

一种自适应的 Web 图像语义自动标注方法^{*}

许红涛, 周向东⁺, 向宇, 施伯乐

(复旦大学 计算机科学技术学院, 上海 200433)

Adaptive Model for Web Image Semantic Automatic Annotation

XU Hong-Tao, ZHOU Xiang-Dong⁺, XIANG Yu, SHI Bai-Le

(School of Computer Science and Technology, Fudan University, Shanghai 200433, China)

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

Xu HT, Zhou XD, Xiang Y, Shi BL. Adaptive model for Web image semantic automatic annotation. *Journal of Software*, 2010,21(9):2183–2195. <http://www.jos.org.cn/1000-9825/3658.htm>

Abstract: This paper proposes a novel adaptive model for Web image semantic automatic annotation. First, the model automatically collects training image data by exploring the associated textual data and the social tagging data of Web images, such as the Flickr's Related Tags. Then, using a newly constrained piecewise penalty weighted regression to combine the adaptive estimation of the weight distribution of associated texts and the prior knowledge constrain together and implement the Web image semantic annotation. The proposed training data auto-generation methods and Web image annotation approaches are tested on a real-world Web image data set and promising results are achieved.

Key words: Web image annotation; training data auto-generation; social Web tagging; image retrieval

摘要: 提出了一种自适应的 Web 图像语义自动标注方法:首先利用 Web 标签资源自动获取训练数据;然后通过带约束的分段惩罚加权回归模型将关联文本权重分布自适应学习和先验知识约束有机地结合在一起,实现 Web 图像语义的自动标注.在 4 000 幅从 Web 获得的图像数据集上的实验结果验证了该文自动获取训练集方法以及 Web 图像语义标注方法的有效性.

关键词: Web 图像标注;训练集自动获取;社会 Web 标签;图像检索

中图法分类号: TP301 **文献标识码:** A

1 介绍

Web 图像的语义分析与标注在 Web 图像检索和 Web 多媒体数据的语义融合等方面具有重要意义,是实现基于语义的 Web 图像检索的关键技术,得到了学术界与产业界的广泛关注.图像标注是指使用语义关键词来表示一幅图像的语义内容.早期的人工标注需要专业人员根据每幅图像的语义标出关键词,费时且具有主观性.近年来,研究者提出了许多自动标注图像语义内容的方法^[1–21],其中根据训练数据进行有指导(/半指导)的图像语

^{*} Supported by the National Natural Science Foundation of China under Grant Nos.60403018, 60773077, 90818023 (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 2008-10-14; Revised 2008-12-03; Accepted 2009-01-15

义学习是研究的主流。

Web 图像通常关联着丰富的文本信息,如图像文件名、替代(ALT)文本、周边文本、所属页面的标题等等(如图 1 所示),图像的语义或多或少地都与这些关联文本相关.目前大部分商业图像搜索引擎(如 Google 图像搜索等)正是利用图像的关联文本信息,将图像检索转换为文本检索.因此,在 Web 图像语义自动标注过程中,除了图像的视觉特征,如何利用 Web 图像的关联文本来提高标注性能是近年来的一个热门研究话题,且取得了一些研究成果^[1-4].然而,多数已有方法或者把所有关联文本作为一个整体,或者仅仅根据先验知识或启发思想对各类关联文本赋予固定的权重.通过观察不难发现,不同的关联文本对预测图像语义的重要性是不同的,而且,随着图像和语义关键词的改变,图像语义与各类关联文本的相关性也呈现不同的分布.如图 2 所示,对于两幅不同的 Web 图像,同一个语义标注词“ocean”在各类关联文本上的分布是不同的;即使对于同一幅 Web 图像,如图 2 中的左边图像,不同的语义标注词(如“beach”和“ocean”)在各类关联文本上的分布也是不同的.因此,本文提出采用有指导的学习方法对 Web 图像语义分布进行自适应估计,以提高 Web 图像语义标注的性能.

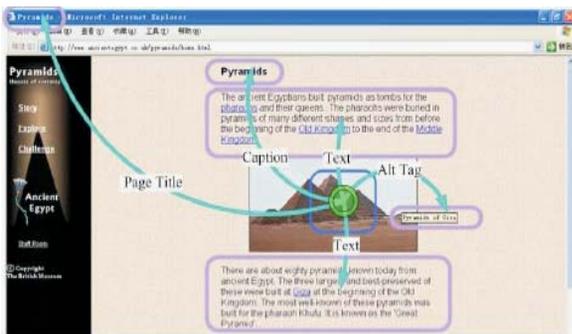


Fig.1 Relationships between Web image and associated texts

图 1 Web 图像与其关联文本之间的关系

Type	Keywords Distribution	Keywords Distribution
FileName	<i>beach20.jpg</i>	<i>siam_ocean.jpg</i>
ALT	<i>Virginia Beach</i>	<i>siam_ocean world</i>
Caption	<i>ocean front at dawn</i>	<i>siam_ocean world</i>
Text	<i>...Virginia Beach...</i>	<i>Siam ocean ...</i>
PageTitle	<i>Atlantic Ocean...</i>	<i>Paragon ... fishes ...</i>
	<i>Virginia Beach:</i>	<i>Preview of siam</i>
	<i>Virginia is for lover</i>	<i>ocean world</i>

Fig.2 Examples of Web images and its associated texts

图 2 Web 图像及其关联文本

训练数据是有指导的权重分布估计和图像语义标注的基础.已有的研究工作大多采用人工的方法来获取训练数据,费时且具有主观性,而且,对于层出不穷的 Web 图像来说,手工获取训练集几乎是不可能的,因此,训练集的自动获取具有重要意义.然而,由于 Web 图像具有数量巨大、种类繁多、质量参差不齐等特点,从 Web 上自动获取训练集的效果并不理想.近年来,随着 Web 社会知识网络(social knowledge network)的迅速发展,利用 Web 图像的人工 tagging 资源的各类应用引起了学术界的极大关注.本文通过对 Flickr 标签(tag)资源的利用,提出了一种基于 Web 图像语义概念空间学习的训练集自动获取方法.我们首先通过挖掘 Web 图像的关联文本自动获取训练图像及其初始类标签,然后将训练图像的初始类标签(关键词)提交给 Flickr^[22],获得其相关标签(related tags,简称 RT)**,构建 Web 图像语义概念空间,并在 Web 图像语义概念空间中考察图像语义之间的相似性,从而对训练图像的类标签进行扩展,极大地提高了训练集的质量.

