

一种基于聚类的社团划分算法

王 伟,李玲娟

(南京邮电大学 计算机学院,江苏 南京 210003)

摘 要: 社团划分是社会网络的一个研究热点。为了快速准确地发现社会网络的社团结构,文中从节点的重要度出发,利用节点之间的相似性,提出了一种基于聚类的社团划分算法—CCDA。其基本思想是每次以节点集合中聚集系数最大的点作为聚类中心,基于最短路径和欧几里得距离计算节点相似度,选择与聚类中心的相似度大于给定阈值的点进行聚类,不断迭代,直至节点集合为空,所产生的各个簇即为不同的社团。对被重复划分的节点,以模块度函数为标准,将节点归属到最合适的社团中。由于该算法每次从重要节点出发,再次选取聚类中心时不需考虑已经被聚类的节点,所以时间复杂度低于 GN 算法和 Newman 算法。将该算法应用于经典的社会网络 Zachary,结果表明了 CCDA 算法对社团划分的有效性。

关键词: 社会网络;社团划分;聚集系数;相似性;聚类

中图分类号: TP311

文献标识码: A

文章编号: 1673-629X(2015)10-0119-04

doi: 10.3969/j.issn.1673-629X.2015.10.026

A Clustering-based Community Division Algorithm

WANG Wei, LI Ling-juan

(College of Computer, Nanjing University of Posts and Telecommunications,
Nanjing 210003, China)

Abstract: Community division has been a research focus in the social network area. In order to quickly and accurately find community structure in the social network, from the importance of nodes and consulting their similarities, propose a clustering-based community division algorithm CCDA. The basic idea of this algorithm is selecting the node owning greater clustering coefficient as the clustering center, calculating similarity by the shortest path and Euclidean distance, putting the node with similarity greater than given threshold to cluster, and iterating the process until the node collection is empty. For the repeated division nodes, the algorithm divides each of them into the most appropriate community by using the module function Q . The clusters generated by the algorithm are corresponding with the communities. Since the algorithm starts from the important node and does not consider those clustered nodes when determining new clustering center, the time complexity of it is lower than GN algorithm and Newman algorithm. The results of applying the algorithm to the classical social network, the Zachary network, show that CCDA is valid in community division.

Key words: social network; community division; clustering coefficient; similarity; clustering

0 引言

在网络理论研究中,复杂网络^[1]是由大量节点和节点之间错综复杂的关系共同组成的网络拓扑结构图。复杂网络往往具有一些特殊的性质,如小世界性效应、无标度特性^[2]。在现实生活中,复杂网络可以描述社会关系网络,如人与人之间的交际关系、论文合作者之间的合作关系、物种捕食之间的食物链关系等。随着对网络性质的物理意义和数学特性的深入研究,

人们发现许多复杂网络都有着一个共同的性质—社团结构。揭示网络中的社团结构,对于了解网络结构与分析网络特性有着十分重要的意义。

对网络中社团的常见定义是基于相对连接密度的描述:网络中的节点可以分为组,组内连接稠密而组间稀疏^[3-4]。即社团内的节点连接紧密而社团之间的节点连接不紧密。

目前,国内外对社团划分展开了大量研究,主要的

收稿日期:2014-06-18

修回日期:2014-10-15

网络出版时间:2015-08-26

基金项目:国家“973”重点基础研究发展计划项目(2011CB302903)

作者简介:王 伟(1989-),男,硕士研究生,研究方向为数据挖掘、大数据、云计算;李玲娟,教授,通讯作者,研究方向为数据挖掘、信息安全、分布式计算。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20150826.1535.004.html>

方法集中在两类:基于优化的算法和基于启发式的算法。

(1) 基于优化的算法:该类算法通过最优化预定义的目标函数来计算复杂网络的社团结构,基于优化 Q 函数的复杂网络总是倾向找到粗糙的而不是精细的网络社团结构。主要代表算法有谱平分法^[5]和 Kernighan-Lin^[6]算法等。

