计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(11):163−167 [doi: 10.15888/j.cnki.csa.007658]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 Node2Vec 的重叠社区发现算法①

陈　卓, 姜　鹏, 袁玺明

(青岛科技大学 信息科学技术学院, 青岛 266061)
通讯作者: 姜　鹏, E-mail: jiangpeng_qust@163.com

摘　要: 针对目前基于种子节点选择的社区发现算法在准确性和复杂度等方面存在的不足, 提出了一种基于 Node2Vec 的重叠社区发现算法. 首先, 使用 Node2Vec 算法学习到网络中每个节点的向量表示, 用以计算节点间的相似度, 其次, 利用节点影响力函数计算节点影响力并找出种子节点, 然后基于每个种子节点进行社区的扩展优化, 最终挖掘出高质量的重叠社区结构. 本文选取多个真实网络进行了对比实验, 结果表明, 本文所提出的算法能够在保证良好稳定性的前提下发现高质量的社区结构.

关键词: Node2Vec; 重叠社区发现; 节点影响力; 种子节点; 社区扩展

引用格式: 陈卓,姜鹏,袁玺明.基于 Node2Vec 的重叠社区发现算法.计算机系统应用,2020,29(11):163–167. http://www.c-s-a.org.cn/1003-3254/7658.html

Overlapping Community Discovery Algorithm Based on Node2Vec

CHEN Zhuo, JIANG Peng, YUAN Xi-Ming

(College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: In view of the shortcomings in accuracy and complexity of community discovery algorithm based on seed node selection, a Node2Vec overlapping community discovery algorithm is proposed. First, the vector representation of each node in the network is learned by using Node2Vec algorithm to calculate the similarity between nodes. Second, the node influence function is used to calculate the node influence and find out the seed node. Then the community extension optimization is carried out based on each seed node. Finally the high quality overlapping community structure is excavated. In this study, several real networks are selected for comparative experiments, and the results show that the proposed algorithm can find high quality community structures under the premise of ensuring sound stability.

Key words: Node2Vec; overlapping community discovery; node influence; seed node; community expansion

现实世界中的很多系统都可以被抽象为复杂网络, 如社交网络、技术网络、生物网络, 这些网络都具有一种普遍的特性——社区结构. 在不同类型的网络中, 社区有着不同的含义, 但是所有社区内部节点间的联系总是比不同社区节点间的联系密切, 准确地发现社区结构是在中观层面上理解复杂网络进而研究复杂系统的有效途径.

社区发现的研究历史可以追溯到 1927 年, Rice 等基于投票模式的相似性发现小的政治团体中的社区[1]. 早期的研究工作大部分都围绕非重叠社区发现展开, 此类算法将复杂网络划分成若干个互不相连的社区结构且一个节点只能隶属于一个社区[2]. 然而, 现实中网络社区之间往往是相互重叠的, 硬划分的社区发现算法无法满足需求, 例如, 在社交网络中, 如果每个社区代表拥有共同兴趣爱好的用户所组成的群体, 则一个用户可以拥有诸多兴趣爱好而隶属于多个社区,

① 基金项目: 国家自然科学基金 (F030810); 山东省重点研发计划 (2018GGX101052)
Foundation item: National Natural Science Foundation of China (F030810); Key Research and Development Program of Shandong Province (2018GGX101052)
收稿时间: 2020-03-12; 修改时间: 2020-04-12, 2020-04-29; 采用时间: 2020-05-10; csa 在线出版时间: 2020-10-29

显然, 重叠的社区结构更能体现出复杂网络的特性, 进而帮助我们从中观层面对复杂系统进行分析.

对复杂网络中重叠社区的发现与研究也因此成为近年来新的研究热点, 而社区发现作为社区分析相关工作的前提, 对于其他领域的研究有着重要影响. 目前, 重叠社区的发现结果可以被应用于情感分析、个性化推荐、实体消歧和链接预测等领域的研究.

