Hooshmand, Sahar; **Abedin, Paniz**; **Külekci, M. Oğuzhan**; **Thankachan, Sharma V.**
I/O-efficient data structures for non-overlapping indexing. (English) Zbl 07300875
Theor. Comput. Sci. 857, 1-7 (2021).

Summary: The non-overlapping indexing problem is defined as follows: pre-process a given text $\mathsf{T}[1,n]$ of length n into a data structure such that whenever a pattern $P[1,m]$ comes as an input, we can efficiently report the largest set of non-overlapping occurrences of P in T. The best-known solution is by Cohen and Porat [ISAAC 2009]. The size of their structure is $O(n)$ words and the query time is optimal $O(m+\mathsf{nocc})$, where nocc is the output size. Later, Ganguly et al. [CPM 2015 and Algorithmica 2020] proposed a compressed space solution. We study this problem in the cache-oblivious model and present a new data structure of size $O(n\log n)$ words. It can answer queries in optimal $O(\frac{m}{B}+\log_B n+\frac{\mathsf{nocc}}{B})I/O$ operations, where B is the block size. The space can be improved to $O(n\log_{M/B} n)$ in the cache-aware model, where M is the size of main memory. Additionally, we study a generalization of this problem with an additional range $[s,e]$ constraint. Here the task is to report the largest set of non-overlapping occurrences of P in T, that are within the range $[s,e]$. We present an $O(n\log^2 n)$ space data structure in the cache-aware model that can answer queries in optimal $O(\frac{m}{B}+\log_B n+\frac{\mathsf{nocc}_{[s,e]}}{B})I/O$ operations, where $\mathsf{nocc}_{[s,e]}$ is the output size.

Reviewer: Reviewer (Berlin)

MSC:

68Q Theory of computing

Keywords:

suffix trees; data structure; string algorithms

Full Text: DOI

References:

[1] Ukkonen, E., On-line construction of suffix trees, Algorithmica, 14, 3, 249-260 (1995) · Zbl 0831.68027

[2] Weiner, P., Linear pattern matching algorithms, (14th Annual Symposium on Switching and Automata Theory. 14th Annual Symposium on Switching and Automata Theory, Iowa City, Iowa, USA, October 15-17, 1973 (1973)), 1-11

[3] Apostolico, A.; Preparata, F. P., Data structures and algorithms for the string statistics problem, Algorithmica, 15, 5, 481-494 (1996) · Zbl 0846.68023

[4] Cohen, H.; Porat, E., Range non-overlapping indexing, (Algorithms and Computation, 20th International Symposium, ISAAC 2009, Honolulu, Hawaii, USA, December 16-18, 2009, Proceedings (2009)), 1044-1053 · Zbl 1273.68097

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

[6] Keller, O.; Kopelowitz, T.; Lewenstein, M., Range non-overlapping indexing and successive list indexing, (Algorithms and Data Structures, 10th International Workshop, WADS 2007, Halifax, Canada, August 15-17, 2007, Proceedings (2007)), 625-636 · Zbl 1209.68160

[7] Nekrich, Y.; Navarro, G., Sorted range reporting, (Algorithm Theory - SWAT 2012 - 13th Scandinavian Symposium and Workshops, Helsinki, Finland, July 4-6, 2012. Proceedings (2012)), 271-282 · Zbl 1347.68343

[8] Ganguly, A.; Shah, R.; Thankachan, S. V., Succinct non-overlapping indexing, (Annual Symposium on Combinatorial Pattern Matching (2015), Springer), 185-195 · Zbl 1432.68089

[9] Ganguly, A.; Shah, R.; Thankachan, S. V., Succinct non-overlapping indexing, Algorithmica, 82, 1, 107-117 (2020) · Zbl 1436.68083

[10] Aggarwal, A.; Vitter, J. S., The input/output complexity of sorting and related problems, Commun. ACM, 31, 9, 1116-1127 (1988)

[11] Frigo, M.; Leiserson, C. E.; Prokop, H.; Ramachandran, S., Cache-oblivious algorithms, ACM Trans. Algorithms, 8, 1, 4:1-4:22 (2012) · Zbl 1295.68236

[12] Frigo, M.; Leiserson, C. E.; Prokop, H.; Ramachandran, S., Cache-oblivious algorithms, (40th Annual Symposium on Foundations of Computer Science, FOCS '99. 40th Annual Symposium on Foundations of Computer Science, FOCS '99, New York, NY, USA, 17-18 October 1999 (1999)), 285-298

[13] Ferragina, P.; Grossi, R., The string b-tree: a new data structure for string search in external memory and its applications, J. ACM, 46, 2, 236-280 (1999) · Zbl 1065.68518

[14] Brodal, G. S.; Fagerberg, R., Cache-oblivious string dictionaries, (Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2006. Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2006, Miami, Florida, USA, January 22-26, 2006 (2006)), 581-590 · Zbl 1192.68169

[15] Roh, K.; Crochemore, M.; Iliopoulos, C. S.; Park, K., External memory algorithms for string problems, Fundam. Inform., 84, 1, 17-32 (2008) · Zbl 1159.68039

[16] Bille, P.; Gørtz, I. L., Substring range reporting, Algorithmica, 69, 2, 384-396 (2014) · Zbl 1360.68375

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

[18] Crochemore, M.; Iliopoulos, C. S.; Kubica, M.; Rahman, M. S.; Walen, T., Improved algorithms for the range next value problem and applications, (STACS 2008, 25th Annual Symposium on Theoretical Aspects of Computer Science, Bordeaux, France, February 21-23, 2008, Proceedings (2008)) · Zbl 1259.68226

[19] Hon, W.; 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

[20] Kopelowitz, T.; Lewenstein, M.; Porat, E., Persistency in suffix trees with applications to string interval problems, (Grossi, R.; Sebastiani, F.; Silvestri, F., String Processing and Information Retrieval, 18th International Symposium, SPIRE 2011, Pisa, Italy, October 17-21, 2011, Proceedings. String Processing and Information Retrieval, 18th International Symposium, SPIRE 2011, Pisa, Italy, October 17-21, 2011, Proceedings, Lecture Notes in Computer Science (2011), Springer), 67-80

[21] Mäkinen, V.; Navarro, G., Position-restricted substring searching, (Correa, J. R.; Hevia, A.; Kiwi, M. A., ATIN 2006: Theoretical Informatics, 7th Latin American Symposium, Valdivia, Chile, March 20-24, 2006, Proceedings. ATIN 2006: Theoretical Informatics, 7th Latin American Symposium, Valdivia, Chile, March 20-24, 2006, Proceedings, Lecture Notes in Computer Science (2006), Springer), 703-714 · Zbl 1145.68392

[22] Afshani, P.; Brodal, G. S.; Zeh, N., Ordered and unordered top-k range reporting in large data sets, (Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2011. Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2011, San Francisco, California, USA, January 23-25, 2011 (2011)), 390-400 · Zbl 1373.68182

[23] Hooshmand, S.; Abedin, P.; Külekci, M. O.; Thankachan, S. V., Non-overlapping indexing – cache obliviously, (Annual Symposium on Combinatorial Pattern Matching, CPM 2018, July 2-4, 2018 - Qingdao, China (2018)), 8:1-8:9 · Zbl 07286734