计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(3):255−259 [doi: 10.15888/j.cnki.csa.006829]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

改进主题模型的短文本评论情感分析①

花树雯, 张云华

(浙江理工大学 信息学院, 杭州 245000)
通讯作者: 花树雯, E-mail: hswlian@qq.com

摘　要: 使用传统的主题模型方法对医疗服务平台中的评论等短文本语料进行主题模型的情感分析时, 会出现上下文依赖性差的问题。提出基于词嵌入的 WLDA 算法, 使用 Skip-Gram 模型训练出的词 *w**替换传统的 LDA 模型中吉布斯采样算法里的词 *w*`, 同时引入参数 *λ*, 控制吉布斯采样时词的重采样的概率. 实验结果证明, 与同类的主题模型相比, 该主题模型的主题一致性高.

关键词: 情感分类; 短文本; 词嵌入; WLDA

引用格式: 花树雯,张云华.改进主题模型的短文本评论情感分析.计算机系统应用,2019,28(3):255–259. http://www.c-s-a.org.cn/1003-3254/6829.html

Short Text Comment Sentiment Analysis of Improved Topic Models

HUA Shu-Wen, ZHANG Yun-Hua

(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: When the traditional topic model method is used to analyze the sentiment of the topic model for short text corpora such as comments in the medical service platform, the problem of poor context dependency may occur. A WLDA algorithm based on word embedding is proposed. The word *w** trained in the Skip-Gram model replaces the word *w*` in the Gibbs sampling algorithm in the traditional LDA model, and the parameter λ is introduced to control the resampling probability of the words during Gibbs sampling. The experimental results show that the subject model has a high degree of consistency compared to similar topic models.

Key words: sentiment classification; short text; word embedding; WLDA

引言

2016 年, Li 等人根据评论语料中的时间、发布人等信息, 为短文本分配不同的权重, 将分配权重后的短文本合并为伪长文本, 将 LDA 模型中的单词 *w* 替换成权重微博链组成的三元组形式<*w*,flag_t,flag_s>, 提出了使用微博链改进的 LDA 主题模型 (WMC-LDA) 对短文本进行分类[1]. 2017 年, Liu 等人尝试使用与训练语料相关的外部语料库进行词嵌入模型的训练, 学习到词语间的语义关系, 作为高斯 LDA 对短文本分析时的词向量的扩充[2]. 2018 年, Bunk 等人提出了 WELDA 模型, 将提取词的先验语义信息的词嵌入模型运行在 LDA 模型词采样的内层, 基于训练语料的词义增强主题模型的训练[3].

综合目前的研究, 现有的短文本主题分类有以下两点不足:

(1) 传统通过利用外部语料扩充词义或者合并短文本的方法提高语料的语义信息, 但是主题模型对训练语料中的词义信息提取不充分.

(2) 主题模型中词嵌入空间的词向量的能力有限, 词嵌入模型运行在吉布斯采样的内层时, 模型的运行效率十分缓慢.

上述存在的问题, 则是本文开展研究的出发点.

① 收稿时间: 2018-09-28; 修改时间: 2018-10-23, 2018-10-29; 采用时间: 2018-10-31; csa 在线出版时间: 2019-02-22

1 相关工作

1.1 LDA 主题模型

LDA 主题模型是 Blei 等人在 03 年提出的, 模型为文档集中的每个文档以概率分布的形式分配多个主题, 每个单词都由一个主题生成[4], LDA 的模型如图 1 所示.

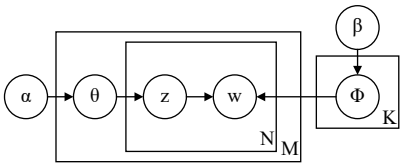

图 1 LDA 模型结构图

图 1 中, α 和 β 表示先验参数, θ 表示从先验参数 α 中提取的主题分布, z 表示从 θ 主题分布中提取的主题, Φ 表示从先验参数 β 中提取的主题 z 对应的词语分布, w 为最后生成的词[5].

LDA 模型中, 词 w 采样是根据主题 z 和模型的先验参数 β, 主题 z 是从先验参数 α 中提取, 所以他们的联合概率分布如式 (1) 所示.

$$p(w,z|\alpha,\beta)=p(w|z,\beta)\,p(z|\alpha) \tag{1}$$

在模型中先验参数 β 服从关于参数 Φ 独立的多项分布, 使用参数 Φ 将式 (1) 更新如下:

$$p(w|z,\Phi)=\prod_{i=1}^{|w|}p(w_i,z_i)=\prod_{i=1}^{|w|}\phi_{s_i,w_i} \tag{2}$$

因为词服从于主题即参数为 w 的多项分布, 所以将上式展开化解如下:

$$p(w|z,\Phi)=\prod_{k=1}^{k}\prod_{t=1}^{v}p(w_i,z_i)\phi_{w,t}^{n_k^w} \tag{3}$$

式中, n_k^w表示 w 在主题 k 中出现的次数, 2.2 节基于上式进行算法改进.

