Semantic Web Service Based Geospatial Knowledge Discovery

Peisheng Zhao, Liping Di
Center for Spatial Information Science and Systems (CSISS)
George Mason University
6301 Ivy Lane Suite 620
Greenbelt, MD 20770
(pzhao, ldi)@gmu.edu

***Abstract* -- Large amount of Earth and space science data has been collecting from various sources. Effective and efficient knowledge discovery from these distributed multi-disciplinary and multi-scale data is becoming a big challenge. It requires the relevant data and processing steps be discovered, accessed and integrated as much as possible. The Semantic Web provides a common interoperable framework in which information is given well-defined meaning such that the data and operations can be used for more effective discovery and integration across various applications. This paper introduces a new approach to distributed data mining for geospatial knowledge discovery based on semantic Web services and their automatic and semi-automatic chaining. In this approach, domain concepts are well defined by geospatial ontology as the basic knowledge, data and data mining processes then are well described by these concepts and served by OGC Web services and semantic Web services. So the whole process of geospatial knowledge discovery can be represented as a service chain in predefined patterns of domain concepts. This approach provides an infrastructure that enables individual data and data mining software not only discoverable and accessible, but also interoperable in order to assemble them automatically or semi-automatically to implement more complicated geospatial knowledge discovery.**

Keywords-Web Service: Semantic Web; Ontology; Data Mining; Service Chain

I. INTRODUCTION

Overwhelming volume of Earth and space science data has being collected from modern-day satellites and other data acquisition systems. The data are processed and managed by a variety of geographically distributed data providers. NASA's Earth Observing System (EOS), for instance, has been generating almost 100 gigabytes of image data per hour on average for the past decade, and releases over 900 Earth science data products at more than a dozen data centers. It is extremely valuable for innovative scientific researches and decision-making processes to extract useful information and knowledge from these distributed massive volumes of data. Data mining, also known as knowledge discovery, can help users detect and interpret patterns and regularities, discover classification rules and infer causation by processing and combining raw data. With complex spatial and/or temporal dynamics, geospatial knowledge discovery involves specifically a complex workflow that commonly requires integration of various data mining algorithms and distributed multi-disciplinary, multi-source, and multi-scale science data. It may also require that the user understand more about the geospatial domain and the variety of data mining techniques than their training provides. Therefore, it is more necessary to develop a novel method to characterize the data mining techniques and relevant datasets for the purpose of easy discovery and accessibility, and more specifically for efficient machine-accessible to implement a task of geospatial knowledge discovery automatically or semi-automatically. The Semantic Web provides a promising common interoperable framework in which information is given well-defined meaning in unambiguous and computer-interpretable form by using ontology such that data and services can be used for more effective discovery, automation, integration, and reuse across various applications. In this paper, we propose a new semantic Web service-based approach to intelligent geospatial knowledge discovery. This approach provides an infrastructure that enables individual data and data mining process not only discoverable and accessible, but also interoperable in order to assemble them automatically or semi-automatically to implement more complicated geospatial knowledge discovery.

The reminder of this paper is organized as follows. In section 2, we discuss the related work. In section 3, we discuss the ontology-based knowledge base which uses ontology to capture domain knowledge and map data mining models to scientific problems. In section 4, we discuss the composition of scientific workflow (i.e. service chain) for geospatial knowledge discovery. And finally in section 5, we present the conclusions and future work.

II RELATED WORK

It is widely recognized that ontology is critical for the development of the Semantic Web. Ontology has originated from philosophy as a reference to the nature and the organization of reality. In general, ontology is a "*specification of a conceptualization*" [1]. In computer science domain, ontology provides a commonly agreed understanding of domain knowledge in a generic way for sharing across applications and groups [2]. Usually ontology is used for [3]:

- **communication** among humans, computational systems, or between humans and computational systems.
- **computational inference** on concept relationships and computational plan, e.g. analyzing the internal structures, algorithms, inputs and outputs of implemented systems in theoretical and conceptual terms.
- **reuse and organization of knowledge**, such as standardizing libraries or repositories of plans and domain information.

In the data mining context, ontology is viewed as a formal structure which encapsulates the semantics of data mining techniques and data sources and relates them to the concepts within disciplines. In [4], ontology for the data mining domain is presented in order to simplify the development of distributed knowledge discovery. This ontology offers a reference model for different kinds of data mining tasks, methodologies and software to help users find the most appropriate solution. It is noted that this ontology is oriented to the general data mining problems so that it only conceptualizes the generic data mining techniques and does not consider domain concepts and data. As scientific problem is more complex, [5] proposes to relate the data to the domain concepts by using ontology and thus users can easily retrieve the relevant data sets to be compared by navigating the ontology. But its work is not concerned with data mining techniques.

