一种结合相关性和多样性的图像标签推荐方法

崔超然 马 军
(山东大学计算机科学与技术学院 济南 250101)

摘 要 为了帮助用户高效地组织和检索图像资源,多数图像分享站点允许用户为图像添加标签.图像标签推荐系统旨在提供一组标签候选项来方便用户完成添加标签的过程.以往的图像标签推荐方法往往利用标签间的共现信息进行标签推荐.但是,由于忽略了图像的视觉内容信息和被推荐标签之间的多样性,以往方法的推荐结果常存在标签歧义和标签冗余的问题.为了解决上述问题,文中提出了一种新的图像标签推荐方法,该方法综合考虑了被推荐标签的相关性和多样性.首先,利用视觉语言模型,该方法分别计算标签与图像的相关性和标签之间的视觉距离.然后,基于上述计算,给出一个贪心搜索算法来找到能合理地平衡相关性和多样性的标签集合,将该集合作为最终的推荐.在 Flickr 数据集上的实验结果表明,该方法在准确率、主题覆盖率和 F_1 测度上均优于目前的代表性方法.

关键词 社会性标注;推荐算法;多样性;视觉语言模型
中图法分类号 TP391 **DOI 号** 10.3724/SP.J.1016.2013.00654

An Image Tag Recommendation Approach Combining Relevance with Diversity

CUI Chao-Ran MA Jun
(*School of Computer Science and Technology, Shandong University, Jinan* 250101)

Abstract To help users organize and retrieve the image resources efficiently, most image sharing sites allow users to annotate the images with tags. Image tag recommendation systems aim to provide a set of tag candidates to facilitate the tagging process done by users. Previous image tag recommendation methods are usually developed based on tag co-occurrence information. However, due to the neglect of the visual information associated with images and the semantic diversity among recommended tags, the recommendation results of previous methods often suffer from the problems of tag ambiguity and redundancy. To solve the above problems, this paper proposes a novel image tag recommendation approach, which considers both the relevance and diversity of the recommended tags. First, the approach employs the visual language model to calculate the relevance between a tag and an image, as well as the visual distance between two tags. Then, according to the above calculations, a greedy search algorithm is proposed to find a tag set as the final recommendation, which reaches a reasonable trade-off between the relevance and diversity. Experiments on Flickr data set show the proposed approach outperforms the state-of-the-art methods in terms of precision, topic coverage and F_1 value.

Keywords social tagging; recommendation algorithm; diversity; visual language model

1 引 言

伴随数字影像技术和互联网的迅速发展,网络中图像的数目呈现爆炸式的增长.为了有效地组织、管理如此大规模的图像资源,图像检索技术应运而生,并受到广泛研究.自 20 世纪 90 年代以来,基于内容的图像检索(Content-based Image Retrieval,

收稿日期:2012-06-27;最终修改稿收到日期:2012-11-08.本课题得到国家自然科学基金(61272240,60970047,61103151)、教育部博士点基金(20110131110028)、山东省自然科学基金(ZR2012FM037)资助.崔超然,男,1987 年生,博士研究生,中国计算机学会(CCF)会员,主要研究方向为信息检索、多媒体内容分析与理解、计算机视觉.E-mail:bruincui@gmail.com.马 军,男,1956 年生,博士,教授,博士生导师,主要研究领域为信息检索、社会网络、机器学习.

CBIR)不断发展，但是由于图像底层视觉特征与高层语义概念之间“语义鸿沟(semantic gap)”[1]的存在，CBIR 的检索性能难以令人满意. 因此，当前的商用图像搜索引擎(Google Images①、Bing Images②)仍是采用基于文本的图像检索(Text-based Image Retrieval，TBIR)方式，通过对图像的文本信息建立索引，利用成熟的文本检索算法，为用户提供图像检索服务，其检索性能依赖于图像相关文本的质量.

近年来，以 Flickr③ 为代表的图像分享网站蓬勃发展. 在 Flickr 中，用户可以为图像定义语义关键字，这些关键字称为图像标签(tag). 图像标签是用户对图像语义内容的描述，为 TBIR 提供了可靠的检索依据. 同时，图像分享网站往往根据标签对图像进行分类和组织，这使得用户乐于为图像添加标签，因为这样做可以使他人更容易地发现图像[2]. 因此，如何帮助用户快速准确地为图像添加标签便成为了一个十分重要的问题，而图像标签推荐系统正是解决这个问题的一个重要方法.

如图 1 所示，图像标签推荐是指在用户为图像添加标签的过程中，根据图像内容和初始标签集合(即用户已经添加的图像标签)，来发现一些新的标签候选项供用户选择. 图像标签推荐系统可以从以下 3 个方面为图像标注工作和后续的图像检索工作提供帮助[3]：(1)促使用户添加更多的标签. 用户在为图像添加标签的过程中，常常无法在短时间内想出大量可用的标签. 而标签推荐系统为用户提供图像标签候选项，减少了用户的工作量，使他们乐意于添加更多的标签. (2)帮助用户使用更加准确专业的标签. 统计显示，在 Flickr 中，经常被使用的标签数目只占全部标签数目的 5% 左右[4]. 许多对某个物体或场景描述更加准确专业的标签，往往由于其在日常生活中使用较少，而被用户忽略. 高质量的标签推荐系统可以根据图像内容，为用户提供更加准确专业的标签，丰富图像标注的词汇量. (3)减少噪音标签的出现. 噪音标签指的是一些拼写错误或者无意义的标签词. 标签推荐系统将添加标签的过程由打字转换为选择，从而有效地避免了噪音标签的出现.

