基于模块度增量的二分网络社区挖掘算法

戴彩艳[1] 陈 崚[2,3] 胡孔法[1]
(南京中医药大学信息技术学院 南京 210016)[1] (扬州大学信息工程学院 江苏 扬州 225009)[2]
(南京大学计算机软件新技术国家重点实验室 南京 210093)[3]

摘 要 针对二分网络的社区挖掘问题,提出了一种基于模块度增量的二分网络社区挖掘算法。该算法假设每个顶点独自构成一个社区,并具有自己的标号。其中,一部分顶点将自己的标号复制并传递到另一部分中的某个顶点上,使之与其位于同一个社区;另一部分的顶点实施同样的操作。如此反复迭代,直至收敛。标号传播时,选择模块度增量最大的边进行传送,使整体模块度不断提高。在真实数据集上进行的测试表明,所提算法能对二分网络进行高质量的社区划分。

关键词 社区挖掘,二分网络,模块度增量,标号传播
中图法分类号 TP301.6 **文献标识码** A

Algorithm for Mining Bipartite Network Based on Incremental Modularity

DAI Cai-yan[1] CHEN Ling[2,3] HU Kong-fa[1]
(College of Information Technology, Nanjing University of Chinese Medicine, Nanjing 210016, China)[1]
(College of Information Engineering, Yangzhou University, Yangzhou, Jiangsu 225009, China)[2]
(State Key Lab of Novel Software Technology, Nanjing University, Nanjing 210093, China)[3]

Abstract Aiming at mining communities from bipartite network, an algorithm based on incremental modularity was proposed. The algorithm assumes that each vertex constitutes a community by itself with its own label. A part of the vertex copies its own label and passes it to a vertex on another part, so that it is located in the same community, and then it performs the same operation on the vertices of another part, and repeats iterations until convergence. In label propagation, the algorithm chooses the edge with the largest incremental modularity, so that the overall modularity is constantly improving. The experimental results on real datasets show that the proposed algorithm can mine high quality communities from bipartite network.

Keywords Mining communities, Bipartite network, Incremental modularity, Label propagation

1 引言

在自然界和人类社会中,各种复杂类型的系统都可以转化成复杂的网络[1],如生物系统[2]、经济系统[3]、群体生态系统[4]以及其他系统[5]。许多复杂系统都可以用网络和图来描述[6-7]。大多数的实际网络都具有社区结构,也就是说一个大的网络可以分成若干个子社区,这些子社区内部的联系是紧密的,而各子社区之间的联系是稀疏的[8]。从这些复杂的网络中挖掘社区并评估社区的质量,对于分析真实世界网络是非常重要的。

复杂网络领域的研究热点是社区结构及社区结构的挖掘。一个社区结构大致可被描述为:在这个社区内部,顶点之间的联系比较紧密;而社区之间的联系是比较稀疏的。因为许多网络都表现出这样的社区结构,所以对这种社区结构进行描述及挖掘极具实用意义。一个社区在结构上是相对独立的,因此它们各自对应一些基本的功能单元。

二分网络是复杂网络中一种重要的表现形式,由两部分不同类型的顶点构成,同一类型的两个顶点不相连。现实世界中的许多网络都呈现出自然的二分结构,比如:作者与他们所发表的论文著作之间形成作者-论文合作网络[9-10],演员与他们所演出的电影作品之间形成演员-事件合作网络[11],投资者与他们持有股份的公司之间形成股份网络[12-13],俱乐部成员与活动网络[14],观众与歌曲网络[15],疾病-基因网络[16]等。图1所示为一个二分网络,在该网络中,方形顶点为同一类型的顶点,圆形顶点为另一同类型的顶点,同类型顶点之间无边相连。从这些二分网络中挖掘社区,对于发现相似顶点

本文受国家自然科学基金(81674099,81503499),江苏省"青蓝工程"资助项目(2016),国家重点研发计划项目(2017YFC1703501,2017YFC1703503,2017YFC1703506),江苏省高校优势学科建设工程项目,江苏省教育信息化研究立项课题(20172097),江苏省现代教育技术研究重点课题(2017-R-54927)资助。

戴彩艳(1985-),女,博士,讲师,主要研究方向为社区挖掘,E-mail:daicaiyan@163.com;**陈 崚**(1956-),男,教授,博士生导师,主要研究方向为数据挖掘、人工智能;**胡孔法**(1970-),男,教授,博士生导师,主要研究方向为数据挖掘。