To provide formal semantic descriptions of NASA data set and scientific concepts, several projects are underway to develop a semantic framework. Described in the OWL language, the ontologies within the Semantic Web for Earth and Environmental Terminology (SWEET) [6] contain several thousand terms spanning a broad extent of Earth system science and related concepts (such as NASA GCMD, ESML, ESMF, grid computing, and OGC). The SWEET provides a high-level semantic description of Earth system science. The ontologies of geographic information metadata (ISO 19115 and FGDC) being developed in [7] [8] add the semantic meanings to the data description by which data sets are explicitly associated with providers, instruments, sensors and disciplines, and the relationships across these concepts. However, the mapping between the metadata ontologies is still in investigation.

III ONTOLOGY-BASED KNOWLEDGE BASE

Knowledge base provides the overall knowledge of the process of geospatial knowledge discovery. In this knowledge base, we use a set of ontologies to capture geospatial domain knowledge and map data mining models to geospatial scientific problems as Fig. 1, i.e. domain terms and concepts, linkage between concepts and datasets, relationships among heterogeneous data, and associations between models and data. The use of ontologies to describe data mining models and their relevant datasets gives well-defined semantic meaning for the diverse data sources and data mining tasks. Thus, the ontology-based knowledge base can help users to efficiently find the best solution and the most appropriate data.

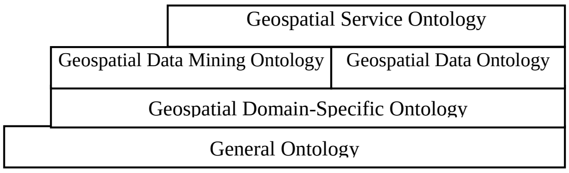
Geospatial Service Ontology
Geospatial Data Mining Ontology
Geospatial Data Ontology
Geospatial Domain-Specific Ontology
General Ontology

Fig. 1. Ontologies for Geospatial Knowledge Discovery

The general ontology is the core upper level vocabulary for all of human consensus reality. It is a common language that all other ontologies must reference. We use the Dublin Core Metadata and OpenCyc as the basis to define upper level concepts and assertions about these concepts.

Geospatial domain ontology aims at providing the core conceptualization and knowledge structure of geospatial domain. For example, “surface water” belongs to “hydrosphere”, and “river” is a kind of “surface water”. It represents the problem space over which the user will query. Other ontologies directly or indirectly incorporate geospatial domain ontology. The SWEET ontology provides an upper-level semantic description of Earth system science. But it mainly concerns with physical property of Earth. We use it as a starting point to reorganize and expand the concepts to cover the NASA scientific researches as much as possible by incorporating the terms in the Global Change Master Directory (GCMD) and the Earth Science Modeling Framework (ESMF). Geospatial domain ontology covers the knowledge about 1) spatial-temporal factors, e.g. location, time and unit, 2) physical facts, e.g. physical phenomena, physical properties and physical substances, 3) disciplines, e.g. scientific domains and projects, and 4) data collection, e.g. instruments, platforms and sensors.

Geospatial data ontology provides an ontological view of NASA distributed heterogeneous data resources. It directly incorporates the domain ontology to link the data with scientific research. NASA has used ECS metadata to describe data in NASA data centers. Geospatial data ontology adds the semantics into the metadata that allows a user to locate data without knowing the exact metadata keywords used by NASA because the terms in query have an equivalent definition in the spatial domain components. As mentioned previously, the metadata ontologies of ISO 19115 and FGDC have been developed. The NASA ECS metadata ontology is developed by incorporating these two ontologies. To provide a unified view of metadata, the semantic relationships among terms in different metadata standards are defined. Thus, there is no distinct boundary across various metadata standards. The user can use any term from any one of the metadata standards to query the data described in any one of metadata standards.

Data mining ontology provides a reference model for different kinds of data mining techniques. It directly incorporates the geospatial domain ontology and geospatial data ontology to associate the data mining techniques with scientific problems and relevant data sources. Based on the ontology described in [4], we develop a data mining ontology oriented toward the research themes in NASA Earth-Sun system. The ontology represents the features of the available data mining techniques, classify their internal structure, and document the relationships and the constraints among them. The following important concepts related to the data mining are included in this ontology.

