A Semantic Enhanced Search for Spatial Web Portals

Wenwen Li[1], Chaowei Yang[1], Rob Raskin[2]
[1] Joint Center for Intelligent Spatial Computing, College of Science, George Mason University
4400 University Dr, Fairfax, VA, U.S. 22030
{wli6,cyang3}@gmu.edu
[2] Jet Propulsion Laboratory, 4800 Oak Grove Dr
Pasadena, CA, U.S. 91109
raskin@jpl.nasa.gov

Abstract

This paper reports our research on utilizing semantic model to improve the searching function within Spatial Web Portals (SWPs). Based on SWEET, We build the domain ontology and implement a semantic inference service. Multiple data resources are bridged to provide cross catalog searches. We expect this research to support spatial search in an intelligent manner.

Keywords

Spatial Web Portals (SWP), Ontology, Intelligent Query, Semantic Search

Introduction

Spatial Web Portal (SWP, Yang et al 2007) is used by the geospatial community to share geospatial data and information including text files, raw and post-processed data, and various geospatial web services. However, the popularity of SWP also brought up problems on how to discover needed data from a variety of geospatial resources and how to visualize the data for multi-purposes. Currently, most search engines inside SWPs are based on keyword matching, which can not effectively 'understand' the meaning of user's queries, especially when a user has limited geospatial knowledge. For example, after California fire, people may ask "What's the air pollution caused by Southern California Fire in 2007?" (Q1), but usually they do not know what may cause air pollution, or what are used in Air Quality (AQ) community to measure the air pollution. The above domain knowledge is hidden behind the query and can not be inferred directly by keyword based search engines. Thus how to provide a 'specialized' answer through 'unspecialized' queries becomes an urgent challenge, which is also known as "Intelligent Question Answering" problem.

To solve the problem, this paper reports our efforts on utilizing semantic web techniques (Berners-Lee, Hendler and Lassila, 2001) to enhance traditional search engines by: 1) building a semantic-based information model, which is a network of concepts with explicit relationships for implementing knowledge reasoning. With this model, a generic query can be explained and inferred to specific information needed. 2) Combining search tools provided within and cross SWPs to integrate heterogeneous data resources, results will include both web pages for presenting text information and spatial maps generated from remote web services for interactive searches.

The Semantic Model

Our model is based on SWEET (Semantic Web for Earth and Environmental Terminology) ontology, where all the concepts are divided into different facets, which include phenomena, property, substance, earth realm, to support reductionism (Raskin and Pan, 2005). As one of the most popular ontology model in Earth Science, SWEET provides an upper-level abstracted expression of the world. Based on formal description logic (DL), it can support terminology reasoning (T-box reasoning). However, only using this model may not be enough for answering query like Q1. First, "Southern California Fire in 2007" is an instance of "wildfire", which can not be inferred if there's no instance stored in the ontology. Second, formal query method is not able to retrieve the real resources (such as reports, documents or other data) even they've reached related node in the ontology model because this task is always partitioned as information retrieval (IR) task. Third, besides information provided in retrieved web pages, people may need more information, such as real-time data or geospatial maps which have sufficient resources in SWP. We build an extended model based on SWEET combining both semantic web and IR techniques to better understand and answer users' queries as follows:

(1) Extracting information from queries and using them as input of our semantic model. Here we do not focus on converting the nature language query to explicit DL expressions, but to extract keywords and do a basic analysis of them. For example, we collect the keywords by providing a structured template for question description,

