一种基于嵌入技术的异构信息网络的快速聚类算法

陈丽敏*①② 杨 静① 张健沛①
①(哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001)
②(牡丹江师范学院计算机系 牡丹江 157012)

摘 要：异构信息网络聚类分析是当前的热点研究问题之一。利用异构信息网络的稀疏性，该文提出一种基于嵌入技术的星型模式的异构信息网络的快速聚类算法。首先从相容的角度将异构信息网络转化为若干个相容的二部图，使用随机映射和一种线性时间求解程序快速计算出每个二部图的近似通勤距离嵌入，每个嵌入都存在一个子集指示目标数据集；然后，使用这些指示子集构建一个通用的聚类模型；最后，将所有指示子集的类设置标号，通过计算指示同一目标对象的指示数据与标号相同类的中心点的加权距离总和，同时划分所有的指示子集，从而快速获得通用模型的极小值。通过理论分析及实验验证，该文算法聚类速度快，聚类准确率高。

关键词：异构信息网络；聚类；通勤距离；嵌入；加权距离总和
中图分类号： TP311 **文献标识码**： A **文章编号**：1009-5896(2015)11-2634-08
DOI: 10.11999/JEIT150106

A Fast Clustering Algorithm Based on Embedding Technology for Heterogeneous Information Networks

Chen Li-min①② Yang Jing① Zhang Jian-pei①
①(*Institute of Computer Science and Technology, Harbin Engineering University, Harbin* 150001, *China*)
②(*Institute of Computer Science and Technology, Mudanjiang Normal University, Mudanjiang* 157012, *China*)

Abstract: Research on clustering heterogeneous information networks is one of the current hotspots. Taking advantages of the sparsity of heterogeneous information networks, a fast clustering algorithm based on embedding technology for heterogeneous information networks of star network schema is proposed in this paper. First, the heterogeneous information network is transformed into some compatible bipartite graphs from the point of compatible view. Then, the approximate commute distance embedding of each bipartite graph is computed via random mapping and a linear time solver, and an indicator subset in each embedding indicates the target dataset. At last, a general model is formulated via all the indicator subsets, and a minimum value of the model is derived by simultaneously clustering all of the indicator subsets using the sum of the weighted distances for all indicators for an identical target object. This proposed algorithm is effective by theory analysis and experimental verification.

Key words: Heterogeneous information network; Clustering; Commute distance; Embedding; Sum of weighted distances

1 引言

信息网络普遍存在，如社会信息网络、DBLP书目网络等。同构信息网络是由一种类型数据构成的，异构信息网络则是由多种类型数据构成的。目前，同构信息网络的聚类研究已经非常丰富[1,2]，但异构信息网络的聚类研究还不多，而异构信息网络聚类分析能更好地理解网络的隐藏结构以及每个类中的数据所代表的角色[3,4]。

异构信息网络中，星型网络模式非常流行，也非常重要。星型网络模式由一个目标对象数据集和多个属性对象数据集构成，关系只存在于目标对象与属性对象之间，属性对象之间不存在关系。

基于相容二部图的思想解决多种异构数据之间的复杂关系是行之有效的，从相容的角度出发，人们先后设计了多种算法解决了多种异构数据协同聚类的问题，其中比较经典的算法有基于半正定的规划算法[5]，基于信息论的算法[6]以及谱聚类算法[7]。这类算法比较通用，但对于异构信息网络而言这些算法的复杂度太高。

基于概率模型的异构信息网络聚类算法NetClus[8]虽然聚类效率较高，但不具有通用性，而且算法的收敛性也不是很稳定，基于概率模型的思想还被用于网络服务聚类[9,10]。ComClus[11]算法是NetClus 的衍生算法，面向包含同构关系与异构关

收稿日期：2015-01-21；改回日期：2015-07-16；网络出版：2015-08-25
*通信作者：陈丽敏 chenlimin_clm@126.com
基金项目：国家自然科学基金(61370083, 61073043, 61073041)；高等学校博士学科点专项科研基金(20112304110011, 20122304110012)
Foundation Items: The National Natural Science Foundation of China (61370083, 61073043, 61073041); The National Research Foundation for the Doctoral Program of Higher Education of China (20112304110011, 20122304110012)

系的信息网络，仍然离不开具体的应用领域，不具有通用性。基于密度的异构网络的子空间聚类算法计算速度较慢[12]，使用语义路径相似性与传统算法结合的异构网络聚类[13,14]也与具体应用领域相关。

异构网络链接推理[15]时，往往需要最初的聚类划分更精确一些，因此高准确率的聚类是十分必要的，但是当网络的数据量非常大时，计算速度又不能太慢。文献[16]在同构数据上使用近似通勤距离嵌入聚类取得了非常好的效果。星型网络模式的异构信息网络能够转换成若干个相容的二部图，使用通勤距离度量二部图各结点之间的关系，能够提高聚类准确率。异构信息网络规模大，但是很稀疏，所以可以使用随机映射和线性时间求解程序[17,18]快速计算出每个二部图的近似通勤距离嵌入。每个嵌入都存在一个子集指示目标数据集，使用这些指示子集构建一个通用的聚类模型。然后将所有指示子集的类设置标号，通过计算每个目标对象的所有指示数据与标号相同的类的中心点的加权距离总和，同时划分所有的指示子集，从而快速获得该通用模型的极小值。本文算法是一个通用的异构信息网络聚类算法，计算速度快，聚类准确率高。

2　二部图的近似通勤距离嵌入

2.1 二部图的通勤距离嵌入

给定 $G_b=\langle V,E\rangle, V(G_b)=X_0\cup X_1$，其中 X_0 与 X_1 为两个不同类型的数据集，若 $E(G_b)=\{<x_i, x_j>\}$，则 $x_i\in X_0, x_j\in X_1$，称 G_b 为二部图。设 $X_0=\{x_1^{(0)},x_2^{(0)},\cdots,x_{n_0}^{(0)}\}$, $X_1=\{x_1^{(1)},x_2^{(1)},\cdots,x_{n_1}^{(1)}\}$，则 G_b 有 $n=n_0+n_1$ 个结点。$\boldsymbol{W}_{n_0\times n_1}$ 为 X_0 与 X_1 的关系矩阵，其中元素 w_{ij} 表示 $<x_i^{(0)},x_j^{(1)}>$ 的权重。由 $\boldsymbol{W}_{n_0\times n_1}$ 可计算 G_b 的 Laplace 矩阵 $\boldsymbol{L}$。设 $\boldsymbol{L}^+$ 是 $\boldsymbol{L}$ 的伪逆矩阵，由文献[19]可知，G_b 的任意两个结点 i,j 之间的通勤距离可通过 $\boldsymbol{L}^+$ 计算。

