Università degli Studi di Palermo

Multicriteria Data Structures

d.m.i matematica e informatica @ unipa

learned.di.unipi.it

5th meeting of the PRIN project

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

22-23 September 2022

University of Palermo, Dipartimento di Matematica ed Informatica – Aula 7

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.

Project 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@di.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
Manuel Striani	manuel.striani@uniupo.it
Alessandro Poggiali	alessandro.poggiali@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
Giosuè Marinò	giosumarin@gmail.com
Flavio Furia	furia.flavio@outlook.it
Alessandro Petrini	alessandro.petrini@unimi.it
Jessica Gliozzo	jessica.gliozzo@unimi.it

Program: 22 September 2022

14.45: Welcome Greetings

14.50: Flavio Furia – UNIMI - "Huffman coding for neural network compression"

15.10: Domenico Amato – UNIPA – "On the suitability of Neural Networks as Building Blocks for the design of efficient Learned Index"

15:30 Antonio Boffa - UNIPI - "Compressed String Dictionaries via Data-Aware Subtrie Compaction"

16.00: Coffee Break

16.20: Alessandro Poggiali – UNIUPO – "Clustering Classical Data with Quantum k-Means"

16.40: Giorgio Vinciguerra - UNIPI – "Advances on learned indexing and compression of integer data"

17.00 -18.00: Project Discussion

Program: 23 September 2022

10.15: Simona Rombo - UNIPA – "Topological ranks reveal functional knowledge encoded in biological networks: a comparative analysis"

10.35: Giosuè Cataldo Marinò - UNIMI – "Guidelines for topology-preserving compression of deep neural networks"

10.55: Planning of Forthcoming Activities

13.00: Lunch

14.00-14.30: Closing Remarks

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 4th meeting

- **[T1]** Study of Learned Bloom Filters [UNIPA, UNIMI]: Study the role of the classifier in LBF with regard to the set size; Investigate the dependence of LBF on the query distribution
  - *This research has been published in ICPRAM 2022, and there are many follow ups that will continue to be researched in the next months, probably after the end of this project.*
- **[T1, T3]** Study the application of classic and learned compressed data structures to real DBs [UNIPI]: *We will aim at finalizing this work by means of a master student*
  - *This work has been pursued independently by a company, named Manticore Search (https://manticoresearch.com/), because they included our PGM-index in their release 5. This shows the industrial impact of PGM-index. On the other hand, as far as our experiments on DBs are concerned, we did not find yet a student that could finalize this work.*
- **[T1,T3]** Compare Deep NN vs FST for string indexing [UNIPI, UNIMI]: Characterize the cases (if any) in which Deep NN are better in terms of space and possibly time than compacted/compressed tries; consider the synergy of multiple Deep NN models; evaluate the construction of an hybrid string index Deep NN+FST.
  - *The research already produced some joint results, but we consider it not yet mature to be published. So it is ongoing.*
- **[T1, T3]** Learned Boosters for Sorted Sets Dictionaries [UNIPA]; Consider the dynamic case and the case of general NN
  - *The case of NN boosters in the design of indexing data structures has been published in EANN 2022, got the Best Paper award, and invited to the special issue of the conference*
  - *The dynamic case is going to be submitted.*
- **[T1, T3]** More research on repetition and linearity aware rank/select dictionaries [UNIPI,UNIPO]:
  - *We extended the experiments of ISAAC 21 and added some theoretical results that has been included into the journal version of this work, now under review.*
- **[T2]** NN compression [UNIMI]: Extending the set of NN compression techniques, considering also convolutional layers, data and problems.
  - *Evaluation of the efficacy of state-of-the-art compression techniques for NNs with respect to the NN layer types and the problem type [journal paper submitted, second round of revision];*
  - *Finalizing the validation of novel NN compression techniques leveraging Huffman's codes and their variants [theoretical and experimental results are ready to be submitted to an international journal]*
  - *The two studies above also involve a use case in medical applications whose results have been submitted to an international journal*
- **[T2, T4]** Build a multicriteria optimiser for NN compression w.r.t. given constraints and criteria (e.g. accuracy compared to the uncompressed model: space, time) [UNIMI, UNIPI,UNIPO]:
  - *UNIMI: Plan to design a multi-criteria optimizer that exploits the studies carried out for task T2 and already described above*

    - *UNIPI,UNIUPO: Finalized the research on NN compression via grammars (i.e. RePair) publishing it on VLDB 2022. We are currently investigating the impact of this approach on ML matrices, and its combination with pruning/quantization.*
- **[T3]** More research on the subtrie compression and collapse technique [UNIPI]: *Deal with the query distribution*
    - *This work led to a paper published in SPIRE 2022. We are currently extending some of the experimental results to be included in the journal version.*
- **[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*
    - *Not yet addressed; new theoretical results suggest that the problem should be addressed in the most general context of Wheeler Graphs, and p-sortable graphs.*
- **[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*
    - *This promising joint research is still ongoing because, although we got several interesting results, unfortunately we did not find yet students that could finalize the work started two years ago.*

During the meeting we also discussed the following topics, sharing our preliminary results and achievements:

- **[T1]** Digging more in the study of learned indexes [UNIPA,UNIPI]:
    - Proposal of two new models derived from RMI that are competitive against the state of the art (i.e., RMI, PGM and RS) [UNIPA]: *Published in AIxIA 2022, and the journal version is under the second round of review.*
    - Choosing the right binary search routine [UNIPA]: *Published in Software Practice and Experience, 2022.*
    - Extending the case of piecewise linear approximations for the design of learned indexes and compressors to non-linear functions [UNIPI]: *Got preliminary results which are going to be finalized into a conference paper under preparation.*
- **[T3,T4]** Compressed views of graphs for efficient IR, discriminative compact patterns, and designing and experimenting on large scale these primitives [UNIPA]
    - *Published in Briefings in Bioinformatics, 2022*
    - *First round of review in Bioinformatics*
    - *Second round of review in BMC BioInformatics*
- **[T3,T4]** Why compression is useful: alignment free studies of k-mer statistics of large scale [UNIPA]
    - *Published in Briefings in Bioinformatics, 2022*

During the meeting we also shared and congratulated about two important achievements by some of the project participants:

(i) Giorgio Vinciguerra won the **2022 PhD Thesis award** from the EATCS Italian Chapter for his research on "*Learning-based compressed data structures*" [T1,T3,T4];

(ii) the paper by Domenico Amato, Giosuè Lo Bosco, Raffaele Giancarlo titled "*On the Suitability of Neural Networks as Building Blocks for the Design of Efficient Learned Indexes*" got the best paper award at **EANN 2022** [T1,T3].