Intelligent Technique to Accomplish a Effective Knowledge Retrieval from Distributed Repositories.

Antonio Martín, Carlos León
Department of Electronic Technology
Higher Technical School of Computer Engineering
Sevilla, Spain
toni@us.es, cleon@us.es

Abstract— **Currently, 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 Intelligence and Semantic Web provide 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. We concentrate on the critical issue of metadata/ontology-based search. More specifically the objective is investigated from a search perspective possible intelligent infrastructures form constructing decentralized digital libraries where no global schema exists. We have used Case Based-Reasoning methodology to develop a prototype for supporting efficient retrieval knowledge from digital library of Seville University. OntoSDL is a collaborative effort that proposes a new form of interaction between people and Digital Libraries, where the latter are adapted to individuals and their surroundings.**

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

I. INTRODUCTION

A Digital Library (DL) enables users to interact effectively with information distributed across a network. These network information systems support search and display of items from organized collections. In the historical evolution of digital libraries the mechanisms for retrieval of scientific literature have been particularly important. Traditional search engines treated the information 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. Thus, sometimes it takes a long time to search for needed information.

Although search engines have developed increasingly effective, information overload obstructs precise searches. Despite large investments and efforts have been made, there are still a lot of unsolved problems. There are a lot of researches on applying these new technologies into current DL information retrieval systems, but no research addresses the semantic and Artificial Intelligence (AI) issues from the whole life cycle and architecture point of view [1]. Our work differs from related projects in that we build ontology-based contextual profiles and we introduce an approaches used metadata-based in ontology search and expert systems [2].

We study improving the efficiency of search methods to search a distributed data space like a DL. The objective has focused on creating technologically complex environments in Education, Learning and Teaching in the DL domain. We presented an intelligent approach to develop an efficient semantic search engine. It incorporates semantic Web and AI technologies to enable not only precise location of DL resources but also the automatic or semi-automatic learning [3]. We focus our discussion on case indexing and retrieval strategies to provide an intelligent application in searching area. For this reason we are improving representation by incorporating more metadata in the information representation. Our objective here is thus to contribute to a better knowledge retrieval in the digital libraries field. Our approach for realizing content based both search and retrieval information implies the application of the Case-Based Reasoning (CBR) technology [4].

The contributions are divided into next sections. In the first section, short descriptions of important aspects in DL domain, the research problems and current work in it are reported. Then, we summarize its main components and describe how can interact AI and Semantic Web to improve the search engine. Third section focuses on the ontology design process and provides a general overview about our prototype architecture. Next, we study the CBR framework jColibri and its features for implementing the reasoning process over ontologies [5]. Obviously, our system is a prototype but, nevertheless, it gives a good picture of the on-going activities in this new and important area. Finally, we present conclusions of our ongoing work on the adaptation of the framework and we outline future works.

II. MOTIVATIION AND REQUIREMENTS

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. Reuse the knowledge is an important area in DL. The Semantic Web provides a common framework that allows knowledge to be shared and reused across community libraries and semantic searchers [6].

This begets new challenges to docent community and motivates researchers to look for intelligent information

 ISBN: 978-1-61208-352-0

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. Thus, ontologies are seen as key enablers for the Semantic Web. The use of AI and ontologies as a knowledge representation formalism offers many advantages in information retrieval [8]. In our work, we analyzed the relationship between both factors ontologies and AI. We have proposed a method to efficiently search for the target information on a DL network with multiple independent information sources [7].

Seville 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. The hypothesis is that with a 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 this particular dominium, a web-based catalogue for the University of Seville. 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. Our objective here is thus to contribute to a better knowledge retrieval in the digital libraries field.

III. THE SYSTEM ARCHITECTURE

In order to support semantic retrieval knowledge in a DL, we develop a prototype named OntoSDL based on ontologies and expert systems. The architecture of our system is shown in Fig.1, which mainly includes three parts: intelligent user interface, ontology knowledge base, and the search engine. Their corresponding characteristics and functions are studied in the following paragraphs.

