# Automatic Discovery of Semantic Structures in HTML Documents

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

## Abstract

*Template-driven HTML documents posses an* implicit, *fixed schema denoting concepts and their relationships in a hierarchical fashion. Discovering this schema remains a relatively unexplored problem. By exploiting a key observation that semantically related items in HTML documents exhibit spatial locality, we develop an algorithm for automatically partitioning them into tree-like semantic structures which expose the implicit schema.*

## 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.*, "Inspectors in Iraq ...") followed by the author of the article (*e.g.*, "By JOHN F. BURNS ..."), followed by a time-stamp and a text summary of the article (*e.g.*, "The new zero-tolerance ..."). 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.

Knowing this schema explicitly has several uses. For example, it eases the task of formulating queries to retrieve

**Figure 1. New York Times Front Page**

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 is the key to transforming legacy HTML documents into more semantics-oriented document formats such as XML [3] and DAML [1]. Yet another application is audio-browseable 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 browseable Web content can significantly expand the reach of the Web to visually challenged individuals.

The important question then is: *Can the implicit schema in template-driven HTML documents be made explicit?* But discovering schema of HTML documents requires gleaning *semantic* information from HTML tags – a task that is difficult if not impossible to accomplish since their primary purpose is presentation of data in the document. Although steady progress is being made in the development of markup languages that facilitates extraction of semantic knowledge (*e.g.*, XML and DAML), HTML documents continue to proliferate.