doi:10.3969/j.issn.1000-436x.2013.07.018

基于成对约束 Info-Kmeans 聚类的图像索引方法

刘文杰[1]，伍之昂[2]，曹杰[2]，潘金贵[1]
(1. 南京大学 软件新技术国家重点实验室，江苏 南京 210046；2. 南京财经大学 江苏省电子商务重点实验室，江苏 南京 210003)

摘 要：针对图像数据噪声大和高维稀疏的特点，提出了一种基于噪声过滤和 Info-Kmeans 聚类的图像索引构建方法。首先，利用余弦兴趣模式过滤噪声。其次，提出了一种新的 Info-Kmeans 聚类算法，该算法不仅避免 KL-divergence 计算过程中的零值困境问题，还能融合以成对约束出现的先验知识。最后，在 LFW 和 Oxford_5K 2 个图像数据集上的实验表明：噪声过滤能显著提高聚类性能；Info-Kmeans 比已有聚类工具具有更优越的性能。

关键词：图像索引；兴趣模式；噪声过滤；聚类分析
中图分类号：TP181 **文献标识码：**A **文章编号：**1000-436X(2013)07-0159-08

Image indexing method based on clustering via Info-Kmeans under pair constraints

LIU Wen-jie[1], WU Zhi-ang[2], CAO Jie[2], PAN Jin-gui[1]
(1. State Key Lab for Novel Software Technology, Nanjing University, Nanjing 210046, China;
2. Jiangsu Provincial Key Laboratory of E-Business, Nanjing University of Finance and Economics, Nanjing 210003, China)

Abstract: Constructing high-quality content-based image indexing is fairly difficult due to the large amount of noise in the data set and the high-dimension and the sparseness of the image data. To meet this challenge, a novel noise-filtering and clustering was proposed using Info-Kmeans based image indexing construction method. Firstly, a noise-filtering method using the cosine interesting patterns was presented. Secondly, a novel Info-Kmeans algorithm was proposed which could avoid the zero-feature dilemma caused by the use of KL-divergence and exploit the prior knowledge in the form of pair constraints. The experimental results on the two image data sets, LFW and Oxford_5K, well demonstrate that: noise filter can improve the clustering performance remarkably and the novel Info-Kmeans algorithm yields better results than the existing clustering tool.

Key words: image indexing; interesting pattern; noise filtering; cluster analysis

1 引言

图像索引旨在根据图片属性为用户提供一组类似图片的访问，高质量的图像索引能大幅度提升图像获取的效果和性能，图像索引方法可以分为 2 类[1]：1)基于概念的方法，根据图片描述的元数据和属性；2)基于内容的方法，根据图片的内部特征。第 1 种方法依赖于人工描述图片建立元数据，其精确度也依赖于元数据的精确度，而且，Internet 上图片量的急剧增加使人工建立元数据变得愈来愈困难。因此，基于内容的图像索引和获取赢得了越来越多的重视。聚类分析(cluster analysis)能根据数据属性的隐含分布情况将相似对象划分到同一簇，是基于内容的构造图像索引的有力手段。

收稿日期：2012-12-01；**修回日期：**2013-02-17
基金项目：国家自然科学基金资助项目 (71072172,61103229)；江苏省省属高校自然科学研究重大基金资助项目(12KJA520001)；国家科技支撑计划基金资助项目(2013BAH16F01)；国家国际科技合作基金资助项目(2011DFA12910)；江苏省自然科学基金资助项目(BK2010373, BK2012863)
Foundation Items: The National Natural Science Foundation of China (71072172, 61103229); Key Project of Natural Science Research of Jiangsu Provincial Colleges and Universities (12KJA520001); National Key Technologies R&D Program of China (2013BAH16F01); International Science & Technology Cooperation Program of China (2011DFA12910); The Natural Science Foundation of Jiangsu Province (BK2010373, BK2012863)

已有的基于内容的图像分析方法大多基于距离学习，即试图基于先验知识学习距离度量，再利用如贝叶斯、k 近邻等经典方法进行分类。而 Mahalanobis 距离是欧拉距离的泛化形式，所以距离学习过程往往归结为训练一个正定对称矩阵(称为 Mahalanobis 矩阵)，再由式(1)得到 Mahalanobis 距离，其中，m_i 是图像示例，$\boldsymbol{A}$ 是 Mahalanobis 矩阵。

$$d_{\boldsymbol{A}}(m_i,m_j)=(m_i-m_j)^{\mathrm{T}}\boldsymbol{A}(m_i-m_j) \tag{1}$$

现有的距离学习方法包括 LMkNN(large margin k nearest neighbor)[2]、LDML(logistic discriminant based metric learning)[3]、MkNN(marginalized k nearest neighbor)[3]、ITML(information-theoretic metric learning)[4]等。本文的思路则与之不同，试图从图像内容特征出发，利用聚类分析将内容相似的图像划分到同一簇[5,6]，在聚类过程融合先验知识以增强聚类的准确度，最终利用聚类结果为图像构建索引。

利用聚类分析构造高质量图像索引主要面临着两大难题：1)模糊图片、不完整图片、稀有类图片等普遍存在，图像数据噪声大；2)图片往往用 BOF(bag-of-feature)模型表示[7]，一副图片被抽象为虚拟词向量，呈现出高维稀疏性。为此，本文提出了一种基于内容的图像索引构建方法，其总体框架如图 1 所示，首先，利用兴趣模式挖掘方法获得图片数据集上的余弦兴趣模式(CIP, cosine interesting patterns)，不包含任何 CIP 的图像被当作噪声数据去掉；其次，在去除噪声后的数据集上，利用 Info-Kmeans 算法在成对约束条件下进行聚类，类标记即是图像索引值。

