CHALLENGES IN DISTRIBUTED INFORMATION SEARCH IN A SEMANTIC DIGITAL LIBRARY

Antonio Martín and Carlos León
Departamento de Tecnología Electrónica, Seville University, Avda. Reina Mercedes S/N, Seville, Spain

Keywords: Ontology, Semantic Web, Retrieval, Case-based Reasoning, Digital Library, Knowledge Management.

Abstract: Nowadays an enormous quantity of heterogeneous and distributed information is stored in the current digital libraries. Access to these collections poses a serious challenge, however, because present search techniques based on manually annotated metadata and linear replay of material selected by the user do not scale effectively or efficiently to large collections. The artificial intelligent and semantic Web provides a common framework that allows knowledge to be shared and reused. In this paper we propose a comprehensive approach for discovering information objects in large digital collections based on analysis of recorded semantic metadata in those objects and the application of expert system technologies. We suggest a conceptual architecture for a semantic and intelligent search engine. OntoFAMA is a collaborative effort that proposes a new form of interaction between people and Digital Library, where the latter is adapted to individuals and their surroundings. We have used Case Based-Reasoning methodology to develop a prototype for supporting efficient retrieval knowledge from digital library of Seville University.

1 INTRODUCTION

In the traditional search engines 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. The useful information buried under the useless information cannot be discovered. It is disconcerting for the end user.

In the historical evolution of digital libraries, the mechanisms for retrieval of scientific literature have been particularly important. There are a lot of researches on applying these new technologies into current Digital Libraries information retrieval systems, but no research addresses the semantic and intelligent artificial issues from the whole life cycle and architecture point of view.

Although search engines have developed increasingly effective, information overload obstructs precise searches. We study improving the efficiency of search methods to search a distributed data space like a Digital Library. It incorporates semantic Web and artificial intelligent technologies to enable not only precise location of digital library resources but also the automatic or semi-automatic learning. Our approach for realizing content based search and retrieval information implies the application of the Case-Based Reasoning (CBR) technology (Toussaint and Cheng, 2006). Our work differs from related projects in that we build an ontology-based contextual profiles and we introduce an approaches used metadata-based in ontology search and expert systems (Warren, 2005).

The contributions are divided into next sections. In the next section, short descriptions of important aspects in Digital Library domain, the general overview about our prototype architecture and current work in it are reported. Then we summarize its main components and describe how can interact Intelligent Artificial and Semantic Web to enhancement a search engine. Next we study the CBR framework jColibri and its features for implementing the reasoning process over ontologies (GAIA, 2009). Finally we present the results of our ongoing work on the adaptation of the framework and we outline the future works.

2 MOTIVATION

In the historical evolution of digital libraries, the mechanisms for retrieval of scientific literature have been particularly important. These network information systems support search and display of items from organized collections. The semantic Web provides a common framework that allows

Martín A. and León C. (2010).
CHALLENGES IN DISTRIBUTED INFORMATION SEARCH IN A SEMANTIC DIGITAL LIBRARY.
In *Proceedings of the 5th International Conference on Software and Data Technologies*, pages 495-498
DOI: 10.5220/0002933104950498

knowledge to be shared and reused across community libraries and semantic searchers (Sure & Studer, 2005).

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 make an effort in this direction by investigating techniques that attempt to utilize ontologies to improve effectiveness in information retrieval (Ding, 2004).

In this paper we study architecture of the search layer in this particular dominium, a web-based catalogue for the University of Seville. The Seville Digital Library (SDL) provides tools that support the construction of online information services for research, teaching, and learning. SDL include services to effectively share their materials and provide greater access to digital content (Witten & Bainbridge, 2003).

3 ONTOFAMA ARCHITECTURE

The architecture of our system named OntoFAMA is shown in figure 1, which mainly includes three parts: search engine, ontology knowledge base, and intelligent user interface.

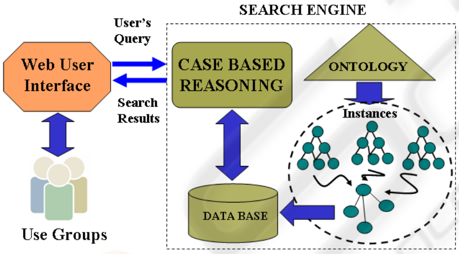


Figure 1: System architecture of OntoFAMA.

