ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 2,Mar. 2021,pp. 314－323
DOI:10. 16337/j. 1004-9037. 2021. 02. 013

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于SimRank全局矩阵平滑收敛的网络社区发现

李维勇[1]，孔　枫[1]，张　伟[2]，陈云芳[2]

(1. 南京信息职业技术学院网络与通信学院,南京 210023;2. 南京邮电大学计算机学院,南京 210023)

摘　要：SimRank方法是一种基于图的拓扑结构信息来衡量任意两个对象间相似程度的方法,针对在真实的大规模社交网络中节点与节点之间的迭代计算过程需要消耗大量的时间,提出了一种基于SimRank全局矩阵平滑收敛的网络社区发现方法(SimRank global smooth convergence,SGSC)。首先,该算法通过经典度量来识别网络中的初始核心节点;然后利用矩阵平滑收敛来计算SimRank得到最终核心节点;最后,基于全局收敛矩阵,将社区聚集在核心节点周围,使用Closeness指数合并两个社区,通过递归的重复该过程,聚类出最终社区。在3种真实的不同规模的社交网络中将SGSC和其他2种具有代表性的方法进行比较,并验证了提出的算法在不同规模的社交网络中社区划分的准确率和算法运行的时间性能上有所提升。

关键词：社区发现;SimRank;矩阵迭代;聚类
中图分类号：TP391　　**文献标志码**:A

Hierarchical Community Detection Based on Global Smooth Convergence Using SimRank

LI Weiyong[1], KONG Feng[1], ZHANG Wei[2], CHEN Yunfang[2]

(1. School of Network and Communication, Nanjing Vocational College of Information Technology, Nanjing 210023, China;
2. School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: SimRank is a method based on the topological structure information of the graph to measure the similarity between any two objects. However, in real large-scale social networks, the iterative computation between nodes is time-consuming. Here we propose a hierarchical community detection algorithm based on global matrix smooth convergence using SimRank, called SGSC. First, the SGSC algorithm identifies the initial core nodes in a network by classical measurement.Then, it smoothly converges a matrix to calculate SimRank to obtain original core nodes. Based on the global convergence matrix, we cluster the communities around the core nodes and use a closeness index to merge two communities. By recursively repeating the process, a dendrogram of the communities is eventually constructed. We validate the performance of SGSC by comparing its results with those of two representative methods for three real-world networks with different scales, and comparison results show that the proposed SGSC algorithm improves the accuracy in community division and reduces running time in social networks of different scales.

基金项目:国家自然科学基金(61672297)资助项目;2019年中国特色高水平高职学校和专业建设计划(教职成函〔2019〕14号)资助项目;2019年度高校“青蓝工程”优秀教学团队(苏教师[2019]3号)资助项目。
收稿日期:2020-12-10;**修订日期**:2021-02-15

Key words: community detection; SimRank; matrix iteration; clustering

引 言

近年来,社区发现成为复杂网络分析的重要任务,社区发现方法在社交网络中受到了极大的关注。社区是社交网络中的一种普遍现象,在社区中的许多成员倾向于形成紧密联系的群体。在不同的情况下,这些群体可以称为社区、集群、内聚子团或模块。社区发现是通过分析网络拓扑结构和网络节点的属性来研究出的某种集群的结构特征从而组成的一种网络节点集合,在该集群中内部节点连接紧密,而外部节点连接稀疏。显然,该集群的内部节点之间的交流比外部节点更加频繁。

如何在复杂网络中找到社区结构已经成为许多领域的热门话题,包括社会学、生物信息学和物理学等。属于同一个社区中的节点具有更大的可能性有着相同或相似的属性:例如,在同一个社交网络中的群体[1]更可能具有共同的兴趣爱好或者背景;在万维网中[2-4]形成的社区结构可能具有共同的主题和相关的页面;在细胞和遗传相关的生物或神经网络中[5-6],社区结构的形成可能暗示细胞存在相似特征。这些网络集合可以帮助简化整个网络的功能分析。

由于社区发现研究具有很重要的意义,目前已有许多社区发现方法被提出。现有的社区发现方法大致分为传统方法、分裂方法、基于模块的方法、谱方法和动态方法等[7]。分裂算法的思想是检测连接不同社区的顶点的边并将其删除,使集群间彼此断开,其中最受欢迎的算法是Girvan和Newman提出的[8-9],他们定义了中介性,并引入了模块度作为网络结构的后继度量,这对于社区发现具有非常重要的意义,并在许多应用场合中获得了成功。模块度是衡量社区结构最著名的质量函数,基于模块度又使用了不同的聚类方法,例如贪心算法[10]、模拟退火方法[11]、外部优化方法[12]、谱优化方法[13]以最大化模块度[14]。一些研究人员还提出了改进的模块度测量方法,例如扩展到有向图,包括有向图的正边和负边概念。谱方法是根据图的特征矩阵、特征向量、特征值来发现社区。动态方法则采用图上的运动过程,例如旋转-旋转交互,随机行走和同步。

