文章编号:1674-7070(2010)01-0019-07

信息网络社区发现研究

黄发良[1,2]　肖南峰[1]

摘要

随着人们生产生活信息化的深入,信息网络的社区发现研究引起越来越多研究者的关注.在对信息网络社区发现研究的基本概念与原理进行简单介绍的基础上,着重对各种发现方法进行分类分析与比较.最后,对信息网络社区发现技术进行了总结与展望.

关键词

信息网络;社区发现;复杂网络;聚类

中图分类号 TP391.41

文献标志码 A

0 引言
Introduction

互联网的迅猛发展极大地推动了社会信息的网络化进程,以即时通讯系统、P2P信息共享网络、博客网络、邮件网络、短信网络与在线聊天室网络等为代表的信息网络已经深入到人们的工作、学习与生活等活动中,是构成信息社区的基础环境.这些各式各样的信息网络承载着人们在生产生活中形成的复杂关系,从这些纷繁芜杂的关系结构发现隐藏的潜在有价值的关系模式是一个非常困难而又很有意义的工作.

信息网络社区发现任务的艰巨性与挑战性主要源自信息网络的复杂性:网络结构的异构性、网络规模的巨大性与网络属性的动态性.网络结构的异构性主要表现在网络节点类型的多样性,相同类型的节点构成同型网络,而不同类型的节点可以构成异质网络,类似的有网络节点间的关系可以是相同类型,也可以是不同类型,还有网络节点关系的方向性与网络节点关系重要性的度量方法的差异性等多种因素影响着网络结构;网络规模的巨大性是互联网信息网络的一个重要特征,传统的社会网络由于数据收集的困难性使得其网络规模比较小,进而使得其表现出的各种特征从统计意义上讲价值不是很大,互联网与万维网的规模巨大性一方面提供了对其拓扑属性进行可靠分析的基础条件,另一方面也向分析算法的计算效率提出更高的要求;本文将网络属性定义为节点属性与节点间关系属性两个方面,信息网络始终处于一个动态的过程,新节点的加入,原有节点的退出,节点关系的建立和消失,网络属性的动态演化对算法的可扩展性与鲁棒性提出巨大的挑战.

信息网络中隐藏的知识吸引着大量来自社会学、物理学与计算机科学等各个不同学科领域中的科研工作者.最早是社会学家注意到社会信息网络的拓扑结构特征,将实际社会关系网络作为研究对象,试图从中发现隐含的社会关系和从信息网络的角度去解释社会现象;物理学研究者主要从复杂网络理论的层面对包括有形实体网络(诸如电力网、因特网、高速公路或地铁系统及神经网络)和抽象空间网络(例如朋友关系网和个体合作网)在内的复杂系统的拓扑结构属性与演化动力特征进行研究,提出了众多用以解释各种物理现象

收稿日期 2009-11-09

资助项目 国家自然科学基金与中国民用航空总局联合资助(60776816);广东省自然科学重点基金(8251064101000005);福建省教育厅科研基金(JA08049).

作者简介

黄发良,男,讲师,博士生,主要研究方向为智能计算及计算机应用.huangfliang@163.com

肖南峰(通信作者),男,博士,教授,主要研究智能计算及计算机应用.xiaonf@scut.edu.cn

1 华南理工大学 计算机科学与工程学院,广州,510006

2 福建师范大学 软件学院,福州,350007

的网络模型,例如小世界模型、无标度网络以及随机网络等;随着以海量数据分析与挖掘为宗旨的数据挖掘技术研究的深入,计算机科学家也逐步投入到这个充满魔力的研究领域中来,并取得了很多有趣的结论,他们主要是根据信息网络的规模巨大性等特点,设计出高效率、高效用并具有一定智能的鲁棒算法.

