软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(5):951−961 [doi: 10.3724/SP.J.1001.2011.03939]

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

从链接密度遍历序列中挖掘网络社团的层次结构*

黄健斌[1+], 孙鹤立[2], Dustin BORTNER[3], 刘亚光[1]

[1](西安电子科技大学 软件学院,陕西 西安 710071)

[2](西安交通大学 计算机科学与技术系,陕西 西安 710049)

[3](Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana 61801, USA)

Mining Hierarchical Community Structure Within Networks from Density-Connected Traveling Orders

HUANG Jian-Bin[1+], SUN He-Li[2], Dustin BORTNER[3], LIU Ya-Guang[1]

[1](School of Software, Xidian University, Xi'an 710071, China)

[2](Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China)

[3](Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana 61801, USA)

+ Corresponding author: E-mail: jbhuang@xidian.edu.cn

Huang JB, Sun HL, Bortner D, Liu YG. Mining hierarchical community structure within networks from density-connected traveling orders. *Journal of Software*, 2011,22(5):951−961. http://www.jos.org.cn/1000-9825/3939.htm

Abstract: This paper proposes a density-based network clustering algorithm, TRAVEL. The algorithm produces a traveling order containing clustering with various densities and finds the optimal clusters in it. The traveling order is subsequently transformed into a data structure of contiguous subinterval heap based on which a clustering algorithm, HCLU, is designed to find the hierarchical cluster boundaries of the network without any user interaction. Experimental results on real-world and computer-generated synthetic networks show that the clustering accuracy of the proposed algorithms is higher than the baseline methods. Furthermore, they are able to produce robust hierarchical communities in various networks with low redundancy in the presence of noise.

Key words: density-based network clustering; hierarchical community detection; hub; outlier

摘 要: 提出一种称为 TRAVEL 的网络聚类算法.它能够产生包含所有可能密度聚类的网络链接遍历序列,并从中自动发现网络的全局优化聚类.然后,遍历序列被转换为连续子区间堆结构.在此基础上,提出一种聚类算法HCLU,可以无须用户干预地从连续子区间堆中自动发现网络的层次聚类边界.在真实网络以及计算机生成的仿真网络数据集上的实验结果表明,所提出的算法比目前的基准方法具有更高的聚类精度.此外,算法能够从各种带有噪声的网络中发现无冗余且鲁棒的层次社团结构.

关键词: 基于密度的网络聚类;层次社团发现;中心点;离群点

中图法分类号: TP311 **文献标识码**: A

* 基金项目: 国家自然科学基金(60933009); 陕西省自然科学基础研究计划(SJ08-ZT14)
收稿时间: 2010-06-20; 定稿时间: 2010-08-13
CNKI 网络优先出版: 2010-11-17 16:53, http://www.cnki.net/kcms/detail/11.2560.tp.20101117.1653.000.html

伴随着网络及相关信息技术的迅猛发展,人类社会已经迈入了网络时代.复杂网络系统随处可见,例如网页链接网络、引文网络、道路交通网络、基因调控网络、蛋白质交互网络等.真实世界网络内在的社团结构是复杂网络最普遍和最重要的拓扑结构属性之一[1,2].网络社团通常代表网络中有意义的模块和实体,例如,万维网中的社团通常对应同主题的网页集合,蛋白质交互网络中的社团往往代表细胞中具有特定功能的一组蛋白质等.社团发现对于揭示复杂网络的内在结构、理解网络的功能特性等均有重要作用.

网络社团发现是一项颇具挑战性的课题,因为社团大小和数目通常是未知的,并且网络内部的链接密度是变化和倾斜的.网络社团结构呈现出显著的层次特征,在低密度大社团内部往往还嵌套有高密度小社团.因而,小社团组合在一起形成大社团,而大社团又可合并为规模更大的社团.此外,复杂网络中节点的角色也是多样的,除了那些与社团内部紧密连接的成员节点以外,网络中还存在其他类型的节点,例如中心点和离群点.中心点在真实复杂网络中往往扮演着非常特殊和重要的角色.例如,网页链接网络中的中心网页可用于改善网页排序并提高搜索的性能[3],虚拟市场和疫病传播网络中的中心节点对于散播观点和疾病起到重要作用[4,5].而离群点则与社团节点边缘连接,可以看作是噪声数据.因此,一个好的网络聚类算法不仅要能够有效发现网络中的层次社团结构,而且要自动识别网络中的中心点和离群点.

目前,已经提出的复杂网络社团发现方法的基本思想大多是依据网络节点的某种内聚性度量指标,递归地对网络进行分裂或合并,进而挖掘出其中的社团结构[6].目前已提出的典型方法有:(1) 图分割法.例如 Kernighan- Lin、比例割、规范割、基于最大流的算法等.(2) 谱聚类法.这类方法采用矩阵分析技术将求解图割函数转化为计算与分析图拉普拉斯矩阵的第二最小特征向量.(3) 基于目标函数优化的聚类.这类方法通过最大化设定的聚类质量评价指标,寻找最优网络聚类.2004 年,Newman 提出了网络模块度(modularity)评价函数[7].典型的模块度优化聚类方法有 FN,CNM[8],BGLL[9]等.目前,模块度优化方法已经成为复杂网络社团发现的一种基准方法,得到了广泛的应用.

