Giulio Ermanno Pibiri

Curriculum Vitae et Studiorum

Contact Information

National Research Council of Italy (CNR)
Institute of Science and Information Technologies "A. Faedo" (ISTI)
Via G. Moruzzi 1, 56124 Pisa, Italy

Email	giulio.pibiri@di.unipi.it
Email	giulio.ermanno.pibiri@isti.cnr.it
Personal page	`http://pages.di.unipi.it/pibiri`
GitHub profile	`https://github.com/jermp`

Personal Information

Place of birth	Bagno a Ripoli (Florence), Italy
Date of birth	13 July 1990

Research Interests

Keywords	Data Structures, Data Compression, Indexing, Efficiency
Short description	The research activity focuses on devising compressed data structures and algorithms to index and search large quantities of data. The proposed solutions are available as research papers and optimized software libraries (written in C++).
Studied problems	Inverted index compression, indexing and estimation of language models, indexing of semantic relations, bitmap compression, rank/select queries, prefix-sums, minimal perfect hashing.

Education

01/11/2015 – 31/10/2018 **PhD in Computer Science (INF/01)**.
- University of Pisa, Pisa, Italy
- Thesis: *Space- and Time-Efficient Data Structures for Massive Datasets*. Defended on 08/03/2019.
- Grade: Excellent.
- Supervisor: Rossano Venturini (`http://pages.di.unipi.it/rossano`)

2012 – 2014 **Master Degree in Computer Science & Networking (class LM18)**.
- University of Pisa and Scuola Superiore Sant'Anna, Pisa, Italy
- Thesis: *Dynamic Elias-Fano Encoding*. Defended on 06/03/2015.
- Grade: 110/110 *summa cum laude*.
- Supervisor: Rossano Venturini (`http://pages.di.unipi.it/rossano`)

2009 – 2012 **Bachelor Degree in Computer Engineering (class L08)**.
- University of Florence, Florence, Italy
- Thesis: *Quantum Computation & Grover's Algorithm*. Defended on 09/10/2012.
- Grade: 110/110 *summa cum laude*.
- Supervisor: Gabriele Vezzosi (`http://www.dma.unifi.it/~vezzosi`)

2004 – 2009 **High School Diploma**.
- Liceo Scientifico Statale Guido Castelnuovo, Florence, Italy
- Grade: 100/100.

Professional Employment

15/03/2021 – present
Postdoctoral Research Fellow in Computer Science.
- Institute of Science and Information Technologies "A. Faedo" (ISTI), National Research Council of Italy (CNR), Pisa, Italy
- Grant issued on the european project ACCORDION with theme "Tecniche algoritmiche per compressione, indicizzazione e ricerca di grandi quantità di dati e progettazione di relative librerie software open source" (Protocollo n. 0000901/2021, 09/03/2021, ISTI 004/2021 - PI).

01/11/2018 – 28/02/2021
Postdoctoral Research Fellow in Computer Science.
- Institute of Science and Information Technologies "A. Faedo" (ISTI), National Research Council of Italy (CNR), Pisa, Italy
- Grant issued on the european project BIGDATAGRAPES with theme "Compressione, indicizzazione e ricerca su grandi collezioni di dati semantici" (Protocollo n. 0003847, 24/10/2018, ISTI 014/2018 - PI). The research activity conducted for this project focused on the design of time and space efficient indexing data structures for structured and unstructured data such as RDF graphs and text documents, including compression techniques for Big data management that support a broad range of analytical queries over arbitrary data dimensions. In particular, it resulted in the development of
  - a novel compressor for inverted indexes;
  - a novel compressed index for RDF data.

