文章编号：1674-8085(2019)02-0058-06

基于中心差异度迭代调整机制的网络社区搜寻算法研究

陶 硕

(马鞍山职业技术学院电子信息系，安徽，马鞍山 243031)

摘 要：为解决当前网络社区搜寻算法存在的节点聚类形成困难，搜寻迭代过于复杂，难以实现社区归属的二次更新等不足，提出了一种基于中心差异度迭代调整机制的网络社区搜寻算法。首先，通过领袖节点重叠度来实现初次社区搜寻裁决，有效降低了重复搜寻的概率，且根据加入节点与领袖节点差异度进行聚类匹配；随后，通过待加入节点与领袖节点之间的交互热度方式进行基于热度机制的聚类递归，实现对搜寻误差的二次校正。仿真实验表明，与当前网络社区搜寻算法中常用的差分迭代阈值裁决机制，混沌度一体化成型迭代机制相比，本文算法具有更高的首次成功率，以及更小的搜寻次数与迭代周期，具有很强的实际部署价值。

关键词：网络社区搜寻；节点聚类；领袖节点；热度机制；聚类匹配；首次成功率

中图分类号：TP393 **文献标识码**：A DOI:10.3969/j.issn.1674-8085.2019.02.011

THE RESEARCH AND SIMULATION OF NETWORK COMMUNITY SEARCH ALGORITHM BASED ON CENTRAL DIFFERENCE ITERATION ADJUSTMENT MECHANISM

TAO Shuo

(Department of Electronic Information, Maanshan Technical College, Ma'anshan, Anhui 243031, China)

Abstract: In order to solve the problems of node clustering in current network community search algorithms, such as the complexity of search iterations and the difficulty of secondary update of community ownership, a network community search algorithm based on the iterative adjustment mechanism of center difference degree is proposed. Firstly, the overlapping degree of leader nodes is used to decide the initial community search, and the difference between join nodes and leader nodes is matched to improve the speed of clustering formation and reduce the search iteration process. In order to improve the accuracy of community discovery, the method of constructing the interaction heat between the node to be added and the leader node is used to improve the accuracy of community discovery. Simulation results show that the proposed algorithm has higher first success rate in comparison with the Differential Iterative Threshold Decision Mechanism and Chaos Integration Forming Iterative Mechanism which are commonly used in current network community search algorithms. The advantages of fewer searches, shorter iteration period and clearer aggregation degree make it valuable for practical deployment.

Key words: network community search; node clustering; leader node; heat mechanism; clustering matching; first success rate

收稿日期：2018-12-01；修改日期：2019-02-12

作者简介：陶硕(1973-)，女，安徽枞阳人，讲师，硕士，主要从事数据挖掘、社网络分析、计算机应用技术等方面的研究(E-mail: taoshuo1973@sina.com).

0 引言

人的本质，不过是各种社会关系的总和，随着大数据技术的不断发展，这种社会关系日趋呈现媒体化、网络化、聚类化的发展趋势[1]。基于关系大数据的用户行为访问数据常常以聚合的形式集中为网络社区，这些网络社区的用户一般具有近似教育背景、类同生活习惯，常常形成特征一致的社区节点：区域内节点联系密切，区域外节点联系稀疏；特别是在电子商务中，往往能够起到相当程度的商业资源发掘及流动性提升的作用[2]。

网络社区搜寻算法能够起到迅速发掘相似社区聚类的功能，近几年来随着用户大数据的崛起，研究者取得了相当多的研究成果。Zeng等人[3]基于领袖节点聚类发掘机制，提出了一种崭新的网络社区搜寻算法，该机制通过周期性挖掘的方式，不断归纳当前网络社区中的热点，且将相似热度的热点进行自适应聚类建模，能够迅速地刷新网络热点，并构建基于领袖节点的社区聚类，具有实现机制简便易行的特点。但该算法需要对网络中几乎全部节点进行信息捕捉，难以适应节点数目动态变化的情形。Zhang等人[4]基于分层聚类算法，提出了一种基于分层重叠机制的网络社区搜寻算法，该算法通过随机选取初始热点的方式，根据分层算法进行分叉扩张，直到网络负载达到固定差异度为止，具有管理便捷且收敛速度较快的特点。但是，该算法对网络搜寻精度要求较高，若初始热点选取出现失误，则极易发生分层分叉故障，导致得到的网络社区结构出现严重的混乱。Wang等人[5]鉴于直接分叉容易出现的失误现象，提出了一种基于标签-邻居节点纠错机制的网络社区搜寻算法，该算法主要通过定义正确-失误标签，并默认相邻节点与本节点具有强相关联的特性，实现对错误分叉进行自纠，具有搜寻精确度高的特性。然而该算法实现机理比较复杂，社区节点数量较多时将很难起到应有的收敛程度。

