软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(11):3481−3491 [doi: 10.13328/j.cnki.jos.005818]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种基于模糊相似关系的局部社区发现方法*

刘井莲[1,2], 王大玲[1], 冯 时[1], 张一飞[1]

[1](东北大学 计算机科学与工程学院,辽宁 沈阳 110169)

[2](绥化学院 信息工程学院,黑龙江 绥化 152061)

通讯作者: 王大玲, E-mail: wangdaling@cse.neu.edu.cn

摘 要: 近几年,在线社交媒体发展飞速,出现了大规模社会网络.传统的基于网络全局结构的社区发现方法难以有效地处理这些大网络.局部社区发现作为一种无需知道网络的全局结构、仅通过分析给定节点的周围节点之间的关系即可找出给定节点所在社区的方法,在社会网络大数据分析中具有重要的应用意义.针对真实世界网络结构中个体间的相似关系是模糊的或不确定性的,提出了一种基于模糊相似关系的局部社区发现方法.首先,采用模糊关系来描述两个节点之间的相似关系,以节点对的相似度作为该模糊关系的隶属函数;然后证明了该关系是一种模糊相似关系,将局部社区定义为给定节点关于模糊相似关系的等价类,进而采用最大连通子图算法求得给定节点所在的社区.分别在仿真网络和真实网络上进行了实验,实验结果表明,该算法能够有效地揭示出给定节点所在的局部社区,相比其他算法,具有更高的 *F-score*.

关键词: 社会媒体网络;局部社区发现;模糊相似关系;社区结构

中图法分类号: TP18

中文引用格式: 刘井莲,王大玲,冯时,张一飞.一种基于模糊相似关系的局部社区发现方法.软件学报,2020,31(11):3481−3491. http://www.jos.org.cn/1000-9825/5818.htm

英文引用格式: Liu JL, Wang DL, Feng S, Zhang YF. Local community discovery approach based on fuzzy similarity relation. Ruan Jian Xue Bao/Journal of Software, 2020,31(11):3481−3491 (in Chinese). http://www.jos.org.cn/1000-9825/5818.htm

Local Community Discovery Approach Based on Fuzzy Similarity Relation

LIU Jing-Lian[1,2], WANG Da-Ling[1], FENG Shi[1], ZHANG Yi-Fei[1]

[1](School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China)

[2](School of Information Engineering, Suihua University, Suihua 152061, China)

Abstract: Online social media has developed rapidly in recent years, and many massive social networks have emerged. Traditional community detection methods are difficult to deal with these massive networks effectively for requiring knowledge of the entire network. Local community detection can find out the community of a given node through the connection relationship between the nodes around the given node without knowledge of the entire network structure, so it is of great significance in social media mining. For the relations between pairs of nodes in real-world networks are fuzzy or uncertain, the similarity relationship between two nodes with fuzzy relation is firstly described, and similarity between nodes as membership function of the fuzzy relation is defined. Then, it is proved that the fuzzy relation is a fuzzy similarity relation, and local community is defined as the equivalence class of the given node about fuzzy similarity relation. Moreover, local community of the given node is discovered by adopting maximal connected subgraph approach. The proposed

* 基金项目: 国家重点研发计划(2018YFB1004700); 国家自然科学基金(61772122, 61872074, 61602103, U1435216); 黑龙江省属高校基本科研业务费项目(KYYWF10236180104)

Foundation item: National Key Research and Development Program of China (2018YFB1004700); National Natural Science Foundation of China (61772122, 61872074, 61602103, U1435216); Fundamental Research Funds for the Provincial University of Heilongjiang Province (KYYWF10236180104)

收稿时间: 2018-05-23; 修改时间: 2018-10-17; 采用时间: 2019-01-07

algorithm is evaluated on both synthetic and real-world networks. The experimental results demonstrate that the proposed algorithm is highly effective at finding local community of the given node, and achieves higher *F-score* than other related algorithms.

Key words: social media network; local community discovery; fuzzy similarity relation; community structure

近年来,以 Facebook、Twitter、新浪微博、微信为代表的在线社交媒体发展飞速.这些社交媒体有上亿的注册用户,创造了海量的内容,用户与用户之间、用户与媒体内容之间、媒体内容与内容之间形成了关系复杂的社会网络[1].这些网络常呈现“同一组内节点连接紧密,不同组间节点连接稀疏”的社区结构特性[2,3].社区结构刻画了网络中连边关系的局部聚集特性.社区发现类似于图划分、聚类的概念,对这个问题的认识,程学旗等人澄清了它们三者之间的区别与联系,并给出了社区发现的问题界定[4].社区划分与图划分的处理对象都是网络,但图划分的目标是把网络中的节点划分成给定个数、大小相同的节点集合,要求划分所切断的边最小,而社区发现是寻找网络中固有的自然划分,而不是按指定条件划分.社区发现和聚类在处理方式上相似,但二者处理的对象不同,聚类所处理的是可以表示成高维向量的属性数据,而社区发现所处理的是表示成网络的关系数据.社区发现的研究具有重要理论价值和实际应用意义,在过去的十几年里受到国内外广大学者的关注.例如,著名学者 Fortunato、刘大有课题组、程学旗课题组等都对社区发现算法研究做了大量工作[4−6].然而,大多学者将网络中个体与群体间相似性关系采用确定性度量,这样可能导致不合理的社区划分.真实世界网络结构中个体间的相似关系可能是模糊的,个体对于社区的隶属关系也可能是不确定的.比如,社交网络中的好友关系就是一种模糊关系,不同用户之间的友好度是不相同的.针对这一特点,产生了一些基于模糊或不确定关系模型的社区发现方法.张泽华等人利用粗糙上下近似算法来研究社区局部扩展过程,提出了一种基于邻域粗糙化的启发式重叠社区扩张方法[7];李刘强等人提出了一种基于模糊层次聚类的重叠社区检测算法[8];宋俐等人提出了一种基于模糊聚类的社区发现算法[9];张云雷等人用社区的下近似集和上近似集来刻画社区的模糊区域,提出了基于粗糙集理论的社区发现方法[10];Wang 等人针对网络结构的复杂性和群体划分的不确定性,采用模糊方法来处理网络节点间的相似性,实现网络社区结构的模糊划分[11].Sun 等人使用模糊关系的操作取代图的遍历搜索来识别社区结构[12].Liu 等人提出一种基于模糊聚类的 K 均值算法来检测复杂网络中的社区结构[13].

