Interoperability and Semantics for Heterogeneous Earthquake Science Data

Anne Yun-An Chen[1], Andrea Donnellan[2], Dennis McLeod[1], Geoffrey Fox[3], Jay Parker[2], John Rundle[5], Lisa Grant[4], Marlon Pierce[3], Miryha Gould[4], Seokkyung Chung[1], Shan Gao[1]

[1]University of Southern California
[2]JPL
[3] Indiana University, [4] University of California, Irvine,[5] University of California, Davis
[1]{yunanche, mcleod, seokkyuc, sgao}@usc.edu

Abstract

We propose to define, design, develop, deploy, and test a data semantics based system to provide interoperability for heterogeneous data in the earthquake science domain. We focus on the database management aspects of the work, including modeling the meaning of the data, providing for web service based access to heterogeneous data sources to scientists to be used in simulations and model development, data mining, etc.

Introduction

In earthquake science, data sources can be categorized into three different kinds: observations, simulations, and hypotheses. Since scientists have their own interpretations and analyses of the raw data and individual databases are distributed, a semantic metadata management system and wrappers for web services are required to support effective information retrieval and web-based search for data of interest to a specific scientist. We propose to construct such a semantic based system to provide interoperability for heterogeneous data, different applications and databases systems, and user-defined packages. Currently, we have created the middleware for two groups of experts in order for them to manage the raw data, retrieve the data, and get the wrapped data for further usage, i.e. in simulation programs. A fault database has been established to handle fault parameters. An initial domain ontology (description of key concepts and inter-relationships in the domain) is being developed by both the computer scientists and earthquake science experts. In what follows, we outline our general approach, and end with a summary of the current and future work.

Essential System Components

The illustration of the whole system is shown in Figure 1.

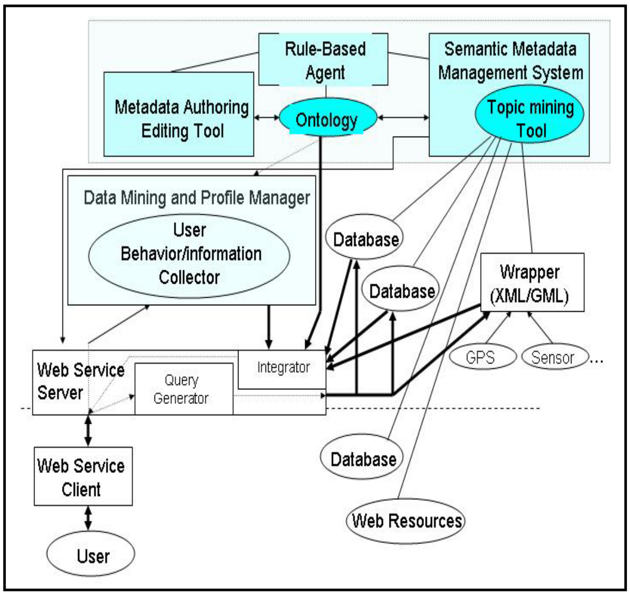

Figure 1 System Illustration

Semantic Metadata Management System

The semantic metadata management system we propose (see the upper portion of Figure 1) has five main functional capabilities:

1. Facilitate domain ontology creation and update,
2. Associate the ontology metadata with observational and hypothetical data,
3. Learn new concepts, relationships, and patterns among the metadata and data,

4. Support user (scientist) data and meta-data discovery/search and,
5. Provide the base for the semantic wrapping of information sources.

Rule-based reasoning agents and a simple metadata authoring/editing tool will accommodate data-mining, learning and updating capabilities for metadata management. We intend to ensure that the structure and format of metadata are compatible with RDF [1], DAML+OIL, and XML with limited "process" [2, 3]. This will make the ontology/metadata portable. All communication of data will utilize XML that is compatible with and to an extent based upon Geography Markup Language (GML) [8]. GML helps to describe the format and transmission of geographic information and ensures that both spatial and non-spatial data can be integrated.

Web Services

Web services use an XML-based protocol and schema of interface definitions to invoke the applications among servers and clients. The protocol provides the information required by the remote services. One of the common web services, SOAP, is the method message procedure and deployed as an application in a web server. Web services cannot be completed without the method interface in the Web Services Description Language (WSDL) [7]. The interface to services is implemented with a web-friendly programming language (such as Java or Python). Web services allow information exchange among different platforms and applications and make remote application invocation possible [9].

Topic Mining

Topic mining is to find (new) concepts and events in a collection or stream of data [6]. Topic mining is able to perform thematic and/or the pattern-oriented trend detection and tracking. Unlike traditional keyword-based search, topic mining provides information upon an event-based point of view and helps to adjust the various interpretations of data for geo-science. "Event" means a certain thing that happens at a certain point of time. For example, if a user searches for "earthquake in Southern California", a typical web search engine would provide the links of general descriptions or the research center information of earthquake in Southern California. However, an event-based search would return facts, i.e. "earthquake on San Andreas fault in May 2002". Furthermore, as shown in Figure 2, topic mining can hierarchically cluster collections of web pages describing earthquake occurrences. It also associates topical terms (from specific to abstract) with each cluster according to a topic.

Current Status

We have created an initial domain ontology and two different databases. One of the databases contains processed data of California faults and the sources are journal articles in the field of paleoseismology. The other contains data of California layers. We've also developed a simple XML-based distributed web service system for this pair of databases.

