

A Comparison between two papers:
“A Flexible Learning System for Wrapper Tables
and Lists in HTML Documents”
and
“A Gateway From HTML to XML”

Sunil Pilani*
Friedrich Dimmel†

15th April 2005

Initial comparison

One of the main differences between these two papers is located in their output type. The Gateway algorithm returns the parsed data as an XML-DOM structure which can be used later in other different engines. The WL² system returns data only as a plain text document without any structure. So it will be difficult too, to use the data furthermore. But with this algorithm and its abilities you can make training sets for specific data retrieval.

Additionally the Gateway program extracts all information from the entire document, whereas the WL² system only extract relevant or user-selected data sets.

Extraction possibilities

In the WL² algorithm the input data is a HTML document. The WL² System can exploit several different representations of a document. For example DOM-level and token-level representations, as well as two-dimensional geometric views of the rendered page and representations of the visual appearances of text. In contrast, the Gateway algorithm uses the DOM structure of the given document,

*Registration No: 9925078; Classification No.: 881; Study: Computer Science

†Registration No.: 0302230; Classification No.: 534; Study: Software and Information Engineering;

transforms it into a T-DOM and finally extracts content data using layout patterns.

Knowledge-based systems

The WL² system is based on the theory of knowledge-based systems. It makes use of predicates, training data and biases. With this methods the system can find and extract the data in a more effective and efficient manner. Based on already aquired knowledge, the system can adapt itself to new tasks.

The Gateway system does not use any knowledge-base system functionality, it claims to work for any page.

Structures

The WL² system does not transform the HTML document into structures or trees, it works with different algorithms, just to extract the relevant data and returns text output.

The Gateway system transform the HTML-DOM into a special T-DOM to build tree structures out of them. With a multi-step topic and content aggregation, important topic nodes are used as “root”-nodes with content data below them. With such a tree, it is easy to generate an XML structure as output.

Table transformation

WL²: The approach that is taken to extract data from tables is based on geometric relationship between data elements rather than semantic relationship between them. An abstract geometric model of each data table is constructed and refined (rationalization, complex cell analysis and nomalization). Then the annotation of the nodes is done by adding attributes directly to the DOM nodes. The builders then model table regularities by accessing attributes in the enriched, annotated DOM tree.

The Gateway algorithm transforms multi-dimensional tables into one-dimensional systems. That means, it repeats table-headings in each line for each cell and marks them as important with a special markup (e.g. bold). So each dataset can be transferred into a single XML structure.

Software solutions

WL², originally a product of WhizBang Labs, was a commercial success and it is still an application in the software solutions of Inxight Inc..

The H2X Gateway is a software product which implements the mentioned features. According to the authors' paper it does its job very well and receives mostly more than 90% correct aggregation.