Chien, Yu-Feng; **Hon, Wing-Kai**; **Shah, Rahul**; **Thankachan, Sharma V.**; **Vitter, Jeffrey Scott**
Geometric BWT: compressed text indexing via sparse suffixes and range searching. (English)
Zbl 1314.68115
Algorithmica 71, No. 2, 258-278 (2015).

Summary: We introduce a new variant of the popular Burrows-Wheeler transform (BWT), called Geometric Burrows-Wheeler Transform (GBWT), which converts a text into a set of points in 2-dimensional geometry. We also introduce a reverse transform, called `Points2Text`, which converts a set of points into text. Using these two transforms, we show strong equivalence between data structural problems in geometric range searching and text pattern matching. This allows us to apply the lower bounds known in the field of orthogonal range searching to the problems in compressed text indexing. In addition, we give the first succinct (compact) index for I/O-efficient pattern matching in external memory, and show how this index can be further improved to achieve higher-order entropy compressed space.

MSC:

68P15 Database theory
68P05 Data structures
68P30 Coding and information theory (compaction, compression, models of communication, encoding schemes, etc.) (aspects in computer science)

Keywords:

text indexing; entropy compression; geometric range searching

Full Text: DOI

References:

[1] Agarwal, P.K.; Erickson, J., Geometric range searching and its relatives, Adv. Discret. Comput. Geom., 23, 1-56, (1999) · Zbl 0916.68031

[2] Aggarwal, A.; Vitter, J.S., The input/output complexity of sorting and related problems, Commun. ACM, 31, 1116-1127, (1998) · Zbl 0929.70018

[3] Aref, W.G.; Ilyas, I.F., SP-gist: an extensible database index for supporting space partitioning trees, J. Intell. Inf. Syst., 17, 215-240, (2001) · Zbl 0998.68050

[4] Arge, L.; Brodal, G.S.; Fagerberg, R.; Laustsen, M., Cache-oblivious planar orthogonal range searching and counting, 160-169, (2005) · Zbl 1380.68137

[5] Arge, L.; Samoladas, V.; Vitter, J.S., Two-dimensional indexability and optimal range search indexing, 346-357, (1999)

[6] Arroyuelo, D.; Navarro, G., A Lempel-Ziv text index on secondary storage, 83-94, (2007) · Zbl 1138.68381

[7] Baeza-Yates, R.; Barbosa, E.F.; Ziviani, N., Hierarchies of indices for text searching, Inf. Syst., 21, 497-514, (1996)

[8] Burrows, M., Wheeler, D.J.: A block-sorting lossless data compression algorithm. Technical report 124, Digital Equipment Corporation, Paolo Alto CA, USA (1994)

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

[10] Clark, D.; Munro, I., Efficient suffix trees on secondary storage, 383-391, (1996) · Zbl 0847.68030

[11] Chien, Y.F.; Hon, W.K.; Shah, R.; Vitter, J.S., Geometric Burrows-Wheeler transform: linking range searching and text indexing, 252-261, (2008)

[12] Chiu, S.Y.; Hon, W.K.; Shah, R.; Vitter, J.S., I/O-efficient compressed text indexes: from theory to practice, 426-434, (2010)

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

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

[15] Ferragina, P.; Venturini, R., A simple storage scheme for strings achieving entropy bounds, 690-696, (2007) · Zbl 1302.68108

[16] Fischer, J.; Gagie, T.; Kopelowitz, T.; Lewenstein, M.; Mäkinen, V.; Salmela, L.; Välimäki, N.N., Forbidden patterns, 327-337, (2012) · Zbl 1353.68066

[17] Gagie, T., Gawrychowski, P.: Linear-space substring range counting over polylogarithmic alphabets. (2012). CoRR. arXiv:1202.3208

[cs.DS]

[18] González, R.; Navarro, G., A compressed text index on secondary memory, 80-91, (2007)

[19] Grossi, R.; Gupta, A.; Vitter, J.S., High-order entropy-compressed text indexes, 841-850, (2003) · Zbl 1092.68584

