基于矩阵分解的二分网络社区挖掘算法

陈伯伦[1] 陈 峻[2,3] 邹盛荣[2] 徐秀莲[4]

(南京航空航天大学计算机科学与技术学院 南京 210016)[1] (扬州大学信息学院计算机系 扬州 225009)[2]
(南京大学软件新技术国家重点实验室 南京 210093)[3] (扬州大学物理科学与技术学院 扬州 225009)[4]

摘 要 二分网络社区挖掘对复杂网络有重要的理论意义和应用价值。提出了一个基于矩阵分解的二分网络社区挖掘算法。该算法首先将二分网络分为两个部分,每个部分尽可能保存完整的社区信息,然后分别对两个部分进行递归的拆分,直至不能拆分为止。在拆分的过程中,应用矩阵分解,使得到的分解能与网络的相关矩阵的行空间尽可能接近,即尽可能保持原图的社区信息。实验结果表明,该算法在不需任何额外参数的情况下,不但能较准确地识别实际网络的社区个数,而且可以获得很好的划分效果。

关键词 二分网络,矩阵分解,社区检测

中图法分类号 TP301.6 **文献标识码** A

Detecting Community Structure in Bipartite Networks Based on Matrix Factorization

CHEN Bo-lun[1] CHEN Ling[2,3] ZOU Sheng-rong[2] XU Xiu-lian[4]

(Department of Computer Science, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)[1]
(Department of Computer Science, Yangzhou University, Yangzhou 225009, China)[2]
(State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing 210093, China)[3]
(College of Physics Science and Technology, Yangzhou University, Yangzhou 225009, China)[4]

Abstract Community detection in bipartite network is very important in the reseach on the theory and applications of complex network analysis. An algorithm for detecting community structure in bipartite networks based on matrix factorization was presented. The algorithm first partitions the network into two parts, each of which can reserve the community information as much as possible, and then the two parts are further recursively partitioned until they can not be partitioned. When partitioning the network, we used the approach of matrix decomposition so that the row space of the associated matrix of the networks can be approximated as close as possible and the community information can be reserved the as much as possible. Experimental results show that our algorithm can not only accurately identify the number of communities of a network, but also obtain higher quality of community partitioning without previously known parameters.

Keywords Bipartite network, Matrix factorization, Detecting community structure

1 引言

在自然界和社会中,许多复杂的系统都可以用网络和图来描述[1,2],而大多数的实际网络都具有社区结构,这样一个大的网络可以分成若干个子社区,在这些子社区的内部的结点相互作用强,而不同社区结点相互作用弱。网络社区结构挖掘算法能够帮助人们理解一个复杂的网络是如何基于一些基本网络构造模块组合而成的,这对于深入理解网络拓扑结构、挖掘隐含模式和预测网络行为都具有十分重要的意义,其广泛应用于多个领域。

自然界中存在的许多网络,都呈现出自然的二分结构,例如科学家与他们所发表的论文著作形成的科学家-论文合作网[3,4]、由电脑终端与数据形成的 P2P 互联网[5]、由演员与他们所演出的电影作品形成的演员合作网[6,7]、投资者与他们持有股份的公司之间形成的股份网[8,9]、某俱乐部中会员与会员参加的活动项目之间形成的活动网[10]、听众与歌曲网络[11]、疾病-基因网络[12]等等。这些网络都可以抽象成一个二部图,我们称之为二分网络。二分网络由两类节点以及两类节点之间的连边组成,而同类节点之间不存在连边。图1所示为一个二分网络,在该网络中,正方形为同一类型的节点,三角型为同一类型的节点,同一类型的节点之间没有连边,不同类型的节点之间可能存在着连边。

二分网络不仅具有普遍性,而且也是复杂网络中的一种重要的网络表现形式,已经成为复杂网络的重要研究对象。

到稿日期:2013-05-20 返修日期:2013-07-20 本文受国家自然科学基金项目(61070047,61070133,61003180),国家重点基础研究发展规划(973)项目(2012CB316003),江苏省自然科学基金项目(BK21010134),江苏省研究生创新基金(CXZZ13_0172)资助。

