There has been much experimental research focus in information retrieval on the processes of text representation, organization and comparison. They do not traditionally care about the individual searcher, preferring instead to focus on the development of global retrieval techniques rather than those adapted for the needs of the individual. We focus on providing a more personalized retrieval experience and on helping users to choose among retrieval alternatives. In order to achieve this goal, we use ontology and resources description to enrich the answer given to the user. We can easily access annotated information by exploiting the Corese semantic search engine (Corby and Faron-Zucker, 2002).

The enriched end-user answer approach shown in Figure 3 consists of creating a hyperdocument composed of the abstracts of documents retrieved by the Corese search engine. Besides, this hyperdocument includes links to additional documents: the original document in PubMed, the submitted query and the interaction informations created on-demand. The author's name, publication date, journal and PubMed identifier are included in the hyperdocument to provide additional useful information .

Our main contribution is an approach to integrate semantic descriptions of gene interactions (Escrire annotations) and the concepts of a domain on-

tology, in order to generate on-demand a virtual interaction information document. This document intends to clarify the concepts involved in the retrieved documents. Figure 4 shows the virtual interaction information document. This document contains specific information such as: genes description –scientific gene names, belonging family, activation or inhibition effects, participant genes names of interactions mentioned in the article.

The challenge is to create such document by adding semantic resource descriptions according to user interests, so as to display them in a manner that facilitates exploration and encourages the user. More specific learning scenarios and profiles must improve the adequacy between the annotation contents and the end-user requests.

4 CONCLUSION AND FUTURE WORK

The innovative aspect of the approach described in this paper and the contribution to the field of adaptive hypermedia documents is the merging of different resource descriptions. This approach provides robustness to end-user answers to a query as well as a way of accessing information annotated. This is done by exploiting the Corese semantic search engine.

From the experience got from this work, we believe that manual annotation of resources is overwhelming to domain experts or teachers when they face a large amount of resources. So, it is imperative to automate as much as possible the extraction of knowledge from structured format documents.

Preliminary evaluations of our prototype have produced encouraging results. So, our future work will focus on an extension of our prototype to analyze additional results on this direction. We would like to apply information filtering techniques in a corporate semantic web. In particular, we are interested in implementing profiles as a correct specification of information interest. These profiles are built based on repeated uses of the system, by a person or persons with long-term goals or information needs.

REFERENCES

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American*, 284(5):35–43.

Corby, O., Dieng-Kuntz, R., and Faron-Zucker, C. (2004). Querying the semantic web with corese search engine. In *Proceedings of the 16th Eureopean Conference on Artificial Intelligence, ECAI'2004, including Prestigious Applicants of Intelligent Systems, PAIS 2004, Valencia, Spain, August 22-27, 2004*, pages 705–709.

Corby, O. and Faron-Zucker, C. (2002). Corese: A corporate semantic web engine. In *Proceedings of the WWW2002 Workshop on Real World RDF and Semantic Web Applications, Honolulu, Hawai, USA*.

Database, M. (2002). http://www.ncbi.nlm.nih.gov/PubMed.

Fensel, D., Angele, J., Decker, S., Erdmann, M., Schnurr, H.-P., Staab, S., Studer, R., and Witt, A. (1999). On2broker: Semantic-based access to information sources at the WWW. In *Proceedings of the World Conference on the WWW and Internet: WebNet*, pages 366–371.

Gaspari, M. and Guidi, D. (June 2003). Towards an ontology-guided search engine. *tecnical Report UBLCS-2003-8*.

Guarino, N., Masolo, C., and Vetere, G. (1999). Ontoseek: Content-based access to the Web. *IEEE Intelligent Systems*, 14(3):70–80.

Kim, H. H. (2005). Ontoweb: Implementing an ontology-based web retrieval system. *Journal of the American Society for Information Science and Technology*, 56(11):1167–1176.

Luke, S., Spector, L., Rager, D., and Hendler, J. (1997). Ontology-based web agents. In *Proceedings of the First International Conference on Autonomous Agents*, pages 59–68.

Martin, P. and Eklund, P. W. (2000). Knowledge retrieval and the World Wide Web. *IEEE Intelligent Systems*, 15(3):18–25.

Medina-Ramírez, C., Corby, O., and Dieng-Kuntz, R. (2002a). A conceptual graph and RDF(S) approach for representing and querying document content. In *Advances in Artificial Intelligence-IBERAMIA 2002, 8th Ibero-American conference on AI. Ganjo Francisco J., Riquelme J. Cristóbal., Toro M. (Eds.). LNCS 2527,Seville, Spain.*, pages 121–130.

Medina-Ramírez, C., Corby, O., and Dieng-Kuntz, R. (2002b). Querying a heterogeneous corporate semantic web: A translation approach. In *Proceedings of the international workshop on "Knowledge Management through Corporate Semantic Webs". During the EKAW conference, Singüenza, Spain.*, pages 53–63.

R., A.-H., Corby, O., Dieng-Kuntz, R., Medina-Ramírez, J. E. R. C., Napoli, A., and Troncy, R. (2000). Three knowledge representation formalisms for content-based manipulation of documents. In *Proceedings of the KR 2002 Workshop on Formal Ontology, Knowledge Representation and Intelligent Systems for the World Wide Web (Semweb)*.

Shadbolt, N., Berners-Lee, T., and Hall, W. (2006). The semantic web revisited. *IEEE Intelligent Systems*, 21(3):96–101.