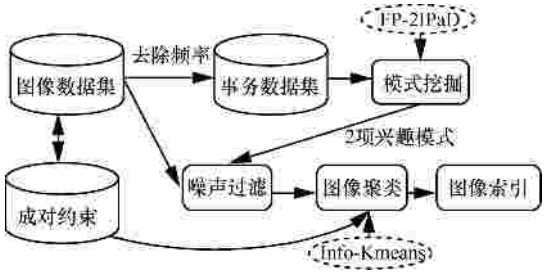

图 1 基于噪声过滤和 Info-Kmeans 聚类的图像索引构建方法总体框架

2 问题描述

本文面向基于 BOF 模型表示的图像数据集，每一行表示一副图片(即一个实例)，由一系列 <*Feature_ID*, *Frequency*> 二元组构成，其中，*Feature_ID* 表示属性 *ID* 号，*Frequency* 表示其出现的次数。数据集的所有属性 *ID* 构成了特征词典，由于特征词典具有高维性，每个实例呈现出稀疏性。设数据集 D 包含 n 个实例：$D=\{x_1,\cdots,x_n\}$，作者试图将其分为 K 个簇，使内容一样的图片被分到同一个簇，然后，以每一簇出现最多数量的图片内容名称命名该簇，即作为索引号。

很多数据集提供了先验知识，如本文实验所使用的 Yahoo 人脸图像数据集，先验知识通常以成对约束的形式出现[3,8]，可分为 must-link 和 cannot-link 2 种约束。如果 2 个实例属于 must-link 约束，那么这 2 个实例在聚类时必须被分配到同一簇中；反之，如果 2 个实例属于 cannot-link 约束，那么这 2 个实例在聚类时必须被分配到不同的簇。为避免 must-link 和 cannot-link 可能存在的冲突，本文的聚类算法仅考虑 must-link 约束。

3 从 Kmeans 到 Info-Kmeans：理论基石

Kmeans 是最著名的聚类方法，将 n 个对象划分为 K 个簇，使得簇内相似度高，而簇间相似度低[8]，Kmeans 以 SSE(sum of squared error)为优化目标，如式(2)所示。

$$E=\sum_{k=1}^{K}\sum_{x\in C_k}dist(x,c_k) \tag{2}$$

其中，*dist* 是欧拉距离，c_k 是类 C_k 的质心。变换不同的距离函数，得到不同类型的 Kmeans 算法，Info-Kmeans 利用 KL-divergence[9]作为距离函数，将 *dist* 函数用 KL-divergence 替代，得到 Info-Kmeans 的优化目标函数为

$$E_{KL}\sum_{k=1}^{K}\sum_{x\in C_k}KL(x\|c_k)$$

$$KL(x\|c_k)\sum_i x_i\log\frac{x_i}{c_{k_i}} \tag{3}$$

其中，x_i 和 c_{k_i} 分别是 x 和 c_k 的属性，高维稀疏数据存在很多缺失属性值，即容易有：$x_i>0$ 且 $c_{k_i}=0$，$KL(x\|c_k)=+8$，这导致难以最小化 E_{KL}，该问题称为零值困境(zero-feature dilemma)。文献[4]阐释了式(3)在概率视图下的含义，即最小化互信息熵的损失，作者借助香农熵获得新的 Info-Kmeans 目标函数来解决零值困境问题。Wu 等人提出 Kmeans 算法距离函数的泛化形式[10]，如式(4)所示。

$$KL(x\|y) = H(y) - H(x) + (x-y)^t \nabla H(y) \quad (4)$$

其中，$H(\cdot)$表示香农熵。于是得到定理 1。

定理 1　Info-Kmean 的目标函数 E_{KL} 等价于式(5)。

$$E'_{KL} = \sum_{k=1}^{K} |c_k| H(c_k) \quad (5)$$

证明　根据式(4) $KL(x\|c_k) = H(c_k) - H(x) + (x-c_k)^t \nabla H(c_k)$，$E_{KL}$ 可以展开为

$$E_{KL} = \sum_{k=1}^{K} \sum_{x \in C_k} KL(x\|c_k)^2 =$$

$$\sum_{k=1}^{K} |c_k| H(c_k) - \sum_{x \in C_k} H(x) -$$

$$\sum_{k=1}^{K} \sum_{x \in C_k} (x-c_k)^t \nabla H(c_k) \overset{\text{def}}{=} E'_{KL} - S_1 - S_2 \quad (6)$$

又因为：$c_k = \sum_{x \in c_k} x$，所以：$\sum_{x \in C_k} (x - c_k) = 0$，$S_2$=0。而 S_1 由数据集对象的固有属性所决定，所以，S_1 为常数。因此，目标函数 E_{KL} 等价于 E'_{KL}，证毕。

E'_{KL} 是本文提出的 Info-Kmeans 算法的目标函数，可以看出，E'_{KL} 仅与 K 个簇聚集的香农熵有关，避免了式(3)计算 KL-divergence 的零值困境问题。

