基于关系矩阵融合的多媒体数据聚类

张 鸿[1)] 吴 飞[2)] 张晓龙[1)]

[1)](武汉科技大学计算机科学与技术学院 武汉 430081)

[2)](浙江大学计算机科学与技术学院 杭州 310027)

摘 要 针对目前多媒体聚类研究中如何挖掘和利用不同数据集之间统计关系的问题,提出一种基于关系矩阵融合的聚类方法,首先,对图像和音频数据集中提取的特征矩阵进行相关性分析和子空间映射,进而在全局范围内对图像相似度、音频相似度以及图像和音频的相关度进行融合与优化,最后,采用基于相似度的循环迭代算法进行图像和音频聚类.对比实验从多个角度验证了文中方法的有效性,并能较好地应用于多媒体交叉检索.

关键词 视听觉特征;关系矩阵;多媒体数据聚类;相关性融合;交叉检索

中图法分类号 TP391 **DOI 号**: 10.3724/SP.J.1016.2011.01705

Multimedia Data Clustering Based on Correlation Matrix Fusion

ZHANG Hong[1)] WU Fei[2)] ZHANG Xiao-Long[1)]

[1)](*College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan* 430081)

[2)](*College of Computer Science and Technology, Zhejiang University, Hangzhou* 310027)

Abstract It is a hot issue to explore statistical correlation between different types of multimedia data, especially in the area of multimedia clustering. In this paper, we propose a multimedia clustering method based on correlation matrix fusion. Visual and auditory feature matrices are firstly initialized and simultaneously mapped into a subspace; Then we utilize correlation fusion strategy on image similarity matrix, audio similarity matrix and image-audio correlation matrix for global reinforcement and optimization; Thirdly, similarity-based clustering method is implemented for image and audio clustering in the subspace. Experiment results are encouraging and show that the performance of our approach is effective. Besides, an interesting experiment of image-audio cross-retrieval validates the applicability of our approach.

Keywords visual-auditory feature; correlation matrix; multimedia data clustering; correlation fusion; cross-retrieval

1 引 言

随着多媒体技术和信息技术的高速发展,Web 页面上各种类型的多媒体数据(如图像、音频、视频)正迅速膨胀,同时,由于多媒体数据本身具有底层视听觉特征异构、高层语义丰富的特点[1],使得对 Web 多媒体数据的有效管理和智能利用十分困难.大量研究工作从数据存储、信息检索、知识挖掘等不同角度关注上述问题,其中,多媒体聚类研究通过机器学习、统计分析等方法,帮助理解多媒体数据集的潜在语义[2],该项研究主要以底层内容特征来表达

收稿日期:2008-11-23;最终修改稿收到日期:2011-08-21.本课题得到国家自然科学基金(61003127,61070068)、湖北省教育厅科学技术研究项目(Q20091101)和武汉科技大学科学基金项目(2008TD04)资助.**张 鸿**,女,1979 年生,博士,副教授,主要研究方向为多媒体内容分析、机器学习、跨媒体检索.E-mail: zhanghong_zju@yahoo.com.cn. **吴 飞** 男,1973 年生,博士,教授,主要研究领域为多媒体分析与检索、统计学习理论. **张晓龙**,男,1963 年生,博士,教授,主要研究领域为机器学习、数据挖掘等.

多媒体样本,学习样本向量在特征空间中的几何结构,从而分析潜在的语义关系,实现多媒体数据集的聚类.

传统的多媒体聚类研究大多是针对单一类型的多媒体数据,如图像聚类[3-4]、音频聚类[5]等.近年来,随着 Web 多媒体数据类型的不断丰富,越来越多的研究者开始关注包括图像、音频等的多媒体数据综合处理和聚类问题;本文的前期工作和相关研究已证明:由于图像、音频等不同类型的多媒体数据可以从视觉、听觉的不同侧面表达相似的语义,视听觉信息彼此具有互补性,这种互补性可用于提高多媒体语义理解的准确率[6-10].

本文在前期工作[6]的基础上,以 Web 页面上获取的图像和音频为训练数据,在特征降维过程中分析了两者间的统计相关性;并在特征子空间中利用图像和图像、图像和音频以及音频和音频之间的多重数据关系进行相关性融合,挖掘潜在的相似语义,并修正相似度矩阵;最后通过基于相似度的循环迭代算法实现图像和音频聚类.对比实验从多方面验证了本文方法的有效性,实验还表明本文方法可成功应用于多媒体检索,实现图像和音频之间的交叉检索.

