Orlandi, Alessio; **Venturini, Rossano**

Space-efficient substring occurrence estimation. (English) Zbl 1336.68319

Algorithmica 74, No. 1, 65-90 (2016).

Summary: In this paper we study the problem of estimating the number of occurrences of substrings in textual data: A text T on some alphabet $\Sigma = [\sigma]$ of length n is preprocessed and an index $\mathcal{I}$ is built. The index is used in lieu of the text to answer queries of the form $\mathsf{Count} \approx (P)$, returning an approximated number of the occurrences of an arbitrary pattern P as a substring of T. The problem has its main application in selectivity estimation related to the *LIKE* predicate in textual databases. Our focus is on obtaining an algorithmic solution with guaranteed error rates and small footprint. To achieve that, we first enrich previous work in the area of compressed text-indexing providing an optimal data structure that, for a given additive error $\ell \geq 1$, requires $\Theta\left(\frac{n}{\ell} \log \sigma\right)$ bits. We also approach the issue of guaranteeing exact answers for sufficiently frequent patterns, providing a data structure whose size scales with the amount of such patterns. Our theoretical findings are supported by experiments showing the practical impact of our data structures.

MSC:

68W32 Algorithms on strings
68P05 Data structures
68P15 Database theory
68P30 Coding and information theory (compaction, compression, models of communication, encoding schemes, etc.) (aspects in computer science)

Keywords:

compressed full-text indexing; pattern matching; full-text indexing; data structures

Software:

PATRICIA

Full Text: DOI

References:

[1] Arroyuelo, D; Navarro, G; Sadakane, K, Stronger Lempel-Ziv based compressed text indexing, Algorithmica, 62, 54-101, (2012) · Zbl 1241.68061

[2] Barbay, J., Gagie, T., Navarro, G., Nekrich, Y.: Alphabet partitioning for compressed rank/select and applications. In: Proceedings of the 21st International Symposium on Algorithms and Computation (ISAAC), pp. 315-326 (2010) · Zbl 1310.68060

[3] Belazzougui, D., Boldi, P., Pagh, R., Vigna, S.: Fast prefix search in little space, with applications. In: Proceedings of the 18th Annual European Symposium on Algorithms (ESA), pp. 427-438 (2010) · Zbl 1287.68188

[4] Belazzougui, D., Navarro, G.: New lower and upper bounds for representing sequences. In: Proceedings of the 20th Annual European Symposium on Algorithms (ESA), pp. 181-192 (2012) · Zbl 1365.68260

[5] Burrows, M., Wheeler, D.: A block sorting lossless data compression algorithm. Technical Report 124, Digital Equipment Corporation (1994)

[6] Chaudhuri, S., Ganti, V., Gravano, L.: Selectivity estimation for string predicates: overcoming the underestimation problem. In: Proceedings of the 20th International Conference on Data Engineering (ICDE), p. 227 (2004)

[7] Elias, P, Efficient storage and retrieval by content and address of static files, J. ACM, 21, 246-260, (1974) · Zbl 0278.68028

[8] Fano, R.M.: On the number of bits required to implement an associative memory. Memorandum 61, Computer Structures Group, Project MAC, MIT, Cambridge, MA (1971) · Zbl 1323.68262

[9] Ferragina, P., González, R., Navarro, G., Venturini, R.: Compressed text indexes: from theory to practice. ACM J. Exp. Algorithmics \textbf{13}, 12-31 (2008) · Zbl 1284.68255

[10] 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

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

[12] Ferragina, P; Venturini, R, The compressed permuterm index, ACM Trans. Algorithms, 7, 10, (2010) · Zbl 1295.68108

[13] Ferragina, P., Venturini, R.: Compressed cache-oblivious String B-tree. In: Proceedings of 21th Annual European Symposium on Algorithms (ESA), pp. 469-480 (2013) · Zbl 1394.68094

[14] Frigo, M; Leiserson, CE; Prokop, H; Ramachandran, S, Cache-oblivious algorithms, ACM Trans. Algorithms, 8, 4, (2012) · Zbl 1295.68236

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

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

[17] Hagerup, T., Tholey, T.: Efficient minimal perfect hashing in nearly minimal space. In: Proceedings of the 18th Annual Symposium on Theoretical Aspects of Computer Science (STACS), pp. 317-326 (2001) · Zbl 0976.68594

[18] Jagadish, H.V., Ng, R.T., Srivastava, D.: Substring selectivity estimation. In: Proceedings of the 18th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of database systems (PODS), pp. 249-260 (1999)

[19] Krishnan, P., Vitter, J.S., Iyer, B.R.: Estimating alphanumeric selectivity in the presence of wildcards. In: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data (SIGMOD), pp. 282-293 (1996)

[20] Manzini, G, An analysis of the Burrows-Wheeler transform, J. ACM, 48, 407-430, (2001) · Zbl 1323.68262

[21] Morrison, DR, PATRICIA: practical algorithm to retrieve coded in alphanumeric, J. ACM, 15, 514-534, (1968)

[22] Navarro, G., Mäkinen, V.: Compressed full text indexes. ACM Comput. Surv. \textbf{39}(1), 2 (2007) · Zbl 1321.68263

[23] Orlandi, A., Venturini, R.: Space-efficient substring occurrence estimation. In: Proceedings of the 30th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS), pp. 95-106 (2011) · Zbl 1336.68319

[24] Russo, L.M.S., Navarro, G., Oliveira, A.: Fully-compressed suffix trees. In: Laber, E.S., Bornstein, C., Nogueira, L.T., Faria, L. (eds.) LATIN 2008: Theoretical Informatics, LNCS, vol. 4957, pp. 362-373. Springer, Berlin (2008) · Zbl 1136.68369

[25] Witten, I.H., Moffat, A., Bell, T.C.: Managing Gigabytes: Compressing and Indexing Documents and Images, 2nd edn. Morgan Kaufmann Publishers, Los Altos, CA (1999) · Zbl 0821.68051