

一种基于适度贪心算法思路的社团划分改进方法

武 澎^{1,2}, 王恒山¹

(1 上海理工大学 管理学院, 上海 200093; 2 河南大学 软件学院, 河南 开封 475001)

摘 要: 当前问题背景下贪心算法存在不足, 由此提出一种适度贪心的算法思路, 引入适度原则, 约束贪婪幅度, 以避免过度贪婪、谬误累积的情况, 并将适度贪心算法思路应用于 FN 社团划分算法的改进研究, 构建了一种新的网络社区划算法. 通过算例对提出的算法进行了测试与分析, 实验结果验证了新算法的可行性与有效性.

关键词: 社团结构; FN 算法; 贪心算法

中图分类号: TP39

文献标识码: A

文章编号: 1000-7180(2014)07-0101-04

Community Division Method in Networks Based on Moderately Greedy Algorithm

WU Peng^{1,2}, WANG Heng-shan¹

(1 School of Management, University of Shanghai for Science and Technology, Shanghai 200093, China;

2 College of Software, Henan University, Kaifeng 475001, China)

Abstract: This paper, based on the deficiency of greedy algorithm, tries to propose a new algorithmic thinking from the perspective of the moderate principle. The moderate principle can avoid the occurrence of excessive greed situation. Based on the lack of FN algorithm, this paper proposes a new algorithm to divide community structure from the perspective of the moderately greedy algorithm. Finally, the new algorithm is illustrated through examples.

Key words: community structure; FN algorithm; greedy algorithm

1 引言

随着人们对网络性质的数学意义以及物理特性的深入分析与探究, 许多专家学者发现现实中的很多网络都具有一个共同的特征: 社团结构^[1-2]. 网络社团的划分研究就是要合理地划分出网络中真实存在的社团结构, 这对于分析网络的拓扑结构、理解网络的功能、发现网络中的隐藏规律和预测网络行为等都有着十分重要的意义与价值^[3].

本文根据当前问题背景下贪心算法的不足, 提出一种适度贪心算法思路, 引入适度原则来对贪婪幅度进行约束, 以避免过度贪婪、谬误累积的情况发生, 并准备将该算法思路运用于 FN 社团划分算法的改进研究, 进而构建一种基于适度贪心算法思路

的社团区划算法, 并通过算例对其进行验证与分析.

2 算法分析及改进

Newman 等针对社团的强制区划问题, 引入模块度^[4]来评价社团区划的合理性. 即假定网络 $G = (V, E)$ 可被分解为 k 个社团 $G_1, \dots, G_k$, 定义一个 $k \times k$ 的对称矩阵 $e = (e_{ij})_{k \times k}$, 其中 e_{ij} 表示网络中连接 G_i 和 G_j 两个社团的连边在所有连边中所占的比例, 这两个节点分别位于第 i 个社团及第 j 个社团. 对角元之和 $Tr e = \sum_{i=1}^k e_{ii}$ 表示连接社团内部节点的边所占的比例, 行(列)元素之和 $a_i = \sum_{j=1}^k e_{ij}$ 表示连接社团 G_i 中节点的边所占的比例, 在此基础上, 可以得

收稿日期: 2013-09-30; **修回日期:** 2013-11-09

基金项目: 国家自然科学基金资助项目(71071098); 上海市一流学科建设项目(S12016LXK)

到网络社团划分的模块度定义: $Q = \sum_{i=1}^k (e_{ii} - a_i^2) = \text{Tr}e - \|e^2\|$, 其中 $\|x\|$ 表示矩阵 x 中所有的元素之和^[2-8].

FN 算法, 即 Newman 快速算法, 该方法其实是一种基于贪婪算法思想的凝聚算法, 具体步骤如下所示.

步骤 1: 初始化网络为 n 个社团, 即每个节点就是一个独立社团. 初始的 e_{ij} 和 a_i 满足下面的条件

$$e_{ij} = \begin{cases} 1/2m, & \text{如果节点 } i \text{ 和 } j \text{ 之间有边相连} \\ 0, & \text{其他} \end{cases}$$

$$a_i = k_i/2m$$

式中, k_i 为节点 i 的度; m 为网络中总的边数.

步骤 2: 依次合并有边相连接的社团对, 并计算合并后的模块度增量 $\Delta Q = e_{ij} + e_{ji} - 2a_i a_j = 2(e_{ij} - a_i a_j)$. 按照贪心算法原理, 每次合并应该沿着使 Q 增大最多或者减少最小的方向进行. 每次合并以后, 对相应的元素 e_{ij} 更新, 并将和 i, j 社团相关的行与列相加.

步骤 3: 重复执行步骤 2, 不断合并社团, 直到整个网络都合成一个社团, 该方法执行结束后可以得到一个社团结构分解的树状图^[2,3,5-9].

