Chapter 8.5
Automatically Extracting and Tagging Business Information for E-Business Systems Using Linguistic Analysis

Sumali J. Conlon
University of Mississippi, USA

Susan Lukose
University of Mississippi, USA

Jason G. Hale
University of Mississippi, USA

Anil Vinjamur
University of Mississippi, USA

ABSTRACT

The Semantic Web will require semantic representations of information that computers can understand when they process business applications. Most Web content is currently represented in formats such as text, that facilitate human understanding, rather than in the more structured formats, that allow automated processing and computer understanding. This chapter explores how natural language processing (NLP) principles, using linguistic analysis, can be employed to extract information from unstructured Web documents and translate it into extensible markup language (XML)—the enabling currency of today's e-business applications, and the foundation for the emerging Semantic Web languages of tomorrow. Our prototype system is built and tested with online financial documents.

INTRODUCTION

Business decision makers demand relevant, accurate, and complete information about the

marketplaces in which they compete. The World Wide Web is a rich but unmanageably huge source of human-readable business information—some novel, accurate, and relevant—some repetitive, wrong, or out of date. As the flood of Web document tops 11.5 billion pages and continues to rise (Gulli & Signorini, 2005), the human task of grasping the business information it bears seems more and more hopeless. Today's Really Simple Syndication (RSS) news syndication and aggregation tools provide only marginal relief to information-hungry, document-weary managers and investors. In the envisioned Semantic Web, business information will come with handles (semantic tags) that computers can intelligently grab onto, to perform tasks in the business-to-business (B2B), business-to-consumer (B2C), and consumer-to-consumer (C2C) environments.

Semantic encoding and decoding is a difficult problem for computers, however, as any very expressive language, for example, English provides a large number of equally valid ways to represent a given concept. Further, phrases in most natural (i.e., human) languages tend to have a number of different possible meanings (semantics), with the correct meaning determined by context. This is especially challenging for computers. As a standard artificial language emerges, computers will become semantically enabled, but humans will face a monumental encoding task. For e-business applications, it will no longer be sufficient to publish accurate business information on the Web in, say, English or Spanish. Rather, that information will have to be encoded into the artificial language of the Semantic Web—another time-consuming, tedious, and error-prone process. Pre-standard Semantic Web creation and editing tools are already emerging to assist early adopters with Semantic Web publishing, but even as the tools and technologies stabilize, many businesses will be slow to follow. Furthermore, a great deal of textual data in the pre-Semantic Web contains valuable business information, floating there along with the out-dated debris. However, the new Web vessels—automated agents—cannot navigate this old-style information. If the rising sea of human-readable knowledge on the Web is to be tapped, and streams of it purified for computer consumption, e-business systems must be developed to process this information, package it, and distribute it to decision makers in time for competitive action. Tools that can automatically extract and semantically tag business information from natural language texts will thus comprise an important component of both the e-business systems of tomorrow, and the Semantic Web of the day after.

In this chapter, we give some background on the Semantic Web, ontologies, and the valuable sources of Web information available for e-business applications. We then describe how textual information can be extracted to produce XML files automatically. Finally, we discuss future trends for this research and conclude.

BACKGROUND

The World Wide Web Consortium (W3C) is leading efforts to standardize languages for knowledge representation on the Semantic Web and is developing tools that can verify that a given document is grammatically correct according to those standards. The XML standard, already widely adopted commercially as a data interchange format, forms the syntactic base for this layered framework. XML is semantically neutral, so the resource description framework (RDF) adds a protocol for defining semantic relationships between XML-encoded data components. The Web ontology language (OWL) adds to RDF tools for defining more sophisticated semantic constructs (classes, relationships, constraints) still using the RDF-constrained XML syntax. Computers can be programmed to parse the XML syntax, find RDF-encoded semantic relationships, and resolve meanings by looking for equivalence relation-