2009 Second International Symposium on Knowledge Acquisition and Modeling

An Ontology-based Semantic Search Approach for Geosciences

Jinhui Xiong, Weizu Huang, Chengzhu Jin
College of Resources and Civil Engineering, Northeastern University, Shenyang, China.
Jinhxiong@yahoo.com.cn, huangweizu@neu.edu.cn, Jinchengzu@neu.edu.cn

***Abstract*—Linguistic and semantic heterogeneities in geosciences are two main impediments in geo-information retrieving with popular search engines or autonomous data catalog from indexing services. In this paper, we proposed a semantic search approach for geosciences, which is applied to the Geoonto website. The search engine is implemented by a smart query agent mapping linguistically between lexicons and ontologies derived from geosciences thesauri, representing geological concepts with semantic relationships, and providing an inference service to retrieve the hierarchy from the geo-ontologies. Geoonto search tool uses these services to refine the users' search request to ensure the accuracy and completeness of query results.**

Keywords-semantic search; geoscience ontology; linguistic mapping.

1. INTRODUCTION

Geo-scientific knowledge differs from that in other domains in terms of complexity, incompleteness, indeterminacy, singularity, explanation, and historicity. This leads to challenges of interoperability between resources from heterogeneous information systems and web pages. Expected increases in both geo-information volume and Web-based accessibility suggest such challenges will continue to rise, exacerbating difficulties in finding, integrating, and using the information[1] . Furthermore, linguistic diversity in autonomous information systems and distributed web resources ultimately conduct the routine information handling and search seriously impeded. The challenges in semantic heterogeneity suggest the need to develop a shared ontology from the conventions of geosciences to supplement both syntactic and schematic interoperability in geosciences[2] .

There are typically two kinds of search approaches. One is the general web search engines, such as Google, AltaVista, Yahoo, Baidu, and the coming Microsoft Bing, which use special robots called web spiders to collect web pages from World Wide Web, and to rank the pages by using metadata extracted from collected HTML pages. Search engines also learn from user behavior to provide intelligent query suggestions, and in some circumstances to group the collected information. This approach retrieves all pages containing keywords submitted, and results in search completeness. The other kind of search approaches is implemented by using the metadata or database. Specifications such as Dublin Core Initiative, Federal Geographic Data Committee (FGDC)[3] , and Geological Information Metadata (DD2006-5) by China Geological Survey (CGS), etc., provide shared metadata schema for heterogeneous scientific data catalogs and databases. Queries on such data catalogs and databases usually generate more accurate search results. The metadata specifications facilitate interoperability between distributed autonomous data catalogs and databases. Nevertheless, these approaches failed in achieving ideal search result with both accuracy and completeness. How to bridge different languages in these search approaches remains a problem to be taken into account.

In this paper, we proposed a semantic search approach (abbreviated as Geoonto) for geosciences implemented at http://www.goeoonto.com. The approach improves the interoperability by addressing the linguistic, semantic, and syntactic heterogeneity in using of popular web search engines. In Section 2, we introduce the ontology concepts and representation language. Section 3 presents a multilingual geology ontology derived from several geo-scientific thesauri. Geoonto architecture and components are elaborated in Section 4. Section 5 gives a brief review to the proposed search approach.

2. ONTOLOGY AND THE SEMANTIC WEB

Ontology is the study of existence, of all kinds of entities - abstract and concrete - that make up of the world[4] . From a philosophic viewpoint, ontology is the study of a priori concepts. A computer-based ontology is defined as an explicit specification of shared concepts and theories, one that represents the intended meaning of a vocabulary[5, 6]. It is machine-processable as well as human-comprehensible[7] . Ontologies have two distinct components. One is the names for important concepts for a specific domain. The other is the relationships between these concepts. A linguistic ontology contains a list of terms in a glossary for a specific domain and relationships between the terms[8] . A mixed ontology is made up of a concept hierarchy. The concept hierarchy, called TBOX in knowledge base, consists of terms with generalization or specification relationships[9] . A disjoint architecture for geosciences could then benefit from a mixed ontology to bridge disjoint source database concepts and terms as well as link to prototypical evidence, and a linguistic ontology that catalogs standard terms and maps them onto concepts in mixed ontology. We constructed a conceptual hierarchy for geosciences and a geological lexicon for term mapping in the Geoonto architecture.