FN 算法 GN 算法相比, 算法的复杂度大大降低, 可以用于分析节点数目高达百万的复杂网络, 然而算法速度的加快却带来了算法精度下降的问题, 一个较为典型的例子就是 FN 算法对 karate club 网络(即 Zachary 空手道俱乐部网络)的划分, 划分速度比 GN 算法有较大改善, 但划分结果却出现了新的划错节点 10. karate club 网络^[10]是复杂网络与社会网络分析领域中经常用到的一个非常经典的测试网络, 该网络中的每个节点代表一个俱乐部的成员, 节点间的连边表示两个成员之间的朋友关系. 该网络作为一个真实的社会关系网络, 常常被用于测试网络社团划分方法的准确性与有效性^[2-6,8,9], karate club 网络的网络拓扑结构图如图 1 所示.

FN 算法对 karate club 网络的划分结果如图 2 所示. 图 2 中, Newman 根据节点应归属的社团不同而对其标识以不同的形状, 我们可以清楚地看到其中节点 10 被错误地划分在左边社团之中, 而这和 karate club 网络的实际分裂情况不符. 若能保持 FN 算法精髓的同时, 弥补其现有不足, 对其算法精度有所改进则算法的整体性能势必会更加令人满意. FN 算法中的合并社团选取策略是其算法的关键所在, 若从此入手应可以对其算法进行有效地改

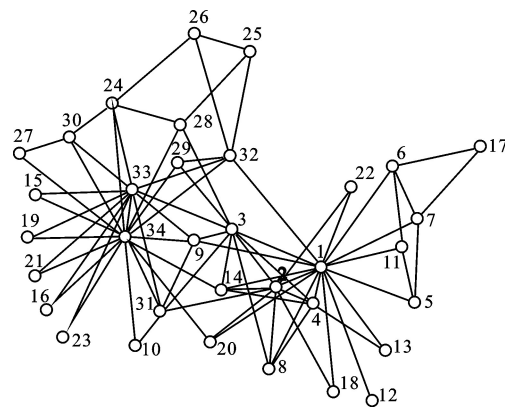


图 1 karate club 网络结构图

进, FN 算法中的合并社团选取策略主要是基于贪心算法的思想.

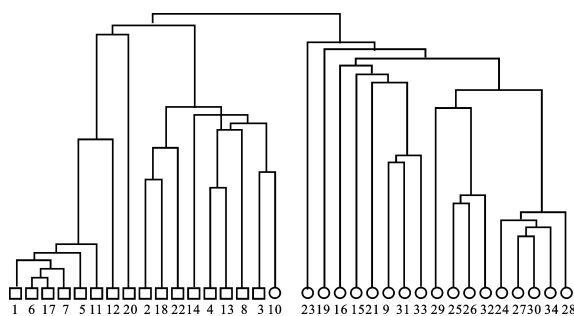


图 2 FN 算法对 karate club 网络的划分^[9]

贪心算法(greedy algorithm), 也称为贪婪算法, 主要是指在对问题求解时, 总是做出在当前看来是最好的选择. 也就是说, 不从整体最优上加以考虑, 所做出的仅是从某种意义上的局部最优解. 贪婪算法常以当前情况为基础作最优选择, 而不考虑各种可能的整体情况. 贪婪算法采用逐步构造最优解的方法, 一般贪婪算法将构造可行解的工作分阶段来完成, 在每个阶段, 选择那些在一定的标准下是局部最优的方案, 期望各阶段的局部最优的选择带来整体最优. 它省去了为找最优解要穷尽所有可能而必须耗费的大量时间, 它采用自顶向下, 以迭代的方法做出相继的贪婪选择^[10-15].

贪心算法其实是一种算法设计思想, 但在当前问题背景下也存在一些不足, 尤其是当上一次迭代结果直接影响下一次迭代结果且出现谬误时, 若采用贪心算法思想进行计算则会造成最终结果的较大误差. 对此, 本文尝试着提出一种适度贪心算法思路来进行补救与归正. 适度贪心算法思路, 即在对问题求解时, 总是做出在当前看来是较好的选择, 和贪婪算法思想一样, 适度贪心算法也采用逐步构造最优解的方法, 在每个阶段, 选择那些在一定的标准下是局部较优的方案, 期望各阶段的局部较优的选择

带来整体最优.适度贪心算法其实就是在贪心算法思想的基础上引入适度原则,约束贪婪幅度,避免过度贪婪、谬误累积,造成最终结果误差较大的情况,其中若适度原则选取合适则可以实现最终结果的理想状态.

