基于扩散理论的动态社区挖掘

马瑞新[1] 邓贵仕[2]
(大连理工大学软件学院 大连 116621)[1] (大连理工大学管理与经济学部 大连 116621)[2]

摘 要 针对动态社区挖掘问题进行分析和研究,基于优先情节和增长定律,根据节点的度数分布,提出以时间为轴动态模拟社会网络的形成演化机制,同时进行社区划分。以 Zachary Club 和 Les Miserables 网络作为实验数据集,对提出的算法进行了实验验证,结果表明,该算法挖掘到的社区都是强连通社区,能够动态、精确地挖掘网络中存在的社区结构,具有较高的实用价值。

关键词 动态挖掘,优先情节,增长定律,时间轴,强连通社区
中图法分类号 TP312 **文献标识码** A

Research of Dynamic Community Discovery Based on Diffusion Theory

MA Rui-xin[1] DENG Gui-shi[2]
(School of Software, Dalian University of Technology, Dalian 116621, China)[1]
(Facility of Management and Economics, Dalian University of Technology, Dalian 116621, China)[2]

Abstract In terms of the dynamic community discovery problem, this article came up with the idea of basing on the priority complex and thegrouth theorem and according to the nodes degree distribution to construct the time axis to dynamically simulate the social network's mechanism of formation and evolution, simultaneously divide the community structures. We used Zachary Club and Les Miserables as the experiment data to test our algorithm. Experimental results show that, the communities that the algorithm gets are all strong connected communities, it is able to dynamically, accurately discover the community structure with high practical value.

Keywords Dynamic discovery, Priority complex, Thegrouth theorem, Time axis, Strong connected communities

1 引言

大量研究表明,深入研究社区结构,在复杂网络中自动搜寻和发现社区具有重要的研究价值[1]。例如社会网络中的社区可以揭示具有相同兴趣爱好、社会背景的团体;引文网络中的社区结构可用于提高文章检索的准确性,实现信息过滤、研究热点追踪和网络情报分析的功能;蛋白质网络中的社区结构可用于发现功能相关的结构单元等等。

传统的社区挖掘注重分析静态的社会网络结构,而忽视了具有能动性的个体,因此,社区挖掘只针对社会网络在某一时刻的静态拓扑结构进行分析与研究,即结构主义的社区挖掘[2]。然而,在真实的社会网络的形成过程中,行动者的属性、认知或行为都会影响其所嵌入的社会网络的形成和演化。文献[3]指出,社会网络属于无尺度网络,大多数节点只有少数几个链接,这些数量众多的较小节点与少数几个巨大的拥有无数链接的中心节点共存,中心节点确定整个网络的拓扑结构。因此,本文提出,社区挖掘的关键在于寻找网络中的关键节点。

根据扩散模型[4]的理论分析可知,互联网上的每一项创新都有精确定义的传播速度,代表了该产品(页面)被用户接收并被介绍给别人的可能性。研究表明,创新的传播速率呈幂率分布[5],即产品刚刚出现的时候传播速率会很快,接下来会逐渐变慢,直至稳定(用户无意之中引用或浏览了该网页)。扩散过程受到两个定律的控制:增长和优先情节[6]。优先情节意味着每个节点吸引链接的能力都与其当前所拥有的链接数量成正比。增长意味着早期出现的节点比晚期出现的节点有更多的机会积累链接,因此,增长特性使先存在的节点拥有无可比拟的优越性,成为链接最为丰富的节点。受到优先情节和增长定律的启发,本文提出,根据某一时刻的静态社会网络结构对社区的形成和演化进行逆向推导,以时间为轴,依据度数的大小判断节点出现时间的早晚,模拟社会网络的形成及演化机制,同时进行社区结构划分。

2 常用社区挖掘算法

复杂网络中社区挖掘的研究起源于社会学的研究工作者 Girvan 和 Newman 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 GN 算法[7]影响最为广泛。然而,GN 算法对计算的需求太高,仅仅能够支持对一万个节点以下的社区进行分析,后来提出的快速 G-N 虽然能够以较低的时间复杂度完成对社区的寻找和划分,但是却以牺牲准确率为代价;Kernighan 算法是一种基于贪婪算法原理[8],在已知网络的确切规模下,将网络分割为两个大小已知的社区的二分法,由于该算法需要较多的先验信息并且不支持对存在奇数个数的社区的挖掘,因此