以上的社区发现算法都是基于网络的全局结构.但随着大规模的 Web 网络和社会网络的出现,这些网络由于规模太大而难以获取其全局结构,或进行全局计算复杂度过高,传统的全局社区发现方法无法有效处理这些大网络.同时也产生了一些应用,例如基于用户偏好的推荐,需要挖掘某个用户所在的局部社区[14].局部社区发现作为一种不需要知道网络的整体结构、仅通过分析给定节点或节点集的周围节点之间的关系就能够找出给定节点所在社区的方法,近年来成为社会网络数据分析中的一个新的热点问题.相比全局社区发现,局部社区发现方法相关的研究较少.基于此,本文针对真实世界网络结构中个体间相似关系的模糊或不确定性,提出一种基于模糊相似关系的局部社区发现算法,将某一节点所在的局部社区转化为该节点关于模糊相似关系 q 水平上的等价类,通过寻找给定节点所在的模糊等价类进行局部社区发现.首先,采用模糊关系来描述两个节点之间的相似关系,用节点对的相似度作为该模糊关系的隶属函数;然后证明了该关系是一种模糊相似关系,将局部社区定义为给定节点关于模糊相似关系的等价类,进而采用最大连通子图算法求得给定节点所在的社区.

综上,本文的主要贡献如下:(1) 提出了采用模糊相似关系来描述网络中两个节点之间的相似关系,并以节点对的相似度作为该模糊关系的隶属函数;(2) 将局部社区定义为给定节点关于模糊相似关系的等价类,给出了局部社区的一种新定义.基于该定义,提出了一种基于最大连通子图的局部社区发现算法;(3) 分别在仿真网络数据集和真实网络数据集上与相关算法进行了对比分析,实验结果表明,本文算法能够有效地揭示出给定节点所在的局部社区,与其他算法相比,具有更高的准确性.

本文第 1 节介绍局部社区发现的问题定义以及相关算法.第 2 节详细介绍本文算法,并通过实例说明本文算法的计算过程.第 3 节介绍在仿真网络数据集和真实网络数据集上的实验结果.第 4 节对本文工作进行总结.

1 相关工作

局部社区发现是复杂网络社区发现问题的子问题,得到的是单一节点所在的社区,而不是网络中的所有社区,因此不要求知道网络的整个结构.一般定义为:对于网络 $G=(V,E)$,V 为节点的集合,E 为边的集合,给定任一个节点 $v(v\in V)$,找出节点 v 在网络中所属的社区 D[15].图 1 给出了一个网络的局部社区示意.

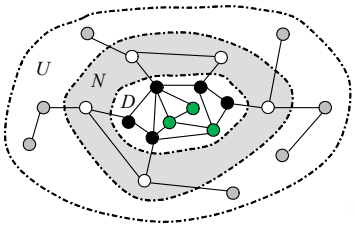


Fig.1 Division of a network into local community D (black nodes and green nodes), D's shell nodes set N (white nodes) and unknown nodes set U (grey nodes)

图 1 网络分为局部社区 D(黑色节点和绿色节点)、D 的外壳节点集合 N(白色节点)以及未知节点集合 U(灰色节点)

针对图 1 所示的网络图,在局部社区发现过程中,网络 G 中的节点主要分为 3 部分:局部社区 D、外壳节点集 N 和未知节点集 U.局部社区 D 中节点及其连接都是已知的,D 中节点可以分为核心节点集 C 和边界节点集 B,其中,C 中节点的邻居节点全都在 D 中,而 B 中节点至少有一个邻居节点不在 D 中.外壳节点集 $N=\{y|y\notin D,(x,y)\in E,x\in D\}$,通过寻找 D 中节点的邻居节点可以找出 N.未知节点集 $U=G-D-N$,在局部社区发现过程中,假设 U 中节点是未知的,即不利用 U 中节点及其连接等信息进行局部社区发现.

Clauset 首次提出局部社区发现问题,并定义了一个社区的局部模块度度量指标 R[16].R 为边界节点集 B 内节点分别与社区 D 和外壳节点集 N 形成边数的比值.计算 R 不需要使用网络的全局信息,该算法通过最大化社区的局部模块度度量指标 R 选取邻近节点的原则构建局部社区.Luo 等人从子图的内度与外度的视角,定义了另一个社区的局部模块度度量指标 M[17],M 为 D 内节点分别与 D 内节点和 N 内节点形成边数的比值.通过迭代选择能使 M 增大的 N 内节点,或者删除能使 M 增大的 D 内的节点的原则,构建局部社区.

以上 R 和 M 这两个指标只考虑与局部社区 D 相关的边的数目,而忽略了不同边的两个节点之间的相似程度是不一样的.Huang 等人首先提出了基于节点间相似度的局部社区发现算法 LTE[18],用社区 D 内形成边的节点对之间的相似度之和与社区 D 内节点和外壳节点集 N 形成边的节点对之间的相似度之和的比值来度量社区的质量,从 N 中选取与社区 D 内节点相似度最大的节点的原则构建局部社区.不同于 LTE 算法,只考虑节点自身及其邻居来度量两个节点间的相似性,Ma 等人还考虑了不直接相连的 d 层邻居($d\geqslant 1$),提出了基于 d 层邻居的局部社区发现算法 GMAC[19].Liu 等人采用节点向量模型来表示网络中的节点,在考虑两个节点的共同邻居的同时,还将邻居节点与它们各自的相似度纳入计算,实现了一种新的局部社区发现算法[20].赵卫绩等人提出了一种加权邻居节点的共同邻居相似度 CNWNN 指标,通过优先考虑与当前局部社区嵌入度最大的节点,逐步找到给定节点所在的社区[21].此外,Panagiotakis 等人提出了基于流传播的局部社区发现算法 FlowPro[22],通过在起始节点发射一定数量的流在网络中传播,流经的节点存储接收到流的 1/2,将另外 1/2 继续传播给自己的邻居节点,最后,以每个节点保存的流的多少来度量与起始节点连接的紧密度,选取流多的节点作为起始节点所在的局部社区.