以及分析网络整体的社区结构是非常重要的。

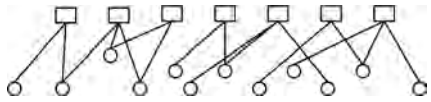

图 1　二分网络示意图

在最初对二分网络社区挖掘的研究中，一种方法是把二分网络映射成单分网络，这会导致信息的大量缺失，使得投影后的网络社区划分结果比原来的二分网络差得多。Guimera 等人[17]证明了这种投影分析对社区进行挖掘会产生错误的结果，甚至影响整个网络的社区结构。Barber[18]展示了对真实的二分网络及其投影的单分网络进行社区检测的不同结果，从而证实了投影分析二分网络的方式不可取。

最近，对于二分网络的社区挖掘，人们提出了许多种成功的算法。为了评估网络社区挖掘结果的质量，Newman[19]介绍了一种量化的方法，称为模块度。Guimera 等人[17]设计了二分模块度，并提出了一种社区挖掘算法，该算法每次只处理一种类型的顶点。Barber[18]拓展了 Newman 的单分网络模块度，提出了一种新的二分模块度，同时采用 Adaptive BRIM 算法来进行社区挖掘，并通过最大化获得二分模块度；但该算法只适用于对小规模的二分网络进行社区挖掘。Murata[20]基于 Newman 的单分网络模块度提出了一种二分模块度，其实验结果与 Newman 的单分网络模块度一致。基于二分模块度，Murata 等人[21]给出了一种新的社区挖掘方法，该方法可以挖掘一对一和多对多的社区关系。

Raghavan 等人[22]介绍了一种创新型和前景广阔的标号传播算法（1abel propagationalgorithm，LPA）来进行社区挖掘。该算法为每个顶点分配一个唯一的标签，在每次迭代过程中，每个顶点尽可能选取它们邻接顶点的标签，最终联系紧密且具有相同标签的顶点形成一个单独的社区。Fujita 等人[23]通过对标号传播算法进行改进，提出了一种更加适合二分网络的算法，该改进算法更适用于大型二分网络的并行实时社区分析。Murata 等人[24]同时提出了 LP&RRIM 算法，该算法是对 BRIM 算法和 LPA 算法的整合与改进，它基于标签传播算法，通过 BRIM 算法进行改善，产生了更好的社区划分，可以在大型二分网络中获得很好的划分质量。

本文提出了一种基于模块度增量的二分网络社区挖掘算法。该算法首先计算二分网络对应的模块度增量矩阵；然后对两种节点中的一类进行聚类，接着将该类顶点标号按模块度增量传递至第二类；对于第二类节点，同样将其顶点标号按模块度增量传递至第一类，直至所有顶点标号不再改变，就可以得到最终的划分结果。本算法不需要对社区的个数进行事先设定，而且得到的社区划分结果的质量较高。

网络的社区挖掘过程就是一个寻找对图 G 的划分方案，使得模块度最大，这实际上就是一个最优化问题。由于划分方案很多，寻找所有的划分方案是一个 NP-完全问题，因此有必要寻找针对二部图的快速且有效的社区挖掘算法来寻找尽可能精确的划分方案。

2　基于模块度增量的二分网络社区挖掘算法

算法的基本思想是：首先，假设每个顶点独自构成一个社区并具有自己的标号；然后，通过标号传播将不同的社区合并，表示顶点所在社区的标号通过顶点间的链接进行传递，U 部分的顶点将自己的标号复制并传递给 V 部分上的一个邻接顶点，使之与其位于同一个社区；接着，V 部分的顶点将自己的标号复制并传递给 U 部分上的一个邻接顶点，使之与其位于同一个社区。如此反复迭代，直到收敛为止。在其中一部分的顶点将自己的标号复制并传递到另一部分上的一个邻接顶点时，实际上是将邻接顶点吸纳到自己所在的社区中。在这种社区结构变化的过程中，须使得所得到的社区结构的模块度有所增加。为此，定义模块度增量的概念来衡量顶点合并到社区后社区结构模块度增加的程度。这样，在每次标号传递中，总是选择模块度增量最大的边进行传送，以使得整体模块度不断提高。

2.1　模块度增量和标号传递

本算法使用 Barber 模块度。假设将网络分为 k 个社区，即 $C_1, C_2, \cdots, C_k$，Barber 的模块度可以有如下的等价形式：

$$Q=\frac{1}{2m}\sum_{j=1}^{k}\sum_{v_i,v_k\in C_j}(a_{ik}-\frac{d_i d_k}{m}) \tag{1}$$