The key idea of semantic Web is the use of machine-processable Web information. Key technologies include explicit metadata, ontologies, logic and inference, and intelligent agents. Ontology needs a formal representation implemented in an ontology language. The most important ontology languages for Web are XML, XML Schema, RDF, RDF Schema, and OWL[10] . We chose OWL to represent the geosciences ontology for the Geoonto search approaches.

3. GEOONTO ONTOLOGY

Geoonto ontology is derived from four thesauri, includeing Asian Multilingual Thesaurus of Geosciences (ATMG), GeoScienceWorld (GSW) Topic Hierarchy,

978-0-7695-3888-4/09 $26.00 © 2009 IEEE
DOI 10.1109/KAM.2009.65

Chinese Geological Lexicon (CGL), and Glossary of Terms in Earth Material Science (GTEMS). We have collected and mined 16,107 lexical terms for lingual mapping between English and Chinese from web resources.

A thesaurus is a word list which allows standardizing terminology, useful for indexing and retrieving information of databases. Terms are related by a similar subject, and grouped into hierarchies and cross-references to other groups of terms which may be relevant to the subject. It provides the user with a single preferred term to describe a subject and allows terms to be selected at a general or specific level, depending on the level of indexing required[11]. The four thesauri mentioned above are introduced in the following text.

3.1 The AMTG and MT

The Asian Multilingual Thesaurus of Geosciences (AMTG) is sponsored by UNESCO and the French Ministry of Foreign Affairs. The AMTG is based on the Multilingual Thesaurus of Geosciences (MT) of the International Union of Geological Sciences/Commission on the Management and Application of Geoscience Information (IUGS/COGEOINFO) and the International Council for Scientific and Technical Information (ICSTI). The main aim of the AMTG project is to facilitate data exchanges between the various SANGIS organizations and linguistic groups in the field of Geosciences. AMTG contains 5,867 terms expressed as descriptors or on-descriptors in 11 languages. They are linked together by three types of relationships[12] :

- hierarchical relationships, which link some terms to other terms expressing more general and more specific concepts, i.e. broader terms and narrower terms. For instance the term *Homo* has for broader term *Hominidae* and for narrower terms: *Homo erectus, Homo habilis, Homo neanderthaliensis* and *Homo sapiens*.
- associative relationships, which link terms to similar terms (related terms). These relationships are established for terms which have the same broader term, for instance: *Homo* has the following related terms: *Australopithecinae* and *Ramapithecinae*.
- equivalence relationships, which link non-preferred terms to preferred terms. This relationship is indicated respectively by the terms Use for and Use. For instance: the non-preferred term *a layer* is linked to the preferred term *crust*.

3.2 GeoScienceWorld (GSW) Topic Hierarchy

GeoScienceWorld (GSW) is a nonprofit corporation formed by a group of leading geoscientific organizations for the purpose of making geoscience research and related information easily and economically available via the Internet[13]. GSW is an unprecedented collaboration of six leading earth science societies and one institute. It is a comprehensive Internet resource for research and communications in the geosciences, built on a core database aggregation of peer-reviewed journals indexed, linked, and inter-operable with GeoRef. The GSW Topic Hierarchy (GSWH) contains 22,532 terms grouped in three topics, Subject, Time and Geography. Each term has a scope note. Hierarchical, associative and equivalence relationships for terms are explicitly defined.

3.3 Geological Lexicons and Mined Terms

The Chinese Geological Lexicon (CGL) is maintained by Gansu Provincial Bureau of Geo-exploration and Mineral Development, China. There are 25,524 terms with scope notes and subject classifications specified. Each tern has its English annotation. But no relationships defined for terms[14] .

The Glossary of Terms in Earth Material Science (GTEMS) is developed by China University of Geosciences (CUG) in 2005 for educational purpose. It contains 5,907 terms listed in Chinese-English topics in earth material science, and 5,339 terms in English-Chinese vocabulary in earth material science[15] . No scope notes and relationships are defined. We have mined 16,107 lexical terms (MLT) for lingual mapping between English and Chinese from web resources.

