基于最大生成树的重叠社区发现算法

郭 娜†, 郑晓艳
(天津职业技术师范大学 信息技术工程学院, 天津 300222)

摘 要:挖掘复杂网络的重叠社区结构对研究复杂系统具有重要的理论和实践意义。针对局部扩展算法(local fitness method,LFM)随机选取种子节点造成的社区结果鲁棒性较低等问题,提出了一种基于最大生成树的重叠社区发现算法:提出一种新颖的边权重定义,将无权的网络转换为带权重的网络,而且该权重真实反映了网络真实结构;提出一种节点影响力计算方法,反映节点在整个网络结构中的重要程度;提出了一种新的生成候选种子集的方法,并借助最大生成树使得到的候选种子节点在网络中更具有代表性;对初始社区划分结果进行优化,避免社区之间重叠度过多。经仿真实验发现,该算法与经典的重叠社区发现算法相比,无论在真实网络还是 LFR 人工网络上,均有良好的表现。

关键词:复杂网络; 社区发现; 重叠社区; 边权; 最大生成树

0 引言

现实中的很多系统都可以用复杂网络来描述。复杂网络中的节点可表示为复杂系统中的个体,节点之间的边是系统中个体之间按照某种规则而自然形成的一种关系,如现实世界中的社交网络[1]、技术网络、生物网络、交通网络[2]等。大量实证研究表明,许多网络是异构的,即复杂网络不是一大批性质相同的节点随机地连接在一起,而是许多类型节点的组合。类型相同的节点之间的连接紧密,而类型不同的节点之间的连接稀疏,把同一类型的节点以及这些节点之间的边所构成的子图称为社区[3]。

在复杂网络搜索和发现社区,有助于人们理解和开发网络,具有重要的社会价值,由此出现了许多社区发现算法。目前大多数算法将一个节点仅归属到一个社区。然而在现实中,事物具有多样性的特点,一种事物往往可归属到不同的类别中,一个节点可属于多个社区,社区间必定存在重叠的现象。所以越来越多的研究者致力于重叠社区结构的研究[4~6]。

目前的社区发现算法主要包括分级聚类算法和图形分割算法两大类。前者是将具有相似性的一类节点归为同一个社区,通过添加或删除边的方法,可以将分级聚类算法分为凝聚算法[7,8]和分裂算法[9,10],如 GN 算法和 Newman 快速算法;后者通常有派系过滤算法(clique percilation method, CPM)[11]和 Kernighan-Lin 算法等。

文献[12]提出的分裂算法已经成为研究网络社区结构的著名算法,并取得了一系列创新型成果[13]。Palla 等人[11]基于凝聚原理,于 2005 年提出了派系过滤算法,开启了重叠社区发现的大门,成为当时社区挖掘的前沿。这种算法通过寻找网络中的极大完全子图来发现社区,但在处理稍大规模的网络时耗时较大,并且对极大子图之外的节点无法判定归属。随后在 2009 年,Lancichinetti 等人[14]提出一种基于局部适应度的重叠社区发现算法 LFM,它也属于凝聚算法,其主要思想是随机选取种子节点,不断扩展能够使社区适应度增益最大的节点,得到种子节点所在的社区集合。搜索过程中,任意节点可以被多次访问,所以会出现一些节点。归属于多个社区的情况,从而发现重叠社区结构。另外它通过设置参数 α 的大小,用于控制发现社区的规模,发现社区的层次结构。但是由于种子节点是随机选取的,可能会出现划分结果不稳定的情况;如果初始种子节点位于稀疏网络的重叠区域,则会出现社区漂移现象;在局部扩展过程中,为了获得最大增益,会有删除节点的操作,可能出现将初始种子节点删除的可能,使得扩展过程不再围绕种子节点,失去社区结构的局部特点。针对 LFM 算法的一些不足,Lee 等人[15]又提出一种以极大团作为初始种子节点进行扩展的社区发现算法 GCE,其在扩展过程中,采用贪心策略逐步添加增益最大的邻居节点,直到适应度函数不再增大为止。但是算法寻找极大子图的代价比较高,而且最终会有一些没有社区归属的节点。