针对当前常见算法存在的不足，设计了一种基于中心差异度迭代调整机制的网络社区搜寻算法。该算法由两个部分构成：通过获取领袖节点的重叠度进行社区归属判断，并通过该重叠度裁决待加入节点与领袖节点间差异度的方式，实现对网络社区的初步定位。通过获取待加入节点与领袖节点间热度进行交互程度判断，并结合裁决差异度，进一步提升社区发现和更新的精度。最后通过仿真实验，验证了本文算法的有效性。

1 网络社区搜寻模型概述

由于网络社区需要处理的是用户及用户行为大数据的集合[6]，实践中将网络、用户析构为图点，图点之间的有向边为用户行为，见图1，表现形式如下：

$$G=<V,E> \qquad (1)$$

其中，G为图点，V为用户，E为用户行为。

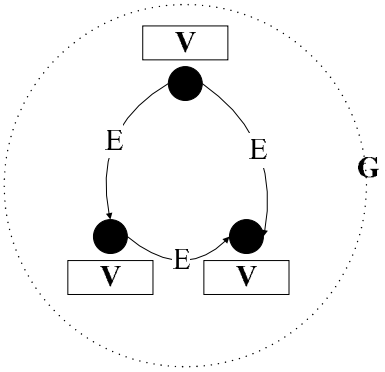

图1 网络社区示意图

Fig.1 Schematic diagram of network community

本文规定 V_i表示图1中第 i 个节点，$d(V_i)$表示第i个节点的权值。在实践中根据节点类型不同，可以作进一步划分：若网络用户属于单一类型，则为低维社区；若网络用户隶属不同的分层分叉结构，则为多维社区[7]。

在网络社区搜寻过程中，$d(V_i)$能够显著体现第i个节点与相邻节点的交互程度：若一个节点的权值越大，说明该节点与其它节点间交互程度也十分密切。此外，不同节点之间的覆盖范围可能有若干重叠之处，见图2。若某个节点的权值较高，则具备被选为领袖节点的价值，能够承担社区的组建任务。

设A和B为两个领袖节点不同的社区，社区的重叠程度定义为 $FG(A,B)$：

$$FG(A,B)=\frac{|A\cap B|-C}{|A\cup B|} \qquad (2)$$

其中C表示A和B的重合节点，若A和B无交集，

则 $FG(A,B)$ 取值为0；若A和B有交集，则 $FG(A,B)$ 取值范围为0-1之间；当A和B重合时，$FG(A,B)$ 取值为1。

一般地，针对多个领袖节点不同的社区而言，其社区重叠程度定义为 $FG(X_1,X_2,\cdots,X_n)$：

$$FG(X_1,X_2,\cdots,X_n)=\frac{\left|X_1\cap\cdots X_n\right|-C_n}{\left|X_1\cup\cdots X_n\right|} \quad (3)$$

其中C表示各领袖节点覆盖范围内重叠节点的总和，获取方式如下：

$$Cn=\oint f(x)(X_1,X\ 2,\cdots,X_n)\mathrm{d}x \quad (4)$$

$F(x)$表示单位阶跃函数，当且仅当某个节点处于覆盖范围时取1：

$$f(x)=1 \quad (5)$$

由模型（2）-（4）可知，由于领袖节点覆盖范围可以处于重叠状态，因此需要考虑不同节点间的连通度问题：即对于新加入网络社区中的节点而言，当且仅当某个节点与某领袖节点之间交互的热度最强时，该节点将加入网络社区，交互热度 $Hot(V_i,C_i)$ 的定义如下：

$$Hot(V_i,C_i)=\sum d(V_i\in C_i) \quad (6)$$

$d(V_i\in C_i)$ 表示对于任意一个 V_i 与社区中任意一个节点 C_i 之间的权值，若对于确定的节点 V_i 而言，对应的交互热度越高，说明该节点与这个社区之间的热度也越强烈。