但是,以上社团发现方法着重关注探测网络中的高密度子图,而忽略了对网络中其他角色节点的识别.Xu 等人提出了一种基于链接密度的网络聚类算法 SCAN[10].对于网络中的任意一对节点,通过定义其基于共享邻居的相似度,使用一个全局的相似度阈值ε,可以发现链接密度高于该阈值的网络社团,并且自动识别网络中的中心点和离群点.但是,SCAN 算法的聚类结果对参数ε非常敏感且参数选择非常困难.此外,该算法无法处理真实复杂网络中社团结构层次嵌套以及密度分布高度倾斜的复杂情况.据此,本文提出一种从网络的链接密度遍历序列中自动挖掘社团层次结构的新算法.其主要创新点如下:

(1) 提出一种自动产生网络节点链接密度遍历序列的 TRAVEL 算法,既可以保存所有可能ε参数对应的网络结构信息,又可以从中自动抽取优化参数ε对应的网络社团;
(2) 提出一种连续子区间堆结构,可从网络链接密度遍历序列中抽取所有连续ε区间及其嵌套层次结构;
(3) 提出一种 HCLU 聚类算法,可从连续子区间堆中自动规约出合理的网络层次社团结构,并且分离出中心点和离群点.

本文第 1 节简单介绍基于密度网络聚类的基本概念.第 2 节阐述链接密度遍历序列的提取算法,并给出全局优化参数ε的确定方法.第 3 节给出社团层次结构规约算法.第 4 节是实验结果及分析.最后,总结了全文.

1 基于密度的网络聚类

本文将网络中的社团看作是由结构相似度大于等于给定阈值ε的一组相互结构可达的节点所构成的聚类.每个聚类中又分为核节点和边界节点,其中,每个核节点最少有μ个与其结构相似度大于等于ε的邻居节点.下面简要介绍基于链接密度的网络聚类相关的基本概念[10−13].

定义 1(结构相似度). 设 $G=\langle V,E,w\rangle$是一个无向网络,w 是边赋权函数.对于任一节点 $u\in V$,u 的结构邻居$\Gamma(u)$是由 u 及与其有公共边的邻接节点构成的集合,即$\Gamma(u)=\{v\in V|\{u,v\}\in E\}\cup\{u\}$.任意两个邻接节点 u 和 v 的结构相似度定义为

$$\sigma(u,v)=\frac{\sum_{x\in\Gamma(u)\cap\Gamma(v)} w(u,x)\cdot w(v,x)}{\sqrt{\sum_{x\in\Gamma(u)} w^2(u,x)}\cdot\sqrt{\sum_{x\in\Gamma(v)} w^2(u,x)}}.$$

对于节点 u,与 u 的结构相似度大于等于ε的邻居节点构成了节点 u 的ε-邻居:

$$\Gamma_\varepsilon(u)=\{v\in\Gamma(u)|\sigma(u,v)\geqslant\varepsilon\}.$$

如果$|\Gamma_\varepsilon(u)|\geqslant\mu$,则 u 是一个核节点,记为$K_{\varepsilon,\mu}(u)$.

以上结构相似度定义扩展自离散余弦相似度,它能够有效地计算无向带权网络中任意两个相邻节点的局部链接密度.这个相似度定义可以替换为其他相似度定义,例如 Jaccard 系数.

定义 2(结构可达). 设$\varepsilon\in\mathbf{R},\mu\in\mathbf{N}$.若$K_{\varepsilon,\mu}(u)\wedge v\in\Gamma_\varepsilon(u)$,则称从节点$u\in V$到节点$v\in V$直接结构可达,记为$u\mapsto_{\varepsilon,\mu}v$.若$\exists\{u_1,\ldots,u_n\}\subseteq V$ 满足 $u=u_1,v=u_n$,且$\forall i\in\{1,2,\ldots,n-1\}$有 $u_i\mapsto_{\varepsilon,\mu}u_{i+1}$,则称从节点 $u\in V$ 到节点 $v\in V$ 结构可达,记为$u\rightarrow_{\varepsilon,\mu}v$.

定义 3(结构连接). 设$\varepsilon\in\mathbf{R},\mu\in\mathbf{N}$.若有 $u\rightarrow_{\varepsilon,\mu}v$ 且 $u\rightarrow_{\varepsilon,\mu}w$,则称 v 和 w 是结构连接的,记为 $v\leftrightarrow_{\varepsilon,\mu}w$.

图 1 给出了网络中结构可达和结构连接关系的示意图,图中任意两个邻居节点之间连接边的长度越长,则两者的相似度越小.假设参数ε=0.75 且μ=4,此时,由于节点 p 与其邻居节点结构相似度大于 0.75 的有 4 个,因此 p 是核节点.对于 p 的邻居节点 q,由于 p 与 q 之间的相似度大于 0.75,因此 p 到 q 是直接结构可达的.从图 1 中容易看出,q 也是核节点;且由于 p 与其邻居节点 r 的相似度大于 0.75,因此 q 到 r 也是直接结构可达的.根据结构可达的定义,p 到 r 是结构可达的.结构可达关系对于核节点之间是自反、对称和传递的.但是,当涉及非核节点时,其对称性通常不再成立.节点 p 到 t 是结构可达的,且节点 p 到 s 也是结构可达的,因此节点 t 与 s 之间是结构连接的.

