软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(3):631−647 [doi: 10.13328/j.cnki.jos.005155]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

复杂网络大数据中重叠社区检测算法*

乔少杰[1], 韩 楠[2], 张凯峰[3], 邹 磊[4], 王宏志[5], Louis Alberto GUTIERREZ[6]

[1](成都信息工程大学 信息安全工程学院,四川 成都 610225)

[2](成都信息工程大学 管理学院,四川 成都 610103)

[3](西南交通大学 信息科学与技术学院,四川 成都 611756)

[4](北京大学 计算机科学技术研究所,北京 100871)

[5](哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150006)

[6](Department of Computer Science, Rensselaer Polytechnic Institute, New York, USA)

通讯作者: 韩楠, E-mail: hannan722@163.com

摘 要: 提出一种新的面向复杂网络大数据的重叠社区检测算法 DOC(detecting overlapping communities over complex network big data),时间复杂度为 $O(n\log^2(n))$,算法基于模块度聚类和图计算思想,应用新的节点和边的更新方法,利用平衡二叉树对模块度增量建立索引,基于模块度最优的思想设计一种新的重叠社区检测算法.相对于传统的重叠节点检测算法,对每个节点分析的频率大为降低,可以在较低的算法运行时间下获得较高的识别准确率.复杂网络大数据集上的算法测试结果表明:DOC 算法能够有效地检测出网络重叠社区,社区识别准确率较高,在大规模 LFR 基准数据集上其重叠社区检测标准化互信息指标 NMI 最高能达到 0.97,重叠节点检测指标 F-score 的平均值在 0.91 以上,且复杂网络大数据下的运行时间明显优于传统算法.

关键词: 复杂网络;大数据;重叠社区检测;模块度;图计算

中图法分类号: TP311

中文引用格式: 乔少杰,韩楠,张凯峰,邹磊,王宏志,GUTIERREZ LA.复杂网络大数据中重叠社区检测算法.软件学报,2017,28(3):631−647. http://www.jos.org.cn/1000-9825/5155.htm

英文引用格式: Qiao SJ, Han N, Zhang KF, Zou L, Wang HZ, Gutierrez LA. Algorithm for detecting overlapping communities from complex network big data. Ruan Jian Xue Bao/Journal of Software, 2017,28(3):631−647 (in Chinese). http://www.jos.org.cn/1000-9825/5155.htm

Algorithm for Detecting Overlapping Communities from Complex Network Big Data

QIAO Shao-Jie[1], HAN Nan[2], ZHANG Kai-Feng[3], ZOU Lei[4], WANG Hong-Zhi[5], Louis Alberto GUTIERREZ[6]

[1](College of Information Security Engineering, Chengdu University of Information Technology, Chengdu 610225, China)

[2](School of Management, Chengdu University of Information Technology, Chengdu 610103, China)

[3](School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China)

[4](Institute of Computer Science and Technology, Peking University, Beijing 100871, China)

[5](Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150006, China)

[6](Department of Computer Science, Rensselaer Polytechnic Institute, New York, USA)

* 基金项目: 国家自然科学基金(61100045, 61363037); 教育部人文社会科学研究规划基金(15YJAZH058); 教育部人文社会科学研究青年基金(14YJCZH046); 成都市软科学项目(2015-RK00-00059-ZF); 四川省教育厅资助科研项目(14ZB0458)

Foundation item: National Natural Science Foundation of China (61100045, 61363037); Planning Foundation for Humanities and Social Sciences of Ministry of Education of China (15YJAZH058); Youth Foundation for Humanities and Social Sciences of Ministry of Education of China (14YJCZH046); Soft Science Foundation of Chengdu (2015-RK00-00059-ZF); Foundation of Educational Commission of Sichuan Province (14ZB0458)

收稿时间: 2016-07-15; 修改时间: 2016-09-14; 采用时间: 2016-11-01; jos 在线出版时间: 2016-11-29

CNKI 网络优先出版: 2016-11-29 13:34:57, http://www.cnki.net/kcms/detail/11.2560.TP.20161129.1334.002.html

Abstract: Currently, the number of Internet users, along with complex networks including online social networks and electronic commerce networks, is growing explosively. To effectively and efficiently detecting overlapping community structure from complex network, big data plays an essential role in point of interest recommendation and hotspot propagation. In this study, a new algorithm over complex networks is proposed to detecting overlapping communities with a time complexity of $O(n\log^2(n))$. The algorithm applies a new method for updating node and edge modularity based on the techniques of modularity clustering and graph computing. Balanced binary tree is used to index the modularity increment, and an overlapping community detection approach is provided based on the idea of modularity optimization to reduce the frequency of node analysis compared to traditional approaches. Experiments are conducted on real complex network big data, and the results show that the DOC algorithm can effectively detect overlapping communities with high accuracy, the normalized mutual information (NMI) can reach to 0.97 in large-scale LFR benchmark datasets, and the overlapping community detecting standard F-score value is averagely higher than 0.91. In addition, the runtime efficiency beats traditional approaches in complex network big data.

Key words: complex network; big data; overlapping community detection; modularity; graph computing

随着互联网、物联网技术的快速发展,事物之间的联系更加紧密,错综复杂的联系形成了多样、多变、规模庞大的网络,例如人际交往形成的复杂社交网络、蛋白质交互网络、基于地理空间的交通网络、城市路线网络等.上述网络因其结构复杂、网络进化、连接和节点的多样性、多重复杂性融合,被称为复杂网络[1].复杂网络在规模与复杂度上的快速增长,演变成网络大数据[2].在现实网络中,社区重叠是复杂网络大数据中另一重要特征,即,不同社区之间具有重叠的节点.重叠社区的检测对于网络结构分析、社区划分等具有重要研究价值和科学意义.值得注意的是,国家重点基础研究发展计划(973)和重大科学研究计划将社交网络结构分析的基础研究作为重要支持方向.

复杂网络大数据中,社区发现算法的研究涉及社会学、生物学、计算机等交叉学科,具有广阔的应用前景.例如在生物学方面,社区检测可以从蛋白质、新陈代谢网络中提取信息,帮助了解生命的奥秘.本文的主要研究动机包括:

1) 早期社区发现研究工作很少考虑重叠节点,建立在节点只属于某一社区的假设之上.然而,在网络大数据中,社区之间重叠是其重要结构特征,考虑网络节点的重叠性可以极大地提高算法的准确性;
2) 传统的非重叠社区发现算法已经不能满足对现实网络分析的要求.现有的重叠社区检测算法时间复杂性较高,应用于大规模复杂网络数据时,其劣势相当明显.当网络节点规模上万,节点连接关系更加复杂的情况下,甚至无法对社区进行划分;
3) 现有重叠社区检测算法很难兼顾算法的准确性和实效性.

针对上述不足,本文的主要贡献包括:基于模块度思想和图论知识,应用新的网络模块度更新方法和社区合并方法,采用平衡二叉树对其进行优化,其节点间模块度增量更新算法时间复杂度仅为 $O(\log^2(n))$,整体算法时间复杂度为 $O(n\log^2(n))$,其中,n 表示节点的个数;在非重叠网络社区检测算法得到的社区划分基础上,提出了一种新的重叠社区检测算法,降低了每个节点识别的时间代价,算法的复杂度仅为 $O(n)$;为该类问题提供一种新的思路,即,将重叠节点的检测作为了一个分类问题进行研究;将所提算法与经典的社区识别算法 COPRA 算法(clustering overlap propagation algorithm)[3]、SLPA 算法(speaker-listener label propagation algorithm)[4]和 CONGA 算法(cluster-overlap newman girvan algorithm)[5]进行了对比实验,从多角度验证所提方法的性能优势.