1 社区的定义
Definition of community

尽管社区发现的研究在复杂网络中有很长时间了,但到目前还没有一个公认的严格定义,在信息网络领域中有这样一个共识:社区内部节点连接紧密而社区间连接松散. 当前的社区定义主要是从网络自我参照与网络属性比较两个角度给出的. 首先是自我参照角度,社区被定义为完全子图,这个定义所有隐含的社区内所有节点都两两相连接的限制条件过于严格,使得其定义的实际应用意义很小,于是出现了多种不同的限制条件弱化的社区定义,如 N-Cliques、N-Clan、N-Club、K-Plex 与 K-Core 等;其次是网络属性比较的角度,这类定义主要通过对网络内部连接数与外部连接数的对比来给出的,代表性的定义有 LS-Set、弱社区,除了这种自身属性比较外,社区图与其对应零模型的比较,网络节点的相似比较都属于这类定义方法.

2 信息网络社区发现的基本方法
Approaches to detecting web community

随着信息网络研究工作的深入,其社区发现方法不断呈现,本文对目前的典型算法初步归类为传统的发现方法、基于分割的方法、基于模块性质量优化的方法、基于动态模型的方法、基于谱分析的方法.

2.1 传统的发现方法

传统的发现方法主要有 3 类:图划分算法、层次聚类法、基于划分的聚类算法. 图划分算法的基本思想是将网络社区发现问题转化为这样一个图论问题:如何将一个网络划分成两个指定大小的社区,并使得割边数最少. 代表算法有 K-L 算法[1]与谱二分法[2]. K-L 算法的基本流程是先将网络随机地划分成指定大小的两个社区,然后进行下面迭代过程:对两个社区进行节点集的交换以使得模块质量 Q 有最大增量,同时,为了避免陷入局部极小,允许有 Q 值负增量的情形,重复上述操作多次并选取其中具有最大 Q 增量的操作执行,得到的结果转入下一轮迭代. 该算法的时间效率高($O(n^2 \log n)$),但对初始划分比较敏感,常用来对其他社区发现算法的结果进行求精. 谱二分法通过对网络邻接矩阵的 Laplacian 矩阵 $\boldsymbol{L}$ 进行谱分析,来计算网络的具有小割的划分方案. 网络的任何二分方案都可以用下标向量 $\boldsymbol{s}$ 来表示(若节点 i 被划分到第 1 个社区则元素 s_i 取值为 1,否则取值为 -1),网络的最小割可以形式化为

$$\boldsymbol{R} = \frac{1}{4}\boldsymbol{s}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{s}, \tag{1}$$

若令

$$\boldsymbol{s} = \sum_i a_i \boldsymbol{v}_i, \tag{2}$$

式中,$\boldsymbol{v}_i$ 是 $\boldsymbol{L}$ 的第 i 个特征向量,则有

$$\boldsymbol{R} = \sum_i a_i^2 \lambda_i. \tag{3}$$

式(3)中,λ_i 是 $\boldsymbol{L}$ 的第 i 个特征值. 直接求解最小割 $\boldsymbol{R}$ 的精确值很困难,但通过矩阵 $\boldsymbol{L}$ 的谱分析可以发现其特征值具有阶梯形分布,这启示本文可以采用如下方法构造 $\boldsymbol{s}$ 来计算 $\boldsymbol{R}$ 的近似值:根据与最接近0的特征值所对应的特征向量的各分量符号构造 $\boldsymbol{s}$(若 $\boldsymbol{v}_2{}^i > 0$,则 $s_i = 1$;反之 $s_i = 0$). 谱二分法具有较高计算效率且在一般情形下可以获得较好的社区结构. 从总的来看,图划分算法还有很多不足,诸如此类算法需要预先确定社区数与各社区大小,这对实际问题来说往往是不现实的,还有图划分是迭代二分的过程,具有不稳定性.

层次聚类法广泛地运用于社会学、生物学与营销学等多个领域中,根据簇形成方式可将其区分为聚合法(将每一个节点初始化为一个簇,然后依据相似度自底向上逐层合并簇直到指定层次)与分裂法(将连通的网络整体视为一个簇,然后依据相似度自顶向下逐层分裂簇直到指定层次). 相似度计算方法主要有簇间节点最短距离、簇间节点最长距离、簇间节点平均距离与簇中心距离,层次聚类过程常用树状图来表示,代表性算法有 Agnes、Birch、Rock 和 Chameleon[3]. 层次聚类算法具有直观易于理解的优势,但它在簇的形成过程中,若发现先前产生的簇结构质量不高,该法却无法回溯进行改进. 故层次聚类法常与其他具有节点迭代划分特征的聚类算法结合使用.