- Scientific discipline: the domain that the data mining technique can be applied, such as solar irradiance and soil moisture. The definition of the domain is documented in spatial domain ontology.
- Scientific mission: the specific goal of a data mining technique, such as classification and clustering.
- Methodology/Algorithm: the type of methodologies and algorithms used in the data mining process, such as neural network and Bayesian network.
- Data source: the type of data and its sources the data mining technique can work on. The data type is defined in data source ontology.
- Mining result: the properties of output of a data mining process, such as running time and accuracy.

Geospatial service ontology enables automatic discovery, invocation and composition of all registered services conforming to the OWL-S specification. It directly incorporates the data mining ontology and data source ontology. The "*profile*" of OWL-S, which describes who provides the service, what the service does, as well as other properties of services, allows the knowledgebase to infer whether or not a particular service is appropriate to a given problem. The "*process model*" of OWL-S, which states the inputs, outputs, preconditions and effects of a service, allows the knowledgebase to figure out whether or not a service meets the requirements as well as the conditions to invoke the service. The "grounding" of OWL-S, which presents the ports, protocols and encoding of invocation, tells knowledgebase how to invoke a service. Moreover, the semantic service description enables us to create composite services.

Since all of the ontologies are represented by OWL, the inference engine in the knowledgebase is an OWL reasoner built on Prolog. Ontological information written in OWL or OWL-S is converted into RDF triples and loaded into the knowledgebase. The engine has built-in axioms for OWL inference rules. These axioms are applied to facts in the knowledgebase to find all relevant entailments such as the inheritance relation between classes that may be not directly in the subclass relationships.

IV COMPOSITION OF SERVICE CHAIN

Geospatial knowledge discovery usually involves a complex scientific workflow which consists of data and process selection, data pre-processing, information and knowledge extraction (data mining), interpretation (post-processing), and visualization. This workflow may have many steps of distributed computation requiring that individual processing steps be discoverable, accessible and interoperable. As Web services are interoperable, it is essential to assemble individual Web services into a service chain to represent a complex geospatial model and process flow to implement geospatial knowledge discovery. Such a method has great potential for reducing the development time and costs. Fig. 2 shows our approach to geospatial knowledge discovery based on semantic Web service. Actually, the process of geospatial knowledge discovery is the process of building service chain in this approach.

All data should be virtually online to allow the data mining system's rapid access. A data service portal that incorporates the catalog service is provided to wrap the online distributed NASA data and make them available through OGC Web Coverage Service, Web Feature Service and Web Map Service. Thus, it is possible for seamless access of geospatial data in a distributed environment with standard interfaces regardless of the low level of accessing details. An important component of this approach is the ability to fuse data or combine evidences from multiple mining activities in order to provide more appropriate data and more accurate picture of the result. This is achieved by the development of a set of data fusion services, i.e. Web Coordinate Transformation Service, Web Image Cutting Service and Web Data Format Translation Service. To make data mining functions/applications describable, discoverable, chainable, and accessible, data mining applications are developed as semantic Web services from scratch or wrap an existing data mining application to become a semantic Web service. All these data mining services are linked to the appropriate problem space and data. Usually catalog services play a 'directory' role: providers advertise the availability of their resources using metadata in a catalog; users can then query the metadata in the catalog to discover interesting resources. The OGC CS/W (Catalog Service for Web) is the *de facto* standard for supporting the discovery of and binding to registered geospatial information resources. By incorporating the ontologies in the knowledgebase, the OGC CS/W in our approach supports the flexible semantic matching of data and data mining techniques regardless of syntactic differences, especially the matching of subsumption hierarchy of data type and service type.

