doi: 10.3969/j.issn.0490-6756.2018.01.015

基于共邻节点相似度的加权网络社区发现方法

刘苗苗[1], 郭景峰[2,3], 马晓阳[2,3], 陈　晶[2,3]
(1.东北石油大学, 大庆 163318; 2.燕山大学信息科学与工程学院, 秦皇岛 066004;
3.河北省虚拟技术与系统集成重点实验室, 秦皇岛 066004)

摘　要: 为实现加权网络的准确划分,发现真实的社区结构,提出一种基于模块度和共邻节点相似性的层次聚类社区划分方法 IEM. 首先,定义两节点间基于共邻节点的相似度. 之后,基于该度量快速聚合当前节点和与其关联紧密度最强的邻居节点以形成初始社区,并进行社区扩展. 最后,以最大化网络模块度为目标进行社区合并以优化划分结果. 算法通过形成初始社区、扩展社区、合并社区三步,实现了加权网络合理有效的社区划分. 以加权模块度作为社区划分质量的评价标准,在多个数据集上的实验结果表明,IEM 算法优于加权 CN、加权 AA、加权 RA. 同时,与 CRMA 算法相比,IEM 算法对加权网络社区划分的有效性和正确性更高.

关键词: 加权网络; 模块度; 共邻节点; 相似度; 社区划分
中图分类号: TP393.01; TP301.6　　**文献标识码**: A　　**文章编号**: 0490-6756(2018)01-0089-10

Community discovery in weighted social networks based on similarities of common neighbors

LIU Miao-Miao[1], *GUO Jing-Feng*[2,3], *MA Xiao-Yang*[2,3], *CHEN Jing*[2,3]
(1. Northeast Petroleum University, Daqing 163318, China;
2. College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China;
3. The Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province, Qinhuangdao 066004, China)

Abstract: In order to divide communities accurately in weighted networks, a hierarchical clustering method IEM based on the similarity and modularity is proposed. Firstly, the similarity of the two nodes is defined based on attributes of their common neighbors. Then, the most closely related nodes are clustered fastly according to their similarity to form the initial community and expand it. Lastly, these communities are merged with the goal of maxmizing the modularity so as to optimize division results. The algorithm achieves more reasonable and effective community division for weighted network by three steps of initializing, expanding and merging communities. Correctness and effectiveness of the algorithm are verified through experiments on many weighted networks using weighted modularity as evaluation index. Results show that IEM is superior to weighted CN, weighted AA and weighted RA. Moreover, it can achieve the higher quality of community division in weighted networks compared with CRMA algorithm.

Keywords: Weighted networks; Modularity; Common neighbors; Similarity; Community division

收稿日期: 2016-11-23
基金项目: 国家自然科学基金(61472340); 国家自然科学青年基金 (61602401)
作者简介: 刘苗苗(1982－), 女, 河南洛阳人, 副教授, 博士生, 研究方向为社会网络分析. E-mail: liumiaomiao82@163.com

1 引 言

社会网络中的社区发现对于理解真实网络的拓扑结构、功能分析、行为预测等具有重要的理论意义和实用价值,被广泛应用于蛋白质功能预测、舆情分析与控制、搜索引擎等众多领域.然而,真实世界中许多网络的边都带有权重.例如,社交网络中个体间的联系存在强弱程度之分;新陈代谢网络中不同反应路径上有着不同的物理流量;交通网络中不同区域之间车的流量也不同.如果用加权网络来描述此类复杂系统,便能够更好地表达个体间的联系强度.加权网络就是边带有权重属性的社会网络[1],用一个数值表达节点间的距离、联系的强弱程度、相似度等不同的含义.例如,航空运输网络中的权重代表两个机场之间的航班次数;移动通信网络中的权重代表两个用户在某个时间段内的通话时长或联系次数;科研合著网络中的权重代表科学家之间的合作次数.在加权网络中,权重不仅表达了节点之间是否存在联系,而且表达了这种联系的紧密程度.根据意义权重可分为相似权和相异权两种.相似权是指权值越大,两节点间距离越近或关系越"亲密".在刻画相似度贡献时,通过权重越大的边相连的两个节点相似度越大.例如,人人网中,权重代表两个人之间的互动次数;互动越频繁,两个人关系越紧密.相异权是指权值越小,两节点间距离越近或关系越"亲密".在刻画相似度贡献时,通过权重越小的边相连的两个节点相似度越大.例如,电力网络中,权重代表两个网点的地理距离,其值越小,两个电力网点间交流越"亲密".加权网络中的权重所包含的数据能够很好地刻画现实系统,帮助理解并分析社会网络的拓扑结构及性质,对于社区划分也具有非常重要的意义.例如,语义网中的权重表达了词条处于同一语义群的可能性大小,在聚类过程中,如果忽略了权重,将丢失大量的重要信息,致使最终聚类结果与实际语义群结构存在较大的差距.此外,许多研究诸如任迎伟[2]等都指出了强关系(即连边权重较大的关系双方)和弱关系(即连边权重较小的关系双方)对于社会网络社区结构分析和研究的重要性.因而,针对加权网络中的社区发现,权重应作为节点聚类的重要考虑因素.本文研究中权重采用相似权.

