基于复杂网络的社区发现算法研究

孟彩霞,李楠楠,张 琰
(西安邮电大学 计算机学院,陕西 西安 700121)

摘 要:近年来,高质量社区的挖掘和发现已经成为复杂网络研究的一个热点。目前大多的社区发现算法主要针对无向网络,但现在的很多真实网络通常都是有向加权的。同时,标签传播算法(LPA)是一种接近线性复杂度的社区发现算法,该算法具有简单高效、不需要提供社区规模和社区个数等先验知识的特点,因而得到了广泛关注和应用。针对有向加权网络,提出了一种基于节点重要性和节点相似性的改进标签传播算法(CRJ-LPA)。该算法综合考虑节点的边权、节点的信息传播能力、节点相似度以及节点集聚系数等因素。算法通过加权的 ClusterRank 获得节点重要性列表用以避免 LPA 中的随机选择;然后,采用 Jaccard 系数度量节点的相似度,结合节点重要性列表计算出一个新的度量 CRJ(重要度和相似度),提高了算法的稳定性。实验结果表明,该算法有效可行,且具有较好的鲁棒性。

关键词:有向加权网络;标签传播;ClusterRank;节点重要性;Jaccard;节点相似度
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2020)01-0082-05
doi:10.3969/j.issn.1673-629X.2020.01.015

Research on Community Detection Algorithm Based on Complex Network

MENG Cai-xia,LI Nan-nan,ZHANG Yan
(School of Computer Science and Technology,Xi'an University of Posts and Telecommunications,Xi'an 700121,China)

Abstract:In recent years,the mining and discovery of high-quality communities has become a hot topic in complex network research. However,most community discovery algorithms are mainly directed at undirected networks,but many real networks are usually directed weighted. At the same time,the label propagation algorithm (LPA) is a community discovery algorithm close to linear complexity. It is simple and efficient,and does not need to provide prior knowledge such as community size and community number,which has been widely concerned and applied. For the directed weighted network,a label propagation algorithm (CRJ-LPA) based on node similarity and node importance is proposed. The node importance list is obtained by weighted ClusterRank to avoid random selection in LPA. Then,the Jaccard coefficient is used to measure the similarity of the nodes. Combined with the node importance list,a new metric CRJ (importance and similarity) is calculated to improve the stability of the algorithm. Experiment shows that the proposed algorithm is feasible and effective with strong robustness.

Key words:directed weighted network;label propagation;ClusterRank;node importance;Jaccard;node similarity

0 引 言

在现实世界中存在各种复杂系统,这些系统通常可以以网络的形式表达,比如常见的电力网络、航空网络以及社交网络等复杂网络。复杂网络具有小世界、无标度、社区结构等许多基本特性,而其中最为重要的特性是社区结构。为了挖掘这些社区结构,可以使用一些不同领域的方法,如数据挖掘中的聚类或图论中的图分区等,挖掘社区结构的过程统称为社区发现[1]。通常将网络表示为图,图中的点表示网络中具体的实体,边表示网络中实体与实体之间的关联[2-3]。大多数关于社区检测的论文使用图作为网络的数学表示,更精确地说是无向图。然而,很多真实网络具有复杂的关系,并且都是有权值和方向的。此外,将有向图转化为无向图会导致信息的丢失,从而使检测到的社区结构没有真正意义[4]。由于很少有文献提出在有向网络中进行社区检测,因此对有向有权的复杂网络进行社区发现是一项艰巨而有意义的任务[5]。

2007 年,Raghavan 等[6]提出了一种标签传播算法

收稿日期:2019-02-20 **修回日期**:2019-06-24 **网络出版时间**:2019-09-24
基金项目:陕西省自然科学基金(2014JM8303);西安邮电大学研究生创新基金(CXL2016-40)
作者简介:孟彩霞(1966-),女,研究生导师,研究方向为大数据处理与数据挖掘;李楠楠(1994-),女,硕士研究生,研究方向为复杂网络和数据挖掘。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190924.1537.058.html

