网络社区动态角色挖掘算法研究*

马瑞新，孟繁成，王涵杨，崔亚杰
（大连理工大学 软件学院，辽宁 大连 116621）

摘　要： 传统的社区挖掘以社区为单位，忽略了社区内部成员的性质和地位。为了提高社区挖掘的精度，为个性化推荐提供一个优化的基础平台，基于优先情节和增长定律，提出了一种新颖的动态角色挖掘算法。首先根据节点的度数分布逆向推导社会网络的形成演化机制，构造网络时间轴；然后根据时间轴逐步向网络中添加新节点，同时进行社区挖掘和角色划分。在人工网络和真实世界网络上进行了多次测试，并与 G-N 算法进行了比较，取得了较好的结果。实验证明，应用动态角色挖掘算法得到的社区都是强连通社区，具有较高的准确性和实用价值。

关键词： 优先情节；增长定律；网络时间轴；角色划分
中图分类号： TP301.6　　**文献标志码：** A　　**文章编号：** 1001-3695(2012)04-1279-03
doi:10.3969/j.issn.1001-3695.2012.04.021

Dynamic role assorted discovery of network community

MA Rui-xin, MENG Fan-cheng, WANG Han-yang, CUI Ya-jie
(*School of Software Technology*, *Dalian University of Technology*, *Dalian Liaoning* 116621, *China*)

Abstract: Traditional community discovery algorithms study network members in community groups, while ignoring the inner members' characteristics and status. In order to improve the discovery accuracy and provide an optimized platform for personalized recommendation system, this paper put forward a novel dynamic community discovery algorithm on the basis of complex priority and the grouth theorem. It firstly derived the social network's mechanism of formation and evolution in negative direction according to the node-degree, at the same time, it constructed the time axis, then gradually put nodes into the network and divided them into different communities and gave distinct roles. It tested the algorithm on both artificial networks and real world networks, and compared with G-N. Experimental results show that, the discovery community are all strong connected communities and this algorithm has great pratical value as well as high accuracy.

Key words: complex priority; thegrouth theorem; network time axis; role assorted

0　引言

复杂网络是描述和理解复杂系统的一种重要表现形式。任何复杂系统都可以从实际问题出发，抽象成由相互作用的个体构成的网络[1]，因此，网络为复杂系统的研究提供了一个崭新、清晰、直观的研究平台。现实世界中存在很多类型的复杂系统，如引文系统[2]、蛋白质大分子系统[3]、食物链系统等，系统中每个独立的个体均可抽象为网络中的节点，节点之间的边则是系统中个体之间按照某种规则形成的关系的表现。

社会网络是对现实生活中人际关系网络的扩展和补充，因此它在人们的日常生活中扮演十分重要的角色。大量研究表明，社会关系网络对某项组织活动的成败起到了决定性的作用；复杂网络具有明显的社区结构[4]。社区结构是网络模块化与异质性的反映，表示真实网络是由许多不同类型节点组合形成的。深入研究网络的社区结构，在复杂网络中自动搜寻或发现社区，具有重要的研究价值[5]。

传统的社区挖掘更加注重某一时间点的静态网络结构，而忽视了个体的能动性和彼此之间的影响力。然而，个体之间的交流互动决定了社会网络的关系结构，因此，只有了解了个体的行为规则，才能更好地进行社区结构分析。大量研究表明，真实网络受到两个定律的控制：优先情节和增长定律[6]。优先情节的内涵在于，节点吸引链接的能力与当前节点拥有的链接数量成正比；增长定律指出，早出现的节点比晚出现的节点有更多的机会积累链接。因此，根据社会网络中节点的度数分布构造时间轴，逆向推导社会网络的形成演化机制；在模拟社会网络的形成过程中，对社会网络中的节点进行角色划分，同时进行社区挖掘。因此，本文提出一种以个体为单位的网络社区动态角色挖掘算法(dynamic role assorted discovery algorithm, DRADA)。

1　社区挖掘

复杂网络中对社区挖掘的分析起源于社会学的研究工作者 Newman 等人[7]及其他相关学者的研究成果。在现有已知的社区发现算法中，以 Newman 提出的基于边中介性的 G-N 算法影响最为广泛。近年来，我国的学者在社区挖掘算法研究方面取得了重大的研究成果。胡健等人[8]提出了一种基于边凝

收稿日期： 2011-09-07；**修回日期：** 2011-10-25　　**基金项目：** 国家自然科学基金资助项目(60803074)

作者简介： 马瑞新(1975-)，男，辽宁庄河人，讲师，博士，主要研究方向为电子商务、社区挖掘、群智能(teacher_mrx@126.com)；孟繁成(1989-)，男，辽宁大连人，本科生，主要研究方向为电子商务；王涵杨(1986-)，男，辽宁大连人，硕士，主要研究方向为电子商务、社区挖掘；崔亚杰(1989-)，男，河南郑州人，硕士，主要研究方向为电子商务、社区挖掘.

