软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(11):3103−3114 [doi: 10.13328/j.cnki.jos.005347]

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

一种加权稠密子图社区发现算法*

杨 贵[1], 郑文萍[1,2], 王文剑[1,2], 张浩杰[2]

[1](山西大学 计算机与信息技术学院,山西 太原 030006)

[2](计算智能与中文信息处理教育部重点实验室(山西大学),山西 太原 030006)

通讯作者: 王文剑, E-mail: wjwang@sxu.edu.cn

摘 要: 目前,针对复杂网络的社区发现算法大多仅根据网络的拓扑结构来确定社区,然而现实复杂网络中的边可能带有表示连接紧密程度或者可信度意义的权重,这些先验信息对社区发现的准确性至关重要.针对该问题,提出了基于加权稠密子图的重叠聚类算法(overlap community detection on weighted networks,简称 OCDW).首先,综合考虑网络拓扑结构及真实网络中边权重的影响,给出了一种网络中边的权重定义方法;进而给出种子节点选取方式和权重更新策略;最终得到聚类结果.OCDW 算法在无权网络和加权网络都适用.通过与一些经典的社区发现算法在 9 个真实网络数据集上进行分析比较,结果表明算法 OCDW 在 *F* 度量、准确度、分离度、标准互信息、调整兰德系数、模块性及运行时间等方面均表现出较好的性能.

关键词: 复杂网络;社区发现;图聚类;重叠聚类;稠密子图

中图法分类号: TP311

中文引用格式: 杨贵,郑文萍,王文剑,张浩杰.一种加权稠密子图社区发现算法.软件学报,2017,28(11):3103−3114. http://www.jos.org.cn/1000-9825/5347.htm

英文引用格式: Yang G, Zheng WP, Wang WJ, Zhang HJ. Community detection algorithm based on weighted dense subgraphs. Ruan Jian Xue Bao/Journal of Software, 2017,28(11):3103−3114 (in Chinese). http://www.jos.org.cn/1000-9825/5347.htm

Community Detection Algorithm Based on Weighted Dense Subgraphs

YANG Gui[1], ZHENG Wen-Ping[1,2], WANG Wen-Jian[1,2], ZHANG Hao-Jie[1]

[1](School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)

[2](Key Laboratory of Computation Intelligence and Chinese Information Processing (Shanxi University), Ministry of Education, Taiyuan 030006, China)

Abstract: Most community detection algorithms in complex networks find communities based on topological structure of the network. Some important information is included in real network data, which represents data reliability or link closeness. Combined these prior information to detect communities might obtain better clustering results. An overlapping community detection on weighted networks (OCDW) is proposed in this study. Edge weight is defined by combining network topological structure and real information. Then, vertex weight is induced by edge weight. To obtain cluster, OCDW selects seed nodes according to vertex weight. After finding a cluster, edges in this cluster reduce their weights to avoid being selected as a seed node with high probability. Compared with some classical algorithms on 9 real networks including 5 unweighted networks and 4 weighted networks, OCDW shows a considerable or better performance on *F*-measure, accuracy, separation, NMI, ARI, modularity and time efficiency.

Key words: complex network; community detection; graph clustering; overlapping clustering; dense subgraph

* 基金项目: 国家自然科学基金(61673249, 61572005); 山西省回国留学人员科研基金(2016-004, 2017-014)

Foundation item: National Natural Science Foundation of China (61673249, 61572005); Shanxi Scholarship Council of China (2016-004, 2017-014)

本文由复杂环境下的机器学习研究专刊特约编辑胡清华教授、张道强教授、张长水教授推荐.

收稿时间: 2017-05-14; 修改时间: 2017-06-16; 采用时间: 2017-08-23

近年来,对各种复杂网络的研究[1−3]是许多领域的研究热点之一,如生物网络、社交网络、电子邮件网络、引文网络等已成为众多学者的主要研究对象.复杂网络中的社区发现不仅有助于深入研究整个网络的功能模块及其演化,而且能够准确地理解并分析复杂系统的拓扑结构及动力学特性,因此具有十分重要的理论意义和应用价值[4,5].

目前,针对复杂网络中的社区发现问题,研究较多的是无向无权网络,包括基于图划分的聚类算法[6]、基于谱分析的聚类算法[7]、基于层次的聚类算法[8]、基于密度的聚类算法[9,10]等.其中,基于密度的聚类算法通过搜索网络中稠密子图[11],能够较好地发现网络中的功能模块,在社区发现中得到了广泛应用.2003 年 Bader 等人提出的 MCODE 算法[12]、2005 年 Palla 等人[13]提出的派系过滤算法(clique percolation method,简称 CPM)、2006 年 Saito 等人[14]提出的 *k*-dense 算法、2009 年 Shen 等人提出的 EAGLE 算法[15]等将网络拓扑结构作为社区划分的依据,对无权网络进行社区划分.

然而,现实复杂网络中的边或顶点中往往包含有一些重要的先验信息,如,高通量实验所得到的蛋白质互作用网络中,边权重往往代表实验可信度;合作网络中的边权重通常代表合作对象的连接紧密程度.2008 年,Blondel 等人基于模块性最优化提出了启发式算法 BGLL[16].2009 年,Liu 等人[17]提出的 CMC 算法通过迭代评分的方法给每条边赋予权重,然后通过搜索网络中的所有极大团并计算其加权密度,最后依据极大团的加权密度定义极大团之间的相互连接度,进而合并极大团得到网络中的社区结构.2011 年,Lee 等人[18]提出的 MDOS 算法在加权网络中通过选取种子节点,并依据子图函数逐步扩展得到稠密子图.Wang 等人基于边聚集系数局部度量提出了适用于无权网络和加权网络的 HC-PIN 算法[19].2014 年,Ren 等人[10]也通过定义边权重将无权网络转化为加权网络,然后结合子图的密度性和模块性定义适应度函数,给出了基于局部适应度的 LF_PIN 算法.但是,这些学者仅考虑基于拓扑结构定义网络中边的权重或者仅考虑网络中具有现实意义的边权重进行社区发现.如何将具有现实意义的边权重与网络拓扑结构相结合,作为社区发现算法的依据,进而发现复杂网络中的重叠社区,是当今社区发现算法研究热点之一.