(LPA),该算法是一种近似线性复杂度的社区发现算法,并且不需要预先知道社区的规模大小和所需要划分的社区个数等,因此受到学者们的广泛关注和应用。但 LPA 在标签传播过程中存在随机性、振荡、不稳定,划分社区效果差等缺点,为此大量研究人员进行了相关研究。Sun 等[7]提出了一种基于 α - degree 邻域影响的标签传播算法,缓解了节点更新中随机更新的问题,提高了算法的稳定性。Yan Xing 等[8]提出了 KBLPA 和 NIBLPA 算法,该算法以 K-shell 算法为依据分析节点的重要性。易秀双等[9]提出了一种基于顶点影响的局部社区发现算法,提高了算法的计算速度和效率。黄佳鑫等[10]在标签传播的思想上综合考虑了节点的重要性和标签的影响力,因此提高了原始标签传播算法的稳定性和准确性。彭磊等[11]依据节点相似度进行更新,提出了 NSLPA 算法。许合利等[12]提出了一种基于核心节点的加权网络中的局部检测算法 CRD-LPA。但是以上这些算法大多数是基于无向图的,因此失去了一些有用的信息,只对社区检测结果进行定量分析。

文中考虑边的方向和权值,将标签传播思想应用于有向加权网络,并且通过加权的 ClusterRank 获得节点重要性列表,以避免 LPA 中的随机选择。其次,采用 Jaccard 系数度量节点的相似度,结合节点重要性列表计算出一个新的度量 CRJ(重要度和相似度),提高算法的稳定性和社区发现质量。

1　标签传播算法

标签传播算法是一种接近线性复杂度的社区发现算法,其基本思想是用已知节点标签信息预测未知节点的标签。

具体算法描述如下:

(1)将所有节点的标签初始化为唯一值,例如初始化节点标签为其 ID 号。

(2)随机地对图中的所有节点进行排序。

(3)根据步骤 2 按顺序更新每个节点,将节点的标签更新为邻居中出现次数最多的标签;若当个数最多的标签不唯一时,随机选一个标签赋给当前节点。

(4)如果网络中的所有节点的标签均稳定不变,则算法终止。否则,返回步骤 2 继续。

基于标签传播算法的社区检测的具体过程如图 1 所示。

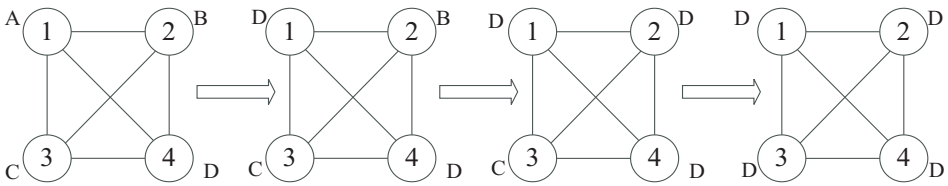

图 1　基于 LPA 的标签传播过程

在图 1 中有四个节点,每个节点 ID 为 1,2,3,4,它们的标签被初始化为 A,B,C 和 D。

在标签传播过程中,节点 1 的标签随机选择为节点 4 的标签 D 后,与节点 2 相邻的节点中,标签 D 的数量最多,因此节点 2 的标签也设置为 D,这样的过程不断持续下去,直到所有可能聚集到一起的节点都具有了相同的社区标签,此时图 1 中所有节点的标签都变成了 D,所有节点都已达到算法的终止然后退出循环。

标签的更新策略分为:同步更新和异步更新。同步更新是指对于节点 x,在第 t 代时,根据其邻居在 $t-1$ 代时的社区标签进行更新。异步更新是指节点 x,在第 t 代时,根据其邻居最新的社区标签进行更新。同步更新标签的方法对于二分或者近似二分的网络来说,可能会导致标签的振荡,所以选择异步更新节点标签的方式。

LPA 算法的随机性有以下两个方面的问题:

(1)节点更新顺序的随机性。每次迭代开始时,都需要重新随机生成节点更新的顺序。但是,这种随机性的方法不仅可以产生最佳值,也可能会产生最差值,因此,增加了算法的不稳定性。

(2)当个数最多的标签不唯一时,标签选择是随机的。这种随机性可能会使得算法的迭代次数增加,并且导致算法不稳定,划分出来的结果也会相对较差。

针对第 1 个问题,提出基于加权的 ClusterRank 算法获得节点重要性列表来依次更新节点,避免随机选择;针对第 2 个问题,采用 Jaccard 系数度量节点的相似度,结合节点重要性列表计算出一个新的度量 CRJ(重要度和相似度),选择度量值最高的标签进行更新,提高算法的稳定性和社区发现质量。

2　CRJ-LPA:改进的标签传播算法

LPA 的效率吸引了众多学者和研究人员的关注和研究。有很多算法可以改善 LPA 的上述问题。NSLPA 算法最大改进之处在随机选择。如果有多个可选标签,则节点将选择相似度的邻居节点的标签,而不是随机选择。此方法在一定程度上避免了 LPA 的随机性问题,

但仍存在逆流问题。CRD – LPA 算法将 ClusterRank 系数与节点局部密度(local density of node,LDN)结合起来进行节点更新。此方法提高了 LPA 的准确性和稳定性,但 CRD 函数降低了节点影响力相同的概率,仍存在随机选择的可能性,同时该算法也忽略了节点边的方向性对结果的影响。

2.1 加权的 ClusterRank 算法

Chen 等[13]根据节点的度和聚类系数对有向复杂网络的节点重要性进行了分析,并以此为基础提出了 ClusterRank 算法。该算法在考虑节点的邻居节点的数量的同时,还考虑到聚类系数对网络中信息传播的巨大影响。ClusterRank 算法是对 LeaderRank 和 PageRank 算法做了进一步的优化和改进[14],但是 ClusterRank 没有考虑网络中节点周围的结构信息和边的权值,因此,无法有效地衡量有向加权网络中节点的重要性。考虑到这个问题,文中结合含权网络中节点强度的定义提出了基于加权的 ClusterRank 算法。

2.1.1 含权网络中的节点强度

在加权定向网络中,节点 v_i 的度称为强度,节点的强度可分为出强度 b_i^{out} 和入强度 b_i^{in}。b_i^{out} 为与相连的出边的权值之和,b_i^{in} 为与相连的入边的权值之和。定义如下所示:

$$b_i^{\text{out}} = \sum_{j \in \Gamma_i^{\text{out}}} w_{ij} \tag{1}$$

$$b_i^{\text{in}} = \sum_{j \in \Gamma_i^{\text{in}}} w_{ji} \tag{2}$$

其中,Γ_i^{out} 为节点 v_i 的出边的集合;Γ_i^{in} 为节点 v_i 的入边的集合。

上面定义的缺点很明显,忽视了节点的度,在网络中往往存在节点的邻居多而节点强度却很小的情况。Garas 等[15]提出了另一种节点强度的定义方式,即用节点的邻居数量和边权重共同表示节点的度值,更加细致地刻画了节点的属性。在这里,节点 v_i 的强度为:

$$k_i' = \left[k_i^{\alpha}\left(\sum_{j}^{k_i} w_{ij}\right)^{\beta}\right]^{\frac{1}{\alpha+\beta}} \tag{3}$$

其中,k_i 为节点 v_i 的度;w_{ij} 为节点 v_i 与其邻居 v_j 之间连边的权值;α 和 β 为自由参数,用来调节度和权值之间的比重。

2.1.2 含权的局部聚类系数

许多社交网络把有向网络从 i 到 j 的连接表示为 j 是 i 的追随者,意味着 j 从 i 接收信息。将 Γ_i 表示为 i 的追随者集合,即 i 的出边集合,并且 i 的追随者之间的相互作用密度可以用 i 的局部聚类系数表示。有向网络的聚类系数定义为:

$$c_i = \frac{|\{e_{jk} \mid j,k \in \Gamma_i\}|}{k_i^{\text{out}}(k_i^{\text{out}} - 1)} \tag{4}$$

其中,k_i^{out} 为节点 i 的出度,即 i 的追随者的数量;$\{e_{jk} \mid j,k \in \Gamma_i\}$ 为 i 的两个追随者之间的连接集合。如果 $k_i^{\text{out}} \leqslant 1$,则 $c_i = 0$。