Inference engine contains a CBR component that automatically searches for similar queries-answer pairs based on the knowledge that the system extracted from the questions text (Bridge and G¨oker, 2006). We used a Case-Based Reasoning (CBR) shell, software that can be used to develop several applications that require cased-based reasoning methodology. In this work we analyzed the CBR object-oriented framework development environments JColibri (Díaz-Agudo and González-Calero, 2007).

OntoFAMA system uses its internal knowledge bases and inference mechanisms to process information about the electronic resources in a Digital Library. At this stage we consider to use ontology as vocabulary for defining the case structure like attribute-value pairs. Ontology will be considered as knowledge structure that will identify the concepts, property of concept, resources, and relationships among them to enable share and reuse of knowledge that are needed to acquire knowledge in a specific search domain.

The acceptability of a system depends to a great extent on the quality of this user interface component. The interfaces provides for browsing, searching and facilitating Web contents and services. The objective of profile intelligence has focused on creating of user profiles: Staff, Alumni, Administrator, and Visitor. The user interface helps to user to build a particular profile that contains his interest search areas in the digital library domain.

4 CASE-BASED REASONING

Case-Based Reasoning is a problem solving paradigm that solves a new problem, in our case a new search, by remembering a previous similar situation and by reusing information and knowledge of that situation. A new problem is solved by retrieving one or more previously experienced cases, reusing the case, revising. The case-based reasoning-cycle in our system may be described by the following processes.

- Retrieval. Main focus of methods in this category is to find similarity between cases.
- Reuse: a complete design where case-based and slot-based adaptation can be hooked is provided.
- Revise the proposed solution if necessary.
- Retain the new solution as a part of a new case.

CBR case data could be considered as a portion of the knowledge (metadata) about an OntoFama object. Every case contains both a solution pointer and a problem description used for similarity assessment. Although CBR claims to reduce the effort required for developing knowledge-based systems substantially compared with more traditional Artificial Intelligence approaches, the implementation of a CBR application from scratch is still a time consuming task.

In this study we used the CBR object-oriented framework development environments JColibri. jColibri is and open source framework and their interface layer provides several graphical tools that help users in the configuration of a new CBR system. Another decision criterion for our choice is the easy ontologies integration. jColibri affords the opportunity to incorporate ontology in the CBR application to use it for case representation and content-based reasoning methods to assess the similarity between them.

5 ONTOLOGY DESIGN

We need a vocabulary of concepts, resources and services for our information system described in the scenario requires definitions about the relationships between objects of discourse and their attributes. We have proposed to use ontology together with CBR in the acquisition of an expert knowledge in a specific domain. We integrated three essential sources to the system: electronic resources, catalogue and personal Data Base.

The W3C defines standards that can be used to design an ontology (W3C, 2009). We wrote the description of these classes and the properties in RDF semantic markup language. We choose Protégé as ontology editor, which supports knowledge acquisition and knowledge base development (Protégé, 2009). It is a powerful development and knowledge-modelling tool with an open architecture. Protégé uses OWL and RDF as ontology language to establish semantic relations. The ontology and its sub-classes are established according to the taxonomies profile, as shown in figure 2.

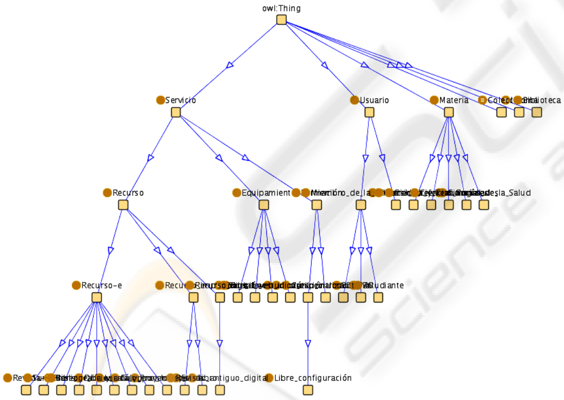


Figure 2: Class hierarchy for the OntoFama ontology.

Our ontology can be regarded as quaternion OntoSearch:={profile, collection, source) where profiles represent the user kinds, resources contains all the services and resources of the digital library and matter cover the different information sources: catalogue, history fond, intranet, Web, etc.

