文章编号:1007-1423(2017)12-0006-04 DOI:10.3969/j.issn.1007-1423.2017.12.002

基于标签传播算法在重叠社区发现中的改进

王茜,方旭

(重庆大学计算机学院,重庆 400044)

摘要:

随着信息技术的发展,复杂网络越来越被广泛关注。在真实网络中,重叠社区发现也越来越普遍,与算法相比,标签传播算法简单。在原有 COPRA 算法的基础上,对标签初始化和标签选择进行改进,提出一种新的重叠社区发现算法。实验表明,该算法挖掘的重叠社区具有更高的模块度,能够更加有效地发现重叠社区结构。

关键词:

复杂网络;重叠社区发现;标签传播

1 重叠社区及其发现算法介绍

随着信息技术的发展,大多数网络都存在社区结构这一特征。现有的社区发现主要分为非重叠社区发现和重叠社区发现两大类,其中非重叠社区发现方法发现的社区是彼此不重叠的,一个节点只能属于一个社区,而重叠社区发现允许一个节点同属于多个社区。

重叠社区与非重叠社区比较具有更现实的意义,首先重叠节点是必然是社区中重要的节点,其次重叠社区反映了更加真实的社会网络结构。那么重叠社区的发现及其算法研究已经成为目前一个热点问题。

2 标签传播算法相关

2.1 标签传播思想

标签传播算法在 2002 年最早被提出,先为任一节点初始化一个标签,然后根据邻接点更新自己标签,如果标签直接越相似,那么邻接点的标签就更易被传播。

2.2 LPA 算法

图 $G=(V,E)$表示有 n 个节点$\{1,2,\cdots,n\}\in V$,m 条边$\{1,2,\cdots,m\}\in E$ 的无向无权复杂网络,标签传播算法根据标签传播的思想,重复迭代更新到算法结束,最后根据节点标签分类。标签传播具体流程如下:

(1)初始化,为节点分配标签,x 节点标签为 $C_x(0)=x$;

(2)随机排列网络中的节点,排列为 X;

(3)对 X 中每个节点 x,根据公式 $C_x(t)=f(C(x_1)(t-1),\cdots,C_{x_m}(t-1),C_{x_{m+1}}(t-1))$对 x 节点的标签进行更新。

(4)判断每个节点的标签是否都与其邻接点中出现次数最多的标签一致,如果是,则算法停止,否则设置 $t=t+1$,继续回到步骤(3)执行;

2.3 COPRA 算法

LPA 算法简单直观,易于理解,而且求解准确性很高,无需指定社区个数等其他任何参数,最主要是算法时间复杂度很低,接近线性。但是 LPA 算法存在两个问题:第一,其稳定性较差,原因是社区间标签易传播,当一个节点存在多个可选标签时,随机地选择其中一个,对于不同的随机选择会产生不同的社区发现结果。第二,在现实生活中,很多节点可能同时属于多个标签,而 LPA 算法是无法挖掘出重叠社区结构的,为此引入了多标签结构 $c(c,b)$。b 表示节点 x 对社区 c 的从属系数,其中 $0\leqslant b\leqslant 1$。通过 $b_t(c,x)$表示在第 t 次迭代节点 x 与社区 c 之间的从属系数,具体计算公式为:

$$b_t(c,x)=\frac{\sum_{y\in N(x)} b_{t-1}(c,y)}{|N(x)|}$$

其中，N(x)表示节点 x 的邻接点集合。

COPRA 算法的流程是：

(1)标签初始化阶段是为任意节点 x 分配标签 $c(x,1)$。

(2)根据邻接点集合标签情况更新 x 节点标签，如果有多个，则随机选取 v 个。其中参数 v 为设定的节点可拥有的最多标签数目，其中 $0<|c(x)|\leqslant v$。

(3)若每个节点的标签都和其邻接点中出现次数最多的标签相同或者算法达到设定的最大迭代次数，则算法停止，否则回到步骤(2)继续执行。

(4)根据节点标签对社区进行划分。

对于任意节点的从属系数应满足：$\sum_{c_x\in C(x)} b_t(c_x,x)=1$，$c_x$ 代表节点 x 可能从属的社区。

2.4 社区质量评价指标

为了评价社区划分情况，我们引入模块度的概念。模块度 Q 的意义是 Q 的值越趋向 1，表示产生的社区划分结构模块化程度越明显。目前为止，评价社区发现算法有很多不同的方法。

Q 函数：为了判断社区划分优劣，一个想法就是保证社区内部边尽可能多。Newman 等人在这种思想的基础上，结合了 GN 算法提出了 Q 函数。具体模块度的计算公式是：$Q=\sum_i (e_{ii}-a_i^2)$。其中 e_{ii} 的值则是指社区 i 内边总数和网络中总边数比。

另一种模块度计算公式是：$Q=\sum_{c\in C}\left(\frac{l_c}{m}-\left(\frac{D_c}{2m}\right)^2\right)$，其中 m 表示图中总边数，l_c 表示 c 中所有内部边的条数，D_c 表示社区 c 中所有顶点度数之和。

