Chapter 3
Using an Ontology-Based Framework to Extract External Web Data for the Data Warehouse

Charles Greenidge
University of the West Indies, Barbados

Hadrian Peter
University of the West Indies, Barbados

ABSTRACT
Data warehouses have established themselves as necessary components of an effective Information Technology (IT) strategy for large businesses. In addition to utilizing operational databases data warehouses must also integrate increasing amounts of external data to assist in decision support. An important source of such external data is the Web. In an effort to ensure the availability and quality of Web data for the data warehouse we propose an intermediate data-staging layer called the Meta-Data Engine (M-DE). A major challenge, however, is the conversion of data originating in the Web, and brought in by robust search engines, to data in the data warehouse. The authors therefore also propose a framework, the Semantic Web Application (SEMWAP) framework, which facilitates semi-automatic matching of instance data from opaque web databases using ontology terms. Their framework combines Information Retrieval (IR), Information Extraction (IE), Natural Language Processing (NLP), and ontology techniques to produce a matching and thus provide a viable building block for Semantic Web (SW) Applications.

DOI: 10.4018/978-1-60960-102-7.ch003
INTRODUCTION

Data warehouses have established themselves as necessary components of an effective IT strategy for large businesses. Modern data warehouses can be expected to handle up to 100 terabytes or more of data (Berson & Smith, 1997; Devlin, 1998; Inmon, 2002; Imhoff et al., 2003; Schwartz, 2003; Day, 2004; Peter & Greenidge, 2005; Winter & Burns, 2006; Ladley, 2007). In addition to the streams of data being sourced from operational databases, data warehouses must also integrate increasing amounts of external data to assist in decision support. It is accepted that the Web now represents the richest source of external data (Zhenyu et al., 2002; Chakrabarti, 2002; Laender et al., 2002), but we must be able to couple raw text or poorly structured data on the Web with descriptions, annotations and other forms of summary meta-data (Crescenzi et al., 2001).

In an effort to ensure the availability and quality of external data for the data warehouse we propose an intermediate data-staging layer called the Meta-Data Engine (M-DE). Instead of clumsily seeking to combine the highly structured warehouse data with the lax and unpredictable web data, the M-DE we propose mediates between the disparate environments.

In recent years the Semantic Web (SW) initiative has focused on the production of “smarter data”. The basic idea is that instead of making programs with near human intelligence, we rather carefully add meta-data to existing stores so that the data becomes “marked up” with all the information necessary to allow not-so-intelligent software to perform analysis with minimal human intervention (Kalfoglou et al., 2004). The Semantic Web (SW) builds on established building block technologies such as Unicode, Uniform Resource Indicators (URIs), and Extensible Markup Language (XML) (Dumbill, 2000; Daconta, Obrst & Smith, 2003; Decker et al, 2000). The modern data warehouse must embrace these emerging web initiatives.

In order to overcome the many technical challenges that remain before the Semantic Web (SW) can be adopted, key problems in Data Retrieval (DR), Information Retrieval (IR), Knowledge Representation (KR) and Information Extraction (IE), must be addressed (Silva and Rocha, 2003; Manning et al., 2008; Horrocks et al., 2005; Zaihrayeu et al., 2007; Buitelaar et al., 2008). The rise of the 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 (Abiteboul et al., 1999; Embley et al., 2005; Etzioni et al., 2008). Without the existence of robust meta-data there is no 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 (Dzbor & Motta, 2006; 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 (Holzinger et al., 2006; Sicilia, 2006; Schreiber & Aroyo, 2008).

Mapping web data to domain ontologies allows 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