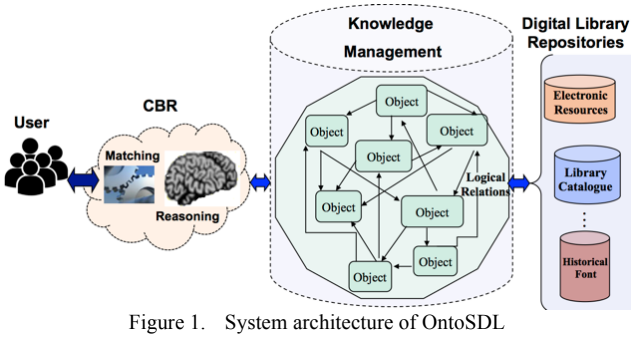

Figure 1. System architecture of OntoSDL

OntoSDL system uses its internal knowledge bases and inference mechanisms to process information about the electronic resources in a DL. 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.

Ontology knowledge base is the kernel part for semantic retrieval information. The metadata descriptions of the resources and library objects (cases) are abstracted from the details of their physical representation and are stored in the case base. Ontology stores information about resources and services where concepts are types, or classes, individuals are allowed values, or objects and relations are the attributes describing the objects.

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. Case base has a memory organization interface that assumes that whole case-base can be read into memory for the CBR to work with it. We used a CBR shell, software that can be used to develop several applications that require cased-based reasoning methodology. Also we have implemented a new interface, which allows retrieving cases enough to satisfy a SQL query. In this work, we analyzed the CBR object-oriented framework development environments JColibri. This framework work as open software development environment and facilitate the reuse of their design, as well as implementations.

The acceptability of a system depends to a great extent on the quality of his user interface component. In our system, the user interacts with the system to fill in the gaps to retrieve the right cases. The interfaces provide for browsing, searching and facilitating Web contents and services. It consists of one user profile, consumer search agent components and bring together a variety of necessary information from different user's resources. The user interface helps to user to build a particular profile that contains his interest search areas in the DL domain. The objective of profile intelligence has focused on creating of user profiles: Staff, Alumni, Administrator, and Visitor.

We have developed a graphical selection interface as illustrated in Fig. 2.

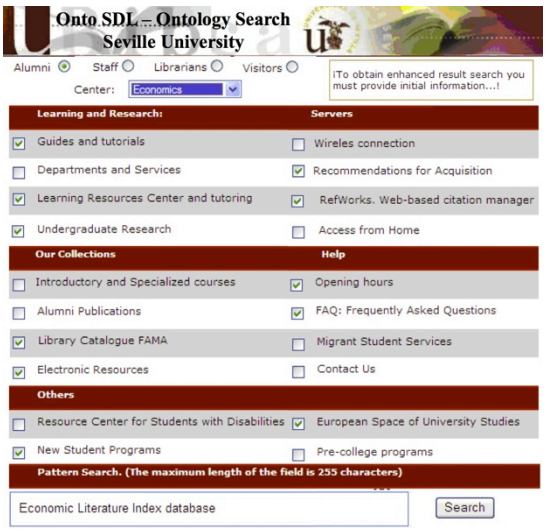

Figure 2. User profiles interface

In an intelligence profile setting, people are surrounded by intelligent interfaces merged. Rather than building static user profiles, contextual systems try to adapt to the user's

current search. OntoSDL monitors user's tasks, anticipates search-based information needs, and proactively provide users with relevant information. Thus creating a computing-capable environment with intelligent communication and processing available to the user by means of a simple, natural, and effortless human-system interaction. 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.

IV. CASE-BASED REASONING INTELLIGENT TECHNIQUE

CBR 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. CBR 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. In our CBR application, problems are described by metadata concerning desired characteristics of a library resource, and the result to a specific search is a pointer to a resource described by metadata. These characterizations are called cases and are stored in a case base. CBR case data could be considered as a portion of the knowledge (metadata) about an OntoSDL object. Every case contains both a solution pointers and problem description used for similarity assessment. Description of the framework case, which is formally described in terms of framework domain taxonomy they are used for indexing cases. The possible solutions described by means of framework instantiation actions and additional information to justifies these steps. The following processes may describe a CBR cycle, Fig. 3:

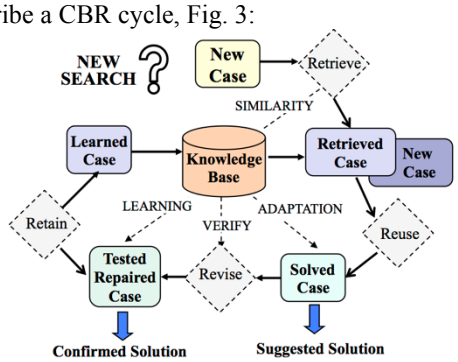

Figure 3. User profiles interface

- Retrieval: main focus of methods in this category is to find similarity between cases. Similarity function can be parameterized through system configuration.
- Reuse: a complete design where case-based and slot-based adaptation can be hooked is provided.
- Revise the proposed solution if necessary. Since the proposed result could be inadequate, this process can correct the first proposed solution.
- 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 knowledge base.

A. CBR Structure

The development of a quite simple CBR application already involves a number of steps, such as collecting case and background knowledge, modeling a suitable case representation, defining an accurate similarity measure, implementing retrieval functionality, and implementing user interfaces. Compared with other AI approaches, CBR allows to reduce the effort required for knowledge acquisition and representation significantly, which is certainly one of the major reasons for the commercial success of CBR applications. Nevertheless, implementing a CBR application from scratch remains a time-consuming software engineering process and requires a lot of specific experience beyond pure programming skills.

Although CBR claims to reduce the effort required for developing knowledge-based systems substantially compared with more traditional AI approaches. The implementation of a CBR application from scratch is still a time consuming task. We present a novel, freely available tool for rapid prototyping of CBR applications. CBR object-oriented framework development environments JColibri have been used in this study. By providing easy to use model generation, data import, similarity modeling, explanation, and testing functionality together with comfortable graphical user interfaces, the tool enables even CBR novices to rapidly create their first CBR applications. Nevertheless, at the same time it ensures enough flexibility to enable expert users to implement advanced CBR applications [9].

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. Our motivation for choosing this framework is based on a comparative analysis between it and other frameworks, designed to facilitate the development of CBR applications. jColibri enhances the other CBR shells: CATCBR, CBR*Tools, IUCBRF, Orenge. 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.

B. Retrieval of similar cases process

The main purpose of establishing intelligent retrieval ontology is to provide consistent and explicit metadata in the process of knowledge retrieval. CBR systems typically apply retrieval and matching algorithms to a case base of past search-result pairs. CBR is based on the intuition that new searches are often similar to previously encountered searches, and therefore, that past results may be reused directly or through adaptation in the current situation. Our system provides multilayer retrieval methods:

 ISBN: 978-1-61208-352-0

1. Intelligent profiles interface: Low-level selection of query profile options, which mainly include the four kinds of user. These users can specify certain initial items, i.e., the characteristics and conditions for a search. For this a statistical analysis has been done to determine the importance values and establishing specified user requirements. User searches are monitored by capturing information from different user profiles. This statistical analysis even can in fact lay the foundation for searches in a particular user profile.

2. Ontology semantic search can query on classes, subclasses or attributes of knowledge base, and matched cases are called back.

3. The retrieval process identifies the features of the case with the most similar query. Our inference engine contains the CBR component that automatically searches for similar queries-answer pairs based on the knowledge that the system extracted from the questions text. The system uses similarity metrics to find the best matching case. Similarity measures used in CBR are of critical importance during the retrieval of knowledge items for a new query. Similarity retrieval expands the original query conditions, and generates extended query conditions, which can be directly used in knowledge retrieval. Unlike in early CBR approaches, the recent view is that similarity is usually not just an arbitrary distance measure, but function that approximately measures utility.

