北京大学学报(自然科学版), 第 49 卷, 第 1 期, 2013 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 49, No. 1 (Jan. 2013)

基于社区节点重要性的社会网络压缩方法

李泓波[1,†] 张健沛[1] 杨静[1] 白劲波[2,3] 初妍[1] 张乐君[1]

1. 哈尔滨工程大学计算机科学与技术学院, 哈尔滨 150001; 2. 哈尔滨工程大学经济管理学院, 哈尔滨 150001;
3. 黑龙江工程学院计算机科学与技术学院, 哈尔滨 150050; † E-mail: islhb@126.com

摘要 针对目前图压缩方法中存在的时间复杂度较高、依赖先验知识设定参数、需要调节的参数过多、压缩有损、忽视网络社区结构等问题, 提出基于社区节点重要性的社会网络压缩方法。该方法由基于贪婪策略的社区发现算法(GS)和社会网络压缩算法(SNC)两部分组成。GS 算法采用拓扑势理论, 不但可以实现社区发现, 而且可挖掘出社区中的重要节点。SNC 算法以网络社区为压缩对象, 在保持社区间的关联关系的前提下实现了无损压缩, 并可在必要时保留社区中的重要节点或基本结构。通过实验, 对方法的可行性和有效性进行了验证。

关键词 社会网络挖掘; 拓扑势; 节点重要性; 无损压缩; 贪婪策略

中图分类号 TP391

Social Network Compression Based on the Importance of the Community Nodes

LI Hongbo[1,†], ZHANG Jianpei[1], YANG Jing[1], BAI Jinbo[2,3], CHU Yan[1], ZHANG Lejun[1]

1. College of Computer Science and Technology, Harbin Engineering University, Harbin 150001; 2. School of Economics and Management, Harbin Engineering University, Harbin 150001; 3. School of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin 150050; † E-mail: islhb@126.com

Abstract In response to the inadequacies of current graph compression methods, such as higher time complexity, dependence on experiences to set parameters, too many parameters to adjust, compression loss, ignoring the community structure of network, a social network compression method is proposed based on the importance of the community nodes. The method include community discovery algorithm (GS) based on greedy strategy and social network compression algorithm (SNC). Adopting topological potential theory GS algorithm is not only capable of discovering communities but also capable of mining important nodes in the communities. SNC algorithm takes communities as targets, achieves lossless compression while maintaining the connections between communities, and keeps important nodes in communities or basic community structure if necessary. The feasibility and effectiveness of the method are verified in experiments.

Key words social network mining; topology potential; importance of nodes; lossless compression; greedy strategy

图压缩(graph compression), 又被称为图简化(graph simplication)或图摘要(graph summarization), 可广泛应用于语义标签网络、重要节点发现、网络检索、网络可视化、网络分析等多个领域。近几年, 相继出现了一些比较典型的图压缩方法[1−6]。目前的图压缩方法大致可以分为无权图压缩和有权图压缩两大类, 采用的压缩方法一般为合并相似性节点。例如, 当节点 A 和节点 B 拥有相同或相似的共

国家自然科学基金(61073043, 61073041, 61100008)、黑龙江省自然科学基金(F200917, F201023, F200901)、高等学校博士学科点专项科研基金(20112304110011)、哈尔滨市优秀学科带头人基金(2010RFXXG002, 2011RFXXG015)和中央高校基本科研业务费专项资金(HEUCF061002)资助

收稿日期: 2012−05−31; 修回日期: 2012−08−29; 网络出版日期: 2012−10−26

网络出版地址: http://www.cnki.net/kcms/detail/11.2442.N.20121026.1755.018.html

同邻居节点时即进行合并，生成所谓的超级节点(supernode)，超级节点之间的边合并为超级边(superedge)。由此，这种方法会在超级节点间产生一些本来在网络中不存在的边，进而造成在解压缩(decompression)时产生误差。因此，这些方法是有损压缩方法。除此而外，这些方法还存在其他的不足，如需要付出较高的时间代价以衡量节点间的相似性，需要先验知识设定合并阈值，需设定较多的参数以满足不同的情况。

随着社会网络时代的来临，对社会网络展开深入研究是时代提出的必然要求。对社会网络进行可视化、知识发现等研究都会涉及社会网络压缩。随着社会网络规模的不断增大，社区发现已经成为社会网络应用过程中一个不可或缺的重要步骤[7]。社区作为社会网络重要的结构特征，在压缩过程中保留其重要节点或基本结构并保持它们之间的关联有着重要的意义和价值。然而，从现存的图压缩方法来看，以社区为压缩对象的研究目前还十分少见。

针对上述这些问题，本文提出一种基于社区节点重要性的社会网络压缩方法。本质上，社会网络压缩也是一种图压缩。但是，为区别于其他不以网络社区为考量前提的方法，我们将该方法称之为社会网络压缩方法。该方法将首先采用拓扑势理论进行社会网络上的社区发现并在此基础上区分社区节点的重要性，然后再进行基于社区节点重要性的压缩。

