基于标签集相关性学习的大规模网络图像在线标注

田 枫 [1, 2] 沈旭昆 [1]

摘 要 传统的网络图像标注方法忽视了标签集整体相关性对标注结果的影响, 导致标签集整体相关性缺乏和语义冗余. 为了解决上述问题, 提出了一种基于标签集相关性学习的大规模网络图像在线语义标注方法. 给出了标签集对图像相关性和标签集内部相关性的概率估计算法, 将上述约束形成一个优化问题, 采用贪心搜索策略获取近似最优解, 找到能合理地平衡上述因素的标签集, 并针对大规模图像集和概念集进行了优化. 真实环境下大规模网络图像集上的测试表明, 相比于目前的代表性网络图像标注方法, 该方法获得的标签集能够更好的描述图像语义, 性能提升明显.

关键词 网络图像标注, 图像语义标注, 标签集相关性, 标签相关性学习

引用格式 田枫, 沈旭昆. 基于标签集相关性学习的大规模网络图像在线标注. 自动化学报, 2014, **40**(8): 1635−1643

DOI 10.3724/SP.J.1004.2014.01635

Large Scale Web Image Online Annotation by Learning Label Set Relevance

TIAN Feng[1, 2] SHEN Xu-Kun[1]

Abstract Traditional web image annotation methods neglect the relevance of the assigned label set as a whole, resulting in the label relevance deficiency and redundancy. To solve the above problems, a novel web image annotation method by learning the label set relevance is proposed, which considers both the relevance of label set to image and the label set internal correlation. Measures that can estimate the above factors are designed, and both the constraints are formulated into a joint framework. Meanwhile, an effective greedy search algorithm is proposed for an approximate optimal label set, which reaches a reasonable trade-off between the relevance of label set to image and internal correlation, and makes the framework more applicable to the data set that contains the large scale concept and images. Experiments on real world web image data sets demonstrate the general applicability of our algorithm. In comparison to the state-of-the-art methods, the proposed approach yields better performance.

Key words Web image annotation, image semantic annotation, label set relevance, label relevance learning

Citation Tian Feng, Shen Xu-Kun. Large scale web image online annotation by learning label set relevance. *Acta Automatica Sinica*, 2014, **40**(8): 1635−1643

网络图像数量的激增, 对图像自动标注的规模化与标注精度提出了更高的要求. 现有标注方法可以简单分为基于模型学习的方法和基于实例检索的方法[1]. 基于模型学习的方法具有较高的标注性能, 但是模型训练的计算复杂性较高, 只适用于处理受限环境下的小规模语义概念集合, 不具备在真实环境下进行大规模语义标注的能力. 基于实例检索的方法将图像标注问题看作图像检索问题, 依据相似图像具有相似语义的前提, 在相似图像间进行标签传播, 是基于数据驱动的无模型方法. 其突破了基于模型学习的方法对训练集合与训练方法的过度依赖, 避免了复杂的参数学习的过程, 可以较灵活地扩展到庞大的互联网数据集中, 适用于大规模数据集的在线标注[2]. Wang 等进行基于文本和内容的检索, 得到语义近似图像集合, 并依据标签进行结果聚类, 得到最终标签[3]. Liu 等依据概率密度估计标签与视觉内容的初始相关度, 并通过在标签相似图上的随机游走获得标签的最终相关分数[4]. Li 等依据邻域内视觉相似样本, 获得近邻标签列表, 然后对标签进行投票, 估计其和目标图像的相关性[5]. 上述基于数据驱动的图像标注方法具有规模化图像标注的能力, 但是相对于受限环境下基于模型学习的方法, 其标注准确率较低. 因此, 真实世界环境下的图像语义标注研究面临下述两个方面的问题: 1) 标注方法的大规模数据处理能力; 2) 标注性能有待提高. 现有研

收稿日期 2012-12-03 录用日期 2013-08-13
Manuscript received December 3, 2012; accepted August 13, 2013
国家高技术研究发展计划 (863 计划) (2009AA012103), 国家自然科学基金 (60533070), 东北石油大学青年科学基金 (2013NQ120) 资助
Supported by National High Technology Research and Development Program of China (863 Program) (2009AA012103), National Natural Science Foundation of China (60533070), and Youth Foundation of Northeast Petroleum University (2013NQ120)
本文责任编委 戴琼海
Recommended by Associate Editor DAI Qiong-Hai
1. 北京航空航天大学虚拟现实技术与系统国家重点实验室 北京 100191 2. 东北石油大学计算机与信息技术学院 大庆 163318
1. State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191 2. School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318

究表明, 标签之间的相关性可以为图像标注过程提供很好的辅助.

Jin 等利用多种词汇语义相似度量方法改善标注结果[6], 但是基于词典的方法无法处理词典不涵盖的标签, 与数据集无关, 对于特定图像可能并不适用. Wang 等提出依据视觉特征构建样本子图, 依据标签相关信息构建标签子图, 然后通过构造二部图集成两部分信息, 在二步图上进行随机游走计算语义组对图像相关性和语义组间相关性[7]. Yang 等提出了一个集成标签相关性和视觉相似性的框架, 其首先依据视觉相似度构造邻接图, 并将图上的传播过程所反映的标签之间相关性嵌入到多标签线性分类器中, 使得图上的标签传播与线性分类器同步学习, 并通过得到的决策面对图像进行分类标注, 实验表明该方法能够较好地揭示标签关联性, 其性能也优于传统的基于图的半监督标注和线性多标签分类的标注方法, 但是上述方法计算复杂, 无法应用到真实环境下的大规模数据集上[8]. 综上所述, 标签相关性能够有效提升图像标注性能, 但是大规模网络图像在线标注的应用背景要求算法兼具有效性和执行效率. 图 1 给出了标签集相关性学习示意图.

