基于核心图的标签传播算法

马杰良[1] 韩 路[2] 潘贞贞[2] 宋 艳[2]
(南京信息工程大学信息与控制学院 南京 210044)[1]
(南京信息工程大学电子与信息工程学院 南京 210044)[2]

摘 要 网络中的社团发现是当前的一个研究热点。在众多社团发现算法中,标签传播算法因简单快速而被广泛应用,但标签传播算法也存在结果稳定性较差的问题。基于此对标签传播算法的初始化过程进行改进,提出了基于核心图的标签传播算法。通过计算图中任意两点的k阶公共邻居,将具有最大相似性的节点及k阶邻居作为初始核心社团,并为其分配初始标签。通过上述过程,提取一些较为紧密的子结构来作为标签传播的初始社团,并给这些结构分配初始社团标签。在真实网络中的实验结果表明,该算法可以大幅提高结果的稳定性。

关键词 社团发现,标签传播,相似性,核心图
中图法分类号 TP301.6 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.1.028

Label Propagation Algorithm Based on Community Core for Community Detection

MA Jie-liang[1] HAN Lu[2] PAN Zhen-zhen[2] SONG Yan[2]
(College of Information and Control, Nanjing University of Information Science and Technology, Nanjing 210044, China)[1]
(College of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China)[2]

Abstract Community detection in networks is a hot research topic currently. Among many community detection algorithms, label propagation algorithm is widely used for it is simple and rapid. But label propagation algorithm also has the problem of poor stability result. Therefore, we improved the initialization process of label propagation algorithm. We proposed a label propagation algorithm based on community core for community detection, and by calculating any two nodes's k order common neighbor, we used the most sililar nodes and it's k order neighbor nodes as the initial community core. According the above process, we got some tight structure as the initial label of label propagation, and assigned the initial community label to these structures. Experimental results in a real network show that the algorithm can improve the stability of the results.

Keywords Community detection, Label propagation, Similarity, Community core

1 引言

自然界与人类社会中诸多系统可以采用复杂网络模型表示,以节点表示系统中实体,边表示实体之间的联系,如人际关系网、论文合作网、电影明星合作网、电子邮件往来网、博客引用网、电话通讯网等。许多复杂网络具有社团结构,社团可看作网络的一个子图,社团内部节点连接紧密,社团间节点连接稀疏。研究表明社会网络[1,2]、生物化学网络[3]都具有明显的社团结构。网络结构与网络功能具有紧密的联系,研究网络社团结构能够揭示隐藏在复杂网络中的规律并有助于行为预测和控制,如Web社团划分[4]、蛋白质网络分析和功能预测等[5]。由此可见,对复杂网络中的社团结构进行发现与分析,可以很好地理解其结构和行为,同时也具有更为重要的研究价值与意义。

许多学者从不同角度对如何发现网络中的社团结构问题进行了研究。目前,人们已经给出了多种社团发现方法。这些方法分为基于优化的社团发现方法、基于启发式策略的社团发现方法及其他方法3类。基于优化的方法有谱方法和模块性最大化方法。基于启发式策略方法有GN算法[6]、WH-Huberman算法[7]、MFC算法[8]、HITS算法[9]、CPM算法[10]等。除此之外,还有其他一些有效的社团发现方法,如基于层次聚类的社团发现方法(凝聚式和分裂式)、基于非负矩阵分解的社团发现方法等。

现有的社团发现算法存在算法复杂度高、需要事先制定社团的数目和预先制定评价指标等缺陷,有的甚至需要给出大致的社团大小,因而限制了算法的实际应用效率。文献[11]提出了一种基于标签传播的社团发现算法LPA(Label Propagation Algorithm, LPA),该算法为每个节点分配一个标签,并随着传播过程更新标签,顶点的标签应该与它的邻居中的具有最多相同标签的顶点保持一致。该算法的优点是计算过程非常简单,计算速度非常快,但缺点是算法的稳定性较差,连续几次运行结果可能会很不相同。针对此问题,本文提