性质 1　$c_{ij}=g_v(l_{ii}^+ + l_{jj}^+ - 2l_{ij}^+)=g_v(\boldsymbol{e}_i-\boldsymbol{e}_j)^{\mathrm{T}}L^+(\boldsymbol{e}_i-\boldsymbol{e}_j)$。其中 l_{ij}^+ 是伪逆矩阵 $\boldsymbol{L}^+$ 的第 (i,j) 个元素，g_v 是二部图 G_b 的权重总和，即 $g_v=\sum w_{ij}$，$\boldsymbol{e}_i$ 是第 i 个元素为 1 的单位列向量，即

$$\boldsymbol{e}_i=\left[\underset{1}{0},\cdots,\underset{i-1}{0},\underset{i}{1},\underset{i+1}{0},\cdots,\underset{n}{0}\right]^{\mathrm{T}}$$

设二部图 G_b 有 s 条边，根据文献[20]，定向 G_b 的边，令

$$B(i,j)=\begin{cases}1, & i\text{是尾 }j\text{是头}\\ -1, & i\text{是头 }j\text{是尾}\\ 0, & \text{其他}\end{cases}$$

其中 i,j 是 G_b 的任意两个结点，则 $\boldsymbol{B}_{s\times n}$ 是一个有向边-点入射矩阵。设 $\widehat{\boldsymbol{W}}_{s\times s}$ 是由边的权值构成的对角矩阵，则 G_b 的 Laplace 矩阵可表示为 $\boldsymbol{L}=\boldsymbol{B}^{\mathrm{T}}\widehat{\boldsymbol{W}}\boldsymbol{B}$。

由性质 1，G_b 的任意两个结点 i,j 的通勤距离 c_{ij} 为

$$\begin{aligned}c_{ij}&=g_v(\boldsymbol{e}_i-\boldsymbol{e}_j)^{\mathrm{T}}\boldsymbol{L}^+(\boldsymbol{e}_i-\boldsymbol{e}_j)\\&=g_v(\boldsymbol{e}_i-\boldsymbol{e}_j)^{\mathrm{T}}\boldsymbol{L}^+\boldsymbol{L}\boldsymbol{L}^+(\boldsymbol{e}_i-\boldsymbol{e}_j)\\&=g_v(\boldsymbol{e}_i-\boldsymbol{e}_j)^{\mathrm{T}}\boldsymbol{L}^+\boldsymbol{B}^{\mathrm{T}}\widehat{\boldsymbol{W}}\boldsymbol{B}\boldsymbol{L}^+(\boldsymbol{e}_i-\boldsymbol{e}_j)\\&=\left[\sqrt{g_v}\widehat{\boldsymbol{W}}^{1/2}\boldsymbol{B}\boldsymbol{L}^+\left(\boldsymbol{e}_i-\boldsymbol{e}_j\right)\right]^{\mathrm{T}}\\&\quad\cdot\left[\sqrt{g_v}\widehat{\boldsymbol{W}}^{1/2}\boldsymbol{B}\boldsymbol{L}^+\left(\boldsymbol{e}_i-\boldsymbol{e}_j\right)\right]\end{aligned}$$

则 c_{ij} 是空间 $\boldsymbol{\psi}=\sqrt{g_v}\widehat{\boldsymbol{W}}^{1/2}\boldsymbol{B}\boldsymbol{L}^+\in\boldsymbol{R}^{s\times n}$ 第 i 个列向量与第 j 个列向量的欧式距离的平方，称 $\boldsymbol{\psi}$ 为二部图 G_b 的通勤距离嵌入。c_{ij} 是 G_b 的两个结点间的平均路径长度，而不是两结点间的最短路径，因此使用通勤距离度量结点间的关系，能够捕获复杂的类，聚类有噪音的数据，鲁棒性更好。那么，使用通勤距离嵌入聚类也能够捕获复杂的类。

2.2 二部图的近似通勤距离嵌入

引理 1[21]　给定向量 $\boldsymbol{V}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n]\in\boldsymbol{R}^s$ 和 $\varepsilon>0$，$\boldsymbol{Q}_{k_r\times s}$ 是行向量独立同分布的随机矩阵，其中 $Q(i,j)=\pm1/\sqrt{k_r}$ 等概率，$k_r=O(\log n/\varepsilon^2)$。则 $\boldsymbol{v}_i,\boldsymbol{v}_j\ \forall\in\boldsymbol{V}$，至少存在 $1-1/n$ 概率满足

$$\begin{aligned}(1-\varepsilon)\left\|\boldsymbol{v}_i-\boldsymbol{v}_j\right\|^2&\le\left\|\boldsymbol{Q}\boldsymbol{v}_i-\boldsymbol{Q}\boldsymbol{v}_j\right\|^2\\&\le(1+\varepsilon)\left\|\boldsymbol{v}_i-\boldsymbol{v}_j\right\|^2\end{aligned}$$

定理 1　给定二部图 $G_b,\varepsilon>0$，矩阵 $\boldsymbol{Y}_{k_r\times n}=\sqrt{g_v}\boldsymbol{Q}\widehat{\boldsymbol{W}}^{1/2}\boldsymbol{B}\boldsymbol{L}^+$，则 $\forall\ i,j\in G_b$，至少存在 $1-1/n$ 概率满足

$$(1-\varepsilon)c_{ij}\le\left\|\boldsymbol{Y}(\boldsymbol{e}_i-\boldsymbol{e}_j)\right\|^2\le(1+\varepsilon)c_{ij}$$