3.4 GeoScienceWorld (GSW) Topic Hierarchy

Geoonto ontology is composed of ATMG concept hierarchy and GSW topic hierarchy. They map with each other through geological lexicons and mined lexical terms. Geoonto ontology is represented in OWL with capacity of inference with reasoner Pellet 1.5. Relationships such as Related and Synonymous are defined as object properties.

The concept hierarchy is built on generalization/ specification relationship. Each concept is a class in OWL, and has its super class. The top class is Thing. Related and Synonymous relationships are defined as transitive and symmetric object properties. These properties may reveal inexplicit meaning in concepts, and thus expands the concept hierarchy.

4. GEOONTO ARCHITECTURE AND COMPONENTS

There are three components in Geoonto tool as presented in Figure 1. These components are detailed next.

4.1 Ontology Inference

The ontology inference component provides ontology inference service (OIS) based on a reasoner Pellet. OIS interacts with Pellet reasoner through OWL API. Pellet is an OWL DL reasoner based on tableaux algorithms. It implements several extensions to OWL DL including combination formalism for OWL-DL ontologies[16] . The reasoner is launched with Geoonto ontology and provides Tbox and Abox queries on the ontology. Tbox queries deal with generalization, specifications and equivalence of a concept. Abox queries retrieve all satisfying instances of a concept and querying for property fillers for an instance. OIS receives query requests from Smart Query Agent and returns hierarchies retrieved from the reasoner.

4.2 Query Analysis

Smart Query Agent is not only a user interface but a query analyzer and a word segmentation processor. It manages requests via the client, runs word segmentation for Chinese requests, constructs a query request for OIS, receives the results from OIS, sends the inference results to search engines and data catalogs, aggregates the retrieved information from search engines and data catalogs, and finally responses the client with the aggregated information.

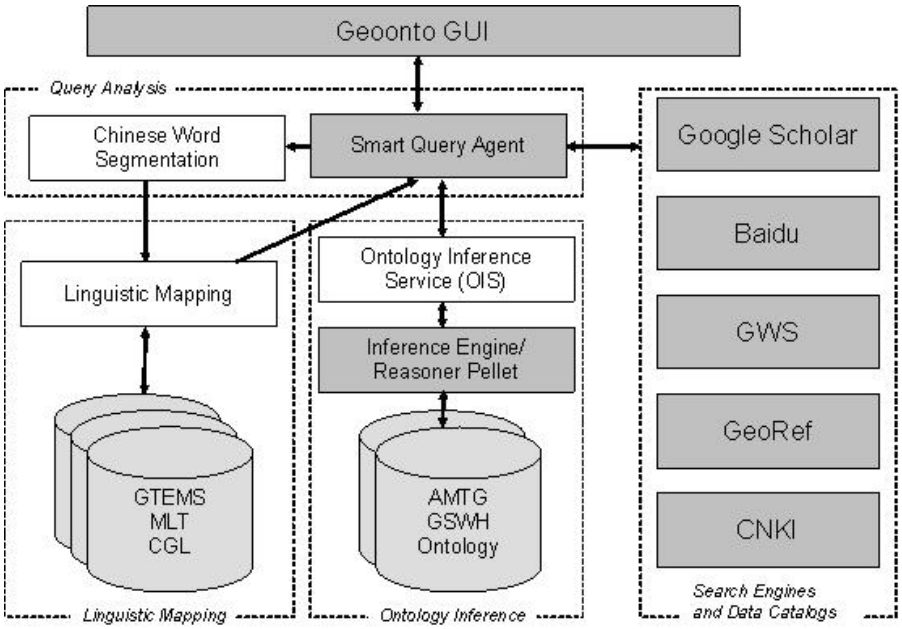

Figure 1. Geoonto architecture

TABLE I. ROLES OF GEOONTO ONTOLOGIES AND MAPPING LEXICONS*

Ontologies/Lexicons	*CH*	*SN*	*GN*	*SP*	*RT*	*SY*	*MP*
AMTG	yes		yes	yes	yes	yes	yes
GSW Topic Hierarchy (GSWH)	yes	yes	yes	yes	yes	yes	yes
Chinese Geological Lexicon (CGL)		yes					yes
Glossary of Terms in EarthMaterial Science (GTEMS)							yes
Mined lexical terms (MLT)							yes

*CH abbreviation for Concept Hierarchy, SN for Scope Notes of Terms, GN for Generalization Relationship, SP for Specification Relationship, RT for Related Relationship, SY for Synonymous Relationship, and MP for linguistic Mapping Role.