到稿日期:2014-02-10 返修日期:2014-05-05 本文受国家自然科学基金(61372128)资助。
马杰良(1964-),男,副教授,主要研究方向为复杂网络、离散系统,E-mail:njkjmjl@163.com;韩 路(1989-),女,硕士生,主要研究方向为复杂网络;潘贞贞(1990-),女,硕士生,主要研究方向为复杂网络;宋 艳(1990-),女,硕士生,主要研究方向为复杂网络。

出了基于核心图的标签传播算法，即把一些较为紧密的子结构找出来，标上相同的标签作为传播的初始标签，而不是给每个节点分配一个标签，通过这些核心图进行标签传播。另外，在传播过程中，如果节点存在多个标签数量同为最大值的情况，那么可以计算节点与社团的相似度，哪个相似度大，就把节点归入那个社团。通过这些方法，能降低 LPA 算法的不稳定性。

2 相关知识

2.1 LPA 算法

基于标签传播的社区发现算法的主要思想是：一个给定的顶点 x 拥有 k 个邻接点 $x_1,x_2,x_3,\cdots,x_k$，每个邻接点都有一个代表其所在社区的标签，x 的标签由它邻接点的标签共同决定，以确定 x 所在的社区。该算法可简述如下：

1)首先为图中的每个顶点分配一个唯一的标签(如一个整数)作为社区的标识，代表所在社区。以标签的顺序排列网络中的节点，并将排序结果放在 X 中。

2)按照 X 中存储的顺序，依次对节点更新。对每个顶点 x 来说，它的标签由所有邻接点的标签中数量最多的那个标签代替，以此更新自己的标签。如果有多个标签的数量同为最大值，则随机选取一个标签作为该顶点的标签。经过若干次的迭代后，每个顶点邻居中的标签变化趋于稳定。

3)将所有具有相同标签的顶点归为一个社区。

然而该算法不能保证迭代过程在若干次之后能够收敛。若算法采取对顶点标签进行同步更新(即在第 t 次迭代时，x 的标签由其邻接点在第 $t-1$ 次迭代过程后的标签所决定)，则这种方式在具有二分结构的图上会导致标签的循环震荡。因此 Raghavan 等人在文献[11]中采取了异步更新的方法，其更新公式如下：

$$C_x(t)=f(C_{x_{i1}}(t),\cdots,C_{x_{im}}(t),C_{x_{i(m+1)}}(t-1),\cdots,C_{x_{ik}}(t-1))$$

其中，$x_{i1},\cdots,x_{im}$ 表示在本次迭代中已经更新过标签的节点；$x_{i(m+1)}(t-1),\cdots,x_{ik}(t-1)$ 表示在此次迭代中标签还没有更新的节点。在 x 的邻接点中，用一部分在第 t 次迭代后的顶点标签以及另一部分在 $t-1$ 次迭代后的标签共同决定 x 的标签，以避免标签震荡现象出现。

在标签传播算法中，网络中每个顶点在初始时都被分配一个唯一的标签，在传播过程中进行动态的标签更新，会导致零散的、孤立的小社区大量出现，迫使一些真正意义上的社区无法生成。由于每个顶点都有唯一的标签，很多影响力较小的顶点在传播过程中会反过来影响一些影响力较大的顶点，使标签传播过程发生一种消耗资源的“逆流”现象。另外，当一个顶点 x 的邻接点中具有多个满足条件的候选标签时，将会随机选择 x 的标签，这种随机性会很大程度地降低算法的稳定性，过多随机性叠加会导致多次实验的结果有相当程度的差异性。

