度中心性节点局部扩展的社区发现算法*

赵卫绩　田　雨　王铁滨　刘井莲
（绥化学院信息工程学院　绥化　152061）

摘　要　针对传统社区发现方法在存在度中心节点的社会网络中表现不佳的问题，提出一种面向度中心性网络的局部社区发现算法。首先，基于最大度k个节点和重叠度阈值形成网络中多个分散的初始社区，然后采用局部社区发现方法，计算每个初始社区的邻居节点的局部模块度增量，每次选取增量最大的节点归入相应社区，以此方法逐步扩展初始社区，直至所有节点划入到社区中。然后，在两个真实网络数据集上与GN和FN算法进行了比较，实验结果表明，该文给出的算法能够揭示出网络中存在的以某个节点为中心的社区结构，相比传统算法，具有更高的准确度。

关键词　社区发现；社会网络；度中心性；局部社区发现；重叠度
中图分类号　TP391　　**DOI**：10. 3969/j. issn. 1672-9722. 2017. 11. 026

Community Detection Algorithm Based on Degree Central Nodes Local Expansion

ZHAO Weiji　TIAN Yu　WANG Tiebin　LIU Jinglian
（Information Engineering College，Suihua University，Suihua　152061）

Abstract　Because traditional community detection methods have poor performance on social networks with centrality，algorithm for discovering community structure in networks with centrality is proposed in this paper. First，based on the top-k maximum degree nodes and overlapping degree threshold，several separate initial community are formed. Then，adopting local community detection method，compute the local modularity incremental quantity of initial communities' neighbor nodes，the node with maximum value is added to the corresponding initial community，repeat these operations until all nodes are assigned to the respective community. We compare our proposed method with GN and FN algorithms on two well-known real-world networks whose community structures are already given. The results of the experiment demonstrate that our algorithm is highly effective at discovering community structure in networks with centrality，and it has high accuracy.

Key Words　community detection，social network，degree centrality，local community detection，overlap degree
Class Number　TP391

1　引言

随着在线社交网络如Facebook、Twitter、新浪微博、微信的兴起，用户在这些网站上的沟通、评价、分享，产生了丰富的内容。在这些网络上，人与人之间、人与内容之间、内容与内容之间形成了海量的关系数据。这些关系数据可以用网络来表示，网络中的节点表示人或内容，边表示节点之间的关系。在这些网络中隐含着很多密集区，区内节点间联系紧密，与区外节点联系稀疏，这样的密集区即为社区[1]。社区发现可用来揭示这类网络中有共同特征的人或内容，为开展个性化服务提供依据，因此社区发现的研究具有理论价值和实际应用意义。

Girvan和Newman首先提出网络中存在着社区结构，并给出了著名的社区发现GN算法[1-2]。该算法首先将整个网络看成一个社区，然后计算网络中每条边的介数并去除介数最大的边，重复以上操作，直至网络中的所有边都被删除。该算法生成的

* **收稿日期**：2017年5月19日，**修回日期**：2017年6月14日
基金项目：绥化学院青年基金重点项目（编号：KQ1301002），2016年黑龙江省大学生创新创业训练计划项目（编号：201610236014）；国家青年科学基金项目（编号：61401185）资助。
作者简介：赵卫绩，男，硕士，讲师，研究方向：数据挖掘、社会网络分析。田雨，男，研究方向：算法设计与分析。王铁滨，女，副教授，研究方向：数据仓库、数据分析。刘井莲，女，博士，讲师，研究方向：社会网络分析与社会媒体挖掘。

