A Semantic Web Approach to Represent and Retrieve Information in a Corporate Memory

Ana B. Rios-Alvarado, R. Carolina Medina-Ramírez, and Ricardo Marcelín-Jiménez

Department of Electrical Engineering
Universidad Autónoma Metropolitana - Iztapalapa
México, DF

Abstract. This paper outlines the first stage of the ODARyT Tool, a semantic web approach to represent and retrieve information in a corporate memory. After a brief discussion about the meet points between semantic web and corporate memories, we describe our semantic approach related to a specific corporate memory located in the Networks and Telecommunication domain. The corporate knowledge is represented in an OWL Ontology, which provides an intuitive way to create annotations of heterogeneous resources composing a Corporate Semantic Web.

1   Introduction

The semantic Web is an extension of the current web in which information is given a well-defined meaning, so as to be accessible and understandable not only by humans but also by computers, thus enabling an extended cooperation among people and machines [1]. This approach relies on ontologies (information exchange and search), semantic annotations (document content representation) and formal knowledge representation languages (for representing these ontologies and annotations). In addition, among the heterogeneous resources belonging, for example to a scientific community or to an enterprise, documents (in digital or paper supports) represent a significant source of knowledge requiring, storage, handling, querying and propagation. In order to better exploit an information repository, we need to understand how its size, topology, and content are evolving [2]. This allows the development of techniques for representing, storing, locating, retrieving information and personalizing the presentation of content resources not only according to user's requirement and/or preferences, but also considering the content of resources [3].

The motivation of our approach is to improve semantic understanding of heterogeneous web resources. This goal can be achieved through a Corporate Semantic Web (CSW)[4]. In this paper we present ODARyT a tool for editing-browsing an ontology and annotating heterogeneous resources composing to a CSW. The rest of this paper is structured as follows. In Section 2, we briefly present the relationship between a CSW and Semantic Web. Sections 3, describe

our approach for representing and retrieving information in a corporate memory. This approach focus on the ODARyT tool. In Section 4, we present some concluding remarks as well as some directions of our work.

2 Corporate Memory

A corporate memory is an explicit, disembodied and persistent representation of knowledge and information in an organization, in order to facilitate their access, share and reuse by members of the organization, for their tasks. In the context of the Semantic Web, there is a meet point between Web and corporate memories: both gather heterogeneous and distributed information and share the same concern about the relevance of information retrieval. Nevertheless, corporate memory has a context, an infrastructure and a scope limited to the organization where they are applied. Formally, a CSW is built up from ontologies, resources (documents or humans) and semantic annotations on these resources (on the document contents or the person features/competences), where these annotations rely on the ontologies[4] .

3 ODARyT Description

In order to handle the resources of a CSW, we need a tool with the following characteristics: *Ontology Edition* (concept descriptions: add, modify and erase), *Browsing* through the ontology and *deployment of definitions*,*annotations Edition* and a *query language* for annotations. A domain ontology and annotations are represented by specific languages that provided the formal syntax required for their representation. In our approach, we use a domain ontology (Networks and Telecommunications) represented in OWL and an annotation database represented in RDF language. This domain ontology is available on demand and it is composed of 315 concepts organized in four branches: *Networks and Telecommunication Services,Digital Communications Systems, Distributed Systems and Documents*.

There are some Integrated Development Environments, for developing Semantic Web applications. For example: Jena [5], Sesame [6], Sewese [7]. Sewese provides reusable, configurable and extensible primitives and components in order to reduce the amount of time spent to develop new semantic web applications.

We propose ODARyT Tool, a web application which provides a mechanism to browse and edit an ontology, edit and query an annotation database of resources composing a corporate memory. It consists of a set of related JSPs (editAnnotation,browseOntology, among others) to handle aspects of presentation, control and business processes of the application. We have used Sewese for building ODARyT interfaces: Ontology-annotation editor and annotation query. Fig. 1 shows the main architectural elements of the currently implemented ODARyT tool and their interactions. In following Sections, we describe these interfaces.
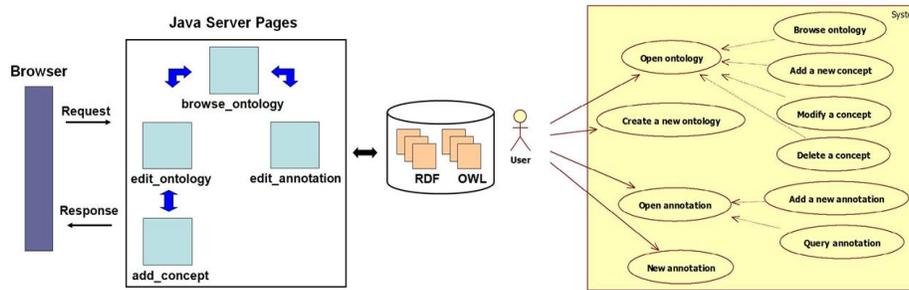
Fig. 1. ODARyT architecture and Use Cases

