软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*,2020,31(2):531−543 [doi: 10.13328/j.cnki.jos.005630]

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

# 基于语义标签生成和偏序结构的图像层级分类*

顾广华[1,2], 曹宇尧[1,2], 李 刚[1,2], 赵 耀[3]

[1](燕山大学 信息科学与工程学院,河北 秦皇岛 066004)

[2](河北省信息传输与信号处理重点实验室(燕山大学),河北 秦皇岛 066004)

[3](北京交通大学 信息科学研究所,北京 100044)

通讯作者: 顾广华, E-mail: guguanghua@ysu.edu.cn; 曹宇尧, E-mail: cyy19921129@163.com

**摘 要**: 智能电子设备和互联网的普及,使得图像数据爆炸性膨胀.为了有效管理复杂图像资源,提出一种基于加权语义邻近集和形式概念偏序结构的图像层级分类方法.首先,根据图像语义相关分数,对不同程度语义设定自适应权系数,从训练图库中构建加权语义邻近集,通过对语义邻近集中图像的词频分布进行判决,自动生成图像的多个语义标签;然后,以每幅图像为对象,以每幅图像自动生成的语义标签为属性,构建形式背景,通过偏序结构算法对复杂图像集进行有效的层级分类.该方法可以得到图像库中图像之间明确的结构关系和图像类别之间的从属关系,为复杂图像大数据进行层级分类管理提供了有效的思路.对 Corel5k、EspGame 和 Iaprtc12 这 3 个数据库进行了图像标注实验,证明了标注的语义完整性和主要语义的准确性;并对 Corel5k 数据库进行了图像的层级分类实验,结果表明,层级分类效果显著.

**关键词**: 加权语义邻近集;词频分布;语义标签;偏序结构;层级分类

**中图法分类号**: TP391

中文引用格式: 顾广华,曹宇尧,李刚,赵耀.基于语义标签生成和偏序结构的图像层级分类.软件学报,2020,31(2):531−543. http://www.jos.org.cn/1000-9825/5630.htm

英文引用格式: Gu GH, Cao YY, Li G, Zhao Y. Image hierarchical classification based on semantic label generation and partial order structure. Ruan Jian Xue Bao/Journal of Software, 2020,31(2):531−543 (in Chinese). http://www.jos.org.cn/1000-9825/5630.htm

## Image Hierarchical Classification Based on Semantic Label Generation and Partial Order Structure

GU Guang-Hua[1,2], CAO Yu-Yao[1,2], LI Gang[1,2], ZHAO Yao[3]

[1](School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China)

[2](Hebei Key Laboratory of Information Transmission and Signal Processing (Yanshan University), Qinhuangdao 066004, China)

[3](Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China)

**Abstract**: The popularity of smart electronic devices and the Internet makes the image data explode. In order to effectively manage the complex image resources, this study proposed an image hierarchical classification method based on a weighted semantic neighborhood set and formal concept partial order structure. Firstly, a weighted coefficient on different semantics is adaptively designed by the image semantic correlation scores, and a weighted semantic neighborhood (WSN) is constructed from the training sets. The semantic labels of

* 基金项目: 国家自然科学基金(61303128); 河北省自然科学基金(F2017203169, F2018203239); 河北省高等学校科学研究重点项目(ZD2017080); 河北省留学回国人员科技活动项目(CL201621)

Foundation item: National Natural Science Foundation of China (61303128); Natural Science Foundation of Hebei Province (F2017203169, F2018203239); Key Research Project of Higher Education of Hebei Province (ZD2017080); Science and Technology Foundation for Returned Overseas Talents of Hebei Province (CL201621)

收稿时间: 2017-10-25; 修改时间: 2017-12-31; 采用时间: 2018-07-25; jos 在线出版时间: 2019-01-21

CNKI 网络优先出版: 2019-01-22 13:49:18, http://kns.cnki.net/kcms/detail/11.2560.TP.20190122.1348.015.html

the images are automatically generated by judging the word frequency distribution of the images in the semantic neighborhood set. Then, the context is built by taking the images as the objects and the semantic labels as the attributes. This study also proposed an efficient hierarchical classification method for complex image dataset based on the partial order structure. The hierarchical classification method can get the explicit structure relation and the subordinate relationship between the image categories, which provides an effective idea for the hierarchical classification management of the complex images of large data. Three datasets Corel5k, EspGame, and Iaprtc12 were labeled by the WSN method. The label result proved the integrity of the image semantics and the accuracy of the main semantics. Further, the Corel5k dataset was performed by the hierarchical classification method. The experimental results showed the significant performance of the hierarchical classification.

**Key words**: weighted semantic neighborhood set; word frequency distribution; semantic label; partial order structure; hierarchical classification

随着互联网技术和智能拍照电子设备的普及,人类生活中所接触到的图像信息日益膨胀,如何实现对这些图像数据的自动分类与管理,显得越来越重要.图像语义分类作为人工智能领域中的重要任务之一,其主流方法是通过对选定的图像集进行学习,训练分类器模型,并对未知图像进行识别分类决策[1].这种分类方法在单目标图像分类中已经得到了很好的实验结果,但在多目标图像分类问题上效果并不太理想.

多目标图像分类的主要难点在于图像中多目标的识别和对多目标图像库的分类方式.多目标识别问题可看作多目标图像的标注问题,目前,多目标图像标注模型有基于分类器的机器学习方法[2−4]和基于 $K$ 近邻(KNN)算法的标签传播算法[5−8].其中,基于分类器的机器学习方法将每个语义作为一个类标签,使用 SVM 决策树或随机森林算法进行学习分类,然后利用分类结果对图像进行标注.这种分类方法在类别较少、且每类训练样本相对均衡的分类任务中可以获得较好的分类效果.但对于多目标图像库来说,图像类别及各类的图像数目都不确定,分类也比较复杂,导致基于分类器的方法难以训练出理想的分类器.基于 KNN 的标签传播算法将图像标注问题转换成了图像检索问题,根据视觉相似度,从训练集中检索出指定数量的图像,并将其关键词传递给待标注图像.这种方法对于多目标图像来说很容易检索到与显著语义相关的图像,但却容易造成非显著(并非不重要)语义标注的缺失.另外,训练集通常为多目标图像,也易出现错误词汇传递问题.针对语义缺失问题,本文提出了一种基于加权语义邻近集(weighted semantic neighborhood,简称 WSN)的图像标注方法.为了保证标注图像时语义的全面性,从图像库的每类图像中各取几幅与待标注图像相似的图像构成语义邻近集.语义邻近集中,每类图像的数目 $K$ 由其与待标注图像相似度而定,相似度越高,$K$ 就越大;之后,对语义邻近集中图像根据相关语义分数选择一部分图像,由这些图像的标注词汇构建初始词集;最后,为了避免标签错误传递问题,本文计算词频分布,设定阈值选择对应的词来标注图像,从而减少词语标注的错误率.

