Ganguly, Arnab; **Shah, Rahul**; **Thankachan, Sharma V.**
Succinct non-overlapping indexing. (English) Zbl 1436.68083
Algorithmica 82, No. 1, 107-117 (2020).

Summary: Text indexing is a fundamental problem in computer science. The objective is to preprocess a text T, so that, given a pattern P, we can find all starting positions (or simply, occurrences) of P in T efficiently. In some cases, additional restrictions are imposed. We consider two variants, namely the *non-overlapping indexing* problem, and the *range non-overlapping indexing* problem. Given a text T having n characters, the non-overlapping indexing problem is defined as follows: pre-process T into a data structure, such that for any pattern P, containing $|P|$ characters, we can report a set containing the maximum number of non-overlapping occurrences of P in T. *H. Cohen* and *E. Porat* [Lect. Notes Comput. Sci. 5878, 1044–1053 (2009; Zbl 1273.68097)] showed that by maintaining a linear space index in which the suffix tree of T is augmented with an $O(n)$ word data structure, a query P can be answered in optimal time $O(|P| + \text{nocc})$, where nocc is the number of occurrences reported. We present the following new result. Let CSA (not necessarily a compressed suffix array) be an index of T that can compute (i) the suffix range of P in $\mathsf{search}(P)$ time, and (ii) a suffix array or an inverse suffix array value in $\mathsf{t_{SA}}$ time. By using CSA alone, we can answer a query P in $\mathsf{search}(P) + \mathsf{sort}(\text{nocc}) + O(\text{nocc} \cdot \mathsf{t_{SA}})$ time. The function $\mathsf{sort}(k)$ denotes the time for sorting k numbers in $\{1, 2, \ldots, n\}$. In the range non-overlapping indexing problem, along with the pattern P, two integers a and $b, b \geq a$, are provided as input. The task is to report a set containing the maximum number of non-overlapping occurrences of P that lie within the range $[a, b]$. For any arbitrarily small positive constant ϵ, we present an $O(n \log^{\epsilon} n)$ word index with $O(|P| + \text{nocc}_{a,b})$ query time, where $\text{nocc}_{a,b}$ is the number of occurrences reported. Our index improves upon the result of Cohen and Porat [loc. cit.].

MSC:

68P05 Data structures
68P15 Database theory
68W32 Algorithms on strings

Keywords:

succinct data structures; range queries; suffix trees; string algorithms

Full Text: DOI

References:

[1] Abouelhoda, Mi; Kurtz, S.; Ohlebusch, E., Replacing suffix trees with enhanced suffix arrays, J. Discrete Algorithms, 2, 1, 53-86 (2004) · Zbl 1115.92303

[2] Alstrup, S., Brodal, G.S., Rauhe, T.: Optimal static range reporting in one dimension. In: Proceedings on 33rd Annual ACM Symposium on Theory of Computing, July 6-8, 2001, Heraklion, Crete, Greece, pp. 476-482 (2001) · Zbl 1323.68536

[3] Baker, B.S.: A theory of parameterized pattern matching: algorithms and applications. In: Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing, May 16-18, 1993, San Diego, CA, USA, pp. 71-80 (1993) · Zbl 1310.68098

[4] Boyer, Rs; Moore, Js, A fast string searching algorithm, Commun. ACM, 20, 10, 762-772 (1977) · Zbl 1219.68165

[5] Brodal, G.S., Fagerberg, R., Greve, M., López-Ortiz, A.: Online sorted range reporting. In: Algorithms and Computation, 20th International Symposium, ISAAC 2009, Honolulu, Hawaii, USA, December 16-18, 2009. Proceedings, pp. 173-182 (2009) · Zbl 1272.68113

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

[7] Cole, R., Gottlieb, L.-A., Lewenstein, M.: Dictionary matching and indexing with errors and don't cares. In: Proceedings of the 36th Annual ACM Symposium on Theory of Computing, Chicago, IL, USA, June 13-16, 2004, pp. 91-100 (2004) · Zbl 1192.68818

[8] Crochemore, M., Iliopoulos, C.S., Kubica, M., Rahman, M.S., Walen, T.: Improved algorithms for the range next value problem and applications. In: STACS 2008, 25th Annual Symposium on Theoretical Aspects of Computer Science, Bordeaux, France, February 21-23, 2008, Proceedings, pp. 205-216 (2008) · Zbl 1259.68226

[9] Farach, M.: Optimal suffix tree construction with large alphabets. In: 38th Annual Symposium on Foundations of Computer Science, FOCS '97, Miami Beach, Florida, USA, October 19-22, 1997, pp. 137-143 (1997)

[10] Ferragina, P.; Manzini, G., Indexing compressed text, J. ACM, 52, 4, 552-581 (2005) · Zbl 1323.68261

[11] Ganguly, A., Shah, R., Thankachan, S.V.: Succinct non-overlapping indexing. In: Combinatorial Pattern Matching—26th Annual Symposium, CPM 2015, Ischia Island, Italy, June 29-July 1, 2015, Proceedings, pp. 185-195 (2015) · Zbl 1432.68089

[12] Ganguly, A., Shah, R., Thankachan, S.V: pBWT: achieving succinct data structures for parameterized pattern matching and related problems. In: Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 397-407. Society for Industrial and Applied Mathematics (2017) · Zbl 1410.68098

[13] Ganguly, A., Shah, R., Thankachan, S.V.: Structural pattern matching-succinctly. In: 28th International Symposium on Algorithms and Computation, ISAAC 2017, December 9-12, 2017, Phuket, Thailand, pp. 35:1-35:13 (2017)

[14] Grossi, R., Vitter, J.S.: Compressed suffix arrays and suffix trees with applications to text indexing and string matching (extended abstract). In: Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing, May 21-23, 2000, Portland, OR, USA, pp. 397-406 (2000) · Zbl 1296.68035

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

[16] 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

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

[18] Karp, Rm; Rabin, Mo, Efficient randomized pattern-matching algorithms, IBM J. Res. Dev., 31, 2, 249-260 (1987) · Zbl 0653.68054

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

[20] Knuth, De; Morris, Jh Jr; Pratt, Vr, Fast pattern matching in strings, SIAM J. Comput., 6, 2, 323-350 (1977) · Zbl 0372.68005

[21] Mäkinen, V., Navarro, G.: Position-restricted substring searching. In: LATIN 2006: Theoretical Informatics, 7th Latin American Symposium, Valdivia, Chile, March 20-24, 2006, Proceedings, pp. 703-714 (2006) · Zbl 1145.68392

[22] Manber, U.; Myers, Ew, Suffix arrays: a new method for on-line string searches, SIAM J. Comput., 22, 5, 935-948 (1993) · Zbl 0784.68027

[23] Navarro, G.; Mäkinen, V., Compressed full-text indexes, ACM Comput. Surv., 39, 1 (2007)

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

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

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