其中 $\boldsymbol{Q}_{k_r\times s}$ 是行向量独立同分布的随机矩阵，$Q(i,j)=\pm1/\sqrt{k_r}$，$k_r=O(\log n/\varepsilon^2)$。

由引理 1 可证明定理 1。由定理 1 知 $c_{ij}\approx\left\|\boldsymbol{Y}(\boldsymbol{e}_i-\boldsymbol{e}_j)\right\|^2$，误差限为 ε。计算 $\boldsymbol{Y}$ 涉及 $\boldsymbol{L}^+$，而直接计算 $\boldsymbol{L}^+$ 复杂度过高。为加快计算速度，令 $\boldsymbol{\theta}=\sqrt{g_v}\boldsymbol{Q}(\widehat{\boldsymbol{W}}^{1/2}\boldsymbol{B})$，则 $\boldsymbol{Y}=\boldsymbol{\theta}\boldsymbol{L}^+$，其等价于 $\boldsymbol{Y}\boldsymbol{L}=\boldsymbol{\theta}$。通过 $\boldsymbol{\theta}$ 的每个行向量 $\boldsymbol{\theta}_i$ 可计算方程组 $\boldsymbol{y}_i\boldsymbol{L}=\boldsymbol{\theta}_i$，其中 $\boldsymbol{y}_i$ 是 $\boldsymbol{Y}$ 的行向量。使用线性时间求解程序 STSolve [17,18]获得的方程组 $\boldsymbol{y}_i\boldsymbol{L}=\boldsymbol{\theta}_i$ 的行向量表示为 $\hat{\boldsymbol{y}}_i$。由文献[20]，因为 $\left\|\boldsymbol{y}_i-\hat{\boldsymbol{y}}_i\right\|_L\le\varepsilon\left\|\boldsymbol{y}_i\right\|_L$，则有

$$(1-\varepsilon)^2c_{ij}\le\left\|\widehat{\boldsymbol{Y}}\left(\boldsymbol{e}_i-\boldsymbol{e}_j\right)\right\|^2\le(1+\varepsilon)^2c_{ij}$$

其中 $\widehat{\boldsymbol{Y}}$ 是由行向量 $\hat{\boldsymbol{y}}_i$ 构成的矩阵。则 $c_{ij}\approx\left\|\widehat{\boldsymbol{Y}}\left(\boldsymbol{e}_i-\boldsymbol{e}_j\right)\right\|^2$，误差限为 ε^2。称 $\widehat{\boldsymbol{Y}}$ 是二部图 G_b 的近似通勤距离嵌入，二部图的近似通勤距离嵌入 (Approximate Commute Distance Embedding of bipartite graph, ACDE)算法如表 1 所示。

表 1 二部图的近似通勤距离嵌入

算法 1 ACDE
输入：关系矩阵 $\boldsymbol{W}_{n_0\times n_1}$。
输出：近似通勤距离嵌入 $\widehat{\boldsymbol{Y}}$。
步骤：(1)由 $\boldsymbol{W}_{n_0\times n_1}$ 计算矩阵 $\boldsymbol{B}$, $\widehat{\boldsymbol{W}}$ 及 $\boldsymbol{L}$;
(2)计算 $\boldsymbol{\theta}=\sqrt{g_v}\boldsymbol{Q}\left(\widehat{\boldsymbol{W}}^{1/2}\boldsymbol{B}\right)$;
(3)通过调用 k_r 次 STSolve 方法[22]计算 $\boldsymbol{y}_i\boldsymbol{L}=\boldsymbol{\theta}_i$ 的每个 $\hat{\boldsymbol{y}}_i$, $1\le i\le k_r$;
(4)输出嵌入 $\widehat{\boldsymbol{Y}}$。

X_0 与 X_1 的数据对象映射到了一个共同的子空间。$\widehat{\boldsymbol{Y}}$ 的前 n_0 个列向量指示数据集 X_0，后 n_1 个列向量指示数据集 X_1。稀疏矩阵 $\boldsymbol{W}_{n_0\times n_1}$ 有 s 个非零元素，第(1)步计算 $\boldsymbol{B}$ 与 $\widehat{\boldsymbol{W}}$ 及 $\boldsymbol{L}$ 的时间复杂度为 $O(2s)+O(s)+O(n)$。因为稀疏矩阵 $\boldsymbol{B}$ 有 $2s$ 个非零元素，对角矩阵 $\widehat{\boldsymbol{W}}$ 有 s 个非零元素，故第(2)步计算 $\boldsymbol{\theta}$ 的时间复杂度为 $O(2sk_r+s)$。使用 STSolve 方法[17,18]计算方程组 $\boldsymbol{y}_i\boldsymbol{L}=\boldsymbol{\theta}_i$ 的一个解需要花费 $\widetilde{O}(s)$ 时间，故第(3)步构造 $\widehat{\boldsymbol{Y}}$ 的时间为 $\widetilde{O}(sk_r)$。则算法 ACDE 的时间复杂度为 $O(2s)+O(s)+O(n)+O(2sk_r+s)+\widetilde{O}(sk_r)=\widetilde{O}(4s+n+3sk_r)$。后续实验证明，在不同的数据集上 k_r 取值都很小。

3 基于近似通勤距离嵌入的异构信息网络聚类

3.1 通用模型的构建

定义 1 给定一个 T+1 种类型的数据集 $\chi=\{X_t\}_{t=0}^{\mathrm{T}}$ 上的信息网络 $G=<V,E,W>$，如果 $\forall e=\langle x_i,x_j\rangle\in E$，那么 $x_i\in X_0$ 且 $x_j\in X_t(t\ne 0)$，则称 G 为星型模式的异构信息网络。X_0 称为目标类型，X_t $(t\ne0)$称为属性类型。

设 $X_t=\{x_1^{(t)},x_2^{(t)},\cdots,x_{n_t}^{(t)}\}$，其中 n_t 是 X_t 的数据对象数目。$\boldsymbol{W}^{(0t)}\in\boldsymbol{R}^{n_0\times n_t}$ 是 X_0 与 X_t 之间的关系矩阵，其中，元素 $w_{ij}^{(0t)}$ 表示边 $<x_i^{(0)},x_j^{(t)}>$ 的权重。G 包含了 T 个 $\{\boldsymbol{W}^{(0t)}\}_{t=1}^{T}$。

