Hybrid Ontology and CBR-based Search in BULCHINO Catalogue

Nadezhda Govedarova, Stanimir Stoyanov, Ivan Popchev

Abstract: *Our ongoing work on the development of CBR-based approaches for realizing content-based search is mainly targeted at discovering new interesting and powerful technological integrations that permit these technologies to exploit each others` strengths in order to mitigate their weaknesses. As a result, we developed a hybrid ontology and CBR-based search engine architecture that takes advantage of three-layered knowledge model for representing the domain specific data and capturing its semantics (data base, metadata repository and ontology) and supports several approaches for realizing the retrieval process over this model (computational-based retrieval, classification-based retrieval and concept-based retrieval). In this paper we present how the architecture could be specified for the Cultural Heritage domain and for BULCHINO catalogue in particular.*

Key words: *search engine, case-based reasoning, content-based retrieval, knowledge management, ontologies, Semantic Web technologies, jCOLIBRI framework, BULCHINO catalogue*

1. Introduction

One of most important aspects of intelligent information retrieval is the ability to understand the semantics of the information, i.e. the intended meaning of terms in special context or application [9]. The formalization and representation of the context and taking advantage of it by performing the retrieval process are, therefore, the main research challenges in designing and implementing content-based search engines. Current studies in this area [4,11,13] focus on the integration of numerous methods and techniques derived from Artificial Intelligence and Knowledge Management areas (e.g., metadata-descriptions, ontologies, similarity measures, and intelligent retrieval mechanisms) to achieve this functionality.

Since intelligent retrieval is one of the main application areas of Case Based Reasoning technology (CBR) [1, 14], semantics formalization in CBR systems has also become a topic of increased research [2, 7, 8, 11, 16]. In CBR, semantics serves as a major source for reasoning, similarity assessment, decision-making and adaptation. Consequently, achieving desired behaviors from CBR systems in these areas will depend on the ability to represent and manipulate information about domain context. A vast field of CBR models with respect to the synergy with other techniques for knowledge representation and retrieval already exists. Our main goal is to discover interesting and powerful new functional integrations that permit these technologies to exploit each others` strengths in order to mitigate their weaknesses.

Our ongoing work [18, 20, 21] aims at the development of a CBR-based search engine that is capable of performing metadata and ontology-based information retrieval in a hybrid fashion. The underlying hypothesis is that case-based reasoning supported by a rich knowledge model is a promising approach for achieving semantics aware search and retrieval. We use the cases as metadata description of the domain specific knowledge and integrate ontology as vocabulary for defining the case structure. Taking this hybrid approach as starting point, we designed an ontology and CBR-based search engine architectur and intend to integrate it in BULCHINO catalogue [19], a Web-based catalogue for electronic representation of Bulgarian cultural heritage. Taking into consideration that cultural heritage collections are usually accompanied by a rich set of metadata, several standards to describe them (CCO[1], CIDOC-CRM[2]) have already been developed and a set of common vocabularies and thesauri (ULAN[3], TGN[4], AAT[5]) are available, we consider this domain as suitable, with respect to knowledge abundance, and challenging enough for conducting our first experiments and tests in it.

---

[1] CCO (Cataloging Cultural objects): http://vraweb.org/ccoweb/cco/index.html

[2] CIDOC-CRM (Conceptual Reference Model, an official standard ISO 21127:2006): http://cidoc.ics.forth.gr

The search engine architecture adopts a knowledge model that is structured into three main layers: the data storage layer (database), the basic metadata descriptions layer (metadata repository, i.e. case base); the top domain conceptualization layer (ontology). It should be noted that different case structures could be defined. Standard cases are composed by several attributes with different simple data types (e.g. String, Integer). If the case structure is designed using types from the ontology for some of the attributes, a hybrid case should be processed. Pure ontology-based cases are also supported. Depending on the case structure the engine assumes three approaches for performing the retrieval process: computationa-based retrieval that implies local similarity assessment (between the attributes with simple types) and global similarity assessment between the cases (for dealing with standard cases); classification-based retrieval that makes use of the Description Logics reasoning capabilities (for dealing with ontological cases) and hybrid retrieval which implies the cooperation of standard retrieval methods with concept-based ones (for dealing with hybrid cases).

