基于 VSM 和 LDA 模型相结合的新闻文本分类研究

彭雨龙
（厦门大学，福建 厦门 361005）

摘　要：针对传统 KNN 算法在处理新闻分类时仅仅考虑文字层面上的相似性，而未涉及语义层面，本文提出了一种基于 VSM 和 LDA 模型相融合的新闻分类算法。首先，在深入研究 VSM 和 LDA 模型的基础上，对新闻文档进行 VSM 和 LDA 主题建模，结合 LDA 模型与 VSM 模型计算文档之间的相似度；其次，以复合相似度运用到基于相似度加权表决的 KNN 算法对新闻报道集合进行分类。实验验证了改进后的相似度计算方法的有效性，实验结果表明改进后的 KNN 算法与传统算法相比，具有较好的效果。
关键词：潜在狄利克雷分布（ LDA）；向量空间模型 (VSM)；文本相似度；KNN 分类
DOI：10.16640/j.cnki.37-1222/t.2016.06.192

1　引言

目前，面对着互联网上各种各样、数量繁多的新闻网页，人们不知道如何选择自己需要和喜爱的新闻。因此，人们越来越迫切地需要一个对新闻进行分类的工具，能够用来快速浏览自己需要的新闻内容。

常见的文本分类技术包括 KNN 算法、贝叶斯算法、支持向量机 SVM 算法以及基于语义网络的概念推理网算法等。KNN 算法在新闻等网页文本分类中有着广泛的应用，他的思想是对于待分类的文本，通过由与该样本最接近的 K 个样本来判断该样本归属的类别 [1]。

本文针对传统 KNN 算法在度量文本相似性时仅仅考虑文字层面的相似性，而未涉及语义层面。首先，对新闻文档进行 VSM 和 LDA 主题建模，结合 LDA 模型与 VSM 模型计算文档之间的相似度；其次，以复合相似度运用到基于相似度加权表决的 KNN 算法对新闻报道集合进行分类。

2　相关工作

2.1　向量空间模型

向量空间模型(VSM：Vector Space Model)由 G.Salton、A. Wong、C. S. Yang[2] 等人于 20 世纪 70 年代提出。向量空间模型（VSM）以特征词作为文档表示的基本单位，每个文档都可以表示为一个 n 维空间向量：$T(F_1,W_1;F_2,W_2;\cdots;F_n,W_n)$，简记为 $T(W_1,W_2,\cdots,W_n)$，F_i 为文档的特征词，W_i 为每个特征词的权重，则 $T(W_1,W_2,\cdots,W_n)$ 为文本 T 的向量表示 [3]。特征词的权重值一般采用 TF*IDF 来计算。

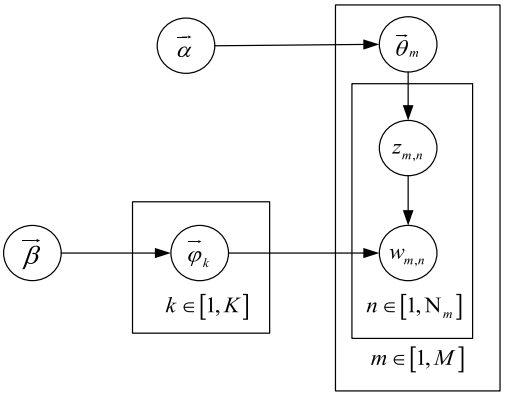

图 1　LDA 概率图模型表示

向量空间模型把文本内容用 n 维空间向量表示，把对文本内容的处理简化为向量空间中的向量运算，并且它以空间上的相似度表达语义的相似度，直观易懂，但向量空间模型并没有考虑到特征词之间的语义关系，可能丢失很多有用的文本信息。

2.2　LDA 主题模型

2.2.1　LDA 主题模型基本思想

主题模型是统计模型的一种，用来发现在文档集合中的抽象主题。LDA（Latent Dirichlet Allocation）是一种文档主题生成模型，也称为一个三层贝叶斯概率模型，包含词、主题和文档三层结构。首次是作为概率图模型由 David Blei、Andrew Ng 和 Michael Jordan 于 2003 年提出 [4]，图 1 为 LDA 的概率图模型。

其中 M 为文档总数，K 为主题个数，N_m 是第 m 个文档的单词总数，β 是每个 Topic 下词的多项分布的 Dirichlet 先验参数，α 是每个文档下 Topic 的多项分布的 Dirichlet 先验参数。$z_{m,n}$ 是第 m 个文档中第 n 个词的主题，$w_{m,n}$ 是第 m 个文档中的第 n 个词。隐含变量 θ_m 和 ψ_k 分别表示第 m 个文档下的 Topic 分布和第 k 个 Topic 下词的分布，前者是 k 维 (k 为 Topic 总数) 向量，后者是 v 维向量 (v 为词典中词项总数)。