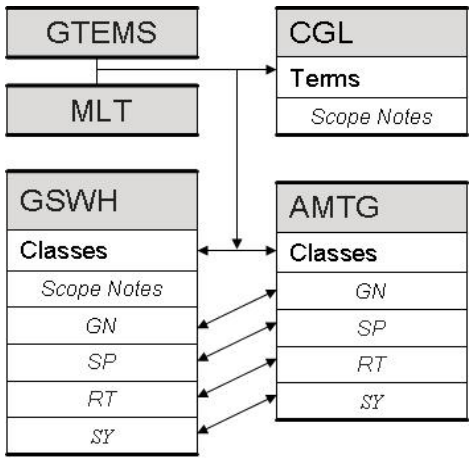

Figure 2. Mapping roles in Geoonto ontology

4.3 Linguistic Mapping

Geoonto tool provides linguistic mapping between English and Chinese concepts. Geoonto ontology is derived from several thesauri listed in the previous section. Each plays a different role in the search tool as shown in table 1. The roles of Geoonto ontologies and lexicons show that every term in Geoonto ontology shares a Chinese scope note as well as an English one when it is mapped onto GSWH and CGL. The terms submitted by a user through Geoonto tool are mapped onto AMTG, GSWH and CGL to get their own hierarchy and scope notes. AMTG maps with GSWH through GN, SP, RT and SY relationships. CGL, GTEMS and MLT serves as agencies in mapping operations. Figure 2 illustrates mapping roles in Geoonto ontologies.

For a specific term, Geoonto tool provides a linguistic mapping from GTEMS or MLT to CGL to get a Chinese scope note. Then the term will be mapped onto GSWH classes to get its English scope note and concept hierarchy. Finally the concept hierarchy will be mapped onto AMTG ontology to get a Chinese concept hierarchy.

5. USE CASE SCENARIOS

5.1 Linguistic Mapping Example

Linguistic mapping results in query suggestions for the user, such as word segmentation suggestions, lingual transaction suggestions, hierarchy in ontology suggestions and reference term suggestions. Each term in hierarchy has its lingual annotation. Other terms which take the

mapping term as prefix or postfix are listed as reference terms, and each has its lingual annotation.

5.2 Semantic Processing Example

The Geoonto website (http://www.geoonto.com) can be used to retrieve concept hierarchy from ontology (Figure 3). The user can pick up a term from the previous linguistic mapping scenario, say *"Homo"* for example, the Geoonto semantic processor component uses the ontology to find generalizations such as *"Hominidae"*, specifications such as *"Homo neanderthaliensis"* , *"Homo sapiens"*, and related terms like *"Abies"* , *"Acantharina"*. Each term has a lingual annotation in the hierarchy. The scope note for the concept is listed in the middle. The user can select these related terms to refine his query request, to retrieve satisfying accuracy result. Generalizations can help the user to receive completeness in search experience. Aggregation of retrieved from web search engines and scientific data catalogs is presented in this scenario.

