**Hon, Wing-Kai**; **Patil, Manish**; **Shah, Rahul**; **Thankachan, Sharma V.**; **Vitter, Jeffrey Scott**
**Indexes for document retrieval with relevance.** (English) Zbl 1394.68127
Brodnik, Andrej (ed.) et al., Space-efficient data structures, streams, and algorithms. Papers in honor of J. Ian Munro on the occasion of his 66th birthday. Berlin: Springer (ISBN 978-3-642-40272-2/pbk). Lecture Notes in Computer Science 8066, 351-362 (2013).

Summary: Document retrieval is a special type of pattern matching that is closely related to information retrieval and web searching. In this problem, the data consist of a collection of text documents, and given a query pattern $P$, we are required to report all the documents (not all the occurrences) in which this pattern occurs. In addition, the notion of relevance is commonly applied to rank all the documents that satisfy the query, and only those documents with the highest relevance are returned. Such a concept of relevance has been central in the effectiveness and usability of present day search engines like Google, Bing, Yahoo, or Ask. When relevance is considered, the query has an additional input parameter $k$, and the task is to report only the $k$ documents with the highest relevance to $P$, instead of finding all the documents that contains $P$. For example, one such relevance function could be the frequency of the query pattern in the document. In the information retrieval literature, this task is best achieved by using inverted indexes. However, if the query consists of an arbitrary string-which can be a partial word, multiword phrase, or more generally any sequence of characters-we cannot take advantages of the word boundaries and we need a different approach.

This leads to one of the active research topics in string matching and text indexing community in recent years, and various aspects of the problem have been studied, such as space-time tradeoffs, practical solutions, multipattern queries, and I/O-efficiency. In this article, we review some of the initial frameworks for designing such indexes and also summarize the developments in this area.

For the entire collection see [Zbl 1270.68019].

**MSC:**

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



**Full Text:** DOI

**References:**

[1] Afshani, P.: On dominance reporting in 3D. In: Halperin, D., Mehlhorn, K. (eds.) ESA 2008. LNCS, vol.˜5193, pp. 41–51. Springer, Heidelberg (2008) · Zbl 1158.68363 · doi:10.1007/978-3-540-87744-8_4

[2] Afshani, P., Brodal, G.S., Zeh, N.: Ordered and unordered top-k range reporting in large data sets. In: SODA, pp. 390–400 (2011) · Zbl 1373.68182 · doi:10.1137/1.9781611973082.31

[3] Aggarwal, A., Vitter, J.S.: The input/output complexity of sorting and related problems. Commun. ACM˜31(9), 1116–1127 (1988) · doi:10.1145/48529.48535

[4] Arge, L., Samoladas, V., Vitter, J.S.: On two-dimensional indexability and optimal range search indexing. In: Proc. 18th Symposium on Principles of Database Systems (PODS), pp. 346–357 (1999) · doi:10.1145/303976.304010

[5] Belazzougui, D., Navarro, G.: Improved compressed indexes for full-text document retrieval. In: Grossi, R., Sebastiani, F., Silvestri, F. (eds.) SPIRE 2011. LNCS, vol.˜7024, pp. 386–397. Springer, Heidelberg (2011) · Zbl 05965196 · doi:10.1007/978-3-642-24583-1_38

[6] Chazelle, B.: Lower bounds for orthogonal range searching: I. the reporting case. J. ACM˜37(2), 200–212 (1990) · Zbl 0696.68051 · doi:10.1145/77600.77614

[7] Chien, Y.-F., Hon, W.-K., Shah, R., Thankachan, S.V., Vitter, J.S.: Geometric burrows-wheeler transform: Compressed text indexing via sparse suffixes and range searching. Algorithmica (2013)

[8] Cohen, H., Porat, E.: Fast set intersection and two-patterns matching. Theor. Comput. Sci.˜411(40-42), 3795–3800 (2010) · Zbl 1207.68270 · doi:10.1016/j.tcs.2010.06.002

[9] Cole, R., Gottlieb, L.-A., Lewenstein, M.: Dictionary matching and indexing with errors and don't cares. In: STOC, pp. 91–100 (2004) · Zbl 1192.68818 · doi:10.1145/1007352.1007374