4　基于噪声过滤和 Info-Kmeans 聚类的图像索引构建方法

4.1　基于兴趣两项集挖掘的噪声过滤

众所周知，频繁模式是指支持度大于阈值的项集，频繁模式未必是用户真正感兴趣的[9,11]，同时，兴趣模式的支持度也未必很高。因此，兴趣模式挖掘，尤其是低支持度的兴趣模式，越来越受到研究者的重视。以往研究已经提出很多衡量兴趣度的指标，包括[12]：Cosine、Interest、Information Gain、Kappa、f-coefficient 等，在众多兴趣指标中，余弦相似度以优良的数学性质得到了更多的关注和应用，比如，余弦相似度满足对称性、空值不变性(null-invariance)[13]以及反交叉支持(anti-cross-support)[14]模式。

本文选择余弦相似度来衡量产生的频繁项集的兴趣度，余弦最初用于衡量 2 个向量的夹角，给定 2 个向量 $\boldsymbol{A}$ 和 $\boldsymbol{B}$，$\cos(\boldsymbol{A},\boldsymbol{B}) = \frac{<\boldsymbol{A},\boldsymbol{B}>}{\|\boldsymbol{A}\|\|\boldsymbol{B}\|}$，其中，“$<>$”表示 2 个向量的内积，“$\|\ \|$”表示向量的值。对于事物数据集，$i_1$ 和 i_2 是 2 个项，$\cos(\{i_1,i_2\}) = \frac{f_{11}}{\sqrt{f_{1+} f_{+1}}}$，$f_{11}$ 是同时包含 i_1 和 i_2 的事务个数，f_{1+}和 f_{+1} 分别是包含 i_1 和 i_2 的事务个数，因此可以用支持度来表达二项集 X={i_1, i_2}的余弦兴趣度。

$$\cos(X) = \frac{supp(X)}{\sqrt{supp(\{i_1\})supp(\{i_2\})}} \quad (7)$$

根据式(7)，容易将二项集推广到多项集，令多项集 X={i_1,…, i_K}, K≥2，可得

$$\cos(X) = \frac{supp(X)}{\sqrt[k]{\prod_{k=1}^{K} supp(\{i_k\})}} \quad (8)$$

根据式(8)，余弦兴趣模式可以定义如下。

定义 1　余弦兴趣模式。令 I 是属性的集合，$J=2^I$ 是 I 的幂集，min_supp 和 min_cos 分别是支持度和余弦相似度阈值，余弦兴趣模式集合定义为

$$F = \{X \subseteq J \,|\, supp(X) \geq min\_supp, \cos(X) \geq min\_cos\} \quad (9)$$

由定义 1 可以看出，余弦兴趣模式实际上是在频繁模式的基础上添加了项集余弦兴趣度的约束。

定义 2　噪声实例。数据集 D={x_1,…,x_n}，将二元组<*Feature_ID, Frequency*>中的 *Frequency* 去掉之后，每个实例 $x_i \in D$ 都能使用 x_i={ w_1^i, w_2^i,…, $w_{|x_i|}^i$} 表示，若 x_i 为噪声，当且仅当：$\forall p_j \in F, p_j \notin X_i$，其中，$F$={$p_1$, p_2,…,p_k}为余弦兴趣模式集合。

由定义 2 可知，噪声实例指不包含任何兴趣项集的实例，显然，只需要得到最短余弦兴趣模式，即兴趣两项集，就可确定噪声数据。

获得余弦兴趣模式最直观的方法是两阶段挖掘：首先，利用经典的 Apriori 或 FP-growth 算法挖掘频繁模式；其次，在频繁模式中选择出兴趣度大于阈值的模式。但是，这种两阶段挖掘方法很难发现低支持度的兴趣模式，因为随着支持度阈值的降低，频繁模式数量急剧增加，Apriori 和 FP-growth 算法时空消耗甚巨。余弦相似度不满足反单调性，难以直接运用于剪枝，幸运的是，以前研究证明了余弦相似度满足条件反单调性[15]，条件反单调性定义如下。

定义 3　条件反单调性。令 I 是属性的集合，$J=2^I$ 是 I 的幂集，度量 M 满足条件反单调性，当且

仅当 $\forall X,Y\in J$，1) $X\subseteq Y$；2)如果 $Y\setminus X\neq\varnothing$ 并且 $\forall i\in X,i'\notin Y\setminus X$，有 $supp(\{i\})\geq supp(\{i'\})$，即必有 $M(X)\geq M(Y)$。

定理 2 余弦相似度满足条件反单调性。

证明 任取多项集 $X=\{i_1,i_2,\cdots,i_k\}$ 及其超集 $Y=\{i_1,\cdots,i_k,\cdots,i_{k+m}\}$，$supp(i_1)\geq supp(i_2)\geq\cdots\geq supp(i_{k+m})$，$m\geq 1$，可得

$$\cos(X)=\frac{supp(X)}{\sqrt[k]{\prod_{j=1}^{k}supp(\{i_j\})}}\geq\frac{supp(Y)}{\sqrt[k]{\prod_{j=1}^{k}supp(\{i_j\})}}\geq$$

$$\frac{supp(Y)}{\sqrt[k+m]{\prod_{j=1}^{k+m}supp(\{i_j\})}}\cos(Y)$$

由定义 3 可知，余弦兴趣模式满足条件反单调性，证毕。

当 FP-Tree 按照支持度递减次序构建时，每条 FP-Tree 路径上的项都按支持度递减，而 FP-growth 算法正是从叶节点向上遍历，因此，一旦发现多项集的余弦兴趣度小于阈值，就再没必要继续向上挖掘，因为该多项集超集的余弦兴趣度也必然小于阈值，余弦相似度就跟支持度一样起到了剪枝作用。