In specifying the software components of the architecture model we have identified the following design goals and challenges:

- Development of the knowledge model. To achieve this goal, we should first of all choose an approach for ontology integration and employment. Then we should define the cooperation principles with the other knowledge layers on the one hand and the retrieval methods on the other hand;
- Development of a reasoning algorithm for analogy-based reasoning over the knowledge model. The main challenge here is to find out how to take advantage of the formalized semantics in the ontology concept model without necessarily applying the DLs reasoning capabilities;
- Development of a model for the adaptation process from the CBR Cycle[14] since the ability of adaptation is one of the main characteristics of intelligent applications and is therefore a subject of interest for us ;
- Design or choice of already available development tools for building CBR-based applications;
- Adaptation of the architecture for a particular domain (initially for the Cultural heritage domain). This implies both defining the case structure and the BULCHINO knowledge model as a whole.

In this paper we focus on the last topic. We made, as well, an analysis of the available software tools for development of CBR systems and chose the jCOLIBRI framework [10, 12] for the technological implementation of the architecture logical model. Our choice is motivated by several aspects, discussed in the following.

The rest of the paper is organized as follows. In the next section we present briefly BULCHINO catalogue, our approach for ontology employment in it and discuss the search scenarios, which could be applied in the catalogue. Section 3 focuses on the architecture of the search engine, where we pay specific attention to the software framework jCOLIBRI, and the supplementary technologies that enables its cooperation with other Semantic Web technologies for dealing with otologies. A case model for cultural heritage domain and its mapping with data, stored in database and ontology is presented in Section 4. Finally, we conclude the paper and outline the directions of our future work in Section 5.

---

[3] ULAN (Union List of Artist Names): http://www.getty.edu/research/conducting_research/vocabularies/ulan
[4] TGN (Thesaurus of Geographic Names): http://www.getty.edu/research/conducting_research/vocabularies/tgn
[5] AAT (Art and Architecture Thesaurus): http://www.getty.edu/research/conducting_research/vocabularies/aat

2. BULCHINO Catalogue

BULCHINO (BULgarian Cultural HIstorical and Natural Objects) is a Web-based catalogue for electronic representation of the Bulgarian culture-historical heritage that has been developed at the E-Commerce Laboratory in the University of Plovdiv. A major aspect to ensure the interoperability of the catalogue is the usage of standard, platform independent technologies – Semantic Web technologies, established standards for describing the cultural objects, such as CCO and CIDOC-CRM and Web services. The access to the available services and electronic content is accomplished via specialized portal, whose architecture is developed in compliance with the referent model of Delphi Group [23].The portal reference architecture includes the following nine different layers: presentation, personalization, e-services control, integration, collaboration, search, categorization, and loop process layer. Our main goal is to improve significantly the efficiency of the search process and extend the search layer as integrating a CBR-based search engine in it, so that intuitive information access and content-based retrieval is provided.

2.1. CCO-based Knowledge Model

The cultural collections in BULCHINO and the objects in them are modeled in compliance with the CCO standard. According to it "a work is a distinct intellectual or artistic creation limited primary to objects and structures made by humans, including built works, visual art works and cultural artifacts" and could be linked with an image (its visual representation). The terminology related to the works and images and used for describing them is stored in authority files [3].