聚系数的社区发现算法,并引入了强连通社区的概念,取得了相当不错的效果;郭崇慧、张娜提出利用共邻矩阵对社会网络进行结构分析和划分,降低了算法的时间复杂度;何晓东等人[9]根据聚类融合的遗传算法对复杂网络中的社区结构进行划分,能够在保证低开销的前提下提高社区划分的模块度。很多学者将群智能算法引入社区挖掘的优化研究中,并且取得了卓越不凡的成就。

近年来,社区挖掘研究在政治、商业、生物、物理等众多学科领域都有极其重要的研究意义。然而,目前存在的社会网络挖掘算法仍有几点不足:现存的社会网络挖掘算法注重对网络拓扑结构的挖掘,缺少对网络节点性质的分析,用户定位模糊,没有实际的应用价值;大多数挖掘算法存在严重的局限性,如针对无向网络、无权网络、静态网络等,算法的适应性、可扩展性不强;传统的社区挖掘算法注重对静态社区结构的研究,缺少对个体行为如何影响社会网络演化的分析;挖掘的结果是以社区为单位,缺少对社区内部成员的分析,缺少差异化的服务,不能满足当前 SNS 网站发展的需求。

2 网络社区动态角色挖掘

2.1 算法思想

角色划分的概念来自于方守兴的特殊人物法则[10],他提出以下三类人在网站的传播过程中起到非常重要的作用:

a)内行,在某些领域具有丰富经验知识的人。此类用户不仅仅有数量众多的邻接用户,而且具有较高的被信任程度或权威度。许多公司的创始人、网站的创始人都是该相关领域的内行,或者有能力召集内行。通过对社会网络的深入研究发现,几乎每个社团组织、社区结构都有属于其内部结构的内核,该内核对社区中其他用户的行为起指导作用,主导社区的形成。

b)联系者,具有广泛社会关系、富有社交天赋的人群,他们的社会关系可能涉及好几个不同的社会圈体。六度分隔理论虽然仍停留在假说阶段,但是其与互联网的紧密结合,体现了巨大的商业价值。从联系者节点中延伸出的弱连接关系,成为了催生 SNS 网站的驱动力所在。

c)推荐者,负责说服目标用户接受信息。然而在真实社会网络中,每一个用户都可以成为推荐人,他们可能是目标用户的最近邻居,也可能是商务网站的销售人员,因此具有模糊不确定性。由于每个目标用户都具有不同的推销者,且推销者对社会网络的结构挖掘、拓扑结构分析和稳定性研究没有实用价值,因此本文的社会网络分析忽略对推销者的深入研究。

这三种人被称做是网络中的信息意见领袖。在口碑传播的过程中,意见领袖的话往往对其他用户形成指导作用。特别是新生用户在不明白如何高效利用网络信息为自己服务的情况下,通过意见领袖提供的建议和心得,可以较快地适应网站和在线社会网络服务。

优先情节和增长定律说明,新节点偏向于与链接多的节点建立连接,因此,早出现的节点会比晚到的链接少的节点增长得快,而且每个节点吸引的链接都与其当前的链接数量成正比。正是由于增长和优先情节的存在,导致了真实网络存在的幂率。根据优先情节和增长的定律,本文假设度数最多的节点出现的时间最早,度数最少的节点出现的时间最晚。在该假设下,以时间为轴,根据现有的静态社会网络拓扑结构逆向模拟该社会网络的生成及演化机制。最早出现的节点最终演化为中心节点。中心节点十分特殊,在任何存在中心节点的网络中,它们都对网络结构起到关键作用,使该网络呈现小世界的特点。中心节点与异乎寻常的多的节点之间存在链接,为系统中任意两个节点创造了联系捷径。因此,对于中心节点的研究具有异乎寻常的重要性。根据中心节点的特殊影响力对社区结构进行挖掘和细化,能够在保证算法效率的前提下取得较高的社区模块度。

2.2 算法详细步骤

在特殊人物法则的影响下,本文对同一社区中的节点进行角色划分,提出社区种子、联系者和普通用户的概念。

社区种子:社区的核心领导者,限定社区形成的主题和目标,并且引导社区内部成员的行为。通过对显示生活中的社会团体进行分析研究,笔者发现,绝大多数社会团体中不仅仅只有一位团体负责人,这些团体负责人如同一个班级中的委员会,对团体内的事务进行表决和管理。因此,本文在社区种子的基础上,提出候选社区种子的概念。