最近由社区发现算法发展起来的图划分开辟了网络分析的新领域。此外,在社区发现算法的研究中,SimRank方法被用来推断网络的结构特性,该方法之所以非常适合社区聚类的原因很明显:节点之间会以更高的概率来吸引和它们具有相同性质的节点。

Ganjaliyev[15]提出了一种通过聚类来识别网络中社区的方法。在这种方法中,存在一组约束条件,即每个数据对象正好分配给一个集群,每个集群至少包含一个对象[16]。选择一定数量的集群,并使所选集群的总权重最大化而使集群之间的相似性最小化。Choudhury等[17]重点研究了一种新的Newman-Girvan算法来检测社交网络中的社区和子社区。该方法已在Zachary空手道俱乐部、Bottlenose海豚网络和大学足球网络等真实的社交网络中展开了实施。Lee和Cunningham[18]描述了在评估大型数据集和小型手工制数据集方面的差异,在较小数据集上运行良好的算法在大型数据集上可能运行的效果较差,因此,他们引入了一个框架,在该框架中,社交网络数据集使用元数据进行注释,该框架基于机器学习:假设如果社区发现算法运行良好,则分类器应该能够使用发现的社区集合来推断与社区结构密切相关的节点属性的缺失值,该节点属性有两个缺陷:不完整和嵌套,他们还解释了挖掘出的数据如何遭受目的收集的数据的影响。Muslim[19]讨论了4种使用节点的结构和属性相似性的社区结构检测方案,每个方案提供了不同的输出,可以将其组合然后在社交网络中找到社区。第1种方案使用节点之间的结构相似性;第2种方案利用节点之间的属性相似性;在第3种方案中利用了节点之间的结构相似性和属性相似性;在第4种方案中仅考虑使用节点之间的属性相似性。Wang等[20]提出了一种基于动态内容的网络社区检测方法NEI-Walk,该篇文章中提出了一种基于内容的网络转换为节点-边缘交互

(Node edge interaction, NEI)网络的方法,其中节点内容、边缘内容和链接结构被无缝地嵌入。首先,基于内容的网络被转换为NEI网络,即一种多模式网络,由2种类型的节点和3种类型的边组成,分别称为n节点和e节点,3种类型的边缘分别表示节点内容相似性,边内容相似性和结构相似性。随着基于内容的网络的发展,提出了一种基于差异活动的方法来逐步维护NEI网络,通过在NEI网络中应用异构随机游走发现潜在社区。Blondel等[21]提出了一种方法,它是一种基于模块化优化的启发式方法。以每个开始节点作为社区,并基于模块度作为来优化应合并的社区标准,在一组节点上重复此操作几次,直到无法进行进一步优化为止。然后在社区图上整体重复该过程。Raghavan等[22]提出了一种基于标签传播的简单方法,最初,图中的每个节点都使用唯一的标签进行初始化,并且在方法的每个步骤中,每个节点都采用其大多数邻居当前具有的标签,无法更改标签时,迭代过程收敛。对于具有相同标签的每组节点形成一个社区。

虽然已有许多不同的社区发现方法被提出,但还有一些尚未解决的问题:当进行大规模的网络分析时,大多数算法效率低下且时间复杂度高。自1998年谷歌网页排名算法PageRank提出以来,相继有SimRank[23]、SimFusion[24]、Penetrating-Rank[25]等基于图拓扑结构的相似度计算模型被提出,其中SimRank被认为是一种比较流行的基于有向图拓扑结构的计算图节点相似度的模型。它的主要思想为:如果两个对象(节点)被相似的对象引用(即有向图中不同节点的入边邻节点相似或相同),那么这两个节点也相似。近年来,基于SimRank的方法被用到了社区发现中,使用SimRank方法计算两个节点之间的相似度,可以计算出具有相同属性信息的节点属于同一个社区的可能性更大,从而可以逐步确定出一个社区结构。

然而面对大规模网络,使用传统的节点之间计算相似度的方法计算消耗的时间过长;但是,对于具有幂律分布[26-27]的网络来说,核心节点起着非常重要的作用。因此,本文提出了一种基于SimRank的层次社区发现算法,该方法有以下3方面创新:

(1) 利用自然的幂律分布和网络中的社区结构,本文用核心节点测量网络中的所有节点迭代的稳定性;

(2) 使用SimRank函数计算节点与节点之间的相似度,通过迭代计算减少的计算复杂度;

(3) 迭代计算后的节点相似度将社区聚集在核心节点周围,根据社区质量评估指标,重复执行社区的合并,以获得合理的社区。

本文首先介绍了社区发现方法研究的相关工作;然后介绍相关的理论基础和概念。进而描述了基于SimRank的全局矩阵平滑收敛的社区发现,并展示实验分析结果。最后,得出结论并建议未来研究方向。

1 SimRank的定义

网络图一般使用$G=(V,E)$来表示,其中$V=\{v_1,\cdots,v_n\}$和$E=\{e_1,\cdots,e_n\}$分别表示节点集合和边的集合,图G的邻接矩阵为$\boldsymbol{A}$,其中$a_{ij}=1$表示节点i和节点j之间存在连接的边,$a_{ij}=0$表示不存在连接的边。在这里图G被认为是无向网络图,所以邻接矩阵$\boldsymbol{A}$是对称矩阵。$d(i)=\sum_j a_{ij}$表示节点i的邻接点的个数。

SimRank是Jeh与Widom于2002年提出的通过图$G=(V,E)$的拓扑结构衡量图中任意2个节点相似度的模型[23]。SimRank计算满足以下2条规则:(1)如果两个不同对象被相似对象引用,则这两个对象相似(递归定义);(2)每个对象与其自身相似度最高(基本情况)。

定义1 SimRank定义的数学表达式为[23]

$$s(a,b)=\begin{cases}1 & a=b\\ \dfrac{c}{|\boldsymbol{I}(a)|\cdot|\boldsymbol{I}(b)|}\sum_{x\in I(a)}\sum_{y\in I(b)}s(x,y) & a\neq b\\ 0 & \boldsymbol{I}(a)=\boldsymbol{\phi}\text{或}\boldsymbol{I}(b)=\boldsymbol{\phi}\end{cases}\tag{1}$$

式中:c为取值0到1之间的阻尼系数,一般取值范围0.6~0.8[10];$\boldsymbol{I}(a)$为节点a的入边邻节点集合中元素的数量。

由于图G的邻接矩阵为$\boldsymbol{A}$,矩阵$\boldsymbol{A}$的列归一化矩阵$\boldsymbol{Q}$,则相似矩阵$\boldsymbol{S}$根据定义1可表示为

$$\boldsymbol{S}=(c\cdot\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{Q})+(1-c)\cdot\boldsymbol{I}\tag{2}$$

式中:$\boldsymbol{Q}^{\mathrm{T}}$为向后转移矩阵$\boldsymbol{Q}=\dfrac{1}{|\boldsymbol{I}(a)|}$的转置,$\boldsymbol{I}$表示单位矩阵,即将计算结果对角元素均取值为1。

式(1)可以用来引入迭代方法来计算:如果$a\neq b$,则有$s_0(a,a)=1$和$s_0(a,b)=0$,对于$k=0,1,2,\cdots$,设(i)$s_{k+1}(a,a)=1$;(ii)如果$\boldsymbol{I}(a)=\boldsymbol{\phi}$或$\boldsymbol{I}(b)=\boldsymbol{\phi}$,$s_{k+1}(a,b)=0$;(iii)否则

$$s(a,b)=\frac{c}{|\boldsymbol{I}(a)|\cdot|\boldsymbol{I}(b)|}\sum\nolimits_{x\in I(a)}\sum\nolimits_{y\in I(b)}s_k(x,y)\tag{3}$$

结果序列$\{s_k(a,b)\}_{k=0}^{\infty}$收敛到$s(a,b)$,得到式(1)的精确解。

对于式(2)可变形为

$$\boldsymbol{S}^{(k+1)}=(c\cdot\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{S}^{k}\boldsymbol{Q})+(1-c)\cdot\boldsymbol{I}\tag{4}$$

Jeh和Widom[23]首先提出:每个节点与自己相似度为1,设置初始SimRank矩阵为$\boldsymbol{S}_0=\boldsymbol{I}$;根据式(4)迭代计算,直至收敛,并且在文献[23]中证明了此算法收敛性。

