Journal of System Simulation

Volume 28 | Issue 11 Article 29

8-13-2020

Multimedia Annotation Refinement Based on Contextual Information Diffusion

Tian Feng
School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China;

Fuhua Shang
School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China;

Multimedia Annotation Refinement Based on Contextual Information Diffusion

Abstract

Abstract: *A data driven multimedia annotation refinement method based on dataset contextual information diffusion was proposed*. The label contextual graph was constructed, and the label correlation can be diffused on textual label space; Multimedia object content relevant graph was constructed. Label contextual graph and multimedia object content relevant graph were mutually reinforced and formulated into a regularized framework. The proposed method incorporated both multimedia content correlation and label contextual information, and the optimization process was solved by approximate solution algorithm. The experimental results on real world dataset show that the proposed method can obviously improve the annotation performance.

Keywords

multimedia annotation, semantic annotation, annotation refinement, contextual diffusion

Recommended Citation

Tian Feng, Shang Fuhua. Multimedia Annotation Refinement Based on Contextual Information Diffusion[J]. Journal of System Simulation, 2016, 28(11): 2860-2867.

第 28 卷第 11 期 系统仿真学报© Vol. 28 No. 11
2016 年 11 月 Journal of System Simulation Nov., 2016

基于上下文语境传播的多媒体语义标注改善

田枫，尚福华
(东北石油大学计算机与信息技术学院，大庆 163318；)

摘要：*提出了一种数据驱动的多媒体对象标注改善方法，利用数据集蕴含的丰富语境相关信息对基本多媒体对象标注结果进行优化。*以标签为节点，相关度为边，构造标签语境相关图，实现概念空间上的相关性传播；将多媒体对象内容特征和文本模态特征互增强过程集成为一个优化框架，通过近似求解策略，实现上下文语境信息传播。该方法充分利用了标签的上下文相关性和多媒体对象的内容相关性。数据集上的实验结果表明，该方法可大幅度提升标注性能。

关键词：多媒体标注；语义标注；语义改善；语境传播
中图分类号：TP391 文献标识码：A 文章编号：1004-731X (2016) 11-2860-08
DOI: 10.16182/j.issn1004731x.joss.201611029

Multimedia Annotation Refinement Based on Contextual Information Diffusion

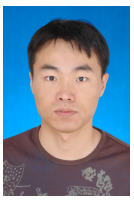
Tian Feng, Shang Fuhua
(School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: *A data driven multimedia annotation refinement method based on dataset contextual information diffusion was proposed*. The label contextual graph was constructed, and the label correlation can be diffused on textual label space; Multimedia object content relevant graph was constructed. Label contextual graph and multimedia object content relevant graph were mutually reinforced and formulated into a regularized framework. The proposed method incorporated both multimedia content correlation and label contextual information, and the optimization process was solved by approximate solution algorithm. The experimental results on real world dataset show that the proposed method can obviously improve the annotation performance.

Keywords: multimedia annotation; semantic annotation; annotation refinement; contextual diffusion

引言

多媒体对象语义标注是实现多媒体语义检索与管理的关键，其任务是给多媒体对象添加反映多媒体内容的文本标签。由于语义鸿沟的存在，多媒体自动标注结果并不能全面，准确地反映多媒体对象的内容，标注改善是多媒体对象标注任务必不可少的环节。Jin 等人利用 WordNet 词典得到标签语义距离，保留高相关度的标签，相关度较低的标签被认为是噪声干扰，予以去除[1-2]。但是这类方法仅考虑标签的字典语义，与数据集统计特性无关。互联网每一天都有新的概念产生，相比于网络词汇和其语义内涵来说，字典的规模和可扩展性是个局限。类似于"ipad"等词汇还没有收录在 WordNet 中，再如，"android"和智能终端具有较强的关联性，但是其在 WordNet 中的解释还停留在"an automaton that resembles a human being"，因此，依据字典得到"apple"与"android"或者"google"的 JCN 相似度均为 0。再如，"street"和"sky"时常出现在同一张照片中，但是这两对标签之间的 JCN 相似度仅为 0.072 5，



收稿日期：2016-04-30 修回日期：2016-07-14；
基金项目：国家自然科学基金(61502094, 61402099)；黑龙江省自然科学基金(F2016002, F2015020)；黑龙江省教育科学规划重点课题(GJB1215019)；
作者简介：田枫(1980-)，男，黑龙江安达，博士，副教授，研究方向为多媒体理解；尚福华(通讯作者 1962-)，男，吉林延吉，教授，博士，研究方向为机器学习。

而在数据集中没有共现的"airport"和"animal"的相似度为 0.080 8，这说明完全依赖词典获取标签之间的相关性无法全面反映数据集中的语境相关信息，语境相关信息不仅包含一些类似于 WordNet 词典中定义的同义关系(比如"football"和"soccer")，上下位关系(如"animal"与"horse"，"car"和"wheel")，还包含语境关联关系(如"horse"和"polo"，"Iran"和"weapon"，"smoke"和"explosion")。数据集所蕴含的语境相关信息越来越受到研究者的重视。如 Wang 等人利用随机游走得到标签之间的统计相关性，优化效果好于基于词典的优化方法[3]。在进一步工作中，他们通过图像的视觉特征定义标签相似度，依据游走的稳定概率进行语义改善[4]。Li 等人通过待优化图像的视觉邻域样本的标签计数对目标图像进行语义改善，优化效果明显[5]。Chen 等人利用 SVM 对样本进行分类，估计标签的初始相关度，然后通过基于图的传播方法提升标注结果。但是其针对每一个概念训练 one-versus-all 分类器的方法，无法在大规模数据集上拓展[6]。Liu 等人通过核密度估计得到标签与视觉特征的初始相关度，并通过标签图上的随机游走获得相关分数，改善标签排序[7]。Xu 等人将图像看做标签组成的文档集合，通过潜在狄利克雷分布模型交替优化标签相似度和相关性[8]。实验结果表明该方法性能略优于文献[7]的方法，但是其图模型的参数估计和主题数目的确定方法导致其无法在真实环境下大规模数据集上应用。Zhu 等人将图像标签相关矩阵分解，得到一个低秩的改善矩阵和一个稀疏的噪声矩阵，但是其要求噪声矩阵要满足稀疏性，标签相关矩阵要低秩，上述因素制约了其在大规模数据集上的应用[9]。文献[10-11]引入标签相关约束，通过基于图的标签传播过程获得标签相关性。文献[12]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得到图像相关性和语义组间相关性。但是同样计算复杂，无法在大规模数据集上的应用。

综上所述，真实数据集上语义概念的丰富性和数据集的规模要求标注改善方法必须具备规模化处理能力。复杂的优化的模型，耗时的参数估计在实际环境下的大规模数据集上并不适用。因此，本文提出了一种基于上下文语境传播的多媒体语义标注改善方法，这种数据驱动的语义改善方法具备规模化处理能力。

1 概念空间上的语义改善

令 $\boldsymbol{Y}_0$ 表示原始的多媒体对象标签矩阵，$\boldsymbol{Y}$ 表示优化后的多媒体对象标签矩阵，$\boldsymbol{Y}_{ij}$ 表示第 j 个标签赋予第 i 个对象的概率，$\boldsymbol{S}$ 表示多媒体对象相关矩阵，$\boldsymbol{S}_{ij}$ 表示两个对象的内容相关度，$\boldsymbol{R}$ 表示标签语境相关矩阵。采用皮尔逊相关系数度量标签 c_i 与 c_j 的相关度：

$$\rho(c_i,c_j)=\frac{1}{n-1}\sum_{k=1}^{n}\left(\frac{\boldsymbol{Y}_{ik}^{\mathrm{T}}-\frac{1}{n}\left\|\boldsymbol{Y}_{ik}^{\mathrm{T}}\right\|_1}{\left(\frac{1}{n-1}\sum_{k=1}^{n}\left(\boldsymbol{Y}_{ik}^{\mathrm{T}}-\frac{1}{n}\left\|\boldsymbol{Y}_{ik}^{\mathrm{T}}\right\|_1\right)^2\right)^{1/2}}\right)\left(\frac{\boldsymbol{Y}_{jk}^{\mathrm{T}}-\frac{1}{n}\left\|\boldsymbol{Y}_{jk}^{\mathrm{T}}\right\|_1}{\left(\frac{1}{n-1}\sum_{k=1}^{n}\left(\boldsymbol{Y}_{jk}^{\mathrm{T}}-\frac{1}{n}\left\|\boldsymbol{Y}_{jk}^{\mathrm{T}}\right\|_1\right)^2\right)^{1/2}}\right)$$

以概念集为顶点，相关度为边，构造语境相关图 $\boldsymbol{R}$。令 $\boldsymbol{Y}_i^{\mathrm{T}}$ 表示 $\boldsymbol{Y}$ 的转置后的第 i 列，其为标签 c_i 的标注结果向量，令 $\boldsymbol{Y}_j^{\mathrm{T}}$ 表示 $\boldsymbol{Y}$ 的转置后的第 j 列，其为标签 c_j 的标注结果向量。标签 c_i 与 c_j 的标注结果向量的差异为 $\left\|\boldsymbol{Y}_i^{\mathrm{T}}-\boldsymbol{Y}_j^{\mathrm{T}}\right\|_2$。如果标签 c_i 与 c_j 语境相关性较强，则对应的标注结果向量的差异应越小，由此，语境相关图 $\boldsymbol{R}$ 上可定义如下损失函数：

$$E(\boldsymbol{R})=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\boldsymbol{R}_{ij}\left\|\frac{\boldsymbol{Y}_i^{\mathrm{T}}}{\sqrt{d(c_i)}}-\frac{\boldsymbol{Y}_j^{\mathrm{T}}}{\sqrt{d(c_j)}}\right\|^2=\mathrm{tr}(\boldsymbol{Y}\boldsymbol{L}\boldsymbol{Y}^{\mathrm{T}})$$

其中，归一化项 $d(c_i)=\sum_{j=1}^{N}\boldsymbol{R}_{ij}$，$\boldsymbol{D}$=diag($d_1$,$d_2$,…,$d_m$)，语境相关图 $\boldsymbol{R}$ 上的拉普拉斯矩阵定义为 $\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{R}\boldsymbol{D}^{-1/2}$。

上述语境相关图 $\boldsymbol{R}$ 的构造方法，形式上与基于图的半监督信息传播和现有的半监督语义改善

第 28 卷第 11 期 系统仿真学报 Vol. 28 No. 11
2016 年 11 月 Journal of System Simulation Nov., 2016

相似，但是本质不同。首先，传统方法是以样本视觉特征为顶点、特征相似性为边构造图，将半监督学习问题定义成一个正则优化问题，包括定义所需优化的目标函数以及使用决策函数在图上的光滑性为指导定义正则化项，最后确定参数，使得决策函数具有两种性质：(1) 决策函数在已标注样本上的输出应尽量与已知标注一致；(2) 决策函数在图上具有光滑性。而概念空间上的语义改善是以概念为节点，以语境相关性为边构造关联图 $\boldsymbol{R}$，通过 $E(\boldsymbol{R})$将 $\boldsymbol{R}$ 的信息传播到基本标注结果 $\boldsymbol{Y}_0$ 中。令 $\nabla_{\boldsymbol{Y}^{\mathrm{T}}} E(\boldsymbol{R}) = \boldsymbol{L}\boldsymbol{Y}^{\mathrm{T}} = 0$，可以得到

$$\boldsymbol{Y}_t^{\mathrm{T}} = \boldsymbol{Y}_{t-1}^{\mathrm{T}} - \alpha \nabla_{\boldsymbol{Y}_{t-1}^{\mathrm{T}}} \boldsymbol{E} = (\boldsymbol{I} - a\boldsymbol{L})^t \boldsymbol{Y}_0^{\mathrm{T}}$$

其中，$a \in [0,1]$ 为步长。对上式做指数展开，并略去高次项，可得

$$\boldsymbol{Y}_t^{\mathrm{T}} = (\boldsymbol{I} - t(\alpha \boldsymbol{L}) + \frac{t^2}{2!}(\alpha \boldsymbol{L})^2 - \frac{t^3}{3!}(\alpha \boldsymbol{L})^3 + \cdots)\boldsymbol{Y}_0^{\mathrm{T}} \approx$$

$$(I - t(\alpha \boldsymbol{L}) + \frac{1}{2}(\alpha t \boldsymbol{L})^2 - \frac{1}{6}(\alpha t \boldsymbol{L})^3)\boldsymbol{Y}_0^{\mathrm{T}}$$

不同于半监督学习过程，上式的求解过程更适用于大规模数据集。

语境相关图 $\boldsymbol{R}$ 受数据集语境制约。例如，训练集中如果存在若干标注有"horse"和"polo"的样本，那么"马"和"马球"之间的相关关系将传播到待优化图像；但是，如果一幅"马"的待优化图像不具有"打马球"的语境，那么这种语义改善对于该图像将失效。如果有成批次的待优化样本提示了新的语境，如"horse"和"grass"的关联，那么这种语境信息应存储到 $\boldsymbol{R}$，即 $\boldsymbol{Y}$ 对 $\boldsymbol{R}$ 应有反馈。

令语境相关图 $\boldsymbol{R}$ 上的拉普拉斯矩阵 $\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{R}\boldsymbol{D}^{-1/2} = \boldsymbol{I} - \boldsymbol{R}^{'}$，则 $E(\boldsymbol{R})$改写为 $E(\boldsymbol{Y}^{\mathrm{T}}, \boldsymbol{R}^{'}) = \frac{1}{2}\mathrm{tr}(\boldsymbol{Y}(\boldsymbol{I} - \boldsymbol{R}^{'})\boldsymbol{Y}^{\mathrm{T}}) = \frac{1}{2}\mathrm{tr}(\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}) - \frac{1}{2}\mathrm{tr}(\boldsymbol{Y}\boldsymbol{R}^{'}\boldsymbol{Y}^{\mathrm{T}})$。令梯度 $\nabla_{\boldsymbol{R}^{'}} \boldsymbol{E} = -\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} = 0$，则 $\boldsymbol{R}_t^{'} = \boldsymbol{R}_{t-1} + \beta \boldsymbol{Y}_{t-1}^{\mathrm{T}}\boldsymbol{Y}_{t-1}$，其中 $\beta \in [0,1]$ 为步长，由 Automatic Local Analysis[13] 可知，$\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}$ 为测试集的语境关系。损失函数 $E(\boldsymbol{Y}^{\mathrm{T}}, \boldsymbol{R}^{'})$ 按照下式展开

$$\begin{cases} \boldsymbol{L}_t = \boldsymbol{I} - \boldsymbol{R}_t^{'} = \boldsymbol{I} - \boldsymbol{R}_{t-1} - \beta \boldsymbol{Y}_{t-1}^{\mathrm{T}}\boldsymbol{Y}_{t-1} = \boldsymbol{L}_{t-1} - \beta \boldsymbol{Y}_{t-1}^{\mathrm{T}}\boldsymbol{Y}_{t-1} \\ \boldsymbol{Y}_t^{\mathrm{T}} = \boldsymbol{Y}_{t-1}^{\mathrm{T}} - \alpha \boldsymbol{L}_{t-1}\boldsymbol{Y}_{t-1}^{\mathrm{T}} \end{cases}$$

算法 1. 概念空间上的语义改善(Semantic context Refinement，SCR).