2 基于核矩阵的子空间映射

特征降维是对多媒体数据向量化表示的重要步骤,也是多媒体内容分析和聚类的前提.传统的方法通常是针对训练样本集提取出的高维特征矩阵进行矩阵分解,并通过线性变换实现维数约减.图像对应非时序性的视觉特征,音频对应时序性的听觉特征,虽然针对两种特征目前都已提出了不同的降维方法[11-12],但很少有研究关注如何对视觉特征矩阵和听觉特征矩阵统一分析和降维,并计算视听觉特征间的潜在相关性.

视觉特征从颜色、纹理、形状等层面描述了图像的视觉属性;音频特征从频域、压缩域等方面描述音频数据的听觉属性.因此,给定 n 幅图像和 n 段音频组成的图像-音频数据库,设视觉特征包括 j 个属性值,即构成 j 维向量,听觉特征构成 k 维向量,则特征提取后得到 $n\times j$ 维的视觉特征矩阵 $\boldsymbol{A}$ 和 $n\times k$ 维的听觉特征矩阵 $\boldsymbol{B}$.在前期研究中提出的基于 CCA (Canonical Correlation Analysis)的相关性保持映射基础上[6],引入核矩阵方法,提出基于 KCCA (Kernel Canonical Correlation Analysis)的同构子空间映射.

首先,通过非线性映射 $\Phi(\boldsymbol{A}),\Psi(\boldsymbol{B})$ 将视觉特征向量和听觉特征向量映射到核空间,并在此空间中采用传统的 CCA 方法计算相关性保持映射的投影向量 $\boldsymbol{W}_{\boldsymbol{A}}=\sum\alpha_i\Phi(a_i)=\Phi(\boldsymbol{A})\boldsymbol{\alpha},\boldsymbol{W}_{\boldsymbol{B}}=\sum\beta_i\Psi(b_i)=\Psi(\boldsymbol{B})\boldsymbol{\beta}$,其中 $\boldsymbol{\alpha}=[\alpha_1,\cdots,\alpha_n]^{\mathrm{T}},\boldsymbol{\beta}=[\beta_1,\cdots,\beta_n]^{\mathrm{T}}$ 表示组合系数;

其次,与 CCA 方法类似,得出目标函数 $\max[\boldsymbol{\alpha}^{\mathrm{T}}\Phi(\boldsymbol{A})^{\mathrm{T}}\Phi(\boldsymbol{A})\Psi(\boldsymbol{B})^{\mathrm{T}}\Psi(\boldsymbol{B})\boldsymbol{\beta}]$,其最优解即为能最大程度保持相关性的投影子空间;引入核矩阵 $\boldsymbol{K}_a,\boldsymbol{K}_b\in R^{n\times n}$,其中 $(\boldsymbol{K}_a)_{ij}=\Phi(a_i)^{\mathrm{T}}\Phi(a_j)$,且 $(\boldsymbol{K}_b)_{ij}=\Psi(b_i)^{\mathrm{T}}\Psi(b_j)$,则可以将目标函数简化为在约束条件 $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}_a\boldsymbol{K}_a\boldsymbol{\alpha}=1,\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{K}_b\boldsymbol{K}_b\boldsymbol{\beta}=1$ 下求解 $\max\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}_a\boldsymbol{K}_b\boldsymbol{\beta}$;

最后,通过拉格朗日乘子法求解上述目标函数,得到组合系数 α_i,β_i 的值,从而,训练集中任意一个图像样本 a 在低维子空间 S^* 中的坐标为 $\boldsymbol{W}_{\boldsymbol{A}}^{\mathrm{T}}\Phi(a)=\sum\alpha_i\Phi(a_i)^{\mathrm{T}}\Phi(a)=\sum\alpha_i\boldsymbol{K}_a(a_i,a)$,音频样本的坐标可以采用类似的方法计算得到.通过上述变换,将矩阵 $\boldsymbol{A}$ 和矩阵 $\boldsymbol{B}$ 转换为 $n\times m(m<j,m<k)$ 维的矩阵 $\boldsymbol{A}^*$ 和 $\boldsymbol{B}^*$.