基于社区质量度量优化类方法,都是基于判断一个节点是不是应该属于当前社区.与之不同的是,本文从关系的视角定义社区,认为社区是具有一定强度的关系构成的.为此,本文采用模糊关系来描述两个节点之间的相似关系,然后证明了该关系是一种模糊相似关系,将局部社区定义为给定节点关于该模糊相似关系 q 水平的等价类,进而采用最大连通子图算法求得给定节点所在的局部社区.

2 基于模糊相似关系的局部社区发现算法

2.1 预备定义

本小节先给出网络、模糊关系 R 等定义,然后给出模糊相似关系视角下的局部社区发现问题的定义.

定义 1(网络[18]). 用图 $G=(V,E)$来表示网络,其中,V 是节点集合,E 是边集合,$|V|$表示节点集合 V 中节点的个数.对于节点 $v(v\in V)$,$\Gamma(v)$表示节点 v 的邻居节点的集合.

$(x,y)\in E$,表示节点 x 和 y 之间存在边;$(x,y)\notin E$,表示节点 x 和 y 之间没有边.x 和 y 之间是否有边相连,这是明确的,可以用“1”或者“0”来刻画.在边的基础上,我们定义一个关系 R,用来表示一条边的两个节点的相似关系.显然,两个节点的相似关系,用“1”或者“0”来刻画是不合适的.为此,我们引入模糊关系,用节点间的相似度作为隶属函数来刻画两个节点的相似关系.

定义 2(模糊关系 R). 设 R 是 V 上的模糊关系,对于任意的节点对$(x,y)\in(V\times V)$,隶属度 $R(x,y)$定义为

$$R(x,y)=\begin{cases}\dfrac{|\Gamma(x)\cap\Gamma(y)|}{\min(|\Gamma(x)|,|\Gamma(y)|)}, & (x,y)\in E\\ 0, & (x,y)\notin E\\ 1, & x=y\end{cases} \tag{1}$$

采用度小的节点的邻居节点集嵌入在度大的节点的邻居节点集中的程度,作为两个节点在同一个社区的隶属度 $R(x,y)$.从公式(1)可得,$R(x,x)=1$,$R(x,y)=R(y,x)$.即模糊关系 R 满足自反性、对称性,所以关系 R 是模糊相似关系.

定义 3(节点关于模糊相似关系 R 的 q 水平的相似类[23]). 节点集合 V 上的模糊相似关系 R,对任意节点 $a\in V$,集合$[a]_{Rq}=\{x|x\in V,R(a,x)\geqslant q\}$,称为元素 a 关于模糊关系 R 的 q 水平的相似类.

设 B_1,B_2 是节点集合 V 上的模糊相似关系 R 上的 q 水平上的 2 个相似类,若 $B_1\cap B_2\neq\varnothing$,则称它们是相似类.将所有与$[a]_{Rq}$ 相似的类合并成一类,就是包含节点 a 的等价类.

基于以上定义,可以得出模糊相似关系视角下的局部社区发现问题可以定义为:设 R 是 V 上的模糊关系,a 是 V 中的任意一个节点,给定阈值 q,a 所在的社区可以看作是节点 a 关于关系 R 的 q 水平上的等价类.从图论的角度,节点 a 所在的社区可以看作是包含节点 a,由隶属度大于等于 q 的边构成的最大连通子图.

2.2 算法描述

传统的模糊聚类算法需要知道网络的全局结构,通过构造模糊相似矩阵,得到网络节点的一个划分,计算量大.而在局部社区发现中,我们的目标只是寻找给定节点所在的社区,而不需要得到网络中所有节点的一个划分.因此,本文采用最大连通子图的方法来进行局部社区发现.将给定节点 v 加入到社区 D,在保证隶属度大于等于 q 的前提下,选择与 D 内节点构成隶属度最大的节点加入到社区 D.重复该过程,直到 D 内节点与 D 外节点的隶属度都小于 q 为止.此时所形成的连通子图即为从节点 v 出发由隶属度大于等于 q 的节点构成的最大连通子图.我们将该最大连通子图看作给定节点 v 的局部社区.本算法的时间复杂度仅取决于所寻找社区的大小,而不取决于整个网络的大小.

算法思路如下.

(1) 初始化:$D=\{v\}$,$N=\Gamma(v)$.

(2) 在保证隶属度大于等于 q 的前提下,从 D 中节点出发,在 N 中寻找与 D 中节点相连且隶属度最大的节点 y,将 y 加入到 D 中.

(3) 将 y 的邻居节点集中不在 D 中的节点加入到 N 中.

(4) 重复步骤(2)和步骤(3),直到 N 为空或 N 中节点与 D 中节点的隶属度都小于阈值 q 为止.

算法描述如算法 1 所示.

算法 1. 基于模糊相似关系的局部社区发现算法.

输入:网络 $G=(V,E)$,初始节点 v,模糊关系 R,阈值 q.

输出:节点 v 所在的社区 D.

