Development of an Ontology-Based Portal for Digital Archive Services *

Ching-Long Yeh
Department of Computer Science and Engineering
Tatung University
40 Chungshan N. Rd. 3rd Sec.
Taipei, 104, Taiwan
chingyeh@cse.ttu.edu.tw

Abstract

In this paper we report employing ontology technology to develop a web portal for the digital archive service. The portal is divided into information provision and information access parts Each part works with an ontology, the knowledge structures commonly used in the digital library environment. The portal plays the role of an information aggregator that collects and unifies the information from different digital libraries and provides users with seamless service interface to access the content in an integrated way.

1 Introduction

Digital libraries manage various kinds of digital contents and provide services for users to navigate, query, use, produce and disseminate the digital resources. In the information environment of digital libraries, it is inconvenient for users to access various kinds of services provided by digital libraries. Therefore it necessary to develop a portal as the integration layer between the content providers, i.e., digital libraries, and users. The portal provides an seamless interface for user to access the services provides by digital libraries. From the technical architecture perspective, the portal plays the role of information aggregator between content providers and users. The

*This research was supported by the Taiwan National Science Council under Contract No. 91-2422-H-036-322

portal collects the semantic contents of digital resources in standards metadata structures, such as Dublin Core [16], from the content providers. To provide higher-level and effective services for users, for example, conceptual search, semantic navigation, personalization, and social interactions, it is necessary to build semantic integration functionality in the portal. The ontology technology is the basis for building such semantic integration functionality. In this paper, we aim at employing the ontology technology to develop a semantic web portal for to integrate the digital libraries funded by the National Digital Archive Program (NDAP)[1].

The backend of the semantic web portal consists of a ontology-based knowledge warehouse and in inference engine, while the front end consists of an interface for content provision and another for user to access the services supported by the portal. In this project, we use DAML+OIL as the description language of ontology. The inference engine is implemented using logic programming inference engine, Prolog [6] and expert system shell, Flex [23]. The content provided by the digital library can be unstructured documents (HTML) to be annotated in RDF [17], structured information, for example, relational database, and tables in HTML documents, to be wrapped into RDF, and other facts entered using ontology editor. All of these are stored in the knowledge warehouse in RDF format. Each of the services in the front end, including conceptual query, semantic navigation, question and answer, personalization, social interactions, is implemented as a set of inference rules represented in Prolog and Flex rules. The service interfaces are then running based on the interpretation of rules by the inference engine. The interoperability between the content providers and the portal is achieved by using standard XML protocol interfaces, SOAP [8] , connecting both ends. The received metadata is stored in the knowledge warehouse using RDF format.

In this paper, we consult the experience of ontology-based portal [11] and the functionality of digital libraries [20, 12] to build the architecture as described in Section 2. In Section 3 we describe the ontology construction and the inference mechanism. Then we describe the status of implementation in Section 4 .

2 Technical Architecture

An ontology represents the common knowledge and interest sharing sharing within a community [19]. As the web evolves into separating semantic content from the presentation, ontology becomes a core component in the development of semantic web portal [11, 9]. Ontology is used as the basis to build intelligent services for user to access the content stored in the portal and the data model to unify diverse information provided by various sources[10, 11]. A typical ontology-based portal system, for example, KA2 [11], consists of a knowledge warehouse and an inference engine, and two function components dealing with information provision and consumption, respectively. An ontology component supports all of the above components with knowledge schema to develop their functions.

In Semantic Web, XML-based ontology language, such as DAML+OIL[14], is used as the description langauge of ontology. The ontology languages are extensions of RDFS [7]; the resulting instances are RDF documents[17]. The information provision component supports different modes of information collection: annotations of unstructured documents, wrapping of semi-structured documents and entering new facts. In each mode, appropriate tools are used to create metadata in RDF format according to the schema of the ontology. The other function component provides various services, for example, conceptual search and ontology-based directory services [11], for users to consume the content collected by the portal. Each service program is developed by consulting the knowledge base consisting of the ontology and the knowledge warehouse and is executed by the inference engine.

The end-users of digital libraries, according to their roles, can be viewed as content providers and content consumers including researchers, educators and learners [12]. In digital library environment [20, 12], the tasks of information provision include (1) annotating the unstructured documents in the NDAP-funded digital libraries; (2) importing metadata records of contents and services from the NDAP-funded digital libraries; and (3) collecting profile information of content consumers and information generated when running social interaction functions as described later. In the ontology-based web portal, the first task can be carried out by employing existing annotation tools summarized in [3]. The second task can be done by using wrapper programs to convert semi-structured documents into RDF ones. The rest tasks are done when executing their respective service functions.

