Munro, J. Ian; **Navarro, Gonzalo**; **Shah, Rahul**; **Thankachan, Sharma V.**
Ranked document selection. (English) Zbl 1435.68078
Theor. Comput. Sci. 812, 149-159 (2020).

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 $w(P, d)$. 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 $O(n \log^{\epsilon} n)$ 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 $O(n \operatorname{polylog} n)$ space for most text families. Finally, we present some additional space-time trade-offs exploring the extremes of those lower bounds.

MSC:

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

Keywords:

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

Full Text: DOI

References:

[1] Biswas, S.; Ganguly, A.; Shah, R.; Thankachan, S. V., Ranked document retrieval for multiple patterns, Theor. Comput. Sci., 746, 98-111 (2018) · Zbl 1408.68052

[2] Biswas, S.; Ku, T.-H.; Shah, R.; Thankachan, S. V., Position-restricted substring searching over small alphabets, J. Discret. Algorithms, 46-47, 36-39 (2017) · Zbl 1375.68230

[3] Chan, T. M.; Durocher, S.; Green Larsen, K.; Morrison, J.; Wilkinson, B. T., Linear-space data structures for range mode query in arrays, Theory Comput. Syst., 55, 4, 719-741 (2014) · Zbl 1319.68062

[4] Chan, Timothy M.; Green Larsen, Kasper; Patrascu, Mihai, Orthogonal range searching on the ram, revisited, (Proceedings of the 27th ACM Symposium on Computational Geometry. Proceedings of the 27th ACM Symposium on Computational Geometry, Paris, France, June 13-15 (2011)), 1-10 · Zbl 1283.68139

[5] Chan, Timothy M.; Wilkinson, Bryan T., Adaptive and approximate orthogonal range counting, ACM Trans. Algorithms, 12, 4, 45:1-45:15 (2016) · Zbl 1421.68017

[6] Crochemore, M.; Iliopoulos, C. S.; Kubica, M.; Sohel Rahman, M.; Tischler, G.; Walen, T., Improved algorithms for the range next value problem and applications, Theor. Comput. Sci., 434, 23-34 (2012) · Zbl 1244.68031

[7] Fredman, M. L.; Willard, D. E., Surpassing the information theoretic barrier with fusion trees, J. Comput. Syst. Sci., 47, 424-436 (1993) · Zbl 0795.68049

[8] Grossi, R.; Orlandi, A.; Raman, R.; Rao, S. S., More haste, less waste: lowering the redundancy in fully indexable dictionaries, (Proc. 26th International Symposium on Theoretical Aspects of Computer Science (STACS) (2009)), 517-528 · Zbl 1236.68064

[9] Hon, W.-K.; Patil, M.; Shah, R.; Thankachan, S. V.; Vitter, J. S., Indexes for document retrieval with relevance, (Space-Efficient Data Structures, Streams, and Algorithms (2013)), 351-362 · Zbl 1394.68127

[10] Hon, W.-K.; Patil, M.; Shah, R.; Wu, S.-B., Efficient index for retrieving top-k most frequent documents, J. Discret. Algorithms, 8, 4, 402-417 (2010) · Zbl 1215.68095

[11] Hon, W.-K.; Shah, R.; Thankachan, S. V.; Vitter, J. S., On position restricted substring searching in succinct space, J. Discret. Algorithms, 17, 109-114 (2012) · Zbl 1267.68102

[12] Hon, W.-K.; Shah, R.; Thankachan, S. V.; Vitter, J. S., Space-efficient frameworks for top-k string retrieval, J. ACM, 61, 2, 9 (2014) · Zbl 1295.68230

[13] Hon, W.-K.; Shah, R.; Vitter, J. S., Space-efficient framework for top-k string retrieval problems, (Proc. 50th IEEE Symposium

on Foundations of Computer Science (FOCS) (2009)), 713-722 · Zbl 1292.68182

[14] Hon, Wing-Kai; Thankachan, Sharma V.; Shah, Rahul; Vitter, Jeffrey Scott, Faster compressed top-k document retrieval, (2013 Data Compression Conference, DCC 2013. 2013 Data Compression Conference, DCC 2013, Snowbird, UT, USA, March 20-22 (2013)), 341-350