基于划分的聚类算法基本流程是将给定大小的

信息网络划分成指定社区数,根据最优化目标函数的原则不断调整各节点的簇分布特征直到迭代条件满足.目标函数的设计原则是使得簇间元素相似度尽可能小而簇内元素相似度尽可能大.该算法中的两个核心问题是簇中心如何表示与簇中心如何有效搜索.对应前一个问题有两种代表性的策略:抽象类中心表示策略与代表元素表示策略.抽象类中心表示的常用策略是对于数值型属性,取簇内数据对应属性值的平均值,而对范畴型属性,取簇内数据对应属性值的众数.采用这种策略的代表算法有 K-Means,K-Modes 和 K-Median,而代表元素表示策略是从簇中选取一个合适的元素作为该簇的代表,典型算法有 K-Medoids 和 PAM[3].簇中心如何有效搜索的问题是算法的可扩展性提出的,即当将原有算法扩展到大数据集时算法的时间性能是否可接受,CLARANS 算法[3]引入随机采样技术来搜索簇中心问题.

2.2 基于分割的方法

分割法源自最朴素的思想,即通过选取具有某种特性的边并删除来实现社区的发现.选择策略可以粗略地分为边介中性大者优先、边聚集系数小者优先、边信息中心度大者优先3大类.

边介中性的定义是经过该边的最短路径数,边介中性大者优先策略的代表是 GN 算法及其变体.GN 算法基本思路是计算边介中性,删除具有最大边介中性的边(若有多条满足此条件的边则随机选取),重复上面两步直到循环结束条件满足[4].GN 算法采用宽度优先搜索策略,效率很高($O(mn)$或 $O(n^2)$),它的提出极大地推动了社区发现领域的发展,在此基础产生了许多 GN 变体.根据电学上的基尔霍夫定律提出基于 Current-Flow 的 GN 算法,该方法时间复杂度为 $O((m+n)n^2)$或 $O(n^3)$.布朗粒子运动启示下的基于随机漫步的 GN 算法,其算法复杂度与前者相同,并且从数学上可以证明二者是等价的[5].考虑到 GN 算法的扩展性,文献[6-7]给出一种选取删除边的新方法,设当前连通子图的大小为 N,若 N 较小则采用原始 GN 算法的删除策略;反之,则随机选取至少 m 个节点,计算其边的介中性直到至少有一条边的介中性大于某个阈值(常为 $10\log N-25$),然后选择具有最大边介中性的边删除.为了克服 GN 不能发现有重叠的社区,文献[8]提出对于具有最大边介中性的边,若其两端点的点介中性比率在一个给定区间内则删除该边,否则只是对具有最高点介中性的点执行分裂操作.文献[9]给出 Split 介中性(各种 Split 产生的虚边的边介中性),删除具有最大 Split 介中性的虚边.文献[10]注意到 GN 算法偏好产生大小不平衡的社区结构,提出在计算边介中性时,只考虑路径的端点只能出现一次的最短路径集.

第2类选择策略主要涉及两个重要的定义:点聚焦系数(以该点为一个顶点的多边形实际个数/该点为一个顶点的多边形最大可能个数)和边聚集系数(包含该边的多边形实际个数/包含该边的多边形最大可能个数),该类算法的基本思路与 GN 算法是相一致的,区别在于选取具有最小边聚集系数的边[11-13].事实上,对比分析边聚集系数与边介中性,可以发现二者在本质上是一对相互对立的概念,值得指出是当多边形的边数逐步增多时,该算法生成的社区结构的局部性逐步减弱而全局特性逐步增强.