分类方式是多目标图像分类的另一难点.对于多目标图像来说,传统的单层分类方式容易造成类别间图像的重复,而且体现不出图像之间明确的结构关系和图像类别之间的从属关系.基于上述问题,本文试图构建一种图像层级分类方式.形式概念分析[9−11]是一种从形式背景进行数据分析和规则提取的强有力工具,对组成本体的概念、属性以及关系等用形式化的语境表述出来,根据语境构造出概念格,即本体,从而清楚地表达出本体的结构.偏序结构[12]是一种基于形式概念分析的数据分析结构算法,具有简洁、层级关系鲜明等优点.属性偏序结构图是将数据中具有某些共同特征的对象聚集到一起,进行数据共性的表达.本文通过构建加权语义邻近集获得复杂图像中多目标的图像标注词汇后,提出了一种基于偏序结构(partial order structure,简称 POS)的多目标图像层级分类方法,以每幅图为对象,以其对应的标签为属性构建形式背景,进行属性偏序结构算法的计算,从而完成多目标图像数据库的层级分类,对多目标图像数据库进行合理的分类整理.方便用户从图库中根据层级关系逐层查找图像,实现对多目标图库更简洁、清晰、有效的层级分类管理.

## 1 语义标签

### 1.1 加权语义邻近集构建

与传统的图像特征相比,卷积神经网络(CNN)[13]提取的深度特征具有更好的泛化能力,在图像分类和图像

检索领域表现惊艳.比如在2014年的ILSVRC[14]比赛上,VGGNet[15]在图像识别竞赛中夺取亚军.研究表明[16,17],从网络结构中全连接层提取的4 096维特征,是一般识别任务的极佳表示.因此,本文使用VGGNet-16[18]中第15层全连接层输出的4 096维特征来对图像进行实验.

每张图像都包含多个语义,而语义又分为主要语义和次要语义.为了更清楚地理解主要语义和次要语义,以图1为例,图中较大方框内标定的内容为主要语义,较小方框内标定的内容为次要语义.在利用特征之间的视觉相似性进行图像检索时,往往出现只能检索主要语义而忽略次要语义的情况.针对这一问题,为了保证语义的完整性,本文依据图像相似度,从每个语义集中选择不同的图像构建语义邻近集合,然后根据深度特征距离得分计算语义邻近集中的图像对待标注图像的贡献值,通过贡献值排序完成图像的初始标注.

Fig.1 Diagram of image semantic distinction

图1 图像语义区分示意图

训练集由图像集 $\boldsymbol{I}$ 和语义关键词集 $\boldsymbol{W}=\{W_1,...,W_m\}$ 组成.将图像集按照图像语义进行分类,为了避免重复计算贡献值,对于包含多种语义的图像,只根据一种语义分类1次,得到语义分类集 $\boldsymbol{G}=\{(W_1,\boldsymbol{I}_1),...,(W_m,\boldsymbol{I}_m)\}$,其中,$\boldsymbol{I}_i$ 为语义关键词 $W_i$ 对应的图像集合.设 $I^*$ 为待标注图像,计算语义集 $\boldsymbol{G}_i=(W_i,I_i)(i=1,2,...,m)$ 中每幅图像与 $I^*$ 的相关分数(相关分数越大,说明该图像与待标注图像越相似)得到的 $\boldsymbol{G}_i$ 相关分类集 $\boldsymbol{\alpha}_i=(\alpha_{ij})$:

$$\alpha_{ij}=\alpha_{I_j,I^*}=\frac{1}{1+\exp(\beta\cdot Dis(I_j,I^*))} \tag{1}$$

其中,$I_j\in\boldsymbol{I}_j$,$\beta=15$,$Dis(I_j,I^*)$ 为 $I_j$ 和 $I^*$ 的视觉距离.

$$Dis(I_j,I^*)=\frac{L(I_j,I^*)}{\max(L(I_j,I^*))},\ L(I_j,I^*)=1-\frac{Cov(I_j,I^*)}{\sqrt{D(I^*)}\sqrt{D(I_j)}} \tag{2}$$

其中,$L(I_j,I^*)$ 为 $I_j$ 和 $I^*$ 的相关距离,$Cov(I_j,I^*)$ 为 $I_j$ 和 $I^*$ 的协方差,$D(I_j)$ 和 $D(I^*)$ 分别为 $I_j$ 和 $I^*$ 的方差.

构建语义邻近集的主要目的是为了保证语义的完整性.进一步地,为了保证图像主要语义的准确性,本文对语义邻近集中各语义类图像数目进行加权.加权系数由各语义类图像与测试图像的相关分数自适应设定(见公式(3)).计算待标注图像与各语义类图像集的语义相关分数集,取每类语义相关分数的最大值组成相关分数向量.依据语义相关性,把相关分数向量分为与主要语义、次要语义和不相干语义分别相对应的3部分,如图2所示.利用公式(3)计算与主要语义、次要语义和不相干语义相对应的权系数$[K_1,K_2,K_3]$,对语义邻近集加权得到加权语义邻近集 $\boldsymbol{S}$,以表征不同语义对待标注图像语义的贡献程度.具体步骤如下.

(1) 计算 $m$ 类语义图像集与待标注图像的相关分数集 $\boldsymbol{A}=\{\boldsymbol{\alpha}_1,...,\boldsymbol{\alpha}_m\}$;取每类相关分数集 $\boldsymbol{\alpha}_i$ 的最大值 $\boldsymbol{M}_i$ 并降序排列,得到语义类图像集的相关分数向量 $\boldsymbol{M}=\{M_1,...,M_m\}$.

(2) 把相关分数向量 $\boldsymbol{M}$ 分3段,分别对应主要语义、次要语义和不相干语义.排序前两类为第1段,对应待测图像的主要语义,记为 $\boldsymbol{M}_1=\{M_1,M_2\}$;排序第3类~第$\lceil(m/2)\rceil$类为第2段,对应待测图像的次要语义,记为 $\boldsymbol{M}_2=\{M_3,...,M_{\lceil(m/2)\rceil}\}$;其他为第3段,对应不相干语义,记为 $\boldsymbol{M}_3=\{M_{\lceil(m/2)\rceil+1},...,M_m\}$.$\lceil\rceil$表示上取整.

(3) 利用公式(3)计算3段相关分数向量的均值和比重,作为主要语义、次要语义和不相干语义对应的语义类图像个数权重 $K_1,K_2,K_3$,$K_1+K_2+K_3=1$.构建加权语义邻近集,表示3种语义对待测图像的重要程度.

$$k_h=mean(\boldsymbol{M}_h),\ K_h=\frac{k_h}{k_1+k_2+k_3} \tag{3}$$

其中,$h=1,2,3$ 分别表示语义类别中主要语义、次要语义和不相干语义这3个语义等级,$k_h$ 表示相应语义等级中

语义相关分数的平均值.需要说明的是,为保证图像语义的完整性和主要语义的准确性,本文设定 $K_1$ 在 0.5~0.7 范围内取值.如果由公式(3)计算出 $K_1$ 的值小于 0.5 或大于 0.7 时,设置权重$[K_1,K_2,K_3]$=[0.7,0.2,0.1],以保证主要语义的正确性和语义邻近集中语义的全面性.图 2 所示为构建语义邻近集的示意图.


