ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.9, September 2008, pp.2449−2460
DOI: 10.3724/SP.J.1001.2008.02449

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于扩展生成语言模型的图像自动标注方法*

王 梅, 周向东+, 张军旗, 许红涛, 施伯乐

(复旦大学 计算机与信息技术系,上海 200433)

Image Auto-Annotation via an Extended Generative Language Model

WANG Mei, ZHOU Xiang-Dong+, ZHANG Jun-Qi, XU Hong-Tao, SHI Bai-Le

(Department of Computing and Information Technology, Fudan University, Shanghai 200433, China)

+ Corresponding author: E-mail: xdzhou@fudan.edu.cn

Wang M, Zhou XD, Zhang JQ, Xu HT, Shi BL. Image auto-annotation via an extended generative language model. *Journal of Software*, 2008,19(9):2449−2460. http://www.jos.org.cn/1000-9825/19/2449.htm

Abstract: In this paper, based on the statistical smoothing strategy, a image region feature generative probability estimation method is proposed by exploiting maximum weight matching algorithm. By further analyzing and measuring the semantic correlations between words based on the training set, a novel image annotation algorithm for adopting the generative model is presented. The first annotation keyword is obtained by using the proposed image region feature generative probability estimation algorithm. Then, a heuristic iterate function is proposed to exploit the keyword semantic correlation. Finally, the semantic correlation between the annotation and the image can be improved by our annotation algorithm. The proposed annotation approach is tested on a real-world image database, and promising results are achieved.

Key words: image annotation; generative model; continuous feature estimation; maximum weight matching; semantic correlation

摘 要: 使用最大权匹配算法,结合统计平滑技术,提出图像区域特征生成概率估计方法,并进一步对训练集中标注词之间的语义相关性(correlation)进行分析与度量,给出一种基于生成模型的图像标注算法.算法使用所提出的基于最大权匹配的图像生成概率估计方法得到较好的起始点,进而设计启发式迭代函数对词与词的相关性加以利用,最终提高标注词与图像的语义相关性.在现实世界图像数据库上的实验结果验证了所提出标注方法的有效性.

关键词: 图像标注;生成模型;连续特征估计;最大权匹配;语义相关性

中图法分类号: TP311 **文献标识码**: A

图像语义的自动标注是实现图像语义检索的关键.标注就是使用语义关键字或标签来表示一幅图像的语义内容,进而将图像检索转化为文本检索.早期手工标注需要专业人员对每幅图像都要标出关键字,费时且具有主观性.图像数量的爆炸性增长促使人们利用各种机器学习算法、统计模型等设计出多种图像自动标注模型[1−16].然而,由于存在语义鸿沟(semantic gap),自动获取图像的语义信息仍然非常困难,图像语义的自动标注性

* Supported by the National Natural Science Foundation of China under Grant Nos.6040301, 860773077 (国家自然科学基金); the National Basic Research Program of China under Grant No.2005CB321905 (国家重点基础研究发展计划(973))

Received 2007-02-14; Accepted 2007-06-29

能亟待提高.

利用统计学习进行图像标注的关键是从训练集中找到视觉特征与标注词之间的关系.一个直观的想法是,同一关键词的视觉特征具有一致性,如“老虎”,其颜色和纹理在视觉特征上保持一致.这样,图像可以被分割成一些带有一定语义含义的局部区域(如采用 Normalized cut 图像分割技术[17]),理想情况下,图像分割后每个区域对应一个清晰的语义对象.因此,通过区域特征的距离计算可以近似度量两个区域(对象)的语义相似性.在此基础上,出现了离散特征模型,如 Duygulu 等人(2002 年)提出的翻译模型[1],对分割后的图像区域特征进行聚类,将连续特征变成离散视觉关键字单词表,图像的标注问题可以看作从图像视觉关键字到语义关键字的翻译过程;Jeon 等人提出的相关模型 CMRM[2](2003 年)利用视觉关键字与语义关键字的相关性(联合概率)进行标注.虽然离散特征模型考虑了对象和区域的语义含义,但这种对视觉特征的离散处理方法会造成视觉特征内容的损失.由于事先很难确定一个理想的聚类粒度,标注结果受离散化效果影响较大[3].

随后出现的连续相关模型 CRM[4]和 MBRM[3](2004 年)使用非参数高斯核进行特征生成概率的连续估计,与离散模型相比,其标注性能有显著提高.然而,上述方法对图像各个区域在生成概率估计中所起作用的复杂性考虑不足.将核密度估计看作区域-图像的相似性度量函数,则在上述模型中,区域与图像的相似性定义为区域与图像各个区域相似性度量的平均值.然而,一幅图像的不同区域对该相似性度量的贡献程度不同,如图像中的“对象”区域比“背景”区域贡献要大,多数已知工作忽视了图像语义相似性度量的复杂性.另一方面,常用的图像分割方法容易产生许多语义不明确的区域,由于这些区域的弱语义特性,使得这类区域经常与代表多种不同语义对象的区域都具有较高的特征相似性,影响图像特征生成概率的估计的准确性.本文通过二分图最大权匹配算法得到图像之间相似区域的匹配结构,在该结构基础上给出一种新的图像特征生成概率的连续估计方法,从而提高了特征生成概率估计的准确性.

出于简化计算的目的,已知工作普遍假设语义关键词之间相互独立.近年来,人们意识到标注时利用词与词之间的相关性(correlation)能够改进标注性能,如“people,beach”,“people,garden”具有较高的相关性,则这两者同时作为图像标注的概率较高;而“grass,tiger”组合成为某一图像的语义标注的概率显然高于“ocean,tiger”.在标注中使用此类相关性信息的代表性工作包括 CLM 模型[5](2004)、TMHD 方法[6](2005)以及 AGAnn[7](2006)方法.CLM 模型使用 EM 算法隐含考虑词与词的相关性;TMHD 方法需借助外部数据源 WordNet,对训练集中有用的信息未充分利用;AGAnn 对自适应图(adaptive graph)标注的结果应用词与词的相关性.本文使用类似文本信息检索中常用的自动局部分析方法(automatic local analysis)[18]针对待标注图像自适应选择局部训练集进行分析,来获得词与词之间相关性的先验知识,并在标注过程中应用该先验知识.

