Key Expression Driven Record Mining for Event Calendar Search

1. Motivation and definition of terms
   - Extraction of data records from websites with event calendars using a one-step mining technique
   - Event calendar records are data records with event details like event title, event location, event date, etc.
   - Usage of language-specific key expressions, e.g., wdAbbrev = (Mo|Di|Mi|Do|Fr|So), and HTML patterns to recognize every single event

2. Website retrieval and page classification
   1. Retrieving websites of opera houses
   2. Classification into pages with event calendar records (a) or without (b)
      - Page contains at least two key expressions
      - Start searching for event calendar records
      - Page contains less than two key expressions
      - Page will be skipped

3. Building up the DOM tree of a selected web page
   - Types of event calendar records:
     a) One single record under one node
     b) All records under one node; each record consists of a set of child nodes

4. Exploiting the structure of event calendar records within the DOM tree
   - Repetition of an HTML tag, e.g., <div>, with non-recurring attributes across all data records, including their text values,
   - Repetition of an HTML tag, e.g., <div>, with recurring attributes across all data records and sometimes incomplete attribute-value-pairs,
   - Missing both tag attribute and value.

5. Decision of smallest maximum data region
   1. Top-down traversal of the DOM tree starting at the root node
   2. Searching for key expressions in order to detect the regularities
   3. Calculating maximum distance of re-appearing key expressions
   4. Detecting the corresponding sub-tree within the HTML structure

6. Conclusion and future work
   - Evaluation results show excellent results of recall (99.56%) and precision (100%)
   - Automated key expression learning and substructure analysis of data records
   - Approach is validated on a much larger test set
   - Test scenario on other websites with event information (e.g., sports)

LWA 2009
Lernen Wissen Adaptivität
21. - 23. September, Darmstadt

Yeong Su Lee
Michaela Geierhos