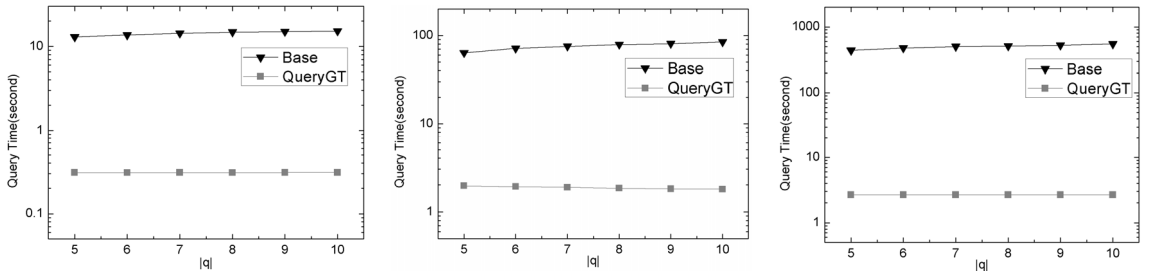
点集的增大,查询结果将有可能变成核值更小、覆盖顶点更多的连通区域,这将使查询结果集变得更大,而两种算法的运行时间均与查询结果集大小相关,故查询时间会随之增加.

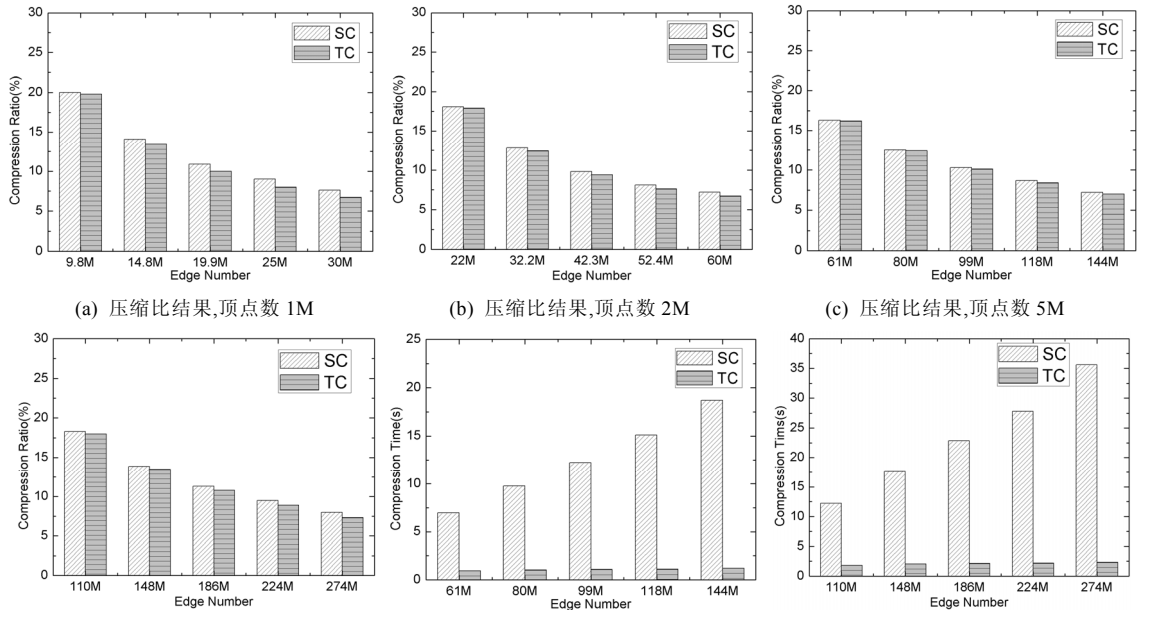


(a) Gowalla 数据集 (b) roadNet-TX 数据集 (c) as-skitter 数据集

Fig.2 Query time of Base vs. QueryGT on real datasets

图 2 真实图数据上查询算法 Base 和 QueryGT 的对比

图 3(a)~图 3(d)给出了 SC 和 TC 算法在虚拟图上的压缩比,为了验证算法的扩展性,分别在顶点数为 1M~10M 时,将虚拟图的边数从 9.8M 增加至 274M.



(a) 压缩比结果,顶点数 1M (b) 压缩比结果,顶点数 2M (c) 压缩比结果,顶点数 5M

(d) 压缩比结果,顶点数 10M (e) 压缩时间,顶点数 5M (f) 压缩时间,顶点数 10M

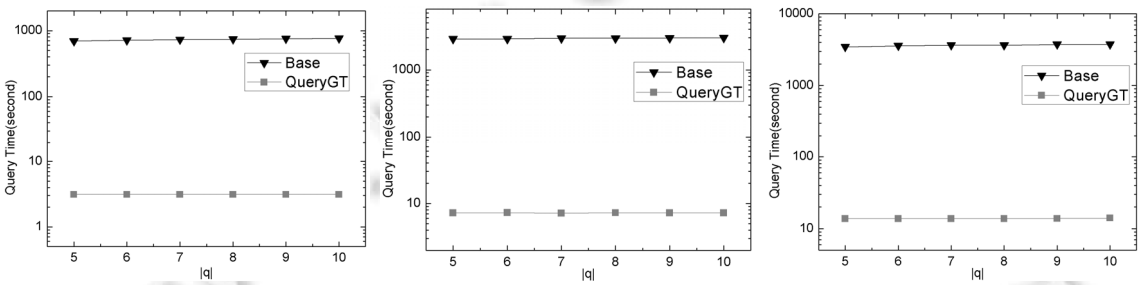
Fig.3 Compression ratio and compression time of SC, TC on synthetic datasets

图 3 虚拟图数据上 SC, TC 算法的压缩比与压缩时间

实验结果表明:图压缩算法在虚拟图上仍然就有很好的压缩效果;且随着平均度数的增加,压缩图的规模更小.即,算法更适用于稠密图.首先可以发现:压缩算法在顶点数为 1M~10M 的虚拟图上都取得了平均低于 12%的压缩比,这说明压缩算法可以将虚拟图有效地压缩;当虚拟图的平均度数为 20 左右时(即边数为 19.9M, 42.3M,99.3M,224M 时),两张虚拟图压缩后的规模均在原始图规模的 10%左右.其次,当顶点数固定、边数增加时,算法的压缩比从 19.8%降至 6.7%,且保持了单调下降的趋势.这说明压缩算法更加适用于稠密图.值得注意的一点是:在虚拟图上,SC 和 TC 算法的压缩效果极为接近,这是由于在偏好模型下,少数度数较大的顶点将构成一

个等价类,而剩余小度数的顶点经过 SC 算法的压缩将非常稀疏,即接近于树,因此 TC 算法无法将上述压缩图进一步大幅度压缩.这说明 SC 算法在本文所使用的虚拟图模型下更加适用.图 3(e)和图 3(f)给出了 SC,TC 算法在虚拟图上的压缩时间,实验结果表明,两种算法的压缩时间仍然都很快速.对于顶点数为 5M 和 10M、边数为 61M~274M 的虚拟图,压缩算法可在 40s 内完成.

图 4 给出了 Base 算法与基于压缩图的 QueryGT 算法在虚拟图上的查询时间对比,其中,查询顶点集是 1 000 组随机生成的、大小从 5 增加至 10 的顶点集,查询时间是 1 000 组查询时间的总和.实验结果表明:(1) QueryGT 的查询效率比 Base 提高了两个数量级;(2) 两种算法的查询时间均随着给定查询顶点集大小$|Q|$的增大而略有增加.首先可以发现:当顶点数分别为 2M,5M 和 10M、边数分别为 22M,80M 和 110M 时,QueryGT 均比 Base 快两个数量级;且对于更加稠密的前者,查询效率提升的更加明显.这是由于图压缩算法对于稠密图会取得更好的压缩效果,从而令查询效率得到更大提升.其次,当查询顶点集大小$|Q|$从 5 增加至 10 时,两种算法的查询时间均有微小的增加.这是因为随着查询顶点集的增大,查询结果将有可能变成核值更小、覆盖顶点更多的连通区域,故查询时间会随之增加.



(a) 顶点数 2M,边数 22M (b) 顶点数 5M,边数 80M (c) 顶点数 10M,边数 110M

Fig.4 Query time of Base vs. QueryGT on synthetic datasets

图 4 虚拟图数据上查询算法 Base 和 QueryGT 的对比

综上所述,本文所提出的图压缩算法无论在真实还是虚拟数据集上都能取得非常可观的压缩效果,且当原始图变得稠密时,算法的压缩效果更加显著.而基于压缩图的查询性能也得到了巨大的提升.