1.2 词嵌入模型和 LDA 模型的对比

词嵌入模型认为可以将语料中的每个单词分配给高维向量空间的实际向量, 通常这个向量空间可以包含 50 到 600 个维度. 提出了 Word2Vec 模型, 在训练过程中, 滑动窗口将覆盖文本和神经网络中的每一个单词的权重以学习预测周围的单词, 通过 PCA 降维, 投射出词嵌入模型和 LDA 模型的两个维度的单词嵌入空间, 通过可视化方法使得词的距离更容易理解. 两点之间的距离越短, 表示词义越相近, PCA 的降维结果如图 2 所示.

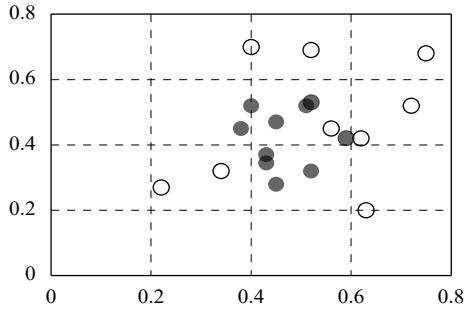

图 2 词向量 PCA 图

选取 LDA 模型中前 10 个单词, 在图 2 中用实心点表示, 空心点表示词向量模型训练出的词向量, 由图可以得出, 实心点在距离上更近, 而空心点之间的距离比实心点较远, 说明词向量训练出的词在词义上更近. Batmanghelich 等人在 NSTM 模型中提出词义的相似性可以通过词向量 $(x_1,x_2,x_3,\cdots,x_n)$ 的余弦距离 cos 来衡量, 余弦的计算如式 (4) 所示.

$$cos=\frac{\sum_{k=1}^{n}x_{1k}x_{2k}}{\sqrt{\sum_{k=1}^{n}x_{1k^2}}\sqrt{\sum_{k=1}^{n}x_{2k^2}}} \tag{4}$$

Batmanghelich 等人的实验证明这种衡量方式, 比通过嵌入模型中的欧几里得距离衡量要准确[6].

2 WELDA 模型的建立

2.1 替换词向量模型构建

词语的关系有相似性和相关性, 语义的相似性关系例如词语‘医生’和‘大夫’, 相关性例如词语‘医生’和‘护士’. 基于词嵌入的模型关注于语义的相似性, 而基于文档的主题模型则擅长捕捉语义的相关性. 考虑到实验的数据量并不十分巨大, 因此使用的 Skip-Gram 模型进行模型的构建.

(1) 语料库通过 Skip-Gram 模型进行词向量训练, Skip-Gram 模型能很好的表示相似的词汇, 使用余弦距离的值计算表示词义的相似性.

表 1 表示实验中在 Skip-Gram 模型下输入语料库后抓取的‘复查’词义相近的词汇.

(2) 模型中, 替换单词 w 的具体做法是, 从 Skip-Gram 模型空间中抽取一个与 w`相近的词向量 w^*,

w*是词嵌入空间中产生的余弦距离上最近的单词, 最后, 替换单词 w`. 例如, 对上文中的‘复查’来说, 替换词新词是‘复诊’.

(3) 借鉴 LFTM 模型的方法, 替换词向量模型时引入了伯努利参数 s~ber(λ), 词的采样可以以一定概率从从词嵌入空间 v 或者从主题分布的词语分布 Φ 中进行采样[7].

表 1 ‘复查’的相近词向量余弦距离示例

词汇	余弦距离
复诊	0.6527
就诊	0.6255
拆线	0.5923
输液	0.5137
抽血	0.5003

替换词向量模型认为, 当参数 λ 为 0 时, 模型则退化为传统的 LDA, 直接从主题分布中提取词, 而参数 λ 不等于 0 时, 以一定概率从词嵌入空间中提取新词, 参数 λ 决定了单词的额外信息泄漏到主题模型中的程度.

2.2 WLDA 模型构建

在 WLDA 模型中, 首先将预处理文本输入到替换词向量模型层 v, 得到训练好的词嵌入空间. 其次, 在模型中加入替换词向量模型层, 最后, 将词 w`输入替换词向量模型层, 模型的结构图如图 3 所示.