  Both results have been published in IEEE Transactions on Knowledge and Data Engineering (TKDE), as the papers
  - "On Optimally Patitioning Variable-Byte Codes"
  - "Compressed Indexes for Fast Search of Semantic Data"

  with the corresponding C++ libraries available on GitHub at
  - `https://github.com/jermp/opt_vbyte`
  - `https://github.com/jermp/rdf_indexes`
- During this period, I also kept doing my own independent research on algorithms and data structures. The studied problems involved: inverted indexes (CSUR 2020), prefix-sums (SPE 2020), bitmap compression (DCC 2021), query auto-completion (SIGIR 2020), rank and select indexes for bitmaps (INFOSYS 2021), and minimal perfect hashing (SIGIR 2021). All works have been published in top-tier conferences/journals. The corresponding software libraries are available from my GitHub page.

01/06/2017 – 31/10/2018
Software Developer.
- Institute of Science and Information Technologies "A. Faedo" (ISTI), National Research Council of Italy (CNR), Pisa, Italy
- Worked for the european project *Large-scale Indie Gaming Analytics* (LIGA). The LIGA project aimed at designing and developing a proof-of-concept platform, customized for the 3D-KUMO use case (`https://www.3dkumo.com`), to analyze the huge volume volume of data generated by the users of indie games (i.e., players) on web portals and social networks.

01/11/2015 – 31/10/2018
PhD Student in Computer Science.
- University of Pisa, Pisa, Italy
- Thesis: *Space- and Time-Efficient Data Structures for Massive Datasets*.
- Supervisor: Rossano Venturini (`http://pages.di.unipi.it/rossano`)
- Worked on inverted indexing, compressed language models, and tries. The thesis is based on the following publications.
  - "On Optimally Partitioning Variable-Byte Codes" (TKDE 2019)
  - "Handling Massive N-Gram Datasets Efficiently" (TOIS 2019)
  - "Fast Dictionary-based Compression for Inverted Indexes" (WSDM 2019)
  - "Inverted Index Compression" (EBDT 2018)
  - "Efficient Data Structures for Massive N-Gram Datasets" (SIGIR 2017)
  - "Dynamic Elias-Fano Representation" (CPM 2017)
  - "Clustered Elias-Fano Indexes" (TOIS 2017)

01/05/2018 – 01/10/2018 **Visiting PhD Student**.
- The University of Melbourne, School of Computing and Information Systems, Melbourne, Australia
- Supervisor: Alistair Moffat (`https://people.eng.unimelb.edu.au/ammoffat`)
- Worked on fast dictionary-based decoding of compressed inverted index data, which resulted in the following publication: "Fast Dictionary-based Compression for Inverted Indexes" (WSDM 2019).

01/04/2018 – 30/04/2018 **Visiting PhD Student**.
- RIKEN Advanced Intelligence Project (AIP), Tokyo, Japan
- Supervisor: Yasuo Tabei (`https://sites.google.com/site/yasuotabei`)
- Worked on various problems, such as, string similarity search, trie indexing, rank/select indexes, and sparse matrix multiplication. Reference publication: "Rank/Select Queries over Mutable Bitmaps" (INFOSYS 2021).

01/04/2015 – 01/07/2015 **Software Engineer Intern at IBM**.
- Rome, Italy
- Supervisor: Alessio Fioravanti
- Worked on the design of the IBM Customer Partnership (Web) Portal for the management of IBM customers and projects.

Teaching Experience

02/2020 – 06/2020 Teacher for *Algorithmics and Laboratory - Corso B, code 008AA*, Bachelor Degree in Computer Science, University of Pisa, Italy

02/2019 – 06/2019 Assistant for *Algorithmics and Laboratory - Corso A, code 008AA*, Bachelor Degree in Computer Science, University of Pisa, Italy

09/2018 – 12/2018 Assistant for *Competitive Programming and Contests, code 645AA*, Master Degree in Computer Science, University of Pisa, Italy

09/2017 – 12/2017 Assistant for *Competitive Programming and Contests, code 645AA*, Master Degree in Computer Science, University of Pisa, Italy

09/2016 – 12/2016 Teacher for *Algorithmics and Laboratory - Corso di recupero, code 008AA*, Bachelor Degree in Computer Science, University of Pisa, Italy

02/2016 – 06/2016 Assistant for *Algorithmics and Laboratory - Corso A, code 008AA*, Bachelor Degree in Computer Science, University of Pisa, Italy

Awards and Grants

2020 *Young Researcher Award* issued by ISTI-CNR.