现有研究提出了计算适用于有向网络和加权网络的局部聚类系数的方法,但这些并不适用于加权定向网络。考虑到这一点,文中融合 Garas 等提出的节点强度概念和信息传播的因素,定义了加权定向网络上的局部聚类系数,如下所示:

$$c_i = \frac{\{w_{jk} \mid j,k \in \Gamma_i\}}{w_i^{\text{out}} k'^{\text{out}}_i (k'^{\text{out}}_i - 1)} \tag{5}$$

$$\text{s.t.} \quad k'^{\text{out}}_i = \left[(k_i^{\text{out}})^{\alpha}\left(\sum_{j}^{k_i^{\text{out}}} w_{ij}\right)^{\beta}\right]^{\frac{1}{\alpha+\beta}}$$

其中,w_i^{out} 为节点 v_i 的出边的权值之和;$\{w_{jk} \mid j,k \in \Gamma_i\}$ 为节点 v_i 的追随者之间连边的权值集合。

2.1.3 含权的 ClusterRank 算法

对于 ClusterRank 只考虑节点的聚类系数,不适用于加权网络的问题,提出了适用于加权定向网络的 ClusterRank 算法。根据式 5 定义的加权定向网络上的局部聚类系数,重新定义了节点 v_i 的 ClusterRank 的评分 s_i:

$$s_i = f(c_i) \sum_{j \in \Gamma i} (k'^{\text{out}}_j + w_{ij}) \tag{6}$$

$$\text{s.t. } f(c_i) = 10^{-c_i}$$

其中,Γ_ι 是节点 v_i 的邻居节点集合;w_{ij} 是节点 v_i 与节点 v_j 直接相连的边的权值;$f(c_i)$ 是节点 v_i 的聚类系数的函数。

在复杂网络中,聚类系数越大,越会阻碍信息的传播,因此随着 c_i 增大的 $f(c_i)$ 值将变小。

2.2 Jaccard 相似度

在复杂网络中,节点之间通常具有一定的相似性,Jaccard 为描述相似度的重要指标。在包含节点集 V 和边集 E 的图 $G(V,E)$ 中,节点 v_i 和节点 v_j 之间的 Jaccard 相似度定义如下:

$$\text{Jaccard}(v_i,v_j) = \frac{|N_i \cap N_j|}{|N_i \cup N_j|} \tag{7}$$

其中,N_i 表示节点 v_i 的邻居节点的集合,Jaccard 的值介于 0 ~ 1 之间,该值越接近 1,表示节点 v_i 和节点 v_j 之间的相似度越高。

在 LPA 算法中,即使通过文中提出的基于加权的 ClusterRank 算法进行节点重要性排序后进行标签的更新,仍然有可能会出现一定的随机选择。因此,定义了一种新的度量 CRJ,通过综合考虑节点重要性和相似性来提高 LPA 算法的准确性,定义如下:

$$\text{CRJ}(i,j) = \frac{S_i}{\sqrt{\sum_{j=1}^{N} S_j^2}} + \frac{J_i}{\sqrt{\sum_{j=1}^{N} J_j^2}} \tag{8}$$

其中,S_i 表示节点 v_i 的重要性,J_i 表示节点 v_i 和节点 v_j 之间的相似度。采用同趋化函数 $g(x)=\frac{x}{\sqrt{\sum x^2}}$ 对 S_i 和 J_i 同时进行处理,使 CRJ (i,j) 能够综合衡量节点重要性 S_i 和相似度 J_i 不同作用的结果,准确地将节点重要性和相似度结合起来。

2.3 CRJ-LPA 算法描述

针对有向加权网络,基于原始的 LPA 算法,文中提出了一种基于节点重要性和相似性的改进 CRJ-LPA 算法。该算法具体步骤如下:

Step1:初始化,根据节点 ID 为每个节点分配一个唯一的标签;

Step2:根据式 6 计算所有节点的重要性,并根据节点重要性由高到低对节点集合 V 进行排序;

Step3:根据式 7 计算节点的相似度;

Step4:从节点集合 V 中依次取出节点进行更新,并且优先更新邻居节点间具有最大影响力的节点,如果出现影响力相同的情况,则根据式 8 计算邻居节点的 CRJ(v,v'),然后将节点 v 的标签更新为具有最高 CRJ(v,v') 的邻居节点 v' 的标签;其次,在标签更新过程中,如果节点的邻居节点中个数最多的标签出现两个或多个时,同样根据 CRJ(v,v') 来更新节点 v 的标签;

