Automatic Annotation of Content-Rich HTML Documents: Structural and Semantic Analysis

Saikat Mukherjee[1], Guizhen Yang[2], and I. V. Ramakrishnan[1]

[1] Department of Computer Science
Stony Brook University, Stony Brook, NY 11794, U.S.A.
{saikat,ram}@CS.StonyBrook.EDU
[2] Department of Computer Science and Engineering
University at Buffalo, Buffalo, NY 14260, U.S.A.
gzyang@CSE.Buffalo.EDU

Abstract. Although RDF/XML has been widely recognized as the standard vehicle for representing semantic information on the Web, an enormous amount of semantic data is still being encoded in HTML documents that are designed primarily for human consumption and not directly amenable to machine processing. This paper seeks to bridge this semantic gap by addressing the fundamental problem of *automatically* annotating HTML documents with semantic labels. Exploiting a key observation that semantically related items exhibit consistency in presentation style as well as spatial locality in template-based content-rich HTML documents, we have developed a novel framework for automatically partitioning such documents into semantic structures. Our framework tightly couples structural analysis of documents with semantic analysis incorporating domain ontologies and lexical databases such as WordNet. We present experimental evidence of the effectiveness of our techniques on a large collection of HTML documents from various news portals.

1 Introduction

The Semantic Web espouses a vision [3] for next-generation information networks where content providers define and share machine processable data on the Web to empower a variety of automated tasks ranging from information integration to Web services. There have been significant activities in building the technological infrastructure for realizing the Semantic Web vision and applications based on it. Notable efforts include developing Semantic Web languages such as DAML+OIL [22, 23] based on Description Logic, and DAML Rules [16], which seeks to extend the Semantic Web with rule-based and object-oriented capabilities. Prototype systems, such as FaCT [21], $\mathcal{F}$LORA-2 [31, 30], and TRIPLE [28], are now available to drive the knowledge representation and reasoning component of the Semantic Web.

Semantic Web documents contain metadata to express the meaning of their content. Tools such as those pioneered by SHOE [20] and OntoBroker [13, 29] facilitate manual annotation of HTML documents with semantic markups. However, an enormous amount of semantic data (such as product descriptions and pricing information) is still being encoded in "plain" HTML documents. These