本文提出一种新的基于生成模型的图像自动标注算法,该算法利用启发式迭代过程进行图像语义标注.在迭代过程中,起始标注的准确估计至关重要,而起始标注词估计的准确程度与图像特征生成概率估计的准确程度紧密相关.本文给出的基于最大权匹配的图像特征生成概率方法提高了特征生成概率估计的准确性,从而为贪心算法提供良好的起点.在此基础上,通过本文给出的词与词相关性的有效度量方法,利用词与词相关性,语义相关的一组词被优先选择为最终标注.结合上述两方面内容,本文算法最终提高了标注性能.使用基准数据集 ECCV2002 对本文提出的标注算法进行检验,与基于图像分割的 MBRM 方法进行比较,本文方法的 recall 和 precision 均有显著提高,分别由 16.1%和 19.0%提高到 19.5%和 21.1%.

本文第 1 节讨论图像标注的相关工作.第 2 节介绍相关模型进行标注的基本框架.第 3 节介绍基于最大权匹配的连续特征密度估计.第 4 节介绍词的语义相似性度量.第 5 节介绍本文设计的启发式贪心算法用于标注的过程.第 6 节讨论实验结果.第 7 节是总结和展望.

1 相关工作

近年来,图像自动标注领域非常活跃,人们利用机器学习方法、统计模型设计出各种不同的学习模型.这些模型主要包括翻译模型[1]、LSA&PLSA 模型[8,9]、相关模型[2−4]、分类模型[10,11]等.机器翻译模型[1]将图像标注过程视为从视觉关键字到文本关键字之间的翻译过程,通过寻找标注词和图像特征之间的关系对待标注图像

进行标注.LSA[8]和 PLSA[9]模型引入隐藏变量在图像特征和词之间建立联系,从而寻找两者共同出现的信息.Monay 等人[9]建立了一对有关联的 PLSA 模型,对文本特征赋予了更高的重要性.将每个标注词看作独立的类,为每个词创建不同的图像分类模型,图像标注也可使用文本分类技术.Srikanth 等人[10]通过 WordNet 建立词的层次结构,进而使用层次化的分类方法进行标注;在最新的图像标注研究中,Jain 等人[11]提出使用多分辨率基于固定网格的图像内容表示方法以及层次增强算法来解决使用图像分类的标注中图像内容表示以及分类器的有效训练等问题.Li 和 Wang[12]设计实现一个在线的快速图像标注系统:ALIPR 标注系统.

在这些模型中,图像特征的相似性度量方式对标注性能的影响至关重要.图像自动标注的机器模型[1]及离散特征模型 CMRM[2],使用离散度量方式.首先,采用图像分割技术生成每幅图像的局部区域,这些图像区域的特征经过聚类,形成带有语义特征的视觉关键字,进而学习视觉关键字与语义关键字之间的关系.但离散化方法会造成视觉特征内容的损失,影响了标注效果.在连续特征模型 CRM[4]和 MBRM[3]中,通过基于核的非参数估计方法提高了图像特征生成概率的估计,但上述连续模型在估计图像生成概率时,对图像各个区域在特征生成概率估计中所起作用的复杂性考虑不足.基于区域的图像相似性度量在基于内容图像检索(CBIR)领域已被研究多年,并存在多种利用区域匹配结构信息来提高图像检索效果的方法[19−21].Wang 等人在 SIMPLIcity[19]中定义了加权的区域相似性之和来度量两幅图像的相似性,最相似的区域有最高匹配优先权.Zhang 在 FUZZYCLUB[20]中对其加以改进,定义一幅图像的某区域和另一幅图像的距离值为该区域和另一幅图像的所有区域中距离度量最小的值,该最小距离值反映区域和图像的最大相似性.这些方法都表明,在图像整体相似性度量中不应忽视图像之间的区域匹配特性.我们认为,这种图像相似区域之间的匹配结构是揭示图像语义相关性的关键,在图像标注领域,估计待标注图像的生成概率时应该充分利用这种匹配信息来提高估计的准确性.

从图像分割的粒度上讲,存在着基于图像分割后的区域和基于固定大小的网格.MBRM[3]中将基于固定大小的网格与基于分割的标注性能进行比较,实验结果显示,基于固定网格图像划分的标注性能优于基于图像分割的标注性能.但由于基于图像分割的标注可以很方便地扩展到对区域(对象)的标注[22]和基于区域(对象)的图像检索[23],故本文仍然关注于基于图像分割的图像标注.

如前所述,训练集中词和词的相关性能够提高标注性能.理想情况是,应该利用词的集合进行标注,此时需穷举标注单词表的所有子集,当标注单词表非常大时,在计算上并不可行.Jin 等人[5]在其标注方法中将该问题放松为估计 $P(\theta_w|I)$——语言模型θ_w 生成图像 I 的标注词的概率,隐含考虑了词和词的相关性.与使用外部知识源 WordNet[6,10]进行词与词相关性不同,该方法在标注过程中利用训练集中词和词的关系进行标注,并且其在一定程度上改善了标注性能.但该文并未给出度量任意两个词的相似性的有效方法,同时对词的“邻居”关系蕴含的语义相似性并未考虑,且由于使用 EM 算法,标注速度较慢.

国内相关的研究工作包括 Li 等人[13]使用条件随机场进行图像标注的半指导学习,以解决自动标注中训练数据稀少的问题.Wang 等人[14]利用 Web 检索和数据挖掘技术设计图像标注系统 AnnoSearch.在该系统中,首先使用 Web 检索得到语义和视觉相似的一组图像,再利用挖掘方法从图像相关的文本描述中得到它们的标注.Hua 等人[15]设计 Web 图像标注系统,利用 Web 图像的上下文信息自动获取相关的语义信息.Liu 等人[7]提出基于流行排序(manifold ranking)学习的图像标注方法,在该方法中,设计 NSC 模式生成自适应相似性图,并通过词的语义相关性信息对自适应图标注的结果进行扩展和不相关词的过滤.

2 标注的基本框架

基于统计生成模型的相关模型使得图像标注性能有较大提高,而本文提出的标注方法在相关模型的基础上进行了进一步扩展,因此,下面首先对图像自动标注的相关模型进行介绍.

2.1 相关模型

相关模型标注的基本思想是:估计概率 $P(w|I)$——给定图像 I 时,单个词 w 作为标注的概率,通过对 $P(w|I)$ 排序选择标注词.

