TOWARDS THE SCHEMA HETEROGENEITY IN DISTRIBUTED DIGITAL LIBRARIES

Hao Ding, Ingeborg T. Sølvberg
Dept. of Computer and Information Science, Norwegian Univ. of Science and Technology

Keywords: Digital Library, Schema, Metadata, Agent, Ontology

Abstract: In this paper, we discussed the problems brought by the schema heterogeneity in DLs, especially those problems found in the application of the OAI-PMH protocol. This paper studies the problem from two perspectives, namely the schema and the architecture respectively. A preliminary architecture is provided that integrates the ontology, agent, P2P together to support the schema mapping. And the semantic negotiation strategy between the heterogenous agents has also been described.

1 INTRODUCTION

With the explosive research in the Semantic Web, many people believe that the Semantic Web may first emerge in controlled communities like DLs because of the reliability of metadata that can be guaranteed. Meanwhile, because DLs could be accessed over the Internet inexpensively and conveniently, the constructions of DLs increase sharply and a number of topics are covered, such as science, history, culture, etc.. Moreover, more and more libraries use Web resources to populate their collections. It thus results in that different DL schema/metadata formats range over not only in the cooperative DLs but also the open-access web-based collections, which definitely increases the difficulty in finding the appropriate information on a specific topic or requirement.

In the past decade, there are many approaches to weave distributed DLs together (for Recall purpose) and alleviate the problem brought by schema variety (for Precision purpose). From the schema perspective, in order to facilitate the federation of distributed DLs or content providers on the Web, it is necessary to have a protocol that can 'harvest' the metadata in different collections. In DL community, two well-known protocols are Z39.50 (Z39.50 protocol) and Open Archive Protocol for Metadata Harvesting (Carl, OAI 2002). The former addresses a number of issues in a more complete manner but it is expensive to adopt. Generally speaking, no matter how great the functionality is, an approach with a high cost of adoption will not be widely used. Z39.50 has rich mechanisms, but it ends with limited distribution, which is contrast to the rapid and broad acceptance of basic web components such as HTTP and HTML (Carl, 2002). OAI-PMH thus aims to establish a low-entry and well-defined interoperability framework applicable across domains (Carl, 2001). It provides an application-independent interoperability framework based on metadata harvesting. Two roles are involved in OAI-PMH – Data Provider and Service Provider. The requirement for metadata (schema) interoperability is addressed by requiring all OAI Data Providers supply a common metadata set – (unqualified) Dublin Core (DCMES, 2003). However, in the current approaches in the metadata harvesting, some problems are brought out in terms of metadata incorrectness (e.g. XML encoding or syntax errors), poor quality of metadata, and metadata inconsistency (MARTIN, 2003). The flexibility in the usage of unqualified DC elements results in that some elements, e.g., 'type', 'format', 'language', etc., which may not share controlled vocabulary that can improve the consistency and then the quality of service (Hyunki, 2003). Furthermore, the simplicity of DC somehow loses the Precision in searching because of its limited description capability. Anyway, the use of Qualified Dublin Core (QDC, 2001) would solve some of these problems, but it will be also expensive to create and deploy as that in Z39.50. From the DL infrastructure perspective, there have been many federated DLs that are implemented in a centralized architecture, which requires a supporting organization to maintain them.

Ding H. and T. Sølvberg I. (2004).
TOWARDS THE SCHEMA HETEROGENEITY IN DISTRIBUTED DIGITAL LIBRARIES.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 307-312

These approaches work well within a controllable organization. For example, the BIBSYS library system (BIBSYS) federates 92 sub libraries that are distributed over the whole Norway in different colleges and universities. Although the sub libraries are geographically distributed, BIBSYS mandates all of them to adopt the BIBSYS-MARC (BIBSYS-MARC, 2001) metadata format. The National library of Norway administrates the centralized library and each sub library is allowed to have additional metadata standards for her own specific usages.

As we argued above, we believe it is almost impossible and impractical for us to create global-applied and unique identifiers (names) for all kinds of objects that we intend to search, browse, or exchange. We also believe that the future DLs will consist of many small or medium sized libraries that can provide specific services for users. Additionally, the users should be able to access not only the cooperative (federated) DLs but also the non-cooperating DLs at the same time.

In this paper, we propose to integrate DL systems in a new manner that combining the semantic negotiation, agent and Peer-to-Peer (P2P) technologies together. Our goal is to let the agent component embedded in different library communicate semantically. The mutually comprehensible agents will help to improve the data quality when harvesting in between.