设向量 $\boldsymbol{z}_i=(z_{i_1},\cdots,z_{i_m})$ 表示 S^* 中的图像或音频样本点,用 d_{ij} 表示任意两个样本点 $\boldsymbol{z}_i$ 和 $\boldsymbol{z}_j$ 在 S^* 中的距离,进而得到所有样本点之间的距离矩阵 $\boldsymbol{D}_{2n\times 2n}=[d_{ij}]$,归一化后表示为 $\boldsymbol{D}^*_{2n\times 2n}=[d^*_{ij}](d^*_{ij}\in(0,1])$,矩阵 $\boldsymbol{D}^*_{2n\times 2n}$ 体现了图像和音频在子空间 S^* 中的几何拓扑关系.

3 基于矩阵融合的相似度优化

由于语义鸿沟的存在[13],矩阵 $\boldsymbol{D}^*_{2n\times 2n}$ 并不能真实地反映样本点在语义上的相似度,即:若样本 $\boldsymbol{z}_i$ 和 $\boldsymbol{z}_j$ 在矩阵 $\boldsymbol{D}^*_{2n\times 2n}$ 中对应的元素值较小,并不能说明两者在语义上是强相关的,反之亦然.针对上述问题,本节提出基于矩阵融合的优化算法,从全局意义上对图像和音频的相似度进行求精.

图像和音频样本共同分布于子空间 S^* 中,这些样本点之间存在的相关性数据关系主要可分为 4 种,如图 1 所示.其中,实线表示子空间中两个样本之间距离较近,为强相关,虚线表示两个样本之间距离较远,为弱相关.这 4 种数据关系在子空间 S^* 中并不是孤立存在的,同时,不同类型的数据关系之间具有互补性和可传递性[6-7,9].

对距离矩阵 $\boldsymbol{D}^*_{2n\times 2n}=[d^*_{ij}]$ 求取倒数,得到关系

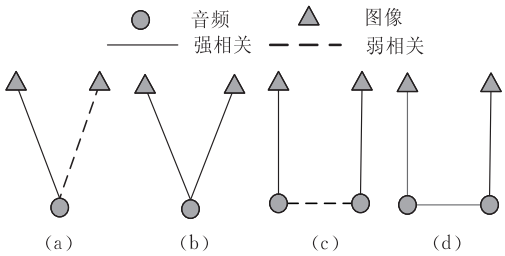


图 1　子空间中图像和音频样本间的相互关系

矩阵 $\boldsymbol{M}$，定义如下：

$$\boldsymbol{M}=\begin{pmatrix}\boldsymbol{M}_{II} & \boldsymbol{M}_{IA}\\ \boldsymbol{M}_{AI} & \boldsymbol{M}_{AA}\end{pmatrix};\ \boldsymbol{M}(i,j)=\boldsymbol{0},\ \text{若}\ \frac{1}{d_{ij}^{*}}<\sigma \tag{1}$$

其中 σ 为常参数，子矩阵 $\boldsymbol{M}_{II}$，$\boldsymbol{M}_{AA}$ 分别是图像相似度矩阵和音频相似度矩阵，子矩阵 $\boldsymbol{M}_{IA}$ 和 $\boldsymbol{M}_{AI}$ 是对称矩阵，表示图像和音频之间的相关程度. 则任意两幅图像 x_i 和 x_j 之间的相似度对应子矩阵 $\boldsymbol{M}_{II}$ 中的元素值 $\boldsymbol{M}_{II}(x_i,x_j)$，采用式(2)对 $\boldsymbol{M}_{II}(x_i,x_j)$ 进行优化：

$$\boldsymbol{M}_{II}^{*}(x_i,x_j)=\lambda\boldsymbol{M}_{II}(x_i,x_j)+(1-\lambda)\cdot\sum\boldsymbol{M}_{IA}(x_i,y_k)\boldsymbol{M}_{IA}(x_j,y_t)\boldsymbol{M}_{AA}(y_k,y_t) \tag{2}$$

式(2)的约束条件为

$$\boldsymbol{M}_{IA}(x_i,y_k)>\varepsilon_1,\ \boldsymbol{M}_{IA}(x_j,y_t)>\varepsilon_2,\ \boldsymbol{M}_{AA}(y_k,y_t)>\varepsilon_3,\text{且}\ \alpha,\varepsilon_1,\varepsilon_2,\varepsilon_3\in(0,1) \tag{3}$$