给定训练集 T,集合的大小记为$|T|$,对训练集中每幅已标注的图像 J_i 可使用图像区域和标注词来表示,如

$J_i=\{f_{i,1},f_{i,2},\ldots,f_{i,m};w_{i,1},w_{i,2},\ldots,w_{i,n}\}$,$m$ 和 n 分别表示图像区域的个数和词的总个数.对不同的图像,m 的个数不一定相等,n 是相等的.$f_{i,j}$ 是区域特征,维数为 D;$w_{i,j}$ 是一个二元变量,表示第 j 个词是否出现在第 i 幅图像中.给定一幅待标注图像 $I=\{f_1,f_2,\ldots,f_t\}$,我们需要估计概率 $P(w|I)$,即给定图像 I 时词 w 作为其标注的概率.

$$P(w\mid I)\propto P(w,I)=\sum_{i=1}^{|T|}P(w,I\mid J_i)P(J_i) \tag{1}$$

即

$$P(w\mid I)\propto P(w,I)=\sum_{i=1}^{|T|}P(w\mid J_i)P(I\mid J_i)P(J_i) \tag{2}$$

即

$$P(w\mid I)\propto P(w,I)=\sum_{i=1}^{|T|}\prod_{j=1}^{t}P(f_j\mid J_i)P(w\mid J_i)P(J_i)\ ,$$

则待求的最佳标注为

$$w^*=\text{argmax}_w P(w|I) \tag{3}$$

假设 $P(J)$服从均匀分布,则需要估计被标注图像 I 的每个区域 f_j 从训练集中任意图像 J 生成的概率 $P(f_j|J)$服从的分布及 w 由 J 生成的概率 $P(w|J)$服从的分布.

2.2 连续特征生成概率估计

在连续模型CRM和MBRM中,对区域特征生成概率 $P(f_j|J)$服从的分布使用基于高斯核的非参数估计得到:

$$P_B(f_j\mid J)=\frac{1}{m}\sum_{k=1,g_k\in J}^{m}\frac{\exp\{-(g_i-f_j)^T\Sigma^{-1}(g_i-f_j)\}}{\sqrt{2^D\pi^D\mid\Sigma\mid}} \tag{4}$$

其中,g_k 表示训练图像 J 的第 i 个区域的特征,m 是 J 中的区域个数.忽略分母的归一化因子,该核密度函数也可以看作特征相似性度量函数,则上式表示区域 f_j 和图像 J 的相似性等于 f_j 和 J 的每个区域作相似性度量,然后取平均.如在图 1 中,估计图像 a 中标识区域从图像 a'中生成的概率时,需将其与图像 a'中背景区域(如 tree 和 sky)进行相似性比较.然而,在相关领域,如图像检索系统 SIMPLIcity[19]及 FUZZYCLUB[20]中,一个基本思想是:区域(对象)与图像的相似性度量取决于对象与图像中对象区域的相似性度量值,与图像中背景关系不大,这样的度量方式更符合人的判断习惯.图像分割后,真正携带语义信息的区域往往在另一幅图像中只与其中语义相似的某个区域相似,按照式(4)计算出的区域生成概率却比较低.由此可见,在考虑分割的语义特性时,式(4)定义的 f_j 与 J 的生成概率计算方法忽视了图像语义相似性度量的复杂性.在 SIMPLIcity 中,定义了加权的区域相似性之和来度量两幅图像的相似性.而 FUZZYCLUB 对其作进一步改进,定义区域与图像的相似性为该区域和另一幅图像的所有区域中相似性度量最大的值,取得了较好的效果.

(a) Image a (b) Image a' (c) Image a (d) Image aa
(a) 图像 a (b) 图像 a' (c) 图像 a (d) 图像 aa

Fig.1 The examples of images with high generating probability
图 1 具有高生成概率的图像举例

2.3 基于最大相似区域的特征生成概率估计

然而,将上述基于区域的图像相似性度量的思想应用于图像标注中并不是一件简单的事情.例如:根据FUZZYCLUB 的思想,我们可定义如下基于最大相似区域的特征生成概率估计:

$$P_S(f_j|J)\propto\max\exp(-(f_j-g_i)^T\Sigma^{-1}(f_j-g_i)) \tag{5}$$

定义区域 f_j 由 J 生成的概率取决于 f_j 由 J 中所有区域距离度量最小的区域生成的概率,即只要 J 中包含与 f_j 较为相似的一个区域,则该生成概率比较高.但是,目前的图像分割技术易产生一些语义信息不明确或根本无意义的小区域,这些区域中多数只是语义对象的组成部分.由于它们的弱语义特性以及不突出的视觉特征,使得它们更容易和另外图像的多个区域都有较高的视觉相似性.根据式(5),这些弱区域有较高的生成概率,从而对图像的生成概率起到误导作用.如图像 a 中弱语义区域(标识区域)容易和图像 a'中标识区域产生较高的视觉相似性,从而产生较高的生成概率,虽然图像 a 与图像 a'从视觉上看并不相似,语义也不相关.而图像 a 和图像 aa 包含共同对象“building”,按式(5)计算出的生成概率却比较低.

为了降低弱语义含义的区域在图像生成概率估计中的作用,我们对区域相似性关系进行结构化,使用最大权匹配算法找到代表两幅图像整体相似性的区域匹配的核心结构,并在该结构的基础上提出一种新的区域生成概率密度估计方法.

3 改进的区域特征生成概率估计与平滑方法

从图像生成的角度来看,图像 I 从另一幅图像 J 生成的概率由 I 中每个区域的联合生成概率决定,若其中某个区域的生成概率太大,将会掩盖其他区域的贡献,从而影响整体图像生成概率的估计的准确性.因此,为了削弱语义区域的影响,同时保留每个区域对图像生成概率的贡献,需考虑如下两个约束:区域与图像相似性仍由区域与其在该图像中的最优匹配来决定;同时,一个区域的最优匹配应尽可能少地共享给其他区域.对区域相似性关系进行结构化,建立图像 I 和 J 的区域匹配结构,使得该结构在满足如上约束时最大化两幅图像的整体相似性十分必要,而该结构可通过基于二分图的最大权匹配算法获得.图 2 举例说明:图像 I(左侧)和图像 J(右侧)经图像分割后的区域分别构成二分图的两组结点,图中加粗的边为该二分图的最大权匹配,最大权匹配中以边相连的区域对互为最优匹配,如图中区域对(b_2, b'_t).除去 J 中区域数小于 I 中区域数的情况,I 中每个区域不能在 J 中有共同的最优匹配.

