Chapter XII

Using Semantic Web Concepts to Retrieve Specific Domain Information from the Web

Rafael Cunha Cardoso, Federal University of Pernambuco and Tools & Technologies (HP-Brazil), Brazil

Fernando da Fonseca de Souza, Federal University of Pernambuco, Brazil

Ana Carolina Salgado, Federal University of Pernambuco, Brazil

Abstract

Currently, systems dedicated to information retrieval/extraction perform an important role on fetching relevant and qualified information from the World Wide Web (WWW). The Semantic Web can be described as the Web’s future once it introduces a set of new concepts and tools. For instance, ontology is used to insert knowledge into contents of the current WWW to give meaning to such contents. This allows software agents to better understand the Web’s content meaning so that such agents can execute more complex and useful tasks to users. This work introduces an architecture that uses some Semantic Web concepts allied to regular expressions (REGEX) in order to develop a system that retrieves/extracts specific domain information from the Web. A prototype, based on such architecture, was developed to find information about offers announced on supermarket Web sites.

Copyright © 2008, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
Introduction

The WWW emerged by the end of the eighties, a time where the impact of this technology on society was unknown (Berners-Lee et al., 1994). The increasing use of the Web has caused a permanent growth in the amount of data available on it. Its specific characteristics have generated a crescent demand for tools specialized in performing efficient management and qualified data retrieval/extraction from Web contents (Baeza-Yates & Ribeiro-Neto, 1999). Efforts to overcome the obstacles created by this ample growth, associated with the desire of inserting some level of intelligence to the retrieval of documents disposed on the Web, have motivated the development of the new generation of the Web: the Semantic Web (Berners-Lee et al., 2001).

It can be thought of as an extension of the current Web where data gains its own meaning. The main objective of this new Web paradigm is to insert some level of knowledge into WWW resources so that software agents can be able to intelligently process Web contents (Hendler, 2001). The Semantic Web is a mesh of information linked up in such a way that can be easily processed by machines on a global scale (Palmer, 2001).

This chapter focuses on some Semantic Web concepts and technologies, particularly ontology and its languages such as DAML+OIL (DAML+OIL 2001) and OWL (OWL, 2004). An architecture that makes use of these concepts to build systems that are capable of retrieving specific domain information from the Web is introduced. A prototype that actually implements the ideas behind the architecture is also presented.

The main objective of such a prototype is to provide a mechanism that efficiently identifies and extracts relevant pieces of information from a set of data from specific knowledge domain. Other technologies are also presented and discussed through the subsections of the chapter to achieve this goal.

Information Description and Retrieval

The rapid growth of the Web has promoted a radical change in the lives of people who have access to it. It is difficult to imagine a modern society living without all of the advantages the Web has brought to mankind. The easy access to the Web allied to its increasing usage has significantly raised the amount of data available on it.

One of the problems caused by such growth is the lack of some kind of description associated to resources spread on the Web. Meaningless contents that populate the current Web have increased several of its problems. Amongst them are: the delay in the information location and the retrieval of a high number of unexpected resources, due to ambiguity problems (Lopatenko, 2001).