以往的图像标签推荐方法[3,5]常利用标签间的共现信息(tag co-occurrence)，将与图像初始标签集合共现相似度高的标签推荐给用户. 图 2 展示了两例利用这种方法获得的推荐结果. 我们认为这种基

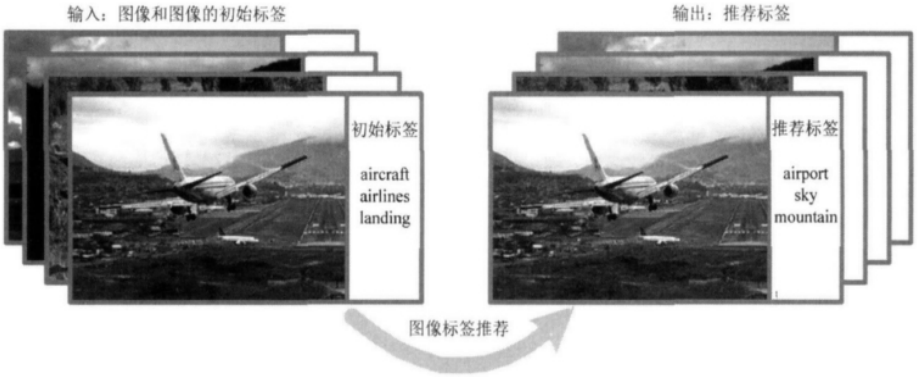

图 1　图像标签推荐的简单图示

图 2　基于标签间共现信息的推荐方法的结果

① http://images.google.com
② http://images.bing.com
③ http://www.flickr.com/

于标签共现信息的推荐方法存在以下两个问题：(1) 标签歧义问题. 由于没有考虑标签与图像内容的相关性，方法的性能易受歧义性标签的影响. 如图 2(a) 所示，由于初始标签‘apple’具有歧义性，在仅考虑标签间共现信息的条件下，推荐方法无法确定图像表达的真正含义，因此会导致一些与图像内容无关的标签(如‘mac’)被推荐. (2) 标签冗余问题. 被推荐的标签往往是初始标签的同义词、近义词，或是与初始标签描述相同概念的关键词，这些标签并不能为用户带来新的信息. 如图 2(b)所示，尽管推荐标签‘auto’，‘automobile’与图像内容具有较高的相关性，但由于初始标签中已经包含‘car’，使这些标签并不能为描述图像内容提供新的信息. 而用户在为这幅图像添加标签时，希望得到的是能从不同角度描述图像内容的标签，如标签‘tree’，‘sky’等.

为了解决上述问题，本文提出了一种新的图像标签推荐方法. 给定一幅图像和它的初始标签集合，该方法希望从剩余的标签中发现满足以下两个条件的一组标签：

(1) 相关性(relevance). 标签与图像所描述的内容有语义上的关联.

(2) 多样性(diversity). 标签能从不同的方面反映图像内容信息.

首先，利用视觉语言模型，分别计算标签与图像的相关性和标签间的视觉距离. 基于此，我们定义了一个标签集合的相关性和多样性. 本文提出的推荐算法的目标是找到一个指定大小的标签集合，使得该集合在相关性和多样性之间达到合理的平衡，将该标签集合推荐给用户. 真实数据集上的实验结果表明，本文的方法在准确率、主题覆盖率和 F_1 测度上均优于目前的代表性方法.

本文第 2 节回顾相关研究工作；第 3 节给出标签与图像的相关性和标签间视觉距离的计算方法；第 4 节论述本文提出的结合相关性与多样性的图像标签推荐算法；相关实验和分析在第 5 节给出；最后总结全文.

2 相关工作

近年来，图像标签推荐问题受到人们的广泛关注. Sigurbörnsson 等人[5]最早研究了 Flickr 中用户的标注行为，统计了用户在标注时常用的标签，并分析了这些标签的语义类别. 基于这些统计分析，他们进一步提出了一种利用标签间共现信息的图像标签推荐方法，选取若干个与图像的初始标签共现相似度高的标签作为推荐结果.

Wu 等人[3]提出了一种基于标签间多模态相关性的图像标签推荐方法. 他们将标签推荐问题转化为一个对标签的排序学习(Learning to Rank)问题. 计算两个标签之间的共现相似度和视觉相似度，并分别作为一种排序特征. 通过 Rankboost[6] 算法融合不同模态的排序特征，从而生成最终的排序函数. 利用该函数对标签排序，将排名高的若干标签推荐给用户.

Liu 等人[7]提出一种基于图像协同的标签推荐方法. 首先，估计标签与图像内容的相关性，根据相关性的高低对每幅图像的初始标签排序. 然后，给定一幅目标图像，在数据集中找到 K 幅与该图像视觉上接近的“邻居”图像，收集每幅“邻居”图像的前 m 个初始标签. 最后，根据标签在数据集中出现的频率，对收集到的 $m\times K$ 个标签排序，将排名高的若干标签推荐给用户.

Ames 等人[2]建立了 ZoneTag 系统. 当用户拍摄了一幅照片后，该系统根据照片的地理位置信息和用户以前使用过的标签，产生一些候选关键词，帮助用户标注该幅图像.

以上这些方法最基本的思路是根据标签与图像的相关性对标签进行排序，将排名高的标签推荐给用户. 但是这种思路没有考虑标签之间的相互关系，从而可能导致推荐结果缺乏多样性. 本文的研究正是从这一角度出发，综合考虑被推荐标签的相关性和多样性，从而改进图像标签推荐的性能.