2.2 k 阶共同邻居节点及节点的相似性

定义 1(图 G 的距离) 即所有节点对之间的平均图距离。互相连接的节点的图距离为 1，直径是最长的任何两个节点之间的距离(即两个最遥远的节点相距有多远)。

定义 2(k 阶公共邻居) 我们用 $G=(V_G,E_G)$ 来表示一个网络，其中 V_G 是图的节点集，节点的个数 $n=|V_G|$，$E_G=(v_i,v_j,w_{ij})$ 表示图含有权重的边集，w_{ij} 表示节点 i,j 之间边的权重。节点 v_1 和 v_2 的 k 阶公共邻居集的定义如下：

$$N_k(v_1,v_2)=\{v|d(v_1,v)=d(v_2,v)=k\}=N(v_1,k)\cap N(v_2,k),k\geqslant 1$$

其中，$N(v_1,k)$ 是指距节点 v_1 为 k 阶的邻居集，而 $N_k(v_1,v_2)$ 是指节点 v_1,v_2 的 k 阶公共邻居集。

定义 3(k 阶共同邻居节点的相似性) 一个网络中 k 的取值是由网络本身的特性决定的，定义如下：

$$\bar{k}=\frac{\sum k_{\max}|N(i)\cap N(j)|}{N(N-1)/2}$$

其中，$k_{\max}|N(i)\cap N(j)|$ 是指节点 i,j 之间拥有最大共同邻居数的阶数。一个网络中的 k 值是指整个网络中所有节点对之间的平均值。

节点间的 k 阶共同邻节点的相似性定义如下：

$$Sim(v_1,v_2)=\frac{|N(v_1,k)\cap N(v_2,k)|}{|N(v_1,k)\cup N(v_2,k)|},k\geqslant 1$$

其中，k 是整数。我们用 k 阶共同邻居节点的相似性找出核心社团，就是为了能够较快、较准确地得到初始核心社团。

定义 4(节点间的相似度)

$$\sigma=\frac{|N(v_i,k)\cap N(v_j,k)|}{|N(v_i,k)\cup N(v_j,k)|-2}$$

其中，$|N(v_i,k)\cap N(v_j,k)|$ 表示节点 i 和节点 j 的共同邻居数，$\sigma\in[0,1]$。本文中，σ 由网络中的平均共同邻居节点来决定，取 $|N(v_i,k)\cap N(v_j,k)|$ 为网络中的平均共同邻居节点数目，它表示节点对之间的平均共同邻居数。

定义 5(节点与社团的相似度) 若将 G 划分为 k 个社团 $C_1,C_2,\cdots,C_k$，设节点 v_i 是社团 C_i 内节点的一个邻居，社团 C_j 内与节点 v_i 相邻的节点个数作为社团对节点 v_i 的支持值，表示如下：

$$sup(C_j,v_i)=\frac{|N(v_i)\cap C_j|}{|N(v_i)\cup C_j|},v_i\in V$$

其中，$|N(v_i)\cap C_j|$ 表示节点 v_i 的邻居与社团 C_j 交集的点数，含义是社团 C_j 内与节点 v_i 相连的节点数。$|N(v_i)\cup C_j|$ 表示节点 v_i 的邻居与社团 C_j 并集的点数。

3 算法描述

3.1 初始核心社团的选取

本文提出的算法是对标签传播算法的初始化过程进行改进，根据定义 3 计算网络中任意两个节点拥有最大公共邻居的平均阶数，将具有最大相似性的节点及 k 阶邻居作为初始核心社团，并为其分配初始标签。

具体算法如下：

1)初始时，计算所有点对的 k-阶公共邻居集。

2)选择具有最大公共 k 阶邻居的点对。计算相似度时，将具有最大相似度点对及其 k 阶公共邻居作为第一个初始核心社团，并赋予标签 1。

3)依据步骤 2 所述的过程，得到另外一个子结构，并将这个子结构中的节点按顺序标号。

4)重复这个过程，直到 X 中没有标过号的节点都不满足 $Sim(v_1,v_2)\geqslant\sigma$，算法结束。$\sigma$ 是节点的相似性指标。