1 相关工作

近年来, 学者们相继提出大量能够识别重叠社区的算法. Palla 等提出一种基于最大团的派系过滤算法 CPM 来分析重叠的社区结构[3], 该算法易受 k 值影响, 且以最大团为种子的方式计算复杂度较高. COPRA 算法[4] 对基于标签传播的非重叠社区发现算法进行改进, 在标签后面附上节点对该标签的归属系数, 以便衡量该节点包含多个社区的信息比重, 在迭代更新节点标签的过程中允许一个节点同时拥有多个标签, 以发现网络中的重叠社区, 该算法每次迭代的时间复杂度接近线性但稳定度较差. 基于链路的重叠社区发现算法首先对网络的边进行聚类, 然后通过收集链路社区内的所有连接的节点进行社区划分, 代表算法为 LINK 算法[5]. 在此基础上, Li 等[6] 提出一种基因表示模型, 通过将链路社区映射成节点社区的方式, 实现对重叠节点的发现. 基于局部社区优化和扩展的方法则从局部社区出发, 基于优化函数进行扩张, 社区间的交叉部分则为重叠节点, 代表算法为 LFM 算法[7], 除此之外, Su 等[8] 提出利用随机游走策略扩展优化的方法. 文献 [9] 在此基础上提出基于种子节点选择的重叠社区发现算法, 首先通过定义的影响力函数选取种子节点, 然后通过吸引力函数以种子节点为核心进行扩展, 发现种子所在的局部社区结构. 其中, 基于种子节点选择和扩展的算法由于稳定性好、效率高而成为主流的社区发现算法. Wang 等[10] 提出一种基于结构中心性的种子选择算法, 实现了一个高覆盖率的朴实算法, 提高了社区发现质量, 但算法不能很好地适用于大规模数据集. 於志勇等提出的 i-SEOCD 算法能够高效地从种子节点出发进行局部扩展, 最终发现稳定的重叠社区[11], 但是该算法在计算节点相似度时只考虑了局部网络, 提高算法执行效率的同时也牺牲了算法准确性.

现有基于种子节点扩展的重叠社区发现算法虽然在稳定性方面表现较好, 但在衡量两节点间关系时, 往往将两节点间是否有连边作为唯一判别标准, 而只考虑狭小作用域范围内的局部信息的做法, 虽然提升了社区发现的效率, 但忽略了网络中更大范围内节点和边因素对社区发现过程的影响, 使得算法在提升效率的同时往往以牺牲部分准确性为代价. 同时, 现有算法在基于种子节点进行社区扩展的过程中, 往往需要不断地迭代计算现有社区与未划分节点间的相似性关系, 计算量大, 不适合进行大规模网络的社区发现.

为了更好地解决以上问题, 本文利用 Node2Vec[12] 算法对网络结构进行学习, 通过控制在游走产生节点序列过程中对深度优先和广度优先的趋向, 将更大范围内的拓扑结构信息体现到节点因素中, 提出了基于 Node2Vec 的重叠社区发现算法, 该算法能够解决现有算法存在的以牺牲准确性来提高效率和不适合大规模数据集的问题.

2 基于 Node2Vec 的重叠社区发现算法

针对以 Jaccard 相似度为指标衡量节点间距离的方法所存在的局限性, 本文采用网络表示学习算法学习到网络中每个节点的向量表示, 针对传统种子节点选择方法稳定性和鲁棒性差的缺点, 提出了新的种子节点选择算法, 并以此为基础进行社区扩张和优化.

2.1 Node2Vec 算法

Perozzi 等[13] 提出了将 Word2Vec 的思想用于图节点表示学习的 Deepwalk 算法, Node2Vec 在此的基础上改变了随机游走的序列生成方式, 通过半监督的方式学习 p, q 两个超参数的值, 控制游走对深度和广度的趋向, 其中 p 控制跳向前节点邻居的概率, q 控制跳向前节点非邻居的概率, 如图 1 所示.

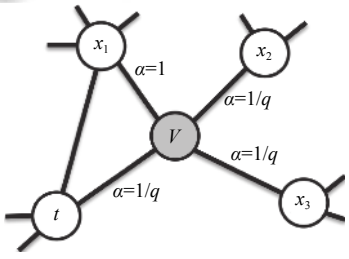

图 1 随机游走过程图

图 1 中, $q > 1$时, 趋向于遍历临近 t 节点的x_1节点, 即趋向于 BFS; $p > 1$时, 趋向于遍历临近 t 节点的x_2或x_3节点, 即趋向于 DFS. 在确定要遍历的邻居节点之后, 采用 skip-gram 模型进行训练进而获得节点的向量表示.

在进行种子节点发现前, 首先利用 Node2Vec 算法对网络结构进行学习, 在学习到网络中每个节点的向量表示后, 对于任意节点u和v, 可利用算法内置的相似

度计算工具计算其在高维空间中的相似度$sim(u,v)$, 其取值范围为 0~1, 通过该方式进一步计算网络中任意节点之间的相似度, 并用相似度矩阵$A_{n\times n}$表示整个网络中节点间的相似度信息, 其中A_{uv}表示节点u和v之间的相似度, 进而可以用$1-sim(u,v)$来表示节点间的相异度.

2.2 种子节点选择算法

通常, 一个网络可以用无向图$G=(V,E)$表示, 其中V表示图中n个节点的集合, E表示图中m条边的集合.