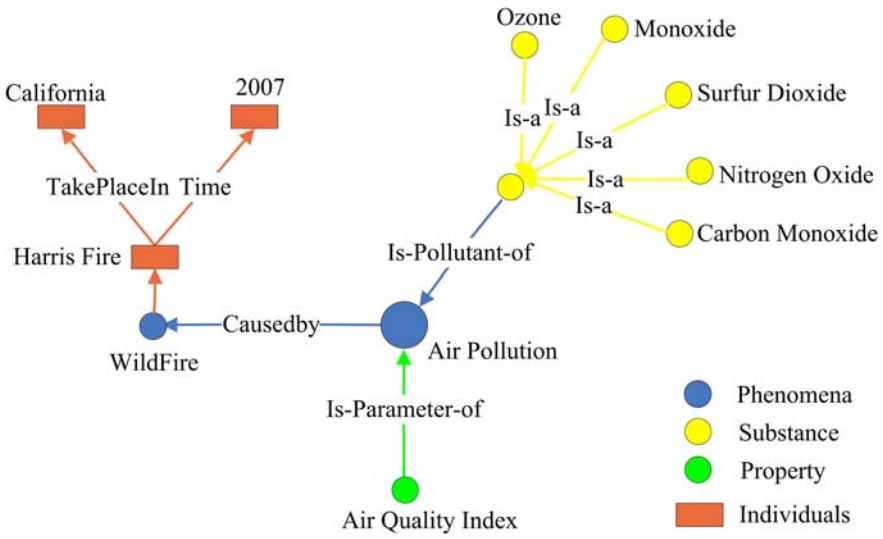

Figure 1. *A graph of ontology fragment pertaining air pollution.*

with which the occurrence of event (such as "Air Pollution"), reason (caused by Fire), place (such as California), and time (2007) could be measured and extracted. In this way, without natural language processing, the system is able to convert a spatial query (Q1) to a logic description Q1*.

Q1*: Air Pollution $\cap\exists$ causedBy.Fire $\cap\forall$ takePlacein.CA $\forall$ Happenedin.2007

(2) Extending SWEET ontology by adding more individuals and roles to support instance reasoning (A-box) to expand user's query. Figure 1 shows an example of the ontology fragment. It's visualized as a labeled graph: nodes in circles denote terminologies and those in rectangles denote instances. Among which, the nodes of the same color means they're in the same facet of the ontology model. Different nodes are connected by certain roles labeled on the arrow. In this paper, we'll take the fragment of ontology shown in Figure1 as a case study instead of talking the entire A-box and T-box of our ontology model. Air quality related terminologies including ("Air Pollution", "WildFire"), pollutant (such as "Ozone", "Carbon Monoxide"), air pollution parameter (such as "Air Quality Index") are defined to support T-box reasoning; while related instances (as events), such as "Harris Fire", "2007", "California" are also defined to support A-box reasoning. By this method, query Q1 can be expanded to DL concept Q1a, Q1b and Q1c by formal query method (Horrocks and Tessaris, 2002):

Q1a: Fire. Name $\cap\exists$ cause.AirPollution $\cap\forall$ takePlacein.CA $\forall$ Happenin.2007

Q1b:

Pollutant $\cap\forall$ isPollutantOf.(AirPollution $\cap\forall$ causedBy.Fire)

Q1c:

Parameter $\cap\exists$ isParameterOf.AirPollution

Where, Q1a aims to find the names<n1, n2…nk> of all the fires that satisfy the restrictions. This is a complex query that will deal with both T-box and A-box reasoning: 1) from the structured query template, the keywords "fire" and "air pollution" can be extracted and we know their relationship as reason (fire) and result (air pollution). 2) the inference is started by event (air pollution) to locate the term in the ontology, then 3) based on the equal relationship (CauseBy) and the known reason "fire", the inference find the term "wildfire" that satisfies this relationship. 4), from Q1a, we know that the objective is to find the specific fire name, so A-box reasoning starts by checking all the instances that wildfire owns to query which one happened in 2007 and took place in California, then "Harris Fire" can be inferred and returned from Q1a.

Q1b and Q1c could be inferred by checking all the roles that are connected with 'AirPollution' to get the pollutants<Po_1, Po_2…, Po_m> and parameter< $Para_1$, $Para_2$…, $Para_n$ >, which are all connected to Term "Air Pollution" directly in our ontology model. The two sub-queries are similar and easier than Q1a because they only deal with T-box reasoning. By those queries,("Ozone", "Monoxide", "Surfur Dioxide", "Nitrogen Oxide", "Carbon Monoxide") can be inferred from Q1b, and "Air Quality Index" can be inferred from Q1c (reference Figure 1).

(3) Coupling keyword groups in an appropriate manner. As discussed in step (2), all the three sub-queries are focusing on different aspects of inferences and each query can return a set of keywords. So after coupling them, such as <"Harris Fire", "Ozone", "Air Quality Index"> or <"Harris Fire", "Carbon Oxide", "Air Quality Index">,