3 改进的标签传播算法

3.1 标签初始化阶段改进

标签传播算法首先为网络中所有的节点分配一个初始标签，然后对这些标签进行迭代更新，如果网络的节点很多，那么更新所有这些标签的时间空间开销就会增大，考虑到在网络中有很多节点的度数较小，它们的更新完全依赖于它们邻节点的标签，可以采用较少初始化标签的量来节省开销和算法的不稳定性。我们提出节点度数 d_0(d_0 为自然数)这一阈值，对节点度数小于 d_0 的节点，其节点标签不予考虑，这样便大大减少了初始化标签的数目，也减少了此类节点的更新量，降低了更新复杂度。至于 d_0 的取值问题，可以结合网络情况，以及实验结果得出最佳阈值。

3.2 标签选择阶段改进

在 COPRA 算法在选择阶段存在随机选择性，这样就很容易造成标签传播算法的不稳定性。为了解决这些问题，本文引进标签综合影响度的概念，决定标签综合影响度的因素有标签从属系数，节点的度之和，边的权重等。根据影响度的排序依次选择影响度较大的节点，减少了算法随机性。针对标签这些影响因子，综合考虑对标签选择过程的影响程度大小，提出以下标签综合影响度的计算公式。假设节点 x 的领节点中的每一个标签 l，其标签综合影响度 sum(l)计算公式如下：

$$\mathrm{sum}(l)=(1-e^{-\sum d(l_i)})+\frac{\sum b(l_i,x)\times w(l_i)}{C(l)}$$

$w(l_i)$代表标签 l 的所在顶点边的权重；$b(l_i,x)$代表每一个标签 l 的顶点对标签 l 的从属系数；$\sum d(l_i)$代表标签 l 的顶点度数总和。

对任一顶点 x，对 sum(l)进行排序，顶点 x 标签为 sum(l)排序中前 v 个 l 的值，然后对这 v 个标签进行归一化处理，最终得到 x 的标签集合。

改进算法完整描述流程图如图 1 所示：

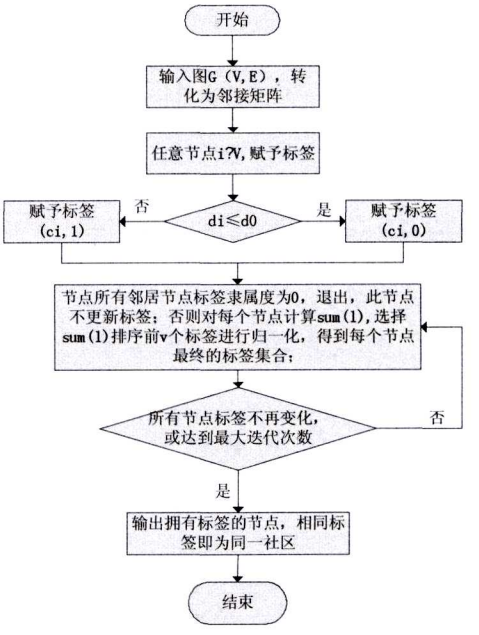

图 1 改进算法流程图

4 实验设计与结果分析

本文实验的硬件为:Intel Core i5-4590 处理器 3.3GHz,16G 内存,算法运行环境为 64 位 Windows 7 和 64 位 Ubuntu 14.0.4 LTS,绘图工具:Visio、Cytoscape。算法实现语言:Java、C++。

数据集一是 Karate Club Network 数据集,Karate 数据集是非官方空手道俱乐部数据集。

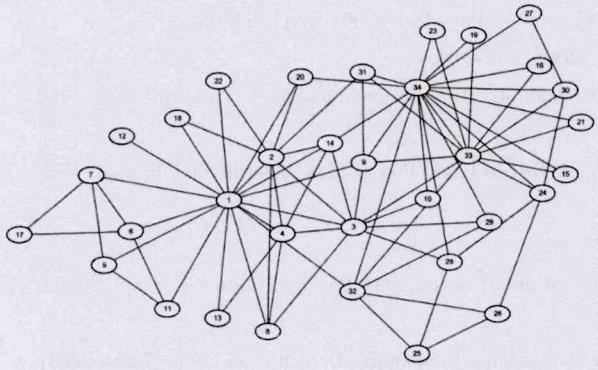

图 2 Karate 网络

表 1 是本文提出的重叠社区标签传播算法在 d=1,v=2 时,由于算法的随机性,重复运行 100 次后取平均值得到的结果,将该算法与 CMP 算法,COPRA 算法进行对比,算法的模块度相比其他算法有所提高,但是由于在标签过滤和标签选择阶段对标签排序,故该算法时间复杂度相对较大。

表 1 Karate数据集实验结果

	CMP	COPRA	本算法	模块度提高率或时间增加率(%)
模块度	0.265	0.459	0.472	2.83%
算法时间	0.098	0.168	0.230	36.9%

数据集二是 Dolphin social network,该数据集是根据海豚群体的交流情况而得到的海豚社会关系网络。这个网络具有 62 个节点,159 条边。节点表示海豚,边表示海豚间的接触关系。