2013 年,Chen 等人[16]提出了一种基于局部中心节点的社区发现算法,根据局部中心节点生成社区。2014 年,刘阳等人[17]提出一种基于边界节点识别的复杂网络局部社区发现算法,通过边界节点识别控制社区范围,给定节点的社区归属。2015 年,Li 等人[18]提出基于谱方法的局部社区发现算法。这些方法都只考虑了网络的拓扑结构,认为网络中的每一条边都具有相同的权重。而在现实世界中,个体之间的联系强弱是有差别的,网络中统一的边权重并不能反映网络的真实结构。而后,一些学者考虑将社区定义为一些相似度高的节点集合。2017 年,Liu 等人[19]提出一种基于节点对相似性的局部社区发现算法。2019 年,文献[20]提出了一种基于结构紧密性的社区发现算法。

针对以上所述的问题,本文提出了一种基于最大生成树的局部扩展算法用于发现重叠社区,主要创新有以下几点:a)提出一种新颖的边权重定义,将无权的网络转换为带权重的网络,权重真实反映了网络真实结构;b)提出一种节点影响力计算方法,反映节点在整个网络结构中的重要程度;c)提出了一种新的生成候选种子集的方法,使得到的候选种子节点在网络中更具有代表性;d)对初始社区结果进行优化,避免了社区之间重叠度过高的情况。实验结果显示,与另外两种经典的重叠社区发现算法相比,无论在真实数据集还是在人工数据集上,本文算法的准确率和鲁棒性都有明显的提高。

1 相关定义

1.1 基本概念

可将复杂网络建模成图 $G(V,E)$,其中 V 是节点的非空有限集,$V=\{v_1,v_2,v_3,\cdots,v_n\}$,$n$ 是网络中节点的个数;E 是边的非空有限集合,$E=\{v_1,v_2,v_3,\cdots,v_m\}$,$m$ 是网络中的边数。所以 $n=|V|$,$m=|E|$。$N(v)$表示在图 G 中节点 V 的相邻节点集合;$k_v=|N(v)|$ 表示节点的度,即与节点 v 相邻的节点集合中的元素总数。

经典三元闭包原则指出:如果 x 和 y 二人拥有共同的好友 a,那么 x 和 y 将有很大可能成为好友,从而形成闭包的三角形 xya。将其推广到复杂网络中:如果两个节点的公共邻居节点越多,那么这两个节点就会越相似,节点间的连接强度越大,如图 1 所示。可以将节点间相似性作为衡量节点连接强度的一个指标。Jaccard 系数[21]的定义为

$$J(V_i,V_j)=\frac{|N(V_i)\cap N(V_j)|}{|N(V_i)\cup N(V_j)|} \tag{1}$$

在一般的无权无向网络中,用邻接矩阵 $\boldsymbol{A}$ 来描述节点之间的连接关系,两个节点 V_i 和 V_j 之间有边相连,对应邻接矩阵 $\boldsymbol{A}[i][j]$ 的值为 1,若无连边,那么值为 0。使用这种传统的邻接矩阵虽然能够很好地表达网络的结构,但是在某种程度上,它并不能真实、具体地反映节点间的连接强度。就好比在科研合作网络中,各个研究者之间合作的论文数量是不一样的,合作的次数越多,研究者的联系更紧密。复杂网络节点间的联系强度很大程度上会影响网络的结构,所以本文提出一种节点间的边权重模型,用来描述节点间的连接强度,反映更真实的网络结构,应用到之后的社区发现过程中。

1.2 边权重模型

复杂网络中,用边权重来表示节点间的连接强度。设 $e=uv\in E$,$A=N(u)\cup N(v)$,$B=N(u)\cap N(v)$,其中 e 为网络 G 的边集 E 中的任意一条边,集合 A 是一条边两个顶点的邻居节点集的并集,集合 B 是两个节点的邻居节点集构成的交集。$|E(A)|(|E(B)|)$的大小在一定程度上反映了其在对应子图的连接强度,因为相对于社区而言,子图内部的边数要高于子图外部。通过式(2),能够生成一个带权图 G^*。

$$W_{uv}=\begin{cases}\dfrac{2|E(A)|}{|A|(|A|-1)}+\dfrac{2|E(B)|}{|B|(|B|-1)} & \exists uv\in E(G)\\ 0 & \neg\exists uv\in E(G)\end{cases} \tag{2}$$

收稿日期: 2019-11-23; **修回日期**: 2020-02-17

作者简介:郭娜(1994-),女(通信作者),河南南阳人,硕士研究生,主要研究方向为数据挖掘、复杂网络社区发现(372491036@qq.com);郑晓艳(1974-),女(蒙古族),内蒙古赤峰人,副教授,硕导,博士,主要研究方向为数据挖掘、复杂网络社区发现、分布式计算.

