Master Thesis Description, status: allocated to Roland Vallery. Language: EN / DE

Tree-based Edit Analysis of Wiki Articles

Summary

Most wikis today use text blobs written in a wiki markup dialect to store and work with articles. We have developed a formal parser that produces a rich tree-based format of wiki content, called Wiki Object Model (WOM). Users of wikis perform edit operations of wiki content (a.k.a. transformations or refactorings) that is carried out in multiple smaller steps. Using the HD tree diff algorithm, this thesis converts a large wiki corpus into a time series of edit scripts and then mines this series for recurring patterns of user behavior. We expect these patterns to reflect the higher-level operations users had in mind when performing the smaller atomic edits.

Work Results

- Literature review
  - Existing work on analysing the English Wikipedia's edit history
  - Relevant work on tree diffs, edit scripts, and pattern recognition algorithms
- Research approach and execution
  - Creation of time series from parsing and diffing Wikipedia article revisions
  - Choice of set of pattern recognition algorithms relevant to the task
  - Application of algorithms to time series data; choice of best performing one
  - (Evaluation of external validity of patterns found by survey or other method)
- Research results
  - Software that creates the time series and mines it for patterns
  - A list of defined recognized patterns ordered by number of occurrences (top 100)

Supervisor

Dipl.-Inf. Hannes Dohrn, hannes.dohrn@fau.de

Prof. Dr. Dirk Riehle, dirk.riehle@fau.de

Open Source Research Group
Computer Science Department
Friedrich-Alexander University

Link to thesis descriptions: http://osr.cs.fau.de/fun
Link to layout of final theses: http://wp.me/pDU66-S1
Link to grading framework for final theses: http://wp.me/pDU66-MF

Link to doc: http://goo.gl/3Uvj0P