针对加权网络的社区发现问题,提出了一种基于种子扩展策略的重叠社区发现算法(overlapping community detection on weighted networks,简称 OCDW).首先给出了一种综合考虑网络自身边信息和网络拓扑结构信息的边权重定义方式,并基于此定义节点的加权度,选择权重最大的节点作为社区种子节点;进而给出了基于社区评估函数的种子扩展策略和权重更新方式,迭代得到加权稠密子图;最后,将重叠率比较高的稠密子图合并为加权中心社区,并根据节点对社区的归属度 $b(v,C)$来度量节点和社区的连接倾向程度,将未聚类节点分配到加权中心社区中,从而得到最终的社区发现结果.

1 背景知识

通常,一个复杂网络可以用图 $G=(V,E)$来表示,其中,节点集 $V=\{v_1,v_2,\ldots,v_n\}$,$n=|V|$;边集 E 中每条边 $e_{i,j}$ 对应 V 中一对顶点(v_i,v_j)之间的连接关系,$m=|E|$.节点 v 的邻域 $N_G(v)=\{u|(u,v)\in E\}$,在不引起混淆的情况下,简记为 $N(v)$.$N(v)$中的节点称为节点 v 的相邻点.节点 v 的度记为 k_v.除非特别指明,本文仅考虑简单无向图,即 $k_v=|N(v)|$.令 $U\subseteq V(G)$,用$[U]_G$ 表示 G 的节点子集 U 的导出子图,记为$[U]$.令 $N_G(U)=\bigcup_{x\in U}N_G(x)$ 表示顶点子集 U 在图 G 中的邻域.令 $N_G^-(U)=N_G(U)\setminus U$.图 1 给出了两个典型复杂网络实例.

为了合理地对复杂网络中的节点进行社区发现,将顶点间的相似性作为衡量节点连接紧密程度的重要标准.Jaccard 度量[20]认为:网络图中两节点间的公共邻接点越多,其在结构上就越相似,连接也就越紧密,如公式(1)所示.

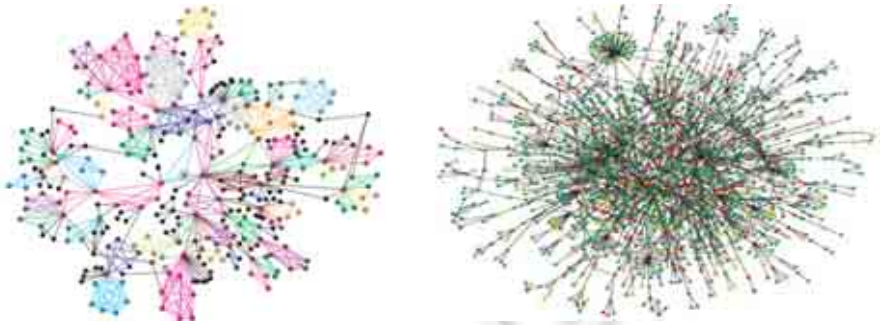
$$Jaccard(v_i,v_j)=\frac{|N(v_i)\cap N(v_j)|}{|N(v_i)\cup N(v_j)|} \tag{1}$$

Hub Promoted 和 Hub Depressed 度量[20]考虑偏好连接对连接紧密度的影响,如公式(2)、公式(3)所示.

$$HP(v_i,v_j)=\frac{|N(v_i)\cap N(v_j)|}{\min\{|N(v_i)|,|N(v_j)|\}} \tag{2}$$

$$HD(v_i,v_j)=\frac{|N(v_i)\cap N(v_j)|}{\max\{|N(v_i)|,|N(v_j)|\}} \tag{3}$$

由于复杂网络构建过程中实验方法的偏差会导致网络拓扑结构中有大量噪声,如由高通量实验构建的蛋白质互作用网络中存在大量的假阳性和假阴性数据,为了更准确地表示网络中边的连接紧密程度,应该综合考虑网络拓扑结构和复杂网络中的边信息.



(a) Co-author 合作网络[21] (b) 酿酒酵母蛋白质互作用网络[22]

Fig.1 Two classical complex networks

图 1 两个典型复杂网络实例

2 边权重定义

现实网络中,边的特性可以通过边权重来体现.如社会网络中,边权重可以表示个体之间连接的紧密及强弱关系;科学家合作网络中的边权重,可以表示科学家之间合作的次数;而蛋白质互作用网络中,边权重往往表示由高通量实验所产生的蛋白质互作用边的可信度[23].在网络中进行社区发现,不仅需要考虑网络的拓扑结构,而且也要考虑节点间边连接的现实意义.

为了能够更加准确地度量节点之间的关联强度,需要先对网络中边的权重进行定义,如公式(4)所示.

$$w'(v_i,v_j)=\alpha\cdot\frac{|N(v_i)\cap N(v_j)|^2}{\min\{|N(v_i)|,|N(v_j)|\}^2}+\beta\cdot\frac{|N(v_i)\cap N(v_j)|^2}{\max\{|N(v_i)|,|N(v_j)|\}^2}+(1-\alpha-\beta)\cdot u(v_i,v_j) \tag{4}$$

其中,$N(v_i)$表示节点 v_i 邻接节点的集合,$N(v_i)\cap N(v_j)$表示节点 v_i 和 v_j 的公共邻接节点集合,$u(v_i,v_j)$表示节点 v_i 和 v_j 在真实网络中边的权重.前两项体现了节点 v_i 和 v_j 在网络拓扑结构方面的相似程度,第 3 项体现了边的真实网络权重.将图 G 的边 $e_{v_iv_j}$ 的最终权重定义为公式(5).

$$w(v_iv_j)=\begin{cases}\varepsilon+\frac{1-\varepsilon}{w'_{avg}}\cdot w'(v_iv_j), & v_iv_j\in E(G)\\ 0, & v_iv_j\notin E(G)\end{cases} \tag{5}$$

其中, $w'_{avg}=\dfrac{\sum_{v_iv_j\in E(G)}w'(v_iv_j)}{|E(G)|}$, 常量$\varepsilon(\in[0,1])$用以区分网络中一对顶点之间是否存在边,此处取$\varepsilon$=0.2.

3 加权网络的重叠社区发现算法 OCDW

基于上一节给出的边权重定义,本文给出一种基于“种子节点扩展策略”的社区发现算法 OCDW,其中包括种子节点的选取、种子扩展过程和社区合并与后处理这 3 个基本过程.

- 首先,根据与节点关联的边权重计算节点权重,选择权重较大的节点作为种子节点;
- 其次,以种子节点为初始点,根据节点对子图的适应度函数将种子节点扩展成局部稠密子图;
- 当所有局部稠密子图构造结束后,将这些子图进行合并处理,得到最终的社区发现结果.

