

基于连边相似度的重叠社区发现算法研究^{*}

施 伟, 傅鹤岗, 张 程

(重庆大学 计算机学院, 重庆 400044)

摘 要: 针对 GN 算法在发现重叠社区时存在的不足, 以及为了降低算法时间复杂度, 提出一种基于网络图中连边相似度划分连边集的重叠社区发现算法 EGN。算法依据网络图的连边集进行划分, 每一条边被划分到某个特定的社区, 而一个节点可以关联多条连边, 因此节点可以被划分到不同的社区, 从而发现重叠社区。EGN 算法首先需要构造网络节点之间连边关系的边图; 然后根据边图中节点的关系计算网络图中连边的相似度, 在节点之间相似度的基础上提出了连边之间相似度的计算方法; 再按照相似度由小到大对边图删除边, 构建出边图的树状图。树状图的每一层对应网络的一个划分, 采用划分密度函数来衡量划分的质量, 以此寻找最优的划分。最后将算法应用到 Zachary 空手道俱乐部网络中, 并与 GN 算法进行对比, 实验结果表明 EGN 算法能够很好地发现重叠社区。

关键词: 社区发现; 重叠社区; 相似度; 划分密度

中图分类号: TP393; TP301.6

文献标志码: A

文章编号: 1001-3695(2013)01-0221-03

doi:10.3969/j.issn.1001-3695.2013.01.056

Overlapping communities detecting based on similarity of edge

SHI Wei, FU He-gang, ZHANG Cheng

(College of Computer Science, Chongqing University, Chongqing 400044, China)

Abstract: To void GN algorithm's problem and reduce the computational complexity, this paper proposed a new algorithm called EGN which based on the partition of edges in network. According to the partition of edges, each edge would be partitioned to an independent community, consequently nodes could be partitioned to multiple communities. It could detect the overlapping community in networks at last. Line graph would be constructed before using EGN algorithm which based on the relationship of nodes in network. And after that, the algorithm calculated the similarity of edges in network. It proposed the method of calculation of edge similarity which based on the similarity of node, and then constructed dendrogram of line graph in the whole process at last. Each layer of the dendrogram was related to the partition of networks. Through dividing the dendrogram in one layer, it detected the communities. It used a function of partition density to measure the capability of partition. At last, it applied this algorithm in the Zachary's Karate Club network, and compared with GN algorithm. The results of experiment show that the EGN algorithm is able to detect overlapping communities very well.

Key words: community detection; overlapping communities; similarity; partition density

现实世界中许多复杂的系统都可以抽象为网络, 如人际关系网^[1]、科学家合作网^[2]、流行病传播网^[3]、蛋白质相互合作网^[4]等。这些网络具有复杂的内部结构, 因此称为复杂网络。通常复杂网络中用节点来表示个体, 节点之间的连边表示个体之间的关系。对于人际关系网来说, 研究不同的人在社会上的交际圈, 可以概括出社会的组成, 对社会的发展提供强有力的参照。对于疾病传播网来说, 研究疾病如何进行传播以及网络结构本身具有的内在性质, 对于控制疾病的传播具有很重要的作用。因此复杂网络研究具有很重要的现实意义。近年来对众多实际网络的研究发现, 它们存在一个共同的特征, 称为社区结构。它是指网络由若干个群或团组成, 每个群的内部节点之间连接相对紧密, 而不同群之间节点的连接比较稀疏^[5]。社区发现近年来成为研究者所关注的焦点, 发现社区的不同结构, 对于了解网络结构和分析网络特征具有重要意义。

目前, 已经提出了众多基于节点集划分的社区发现算法。例如基于贪婪算法原理的 Kernighan-Lin 算法^[6]、基于谱思想的谱平分算法^[7]、基于分裂思想的 GN 算法^[5]、基于凝聚思想

的 Newman 快速算法^[8]等。社区发现的最终目的是将网络划分为若干个互相分离的社区。上面介绍的算法将每个节点严格地划分到某个社区, 导致不能发现重叠社区。在实际网络中, 某些节点可能同时属于多个社区, 即网络是由许多彼此重叠相互关联的社区组成^[9]。例如, 在人际关系网络中, 每个人根据不同的分类方法都会属于多个不同的社区(如学校、家庭、朋友等)。针对上述存在的问题, 本文在 GN 算法的基础上进行改进, 按照网络图中的连边集进行划分, 提出了基于连边相似度的重叠社区发现算法 EGN。

1 网络边图及连边相似度构造方法

1.1 基于关联矩阵构造边图

通常网络图 G 采用节点的邻接矩阵表示, 而邻接矩阵只能表示图 G 中节点与节点的关系, 不能表示连边的相关信息。因此本文采用关联矩阵表示 G 。

定义 1 网络图 G 的关联矩阵 B 为 $N \times L$ 矩阵, 其中 N 表

收稿日期: 2012-05-26; 修回日期: 2012-07-02 基金项目: 国家科技支撑计划资助项目(2011BAH25B01)

作者简介: 施伟(1987-), 男, 河南新乡人, 硕士研究生, 主要研究方向为数据挖掘、社区发现(weishi2007@163.com); 傅鹤岗(1950-), 男, 副教授, 主要研究方向为电子商务、软件工程; 张程(1977-), 男, 讲师, 博士, 主要研究方向为移动智能。

示 G 中的节点个数, L 表示 G 中的连边数目。如果节点 i 与连边 α 相关联, 则 $B_{i\alpha} = 1$; 否则 $B_{i\alpha} = 0$ 。

关联矩阵表示图 G 中节点与连边的关系, 因此通过图 G 的关联矩阵可以推导出每个节点度 k_i 和每条连边所关联的节点数目 k_α , 即

$$k_i = \sum_{\alpha} B_{i\alpha}, k_{\alpha} = \sum_i B_{i\alpha} \quad (1)$$

从关联矩阵中可以推导出节点与节点、连边与连边的关系。设网络图 G 的邻接矩阵为 A , A 是一个 $N \times N$ 的对称矩阵。文献[10]提供了关联矩阵到邻接矩阵的转换式:

$$A_{ij} = \sum_{\alpha} B_{i\alpha} B_{j\alpha} - k_i \delta_{ij} \quad (2)$$

其中, 当 $i=j$ 时, $\delta_{ij} = 1$, 否则 $\delta_{ij} = 0$ 。

定义 2 设 G 是一个图, $L(G)$ 是另一个图, 满足 $V(L(G)) = E(G)$, $L(G)$ 中两节点相邻当且仅当它们是 G 中的两条相邻的边, 则称 $L(G)$ 是 G 的边图。

网络图的边图采用邻接矩阵 C 表示, C 是一个 $L \times L$ 的对称矩阵。由上面提出的关联矩阵和邻接矩阵的转换式, 推导出由关联矩阵到边图的邻接矩阵的转换式:

$$C_{\alpha\beta} = \sum_i B_{i\alpha} B_{i\beta} (1 - \delta_{\alpha\beta}) \quad (3)$$

其中, 当 $\alpha = \beta$ 时, $\delta_{\alpha\beta} = 1$, 否则 $\delta_{\alpha\beta} = 0$ 。

边图理论因其具备许多优良的性质已得到广泛的研究。Whitney 唯一性定理指出图 G 的结构完全可以从边图 $L(G)$ 中推导出来。因此将网络图的关联矩阵转换为边图的邻接矩阵不会丢失网络图的信息。

1.2 基于边图的连边相似度构造

相似度的概念在社区发现研究中得到广泛应用, 主要是针对节点的相似度^[11,12]。在本文的边图中, 每个节点对应于网络图 G 中的每一条边。边图中两个节点之间有连边, 当且仅当这两个节点对应的边在网络图 G 中有公共的节点。针对边图中相互关联的节点来构造网络图 G 中连边的相似度。一般来说, 如果网络图中的两条边有相同或相近的节点集合与之关联, 那么这两条边被认为是相似的。基于该思想, 提出连边相似度的计算方法。

定义 3 网络图 G 的节点 i 的广泛邻居节点是指包括自己在内的, 与其距离为 1 的节点集合。记为 $n_+(i)$ 。

$$n_+(i) = \{x | d(i, x) \leq 1\} \quad (4)$$

其中, $d(i, x)$ 表示节点 i 到节点 x 的最短路径长度。

设网络图 G 中连边 e_{ik} 与 e_{jk} 共享一个节点 k , 对应于边图中这两个节点有连边。根据节点广泛邻居节点的概念, 下面给出连边相似度的计算式:

$$S(e_{ik}, e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|} \quad (5)$$

其中, $|x|$ 表示集合 x 的元素个数。

2 基于连边相似度的社区发现算法

构造完边图及连边的相似度以后, 再对边图进行聚类分析, 发现网络的社区。本文采用层次的分裂式聚类算法进行社区划分, 用连边的相似度替代 GN 算法中的边界数, 对 GN 算法进行改进, 提出一种新的基于连边分裂思想的 EGN 算法。

2.1 GN 算法存在的不足

GN 算法是基于边界数的层次分裂算法, 每一条边的边界数是指网络所有最短路径中经过该边的路径数目。GN 算法的主要思想如下: a) 计算网络中每一条边的边界数; b) 将边界数最大的边从网络中删除; c) 重新计算网络中剩余边的边界

数; d) 重复 b) 步, 直到所有的边都删除为止。

该算法在分裂的过程中要不断地计算每一条边的边界数, 过程比较复杂, 计算的时间复杂度也很高。而且该算法是基于节点集进行划分的, 这样的划分使得网络中的每一个节点只能属于某个独立的社区, 而实际网络通常存在重叠社区, 因此不能保证最终得到社区的准确性。

2.2 改进算法——EGN

本文以网络图的连边集进行划分, 最终使得连边属于某个特定的社区, 而连边所对应的节点可能同时属于多个社区, 进一步发现重叠社区。通过计算边图中节点间的相似度, 即对应于网络图中连边的相似度, 用连边相似度替代 GN 算法中的边界数, 基于连边相似度的思想对 GN 算法进行改进。这样的改进不仅降低了计算的复杂度, 而且能够发现重叠社区。如果边图中两个节点之间存在连边, 则计算出这两个节点对应于网络图中的两条边的相似度, 然后用连边的相似度替代边界数作为网络划分的指标对边网络图进行层次分裂, 即得到改进的 EGN 算法。其具体思想如下: a) 根据边图每一条边对应的一对节点, 计算出原网络图中对应连边的相似度; b) 删除边图中相似度最小的一对节点所关联的边; c) 重复步骤 b), 直到边图中所有的边都删除为止。

相对于 GN 算法每次需要重新计算边界数, EGN 算法无须重复计算连边的相似度, 这样大大减少了计算过程, 提高了算法执行的时间。EGN 算法每次删除边图中连边得到网络的一个连边的划分, 将边转换为对应的节点即可得到对应的社区。EGN 算法执行完毕以后, 最终使得网络图中每一条连边属于一个独立的社区, 进而使得两个节点属于一个社区。这样构成了一棵连边集的层次树状图, 每一层对应边图的划分, 即网络社区。为了寻找相对准确的划分, 需要提供一个衡量划分的标准, 对树状图进行分裂。Newman 等人提出的模块度衡量标准只适用于节点集的划分情况, 不适合以连边为集合的划分方法^[4], 因此需要一种新的度量标准进行社区划分的度量。本文提出一个不受网络大小限制的基于社区内部连边密度的划分密度(partition density)函数来度量网络划分的质量。假设一个网络图 G 共有 N 个节点, M 条连边, 构造 G 对应的边图 L , 对 L 按照 EGN 算法进行层次划分得到树状图。在树状图的第 C 层有 $\{p_1, \dots, p_c\}$ 共 c 个划分。第 c 个划分 p_c 共有 $m_c = |p_c|$ 条连边, 共包含 $n_c = |\cup_{e_{ij} \in p_c} \{i, j\}|$ 个节点, 且 $\sum_c n_c \geq N$ 。EGN 算法针对的是网络图的边图, 每一个划分都是网络图中的连边集合, 对应于网络中各个节点之间的关系。社区内部之间节点相互间有关联, 因此对应划分的图是连通图。根据图论的知识可知, n_c 个节点要构成连通图至少需要 $n_c - 1$ 条边, 而 n_c 个节点最多可以构成 $\frac{n_c(n_c - 1)}{2}$ 条连边。给出划分 p_c 的密度 D_c 的计算式:

$$D_c = \frac{m_c - (n_c - 1)}{\frac{n_c(n_c - 1)}{2} - (n_c - 1)} \quad (6)$$

如果 $n_c = 2$, 则 $D_c = 0$ 。

第 C 层的整体划分密度 D 是该层各个划分密度 D_c 的加权平均值, 则 D 的计算公式如下:

$$D = \frac{\sum_c D_c m_c}{\sum_c m_c} = \frac{2}{M} \sum_c m_c \frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)} \quad (7)$$

其中 $M = \sum_c m_c$ 。

划分密度越大说明划分的效果越好^[13,14], 即发现的社区更加准确。有了划分度量方法, 发现重叠社区的过程是: 先采

用 EGN 算法构造出边图的树状图,然后利用划分密度函数计算出树状图中每一层的划分密度,最后选择划分密度最大的一层进行分裂,得到最终的网络社区。

3 实验结果与分析

本章采用 Zachary 空手道俱乐部网络数据集^[15]进行实验。根据实验结果对比本文提出的 EGN 算法与传统的 GN 算法,实验结果表明 EGN 算法能够有效地发现重叠社区。

在复杂网络社区发现的研究中,Zachary 空手道俱乐部网络是经常使用的经典数据集,它反映了美国一个大学中的空手道俱乐部成员间的相互社会关系。该数据集包含 34 个节点、78 条连边,其中节点表示俱乐部的成员,连边表示成员之间的关系,网络结构如图 1 所示。方形和圆形的节点分别代表了分裂后俱乐部中的各个成员。

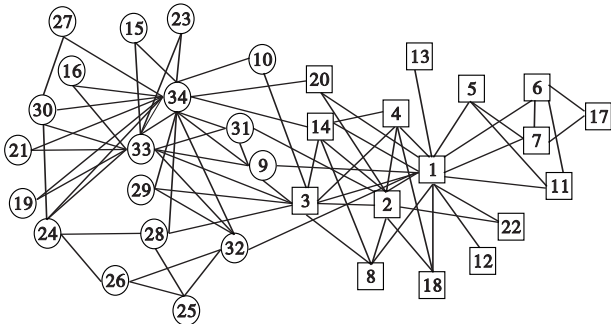


图1 Zachary空手道俱乐部网络

本文分别将 GN 和 EGN 算法应用到 Zachary 空手道俱乐部网络中。由 GN 算法得到两个独立的社区,如图 2 所示,每个圆圈中的节点集合表示一个社区。其中节点 3 本身就有歧义,同时属于两个社区,还有如节点 9、10 等也同时属于两个社区,但是只能将其划分到一个独立的社区中。由实验结果可知,GN 算法不能准确地对节点进行划分,不能发现重叠社区。

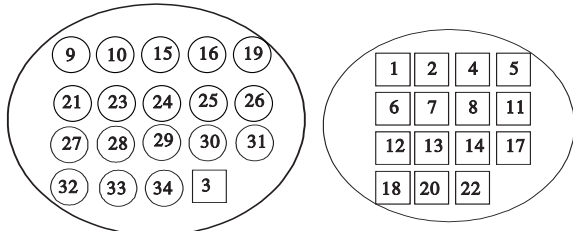


图2 GN算法发现的网络社区

采用 EGN 算法进行实验,先根据 Zachary 空手道俱乐部网络数据集构造出以连边为顶点的边图,本实验构造的边图节点数为 78,连边数为 528;然后按照 EGN 算法对边图进行划分,构造出层次树状图,计算每一层的划分密度并保存;最后求得划分密度最大值 $D=0.026\ 152\ 3$,此时得到最好的划分,对应连边的社区数目为 4。如图 3 所示,其中不同的颜色的线条表示不同的划分(见电子版)。每一条边对应两个节点,将边转换为对应的节点,即得到最终的重叠社区,如图 4 所示。

从 EGN 算法的实验结果可以看出,它能够发现重叠社区,发现的社区结构比 GN 算法发现的社区结构要更加准确,符合实际的情况。

4 结束语

本文对传统以节点集进行划分的 GN 算法进行改进,提出了以节点之间的连边进行划分的 EGN 算法,并提出了连边相似度的计算公式及划分衡量标准划分密度函数的定义及计算

