A Framework for Intelligent Geospatial Knowledge Discovery at Web Service Environment

Prof. Liping Di
Laboratory for Advanced Information Technology and Standards (LAITS)
George Mason University
9801 Greenbelt Road, Suite 316-317
Lanham, MD 20706
ldi@gmu.edu
301-552-9496
Introduction

• Geospatial data is the major type of data that human beings has collected.
  – more than 80% of the data are geospatial data.

• Image/gridded data is the dominant form of geospatial data in terms of volume.
  – Most of those data are collected by the Earth Observing (EO) community.

• Geospatial data will grow to ~exabyte very soon.
  – NASA EOSDIS has more than two petabytes of data in archives and is added more than 2 terabytes of new data per day.
  – Application data centers: 10’s of terabytes of geospatial data.
  – Tens of thousands of datasets on-line now.

• How to effectively, wisely, and easily use the geospatial data is the key geospatial technology issue that we have to solve.
Steps from Geospatial Data to Information & Knowledge

• In order for the geospatial data to be useful, they have to be converted to user-specific information and knowledge.

• Three steps in the processes for conversion of geospatial data into information and knowledge
  – Geoquery: to discover the data needed for the project and obtain the data.
  – Geo-assembly: assemble data from multiple sources into a homogeneous form for analysis.
  – Geo-analysis/modeling: geospatial information extraction (GIE) and geospatial knowledge discovery (GKD).

• Currently, a typical geospatial project requires data from multiple sources, therefore the user needs to query and order data from multiple data systems for data finding and ordering.

• The geo-assembly and geo-analysis/modeling are taken place at user’s site with the available local resources.
The Problems

- The conversion from Geospatial data to knowledge requires the users at their local site having:
  - Significant amount of knowledge
    - Domain knowledge for information/knowledge extraction from raw data
    - Domain knowledge on the geospatial data processing/formats
  - Significant amount of computer hardware and software resources.
  - As a result, currently the use of geospatial data is very expensive
- Most geo-imagery will never be directly analysed by humans
  - Human attention is the scarce resource, insufficient to analyse petabytes of geospatial data.
  - Many datasets have not been analysed once before they are archived.
- The fundamental problem is that current data and information systems at data providers’ site can only provide on-line data ordering and access at best, not the user-specific information and knowledge.
  - Rich in geospatial data but poor in up-to-date geospatial information and knowledge.
What Do Users Want?

- They want to have ready-to-use geospatial information and knowledge that can answer their specific application questions.
- Most of users don’t care very much about whether or not the answer is derived from Landsat TM, or SPOT, or field observations, as long as the users can obtain the accurate answers easily from geospatial information systems that meet their application requirements.
- Instead of only experts can use the geospatial data, everyone, from students to decision-makers can obtain and use the geospatial information and knowledge easily.
What Do We Need to Do?

• As geospatial technology developers, we need to:
  − Make the geospatial information the mainstream information so that users can easily obtain the ready-to-use customized geoinformation and knowledge if they want.

• What we need to develop is a system that can automatically convert the geospatial data to user-specific geoinformation and knowledge.
  − Automate the process from geospatial data to information to knowledge based on users’ requirements.
  − Change the paradigm of geospatial information process from everything locally owned and operated to large-scale sharing of geospatial data, information, software, hardware and other resources over the web.
  − The key for enabling the sharing is the standard-based service
The research objectives

• Develop an interoperable system framework for enabling
  – choreographed intelligent web services for geospatial knowledge discovery;
  – the dissemination of personalized, on-demand geospatial data, information and knowledge.

• The system can answer many “what if” questions through
  – automatically and intelligently chaining individual service modules to form a complex geospatial model;
  – matching the input data with the model;
  – executing the model to deliver the answer to the users.

• The system has capabilities for self-improvement and evolution through community-involved open, accumulated approach for sharing geospatial knowledge.
  – allowing individuals to contribute geospatial web-service modules and models that can be reused by others in the community.
  – The contributed modules/models become a part of the system capability.
Technology foundation

• The geo-object and geo-tree concepts developed by the PI, which provide the mechanism for
  – unifying the representation of geospatial data, information, and knowledge
  – capturing the process of geospatial knowledge discovery
  – managing geospatial knowledge
  – enabling the dynamic reuse of the geospatial knowledge in multiple applications.

• The geospatial interoperability standards developed by ISO, FGDC, and OGC;
  – enabling the development of chainable services

• Interoperable geospatial web services and Grid technology.
  – providing the framework for deploying distributed services.

• Semantic Web and Artificial Intelligence (AI) technology
  – enabling the automatic construction of solutions to users’ questions.
The Service-Oriented Architecture (SOA)

- A service is a well-defined set of actions in a computer system. It is self-contained, stateless, and does not depend on the state of other services.
- Stateless means that each time a consumer interacts with a service, an action is performed. After the results of the service invocation have been returned, the action is finished. There is no assumption that subsequent invocations are associated with prior ones.
- Services are one of the key components in the service-oriented architecture.
- In the service-oriented architecture, the description of a service is essentially a description of the messages that are exchanged between the consumer and the service.
- The real power of SOA is that individual standard-based services can been chained together to solve complex tasks dynamically.
- Implementation of services in the web environment is the web services and in the Grid environment is the Grid services.
Basic Service Operations

