Toward the Semantic Deep Web

James Geller, New Jersey Institute of Technology
Soon Ae Chun, College of Staten Island, City University of New York
Yoo Jung An, Fairleigh Dickinson University

The Semantic Deep Web fuses aspects of the Semantic Web with the use of ontology-aware browsers to extract information from the Deep Web.

The World Wide Web is arguably the greatest technological success in history. Starting from zero in 1990, it grew to 16 million pages by the end of 1995. In 2008, more than 1.4 billion webpages are accessible to anybody with an Internet connection and a computer.

With the explosion of the Web, indexing and searching the content of all Web documents has become a perpetual challenge, in spite of the continuous technological advancements of Web search engines. However, this challenge is even greater when considering the Web data *not* accessible by search engines.

Many organizations generate back-end data that is dynamically retrieved through Web-form-based interfaces and thus not indexed by conventional search engines. This hidden, invisible, and nonindexable content is called the Deep Web, and its size is estimated to be tens of thousands of times larger than the surface Web.

The term "Before Web" underscores the momentous transformation in information-access behavior occasioned by the Web. Today, people commonly rely on the Web to find various products and services, telephone numbers and addresses, map locations, flight information, movie showtimes, jobs, and even dates. However, the Web is inherently unintelligent and using it requires human interaction and intellect.

Web users must guess which search terms will lead to the satisfaction of their information need. They must evaluate the search results and decide which of the snippets presented by the search engine is most promising. If the search returns no hits, or if no promising snippets appear among the first few pages of hits, users must refine the search terms. Users continue executing this "loop" until it satisfies their information need or they determine that no existing document meets their expectations.

THE SEMANTIC WEB

Originally proposed in 2001 by Tim Berners-Lee, James Hendler, and Ora Lassila, the Semantic Web attempts to overcome this problem by extending some of the intelligent behavior currently performed by humans to the Web, making searching easier and more productive.

A major goal of the Semantic Web is to facilitate the automation of e-business processes and services: Software agents (softbots) with rich semantic knowledge and reasoning capabilities automatically roam the Web, find data and services, and combine them to achieve business goals.

However, a large amount of information on the Web, especially valuable e-commerce data, is invisible to search engines, Web crawlers, and agent programs and only accessible through Web services or Web-form interfaces.

The need for large ontologies—representations of human-like knowledge in computational graph sctructures—in a standard format such as the Resource Description Framework (RDF) and the Web Ontology Language (OWL) has likewise hampered the Semantic Web's progress. Ontology development requires domain-specific knowledge, often lodged in some expert's mind or in textual documents. Capturing and extracting this knowledge and representing it in a computer-processable format remain difficult tasks.

Building software agents that can reason and make inferences, using semantics provided by ontologies, is also challenging.

THE SEMANTIC DEEP WEB

To address the problems associated with accessing rich, structured back-end data as well as ontology construction and use, we propose the Semantic Deep Web. As the name suggests, the Semantic Deep Web consists of elements from both the Deep Web and the Semantic Web, especially the hidden back-end data sources, the interface or Deep Web services that access these data sources, and programs to manipulate ontologies.

It's important not to confuse the Semantic Deep Web with the Deep Semantic Web, which is part of the Semantic Web's original vision. Whereas the Deep Semantic Web refers to the more complex and AI-oriented levels in the so-called Semantic Web layer cake, the Semantic Deep Web fuses aspects of the Semantic Web with the use of ontology-aware browsers to extract information from the Deep Web.

The primary goals of the Semantic Deep Web are to access Deep Web data through various Web technologies and to realize the Semantic Web's vision by enriching ontologies using this data. Its research areas include

- information extraction from the Deep Web, especially e-commerce sites;
- semantic annotation and indexing of the Deep Web;
- Deep Web schema understanding based on data semantics;
- Semantic Deep Web search engines;
- Semantic Deep Web data fusion and interoperation;
- semantic browsing and visualization of the Deep Web;
- semiautomatic ontology generation from the Deep Web;
- quality of ontology measurements; and
- quality of Semantic Deep Web search and ranking measurements.

The Semantic Deep Web uses two main approaches to access the Deep Web using Semantic Web technologies. Both require a Semantic Deep Web crawler.

The first approach, which we refer to as *ontology plug-in search*, is to enrich a domain ontology with Deep Web data semantics so that it can be used to refine user search queries processed by a conventional search engine such as Google. The other approach, called *Deep Web service annotation*, is to annotate Deep Web services (Deep Web form sites) with Deep Web data semantics; these semantically annotated documents are searchable using a Semantic Web search engine such as Swoogle (http://swoogle.umbc.edu).

It's semantically easier to derive ontologies from Deep Web data sources than from unconstrained natural-language text documents.

ONTOLOGY PLUG-IN SEARCH

Ontologies are central building blocks of the Semantic Web, allowing reasoning capabilities and automation. They help softbots by serving as repositories of knowledge about synonyms, homonyms, general versus specific concepts, concept properties, semantic relationships between concepts, and in many cases axioms or rules associated with concepts. For example, if a user is looking for a car, a search engine that recognizes "automobile" as a synonym for "car" will report back more relevant webpages.