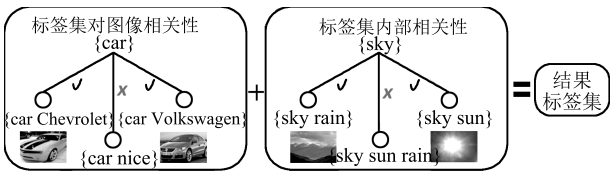


图 1 标签集相关性学习示意图

Fig. 1 Illustration of label set relevance learning

标签 “car” 能够有效描述图 1 中的第 1 幅图像, 相对而言, 标签 “nice”, “photo”, “cannon” 的相关度较低, 而具有较强相关性的标签 “Chevrolet” 却并没有出现在标签集中, 因此标签集 {“car”, “nice”, “photo”, “cannon”} 整体上是相关性缺乏的. 标签集 {“apple”, “iphone”, “fruit”} 可能赋给一幅具有 “水果” 语义的图像, 因为标签 “apple” 与标签 “fruit” 共现频率较高, 标签 “apple” 与标签 “iphone” 共现频率较高, 但是三个标签构成的标签集整体使得其具有歧义性, 无法确定图像内容是电子产品, 还是水果, 因此 {“apple”, “iphone”, “fruit”} 是语义不一致的. 类似的例子是, 标签集 {“sky”, “sun”, “rain”} 可能赋给图 1 中的第 3 幅图像, 因为标签 “sky” 与标签 “sun” 的相似度较高, 标签 “sky” 与标签 “rain” 相似度较高, 但是标签 “rain” 对标签集 {“sky”, “sun”} 相关性较低, 因此标签集 {“sky”, “sun”, “rain”} 是语义不相关的. 这些问题只有当我们从标签集整体的角度来考虑的时候才会出现. 上述问题导致现有的数据驱动的标注方法在实际应用中性能有待提高, 而且, 这些问题的存在, 也使得用户对标注结果的直观感受不好. 由此, 本文提出一种基于标签集相关性学习的网络图像语义在线标注方法, 不同于传统方法, 本文考虑的是标签集整体的相关性. 该相关性包含两方面的含义: 1) 标签集对图像的相关性. 2) 标签集内部标签的相关性. 利用散度变化量约束标签集对图像相关性, 依据标签相关性分析约束候选标签集的内部相关性, 进而将两方面约束形成一个优化问题, 获得的标签集能够更好地描述图像语义. 但是该问题的求解是一个非线性整数规划问题, 属于 NP-Hard 类的最优化组合问题, 候选标签集将随着标签规模呈指数增长, 不存在多项式时间内的精确求解算法. 为了提高该方法在大规模网络数据集上的标注效率, 我们针对大规模图像和概念集合进行了优化, 利用贪心搜索算法来求解该问题的近似最优解. 第 1 节首先介绍了标签集相关性学习框架, 着重介绍了标签集对图像相关性概率估计算法和标签集内部相关性概率估计算法, 第 2 节对算法复杂性进行了分析, 并给出了针对大规模数据集的优化方法, 第 3 节对算法实际运行中的参数调节、性能与效率进行了分析.

1 标签集相关性学习框架

令 X 表示图像集合, W 表示标签集合, W_q 表示包含 q 个标签的候选标签集合. 给定图像 $x \in X$, 定义 W_q 对 x 的相关度为 $F(W_q, x)$, 优化目标为

$$\begin{aligned}\widehat{W_q} = {} & \arg\max F(W_q, x) = \\ & \arg\max_{W_q \subset W} (\alpha \cdot r_{\text{set-to-image}}(W_q, x) + \\ & (1-\alpha) \cdot c_{\text{set}}(W_q)) \end{aligned} \tag{1}$$

其中, $r_{\text{set-to-image}}(W_q, x)$ 表示标签集 W_q 对目标图像 x 的相关性, $c_{\text{set}}(W_q)$ 表示标签集 W_q 的内部相关性, $\widehat{W_q}$ 为 x 的理想标签集, α 为平衡因子. 由于候选标签集随着标签集规模呈指数增长, 式 (1) 求解的复杂度过高, 所以我们采用一种贪心搜索策略获得式 (1) 的近似解, 在第 q 次迭代的时候, 选择和 x 最相关, 且和 W_{q-1} 最相关的标签 w, 即

$$\begin{aligned} & r_{\text{set-to-image}}(W_q, x) \approx \\ & \quad r_{\text{set-to-image}}(W_{q-1}, x) + r(w, x) \end{aligned} \tag{2}$$

式 (2) 中, $W_q = W_{q-1} \bigcup \{w\}$, $r(w, x)$ 表示标签 w 对图像 x 的相关性. 同理, W_q 内部相关性近似为

$$c_{\text{set}}(W_q) \approx c_{\text{set}}(W_{q-1}) + c(w, W_{q-1}) \tag{3}$$

式 (3) 中, $c_{\text{set}}(W_{q-1})$ 表示 W_{q-1} 的内部相关性, $c(w, W_{q-1})$ 表示标签 w 对标签集 W_{q-1} 的相关性.

在第 q 次迭代时, 寻找能够最大限度地增加标签集相关性的标签, 并将其加入到 W_{q-1} 中. 综合式 (2) 与式 (3), 优化目标 (1) 改写为

$$\begin{aligned}\widehat{w} = \arg\max_{w\in W\backslash W_{q-1}} F(w,x) = \\ \arg\max_{w\in W\backslash W_{q-1}} (\alpha\cdot r(w,x) + \\ (1-\alpha)\cdot c(w,W_{q-1}))\end{aligned} \tag{4}$$

