

文章编号:1001-9081(2012)04-1097-04

doi:10.3724/SP.J.1087.2012.01097

# 图像搜索结果的重叠层次聚类与代表点展现

谷瑞军<sup>1\*</sup>, 陈圣磊<sup>1</sup>, 陈耿<sup>1,2</sup>, 汪加才<sup>1</sup>

(1. 南京审计学院 信息科学学院, 南京 210029; 2. 江苏大学 计算机科学与通信工程学院, 江苏 镇江 212013)

(\* 通信作者电子邮箱 grj79@hotmail.com)

**摘要:**针对图像聚类中面临的高维、准确度低、部分重叠等问题,提出了一种高效的基于链接层次聚类的多标记图像聚类。该方法通过图像距离计算相似度,通过链接聚类检测重叠簇。从而每个图像可能归属于多个簇,使得簇标签的意义更明确。为了检验方法的有效性,对通过搜索引擎检索特定关键词返回的图片数据集进行聚类。结果表明,该方法能有效发现具有重叠划分的簇,且簇的意义比较明确。

**关键词:**图像聚类;链接聚类;多簇划分;图像距离

**中图分类号:** TP391.41    **文献标志码:**A

## Hierarchical overlapping clustering and exemplar visualization of images returned by search engine

GU Rui-jun<sup>1\*</sup>, CHEN Sheng-lei<sup>1</sup>, CHEN Geng<sup>1,2</sup>, WANG Jia-cai<sup>1</sup>

(1. School of Information Science, Nanjing Audit University, Nanjing Jiangsu 210029, China;

2. School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang Jiangsu 212013, China)

**Abstract:** To resolve the problems of high dimensionality, low accuracy and overlapping in image clustering, an effective link-clustering based image multiple-cluster partition method was proposed in this paper. This method utilized image distance to measure similarity and identified overlapping clusters by using link-clustering. As a result, an image may be partitioned into multiple clusters, and this multiple-cluster partition makes each cluster more specific compared with others. To validate this method, experiments were carried out on the datasets returned by search engine when searching for some key words. The result shows that the proposed method can find explicit clusters with partial overlapping.

**Key words:** image clustering; link clustering; multiple-cluster partition; image distance

随着图像数量的急剧增长,图像聚类<sup>[1-9]</sup>已成为将大量图像划分为少数有意义分组(簇)的重要技术。通常情况下,通过图像搜索引擎返回的搜索结果包含多个主题。将结果组织为不同语义的簇有利于用户的浏览。然而,对于图像聚类存在很多挑战,例如高维灾难、可伸缩性差、准确度低、簇意义模糊、簇部分重叠、图像关键特征不易提取等。本文关注的问题是如何对 Web 图像的搜索结果进行聚类。为了提高图像聚类的结果质量,本文提出一种高效的基于链接层次聚类的多标记图像聚类,该方法通过图像距离计算相似度,通过链接聚类检测重叠簇,从而每个图像可能归属于多个簇,使得簇标签的意义更明确。为了检验方法的有效性,选择“flower”、“Afric”等几个关键词通过搜索引擎进行图片搜索,并取前 25 张图片进行聚类,结果表明,该方法能有效发现具有重叠划分的簇,且簇的意义比较明确。

## 1 相关工作

近年来,数字媒体技术的发展使得数字图像的制作和传播越来越容易,有效的图像搜索已成为多媒体搜索领域的重要研究课题之一。精确图像搜索是当前研究的焦点,例如使用 PageRank<sup>[6]</sup>算法根据图像的纹理信息和内容搜索用户感兴趣的图像。为了实现 Web 图像检索结果的聚类,文献[8]定义了单词与图像节点之间的异构链接以及单词节点之

间的同构链接,提出并定义了单词可见度这一属性,并将其集成到传统的 TF-IDF 模型中以挖掘单词—图像之间关联的权重,应用复杂图聚类和二部图协同谱聚类等算法验证了在图模型上引入两种相关性关联的有效性,达到了改进了 Web 图像聚类性能的目的。然而,图像的搜索结果往往数量巨大,如何进行有效的聚类,提供给不同兴趣爱好的用户进行选择需要进一步研究。