2017 *SIGIR Student Travel Grant* issued by ACM SIGIR.

2015 *PhD Scholarship* issued by the University of Pisa, Department of Computer Science.

2015 *Master Degree Award: Best Performance a.y. 2013/2014* issued by Scuola Superiore Sant'Anna.

2015 *Best Master Thesis Award in Theoretical Computer Science*, issued by the Italian chapter of the European Association for Theoretical Computer Science (EATCS).

Publications

2021 Giulio Ermanno Pibiri and Roberto Trani, *Parallel and External-Memory Construction of Minimal Perfect Hash Functions with PTHash*. CoRR, `https://arxiv.org/abs/2106.02350`, pages 12.

2021 Giulio Ermanno Pibiri and Roberto Trani, *PTHash: Revisiting FCH Minimal Perfect Hashing*. ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 10.

2021 Giulio Ermanno Pibiri and Shunsuke Kanda, *Rank/Select Queries over Mutable Bitmaps*. Information Systems (INFOSYS), pages 21.

2021 Raffaele Perego, Giulio Ermanno Pibiri and Rossano Venturini, *Compressed Indexes for Fast Search of Semantic Data*. IEEE International Conference on Data Engineering (ICDE), pages 2.

2021 Giulio Ermanno Pibiri, *Fast and Compact Set Intersection through Recursive Universe Partitioning*. IEEE Data Compression Conference (DCC), pages 10.

2020 Giulio Ermanno Pibiri and Rossano Venturini, *Techniques for Inverted Index Compression*. ACM Computing Surveys (CSUR), pages 36.

2020 Giulio Ermanno Pibiri and Rossano Venturini, *Practical Trade-Offs for the Prefix-Sum Problem*. Software: Practice and Experience (SPE), pages 29.

2020 Simon Gog, Giulio Ermanno Pibiri and Rossano Venturini, *Efficient and Effective Query Auto-Completion*. ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 10.

2020 Giulio Ermanno Pibiri and Rossano Venturini, *Succinct Dynamic Ordered Sets with Random Access*. CoRR, `https://arxiv.org/abs/2003.11835`, pages 15.

2020 Raffaele Perego, Giulio Ermanno Pibiri and Rossano Venturini, *Compressed Indexes for Fast Search of Semantic Data*. IEEE Transactions on Knowledge and Data Engineering (TKDE), pages 12.

2019 Giulio Ermanno Pibiri. *On Implementing the Binary Interpolative Coding Algorithm*. Tech Report, 8 pages.

2019 Giulio Ermanno Pibiri. *Space- and Time-Efficient Data Structures for Massive Datasets*. Ph.D. Thesis, University of Pisa, 210 pages.

2019 Giulio Ermanno Pibiri and Rossano Venturini, *On Optimally Partitioning Variable-Byte Codes*. IEEE Transactions on Knowledge and Data Engineering (TKDE), pages 12.

2019 Giulio Ermanno Pibiri and Rossano Venturini, *Handling Massive N-Gram Datasets Efficiently*. ACM Transactions on Information Systems (TOIS), pages 41.

2019 Giulio Ermanno Pibiri, Matthias Petri, Alistair Moffat, *Fast Dictionary-based Compression for Inverted Indexes*. ACM Conference on Web Search and Data Mining (WSDM), pages 9.

2018 Giulio Ermanno Pibiri and Rossano Venturini, *Inverted Index Compression*. Encyclopedia of Big Data Technologies (EBDT), pages 8.

2017 Giulio Ermanno Pibiri and Rossano Venturini, *Efficient Data Structures for Massive N-Gram Datasets*. ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 10.