其中，m 为网络中边的总数，d_i 为顶点 v_i 的度，a_{ik} 为邻接矩阵中的元素。

式(1)的每一项 $a_{ik}-\frac{d_i d_k}{m}$ 中，a_{ik} 表示在同一社区中的顶点 u_i 和 v_k 间实际存在的链接数，$\frac{d_i d_k}{m}$ 表示顶点 u_i 和 v_k 间随机连接所产生的边数的期望值。由式(1)可以看出，如果顶点 u_i 和 v_k 被划入同一个社区，则 $a_{ik}-\frac{d_i d_k}{m}$ 被加入到整体的模块度 Q 中。因此，要将 $a_{ik}-\frac{d_i d_k}{m}$ 较大的顶点对 u_i 和 v_k 划入同一个社区，才能使得整体的模块度 Q 增大。称 $\Delta Q_{ik}=a_{ik}-\frac{d_i d_k}{m}$ 为顶点对 ui 和 v_k 模块度的增量。

定义 ΔQ 为一个 $l * m$ 阶的模块度增量的矩阵，l 是二部图一边的顶点数，m 是另一边的顶点数。ΔQ 中的 (i,j) 元素 ΔQ_{ij} 为 U 中顶点 v_i 与 V 中顶点 v_j 合并入一个社区后模块度的增量，被定义为：

$$\Delta Q_{ij}=\begin{cases}\frac{1}{m}(1-\frac{d_i d_j}{m}), & \text{若}(v_i,v_j)\in E\\ 0, & \text{否则}\end{cases} \tag{2}$$

在顶点 u_i 上进行标号传递时，算法总是选择使得 ΔQ_{ik} 最大的邻接顶点 v_k 进行传递。

标号传递的规则如下：

设 x 为 V 部一顶点，其部顶点集合表示为 $N(x)=\{y_1,\cdots,y_n\}$，记 $S_i=\sum_{\substack{y_i\in C_i\\ y_i\in N(x)}}\Delta Q(x,y_j)$ 为 x 邻接节点中属于 C_i 类的节点与 x 连边上的模块度增量的和，则节点 x 的标号为：

$$l(x)=\arg\max_{1\leqslant i\leqslant k} s_i \tag{3}$$

2.2　算法框架

综上所述，本文所提出的算法的框架如下。

算法 1　LP_CD（基于标签传递的社区检测算法）

输入：二部网络 G=(U,V,E)，其中 U=(u_1,u_2,…,u_m)，V=(v_1,v_2,…,v_n)；二部网络的邻接矩阵 A

输出：顶点 v_i 上的标号 l(v_i)，表示 v_i 所属的社区，i=1,2,…,n；顶点 u_i 上的标号 l(u_i)，表示 u_i 所属的社区，i=1,2,…,m

开始

1. 根据式(2)计算 ΔQ 矩阵中的所有元素 ΔQ_{ij}；

2. /* 对 U 部顶点进行初始聚类，设每个顶点即为一个类，即记 u_i 的标号 $l(u_i)=i$ */

for i=1 to m do

$l(u_i)=i$

endfor i

3. repeat

/* U 部顶点标号按模块度增量传递至 V 部顶点 */

for i=1 to m do

for k=1 to m do

$S_k=\sum_{\substack{u_j\in C_k\\ u_j\in N(v_i)}}\Delta Q(v_i,u_j)$；

endfor k；

$l(v_i)=\arg\max_k s_k$；

endfor i

/* V 部顶点标号按模块度增量传递至 U 部顶点 */

for i=1 to m do

for k=1 to m do

$S_k=\sum_{\substack{v_j\in C_k\\ v_i\in N(u_i)}}\Delta Q(u_i,v_j)$；

endfor k；

$l(u_i)=\arg\max_k s_k$；

endfor i

until 所有顶点标号不再改变

结束

2.3 “抖动”现象及其避免

在标号传递过程中，如果传递到 S_i 的值有多个是相同的，此时若不能确定性地选取，而是随机取值，就会出现“抖动”现象。

例：图 2 中顶点 y 在接收 a 或 b 的标号时如果是随机的，那么其对应的概率应该为 0.5。若第一次选 a 的标号“1”，则 x,y,z 的标号为(1,1,2)。再次传递至 a,b 后，a,b 的标号仍为(1,2)。若再次传递至 y 时，y 选 b 的标号“2”，则 x,y,z 的标号为(1,2,2)。这时就会出现 y 标号交替地取 1 或 2，即出现抖动现象。

