

一种面向度中心性及重叠网络社区的发现算法

刘井莲^{1,2} 王大玲^{1,3} 赵卫绩² 冯 时^{1,3} 张一飞^{1,3}

(东北大学信息科学与工程学院 沈阳 110819)¹ (绥化学院信息工程学院 绥化 152061)²

(东北大学医学影像计算教育部重点实验室 沈阳 110819)³

摘 要 针对社会网络中存在较多以度中心节点为中心并且具有多社区重叠节点的网络社区结构,提出了一种面向度中心性及重叠网络社区的两阶段发现算法。第一阶段发现初始社区:选取度最大的 Top-k 个节点作为候选中心节点,并将每个节点与其邻居节点形成候选初始社区,其中如果某候选社区与已形成的初始社区的重叠度低于阈值,则形成一个新的初始社区;第二阶段调整社区划分:通过偏离度机制进行调整,将偏离度最大值对应的节点划分到连接紧密的相应社区内,形成最终社区划分。实验表明,该方法不仅能够揭示网络中以某个节点为中心的密集社区结构,还能有效处理初始社区不同程度的重叠问题。相比现有算法,所提方法对预先输入的候选初始社区数 k 值不敏感,并具有较高的准确性和灵活性。

关键词 社会网络,社区发现,度中心性,重叠度,偏离度

中图分类号 TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.3.006

Algorithm for Discovering Network Community with Centrality and Overlap

LIU Jing-lian^{1,2} WANG Da-ling^{1,3} ZHAO Wei-ji² FENG Shi^{1,3} ZHANG Yi-fei^{1,3}

(School of Information Science and Engineering, Northeastern University, Shenyang 110819, China)¹

(College of Information Engineering, Suihua University, Suihua 152061, China)²

(Key Laboratory of Medical Image Computing of Ministry of Education, Northeastern University, Shenyang 110819, China)³

Abstract Many social networks have central nodes and overlap nodes in more than one initial community, so a two-stage algorithm for discovering network community with centrality and overlap was proposed in this paper. In the first stage, initial communities are found. First, the top-k maximum degree nodes are chosen as candidate central nodes and the central nodes with their own neighbor nodes form separate candidate initial communities, and then the overlap degree of the candidate initial communities is computed one by one with generated initial communities. If all the overlap degrees are less than a given threshold, a new initial community is formed. In the second stage, the community division is adjusted. A concept of deviate degree is defined, and the corresponding nodes are merged to a closely linked community with maximum deviate degree. Finally community division is formed. Experimental results show that it can not only reveal the tightly-knit network communities in network with centrality, but also deal with problems of overlap initial communities effectively. Compared with existing algorithms, the algorithm in this paper is not sensitive to the prior number of candidate initial communities k , and has a high accuracy and flexibility.

Keywords Social network, Community discovery, Degree centrality, Overlap degree, Deviate degree

社会媒体和社交网络中人与人之间的关系、资源与资源之间的关系以及人对资源的操作形成了海量复杂多样的关系数据。在这些关系大数据背后,常常存在着相互紧密联系的密集区,区内节点间联系紧密,区外节点间联系稀疏,这样的密集区即为社区^[1]。社区可用于揭示有共同兴趣、爱好、教育背景或工作背景等共同特征的社会团体^[2],也可以用于揭示相互关联紧密的资源或商品信息,例如在商品推荐中,可以用于分析哪几样商品关联紧密,常被一起购买,这些关系紧密的商品就是一个商品社区,而传统关联规则算法则分析两种商

品或两个事件的关联程度。因此,社区发现的研究具有理论价值和实际应用意义。

1 网络社区结构与相关的发现方法

社区发现在过去的十几年里受到国内外研究者的广泛关注,并取得了很多开创性研究成果。2002年, Girvan 在文献[1]中首次提出社区发现算法,即著名的 GN 算法,引发了社区发现的研究热潮^[2-9]。传统社区发现研究的是非重叠社区结构,而在实际生活中,常常存在重叠社区结构,例如一个人

到稿日期:2015-03-10 返修日期:2015-06-16 本文受国家自然科学基金(61370074, 61402091)资助。

刘井莲(1980—),女,博士生,主要研究方向为社会网络分析与社交媒体挖掘, E-mail: datamining@163.com;王大玲(1962—),女,教授,博士生导师,主要研究方向为社交媒体挖掘、搜索与推荐;赵卫绩(1980—),男,讲师,主要研究方向为数据挖掘;冯 时(1981—),男,博士,讲师,主要研究方向为情感分析与观点挖掘;张一飞(1977—),女,博士,讲师,主要研究方向为多媒体数据挖掘。

可能既属于篮球俱乐部,又属于乒乓球俱乐部。为此,2005年,Palla在文献[5]首次提出基于派系过滤的重叠社区发现算法,开辟了重叠社区结构的研究热潮。目前已有基于团渗理论、局部扩张、链接划分等多种重叠社区挖掘算法^[5-9]被提出。