2017 Giulio Ermanno Pibiri and Rossano Venturini, *Dynamic Elias-Fano Representation*. Annual Symposium on Combinatorial Pattern Matching (CPM), pages 14.

2017 Giulio Ermanno Pibiri and Rossano Venturini, *Clustered Elias-Fano Indexes*. ACM Transactions on Information Systems (TOIS), volume 2, pages 33.

Software

GitHub profile `https://github.com/jermp`

Data Structures

At my GitHub profile you can find efficient C++ implementations of the following data structures (see also related publications):

- Inverted Indexes (TOIS 2017, TKDE 2019, WSDM 2019, SIGIR 2020, CSUR 2020)
- Tries (SIGIR 2017, TOIS 2019, TKDE 2020)
- Compressed Bitmaps (DCC 2021)
- Mutable Bitmaps with Rank/Select (INFOSYS 2021)
- Segment-Trees and Fenwick-Trees (SPE 2020)
- Minimal Perfect Hash Functions (SIGIR 2021)

A more detailed list follows below.

pthash PTHash: Fast and compact minimal perfect hash functions.
Reference publications: SIGIR 2021, arXiv 2106.02350 2021.

rank_select Mutable bitmaps with support for Rank and Select queries.
Reference publication: INFOSYS 2021.

psds A range of tree-shaped data structures for maintaining prefix-sums, including:

- binary Segment-Tree (top-down and bottom-up),
- b-ary Segment-Tree,
- Fenwick-Tree,
- b-ary Fenwick-Tree,
- blocked Fenwick-Tree,
- truncated Fenwick-Tree.

Reference publication: SPE 2020.

autocomplete Efficienct and effective autocompletion framework, based on forward/inverted indexes, succinct RMQ, and string dictionaries (Front-Coding and tries).
Reference publication: SIGIR 2020.

2i_bench A benchmarking suite for inverted index data structures, featuring the following compressors:

- Elias-Fano and partitioned Elias-Fano,
- Opt-PFor-Delta,
- Binary Interpolative,
- QMX,
- Simple family,
- Variable-Byte family, including Opt-VByte,
- Gamma, Delta, Rice, Zeta,
- DINT.

Reference publication: CSUR 2020.

interp An efficient implementation of the Binary Interpolative Coding algorithm.

s_indexes Compressed bitmap indexes that support fast intersection and union.
Reference publication: DCC 2021.

rdf_indexes Trie-based indexes for semantic data like RDF triples.
Reference publication: TKDE 2020.

dint DINT: fast and compact dictionary-based decoder for inverted lists.
Reference publication: WSDM 2019.

opt_vbyte Optimal partitioning of inverted lists compressed using binary vectors and point-wise encoders, like Variable-Byte.
Reference publication: TKDE 2019.

tongrams Fast language model queries and estimation in compressed space.
Reference publications: SIGIR 2017, TOIS 2019.

clustered_indexes Clustered Elias-Fano inverted indexes.
Reference publication: TOIS 2017.

Miscellanea

essentials A C++ library providing essential core utilities for data structure design and benchmarking. More precisely:

- benchmarking facilities, including: messages displaying local time, configurable timer class, function to prevent code elision by compiler, simple creation and printing of json documents;
- functions to serialize-to and load-from disk data structures,
- functions to compute the numbr of bytes consumed by data structures,
- support for creating, removing, and iterate inside directories,
- transparent support for contiguous memory allocation.

cmd_line Command line parser for C++17. It offers all handy features in just 150 lines of code.

mm_file A self-contained, header-only, implementation of memory-mapped files in C++ for both reading and writing.

Talks

04/2020 *Compressed Indexes for Fast Search of Semantic Data*. ICDE conference presentation. Virtual event.

03/2020 *Fast and Compact Set Intersection through Recursive Universe Partitioning*. DCC conference presentation. Virtual event.

04/03/2020 *Efficiency for Real-World Applications* Seminar. ISTI-CNR. Virtual event.

27/07/2020 *Efficient and Effective Query Auto-Completion*. SIGIR conference presentation. Virtual event.