陈伯伦(1986-),男,博士,主要研究方向为复杂网络的链路预测,E-mail:chenbolun1986@163.com;陈 峻(1951-),男,教授,博士生导师,主要研究方向为数据挖掘、体系结构、并行计算;邹盛荣(1968-),男,副教授,硕士生导师,主要研究方向为复杂网络、算法分析、形式化方法;徐秀莲(1978--),女,博士,主要研究方向为复杂网络计算。

近年来，二分网络的研究遍及各个领域，学者们关注于网络的二分性和它的社区特性，因而一些更深层的网络性质被揭示出来。可以说，二分网络的研究为复杂网络的研究提供了新视角。对二分网络的社区挖掘在学术圈的探测、网页推荐、疾病诊断等方面有很多重要的应用。

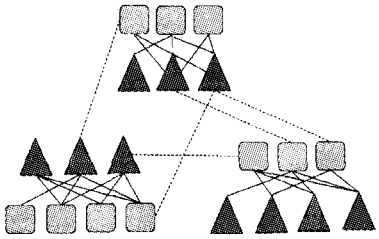

图1　二分网络的一个示例

二分网络社区发现在最初有两种思路：一个思路是把二分网络映射成为单分网络，利用较为成熟的单分网络社区发现算法进行社区划分。其投影方法也分为加权投影和无权投影，但实验证明无论何种投影方式都会导致信息缺失，从而导致社区划分失准。另一个思路则是直接在原始二分网络结构上做社区划分，因此需要对二分网络的网络性质进行新的研究，进而仿照单分网络的社区发现思想进行二分网络社区发现。目前很多研究者将注意力逐渐转向了对二分网络特性的研究。若对二分网络的统计性质做出延伸性的定义，则可以在这些性质的基础上得到相应的社区发现算法。此后，在原始二分网络上的各种统计性质被一一提出[13,14]，如集聚系数、边介数、度和度分布、重叠性、基于拓扑特征等等。一些研究者采用了基于边聚集系数的层次聚类方法[14,15]，他们把二分模块度作为算法的停止条件；Barber[16]提出了一种有效的Brim算法，用于对小规模的二分网络进行社区检测；Lehmann[17]等人提出了一种扩展的 k 团体算法来对二分网络进行检测；近来 Raghavan[18] 等人提出了标号传播的方法(LPA)来进行社区检测，后来 Tsuyoshi Murata[19] 等对标号传播的方法进行了改进，对二分网络进行了有效的划分；N Du 等人提出了基于最大二分子团过滤的 BiTector 算法[20]、王洋等人提出的基于比较性定义和社区作用力的聚类方法[21]都对二分网络进行了划分。

在本文中，我们从矩阵分解的角度提出了一个基于矩阵分解的二分网络社区挖掘算法。该算法首先将二分网络分为两个部分，每个部分尽可能保存完整的社区信息，然后分别对两个部分进行递归的拆分，直至不能拆分为止。在拆分的过程中，我们应用矩阵分解，使得到的分解能与原矩阵的行空间尽可能接近，即尽可能保持原图的社区信息。该算法在不需任何额外参数的情况下，不但能较准确地识别实际网络的社区个数，而且能获得很好的划分效果。

2　二分网络的社区及其模块度

二分网络可以表示成一个二部图 $G=(U,V,E)$，G 的顶点分为 U 和 V 两个部分，E 为 G 的边的集合。在集合 U(或 V)内的顶点之间不存在边相连，对 E 中所有的边 (u_i,v_j)，必有 $u_i\in U,v_j\in V$。因此二部图的邻接矩阵呈如下形式：

$$\tilde{A}=\begin{bmatrix} O & A \\ A^{\mathrm{T}} & O \end{bmatrix} \tag{1}$$

设 U 和 V 两个部分分别有 m、n 个顶点，则非零子矩阵 A 为一个 $m*n$ 阶矩阵。由于二部图 G 的邻接矩阵 $\tilde{A}$ 是对称的，我们可仅使用矩阵 A 来表示二部图 G，称 A 为二部图 G 的关系矩阵。A 的每一行表示 U 中一个顶点的链接方式，每一列表示 V 中一个顶点的链接方式。