在大规模复杂的社会网络中,还存在着基于某些中心性节点的社区结构,例如明星粉丝团、传销团、恐怖分子团等。该类网络中的核心节点,也称为领袖节点或中心性节点,相比网络中其它参与者节点具有更强的影响力。近年来产生了一些基于中心性节点的社区发现算法^[10-15]。2004年,Costa在文献[10]首次提出基于度中心性的社区发现——Hub社区发现算法,但该算法需要预先知道社区数目。2009年,Lancichinetti在文献[8]首次提出重叠层次结构的社区发现算法,即著名的LFM算法,它通过随机选取网络中的初始节点,根据定义的局部适应度函数进行社区的逐层扩张,直到适应度函数为负时停止,从而得到种子节点所在的局部社区。该算法的缺点是随机选取种子节点,这样选择容易导致错误选取种子而不能得到合理的社区结构。2010年,Khorasgani在文献[11]将社区定义为领袖节点与其跟随者的集合,但是该算法需要社区个数的先验知识,对社区个数 k 值敏感。2011年,Wang等人在文献[12]提出基于核节点来揭示隐藏的社区结构。2012年,Coscia在文献[13]也提出基于某个节点与其邻居节点组成的子网分析,然后通过标签传播算法和一系列合并操作来发现社区。2012年,Zhang在文献[14]提出先发现核心节点,然后通过扩张核心节点的极大子图来获取社区结构。2012年,潘磊在文献[15]提出基于领袖节点扩张的重叠社区发现方法。

针对网络中同时存在中心性节点及重叠节点的情形,考虑到度中心性节点的影响力,本文提出一种面向度中心性及重叠网络的社区发现算法。算法分为两个阶段:第一阶段首先选取度最大的前 k 个节点作为候选中心节点,然后将度最大节点与其邻居节点构成第一个初始社区,按照度递减次序依次选取下一个节点,将该节点与其邻居节点形成候选初始社区,如果该候选社区与已形成的初始社区重叠度低于阈值,则形成一个新的初始社区,该阶段有效处理了社区的重叠问题;第二阶段针对初始社区中的重叠节点和不在初始社区中的节点,提出一个偏离度的度量机制,每次将偏离度最大值对应的节点划分到连接紧密的相应社区内,然后更新社区内节点和边界节点或邻居节点集,最终形成社区划分。实验表明,该方法不仅能够实现对网络的准确划分,还能有效处理候选初始社区不同程度的重叠问题,相比TopLeader算法,对预置输入的候选初始社区数 k 值不敏感,并具有较高的模块度值和ARI值。

本文第2节介绍社区的定义及重叠度、偏离度的计算方法,在此基础上给出面向度中心性的社区发现算法;第3节给出真实网络数据的测试与比较结果;最后总结全文。

2 面向度中心性的网络社区发现

2.1 相关概念

(1) 社区

网络是将复杂系统中的实体抽象为节点,将实体之间的关系抽象为边而形成的抽象模型,通常用图 $G=(V,E)$ 的形式来表示,其中 V 是节点集合, E 是边集合, $|V|$ 表示图中节

点的个数, V_i 表示图中第 i 个节点, $d(V_i)$ 表示节点 V_i 的度。根据网络中节点类型单一与否,可分为单模网络和多模网络。

社区是图的子图,用来表示网络中存在的“组内节点连接紧密,组间节点连接稀疏”的密集子图^[1]。社区内的节点根据与社区外节点是否有边,可分为边界节点集Boundary和核心节点集Core。所谓社区发现,是指在图 G 中确定 $k(k \geq 1)$ 个社区 $C=\{C_1, C_2, \dots, C_k\}$,使得各社区的节点集构成节点集 V 的一个覆盖。若任意两个社区的节点集的交集都为空,则称 C 为非重叠社区;否则称为重叠社区。本文所研究的社区发现算法针对的是单模网络的非重叠社区。

(2) 重叠度

在社交网络分析中,节点的度能说明一个节点的活跃程度和被关注度,以度大的节点为中心最有可能形成一个密集社区。但是,度大的节点并不一定能够形成以自己为中心的社区,因为如果两个节点的邻居相同,或者一个节点的邻居包含于另一个节点的邻居,那么这两个节点属于同一个社区。因此,为了判断度小的节点形成的候选初始社区能否成为初始社区,本文采用重叠度来度量一个社区包含于另一个社区的程度。

设 A, B 是两个社区,则这两个社区的重叠度 $over(A, B)$ ^[15]定义为:

$$over(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)} \quad (1)$$

$over(A, B)$ 用来表示社区 A 和 B 的重叠程度,取值范围为 $[0, 1]$ 。当 $A \cap B = \emptyset$ 时, $over(A, B)$ 取值为0;当 $A \subseteq B$ 或者 $B \subseteq A$ 时, $over(A, B)$ 取值为1;一般情况下, $over(A, B)$ 取值在 $(0, 1)$ 之间。