We used a computational based retrieval where numerical similarity functions are used to assess and order the cases regarding the query. The retrieval strategy used in our system is nearest-neighbor approach. This approach involves the assessment of similarity between stored cases and the new input case, based on matching a weighted sum of features. A typical algorithm for calculating nearest neighbor matching is next:

$$similarity(Case_I, Case_R) = \frac{\sum_{i=1}^{n} w_i^{\times sim\left(f_i^I, f_i^R\right)}}{\sum_{i=1}^{n} w_i} \qquad (1)$$

Where w_i is the importance weighting of a feature (or slot), *sim* is the similarity function of features, and f_i^I and f_i^R are the values for feature *i* in the input and retrieved cases respectively.

The use of structured representations of cases requires approaches for similarity assessment that allow to compares two differently structured objects, in particular, objects belonging to different object classes. An important advantage of similarity-cased retrieval is that if there is no case that exactly matches the user's requirements, this can show the cases that are most similar to his query.

V. ONTOLOGY DESIGN AND DEVELOPMENT

We need a vocabulary of concepts, resources, and services for the knowledge system described. This scenario requires definitions about the relationships between objects of discourse and their attributes [10]. We have proposed to use ontology together with CBR in the acquisition of the knowledge in the specific DL domain. The primary information managed in the OntoSDL domain is metadata about library resources, such as books, digital services and resources, etc. 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 [11]. We wrote the description of these classes and the properties in RDF semantic markup language. RDF is used to define the structure of the metadata describing DL resources. OntoSDL 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. Our ontology can be regarded as triplet OntoSearch:={profile, collection, source) where profiles represent the user kinds, collection contains all the services and sources of the DL, and source cover the different information root: catalogue, history fond, intranet, Web, etc. We choose Protégé as our ontology editor, which supports knowledge acquisition and knowledge base development [12]. It is a powerful development and knowledge-modeling tool with an open architecture. Protégé uses OWL and RDF as ontology language to establish semantic relations.

In order to realize ontology-based intelligent retrieval, we need to build case base of knowledge with inheritance structure. The ontology and its sub-classes are established according to the taxonomies profile, as shown in Fig. 4.

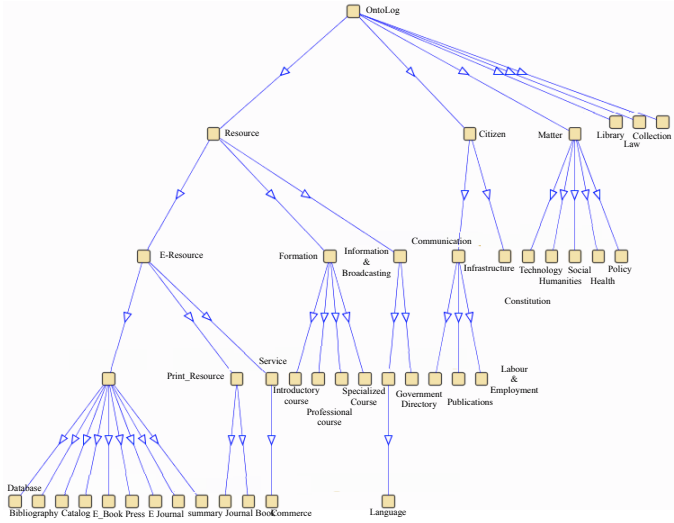

Figure 4. Class hierarchy for the OntoSDL ontology

This shows the high level classification of classes to group together OntoSDL resources as well as things that are related with these resources. As shown the Fig. 4, profile ontology includes several attributes like Electronic_Resources, Digital_Collections, Catalogue, Science_Resources, etc. After ontology is established, we need to add enough initial instances and item instances to knowledge base. For this purpose we followed these steps:

 ISBN: 978-1-61208-352-0

first we choose a certain item, and create a blank instance for item; second the domain expert, in this case the librarian fills blank units of instance according the domain knowledge. To finish, 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. This is sufficient for our proof-of-concept demonstration, but would not be sufficiently efficient to access large resource sets. Each case contains a set of attributes concerning both metadata and knowledge. However, our prototype is currently being extended to enable efficient retrieval directly from a database, which will enable its use for large-scale sets of resources.