当前,有关加权网络的社区发现方法已经有了一定的研究,相关算法主要有:Newman将无权网络中的GN算法拓展到加权网络中,将边介数替换为关于权重的边介数,提出了WGN算法[1].郭崇慧等[3]利用多重图的思想,将无权网络中基于共邻矩阵的方法扩展到加权网络中用于社区发现.Jaho等人[4]提出了一种基于兴趣相似度的框架,用于加权图中的社区发现.该文中边的权重是基于用户间兴趣标签的相似度定义的,然而该方法仅对具有较高兴趣相似度的节点聚类效果较好.Kumar等人[5]提出一种基于密度的社区挖掘方法.在该文的Twitter数据集中,twitter是节点,两节点间的retweets是权重,然后基于可变密度进行聚类.但实验结果不理想,在Twitter数据集中仅发现了少量社区.Liu等人[6]提出了基于社区间吸引力的大型社会网络加权图聚类算法,然而该算法仅对稀疏的加权网络聚类效果较好.Sharma[7]提出加权符号网络社区挖掘算法AGMA,依据链接类型和链接权重将加权图划分为若干个社区.Lu等人[8]提出了一种基于Conductance的加权网络社区发现算法I^2C,将具有最大归属度的边加入社区,并利用社区导率来判定是否形成新的社区.Han等[9]提出了改进的CNM算法,通过定义关于权重大小的增量模块度来实现加权网络中的社区发现,但算法复杂度接近$O(\mathrm{mnlb}^2(n))$.王坤等[10]提出了一种基于相似度的中心聚类算法SCC,选取中心度大的节点作为社区中心,然后基于节点归属度实现了加权网络中的社区发现.王佳嘉[11]提出了针对节点数、边数、权重和社区数量随时间变化的动态加权网络社区发现方法.Chen等人[12]通过发现初始部分社区并扩展初始社区,实现了加权网络中的重叠社区发现.林旺群等[13]提出了一种基于带权图并行分解的层次化社区发现方法,有效避免了传统基于模块度的社区发现方法倾向于发现相似规模社区的弊端.王书凯[14]提出了一种分裂式的基于贪心选择策略的算法,用于加权网络的社区发现.然而,在Zachary's Karate club以及Dolphins等加权网络上的实验显示,划分结果与真实社区结构有一定的偏差.詹培森[15]提出了一种加权网络局部社区发现算法,以局部极大权值点为起始节点,通过逐步添加节点来发现局部社区.姚宏亮等人[16]提出了一种复合加权模型以有效表示股票网络的活跃性,进而实现了复合加权网络的社团划分.Saha等人[17]基于模糊聚类的概念提出了社区软划分的方法,解决了加权网络的重叠社区发现问题.赵健等人[18]通过计算网内移动节点间的最优路径树、相似指数和社区离散指数等参数,实现了

有向加权网络中服务社区的合理划分. 季大祥[19]提出了一种扩展的加权模块度计算方法，并通过实验验证该方法能够更好地结合节点间连接权重对社团结构进行度量. 姚尊强[20]提出了一种加权短消息网络中的社区发现方法，并对基于节点局部结构信息的加权 CN、加权 AA、加权 RA 相似度指标进行了研究. Guo 等人[21]针对 AGMA 算法存在的问题进行改进，提出了一种加权网络社区发现算法 CRMA.

一个好的社区划分算法应该同时满足两点：较高的社区划分准确率和较低的计算复杂度. 然而，现有大多数算法很难同时达到以上两点. 鉴于层次聚类法是经典的准确率较高的社区划分算法，本文首先定义了基于共邻节点的相似度进行节点快速聚类，并以最大化加权网络模块度 Q_w[18] 为目标进行社区合并，提出了一种划分准确率较高、复杂度较低的凝聚算法 IEM (Initialize-Extend-Merge)，用于加权网络中的社区发现. 论文首先分析了加权网络社区发现研究现状，之后描述了 IEM 算法的核心思想、相关定义及实现过程，最后通过实验对算法的有效性和正确性进行了验证，实验结果也显示了所提算法的优越性.

2　基于共邻节点的加权相似度定义

2.1　算法思想

加权网络中的社区划分不仅要考虑节点间是否关联，还要考虑关联的紧密程度，将每个节点划分至与其连接紧密度最大的社区，使得同一社区内连接稠密的同时，节点间的连接紧密度也较大. 而基于相似性的层次聚类社区发现算法认为，两节点间相似度越大，两者属于同一社区的可能性越高. 因此，算法的关键是有效捕获影响节点相似度的网络拓扑属性，合理定义节点间的相似性度量，进而完成节点聚类及社区发现.

为降低复杂度，相似性度量的定义中只考虑了两节点及其共同邻居节点的度、强度、边权信息对于两节点的相似度影响. 算法认为，若两节点不直接相连，则两者相似度为 0；若两节点直接相连，则两者的相似度取决于其共同邻居节点对两者的相似性贡献.

首先，在度量共邻节点对于两节点的相似性贡献时，认为两节点拥有的共同邻居数目越多，两者相似度越高. 因此，算法考虑了两节点的所有共同邻居对于两者的相似性贡献之和. 然而，在加权网络中，度数相同的两节点其强度不一定相同，反之亦然. 在此，算法认为，度数越小、强度越大的共同邻居对其连接的两节点的相似性贡献要大于度数大而强度小的共邻节点的作用. 基于此，定义节点的单位权重，用来度量该节点对其所连接的节点对的相似性贡献. 节点的单位权重应与该节点的强度成正比，与其度数成反比. 也即，单位权重越大的共邻节点对其所连接的两节点的相似度贡献越大.

其次，在度量与共邻节点相连的两条边的权重对于两节点的相似度贡献时，认为与共邻节点相连的两条边的权重之和在两节点所有连边的权重之和中所占的比重越大，两节点相似度越高. 基于此，定义共邻节点作用系数，用来度量某共邻节点在其所连接的两节点的所有邻居节点中所起的相似性贡献程度. 两节点的共邻节点作用系数等于两者与该共邻节点的两条连边的权重之和比上两者与它们所有邻居节点连边的权重之和. 其值越大，该共邻节点在所有邻居节点中对于两节点的相似度贡献越大.

再次，在定义节点的单位权重以及共邻节点作用系数的基础上，提出共邻节点连接强度的概念，其值等于该共邻节点的作用系数与其单位权重的乘积. 该值越大，表明该共邻节点对其所连接的两节点的相似度贡献越大.

