SEAN: A System for Semantic Annotation of Web Documents

Amarjeet Singh* **Saikat Mukherjee*** **I. V. Ramakrishnan*** **Guizhen Yang†** **Zarana Shah***

*Department of Computer Science
Stony Brook University, Stony Brook, NY 11794

†Department of Computer Science and Engineering
University at Buffalo, Buffalo, NY 14260

Semantic Web documents use metadata to express the meaning of the content encapsulated within them. Although RDF/XML has been widely recognized as the standard vehicle for describing metadata, an enormous amount of semantic data is still being encoded in HTML documents that are designed primarily for human consumption. Tools such as those pioneered by SHOE [Heflin *et al.*, 2003] and Ontobroker [Fensel *et al.*, 1998] facilitate manual annotation of HTML documents with semantic markups.

Figure 1: New York Times front page

In this demo we will present SEAN, a system for *automatically annotating* HTML documents. It is based on the idea that well-organized HTML documents, especially those that are machine generated from templates, contain rich data denoting semantic concepts (*e.g.*, "News Taxonomy" and "Major Headline News") and concept instances. These kinds of documents are increasingly common nowadays since most Web sites (*e.g.*, news portals, product portals, etc.) are typically maintained using content management software that creates HTML documents by populating templates from backend databases. For instance, observe that the Web page in Figure 1 has a news taxonomy (on the left), which does not change, and a template for major headline news items. Each of these items begins with a hyperlink labeled with the news headline (*e.g.*, "White House ..."), followed by the news source (*e.g.*, "By REUTERS ..."), followed by a timestamp and a text summary of the article (*e.g.*, "The White House today ...") and a (variable) number of pointers to related news. These concepts and concept instances can be organized into a semantic partition tree (such as the one shown in Figure 2, which represents the "semantics" of the HTML document.

In a semantic partition tree each partition (subtree) consists of items related to a semantic concept. For example, in Figure 2 all the major headline news items are grouped under the subtree labeled "Major Headline News". There are two main tasks underlying the creation of a semantic partition tree from an HTML document: (i) identify segments of the document that correspond to semantic concepts; and (ii) assign labels to these segments. Informally, we say that several items are semantically related if they all belong to the same concept.

SEAN automatically transforms well-structured HTML documents into their semantic partition trees by exploiting two key observations. First, *semantically related items exhibit consistency in presentation style*. For instance, observe in Figure 1 the presentation styles of those items in the news taxonomy on the left. The main taxonomic items, "NEWS", "OPINION", "FEATURES", etc., are all presented in bold font. All the subtaxonomic items (*e.g.*, "International", "National", "Washington", etc.) under the main taxonomic item (*e.g.*, "NEWS") are hyperlinks. A similar observation can also be made on all the major headline news items. Second, *semantically related items exhibit spatial locality*. For example, when rendered in a browser, all the taxonomic items are placed in close vicinity occupying the left portion of the page. Specifically, in the DOM tree corresponding to the HTML document in Figure 1 all the items in the news taxonomy will be grouped together under one single subtree.

The first observation leads to the idea of associating a type with every leaf node in the DOM tree. The type of a leaf node consists of the root-to-leaf path of this node in the DOM tree and captures the notion of consistency in presentation style. The second observation motivates the idea of propagating types bottom-up in the DOM tree and discovering structural recurrence patterns for semantically related items at the root