Munro, J. Ian; **Navarro, Gonzalo**; **Nielsen, Jesper Sindahl**; **Shah, Rahul**; **Thankachan, Sharma V.**

Top-k term-proximity in succinct space. (English) Zbl 1366.68039

Ahn, Hee-Kap (ed.) et al., Algorithms and computation. 25th international symposium, ISAAC 2014, Jeonju, Korea, December 15–17, 2014. Proceedings. Cham: Springer (ISBN 978-3-319-13074-3/pbk; 978-3-319-13075-0/ebook). Lecture Notes in Computer Science 8889, 169-180 (2014).

Summary: Let $\mathcal{D} = \{\mathsf{T}_1, \mathsf{T}_2, \ldots, \mathsf{T}_D\}$ be a collection of D string documents of n characters in total, that are drawn from an alphabet set $\Sigma = [\sigma]$. The *top-k document retrieval* problem is to preprocess $\mathcal{D}$ into a data structure that, given a query $(P[1..p], k)$, can return the k documents of $\mathcal{D}$ most relevant to pattern P. The relevance is captured using a predefined ranking function, which depends on the set of occurrences of P in T_d. For example, it can be the term frequency (i.e., the number of occurrences of P in T_d), or it can be the term proximity (i.e., the distance between the closest pair of occurrences of P in T_d), or a pattern-independent importance score of T_d such as PageRank. Linear space and optimal query time solutions already exist for this problem. Compressed and compact space solutions are also known, but only for a few ranking functions such as term frequency and importance. However, space efficient data structures for term proximity based retrieval have been evasive. In this paper we present the first sub-linear space data structure for this relevance function, which uses only $o(n)$ bits on top of any compressed suffix array of $\mathcal{D}$ and solves queries in time $O((p+k)\operatorname{polylog} n)$.

For the entire collection see [Zbl 1318.68007].

MSC:

68P20 Information storage and retrieval of data
68P05 Data structures

Full Text: DOI