ABSTRACT

At the present time, systems dedicated to information retrieval/extraction perform an important role 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, ontologies are being used to insert knowledge into contents of the current WWW, giving meaning to such contents. This allows software agents to better understand the Web’s content so that such agents execute more complex and useful tasks for 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 an architecture was developed in order to find information about offers announced on supermarket Web sites.

Keywords: information retrieval; ontologies; Semantic Web; Web-based applications; Web programming

INTRODUCTION

The WWW appeared at the end of the 1980s, a time when the impact of this technology on society was unknown (Berners-Lee, Cailliau, Luotenem, Nielsen, & Secret, 1994). The Web’s increasing use has caused a permanent growth in the amount of data available. The Web’s specific characteristics have generated a great need for tools specialized in executing efficient management and qualified data retrieval/extraction from the Web (Baeza-Yates & Ribeiro-Neto, 1999). Efforts to overcome the obstacles caused by this uncontrolled growth resulted in the development of the new generation of the Web, called Semantic Web (Berners-Lee, Hendler, & Lassila, 2001).

The Semantic Web is considered an extension of the current Web, where data gain their own meanings. The objective
of this new Web paradigm is to insert knowledge into WWW resources so that software agents can be able to understand and process Web contents in a clever way (Hendler, 2001).

This article introduces an architecture that uses Semantic Web concepts (e.g., ontologies, DAML+OIL) (Decker, Mitra, & Melnik, 2002; DAML+OIL, 2004) allied to known techniques (e.g., Regular Expressions) in order to retrieve and then extract information from Web contents.

In order to experiment with the Semantic Web concepts used in this work, a prototype based on the proposed architecture was developed. The main objective of such a prototype is to provide a mechanism that identifies and extracts relevant information from a set of data from a specific knowledge domain. The domain shall have its main concepts modeled and implemented as a Semantic Web Ontology.

This article is organized as follows. The second section is dedicated to the motivations of this work. The third section presents related works, and the fourth section introduces the system’s architecture. The fifth section focuses on the techniques that were used to develop the prototype, and the sixth section discusses the case study. Finally, the seventh section concludes the article.

**MOTIVATION**

The rapid growth of the Web has resulted in a radical change in the lives of people who have access to this communication medium. It is difficult to imagine modern society without all the advantages that the Web has introduced. Easy access to the Web has raised significantly the amount of data available on it.

One of the problems caused by the Web’s growth is that most of the resources available on it do not have any description associated with them. These meaningless contents that populate the current Web cause some problems, such as delay in the information location and retrieval of a high number of unexpected resources due to ambiguity problems (Lopatenko, 2001).

In this scenario, where unstructured and meaningless data predominate, tools specialized in collecting qualified information on the Web have become an enormous challenge to the scientific community. The effectiveness of these tools is linked directly to the way these resources were described before being published on the Web. Such tools are denominated as search tools and may be classified into two main classes:

**Research in Directories.** Systems introduced when the Web was small enough to be collected in a non-automatic way. Documents are classified manually according to a taxonomy.

**Search Engines.** Systems that valorize the database size instead of the information quality they retrieve. In this kind of system, resources are gathered by software agents that traverse the Web collecting relevant data.

These classes of software mechanisms both are guided by searches based on the meaning inlaid in the keywords provided by its users. Such search tools do not consider the semantic aspects involved in the keywords that were submitted to the search, analyzing just syntactic aspects. This results in non-trusted searches that do not completely fulfill the user’s expectations.

Another research area, complementary to Information Retrieval (Baeza-Yates & Ribeiro-Neto, 1999), is the Information Extraction area (Adams, 2004). Its main objective is to extract relevant information