1 相关工作

复杂网络大数据中,社区发现算法是根据网络的拓扑结构以及节点属性的相似性,将网络进行模块划分的方法,通常情况下,找到这类划分方法精确解是一个 NP 难问题.按照是否考虑重叠节点划分将社区发现算法分为两类:未考虑重叠节点和考虑重叠节点社区发现算法,主要工作如下.

(1) 未考虑重叠节点的社区发现算法

基于模块度最优思想的凝聚类算法成为目前网络社区挖掘方法的主流,经典算法包括 Fast-Newman 算法[6]和 CNM 算法[7]等.Fast-Newman 算法基于最大化模块度的贪婪思想,归并能够产生最大模块度增量的两个社区,

直到任意两个社区所产生的模块度增量均小于0时终止.Clauset等人[8]提出了CNM算法,对Fast-Newman算法中节点归并操作进行优化,同时采用大根堆等数据结构对算法进行改进.Zhang等人[9]重新设计模块度的求解方法,改进了Fast-Newman算法的更新策略,较好地提高了社区识别的效率.Blondel等人[10]对模块度增量的求解方法进行改进,设计新的模块度增量计算公式,迭代合并社区,取得了良好的社区识别效果. Oliveira等人[11]利用改进的Kuramoto耦合振子同步模型,从动力学因素分析网络,实现了复杂网络中的社区发现.基于硬聚类算法思想,Chen等人[12]提出了一种在初步划分下的重叠社区检测算法,对节点隶属度计算进行改进,优点在于最大化重叠社区划分的模块度,但计算过程复杂度较高,不适用于复杂网络大数据.基于标签传播的社区识别经典算法是未考虑重叠社区检测的硬聚类算法RAK[13],算法通过赋予每个节点一个社区标签,不断在邻接节点间进行传播,最终收敛到一个较好的解.除了上述算法,还有一些社区划分效果较好的算法,如LFM算法[14],通过设定适应度函数,每次以一个节点作为开始,不断扩充自身的社区,使得适应度函数值达到局部最大值,能够较好地检测复杂网络的社区结构.但是由于其适应度函数的选取将会影响社区的大小,使得算法不稳定.Staudt等人[15]基于内存共享的思想提出了并行标签传播算法,对大规模网络数据取得较好的识别效果.

(2) 重叠社区检测算法

Gregory等人对G-N算法进行改进,提出了CONGA算法[5],通过对节点自身进行分裂,在两个分裂节点间添加虚边,所提方案能够较准确地检测重叠社区,但是由于该算法的时间复杂度较高,实际应用价值不高.近年来,研究人员针对RAK改进设计了多种能够检测重叠社区的方法,其中代表性算法包括COPRA[3]及算法SLPA[4]. COPRA算法通过设置参数来确定概率接受准则,进而通过每次传播的标签概率是否满足相应概率接受准则来确定是否接受该标签,当所有标签都无法满足概率接受准则的时候,就随机选取概率最大的标签之一,使得算法能够检测重叠社区.但是由于随机选取标签的原因,将会导致算法收敛性能较差,社区划分的结果不够稳定. SLPA算法增强了标签传播的收敛性能,但是由于要设置不同的概率准则以及参数才能选取相应的标签,导致算法在实际应用过程中需要对不同的网络进行参数调节,适用范围较窄.

为了克服传统社区发现算法复杂度较高等不足,本文提出了一种新的考虑重叠社区检测的大规模复杂网络社区划分方法,在极大地降低算法运行时间复杂度的同时,保证重叠社区检测的准确性.

2 基本概念

本节对算法中使用的基本概念进行描述.

定义 1(复杂网络). $CN=(V,E), V=\{\sum v_i \mid i=1,2,...,n\}$ 表示网络中节点的集合, $E=\{\sum e_i \mid i=1,2,...,m\}$ 表示网络中边的集合.其中,$v=(o,a)$表示节点,o 表示网络中的实体,a 表示实体的属性;$e=(p,q,r)$表示网络中一条边,p 和 q 分别表示两个不同实体,r 表示实体间的关系.复杂网络的复杂特性表现在:

1) 结构复杂:节点数目巨大,网络结构呈现多种不同特征;
2) 网络进化:节点或连接的产生与消失;
3) 连接多样性:节点之间的连接权重存在差异,且有可能存在方向性;
4) 动力学复杂性:节点集可能属于非线性动力学系统,例如节点状态随时间发生复杂变化;
5) 节点多样性:复杂网络中的节点可以代表任何事物;
6) 多重复杂性融合:上述多重复杂性相互影响,导致更为难以预料的结果.

定义 2(复杂网络大社区). 大社区是复杂网络中节点的集合,社区内部节点连接紧密,社区之间节点连接较为稀疏.随着复杂网络规模增大,社区也不断增大(velocity),社区内节点数目不断增多(volume),社区内部和社区之间的关系变得非常复杂,具备网络大数据的复杂性,即,数据类型的复杂性、数据结构的复杂性和数据内在模式的复杂性,复杂关系中隐藏着有价值的社会关系网交互信息(value).大的社区不断吞并小社区,小社区不断消亡,社区变得更加广阔(variety),此类社区结构称为复杂网络大社区.

定义 3(重叠社区). 重叠社区是网络中节点的集合,社区内节点同时隶属于多个不同的社区,社区内部节点间的联系较为紧密,而属于不同社区的节点之间的联系较为稀疏,此类社区称为重叠社区.

如图 1 所示,节点 5 同时属于社区 1 和社区 2,节点 8 同时属于社区 2 和社区 3,图中的 3 个社区被称为重叠社区.

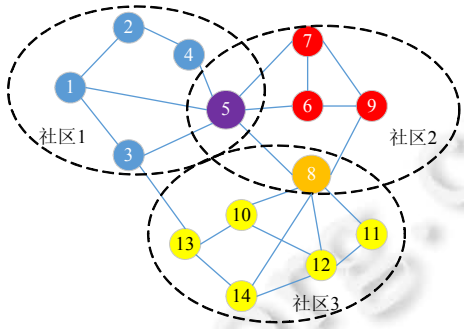


Fig.1 Example of overlapping communities

图 1 重叠社区举例

传统的重叠社区检测算法无法高效地对复杂网络中重叠节点进行检测,主要困难在于:传统算法时间性能较差,无法应对节点数目众多以及边信息复杂的网络,当节点规模较大时,往往导致算法无法正常运行.

定义 4(节点度). 已知一个节点数为 n、边数为 m 的复杂网络 $CN(V,E)$中,V 表示节点集合,E 表示边的集合.如果节点 u 与节点 v 中间有边相连,则 e_{uv}=1;否则,e_{uv}=0.节点 u 的度 k_u 表示为 $k_u=\sum_{v\in V}e_{uv}$, 表示与节点 u 相连的边的数目.

注意:下文出现的参数 n 和 m 均表示网络中节点和边的数量.

定义 5(节点隶属度). 对于重叠网络中节点 u 和社区 c,节点隶属度表征节点对社区的隶属程度,定义为

$$B(u,c)=\frac{k_u^c}{k_u} \tag{1}$$

其中,k_u^c 表示节点 u 在社区 c 中关联边的度,k_u 表示节点 u 的度.

从图 1 可以看出,节点 8 对社区 2 和社区 3 的隶属度分别为 0.33 和 0.67.