方法:

```
1)  initialize D={v},N=Γ(v);
2)  while (N!=NULL)
3)     dic={·};
4)     for any node a∈D
5)        for any node b∈N
6)           if R(a,b)≥q then
7)              dic[(a,b)]=R(a,b);
8)           End If
9)        End For
10)    End For
11)    if (dic==NULL)
12)       break;
13)    else
14)       find (x,y) such that R(x,y) is maximum;
15)       D=D∪{y};
16)       update N;
17)    End If
18) End While
19) return D;
```

第 1 行初始化局部社区 D 和外壳节点集 N.第 2 行~第 18 行通过最大连通子图的方法寻找局部社区,其中,第 3 行~第 10 行计算 D 中节点与 N 中节点形成边的节点对的隶属度,将隶属度大于等于 q 的边以及该边的隶属度保存到字典类型变量 dic 中;第 11 行~第 17 行选择隶属度最大的边并入连通子图,并更新外壳节点集 N.第 19 行返回找到的局部社区 D.

2.3 时间复杂度分析

本文算法采用整型数据类型来表示网络 G 中的节点,每个节点被赋予一个唯一的整型序号来表示,采用关于节点的哈希表来存储整个网络,每个节点关联了一个向量,向量中存储按照由小到大排序好的邻居节点集.假定网络中节点度的平均值为 k,那么查找任一节点的邻居节点集的时间复杂度为 $O(1)$.计算两个节点 a 和 b 的共同邻居时,采用两个“指针”分别指向 $\Gamma(a)$和 $\Gamma(b)$遍历求交集的方法,该操作的时间复杂度为 $O(2k)$.

由于局部社区发现是在网络中给定节点的周围区域计算,不是建立在整个网络结构的基础上,因此算法的时间复杂度与所得到的社区的规模有直接关系,而不依赖于网络 G 中节点数$|V|$、边数$|E|$的大小.算法 1 中,最频繁的操作是第 6 行,计算节点 a,b 的隶属度 $R(a,b)$,该操作的时间复杂度为 $O(2k)$.假定最后得到的社区中节点个数为$|D|$,则算法在执行过程中需要计算$|D|\cdot k$ 条边的隶属度,即第 6 行在整个算法中重复的次数为$|D|\cdot k$.因此,算法的时间复杂度为 $O(2k^2\cdot|D|)$.

2.4 实例分析

为了更加直观地理解本文提出的局部社区发现算法,采用了由 9 个节点和 14 条边组成的小型网络[3],来说明本文算法的计算过程.网络连接情况如图 2 所示,真实社区划分{1,2,3,4}和{5,6,7,8,9}.

计算流程:设置阈值 q=0.3,节点 1 作为初始节点,首先将节点 1 加入到局部社区 D 中,即 D={1},节点 1 与其邻居节点 2~节点 4 的模糊关系隶属度 $R(1,2)$=0.5,$R(1,3)$=0.667,$R(1,4)$=0.333,均大于阈值 q,选择隶属度最大的节点 3 加入到 D 中,即 D={1,3}.计算 D 内节点与其邻居节点 2、节点 4 的模糊关系隶属度 $R(1,2)$=0.5,$R(3,2)$=0.5,

$R(1,4)=0.333$,$R(3,4)=0.333$,均大于阈值 q,选择隶属度最大的节点 2 加入到 D 中,即 $D=\{1,3,2\}$.以此类推,可将节点 4 加入到 D 中,即 $D=\{1,3,2,4\}$.然后计算社区 D 与其邻居节点的模糊关系隶属度 $R(4,5)=0.25$,$R(4,6)=0.25$,由于没有大于阈值 q 的邻居节点,因此算法结束,返回 $D=\{1,2,3,4\}$作为寻找到的局部社区.该结果与真实社区结果相同.

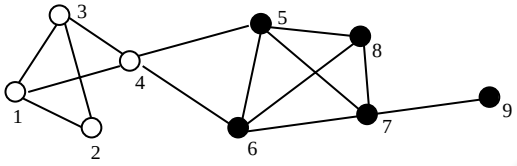


Fig.2 Small social network with 9 nodes and 14 edges

图 2 一个具有 9 个节点和 14 条边的小型社会网络

3 实验与结果

3.1 相关算法及评价指标

为了测试本文算法的有效性,与以下 4 种算法进行了比较分析.

(1) Clauset 算法[16]是第一个局部社区发现算法,通过最大化社区模块度 R 构建局部社区.

(2) LWP 算法[17]是一种著名的局部社区发现算法,通过优化社区模块度 M 寻找具有最大社区模块度的子图,从而发现一个节点所属的社区.

(3) GMAC 算法[19]是马连航等人提出的一种利用 d 层邻居节点采用相似度方法寻找与查询节点相似度最大的节点集合.对于参数 d 设置为作者推荐的 3.

(4) FlowPro 算法[22]是一种基于流传播的局部社区发现算法,起始节点发射一定数量的流在网络中传播,根据与起始节点在同一个社区中的节点获得的流多于社区外节点的原则确立起始节点所在的局部社区.

我们使用查准率 *Precision*、查全率 *Recall*、查准率和查全率的调和均值 *F-score*[19]来衡量局部社区发现算法的有效性.对于节点 v,假设节点 v 所在的真实社区的节点集合为 C_i;从节点 v 出发,局部社区发现算法找到的社区中的节点集合为 D,那么,

$$Precision=\frac{|D\cap C_i|}{|D|} \tag{2}$$

$$Recall=\frac{|D\cap C_i|}{|C_i|} \tag{3}$$

$$F\text{-}score=2\times\frac{Precision\times Recall}{Precision+Recall} \tag{4}$$

单一的查准率 *Precision* 或查全率 *Recall* 指标取值高低,反映不了算法有效性的高低.在很多场景下,随着 *Precision* 的降低,*Recall* 指标会升高;反之,随着 *Precision* 的升高,*Recall* 指标也会降低.因此,我们采用综合了这两个指标的调和均值 *F-score* 来衡量算法有效性的高低.