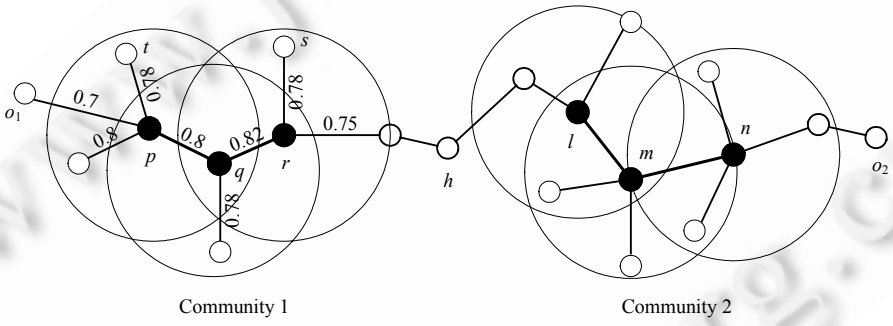


Fig.1 Relationship of structural reachability and connectivity between the nodes in an example network which has two structure-connected communities, as well as one hub and two outliers

图 1 一个具有两个社团以及一个中心点和两个离群点的示例网络中,节点间的结构可达和结构连接关系

定义 4(结构连接社团). 设集合 $C[u]\subseteq V$,$C[u]$是由$K_{\varepsilon,\mu}(u)\in V$表示的社团,当且仅当满足以下条件:(1) $u\in C[u]$;(2) $\forall v\in V,u\rightarrow_{\varepsilon,\mu}v\Rightarrow v\in C[u]$;(3) $|C[u]|\geqslant\mu$.

根据以上关于结构连接社团的定义,网络中的一个社团 C 恰好包含其中与任意核节点结构可达的那些节点.同一个社团内的节点间是结构连接的.

定义 5(中心点和离群点). 给定参数ε和μ,设$CR_{\varepsilon,\mu}$是一个网络中所有社团组成的集合.节点$h\in V$是一个中心点当且仅当满足以下条件:(1) $\forall C[u]\in CR_{\varepsilon,\mu},h\notin C[u]$,即 h 不属于任何聚类;(2) $\exists C,D\in CR_{\varepsilon,\mu},C\neq D,u\in C\wedge v\in D$,满足$h\in\Gamma(u)\wedge h\in\Gamma(v)$,即 h 邻接多个聚类中的节点.不属于任何聚类的非中心点称为离群点.

如图 1 所示,不属于任何社团的节点 h 连接着相邻两个社团中的节点,因此它是一个中心点;而节点 o_1 和 o_2 虽然也不属于任何社团,但它们均最多与 1 个社团连接,因此这两个节点是离群点.如果允许中心点属于多个相

邻社团,本文提出的方法可以很方便地用于发现网络中的重叠社团.

2 从网络的链接密度遍历序列中提取全局优化ε聚类

本节介绍 TRAVEL 算法,它用于高效地计算网络的局部结构相似度阈值,以及发现全局优化ε聚类.

2.1 产生网络链接密度遍历序列

定义 6(核相似度). 给定节点 $u\in V$,u 的核相似度为

$$CS(u)\equiv\begin{cases}\max\{\varepsilon\in R^+ : |\{v\in\Gamma(u):\sigma(u,v)\geqslant\varepsilon\}|\geqslant\mu\}, & |\Gamma(u)|\geqslant\mu\\ 0, & \text{otherwise}\end{cases},$$

如果$|\Gamma(u)|\geqslant\mu$,则节点 u 的核相似度是使节点 u 满足 $|\Gamma_{\hat{\varepsilon}}(u)|\geqslant\mu$ 成为核节点的最大相似度阈值 $\hat{\varepsilon}$;否则,u 的核相似度等于 0.

定义 7(可达相似度). 给定节点 $u,v\in V$,u 节点到 v 的可达相似度为

$$RS(u,v)\equiv\min\{CS(u),\sigma(u,v)\}.$$

节点 u 到 v 的可达相似度是满足 u 是核节点且 u 到 v 直接结构可达的最大相似度阈值 $\tilde{\varepsilon}$.

为了捕获所有可能的ε聚类,TRAVEL 算法预先设定一个相对较小的初始相似度阈值ε_0.算法使用一个优先级队列并初始化优先级队列 Q 为空,Q 中的节点始终按其优先级降序排序,.若 Q 队空,则从网络中任意选择一个未访问节点 v,将 v 以其核相似度作为优先级插入队 Q 中;否则,将 Q 中的队头节点 h 出队.此时,若 h 是核节点,则将 $\Gamma_{\varepsilon_0}(h)$ 中未访问节点 w 以及 h 到 w 的可达相似度插入队 Q 中.重复上述过程,直至网络中所有节点均被访问.TRAVEL 算法的详细过程如算法 1 所示.图 2 显示了一个简单网络的链接密度排序点图,其中,3 个明显分隔的“山峰”表明网络中存在 3 个显著社团,而中心点和离群点则位于 3 个隆起之间的“低洼”地带.

Fig.2 Graph of density-connected traveling order in a network with three significant communities

图 2 具有 3 个明显社团的网络链接密度排序图

算法 1. *TRAVEL*(G,ε_0,μ).

输入:网络 $G=\langle V,E\rangle$,最小相似度阈值ε_0,最小邻居节点数μ.

输出:网络链接密度序列 S.

```
(1)  PriorityQueue Q;                    //初始化动态优先级队列 Q,其中元素始终按可达相似度由大到小排列
(2)  long i=0; Node S[N];
(3)  for all v∈V
(4)      if v.visited then continue;
(5)          Q.Insert(v,CS(v));                    //将未访问节点入队
(6)      while Q.size()≠0 do                       //队不为空
(7)          S[++i]=Q.HeadOut();                   //队头节点出队
(8)          S[i].visited=ture;
(9)          if currNode.isCoreNode(G,ε0,μ) then   //将出队核节点的未访问可达邻居节点入队
(10)             Q.NeighborEnquere(G,currNode,ε0,μ);
(11)     end while
(12) end for
```

(13) **return** S;

可以证明,以上 TRAVEL 算法输出的长度为 n 的序列 S 中包含了所有可能ε对应的网络聚类信息.

