Computers & Geosciences 36 (2010) 270–281

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Augmenting geospatial data provenance through metadata tracking in geospatial service chaining

Peng Yue [a,*], Jianya Gong [a], Liping Di [b]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan 430079, China
[b] Center for Spatial Information Science and Systems (CSISS), George Mason University, 6301 Ivy Lane, Suite 620, Greenbelt, MD 20770, USA

ARTICLE INFO

Article history:
Received 18 April 2008
Received in revised form
6 August 2009
Accepted 20 September 2009

Keywords:
Geospatial Web Service
Service chaining
Data provenance
Metadata tracking
GIS
Semantic Web

ABSTRACT

In a service-oriented environment, heterogeneous data from distributed data archiving centers and various geo-processing services are chained together dynamically to generate on-demand data products. Creating an executable service chain requires detailed specification of metadata for data sets and service instances. Using metadata tracking, semantics-enabled metadata are generated and propagated through a service chain. This metadata can be employed to validate a service chain, e.g. whether metadata preconditions on the input data of services can be satisfied. This paper explores how this metadata can be further exploited to augment geospatial data provenance, i.e., how a geospatial data product is derived. Provenance information is automatically captured during the metadata tracking process. Semantic Web technologies, including OWL and SPARQL, are used for representation and query of this provenance information. The approach can not only contribute to the automatic recording of geospatial data provenance, but also provide a more informed understanding of provenance information using Semantic Web technologies.

1. Introduction

1.1. Metadata tracking in geospatial service chaining

Web services technologies have shown promise for providing heterogeneous data from distributed data archive centers for open worldwide use. Previously stand-alone geo-processing functions are now being wrapped as interoperable web services that can be chained to support a “Cyberinfrastructure for e-Science” (Hey and Trefethen, 2005). The Open Geospatial Consortium (OGC) is developing geospatial Web services standards by adapting or extending common-purpose Web service standards. Through the OGC Web Services (OWS) testbeds, OGC has been developing a series of interface specifications under the OGC Abstract Service Architecture (Percivall, 2002), including Web Feature Service (WFS), Web Map Service (WMS), Web Coverage Service (WCS), Catalogue Services for the Web (CSW), and Web Processing Service (WPS). To solve a complex real world problem in a service-oriented environment, multiple services must be chained together. Although the manual composition of service chains is useful, it needs considerable time and requires users to be both domain and technical experts. The wide use of accessible geospatial data and services over the Web requires a certain degree of automation for service composition.

Automatic service composition is a hot research topic in computer science (Srivastava and Koehler, 2003; Rao and Su, 2004). It consists of three phases: (1) process modeling, which involves generating an abstract composite process model consisting of the control flow and data flow among atomic processes; (2) process model instantiation, where the abstract process is instantiated into a concrete workflow or executable service chain; and (3) workflow execution, where the chaining result or workflow is executed in a workflow engine to generate on-demand data products. In the geospatial domain, the process model is a geo-processing workflow, which transforms source data into value-added data products. Each process node (i.e., atomic process) in the process model represents one type of geospatial service. All share the same functional behaviors: functionality, input and output. The descriptions of these behaviors can use service ontologies from the Semantic Web (Berners-Lee et al., 2001). Many approaches are available for generating a process model based on the service ontologies using Artificial Intelligence (AI) planning methods (Peer, 2005). A process model contains knowledge about how to generate a data product. Since this data product does not really exist in any archive, it is regarded as a *virtual data product*. This virtual data product represents a geospatial data type, not an instance (an individual data set), that the process model can produce. It can be materialized on demand for users when all required

[*] Corresponding author. Tel.: +86 27 68778755
E-mail address: geopyue@gmail.com (P. Yue).

0098-3004/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2009.09.002

geo-processing services and input data are available. The materialization of a virtual data product requires metadata specifications such as spatial bounding box and spatial projection. The term *metadata* in metadata tracking means descriptive information for data products such as that defined in ISO 19115 (ISO/TC 211, 2003). Through propagating these specifications to each process node of a process model, the whole process model is instantiated. Therefore, metadata tracking, the generation and propagation of geospatial metadata through the process model, is a key step towards the instantiation of the process model.

1.2. Geospatial data provenance

As the number of geospatial services grows with the wider integration of geospatial services, it becomes important to automate the recording of data provenance rather than relying on manual work (Foster, 2005). Data provenance, also referred to as lineage, contains information about the sources and production processes used in producing a data product (ISO/TC 211, 2003). With the development of multi-sensor and multi-platform technologies, the processing and transformation of multi-resolution and multi-spectral images becomes more and more frequent and complex. Therefore, data provenance is important to help users make decisions about the quality of derived data products, discover dependencies among data and services, or re-enact the process of derivation of data products.