1.1 标签集对图像相关性概率估计

如果标签集能够明确地描述图像的视觉内容, 则该标签集与图像相关. 两种情况会导致标签集对图像的描述不明确. 第一种情况是词汇的多义性, 例如标签集 {“apple”, “photo”} 对于目标图像而言就具有多义性, 无法明确图像中的视觉内容是“水果” 还是 “电脑”. 而 {“apple”, “computer”} 或者 {“apple”, “fruit”} 两个标签集就不存在这个问题. 第二种情况是标签集合的相关性不足, 如标签集 {“car”} 的相关性不足, 如果标签集中含有标签 “Chevrolet” (雪弗兰) 或者 “Volkswagen” (大众), 则标签集对图像的相关性会增加较多, 相对而言, 候选标签 “nice” 或者 “cool” 加入到标签集中, 标签集对图像的相关性增加较少. 因此, 在标签集框架下讨论标签对图像的相关性, 主要考虑将该标签加入到标签集后, 能带来多少相关性的增加. 如果其带来较大的相关性增加, 必然导致词汇表中剩余标签对候选标签集的后验概率有较大变化, 所以可以依据后验概率的变化量来选择标签. 考虑如下场景, 当前标签集为 {“car”}, 要从 “yellow”, “Chevrolet”, “Volkswagen”, “Germany”, “US” 等标签中选择 1 个标签加入到标签集中, 这些候选标签和 “car” 的共现频率均较高, 但是, 当 “Chevrolet” 加入到标签集后, 标签集拓展为 {“car”, “Chevrolet”}, 这时 “Volkswagen” 或者 “Germany” 和标签集 {“car”, “Chevrolet”} 的共现频率就会非常低, 后验概率的变化量也较大, 这也体现了标签集 {“car”, “Chevrolet”} 比原有的标签集 {“car”}对图像的相关性更强. 如果选择 “yellow” 加入到候选标签集合中, 则后验概率的变化就会较小, 这也体现了标签集 {“car”, “yellow”} 和标签集 {“car”, “Chevrolet”} 相比, 相关性较弱. 假设包含 $q-1$ 个标签的标签集 $W_{q-1}\subseteq W$, 令 $w_i\in V$, 表示 W 中剩余标签, 其中 $V = W\backslash W_q$. 标签 w_i 的先验概率为

$$p(w_i) = \frac{\sum\limits_{w_i\in W} co(w_i,w_j)}{\sum\limits_{w_k\in W, w_j\in W} co(w_k,w_j)}$$

后验概率为

$$p(w_i|w_j) = \frac{co(w_i,w_j)}{\sum\limits_{w_a\in W} co(w_a,w_j)}$$

其中, $co(a,b)$ 表示两个标签的共现次数. 选择 w_i 加入到标签集 W_{q-1} 中后, 标签集拓展为 $W_q^i = W_{q-1}\cup\{w_i\}$. 假设标签之间彼此独立, 即

$$\begin{aligned}p(w_i|W_{q-1}) = \frac{p(W_{q-1}|w_i)p(w_i)}{p(W_{q-1})} = \\ \frac{p(w_i)\prod\limits_{w_a\in W_{q-1}} p(w_a|w_i)}{\sum\limits_{w_j\in W} p(w_j)\prod\limits_{w_a\in W_{q-1}} p(w_a|w_j)}\end{aligned} \tag{5}$$

选择 w_i 对标签集 W_{q-1} 进行扩展前后的词汇表中剩余标签 w' 对标签集 W_q^i 的 Kullback-Leibler 散度为

$$\begin{aligned}KL(p(W|W_q^i), p(W|W_{q-1})) = \\ KL(p(W|W_q^i)||p(W|W_{q-1})) + \\ KL(p(W|W_{q-1})||p(W|W_q^i)) = \\ \sum_{w'\in W} p(w'|W_q^i)\log\frac{p(w'|W_q^i)}{p(w'|W_{q-1})} + \\ \sum_{w'\in W} p(w'|W_{q-1})\log\frac{p(w'|W_{q-1})}{p(w'|W_q^i)}\end{aligned} \tag{6}$$

选择不同的标签对当前标签集进行扩展, 会导致式 (6) 的散度值变化, 变化量较大的标签意味着扩展后的标签集相关性更强, 即

$$\begin{aligned}r(w,x)\propto f(KL(p(W|W_q^i), p(W|W_{q-1}))) = \\ f(KL(p(W|W_q^i)||p(W|W_{q-1})) + \\ KL(p(W|W_{q-1})||p(W|W_q^i)))\end{aligned} \tag{7}$$

式 (7) 中, $f(\cdot)$ 是一个单调递增函数. 由于式 (7) 选择的标签并没有考虑到标签和图像内容的相关性, 所以, 对于标签集“car”, 可能选择具有大信息量的标签 “cup”, 因为选择该标签会导致词汇表中剩余标签对标签集的后验概率发生较大变化. 所以式 (7) 改写为

$$\begin{aligned}r(w,x)\propto \\ f(KL(p(W|W_q^i), p(W|W_{q-1})))\cdot p(w|x)\end{aligned} \tag{8}$$

式 (8) 中, $p(w|x)$ 为图像 x 具有标签 w 的概率, 可过滤与图像内容不相关的标签. 由于 $p(w|x)\propto p(w,x)$, 式 (8) 进一步改写为

$$\begin{aligned}r(w,x)\propto \\ f(KL(p(W|W_q^i), p(W|W_{q-1})))\cdot p(w,x)\end{aligned} \tag{9}$$

式 (9) 中, 通过图像 x 的 k 邻域内样本求得 $p(w,x)$:

$$p(w,x)=\sum_{i=1}^{k}p(w,x|x_i)p(x_i)=\sum_{i=1}^{k}p(w|x_i)p(x|x_i)p(x_i)\approx\frac{1}{k}\sum_{i=1}^{k}\delta(w,x_i)\mathrm{sim}_{\mathrm{visual}}(x,x_i)\tag{10}$$

式 (10) 中, $\delta(w,x_i)$ 为指示变量, 如果 x_i 具有标签 w, 则 $\delta(w,x_i)$ 为 1, 否则为 0. $\mathrm{sim}_{\mathrm{visual}}(x,x_i)$ 为图像视觉相似度. 由于式 (10) 只是利用了图像的视觉特征, 因此得到的近邻样本只是视觉相似, 而图像的文本标签没有参与相似度计算, 得到的邻域内含有噪声. 由于测试图像不具有标签, 而只能得到其视觉特征, 所以我们首先选取图像 x 的 K ($K>k$) 个视觉相似近邻, 邻域内样本之间采用其文本标签向量得到相似度, 这样 K 邻域内样本 x_i 和图像 x 的相似度为

$$\mathrm{sim}(x_i,x)=\frac{1}{K}\sum_{x_j\in N(x)}\mathrm{sim}_{\mathrm{text}}(x_i,x_j)\mathrm{sim}_{\mathrm{visual}}(x,x_j)\tag{11}$$

式 (11) 中, $N(x)$ 图像 x 的 K 邻域, K 邻域内样本相似度 $\mathrm{sim}_{\mathrm{text}}(x_i,x_j)$ 由对应标签向量的余弦距离得到. 依据式 (11) 选择 top-k 个语义相似图像作为其语义近邻, 参与式 (10) 的计算.

