Ontology and Metadata Based Semantic Search System

Hongshan Yang[1], Guoxiang Wang[2], Jiaxun Chen[1]
[1]College of Information Sciences and Technology, DongHua University
[2]Aviation University of Air Force
P. R. China
[1]No.1882 West YanAn Rd., Shanghai
P. R. China

Abstract: - Semantic search is valuable and has big developing foreground. Metadata has become one of the most important technologies supporting information searching in the last twenty years. Metadata provide the structured and standard information about its describing object including the content introduction, the background, the physical property and the using restriction etc. Ontology is a useful tool to endue the computer with understanding the semantic of data. In this paper, an ontology based metadata scheme and semantic search system are put forward. A prototype of the semantic search system with an ontology mapping server is realized.

Key-Words: - Ontology, Metadata, Semantic Search, Ontology Mapping, Dublin Core, Digital Library

1 Introduction

Information searching technology got a very high development in the last ten years. The searching engine is advanced very quickly. The searching range is enlarging and the searching precision is increasing every day. Metadata plays a very important role in information searching. The resource creator will give metadata to make his/her resource being found easily. The searching service provider marks the HTML pages with metadata to increase the searching speed and precision.

Metadata has great importance in digital library. Traditional library realized the book searching by book card and book category. Based on metadata digital library can provide more indexes for information searching, including title, keywords, creator, abstract, publishing date and so on. At present the digital library is mostly realized by condition matching search mechanism, which can only support exact search and exact matching and have no idea of semantic information searching. The metadata is only a structured and standard information without semantic or it can be say the computer/search engine can not understand the semantic of metadata.

Metadata, which origins from the book card, always plays an important role in information system. The digital information in the world is variety and large-amounted. The data is enlarged very quickly and the data is updated every moment, which bring a big challenge in using and management the information. Metadata just provides the function of description, management, catalogue, search and locating on a great deal of information. So it has become one of the core technologies of information processing, discovery and utilization.

Ontology is a tool of realizing semantic. It provides a common knowledge sharing space, which makes the different systems or organizations have consistent and unified understanding on information semantic. Using ontology to building the semantic oriented metadata can support the concept modeling, information searching and exchanging. Ontology provides the search engine with the functionality of a semantic match. It is different from traditional search engines that can only completed directly and exactly search.

In this paper, we present the ontology based metadata and construct a semantic information search system. In Section 2, we introduce related works about metadata and ontology. In Section 3, we present the metadata ontology model. In Section 4, we design the ontology based metadata. In Section 5, we describe the semantic searching system. In section 6, we introduce the prototype system development and the practice result in digital library.

2 Related Works

Along with the information increasing at very high speed, especially the information on the Internet is increasing in geometric series. Metadata is more and more important in information resource management, sharing, searching and applying. Also, in many condition, metadata itself has become an absolutely necessary part of information resource. Metadata standard and technology researching upsurge has been continued for more than 20 years. Now series of metadata standards have been developed, which

adapt in different areas and satisfy their requirements. The earliest metadata standard in digital library domain is MARC (MAchine-Readable Cataloging), which defines a data format that emerged from a Library of Congress-led initiative. It provides the mechanism by which computers exchange, use and interpret bibliographic information, and its data elements make up the foundation of most library catalogs used today [1]. MARC has play an important role in the digital library development, but it has obvious disadvantage that the recording format is too complexity and the workload is tremendous. Dublin Core (DC) is another widely used metadata standard, which defines 15 core elements and some qualifiers to describe common entity [2]. England e-Government Metadata Standard (eGMS) expands some elements and qualifiers based on DC to fulfill the requirements of government resource description [3]. Australian Government Information Locate Service (AGLS) is the standard of Australia in locating government information and services [4], which is also based on the DC. Chinese Digital Library Standards are series of standard based on DC, which using in describing degree thesis, periodical paper, audio resource, rubbings, and ancient books etc [5].

In addition, metadata standards are established in other domain to manage and share the resources. They includes: metadata standard describing geographic data [6], the standard of Categories for the Description of Works of Art (CDWA) [7], and the standard for describing archive and manuscript collection: Encoded Archival Description (EAD) [8].

Ontology is a philosophy concept originally, which is first used by computer scientist in Artificial Intelligence to represent the common concept. Ontology has been researched in widely areas and many scholars have given multiple definitions from different aspects. But the most popular and accepted is the “specification of a conceptualization” given by Gruber [9].In information domain ontology provides a common knowledge basis for all to send message, cooperate and sharing information.

Ontology was also being proposed to use in the Semantic Web [10] to make the computer understand the information on the Web, so that it can semantic search and integrate the information. In digital library, ontology has been researched and used: construct formal digital library ontology [11] and use ontology in taxonomy of metadata [12].