本实验设置 d=1,v=2 时,重复运行 100 次后取平均值得到的结果,表 2 为该算法与原算法的模块度和时间复杂度对比分析,算法的模块度提高,时间复杂度增大。

表 2 Dolphin 数据集实验结果

	COPRA	本算法	模块度提高率或时间增加率(%)
模块度	0.365	0.372	1.92%
算法时间	0.276	0.306	10.9%

数据集三是人工合成数据集,选取合成数据集为 68 个顶点,440 条边,每条边的权重为 1 到 10 之间的正整数的数据集。数据集网络图如图 3 所示:

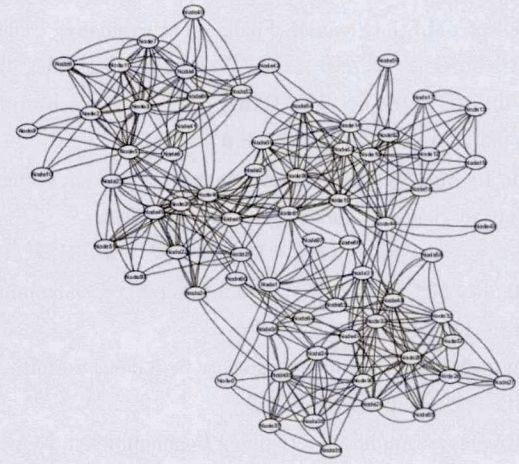

图 3 人工合成网络

针对人工网络设置 d=2,v=2 时,重复运行 100 次后取平均值得到的结果如表 3 所示:

表 3 人工数据集实验

	COPRA	本算法	模块度提高率或时间增加率(%)
模块度	0.472	0.519	9.9%
算法时间	0.320	0.430	34.4%

5 结语

本文从初始化和标签选择两个阶段对 COPRA 算法进行改进。首先针对度数小于设定阈值的节点进行预处理,以此来提高算法执行的效率,最后在多个标签可选情况下,对节点影响度进行排序,从而减小算法随机性,提高了算法的稳定性。本文采用了三个数据集进行试验测试,分别对参数 d 和 v 进行设定,重复运行算法多次,对运行结果进行分析处理。结果表明,改进的 COPRA 算法模块度和稳定性有所提高,但是由于算法在预处理阶段和标签排序方面的改进,使得算法时间

复杂度增加，最终算法能够挖掘出重叠的高质量社区结构。本文与 COPRA 算法一样依然采用同步更新策略，为了使算法更加稳定，可以考虑综合两种传播方式进行改进。

参考文献：

[1] Girvan M, Newman M.E.J. Community Structure in Social and Biological Networks[J]. PNAS,99(12),2002.

[2]Weiss R.s, E Jaeobson.Am.Soeiol[J]. Rev,20,661,1955.

[3]Wu F,Huberman B A.Find Communities in Linear Time: A Physics Approach[J]. Euro.Phys.JB,38:331-338,2003.

[4]Santo Fortunato. Community Detcetion in Graphs[J]. Physics.2010.

[5]Newman M.E.J,Girvan M. Finding and Evaluating Community Structure in Networks[J]. Phys Rev E 69,026112,2004.

[6]汪小帆,李翔,陈关荣. 复杂网络理论及其应用[M]. 北京:清华大学出版社,2006,162-191.

[7]PALLA G, DERENYI I, FARKAS I, et al. Uncovering the Over-Lapping Community Structure of Complex Networks in Nature and Society[J]. Nature,2005,435: 814-818.

[8]Xie Jie-rui,Szymanski B.k Community Detection Using a Neighborhood Strength Driven Label Propagation Algorithm[C]. Proc of IEEE Network Science Workshop,:188-195,2011.

[9]沈海燕,李星毅. 一种新的重叠社区发现算法[J].软件导刊,2015,14(4),59-62.

[10]Barber M J,Clark J W. Detecting Network Communities by Propagating Labels Under Constraint[J]. Physical Review E,2009,80(2): 129-139.

[11]Wu Zhihao,Lin Youfang,Gregory S. Balanced Multi-Label Propagation for Overlapping Community Detection in Social Networks[J]. JCST,2012,27(3): 468-479.

[12]Gregory S. Finding Overlapping Communities in Networks by Label Propagation[J]. New J Physics,2010,12(10):103-118.

作者简介：

王茜(1964-),女,重庆人,教授,研究方向为网络安全、电子商务、数据挖掘

方旭(1991-),男,湖北黄冈人,硕士,研究方向为数据挖掘、社区发现

收稿日期:2017-03-14　　修稿日期:2017-04-05

Improvement of Overlapping Community Detection Based on Label Propagation Algorithm

WANG Qian,FANG Xu

(College of Computer Science, Chongqing University,Chongqing 400044)

Abstract:

With the development of information technology, more and more attention has been paid to complex network. In real network, overlapping community detection is common, comparing with other algorithm, LPA is simple. Proposes a new algorithm to detect overlapping community on the basis of the original COPRA algorithm. Experimental results show that the algorithm has higher modularity, it can discover the overlapping community structure effectively.

Keywords:

Complex Network; Overlapping Community Detection; Label Propagation