结果是一棵层次聚类树,得到了网络的很多划分。为了找到网络的最佳划分,Newman 和 Girvan 给出了模块度的概念[2],模块度越大意味着相应社区划分越好,为从多个社区划分结果中选择最佳的划分提供了依据。之后 Newman 基于模块度的概念提出了一种被称为 FN 的快速社区发现算法[3]。该算法首先将各个节点看成一个社区,然后不断合并能够使模块度增量最大的两个社区成为一个新社区,直至所有的节点成为一个社区。文献[4]将网络看作一个动力学系统,每个节点与周围节点进行交互,通过模拟网络中节点间的距离变化动态地发现社团结构。除了以上方法,还有一类基于度中心节点的聚类方法,文献[5]提出 Top Leaders 算法,该算法将社区定义为领袖节点与其跟随者节点组成的集合。文献[6]提出从度中心节点出发进行社区发现可以保证社区发现的准确度,基于该思想,文献[7]提出了一种基于度中心节点的社区发现算法,该算法将相异度大的中心节点的集合组成核心节点集,然后采用相似性度量方法将网络中其他的节点划分到最相似的核心节点所在的社区。同 Top Leaders 算法一样,该算法在处理网络中的其他节点的划分时,也是采用基于全局的相似度计算方法,需要计算节点到每一个初始社区的相似度,时间复杂度高。文献[8]提出一种面向度中心性及重叠网络社区的发现算法,考虑到度中心性节点的影响力以及重叠节点与不同社区的连接紧密程度,实现了同时具有度中心性节点和重叠节点网络的准确划分。

文献[9]提出了局部社区发现方法,从网络中的一个节点出发可以找到其所在的社区。定义了局部模块度 R,用来衡量一个已有社区加入一个邻居节点后该社区的模块度增减情况。文献[10]用一个节点的 d 层邻居节点的集合来表示该节点,通过节点的相似性给出了称为 GMAC 的局部社区发现算法。文献[10]基于随机游走方法给每个节点赋一个权重,通过定义了一个有偏向的密度指标来寻找局部社区。该类方法基于局部计算,时间复杂度低,而且当起始节点处于社区的中心位置,该算法的准确性很高[6]。

结合度中心性节点算法和局部社区发现算法的优点,本文提出一种基于度中心节点局部扩展的社区发现算法,首先,基于最大度 k 个节点和重叠度阈值形成网络中多个分散的初始社区,然后,采用局部社区发现方法,计算每个初始社区的邻居节点的局部模块度增量,每次选取增量最大的节点归入相应社区,更新该初始社区的邻居节点及其局部模块度增量,以此方法逐步扩展初始社区,直至所有节点划入到社区中。本文算法采用局部社区发现方法,从初始社区局部扩展,获取全局上的社区模块度增加最大的节点,以局部的计算量取得了全局计算的效果,保证社区划分的准确性。

2 度中心性节点局部扩展的社区发现

2.1 问题定义及相关概念

在描述本文算法之前,先给出问题定义及相关概念。

1) 问题定义

通常用图 $G=(V,E)$ 来表示一个无向网络,其中 V 是图中节点的集合,E 是图中边的集合,$n=|V|$ 是图中节点的数目,$d(v_i)$ 表示节点 v_i 的度。社区发现是将图 G 分为 $m(m \geq 1)$ 个连通内部连接紧密的子图,而子图间连接稀疏,可表示为:$C=\{V_1, V_2, \cdots, V_m\}$

$V_i(1 \leq i \leq m)$ 是图 G 的一个连通子图的节点集合,$V_1 \cup V_2 \cup \cdots \cup V_m = V$。若任意两个社区的节点集合的交集为空,则称 C 为图 G 的一个非重叠社区划分,否则 C 为图 G 的一个重叠社区划分。本文关注的是非重叠社区划分。

2) 重叠度

本文采用重叠度来度量一个候选初始社区包含于另一个候选初始社区的程度。如果两个候选初始社区的重叠度较大,则节点个数较小的候选初始社区不能成为初始社区。

设 A、B 是两个候选初始社区,则这两个候选初始社区的重叠度 $over(A,B)$[12] 定义为

$$over(A,B)=\frac{|A \cap B|}{\min(|A|,|B|)} \tag{1}$$

3) 局部模块度 R[8]