马瑞新(1975—),男,博士,讲师,主要研究领域为电子商务、社会网络分析、群智能,E-mail:kara0807@gmail.com。

在实际网络中难以得到较好的应用；Radichil[9]等人在GN算法的基础上提出基于网络中三角环数量的快速分裂算法，但如果Web网络中存在的三角形数目很少，那么相应的边聚集系数就会很小，算法就无法正确地搜索网络中的社区，特别是针对树形结构的网络。因此，尽管人们对于复杂网络的社区发现问题已进行了大量的研究，但是仍然存在一些目前无法解决的基本问题，如社区的概念虽然使用广泛，但却缺少严格的数学定义；多数算法都针对静态网络，不能动态地分析网络结构变化；此外，现有的社区挖掘算法仅仅根据网络的拓扑图进行结构挖掘，忽略了个体对社区形成的作用。因此，复杂网络中社区发现的研究还未成体系，很多工作有待完善。

3 基于扩散理论的动态社区挖掘

在本文中，使用向量空间模型[10]表示网络中节点的社会关系，同时使用向量描述社区的关系特征，以便进行相似性度量。社区的关系特征向量由社区内节点的关系向量共同组成。

3.1 算法思想

虽然大多数真实网络存在巨大差异，但它们都有一个共同之处：增长。它们从少数几个节点开始，随着节点的不断增长，网络的规模与日俱增，逐渐达到当前的数量。在增长的过程中，不断地与其他节点建立链接。优先情节便是指，在建立链接的过程中，如果同时面对的两个节点中，一个的链接数量是另外一个的两倍，那么选择它的概率将是另外一个的两倍。本文受优先情节和增长定律的启发，提出链接数量众多的节点出现的时间要远远早于只有少数几个链接的节点。因此，本文根据真实社会网络中节点的度数分布，判断其出现时间的早晚，以时间为轴，模拟社会网络的形成及演化机制。

社会网络中节点的度数分布同时符合二八法则[11]，即20%的节点拥有80%的链接数量。将社会网络中度数排名前20%的用户称为核心节点，并对其采取特殊的相似性度量标准。

3.2 算法基本步骤

算法分为两个过程，基础划分过程和新增成员动态规划过程。基础划分是对于网络中现有的成员进行社区划分，得到基本的社区结构；而新增成员动态划分是在动态检测网络变化的过程中，对网络中新出现的成员及时地进行社区动态规划，以便更好地进行用户定位，从而为其提供更加优秀的个性化信息服务。

基础划分过程如下所示：

(1)将所有粒子按照度数的降序排列组成一张列表Ldegree，社区集合{SN}初始化为空；

(2)将列表中的第一个节点取出作为第一个社区的成员；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$\bar{\delta}$，则该节点成为新社区的第一个成员，否则，将其加入到已有社区中，社区关系向量$V_{SN}=V_{SN}+V_i$；

(3)重复进行第(2)步，直至列表末端。

增长是真实网络最基本的特性之一，为了动态地进行社区规划，本文提出，对于新出现的任意节点在其加入网络的同时，根据链接状况对其进行社区划分。由于优先情节的存在，新加入节点的网络初始链接并不能代表其真正的所属社区，因此，当到达基础划分周期时，将重新对当前网络中存在的所有节点进行社区划分。一般而言，搜索引擎的搜索周期为一个月，即每隔一个月，搜索引擎就会采集一次网络上的网页，形成网络的拓扑结构并对网页进行排序分析。因此，本文中设定基础划分周期为一个月。

在上述过程中需要注意的是，Top20%的核心用户的最小相似性阈值是不断变化的，本文设定核心用户的最小相似性阈值为$\bar{\delta}=1/(N_{SNi}+1)$，$N$为$SNi$的规模。而对于Top20%以外的成员，其相似性阈值为$\bar{\delta}=1/n$，n为所研究的社会网络的规模。

从算法的步骤可以看出，晚到的节点不断地与早到的节点进行相似性度量，若彼此之间兴趣相似、爱好相同，则加入早到的节点所在的社区；否则，其独立门派，建立新的社区。真实的社会网络也是如此，20年前，万维网只有一个节点，即蒂姆.伯纳斯编写的第一个网页。虽然现在的互联网上已经有超过10亿左右的网页，但它依然在一个节点一个节点地增长，日复一日，年复一年，始终如此。