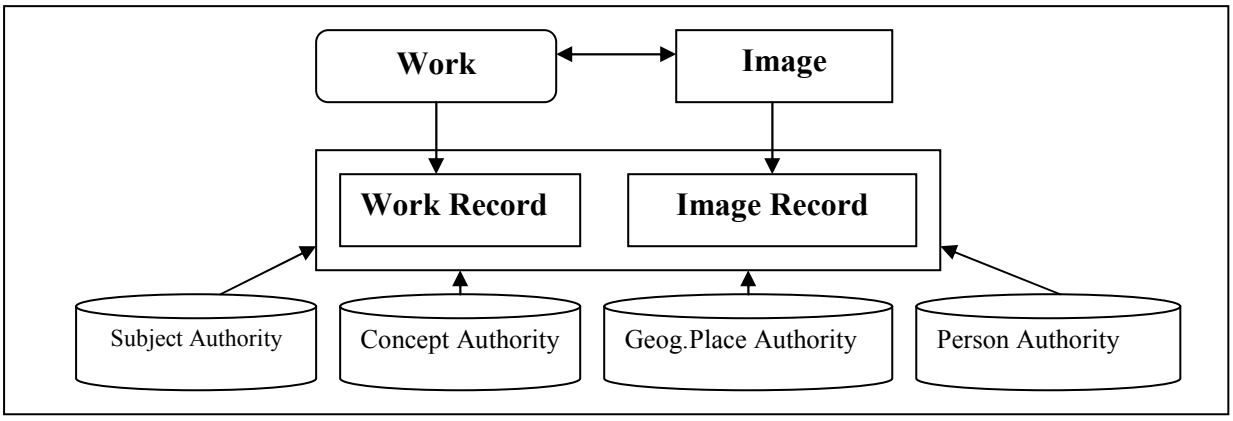

Fig. 1: Knowledge Classification Model

- *Work record* – the data is stored using required and additional recommended elements that give concrete information about the cultural-historical work;
- *Subject authority* – contains subjects from the iconographic terminology, names from literature, mythological and religious nature, historical event;
- *Concept authority* – contains terminology, needed for the description of work of art, the material that the work is made of, the activities involved with the work, style, role of the creator, etc;
- *Geographic place authority* – contains information about geographical locations of the cultural works and their creators;
- *Personal and corporate name authority* – contains information about architectural work, the individuals and corporations, connected with the cultural-historical work;
- *Image record* – contains an image of the cultural work.

2.2. Search Scenarios

Regarding the distributed collections of cultural objects, available in the catalogue, the search methods could be applied locally to every single collection and underlying work record (local queries) or globally to all collections supported in the catalogue (global queries). Taking into consideration these use cases, we have to develop search strategies which allow retrieval of single objects as well as of integrate tourist routes (a set of objects). We developed an approach, which implies the application of the CBR technology for metadata characterization of the cultural objects (presented as cases and stored in a case base), where the structure of the case is based on the defined by the Work record description, and the integration of ontology for semantics formalization of these characterizations. It, however does not provide any insight into the ontology architecture with respect to the distributed information sources, neither gives any directions for the search strategy. In the following we present our concept for employment of ontology in the knowledge model of BULCHINO catalogue and in Section 3 and 4 outline our basic ideas concerning the search strategy.

2.3. Ontology Employment Approaches

The typical information integration system uses ontologies to explicate the contents of an information source, mainly by describing the intended meaning of table and data field names [9]. For this reason we decided to integrate ontology in our knowledge model as a supplement which resembles and extends the structure of the information sources.

In general, three different groups of ontology-based integration approaches can be identified [9]:

- *Single-ontology approaches* – uses one global ontology providing a shared vocabulary for the specification of the semantics. All sources are related to the one global ontology;
- *Multiple-ontology approaches* – each information source is described by its own ontology;
- *Hybrid approaches* - the semantics of each source is described by its ontology, which are built upon one global shared vocabulary (Fig.2 A).