最后，将两节点的所有共邻节点连接强度之和作为两者基于共邻节点的相似度. 在此，有一种特殊情况，当两节点没有共同邻居节点时，引入节点对边权强度的概念作为两者的相似性度量，并将其定义为两节点连接边的权重与两节点所有连边权重之和的比值. 节点对边权强度越大，表明两节点连接越紧密，两者相似度越高.

基于以上定义，计算网络中相连节点间的相似度，将当前节点和与其拥有最大相似度的邻居节点快速聚类，从而实现社区的初步划分. 之后，逐步合并能使网络模块度增大的两个社区，直到形成最终的社区结构.

2.2　相关定义

为准确描述算法，给出如下相关定义及说明：无向加权网络 $G=(V,E,W)$，V 为顶点集，E 为边集，W 为权重集合. $\forall x,y\in V$，$\Gamma(x)$ 代表节点 x 的邻居节点集合，e_{xy} 代表节点 x 与节点 y 之间的连边，w_{xy} 代表连边 e_{xy} 的权重.

定义 1　节点的强度. 设 $G=(V,E,W)$，$\forall x\in V$，将节点 x 的强度定义为与该节点相连的所有边

的权重之和,如式(1)所示,记作 $s(x)$.

$$s(x) = \sum_{z\in\Gamma(x)} w_{xz} \tag{1}$$

节点强度描述了该节点与周围网络连接的紧密程度.当加权网络退化为无权网络时,所有边权均为 1,此时节点的强度就等于节点的度.

定义 2 节点的单位权重.设 $G=(V,E,W)$, $\forall x\in V$,将节点 x 的单位权重定义为该节点所有连边权重的平均值,如式(2)所示,记作 $u(x)$.

$$u(x) = \frac{s(x)}{k(x)} \tag{2}$$

其中,$s(x)$为节点 x 的强度;$k(x)$为节点 x 的度,也即与节点 x 相连的边数.

定义 3 共邻节点作用系数.设 $G=(V,E,W)$,$x,y\in V$,$\forall\ z\in V\cap\Gamma(x)\cap\Gamma(y)$,将共邻节点 z 对于节点对$<x,y>$的作用系数定义为两节点与其共邻节点两条连边的权重之和与该节点对与其所有邻居节点连边的权重之和的比值,如式(3)所示,记作 $\varepsilon_z^{CN}(x,y)$.

$$\varepsilon_z^{CN}(x,y) = \frac{w_{xz}+w_{zy}}{s(x)+s(y)-w_{xy}} \tag{3}$$

定义 4 共邻节点连接强度.设 $G=(V,E,W)$,$x,y\in V$,$\forall\ z\in V\cap\Gamma(x)\cap\Gamma(y)$,将共邻节点 z 对于节点对$<x,y>$的连接强度定义为该共邻节点的单位权重与其作用系数的乘积,如式(4)所示,记作 $\mathrm{Sim}_z^{CN}(x,y)$.

$$\mathrm{Sim}_z^{CN}(x,y) = u(z)*\varepsilon_z^{CN}(x,y) \tag{4}$$

其中,$u(z)$为节点 z 的单位权重,$\varepsilon_z^{CN}(x,y)$为共邻节点 z 对于节点对$<x,y>$的作用系数.共邻节点连接强度描述了该共邻节点对其所连接的两个节点的相似性贡献.共邻节点的单位权重及作用系数越大,该共邻节点对其所连接的两节点相似性贡献越大,两节点相似度越高.

定义 5 节点对边权强度.设 $G=(V,E,W)$, $\forall x,y\in V$,将节点对$<x,y>$的边权强度定义为两节点连接边的权重与两者和其所有邻居节点连边的权重之和的比值,如式(5)所示,记作 $sw(x,y)$.节点对的边权强度度量了两节点之间连接边的权重在两节点所有邻居节点连接边的权重之和中所占的比重.该值越大意味着两节点间连接紧密强度越大,两者相似度越高.

$$sw(x,y) = \frac{w_{xy}}{s(x)+s(y)-w_{xy}} \tag{5}$$

定义 6 基于共邻节点的加权相似度.设 $G=(V,E,W)$,$\forall$ x,$y\in V$,将节点对$<x,y>$基于共邻节点的加权相似度定义为该节点对所有共邻节点连接强度之和,如式(6)所示,记作 Sim_{xy}^{IEM}.

$$\mathrm{Sim}_{xy}^{IEM} = \begin{cases} 0, & e_{xy}\notin E \\ sw(x,y), & e_{xy}\in E\wedge\Gamma(x)\cap\Gamma(y)=\Phi \\ \sum\limits_{z\in\Gamma(x)\cap\Gamma(y)} \mathrm{Sim}_z^{CN}(x,y), & e_{xy}\in E\wedge\Gamma(x)\cap\Gamma(y)\neq\Phi \end{cases} \tag{6}$$

3 IEM 算法

3.1 算法描述

基于上述加权相似度指标的定义,采用层次聚类的思想,提出基于共邻节点相似度的加权网络社区发现算法 IEM.算法主要包括三大部分:形成初始社区、扩展社区和合并社区,具体步骤如下.

I:形成初始社区.

(1) 给定加权网络图,按照式(6)计算网络中任意节点对之间的相似度,存于矩阵中.将每个节点看做一个社区,随机选择一个节点 v_i 作为起始节点,并将其置为当前节点.

(2) 找到与当前节点拥有最大相似度的节点 v_j,将包含 v_j 的社团与包含当前节点的社团合并,形成一个社区,并将其作为当前社区,同时将 V_j 作为当前节点.

II:扩展社区.

(3) 找到与当前节点具有最大相似度的节点 v_k 作为下一个要进行聚类处理的节点.如果 v_k 不属于当前社区,则将 v_k 聚类至当前社区,同时将 v_k 置为当前节点。否则意味着 v_k 已经被包含在当前社区中,此时,随机选择一个未被访问过的节点作为当前节点,并跳转到步骤(2).

(4) 重复执行步骤(3),直到所有节点被访问过,得到社区扩展后网络的初始划分结果.