输入：初始化多媒体对象标签矩阵 $\boldsymbol{Y}_0$；语境相关图 $\boldsymbol{R}$；参数α，β;

输出：优化后的多媒体对象标签矩阵 $\boldsymbol{Y}$

1：初始化 $\boldsymbol{R}$ 的拉普拉斯矩阵 $\boldsymbol{L}_0 = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{R}\boldsymbol{D}^{-1/2}$，$t$=1；

2：Repeat

3：$\boldsymbol{L}_t = \boldsymbol{L}_{t-1} - \beta \boldsymbol{Y}_{t-1}^{\mathrm{T}}\boldsymbol{Y}_{t-1}$；

4：$\boldsymbol{Y}_t^{\mathrm{T}} = \boldsymbol{Y}_{t-1}^{\mathrm{T}} - \alpha \boldsymbol{L}_{t-1}\boldsymbol{Y}_{t-1}^{\mathrm{T}}$；

5：t=t+1；

6：Until 满足收敛准则

7：Return $\boldsymbol{Y}_t$

2 多模态互增强的语义改善

算法 1 通过概念之间的相关性进行语义改善。但是其只考虑了标签之间的相关性，并没有考虑多媒体对象内容对语义改善效果的影响。利用语境相关性进行语义改善，应该同时考虑多媒体对象的内容相关性和标签间的语境相关性。多模态互增强语义的基本思想是：如果多媒体对象内容相关性较强，那么在统计意义上，对应概念之间的语境相关性应加强；另一方面，如果标签间语境相关性较强，那么对应多媒体对象的内容相关性应加强。反过来也同样成立，即如果两个对象的内容相关性较弱，那么对应概念之间的相关性应适当减弱；另一方面，如果两个标签之间语境相关性较弱，那么对应对象的内容相关性也应适当减弱。通过内容相关性与文本模态相关性的互增强过程，对基本标注结果进行优化。相关性的增强只是相对具体的语境，并不意味着多媒体对象之间，或者标签之间确实具有普适的相关性。实际可能出现如下情况，即内容特征较相关的图像，对应的文本模态表示，即标签向量之间的相关性其实较弱，经过上述的互增强过程后，标签之间的语境相关性比原来增强了。这种情况有可能存在，但是本文的相关性增强，只局限在这两个对象内容相关性较强的语境下。

在概念空间的损失函数上扩充第 2 个损失项

$$E(\boldsymbol{S})=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{S}_{ij}\left\|\frac{\boldsymbol{Y}_i}{\sqrt{d(x_i)}}-\frac{\boldsymbol{Y}_j}{\sqrt{d(x_j)}}\right\|^2=$$

$$\mathrm{tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{L}^{\mathrm{v}}\boldsymbol{Y})=\mathrm{tr}(\boldsymbol{L}^{\mathrm{v}}(\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}))$$

其中，归一化项 $d(x_i)=\sum_{j=1}^{n}\boldsymbol{S}_{ij}$，$\boldsymbol{D}$=diag($d_1$,$d_2$,…,$d_n$)，内容特征相关图 $\boldsymbol{S}$ 上的拉普拉斯矩阵定义为 $\boldsymbol{L}^{\mathrm{v}}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}$。将上节的损失函数重新表示为：

$$E(\boldsymbol{R})=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\boldsymbol{R}_{ij}\left\|\frac{\boldsymbol{Y}_i^{\mathrm{T}}}{\sqrt{d(c_i)}}-\frac{\boldsymbol{Y}_j^{\mathrm{T}}}{\sqrt{d(c_j)}}\right\|^2=\mathrm{tr}(\boldsymbol{Y}\boldsymbol{L}^{\mathrm{w}}\boldsymbol{Y}^{\mathrm{T}})$$

其中，归一化项 $d(c_i)=\sum_{j=1}^{N}\boldsymbol{R}_{ij}$，$\boldsymbol{D}$=diag($d_1$,$d_2$,…,$d_m$)，语境相关图 $\boldsymbol{R}$ 上的拉普拉斯矩阵定义为 $\boldsymbol{L}^{\mathrm{w}}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}$。合并两个损失项，得损失函数为：

$$E=\mu\mathrm{tr}(\boldsymbol{L}^{\mathrm{v}}(\boldsymbol{Y}\boldsymbol{Y}^{T}))+\delta\mathrm{tr}(\boldsymbol{L}^{\mathrm{w}}(\boldsymbol{Y}\boldsymbol{Y}^{T}))$$