[15] Jørgensen, A. G.; Larsen, K. G., Range selection and median: tight cell probe lower bounds and adaptive data structures, (Proc. 22nd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA) (2011)), 805-813 · Zbl 1373.68196

[16] Lee, D. T.; Wong, C. K., Worst-case analysis for region and partial region searches in multidimensional binary search trees and balanced quad trees, Acta Inform., 9, 23-29 (1977) · Zbl 0349.68016

[17] Mäkinen, V.; Navarro, G., Position-restricted substring searching, (Proc. 7th Latin American Symposium on Theoretical Informatics (LATIN) (2006)), 703-714 · Zbl 1145.68392

[18] Munro, J. I., Tables, (Proc. 16th Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS) (1996)), 37-42

[19] Munro, J. I.; Navarro, G.; Nielsen, J. S.; Shah, R.; Thankachan, S. V., Top-k term-proximity in succinct space, Algorithmica, 78, 2, 379-393 (2017) · Zbl 1370.68075

[20] Munro, J. I.; Navarro, G.; Shah, R.; Thankachan, S. V., Ranked document selection, (Proc. 15th Scandinavian Symposium on Algorithmic Theory (SWAT) (2014)), 344-356 · Zbl 1416.68064

[21] Ian Munro, J.; Navarro, Gonzalo; Sindahl Nielsen, Jesper; Shah, Rahul; Thankachan, Sharma V., Top- k term-proximity in succinct space, (Algorithms and Computation - 25th International Symposium, Proceedings. Algorithms and Computation - 25th International Symposium, Proceedings, ISAAC 2014, Jeonju, Korea, December 15-17 (2014)), 169-180 · Zbl 1366.68039

[22] Muthukrishnan, S., Efficient algorithms for document retrieval problems, (Proc. 13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA) (2002)), 657-666 · Zbl 1093.68588

[23] Navarro, G., Spaces, trees and colors: the algorithmic landscape of document retrieval on sequences, ACM Comput. Surv., 46, 4 (2014), article 52 · Zbl 1305.68078

[24] Navarro, G.; Nekrich, Y., Top-k document retrieval in optimal time and linear space, (Proc. 23rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA) (2012)), 1066-1078 · Zbl 1422.68063

[25] Navarro, G.; Nekrich, Y., Time-optimal top-k document retrieval, SIAM J. Comput., 46, 1, 89-113 (2017) · Zbl 1359.68053

[26] Navarro, G.; Sadakane, K., Fully-functional static and dynamic succinct trees, ACM Trans. Algorithms, 10, 3 (2014), article 16 · Zbl 1333.68084

[27] Navarro, G.; Thankachan, S. V., Faster top-k document retrieval in optimal space, (Proc. 20th International Symposium on String Processing and Information Retrieval (SPIRE) (2013)), 255-262

[28] Navarro, G.; Thankachan, S. V., New space/time tradeoffs for top-k document retrieval on sequences, Theor. Comput. Sci., 542, 83-97 (2014) · Zbl 1317.68049

[29] Nekrich, Y.; Navarro, G., Sorted range reporting, (Proc. 13th Scandinavian Symposium on Algorithmic Theory (SWAT) (2012)), 271-282 · Zbl 1347.68343

[30] Okajima, Y.; Maruyama, K., Faster linear-space orthogonal range searching in arbitrary dimensions, (Proc. 17th Workshop on Algorithm Engineering and Experiments (ALENEX) (2015)), 82-93 · Zbl 1430.68076

[31] Russo, L.; Navarro, G.; Oliveira, A., Fully-compressed suffix trees, ACM Trans. Algorithms, 7, 4 (2011), art. 53 · Zbl 1295.68103

[32] Shah, R.; Sheng, C.; Thankachan, S. V.; Vitter, J. S., Top-k document retrieval in external memory, (Proc. 21st Annual European Symposium on Algorithms (ESA) (2013)), 803-814 · Zbl 1394.68129

[33] Tsur, D., Top-k document retrieval in optimal space, Inf. Process. Lett., 113, 12, 440-443 (2013) · Zbl 1371.68071

[34] Weiner, P., Linear pattern matching algorithms, (Proc. 14th Annual IEEE Symposium on Switching and Automata Theory (1973)), 1-11