Ⅲ:合并社区.

在上述节点聚类及社区扩展两步操作中,若网络中存在一些节点对,与该节点对中其中一个节点拥有最大相似度的节点均为该节点对中的另一个节点,则聚类结果会形成许多由这些两两节点组成

的小社区，导致网络模块度较低. 此时，采用最大化模块度的思想对社区的初始划分结果进行优化.

(5) 计算当前网络的模块度 Q_{current}.

(6) 针对当前网络的社区结构，计算合并当前任意两社区之后网络的模块度来更新模块度矩阵 Q，即 Q_{ij} 为合并第 i 个社区和第 j 个社区之后所得网络的模块度.

(7) 取 Q_{ij} 的最大值 $\max(Q_{ij})$.

(8) 若 $\max(Q_{ij})>Q_{\text{current}}$，则合并对应的两个社区 i 和社区 j，之后更新合并后的社区结构，并将 $\max(Q_{ij})$ 赋值给 Q_{current}.

(9) 重复执行步骤(6)～(8)，直到 $\max(Q_{ij})\leqslant Q_{\text{current}}$，代表合并当前任意两社区之后网络的模块度都小于当前网络的模块度，意味着合并操作结束，最终社区结构形成.

3.2 算法实现

算法 1　IEM algorithm：

输入：无向加权网络图 $G=(V,E,W)$

输出：G 的社区划分结果 $G=\{C_1,C_2,\cdots,C_p\}$，满足 $V=C_1\cup C_2\cup\cdots C_p$ 且 $\forall\ i\neq j,\ C_i\cap C_j=\varnothing$

1) ReadFile from dataset. txt;

2) Return adjMatrix A;

3) for each $v_i\in V$ do get $\Gamma(v_i)$ endfor;

4) New a SimMatrix S;

5) for i, j=1 to n do $S(i,j)=S_{ij}^{IEM}$ endfor;

6) for each $v_i\in V$ do new List Slist, simlist(v_i);

7) for each $v_j\in\Gamma(v_i)$ do

get $\max(S(i,j))$; simlist(v_i). add(v_j)

endfor;

8) SList. add(v_i, simlist(v_i)) endfor;

9) p=0;

10) select a node $v_i\in V$ where v_i. visited=false; currentnode=v_i; v_i. visited=true;

11) p++;

12) new a List community C_p;

13) C_p. add(currentnode);

14) get node_max from SList(currentnode, simlist(currentnode)); v_j=node_max;

15) C_p. add(v_j); currentnode=v_j; v_j. visited=true;

16) get node_max from SList; v_k = node_max;

17) if v_k not in C_p{Cp. add (v_k) currentnode =v_k and goto 16)};

18) else goto 10);

19) until all node v_n in V v_n. visited=true;

20) get $G_{\text{current}}=\{C_1,C_2,\cdots,C_p\}$;

21) Cal $Q_w(G_{\text{current}})$;

22) Initialize Matrix Q;

23) for i, j=1 to p do

$Q_{ij}=Q_w(G.\text{Merge}(C_i,C_j))$ endfor;

24) if $\max(Q_{ij})>Q_w(G_{\text{current}})$

{$C_i=C_i\cup C_j$; $Q_w(G_{\text{current}})=\max(Q_{ij})$; goto 23)};

25) else Return $G=\{C_1,C_2,\cdots,C_p\}$.

4　实验与分析

在多个模拟及真实加权网络上对算法的正确性和有效性进行了验证，以加权模块度为评价算法社区划分质量的指标，与加权 CN、加权 AA、加权 RA 以及 CRMA 算法进行了划分结果的对比分析，结果显示 IEM 算法对于加权网络的社区划分效果优于其它算法.

4.1 数据集描述

为验证 IEM 算法性能，采用了 1 个仿真数据集，并从 http://konect. uni-koblenz. de/networks/下载得到 5 个真实加权网络进行了实验分析，各数据集信息如下：

(1) 仿真数据集：来自文献[22]中的一个仿真加权网络，共包含 36 个节点和 70 条边，权重范围为(0,1]，其拓扑结构如图 1 所示.

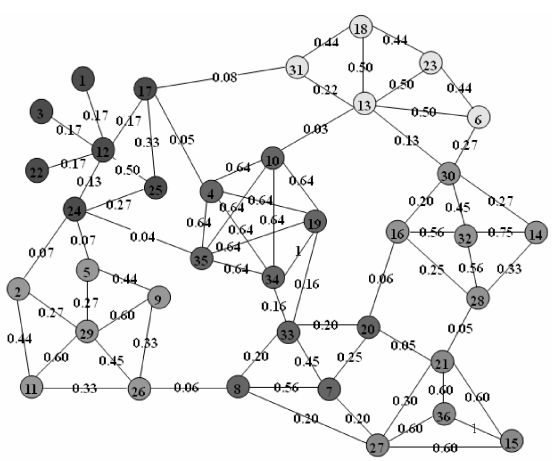

图 1　一个仿真加权全正网络[22](IEM 社区划分结果)

Fig. 1　An artificial weighted network[22]

(2) Zachary's Karate club：由 Wayne Zachary 从 1970～1972 年间观察了美国一所大学空

手道俱乐部成员之间的交往情况,经数据汇总构造出的成员关系网络,共包含 34 个节点和 78 条边.其中,节点代表俱乐部成员;边表示两个成员在俱乐部外私下交往较为密切;权重代表两成员间交往的密切程度.

(3) Les Miserables 人物关系网络:由 Knuth 通过对长篇小说《悲惨世界》中的人物关系进行分析,统计出同时出现的人物,最终构成的一个人物关系网络,共包含 77 个节点和 254 条边.其中,每个节点代表一个人物;节点之间有连边表示两个人物在同一场景中出现过;权重表示两个人物同时出现的次数.