其中，参数μ，δ为正则参数，平衡多媒体内容特征模态和文本模态。该目标函数将内容相关图 $\boldsymbol{S}$，语境相关图 $\boldsymbol{R}$，多媒体对象的标签向量表示矩阵 $\boldsymbol{Y}$，及文本特征的内容特征表示矩阵 $\boldsymbol{Y}^{\mathrm{T}}$ 集成为一个优化框架，达到内容特征模态和文本模态互增强的目的。需要说明的是，令 $\mu=0$，则目标函数退化为只考虑文本模态的语境相关性；令 $\delta=0$，则目标函数退化为只考虑内容特征模态的相关性。考虑到语义改善的结果 $\boldsymbol{Y}$ 和基本标注结果矩阵 $\boldsymbol{Y}_0$ 的近似一致性假设，得到如下目标函数：

$$\min\{\mu\mathrm{tr}(\boldsymbol{L}^{\mathrm{v}}(\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}))+\delta\mathrm{tr}(\boldsymbol{L}^{\mathrm{w}}(\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}))+\left\|\boldsymbol{Y}-\boldsymbol{Y}_0\right\|_F^2\}$$

$$\text{s.t. } \boldsymbol{Y}_{ij}\geqslant 0$$

为了使得该方法适用于大规模数据集，在迭代求解过程中，每次只优化 $\boldsymbol{Y}$ 的一个行向量，而其余行固定。因此，上式的求解形式表示为

$$\min_{\boldsymbol{Y}_i}\left\{\begin{array}{l}\boldsymbol{Y}_i(\delta\boldsymbol{L}^{\mathrm{w}}+(\mu\boldsymbol{L}_{ii}^{\mathrm{v}}+1)\boldsymbol{I})\boldsymbol{Y}_i^{\mathrm{T}}+\\ 2\mu\sum_{j=1,j\neq i}^{m}(\boldsymbol{L}_{ij}^{\mathrm{v}}\boldsymbol{Y}_j\boldsymbol{Y}_i^{\mathrm{T}}-2\boldsymbol{Y}_0^{i}\boldsymbol{Y}_i^{\mathrm{T}})\end{array}\right\}$$

$$\text{s.t. } \boldsymbol{Y}_{ij}>0$$

其中，$\boldsymbol{Y}_0^{i}$ 表示初始标注矩阵的第 i 行。该问题是一个带约束条件的二次规划问题，如果直接求解这个问题，类似于内插法的时间复杂度为 $O(m^3)$，m 为概念集的规模，采用二次规划问题的传统求解方法求解目标计算复杂性也较高。所以，首先放松非负约束，则目标函数对 $\boldsymbol{Y}_i$ 的梯度为

$$\nabla_{\boldsymbol{Y}_i}E=\boldsymbol{Y}_i(\delta\boldsymbol{L}^{\mathrm{w}}+(\mu\boldsymbol{L}_{ii}^{\mathrm{v}}+1)\boldsymbol{I})+\mu\sum_{j=1,j\neq i}^{m}(\boldsymbol{L}_{ij}^{\mathrm{v}}\boldsymbol{Y}_j^{\mathrm{T}}-\boldsymbol{Y}_0^{i})。$$

令 $\nabla_{\boldsymbol{Y}_i}E=0$，则

$$\boldsymbol{Y}_i=(\boldsymbol{Y}_0^{i}-\mu\sum_{j=1,j\neq i}^{m}(\boldsymbol{L}_{ij}^{\mathrm{v}}\boldsymbol{Y}_j^{\mathrm{T}}))(\delta\boldsymbol{L}^{\mathrm{w}}+(\mu\boldsymbol{L}_{ii}^{\mathrm{v}}+1)\boldsymbol{I})^{-1}$$

这是目标函数在放松约束条件下的解析解，由于网络数据集规模较大，矩阵的求逆运算并不可行。对上式整理得

$$\boldsymbol{Y}_i=(\boldsymbol{Y}_0^{i}-\mu\sum_{j=1,j\neq i}^{m}(\boldsymbol{L}_{ij}^{\mathrm{v}}\boldsymbol{Y}_j^{\mathrm{T}}))(\boldsymbol{I}+(\delta\boldsymbol{L}^{\mathrm{w}}+\mu\boldsymbol{L}_{ii}^{\mathrm{v}}))^{-1}$$

其中含有形如$(\boldsymbol{I}+\boldsymbol{M})^{-1}$ 的求逆运算，对这种矩阵按照下式做逆做泰勒展开：

$$(\boldsymbol{I}+\boldsymbol{M})^{-1}=\boldsymbol{I}+\sum_{i=1}^{\infty}(-1)^i\boldsymbol{M}^i$$

得到下式：

$$(\delta\boldsymbol{L}^{\mathrm{w}}+(\mu\boldsymbol{L}_{ii}^{\mathrm{v}}+1)\boldsymbol{I})^{-1}\approx$$

$$\frac{1}{\mu\boldsymbol{L}_{ii}^{\mathrm{v}}+1}(I+\sum_{j=1}^{p}(-1)^j(\frac{\delta\boldsymbol{L}^{\mathrm{w}}}{\mu\boldsymbol{L}_{ii}^{\mathrm{v}}+1})^j)$$

因此，目标函数在放松非负约束下的近似解为：

$$\boldsymbol{Y}_i\approx\frac{(\boldsymbol{Y}_0^{i}-\mu\sum_{j=1,j\neq i}^{m}(\boldsymbol{L}_{ij}^{\mathrm{v}}\boldsymbol{Y}_j^{\mathrm{T}}))(\boldsymbol{I}+\sum_{j=1}^{p}(-1)^j(\frac{\delta\boldsymbol{L}^{\mathrm{w}}}{\mu\boldsymbol{L}_{ii}^{\mathrm{v}}+1})^j)}{\mu\boldsymbol{L}_{ii}^{\mathrm{v}}+1}$$

其中 p 为常量，实际求解过程中高阶项略去，$\boldsymbol{L}^{\mathrm{w}}$ 可依据训练集预估计并缓存。放松约束的条件下得到的解有可能会违反目标函数的非负约束，对于负值，直接令其投影$\boldsymbol{Y}_{ij}=0$。由于网络数据集上的初始标注结果 $\boldsymbol{Y}$ 的稀疏性很高，所以上式的求解速度很快。算法描述如下：

算法 2. 多模态互增强语义改善算法 (Enhanced Multimodal Refinement，EMR).

