ARCHITECTURE OF A SYSTEM FOR DOCUMENT RETRIEVAL USING SEMANTIC METADATA

Juan L. Dinos Rojas
Advisor: Dr. Fernando Vega Riveros

Electrical and Computer Engineering Department
University of Puerto Rico, Mayaguez Campus
Mayaguez, Puerto Rico 00680
Juan.Larry@ece.uprm.edu, fvega@ece.uprm.edu

ABSTRACT

In this article presents the architecture of a system for document or information retrieval using semantic metadata. The resource metadata are represented using RDF (Resource Description Framework) and RDF(S) (Resource Description Framework Schema). Both standards possess elements that allow the representation of concepts, their relationships and their attributes. Another important element within the investigation was the construction of an ontology that is used to construct the knowledge base.

To retrieve these resources, which could be located in different knowledge bases, a distributed agent-based architecture was designed. In this architecture, each agent is in charge of different tasks such as user interaction, ontology retrieval and document search. The Java Agent Development Framework (JADE) was used for implementation of agent-based architecture.

1. INTRODUCTION

Within a global market, the new competitive differential depends on the creation and application of knowledge. In a post-industrial society, knowledge and communications rather than natural resources and physical labor become the sources of wealth [Ungson99]; but knowledge has the peculiarity of not being worn out by being used; on the contrary, the more knowledge is used, the richer it gets and the more there is, since new insights may develop and new knowledge is likely to accumulate. As a consequence, a central task to augmenting wealth is not only the efficient use of scarce resources but also the encouragement of active cooperation for the generation of new knowledge [Adler89].

The current growth of the Internet has enabled access to very large volumes of information resources located in different and heterogeneous systems. Access to this information is realized through browsers which use HTML (Hypertext Markup Language) to display the information. The main disadvantage of HTML is that it does not allow for an adequate structuring of the information. Thus, the W3C has developed another alternative to code web information, to give more meaning to and reduce ambiguity from the information resources. These standards are RDF and RDF(S), which offer primitives for defining knowledge models that are closer to frame-based approaches [Gomez02]. RDF(S) is widely used as a representation format in many tools and projects.

In addition to technologies like RDF[Lassila99] and RDFS[Brickley99] to describe document metadata, other important technologies such as software agents present important features that can be used for document retrieval, e.g. autonomy, pro-activity, cooperation, etc. The vision of intelligent agents is very convincing and many people believe that these agents will become necessary as complexity within the World Wide Web grows [Hendler99]. These agents will help find information when provided with only some words to the search engines, achieving the desired results in a more effective way.

The advantages that can come from the combination of these technologies to the document retrieval are significant. In section 2 of this paper we will discuss the motivation and problem formulation. In section 3 a description of the ontology representation for the documents is given. In Section 4 the ontology representation using RDF and RDFS is presented. Section 5 describes the distributed agent-based architecture for document retrieval. Section 6 presents the conclusions.

2. PROBLEM FORMULATION AND MOTIVATION

Today technologies, initiatives and strategies such as the DCMI [Beckett02], RDF [Lassila99], and OIL [Horrocks00] exist or are being developed which allow identifying and describing knowledge and information resources. It is therefore important that knowledge management technologies and strategies be researched, developed and applied in education to prepare the future workforce for the new economic model and simultaneously enrich their learning environment.

The standards mentioned before such as the DCMI, RDF, RDFS, etc. are important to describe resources such as homework assignments, essays, reports and others used by students and professors. It is important to develop tools that also allow managing these resources and the ontologies constructed based on these standards.

This stated necessity generates a justifiable motivation for the development of a system to assist students and professors in storing and retrieving information resources in their learning/teaching process.

3. ONTOLOGY FOR THE DOCUMENT REPRESENTATION

The students consult different sources of information in their learning process such as elaborating their homework assignments and studying for their tests. At the present time many of these sources of information are in the internet or data bases in digital format, which facilitate their search and use.

When searching for a document the students resort to search engines that at the moment exist like Google, Altavista, etc., where keywords are used to look for information. As result a great amount of links to documents may be obtained. Within this search process the students generally enter related concepts or subjects that could identify more specifically documents or subjects that need to be found in the internet or data bases.

A suitable representation of these topics or concepts and the relations that can exist between these topics, would help to organize different documents and facilitate a more effective search. Ontologies "would allow us to represent the knowledge of the real world through related entities and objects" [Gruber93]. Through the ontologies we will be able to identify the concepts or topics that the students use, at the time of making the search of a document.

3.1 Ontology of documents

Ontologies play an important role in knowledge representation, which allows us to identify classes in a subclass-hierarchy. Each class is characterized by properties that are shared/inherited by all elements in that class [Lassila00]. Our investigation began with the identification of the main concepts and organized the different documents that the students use.