3 标签的相关性和标签间的视觉距离

本节利用视觉语言模型，分别计算标签与图像的相关性和标签之间的视觉距离. 首先利用数据集来学习每个标签的视觉语言模型，通过该模型表达标签代表的视觉概念(第 3.1 节). 然后结合标签与初始标签集合的共现相似度和标签与图像的视觉相似度，来计算标签与图像的相关性(第 3.2 节). 最后通过两个标签的视觉语言模型间的 Jensen-Shannon 散度来计算它们之间的视觉距离(第 3.3 节).

3.1 标签的视觉语言模型

本文使用视觉语言模型(Visual Language Model, VLM)[8]来表达标签代表的视觉概念. VLM 是对传统的统计语言模型的扩展，用基于图像的视觉词袋(Bag-of-Visual-Words)[9]表示. VLM

认为图像中的视觉词在空间上是相互依赖的，相邻视觉词的排列遵守某种视觉语法，而一种视觉概念可以通过特定的视觉语法来表达.

给定一个标签 t，设数据集中含有标签 t 的图像的集合为 S_t. 图 3 展示了 t 的 VLM 的生成过程. 将 S_t 中的每幅图像分成很多相同大小且互不遮挡的图像块(patch)，在每个图像块上提取出相同维度的特征描述向量，并利用聚类方法将该特征编码为一个视觉词. VLM 假设图像中的视觉词是按照自左而右，自上而下的顺序生成的，因而一幅图像被表示为一个视觉词序列，每一个视觉词的出现条件依赖于它之前的视觉词. 标签 t 的 VLM 通过估计 S_t 中视觉词出现的条件概率分布来得到视觉词间的依赖关系，而这种依赖关系即反映了标签 t 代表的视觉概念.

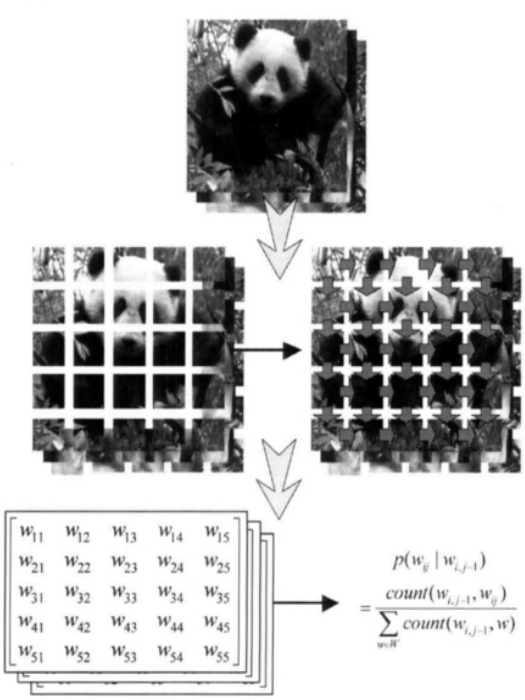

图 3　二元视觉语言模型的生成过程示意图

在估计条件概率时，考虑前面 N 个视觉词的模型，称为 N 元视觉语言模型(N-gram Visual Language Model). 出于性能和效率的综合考虑，本文采用二元视觉语言模型(Bigram Visual Language Model，Bigram VLM)，即认为当前视觉词的出现仅依赖于它左边的视觉词

$$p(w_{ij}\mid w_{11},w_{12},\cdots,w_{mn})=p(w_{ij}\mid w_{i,j-1}) \quad (1)$$

w_{ij} 表示图像中的第 i 行 j 列的视觉词，$w_{11},w_{12},\cdots,w_{mn}$ 是 w_{ij} 之前的视觉词序列.

估计 $p(w_{ij}\mid w_{i,j-1})$ 的最简单方法是极大似然估计(Maximum Likelihood Estimation，MLE)，令 $count(w_{i,j-1},w_{ij})$ 表示二元语法 $w_{i,j-1}w_{ij}$ 的出现次数，W 表示不同视觉词的集合，有

$$p(w_{ij}\mid w_{i,j-1})=\frac{count(w_{i,j-1},w_{ij})}{\sum_{w\in W} count(w_{i,j-1},w)} \quad (2)$$

由于数据稀疏性，训练集可能无法涵盖所有的二元语法，直接使用 MLE 会导致 $p(w_{ij}\mid w_{i,j-1})=0$ 的情况发生，因此需要进行平滑处理. 本文采用如下的平滑方法，该方法将概率回退技术和概率折扣技术相结合[10].

$$p(w_{ij}\mid w_{i,j-1})=\begin{cases}\beta\times p(w_{ij}), & count(w_{i,j-1},w_{ij})=0\\ d\times\dfrac{count(w_{i,j-1},w_{ij})}{\sum_{w\in W} count(w_{i,j-1},w)}, & count(w_{i,j-1},w_{ij})\neq 0\end{cases} \quad (3)$$

$$\beta=\frac{1-\sum_{count(w_{i,j-1},w)>0} p(w\mid w_{i,j-1})}{1-\sum_{count(w_{i,j-1},w)>0} p(w)} \quad (4)$$

$$d=1-\frac{n_1}{R} \quad (5)$$

式(3)中，若二元语法 $w_{i,j-1}w_{ij}$ 没有出现在训练集中，则利用概率回退技术通过一元语法 w_{ij} 的分布来计算它的条件概率，β 为回退因子. 若二元语法 $w_{i,j-1}w_{ij}$ 在训练集中出现，则利用概率折扣技术来降低它的条件概率估计值. d 是线性折扣因子，如式(5)所示，n_1 表示出现次数为 1 的视觉词的个数，R 表示不同视觉词的总数. 文献[11]对不同的概率折扣技术进行了比较，实验结果表明采用线性折扣的 VLM 可以取得更好的性能.

