Journal of System Simulation

Volume 27 | Issue 1 Article 20

9-2-2020

Community Detection Algorithm in Multi-Relational Networks

Jingping Yu
1. School of Engineering Research Institute, Jiangxi University of Science and Technology, Ganzhou 341000, China;;

Zheng Jie
2. School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China;

Guixiang Zhu
2. School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China;

Community Detection Algorithm in Multi-Relational Networks

Abstract

Abstract: In view of the traditional community detection algorithms being mainly applied to single relational networks, ignoring the interaction of relationship in the multi-relational networks, being unable to distinguish the importance of each relation for community detection, *a novel algorithm called InteractRank was proposed. Based on the node and the relation of ranking model, the algorithm could transform multi-relational network into single relational network. Combined the PageRank algorithm and the random walk model, the algorithm considered the connection within groups and between groups in multi-relational networks. After transforming into single relational networks, spectral clustering algorithm was adopted to detect community.* Through the simulation experiments on the standard UCI dataset, InteractRank indicates to be effective to community detection in multi-relational networks.

Keywords

multi-relational networks, community detection, PageRank, random walk model, spectral clustering

Recommended Citation

Yu Jingping, Zheng Jie, Zhu Guixiang. Community Detection Algorithm in Multi-Relational Networks[J]. Journal of System Simulation, 2015, 27(1): 147-154.

第 27 卷第 1 期 系统仿真学报© Vol. 27 No. 1
2015 年 1 月 Journal of System Simulation Jan., 2015

基于多关系网络的社区检测算法

喻金平[1]，郑杰[2]，朱桂祥[2]
(1. 江西理工大学工程研究院，赣州 341000；2. 江西理工大学信息工程学院，赣州 341000)

摘要：针对传统的社区检测算法主要适用于单关系网络，忽略了多关系网络中各关系间的相互影响，不能区分出各关系对于社区检测的重要性等问题，*提出一种基于节点和关系联合排名模型，能够将多关系合并为单关系的 InteractRank 算法。该算法在多关系网络中结合 PageRank 算法以及随机游走模型的思想，考虑了多关系网络中各关系内和关系间个体的联系。同时，利用谱聚类对 InteractRank 算法得到的单关系网络进行聚类，用于社区检测。*通过 UCI 标准数据集上的仿真实验表明：InteractRank 算法能够在多关系网络进行有效的社区检测。

关键词：多关系网络；社区检测；PageRank；随机游走模型；谱聚类
中图分类号：TP393 文献标识码：A 文章编号：1004-731X (2015) 01-0147-08

Community Detection Algorithm in Multi-Relational Networks

Yu Jingping[1], *Zheng Jie*[2], *Zhu Guixiang*[2]
(1. School of Engineering Research Institute, Jiangxi University of Science and Technology, Ganzhou 341000, China;
2. School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China)

Abstract: In view of the traditional community detection algorithms being mainly applied to single relational networks, ignoring the interaction of relationship in the multi-relational networks, being unable to distinguish the importance of each relation for community detection, *a novel algorithm called InteractRank was proposed. Based on the node and the relation of ranking model, the algorithm could transform multi-relational network into single relational network. Combined the PageRank algorithm and the random walk model, the algorithm considered the connection within groups and between groups in multi-relational networks. After transforming into single relational networks, spectral clustering algorithm was adopted to detect community.* Through the simulation experiments on the standard UCI dataset, InteractRank indicates to be effective to community detection in multi-relational networks.

Keywords: multi-relational networks; community detection; PageRank; random walk model; spectral clustering

引言

复杂网络，是由众多的节点或顶点集通过关系复杂的线或边连接而成的网络。现实世界中有很多这样的复杂网络模型，比如人际关系网、互联网、食物链网以及社交网。它们都有着共同的性质，如“小世界”效应[1]和无标度网络[2]概念的提出。随着近年来对其研究的深入，人们发现许多复杂网络都表现出社团结构的特征，即整个网络由若干个社团或簇构成，社团内部节点的连接比较紧密，而社团间节点的连接比较松散。自 2002 年 Girven 和

作者简介：喻金平(1964-)，男，江西南昌人，硕士，教授，研究方向为数据挖掘；郑杰(1990-)，男，安徽六安人，硕士，研究方向为数据挖掘、复杂网络；朱桂祥(1988-)，男，江苏扬州，硕士，研究方向为社交网络、数据挖掘。
收稿日期：2014-01-11 修回日期：2014-03-03
基金项目：江西省教育厅自然科学基金资助项目(DJJ12346)；江西省研究生创新专项基金资助项目(YC2013-S198)。

Newman 提出基于边介数的 GN 算法[3]以来，社区挖掘便在国际上掀起了一股研究热潮，吸引了来自计算机科学、生物学和社会学等众多领域的研究者，新的理论、算法、应用领域层出不穷。

