Combining Text Compression and String Matching: The Miracle of Self-Indexing

Gonzalo Navarro*

Department of Computer Science, University of Chile.
Blanco Encalada 2120, Santiago, Chile. gnavarro@dcc.uchile.cl

This decade has witnessed the raise of what I consider the most important breakthrough of modern times in text compression and indexed string matching. *Self-indexing* is the mechanism by which a text is simultaneously compressed and indexed, so that the self-index occupies space close to that of the compressed text, provides random access to any part of it, and in addition supports efficient indexed pattern matching. Thus a self-index can replace the text by a compressed version with enhanced search functionalities. Self-indexing builds on a large base of *compressed data structures*, which is another fascinating algorithmic area that has appeared two decades ago with the aim of obtaining compact representations of classical data structures. Although they usually require more instructions than their classical counterparts to operate, they can benefit from the memory hierarchy. This is particularly noticeable when they can operate in main memory in cases where the classical structures require disk storage.

My aim in this talk is to present a thin "vertical" slice of this construction, so that there is time to visualize in sufficent detail a complete solution from the basics to the final result. I will start with a plain and a compressed solution to provide *rank* on bitmaps, a simple operation of counting the number of 1s up to a given position, with a surprising number of applications. I will then introduce *wavelet trees*, which constitute a sort of self-index for sequences, supporting operation *rank* for the alphabet symbols. Then I will explain the Burrows-Wheeler Transform and the FM-index concept, which coupled with wavelet trees offer a fully-functional self-index. Finally, I will show how this simple combination is able of achieving high-order compression of a text, and will give some insights on recent work around indexing highly repetitive sequence collections, such as DNA and protein databases, versioned data, and temporal text databases. I will conclude by posing some open challenges.

* Partially funded by Millennium Institute for Cell Dynamics and Biotechnology (ICDB), Grant ICM P05-001-F, Mideplan, Chile.

Gonzalo Navarro: Combining Text Compression and String Matching: The Miracle of Self-Indexing, p. 1.
Proceedings of PSC 2009, Jan Holub and Jan Žďárek (Eds.), ISBN 978-80-01-04403-2