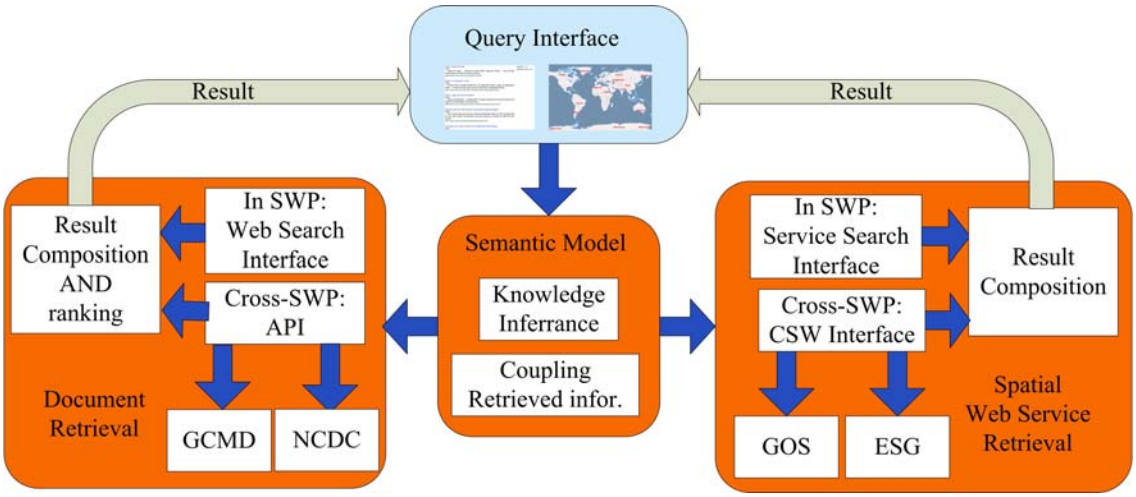

Figure 2. *Integral Infrastructure for knowledge discovering*.

and input the new groups of keywords into keyword-based data sources, the chance of getting more accurate information is increased.

(4) Dispatching the coupled keywords into search interface. Usually, there are two types of search tools provided within the SWP: one is IR based search tools for searching web documents. Considering the characteristic of spatial issues, most of the SWPs provides spatial web service search engine, such as those in ESIP's Earth Information Exchange Portal (EIE) (Yang et al., 2007) and NASA's Earth Science Gateway (ESG) (Bambacus and Evans, 2005). Thus related information found through them can be visualized into two types: text and generated geospatial map (Figure 2 "in SWP" box).

Integrating with Heterogeneous Resources cross SWPs.

Resources from single SWP may not have sufficient information for certain application. We try to build connections cross multi-SWPs to extend searching scope. As Figure 2 shows, after reasoning from semantic model (middle part), the system will redirect the keywords to GCMD and NCDC's web portal to retrieve more related web documents through the Application Programming Interfaces (APIs) (left part). Meanwhile, for discovering more geospatial map, we rely on OGC's Web Catalog Service (CSW) provided by Geospatial One Stop (GOS) and ESG for searching related layers of geospatial web services based on the combined keywords. For CSW search, we can not send the keywords directly to the interface the same as web document search, but the keywords should be encapsulated as filters in a XML-based or KVP (Key Value Pairs)-based query according to OGC CSW specification (Nebert and Whiteside, 2005). Returned map layers together with those got from local SWP will be integrated into a single result set and visualized in the query interface (right part). In this way, geospatial resources located in distributed environment could be seamlessly integrated and serve as a solid information base to support geospatial information discovery.

Conclusion and Future Work

The research addresses intelligent question answering in geospatial sciences. A semantic enabled model extending the search capabilities of existing methods in SWPs is able to answer more complex queries. The latter part presents how to build connections cross popular SWPs to integrate and interoperate information seamlessly.

Knowledge discovering system for Air Quality is being researched to support effective decision making. In the future, we'll try to improve intelligent reasoning and develop systems that could support more geospatial information resources.

Acknowledgments

Research reported is supported by NASA grant NNX07AD99G.

References

Berners-Lee, T.; Hendler, J. and Lassila, O. 2001. The Semantic Web. *Scientific American* 284, 34–43.

Bambacus, M.J. and Evans J. 2005. NASA's Earth-Sun System Gateway: An Open Standards-based Portal to Geospatial Data and Services. In proceedings of IEEE International Symposium of Geoscience and Remote Sensing, 4228- 4231.

Horrocks I. and S. Tessaris, 2002. Querying the semantic web: A formal approach. In Proceedings of International

Semantic Web Conference, 177-191, 2002.

Nebert, D., and Whiteside, A. 2005. Catalog Services, Version 2, OGC Implementation Specification. Available online at: http://portal.opengeospatial.org/files/?artifact_id=20555 , OGC (accessed 1 November, 2007).

Raskin R. and Pan M. Knowledge Representation in the Semantic Web for Earth and Environmental Terminology (SWEET), *Computer and Geosciences*, 31(9): 1119-1125, 2005.

Yang C.; Evans J.; Cole M.; Alameh N.; Marley S. and Bambacus M., 2007. The Emerging Concepts and Applications of the Spatial Web Portal, *Photogrammetry Engineering &Remote Sensing*, 73(6):691-698.

Yang, C.; LI, W.; Xie J.; Zhou B.; 2007. Distributed Geospatial Information Processing - Sharing Distributed Geospatial Resources to Support the Digital Earth, *International Journal of Digital Earth*. (in press).