社区发现的算法有很多，其中大部分算法是针对单关系网络。它们假设网络中各关系之间相互独立，没有考虑到不同关系对网络中节点的影响。然而这一假设并不符合真实网络结构[4]。比如微博网络中存在转发网络、回复网络、复制网络、阅读网络共 4 种关系网络，类似于 PageRank 算法[5]，用户在沿着网络随机游走的同时，也会以一定概率随机跳转到其他节点。网络中用户 A 和 B 在回复网络中会产生相互回复的关系，但是也会以一定概率受对方的影响跳转到转发网络中转发对方的微博[6]。所以，在多关系网络中，各关系之间应该是相互影响的。近年来，对于多关系网络挖掘以及社区检测，吸引了众多学者对其研究。文献[7]采用基于线性回归的多关系抽取方法，将多关系社会网络引入到数据挖掘领域，文中方法在基于用户查询信息的基础上对网络中的多个关系进行优化匹配，从而抽取出一种最优线性组合的关系进行社区挖掘。文献[8]基于结构化特征将用于单关系网络的社区挖掘方法拓展到多关系网络挖掘，在没有先验信息的条件下，使用一种跨关系网络验证的方法评价各种集成单关系网络的方法的性能。文献[9]通过建立具有社区结构的有界信任舆论涌现模型，并采用 Multi-Agent 方法对模型仿真发现，具有社区结构的有界信任基本模型终态与社区间连通性无关而只取决于交互阈值。该研究有利于理解社团结构对舆论事件涌现的影响。文献[10]针对流行病传播动力学问题，提出一种双关系网络模型(工作-朋友关系网)，研究多关系对于流行病传播动力学行为的影响，指出在全接触模式下，多关系的存在会显著降低网络中的爆发阈值，可以使疾病更容易流行而难以控制。这一研究为多关系网络提供了一个新的视角。

在多关系网络日益普遍的大背景下，为了解决传统算法在处理多关系向单关系的转换过程中，可能会存在数据规模增加、潜在社区信息缺失以及数据复制等问题，本文首先提出 InteractRank 算法，该算法将多关系网络合并为单关系网络。然后，本文利用合并后的单关系网络作为输入，采用谱聚类算法对关系网络进行聚类，完成社区检测工作。

1 相关理论简介

1.1 多关系网络与社区检测

当前社交网络的研究对象主要是只有一种关系的社会网络，即同质网络，如 PageRank 算法中认为网页间只存在超链接这样一种关系。然而在现实网络中，实体间总是存在多种多样的关系，不同的关系在不同的情况下扮演不同的角色，每一种关系都形成一个关系网络，这种包含多个关系的网络可以称为多关系网络[11]。具体定义如下：

定义 1：一个包含 m 个关系的多关系网络可以抽象地用一个由点集 V 和边集 E 组成的图组 $G_r=(V, E^{(r)})$表示，G 中的节点表示网络中的对象(个体)，边表示网络中个体之间的联系或者相互作用的连接关系。其中 $r=1, 2, \ldots, m$，$V(G)=\{1, 2, \ldots, n\}$代表含有 n 个元素的节点集合，$E^{(r)}(G)=\{(i, j)|i, j\in V(G_r)\}$是在关系 r 上网络的邻接矩阵。如果节点 i 和 j 之间有边，则 $E^{(r)}_{i,j}=1(i\neq j)$，否则 $E^{(r)}_{i,j}=0(i\neq j)$。

多关系网络的表示不同于单关系网络。传统的单关系网络可以看成是一个有向或无向图，其由点集以及点与点之间的边集共同组成，记为 $G=(V, E)$。其中 V 代表网络中的顶点集，E 为网络中的边集，且 $E\subseteq(V\times V)$。在单关系网络中，虽然 G 中各边可以被赋予不同的权值，但是只能表示一种关系。在多关系网络中，由定义 1 可知其定义为一个图组 $G_r=(V, E^{(r)})$，这里可以使用一个 $n\times n\times m$ 的立方体张量表示[12]，记为 $\boldsymbol{S}=(\boldsymbol{s}_{i,j,r})$。$\boldsymbol{s}_{i,j,r}$ 表示节点 i 和节点 j 在关系 r 上的关系权重值，如果节点 i 能通过关系 r 与节点 j 相连，则 $\boldsymbol{s}_{i,j,r}\neq 0$，反之，$\boldsymbol{s}_{i,j,r}=0$。张量 $\boldsymbol{S}$ 可以看作是由 m 种不同关系的单关系网络合成的多关系网络。

社区检测的目的是为了寻求一个较好的 p 分割，p={C_1，C_2，…，C_k}，且 $C_1 \cup C_2 \cup \ldots \cup C_k \subseteq V$。其中 C_k 是第 k 个社区划分，k 是要检测的社区个数。

社区检测的算法大致可以分为两类：一类是通常采用自底向上的策略发现社区。这类算法首先从一个节点或是一组节点开始，通过 Jaccord 系数或余弦相似度准则在整个网络搜索合并具有相同特征的实体，找到有意义的社区，同时可以有效的避免网络中存在弱连接节点的影响，但其对于大数据的处理执行效率不高。另一类是基于全局模型的算法如混合模型、K-means 模型、谱聚类。这些算法基于对网络拓扑结构的全局考虑，通过启发式过程在网络中迭代优化得到一个较优的网络划分从而发现社区。该类算法不能很好的区分网络中的弱连接的点。