同时，本文的噪声过滤方法是将包含任意模式的实例保留，那么，包含多项集的实例一定包含该多项集的所有两项子集。如果将多项集 Y 中项按支持度递减次序排列，得到：$Y=\{i_1,i_2,\cdots,i_k\}$，且 $supp(i_1)>supp(i_2)>\cdots>supp(i_k)$，由条件反单调性，$\cos(\{i_{k-1},i_k\})>\cos(Y)$，所以，两项集 $\{i_{k-1},i_k\}$ 一定是兴趣模式，而且在噪声过滤问题中，两项集 $\{i_{k-1},i_k\}$ 足以代表其超集 Y。此外，仅挖掘兴趣两项集也大幅度提高了算法效率。作者基于 FP-growth 算法提出 FP-2IpaD (FP-growth based 2-itemsets interesting pattern discovery)算法，其伪代码如图 2 所示。

第 1)~7)行为单前缀路径的 FP-Tree 挖掘，与经典 FP-Growth 算法不同的是，FP-2IPaD 仅考虑两项子集，这样使时间复杂度降低到 $O(C_{|P|}^{2})$。第 8)~20)行为多路径部分，在 FP-2IPaD 的递归调用过程中，不断产生新的两项频繁项集。在第 14)~18)行，如果新的项集产生，就将它加入到 F 中。

4.2 基于成对约束的 Info-Kmeans 聚类

本文所提出的 Info-Kmeans 聚类算法将 n 个实例划分到 K 个簇，使得目标函数 E'_{KL} 最小，本质上，该优化过程属于 NP_hard 问题，作者基于随机抽样方法，每次都将使得 E'_{KL} 下降最多的簇标号赋给实例，经过上述贪心过程的反复迭代，试图得到最优解。同时，对于受 must-link 约束的实例，Info-Kmeans 算法为该 must-link 的实例集合寻求局部最优簇标记。

```
输入: Tree:数据集 D 的 FP-Tree; min_supp: support 阈值;
      min_cos: cosine 阈值
输出: F: 余弦兴趣两项集
FP-2IPaD(Tree,a, min_supp, min_cos)
1) if Tree 包括单个路径
2)    设 P 为 Tree 的单路径部分，Q 为 Tree 的多路径部分；
3)    for 路径 P 中所有的两项集组合
4)       if cos(ß∪a)≥min_cos
5)          Add ß∪a to F;
6)       end if
7)    end for
8) else if Q?null
9)    for Q 中的每一项 a_i
10)      产生模式 ß= a_i∪a 满足 supp(ß)=supp(a_i);
11)      if a = null
12)         构建 ß 的条件模式基及 ß 的条件 FP 树 Tree_ß;
13)      end if
14)      if Tree_ß?null
15)         call FP-2IPaD(Tree_ß, ß, min_supp, min_cos);
16)      else if (|ß|=2 && cos(ß)≥min_cos)
17)         Add ß to F
18)      end if
19)   end for
20) end if
```

图 2 FP-2IPaD 算法

算法 2 描述了 Info-Kmeans 算法的主过程，算法 3 描述了 GroupAssign 子过程。利用并查集建立 must-link 约束关系，$father[n]$表示实例的父亲结点，布尔数组 $head[n]$表示实例是否为根。图 3 的 3)~15)行是 n 次随机抽样，第 5)行和第 6)行分别判断当前实例 x_j 是否受到 must-link 约束及是否为根结点，第 7)行找到与 x_j 在同一簇的所有实例，第 8)行调用 GroupAssign 子过程为该 must-link 实例集合找到局部最优簇标记。若当前实例 x_j 不受到 must-link 约束，第 12)行调用 GroupAssign 子过程仅为 x_j 赋簇标记。Info-Kmeans 算法有 2 个收敛条件：1) n 个实例的簇标记不再发生变化(见 16)~18)行)；2)迭代轮次达到 $maxIter$(第 2)行)。

```
输入: D:数据集; K:簇的数目; maxIter:最大迭代轮数
输出: Val*(E'KL):目标函数值; Lab*[n]: n 个实例的簇标号
Info-Kmeans(D,K,maxIter)
1) 读取并初始化数据集 D;
2) for i ? 1 : maxIter
3)    for j ? 1 : n
4)       xj? RandomSampling(D);
5)       if father[xj] ?- 1
6)          if head[xj] = 1
7)             instances[num] ? findConstraints(xj);
8)             minc ? GroupAssign(instances, num, Lab[n]);
9)             minc? instance[num]更新 Val(E'KL);
10)         end if
11)      else
12)         minc ? GroupAssign(xj, 1, Lab);
13)         minc? xj, 更新 Val(E'KL);
14)      end if
15)   end for
16)   if Lab[D]未发生变化
17)      break;
18)   end if
19) end for
```