我们要对 G 进行社区挖掘，就是要将其分解为若干个子图 $G_1,G_2,\cdots,G_p$，其中 $G_i=(U_i,V_i,E_i)$，$(i=1,2,\cdots,p)$，$U_i\subset U,V_i\subset V$，$\bigcup_{i=1}^{p}U_i=U$，$\bigcup_{i=1}^{p}V_i=V$，对 E_i 中所有的边 (u,v)，必有 $u\in U_i,v\in V_i$。因此，每一个子图 G_i 也是一个二部图，构成为一个社区。我们要求在每一个社区图 G_i 内部顶点间的链接尽可能地多，而不同社区 G_i 和 G_j 的顶点间的链接尽可能地少。

由于最佳社区的个数事先是没有给出的，社区发现算法的结果的优劣很难有一个统一的标准来衡量。对于单部的网络，Newman 等人提出用模块性(Modularity)[22]来衡量对图聚类的质量。模块度的物理意义是：网络中用连接属于同一个社区的节点的边比例，减去在同样社区结构下随机连接这两个节点边的比例的期望值。如果社区内部边连接的比例值不大于任意情况下连接时的期望值，则模块度为0，模块度值越大则表明社区结构越明显。如果模块度为1，则表示几乎所有社区之间是没有连接的。一般在实际网络中，模块度值取0.3～0.7之间即认为有明显的社区结构。因此这就成了一个最优化的问题，即寻找图的一个划分，使得模块度值最大。

对于一个二分网络也同样可以通过基于二分网络的模块度来进行衡量。Guimerà 等人[23]以及 Barber 等人[16]分别提出了两种二分网络模块度的定义及其相应的优化算法，两种定义同样都是以连接属于同一个社区的节点的边来衡量，但区别在于其最大化的表达途径各不相同。本文中，我们采取 Barber 等人提出的二分网络模块度为评价标准。

设 G 的两个部分的顶点分别为 $U=(u_1,u_2,\cdots,u_m)$ 和 $V=(v_1,v_2,\cdots,v_n)$，对于某种社区划分方案，Barber 等人的二分网络模块度定义为：

$$Q=\frac{1}{|E|}\sum_{i=1}^{m}\sum_{j=1}^{n}(A_{ij}-P_{ij})\delta(u_i,v_j) \tag{2}$$

式中，$P_{ij}=\frac{d_{ui}d_{vj}}{m}$ 为该网络随机模型中节点 u_i、v_j 存在连接概率的期望值，d_{ui}、d_{vj} 分别为节点 u_i、v_j 的度。$|E|$ 为网络图中边的总数。$\delta(u_i,v_j)$ 定义为：

$$\delta(u_i,v_j)=\begin{cases}1, & u_i,v_j \text{ 属于同一个社区}\\ 0, & \text{否则}\end{cases}$$

如上 Barber 等人的二分网络模块度定义与 Newman 等人提出的单部图上的模块性的基本思想是一致的。在式(2)中，A_{ij} 表示在同一社区的顶点 u_i 和 v_j 间实际存在的链接的个数，P_{ij} 表示顶点 u_i 和 v_j 间随机连接所产生的边的期望值。社区挖掘的任务就是寻找一种对 G 的划分，使得模块度 Q 达到最大值，这实际上是一个优化问题。由于划分的方案很多，这实际上是一个 NP-完全问题，因此，有必要研究快速有效的方法对二分网络进行社区挖掘。

3　算法的基本思想与框架

我们的基本思想是：首先将二部图 $G=(U,V,E)$ 拆分为两个部分：$G_0=(U_0,V_0,E_0)$；$G_1=(U_1,V_1,E_1)$，使它们尽可