3.1 种子节点的选取

种子节点应该位于最终社区对应子图的拓扑中心位置,通常,这类节点在网络中重要性比较高且位于网络中相对稠密的区域中.种子节点的选择要有足够的代表性,能够使得社区发现算法在尽可能少的迭代过程中给出网络中大多数节点的社区归属,因此,种子节点之间在网络结构上应该相距较远.基于这些原则,将网络中节点 $v_i \in V(G)$的加权度作为其权重定义,如公式(6)所示:

$$wd(v_i)=\sum_{v_j \in N(v_i)} w(v_i v_j)\cdot k_j \tag{6}$$

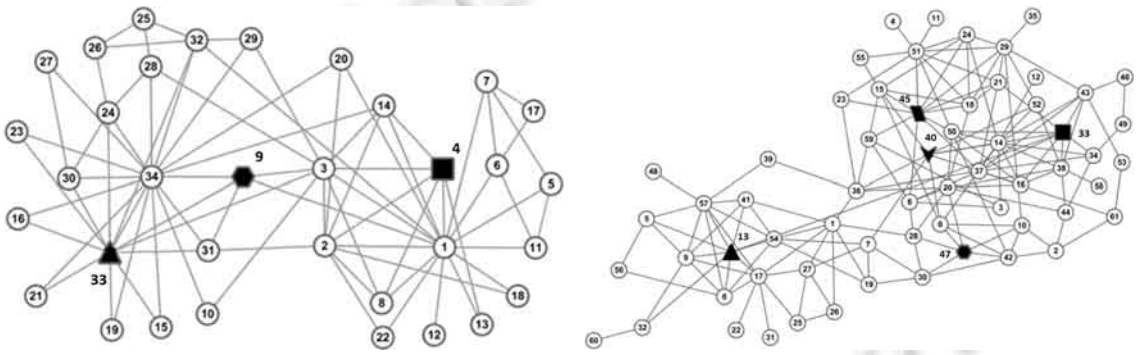
其中,$N(v_i)$表示节点 v_i 的邻接点集合,k_j 表示节点 v_j 的度.节点 v_i 的加权度 $wd(v_i)$与其邻点的度数以及连边的权重成正相关,反映了 v_i 在网络中的重要程度.此处选择权重最大的节点作为第 1 个种子节点.

为使网络中尽可能多的节点有社区归属,在选择下一个社区的种子节点时,应该降低那些已经成为种子的节点选择概率.因此,本算法在找到一个稠密子图之后,减少该子图中的边权重,并重新计算相关节点的加权度.若某节点的加权度变化过大,则不应选作其他社区的种子节点.因此定义节点的加权度变化率,如公式(7)所示.

$$cr(v)=1-\frac{wd'(v_i)}{wd(v_i)} \tag{7}$$

其中,$wd(v)$表示节点 v 在原始网络中的加权度,$wd'(v)$表示节点 v 在更新子图边权重之后的加权度.如果节点 v 的加权度变化率 $cr(v)>\theta$,则节点 v 不能再次被选为种子节点,本文设定 θ=0.3.

图 2 为空手道俱乐部和海豚社交网络在算法执行完之后的所有种子节点,用黑色节点表示.可以看出,图中种子节点选择数量适中且具有较好的代表性.



(a) 空手道俱乐部 (b) 海豚社交网络

Fig.2 Seed nodes in Zachary's karate club and dolphins social network obtained by OCDW

图 2 OCDW 算法在 Zachary 空手道俱乐部和海豚社交网络中得到的种子节点

3.2 种子扩展策略

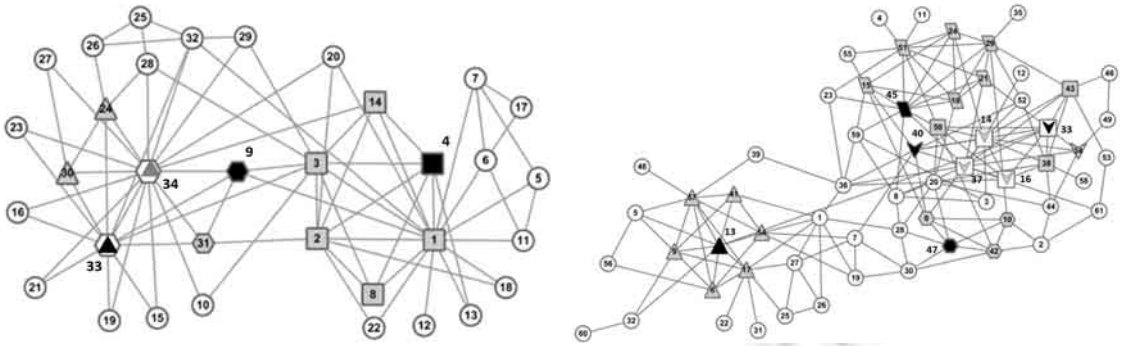
为了得到当前种子节点附近的一个稠密子图,首先给出子图的适应度函数以评价其稠密程度.假设 S 是图 G 的一个连通子图,V_S 表示子图 S 的顶点集,E_S 表示 S 的边集,令 $n_s=|V_S|$,$m_s=|E_S|$.在 S 中添加 $\frac{n_s(n_s-1)}{2}-m_s$ 条边,可以得到一个完全图 S',把新添加的权重设定为图 G 中边的平均权重.通过考虑 S 中已有边与新添边权重的差异来评价 S 的稠密程度.公式(8)给出了子图 S 的社区评估函数 $f(S)$.

$$f(S)=\sum_{v_i v_j \in E(S)} w(v_i v_j)-\frac{1}{2|E(G)|}[n_s(n_s-1)-2m_s]\cdot\sum_{v_i v_j \in E(G)} w(v_i v_j) \tag{8}$$

显然,$f(S)$越大,子图 S 在给定权重意义下越稠密.根据定义可知,若 S 中只有 1 个节点,则 $f(S)$=0.

对节点 $v \notin V_s$,定义 v 对 S 的适应度函数 $\delta_S(v)=f(S\cup\{v\})-f(S)$.在种子节点扩展步骤,将满足 $\delta_S(v)>0$ 且适应度函数最大的节点扩展到当前子图 S 中,此过程迭代进行,直至找不到满足条件的节点,从而得到一个当前稠密子图.图 3 为空手道俱乐部和海豚社交网络通过算法依据子图适应度函数将种子节点扩展得到的稠密子图,不同形

状的灰色节点表示不用的稠密子图.