联系者:负责不同社区之间的消息传播和互动。如果没有联系者,信息就不能像病毒一样席卷整个网络。普通成员通过与本社区内联系者的链接而与整个世界联系到一起。

基于角色划分的动态社区挖掘算法的详细步骤如下:

a)将所有节点按照度数的降序排列组成一张列表 L_{degree},社区集合 $\{SN\}$ 初始化为空;将列表中的第一个节点取出,作为第一个社区的成员。

b)从第二个节点开始对列表从上往下进行检查,如果列表中的自由节点 i 与已发现的社区 SN 之间的相似性小于最小相似性阈值 $\bar{\delta}$,该节点成为新社区的第一个成员,将其加入社区种子集合中;若自由节点仅仅与一个已存社区的余弦相似性大于最小相似性阈值 $\bar{\delta}$,且该自由节点是社区中的第二个成员,则标志为候选社区种子,否则,仅将其加入社区中,社区关系向量 $V_{SN} = V_{SN} + V_i$;若节点与多个社区的余弦相似性均大于最小相似性阈值,则将节点加入到与之关系最相近的社区中,并将其标志为社区联系者。

c)重复进行 b),直至列表末端。

从上述过程中可以看出,度数最高的节点首先出现在社会网络中,新加入的每一个节点都会与已存在的节点进行相似性度量,判断自己是不是与已存在的节点属于一类。如果是,就加入到已存在的节点团体中;否则,自成一派。在真实的社会网络中,每时每刻都有新的节点出现,虽然算法不能断定其所属社区,但是可以根据每个节点的行为特征和属性,对其进行预测和引导,促进社会网络的收敛。

二八法则同样适用于社会网络,即 20% 的节点拥有 80% 的链接,而 80% 的节点仅仅拥有 20% 的链接。因此,本文提出,度数分布排名前 20% 的用户成为网络中的核心节点。核心节点的关系链接众多,因此,在进行最小相似性阈值判断时,要区别对待。本文设定,$\bar{\delta}(\text{core}) = 0.5$,$\bar{\delta}(\text{normal}) = 1/n$,$n$ 为社会网络的规模,即核心节点具有大于 50% 的相似度才将其归为一类;否则,它将引导新的社区形成。

2.3 结果分析及对比

1)划分结果对比

本文使用悲惨世界数据集[11]和 SNS 科技论文管理平台数据集作为复杂社会网络的测试数据集,对网络社区动态角色挖

掘算法进行实验验证。

Knuth 根据 Victor Hugo(维克多 · 雨果)的小说《Les Misèrables》整理了其中的人物关系网络。网络中的节点表示小说中的角色,边表示两个角色同时出现在一幕或多幕中。网络中有五个主要人物:主人公 Jean Valjean(节点 11)、探长 Javert(节点 27)、神父 Bishop Myriel(节点 0)、女工 Fantine(节点 23)及其女儿 Cosette(节点 26)。研究人际网络中关键的边(即人物之间的联系)对网络整体性能的影响,对舆情和疾病等的传播具有非常重要的意义。

为了更好地检测社区挖掘的效果,笔者在研究过程中搭建了一个 SNS 科技论文管理平台[12],该平台以科技论文类网站为研究对象,分析科研人员对论文网站的服务需求,设计并实现了个性化定制服务与个性化推荐服务为一体的功能。为更好地验证本文算法的有效性,笔者从该平台中随机抽取了 99 名用户组成了一个全新的社会网络测试数据集,如图 1 所示。