5)对得到的这些小的社团进行合并，如果任意两个社团满足 $|C_i\cap C_j|\geqslant\frac{1}{2}|C_i\cup C_j|$，则对这两个小社团进行合并。

经过这些计算，我们得到若干个局部核心子结构，这些核心图连接较为紧密，将其作为初始核心社团，并作为标签传播的初始标签，成为后面标签传播的出发点。

3.2 基于标签传播算法的社团划分

1)我们将从 3.1 节得到的若干个标过号的子结构作为标签传播的初始标签状态。

2)对网络中剩余的节点进行排序，剩余节点按照与初始核心社团的 k 阶邻居进行排序，k 从小到大，生成 X。

3)for $(x\in X)C_x(t)=f(C_{x_{i1}}(t),\cdots,C_{x_{im}}(t),C_{x_{i(m+1)}}(t-1),\cdots,C_{x_{ik}}(t-1))$

其中，$x_{i1}\cdots x_{im}$ 表示在本次迭代中已经更新过标签的节点；$x_{i(m+1)}(t-1)\cdots x_{ik}(t-1)$ 表示在此次迭代中标签还没有更新的节点。f 为返回当前邻节点中出现次数最多的标签。若次数最多的标签不止一个，按定义 3 的公式计算节点与社团的相似度，选择相似性最高的社团具有的标签作为节点 x 的标签。

4)若所有节点的 *label* 都不再变化则结束，否则 $t=t+1$ 并且返回步骤 3)。

4 实验与仿真结果

为了验证算法的有效性，本文用 Matlab 进行仿真，采用几个经典的网络数据集进行测试。以下所有图中用不同颜色表示节点所属的真实社团，以节点位置表示划分结果。在测试过程中，LPA 算法和我们改进的算法都有随机更新节点的顺序。由于存在这种情况，多次运行这两种算法时会产生不同的结果，因此在每个网络上分别运行 10 次，取均值。

4.1 Zachary 空手道俱乐部网络

第一个数据集采用 Zachary 空手道俱乐部网络[12]，它是美国一所大学空手道俱乐部成员间关系的网络。Zachary Wayne 通过两年的观察，获取了 34 个成员的社会交往关系，从而得到了一个网络。在这个网络中，每个成员代表一个节点，如果两个成员在俱乐部内或在俱乐部外有社会交往关系，则这两个成员对应的顶点之间有一条边相连。由于领导人的矛盾，俱乐部分裂为以校长和主管为中心的两个小俱乐部。本算法划分结果如图 1 所示。

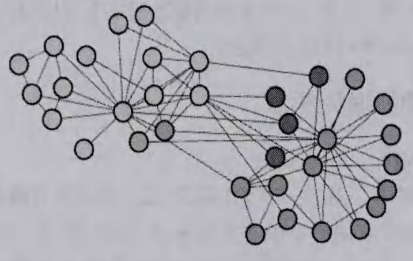

图 1 空手道俱乐部网络社团划分

在实验室过程中，我们能够得到两个核心图。通过这个两个核心图进行标签传播，可以得到图 1 的划分效果。结果表明所有节点都已正确划分。而在原始标签传播算法的结果中，该数据集在实验中多次被分为 3 个社团，并且不同实验之间的结果差别较大。

4.2 海豚社会关系的网络

第二个数据集是关于生活在新西兰神奇湾的 62 只海豚的社会关系网络[13]，网络中节点为海豚，节点之间的边为海豚之间的社会关系，由于一只关键海豚的离开，它们将自动划分为两个较小的族群。图 2 是用我们的算法对该网络进行划分的结果。