1 拓扑势理论简介

拓扑势理论[8]来源于核子物理学，用于指导社会网络上的社区发现。在核子物理学中，核子间的非接触式相互作用通过核子场进行刻画。拓扑势理论借鉴了核子场理论，认为存在直接或间接关系的节点间存在相互的作用力。这种节点间的作用力，在拓扑势理论中被称为拓扑势，像核子间的作用力一样，随着距离的增加而不断衰减，直至衰减为零。拓扑势理论中节点间的距离指的是节点间的拓扑距离，而非核子场理论中的欧氏距离。在拓扑势理论中规定一个节点的拓扑势即为其他节点对其作用力的加和。

因此，如果一个节点拥有较高的拓扑势，那么一定说明它的周围围绕着较多的有直接联系或间接联系的节点。由于富人俱乐部现象在社会网络中是普遍存在的，所以社会网络一般不为均匀网络或规则网络，在整个网络中总会存在一些具有高拓扑势的局部极值节点。这些节点及其周边节点通过它们之间的联系紧密地缠绕在一起，形成社区结构。如果找到这些拥有较高拓扑势的极值节点，就相当于找到了社区的代表点，如能进一步以某种比较合适的方式从社区代表点开始搜索与之联系相对紧密的节点，就可实现社区发现。容易看出，拓扑势社区发现方法明显区别于那些需要指定社区数目和社区规模的社区发现方法[9−10]。

下文中涉及的符号、公式和概念请参阅文献[8, 11]。

2 社区节点重要性分析

从社区构成的层面来说，应用拓扑势方法发现的社区中的节点的重要性是存在差别的。为了说明社区节点的重要性差异，我们首先给出如下的定理和推论。

定理 1 设节点 u 和 v 处于某社会网络中社区代表点 v^* 的一条吸引链上，且 u 位于 v^* 的第 a 跳，v 位于 v^* 的第 $a+1$ 跳，$a=0,1,2,\ldots,h-1$，则 u 和 v 对 v^* 的拓扑势贡献量比值 $R_{u\leftarrow v}(a,a+1)=\mathrm{e}^{\frac{2a+1}{\sigma_{\mathrm{opt}}^2}}$。

证明 由文献[11]的式(7)可推知任意一个处于吸引链上的节点 p 对社区代表点 v^* 的拓扑势贡献量为

$$A_{v^*\leftarrow p}(\sigma_{\mathrm{opt}},l)=\frac{1}{n}\mathrm{e}^{-\left(\frac{l}{\sigma_{\mathrm{opt}}}\right)^2}, \tag{1}$$

l 为 p 离开 v^* 的最小跳数。

由式(1)，易知 u 和 v 对 v^* 的拓扑势贡献量分别为 $A_{v^*\leftarrow u}(\sigma_{\mathrm{opt}},a)=\frac{1}{n}\mathrm{e}^{-\left(\frac{a}{\sigma_{\mathrm{opt}}}\right)^2}$ 和 $A_{v^*\leftarrow v}(\sigma_{\mathrm{opt}},a+1)=\frac{1}{n}\mathrm{e}^{-\left(\frac{a+1}{\sigma_{\mathrm{opt}}}\right)^2}$，因此二者的贡献量之比为

$$R_{u\leftarrow v}(a,a+1)=\frac{A_{v^*\leftarrow u}(\sigma_{\mathrm{opt}},a)}{A_{v^*\leftarrow v}(\sigma_{\mathrm{opt}},a+1)}=\mathrm{e}^{\frac{2a+1}{\sigma_{\mathrm{opt}}^2}}。$$

推论 1 设节点 u 和 v 处于某社会网络中社区代表点 v^* 的一条吸引链上，且 u 位于 v^* 的第 a 跳，v 位于 v^* 的第 $a+1$ 跳，$a=0,1,2,\ldots,h-1$，则 u 和 v 对 v^* 的拓扑势贡献量比值 $R_{u\leftarrow v}(a,a+1)>1$。

证明 由定理 1 知 $R_{u\leftarrow v}=\mathrm{e}^{\frac{2a+1}{\sigma_{\mathrm{opt}}^2}}$，又已知 $a=0,1,2,\ldots,h-1$，$\sigma_{\mathrm{opt}}>0$，于是就有 $2a+1>0$，$\sigma_{\mathrm{opt}}^2>0$，$\frac{2a+1}{\sigma_{\mathrm{opt}}^2}>0$，$R_{u\leftarrow v}(a,a+1)=\mathrm{e}^{\frac{2a+1}{\sigma_{\mathrm{opt}}^2}}>1$。

推论 2　设节点 u，v 和 w 处于某社会网络中社区代表点 v^* 的一条吸引链上，且 u 位于 v^* 的第 a 跳，v 位于 v^* 的第 $a+1$ 跳，w 位于 v^* 的第 $a+2$ 跳，$a=0, 1, 2, ..., h-2$，则有

$$R_{v\leftarrow w}(a+1,a+2) > R_{u\leftarrow v}(a,a+1)。$$

证明　因为 $\sigma_{\text{opt}}^2>0$，a 为非负整数，且对一个给定网络来说 σ_{opt} 为一定值，所以 $\frac{2(a+1)+1}{\sigma_{\text{opt}}^2} > \frac{2a+1}{\sigma_{\text{opt}}^2}$。由定理 1 知 $R_{u\leftarrow v}(a,a+1)=\mathrm{e}^{\frac{2a+1}{\sigma_{\text{opt}}^2}}$，$R_{v\leftarrow w}(a+1,a+2) = \mathrm{e}^{\frac{2(a+1)+1}{\sigma_{\text{opt}}^2}}$，又知 e^x (x>0)是一个严格单调增函数，因此 $R_{v\leftarrow w}(a+1,a+2) > R_{u\leftarrow v}(a,a+1)$。