3 Metadata Ontology Model

Metadata is a kind of knowledge, which is a structured description of resource. So metadata standard can be represented by ontology. We chose DC as the instance to build resource description metadata ontology. The regular of transforming metadata standard to ontology is that defining the description object as Concept (Class), named Resource; defining the elements of metadata as the Properties of Resource; defining the refinement of the elements as the Subproperty and defining the encoding schema as the constraint of property. Fig.1 shows the metadata ontology model.

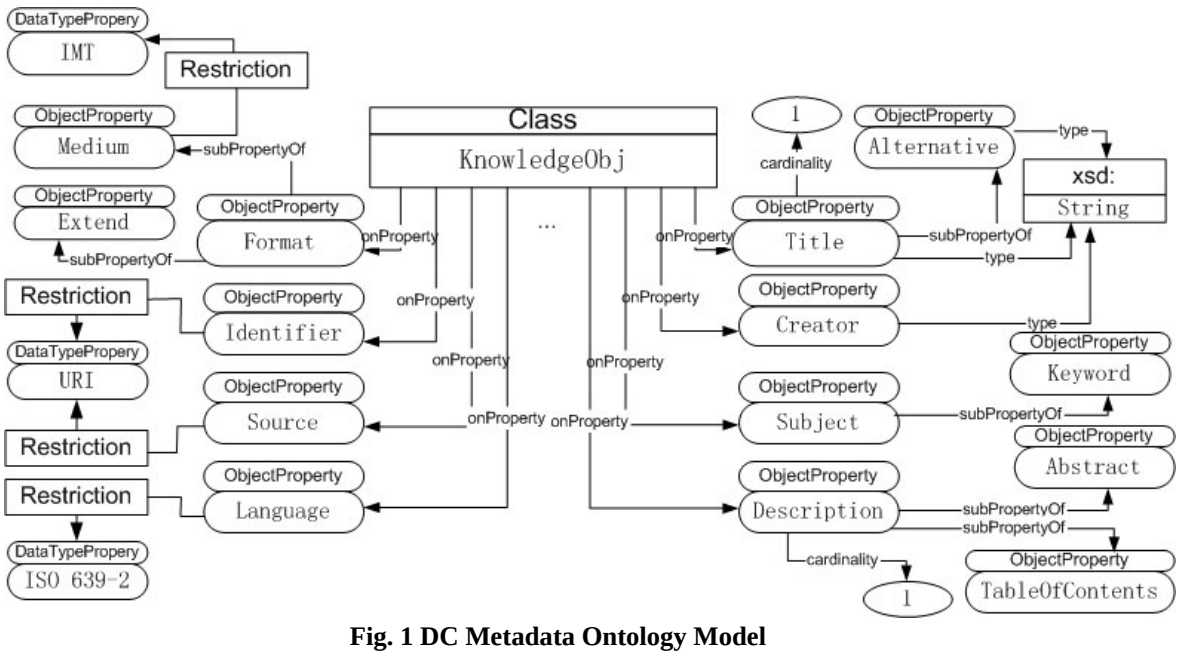

Fig. 1 DC Metadata Ontology Model

The DC metadata ontology only has one Class (Resource), which represents all kinds of objects being described by DC. The DC elements (Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Source, Relation, Rights, Identifier, Language, Coverage) are defined as ObjectProperty. The refinements and qualifiers of element, such as Alternative-refinement Title, tableOfContents and abstract-qualifier of Description etc are defined as Subproperty of the ObjectProperty. The encoding schemes of element and qualifiers, such as LCSH-encoding scheme of Subject, IMT-encoding scheme of Medium and BOX-encoding scheme of Coverage etc are defined as the Restriction of ObjectProperty or Subproperty.

In the Fig.1, we only draw 8 core element and some of their subproperties. Most components of the ontology are elided, because the figure can't hold them. But they are all detailed defined in the ontology RDFS [13]: all the metadata elements are defined into properties; all the refinements are defined into subproperties; and all encoding schema are defined into constraints. The datatype and cardinality are also defined.

As the DC metadata srandard is transformed into DC ontology, metadata records become instances of the metadata ontology. We develop the RDFS for DC metadata ontology. So that the metadata instance structure can be produced automatically, the syntax and restriction can be checked according to the RDFS by metadata editor tool. The editor will be introduced in detail in section 6.

4 Ontology Based Metadata

Ontology based metadata is building metadata upon ontologies, so that the systems and organizations can understand the semantic of metadata. It will avoid the heterogeneous and multi-sense of metadata. The ontology used in constructing metadata is often not only one, but a few of them merged or combined together. Fig. 2 show the relationship of multiple ontologies in ontology based metadata.