3.2 仿真网络数据集实验

首先,在 LFR 基准网络上测试本文算法的有效性.LFR 基准网络是由 Lancichinetti[24]提出,被广泛应用于社区发现算法测试中[18−20].LFR 基准网络节点度和社区规模均服从幂率分布,更接近于现实世界的真实网络.LFR 网络生成程序的常用参数见表 1.

本次实验中,参数设置为:n=5000,k=10,k_{max}=50,mu 取值为 0.1,0.15,...,0.5 共计 10 个值,相邻的两个取值之间间隔 0.05,其他参数使用默认值.混合参数 mu 为每个节点与社区外节点链接数目占该节点度的比例,例如,mu=0.1 表示社区中节点出发的边 90%在社区内部,10%连接社区外的节点.所以,随着 mu 值的增大,每个社区内

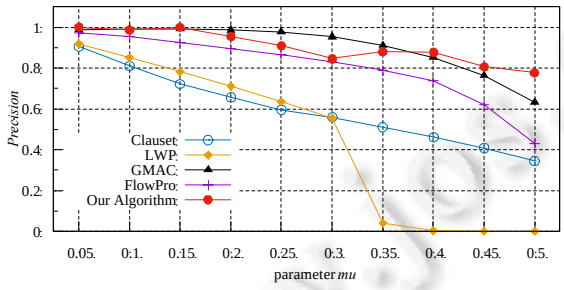
的节点与社区外的节点连接的比例逐渐增加,导致社区发现的难度越来越高.

Table 1 Parameters of LFR

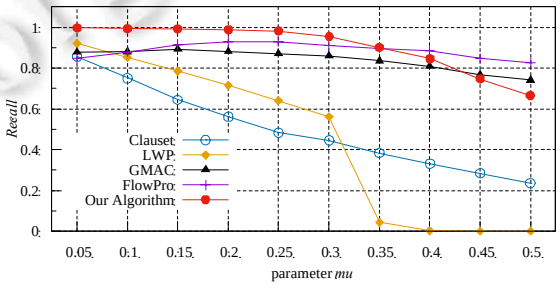
表 1 LFR 参数

参数	含义
n	网络中的节点个数
k	节点平均度
mu	混合参数
k_{max}	节点最大度
C_{min}	最小社区规模
C_{max}	最大社区规模
t_1	节点度的幂率分布指数
t_2	社区规模的幂率分布指数

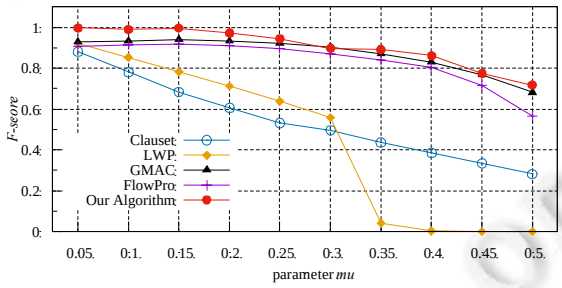
在实验中,我们以这些网络中的每一个节点作为起始节点,进行一次局部社区发现实验,在每个网络上重复 5 000 次实验,以这些实验的平均 Precison 和 Recall 作为该数据集的实验结果.在实验中,我们设置参数 q 为 0.2,实验结果如图 3 所示.



(a) *Precision* 指标的比较结果



(b) *Recall* 指标的比较结果



(c) *F-score* 指标的比较结果

Fig.3 Experimental results on LFR network datasets

图 3 在 LFR 网络数据集上的实验结果

当 mu=0.05 时,5 种算法的表现虽有一点差异,但它们的 *Precision*、*Recall* 和 *F-score* 都很高,算法的性能都很高.随着 mu 值的不断增大,5 种算法的性能都表现出了不同程度的下降,其中,Clauset 和 LWP 两种算法性能下降得最快.当 mu>0.3 后,LWP 算法的性能迅速下降,*Precision*、*Recall* 和 *F-score* 这 3 个指标都接近于 0 或者等于 0.这是因为 LWP 算法要求获得社区的 M 指标大于 1,也就是要求社区内节点形成的边数要多于社区内节点与社区外节点形成的边数.当 mu=1/3 时,社区内节点形成的边数等于社区内节点与社区外节点形成的边数;当 mu>1/3 后,也就意味着社区内节点形成的边数要小于社区内节点与社区外节点形成的边数,在这种情况下,几乎没有社区能够满足 M>1 这个条件,因此 LWP 算法的各指标近似于 0 或等于 0.这一结果与文献[18,20]的实验结果相同.