输入：多媒体对象标签矩阵 $\boldsymbol{Y}_0$；语境相关图 $\boldsymbol{R}$；参数μ，δ，p，k；

输出：优化后的多媒体对象标签矩阵 $\boldsymbol{Y}$

1：将具有初始标签的多媒体对象集合聚为 k 个簇；

第 28 卷第 11 期 系统仿真学报 Vol. 28 No. 11
2016 年 11 月 Journal of System Simulation Nov., 2016

2：For 每个簇 do

3：初始化 $\boldsymbol{R}$ 的拉普拉斯矩阵 $\boldsymbol{L}^{\mathrm{w}}$；

4：构造视觉相关图 $\boldsymbol{S}$，初始化 $\boldsymbol{S}$ 的拉普拉斯矩阵 $\boldsymbol{L}^{\mathrm{v}}$；

5：Repeat

$$\boldsymbol{Y}_i \approx \frac{(\boldsymbol{Y}_0^i - \mu \sum_{j=1, j\neq i}^{m} (\boldsymbol{L}_{ij}^{\mathrm{v}} \boldsymbol{Y}_j^{\mathrm{T}}))(\boldsymbol{I} + \sum_{j=1}^{p} (-1)^j (\frac{\delta \boldsymbol{L}^{\mathrm{w}}}{\mu \boldsymbol{L}_{ii}^{\mathrm{v}} + 1})^j)}{\mu \boldsymbol{L}_{ii}^{\mathrm{v}} + 1}$$

6：Until 满足收敛准则

7：End For

8：Return $\boldsymbol{Y}$

在算法 2 的优化过程中，考虑到网络数据集规模较大，因此直接在原始矩阵 $\boldsymbol{Y}$ 上构造图和进行优化效率还是较低，因此实际求解过程中，我们将数据集按照语义相似度进行簇划分，不同簇之间认为语境不相关，即假设簇间相关度为 0。定义两个多媒体对象 o_i 与 o_j 的语义相似度为

$$s(\mathrm{o}_i, \mathrm{o}_j) = \frac{1}{2|\Omega_i|} \sum_{k=1}^{|\Omega_i|} \max_{t \in \Omega_j} sim(t_k^i, t) + \frac{1}{2|\Omega_j|} \sum_{k=1}^{|\Omega_j|} \max_{t \in \Omega_i} sim(t_k^j, t)$$

其中，$\Omega_i = \{t_k^i\}_{i=1}^{|\Omega_i|}$，为 o_i 的标签集合，t_k^i 为 o_i 标签集合中第 k 个标签。$\Omega_j = \{t_k^j\}_{i=1}^{|\Omega_j|}$ 为 o_j 的标签集合，标签与标签的距离采用 Google 距离。上式的定义满足下述条件：

(1) $s(o_i, o_j) = s(o_j, o_i)$，即满足对称性；

(2) $s(o_i, o_j) = 1$ if $\Omega_i = \Omega_j$，即两个多媒体对象的标签完全相同，则相似度为 1；

(3) $s(o_i, o_j) = 0$ if $s(t', t'') = 0 \quad \forall t' \in \Omega_i \quad \forall t'' \in \Omega_j$，即仅当两个对象没有相同标签，其相似度为 0。

对原始数据集聚类为若干个语义簇后，拉普拉斯矩阵 $\boldsymbol{L}^{\mathrm{v}}$ 和 $\boldsymbol{L}^{\mathrm{w}}$ 的维度降低较为明显。

3 实验结果与分析

采用 NUS-WIDE[14]与 Flickr 25 K[15]数据集。NUS-WIDE 数据集来自 Flickr 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 Web 中海量图像的真实情况，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。两个数据集的图像均来自 Flickr 图像共享网站。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 COLOR64 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，平均标签召回率，F1 值。

概念空间优化 SCR 的梯度下降求解过程需要确定步长，算法 1 中设定步长 $a \in [0,1]$，a 取值越小，达到稳定状态耗时越多，实验中设置 $a = \beta = 0.04$。图 1 记录了 SCR 的 F1 值随着迭代次数变化情况。从图中可以看到，随着迭代次数增加，F1 值增长态势明显，说明了概念空间的语境相关信息对于语义改善是有效的，而且当迭代次数大于 60 的时候，F1 较为稳定，因此在后续对比实验中，迭代次数固定为 60。