1.3　最大生成树

一个连通且不存在回路的图称为树,如果图 G 的生成子图 T 是树,则称 T 为 G 的生成树,一个加权的最大生成树是 G 的具有最大权值的生成树。例如,图 2 是我国北方某地区城市高速公路交通图,边上的权值代表该路线的车流量。因为雨雪天气,需要及时清理路上的积雪,所以首先要考虑清理这个城市车流量大的道路的积雪,才能够确保整个交通网络快速地恢复良好的通行。因为考虑到种子节点位于网络中的关键位置,所以本文考虑用最大生成树来简化网络结构,提取出网络结构中影响力较大边的集合,为下一步计算节点的影响力做铺垫。本文使用克鲁斯卡尔算法来生成带权网络 G^* 的最大生成树 T。

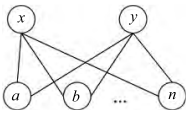

图1　节点x和y的公共邻居

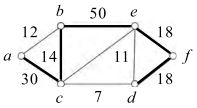

图2　北方某地区城市高速公路交通图
(粗线为使用克鲁斯卡尔算法
得到的车流量大的主要干线)

1.4　节点影响力

假设树 $T(V,E)$ 中,对任意一个节点 v,它的影响力定义为在树 T 中与节点 v 相连的边的权重之和,为

$$I(v)=\sum_{\{u|uv\in E(T)\}}\omega_{uv} \tag{3}$$

节点影响力也是衡量网络集团化的重要参数,节点在网络中的连边越多、连边的边权重越大,会使节点在整个网络中的局部聚集性越高,地位越关键,成为干扰点和离散点的概率越小。

1.5　社区适应度

给定图 $G=(V,E)$,社区 $C\subset G$,社区 C 的适应度函数为

$$F_{|C|}=\frac{C_{\text{kin}}}{(C_{\text{kin}}+C_{\text{kout}})^{\alpha}} \tag{4}$$

其中:C 代表当前的社区;C_{kin} 和 C_{kout} 分别代表当前社区内部节点的度之和与当前社区中每个节点与网络其他部分相连的边数;α 是控制社区规模的参数。$F_{|C|}$ 反映了社区内连接强度和社区与外部网络连接强度的相对关系,该值越大,代表当前社区的社区结构越明显。

1.6　添加节点适应度

给定图 $G=(V,E)$,社区 $C\subset G$,节点 v 是社区 C 的邻居集中的节点。将 $F_{(|C|+v)}$ 的计算代入式(4)。式(5)表示在相同的参数 α 的情况下,当前社区加入节点 v 时社区适应度的改变量。由于社区适应度一定程度上反映了社区内部与社区间的连接强度,所有只有 $F(_C^{v+})$ 为正值的邻居节点才能加入到当前社区中。

$$F(_C^{v+})=F_{(|C|+v)}-F_{(|C|)} \tag{5}$$

1.7　删除节点适应度

给定图 $G=(V,E)$,社区 $C\subset G$,节点 v 是社区 C 中的节点。$F_{(|C|-v)}$ 的计算代入式(4)。式(6)表示在相同的参数 α 的情况下,删除当前社区节点 v 时社区适应度的改变量。由于社区适应度一定程度上反映了社区内部与社区间的连接强度,所以 $F(_C^{v-})$ 为负值的节点需要从当前社区删除。

$$F(_C^{v-})=F_{(|C|)}-F_{(|C|-v)} \tag{6}$$

1.8　社区重叠度

对于初始社区集合 $\{c_1,c_2,\cdots,c_k\}$ 中的任意两个社区 C_a 和 C_b,社区重叠度定义为

$$\theta(C_a,C_b)=\frac{|C_a\cap C_b|}{\min(|C_a|,|C_b|)} \tag{7}$$

其中:$\theta(C_a,C_b)$ 表示社区 C_a 和 C_b 的重叠度,取值为[0,1],值越大,说明重叠度越大。

2　算法思想与分析

基于最大生成树的重叠社区发现算法主要包括生成候选种子集、生成初始社区划分结果、社区优化三个阶段:a)根据提出的边权重模型,将无权网络转换为加权网络,生成加权网络的最大生成树,得到原始网络中节点的影响力排序结果,得到候选种子集;b)按影响力排序结果选择种子,生成围绕该种子的适应度最大的社区,当所有社区构造完成之后,即得到初始社区划分结果;c)针对初始社区划分结果,将重叠度较大的社区之间进行合并。