2.2 提取全局优化ε网络聚类

文献[11]中给出了一种从 k-近邻曲线中发现"膝盖"值的参数ε手工设定方法.然而事实上,很多情况下难以自动从网络的 k-近邻曲线发现膝盖值,本文提出一种自动发现全局优化ε的算法.由于网络的 k-近邻曲线是单调递增的,并且膝盖假设的前提为曲线是凸的.假定选择其中任一排序输出的相似度值ε_i,从最小值ε_1 和最大值ε_n分别画一条连接ε_i 的直线,那么曲线的"膝盖"应该是其中夹角最小者.使用以下公式来探测两条直线的实际夹角θ.

$$\tan\theta=\frac{|m_1-m_2|}{1+m_1\cdot m_2},$$

其中,$m_1=(\varepsilon_i-\varepsilon_1)/i$ 且 $m_2=(\varepsilon_n-\varepsilon_i)/(n-i)$是线段的倾斜度.使用点对点方法有

$$\theta_i=\arctan\frac{(n-i)(\varepsilon_i-\varepsilon_1)(\varepsilon_n-\varepsilon_i)}{i(n-i)+(\varepsilon_i-\varepsilon_1)(\varepsilon_n-\varepsilon_i)}.$$

使用二叉搜索可以在 $O(\log n)$时间内完成对最大θ_i 的查找.为了提高算法的鲁棒性,可选择多个不同的启始搜索点,然后选择出现次数最频繁的值作为优化ε.本文将优化ε搜索过程与 TRAVER 算法中的网络遍历完全结合在一起,全局最优ε可以在 $O(n\log n)$时间内自动发现.

3 密度嵌套的社团层次结构提取

以下提出一种基于链接密度遍历序列的无参数网络聚类方法,以获得网络社团的层次嵌套结构.

3.1 构建连续子区间堆

定义 8(连续ε区间). 给定网络 $G=\langle V,E\rangle$,参数$\mu\in N,\varepsilon_0>0$ 以及遍历函数 S.从 a 到 b 的连续闭区间$[a,b]\subseteq\{1,2,\ldots,|V|\}$,若满足:(1) $\forall i\in[a,b],Sim_G(i)\geqslant\varepsilon$;(2) $\nexists[c,d]\supset[a,b]$ 满足$\forall i\in[c,d],Sim(i)\geqslant\varepsilon$,则称$[a,b]$是一个连续$\varepsilon$区间,记为 $Contig_\varepsilon(a,b)$.

连续ε区间表示可达相似度大于等于ε的一个极大连续位置索引子集.对于连续 $\hat{\varepsilon}$ 区间 $Contig_{\hat{\varepsilon}}(a,b)$,在当前$\varepsilon$参数值小于等于 $\hat{\varepsilon}$ 的情况下,区间$[a,b]$内所有位置索引对应的网络节点在同一个聚类中.

定义 9(子区间). 给定一个连续ε区间 $Contig_\varepsilon(a,b)$,区间$[c,d]$是一个连续区间$[a,b]$的子区间,当且仅当$[c,d]\subset[a,b]$且$\exists\varepsilon'>\varepsilon$满足$[c,d]$是一个连续$\varepsilon'$区间,记为$[c,d]\prec[a,b]$.

连续子区间可以存放于一个树结构中,而子区间偏序关系使得该树结构是一个堆,称为连续子区间堆.通过以下过程构建连续子空间堆:首先,序列 S 中第 1 个被访问的节点成为某连续区间的下界,此时上界尚不可知,因此先将此节点入栈;从 S 中读入后续节点,它也被视为一上界不知的连续区间的下界.此时,如果当前所访问节点的可达相似度值ε大于等于栈顶元素,则将其入栈;否则,可达相似度值大于等于ε的栈顶元素均出栈.当前连续子区间的上界即为目前所读元素的位置索引.这个连续子区间将被视为后一个出栈的连续子区间的孩子节点.当最后一个被访问节点读入,栈中所剩的所有连续区间将作为堆顶的孩子,最终形成整个堆结构.

3.2 探测聚类边界

上一节显示了如何从TRAVEL产生的遍历序列中提取所有可能的ε聚类对应的嵌套层次树.但是,其中包含大量差别很小的聚类,下面介绍优化层次聚类的确定方法.

定义 10(可达相似度的导数). $Sim(i)$的导数定义为

$$Sim'(i)\equiv Sim(i)-Sim(i-1).$$

定义 11(聚类边界条件). 连续区间$[a,b]$对应一个聚类仅当 $Contig_\varepsilon(a,b)$,且满足 $Sim'(a)$是一局部最大值以及 $Sim'(b)$是一局部最小值.

下面使用权重方法来确定一个节点是否为聚类边界.设 n_k 是被评价的节点,其父节点 n_{k-1} 与 n_k 的相关度可

以用两者宽度的比例 w_k/w_{k-1} 来表示.因此,如果这个比例比较高(例如 0.95),那么 n_{k-1} 对于确定 n_k 是否为聚类边界将有很高的相关度.但是,如果这个比例较低(例如 0.40),那么 n_{k-1} 对于确定 n_k 是否为聚类边界的相关度就较低.此外,设 n_k 的祖先 $n_{k-d}(d<k)$的相关度定义为

$$R(w_{k-d},w_k)\equiv\frac{w_k}{w_{k-1}}\frac{w_{k-1}}{w_{k-2}}\cdots\frac{w_{k-d+1}}{w_{k-d}}=\frac{w_k}{w_{k-d}}.$$

n_{k-d} 对于 n_k 的相关度是从 n_k 到 n_{k-d} 的路径上成对相邻节点间相关度的乘积.因此,如果中间节点都具有高的成对相关度,那么它们会形成一个链,n_{k-d} 对于 n_k 也有高的相关度.然而,如果任一中间节点对相关度较低,将会打破这个链,所有后续节点的相关度都将减弱.