公式,从而发现重叠社区,并将算法应用到真实的数据集 Zachary 空手道俱乐部网络中。实验表明 EGN 算法能够很好地发现重叠社区。未来的工作将致力于提高算法的工作效率和发现社区的准确度。

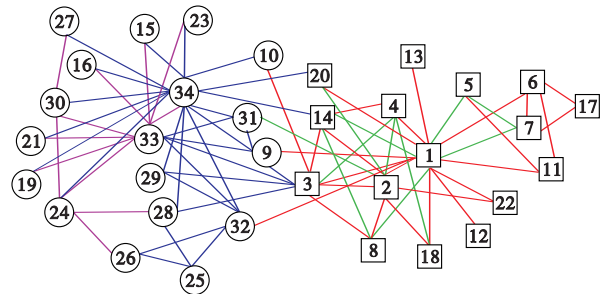


图3 按照连边集的网络划分

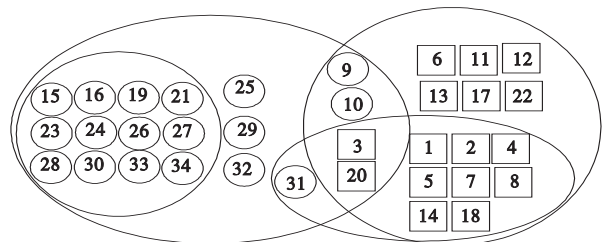


图4 边集对应的节点划分

参考文献:

- [1] AMARAL L A N, SCALA A, BARTHELEMY M, *et al.* Classes of small-world networks[J]. *Proceedings of the National Academy Sciences USA*, 2000, 97(21): 11149-11152.
- [2] REDNER S. How popular is your paper? An empirical study of the citation distribution[J]. *European Physical Journal B*, 1998, 4(2): 131-134.
- [3] PASTOR-SATORRAS R, VESPIGNANI A. Epidemic dynamics and endemic states in complex networks[J]. *Physical Review E*, 2001, 63(6): 066117.
- [4] DREWES G, BOUWMEESTER T. Global approaches to protein-protein interactions[J]. *Current Opinion in Cell Biology*, 2003, 15(2): 199-205.
- [5] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. *Proceedings of the National Academy Sciences USA*, 2002, 99(12): 7821-7826.
- [6] KERNIGHAN B W, LIN S. An efficient heuristic procedure for partitioning graphs[J]. *Bell System Technical Journal*, 1970, 49(2): 291-307.
- [7] POTHEN A, SIMON H, LIOU K P. Partitioning sparse matrices with eigenvectors of graphs[J]. *SIAM Journal on Matrix Analysis and Application*, 1990, 11(3): 430-452.
- [8] NEWMAN M E J. Fast algorithm for detecting community structure in networks[J]. *Physical Review E*, 2004, 69(6): 066133.
- [9] CLAUSET A, NEWMAN M E J, MOORE C. Finding community structure in very large networks[J]. *Physical Review E*, 2004, 70(6): 066111.
- [10] ZHOU Tao, REN Jie, MEDO M, *et al.* Bipartite network projection and personal recommendation[J]. *Physical Review E*, 2007, 76(4): 046115.
- [11] LEICHT E A, HOLME P, NEWMAN M E J. Vertex similarity in networks[J]. *Physical Review E*, 2006, 73(2): 026120.
- [12] HU Yan-qing, LI Meng-hui, ZHANG Peng, *et al.* Community detection by signaling on complex networks[J]. *Physical Review E*, 2008, 78(1): 016115.
- [13] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. *Physical Review E*, 2004, 69(2): 026113.
- [14] AHN Y Y, BAGROW J P, LEHMANN S. Communities and hierarchical organization of links in complex networks[EB/OL]. (2009-03-19). http://www.bradblock.com/S3-website-us-west-1.amazonaws.com/Communities-and-Hierarchical-Organization-of_Links_in_Complex-Networks.pdf.
- [15] SHANG Ming-sheng, CHEN Duan-bing, ZHOU Tao. Detecting overlapping communities based on community cores in complex networks[J]. *Chinese Physics Letters*, 2010, 27(5): 058901.