推论 3　设节点 u，v 和 w 处于某社会网络中社区代表点 v^* 的一条吸引链上，且 u 位于 v^* 的第 a 跳，v 位于 v^* 的第 $a+1$ 跳，w 位于 v^* 的第 $a+2$ 跳，$a=0, 1, 2, ..., h-1$，则有

$$R_{v\leftarrow w}(a+1,a+2)=\mathrm{e}^{\frac{2}{\sigma_{\text{opt}}^2}}R_{u\leftarrow v}(a,a+1)。$$

证明　由定理 1 知 $R_{u\leftarrow v}(a,a+1)=\mathrm{e}^{\frac{2a+1}{\sigma_{\text{opt}}^2}}$，$R_{v\leftarrow w}(a+1,a+2)=\mathrm{e}^{\frac{2(a+1)+1}{\sigma_{\text{opt}}^2}}$，所以 $R_{v\leftarrow w}(a+1,a+2)\ /\ R_{u\leftarrow v}(a,a+1) = \mathrm{e}^{\frac{2(a+1)+1}{\sigma_{\text{opt}}^2}} \Big/ \mathrm{e}^{\frac{2a+1}{\sigma_{\text{opt}}^2}} = \mathrm{e}^{\frac{2}{\sigma_{\text{opt}}^2}}$，即

$$R_{v\leftarrow w}\ (a+2,a+1)=\mathrm{e}^{\frac{2}{\sigma_{\text{opt}}^2}}R_{u\leftarrow v}(a,a+1)。$$

推论 4　设节点 u，v，x 和 y 处于某社会网络中社区代表点 v^* 的一条吸引链上，且 u 位于 v^* 的第 a 跳，v 位于 v^* 的第 $a+1$ 跳，x 位于 v^* 的第 b 跳，y 位于 v^* 的第 $b+1$ 跳，$a, b=0, 1, 2, ..., h-1$ 且 $b>a$，则有 $R_{x\leftarrow y}(b,b+1)=\mathrm{e}^{\frac{2(b-a)}{\sigma_{\text{opt}}^2}}R_{u\leftarrow v}(a,a+1)$。

证明　由定理 1 知 $R_{x\leftarrow y}(b,b+1)=\mathrm{e}^{\frac{2b+1}{\sigma_{\text{opt}}^2}}$，$R_{u\leftarrow v}(a,a+1)=\mathrm{e}^{\frac{2a+1}{\sigma_{\text{opt}}^2}}$，所以 $R_{x\leftarrow y}(b,b+1)/R_{u\leftarrow v}(a,a+1) = \mathrm{e}^{\frac{2b+1}{\sigma_{\text{opt}}^2}} \Big/ \mathrm{e}^{\frac{2a+1}{\sigma_{\text{opt}}^2}} = \mathrm{e}^{\frac{2(b-a)}{\sigma_{\text{opt}}^2}}$，即 $R_{x\leftarrow y}(b,b+1) = \mathrm{e}^{\frac{2(b-a)}{\sigma_{\text{opt}}^2}} R_{u\leftarrow v}(a,a+1)$。

表 1 列出了若干网络中距离相差一跳的节点对代表点的贡献量比值，可用于验证上述定理和推论的正确性。定理 1 及其推论充分说明，在拓扑势方法发现的社区中，近邻节点较之远邻节点对代表点拓扑势的贡献更大，随着与代表点距离的加大，节点的贡献量指数倍下降。因此，拥有局部极值的社区代表点，其近邻节点的数量更多，它们之间的联系也更紧密，并形成了社区的核心结构。远邻节点对代表点的拓扑势的贡献则相对小得多，数量也相对得少，它们之间的联系也更加稀疏。综上所述，从社区构成的层面来说，代表点的近邻节点较之远邻节点具有更高的重要性。

上述的结论与人们对社区的直观感受是完全一致的。另外，我们还可以从其他一些层面来说明社区代表点的近邻节点的重要性要高于远邻节点的重要性。例如，研究发现既鲁棒又脆弱是复杂系统和复杂网络的基本特征之一，而引发复杂网络脆弱性的一个重要手段是有意识地对拥有高度数的节点发起攻击[17]。在这种攻击策略下，网络的连通性受到破坏，信息流通受到阻滞。按照这种攻击策略，有意识地攻击高度数节点的一跳邻居，同样能破坏网络的连通性；相反地，如果对高度数节点的远邻进行攻击，则很难破坏网络的连通性。这说明高度数节点及其近邻在网络中具有比远邻节点更高的重要性。实证研究[8,11,18]及前述的定理和推论都表明拓扑势方法发现的社区代表点即为网络中的高度数节点，因此利用拓扑势方法发现的社区中的节点的重要性也存在高低之分。