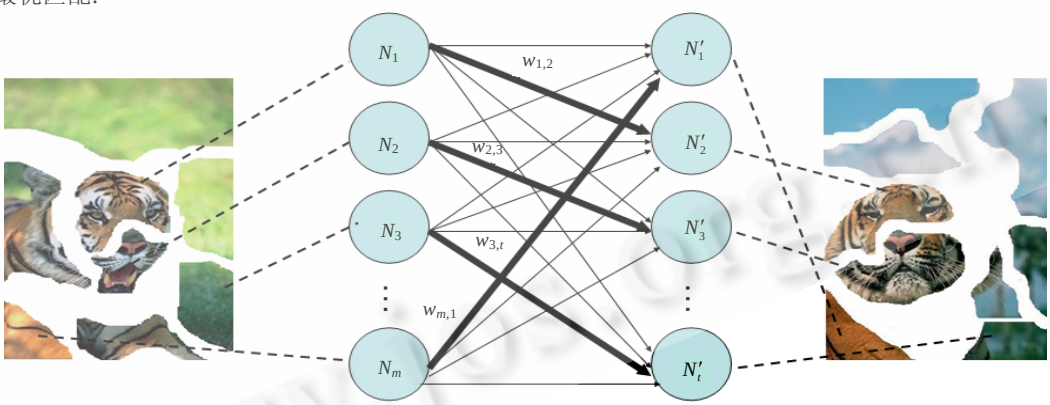

Fig.2 Maximum weight matching, the thick edges denote the optimal matching, the weight associated with the edge denotes the similarity between the region in image I and image J

图 2 最大权匹配,图中加粗的边即找到的最优匹配,该边上的权值表示 I 中区域由 J 生成的概率

设图像 I,J 经过图像分割后形成的区域分别是$\{b_1,\ldots,b_m\}$和$\{b'_1,\ldots,b'_t\}$.建立两个结点集 $N=\{N_1,\ldots,N_m\}$和 $N'=\{N'_1,\ldots,N'_t\}$,其结点分别对应 I 和 J 的区域,建立二分图 $G=\{N\cup N',E\}$,$E=N\times N'$.边 (N_i,N'_j) 表示可能的区域匹配对,边上的权值由 N_i 对应区域 b_i 从 N'_j 对应区域 b'_j 生成的非参估计值给出,该值也可以理解为区域匹配对的相似程度:

$$w_{i,j}=\frac{\exp\{\{-(g_i-f_j)^T\Sigma^{-1}(g_i-f_j)\}\}}{Z_w} \tag{6}$$

其中,g_i,f_j 分别对应 b_i, b'_j 的 D 维视觉特征,Z_w 是归一化因子.

在构造如上带权完全二分图后,接下来给出该二分图下匹配与最大权匹配的定义:一个完全二分图的匹配 M 是指 E 的一个子集,该子集中任意两条边不含公共结点.完全二分图的最大权匹配是一个匹配,同时满足如下两个条件:N'中结点均包含在该匹配中;该匹配包含的边权的和最大,即

$$\max\left(\sum_{N_i\in N} w_{i,\pi(i)}\right) \tag{7}$$

如上定义的最大权匹配在结点集 N 和 N'包含的结点数目相同时成立.但在图像标注中,图像分割后的区域个数不一定完全相同,当图像 I 的区域个数大于 J 时,既要求最大权匹配的任意两条边不含公共结点,又要求 N 中结点全部包含在该匹配中,找不到这样的匹配.此时将条件修改为:保证 N 中结点在 N'中共享最少公共最优匹配结点的同时,最大化所有权值的和.此时,最大权匹配在最大化两幅图像整体相似性的前提下减少区域对整体相似性的贡献次数,较好地满足了我们的需要.

求解上述最大权匹配问题的常用方法是匈牙利算法[24],通过匈牙利算法可以找到 I 中区域在 J 中的最优匹配,这样,区域 b_i 由图像 J 生成的概率定义为 b_i 从其在 J 中的最优匹配生成的概率,即

$$P_{opt}(b_i|J)=w_{i,k},k=\pi(i) \tag{8}$$

但上式得到的概率值是仅用最优匹配区域进行核密度估计的结果,由于估计样本集非常稀疏(仅包含一个样本,最优匹配区域),因此会导致低偏差、高方差,从而总体误差较大[25].为此,我们使用 Jelinek-Mercer 平滑方法[26]进行平滑.具体做法:将整幅图像的所有区域作为背景集,在此基础上,做基于核的非参数概率密度估计 $P_B(b_i|J)$,最终可以得到:

$$P(b_i|J)=\lambda P_{opt}(b_i|J)+(1-\lambda)P_B(b_i|J) \tag{9}$$

其中,λ是平滑因子,最优值可在验证集中确定,本文在实验中设为 0.7.通过区域匹配,在对图 1 中图像 a 进行标注时,减少了弱语义区域对整体生成概率的贡献,从而为其标注正确的语义标签“building”.

4 词的语义相关性度量

文本检索中,自动局部分析[18]利用与初始查询项(query term)相关度最高的局部文档集合中词与词的共现(co-occurrence)相关性关系,进行查询扩展.在图像标注中,将训练集中每幅图像看成包含标注词的文档,本文采用类似方法,针对待标注图像自适应构造局部训练集进行语义相关性度量,从而将主题相关的词加入待标注图像的标注中.具体如下:

对于给定待标注图像,我们利用视觉相似性选择前 K 个具有最大视觉生成概率的训练图像组成其局部视觉邻域.将邻域中每幅训练图像用标注词序列表示.建立矩阵 $M=K\times|T|$,M_{ij} 表示第 i 个词是否出现在第 j 幅图像中.例如,当标注单词表包含 5 个单词:V={steet,bridge,bus,train, people},局部训练集合中共有 4 幅图像:$T=\{J_1,J_2,J_3,J_4\}$,标注与训练图像的关系由图 3 左矩阵表示:

	J_1	J_2	J_3	J_4
w_1 (street)	1	1	1	0
w_2 (bridge)	0	0	0	1
w_3 (bus)	1	1	0	0
w_4 (train)	0	0	0	1
w_5 (people)	1	1	0	1

⇒ Co-occurrence correlation

	w_1	w_2	w_3	w_4	w_5
w_1	3	0	2	0	2
w_2	0	1	0	1	1
w_3	2	0	2	0	2
w_4	0	1	0	1	1
w_5	2	1	2	1	3

Fig.3 The example of word correlation analysis

图 3 词的语义相关性分析举例

将矩阵的每一行看作一个向量 $\vec{w}$,第 i 行表示词 w 在训练集中的出现模式.矩阵的第 j 列表示图像 J_j 的标注,若 w_i 是图像 J_j 的标注,则 M_{ij} 为 1,否则为 0.此时,词 w_i 与 w_j 的共现相关性可由相应行向量的内积 $\vec{w}_i\cdot\vec{w}_j$ 来度量.