1.2 随机游走模型与 PageRank 算法

定义 2：(随机过程)设(Ω，F，P)为一概率空间，T 为一参数集，R 为实数域，$T \subset R$，若对每一个 $t \in T$，均有定义在(Ω，F，P)上的一个随机变量 $X(\omega, t)$，($\omega \in \Omega$)与之对应，则称 $X(\omega, t)$为(Ω，F，P)上的一个随机过程。

定义 3：(马尔可夫链)对取值于有限状态空间 E 的随机过程{X_n，n=1，2，…}，若随机变量下一时刻的状态只与当前状态有关，即如公式(1)所示：

$$P(X_{n+1}=j \mid X_n=i_n, X_{n-1}=i_{n-1}, \cdots, X_1=i_1, X_0=i_0) = P(X_{n+1}=j \mid X_n=i_n) \quad (1)$$

其中：j，i_0，i_1，…，i_{n-1} 为相对应时刻的状态，这样的随机过程称为马尔可夫(Markov)链。

给定图 G=(V，E)中，随机游走模型的基本思想是，假设一个游走者从 v_0 点开始遍历此图，游走者将以概率 1-γ 游走到 v_0 点的近邻顶点，以概率 γ 随机跳跃到图中任一顶点，称 γ 为跳转发生概率[13]。每次游走所得到的概率分布都反应了图 G 中每个顶点被访问的概率。当时间 $t \to \infty$时，图上的随机游走的概率分布会收敛到一个极限概率分布，这个极限称为平稳分布。

随机游走的模型应用广泛，PageRank 算法就是其典型应用之一。在单关系网络中，PageRank 算法是运用随机游走模型计算网络中的所有节点和关系的权值，得到节点和关系的联合排名。本文提出的 InteractRank 算法是将 PageRank 的思想拓展到多关系网络中，在随机游走过程中迭代得到平稳分布的节点和关系的联合排名值。在迭代过程中利用公式(1)使本文算法具有记忆性，根据上次的迭代结果判断计算出当前迭代的节点和关系联合排名，使得排名值更准确，从而有效的将多关系网络合并为单关系网络，并用于社区检测，得出各关系在社区检测中的重要性比例。

2 InteractRank 算法

2.1 InteractRank 的算法简介

InteractRank 算法以张量 $\boldsymbol{S}$ 作为输入，迭代输出节点和关系的概率分布。设 $\boldsymbol{p}$=(p_1，p_2，…，p_n)和 $\boldsymbol{q}$=(q_1，q_2，…，q_m)分别代表多关系网络中节点和关系的概率分布向量，X_t 和 Y_t 分别表示在时刻 t 访问某节点和某边的随机变量。根据 Markov 链有，对于给定节点 j 的近邻节点 i 到达节点 j 的跳转概率如公式(2)所示：

$$Prob_n(X_t=j \mid X_{t-1}=i) = Prob_n(j \mid i) = \frac{\sum_d \boldsymbol{q}_d \cdot \boldsymbol{s}_{i,j,r}}{\sum_l \sum_r \boldsymbol{q}_r \cdot \boldsymbol{s}_{i,l,r}} \quad (2)$$

公式(2)表示从节点 i 到节点 j 的转移概率，与 PageRank 不同，在多关系网络中，节点的权重不仅与其近邻节点相关，也受具有不同权重的关系的影响。节点 i 选择关系 r 的跳转概率如公式(3)所示：

$$Prob_d(Y_t=r \mid X_{t-1}=i) = Prob_d(r \mid i) = \frac{\boldsymbol{q}_r \cdot \sum_j \boldsymbol{s}_{i,j,r}}{\sum_r \boldsymbol{q}_r \sum_j \boldsymbol{s}_{i,j,r}} \quad (3)$$

从公式(2)，(3)可以看出，与节点 i 相邻的其他近邻节点以及具有不同权重的关系都会影响节点 i 在多关系网络中的权重。通过公式(2)，(3)可以体现并验证多关系网络中节点和关系间是交互影响的。根据 Markov 链的平稳分布的性质可知，在多关系网络中，类似于 PageRank 中的随机游走

过程，当时间 t 趋于无限大时，$\boldsymbol{p}$ 和 $\boldsymbol{q}$ 会收敛到一个极限概率分布，如公式(4)所示：

$$\boldsymbol{p}_j = \lim_{t\to\infty} Prob(X_t = j), \boldsymbol{q}_r = \lim_{t\to\infty} Prob(Y_t = r) \tag{4}$$

