Hon, Wing-Kai; **Ku, Tsung-Han**; **Lam, Tak-Wah**; **Shah, Rahul**; **Tam, Siu-Lung**; **Thankachan, Sharma V.**; **Vitter, Jeffrey Scott**

Compressing dictionary matching index via sparsification technique. (English) Zbl 1322.68071

Algorithmica 72, No. 2, 515-538 (2015).

Summary: Given a set $\mathcal{D}$ of patterns of total length n, the dictionary matching problem is to index $\mathcal{D}$ such that for any query text T, we can locate the occurrences of any pattern within T efficiently. This problem can be solved in optimal $O(|T| + \text{occ})$ time by the classical AC automaton [*A. V. Aho* and *M. J. Corasick*, Commun. ACM 18, 333–340 (1975; Zbl 0301.68048)], where occ denotes the number of occurrences. The space requirement is $O(n)$ words which is still far from optimal. In this paper, we show that in many cases, sparsification technique can be applied to improve the space requirements of the indexes for the dictionary matching and its related problems. First, we give a compressed index for dictionary matching, and show that such an index can be generalized to handle dynamic updates of $\mathcal{D}$. Also, we give a compressed index for approximate dictionary matching with one error. In each case, the query time is only slowed down by a polylogarithmic factor when compared with that achieved by the best $O(n)$-word counterparts.

MSC:

68P30 Coding and information theory (compaction, compression, models of communication, encoding schemes, etc.) (aspects in computer science)
68P10 Searching and sorting

Cited in **2** Documents

Keywords:

data compression; dictionary matching; text indexing; sparsification technique

Full Text: DOI

References:

[1] Aho, A.; Corasick, M., Efficient string matching: an aid to bibliographic search, Commun. ACM, 18, 333-340, (1975) · Zbl 0301.68048

[2] Alstrup, S.; Husfeldt, T.; Rauhe, T., Marked ancestor problems, 534-544, (1998)

[3] Amir, A.; Farach, M., Adaptive dictionary matching, 760-766, (1991)

[4] Amir, A.; Farach, M.; Galil, Z.; Giancarlo, R.; Park, K., Dynamic dictionary matching, J. Comput. Syst. Sci., 49, 208-222, (1994) · Zbl 0942.68783

[5] Amir, A.; Farach, M.; Idury, R.; Poutre, A.L.; Schaffer, A., Improved dynamic dictionary matching, Inf. Comput., 119, 258-282, (1995) · Zbl 0832.68033

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

[7] Arge, L.; Vitter, J.S., Optimal external memory interval management, SIAM J. Comput., 32, 1488-1508, (2003) · Zbl 1030.68027

[8] Belazzougui, D., Succinct dictionary matching with no slowdown, 88-100, (2010) · Zbl 1286.68521

[9] Bender, M.A.; Cole, R.; Demaine, E.D.; Farach-Colton, M.; Zito, J., Two simplified algorithms for maintaining order in a List, 152-164, (2002) · Zbl 1019.68527

[10] Bender, M.A.; Farach-Colton, M.; Pemmasani, G.; Skiena, S.; Sumazin, P., Lowest common ancestors in trees and directed acyclic graphs, J. Algorithms, 57, 75-94, (2005) · Zbl 1085.68103

[11] Chan, H.L.; Hon, W.K.; Lam, T.W.; Sadakane, K., Compressed indexes for dynamic text collections, ACM Trans. Algorithms, 3, 2, (2007) · Zbl 1321.68261

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

[13] Cole, R.; Gottlieb, L.-A.; Lewenstein, M., Dictionary matching and indexing with errors and don't cares, 91-100, (2004) · Zbl 1192.68818

[14] Dietz, P.F.; Sleator, D.D., Two algorithms for maintaining order in a List, 365-372, (1987)

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

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

[17] Ferragina, P.; Venturini, R., A simple storage scheme for strings achieving entropy bounds, Theor. Comput. Sci., 372, 115-121, (2007) · Zbl 1110.68029

[18] Ferragina, P.; Muthukrishnan, S.; Berg, M., Multi-method dispatching: a geometric approach with applications to string matching problems, 483-491, (1999) · Zbl 1345.68103

[19] Fischer, J.; Heun, V., Space-efficient preprocessing schemes for range minimum queries on static arrays, SIAM J. Comput., 40, 465-492, (2011) · Zbl 1222.05024

[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] Hagerup, T.; Miltersen, P.B.; Pagh, R., Deterministic dictionaries, J. Algorithms, 41, 69-85, (2001) · Zbl 1002.68503

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

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

[24] Hon, W.K.; Ku, T.H.; Shah, R.; Thankachan, S.V.; Vitter, J.S., Faster compressed dictionary matching, 191-200, (2010) · Zbl 1259.68259

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

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

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

[28] McCreight, E.M., Priority search trees, SIAM J. Comput., 14, 257-276, (1985) · Zbl 0564.68050

[29] Overmars, M.H., Efficient data structures for range searching on a grid, J. Algorithms, 9, 254-275, (1988) · Zbl 0637.68067

[30] Sadakane, K., Compressed suffix trees with full functionality, Theory Comput. Syst., 41, 589-607, (2007) · Zbl 1148.68015

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

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