然而,由于 Web 数据的多样性和复杂性,自动获取训练数据的数量仍然是有限的,单纯使用标准的回归模型难以满足 Web 图像语义分布估计的要求.关联文本权重分布的自适应学习是基于这样一个假设:我们能够找到和待标注图像具有相似权重分布的训练图像.然而,由于训练集的有限性、Web 数据的多样性等问题,我们找到的“相似”的训练图像并不总是能够有效地代表待标注图像,使得估计出的权重分布可能是不准确的,甚至是错误的.这种数据局限性引起的统计偏差,往往需要全局性的或额外的先验知识加以纠正.因此,为了提高关联文本权重估计的准确性,我们提出使用带约束的分段惩罚加权回归模型将关联文本权重分布估计和先验知识

** RT can be obtained by using Flickr's APIs: flickr.tags.getRelated. It returns a list of tags related to the given tag, based on clustered usage analysis --refer to: <http://www.flickr.net/services/api/flickr.tags.getRelated.html>

约束有机地结合起来,以先验知识来指导权重分布的估计过程.例如,一般情况下,根据网页设计的习惯和原则,ALT 文本对预测图像语义来说是比较重要的,然而,确实也存在不少训练图像缺乏 ALT 文本,因此,很可能出现根据训练图像估计出的权重分布中,ALT 文本对应的权重过小的情况.如果我们根据先验知识对权重分布估计进行一些约束,如 ALT 文本对应的权重不应该太小,则权重分布估计的过程将更加趋于合理性.

本文提出了一种自适应的 Web 图像语义自动标注方法.该方法综合考虑了 Web 图像的视觉特征和文本特征对预测图像语义的贡献.特别地,本文提出了一种新的带约束的分段惩罚加权回归模型,从而将关联文本权重分布估计和先验知识约束有机地结合起来,对 Web 图像语义在其关联文本上的分布自适应地进行建模,显著地提高了 Web 图像语义标注的性能.在 4 000 幅从 Web 获得的图像数据集上的实验结果验证了本文提出的训练集自动获取方法及 Web 图像语义标注方法的有效性.

本文工作主要贡献如下:

1) 提出了一种基于 Web 图像语义概念空间学习的训练集自动获取方法.该方法通过在 Web 图像语义概念空间中度量图像语义之间的相似性来扩展训练图像的类标签,有效地提高了训练集的质量.

2) 提出了一种自适应的 Web 图像语义自动标注方法.该算法综合考虑 Web 图像的视觉特征和文本特征对预测图像语义的贡献,并提出了带约束的分段惩罚加权回归模型,有效地将关联文本权重分布估计和先验知识约束结合在一起,自适应地学习图像语义(标注词)在 Web 图像关联文本上的分布,有效地提高了图像标注的性能.

3) 在真实的 Web 图像数据集上对本文提出的训练集自动获取方法和 Web 图像语义自动标注方法进行实验,实验结果验证了它们的有效性.

本文第 2 节给出相关工作.第 3 节详细介绍训练集的自动获取方法.第 4 节给出自适应的 Web 图像语义自动标注方法.第 5 节是实验结果及讨论.第 6 节是总结和展望.

2 相关工作

近年来,图像语义自动标注领域非常活跃,人们利用机器学习、统计模型等设计出各种不同的图像语义自动标注模型,主要可以分为两大类:基于概率模型的方法和基于分类的方法.基于概率模型的方法主要是从已标注训练图像集中直接或间接地学习图像(图像区域)视觉特征和语义概念(语义标注词)之间的联合概率分布,然后利用该概率分布对待标注图像进行语义标注,其代表性模型包括 Co-occurrence 模型^[5]、翻译模型^[6]、LDA 模型^[7]、CMRM^[8]、CRM^[11]、MBRM^[12]等;而基于分类的方法则将每一个语义概念(语义标注词)看作一个被混合模型刻画类,从而将图像语义自动标注转化为多类分类问题,语言索引方法^[9]、基于 SVM 的方法^[10]、多实例学习方法^[21]都属于这一类.然而,这些方法大都忽略了 Web 图像关联的丰富的文本信息,并不适合标注 Web 图像.而且,相对于 Web 图像标注中可能存在的几乎是无限的候选关键词集合,大多数已知的标注方法仅能对其很少的一部分候选语义关键词进行建模.因此,这些标注方法大都不能直接用于 Web 图像语义标注任务.

Web 图像通常伴随着丰富的文本信息,其语义内容或多或少地都与这些关联文本信息有关,因此,利用 Web 图像的关联文本揭示其语义内容是 Web 图像语义自动标注的一种重要手段.在商业领域,许多图像搜索引擎,如 Google 和 Yahoo 图像搜索,正是利用 Web 图像的关联文本将图像搜索转换为文本搜索的.在学术领域,Sanderson 等人^[19]较早地利用 Web 图像关联的文本信息对图像语义内容进行建模,它不考虑关联文本的结构,将所有文本看作一个词集;Shen 等人^[20]开始考虑使用 Web 图像关联的更详细的文本信息,如图像文件名、替代(ALT)文本、周边文本、页面的标题等,并建立了一个简单的模型将这些相关的文本内容组合起来;Wang 等人^[2]提出了一个基于搜索和数据挖掘技术的 Web 图像语义自动标注系统——AnnoSearch,它首先根据用户提供的语义关键词,利用基于文本的图像搜索技术获得和待标注图像语义相似的图像集,然后在该图像集上执行基于内容的图像检索(CBIR),从而获得与待标注图像语义和视觉上都相似的图像集,最后从这些图像的关联文本中挖掘待标注图像的语义标注词,但该系统要求至少一个准确的初始关键词作为种子来执行基于文本的图像搜索,如果没有初始关键词,或初始关键词是不正确的,标注的性能将极大地降低.Li 等人^[1]提出了一种搜索和挖掘相结合

(searching and mining)的 Web 图像语义自动标注方法,首先利用基于内容的图像检索(CBIR)方法得到一个与待标注图像视觉相似的 Web 图像集合,然后利用搜索结果聚类技术,从这些 Web 图像的关联文本中提取最有代表性的关键词作为待标注图像的语义标注词,但其标注性能受图像搜索阶段的影响很大,而基于内容的图像搜索(CBIR)的准确度一般比较低,因此,该方法的整体标注性能仍有待提高.Feng 等人^[3]分别基于 Web 图像的视觉特征和文本特征建立两个分类器,并且假定它们是正交的,然后利用 Co-training 的方法来学习 Web 图像的语义标注词.Viecent 等人^[4]分别基于 Web 图像的视觉特征和文本特征建立两个模型,然后将两个模型加权起来得到最终的语义标注模型,然而,我们很难确定两个模型各自所占的比重.而且,文献[4]中利用决策树模型对 Web 图像语义在其关联文本上的分布进行建模,训练完成后,决策树模型将不再改变,但训练出的固定决策树模型并不能够准确地表示所有待标注图像的语义分布,从而影响了 Web 图像语义标注的性能.总之,多数已有的利用 Web 图像关联文本进行标注的方法大都或者把所有关联文本作为一个整体,或者仅仅根据先验知识或启发想法对各类关联文本赋予固定的权重.然而,不同的关联文本对预测图像语义的重要性是不同的,而且,随着图像和语义关键词的改变,图像语义与各类关联文本的相关性也呈现不同的分布.因此,在我们先前的工作中,已提出了一种基于关联文本权重分布自适应学习的 Web 图像语义自动标注方法.然而,通过对标注结果的分析,我们发现由于训练集的有限性、Web 数据的多样性等因素,估计出的权重分布有可能是不准确的,甚至是错误的.为了纠正这种错误,我们需要利用先验知识来指导自适应估计,就我们所了解,还没有工作将先验知识和关联文本权重分布的自适应学习有效地进行结合.本文提出采用有指导的学习方法来自适应地学习每一幅图像和每一个候选语义标注词对应的 Web 图像语义分布,并且使用先验知识来指导语义分布的自适应估计过程,有效地将先验知识和图像对应的特定语义分布的学习统一起来,从而极大地提高了 Web 图像语义标注的性能.