在多关系网络中，由于 $\boldsymbol{p}_i$ 和 $\boldsymbol{q}_d$ 是耦合的，所以很难得到 $\boldsymbol{p}_i$ 和 $\boldsymbol{q}_d$ 以及它们的联合概率分布 $Prob(X_t = j, X_{t-1} = i)$ 和 $Prob(Y_t = r, X_{t-1} = i)$，传统的算法仅仅假设点概率分布和关系概率分布是相互独立的，可以表达为如公式(5)，(6)所示：

$$Prob(X_t = j, X_{t-1} = i) = Prob(X_t = j)\,Prob(X_{t-1} = i) \tag{5}$$

$$Prob(Y_t = r, X_{t-1} = i) = Prob(Y_t = r)\,Prob(X_{t-1} = i) \tag{6}$$

从公式(5)，(6)可以看出，节点选择关系的概率与整个网络的相似结构无关，这在多关系网络是不合理的。举例来说，假设关系 k 在之前的跳转迭代中获得了很高的权值，其中节点 i 与其近邻节点的相似度很低，在这种情况下，传统算法仍假定每个节点跳转到关系 k 时的概率是相等的，即不考虑前次迭代跳转的影响，对于与近邻节点相似度低的节点 i 以及其他节点没有区别处理，这样会导致信息缺失，不能准确反映出多关系网络中的个体差异，影响后面的社区检测的准确性。所以，在 InteractRank 算法中，需要考虑与近邻节点拥有较低相似度的节点 i 在跳转到关系 k 时的概率应该较小，即拥有不同的相似度的节点的跳转概率也不同。这里，InteractRank 的两个联合概率分布模型重新定义如公式(7)，(8)所示：

$$Prob(X_t = j, X_{t-1} = i) = Prob(X_{t-1} = i)\cdot Prob(X_t = j \mid X_{t-1} = i) \tag{7}$$

$$Prob(Y_t = r, X_{t-1} = i) = Prob(X_{t-1} = i)\cdot Prob(Y_t = r \mid X_{t-1} = i) \tag{8}$$

InteractRank 算法还考虑了节点和关系的先验概率对得到 $\boldsymbol{p}$ 和 $\boldsymbol{q}$ 的影响，从而 InteractRank 提出公式(9)，(10)用来迭代计算节点和关系的联合排序的权值：

$$\boldsymbol{p}_j^{t+1} = \alpha \sum\nolimits_i \boldsymbol{p}_j^t \cdot Prob_n^t(j \mid i) + (1-\alpha)\cdot \boldsymbol{p}_j^* \tag{9}$$

$$\boldsymbol{q}_r^{t+1} = \beta \sum\nolimits_i \boldsymbol{p}_j^t \cdot Prob_d^t(r \mid i) + (1-\beta)\cdot \boldsymbol{q}_r^* \tag{10}$$

其中：α，$\beta \in [0, 1]$是用来调整网络结构和先验概率影响比例的影响因子。$\boldsymbol{p}^* = (p_1^*, p_2^*, \cdots, p_n^*)$ 和 $\boldsymbol{q}^* = (q_1^*, q_2^*, \cdots, q_m^*)$ 分别表示节点和关系的先验分布向量。为了能够处理如下情况，给定节点 j，有 $\boldsymbol{s}_{i,j,r} = 0$，从而有 $\text{Prob}_n(j|i) = \text{Prob}_d(r|i) = 0$，此时该节点则无法完成迭代[5]，就会产生信息遗漏，无法涵盖整个网络结构信息，本文在 InteractRank 算法中加入节点和关系的先验概率避免该情况的出现以提高 InteractRank 算法的健壮性。

2.2 InteractRank 算法流程

结合前面对 InteractRank 算法的介绍可知，对于给定节点 i，结合随机游走模型，可以通过公示(2)计算其跳转到节点 j 的概率，通过公式(3)可以计算其跳转到关系 r 的概率。InteractRank 算法主要通过公式(9)，(10)迭代计算新的 $\boldsymbol{p}^{t+1}$，$\boldsymbol{q}^{t+1}$，直到算法收敛，得到节点和关系的平稳分布。InteractRank 算法的流程如下所示：

输入：$\boldsymbol{S}$：关系张量

ε：聚合收敛阈值

$\boldsymbol{p}^*$，$\boldsymbol{q}^*$：节点和关系的先验分布

α，β：两个平衡影响因子

输出：$\boldsymbol{p}$，$\boldsymbol{q}$：节点和关系聚合时的平衡分布

步骤 1：设置迭代次数 t=1，并随机初始化 $\boldsymbol{p}^t$ 和 $\boldsymbol{q}^t$，且 $\sum_{i=1}^{n} \boldsymbol{p}_i = 1$，$\sum_{q=1}^{m} \boldsymbol{q}_r = 1$；

步骤 2：利用公式(8)、(9)计算$||\boldsymbol{p}^t-\boldsymbol{p}^{t-1}||+||\boldsymbol{q}^t-\boldsymbol{q}^{t-1}||$的值，$t$=$t$+1；

步骤 3：如果$||\boldsymbol{p}^t-\boldsymbol{p}^{t-1}||+||\boldsymbol{q}^t-\boldsymbol{q}^{t-1}||>\varepsilon$，跳到步骤 2，否则，算法结束，输出 $\boldsymbol{p}$ 和 $\boldsymbol{q}$ 的值。