Fig.2 Construction of weighted semantic neighborhood sets

图 2 构建加权语义邻域集

### 1.2 语义标签生成

构建加权语义邻近集以后,计算待标注图像 $I^*$与语义邻近集 $\boldsymbol{S}$ 的相关分数集 $\boldsymbol{a}_s$,选择 $\boldsymbol{a}_s$ 中前 $s$ 个最大值对应的图像,这些图像对应的关键词构成初始词集 $\boldsymbol{W}_s$.构成初始词集 $\boldsymbol{W}_s$ 的 $s$ 张图像中几乎每张图像都包含待标注图像 $I^*$的主要语义,其中部分图像也包含 $I^*$的次要语义.虽然这些图像中有可能包含不相干语义,但不同图像包含同一种不相干语义的概率非常低,所以词集中主要语义和次要语义对应词汇的词频要远远大于不相干语义对应词汇的词频.因此,本文通过对词集 $\boldsymbol{W}_s$ 设置合适的词频阈值 $f$,选择合适的关键词构成待标注图像 $I^*$的标注关键词集 $W^*$,可以避免错误词汇的传递.算法的过程如图 3 所示.

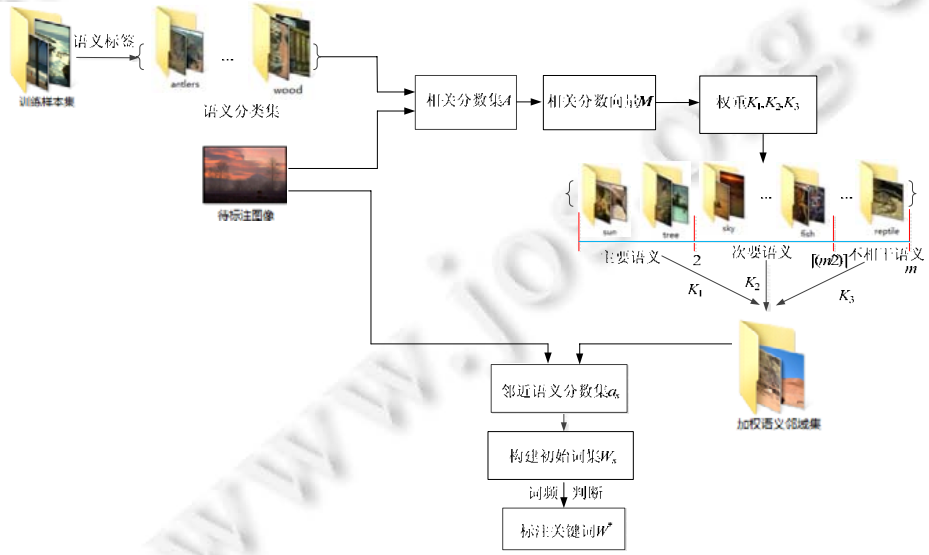



Fig.3 Diagram of semantic label generation based on WSN

图 3 基于 WSN 的语义标签生成框图

## 2 基于偏序结构的图像层级分类

### 2.1 形式背景

数据集可以由对象及其属性构建的形式背景来表示.

**定义 1**. 设 $\boldsymbol{U}$ 是对象的集合,$\boldsymbol{V}$ 是属性的集合,$\boldsymbol{I}$ 是两个集合 $\boldsymbol{U}$ 与 $\boldsymbol{V}$ 间的关系,则称三元组 $\boldsymbol{K}=(\boldsymbol{U},\boldsymbol{V},\boldsymbol{I})$为一个形式背景(简称背景).$(u,v)\in\boldsymbol{I}$ 表示对象 $u$ 具有属性 $v$.背景可以用一个矩形表来表示,如表 1 所示,每一行为一个对象,每一列为一个属性.若 $u$ 行 $v$ 列的交叉处为 1,则表示对象 $u$ 具有属性 $v$;若 $u$ 行 $v$ 列的交叉处为 0,则表示对象 $u$ 不具有属性 $v$.

将多目标图像库中的每个图像生成对应的标签以后,以每幅图像为对象,以标签为属性,将图像数据集转化为形式背景.以图 4 为例简要说明.6 幅图为对象,标签 outdoor、indoor、bird 等作为属性,建立形式背景见表 1,其中,数字 1 表示图片包含相应的标签,0 表示不包含.


Fig.4 Example of multi-objective image

图 4 多目标图像示例图

**Table 1** Context corresponding to Fig.4

**表 1** 图 4 对应的形式背景

| | Outdoor | Indoor | Bird | Sky | Dog | People | Working | Street | Subway | Bed |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 4 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 6 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |

### 2.2 偏序结构

针对由图像库构建的形式背景进行偏序结构运算,并对图库进行层级分类.定义属性度 $D_a$ 为属性 $a$ 对应的对象数,属性贡献度 $B_{a,b}$ 为属性 $a$ 对应的对象中不包含属性 $b$ 的对象数.首先,根据每个属性的属性度对属性进行降序排序;然后,依次计算两个属性之间的属性贡献度,并根据属性贡献度的大小再进行降序排列.表 2 所示为表 1 重排后的形式背景.

**Table 2** Rearranged context

**表 2** 重排后的形式背景

| | Outdoor | Indoor | Dog | People | Working | Bed | Bird | Sky | Street | Subway |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 3 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 6 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |

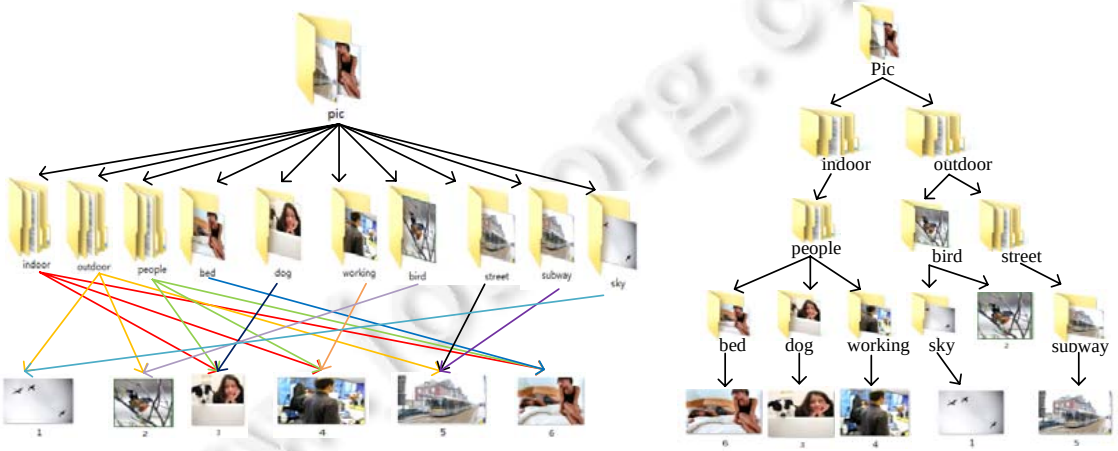
**定义 2**. 如果形式背景 $K=(U,V,I)$和 $K_1=(U_1,V_1,I_1)$满足 $U_1\subseteq U,V_1\subseteq V$,进而推出 $I_1\subseteq I$,则 $K_1$ 为 $K$ 的子背景,记为 $K_1\subseteq K$.