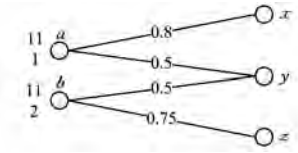

图 2 “抖动”现象示例图

为避免“抖动”现象，当在顶点 y 所连接的顶点 $N(y)$ 中出现 $\max S_i$ 最大值相同时，可以考查这些顶点的重要值。设 S_i 取最大值的顶点有 $x_1,x_2,\cdots,x_y$，每一个 x_i 的部集为 $N(x_i)$。

计算 $Score(x_i)=\dfrac{\Delta Q(y,x_i)}{\sum_{y_l\in N(x_i)}\Delta Q(y_l,x_i)}$

2.4 收敛性的证明

定理 1 算法 LP_CD 的迭代过程肯定收敛，即所有顶点标号在经过有限次迭代后会稳定在一个值。

证明：首先，由算法 LP_CD 的描述易见，对于 $G=(U,V,E)$ 的每一个顶点 x，设它每一次接收另一部顶点的一种标号，由于防“抖动”的措施，x 每次所接收到的标号是不一样的。由于标号的种类不会超过 m 和 n 的最小值，因此 x 最终会固定在某一个标号，而且必然会接收某固定顶点 y 的标号。为此，可以构建一个二分网络 $G'=(U,V,E')$，其中，U,V 是两种类型的节点，E' 是边的集合，为 E 的子集。设 $x\in U$ 每一次接收另一部顶点 $y\in V$ 的标签，从而会对应一条从 y 到 x 的直接边 (y,x)。设 $|U|=m,|v|=n,|E'|=m+n$，即共有 $m+n$ 条边。

可以证明，在 G' 中除了形如 $y\to x\to y$ 的环外，不可能有长度大于 2 的环。例如，出现图 3 所示的长度为 4 的环(在二部图中，环的长度只能为偶数)。

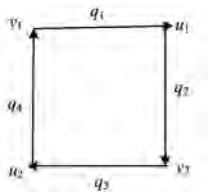

图 3 长度为 4 的环的示例图

由 v_1 点可知：$q_4>q_1$ (a)

由 u_1 点可知：$q_1>q_2$ (b)

由 v_2 点可知：$q_2>q_3$ (c)

由 u_2 点可知：$q_3>q_4$ (d)

由(b)，(c)，(d)可知 $q_1>q_4$，这与(a)矛盾。因此，G' 中不可能有长度大于 2 的环。

设在 G' 的顶点集合 $U_1\subset U,V_1\subset V$ 中，对于 $U_1\cup V_1$ 中任意的两顶点 x,y，存在一条由 x 到 y 的有向路径且其中的所有节点都属于 $U_1\cup V_1$，那么称 $U_1\cup V_1$ 为 G' 的一个传递集。算法 LP_CD 的传递过程实际上将 G' 的顶点划分成若干个传递集 $C_1,C_2,\cdots,C_k$，且满足：$C_i,C_j(i\neq j)$ 中任意两个顶点之间不存在有向边相连，而且每一个 C_i 中不可能有长度大于 2 的环。标号的传播只能在一个传递集中完成，直到传递集中所有的标号完全一样为止。如果 C_i 中不存在环 $x\to y\to x$，则标号传递可以在 $|C_i|$ 步内使得传递集中所有的标号完全一样。如果 C_i 中存在环 $x\to y\to x$，那么一次标签传递之后，点 x,y 的标签将变为一致，且以后也不会再改变，进而标号传递也将在 $|C_i|$ 步内使得传递集中所有的标号完全一样。因此，算法 LP_CD 的迭代过程肯定收敛。

证毕。

推论 上述迭代可以在 $m+n$ 步内收敛。

证明：C_i 是 G' 的最大传递集，且长度满足 $|C_i|\leqslant m+n$。因此，上述迭代在 $m+n$ 步内收敛。

相比于其他算法，基于模块度增量的算法具有如下优点：

1)它无须事先确定社区的个数。在大部分已有的社区挖掘算法中，必须要预先知道社区的个数，但这在实际应用问题中是不现实的。因为最佳社区的个数本身也是一个需要优化的参数，它是由网络的拓扑结构所决定的。在大部分已有的社区挖掘算法中，都是人为给定一个社区个数，这样的值不一定是最优的，从而存在一个“分辨率”问题。如果个数过多，即分辨率较细，社区划分得太分散；如果个数过少，即分辨率较粗，社区划分得太粗糙。而文中所提算法将对社区个数的优化与社区挖掘过程合并到一起进行，可以得到模块度最佳的社区个数和相应的划分。

2)它可以保证有较大的模块度值。算法的每一步迭代都在以增大模块度的值为目标，因此整个迭代过程可以保证整体模块度在不断增加，这是一个对模块度不断优化的过程，使得最终结果有较大的模块度值。

