Journal of System Simulation

Volume 29 | Issue 5 Article 13

6-3-2020

Hierarchical Agglomerative Community Detection Algorithm Based on Similarity Modularity

Wenwei Zhan
College of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China;

Jingke Xi
College of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China;

Zhixiao Wang
College of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China;

Hierarchical Agglomerative Community Detection Algorithm Based on Similarity Modularity

Abstract

Abstract: Fast Unfolding is a hierarchical community detection algorithm based on modularity. It runs very fast, but the accuracy needs to be improved. Because the algorithm adopts traditional modularity to merger communities, it only considers node link information and ignores the neighbor nodes. Therefore, two nodes that have common neighbors and weak link information may not be merged, thus affecting the accuracy. In view of the shortcomings, *a hierarchical agglomerative community detection algorithm based on similarity modularity was proposed through introducing optimized similarity to improve the modularity*. It adopts NMI as the accuracy measurement. Experiments on the real network and LFR synthetic network show that the accuracy of detecting community is obviously improved.

Keywords

fast unfolding algorithm, modularity, similarity, community detection

Recommended Citation

Zhan Wenwei, Xi Jingke, Wang Zhixiao. Hierarchical Agglomerative Community Detection Algorithm Based on Similarity Modularity[J]. Journal of System Simulation, 2017, 29(5): 1028-1032.

第 29 卷第 5 期 系统仿真学报© Vol. 29 No. 5
2017 年 5 月 Journal of System Simulation May, 2017

基于相似性模块度的层次聚合社区发现算法

占文威，席景科，王志晓
（中国矿业大学计算机学院, 江苏 徐州 221116）

摘要：Fast Unfolding 是一种基于模块度优化的层次聚合社区发现算法，其优点是运行速度很快，不足之处是准确度有待提升，这是因为该算法采用传统模块度作为合并社区的衡量标准，而传统模块度函数在计算时只考虑节点间的链接信息，忽略邻居节点的影响，导致会出现两个节点共同邻居较多但由于节点间链接信息较弱不能被合并的情况，从而影响结果的准确度。针对该不足之处，*通过引入优化后的相似度来改进 Fast Unfolding 算法的模块度函数，提出一种基于相似性模块度的层次聚合社区发现算法*，并采用归一化互信息量即 NMI(Normalized Mutual Information)作为评价算法准确性的指标，在真实网络和 LFR(Lancichinetti Fortunato Radicchi)人工合成网络上进行实验，结果表明改进算法检测社区结构的准确度有明显改善。

关键词：Fast Unfolding 算法；模块度；节点相似度；社区发现
中图分类号：TP393 文献标识码：A 文章编号：1004-731X (2017) 05-1028-06
DOI: 10.16182/j.issn1004731x.joss.201705013

Hierarchical Agglomerative Community Detection Algorithm Based on Similarity Modularity

Zhan Wenwei, *Xi Jingke*, *Wang Zhixiao*
(College of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: Fast Unfolding is a hierarchical community detection algorithm based on modularity. It runs very fast, but the accuracy needs to be improved. Because the algorithm adopts traditional modularity to merger communities, it only considers node link information and ignores the neighbor nodes. Therefore, two nodes that have common neighbors and weak link information may not be merged, thus affecting the accuracy. In view of the shortcomings, *a hierarchical agglomerative community detection algorithm based on similarity modularity was proposed through introducing optimized similarity to improve the modularity*. It adopts NMI as the accuracy measurement. Experiments on the real network and LFR synthetic network show that the accuracy of detecting community is obviously improved.

Keywords: fast unfolding algorithm; modularity; similarity; community detection

引言

现实世界中的复杂网络如电信网络、生物网络、人际关系网络等有重要的理论及应用价值，社区结构是解析复杂网络的一个重要特性，因此人们提出了大量挖掘社区结构的社区发现算法。

目前国内外常见的社区发现算法大致可分为 3 类：基于图分割的方法、层次聚类方法、极值优化方法。基于图分割的算法比较经典的有谱二分法[1]和 KL(Kernighan Lin)算法[2]。谱二分法利用网

收稿日期：2016-04-30 修回日期：2016-08-08;
基金项目：国家自然科学基金(61402482), 中国博士后基金(2015T80555), 江苏省博士后基金(1501012A);
作者简介：占文威(1990-)，男，湖北，硕士，研究方向为数据挖掘；席景科(1972-)，男，河南，博士，副教授，研究方向为数据挖掘。

