zbMATH Open
THE FIRST RESOURCE FOR MATHEMATICS

Biswas, Sudip; **Ganguly, Arnab**; **Shah, Rahul**; **Thankachan, Sharma V.**
Ranked document retrieval for multiple patterns. (English) Zbl 1408.68052
Theor. Comput. Sci. 746, 98-111 (2018).

Index data structures for full-text search are very important in information retrieval with applications ranging from web search to various bioinformatics settings. One of the most important questions is to find the most relevant documents based on multiple patterns. The state-of-the-art practical techniques (e.g., inverted indices) do not give good performance guarantees in the worst case.

This paper represents a step in closing this gap between theory and practice by giving improved time/space bounds for this problem. The techniques build heavily on known advanced data structures in stringology, in particular suffix trees and least-common-ancestor data structures.

Reviewer: Peter Sanders (Karlsruhe)

MSC:

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

Cited in **2** Documents

Keywords:

suffix tree; suffix array; weighted ancestor query; compressed suffix array; succinct encoding

Full Text: DOI

References:

[1] Afshani, Peyman; Nielsen, Jesper Sindahl, Data structure lower bounds for document indexing problems, (43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy, (2016)), 93:1-93:15 · Zbl 1388.68026

[2] Alstrup, Stephen; Brodal, Gerth Stølting; Rauhe, Theis, Optimal static range reporting in one dimension, (Proceedings on 33rd Annual ACM Symposium on Theory of Computing, July 6-8, 2001, Heraklion, Crete, Greece, (2001)), 476-482 · Zbl 1323.68536

[3] Belazzougui, Djamal; Navarro, Gonzalo, Alphabet-independent compressed text indexing, ACM Trans. Algorithms, 10, 4, 23, (2014) · Zbl 1325.68307

[4] Biswas, Sudip; Ganguly, Arnab; Shah, Rahul; Thankachan, Sharma V., Forbidden extension queries, (35th IARCS Annual Conference on Foundation of Software Technology and Theoretical Computer Science, FSTTCS 2015, December 16-18, 2015, Bangalore, India, (2015)), 320-335 · Zbl 1366.68029

[5] Biswas, Sudip; Ganguly, Arnab; Shah, Rahul; Thankachan, Sharma V., Ranked document retrieval with forbidden pattern, (Combinatorial Pattern Matching - 26th Annual Symposium, CPM 2015, Ischia Island, Italy, June 29-July 1, 2015, Proceedings, (2015)), 77-88 · Zbl 1432.68120

[6] Biswas, Sudip; Patil, Manish; Shah, Rahul; Thankachan, Sharma V., Succinct indexes for reporting discriminating and generic words, (String Processing and Information Retrieval - 21st International Symposium, SPIRE 2014, Ouro Preto, Brazil, October 20-22, 2014, Proceedings, (2014)), 89-100 · Zbl 1330.68054

[7] Cohen, Hagai; Porat, Ely, Fast set intersection and two-patterns matching, Theoret. Comput. Sci., 411, 40-42, 3795-3800, (2010) · Zbl 1207.68270

[8] Cormen, Thomas H.; Stein, Clifford; Rivest, Ronald L.; Leiserson, Charles E., Introduction to algorithms, (2001), McGraw-Hill Higher Education · Zbl 1047.68161

[9] Durocher, Stephane; Shah, Rahul; Skala, Matthew; Thankachan, Sharma V., Linear-space data structures for range frequency queries on arrays and trees, (Mathematical Foundations of Computer Science 2013 - 38th International Symposium, MFCS 2013, Klosterneuburg, Austria, August 26-30, 2013, Proceedings, (2013)), 325-336 · Zbl 1400.68062

[10] Elias, Peter, Efficient storage and retrieval by content and address of static files, J. ACM, 21, 2, 246-260, (1974) · Zbl 0278.68028

[11] Fano, Robert Mario, On the number of bits required to implement an associative memory, (1971), Massachusetts Institute of Technology, Project MAC

[12] Ferragina, Paolo; Manzini, Giovanni, Opportunistic data structures with applications, (41st Annual Symposium on Foundations of Computer Science, FOCS 2000, 12-14 November 2000, Redondo Beach, California, USA, (2000)), 390-398

© 2021 FIZ Karlsruhe GmbH

[13] Ferragina, Paolo; Manzini, Giovanni, Indexing compressed text, J. ACM, 52, 4, 552-581, (2005) · Zbl 1323.68261

[14] Fischer, Johannes; Gagie, Travis; Kopelowitz, Tsvi; Lewenstein, Moshe; Mäkinen, Veli; Salmela, Leena; Välimäki, Niko, Forbidden patterns, (LATIN 2012: Theoretical Informatics - 10th Latin American Symposium, Arequipa, Peru, April 16-20, 2012, Proceedings, (2012)), 327-337 · Zbl 1353.68066

[15] Gawrychowski, Pawel; Lewenstein, Moshe; Nicholson, Patrick K., Weighted ancestors in suffix trees, (Algorithms - ESA 2014 - 22nd Annual European Symposium, Wroclaw, Poland, September 8-10, 2014, Proceedings, (2014)), 455-466 · Zbl 1425.68087

[16] Golynski, Alexander; Munro, J. Ian; Rao, S. Srinivasa, Rank/select operations on large alphabets: a tool for text indexing, (Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2006, Miami, Florida, USA, January 22-26, 2006, (2006)), 368-373 · Zbl 1192.68800