1. Publish
2. Find
3. Bind
4. Chain
Major Web Service Standards

- **SSL**
- **WSDL**
- **XML**
- **UDDI**
- **SOAP**
- **BEPL**

**Security Stack**
- HTTP
- XML

**Service Description Stack**
- XML

**Service Discovery Stack**
- XML

**Service Binding Stack**
- HTTP

**Modeling & Workflow Stack**
- XML

Chain
Distributed Geospatial Services

- Geospatial services are those handling the geospatial data and information.
- It is envisioned that in the near future many people will set up various geospatial services on the web just like today’s websites
  - Some may provide data while others may provide information processing services (e.g., classification) or geospatial brokerage services.
  - Because those services are set up by many different people over the web, they are naturally distributed.
- If geospatial service standards (mainly for the interface standards) are available and most of the people follow the standards for setting their services, individual services can be chained together to solve complex geospatial problems.
  - The problems will be solved over the web instead of locally.
  - The user will obtain the result instead of handle the details of data and process transactions.
Solve complex geospatial problems by using distributed geospatial services:

The Geo-object and Geo-tree Concepts

archived geo-object
user geo-object
Intermediate geo-object

User-Requested object

User-Obtained geo-object

Geospatial web/Grid services
Automated data transformation service (WGS/WFS)
Relationship between Geo-tree and Service Chain

- Geo-tree is a conceptual geospatial process model representing the step-by-step how a geo-object (representing either GI or GK) is derived.
  - representing the domain knowledge needed for analyzing geospatial data to derive user-specific GI or GK.
  - A library of Geo-trees represents the enterprise/community knowledge since it is contributed by the members of them.
- The root of a geo-tree is a virtual product that the Geo-tree can derive, and all sub-trees are also virtual products.
- A service chain is the instantiation of the Geo-tree. If a Geo-tree can be instantiated, then all virtual products in the tree can be produced on demand.
Types of Service Chaining

• Three types of service chaining are defined in ISO 19119:
  – User-defined (transparent) – the Human user defines and manages the chain.
  – Workflow-managed (translucent) – the Human user invokes a service that manages and controls the chain, where the user is aware of the individual services in the chain.
  – Aggregate (opaque) – the Human user invokes a service that carries out the chain, where the user has no awareness of the individual services in the chain.

• An intelligent geospatial system based on distribute geospatial web services needs to support all three chaining methods.
Construction of Service Chains

- The first type of chaining allows users to construct a geospatial model to be run in the system
  - Require domain knowledge—for expert to contribute their domain knowledge.
  - The knowledge is kept in the Geo-tree/service chain.
- The second type of chaining basically is to use existing geo-tree to materialize a virtual object.
  - Anyone can use this type of chaining to produce a virtual product on demand.
  - Anyone can use but it is not able to produce a product who’s geo-tree doesn't already exist in a data/information system.
- The third type of chaining require the system to be intelligent enough to automatically form a geo-tree/service chain by decomposing user’s query.
  - require the domain knowledge
  - require the automated reasoning.
  - Anyone can use and can produce a new product based on users’ query automatically.
- The first two types of chains do not require significant machine intelligence.
  - Current technology is enough for implementing such chaining approach.
- The third one requires significant machine intelligence
  - Current technologies are not able to provide such kind of chaining.
  - Significant research is needed.
Features of Service-based Geospatial Systems

• Evolvable
  – The system will grow its capabilities with more and more service modules and models become available.

• Truly open, distributed system
  – anyone can set up a service and to become a part of the system after proper registration of his services.

• Build by the community for the community
  – User community involvement of the development is one of the keys for the success of the system since they can contribute service modules or models to the system.
  – A contributor is not necessary to set up the services by using their resources. The contributions can be hosted anywhere on the web.

• Following standards for building the service modules and models are the key for the system to work.
The Key Standard Environments

• The distributed geospatial services have to be built on the two key standard environment.
  – common data environment
  – common service environment

• The common data environment provides the standard interfaces for data discovery and access from multiple data archives.
  – upon that the service can be built independent of data sources and formation.
  – The OGC WCS, WFS, WMS, and WRS are the fundamental standards for building a common data environment.

• The common service environment provides standard interfaces and methods for service declaration, discovery, chaining, binding, and execution.
  – foundation for interoperating and chaining services.
  – The W3C provides the base standards and OGC provides the geospatial extensions.
From Geospatial model to user geo-object

Knowledge Capture phase

User query Phase

User retrieval phase

Geospatial Model → Virtual geo-object → Logical Workflow → Concrete Workflow → Workflow execution → user geo-object
Research issues

• The goal of distributed geospatial web services is to provide a geospatial knowledge system that can provide answers to many different geospatial questions automatically.

• Four key issues for geospatial web services
  – How to automatically decompose user’s query (user object) to construct the geo-tree based on distributed data and service catalogs?
  – How to represent the geo-trees in computer understandable and executable workflows?
  – How to manage, share, and reuse geo-trees that represent the geospatial knowledge of a specific domain?
  – How to execute the geo-tree at the distributed, web service environment automatically to derive the product that exactly meets the user’s query.

• This first one requires geospatial domain knowledge while all other three are common in the broad web service community.
  – Ontology-based, rule-constraint machine reasoning may be the solution for the issues.
  – AI community is working on this issue. And we will adopt geospatial ontologies to test the AI solution.

• The other three issues are also concerned by W3C and OGC. And we will follow closely the progresses made by them.