2.2.2　Gibbs 抽样

Gibbs Sampling 是马尔科夫链蒙特卡洛算法的一个实例。该算法每次选取概率向量的一个维度，给定其他维度的变量值采样当前维度的值，不断迭代至收敛输出待估计的参数 [5]。

从 2.2.1 中可知，$z_{m,n}$、θ_m 和 ψ_k 变量都是未知的隐含变量，也是我们需要根据观察到的文档集合中的词来学习估计的。

学习步骤如下：

（1）应用贝叶斯统计理论中的标准方法 [6]，推理出有效信息 P(w|T)，确定最优主题数 T，使模型对语料库数据中的有效信息拟合达到最佳。

（2）初始时为文本中的每个词随机分配主题 $Z^{(0)}$，统计第 z 个主题下的词项 t 的数量，以及第 m 篇文档下出现主题 z 中的词的数量。

（3）每一轮计算 $p(z_i|z_{-i},d,w)$ 这里 i=(m,n) 是一个二维下标，对应于第 m 篇第 n 个词，即排除当前词的主题分配，根据其他所有词的主题分配估计当前词分配给各个主题的概率，根据这个概率分布，为该词采样一个新的主题 $Z^{(1)}$。同样更新下一个词的主题。直到每个文档下 Topic 分布 θ_m 和每个 Topic 下词的分布 ψ_k 收敛。

$p(z_i|z_{-i},d,w)$ 称 Gibbs 更新规则，计算公式如下：

$$p(z_i = k \mid \vec{z}_{\neg i}, \vec{w}) \propto \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\sum_{k=1}^{K}(n_{m,\neg i}^{(t)} + \alpha_k)} \cdot \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V}(n_{k,\neg i}^{(t)} + \beta_t)}$$

3 基于 VSM 和 LDA 模型的新闻分类

3.1 基于 VSM 和 LDA 模型的文本相似度计算

（1）对于文档 d_i，d_j，由向量空间模型（VSM）进行预处理，得到的文本的特征词向量 $d_{i\_VSM}=(w_1,w_2,\cdots,w_N)$ 和 $d_{j\_VSM}=(w'_1,w'_2,\cdots,w'_N)$，N 为特征词个数。

（2）由 LDA 模型进行预处理，得到文本 - 主题向量为 $d_{i\_LDA}=(t_1,t_2,\cdots,t_k)$ 和 $d_{j\_LDA}=(t'_1,t'_2,\cdots,t'_k)$，K 为主题个数。

（3）由公式（3-1）计算基于向量空间模型的相似度，取两个向量之间的夹角余弦：

$$Sim_{VSM}(d_i,d_j)=\frac{d_{i\_VSM}\times d_{j\_VSM}}{\left|d_{i\_VSM}\right|\times\left|d_{j\_VSM}\right|} \quad (3\text{-}1)$$

（4）由公式（3-2）计算基于 LDA 模型相似度，同样取两个向量之间的夹角余弦：

$$Sim_{LDA}(d_i,d_j)=\frac{d_{i\_LDA}\times d_{j\_LDA}}{\left|d_{i\_LDA}\right|\times\left|d_{j\_LDA}\right|} \quad (3\text{-}2)$$

（5）将二者线性组合，得到最终文档 d_i、d_j 的复合相似度为 [7]：

$$Sim(d_i,d_j)=\lambda\times Sim_{VSM}(d_i,d_j)+(1-\lambda)\times Sim_{LDA}(d_i,d_j),\lambda\in(0,1) \quad (3\text{-}3)$$

3.2 基于 VSM 和 LDA 模型的新闻文本分类

本文改进的 KNN 算法的具体过程如下 [8]：

输入：待分类新闻文本 d 和已知类别的新闻文本 D；

输出：待分类新闻文本 d 的可能类别。

（1）对 d 和 D 集合进行预处理，构建其特征向量和主题向量；

（2）对 d 中的每个新闻文本，采用公式（3-3）计算其于 D 中每个新闻文本的相似度；

（3）从中选择与 d 相似度最大的 K 个文本；

（4）对于待分类文本的 K 个邻居，依次按公式 (3-4) 进行计算 d 隶属每个类别的权重。

$$W(d)=\sum T_j(d_i)*Sim(d,d_i) \quad (3\text{-}4)$$

其中，y 表示 d 的特征向量，$T_j(d_i)$ 表示指示函数，指示是否是同一类别，即 d_i 是否属于 Cj，若是，则值为 1，否则为 0。$Sim(d,d_i)$ 表示待分类文本与邻居 di 的复合相似度。