2.1　候选种子集

针对随机选取种子节点造成算法结果鲁棒性较低的情况,考虑计算网络中各个节点的影响力,让网络中具有代表性且具有高影响力的节点作为种子节点。通常,这类节点在最终生成的社区子图拓扑的中心位置,选取这类节点作为网络中的种子节点,能够使算法以尽可能少的迭代次数获得社区划分。具体步骤如下:

算法 1　候选种子集生成算法
输入:$G=(V,E)$。
输出:candidate = $\{v_1,v_2,\cdots,v_n\}$。
1　initialize, let candidate = ∅
2　for $e_{uv}\in E$
3　　caculate w_{uv}
4　generate W_{uv}
5　generate $G^*=(V,E,W)$ use W_{uv}
6　T = Kruskal(G^*)
7　for v_i in T
8　　caculate $I(v_k)$
9　rank all v_k in non-increasing order
　　//按影响力排序,即 $I(v_1)\geqslant I(v_2)\geqslant\cdots\geqslant I(v_n)$
10　return result candidate = $\{v_1,v_2,\cdots,v_n\}$

2.2　生成初始社区集合

在生成社区阶段,本文采用最大化社区适应度函数的方法对种子节点进行扩展,以期能得到当前种子节点附近的社区适应度最大的集合。在该阶段,通过批量扩展社区,使当前社区适应度最大;此后,考虑逐个删除某个节点,使社区适应度最大。算法具体步骤如下:

算法 2　初始社区集生成算法
输入:$G=(V,E)$, candidate = $\{V_1,V_2,\cdots,V_n\}$,社区规模参数 α。
输出:$C=\{C_1,C_2,\cdots,C_n\}$。
1　while candidate ≠ ∅ do
2　　for v_i from candidatein order
3　　candidate←candidate − $\{v\}$
4　　C_k←$C_k\cup\{v\}$
5　　update N_{set} of C_k, calculate $F_{|C_k|}$ using eq(4).
　　//C_k 为当前社区集合,N_{set} 为当前社区的邻居集
6　　while $N_{\text{set}}\neq\varnothing$ do
7　　　for node $\in N_{\text{set}}$
8　　　　$M=\varnothing$
9　　　　caculate $F(C^{+\text{node}})$
10　　　if $F(C^{+\text{node}})>0$
11　　　　M←$M\cup\{\text{node}\}$
12　　　C_k←$C_k\cup M$
13　　end while
14　　while $C_k\neq\varnothing$ do
15　　　for $v\in C_k$, calculate $F(C^{-V})$
16　　　if $F(C^{-V})<0$
17　　　　C_k←$C_k-\{v\}$
18　　end while
19　　let C_i←C_k, candidate = candidate − C_k
20　　$k=k+1$
21　end while

2.3　社区结果优化

重叠社区发现算法可以允许不同的两个社区包含若干个公共节点,所以得到的初始社区之间存在一定的相似性。但如果两个社区的重叠性过高,会出现过度重叠的现象,需要对重叠度过高的社区进行合并。算法描述如下:

算法 3　社区优化算法
输入:$C=\{C_1,C_2,\cdots,C_n\}$;重叠度阈值为 0.4。
输出:$\{C_1,C_2,\cdots,C_m\}$。
1　for C_a in C
2　　for C_b in $\{C_{a+1},C_{a+2},\cdots,C_n\}$
3　　　caculate $\theta(C_a,C_b)$
4　　　if $\theta\geqslant 0.4$
5　　　　$C_a=\{C_a\cap C_b\}$
6　　　　remove C_b

3　评价

3.1　实验数据

3.1.1　真实数据

本文采用多个不同规模的经典真实数据集测试,如表 1 所示。

3.1.2　LFR 人工数据集

如今 LFR(Lancichinetti Fortunato Radicchi)基准网络[26]被认为是社区检测的标准测试网络,其特征在于节点度和社区规模的非均匀性分布。可以用 LFR 生成一个合成网络,这种基准图拥有真