传统上,由于很多搜索引擎返回成千上万的图片排序列表,因此带给用户不好的用户体验。目前解决的方法是根据不同的视角将搜索结果进行重新组织,划分为不同的分组。文献[9]提出一个使用可视化、纹理和链接分析的层次式聚类方法。通过使用基于视觉的页面分割算法,网页被划分成块,图像的纹理和链接信息可以从包含该图像的块中提取。通过使用块层次的链接分析技术,可以构建以图像为节点的图。然后,应用谱技术发现图像 Euclidean 嵌入,该嵌入反映了图的结构。对于每幅图像,该方法中给出 3 种表示,即视觉特征、纹理特征、基于表示的图。然后,使用谱聚类技术,将搜索结果聚类为不同语义的簇。文献[3]借鉴近邻传播聚类的思想,设计了一种稀疏、快速的近邻传播算法,可以发现图像搜索结果的代表点,从而更好地展示搜索结果。在真实数据集上的实验结果证明了该方法在视觉表达和定量分析上的有效性。上述方法多是采用经典的聚类算法,虽然可以生成有

收稿日期:2011-10-08;修回日期:2011-12-03。

基金项目:国家自然科学基金资助项目(70971067/C0112);国家社会科学基金资助项目(10BGL016)。

作者简介:谷瑞军(1979 -),男,山东菏泽人,讲师,博士,CCF 会员,主要研究方向:图像检索、数据挖掘; 陈圣磊(1977 -),男,山东兗州人,讲师,博士,主要研究方向:机器学习; 陈耿(1965 -),男,江苏南京人,教授,博士生导师,CCF 高级会员,主要研究方向:计算机审计、数据挖掘; 汪加才(1962 -),男,江苏连云港人,教授,博士,主要研究方向:商务智能。

划分的结果,但如何呈现给用户依然是一个开放性问题。另外,聚类后,每个图像只能划分到一个簇,即硬划分。事实上,一个图像可能包含多种语义,即可以同时归属于多个簇。

## 2 基于链接聚类的图像多重划分

针对图像聚类中面临的高维、准确度低、部分重叠等问题,提出了一种高效的基于链接层次聚类的多标记图像聚类。该方法称为基于链接聚类的图像多重划分( Link-clustering based Image Multiple-clusters Partition, LIMP)。描述如下。

第 1 步 考虑到像素间的空间相关信息,计算图像间的图像距离。

第 2 步 计算图像间的相似度,设定阈值,若两个图像间的相似度大于阈值,则认为两者之间存在边,边的权重为相似度,从而构造出一个加权无向图。

第 3 步 应用基于边相似度的链接层次聚类对加权无向图的边集进行聚类,得到一个基于边集的簇划分。因为一个节点可归属多个边,从而可以归属于多个簇。

第 4 步 根据簇内节点的权重和进行排序,找到权重和最大的节点作为该簇的代表点。

### 2.1 图像距离计算

与相似度计算中常用的欧氏距离不同,图像聚类( Image Distance, IMD)<sup>[10]</sup> 考虑到像素间的空间相关信息。因此,IMD 对图像小的扰动更具鲁棒性。假设  $\mathbf{x}, \mathbf{y}$  是两幅大小为  $m \times n$  的图像,其中  $\mathbf{x} = (x_1, x_2, \dots, x_{mn})$ ,  $\mathbf{y} = (y_1, y_2, \dots, y_{mn})$ , 欧氏距离  $d_E^2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{mn} (x_i - y_i)^2$ 。设  $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{mn}$  是  $mn$ -维图像空间的基, 度量系数  $q_{ij}$  定义如下:

$$q_{ij} = \langle \mathbf{e}_i, \mathbf{e}_j \rangle = \sqrt{\langle \mathbf{e}_i, \mathbf{e}_i \rangle} \sqrt{\langle \mathbf{e}_j, \mathbf{e}_j \rangle} \cdot \cos \theta_{ij} \quad (1)$$

