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

Top-k term-proximity in succinct space. (English) Zbl 1370.68075

Algorithmica 78, No. 2, 379-393 (2017).

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 $(P[1 \ldots p], k)$, can return the k documents of $\mathcal{D}$ most relevant to the 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 the general top-k document retrieval 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 $O((p+k)\,\text{polylog}\,n)$ time. We also show that scores that consist of a weighted combination of term proximity, term frequency, and document importance, can be handled using twice the space required to represent the text collection.

MSC:

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

Cited in **1** Document

Keywords:

document indexing; top-k document retrieval; ranked document retrieval; succinct data structures; compressed data structures; compact data structures; proximity search

Software:

Wumpus

Full Text: DOI

References:

[1] Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval, 2nd edn. Addison-Wesley, Reading (2011)

[2] Belazzougui, D., Navarro, G.: Alphabet-independent compressed text indexing. In: Proceedings of the 19th ESA, pp. 748-759 (2011) · Zbl 1325.68307

[3] Belazzougui, D; Navarro, G; Valenzuela, D, Improved compressed indexes for full-text document retrieval, J. Discrete Algorithms, 18, 3-13, (2013) · Zbl 1268.68075

[4] Benson, G; Waterman, M, A fast method for fast database search for all k-nucleotide repeats, Nucleic Acids Res., 22, 4828-4836, (1994)

[5] Broschart, A., Schenkel, R.: Index tuning for efficient proximity-enhanced query processing. In: INEX, pp. 213-217 (2009) · Zbl 1243.68161

[6] Büttcher, S., Clarke, C.L.A., Cormack, G.: Information Retrieval: Implementing and Evaluating Search Engines. MIT Press, Cambridge (2010) · Zbl 1211.68176

[7] de Berg, M., van Kreveld, M., Overmars, M., Schwarzkopf, O.: Computational Geometry: Algorithms and Applications, 3rd edn. Springer, Berlin (2008) · Zbl 0939.68134

[8] Ferragina, P; Grossi, R, The string B-tree: a new data structure for string search in external memory and its applications, J. ACM., 46, 236-280, (1999) · Zbl 1065.68518

[9] Ferragina, P., Manzini, G., Mäkinen, V., Navarro, G.: Compressed representations of sequences and full-text indexes. ACM Trans. Algorithms \textbf{3}(2), Art. No. 20 (2007) · Zbl 1321.68263

[10] Gagie, T; Navarro, G; Puglisi, SJ, New algorithms on wavelet trees and applications to information retrieval, Theor. Comput.

Sci., 426-427, 25-41, (2012) · Zbl 1243.68161

[11] Gusfield, D.: Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology. Cambridge University Press, Cambridge (1997) · Zbl 0934.68103

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

[13] Hon, W.-K., Shah, R., Thankachan, S.V., Vitter, J.S.: Faster compressed top-k document retrieval. In: Proceedings of the 23rd DCC, pp. 341-350 (2013)

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

[15] Hon, W.-K., Shah, R., Vitter, J.S.: Space-efficient framework for top-k string retrieval problems. In: Proceedings of the 50th FOCS, pp. 713-722 (2009) · Zbl 1292.68182

[16] Manber, U; Myers, G, Suffix arrays: a new method for on-line string searches, SIAM J. Comput., 22, 935-948, (1993) · Zbl 0784.68027

[17] Manzini, G, An analysis of the Burrows-Wheeler transform, J. ACM., 48, 407-430, (2001) · Zbl 1323.68262

[18] Munro, J.I., Navarro, G., Nielsen, J.S., Shah, R., Thankachan, S.V.: Top-k term-proximity in succinct space. In: Proceedings of the 25th ISAAC, pp. 169-180 (2014) · Zbl 1366.68039

[19] Muthukrishnan, S.: Efficient algorithms for document retrieval problems. In: Proceedings of the 13th SODA, pp. 657-666 (2002) · Zbl 1093.68588

[20] Navarro, G, Spaces, trees and colors: the algorithmic landscape of document retrieval on sequences, ACM Comput. Surv., 46, art. no. 52, (2014) · Zbl 1305.68078

[21] Navarro, G; Mäkinen, V, Compressed full-text indexes, ACM Comput. Surv., 39, art. no. 2, (2007) · Zbl 1321.68263

[22] Navarro, G., Nekrich, Y.: Top-k document retrieval in optimal time and linear space. In: Proceedings of the 23rd SODA, pp. 1066-1078 (2012)

[23] Navarro, G., Russo, L.: Fast fully-compressed suffix trees. In: Proceedings of the 24th DCC, pp. 283-291 (2014)

[24] Navarro, G., Thankachan, S.V.: Faster top-k document retrieval in optimal space. In: Proceedings of the 20th SPIRE, LNCS 8214, pp. 255-262 (2013) · Zbl 1406.68022

[25] Navarro, G., Thankachan, S.V.: Top-k document retrieval in compact space and near-optimal time. In: Proceedings of the 24th ISAAC, LNCS 8283, pp. 394-404 (2013) · Zbl 1406.68022

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

[27] Nekrich, Y., Navarro, G.: Sorted range reporting. In: Proceedings of the 13th SWAT, LNCS 7357, pp. 271-282 (2012) · Zbl 1347.68343

[28] Pătraşcu, M.: Succincter. In: Proceedings of the 49th FOCS, pp. 305-313 (2008)

[29] Raman, R; Raman, V; Srinivasa, SR, Succinct indexable dictionaries with applications to encoding k-ary trees, prefix sums and multisets, ACM Trans. Algorithms, 3, art. no. 43, (2007) · Zbl 1446.68046

[30] Schenkel, R., Broschart, A., Hwang, S.-W., Theobald, M., Weikum, G.: Efficient text proximity search. In: SPIRE, pp. 287-299 (2007) · Zbl 0784.68027

[31] Shah, R., Sheng, C., Thankachan, S.V., Vitter, J.S.: Top-k document retrieval in external memory. In: Proceedings of the 21st ESA, LNCS 8125, pp. 803-814 (2013) · Zbl 1394.68129

[32] Weiner, P.: Linear pattern matching algorithm. In: Proceedings of the 14th Annual IEEE Symposium on Switching and Automata Theory, pp. 1-11 (1973)

[33] Yan, H., Shi, S., Zhang, F., Suel, T., Wen, J.-R.: Efficient term proximity search with term-pair indexes. In: CIKM, pp. 1229-1238 (2010)

[34] Zhu, M., Shi, S., Li, M., Wen, J.-R.: Effective top-k computation in retrieving structured documents with term-proximity support. In: CIKM, pp. 771-780 (2007) · Zbl 1243.68161

[35] Zhu, M., Shi, S., Yu, N., Wen, J.-R.: Can phrase indexing help to process non-phrase queries? In: CIKM, pp. 679-688 (2008) · Zbl 1323.68262