UNIVERSITÀ DI PISA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Dottorato di Ricerca in Ingegneria dell'Informazione

Doctoral Course

"Theory & Practice of Data Compression"

Giulio Ermanno Pibiri
ISTI-CNR, Pisa, Italy
giulio.ermanno.pibiri@isti.cnr.it

Short Abstract: The need of storing data in compact form is increasingly important for the ever-growing rate of data produced on a daily basis. To keep up with this data explosion phenomenon, data compression is a mandatory step to deliver good quality of service in concrete applications. In this introductory course you will learn about fundamental data compression algorithms that are all widely adopted in practice by tools that we use every day, like filesystems, computer networks, search engines, databases, and so on. These algorithms have now become indispensable knowledge across many fields in computing, including Information Retrieval, Machine Learning, Natural Language Processing, Applied Physics, and Bioinformatics. To better grasp the beauty behind data compression, we will also learn how to implement some of these algorithms in C++ through several "hands-on" sessions.

Course Contents in brief:

1. Introduction
    - What is and Why Data Compression?
    - Motivations
    - Technological Limitations: Memories and Hierarchies
    - Applications
    - Basic Notions: Entropy, Information-Content, Data-Redundancy, Compression-Ratio

2. Integer Codes
    - Basic Notions: Distributions, Kraft-McMillan Inequality

- Run-Length Encoding, Gamma, Delta, Golomb, Rice, Zeta, Fibonacci, Variable-Byte
- Encoding/Decoding of Prefix-Free Codes

3. Lab Session 1 on Integer Codes

4. List Compressors
- Basic Notions: Combinatorial Lower Bound
- Binary Packing, Simple, PForDelta, Elias-Fano, Interpolative, Directly-Addressable, Hybrid
- Inverted Indexes and Social Networks

5. Lab Session 2 on List Compressors

6. Statistical Compressors
- Shannon-Fano, Huffman, Arithmetic Coding, Asymmetric Numeral Systems

7. Dictionary-Based Compressors
- LZ77, LZ78, LZW, variants: gzip, LZO, Zstd

Total # of hours of lecture: 20

References:

[1] Robert Sedgewick and Kevin Wayne. 2011. Algorithms (4-th edition). Addison-Wesley Professional, ISBN 0-321-57351-X.

[2] David Salomon. 2007. Variable-length Codes for Data Compression. Springer Science & Business Media, ISBN 978-1-84628-959-0.

[3] Alistair Moffat and Andrew Turpin. 2002. Compression and coding algorithms. Springer Science & Business Media, ISBN 978-1-4615-0935-6.

CV of the Teacher

Giulio Ermanno Pibiri (http://pages.di.unipi.it/pibiri) is a Post-Doctoral Research Fellow in Computer Science, currently affiliated to the HPC-Lab, ISTI-CNR (Pisa, Italy). He obtained a PhD in Computer Science in 2019, from the University of Pisa. His research activity focuses on data indexing, i.e., compressing data to make queries efficient over large-scale datasets. He has extensive experience in low-level programming (C/C++) and software optimization. Starting from 2017, he authored more than 15 research papers on compressed data structures in top-tier venues like SIGIR, WSMD, TOIS, and TKDE. He is part of the program committees of the most prestigious conferences in Information Retrieval, such as ACM SIGIR and ACM WSDM, and has organised some, like ESA 2020, CPM 2019, and SPIRE 2017. He taught courses about programming and algorithms to graduate and undergraduate students at the University of Pisa.

Room and Schedule

Room: *Aula Riunioni del Dipartimento di Ingegneria dell'Informazione, Via G. Caruso 16, Pisa – Ground Floor*

Schedule: 11/04 - 15/04, 14:00 - 18:00