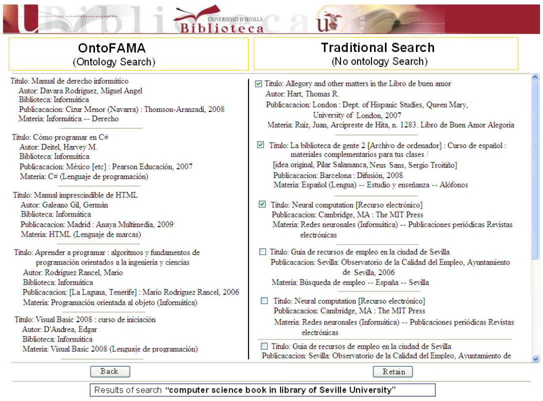
After ontology is established, we need to add enough initial instances and item instances to knowledge base. For this purpose we choose a certain item, and create a blank instance for item. Then the domain expert, in this case the librarian fills blank units of instance according the domain knowledge (Horridge and Knublauch, 2004). Then the library of cases (the "case base") is generated from a file store where each case is represented with RDF syntax. 1100 cases were collected for user profiles and their different resources and services. Each case contains a set of attributes concerning both metadata and knowledge.

6 CONVERSATION INTERFACE

As we have seen in previous sections our system has a graphical user interface for determining initial user requirements in search. Rather than building static user profiles, contextual systems try to adapt to the user's current search. OntoFAMA monitors user's tasks, anticipates search-based information needs, and proactively provide users with relevant information. This configuration contains the user requirements most typically described the relative needs, tasks, and goals of the user for an individual search.

The user enters query commands and the system asks questions during the inference process. Besides, the user will be able to solve new searches for which he has not been instructed, because the user profiles what he has learnt during the previous searchers.

The user inputs the keywords in the user profile interface. Suppose the user is looking for some resource about "Computer Science electronic resource" in the library digital domain of Seville. The required resources should contain some knowledge about "Computer Science" and related issues. After searching, some resources are returned as results, which are shown in figure 3.



OntoFAMA
(Ontology Search)

Título: Manual de derecho informático
Autor: Davara Rodriguez, Miguel Angel
Biblioteca: Informática
Publicacacion: Cizur Menor (Navarra) : Thomson-Aranzadi, 2008
Materia: Informática -- Derecho

Título: Cómo programar en C#
Autor: Deitel, Harvey M.
Biblioteca: Informática
Publicacacion: México [etc] : Pearson Educación, 2007
Materia: C# (Lenguaje de programación)

Título: Manual imprescindible de HTML
Autor: Galeano Gil, Germán
Biblioteca: Informática
Publicacacion: Madrid : Anaya Multimedia, 2009
Materia: HTML (Lenguaje de marcas)

Título: Aprender a programar : algoritmos y fundamentos de programación orientados a la ingenieria y ciencias
Autor: Rodriguez Rancel, Mario
Biblioteca: Informática
Publicacacion: [La Laguna, Tenerife] : Mario Rodriguez Rancel, 2006
Materia: Programación orientada al objeto (Informática)

Título: Visual Basic 2008 : curso de iniciación
Autor: D'Andrea, Edgar
Biblioteca: Informática
Materia: Visual Basic 2008 (Lenguaje de programación)

Back

Traditional Search
(No ontology Search)

☑ Título: Allegory and other matters in the Libro de buen amor
Autor: Hart, Thomas R.
Publicacacion: London : Dept. of Hispanic Studies, Queen Mary, University of London, 2007
Materia: Ruiz, Juan, Arcipreste de Hita, n. 1283. Libro de Buen Amor Alegoria

☑ Título: La biblioteca de gente 2 [Archivo de ordenador] : Curso de español : materiales complementarios para tus clases / [idea original, Pilar Salamanca, Neus Sans, Sergio Troitiño]
Publicacacion: Barcelona : Difusión, 2008
Materia: Español (Lengua) -- Estudio y enseñanza -- Alófonos

☑ Título: Neural computation [Recurso electrónico]
Publicacacion: Cambridge, MA : The MIT Press
Materia: Redes neuronales (Informática) -- Publicaciones periódicas Revistas electrónicas