[10] Culpepper, J.S., Navarro, G., Puglisi, S.J., Turpin, A.: Top-k ranked document search in general text databases. In: de Berg, M., Meyer, U. (eds.) ESA 2010, Part II. LNCS, vol.˜6347, pp. 194–205. Springer, Heidelberg (2010) · Zbl 1287.68035 ·

doi:10.1007/978-3-642-15781-3_17

[11] Ferragina, P., Manzini, G.: Indexing compressed text. J. ACM˜52(4), 552–581 (2005) · Zbl 1323.68261 · doi:10.1145/1082036.1082039

[12] Fischer, J., Gagie, T., Kopelowitz, T., Lewenstein, M., Mäkinen, V., Salmela, L., Välimäki, N.: Forbidden patterns. In: Fernández-Baca, D. (ed.) LATIN 2012. LNCS, vol.˜7256, pp. 327–337. Springer, Heidelberg (2012) · Zbl 1353.68066 · doi:10.1007/978-3-642-29344-3_28

[13] Gagie, T., Karhu, K., Navarro, G., Puglisi, S.J., Sirén, J.: Document listing on repetitive collections. In: Fischer, J., Sanders, P. (eds.) CPM 2013. LNCS, vol.˜7922, pp. 107–119. Springer, Heidelberg (2013) · Zbl 1381.68076 · doi:10.1007/978-3-642-38905-4_12

[14] Gagie, T., Navarro, G., Puglisi, S.J.: Colored range queries and document retrieval. In: Chavez, E., Lonardi, S. (eds.) SPIRE 2010. LNCS, vol.˜6393, pp. 67–81. Springer, Heidelberg (2010) · Zbl 05803982 · doi:10.1007/978-3-642-16321-0_7

[15] Gagie, T., Navarro, G., Puglisi, S.J.: New algorithms on wavelet trees and applications to information retrieval. Theor. Comput. Sci.˜426, 25–41 (2012) · Zbl 1243.68161 · doi:10.1016/j.tcs.2011.12.002

[16] Grossi, R., Vitter, J.S.: Compressed suffix arrays and suffix trees with applications to text indexing and string matching. SIAM J. Comput.˜35(2), 378–407 (2005) · Zbl 1092.68115 · doi:10.1137/S0097539702402354

[17] Hon, W.-K., Patil, M., Shah, R., Wu, S.-B.: Efficient index for retrieving top-k most frequent documents. J. Discrete Algorithms˜8(4), 402–417 (2010) · Zbl 1215.68095 · doi:10.1016/j.jda.2010.08.003

[18] Hon, W.-K., Shah, R., Thankachan, S.V.: Towards an optimal space-and-query-time index for top-k document retrieval. In: Kärkkäinen, J., Stoye, J. (eds.) CPM 2012. LNCS, vol.˜7354, pp. 173–184. Springer, Heidelberg (2012) · Zbl 1358.68092 · doi:10.1007/978-3-642-31265-6_14

[19] Hon, W.-K., Shah, R., Thankachan, S.V., Vitter, J.S.: String retrieval for multi-pattern queries. In: Chavez, E., Lonardi, S. (eds.) SPIRE 2010. LNCS, vol.˜6393, pp. 55–66. Springer, Heidelberg (2010) · Zbl 05803981 · doi:10.1007/978-3-642-16321-0_6

[20] Hon, W.-K., Shah, R., Thankachan, S.V., Vitter, J.S.: Document listing for queries with excluded pattern. In: Kärkkäinen, J., Stoye, J. (eds.) CPM 2012. LNCS, vol.˜7354, pp. 185–195. Springer, Heidelberg (2012) · Zbl 1358.68093 · doi:10.1007/978-3-642-31265-6_15

[21] Hon, W.-K., Shah, R., Thankachan, S.V., Vitter, J.S.: Faster compressed top-k document retrieval. In: DCC (2013)

[22] Hon, W.-K., Shah, R., Vitter, J.S.: Space-efficient framework for top-k string retrieval problems. In: FOCS 2009, pp. 713–722 (2009) · Zbl 1292.68182 · doi:10.1109/FOCS.2009.19

[23] Hon, W.-K., Shah, R., Vitter, J.S.: Compression, indexing, and retrieval for massive string data. In: Amir, A., Parida, L. (eds.) CPM 2010. LNCS, vol.˜6129, pp. 260–274. Springer, Heidelberg (2010) · Zbl 1286.68118 · doi:10.1007/978-3-642-13509-5_24