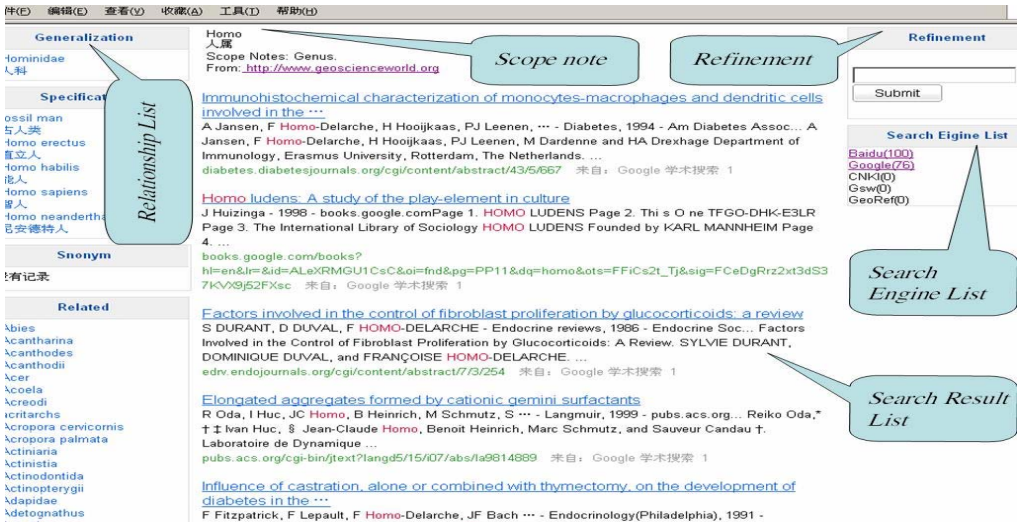

Figure 3. Semantic Processing Example.

6. CONCLUSIONS AND FUTURE WORK

The Geoonto search approach is based on a smart query agent, which retrieves information from popular web search engines or autonomous data catalogs/databases, integrates ontologies with request analyzing and associates semantic information in the search process, and generates a refined query string. Users can use the query string to extend their further requests. Linguistic Mapping and ontology mapping benefit users in covering lingual heterogeneities. Semantic heterogeneities in geosciences, due to its characteristic of complexity, incompleteness, indeterminacy, singularity, explanation, and historicity, can be hurdled with concept hierarchies derived from the ontology inference. The approach has been applied to provide online search service and has received some positive responses. Our future work will focus on evolvement of geo-ontology and aggregation of geosciences knowledge.

REFERENCES

[1] Brodaric, B. and M. Gahegan, Representing geoscientific knowledge in cyberinfrastructure: Some challenges, approaches, and implementations, in Geoinformatics: Data to Knowledge: Geological Society of America Special Paper 397, A. K. Sinha, Editor. 2006, Geological Society of America. p. 1-20.

[2] Breunig, M., An approach to the spatial data and systems integration for a 3D geo-information system. Computers & Geosciences, 1999. 25(1): p. 39-48.

[3] Ramachandran, R., et al., Ontology-based semantic search tool for atmospheric science, in 22nd International Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology. 2006: Atlanta.

[4] Sowa, J.F., Knowledge Representation: Logical, Philosophical, and Computational Foundations. 2000, New York: Brooks/Cole. 51,493-496, 594.

[5] Gruber, T.R., A translation approach to portable ontology specifications. Knowledge Acquisition, 1993. 5: p. 199-220.

[6] Guarino, N., Formal Ontology and Information Systems, in Formal Ontology in Information Systems, N. Guarino, Editor. 1998, IOS Press: Amsterdam. p. 3-15.

[7] Smith, B., Basic Concepts of Formal Ontology, in Formal Ontology in Information Systems, N. Guarino, Editor. 1998, IOS Press: Amsterdam. p. 19-28.

[8] Jackson, A.J., Glossary of Geology, 4th ed. 1997: Alexandria, Virginia, American Geological Institute. 769p.

[9] Nardi, D. and R.J. Brachman, An Introduction to Description Logic, in The Description Logic Handbook, F. Baader, et al., Editors. 2002, Cambridge University Press. p. 17-19.

[10] Antoniou, G. and F. van Harmelen, A Semantic Web Primer (Cooperative Information Systems). 2004, Cambridge MA, USA.: MIT Press. 18-19.

[11] Austin, D. and P. Dale, Guidelines for the Establishment and Development of Multilingual Thesauri, 2nd ed. 1981, Paris: Unesco. 64.

[12] CCOP and CIFEG, eds. Asian Multilingual Thesaurus for Geoscience. 2006. 633p.

[13] GeoScienceWorld, GSW Topic Hierarchy.

[14] Gansu BGMD, Chinese Geological Lexicon, http://www.gsdkj.net/pro/view.php.

[15] Sang Kanglong, et al., Glossary of Terms in Earth Material Science. 2005: China University of Geosciences. 365p.

[16] Sirin, E., et al., Pellet: A Practical OWL-DL Reasoner. 2005, Technical report, University of Maryland Institute for Advanced Computes Studies (UMIACS).