式(2)中实参 λ 是权重参数，实参 $\varepsilon_1,\varepsilon_2,\varepsilon_3$ 用于控制音频样本 y_k 和 y_t 的选择，y_k 和 y_t 是用于传递相似度的媒介，故参数 $\varepsilon_1,\varepsilon_2,\varepsilon_3$ 称为传递因子；$\boldsymbol{M}_{IA}(x_i,y_k)>\varepsilon_1$ 和 $\boldsymbol{M}_{IA}(x_j,y_t)>\varepsilon_2$ 表示图像 x_i 和音频 y_k 以及图像 x_j 和音频 y_t 之间具有强相关性，如图 1(b)～(d)所示；$\boldsymbol{M}_{AA}(y_k,y_t)>\varepsilon_3$ 要求音频样本 y_k 和 y_t 是强相关的，如图 1(b)、(d)所示(其中图 1(b)表示 y_1 和 y_2 是同一样本的特殊情况)；符号 $\sum$ 表示对所有符合约束条件的音频 y_k 和 y_t 计算 $\boldsymbol{M}_{IA}(x_i,y_k)\boldsymbol{M}_{IA}(x_j,y_t)\boldsymbol{M}_{AA}(y_k,y_t)$，对结果进行累加. 实验中传递因子 ε_1 的取值为

$$\varepsilon_1=\frac{1}{2}\left(\frac{1}{n}\sum_{m=1}^{n}\boldsymbol{M}_{IA}(x_i,y_m)+\max(\boldsymbol{M}_{IA}(x_i,\forall y))\right) \tag{4}$$

其中 n 为音频样本数量. 传递因子 ε_2 的计算与 ε_1 类似；参数 ε_3 依照式(4)从子矩阵 $\boldsymbol{M}_{AA}$ 计算得出. 同理，式(2)可用于修正音频和音频样本之间的相似度 $\boldsymbol{M}_{AA}$，修正后的相似度矩阵记为 $\boldsymbol{M}_{II}^{*}$，$\boldsymbol{M}_{AA}^{*}$.

式(2)的计算结果更接近图像 x_i 和 x_j 在语义上的相似程度，这是因为：式(2)从多重数据关系的角度，综合分析了子空间中多媒体样本点间的相似度，将图像和音频以及音频和音频之间的相关性，融入到图像相似度的计算和优化. 另外，从矩阵优化的角度而言，子矩阵 $\boldsymbol{M}_{II}$，$\boldsymbol{M}_{AA}$，$\boldsymbol{M}_{IA}$($\boldsymbol{M}_{AI}$)分别表示子空间中 3 种单一类型的多媒体数据关系，具有稀疏性，同时，这 3 种数据关系之间具有互补性和可传递性，式(2)将子矩阵 $\boldsymbol{M}_{AA}$，$\boldsymbol{M}_{IA}$ 中潜在的互补信息融入到子矩阵 $\boldsymbol{M}_{II}$ 中，提高了矩阵 $\boldsymbol{M}_{II}$ 的密集度，也是对矩阵 $\boldsymbol{M}_{II}$ 所表达数据关系的优化.

4　聚类算法

由于在矩阵融合过程中图像和音频在子空间 S^{*} 中的坐标未变，不能以坐标值为多媒体聚类的输入条件. 并且，一些传统的聚类方法，如 Kmeans 聚类[3]，在初始状态下需要指定聚类中心，而聚类中心的选择将会对结果造成较大影响. 本文受到 AP (Affinity Propagation)聚类算法[4]的启发，采用基于相似度的循环迭代方法进行聚类. AP 算法利用数据点之间的相似度，通过循环迭代自动计算聚类质心和及其隶属数据点，最初被用于文本数据分析、人脸识别等领域.

设图像数据集 X 分布于无向加权图 G 中，且 $x_i,x_j\in X$ 之间的权重对应于图像相似度矩阵中的元素值 $\boldsymbol{M}_{II}^{*}(x_i,x_j)$，节点 x_i,x_j 之间存在一条无向加权边当且仅当 $\boldsymbol{M}_{II}^{*}(x_i,x_j)<\xi$，其中 ξ 为常参数. 图像聚类质心的计算如下，节点 $node(i)$ 向其相邻节点 $node(j)$ 发送实数值的消息 $r(i,j)$，表示节点 $node(i)$ 选择节点 $node(j)$ 作为质心的概率，计算如下：

$$r(i,j)=\boldsymbol{M}_{II}^{*}(i,j)-\max_{j\neq j'}\{a(i,j')+\boldsymbol{M}_{II}^{*}(i,j')\} \tag{5}$$