选择最少的前 $n$ 列属性进行求并运算得到一个全 1 序列,比如表 2 中,$n$=2.表示前 $n$ 个属性为能够包含所有图像的最少的标签,将这些标签建立文件夹作为第 1 层分类.分别依次选择这 $n$ 个属性中为 1 的行向量(为了避免重复运算,若 $n$>1,还要满足行向量中的 1:$n$–1 列中不包含 1)构成子背景.

子背景去掉全 1 列向量,如果含有全 0 行向量,说明这一行对应的图像已经没有了可以再分类的标签,则将对应图片保存到当前文件夹下,完成此图像的分类.

将每个分类属性下的子背景重复以上计算,直到所有子背景为全 0 矩阵为止,完成对图库的全部分类.

对于多目标图像,为了更清楚地说明基于本文偏序结构算法的层级分类效果相对于单层分类的优势,将图 4 中 6 张图片进行偏序结构算法层级分类和传统的单层分类,分类效果如图 5 所示.




(a) 单层分类图例 (b) 层级分类图例

Fig.5 Example for one-level classification and hierarchy classification

图 5 单层分类和层级分类示例

在图 5 中,每个箭头下的标签或图像表示上个标签文件包含的类别或图像.图 5(a)所示为 6 张示例图像按照图像类别标签的单层分类结果,图 5(b)为层级分类结果.为了更清楚地表示分类效果,图 5(a)中图像所属的每个类别用不同颜色的箭头标出.从图 5(a)中可以看出,每张多目标图像都不同程度地进行了多次分类运算,并且从箭头的交叉上可以看出分类效果混乱.如果要对图像进行多标签检索,比如要查找同时包含 people 和 dog 的图像,就需要查找 people 和 dog 文件下重复的图像,这在操作上非常不便.而从图 5(b)中可以看到,根据多目标图像类别标签之间的包含关系进行层级分类后,其分类结构非常清晰,每张图像都不会出现重复分类的问题.对层级分类进行多标签检索时,可以按照层级关系逐层查找,比如寻找同时包含 people 和 dog 的图像,可以先找到包含标签 people 的分类支路,然后再寻找标签 dog 下的图像即可,这样就避免了对其他支路标签的查找.如果只对 1 个类别标签进行检索,比如 bird,虽然图 5(a)的单层分类方式可以快速地从 bird 文件下快速找出所有包含 bird 标签的图像,但无法判断 bird 标签在这些图像的所有标签中所占的比重(或者说无法判断除 bird 标签以外还有没有其他标签以及其他标签的个数).如果利用图 5(b)的层级分类结构对 bird 簇下的图像进行逐层查找,不但可以找出包含 bird 标签的所有图像,而且可以按照标签在图像中所占的比重对其排序,比如图 5(b)中图像 2 的标签中,bird 所占比重为 1/2;图像 1 的标签中,bird 所占比重为 1/3.

多目标图像含有多个语义标签,对于多目标图像,仅按照单个标签类别分类是不合理的,如图 5(a)中的图像 3、图像 4、图像 6,虽然图像中都包含 person 标签,但严格来说,3 张图像语义表达并非完全一致.为了更加合理地对多目标图像进行类别描述,本文结合层级分类结构提出组合类的概念.在构造层级结构时,本文的 POS 算法

将数据集中所有标签按照属性度和标签之间的关联规则进行有序排列,并且将具有相同标签排序的图像聚集到一起.从图 5(b)中可以看到,在层级分类结构中,每类多目标图像都对应层级结构的一个支路,本文将每个支路上标签的有序排列称为图像的组合类别(比如{indoor,people,bed},{outdoor,bird,sky}等).如果存在组合类别 $C_a \cup C_b = C_a$,那么组合类别 $C_a$ 称为组合类别的子类,$C_b$ 称为 $C_a$ 的父类.如图 5(b)中,图像 1 对应的组合类为{outdoor,bird,sky},图像 2 对应的组合类为{outdoor,bird},则组合类{outdoor,bird,sky}为子类,组合类{outdoor,bird}为父类.在层级分类中,利用组合类别描述多目标图像要比单层分类更加准确,而且层级分类可以清晰地看出图像分类的分支结构,从而得到标签语义类之间的从属关系.

## 3 实验结果与分析

### 3.1 实验数据集

本文实验采用 Corel5k[19]、EspGame[8]和 Iaprtc12[8]这 3 个在图像标注领域常用的数据集.Corel5k 是图像标注常用的标准图像库,其中包含自然风光、人文生活等多目标图像,共包含大小为 192×168 的 5 000 幅彩色图像,每幅图像都被人工标注 1 个~5 个标签,平均每张图像能被 3.4 个标签标注,总共有 260 个标签出现.EspGame 数据集是从在线交互式游戏中获取的 20 770 张图像,其中包括商标、绘图、风景和人物照片等.其中,每幅图像从游戏中提取 1 个~15 个标签,平均每幅图包含 4.7 个语义标签,总共有 268 个标签注释.Iaprtc12 数据集由 19 627 张图像构成,涵盖运动、行为、人物、动物、城市、风景等方面.其关键词是从图像的标语或字幕中提取.在 Iaprtc12 数据集中,图像包含语义标签数最多为 23 个,最少为 1 个,平均每幅图像有 5.7 个语义标签,所有标签数为 291.在 3 个数据集中,EspGame 和 Iaprtc12 数据集在图像和标签数上都大于 Corel5k,其中,Iaprtc12 图像中包含的标签数最多,标注难度最大.数据集中标签和图像的详细信息见表 3.

**Table 3** Information of three benchmark databases

**表 3** 3 个基准数据集的信息

| 数据集 | Corel5k | EspGame | Iaprtc12 |
|---|---|---|---|
| 图像数 | 5 000 | 20 770 | 19 627 |
| 标签数 | 260 | 268 | 291 |
| 训练图像数 | 4 500 | 18 689 | 17 665 |
| 测试图像数 | 500 | 2 081 | 1 962 |
| 每幅图的平均标签数 | 3.4 | 4.7 | 5.7 |
| 每个标签平均出现次数 | 58.6 | 326.7 | 347.7 |

### 3.2 基于WSN的图像标注

实验选用准确率 $P_a$、召回率 $R$、$F$ 值和被准确预测过的标签数 $N^+$作为评判标准:

$$P_a = \frac{1}{M}\sum_{j=1}^{M}\frac{Correct(w_j)}{Predicted(w_j)}\times 100\%,\ R = \frac{1}{M}\sum_{j=1}^{M}\frac{Correct(w_j)}{Ground(w_j)}\times 100\%,\ F = \frac{2P_aR}{P_a+R}\times 100\% \tag{4}$$

其中,$M$ 为测试集中待标注图像的总数目,$Correct(w_j)$为第 $j$ 幅图像生成的所有关键词 $w_j$ 中预测正确的数目,$Predicted(w_j)$为第 $j$ 幅图像生成所有关键词 $w_j$ 的数目,$Ground(w_j)$为测试集中第 $j$ 幅图像的实际标注数.