目标数据集 X_0 与属性数据集 X_t 构成一个二部图 $G^{(0t)}$，一个二部图 $G^{(0t)}$ 对应一个关系矩阵 $\boldsymbol{W}^{(0t)}$。由算法 ACDE 计算二部图 $G^{(0t)}$ 的近似通勤距离嵌入 $Y^{(0t)}=\{y_1^{(0t)},y_2^{(0t)},\cdots,y_{n_0+n_t}^{(0t)}\}$，其中 $Y^{(0t)}$ 的前 n_0 个数据指示目标数据集 X_0，表示为 $Y_t^{(0)}$，后 n_t 个数据指示属性数据集 X_t，表示为 $Y^{(t)}$，称 $Y_t^{(0)}$ 与 $Y^{(t)}$ 为指示子集。$y_i^{(t)}\in Y_t^{(0)}$ 指示 X_0 的第 i 个对象，称 $y_i^{(t)}$ 为指示数据，$1\le i\le n_0$。$Y_t^{(0)}$ 的指示数据与 X_0 的对象一一对应。G 包含 T 个二部图，T 个二部图对应 T 个近似通勤距离嵌入，则目标数据集 X_0 被 T 个指示子集 $Y_t^{(0)}$ 所指示，X_0 的每个对象被 T 个指示数据所指示。

$\beta^{(t)}$ 是关系 $\boldsymbol{W}^{(0t)}$ 的权重，其中 $\sum_{t=1}^{T}\beta^{(t)}=1$，$\beta^{(t)}>0$。目标数据集 X_0 划分为 K 个类。指示同一个目标对象的指示数据属于 T 个不同的类，这 T 个不同的类分别属于 T 个不同的指示子集，将这 T 个类设置相同的类标号。由指示 X_0 的 T 个指示子集 $Y_t^{(0)}$ 构建模型：

$$F=\sum_{t=1}^{T}\left(\beta^{(t)}\sum_{i=1}^{n_0}\left(\gamma_{ij}\left(\sum_{j=1}^{K}\left\|y_i^{(t)}-\omega_j^{(t)}\right\|^2\right)\right)\right)\qquad(1)$$

其中 $\omega_j^{(t)}$ 是指示子集 $Y_t^{(0)}$ 的第 j 个类的中心点，指示函数 $\gamma=\{\gamma_{ij}\}_{i=1}^{n_0}$ 与目标数据集 X_0 的数据对象一一对应，当指示 X_0 的第 i 个对象的各指示数据 $y_i^{(t)}$ 属于 $Y_t^{(0)}$ 的第 j 个类时，$\gamma_{ij}=1$，否则 $\gamma_{ij}=0$。

从相容的角度，式(1)目标函数 F 取得最小值，则目标数据集 X_0 聚类达到最佳。显然，式(1)的全局最优解是 NP 难问题。

3.2 快速算法的推理

3.2.1 类标号设置 若 $\{Y_t^{(0)}\}_{t=1}^{T}$ 的类的标号已知，则 F 的极小值的求解过程简化。不妨设 $q_1,q_2\in X_0$，$\{y_1^{(t)},y_2^{(t)}\in Y_t^{(0)}\}_{t=1}^{T}$，$\{y_1^{(t)}\}_{t=1}^{T}$ 指示 q_1，$\{y_2^{(t)}\}_{t=1}^{T}$ 指示 q_2。因为指示同一个目标对象的指示数据所属的类具有相同的类标号，若已知 $\{y_1^{(t)}\}_{t=1}^{T}$ 中的一个指示数据属于第 j 个类，则所有 $\{y_1^{(t)}\}_{t=1}^{T}$ 在各自的指示子集均属于第 j 个类；若已知 $\{y_1^{(t)}\}_{t=1}^{T}$ 属于第 j 个类，则 $\{y_2^{(t)}\}_{t=1}^{T}$ 在各自的指示子集或者都属于第 j 个类，或者都不属于第 j 个类。

$\{Y_t^{(0)}\}_{t=1}^{T}$ 的每个类都有一个初始中心点，在目标数据集 X_0 中随机选择 K 个对象，指示这 K 个对象的指示数据在各自的指示子集中作为 K 个类的初始中心点，指示同一个目标对象的中心点，令其所在的类的标号相同，从而完成各指示子集的类标号的设置。则其他指示同一个目标对象的指示数据或者都属于第 j 个类，或者都不属于第 j 个类，$1\le j\le K$。

3.2.2 加权距离总和 X_0 的一个对象被 T 个指示数据所指示，这 T 个指示数据到各自嵌入子集的类的中心点的距离都影响着该目标对象所属类的分配。指示数据到中心点的距离的权重由其指示子集的权重决定，指示子集的权重就是相应的关系矩阵的权重。

不妨设 $q_i\in X_0$，$\{y_i^{(t)}\in Y_t^{(0)}\}_{t=1}^{T}$，$\{y_i^{(t)}\}_{t=1}^{T}$ 指示 q_i，指示数据 $y_i^{(t)}$ 到 $Y_t^{(0)}$ 的第 j 个类的中心点的加权距离为 $\beta^{(t)}\left\|y_i^{(t)}-\omega_j^{(t)}\right\|^2$。加权距离总和 $\mathrm{dis}=\sum_{t=1}^{T}\beta^{(t)}\left\|y_i^{(t)}-\omega_j^{(t)}\right\|^2$ 决定了 q_i 所属的类。

$$j=\arg\min\sum_{t=1}^{T}\beta^{(t)}\left\|y_i^{(t)}-\omega_j^{(t)}\right\|^2 \tag{2}$$

j 就是 q_i 所属类的标号，也是 $\{y_i^{(t)}\}_{t=1}^{T}$ 所属类的标号。

3.2.3 F 的极小值求解　F 可以进一步表示为

$$\begin{aligned}F&=\sum_{t=1}^{T}\left(\beta^{(t)}\sum_{i=1}^{n_0}\left(\gamma_{ij}\left(\sum_{j=1}^{K}\left\|y_i^{(t)}-\omega_j^{(t)}\right\|^2\right)\right)\right)\\&=\sum_{i=1}^{n_0}\left(\gamma_{ij}\left(\sum_{j=1}^{K}\left(\sum_{t=1}^{T}\beta^{(t)}\left\|y_i^{(t)}-\omega_j^{(t)}\right\|^2\right)\right)\right)\end{aligned} \tag{3}$$