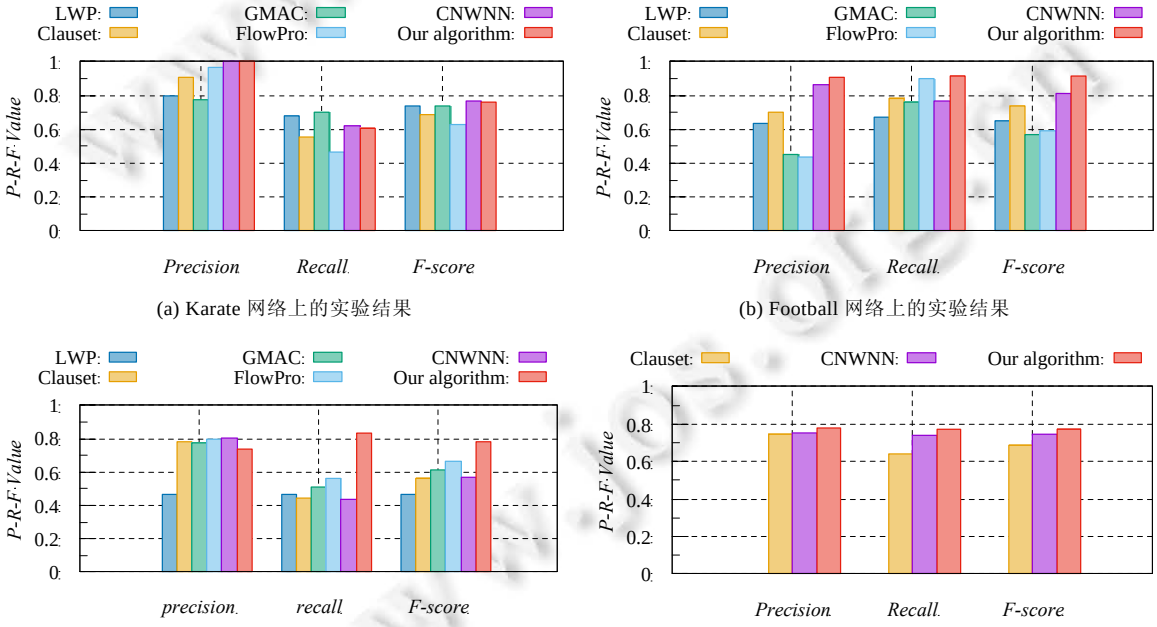
本文算法、GMAC、FlowPro 这 3 种算法的表现明显好于其他两种算法.虽然在 $mu\in\{0.1,0.2,0.25,0.3,0.35\}$ 时,本文算法的 *Precision* 略低于 GMAC,在 $mu\in\{0.45,0.5\}$时,本文算法的 *Recall* 指标略低于 GMAC 算法,在 $mu\in\{0.4,0.45,0.5\}$时,本文算法的 *Recall* 指标略低于 FlowPro 算法,但只有在 mu=0.3 时,本文算法的 *F-score* 略低于 GMAC,在其他 9 个数据集上效果都好于 GMAC 算法.基于模糊相似关系的局部社区发现算法每次选择的都是与已发现社区内节点隶属度最大的节点,优于其他方法,因此得到了更好的结果.

综上所述,本文算法在 LFR 基准网络上的表现优于其他 4 种算法.

3.3 在真实网络数据集上的仿真实验

上一节我们给出了在仿真网络数据集上的实验结果,本节我们在 4 个公认的用于测试社区发现算法有效性的真实网络数据集上测试本文算法.这 4 个真实数据集分别为:(1) Karate 数据集[25],该网络是 Zachary 在 1970~1972 年间观察美国一所大学的空手道俱乐部成员之间的关系形成的网络,该网络中包含 34 个节点和 78 条边,后来因为主管节点 34 和教练节点 1 之间发生分歧,分裂成两个小型的俱乐部;(2) Football 数据集[4],该网络是美国 NCAA 大学橄榄球 2000 年秋季常规赛 I~A 分区学校间的比赛网络数据,该网络中包含 115 个节点和 613 条边;(3) Polbooks 数据集[26],该网络是 Krebs 建立的美国政治图书网络,该网络中包含 105 个节点和 441 条边;(4) DBLP 数据集[27],该网络是一个科学家合作网络,网络中的边表示两人至少合作发表过一篇论文,该网络中包含 317 080 个节点和 1 049 866 条边.这 4 个网络都具有已知的社区结构.

采取与 LFR 数据集上相同的实验方法,我们以这些网络中的每一个节点作为起始节点,进行一次局部社区发现实验,在每个网络上重复 n(n 为网络中的节点个数)次实验,以这些实验的平均 *Precison* 和 *Recall* 作为每一个数据集的实验结果.由于 DBLP 数据集节点个数太多,LWP、GMAC、FlowPro 这 3 种算法运行时间过长,在该数据集上,我们不与这 3 个算法对比.为此,我们增加了与算法 CNWNN[21]的对比.在 Karate 数据集上,我们设置 q=0.5,实验结果如图 4(a)所示.在 Football 数据集上,我们设置 q=0.4,实验结果如图 4(b)所示.在 Polbooks 数据集上,我们设置 q=0.3,实验结果如图 4(c)所示.在 DBLP 数据集上,我们设置 q=0.7,实验结果如图 4(d)所示.