实网络的特征,即节点度和社区规模都具有异质性。网络结构由如下的可调参数控制:N 为生成网络的节点个数;k 为网络节点的平均度;k_{max} 为最大度;t_1 为度分布的指数;t_2 为社区大小分布的指数;c_{min} 为最小社区顶点个数;c_{max} 为最大社区的顶点个数,mu 为混合参数,其取值在(0,1)。每一个顶点在社区内部的边是其度数的 $\mu-1$ 倍,而与外部社区顶点相连的边是其度的 μ 倍。表 2 是合成的人工数据集。

表 1 真实数据集参数

数据集	点数	边数	社区数	说明
karate[22]	34	78	2	空手道俱乐部网络
dolphins[23]	62	159	2	海豚社会关系网络
lesmis[24]	77	256	4	《悲惨世界》人物关系网络
football[25]	115	613	12	美国大学生橄榄球联赛网络
polbooks	105	441	3	美国政治书网络
netscience[1]	1 589	2 742	–	科学家合作网络

表 2 LFR 人工数据集参数

参数	D_1	D_2	D_3
网络顶点数 N	100	100	100 ~ 1 000
平均度 k	4	4	4
最大度 k_{max}	10	10	10
混合参数 mu	0.1	0.1	0.1
度分布指数 t_1	2	2	2
社区大小分布指数 t_2	1	1	1
最小社区顶点数 c_{min}	5	6	6
最大社区顶点数 c_{max}	30	30	30
重叠节点数 O_n	10	5 ~ 20	10
重叠节点所属社区数 O_m	2 ~ 5	3	2

3.2 评价指标

3.2.1 模块度(EQ)

最早 Chauset 等人[27]提出了模块度函数 Q,成为评价非重叠社区的重要指标,被许多研究人员广泛使用。针对重叠社区的模块度评价,使用 Shen 等人[28]提出的公式,表示为

$$EQ=\frac{1}{2m}\sum_i\sum_{v\in C_i,w\in C_i}\frac{1}{O_vO_w}\left[A_{vw}-\frac{k_vk_w}{2m}\right] \tag{8}$$

其中:v、w 是同一个社区的两个节点;m 是网络 G 的总边数;k 是网络中包含的总的社区数量;O_v 代表节点 v 所属的社区数量;$\boldsymbol{A}_{vw}$ 是图 G 邻接矩阵 $\boldsymbol{A}$ 的元素。当网络中社区内部链接越紧密而社区之间连接越稀疏的时候,EQ 的值就越大,算法划分的效果就越好,EQ 的取值在[0,1]。计算模块度时,并不需要与网络已知的社区结构对照,因为大多数网络的社区结构是未知的,所以与其他评价指标相比,模块度的适用范围更广,所以本文主要使用重叠社区模块度作为评价指标。

3.2.2 标准互信息(NMI)

本文用标准互信息来评估人工网络的社区划分质量,最早是由 Lancichinetti 改进了适用于重叠社区的标准互信息[29],但由于该方法在两种划分结果的社区数量差异较大时,会出现对两个集群相似性评价过高的情况。为了避免这种情况,使标准互信息指标更合理,本文用文献[30]提出的标准互信息评价标准。

$$NMI=\frac{\frac{1}{2}[H(X)+H(Y)-H(X|Y)-H(Y|X)]}{\max(H(X),H(Y))} \tag{9}$$

其中:$H(X)$和$H(Y)$为随机变量 X 和 Y 的信息熵;$H(X|Y)$和$H(Y|X)$为随机变量 X、Y 的条件熵。

3.3 实验条件

本文算法和参照实验的另外两种算法均存在一些可变参数;在 LFM 算法中,参数 α 控制在 0.6 ~ 1,通过多次实验将 α 调整到最合适的值,使衡量指标达到最大值。另外,由于 LFM 算法结果随机性较大,随机选取 10 次运行结果的平均值作为其结果。在本文算法中,设置的 α 参数与 LFM 算法相同,另外重叠度阈值设置为 0.45。在CPM 算法中,全耦合网络节点数 k 设置为 3。

3.4 实验结果

由图 3 可知,本文算法要优于另外两种算法,对于大部分真实网络,本文算法相对于 LFM 和 CPM 算法具有更高的模块度和更高的稳定性,得到的划分结果更接近于真实网络。本文算法的指标明显高于 CPM 算法。另外,在算法质量优于 LFM 的情况下,具有高鲁棒性,避免 LFM 算法稳定性很低的情况,也不会出现种子随机选取陷入死循环的情况。