Unfortunately, building large ontologies is labor-intensive and fiendishly difficult. Researchers have made some progress in automatically generating ontologies using algorithms that mine concepts and their relationships from text documents, but detecting and reliably extracting concepts and their relationships remains an AI-complete problem comparable in difficulty to the natural-language-processing problem.

It's semantically easier to derive ontologies from Deep Web data sources, especially well-structured relational back-end databases, than from unconstrained natural-language text documents.

For example, a database table with a column called CityNames will contain a homogeneous set of terms with similar meanings such as New York, Los Angeles, London, Paris, and Rome. While the column name itself is typically not accessible, the data in the table might be found by repeated probing queries, and the label next to the input field in a Web form ("departure city") is often a good proxy for the column name.

Deep Web-based ontology construction consists of first harvesting concepts—that is, Deep Web form attributes—from many websites and then iteratively interweaving these concepts into the IS-A skeleton (hyponym structure) of the well-known WordNet (http://wordnet.princeton.edu) ontology. The underlying assumption is that these form attributes reflect the schema of the underlying Deep Web data sources.

These two processes semiautomatically generate a domain ontology that is subsequently enriched with instances and semantic relationships between the instances extracted from Deep Web data.

The enriched domain ontology provides information for an ontology-enabled search engine, which informs users about domain-specific terms and relationships they might find useful when further specifying their information needs.

The ontology plug-in for the Deep Web search must identify a relevant domain ontology for user search terms, and this approach promises to return more relevant Deep Web search results than just using keyword-based search.

DEEP WEB SERVICE ANNOTATION

A key question is how to enrich Deep Web service descriptions to reflect the services' dynamic content. What quantities of dynamic back-end information ensure that a Deep Web service reliably represents the contents in the Deep Web?

The Semantic Deep Web crawler first samples the data sources *N* different times. It then constructs a Deep Web data signature with a multidimensional distribution summary of the data. This includes the thematic distribution and geospatial coverage of data using clustering and frequency distribution analyses.

This semantic data signature is represented in a semantic annotation document in RDF, OWL-S, the Web Service Modeling Language (WSML), or Semantic Annotations for WSDL and XML Schema (SAWSDL).

A Semantic Web search engine like Swoogle can search such a representation and index it to locate Deep Web sites and services. In this way, Deep Web services become searchable not just by input/output parameters but by their thematic (semantic) content and other contextual information residing in Deep Web data sources.

A Semantic Deep Web search engine should consider semantic as well as string matches. It can compute semantic similarity using semantic equivalence, overlap, generality, and specificity operators derived from more primitive operators such as "equalTo," "instanceOf," "parentOf," "siblingTo," "childOf," and "synonymOf." For example, the search engine could return hits for terms that match linguistically or structurally, or are synonymous.

The semantic matching process should also utilize the frequency distributions of search terms or similar terms, contexts, and relationships in the semantic annotation files.

The search engine can rank the relevance of Deep Web search results based on thematic and spatiotemporal distributions of semantically similar content. If the query is mostly about Deep Web content, it can assign a higher weight to content-related semantic annotation matching. If the query is mostly about the service descriptors—for example, quality of services or provider types—it can give more weight to service descriptor-based matching.

REALIZING THE VISION

To realize the Semantic Deep Web vision, researchers must integrate Semantic Web and Deep Web technologies. Activities include

- developing ontology-aware, high-quality Web search engines;
- building large ontologies from Deep Web sites, starting with all e-commerce subdomains;
- gaining acceptance of an "open source attitude" in the e-commerce realm to make building Deep Web ontologies easier by accessing currently securely locked data sources;
- creating libraries of semantic crawlers for the purpose of extracting back-end database information; and
- building comprehensive index structures for Deep Web sites.

Building research communities to tackle the challenges inherent in the overlap of the Semantic Web and the Deep Web is also desirable. Toward that end, a first-of-its-kind international workshop titled "The Semantic Web Meets the Deep Web" was held in July 2008 in Washington, DC, in conjunction with the joint 10th IEEE Conference on E-Commerce Technology and the 5th IEEE Conference on Enterprise Computing, E-Commerce, and E-Services. Details are available at www.cis.njit.edu/~oohvr/SemanticDeepWebWS.

Whether the proposed Semantic Deep Web would ultimately combine the beneficial aspects of the Semantic Web and the Deep Web, or would only inherit the challenges from both worlds, remains an open question. Both research community support and industry buy-in are required to make the Semantic Deep Web a success. ■

James Geller is a professor in the Department of Computer Science at the New Jersey Institute of Technology, where he also directs the Semantic Web and Ontologies Lab. Contact him at geller@njit.edu.

Soon Ae Chun is an assistant professor of information systems in the Department of Business at the College of Staten Island, City University of New York. Contact her at chun@mail.csi.cuny.edu.

Yoo Jung An is a visiting professor in the Gildart Haase School of Computer Sciences and Engineering at Fairleigh Dickinson University. Contact her at yoojungan@gmail.com.

Editor: Richard G. Mathieu, Dept. of Computer Information Systems and Management Science, College of Business, James Madison Univ., Harrisonburg, VA; mathierg@jmu.edu