Ph: 9585554590, 9585554599
Email: support@salemsmartreach.com
URL: www.salemsmartreach.com

Large-Scale Pattern Search Using Reduced-Space On-Disk Suffix Arrays

Abstract:

The suffix array is an efficient data structure for in-memory pattern search. Suffix arrays can also be used for external-memory pattern search, via two-level structures that use an internal index to identify the correct block of suffix pointers. In this paper, we describe a new two-level suffix array-based index structure that requires significantly less disk space than previous approaches. Key to the saving is the use of disk blocks that are based on prefixes rather than the more usual uniform-sampling approach, allowing reductions between blocks and subparts of other blocks. We also describe a new in-memory structure—the condensed BWT— and show that it allows common patterns to be resolved without access to the text. Experiments using 64 GB of English web text on a computer with 4 GB of main memory demonstrate the speed and versatility of the new approach. For this data, the index is around one-third the size of previous two-level mechanisms; and the memory footprint of as little as 1% of the text size means that queries can be processed more quickly than is possible with a compact FM-INDEX.