本文提出 HCLU 聚类算法从两个方向计算来确定聚类边界,算法 2 给出详细实现过程.在 ContigHeap 的根节点上调用 $HCLU(\mu,0,0,0,0,\&ts_0=0,\&ts_1=0,\&ts_2=0)$,算法从连续子区间堆 ContigHeap 中删除所有冗余的连续子区间.图 3 和图 4 分别显示了从 DBLP 合著网络的部分链接遍历序列中发现的连续子区间堆以及 HCLU 产生的层次聚类边界.

Fig.3 Contiguous subintervals found by BuildContigHeap

图 3 BuildContigHeap 算法发现的连续子区间

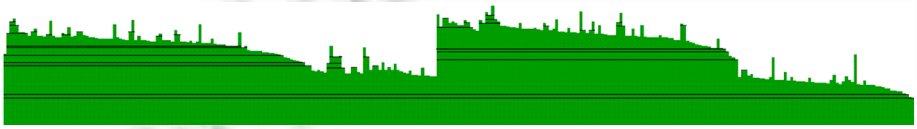


Fig.4 Cluster boundaries pruned by algorithm HCLU

图 4 HCLU 算法修剪后的聚类边界

算法 2. $HCLU(\mu,s_0,s_1,s_2,parEps,\&cs_0,\&cs_1,\&cs_2)$.

输入:聚类最小节点数μ,根到叶子带权矩 s_0~s_1,父节点的ε值 $parEps$.

输出:叶子到根带权矩 cs_0~cs_1.

(1) $\Delta\varepsilon=\varepsilon-parEps$; //计算密度间隔

(2) $w=b-a$;

(3) $active=(w+1\geqslant\mu)$;

(4) **if** *active* and $s_0\neq0$ **then** //迭代修剪

(5) $m=s_1/s_0$; //探测聚类边界

(6) $s=(1/s_0)\sqrt{s_0s_2-s_1^2}$;

(7) $active=(\Delta\varepsilon-m\geqslant s)$;

(8) $s_0=1/w$; $s_1=\Delta\varepsilon\cdot s_0$; $s_2=\Delta\varepsilon\cdot s_1$;

(9) $ts_0=ts_1=ts_2=0$;

(10) **for** all children **do** //对孩子节点递归处理

(11) $child.HCLU(minPts,s_0,s_1,s_2,\&ts_0,\&ts_1,\&ts_2)$;

(12) **if** *active* and $ts_0\neq0$ **then**

(13) $m=ts_1/ts_0$;

(14) $s=(1/ts_0)\sqrt{ts_0ts_2-ts_1^2}$;

(15) $active=(\Delta\varepsilon-m\geqslant s)$;

(16) **endif**

(17) $cs_0=cs_0+ts_0+w$; $cs_1=cs_1+ts_1+\Delta\varepsilon\cdot w$; $cs_2=cs_1+ts_1+\Delta\varepsilon\cdot w$;

(18) **endfor**

(19) **endif**

4 实验结果与分析

本节对所提出的 TRAVEL+HCLU 算法与基于密度的网络聚类算法 SCAN[10]以及基于模块度优化的 CNM[8]和 BGLL[9]算法进行了综合评测,其中,BGLL 是当前准确度最高的网络社团发现基准算法之一.TRAVEL+HCLU 和 SCAN 算法采用 Visual Studio.Net 2005 C#语言实现,CNM 和 BGLL 算法则分别采用 Clauset 和 Blondel 等提供的 C++源代码.所有实验均在配置 PIV 2.4GHz CPU 和 2G 内存的 PC 机上完成.

4.1 实验数据集

为了评价 TRAVEL 算法发现全局优化ε聚类和层次聚类的能力,使用了两个真实网络数据集:US Political Books 网络(简称 Polbooks)和 DBLP 论文合著网络(简称 DBLP-Coauthorship).本文还使用 Lancichinetti 等人[13,14]开发的工具软件产生了一些基准网络数据集,用于详细分析算法针对不同结构网络的聚类性能.本文使用节点数为 5 000 和 50 000 的两种规模的基准网络,对于每一种网络,分别产生混合系数从 0.1 到 0.8 且间隔为 0.05 的 15 个不同网络.表 1 给出了这些数据集的详细参数设置情况,其中,网络命名为 Benchmark-n,n 是节点个数,m 是边数的平均值,参数 k 表示节点的平均度数,max$k$ 表示节点的最大度数,mu 表示混合参数,min$c$ 表示最小社团节点数,max$c$ 表示最大社团节点数.

Table 1 Parameters of the computer-generated networks for accuracy evaluation

表 1 用于聚类精度评价的计算机生成网络参数

Dataset	n	m	k	max$k$	min$c$	max$c$
Benchmark-5000	5 000	49 227	20	50	10	50
Benchmark-50000	50 000	998 069	40	150	20	100

4.2 评价指标

本文使用规范化互信息(normalized mutual information,简称 NMI)来评价不同方法的聚类质量,该指标基于含混矩阵 N.NMI 定义为

$$NMI=\frac{-2\sum_{i,j}N_{ij}\log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right)}{\sum_{i}N_{i.}\log\left(\frac{N_{i.}}{N}\right)+\sum_{j}N_{.j}\log\left(\frac{N_{.j}}{N}\right)},$$

其中,N_{ij} 是聚类 X_i 和 Y_j 中公共的节点数,$N_{i.}$ 是 N 中第 i 行求和,N_{j} 是 N 中第 j 列求和.NMI 的取值在 0,1 之间,取 0 时表示两种结果完全不一致,取 1 时表示完全一致.