3 训练集的自动获取

作为有指导的 Web 图像语义自动标注的基础,自动获取高质量的 Web 训练图像是至关重要的.本节首先给出一个启发式的 Web 图像训练集的自动获取方法(training generation,简称 Trg);然后,在 Trg 方法的基础上,提出了两种通过 Web 图像语义概念空间学习来提高训练集质量的方法.

3.1 训练集的自动获取方法

训练集自动获取方法的基本思想类似于 TF/IDF 的思想^[22].在估计关键词 w 作为图像语义标注词的概率时,我们考虑两种词频: w 在某类关联文本(如图像名)中出现的频率和 w 在不同类型的关联文本中的出现频率(即包含 w 的关联文本的类个数).而我们利用的启发式思想是词频越高的关键词就越有可能作为对应图像的类标签.

记图像 I 的第 $i(i=1, \dots, m, m$ 是关联文本的类个数)类关联文本为 T_i , 出现在图像 I 的关联文本中的所有关键词的集合为 $WS(I)$, 则对任意 $w \in WS(I)$, w 作为图像 I 的类标签的置信度定义如下:

$$Conf(w, I) = \frac{df(w)}{m} \times \sum_{i=1}^m \alpha_i \times \frac{tf(w, T_i)}{|T_i|} \quad (1)$$

其中, $df(w)$ 表示关键词 w 出现的关联文本的类个数, $tf(w, T_i)$ 表示关键词 w 在关联文本 T_i 中的词频; $|T_i|$ 表示关联文本 T_i 中出现的关键词的总个数; $\alpha_i (\sum \alpha_i = 1)$ 表示关联文本 T_i 的权重.

给定置信度阈值 η , 图像 I 的初始类标签集合定义如下:

$$BA(I) = \{w | w \in WS(I) \& Conf(w, I) \geq \eta\} \quad (2)$$

记图像数据集为 L , 则训练集 L_{train} 和测试集 L_{test} 定义如下:

$$L_{train} = \{I | I \in L \& BA(I) \neq null\} \quad (3)$$

$$L_{test} = L \setminus L_{train} \quad (4)$$

3.2 基于 Web 图像语义概念空间学习的类标签扩展

类似于传统的 Web 图像分析模型,第 3.1 节给出的训练集自动获取方法是一类单向的方法,即图像的类标签完全是从其关联文本中提取出来的.然而,这类方法得到的训练图像的类标签是不完全的.为了进一步提高自

动获取的训练集质量,本节利用 Web 图像人工 tagging 资源来构建 Web 图像语义概念空间,并通过在 Web 图像语义概念空间中考察图像语义概念之间的相似性来扩展训练图像的类标签,从而提高训练集的质量.

3.2.1 Web 图像语义概念空间的构建

公共的 Web 社会知识网络 Flickr 中存在着大量的标签(关键词),通过 Flickr 提供的 API 接口我们可以得到这些标签(关键词)在表达图像语义时的关系,这种关系不同于从各种语法词典(如 WordNet)中得到的关键词之间的关系,它们真实地反映了这些关键词在图像语义空间中的关系,因此,我们可以将 Web 社会知识网络 Flickr 看作一个 Web 图像语义概念空间(ICS),并通过在 ICS 中考察图像语义之间的关系来扩展训练图像的类标签,提高训练集 L_{train} 的质量.

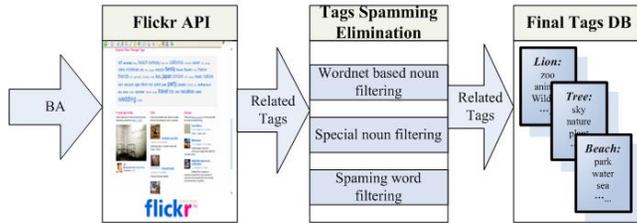


Fig.3 Construction of Web images semantic concept space

图 3 Web 图像语义概念空间的构建

图 3 给出了 Web 图像语义概念空间的构建过程.我们首先将训练图像的初始类标签集合(BA)中的每一个关键词 w 提交给 Flickr,从而获得 w 在 ICS 中的图像语义概念邻域(ICN),即 w 在 Flickr 中的相关标签(related tags).需要强调的是,关键词 w 的图像语义概念邻域(ICN)是由那些经常和 w 一起来表示同一幅图像语义的概念(关键词)组成的.

垃圾标签(Spamming)是考察 Web 资源时必须考虑的问题,即图像语义概念空间中包含许多噪音数据,如一些不合法的单词等,这些噪音数据不应该存在于图像的语义空间中,因此,我们需要对图像语义概念空间进行过滤,以构建更合理的 Web 图像语义概念空间,从而更好地度量图像语义之间的关系.如图 3 所示,本文对图像语义概念空间进行了 3 种过滤:(1) 名词提取;(2) 专有名词过滤,如 UK 等;(3) “不合法”词语过滤,如包含数字的词语(如 a123 等)和长度小于等于 2 的词(如 bw 等),这些词语没有明确的含义,不可能作为图像的语义标注.注意,在一些专门识别图像中人物的名字、事件发生的时间、地点等工作中,专有名词是有用的,但在我们的工作中,专有名词作为噪音而被过滤掉了.

3.2.2 基于公共图像概念邻域类标签扩展(common image concept neighbor,简称 CICN)

记关键词 w 在 Web 图像语义概念空间中的图像语义概念邻域为 $ICN(w)$,则图像 I 在 ICS 中的公共概念邻域定义如下:

$$CICN(I) = \{w | w \in \bigcap_{w_i \in BA(I)} (ICN(w_i))\} \tag{5}$$

由于 $ICN(w)$ 是由那些经常和 w 一起表示同一幅图像语义的概念(关键词)组成的,即 $ICN(w)$ 中的关键词在 Web 图像语义概念空间距离 w 都很近,则由公共概念邻域的定义可以看出, $CICN(I)$ 中的概念(关键词)和图像 I 已知的标注都很相似,很有可能较准确地表示图像 I 的语义,因此,我们可以利用 $CICN(I)$ 来扩展图像 I 的类标签.经过公共图像概念邻域扩展后,图像 I 的类标签集合定义如下:

$$An(I) = BA(I) \cup CICN(I) \tag{6}$$

3.2.3 基于图像语义概念关系图的类标签扩展(image concept relation graph,简称 ICRG)

尽管基于公共图像概念邻域的类标签扩展方法可以提高自动获取的训练集质量,但仍然存在如下缺点:(1) 当 $BA(I)$ 中的关键词数量较大时,图像 I 的公共图像概念邻域中的概念很少,甚至为空,从而起不到扩展的作用;(2) 没有考虑概念之间的隐含关系,即如果 $w_1 \in ICN(w_2)$, $w \in ICN(w_1)$, $w \notin ICN(w_2)$ 成立,则 w 和 w_2 之间也存在着某种语义关系;(3) 忽略了 ICS 中概念之间存在的不同关系.因此,为了进一步提高训练集的质量,我们又提

出了基于图像语义概念关系图的类标签扩展方法.