图 3　Info-Kmeans **主算法**

GroupAssign **子过程能为一个或多个实例找到使得** E'_{KL} **下降最多的簇标号，核心在于图** 4 **第** 5) **行** $?_k(instances[j])$**的计算上，**$?_k(instances[j])$**表示实例** x **从原来簇(记为** c'_k **)移到目标簇(记为** c_k **)时，**E'_{KL} **的变化量，如式**(10)**所示。**

$$?_k(x') = E'_{KL}(\text{new}) - E'_{KL}(\text{old}) =$$
$$(|c'_k|-1)H\left(\frac{\sum_{x\in c'_k} p(Y|x) - p(Y|x')}{|c'_k|-1}\right) - |c'_k|H\left(\frac{\sum_{x\in c'_k} p(Y|x)}{|c'_k|}\right) +$$
$$(|c_k+1|)H\left(\frac{\sum_{x\in c_k} p(Y|x) + p(Y|x')}{|c'_k|+1}\right) - |c_k|H\left(\frac{\sum_{x\in c_k} p(Y|x)}{|c_k|}\right) \quad (10)$$

```
输入: inst:实例数组; num:正对约束的实例数目; Lab[n]: n 个
      实例的当前簇标号
输出: minc:最优簇标号
GroupAssign(inst, num, Lab[n])
1) objVarOld ? 0;
2) for k ? 1: K
3)    ? objVal(k) ? 0;
4)    for j ? 1: num
5)       ? objVal(k) ? ? objVal(k)+? k(inst[j]);
6)    end for
7)    if ? objVal(k) < objVarOld
8)       objVarOld ? ? objVal(k), minc? k;
9)    end if
10) end for
11) return minc
```

图 4　GroupAssign **子过程**

5　实验及分析

5.1 实验设计与数据集

数据集　**实验中采用** 2 **个数据集的描述信息，表** 1 **给出了** 2 **个数据集的特征描述。人脸图像数据集** LFW [注1](labeled faces in the wild)**被用来研究人面识别。该数据集包含** 13 233 **张从互联网上搜集的人脸图像，其中** 1 680 **个人的图片出现了** 2 **次及** 2 **次以上。作者的实验利用到的** LFW View1 [注2]**是随机将整个数据集分为** 10 **份，每份数据集各包含** 110 **对正对约束。将其中的** 9 **份作为训练集，** 1 **份作为测试集。牛津地标数据集** Oxford_5K [注3]**是从** Flickr **上将** 11 **个特殊的牛津地标名作为关键字搜索得来的。该数据集中包括** 11 **个不同的地标，共** 5 062 **张图像。值得注意的是，** 2 **个数据的特征并不相同，**LFW **抽取人脸上** 9 **个特征点(如图** 5 **所示)，**3 456 **维特征向量由** 9 (point)×3 (scale)×8 (direction)×16(SIFT vector)**构成；而** Oxford_5K **则依据** BOF **模型生成的一般过程**[16]**，先获取图像分片，再学习虚拟词典**(visual vocabulary)**，该数据集的虚拟词典维数是** 658 346**，最后将图片用虚拟词典中的虚拟词**(visual words)**表达，得到高维稀疏向量。**

图 5　**人脸图像上的** 9 **个特征点**

表 1　　**实验数据集描述特征**

数据集	实例数量	特征数量	簇数量	密度%
LFW	13 233	3 456	**未知**	11.886
Oxford_5K	5 062	658 346	11	0.022 8

注1　http://vis-www.cs.umass.edu/lfw/。
注2　http://vis-www.cs.umass.edu/lfw/#views。
注3　http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/。

工具　作者使用了FP-2IPaD算法、Info-Kmeans聚类算法及CLUTO[注4]。基于Bolgelt提供的FP-growth算法开源代码，利用C语言实现了用于噪声过滤的两项兴趣模式挖掘算法FP-2IPaD。作者还利用C语言实现了4.2节提出的Info-Kmeans聚类算法，并用Python脚本完成数据处理。为论证所提出的聚类算法的优势，将其与Karypis实验室开发的著名高维数据聚类工具CLUTO做比较，CLUTO基于传统Kmeans方法，以余弦作为距离函数[17]。

评价指标　实验中采用NMI作为评价指标。NMI(normalized mutual information)是评价聚类性能最常用的指标[18,19]，本文实验沿用NMI指标，如式(11)所示。

$$NMI\left(\frac{I(K,L)}{\sqrt{H(K)H(L)}}\right) \tag{11}$$

其中，K为簇数组，L是正确的簇标记。*NMI*取值范围是[0,1]，*NMI*越大说明聚类效果越好。当然，评价聚类性能的指标很多，比如熵、纯度、互信息、Jaccard系数等，文本聚类上的结果表明这些指标呈现出相似趋势[20]，*NMI*足以评价聚类质量。

5.2　实验结果分析

5.2.1　噪声过滤效果测试

首先探究噪声过滤对图像聚类的影响。实验选取的2个数据集都包含大量噪声数据：1) Oxford_5K是由17个关键词从Flickr中查询得到，由于有的关键词含义过于宽泛(如：Oxford和New Oxford等)，使得Oxford_5K数据集包含了大量与牛津地标图片无关的噪声数据，图6给出了正常图片与噪声图片的例子，可以看出，很多噪声图片与牛津地标风马牛不相及；2) LFW数据集包含很多只有一幅人脸图像的人，如图7所示，5 749个人中有4 069人仅有一幅人脸图像，这种稀有类将严重影响聚类性能[21]。