定义 6(模块度). 网络的模块度 Q 定义为[6]

$$Q=\sum_{i=1}^{k}(e_{ii}-a_i^2) \tag{2}$$

$$Q=\frac{1}{2m}\sum_{c\in C}\sum_{u,v\in V}\delta_{cu}\delta_{cv}\left(A_{uv}-\frac{k_uk_v}{2m}\right) \tag{3}$$

公式(2)和公式(3)等价.公式(2)中:e_{ii} 表示在一个社区 i 内部,连接两个顶点的边数占网络总边数的比例;a_i 表示与社区 i 中节点相连,但另一个节点不属于社区 i 的边数占网络总边数的比例;k 表示社区数目.在公式(3)中,A 表示网络的邻接矩阵,k_u 和 k_v 分别表示节点 u 和 v 的度,C 表示所有社区构成的集合,m 表示网络的总边数.如果节点 u 在社区 c 中,则 δ_{cu}=1;否则,δ_{cu}=0.

模块度表征了网络中节点社区划分的精确程度.当一个网络在某种模块划分下的模块度 Q 达到最大值时,表明网络社区划分达到了最佳.然而,求解在所有网络社区划分状况下的模块度 Q 是一个 NP 难题,因此,Newman 提出了模块度增量的概念,其意义是:当社区 i 和社区 j 进行合并时,将产生的模块度增量定义为[7]

$$\Delta Q=2(e_{ij}-a_i\times a_j) \tag{4}$$

其中,$a_i=k_i/2m$,k_i 表示社区合并后的节点度数,m 为网络中边的数量.

3 复杂网络大数据中重叠社区检测算法

本文提出的面向复杂网络的重叠社区检测算法(detecting overlapping communities in complex networks,简

称 DOC)包含两个主要部分.

(1) 基于模块度理论和图论知识,在 Fast-Newman 算法的基础上设计新的网络模块度更新方法和社区合并方法.为了提高网络大数据挖掘效率,设计平衡二叉树索引模块度增量,得到无重叠的社区;

(2) 根据节点隶属度定义对网络中节点进行分类,对不同隶属度下的节点进行讨论,从而获得重叠节点部分.DOC 算法的工作原理如图 2 所示.

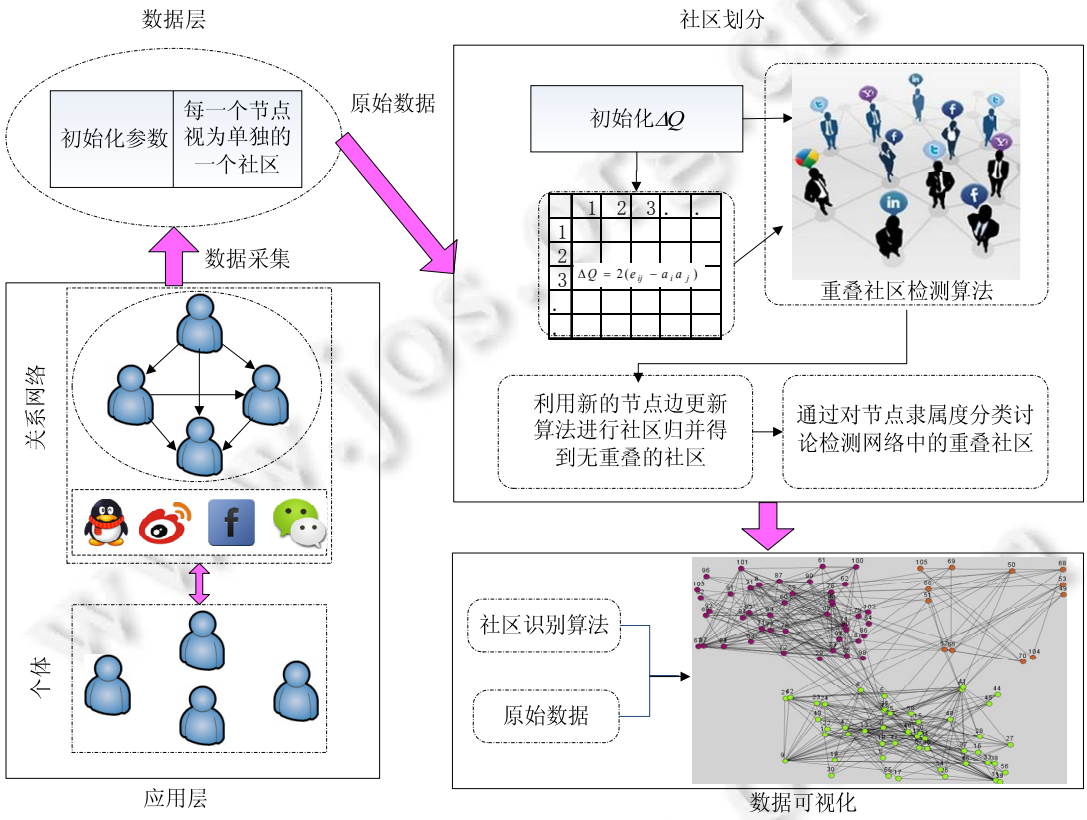


Fig.2 Working mechanism of the DOC algorithm

图 2 DOC 算法工作原理图

如图 2 所示:DOC 算法首先基于大规模复杂网络提取网络信息,形成由节点和边构成的网络;然后,基于 Fast-Newman 算法,利用新的节点归并方法以及新的数据结构对其进行优化,这一步未考虑重叠节点的情况;再者,利用重叠社区检测算法得到新的社区划分结果;最后,对社区划分的结果可视化输出.

3.1 理论基础

本节将通过概率论以及图计算理论介绍重叠社区检测算法的基本思想.利用模块度进行社区划分的基本思路为:如果一个网络的子图中节点之间的连接强度远大于网络随机划分下子图中节点之间连接的强度,那么可以认为该子图可以作为网络的一个社区.通过概率量化节点间的连接强度的过程如下.

对于随机无向图,图中任意两个节点之间具有连接边的概率 P_{ij} 是

$$P_{ij}=\frac{k_i\times k_j}{(2m)^2} \tag{5}$$

其中,k_i 表示节点 i 的度数,k_j 表示节点 j 的度数,m 表示图中的边的数量.

对于网络图 $CN(E,V)$,对于其中任意一个社区,能够计算出实际社区内部节点连接强度相对于随机划分下

社区内部节点间的连接强度的差值,节点间的连接强度反映到概率上的表达形式如下:

$$Q_c = \sum_{i,j \in c} \left[\frac{A_{ij}}{2m} - P_{ij} \right] \tag{6}$$

其中,A_{ij}表示在节点 i 和 j 之间的邻接关系,如果 i 和 j 之间有边连接,则 A_{ij}=1;反之,A_{ij}=0.

根据公式(6),可以得到基于整个网络的模块度[8]:

$$Q = \frac{1}{2m} \sum_{c \in C} \sum_{u,v \in V} \left(A_{uv} - \frac{k_u k_v}{2m} \right) \tag{7}$$

其中,A 表示网络邻接矩阵,k_u 和 k_v 表示节点 u 和 v 的度,C 表示所有社区构成的集合,m 表示网络边数量.

上述方法仅针对未考虑重叠节点的社区划分,其含义是:对于每一种社区划分方式,模块度越大,越能够证明社区划分方式的合理性.然而,当允许一个节点同时属于多个社区的时候,上述模块度将不再适用.因此,此处引入节点的模糊隶属函数.模糊隶属函数值表示每个特征向量同时属于多个聚类的“某种程度”,区间[0,1]中的隶属度函数相应量化这个程度.可以得到以下基于重叠社区检测的模块度表达式:

$$Q = \frac{1}{2m} \sum_{c \in C} \sum_{u,v \in V} \delta_{cu} \delta_{cv} \left(A_{uv} - \frac{k_u k_v}{2m} \right) \tag{8}$$

注意,公式(8)相对于公式(7)新增的两个隶属度因子δ_{cv}和δ_{cu}分别表示节点 v 和节点 u 相对于社区 c 的隶属程度.Nicosia 等人[16]通过大量实验对比不同的模糊隶属函数,认为模糊隶属函数使用公式(9)最为合适.

$$\delta_{cv} = \frac{1}{1 + \exp(-120 \times B(v,c) + 60)} \tag{9}$$

其中,$B(v,c)$表示节点 v 隶属于社区 c 的隶属度.