如果两个度中心节点能够形成初始社区,那么这两个节点显然属于不同的两个社区,它们与其邻居节点形成集合的重叠度应较小。基于此,本文在后面的算法1中基于一定的重叠度阈值,有效处理了候选初始社区的重叠问题。

(3) 偏离度

由于社区间与社区外的一个节点可能与不同社区多个节点相连,因此需要考虑该节点与哪个社区联系更紧密。基于此,本文定义一个连接度的公式来度量节点与形成中的社区之间的连接紧密程度。节点 V_i 与社区 C_j 连接度 $conn(V_i, C_j)$ 的定义如下:

$$conn(V_i, C_j) = \frac{E_{Core}}{E_{Core} + E_{Boundary} + E_{Shell}} \quad (2)$$

E_{Core} 为节点 V_i 与社区 C_j 的核心节点集Core连接的边数, $E_{Boundary}$ 为 V_i 与 C_j 的边界节点集Boundary连接的边数, E_{Shell} 为 V_i 与 C_j 外的节点集Shell连接的边数。

对于具有 k 个社区的整个社区集合 C ,节点 V_i 与 k 个社区的平均连接度 $avg(V_i, C)$ 定义为:

$$avg(V_i, C) = \frac{\sum_{j=1}^k conn(V_i, C_j)}{k} \quad (3)$$

节点到一个社区的连接度必须大于等于 avg 才有可能将节点划分到这一社区中。为了快速识别节点的归属,本文将节点 V_i 到社区 C_j 的偏离度定义为:

$$dev(V_i, C_j) = conn(V_i, C_j) - avg(V_i, C) \quad (4)$$

从式(4)中不难发现,偏离度越大表明该节点与对应社区连接越紧密。因此,本文在后面的社区发现算法2中,将偏离度最大节点优先划入到相应社区内。

2.2 算法思想

初始社区中心节点的选取和节点归属哪个社区是社区发现要解决的两个重要问题。Khorasgani 提出了一个著名的 TopLeader 方法^[11],即将社区看作是由领袖节点和跟随者节点组成的集合,在领袖节点的选取上提出把度最大且公共邻居节点少于 5 个的 k 个节点作为领袖节点的策略,这种策略使得成为社区中心的节点属于同一社区的可能性逐渐降低,社区发现的准确度和算法的稳定性提高,但该算法需要社区个数先验知识,对 k 值敏感。为解决这一问题,并提高社区发现的稳定性,本文提出的面向度中心性节点及重叠网络社区的发现算法首先选取度(即前述关于图定义中的 $d(V_i)$)最大的前 k 个节点作为候选中心节点,然后将度最大节点与其邻居节点构成第一个初始社区,按照度递减次序依次选取下一个节点,该节点与其邻居节点形成候选初始社区,如果该候选社区与已形成的初始社区重叠度低于阈值,则形成一个新的初始社区;对于初始社区中的重叠节点和不在初始社区中的节点,采用本文提出的偏离度概念(见式(4))进行度量,每次将偏离度最大值对应的节点划分到连接紧密的相应社区内。

2.3 算法描述

算法 1 形成初始社区。首先,获取网络中 top k 个度中心性节点,将这 k 个节点与其邻居节点形成候选初始社区;然后基于式(1),计算任意两个候选初始社区的重叠度,基于重叠度阈值 q ,形成网络中分散的多个初始社区。其有效处理了候选初始社区的重叠问题,对输入的候选初始社区数 k 不敏感。具体算法如下。

算法 1 发现初始社区

输入:网络 $G=(V,E)$,候选初始社区个数 k ,重叠度阈值 q ;

输出:初始社区划分 C

方法:

- 1) For $i=1$ to $|V|$ 计算节点 V_i 的度 $d(V_i)$;
- 2) 将度最大的 k 个节点依次保存到数组 top 中; // 结果为 $d(top[1]) \geq d(top[2]) \geq \dots \geq d(top[k])$
- 3) $k'=0$;
- 4) For $i=1$ to k
- 5) $tmpc=\{top[i]\}$; // $tmpc$ 表示候选初始社区
- 6) For $j=1$ to $|V|$ If $(top[i], V_j) \in E$ $tmpc=tmpc \cup \{V_j\}$;
- 7) For $j=1$ to k'
- 8) 计算 $over(tmpc, C_j)$; // 应用式(1)
- 9) If $over(tmpc, C_j) > q$ break;
- 10) EndFor
- 11) If $j=k'+1$
- 12) $k'++$;
- 13) $C_{k'}=tmpc$;
- 14) EndIf
- 15) EndFor
- 16) $C=\{C_1, C_2, \dots, C_{k'}\}$;

在算法 1 中,行 1)和 2)计算图中每个节点的度,选取度最大的 k 个节点依次保存到数组 top 中,目的是选择度大且与已形成的初始社区重叠度低的节点作为初始社区中心节点。行 3)–15)的循环中, $i=1, k'=1, top[i]$ 中的节点及其邻居节点集合构成 C_1 ; $i++$, $top[i]$ 中的节点及其邻居节点集合构成候选初始社区 $tmpc$ 。行 7)–10)计算 $tmpc$ 与已生成的初始社区 $C_1-C_{k'}$ 的重叠度,如果大于重叠度阈值 q ,则退出循环。行 11)–14)中,如果重叠度均小于阈值 q ,则 $k'++$,