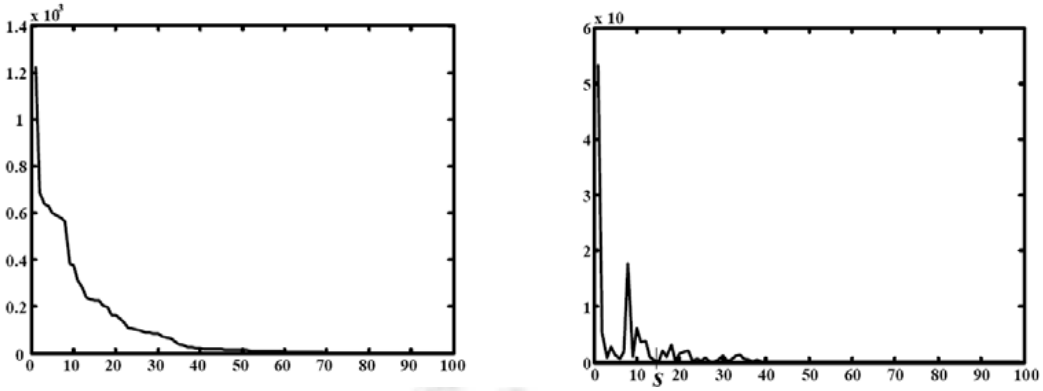
参数设定:因为每张测试图像包含的语义种类和数量都不相同,构成语义集的权重 $K_1,K_2,K_3$ 也不相同.为了确保每张测试图像的自适应能力,本文通过实验来验证设置阈值 $f$ 和 $s$ 的方法.

(1) 设置参数 $s$

构建词集时,从加权语义邻近集中选取与测试图像语义相似度最高的前 $s$ 张图像,其对应的关键词构成待标注图像的初始词集.设置 $s$ 的目的主要是为了避免不相干语义的错误传递.

由第 1.2 节可知,构成初始词集图像的选择是依据测试图像与语义邻近集图像的相关分数.由于每张待标注图像生成语义邻近集时的权重不同,所以与每个语义邻近集相关分数的分布也不同.为了使 $s$ 能够自适应语义邻近集权重的变化,本文按照语义邻近集中相关分数的分布自适应取值 $s$.图 6(a)为语义邻近集中每张图像与

待标注图像相关分数的分布情况,图 6(b)为相关分数的差分分布,横轴为语义邻近集中图像的序号,纵轴为相关分数值.在图 6(a)中可以看出语义邻近集中图像相关分数的分布,相关分数越高,其对测试图像的语义影响越重要.主要语义对测试图像的语义影响最大,次要语义次之,不相干语义影响最小,几乎为零.也就是说,主要语义和次要语义对应的相关分数曲线波动较大,不相干语义波动平缓,如图 6(b)所示.由此,本文依据相关分数波动程度选择合适的分界点 $s$.为了避免 $s$ 过大或过小影响标注效果,本文将 $s$ 的取值范围设定在 5~20.如果 $s$ 值超出了设定范围,则将 $s$ 设置为常数 13.



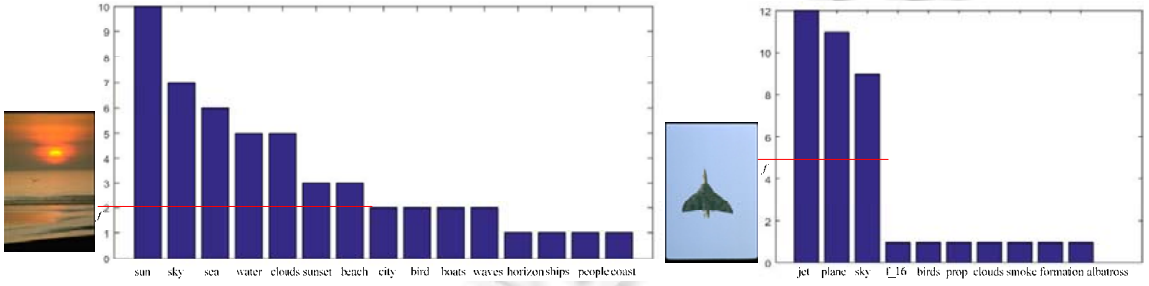

(a) 相关分数分布 (d) 相关分数差分

Fig.6 Correlation score analysis in semantics neighbors

图 6 语义邻近集中相关分数的分析

(2) 设置参数 $f$

在构造初始词集时,与测试图像相似的图像也很有可能会传递不相干的语义词汇.但是因为不同图像传递的不相干语义相同的概率很小,所以在词集中主要语义和次要语义的词频会远远大于不相干语义.因此,设置阈值 $f$ 可以有效地防止不相干语义的传递.为了使每个初始词集设置合适的词频阈值 $f$,本文分析了初始词集中词频分布与主要语义权重 $K_1$ 的关系.图 7 为不同 $K_1$ 对应初始词集的词频分布,其中,横轴为初始词集中的关键词,纵轴为关键词对应的词频.图 7(a)为 $K_1$ 相对较小时(取值 0.6)初始词集的词频分布,图 7(b)为 $K_1$ 较大时(取值 0.7)初始词集的词频分布.由图 7 可以看出,阈值 $f$ 的大小和主要语义的权重 $K_1$ 有正比关系.因为图像语义越单一,其主要语义所占的比重就越大,主要语义和次要语义与不相干语义词频的差值也越高,所以阈值 $f$ 也就越大.