在网络中, 节点u的邻居集合$N(u)$定义如下:

$$N(u)=\{u:v\in V,(u,v)\in E\} \tag{1}$$

节点u对节点v的影响力用$F(u,v)$表示如下:

$$F(u,v)=\frac{D(u)D(v)}{(1-sim(u,v))^2} \tag{2}$$

其中, $D(u)$和$D(v)$分别表示节点u和v的度, $sim(u,v)$表示u, v节点的相似度, 可通过 Node2Vec 生成的节点向量计算得到, $1-sim(u,v)$表示为两节点间的距离, 距离越远, u对v的影响力越小.

节点u的影响力值通过以下公式计算得到:

$$F(u)=\sum_{v\in N(u)}\frac{D(u)D(v)}{(1-sim(u,v))^2} \tag{3}$$

节点影响力的大小与其邻居节点的数量、度数以及相异度有关, 影响力越大, 节点越有机会成为种子节点.

在种子节点选择算法中, 首先根据节点的向量计算所有节点的影响力值, 如果某节点的影响力值比其所有邻居节点的影响力值都大, 则将该节点加入到种子节点的集合中. 算法 1 中列出了种子节点选择算法的伪代码, 其中 2–4 行利用定义的节点影响力计算公式计算出每个几点的影响力值, 5–9 行将每个节点的影响力值与其所有邻居的影响力值进行比较, 若邻居节点中没有比当前节点影响力值大的节点, 则将该节点加入到种子节点集合中.

算法 1. 种子节点选择算法

输入: 无向图$G=(V,E)$; 相似度矩阵$A_{n\times n}$.
输出: 种子节点集合 S.

```
1. S←∅
2. for u∈V do
3.    利用式 (3) 计算 F(u)
4. end for
5. for u∈V do
6.    if ∀v∈N(u) and F(u)≥F(v)
7.       S←S∪u
8.    end if
9. end for
10. return S
```

2.3 社区扩展算法

针对现有算法在社区扩张过程中重复计算量大的问题, 本文在得到分布均匀、影响力大的种子节点之后, 充分利用前一阶段计算所得的节点相似度矩阵, 从每个种子节点出发进行社区扩展, 首先, 以集合中的每个种子节点为核心构建社区, 若节点与种子节点的相似度大于阈值ε, 则将该节点划入该种子节点所属的社区, 然后, 对于尚未被划分的节点, 比较其与各个种子节点的相似度, 选取与之最相似的种子节点, 加入其所在社区, 最终完成社区的划分.

算法 2 中列出了社区扩展算法的伪代码, 2–4 行首先将所有节点标记为 false, 5–13 行分别以每个种子节点为核心进行社区扩展, 并将被划分的节点标记为 true, 此过程中以各节点为核心的社区独立进行扩展, 能够很好地根据阈值ε的大小控制重叠节点的规模, 阈值ε越小, 发现重叠节点的几率越大, 14–19 行将一轮划分结束后没有归属的节点分离出来, 20–25 行则对标记为 false 的节点进行处理, 选择与之相似度最高的种子节点所在的社区作为其社区归属, 最终经过两个阶段的处理, 得到最终的社区.

算法 2. 社区扩展算法

输入: 无向图$G=(V,E)$; 相似度矩阵$A_{n\times n}$; 种子集合 S.
输出: 社区结构 C.

```
1. C←∅
2. for u∈V do
3.    Label(u)=false
4. end for
5. for seed in S do
6.    CS←∅,CS←CS∪{seed}
7.    for u∈V and u∉S
8.       if A[seed][u]≥ε
9.          CS←CS∪{U},Label(u)=true
10.      end if
11.   end for
12.   C←C∪CS
13. end for
14. R←∅
15. for node in V
16.   if Label(node)=false
17.      R←R∪{node}
18.   end if
19. end for
20. for v in R
21.   for seed in S
22.      CS_num = argmaxA[seed][v]
23.   end for
24.   CS←CS∪{v}, Label(v)=true
2.5 end for
26. return C
```

通常情况下, 选择合适的阈值ε能够使得大部分节点经过第一阶段的处理能够划入相应的社区, 阈值ε越小, 需要进行第二阶段处理的节点越少, 但也会导致社区之间重叠度很高. 本文在基于种子节点进行社区扩展的过程中, 充分利用前阶段的计算结果, 将迭代更新的过程简化成了寻址过程, 完美状态下, 只需进行$k\times n$次计算即可完成社区检测, 最坏情况下, 则需进行$2\times k\times n$次计算, 其中k表示种子节点个数, n表示网络中节点的个数, 且二者间满足$k << n$, 总体复杂度为$O(n)$, 优于现有的时间复杂度为$O(n\log n)$的社区扩展算法.