$tmpc$ 成为新的初始社区 $C_{k'}$ 。行 16)中, $C_1, C_2, \dots, C_{k'}$ 形成初始社区划分 C 。算法 1 保证了初始社区之间的低重叠度,提高了社区发现的稳定性。

算法 2 解决 k' 个初始社区中重叠节点和初始社区外节点的归属问题。基于式(4),计算初始社区中的重叠节点和初始社区外节点的偏离度,将偏离度最大的节点优先划入到相应社区,不断重复该步骤,直到网络中所有节点划分到与其连接紧密的相应社区内。充分考虑到社区的“社区内节点连接紧密,社区间节点连接稀疏”的特性,从全局角度考虑将节点划入到与其连接最紧密的社区内,提高了社区发现的准确度。具体算法如下。

算法 2 形成最终社区划分

输入:网络 $G=(V,E)$,初始社区 $C=\{C_1, C_2, \dots, C_{k'}\}$;

输出:最终划分结果 $C=\{C_1, C_2, \dots, C_{k'}\}$

方法:

- 1) For $i=1$ to k'
- 2) For $j=1$ to $|C_i|$
- 3) $v=C_i[j]$; // $C_i[j]$ 表示初始社区 C_i 中的第 j 个节点
- 4) $Tidset(v)=Tidset(v) \cup \{i\}$; // $Tidset(v)$ 表示包含节点 v 的所有初始社区(编号)的集合
- 5) EndFor
- 6) EndFor
- 7) For $j=1$ to $|V|$ if $|Tidset(V_j)| > 1$ $B=B \cup \{V_j\}$;
- 8) For $j=1$ to k' $C_j=C_j-B$;
- 9) While $|B| > 0$
- 10) 初始化二维数组 $M[|B|, k']$; // $M[i, j]$ 为 B 中第 i 个顶点 B_i 到社区 C_j 的偏离度 $dev(B_i, C_j)$
- 11) 计算 M 中取值最大的元素所在的行和列,标记为 $ridx, cidx$;
- 12) $C_{cidx}=C_{cidx} \cup \{B_{ridx}\}$; // 将 B 中第 $ridx$ 个节点并入 C 中第 $cidx$ 个社区
- 13) $B=B-\{B_{ridx}\}$;
- 14) EndWhile
- 15) $C=\{C_1, C_2, \dots, C_{k'}\}$; // 更新后的 $C_1, C_2, \dots, C_{k'}$ 形成最终社区 C

在算法 2 中,行 1)–7)首先计算出初始社区重叠节点集合 B 。行 8)将重叠节点从每一个初始社区中删除。行 9)–14)计算 B 中节点到 k' 个初始社区的偏离度,将偏离度最大值的节点加入到连接紧密的相应社区中,然后更新 C 和 B ,重复上述步骤,直到重叠节点集合 B 为空。对于不在初始社区中的节点,同样采用行 9)–14)的方法,将不在初始社区中的节点加入到连接紧密的初始社区。在行 15),更新后的 $C_1, C_2, \dots, C_{k'}$ 形成的社区 C 是对网络 G 的最终划分结果。

3 实验分析

为了验证算法的可行性和有效性,本节首先在 Karate^[16]数据集上验证所提算法的准确性,之后与相关算法在数据集 Karate 和 Strike^[17]上进行比较。

3.1 算法测试

本文采用经典数据集 Karate,它是美国大学空手道俱乐部成员间的社会关系,该网络图中包含 34 个节点和 78 条边,其中每个节点表示一个俱乐部成员,节点间有连接表示两个成员之间经常在一起参加俱乐部活动。在该俱乐部,因为主管节点 34 和教练节点 1 之间发生分歧而分裂成以两个节点为核心的俱乐部。该俱乐部成员的社会关系网具体如图 1 所示。

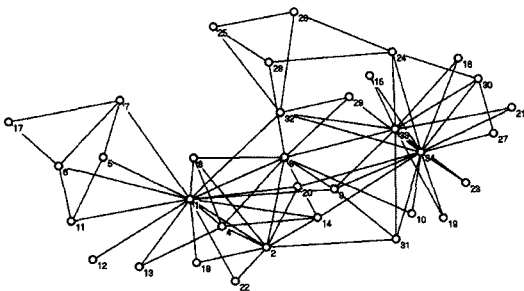


图1 Karate网络图

(1)算法流程分析(设置阈值 $q=0.42$, 候选初始社区数 $k=4$)

