External memory construction of full-text indexes

Andreas Crauser and Paolo Ferragina

Max Planck Institut für Informatik
Saarbrücken, Germany
email:{crauser,paolo}@mpi-sb.mpg.de

Abstract

The construction of full-text indexes on very large text collections is nowadays a hot problem. The suffix array [Manber-Myers, 1993] is one of the most interesting full-text indexing data structures due to its simplicity, space efficiency and powerful/fast search operations supported. In this paper we analyze, both theoretically and experimentally, the I/O-complexity and the working space of some state-of-the art algorithms for constructing suffix arrays on large text collections. Moreover, we introduce three new efficient algorithms which are flexible enough to be adapted to work also in distributed memory systems without any loss in efficiency. We perform a set of experiments based on different text collections (English texts and Amino-acid sequences) and give a precise hierarchy of suffix-array construction algorithms according to their working-space vs. construction-time tradeoff. At the best of our knowledge, this is the first study which provides a wide spectrum of possible approaches to the construction of suffix arrays in external memory, and thus it should be helpful to anyone who is interested in building full-text indexes on very large text collections.