Fig. 2 Relationship of Ontologies Used in Metadata

4.1 Ontology Combining and Merging

DC as the common metadata standard, it has to be expanded, when describing special domain resources. The same is true that only one ontology can't cover the knowledge of all domains. So we often build the metadata based on multi-ontologies or merge a few ontologies into a large and width one. In our research, we chose the Cyc ontology as the upper ontology in building the metadata. The domain of Cyc ontology includes all of human consensus reality. It defines 47,000 concepts and 306,000 assertions about these concepts to interrelate, constrain and define them [14]. Besides the upper ontology domain ontologies are needed to represent the detailed semantic of the special domain. Cyc ontology covers all domains, so the domain ontology always have intersection or same definitions of class and proper, the over defined part is the interface of the ontologies merging or combination.

4.2 Ontology and Taxonomy

Ontologies can be used with the taxonomy and thesaurus in metadata. Whereas the taxonomy can be regard as a simple ontology, which only lists some of the entities without defining the relation and property. We have expand DC with a qualifier DC:subject:catogry, which give the rough classification of the resource. We mainly use Cyc as the encoding scheme of DC:subject:catogry and DC:subject:keyword. But other elements, such as DC:date and DC:coverage, reference the class Date and Position defined in Cys ontology. When recording the metadata, workers can select the category of resource and then select the keyword from this category. We have embedded the Cyc ontology into the metadata editor and provide the hierarchy of category and keyword.

The elements value such as DC:category, DC:keyword and DC:coverage etc comes from or relates to the ontology, when the ontology evolved as the time change. The metadata will automatically evolve as the meaning of category expanded and the keyword added etc. But we still need domain ontologies to express the special and detail character of the resource, which has not been defined in upper ontology as Cyc. The relation between upper ontology and the domain ontology can be coordinate or merge them together just as shown in Fig.2.

5 Semantic Searching System

Ontology-based metadata can support the semantic search of information resources. System can analyze,

formulate and optimize the implicit semantic of user's query condition combining with the ontologies. The semantic search can make the search widely or deeply by searching the base class, relevant class or searching the special subclass of the user's query class. In this paper, we propose two schemes to realize the semantic search. They are shown in Fig. 3 and Fig. 4.

5.1 Ontology Mapping Server Based Semantic Search

In the first scheme, we realize the semantic search by adding an ontology mapping server between user and the search engine. The user's query condition will be first process by the ontology server. It formulates the keywords with knowledge ontology and analyzes the query condition with user ontology to expose the implicit semantic. The query condition submitted to the search engine will become normative and include the implicit semantic. Then search engine searches the metadata database and feedback the relevant resources to the user. The system architecture and process is shown in Fig.3.

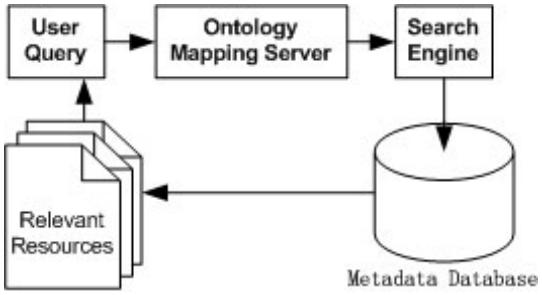

Fig.3 Ontology Server Based Semantic Search System Framework

The ontology mapping mechanism is the core of the system. It includes multi operations in mapping the user's input into concepts. For example, the user input "crocodile, living area". The ontology server will expand the crocodile into the concept of crocodile and the SubConcept of crocodile (alligator, caiman and gavial etc). Because the user give a width concept (crocodile), the system think the user want to know the living area of all kinds of crocodiles, so the system refine and expand the concept of crocodile to get more results.

5.2 Ontology Server and Agent Based Semantic Search

In the second scheme, we develop a user agent system between user and the search engine. The ontology server is still worked, which support the user agent to formulate the searching condition. The agent interacts with the ontology terminology server. It formulates the user's query according to the knowledge ontology and user ontology.

But the agent has another role that it can filter the searching result. After the search engine returns the searching result, the agent filters the data and discards the unfit records. Agent stores the user's information (profession, specialty, preference etc) and it saves and studies the user's feedback of the searching result, so it can evolve automatically. It will formulate the user's search condition properly and filter the search precisely. Agent can understand the user more and more along with the user use the system. The system and process is as shown in Fig. 4.

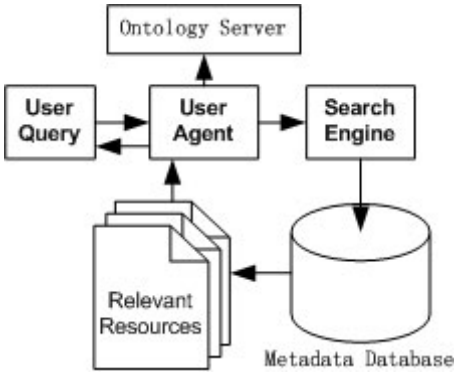

Fig. 4 Ontology and Agent Based Semantic Search System

The emphasis of ontology server based semantic searching is the ontology mapping and formulating method to the user's query. But, the performance of user agent based semantic searching lies on the veracity of agent formulating. Also, the agent evolution should be carefully considered to continuously increasing the performance of agent efficiency.