The services in digital library environment can be characterized as supporting discovery, use, tailoring and social interactions [20, 12]. The discovery service is concerned with finding the contents and services of user's needs and interests. The ontology-based web portal provides ontology-based query and navigation services to facilitate users achieving their goals. The query service is executing by invoking the inference engine and consulting the knowledge base. Users can specify their requirements by using conceptual constraints or forms provided by the system. The navigation service

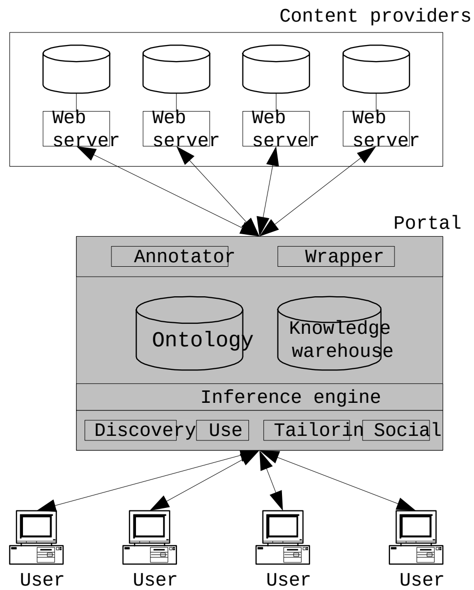

Figure 1: A high-level view of the components of the ontology-based portal.

provides users with directories dynamically computed by consulting the ontology schema and user's profiles. In this paper, we only consider simple use of discovered resources simply by browsing.

The tailoring services are the adaptations of services for personal or group purposes. They are made by consulting the profile information of individuals or group of users. The services of social interactions are of particular interests in education [12]. They are achieved by supporting, for example, functions of electronic discussion groups, bulletin boards for posing and answering questions, collaboration of content creation, and message filtering for target audience. By using the knowledge of ontology, both of the tailoring and social interactions service can be made more effective and efficient.

According to the above analysis, the high-level view of the architecture of the ontology-based portal can be summarized as shown in Figure 1. The central part of the portal is a a knowledge base consisting of ontology and knowledge warehouse. The information provision component consists programs for annotation and metadata wrapping. In the front end, the discovery, use, tailoring and social interaction functions are implemented based on an inference engine.

3 Ontology Construction and Inference Engine

As shown in Figure 1, the semantic content and the control are separated. The ontology and the knowledge warehouse are the schema and facts of the semantic content. The inference engine controls the interpretation of rules specified in the service programs. In this section we describe the ontology component and the inference engine.

3.1 Ontology construction

The end-users of digital libraries, according to their roles, can be viewed as content providers and content consumers including researchers, educators and learners [12]. The end-users are thus modelled as two conceptual hierarchies, organization and person. In this paper, we adapt for our purpose from the ontology developed in the KA2 project [5, 22]. KA2 is a ontology-based portal for the community of knowledge acquisition. In addition to the above two concepts, the KA2 ontology has other conceptual models for describing publications, projects, events, research topics and products in the field of knowledge acquisition. These are consulted to construct other conceptual models of the ontology component. For example, the publications model provide useful idea to classify the article types of web contents; the events model are used to categorize the event types happened in the social interaction function.

To integrate the ontology with metadata from the content providers, we add metadata schema generally used in the field of digital libraries, Dublin Core [5], using the syntax of the ontology language. The conceptual models in the ontology are used to create data instances that describe objects using attribute-value pairs. Usually categorization of subjects are helpful for users to find out what they want in a precise way. We employ the classification system from the Open Directory Project [2], a publicly comprehensive human-edited directory of the Web. The classification are used to enrich the semantic contents of entities stored in the portal.

The above conceptual models are concerned with describing the facts about objects. In this paper we are also interested in the definition of services in the ontology component. At present we are employing the technique of content planning commonly used in the field of natural language generation [18] to organize the selected content according to user's model and contextual information. We design a number of content planning rules as the basis to support personalized service of content selection and organization. Furthermore, the same idea can be used to develop content organization services for education purpose. The content planning rules are represented using the service ontology language DAML-S [4].

3.2 Inference engine

The inference capability, based on the ontology schema, derives the facts implicit in the knowledge warehouse. At present, the inference part of ontology definition is not yet finished. Furthermore, no inference engine for the DAML ontology language is available for the moment. We therefore choose Flex [23], a knowledge-based system toolkit to implement the inference mechanism of the portal.

Flex is based on the logic programming system, Prolog, and supports frame representation, which is similar to the inference engine used in other ontology-based portal [11, 13]. Furthermore, it supports forward and backward chaining inferences, which enables us to develop programs at conceptual level. Using Flex, we can access persistent databases using standard database connection interfaces. In addition to local GUI, it has utilities to make programs interact with outside world, such as the connection with web server, TCP/IP and agent libraries.

To make use of Flex, we first convert the ontology definitions into the frame representation of Flex. Each service program is implemented as a combination of backward and forward chaining rules as appropriate. For example, the concept-constraint search can be modelled by using a number of goal-driven backward chaining rules [15]. Personalization function is a kind of configuration problem, which is suitable to be modelled as a set of forward chaining rules.

4 Implementation

The first year of our project of using the ontology technology in digital library application, funded by the NDAP, focused on the technical design and implementation of the prototype of a semantic web portal, in order to obtain the necessary technical basis for the development of the semantic-integrated portal. We have designed and implemented an ontology-based portal for the library service [24] and an RDF triple store with conceptual search service [25]. In the second year of the project, from Mar of 2003, we focus on the development of the ontology-based portal using knowledge engineering approach [21, 11].