搜寻社区时，节点可能有多个交互热度，即可能有 k 个社区均与节点 V_i 存在数据交互关系[8]，规定平均热度指标如下：

$$\overline{Hot}(V_i,C_i)=\frac{\sum Hot(V_i,C_i)}{k} \quad (7)$$

对于任意节点 V_i 而言，其交互热度与平均热度之差，即为该节点与确定的社区 C_i 之间的差异度，该差异度越大，说明节点加入该社区的可能性越小，差异度定义如下：

$$Kill(V_i)=Hot(V_i,C_i)-\overline{Hot}(Vi,Ci) \quad (8)$$

2 本文网络社区搜寻算法设计

考虑到当前网络社区搜寻算法存在诸如差异度机制弱反馈效应强[9]、收敛时间短等不足问题[10]，本文构造了中心差异度迭代调整机制来实现网络社区的准确搜寻，主要分为两个过程：①基于重叠度-差异度筛选机制的初始搜寻，该过程主要通过获取领袖节点间重叠度来计算差异度，对待加入节点进行初始社区搜寻。②基于差异度-交互热度的节点精度二次搜寻，该过程主要针对节点与领袖节点之间数据交互的差异度-热度进行综合裁决，以便对错误社区进行二次修正，降低节点搜寻网络社区过程中存在的误差。

2.1 基于重叠度-差异度筛选机制的初始搜寻

对待加入社区的节点 V_i，首先搜寻网络中热度最高的前 k 个领袖节点，获取方式参照模型（7），形成初始社区聚类 Ω：

$$\Omega=<(C_1,C_2,\cdots,C_k)> \quad (9)$$

其中 C_i 表示该聚类中已含的领袖节点。

按模型（3）-（5），获取初始社区聚类 Ω 的重叠度 $FG(\Omega)$：

$$FG(\Omega)=\frac{\left|C_1\cap\cdots C_n\right|-C_n}{\left|C_1\cup\cdots C_n\right|} \quad (10)$$

按模型（8）所示，获取社区聚类 Ω 的差异度 $Kill(\Omega)$：

$$Kill(\Omega)=\frac{1}{k}\sum Hot(V_i,C_i)-\overline{Hot}(V_i,C_i) \quad (11)$$

当节点 V_i 加入的社区热度同时小于模型（10）所获取的聚类 Ω 的重叠度 $FG(\Omega)$，见图2，将该节点纳入这个社区。

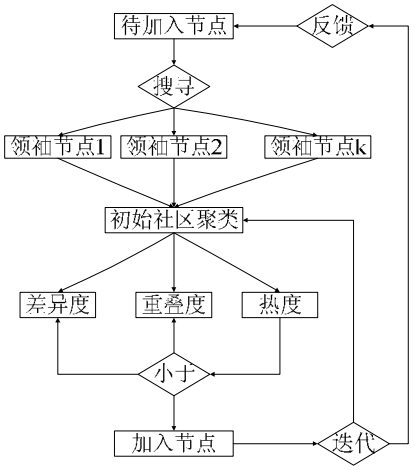

图2 基于重叠度-差异度筛选机制的初始搜寻

Fig.2 Initial search based on overlap degree difference mechanism.