以社区SN和粒子i为例，粒子与社区的余弦相似性[12]根据式(1)进行计算：

$$\mathrm{Sim}(SN,i)=\cos(SN,i)=\frac{V_{SN}\cdot V_i}{|V_{SN}|\cdot|V_i|} \tag{1}$$

以图1为例，构建其关系矩阵，如图2所示。

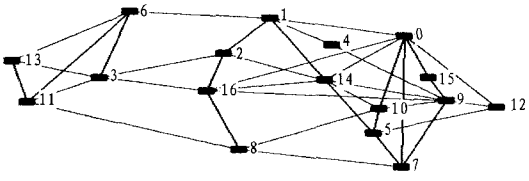

图1 小型社会网络模型

Relation-Matrix=

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	1	1	0	0	0	0	0	1	0	0	1	0	1	0	1	1	1
1	1	1	1	0	1	0	1	0	0	0	0	0	0	0	1	0	0
2	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1
3	0	0	1	1	0	0	1	0	0	0	0	1	0	1	0	0	1
4	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
5	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	1	0
6	0	1	0	1	0	0	1	0	0	0	0	1	0	1	0	0	0
7	1	0	0	0	0	0	1	0	1	1	1	0	0	0	0	0	0
8	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	0	1
9	0	0	0	0	0	1	0	0	1	0	1	1	0	1	0	1	1
10	1	0	0	0	0	0	1	0	0	1	1	1	0	0	0	1	0
11	0	0	0	0	1	0	0	1	0	1	0	0	1	0	1	0	0
12	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0
13	0	0	0	1	0	0	1	0	0	0	0	1	0	1	0	0	0
14	0	1	1	0	0	1	0	0	0	1	1	0	0	0	1	0	1
15	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
16	1	0	1	1	0	0	0	0	1	1	0	0	0	0	1	0	1

图2 小型社会网络邻接关系矩阵

图3展示了基于扩散理论的动态社区挖掘算法的结果。

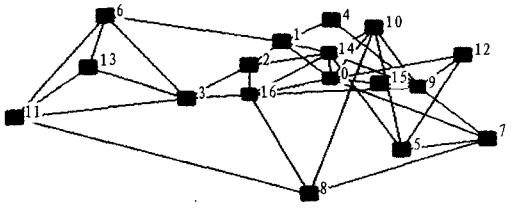

图3 基于扩散理论的动态社区挖掘结果(与G-N算法相同)

3.3 复杂度对比

基于扩散理论[13]的动态社区挖掘算法以节点为单位，将边作为节点的一个属性-度数进行分析和计算，采用快速排序法对节点进行排序，极大地降低了算法的时间复杂度。然而

(下转第287页)

[9] Jing L, Ng M K, Huang J Z. An entropy weighting k-means algorithm for subspace clustering of high-dimensinoal sparese data [J]. IEEE Trans. on Knowledge and Data Engineering, 2007, 19 (8): 1026-1041

[10] 李俊金，向阳，芦英明，等. 粒子群聚类算法综述[J]. 计算机应用研究，2009，126(112)

[11] Liang J J, Qin A K, Suganthan P N, et al. Comprehensive learning particle swarm optimizer for global optimization of multimodal functions[J]. IEEE Transactions onEvolutionary Computation, 2006, 10(3): 281-295

[12] 陈黎飞，郭躬德，姜青山. 自适应的软子空间聚类算法[J]. Journal of Software, 2010, 11(21): 2513-2523

(上接第 278 页)

本文算法为了提高社区挖掘的精确度，在每个基础周期都会重新运行算法，并且进行新的社区划分。这一行为虽然在一定程度上提高了算法的时间复杂度，但是与其他算法相比，仍然具有时间的优势。表 1 中，T 代表算法的总运行时间，t 代表算法的基础周期，则结果如表 1 所列。

表 1 时间复杂度对比分析

算法名称	参考文献	时间复杂度
N-G 算法	[3]	$O(m^2n)$
G-N 算法	[2]	$O(n^2m)$
BB 算法	[7]	$O(n^3)$
本文算法		$O(\frac{T}{t}nlog(m+n))$

4 实验结果分析对比

本文基于现有的静态社会网络，以时间为轴模拟真实社会网络的动态形成过程。根据节点的度数分布，逆向推倒社会网络中节点的出现顺序。采用 Matlab 编程，取得如图 4 和图 5 所示的实验结果。

