Querying for Relations from the Semi-structured Web (Keynote Address, COMAD 2009)

Sunita Sarawagi
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay
sunita@iitb.ac.in

Abstract

We present a class of web queries whose result is a multi-column relation instead of a collection of unstructured documents as in standard web search. The user specifies the query either via a few example records, or a text description of columns of the relation. Starting from this seed, we show how to compile the result from several, possibly overlapping, tables and lists on the web. Many challenges arise in the process. First, we need to be able to extract structured records from HTML pages with little user supervision. We present algorithms for jointly aligning arbitrary record sets on the web with the query table. We adapt state of the art extraction models like Conditional Random Fields to exploit inter and intra source regularity in a unified framework. Second, we need to be able to consolidate the results from several sources in the face of missing columns, noisy extractions, and zero human supervision. We show how a suitably designed Bayesian networks allows us to compose a resolver from a library of type-specific similarity functions and table statistics. Finally, we discuss the problem of ranking the result rows by their estimated membership in the hidden target relation.