INTELLIGENT INFORMATION PROCESSIG IN A DIGITAL LIBRARY USING SEMANTIC WEB

Antonio Martín, Carlos León
Dpto. Tecnología Electrónica. Sevilla Universtiy
Sevilla,Spain
toni@us.es; cleon@us.es

Abstract

With the explosive growth of information, it is becoming increasingly difficult to retrieve the relevant documents with current search engine only. The information is treated as an ordinary database that manages the contents and positions. To the individual user, there is a great deal of useless information in addition to the substantial amount of useful information.

This begets new challenges to docent community and motivates researchers to look for intelligent information retrieval approach and ontologies that search and/or filter information automatically based on some higher level of understanding are required. We study improving the efficiency of search methods and classify the search patrons into several models based on the profiles of agent based on ontology.

We have proposed a method to efficiently search for the target information on a Digital Library network with multiple independent information sources. This paper outlines the development of an expert prototype system based in an ontology for retrieval information of the Digital Library University of Seville. The results of this study demonstrate that by improving representation by incorporating more metadata from within the information and the ontology into the retrieval process, the effectiveness of the information retrieval is enhanced. We used Jcolibri and Prótége for developing the ontology and creation the expert system respectively.

1. INTRODUCTION

In the current digital libraries and Internet the information is treated as an ordinary database that manages the contents and positions. The result generated by the current search engines is a list of Web addresses that contain or treat the pattern. A main bottleneck here is the lack of uniform semantics interpretation standards and technologies. Hence, interoperability can be only achieved by some kind of semantics unification.

Although search engines have developed increasingly effective, information overload obstructs precise searches. We make an effort in this direction by investigating techniques that attempt to utilize the Artificial Intelligent and the ontology as a knowledge representation formalism to improve effectiveness in information retrieval. The hypothesis is that with case-based reasoning (CBR) expert system and by incorporating limited semantic knowledge, it is possible to improve the effectiveness of an Information retrieval system.

In this paper we study architecture of the search layer in a particular dominion, a web-based catalogue for the University of Seville. We provide an insight of the technical aspects of the application by enumerating the technological requirements and the associated architecture, where we pay specific attention to the CBR framework jCOLIBRI, GAIA (1) and its features for implementing the reasoning process over ontologies. We outline its main components and describe how can interact Intelligent Artificial and Semantic Web to enhancement a search engine. The paper is organized as follows. Next Section describes the setting of Digital Library domain, the research problems and current work in it. Then we present briefly the OntoFAMA and our concept for extending the search layer of the system by integrating an intelligent service in it to perform the search process over different types of information repositories. We present the results of our ongoing work on the adaptation of the framework. Finally we outline the directions of our future work.

2. MOTIVATION AND REQUIREMENTS

Digital Libraries is a privileged domain for the application of innovative, knowledge intensive services that provide a flexible and efficient method for searching information and guarantee the user with a set of results actually related to his/her interest. The Sevilla Digital Library (SDL) is dedicated to the production, maintenance, delivery, and preservation of a wide range of high-quality networked resources for scholars and students at University and elsewhere. SDL provides tools that support the construction of online information services for research, teaching, and learning, including services that enable the Sevilla University libraries to effectively share their materials and provide greater access to digital content.

A major aspect to ensure the interoperability of the services is the usage of standard, platform independent technologies, Semantic Web technologies, established standards for describing the library objects and web services. Our efforts aim to improve significantly the efficiency of the search process and extend the search layer with an intelligent e-service, so that intuitive information access and content-based retrieval is

available. We developed four user profiles based on ontologies: Staff, Alumni, Administration, and visitor. These profiles are used to specify the search results. The profile is going to satisfy the quality of information for a specific user, Figure 1.

Figure 1 Servers and resources associated with a Profile

3. ONTOLOGY DEVELOPMENT

Our ontology is organized into multiple sub-names or ontologies. OntoFAMA project contains a collection of codes, visualization tools, computing resources, and data sets distributed across the grids, for which we have developed a well-defined ontology using RDF language [2]. Figure 2 shows the high level classification of classes to group together OntoFAMA resources as well as things that are related with these resources.