3.2　标签与图像的相关性

给定一幅图像 I 和它的初始标签集合 T_i，对一个标签 t，分别计算 t 与 T_i 的共现相似度和 t 与 I 的视觉相似度，两者共同衡量了 t 与 I 的相关性.

3.2.1　标签共现相似度计算

用户在为图像添加标签时，总是倾向于使用能够反应图像内容的标签，如果两个标签总是被同时添加给图像，则说明两个标签所代表的概念更有可能一同出现. 因此，若 t 与 T_i 有较高的共现相似度，则 t 更有可能会反映 I 的内容. 定义两个标签 t_i 和 t_j 的共现相似度为

$$r(t_i,t_j)=\frac{|t_i\cap t_j|}{|t_i|} \quad (6)$$

$|t_i|$ 表示数据集中包含标签 t_i 的图像数目. 直观地说，$r(t_i,t_j)$ 代表了当图像获得标签 t_i 后仍能获得标签 t_j 的可能性. 基于该定义，标签 t 与初始标签集合

T_i的共现相似度$C(T_i,t)$被定义为 t 与每一个初始标签的共现相似度之和

$$C(T_i,t)=\sum_{t_i\in T_i} f(r(t_i,t)) \tag{7}$$

$f(\cdot)$是一个单调递增的平滑函数.

3.2.2 标签视觉相似度计算

在计算标签与图像的相关性时,若仅仅考虑共现相似度,结果会受到标签歧义问题的影响. 例如,标签'sun'和'java'具有较高的共现相似度,并且'sun'是一幅图像的初始标签,但若该图像描绘的是有关日出的场景,那么标签'java'显然与图像内容无关. 因此,为了避免受到标签歧义问题的影响,需要计算标签与图像内容的视觉相似度.

基于 3.1 节介绍的知识,图像 I 可以被表示为一个视觉词序列 $I=w_{11},w_{12},\cdots,w_{pq}$,计算标签 t 的 Bigram VLM 产生这个视觉词序列的似然:

$$\begin{aligned} p(I\mid t)&=\prod_{w_{ij}\in I} p(w_{ij}\mid w_{11},w_{12},\cdots,w_{mn},t)\\ &=\prod_{w_{ij}\in I} p(w_{ij}\mid w_{i,j-1},t) \end{aligned} \tag{8}$$

$p(w_{ij}\mid w_{11},w_{12},\cdots,w_{mn},t)$表示在 t 的 VLM 中,w_{ij}依赖于它的前序视觉词出现的条件概率. 直观地说,$p(I\mid t)$反映了依据标签 t 代表的视觉概念主题,"创作"出图像 I 的可能性. 本文根据 $p(I\mid t)$将标签 t 与图像 I 的视觉相似度 $V(I,t)$定义为

$$V(I,t)=g(p(I\mid t)) \tag{9}$$

$g(\cdot)$是一个单调递增的平滑函数.

结合上述两种相似度,我们最终将标签 t 与 I 的相关性 $R(I,t)$定义为

$$R(I,t)=\eta C(T_i,t)+(1-\eta)V(I,t) \tag{10}$$

参数 $\eta(0<\eta<1)$是调节两种相似度权重的系数,我们将在实验部分讨论 η 的取值变化对结果的影响.

3.3 标签间的视觉距离

以往的图像标签推荐方法仅仅考虑被推荐标签与图像的相关性,而忽略了它们之间的相互关系,使得被推荐的标签往往代表相同或相近的概念. 而一幅图像常包含多种概念(如不同的物体等),利用以往方法得到的推荐结果可能无法全面地反映图像内容信息.

本文提出的图像标签推荐算法希望被推荐的标签能从不同的方面反映图像内容信息,即推荐结果具有较好的多样性. 为此,我们首先计算两个标签之间的视觉距离. 由 3.1 节可知,标签 t 的 VLM 估计了包含 t 的所有图像中视觉词出现的条件概率分布,这个分布表达了视觉词在空间上的相互依赖关系,能反映 t 代表的视觉概念. 因此,我们可以通过计算两个标签的视觉词分布的 Jansen-Shannon 散度,来度量两者间的视觉距离. 给定两个标签 t_i,t_j,将它们之间的视觉距离 $D(t_i,t_j)$定义为

$$D(t_i,t_j)=\frac{1}{2}l(KL(d_i\parallel d_j)+KL(d_j\parallel d_i)),$$

$$KL(d_i\parallel d_j)=\sum_{w_m,w_n\in W} p(w_m\mid w_n,t_i)\log\frac{p(w_m\mid w_n,t_i)}{p(w_m\mid w_n,t_j)} \tag{11}$$

$p(w_m\mid w_n,t_i)$表示在 t_i 的 Bigram VLM 中,视觉词 w_m依赖于视觉词 w_n出现的条件概率. $l(\cdot)$是一个单调递增的平滑函数. 相比其它的距离测度[12],这种视觉距离的计算方式能有效地反映两个标签代表的视觉概念间的差异[3].

4 图像标签推荐算法

结合上述的标签与图像的相关性和标签间的视觉距离,本节介绍结合相关性和多样性的图像标签推荐算法. 我们首先定义一个标签集合的相关性和多样性,然后利用贪心搜索算法找到能合理地平衡相关性和多样性的标签集合,将该集合作为最终的推荐结果.