由如上度量方式可得图 2 右矩阵,可以看出“street,bus”,“street,people”相关性较高,即“street,bus”,“street,people”频繁共同出现,则它们对未标注图像进行标注时共同出现的概率较高

鉴于 M 矩阵非常稀疏,为了避免在下面的计算中出现多 0 的情况,我们首先将整个局部训练集当成背景集,对 M_{ij} 使用 Jelinek-Mercer 方法[26]进行平滑.接下来,通过词在局部训练集中的共现模式进行语义相关性度量,词

w_u 与 w_v 的共现相关性可按下式定义:

$$c_{w_u,w_v}=\sum_{J_j\in T}M_{uj}\times M_{vj} \tag{10}$$

对 c_{w_u,w_v} 进行归一化处理:

$$s_{w_u,w_v}=\frac{c_{w_u,w_v}}{c_{w_u,w_u}+c_{w_v,w_v}-c_{w_u,w_v}} \tag{11}$$

将所有 w_u 的相关性值组成向量 $\vec{s}_{w_u}$,即 $\vec{s}_{w_u}=(s_{w_u,1},s_{w_u,2},...,s_{w_u,n})$.同样地,将所有 w_v 的相关性值组成向量 $\vec{s}_{w_v}$,即 $\vec{s}_{w_v}=(s_{w_v,1},s_{w_v,2},...,s_{w_v,n})$,通过 $\vec{s}_{w_u}$ 和 $\vec{s}_{w_v}$ 的内积可度量 w_u 和 w_v 的上下文共现模式:

$$Sim(w_u,w_v)=\frac{\vec{s}_{w_u}\cdot\vec{s}_{w_v}}{|\vec{s}_{w_u}|\times|\vec{s}_{w_v}|} \tag{12}$$

若 w_u 和 w_v 具有相似的“邻居”,即它们出现的上下文较相似,则得到的 $Sim(w_u,w_v)$值较大.如上例中词“bridge”和“train”,虽然共现频率低,但它们具有相似的邻居关系,则这两个词也具有较强的语义相关性.最终的 $Sim(w_u,w_v)$反映了 w_u 和 w_v 的语义相关性.

5 本文标注方法

至此,本文给出新的图像区域特征生成概率估计方法,并且进行词与词相关性的有效度量.接下来,本文设计标注算法,同时利用这两方面的信息改进标注效果.为了说明算法的有效性,本文首先将独立词标注的相关模型扩展为基于词集标注,进而设计启发式迭代函数对目标函数进行求解.在启发式迭代过程中,新的图像区域生成概率估计提高为迭代选择一个好的起点,进而词与词相关性保证后续标注紧密耦合,最终提高图像标注与图像特征的语义相关性.

5.1 标注目标函数

基于统计生成模型的相关模型假设词与词独立,标注过程中每次选择一个独立的词进行标注,此时,标注目标函数如式(3)所示.本文标注算法将其扩展为对词集进行标注.设 S_k 是标注单词表 V 的一个大小为 k 的子集,则给定待标注图像 I,S_k 是其标注的概率为

$$\begin{aligned}&P(S_k\mid I)\Leftrightarrow P(S_k,I)\\&P(S_k,I)=P(w_k,S_{k-1},I)=P(w_k\mid I,S_{k-1})P(S_{k-1},I)\end{aligned} \tag{13}$$

其中,$S_{k-1}\cup w_k=S_k$,$k\geq 1$,则待求的最优标注为

$$S_k^*=\arg\max_S P(S_k,I) \tag{14}$$

很明显,该问题不是一个简单的线性规划问题,因此很难找到分析解决方案.并且,当$|V|$非常大时,使用穷举搜索的方法在计算上并不可行.我们使用贪心算法得到近似最优解.对上式两边同取对数得到:

$$\log P(S_k,I)=\log P(S_{k-1},I)+\log P(w_k|I,S_{k-1}) \tag{15}$$

此时,令 $f(S_k)=\log P(S_k,I)$,由于对数函数的保序性,对式(13)的求解等价于对 $f(S_k)$的求解.此时,可定义贪心算法迭代公式如下:

$$f(S_k)=f(S_{k-1})+\log P(w_k|I,S_{k-1}) \tag{16}$$

则贪心算法每一步要找的 w^*满足下式:

$$\begin{aligned}w^*&=\arg\max_w f(S_k)-f(S_{k-1})\\&=\arg\max_w \log P(w\mid I,S_{k-1})\\&=\arg\max_w P(w\mid I,S_{k-1})\end{aligned} \tag{17}$$

假设 I 与 S_{k-1} 相互独立,则有 $P(w\mid I,S_{k-1})=\frac{1}{P(w)}P(w\mid I)P(w\mid S_{k-1})$,设 $P(w)$服从均匀分布,则对于所有的词 w,$P(w)$均相同,此时有

$$w^*=\arg\max P(w|I)P(w|S_{k-1})=\arg\max P(w,I)P(w,S_{k-1}) \tag{18}$$

其中,对 $P(w|S_{k-1})$的极大似然估计为 $P_M(w|S_{k-1})=\frac{\#\{J|w,S_{k-1}\in J\}}{\#\{J|S_{k-1}\in J\}}$,$\#\{J|w,S_{k-1}\in J\}$表示训练集中标注同时包含 w,S_{k-1} 的图像数.对于有限的训练集,当$|S|$增大时,该 w 和 S_{k-1} 共同出现的次数非常少,因此,该概率将出现很多为 0 的情况.然而,在当前模型中未出现的情况不代表以后也不会出现,因此须对其进行有效平滑.在信息检索中,平滑通常在一个大的背景集合中进行,该背景集合包含更多信息,从而为在当前模型中未出现的情况分配一个小的概率,调整当前极大似然估计使其更加准确[26].如我们可以选择一个更大的训练图像集加以平滑.然而,根据 Zipf 法则,总有很多词在图像中出现的次数很少[8],这些不常出现的词的组合共同出现的次数更少,因此,即便使用更大的背景集,仍然无法解决上述稀疏问题.为此,我们定义如下方法为极大似然估计 $P_M(w|S_{k-1})$中概率为 0 的项分配调整值,以达到平滑的目的.