1.2 标签集内部相关性概率估计

通过求解标签集对图像相关性, 可以选择和目标图像相关的标签, 并将其加入到标签集中. 但是这隐含了一个假设, 即标签之间彼此独立. 该假设忽略了标签集内部标签的相关性. 假设图像的标签集为 {“sky”, “sun”}, 则标签 “rain” 很可能会被加入到标签集中, 因为对于该标签集, 标签 “rain” 的加入将增加相关性, 并且 “rain” 与 “sky” 也具有较强的相关性. 但是标签 “rain” 与标签集 {“sky”, “sun”} 整体语义不相关, 将其加入到标签集中并不合理, 应该过滤掉该标签. 当然, 也可能存在这样的图像, 符合标签集 {“sky”, “sun”,“rain”} 的描述, 我们这里只是从统计意义上进行相关性估计, 即对多数情况适用. 此外, 和标签集内标签相关性较大的标签应该加入到标签集合中, 如标签 “pet” 与标签集 {“animal”,“dog”} 整体语义一致, 则其可以加到该标签集中. 所以式 (2) 中标签 w 对标签集 W_{q-1} 的相关性 $c(w,W_{q-1})$ 定义为对应向量的相似度

$$c(w,W_{q-1})=\mathrm{sim}_{\mathrm{text}}(V_w,V_{W_{q-1}})\tag{12}$$

式 (12) 中, 标签集 W_{q-1} 与标签 w 表示为两个向量 $V_{W_{q-1}}$ 与 V_w, 采用余弦相似度作为相关性度量. 为了得到标签集 W_{q-1} 的向量 $V_{W_{q-1}}$, 首先为集合中每一个标签 w 构造一个向量 V_w. 受自然语言处理中 “词袋模型” 启发, 将每幅图像看作一个包含其附属标签的文档, 构造 “标签–图像” 矩阵 $M_{|W|\times n}$, 其中 $|W|$ 为词汇表规模, n 为图像集规模, M 的第 i 行表示标签 w_i 在图像中出现的情况, 第 j 列为图像 x_j 的标签向量 ($M_{i,j}=1$ 表示 x_j 具有标签 w_i, $M_{i,j}=0$ 表示图像 x_j 不具有标签 w_i). 进而得到 “标签–标签” 矩阵 $U_{|W|\times|W|}=MM^{\mathrm{T}}$, U_{ij} 为标签 w_i 与 w_j 之间的相关性. 对矩阵元素规范化处理, 即 $U_{ij}=\frac{U_{ij}}{U_{ii}+U_{jj}-U_{ij}}$, 去除高频标签引起的偏差, 最终得到的 U 的第 i 行向量 U_i 为标签 w_i 的邻域向量, 则两个标签 w_i 与 w_j 的相关性可由对应的邻域向量 U_i 与 U_j 得到. 令标签 w_i 的邻域向量

$$V_{w_i}=\langle v_{i1},\cdots,v_{i|W|}\rangle=\langle\beta_iU_{i,1},\cdots,\beta_iU_{i,|W|}\rangle$$

标签 w_j 的邻域向量

$$V_{w_j}=\langle v_{j1},\cdots,v_{j|W|}\rangle=\langle\beta_jU_{j,1},\cdots,\beta_jU_{j,|W|}\rangle$$

其中, β 表示标签对目标图像的权重, 依据式 (16) 可得. 合并两个标签邻域向量为标签集邻域向量 $V_{w_i\bigcup w_j}$, 其向量元素

$$v_k=\begin{cases}\dfrac{\max(\beta_i,\beta_j)}{\beta_j}(v_{ik}+v_{jk}), & v_{ik}v_{jk}\neq 0\\ v_{ik}+v_{jk}, & \text{否则}\end{cases}\tag{13}$$

其中, $\max(\beta_i,\beta_j)$ 表示对于目标图像而言, 重要标签具有更大的权重. 逐次将标签合并, 得到组合概念

$$W_q=\bigcup_{w_i\in W_q}w_i=\left(\bigcup_{w_i\in W_{q-1}}w_i\right)\bigcup w_q$$

同理, 依据式 (13) 逐次对标签集 W_q 中标签的邻域向量进行合并, 最终得到标签集 W_{q-1} 的邻域向量 $V_{W_{q-1}}$.

2 算法复杂度分析与优化

如前所述, 理想的标注结果应该兼具标签集对图像相关性和标签集内部相关性, 即

$$W_q^*=\arg\max(F(W_q,x)),\quad W_q\subset W\tag{14}$$

式 (14) 的求解是一个非线性整数规划问题, 不存在多项式时间内的精确求解算法. 考虑到网络数据集规模较大, 候选标签集随着词汇表的规模呈指数增长, 因此, 通过贪心搜索算法来求解该问题的近似最优解, 构造选标签集 W_q. 求解过程如算法 1 所示.

初始时, 将 W_q 初始化为空集. 依据式 (11) 获得目标图像 x_t 的 k 个语义近邻, 然后依据式 (16) 选择一个和目标图像 x_t 相关性最大的一个标签 w_i, 将 w_i 作为第一个标签加入到 W_q 中. 然后, 算法迭代地寻找剩余的标签, 每轮迭代在除 W_q 外的剩余标签中寻找标签 w_r, 按照式 (2) 找到相关性得分最高的标签 w_r, 加入到当前的标签集 W_q 中.

算法 1. 基于标签集相关性学习的图像语义标注算法 LSLabel

输入. 测试图像 x_t; 训练集 T; 标签数量 q.

输出. x_t 的标签集 W_q.

1) 初始化 $W_q = \varnothing$;

2) 依据式 (11) 得到 x_t 的 k 近邻; 依据式 (16) 选出标签 w_i;

3) $W_q = W_q \bigcup \{w_i\}$;

4) For $i = 2$ to q do

5) 从 $W \backslash W_q$ 中选出标签 w_r, 满足

$w_r = \arg \max\limits_{w \in W/W_i} \alpha \cdot r(w,x) + (1-\alpha) \cdot c(w, W_{q-1}))$;

6) $W_q = W_q \bigcup \{w_r\}$;

7) End For;

8) Return W_q.