(2) 启发式算法:该类算法将复杂网络社团划分问题转化为预定义的启发式规则的设计问题。其特点是:基于某些直观的假设来设计启发式算法,对于大部分的社会网络,能快速地找到最优解或者近似最优解。主要算法有 GN (Girvan-Newman) 算法^[7]和 Newman 贪婪算法^[8-9]等。

这些社团结构划分算法都有各自的优缺点,比如谱平分法,优点是具有严密的数学理论,且划分的速度很快,但缺点也很明显,对于那些社团结构不是很明显的社会网络,划分存在较大误差。Kernighan-Lin 算法采用的是一种贪婪算法,需要事先知道社团的数目,如果事先不知道,划分结果就不准确。GN 算法根据边介数自顶向下对社团结构进行分裂,且每分裂一次都得重新计算边介数,时间开销比较大。

基于启发式算法的思想,文中提出一种基于聚类的社团划分算法(Clustering based Community Division Algorithm, CCDA)。该算法通过选取聚集系数大的节点作为聚类中心,通过比较节点之间的相似性,进行节点的聚类,然后再选取剩余节点中聚集系数最大的节点作为聚类中心进行聚类,重复此过程,直到所有节点都被划分为止。最后即可得到网络的社团结构。

1 相关定义

为了更加方便地探索复杂网络中社团的结构,研究者们给出了一些量化的社团定义,例如节点连接度、强社团、聚集系数、模块度函数的定义等。相似性是文中提出的社团划分算法中聚类的重要参考依据,所以文中也给出了相似度的定义。

(1) 节点连接度:被定义为与该节点有边相连的其他节点的数目,也称节点的度。

(2) 强社团:满足下面条件的社团 ζ 为强社团。

$$k_i^{\text{in}}(\zeta) > k_i^{\text{out}}(\zeta) \quad \forall i \in \zeta$$

式中, $k_i^{\text{in}}(\zeta)$ 表示节点 i 与 ζ 内部节点连接的度; $k_i^{\text{out}}(\zeta)$ 表示节点 i 和 ζ 外部节点连接的度。即在强社团 ζ 中,任一节点与 ζ 的内部节点连接的度一定比其与 ζ 外部节点的连接的度要大。

(3) 节点聚集系数:节点的聚集系数 C_i 定义为:

$$C_i = E_i / T_i, C_i \in [0, 1] \quad (1)$$

其中,设节点 i 的度是 k_i , E_i 表示节点 i 的 k 个邻

居节点之间实际的连接边数; T_i 表示节点 i 的 k 个邻居节点可能形成的最大连接数, $T_i = k * (k - 1) / 2$ 。当 $C_i = 1$ 时,表示节点 i 的所有邻居节点都相连,也表示节点 i 处于中心地位。

(4) 模块度函数 Q ^[10-11]: Q 函数是现今最广泛使用的衡量社团结构划分好坏的指标,下面给出一种最常见的模块度函数 Q 的定义:

$$Q = \sum_{s=1}^m \left[\frac{l_s^{\text{in}}}{L} - \left(\frac{d_s}{2L} \right)^2 \right] \quad (2)$$

其中, l_s^{in} 表示社团内连接边的数量; d_s 是社团内所包含节点的度; L 是总的连边数。一般来说, Q 函数值越大,所划分出的社团结构越好。

(5) 网络节点的相似度。

文中提出的 CCDA 算法中,节点相似性的度量依据的是越相似的两个节点到网络结构中其他节点的最短路径应该越接近这一思想。另外,文中用欧氏距离定量地计算节点的相似度。

a. 假设 G 是一个具有 M 个节点的无向图,设 $G = \{x_1, x_2, \dots, x_m\}$, 另外定义 x_{ik}, x_{jk} 分别表示节点 i, j 到节点 k 的最短路径,则节点 i, j 的欧氏距离为:

$$D(x_i, x_j) = \sqrt{\sum_{k=1}^M (x_{ik} - x_{jk})^2} \quad (3)$$

b. 由于欧氏距离和相似度正好成反比,欧氏距离越小,代表两个节点的相似度越大,另外为了保证相似度的取值范围为 $(0, 1]$, 所以定义相似度为:

$$S(x_i, x_j) = \frac{1}{1 + D(x_i, x_j)} \quad (4)$$

2 算法描述与分析

在 CCDA 算法中,主要使用的是凝聚式聚类的思想,将凝聚式聚类的过程理解成节点之间自我组合的过程。最初阶段,把网络结构的每个节点都当成一个个单独的社团或者个体。每次进行迭代的时候,从与之相连的节点的集合中选择一个相似度系数最大或者“距离”最小的节点合并,不断进行迭代,直到所有节点都被划分为止。

2.1 算法描述

算法:CCDA。

输入:一个无向无权网络 $G = \langle V, E \rangle$, V 是网络中的节点集合, E 是网络中的边集合;

输出:网络的社团结构。

步骤:

Step1:初始化网络,按式(1)计算节点的聚集系数,存入一个键-值对的 map 数据结构中。

Step2:根据式(4),计算所有节点之间的相似度 $S(x_i, x_j)$, 结果存入对应的相异度矩阵中;计算所有节

点的平均相似度值 D , 作为划分阈值。

Step3: 选聚集系数最大的节点为聚类中心。

Step4: 将所有与该聚类中心的相似度大于 D 的节点加入该社团, 并在节点集中删除这些节点。

Step5: 如果节点集中节点数为 0, 跳到 Step6; 否则重复 Step3 到 Step5。

Step6: 如果所形成的社团中没有被重复划分的节点, 算法结束。否则, 按式(2)分别计算重复划分节点在对应社团的模块度函数 Q , 将该节点归于 Q 值大的社团。

CCDA 的具体流程如图 1 所示。

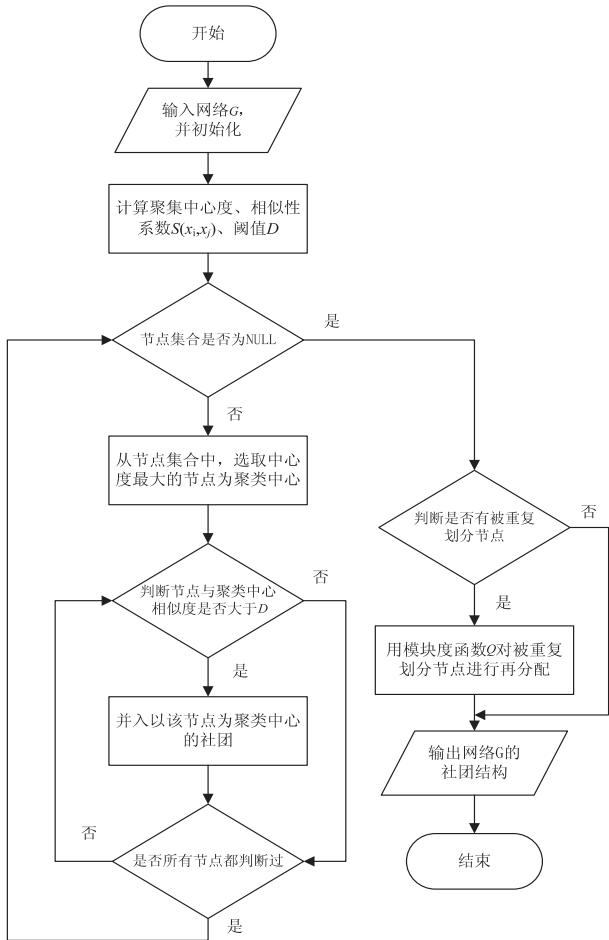


图 1 CCDA 算法的具体流程

2.2 算法分析

(1) 与 Kernighan-Lin 和谱平分法的比较。

Kernighan-Lin 算法用试探性的方法把社团划分成两个社团, 所以它只适用于已知社团结构为两个的复杂网络; 而文中算法无需知道社团个数, 且能动态划分出社团的个数。另外, 相比于谱平分法对于那些社团结构十分明显的复杂网络有着较好的划分结果, 文中算法则不要求社团结构是明显的。