综上, 基于 Node2Vec 的重叠社区发现算法整体流程大致分为以下 3 个步骤:

首先, 利用 NodeVec 算法对网络结构进行学习, 得到包含丰富拓扑结构信息的节点的向量表示, 基于节点向量值计算每对节点间的相似度, 用一个$n\times n$阶矩阵来表示网络结构中所有 n 个节点间的相似度值.

然后, 利用前一阶段计算得到的节点相似度, 根据定义的节点影响力公式筛选出能够独立领导社区的种子节点集合.

最后, 以种子节点为核心, 分阶段进行社区扩展, 首先通过比较每个种子节点与所有非种子节点间相似度与给定阈值的大小关系初步扩展社区, 然后对于未被划分的节点, 选择与之相似度最大的种子节点, 划入其所属社区, 直至所有节点都有至少一个社区归属, 重叠社区检测完毕.

3 实验

为验证算法的相关性能, 在多个不同规模的真实数据集上与其他经典重叠社区发现算法进行对比实验, 待比较的算法分别是 CPM、LINK、COPRA 和 LFM 算法.

3.1 实验数据集

分别选取不同类型不同数量级的 5 个真实网络数据集, 具体包括美国空手道俱乐部网络 Karate[14]、海豚关系网 Dolphins[15]、大学生足球联赛网络 Football[16]、欧洲研究机构电子邮件网络 Email-EU[17] 和高能物理范畴论文引用关系网 Ca-HepPh[18], 各网络规模如表 1.

3.2 评价指标

由于社区划分没有标准的结果, 对于真实数据集, Newman 提出的模块度函数[19] 被广泛认可, 但该评价标准并不能很好地适用于重叠社区, Shen 等在此基础上提出了能衡量重叠社区划分结果的重叠模块度函数[20], 定义如下:

$$EQ=\frac{1}{2m}\sum_{i}\sum_{u\in c_i,v\in c_i}\frac{1}{Q_uQ_v}\left[A_{uv}-\frac{k_uk_v}{2m}\right] \tag{4}$$

其中, m表示网络中的总边数, A为网络的邻接矩阵, k_u为节点u的度数, Q_u表示节点u所属的社区数量.

表 1 真实数据情况表

数据集	节点	边
Karate	34	78
Dolphins	62	159
Football	115	616
Email-EU	1005	25 571
Ca-HepPh	12 008	118 521

3.3 实验结果

本文提出的算法在基于种子节点进行社区扩展时, 社区划分结果易受阈值ε大小的影响, 故首先在不同数据集上在不同阈值ε的指引下进行社区划分, 通常情况下阈值ε的取值范围为 0.3~0.7, 取步长$h=0.1$进行实验, 社区划分结果随阈值ε大小改变而变化的情况如图 2

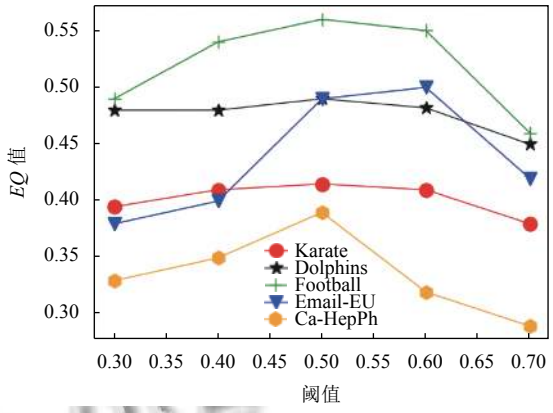

图 2 重叠模块度随阈值变化图

从图 2 可以看出, 在不同数据集上, 模块度总是在$\varepsilon=0.5$左右的位置取到峰值, 这也说明阈值ε对划分结果的影响趋势在所有数据值上是大致相当的.

将本文算法与其他 4 个经典的重叠社区发现算法在不同规模不同类型的数据集上进行对比实验 (阈值取$\varepsilon=0.5$), 结果如表 2 所示.

在大多数数据集上, 本文算法均取得了最高的模块度值, 尤其是在 Emai-EU 和 Ca-HepPh 两个大规模的数据集上, 分别取得了接近 0.5 和 0.4 重叠模块度值, 所发现的社区质量明显优于其他算法, 实验证明, 使用 Node2Vec 算法将更大作用域范围内的网络信息映射到节点向量中的方式, 能够有效地避免范围限制所带来的准确率方面的牺牲, 提升社区发现质量, 在规模大、结构复杂的网络上, 提升效果格外显著.