The following of the paper is organized as follows: Section 2 introduces the related works in schema interoperability; Section 3 provides a multi-agent based P2P architecture for distributed DLs in which heterogeneous agents can communicate for ontology-based negotiating for understanding the meaning of different schema if there are. Discussion and Conclusion come in the final section.

2 RELATED WORK

Mappings between heterogeneous schemas have been studied for quite a while.

A framework for dealing with heterogeneous OSM schemas is presented in (Biskup, 2003). OSM models contain objects, their relationships and a predicate calculus for expressing constraints. The global schema is defined ontologically and independent from the source schemas. Interaction with an administrator is assumed (however not required) for setting up deterministic mappings between objects (and relations, respectively).

TSIMMIS (Chawathe, 1994) is one of the early systems integrating heterogeneous digital libraries. Schema mappings are defined in a textual format with actions which are executed when a corresponding template matches a query.

With the growing popularity of XML, mappings between different DTDs are also investigated. Due to the deterministic nature of XML, uncertainty is not supported by any of these approaches. A tree-grammar-based approach for inducing integrated views (XML-QL templates which can be used for stating user queries) for XML data with heterogeneous DTDs is presented in (Jeong, 1995). Type trees derived from the source DTDs are converted into a tree automaton. States belonging to similar types are merged to obtain a minimized integrated view.

MIND (Henrik, 2003) uses probabilistic logics for uncertain schema mapping. They mapped DAML+OIL into the probabilistic Datalog (Norbert, 2000) and use XSLT for actually transforming queries and documents.

National Science Digital Library (NSDL) adopts eight native metadata standards. The collections selected for inclusion in the NSDL have metadata conforming to the common or well-established standards, if they have metadata at all. If they have, the systems will automatically crosswalk native metadata to qualified Dublin Core (QDC, 2001), which will provide a lingua franca for interoperability. If not, the systems will processes content and generate metadata automatically (Carl, 2002).

3 SEMANTIC NEGOTIATION IN AGENT P2P-BASED DISTRIBUTED DLS

3.1 Architecture

Basing on the aforementioned arguments, we propose to adopt an agent P2P-based platform where ‘harvesting agents’ can harvest the metadata from other libraries in a semantic negotiation mechanism.

Agents are autonomous program units capable of working towards a set of goals. In multi-agent system, cooperating agents need a shared set of conventions (Wooldridge, 1995). The legacy approach is to agree upon a set of conventions,

particularly, a set of domain ontology beforehand, and then embed them into the agent communication protocols. In constructing agent-based distributed DLs, several open problems are still inherited as follows.

It is hard to have a world-wide consensus ontology base as mentioned above and hence it is groundless to have an associated language for every possible domain of multi-agent application.

Agent P2P-based DLs systems are open system because they consist not only the cooperative DLs but also the non-cooperating DLs. This means that the conventions can not be defined once and for all but are expected to expand as new needs arise.
Agent P2P-based DLs are typically distributed systems. There is no central control server.

So, there should be a shared lexicon for the involved agents to communicate a description. We believe that a co-evolutionary coupling on ontology and agent communication language will help improve the coordination in distributed DLs.

Figure 1 illustrates a general sketch of the architecture we propose.

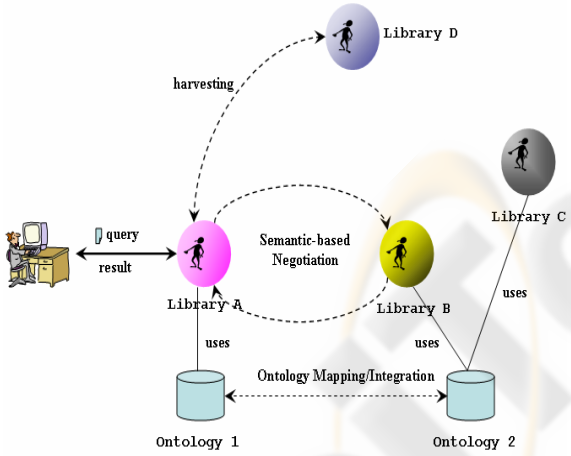

Figure 1: Semantics-based Interoperability in Distributed DLs

The involved agents are autonomous and they can be cooperative or not (e.g., Library B and Library D), which is well-suited for the real-world situations. According to the figure, Library A and Library D share a common metadata format, say DC (DCMES, 2003), so A can directly harvest the metadata records from D. However, B does not support DC format, but the Encoded Archival Description (EAD, 2002). It thus needs the schema mapping from EAD to DC if A wants also to harvest metadata records from B. So, the agent in A can be activated to negotiate with agent in A for the schema mapping in between (details about the negotiation are described in next sub-section).