(4) Train Bombing:是 2004 年 3 月发生在西班牙首都马德里的火车爆炸案恐怖分子关系网络,共包含 64 个节点和 243 条边.其中,节点代表恐怖分子;边代表恐怖分子在爆炸案中有协作、信任或通讯联系等关系;权值代表恐怖分子接触的频率.

(5) US Airport:是一个美国航空运输网络,共包含 332 个节点和 2126 条边.其中,节点代表机场;边代表两个机场之间有一条航线;权重代表两个机场之间的飞行次数.

(6) Net Science:科研人员之间关于合作发表论文的科学家合作网络,共包含 379 个节点和 914 条边.其中,节点代表研究人员;边代表研究人员之间的合作关系,如曾经一起发表过论文等;权重代表两个科研人员之间的合作次数.

4.2 评价指标

在社会网络社区发现研究中,模块度是一种被广泛采纳的、衡量社区结构优劣的度量指标,其值通常在 0.3~0.7 之间.对于网络的某个社区划分而言,模块度越大表示网络社区结构的强度越强,也即划分结果越准确.本文采用文献[19]中所述的加权网络模块度 Q_w 作为度量指标对算法社区划分性能进行了评价分析,其定义如式(7)所示.

$$Q_w = \frac{1}{2W}\sum_{ij}\left(w_{ij} - \frac{w_i w_j}{2W}\right)\delta(C_i, C_j) \qquad (7)$$

其中,w_{ij} 为节点 v_i 与 v_j 连接边的权重;$w_i = \sum_j w_{ij}$ 和 $w_j = \sum_i w_{ij}$ 分别为节点 v_i 与 v_j 的权重,即前文所述的节点的强度;$W = \sum_{ij} w_{ij}$ 为整个网络中所有连边的权重之和;$\delta(C_i, C_j)$ 为罚函数,当 v_i 与 v_j 属于同一社区时,其值为 1,否则为 0.若网络中所有边权都为 1,则加权模块度 Q_w 将与最原始的无权网络模块度 Q 相同.

4.3 实验结果对比分析

针对上述数据集,将本文所提 IEM 算法与文献[19]中所述的 3 个基准相似性指标加权 CN、加权 AA 和加权 RA 以及本文作者在文献[21]中针对 AGMA 算法改进之后所提的 CRMA 算法进行了社区划分结果的实验对比.5 种算法在 6 个数据集上所得的社区划分结果见表 1.其中,第 1 列列出了实验所用的 6 个加权网络数据集;第 2 列列出了每个网络的性质,其中 $|V|$ 代表节点数,$|E|$ 代表边数;第 3 列至第 7 列列出了实验中对比的 5 个算法,其社区发现结果用"社区数目/加权模块度"来表示.

表 1 5 种算法在 6 个数据集上社区划分实验结果

Tab.1 Community discovery results of five algorithms in six data sets

	$\|V\|$ / $\|E\|$	加权 CN	加权 AA	加权 RA	CRMA	IEM
(1)仿真数据集	36 / 70	7 / 0.8039	7 / 0.8039	7 / 0.8039	7 / 0.8039	7 / 0.8039
(2) Karate Club	34 / 78	2 / 0.4547	2 / 0.4547	2 / 0.4547	2 / 0.4547	4 / 0.4950
(3) Les Miserables	77 / 254	1 / 0.0350	3 / 0.4185	3 / 0.4577	9 / 0.5222	5 / 0.5427
(4) Train Bombing	64 / 243	1 / 0.0303	4 / 0.3626	4 / 0.3604	4 / 0.4420	5 / 0.4579
(5) US Airport	332 / 2126	2 / 0.0174	3 / 0.0987	3 / 0.1039	4 / 0.1347	4 / 0.1932
(6) Net Science	379 / 914	8 / 0.6045	19 / 0.8453	18 / 0.8499	21 / 0.8430	19 / 0.8512

从表 1 可知:

(1) 对于第一个数据集,5 种算法社区划分结果完全相同,如图 1 所示.在此说明,在本文所有图中,根据算法划分结果,位于不同社区的节点分别用不同的颜色表示,以示区别.

(2) 对于 Zachary's Karate club 网络,前四个算法社区划分结果完全相同,都将其划分为 2 个社区,如图 2 所示.而 IEM 算法将其划分为 4 个社区,如图 3 所示,网络模块度对比其它四个算法提高了 11.11%.这里,需要强调的是网络模块度的值以及网络中社区的数目,会因不同的算法或者不同的实现方法而有所不同.一般情况下,较大的模

块度的值对应于相对准确的社区数目以及更接近于真实的社区结构.