给定这 T 个指示子集的类的初始中心点 $\{\omega_j^{(t)}\}_{j=1}^{K}$，$1\le t\le T$，首先由式(2)划分指示子集 $\{Y_t^{(0)}\}_{t=1}^{T}$ 的类，并记式(3)目标函数 F 的值为 F_1。$\{Y_t^{(0)}\}_{t=2}^{T}$ 的类的中心点不变，然后计算 $Y_1^{(0)}$ 的每个类的新中心点 $\{\widehat{\omega}_j^{(1)}\}_{j=1}^{K}$，新中心点取值该类所有指示数据的平均值，$\gamma_{ij}$ 不变，记式(3)目标函数 F 的值为 F_2。

定理 2

$$F_2=\sum_{i=1}^{n_0}\left(\gamma_{ij}\sum_{j=1}^{K}\left(\beta^{(1)}\left\|y_i^{(1)}-\widehat{\omega}_j^{(1)}\right\|^2+\sum_{t=2}^{T}\left(\beta^{(t)}\left\|y_i^{(t)}-\omega_j^{(t)}\right\|^2\right)\right)\right)\le F_1$$

证明　因为 γ_{ij} 不变，故

$$\begin{aligned}F_2=&\sum_{i=1}^{n_0}\left(\gamma_{ij}\left(\sum_{j=1}^{K}\left(\beta^{(1)}\left\|y_i^{(1)}-\widehat{\omega}_j^{(1)}\right\|^2\right)\right)\right)\\&+\sum_{i=1}^{n_0}\left(\gamma_{ij}\left(\sum_{j=1}^{K}\sum_{t=2}^{T}\left(\beta^{(t)}\left\|y_i^{(t)}-\omega_j^{(t)}\right\|^2\right)\right)\right)\end{aligned}$$

因为 $\{Y_t^{(0)}\}_{t=2}^{T}$ 的类的中心点不变，所以 $\sum_{i=1}^{n_0}\left(\gamma_{ij}\left(\sum_{j=1}^{K}\sum_{t=2}^{T}\left(\beta^{(t)}\left\|y_i^{(t)}-\omega_j^{(t)}\right\|^2\right)\right)\right)$ 为固定值。而

$$\sum_{i=1}^{n_0}\left(\gamma_{ij}\left(\sum_{j=1}^{K}\left(\beta^{(1)}\left\|y_i^{(1)}-\widehat{\omega}_j^{(1)}\right\|^2\right)\right)\right)\le\sum_{i=1}^{n_0}\left(\gamma_{ij}\left(\sum_{j=1}^{K}\left(\beta^{(1)}\left\|y_i^{(1)}-\omega_j^{(1)}\right\|^2\right)\right)\right)$$

可见重新确定子集 $Y_1^{(0)}$ 的类的中心点，$F_2\le F_1$。

而当 $Y_1^{(0)}$ 的类替换了新的中心点 $\{\widehat{\omega}_j^{(1)}\}_{j=1}^{K}$，$\{Y_t^{(0)}\}_{t=2}^{T}$ 的中心点不变，由式(2)重新划分 $\{Y_t^{(0)}\}_{t=1}^{T}$ 类，记对应的式(3)目标函数 F 的值为 F_3，则有 $F_3\le F_2$。

由式(2)划分 $\{Y_t^{(0)}\}_{t=1}^{T}$ 的类，确定 $Y_1^{(0)}$ 的类的新中心点 $\{\widehat{\omega}_j^{(1)}\}_{j=1}^{K}$，用新的中心点代替原中心点 $\{\omega_j^{(1)}=\widehat{\omega}_j^{(1)}\}_{j=1}^{K}$，并重新划分 $\{Y_t^{(0)}\}_{t=1}^{T}$ 的类；然后逐一对 $\{Y_t^{(0)}\}_{t=2}^{T}$ 执行相同的操作。重复上述操作，直到式(3)收敛，则得到 F 的局部极小值。基于嵌入技术的异构信息网络快速聚类算法(Fast Clustering Algorithm based on Embedding Technology for heterogeneous information networks, FCAET)如表 2 所示。

表 2 基于嵌入技术的异构信息网络快速聚类算法

算法 2　FCAET

输入：$\{\boldsymbol{W}^{(0t)}\in\boldsymbol{R}^{n_0\times n_t}\}_{t=1}^{T},\{\beta^{(t)}>0\}_{t=1}^{T}$，聚类数 K。

输出：目标数据集 X_0 的类。

步骤：(1)for　t=1:T do

{ (a)由算法 1 计算二部图 $G^{(0t)}$的近似通勤距离嵌入；

(b)确定指示 X_0 的指示子集 $Y_t^{(0)}$；}

(2)初始化 $\{Y_t^{(0)}\}_{t=1}^{T}$ 的 K 个类的初始中心点 $\{\omega_j^{(t)}\}_{j=1}^{K}$，并建立类标号；