5 结束语

本文研究了基于图压缩技术的 k-MSC 查询处理算法,提出了图压缩算法 SC 及在查询转换算法,证明了基于图压缩算法 SC 的查询的正确性.考虑到 k-MSC 查询仅需要找到符合要求的连通区域,本文提出了图压缩算法 TC 将 SC 算法得到的压缩图进一步压缩为树.本文证明了基于压缩树的查询的正确性,并给出基于压缩树的无需解压缩的查询算法.通过真实和虚拟数据上的实验结果表明:所提出图压缩算法的压缩比平均可达到 12%;而对于稠密图,算法将取得更好的接近 10%的压缩比.基于压缩算法的查询效率也得到了很好的提升,与直接在原始图上查询的 Base 算法相比,查询效率提高了 1~2 个数量级.在今后的工作中,我们将进一步探讨针对本文压缩方法的更新技术.

致谢 在此,我们向曾经对本文提出宝贵审稿建议的审稿专家以及哈尔滨工业大学计算机科学与技术学院的李建中教授表示衷心的感谢.

References:

[1] Smith C. By the numbers: 98 amazing facebook statistics. DMR, 2014.

[2] Agrawal R, Rajagopalan S, Srikant R, Xu Y. Mining newsgroups using networks arising from social behavior. In: Proc. of the WWW 2003. 2003.

[3] Lappas T, Liu K, Terzi E. Finding a team of experts in social networks. In: Proc. of the KDD 2009. 2009.

[4] Yan X, Zhou XJ, Han J. Mining closed relational graphs with connectivity constraints. In: Proc. of the KDD 2005. 2005.

[5] Asthana S, King OD, Gibbons FD, Roth FP. Predicting protein complexmembership using probabilistic network reliability. Genome Research, 2004,14(6):1170−1175.

[6] Hanneman RA, Riddle M. Introduction to Social Network Methods. University of California, Riverside, 2005.

[7] Seidman SB. Network structure and minimum degre. Social Networks, 1983,5:269−287.

[8] Bollobas B. The evolution of sparse graphs, in graph theory and combinatorics. In: Proc. of the Cambridge Combinatorial Conf. in Honor of Paul Erdos. Academic Press, 1984. 35−57.

[9] Kortsarz G, Peleg D. Generating sparse 2-spanners. Journal of Algorithms, 1994,17(2):222−236.

[10] Andersen R, Chellapilla K. Finding dense subgraphs with size bounds. In: Proc. of the WAW. 2009. 25−37.

[11] Batagelj V, Zaversnik M. An $o(m)$ algorithm for cores decomposition of networks. Computer Science, 2003,1(6):34−37.

[12] Boldi P, Vigna S. The webgraph framework i: Compression techniques. In: Proc. of the 13th Int'l Conf. on World Wide Web. ACM Press, 2004. 595−602.

[13] Chierichetti F, Kumar R, Lattanzi S, Mitzenmacher M, Panconesi A, Raghavan P. On compressing social networks. In: Proc. of the 15th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2009. 219−228.

[14] Maserrat H, Pei J. Neighbor query friendly compression of social networks. In: Proc. of the 16th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2010. 533−542.

[15] Fan W, Li J, Wang X, Wu Y. Query preserving graph compression. In: Proc. of the 2012 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2012. 157−168.

[16] Buneman P, Grohe M, Koch C. Path queries on compressed XML. In: Proc. of the 29th Int'l Conf. on Very Large Data Bases—Vol.29. In: Proc. of the VLDB Endowment. 2003. 141−152.

[17] Navlakha S, Rastogi R, Shrivastava N. Graph summarization with bounded error. In: Proc. of the 2008 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2008. 419−432.

[18] Newman ME, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004,69(2):026113.

[19] Huang X, Cheng H, Qin L, Tian W, Yu JX. Querying k-truss community in large and dynamic graphs. In: Proc. of the SIGMOD 2014. 2014.

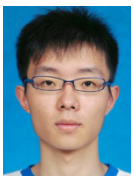
[20] Cui W, Xiao Y, Wang H, Lu Y, Wang W. Online search of overlapping communities. In: Proc. of the SIGMOD 2013. 2013.

[21] Chang L, Lin X, Qin L, Yu JX, Zhang W. Index-Based optimal algorithms for computing steiner components with maximum connectivity. In: Proc. of the 2015 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2015. 459−474.

[22] Gibbons A. Algorithmic Graph Theory. Cambridge University Press, 1985.

[23] Aho AV, Hopcroft JE, Ullman JD. On finding lowest common ancestors in trees. In: Proc. of the STOC'73. 1973.

[24] Bollobás B, Riordan O. The diameter of a scale-free random graph. Combinatorica, 2004,24(1):5−34.



李鸣鹏(1989－),男,黑龙江勃利人,硕士,主要研究领域为图数据的查询处理.

邹兆年(1979－),男,博士,讲师,CCF 会员,主要研究领域为图数据挖掘.

高宏(1966－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据挖掘,无线传感器网络.