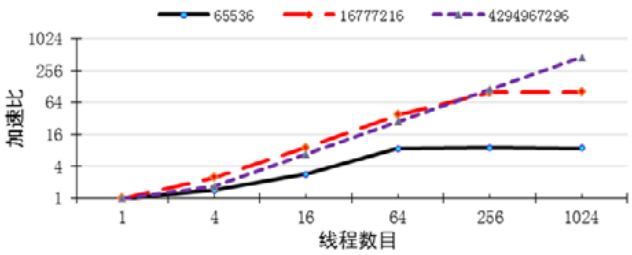


Fig.17 Variation of speedup with the number of threads for parallel SGPM-SAI

图 17 Parallel SGPM-SAI 算法的加速比随线程数目的变化

7 结论与讨论

带通配符的模式匹配是一个经典的研究问题,带有可变间隙约束的模式匹配则是近年来比较热门的研究方向.为高效求解带稀疏间隙约束条件下模式匹配的完备解,本文提出了一种基于图索引的匹配算法,即 SGPM-SAI 算法.该算法通过对文本串预处理,建立一种称为 W-SAM 的索引结构,然后采用我们所提出的模式匹配算法进行匹配,并返回匹配结果的完备解.通过对比实验,在不考虑预处理时间的情况下,对于固定间隙约束的模式匹配,相比几种现有的典型匹配算法,SGPM-SAI 算法性能高出 1~3 个数量级;而对于可变间隙约束的模式匹配,SGPM-SAI 算法性能则可高出 3~5 倍.SAIL 具有良好的时间性能来解决满足 one-off 条件下的模式匹配问题,本文选取 SAIL 算法的一种优化算法 SAIL-Gen 作为基准比较方法进行了对比,结果表明:当间隙约束比较稀疏时,SGPM-SAI 算法的性能要显著优于 SAIL-Gen 算法.为有效利用现代处理器的大规模并行处理能力,本文提出了 SGPM-SAI 算法的并行优化方案.实验结果表明:在间隙可变度较大时,Parallel SGPM-SAI 算法的加速效果显著,且具有良好的并行可扩展性,非常适宜利用现代处理器的大规模并行处理能力.

虽然我们的模式匹配算法相比经典算法具有更加优越的性能,但代价是算法需要较长的时间来对文本进行预处理,以及较大的内存空间来存储图结构的数据索引.因此,SGPM-SAI 算法比较适用于待匹配的文本较为稳定、而用来查找的模式串变动较为频繁的应用场景,如数据仓库 OLAP 中对历史数据的模糊查询.在下一步的研究工作中,如何进一步压缩创建文本索引的时间和空间代价,将是我们研究的重点.

致谢 本文部分工作是在作者访问中国人民大学的萨师煊大数据管理和分析中心时完成的,该中心获国家高等学校学科创新引智计划(111 计划)的资助.

References:

[1] Fischer MJ, Paterson MS. String-Matching and other products. In: Proc. of the 7th SIAM AMS Complexity of Computation. Cambridge, 1974. 113−125.

[2] Qiang JP, Xie F, Gao J, *et al*. Pattern matching with arbitrary-length wildcards. Acta Automatica Sinica, 2014,40(11):2499−2511 (in Chinese with English abstract). [doi: 10.3724/SP.J.1004.2014.02499]

[3] Clifford P, Clifford R. Simple deterministic wildcard matching. Information Processing Letters, 2007,101(2):53−54. [doi: 10.1016/j.ipl.2006.08.002]

[4] Cole R, Hariharan R. Verifying candidate matches in sparse and wildcard matching. In: Proc. of the Annual ACM Symp. on Theory of Computing. 2002. 592−601. [doi: 10.1145/509907.509992]

[5] Kalai A. Efficient pattern-matching with don't cares. In: Proc. of the 13th Annual ACM-SIAM Symp. on Discrete Algorithms. Philadelphia, 2002. 655−656.