能保存 G 中完整的社区信息。这里，G_0 和 G_1 也为二部图，且满足：$U_0 \cup U_1 = U, V_0 \cup V_1 = V$。为了检测可能重叠的社区，允许 U_1 和 U_2 之间、V_1 和 V_2 之间有交集。然后再对 G_0 和 G_1 进行递归的拆分，直至不能拆分为止。为了获得质量较高的社区拆分结果，用模块度来检测拆分效果，使得每次拆分都能提高模块度。如果继续拆分不能使得模块度得到提高，则拆分过程停止。

由于在式(1)中所示的二部图 G 的邻接矩阵 $\tilde{A}$ 是对称的，我们可仅使用关系矩阵 A 来表示二部图 G。A 的每一行表示 U 中一个顶点的链接方式，每一列表示 V 中一个顶点的链接方式。采用二部图 G 进行社区挖掘，本质上是对矩阵 A 的行列进行双向聚类或主成分分析的过程。为降低计算量，我们注意到 A 是一个二值矩阵，可以采用对 A 进行一维分解的方法，这个分解过程是在 A 的行(或列)中抽取具有代表性的模式，再根据各个行(或列)相对于该模式的距离将它们一分为二。将模式用二值行向量 y^T 表示，另使用二值向量 x 表示各行对模式的隶属关系，用向量 x 将顶点分为两部分。在每次拆分时，要求得向量 x、y，使 $\| A - xy^T \|^2$ 极小，也就是说使得矩阵 $A - xy^T$ 的零元素尽可能少。

因此，我们的算法 MP(Matrix partition)的总体框架如下：

Algorithm MP(G,A,P);
Input:G:二部图 G,其相关矩阵为 A_{m*n};
Output:P=$\{C_1, C_2, \cdots, C_n\}$:对 G 的社区划分;
Begin
1. DecompositionA (A,x,y);
/*对 A 进行一维向量分解，即寻找向量 x、y，使 $\| A - xy^T \|^2$ 极小;*/
2. DecompositionG (G,x,G_0,G_1);
/*将 G 中的顶点根据 x 分为两个二部图 G_0、G_1，它们的相关矩阵分别为 A_0 和 A_1;*/
3. 计算 G_1、G_0 的模块度 $Q(G_1)$、$Q(G_0)$;
4. 如果划分后的 G_1、G_0 的模块度的和大于划分前 G 的模块度，那么对 G_0、G_1 进行划分:
MP(G_1,A_1,P_1);MP(G_0,A_0,P_0);
5. 如果划分后的 G_1、G_0 的模块度的和小于划分前 G 的模块度，那么停止划分;
6. 输出对 G 的社区划分;
End

4 对二部网络的分解

算法 MP(G,A,P)的第一行调用过程 DecompositionA(A,x,y)，对二部图 G 的关系矩阵 A 进行一维向量分解，即寻找向量 x、y，使得 $\| A - xy^T \|^2$ 极小。因为 $\| A - xy^T \|_2^2 = \| A \|^2 - 2x^T Ay + \| x \|_2^2 \| y \|_2^2$，要使得 $\| A - xy^T \|^2$ 极小，只要使 $2x^T Ay - \| x \|_2^2 \| y \|_2^2$ 最大。我们采取迭代的方法，首先选定一个初始向量 y，以求得使 $2x^T Ay - \| x \|_2^2 \| y \|_2^2$ 最大的 x。在此 x 向量的基础上，求得使 $2x^T Ay - \| x \|_2^2 \| y \|_2^2$ 最大的 y。如此迭代下去，直到所得到的向量 y 不再变化为止。

因此，算法 DecompositionA(A,x,y)的描述如下：

Algorithm DecompositionA(A,x,y)
Input:A_{m*n}:关系矩阵;
Output:x_{m*1}、y_{n*1}:使得$||A - xy^T||^2$极小的向量;
Begin
1. 选取一个初始向量 y;
2. Repeat
2.1 Y′=y;
2.2 Z_y=Ay、n_y=$\| y \|_2^2$;
2.3 取使 $2x^T Z_y - n_y \| x \|_2^2$ 最大的 x;
2.4 $Z_x = x^T A$;$n_x = \| x \|_2^2$;
2.5 取使 $2Z_x y - n_x \| y \|_2^2$ 最大的 y;
Until y=y′
End