Step5:如果网络中的所有节点的标签均稳定不变,则循环停止并退出算法。否则,跳转到 Step4 继续循环。

3 实验结果与分析

选取 Lesmis 与 Celegansneural 两种国际上公认的真实数据集,对 CRJ-LPA 算法进行测试。算法的实验环境为 Python3.5 软件,硬件配置为 i5-3230M,RAM:4.00G;软件配置:64 位 WIN7 操作系统。

3.1 有向加权网络模块度

文献[13]中 Newman 和 Girvan 提出了模块度的概念,后来作为衡量社区算法性能的公认评价标准。再后来,Newman 等将其拓展到有向、加权网络上[16],定义如下:

$$Q=\frac{1}{w}\sum_{ij}\left(w_{ij}-\frac{w_i^{out}w_j^{in}}{w}\right)\delta(c_i,c_j) \tag{9}$$

其中,w 为网络中所有边的权值之和,w_{ij} 为节点 v_i 和 v_j 之间连边的权值,w_i^{out} 为节点 v_i 出边权值之和,w_j^{in} 为节点 v_j 入边权值之和,c_i 为节点 v_i 所在社区,c_j 为节点 v_j 所在社区。若 c_i 与 c_j 相等,则 $\delta(c_i,c_j)$ 的值为 1,否则为 0。模块度用来衡量社区结构性的强弱,其值越接近 1,表示划分出的社区结构越强,划分的结果越好。通常采用模块度作为社区发现算法的评价标准,该值在[0.3,0.7]的区间内,表示社区划分质量较好。

3.2 实验结果

数据集 Lesmis 与 Celegansneural 是两种有向有权复杂网络数据集,其基本信息如表 1 所示。

表 1　真实复杂的网络数据集信息

数据集	节点数	边数
Lesmis	$N=77$	$E=254$
Celegansneural	$N=297$	$E=2\ 359$

在这两种真实数据集上对算法进行分析与验证,并且根据模块度来衡量算法划分的社区结构的优劣。同时,将文中算法 CRJ-LPA 与传统 LPA 算法(如 LPA、NSLPA、KBLPA 算法)进行比较。不同算法分别在数据集上进行运算后的模块度如表 2 所示。

表 2　算法模块度的比较

算法	Lesmis	Celegansneural
LPA	0.133 0	0.019 9
NSLPA	0.110 4	0.023 0
KBLPA	0.150 8	0.072 9
CRJ-LPA	0.412 7	0.359 6

通过对表 2 中的实验数据进行分析可以看出,与传统的 LPA、NSLPA、KBLPA 算法相比,文中算法发现的社区结构的平均模块度最大。从上述精准的数字描述可以看出,文中算法在这两种有向有权复杂网络数据集上比传统 LPA 等算法在性能上有明显提升,且模块度的值均在良好社区结构的模块度区间[0.3,0.7]范围内。因此,文中算法划分的社区结构良好,且算法准确性和稳定性较高。文中算法与 LPA 算法对 Lesmis 数据集划分的结果如图 2 与图 3 所示。

图 2 和图 3 将位于不同社区的节点用直线分隔开,并且通过两幅图的对比可以得出,文中算法发现的社区较传统 LPA 算法所得社区数量多,且较为稳定,没有超大社区。

4 结束语

针对有向加权网络,提出了一种基于节点重要性和节点相似性的改进标签传播算法(CRJ-LPA)。该算法综合考虑了复杂网络中边的权值和方向性,并且采用标签传播的思想进行社区发现。首先,通过有向加权的 ClusterRank 算法获得节点的重要性排序列表,然后根据此顺序更新节点标签,提高社区结构的划分质量;其次,在节点更新过程通过节点重要性和相似性计算出一个新的度量 CRJ,以此来避免原始 LPA 中的随机选择,有效克服了传统标签传播算法的随机性。通过真实数据集对算法进行测试,发现该算法具有较