其中: $\langle \mathbf{e}_i, \mathbf{e}_j \rangle$  表示  $\mathbf{e}_i$  和  $\mathbf{e}_j$  的内积,  $\theta_{ij}$  是  $\mathbf{e}_i$  和  $\mathbf{e}_j$  的夹角, 图像  $\mathbf{x}, \mathbf{y}$  间的图像距离定义为:

$$d_{\text{IMD}}^2(\mathbf{x}, \mathbf{y}) = \sum_{i,j=1}^{mn} q_{ij} (x_i - y_i) (x_j - y_j) = (\mathbf{x} - \mathbf{y})^\top \mathbf{Q} (\mathbf{x} - \mathbf{y}) \quad (2)$$

其中:  $\mathbf{Q} = (q_{ij})_{mn \times mn}$  是对称的正定矩阵,  $q_{ij}$  可表示为一个 Gaussian 函数:

$$q_{ij} = \frac{1}{2\pi\sigma^2} \sum_{i,j=1}^{mn} \exp\left\{-\frac{|P_i - P_j|^2}{2\sigma^2}\right\} \quad (3)$$

其中  $|P_i - P_j|$  是像素  $P_i$  和  $P_j$  间的距离。设  $\sigma = 1$ , 有:

$$d_{\text{IMD}}^2(\mathbf{x}, \mathbf{y}) = \frac{1}{2\pi} \sum_{i,j=1}^{mn} \exp\left\{-\frac{|P_i - P_j|^2}{2}\right\} (x_i - y_i) (x_j - y_j) \quad (4)$$

为减少计算复杂度,  $\mathbf{Q}$  分解为:

$$\mathbf{Q} = \boldsymbol{\Gamma} \boldsymbol{\Lambda} \boldsymbol{\Gamma}^\top = (\boldsymbol{\Gamma} \boldsymbol{\Lambda}^{1/2} \boldsymbol{\Gamma}^\top) (\boldsymbol{\Gamma} \boldsymbol{\Lambda}^{1/2} \boldsymbol{\Gamma}) = \boldsymbol{\Lambda}^{1/2} \mathbf{Q} \boldsymbol{\Lambda}^{1/2}$$

其中:  $\boldsymbol{\Lambda}$  是由  $\mathbf{Q}$  的特征值构成的对角矩阵,  $\boldsymbol{\Gamma}$  是由  $\mathbf{G}$  的特征向量构成的正交矩阵。设  $\mathbf{u} = \boldsymbol{\Lambda}^{1/2} \mathbf{x}$ ,  $\mathbf{v} = \boldsymbol{\Lambda}^{1/2} \mathbf{y}$ , 于是:

$$d_{\text{IMD}}^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^\top \boldsymbol{\Lambda}^{1/2} \boldsymbol{\Lambda}^{1/2} (\mathbf{x} - \mathbf{y}) = (\mathbf{u} - \mathbf{v})^\top (\mathbf{u} - \mathbf{v}) \quad (5)$$

该式具有欧氏距离的形式。因为  $\mathbf{Q}$  仅和图像的大小有关, 和图像的内容无关,可在执行具体图像变换前预先计算  $\mathbf{Q}$ 。图像距离的计算如算法 1 所示。

### 算法 1 Computer image distance。

- 1) 计算  $\mathbf{Q}: q_{ij} = \frac{1}{2\pi\sigma^2} \sum_{i,j=1}^{mn} \exp\left\{-\frac{|P_i - P_j|^2}{2\sigma^2}\right\}$ 。

2) 计算 RGB 空间欧氏距离。

3) 计算图像距离:

$$d_{\text{IMD}}^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^\top \mathbf{Q}^{1/2} \mathbf{Q}^{1/2} (\mathbf{x} - \mathbf{y}) = (\mathbf{u} - \mathbf{v})^\top (\mathbf{u} - \mathbf{v})$$