通过 InteractRank 算法我们可以得到节点和关系的平稳分布，在考虑点和关系相互影响相互作用的前提下，将多关系网络合并为单关系网络。

从 InteractRank 算法流程可以看出，算法是通过迭代计算得到节点和关系聚合时的平衡分布。InteractRank 算法中时间复杂度最高的步骤为步骤 2，因为向量 $\boldsymbol{p}$ 和 $\boldsymbol{q}$ 都是稀疏矩阵，假设两个稀疏

矩阵中非零元素的最大维度为 N，则在计算法 $\|\boldsymbol{p}^t-\boldsymbol{p}^{t-1}\|+\|\boldsymbol{q}^t-\boldsymbol{q}^{t-1}\|$的时间复杂度为 $O(N^2)$，算法在未满足步骤 3 中$\|\boldsymbol{p}^t-\boldsymbol{p}^{t-1}\|+\|\boldsymbol{q}^t-\boldsymbol{q}^{t-1}\|>\varepsilon$ 条件时，会一直迭代执行 t 次，所以 InteractRank 算法的时间复杂度为 $O(tN^2)$。

3 实验结果及其分析

3.1 InteractRank 算法验证

为了体现算法的有效性，鉴于真实世界中没有提供有关成员关系的真实信息，所以，这里我们采用 UCI 数据中两个合成的数据集来验证 InteractRnak 算法的性能。两个数据集分别是 Iris 数据集和 Breast 数据集，2 个数据集中的每一个实体代表一个节点，每一个属性代表一个关系，它们的特征如表 1 所示：

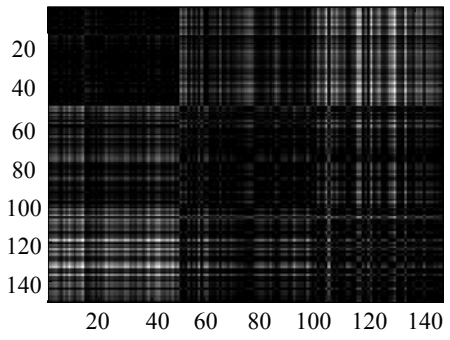
表 1 Iris，Breast 数据集

数据集名称	样本个数	属性个数	社区个数
Iris	150	4	3
Breast	699	9	2

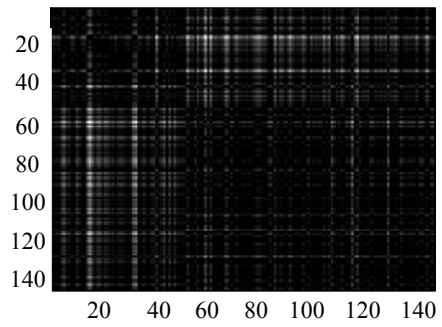
为了简便起见，这里设 $\alpha=\beta=0.5$，$\boldsymbol{p}_i^*=1/n(1\leqslant i\leqslant n)$，$\boldsymbol{q}_d^*=1/m(1\leqslant d\leqslant m)$，$\varepsilon$=1×10e-4。并定义 $\boldsymbol{p}_{i,r}$ 和 $\boldsymbol{p}_{j,r}$ 分别代表在关系 r 上节点 i, j 的值。利用公式(11)计算每个节点在各个关系上的相似度：

$$\boldsymbol{s}_{i,j,r}=1-\frac{r_{i,j}}{1+r_{max}},\quad r_{i,j}=(\boldsymbol{p}_{i,r}-\boldsymbol{p}_{j,r})^2 \qquad (11)$$

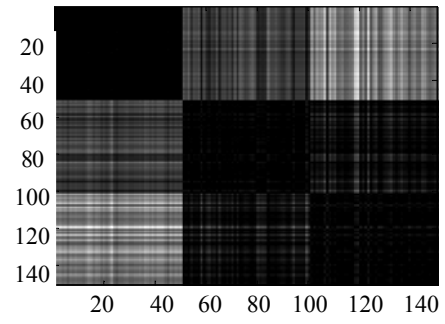
通过公式(11)可以得到在每个关系 r 上的张量矩阵 ***Tensor_r***，***Tensor_r*** 是一个对称矩阵，矩阵里每个元素分别代表 $\boldsymbol{p}_{i,r}$ 和 $\boldsymbol{p}_{j,r}$ 在关系 r 上的空间相似度。实验结果中，用图形的明亮度来表示 $\boldsymbol{p}_{i,r}$ 和 $\boldsymbol{p}_{j,r}$ 的相似程度，方块区域内亮度越高，表明社区内节点之间相似度越高，即明亮方块区域可有可能就是潜在的社区，方块区域间亮度差越大表明社区间划分越准确，这也符合社区结构中“社区内部连接紧密，社区之间连接稀疏”的原则。此外，从各关系明亮程度可以看出其对社区检测的权重比例。其中图 1~2 展示了 Iris 数据集的 4 个关系的张量矩阵和 Breast 数据集的 9 个关系的张量矩阵对于社区检测的影响程度。