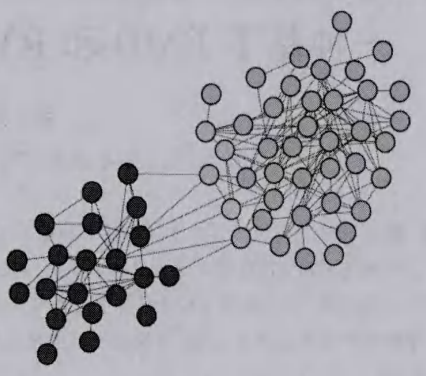

图 2 海豚社会网络

海豚社会关系的网络的实验中，我们也得到两个核心社团(如图 3 所示)，并且是较大的两个核心社团，作为后面标签传播的初始值。通过图 2 与实际的社团的对比，我们可以看出这个算法所发现的两个社团与现实中的社团是一样的。

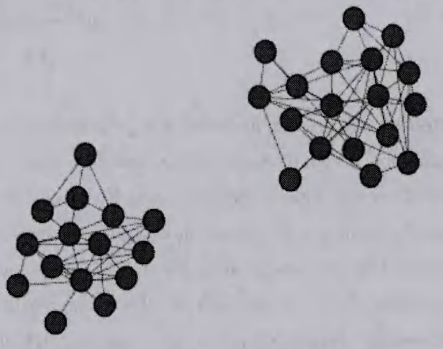

图 3 海豚社会网络所得的两个核心图

4.3 美国大学足球队比赛网络分析

分析的第 3 个网络是 2000 年秋美国大学足球队常规赛的比赛网络[14]，这个网络包含 115 个表示 115 支足球队的节点，以及 613 条表示这些足球队之间有比赛的边。这 115 个足球队来自 12 个联盟，每个小组由 8～12 支球队组成，同一联盟内的球队之间的比赛频繁，而不同联盟的球队之间赛事较少。图 4 是我们对该网络划分的结果。

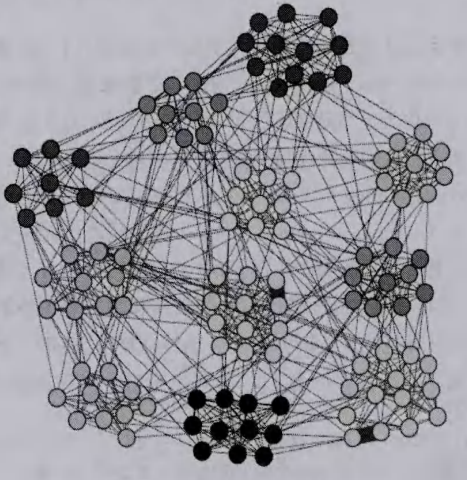

图 4 足球队比赛网络划分

(下转第 148 页)

[8] Li Xing, Fu Wen-xiu. Efficient RFID Data Cleaning Method[J]. Telkomnika Indonesian Journal of Electrical Engineering, 2013: 1707-1713

[9] Li Ling-juan, Liu Tao, Rong Xiang, et al. An Improved RFID Data Cleaning Algorithm Based on Sliding Window[C]// IOT (Internet of Things) Workshop. 2012: 262-268

[10] 王妍,石鑫,宋宝燕. 基于伪事件的 RFID 数据清洗方法[J]. 计算机研究与发展,2009,46(z2):270-274

[11] 谷峪,于戈,张天成. RFID 复杂事件处理技术[J]. 计算机科学与探索,2007,1(3):255-267

(上接第 121 页)

在实验中,一共发现了 11 个分组。与实际的 12 个分组有点区别,这是因为第 6 组中的队伍都是独立的队伍,它们并没有相互比赛,因此在本算法中,这一组被归类到与其比赛次数最多的分组中,因而出现了与实际分组不同的情况。由此我们可以看到本算法在橄榄球联赛数据集上的社团发现结果也是非常符合现实的,同时不会出现多次实验结果不同的现象,非常稳定。

本文针对两种算法在足球队比赛网络上运行 10 次的时间情况绘制了对比折线图,如图 5 所示。为了验证算法对一个随机网络的有效性,我们选了一个有 2000 个节点的网络进行测试,具体运行时间对比如图 6 所示。算法的准确性如表 1 所列。从中可以看出,我们的算法与 LPA 算法相比在时间上没有太多的增加,且提高了准确率。