3.2.3.1 图像语义概念关系图的构建

图像语义概念关系图是一个有向加权图 (V, E) ,其中顶点集合 V 是 ICS 中所有概念的集合,顶点 w 到 w' 存在一条有向边 $e_{ww'} \in E$ 当且仅当 $w' \in ICN(w)$.边 $e_{ww'}$ 具有权重 $p(w'|w)$,即当已知 w 是图像的语义标注词时, w' 作为图像的语义标注词的概率.为了进一步考虑概念之间的隐含关系,我们定义关键词 w 在 Web 图像语义概念空间中的扩展图像语义概念邻域为 $EICN(w)$ 如下:

$$EICN(w) = \{w' | w' \in ICRG \ \& \ ICRG \text{中存在} w \text{到} w' \text{的一条路径}\} \quad (7)$$

根据图像概念(关键词) w 和 w' 之间存在的不同关系,权重 $p(w'|w)$ 分别计算如下:

- 共现关系:如果 $w' \in EICN(w)$ 且 $w \in EICN(w')$,则称图像概念 w 和 w' 之间存在共现关系,即当已知 w 是图像的语义标注词时, w' 也是图像的语义标注词,反之亦然.因此有:

$$p(w'|w) = p(w|w') = 1 \quad (8)$$

- 导出关系:如果 $w' \in EICN(w)$ 但 $w \notin EICN(w')$,则称图像概念 w 和 w' 之间存在导出关系,即当已知 w 是图像的语义标注词时, w' 也很有可能是图像的语义标注词.此时, $p(w'|w)$ 定义如下:

$$p(w'|w) = e^{-dis(w,w') \times \frac{|EICN(w)|}{|ICS|}} \quad (9)$$

其中, $dis(w,w')$ 是在图 $ICRG$ 中顶点 $w \sim w'$ 之间的最短路径长度,可以使用Dijkstra算法^[23]来计算. $|EICN(w)|$ 和 $|ICS|$ 分别是概念 w 的扩展图像概念邻域中概念的个数和整个图像语义概念语义空间中概念的个数.

- 耦合关系:如果 $w' \notin EICN(w)$ 且 $w \notin EICN(w')$,则称图像概念 w 和 w' 之间存在耦合关系.该类关系主要起到一种平滑作用,即当 $w' \notin EICN(w)$ 且 $w \notin EICN(w')$ 成立时,不表示 w 和 w' 之间没有关系,而是以某种较弱的耦合关系存在.此时, $p(w'|w)$ 定义如下:

$$p(w'|w) = p(w|w') = e^{-\max_dis \times \frac{|EICN(w) \cup EICN(w')|}{|ICS|} \times \frac{|EICN(w) \cap EICN(w')|}{|EICN(w) \cap EICN(w')|}} \quad (10)$$

其中, $\max_dis = \max_{w', w'' \in ICS} \{dis(w', w'')\}$.公式(10)表明,如果概念 w 和 w' 的扩展图像语义概念邻域重合越大,则它们之间的耦合性就越强.

3.2.3.2 基于图像语义概念关系图的类标签扩展

构建了图像语义概念关系图,我们就可以根据 $ICRG$ 来度量语义概念 w 和图像 I 之间的语义相关性.语义概念 w 和图像 I 之间的语义相关性 $relation(w, I)$ 定义如下:

$$relation(w, I) = relation(w, BA(I)) = \frac{1}{|BA(I)|} \sum_{w' \in BA(I)} p(w|w') \quad (11)$$

给定语义相关性阈值 β ,则图像 I 的扩展类标签集合 $EA(I)$ 定义如下:

$$EA(I) = \{w | w \in ICS \ \& \ relation(w, I) \geq \beta\} \quad (12)$$

则经过图像语义概念关系图扩展后图像 I 的类标签集合定义如下:

$$An(I) = BA(I) \cup EA(I) \quad (13)$$

4 自适应的Web图像语义标注

获取训练集后,我们就可以使用有指导的学习方法来自动标注 Web 图像.本节将详细介绍自适应的 Web 图像语义自动标注方法.特别地,在考虑图像文本特征对预测图像语义的贡献时,我们利用带约束的分段惩罚加权回归模型来自适应地学习图像语义在其关联文本及其结构关系上的分布.

4.1 基本框架

L_{train} 中的图像 J_i 由语义标注词、视觉特征和关联文本特征表示,即 $J_i = \{W_i, V_i, T_i\}$,其中 $W_i = \{w_{i1}, \dots, w_{il}\}$ 含有 l 个语义关键词, w_{ij} 是一个二元变量,表示第 j 个语义关键词是否为第 i 幅图像的语义标注词; $V_i = \{f_{i1}, \dots, f_{im}\}$ 表

示图像 m 个子区域的视觉特征 $f_{i,j}$ 的集合; $T_i = \{T_{i1}, \dots, T_{im}\}$ 表示图像的 n 类关联文本特征.

对于待标注 Web 图像 $I = \{V, T\}$, 不失一般性, 我们可以假设图像 I 的视觉特征 V 和文本特征 T 相互独立. 则关键词 w 作为图像 I 的语义标注的概率 $P(w|I)$ 定义如下:

$$P(w|I) = P(w|I_V, I_T) = \frac{P(w|I_V)P(w|I_T)}{P(w)}, \quad (14)$$

其中, I_V, I_T 分别表示图像 I 的视觉特征和关联文本特征. 假设 $P(w)$ 满足均匀分布, 则

$$w^* = \arg \max_w \{P(w|I_V)P(w|I_T)\} \quad (15)$$

公式(15)表明, 为了标注图像 I , 我们需要估计两个生成概率: 基于视觉特征的生成概率 $P(w|I_V)$ 和基于文本特征的生成概率 $P(w|I_T)$.

4.2 基于文本特征的生成概率估计

4.2.1 目标函数

令 $H(T)$ 表示扩展函数的集合, 即关联文本 T 及其结构关系, $\omega = \{\omega_1, \dots, \omega_N\}$ 表示 $H(T)$ 对预测图像 I 的语义的重要性(文本权重), N 是图像文本特征的维数, $X_j = p(w|h_j(T))$ 表示 $h_j(T) \in H(T)$ 对语义概念 w 的语义贡献. 记候选语义标注词 w_i 对应的文本权重和语义贡献分别为 $\omega(w_i)$ 和 X_i , 则基于文本特征的生成函数 $P(w|I_T)$ 定义如下:

$$P(w_i | I_T) \propto P(w_i, I_T) = \sum_{j=1}^N \omega_j(w_i) X_{ij} \quad (16)$$

4.2.2 基扩展和生成概率估计

来自于不同结构的关联文本对预测图像语义的贡献可能非常复杂, 因此, 我们有必要考虑这些关联文本之间的高阶结构关系对预测图像语义的贡献. 对于图像 I 的关联文本 $T = \{T_1, \dots, T_n\}$, 其 k 阶结构关系记为 $ST^k = \{(T_1 T_2 \dots T_k), \dots, (T_{n-k+1} \dots T_n)\}$. 2 阶结构关系是最简单的高阶关系, 即 $ST^2 = \{(T_1 T_2), (T_1 T_3), \dots, (T_{n-1} T_n)\}$.

