ISSUES IN EXTRACTING INFORMATION FROM THE WEB

(Extended Abstract)

William W. Cohen
WhizBang Labs - Research
4616 Henry Street, Pittsburgh PA 15213
wcohen@whizbang.com

Text extraction is the process of pulling out certain substrings from a document, and (possibly) storing these extracted substrings into a database in a systematic way. For instance, a text extraction system might examine "help wanted" ads and construct database records containing a job title, a job location, and an employer's name and address.

Extraction from natural language text is often facilitated by linguistic information, like part of speech and shallow parses. Unfortunately, much of the information on the web is not presented in grammatical running text, but instead is presented in short snippets of text organized into lists and tables. This makes use of linguistic information problematic. In extracting text from such "semi-structured" documents, information about *document formatting* can sometimes play the same role that linguistic information does in grammatical documents; however, a number of technical issues arise in doing this.

One issue is that, just as natural language sentences can be parsed multiple ways, formatting structures can be syntactically ambiguous. For instance, complex tables often can be interpreted in many different ways, only some of which are semantically correct. This leads to an unfortunate cyclic dependency, in which semantic analysis (extraction) requires formatting analysis, which in turn requires semantic analysis.

A second issue is that on many web sites, there are strong regularities in the types of formatting structures used. The situation is roughly analogous to a parsing problem in which thousands of documents must be parsed, but each document is written in its own distinctive sublanguage of English. Extraction from a web site can often be facilitated if one can infer the local formatting sublanguage. However, this local sublanguage inference problem also typically requires some semantic understanding, again leading to a cyclic dependency.

Fortunately, methods exist for breaking these cycles. "Shallow" format analysis can be used as a set of features for a learning system. This allows the learning method to resolve possible ambiguities. Local sublanguage learning can be addressed by either probabilistic or iterative techniques. In my talk I will present a number of short case studies of these techniques, and explain how they relate.