3　方法基本思想及算法描述与分析

3.1　方法基本思想

从前面的分析可以知道，在拓扑势方法发现的

表 1　若干网络的 $R_{u\leftarrow v}(a,a+1)$ 值

Table 1　$R_{u\leftarrow v}(a,a+1)$ values of several networks

节点 u 离开代表点 v^*的跳数 a	节点 v 离开代表点 v^*的跳数 $a+1$	空手道俱乐部网络[12] (σ_{opt}=1.0204)	海豚社会网络[13] (σ_{opt}=1.1782)	邻词接网络[14] (σ_{opt}=1.0043)	场景人物网络[15] (σ_{opt}=1.0435)	美国政治书籍[16] (σ_{opt}=0.9803)
1	2	17.8365	8.6810	19.5772	15.7225	22.6869
2	3	121.7630	36.6679	142.2059	98.6741	181.8129
3	4	831.2309	154.8820	1.0330×10^3	619.2763	1.4571×10^3

社区中代表点的邻居节点的重要性是逐跳降低的。因此, 本文压缩方法将首先基于拓扑势理论进行社区发现, 然后再进行网络压缩。本文将网络压缩定义如下。

定义 1 给定网络 $G=(V,E)$ (V 和 E 分别为 G 的顶点集合和边集合), 若存在一个网络 $G'=(V',E')$, 且有 $V'\subset V$ 及 $E'\subset E$, 那么就称 G' 为 G 的一个网络压缩。

本文方法采用相对代表点从外向内压缩的方式进行, 最多可压缩到网络中只剩下代表点。该方法的优势之一体现为: 在压缩过程中不但可以压缩掉一些相对不重要的节点, 有效降低网络规模, 同时也可在必要时保留社区中的重要节点或社区的基本结构。

区别于一般的图压缩方法, 本文方法在压缩过程中无须用户指定经验参数, 而只须在方法自动确定的优化影响范围 h 的指导下指定要压缩到的跳数即可。图 1 表现了一个具有优化影响范围为 2 跳的社区的压缩示意图, 单向箭头代表从某跳可压缩至另一跳。

3.2 算法描述

3.2.1 算法基本数据结构

为了实现无损压缩, 本文方法在社区发现过程中即标明所有节点距离代表点的跳数, 这些标记写在存储社区的链表结构中。由于在压缩过程中有可能失去社区间的关联关系, 所以又设计了存储社区间关系的链表。方法中用到的基本数据结构如下。

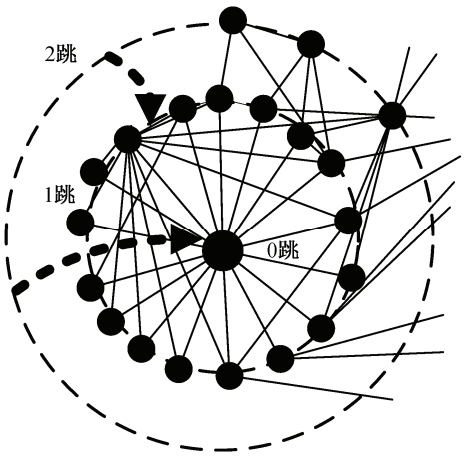

图 1 社区压缩示意图

Fig. 1 Community compression schematic diagram

```
//存储各个社区的链表结构
typedef struct CommNode{
    int Node
    int hop
    struct CommNode *nextnode
}CommNode
typedef struct {
    int RepNode
    int totalhop
    CommNode *FirstNode
}VexNode,CommunityArr[maxSize]
//定义图的类型
typedef struct{
    int adjMatrix[maxSize][maxSize]
    double potential[maxSize]
    CommNode *RepSet
    int TagArr[maxSize]
    CommunityArr Comm
    MulAttrNodeArr BoundNode
    CommRelation *CommR
}Graph
//存储社区间关系的链表类型
typedef struct CommRelation{
    int c1
    int c2
    struct CommRelation *next
}CommRelation
//存储边界结点的链表结构
typedef struct RepNode{
    int RepNodeNum
    double probability
    struct RepNode *next
}RepNode
typedef struct MulAttrNode{
    int NodeNum
    RepNode *FirAttrNode
}MulAttrNode, MulAttrNodeArr[maxSize]
```

3.2.2 基于贪婪策略的社区发现算法 GS

目前的社会网络发现方法, 在对社区中某些节点(特别是处于边缘的节点)进行处理时, 实质上是人为地割裂了这些节点与另一社区中节点的联系。为了比较形象地说明问题, 将图 2 所示的空手道网络看成犯罪网络, 而且重点考察犯罪嫌疑人节点 3。节点 3 是一个比较容易被误分的节点, 一般认为它应归入右边的社区, 但有的算法将其归入左边的社区(如 GN 算法[9]、模块度优化算法[19]和谱二分法[20]), 还有的算法将其既归入左边社区又归入右边的社区(如一些重叠社区算法[8])。然而, 不论将其归入哪个社区, 都是将节点 3 作为社区的边缘节点, 而不能将它和那些与它联系相对紧密的另一个社区中的节点放入同一个范围内进行分析。在对犯罪网络分析时, 上述传统做法极易忽略掉一些重要线索, 割裂一些节点与另一社区(团伙)的联系, 导

