Master Thesis Description: Daniel Knogl, status: assigned, language: [DE | EN]

Design and Implementation of Graph-DB based Storage for Wikipedia Articles

Summary

The English Wikipedia alone contains roughly 38 million pages of which 5 million are articles with an average revision count of 21 revisions per article, totalling ~800 million revisions as of February 2016. An article can be parsed into a document tree (we use the Wiki Object Model). Using HDDiff, a graph-based differencing algorithm, a mapping between the old revision's nodes and the new revision's nodes can be established. The goal of this thesis is to evaluate the storage costs and query/retrieval performance if one were to store the entire English Wikipedia as time-series graph.

Work Results

- Literature review
    - Definition of requirements and quality criteria
    - Investigate available graph DB software for suitability (e.g. license, maturity, support)
    - Conceptual evaluation of alternatives
- Design and implementation of graph DB based revision storage
    - Design and document database schema
    - Implement middleware for storing and retrieving revisions from graph DB
    - Implement Mediawiki dump loader (Wiktionary-en, Wikipedia-en)
    - REST-API that implements selected use cases
        - Get resource (subtree)
        - Get commits (for resource)
        - Get outgoing-links (for resource)
        - Optional: Get history of resource
    - Implement performance evaluation framework
- Evaluation and discussion of results
    - Storage and retrieval of articles to and from the graph DB demonstrated by integration tests

    - Evaluation and discussion of costs and performance of graph DB based implementation

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

More information: http://osr.cs.fau.de/theses/resources/

Read the description on UnivIS