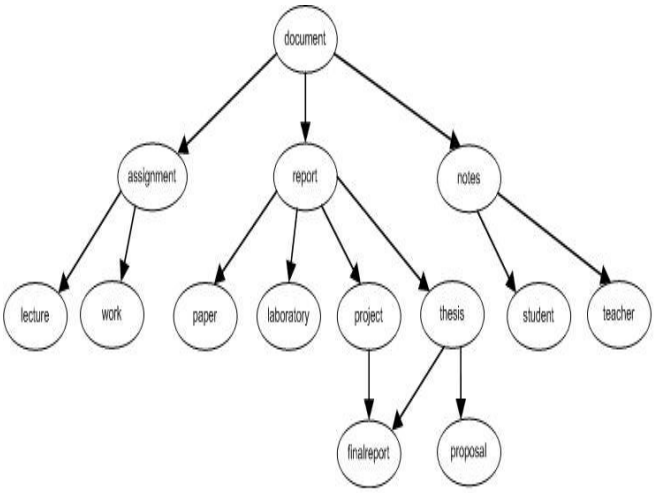

Figure 1. Ontology of Documents

Figure 1 shows the ontology, where concepts like "laboratory reports", "reading assignments" and "homework assignments" can be identified. This ontology is used to construct the knowledge base.

3.2 An example of ontology

We take as a case study the concepts related to "The Knowledge Representation" for the construction of a ontology, where some concepts or subjects and their relations are identified. In Figure 2 we show part of this ontology.

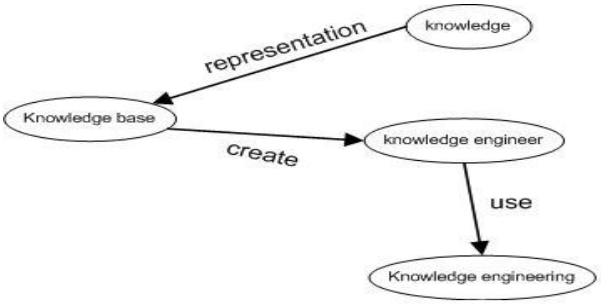

Figure 2. Example of Subject Ontology

In this example we have the ontology, where the concepts like "knowledge base", "knowledge", and "knowledge engineer" as well as relations like "representation" and "create" were identified. RDFS allows more varied and

versatile representation of relationships beyond the class-subclass relationship commonly used in the object oriented paradigm.

4. RDFS MODEL

The Resource Description Framework and RDF Schema were developed by the W3C to allow the specification of the semantics of data based on XML, in a standardized, interoperable manner [Gomez02].

The Standard "Resource Description Framework Schema", RDFS [Brickley99], is a model formed by three elements: "resource", "property" and "value", which allow expressing the metadata identified for the information resources.

In this stage of the investigation we focused on how to represent the knowledge using these mark-up languages. This will allow the reusability and sharing of this knowledge between different users. The subject ontology can be expressed through RDFS.

As an illustrative example lets consider the representation of the concept "KnowledgeEngineer" and the relation that exists with other concepts such as "KnowledgeBase", "Knowledge" and "KnowledgeEngineering". These represent the objects in an RDF sentence. The expressions "cr:create", "cr:obtain" and "cr:use" represent the predicates. The sentence <cr:linkDocument rdf:resource="#document1>" identifies a document related to the topic or concept.

```
<cr:Topic rdf:ID="KnowledgeEngineer">
    <cr:create rdf:resource="#KnowledgeBase" />
    <cr:Obtain rdf:resource="#Knowledge" />
    <cr:use rdf:resource="#KnowledgeEngineering" />
    <cr:linkDocument rdf:resource="#document1" />
    <cr:linkDocument rdf:resource="#document2" />
</cr:Topic>
```

The sentence that identifies the document makes reference to the resource "#document1" that contains information about the document. Additionally the attributes of a document are shown.

```
<res:document rdf:ID="document1">
  <res:title>Title of Document</res:title>
  <res:author>Juan L. Dinos</res:author>
  <res:url>http://www.ece.uprm.edu/s012127/document1.pdf</res:utl>
</res:document>
```

Each document is identified by a type of document which will be represented by "rdf:document". Additionally the characteristics of a document are represented by "res:title", "res:author", "res:url". At the moment the metadata represent basic characteristics that identify a document, but can be extended depending on the type of use and additional requirements of the system.

5. ARCHITECTURE FOR DOCUMENT RETRIEVAL

We propose an agent-based architecture for the document retrieval. Agent features such as autonomy, asynchronous, fault tolerant, etc are adequate effective for retrieving resources such as documents from different knowledge bases. For the implementation of agents, the Java Agent DEvelopment Framework (JADE) was used [Bellifemine99]. This framework efficiently manages message passing between agents, and implements the FIPA specification where the components are clearly identified and integrated (interaction protocols, envelop, ACL, content languages, encoding schemes, ontologies and, finally, transport protocols).