算法 DecompositionA(A,x,y)中第一步选取初始二值向量 y。具体的方法为：取 A 的行中与其它结点总体相似度最大的某一行作为初始向量 y。为此，我们对 A 中的所有行对之间计算其相似度。设第 i 行和第 j 行之间的相似度为 r_{ij}，它们之间的海明距离为 h_{ij}，我们定义 $r_{ij} = n - h_{ij}$，这里 n 为行向量的维数。r_{ij} 为 A 的行所代表的一类顶点中第 i 顶点和第 j 顶点与另一类顶点相连的次数，我们称其为第 i 顶点和第 j 顶点构成三角元(triple)的个数。对第 i 顶点定义 $r_i = \sum_j^n r_{ij}$ ($i = 1,2,\cdots,m$)，r_i 为结点 i 和其它结点构成三角元个数之和，我们取 r_i 最大的行向量作为 y^T。

在上述算法中，步骤 2.3 和步骤 2.5 计算使 $2x^T Z_y - n_y \| x \|_2^2$ 最大的二值向量 x 和使 $2Z_x y - n_x \| y \|_2^2$ 最大的二值向量 y，即求使 $2x^T Ay - \| x \|_2^2 \| y \|_2^2$ 最大的二值向量 x 或 y。我们以步骤 2.3 求二值向量 x 为例说明其具体的方法。我们逐项比较 Z_y(即 Ay)的第 i 个元素 $Z_y(i)$：如果 $2Z_y(i) > n_y$，则置 $x_i = 1$，否则 $x_i = 0$。

定理 1 对于 $m*n$ 阶二值矩阵 A 和 n 维二值向量 y，定义 m 维二值向量 x：若 m 维二值向量 Z_y(即 Ay)的第 i 个元素 $Z_y(i)$ 满足 $2Z_y(i) > n_y$，则 $x_i = 1$；否则 $x_i = 0$。因此 x 为使 $2x^T Ay - \| x \|_2^2 \| y \|_2^2$ 达到最大值的二值向量。

证明：因为 $2Z_y(i) - n_y > 0 (i = 1,2,\cdots,m)$，即 $2Ay(i) - \| y \|_2^2 > 0$，则 $2x^T Ay - \| x \|_2^2 \| y \|_2^2 > 0$。如果 $2Z_y(i) - n_y > 0$，此时 $x_i = 1$，可以保证 $2x^T Ay - \| x \|_2^2 \| y \|_2^2$ 在第 i 项上的值是正的，而且在 x 为二值向量的前提下，该项值已达到最大；反之，如果 $x_i = 0$，显然失去这一项，值会变小。如果 $2Z_y(i) - n_y < 0$，此时 $x_i = 0$，使得 $2x^T Ay - \| x \|_2^2 \| y \|_2^2$ 在第 i 项上的值为 0，可以保证不会使该值降低；反之，如果 $x_i = 1$，显然该值会变小。所以综上所述，x 为使 $2x^T Ay - \| x \|_2^2 \| y \|_2^2$ 达到最大值的二值向量。

由此可知，DecompositionA(A,x,y)算法求得的二值向量 x 和 y 可以使 $2x^T Ay - \| x \|_2^2 \| y \|_2^2$ 最大，即 $\| A - xy^T \|^2$ 极小。

在求得二值向量 x 和 y 后，算法 MP(G,A,P)的第 2 行调用过程 DecompositionG(G,x,G_0,G_1) 将 G 中的顶点根据二值向量 x 分量的值分为两个二部图 G_0、G_1。用 V_0、E_0 分别表示二部图 G_0 的结点集合和边的集合，V_1、E_1 分别表示二部图 G_1 的结点集合和边的集合。将 G 中的顶点分为两个二部图 G_0、G_1 的算法 DecompositionG(G,G_0,G_1)描述如下：