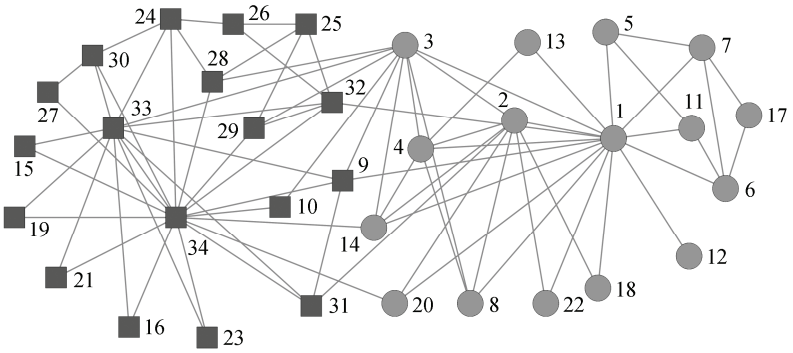

图 2　空手道俱乐部网络
Fig. 2　Karate club network

致信息损失。

另外, 由于目前的拓扑势社区方法中用于决定重叠节点社区归属的效益函数过于严格, 导致发现的社区中重叠节点过于稀少, 与现实世界中普遍存在的个体同时隶属于多个社区的情况存在较大出入。

为了克服上述问题, 本文提出基于贪婪策略的社区发现算法GS(greedy strategy), 该方法在确定了社区代表点之后, 以贪婪策略沿着吸引链遍历被代表点吸引的所有节点。GS 算法的具体描述如下。

```
Input network G = (V, E), |V| = n, |E| = m
Output community Ci  (i 为代表点编号)
InitComm( ); InitTag(TagArr,−1); InitBoundNode( )
for (each node v in RepSet){
  i = GetRepNode(v); j = 0; InsertComm(i, i, j)
 InitQueue(Q1); InitQueue(Q2); EnQueue(&Q1, i);
          TagArr[i] = i
 while(Q1 is not empty){
   j ++
   while (Q1 is not empty){
   DeQueue(&Q1, &u)
      for (拓扑势小于等于 u 的邻居节点 w)
         if(TagArr[w] != i){
           InsertComm(i, w, j)
           if(TagArr[w] == −1){TagArr[w] = i;
           EnQueue(&Q2, w);}
           else{if(社区 TagArr[w]和社区 i 之间的关联
                关系不在社区关系链表 CommR 中)
             InsertRelation(CommR, TagArr[w], i)
               if(BoundNode[w] is empty)
                  InsertBoundNode(w,TagArr[w])
               InsertBoundNode(w, i)
               TagArr[w] = i; EnQueue(&Q2, w)
                }
           }
      }
     while(Q2.front !=Q2.rear) {DeQueue(&Q2, &u);
       EnQueue(&Q1, u);}
   }
  G.Comm[i].totalhop = j
 }
```

3.2.3　社会网络压缩算法 SNC

在 GS 算法发现社区的基础上, 社会网络压缩算法 SNC(social network compression)通过交互方式获得要压缩到的跳数, 然后再进行压缩操作。SNC 算法的具体描述如下。

```
Input network G = (V,E), |V| = n, |E| = m,OptSigma
Output compressed community Ci (i 为社区代表节点
的编号, 对于每个社区 Ci, 只显示用户指定跳数以内的
节点)
DiscoverCommunity( )
h = (int)(3*OptSigma/sqrt(2))
cout<<"\n 当前网络的优化影响范围为: "<<h
cout<<"\n 请输入要显示的跳数"<<"(<="<<h<<"): "
cin>>hop
for(i = 0; i < maxSize; i ++){
    p = G.Comm[i].FirstNode
    while(p) {
        if(p->hop <= hop) display(p->Node)
        p = p->nextnode;
    }
}
r = G.CommR->next
while(r){
    display(r->c1, r->c2)
    r = r->next
}
```

3.2.4　算法时间复杂度分析

本文方法涉及用于进行社区发现的 GS 算法和进行社会网络压缩的 SNC 算法。相比较而言, GS 算法的时间复杂度要高一些。由于 GS 算法在最坏的情况下也不会超过 $O(n^2)$ (n为当前网络中的节点个数), 所以本文方法的时间复杂度不会超过 $O(n^2)$。

4　实验及分析

4.1　实验

为验证本文方法的可行性和有效性, 在空手道俱乐部网络[12]和海豚社会网络[13]两个被广泛使用

的数据集上对方法进行测试。两个网络中的节点编号与 Newman[16]提供的编号保持一致。

4.1.1 空手道俱乐部网络上的压缩实验

空手道俱乐部网络是 Zachary 依据其成员的交往情况绘制的。该俱乐部由于纠纷最终分裂为以教练和主管各自为核心的两个团伙，如图 2 所示。

应用 GS 算法在空手道俱乐部网络上进行社区发现，结果如图 3 所示。在图 3 中圆形和方形图标分别用于标示两个不同的社区，大图标用于标示社区的代表点，三角形图标用于标示社区间的重叠节点。图 4~6 中图标的含义与图 3 相同。