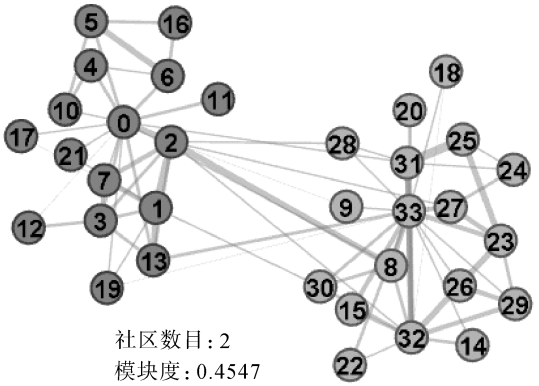

图 2　CRMA(加权 CN/AA/RA)对 Karate club 划分结果

Fig. 2　Communities division results of CRMA (weighted CN/AA/RA) in Karate Club

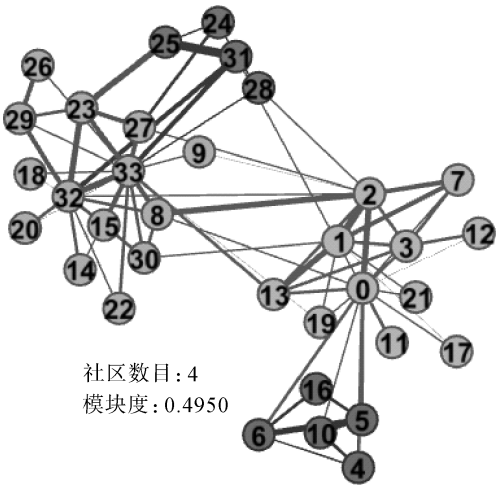

图 3　IEM 算法对 Karate club 划分结果

Fig. 3　Communities division results of IEM in Karate Club

(3) 对于 Les Miserables 和 Train Bombing 两个网络，5 种算法对应的社区划分结果各不相同. 这两个数据集都较为稀疏，导致仅考虑与共同邻居节点连边权重的加权 CN 算法在此类数据集上划分效果欠佳. 加权 AA 和加权 RA 两种算法在这两个数据集上社区划分结果及模块度相差较小，CRMA 与 IEM 算法划分质量明显优于前三种算法. 以加权模块度为评价标准，IEM 算法针对这两个网络的社区划分质量最高.

(4) 对于 US Airport 网络，5 种算法对应的社区划分结果也各不相同，且模块度都较低. 该数据集图密度为 0.039，网络中节点的加权平均度数只有 0.924. 由于网络中许多节点对没有共同邻居，导致 5 种算法针对此数据集划分结果都欠佳. 但总体而言，IEM 算法社区划分结果最佳.

(5) 对于 Net Science 而言，虽然数据集较为稀疏，但网络中节点平均度数为 4.823，平均强度为 2.583，聚类系数为 0.798，因此各算法社区划分效果整体都较好. 其中，加权 CN 划分质量最差，在社区数目和模块度上与其它算法差异较大. 其余 4 种算法划分结果稍有差异，模块度也相差较小. 从实验统计结果来看，IEM 算法针对此数据集的社区划分质量优于前 4 种算法，进一步验证了 IEM 算法综合共邻节点的度、强度、连接强度等定义两节点相似度的正确性.

此外，针对实验所用数据集，IEM 算法划分效果都优于本文作者之前所提的 CRMA 算法. 两种算法针对第 3 个至第 6 个数据集的社区划分结果分别见图 4～图 11.

由于 CRMA 算法是针对 AGMA 算法的改进，考虑了连边符号信息，因此该算法在加权符号网络中能够获得更好的划分效果. 虽然 CRMA 算法仍适用于加权全正网络，但是每种算法都是针对其目标设计实现的，在不同性质的数据集上算法的划分效果会有所不同. 就仅包含正连接的加权网络而言，IEM 社区划分质量较 CRMA 算法更加合理有效.