络结构的 Laplace 矩阵中不为零的特征值所对应的特征向量和同一个社区内的节点对应的元素近似值相等的原理对网络社区进行划分；KL 算法先定义一个增益函数 Q，表示社区内部边数与社区之间边数的差值，然后将所有节点划分到两个大小相同的社区，再不断交换两个社区之间的节点以优化 Q 值，直到其中一个社区所有节点都已交换为止，Q 值越大说明划分越优，然后对子社区进行迭代，达到指定的社区数后停止。层次聚类算法可分为聚合算法和分离算法两类。聚合算法将初始每个节点当作单独的社区，然后根据某种标准进行社区合并，最终得到划分的社区。分离算法则刚好相反，初始所有节点当作一个社区，然后根据某个标准，不断的从中删掉边，直到不能删除为止。GN(Girvan Newman)算法[3]是一个经典的层次聚类算法，该算法用边介数来衡量边的强弱，然后不断删掉边介数最大的边，直到无边可删。极值优化算法先定义一个目标函数，然后通过不断优化目标函数来寻找最优社区结构的划分。典型的有 Newman 快速算法[4]，它采用社区质量模块度函数作为目标函数，将使 Q 值最大的划分作为最优结果。

Fast Unfolding 算法[5]是一个典型的层次聚类算法，它采用 Newman 提出的模块度函数[6]为主要优化目标函数。该算法的步骤简单直观，易于实现，并且计算效率很高，可以快速处理节点较多的大规模网络。另外，由于 Fast Unfolding 算法采用层次聚类的方式，在一定程度上避免了基于模块度优化所带来的分辨率问题[7]。该算法被社区发现领域内的知名学者 Fortunato 等[8]推荐为目前性能最好的基于模块度优化的社区发现算法。但由于 Fast Unfolding 算法采用传统模块度作为合并社区的衡量标准，而传统模块度函数在计算时只考虑节点间的链接信息，忽略周围邻居节点的影响，导致会出现两个节点共同邻居较多但由于节点间链接信息较弱不能被合并的情况，从而影响最终结果的准确度。本文引进节点相似度并加以适当改进，然后重新定义模块度函数，最后提出基于相似性模块度的层次聚合社区发现算法，简称 HASM 算法(Hierarchical agglomerative community detection algorithm based on similarity modularity)。

1 基于相似性模块度的层次聚合社区发现算法

1.1 节点相似度

节点相似度用于表示两个节点之间的紧密程度。影响两个节点间相似度的因素通常有节点属性，节点间的共同邻居，节点间的链接关系如权重、最短路径长度、独立路径数等等。实际应用中常采用的方法有：余弦相似度、Jaccard 相似系数、Adamic-Adar[9]指数等。这 3 个方法均从节点间的共同邻居来衡量两个节点的相似度，不同之处在于前两者只考虑共同邻居的数量，而 Adamic-Adar 指数则还考虑了共同邻居的度，共同邻居的度越大，对节点相似度的贡献越小。这实际上很符合实际情况，如微博大 V 通常被很多普通会员关注，而普通会员之间往往没有互相关注，因此本文采用 Adamic-Adar 指数来计算节点相似度，其定义为：

$$s_{ij}=\sum_{c\in N(i)\cap \mathrm{N}(j)}\frac{1}{\log k_c} \tag{1}$$

式中：$N(i)$为节点 i 的邻居集合；c 为节点 i 和 j 的共同邻居；k_c 为节点 c 的度。式(1)存在 2 个问题：(1) 没有考虑边的权重；(2) 当两个直接相连的节点没有共同邻居时其相似度为 0，这显然不符合事实。为了解决以上两个问题，现在对式(1)进行改进，改进后的相似度如式(2)所示：

$$s_{ij}=\frac{\sum_{c\in N(i)\cap N(j)}\frac{1}{\log w_c}}{\sum_{c\in N(i)\cap N(j)}\frac{1}{\log w_c}+1}w_{ij}+w_{ij} \tag{2}$$

式中：$N(i)$为节点 i 的邻居集合；W_c 为与节点 c 相连的边的权重之和；W_{ij} 为节点 i 和 j 相连的边的权重。式(2)的第一项反映的是节点 i 和 j 的共同邻居对 i 和 j 相似度的贡献值，共同邻居越多或共同邻居自身权重之和越小则贡献值越大。另外，当两个

https://dc-china-simulation.researchcommons.org/journal/vol29/iss5/13
DOI: 10.16182/j.issn1004731x.joss.201705013