### 2.2 基于相似度构造加权图

图像距离可以表示两个图像间的相似程度,即距离越近,相似度越大。图像  $\mathbf{x}_i$  和  $\mathbf{x}_j$  间的相似度表示为:

$$s_{ij} = e^{-d_{\text{img}}^2(\mathbf{x}_i, \mathbf{x}_j) / \langle d_{\text{img}} \rangle} \quad (6)$$

其中  $\langle d_{\text{img}} \rangle$  为所有图像距离的平均值。如果将图像  $\mathbf{x}$  看作图的节点,则可将相似度视为节点间边的权重,从而可以构造一个加权的无向图。算法描述如算法 2 所示。

### 算法 2 Construct weighted graph。

1) 计算形似度矩阵  $S: s_{ij} = e^{-d_{\text{img}}^2(\mathbf{x}_i, \mathbf{x}_j) / \langle d_{\text{img}} \rangle}$ 。

2) 节点  $\mathbf{x}_i$  和  $\mathbf{x}_j$  间构造边  $e_{ij}$ , 权重为  $w_{ij} = \begin{cases} s_{ij}, & s_{ij} > t \\ 0, & \text{其他} \end{cases}$ ;  $t$  取  $s_{ij}$  的均值。

3) 构造加权无向图  $G: G = (E, V)$ ,  $E$  为边集合,  $V$  为节点集合。

### 2.3 基于加权边的链接聚类

链接聚类<sup>[11-13]</sup> 是一种发现重叠社团的聚类算法。该算法将节点间的链接(边)作为聚类对象,计算链接之间的相似度,然后采用凝聚式层次聚类,最终形成一个大社团。根据划分密度选择最佳的簇划分。聚类结果中,节点属于不同的边,因而可以同时属于不同的簇,最终得到的社团可以部分重叠。上述方法构造的网络有加权的,因而需做如下改进。

设向量如下:

$$\mathbf{a}_i = (\bar{A}_{i1}, \bar{A}_{i2}, \dots, \bar{A}_{in}) ; \bar{A}_{ij} = \frac{1}{k_i} \sum_{i' \in n(i)} w_{ii'} \delta_{ij} + w_{ij}$$

其中,  $w_{ij}$  是边  $e_{ij}$  的权重,  $n(i) = \{j | w_{ij} > 0\}$  是节点  $i$  的紧邻集  $k_i = n(i)$ 。如果  $i = j$ ,  $\delta_{ij} = 1$ , 反之为 0。边  $e_{ik}$  和边  $e_{jk}$  间的相似度为:

$$s'(e_{ik}, e_{jk}) = \frac{\mathbf{a}_i \cdot \mathbf{a}_j}{|\mathbf{a}_i|^2 + |\mathbf{a}_j|^2 - \mathbf{a}_i \cdot \mathbf{a}_j} \quad (7)$$

**定义 1 划分密度。**对于有  $M$  条边的图,  $\{P_1, P_2, \dots, P_c\}$  是一个包含  $C$  个子集的边划分。子集  $P_c$  包含  $m_c = |P_c|$  条边和  $n_c = |\bigcup_{e_{ij} \in P_c} \{i, j\}|$  个节点。接着定义:

$$D_c = \frac{m_c - (n_c - 1)}{n_c(n_c - 1)/2 - (n_c - 1)} \quad (8)$$

特别地,当  $n_c = 2$ ,  $D_c = 0$ 。划分密度  $D$  定义为以簇边数  $m_c$  为权重的  $D_c$  的平均值:

$$D = \frac{1}{M} \sum_c m_c D_c = \frac{2}{M} \sum_c m_c \frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)} \quad (9)$$

层次聚类方法不断合并分组直到所有元素均为同一个簇的成员。聚类过程的历史存储到树状图中,该图包含了层级簇结构的所有信息。划分密度有全局最大值,取值范围为  $[-2/3, 1]$ 。当  $D = 1$  时,每个簇均为完全连通的派系,当  $D = -2/3$  时,每个簇均为包含两个互不相连的边。算法描述如算法 3 所示。