Algorithm DecompositionG(G,G_0,G_1)
Input:G:二部图，其相关矩阵为 A_{m*n}，顶点集合为 V，边集合 E;G 中有两类顶点，第一类有 m 个顶点，第二类有 n 个顶点;

Output：G_0，G_1：拆分后得到的两个二部图，$G_0=(U_0,V_0,E_0)$；$G_1=(U_1,V_1,E_1)$；

Begin：

1. for i=1 to m do

2. if $x_i=1$ then

将 U 中的第 i 个顶点 u_i 放入集合 U_1；

对于所有满足$(u_i,v_i)\in E$ 的 V 中顶点 v_j 放入集合 V_1 中；

将连接 u_i 和 v_j 的边放入集合 E_1 中；

3. else

将 u_i 放入集合 U_0；

对于所有满足$(u_i,v_i)\in E$ 的 V 中顶点 v_j 放入集合 V_0 中；

将连接 u_i 和 v_j 的边放入集合 E_0 中；

4. endif

5. endfor

End

5 实验结果与分析

为了测试上述算法的性能，我们用计算机生成的数据和实际数据对算法 MP 进行了实验，并与边聚集系数算法[14]、Brim 算法[16]、LPA 算法[18]、Davis 方法[24]与本文提出的算法进行了分析比较。实验在内存为 2.5G、CPU 为 CoreDuo 上运行，运行的操作系统为 WINDOWS7，使用 Java 编程。

5.1 在计算机生成的数据集上的测试

为了验证我们算法的正确性，用计算机生成大量代表二分网络的数据集，每一个二分网络包含了 128 个结点，把这些结点平均分成两个不同的大类，每一大类分别为 64 个结点，把二分网络平均分成 4 个社区，每一个社区有 32 个节点，其中每一个类型的节点都为 16 个。设置每个顶点的度 Z 为 16。Z_{in} 为顶点连接自己所属社区的度，Z_{out} 为连接其余社区的度，即 $Z=Z_{in}+Z_{out}$。同一类型的节点之间没有连接。计算了对于不同的 Z_{out}，将我们的算法 MP 和文献[14]提出的基于边集聚系数的聚类算法、基于拓扑特征算法的划分社区正确性进行了比较，结果如图 2 所示。

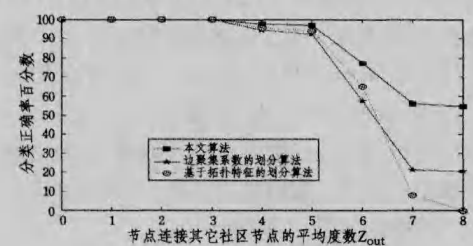

图 2 顶点分类正确性的比较

从图 2 中可以看出，我们的算法在 $Z_{out}<6$ 时聚类正确率达到 98%以上，比其余两种算法效果要好，而且我们的算法在 $6<Z_{out}<8$ 时聚类正确率也明显好于其余的两种算法。

5.2 在 Southern Women 数据集上的测试

我们还采用 Southern Women Data[24] 数据集进行实验。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。用我们的算法 MP 进行实验的结果如图 3 所示。其中社区以不同的颜色区分，顶点类型以图形形状区分，圆型代表妇女，正方型代表活动。

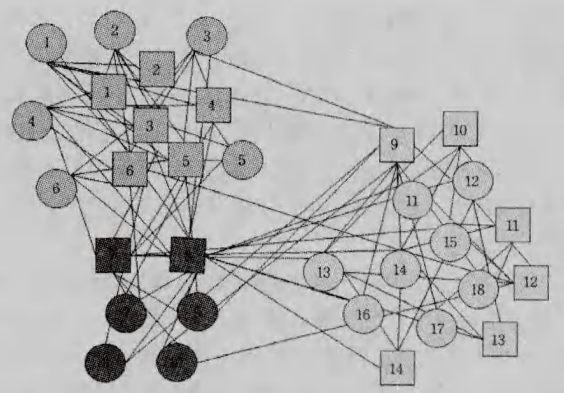

图 3 本文算法对妇女-活动网络划分的社区结构图

实验结果显示网络被划分为两个社区，其中妇女 1－6 与活动 1－6 被划分为一个社区，妇女 7－10 与活动 7、8 被划分为另一个社区，妇女 11－18 与活动 9－14 为第三个社区。用 Barber 提出的二分网路模块度进行计算，得出 $Q=0.3364$。我们也用 Brim 算法[16]、LPA 算法[18]、Davis 方法[24]与本文提出的 MP 算法进行了分析比较，比较的结果如图 4 所示。