其中 $a(i,j)$ 是节点 $node(j)$ 向 $node(i)$ 发送实数值消息，表示 $node(j)$ 能够成为 $node(i)$ 的质心的概率，$a(i,j)$ 的值初始化为零，$a(i,j)$ 的计算如下所示：

$$a(i,j)=\min\left\{0,r(j,j)+\sum\max\{0,r(i',j)\}\right\} \tag{6}$$

式(5)、(6)在整个数据集内迭代进行，直到达到收敛状态，即 $r(i,j)$ 和 $a(i,j)$ 的变化小于规定的阈值. 图像聚类质心的计算过程反映了整个数据集范围内的累积关系，即 $r(i,j)$ 的值，同时 $a(i,j)$ 的不断更新也反映了某个图像数据点能够成为一个合适的聚类质心的累积概率. 因此，对于图像数据点 $node(i)$，

若 $node(j)$能够使 $a(i,j)+r(i,j)$取得最大值，则 $node(j)$为 $node(i)$的聚类质心. 音频的聚类采用与图像类似的方法得到.

5 实验结果与分析

5.1 数据集和特征选择

为验证上述算法的有效性，在 Windows XP 下用 VC6.0 实现了一个原型系统，实验从 Web 页面上采集了 20 个语义类别的图像和音频作为训练数据集，例如：爆炸、鸟类、汽车、老虎、狗、海豚、闪电等类别，其中每个类别中包括 100 幅图像和 60 段音频例子. 实验数据主要来源于网站 http://www.animalbehaviorarchive.org、http://encarta.msn.com 和 http://image.baidu.com/，部分图像来自于 Corel 图像数据集.

实验提取的图像特征包括 256-d HSV 颜色直方图、64-d LAB 颜色聚合向量以及 32-d Tamura 方向度；音频特征包括 4 个 Mpeg 压缩域特征：质心(Centroid)、衰减截至平率(Rolloff)、频谱流量(Spectral Flux)和均方根(RMS).

音频是时序性数据，对持续时间不同的音频样本提取得到的特征向量维数也不同，文本收集的音频例子持续时间均不超过 7 秒钟，并使用前期工作中的模糊聚类方法[6]，对初始音频特征提取相同数目的聚类质心作为音频索引. 此外，基于相关性分析视听觉特征降维和坐标计算用 Matlab 7.0 完成，计算复杂度将随着特征维数的降低而减小.

5.2 性能评价公式

实验从训练数据集中选取 $k(2\leq k\leq 20)$个语义类别的图像和 $k(2\leq k\leq 20)$个语义类别音频数据进行聚类，并设定输出的图像聚类个数和音频聚类个数均等于选定的 k 值.

设某个图像样本 x 和音频样本 y 均是从语义类别为 g_i的数据集中选取，x 的聚类结果是被划分到类别 r_i中，y 被划分到类别 s_i中，则图像聚类准确率 I_{Accuracy}和音频聚类准确率 A_{Accuracy}计算如下：

$$\begin{cases} I_{\text{Accuracy}}=\dfrac{1}{n}\sum_{i=1}^{n}\delta(g_i,map(r_i)) \\ A_{\text{Accuracy}}=\dfrac{1}{m}\sum_{i=1}^{m}\delta(g_i,map(s_i)) \end{cases},$$

$$\delta(j,k)=\begin{cases}1, & j=k\\ 0, & \text{其它}\end{cases} \tag{7}$$

其中 n,m 分别表示本次聚类实验中图像样本总数和音频样本总数，$map(x)$是将语义类别标签映射到聚类结果上的最优映射函数，本文用文献[14]中的 Kuhn-Munkres 方法获取最优映射.

5.3 参数选取

相关性融合算法中传递因子 $\varepsilon_1,\varepsilon_2,\varepsilon_3$的取值根据式(4)计算得出，$\varepsilon_3=0.72$，$\varepsilon_1,\varepsilon_2$没有固定取值，而是随图像样本集的变化而变化(参见第 3 节)；此外，权重参数 λ 直接影响了相似度修正结果$\boldsymbol{M}_{II}^{*},\boldsymbol{M}_{AA}^{*}$.

为优化 λ 取值，实验从值域(0,1)范围内选取不同数值进行测试，如图 2 所示. 并且对于 λ 的每个取值，按照 5.2 节中的方法，从训练数据集中选取 k 个语义类别的图像和 k 个语义类别音频数据进行聚类，计算相应的 I_{Accuracy}和 A_{Accuracy}，图 2 的结果是 $k=3,4,5,6$ 的情况下得到的性能均值.