第3个选择策略[14]是从信息传播的角度来定义边的特性,边的信息中心度定义为信息在网络中传播效率在边删除前后的差,传播效率是指所有节点之间的距离的倒数的平均值.算法在迭代过程中选取具有最大信息中心度的边删除,与 GN 相比较,该算法发现的社区结构与 GN 算法的结果一致的程度非常高,但其效率比较低($O(m^3n)$或 $O(n^4)$).

2.3 基于模块性质量优化的方法

基于模块性质量优化算法的优劣取决于模块性质量测度函数的优化策略与模块性质量测度函数的设计思想.模块性质量函数是建立在随机网络中不存在明显的社区结构的假设上,可以通过与其对应空模型的比较来确定当前网络社区结构的质量.该函数的两种不同形式:

$$Q = \frac{1}{2m}\sum_{ij}(A_{ij}-P_{ij})\delta(C_i,C_j); \tag{4}$$

$$Q = \sum_{c=1}^{n_c}\left[\frac{l_c}{m}-\left(\frac{d_c}{2m}\right)^2\right]. \tag{5}$$

式中:$\boldsymbol{A}$ 是邻接矩阵;P_{ij}是随机网络中节点 i 与节点 j 之间的平均连接数;d_c 为簇 c 的度,l_c 为簇 c 的内部边数.优化策略大致可以分为贪心优化,谱优化与超启发式优化等.在这类基于模块性质量优化算法中,有些效率优先,有些准确率为重,其他则二者兼顾.

贪心优化策略的思想非常简单,就是初始化网

络为 n 个簇,n 为网络的大小,计算各边加入所带来的 Q 值增量并选取能使 Q 值增量最大的边加入. 该算法的效率为 $O((m+n)n)$ 或 $O(n^2)$. 考虑到实际信息网络具有很大的稀疏性且社区结构具有层次性,文献[14]设计了 Max-Heaps 这种精巧的数据结构将算法复杂度降低到 $O(md\log n)$. 采用普通贪心优化 Q 值策略的社区发现算法有着"对大社区偏好"的特点,而现实复杂网络中的社区具有较大的异质性,即各个社区大小具有较大差异. 为了解决这个问题,文献[15]提出对 Q 值增量进行规范化处理,合并社区 i 与社区 j 产生的 Q 值增量

$$\mathrm{d}Q_{ij} = 2\left(e_{ij} - \frac{a_i a_j}{2L_{\text{total}}}\right), \tag{6}$$

式中:

$$a_i = \sum_j e_{ij}; \quad e_{ij} = \frac{L_{ij}}{L_{\text{total}}},$$

L_{ij}是社区 i 与社区 j 之间的边数,L_{total}是网络中的总边数.

由于 Newman 法将各个社区视为同等重要,这导致该方法对大社区的偏好,为了避免问题的发生,引入规范化的 Q 值增量

$$\mathrm{d}Q'_{ij} = \frac{\mathrm{d}Q_{ij}}{a_i}2\left(e_{ij} - \frac{a_i a_j}{2L_{\text{total}}}\right). \tag{7}$$

选取具有最大 $\mathrm{d}Q_{ij}$' 的社区 i 与社区 j 进行合并. 注意,由于 $\mathrm{d}Q_{ij}' = \mathrm{d}Q_{ji}'$,实现时二者都要考虑,$Q$ 值的计算是用规范前的 $\mathrm{d}Q$ 来计算的,而不是规范后的 $\mathrm{d}Q'$. WT 算法[16]认为 Newman 法扩展性能不好是由于该方法生成的树状图具有不平衡性,条件$d = \log n$ 不再成立,从而导致该算法的复杂度常处于最坏情形,进而提出利用合并比 ratio 对 Q 值增量进行加权

$$\text{ratio}(c_i, c_j) = \min(|c_i|/|c_j|, |c_j|/|c_i|), \tag{8}$$

为了因过早合并而偏好大的社区,文献[17]将层次聚类思想与基于划分的聚类思想相结合,提出在迭代合并的过程中允许多个社区合并而不仅仅是两个社区的合并. 基于谱优化 Q 值的社区发现问题可以形式化为这样一个条件优化问题

$$Q = \frac{1}{4}\boldsymbol{s}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{s}, \quad \text{s.t.}\, s_i \in \{+1, -1\}. \tag{9}$$

式中,$\boldsymbol{B}$ 为模块性矩阵,$B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$,其中 k_i, k_j 分别指节点 i 与节点 j 的强度.