第 29 卷第 5 期 系统仿真学报 Vol. 29 No. 5
2017 年 5 月 Journal of System Simulation May, 2017

节点之间有边时，那么这两个节点的共同邻居对相似度的贡献应比边的权重对相似度的贡献要小，即直接相连的节点更有可能被划分到一个社区中，因此将共同邻居产生的相似度归约到0-1空间并将其作为权重的比例。将式(2)化简后的相似度如式(3)所示：

$$s_{ij}=(2-\frac{1}{\sum_{c\in N(i)\cap \mathrm{N}(j)}\frac{1}{\log w_c+1}})w_{ij} \tag{3}$$

式(3)即为两个直接相连的节点间的相似度，由于有边相连的节点相似度总是大于没有边相连的节点相似度，因此只需对直接相连的节点计算相似度，这样减少不必要的计算并能保证算法的准确性。

1.2 基于节点相似度的模块度函数

模块度函数是 Newman 提出的一种社区质量评价标准，是目前应用最广的社区质量评价标准。在一个有权的网络中，模块度函数的定义为：

$$Q=\frac{1}{2m}\sum_{i,j}\left[A_{i,j}-\frac{k_i k_j}{2m}\right]\partial(C_i,C_j) \tag{4}$$

式中：$A_{i,j}$表示节点 i 与 j 连边的权值；k_i表示和节点 i 相连的边的权值之和；C_i表示 i 所属的社团。$\partial(C_iC_j)$表示 C_i与 C_j是否为同一个社团，如果 C_i与 C_j为同一个社团此为 1，否则为 0；m 为所有权重之和。在这里 Q 值越大，划分的社区结果越合理。

在 Fast Unfolding 算法合并社区的过程中，优先选择有边相连且权重大的节点，那么有可能会出现一个点与另外两个点的连边的权重相同的情况，这时候如果随机选择其中一个，可能选择的节点与原先节点的共同邻居不如未选择的节点多，即相似度不如未选择的大，从而造成误差。另外，还有一种情况是权重小的一方其共同邻居多从而使得整体相似度大于另一方，而原本的算法选择的是权重大的一方。这两种情况的出现是因为在合并社区并计算模块度时，仅仅考虑节点间的权重而没有考虑共同邻居的影响，从而使得划分结果的正确率有所下降。基于以上分析，现将式(3)的相似度计算公式引入式(4)的模块度计算公式中，得到基于相似度的模块度函数，其定义如下：

$$Q=\frac{1}{2s}\sum_{i,j}\left[S_{ij}-\frac{S_i S_j}{2s}\right]\partial(C_iC_j) \tag{5}$$

式中：s 表示所有节点相似度之和；S_{ij}表示节点 i 和 j 的相似度。

1.3 HASM 算法

HASM 算法在 Fast Unfolding 算法的基础上进行优化改进，首先引入节点相似度并进行优化改进，然后将改进后的相似度与模块度函数进行结合，得到基于相似度的模块度函数，整个算法分为 3 个步骤：

(1) 对于一个有 N 个节点的初始网络，利用式(3)计算所有直接相连的点之间的相似度。

(2) 将网络中的每个节点作为单独社区，然后找到其邻居，优先选择相似度大的节点，将该节点分别加入选择节点所在的社区，计算模块度的增量，若模块度值增加且为正，则将该节点加入增值最大的那个社区。

(3) 步骤 2 完成后，整个网络已经得到初步的划分，这时将初步划分好的每个社区当作一个节点，边则代表 2 个社区之间有边相连的点之间的相似度之和。重复执行上述两个步骤直到没有模块度增加，迭代结束。

2 实验分析

为了评估算法的准确度，采用文献[10]提出的社区发现算法计算精度评价指标——归一化互信息量(Normalized Mutual Information，简称 NMI)作为评价算法准确性的指标。NMI 是评价社区划分精度常用的指标，很多文献[11-13]都采用了该指标。NMI 的值介于 0 和 1 之间，值越大说明两个社区划分的结果越相似，值为 1 时则两个社区划分结果完全相同。