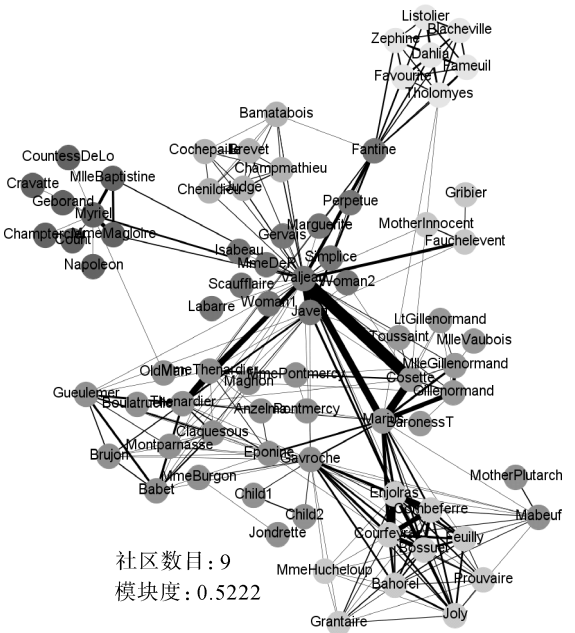

图 4　CRMA 算法对 Les Miserables 划分结果

Fig. 4　Division results of CRMA in Les Miserables

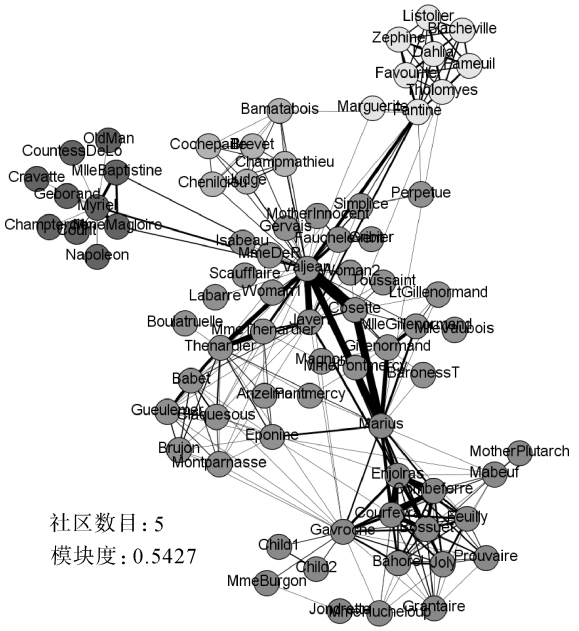

图 5 IEM 算法对 Les Miserables 划分结果
Fig. 5 Division results of IEM in Les Miserables

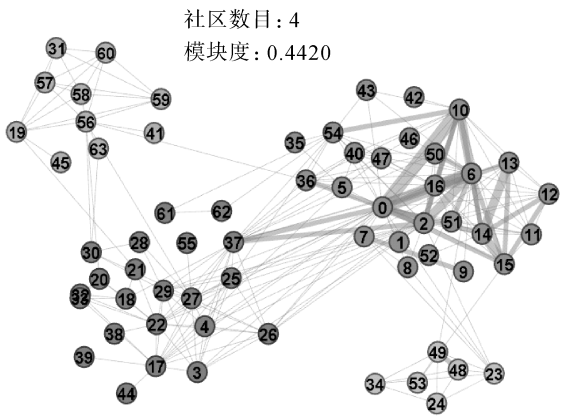

图 6 CRMA 算法对 Train Bombing 划分结果
Fig. 6 Division results of CRMA in Train Bombing

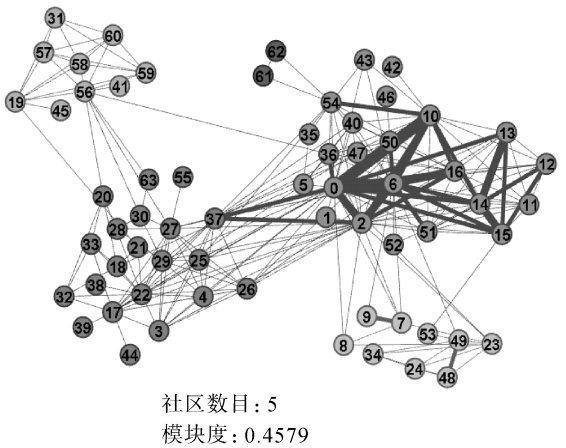

图 7 IEM 算法对 Train Bombing 划分结果
Fig. 7 Division results of IEM in Train Bombing

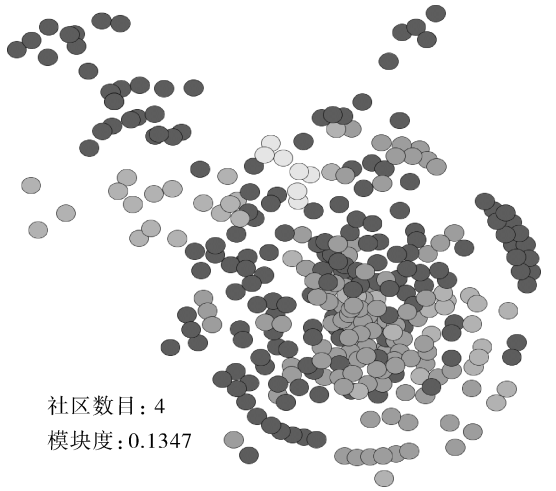

图 8 CRMA 算法对 US Airport 划分结果
Fig. 8 Division results of CRMA in US Airport

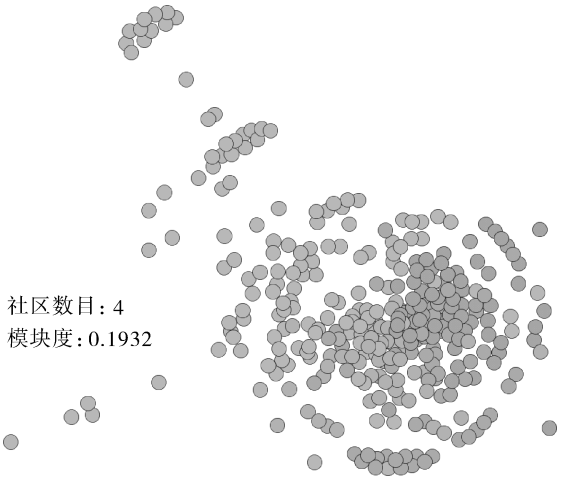

图 9 IEM 算法对 US Airport 划分结果
Fig. 9 Division results of IEM in US Airport

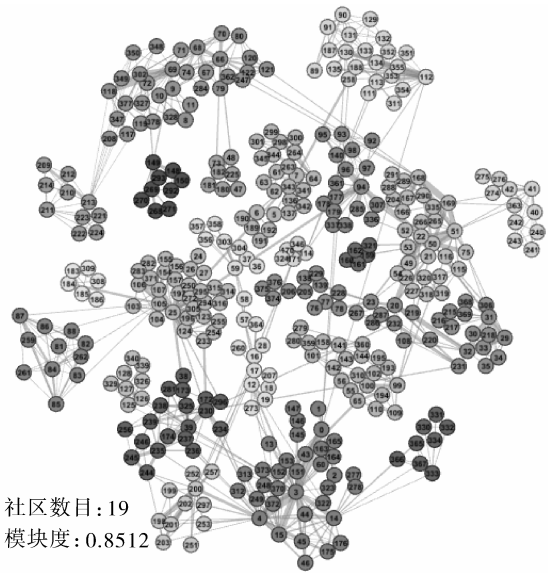

图 10 CRMA 算法对 Net Science 划分结果
Fig. 10 Division results of CRMA in Net Science

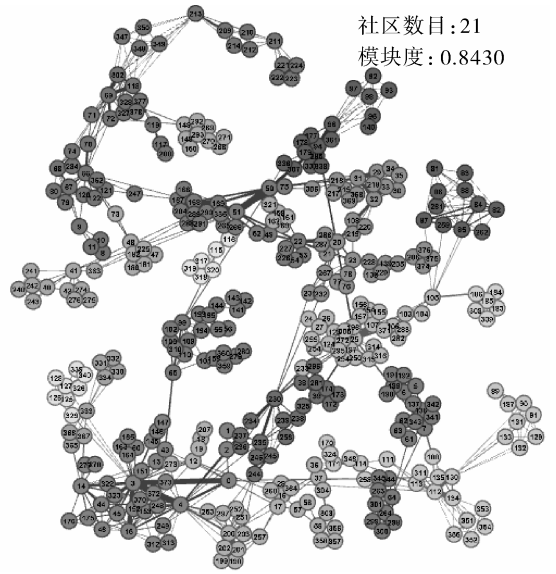

图 11　IEM 算法对 Net Science 划分结果
Fig. 11　Division results of IEM in Net Science

总而言之，IEM 算法优于其他四种算法. 虽然，该算法针对后 5 个网络的社区划分结果与其他 4 种算法不同，但模块度都是最高的. 且算法针对 Karate club、US Airport 和 Net Science 三个网络的社区划分结果与文献[23]中所述的 fastgreedy、betweenness、label 等经典社区发现方法针对这 3 个数据集的划分结果相比，在社区数目和模块度上基于吻合. 即，以加权模块度为评价指标，IEM 算法社区划分效果最佳.