[17] Grossi, Roberto; Vitter, Jeffrey Scott, Compressed suffix arrays and suffix trees with applications to text indexing and string matching (extended abstract), (Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing, May 21-23, 2000, Portland, OR, USA, (2000)), 397-406 · Zbl 1296.68035

[18] Gusfield, Dan, Algorithms on strings, trees, and sequences - computer science and computational biology, (1997), Cambridge University Press · Zbl 0934.68103

[19] Hon, Wing-Kai; Shah, Rahul; Thankachan, Sharma V.; Vitter, Jeffrey Scott, String retrieval for multi-pattern queries, (String Processing and Information Retrieval - 17th International Symposium, SPIRE 2010, Los Cabos, Mexico, October 11-13, 2010, Proceedings, (2010)), 55-66

[20] Hon, Wing-Kai; Shah, Rahul; Thankachan, Sharma V.; Vitter, Jeffrey Scott, Document listing for queries with excluded pattern, (Combinatorial Pattern Matching - 23rd Annual Symposium, CPM 2012, Helsinki, Finland, July 3-5, 2012, Proceedings, (2012)), 185-195 · Zbl 1358.68093

[21] Hon, Wing-Kai; Shah, Rahul; Thankachan, Sharma V.; Vitter, Jeffrey Scott, Space-efficient frameworks for top-\textit{k} string retrieval, J. ACM, 61, 2, 9, (2014) · Zbl 1295.68230

[22] Hon, Wing-Kai; Shah, Rahul; Vitter, Jeffrey Scott, Space-efficient framework for top-\textit{k} string retrieval problems, (50th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2009, October 25-27, 2009, Atlanta, Georgia, USA, (2009)), 713-722 · Zbl 1292.68182

[23] Kopelowitz, Tsvi; Pettie, Seth; Porat, Ely, Higher lower bounds from the 3SUM conjecture, (Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016, (2016)), 1272-1287 · Zbl 1410.68147

[24] Larsen, Kasper Green; Munro, J. Ian; Nielsen, Jesper Sindahl; Thankachan, Sharma V., On hardness of several string indexing problems, (Combinatorial Pattern Matching - 25th Annual Symposium, CPM 2014, Moscow, Russia, June 16-18, 2014, Proceedings, (2014)), 242-251 · Zbl 1407.68229

[25] Matias, Yossi; Muthukrishnan, S.; Sahinalp, Süleyman Cenk; Ziv, Jacob, Augmenting suffix trees, with applications, (Algorithms - ESA '98, 6th Annual European Symposium, Venice, Italy, August 24-26, 1998, Proceedings, (1998)), 67-78

[26] Munro, J. Ian, Tables, (Foundations of Software Technology and Theoretical Computer Science, 16th Conference, Hyderabad, India, December 18-20, 1996, Proceedings, (1996)), 37-42

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

[28] Muthukrishnan, S., Efficient algorithms for document retrieval problems, (Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms, January 6-8, 2002, San Francisco, CA, USA, (2002)), 657-666 · Zbl 1093.68588

[29] Navarro, Gonzalo, Spaces, trees, and colors: the algorithmic landscape of document retrieval on sequences, ACM Comput. Surv., 46, 4, 52, (2013) · Zbl 1305.68078

[30] Navarro, Gonzalo; Mäkinen, Veli, Compressed full-text indexes, ACM Comput. Surv., 39, 1, (2007) · Zbl 1321.68263

[31] Navarro, Gonzalo; Nekrich, Yakov, Top-\textit{k} document retrieval in optimal time and linear space, (Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2012, Kyoto, Japan, January 17-19, 2012, (2012)), 1066-1077 · Zbl 1359.68053

[32] Navarro, Gonzalo; Thankachan, Sharma V., New space/time tradeoffs for top-\textit{k} document retrieval on sequences, Theoret. Comput. Sci., 542, 83-97, (2014) · Zbl 1317.68049

[33] Navarro, Gonzalo; Thankachan, Sharma V., Bottom-\textit{k} document retrieval, StringMasters 2012; 2013 Special Issue (Volume 2), J. Discrete Algorithms, 32, 69-74, (2015) · Zbl 1328.68057

[34] Patil, Manish; Thankachan, Sharma V.; Shah, Rahul; Hon, Wing-Kai; Vitter, Jeffrey Scott; Chandrasekaran, Sabrina, Inverted indexes for phrases and strings, (Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011, (2011)), 555-564

[35] Sadakane, Kunihiko; Navarro, Gonzalo, Fully-functional succinct trees, (Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010, (2010)), 134-149 · Zbl 1288.05046

[36] Shah, Rahul; Sheng, Cheng; Thankachan, Sharma V.; Vitter, Jeffrey Scott, Top-\textit{k} document retrieval in external memory, (Algorithms - ESA 2013 - 21st Annual European Symposium, Sophia Antipolis, France, September 2-4, 2013, Proceedings, (2013)), 803-814 · Zbl 1394.68129

This reference list is based on information provided by the publisher or from digital mathematics libraries. Its items are heuristically matched to zbMATH identifiers and may contain data conversion errors. It attempts to reflect the references listed in the original paper as accurately as possible without claiming the completeness or perfect precision of the matching.