3)算法的复杂度较低。设 $m<n$，不难分析出，算法 LP_CD 的复杂度为 $O(m^2)$，比其他算法的低。

3　实验

3.1　实验环境

为了验证本算法的性能，在 3 个真实数据集上进行了测试。所有的实验都是基于奔腾 IV，Windows XP，1.7 GB 内存，并使用 VC++ 6.0 编程实现的。

3.2　实验结果

3.2.1　在 Southern Women 数据集上的测试

首先在 Southern Women 数据集上[25]进行了测试，这个数据集具有明显的社区结构，因此被广泛地用于测试分析。该数据集是由 18 个妇女和 14 个活动构成的二分网络，如果妇女参加了某个活动，那么网络中所对应的顶点之间存在着连接。Southern Women 数据集的组成成分如表 1 所列。

表 1　Southern Women 数据集的组成成分

妇女数	事件数	边	测试边	训练边	潜在边
18	14	93	10	83	68

首先，对这 18 个妇女和 14 个活动进行编号，1－18 表示 18 个妇女，19－32 表示 14 个活动；其次，使用基于模块度增量的二分社区挖掘算法 LP_CD 对 Southern Women 二分网络进行社区划分。LP_CD 算法将 Southern Women 二分网络划分成两个社区，即{妇女 1－9，活动 19－26}以及{妇女 10－18，活动 27－32}，此时对应的模块度为 0.585964。

Davis 等人使用常规的人种学知识将妇女划分为两个社区，分别是妇女 1－9 以及妇女 9－18，其中 9 是两个社区的共有项。于是，有研究者提出将妇女 1－9 划分为一个社区，并命名为“Davis 1”；将妇女 10－18 划分为另一个社区，并命名为“Davis 2”[26]。

另外，利用 Baber 提出的 BRIM 算法[18]进行社区划分，得到 4 个社区，分别是{妇女 1－6，活动 19－24}，{妇女 7，9，10，活动 25－26}，{妇女 8，16－18，活动 27－29}以及{妇女 11－15，活动 28，30－32}。Murata 提出使用 LPA 算法[22]将这个二分网络分为两个社区，它们分别是{妇女 1－7，9，活动 19－26}以及{妇女 8，10－18，活动 27－32}。将本文提出的算法(LP_CD)与 BRIM，Davis1，Davis2，LPA 以及 IRBC[27]进行比较。图 4 显示了使用不同算法对 Southern Women 二分网络进行社区划分得到的模块比较情况。

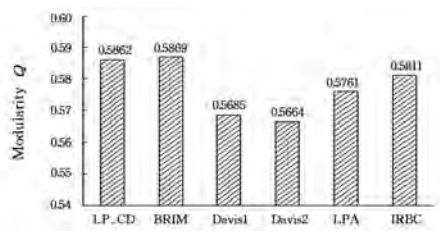

图 4　6 种不同算法对 Southern Women 进行划分得到的社区模块度

从图 4 中可以发现，除了 BRIM，LP_CD 得到的模块度是最高的。BRIM 虽然获得了最高的模块度，但是它是有前提的，即在社区划分之前首先要设定数据集只能被划分为 4 个社区；LP_CD 在划分前却无需确定划分的社区个数，而其模块度与 BRIM 的相差无几，只低了 0.0007。

3.2.2　在 Scotland 连锁企业数据集上的测试

本文还采用了 20 世纪初 Scotland 连锁企业的数据集[28]进行测试。该集合收集了 Scotland 早期的 108 家公司和 136 位股东之间的关系，每一位股东可能在不同的公司任职，每一家公司也可能有不同的股东，这样公司与股东之间就形成了二分网络的关系。与 Southern Woman 数据集不同，该数据集是非连通图，有许多离群点。表 2 则显示了 Scotland 中最大联通图的组成成分。

表 2　Scotland 中最大联通图的组成成分

法人	公司	边	测试边	训练边	潜在边
131	86	348	35	313	1423

实验中，对表 2 中的最大组成成分进行了社区划分，得到 16 个公司社区以及 22 个法人社区，对应的模块度为 0.4144。公司社区的个数和法人社区的个数是不等的，也就是说此时挖掘出的社区不再是一对一的关系，而是多对多的关系。

使用 BRIM 进行划分，得到 34 个社区，其中包含 17 个法人社区和 17 个公司社区。Baber 算法的前提是假设两种节点类型之间存在的关系是一对一的关系；并设定当社区个数少于 20 时，BRIM 算法可以获得最大的社区模块度 0.4152。LP_CD 算法的社区挖掘结果的模块度与 BRIM 相近，但是本文提出的算法无需事先设定社区的数目。而使用 LPA 算法对该数据集的社区进行挖掘时，对应的模块度为 0.4088，明显低于 LP_CD。