2 SGSC算法

2.1 概览

在真实的社交网络交往过程中,人们的初始行为通常具有以下特征:存在着随机性,但是,随着时间的流逝,社会关系会趋于稳定状态,逐渐形成相对稳定的社会圈子,通常称这种现象为社区。受此现象观察启发,本文采用SimRank全局矩阵平滑收敛的网络社区发现方法(SimRank global smooth convergence, SGSC)算法来作为衡量社交网络节点的模型,来模拟社交网络中关系的相似性。

由于社交网络具有大量节点,通常时间复杂度计算会随网络规模呈指数增长。为了降低计算的复杂性,本文将从两方面来进行:(1)将社交网络视为粗粒度单位而不是细粒度的单元或节点,这意味着是可以进行社区聚类的;(2)提出了基于SimRank的矩阵收敛的概念,这是计算节点与节点的相似度所需要的;并且本文提出的方法是不需要设置初始的社区数量,因而对核心节点的初始选择不敏感。

本文算法的主要步骤如下:

步骤1 根据节点的中心度指标选择初始核心节点;

步骤2 计算SimRank函数,利用节点之间的相似度进行矩阵迭代收敛;

步骤3 对于每个非核心节点,选择离的最近的真正核心节点中的核心节点,并加入该集合,因此,围绕核心节点形成初始社区;

步骤4 重复此过程以评估形成的社区质量,并合并紧密的社区。

2.2 选择核心节点

通常网络是由一些社区组成,并且每个社区都有一个核心节点,在此基础上提出合理的假设,只要核心节点可以到达其他社区,就可以考虑一个社区的所有节点可以到达另一个社区的所有节点,这意

味着整个网络都是连接的。这种情况类似于首都和每个省份的大城市之间都有联系,所以全国各地的所有城市都是相互联系。多数情况下社交网络中的节点遵循幂律分布,因此只有极少数的节点互连着其他节点,称为核心节点。选择初始核心节点的方法至关重要,必须避免选择错误的核心节点并不错失真实的核心节点。

利用传统的衡量社交网络中节点重要性的指标有以下几种方法:度中心性、紧密度中心性、中介中心性和特征向量中心性。在本文中,采用度中心性作为最基本的指标来识别网络中的社区。这样做有2个原因:(1)计算度中心性比较简单;(2)度中心性和迭代过程中使用的矩阵是一致的。

注意到度中心性不仅反映了每个节点和其他节点的相关性,并且也与和网络规模相关,即网络中的节点数。随着网络大小的增加,度中心性的最大值也可能在增加,为了消除度中心性对网络规模的影响,定义节点 i 的重要性如下

$$\text{Important}(v_i)=d(i)=\frac{\sum_{j}^{n}a_{ij}}{n-1} \tag{5}$$

若一节点 i 满足 $\text{Important}(v_i)\geqslant\tau$,通过式(5)选择出来的节点集合称为初始核心节点,初始核心节点集合定义为

$$\text{CenterSet}=\left\{v_i\middle|\ \text{Important}(v_i)\geqslant\tau\right\}$$

式中 τ 为一预设阈值。

根据SimRank的特点,随着迭代趋于无穷大,矩阵不断趋于最终的稳定值。在迭代过程中,一个节点到另一个节点的集合在不断增加或逐渐稳定在一个较大的值,这表明目标节点更具有影响力,因此是真实的核心节点。因此,使用经典的度中心性进行度量,解决选择初始核心节点集,如果设置的阈值更高,初始核心节点集合相对小;如果将其设置为较低,则初始核心节点集合比较大。但是,此更改对确定最终的核心节点影响不大。

2.3 选择社区合并

确定迭代的步数之后,可以确保矩阵迭代的稳定性,并且可以确定核心节点距离每个最近的非核心节点,然后将非核心节点放入其最近的核心节点所在的社区中。

最初的社区围绕核心节点形成,算法1中的算法需要进一步的聚集才能形成层次结构,涉及两个步骤:(1)选择2个要合并的社区并计算它们的相似性;(2)确定聚类过程是否应停止。由于传统衡量社区划分质量的指标,例如:模块 Q,会涉及更多的计算复杂度。从这个角度来看,利用社区紧密度指标,以确定两个社区之间的相似性,定义如下

$$\text{Closeness}(\boldsymbol{C}_i,\boldsymbol{C}_j)=\frac{\dfrac{\text{Edges}_{\text{internal}}(\boldsymbol{C}_i\cup\boldsymbol{C}_j)}{\text{Edges}_{\text{external}}(\boldsymbol{C}_i\cup\boldsymbol{C}_j)}}{\dfrac{1}{2}\cdot\left(\dfrac{\text{Edges}_{\text{internal}}(\boldsymbol{C}_i)}{\text{Edges}_{\text{external}}(\boldsymbol{C}_i)}+\dfrac{\text{Edges}_{\text{internal}}(\boldsymbol{C}_j)}{\text{Edges}_{\text{external}}(\boldsymbol{C}_j)}\right)} \tag{6}$$