(2) 时间复杂度的分析。

CCDA 算法首先计算两两节点之间的相似度, 有 n 个节点, 所以需要 $n * n$ 次循环, 这部分的时间复杂度

是 $O(n^2)$ 。然后选取聚集系数最大的节点, 并比较该节点与其他节点之间的相似度是否大于阈值, 这部分的时间复杂度是 $O(n^2)$ 。又由于每次已划入社团的节点在集合中被删除, 找中心节点时不需再比较。所以本算法的平均时间复杂度小于 GN 算法和谱平分法的时间复杂度 $O(n^3)$ 。

3 算法应用及结果分析

为了验证所提出的 CCDA 算法的可行性以及划分的准确性, 文中选择对经典的社会网络 Zachary 空手道俱乐部成员网络^[12-14]进行基于 CCDA 算法的划分。

(1) Zachary 空手道俱乐部成员网络。

Zachary 空手道俱乐部成员网络是社会学家 Zachary 观察美国一所大学的空手道俱乐部成员之间的相互社会关系给出的结果, 该俱乐部因内部分歧, 一部分成员被教练带走组成了一个新的俱乐部。图 2 是 Zachary 俱乐部划分成的两个不同的社团, 该社团结构由 34 个节点、78 条边组成, 节点代表俱乐部的成员, 边代表成员之间的关系。

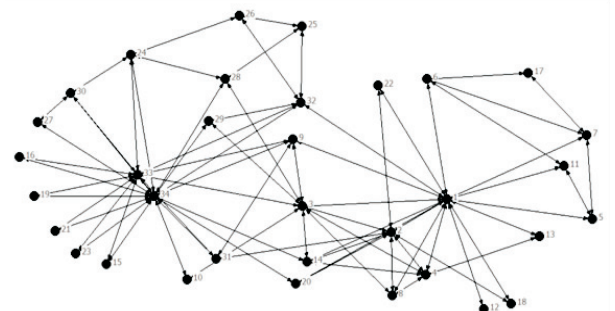


图 2 Zachary 俱乐部网络图

(2) 实验结果。

文中选用 Pentium T4200 处理器、2 G 内存以及 Windows 7 操作系统作为运行平台, 在 Myeclipse 上运行, 实验结果如表 1 所示。下划线处为社团中心。社团 1 的中心为节点 34, 社团 2 的中心为节点 1。

表 1 基于 CCDA 算法的社团划分实验结果

划分出的社团		社团的组成
中间结果	社团 1	9, 15, 16, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34
	社团 2	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 17, 18, 20, 22, 29
最终结果	社团 1	15, 16, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34
	社团 2	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 17, 18, 20, 22

从表 1 可以看出, CCDA 算法的划分结果基本与实际网络相符。只有节点 9 和节点 29 被重复划分到两个社团中。从实际网络中可以发现, 这两个节点分

别处于两个社团的边界处,属于边界节点。

对被重复划分到两个社团中的节点 9 和节点 29,再结合模块度 Q 函数的计算公式,算出节点 9 在两个社团的模块度 Q 值分别为 $Q_1=0.389$, $Q_{34}=0.366$,所以节点 9 属于社团 2。同理,节点 29 属于社团 1。可见,CCDA 算法的划分结果与实际网络结构相符。

针对同一数据集,文中还将该算法与两大经典的社会网络划分算法 GN (Girvan–Newman) 算法和 Newman 贪婪算法做了对比,结果见表 2。

表 2 CCDA 算法与其他算法的对比

算法	准确度/%	运行时间/ms
GN 算法	97.06	2 063
CCDA 算法	94.12	22
Newman 算法	73.53	35

从表 2 可以看出,文中设计的 CCDA 算法的准确度虽然不及 GN 算法,但是算法运行时间远低于 GN 算法;并且 CCDA 算法在准确度和运行时间方面都优于 Newman 算法,进一步表明了文中算法的可行性和正确性。