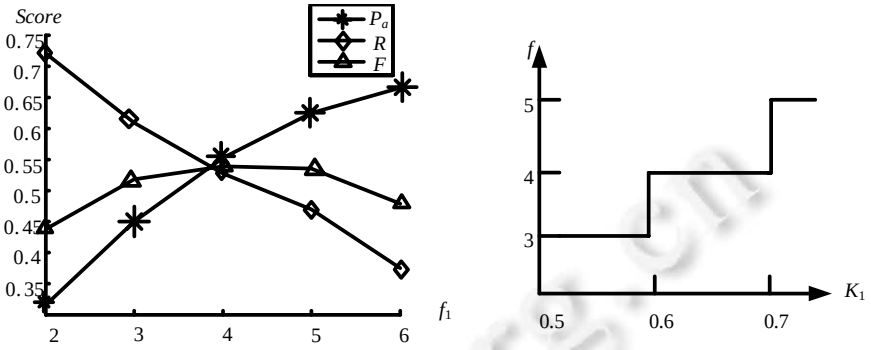

(a) 语义邻近集权重 6:3:1 (b) 语义邻近集权重 7:2:1

Fig.7 Frequency distribution of different semantic neighbors' weights

图 7 不同语义邻近集权重的词频分布

虽然根据 $K_1$ 可以大致判断 $f$ 随语义邻近集的波动,但是并不能准确无误地确定最小相关语义的词频.本文先对所有语义邻近集设置词频参数 $f_1$,通过调整词频 $f_1$,可以推测词频阈值 $f$ 对实验结果的影响.图 8(a)为标注结

果随 $f_1$ 的变化情况,横轴为 $f_1$,纵轴为 $P_a$、$R$ 和 $F$ 值.从图 8(a)可以看到,随着阈值 $f_1$ 的增加,准确率 $P_a$ 逐渐上升,召回率 $R$ 逐渐下降.为了避免准确率和召回率过低,同时也为了使词频阈值适应不同语义近邻集权重的变化,本文根据 $K_1$ 与阈值 $f$ 的正比关系,设计 $f$ 与 $K_1$ 的分段函数如图 8(b)所示,在范围 3~5 内设计词频阈值 $f$ 与 $K_1$ 的分段函数,以适应语义邻近权重对词集词频的变化.




(a) $P,R,F$ 随 $f_1$ 的变化曲线 (b) $f$ 关于 $K_1$ 的分段函数

Fig.8 Design of piecewise function $f$

图 8 设计分段函数 $f$

为了证明 WSN 算法的有效性,本文在 Corel5k,EspGame 和 Iaprtc12 这 3 个数据集上与其他一些经典的图像标注方法进行比较,对比结果见表 4.

**Table 4** Performance comparison of several image label methods on three datasets

**表 4** 在 3 个数据集上几种图像标注方法的性能比较

| | Corel5k | | | | EspGame | | | | Iaprtc12 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| method | $P_a$ | $R$ | $F$ | $N^+$ | $P_a$ | $R$ | $F$ | $N^+$ | $P_a$ | $R$ | $F$ | $N^+$ |
| MBRM[19] | 0.24 | 0.25 | 0.24 | 122 | 0.18 | 0.19 | 0.18 | 209 | 0.24 | 0.23 | 0.23 | 223 |
| GS[20] | 0.30 | 0.33 | 0.31 | 146 | – | – | – | – | 0.32 | 0.29 | 0.30 | 252 |
| JEC[7] | 0.27 | 0.32 | 0.29 | 139 | 0.24 | 0.19 | 0.21 | 222 | 0.29 | 0.19 | 0.23 | 211 |
| LM3L[21] | 0.33 | 0.37 | 0.35 | 146 | 0.40 | 0.26 | 0.32 | 239 | 0.44 | 0.28 | 0.34 | 242 |
| RNN[22] | 0.31 | 0.39 | 0.35 | 153 | – | – | – | – | 0.33 | 0.31 | 0.32 | 255 |
| TagProp[8] | 0.33 | 0.42 | 0.37 | 160 | 0.39 | 0.27 | 0.32 | 239 | 0.46 | 0.35 | 0.40 | 266 |
| CCA-KNN[23] | 0.42 | **0.52** | 0.46 | **201** | 0.46 | 0.36 | 0.40 | **260** | 0.45 | 0.38 | 0.41 | **278** |
| Multi-label CNN[24] | 0.41 | 0.35 | 0.38 | **–** | – | – | – | – | – | – | – | – |
| WSN | **0.60** | 0.47 | **0.53** | 122 | 0.45 | **0.38** | **0.41** | 230 | **0.48** | **0.46** | **0.47** | 252 |

从表 4 可以看出,本文提出的基于 WSN 的图像标注方法在 3 个数据集中都取得了最好的 $F$ 值,说明本文方法整体上优于其他方法.虽然在图像数和平均标签数较小的 Corel5k 数据集上,本文方法召回率 $R$(0.47)略低于 CCA-KNN 方法(0.52),但其准确率 $P_a$ 高达 0.60,明显优于其他方法.这是因为本文在构建语义邻近集时对主要语义、次要语义和不相干语义设置不同的权重,保证了主要语义的准确性,并且自适应选择阈值 $s$ 和词频 $f$ 可以有效地防止错误语义传递,所以准确率 $P_a$ 有着明显的优势.

在图像数和平均标签数较大的 Iaprtc12 和 EspGame 数据集上,本文基于 WSN 的图像标注方法有着更佳的表现.除了在 EspGame 数据集中,其准确率(0.45)和 CCA-KNN 方法的准确率(0.46)持平以外,准确率和召回率指标均高于其他方法,特别是召回率 $R$ 在 Iaprtc12 数据集中表现尤为突出(0.46).由表 3 可知,Iaprtc12 数据集中每幅图的平均标签数(5.7)最高,本不易得到较高的召回率,而本文方法通过构建语义邻近集,保证了语义标签的全面性,取得了远优于其他方法的召回率.

由表 4 可以看出,基于 WSN 的图像标注方法得到的标注词汇个数 $N^+$值在所有图像标注方法中位于中游.这是因为本文采用语义加权的方式,对每幅图像都对其主要语义采用较大的权重,不可避免地会损失一些非主要语义对应的标注词汇,所以在所有测试图像上得到的 $N^+$值会有相对不太高的结果.但是从召回率 $R$ 的指标上

来看,本文基于WSN的图像标注方法对每幅待标注图像生成语义标签的全面性高于其他方法.综合所有判别标准,基于 WSN 的图像标注方法在提高图像标注性能上有着比较明显的意义.

表 5 为标注实例,其中,第 2 列为 3 个数据集中 6 张复杂度较高的测试图像,第 3 列为测试图像人工标注的真值,第 4 列为通过 WSN 算法对测试图像生成的关键词.第 4 列中加粗的词汇为 WSN 生成与真值相同的词汇,斜体词汇为 WSN 生成的词汇中语义正确但在真值中不存在的词汇.由表 5 中结果可以看出,通过 WSN 算法,不但可以对图像的主要语义和次要语义进行正确的标注,而且可以扩展人工标注词汇.

**Table 5** Annotation instance of the WSN method on three database

**表 5** WSN 方法在 3 个数据集上的标注实例

| 数据集 | Corel5k | | EspGame | | Iaprtc12 | |
|---|---|---|---|---|---|---|
| 图像 |  |  |  |  |  |  |
| 人工标注 | water, beach, people, kauai | water, beach, people, sunset | home, house, road, roof, street, tree, white | black, dog, grass, green, guy, man, run, shoes, white | building, bush, slope, square | bag, lookout, man, rock sky, tee_shirt, tree, waterfall, woman |
| WSN 标注 | **water**, **beach**, **people**, *sand*, *sky* | **water**, **sunset**, *sky*, *sun*, *clouds*, *horizon* | *grass*, *green*, **house**, **tree**, *sky*, **street** | **black**, **dog**, **grass**, **green**, **man**, *tree*, **white** | *river*, **rock**, **waterfall**, *jungle*, *man*, *view*, **sky**, *slope*, *mountain* | *front*, *people*, **square**, **building**, *tourist*, *group*, *lamp*, *street*, *sky*, desert |

### 3.3 基于POS的层级分类

在基于偏序结构算法的多目标图像层级分类实验中,使用准确率 $P_c$、召回率 $R_c$、分类最高层级 $H_{\max}$、最小层级 $H_{\min}$ 和图像库中分类第 1 层的标签数目 $L_1$ 来评判层级分类实验.