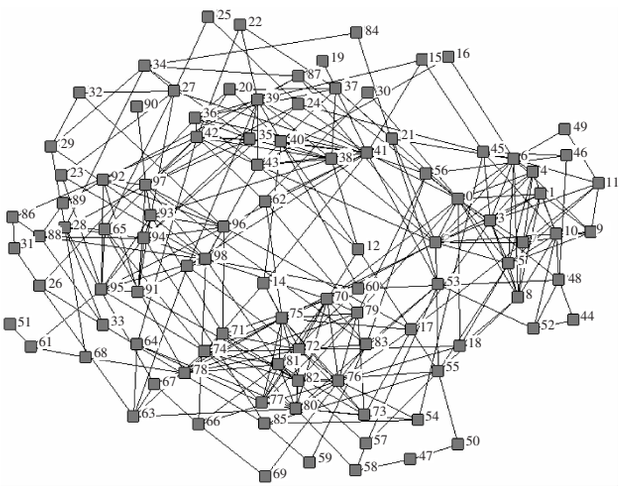

图 1　SNS 科技论文管理平台数据集

图 2、3 分别展示了本文算法及 G-N 算法对悲惨世界数据集和 SNS 科技论文管理平台数据集的划分结果。

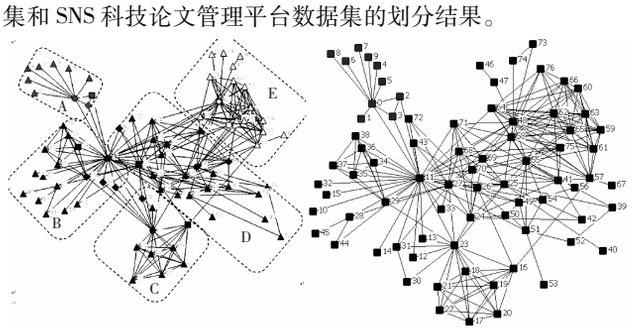

图 2　DRADA 与 G-N 算法对悲惨世界数据集的划分结果对比

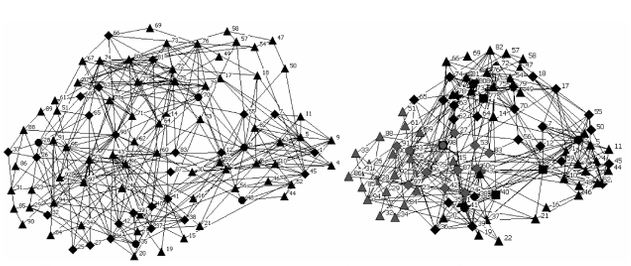

图 3　G-N 算法与 DRADA 对 SNS 科技论文数据集的划分对比

对比图 2、3 可以发现,DRADA 算法能够在无任何先验信息的情况下,降低社区挖掘的粒度,细化社区挖掘的精度。

此外,DRADA 算法能够有效地寻找社区中的社区核心人员和联系者。这些核心人员的发现对于个性化推荐系统的实现以及对新产品的推广都有十分重大的影响。

2)模块度对比

模块度作为衡量无关系加权网络结构划分的量化标准,受到了广大社会网络研究学者的认可。本文根据模块度的计算方法对不同算法的社区划分结果进行度量,结果如表 1 所示。

表 1　社区挖掘结果的模块度对比

name	Karate Club	Dolphins	SNS 科技论文数据集	Les Misèrables
G-N	0.395	0.381	0.046	0.082
Fast G-N	0.360	0.379	0.041	0.074
Polish	0.358	0.226	0.027	0.107
DRADA	0.401	0.379	0.301	0.486

表 1 中的内容为 DRADA 算法与其他经典社区算法的模块度对比结果。与其他经典社区挖掘算法相比,动态角色挖掘取得了最高的模块度,并且网络越复杂,算法的效果越明显。此外,DRADA 根据自由节点与社区特征向量的邻接相似性进行判断和规划,社区中每一个节点的对内连接都多于对外连接,因此,最终得到的都是强连通社区。

3　结束语

动态角色挖掘算法受到社会分化和特殊人物法则的启发,基于优先情节和增长定律,根据网络中节点的核心程度对其进行角色划分,不仅能够有效准确地寻找社会网络中的结构团体,而且对社区内部进行层次划分,有助于人们深入理解复杂社会网络的内在机理和系统结构。

参考文献:

[1] 高霖. 社会网络动态性及网络环境中的分布式搜索策略研究[D]. 合肥:中国科学技术大学, 2009.

[2] 艾伯特—拉斯洛 · 巴拉巴西. 链接网络新科学[M]. 徐彬,译. 长沙:湖南科学技术出版社, 2007.

[3] 马汀 · 奇达夫,蔡文彬. 社会网络与组织[M]. 王凤彬,朱超威,译. 北京:中国人民大学出版社,2007.

[4] BARABASI A L, ALBERT R. Emergence of scaling in random networks[J]. Science,1999,286(5439):509-512.

[5] ZHANG Yu-zhou, WANG Jian-yong, WANG Yi, *et al*. Parallel community detection on large networks with propinquity dynamics[C]// Proc of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM Press, 2009:997-1006.

[6] KRAPIVSKY P L, REDNER S, LEYVRAZ F. Connectivity of growing random networks[J]. Physical Review Letters, 2000, 85(21): 4629-4632.

[7] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2): 026113.

[8] 胡健,董跃华, 杨炳儒. 大型复杂网络中社区结构发现算法[J]. 计算机工程,2008,34(19):92-93,100.

[9] 何晓东,周栩,王佐. 复杂网络社区挖掘——基于聚类融合的遗传算法[J]. 自动化学报,2010,36(8):1160-1170.

[10] 王娴,谢弛,荣雪,等. SNS 网站运营的现状和未来趋势研究[EB/OL]. (2008-12-03). http://media. people. com. cn/GB/22114/119489/140165/8454258. html.

[11] KNUTH D E. The Stanford graphbase: a platform for combinatorial computing[M]. Boston: Addison-Wesley, 2009.

[12] http://www. linkscholar. com/[EB/OL].