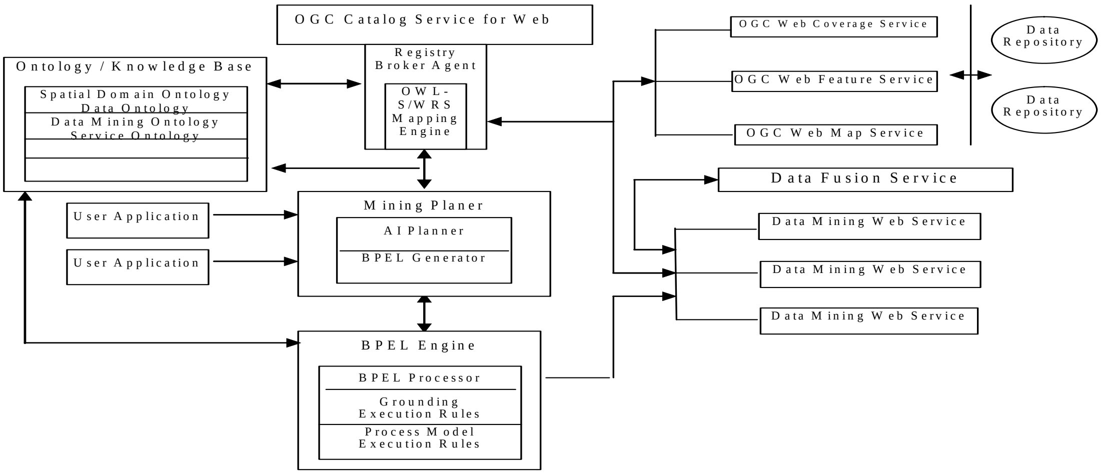

Fig. 2. Geospatial knowledge discovery: a semantic Web service-based approach

Assembling individual services into a more complex Web process (e.g. service chain) to achieve desired results proves to be essential for the knowledge discovery. With the help of ontologies in the knowledgebase, we aim at realizing three types of service chain defined in [9].

- **User-defined (transparent):** the user queries the catalog service with more specific details on the different data mining resources to define and manage the service chain.
- **Workflow-managed (translucent):** the user queries the catalog service for a given problem, and then the knowledgebase assists the user to select and configure the most suitable data mining solution in each step of chaining process.
- **Aggregate (opaque):** the user presents a problem, and then the knowledgebase incorporates the catalog service to build a service chain with the best data mining solution without user's intervention.

Ontology plays a very important role in the chaining process. It suggests what to do and what to use on the basis of users' requirements, e.g. locating data sources on a specific topic and finding data mining services for a desired data mining algorithm, a particular method, or a specified data mining task. By translating user's goal into sub-problems that can be solved by available data mining services – the construction of service chains, a "*Mining Planer*", based on AI planning algorithms, incorporates the ontologies to layouts the service chain automatically or semi-automatically. We adopt the Business Process Execution Language (BPEL), which is widely used in e-business, to represent service chains. A "*BPEL Engin*e", which is a BPEL process manager for integrating services into collaborative and transactional processes within a Service Oriented Architecture (SOA), is used to mange, deploy and execute the service chains.

V CONCLUSION

The most significant distinction of the proposed approach is to use semantic Web service to bridge the growing interoperable gap between data collection and analysis that hinders geospatial knowledge discovery. In this approach, domain concepts are well defined by geospatial ontology as the basic knowledge, data and data mining processes are well described by these concepts and served by OGC Web services and semantic Web services. So the whole process of geospatial knowledge discovery is to build a service chain in predefined patterns of domain concepts automatically or semi-automatically. This approach provides a prominent mechanism that enables scientists and decision-makers to fully exploit the potential of the geospatial data and data mining technologies.

Ontology plays a critical role in our approach. However, it is impossible to exhaustively put all relevant geospatial information into ontology. In the next step, we will investigate more existing geospatial ontologies and standards to sketch geospatial space precisely and elaborate the relationships inherent in the nature of geospatial data and data mining technologies.

REFERENCES

[1] T. Gruber, "A translation approach to portable ontologies", in *Knowledge Acquisition*, Vol. 5, Issue 2, 1993, pp. 199-220.

[2] B. Chandrasekaran, T. Johnson, V. Benjamins, "Ontologies: what are they? why do we need, them?", in *IEEE Intelligent Systems and Their Applications*. Vol. 14, Issue 1, 1999, pp. 20-26.

[3] Ontology (Special Issue on), in *CACM*, Vol. 45, Issue 2, 2002.

[4] M. Cannataro, C. Comito, "A Data Mining Ontology for Grid Programming", in *Proceedings of 1st Int. Workshop on Semantics in Peer-to-Peer and Grid Computing*, Budpest, Hungary, 2003.

[5] S. Tadepalli. A. Sinha, N. Ramakrishnan, "Ontology Driven Data Ming for Geosciences", in *Proceedings of 2004 AAG Annual Meeting*, Denver, USA, 2004.

[6] R. Raskin, "Enabling Semantic Interoperability for Earth Science Data". http://sweet.jpl.nasa.gov/EnablingFinal.doc

[7] A. Islam, L. Bermudez, B. Beran, S. Fellah, M. Piasecki, "Ontology for Geographic Information – Metadata", 2003. http://loki.cae.drexel.edu/~wbs/ontology/iso-19115.htm

[8] P. Zhao, "Geospatial Ontology", 2005 http://geobrain.laits.gmu.edu/ontology/

[9] ISO, "Geographic information – Services"