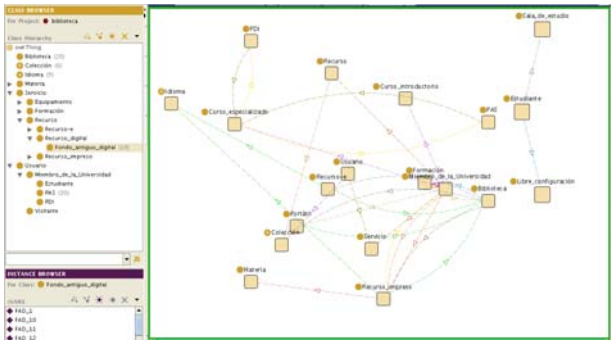

Figure 2 Class hierarchy for the OntoFama ontology

These are used to describe characteristics of the instances of users and resource. User Ontology define user groups and relationships between them: Teaching and research staff, administrative staff, Students and External Users (visiting students, visiting scholars, research staff external, etc) and Service Ontology that define characteristics of services in a specific work area: provider, identifying data, information resources structure, etc. For the construction of the ontology we followed next steps: determine the domain and scope of the ontology, enumerate important terms the ontology, define the classes and class hierarchy, define properties of classes, and define the facets of the slots, and generating the Ontology instances. For the manual generation and modelling of the domain ontology we chose the Protégé editor [3].

4. OntoFAMA ARCHITECTURE

The proposed architecture is based on our approach for realization of content based retrieval information by means of metadata characterizations and domain ontology inclusion, Figure3. It implies to use ontology as vocabulary to define complex, multi-relational case structures to support the CBR processes.

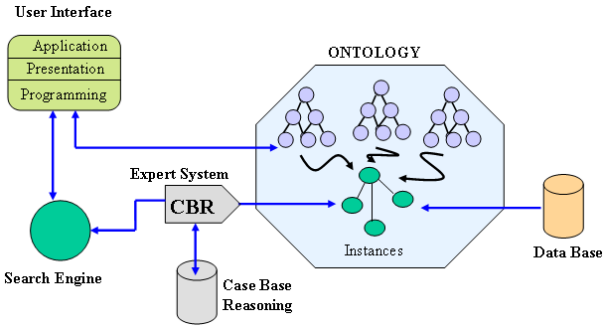

Figure 3 OntoFAMA Architecture.

The metadata descriptions of the resources and library objects (cases) are abstracted from the details of their physical representation in the Electronic Catalogue and are stored in the case base. This way the same methods can operate over different types of information repositories. The mapping between the two layers is realized by connectors. These Connectors read the values of the data base columns and ontology, and return them to the application, i.e. assign them to the attributes of the case. Basing on the same idea, the case base implements a common interface for the similarity methods to assess the cases. This way the organization and indexation of case base will not affect the implementation of the reasoning methods.

OntoFAMA contains a CBR component that automatically searches for similar queries-answer pairs based on the knowledge that the system extracted from the questions text. CBR is one of most successful applied AI technologies of recent years and is widely discussed in the literature as a technology for building information systems to support knowledge management, where metadata descriptions for characterizing knowledge items are used. Current research shows how these systems can benefit from a standardized shared knowledge representation that implies unambiguous interpretation of cases and in this way enable the development of systems that are able to search across multiple case-bases [4]. CBR is a problem solving paradigm in Artificial Intelligent where the problems are solved on previously experienced similar problems. In OntoFAMA CBR is based on the intuition that new searches are often similar to previously encountered searches, and therefore, that past solutions (results) may be reused directly or through adaptation in the current situation.