(a) Karate 网络上的实验结果

(b) Football 网络上的实验结果

(c) Polbooks 网络上的实验结果

(d) DBLP 网络上的实验结果

Fig.4 Experimental results on real network datasets

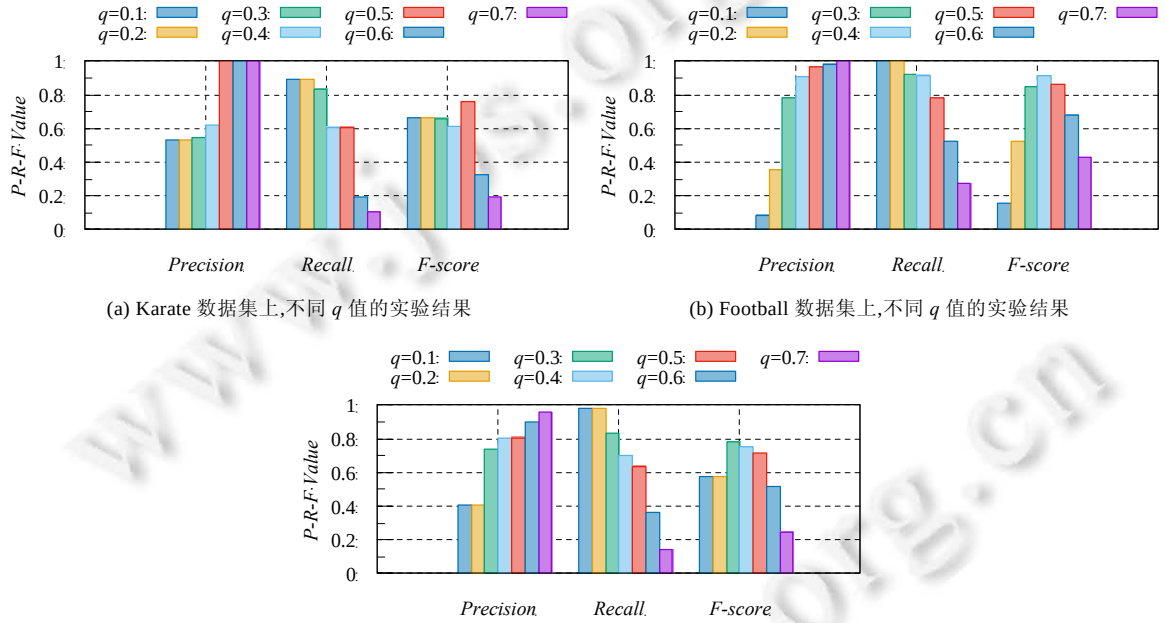
图 4 在真实网络数据集上的实验结果

在 Karate 数据集上,本文算法和 CNWNN 算法的 *Precision* 指标最高,*Recall* 指标低一点,CNWNN 算法的 *F-score* 指标最高,本文算法次之.在 Football 数据集上,本文算法的 *Precision* 和 *Recall* 都高于其他 5 种算法,因此,*F-score* 明显高于其他 5 种算法.在 Polbooks 数据集上,本文算法的 *Recall* 指标最高,虽然 *Precision* 略低于 Clause、GMAC、FlowPro、CNWNN 算法,但 *F-score* 指标明显高于其他 5 种算法.在 DBLP 数据集上,本文算法的 *Precision*、*Recall*、*F-score* 都高于 Clauset 和 CNWNN 算法.与在仿真网络数据集上的结果相同,基于模糊相似关系的局部社区发现算法选择邻居节点的方法优于其他算法.

综合以上分析可以得出,在这 4 个真实网络数据集上,本文算法表现最好.

3.4 实验参数讨论

本小节,我们讨论参数 q 的变化对本文算法的影响.在 Karate、Football、Polbooks 这 3 个数据集上,分别设置 q 值为 0.1,0.2,...,0.7 等 7 个值.采取与之前相同的实验方法,对于每个数据集上的每一个 q 值,我们以网络中的每一个节点作为起始节点,在网络上重复 n(n 为网络中的节点个数)次实验,以这些实验的平均 *Precison* 和 *Recall* 作为当前数据集上的当前 q 值的实验结果.在每一个数据集上观察在参数 q 变化的情况下,*Precision*、*Recall* 和 *F-score* 指标如何变化.实验结果如图 5 所示.



(a) Karate 数据集上,不同 q 值的实验结果

(b) Football 数据集上,不同 q 值的实验结果

(c) Polbooks 数据集上,不同 q 值的实验结果

Fig.5 Experimental results with different q on real-world network datasets

图 5 真实网络数据集上不同 q 值的实验结果

从图 5 可观察到,随着参数 q 的逐渐增大,本文算法在 Karate、Football 和 Polbooks 数据集上都呈现出了一致的规律:*Precision* 指标取值逐渐增大,而 *Recall* 指标取值逐渐降低;*F-score* 指标在 q 逐渐增大过程中,基本呈现出先增大后减少的趋势.参数 q 的增大,要求只有相似度更高的节点才能被加入到同一社区中,返回的节点是较低 q 值返回节点序列的前一部分,节点个数在减少,其中,正确的节点个数也随之减少,因此查全率 *Recall* 在降低,同时查准率 *Precision* 在提高.在算法的实际应用时,可以利用这一规律,调整 q 值,得到理想的社区.

4 结论与未来工作展望

近年来,随着互联网和电子通信技术的快速发展,产生了规模庞大的动态的网络大数据,传统的全局社区发

现方法由于难以获取其全局结构,无法有效处理这些动态的大型复杂网络.因此,近些年,局部社区发现算法引起了广大学者的关注.针对复杂社会网络中个体间的相似关系是模糊的或不确定的,本文提出一种基于模糊相似关系的局部社区发现方法.首先,采用模糊关系来刻画一条边的两个节点之间的相似关系.然后证明了该模糊关系为模糊相似关系,将某一节点所在的社区转化为节点关于模糊相似关系 q 水平上的等价类,通过寻找最大连通子图方法得到某一节点所在的局部社区.与其他算法相比,本文算法在仿真网络数据集和真实网络数据集上都取得了良好的效果.此外,还观察了参数 q 的变化对本文算法的影响,给参数设置提供了依据.

为了适应社会媒体和社交网络快速发展的需要,我们下一步的工作将针对动态的、异构的、内容与链接结合的复杂大型真实社会网络,设计出快速、高准确度和无监督的局部社区发现算法.

References:

[1] Zhao S, Liu XM, Duan Z, Zhang YP, Tang J. A survey on social ties mining. Chinese Journal of Computers, 2017,40(3):535−555 (in Chinese with English abstract).

[2] Girvan M, Newman MEJ. Community structure in social and biological networks. Proc. of the National Academy of Sciences of the United States of America, 2002,99(12):7821−7826.

[3] Qiao SJ, Han N, Zhang KF, Zou L,Wang HZ, Gutierrez LA. Algorithm for detecting overlapping communities from complex network big data. Ruan Jian Xue Bao/Journal of Software, 2017,28(3):631−647 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5155.htm [doi: 10.13328/j.cnki.jos.005155]

[4] Cheng XQ, Shen HW. Community structure of complex networks. Complex Systems and Complexity Science, 2011,8(1):57−70 (in Chinese with English abstract).

[5] Liu DY, Jin D, He DX, Huang J, Yang JN, Yang B. Community mining in complex networks. Journal of Computer Research and Development, 2013,50(10):2140−2154 (in Chinese with English abstract).

[6] Fortunato S, Hric D. Community detection in networks: A user guide. Physics Reports, 2016,659:1−44.

[7] Zhang ZH, Miao DQ, Qian J. Detecting overlapping communities with heuristic expansion method based on rough neighborhood. Chinese Journal of Computers, 2013,36(10):2078−2086 (in Chinese with English abstract).

[8] Li LQ, Gui XL, An J, Sun Y. Overlapping community detection algorithm based on fuzzy hierarchical clustering in social network. Journal of Xi'an Jiaotong University, 2015,49(2):6−13 (in Chinese with English abstract).

