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, 10. Januar 2020

BSc Thesis
Topic: Integrating the RCAS Index with the Software Heritage Archive

The Software Heritage Archive [1, 3] is an attempt to collect all publicly accessible software source code. It is the largest archive of its kind and archives more than 90 million repositories with 6 billion source code files. The data is modeled as a graph that encodes, among others, the directory structure of the archived repositories as well as the contents of each source code file. Querying such a large database is challenging and requires specialized index structures.

The Robust Content-And-Structure (RCAS) index [4] is a novel index for semi-structured hierarchical data. Unlike pure content indexes or pure structure indexes, the RCAS index is designed to answer queries efficiently that contain a value predicate on the content of some attribute (e.g., file size) and a path predicate on the location of this attribute in the hierarchical structure of the data. An example of such a Content-and-Structure (CAS) query is to find all files that are larger than 10MB and that are stored in the home directory of a user.

The goal of this Bachelor thesis is to first implement the RCAS index and then integrate it with the Software Heritage Archive. In particular, the student should explore how the RCAS index can be used to index the Software Heritage Archive and what kind of queries are best supported by the RCAS index.

Tasks

1. Study the relevant literature [4] to understand how the RCAS index interleaves paths and values in the index.
2. Implement the RCAS index. The implementation must support (a) the bulk construction of the index for given a dataset and (b) querying the index for given a CAS query. Another possible angle to explore is a more space-efficient implementation of RCAS using techniques from [2].

3. Familiarize yourself with the Software Heritage Archive [1, 3]. Download a subset of the dataset and explore its structure. Which queries on this dataset can be supported by the RCAS index?
4. Integrate your RCAS implementation from task 2 with the dataset collected in task 3.
5. Conduct an experimental evaluation of your implementation. The evaluation should include runtime measurements for the queries selected before. Additionally, evaluate the space consumption of the RCAS index on the dataset.
6. Write the thesis (approximately 50 pages).
7. Present the thesis in a DBTG meeting (25 minutes presentation).

References

[1] R. Di Cosmo and S. Zacchiroli. Software heritage: Why and how to preserve software source code. In *iPRES 2017: 14th International Conference on Digital Preservation*, 2017.

[2] 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.

[3] A. Pietri, D. Spinellis, and S. Zacchiroli. The software heritage graph dataset: Large-scale analysis of public software development history. In *MSR 2020: The 17th International Conference on Mining Software Repositories*. IEEE, 2020.

[4] 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: Kevin Wellenzohn (wellenzohn@ifi.uzh.ch)

Start date: 27 January 2020

End date: 27 July 2020

University of Zurich
Department of Informatics

Prof. Dr. Michael Böhlen
Professor