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, 18. März 2020

MSc Thesis
Topic: Supporting Updates in the RCAS Index

The Robust Content-And-Structure (RCAS) index [2] 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.

Currently, the RCAS index only supports bulk-loading, while the support for updates (i.e., insertions and deletions) is still in its early stages [3]. Updating the RCAS index efficiently is challenging because as keys are inserted/deleted, the positions of the discriminative bytes can change on which the dynamic interleaving of the indexed keys is based.

Two approaches to update the RCAS index are described in [3]. The first (eager) approach maintains the dynamic interleaving but needs to re-structure large parts of the index, which is expensive. The second (lazy) approach can be implemented efficiently but sacrifices the alternating structure of the dynamic interleaving, which negatively affects the robustness of the RCAS index.

The goal of this Master thesis is to explore, implement, and evaluate new ways to update the RCAS index that strike a balance between the update and query performance of the index.

Tasks

1. Study the relevant literature: (a) [2] to understand how the RCAS index interleaves paths and values in the index, (b) [3] describes the problem of updating the RCAS index, and (c) [1] describes the trie structure (ART) on which the RCAS index is based.
2. Implement the eager and lazy approaches to update the RCAS index described in [3]. These approaches serve as baseline for the update and query performance in the experimental evaluation.
3. Investigate and implement new approaches to update the RCAS index efficiently and still provide good query performance.
4. Conduct an experimental evaluation in which the proposed approach is compared to the eager and lazy techniques. Evaluate the update cost and query performance of these approaches.
5. Write the thesis.
6. Present the thesis in a DBTG meeting (25 minutes presentation).

References

[1] V. Leis, A. Kemper, and T. Neumann. The adaptive radix tree: Artful indexing for main-memory databases. In *ICDE'13*, pages 38–49, Washington, DC, USA, 2013. IEEE Computer Society.

[2] 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.

[3] K. Wellenzohn, M. H. Böhlen, and S. Helmer. Updating the RCAS index. To be published.

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

Start date: 1 April 2020

End date: 30 September 2020

University of Zurich
Department of Informatics

Prof. Dr. Michael Böhlen
Professor