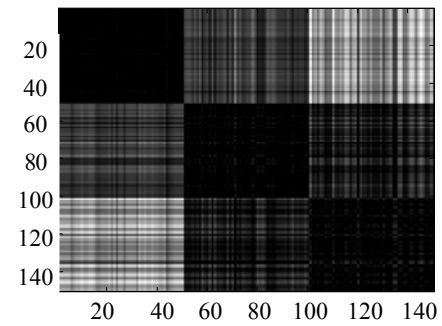
(a) 关系 1



(b) 关系 2

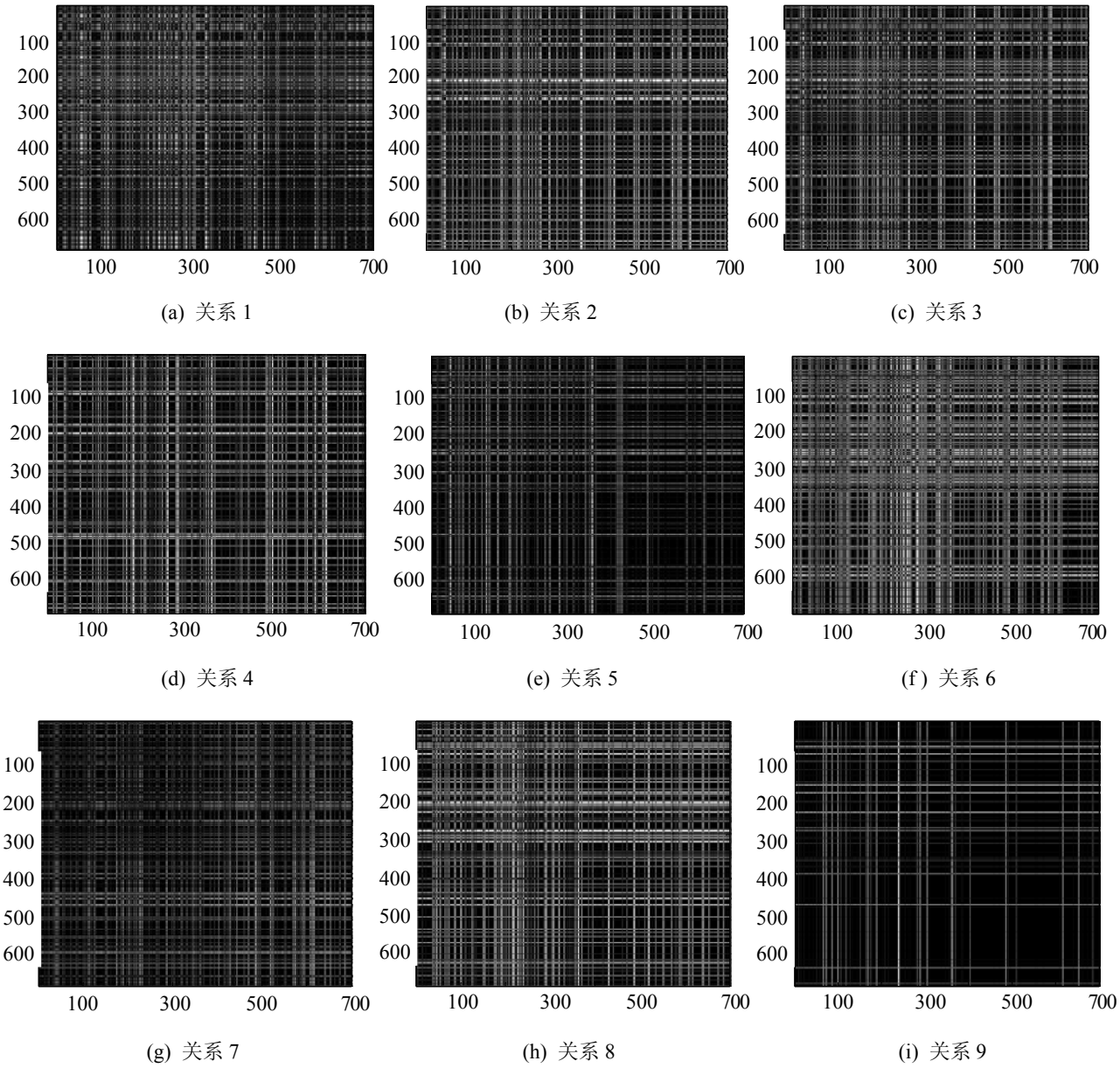


(c) 关系 3



(d) 关系 4

图 1 在 Iris 数据集上的 4 个关的密切度矩阵



(a) 关系 1 (b) 关系 2 (c) 关系 3

(d) 关系 4 (e) 关系 5 (f) 关系 6

(g) 关系 7 (h) 关系 8 (i) 关系 9

图 2 Breast 数据集上 9 个关系的密切度矩阵

从图 1 可以看出，关系 3 和 4 包含了更多明亮的方块区域，且各方块区域间明亮对比差别较大，而关系 1 和 2 则相对较小，所以，关系 3 和 4 在社区检测上相比于关系 1 和 2 具有更重要的意义，也占有更大的权重。同理，由图 2 可知，关系 6 中呈现的明亮的方块区域比其他 8 个关系的都要明显，可以发现关系 5，7，9 中明亮区域并不明显，也就是说，关系 6 在社区检测中相较其他 8 个关系更具有意义和区分度。下面本文分别对两个数据集应用 InteractRank 算法计算各关系的平衡分布，得到结果如图 3 所示。

