Navarro, Gonzalo; **Thankachan, Sharma V.**

Bottom-k document retrieval. (English) Zbl 1328.68057

J. Discrete Algorithms 32, 69-74 (2015).

Summary: We consider the problem of retrieving the k documents from a collection of strings where a given pattern P appears least often. This has potential applications in data mining, bioinformatics, security, and big data. We show that adapting the classical linear-space solutions for this problem is trivial, but the compressed-space solutions are not easy to extend. We design a new solution for this problem that matches the best-known result when using $2|\mathsf{CSA}| + o(n)$ bits, where CSA is a compressed suffix array. Our structure answers queries in the time needed by the CSA to find the suffix array interval of the pattern plus $O(k \lg k \lg^{\varepsilon} n)$ accesses to suffix array cells, for any constant $\varepsilon > 0$.

MSC:

68P20 Information storage and retrieval of data
68P05 Data structures
68W32 Algorithms on strings

Keywords:

compact data structures; document retrieval; string collections

Software:

Wumpus

Full Text: DOI

References:

[1] Baeza-Yates, R.; Ribeiro-Neto, B., Modern information retrieval, (2011), Addison-Wesley

[2] Belazzougui, D.; Navarro, G.; Valenzuela, D., Improved compressed indexes for full-text document retrieval, J. Discrete Algorithms, 18, 3-13, (2013) · Zbl 1268.68075

[3] Burns, L.; Hellerstein, J. L.; Ma, S.; Perng, C. S.; Rabenhorst, D. A.; Taylor, D., A systematic approach to discovering correlation rules for event management, (Proc. 7th IFIP/IEEE Intl. Symp. on Integrated Network Management, (2001))

[4] Büttcher, S.; Clarke, C.; Cormack, G., Information retrieval: implementing and evaluating search engines, (2010), MIT Press · Zbl 1211.68176

[5] Cagliero, L.; Garza, P., Infrequent weighted itemset mining using frequent pattern growth, IEEE Trans. Knowl. Data Eng., 26, 4, 903-915, (2014)

[6] Chan, T. M.; Durocher, S.; Skala, M.; Wilkinson, B., Linear-space data structures for range minority query in arrays, (Proc. SWAT, (2012)), 295-306 · Zbl 1318.68068

[7] Dong, X., Mining interesting infrequent and frequent itemsets based on minimum correlation strength, (Proc. AICI, LNAI, vol. 7002, (2011)), 437-443

[8] Durocher, S.; Shah, R.; Skala, M.; Thankachan, S. V., Linear-space data structures for range frequency queries on arrays and trees, (Proc. MFCS, (2013)), 325-336 · Zbl 1400.68062

[9] El-Falah, T.; Lecroq, T.; Elloumi, M., Extraction of infrequent simple motifs from a finite set of sequences using a lattice structure, Recent Pat DNA Gene Seq., 7, 2, 123-127, (2013)

[10] Fischer, J.; Gagie, T.; Kopelowitz, T.; Lewenstein, M.; Mäkinen, V.; Salmela, L.; Välimäki, N., Forbidden patterns, (Proc. 10th LATIN, LNCS, vol. 7256, (2012)), 327-337 · Zbl 1353.68066

[11] Gagie, T.; Kärkkäinen, J.; Navarro, G.; Puglisi, S. J., Colored range queries and document retrieval, Theor. Comput. Sci., 483, 36-50, (2013) · Zbl 1292.68045

[12] Gupta, A.; Mittal, A.; Bhattacharya, A., Minimally infrequent itemset mining using pattern-growth paradigm and residual trees, (2012), CoRR

[13] Haglin, D.; Manning, A., On minimal infrequent itemset mining, (Proc. DMIN, (2007)), 141-147

[14] Herold, J.; Kurtz, S.; Giegerich, R., Efficient computation of absent words in genomic sequences, BMC Bioinform., 9, 167, (2008)

[15] Hon, W.-K.; Shah, R.; Thankachan, S.; Vitter, J., Faster compressed top-k document retrieval, (Proc. 23rd DCC, (2013)), 341-350

[16] Hon, W.-K.; Shah, R.; Vitter, J., Space-efficient framework for top-\textit{k} string retrieval problems, (Proc. 50th FOCS, (2009)), 713-722 · Zbl 1292.68182

[17] Ji, Y.; Ying, H.; Tran, J.; Dews, P.; Mansour, A.; Massanari, R. M., A method for mining infrequent causal associations and its application in finding adverse drug reaction signal pairs, IEEE Trans. Knowl. Data Eng., 25, 4, 721-733, (2013)

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

[19] Navarro, G., Spaces, trees and colors: the algorithmic landscape of document retrieval on sequences, ACM Comput. Surv., 46, 4, (2014), art. 52 · Zbl 1305.68078

[20] Navarro, G.; Mäkinen, V., Compressed full-text indexes, ACM Comput. Surv., 39, 1, (2007), art. 2 · Zbl 1321.68263

[21] Navarro, G.; Nekrich, Y., Top-\textit{k} document retrieval in optimal time and linear space, (Proc. 23rd SODA, (2012)), 1066-1078

[22] Navarro, G.; Thankachan, S., Faster top-\textit{k} document retrieval in optimal space, (Proc. 20th SPIRE, LNCS, vol. 8214, (2013)), 255-262

[23] Navarro, G.; Thankachan, S., Top-\textit{k} document retrieval in compact space and near-optimal time, (Proc. 24th ISAAC, LNCS, vol. 8283, (2013)), 394-404 · Zbl 1406.68022

[24] Navarro, G.; Valenzuela, D., Space-efficient top-\textit{k} document retrieval, (Proc. 11th SEA, (2012)), 307-319

[25] Rahman, A.; Ezeife, C. I.; Aggarwal, A. K., WiFi miner: an online apriori-infrequent based wireless intrusion system, (Proc. Knowledge Discovery from Sensor Data, LNAI, vol. 5840, (2010)), 76-93

[26] Raman, R.; Raman, V.; Srinivasa Rao, S., Succinct indexable dictionaries with applications to encoding \textit{k}-ary trees, prefix sums and multisets, ACM Trans. Algorithms, 3, 4, (2007), art. 43 · Zbl 1093.68582

[27] Sadakane, K., Succinct data structures for flexible text retrieval systems, J. Discrete Algorithms, 5, 12-22, (2007) · Zbl 1137.68360

[28] Shah, R.; Sheng, C.; Thankachan, S. V.; Vitter, J., Top-k document retrieval in external memory, (Proc. 21st ESA, LNCS, vol. 8125, (2013)), 803-814 · Zbl 1394.68129

[29] Tsur, D., Top-k document retrieval in optimal space, Inf. Process. Lett., 113, 12, 440-443, (2013) · Zbl 1371.68071

[30] Vens, C.; Danchin, E.; Rosso, M. N., Identifying proteins involved in parasitism by discovering degenerated motifs, (Proc. 4th International Workshop on Machine Learning in Systems Biology, (2010)), 81-84

[31] Weiner, P., Linear pattern matching algorithm, (Proc. 14th Annual IEEE Symposium on Switching and Automata Theory, (1973)), 1-11

[32] Wu, X.; Zhang, C.; Zhang, S., Efficient mining of both positive and negative association rules, ACM Trans. Inf. Syst., 22, 3, 381-405, (2004)