好的可行性和准确性,能够准确地衡量节点的重要性,而且与 LPA 算法具有相似的时间复杂度。

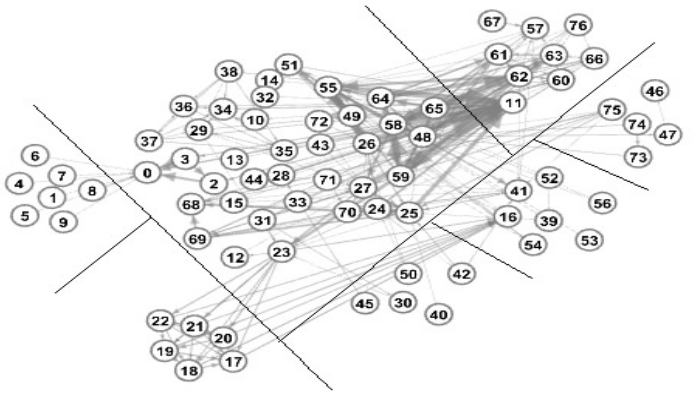

图 2 文中算法效果

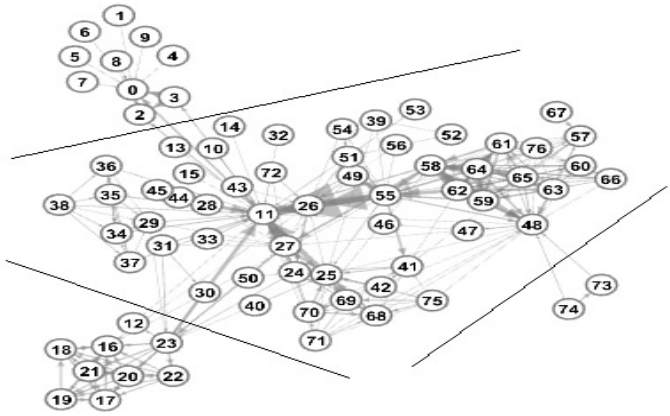

图 3 LPA 算法效果

参考文献:

[1] AGRESTE S, MEO P D, FIUMARA G, et al. An empirical comparison of algorithms to find communities in directed graphs and their application in web data analytics[J]. IEEE Transactions on Big Data, 2017, 3(3): 289-306.

[2] NEWMAN M E J. The structure and function of complex networks[J]. SIAM Review, 2003, 45(2): 167-256.

[3] 王 丹,刘发升. 复杂网络的社区发现算法研究[J]. 计算机时代,2009(3):57-59.

[4] LEICHT E A, NEWMAN M E J. Community structure in directed networks[J]. Physical Review Letters, 2008, 100(11): 118703.

[5] 杨晓光,朱保平. 基于复杂网络的社区发现算法[J]. 南京理工大学学报:自然科学版,2016,40(3):267-271.

[6] RAGHAVAN U N, ALBERT R, KUMARA S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E, 2007, 76(3): 036106.

[7] SUN Heli, HUANG Jianbin, ZHONG Xiang, et al. Label propagation with α-degree neighborhood impact for network community detection[J]. Computational Intelligence and Neuroscience, 2014(1): 130689.

[8] XING Yan, MENG Fanrong, YONG Zhou, et al. A node influence based label propagation algorithm for community detection in networks[J]. The Scientific World Journal, 2014, 2014: 627581.

[9] 易秀双,韩业挺,王兴伟. 一种基于节点影响力的局部社区发现算法[J]. 小型微型计算机系统,2013,34(9):1975-1979.

[10] 黄佳鑫,郭 昆,郭 红. 融入节点重要性和标签影响力的标签传播社区发现算法[J]. 小型微型计算机系统,2015,36(6):1171-1175.

[11] 彭 磊. 基于标签传播的社区发现算法的研究[D]. 西安:西安电子科技大学,2015.

[12] 许合利,宁念文,牛丽君. 一种结合节点局部影响力的标签传播算法[J]. 小型微型计算机系统,2017,38(6):1299-1304.

[13] CHEN Duanbing, GAO Hui, LÜ Linyuan, et al. Identifying influential nodes in large-scale directed networks: the role of clustering[J]. PloS One, 2013, 8(10): e77455.

[14] 任晓龙,吕琳媛. 网络重要节点排序方法综述[J]. 科学通报,2014,59(13):1175-1197.

[15] GARAS A, SCHWEITZER F, HAVLIN S. A k-shell decomposition method for weighted networks[J]. New Journal of Physics, 2012, 14(8): 083030.

[16] NEWMAN M E J. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2): 026113.