同样地，将 LP_CD 与 BRIM，Davis1，Davis2 以及 LPA 算法的模块度进行比较。图 5 显示了 6 种不同算法划分得到的社区对应的模块度情况。

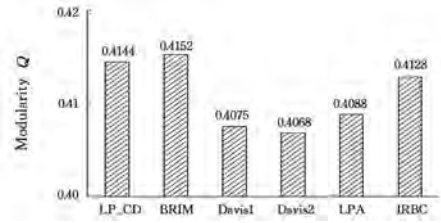

图 5　6 种不同算法对 Scotland 进行划分得到的社区模块度

从图 5 中可以看出，LP_CD 的模块度与 BRIM 的相近，只相差了 0.0008，但是高于 Davis1，Davis2 以及 LPA 算法的模块度，比其中最高的 IRBC 还高了 0.0016。因此，LP_CD 在事先无需确定社区个数的情况下，可以获得更高的模块度，确保了挖掘结果的质量。

3.2.3　在 MovieLens 数据集上的测试

最后在基准数据集 MovieLens[29]上进行了测试。该数据集由 943 个影评人和 1682 部电影组成。该集合收集了多个用户对多部电影的评级数据，每个影评人至少评价 20 部电影，投票级别为离散的 1－5 级。MovieLens 数据集的组成成分如表 3 所列。

表 3　MovieLens 数据集的组成成分

影评人数	电影数	边	测试边	训练边
943	1682	82520	8929	73591

实验中，对 MovieLens 数据集的组成成分进行了社区划分，得到 122 个影评人社区以及 261 个电影社区，对应的模块度为 0.4013。影评人社区的个数和电影社区的个数是不等的，挖掘出的社区不是一一对应的关系，而是多对多的关系。

使用 BRIM 进行划分，得到 390 个社区，其中包含 195 个法人社区和 195 个公司社区。但是，该算法的前提是假设两种节点类型之间存在的关系是一对一的。通过设定社区个数少于 200，可以使用 BRIM 算法获得最大的社区模块度 0.4084。LP_CD 算法的社区挖掘结果对应的模块度与 BRIM 接近，但是前者无需事先设定社区的数目。使用 LPA 算法对该数据集的社区进行挖掘时，对应的模块度为 0.3797；而使用 IRBC 获得的模块度则优于 LPA，为 0.3879。

同样地，将 LP_CD 与 BRIM，Davis1，Davis2，LPA 以及 IRBC 算法的模块度进行比较。图 6 显示了 6 种不同算法划分得到的社区对应的模块度情况。

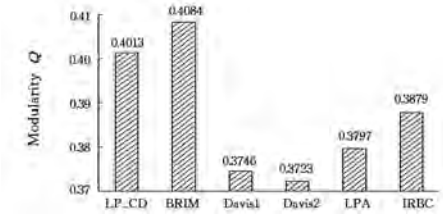

图 6 6 种不同算法对 MovieLens 进行划分得到的社区模块度

从图 6 中可以看出，LP_CD 的模块度与 BRIM 相近，只相差了 0.0071，但是高于 Davis1，Davis2，LPA 以及 IRBC 算法的模块度，比其中最高的 IRBC 高出 0.0134。因此，LP_CD 在事先无需确定社区个数的情况下，可以获得更高的模块度，确保了挖掘结果的质量。

3.2.4 对"抖动"处理的测试

通过 2.3 节的描述可知，在标号传递过程中可能出现"抖动"现象。为避免这种现象，可以在某顶点 y 所连接的顶点 $N(y)$ 中，当出现的最大值相同时，考查这些顶点的重要值。下面通过实验对使用"抖动"处理法与不使用"抖动"处理法的社区划分过程进行了测试。

首先，在 Scotland 数据集上，主要对添加"抖动"处理方法 LP_CD1 和不加"抖动"处理方法 LP_CD2 得到稳定模块度的时间先后进行了测试。图 7 显示了两者模块度达到稳定的时间情况。

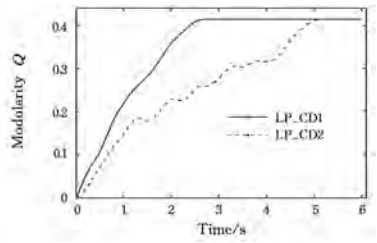

图 7 在 Scotland 数据集上有/无"抖动"处理的情况比较

从图 7 中可以看出，采用了"抖动"处理的 LP_CD1 在 2.5 s 时模块度就达到了稳定，而没有采用"抖动"处理的 LP_CD2 在 5 s 才达到稳定。出现这种现象的原因是：后者在标签传递过程中出现"抖动"现象时没有采取措施，而是任其继续往复传播，由此耗费了较多时间。