4.4　算法复杂度分析

IEM 算法使用矩阵存储图边关系，空间复杂度为 $O(n^2)$. 计算了网络中直接相连的任意节点对之间的相似性值并存至相似度矩阵中，计算复杂度为 $O(m)$，空间复杂度为 $O(n^2)$. 其次，遍历相似性矩阵，使用 list(i,arraylist)存储与网络中每个节点 v_i 相似度最高的节点编号，以便在算法前两个阶段快速从 list 中提取到连接紧密度最大的节点对的编号完成聚类，计算复杂度为 $O(n\log_2 n)$，空间复杂度为 $O(2n)$. 最后，社区合并阶段，计算了初始划分结果中 p 个社区两两合并后的网络模块度，复杂度为 $O(p^2)$. 与其它几种算法相比，本算法计算复杂度略微有所提高，但算法在达到较高社区划分准确率的前提下，仍可保证时间上的可行性和有效性.

5　结　论

有效捕获两节点的共同邻居属性对于两者的相似性影响，提出基于共邻节点相似性的加权网络社区发现方法 IEM. 依据相似度的定义完成当前节点与其邻居节点的聚类，形成初始社区并进行社区扩展，最后基于模块度优化的思想对社区进行合并. 改进了加权 CN、加权 AA、加权 RA 等相似性指标的不足之处，实验结果显示所提算法对于加权网络社区划分的有效性和较高的划分质量. 针对大规模数据集，改进算法，降低社区合并阶段的计算复杂度以提高算法执行效率是下一步的研究内容. 此外，鉴于模块度在社区划分方面存在的分辨率问题，综合采用其它度量标准对算法性能进行有效评估也是下一步工作.

参考文献：

[1]　Newman M E J. Analysis of weighted networks [J]. Phys Rev E, 2004, 70: 56133.

[2]　任迎伟，李静. 创业过程组织社会网络动态演进机理研究[J]. 四川大学学报：哲学社会科学版，2013，188：104.

[3]　郭崇慧，张娜. 基于共邻矩阵的复杂网络社区结构划分方法[J]. 系统工程理论与实践，2010，30：1077.

[4]　Jaho E, Karaliopoulos M, Stavrakakis I. ISCoDe: a framework for interest similarity-based community detection in social networks [C]// Proceedings of 2011 IEEE Conference on Computer Communications Workshops. Shanghai, China: IEEE INFOCOM, 2011.

[5]　Subramani K, Velkov A, Ntoutsi I, *et al*. Density-based community detection in social networks[C]// Proceedings of 2011 IEEE 5th International Conference on Internet Multimedia Systems Architecture and Application. Bangalore, India: IEEE IMSAA, 2011.

[6]　Liu R F, Feng S, Shi R S, *et al*. Weighted graph clustering for community detection of large social networks [J]. Procedia Comput Sci, 2014, 31: 85.

[7]　Sharma T. Finding communities in weighted signed social networks[C]//Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Istanbul, Turkey: IEEE ASONAM, 2012.

[8]　Lu Z, Wen Y, Cao G. Community detection in weighted networks: algorithms and applications [C]// Proceedings of 2013 IEEE International Conference on Pervasive Computing and Communications. San Dieg: IEEE, 2013, 26: 179.

[9] Han H, Wang J, Wang H. A new algorithm to detect community structures on weighted network [C]// Proceedings of 2010 International Conference on Computational Intelligence and Software Engineerin. Wuhan: IEEE, 2010: 1.

[10] 王坤, 吕光宏, 梁召伟, 等. 基于相似度的加权复杂网络社区发现方法[J]. 四川大学学报:自然科学版, 2014, 51: 1170.

[11] 王佳嘉. 动态复杂网络社区发现算法研究及应用[D]. 大连: 大连理工大学, 2014.

[12] Chen D, Shang M, Lv Z, *et al*. Detecting overlapping communities of weighted via a local algorithm [J]. Phys Stat Mech Appl, 2010, 389: 4177.

[13] 林旺群, 卢风顺, 丁兆云, 等. 基于带权图的层次化社区并行计算方法[J]. 软件学报, 2012, 3: 1517.

[14] 王书凯. 基于交互模块度的带权图网络社区发现[D]. 昆明: 云南大学, 2014.

[15] 詹培森. 加权复杂网络的局部社区发现算法并行化研究与实现[D]. 广州: 华南理工大学, 2013.

[16] 姚宏亮, 罗明伟, 李俊照, 等. 复合加权股票网络的活跃性层次聚类[J]. 计算机科学与探索, 2014, 8: 207.

[17] Saha T, Domeniconi C, Rangwala H. Detection of communities and bridges in weighted networks [C]// Proceedings of Machine Learning & Data Mining in Pattern Recognition. New York, USA: Petra Perner, 2011.

[18] 赵健, 安健. 群智感知服务中一种面向有向加权网络的社区发现算法[J]. 计算机应用研究, 2014, 31: 3795.

[19] 季大祥. 社交网络中的社团发现与度量研究[D]. 济南: 山东大学, 2014.

[20] 姚尊强. 加权复杂网络的分析和预测[D]. 青岛: 青岛理工大学, 2012.

[21] Guo J F, Liu M M, Liu L L, *et al*. An improved community discovery algorithm in weighted social networks [J]. ICIC Express Lett, 2016, 10: 35.

[22] Yang B, Liu J M. An autonomy oriented computing approach to distributed network community mining [C]// Proceedings of the First International Conference on Self-Adaptive and Self-Organizing Systems. Boston, USA: IEEE SASO, 2007.

[23] 刘旭. 基于目标函数优化的复杂网络社区结构发现[D]. 长沙: 国防科技大学, 2012.