社区 D 中的节点可以分为两部分,那些与 D 外节点有连接的节点称为边界节点,与 D 外节点没有连接的节点称为内部节点。边界节点的集合,用 B 来表示。B_{in} 为 B 中节点与 D 内节点连接的边数,B_{out} 为 B 中节点与 D 外节点连接的边数。局部模块度 R 定义为

$$R=\frac{B_{in}}{B_{in}+B_{out}} \tag{2}$$

2.2 算法描述

针对存在度中心节点的社会网络,本文充分利用从中心点出发更容易得到正确的社区以及局部方法时间复杂度低的优点,提出一种基于度中心节

点局部扩展的两阶段社区挖掘算法。首先，基于重叠度公式 $over(A,B)$，采用文献[8]中算法 1 发现 k' 个初始社区，然后采用局部社区扩展方法，从这 k' 个初始社区出发，解决不在 k' 个初始社区中的节点归属问题。首先计算 k' 个初始社区的邻居节点集 $Neighbors$，其中初始社区 C_i 的邻居节点集为 $Neighbors_i$。对于 i 从 1~ k'，计算 $Neighbors_i$ 中每一个节点如果加入 C_i 的模块度增量，保存到 $DeltaR_i$ 中。选择 $DeltaR_1, DeltaR_2, \cdots, DeltaR_{k'}$ 中取值最大的模块度增量对应的节点并入相应的初始社区，如果该节点也出现在其他初始社区的邻居节点中，则将之删除，重复该操作，直至所有节点都划分到社区中。具体算法为

输入：网络 $G=(V,E)$，初始社区 $C=\{C_1, C_2, \cdots, C_{k'}\}$；

输出：最终划分结果 $C=\{C_1, C_2, \cdots, C_k\}$

方法：

/*计算初始社区 C_i 的邻居节点集 $Neighbors_i$*/

for every $C_i \in C$

for every $v_j \in C_i$

$Neighbors_i = Neighbors_i \cup nbrs(G, v_j)$；/*函数 $nbrs(G, vi)$ 以集合的形式返回节点 v_i 的所有邻居节点*/

end for

$Neighbors_i = Neighbors_i - C_i$；

end for

/*计算初始社区 C_i 邻居节点集 $Neighbors_i$ 中每个节点的局部模块度增量*/

for every $Neighbors_i \in Neighbors$

for every $v_j \in Neighbors_i$

计算 v_j 相对于社区 C_i 的局部模块度增量，用 dr 表示；//应用公式(2)

