Algorithms and data structures for massive data: what's next?

Paolo Ferragina

Dipartimento di Informatica
University of Pisa, Italy
ferragina@di.unipi.it

Abstract. In this talk I'll survey the last 30 years of data-structure design, showing that the storage and indexing of large datasets has led algorithm and software developers to optimize, in the course of these years, several computational resources, such as time, space, I/Os, compression, just to name a few. One of the key results of this impressive flow of research has been that, nowadays, it is known how to index almost any (complex) data type in compressed space and to efficiently support various kinds of query operations. In many cases, the asymptotic performance obtainable over the compressed data is equivalent to the one achievable over the original (and uncompressed) raw data; sometimes, compression induces even a speed-up in practice because of a better use of memory/IO/CPU resources.
Recently, the advent of (big) data centers and the ubiquitous use of mobile devices, has raised the attention toward another resource: energy efficiency. This triggered a significant amount of research in all areas of IT, leading to believe that improvements in the energy efficiency of computing devices will be much more dramatic, and eventually have much greater impact, than in other areas of technology. However, the average power consumption and computation rates of computing devices are intricately tied together, making it difficult to speak of power complexity in isolation. So the next challenge will be, in my opinion, to design algorithms and data structures which optimize, or trade in a principled way, various computational resources simultaneously. This is what system engineers are addressing everyday by means of proper heuristics. But we will argue in this talk that this design is sophisticated to be done in a principled way, and needs a joint effort with other CS fields, such as Operational Research and Graph Theory.