3.1 Ontology browser

This interface contains a main menu with options: New Ontology, Open Ontology, and Open New Annotation. The New Ontology option creates a new ontology, and only requires the name for the new ontology. The Open Ontology option displays in a list of links, the ontology concepts, and allow us to navigate through these concepts, showing for example: its definition, its parents and its child concepts. There are others ontology editors like Protege, NeOn, or Top-Braid which are used by ontology experts, the main advantage that ODARyT tool offers is to query and to locate the resources annotated by non ontology experts.

3.2 Ontology and Annotations Editor

The Editor interface contains three sections: Add, Edit and Delete a concept. In the Add Concept option, the user can create a new concept belonging to the ontology by introducing: ID (identification number), name, parent (parent concept ID), comment (definition of a concept). In the Edit concept option, the user should indicate the ID for the concept, that she/he wants modify. Some elements to introduce are the new concept name, new parents or new comment. The annotation authoring process is assisted by the use of a controlled vocabulary (ontology).The user can introduce the information about the resource, which could be: a) Annotation name, b) Author name, c)Onto keywords (ontology concepts), d) Other keywords, e) File name corresponding to the resource to be annotated , f) format (PDF, word, txt, etc), and g) year. The annotation include concepts selected from the ontology navigation area. At the moment,concepts are linked to annotations in a manual way. The non-expert user have the ontology as a guide to use these concepts for describing heterogeneous resources. The annotations are useful in a document retrieval scenario.

3.3 Annotation Database and Queries

This interface contains a section, where the user should introduce a query in SPARQL syntax, choose a display format and retrieve the relevant resources

from the CSW. The query result could be shown in different formats: in a table or in a list, according to the option selected. A set of style sheets is handled by the interface. ODARyT facilitates the tasks for editing, consulting, and navigating the CSW resources for experts and non-experts in the domain of Networks and Telecommunications. In this scenario, a user can contribute to increase the annotations database by introducing new semantic descriptions of resources: sound files, video, images, documents (html, XML, txt, etc). Finally, ODARyT tool could be use by other research groups from differents knowledge domains. However, for a common user we consider that *Annotation Database and Queries* interface could be improved by providing a translator from natural language to SPARQL.

4   Conclusions and Further Work

Organizations of any kind deal with the continuous growth of heterogeneous resources. Corporate knowledge management should enhance among enterprise members: learning, information interchange, improve information retrieval and information distribution. We have presented ODARyT a semantic web approach to represent and retrieve resources from a corporate memory. We evaluate our prototype within a representative group of experts as well as a group of non-experts in the domain of Networks and Telecommunications. Preliminary evaluations of our prototype have produced encouraging results and show that using the implicit information in the ontology and in annotations is well suited for needs of information retrieval and domain qualification. So, our future work will focus on work in an annotation authoring process assisted by offering concept-driven suggestions and appropriate help, based on a domain ontology. We are carrying out some improvements in our prototype such as: to hide technical issues related to SPARQL in the annotation database interface, we consider to exploit the OWL richness by defining transitivity and inverse relationships in the ontology proposed.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American. 2001
2. Levene,M., Poulovassilis, A.: Computer Networks Special Issue (Web Dynamics), vol 50, Issue 10, ISSN 1389-1286, 2006
3. Rios-Alvarado, A.B., Medina-Ramírez, R.C., Marcelín-Jiménez, R.: Fast semantic retrieval in a P2P corporate memory, WebMedia 2008, 26-29 Octubre, Vila Velha, ES, Brasil. ISBN 978-85-7669-199 0
4. Dieng-Kuntz, R.: Corporate Semantic Webs, ERCIM News No. 51, pp 19-21. 2002.
5. K. Wilkinson, C. Sayers, H. Kuno, D. Reynolds: Efficient RDF Storage and Retrieval in Jena2, HP Laboratories Technical Report HPL-2003-266
6. J. Broekstra, A. Kampman, and F. van Harmelen : Sesame: A generic architecture for storing and querying rdf and rdf schema. In The Semantic Web - ISWC 2002, volume 2342 of Lecture Notes in Computer Science, pages 54–68. Springer, 2002
7. Sewese http://www-sop.inria.fr/acacia/soft/sewese/, 2007.