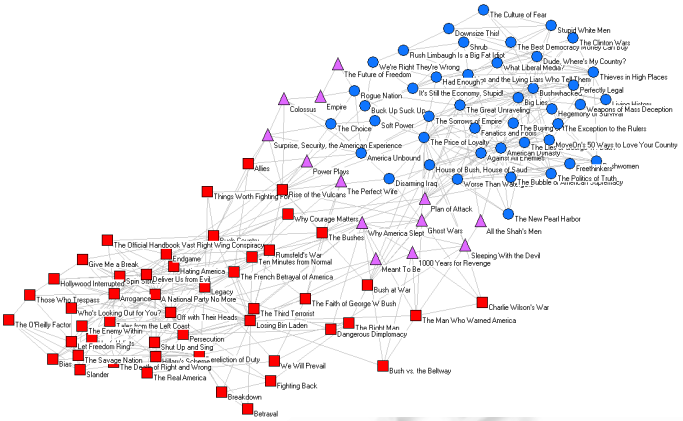
4.3 真实网络上的聚类性能评价

4.3.1 Polbooks 网络

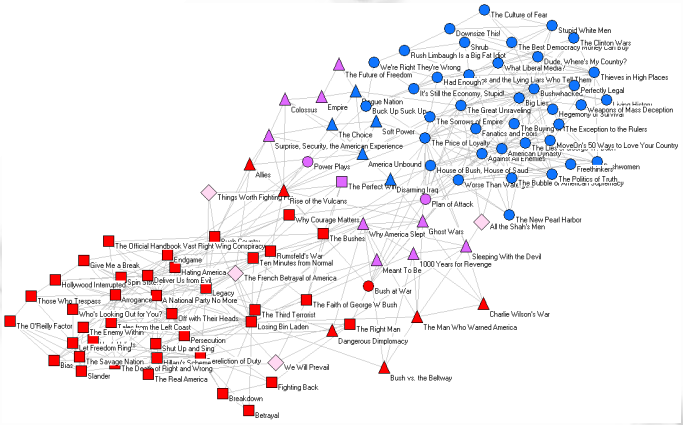
Polbooks 网络中包含 105 个节点和 441 条边.网络中的每个节点表示一本美国出版的关于政治主题的图书,如果两本书被同一消费者购买,则在这两个节点之间连有一条边.Newman 通过研究认为,该网络中存在 3 个不同社团,并对其中的节点给出了 Liberal,Neutral,Conservative 这 3 种不同类型的标注.

图 5(a)和图 5(b)分别显示了 Polbooks 网络的标准社团结构以及 TRAVEL 算法在该网络上的聚类结果.TRAVEL 算法准确识别出 3 个社团.正方形、圆形和三角形这 3 种不同的形状分别代表 Conservative,Liberal 和 Conservative 类图书所代表的社团.此外,算法还识别出 3 个用菱形节点表示的中心点.在该网络中,SCAN 算法在

参数设置为ε=0.43 且μ=2 时发现 4 个社团,同时还识别出 3 个中心点和两个离群点.基于模块度优化的算法 CNM 和 BGLL 在该网络中分别发现了 4 个和 5 个社团.



(a) Ground truth of community structure in Polbooks network
(a) Polbooks 网络的真实社团结构



(b) Communities discovered by TRAVEL algorithm in Polbooks network
(b) TRAVEL 算法在 Polbooks 网中发现的社团

Fig.5 Community structure in Polbooks network and the clustering result of TRAVEL algorithm on it
图 5 Polbooks 网络中的标准社团结构以及 TRAVEL 算法在其上的聚类结果

TRAVEL 算法在 Polbooks 网络上取得了最好的聚类结果,这是由于 TRAVEL 算法采用了基于链接密度的聚类原理且能够自动发现全局优化的参数ε.此外,算法对μ参数并不敏感,对于网络聚类通常设$\mu\leqslant$5 即可.

4.3.2 DBLP-Coauthorship 网络

下面使用 DBLP-Coauthorship 网络评测 TRAVEL+HCLU 算法的层次聚类质量.DBLP-Coauthorship 网络是从 2007 年 DBLP 计算机科学在线论文网络中抽取的包含数据库、信息检索、数据挖掘和机器学习这 4 个研究方向学者合著信息的一个真实网络.该网络包含 28 702 个节点和 66 832 条边,其中,每个节点对应一个不同的论文作者,两个节点之间的边表示这两个节点之间有合作撰写的论文,边上的权值表示合著论文的篇数.

为了显示 HCLU 所发现层次聚类的合理性,从聚类结果中选出一些具有代表性的聚类,如图 6 所示.受论文篇幅所限,这里无法将整个聚类结果完全展现,图 6 中共给出两层聚类、11 个社团,每个社团表示了一组具有共同研究兴趣并合著论文的学者.从图中可以观察到,本文提出的层次聚类算法能够从大规模学术协作网络中发

现有意义的社团层次结构.例如,图6中上层的大社团有包含Klein和Jordan等著名学者的机器学习领域的社团,包含Callan和Carbonell等人的信息检索领域的社团,包含Ullman和Gray等人的数据库领域的社团,以及包含Aggarwal和Karypis等人的数据挖掘领域的社团;而下层的小社团表示了联系更为紧密的学术团体,其中包括Kumar,Kleignberg等一些知名学者及其同事和学生组成研究组.在发现社团的同时,本文提出的算法还能有效地发现网络中的中心点和离群点,在学者名字前分别用菱形和三角形标注.例如,在这个网络中Yu,Han,Garcia-Molina等知名科学家与很多研究组合作发表了大量的学术论文,这些学者被识别为中心点,或者可以认为他们属于多个社团从而形成重叠社团;相反,那些仅与某个特定研究组合作发表了少量论文的学者则被识别为离群点.

