An Ontology for Integrated Searching of Hypermedia in a Distributed Server Environment

C. Alexakos[1], B. Vassiliadis[2], K. Votis[1] and S. Likothanassis[1]

[1] Pattern Recognition Lab., Dept. of Computer Engineering and Informatics,
University of Patras, Greece,
{alexakos,botis,likothan}@ceid.upatras.gr
[2] Computer Science, Hellenic Open University, Greece,
bb@eap.gr

Abstract

One of the challenges of searching in a distributed environment is to achieve results of high relevancy to user queries. This means that differences in the description of data stored in different locations have to be overcome. In this paper we present an advanced search mechanism for locating digital content stored in distributed servers using a semantic representation mode. Ontologies, that form a scalable description of knowledge domains, facilitate searching tasks in the metadata level, where information concerning web digital content is published, managed and stored.

Keywords: Information Retrieval, Hypermedia, Ontologies, Distributed Servers.

1. Introduction

Different ways of understanding (and describing) things is the source of much havoc on both the real and the virtual world. In the first case, it makes the life of people difficult and in the second, the life of search engine developers [Rutledge et al. (2003)]. The problem of heterogeneity of information in on-line digital libraries is actually a problem of description and not the information per-se. It is difficult for people to understand each others descriptions of information on-line and much more difficult for machines to do so.

In this work we tangle with the problem of searching amongst distributed servers that hold hypermedia objects, having in mind the problem of description heterogeneity. Hypermedia, being complex in nature are much more difficult to be described by way of analyzing their content. In the contrary, text (and hypertext) can be described by

keywords that are easily extracted. In order to enable machine-to-machine interaction that in turn will facilitate truly efficient searching in a distributed environment; we turn to semantics and more specifically ontologies for facilitating searching. We describe a scheme for searching hypermedia sources using a multi-layer ontology. Searching tasks are carried out in the metadata level, where information concerning hypermedia objects is published, managed and stored. In section 2 we briefly revisit the most significant works in the field while in sections 3 and 4 we present the ontological search model. In sections 5 and 6 the services and architecture of a draft system are discussed and in section 7, a use case is presented.

2. Related Work in Semantic-enhanced searching

In this work we view hypermedia searching as a classic searching process of complex digital artifacts on-line. Unlike multimedia information retrieval, which takes place usually on a local server (or a cluster of local servers), we focus only on semantically enhanced searching over the internet or a very large corpus of distributed servers.

The need for good search and query mechanisms has been an early goal for the hypermedia community [Vitali and Bieber (1999); Wiil et al. (1999)]. Unlike hypertext, content analysis of hypermedia is much more difficult. Images, video and audio need special analysis techniques which are time-consuming and not as efficient as users are used to. This is the reason why commercial multimedia search engines have not appeared yet; the Google, being a tremendous success in hypertext searching, employees an image search engine that relies again on hypertext analysis and not analysis of images themselves. This indicates that on-line analysis of hypermedia content per se is not ready for wide use yet.

With the advent of the Semantic Web and its enabling technologies, a plethora of techniques, ontologically principled mechanisms and frameworks have been presented for hypermedia searching. Proposals such as the one of [Montero et al. (2003)] suggest that semantics should be included in the conceptual modelling stage of hypermedia production. Topia [Rutledge et al. (2003)] is an architecture for domain-independent processing of semantics and discourse into hypermedia presentations. In [Larsen, and Bouvin (2004)], a framework and a searching algorithm for locating distributed hypermedia in a P2P network is presented. Approaches such as semantically indexed hypermedia [Tudhope and Cunliffe (1999)] and ontology-based linking [Carr et al. (2001)] are also worth mentioning.

Furthermore, the introduction of the Semantic Web and its, nearly, unanimous approval has driven towards the representation of ontologies in semantics. For this purpose, a variety of semantic mark-up languages has been developed, based on the dominant XML standard. The most prominent ontology markup languages are DAML+OIL and its successor OWL, which is build on top of RDF Schema. The integration of hypermedia and semantic web technologies has been proposed both for

standard [Nanard et al. (2003)] and open hypermedia configurations [Dolog et al. (2003)], for ontology matching [Doan et al. (2003); Choi et al. (2006)] and management [Maedche et al. (2003)] while searching in distributed environments. Ontology schemes have been successfully used/designed for centralised management of images [Mezaris et al. (2003)], XML document schemata [Lu et al. (2003)], MPEG-7 semantically enriched content [Westermann and Klas (2003)] and query processing in P2P [Stuckenschmidt (2005)].