2.2 基于差异度-交互热度的社区归属二次搜寻

通过重叠度-差异度筛选机制的初始搜寻虽然能够实现对网络社区的初步搜寻，但当领袖节点数量较多时，该方法难以实现网络社区的精确化，因此，本文基于差异度-交互热度，构建了节点精度二次搜寻，过程如下：

Step 1：按模型（8）所示，获取任意节点V_i与确定的社区C_i之间的差异度$Kill(V_i)$；

Step 2：按模型(7)与模型(10)所示，获取任意节点V_i与确定的社区C_i间的平均热度$\overline{Hot}(V_i, C_i)$；

Step 3：在节点的搜寻周期内，节点V_i逐个按照模型(3)~ (5)所示，匹配交互热度$\overline{Hot}(V_i, C_i)$。当差异度$Kill(V_i)$与平均热度$\overline{Hot}(V_i, C_i)$之差小于交互热度$\overline{Hot}(V_i, C_i)$时，见图3，本周期内的二次搜寻流程结束。

Step 4：在下一轮周期开始时，节点V_i逐个按照Step1-Step3所示，对所处社区进行二次搜寻，算法结束。

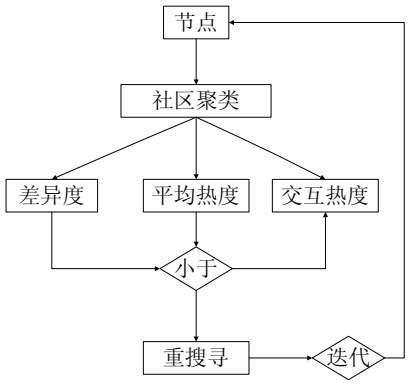

图3 基于差异度-交互热度的节点精度二次搜寻的过程

Fig.3 The search process of node accuracy based on the difference degree of interaction heat.

2.3 复杂度分析

本文算法复杂度主要集中在初始社区聚类形成步骤以及节点更新步骤，两步均为二叉搜寻算法，需要经过m个聚类及n个周期才能完成搜索，因此本文算法的时空复杂度为$o(n)+o(m)$：其中，常用的差分迭代阈值裁决机制[11]（Differential Iterative Threshold Decision Mechanism，DITD机制）需要进行2步迭代，迭代过程采取二叉搜寻算法，因此该机制的时空复杂度为$o(n)+o(m)$，与本文算法持平。混沌度一体化成型迭代机制[12]（Chaos Integration Forming Iterative Mechanism，CIFI机制），虽然采取一体化机制，但成型过程采用冒泡算法，且一旦发生搜寻失误将重新进行搜寻，故该算法的时空复杂度为$mo(n^2)$，其中m为错误发生频率，显然该算法的时空复杂度要高于本文算法。

3 仿真实验

3.1 仿真环境参数设置

为便于对本文算法的优越性进行仿真，采取MATLAB仿真实验环境，针对当前网络社区搜寻算法中常用的差分迭代阈值裁决机制[11]（Differential Iterative Threshold Decision Mechanism，DITD机制），混沌度一体化成型迭代机制[12]（Chaos Integration Forming Iterative Mechanism，CIFI机制）进行仿真对比。从首次成功率、搜寻次数、迭代周期、聚集度四个指标进行仿真对比，详细参数如下：

表 1 仿真参数

Table1 Simulation parameters

参数	数 值
社区覆盖(m^2)	12400*24400
社区搜寻周期(min)	不高于 128s
领袖节点密度	不高于 100
覆盖半径(hop)	不高于 8
节点个数	不低于 1024
聚类类型	简单聚类
聚类个数	不超过 12

3.2 首次成功率

图4显示了本文算法与DITD机制、CIFI机制的首次成功率的测试数据。由图可知，在不同的领袖节点数量下，本文算法的首次成功率始终处于较高的水平，当节点数量为60个时，其成功率达到90.12%左右。而DITD机制、CIFI机制的首次成功率始终要低于本文算法。这是由于本文通过重叠度-差异度筛选机制的初始搜寻流程，能够显著地降低节点归属到领袖节点重叠区域的可能，且可以采取差异度的方式对初始社区进行进一步搜寻。此外，本文算法基于热度思想，对初始搜寻过程中的错误进行二次修正，能够显著减少误搜寻的可能性，因而首次成功率较高。DITD机制虽然对搜寻过程采取基于差分迭代的方式降低搜寻失败的概率，然而该

机制实现过程比较复杂，且差分迭代过程中需要网络环境保持平稳状态，若网络发生抖动，将会导致算法搜寻的成功率受到很大的影响。CIFI机制主要采用一次成型机制，没有设计精度提升流程对搜寻过程进行精度提升，因此首次成功率较低。