$DeltaR_i = DeltaR_i \cup \{dr\}$；

end for

end for

while $|C| < |G|$

计算 $DeltaR$ 中取值最大的元素所在的 $DeltaR_x$，以及在 $DeltaR_x$ 中的序号 y；

$C_x = C_x \cup \{Neighbors_x[y]\}$；

更新 C_x 的邻居节点集 $Neighbors_x$ 及相应的 $DeltaR_x$；

for every $C_i \in C \ \& \ i! = x$

if $Neighbors_x[y] \in Neighbors_i$

从 $Neighbors_i$ 中删除 $Neighbors_x[y]$ 节点；

end if

end for

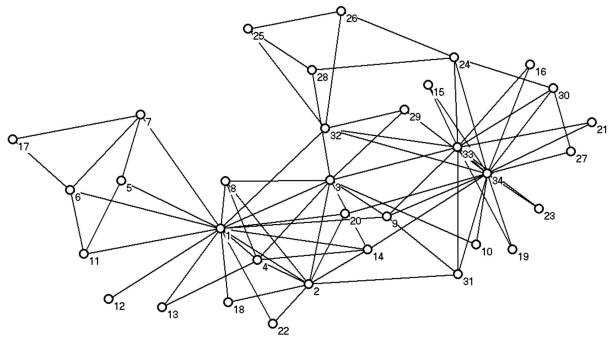
end while

3　实验分析

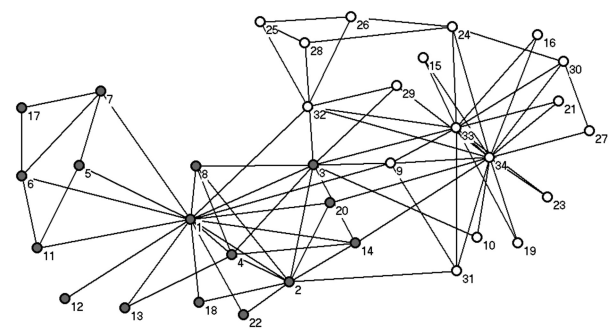
为了测试本文算法的准确性，在 Karate[13] 和 Dolphins[14] 两个真实网络数据集上与 GN 和 FN 算法进行了对比。为了衡量各个算法的效果，我们采用社区划分的两个常用指标归一化互信息 NMI[15] 和 ARI(Adjusted Rand Index)[16] 来衡量各算法得到的社区划分与真实社区划分的相似性，NMI 和 ARI 取值越大，表示与真实社区划分越吻合。

1) 在 Karate 数据集上的实验

Karate 数据集是美国大学空手道俱乐部成员间的社会关系，该网络图中包含 34 个节点，78 条边，其中每个节点表示一个俱乐部成员，节点间连接表示两个成员之间经常在一起参加俱乐部活动，在该俱乐部，因为主管节点 34 和教练节点 1 之间发生分歧而分裂成以两个节点为核心的俱乐部。该俱乐部成员的社会关系网具体如图 1(a) 所示，社区划分如图图 1(b) 所示。



(a)



(b)

图 1　Karate 俱乐部网络及其社区结构

以 k=6，q=0.25，本文算法得到 2 个社区，分别为[1,2,3,4,5,6,7,8,11,12,13,14,17,18,20,22]和[9,10,15,16,19,21,23,24,25,26,27,28,29,30,31,32,33,34]，该结果与真实结果完全一样。而 GN 错划了节点 3，FN 算法错划了节点 10。比较结果如图 2，本文算法的 NMI 和 ARI 值均为 1，而 GN 算法的 NMI、ARI 值分别为 0.8365、0.8823，FN 算法的 NMI 和 ARI 分别为 0.8372、0.8823。本文算法在该数据集上的准确度高于 GN 和 FN 算法。

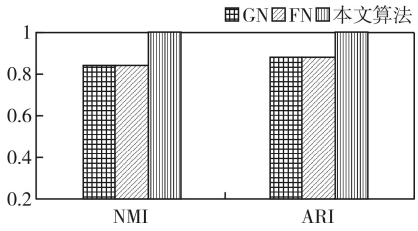
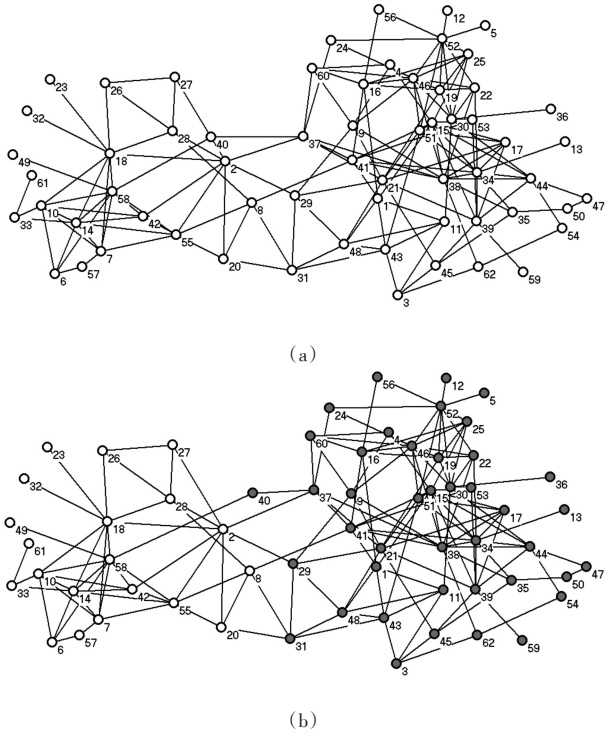


图2 在Karate数据集上的比较结果

2) 在Dolphins数据集上的实验

海豚关系网是社会网络分析中的一个真实网络,常用于测试社区挖掘算法的有效性。这是Lusseau等对栖息在新西兰Doubtful Sound峡湾的62只海豚进行长达7年的观察,并构造海豚关系网如图3(a)所示,图中共有62个节点,159条边。每个节点表示一个海豚,边表示两个海豚接触频繁。图3(b)为该网络的社区结构。