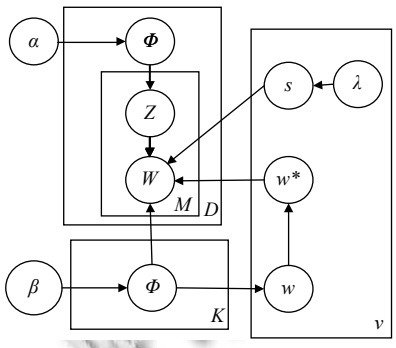

图 3 WLDA 模型结构图

WLDA 模型生成过程如下:

(1) 选择文档集合中的主题 k=1,…, k;

(2) 选择单词分布 Φ_k~Dir(β);

(3) 对每篇文档 d=1,…, M:

1) 生成文档主题分布 θ_d~Dir(α);

2) 对文档中的每个词 i=1,…,N_d;

① 生成词的主题 z_{di}~Mult(θ_d);

② 选择 w~Mult(Φz_{di}), $\Psi_{d,i}$~Ber (λ), 如果 $\Psi_{d,i}$=1, 替换新单词 w^*.

替换词 w 为在上述替换词向量模型中抓取相似的单词 w^*, 用$n_{-i,j}^{(w_i)}$表示 w_i 被分配给话题 j 的次数, 根据步骤 a 中得到的公式, 以及贝叶斯法则和 Diri 先验, 将公式推导如下.

$$x_{w,j}=\frac{n_{-i,j}^{(w_i)}+\beta}{n_{-i,j}^{(\cdot)}+w\beta} \tag{5}$$

同理用表示$n_{-i,j}^{(d_i)}$是文档 d_i 被分配给话题 j 的次数. $p(z|\alpha)$ 推导计算, 得出公式:

$$\theta_{d,j}=\frac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i.}^{(d_i)}+k\alpha} \tag{6}$$

更新吉布斯采样器如式 (7) 所示.

$$p(z_{d,i}=k|z_{-(d,i)},w)\propto\frac{n_{-i,j}^{(w_i)}+\beta}{n_{-i,j}^{(\cdot)}+w\beta}\frac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i.}^{(d_i)}+k\alpha} \tag{7}$$

其中, 基于伯努利分布, 从替换词向量模型层 v 中采样词 w^*, 交换当前单词 w`的新主题的分布, 由于词向量训练并不运行在吉布斯采样的内层, 而是在词向量模型训练好之后, 主题模型在词采样阶段从词嵌入空间中以一定概率提取词义相近的词进行替换.

由此在理论上来说, 词的替换使该模型的主题的困惑度下降, 而在外部训练好词嵌入空间, 使 WLDA 模型的运行效率更高.

3 实验与分析

3.1 实验环境及数据预处理

实验硬件环境为酷睿 i7 处理器, 运行内存为 16 GB, 操作系统为 Win10, 实验的软件是 Eclipse,采用的语言是 Python.

实验数据处理分为以下两步:

(a) 在挂号网上爬取出评论数据, 去除标点符号.

(b) 使用结巴分词, 进行停用词处理和将语料库进行分词.

分词得到的 txt 局部文本如图 4 所示.

3.2 实验内容及结果分析

实验分为 2 个部分.

(a) 配置 λ 参数, 找出合适的重采样概率 λ.

(b) 基于 WLDA 的进行情感词抽取并和其他模型进行实验对比.

实验中我们采用 Perplexity(困惑度) 值作为评判标准, 式 (8) 为 Perplexity 的计算公式[7].

$$Perplexity(D_{\text{test}}) = \exp\left\{-\frac{\sum_{D=1}^{M}\log p(w_d)}{\sum_{D=1}^{M}N_D}\right\} \quad (8)$$

其中, M 代表测试预料集的文本数量, N_d 代表第 d 篇文本的大小 (即单词的个数), $p(w_d)$ 代表的是文本的概率[8]. 如果重采样的参数等于 1, 则实验中使用的为标准的 LDA, 当重采样次数等于 0 时, 文档中所有的词全部是从词嵌入的空间中抽取. Perplexity 对比的数据如图 5 所示.