基于上述理论基础,可以得到以下推论.

推论 1. 针对某一社区,将其他社区的某一节点 u 加入该社区后,所产生的重叠社区模块度增量为[8]

$$\Delta Q_{ov} = \sum_{v \in C} \left[\frac{A_{uv}}{2m} - \frac{k_u \times k_v}{4m^2} \right] \tag{10}$$

推论 2. 对于有效重叠社区划分方法,应当满足重叠社区模块度不小于未考虑社区重叠的模块度,即满足:

$$\sum_{c \in C} \sum_{u,v \in V} \delta_{cu} \delta_{cv} \left(A_{uv} - \frac{k_u k_v}{2m} \right) \geqslant \sum_{c \in C} \sum_{u,v \in V} \left(A_{uv} - \frac{k_u k_v}{2m} \right) \tag{11}$$

根据推论 2,引出本文算法的基本思路:首先,通过传统的未考虑重叠社区的算法获取基本的网络划分;然后,通过最优化重叠社区模块度获得重叠社区节点部分.

3.2 基于模块度的社区划分方法

本节将介绍 DOC 算法的第 1 个步骤,不考虑社区重叠部分的社区划分方法,通过设计新的节点和边的更新方法,针对复杂网络大数据,利用平衡二叉树数据结构进行优化,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模块度增量的复杂度大为降低,时间复杂度为 $O(n\log^2(n))$.下面首先对节点和边的更新算法进行介绍.

Fast-Newman 算法在网络中节点 i 和节点 j 进行归并后,需要计算当前情况下两两节点归并所产生的模块度增量ΔQ_{ij}.本文针对这一过程进行改进:已知归并的两个社区为 i 和 j,其归并后的新社区编号为 k,此时,对于网络任意两个社区 p 和 q,当计算社区 p 和 q 归并的模块度增量时,其归并的模块度增量有以下 3 种情况(节点更新规则).

1) 当社区 p 和社区 i,j 均有连接或均无连接时,合并社区 p 和社区 k 产生的模块度增量为$\Delta Q_{pk}=\Delta Q_{pi}+\Delta Q_{pj}$.

证明:

∵由模块度定义可知,模块度增量与合并社区始末的网络社区划分情况有关;

考虑 3 个社区,即 i,j,p 合并所产生的模块度增量为ΔQ_{ijp},则在社区 i,j 合并为社区 k 后,社区 k 和社区 p 合并所产生的模块度增量公式:$\Delta Q_{pk}=Q_{末}-Q_{始}=(\Delta Q_{ijp}+Q_0)-(\Delta Q_{ij}+Q_0)$;

$\therefore \Delta Q_{pk}=\Delta Q_{ijp}-\Delta Q_{ij}$.

$\because \Delta Q_{ijp}=\Delta Q_{ij}+\Delta Q_{pj}+\Delta Q_{pi}$;

$\therefore \Delta Q_{pk}=\Delta Q_{pi}+\Delta Q_{pj}$. □

2) 当社区 p 仅与社区 i 有连接时,合并社区 p 和 k 产生的模块度增量公式为$\Delta Q_{pk}=\Delta Q_{pi}-2a_pa_j$.

证明:

$\because \Delta Q_{pk}=\Delta Q_{ijp}-\Delta Q_{ij}=(\Delta Q_{ij}+\Delta Q_{pj}+\Delta Q_{pi})-\Delta Q_{ij}$ 且$\Delta Q_{pj}=2(e_{pj}-a_pa_j)=-2a_pa_j$;

$\therefore \Delta Q_{pk}=\Delta Q_{pj}+\Delta Q_{pi}=\Delta Q_{pi}-2a_pa_j$. □

3) 当社区 p 仅与社区 j 有连接时,合并社区 p 和 i 产生的模块度增量公式为$\Delta Q_{pk}=\Delta Q_{pj}-2a_pa_i$.

DOC 算法利用新的节点和边更新算法以及平衡二叉树,使得每次更新边和节点时不需要计算全部模块度增量.在社区识别时,首先需要对参数初始化,如算法 1 所示,主要步骤:(1) 计算网络中各节点的度,保存在节点度数向量 k 中(第 1 行~第 4 行);(2) 计算各个节点合并产生的模块度增量,保存在矩阵ΔQ中(第 5 行~第 9 行);(3) 初始化平衡二叉树森林 T(见算法 2),使得每棵平衡二叉树 T_i 存储了ΔQ_i(第 10 行),并输出各参数(第 11 行).

算法 1. 参数初始化.

输入:复杂网络的邻接矩阵 A;

输出:各节点度数向量 k,ΔQ 矩阵以及初始化平衡二叉树 T.

```
1.  k_i=0;
2.  for i:=1 to n
3.      for j:=1 to n
4.          k_i=k_i+A_ij;
5.  for i:=1 to n
6.      for j:=1 to n
7.          if (A_ij≠0) then
8.              continue;
9.          ΔQ_ij=1/2m-k_ik_j/(2m)^2;
10. T=InitialBTree();
11. Output k,ΔQ 矩阵,T;
```

算法 2. 建立和更新平衡二叉树森林.

输入:初始ΔQ 矩阵;

输出:存有每行ΔQ 的平衡二叉树森林.

```
1.  for i:=1 to n
2.      for j:=1 to log2(n)
3.          InsertNode(ΔQ_ij);
4.          AdjustAVL(T_i);
5.  DeleteTree(T_j);
6.  for u:=1 to n
7.      P=Search(T_u,A(u,j));
8.      DeleteNode(P);
```

算法 2 的主要步骤如下:

(1) 平衡二叉树森林的建立算法:将ΔQ 矩阵的每一行均插入相应的平衡二叉树中,对平衡二叉树进行平衡调整(第 1 行~第 4 行);

(2) 平衡二叉树森林的更新算法:依据ΔQ 矩阵更新平衡二叉树森林.具体操作为:删除第 j 棵平衡二叉树(第 5 行),根据节点更新规则更新第 i 棵平衡二叉树.针对任意平衡二叉树 u,根据ΔQ 矩阵中ΔQ_{ui} 以及

ΔQ_{uj},依据更新规则更新、删除平衡二叉树中相应值(第 6 行~第 8 行).

初始化主要参数之后则进入第 2 个步骤,非重叠社区的检测,算法如下所示.

算法 3. 非重叠社区检测.

输入:邻接矩阵 A,ΔQ 矩阵,n 棵平衡二叉树构成的森林 F,节点度数向量 k;

输出:非重叠社区存储向量集合 C.

```
1.  for i:=1 to n
2.      C_i=i;
3.  while max{ΔQ}>0 do
4.      Merge(C_i,C_j);
5.      k_i=k_i+k_j;
6.      Update(A);
7.      Update(F);
8.      Update(ΔQ);
9.  Output C;
```

算法 3 首先将各个节点作为单独社区,即,赋予各个社区相应节点编号(第 1 行、第 2 行);从ΔQ 矩阵中查找最大的模块度增量所需合并的两个社区 i 和 j 的编号,将对应的两个社区合并(第 3 行、第 4 行);更新节点度数向量和邻接矩阵(第 5 行、第 6 行),依据ΔQ 矩阵利用新的规则(第 3.2 节第 3 段开始讨论的 3 种情况)更新平衡二叉树(第 7 行),更新模块度增量矩阵(第 8 行);重复进行上述步骤,直到最大模块度增量为负数;输出划分后的社区(第 9 行).