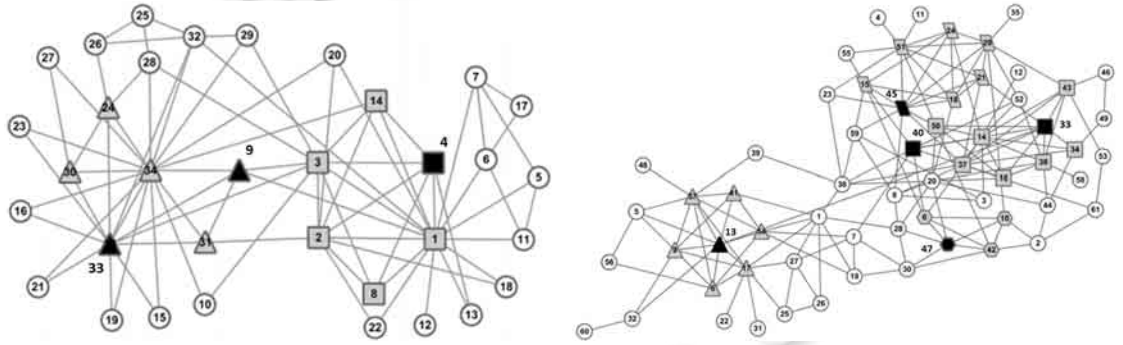
(a) 空手道俱乐部 (b) 海豚社交网络

Fig.3 Weighted dense subgraphs in Zachary's karate club and dolphins social network obtained by OCDW

图 3 OCDW 算法在空手道俱乐部和海豚社交网络中得到的加权稠密子图

3.3 合并优化社区与未聚类节点处理

通过上述步骤之后,我们可以找到很多稠密子图,但是存在一些稠密子图之间重叠节点较多的情况,需要将重叠节点较多的稠密子图进行合并,得到最终社区发现结果.如图 3(a)所示,节点{9,31,33,34}(图中六边形节点)构成的稠密子图与由节点{24,30,33,34}(图中三角形节点)构成的稠密子图之间有一半节点重合,需要进行合并.本文中判断两个稠密子图合并的条件为两个子图间重叠节点数不小于较小稠密子图节点数一半时,将二者合并得到中心社区,称为加权中心社区.图 4(a)和图 4(b)分别给出了空手道俱乐部和海豚社交网络合并后所得到的加权中心社区.



(a) 空手道俱乐部 (b) 海豚社交网络

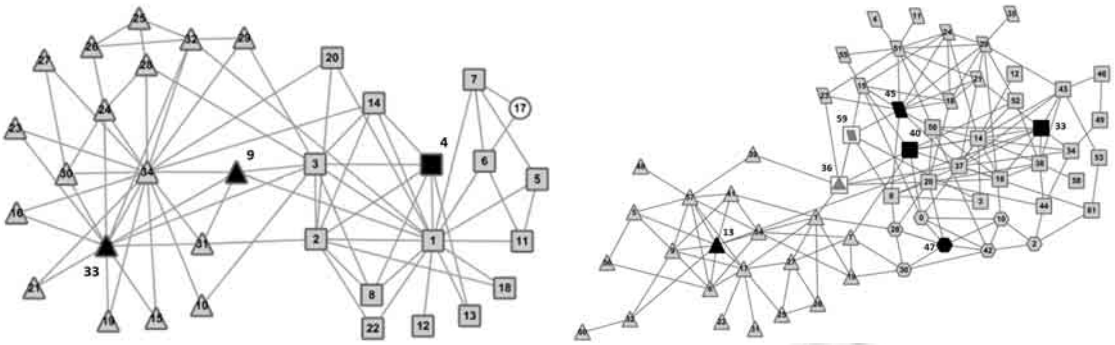
Fig.4 Weighted core communities in Zachary's karate club and dolphins social network obtained by OCDW

图 4 OCDW 算法在空手道俱乐部和海豚社交网络中得到的加权中心社区

OCDW 算法得到的加权中心社区包括了原始网络中的一些局部稠密子图,但并未完全覆盖整个网络,仍然存在一些未聚类节点(如图 4 中的空白节点所示).理想的社区发现结果应该使网络中尽量多的节点有最终的社区归属,因此,根据节点与中心社区连接紧密程度定义未聚类对中心社区的归属度,如公式(9)所示.

$$b(v,C)=\gamma\cdot\frac{\sum_{x\in N(v)\cap C}w(x,v)}{\sum_{x\in C}w(x,v)}+(1-\gamma)\cdot\frac{\sum_{x\in N(v)\cap C}wd(x)}{\sum_{x\in C}wd(x)} \tag{9}$$

基于公式(9)给出的节点对社区归属度的定义,迭代地将未聚类节点划分到已有的中心社区中,对中心社区进行扩展,得到更加合理的聚类结果(详细的中心社区扩展策略见算法 1 的步骤 7 和步骤 8).图 5 给出了空手道俱乐部数据和海豚社交网络的最终聚类结果.



(a) 空手道俱乐部 (b) 海豚社交网络

Fig.5 Final communities in Zachary's karate club and dolphins social network obtained by OCDW

图 5 OCDW 算法在空手道俱乐部和海豚社交网络中得到的最终聚类结果

3.4 算法描述

算法 **1**. 加权网络重叠社区发现算法.

输入:网络 $G=(V,E,U)$,其中,V 和 E 分别表示 G 的节点集和边集,U 是 G 的权重矩阵.

输出:社区集合 $C=\{C_1,C_2,\ldots,C_k\}$.

过程:

步骤 1. 令 $n=|V|,m=|E|,t=0,F=V,\alpha=\dfrac{2m}{n(n-1)}$.

步骤 2. 如果$\alpha<0$,则转步骤 9.令 $C=\varnothing,\beta=0.7-\alpha$.

根据公式(4)、公式(5)计算图 G 中每条边的综合权重矩阵 $W_{n\times n}(G)$.

步骤 3. 根据公式(6)计算 V 中节点的加权度 $wd(v)$,假设 $wd(v_1)\geq wd(v_2)\geq\ldots\geq wd(v_n)$.

步骤 4. 如果 F 为空集,则转步骤 5;否则,令 $t=t+1,C_t=\varnothing$.

步骤 4.1. 选择集合 F 中权重最大的一个节点 v,令 $F=F-\{v\}$,且 $C_t=C_t\cup\{v\}$.