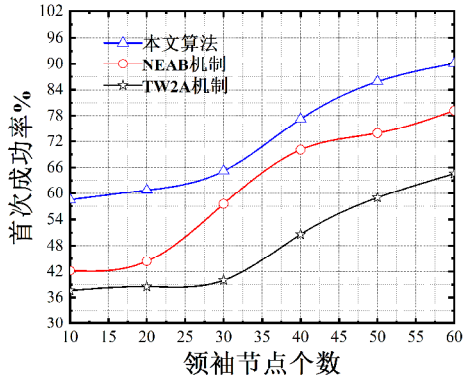

图4 首次成功率仿真

Fig.4 First success rate simulation

3.3 搜寻次数

图5显示了本文算法与DITD机制、CIFI机制在领袖节点不断增加的情况下搜寻次数测试结果。由图可知，本文算法的搜寻次数始终处于较低的水平，DITD机制、CIFI机制需要更多的搜寻次数才能达到性能的稳定。这是由于本文算法的首次成功率较高，且通过差异度方式对初始社区进行进一步搜寻，精度要好于对照组算法，因此需要搜寻的次数较低。DITD机制搜寻过程比较复杂，对环境的适应性较差，需要更多的搜寻次数才能实现对网络社区的精确搜寻。CIFI机制虽然能够进行一次成型，然而由于该算法没有二次精度提升机制，一旦发生搜寻错误将会再次进行搜寻流程，因此搜寻次数要高于本文算法。

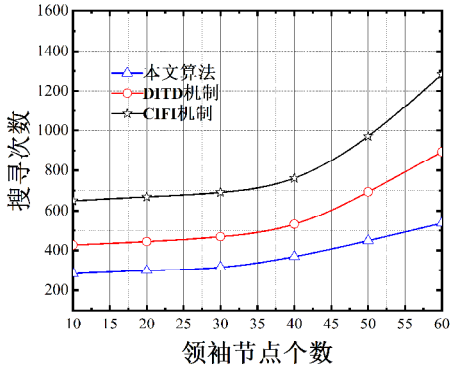

图5 搜寻次数仿真

Fig.5 Simulation of search times

3.4 迭代周期

图6显示了本文算法与DITD机制、CIFI机制在领袖节点不断增加的情况下迭代周期测试结果。由图可知，本文算法的迭代周期要显著低于对照组算法，这是由于本文算法的首次成功率及搜寻次数均要显著好于对照组算法，因此成功搜寻社区过程中的迭代周期较短。DITD机制、CIFI机制由于在首次成功率及搜寻次数性能上要低于本文算法，DITD机制由于搜寻过程非常复杂，迭代过程准确度较差，因此迭代周期较长。CIFI机制虽然能够一次性地实现社区搜寻，然而由于该算法精度较低，迭代过程中需要不断修正社区，因此迭代周期较本文算法逊色。

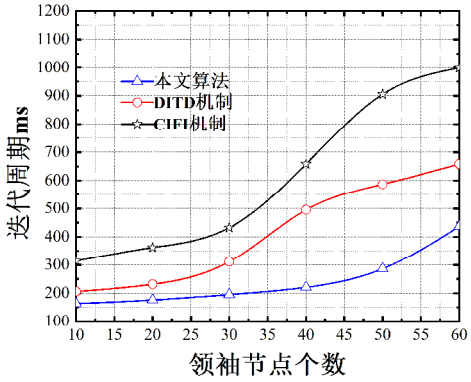

图6 迭代周期仿真

Fig.6 Iteration cycle simulation

3.5 聚集度

图7（a）为初始信源，图7（b）-（d）为本文算法、DITD机制、CIFI机制在聚集度上测试结果。为对领袖节点进行仿真，采取文献[13]使用的广义熵的模糊聚类生成初始信源，错误节点以噪声形式进行掩盖，而正确的领袖节点以初始信源上像素点的形式存在，见图7(a)。迭代时间设定为2h。由图可知，本文算法保持了初始信源的大部分特征，无色彩失真，只存在轻微的噪声，见图7(b)。而DITD机制虽然细节清晰，但其色彩信息出现较大的失真，见图7(c)。CIFI机制的聚集度不理想，其输出图像较为模糊，丢失了部分细节，见图7(d)。这显示了所提方案具有良好的聚集度性能。这是由于本文算法搜寻成功度较高，且通过差异度-交互热度的节点精度二次搜寻流程能够迅速聚集热点并进行精度提升，因此聚集度高。DITD机制采用差分机制，迭代过程复杂，且出现错误的情况下难以聚集热