Open problems arise when integration is focused on the contents and not on the heterogeneous semantic metadata: lack of adequate knowledge representation methods, time delays in searching distributed hypermedia sources and restrictions due to the specificity of current ontology schemes used for searching [Klampanos et al., (2006)]. One of the initial solutions to distributed hypermedia content integration was the use of a single global ontology scheme in order to describe all the contents of hypermedia systems. This centralised approach has proven to be inflexible since sources are exponentially increasing affecting query response times. On the other hand, content-based integration faces three distinct difficulties. First, the utilization of a single description scheme forces the usage of a specific ontology for the semantic annotation of hypermedia contents. Second, when the hypermedia information is provided by distributed servers, the amount of semantic descriptions is increasing according to content volume. Third, in some cases ontologies used are domain-specific. Specificity deters widespread use.

3. Ontological Search Model

Our approach introduces an integrated search model that is utilized based on the metadata information that is describing the digital content is its server. The utilization of ontologies in the proposed model is based on the feature that they give a specific and structural description of a domain, where its terms and their relationship is clearly defined.

The searching model is composed of two functional parts, the first part is the ontology scheme and the second is a set of instances of the ontology scheme containing the metadata of the integrated digital data. The distinction of these two parts is similar with the T-Box and A-Box distinction which is drawn in Description Logics [Baader et al. (2003)].

The ontology scheme, that is called Integrated Data Ontology (IDO), is primarily used as a "catalogue" for the type of contents and the domain of knowledge that are integrated. Knowledge domains are specified in classes and subclasses providing a hierarchical model presenting all the knowledge fields that are included in the hypermedia contents of the distributed servers. There are also a number of properties denoting the relationship between classes. The semantic representation of knowledge domains follows the same level of detail with the semantic metadata schemes used by

the hypermedia servers. For example, in a system that provides books the metadata is structured by elements as “Book Title”, “Book Author” and “Book Abstract”. According to IDO scheme there is class “Book” with the corresponding metadata elements as properties, namely there are properties of the class “Book” as “Title”, “Author” and “Abstract”. This ensures that all terms and their relationships utilized by each digital content provider server separately are included in the ontology scheme.

The set of instances, named as IDO Instances, comprises the content of the metadata information provided for the description of the distributed digital data in terms of the IDO ontology scheme. The IDO instances are constructed according to mapping information that depicts the association of the metadata elements to the terms of the IDO scheme. The content of the metadata is transformed by creating instances of the classes of the IDO scheme and fill them with the metadata data according to the mapping information. In the example of books, there is an IDO instance of each book where the property of “Title” takes the values of the metadata element “Book Title”. This transformation simplifies the searching procedure by carrying out the search in ontology instances instead of composing different search queries to the distributed data systems.

The semantics of the Integrated Data Ontology and IDO Instances are composed according to the Ontology Web Language (OWL) [W3C (2004)] standard, a recommendation from W3C as a semantic mark-up language for representing ontologies based on the dominant XML/RDF standard.

4. Constructing the Search Model

As presented above our approach introduces an ontological search model in order to both index the digital data and integrate the metadata description information. In this context, a complete description of different digital contents is possible. What is still missing is the introduction of the metadata information to the semantic information represented in this model. A methodology is required to address the introduction needs between the different implemented metadata information of a digital content provider system and the proposed ontological model providing the aforementioned semantic information.

The first step in this methodology concerns the composition of the Integrated Data Ontology (IDO) with additional information provided from the digital catalogue and the metadata structure of each digital content system. This step involves the enrichment of the Integrated Data Ontology with new classes and properties in case there do not exist in the already developed ontology scheme.

The second step in our methodology regards the transformation of the actual metadata information in the different digital content systems according to the Integrated Data

Ontology scheme. This transformation leads to the enhancement of the IDO Instances with the metadata information of the data that is stored in a new imported system. This step includes the mapping of the terms of the systems metadata structure to the already developed ontology terms in the Integrated Data Ontology. This mapping is related to the description, in ontology terms, of the metadata content (e.g. which server provides what). Furthermore, an update of the IDO Instances with data, according to the mapping information of the actual metadata content, takes place.

This methodology leads to the desirable unification of the ontological search model with the metadata described digital contents, and thus contributes to the creation of a flexible and reusable ontology search scheme. It is essential to mention that the same methodology is adopted in the case of digital content update in one of the distributed servers. The synthesis of the Integrated Data Ontology and the IDO Instances constructs a model that contains all the necessary semantic information for searching and locating the desirable data avoiding the executing of a query to the distributed servers.

5. Creating the appropriate Web Services

The ontological search model presented in the previous section utilizes an integration solution to the searching process in an amount of data located in distributed server. The next step in our scheme is to present the data extraction and transfer from the corresponding system. In this context, we propose the use of web services implementation in order to open up the distributed digital content servers.