第一步:初始社区划分,度最大的4个节点分别是1,34,33,3。按照算法1,计算初始社区划分。以1为中心形成的初始社区 $C_1=\{1,2,3,4,5,6,7,8,9,11,12,13,14,18,20,22,32\}$ 。以34为中心形成候选初始社区 $\{9,10,14,15,16,19,20,21,23,24,27,29,30,31,32,33,34\}$,与 C_1 的重叠度为0.235,小于阈值 q ,生成新的初始社区 C_2 。以33为中心形成的候选初始社区 $\{3,9,15,16,19,21,23,24,30,31,32,33,34\}$,与 C_2 的重叠度为0.923,不生成新的初始社区。以3为中心形成的候选初始社区 $\{1,2,3,4,8,9,10,14,28,29,33\}$,与 C_1 的重叠度为0.636,不生成新的社区。 C_1, C_2 形成初始社区划分 C 。

第二步:处理初始社区重叠节点的归属,其中 $B=\{9,14,20,32\}$, $C_1=C_1-B$, $C_2=C_2-B$, 计算重叠节点到每个初始社区的偏离度,具体如表1所列。

表1 重叠节点偏离度

节点 社区	9	14	20	32
C_1	-0.1000	0.3000	0.1667	-0.1667
C_2	0.1000	-0.3000	-0.1667	0.1667

节点14到 C_1 的偏离度最大,因此 $C_1=C_1 \cup \{14\}$, 然后更新重叠节点集 $B=\{9,20,32\}$ 和 C , 重新计算偏离度。同理,依次分别将节点32,20,9划分到 C_2, C_1, C_2 中。对于不在 C_1 和 C_2 中的节点 $\{17,25,26,28\}$, 采用同样的方法将节点17划分到 C_1 , 节点25,26和28划分到 C_2 中。最终形成两个社区:以节点1为中心的16个节点组成的社区 C_1 即 $\{1,2,3,4,5,6,7,8,11,12,13,14,17,18,20,22\}$, 和以节点34为中心的18个节点组成的社区 C_2 即 $\{9,10,15,16,19,21,23,24,25,26,27,28,29,30,31,32,33,34\}$, 这一结果如图2所示。

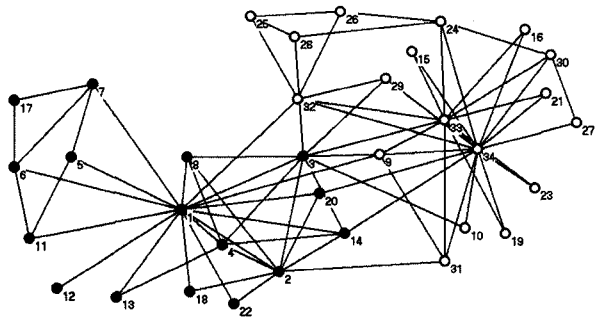


图2 Karate社区划分图

(2)实验分析

从上面计算流程不难发现, k 值选取越大,形成的候选初

始社区数越多,当输入连续的 k 值产生相同社区划分结果时,划分大多为合理划分。在 Karate 数据集上,经过多次反复实验发现,当 q 值设置为 $0.25 \sim 0.42$, k 值设置为 $2 \sim 34$ 时,所得到的都是如图2的划分结果,算法稳定性很好,对预置社区数 k 值不敏感。这是因为算法要求候选初始社区与已有初始社区的重叠度大于一定阈值才能成为初始社区。如果实际网络中存在以度小节点为中心节点的小社区时,在逐渐增加 k 值的情况下,该小社区由于与其它初始社区的重叠度低会成为初始社区。在 Karate 数据中由于不存在以度小节点为中心的小社区,因此 k 值从2增加到34所得的结果均相同。

(3)重叠节点归属分析

在复杂网络中常常存在同时与多个社区都有紧密连接的重叠节点。重叠社区发现和非重叠社区发现两类算法对重叠节点的处理方法不一样。前者考虑到重叠节点同时与多个社区连接紧密,将这些重叠节点同时划分到相应的多个社区内;而后者是把这些重叠节点划分到相对连接最紧密的社区内。两类算法都对 Karate 网络数据集进行了实验。采用重叠社区发现的文献[8]和文献[18]都将 Karate 网络中重叠节点3和节点10同时划入到不同的两个社区。采用非重叠社区发现的GN算法[1]将节点3误划,FN算法[4]将节点10误划,沈华伟在文献[3]也谈到将节点3误划,这些重叠节点被误划的原因是忽略了网络中不同节点所扮演的不同角色以及不同的影响力。例如重叠节点10在 C_1 和 C_2 中虽各有一个邻居,但在 C_2 中的邻居节点是俱乐部的主管节点34,而在 C_1 中邻居节点是俱乐部的普通成员,从度中心性节点的重要性和影响力考虑,节点10应划入到 C_2 中。本文正是考虑到度中心性节点的影响力以及重叠节点与不同社区的连接紧密程度,实现了同时具有度中心性节点和重叠节点网络的准确划分。

3.2 与相关算法进行比较

为将本文方法与相关工作进行比较,首先给出两个度量指标。

(1)模块度函数 Q

模块度最早由 Newman 和 Girvan 提出[19]。模块度函数 Q 用来衡量社区划分的质量,当同一社区内节点连接越紧密,不同社区间节点连接越稀疏时, Q 函数取值越大。