基扩展方法可以用来向线性模型中加入一些非线性的因素, 本文利用基扩展方法来考虑关联文本的高阶结构关系的影响. 定义扩展函数集合 $H(T)$ 来统一地表示图像的关联文本及其高阶结构关系, 为简单起见, 本文仅考虑关联文本 T 及其 2 阶关系对图像语义的贡献. 则转换函数 $h_i(T) \in H(T)$ 定义如下:

$$h_j(T) = \begin{cases} T_j, & j = 1, \dots, n \\ T_i T_l, & (i \neq l) \leq n, n < j \leq N, T_i T_l = T_i T_j \end{cases} \quad (17)$$

其中, $N = n(n+1)/2$ 是扩展后图像文本特征的维数.

不失一般性, 可以假设关键词 w 出现在文本 T_i 和 $T_l (i \neq j)$ 中是相互独立的, 则关键词 w 从 $h_i(T)$ 中生成的概率 $p(w|h_i(T))$ 可以估计如下:

$$\hat{p}(w|h_j(T)) = \begin{cases} p(w|T_j), & j = 1, \dots, n \\ p(w|T_i)p(w|T_l), & i \neq l \leq n, n < j \leq N \end{cases} \quad (18)$$

其中, $p(w|T_i)$ 表示关键词 w 从文本 T_i 中生成的概率, 可以利用极大似然估计来估计.

4.2.3 文本权重的带约束分段惩罚加权回归估计

根据公式(16)可知, 估计生成概率 $p(w|I_T)$ 的关键在于自适应地学习文本权重 $\omega(w)$. 传统的图像标注工作大都假设视觉特征相似的图像语义也很相似, 然而受语义鸿沟的影响, 标注的效果一直不理想. 而目前大多数商业图像搜索引擎, 如 Google image, Alta Vista 和 Yahoo image 等, 则假设处于相似的环境中的 Web 图像(这些图像具有相似的关联文本)具有相似的语义, 并取得了可以接受的查询结果. 事实上, 视觉特征和关联文本只是从不同的方面反映 Web 图像的语义, 我们有理由相信视觉特征和关联文本都相似的 Web 图像更可能具有相似的语义. 因此, 本文假设图像语义在其关联文本上的分布符合一定的统计规律, 即视觉特征和关联文本相似的图像, 其语义在关联文本上的分布也具有一定的相似性.

本文使用有指导的方法来估计关联文本 $H(T)$ 对 Web 图像语义推断的重要性(文本权重), 基本思想是首先在训练图像集中找出待标注图像 I 的相似邻域 $neighbor(I)$, $neighbor(I)$ 中每一幅图像对应的文本权重分布都在

一定程度上逼近图像 I 的真实分布,因此,我们可以使用回归的方法,通过拟合 $neighbor(I)$ 中图像对应的文本权重分布来预测图像 I 的真实分布.由于 $neighbor(I)$ 中图像对应的文本权重分布逼近图像 I 的真实分布的程度有一定的差异,我们需要为 $neighbor(I)$ 中的图像赋予不同的权重;同时,为了减少回归模型的预测误差,我们利用 L_2 正则化方法对回归系数进行惩罚,特别地,为了区分关联文本和其高阶结构关系对图像语义的贡献,我们对回归系数的不同子集进行不同程度的惩罚,因此,本文提出使用带约束的分段惩罚加权回归来估计图像 I 的关联文本权重分布,具体如下:

作为确定图像 I 的相似邻域 $neighbor(I)$ 的基础,本文使用生成概率估计的方法来度量图像之间的相似性.假设 Web 图像的视觉特征和文本特征相互独立,则图像 I 由 J 生成的概率定义如下:

$$P(I|J) = P(I_V, I_T | J) = P(I_V | J)P(I_T | J) = P(I_V | J_V)P(I_T | J_T) \quad (19)$$

其中, $P(I_V | J_V)$ 表示图像 I 在视觉上由图像 J 生成的概率,可以使用区域生成概率的乘积进行估计,而区域生成概率则使用非参数高斯核估计^[12]进行估计; $P(I_T | J_T)$ 表示图像 I 的文本特征 T_i 从图像 J 的文本特征 T_j 中生成的概率,可以使用最大似然估计方法进行估计.

给定待标注图像 I 和候选语义标注词 w ,记 $neighbor(I)$ 中标注词包含 w 的图像集合为 $neighbor(I, w)$.本文以 $neighbor(I, w)$ 作为训练集来估计对应于图像 I 的位置权重分布 $\omega(w)$. $neighbor(I, w)$ 中图像 J_i 对估计 I 的语义贡献是不同的,记图像 J_i 对应的权重为 μ_i ,即图像 J_i 和 I 的相似度.令 D_s 表示对应于 s 阶结构关系的系数集合, D_{pre} 是已知的比较重要的系数集合,则带约束的分段惩罚加权回归估计定义如下:

$$\hat{\omega}(w) = \arg \min_{\omega(w)} \left\{ \sum_{i=1}^K \mu_i (y_i - \omega_0(w) - \sum_{j=1}^N X_{ij} \omega_j(w))^2 \right\} \quad (20)$$

$$\text{受限于: } \sum_{j \in D_s} \omega_j(w)^2 \leq t_s (s=1, \dots, k), \sum_{j \in D_{pre}} \omega_j(w)^2 \geq t_{pre}$$

其中, $X_{ij} = p(w | h_j(T))$ 是回归模型的输入值,表示关键词 w 从文本特征 $H(T)$ 中生成的概率; y_i 是回归模型的预测值,表示候选语义标注词作为图像语义标注的概率,由于已知 w 是图像 J_i 的语义标注词,因此预测值 $y_i = p(w | J_i) = 1$.

式(21)等价于如下的带约束的分段惩罚加权残差平方和:

$$\hat{\omega}(w) = \arg \min_{\omega(w)} \left\{ \sum_{i=1}^K \mu_i \left(y_i - \omega_0(w) - \sum_{j=1}^N X_{ij} \omega_j(w) \right)^2 + \sum_{s=1}^k \gamma_s \sum_{j \in D_s} \omega_j(w)^2 - \sum_{j \in D_{pre}} \gamma_{pre} \omega_j(w)^2 \right\} \quad (21)$$

其中, $\gamma = \{\gamma_1, \dots, \gamma_k\}$, $\gamma_1 \geq \dots \geq \gamma_k$ 和 γ_{pre} 是惩罚因子.