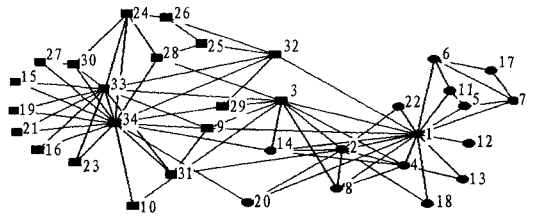

图 4 本文算法对 Zachary Club 的划分结果

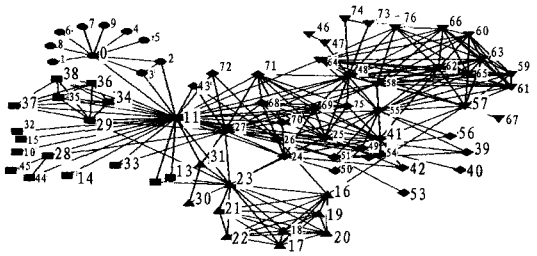

图 5 本文算法对 Les Miserables 的划分结果

根据节点的度数分布，以时间为轴分析 Les Miserables 网络的形成演化过程。

图 6 中，横轴代表时间，纵轴代表节点度数，竖轴表示具有不同度数的节点的数量。

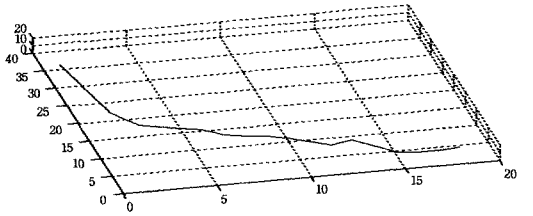

图 6 悲惨世界社会网络的数字化形成过程

从图 6 可以看出，度数最大的节点最先出现，然后陆陆续续有新的节点加入到悲惨世界数据集中，后来的节点都偏向于与度数大的节点建立链接，因此导致最先出现的节点度数的增长最快。同时，上图阐释了社会网络中节点的度数符合幂分布规律[14]。

结束语 通过对社会网络形成集演化机制的研究，本文提出了一种基于扩散理论的动态社区挖掘。利用某一时刻的静态社会网络结构模型进行逆向推理分析，模拟真实社会网络的形成；同时，不断地对新加入的节点进行关系分析和相似性度量，进而将节点划分到其所属社区中。实验指明，该算法不仅具有较高的模块度，而且能够保持较低的时间开销，大大降低了算法的复杂度，具有较好的实用价值。

参 考 文 献

[1] Dorogovtsev S N, Mends J F F. Evolution of Networks: From biological Nets to the Internet and WWW[M]. Oxford University Press, Oxford, 2003

[2] Friedkin N. A structure theory of social influence[M]. Cambridge University Press, Cambridge, 1998

[3] Barabasi A L, Albert. Emergence of scaling in random networks [R]. Science, 1999, 286(5439): 509-512

[4] 郭鸿，周娅. Web 结构挖掘中 HITS 算法的改进[J]. 信息化纵横，2009，16：70-73

[5] Madria S K. Research issues in Web data Mining[J]. Data Warehousing and Wlege discovery, 1999, 99: 303-312

[6] Bianconi G, Barabsi A. Competition and Multiscaling in Evolving Networks[J]. Europhysics Letters, 2001, 5: 436-442

[7] Newman M E J, Girvan M. Finding and Evaluating community structure in networks[R]. E69, 026113, Physics Review, 2004

[8] Kernighan B W, Lin S. A efficient heuristic procedure for partitioning graphs[J]. Bell System technical Journal, 1970, 49(2): 291-307

[9] Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks[C]// Proceedings of the National Academy of Sciences. USA, 2004: 2658-2663

[10] 高琪，张永平. PageRank 算法中主题漂移的研究[J]. 网络与通信，2010，3(3)：117-119

[11] Krapivsky P L, Render S. Connectivity of Growing Random Networks[R]. Physical Review Letters, 2000, 85: 4629-4632

[12] Wolf J, Aggarwal C, Wu K L, et al. Horting Hatches an Egg: A New Graph-Theoretic Approach to Collaborative Filtering [C]// Proceeding of the 5th ACM SIGKDD Conf on Knowledge Discovery and Data Mining. 1999: 201-212

[13] 艾伯特. 巴拉巴西. 链接网络新科学[M]. 徐彬，译. 长海：湖南科学技术出版社，2007

[14] Albert R, Barabasi A L. Statistical mechanics of complex networks[R]. Reviews of Modern Physics, 2002: 74