2.1 美国政治书籍网络

美国政治书籍网络也称 polBooks 网络，是由

2004 年美国总统大选时出版并在亚马逊网上书店销售的关于美国政治的书籍所构成的网络。在该网络中，节点代表书籍并带有代表政治立场的标签，分别是自由派、中立派和保守派；两个节点间的边表示同时购买这两本书籍的读者。该网络包含 105 个节点和 441 条边，根据书籍的政治立场被分为 3 个部分。Newman 快速算法、Fast Unfolding 算法以及 HASM 算法在 polBooks 网络上的运行结果如表 1 所示。

表 1 polBooks 网络运行结果

Tab. 1 Experimental results on polBooks network

算法	NMI	Max(Q)	Communities
Newman 快速算法	0.752 3	0.451 2	3
Fast Unfolding 算法	0.774 5	0.498 6	3
HASM 算法	0.800 5	0.553 4	3

表 1 中的 Max(Q)是算法得到的最大模块度值，Communities 是最大模块度值所对应的划分社区的个数。

表 1 中 3 个算法均将 polBooks 网络划分成 3 个社区，但 NMI 值和 Q 值均不相同。模块度能够反映社区结构的合理性，NMI 值能够反映 2 个社区划分结果的相似程度，HASM 算法的 Q 值在三者中均为最大说明 HASM 算法得到了更为合理的社区结构，NMI 值最大说明与原本划分结果更为接近，从而证明基于相似度的模块度函数比传统模块度函数更具有优势。HASM 算法在 polBooks 网络上的具体划分情况如图 1 所示。

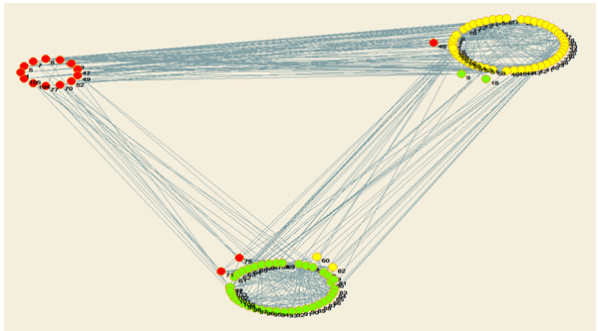

图 1 HASM 算法划分 polBooks 网络的结果

Fig. 1 The result on polBooks network by HASM

图 1 中黄色节点表示保守派，绿色节点表示自由派，红色节点表示中立派。可以看出一共有 7 个节点与标准结果有出入，其中节点 48 原本属于保守派以及节点 71 和 75 原本属于自由派，但 HASM 算法将其划分到中立派中，实际上这种情况是很有可能发生的，即一个持中立政治立场的读者不仅会购买代表中立立场的书籍，而且很可能同时购买代表保守立场或自由主义立场的书籍。其他四个节点的错误划分则属于正常误差。polBooks 网络上的实验结果表明 HASM 算法在该网络上的划分效果优于 Newman 快速算法和 Fast Unfolding 算法。

2.2 LFR 网络

LFR(Lancichinetti Fortunato Radicchi)是一种常用的人工合成网络，可作为社区发现算法的基准网络。LFR 网络的生成程序通常提供 8 个设定参数用于改变社区结构，分别如下：节点总个数(N)，节点平均度数(k)，最大度数(k_{max})，节点幂律分布指数(α)，社区规模幂律指数(β)，社区结构清晰度(γ)，社区规模的最小值(C_{min})，社区规模的最大值(C_{max})。

首先随机生成一个网络，观测各算法的运行效果，参数的设定参考 GN 基准网络[2]并进行适当修改，具体配置如表 2 所示，实验结果如表 3 所示。

表 2 LFR 设定参数

Tab. 2 LFR parameters

N	k	k_{max}	α	β	γ	C_{min}	C_{max}
200	9	20	1	1	0.5	15	20

表 3 LFR 网络运行结果

Tab. 3 The results running on LFR network

算法	NMI	Max(Q)	Communities
Newman 快速算法	0.568 5	0.225 0	12
Fast Unfolding 算法	0.680 5	0.390 3	9
HASM 算法	0.731 5	0.473 2	10

