Chapter 13
An Ontology-Based Extraction Framework for a Semantic Web Application

Hadrian Peter
University of the West Indies, Barbados

Charles Greenidge
University of the West Indies, Barbados

ABSTRACT

The Semantic Web vision is rapidly becoming a mainstream reality, but obstacles remain in the way. A major challenge is the adoption of practical Semantic Web applications and the production of vast stores of ubiquitous meta-data which is needed to allow robust inference engines to attain the goals of machine readability of web documents. The authors propose the Semantic Web Applications (SEMWAP) framework which facilitates semi-automatic matching of instance data from opaque web databases using lightweight ontology terms. This framework combines information retrieval, information extraction, natural language processing, and ontology techniques to produce a matching and thus provides a viable building block for Semantic Web applications. To experimentally investigate the characteristics and limitations of the SEMWAP framework, a prototype system called the Semantic Ontological Data Labeler (SODL) was constructed.

INTRODUCTION AND BACKGROUND

The hunt is on for the fabled "killer application" for the Semantic Web (SW) that is expected to remove the remaining hurdles that stand in the way of the full viability of the SW (Antoniou & van Harmelen, 2008). We propose a modest, yet robust, framework called SEMWAP (Semantic Web application) that may be used in constructing such applications, providing an essential building block for SW mavens, practitioners, and researchers.

In order to overcome the many technical challenges that remain before the SW can be adopted, the key problems in data retrieval (DR), information retrieval (IR), knowledge representa-

DOI: 10.4018/978-1-4666-1873-2.ch013

tion (KR) and information extraction (IE), must be addressed (Manning et al., 2008). The rise of the World Wide Web, with its vast data stores, has served to highlight the twin problems of information overload and search (Lee et al., 2008). To address these limitations smarter software is needed to sift through increasing Web data stores and the data itself must be adequately marked-up with expressive meta-data to assist the software agents. A major hindrance to the full adoption of the SW is that much data is in a semi-structured or unstructured form and lacking adequate meta-data (Etzioni et al., 2008). Without the existence of robust meta-data there is little opportunity for SW inferencing mechanisms to be deployed.

To overcome this hindrance new web tools are being developed, with SW technologies already integrated into them, which will facilitate the addition of the necessary mark-up (Shchekotykhin et al., 2007). Beyond this there is an Information Extraction issue that must be tackled so that older web data, or data currently managed by older tools, can be correctly identified, extracted, analysed, and ultimately semantically marked up. In traditional Artificial Intelligence (AI) (Russell & Norvig, 2003) much work has been done in the field of ontological engineering where there is an attempt to model concepts of the real world using precise mathematical formalisms (Sicilia, 2006).

Our paper focuses on mapping web data to domain ontologies, allowing several Information Extraction issues to be directly addressed. We use a variety of techniques to make sense of the structure and meaning of the Web data, ultimately providing a match to a domain ontology. In particular the WordNet lexical database (Euzenat & Shvaiko, 2007) is used to facilitate some basic matching activities. We also make use of current search engine capability in our ontology mapping process. Allowing search engine inputs helps us to align the matching process with data as it exists online, rather than as construed in some selectively crafted catalog which may not be representative of web data (Schoop et al., 2006).

The motivation for our ontology-based framework is a number of earlier approaches to information extraction on the Web. Hand-written wrappers are the first approach to IE on the Web but the well known limitations of such approaches are robustness and scalability hurdles (Shen et al., 2008). The literature on ontology-driven IE on the Web is rather sparse, however, there is a growing body of literature which grapples with ontology-based matching of data on the Web (Isaac et al., 2007).

Work done on Dutch collections show that simple Jaccard based measures (Euzenat & Shvaiko, 2007) can give adequate results. Hu and Qu (2007) studied simple matches between a relational schema and a Web Ontology Language (OWL)-based ontology, ultimately constructing the MARSON system for performing these mappings. Shchekotyhin (2007) argued about the general unsuitability of traditional Natural Language Processing (NLP), IE and clustering techniques for ontology instantiation from web pages, and proposes an ontology modeling system for the identification/extraction of instance data from tabular web pages.

To experimentally investigate the characteristics and limitations of the SEMWAP framework we constructed a prototype system called the Semantic Ontological Data Labeler (SODL). This prototype mirrors the phases as outlined by our algorithm for labeling instance data based on lexico-semantic matches between data and terms belonging to a lightweight ontology (Greenidge, 2009).

The rest of the paper is organized as follows. First we identify the related work that has influenced the framework, and examines the characteristics of the Semantic Web relevant to our framework. Next, we outline the important elements of our framework, including the detailed algorithm. A validation of the framework is provided including an explanation of the underlying