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

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

Zürich, October 20, 2020

MSc Project
Topic: Implementing learned indexes on 1-dimensional and 2-dimensional data

Kraska et al. [1] show that the B-tree, an index for 1-dimensional sorted data, can be replaced by a learned index, i.e., a machine learning model. The authors argue that a fine-tuned learned index not only outperforms B-trees for point queries, but also is smaller than B-trees. Recent work [2] proposes LISA: a learned index for spatial data to replace 2-dimensional indices (e.g., R-trees and K-D-trees) on spatial data. LISA, the learned index on 2-dimensional data, extends the learned index method for 1-dimensional data [1] to support KNN and range queries, and it handles updates.

This goal of this master project is to implement the learned indexes [1, 2] and evaluate the implementations.

Tasks

- **Task 1: Literature Review**
  - Study the relevant research work on learned indexes. In particular, study the learned index for B-Trees [1] and the learned index for spatial data [2].
  - Study relevant data structures (e.g., B-trees for 1-dimensonal data, R-trees and KD-trees for 2-dimensonal data, etc) as well as point, range and KNN queries.
- **Task 2: Implementing a learned index for 1-dimensional data**
  - Implement the B-tree and point queries.
  - Implement a baseline learned index [1] that supports point queries.
  - Implement some (or all) optimized learned indexes using the Learning Index Framework and the Recursive Model Index together with search strategies [1] that support point queries.

  - Use the generated dataset from `https://github.com/stanford-futuredata/index-baselines` (visited at 20 October, 2020)
  - Evaluate your implementations. Compare space assumptions of the implemented learned indexes and B-tree. Analyse running times of building and training and the implemented learned indexes and B-tree. Analyse querying times of the implemented learned indexes and B-trees. Identify the limatations of the learned index.
- **Task 3: Implement a learned index for 2-dimensional data**
  - Implement either R-trees or K-D-trees that supports point, range and KNN queries.
  - Implement all components of LISA [2] that supports point, range and KNN queries.
  - Use 2-dimensional point data as the dataset.
  - Evaluate your implementations. Compare space assumptions of the implemented LISA and the chosen spatial index. Analyse running time of building and training the implemented LISA and chosen spatial index (either R-trees or K-D-trees). Analyse querying times of the implemented LISA and the chosen spatial index. Identify the limatations of LISA.
- **Task 4: Convolutional Neural Networks (CNNs) as learned indexes**
  - Conduct experimental studies on indexing 1-dimensional and 2-dimensional data using CNNs.
- **Task 5: Write the report**
  - Summarize your implementations and results in a final report.

References

[1] T. Kraska, A. Beutel, E. H. Chi, J. Dean, and N. Polyzotis. The case for learned index structures. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, page 489–504. Association for Computing Machinery, 2018.

[2] P. Li, H. Lu, Q. Zheng, L. Yang, and G. Pan. Lisa: A learned index structure for spatial data. SIGMOD '20, page 2119–2133. Association for Computing Machinery, 2020.

Supervisor: Qing Chen (qing@ifi.uzh.ch), Michael Böhlen (boehlen@ifi.uzh.ch)

Start date: October 20, 2020

End date: October 19, 2021

University of Zurich
Department of Informatics

Prof. Dr. Michael Böhlen
Professor