Munro, J. Ian; Navarro, Gonzalo; Shah, Rahul; Thankachan, Sharma V.

Ranked document selection. (English) Zbl 1435.68078

Summary: Let \( \mathcal{D} \) be a collection of string documents of \( n \) characters in total. The top-\( k \) document retrieval problem is to preprocess \( \mathcal{D} \) into a data structure that, given a query \((P, k)\), can return the \( k \) documents of \( \mathcal{D} \) most relevant to pattern \( P \). The relevance of a document \( d \) for a pattern \( P \) is given by a predefined ranking function \( \omega(P, d) \). Linear space and optimal query time solutions already exist for this problem. In this paper we consider a novel problem, document selection, in which a query \((P, k)\) aims to report the \( k \)th document most relevant to \( P \) (instead of reporting all top-\( k \) documents). We present a data structure using \( O(n \log^* n) \) space, for any constant \( \epsilon > 0 \), answering selection queries in time \( O(\log k / \log \log n) \), and a linear-space data structure answering queries in time \( O(\log k) \), given the locus node of \( P \) in a (generalized) suffix tree of \( \mathcal{D} \). We also prove that it is unlikely that a succinct-space solution for this problem exists with poly-logarithmic query time, and that \( O(\log k / \log \log n) \) is indeed optimal within \( O(n \text{polylog } n) \) space for most text families. Finally, we present some additional space-time trade-offs exploring the extremes of those lower bounds.

MSC:

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

Keywords:
document indexing; top-k document retrieval; succinct data structures

Full Text: DOI

References:
