The Emergent Semantic Web: Managing Data Efficiently

Mahesh Raisinghani  
*University of Dallas*

Tapas Sahoo  
*University of Dallas*

Follow this and additional works at: http://aisel.aisnet.org/amcis2004

Recommended Citation  
http://aisel.aisnet.org/amcis2004/3

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in AMCIS 2004 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
The Emergent Semantic Web: Managing Data Efficiently

Mahesh S. Raisinghani  
University of Dallas, Graduate School of Management  
mraising@gsm.udallas.edu

Tapas Sahoo  
University of Dallas, Graduate School of Management  
tapas_s@hotmail.com

ABSTRACT
The Semantic Web is a vision that will extend the current Web to give well-defined meaning to information, enabling computers and people to work in better cooperation. This is a collaborative effort between the World Wide Web Consortium (W3C) Semantic Web Activity and a large number of researchers and industrial partners. This effort will enable data to be used for effective discovery, automation, integration, and reuse across applications. Ontologies are being developed to make computers knowledgeable about data. However, one of the major challenges is to build a common Ontology for the entire Web. This paper discusses the purpose, present state and future challenges for the Semantic Web revolution that is emerging. It addresses issues involved in effectively storing and managing data on the Web in order to achieve its potential as the largest repository of information.

Keywords
Semantic Web, URI, RDF, Ontology, Web Services, Inference Engines

INTRODUCTION
The rich presentation capability of the Web makes it highly readable for humans; but adds no meaning to the information, when read by computers. For example, a HTML page might present the resume of a person with his name in bold capital letters, but there is no way for the computer to interpret the “name” as a name. Also, there may be several such HTML pages containing various types of information about this person, but the computer cannot link them together to summarize the information. It still needs human intervention to interpret, filter, and assemble the relevant information.

As the next evolution of the current Web, the “Semantic Web”, would contain information that has well-defined meaning. This added meaning to data, called “metadata”, would enable computers and people to work in cooperation (Hendler et al., 2002). We first present an overview of the architecture of the Semantic web and then discuss issues about managing data. Finally, we conclude with various challenges ahead and the next practices in the development of Semantic Web.

OVERVIEW OF THE ARCHITECTURE
Figure-1 displays the layered architecture of the Semantic Web. Its major components are described in the following section.
Uniform Resource Identifiers (URIs)

Each item on the Web is considered as a resource and URIs are used to uniquely identify each resource (Swartz, 2002). URIs can be assigned to real world objects like persons, places, books etc. The most common form of URI is the URL, which represents the address of a unique webpage on the Internet. However, the primary function of a URI is to identify a resource, not to tell how to get a specific file on the Web.

Resource Description Framework (RDF)

RDF is the standard followed by W3C to process metadata on the Semantic Web (Brickley, & Guha, 2000). RDF is a framework to create statements about resources in a machine-readable format and is based on the idea of identifying things using URIs, and describing resources in terms of simple properties and property values. A RDF statement has three parts: a subject, a predicate and an object, each represented by its URI (http://www.w3.org/TR/2004/REC-rdf-primer-20040210/#basicconcepts).

Ontologies

An Ontology enumerates and gives semantic descriptions of concepts, defining domain-relevant attributes and various relationships among them. For example, an ontology describing wines will include concepts like vintages, wine regions, wineries, and grape varieties. It will also include relations such as produced by, made from, color, year, and body of wine (Noy, 2003).

Logic

One of the biggest challenges in making the computers more intelligent is to enable them to apply logical principles and draw conclusions by inference. This is one of the key research areas in the field of Artificial Intelligence.
Trust
The freedom of expression on the Web can lead to its misuse and result in mistrust. If the Web has to become the single source of all information, it has to provide a mechanism of proving trustworthiness. Digital signature is the mechanism that is going to be used to provide proof that a certain person wrote (or agrees with) a document or statement.

Inference Engines
Inference Engines will process the available knowledge and most importantly, will deduce new knowledge from the knowledge already specified to them. They will apply reasoning over the information that is coming in from all over the Web.

Web Services
The Web is evolving into a provider of automated services, in different forms such as flight-booking programs, and a variety of e-commerce and business-to-business applications. Intelligent software agents are being able to automatically locate and execute various Web services available on the Web.

MANAGING DATA ON THE SEMANTIC WEB
Even though the Web uses many databases at the back end, the structure and semantics of the data is lost when it is squeezed into the Web (Marchiori, 2001). Inside a database, any piece of information has a well-defined meaning, in terms of entities and relationships. However, once this information is published in a Web page, both the meaning of the data and relationship between them is lost. Only a human with the domain knowledge can infer that relationship. To solve this problem, it is important to design the Semantic Web as an efficient and scalable data store, and try to minimize the loss of metadata along the passage to and from the Web.