算法 1 主要计算步骤为按照式 (9) 估计 $r(w,x)$, 其时间复杂度为 $\mathrm{O}(|W|^2)$, 算法 1 的总时间复杂度为 $\mathrm{O}(q \cdot |W|^3)$. 在实际计算时, q 的值一般较小, 故算法的运行时间主要依赖于数据集中标签数量. 由于网络图像集合一般词汇表规模庞大, 所以算法 1 不能够有效求解. 考虑到标签集中大部分标签和目标图像是不相关的, 相应的相关性无需计算, 对算法 1 采用如下优化策略. 首先, 按照式 (11) 求取目标图像的 k 个语义近邻, 令其邻域集合的标签构成集合 Γ, 式 (2) 中候选标签只在 Γ 中选取, 而不从全局词汇表 W 中选取. 即式 (2) 改写为

$$\hat{w} = \arg \max_{w \in \Gamma / W_{q-1}} (\alpha \cdot r(w,x) + (1-\alpha) \cdot c(w, W_{q-1})) \tag{15}$$

按照式 (16) 求解标签 w_i 对图像 x 的相关性:

$$\beta_i = \frac{0.5k - |count_{\mathrm{neighborhood}}(w_i) - 0.5k|}{\log(|count(w_i)| + 1)} \cdot p(w,x) \tag{16}$$

式 (16) 中, $count_{\mathrm{neighborhood}}(w_i)$ 表示图像 x 的 k 近邻 (按照式 (11) 获得) 中具有标签 w_i 的样本数, $count(w_i)$ 为数据集中具有标签 w_i 的图像数, $p(w,x)$ 为图像 x 具有标签 w 的概率, 由式 (10) 可得, 常量 $0.5k$ 用以降低高频标签影响, 缓解标签分布不平衡带来的偏差. 按照 β 值取得相应图像的 top-m 个标签构成集合 Γ^*, 则式 (9) 改写为

$$\begin{aligned} r(w,x) \propto\ & f(KL(p(\Gamma^*|W_q^i), p(\Gamma^*|W_{q-1}))) \cdot p(w,x) = \\ & f\left(\sum_{w' \in \Gamma^*} p(w'|W_q^i) \log \frac{p(w'|W_q^i)}{p(w'|W_{q-1})} + \right. \\ & \left. \sum_{w' \in \Gamma^*} p(w'|W_{q-1}) \log \frac{p(w'|W_{q-1})}{p(w'|W_q^i)} \right) \cdot p(w,x) \end{aligned} \tag{17}$$

同时, 式 (5) 改写为

$$\begin{aligned} p(w_i|W_{q-1}) &= \frac{p(W_{q-1}|w_i)p(w_i)}{p(W_{q-1})} = \\ & \frac{p(w_i) \prod\limits_{w_a \in W_{q-1}} p(w_a|w_i)}{\sum\limits_{w_j \in \Gamma^*} p(w_j) \prod\limits_{w_a \in W_{q-1}} p(w_a|w_j)} \end{aligned} \tag{18}$$

优化目标式 (4) 改写为

$$\hat{w} = \arg \max_{w \in \Gamma^* \backslash W_{q-1}} F(w,x)$$

这样, $r(w,x)$ 时间复杂度降低为 $\mathrm{O}(m^2)$, 其中 $m = |\Gamma^*|$ 为集合 Γ^* 含有的标签数量, $F(x,w)$ 复杂度降低为 $\mathrm{O}(m^2|\Gamma|)$, 算法 1 的复杂度降低为 $\mathrm{O}(qm^2 \times |\Gamma|)$, 由于 m 与 $|\Gamma|$ 远小于词汇表规模, 所以该算法对于大规模数据集是有效的.

3 实验结果及分析

采用 NUS-WIDE[9] 与 Flickr 25K[10] 数据集. NUS-WIDE 数据集来自 Flickr 图像共享网站约 5 000 名用户提供的 269 648 幅图像和 425 059 个不同的标签, 图像内容包含丰富多样的物体和场景, 反映了网络图像集的真实情况, 我们以 5 018 个基准标签进行测试. Flickr 25K 数据集包含 1 386 个频率大于 20 的标签和 25 000 幅图像和组成. NUS-WIDE 中取 200 000 幅图像为训练集, 其余为测试集. Flickr 25K 中取 20 000 幅图像为训练集, 其余为测试集. 对于测试集中的每一幅图像, 通过不同的标注方法产生标签, 6 名志愿者独立的对标签的相关性进行判断, 最后通过投票确定标签是否与图像内容相关. 对图像提取征 64 维颜色特征, 包含 44 维颜色相关图、14 维颜色纹理矩、6 维颜色矩构成, 384 维 GIST 特征, PCA 降维后的 30 维 Dense-Surf 特征. 采用图像标注方法中常用的评价指标进行评测, 包括平均标签准确率 (P)、平均标签召回率 (R) 和 F1 值, 同时我们还测试了平均图像准确率和平均图像召回率. 令 $\#(s)$ 表示标注结果中包含标签 w 的样本数, $\#(c)$ 为正确标注的样本数, $\#(t)$ 为测试集中包含标签 w 的样本总数, 则 $\mathrm{P}(w) = \#(c)/\#(s)$,

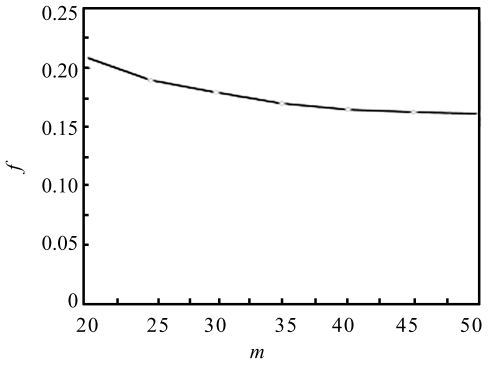
$\mathrm{R}(w) = \#(c)/\#(t)$, 对测试集所有标签的上述度量求取均值得到平均标签准确率 (Average precision of label, APL) 和平均标签召回率 (Average recall of label, ARL) 作为评价指标, $\mathrm{F1} = \frac{2\times APL\times ARL}{APL+ARL}$. 令 q 表示一幅图像 I 标注结果中的标签数, $\#(c)$ 为图像 I 正确标注的标签数, $\#(t)$ 为 I 的基准标签数, 则 $\mathrm{P}(I) = \#(c)/q$, $\mathrm{R}(I) = \#(c)/\#(t)$, 对测试集所有标签的上述度量求取均值得到平均图像准确率 API 和平均图像召回率 (Average recall of image, ARI).