(3) do

{ for　t=1:T do

{ (a)由式(2)确定 $\{Y_t^{(0)}\}_{t=1}^{T}$ 的 K 个类；

(b)重新确定 $Y_t^{(0)}$ 每个类的新的中心点 $\{\widehat{\omega}_j^{(t)}\}_{j=1}^{K}$；

(c) $\{\omega_j^{(t)}=\widehat{\omega}_j^{(t)}\}_{j=1}^{K}$；}

} while 式(3)收敛；

(4)输出目标数据集 X_0 的类。

由算法 1 的时间复杂度分析可知算法 2 第(1)步的时间复杂度为 $\widetilde{O}\left(\sum_{t=1}^{T}(4s_t+n_t+3s_tk_r)\right)$，其中 T 是异构信息网络的二部图的数目，k_r 是指示子集 $Y_t^{(0)}$ 数据的维度，n_t 与 s_t 是第 t 个二部图的结点数与边数。第(2)步仅仅花费 $O(K)$时间，是常量。第(3)步花费 $O(uTKk_rn_0)$时间，其中 K 是聚类数目，n_0 是 X_0 的对象数目，u 是式(3)收敛的迭代次数。所以算法 FCAET 的时间复杂度为 $\widetilde{O}\left(\sum_{t=1}^{T}(4s_t+n_t+3s_tk_r)\right)+O(uTKk_rn_0)$，其中 k_r,u 都很小，而 T 与 K 是常量。

4　实验

4.1 实验数据

从 DBLP 选取真实数据建立实验数据集，DBLP 是一个典型的异构信息网络，其中包括 4 种类型数据对象，分别命名为 papers, authors, terms 和 venues。首先抽取一个小数据集 S_1，即文献[8]使用的称为“four-area dataset”的数据集。小数据集 S_1 选取了 4 个学术区域，这 4 个区域为：database, data mining, information retrieval 及 machine learning。每个区域取 5 个有代表性的会议，共 20 个会议，20 个会议的所有 authors, papers 及出现在论文题目中的所有 terms。

本文又抽取 2008～2012 年的 8 个学术区域的

DBLP 数据作为另外一个测试数据集，这 8 个学术区域为：databases, data mining & machine learning, information retrieval, computer graphics, computer network, information security, computer architecture, software engineering & programming language。数据集共包括 80 个 venues(每个区域选取 10 个 venues), 21,271 个 papers, 45, 576 个 authors 及 42, 962 个 terms。这个大的数据集称为 S_2。

当分析 papers 时，papers 作为目标数据集，其他为属性数据集。因为 DBLP 提供了非常有限的引用信息，故 papers 之间不设置直接的链接。当分析 authors 时，authors 作为目标数据集，papers 与 venues 作为属性数据集。由于合作关系，authors 之间存在直接的链接。故 authors 还作为目标数据集的另外一个属性数据集。

本文算法均采用文献[22]的一种近乎线性时间的求解程序计算算法中的嵌入数据集，该方法用于对角占优矩阵，网址 http://www.cs.cmu.edu/~jkoutis/cmg.html。实验的运行环境为 Inter Pentium III 处理器，2 GB 内存，Windows XP 操作系统，Matlab 编程。

4.2 关系矩阵的确定

当 papers 为目标数据集, authors, venues 和 terms 为属性数据集。X_0 表示目标数据集 papers, X_1, X_2 与 X_3 分别表示属性数据集 authors, venues 与 terms。X_0 与 $\{X_t\}_{t=1}^3$ 的关系矩阵为 $\{\boldsymbol{W}^{(0t)}\}_{t=1}^3$，其中 $\{\boldsymbol{W}^{(0t)}\}_{t=1}^3$ 的元素为

$$w_{ij}^{(0t)}=\begin{cases}1, & \text{如果} i\in X_0, j\in X_1\cup X_2,\\ & \quad \text{结点 } i \text{ 连接结点 } j\\ p, & \text{如果} i\in X_0, j\in X_3,\\ & \quad \text{结点} i \text{在结点} j \text{中出现 } p \text{ 次}\\ 0, & \text{其他}\end{cases}$$

当 authors 为目标数据集, papers 和 venues 为属性数据集。因为作者之间存在合作关系，authors 也是另外一个属性数据集。X_0 表示 authors, X_1 与 X_2 分别表示 papers 与 venues。$\boldsymbol{W}^{(0t)}$ 为 X_0 与 X_t 的关系矩阵，$0\le t\le 2$。其中 $\{\boldsymbol{W}^{(0t)}\}_{t=0}^2$ 的元素为

$$w_{ij}^{(0t)}=\begin{cases}1, & \text{如果} i\in X_0, j\in X_1\cup X_2, \text{结点} i \text{连接结点} j\\ p, & \text{如果} i\in X_0, j\in X_0, \text{结点} i \text{与} j \text{合作} p \text{篇文章}\\ 0, & \text{其他}\end{cases}$$

实验的所有算法均采用相同的关系矩阵。

4.3 参数分析

4.3.1 参数 k_r 分析 k_r 在不同的数据集上取值都非常小[16]。文献[16]实验证明对同构数据集，当 $k_r\ge 50$ 时，准确率曲线已经很平滑。取小数据集 S_1 来比较参数 k_r 的变化对异构信息网络数据聚类准确率的影响。使用算法 FCAET 聚类 papers 时，$\{\boldsymbol{W}^{(0t)}\}_{t=1}^3$ 的权重分别取 $\beta^{(1)}$=0.3，$\beta^{(2)}$=0.4，$\beta^{(3)}$=0.3。聚类 authors 时，$\{\boldsymbol{W}^{(0t)}\}_{t=0}^2$ 的权重分别取 $\beta^{(1)}$=0.4，$\beta^{(2)}$=0.2，$\beta^{(3)}=0.4$。迭代次数 $u=40$。k_r 对算法准确率的影响如图 1，图 2 所示。

实验说明当 k_r>50 时准确率曲线已经趋于平滑。因此取 k_r=60 是很适合的。k_r 很小，且对算法 FCAET 的计算速度基本没有影响。这也是算法 FCAET 的一个优点，即在准确性方面对参数 k_r 不敏感。其他实验也取相同的关系矩阵权重及 k_r=60。

4.3.2 迭代次数 u 分析 取小数据集 S_1 比较迭代次数 u 对异构信息网络数据聚类准确率的影响。迭代参数 u 对 papers 及 authors 聚类准确率的影响如图 3，图 4 所示。当 $u=30$ 时，算法就已经收敛，说明本文算法收敛速度非常快，本文其他实验均取 $u=40$。

4.4 聚类稳定性对比

本文选择复杂度较低的通用算法 CIT[6]，基于排序的算法 NetClus[8] 与本文算法 FCAET 做稳定性比较。算法 ComClus 是 NetClus 的衍生算法，性质类似，这里不做分析。本次实验在小数据集 S_1 上聚类目标数据集 papers，以比较 3 个算法 CIT,

