Hon, Wing-Kai; **Shah, Rahul**; **Thankachan, Sharma V.**; **Vitter, Jeffrey Scott**

On position restricted substring searching in succinct space. (English) Zbl 1267.68102

J. Discrete Algorithms 17, 109-114 (2012).

Summary: We study the position restricted substring searching (PRSS) problem, where the task is to index a text $T[0 \dots n-1]$ of n characters over an alphabet set Σ of size δ, in order to answer the following: given a query pattern P (of length p) and two indices ℓ and r, report all $occ_{\ell,r}$ occurrences of P in $T[\ell \dots r]$. Known indexes take $O(n \log n)$ bits or $O(n \log^{1+\epsilon} n)$ bits space, and answer this query in $O(p + \log n + occ_{\ell,r} \log n)$ time or in optimal $O(p + occ_{\ell,r})$ time respectively, where ϵ is any positive constant. The main drawback of these indexes is their space requirement of $\Omega(n \log n)$ bits, which can be much more than the optimal $\log \delta$ bits to store the text T.

This paper addresses an open question asked by *V. Mäkinen* and *G. Navarro* [Lect. Notes Comput. Sci. 3887, 703–714 (2006; Zbl 1145.68392)], which is whether it is possible to design a succinct index answering PRSS queries efficiently. We first study the hardness of this problem and prove the following result: a succinct (or a compact) index cannot answer PRSS queries efficiently in the pointer machine model, and also not in the RAM model unless bounds on the well-researched orthogonal range query problem improve. However, for the special case of sufficiently long query patterns, that is for $\Omega(\log^{2+\epsilon} n)$, we derive an $|CSA_f| + |CSA_r| + o(n)$ bits index with optimal query time, where $|CSA_f|$ and $|CSA_r|$ are the space (in bits) of the compressed suffix arrays (with $O(p)$ time for pattern search) of T and $\overleftarrow{T}$ (the reverse of T) respectively.

The space can be reduced further to $|CSA_f| + o(n)|$ bits with a resulting query time will be $O(p + occ_{\ell,r} + \log^{3+\epsilon} n)$. For the general case, where there is no restriction on pattern length, we obtain an $O(\frac{1}{\epsilon 3} n \log \delta)$ bits index with $O(p + occ_{\ell,r} + n^{\epsilon})$ query time. We use suffix sampling techniques to achieve these space-efficient indexes.

MSC:

68P05	Data structures
68P10	Searching and sorting

Cited in **9** Documents

Keywords:

succint data structures; pattern matching; range searching

Full Text: DOI

References:

[1] D. Belazzougui, G. Navarro, Alphabet-independent compressed text indexing, in: ESA, 2011, pp. 748-759. · Zbl 1325.68307

[2] Bentley, J.L.; Maurer, H.A., Efficient worst-case data structures for range searching, Acta informatica, 13, 155-168, (1980) · Zbl 0423.68029

[3] P. Bille, L.L. Gørtz, Substring range reporting, in: CPM, 2011, pp. 299-308. · Zbl 1339.68049

[4] T.M. Chan, K.G. Larsen, M. Patrascu, Orthogonal range searching on the RAM, revisited, in: SoCG, 2011, pp. 1-10. · Zbl 1283.68139

[5] Chazelle, B., Lower bounds for orthogonal range searching, I: the reporting case, Journal of the ACM, 37, 200-212, (1990) · Zbl 0696.68051

[6] Y.F. Chien, W.K. Hon, R. Shah, J.S. Vitter, Geometric Burrows-Wheeler transform: linking range searching and text indexing, in: DCC, 2008, pp. 252-261.

[7] M. Crochemore, C.S. Iliopoulos, M. Kubica, M.S. Rahman, T. Walen, Improved algorithms for the range next value problem and applications, in: STACS, 2008, pp. 205-216. · Zbl 1259.68226

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

[9] Gagie, T.; Gawrychowski, P., Linear-space substring range counting over polylogarithmic alphabets. corr, (2012)

[10] Grossi, R.; Vitter, J.S., Compressed suffix arrays and suffix trees with applications to text indexing and string matching, SIAM journal on computing, 35, 2, 378-407, (2005) · Zbl 1092.68115

Edited by FIZ Karlsruhe, the European Mathematical Society and the Heidelberg Academy of Sciences and Humanities
© 2021 FIZ Karlsruhe GmbH

[11] W.K. Hon, T.H. Ku, R. Shah, S.V. Thankachan, J.S. Vitter, Compressed text indexing with wildcards, in: SPIRE, 2011, pp. 267-277. · Zbl 1280.68305

[12] W.K. Hon, R. Shah, J.S. Vitter, Ordered pattern matching: towards full-text retrieval, Technical Report TR-06-008, Purdue University, March 2006.

[13] T. Kopelowitz, M. Lewenstein, E. Porat, Persistency in suffix trees with applications to string interval problems, in: SPIRE, 2011, pp. 67-80.

[14] V. Mäkinen, G. Navarro, Position-restricted substring searching, in: LATIN, 2006, pp. 703-714. · Zbl 1145.68392

[15] Mäkinen, V.; Navarro, G., Compressed full-text indexes, ACM computing surveys, 39, 1, (2007) · Zbl 1321.68263

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

[17] McCreight, E.M., A space-economical suffix tree construction algorithm, Journal of the ACM, 23, 2, 262-272, (1976) · Zbl 0329.68042

[18] P. Weiner, Linear pattern matching algorithms, in: SWAT, 1973, pp. 1-11.

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.

© 2021 FIZ Karlsruhe GmbH