

PS on Web Information Extraction

Differences and similarities in the methods of data extraction presented in papers #19 & #20

by Redlingshofer Leopold e0325929@student.tuwien.ac.at
Seferovic Edvin e0325189@student.tuwien.ac.at

April 2005

Wrapper Generation

Data-rich section extraction works in a similar way by using in depth first order algorithm of the DOM tree. Badly-formed HTML documents will be transformed into well formed in [20]. The wrapper generation in ADead[20] is more simple, though it will be interesting if the generated minimum extract tree applies to similar pages with small deviation in data content/schema. It seems that this is an advantage of this method because all data is correctly sorted into the DOM-tree. The wrapper generation in [19] is more complicated and less DOM sub trees are thrown away. The complex structure (building a token suffix tree, building a pattern tree,..) in [19] is needed for generating of a flat table structure. With other tools or human understanding it would be possible to reverse engineer the original (relational) database. Optional attributes and disjunctions that are not found in training data set will not be entered into the table because of unsuitable regular expression.

Data Extraction

In ADead system[20] new elements will not be lost because all of the TD-pattern separation of the output slot which are entered into XML. In [19] a data tree is used to deduce a table which contains all data.

Label Assignment/Data Annotation

Both methods base on the fact that the annotation text is visually close to data elements. In [20] the minimum extract tree is used, in [19] the form elements are used for annotation as they are connection to the hidden web. In [19] several heuristics are used for label assignment of extracted data, and [20] actually does not describe any other heuristics then the computation of association possibility.

Implementation

ADead[20] system seems to work well with RISE pages according to the test that authors made. Anyhow it would be very interesting to see those two systems in praxis and to compare them in many aspects (correctly extracted/annotated data, speed, use memory etc.).