作者设置不同的支持度和余弦兴趣度门槛得到不同的实验场景，然后利用兴趣模式过滤噪声，再分别用Info-Kmeans和CLUTO两种工具对去噪后的数据聚类，计算*NMI*指标，表2给出了2个数据集8个场景的结果。场景1是原始数据集，2种工具得到的*NMI*值都极低，说明噪声数据确实极大地降低了聚类性能。利用余弦兴趣模式去噪声之后，*NMI*值明显升高，这说明本文提出的噪声过滤方法能有效地提高聚类性能。图7比较了LFW数据集上去噪声前后人脸图像数量分布情况，可以看出，利用场景3的参数设置进行噪声过滤后，仅有一幅人脸图像的人数急剧降低，说明噪声过滤能将大部分影响聚类性能的稀有类去除，这也正是聚类性能得到提升的原因。

图6　牛津地标数据集上地标图像与噪声

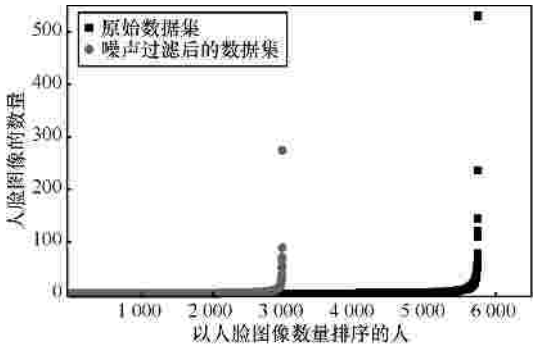

图7　人脸图像数据集上去噪声前后人脸图像数量分布

5.2.2　Info-Kmeans和CLUTO聚类性能比较

本节将所提出的Info-Kmeans与CLUTO进行聚类性能上的比较，CLUTO是针对高维稀疏数据设计的著名聚类工具，在文本聚类上表现出卓越的性能[17]，实验中CLUTO参数设置为：*clmethod*=direct，*sim*=cosine，*ntrials*=1，*colmodel*= none。将Info-Kmeans和CLUTO分别重复聚类10次，取平

注4　http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview。

表 2　不同参数设置下的实验结果 Dataset

数据集	场景	参数设置		模式数量	图像数量	*NMI*	
		min_supp	*min_cos*			Info-Kmeans	CLUTO
Oxford_5K (K=11)	场景 1	—	—	—	5 060	0.185	0.157
	场景 2	0.08%	0.45	3 553	1 352	0.594	0.534
	场景 3	0.09%	0.47	3 383	1 009	0.625	0.483
	场景 4	0.10%	0.48	839	559	0.397	0.331
LFW (K=800)	场景 1	—	—	—	13 233	0.176	0.184
	场景 2	20%	0.50	907	11 389	0.801	0.678
	场景 3	35%	0.55	321	6 235	0.846	0.692
	场景 4	40%	0.60	21	3 653	0.885	0.703

均 *NMI* 指标值。表 2 给出不同实验场景下 2 种工具所得到的 NMI 对比，可以看出，Info-Kmeans 的聚类效果优于 CLUTO。

由于 LFW 的簇数目未知，作者改变簇数目 K，比较了 Info-Kmeans 和 CLUTO 在 Case2 和 Case4 的聚类性能，结果如图 8 所示，Info-Kmeans 聚类性能仍然优于 CLUTO，随着 K 的增大，*NMI* 增大，原因在于人数量远大于 K，随着簇数目的增加，同个人的多幅人脸图像更容易被划分到同一簇，而 K 较小时，人脸形似的不同人图片容易被分到同一簇。

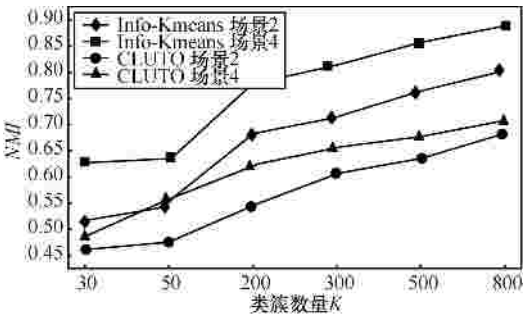

图 8　人脸图像数据集上聚类算法性能对比

对于 Oxford_5K 数据集，簇数目 K 已知，聚类算法往往受到初始随机分配簇标记的影响，聚类结果具有随机性。因此，作者比较了 Info-Kmeans 和 CLUTO 聚类的稳定性，结果如图 9 所示，图中 Info-C2 表示 Info-Kmeans 的场景 2，图中竖线表示 10 次聚类结果中的最小 *NMI* 值和最大 *NMI* 值，与竖线垂直的横线表示平均值。可以看出，Info-Kmeans 比 CLUTO 具有更好的稳定性，这是由 Info-Kmeans 多轮的随机抽样和迭代优化所决定的。

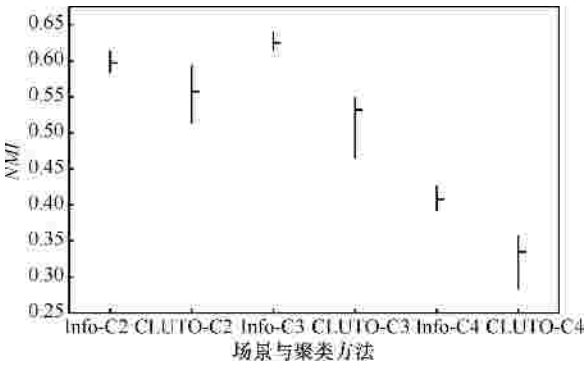

图 9　牛津地标数据集上聚类算法稳定性对比

5.2.3　基于聚类结果识别图像的例子

基于 Info-Kmeans 得到的聚类结果，如果以每个簇出现最多图片的名字来命名该簇，就能获得每幅图片的标记，LFW 数据集上的标记是人名，而 Oxford_5K 数据集上的标记是地标建筑名称。图 10 给出 Oxford_5k 数据集上 2 个簇的例子，矩形框内表示识别打错标记的图片，以 All Souls 为例，表示该簇中共有 63 幅图片，其中 59 幅是 All Souls 的图片。结合 Info-Kmeans 聚类结果的 *NMI* 指标，Info-Kmeans 能够将大部分同个地标的图片划分到同一簇，则可以用该簇中出现最多的图片来命名该簇，从而为同一簇中的所有图片建立索引，满足图片查询获取及其他应用的需求。