In [W3C, (2003)], WSs are referred as "software applications identified by a Uniform Resource Identifier (URI), whose interfaces and binding are capable of being defined, described and discovered by XML artifacts and support direct interactions with other software applications using XML-based messages via Internet-based protocols". WSs are reusable software building blocks distributed over the Web easily accessible via widespread protocols like HTTP and SMTP. They are loosely coupled, communicating through XML based documents.

According to the prospects that are supported by the web services, each of the integrated servers is enhanced with SOAP in order to provide a flexible and standard based service for transferring the desired digital content. A client script is also installed in the integrated middleware in order to invoke the provided web services from the distributed systems.

6. The Searching Process

The proposed scalable scheme provides the necessary supporting mechanism to a search engine that helps navigating through the ontology terms and instances fast and

efficiently. It also transfers the resulted data from the associated systems. The basic concept is the separation of the search process in three steps.

In the first step, the engine searches in the Integrated Data Ontology aiming to find the proper general domain (or domains) where the results may lay and the appropriate structural ontology terms that describe the corresponding semantic metadata structure.

The search engine extracts the basic structural elements of the semantic metadata that will be used in the hypermedia content searching. In the second step, a query is composed according to the RDF Data Query Language (RDQL) querying language for searching in the IDO Instances. RDQL is querying language for RDF structured documents, which is the language representing the OWL individuals (instances) of OWL Ontologies. The RDQL queries consist of a graph pattern, expressed as a list of triple patterns and also use a set of constraints on the values of those variables, and a list of the variables required in the answer set.

In the final step, the elaboration of the results of the RDQL query takes place in order to locate the servers that contain the resulted data. Furthermore, the appropriate client scripts are executed in order to invoke the corresponding web services and acquire the desirable digital content.

7. The Architecture

The use of the ontology search model requires a specific architecture that makes it possible to integrate the different semantically annotated digital data residing in different servers in a flexible and interoperable way. Our approach introduces a middleware for developing the multilayer ontology scheme and managing the search queries from the users. This middleware involves the following functional elements:

- The *Ontology Model Constructor* that provides the mechanism and the graphical interface for composing the Integrated Data Ontology and creates the mapping information between the ontology terms and the metadata content. It also provides a functional module for transforming the actual metadata information in order to upgrade the IDO Instances according to the corresponding mapping information.
- The *Query Graphical Interface* that provides the appropriate search forms to the user of the integrated system.
- The *Ontology Model Query Tool* that is responsible for the execution of the two first steps of the search process to the ontology search model.
- The *Integrated Systems Data Mediator* is the platform which accepts the results of the queries in the ontology search model and invokes the appropriate web services in order to acquire the desirable data.

The integration consists of two discrete phases: a) the Construction Search Model Phase where utilization of the functionalities of the Ontology Model Constructor takes place, the Integrated Data Ontology is defined and the IDO Instances are updated, and b) the Search Process Phase where a guest user uses the Query Graphical Interface to compose queries that are processed by the Ontology Model Query Tool. The Integrated Systems Data Mediator is responsible for allocating and bringing the results to the user. The overall architecture is depicted in Figure 1.