### 算法 3 Weighted link based clustering。

1) 计算加权边间的形似度矩阵  $S'$ :

$$s'(e_{ik}, e_{jk}) = \frac{\mathbf{a}_i \cdot \mathbf{a}_j}{|\mathbf{a}_i|^2 + |\mathbf{a}_j|^2 - \mathbf{a}_i \cdot \mathbf{a}_j}$$

2) 使用单链层次聚类对边集进行聚类。

3) 计算划分密度,当  $D$  取最大式,找到最佳的边集划分  $\{P_1, P_2, \dots, P_c\}$ 。

### 2.4 簇代表点的选取与可视化

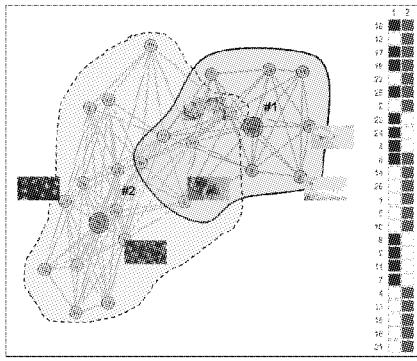
对图像进行链接聚类后,每个簇包含的边互不重叠。现

根据边的归属,将边对应的节点划分到同一簇。然后根据簇内节点的权重和进行排序,找到权重和最大的节点作为该簇的代表点。算法描述如算法4所示。

#### 算法4 Cluster exemplar selection。

- 1) 根据边的归属,重新划分节点的归属:  $\tilde{P}_c = \cup \{\langle x_i, x_j \rangle | e_{ij} \in P_c\}$ 。
- 2) 选择簇代表点:  $E_m = x_i$  where  $\max \left( \sum_{e_{ij} \in c_m} w_{ij} \right)$ ,  $E_m$  代表第  $m$  个簇的代表点。

3) 簇间关系的可视化:根据节点的归属和簇代表点对图



(a) 聚类结果

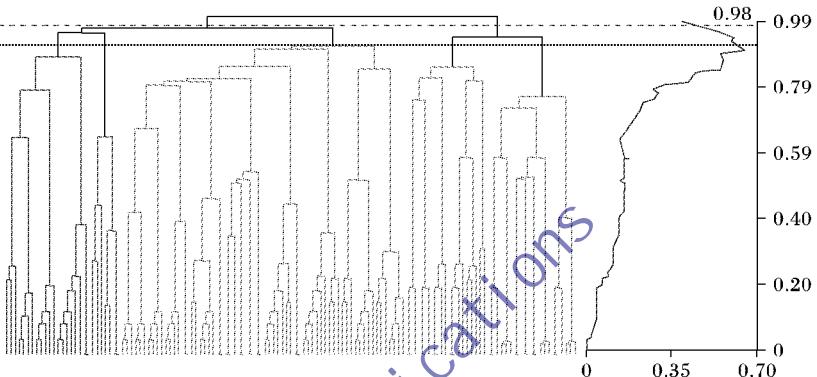
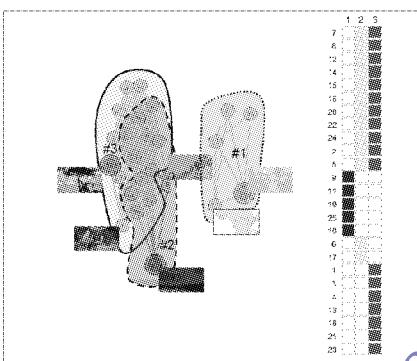


图1 “flower”聚类结果与层次聚类树状图 (edges=133,nodes=25)



(a) 聚类结果

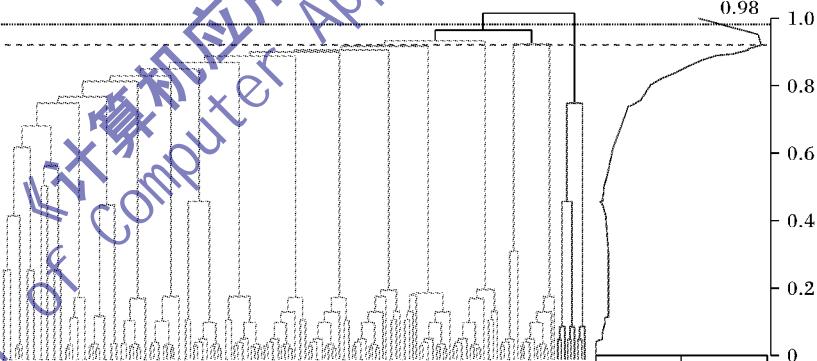


图2 “Afric”聚类结果与层次聚类树状图 (height = 0.92)