图 10 Oxford_5k 数据集上 Info-Kmeans 获得的类簇例子

6 结束语

本文提出了一种基于噪声过滤和 Info-Kmeans 聚类的图像索引构建方法。首先，提出基于 FP-Tree 的两项余弦兴趣模式挖掘算法，利用余弦兴趣模式过滤噪声。其次，提出一种新的 Info-Kmeans 聚类算法，将 KL-divergence 计算等价变换为香农熵的增量计算，从而避免了 KL-divergence 计算过程中的零值困境问题，同时，该算法融合了以成对约束出现的先验知识。最后，在 LFW 和 Oxford_5K 两个图像数据集上的实验表明：1) 噪声过滤能显著提高聚类性能；2) Info-Kmeans 比著名聚类工具 CLUTO 具有更优越的性能。

在今后的工作中主要是将这种基于内容的图像索引构建方法扩展到社会网络分析领域，面向更加丰富的社会媒体(social media)及社区进行根据内容特征的聚类分析。

参考文献：

[1] YANG C W, SHEN J J. Recover the tampered image based on VQ indexing[J]. Signal Processing, 2010, 90(1):331-343.

[2] WEINBERGER K, BLITZER J, SAUL L. Distance Metric Learning for Large Margin Nearest Neighbor Classification[M]. Cambridge: MIT Press, 2006.

[3] GUILLAUMIN M, VERBEEK J, SCHMID C. Is that you? metric learning approaches for face identification[A]. Proceedings of the International Conference on Computer Vision[C]. Kyoto,2009. 498-505.

[4] DAVIS J, KULIS B, JAIN P, *et al.* Information-theoretic metric learning[A]. Proceedings of the 24th International Conference on Machine Learning[C]. New York, USA, 2007. 209-216.

[5] CAO J, WU Z A, WU J J, *et al.* Towards information-theoretic *K*-means clustering for image indexing[J]. Signal Process ng, 2013, 93(7):2026-2037.

[6] **李国波，陈钢，吴百锋. 基于特征聚类的图像错误检测及掩盖算法**[J]. **通信学报**, 2010, 31(12):1-11.
LI G B, CHEN G, WU B F. Error detection and concealment based on characteristic clustering of image[J].Journal on Communica ions, 2010, 31(12):1-11.

[7] YANG J C, YU K, GONG Y, *et al.* Linear spatial pyramid matching using sparse coding for image classification[A]. Proceedings of the IEEE Int'l Conf on Computer Vision and Pattern Recognition[C]. Miami,USA, 2009. 1794-1801.

[8] **尹学松，胡恩良，陈松灿. 基于成对约束的判别型半监督聚类分析**[J]. **软件学报**, 2008, 19(11):2791-2802.
YIN X S, HU E L, CHEN S C. Discriminative semi-supervised clustering analysis with pairwise constraints[J]. Journal of Software, 2008, 19(11):2791-2802.

[9] TAN P N, STEINBACH M, KUMAR V. Introduction to Data Mining[M]. Boston: Addison-Wesley, 2005.

[10] WU J J, XIONG H, CHEN J, *et al.* A generalization of proximity functions for *K*-means[A]. Proceedings of the IEEE International Conference on Data Mining[C]. Omaha, 2007. 361-370.

[11] **付爱英，曾勋炜，徐知海等. 基于聚类的关联规则挖掘算法的研究及应用**[J]. **通信学报**, 2006, 27(z1):177-180.
FU A Y, ZENG Q W, XU Z H, *et al.* Research and application of the algorithms for mining association rules based on clustering[J]. Journal on Communications, 2006, 27(z1):177-180.

[12] WU J J, ZHU S W, XIONG H, *et al.* Adapting the right measures for pattern discovery: a unified view[J]. IEEE Trans on Systems, Man, and Cybernetics, 2012, 42(4):1203-1214.

[13] TAN P N, KUMAR V, SRIVASTAVA J. Selecting the right interestingness measure for association patterns[A]. Proceedin of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. Edmonton, 2002. 32-41.

[14] XIONG H, TAN P N, KUMAR V. Hyperclique pattern discovery[J]. Data Mining and Knowledge Discovery Journal, 2006, 13(2): 219-242.

[15] WU J J, ZHU S W, LIU H F, *et al.* Cosine interesting pattern discovery[J]. Information Sciences, 2012, 184(1):176-195.

[16] LI F F, PIETRO P. A Bayesian hierarchical model for learning natural scene categories[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. San Diego, USA, 2005. 524-531.

[17] CLUTO a clustering toolkit [EB/OL]. www.cs.umn.edu/~cluto, 2012.

[18] ZHONG S, GOHOSH J. Generative model-based document clustering: a comparative study[J]. Knowledge and Information Systems, 2005, 8(3):374-384.