对于这个优化问题,可以这样构造 $\boldsymbol{s}$ 来求得 Q 值的近似解. 若最大特征值对应的特征向量的第 i 个分量大于 0,则 $s_i = 1$,反之 $s_i = -1$,基于上述理论可以进行二分法社区发现. 二分法在本质上是只利用主特征信息而将其他有用特征信息丢失了,这降低了结果社区结构的质量. 文献[18]提出类 PCA 方法来充分利用模块性矩阵 $\boldsymbol{B}$ 的特征信息. 文献[19]提出在二分过程中增添"最终调整"操作,使得顶点既可以在现有社区间移动又可以形成新的社区.

超启发式规则与方法的提出是为了克服基本局部搜索容易陷入局部极值的不足而提出来的,目前用于 Q 值优化的超启发式优化算法有遗传算法、模拟退火、极值优化与粒群优化等.

自从经典 Q 值表达式产生以来,它一直处于变化革新之中,呈现出很多 Q 值扩展版本,如加权图版本[20]、有向图版本[21]、有向加权图版本[21]、基于空模型比较的版本[22]、带符号加权图版本与二部图版本[23].

2.4 基于动态模型的方法

信息网络社区结构的形成与演变是一个动态时变的过程,基于动态模型的社区发现方法主要是通过对该过程的动态特征建模来实现社区结构的发现,目前主要 3 类模型:旋转模型、随机漫步模型与同步机制.

旋转模型的主要思想是通过网络节点与 Potts 自旋系统的旋转体对应将信息网络中的一个社区映射到一个处于基态或哈密顿极小态的磁畴. 文献[24]最早将网络节点表示为具有 q 状态的自旋体,社区结构质量函数等价地表示为自旋系统的能量函数,采用 SA 技术搜索系统的基态或能量极小态,每一次迭代都能得到一个社区结构,对于给定迭代次数中可以构建一个节点共现矩阵,最后可用街区模型化类似技术对矩阵重排来得到最终的社区结构. 与此类似,文献[25]提出 FRFIM 模型,该模型对 Zachary 网络节点关系断裂进行模拟;文献[26]从晶格系统的角度提出了社区发现的一个泛框架,基于模块性优化的策略与文献[24]中的 H 只是该框架下的一个特例;文献[27]发现旋转模型法也具有"粒度受限"的问题,针对此问题,最近的文献[28]从直接进行能量比较的角度设计能量函数而不是从当前社区结构与随机模型比较的角度去设计能量函数,提出如下算法:随机初始化社区结构,for 每一节点 do {寻找使能量减少最多的移动策略,移动并做移动标记},重复上述 for 循环直到迭代次数满足或达

到局部最小能量值. 对当前各社区检测,看是否存在能产生更低能量的簇合并,若有则合并且重复 for 循环. 重复上述过程 t 次选取具有最低能量的解作为其最佳解. 该算法具有优于当前其它的精度并且具有很好的抗噪性能.