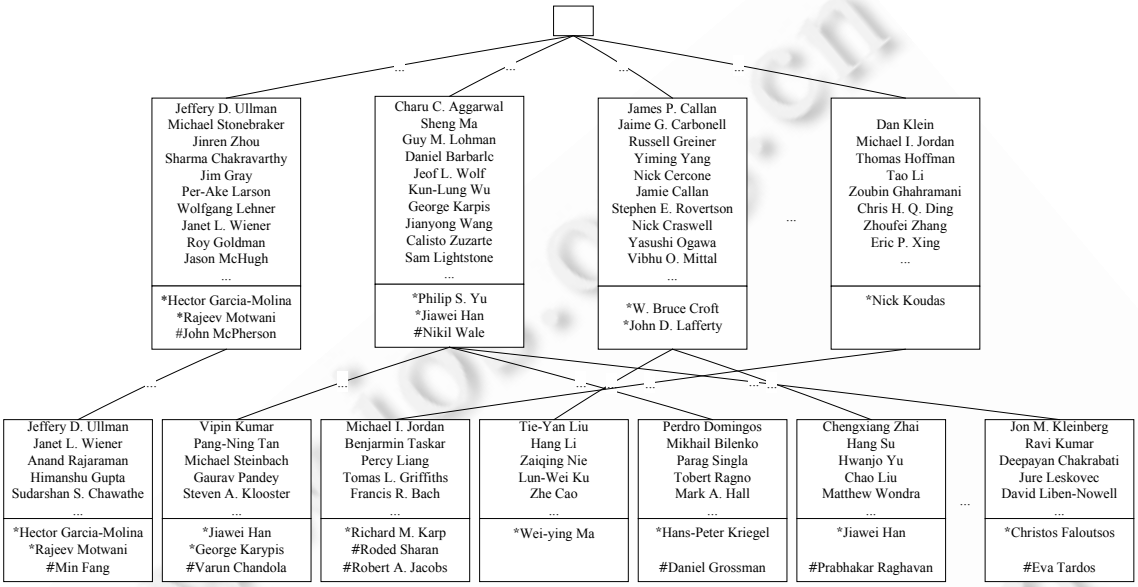


Fig.6 Eleven real communities in two layers found by the HCLU algorithm on the DBLP-Coauthorship network

图6 HCLU算法在DBLP-Coauthorship网络中发现的两层11个真实社团

4.4 人工合成网络上的聚类性能评价

为了全方位地比较算法的性能,采用Lancichinetti等人[13]设计的基准网络来深入评测以上3种算法的聚类准确性.这些生成网络满足与真实网络类似的无标度特性以及内在社团结构特征.

图7给出TRAVEL,BGLL和CNM等3种算法在生成基准网络上聚类结果的NMI值.从中可以看出,当混合系数*mu*<0.5时,TRAVEL算法聚类结果的NMI均等于或接近1.这表明,当网络模块化较清晰时,算法能够完全正确地识别网络的内在社团结构.TRAVEL算法在节点数为50 000网络上的性能要明显高于5 000网络.这是因为50 000网络中社团规模与节点数的比值较小,因而社团之间的边界更清晰且重叠点较少.当混合系数*mu*>0.5时,BGLL算法在5 000网络上的性能要略好于TRAVEL,表明BGLL算法在处理模糊社团识别方面有更强的能力.但是,BGLL算法在大规模网络上的性能明显不及TRAVEL.经测试表明,BGLL算法在这些网络上发现的社团数量通常比标准值小很多,并且趋向于将小的社团合并为规模更大的社团.因此,分辨率局限问题[15]是导致BGLL算法在大规模小团网络上性能下降的主要原因.而CNM算法的聚类准确率较低,并随着*mu*的增大性能迅速下降,这表明使用贪婪模块度最大的层次聚类算法的性能较差.当*mu*>0.5时,以上3种算法的NMI值均明显呈下降趋势.但是,即使在*mu*=0.8时,TRAVEL算法在50 000网络上聚类结果的NMI值依然保持在0.7以上,这表明,本文提出的算法具有很好的鲁棒性.

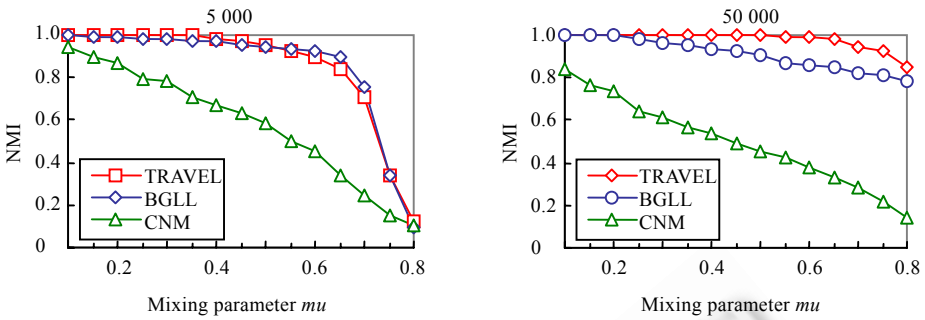


Fig.7 NMI values for the clustering results of three algorithms on the computer-generated benchmark networks