图 4 中用到的网络是人工生成的 LFR 人工基准网络 D_1,随着逐渐增大 O_m,使重叠节点可以隶属多个社区。从图中可以明显地看出,不管在哪个网络中,本文算法都要明显优于另外两种算法,具有相对较高的模块度。从整体上来看,这三种算法随着 O_m 的增大,模块度的大小在逐渐下降。主要的原因是不同算法在 O_m 和社区数量上表现不同。

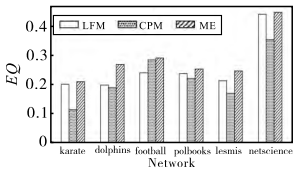

图3 真实网络中各种算法的模块度

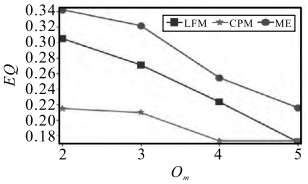

图4 D_1网络上的模块度

图 5 中用到的网络是人工生成的 LFR 人工基准网络 D_2,分别设置不同的 O_n,使网络的重叠节点比例不断增大。从图中可以明显地看出,随着网络中重叠节点的增多,三种算法的模块度指标整体呈现出下降的趋势,但是无论网络中重叠节点的比例有多大,本文算法的模块度指标始终优于另外两种算法。

图 6 中用到的网络是人工生成的 LFR 人工基准网络 D_3,网络中的节点个数 N 设为 100 ~ 1 000。很显然,对于这种中小规模网络,随着网络中节点数目的增多,算法的模块度指标整体出现增长趋势最终呈现出稳定趋势;而且,无论哪种规格的网络,本文算法的模块度始终是最高的。从图中可以看到,在社区规模为 700 时,三种算法的模块度指标均达到峰值。

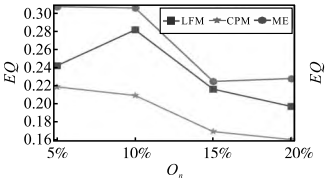

图5 D_2网络上的模块度

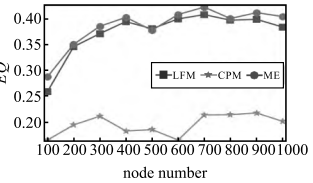

图6 D_3网络上的模块度

图 7 是在真实数据集上的标准互信息,明显可以看出,本文算法划分结果与真实的划分结果更加接近。

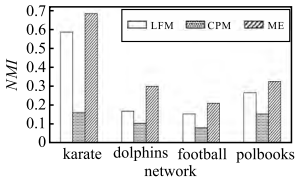

图7 真实数据集上的标准互信息

4 结束语

本文提出了一种基于最大生成树的重叠社区发现算法,首先针对现实网络中边可能带有表示连接紧密程度的权重,将无权网络转换为带权网络;进而结合生成树算法,获得网络中节点的影响力排名,得到候选种子集;然后根据社区适应度最优策略生成初始社区,最终通过社区优化得到社区划分结果。分别在真实网络和人工网络上的实验显示了本文算法较另外两种重叠社区发现算法而言,具有更高的稳定性和准确率。

参考文献:

[1] Newman M E J. The structure of scientific collaboration networks[J]. Proceedings of the National Academy of Sciences of the United States of America,2001,98(2):404-409.

[2] 谭红叶,吴永科,张虎,等. 面向复杂有权网络的社区发现方法研究[J]. 中文信息学报,2018,32(8):111-119.

[3] Schweitzer F, Fagiolo G, Sornette D, *et al*. Economic networks: the new challenges[J]. Science,2009,325(5939):422-425.

[4] Shahmoradi M R, Ebrahimi M, Heshmati Z, *et al*. Multilayer overlapping community detection using multi-objective optimization[J]. Future Generation Computer Systems,2019,101:221-235.

[5] Teng Xiangyi, Liu Jing, Li Mingming. Overlapping community detection in directed and undirected attributed networks using a multiobjective evolutionary algorithm[J/OL]. IEEE Trans on Cybernetics, 2019. http://doi.org/10.110q/tcyb.2019.2931983.

[6] 李东,程鸣权,徐杨,等. 基于平均互信息的最优社区发现方法[J]. 中国科学:信息科学,2019,49(5):613-629.

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

[8] Radicchi F, Castellano C, Cecconi F, *et al*. Defining and identifying communities in networks[J]. Proceedings of the National Academy of Sciences,2004,101(9):2658-2663.