随机漫步模型算法设计的基本原则是:若网络具有很强的社区结构性则随机漫步者多数时间是停留在社区内部边上. 文献[29]第 1 次运用布朗粒子运动来测度节点间的距离,将节点 i 到节点 j 的距离 d_{ij}定义为布朗粒子从节点 i 运动到节点 j 平均需要的步数,给出节点 i 的全局吸引子(在整个网络中距离节点 i 最近的节点)与节点 i 的局部吸引子(在节点 i 的某个邻域内距离节点 i 最近的节点)两个定义,并构造基于局部(或全局)吸引子的社区发现算法. 紧接着文献[30]借鉴观景法思想(如果两个节点处于同一个社区内,二者对整个网络产生的视图差异应该非常小),节点 i 对应的视图为 $\{d_{i1},\cdots,d_{i,i-1},d_{i,i+1},\cdots,d_{iN}\}$,给出节点 i 与节点 j 的非相似度计算方法. 文献[31]提出了一种对具有共同邻居有偏好的布朗运动,即若两个节点之间的共同邻居越多则两者之间的跳变概率越大并在此基础上给出算法 Netwalk. 文献[32]则认为两个节点之间距离是与在固定跳变次数 t 能从节点 i 跳到节点 j 的概率 P_{ij}^t相关的,并给出节点距离与社区距离. 文献[33]提出了一种建立在类似信息传播的信号发送过程之上的社区发现算法.

同步现象广泛存在于自然界与社会体系之中,若将网络节点映射到具有不同初相的振荡器,则可利用同步机制相关原理进行社区发现的研究. 文献[34]在 Kuramoto 模型的基础上给出 t 时刻的节点相关度矩阵,该矩阵的元素为

$$\rho_{ij}(t) = \cos(\theta_i(t) - \theta_j(t)). \qquad (10)$$

式中,$\theta_i(t)$是指第 i 个振荡器在时刻 t 时的相角,对此矩阵二值化可得动态关联矩阵,对动态关联矩阵进行谱分析可得社区数. 文献[35]在评价改变率(Opinion Changing Rate,OCR)模型的基础上给出振荡器相位演化的动态机制

$$\dot{x}_i(t) = w_i + \frac{\sigma\sum_{j\in Ni} b_{ij}^{\alpha(t)} \cdot \sin(x_j - x_i)\beta e^{-\beta|x_j - x_i|}}{\sum_{j\in N_i} b_{ij}^{\alpha(t)}}. \qquad (11)$$

式中:σ 为连接强度;β 为常量,当两个振荡器的相位距离大于 β 时则这两个振荡器的交互关闭;b_{ij}是边 E_{ij}的介中性,在 t 时所有具有相同频率 $\dot{x}_i(t)$的振荡器聚为一类. 从系统完全同步($\alpha(0)=0$)开始,逐步递减,每一次迭代对应一个划分,形成一个层次聚类树状图. 算法时间复杂度为 $O(mn)$或 $O(n^2)$.

2.5 基于谱分析的方法

谱分析社区发现方法的主要思想是对网络邻接矩阵或其变体,诸如转移矩阵 $\boldsymbol{T}$(利用行元素和去归一化的邻接矩阵),Laplace 矩阵,右随机矩阵 $\boldsymbol{R}$($\boldsymbol{T}$ 的转置)等,进行特征谱分析来推导出网络的结构层次分析. 最初,研究人员利用转移矩阵的特征向量提取网络的社区结构信息,通过分配比的计算可以确定一个特征向量是否为"固定的",进一步可以确定对应社区的大小. 文献[36]在 Laplace 矩特征谱的基础上提出了一种非常精妙的方法,它注意到在特征向量的各分量中,同一社区中的节点所对应位置的分量值非常接近这样一个事实,将网络节点嵌入到 M(M 为特征向量的个数)维空间中去,然后计算这些节点在这个空间中的距离,根据距离进行聚类. 文献[37]将信息网络视为电流网络并给出节点对之间的有效电阻,跳变概率与广义距离,在此基础上进行层次聚类,该算法的效率比较低($O(n^3)$).

由于实际网络的社区结构并非很清晰,仅仅依靠主特征向量不能够很好发现真正的底层社区结构,文献[38]基于同一社区中的节点所对应位置的分量值相关性非常强这一事实,将其他 M 个特征向量引入计算节点相似度.