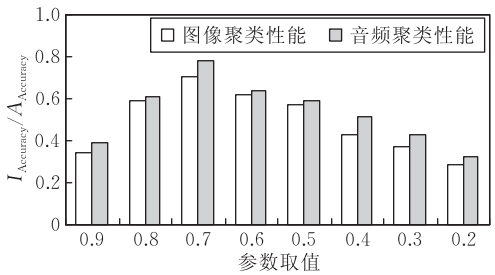


图 2 权重参数 λ 对聚类性能的影响

可见，当权重参数 $\lambda=0.7$ 时，图像聚类性能 I_{Accuracy}和音频聚类性能 A_{Accuracy}均达到最优.

5.4 对比实验和分析

在确定参数取值之后，为验证本文方法的有效性和优越性，实验分别采用下列 4 种方法，对 5.2 节中选取的 k 个语义类别的图像和 k 个语义类别的音频数据集进行聚类：

(1)本文方法. 用本文的方法对实验数据进行视听觉特征统一降维和子空间映射、矩阵融合和基于相似度的数据聚类；

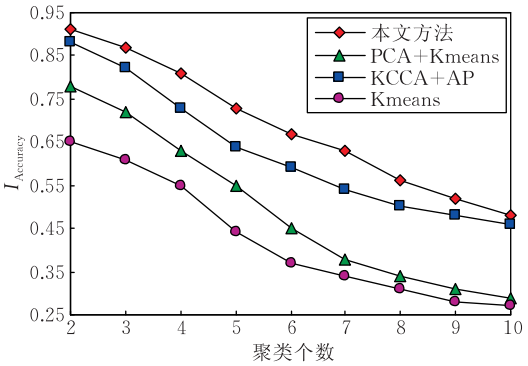
(2)KCCA+AP. 为了验证本文方法在子空间中进行矩阵融合的有效性，首先采用 KCCA 方法进行子空间映射，然后直接采用 AP 算法[4]分别对图像和音频聚类；

(3)PCA+Kmeans. PCA(Principal Component Analysis)[15]是一种经典的多媒体特征分析方法，首先对 5.1 节中提取的视觉特征和音频特征分别用 PCA 方法进行主成分提取和去噪，然后用传统的 Kmeans[3]聚类方法分别进行图像和音频聚类；

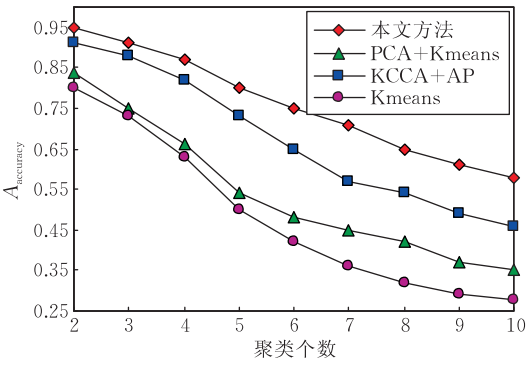
(4)Kmeans. 直接使用 Kmeans 聚类方法分别进行图像聚类和音频聚类.

上述实验中参数 k 的取值为[2,10]范围内的所有整数，并对 k 的每个取值，随机选择 5 次数据集进行聚类实验，最后计算性能均值. 其中 80% 的数据用作训练数据，其余 20% 作为测试数据(Kmeans 方法除外，无训练过程).

图 3 显示了上述 4 种方法所得聚类结果的 I_{Accuracy}，A_{Accuracy} 性能平均值. 从图 3 可以看到，本文的方法在图像聚类准确率 I_{Accuracy} 和音频聚类准确率 A_{Accuracy} 方面均优于其它 3 种方法.



(a) 图像聚类结果



(b) 音频聚类结果

图 3　4 种方法得到的聚类性能对比

几类方法中，PCA＋Kmeans 方法的 I_{Accuracy} 性能虽然相对较低，但仍优于 Kmeans 方法，这是因为 PCA 方法的使用去除了视觉特征中的噪声；PCA＋Kmeans 方法和 Kmeans 方法的 A_{Accuracy} 性能较为接近，这是因为在 5.1 节中已采用文献[3]中的模糊聚类方法，对初始的时序性听觉特征进行索引.

此外，本文方法明显优于 KCCA＋AP 方法，这是因为聚类性能的高低很大程度上取决于输入条件，即：相似度矩阵，也正说明了本文提出的特征子空间中相关性融合算法能够优化图像和音频数据集的相似度矩阵，使之更加符合高层语义关系.

图 4 是当 $k=5$ 时，聚类结果的一个示例图，其中每一列是一个聚类类别中的图像示例，并按照与聚类中心的相似度进行排序，包括前 5 个正确结果以及一个错误结果. 实验结果说明了，关系矩阵融合方法挖掘了图像和音频数据集在特征子空间中的潜在关系，可有效用于语义理解和聚类.