The Web also needs the ability to easily retrieve data stored in different pages. The gap between the databases and Web pages should be minimized or eliminated to achieve this goal. The Web can achieve its potential of a distributed database if collection of Web pages become databases themselves.

A MODEL FOR SEARCHING DYNAMIC PAGES
The search engines build content dynamically based on user queries. These pages send the queries to the back end database and retrieve the information requested by users. Even though the dynamic pages do not contain any data, they do know what type of data to retrieve and contain the query that gets executed. In case of relational databases, this query is written using the SQL.

An intelligent agent can make use of this information and even execute the query to retrieve the information. Figure 2 illustrates an example of how agents can be used to extract the information from two dynamic Web pages and merge them together to satisfy a user’s request.
The collection of dynamic Web pages, referred to as the “Deep Web” is estimated to be approximately 500 times the size of the “Surface Web” (Bergman, 2001). As most of the existing search engines fail to index this vast portion of the Web, there arises the need for intelligent tools that can extract and integrate the results from dynamic Web pages. The key prerequisite for this kind of information integration is obtaining the schema of these Web sites. One of the recent research works in this area (Wang and Lochovsky, 2003) describes the “DeLa” (Data Extraction and Label Assignment) system that can capture the schema and extract data from dynamic Web pages with complex HTML search forms. Query languages, in general, fetch a set of entities joined by one or more relationships specified by the users. However, given a set of entities, the query languages fail to identify the relationship between them. Anyanwu, & Sheth (2003) have presented the notion of semantic associations to potentially eliminate this limitation.

Aberer et al. (2003) describe a novel approach for sharing metadata among various data sources with local schemas in a bottom-up, semiautomatic manner without relying on pre-existing, global semantic models. In this approach, participating data sources incrementally exchange schema translation between each other and eventually building the semantics that all participants understand.

**CHALLENGES AND NEXT PRACTICES**

Realizing the vision of a single ontology is far from reality due to differences in cultural, organizational, or administrative behavior. It is more likely that there will be more than one ontology even for the same domain on the Semantic Web (Noy, 2003). There are two significant areas of concern in the development of ontologies. The first one is the huge difference between the RDF framework and majority of the existing data sources. While RDF is focused on identifying the domain structure, most of today’s data sources tend to focus more on the document structure around important objects. The second challenge is to deal with numerous ontologies and schemas that are likely to emerge in any particular domain. The Semantic Web needs a flexible architecture and efficient algorithms that could mediate between different but related ontologies and schemas.

Many research organizations are currently working to develop systems that address the aforementioned problems. One such system, called “Piazza” (Halevy et. al., 2003) provides algorithms to map the domain and document structures, as well as enabling interoperability of XML data with RDF data. With the likely proliferation of ontologies, there is an eminent need for a centralized ontology registry that should maintain the list of available ontologies in different domains. Maedche et al. (2003) have provided an integrated approach for establishing such an infrastructure for searching, reusing, and evolving distributed ontologies.

The vision of Semantic web heavily depends on meta-data. Ongoing research activities are focused on two approaches to define meta-data. The first approach is to create ontologies that will store metadata about entities and provide them when...
needed. The second approach is to annotate web pages with semantic tags. Dill et al. (2003) have proposed an application called “SemTag” that performs automatic semantic tagging. They have also developed a platform called “Seeker” that can be shared by different tagging applications.

Most of the approaches for creating metadata from existing information have focused on static Web pages, but not dynamic pages. One possible approach is to annotate the information inside the database rather than annotating the Web pages. Handschuh et al. (2003) have proposed a framework called “Deep Annotation”, to provide semantic annotation for large sets of data. This process leaves semantic data inside database systems where it can be handled best.

This study can be developed to address the following research questions:

What are the structures, architectures, models, methods, languages and tools for usable description of and access to the contents of Web documents and their interrelationships? and

How can query and transaction management, resource identification and search methods, agent technologies, and harmonization and analysis of diverse contents be optimally incorporated?

CONCLUSION

The Semantic Web will eliminate many of the drawbacks of the current Web by making it understandable by machines. This can be achieved by adding meaning to data by using frameworks such as URIs, RDF, and ontologies. The Semantic Web has the potential of becoming a huge distributed database. With the development of the right infrastructure, the main drawback of the current search engines to retrieve dynamic Web pages can be eliminated.

The Semantic Web provides an infrastructure that would enable different consumers and producers of data to interoperate in an automated fashion. For example, databases, services, programs, and digital devices can consume and produce data on the Web. Software agents can use this information to search, filter and prepare information in new and exciting ways to assist the web user. Semantic Web has a long way to go, but a quite bit of progress have already been made. There are endless possibilities, and the journey itself is worth the effort even if all the goals are not achieved.

REFERENCES