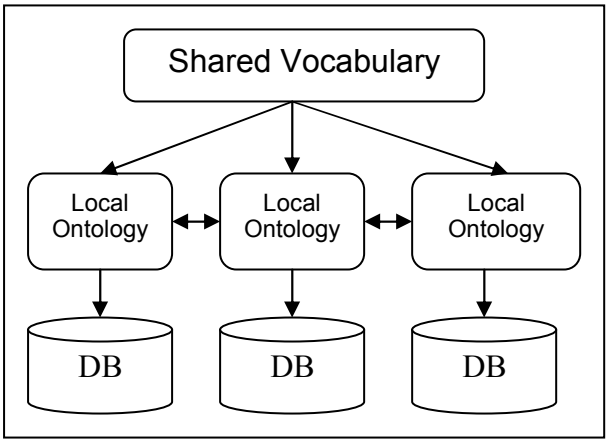

Fig. 2 A: Hybrid Approach

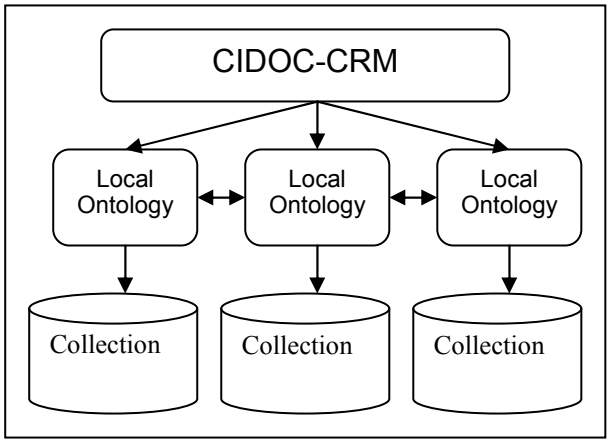

Fig. 2 B: Hybrid Approach in BULCHINO

The ability to exchange information at run time, also known as interoperability, is an important feature of BULCHINO, which allows the exchange of information between it and different cultural institutions, museums, universities, etc. It enables, as well, the usage of the catalogue in different areas and its inclusion in the European and worldwide cultural networks. In order such interaction to be possible, the system has to transfer its information in a common data framework (global shared vocabulary).

Taking into account the classification of data in BULCHINO in different thematic collections of cultural objects on the one hand and the required interoperability of the catalogue on the other hand, we consider developing the ontology infrastructure, basing on the third (hybrid) group of approaches (Fig. 2 B).

3. CBR-based Search Engine Architecture

For the implementation of the engine prototype we chose the framework jCOLIBRI-a Java-based framework that supports the development of knowledge intensive CBR systems and help in the integration of ontology in them. 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: CAT-CBR[5], CBR*Tools[15], IUCBRF [17], Orenge [26] in several aspects: availability (open source framework), implementation (the Java implementation is one of our main requirements with respect to the easy integration in the BULCHINO system which is implemented in J2EE [22] environment), GUI (the provided graphical tools facilitate the system design). Another decision criterion for our choice is connected with the fact that jCOLIBRI affords the opportunity to incorporate ontology in the CBR system to use it for case representation and content-based reasoning methods to assess the similarity between them. The ontology support of jCOLIBRI is built arround the OntoBridge library [25] that was developed by the GAIA research group [28] to easily manage ontologies and DLs (Description Logics) reasoners. It is base on the Jena framework for development of Semantic Web applications [24] and makes possible the connection with several DLs reasoners. Some of the actively supported once are presented in [6]: CEL, FaCT++, fuzzyDL, KAON2, MSPASS, Pellet, QuOnto. We chose Pellet [27] as inference engine in our architecture for several reasons. Apart from being open source and Java based, it supports Jena interface what enables the direct connection with the OntoBridge library of jCOLIBRI. For the manual generation and modeling of the domain ontology we chose the Protégé editor.

3.1. Data Types and Structures

The proposed architecture (Fig.2) is based on our approach for realization of content based retrieval of cultural objects by means of metadata characterizations and domain ontology inclusion. It implies to use ontology as vocabulary to define complex, multi-relational case structures to support the CBR processes. 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.

Except from the Concept type, the architecture takes advantage, as well, of another feature of jCOLIBRI framework – the two-layer organization of the case base. The metadata descriptions of the cultural objects (cases) are abstracted from the details of their physical representation in the Electronic Catalogue (Persistency media) and are stored in the case base (In-memory organization). This way the same methods can operate over different types of information repositories. The mapping (the process is described in the following) between the two layers is realized by connectors – objects that 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.

3.2. Retrieval Approaches

Our current work is focused mainly on the realization of the retrieval step from the CBR cycle. The main aim of the retrieval process is to obtain the most similar cases, given a query, where the underlying similarity functions (similarity model) play a crucial role. Depending on the case structure (the types of the attributes) three different retrieval approaches are supported by jCOLIBRI:

- *Computational-based retrieval* – is based on numeric similarity functions that assess the similarity between cases (global similarity), entirely composed of attributes with simple types (local similarity). An example (PersonExample) for application of this approach we presented in [21];
- *Classification-based retrieval* – is based on DLs classification capabilities, where two approaches are possible: concept classification and instance recognition;
- *Concept-based retrieval* – is based on numeric similarity functions, applied over ontology. When dealing with ontologies the concept hierarchy influences the similarity assessment, since the class hierarchy contains knowledge about the similarity of the objects. Four functions are provided by jCOLIBRI to compute the concept-based similarity that depends on the location of the cases in the ontology [11, 12].