3.3 重叠社区检测算法

本节将介绍 DOC 算法的第 2 步,即重叠社区检测算法,基于第 3.2 节初步未考虑重叠社区的社区检测算法的聚类结果,对网络中各个节点进行隶属度讨论,从而获得网络的重叠社区划分.

根据第 3.1 节中介绍的重叠社区模块度公式(8),可以得到结论:当节点对某社区隶属度大于 0.55 时,其节点相应模糊隶属度函数公式(9)的值大于 0.997 5,接近于 1;当节点对某社区的隶属度小于 0.4 时,其模糊隶属度函数公式(9)的值小于 10^{-6},接近于 0.基于最优化重叠社区模块度的思想,可以通过对节点的隶属度分成 3 类讨论,最终确定各节点的社区.算法 4 通过计算节点对社区的隶属度,进而实现重叠节点检测,如下所示.

算法 4. 重叠社区检测.

输入:非重叠社区向量集合 C,社区数 p;

输出:重叠社区向量集合 C'.

```
1.  for i:=1 to p   //p 表示算法 3 划分后的社区数量
2.    for each v∈C_i   //遍历社区 C_i 中每一个节点
3.      for k:=1 to p
4.        if k==i then
5.          continue;
6.        for each u∈C_k
7.          B(v,C_k)=B(v,C_k)+A_vu/k_v;   //计算节点 v 相对社区 C_k 的隶属度
8.        if B(v,C_k)>0.55 then   //对应节点对社区隶属情况 1
9.          if v∉C_k then
10.           add this node to C_k;
11.       else
12.         if B(v,C_k)>=0.4 then   //对应节点对社区隶属情况 2
13.           if ΔQ_ov>0 then
```

14. add this node to C_k;
15. **else if** $v \in C_k$ **then**
16. separate this node from C_k;
17. **else**
18 **if** $v \in C_k$ THEN //对应节点对社区隶属情况 3
19. separate this node from C_k;
20. Output C';

算法 4 中,节点对某社区的隶属进行的讨论分成 3 种情况.

(1) 当节点对某社区的隶属度大于 0.55 时:若该社区为原始划分社区,则不必进行操作;若该社区不为原始划分社区,则将该节点加入相应社区;

(2) 当节点对某社区的隶属度介于[0.4,0.55]时,计算节点加入相应社区是否使重叠社区的模块度增加:若增加,则将该节点加入该社区;反之,该社区将不能拥有这一节点作为其社区元素.其中,应用公式(10)计算该节点加入相应社区所产生的模块度增量;

(3) 当节点对某社区的隶属度小于 0.4 时,该节点不能作为该社区的元素.

3.4 算法时间复杂度分析

为了说明 DOC 算法具有较好的时间性能,可用于处理网络大数据,本节将分析 DOC 算法的时间复杂性.首先获取非重叠社区算法:初始化ΔQ矩阵时间复杂度为 $O(m)$;初始化ΔQ平衡二叉树时间复杂度约为 $O(n\log(n))$,更新平衡二叉树复杂性是 $O(\log(n))$;选取最大模块度增量进行社区合并的算法时间复杂度为 $O(n\log^2(n))$.综合上述步骤,获取非重叠社区算法时间复杂度为 $O(n\log^2(n))$.值得注意的是:大多数社区发现算法并未考虑参数初始化过程时间复杂性(本文算法 1 的第 2 步~第 9 步参数初始化的复杂性是 $O(n^2)$),为了保证算法可比性,本文忽略参数初始化的时间复杂度.其次检测重叠社区算法:算法对每个节点对不同社区的隶属度进行讨论,通过分析可以得到算法 4 的时间复杂度为 $O(n\times p^2)$,因为社区的数目 p 为常数,所以重叠社区检测算法复杂度可粗略估计为 $O(n)$.综上,DOC 算法的总时间复杂度为 $O(n\log^2(n)+n)$,近似为 $O(n\log^2(n))$.

4 实验分析

4.1 数据集描述

为了验证本文所提的重叠社区检测算法社区划分的准确性和时间性能,实验采用不同的复杂网络大数据集进行实验:(1) 使用 LFR 基准程序生成的人工模拟的大规模复杂网络数据集[17];(2) 从 SNAP 网站上获取的 3 个真实社交网络.LFR 基准程序是由 Lancichinetti 等人提出生成人工模拟网络的程序,该程序在生成基准网络的同时生成含有类标的社区结果.本文算法生成网络参数设置见表 1,其中,社区混合参数μ取值范围为(0,1],表征了社区结构的明显程度,参数值越大,说明社区结构越不明显.

本文实验通过 LFR 基准程序生成了 2 个基准网络,分别为:

(1) 节点数目为 1 000~10 000 的 10 个大规模网络图,其社区混合参数μ=0.3,其社区重叠节点个数为 10~100 个,节点可归属社区个数为 2;

(2) 节点数量为 5 000、节点可归属社区个数为 2~8 个不等的网络.

表 2~表 4 给出了实验中所用数据集的详细信息.

Table 1 Parameter setting of LFR benchmark network generation

表 1 LFR 基准网络生成参数说明

参数	说明
n	网络的节点数目
k	网络中节点的平均度数
$C_{\min}$	最小社区的节点数目
$C_{\max}$	最大社区的节点数目
O_n	重叠节点的个数
om	重叠节点所从属的社区个数
μ	社区混合参数

Table 2 LFR benchmark network dataset *A*

表 2 LFR 基准网络数据集 *A*

网络	节点数	边数	μ	节点平均度	重叠节点个数
A-1k	1 000	7 471	0.3	14.942	10
A-2k	2 000	14 634	0.3	14.634	20
A-3k	3 000	21 800	0.3	14.533	30
A-4k	4 000	29 648	0.3	14.824	40
A-5k	5 000	36 636	0.3	14.654	50
A-6k	6 000	43 557	0.3	14.519	60
A-7k	7 000	51 437	0.3	14.696	70
A-8k	8 000	58 242	0.3	14.561	80
A-9k	9 000	65 350	0.3	14.522	90
A-10k	10 000	73 044	0.3	14.609	100

Table 3 LFR bechmark network dataset *B*

表 3 LFR 基准网络数据集 *B*

网络	节点数	μ	平均节点度数	节点可归属社区个数
*B*1-1	5 000	0.1	9.96	2
*B*1-2	5 000	0.1	9.91	3
*B*1-3	5 000	0.1	9.97	4
*B*1-4	5 000	0.1	10.0	5
*B*1-5	5 000	0.1	9.83	6
*B*1-6	5 000	0.1	10.1	7
*B*1-7	5 000	0.1	9.88	8
*B*2-1	5 000	0.3	9.99	2
*B*2-2	5 000	0.3	9.91	3
*B*2-3	5 000	0.3	9.99	4
*B*2-4	5 000	0.3	10.1	5
*B*2-5	5 000	0.3	10.0	6
*B*2-6	5 000	0.3	9.94	7
*B*2-7	5 000	0.3	9.84	8

Table 4 SNAP real networks

表 4 SNAP 真实网络数据集

网络	节点数	边数
Ca-GrQc	5 242	14 496
Ca-HepPh	12 008	118 521
Ca-HepTh	9 877	25 998

本文所有算法利用面相对象 Java 语言编程实现,硬件平台为 2.5GHz Intel Core i5 的 CPU,内存为 16GB,运行在 Apple 的 OS X 操作系统上.使用的对比算法包括:

(1) 基于标签传播思想的 COPRA 算法[3]和 SLPA 算法[4];

(2) 基于分裂思想的 CONGA 算法[5],对比 CONGA 各种划分选取最好准确度作为最终结果.