6 System Realization and Application

In order to verify the above design, we develop a prototype of metadata management and searching system for digital library. It providers ontology and metadata based information search service for the university teachers and students.

We built the prototype system by 4 steps. Firstly, we expanded DC and represented it into ontology (RDFS) by Protégé-2000. Secondly, we developed a multi-schema and multi-ontology metadata editor, which can produce different format metadata according to the inputted schema. The editor has integrated a lot of encoding scheme, such as LCSH,

IMT, RFC1766, URI, ISO3166 and W3CDTF etc. Some of these encoding schemes make user choose the proper value of elements, some of them prescribe the format of values. We used the editor to record the metadata of resources. Thirdly, we developed the relational not XML database to store the metadata, because the RDF model (Object-Attribute-Value triples) can map very directly to the relational database model and relational database is easy to deploy. The library in our university is use DB2 relational database, so it is easy to exchange data between the former system and the semantic searching system. Fourthly, we built an ontology server to formulating and optimizing the user's query. It can map the searching keyword into ontology concept and analyze the relation and the implicit semantic in the query with the user ontology. We integrated Cyc ontology and some domain ontology (now is Computer Science and Biology) into the ontology server. We set the principle of processing the user query as: if the searching keyword is a very widely concept, we will limit the searching areas that relate to the user speciality and research direction; if user input a very narrow keyword, system will expand the concept considered the user ontology, such expand is as searching the base Class, SubClass and similar Class that related to user speciality. Also, the searching result is ordered according to the matching degree, user interests and speciality.

We have applied the system in the library and make some statistics to verify its efficiency. We found that the Recall Ratio, Precision Ratio increased obviously. And it is recognized by users. Response Time has little change when the User Effort is light. But with heavy User Effort, the Response Time will prolong. In conclusion, the system has some advantages and can fulfill the requirement of teachers and students on scientific and research information. It gives us hope of building a full semantic and universal searching system for digital library. But it is only a prototype system, There is still a lot of work should be done. Its domain ontologies, which only include Computer Science and Biology, should be added. The ontology mapping precision and the system function need to be enhanced.

References:

[1] Network Development and MARC Standards Office, *Understanding MARC authority records*, Library of Congress, 2004, 2nd. ed. ISBN: 0-8444-1113-2.

[2] DCMI, *The Dublin Core Metadata Initiative*, http://dublincore.org/.

[3] Cabinet Office, *e-Government Metadata Standard Version 2*, http://www.govtalk.gov.uk/documents/eGMS version 3.doc.

[4] National Archives of Australia, *Australian Government Information Locate Service*, http://www.naa.gov.au/agls.

[5] Xiao Long, Zhao Liang, Feng Xiangyun et al., *Chinese Digital Library Standards*, http://cdls.nstl.gov.cn/cdls2/w3c/2003/SpcMetadata/ (in Chinese).

[6] FGDC, *Content Standard for Digital Geospatial Metadata*, http://www.fgdc.gov/metadata/metadata.html

[7] J. Paul Getty Trust, *Categories for the Description of Works of Art (CDWA)*, http://www.getty.edu/research/conducting_research/standards/cdwa/8_printing_options/definitions.pdf.

[8] Network Development and MARC Standards Office, Encoded Archival Description (EAD) Version 2002, http://www.loc.gov.library.unl.edu/ead/.

[9] T. R. Gruber, *Translation approach to portable ontologies, Knowledge Acquisition*, 5(2):199-220 1993.

[10] Berners-Lee, T., Hendler, J., and Lassila, O., *Semantic Web. Scientific American,* 284, 5(2001) 34–43, 2001.

[11] Marcos André Gonçalves, Layne T. Watson, and Edward A. Fox,, T*owards a Digital Library Theory: A Formal Digital Library Ontology*, ACM SIGIR Mathematical/Formal Methods in Information Retrieval, MF/IR 2004, Sheffield, UK.

[12] Liu Ying, Zhan Meng, *A Review of Ontology Research on Digital Library*, Library Journal, (in Chinese), 2005.

[13] Tom Barrett et al, *RDF Representation of Metadata for Semantic Integration of Corporate Information Resources*, Proc. WWW-2002, Hawaii.

[14] Ramachandran, Deepak, P. Reagan, K. Goolsbey, *First-Orderized ResearchCyc: Expressivity and Efficiency in a Common-Sense Ontology*, AAAI Workshop on Contexts and Ontologies: Theory, Practice and Applications. Pittsburgh, Pennsylvania, 2005.