步骤 4.2. 令 $N(C_t)$表示当前集合 C_t 的邻接点集合,对 $N(C_t)$中的每个节点 u,计算 u 对 C_t 的影响度 $\delta_{C_t}(v)=f(C_t\cup\{u\})-f(C_t)$,其中,$f(\cdot)$由公式(8)给出.

步骤 4.3. 令节点 x 是 $N(C_t)$中对 C_t 影响度最大的节点且满足 $\delta_{C_t}(x)>0$,即 $\delta_{C_t}(x)=\max_{v\in N(C_t)}\delta_{C_t}(v)$.

步骤 4.4. 若节点 x 存在,则令 $C_t=C_t\cup\{x\}$,则转步骤 4.3.

步骤 4.5. 若$|C_t|\leq 3$,令 $t=t-1$,转步骤 4.

步骤 4.6. 将 C_t 的导出子图$[C_t]$识别为稠密子图.对$[C_t]$中的每条边 $xy\in E([C_t])$,使得 $w(xy)=\dfrac{w(xy)}{\sqrt{|C_t|}}$.

步骤 4.7. 对每个节点 $x\in C_t$,令 $wd'(x)=\sum_{v\in N(x)}w(vx)\cdot k_v$,根据公式(7)计算节点加权度变化率,若 $cr(x)\leq\theta$,则令 $F=F-\{x\}$;转步骤 4.

步骤 5. 对每一个稠密子图 $C_i\in\{C_1,\ldots,C_t\}$,计算 C_i 与其他子图的重叠度.若 C_i 与子图 $C_j(i<j)$节点的重叠度 $|C_i\cap C_j|/\min\{|C_i|,|C_j|\}\geq 0.5$,则令 $C_i=C_i\cup C_j$.

步骤 6. 对合并后的稠密子图重新编号,得到中心社区集合为 $\{C_1^{(0)},\ldots,C_S^{(0)}\}$.若 $s<1$,则令$\alpha=\alpha-0.03$,转步骤 2.

步骤 7. 令集合 $T^{(0)}=V-\bigcup_{i=1}^{S}C_i$, $\mu_0=0.7,t=0$.

步骤 8. 令 $t=t+1$,对每个社区 $C_1^{(t-1)},\ldots,C_S^{(t-1)}$, 执行以下操作.

步骤 8.1. $C_i^{(t)}=C_i^{(t-1)}(1\leq i\leq s)$, 对任意元素 $v\in N(C_i^{(t-1)})\cap T^{(t-1)}$, 根据公式(9)计算 $b(v,C)$.如果 $b(v,C)\geq\mu_{t-1}$,则 $C_i^{(t)}=C_i^{(t)}\cup\{v\}$.

步骤 8.2. 令$\mu_t=\mu_{t-1}-0.1, T^{(t)}=V-\bigcup_{i=1}^{S}C_i^{(t-1)}$, 若$\mu_t\geq 0.3$ 且 $T^{(t)}\neq\varnothing$,则返回步骤 8.

步骤 9. 结束,输出社区集合 $\{C_1^{(t)},...,C_S^{(t)}\}$.

加权网络重叠社区发现算法 OCDW 给出了基于贪心搜索策略的社区发现算法,其中,步骤 1~步骤 3 实现了对网络中节点和边权重赋初值,平均时间复杂度与网络中的总边数成正比,即 $O(c|E|)$.步骤 4 迭代实现了种子节点的选择与获得中心社区,每次迭代会遍历网络中的所有边,代价为 $O(|E|)$.中心社区扩展迭代次数与中心点的选择密切相关,假设迭代次数为 t,则步骤 4 的总代价为 $O(t|E|)$.步骤 5 实现了中心社区合并操作,假设有 s(通常 $s<<|V|$)个中心社区,则时间代价为 $O(s^2)$.步骤 7、步骤 8 实现了未聚类点分配过程,时间代价至多为 $O(|V|)$.综合以上分析,算法 OCDW 的总平均时间复杂度为 $O(ct|E|+|V|)$,其中,c 和 t 是常数.

4 实验与结果分析

在空手道俱乐部(zachary's karate club)[24]、海豚社交网络(dolphins social network)[25]、大学生足球网络(American college football network)[26]、Polbooks(books about US politics dataset,http://www.orgnet.com/)、Adjnoun[26]和 Email[27]这 6 个无权网络数据集以及 Les Misėrables[28]、合作网络(NetScience)[28]和 Hep-th[29]这 3 个加权网络数据集上对算法 OCDW 进行评测,数据基本情况见表 1.

Table 1 Experimental datasets

表 1 实验数据集

数据集	是否加权	顶点数	边数	参考社区数
Karate	否	34	78	**2**
Dolphins	否	62	159	**2**
Football	否	115	613	**12**
Polbooks	否	105	441	**3**
Adjnoun	否	112	425	**2**
Email	否	1 133	5 451	–
Les misérables	是	77	254	–
NetScience	是	1 589	2 742	–
Hep-th	是	8 361	15 751	–

4.1 评价指标

设 OCDW 算法的社区发现结果为 $C=\{C_1,\ldots,C_s\}$,有标签数据集上的原始划分结果为 $O=\{O_1,\ldots,O_t\}$.对集合 C_i 和 $O_j(1\leq i\leq s,1\leq j\leq t)$,令 $T_{i,j}=|C_i\cap O_j|, b_i=\sum_{j=1}^{t}T_{i,j}, d_j=\sum_{i=1}^{s}T_{i,j}$. 本文采用以下指标对算法性能进行度量.

(1) 准确度(Acc)[30]用来度量两个参照集的匹配程度,其定义如公式(10)所示.

$$Acc=\sqrt{\frac{\sum_{i=1}^{S}\max_j\{T_{i,j}\}}{\sum_{i=1}^{S}b_i}\cdot\frac{\sum_{j=1}^{t}\max_i\{T_{i,j}\}}{\sum_{j=1}^{t}|O_j|}} \tag{10}$$

(2) 分离度(Sep)[31]用来度量社区发现结果与原始数据划分的一致性,其定义如公式(11)所示.

$$Sep=\sqrt{\frac{\left(\sum_{i=1}^{S}\sum_{j=1}^{t}Sep_{i,j}\right)^2}{s\times t}} \tag{11}$$

其中, $Sep_{i,j}=\frac{T_{i,j}^2}{b_i\cdot d_j}$ 表示已发现社区 C_i 与原始社区 O_j 的一致度.