In the architecture, the semantics-based negotiation mechanism happens between two heterogeneous agents that embedded in different library system. We have not chosen the *pure* P2P infrastructure because the current searching methods, such as the JXTA search protocol, assume that all providers are cooperative, thus they need to thus provide complete, reliable resource descriptions. But it is impracticable in some federated DLs environment that many libraries consider their rich metadata to be an important asset and only permit the 'privilege' users to access their collections (Carl, 2002). Thus, we propose to import the agent technology because it can support the communications between two libraries without reference to that they are cooperative or not. Furthermore, the agent-based communication mechanism and technology is fairly mature and is especially suitable for the explanations on a specific schema (negotiation). The major overhead may come from negotiation.

On the other side, dissimilar with the classical adoption of multi-agents system in DLs, e.g., the UMDL agent at University Michigan (William, 1995), which has a mediator for facilitating communication between agents, we plan to integrate the mediating functionality into an agent's own capabilities. Such that it will help keep track of an agent's neighbourhood and cache locations of other agents. In this way an agent P2P network is formed and a central bottleneck of the system is alleviated.

The major characteristics of the proposed approach are:

- No central control server. The agents have to coordinate by themselves in a self-adaptive fashion.
- The ontology remains adaptive. New coming DL system which contains different metadata or no metadata at all may require it to induce the meaning of terms in a specific schema.
- Library systems can join and leave freely as that in the P2P network.

Currently, in DL community, there has not been much done in bringing together P2P networks and agents for semantics-based interoperability. Thus, putting together P2P, agent and semantics is an unexploited research topic. And we believe it is a worthwhile research to go further.

3.2 The Role of Ontologies in DLs

Before we describe the semantic negotiation strategy between two heterogeneous agents, it is necessary for us to re-visit the role of ontologies in DLs.

According to aforementioned discussion, we believe that in the development of future digital libraries, the deployment of careful generated ontologies or thesauri will offer higher reliability and quality for the DL services. Furthermore, based on the adoption of ontologies, it will also help make mapping among related schema or integrate various schema into a repository to support the content-based retrieval. In fact, DL researchers have implicitly applied the idea of ontologies in DLs, for example, the process of classification on digital records. But there is still a long way to go to realize the ontology-based harvesting, searching and browsing, etc in DLs.

As concerning Ontology itself alone, James Hendler states that the Semantic Web will contain a great number of small possibly mutually inconsistent ontological components that consist largely of pointers to each other instead of few large and consistent ontologies (James, 2001). Currently, the most promising approach for the comparably 'large' standard ontologies is the effort to clean-up, refine, validate and merge the existing resources, e.g. WordNet (http://www.cogsci.princeton.edu/wn), HowNet(http://www.keenage.com/zhiwang/ezhiwang.html),CoreLex(http://www.cs.brandies.edu/~paulb/CoreLex/overview.html), the publicly accessible part of Cyc (http://www.cyc.com/), etc., for the practical application, like ontology/metadata mapping in DLs. There is available program for helping validating designed ontologies (Nicola, 2002).

According to the well-know '5 papers on Wordnet' (Miller, 1990), the essential part of concepts are:

- Synonymy(similar concept): <*creator*, *maker*>
- Hyponymy(narrower-broader/ISA): <*designer* is a *creator*>, <*creator* is *person*>
- Meronymy(part-of/HASA): <*creator* has *personality*>
- Derivationally related terms/concepts: <*creator* RELATEDTO *create*(verb)>

A number of papers in the DL and IR communities have described the considerable improvement obtained by adopting synonymy and hyponymy. For example, in the application of query expansion. This paper is yet not another endeavour to propose new approaches for performance improvement. Rather, it concentrates on how we can incorporate them into distributed DLs and alleviate the problems brought by schema heterogeneity. The following section will concentrate on the semantic negotiation strategy.

3.3 Semantic Negotiation Strategy

Semantic Negotiation is a general purpose mechanism that can be used in many different contexts for exchanging schemas information and description. In the procedure of negotiation, the agent on the Service Provider (SP, the same meaning as that in OAI-PMH) is expected to interpret/understand the schema formats on the heterogeneous Data Provider (DP, also from OAI-PMH). The process is as follows:

1). When $agent_{sp(i)}$ asks $agent_{dp(j)}$ for the schema format information, $agent_{dp(j)}$ sends $agent_{sp(i)}$ a list of terms, using the description based on a lexical base, for example, Wordnet. And the latter should also support such a kind of lexical base. The reason for doing so is that it is almost impossible for two agents to mutually comprehend and exchange data without *any* shared vocabulary or thesauri.

