

Hon, Wing-Kai; Lam, Tak-Wah; Shah, Rahul; Thankachan, Sharma V.; Ting, Hing-Fung; Yang, Yilin

**Dictionary matching with a bounded gap in pattern or in text.** (English) Zbl 1391.68129

Algorithmica 80, No. 2, 698-713 (2018).

**Summary:** A gap is a sequence of don't care characters. In this paper, we study two variants of the dictionary matching problem, where gaps may be present in the patterns or in the text. The first variant, called *dictionary matching with one gap*, considers indexing a collection  $\mathcal{D}$  of  $d$  one-gap-patterns, where the  $i$ th pattern is of the form  $P_i[\alpha_i, \beta_i]Q_i$  with  $P_i$  and  $Q_i$  are strings drawn from an alphabet  $\Sigma$  and  $[\alpha_i, \beta_i]$  denote the lower and upper bounds on the gap length. The target is to allow a user to efficiently identify all substrings of a query text  $T$  that match with any one-gap-pattern in the collection. We present a linear space solution for answering the above dictionary matching query in time  $O(|T|\gamma \log \lambda \log d + \text{occ})$ , where  $\gamma$  denotes the number of distinct gap lengths,  $\lambda$  denotes the number of distinct lower and upper bounds of gap lengths, and the  $\text{occ}$  is the output size. The query time can be improved to  $O(|T|\gamma + \text{occ})$  using  $O(d^{1+\epsilon})$  extra space, where  $\epsilon > 0$  is an arbitrarily small constant. Additionally, we show a succinct-space index offering a space-time tradeoff. In the special case where parameters  $\alpha_i$  and  $\beta_i$ 's for all the patterns are same, our results improve upon the work by A. Amir et al. [Theor. Comput. Sci. 589, 34–46 (2015; Zbl 1319.68106)]. The second variant, called *dictionary matching with one missing substring*, is a new problem in which a gap of bounded length may be present in the text substring when it is being matched. We show that this problem can be solved by using a similar framework. Furthermore, by applying a centroid path decomposition on the *failure tree*, we obtain a space-time tradeoff result, which will be suitable when the dictionary contains only *short* patterns, or when index space is a critical concern.

**MSC:**

68W32 Algorithms on strings  
68P05 Data structures

Cited in **3** Documents

**Keywords:**

dictionary matching; bounded gap; stabbing query; indexing

**Full Text:** [DOI](#)

**References:**

- [1] Afshani, P., Arge, L., Larsen, K.G.: Higher-dimensional orthogonal range reporting and rectangle stabbing in the pointer machine model. In: Proceedings of ACM Symposium on Computational Geometry (SoCG), pp. 323-332 (2012) · [Zbl 1293.68278](#)
- [2] Aho, AV; Corasick, MJ, Efficient string matching: an aid to bibliographic search, Commun. ACM, 18, 333-340, (1975) · [Zbl 0301.68048](#)
- [3] Amir, A., Farach, M.: Adaptive dictionary matching. In: Proceedings of IEEE Annual Symposium on Foundations of Computer Science (FOCS), pp. 760-766 (1991) · [Zbl 1211.68141](#)
- [4] Amir, A; Farach, M; Idury, RM; Poutré, JAL; Schäffer, AA, Improved dynamic dictionary matching, Inf. Comput., 119, 258-282, (1995) · [Zbl 0832.68033](#)
- [5] Amir, A; Keselman, D; Landau, GM; Lewenstein, M; Lewenstein, N; Rodeh, M, Text indexing and dictionary matching with one error, J. Algorithms, 37, 309-325, (2000) · [Zbl 0966.68062](#)
- [6] Amir, A., Levy, A., Porat, E., Shalom, B.R.: Dictionary matching with one gap. In: Proceedings of Annual Symposium on Combinatorial Pattern Matching (CPM), pp. 11-20 (2014) · [Zbl 1390.68781](#)
- [7] Amir, A; Levy, A; Porat, E; Shalom, BR, Dictionary matching with a few gaps, Theor. Comput. Sci., 589, 34-46, (2015) · [Zbl 1319.68106](#)
- [8] Belazzougui, D.: Succinct dictionary matching with no slowdown. In: Proceedings of Annual Symposium on Combinatorial Pattern Matching (CPM), pp. 88-100 (2010) · [Zbl 1286.68521](#)
- [9] Boyer, RS; Moore, JS, A fast string searching algorithm, Commun. ACM, 20, 762-772, (1977) · [Zbl 1219.68165](#)
- [10] Chan, T.M., Larsen, K.G., Patrascu, M.: Orthogonal range searching on the RAM, revisited. In: Proceedings of ACM Symposium on Computational Geometry (SoCG), pp. 1-10 (2011) · [Zbl 1283.68139](#)

