University of Zurich^UZH

Department of Informatics

University of Zürich
Department of Informatics
Binzmühlestr. 14
CH-8050 Zürich
Phone. +41 44 635 43 11
Fax +41 44 635 68 09
www.ifi.uzh.ch/dbtg

UZH, Dept. of Informatics, Binzmühlestr. 14, CH-8050 Zürich

Prof. Dr. Michael Böhlen
Professor
Phone +41 44 635 43 33
Fax +41 44 635 68 09
boehlen@ifi.uzh.ch

Zürich, 26. März 2020

Master Basic Module
Topic: Applying Bit-Wise Dynamic Interleaving in the RCAS Index

The Robust Content-And-Structure (RCAS) index [1] is a novel index for semi-structured hierarchical data. Unlike pure content indexes or pure structure indexes, the RCAS index combines the content and structure of the data in a single index by *interleaving* paths and values. At the core of the RCAS index is a new interleaving scheme, called Dynamic Interleaving, that adapts to the distribution of the data and interleaves paths and values at their discriminative bytes. The discriminative byte of a set of byte-strings is the first byte after the longest common prefix of the strings, i.e., the first byte for which the strings differ. Interleaving paths and values at their discriminative bytes ensures robust query performance for Content-and-Structure CAS queries that consist of a path predicate and a value predicate. An example of such a CAS query is to find all files that are larger than 10MB and that are stored in the home directory of a user.

The RCAS index partitions the data such that all indexed keys are grouped together that have the same value for the discriminative path or value byte. Each partitioning step produces between 2 and 2^8 partitions. Experiments have shown that in practice partitioning the data in the path dimension produces few large partitions, while partitioning the data in the value dimension yields many small partitions [1]. This imbalance may negatively affect the robustness of the RCAS index, since smaller partitions can prune the search space faster during query processing.

The goal of this Master Basic Module is to implement and evaluate a version of the RCAS index that uses *bit-wise* dynamic interleaving. A bit-wise partitioning of the keys produces exactly two partitions. As a result, the imbalance between the path and value dimensions may be avoided.

Tasks

1. Study the relevant literature [1] to understand how the RCAS index interleaves paths and values at their discriminative bytes in the index.
2. Implement a version of the RCAS index that applies the *bit-wise* dynamic interleaving. The implementation should support
   - (a) Bulk-loading
   - (b) Evaluation of simple CAS queries. The path predicate needs to only support the descendant axis at the end of the query path.
3. Summarize your work in a short report.

Optional Tasks

1. Adapt the RCAS implementation to allow the descendant axis anywhere in the query path.
2. Conduct an experimental evaluation in which you compare the query performance of the *bit-wise* and the *byte-wise* RCAS index.

References

[1] K. Wellenzohn, M. H. Böhlen, and S. Helmer. Dynamic interleaving of content and structure for robust indexing of semi-structured hierarchical data. To be published.

Supervisor: Sven Helmer (helmer@ifi.uzh.ch), Kevin Wellenzohn (wellenzohn@ifi.uzh.ch)

Start date: March 1, 2020

End date: June 1, 2020

Oral exam: June 2, 2020, 14:00-14:30, BIN 2.E.13

University of Zurich
Department of Informatics

Prof. Dr. Michael Böhlen
Professor