3.3. Components

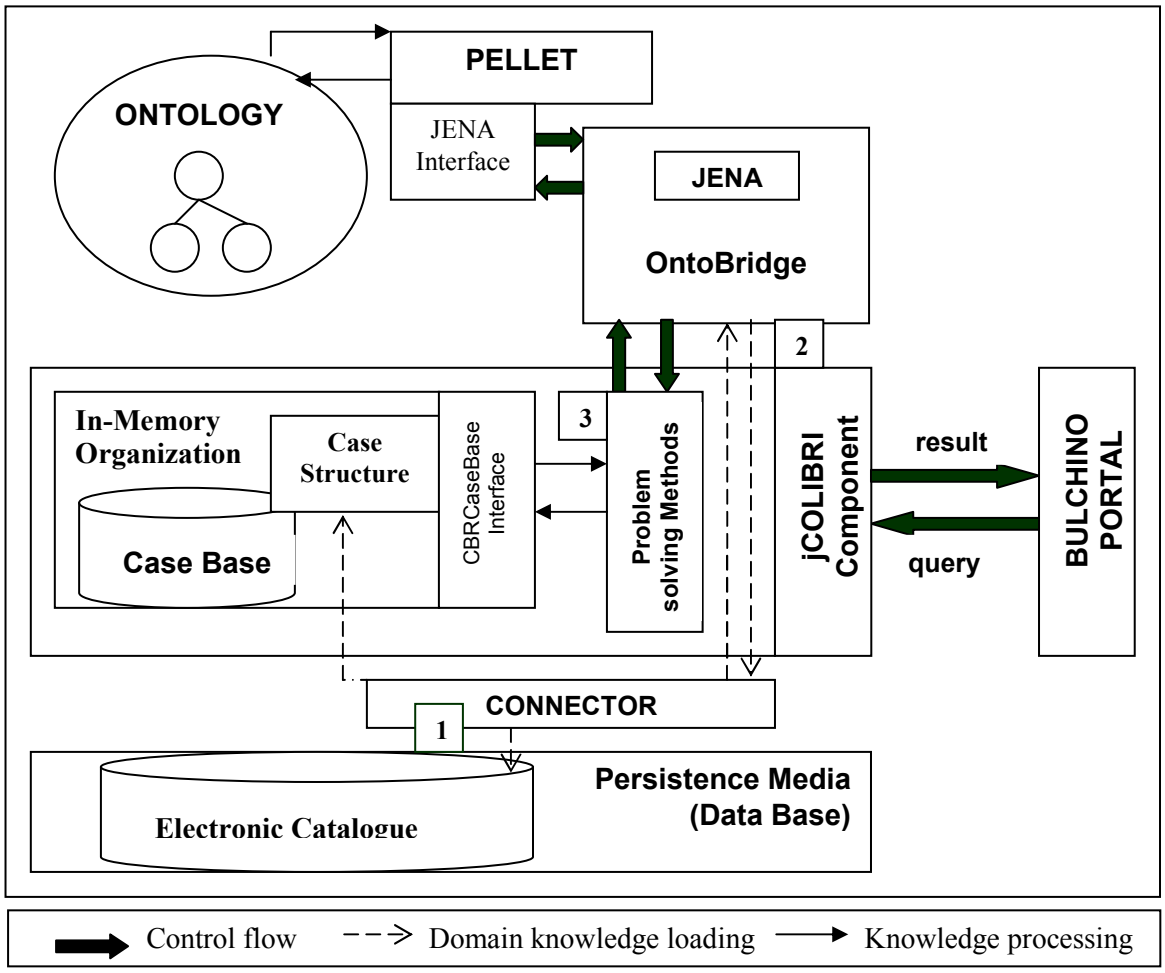

Fig.3: CBR-based Search Engine Architecture

- Persistence Media - in this case we are not storing the whole case base into the ontology:

- Data Base continues storing the values of the attributes. But now the column that stores the attribute linked with a concept is going to have the names (stored as strings of characters) of the instances of this concept.