式中:$\text{Edges}_{\text{internal}}(\boldsymbol{C}_i\cup\boldsymbol{C}_j)$ 和 $\text{Edges}_{\text{external}}(\boldsymbol{C}_i\cup\boldsymbol{C}_j)$ 分别表示新的合并后的社区的内部边和外部边,$\text{Edges}_{\text{internal}}(\boldsymbol{C}_i)$ 和 $\text{Edges}_{\text{external}}(\boldsymbol{C}_i)$ 分别表示 $\boldsymbol{C}_i$ 社区的内部边和外部边,当 $\text{Closeness}(\boldsymbol{C}_i,\boldsymbol{C}_j)$ 大于某个阈值时,将两个社区进行合并,一般将该阈值设置为1~2之间,根据不同的网络可适当调整该阈值,算法1详细给出了SGSC算法的整个过程。

算法1 SGSC

输入:图 G,核心节点集合CenterSet,参数 α,迭代次数 k,阻尼系数 c

输出:社区划分集合

1:for k:

2: 计算$S=(c\cdot \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{Q})+(1-c)\cdot \boldsymbol{I}$直至收敛

3: end for

4: if $S>\alpha$:

5: CenterSet=$[v_i]$

6: for i in CenterSet:

7: for j in图G中除核心节点外的每个节点v_j:

8: 计算Closeness($\boldsymbol{C}_i,\boldsymbol{C}_j$) 直至无法计算

9: end for

9: if Closeness($\boldsymbol{C}_i,\boldsymbol{C}_j$)$>\omega$:

10: 将相应的社区节点进行合并

11: 返回初始社区划分集合C

12:for 社区C中的任意两个社区$\boldsymbol{C}_i,\boldsymbol{C}_j$

13: 重复执行8到10

14:返回社区划分集合

2.4 验证指标

为了评估不同算法的有效性,本文采用Newman-Girvan模块度这个指标来评估网络中社区的结构强度:对于整个网络来说,cp为通过某种算法事先得到的社交网络划的社区划分。cp的Newman-Girvan模块度定义为[8]

$$Q_{cp}=\left[a_{vw}-\frac{\boldsymbol{d}_v\cdot\boldsymbol{d}_w}{(2m)(2m)}\right]\delta(\boldsymbol{C}_v,\boldsymbol{C}_w)=\sum_{i=1}^{C}e_{ii}-a_i^2 \tag{7}$$

在本文中节点v和节点w的值$a_{vw}=1$表示有向图中的权值,如果$\delta(\boldsymbol{C}_v,\boldsymbol{C}_w)=1$表示节点$v$和节点$w$属于同一个社区,如果$\delta(\boldsymbol{C}_v,\boldsymbol{C}_w)=0$则不属于同一个社区。$\boldsymbol{C}_v$和$\boldsymbol{C}_w$分别代表节点$v$和节点$w$属于的社区。$e_{ij}=\frac{A_{vw}}{2m}$表示节点$i$的终端和另一个社区$j$相连的边缘分数。$a_i=\frac{d_i}{2m}=\sum_j e_{ij}$是连接到社区$i$中节点的边的比例。

3 实验分析

为了验证SGSC算法在不同类型的网络中的性能,本文使用了3个数据集进行分析:American College Football网络[28]、Facebook网络[29]以及欧洲Deezer社交网络[29],如表1所示。American College Football网络是具有115个节点和613条边的小组织的网络;Facebook网络包含来自Facebook的“圈子”(或“朋友列表”),它是由4 039个节点和88 234条边组成的中等规模的社交网络;Deezer用户的社交网络的节点是来自欧洲国家的Deezer用户,边缘是他们之间的相互关注者关系,它是由28 281个节点和92 752条边组成的大规模网络。通过实验,将提出的SGSC算法和GN[8]以及Newman快速算法[10]进行比较。

表1 测试数据集

Table 1 Test datasets

网络	节点数	边数
American College Football网络	115	613
Facebook网络	4 039	88 234
欧洲Deezer社交网络	28 281	92 752