从图 3 可以看出经过 InteractRank 算法得到的关系平衡分布与图 1 和图 2 中反映的密切度矩阵是一致的，这证明了 InteractRank 算法的合理性和有效性。

3.2 谱聚类对于社区检测的验证

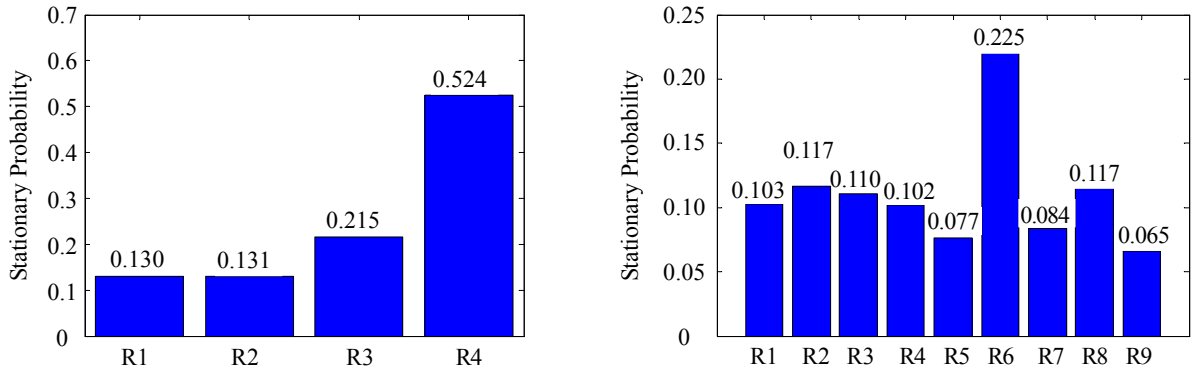
社区检测问题可以转化为聚类问题，而谱聚类就是一个很好的方法。谱聚类的思想是将样本看作顶点，样本间的相似度看作带权值的边，从而将聚

类问题转换为图分割问题，找到一种图分割的方法使得连接不同组的边的权重尽可能低，即组间相似度尽可能低，组内边的权重尽可能高，即组内相似度尽可能高，这与社区检测的目标是一致的。这里，将谱聚类与 K-means 聚类算法进行对比，显示谱聚类在社区检测中的优越性。

为评估算法检测出的社区，使用标准化的信息索引值(Normalized Mutual Information， NMI)作为评判标准，NMI 定义如公式(11)所示：

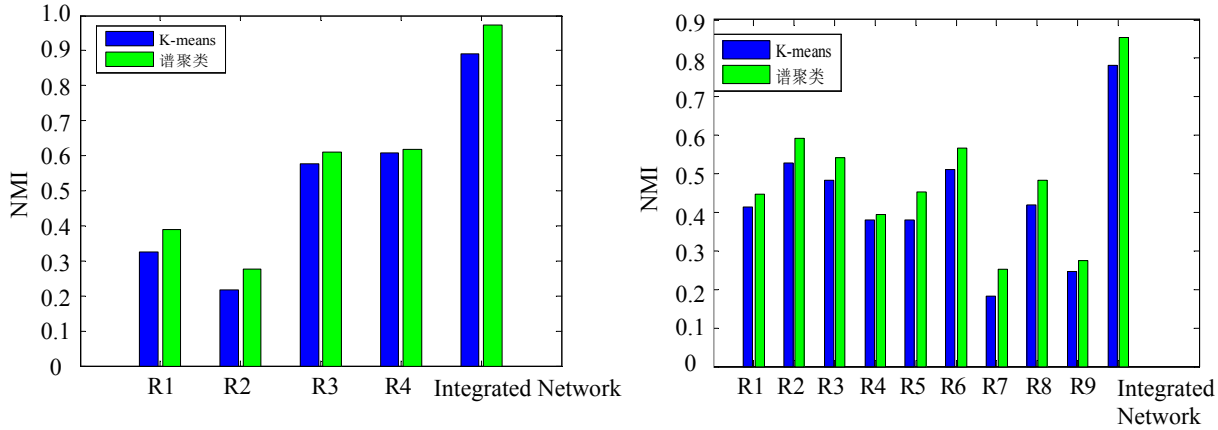
$$NMI(L;G)=\frac{I(L;G)}{\sqrt{H(L)H(G)}} \tag{11}$$

其中：L 和 G 分别代表通过社区检测方法得到的网络划分和真实的网络社区划分。$I(L;G)$代表 L 和 G 的互信息。图 4 展示了 InteractRank 算法能够将一个多关系网络合并为一个单关系网络，合并的单关系网络在社区检测划分上比原始网络结构中任一单关系网络得到的社区划分都要清晰，并且可以发现谱聚类的社区检测性能优于 K-means 聚类算法。