(3) 度量(F-measure)[32]通过精度(precision)和召回率(recall)的调和平均值来度量社区发现结果与原始社区的匹配程度,其定义如公式(12)所示(此处θ=0.25).

$$F\text{-}measure=\frac{2\cdot Precision\cdot Recall}{Precision+Recall} \tag{12}$$

其中,

$$Precision=\frac{|\{C_i\in C:\exists O_j\in O\wedge NA(C_i,O_j)\geq\theta\}|}{|C|},$$

$$Recall=\frac{|\{C_j\in C:\exists O_i\in O\wedge NA(C_i,O_j)\geq\theta\}|}{|O|},$$

$$NA(C_i,O_j)=\frac{(T_{i,j})^2}{|C_i|\cdot|O_j|}.$$

(4) 标准互信息(NMI)通过熵来度量两个数据分布的吻合程度,此处用来度量社区发现的结果与原始社区划分的吻合程度,其定义如公式(13)所示.

$$NMI=\frac{2\sum_{i=1}^{s}\sum_{j=1}^{t}T_{i,j}\log\frac{nT_{i,j}}{b_id_j}}{-\sum_{i=1}^{s}b_i\log\frac{b_i}{n}\sum_{j=1}^{t}d_j\log\frac{d_j}{n}} \tag{13}$$

(5) 调整兰德系数(ARI)是由 Hubert and Arabie 提出的,通过统计划分正确的社区中元素的对数来度量算法社区发现结果和原始社区的一致性,定义如公式(14)所示.

$$ARI=\frac{\sum_{i=1}^{s}\sum_{j=1}^{t}\binom{T_{i,j}}{2}-\frac{\sum_{i=1}^{s}\binom{b_i}{2}\sum_{j=1}^{t}\binom{d_j}{2}}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_{i=1}^{s}\binom{b_i}{2}+\sum_{j=1}^{t}\binom{d_j}{2}\right]-\frac{\sum_{i=1}^{s}\binom{b_i}{2}\sum_{j=1}^{t}\binom{d_j}{2}}{\binom{n}{2}}} \tag{14}$$

其中,n 为网络总节点数.$ARI\in[-1,1]$,其值越大,说明社区发现结果与原始社区越吻合.

(6) 模块性(EQ).本文用 Shen 等人提出的重叠社区模块性[15]度量算法在无标签数据集上的社区发现质量,如公式(15)所示.

$$EQ=\frac{1}{2}\sum_{i=1}^{s}\left(\sum_{x\in C_i}\sum_{y(\neq x)\in C_i}\frac{1}{O_xO_y}\left[A_{xy}-\frac{k_xk_y}{2m}\right]\right) \tag{15}$$

其中,o_x 表示在最终社区发现结果中节点 x 属于的社区数,A_{xy} 是原始网络邻接矩阵,k_x 表示节点 x 的度数,m 是原始网络总边数.

4.2 实验结果与分析

在 5 个带社区标签的数据集空手道俱乐部(karate)、海豚社交网络(dolphins)、大学生足球网络(football)、Polbooks 和 Adjnoun 上,将本文算法 OCDW 与经典的社区发现算法 k-dense[14],CPM[13],MCODE[12],HC-PIN[19],MDOS[18]和 BGLL[16],从 F 度量、准确度(ACC)、分离度(Sep)、标准互信息(NMI)、调整兰德系数 ARI 这些方面进行比较,结果如表 2 和图 6 所示.

尽管 k-dense 算法和 CPM 算法在一些指标中表现突出,然而其运行结果中存在大量的未聚类节点.本文算法 OCDW 能够将网络中的绝大部分节点进行社区指派,在 5 个带标签数据集实验中,各项指标均名列前茅,表明其聚类结果与网络原始社区最为接近,具有较高的综合性能.

在 4 个无标签数据集 Email,Les Misérables,NetScience 和 Hep-th 上,将本文 OCDW 算法在模块性及运行时间方面与其他算法进行比较,结果见表 3.由于算法 OCDW 在不影响聚类效果的前提下对种子节点的选择范围进行了控制,因此算法的运行时间也得到了很好的改进.综合考虑模块性和时间性能,本文算法 OCDW 算法表现良好.

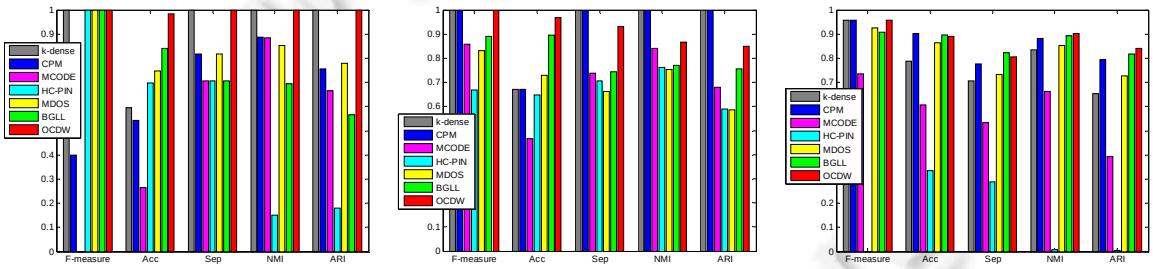
总体而言,本文提出的 OCDW 算法在社区发现过程中综合考虑了网络的拓扑结构和原始权重信息,能够给

出贴近现实网络的社区发现结果,并且具有较好的时间性能和模块性.

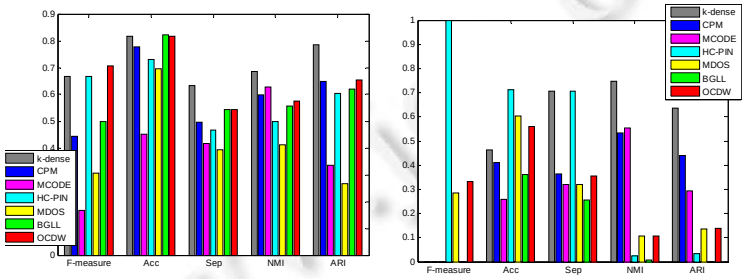
Table 2 Comparison of OCDW with some classical algorithms