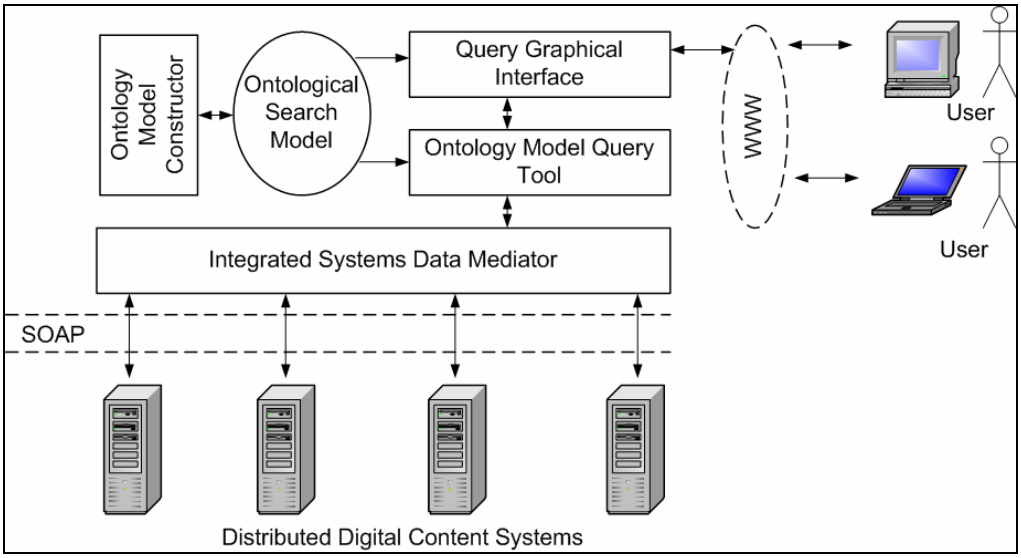

Figure 1. *Proposed Architecture*

8. A Use case

In this section we present a use case that makes use of our architecture. We have chosen two digital hypermedia servers with cultural content (digital images and hypertext). These two digital content servers contain cultural content as books, paintings and architectural monuments. The content in both servers is indexed and described with two different XML metadata schemes, with the first server using the Dublin Core standard. For the simplicity of the presentation of the use case, a demonstration of a searching query about finding books of a specific author will be used for the presentation of the two phases of our technique.

8.1 Construction Search Model Phase

In this section we present a part of the composed Integrated Data Ontology that is describing the digital data related to books. The ontology scheme contains three classes named “Cultural_Heritage”, “Regional_Heritage” and “Book”. The relationships between these classes are represented by the object property

"HasPublications" as depicted in Figure 2. The "Book" class is characterized by a set of data properties: "Title", "Abstract", "Author" and "NumberPages".

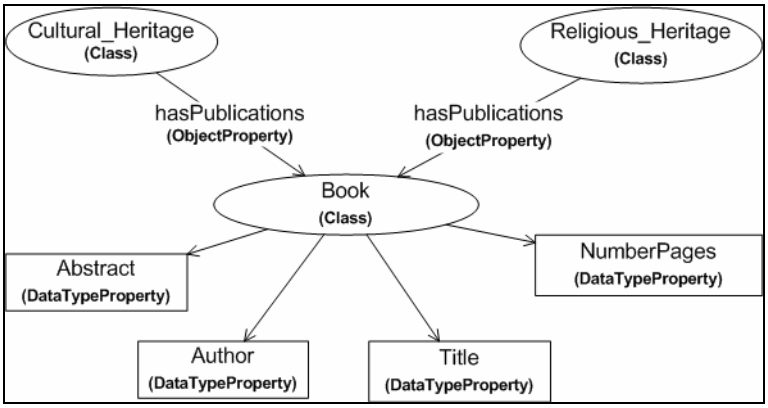

Figure 2. *Integrated Data Ontology for books*

The metadata of the two e-culture servers are mapped to the ontology. The mapping includes the association of the elements of metadata schemes to the data properties of the "Book" class. For example, according to the Dublin Core standard, each book is described by the metadata elements "title", "creator", "description", etc. Using the Ontology Model Constructor the following mapping information is created: "title" is associated with the property "Title", "creator" with "author" and "description" with "abstract". According to this mapping information, the metadata contents from the two systems are transformed in order to upgrade the IDO Instances set. The execution of this action denotes the completeness of the Construction Search Model Phase.

8.2 Search Process Phase

The Process Phase in this use case includes the search process that will query the integrated digital content servers to retrieve the data that is associated with the books of a particular author named "Lord Byron". A guest user queries the system using the forms of the user interface and query data are fed to the Ontology Model Query Tool. The user's query is transformed to a RDQL query that is depicted in Table 1.

Table 1. *Search Book RDQL query*

```
SELECT  ?booktitle
WHERE
(?book ido:Author ?x_author)
(?book ido:Title ?booktitle)
```

```
AND ?x_author == “Lord Byron”


USING ido FOR “http://prlab.ceid.upatras.gr/ido#”
```

The result of the query is used as an input to the Integrated Systems Data Mediator that stores an identifier for each instance in the IDO Instances. This instance is used for recognizing the server where the corresponding data are located. The two servers that are used in this use case have been integrated utilizing web services technology. There are, in both cases, SOAP-enhanced servers that publish information retrieval services at the web. Also, the Integrated Systems Data Mediator stores the corresponding client scripts in order to invoke the provided web services of the servers. The Integrated Systems Data Mediator locates the server (or servers) where the data is stored. The last step is to invoke the appropriate web method from the two digital libraries for data included in the query result. The data is transferred from the two systems and is presented to the user.

9. Conclusion

We presented an Ontology-based Integration Mechanism for integrating distributed digital artefacts. The process introduces an ontological search model in order to both indexing the digital data and integrating the metadata description information. A search methodology process has been defined for the desirable unification of the ontology model and a prototype based on searching a set of distributed cultural heritage databases has been implemented to test the feasibility of the mechanism. As we progressed through the implementation some interesting issues emerged. We plan to extend our current model in order to incorporate additional types of edges and domain functions into the ontology model.