在 GS 算法发现的社区上，应用 SNC 算法对发现的社区进行 2 跳、1 跳和 0 跳压缩，压缩结果如图 4~6 所示。图 4~6 中的双向箭头用于标示两个社区间的关联关系。

4.1.2 海豚社会网络上的压缩实验

图 7 所示的海豚社会网络反映了两个海豚家族成员间的交互关系，较大家族中成员数为 42 名，较小家族中成员数为 20 名。

应用 GS 算法在海豚社会网络上进行社区发现，结果如图 8 所示。图 8 中圆形、方形图标和星形图标分别用于标示 3 个不同的社区，大图标用于标示

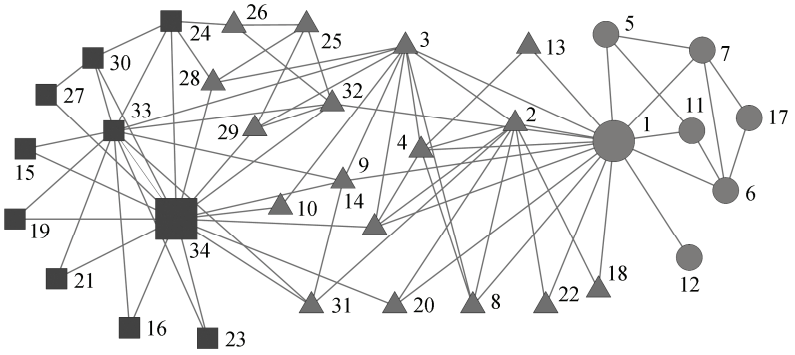

图 3 GS 算法在空手道俱乐部网络上发现的社区

Fig. 3 Communities discovered by GS algorithm on karate club network

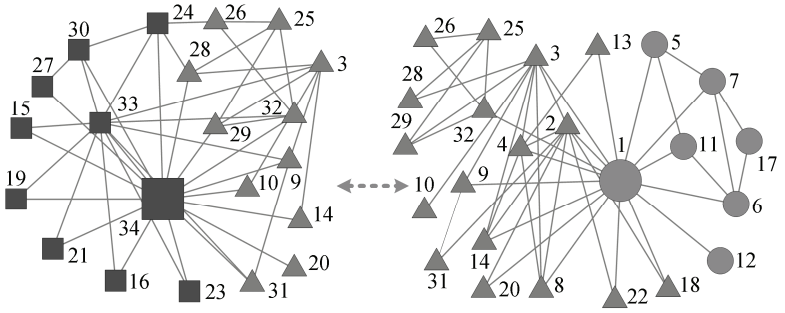

图 4 空手道俱乐部网络上的 2 跳压缩

Fig. 4 Two hops compression on karate club network

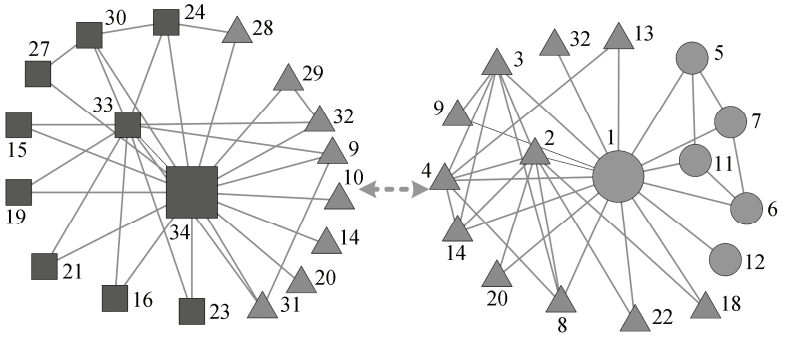

图 5 空手道俱乐部网络上的 1 跳压缩

Fig. 5 One hop compression on karate club network

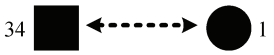

图 6　空手道俱乐部网络上的 0 跳压缩
Fig. 6　Zero hops compression on karate club network

社区的代表点, 三角形图标用于标示社区间的重叠节点。图 9~11 中图标的含义与图 8 相同。

在 GS 算法发现的社区上, 应用 SNC 算法对发现的社区进行 2 跳、1 跳和 0 跳压缩, 压缩结果如图 9~11 所示。图 9~11 中的双向箭头也用于标示两个社区间的关联关系。

4.2　实验分析

经典数据集上的实验表明, 应用基于贪婪策略的社区发现算法 GS 和社会网络压缩算法 SNC 后, 社区中的节点数目可以得到有效的压缩。表 2 列出空手道俱乐部网络和海豚社会网络中各社区在 2 跳、1 跳和 0 跳的压缩率数据(第 2 列中社区的编号与社区代表点的编号保持一致)。本文中的压缩率定义如下。

定义 2　如果网络 $G'=(V',E')$ 是网络 $G=(V,E)$ 的一个网络压缩, 则 $R=\dfrac{|G|-|G'|}{|G|}$ 称为网络 G 的压缩率。