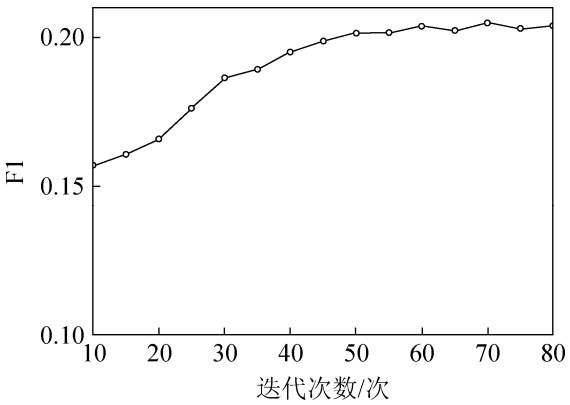


图 1 SCR 性能随迭代次数变化情况

算法 2 中，正则参数μ和δ用于平衡内容特征模态和文本模态，其比例对 EMR 性能影响较大。固定μ为 0.5，记录δ不同取值下算法的 F1 值，如图 2 所示。在曲线的大致中部位置，系统性能较高，模态互增强效果较为显著，而在δ小于 0.6 和大于 1.2 时，F1 值下降较为明显。这说明两者的比例固定为 0.5 较为合适，否则模态互增强的效果就会退化。因此，μ设定为经验值 0.5，δ设定为 1.0。算法 2 中簇数量设定为 120，常量 p 设定为 5。

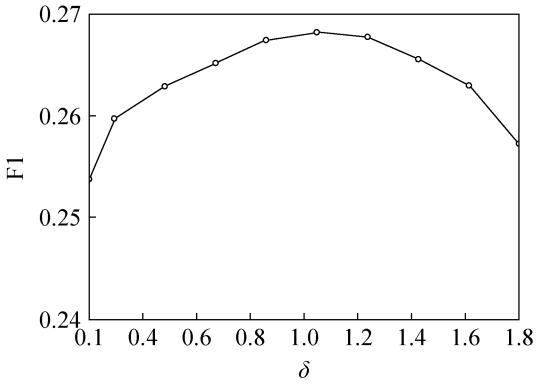


图 2 EMR 性能随δ变化变化情况(μ=0.5)

对 5 种标注优化方法进行比较，包括基于马尔可夫随机游走模型的标注优化 RWR[3]，基于语境相关传播的优化方法 SCR(Semantic context Refinement，算法 1)，基于内容的标注优化方法 CIAR[4]，基于标签相似图随机游走的优化方法 RWTR[7]，基于模态互增强的优化方法 EMR(Enhanced Multimodal Refinement，算法 2)。公平起见，我们以近邻相关学习[5]产生的标注结果为候选标签，然后对基本标注进行优化。表 1 给出了不同方法在 NUS-WIDE 数据集上优化后的性能对比结果。

表 1 评测集上的 F1 值

方法	F1
NV	0.160
RWR	0.186
SCR	0.203
CIAR	0.234
RWTR	0.244
EMR	0.268

SCR 利用语境相关图进行标签相关关系传播，与利用标签相似图上的随机游走的 RWR 相比较，性能提升 16.5%。但是，上述方法均忽略了媒体对象的内容特征。CIAR 利用了图像的内容信息，通过待优化图像的视觉特征定义标签相似度，以马尔可夫随机游走的稳定状态概率较大的标签作为最终改善结果，其性能较 SCR 提升 15.1%。RWTR 与 CIAR 均采用生成模型的优化框架，不同于 CIAR，RWTR 利用了图像的视觉内容信息和标签共现信息构造标签相关图，通过图上的随机游走得到最终的标注结果，其性能较 CIAR 提升 4.2%。EMR 方法取得最好的性能，说明内容特征模态和文本模态互增强机制是有效的。

算法 1 复杂度为 $O(m^2n)$，其中，m 为概念集规模，n 为图像集规模。对比方法中的 RWTR 的复杂度为 $O(mn^2)$，因此，如果概念集合固定，SCR 方法随优化图像的数量线性增长，对于大规模图像集合而言更为适用。SCR 运行时，可将图像标签矩阵分解，因此，SCR 更胜任大规模概念集上的标签相关性传播任务。随机从图像集中选取 18 000 幅图像，相似度矩阵和相关性矩阵均预计算并缓存。图 3 记录了 EMR 和 RWTR 的运行时间随数据集规模变化的情况。从图中可以看出 RWTR 对数据集规模近似呈现多项式复杂度 $O(n^{0.61})$，而 EMR 为近似线性复杂度 $O(n^{0.43})$，EMR 方法在运行效率上优于 RWTR。图 4 记录了随着簇数量变化的 EMR 方法的性能，随着簇数量的变化，EMR 算法性能变化幅度在 0.5%以内，因此 EMR 对簇数量较为稳定。由算法 2 可知，EMR 运行时间随着簇数量线性变化。对数据集进行聚类分割的策略忽略了簇间样本的相似性，如果簇数量过多，将导致系统性能下降。

图 5 记录了 EMR 性能随 k 的变化情况。从结果中可以看到，随着 k 的增长，相关图的邻接信息逐渐丰富，EMR 性能随之升高，但是当 k 超过 6 的时候，F1 增速放缓，因为邻接信息已经较为充分，而当 k 大于 10 后，EMR 性能逐渐降低，这是

https://dc-china-simulation.researchcommons.org/journal/vol28/iss11/29
DOI: 10.16182/j.issn1004731x.joss.201611029