(下转第 180 页)

已经有研究者基于 Spark 内存计算框架开展了生物大数据处理与分析方面的工作,但是研究尚不成熟,其中需要解决的一个关键问题是串、并行处理结果如何保持一致,二者存在差异的原因是什么。本文实验结果表明,在串并行 TopHat 执行序列比对方面二者存在约 0.03% ~2.3% 的差异,其原因尚不明确,是后续值得进一步研究的内容;在检测显著差异表达基因/转录本方面,串并行 Cufflinks 能够取得约 88% 的一致性,这对于评价本文所提出的并行方案的结果的可靠性具有重要意义。

在基于 Spark 或 MapReduce 平台对传统串行软件进行并行化改写或调用的过程中,由于串行软件自身效率不高而使得并行效率难以充分提升,这个瓶颈问题有待今后进一步研究解决。

致谢 感谢西北农林科技大学园艺学院管清美教授提供的拟南芥 RNA-seq 数据。

参考文献:

[1] Decap D, Reumers J, Herzeel C, *et al*. Halvade: scalable sequence analysis with MapReduce[J]. Bioinformatics, 2015, 31(15): 2482-2488.
[2] Decap D, Reumers J, Herzeel C, *et al*. Halvade-RNA: parallel variant calling from transcriptomic data using MapReduce[J]. PLoS ONE, 2017, 12(3): e0174575.
[3] Pandey R V, Schlotterer C. DistMap: a toolkit for distributed short read mapping on a Hadoop cluster[J]. PLoS ONE, 2013, 8(8): e72614.
[4] 杨晓亮. MapReduce 并行计算应用案例及其执行框架性能优化研究[D]. 南京:南京大学, 2012.
[5] 林子雨. 大数据技术原理与应用:概念、存储、处理、分析与应用[M]. 2 版. 北京:人民邮电出版社, 2017: 174-175.
[6] Yang A, Troup M, Lin Peijie, *et al*. Falco: a quick and flexible single-cell RNA-seq processing framework on the cloud[J]. Bioinformatics, 2017, 33(5): 767-769.
[7] Abuín J M, Pichel J C, Pena T F, *et al*. SparkBWA: speeding up the alignment of high-throughput DNA sequencing data[J]. PLoS ONE, 2016, 11(5): e0155461.
[8] Klein M, Sharma R, Bohrer C H, *et al*. Biospark: scalable analysis of large numerical datasets from biological simulations and experiments using Hadoop and Spark[J]. Bioinformatics, 2017, 33(2): 303-305.
[9] Zou Quan, Hu Qinghua, Guo Maozu, *et al*. HAlign: fast multiple similar DNA/RNA sequence alignment based on the centre star strategy[J]. Bioinformatics, 2015, 31(15): 2475-2481.
[10] Wan Shixiang, Zou Quan. HAlign-II: efficient ultra-large multiple sequence alignment and phylogenetic tree reconstruction with distributed and parallel computing[J]. Algorithms for Molecular Biology, 2017, 12(9): article No. 25.
[11] Trapnell C, Roberts A, Goff L, *et al*. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks[J]. Nature Protocols, 2012, 7(3): 562-578.
[12] Wasserstein R L, Lazar N A. The ASA statement on *p*-values: context, process, and purpose[J]. The American Statistician, 2016, 70(2): 129-133.
[13] Abuin J M, Pichel J C, Pena T F, *et al*. BigBWA: approaching the Burrows-Wheeler aligner to big data technologies[J]. Bioinformatics, 2015, 31(24): 4003-4005.
[14] Kim D, Langmead B, Salzber S L. HISAT: a fast spliced aligner with low memory requirements[J]. Nature Methods, 2015, 12(4): 357-360.

(上接第 165 页)

[6] Zhu Z, Cao S. Back-stepping sliding mode control method for quadrotor UAV with actuator failure[J]. Journal of Engineering, 2019, 2019(22): 8374-8377.
[7] Okyere E, Bousbaine A, Poyi G T, *et al*. LQR controller design for quadrotor helicopters[J]. Journal of Engineering, 2019, 2019(17): 4003-4007.
[8] 余润芝,赵文龙,程若发. 四旋翼飞行器的神经网络 PID 控制算法研究[J]. 现代电子技术, 2019, 42(10): 108-112.
[9] 李砚浓,李汀兰,姜艺,等. 基于 RBF 神经网络自适应 PID 四旋翼飞行器控制[J]. 控制工程, 2016, 23(3): 378-382.
[10] Rosales C, Tosetti S, Soria C, *et al*. Neural Adaptive PID control of a quadrotor using EFK[J]. IEEE Lat-in America Trans, 2018, 16(11): 2722-2730.
[11] 刘金锟. 先进 PID 控制 MATLAB 仿真[M]. 4 版. 北京:电子工业出版社, 2016.