表 2 OCDW 与一些经典算法比较结果

Data set	Index	*k*-dense	CPM	MCODE	HC-PIN	MDOS	BGLL	OCDW
Karate	*F*-measure	1.000 0	0.400 0	0.000 0	1.000 0	1.000 0	1.000 0	**1.000 0**
	Acc	0.594 1	0.542 3	0.265 6	0.696 3	0.747 5	0.840 1	**0.985 2**
	Sep	1.000 0	0.816 4	0.707 1	0.707 1	0.816 4	0.707 1	**1.000 0**
	NMI	1.000 0	0.888 0	0.883 8	0.149 2	0.851 8	0.695 6	**1.000 0**
	ARI	1.000 0	0.757 0	0.664 2	0.179 6	0.780 1	0.566 9	**1.000 0**
Dolphins	*F*-measure	1.000 0	1.000 0	0.857 1	0.666 7	0.833 3	0.888 9	**1.000 0**
	Acc	0.672 0	0.672 0	0.465 2	0.647 5	0.728 4	0.894 8	**0.968 4**
	Sep	1.000 0	1.000 0	0.737 7	0.707 1	0.662 2	0.743 5	**0.930 6**
	NMI	1.000 0	1.000 0	0.840 6	0.760 6	0.753 0	0.769 9	**0.868 0**
	ARI	1.000 0	1.000 0	0.680 4	0.588 8	0.584 8	0.756 8	**0.849 1**
Football	*F*-measure	0.956 5	0.956 5	0.736 8	0.000 0	0.924 9	0.909 0	**0.956 5**
	Acc	0.787 3	0.901 3	0.608 0	0.336 2	0.864 4	0.895 2	**0.890 7**
	Sep	0.705 4	0.777 0	0.533 3	0.288 6	0.732 4	0.822 3	**0.805 5**
	NMI	0.834 9	0.882 2	0.662 8	0.007 0	0.852 7	0.892 2	**0.900 7**
	ARI	0.654 8	0.794 6	0.392 2	0.003 0	0.725 8	0.816 5	**0.839 5**
Polbooks	*F*-measure	0.666 6	0.444 4	0.166 6	0.666 6	0.307 6	0.500 0	**0.705 8**
	Acc	0.816 6	0.776 8	0.452 4	0.729 5	0.696 5	0.823 1	**0.816 6**
	Sep	0.633 8	0.495 2	0.415 8	0.466 1	0.394 6	0.542 6	**0.542 8**
	NMI	0.685 9	0.600 0	0.628 2	0.497 8	0.412 8	0.555 8	**0.573 9**
	ARI	0.786 2	0.649 5	0.336 0	0.604 8	0.267 8	0.620 4	**0.653 3**
Adjnoun	*F*-measure	0.000 0	0.000 0	0.000 0	1.000 0	0.285 7	0.000 0	**0.333 3**
	Acc	0.462 9	0.409 5	0.259 6	0.713 3	0.605 2	0.361 3	**0.559 8**
	Sep	0.707 1	0.364 6	0.320 2	0.707 1	0.320 1	0.254 8	**0.356 4**
	NMI	0.747 6	0.532 9	0.555 1	0.025 2	0.106 0	0.007 3	**0.107 4**
	ARI	0.636 7	0.438 9	0.292 4	0.033 9	0.135 5	−0.010 6	**0.139 0**



(a) Karate (b) Dolphins (c) Football



(d) Polbooks (e) Adjnoun

Fig.6 Comparison results on *F*-measure, accuracy, separation, NMI and ARI on 5 real network datasets

图 6 在 5 个真实网络数据集上,*F* 度量、准确度、分离度、标准互信息以及调整兰德系数实验比较结果

Table 3 Modularity (EQ) and running time of different algorithms on 4 real network datasets

表 3 在 4 个真实网络数据及上模块性(EQ)、运行时间指标的比较结果

Data set	Index	*k*-Dense	CPM	MCODE	HC-PIN	MDOS	BGLL	OCDW
Les misérables	EQ	0.154 1	0.299 8	0.338 2	0.248 1	0.187 8	**0.565 0**	**0.463 0**
	Time (s)	0.2	3	0.9	0.8	1.6	**0.1**	**0.2**
Email	EQ	0.256 1	0.265 3	0.100 4	0.001 7	0.206 3	0.289 2	**0.350 1**
	Time (s)	15	59	23	18	67	3	**12**
NetScience	EQ	0.683 3	0.591 5	0.648 1	0.661 5	0.408 9	0.636 6	**0.695 7**
	Time (s)	13	87	36	26	81	10	**11**
Hep-th	EQ	0.387 2	0.538 7	0.303 0	0.623 0	0.361 2	0.602 8	**0.630 5**
	Time (s)	257	392	315	279	3471	**36**	**206**

5 总 结

本文提出了一种基于种子-扩展策略的面向复杂网络中加权稠密子图的社区发现算法 OCDW,综合考虑网络拓扑结构和现实网络权重,对网络中的边权重进行定义;通过搜索网络中的在加权意义下的相对稠密子图得到加权中心社区;评估未聚类节点与加权中心社区的归属度,对未聚类节点进行社区指派.通过与经典的社区发现算法 *k*-dense,CPM,MCODE,HC-PIN,MDOS,BGLL 等算法在 9 个真实网络数据集上进行分析比较,结果表明,本文提出的算法 OCDW 能够将网络中的绝大部分节点进行社区指派,并在 *F* 度量、准确度、分离度、标准互信息、调整兰德系数、模块性及运行时间等方面均表现出较好的性能.

References:

[1] Bai L, Cheng XQ, Liang JY, Guo YK. Fast graph clustering with a new description model for community detection. Information Sciences, 2017,388-389:37−47. [doi: 10.1016/j.ins.2017.01.026]

[2] Atay Y, Koc I, Babaoglu I, Kodaz H. Community detection from biological and social networks: A comparative analysis of metaheuristic algorithms. Applied Soft Computing, 2017,50:194−211. [doi: 10.1016/j.asoc.2016.11.025]

[3] Liu BY, Wang CR, Wang C, Wang JW, Wang XW, Huang M. Microblog community discovery algorithm based on dynamic topic model with multidimensional data fusion. Ruan Jian Xue Bao/Journal of Software, 2017,28(2):246−261 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5116.htm [doi: 10.13328/j.cnki.jos.005116]

[4] Huang FL, Yu G, Zhang JL, Li CX, Yuan CA, Lu JL. Mining topic sentiment in micro-blogging based on micro-blogger social relation. Ruan Jian Xue Bao/Journal of Software, 2017,28(3):694−707 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5157.htm [doi: 10.13328/j.cnki.jos.005157]

[5] Wilson SJ, Wilkins AD, Lin CH, Lua RC, Lichtarge O. Discovery of functional and disease pathway by community detection protein-protein interaction networks. Pacific Symp. on Biocomputing, 2016,22:336−347.