本文使用的是Intel Pentium四核处理器,带有8 GB DDR3内存,Windows 10操作系统和Python3

的Numpy图分析工具和Matplotlib数据分析工具。

3.1 初始核心节点的选择

如前文所述,选择初始核心节点的方法是使用度中心性阈值的标准化公式,该值的范围为0.0～1.0,0.0表示与任何节点都没有联系(例如一个孤点),1.0表示与每一个节点都有直接联系。在社会网络中,标准化的行为人的度中心性测量行为人在诸多关系中的参与程度,得到高分的行为人是网络中最显眼的参与者。如果标准化度中心性值越接近1.0,那么行为人在关系网络中的参与度越高。在American College Football网络中使用该方法确定了初始核心节点设置为{1,0,3,2,5,6,7,15,53,67,82,88,104,43},根据不同的阈值设定,虽然选出的核心节点个数不一样,但是,经过多步骤迭代,真正的核心节点仍然是节点7、51、18和节点43。尽管未选择节点51和节点18作为初始核心迭代计算后,但是经过迭代之后,真正的核心节点仍然节点7、51、18和节点43。表2记录了选择不同的阈值对应的核心节点的个数以及最终确定的核心节点。

表2 阈值的选择对于核心节点的影响(American College Football网络)

Table 2 Effect of the threshold value on core nodes (American College Football network)

阈值选择	初始核心节点个数	最终核心节点
度中心性>0.10	14	7、51、18、43
度中心性>0.09	72	7、51、18、43
度中心性>0.08	101	7、51、18、43

对于中等规模的网络Facebook网络来说,重复同样的方式进行测试,表3记录了不同阈值的选择对于核心节点的影响,尽管初始核心节点不一样,但得到的最终核心节点始终是{8,0,58,351,688,107,726,1 022,1 375,1 394,348,1 594,1 609,1 680,351,171,2 068,2 427,2 618,2 727,352,3 226,3 303,3 334,3 405,1 821,1 825,2 087,2 088,3 486,3 559,3 573,3 582,3 626,3 639,3 645,3 687,3 697,3 703,3 712,3 804,3 926,3 955,1 827,1 830,3 985,3 996,3 998,4 000,4 001,4 002,4 015,4 018,4 024,4 028,4 031,1 858,1 902,1 915,1 973,1 978,1 831,1 993,2 001,1 843,2 031},从实验的结果来看,本文提出的SGSC算法对于最终的核心节点的确认是不会受到任何影响的。

表3 阈值的选择对于核心节点的影响(Facebook网络)

Table 3 Effect of the threshold value on core nodes (Facebook network)

阈值选择	初始核心节点个数	最终核心节点个数
度中心性>0.10	4	66
度中心性>0.05	36	66
度中心性>0.03	352	66

3.2 精确度

本文还设计了一个实验来评估算法的社区发现的能力,使用0,1,2,3,…作为American College Football网络、Facebook网络以及欧洲Deezer社交网络节点的编号,根据提出的SGSC算法,可以将该网络划分成12个社区,{3,5,10,11,40,52,72,74,81,84,98,102,107},{0,4,9,16,23,41,90,93,104},{1,25,33,37,45,89,103,105,109},{12,14,18,26,31,34,36,38,42,43,54,61,71,85,99},{46,49,53,67,73,83,88,110,114},{6,13,15,47},{44,48,57,66,75,86,91,92,97,112},{17,20,27,56,58,59,62,63,65,70,76,87,95,96,113},{24,28,69},{7,8,21,22,50,51,68,77,78,108,111},{19,29,30,35,55,79,80,82,94,101},和{2,32,39,60,64,100,106},和真实的American College Football网络社区划分相比,SGSC算法对于社区的划分个数完全一致,并且对于相应的节点属于相应的社区也是划分一致。

对于中等规模的Facebook网络和大规模的欧洲的Deezer社交网络,表4,5记录了这3种算法的社区划分结果以及模块度。通过对这2种网络的聚类出的社区模块度的比较,从表4,5中可以看出,随着

网络规模的增长,SGSC算法表现出明显的优越性,GN算法则在中等规模和大规模的网络中已经无法得出聚类的结果。由于GN算法是一种性能不太好的基础算法,因而尤其不适合处理大规模的真实社交网络。

表4 Facebook网络中社区划分结果以及模块度

Table 4 Community division results and modularity of the divided communities (Facebook network)

算法	社区个数	算法模块度
SGSC算法	15	0.485 1
GN算法	无	无
Newman快速算法	8	0.279 6

表5 Deezer社交网络中社区划分结果以及模块度

Table 5 Community division results and modularity of the divided communities (Deezer social network)