下面在 MovieLens 数据集上继续对 LP_CD1 和 LP_CD2 进行测试，结果如图 8 所示。

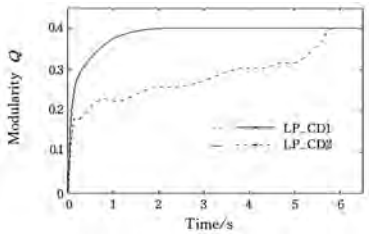

图 8 在 MovieLens 数据集上有/无"抖动"处理的情况比较

从图 8 中可以看出，LP_CD1 在 30 s 时模块度就达到了稳定，而 LP_CD2 在近 125 s 时才达到稳定。原因同样是后者在标签传递过程中出现"抖动"现象时没有采取措施，因此耗费了较多时间。

由以上两个实验可以看出，对"抖动"现象进行有效处理可以节约模块划分的时间，特别是当数据集比较大且抖动出现次数多时，有"抖动"处理的优势较为明显。如在 MovieLens 数据集上，有"抖动"处理与无"抖动"处理方法相比，模块度达到稳定的时间少了很多。

结束语 本文提出了一种基于模块度增量的二分网络社区挖掘算法，主要采用标号传播的方法，在标号传播过程中选择模块度增量最大的边进行传送，从而使整体模块度不断提高。在此过程中，还对传递中出现的"抖动"现象进行了处理。为了验证算法的正确性和有效性，在真实数据集上对算法进行了测试。实验结果表明，该算法能够从二分网络中抽取社区，获得高质量的网络社区，而"抖动"处理的使用则可以加快社区划分的速度。下一步准备基于标号传播的方法研究如何有效挖掘动态二分网络中的社区。

参 考 文 献

[1] BECKETT S J. Improved community detection in weighted bi partite networks[J]. Royal Society Open Science, 2016, 3(1): 140536.

[2] CUI Y Z, WANG X Y. Uncovering overlapping community structures by the key bi-community and intimate degree in bi-partite networks[J]. Physica A: Statistical Mechanics and Its Applications, 2014, 407(407): 7-14.

[3] 陈伯伦，陈峻，邹盛荣，等. 基于矩阵分解的二分网络社区挖掘算法[J]. 计算机科学，2014，41(2)：55-58，101.

[4] 辛宇，杨静，谢志强. 一种面向语义重叠社区发现的 Link-Block 算法[J]. 软件学报，2016，27(2)：363-380.

[5] HORVAT E A, ZWEIG K A. A fixed degree sequence model for the one-mode projection of multiplex bipartite graphs[J]. Social Network Analysis and Mining, 2013, 3(4): 1209-1224.

[6] 刘大有，金弟，何东晓，等. 复杂网络社区挖掘综述[J]. 计算机研究与发展，2013，50(10)：2140-2154.

[7] BECKETT S J. Improved community detection in weighted bi partite networks[J]. Royal Society Open Science, 2016, 3(1): 140536.

[8] LI J, WANG X, CUI Y. Uncovering the overlapping community structure of complex networks by maximal cliques[J]. Physica A Statistical Mechanics & Its Applications, 2014, 415: 398-406.

（下转第 464 页）

行分裂和合并,以解决过细分簇的问题,从而实现了对二分 K 均值聚类过程的改进。

参 考 文 献

[1] HAN J W,KAMBER M,PEI J. Data mining:concepts and techniques(3rd ed)[M]. Burlington:Elsevier Science,2011.

[2] ILLHOI Y,HU X H. A comprehensive comparison study of document clustering for a biomedical digital library MEDLINE [C]// Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries. New York,USA:ACM,2006:220-229.

[3] SILVA J D A,HRUSCHKA E R. Extending k-Means-Based Algorithms for Evolving Data Streams with Variable Number of Clusters[C]// International Conference on Machine Learning and Applications and Workshops. 2011:14-19.

[4] SAVARESI S M,BOLEY D. On the Performance of Bisecting K-Means and PDDP[C]// Proc. of the 1st SIAM International Conference on Data Mining. Chicago,USA:2001:1-14.

[5] 刘广聪,黄婷婷,陈海南. 改进的二分 K 均值聚类算法[J]. 计算机应用与软件,2015,32(2):261-263.