定义 $relation(w,S_{k-1})$,度量词 w 与词集 S_{k-1} 的相关性:

$$\begin{cases} relation(w)=1, & |S|=1 \\ relation(w,S_{k-1})=\sum_{w'\in S_{k-1}} Sim(w,w'), & |S|>1 \end{cases} \tag{19}$$

其中,$Sim(w,w')$是在第 4 节介绍的词 w 和 w'的语义相似性,为保证 $relation(w,S_{k-1})$在[0,1]之间且满足概率性质,即 $\sum_{w\in V,w\notin S_{k-1}} relation(w,S_{k-1})=1$,对其进行归一化.令归一化后的 $relation(w,S_{k-1})$作为对概率 $P(w|S_{k-1})$的平滑.即

$$P(w|S_{k-1})=(1-\gamma)P_M(w|S_{k-1})+\gamma relation(w,S_{k-1}) \tag{20}$$

其中,γ为平滑因子,当$|S|$增大时,$relation(w,S_{k-1})$将对该概率 $P(w|S_{k-1})$起决定作用,此时,γ接近于 1.

对 $p(w,I)$的计算采用类似于 MBRM[3]中的做法,由于每个词在一幅图像的标注中只出现 1 次,因此,对词的分布使用二项式分布,具体如下:

$$p(w,I)=\sum_{i=1}^{|T|}p(w,I|J_i)p(J_i)=\sum_{i=1}^{|T|}\prod_{j=1}^{m}p(f_j|J_i)p(w|J_i)\prod_{w\neq w'}(1-p(w'|J_i)) \tag{21}$$

对 $p(w|J_i)$的平滑,使用二项式分布的共轭分布——Beta 先验[3].

5.2 标注过程

本文提出的标注算法其迭代过程如下.

算法 1. Anno_Ext.

输入:待标注图像 I,标注单词表 V,关键词相关性度量值 Sim,固定标注长度 k;

算法过程:

- 初始化 $S_0=\varnothing$
- For i=1,2,…, $|T|$
  - 根据式(9)计算生成概率 $P(I|J_i)$
- EndFor
- 选择前 K 个具有最大 $P(I|J_i)$的训练图像组成局部视觉邻域,进行关键词语义相关性分析
- For i=1,2,…,k
  - 计算 $w^*=\arg\max f(S_{i-1}\cup w)-f(S_{i-1})$
  - 令 $S_i=S_{i-1}\cup w^*$
- EndFor

输出:标注 S.

通过下面的式子,我们可以进一步分析贪心算法每一步找到的 w^*

$$w^*=\arg\max f(S_{i-1}\cup w)-f(S_{i-1})=\arg\max p(w,I)p(w,S_{i-1}) \tag{22}$$

此时,S_{i-1} 表示确定已标注词.可以看出,本文的标注方法具有以下特点:

- 由于 $P(w|S_{i-1})$受 $relation(w,S_{i-1})$主导,因此,本文标注算法能够有效地将提高生成概率估计以及词的相关性度量结合在一起以改进标注结果.

- 本文迭代标注算法第 1 个词的选择仅取决于词 w 和待标注图像 I 的视觉特征的联合概率 $p(w,I)$,因此,改进的生成概率估计将为应用词的相关性提供良好的起点.
- 在良好起点的基础上,本文通过度量词的语义相关性,在迭代过程中考虑与已标注词的语义相关性,从而通过关键词的语义相关性减少因不准确的特征估计带来的错误标注.

6 实 验

6.1 实验建立

我们在实验中使用的 Corel 数据集取自 ECCV 2002 基准数据集[1].该数据集包括 5 000 幅图像,来自 50 个 Corel Stock Photo CDs.每个 CD 目录下包含同一主题的 100 幅图像.每幅图像与 1~5 个标注词关联,共有 374 个词.我们将数据集分为 3 部分:训练集 4 000 幅,验证集 500 幅,测试集 500 幅图像.其中,验证集包括每个目录下的 10 幅图像,主要用来确定模型参数.参数确定后,验证集的数据加到训练集中形成新的训练集重新训练模型.与之前的方法一样,我们主要用检索单个词的查全率、查准率和 F_1 来度量标注的性能好坏.给定查询词 w,若存在测试集中手工标注结果中包含 w 的图像个数为 $|W_G|$,使用自动标注模型的标注结果中包含该词的图像个数为 $|W_M|$,其中,$|W_C|$ 个是正确的,则

$$\text{Recall}=\frac{|W_C|}{|W_G|},\ \text{Precision}=\frac{|W_C|}{|W_M|},\ F_1=\frac{2\times Recall\times Precision}{Recall+Precision}.$$

Recall 度量出对单个词查询的完整性,Precision 度量查询的精度,平均的查准率和查全率则反映出标注整体的性能.

由于连续相关模型 MBRM 代表了目前较好的标注水平,实验中,我们将本文标注算法与 MBRM 标注模型进行了比较.由于本文研究工作是建立在图像分割的基础上,故下面的实验比较若无特殊声明,则 MBRM 基于图像分割,非参数核密度估计均选择高斯核.我们使用相同的训练集对这些模型进行训练,并在相同的测试集上进行测试.固定标注长度 k 设为 5. Recall 和 Precision 均是在整个测试集出现的 263 个词上计算得到的.其中,本文词的相关性估计局部邻域个数 K 选择为 25.

6.2 性能度量

6.2.1 本文标注算法整体性能度量

表 1 是本文标注算法与 MBRM 的标注结果的比较.其中,Anno_MSR 表示在本文提出的启发式迭代标注框架中,使用最大相似区域的特征生成概率估计结合词与词相关性;Anno_MWRM 表示使用基于最大权匹配的特征生成概率估计结合词与词相关性.可以看到,与 MBRM 相比,本文提出标注方法在 Recall 和 Precision 上均有所提高,最大提高幅度分别为 21%和 11%.使用最大相似区域的特征生成概率估计结合词与词相关性,Recall 和 Precision 分别由 16.1%和 19.0%提高到 17.8%和 21.6%.而使用最大权匹配的密度估计结合词与词相关性,Recall 和 Precision 分别提高到 19.5%和 21.1%.由于本文提出的算法中有两方面改进,一是连续特征估计方法的改进,二是对词的相关性度量,因此需分别检查这两方面的有效性,并分析这两方面对改进标注性能所起到的作用.