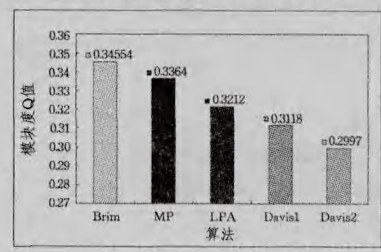

图 4 不同算法对 Southern women 数据集划分模块度的值

由图 4 可见，除了 Brim 算法，我们的 MP 算法的模块度比其他算法都高。Brim 算法所获得的模块度值是最大值时，它将数据集分成了 4 个社区，这 4 个社区需要进行人工的定义。而我们的 MP 算法不需要事先知道划分的社区个数，所得到的模块度和 Brim 算法所获得的模块度的值很接近，只相差 0.009。我们的算法比 LPA 算法的效果要好，Davis 方法在划分的过程中，妇女 9 被当成是重叠节点，属于两个社区。为方便比较，将该顶点的划分分别做了两种方案：一种方案在图 4 中记为 Davis1，表示将节点 9 同活动 1－8 合并的结果；另一种方案在图 4 中记为 Daviss2，表示同活动 10－18 合并的结果。但是我们的算法所获得的结果比这两者的结果都要好。因此，我们的算法在不需要知道社区划分的个数的情况下，可以得到较高的模块度，划分的质量也比较好。

5.3 在 Scotland 公司法人关系数据集上的测试

我们还采用了 20 世纪初苏格兰连锁企业的数据集[25]进行了测试。该集合收集了苏格兰早期的 108 个公司和 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。但与 Southern Woman 数据集不同，该数据集是非连通图，因此在实验之前首先提取其中的一个最大连通子图，该图由 131 位股东和 86 家公司组成。

通过实验最终将整个网络划分为 35 个社区，并得到模块度 0.5918。Barber[16] 用 Brim 算法对此数据集做实验时发现，如果控制社区的个数在 30 以内时，Brim 算法将会获得最

（下转第 101 页）

参 考 文 献

[1] Wu De-kai. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora[J]. Computational Linguistics, 1997, 23(3): 377-403

[2] Chiang D. A hierarchical phrase-based model for statistical machine translation[C]//Proceedings of ACL. 2005: 263-270

[3] Yamada K, Knight K. A syntax-based statistical translation model[C]//Proceedings of ACL. Toulose, France, 2001: 523-530

[4] Galley M, Hopkins M, Knight K, et al. What's in a translation rule? [C]//Proceedings of the 2004 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Boston, Massachusetts, 2004: 273-280

[5] Lin D, Cherry C. Word alignment with cohesion constraint[C]// Proceedings of the HLT-NAACL. Edmonton, Canada, 2003: 49-51

[6] Quirk C, Menezes A, Cherry C. Dependency treelet translation: Syntactically informed phrasal SMT[C]//Proceedings of ACL. Ann Arbor, Michigan, 2005: 271-279

[7] 熊德意. 基于括号转录语法和依存语法的统计机器翻译研究[D]. 北京：中国科学院计算技术研究所，2007

(上接第 58 页)

大的模块度的值，我们算法和 Brim 算法获得的模块度的值比较接近，但是我们不需要实现控制社区的个数。LPA 算法在求解次问题时，获得的最大模块度值为 0.5782。可以看出我们的算法要优于 LPA 算法。