[6] Newman MEJ. Community detection and graph partitioning. Europhysics Letters, 2013,103(2):No.28003. [doi: 10.1209/0295-5075/103/28003]

[7] Newman MEJ. Spectral methods for community detection and graph partitioning. Physical Review E, 2013,88(4):No.042822. [doi: 10. 1103/PhysRevE.88.042822]

[8] Lin CC, Kang JR, Chen JY. An integer programming approach and visual analysis for detecting hierarchical community structures in social networks. Information Sciences, 2015,299:296−311. [doi: 10.1016/j.ins.2014.12.009]

[9] Bai XY, Yang PL, Shi XH. An overlapping community detection algorithm based on density peaks. Neurocomputing, 2017,226: 7−15. [doi: 10.1016/j.neucom.2016.11.019]

[10] Ren J, Wang JX, Li M, Wang LS. Identifying protein complexes based on density and modularity in protein-protein interaction network. BMC Systems Biology, 2013,7(Suppl 4):No.S12. [doi: 10.1186/1752-0509-7-S4-S12]

[11] Li XL, Foo CS, Ng SK. Discovering protein complexes in dense reliable neighborhoods of protein interaction networks. Computational Systems Bioinformatics, 2007,6:157−168. [doi: 10.1142/9781860948732_0019]

[12] Bader GD, Hogue CWV. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics, 2003,4(1):No.2. [doi: 10.1186/1471-2105-4-2]

[13] Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 2005,435(7043):814−818. [doi: 10.1038/nature03607]

[14] Saito K, Yamada T, Kazama K. Extracting communities from complex networks by the *k*-dense method. In: Patist JP, ed. Proc. of the 6th IEEE Int'l Conf. on Data Mining Workshops (ICDMW 2006). Piscataway: IEEE Press, 2006. 300−304. [doi: 10.1109/ICDMW.2006.76]

[15] Shen HW, Cheng XQ, Cai K, Hu MB. Detect overlapping and hierarchical community structure in networks. Physica A, 2009, 388(8):1706−1712. [doi: 10.1016/j.physa.2008.12.021]

[16] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008,2008(10):10008−100019. [doi: 10.1088/1742-5468/2008/10/P10008]

[17] Liu GM, Wong L, Chua HN. Complex discovery from weighted PPI networks. Bioinformatics, 2009,25(15):1891−1897. [doi: 10.1093/bioinformatics/btp311]

[18] Lee AJT, Lin MC, Hsu CM. Mining dense overlapping subgraphs in weighted protein-protein interaction networks. BioSystems, 2011,103(3):392−399. [doi: 10.1016/j.biosystems.2010.11.010]

[19] Wang JX, Li M, Chen JE, Pan Y. A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks. IEEE/ACM Trans. on Computational Biology and Bioinformatics, 2011,8(3):607−620. [doi: 10.1109/TCBB.2010.75]

[20] Lü LY, Zhou T. Link prediction in complex networks: A survey. Physica A, 2011,390(6):1150−1170. [doi: 10.1016/j.physa.2010.11.027]

[21] Palla G, Barabási AL, Vicsek T. Quantifying social group evolution. Nature, 2007,446:664−667. [doi: 10.1038/nature05670]

[22] Jeong H, Mason SP, Barabási AL, Oltvai ZN. Lethality and centrality in protein networks. Nature, 2001,411(6833):41−42. [doi: 10.1038/35075138]

[23] Wang J, Liang JY, Zheng WP. A graph clustering method for detecting protein complexes. Journal of Computer Research and Development, 2015,52(8):1784−1793 (in Chinese with English abstract).

[24] Zachary WW. An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 1977,33(4):452−473. [doi: 10.1086/jar.33.4.3629752]

[25] Lusseau D, Newman MEJ. Identifying the role that animals play in their social networks. Proc. of the Royal Society B: Biological Sciences, 2004,271(Suppl 6):S477−S481. [doi: 10.1098/rsbl.2004.0225]

[26] Newman MEJ. Finding community structure in networks using the eigenvectors of matrices. Physical Review E, 2006,74(3):No.036104. [doi: 10.1103/PhysRevE.74.036104]

[27] Guimerà R, Danon L, Diaz-Guilera A, Giralt F, Arenas A. Self-Similar community structure in a network of human interactions. Physical Review E, 2003,68(6):No.065103. [doi: 10.1103/PhysRevE.68.065103]

[28] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004,69(2):No.026113. [doi: 10.1103/PhysRevE.69.026113]

[29] Newman MEJ. The structure of scientific collaboration networks. Proc. of the National Academy of Sciences of the United States of America, 2001,98(2):404−409. [doi: 10.1073/pnas.98.2.404]

[30] Brohée S, Helden JV. Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics, 2006,7(1):No.488. [doi: 10.1186/1471-2105-7-488]

[31] Qin GM, Gao L. Spectral clustering for detecting protein complexes in protein-protein interaction (PPI) networks. Mathematical and Computer Modelling, 2010,52(11):2066−2074. [doi: 10.1016/j.mcm.2010.06.015]

[32] Li XL, Wu M, Kwoh CK, Ng SK. Computational approaches for detecting protein complexes from protein interaction networks: A survey. BMC Genomics, 2010,11(Suppl 1):No.S3. [doi: 10.1186/1471-2164-11-S1-S3]

附中文参考文献:

[3] 刘冰玉,王翠荣,王聪,王军伟,王兴伟,黄敏.基于动态主题模型融合多维数据的微博社区发现算法.软件学报,2017,28(2):246−261. http://www.jos.org.cn/1000-9825/5116.htm [doi: 10.13328/j.cnki.jos.005116]

[4] 黄发良,于戈,张继连,李超雄,元昌安,卢景丽.基于社交关系的微博主题情感挖掘.软件学报,2017,28(3):694−707. http://www.jos.org.cn/10 00-9825/5157.htm [doi: 10.13328/j.cnki.jos.005157]

[23] 王杰,梁吉业,郑文萍.一种面向蛋白质复合体检测的图聚类方法.计算机研究与发展,2015,52(8):1784−1793.

杨贵(1975－),男,山西大同人,博士生,CCF 专业会员,主要研究领域为生物信息学.

王文剑(1968－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为机器学习,数据挖掘.

郑文萍(1979－),女,博士,副教授,CCF 专业会员,主要研究领域为机器学习,生物信息学.

张浩杰(1991－),男,硕士,主要研究领域为生物信息学.