This paper describes a synergistic effort between automatic service composition and data provenance. Most existing work on data provenance is in the domain of general information (Bose and Frew, 2005; Simmhan et al., 2005; Miles et al., 2007; da Silva et al., 2003; Zhao et al., 2004; Foster et al., 2002; Golbeck and Hendler, 2007) and does not include content specific to the geospatial domain, such as geospatial metadata standards. Although there has been some work in the geospatial domain (Lanter, 1991, 1992; Alonso and Hagen, 1997; Frew et al., 2001, 2007; Wang et al., 2008; Tilmes and Fleig, 2008), it did not consider the service-oriented environment enabled by OGC Web service standards. The emergence of Semantic Web technologies, including the Resource Description Framework (RDF) (Klyne and Carroll, 2004), the Web Ontology Language (OWL) (Dean and Schreiber, 2004), and the SPARQL Protocol and RDF Query Language (SPARQL) (Prud'hommeaux and Seaborne, 2006), provides a way to connect data for more effective discovery and integration, and thus shows considerable promise for new approaches to geospatial data provenance. In addition, the previous work focused mostly on analyses of provenance information that was created during execution, rather than on metadata generated before execution (Kim et al., 2006). The geospatial metadata generated during process model instantiation provides a context for evaluating the quality and reliability of the data product before the intensive execution of the workflow, thus contributing to the data product's provenance. Therefore, this paper presents how to interleave the Semantic Web approaches for data provenance with metadata tracking to record and query provenance information generated in instantiating a process model. The contributions of this paper are: (1) a model and semantic representation for geospatial data provenance that integrates the geospatial metadata standard and process models for geo-processing services and service chains; (2) automatic capture of geospatial data provenance through metadata tracking in the phase of process model instantiation; and (3) support to the storage and query of geospatial data provenance through Semantic Web technologies.

The remainder of the paper is organized as follows. Section 2 introduces a use case to help in understanding our work. The primary challenges for research on geospatial data provenance in a service-oriented environment are described in Section 3. In Section 4, the approach for addressing these challenges is presented, including a semantic representation of geospatial data provenance, a metadata-tracking component, and extension of the metadata-tracking component to support automatic recording and querying of geospatial data provenance. The work is compared with related work in Section 5, and conclusions and pointers to future work are given in Section 6.

2. A use case

An Earth science application serves as an example to help understand metadata tracking during service chaining and to illustrate how metadata tracking can contribute to data provenance. The application is wildfire prediction from weather and remote sensing data. The wildfire prediction process uses a variety of geospatial data items when creating the wildfire prediction product. This input data consists of the Leaf Area Index (LAI), Fraction of Photosynthetically Active Radiation (FPAR), Land Cover/Use Types (LULC), daily maximum temperature, daily minimum temperature, and precipitation.

The process model is derived in the process modeling phase. Its representation is formalized through an ontology approach. In information sciences, an ontology is a formal, explicit specification of a conceptualization that provides a common vocabulary for a knowledge domain and defines the meaning of the terms and the relations between them (Gruber, 1993). Ontologies are crucial to making the semantics of the exchanged content machine-understandable. OWL, recommended by W3C as the standard Web ontology language, is designed to enable the creation of ontologies and the instantiation of these ontologies in the description of resources. Therefore, process models for geo-processing workflows are addressed through the introduction and design of OWL-based ontologies conveying semantic information on geospatial services and data. The following ontology entities are linked to the process model for wildfire prediction in the upper part of Fig. 1: "WildFirePrediction" for the semantics of service functions, "FPAR", "LAI", "IGBP_CLASS[1]", "Maximum_Temperature", "Minimum_Temperature", and "Precipitation_Amount" for the semantics of input data, and "Wildfire_Danger_Index" for the semantics of output data.

We can refer to this process model as a virtual data product for wildfire prediction. Therefore, an instance of this virtual data product can be generated with metadata specifications through the materialization process. For example, a user provides the spatial (e.g. Bakersfield, CA, United States) and temporal (e.g. August 26, 2006) information. A semantically augmented geospatial catalogue service (Yue et al., 2006) can be used to automatically determine that the National Oceanic & Atmospheric Administration (NOAA) National Digital Forecast Database[2] (NDFD) can provide the weather data (MAXT, MINT and QPF) and National Aeronautics and Space Administration (NASA) Earth Observing System (EOS) Moderate Resolution Imaging Spectroradiometer (MODIS)[3] products can provide FPAR, LAI, and LULC.

[1] Land cover classes defined by the International Geosphere–Biosphere Program (IGBP).

[2] The operational NDFD data provided by the NOAA National Weather Service (NWS) are stored in the GRIB2 data format with a Lambert conformal coordinate reference system and a spatial resolution of 5-km.

[3] The operationally available NASA data in the Land Processes Distributed Active Archive Center (LPDAAC) are stored in HDF-EOS data format, and in a sinusoidal grid coordinate reference system at a spatial resolution of 1-km. The MODIS grids are stored as tiles, each covering approximately 1200 × 1200 square kilometers.