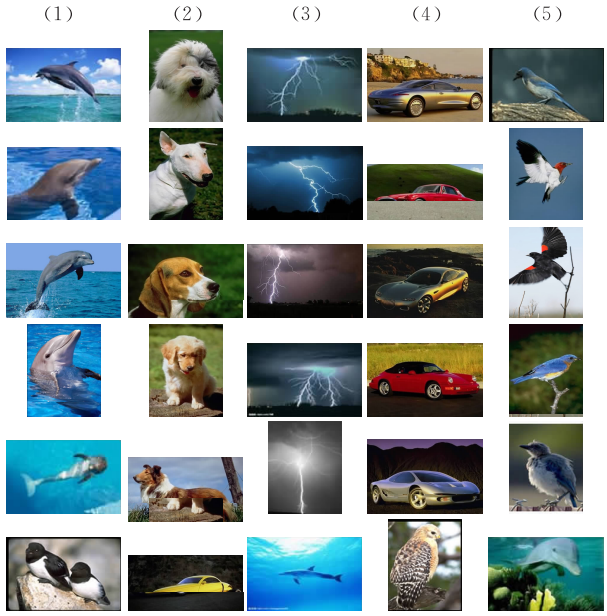


图 4　图像聚类结果示例($k=5$)

5.5　图像-音频交叉检索结果

为进一步验证本文方法的适用性，在上述实验的基础上，还设计了图像-音频交叉检索实验，步骤如下：

(1) 求解相似音频. 用户提交一个图像样本 r 作为查询例子，检索系统从音频聚类结果中找到类别标签与 r 相一致的音频类 $\Omega=\{y_1,\cdots,y_i,\cdots,y_p\}$ (y_i 表示音频样本)；

(2) 排序. 从相关性矩阵 $\boldsymbol{M}_{IA}$ (参见第 3 节)中找到 r 与 y_i 的关系值，以降序输出数据集 Ω 中的音频，作为检索结果.

表 1 显示了当参数 $k\in[5,6,7,8,9]$ 时，对检索结果进行两轮相关反馈(采用文献[16]中的反馈策略)后得到的平均结果，其中每一行列出了返回结果个数分别为 $n=5,10,15,20,25,30,35$ 时，正确结果的个数. 可见，本文的方法应用于图像-音频之间的交叉检索，可以取得较好的结果，当样本类别数 $k=5$，返回结果个数为 $n=15$ 时，正确结果的个数为 13.12.

表 1　以图像为查询例子检索音频的平均结果

k	平均正确结果个数						
	$n=5$	$n=10$	$n=15$	$n=20$	$n=25$	$n=30$	$n=35$
5	4.41	8.69	13.12	16.82	19.54	22.87	23.45
6	4.31	8.21	12.65	15.27	18.11	19.87	21.13
7	4.12	7.13	10.93	14.15	17.21	18.41	19.35
8	3.95	6.84	10.24	12.74	14.13	15.14	16.05
9	3.25	6.25	9.26	11.67	12.45	13.74	14.23

6 结 论

不同于传统的多媒体聚类研究,本文提出的方法可以同时用于图像和音频两种不同量纲的多媒体数据,创新之处在于:将图像和音频数据同时映射为特征子空间中的样本点,综合利用子空间中不同数据集之间的多重相关性,优化多媒体语义的学习结果,并通过基于相似度的循环迭代算法实现了图像和音频的聚类.本文方法考虑了目前 Web 多媒体数据中图像和音频共存的现实情况,突破了传统聚类方法对不同类型数据集之间相关性融合分析上的局限性.

实验从多方面验证了本文方法的有效性,并给出了应用于多媒体检索领域的实例和性能分析结果.局限性在于,当数据集达到海量时,本文方法缺少有效索引机制,难以快速处理 Web 上的海量多媒体信息.因此,进一步研究工作包括:大规模数据集的多层索引和响应时间优化等问题.