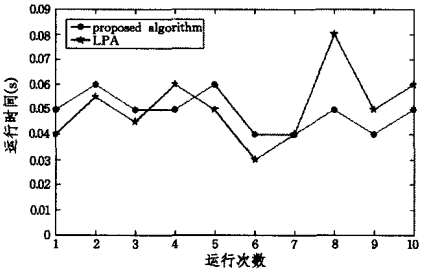

图 5 足球队比赛网络的运行时间对比

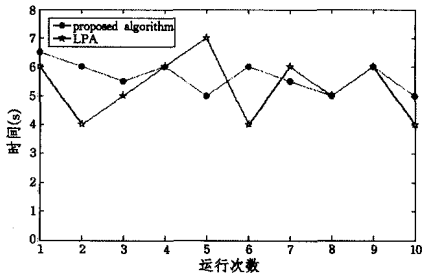

图 6 随机网络的运行时间对比

表 1 算法的准确率比较

具体的网络	LPA 算法	改进的 LPA 算法
空手道俱乐部网络	80.4%	97.2%
海豚社会网络	85.6%	96.3%
足球队比赛网络	86.1%	93.2%
随机网络	88.2%	91.3%

结束语 本文对 LPA 算法的初始条件进行了改进,提出了基于核心图的标签传播算法,即利用节点的 k 阶共同邻居,找出网络中的初始核心社团。用这些初始核心社团作为标签传播的初始状态,不仅可以显著提升算法的稳定性,在许多情况下还能够提高算法所发现的社区结构的质量。由于对网络进行初始化,算法会花费一些时间,改进后的 LPA 算法在时间上并没有太多的增加,而且在提高了准确率的前提下,增加运算时间是可以接受的。

参 考 文 献

[1] Zhao Y P, Levina E, Zhu J. Community extraction for social networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2011, 108(18): 7321-7326

[2] Kelley S, Goldberg M, Magdon-Ismail M, et al. Defining and discovering communities in social networks[J]. Handbook of Optimization in Complex Networks, 2012, 57(2): 139-168

[3] Angeles S M, Boguna M, Sagues F. Uncovering the hidden geometry behind metabolic networks[J]. Molecular BioSystems, 2012, 8(3): 843-850

[4] Ino H, Kudo M, Nakamura A. Partitioning of Web graphs by community topology[C]// Proceedings of the 14th International Conference on World Wide Web. New York: ACM, 2005: 661-669

[5] Farutin V, Robison K, Lightcap E, et al. Edge-count probabilities for the identication of local protein communities and their organization[J]. Proteins: Structure, Function, and Bioinformatics, 2006, 62(3): 800-818

[6] Newman M E J. Modularity and communities structure in networks [J]. Proceedings of the National Academy of Science, 2006, 103(23): 8577-8582

[7] Guimera R, Amaral L. Functional cartography of complex metabolic networks[J]. Nature, 2005, 433(7028): 895-900

[8] Flake G W, Lawrence S, Giles C L, et al. Self-organization and identification of Web communities[J]. IEEE Computer, 2002, 35(3): 66-71

[9] Kleinberg J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM, 1999, 46(5): 604-632

[10] Palla G, Derenyi I, Farkas I, et al. Uncovering the overlapping community structures of complex networks in nature and society [J]. Nature, 2005, 435(7043): 814-818

[11] Raghavan U N, Albert P, Kumara S. Near linear time algorithm to detect community structure in larger-scale networks[J]. Physical review E, 2007, 76(3): 036106

[12] Zachary W W. An information flow model for conflict and fission in small groups[J]. Journal of Anthropological Research, 1977, 33(4): 452-473

[13] Lusseau D. The emergent properties of a dolphin social network [J]. Proceedings of the Royal Society B: Biological Sciences, 2003, 270(S2): 186-188

[14] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proc. Natl. Acad. Sci., 2002, 99: 7821-7826