4.1 标签集合的相关性与多样性

在以往的图像标签推荐算法中,标签推荐问题往往被转换为一个根据标签与图像的相关性对标签进行排序的问题,算法将排名较高的标签推荐给用户. 本文提出的图像标签推荐算法考虑了被推荐标签之间的相互关系,因而算法的目标是推荐一个指定大小的标签集合.

结合前一节介绍的内容,给定目标图像 I,对于一个大小为 N 的候选标签集合 S_T,将 S_T中标签与 I 的平均相关性定义为衡量 S_T的相关性 $Rel(S_T)$的指标.

$$Rel(S_T)=\frac{\sum_{t\in S_T} R(I,t)}{N} \tag{12}$$

$R(I,t)$由式(10)定义. 将 S_T中两个标签间的平均视觉距离定义为衡量 S_T的多样性 $Div(S_T)$的指标.

$$\begin{aligned} Div(S_T)&=\frac{\sum_{t_i,t_j\in S_T,t_i\neq t_j} D(t_i,t_j)}{C_N^2}\\ C_N^2&=\frac{N(N-1)}{2} \end{aligned} \tag{13}$$

$D(t_i,t_j)$由式(11)定义. 进一步地,以两个指标的加权

和作为 S_T 相关性和多样性的平衡程度得分 $F(S_T)$，有

$$F(S_T)=\lambda Rel(S_T)+(1-\lambda)Div(S_T) \quad (14)$$

参数 $\lambda(0<\lambda<1)$ 用来控制在计算得分时相关性和多样性所占的比重，我们将在本文的实验部分对 λ 取值的影响进行讨论.

4.2　推荐算法描述及时间复杂度分析

在进行图像标签推荐时，本文提出的算法希望找到一个能合理地平衡相关性与多样性的标签集合. 给定目标图像 I 和它的初始标签集合 T_i，算法在剩余标签中，选取相关性与多样性平衡程度得分最高的一个标签集合 S_T^*，将它推荐给用户，即

$$S_T^*=\arg\max_{S_T} F(S_T),\ S_T\subset T\setminus T_i \quad (15)$$

T 代表数据集中全部标签的集合.

式(15)的求解是一个典型的非线性整数规划问题[13]，属于 NP-Hard 类的最优化组合问题，不存在多项式时间内的精确求解算法. 因此，我们利用一个贪心搜索算法来求解该问题的近似最优解，求解过程如算法 1 所示. 初始时，将 S_T^* 初始化为空集(算法第 1 行). 首先，算法在除 T_i 外的剩余标签中，找到与图像的相关性最大的一个标签 t_i，将 t_i 作为第一个标签加入到 S_T^* 中(算法第 2 行～3 行). 然后，算法迭代地寻找剩余的 $N-1$ 个标签. 每轮迭代在除 T_i 和 S_T^* 外的剩余标签中找到标签 t_r，t_r 是加入当前的 S_T^* 后使其得分最大的一个标签，将 t_r 加入到 S_T^* 中(算法第 4 行～7 行). 最终，集合 S_T^* 包含 N 个标签，返回 S_T^* 作为推荐结果.

在推荐算法开始前，首先离线地训练出数据集中各标签的 VLM 并计算任意两个标签间的共现相似度和视觉距离. 算法 1 的时间复杂度为 $O(MN^2)$. 其中 N 为希望得到的推荐标签的数量，M 为数据集中全部标签的数量. 在实际计算时，N 的值一般较小(实验中取 $N=10$)，故算法的运行时间主要依赖于数据集中全部标签的数量. 一种有效的提高运行效率的方法是，在计算时首先将全部标签按照其与图像的相关性排序，然后按照性能需要，选择排序靠前的若干标签继续进行算法 1 中的计算.

算法 1.　基于贪心搜索的标签推荐算法.

输入：训练集中的全部标签 T，一幅图像 I，I 的初始标签集合 T_i，希望得到的推荐标签的个数 N

输出：大小为 N 的推荐标签集合 S_T^*

1. 初始化 $S_T^*=\varnothing$；
2. 从 $T\setminus T_i$ 中选出标签 t_i，t_i 满足
$$t_i=\arg\max_{t_c} R(I,t_c);$$
3. $S_T^*=S_T^*\cup\{t_i\}$；
4. for $i=2$ to N do
5. 从 $T\setminus\{T_i\cup S_T^*\}$ 中选出标签 t_r，t_r 满足
$$t_r=\arg\max_{t_c}\left[\lambda R(I,t_c)+\frac{1-\lambda}{|S_T^*|}\sum_{t_s\in S_T^*}D(t_s,t_c)\right];$$
6. $S_T^*=S_T^*\cup\{t_r\}$；
7. end for
8. return S_T^*.

5　实验结果与分析

5.1　实验设置

为了验证本文提出的方法的有效性，我们采用 NUS-WIDE 数据集[14]作为实验数据集，该数据集来自 Flickr 中约 5000 名用户提供的 269648 幅图像和 425059 个不同的标签，图像内容包含丰富多样的物体场景，反映了 Web 中海量图像的真实情况. 由于 NUS-WIDE 数据集包含了大量的噪音标签，因此我们首先对数据集中的标签进行了过滤操作. 去除未被 WordNet 索引或者出现次数小于 50 的标签，并且对剩余标签进行词干化，最终保留了 4377 个不同的标签.

图 4 给出数据集中各标签出现的次数统计，从中可以看出：它们近似呈现长尾分布特性[5]. 其中，出现次数大于 5000 的标签不足 1%，这些标签多代表较为常见通用的概念(如‘nature’，‘color’等). 出现次数大于 500 次的标签仅有 20%，超过一半的标签的出现次数小于 100. 许多出现次数较少的标签常可以准确地描绘某个特定场景或物体(如‘purple’‘puss’等).

