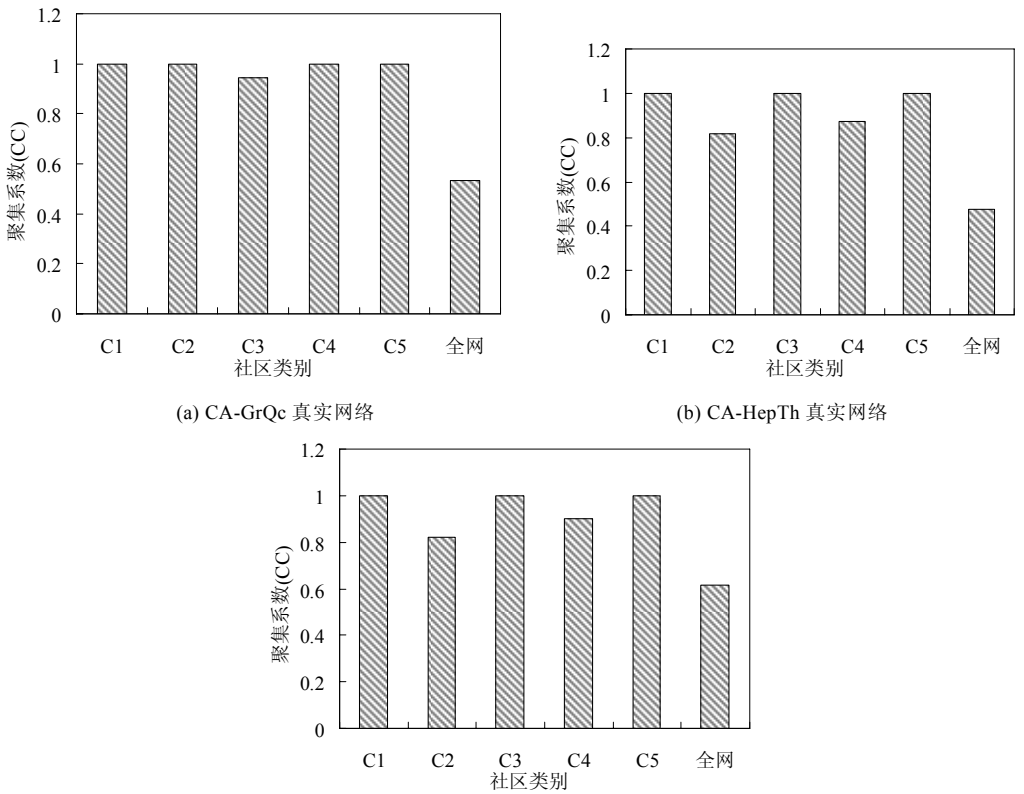
$$\bar{C}=\frac{1}{n}\sum_{i=1}^{n}C_i \tag{15}$$

若一个社区内所有节点的平均 *CC* 远大于整个网络作为一个社区时所有节点的平均 *CC*,说明所识别出的社区结构是有意义的,进而说明了社区识别质量很高.

以图 7(a)为例进行说明,图 7(b)、图 7(c)的情况类似.图 7(a)显示了 CA-GrQc 随机选取社区的聚集系数及全网聚集系数对比,其中:社区 *C*1,*C*2,*C*4,*C*5 的聚集系数达到 1,说明这一社区内部任意一个节点的邻接节点互相之间均有边相联系;*C*3 社区的聚集系数也达到了极高的 0.944 4.同时,全网的聚集系数为 0.532 0.说明网络的社区特征并不是很明显的情况下,算法仍能较准确地识别网络的社区结构,具有较高的识别质量.



(a) CA-GrQc 真实网络

(b) CA-HepTh 真实网络

(c) CA-HepPh 真实网络

Fig.7 Clustering coefficient of DOC algorithm under SNAP real networks

图 7 DOC 算法在 SNAP 真实网络大数据下社区聚集系数

4.4 算法运行时间性能分析

本节将通过对比不同算法在 LFR 基准数据集上的实验效果来验证本文所提算法的时间性能优势.

图 8 表示在 LFR 基准数据集(数据集 *A*)上 DOC 算法的运行时间近似线性变化,与本文第 3.4 节所提的时间复杂度 $O(n\log^2(n))$吻合,明显优于传统算法的时间复杂度.其他数据集上的实验结果类似,这里不再赘述.图 8 说明,DOC 相比于 COPRA 和 SLPA 算法在时间性能上有极大的提高.原因在于:DOC 算法通过建立平衡二叉树、对模块度增量建立索引,使得每次算法寻找最大的模块度增量的复杂程度降低.因为 CONGA 算法时间复杂度较高,为 $O(m^2n)$,最坏情况为 $O(m^3)$,算法运行时间性能较差,本节实验没有给出算法运行时间.

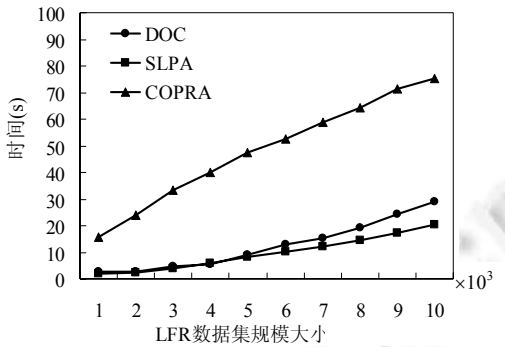


Fig.8 Runtime comparison of different algorithms under LFR benchmark network dataset *A*

图 8 不同算法在 LFR 基准数据集 *A* 上运行时间对比

5 结束语

本文提出了针对复杂网络大数据的重叠社区检测算法,算法基于模块度和图计算的思想,采用了新的节点和边的更新方法,利用平衡二叉树优化经典无重叠社区发现 Fast-Newman 算法,提高了节点更新的效率.进行大量实验结果后的表明:本文所提出的算法能够准确地检测重叠社区节点,同时极大地降低了算法的时间复杂度.