(a)

(b)

图3 海豚网络及其社区结构

以k=6,q=0.18,本文算法得到[0,2,3,4,8,10,11,12,14,15,16,18,20,21,23,24,28,29,30,33,34,35,36,37,38,40,42,43,44,45,46,47,49,50,51,52,53,55,58,59,61]和[1,5,6,7,9,13,17,19,22,25,26,27,31,32,39,41,48,54,56,57,60]两个社区,相比真实社区只错划了节点39。GN算法同样错划了节点39,而FN算法错划了28和30两个节点。比较结果如图4,本文算法与GN算法的NMI和ARI值相同,分别为0.8888、0.9348,FN算法的NMI和ARI值分别为0.8141、0.8721。本文算法与GN算法在该数据集上的准确度都高于FN算法。

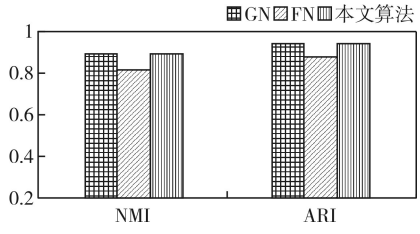


图4 在海豚数据集上的比较结果

4 结语

社区发现能够揭示出网络中隐含着的社区结构,已经成为一种分析社会网络的重要手段。本文针对存在度中心节点的社会网络,提出一种面向度中心性网络的社区发现算法。首先基于最大度节点和重叠度阈值形成网络中多个分散的初始社区。然后采用局部社区发现方法,从所有初始社区的邻居节点中选择局部模块度增量最大的节点归入相应社区,以此方法逐步扩展初始社区,直至所有节点划入到社区中。最后,在真实网络数据集上与GN和FN算法进行了比较,实验结果表明,本文算法具有更高的准确度。

参考文献

[1] Girvan M, Newman MEJ. Community structure in social and biological networks[C]//Proceedings of the National Academy of Sciences of the united States of America. 2002,99(12):7821-7826.

[2] Newman MEJ, Girvan M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2):026113.

[3] Newman MEJ. Fast algorithm for detecting community structure in networks[J]. Physical Review E, 2004, 69(6):066133-1--066133-5.

[4] Shao J, Han Z, Yang Q, Zhou T. Community Detection based on Distance Dynamics[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015:1075-1084.

[5] Khorasgani RR, Chen J, Zaïane O R. Top Leaders Community Detection Approach in Information Networks[C]// 4th SNA-KDD Workshop on Social Network Mining and Analysis. Washington, DC, USA,2010:2319-7323.

[6] Chen Q, Wu T: A method for local community detection by finding maximal-degree nodes[C]//ICMLC 2010:8-13.

[7] 牛冬冬,陈鸿昶,金鑫,等. 基于核心节点的复杂网络社区划分算法[J]. 计算机工程与设计, 2013, 12: 4089-4093.

NIU Dongdong, CHEN Hongchang, JIN Xin, et al. Com-

(下转第2233页)

发现了它的不足。从而针对它的缺点进行了改进，提出了新的系统，完善了基于位置查询的功能。从进行的实验中，知道该系统能够回答至少70%的问题。此外，对于娱乐、美食、夜生活等方面的问题，玩转四方的用户比自由用户回答的更符合询问者的要求。最后，社交软件方面为我们提供了散播问题的平台，这有利于我们快速找到问题的答案。

参考文献

[1] Howe, Jeff. The Rise of Crowdsourcing[J]. 06 Jenkins H Convergence Culture Where Old & New Media Collide, 2006,14(14):1-5.

[2] Howe J. Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business[J]. American Journal of Health-System Pharmacy,2009,67(18):1565-1566.

[3] Kittur A, Chi E H, Suh B. Crowdsourcing user studies with Mechanical Turk[C]// CHI 08: Sigchi Conference on Human Factors in Computing Systems. ACM, 2008: 453-456.

