Semantic Web Fundamentals

Grigoris Antoniou
Institute of Computer Science, FORTH, Greece

Vassilis Christophides
Institute of Computer Science, FORTH, Greece

Dimitris Plexousakis
Institute of Computer Science, FORTH, Greece

Martin Doerr
Institute of Computer Science, FORTH, Greece

INTRODUCTION

World Wide Web: A Critical View

The World Wide Web (Berners-Lee, Cailliau, & Groff, 1992; Berners-Lee, 1999) has changed the way people communicate with each other and the way business is conducted. It lies at the heart of a revolution that is currently transforming the developed world toward a knowledge economy (Neef, 1997), and more broadly speaking, to a knowledge society.

Most of today's Web content is suitable for human consumption. Even Web content that is generated automatically from databases is usually presented without the original structural information found in databases. Typical uses of the Web today involve humans seeking and consuming information, searching and getting in touch with other humans, and reviewing the catalogs of online stores and ordering products by filling out forms.

These activities are not particularly well supported by software tools. Apart from the existence of links that establish connections between documents, the main valuable, indeed indispensable, tools are search engines. Keyword-based search engines, such as AltaVista (www.altavista.com), Yahoo (www.yahoo.com), and Google (www.google.com; Page & Brin, 1998), are the main tool for using today's Web. It is clear that the Web would not have been the huge success it was, were it not for search engines. However, there are serious problems associated with their use; the most important ones are listed in Table 1.

BACKGROUND

The Semantic Web Vision

The Semantic Web (Berners-Lee, Hendler, & Lassila, 2001; Davis, Fensel, & van Harmelen, 2002; Fensel et al., 2002; Antoniou & van Harmelen, 2004) proposes to overcome the difficulties listed above by making Web content machine processable. The key point is that the semantics (meaning) of Web content must be explicitly represented and processed. This aim will be achieved by combining the following technologies:

Table 1. A list of problems associated with keyword-based search engines

High recall, low precision Too many, mostly irrelevant pages are retrieved
Low or no recall Key relevant pages are not retrieved
Sensitivity to chosen vocabulary Slight changes in vocabulary may cause significant changes in results
Results are single Web pages Information may be spread across various pages
Human involvement is necessary To interpret retrieved pages, and to combine information

- *Explicit Metadata*: Web content will carry its meaning "on its sleeve" through appropriate semantic markup.
- *Ontologies*: They will describe semantic relationships between terms and will serve as the foundation for establishing shared understanding between applications.
- *Logical Reasoning*: Automated reasoning-enabled tools will make use of the information provided by metadata and ontologies.

More on ontologies and reasoning are found in the following sections. As a simple example, suppose that you are searching for photos of an orange ape in an annotated online collection of digital photos. Suppose that picture 1 is annotated as "playing orangutan." Then this picture can be retrieved on the Semantic Web, although its annotation does not contain the words "orange" or "ape" (so, a keyword-based search would fail). This can be achieved through interplay of (a) the annotation, (b) information contained in an ontology about animals that states that orangutans are apes and are orange, and (c) reasoning that combines the above information to conclude that the picture is relevant to the user's query.

The Semantic Web vision was created by Tim Berners-Lee, the person who created the WWW. The Semantic Web activities are coordinated by the World Wide Web Consortium (http://www.w3.org/2001/sw/). Table 2 collects a few critical issues of Web technologies addressed by the Semantic Web initiative.

FUTURE TRENDS

Ontologies

An ontology is a formalization of a shared conceptualization of a particular domain. It supports interoperability between applications at the semantic level (the meaning of information) and also allows people to talk about objects of common interest. Typically, it describes the objects of the domain and the relationships that hold between them.

Objects of the same kind are organized in so-called classes that are collections of objects sharing certain characteristics. For example, in a university domain, classes may be professors, students, administrative and technical staff, courses, lecture theaters, etc. Individual objects (called *resources* in Web terminology) can be declared to be instances of a certain class.

Once we have defined classes, we can also define relationships between them. One particular kind of relationship is a class hierarchy. A class A is a subclass of a class B (and B is a superclass of A) if every instance of A is also an instance of B. For example, the class of professors is a subclass of the class of all university employees. Figure 1 shows a sample class hierarchy for the university domain.

The subclass relationship is a general-purpose relationship between classes. Further relationships, called *properties*, can be defined by the user. In the university domain, such properties might be that a course is taught by a particular professor, that a student takes a course, and that a professor is head of a department.

The interplay of classes and properties opens interesting modeling possibilities. For one, it is possible to define domain and range restrictions. For example, one can specify that a course can only be taught by an academic staff member (*range restriction*) and that only a course, and not, say, a lecture theater, can be taught (*domain restriction*).

Moreover, the valuable concept of *inheritance* can be utilized. Suppose that person *X* is declared to be an associate professor. Then *X* is allowed to teach a course, because he inherits this possibility from its superclass academic staff member. This way, we can avoid adding superfluous information to the ontology (*X* is an academic staff member; *X* is a staff member).

Table 2. Critial issues of Web technologies (Khosrow-Pour, 2004) addressed by the Semantic Web

Cyberloafing Surfing the Internet, wasting time, and accessing inappropriate materials
Flooding of the Web with content Including information that is not helpful
Having inadequate search facilities on the WWW Lack of high-level query language search engines for locating, filtering, and presenting information
Maintaining integrity of data Maintaining up-to-date and accurate information on the site for viewers to use