Q 的函数表达式为:

$$Q = \frac{1}{2m} \sum_{vw} [A_{vw} - \frac{d(v)d(w)}{2m}] \delta(C_v, C_w) \quad (5)$$

其中, v, w 为节点; A_{vw} 为邻接矩阵 A 中的元素。若节点 v, w 相连,则 $A_{vw}=1$, 否则 $A_{vw}=0$; $m = \frac{1}{2} \sum_{ij} A_{ij}$ 为图中边的数目; C_v 为节点 v 所属的社区; $\delta(C_v, C_w)$ 表示节点 v 所属的社区 C_v 与节点 w 所属的社区 C_w 是否相同,相同时为1,否则为0; $d(v)$ 是节点 v 的度; $\frac{d(v)d(w)}{2m}$ 表示在随机网络中节点 v 和节点 w 之间存在连边的可能性。

(2)ARI

ARI 全称为 Adjust Rand Index[11]。

设对象集 $S=\{O_1, O_2, \dots, O_n\}$, $U=\{u_1, u_2, \dots, u_R\}$ 和 $V=\{v_1, v_2, \dots, v_c\}$ 是 S 的两个划分。

$$ARI = \frac{\binom{n}{2} (a+d) - [(a+b)(a+c) + (c+d)(b+d)]}{\binom{n}{2} - [(a+b)(a+c) + (c+d)(b+d)]} \quad (6)$$

其中, $\binom{n}{2}$ 表示可以组成的节点对的总数, a 表示在 U, V 中所属的群组都相同的节点对总数, b 表示在 U 中所属的群组相同而在 V 中所属的群组不同的节点对总数, c 表示在 U 中所属的群组不同而在 V 中所属的群组相同的节点对总数, d 表示在 U, V 中所属的群组都不同的节点对总数。

采用式(5)所示的模块度和式(6)所示的 ARI 两个指标, 将本文提出的算法与著名的 CFinder^[5]、FastModularity^[4] 和 TopLeader^[11] 算法在社区发现经典测试数据集 Karate 和 Strike 上进行比较, 结果如图 3 所示。

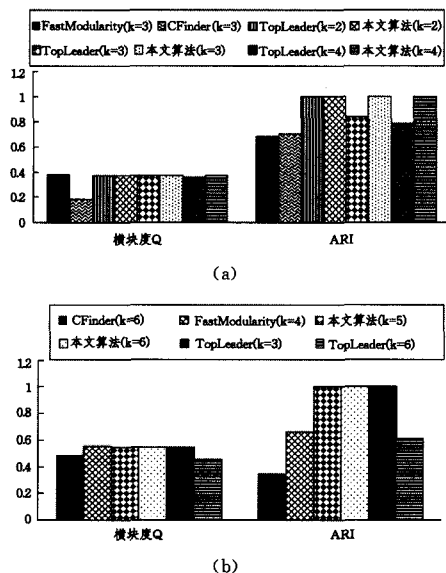


图 3 不同算法的比较

从图 3 的运行结果来看, 对于给定的不同 k 值, 本文算法大多时候能得到较高的模块度值和 ARI 值, 并且对社区数 k 值不敏感, 能够有效处理候选初始社区的重叠问题, 获取合理的社区结构。在 TopLeader 算法中, 当输入社区数 k 与实际社区数相符时, 能够得到与实际相符的划分结果, 但很难预先知道一个网络的真实社区个数。而在本文算法中, 如果预先不知道准确的社区数, 可以输入多个不同 k 值, 本文输入的 k 值仅仅表示候选初始社区数, 如果 k 值较大, 必然会形成较多个重叠的候选初始社区。由于对高于一定重叠度阈值的两个候选初始社区进行相应重叠处理, 因此, 实际获取的初始社区数 k' 一般小于候选初始社区数 k 。重叠度阈值的设定与网络的结构和规模有很大关系, 要通过多次实验、观察划分结果的模块度值来调整, 选择最佳结果。经过多次反复实验发现, 大多网络的重叠阈值设定在 0.25 左右时能够得到合理划分。

结束语 本文在分析当前社区发现算法的基础上, 针对社会网络分析中常常存在一些以某个节点为中心的密集区并具有重叠节点的网络结构, 提出一种面向此类结构社区的发现算法, 该算法通过获取网络中分散的中心节点来处理社区间的重叠问题, 通过引入偏离度概念解决了网络中不确定节点的归属问题。在经典数据集上的仿真实验表明, 本文算法能够发现与实际相符的社区结构, 具有较高的准确性和灵活性。与和其相似的 TopLeader 算法相比, 本文算法对预置的社区数 k 不敏感。