- Ontology stores information about cultural objects where concepts are types, or classes, individuals are allowed values, or objects and relations are the attributes describing the objects. Knowledge is formally represented and stored in OWL (Web Ontology Language) ontology that is managed with the Pellet reasoner;

- jCOLIBRI Component - is designed on the basis of jCOLIBRI framework.

It is the core of the architecture and is responsible for the knowledge and queries management. Since the problem solving methods are domain independent, the domain specific information should be first loaded from the persistence media so that processing with it is possible. The data base connector will read the values in the table (1) and if encounters a concept typed attribute it looks for an instance with the same name in the Ontology (through OntoBridge) (2). Once found the connector will fill the values of the attribute of each case with the corresponding instances of the ontology, loaded by the Pellet reasoner. It is used as well by the methods (3) to compute the content-based similarity between the concept typed attributes;

- OntoBridge - is a Java library that eases the management of the ontology in a jCOLIBRI-based application. It uses Jena library to implement most of the required methods for accessing the ontology, loaded in the reasoner. With this extension the jCOLIBRI component can acquire domain knowledge from ontology and achieve this way uniform case representation, what will enhance the interoperability of the whole system. This, namely, was our main goal when designed the jCOLIBRI component as BULCHINO search layer extension;

- Pellet – is an open source Java-based OWL-DL reasoner.

4. Example

According to the search scenarios, presented in Section 2.2, in BULCHINO catalogue retrieval of single objects as well as of an integrate tourist route is possible. Since we have not defined, yet, the structure and formal representation of the tourist route concept for this particular system, we propose to illustrate the practical applicability of our approach for content-based retrieval by applying it for retrieving initially only single objects. In this Section we present an abstract model of the case which we intend to use as metadata description of the cultural objects in BULCHINO catalogue.

4.1. Case Structure

CCO recommends a minimal set of core metadata elements, which comprise the most important descriptive information, necessary to make a record for a work and an image [3]. The structure of the case in our example consists of 9 attributes, which we define on the basis of this recommended set of elements. The mapping between the case structure and the instances of its attributes, stored both in Data Base and ontology, is described in Section 3.1 and 3.3.

For the similarity assessment between content-based attributes we apply specific functions (inter-class and intra-class similarity functions) [12] that compute the similarity on the basis of the depth of the given attributes in the ontology hierarchy.