4.2 社区检测准确性分析

本节从 4 个指标对比分析复杂网络大数据社区检测算法的准确性,具体内容如下.

(1) 标准化互信息指标 NMI(normalized mutual information)[18].

这一指标对于人工模拟的网络,即已知标准划分情况的网络能够非常好地判断算法的重叠社区检测准确度,其表达式如公式(12)所示.

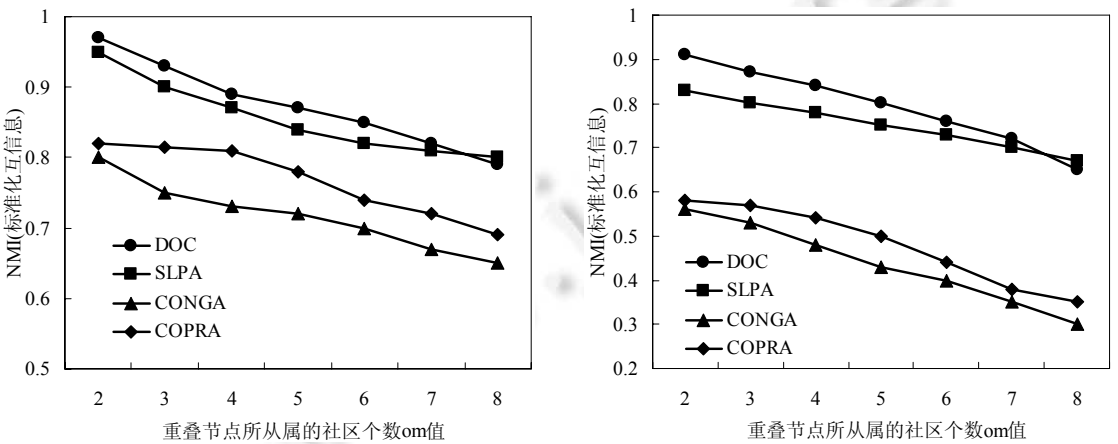
$$I(A,B)=\frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B}N_{ij}\log\left(\frac{N_{ij}\times N}{N_{i\bullet}\times N_{\bullet j}}\right)}{\sum_{i=1}^{C_A}N_{i\bullet}\log\left(\frac{N_{i\bullet}}{N}\right)+\sum_{j=1}^{C_B}N_{\bullet j}\log\left(\frac{N_{\bullet j}}{N}\right)} \tag{12}$$

其中,C_A 表示标准社区划分结果,C_B 为算法所得到的社区划分结果,矩阵 N 的行对应标准的社区检测结果,矩阵 N 的列对应算法得到的社区检测结果,第 i 行的总和记作 $N_{i\bullet}$,第 j 列的总和记作 $N_{\bullet j}$.对公式进行分析可知:

- 当算法得到的社区划分结果和标准社区划分结果一致时,NMI 指标的值等于 1;
- 当算法得到的社区结果完全和标准社区划分相反时,例如划分得到了所有节点在一个社区之中,此时,NMI 指标的值将等于 0.

图 3 的实验分别基于表 2 的 LFR 基准网络 B 中的 $B1$,$B2$ 数据集进行,其中,数据集 $B1$ 的社区混合参数值为 0.1,数据集 $B2$ 的社区混合参数值为 0.3.通过图 3 可以得到如下结论.

1) 在节点归属社区个数变化情况下,DOC 算法的 NMI 值最高能达到 0.97,在 $B1$ 数据集上(μ=0.1)产生的 *NMI* 指标平均相对 SLPA 提高了 2.1%,相对于 CONGA 算法提高了 17.9%,相对于 COPRA 算法提高了 12.2%;在 $B2$ 数据集上(μ=0.3),相对于 SLPA 算法,NMI 准确率提高了 5.2%,相对于 CONGA 算法,NMI 准确率提高了 45.0%,相对于 COPRA 算法提高了 39.5.2%.DOC 算法较好的原因在于:在初步检测社区的基础上,对每个节点进行讨论,确保每个节点能够被划分到正确社区,而其他算法未对节点单独进行隶属度分析;
2) 随着社区混合参数的增加,社区结构将不明显,这使得各算法社区检测的能力下降.同时,随着节点归属社区数目的增加,各算法检测的准确下降趋势更明显.这与现实情况相符:当网络社区结构较弱,节点归属社区数目越多,算法对社区检测的难度会增大;
3) 实验结果同样表明:DOC 算法在社区结构变化情况下,始终具有较高的社区检测准确率.



(a) LFR 基准网络 $B1$ (b) LFR 基准网络 $B2$

Fig.3 Comparison of the NMI value under LFR benchmark network datasets among different algorithms

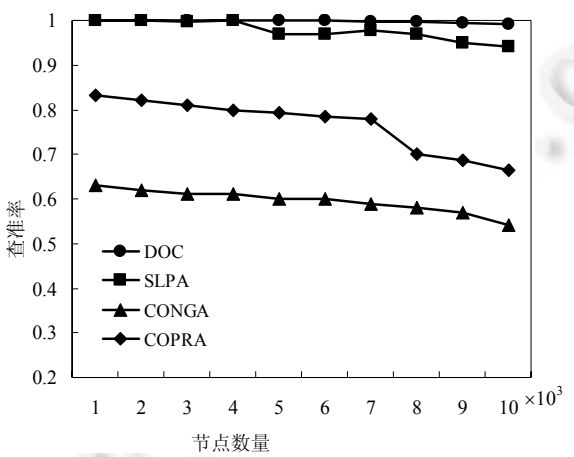
图 3 不同算法在 LFR 基准网络数据集下 NMI 值对比

(2) 查准率.

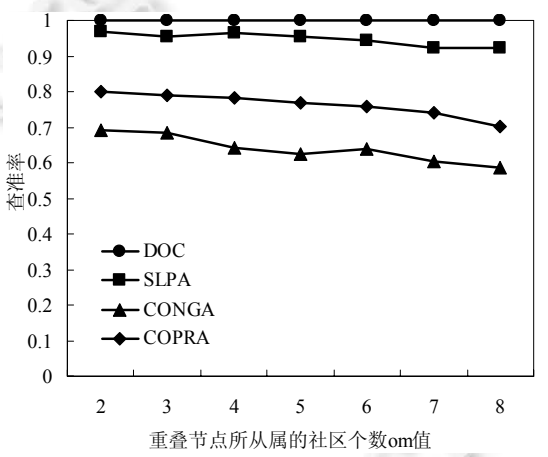
查准率表征了社区划分算法正确识别节点占算法识别出节点的比例.图 4 实验数据集 A 节点数从 1 000 增

加到 10 000,*B*1 和 *B*2 数据集是由 LFR 基准程序生成的混合参数为 0.1 和 0.3 含 5 000 个节点的数据集.

