

First How does XWrap-works

- § XWRAP works in four Steps
- § Syntactical Structure Normalization
- § Information Extraction
- § Code generation
- § Testing and Packaging

XWrap is a supervised Wrapper

- § XWrap is not so „intelligent“ to work without User interaction.
- § XWrap need´s „help“ in three steps
- § The User Interaction is needed by
 - § Region Extracting
 - § Hierarchical Structure Extraction
 - § Testing and packing

The Four Steps

1.Step Syntactical Structure input

- § Prepares and set´s up the environment
- § The syntactical normalizer accepts an URL and establishes the connection to the WEB-Page
- § Fetches the Page
- § Cleans the HTML from bad tags

2. Step Information Extracting

- § All the logical transforming, parsing etc...
- § This is the main component of XWrap
- § The four big parts of Information Extracting
 - § Pre-processing
 - § Region Extracting
 - § Semantic token Extracting
 - § Hierarchical Structure Extraction

2.1. Pre-processing

- § Clean HTML doc is feed to an language-compliant tree parser
- § Character by Character parsing the document to units called syntactical tokens into a parsing tree
 - § Each node representing syntactical token
 - § Each tag Node represents a pair of HTML Tags
 - §

 - § </br>
 - § Non leaf nodes are tags
 - § All leaf nodes are text strings

2.2. Region Extracting

The region extractor asks the User to highlight the start tag of an important element->end tag of the semantic token, find by XWrap

- § Highlighting the entire region
 - § Additional region extractor computes type
 - § Sub regions
 - § Derives the set of region extracting rules
 - §
- § For each region a special set of rules is used
- § Result: Cuts the regions of interest out of the tree

2.3. Semantic Token Extracting

Walks through the tree structure generated by the semantic tokens

- § Generates a set of semantic tokens of interest
- § Uses a CSV with all the paired element types and values
- § XWrap supports some delemeters in CSV
 - § Comma „,“
 - § Semicolon „;“
 - § Pipes „|“
- § S-token extractor walks successiv through tree nodes
- § Starting at first leaf node not grouped into a token

2.4. Hierarchical Structure Extraction

- § This step determinates nesting hierarchy
 - § What kind of hirarchy
 - § What are the top level sections (tables)
 - § What are the subsections (rows or columns)
 - § Wich sections belongs together
 - § Using the XML-templates
- § Outcame of this Section
 - § Set of hierarchical structure extraction rules in a context-free grammar
 - § Simple heuristic for feedback-driven interaction with the user

3.Step Code generation

- § XWrap uses three sets of rules
- § Uses an implemented sets of code for semantic knowledge in XML-format (Templates)

4.Step

- § Testing and PackagingFor each URL, the Wrapper goes through syntactical structure
- § User is able to see the results
- § User is satisfied, the program ends
- § User is not satisfied with result, then improve the rules of extraction

Weak points of XWrap

- § Better tools for information extracting rules
- § To hone XWrap template language

Improved XWrap called XWrap-Elite

- § Not a semiautomatic Wrapper
- § No user interaction