我们通过 $\bar{y} = \sum_{i=1}^K y_i / K$ 来估计 $\omega_0(w)$,其余的系数使用中心化的 X_{ij} ,通过无截距的加权回归估计得到.假设已进行中心化,则式(23)可以写成矩阵的形式:

$$RSS(\lambda) = \mu(y - X\omega(w))^T (y - X\omega(w)) + \sum_{s=1}^k \gamma_s \omega'_s(w)^T \omega'_s(w) - \gamma_{pre} \omega'_{pre}(w)^T \omega'_{pre}(w) \quad (22)$$

其中,当 $\omega_j \in D_s$ 时 $\omega'_s(w) = \omega_j$,其他情况下 $\omega'_s(w) = 0$,同样,当 $\omega_j \in D_{pre}$ 时, $\omega'_{pre}(w) = \omega_j$,其他情况下 $\omega'_{pre}(w) = 0$,则回归模型的解为

$$\hat{\omega} = \left(\mu X^T X + \sum_{s=1}^k \gamma_s I_s - \gamma_{pre} I_{pre} \right)^{-1} X^T y \quad (23)$$

其中,矩阵 I_s 和 I_{pre} 均为 $N \times N$ 的矩阵,除了对角线上的元素,其他元素均为 0,当 $\omega_j \in D_s$ 时, $I_s(j, j) = 1$,否则为 0,同样,当 $\omega_j \in D_{pre}$ 时, $I_{pre}(j, j) = 1$,否则为 0.

4.3 基于视觉特征的生成概率估计

给定一幅待标注图像 I 和关键词 w , w 从视觉特征上作为图像 I 的语义标注词的概率 $P(w | I_V)$ 定义如下^[24]:

$$P(w | I_V) \propto P(w, I_V) = \sum_{i=1}^{|I|} P(w, I_V | J_i) P(J_i) = \sum_{i=1}^{|I|} P(w | J_i) P_V(I | J_i) P(J_i) \quad (24)$$

其中, $P_V(I|J_i)$ 表示仅仅考虑图像的视觉特征时, 图像 I 从图像 J_i 中生成的概率, 假设 $P(J_i)$ 服从均匀分布.

假设图像分割后各区域相互独立, 则 $P_V(I|J_i)$ 等于各区域生成概率的乘积, 区域 f_i 由图像 J_i 生成的概率 $P_V(f_i|J)$ 使用核密度非参数估计^[3], 如下:

$$P_V(f_j | J) = \frac{1}{m} \sum_{k=1, g_k \in J}^m \frac{\exp\{-(g_{ik} - f_j)^T \Sigma^{-1} (g_{ik} - f_j)\}}{\sqrt{2^D \pi^D |\Sigma|}} \quad (25)$$

其中, g_{ik} 表示训练图像 J_i 的第 k 个区域的视觉特征, m 是 J_i 中区域个数.

对词的估计 $P(w|J)$, 我们使用如下二重平滑:

$$P(w | J) = \lambda P_M(w | J) + (1 - \lambda) P_S(w | J) \quad (26)$$

其中, $P_M(w|J)$ 是指对 $P(w|J)$ 的极大似然估计, $P_S(w|J)$ 是把整个训练集作为背景得到的概率估计值, λ 是平滑因子.

4.4 标注算法

本文提出自适应的 Web 图像标注方法的标注过程见算法 1.

算法 1. 自适应的 Web 图像标注算法.

输入: 训练集合 L , 待标注图像 I .

输出: 待标注图像 I 的语义标注.

算法过程:

1. 计算 I 从 L 中每幅图像生成的概率 $P(I|J), J \in L$.
2. 选择前 K 个最高生成概率训练图像, 得到 I 的 $neighbor(I)$.
3. 根据公式(16), 计算语义关键词 w 由图像 I 的文本特征生成的概率 $P(w|I_T)$.
4. 根据公式(25), 计算语义关键词 w 由图像 I 的视觉特征生成的概率 $P(w|I_V)$.
5. 根据公式(15), 计算语义关键词 w 作为图像 I 的语义标注的概率, 选择概率最大的前 k 个语义关键词作为图像 I 的语义标注.

时间复杂度分析: 令 n 表示训练集的个数, 则算法 1 所列前两步时间复杂度均为 $O(n)$. 第 3 步的时间复杂度由对邻域中图像的生成概率估计的时间复杂度+分段惩罚加权回归估计的时间复杂度组成. 令 t 表示文本特征类型的个数, 邻域中图像个数为 K , 则对邻域中图像的生成概率估计的时间复杂度为 $O(t^2K)$; 岭回归的时间复杂度为 $O(K)$, 因此第 3 步的时间复杂度为 $O(t^2K) + O(K)$. 假设图像被划分为 m 块, 则第 4 部分的时间复杂度为 $O(m^2K)$. 假设 p 为单词表的个数, 则算法第 5 步的时间复杂度为 $O(p)$. 由于 K, t, m 和 p 相对 n 非常小, 且这些参数值是不随着数据集大小而变化的, 因此, 算法 1 的复杂度主要由 $O(n)$ 主导.

5 实验

5.1 实验数据及设置

我们通过向 yahoo 搜索引擎提交查询关键词得到一个 Web 页面集 P , 然后使用一个“轻量级”的页面解析程序包 HTMLParser^[25] 将集合 P 中的页面转换成 DOMTree, 最后通过对 DOMTree 的遍历获得图像集 L 及其关联文本. 本文实验中提交的查询词有 *beach, bear, birds, bridge, building, car, Egypt pyramid, flower, great wall, tree, tiger, whale* 等. 图像集 L 共包含 3 550 图像, 我们为每一幅图像生成 200×200 的缩略图, 然后利用基于固定大小网格的方法将每一个缩略图分割为 36 块, 每一块根据 MPEG7 标准提取 528 维的特征向量, 共同组成图像的视觉特征; 图像的关联文本特征共包含来自于 HTML 页面结构中 5 个不同位置的文本, 分别是图像文件名文本、图像的替代(ALT)文本、标题文本、周边文本、页面标题文本, 记为 T_1, \dots, T_5 . 根据第 3 节介绍的训练数据集自动生成方法, 最终得到 520 幅训练图像, 其余的为测试图像. 人工为每一幅测试图像赋予 1~7 个标注词, 最终得到 67 个标注关键词.

在训练集自动生成过程中, 我们考虑除周边文本以外的其他 4 类关联文本, 4 类文本的权重 $\alpha_i = 0.25 (i=1, \dots, 4)$, 置信度阈值 $\eta = 0.2$, 语义相关性阈值 $\beta = 0.8$. 我们将得到的训练集平均分成训练、验证两部分以

确定标注模型的一些参数,如平滑参数 λ ,惩罚参数 λ_1 、 λ_2 和惩罚因子 λ_{pre} .在本文实验中,通过验证,平滑参数 λ 取0.6,惩罚参数 $\gamma_1=0.75$, $\gamma_2=0.3$,惩罚因子 $\gamma_{pre}=0.15$.在实验中,我们使用的先验知识是ALT文本及其相关高阶关系比较重要.