图 7 3 种算法在计算机生成的基准网络上聚类结果的 NMI 值

4.5 算法时间复杂度分析

本文在 Barabasi-Albert 无标度模型的网络生成模型上测试了 TRAVEL,HCLU 和 SCAN 算法的性能,因为很多真实世界的网络均有与 BA 模型类似的长尾现象.所生成的网络的节点数和边数逐步增长,对于每个网络,其边数的增长率与节点数呈线性关系,边数增长率约节点数增长率的 10 倍.在每一种规模的网络上均完成 10 次运行取结果的平均值,对于 SCAN 算法则采用不同的ε参数值重复运行.图 8 显示了在 BA 模型上的运行时间(μ=5).当参数ε增加时,SCAN 的运行时间轻微增加,而 TRAVEL 算法的运行时间显著下降.在这种情况下,SCAN 和 TRAVEL 运行时间相对于边数呈高于线性增长.TRAVEL 在针对于不同参数ε所有的测试中需要的时间略高于 SCAN.对于 30 万个节点的网络,TRAVEL 比 SCAN 慢 5.7%.而 BuildContigHeap 和 HCLU 的最长运行时间仅为 15.6s.因而,整个算法的总体时间复杂度与 SCAN 算法相同,为 $O(m)$.

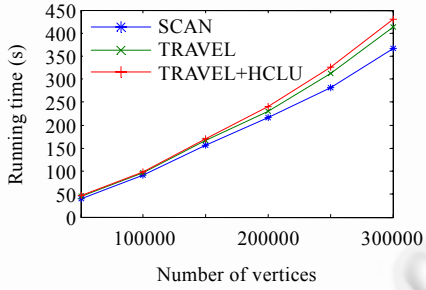


Fig.8 Comparison of the running time for algorithms TRAVEL, TRAVEL+HCLU and SCAN

图 8 TRAVEL,TRAVEL+HCLU 和 SCAN 算法的运行时间比较

5 结 论

本文提出 TRAVEL 算法,从任意一点出发,通过遍历网络密度链接序列,并采用夹角法从网络的 k-近邻排序曲线中自动探测网络优化的全局ε参数值.然后,将链接密度遍历序列转换为一个连续子区间堆 ContigHeap,其中,每个节点表示一个相对于特定参数ε网络聚类.在此基础上提出算法 HCLU,通过遍历堆修剪无效的聚类边界,可以产生网络的社团层次结构.实验结果显示,本文提出的方法有较高的聚类性能,且能够自动产生合理的网络层次与可变聚类.算法的时间复杂度为 $O(m)$,与 SCAN 算法相当.在后面的工作中,将采用本文提出的方法分析不同领域的复杂网络,并探索大规模网络社团的在线分析方法.

References:

[1] Guimera R, Amaral LN. Functional cartography of complex metabolic networks. Nature, 2005,433(7028):895−900. [doi: 10.1038/nature03288]

[2] Girvan M, Newman MEJ. Community structure in social and biological networks. Proc. of the National Academy of Sciences, 2002,9(12):7821−7826.

[3] Kleinberg JM. Authoritative sources in a hyperlinked environment. In: Howard K, ed. Proc. of the 9th ACM-SIAM Symp. on Discrete Algorithms. Philadelphia: SIAM, 1998. 668−677.

[4] Domingos P, Richardson M. Mining the network value of customers. In: Lee D, ed. Proc. of the 7th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM, 2001. 57−66. [doi: 10.1145/502512.502525]

[5] Wang Y, Chakrabart D, Wang C, Faloutsos C. Epidemic spreading in real networks. ACM Trans. on Information and System Security, 2008,10(4):1−26. [doi: 10.1109/RELDIS.2003.1238052]

[6] Yang B, Liu DY, Liu JM, Jin D, Ma HB. Complex network clustering algorithms. Journal of Software, 2009,20(1):54−66 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/20/54.htm [doi: 10.3724/SP.J.1001.2009.00054]

[7] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004,69(2):026113. [doi: 10.1103/PhysRevE.69.026113]

[8] Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Physical Review E, 2004,70(6):66111. [doi: 10.1103/PhysRevE.70.066111]

[9] 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):P10008. [doi: 10.1088/1742-5468/2008/10/P10008]

[10] Xu X, Yuruk N, Feng Z, Schweiger TAJ. SCAN: A structural clustering algorithm for networks. In: Berkhin P, ed. Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM, 2007. 824−833.

[11] Ester M, Kriegel HP, Jorg S, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis E, ed. Proc. of the 2nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM, 1996. 226−231.

[12] Ankerst M, Breunig MM, Kriegel HP, Sander J. Optics: Ordering points to identify the clustering structure. In: Davidson SB, Faloutsos C, eds. Proc. of the 1999 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM, 1999. 49−60. [doi: 10.1145/304182.304187]

[13] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. Physical Review E, 2008, 78(4):046110. [doi: 10.1103/PhysRevE.78.046110]

[14] Lancichinetti A, Fortunato S. Community detection algorithms: A comparative analysis. Physical Review E, 2009,80(5):056117. [doi: 10.1103/PhysRevE.80.056117]

[15] Fortunato S, Barthelemy M. Resolution limit in community detection. Proc. of the National Academy of Sciences, 2007,104(1): 36−41. [doi: 10.1073/pnas.0605965104]

附中文参考文献:

[6] 杨博,刘大有,Liu JM,金第,马海宾.复杂网络聚类方法.软件学报,2009,20(1):54−66. http://www.jos.org.cn/1000-9825/20/54.htm [doi: 10.3724/SP.J.1001.2009.00054]

黄健斌(1975－),男,湖北随州人,博士,副教授,CCF 会员,主要研究领域为数据挖掘,信息网络分析.

Dustin BORTNER(1982－),男,硕士,主要研究领域为数据挖掘.

孙鹤立(1983－),女,博士,讲师,主要研究领域为数据挖掘,信息检索.

刘亚光(1986－),男,硕士生,主要研究领域为数据挖掘.