CBR systems typically apply retrieval and matching algorithms to a case base of past problem-solution pairs. These characterizations are called cases and are stored in a case base. Standard cases are composed by several attributes with different simple data types (Integer, String). We use the Concept data type supported by the jCOLIBRI framework to indicate that an attribute is

going to represent a concept of the ontology. The values of this attribute are going to be the corresponding instances of the linked concept.

The case-based reasoning-cycle in our system may be described by the following processes:

1. Retrieve the most similar cases. During this process, the CB Reasoner searches the database to find the most approximate case to the current search.

2. Reuse the cases. This process includes using the retrieved case and adapting it to the new situation. The reasoner might propose a solution.

3. Revise the proposed solution if necessary. Since the proposed solution could be inadequate, this process can correct the first proposed solution.

4. Retain the new solution as a part of a new case. This process enables CBR to learn and create a new solution that should be added to the case base.

5. EXPERIMENT RESULTS

Experiments have been carried out in order to test the efficiency of Artificial Intelligent and Ontologies in retrieval information. Our system has a graphical user interface for determining initial user requirements early in search. This operation permit that useless information in search engine process can be reduced or completely avoided. Managing user requirements by placing focus on identifying, gathering, and documenting essential information is a specialised work area or user profiles.

The main goal is to check if the mechanism of query formulation, assisted by an interface, gives a suitable tool for augmenting the number of significant documents extracted from the Digital Library, to be stored in the CBR. The user begins the search devising the starting query. The outcomes represented in the following Figure 4 display the number of important documents retrieved in OntoFama and the total number of documents retrieved in a traditional search engine.

Figure 4 Search engine results page.

The results include a list of web pages with titles, a link to the page, and a short description showing where the keywords have matched content within the page.

6. CONCLUSIONS

An ontology and integrated intelligent system architecture for search operation support system and its implementation platform have been developed in this paper. We presented an ontology based in Artificial Intelligent architecture for knowledge management in the Sevilla Digital Library. A crucial role in it plays the jCOLIBRI-based and Protégé components that are the cornerstone in the proposed architecture. Its system functions include data processing and intelligent information retrieval.

We described an effort to design and develop a prototype for management the resources in a library such as OntoFAMA project, and to exploit them to aid users as they select resources. It introduced a prototype web-based CBR retrieval system which operates on an RDF file store. This system combines RDF representation and CBR recommendation methodology to do code selection for the resources codes; thus it applies a CBR approach with RDF data model.

Furthermore an intelligent agent was illustrated for assisting the user by suggesting improved ways to query the system on the ground of the resources in a Digital Library according to his own preferences, which come to represent his interests. Future work will concern the exploitation of information coming from others libraries and services and further refine the suggested queries, to extend the system to provide another type of support, as well as to refine and evaluate the system through user testing.

References

[1] Grupo GAIA de la Universidad Complutense de Madrid. Distribución del entorno de desarrollo jCOLIBRI con licencia LGPL, http://gaia.fdi.ucm.es/grupo/projects/jcolibri/, 2009.

[2] W3C, RDF Vocabulary Description Language 1.0: RDF Schema. http://www.w3.org/TR/rdf-schema/, 2009

[3] PROTÉGÉ, The Protégé Ontology Editor. <http://protege.stanford.edu/>, 2009.

[4] Bridge, M. H. G¨oker, L. McGinty, and B. Smyth. Case-based recommender systems. Knowledge Engineering Review, 20(3):315–320, 2006.

[5] Taniar, D. and Wenny Rahayu, J.. Web semantics and ontology. Hershey, PA: Idea Group Pub, 2006.

Biography

Photograph of the Author *(Optional)*

Antonio Martin obtained his PhD in Computer Science from de University of Sevilla. Our areas of investigation are the applications of expert systems. Dr. Martin is Professor of Electronic Engineering and head of technological units in the Seville University Library.

Dr. Carlos León received his Physical Electronics degree and his PhD Computer Science from the Sevilla University. He is Professor of Electronic Engineering and head of Computer and Telecommunication Service of Seville University. His areas of research are expert systems, neural networks, data, mining, etc.