References

Carr, L., Hall, W., Bechhofer, S., Goble, C., (2001). *Conceptual linking: ontology-based open hypermedia*. Proceedings of the 10th international conference on World Wide Web, pp. 334 – 342.

Choi, N., Song, I., and Han, H. 2006. *A survey on ontology mapping*. SIGMOD Rec. 35,3.

Doan, A., Madhavan, J., Dhamankar, R., Domingos, P., Halevy, A., (2003). *Learning to match ontologies on the Semantic Web*. The VLDB Journal - The International Journal on Very Large Data Bases, Vol. 12(4), pp. 303 – 319.

Dolog, P., Henze, N., Nejdl, W., (2003). *Logic-Based Open Hypermedia for the Semantic Web*. In Proc. of International Workshop on Hypermedia and the Semantic Web, Hypertext 2003 Conference, Nottingham, UK.

Klampanos, A., Poznański, V., Jose, J.M., Dickman, P. (2005). *A suite of testbeds for the realistic evaluation of peer-to-peer information retrieval systems*. Lecture Notes in Computer Science, 3408, pp.38-51.

Larsen, R.D., Bouvin, N.O., (2004). *HyperPeer: searching for resemblance in a P2P network*. Proceedings of the 15th ACM conference on Hypertext & Hypermedia, pp. 268 – 269.

Li, WS., Candan, K.S. (1999). *Integrating content search with structure analysis for hypermedia retrieval and management*. ACM Computing Surveys (CSUR), Volume 31, Issue 4es, Article No. 13.

Lu, E.J.L., Jung, Y.M., (2003). *XDSearch: an efficient search engine for XML document schemata*. Expert Systems with Applications, 24, pp. 213–224.

Maedche, A., Motik, B., Stojanovic, L., (2003). *Managing multiple and distributed ontologies on the Semantic Web*. The VLDB Journal - The International Journal on Very Large Data Bases, Vol. 12(4), pp. 286 – 302.

Mezaris, V., Kompatsiaris, I., Strintzis, M.G., (2004). *Region-based Image Retrieval using an Object, Ontology and Relevance Feedback*. Eurasip Journal on Applied Signal Processing, Vol. 2004 (6), pp.886-901.

Montero, S., Diaz, P., Aedo, I., Dodero, J.M., (2003). *Toward Hypermedia Design Methods for the Semantic Web*. Proceedings of the 14th International Workshop on Database and Expert Systems Applications, pp. 762.

Nanard, M., Nanard, J., King, P., (2003). *IUHM: a hypermedia-based model for integrating open services, data and metadata*, Proceedings of the 14th ACM conference on Hypertext and hypermedia, pp. 128 – 137.

Rutledge, L., Alberink, M., Brussee, R., Pokraev, S., van Dieten, W., Veenstra, M., (2003). *Hypermedia semantics: Finding the story: broader applicability of semantics and discourse for hypermedia generation*. Proceedings of the 14th ACM conference on Hypertext and hypermedia, pp. 67 – 76.

Stuckenschmidt, H., Giunchiglia, F., van Harmelen, F., (2005). *Query Processing in Ontology-Based Peer-to-Peer Systems. Ontologies for Agents: Theory and Experiences*, Whitestein Series in Software Agent Technologies.

Tudhope, D., Cunliffe, D., (1999). *Semantically indexed hypermedia: linking information disciplines*. ACM Computing Surveys (CSUR), Volume 31, Issue 4es, article No. 4.

Vitali, F., Bieber, M., (1999). *Hypermedia on the Web: what will it take?*. ACM Computing Surveys (CSUR), Volume 31, Issue 4es, article No. 31.

Westermann, U., Klas, W., (2003). *An Analysis of XML Database Solutions for the Management of MPEG-7 Media Descriptions*. ACM Computing Surveys (CSUR), Vol. 35(4), pp. 331–373.

Wiil, U.K., Nürnberg, P.J., Leggett, J.J., (1999). *Hypermedia research directions: an infrastructure perspective*. ACM Computing Surveys (CSUR), Volume 31, Issue 4es, Article No. 2.

Baader, F., D. Calvanese, D. McGuiness, D. Nardi, P. Patel - Schneider, *Description Logic Handbook - Theory, Implementation and Applications*, Cambridge University Press.

W3C, (2004). Web Ontology Language (OWL), W3C, http://www.w3.org/2004/OWL/

W3C (2003). *Web Services Architecture*. W3C Working Draft 14 May 2003