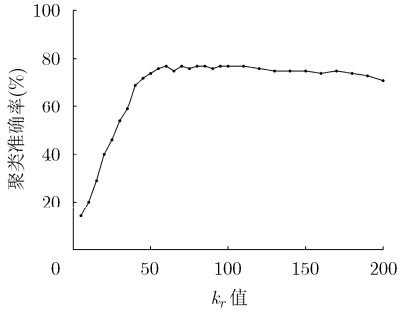

图 1 k_r 对 papers 的影响

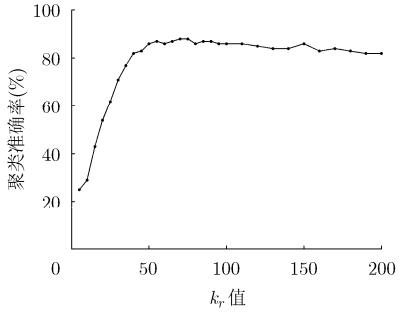

图 2 k_r 对 authors 的影响

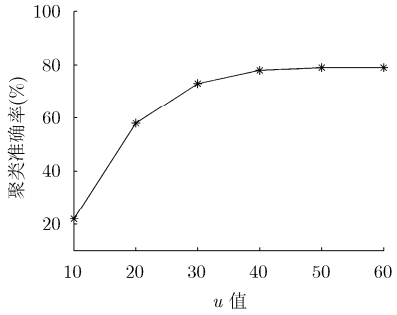

图 3　u 对 papers 的影响

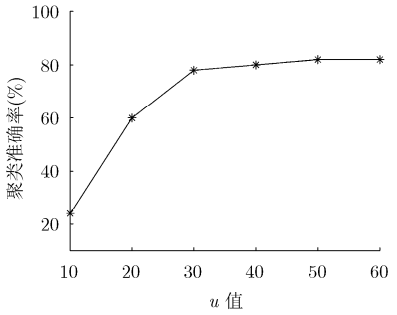

图 4　u 对 authors 的影响

NetClus 及 FCAET 的稳定性。3 个算法的 10 次实验准确率如图 5 所示，虽然算法 FCAET 和算法 NetClus 的计算速度都很快，但图 5 的实验数据说明本文算法 FCAET 的稳定性更好，初始中心点的选择对聚类结果影响不大，而算法 NetClus 不是很稳定，初始类的划分对算法 NetClus 的函数及收敛速度影响非常大。当异构信息网络稀疏时，算法 CIT 的收敛速度比较慢，当迭代次数 u 很小时，聚类精度并不高。

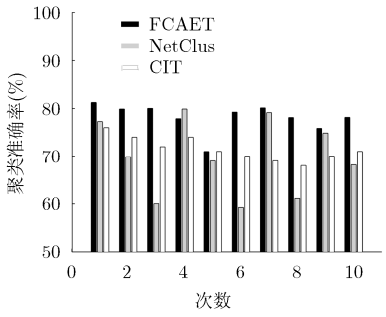

图 5　3 个算法 10 次实验聚类准确率比较

4.5 聚类准确率及计算速度的对比

基于半正定的规划算法[5]以及谱聚类算法[7]时间复杂度过高，不适合聚类规模较大的异构信息网络。本文选择复杂度较低的通用算法 CIT[6]，基于排序的算法 NetClus[8]，及 NetClus 的衍生算法 ComClus[11]与本文算法 FCAET 做准确率及计算速度的比较。算法 NetClus, ComClus 的参数均取文献[8]给定的实验参数。每个算法都涉及到初始划分，初始划分影响聚类结果，因此，对于每个目标数据集，所有算法均做 3 次实验，3 次实验中聚类准确率最高的一次作为该算法的准确率，本次实验的计算速度作为该算法的计算速度。聚类准确率对比结果如表 3 所示，表 3 数据说明本文算法的聚类准确率高于其他算法。算法 CIT 计算复杂度 $O(n^2)$，当异构信息网络稀疏时，算法 CIT 的收敛速度比较慢，故聚类精度并不高。算法 NetClus 只使用了异构数据关系，故聚类精度较低。算法 ComClus 利用了同构数据的关系，聚类精度高于 NetClus 的聚类精度，但也增加了计算复杂度。本文算法是一种基于图嵌入技术的聚类算法，而且采用通勤距离度量能够更加充分表现数据之间的关系，并且收敛性速度不受关系矩阵的稀疏性影响，同时也能够利用目标数据集的同构关系，故聚类精度高。计算速度对比结果如表 4 所示，表 4 数据说明本文算法的计算时间基本等同于算法 NetClus 的计算时间。但本文算法 FCAET 通用性更强，可以用于任何一个星型网络模式的异构信息网络的聚类。而算法 NetClus 及 ComClus 需要根据具体应用领域设计函数，依赖于数据特性，不具有普遍性。

4.6 FCAET 运行时间分析

算法 FCAET 在两个数据集上的运行时间分布情况如表 5 所示，实验说明本文算法的运行效率是很高的。3 个指示子集的串行计算速度占程序运行时间 50%左右。若并行计算 3 个指示子集，速度会更快。求解模型极小值也可采用并行执行以提高计算速度。

表 3 聚类准确率比较(%)

目标对象	CIT	NetClus	ComClus	FCAET
S_1 的 papers	73.91	71.54	72.83	78.87
S_1 的 authors	74.41	69.13	74.91	81.33
S_2 的 papers	70.84	71.28	72.93	76.36
S_2 的 authors	71.02	68.29	73.01	77.94

表 4 计算速度比较(s)

目标对象	CIT	NetClus	ComClus	FCAET
S_1 的 papers	78.5	37.3	40.3	37.1
S_1 的 authors	79.8	36.9	39.8	38.3
S_2 的 papers	1459.3	792.6	817.3	798.4
S_2 的 authors	1474.7	733.7	771.4	764.9

表 5 算法 FCAET 运行时间分配情况(s)