表 2 真实数据集对比结果表

算法	Karate	Dolphins	Football	Email-EU	Ca-HepPh
CPM	0.187	0.362	0.560	0.375	0.292
Link	0.159	0.003	0.010	0.113	0.132
COPRA	0.342	0.482	0.485	0.403	0.328
LFM	0.317	0.345	**0.572**	0.431	0.363
本文算法	**0.415**	**0.484**	0.563	**0.494**	**0.392**

4 结论与展望

本文提出了一种基于 Node2Vec 的重叠社区发现算法, 首先获得网络结构的向量表示并计算节点之间的相似度值, 利用定义的影响力函数选择出种子节点, 然后以每个种子节点为核心进行社区扩张. 本文选取了不同类型不同规模的真实网络数据集, 并在这些数据集上将本文算法与其他类经典重叠社区发现算法进行对比性实验, 实验结果表明, 本文算法在大部分数据集尤其是大规模数据集上表现出了明显的优势. 后续工作将提高算法的性能, 降低算法复杂度, 并将算法应用到动态社区发现研究中.

参考文献

1 Rice SA. The identification of blocs in small political bodies. American Political Science Review, 1927, 21(3): 619–627. [doi: 10.2307/1945514]

2 骆志刚, 丁凡, 蒋晓舟, 等. 复杂网络社团发现算法研究新进展. 国防科技大学学报, 2011, 33(1): 47–52. [doi: 10.3969/j.issn.1001-2486.2011.01.011]

3 Palla G, Derényi I, Farkas I, *et al*. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 2005, 435(7043): 814–818. [doi: 10.1038/nature03607]

4 Gregory S. Finding overlapping communities in networks by label propagation. New Journal of Physics, 2010, 12(10): 103018. [doi: 10.1088/1367-2630/12/10/103018]

5 Ahn YY, Bagrow JP, Lehmann S. Link communities reveal multiscale complexity in networks. Nature, 2010, 466(7307): 761–764. [doi: 10.1038/nature09182]

6 Li MM, Liu J. A link clustering based memetic algorithm for overlapping community detection. Physica A: Statistical Mechanics and its Applications, 2018, 503: 410–423. [doi: 10.1016/j.physa.2018.02.133]

7 Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics, 2009, 11(3): 033015. [doi: 10.1088/1367-2630/11/3/033015]

8 Su YS, Wang BJ, Zhang XY. A seed-expanding method based on random walks for community detection in networks with ambiguous community structures. Scientific Reports, 2017, 7: 41830. [doi: 10.1038/srep41830]

9 齐金山, 梁循, 王怡. 基于种子节点选择的重叠社区发现算法. 计算机应用研究, 2017, 34(12): 3534–3537, 3568. [doi: 10.3969/j.issn.1001-3695.2017.12.003]

10 Wang XF, Liu GS, Li JH. Overlapping community detection based on structural centrality in complex networks. IEEE Access, 2017, 5: 25258–25269. [doi: 10.1109/ACCESS.2017.2769484]

11 於志勇, 陈基杰, 郭昆, 等. 基于影响力与种子扩展的重叠社区发现. 电子学报, 2019, 47(1): 153–160. [doi: 10.3969/j.issn.0372-2112.2019.01.020]

12 Grover A, Leskovec J. Node2Vec: Scalable feature learning for networks. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA. 2016. 855–864.

13 Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online learning of social representations. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA. 2014. 701–710.

14 Zachary WW. An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 1977, 33(4): 452–473. [doi: 10.1086/jar.33.4.3629752]

15 Lusseau D, Schneider K, Boisseau OJ, *et al*. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. Behavioral Ecology and Sociobiology, 2003, 54(4): 396–405. [doi: 10.1007/s00265-003-0651-y]

16 Girvan M, Newman MEJ Newman. Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12): 7821–7826. [doi: 10.1073/pnas.122653799]

17 Yin H, Benson AR, Leskovec J, *et al*. Local higher-order graph clustering. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, Nova Scotia, Canada. 2017. 555–564.

18 Leskovec J, Kleinberg J, Faloutsos C. Graph evolution: Densification and shrinking diameters. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 2. [doi: 10.1145/1217299.1217301]

19 Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004, 69(2): 026113. [doi: 10.1103/PhysRevE.69.026113]

20 Shen HW, Cheng XQ, Cai K, *et al*. Detect overlapping and hierarchical community structure in networks. Physica A: Statistical Mechanics and its Applications, 2009, 388(8): 1706–1712. [doi: 10.1016/j.physa.2008.12.021]