对于多语义目标图像,每个语义类别都有意义.所以,在计算 POS 层级分类准确率 $P_c$ 和召回率 $R_c$ 时按照第 2.2 节中组合类别的方法,通过对类别元素的组合,计算出所有组合类别下图像的准确率和召回率.

$$P_c=\frac{1}{N}\sum_{j=1}^{N}\frac{Correct(c_j)}{Class(c_j)}\times 100\%,\ R_c=\frac{1}{N}\sum_{j=1}^{N}\frac{Correct(c_j)}{Ground(c_j)}\times 100\% \tag{5}$$

其中,$N$ 为通过 POS 算法分类后的所有组合类数,$Correct(c_j)$为第 $j$ 个组合类中分类正确的图像个数,$Class(c_j)$为第 $j$ 个组合类中的总图像数,$Ground(c_j)$为第 $j$ 个组合类真值中的总图像数.

基于 POS 的图像分类是依据 WSN 方法生成的标签为属性进行具有层级结构的分类,所以 WSN 生成的标签其准确性非常重要.基于 POS 的层级分类是按照组合类的方式计算准确率,意味着组合类中只要有一个类别元素判别错误,则其所有分支类别结果都会判别为错.如图 9 中如果第 1 层中标签“water”判别错误,其簇下的所有组合类都将判别错误.这样,人工标注中语义的不全面性将会对 POS 的分类结果造成很大的影响.从表 5 中可以看出,基于WSN的图像标注方法生成了很多人工标注中不存在但却是正确的词汇.为了更加准确地描述POS的层级分类效果,本文找了 10 个人对数据量较少的 Corel5k 数据集生成的 WSN 词汇进行判别,若有 8 人认为标签正确,本文就认为此标签是正确的.经过人工判别,WSN 生成词汇的准确率为 0.91.本文在 Corel5k 数据集上做了分类测试,并提供了 Corel5k 数据集的分类数据层级结构图,如表 6 和图 9 所示.

**Table 6** Result of hierarchical classification

**表 6** 层级分类结果

| $P_c$ | $R_c$ | $H_{\max}$ | $H_{\min}$ | $L_1$ |
|---|---|---|---|---|
| 0.79 | 0.82 | 7 | 1 | 32 |

由表 6 可见,基于 POS 层级分类后的分类准确率 $P_c$ 为 0.79,召回率 $R_c$ 为 0.82,层级分类的准确率 $P_c$ 和召回率 $R_c$ 与生成标签的准确率 $P_a$ 有着直接的关系.这是因为基于属性偏序结构算法的多目标图像层级分类是将 WSN 生成的标签作为属性来进行分类的.Corel5k 数据库中 500 张测试图,语义最复杂的图像被分为 7 层,说明

WSN 对其生成了 7 个语义词汇;语义最简单的图像被分为 1 层,说明 WSN 对其生成了 1 个语义词汇.层级分类结构中第 1 层共有 32 个标签,这 32 类标签可以包含数据集中的所有图像,是数据集中最具有概括性的标签.具体层级分类结构如图 9 所示.

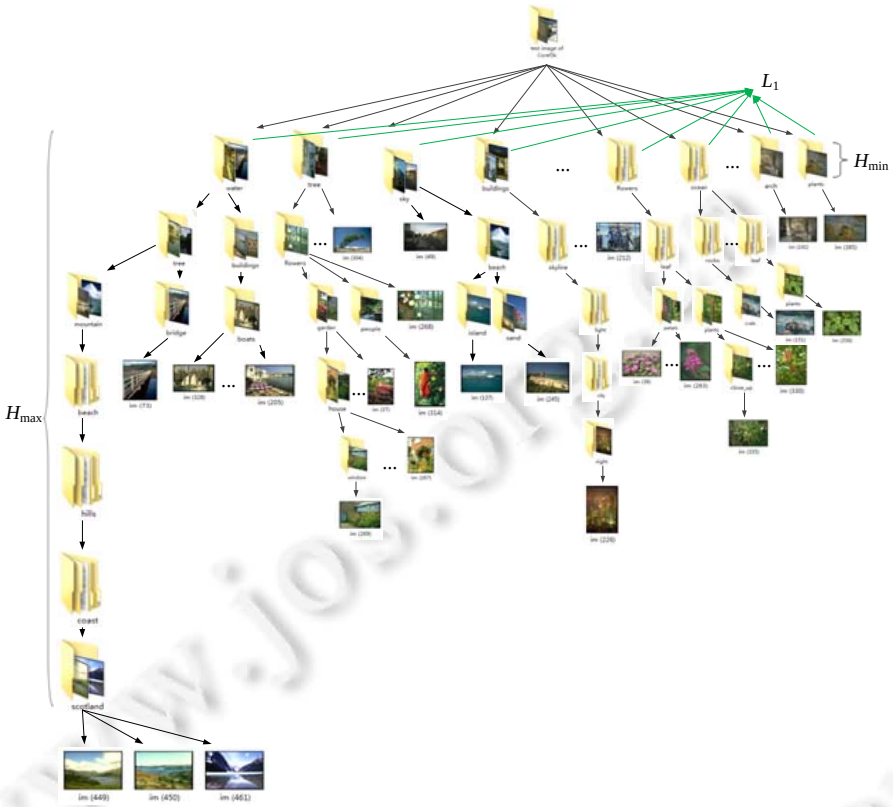



Fig.9 Classification structure of 500 testing image from Corel5k

图 9 Corel5k 中 500 张测试图的分类结构

图 9 是对 Corel5k 数据库中 500 张测试图通过 WSN 标注和基于 POS 后的层级分类结构图.从图 9 中可见,图像的分类层级一共有 7 层,即 $H_{max}$=7,第 7 层为层级分类中最大层级,代表着图像库中语义最复杂图像的分类层数,比如图 9 中的图像 im(449)、im(450)和 im(461),它们一共生成了 7 个语义标签.而且从图 9 中也可以看出,有一些图像之间包含着部分共同标签,如 im(37)和 im(267)二者之间具有 3 个共同的语义标签(tree,flower,garden),表征着它们的共有属性.这是一种新型的分类方式,根据属性先共性、后异性,逐层细化,层级分类.对图像进行检索时,可以根据其共性比如“garden”,能够同时检索出来图像 im(37)和 im(267),也可以根据其特异属性区分检索“house”区分出 im(267).对应于最大分类层级,最小分类层级代表着图像库中语义复杂度最小图像的分类层级,如图 9 所示,最小分类层级为 1,即 $H_{min}$=1,表示单语义标签图像.比如,图像 im(141)只有 1 个标签“arch”.$L_1$ 为层级结构中第 1 层的分类标签个数,第 1 层的分类标签为概括性很强的词汇,涵盖了图库中的所有图像,对数据集的描述有着重要意义.