[6] Wu XD, Zhu XQ, He Y, Arslan AN. PMBC: Pattern mining from biological sequences with wildcard constraints. Computers in Biology and Medicine, 2013,43(5):481−492. [doi: 10.1016/j.compbiomed.2013.02.006]

[7] Sitaridi EA, Ross KA. GPU-Accelerated string matching for database applications. VLDB Journal, 2015. 1−22. [doi: 10.1007/s00778-015-0409-y]

[8] Xin C, Jia XF, Wu YX, *et al*. Strict pattern matching with general gaps and one-off condition. Ruan Jian Xue Bao/Journal of Software, 2015,26(5):1096−1112 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4707.htm [doi: 10.13328/j.cnki.jos.004707]

[9] Chen G, Wu XD, Zhu XQ, Arslan AN, He Y. Efficient string matching with wildcards and length constraints. Knowledge and Information Systems, 2006,10(4):399−419. [doi: 10.1007/s10115-006-0016-8]

[10] Wu YX, Liu YW, Guo L, Wu XD. Subnettrees for strict pattern matching with general gaps and length constraints. Ruan Jian Xue Bao/Journal of Software, 2013,24(5):915−932 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4381.htm [doi: 10.3724/SP.J.1001.2013.04381]

[11] Wu XD, Xie F, Qiang JP. Pattern Matching with Wildcards and Length Constraints. Beijing: Science Press, 2016.

[12] Bille P, G¨ortz IL, Vildh´oj HW, Wind DK. String matching with variable length gaps. Theoretical Computer Science, 2012,443(1): 25−34. [doi: 10.1016/j.tcs.2012.03.029]

[13] Wu YX, Shen C, Jiang H. Strict pattern matching under non-overlapping condition. Science China Information Sciences, 2017, 60(1): 1−16. [doi: 10.1007/s11432-015-0935-3]

[14] Ding BL, Lo D, Han JW, Khoo SC. Efficient mining of closed repetitive gapped subsequences from a sequence database. In: Proc. of the 25th Int'l Conf. on Data Engineering. Shanghai: IEEE, 2009. 1024−1035. [doi: 10.1109/ICDE.2009.104]

[15] Wang H, Wang HP, Wu XD. Models for pattern matching with wildcards and length constraints. Computer Science, 2016,43(4): 279−283 (in Chinese with English abstract).

[16] Manber U, Baeza-Yates R. An algorithm for string matching with a sequence of dont cares. Information Processing Letters, 1991, 37(3):133−136. [doi: 10.1016/0020-0190(91)90032-D]

[17] Navarro G, Raffinot M. Fast and simple character classes and bounded gaps pattern matching, with applications to protein searching. Journal of Computational Biology, 2003,10(6):903−923. [doi: 10.1089/106652703322756140]

[18] Knuth DE, Jr Morris JH, Pratt VR. Fast pattern matching in strings. SIAM Journal on Computing, 1977,6(1):323−350. [doi: 10.1137/0206024]

[19] Boyer RS, Moore JS. A fast string searching algorithm. Communications of the ACM, 1977,20(10):762−772. [doi: 10.1145/359842.359859]

[20] Karp RM, Rabin MO. Efficient randomized pattern-matching algorithms. IBM Journal of Research and Development, 1987,31(2): 249−260. [doi: 10.1147/rd.312.0249]

[21] Aho AV, Corasick MJ. Efficient string matching: An aid to bibliographic search. Communications of the ACM, 1975,18(6): 333−340. [doi: 10.1145/360825.360855]

[22] Wu S. A fast algorithm for multi-pattern searching. Technical Report, Report TR-94-17, Tucson: Department of Computer Science, University of Arizona, 1994.