图1(a)为flower图像的聚类结果,为2个相互重叠的簇。其中簇1包含12幅图像,簇2包含18幅图像,簇1和簇2重叠区包含5幅图像。可以看出,簇1中的图像背景较浅,且包含花朵较少,簇2中的图像背景较深,且包含花朵较多。簇1和簇2重叠的区域则兼顾了簇1和簇2的特征,即图像背景较浅,但包含的花朵较多。图1(b)为对应的层次聚类树状图,当划分密度取最大值时,聚类结果为5个簇,但簇意义不明确。本实验取  $height = 0.98$ ,聚类结果为2个簇,簇的意义比较明确。

图2(a)为Afric图像的聚类结果,为3个有部分重叠的簇。其中簇1包含5幅图像,簇2包含13幅图像,簇3包含18幅图像,簇2和簇3重叠区包含11幅图像。可以看出,簇3中的图像多为包含动物的风景图,簇2中的典型图片则为纯粹的风景图,簇2和簇3重叠较多,划分边界不明显。簇1中则全是非洲地图。图2(b)为对应的层次聚类树状图,当划分密度取最大值时,  $height = 0.92$ ,聚类结果为3个簇,簇2和簇3的边界不清楚,意义区分不明显。如提升至  $height = 0.98$ ,簇2和簇3合并,则划分为聚类意义非常明确的两个簇。

为定量分析图像的聚类效果,引入社团质量(Community Quality)<sup>[11]</sup>,定义为:

像数据进行可视化。

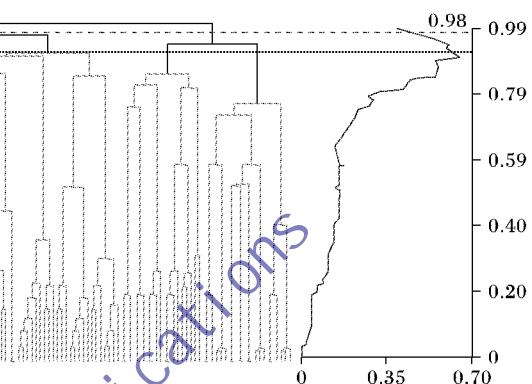
## 3 实验分析

### 3.1 数据集的选取与处理

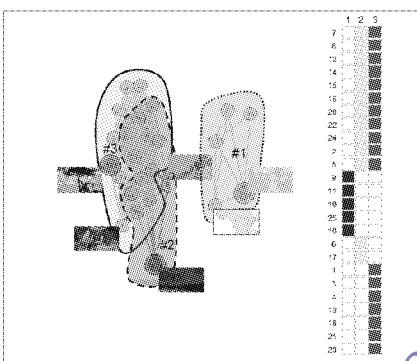
为了检验方法的有效性,选择“flower”、“afric”等几个关键词通过Google进行图片搜索,并取前25张图片作为测试数据集。为便于计算图像间的相似度,对图像进行剪切和缩放处理,确保每个图像的尺寸相同。

### 3.2 实验结果与分析

对处理过的图像使用LIMP进行聚类,结果如图1、2所示。



(b) 层次聚类树状图(edges=133,nodes=25)



(b) 层次聚类树状图(edges=174,nodes=25)