[9] Song L, Xie G, Yang YY. Community partition algorithm based on fuzzy clustering. Computer Engineering, 2016,42(8):126−133 (in Chinese with English abstract).

[10] Zhang YL, Wu B, Liu Y. A novel community detection method based on rough set *K*-means. Journal of Electronics and Information Technology, 2017,39(4):770−777 (in Chinese with English abstract).

[11] Wang XF, Liu GS, Li JH. A detecting community method in complex networks with fuzzy clustering. In: Proc. of the Int'l Conf. on Data Science and Advanced Analytics. 2014. 484−490.

[12] Sun PG, Gao L, Han SS. Identification of overlapping and non-overlapping community structure by fuzzy clustering in complex networks. Information Sciences, 2011,181(6):1060−1071.

[13] Liu Q, Peng ZM, Gao Y, Liu Q. A new *K*-means algorithm for community structures detection based on fuzzy clustering. In: Proc. of the IEEE Int'l Conf. on Granular Computing. 2012. 1−5.

[14] Zhao YL, Nie LQ, Wang XY, Chua TS. Personalized recommendations of locally interesting venues to tourists via cross-region community matching. ACM Trans. on Intelligent Systems & Technology, 2014,5(3):1−26.

[15] Liu Y, Ji XS, Liu CX. Detecting local community structure based on the identification of boundary nodes in complex networks. Journal of Electronics and Information Technology, 2014,36(12):2809−2815 (in Chinese with English abstract).

[16] Clauset A. Finding local community structure in networks. Physical Review E, 2005,72(2):Article No.026132.

[17] Luo F, Wang JZ, Promislow E. Exploring local community structures in large networks. Web Intelligence and Agent Systems, 2008, 6(4):387−400.

[18] Huang JB, Sun HL, Liu YG, Song QB, Weninger T. Towards online multiresolution community detection in large-scale networks. Plos One, 2011,6(8):Article No.e23829.

[19] Ma LH, Huang H, He QM, Chiew K, Wu JN, Che YZ. GMAC: A seed-insensitive approach to local community detection. In: Proc. of the Int'l Conf. on Data Warehousing and Knowledge Discovery. 2013. 297−308.

[20] Liu JL, Wang DL, Feng S, Zhang YF, Zhao WJ. A novel approach of discovering local community using node vector model. In: Proc. of the Web Information Systems Engineering (WISE). 2016. 513−521.

[21] Zhao WJ, Zhang FB, Liu JL. A novel local community detection algorithm based on common neighbors similarity measurement with weighted neighbor nodes. Journal of Nanjing University (Natural Science), 2018,54(4):751−757 (in Chinese with English abstract).

[22] Panagiotakis C, Papadakis H, Fragopoulou P. Local community detection via flow propagation. In: Proc. of the IEEE ACM Int'l Conf. on Advances in Social Networks Analysis and Mining. 2015. 81−88.

[23] Zhang XH, Pei DW, Dai JH. Fuzzy Mathematics and Rough Set Theory. Beijing: Tsinghua University Press, 2013. 141−143 (in Chinese).

[24] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. Physical Review E, 2008, 78(4):Article No.046110.

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

[26] Newman MEJ. Modularity and community structure in networks. Proc. of the National Academy of Science of the United States of America, 2006,103(23):8577−8582.

[27] Yang J, Leskovec J. Defining and evaluating network communities based on ground-truth. Knowledge and Information Systems, 2012,42(1):181−213.

附中文参考文献:

[1] 赵姝,刘晓曼,段震,张燕平,唐杰.社交关系挖掘研究综述.计算机学报,2017,40(3):535−555.

[3] 乔少杰,韩楠,张凯峰,邹磊,王宏志,Gutierrez LA.复杂网络大数据中重叠社区检测算法.软件学报,2017,28(3):631−647. http://www.jos.org.cn/1000-9825/5155.htm [doi: 10.13328/j.cnki.jos.005155]

[4] 程学旗,沈华伟.复杂网络的社区结构.复杂系统与复杂性科学,2011,8(1):57−70.

[5] 刘大有,金弟,何东晓,黄晶,杨建宁,杨博.复杂网络社区挖掘综述.计算机研究与发展,2013,50(10):2140−2154.

[7] 张泽华,苗夺谦,钱进.邻域粗糙化的启发式重叠社区扩张方法.计算机学报,2013,36(10):2078−2086.

[8] 李刘强,桂小林,安健,孙雨.采用模糊层次聚类的社会网络重叠社区检测算法.西安交通大学学报,2015,49(2):6−13.

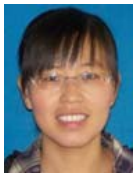
[9] 宋俐,谢刚,杨云云.基于模糊聚类的社团划分算法.计算机工程,2016,42(8):126−133.

[10] 张云雷,吴斌,刘宇.一种新的基于粗糙集 *K*-均值的社区发现方法.电子与信息学报,2017,39(4):770−777.

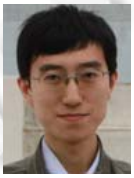
[15] 刘阳,季新生,刘彩霞.一种基于边界节点识别的复杂网络局部社区发现算法.电子与信息学报,2014,36(12):2809−2815.

[21] 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.南京大学学报(自然科学),2018,54(4):751−757.

[23] 张小红,裴道武,代建华.模糊数学与 Rough 集理论.北京:清华大学出版社,2013.141−143.



刘井莲(1980－),女,博士生,副教授,主要研究领域为社会网络分析,社会媒体挖掘.



冯时(1981－),男,博士,副教授,CCF 专业会员,主要研究领域为观点挖掘,情感分析,对话生成.

王大玲(1962－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为社会媒体处理,情感分析,数据挖掘,信息检索.

张一飞(1977－),女,博士,讲师,CCF 专业会员,主要研究领域为机器学习,多媒体数据处理.