点，因而聚集度较低，图样出现了扭曲变形现象。CIFI机制仅采取一次成型机制，无法对错误的搜寻点进行过滤，因而聚集度也要低于本文算法，图样显示了严重的模糊现象。

(a)初始信源 (b)本文算法

(c)DITD机制 (d)CIFI机制

图7 不同算法的聚集度测试结果

Fig.7 Test results of aggregation of different algorithms

4 结束语

考虑到当前网络社区搜寻算法存在一些不足之处，提出了一种全新的基于中心差异度迭代调整机制的网络社区搜寻算法，该算法通过重叠度-差异度筛选机制的初始搜寻流程，能够实现较高的首次成功率，且迭代周期短，搜寻次数低。此外，本算法针对当前机制均存在二次更新困难的不足，通过差异度-交互热度的节点精度二次搜寻流程，大大提高了社区归属的准确度。仿真实验也表明本文算法对DITD机制、CIFI机制具有较大的优势。

下一步，针对本文算法对流动性强的环境适应性不足的问题，该问题主要是由于节点高速流动状态时拓扑结构发生剧烈变化，特别是在聚类变动较快的情况下存在成功率下降，将通过引入超混沌流动性一体化控制机制，改善本文算法在高流动性条件下存在的精度不足的问题，进一步提升本文算法的适应性能。

参考文献：

[1] 乔少杰,郭俊,韩楠,等. 大规模复杂网络社区并行发现算法[J]. 计算机学报,2017,40(3):687-700.

[2] 张皓,王明斐,陈艳浩. 基于Kullback-Leibler距离的二分网络社区发现方法[J]. 计算机应用研究,2017,34(5):1480-1483.

[3] Zeng Z, Jiang x. Richard N. Discovering Causal Interactions Using Bayesian Network Scoring and Information Gain[J]. BMC Bioinformatics,2016,17(1):1021-1036.

[4] Zhang S, Jin L, Lin L. Discovering Similar Chinese Characters in Online Handwriting with Deep Convolutional Neural Networks[J]. International Journal on Document Analysis and Recognition (IJDAR), 2016,19(3): 237-252.

[5] Wang T S, Lin H T, Wang P. Weighted-Spectral Clustering Algorithm for Detecting Community Structures in Complex networks[J]. Artificial Intelligence Review, 2017,47(4): 463-483.

[6] Yang J L, Huang T, Song W M. Discover the Network Mechanisms Underlying the Connections Between Aging and Age-Related Diseases.[J]. Scientific reports, 2016, 23(6): 32-56.

[7] Rocco C M, Moronta J, Ranmirez-Manque Z, et al. Effects of Multi-state Links in Network Community Detection[J]. Reliability Engineering and System Safety,2017,163(12): 46-56.

[8] Wei Z, Xiao K Z, Zhao K. Analysis of Associtivity and Community Structure in Mobile Social Networks[J]. Procedia Computer Science,2017,107(6): 630-635.

[9] Chen Y, Wang x, Xiang x, et al. Overlapping Community Detection in Weighted Networks Via a Bayesian Approach[J]. Physica A: Statistical Mechanics and its Applications, 2017,468(12): 790-801.

[10] 赵森严. 基于移动代理的无线传感器网络路由算法研究[J]. 井冈山大学学报:自然科学版,2017,38(5):46-49.

[11] Lin C C, Kang J R, Chen J Y. An Integer Programming Approach and Visual Analysis for Detecting Hierarchical Community Structures in Social Networks[J]. Information Sciences, 2015, 299(3): 25-33.

[12] Zhang Z, Wang Z. Mining Overlapping and Hierarchical Communities in Complex Networks[J]. Physica A: Statistical Mechanics and its Applications,2015,421(41):296-311.