Multicriteria Data Structures

learned.di.unipi.it

4th meeting of the PRIN project

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

9 FEBRUARY 2022

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
Antonio Boffa	antonio.boffa@phd.unipi.it
Andrea Guerra	andrea.guerra@phd.unipi.it
Francesco Tosoni	francesco.tosoni@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
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

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: 9 february 2022 *(video conference)*

9:00 **Welcome** (*10 min*)

9:10 **UNIPA past and ongoing activity round-up [Tasks T1, T2, T3,T4]**
Raffaele Giancarlo (*15 min*)

9:25 **Neighborhood based approaches for the prediction of lncRNA-Disease associations from tripartite graphs [Tasks T3, T4]**
Mariella Bonomo (*20 min*)

9:45 **UNIMI past and ongoing activity round-up [Tasks T1, T2, T3]**
Marco Frasca (*25 min*)

10:10 **The role of classifiers and query distribution in Learned Bloom Filters [Task T1]**
Dario Malchiodi (*30 min*)

10:40 **Break (10 min)**

10:55 **UNIPO past and ongoing activity round-up [Tasks T1, T2, T3]**
Giovanni Manzini (*15 min*)

11:10 **UNIPI past and ongoing activity round-up [Tasks T1, T2, T3, T4]**
Paolo Ferragina (*15 min*)

11:30 **Scaling compression to massive matrices [Tasks T2, T4]**
Francesco Tosoni (*30 min*)

12:00 **Repetition- and linearity-aware rank/select dictionaries [Tasks T3, T4]**
Giorgio Vinciguerra (*20 min*)

12:20 ***Indexing and compressing regular languages (Invited talk)***
Nicola Prezza (*30 min*)

14:30 **Wrap-up, discussion on next events and research achievements**

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.

Results achieved on the research planned in the 3rd meeting

- **[T1]** Use the fixed-len or variable-len bucketing of the universe to boost the performance of (classic vs learned) data structures [UNIPA].
  - *Manuscript in Preparation. Results available in DR. Amato PHD dissertation*
- **[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]
  - *Discovered that NN models could reproduce but not improve the PGM results, so that we abandoned this research line*
- **[T2]** Extending the set of NN compression techniques, considering also convolutional layers [UNIMI].
  - *Manuscript in preparation*
- **[T2, T4]** New approach based on grammar compression that improves Compressed Linear Algebra [UNIPI, UNIPO]
  - *Results submitted to VLDB '22*
- **[T1, T3]** Study new/engineered compressed and data-aware index for strings, possibly in a real DB scenario [UNIPI]
  - *Results submitted to SIGIR '22*
- **[T1, T2]** Study the combination of repetitiveness (LZ-like) and approximate linearity in the data (PGM-like) [UNIPI, UNIPO]
  - *Paper accepted to ISAAC '21, now journal version under preparation*
- **[T1]** Study of Learned Bloom Filters [UNIPA,UNIMI]
  - *On the role of classifiers in Learned Bloom Filters-presented at ICPRAM 22- More work in progress.*
- **[T1, T3]** Study a learned index for strings based on DNN [UNIPI, UNIMI]
  - *Work in progress, some preliminary results available*
- **[T2]** Extending and further validating a lossless storage for DNN [UNIMI].
  - *Work in progress. Almost completed.*
- **[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]
  - *No progress due the delay in finalizing the preparatory studies.*
- **[T3]** Use of Wheeler Automata for the design of Multicriteria succinct indices [UNIPO]
  - *No progress since no postdoc applied for the position*
- **[T4]** PPM versus NN [UNIPI, UNIPA, UNIPO]
  - *Still in the process of finalizing this research, via a master student*
- **[T1, T3]** Study the application of classic and learned compressed data structures to real DBs [UNIPI]
  - *Still in the process of finalizing this research, via a master student*

Some notes on the 4th meeting, and planned activities for the remaining period of the project

- **[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, UNIPI]
  - *Apply grammar compression to NN matrices, considering the tradeoffs with quantization and pruning.*
- **[T3]** Design of Multicriteria succinct indices using BWT variants [UNIPI, UNIPO, UNIPA]
  - *This year we will aim at finalizing this work by means of a master student interested in pursuing this research*
- **[T4]** PPM versus NN [UNIPI, UNIPA, UNIPO]
  - *This year we will aim at finalizing this work by means of a master student interested in pursuing this research*
- **[T1, T3]** Study the application of classic and learned compressed data structures to real DBs [UNIPI]
  - *This year we will aim at finalizing this work by means of a master student interested in pursuing this research*
- **[T2]** NN compression [UNIMI]
  - Extending the set of NN compression techniques, considering also convolutional layers, data and problems .
- **[T1]** Study of Learned Bloom Filters [UNIPA, UNIMI]
  - Study the role of the classifier in the LBF with regard to the set size
  - Investigate the dependence of LBF on the query distribution
- **[T1,T3]** Compare DNN vs FST for string indexing [UNIPI, UNIMI]
  - Characterize the cases (if any) in which DNN are better in terms of space (and time?)
  - Consider the synergy of multiple DNN models
  - Evaluate the construction of an hybrid string index DNN+FST
- **[T1, T3]** Learned Boosters for Sorted Sets Dictionaries [UNIPA]
  - Consider the dynamic case
- **[T3]** More research on the subtrie compression and collapse technique [UNIPI]
  - Deal with the query distribution
- **[T3, T4]** More research on repetition and linearity aware rank/select dictionary [UNIPI]
  - Study efficient construction algorithms
- **[T2, T4]** Matrix compression for linear algebra approach [UNIPI, UNIPO]
  - Investigate matrices coming from real ML models
  - Investigate the design of other column permuting algorithms, such as 2-opt and 3-opt for TSP