[19] VINH N X, EPPS J, BAILEY J. Information theoretic measures for clusterings comparison: is a correction for chance necessary?[A]. Proceedings of the International Conference on Machine Learning[C]. Montreal, 2009. 1073-1080.

（下转第 173 页）

2nd International Symposium on Intelligence Information Processing and Trusted Computing[C]. 2011.139-142.

[4] **黄海生，王汝传. 基于隶属云理论的主观信任评估模型研究**[J]. **通信学报**, 2008, 29(4):13-19.
HUANG H S, WANG R C. Subjective trust evaluation model ba d on membership cloud theory[J]. Journal on Communication, 2008, 29(4): 13-19.

[5] **李致远，王汝传**. P2P **电子商务环境下的动态安全信任管理模型**[J]. **通信学报**, 2011, 32(3):50-59.
LI Z Y, WANG R C. Dynamic secure trust management model for P2P e-commerce environments[J]. Journal on Communications, 2011, 32(3): 50-59.

[6] MA S N, HE J S, GAO F. A trust quantification method based on grey fuzzy theory[A]. International Conference on Security of Information and Networks[C]. 2010.27-31.

[7] **陈超，王汝传，张琳. 一种基于开放式网络环境的模糊主观信任模型研究**[J]. **电子学报**, 2010, 38(11):2505-2509.
CHEN C, WANG R C, ZHANG L. The research of subjective ust model based on fuzzy theory in open networks[J]. Acta lectronica Sinica, 2010, 38(11):2505-2509.

[8] **朱友文，黄刘生，陈国良等. 分布式计算环境下的动态可信度评估模型**[J]. **计算机学报**, 2011, 34(1):55-64.
ZHU Y W, HUANG L S, CHEN G L, *et al*. Dynamic trust evaluation model under distributed computing environment[J]. Chinese Journal of Computers, 2011, 34(1):55-64.

[9] **蒋黎明，张宏，张琨. 开放系统中一种基于模糊修正的证据信任模型**[J]. **电子与信息学报**, 2011, 33(8):1930-1936.
JIANG L M, ZHANG H, ZHANG K. An evidential trust model with fuzzy adjustment method for open systems[J]. Journal of Electronics & Information Technology, 2011, 33(8):1930-1936.

[10] **杨凯，马建峰，杨超. 无线网状网中基于** D-S **证据理论的可信路由**[J]. **通信学报**, 2011, 32(5):89-96.
YANG K, MA J F, YANG C. Trusted routing based on D-S evidence theory in wireless mesh network[J]. Journal on Communications, 2011, 32(5):89-96.

[11] **田春岐，邹仕洪，王文东. 一种新的基于改进型** D-S **证据理论的** P2P **信任模型**[J]. **电子与信息学报**, 2008, 30(6):1480-1484.
TIAN C Q, ZOU S H, WANG W D. A new trust model based on advanced D-S evidence theory for P2P networks[J]. Journal of Electronics & Information Technology, 2008, 30(6):1480-1484.

[12] JIANG L, XU J, ZHANG K. A new evidential trust model for open distributed systems[A]. Expert Systems with Applications[C]. 2012. 3772-3782.

[13] **张琳，王汝传，张永平. 一种基于模糊集合的可用于网格环境的信任评估模型**[J]. **电子学报**, 2008, 36(5):862-868.
ZHANG L, WANG R C, ZHANG Y P. A trust evaluation model based on fuzzy set for grid environment[J]. Acta Electronica Sinica, 2008, 36(5):862-868.

[14] **王万森. 人工智能原理及其应用**[M]. **北京: 电子工业出版社**, 2006.
WANG W S. Artificial Intelligence Principle and Application[M]. Beijing: Electronic Industry Press, 2006.

作者简介:

张琳（1980-），女，江苏丰县人，博士后，南京邮电大学副教授、硕士生导师，主要研究方向为网格计算、网络安全、信任、可信计算等。

刘婧文（1989-），女，江苏连云港人，南京邮电大学硕士生，主要研究方向为服务计算、信息安全、信任、可信计算等。

王汝传（1943-），男，安徽合肥人，南京邮电大学教授、博士生导师，主要研究方向为计算机软件、计算机网络和网格、信息安全、无线传感器网络、移动代理和虚拟现实技术等。

王海艳（1974-），女，江苏盐城人，南京邮电大学教授、硕士生导师，主要研究方向为信息安全、计算机软件、可信计算等。

（上接第 166 **页）**

[20] CAO J, WU Z A, WU J J, *et al*. SAIL: summation-based incremental learning for information-theoretic text clustering[J]. IEEE Trans on Cybernetics, 2013, 43(2):570-584.

[21] WU J J, XIONG H, CHEN J. COG: local decomposition for rare class analysis[J]. Data Mining and Knowledge Discovery, 2010, 20(2): 191-220.

作者简介：

刘文杰（1988-），男，湖北黄石人，南京大学硕士生，主要研究方向为数据挖掘、多媒体技术。

伍之昂[通信作者]（1982-），男，江苏宜兴人，博士，南京财经大学副教授，主要研究方向为网络计算、数据挖掘和推荐系统。E-mail: zawuster@gmail.com。

曹杰（1969-），男，江苏姜堰人，博士，南京财经大学教授、博士生导师，主要研究方向为商务智能和数据挖掘。

潘金贵（1952-），男，江苏南京人，博士，南京大学教授、博士生导师，主要研究方向为多媒体技术。