目标对象	嵌入计算时间	聚类时间	时间总和
S_1 的 papers	19.6	17.5	37.1
S_1 的 authors	18.1	20.2	38.3
S_2 的 papers	393.8	405.6	798.4
S_2 的 authors	376.4	388.5	764.9

5 结束语

本文提出的异构信息网络快速聚类算法不同于以往的异构数据聚类算法，本文算法根据每个关系矩阵先求解指示目标数据集的每个指示数据集，然后根据指示数据之间的关系建立模型，从而得到目标数据集的划分。本文算法通用性强、稳定性好，同时计算速度快而且聚类准确率高，非常适合异构信息网络的聚类。而以往的聚类算法要么时间复杂度过高，要么通用性不强，聚类结果不稳定。通过理论分析及实验验证，本文算法的计算速度和聚类准确率满足要求。本文算法中各关系矩阵的权重影响着目标数据集的划分，这些权重还不能自适应确定，需要进一步研究。当异构数据关系不再稀疏，网络模式不再是星型的，如何快速划分任意模式的大规模异构信息网络，都是亟待解决的问题。

参 考 文 献

[1] 肖杰斌, 张绍武. 基于随机游走和增量相关节点的动态网络社团挖掘算法[J]. 电子与信息学报. 2013, 35(4): 977–981.
Xiao Jie-bin and Zhang Shao-wu. An algorithm of integrating random walk and increment correlative vertexes for mining community of dynamic networks[J]. *Journal of Electronics & Information Technology*, 2013, 35(4): 977–981.

[2] 陈季梦, 陈家俊, 刘杰, 等. 基于结构相似度的大规模社交网络聚类算法[J]. 电子与信息学报. 2015, 37(2): 449–454.
Chen Ji-meng, Chen Jia-jun, Liu Jie, *et al.*. Clustering algorithms for large-scale social networks based on structural similarity[J]. *Journal of Electronics & Information Technology*, 2015, 37(2): 449–454.

[3] Sun Y and Han J. Mining heterogeneous information networks: principles and methodologies[J]. *Proceedings of Mining Heterogeneous Information Networks*: *Principles and Methodologies*, 2012, 3(2): 1–159.

[4] Huang Y and Gao X. Clustering on heterogeneous networks [J]. *Wiley Interdisciplinary Reviews*: *Data Mining and Knowledge Discovery*, 2014, 4(3): 213–233.

[5] Gao B, Liu T Y, Zheng X, *et al.*. Consistent bipartite graph co-partitioning for star-structured high-order heterogeneous data co-clustering[C]. Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, 2005: 41–50.

[6] Gao B, Liu T, and Ma W-Y. Star-structured high-order heterogeneous data co-clustering based on consistent information theory[C]. Proceedings of the 6th International Conference on Data Mining (ICDM 2006), Hong Kong, 2006: 880–884.

[7] Long B, Zhang Z M, Wu X, *et al.*. Spectral clustering for multi-type relational data[C]. Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, 2006: 585–592.

[8] Sun Y, Yu Y, and Han J. Ranking-based clustering of heterogeneous information networks with star network schema[C]. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, 2009: 797–806.

[9] Li P, Wen J, and Li X. SNTClus: a novel service clustering algorithm based on network analysis and service tags[J]. *Przeglad Elektrotechniczny*, 2013, 89(1): 208–210.

[10] Li P, Chen L, Li X, *et al.*. RNRank: Network-Based Ranking on Relational Tuples[M]. Boston: Behavior and Social Computing, Springer International Publishing, 2013: 139–150.

[11] Wang R, Shi C, Philip S Y, *et al.*. Integrating Clustering and Ranking on Hybrid Heterogeneous Information Network[M]. Berlin: Advances in Knowledge Discovery and Data Mining, Springer Berlin Heidelberg, 2013: 583–594.

[12] Boden B, Ester M, and Seidl T. Density-Based Subspace Clustering in Heterogeneous Networks[M]. Berlin: Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelberg, 2014: 149–164.

[13] Meng Q, Tafavogh S, and Kennedy P J. Community detection on heterogeneous networks by multiple semantic-path clustering[C]. 2014 6th IEEE International Conference on Computational Aspects of Social Networks (CASoN), Porto, 2014: 7–12.

[14] Meng X, Shi C, Li Y, *et al.*. Relevance Measure in Large-scale Heterogeneous Networks[M]. Boston: Web Technologies and Applications, Springer International Publishing, 2014: 636–643.

[15] Aggarwal C C, Xie Y, and Philip S Y. On dynamic link inference in heterogeneous networks[C]. SIAM International Conference on Data Mining, Anaheim, 2012: 415–426.

[16] Khoa N L D and Chawla S. Large Scale Spectral Clustering Using Resistance Distance and Spielman-teng Solvers[M]. Berlin: Discovery Science, Springer Berlin Heidelberg, 2012: 7–21.

[17] Spielman D A and Teng S H. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving

linear systems[C]. Proceedings of the 36th Annual ACM Symposium on Theory of Computing, Chicago, 2004: 81–90.

[18] Spielman D A and Teng S H. Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems[J]. *SIAM Journal on Matrix Analysis and Applications*, 2014, 35(3): 835–885.

[19] Fouss F, Pirotte A, Renders J M, *et al*.. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2007, 19(3): 355–369.

[20] Spielman D A and Srivastava N. Graph sparsification by effective resistances[J]. *SIAM Journal on Computing*, 2011, 40(6): 1913–1926.

[21] Achlioptas D. Database-friendly random projections[C]. Proceedings of the 20th ACM Sigmod-Sigact-Sigart Symposium on Principles of Database Systems, New York, 2001: 274–281.

[22] Koutis I, Miller G L, and Tolliver D. Combinatorial preconditioners and multilevel solvers for problems in computer vision and image processing[J]. *Computer Vision and Image Understanding*, 2011, 115(12): 1638–1646.

陈丽敏：　女，1970 年生，副教授，博士生，研究方向为数据挖掘、机器学习.

杨　静：　女，1962 年生，教授，博士生导师，研究方向为数据库与知识库.

张健沛：　男，1956 年生，教授，博士生导师，研究方向为数据库与知识库.