On the Power of Semantic Partitioning of Web Documents

Guizhen Yang Saikat Mukherjee
Wenfang Tan I.V. Ramakrishnan
Department of Computer Science
State University of New York at Stony Brook
Stony Brook, NY 11794-4400
{guizyang, saikat, wtan, ram}@cs.sunysb.edu

Hasan Davulcu
Department of Computer Science
and Engineering
Arizona State University
Tempe, AZ 85287-5406
hdavulcu@asu.edu

1 Introduction

A growing number of Web sites are maintained by content management software and thus a large number of Web pages are machine-generated via templates. Normally in such Web pages there is *implicitly* a fixed "schema" and what changes is the content. Informally a schema for a Web page represents concepts and relationships among them in a hierarchical fashion. For example, Figure 1 is a screen shot of the New York Times front page (see http://www.nytimes.com). Observe that this page includes: (i) a taxonomy of items such as "NEWS" (consisting of hyperlinks labeled with "International", "National", ...), "OPINION" (consisting of hyperlinks "Editorial/Op-Ed", ...), etc.; (ii) several headlines of news articles where each article begins with a hyperlink labeled with the news headline (*e.g.*, "Bush tells Nation ...") followed by the author of the article (*e.g.*, "By Richard W. Stevenson ..."), followed by a time-stamp and a text summary of the article (*e.g.*, "President Bush portrayed ..."). The schema for this fragment of the New York Times front page therefore includes the taxonomy (which does not change) and the template for the news article. We should point out that the schema will also include several additional elements pertaining to other content appearing in the page.

Figure 1: New York Times Front Page

The important question then is: *Can the implicit schema in template-driven HTML documents be made explicit?* We formulate the problem of schema discovery from HTML documents as one of discovering semantic structures in Web documents and partitioning them accordingly. Our objective is to take a HTML document generated by a template and automatically partition it into semantically meaningful clusters via structural and semantic analysis. Each partition will consist of items related to a semantic concept. For example, Figure 2 is such a tree corresponding to the New York Times front page in Figure 1. Observe in this figure that the headline news items are all grouped under the "Headline News" category.

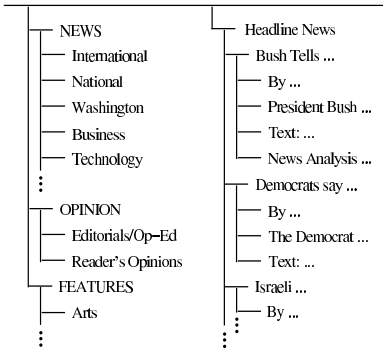

Figure 2: Screen Shot of the Semantic Partition Tree for New York Times Front Page

Semantic partitioning has several important and powerful implications in practice. First, it eases the task of formulating queries to retrieve data from Web documents. In the New York Times example, one can pose a query to retrieve all the links under the "NEWS" item in the taxonomy. Knowledge of the schema made explicit via semantic partitioning is the key to transforming legacy HTML documents into more semantics-oriented document formats such as XML [XML, 2003] and DAML [DAML, 2000]. Yet another application is audio-browsable Web content. By putting a dialog interface to the content of a Web page which is reorganized based on the knowledge of its schema, a user can easily browse its content using audio. More generally a Web site itself can be navigated using voice commands. Audio browsable Web content can significantly expand the reach of the Web to visually challenged individuals. Finally semantic partitioning can enable the creation of self-repairing wrappers, the technology that