[6] VAMSI K B S,SATHEESH P,SUNEEL K R. Comparative Study of K-means and Bisecting K-means Techniques in Wordnet Based Document Clustering[J]. International Journal of Engineering and Advanced Technology,2012,1(6):119-234.

[7] 张军伟,王念滨,黄少滨,等. 二分 K 均值聚类算法优化及并行研究[J]. 计算机工程,2011,37(17):23-25.

[8] 裘国永,张娇. 基于二分 K-均值的 SVM 决策树自适应分类方法[J]. 计算机应用研究,2012,29(10):3685-3709.

[9] STEINBACH M,KARYPIS G,KUMAR V. A Comparison of Document Clustering Techniques[C]// Proc. of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston,USA,2000:525-526.

[10] LIU X Z,FENG G C. Kernel Bisecting K-Means Cluster- ing for SVM Training Sample Reduction[C]// Proc. of the 19th International Conference on Pattern Recognition. Tampa, USA, 2008:1-4.

[11] 戴东波,汤春蕾,熊赟. 基于整体和局部相似性的序列聚类算法[J]. 软件学报,2010,21(4):702-717.

(上接第 446 页)

[9] HIMMELSTEIN D S,BARANZINI S E. Heterogeneous Network Edge Prediction:A Data Integration Approach to Prioritize Disease-Associated GenesJ]. Plos Computational Biology,2015, 11(7):e1004259.

[10] IBRAHIM N M A,CHEN L. Link prediction in dynamic social networks by integrating different types of information[J]. Applied Intelligence,2015,42(4):738-750.

[11] 王祯骏,王树徽,张维刚,等. 基于社交内容的潜在影响力传播模型[J]. 计算机学报,2016,39(8):1528-1539.

[12] GAO F,MUSIAL K,COOPER C,et al. Link prediction methods and their accuracy for different social networks and network metrics[J]. Scientific Programming,2015,2015:1-13.

[13] CHEN G,WANG Y. Community detection in complex networks using extremal optimization modularity density[J]. Journal of Huazhong University of Science & Technology,2011,39(4): 82-85.

[14] LI Z,ZHANG S,ZHANG X. Modularity and community detection in bipartite networks[J]. American Journal of Operations Research,2015,5(5):421-434.

[15] BAKER A. Complexity,Networks,and Non-Uniqueness [J]. Foundations of Science,2013,18(4):687-705.

[16] KAYA B,POYRAZ M. Age-series based link prediction in evolving disease networks[J]. Computers in Biology and Medicine, 2015,63:1-10.

[17] GUIMERA R,SALES-PARDO M,AMARAL L A. Module identification in bipartite and directed networks[J]. Physical Review E,2007,76(2):066102.

[18] MICHAEL J. BARBER. Modularity and community detection in bipartite networks[J]. Physical Review E,2007,76(2):066102.

[19] EWMAN M E J. The Structure and Function of Complex Networks [J]. Siam Review,2003,45(2):167-256.

[20] MURATA T. Detecting communities from bipartite networks based on bipartite modularities[C]// 2009 International Conference on Computational Science and Engineering. 2009:50-57.

[21] LIU X,MURATA T. How does label propagation algorithm work in bipartite networks[C]// 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology(WI-IAT'09). 2009:5-8.

[22] RAGHAVAN U N,ALBERT R,KUMARA S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics,2007,76(32):036106.

[23] FUJITA S,FUJINO A. Word Sense Disambiguation by Combining Labeled Data Expansion and Semi-Supervised Learning Method[J]. Acm Transactions on Asian Language Information Processing,2013,12(2):1-26.

[24] LIU X,MURATA T. Community Detection in Large-scale Bipartite Networks[C]// IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. 2009:50-57.

[25] DAVIS A,GARDNER B B,GARDNER M R. Deep South[M]. University of Chicago Press,1941.

[26] ERGUN G. Human sexual contact network as a bipartite graph [J]. Physica A,2002,308:483-488.

[27] ZHANG P,WANG D,XIAO J. Improving the recommender algorithms with the detectedcommunities in bipartite networks [J]. Physica A,2017,471:147-153.

[28] SCOTT J,HUGHES M. The Anatomy of Scottish Capital:Scottish Companies and Scottish Capital[J]. Economic History Review,1980,3(4):1900-1979.

[29] KARUMUR R P,NGUYEN T T,KONSTAN J A. Exploring the Value of Personality in Predicting Rating Behaviors:A Study of Category Preferences on MovieLens[C]// ACM Conference on Recommender Systems. 2016:139-142.