Hon, Wing-Kai; **Ku, Tsung-Han**; **Shah, Rahul**; **Thankachan, Sharma V.**; **Vitter, Jeffrey Scott**
Compressed text indexing with wildcards. (English) Zbl 1280.68305
J. Discrete Algorithms 19, 23-29 (2013).

Summary: Let $T = T_1\phi^{k_1}T_2\phi^{k_2}\cdots\phi^{k_d}T_{d+1}$ be a text of total length n, where characters of each T_i are chosen from an alphabet Σ of size σ, and ϕ denotes a wildcard symbol. The text indexing with wildcards problem is to index T such that when we are given a query pattern P, we can locate the occurrences of P in T efficiently. This problem has been applied in indexing genomic sequences that contain single-nucleotide polymorphisms (SNP) because SNP can be modeled as wildcards. Recently, *A. Tam* et al. ["Succinct text indexing with wildcards", in: SPIRE 2009, 39–50 (2009)] and *C. Thachuk* [Lect. Notes Comput. Sci. 6661, 27–40 (2011; Zbl 1339.68339)] have proposed succinct indexes for this problem. In this paper, we present the first compressed index for this problem, which takes only $nH_h + o(n\log\sigma) + O(d\log n)$ bits of space, where H_h is the hth-order empirical entropy ($h = o(\log_\sigma n)$) of T.

MSC:

68W32 Algorithms on strings
68R05 Combinatorics in computer science
68P15 Database theory
68U15 Computing methodologies for text processing; mathematical typography

Cited in **4** Documents

Keywords:

approximate pattern matching; wildcards; compressed text indexing

Full Text: DOI

References:

[1] Amir, A.; Keselman, D.; Landau, G. M.; Lewenstein, M.; Lewenstein, N.; Rodeh, M., Text indexing and dictionary matching with one error, Journal of Algorithms, 37, 2, 309-325, (2000) · Zbl 0966.68062

[2] D. Belazzougui, Succinct dictionary matching with no slowdown, in: CPM, 2010, pp. 88-100. · Zbl 1286.68521

[3] M. Burrows, D.J. Wheeler, A block-sorting lossless data compression algorithm, Technical Report 124, Digital Equipment Corporation, Paolo Alto, CA, USA, 1994.

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

[5] 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.

[6] R. Cole, L.-A. Gottlieb, M. Lewenstein, Dictionary matching and indexing with errors and don't cares, in: STOC, 2004, pp. 91-100. · Zbl 1192.68818

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

[8] Ferragina, P.; Venturini, R., A simple storage scheme for strings achieving entropy bounds, Theoretical Computer Science, 372, 1, 115-121, (2007) · Zbl 1110.68029

[9] Ferragina, P.; Manzini, G.; Mäkinen, V.; Navarro, G., Compressed representations of sequences and full-text indexes, ACM Transactions on Algorithms, 3, 2, (2007) · Zbl 1321.68263

[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

[11] Y. Han, Deterministic sorting in $O(n \log \log n)$ time and linear space, in: STOC, 2002, pp. 602-608. · Zbl 1192.68196

[12] W.K. Hon, T.W. Lam, R. Shah, S.L. Tam, J.S. Vitter, Compressed index for dictionary matching, in: DCC, 2008, pp. 23-32.

[13] W.K. Hon, R. Shah, S.V. Thankachan, J.S. Vitter, On entropy-compressed text indexing in external memory, in: SPIRE, 2009, pp. 75-89.

[14] W.K. Hon, T.H. Ku, R. Shah, S.V. Thankachan, J.S. Vitter, Faster compressed dictionary matching, in: SPIRE, 2010, pp. 191-200. · Zbl 1259.68259

[15] G. Jacobson, Space-efficient static trees and graphs, in: FOCS, 1989, pp. 549-554.

[16] J. Kärkkäinen, E. Ukkonen, Sparse suffix trees, in: COCOON, 1996, pp. 219-230.

[17] T.W. Lam, W.K. Sung, S.L. Tam, S.M. Yiu, Space-efficient indexes for string matching with don't cares, in: ISAAC, 2007, pp. 846-857. · Zbl 1193.68293

[18] 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

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

[20] Nekrich, Y., Orthogonal range searching in linear and almost-linear space, Computational Geometry, 42, 4, 342-351, (2009) · Zbl 1170.68012

[21] Raman, R.; Raman, V.; Rao, S. S., Succinct indexable dictionaries with applications to encoding \textit{k}-ary trees, prefix sums and multisets, ACM Transactions on Algorithms, 3, 4, (2007) · Zbl 1093.68582

[22] A. Tam, E. Wu, T.W. Lam, S.M. Yiu, Succinct text indexing with wildcards, in: SPIRE, 2009, pp. 39-50.

[23] C. Thachuk, Succincter text indexing with wildcards, in: CPM, 2011, pp. 27-49. · Zbl 1339.68339

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

[25] Willard, D. E., Log-logarithmic worst-case range queries are possible in space $\Theta(N)$, Information Processing Letters, 17, 2, 81-84, (1983) · Zbl 0509.68106