2.6 各种方法的比较分析

正如"No Free Lunch Theorems"所言,最好的社区发现方法往往总是局部适用且与应用场景相关的,每类算法都有其优势与不足. 以图划分、层次聚类和划分法为代表的传统发现方法有着思想简单直观、易于实现且都具有良好的数学理论基础,但也有其缺陷,如图划分法与 K-Means 算法需要事前指定网络中存在的社区数,而这种先验知识的获取往往是很困难的,甚至是不可能的. 层次聚类在构建"树形图"时无法修正之前的错误划分. 分割算法产生的社区质量严重依赖于分裂标准的选取,不论是边中心度、边聚集系数还是边信息中心度都有其适用的场合,分裂标准的选取没有实现任务的自适应. 源于谱图理论的谱聚类具有诸多优势,如仅与数据点的数目有关,而与数据对象维数无关,这使得有效地避免了"维数灾难"问题. 它不对数据的全局结构作假

设,可以避免“局部最优”的问题,但它同样存在 Laplacian 矩阵或其变体的“迹优化”问题. 影响 Q 值优化法性能的一个重要因素是 Q 值表达式,尽管 Q 值表达式一直在改进,但没有改变模块性质量优化的本质缺陷,所有此类算法都坚持 Q 值越大则所发现的社区结构就越好,然而事实并非如此,究其原因是此类算法建立在随机图没有社区结构的假设之上,而此假设并非永真. 还有一个众所周知的“粒度有限”的缺点,即算法对于阶数比较小但却是定义良好的社区无能为力. 到目前为止,只有 Danon 等人对社区发现各类算法进行系统的量化比较研究[39].

3 结束语
Concluding remarks

随着人们生产生活的信息化与分工协作的深入,信息网络的社区发现研究吸引着越多的研究者投入到这个领域中来,经过来自不同领域的科研学者协同努力,该领域已有数目繁多的理论算法与原型系统. 但是,其中还有很多的问题需要解决. 首先,对社区的定义缺乏统一的认识;其次是对空模型的定义缺乏统一的认识;另外是模型的选择缺乏理论性的指导,面对众多的模型,应用系统的开发者该如何去选取适合当前问题的模型,这给开发者们提出了一个很大的难题,目前这还没有引起信息网络研究领域足够的重视,其实这是一个很有现实意义的课题;最后就是目前信息网络中社区发现模型多数依赖节点关系,其实节点本身属性对社区结构也有着不可忽视的作用.