未来工作包括:将算法应用于为真实世界的各类复杂网络大数据中,提供社区识别服务,进而给用户提供包括兴趣点推荐等多种个性化服务.因为由于网络用户信息的不断更新,设计新算法实现社区的实时检测.此外,将在静态社区检测基础上设计动态网络社区检测算法,提高社区检测算法在实际网络中的应用价值.

References:

[1] Barabási A, Albert R, Jeong H, Bianconi G. Power-Law distribution of the World Wide Web. Science, 2000,287(5461):2115. [doi: 10.1126/science.287.5461.2115a]

[2] Wang YZ, Jin XL, Cheng XQ. Network big data: Present and future. Chinese Journal of Computers, 2013,36(6):1125−1138 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2013.01125]

[3] Gregory S. Finding overlapping communities in networks by label propagation. New Journal of Physics, 2010,12(10):103018. [doi: 10.1088/1367-2630/12/10/103018]

[4] Xie JR, Szymanski BK, Liu XM. SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In: Proc. of the 2011 IEEE 11th Int'l Conf. on Data Mining Workshops. Washington: IEEE, 2011. 344−349. [doi: 10.1109/ICDMW.2011.154]

[5] Gregory S. An algorithm to find overlapping community structure in networks. In: Proc. of the European Conf. on Principles of Data Mining and Knowledge Discovery. Berlin, Heidelberg: Springer-Verlag, 2007. 91−102. [doi: 10.1007/978-3-540-74976-9_12]

[6] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physcial Review E, 2004,69(2):026111. [doi: 10.1103/PhysRevE.69.026113]

[7] Newman MEJ. Fast algorithm for detecting community structure in networks. Physical Review E, 2004,69(6):066133. [doi: 10.1103/PhysRevE.69.066133]

[8] Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Physcial Review E, 2004,70(6):066111. [doi: 10.1103/PhysRevE.70.066111]

[9] Zhang XW, You HB, Zhu W, Qiao SJ, Li JW, Gutierrez LA, Zhang Z, Fan XN. Overlapping community identification approach in online social networks. Physica A, 2015,421:233−428. [doi: 10.1016/j.physa.2014.10.095]

[10] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008,30(2):155−168. [doi: 10.1088/1742-5468/2008/10/P10008]

[11] Oliveira JEM, Quiles MG. Communities detection in complex networks using coupled kuramoto oscillators. In: Proc. of the 14th Int'l Conf. on Computational Science and Its Applications. Berlin, Heidelberg: Springer-Verlag, 2014. 85−90. [doi: 10.1109/ICCSA.2014.25]

[12] Chen DB, Shang MS, Lv ZH, Fu Y. Detecting overlapping communities of weighted networks via a local algorithm. Physica A, 2010,389(19):4177−4187. [doi: 10.1016/j.physa.2010.05.046]

[13] Raghavan UN, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E, 2007,76(3):036106. [doi: 10.1103/PhysRevE.76.036106]

[14] Lancichinetti A, Fortunato S, Kertesz J. Detecting the overlapping and hierarchical community structure of complex networks. New Journal of Physics, 2008,11(15):19−44. [doi: 10.1088/1367-2630/11/3/033015]

[15] Staudt CL, Meyerhenke H. Engineering parallel algorithm for community detection in massive networks. IEEE Trans. on Parallel and Distributed Systems, 2015,27(1):171−184. [doi: 10.1109/TPDS.2015.239063]

[16] Nicosia V, Mangioni G, Carchiolo V, Malgeri M. Extending the definition of modularity to directed graphs with overlapping communities. Journal of Statistical Mechanics: Theory and Experiment, 2009,2009(3):P03024. [doi: 10.1088/1742-5468/2009/03/P03024]

[17] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graghs for testing community detection algorithm. Physical Review E, 2008, 78(4):046110. [doi: 10.1103/PhysRevE.78.046110]

[18] Danon L, Diaz-Guilera A, Duch J, Arenas A. Comparing community structure identification. Journal of Statistical Mechnics Theroy and Experiment, 2005,2005(9):P09008. [doi: 10.1088/1742-5468/2005/09/P09008]

附中文参考文献:

[2] 王元卓,靳小龙,程学旗.网络大数据:现状与展望.计算机学报,2013,36(6):1125−1138. [doi: 10.3724/SP.J.1016.2013.01125]

乔少杰(1981－),男,山东招远人,博士,教授,CCF 高级会员,主要研究领域为大规模社区发现,移动对象数据库,轨迹数据挖掘.

韩楠(1984－),女,博士,讲师,主要研究领域为社区发现,移动对象数据库,生物信息学.

张凯峰(1994－),男,主要研究领域为社区发现.

邹磊(1981－),男,博士,副教授,CCF 高级会员,主要研究领域为图数据管理.

王宏志(1978－),男,博士,教授,CCF 高级会员,博士生导师,主要研究领域为大数据,数据管理.

Louis Alberto GUTIERREZ (1980－),男,博士,Researcher,主要研究领域为数据挖掘.