Multicriteria Data Structures

learned.di.unipi.it

3rd meeting of the PRIN project

"Multicriteria Data Structures and Algorithms: from compressed to learned indexes, and beyond"

12 MARCH 2021

Video conference

The ever growing need to efficiently store, retrieve and analyze massive datasets, originated by very different sources, is currently made more complex by the different requirements posed by users and applications. Such a new level of complexity cannot be handled properly by current data structures for Big Data problems.

To successfully meet these challenges, we propose a new generation of "Multicriteria Data Structures and Algorithms" that originate from some recent and preliminary results of the proponents. The "multicriteria" feature refers to the fact that we seamlessly integrate, via a "principled" optimization approach, modern compressed data structures with new, revolutionary, data structures "learned" from the input data by using proper machine-learning tools. The goal of the optimization is to select, among a family of properly designed data structures, the one that "best fits" the multiple constraints imposed by its context of use, thus eventually "dominating" the multitude of trade-offs currently offered by known solutions.

In this project, we will lay down the theoretical and algorithmic-engineering foundations of this novel research area, which has the potential of supporting innovative data-analysis tools and data-intensive applications.

Participants

Unit 1 - *Università di Pisa*

Paolo Ferragina (PI)	paolo.ferragina@unipi.it
Davide Bacciu	davide.bacciu@unipi.it
Antonio Boffa	antonio.boffa@phd.unipi.it
Antonio Carta	antonio.carta@di.unipi.it
Francesco Tosoni	francesco.tosoni@phd.unipi.it
Andrea Valenti	andrea.valenti@phd.unipi.it
Giorgio Vinciguerra	giorgio.vinciguerra@phd.unipi.it

Unit 2 - *Università degli Studi di Palermo*

Raffaele Giancarlo (PI)	raffaele.giancarlo@unipa.it
Domenico Amato	domenico.amato01@unipa.it
Mariella Bonomo	mariella.bonomo@unipa.it
Giosuè Lo Bosco	giosue.lobosco@unipa.it
Simona Ester Rombo	simonaester.rombo@unipa.it
Armando La Placa	armando.laplaca@community.unipa.it
Gennaro Grimaudo	gennaro.grimaudo@unipa.it

Unit 3 - *Università degli Studi del Piemonte Orientale "Amedeo Avogadro"-Vercelli*

Giovanni Manzini (PI)	giovanni.manzini@uniupo.it
Lavinia Egidi	lavinia.egidi@uniupo.it
Manuel Striani	manuel.striani@uniupo.it

Unit 4 - *Università degli Studi di Milano*

Marco Frasca (PI)	marco.frasca@unimi.it
Dario Malchiodi	dario.malchiodi@unimi.it
Giorgio Valentini	giorgio.valentini@unimi.it
Marco Mesiti	marco.mesiti@unimi.it
Alessandro Petrini	alessandro.petrini@unimi.it
Giosuè Marinò	giosumarin@gmail.com
Jessica Gliozzo	jessicagliozzo@gmail.com

Program

Friday, March 12, 2021

16:00 Welcome (*10 min*)

16:10 UniPa past and ongoing activity round-up [Tasks T1, T2, T3,T4]
Raffaele Giancarlo (*10 min*)

16:20 Developments on Theoretic and Practical Aspects of Learned Table Search and Indexing [Tasks T1, T3]
Domenico Amato (*20 min*)

16:40 UniMi past and ongoing activity round-up [Tasks T1, T2]
Marco Frasca (*15 min*)

16:55 Compression strategies and space-conscious representations for deep neural networks [Task T2]
Alessandro Petrini / Giosuè Marinò (*15 min*)

17:10 Break (10 min)

17:20 UniPO past and ongoing activity round-up [Tasks T1, T2, T3]
Giovanni Manzini (*15 min*)

17:35 Space Efficient Merging of Compressed Indices [Task T3]
Lavinia Egidi (*15 min*)

17:50 UniPi past and ongoing activity round-up [Tasks T1, T2, T3, T4]
Paolo Ferragina (*15 min*)

18:05 Learned compressed rank/select dictionaries [Tasks T3, T4]
Antonio Boffa (*15 min*)

18:20 Wrap-up, discussion on next events

Main tasks of the project

[T1] Classic Data Structures vs Purely Learned Indexes.

[T2] Compressed ML models.

[T3] Multicriteria Data Indexing.

[T4] Multicriteria Data Compression.

Some notes on the meeting

- **[T1]** Use the fixed-len or variable-len bucketing of the universe to boost the performance of (classic vs learned) data structures [UniPA]
- **[T1]** Mapping PGM into one NN and find an NN structure which is smoother in the learning process and can improve PGM performance in time and space [UniMI]
- **[T2]** Extending the set of NN compression techniques, considering also convolutional layers [UniMI]
- **[T2, T4]** Build a multicriteria optimiser for NNs across a range of possible compression techniques wrt given constraints and criteria (e.g. accuracy compared to the uncompressed model, space, time) [UniMI, UniPO]
- **[T4]** PPM versus NN [UniPI, UniPA, UniPO]
- **[T2]** New approach based on grammar compression that improves Compressed Linear Algebra [UniPI, UniPO]
- **[T3]** Use of Wheeler Automata for the design of Multicriteria succinct indices [UniPO]
- **[T1, T3]** Study new/engineered compressed and learned index for strings, possibly in a real DB scenario [UniPI]
- **[T1, T2]** Study the combination of repetitiveness (LZ-like) and approximate linearity in the data (PGM-like) [UniPI]
- **[T1, T3]** Study the application of classic and learned compressed data structures to real DBs [UniPI]