(5) 比较每个类的权重，将权重最大的类别定为 d 的类别。转入 (2) 直至所有待分类文本分类完成。

4 实验结果及分析

4.1 文本分类的性能评价

评价文本分类算法的有两个指标：准确率（Precision）和召回率（Recall）。由于准确率和召回率是分别从两个不同的方面来评价分类效果，所以一般采用 F_measure 来评估分类效果，如公式 4-1。

$$F\_measure=\frac{2*Precision*Recall}{Precision+Recall} \quad (4\text{-}1)$$

4.2 文本分类实验结果及分析

本实验语料采用搜狗实验室文本分类语料库，选取军事、体育、旅游、教育、娱乐、财经六个类别，每个类别下挑选 200 篇文章，总共 1200 篇，其中训练集占 1/3，首先，针对不同的 K 值下的分类效果找出最佳的 K 值，然后，对传统 KNN 算法和基于相似度加权的 KNN 算法进行对比试验。传统的 KNN 算法的权重计算方法如公式 4-2 所示：

$W(d)=\sum T_j(d_i)*Sim_{VSM}(d,d_i)$（$Sim_{VSM}(d,d_i)$ 为公式 3-1 所求）(4-2)

最终确定实验的参数如下：KNN 的 K 值取 20，主题数 K=30，Dirichlet 先验参数选取经验值 α =1，β =0.01，Gibbs 抽样次数设为 5000；VSM 和 LDA 模型线性结合参数 λ 设置为 0.8，实验效果如图 2 所示。

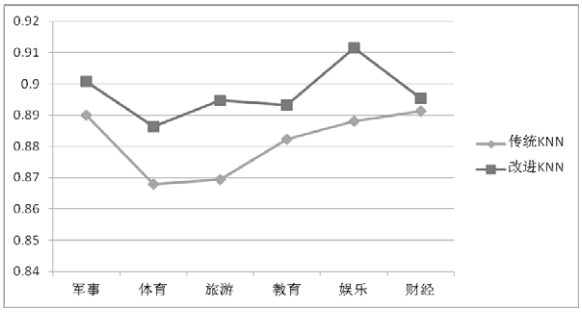

图 2 改进后的 KNN 分类算法与传统 KNN 分类算法的 F_measure 值

从图 2 中可以看出，改进后的 KNN 分类算法在军事、体育、旅游、教育、娱乐、财经六个方面都较传统 KNN 分类算法好一些，因为，传统 KNN 算法只是单纯第从文字层面来计算两段文本之间的距离，而将 VSM 结合 LDA 模型后，既可以较完整地保留文本的信息，又可以提取语义层面的信息，这样能更精确地计算两段文本之间的相似度。

5 总结与展望

本文提出了基于 VSM 和 LDA 模型相结合的 KNN 分类算法，与传统 KNN 分类算法相比，引进了 LDA 模型，从而在计算两段文本之间的距离时融合了语义层面的相似度，在相似度计算方法上进行了改进，实验也验证了改进后算法的有效性。

由于当前所用的中文语料库还有待完善，本文选用的搜狗实验室文本语料库，主题数较少，使得 LDA 主题模型的作用不太明显，后续将考虑使用爬虫程序从各大新闻网站上选取一些语料库的来源。

参考文献：

[1] 张宁．使用 KNN 算法的文本分类 [J]．计算机工程，2005(04)．

[2]G.Salton,A.Wong,C.S.Yang.A Vector Space Model for Automatic Indexing[J].Communications of the ACM: Volume 18 Issue 11,1975(11).

[3] 王萌，何婷婷，姬东鸿，王晓荣．基于 HowNet 概念获取的中文自动文摘 [J]．中文信息学报，2005,19(03):87-93.

[4]Blei D M, Ng A Y, Jordan M I.Latent dirichlet allocation[J].the Journal of machine Learning research, 2003(03):993-1022.

[5] 赵爱华，刘培玉，郑燕．基于 LDA 的新闻话题子话题划分方法 [J]．小型微型计算机系统，2013(04)．

[6] 董婧灵，李芳，何婷婷．基于 LDA 模型的文本聚类研究 [G].2011.

[7] 王爱平，徐晓艳，国玮玮，李仿华．基于改进 KNN 算法的中文文本分类方法 [J]．微型机与应用，2011(18)．

作者简介：彭雨龙（1989-），男，湖南邵阳人，硕士研究生，研究方向：数据挖掘。