由于两个网络的优化影响范围都为 h=2, 因此可能导致某些社区中的节点恰好无法吸引其他社区中的节点, 从而其压缩率为 0, 如空手道俱乐部网络的社区 C_1 在压缩至 2 跳时的压缩率即属于此种情况。一般情况下, 处于优化社区中的一些节点仍然会有吸引其他社区节点的能力, 因此空手道俱乐部网络中的社区 C_{34} 和海豚社会网络中的社区 C_{14}, C_{17} 和 C_{20} 在压缩至 2 跳时的压缩率都不为 0, 最高的压缩率可达 0.4314。与文献[8]中识别出的社区进行对比, 在压缩至 1 跳时压缩后的社区仍然保持了基本结构或者保留了重要节点, 而此时最大的压缩率最高可 0.75, 最低也为 0.2917。在压缩至 0 跳时, 各个社区的压缩率达到了最高, 都在 0.95 以上。

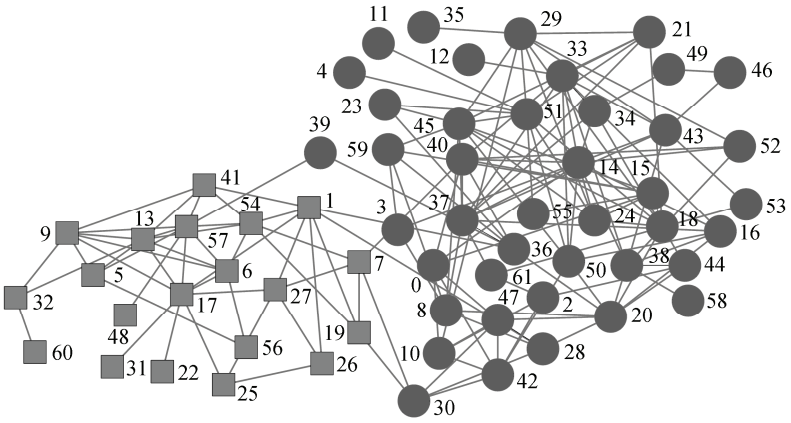

图 7　海豚社会网络
Fig. 7　Dolphin social network

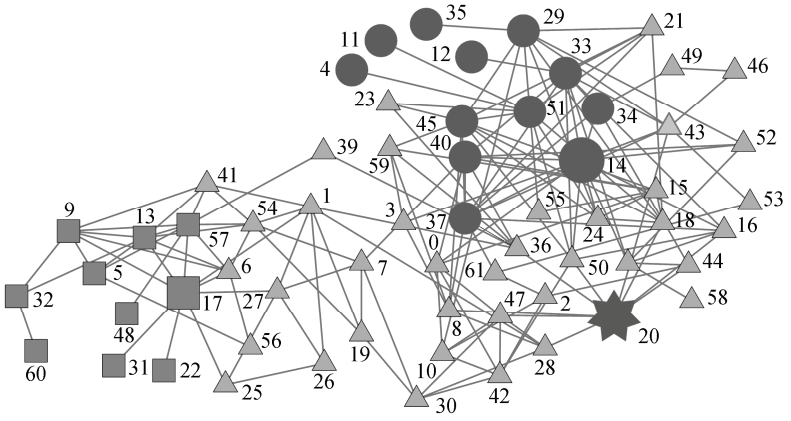

图 8　GS 算法在海豚社会网络上发现的社区
Fig. 8　Communities discovered by GS algorithm on dolphin social network