我们采用 *precision*, *recall*, *F1* 这3个度量值来验证本文标注算法的有效性,定义如下:

$$precision = \frac{|correct|}{|predicted|} \times 100\%, recall = \frac{|correct|}{|ground - truth|} \times 100\%, F1 = \frac{2 \times precision \times recall}{precision + recall},$$

其中, $|correct|$ 表示预测结果中正确的标注词个数, $|predicted|$ 表示预测结果中标注词的个数, $|ground - truth|$ 表示所有正确的标注词个数. *Recall* 度量对单个词查询的完整性, *precision* 度量查询的精度, 平均的查准和查全率则反映标注整体的性能. 我们使用相同的训练集对标注模型进行训练, 并在相同的测试集上进行测试. 固定标注长度设为5.

5.2 实验结果

5.2.1 本文标注算法的整体性能

图4给出了本文提出的自适应Web图像语义自动标注方法(ModelAdap)和 ModelFMD、ModelNTC方法的标注结果比较,其中 ModelFMD^[4]在训练阶段提前学习出一个固定的语义分布模型,然后利用该模型标注所有的测试图像,在本组实验中,ModelFMD模型中视觉模型和文本模型的权重分别取0.3和0.7. ModelNTC则是将各种类型的关联文本看作一个整体的Web图像标注方法.

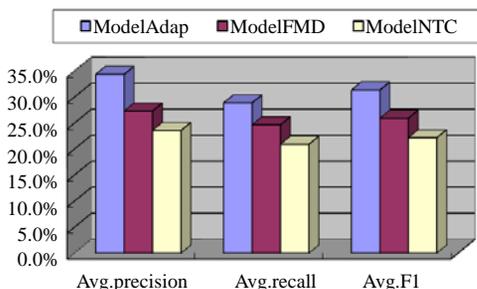


Fig.4 Annotation performance comparison between our algorithm ModelAdap, ModelFMD and ModelNTC

图4 标注算法 ModelAdap, ModelFMD 和 ModelNTC 的标注性能比较

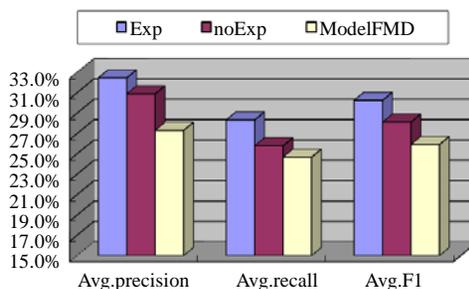


Fig.5 Effectiveness of basic expansion

图5 基扩展的有效性检验

由图4我们可以看出,与 ModelFMD、ModelNTC相比,本文提出的标注方法在 *recall*, *precision* 和 *F1* 上均有所提高,其中 *F1* 分别从 22.2%和 25.9%提高到了 31.4%.这说明与将所有关联文本作为一个整体,或者仅仅根据先验知识对各类关联文本赋予固定的权重标注方法相比,本文自适应地为每一幅图像和每一个标注词学习出特定的权重分布的做法更加合理,从而可以有效地提高标注性能.由于在本文提出的标注方法,一是对考察基本类型的关联文本之间高阶关系,另一个是带约束的分段惩罚加权回归,因此需分别检查这两个方面的有效性,并分析这两个方面对改进标注性能所起作用.

5.2.2 基扩展的有效性验证

图5给出了考察基本类型关联文本之间高阶关系的实验结果,在使用标准的岭回归来学习权重分布的前提下,将本文利用基扩展考察基本类型关联文本之间高阶关系对图像语义的贡献的标注方法(Exp)分别与本文标注框架下仅仅考虑基本类型关联文本的标注方法(noExp)和仅仅考虑基本类型关联文本的 ModelFMD^[4]模型进行比较.其中,ModelFMD的参数同上一个实验.从图5可以看出,在都不考虑关联文本之间的高阶关系的情况下,本文标注模型(noExp)的标注性能要优于 ModelFMD, *F1* 从 25.9%提高到了 28.1%,这是因为本文的模型自适应地为每一幅图像学习关联文本的权重分布,而 ModelFMD 则是在训练阶段提前训练一个固定的语义分布模型.而通过基扩展方法加入关联文本之间的高阶结构关系的贡献时,本文标注模型的标注性能得到了有效地提

高, F1 从 28.1%提高到了 30.2%,这是因为通过关联文本之间的高阶关系,我们可以更准确地预测标注词和图像之间的相关性,从而提高图像标注的准确性.

5.2.3 带约束的分段惩罚加权回归的有效性验证

图 6 给出了利用带约束的分段惩罚加权回归模型来自适应地学习图像的语义在关联文本及其结构上分布的实验结果.本实验主要验证带约束的分段惩罚加权回归在标注 Web 图像过程中的有效性.实验中主要考虑两个贡献点:(1) 分段惩罚加权回归(PPWR);(2) 回归过程中的先验知识约束(Ctr).

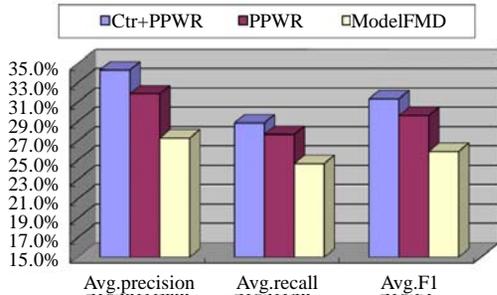


Fig.6 Effectiveness of constrained piecewise penalty weighted regression

图 6 受约束的分段惩罚加权回归的有效性检验

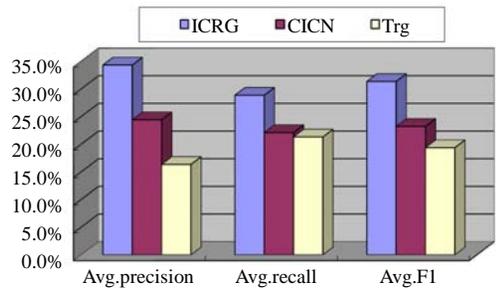


Fig.7 Effectiveness of Flickr based label expansion in auto-generation of training set

图 7 基于 Flickr 的类标签扩展在自动获取训练集中的有效性验证

从图 6 可以看出,先验知识约束和分段惩罚加权回归对标注都起到了积极作用,F1 从 25.9%分别提高到了 29.6%和 31.4%.这说明在 Web 图像语义标注过程中,关联文本和其高阶结构信息对于预测图像的语义是具有不同贡献的,对其分别进行相应的惩罚的做法更适合于估计文本权重分布.同时,本组实验也验证了通过先验知识的指导,我们可以得到更加合理的文本权重分布,从而提高图像标注的性能.

5.2.4 Flickr 标签资源在训练集自动获取中的有效性验证

本文基于 Flickr 标签资源来构建 Web 图像语义概念空间,并通过 Web 图像语义概念空间学习来提高训练集的质量.为了验证 Flickr 标签资源在训练集自动获取过程中的有效性,本组实验利用基于 Flickr 类标签扩展的训练集自动获取方法(C1CN 和 ICRG)和基准方法(Trg)获得 3 个训练集,并比较了 WAIA_AdP 标注方法在 3 个训练集上的标注性能.图 7 给出了标注性能比较结果.