按表 2 设定参数生成的网络一共有 11 个社区，由于 γ 值较大，即社区结构较为模糊，因此 3 个算

法的模块度均偏小，但 HASM 算法的 NMI 值和 Q 值均最大，说明其受社区结构清晰度影响最小，划分效果最好。

为了测试 HASM 算法在不同社区结构清晰度下的划分效果，按照表 2 的设定参数通过改变 γ 值来生成社区结构清晰度不同的多个网络。

3 种算法在不同 γ 值下得到的 NMI 值如图 2 所示。

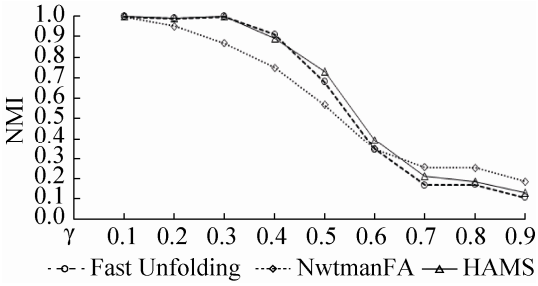

图 2 三种算法的 NMI 值随 γ 的变化趋势
Fig. 2 NMI change trend with γ of three algorithms

从图 2 中可以看出，当 γ≤0.3 时，Fast Unfolding 算法和 HASM 算法的 NMI 值基本为 1，能够完全正确地检测出网络的社区结构，而 Newman 快速算法的表现相对较差，但整体上也能较为准确地识别出网络的社区结构，只是存在数个节点的出入。当 0.3<γ≤0.5 时，这时随着社区结构模糊度增大，3 种算法的 NMI 值均减小，Fast Unfolding 算法和 HASM 算法的 NMI 值均在 0.7 以上，说明此时这两种算法依然能够准确地识别出社区结构，HASM 算法的 NMI 值比 Fast Unfolding 算法略高，而 Newman 快速算法表现相对较差，NMI 值大幅降低。当 0.5<γ≤0.6 时，社区结构模糊度进一步增大，3 种算法的 NMI 值均下降较快，此时检测社区结构的效果开始不理想，HASM 算法的表现相对较好。当 γ>0.7 时，3 种算法的 NMI 值显著下降，检测社区结构的效果较差，此时 Newman 快速算法的表现相对略好，HASM 算法其次。总体来看，当社区结构不是太模糊时，HASM 算法在 3 种算法中表现最好，能够较为准确地检测出网络的社区结构，Newman 快速算法则表现最差；当社区结构非常模糊时，3 种算法的效果均不理想，Newman 快速算法表现相对稍好，但意义不大。综上所述，HASM 算法在社区结构不是非常模糊的情况下，其检测社区结构的效果较好且优于 Fast Unfolding 算法和 Newman 快速算法。

3 结论

本文首先引入 Adamic-Adar 相似度并进行改进优化，然后将改进后的相似度与 Fast Unfolding 算法的模块度函数进行结合，提出基于相似性模块度的 HASM 算法。在实验环节中，采用 NMI 作为评价算法划分社区结果的优劣指标，并将 HASM 算法、Fast Unfolding 算法和 Newman 快速算法应用在 polBooks 真实网络以及 LFR 人工网络上，结果表明 HASM 算法的 NMI 值相对较高，证明基于相似性模块度的 HASM 算法划分社区的准确度较高。

参考文献:

[1] Pothen A, Simon H D, Liou K P. Partitioning sparse matrices with eigenvectors of graphs [J]. Siam Journal on Matrix Analysis and Applications (S0895-4798), 1990, 11(3): 430-452.

[2] Kemighan B W, Lin S. An efficient heuristic procedure for partioning graphs [J]. Bell System Technical Journal (S0005-8580), 1970, 49(2): 291-307.

[3] Girvan M, Newman M E J. Community structure in social and biological networks [J]. Proceedings of the National Academy of Sciences of the United States of America (S0027-8424), 2002, 99(12): 7821-7826.

[4] Newman M E. Fast algorithm for detecting community structure in networks [J]. Physical Review E Statistical Nonlinear and Soft Matter Physics (S1539-3755), 2004, 69(6): 066133.

[5] Blondel V D, Guillaume J l, Lambiotte R, et al. Fast unfolding of communities in large networks [J]. Journal of Statistical Mechanics: Theory and Experiment (S1742-5468), 2008, (10): 10008.

[6] Newman M E J, Girvan M. Finding and evaluating community structure in networks [J]. Physical Review E Statistical Nonlinear and Soft Matter Physics (S1539-3755), 2004, 69(2 Pt 2): 026113.

（下转第 1040 页）