VI. EXPERIMENTAL EVALUATION

Experiments have been carried out in order to test the efficiency of AI and ontologies in retrieval information in a DL. These are conducted to evaluate the effectiveness of run-time ontology mapping. The main goal has been to check if the mechanism of query formulation, assisted by an agent, gives a suitable tool for augmenting the number of significant documents, extracted from the DL to be stored in the CBR. The user begins the search devising the starting query. Suppose the user is looking for some resource about "Computer Science electronic resource" in the library digital domain of Seville, Fig. 5.

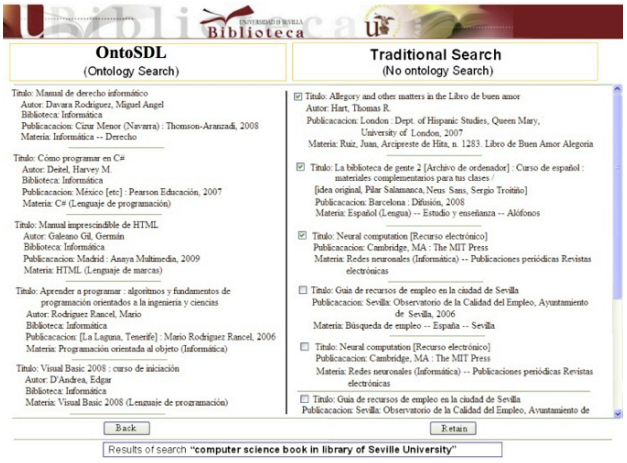

Figure 5. Search engine results page

The user inputs the keywords in the user profile interface. The required resources should contain some knowledge about "Computer Science" and related issues. After searching, some resources are returned as results. 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.

We have compared our prototype with some semantic search engines like Hakia, Lexxe, SenseBot, etc. However, we have focused in Google because is the world's dominant search engine and Google has made significant inroads in semantic indexing in search. It is a fact that deep inside Google is based on breakthrough semantic search techniques that are transforming Google's search results [13].

For our experiments we considered 50 users with different profiles. Therefore, we could establish a context for the users, they were asked to at least start their essay before issuing any queries to OntoSDL. They were also asked to look through all the results returned by OntoSDL 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 must consider that retrieved documents relevance is subjective. That is different people can assign distinct values of relevance to a same document. In our study, 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 OntoSDL. Next, 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).

TABLE I. ANALYSIS OF RETRIEVED DOCUMENTS RELEVANCE FOR SELECT QUERIES

	Excellent	Good	Acceptable	Poor
OntoSDL	7,50%	41,50%	40,60%	10,40%
Google	2,60%	27,90%	43,40%	26,10%

After the data was collected, we had a log of queries averaging 5 queries per user. Of these queries, some of them had to be removed, either because there were multiple results clicked, no results clicked, or there was no information available for that particular query. The remaining queries were analyzed and evaluated. These results are presented in Fig. 6.

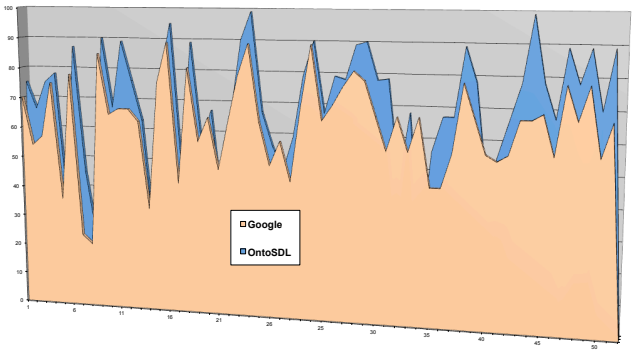

Figure 6. Search engine results page

In our study DL domain we can observe the best final ranking was obtained for our prototype OntoSDL and an interesting improvement over the performance of Google. Test of significance is the analysis of the number of

searches that have been resolved satisfactory by OntoSDL. As noted in Table 1 our system performs satisfactorily with about a 91.6% rate of success in real cases.