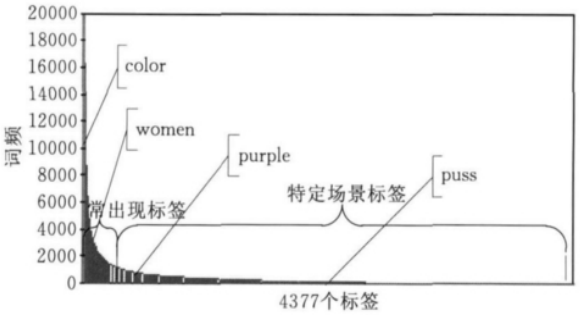

图 4　数据集中标签出现次数统计

实验中，为了减小图像尺寸变化对结果的影响，所有图像均被调整为 320×320 像素的尺寸. 对于每一幅图像，将其均匀地划分为多个 8×8 像素的图像块，在每个图像块上提取出 8 维的纹理梯度直方图[15]作为特征描述向量. 这种特征具有低维度和尺度不变性的特点，利用该特征的 VLM 可以取得更好的性能[16]. 在建立视觉词典时，词典的大小设为 300.

分别随机地选取 500 幅图像作为验证集和测试

集,验证集用来确定方法中参数的最佳取值,测试集用来评估方法的性能.利用剩余的全部图像训练标签的 VLM 并计算标签间的共现相似度和视觉距离.式(7)、(11)中的平滑函数被定义为标准 sigmoid 函数;式(9)中的平滑函数被定义为对数线性平滑函数[17].

对于验证集和测试集中的每一幅图像,不同的推荐方法均产生 10 个推荐标签,3 名志愿者独立的对标签的相关性进行判断,最后,使用投票的方法确定标签是否与图像内容相关.实验中,我们统计了志愿者两两之间的 Cohen's kappa 统计量[18],计算得到三人 Cohen's kappa 系数的平均值为 0.77,超过一致性判优边界(0.75),这说明志愿者在判断推荐标签的相关性时取得了较好的一致性,证明了本实验人工评判的可信性[19].

5.2 评价指标

为了评价不同图像标签推荐方法的性能,我们采用以下 3 个评价测度来衡量一幅图像的推荐结果的质量.

(1) **准确率**(Precision).令 TP_i 表示推荐结果中相关标签的数目,N 表示推荐标签的总数,则推荐结果的准确率为

$$Precision=\frac{TP_i}{N} \tag{16}$$

(2) **主题覆盖率**(Topic Coverage, T-coverage).类似于文献[20]中的 $S\text{-}recall$ 测度,主题覆盖率衡量了推荐结果中相关标签的语义多样性,其值为结果中相关标签能覆盖的语义主题的比例.

$$T\text{-}coverage=\frac{|\bigcup_{i=1}^{K}topic(t_i)|}{N_t} \tag{17}$$

式中 K 表示结果中相关标签的个数,t_i 表示第 i 个相关标签,$topic(t_i)$ 是 t_i 对应的语义主题,N_t 是与图像相关的语义主题的总数.为了确定 t_i 和 N_t,我们采用 pooling 技术[21],将图像的初始标签、不同推荐方法推荐的相关标签集合在一起.根据标签的语义,志愿者手工地对这些标签进行聚类,将标签的类别作为它的语义主题,在聚类时不限制类别的个数.

(3) F_1 值.结合以上两种测度来综合评价

$$F_1=\frac{2\times Precision\times T\text{-}coverage}{Precision+T\text{-}coverage} \tag{18}$$

我们为出现在验证集和测试集中的每幅图像的推荐结果计算上述 3 个测度,最终把得到的结果对集合中的所有图像取平均,以此作为评价指标.

5.3 实验结果分析

5.3.1 参数设置对方法性能的影响

下面,我们将通过实验来考察方法中涉及到的两个参数对性能的影响,它们分别是式(10)中的参数 η 和式(14)中的参数 λ.

在计算标签与图像的相关性时,η 用于调节共现相似度和视觉相似度的权重.首先我们设 $\lambda=0.5$,然后分别在 η 取不同值时,观察方法在验证集上的性能,图 5 展示了实验结果.从图中可以看出,当 $\eta=0.5$ 或 0.6 时,方法的性能最好.这说明在计算标签与图像的相关性时,应较为平衡地分配共现相似度和视觉相似度所占的比重.实验中,当 λ 在 $[0.4,0.8]$ 范围内变化时,η 的最佳取值并未受到明显影响.因此,在后面的实验中,我们设 $\eta=0.5$.

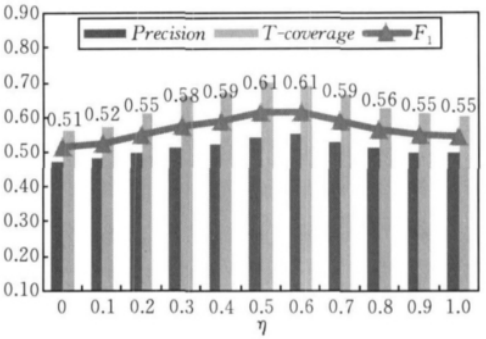

图 5 η 变化对方法性能的影响

在式(14)中,λ 用于调节相关性和多样性所占的比重.为了清楚地了解 λ 的影响,我们同样在 λ 取不同值时,计算方法在验证集上取得的实验结果.直观地说,如果 λ 太小,则推荐结果中容易引入无关标签;相反,如果 λ 太大,那么推荐结果中有可能出现语义冗余的标签.在这两种情况下,方法均无法获得最好的性能.λ 的变化对方法性能的影响如图 6 所示,从中可以得到相同的结论.当 $\lambda=0.6$ 或 0.7 时,方法的性能最好,之所以会这样,主要原因是我们在评价方法时只考虑相关标签的多样性,方法会在保证推荐结果的相关性较高的情况下得到最好性能.在后面的实验中,我们设 $\lambda=0.6$.