2). if $agent_{sp(i)}$ does not understand the description, it responds with an error code indicating that the description can not be understood. In this case, it lists the particular terms not understood. Based on this feedback, $agent_{dp(j)}$ can try to provide a description that the server is more likely to understand.

3). if the $agent_{sp(i)}$ partially understands the description, that is, there are some mismatching terms, it returns an error code saying so. It can optionally also tell the $agent_{dp(j)}$ which part of the description was not satisfied by any of the terms.

4). if the $agent_{sp(i)}$ understands the description, it returns the confirmation to $agent_{dp(j)}$. In the case where the answer is a list of resources, the answer may include additional data about each resource, which the $agent_{sp(i)}$ may cache, in anticipation of future queries about these resources.

The sequence diagram is illustrated in Figure 2.

Let us take a simple example, if $agent_{sp(1)}$ on Library A queries $agent_{dp(2)}$ on Library B for the metadata schema, $agent_{dp(2)}$ then responds his metadata format in which there is one term – 'author' that $agent_{sp(1)}$ does not understand. Thus $agent_{sp(1)}$ sends a feedback to $agent_{dp(2)}$, claiming that unknown term. Based on the feedback, $agent_{dp(2)}$ provides a

description (see below) that is generated from the prerequisite query on Wordnet.

From the fragment of description, $agent_{sp(1)}$ finds that 'creator' is just one of the elements in DC that Library A supports. Thus he responds which he understands the term successfully and cache the mapping for the application later on between Library A and B. Hereby, the mapping should be focused on specific relationships among specific libraries.

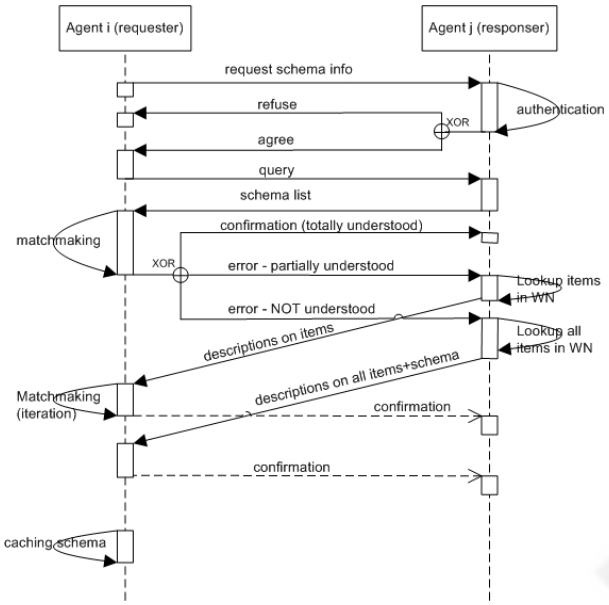

Figure 2: The Sequence Diagram for the Semantic Negotiation

4 DISCUSSION AND CONCLUSION

Even if the WWW contains more information than any single traditional library, it can not substitute the traditional library because it lacks these services (particularly organization and sophisticated search support) (William, DLib1995). No one is disassembling their libraries because of WWW yet. On the other side, because the webpage/media editing tools become better and access to networks becomes easier and cheaper, there will be millions of content suppliers. The sharply increased public DLs available on the Web are just a good proof for it. However, people also find the difficulties in finding the appropriate information because of the voluminous collections and hence the problems in locating the proper repositories. And the key issue in the problem comes from the schema heterogeneity.

Many approaches in DL community have been carried out to investigate the problem. There are also many practical DL systems appear. Some of the solutions create an integrated and global schema set that may include exactly one (e.g. MARC21) or several metadata formats (e.g. Dublin Core, Encoded Archival Description, etc.). The individual library thus maps its local metadata format to the global one. If the global metadata set contains just one format, such as the BIBSYS-MARC in BIBSYS, all of the cooperative DLs should abide by the BIBSYS-MARC format respectively although they can extend some items locally. As to a metadata set that may hold several schema formats, like NSDL, which adopts eight metadata standards. The collections selected for inclusion in the NSDL have metadata conforming to the common or well-established standards.

Such approaches will be unavoidably faced with the problem in scalability, specifically, in the situations when libraries join and leave. These cooperative libraries will take pains in adjusting the global view of the metadata set or reformatting the local metadata standards. The UMDL adopts the agent technology in the DL development, bearing the intention to create a flexible software architecture that can federate as many content suppliers, information-organizational schemas, and service providers as possible, and yet scale to the extremely large size needed to support the DLs in the future (William, DLib1995).