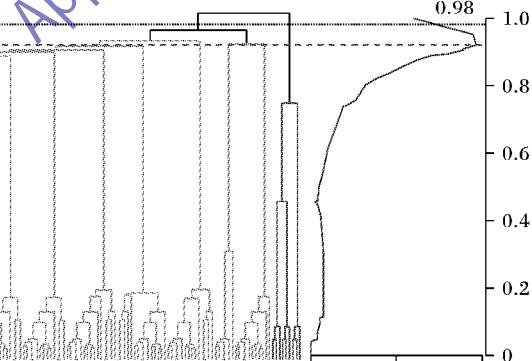


图2 “Afric”聚类结果与层次聚类树状图 (height = 0.92)

$$Q = 1 - \frac{\langle d(i, j) \rangle_{\text{all } i, j \text{ within same community}}}{\langle d(i, j) \rangle_{\text{all pairs } i, j}} \quad (10)$$

$Q$ 值越大,表明社团的结构划分越好。选择flower数据集,基于图像距离(ImgDist) LIMP方法和使用欧氏距离(EucDist)的普通方法对比如图3所示。可以看出,使用基于图像距离(ImgDist) LIMP方法得到的图像的聚类效果较好。

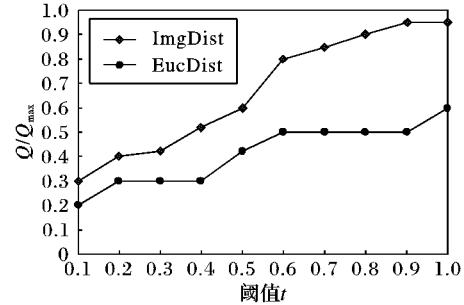


图3 社团质量对比

## 4 结语

本文针对图像聚类中面临的高维、准确度、部分重叠等问题,提出了一种高效的基于链接层次聚类的多标记图像聚类。

该方法通过图像距离计算图像间的相似度，并构造出一个加权无向图。然后，应用基于边相似度的链接层次聚类对加权无向图的边集进行聚类，得到一个基于边集的簇划分。因为一个节点可以属于多个边，从而实现了图像的多重划分。实验结果表明该方法能有效发现具有重叠划分的簇，且簇的意义比较明确。

如能将搜索引擎返回的有关图片的语义信息作为度量图片间相似度一部分，聚类结果的准确度将会进一步提升。目前，通过可视化结果对该方法进行定性分析和简单的定量分析，如何对上述结果进行更深入的定量分析，是下一步要做的工作。另外，实验中并没有采用最大划分密度所对应的 *height* 值进行簇划分，而是人工选取 *height* 值进行划分。也就是说，理论上最大划分密度所对应的簇划分内聚性最好，但实际上有时簇间重叠度过大将导致簇的直观意义不明确，不适合图像的可视化展现。下一步将研究如何在簇内聚性和簇间重叠度之间进行综合度量，结合划分密度和重叠度，自动进行最优的簇划分。

#### 参考文献：

- [1] 韩敏, 范剑超. 单点逼近型加权模糊 C 均值算法的遥感图像聚类应用[J]. 中国图象图形学报, 2009, 14(11): 2333–2340.
- [2] 谷瑞军, 须文波. 基于核方法的彩色图像量化研究[J]. 计算机应用, 2006, 26(9): 2063–2064.
- [3] JIA Y Q, WANG J D, ZHANG C S, et al. Finding image exemplar-susing fast sparse affinity propagation [C]// Proceedings of the 13th Annual ACM International Conference on Multimedia. New York: ACM Press, 2008: 639–642.
- [4] 唐敏, 阳爱民. 一种高效的图像数据库检索方法[J]. 计算机应用, 2008, 28(6): 1454–1456.
- [5] 谷瑞军, 叶宾, 须文波. 基于谱聚类的两阶段颜色量化算法[J]. 中国图象图形学报, 2007, 12(10): 922–925.
- [6] JING Y, BALUJA S. Pagerank for product image search [C]// Proceedings of ACM WWW'08. New York: ACM Press, 2008: 307–316.
- [7] 吴飞, 韩亚洪, 庄越挺, 等. 图像-文本相关性挖掘的 Web 图像聚类方法[J]. 软件学报, 2010, 21(7): 1561–1575.
- [8] CAI D, HE X, LI Z, et al. Hierarchical clustering of WWW image search results using visual, textual and link information [C]// Proceedings of the 12th Annual ACM International Conference on Multimedia. New York: ACM Press, 2004: 952–959.
- [9] 路晶, 马少平. 基于多例学习的 Web 图像聚类[J]. 计算机研究与发展, 2009, 46(9): 1462–1470.
- [10] WANG L, ZHANG Y, FENG J. On the Euclidean distance of images [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1334–1339.
- [11] AHN Y Y, BAGROW J P, LEHMANN S. Link communities reveal multiscale complexity in networks [J]. Nature, 2010, 466(7307): 761–764.
- [12] EVANS T S, LAMBIOTTE R. Line graphs, link partitions and overlapping communities [J]. Physical Review E, 2009, 80(1): 016105.
- [13] KALINKA A T, TOMANCAK P. Linkcomm: An R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type [J]. Bioinformatics, 2011, 27(14): 2011–2012.