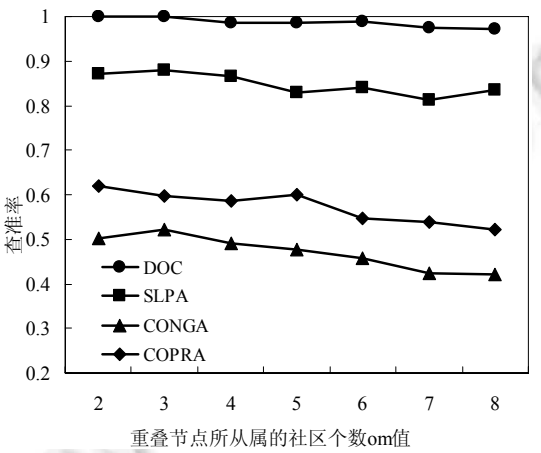
图 4 分别给出在 LFR 基准程序数据集 *A*,*B*1,*B*2 下,各种算法的查准率对比结果.查准率越高,说明算法在应用于推荐系统中时具有很好的预测用户喜好的功能.实验结果表明:在社区节点数目不断变化(数据集 *A*)、社区结构明显程度不同及重叠节点所属社区数目变化(数据集 *B*)的情况下,均能够取得较高的查准率.其中:DOC 算法在数据集 *A* 上相对于 SLPA 算法的查准率提高了 2.1%,相对于 CONGA 算法提高了 40.4%,相对于 COPRA 算法提高了 23.1%;在数据集 *B*1 上,相对于 SLPA 算法,查准率提高了 5.2%,相对于 CONGA 算法提高了 36.1%,相对于 COPRA 算法,在查准率上提高了 23.7%;在数据集 *B*2 上,相对于 SLPA 算法,在查准率上平均提高了 14.1%,相对于 CONGA 算法平均提高了 52.3%,相对于 COPRA 算法平均提高了 41.9%.原因在于:DOC 算法在初步划分的社区基础上增加了对节点进行分类的操作,使得对节点的划分更加精准.



(a) LFR 基准网络 *A*



(b) LFR 基准网络 *B*1



(c) LFR 基准网络 *B*2

Fig.4 Comparison of precision under LFR benchmark network datasets among different algorithms

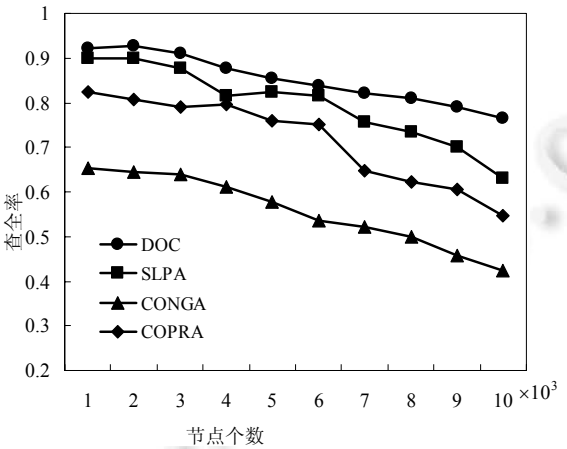
图 4 不同算法在 LFR 基准网络数据集下查准率指标对比

(3) 查全率.

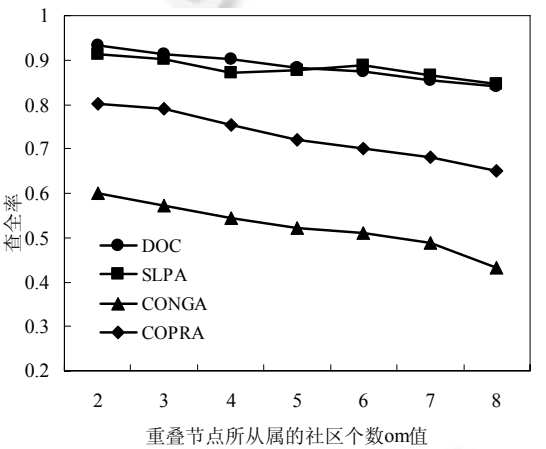
查全率表征了算法给出社区划分后正确识别的节点占所有节点的比率.

图 5 表示了在 LFR 基准程序生成的 *A*,*B*1,*B*2 数据集下,不同算法的查全率对比结果.实验结果表明:在社区节点数目不断变化(数据集 *A*)、社区结构明显程度不同及重叠节点所属社区数目变化(数据集 *B*)下,DOC 算法

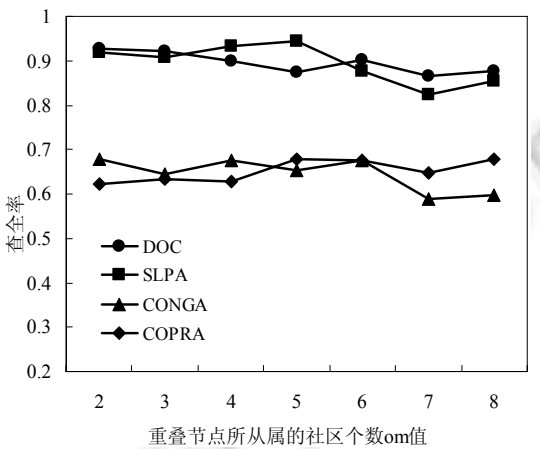
均具有较高的查准率.其中:在数据集 *A* 上,相对于 SLPA 算法平均提高了 6.6%,相对于 CONGA 算法平均提高了 34.6%,相对于 COPRA 算法,在查全率上平均提高了 16.0%;在数据集 *B*1 上,相对于 SLPA 算法平均提高了 0.6%,相对于 CONGA 算法平均提高了 40.9%,相对于 COPRA 算法平均提高了 17.8%;在数据集 *B*2 上,相对于 SLPA 算法平均提高了 0.2%,相对于 CONGA 算法平均提高了 28.0%,相对于 COPRA 算法平均提高了 27.2%.原因在于:DOC 算法是采用了优化后的 Fast-Newman 算法进行了初步的社区检测,使得算法针对不同网络的考虑更加充分,社区识别更加全面.相对于 SLPA 以及 COPRA 算法利用标签传播思想和 CONGA 算法利用节点分裂思想的算法具有更高的查全率.



(a) LFR 基准网络 *A*



(b) LFR 基准网络 *B*1



(c) LFR 基准网络 *B*2

Fig.5 Comparison of recall under LFR benchmark network datasets among different algorithms