[4] Kamar E, Horvitz E. Collaboration and Shared Plans in the Open World: Studies of Ridesharing [C]// IJCAI 2009, Proceedings of the, International Joint Conference on Artificial Intelligence, Pasadena, California, Usa, July. 2009:187-194.

[5] Chen J, Subramanian L, Brewer E. Sms-based web search for low-end mobile devices[C]// International Conference on Mobile Computing and Networking, MOBICOM 2010, Chicago, Illinois, Usa, September. 2010: 125-136.

[6] Chow C Y, Bao J, Mokbel M F. Towards location-based social networking services[C]// International Workshop on Location Based Social Networks, Lbsn 2010, November 2, 2010, San Jose, Ca, Usa, Proceedings. 2010: 31-38.

[7] Davidov D, Tsur O, Rappoport A. Semi-supervised recognition of sarcastic sentences in twitter and amazon[J]. Conll, 2010:107-116.

[8] Demirbas M, Bayir M A, Akcora C G, et al. Crowd-sourced sensing and collaboration using twitter[C]// World of Wireless Mobile and Multimedia Networks. IEEE, 2010:1-9.

[9] Lange T, Kowalkiewicz M, Springer T, et al. Overcoming challenges in delivering services to social networks in location centric scenarios.[C]// International Workshop on Location Based Social Networks, Lbsn 2009, November 3, 2009,Seattle,Washington,Usa,Proceedings.2009:92-95.

[10] Roussopoulos N, Kelley S, Vincent F. Nearest Neighbor Queries[J]. Acm Sigmod Record, 1995, 24(2):71-79.

[11] Ledlie J, Odero B, Minkov E, et al. Crowd translator: on building localized speech recognizers through micropayments[J]. Acm Sigops Operating Systems Review, 2010, 43(4):84-89.

[12] Von Ahn L, Liu R, Blum M. Peekaboom: a game for locating objects in images[C]// Sigchi Conference on Human Factors in Computing Systems. ACM, 2006:55-64.

(上接第 2210 页)

plex network community detection algorithm based on core nodes[J]. Computer engineering and Design, 2013, 34(12):4089-4093.

[8] 刘井莲，王大玲，冯时，等. 一种面向度中心性及重叠网络社区的发现算法[J].计算机科学,2016,43(3):33-37.
LIU Jinglian, WANG Daling, FENG Shi, et al.Algorithm for discovering network community with centrality and overlap[J].Computer Science,2016,43(3):33-37.

[9] Liu J L, Wang D L, Feng S, Zhang Y F, Zhao F J. Algorithm for discovering network community with centrality and overlap[J], Computer Science,2016,43(3):33-37.

[10] Clauset A: Finding local community structure in networks[J]. Physical Review E, 2005, 72(2):026132.

[11] Ma L, Huang H, He Q, Chiew K, Wu J, Che Y: GMAC: A Seed-Insensitive Approach to Local Community Detection[C]// DaWaK 2013: 297-308.

[12] Wu Y, Jin R, Li J, Zhang X. Robust local community detection: on free rider effect and its elimination[C]// VLDB, 2015:798-809.

[13] PAN L, DAI C, WANG C J, et al. Overlapping community detection via leader-based local expansion in social Nnetworks[C]//Proceedings of IEEE 24th International Conference on Tools with Artificial Intelligence (ICTAI12). Athens, Greece: Ioannis Vlahava, Sotirios G. Ziavras,2012: 397-404.

[14] Zachary WW: An information flow model for conflict and fission in small groups[J]. Journal of Anthropological Research, 1977, 33(4): 452-473.

[15] Lusseau D, Schneider K, Boisseau O J, et al. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations—Can geographic isolation explain this unique trait?[J]. Behavioral Ecology and Sociobiology, 2003,54(4):396-405.

[16] DANON L, DÍAZ-GUILERA A, DUCH J, et al. Comparing community structure identification[J]. Journal of Statistical Mechanics,2005,9(9):P09008-1-P09008-10.

[17] Santos J M, Embrechts M. On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification[C]//Artificial Neural Networks-ICANN. Springer Berlin Heidelberg, 2009:175-184.