参 考 文 献

[1] Lew M, Sebe N, Djeraba C, Jain R. Content-based multimedia information retrieval: State-of-the-art and challenges. ACM Transactions on Multimedia Computing, Communication and Applications, 2006, 2(1): 1-19

[2] Bekkerman R, Jeon J. Multi-modal clustering for multimedia collection//Proceedings of the CVPR. Minneapolis, USA, 2007: 1-8

[3] McLachlan G J, Basford K E. Mixture models: Inference and applications to clustering. Statistics: Textbooks and Monographs, New York, 1988

[4] Frey Brendan J, Dueck Delbert. Clustering by passing messages between data point. Science, 2007, 315: 972-976

[5] Guo G D, Li S Z. Content-based audio classification and retrieval by support vector machines. IEEE Transactions on Neural Network, 2003, 14(1): 209-115

[6] Zhang Hong, Wu Fei, Zhuang Yue-Ting, Chen Jian-Xun. Cross-media retrieval method based on content correlations. Chinese Journal of Computers, 2008, 31(5): 820-826 (in Chinese)
(张鸿, 吴飞, 庄越挺, 陈建勋. 一种基于内容相关性的跨媒体检索方法. 计算机学报, 2008, 31(5): 820-826)

[7] Yang Yi, Xu Dong, Nie Feiping et al. Ranking with local regression and global alignment for cross-media retrieval//Proceedings of the ACM Multimedia Conference. Beijing, China, 2009: 175-184

[8] Wu Fei, Zhang Hong, Zhuang Yueting. Learning semantic correlations for cross-media retrieval//Proceedings of the International Conference on Image Processing. Atlanta, USA, 2006: 1465-1468

[9] Yang Yi, Zhuang Yueting, Wu Fei, Pan Yunhe. Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. IEEE Transactions on Multimedia, 2008, 10(3): 437-446

[10] McGurk Harry, MacDonald John. Hearing lips and seeing voices. Nature, 1976, 264: 746-748

[11] Wu Yi, Chang Edward Y, Chang Kevin Chen-Chuan, Smith John R. Optimal multimodal fusion for multimedia data analysis//Proceedings of the ACM Multimedia Conference. New York, USA, 2004: 572-579

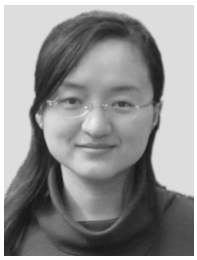
[12] Seung H S, Lee D. The manifold ways of perception. Science, 2000, 290(5500): 2268-2269

[13] Zhao R, Grosky W I. Negotiating the semantic gap: From feature maps to semantic landscapes. Pattern Recognition, 2002, 35(3): 593-600

[14] Lovasz L, Plummer M. Matching Theory. Holland: Elsevier Science Publishers B. V., 1986

[15] Joliffe I. Principal Component Analysis. New York: Springer-Verlag, 1986

[16] Rui Yong, Huang Thomas S, Ortega Michael, Mehrotra Sharad. Relevance feedback: A power tool in interactive content-based image retrieval. IEEE Transactions on Circuits and Systems for Video Technology, 1998, 8(5): 644-655



ZHANG Hong, born in 1979, Ph. D., associate professor. Her research interests include content-based multimedia analysis, machine learning and cross-media retrieval.

WU Fei, born in 1973, Ph. D., professor. His research interests include content-based multimedia retrieval and statistical learning theory.

ZHANG Xiao-Long, born in 1963, Ph. D., professor. His research interests include machine learning and data mining.

Background

Multimedia data clustering is a hot research topic in content-based multimedia analysis and semantic understanding. Most research works focused on how to actually learn data correlation within single modality, and proposed effective

clustering algorithms, such as image clustering, audio clustering and video clustering. Little of them concerned data clustering algorithms for multimedia data of different modalities. However, in some cases multimedia data of different modalities co-exist, such as webpage and multimedia document, and different multimedia data represent high-level semantics from different aspects. So it is interesting to mine underlying cross-model correlation and utilize complementary information among different modalities. Especially for multimedia data clustering, above issues are important to calculate cluster centers. Main challenges for multimedia clustering on different modalities include the heterogeneity between low-level content features of different modalities and correlation measure between different modalities. This paper proposes a multimodal data clustering algorithm based on correlation matrix fusion. Considering image and audio are two typical kinds of multimedia data, our algorithm is described and tested based on image-audio database. This paper constructs an isomorphic subspace which solves heterogeneity problem and enables cross-modal correlation learning, calculates image and audio cluster centers by similarity-based method. This paper is supported by National Natural Science Foundation of China (Nos. 61003127, 61070068). These projects focus on multimodal feature analysis and cross-media retrieval. The research team has been focused on content-based cross-media retrieval, data clustering, feature analysis, and has published some papers. This paper solves multimodal feature analysis and semantic understanding for data clustering, which is an important issue for the projects.