图 5 不同算法在 LFR 基准网络数据集下查全率指标对比

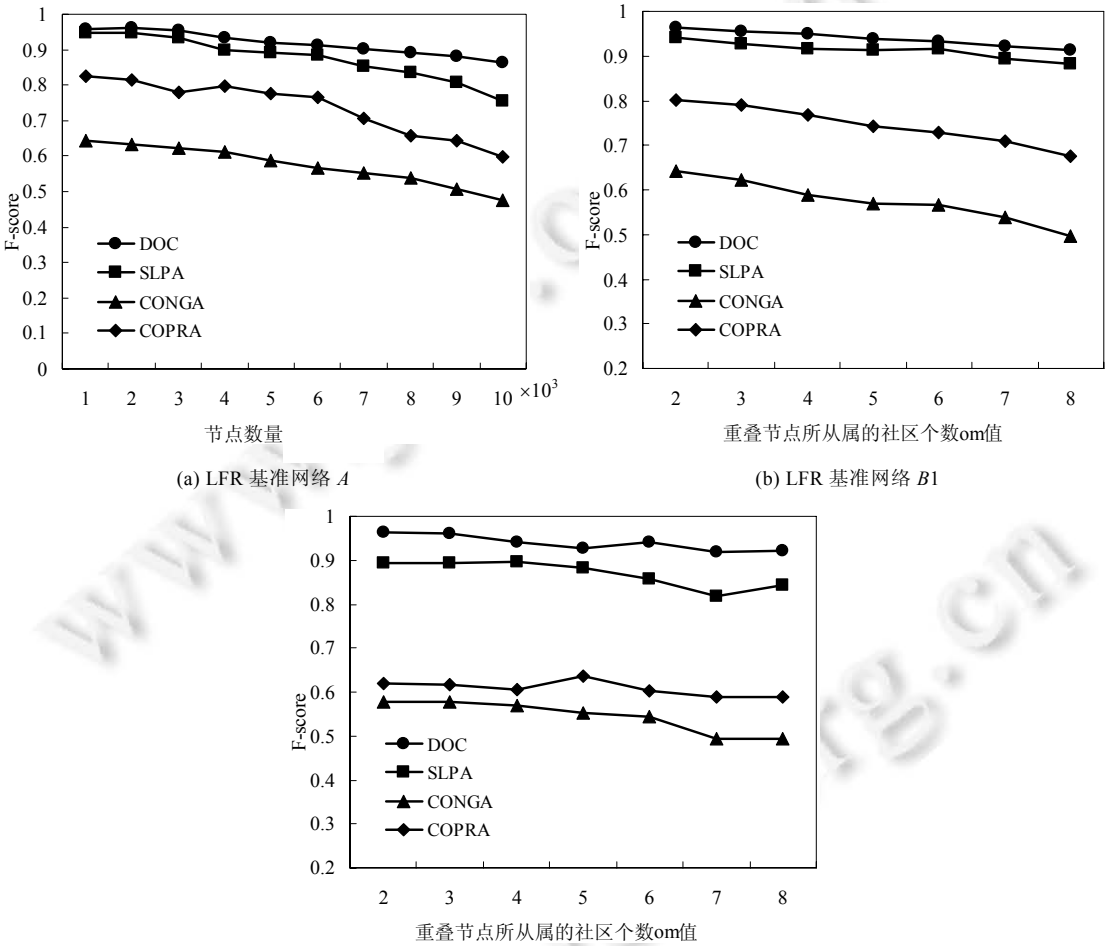
(4) 综合指标 F-score.

F-score 能够描述重叠节点检测的综合准确性,定义如公式(13)所示.

$$F\text{-}score=\frac{2\times recall\times precision}{precision+recall} \tag{13}$$

图 6 所示实验数据集基于 LFR 基准程序生成的 *A*,*B*1,*B*2 数据集生成.DOC 算法的 F-score 值在不同数据集上的平均值均高于 0.91.在数据集 *A* 上,相对于 SLPA 算法,F-score 平均提高了 4.7%,相对于 CONGA 算法平均

提高了 37.5%,相对于 COPRA 算法平均提高了 19.7%.在数据集 *B*1 上,相对于 SLPA 算法平均提高了 2.8%,相对于 CONGA 算法平均提高了 38.7%,相对于 COPRA 算法平均提高了 20.7%.在数据集 *B*2 上,相对于 SLPA 算法平均提高了 7.4%,相对于 CONGA 算法平均提高了 42.1%,相对于 COPRA 算法平均提高了 35.2%.DOC 算法的优势得益于应用了新的节点和边的更新策略,提高了算法的查全率.此外,本文针对初步划分的节点重新进行分类处理,使得最终的查准率大大提高.



(a) LFR 基准网络 *A*

(b) LFR 基准网络 *B*1

(c) LFR 基准网络 *B*2

Fig.6 Comparison of F-score vlaue under LFR benchmark network datasets among different algorithms

图 6 不同算法在 LFR 基准网络数据集下 F-score 值对比

4.3 社区识别质量分析

网络社区结构和搜索性能与聚类系数密切相关,本文使用其衡量算法在复杂网络大数据中社区识别质量.

定义 7(聚集系数). 用于描述网络的聚集度,计算公式如下.

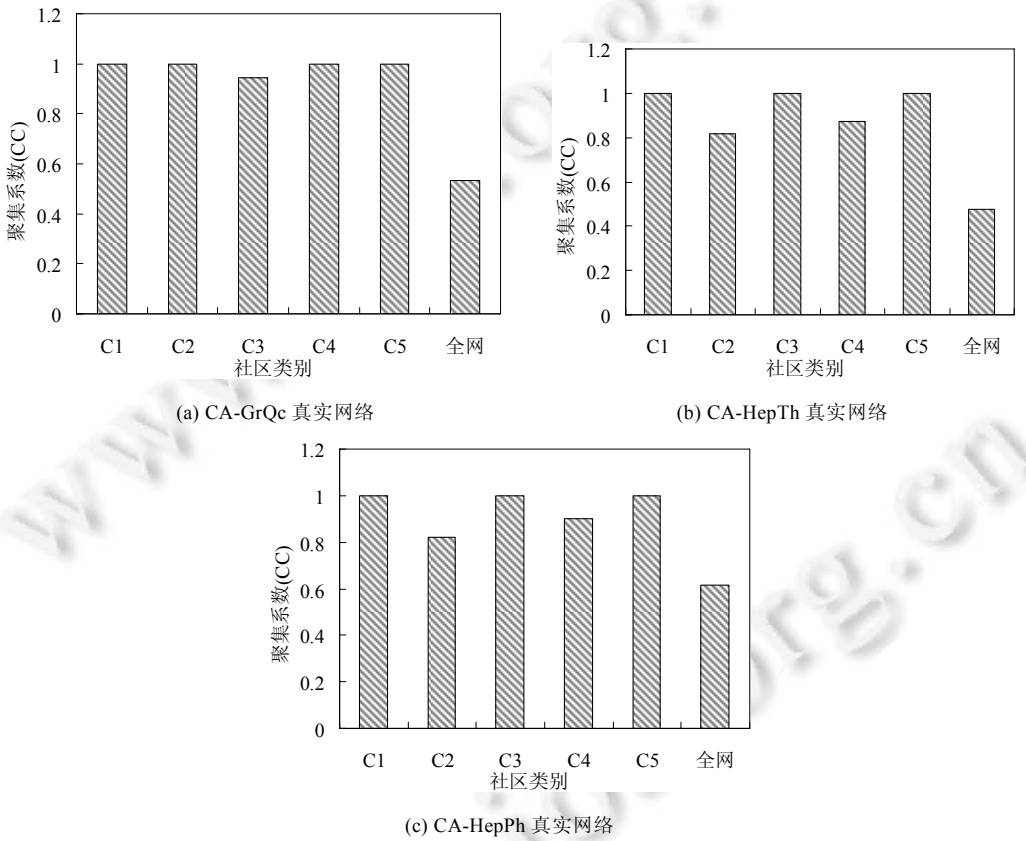
$$C_i = \frac{2|e_{jk}|}{k_i \times (k_i - 1)} \tag{14}$$

其中,e_{jk} 表示节点 i 的相邻节点 j 和节点 k 之间的连接边,k_i 表示节点 i 的度.聚集系数 $C(i)\in[0,1]$,当 $C(i)=1$ 时,表示该社区是一个完全图.整个网络的聚集系数(全网 *CC*)是所有节点聚集系数的平均值,即:

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

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.

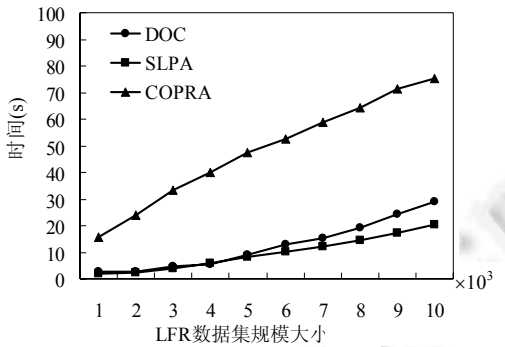


Fig.8 Runtime comparison of different algorithms under LFR benchmark network dataset *A*

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.