[24] Culpepper, M.P.J.S., Scholer, F.: Efficient in-memory top-k document retrieval. In: SIGIR (2012) · doi:10.1145/2348283.2348317

[25] Karpinski, M., Nekrich, Y.: Top-k color queries for document retrieval. In: SODA, pp. 401–411 (2011) · Zbl 1373.68197

[26] Konow, R., Navarro, G.: Faster Compact Top-k Document Retrieval. In: DCC (2013) · Zbl 1369.68171 · doi:10.1109/DCC.2013.43

[27] Matias, Y., Muthukrishnan, S.M., Şahinalp, S.C., Ziv, J.: Augmenting suffix trees, with applications. In: Bilardi, G., Pietracaprina, A., Italiano, G.F., Pucci, G. (eds.) ESA 1998. LNCS, vol.˜1461, pp. 67–78. Springer, Heidelberg (1998) · doi:10.1007/3-540-68530-8_6

[28] Muthukrishnan, S.: Efficient algorithms for document retrieval problems. In: SODA, pp. 657–666 (2002) · Zbl 1093.68588

[29] Navarro, G.: Spaces, trees and colors: The algorithmic landscape of document retrieval on sequences. CoRR, abs/1304.6023 (2013)

[30] Navarro, G., Nekrich, Y.: Top-k document retrieval in optimal time and linear space. In: SODA, pp. 1066–1077 (2012)

[31] Navarro, G., Puglisi, S.J.: Dual-sorted inverted lists. In: Chavez, E., Lonardi, S. (eds.) SPIRE 2010. LNCS, vol.˜6393, pp. 309–321. Springer, Heidelberg (2010) · Zbl 05804008 · doi:10.1007/978-3-642-16321-0_33

[32] Navarro, G., Puglisi, S.J., Valenzuela, D.: Practical compressed document retrieval. In: Pardalos, P.M., Rebennack, S. (eds.) SEA 2011. LNCS, vol.˜6630, pp. 193–205. Springer, Heidelberg (2011) · Zbl 05906096 · doi:10.1007/978-3-642-20662-7_17

[33] Navarro, G., Thankachan, S.V.: Faster top-k document retrieval in optimal space (submitted) · Zbl 1406.68022

[34] Navarro, G., Valenzuela, D.: Space-efficient top-k document retrieval. In: Klasing, R. (ed.) SEA 2012. LNCS, vol.˜7276, pp. 307–319. Springer, Heidelberg (2012) · Zbl 06069616 · doi:10.1007/978-3-642-30850-5_27

[35] Nekrich, Y., Patil, M., Shah, R., Thankachan, S.V., Vitter, J.S.: Top-k categorical range maxima queries (submitted)

[36] Patil, M., Thankachan, S.V., Shah, R., Hon, W.-K., Vitter, J.S., Chandrasekaran, S.: Inverted indexes for phrases and strings. In: SIGIR, pp. 555–564 (2011) · doi:10.1145/2009916.2009992

[37] Sadakane, K.: Succinct data structures for flexible text retrieval systems. J. Discrete Algorithms˜5(1), 12–22 (2007) · Zbl 1137.68360 · doi:10.1016/j.jda.2006.03.011

[38] Shah, R., Sheng, C., Thankachan, S.V., Vitter, J.S.: On optimal top-k string retrieval. CoRR, abs/1207.2632 (2012) · Zbl 1394.68129

[39] Tsur, D.: Top-k document retrieval in optimal space. Inf. Process. Lett.˜113(12), 440–443 (2013) · Zbl 1371.68071 · doi:10.1016/j.ipl.2013.03.012

[40] Välimäki, N., Mäkinen, V.: Space-efficient algorithms for document retrieval. In: Ma, B., Zhang, K. (eds.) CPM 2007. LNCS, vol.˜4580, pp. 205–215. Springer, Heidelberg (2007) · Zbl 1138.68401 · doi:10.1007/978-3-540-73437-6_22

[41] Vitter, J.S.: Compressed data structures with relevance. In: CIKM, pp. 4–5 (2012) · doi:10.1145/2396761.2396765