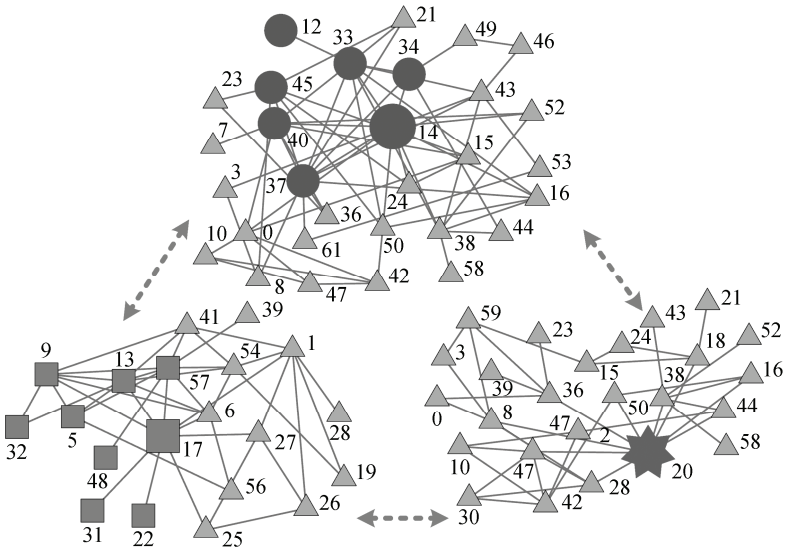

图 9　海豚社会网络上的 2 跳压缩图

Fig. 9　Two hops compression on dolphin social network

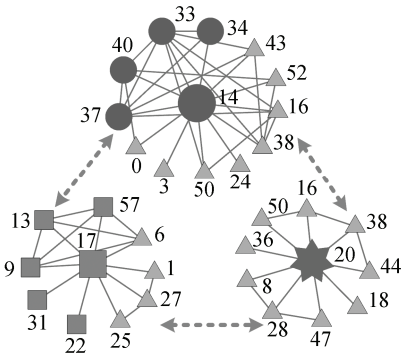

图 10　海豚社会网络上的 1 跳压缩

Fig. 10　One hop compression on dolphin social network

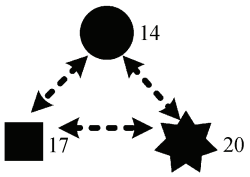

图 11　海豚社会网络上的 0 跳压缩

Fig. 11　Zero hops compression on dolphin social network

5　结语

随着图压缩方法和技术在语义标签网络、网络检索等众多领域的应用越来越广泛，相关研究日益受到关注。针对图压缩方法中存在的时间复杂度较高、依赖先验知识设定参数、需要调节的参数过多、压缩有损以及忽视网络社区结构等问题，本文提出了一种新的压缩方法——基于社区节点重要性的社会网络压缩方法。在提出与拓扑势方法发现的社区中节点重要性相关的定理和推论的基础上，该方法通过基于贪婪策略的 GS 算法进行社区发现，挖掘社区中的不同层次的重要性节点，然后通过社会网络压缩算法 SNC，依据节点的重要性对社区进行压缩。方法的可行性和有效性通过在经典数据集上的实验进行了验证。实验结果表明，该方法不但在压缩过程中可以保持社区间的关联关系，而且具有比较理想的社区压缩率，最高可达 0.95 以上，并且可以在需要时保留社区中的重要节点或社区基本结构。

表 2　社区压缩率数据表

Table 2　Community compression rate list

网络名称	社区名称	社区节点数目	压缩所至跳数		
			2	1	0
空手道俱乐部网络	C_1	24	0	0.2917	0.9583
	C_{34}	27	0.2222	0.3333	0.9630
海豚社会网络	C_{14}	51	0.4314	0.7451	0.9811
	C_{17}	23	0.1304	0.5652	0.9565
	C_{20}	40	0.3750	0.7500	0.9750

参考文献

[1] Tian Y, Hankins R A, Patel J M. Efficient aggregation for graph summarization // Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: Association for Computing Machinery, 2008: 567–579

[2] Hauguel S, Zhai Chengxiang, Han Jiawei. Parallel PathFinder algorithms for mining structures from graphs // 2009 Ninth IEEE International Conference on Data Mining. Miami: Institute of Electrical and Electronics Engineers Inc, 2009: 812–817

[3] Navlakha S, Rastogi R, Shrivastava N. Graph summarization with bounded error // 2008 ACM SIGMOD International Conference on Management of Data. New York: Association for Computing Machinery, 2008: 419–432

[4] Zhang Ning, Tian Yuanyuan, Patel J M. Discovery-driven graph summarization // 26th IEEE International Conference on Data Engineering. Long Beach: IEEE Computer Society, 2010: 880–891

[5] Toivonen H, Zhou Fang, Hartikainen A, et al. Compression of weighted graphs // The 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. San Diego: Association for Computing Machinery, 2011: 965–973

[6] Toivonen H, Mahler S, Zhou F. A frame work for path-oriented network simplification // Advances in Intelligent Data Analysis IX. Berlin: Springer-Verlag, 2010: 220–231

[7] Xu J, Chen H. CrimeNet explorer: a framework for criminal network knowledge discovery. ACM Transactions on Information Systems, 2005, 23(2): 201–226

[8] 淦文燕, 赫南, 李德毅, 等. 一种基于拓扑势的网络社区发现方法. 软件学报, 2009, 20(8): 2241–2254

[9] Girvan M, Newman M E J. Community structure in social and biological networks. Natl Acad Sci, 2002, 99(12): 7821–7826

[10] Fortunato S, Latora V, Marchiori M. Method to find community structures based on information centrality. Physical Review E: Statistical, Nonlinear, and Soft Matter Physics, 2004, 70(5): 056104

[11] Zhang Jianpei, Li Hongbo,Yang Jing, et al. Community discovery method with uncertainty measure of overlapping nodes based on topological potential. Journal of Harbin Institute of Technology: New Series, 2012, 19(2): 16–22

[12] Zachary W W. An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 1977, 33(4): 452–473

[13] Lusseau D, Schneider K, Boisseau O J, et al. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. Behav Ecol Sociobiol, 2003, 54(4): 396–405

[14] Newman M E J. Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74, 2006: 036104

[15] Knuth D E. The Stanford graph base: a platform for combinatorial computing. Massachusetts: Addison-Wesley, 1993

[16] Newman M E J. Network data [EB/OL]. (2011–06–14)[2012–08–14]. http://www-personal.umich.edu/~mejn/netdata/

[17] Albert R, Jeong H, Barabasi A L. Error and attack tolerence of complex networks. Nature, 2000, 406: 378–382

[18] Han Yanni, Li Deyi, Wang Teng. Identifying different community members in complex networks based on topology potential. Frontiers of Computer Science in China, 2011, 5(1): 87–99

[19] Newman M E J, Girvan M. Finding and evaluating community structure in networks. Phys Rev E, 2004, 69: 026113

[20] Pothen A, Simon H D, Liu K P. Partition sparse matrices with eigenvectors of graphs. SIAM Journal of Matrix Analysis and Application, 1990, 11(3): 430–452