3.1 参数调节与估计

由于大多数标签对标签集相关性的估计值大于 0.3, 而对应的散度值却小于 0.1, 所以为了使得式 (4) 的两个约束项取值范围相近, 将函数 $f(\cdot)$ 定义为 $(\cdot)^{0.5}$, 即式 (17) 改写为

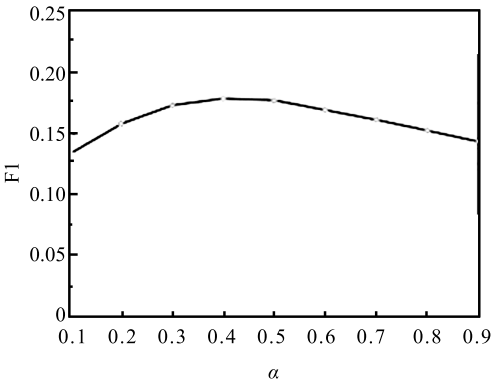
$$r(x,w) \propto (KL(p(\Gamma^*|W_q^i), p(\Gamma^*|W_{q-1})))^{0.5} \cdot p(w,x) \tag{19}$$

式 (18) 中, Γ^* 的大小 m 对 KL 散度值具有影响, 图 2 (a) 记录了随 m 取值变化的 $f(\cdot)$ 曲线, 可以看到, 当 m 大于 40 的时候, 值趋于稳定, 所以实验中 Γ^* 的规模 m 取值为 40. 式 (4) 中参数 α 是平衡项, 用来平衡在给定图像的情况下, 标签集 W_q 对图像相关性和标签集 W_q 内部相关性. 式 (11) 中参数 K 与式 (10) 中参数 k 用来进行邻域样本选取, 实验中设置 $K = 70$, $k = 35$, 然后记录 α 在 {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9} 中取值的情况下算法的 F1 值, 如图 2 (b) 所示, 当 α 在 $0.3\sim0.5$ 之间变化的时候, 性能达到最佳, 设置为 0.4. 图 2 (c) 给出了固定 α 为 0.4 时, K 和 k 在 {50, 25; 70, 35; 90, 45; 120, 60; 150, 75} 中变化的情况下, LSLabel 的 F1 值, 可以看到, 随着 K 和 k 的增加, 性能逐渐升高, 因为邻域范围限制过小会导致参与式 (10) 和式 (11) 的样本数量过少, 但是当值超过 90 和 45 后, 系统性能开始下降, 这是因为邻域范围过大, 目标图像的近邻样本增多, 引入了噪声, 噪声样本和其附属标签参与了参与式 (10) 和式 (11) 的计算, 导致系统性能下降, 所以固定 K 为 90 且 k 为 45.



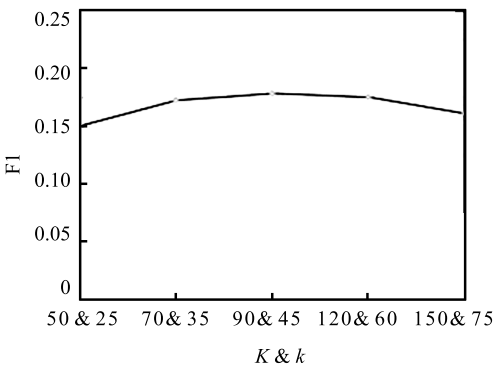
(a) 随 m 变化的 f 曲线

(a) f curve with respect to m



(b) 随 α 变化的 F1 曲线

(b) F1 score with respect to α



(c) 随 K 和 k 变化的 F1 曲线

(c) F1 score with respect to K and k

图 2 系统参数取值变化对性能影响

Fig. 2 Performance with respect to various parameter values

3.2 评测集上算法性能与效率分析

表 1 和表 2 记录了 NUS-WIDE 数据集上只采用 "标签集对图像相关性" 的方法 LSLabel(r), 将 "标签集对图像相关性" 和 "标签集内部相关性" 融合的方法 LSLabel($r + c$), 及在此基础上依据大规模概念集合进行优化的方法 LSLabel($r + c + o$) 的平均图像准确率与平均图像召回率. 可以看到, 集成 "标签集对图像相关性" 和 "标签集内部相关性"的方法 LSLabel($r + c$) 较 LSLabel(r) 在平均图像准确率上平均提高 4.9 %, 在平均图像召回率上平均提高 2.1 %, 可得出结论, 我们在标签集相关性上集成两方面约束的策略是有效的. 从实验结果还可以看到, 针对大规模数据集优化后的方法 LSLabel($r+c+o$) 在平均图像准确率和召回率上较 LSLabel($r + c$) 略