第 28 卷第 11 期 系统仿真学报 Vol. 28 No. 11
2016 年 11 月 Journal of System Simulation Nov., 2016

因为邻域范围过大，会引入相应的噪声干扰并降低系统的性能。因此，实验中固定 k 为 10。

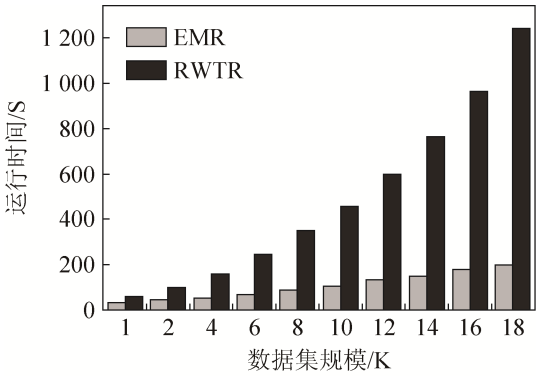


图 3 不同数据集规模下的运行时间比较

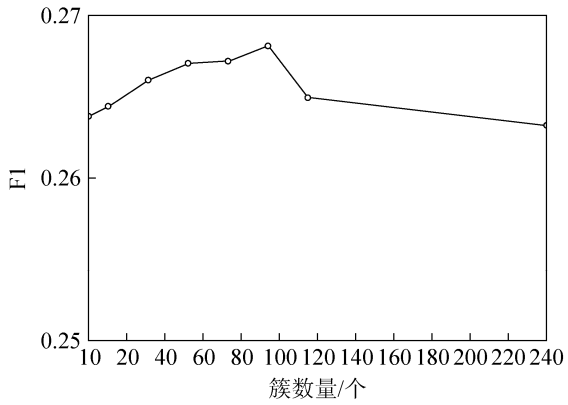


图 4 不同簇数量下 EMR 性能

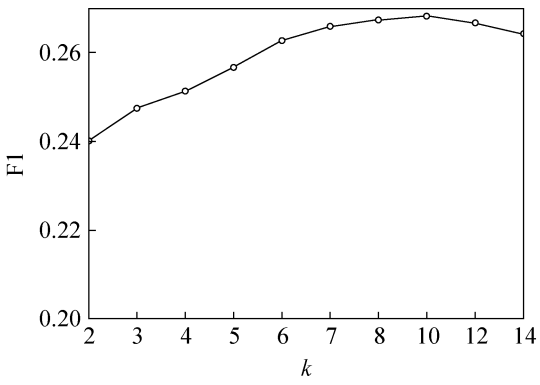


图 5 图稀疏性对性能影响

4 结论

本文提出了一种基于上下文语境传播的多媒体语义标注改善方法。构建标签语境相关图和多媒体对象内容相关图，将标注改善问题描述为正则化框架下的优化问题，通过该问题的求解获得改善的标注结果。该方法是一种数据驱动的语义改善方法，充分利用了数据集蕴含的上下文语境相关信息，保证了多媒体对象的内容相关性和标签的语义关联性。实验结果表明，该方法适用于大规模数据集上的语义改善任务，而且采用数据集聚类和约束放松的策略可进一步提高模型的标注效率，其性能也优于其它对比方法。

参考文献：

[1] Jin Y H, Khan L, Wang L. Image annotations by combining multiple evidence & Wordnet [C]// Proceedings of the 13th annual ACM international conference on Multimedia. New York, USA: ACM, 2005: 706-715.

[2] Jin YH, Khan L, Prabhakaran B. Knowledge Based Image Annotation Refinement [J]. Journal of Signal Processing Systems, 2010, 58(3): 387-406.

[3] Wang CH, Feng J, Zhang L, et al. Image annotation refinement using random walk with restarts [C]// Proceedings of the ACM International Conference on Multimedia. New York, USA: ACM, 2006: 647-650.

[4] Wang CH, Feng J, Zhang L, et al. Content-Based Image Annotation Refinement [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE Computer Society, 2007: 1-8.

[5] Li XR, Snoek C.G.M, Worring M. Learning social tag relevance by neighbor voting [C]// IEEE Transactions on Multimedia. USA: IEEE, 2009: 1310-1322

[6] Chen L, Xu D, Tsang IW. Tag-based Web Photo Retrieval Improved by Batch Mode Re-Tagging [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE Computer Society, 2010: 3440-3446.

[7] Liu D, Hua XC, Yang LJ, et al. Tag ranking [C]// Proceedings of the Intenational World Wide Web Conference. New York, USA: ACM, 2009: 351-360.

[8] Xu H, Wang JD, Hua XS. Tag Refinement by Regularized LDA [C]// Proceedings of the ACM Multimedia Conference. New York, USA: ACM Press, 2009: 573-576.

[9] Zhu GY, Yan SC, Ma Y. Image tag refinement towards low-rank, content-tag prior and error sparsity [C]// Proceedings of the ACM Multimedia Conference. New York, USA: ACM Press, 2010: 461 -470.

（下转第 2877 页）