参考文献
References

[1] Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs[J]. Bell Sys Tech J, 1970, 49(2): 291-308
[2] Barnes E R. An algorithm for partitioning the nodes of a graph [J]. SIAM J Alg Discr Meth, 1982, 4(3): 541-550
[3] Jiawei H, Micheline K. Data mining: Concepts and techniques [M]. 2nd Ed. Morgan Kaufmann, 2006
[4] Girvan M, Newman M E. Modularity and community structure in networks[J]. PNAS, 2002, 99(12): 7821-7826
[5] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Phys Rev E, 2004, 69(2): 026113
[6] Tyler J, Wilkinson D, Huberman B. Email as spectroscopy: Automated discovery of community structure within organizations[C] // International Conference on Communities and Technologies, 2003: 81-96
[7] Wilkinson D M, Huberman B A. A method for finding communities of related genes[J]. PNAS, 2004, 101(S1): 5241-5248
[8] Pinney J W, Westhead D R. Betweenness-based decomposition methods for social and biological networks[J]. Interdisciplinary Statistics and Bioinformatics, 2006: 87-90
[9] Gregory S. An algorithm to find overlapping community structure in networks[C], PKDD, 2007: 91-102
[10] Chen J, Yuan B. Detecting functional modules in the yeast protein-protein interaction network[J]. Bioinformatics, 2006, 22: 2283-2290
[11] Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks[J]. PNAS, 2004, 101: 2658-2663
[12] Zhang P, Wang J, Li X, et al. Clustering coefficient and community structure of bipartite networks[J]. Physica A: Statistical Mechanics and its Applications, 2008, 387(27): 6869-6875
[13] Fortunato S, Latora V, Marchiori M. Method to find community structures based on information centrality[J]. Physical Rev E 2004, 70(5): 056104
[14] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks[J]. Phys Rev E, 2004, 70(5): 066111
[15] L Danon, Guilera A D, Arenas A. The effect of size heterogeneity on community identification in complex networks[J]. J Stat Mech, 2006, 11(10): 11010
[16] Ken W, Toshiyuki T. Finding community structure in mega-scale social networks[C] // Procedings of the 16th International Conference of World Wide Web, 2007: 1275-1276
[17] Blondel V D. Fast unfolding of community hierarchies in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008(10): 10008
[18] Newman M E J. Finding community structure in networks using the eigenvectors of matrices[J]. Physical Rev E, 2006, 74(3): 036104
[19] Sun Y. Improved community structure detection using a modified fine tuning strategy[J], Euro phys Lett, 2009, 86(2): 28004
[20] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Phys Rev E, 2004, 69: 066133.
[21] Arenas A. Community structure in directed networks[J]. New J Phys, 2007, 9: 176
[22] Marco G, Robert G, Dorothea W. Significance-driven graph clustering[C] // AAIM, 2007: 11-26
[23] Barber M J. Modularity and community detection in bipartite network[J]. Phys Rev E, 2007, 76(6): 066102
[24] Reichardt J, Bornholdt S. Detecting fuzzy community structures in complex networks with a Potts model[J]. Phys Rev Lett, 2004, 93(21): 218701
[25] Son S W, Jeong H, Noh J D. Random field Ising model and community structure in complex networks[J], Eur Phys J: B, 2006, 50(3): 431-437
[26] Reichardt J, Bornholdt S, Statistical mechanics of community detection[J]. Phys Rev E, 2006, 74(5): 016110
[27] Jussi M. Kumpula, Jari S, et al. Limited resolution in complex network community detection with Potts model approach[J]. The European Physical Journal: B, 2007, 56(1): 41-45
[28] Peter R, Zohar N. A highly accurate and resolution-limit-free Potts model for community detection[J]. arXiv, 2008(3): 2548v2
[29] Zhou H, Distance, dissimilarity index and network community structure[J]. Phys Rev E, 2003, 67(6): 061901

[30] Zhou H. Network landscape from a Brownian particle's perspective [J]. Physical Rev E,2003,67(2):041908
[31] Zhou H,Lipowsky R. Network Brownian motion:A new method to measure vertex-vertex proximity and to identify communities and subcommunities[C]//ICCS 2004:1062-1069
[32] Pascal P,Matthieu L. Computing communities in large networks using random walks[C]//ISCIS 2005:284-293
[33] HU Yanqing,LI Menghui,ZHANG Peng,et al. Community detection by signaling on complex networks[J]. Phys Rev E,2008,78(1):016115
[34] Arenas A,Díaz G A,Perez V C J. Synchronization reveals topological scales in complex networks[J]. Phys Rev Lett,2006,96(11):114102
[35] Boccaletti S. Detecting complex network modularity by dynamical clustering[J]. Phys Rev E,2007,75(4):045102
[36] Donetti L,Muñoz M A. Detecting network communities:a new systematic and efficient algorithm[EB/OL]. (2004-10-12). http://arXiv.org/pdf/cond-mat/0404652
[37] Alves N A. Unveiling community structures in weighted networks [J]. Phys Rev E,2007,76(3):036101
[38] Capocci A,Servedio V D P,Caldarelli G,et al. Detecting communities in large networks[J]. Physica A,2005,352:669-676
[39] L Danon,Diaz G A. Comparing community structure identification [J]. Journal of Statistical Mechanics:Theory and Experiment,2005,29(9):09008

Research on community detection in information network

HUAN Faliang[1,2] XIAO Nanfeng[1]
1 School of Computer Science & Engineering,South China University of Technology,Guangzhou 510006
2 Faculty of Software,Fujian Normal University,Fuzhou 350007

Abstract With the deepening informationization of people's production and daily life,community detection in information network is drawing more and more attention of researchers in different domains. Based on the brief introduction to elementary concepts and principles related to the study,this paper puts emphasis on the comparison between and analysis of various detection methods,and in the end summarizes the techniques and trends of community detection in information network.

Key words information network;community detection;complex network;clustering