表 1　平均图像准确率

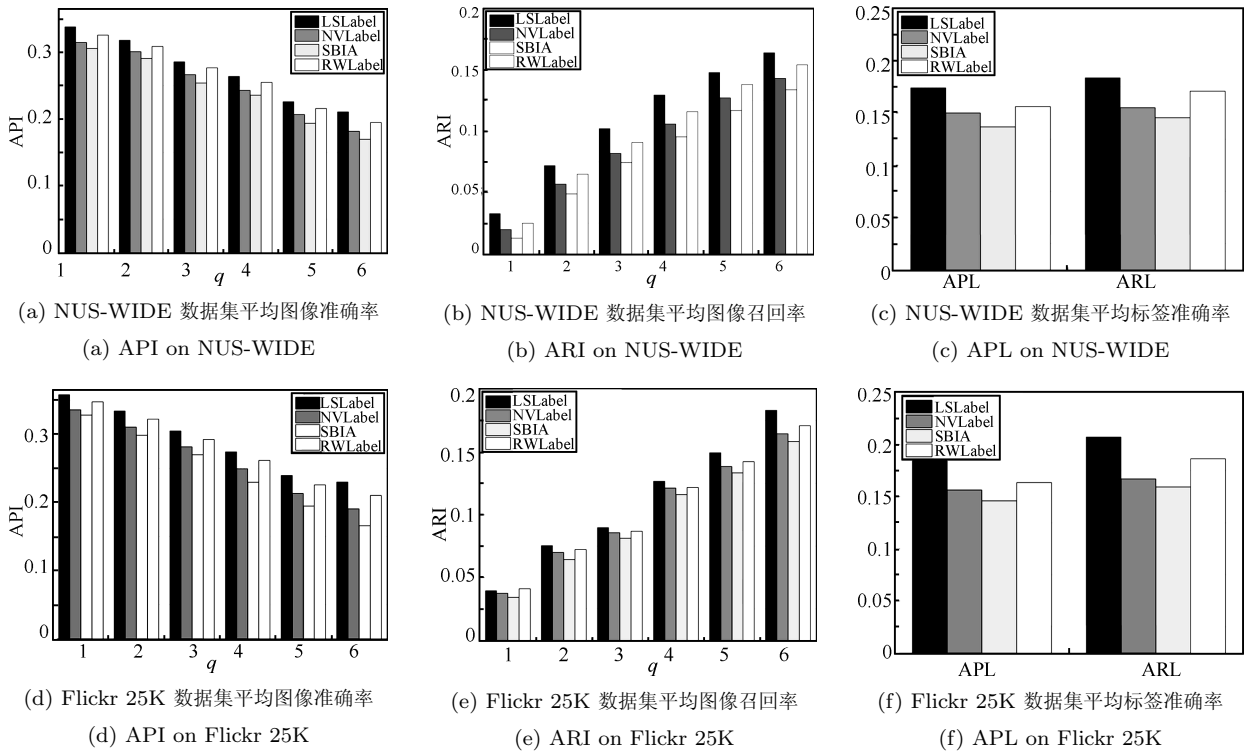
Table 1　Average precision per image

	$q=1$	$q=2$	$q=3$	$q=4$	$q=5$	$q=6$
LSLabel(r)	0.307	0.278	0.245	0.228	0.195	0.168
LSLabel($r+c$)	0.340	0.321	0.290	0.268	0.231	0.217
LSLabel($r+c+o$)	0.338	0.318	0.285	0.264	0.226	0.211

表 2　平均图像召回率

Table 2　Average recall per image

	$q=1$	$q=2$	$q=3$	$q=4$	$q=5$	$q=6$
LSLabel(r)	0.029	0.066	0.094	0.116	0.133	0.148
LSLabel($r+c$)	0.034	0.079	0.110	0.135	0.154	0.169
LSLabel($r+c+o$)	0.033	0.075	0.107	0.130	0.148	0.164



(a) NUS-WIDE 数据集平均图像准确率　(b) NUS-WIDE 数据集平均图像召回率　(c) NUS-WIDE 数据集平均标签准确率

(a) API on NUS-WIDE　(b) ARI on NUS-WIDE　(c) APL on NUS-WIDE

(d) Flickr 25K 数据集平均图像准确率　(e) Flickr 25K 数据集平均图像召回率　(f) Flickr 25K 数据集平均标签准确率

(d) API on Flickr 25K　(e) ARI on Flickr 25K　(f) APL on Flickr 25K

图 3　标注性能比较

Fig. 3　Comparison of the annotation performance

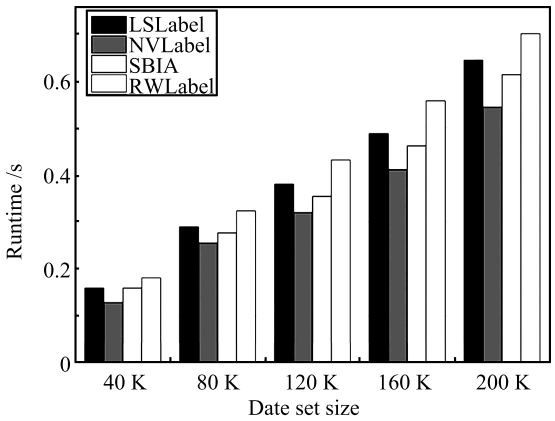
有降低, 但是 LSLabel($r+c+o$) 具备大规模概念集合的在线处理能力. 比较 4 种在线图像标注算法, 包括基于搜索的图像标注方法 (SBIA)[3]、基于标签相关图随机游走的标注方法 (RWLabel)[4]、基于概率相关性估计的标注算法 (NVLabel) (RWLabel 与 SBIA 所需初始标签由 NVLabel 提供)[5] 和基于标签集相关性学习的标注算法 (为便于表示, 将 LSLabel($r+c+o$) 简记为 LSLabel). 图 3 (a) ∼ 3 (f) 分别给出了两个标准评测集上, 标签输出数量从 1 到 6 变化的情况下的平均图像准确率与平均图像召回率. 在 NUS-WIDE 数据集上, 在标签集规模限定为 6 时, LSLabel 在平均图像标注准确率 API 上高出 NVLabel 16 %, 高出 SBIA 24 %, 高出 RWLabel 8.4 %, 在平均图像标注召回率 ARI

上高出 NVLabel 14.7 %, 高出 SBIA 22.6 %, 高出 RWLabel 6.4 %. 在 Flickr 25K 数据集上, 在标签集规模限定为 6 时, LSLabel 在平均图像标注准确率 API 上高出 NVLabel 20.6 %, 高出 SBIA 38.6 %, 高出 RWLabel 9.4 %, 在平均图像标注召回率 ARI 上高出 NVLabel 11.2 %, 高出 SBIA 15.6 %, 高出 RWLabel 7.1 %. LSLabel 标注效果明显优于其他 3 种在线标注算法. 其原因在于, SBIA 依据相似图像结果聚类, 对于超过阈值的簇主题标签进行输出, 但是无法确定合理的簇数量. NVLabel 利用概率相关性估计标签相关性, 倾向于输出相似图像集合的高频标签, 而依据概率密度估计与随机游走的 RWLabel 与 NVLabel 相同, 所以这两种方法都倾向于利用高频标签进行标注. 基于标签相关性学习的 LSLabel 方法将候选标签集看作一个整体, 标签集具备与目标图像的相关性, 集合内标签具备语义相关性. 从实验结果还可以看到, 随着标注结果中标签数量的增加, 噪声标签也随之增加, 所有方法的标注准确率均下降. 图 3 (c) 与图 3 (f) 分别给出了标签集规模为 6 时, 两个评测集上各种方法的平均标签准确率 (APL) 和平均标签召回率 (ARL). 对比其他在线标注方法, LSLabel 在 NUS-WIDE 数据集上平均标签准确率 (APL) 提高了 11.4 %～27.2 %, 在平均标签召回率 (ARL) 上提高了 7.4 %～26.2 %. LSLabel 在 Flickr 25K 数据集上平均标签准确率 (APL) 提高了 12.6 %～26.2 %, 平均标签召回率 (ARL) 提高了 11.2 %～30.0 %.