Table 1 The annotation performance comparison between our algorithm and MBRM

表 1 本文标注算法与 MBRM 的比较

	Ave.Recall (%)	Ave.Precision (%)
MBRM	16.1	19.0
Anno_MSR	17.8	21.6
Anno_MWRM	19.5	21.1

6.2.2 基于最大权匹配的连续特征估计有效性检验

表 2 给出特征生成概率估计方法的实验结果,假设在词和词独立的情况下,将本文基于最大权匹配的连续特征估计方法(MWRM)分别与离散特征估计(CMRM)、连续非参估计方法(MBRM)、基于最大相似区域的连续特征估计方法(MSR)进行比较.其中,MSR 连续特征估计方法使用式(5)来计算,并使用式(9)进行平滑.从图 4

看出,基于连续特征估计的标注性能远远高于离散特征模型,而单纯使用 MSR,其标注效果略低于 MBRM.这是由于离散特征模型过多地依赖于聚类粒度的选择,事先很难选择一个合适的聚类粒度;而基于 MSR 的生成概率估计方法无法对弱语义区域进行有效处理.本文提出的基于最大权匹配的特征生成概率估计通过最大权匹配结构,从整体上降低了无关区域在相似性度量中所起的作用,从而使生成概率估计更加准确,也使得图像标注的性能有所提高,其 Recall 和 Precision 分别由 16.1%和 19.0%提高到 18.3%和 19.8%.

Table 2 The effectiveness of maximum weight matching based probability estimation

表 2 基于最大权匹配的连续特征密度估计有效性检验

	Ave.Recal *l*(%)	Ave.Precision (%)
CMRM	10.7	8.7
MBRM	16.1	19.0
MSR	15.8	18.1
MWRM	18.3	19.8

6.2.3 词与词相关性度量有效性检验

在本组实验中,我们对区域特征的生成概率使用类似相关模型的做法,即使用整幅图像作为样本集作非参数估计,在词的估计上与其假设词和词独立不同,将词与词语义相关性度量值应用到本文的贪心迭代过程中标注,其标注性能见表 3.图中,WCor 表示本文方法,可以看到,Recall 和 Precision 分别由 16.1%和 19.0%上升到了 18.6%和 19.8%.

Table 3 The effectiveness of words correlation measure

表 3 词与词相关性有效性检验

	Ave.Recall (%)	Ave.Precision (%)
MBRM	16.1	19.0
WCor	18.6	19.8

6.2.4 基于最大权匹配的特征生成概率估计为迭代算法提供较好起点有效性验证

在第 5.2 节中,我们分析了第 1 个语义关键词与视觉特征生成概率估计紧密相关.为了验证改进的特征生成概率是否能够为本节启发式迭代算法提供较好的起点,将词的相关性度量分别应用于以往连续特征估计方法(MBRM+Cor),基于最大相似区域的特征生成概率估计(MSR+Cor)以及基于最大权匹配的特征生成概率估计(MWRM+Cor)中,表 4 给出了实验结果.结合表 2 可以看出,单纯使用基于最大相似区域密度估计方法估计区域特征生成概率,其标注性能略低于 MBRM,但从表 4 看到,结合词与词相关性后,Precision 有所提高.而单纯使用基于最大权匹配的连续特征估计方法其标注性能有所提高,可以看出,词与词的相关性度量与基于最大权匹配的密度估计对标注都起到积极作用.而如表 4 第 3 行所示,使用基于最大权匹配的特征生成概率估计词的相关性分析,整体的 Recall 和 Precision 均有显著提高,说明基于最大权匹配的特征估计方法能够为本文启发式贪心算法找到好的起点,使得词的相关性信息应用在较为正确的基础上,进而两者相互结合共同提高整体标注性能.

Table 4 The effectiveness of the start point provided by MWRM-based probability estimation

表 4 基于最大权匹配的特征生成概率估计为迭代算法提供较好起点有效性验证

	Ave.Recall (%)	Ave.Precision (%)	F_1
MBRM + Cor	18.6	19.8	0.192
MSR + Cor	17.8	21.6	0.195
MWRM + Cor	19.5	21.1	0.203

6.2.5 算法时间复杂度分析

算法 1 的时间复杂度主要取决于每次迭代过程中 w^*的计算,而 w^*的计算可分为下面 3 部分时间复杂度之和:图像特征生成概率估计($p(I|J)$)、词的生成概率估计($p(w|J)$)以及词的相关性度量.由于图像生成概率等于区域生成概率的乘积,而每幅图像包含的区域个数固定,因此,第 1 部分时间复杂度主要取决于区域特征生成概率估计.设训练样本数为 n,本文在生成概率计算时使用匈牙利算法得到区域最大权匹配.因此,第 1 部分总的时间复杂度为 $O(n)\times O(Hun)$,其中,$O(Hun)$表示匈牙利算法的时间复杂度,有 $O(Hun)=O(|V|(|E|+|V|\log|V|))$,$|V|$是结点数,

$|E|$为边数.一般情况下,图像区域分割的数目是非常有限的,如本文实验中图像分割区域数最大为 10,即结点数不超过 20,远远小于 n.因此,生成概率估计的时间复杂度为 $O(n)$.设标注单词表中词的个数为 m,则词的生成概率估计时间复杂度为 $O(mn)$,词的相关性度量时间复杂度为 $O(m^2)$.得到 3 部分时间复杂度的和为

$$O(n)+O(mn)+O(m^2)).$$

通常情况下,单词表中的单词数目远远小于训练样本数,因此,该时间复杂度由 $O(n)$主导.当固定标注长度为 k 时,算法 1 总的复杂度为 $kO(n)=O(n)$.

7 总结和展望

相关模型在图像标注领域显示了其良好的性能,但其隐含的定义区域与图像的相似性为区域与图像中所有区域的平均相似性,忽略了对象与区域的语义特性.本文通过二分图最大权匹配算法得到两幅图像区域匹配的核心结构,在此基础上,结合统计平滑技术提出基于最大权匹配的区域密度概率估计方法.另一方面,相关模型中假设词与词相互独立,但实践表明,词与词的相关性能够改进标注性能.本文针对训练集,对词与词语义相关性进行有效度量,并设计启发式贪心算法对词与图像特征的联合概率以及词与词的关系综合考虑.基于最大权匹配与平滑技术的区域密度概率估计方法为启发式贪心算法找到了好的起始点,词与词相关性使得后续标注词语义紧密相关.实验表明,本文所提出方法使得标注性能有显著提高.鉴于词和词的关系在标注过程中所起的重要作用,而本文仅对训练集进行分析,训练集中包含的图像是有限的,如此学到的知识是有限的.在下一步的工作中,我们考虑结合外部知识源,如 WordNet,学习到更多关于词和词相关性的知识.另外,还考虑将现有工作扩展到基于区域(对象)的图像标注和检索中.