进一步的工作包括针对复杂的网络结构, 设计出快速、高准确度和无监督(不依赖先验知识, 对参数不敏感)的社区发

现算法。同时, 为了适应社会媒体和社交网络快速发展的需要, 也将进一步开展对动态网络、异构网络、内容与链接结合等复杂网络的社区发现算法的研究与设计。

参考文献

- [1] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12): 7821-7826
- [2] Gan W Y, He N, Li D Y, et al. Community Discovery Method in Networks Based on Topological Potential[J]. Journal of Software, 2009, 20(8): 2241-2254(in Chinese)
淦文燕, 赫南, 李德毅, 等. 一种基于拓扑势的网络社区发现方法[J]. 软件学报, 2009, 20(8): 2241-2254
- [3] Shen H W, Cheng X Q, Chen H Q, et al. Information bottleneck based community detection in networks[J]. Chinese Journal of Computer, 2008, 31(4): 677-686(in Chinese)
沈华伟, 程学旗, 陈海强, 等. 基于信息瓶颈的社区发现[J]. 计算机学报, 2008, 31(4): 677-686
- [4] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E, 2004, 69(6): 066133-1-066133-5
- [5] Palla G, Derényi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435(7043): 814-818
- [6] Shen H W, Cheng X Q, Cai K, et al. Detect overlapping and hierarchical community structure in networks[J]. Physica A, 2009, 388(8): 1706-1712
- [7] Evans T S. Clique graphs and overlapping communities[J]. Journal of Statistical Mechanics Theory and Experiment, 2010, 2010(3): 257-265
- [8] Lancichinetti A, Fortunato S, Kertesz J. Detecting the overlapping and Hierarchical community structure in complex networks[J]. New Journal of Physics, 2008, 11(15): 19-44
- [9] Evans T S, Lambiotte R. Line graphs, link partitions and overlapping communities[J]. Physics Review E, 2009, 80(1): 145-148
- [10] Costa L F. Hub-based Community Finding [OL]. [2015-06-15]. <http://wenku.baidu.com/view/ca8a35e981c758f5f61f6767.html>
- [11] Khorasgani R R, Chen J, Zaiane O R. Top Leaders Community Detection Approach in Information Networks[C]// Proceedings of the 4th Workshop on Social Network Mining and Analysis. Washington DC, USA, 2013
- [12] Wang L, Lou T, Tang J, et al. Detecting Community Kernels in Large Social Networks[C]// The 11th IEEE International Conference on Data Mining. Vancouver, BC, Canada, 2011: 784-793
- [13] Coscia M, Rossetti G, Giannotti F, et al. DEMON: a local-first discovery method for overlapping communities[C]// The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 615-623
- [14] Zhang T, Wu B. A Method for Local Community Detection by Finding Core Nodes[C]// International Conference on Advances in Social Networks Analysis and Mining. Istanbul, Turkey, 2012: 1171-1176

因此 $B^* = (0, 1)/y_1 + (0, 1, 0, 7)/y_2 + (0, 2, 0, 7)/y_3 + (0, 6, 0, 3)/y_4 + (1, 0)/y_5$ 。

与 $B = \text{“高”}$ 进行语义匹配^[14, 15] 比较可知, B^* 为“较高”, 此时应该适当加热, 使室温变得较高。

结束语 针对直觉模糊推理问题, 通过求出直觉模糊对的强真度和真值支持度, 将其分别用模糊推理方法进行推理, 进而给出了一种新的推理方法。模糊推理方法在推理过程中只考虑了隶属度, 而本文所提出的方法不仅考虑了隶属度, 同时也考虑了非隶属度, 因而推理过程更加准确; 直觉模糊推理方法在推理过程中需要同时对隶属度和非隶属度进行运算, 而本文所提出的方法在推理过程中引入强真度和真值支持度, 使复杂的直觉模糊推理转换为简单的模糊推理, 因而推理过程更加简便。该方法为直觉模糊推理问题提供了一种有效、简便的解决方法, 因此, 它是对直觉模糊推理有效扩展的新尝试, 且取得了良好的效果。