图 4 (a) 与图 4 (b) 给出了各种方法在标注标签数量为 6 时, NUS-WIDE 数据集上训练集规模从 40 000 到 2 000 000, Flickr 25K 数据集上训练集规模从 8 000 到 20 000 的情况下算法运行时间的对比结果. 可以看到, 运行时间随着图像规模增长而增加. LSLabel 方法的运行时间低于 RWLabel, 高于 SBIA, NVLabel 算法运行时间最少, RWLabel 耗时最长. 该结果与各种方法的算法时间复杂度一致. RWLabel 的复杂度为 $\mathrm{O}(k \cdot t^2 \cdot n^2 + m \cdot k \cdot n^2)$, 其中 m 为迭代次数, n 为每幅图像的平均标签数量, t 为每标签的平均图像数量, 由于 t 通常较高, 所以其运行效率较低, 而 NVLabel 的复杂度为 $\mathrm{O}(k \cdot t_{\max})$, 其中 k 为邻域样本数量, $t_{\max}$ 为每幅图像的最大标签数量, 而 SBIA 因为对相似图像结果聚类, 所以耗时高于 NVLabel.

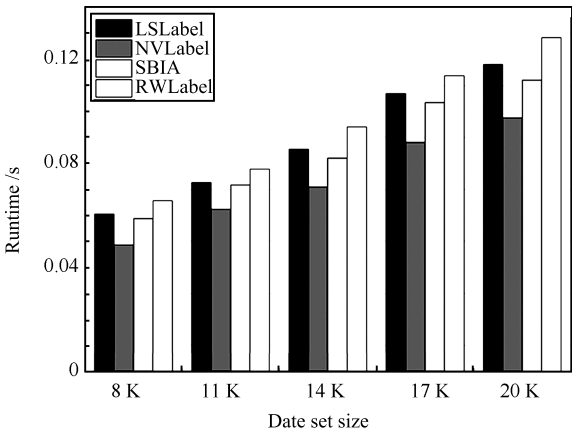
4 结论

现有的图像标注方法依据单个标签与图像的相关性获得最终标注结果, 获得的标签集存在相关性缺乏和语义歧义等问题. 针对大规模网络图像的在线自动标注问题, 本文提出的基于标签集相关性学习的图像标注方法将标签集相关性分解为标签集对图像相关性和标签集的内部相关性, 并将标签集对图像相关性和标签集内部相关性两方面约束形成一个优化问题, 利用贪心搜索获得该问题的近似最优解, 在此基础上给出了大规模图像和概念集合上的优化求解策略, 获得的标签集能够更好地描述图像语义. 网络数据集上的测试表明了方法的有效性.



(a) NUS-WIDE 数据集运行时间

(a) Runtime on NUS-WIDE dataset



(b) Flickr 25K 数据集运行时间

(b) Runtime on Flickr 25K dataset

图 4 方法执行效率

Fig. 4 Efficiency of various methods

References

1 Zhang D S, Islam M M, Lu G J. A review on automatic image annotation techniques. *Pattern Recognition*, 2012, **45**(1): 346−362

2 Wang M, Ni B B, Hua X S, Chua T S. Assistive tagging: a survey of multimedia tagging with human-computer joint exploration. *ACM Computing Surveys*, 2012, **44**(4): 1−24

3 Wang X J, Zhang L, Li X R, Ma W Y. Annotating images by mining image search results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, **30**(11): 1919−1932

4 Liu D, Hua X C, Yang L J. Tag ranking. In: Proceedings of the 2009 International World Wide Web Conference. New York, USA: ACM, 2009. 351−360

5 Li X R, Snoek C G M, Worring M. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, 2009, **11**(7): 1310−1322

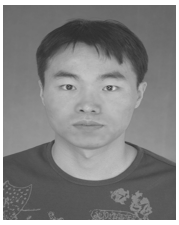
6 Jin Y, Khan L, Prabhakaran B. Knowledge based image annotation refinement. *Journal of Signal Processing Systems*, 2010, **58**(3): 387−406

7 Wang H, Huang H, Chris H Q D. Image annotation using bi-relational graph of images and semantic labels. In: Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2011. 793−800

8 Yang Y, Wu F, Nie F P, Shen H T, Zhuang Y, Hauptmann A G. Web and personal image annotation by mining label correlation with relaxed visual graph embedding. *IEEE Transactions on Image Processing*, 2012, **21**(3): 1339−1351

9 Chua T S, Tang J H, Hong R C, Li H J, Luo Z P, Zheng Y T. NUS-WIDE: A real-world web image database from National University of Singapore. In: Proceedings of the 2009 ACM Conference on Image and Video Retrieval. New York, USA: ACM, 2009. 1−9

10 Huiskes M J, Lew M S. The MIR Flickr retrieval evaluation. In: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval. New York, USA: ACM, 2008. 39−43



田 枫 东北石油大学计算机与信息技术学院副教授. 2014 年获得北京航空航天大学博士学位. 主要研究方向为跨媒体理解和多媒体数据挖掘. 本文通信作者. E-mail: tianfeng80@gmail.com

(**TIAN Feng** Associate professor at Northeast Petroleum University. He received his Ph. D. degree from Beihang University in 2014. His research interest covers cross media understanding and multimedia mining. Corresponding author of this paper.)

沈旭昆 北京航空航天大学计算机学院教授. 主要研究方向为虚拟现实与可视化, 计算机视觉, 多媒体内容管理.
E-mail: xkshen@buaa.edu.cn

(**SHEN Xu-Kun** Professor at Beihang University. His research interest covers visual reality, computer vision, and multimedia content management.)