References:

[1] Duygulu P, Barnard K, de Freitas JFG, Forsyth DA. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Heyden A, ed. Proc. of the European Conf. on Computer Vision. Berlin: Springer-Verlag, 2002. 97−112.

[2] Jeon J, Lavrenko V, Manmatha R. Automatic image annotation and retrieval using cross-media relevance models. In: Proc. of the Int'l ACM SIGIR. Toronto: ACM Press, 2003. 119−126.

[3] Feng SL, Manmatha R, Lavrenko V. Multiple Bernoulli relevance models for image and video annotation. In: Proc. of the IEEE Conf. Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2004. 1002−1009.

[4] Lavrenko V, Manmatha R, Jeon J. A model for learning the semantics of pictures. In: Sebastian T, Lawrence KS, Bernhard S, eds. Proc. of the Neural Information Processing Systems (NIPS). Vancouver, Whistler: MIT Press, 2004. 553−560

[5] Jin R, Chai JY, Si L. Effective automatic image annotation via a coherent language model and active learning. In: Henning S, Nevenka D, eds. Proc. of the Int'l Conf. on ACM Multimedia. New York: ACM Press, 2004. 892−899.

[6] Jin Y, Khan L, Wang L, Awad M. Image annotations by combining multiple evidence & WordNet. In: Zhang HZ, Chua TS, eds. Proc. of the ACM Int'l Conf. on Multimedia. Singapore: ACM Press, 2005. 706−715.

[7] Liu J, Li MJ, Ma WY, Liu QS, Lu HQ. An adaptive graph model for automatic image annotation. In: James ZW, Nozha B, eds. Proc. of the ACM SIGMM Int'l Workshop on Multimedia Information Retrieval. Santa Barbara: ACM Press, 2006. 61−69.

[8] Monay F, Gatica-Perez D. On image auto-annotation with latent space models. In: Lawrence AR, Harrick MV, Thomas P, Prashant JS, John RS, eds. Proc. of the ACM Int'l Conf. on Multimedia. Berkeley: ACM Press, 2003. 275−278.

[9] Monay F, Gatica-Perez D. PLSA-Based image auto annotation: Constraining the latent space. In: Henning S, Nevenka D, eds. Proc. of the Int'l Conf. on ACM Multimedia. New York: ACM Press, 2004. 348−351.

[10] Srikanth M, Varner J, Bowden M, Moldovan D. Exploiting ontologies for automatic image annotation. In: Ricardo ABY, Nivio Z, Gary M, Alistair M, John T, eds. Proc. of the SIGIR. Salvador: ACM Press, 2005. 552−558.

[11] Gao YL, Fan JP, Xue XY, Jain R. Automatic image annotation by incorporating feature hierarchy and boosting to scale up SVM classifiers. In: Klara N, Matthew T, Yong R, Wolfgang K, Ketan MP, eds. Proc. of the ACM Int'l Conf. on Multimedia. Santa Barbara: ACM Press, 2006. 901−910.

[12] Li J, Wang JZ. Real-Time computerized annotation of picture. In: Klara N, Matthew T, Yong R, Wolfgang K, Ketan MP, eds. Proc. of the ACM Int'l Conf. on Multimedia. Santa Barbara: ACM Press, 2006. 911−920.

[13] Li W, Sun MS. Semi-Supervised learning for image annotation based on conditional random fields. In: Hari S, Milind RN, John RS, Yong R, eds. Proc. of the Conf. on Image and Video Retrieval. LNCS, 2006. 463−472.

[14] Wang XJ, Zhang L, Jing F, Ma WY. AnnoSearch: Image auto-annotation by search. In: Hari S, Milind RN, John RS, Yong R, eds. Proc. of the Conf. on Image and Video Retrieval. LNCS, 2006. 1483−1490.

[15] Hua ZG, Wang XJ, Liu QS, Lu HQ. Semantic knowledge extraction and annotation for Web images. In: Zhang HZ, Chua TS, eds. Proc. of the ACM Int'l Conf. on Multimedia. Singapore: ACM Press, 2005. 467−470.

[16] Shi R, Chua TS, Lee CH, Gao S. Bayesian learning of hierarchical multinomial mixture models of concepts for automatic image annotation. In: Hari S, ed. Proc. of the Conf. on Image and Video Retrieval. LNCS, 2006. 102−112.

[17] Shi J, Malik J. Normalized cuts and image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000, 22:888−905.

[18] Yates RB, Neto BR. Modern Information Retrieval. New York: ACM Press, 1999. 123−129.

[19] Wang JZ, Li J, Wiederhold G. SIMPLIcity: Semantics-Sensitive integrated matching for picture LIbraries. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2001,23(9):947−962.

[20] Zhang RF, Zhang ZF (Mark). A clustering based approach to efficient image retrieval. In: Proc. of the 14th IEEE Conf. on Tools with Artificial Intelligence (ICTAI). Washington: IEEE Computer Society, 2002. 339−346.

[21] Wang T, Rui Y, Sun JG. Contraint based region matching for image retrieval. Int'l Journal of Computer Vision, 2004,56(1-2): 37−45.

[22] Bi JB, Chen YX. A sparse support vector machine approach to region-based image categorization. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. San Diego: IEEE Computer Society, 2005. 1121−1128.

[23] Zhang Q, Goldman SA, Yu W, Fritts J. Content-Based image retrieval using multiple-instance learning. In: Claude S, Achim GH, eds. Proc. of the 19th Int'l Conf. on Machine Learning. 2002. 682−689.

[24] Kuhn HW. The hungarian method for the assignment problem. Naval Research Logistics Quarterly, 1955,2:83−97.

[25] Hastie T, Tibshirani R, Friedman J. The Element of Statistical Learining; Data Mining, Inference, and Prediction. New York: Springer-Verlag, 2001. 172−173.

[26] Zhai CX, Lafferty J. A study of smoothing methods for language models applied to information retrieval. ACM Trans. on Information Systems, 2004,22(2):179−214.

王梅(1980－),女,上海人,博士生,主要研究领域为多媒体数据库,信息检索.

许红涛(1980－),男,博士生,主要研究领域为 Web 数据管理.

周向东(1969－),男,博士,副教授,主要研究领域为数据库,信息检索.

施伯乐(1935－),男,教授,博士生导师,CCF 高级会员,主要研究领域为数据库理论与应用.

张军旗(1979－),男,博士生,主要研究领域为多媒体数据库,信息检索.