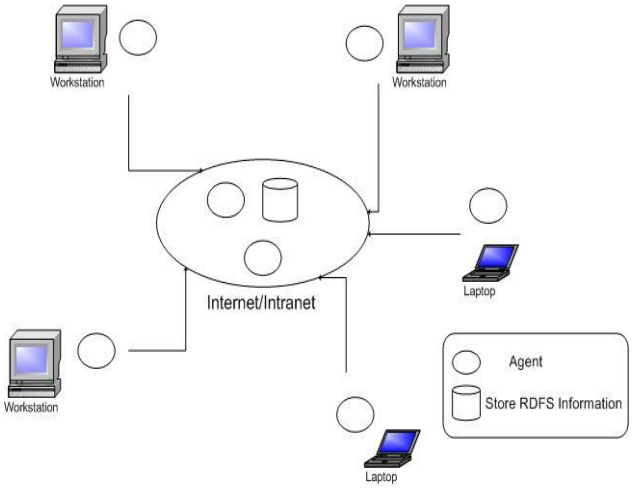

Figure 3. Agent-Based Architecture for the Document Retrieval

The architecture shown in Figure 3 has as main component the "UserAgent". This agent interacts directly with the user to allow sharing information with other agents. Additionally "UserAgent" interacts with the agents "SearchAgent" and "OntologyAgent".

The "OntologyAgent" allows the recovery of different ontologies and its structure. "SearchAgent" is in charge of retrieving the metadata of documents. In this architecture the "RDF Data Query Language (RDQL)" was used to retrieve the metadata of documents.

6. CONCLUSIONS

The use of ontologies is very important and is of great help to enable the identification the concepts used more often in the knowledge domain that the students and professors use. The ontologies allow identifying the properties and relationships existing among the concepts in a given domain, becoming a powerful knowledge representation metadata for knowledge repositories.

Technologies like RDF, RDFS allow the construction of sophisticated representations of knowledge, besides providing a rich set of primitives and elements to represent the different kinds of resources that are part of a teaching-learning process.

The software agent technology offers an enormous potential for the task of managing the information that students use and generate, where tasks such as document search and retrieval can be delegated to agents.

REFERENCES

[Adler89] Adler, P.S., "When knowledge is the critical resource, knowledge management is the critical task", IEEE Trans. On Engineering Management, vol. 36 No. 2, pp 87-94, May 1989.

[Beckett02] Beckett, D., E. Miller, and D. Brickley, "Expressing Simple Dublin Core in RDF/XML", http://dublincore.org/documents/2002/07/31/dcmes-xml/, 2002.

[Bellifemine99] Bellifemine, F., A. Poggi, and G. Rimasa, "JADE – A FIPA – compliant agent framework", Proceeding of PAAM'99, London, April 1999.

[Brickley99] Brickley, D. and R.V. Guha, "Resource Description Framework (RDF) Schema Specification", W3C Proposed Recommendation, www.w3.org/TR/PR-rdf-schema, Mar. 1999.

[Gomez02] Gomez-Perez, A., O. Corcho, "Ontology Languages for the Semantic Web", Intelligent Systems, vol. 17 issue: 1, Jan-Feb 2002.

[Gruber93] Gruber, T., "Toward Principles for the Design of Ontologies Used for Knowledge Sharing", Technical Report KSL-93-04, Knowledge Systems Laboratory, Stanford University, CA, 1993.

[Hendler99] Hendler, J., "Is There an Intelligent Agent in Your Future?",http://www.nature.com/nature/webmatters/agents/agents.html, March 1999.

[Horrocks00] Horrocks, I. et al., "OIL in a Nutshell", Proc. ECAI '00 Workshop on Application of Ontologies and PSMs, Berlin, Germany, pp. 4.1–4.12, 2000.

[Lassila99] Lassila, O., and R. Webick, "Resource Description Framework (RDF) Model and Syntax Specification.", W3C Recommendation, www.w3.org/TR/PR-rdf-syntax, Jan. 1999

[Lassila00] Lassila, O., F. van Harmelen, I. Horrocks, J. Hendler, and D.L. McGuinness, "The Semantic Web and its Languages", Intelligent Systems, IEEE, vol. 15 Issue: 6, pp. 67 – 73, Nov. - Dec. 2000.

[Ungson99] Ungson, G.H. and J. D. Trudel, "The emerging knowledge-based economy", IEEE Spectrum, http://www.spectrum.ieee.org/spectrum/may99/features/engi.html, May 99.