现将适度贪心算法思路应用于 FN 算法的改进研究,经过分析可以看出, FN 算法步骤二中“每次合并应该沿着使 Q 增大最多或者减少最小的方向进行”是算法的关键.原算法是每次选取 $\max(\text{delta}Q)$ 对应的社团对来进行合并,出现了过度选取、错误划分的情况,若是能在算法理念大方向不变的情况下,对其每次的合并社团选取策略进行改进则有可能实现算法精度上的进步,本文起初是每次将的 $\text{delta}Q$ 矩阵的各列按照其 $\text{delta}Q$ 值的大小进行排序,然后引入一个算法调节因子 factor 来充当适度贪心算法中的适度原则,将其和 $\max(\text{delta}Q)$ 的乘积对应的社团对进行合并,以实现约束贪婪幅度的约束,进而期望避免过度贪婪、错误划分的情况.经过试验,对于 Zachary 空手道网络自选调节因子 factor 取值为 0.9 左右时算法的社团划分效果较好,可以有效的排除歧义节点的干扰,实现正确的划分归类.而对于其他网络,则 factor 取 0.9 的效果则未必较好,若想得到理想的划分效果,则需重新选取的合适调节因子取值.这样一来,虽然当前策略改进了原 FN 算法的精度,但又引入了一个自选调节因子 factor ,算法精度提升的同时,算法适应性则降低了.如何解决调节因子的合适取值则成为当前算法改进的主要问题.经过分析可以得出, FN 算法在执行过程中,合并社团是两两合并,而且合并的两个社团地位并不完全等同,一个社团为主出发社团,是合并动作的起始社团,另一个社团为被选取社团,在此我们将社团合并时的主出发社团定义为合并起始社团,合并时的另一个社团定义为合并被选社团.后经研究发现,合并前起始社团的节点个数与合并后社团的节点个数之比可以作为一个很好的自适应调节因子,来充当适度贪心算法中的适度原则,对每次迭代中的贪婪幅度进行约束,以规避过度贪婪、谬误累积、错误划分的情况,进而实现较为理想的最终结果.本文提出的基于适度贪心算法思路的社团划分改进方法的部分实现代码如下所示:

```
deltaQ=2*(e(x,y)-a(x)*a(y));
alll=find(clusters(2,:)==x);
arrr=find(clusters(2,:)==y);
att=length(alll);
```

```
ayyy=length(alll)+length(arrr);
auuu=att/ayyy;
aiii=[aiii[x,y;deltaQ*auuu]];
[max deltaQ,id]=max(aiii(3,:));
id=id(1);
x=aiii(1,id);
y=aiii(2,id);
```

3 验证与分析

为了验证新算法的性能,我们利用人工生成网络和真实世界网络对新算法进行测试与分析.算法实验环境为:处理器 Intel(R) Pentium(R) D CPU 3.20 GHz,内存 1 GB,硬盘 160 GB,操作系统为 Microsoft Windows XP,编程语言为 Matlab 7.0.

3.1 人工网络算例

本文根据需要生成一个具有明显社团结构的人工网络(简称三团网络),其网络拓扑结构如图 3 所示.

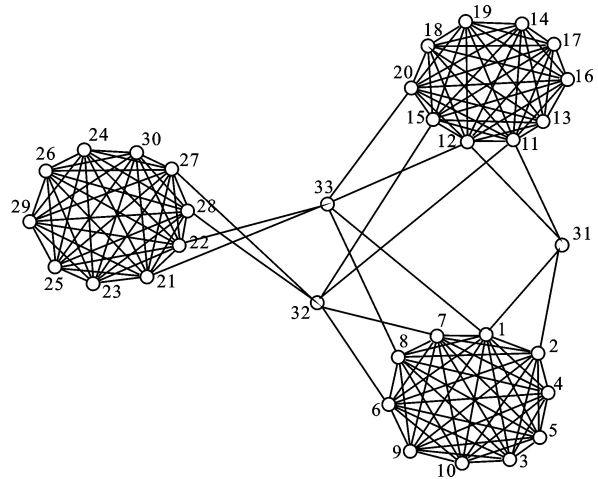


图 3 三团网络结构图

本文提出的改进新算法对三团网络的划分结果如图 4 所示.

由图 4 中的划分结果可以看出本文所提出的新算法对网络社团的划分效果较为理想,划分结果验证了新算法的可行性与有效性.

3.2 真实网络算例

karate club 网络是当前网络社团划分公认的测试网络,而且原 FN 算法对该网络划分存在错误.本文提出的改进新算法对 karate club 网络的划分结果如图 5 所示.

参照图 2 与图 5,可以看出本文提出的改进新算法有效地弥补了原 FN 算法的不足,成功地将错误划分点 10 进行了正确归类,新算法的划分结果与

karate club 网络的真实情况完全一致,再次验证了本文所提改进新算法的可行性与有效性.

