Multicriteria Data Structures

Kickoff meeting del progetto PRIN

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

14 OTTOBRE 2019 / 15:00 - 19:30

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.

Partecipanti

Unità 1 - *Università di Pisa*

Paolo Ferragina (PI)	paolo.ferragina@unipi.it
Davide Bacciu	davide.bacciu@unipi.it
Antonio Carta	antonio.carta@di.unipi.it
Luca Oneto	luca.oneto@unipi.it
Andrea Valenti	andrea.valenti@phd.unipi.it
Giorgio Vinciguerra	giorgio.vinciguerra@phd.unipi.it

Unità 2 - *Università degli Studi di Palermo*

Raffaele Giancarlo (PI)	raffaele.giancarlo@unipa.it
Domenico Amato	domenico.amato01@unipa.it
Andrea De Salve	andrea.desalve@unipa.it
Giosuè Lo Bosco	giosue.lobosco@unipa.it
Simona Ester Rombo	simonaester.rombo@unipa.it

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

Giovanni Manzini (PI)	giovanni.manzini@uniupo.it
Lavinia Egidi	lavinia.egidi@uniupo.it

Unità 4 - *Università degli Studi di Milano*

Marco Frasca (PI)	marco.frasca@unimi.it
Giorgio Valentini	valentini@di.unimi.it
Dario Malchiodi	malchiodi@di.unimi.it

Ordine del giorno

15:00 Presentazioni (*15 min*)

15:15 Attività di ricerca UNIMI, Marco Frasca (*45 min*)

16:00 Attività di ricerca UNIPA, Raffaele Giancarlo (*45 min*)

16:45 Attività di ricerca UNIUPO, Giovanni Manzini e Lavinia Egidi (*45 min*)

17:30 Attività di ricerca UNIPI, Davide Bacciu, Paolo Ferragina e Giorgio Vinciguerra (*45 min*)

18:15 Discussione e pianificazione attività future

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

UNIMI: T1-D1.1 COMPRESSION OF NN

Preliminary experiments about learned indexes with neural networks models (UNIPA) + NN compression techniques (UniMI) to reduce the space occupancy and their comparison with some dictionary indexes.

UniMI presented some experiments about an input sequence uniformly extracted, they showed that the compression of the NN achieves also improvements in the error, not only in the space, and this is surprising.

They also started to look at the work to the RMI (in the sense of Kraska te al.), which has been indicated as part of our Task 3, in order to understand how their compression impact onto the RMI.

GOAL for T1 - D1.1 and D1.2:

- Understand how the results change by varying the input, in terms of space and accuracy
- They could incorporate the error accuracy on the space occupancy, in a sort of multi-criteria view, but they need to extend the loss function to the maximum error
- Extend the learned indexes to more levels (a là RMI)

UNIMI: T2-D2.1 DEVELOP COMPRESS ML MODELS AND RELATION

Preliminary study of compression algorithms for complex ML models. They wish to study the relation between the structure of the ML model and the achieved compression ratio, and the kind of problem dealt with.

GOAL for T2 - D2.1 and D2.2:

- Study novel techniques to compress ML models, by possibly exploiting some very recent sparsification techniques which occur at training time, thus possibly interesting from our multi-criteria point of view
- Experiment with different problem categories

UNIPA: T1-D1.1 and D1.2

The key addressed question is: A single NN what can do in terms of dictionary problem with respect to classic data structures?

They currently found that classic data structures are 10x faster than NN-based approaches (with more levels). The case of linear NN-models are more competitive and thus make everything hope for more, but the contours are not clear and they think that a more methodological approach is needed.

GOAL for T1:

- Propose a methodological approach for studying the comparison between classic indexes.
- Study the role of NN as booster of classic data structures. For now they studied the binary search, the question is how to extend to other data structures (results are known for Bloom Filters, Mitzenmacher).

UNIPA: T2-D2.1 and D2.2

They are mainly studying Spectral Kernels and Compressed Amnesic Probabilistic Automata.

UNIPA: T4-D4.1 and D4.2

They are mainly studying and designing novel classes of multi- criteria and ML-based data compressors (mainly lossy): with the goal of preserving the biological functionality.

Another issue is to study the relation between PPM versus DeepZip: namely, compare the prediction performance starting with an uncompressed trie, test it and if effective in prediction, study its compression (thus following the idea of "boosting").

UNIUPO:

They have summarized some results in BioInformatics that have been obtained recently by the group about some compression and string search algorithms. They aim to see whether some learned-based approaches could be applied on these problems to improve them.

Then, they moved to illustrate the activities more related to the project, thus sketching preliminary ideas on Learned FM-index, Compressed Linear Algebra, Probabilistic Suffix Tree, PPM versus DeepZip, ML security. Most of these studies are in collaboration with other partners: UniPI and UNIPA. UNIMI claims to be interested in Compressed Linear Algebra.

UNIPI:

They recalled the literature on learned indexes and then they explained the preliminary results they got : called, PGM-index; and then they illustrate the open problems.

The presentation was closed by discussing several open problems (and possible collaborations in brackets) on benefit of using more powerful models (UNIPA), how to manage insertion and

deletion in the PGM-index, justify theoretically the improvement of learned indexes, extend PGM to strings or FM-index (UniPMO), how to design hybrid indexes (UNIPA, UNIMI).

Actions to take

1. Create the site of the project
2. Create a shared folder where partners can store their papers, SW and datasets
3. Plan the meeting of February that will take place in Pisa