(上接第 172 页)

[9] Fortunato S, Latora V, Marchiori M. A method to find community structures based on information centrality[J]. Physical Review E, 2004, 70(5): 056104.
[10] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks[J]. Physical Review E: Statistical Nonlinear and Soft Matter Physics, 2004, 70(6): 066111.
[11] Palla G, Deranyi I, Farkas I, *et al*. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435(7043): 814-818.
[12] Mark Newman E J. The structure and function of networks[J]. Computer Physics Communications, 2002, 147(1): 40-45.
[13] Newman M E J. The structure and function of complex networks[J]. SIAM Review, 2003, 45(2): 167-256.
[14] Lancichinetti A, Fortunato S, Kertesz J. Detecting the overlapping and hierarchical community structure in complex networks[J]. New Journal of Physics, 2009, 11(3): 033015.
[15] Lee C, Reid F, Mcdaid A, *et al*. Detecting highly overlapping community structure by greedy clique expansion[C/OL]//Proc of International Conference on Paper Presented at SNA-KDD Workshop. Washington, NJ: ACM Press, 2010. (2010-02-09)[2010-06-15]. https://arxiv. xilesou. top/abs/1002. 1827.
[16] Chen Qiong, Wu Tingting, Fang Ming. Detecting local community structures in complex networks based on local degree central nodes[J]. Physica A: Statistical Mechanics and Its Applications, 2013, 392(3): 529-537.
[17] 刘阳,季新生,刘彩霞. 一种基于边界节点识别的复杂网络局部社区发现算法[J]. 电子与信息学报, 2014, 36(12): 2809-2815.
[18] Li Yixuan, He Kun, Bindel D, *et al*. Uncovering the small community structure in large networks: a local spectral approach[EB/OL]. (2015-09-25). https://arxiv. org/abs/1509. 07715.
[19] Liu Jinglian, Wang Daling, Zhao Weiji, *et al*. A unified framework of lightweight local community detection for different node similarity measurement[C]//Proc of Chinese National Conference on Social Media Processing. Singapore: Springer, 2017: 283-295.
[20] 潘剑飞,董一鸿,陈华辉,等. 基于结构紧密性的重叠社区发现算法[J]. 电子学报, 2019, 47(1): 147-154.
[21] Lu Linyuan, Zhou Tao. Link prediction in complex networks: a survey[J]. Physica A: Statistical Mechanics and Its Applications, 2011, 390(6): 1150-1170.
[22] Zachary W W. An information flow model for conflict and fission in small groups1[J]. Journal of Anthropological Research, 1976, 33(4): 452-473.
[23] Lusseau D, Schneider K, Boisseau O J, *et al*. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations[J]. Behavioral Ecology & Sociobiology, 2003, 54(4): 396-405.
[24] Knuth D E. The Stanford GraphBase: a platform for combinatorial computing[M]. Nork York: ACM Press, 1993.
[25] Girvan M, Newman M E J. Community structure in social and biological networks[C]//Proc of National Academy of Sciences of the United States of America. 2002.
[26] Lancichinetti A, Fortunato S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities[J]. Physical Review E, 2009, 80(1): 016118.
[27] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks[EB/OL]. (2004-08-30). http://doi. org/10.1103/physreve. 70. 066111.
[28] Shen H, Cheng X, Cai K, *et al*. Detect overlapping and hierarchical community structure in networks[J]. Physica A: Statistical Mechanics and its Applications, 2009, 388(8): 1706-1712.
[29] Lancichinetti A, Fortunato S, Kertesz J. Detecting the overlapping and hierarchical community structure in complex networks[J]. New Journal of Physics, 2009, 11(3): 033015.
[30] Mcdaid A F, Greene D, Hurley N. Normalized mutual information to evaluate overlapping community finding algorithms[EB/OL]. (2013-08-02). https://arxiv. org/abs/1110. 2515.