参考文献

- [1] Zadeh L A. Fuzzy sets[J]. Information and Control, 1965, 8(3): 338-353
 - [2] Atanassov K. Intuitionistic fuzzy sets[J]. Fuzzy Sets and Systems, 1986, 20(1): 87-96
 - [3] Atanassov K. More on intuitionistic fuzzy sets[J]. Fuzzy Sets and Systems, 1989, 33(1): 37-45
 - [4] Li You-yu, Zhang Xing-fang. Research on fuzzy CRI method and fuzzy reasoning[J]. Fuzzy Systems and Mathematics, 2013, 27(4): 42-47 (in Chinese)
李友雨, 张兴芳. 模糊 CRI 算法及模糊推理的有效性研究[J]. 模糊系统与数学, 2013, 27(4): 42-47
 - [5] Wang Chao. Triple I method with the reductive property of general fuzzy inference[J]. Journal of Chongqing Normal University(Natural Science), 2013, 30(3): 103-106 (in Chinese)
王超. 一般模糊推理具有还原性的全蕴涵三 I 算法[J]. 重庆师范大学学报(自然科学版), 2013, 30(3): 103-106
 - [6] Chen Tu-yun, Meng Yan-ping. The reasoning method by fuzzy set similarity degree[J]. Journal of Engineering Mathematics, 2005, 22(2): 346-348 (in Chinese)
陈图云, 孟艳平. 模糊集相似度限定推理方法[J]. 工程数学学报, 2005, 22(2): 346-348
 - [7] Lei Ying-jie, Wang Bao-shu, Lu Yan-li. Approximate reasoning method based on intuitionistic fuzzy logic[J]. Control and Decision, 2006, 21(3): 305-310 (in Chinese)
雷英杰, 王宝树, 路艳丽. 基于直觉模糊逻辑的近似推理方法[J]. 控制与决策, 2006, 21(3): 305-310
 - [8] Lei Ying-jie, Wang Jing-yu, Ji Bo, et al. Technique for intuitionistic fuzzy reasoning with truth qualifications[J]. Systems Engineering and Electronics, 2006, 28(2): 234-236 (in Chinese)
雷英杰, 汪竞宇, 吉波, 等. 真值限定的直觉模糊推理[J]. 系统工程与电子技术, 2006, 28(2): 234-236
 - [9] Wang Yi, Lei Ying-jie. Techniques for interpolation reasoning based on intuitionistic fuzzy logic[J]. Systems Engineering and Electronics, 2008, 30(10): 1944-1948 (in Chinese)
王毅, 雷英杰. 基于直觉模糊逻辑的插值推理方法[J]. 系统工程与电子技术, 2008, 30(10): 1944-1948
 - [10] Wang Yi, Liu San-yang, Zhang Wen, et al. Intuitionistic fuzzy similarity measures reasoning method based on inclusion degrees[J]. Systems Engineering and Electronics, 2014, 36(3): 494-500 (in Chinese)
王毅, 刘三阳, 张文, 等. 基于包含度的直觉模糊相似度量推理方法[J]. 系统工程与电子技术, 2014, 36(3): 494-500
 - [11] Yang Hui-cheng, Chen Hao. Intuitionistic fuzzy approximate reasoning based on intuitionistic fuzzy operation[J]. Application Research of Computer, 2011, 28(1): 102-104 (in Chinese)
杨会成, 陈昊. 基于直觉模糊蕴涵式运算直觉模糊近似推理方法[J]. 计算机应用研究, 2011, 28(1): 102-104
 - [12] Lv Da-jiang, Shi Zhi-han, Lei Ying-jie, et al. Intuitionistic fuzzy inference emulator based on Mamdani implication relation[J]. Computer Engineering and Design, 2010, 31(11): 2539-2542 (in Chinese)
吕大江, 石志寒, 雷英杰, 等. 基于 Mamdani 蕴含关系的直觉模糊推理仿真[J]. 计算机工程与设计, 2010, 31(11): 2539-2542
 - [13] Zou Li, Tan Xue-wei, Zhang Yun-xia. Knowledge reasoning based on linguistic truth-valued intuitionistic fuzzy logic[J]. Computer Science, 2014, 40(1): 134-137 (in Chinese)
邹丽, 谭雪微, 张云霞. 语言真值直觉模糊逻辑的知识推理[J]. 计算机科学, 2014, 40(1): 134-137
 - [14] Lei Ying-jie, Zhao Ye, Wang Tao. On the measurement of similarity on semantic match for intuitionistic fuzzy[J]. Journal of Air Force Engineering University(Natural Science), 2005, 6(2): 70-73 (in Chinese)
雷英杰, 赵晔, 王涛. 直觉模糊语义匹配的相似性度量[J]. 空军工程大学学报(自然科学版), 2005, 6(2): 70-73
 - [15] Lei Ying-jie, Wang Tao, Zhao Ye. On the semantic distance and near compactness for intuitionistic fuzzy match[J]. Journal of Air Force Engineering University(Natural Science), 2005, 6(1): 69-72 (in Chinese)
雷英杰, 王涛, 赵晔. 直觉模糊匹配的语义距离与贴近度[J]. 空军工程大学学报(自然科学版), 2005, 6(1): 69-72
-
- (上接第 37 页)
- [15] Pan L, Dai C, Wang C J, et al. Overlapping Community Detection via Leader-Based Local Expansion in Social Networks[C]//IEEE 24th International Conference on Tools with Artificial Intelligence. Athens, Greece, 2012: 397-404
 - [16] Zachary W W. An information flow model for conflict and fission in small groups[J]. Journal of Anthropological Research, 1977, 33(4): 452-473
 - [17] Nooy W D, Mrvar A, Batagelj V. Informal communication within a sawmill strike[OL]. [2015-06-15]. <http://vlado.fmf.uni-lj.si/pub/networks/data/esna/strike.htm>
 - [18] Zhang Z H, Miao D Q, Qian J. Detecting Overlapping Communities with Heuristic Expansion Method Based on Rough Neighborhood[J]. Chinese Journal of Computers, 2013, 36(10): 2078-2086 (in Chinese)
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[J]. 计算机学报, 2013, 36(10): 2078-2086
 - [19] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2): 026113