- [11] Chazelle, B, Filtering search: a new approach to query-answering, *SIAM J. Comput.*, 15, 703-724, (1986) · [Zbl 0612.68088](#)
- [12] Cole, R., Gottlieb, L., Lewenstein, M.: Dictionary matching and indexing with errors and don't cares. In: *Proceedings of Annual ACM Symposium on Theory of Computing*, pp. 91-100 (2004) · [Zbl 1192.68818](#)
- [13] Feigenblat, G., Porat, E., Shifman, A.: An improved query time for succinct dynamic dictionary matching. In: *Proceedings of Annual Symposium on Combinatorial Pattern (CPM)*, pp. 120-129 (2014) · [Zbl 1407.68108](#)
- [14] Ferragina, P; Manzini, G, Indexing compressed text, *J. ACM*, 52, 552-581, (2005) · [Zbl 1323.68261](#)
- [15] Fredriksson, K; Grabowski, S, Efficient algorithms for pattern matching with general gaps, character classes, and transposition invariance, *Inf. Retr.*, 11, 335-357, (2008)
- [16] 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](#)
- [17] Haapasalo, T., Silvasti, P., Sippu, S., Soisalon-Soininen, E.: Online dictionary matching with variable-length gaps. In: *Proceedings of Symposium on Experimental Algorithms (SEA)*, pp. 76-87 (2011) · [Zbl 0966.68062](#)
- [18] Hofmann, K; Bucher, P; Falquet, L; Bairoch, A, The PROSITE database, its status in 1999, *Nucleic Acids Res.*, 27, 215-219, (1999)
- [19] Hon, WK; Ku, TH; Lam, TW; Shah, R; Tam, SL; Thankachan, SV; Vitter, JS, Compressing dictionary matching index via sparsification technique, *Algorithmica*, 72, 515-538, (2015) · [Zbl 1322.68071](#)
- [20] Hon, W.K., Ku, T.H., Shah, R., Thankachan, S.V., Vitter, J.S.: Compressed dictionary matching with one error. In: *Proceedings of IEEE Data Compression Conference (DCC)*, pp. 113-122 (2011) · [Zbl 1259.68259](#)
- [21] Hon, WK; Ku, TH; Shah, R; Thankachan, SV; Vitter, JS, Faster compressed dictionary matching, *Theor. Comput. Sci.*, 475, 113-119, (2013) · [Zbl 1259.68259](#)
- [22] Hon, W.K., Lam, T.W., Shah, R., Tam, S.L., Vitter, J.S.: Compressed index for dictionary matching. In: *Proceedings of IEEE Data Compression Conference (DCC)*, pp. 23-32 (2008) · [Zbl 1319.68106](#)
- [23] Hon, W.K., Lam, T.W., Shah, R., Tam, S.L., Vitter, J.S.: Succinct index for dynamic dictionary matching. In: *Proceedings of International Symposium on Algorithms and Computation (ISAAC)*, pp. 1034-1043 (2009) · [Zbl 0372.68005](#)
- [24] Karp, RM; Rabin, MO, Efficient randomized pattern-matching algorithms, *IBM J. Res. Dev.*, 31, 249-260, (1987) · [Zbl 0653.68054](#)
- [25] Karpinski, M., Nekrich, Y.: Space efficient multi-dimensional range reporting. In: *Proceedings of Annual International Conference on Computing and Combinatorics (COCOON)*, pp. 215-224 (2009) · [Zbl 1248.68524](#)
- [26] Knuth, DE; Morris, JH; Pratt, VR, Fast pattern matching in strings, *SIAM J. Comput.*, 6, 323-350, (1977) · [Zbl 0372.68005](#)
- [27] Kucherov, G; Rusinowitch, M, Matching a set of strings with variable length don't cares, *Theor. Comput. Sci.*, 178, 129-154, (1997) · [Zbl 0901.68037](#)
- [28] Lewenstein, M.: Dictionary matching. In: *Encyclopedia of Algorithms*, pp. 533-538 (2016) · [Zbl 1211.68141](#)
- [29] Manber, U; Myers, EW, Suffix arrays: a new method for on-line string searches, *SIAM J. Comput.*, 22, 935-948, (1993) · [Zbl 0784.68027](#)
- [30] Mehldau, G; Myers, G, A system for pattern matching applications on biosequences, *Comput. Appl. Biosci.*, 9, 299-314, (1993)
- [31] Navarro, G; Raffinot, M, Fast and simple character classes and bounded gaps pattern matching, with applications to protein searching, *J. Comput. Biol.*, 10, 903-923, (2003)
- [32] Sleator, DD; Tarjan, RE, A data structure for dynamic trees, *J. Comput. Syst. Sci.*, 26, 362-391, (1983) · [Zbl 0509.68058](#)
- [33] Weiner, P.: Linear pattern matching algorithms. In: *Proceedings of Annual Symposium on Switching and Automata*, pp. 1-11 (1973)
- [34] Zhang, M; Zhang, Y; Hu, L, A faster algorithm for matching a set of patterns with variable length don't cares, *Inf. Process. Lett.*, 110, 216-220, (2010) · [Zbl 1211.68141](#)

This reference list is based on information provided by the publisher or from digital mathematics libraries. Its items are heuristically matched to zbMATH identifiers and may contain data conversion errors. It attempts to reflect the references listed in the original paper as accurately as possible without claiming the completeness or perfect precision of the matching.