17/09/2019 *Compressed Indexes for Fast Search of Semantic Data*. IIR conference presentation. Department of Information Engineering, Padova, Italy.

07/06/2019 *Ordered Set Problems*. Seminar. ISTI-CNR, Pisa, Italy.

08/03/2019 *Space- and Time-Efficient Data Structures*. PhD thesis defense. The University of Pisa, Pisa, Italy.

12/02/2019 *Fast Dictionary-based Compression for Inverted Indexes*. WSDM conference presentation. Melbourne Exhibition Center, Melbourne, Australia.

01/02/2019 *Indexing Compressed Data for Fast Retrieval*. Talk. The University of Pisa, Pisa, Italy.

15/11/2018 *Space- and Time-Efficient Data Structures*. PhD research results. The University of Pisa, Pisa, Italy.

29/10/2018 *Effective Web Graph Representations*. Seminar. The University of Pisa, Pisa, Italy.

17/05/2018 *On Optimally Partitioning Variable-Byte Index Data*. Seminar. RMIT University, Melbourne, Australia.

10/04/2018 *Elias-Fano Encoding: a powerful tool for data structure design*. Seminar. RIKEN AIP, Tokyo, Japan.

10/10/2017 *Space- and Time-Efficient Data Structures*. PhD research results. The University of Pisa, Pisa, Italy.

10/08/2017 *Efficient Data Structures for Massive N-Gram Datasets*. SIGIR conference presentation. Keio Plaza Hotel, Tokyo, Japan.

06/07/2017 *Dynamic Elias-Fano Representation*. CPM conference presentation. University Library of Warsaw, Warsaw, Poland.

06/06/2017 *Efficient Data Structures for Massive N-Gram Datasets*. IIR conference presentation. Università della Szizzera Italiana, Lugano, Switzerland.

17/10/2016 *Space- and Time-Efficient Data Structures* PhD thesis proposal. The University of Pisa, Pisa, Italy.

21/06/2016 *Elias-Fano Encoding: succinct representation of monotone integer sequences with search operations*. Seminar. The University of Pisa, Pisa, Italy.

Professional Activities

2021 Member of the Program Committee of the 30-th edition of the International ACM Conference on Information and Knowledge Management (CIKM 2021).

2021 Member of the Program Committee of the 44-th edition of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021).

2021 Member of the Program Committee of the 43-rd European Conference on Information Retrieval (ECIR 2021).

2020 Member of the Program Committee of the 14-th International ACM Conference on Web Search and Data Mining (WSDM 2021).

2020 Member of the Program Committee of the 29-th edition of the International ACM Conference on Information and Knowledge Management (CIKM 2020).

2020 Member of the Organizing Committee of the 28-th edition of the Annual European Symposium on Algorithms (ESA 2020).

2020 Member of the Program Committee of the 43-rd edition of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020).

2019 Member of the Program Committee of the 42-nd edition of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019).

2019 Member of the Organizing Committee of the 30-th edition of the International Symposium on Combinatorial Pattern Matching (CPM 2019).

2018 Member of the Program Committee of the 2-nd edition of the Workshop on Knowledge Graphs and Semantics for Text Retrieval and Analysis (KG4IR), in conjuction with ACM SIGIR 2018.

2017 Member of the Organizing Committee of the 24-th edition of the International Symposium on String Processing and Information Retrieval (SPIRE 2017).

2016 Student volunteer for the organization of the 39-th edition of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016).

2016 – present Anonymous reviewer for the following conferences/journals: SIGIR, WSDM, WWW, CIKM, TALG, ESA, INFOSYS, SPE, CPM, DCC, ECIR, SPIRE, Algorithmica.

Languages

Italian **Native** *CEFR level: C2*

English **Fluent** *CEFR level: C1*

2018 **TOEFL iBT in English**.
100 (HIGH level)

2008 **First Certificate in English (Level B2)**.
University of Cambridge, Cambridge, United Kingdom