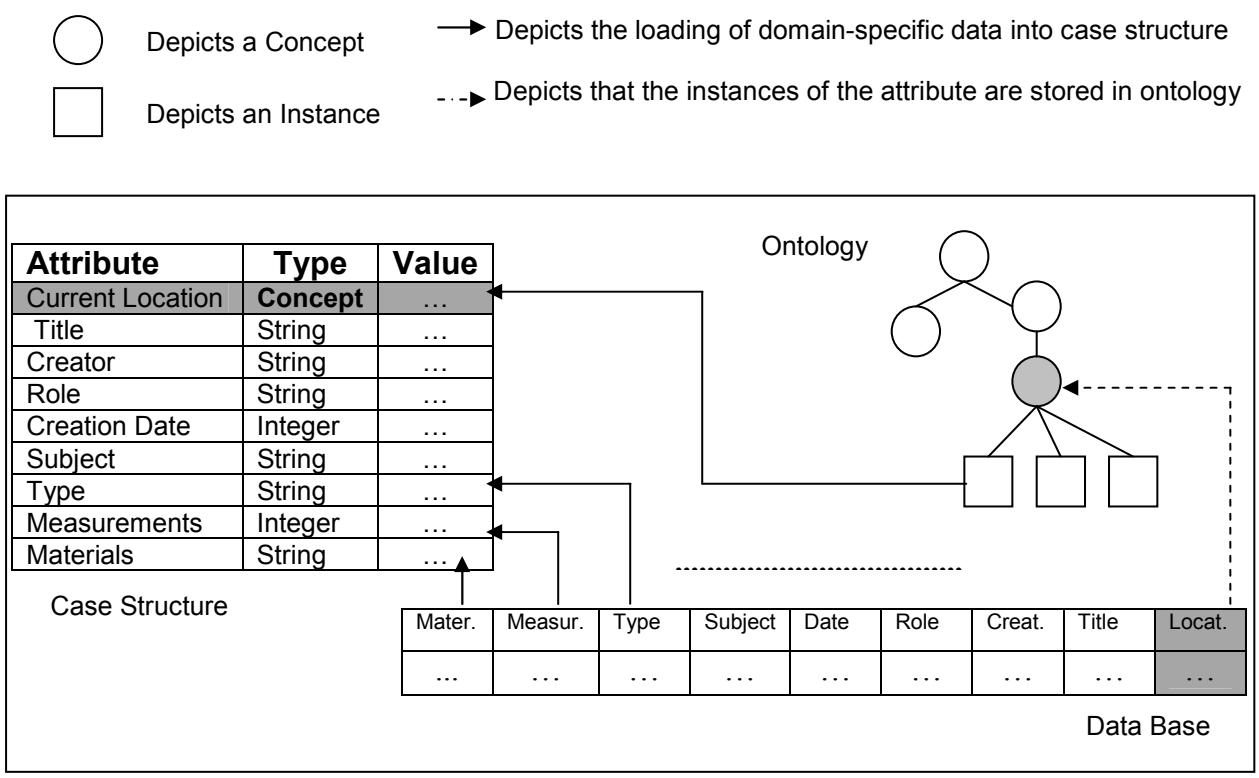

Fig.4: Mapping between Data Base, Case Structure and Ontology

5. Conclusions and Future Work

We believe that by combining analogical processing (case based reasoning), large knowledge bases, structured and presented in compliance with the up to date Semantic Web technologies and standard approaches for information retrieval, we can develop new powerful tools for intelligent information retrieval. The main contribution of this paper is the presentation search engine architecture that adopts an approach for retrieving information, based on case-based metadata characterisations of the information and ontology conceptualisation of the domain at hand - Cultural heritage domain.

Future work includes the specification of the architecture and the integration of actual cultural content in the developed knowledge models. We are investigating as well the addition of supplementary functionality of the engine and namely the generation of the tourist routs, which structure should be designed in compliance with the already developed knowledge models. This will help us to gather the experience and the experimental data about the architecture that will be necessary to evolve it, as well as discovering what new problems lie ahead.

Acknowledgements

The authors wish to acknowledge the financial support of the Bulgarian Ministry of Education and Science and Research. Project Ref.No.MI-1502/2005.

References

[1] Aamodt, A., Plaza, E., Case-base reasoning: Foundational issues, methodological variations, and system approaches, AI Communications, 7(1), 1994

[2] Anders Kofod-Petersen, Challenges in CBR for Context Awareness in Ambient Intelligent Systems, Proceedings of the 1st Workshop on Case-based Reasoning and Context Awareness (CACOA) co-located with the 8th European Conference on Case-Based Reasoning (ECCBR) September 5, 2006, Ölüdeniz/Fethiye, Turkey

[3] Cataloging Cultural Objects, A guide to Describing Cultural Works and Their Images, 2005, http://www.vraweb.org/ccoweb/cco/index.html (to date)

[4] C. Doulaverakis, E. Nidelkou, A. Gounaris, Y. Kompatsiaris, An Ontology and Content Based Search Engine For Multimedia Retrieval, 10th East-European Conference on Advances in Databases and Information Systems, ADBIS 2006, Thessaloniki, Hellas, 3-7 September, 2006

[5] C.Abasolo, E.Plaza, and J.-L.Arcos, Components for case-based reasoning systems, Lecture Notes in Computer Science, 2504, 2002

[6] Description Logics Reasoners, http://www.cs.man.ac.uk/~sattler/reasoners.html

[7] Díaz-Agudo, B., González-Calero, P.A., An Architecture for Knowledge Intensive CBR Systems, In Blanzieri, E., Portinale, L., (Eds.): Advances in Case-Based Reasoning (Procs. of the 5th European Workshop on Case-Based Reasoning, EWCBR 2000), Lecture Notes in Artificial Intelligence, 1898, Springer, 2000

[8] Gómez-Gauchía, H., Díaz-Agudo, B., González-Calero, P.A., "Ontology-Driven Development of Conversational CBR Systems". In Roth-Berghofer, T.R., Göker, M.H., Güvenir, H.A., (Eds.): Advances in Case-Based Reasoning, Procs. of the 8th European Conference on Case-Based Reasoning, ECCBR 2006. Lecture Notes in Artificial Intelligence, 4106, Springer, 2006