□ Título: Guia de recursos de empleo en la ciudad de Sevilla
Publicacacion: Sevilla: Observatorio de la Calidad del Empleo, Ayuntamiento de Sevilla, 2006
Materia: Búsqueda de empleo -- España -- Sevilla

□ Título: Neural computation [Recurso electrónico]
Publicacacion: Cambridge, MA : The MIT Press
Materia: Redes neuronales (Informática) -- Publicaciones periódicas Revistas electrónicas

□ Título: Guia de recursos de empleo en la ciudad de Sevilla
Publicacacion: Sevilla: Observatorio de la Calidad del Empleo, Ayuntamiento de

Retain

Results of search **"computer science book in library of Seville University"**

Figure 3: 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.

7 EVALUATION

Experiments have been carried out in order to test the efficiency of artificial intelligent and ontologies in retrieval information in a digital library. For our experiments we considered 50 users with different profiles. So that we could establish a context for the users, they were asked to at least start their essay before issuing any queries to OntoFAMA. They were also asked to look through all the results returned by OntoFAMA before clicking on any result. We compared the top 10 search results of each keyword phrase per search engine. Our application recorded which results on which they clicked, which we used as a form of implicit user relevance in our analysis. We have agreed different values to measure the quality of retrieved documents, excellent, good, acceptable and poor.

In each experiment we report the average rank of the user-clicked result for our baseline system, Google and for our search engine OntoFAMA. Then we calculated the rank for each retrieval document by combining the various values and comparing the total number of extracted documents and documents consulted by the user (table 1). We can observe the best final ranking was obtained for our prototype OntoFAMA and an interesting improvement over the performance of Google.

Table 1: Analysis of retrieved documents relevance.

	Excellent	Good	Acceptable	Poor
OntoFAMA	5,5 %	39,3 %	40,6 %	14,4 %
Google	2,7 %	31 %	44,8 %	21,3 %

8 CONCLUSIONS

We have investigated how the semantic technologies can be used to provide additional semantics from existing resources in digital libraries. For this purpose we presented a system based in ontology and artificial intelligent architecture for knowledge management in the Seville Digital Library. Our study addresses the main aspects of a semantic Web information retrieval system architecture trying to answer the requirements of the next-generation semantic Web user.

To put our aims into practice we should first of all develop the domain ontology and study how the content-based similarity between the concepts typed attributes could be assessed in CBR system. A dedicated inference mechanism is used to answer queries conforming to the logic formalism and terms defined in our ontology. We have been working on the design of entirely ontology based structure of the case and the development of our own reasoning methods in jColibri to operate with it. 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

Toussaint, J., Cheng, K., 2006. *Web-based CBR as a tool with the application to tooling selection.* International Journal of Advanced Manufacturing Technology.

Warren, P. 2005. *Applying semantic technologies to a digital library: a case study*" Library Management Journal, Emerald.

GAIA - Group for Artificial Intelligence Applications, 2009. *jCOLIBRI project - Distribution of the development environment with LGPL*, http://gaia.fdi.ucm.es/grupo/projects/. Complutense University of Madrid.

Sure, Y., and Studer, R., 2005. *Semantic web technologies for digital libraries*. Library Management Journal, Emerald, vol. 26.

Ding, H., 2004.*Towards the metadata integration issues in peer-to-peer based digital libraries*. GCC (H. Jin, Y. Pan, N. Xiao, and J. Sun, eds.), vol. 3251 of Lecture Notes in Computer Science, Springer.

Witten, I. H., and Bainbridge, D., 2003. *How to Build a Digital Libary.* Morgan Kaufmann.

Bridge, M., G¨oker, H., McGinty,L., Smyth, B. 2006.*Case-based recommender systems.* Knowledge Engineering Review.

Díaz-Agudo, B., González-Calero, P.A., Recio-García, J., Sánchez-Ruiz, A., 2007. *Building CBR systems with jColibri.* Journal of Science of Computer Programming.

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

PROTÉGÉ, 2009. *The Protégé Ontology Editor and Knowledge Acquisition System*. <http://protege.stanford.edu/>.

Horridge, M., Knublauch, H. 2004. *A Practical Guide To Building OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tools*. The University Of Manchester. Reino Unido.