算法	社区个数	算法模块度
SGSC算法	79	0.386 5
GN算法	无	无
Newman快速算法	118	0.142 5

同时,本文还比较了3种不同的网络中社区划分的模块度值,图1记录了在3种不同的网络中社区划分的模块度值比较,可以发现GN算法在Facebook网络中已经无法聚类出社区,因而它的社区模块度为0,对于大规模Deezer社交网络的社区模块度也同样为0。因此可以看出无论是在小规模American College Football网络中还是在大规模的Deezer社交网络中,本文所提出算法划分出的社区模块度均优于GN算法和Newman快速算法。

虽然本文提出的算法最终对于社区划分的模块度的值不是非常理想,但从图2还是可明显看出,由于GN算法在中等规模的Facebook网络中和Deezer网络中已无法聚类出社区,所以GN算法运行的时间必然超过10 000 s,虽然在小规模网络中,Newman快速算法的运行时间复杂度和本文提出的SGSC算法几乎相差不大,但随着网络规模的增大,本文算法的优势逐渐体现,在大规模Deezer社交网络中,本文提出的SGSC算法明显优于Newman快速算法。

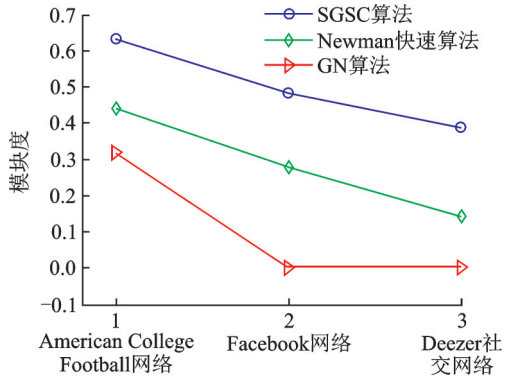

图1 不同的网络中社区划分的模块度

Fig.1 Modularity of the divided communities in different networks

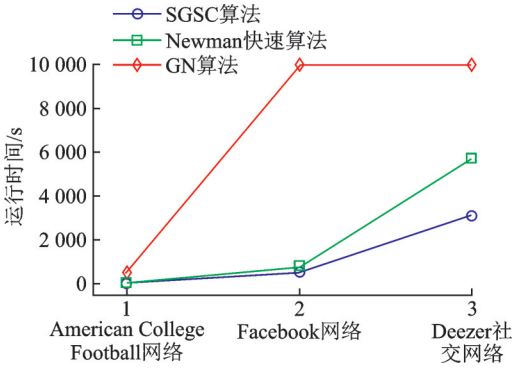

图2 不同的网络中各算法运行的时间

Fig.2 Comparisons of complexity

4 结束语

本文中提出了一种基于SimRank的全局矩阵平滑收敛的社区发现算法,利用SimRank的方法计算节点与节点之间的相似性而非传统计算节点相似度的方法,SGSC方法基于矩阵的迭代收敛聚类出所

有的初始社区,最初的社区划分利用核心节点代表全局结构信息。作为最初的社区,仅仅通过合并核心节点周围的节点,直接使用局部信息以迅速形成小型社区,无需重新计算节点属于哪个社区,从而大大改善算法的效率。今后将重点研究如何确定SimRank的迭代停止条件并结合真实的社交网络的特征信息以更好地进行社区聚类。

参考文献:

[1] ALZAHRANI T, HORADAM K J. Community detection in bipartite networks: Algorithms and case studies[M]//Complex systems and networks. Berlin: [s.n.], 2016: 25-50.

[2] FU J, ZHANG W, WU J. Identification of leader and self-organizing communities in complex networks[J]. Scientific Reports, 2017, 7(1): 704.

[3] 张波,向阳,黄震华.一种社交网络中的个体间推荐信任度计算方法[J].南京航空航天大学学报,2013,45(4): 563-569.
ZHANG Bo, XIANG Yang, HUANG Zhenhua. Recommended trust computation method between individuals in social network site[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2013, 45(4): 563-569.

[4] 张伟,祁德昊,陈云芳.大规模网络中基于LDA模型的重叠社区发现[J].南京邮电大学学报(自然科学版),2018,38(3): 54-64.
ZHANG Wei, QI Dehao, CHEN Yunfang. Overlapping community detection algorithm based on LDA in large scale networks [J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 2018, 38(3): 54-64.

[5] KONG P, HUANG G, LIU W. Identification of protein complexes and functional modules in E. coli PPI networks[J]. BMC Microbiology, 2020, 20(1): 1-9.