[9] Heiner Stuckenschmidt, Frank van Harmelen, Information Sharing on the Semantic Web, Springer Verlag, 2005, ISBN 3-540-20594-2

[10] jCOLIBRI Theoretical Foundations, http://gaia.fdi.ucm.es/projects/jcolibri/docs.html/

[11] Juan A. Recio-García, Belén Díaz-Agudo, Pedro A. González-Calero, and Antonio Sánchez, Ontology based CBR with jCOLIBRI, Applications and Innovations in Intelligent Systems XIV.Proceedings of AI-2006, the Twenty-sixth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, pages 149–162, Cambridge, United Kingdom, December 2006. Springer

[12] Juan A. Recio-García, Belén Díaz-AgudoPedro, González-Calero, jCOLIBRI2 Tutorial, Document version 1.1, Jenuary 22, 2008

[13] Forbus, K., Birnbaum, L., Baker, J., Wagner, E., Witbrock, M., Analogy, Intelligent IR, and Knowledge Integration for Intelligence Analysis: Situation Tracking and the Whodunit problem, in: Proceedings of the 2005 International Conference on Intelligence Analysis

[14] Lenz, M., Bartsch-Sporl, B., Burkhard, H., and Wess, S. Case-Based Reasoning Technology – From Foundation to Applications,Lecture Notes in Artificial Intelligence 1400, Springer Verlag, 1998

[15] M.Jaczynski and B.Trousse, CBR*Tools: An object-oriented framework for the design and the implementation of case-based reasoners, In Proceedings of the 6th German Workshop on Case-Based Reasoning, 1998

[16] Ralph Bergman, Martin Schaaf, On the Relation between Structural Case-Based Reasoning and Ontology-Based Knowledge management, In Proc. of German Workshop On Experience Management, April, 2003

[17] S.Bogaerts and D.Leake, IUCBRF: A Framework For Rapid And Modular Case-Based Reasoning System Development, November, 2004

[18] Stoyanov S., I. Popchev, D. Chaushkova, M. Trendafilova, A Case-based reasoning Approach for Development of Intelligent Services. Journal “Information Technologies and Control”, No. 3/2004, Year II, Pp. 31-34, ISSN 1312-2622

[19] S. Stoyanov, M. Trendafilova, E-Catalogue “Culture-historical heritage and nature objects in Bulgaria”, Conference “New education technologies”, 16-17 May, 2003, Sofia, 289 – 298 (in Bulgarian)
[20] S.Stoyanov, M. Trendafilova, CBR-Search in Electronic Catalogues, In Proc. of the International Conference “Automatics and Informatics ‘03”, vol.1, Pp.65-68, 6-8 October, 2003, Sofia, Bulgaria. ISBN 954-9641-34-1
[21] S.Stoyanov, N.Govedarova, I.Popchev, CBR-based Search in BULCHINO Catalogue, In Proc CS&P`07, vol. 2, Pp 521-533

[22] http://java.sun.com/j2ee/1.4/docs/tutorial/doc (to date)
[23] http://www.delphigroup.com (to date)
[24] http://jena.sourceforge.net/index.html (to date)
[25] http://gaia.fdi.ucm.es/grupo/projects/ontobridge (to date)
[26] http://www.ovitas.com/PDF/orengeWhitepaper.pdf (to date)
[27] http://pellet.owldl.com (to date)
[28] http://gaia.fdi.ucm.es/index.html (to date)

About the Authors

Nadezhda Govedarova, PhD student, Institute of Information Technologies, Bulgarian Academy of Science, Phone + 359 885 945 669, E-mail: nadig@abv.bg

Assoc.Prof. Stanimir Stojanov, PhD, E-commerce Laboratory, University of Plovdiv, Phone: +359 888 318 164, E-mail: csstani@pu.acad.bg

Academician Ivan Popchev, Dr.Sc., Institute of Information Technologies, Bulgarian Academy of Science, Phone: +359 2 979 32 33, E-mail: ipopchev@iit.bas.bg