[23] Bellekens X, Atkinson R, Andonovic I, *et al*. Investigation of GPU-based pattern matching. In: Proc. of the Post Graduate Symp. on the Convergence of Telecommunications, Networking and Broadcasting. 2013. [doi: 10.6084/m9.figshare.3821922.v1]

[24] Lin CH, Tsai SY, Liu CH, *et al*. Accelerating string matching using multi-threaded algorithm on GPU. In: Proc. of the IEEE Global Telecommunications Conf. IEEE, 2010. 1−5. [doi: 10.1109/GLOCOM.2010.5683320]

[25] Xu D, Zhang H, Fan Y. The GPU based high-performance pattern-matching algorithm for intrusion detection. Journal of Computational Information Systems, 2013. 3791−3800. [doi: 10.12733/jcis5781]

[26] Cole R, Gottlieb LA, Lewenstein M. Dictionary matching and indexing with errors and dont cares. In: Proc. of the 36th Annual ACM Symp. on Theory of Computing. New York: ACM Press, 2004. 91−100. [doi: 10.1145/1007352.1007374]

[27] Karkkainen J, Sanders P. Simple linear work suffix array construction. In: Proc. of the Int'l Colloquium on Automata Languages and Programming. 2003. 943−955. [doi: 10.1007/3-540-45061-0_73]

[28] Ko P, Aluru S. Space efficient linear time construction of suffix arrays. In: Proc. of the Combinatorial Pattern Matching. 2003. 200−210. [doi: 10.1007/3-540-44888-8_15]

[29] Khancome C, Boonjing V. New Hashing based multiple string pattern matching algorithms. In: Proc. of the Int'l Conf. on Information Technology: New Generations. 2012. 195−200. [doi: 10.1109/ITNG.2012.34]

[30] Blumer A, Blumer J, Haussler D, *et al*. The smallest automaton recognizing the subwords of a text. Theoretical Computer Science, 1985,40(1):31−55. [doi:10.1016/0304-3975(85)90157-4]

[31] Crochemore M. Transducers and repetitions. Theoretical Computer Science, 1986,45(1):63−86. [doi: 10.1016/0304-3975(86)90041-1]

[32] Inenaga S, Takeda M, Shinohara A, *et al*. The minimum DAWG for all suffixes of a string and its applications. LNCS, 2002. 153−167. [doi: 10.1007/3-540-45452-7_14]

[33] Inenaga S, Bannai H, Shinohara A, *et al*. Discovering best variable-length-don't-care patterns. In: Proc. of the Discovery Science. 2002. 86−97. [doi: 10.1007/3-540-36182-0_10]

[34] Lothaire M. Algebraic Combinatorics on Words. Cambridge University Press, 2002.

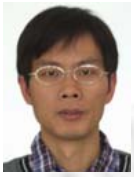
附中文参考文献:

[2] 强继朋,谢飞,高隽,等.带任意长度通配符的模式匹配.自动化学报,2014,40(11):2499−2511. [doi: 10.3724/SP.J.1004.2014.02499]

[8] 柴欣,贾晓菲,武优西,等.一般间隙及一次性条件的严格模式匹配.软件学报,2015,26(5):1096−1112. http://www.jos.org.cn/1000-9825/4707.htm [doi: 10.13328/ j.cnki.jos.004707]

[10] 武优西,刘亚伟,郭磊,吴信东.子网树求解一般间隙和长度约束严格模式匹配.软件学报,2013,24(5):915−932. http://www.jos.org.cn/1000-9825/4381.htm [doi: 10.3724/SP.J.1001.2013.04381]

[15] 汪浩,王海平,吴信东.带有通配符和长度约束的模式匹配问题求解模型.计算机科学,2016,43(4):279−283.

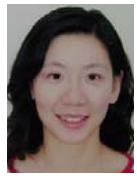


周开来(1978－),男,湖北南漳人,博士,副教授,主要研究领域为数据库,数据挖掘,并行计算.

陈 红(1965－),女,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为数据仓库与数据挖掘,物联网中的数据管理.

熊子绎(1995－),男,硕士生,主要研究领域为数据库,并行计算.

李翠平(1972－),女,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为社会网络分析,社会推荐,大数据分析和挖掘.



孙 辉(1977－),女,博士,讲师,CCF 专业会员,主要研究领域为数据库与数据挖掘,并行计算.