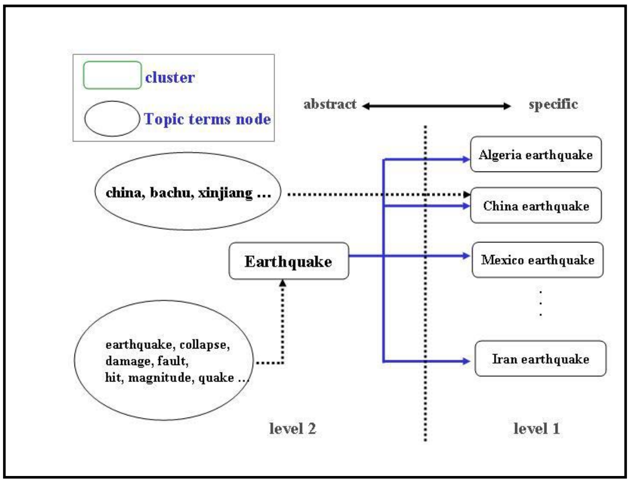

Figure 2 An Example of Topic Mining

Domain Ontology

Our basic ontology model defines a collection of concepts and interrelationships among those. We support the semantics of three key kinds of generic inter-relationships for which the system "knows" the meaning: Is-A, Part-Of, and Instance-Of [4]. These represent the object-based primitives in semantic data models [5], and form the starting point for our ontology model. We intend to explore expanding this set with other generic semantic primitives, as well as possibly some domain-specific ones.

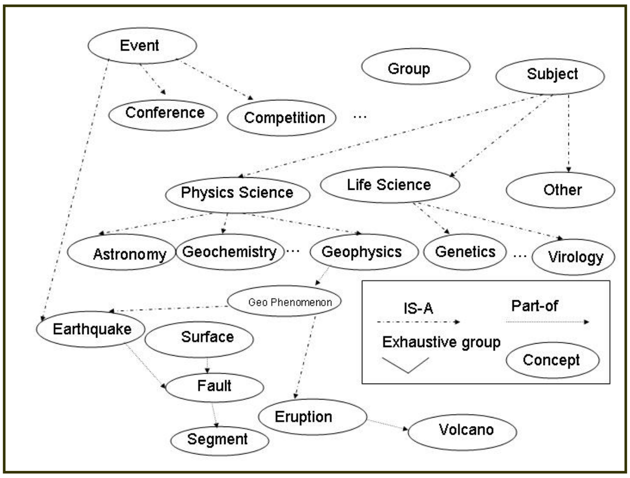

Figure 3 Example ontology of earthquake domain

Middleware and Web Services

At present, we employ a SOAP server on Tomcat. Several clients for one of our web services using the WSDL interface have been developed. These clients are used to generate SOAP requests. Our users now can use the client stub to request the information and extract the results for literature references, data integration, and even graphical simulations with virtual reality tools.

Conclusions and Future Work

The nature of data in the earthquake science domain is full of variety and the interpretations of data differ from resource to resource, and scientist to scientist [10]. Therefore, our system must provide integration portability to manage the interoperability for heterogeneous data. At the same time, our system has the most interaction in these specific areas: "metadata services", "federated database system", "data assimilation", "data mining", and "web services". Our semantic metadata management system includes rule-based reasoning agents and a simple metadata authoring/editing tool to modify the ontology; topic mining technique also plays an important role for dynamic ontology management.

Acknowledgement This work was supported by NASA's Computational Technologies Project. Portions of this work were carried out by the Jet Propulsion Laboratory, California Institute of Technology under contract with NASA.

References

1. S. Decker, D. Brickley, J. Saarela, and J. Angele, "A Query and Inference Service for RDF", *Proceedings of QL'98 - The Query Language Workshop*, December 1998.
2. D. Fensel, F. van Harmelen, I. Horrocks, D. McGuiness, and P. Patel-Schneider, "OIL: An Ontology Infrastructure for the Semantic Web", *IEEE Intelligent Systems*, 16(2): 38-45, March/April 2001.
3. J. Heflin and J. Hendler, "A Portrait of the Semantic Web in Action, Intelligent Systems", *IEEE Expert* , 16(2): 54 -59, March-April 2001.
4. L. Khan, D. McLeod, and E. Hovy, "Retrieval Effectiveness of an Ontology-Based Model for Information Selection", *The VLDB Journal*, 2003 (to appear).
5. G. Aslan and D. McLeod, "Semantic Heterogeneity Resolution in Federated Database by Metadata Implantation and Stepwise Evolution", *The VLDB Journal*, 18(2), October 1999.
6. S. Chung, and D. McLeod. "Dynamic Topic Mining from News Stream Data", *Proceedings of ODBASE'03*, November 2003 (to appear).
7. Web Services Description Language, <http://www.w3c.org/TR/wsdl>.
8. http://opengis.net/gml/01-029/GML2.html.
9. M. Pierce, C. Youn, and G. Fox. "Application Web Services" Internal Comm. Grid Lab Report. <http://www.servogrid.org/slide/GEM/Interop/AWS2.doc>.
10. L. B. Grant and M. M. Gould, "Assimilation of Paleoseismic Data for Earthquake Simulation", *Pure and Applied Geophysics*, 2003 (to appear).