We start from the identification of technical architecture and analysis of requirements. Then we design the web site and ontology, including the static and dynamic parts, according to the result of requirement analysis. We then implement a portal according to the preceding conceptual design and test its functionality. The web site and ontology are modified during this step until they reach a stable state. Then we move the pilot implementation into practical operation and do necessary management and maintenance.

At present, we have gathered sufficient knowledge from the first step to carry out the construction of ontology in the second step. We are implementing and testing the conceptual search program using the ontology and self-created RDF data. We are further implementing personalization and social interaction functions using Flex.

The information provision part has different types of data creation. Within the portal, we use ontology authoring tools, such as Protege-2000 [26], to create facts in RDF. As for the unstructured documents, we use annotation tools listed in [3] to annotate the content. For the problem of interoperability between the content providers and the portal, we are investigating whether using existing standard protocol, like OAI-PMH [27], or designing one using standard XML protocol [8] carrying metadata in XML.

5 Conclusion

In this paper we separate the semantic content the function control components to build a portal for digital library environment. The semantic contents can be made in different ways according to the schema of ontology and it results in the same format. With the separation, the portal can be divided into two parts, information provision and information access, working independently to each other. It also facilitates the development of services at the conceptual level, for example, the conceptual search and semantic navigation of the discovery activity. Furthermore, we are able to develop advanced services based on such knowledge base.

References

[1] National Digital Archives Program, http://www.ndap.org.tw/.

[2] Open Directory Project. http://dmoz.org/.

[3] SW Annotation and Authoring. http://annotation.semanticweb.org/.

[4] Anupriya Ankolenkar, et al. DAML-S: web service description for the Semantic Web. In *The Semantic Web - ISWC 2002*. Springer. 2002.

[5] R. Benjamins, D. Fensel, S. Decker, and Gomez-Perez. KA2: building ontologies for the Internet: a mid term report. *International Journal of Human Computer Studies*. pp. 687-712. 1999.

[6] I. Bratko. *Prolog Programming for Artificial Intelligence, 3rd ed.*. Addison-Wesley. 2001.

[7] D. Brickley and R.V. Guha. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Working Draft 12 November 2002. W3C.

[8] D. Box et al. Simple Object Access Protocol (SOAP) 1.1, W3C Note 08 May 2000. W3C

[9] R. Scott Cost et al. ITtalks: a case study in the Semantic Web and DAML+OIL. *IEEE Intelligent Systems Special Issue*. 2002.

[10] S.Staab et al. AI for the web - ontology-based community web portals. *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on on Innovative Applications of Artificial Intelligence*, Austin, Texas, USA. 2000.

[11] S.Staab et al. Semantic community web portals. *WWW9*. Amsterdam. 2000.

[12] D. Fulker and G. Janee. Components of an NSDL architecture: technical scope and functional model. *JCDL '02*. ACM. 2002.

[13] A. Gupta, B. Ludascher and R. W. Moore. Ontology services for curriculum development in NSDL. *JCDL '02*. ACM. 2002.

[14] I. Horrocks, F. van Harmelen and P. Patel-Schneider. DAML+OIL. DAML Program. 2001.

[15] Amzi! inc. *Building Expert Systems in Prolog*. Springer-Verlag. 1989.

[16] S. Kokkelink and R. Schwanzl. Expressing Qualified Dublin Core in RDF / XML. Dublin Core Metadata Initiative. 2002. http://dublincore.org/documents/dcq-rdf-xml/.

[17] O. Lassila and R. R. Swick. Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation 22 February 1999.

[18] K. R. McKeown. *Text Generation*. Cambridge University Press. 1985.

[19] N. F. Noy and D. L. McGuinness. Ontology development 101: a guide to creating your first ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880. 2001.

[20] A. Powell and L. Lyon. The DNER technical architecture: scoping the information environment. 2001. http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/dner-arch.html.

[21] R. Studer, V. R. Benjamins and D. Fensel. Knowledge engineering: principles and methods. *Data Knowledge Engineering*. 1998.

[22] Y. Sure. KA2 - Knowledge Acquisition Community Ontology. http://ontobroker.semanticweb.org/ontos/ka2.html.

[23] P. Vasey. *flex Expery System Toolkit, version 1.2*. Logic Programming Associates Ltd. London, England. 1989.

[24] C. L. Yeh and C. G. Chen. Design and implementation of an ontology-based web portal. *Proceedings of the First Workshop of Digital Archive Technology*. Taipei, Taiwan. 2002.

[25] C. L. Yeh and R. F. Lin. Design and implementation of an RDF triple store. *Proceedings of the First Workshop of Digital Archive Technology*. Taipei, Taiwan. 2002.

[26] The Protege Project. http://protege.stanford.edu/.

[27] Carl Lagoze, et al. The open archives initiative protocol for metadata harvesting. OAI. Jun. 2002.