(a) Iris 各关系的平衡分布　　(b) Breast 各关系的平衡分布

图 3　Iris，Breast 各关系的平衡分布



(a) Iris 数据集上合并后单关系的性能比较　　(b) Breast 数据集上合并后单关系的性能比较

图 4　Iris，Breast 数据集上合并后单关系的性能比较

综上所述可以看出，通过 InteractRank 算法迭代得到的单关系网络在社区检测上具有更高的准确率，检测到的社区符合社区内个体联系紧密，社区间联系稀疏。从图 1，2，4 可以看出，合并后的单关系在社区检测方面比仅仅使用某一单关系用于社区发现的准确率都要高。并且可以得到多关系网络中各个单关系对于整个社区检测的重要性。同时可以发现，谱聚类在社区发现的聚类问题上的准

确性要优于 K-means 算法，能够得到质量更高、划分准确的社区。

4 结论

鉴于传统社区检测算法都是假定各关系之间是相互独立的，且对于社区检测的重要性相同，从而无法反映出社区的潜在信息和社区间的重要关系。本文提出一种 InteractRank 算法，综合考虑节点和关系之间的相互关系，通过迭代计算得到节点和关系的平稳分布，实现了从多关系到单关系的融合以及得出不同关系对于社区检测具有不同的重要性的结论。并通过实验结果表明合并后的单关系能够很好地用于社区检测。但是，该算法也有其自身的局限性和不足，在跳转概率中一些参数还是需要依靠经验值进行设定，另外使用原始谱聚类进行社区检测的准确性还有待提高。如何获取到 InteractRank 算法中各参数的准确值以及结合其他改进的聚类算法用于社区检测，提高社区检测的准确度，将是本文下一步的研究内容。

参考文献：

[1] Watts D J, Strogatz S H. Collective dynamics of ‘small-world’ networks [J]. Nature (S0028-0836), 1998, 393(6684): 440-442.

[2] Barabási A L, Albert R. Emergence of scaling in random networks [J]. Science (S0036-8075), 1999, 286(5439): 509-512.

[3] Girvan M, Newman M E J. Community structure in social and biological networks [J]. Proceedings of the National Academy of Sciences (S0036-8075), 2002, 99(12): 7821-7826.

[4] Dai B T, Chua F C T, Lim E P, *et al*. Structural Analysis in Multi-Relational Social Networks [C]// SDM. USA: SIAM, 2012: 451-462.

[5] Page L, Brin S, Motwani R, *et al*. The PageRank citation ranking: bringing order to the web [M]. USA: Stanford InfoLab., 1999.

[6] 丁兆云, 周斌, 贾焰, 等. 微博中基于多关系网络的话题层次影响力分析 [J]. 计算机研究与发展, 2013, 50(10): 2155-2175. (Ding Zhaoyun, Zhou Bin, Jia Yan, *et al*. Topical Influence Analysis Based on the Multi-Relational Network in Microblogs [J]. Journal of Computer Research and Development, 2013, 50(10): 2155-2175.)

[7] Cai D, Shao Z, He X, *et al*. Mining hidden community in heterogeneous social networks [C]// Proceedings of the 3rd International Workshop on Link Discovery. USA: ACM, 2005: 58-65.

[8] Tang L, Wang X, Liu H. Community detection via heterogeneous interaction analysis [J]. Data Mining and Knowledge Discovery (S1384-5810), 2012, 25(1): 1-33.

[9] 司夏萌, 刘云, 丁飞, 等. 具有社团结构的有界信任舆论涌现模型研究 [J]. 系统仿真学报, 2009, 21(23): 7644-7647. (Si Xiameng, Liu Yun, Ding Fei, *et al*. Research on Bounded Confidence Consensus Emergency Model with Community Structure [J]. Journal of System Simulation (S1004-731X), 2009, 21(23): 7644-7647.)

[10] 李睿琪, 唐明, 许伯铭. 多关系网络上的流行病传播动力学研究 [J]. 物理学报, 2013, 62(16): 168903. (Li Ruiqi, Tang Ming, Hui Pakming. Epidemic spreading on multi-relational networks [J] Acta Phys. Sin, 2013, 62(16): 168903.)

[11] Berlingerio M, Coscia M, Giannotti F, *et al*. Multidimensional networks: foundations of structural analysis [J]. World Wide Web (S1386-145X), 2013, 16(5-6): 567-593

[12] Rodriguez M A, Shinavier J. Exposing multi-relational networks to single-relational network analysis algorithms [J]. Journal of Informetrics (S1751-1577), 2010, 4(1): 29-41.

[13] 郑伟, 王朝坤, 刘璋, 等. 一种基于随机游走模型的多标签分类算法 [J]. 计算机学报, 2010, 33(8): 1418-1426. (Zheng Wei, Wang Chaokun, Liu Zhang, *et al*. A Multi-Label Classification Algorithm Based on Random Walk Model [J]. Chinese Journal of Computers, 2010, 33(8): 1418-1426.)