[6] 赵卫绩, 张凤斌, 刘井莲. 一种基于加权共同邻居相似度的局部社区发现算法[J]. 南京大学学报(自然科学版), 2018, 54(4): 93-99.
ZHAO Weji, ZHANG Fengbin, LIU Jinlian. A novel local community detection algorithm based on common neighbors similarity measurement with weighted neighbor nodes[J]. Journal of Nanjing University (Natural Science Edition), 2018, 54(4): 93-99.

[7] DI I M, GAMBOSI G, ROSSI G, et al. Min-max communities in graphs: Complexity and computational properties[J]. Theoretical Computer Science, 2016, 613: 94-114.

[8] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826.

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

[10] CEBO D. The development and expansion of the research fronts in HIV/AIDS research: Fast algorithm for detecting community structure in networks[J]. Current Topics in Microbiology and Immunology, 2019, 15: 47-53.

[11] HE J, CHEN D, SUN C. A fast simulated annealing strategy for community detection in complex networks[C]//Proceedings of 2016 2nd IEEE International Conference on Computer and Communications (ICCC). [S.l.]: IEEE, 2016: 2380-2384.

[12] DUCH J, ARENAS A. Community detection in complex networks using extremal optimization[J]. Physical Review E, 2005, 72(2): 027104.

[13] LESKOVEC J, LANG K J, MAHONEY M. Empirical comparison of algorithms for network community detection[C]//Proceedings of the 19th International Conference on World Wide Web. New York: ACM, 2010: 631-640.

[14] NEWMAN M E J. Equivalence between modularity optimization and maximum likelihood methods for community detection[J]. Physical Review E, 2016, 94(5): 052315.

[15] GANJALIYEV F. New method for community detection in social networks extracted from the Web[C]//Proceedings of 2012 IV International Conference "Problems of Cybernetics and Informatics"(PCI). [S.l.]: IEEE, 2012: 1-2.

[16] KHATOON M, BANU W A. A survey on community detection methods in social networks[J]. International Journal of Education and Management Engineering, 2015, 5(1): 8.

[17] CHOUDHURY D, BHATTACHARJEE S, DAS A. An empirical study of community and sub-community detection in social networks applying Newman-Girvan algorithm[C]//Proceedings of 2013 1st International Conference on Emerging Trends and Applications in Computer Science. [S.l.]: IEEE, 2013: 74-77.

[18] LEE C, CUNNINGHAM P. Community detection: Effective evaluation on large social networks[J]. Journal of Complex Networks, 2014, 2(1): 19-37.

[19] MUSLIM N. A combination approach to community detection in social networks by utilizing structural and attribute data[J]. Social Networking, 2016, 5(1): 11-15.

[20] WANG C D, LAI J H, YU P S. NEIWalk: Community discovery in dynamic content-based networks[J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(7): 1734-1748.

[21] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10): P10008.

[22] 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.

[23] JEH G, WIDOM J. SimRank: A measure of structural-context similarity[C]//Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2002: 538-543.

[24] XI W, FOX E A, FAN W, et al. Sim-fusion: Measuring similarity using unified relationship matrix[C]//Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2005: 130-137.

[25] ZHAO P, HAN J, SUN Y. P-rank: A comprehensive structural similarity measure over information networks[C]//Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York: ACM, 2009: 553-562.

[26] ADAMIC L A, HUBERMAN B A, BARABÁSI A L, et al. Power-law distribution of the world wide web[J]. Science, 2000, 287(5461): 2115.

[27] FALOUTSOS M, FALOUTSOS P, FALOUTSOS C. On power-law relationships of the internet topology[J]. ACM SIGCOMM Computer Communication Review, 1999, 29(4): 251-262.

[28] NEWMAN M. American college football[EB/OL]. (2002-01-01) [2013-4-19]. http://www-personal. umich. edu/~mejn/netdata/.

[29] LESKOVEC J. Social circles: Facebook [EB/OL]. (2012-01-01) [2012-12-3]. http://snap.stanford.edu/data/index.html.

作者简介:

李维勇(1976-),**通信作者**,男,副教授,研究方向:大数据分析、社会网络分析,E-mail:liwy@njcit.cn。

孔枫(1992-),女,硕士研究生,研究方向:社会网络、网络社区检测。

张伟(1973-),男,博士,教授,研究方向:社会网络分析、隐私保护、恶意代码分析。

陈云芳(1976-),男,博士,副教授,研究方向:网络安全、社会网络、大数据分析。

(编辑:张彤)