(上接第 1093 页)

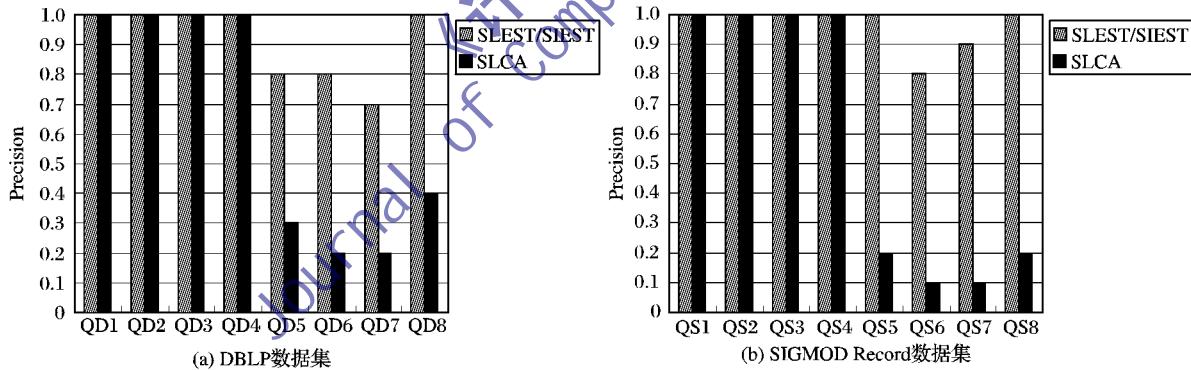


图 4 在 DBLP 和 SIGMOD Record 上的查准率

## 5 结语

本文分析了 LCA 相关概念在 XML 关键字查询中的主要弊端，即不能根据查询关键词之间的语义捕获用户的查询意图，以至于查询结果中存在大量的与查询无关的没有意义的节点。本文借鉴已有文献中提出的以实体子树作为返回结果的思想，提出了最小最低实体子树和最小相关实体子树，由于实体子树比单个节点所携带的信息量多，所提出的算法能很好地捕获到查询关键字之间的 IDREF 语义关系。通过实验证明，本文提出的算法比主流的以 SLCA 作为返回结果的算法的准确率高。

#### 参考文献：

- [1] 李辛. 基于语义相关性的 XML 关键字查询的研究与实现[D]. 北京: 北京交通大学, 2009.

- [2] GUO L, SHAO F, BOTEV C, et al. XRANK: Ranked keyword search over XML documents [C]// Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2003: 16–27.
- [3] LIU Z Y, CHEN Y. Identifying meaningful return information for XML keyword search [C]// Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2007: 329–340.
- [4] 孔令波, 唐世渭, 杨冬青, 等. XML 信息检索中最小子树根节点问题的分层算法[J]. 软件学报, 2007, 18(4): 919–932.
- [5] 吉聪睿, 邓志鸿, 唐世渭. 基于 Nearest Pair 的 XML 关键词检索算法[J]. 软件学报, 2009, 20(4): 910–917.
- [6] XU Y, PAPAKONSTANTINOU Y. Efficient keyword search for smallest LCAs in XML databases [C]// Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2005: 537–538.