Self-Indexing XML

Gonzalo Navarro

Universidad de Chile
gnavarro@dcc.uchile.cl

Self-indexing is a technology that integrates text compression and text indexing, such that a text collection can be simultaneously compressed and indexed. The resulting representation, called a self-index of the text, takes space close to that of the compressed text, is able of reproducing any text substring, and offers indexed searching of the collection. This has been a major breakthrough in the last decade, allowing one to handle huge text collections within main memory and representing them succinctly.

In this talk I will, besides presenting the basics of this technique, discuss how it can be extended to index XML collections, where, on one hand, the text has structure and, on the other hand, we wish to support much more complex query languages, XPath at least. I will first describe two plug-and-play solutions, where a text self-index is coupled with compressed data structures that handle trees and sequences. Then I will introduce two more innovative solutions, where the compressed data structures are designed specifically for this problem. This area is full of open challenges and I will conclude by enumerating the most relevant ones.