Another important aspect of the design and implementation of an intelligent system is determination of the degree of speed in the answer that the system provides. During the experimentation, heuristics and measures that are commonly adopted in information retrieval have been used. While the users were performing these searches, an application was continually running in the background on the server, and capturing the content of queries typed and the results of the searches. Statistical analysis has been done to determine the importance values in the results. We can establish that speed in our system improves the proceeding time and the average of the traditional search engine. The results for OntoSDL are 9.15% better than proceeding time and 11.9% better than executing time searches/sec in the traditional search engines.

VII. CONCLUSION AND FUTURE WORK

We have investigated how semantic technologies and AI can be used to provide additional semantics from existing resources in digital libraries. We described an effort to design and develop a prototype for management the resources in a library such as OntoSDL project, and to exploit them to aid users as they select resources. Our study addresses the main aspects of a Semantic Web knowledge retrieval system architecture trying to answer the requirements of the next-generation Semantic Web user. This scheme is based on the next principle: knowledge items are abstracted to a characterization by metadata description and it is used for further processing.

For this purpose we presented a system based in ontology and AI architecture for knowledge management in the Seville DL. 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. It introduced a prototype web-based CBR retrieval system, which operates on an RDF file store. 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 DL according to his own preferences, which come to represent his interests.

Finally, the study analyzes the implementation results, and evaluates the viability of our approaches in enabling search in intelligent-based digital libraries. The results 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. 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. It is also necessary the development of an authoring tool for user authentication, efficient ontology parsing and real-life applications.

REFERENCES

[1] D. Govedarova, S. Stoyanov, and I. Popchev, “An Ontology Based CBR Architecture for Knowledge Management in BULCHINO Catalogue” International Conference on Computer Systems and Technologies”, 2008.

[2] P. Warren. “Applying semantic technologies to a digital library: a case study”, "Applying semantic technology to a digital library: a case study", Library Management, Vol. 26 Iss: 4/5, pp.196 – 205, 2005.

[3] H. Stuckenschmidt and F. Harmelen, “Ontology-based metadata generation from semi-structured information”, K-CAP '01: Proceedings of the 1st international conference on Knowledge capture, 2001, pp. 163-170, doi:10.1145/500737.500763

[4] J. Toussaint and K. Cheng, “Web-based CBR (case-based reasoning) as a tool with the application to tooling selection”, The International Journal of Advanced Manufacturing Technology, 29(1-2): pp. 24 - 34, May 2006, DOI: dx.doi.org/10.1007/s00170-004-2501-0

[5] GAIA - Group for Artificial Intelligence Applications. “*jCOLIBRI project - Distribution of the development environment with LGPL”*, http://gaia.fdi.ucm.es/grupo/projects/,Complutense University of Madrid, January, 2014.

[6] Y. Sure and R. Studer, “Semantic web technologies for digital libraries”, Library Management Journal, Emerald, Library Management, Vol. 26 Iss: 4/5, 2005, pp.190 - 195

[7] H. Ding, “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, 2004, pp 851-854.

[8] M. Bridge, H. G¨oker, L McGinty, and B. Smyth, “Case-based recommender systems”, The Knowledge Engineering Review archive, Volume 20 Issue 3, September 2005, Pages 315 - 320, doi>10.1017/S0269888906000567

[9] B. Díaz-Agudo, P.A. González-Calero, J.Recio-García, and A Sánchez-Ruiz, “Building CBR systems with jColibri”, Journal of Science of Computer Programming, Volume 69, Issues 1–3, 1 December 2007, Pages 68–75, doi: dx.doi.org/10.1016/j.scico.2007.02.004.

[10] S. Staab and R. Studer, “Handbook on Ontologies. International Handbooks on Information Systems”, Springer, Berlin, 2005.

[11] W3C, *RDF Vocabulary Description Language 1.0: RDF Schema,* http://www.w3.org/TR/rdf-schema/, January, 2014.

[12] PROTÉGÉ, *The Protégé Ontology Editor and Knowledge Acquisition System,* <http://protege.stanford.edu/>, January, 2014.

[13] D. Amerland, “Google Semantic Search: Search Engine Optimization (SEO) Techniques That Get Your Company More Traffic, Increase Brand Impact and Amplify Your Online Presence, Que Publishing Kindle Edition, July, 2013.

 ISBN: 978-1-61208-352-0