另外,通过观察图 9 发现一个很有意思的现象:对于每一个语义节点来说,下一层所属的标签中,其概括性从左到右越来越弱,也即特异性越来越强.比如,层级结构第 1 层的语义标签中,“water”的概括性要高于“flower”,也就意味着“water”的语义簇要比“flower”的语义簇复杂,如图 9 所示.

## 4 结 论

为了解决多目标图像标注时语义缺失问题和错误标签传递问题,本文提出了基于 WSN 的图像标注算法,通过构建加权语义邻近集解决了语义缺失问题,从而保证了图像标注时的语义全面性.在生成标签集合后,通过词频判断选择合适的标注词汇,避免了错误标签传递,使得图像标注的准确率和召回率都有了很大的改善.利用WSN模型将图库中的多目标图像生成相应的标签后,本文利用POS算法对其进行有效的层级分类.这种分类方式可以根据图像复杂度对图像库生成明确的层级结构,在分类结构中将图像的相同语义标签共用,可以减小结构复杂度,并且方便通过相同语义标签对图像进行检索.另外,还可以从层级结构中找到图像概括性最强和特异性最强的标签,为图像语义检索提供了一种新的思路.但是图像标注时,对同一个目标可能会出现同义词的问题,比如“people”和“person”,这对 WSN 方法和 POS 分类都会增加难度,这也是本文下一步工作要解决的问题.

**References**:

[1] Cao J, Mao D, Cai Q, Hai-Sheng L, Jun-Ping DU. A review of object representation based on local features. Journal of Zhejiang University-Science C (Computers & Electronics), 2013,14(7):495−504.

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

[3] Qiu ZY, Fang Q, Sang JT, Xu CS. Regional context-aware image annotation. Chinese Journal of Computers, 2014,37(6); 1390−1397 (in Chinese with English abstract).

[4] Gao Y, Fan J, Jain R. Automatic image annotation by incorporating feature hierarchy and boosting to scale up SVM classifiers. In: Proc. of the ACM Int’l Conf. on Multimedia. ACM Press, 2006. 901−910.

[5] Wagstaff K, Cardie C. Clustering with instance-level constraints. In: Proc. of the Int’l Conf. on Machine Learning. 2000. 1103−1110.

[6] Kumar M, RathN, Swain A, Rath SK. Feature selection and classification of microarray data using MapReduce based ANOVA and *K*-nearest neighbor. Procedia Computer Science, 2015,54:301−310.

[7] Makadia A, Pavlovic V, Kumar S. A new baseline for image annotation. In: Proc. of the European Conf. on Computer Vision. Springer-Verlag, 2008. 316−329.

[8] Guillaumin M, Mensink T, Verbeek J, Schmid C. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. IEEE Int’l Conf. on Computer Vision, 2010,30(2):309−316.

[9] Tsoumakas G, Katakis I, Vlahavas I. Mining multi-label data. In: Data Mining and Knowledge Discovery Handbook. Springer-Verlag, 2010. 667−685.

[10] Poelmans J, Ignatov DI, Kuznetsov SO, Duido D. Formal concept analysis in knowledge processing: A survey on applications. Expert Systems with Applications an Int’l Journal, 2013,40(16):6601−6623.

[11] Thomas JJ, Cook KA. A visual analytics agenda. IEEE Computer Graphics & Applications, 2006,26(1):10−13.

[12] Hong WX, Luan JM, Zhang T, Li SX, Yan EL. Knowledge discovery method based on partial sequence structure theory. Journal of Yanshan University, 2014,38(5):395−402 (in Chinese with English abstract).

[13] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proc. of the Int’l Conf. on Neural Information Processing Systems. Curran Associates Inc., 2012. 1097−1105.

[14] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstin M, Berg AC, Li FF. ImageNet large scale visual recognition challenge. Int’l Journal of Computer Vision, 2015,115(3):211−252.

[15] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proc. of the 3rd Int’l Conf. on Learning Representations. 2015.

[16] Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T. DeCAF: A deep convolutional activation feature for generic visual recognition. In: Proc. of the 31st Int’l Conf. on Machine Learning 2014. New York: ACM Press, 2014. 647−655.

[17] Razavian AS, Azizpour H, Sullivan J, Carlsson S. CNN features off-the-shelf: An astounding baseline for recognition. In: Proc. of the Computer Vision and Pattern Recognition Workshops. IEEE, 2014. 512−519.

[18] Tang Y, Wu X. Scene text detection and segmentation based on cascaded convolution networks. IEEE Trans. on Image Processing a Publication of the IEEE Signal Processing Society, 2017,26(3):1509−1520.

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

[20] Zhang S, Huang J, Huang Y, Yu Y, Li H, Metaxas DN. Automatic image annotation using group sparsity. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. New York: IEEE, 2010. 3312−3319.

[21] Hariharan B, Zelnik-Manor L, Vishwanathan SVN, Varma M. Large scale max-margin multi-label classification with priors. In: Proc. of the Int'l Conf. on Machine Learning. 2010. 423−430.

[22] Cui C, Ma J, Tao L, Wang X, Ren Z. Ranking-oriented nearest-neighbor based method for automatic image annotation. In: Proc. of the Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval 2013. New York: ACM Press, 2013. 957−960.

[23] Murthy VN, Maji S, Manmatha R. Automatic image annotation using deep learning representations. In: Proc. of the ACM Int'l Conf. on Multimedia Retrieval. ACM Press, 2015. 603−606.

[24] Li JC, Yuan C, Song Y. Multi-label image annotation based on convolution neural network. Computer Science, 2016,43(7):41−45 (in Chinese with English abstract).

**附中文参考文献:**

[3] 邱泽宇,方全,桑基韬,徐常胜.基于区域上下文感知的图像标注.计算机学报,2014,37(6):1390−1397.

[12] 洪文学,栾景民,张涛,李少雄,闫恩亮.基于偏序结构理论的知识发现.燕山大学学报,2014,38(5):395−402.

[24] 黎健成,袁春,宋友.基于卷积神经网络的多标签图像自动标注.计算机科学,2016,43(7):41−45.

**顾广华**(1979－),男,河南濮阳人,博士,教授,CCF 专业会员,主要研究领域为图像分类,图像理解,图像检索.

**李刚**(1979－),男,博士,副教授,主要研究领域为模式识别,形式概念分析.

**曹宇尧**(1992－),男,硕士,主要研究领域为图像分类,形式概念分析,图像检索.

**赵耀**(1967－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为跨媒体信息处理,基于内容的图像与视频检索.