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

[21] Guttman, A., R-trees: a dynamic index structure for spatial searching, 47-57, (1984)

[22] Hellerstein, J.M.; Naughton, J.F.; Pfeffer, A., Generalized search trees for database systems, 562-573, (1995)

[23] Hon, W.K.; Lam, T.W.; Shah, R.; Lung, S.L.; Vitter, J.S., Succinct index for dynamic dictionary matching, 1034-1043, (2009)

[24] Hon, W.K.; Lam, T.W.; Shah, R.; Lung, S.L.; Vitter, J.S., Compressed index for dictionary matching, 23-32, (2008)

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

[26] Hon, W.K.; Shah, R.; Thankachan, S.V.; Vitter, J.S., On entropy-compressed text indexing in external memory, 75-89, (2009)

[27] Hon, W.K.; Ku, T.H.; Shah, R.; Thankachan, S.V.; Vitter, J.S., Compressed text indexing with wildcards, 267-277, (2011) · Zbl 1280.68305

[28] Hon, W.K.; Ku, T.H.; Shah, R.; Thankachan, S.V.; Vitter, J.S., Compressed dictionary matching with one errors, 113-122, (2011)

[29] Hon, W.K.; Shah, R.; Vitter, J.S., Compression, indexing, and retrieval for massive string data, 260-274, (2010) · Zbl 1286.68118

[30] Jacobson, G., Space-efficient static trees and graphs, 549-554, (1989)

[31] Kanth, K.V.R.; Singh, A.K., Optimal dynamic range searching in non-replicating index structures, 257-276, (1999)

[32] Kärkkäinen, J.; Ukkonen, E., Sparse suffix trees, 219-230, (1996)

[33] Kolpakov, R.; Kucherov, G.; Starikovskaya, T.A., Pattern matching on sparse suffix trees, (2011)

[34] Mäkinen, V., Navarro, G.: Compressed full-text indexes. ACM Comput. Surv. \textbf{39}(1) (2007) · Zbl 1138.68381

[35] Mäkinen, V., Navarro, G.: Dynamic entropy-compressed sequences and full-text indexes. Technical report TR/DCC-2006-10, University of Chile (2006)

[36] Mäkinen, V.; Navarro, G., Position-restricted substring searching, 703-714, (2006) · Zbl 1145.68392

[37] Mäkinen, V.; Navarro, G.; Sadakane, K., Advantages of backward searching-efficient secondary memory and distributed implementation of compressed suffix arrays, 681-692, (2004) · Zbl 1116.68408

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

[39] McCreight, E.M., A space-economical suffix tree construction algorithm, J. ACM, 23, 262-272, (1976) · Zbl 0329.68042

[40] Munro, J.I., Tables, 37-42, (1996)

[41] Russo, L.M.S.; Navarro, G.; Oliveira, A.L., Fully compressed suffix trees, ACM Trans. Algorithms, 7, 53, (2011) · Zbl 1295.68103

[42] Sadakane, K.: Compressed suffix trees with full functionality. Theory Comput. Syst. 589-607(2007) · Zbl 1148.68015

[43] Samet, H., The quadtree and related hierarchical data structures, ACM Comput. Surv., 16, 187-260, (1984)

[44] Subramanian, S.; Ramaswamy, S., The P-range tree: a new data structure for range searching in secondary memory, 378-387, (1995) · Zbl 0848.68029

[45] Thankachan, S.V., Compressed indexes for aligned pattern matching, 410-419, (2011)

[46] Weiner, P., Linear pattern matching algorithms, 1-11, (1973)

[47] Willard, D.E., Log-logarithmic worst-case range queries are possible in space 0(N), Inf. Process. Lett., 17, 81-84, (1983) · Zbl 0509.68106

[48] Yu, C.C.; Hon, W.K.; Wang, B.F., Efficient data structures for orthogonal range successor problem, 96-105, (2009) · Zbl 1248.68175