结束语 本文提出了一个基于矩阵分解的二分网络社区挖掘算法。该算法首先将二分网络分为两个部分，每个部分尽可能保存完整的社区信息，然后分别对两个部分进行递归的拆分，直至不能拆分为止。在拆分的过程中，我们应用矩阵分解，使得到的分解能与网络的相关矩阵的行空间尽可能接近，以尽可能保持原图的社区信息。实验结果表明，该算法在不需任何额外参数的情况下，不但能较准确地识别实际网络的社区个数，而且可以获得很好的划分效果。

参 考 文 献

[1] Newman M E J. The structure and function of complex networks[J]. SIAM Rev., 2003(45): 16-256

[2] Strogatz S H. Exploring complex networks [J]. Nature, 2001 (410): 268-276

[3] Newman M E J. Scientific collaboration networks. I. network construction and fundamental results[J]. Physical Review E, 2001, 64: 016131

[4] Newman M E J. Scientific collaboration networks. II. shortest paths, weighted networks, and centrality[J]. Physical Review E, 2001, 64: 016132

[5] Le Blond S, Guillaume J L, LatapyM. C lustering in P2P exchanges and consequences on performances[C]// Castro M, Renesse R. Peer- to-Peer Systems IV. Berlin: Heidelberg, 2005: 193-204

[6] Watts D J, Strogatz S H. Collective dynamics of small world networks[J]. Nature, 1998, 393: 440-442

[7] 刘爱芬，付春花，张增平，等. 中国大陆电影网络的实证统计研究[J]. 复杂系统与复杂性科学，2007，4(3)：10-16

[8] Robins G, Alexander M. Small worlds among interlocking directors: network structure and distance in bipartite graphs[J]. Computational & Mathematical organization Theory, 2004, 10: 69-94

[9] Battiston S, Catanzaro M. Statistical properties of corporate board and director networks[J]. European Physics Journal B, 2004, 38: 345-352

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

[11] Lambiotte R, Ausloos M. Uncovering collective listening habits and music genres in bipartite networks[J]. Physical Review E, 2005, 72: 066107

[12] 陈文琴，陆君安，梁佳. 疾病基因网络的二分图投影分析[J]. 复杂系统与复杂性科学，2009，6(1)：13-19

[13] Lind P G, Gonzalez M C, Herrmann H J. Cycles and clustering in bipartite networks[J]. Physical Review E, 2005, 72: 056127

[14] Zhang P, Wang J L, Li X J, et al. Clustering coefficient and community structure of bipartite networks[J]. Physica A, 2008, 387: 6869-6875

[15] 吴亚晶，狄增如，等. 基于资源分布矩阵的二分网聚类方法[J]. 北京师范大学学报：自然科学版，2010，46(5)：643-646

[16] Michael J. Barber, Modularity and community detection in bipartite networks[J]. Phys. Rev. E 76, 066102, 2007

[17] Lehmann S, Schwartz M, Hansen L K. Biclique communities[J]. Phys. Rev. E, 2008, 78(1): 016108

[18] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks[J]. Phys. Rev. E, 2007, 76(3): 036106

[19] Liu X, Murata T. How does label propagation algorithm work in bipartite networks? [C]//Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT '09. Washington, DC, USA, 2009

[20] Du N, Wang B, Wu B, et al. Overlapping Community Detection in Bipartite Networks[C]//WI. 2008: 176-179

[21] 王洋，狄增如，樊琪. 二分网络社团结构的比较性定义[J]. 复杂系统与复杂性科学，2009，6(4)：40-44

[22] Newman M E J. Modularity and Community Structure in Networks [J]. Proc Natl Acad Sci USA, 2006, 103(23): 8577- 8582

[23] Guimera R, Sales-Pardo M, Amaral L A. Module identification in bipartite and directed networks[J]. Physical Review E, 2007, 76: 036102

[24] Davis A, Gardner B B, Gardner M R. Deep South [M]. Chicago: The University of Chicago Press, 1941

[25] Scott J, Hughes M. The Anatomy of Scottish Capital: Scottish Companies and Scottish Capital[C] // CroomHelm. London, 1980: 1900-1979