4 结束语


复杂网络中存在很多的社团发现算法,文中提出的基于聚类的社团划分方法从节点的重要性出发,并利用节点之间的相似度对社团进行划分,无需预设社团个数,也不要求社会网络有明显结构,能合理处理边界节点,聚类准确度较高且时间复杂度较低。在 Zachary 空手道俱乐部成员网络上的划分结果证明了该算法的综合优势。

参考文献:

[1] Dorogovtsev S N, Mendes J F F. Evolution of networks[J].
+++++
(上接第 118 页)
[19] Shirai H, Sato R, Otoi. K. Electromagnetic wave propagation estimation by 3-D SBR method [C]//Proc of international conference on electromagnetics in advanced applications. Torino;IEEE,2007:129-132.
[20] Son H W, Myung N H. A deterministic ray tube method for microcellular wave propagation prediction model [J]. IEEE Transactions on Antennas and Propagation,1999,47(8):1344-1350.
[21] Chen S H, Jeng S K. SBR image approach for radio wave propagation in tunnels with and without traffic[J]. IEEE Transactions on Vehicular Technology,1996,45(3):570-578.

Adv Phys,2002,51(4):1079-1187.
[2] Widmer J, Renda R, Mauve M. A survey on TCP-friendly congestion control[J]. IEEE Network,2001,15(3):28-37.
[3] Newman M E J. Modularity and communities structures in networks[J]. Proc of the National Academy of Science,2006,103(23):8577-8582.
[4] 刘 瑶. 社会网络特征分析与社团结构挖掘[D]. 成都:电子科技大学,2013.
[5] Fiedler M. Algebraic connectivity of graphs[J]. Czech Math J,1973,23:298-305.
[6] Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs[J]. Bell Labs Technical Journal,1970,49(2):291-307.
[7] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proc of the National Academy of Science,2002,99(12):7821-7826.
[8] 吴 鹏,李思昆. 适于社会网络结构分析与可视化的布局算法[J]. 软件学报,2011,22(10):2467-2475.
[9] Newman M E J. Fast algorithm for detecting community structure in networks [J]. Physical Review E, 2004, 69 (6): 066133.
[10] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical Review E,2004,69(2):026113.
[11] 周兰娟. 分级聚类算法在科研网络社团划分中的应用 [D]. 济南:山东师范大学,2013.
[12] 赖大荣. 复杂网络社团结构分析方法研究 [D]. 上海:上海交通大学,2011.
[13] Zachary W. An information flow model for conflict and fission in small groups [J]. Journal of Anthropological Research, 1977,33(4):452-473.
[14] 臧 丽,王 红,杨通辉. 基于改进的 ACCA 的复杂网络社团结构发现[J]. 计算机技术与发展,2012,22(10):129-132.
[22] Chen S H, Jeng S K. An SBR/image approach for radio wave propagation in indoor environments with metallic furniture [J]. IEEE Transactions on Antennas and Propagation,1997,45(1):98-106.
[23] Hammoudeh A K. Millimetric wavelengths radio wave propagation for line of sight indoor microcellular mobile communications[J]. IEEE Transactions on Vehicular Technology,1995,44(3):449-460.
[24] Kanatas A, Kountouris I D, Kostaras G B, et al. A UTD propagation model in urban microcellular environments [J]. IEEE Trans on Vehicular Technology,1997,46(1):185-193.

一种基于聚类的社团划分算法

作者: [王伟](#), [李玲娟](#), [WANG Wei](#), [LI Ling-juan](#)
作者单位: [南京邮电大学 计算机学院, 江苏 南京, 210003](#)
刊名: [计算机技术与发展](#) 
英文刊名: [Computer Technology and Development](#)
年, 卷(期): 2015(10)

引用本文格式: [王伟](#). [李玲娟](#). [WANG Wei](#). [LI Ling-juan](#) 一种基于聚类的社团划分算法[期刊论文]-[计算机技术与发](#)
[展](#) 2015(10)