However, UMDL has not utilized the emerging Semantic Web technology, which is widely accepted that it can offer some semantic groundings. In the distributed DLs, the profitable area is to embed the semantic negotiation strategies into the agent communication policies.

In this paper, we firstly discussed the problems brought by the schema heterogeneity in DLs. Many problems in the implementation of OAI-PMH protocol have also reported their findings in this issue. We believe that the future DLs could not be accomplished without an adoption of a careful design of ontologies. The essential types of ontologies that could improve schema mapping were also presented. In order to have a platform for the semantic-based agent communication in distributed DLs environment, we proposed a preliminary architecture that integrates the ontology, agent, P2P together to support the schema mapping. The semantic negotiation strategy has also been provided. We are aware that there are many open questions, so this work should be considered a

stepping stone. And, it is a worthwhile research to go further.

REFERENCE

BIBSYS, the Norwegian library automation network, http://www.bibsys.no

BIBSYS-MARC : Bibliografisk format. BIBSYS, 2001. http://www.bibsys.no/handbok/marc/marc.pdf

Biskup,J. and Embley, D. W., Extracting information from heterogeneous information sources using ontologically specified target views. *Information Systems*, 28(3):169–212, 2003.

Carl Lagoze, Herbert Van de Sompel, 2002, The Open Archives Initiative Protocol for Metadata Harvesting, http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

Carl Lagoze, Herbert Van de Sompel: The open archives initiative: building a low-barrier interoperability framework. *JCDL 2001*: 54-62

Carl Lagoze, William Y. Arms, Stoney Gan, Diane Hillmann, Christopher Ingram, Dean B. Krafft, Richard J. Marisa, Jon Phipps, John Saylor, Carol Terrizzi, Walter Hoehn, David Millman, James Allan, Sergio Guzman-Lara, Tom Kalt: Core services in the architecture of the national science digital library (NSDL). *JCDL 2002*: 201-209.

Chawathe, S., Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J., and Widom, J.. The TSIMMIS project: Integration of heterogeneous information sources. *In 16th Meeting of the Information Processing Society of Japan*, pages 7–18. Tokyo, Japan, 1994.

Dublin Core Metadata Element Set (DCMES), 2003, Version 1.1: Reference Description, http://dublincore.org/documents/dces/

Dublin Core Qualifiers (QDC), 2001, http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/

Encoded Archival Description (EAD), Version 2002, http://lcweb.loc.gov/ead/

Henrik Nottelmann, Norbert Fuhr: Combining DAML+OIL, XSLT, and Probabilistic Logics for Uncertain Schema Mappings in MIND. *ECDL 2003*: 194-206, Trondheim, Norway.

Hyunki Kim, Chee-Yoong Choo, Su-Shing Chen: An Integrated Digital Library Server with QAI and Self-Organizing Capabilities. *ECDL 2003*: 164-175, Trondheim, Norway.

James Hendler, Agents and the Semantic Web. *IEEE Intelligent Systems*, 16(2):30-37, 2001.

Jeong, E. and Hsu, C.-N.. Induction of integrated view for XML data with heterogeneous DTDs. In Paques et al. [17], pages 151–158.

Martin Halbert, Joanne Kaczmarek, and Kat Hagedorn, Findings from the Mellon Metadata Harvesting Initiative. *ECDL2003*, pp.58-69, Trondheim, Norway.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Five Papers on WordNet. *Special Issue of International Journal of Lexicography*, 3 (4). (1990)

National Science Digital Library (NSDL), http://nsdl.org

Nicola Guarino, Christopher Welty, Evaluating Ontological Decisions with ONTOCLEAN, *Communication of the ACM*, Feb. 2002/ Vol.45. No.2.

Norbert Fuhr: Probabilistic datalog: Implementing logical information retrieval for advanced applications. *JASIS* 51(2): 95-110 (2000)

William P. Birmingham, Edmund H. Durfee, Tracy Mullen, Michael P. Wellman, The Distributed Agent Architecture of the University of Michigan Digital Library (Extended Abstract), *AAAI Spring Symposium on Information Gathering*, 1995.

William P. Birmingham, An Agent-Based Architecture for Digital Libraries, *DLib Magazine*, July 1995.

Wooldridge, M. and N.R. Jennings, Intelligent Agents: Theory and Practice. *Knowledge Engineering Review*, 1995, 10(2).

Z39.50 protocol, http://www.loc.gov/z3950/