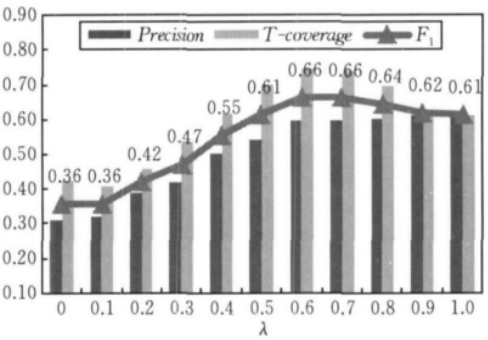

图 6 λ 变化对方法性能的影响

5.3.2 相关方法的比较与分析

在本组实验中,我们将通过与其它几种方法的

比较，验证本文提出的结合相关性与多样性的图像标签推荐方法的有效性. 这里，涉及到的方法包括：TC(利用标签共现信息的图像标签推荐)[5]、MRR(基于多模态相关性的图像标签推荐)[3]、CR(基于图像协同的标签推荐)[7]、RD(本文提出的结合相关性与多样性的图像标签推荐).

图 7 显示了 4 种方法在测试集上的实验结果. 可以看到，MRR 获得了最高的准确率，该方法的优势主要在于考虑了图像与标签间多模态的相关性，并利用 Rankboost 算法将它们融合在一起. RD 在准确率上比 MRR 有略微的降低，但较 TC 仍提高了 7%，这主要是因为 RD 在计算相关性时结合了共现相似度和视觉相似度. CR 的准确率较低，可能是因为数据集中的图像丰富多样，无法较准确地找出语义相似的图像. 相比于其它 3 种方法，RD 在主题覆盖度上取得了最好的性能，分别超出了 15%，10%和 13%，这证明 RD 可以更好地确保推荐结果的多样性. 综上可以看到，本文提出的方法较好地平衡了推荐结果的相关性和多样性，在 4 种方法中也取得了最高的 F_1 值.

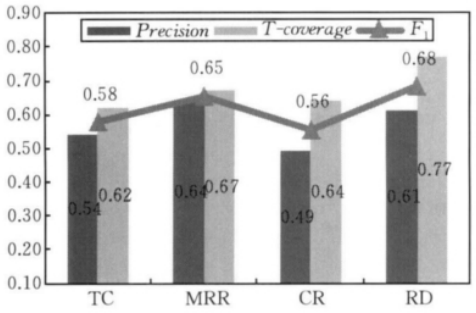

图 7　测试集上的性能比较

为了更加深入地观察不同方法产生的推荐标签，我们统计了各方法的推荐结果中不同的相关标签的数目，并且计算最常出现的 50 个相关标签的出现次数占总数的比例，表 1 展示了比较结果. 可以看到，相比于其它 3 种方法，RD 使用了更加丰富的词汇，推荐结果中不同的相关标签的数目是其它方法的近两倍. 在利用 TC 得到的相关标签中，最常出现的 50 个标签的出现次数约占总数的 60%，这说明 TC 倾向于集中使用少量标签. 这些标签往往代表一些较为通用的概念(如'nature'、'landscape')，尽管许多图像都与这些概念相关，但由于图像内容丰富多样，这些标签往往无法精确地描绘出图像反映的具体信息. 而在 RD 的结果中，相关标签分布更为均匀，最常出现的 50 个标签的出现次数仅占总数的 15%，是 4 种方法中最低的.

表 1　不同方法推荐结果的数据统计

方法	不同相关标签数目/个	最常出现的50个标签所占比例/%
TC	320	60.34
MRR	359	43.13
CR	287	34.59
RD	**667**	**15.64**

图 8 展示了 RD 的推荐结果. 对于每幅图像，图中列出了它的初始标签和推荐标签. 可以看到，一方

图 8　本文方法的推荐结果

面,推荐标签可以更加具体地刻画初始标签代表的概念,如图 8(a)中的推荐标签‘dancer’就是对初始标签‘girl’、‘people’的进一步说明;另一方面,当初始标签较少时,推荐标签可以表达初始标签未能反映的物体或场景,如图 8(f)中的推荐标签‘sky’、‘grass’.综上说明,RD 方法的推荐结果力图从与初始标签不同的角度,基于图像的视觉特征与初始标签的共现概率以及语义的多样性等方面,为用户提供新的图像标注选择.

6 结 语

本文对图像标签推荐技术进行了研究,提出了一种结合相关性与多样性的图像标签推荐方法,解决了传统方法中标签歧义与标签冗余的问题.方法定义了一个标签集合的相关性和多样性,并选取一个能合理地平衡相关性和多样性的标签集合推荐给用户.实验结果表明,本文提出的方法一方面提高了推荐结果与图像的相关性,另一方面使推荐结果能较全面地反映图像内容.

我们未来拟开展的研究包括:(1)将视觉主题模型与方法中的视觉语言模型相结合以进一步提升推荐性能;(2)将该方法应用到其它类型的社会性共享资源中.