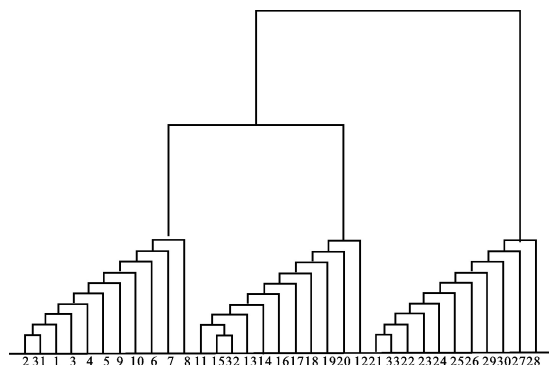


图4 本文新算法对三团网络的划分

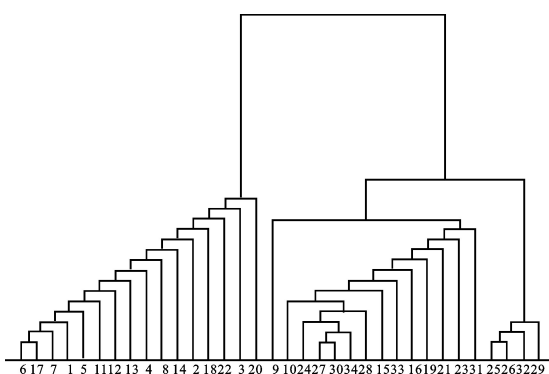


图5 本文新算法对 karate club 网络的划分

4 结束语

本文提出一种适度贪心算法思路,即引入适度原则,对每次迭代中的贪婪幅度进行约束,以避免过度贪婪、谬误累积、最终结果误差较大的情况,并在此基础上,构建了一种基于适度贪心算法思路的社团划分算法,并通过算例对新算法进行了验证与分析.

参考文献:

- [1] 汪小帆,李翔,陈关荣. 复杂网络理论及其应用[M]. 北京:清华大学出版社,2006.
- [2] 解肖,汪小帆. 复杂网络中的社团结构分析算法研究

综述[J]. 复杂系统与复杂性科学,2005,2(3):1-12.

- [3] 金弟,杨博,刘杰,等. 复杂网络簇结构探测——基于随机游走的蚁群算法[J]. 软件学报,2012,23(3):451-464.
- [4] Newman M E J, Girvan M. Finding and evaluating community structure in networks [J]. Physical Review E, 2004(69):026113.
- [5] 解肖. 复杂网络的社团结构建模与分析[D]. 上海:上海交通大学,2007.
- [6] 莫春玲. 复杂网络中聚类方法及社团结构的研究[D]. 武汉:武汉理工大学,2007.
- [7] 王波. 基于派系的复杂网络及其在公交网络上的应用研究[D]. 杭州:浙江工业大学,2009.
- [8] 邓智龙,淦文燕. 复杂网络中的社团结构发现方法[J]. 计算机科学,2012,39(6):103-108.
- [9] Newman M E J. Fast algorithm for detecting community structure in networks [J]. Physical Review E, 2004(69):066133.
- [10] Zachary W. An information flow model for conflict and fission in small groups [J]. Journal of Anthropological Research, 1977,33(4):452-473.
- [11] 王铮,吴静. 计算地理学[M]. 北京:科学出版社,2011.
- [12] 百度百科. 贪心算法[EB/OL]. [2013-04-20]. <http://baike.baidu.com/view/298415.htm?fromId=1628576>.
- [13] 付俐. 马钢铁矿石采购物流成本核算与优化控制研究[D]. 马鞍山:安徽工业大学,2012.
- [14] 汪晓银. 数学建模与数学实验[M]. 北京:科学出版社,2010.
- [15] 李博权,李绪志,王红飞,等. 贪婪算法与动态规划结合的任务规划方法[J]. 微电子学与计算机,2013,30(2):144-147.

作者简介:

武 澎 男,(1981-),博士研究生,讲师. 研究方向为社团划分、超网络、网络分析等.

王恒山 男,(1948-),教授,博士生导师. 研究方向为管理信息系统、复杂网络等.

(上接第 100 页)

- [5] Ferrara A, Green M, Hohenberger S. et al. Practical short signature batch verification [C]//Proceedings of Topics in cryptology-CT-RSA 2009. Berlin: Springer Heidelberg, 2009:309-324.
- [6] Standards for Efficient Cryptography. SEC1. Elliptic curve cryptography [S]. Mississauga: Certicom Corp,2000.

作者简介:

杜大海 男,(1982-),博士,工程师. 研究方向为信息安全、物联网安全.

范 红 女,(1969-),博士,副研究员. 研究方向为信息安全.

王 冠 女,(1985-),硕士,助理工程师. 研究方向为信息安全.

李程远 男,(1984-),硕士,工程师. 研究方向为信息安全.