由图 7 我们可以得到如下结论:(1) 标注算法在训练集 C1CN 和 ICRG 上的标注性能都要远远好于在 Trg 上的性能,F1 从 19.4%分别提高到了 23.2%和 31.4%,这说明我们可以通过 Web 图像语义空间学习来提高自动获取的训练集的质量,进而提高 Web 图像语义自动标注的性能.通过对训练图像的观察,我们发现这是因为仅仅通过考察 Web 图像的关联文本获得的训练图像的标注是不完全的,即仅得到了图像很少一部分语义标注,而通过 Web 图像语义空间学习可以为图像赋予更多正确的语义标注词.(2) 标注算法在训练集 ICRG 上的标注性能要远远好于在 C1CN 上的性能,F1 从 23.2%提高到了 31.4%,说明相对于公共图像概念邻域,图像语义概念关系图更适合于训练图像的分类标签扩展.这是因为公共图像概念邻域中的概念一般较少,特别是当训练图像的初始类标签较多时,公共图像概念邻域可能包含很少概念,甚至为空,而起不到扩展的效果,而通过选择合适的语义相关性阈值,图像语义概念关系图可以很好地发现那些初始类标签中遗漏的语义标注词.

6 总结和展望

Web 上出现的海量图像资源很早就引起了研究者的兴趣,而 Web 图像语义自动标注则是管理和检索这些快速增长的海量数据的一种有效途径,尽管已经取得了一些研究成果,Web 图像语义自动标注的性能仍远远没有达到实用的要求.Web 图像语义自动标注中有两个关键的问题:如何自动获取高质量的训练集和如何根据训练集来准确地标注未知图像,本文对这两个问题进行了研究.对于前一个问题,本文提出并验证了可以通过考察 Web 图像的关联文本和 Web 图像人工 tagging 资源来获取高质量的训练数据.同时,我们也提出了一种自适应的

Web 图像语义自动标注方法.本文的标注方法综合考虑了 Web 图像的视觉特征和关联文本对预测图像语义的贡献,特别地,我们通过带约束的分段惩罚加权回归模型将关联文本权重分布估计和先验知识约束有机地结合在一起,自适应地对图像语义在其关联文本上的分布进行建模.实验证明本文所提方法使得 Web 图像语义自动标注性能有显著提高.

鉴于 Web 图像语义在其关联文本上分布的复杂性,在下一步的工作中,我们将考察其他模型,如广义加法模型和 boosting 方法等,以更好地对图像语义在其关联文本上的分布进行建模;同时,我们也将考虑更合理的先验知识约束.另外,由于 Web 数据中存在大量的噪音,我们将考虑 Web 标签数据和一些比较标准的知识源(如 WordNet)的结合,从而更好地度量图像语义概念之间的关系.

References:

- [1] Li XR, Chen L, Zhang L, Lin FZ, Ma WY. Image annotation by large-scale content-based image retrieval. In: Nahrstedt K, *et al.*, ed. Proc. of the 14th ACM Int'l Conf. on Multimedia. Santa Barbara: ACM Press, 2006. 607–610.
- [2] 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. Image and Video Retrieval. 2006. 1483–1490.
- [3] Feng HM, Shi R, Chua TS. A bootstrapping framework for annotating and retrieving WEB images. In: Schulzrinne H, *et al.*, eds. Proc. of the 12th ACM Int'l Conf. on Multimedia. New York: ACM Press, 2004. 960–967.
- [4] Tseng VS, Su JH, Wang BW, Lin YM. WEB image annotation by fusing visual features and textual information. In: Proc. of the 2007 ACM Symp. on Applied Computing, Symposium on Applied Computing. New York: ACM Press, 2007. 1056–1060.
- [5] Mori Y, Takahashi H, Oka R. Image-to-word transformation based on dividing and vector quantizing images with words. In: Proc. of the 1st Int'l Workshop on Multimedia Intelligent Storage and Retrieval Management. 1999.
- [6] Duygulu P, Barnard K, de Freitas JFG, Forsyth DA. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Proc. of the European Conf. on Computer Vision. 2002. 97–112.
- [7] Blei D, Jordan M. Modeling annotated data. In: Proc. of the Int'l ACM SIGIR. Toronto: ACM Press, 2003. 127–134.
- [8] 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.
- [9] Li J, Wang J. Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2003,25(19):1075–1088.
- [10] E. Chang, G. Kingshy, G. Sychay, and G. Wu. Cbsa: Content-Based soft annotation for multimodal image retrieval using Bayes point machines. IEEE Trans. on CSVT, 2003,13(1):26–38.
- [11] 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 and Whistler: MIT Press, 2004. 553–560.
- [12] 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.
- [13] Carneiro G, Vasconcelos N. Formulating semantic image annotation as a supervised learning problem. In: Proc. of the 23d Conf. on Computer Vision and Pattern Recognition. San Diego: IEEE Computer Society Press, 2005.
- [14] 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.
- [15] Gao S, Wang DH, Lee CH. Automatic image annotation through multi-topic text categorization. In: Proc. of. the Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). Toulouse: IEEE Computer Society, 2006. 377–380.
- [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. Image and Video Retrieval. 2006. 102–112.
- [17] Yang CB, Dong M. Region-Based image annotation using asymmetrical support vector machine-based multiple-instance learning. In: Proc. of the 2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. New York: IEEE Computer Society, 2006. 2057–2063.

[18] Carneiro G, Chan AB, Moreno PJ, Vasconcelos N. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2007,29(3):394–410.

[19] Sanderson HM, Dunlop MD. Image retrieval by hypertext links. In: *Proc. of the 20th Annual Int'l ACM SIGIR*. Philadelphia: ACM Press, 1997. 296–303.

[20] Shen HT, Qoi BC, Tan KL. Giving meaning to WEB images. In: *Proc. of ACM Int'l Conf. on Multimedia*. ACM Press, 2000. 39–47.

[21] Yang C, Dong M. Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. In: *Proc. of the 2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*. New York: IEEE Computer Society, 2006. 2057–2063.

[22] Yates RB, Neto BR. *Modern Information Retrieval*. New York: ACM Press, 1999. 123–129.

[23] Dijkstra E. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1959. 269–271.

[24] Zhou XD, Wang M, Zhang Q, Zhang JQ, Shi BL. Automatic image annotation by an iterative approach: Incorporating keyword correlations and region matching. In: *Proc. of the 6th ACM Int. Conf. on Image and Video Retrieval (CIVR'07)*. Amsterdam: ACM Press, 2007. 25–32.

[25] <http://htmlparser.sourceforge.net>



许红涛(1980—),男,博士,主要研究领域为多媒体数据库,信息检索.



向宇(1985—),男,硕士生,主要研究领域为多媒体数据库,信息检索.



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,主要研究领域为数据库理论与应用.