致 谢 在此,我们向对本文工作给予建议、帮助的老师和同学,尤其是山东大学信息检索研究组的高帅同学和韩晓晖同学,表示感谢!

参 考 文 献

[1] Smeulders A W M, Worring M, Santini S, Gupta A, Jain R. Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(12): 1349-1380

[2] Ames M, Naaman M. Why we tag: Motivations for annotation in mobile and online media//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. San Jose, USA, 2007: 971-980

[3] Wu L, Yang L, Yu N, Hua X S. Learning to tag//Proceedings of the 18th International Conference on World Wide Web. Madrid, Spain, 2009: 361-370

[4] Akbas E, Yarman Vural F T. Automatic image annotation by ensemble of visual descriptors//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, 2007: 1-8

[5] Sigurbjörnsson B, van Zwol R. Flickr tag recommendation based on collective knowledge//Proceedings of the 17th International Conference on World Wide Web. Beijing, China, 2008: 327-336

[6] Freund Y, Iyer R, Schapire R E, Singer Y. An efficient boosting algorithm for combining preferences. The Journal of Machine Learning Research, 2003, 4: 933-969

[7] Liu D, Hua X S, Yang L, Wang M, Zhang H J. Tag ranking//Proceedings of the 18th International Conference on World Wide Web. Madrid, Spain, 2009: 351-360

[8] Wu L, Li M, Li Z, Ma W Y, Yu N. Visual language modeling for image classification//Proceedings of the international workshop on multimedia information retrieval. Augsburg, Germany, 2007: 115-124

[9] Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos//Proceedings of the 9th IEEE International Conference on Computer Vision. Nice, France, 2003: 1470-1477

[10] Katz S. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Transactions on Acoustics, Speech and Signal Processing, 1987, 35(3): 400-401

[11] Tirilly P, Claveau V, Gros P. Language modeling for bag-of-visual words image categorization//Proceedings of the 2008 international conference on content-based image and video retrieval. Niagara Falls, Canada, 2008: 249-258

[12] Koloniari G, Petrakis Y, Pitoura E, Tsotsos T. Query workload-aware overlay construction using histograms//Proceedings of the 14th ACM international conference on information and knowledge management. Bremen, Germany, 2005: 640-647

[13] Boyd S, Vandenberghe L. Convex Optimization. Cambridge, England: Cambridge University Press, 2004

[14] Chua T S, Tang J, Hong R, Li H, Luo Z, Zheng Y. NUS-WIDE: A real-world web image database from National University of Singapore//Proceedings of the ACM International Conference on Image and Video Retrieval. Santorini, Greece, 2009: 1-9

[15] Dalal N, Triggs B. Histograms of oriented gradients for human detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2005: 886-893

[16] Wu L, Hua X S, Yu N, Ma W Y, Li S. Flickr distance//Proceedings of the 16th ACM international conference on Multimedia. Vancouver, British Columbia, 2008: 31-40

[17] Chen C, Mangasarian O L. A class of smoothing functions for nonlinear and mixed complementarity problems. Computational Optimization and Applications, 1996, 5(2): 97-138

[18] John S U. Diversity of decision-making models and the measurement of interrater agreement. Psychological Bulletin, 1987, 101(1): 140-146

[19] Landis J R, Koch G G. The measurement of observer agreement for categorical data. Biometrics, 1977, 33(1): 159-174

[20] Zhai C X, Cohen W W, Lafferty J. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval//Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. Toronto, Canada, 2003: 10-17

[21] Manning C D, Raghavan P, Schtze H. Introduction to Information Retrieval. Cambridge, England: Cambridge University Press, 2008

CUI Chao-Ran, born in 1987, Ph.D. candidate. His research interests include information retrieval, multimedia content analysis and understanding, and computer vision.

MA Jun, born in 1956, professor, Ph. D. supervisor. His research interests include information retrieval, social network and machine learning.

Background

With the rapid development of Internet and Web 2.0 technology, a large number of community contributed images have been produced and shared on the Web. Quite a few representative Web 2.0 websites, such as Flickr and Pinterest, not only provide users interfaces of image sharing, but also allow users to collaboratively describe the resources with their own tags through social tagging services. From the perspective of critical web applications such as keyword-based image search engines, image tags are indispensable for image indexing and retrieval. Moreover, recent user studies reveal that users are willing to tag their photos with the motivation to make them better accessible to the general public. In this paper, the authors focus on helping users in the tagging phase by developing a novel image tag recommendation approach.

Image tag recommendation is to recommend more tags for an image based on the existing clues, including tags, surrounding text and visual content information. Previous image tag recommendation approaches are usually performed by ranking the related tags based on the tag co-occurrence information. However, due to the neglect of the visual information associated with images and the semantic diversity among recommended tags, the recommendation results from previous approaches often suffer from the problems of tag ambiguity and redundancy. In light of the two problems, the authors propose an image tag recommendation approach, which considers both the relevance and diversity of the recommendation results. To this end, they employ the visual language model to calculate the visual relevance between a tag and an image, as well as the visual distance between two tags. The goal of the approach is to find a tag set as the final recommendation, which reaches a trade-off between the relevance and diversity. A greedy search algorithm is proposed to achieve the objective.

This work is supported by the National Natural Science Foundation of China (Nos. 61272240, 60970047, 61103151), the Doctoral Fund of Ministry of Education of China (20110131110028). These projects aim to study the social media information processing oriented social networks under the application backgrounds of multi-document summarization and image annotation.

The group is dedicated to research of new theories, algorithms and systems for information retrieval, multimedia content analysis and understanding, and social media information processing. Related papers have been published in reputable domestic and international journals and conferences.