Lempel?Ziv compressed structures for document retrieval

Ferrada, Héctor

Navarro, Gonzalo

Document retrieval structures index a collection of string documents, to retrieve those that are relevant to query strings p: document listing retrieves all documents where p appears; top-k retrieval retrieves the k most relevant of those. Classical structures use too much space in practice. Most current research uses compressed suffix arrays, but fast indices still use 17?21 bpc (bits per character), whereas small ones take milliseconds per returned answer. We present the first document retrieval structures based on Lempel?Ziv compression, precisely LZ78. Our structures use 7?10 bpc and dominate a large part of the space/time tradeoffs. They also enable more efficient partial or approximate answers: our document listing outputs the first 75%?80% of the answers at a rate of one per microsecond; for top-k retrieval we return a result of 90% quality at the same rate and using just 4?6 bpc. This outperforms current indices by a wide margin.