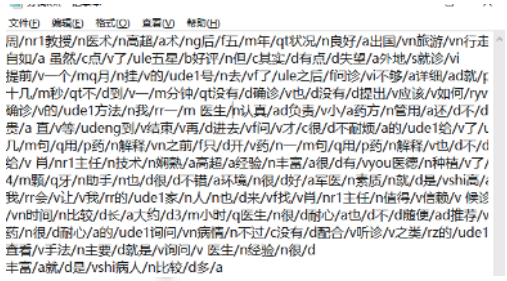

图 4 分词得到的 txt 文本局部图

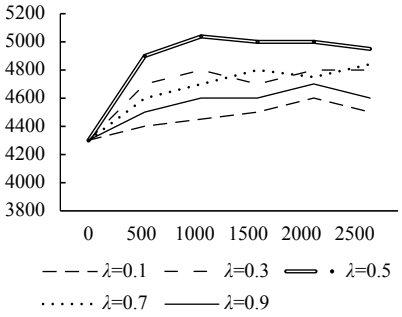

图 5 Perplexity 值对比

图 5 中的 λ 为重采样次数, 横坐标为模型的迭代次数, 纵坐标为困惑度, 实验得出当收敛次数需要小于 1000 次, 重新采样次数为 0.5 时, 模型的困惑度较小.

DMM 模型通过假设每个短文本只包含一个主题[8], 15 年, das 等人首次提出了高斯 LDA 模型, 使用词向量代替离散的值[9], 这两个模型都在一定程度上, 解决了短文本的上下文依赖性差的问题. 实验选择 DMM 模型, 高斯 LDA 模型和重采样概率为 0.5 的 WLDA 模型进行对比.

针对测试的评论数据, 使用 PMI 来量化这三个主题模型中的主题质量. PMI(主题一致性标准) 常常被用来量化主题模型中的主题的质量, PMI 的定义如式 (9) 所示[9].

$$PMI-Score(k) = \frac{1}{T(T-1)}\sum_{1<i<j<T}PMI(w_i,w_j) \quad (9)$$

其中, $PMI(w_i,w_j) = \log\frac{p(w_i,w_j)}{p(w_i)p(w_j)}$, $p(w_i)$ 是单词在外部语料文档中出现的概率, PMI 的值越大, 说明该模型训练出的主题的相关性越强.

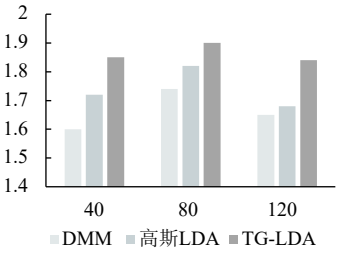

图 6 模型的 PMI 对比

实验结果表明, WLDA 模型的表现要优于高斯 LDA 模型模型, 困惑度最小, 这一点得益于 WLDA 在吉布斯采样阶段, 选择词嵌入空间的词向量 w^*, 对单词 w`选择性替换, 而替换的词向量提高了模型训练中词向量的相似性, 补充了上下文的语义, 当模型中的主题数为 120 时, 模型的 PMI 值变低, 是由于替换的词向量的质量不高, 对短文本的主题学习造成了影响.

运行时间如表 2 所示.

表 2 运行时间表 (单位: min)

方法	运行时间
高斯 LDA	4283
WLDA,λ=0.5	206
DMM	200

DMM 模型的运行时间最短, 但是由于 DMM 模型假设每个短文本只包含一个主题, 这个假设十分不严谨, 因此, DMM 的 PMI 值远远小于 WLDA 模型.

4 结束语

本文提出了一种基于主题模型的短文本评论情感分析模型, 通过在某医院的评论数据上实验, 证明了该模型对主题词的分类更加的突出, 并且有较高的主题一致性.

在下一步工作中, 将进一步研究降低模型的时间复杂度, 提高模型的运行效率.

参考文献

1 李鹏, 于岩, 李英乐, 等. 基于权重微博链的改进 LDA 微博主题模型. 计算机应用研究, 2016, 33(7): 2018–2021. [doi: 10.3969/j.issn.1001-3695.2016.07.021]

2 Das R, Zaheer M, Dyer C. Gaussian LDA for topic models with word embeddings. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China. 2015. 167–182.

3 Bunk S, Krestel R. WELDA: Enhancing topic models by incorporating local word context. Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries. Fort Worth, TX, USA. 2018. 78–86.

4 韩忠明, 李梦琪, 刘雯, 等. 网络评论方面级观点挖掘方法研究综述. 计算机学报, 2018, 29(2): 417–441.

5 彭三春. 基于 RNN 和 LDA 模型的商品评论情感分类研究[硕士学位论文]. 杭州: 浙江理工大学, 2018.

6 黄发良, 于戈, 张继连, 等. 基于社交关系的微博主题情感挖掘. 软件学报, 2017, 28(3): 694–707.

7 李勇敢, 周学广, 孙艳, 等. 中文微博情感分析研究与实现. 软件学报, 2017, 28(12): 3183–3205.

8 Wei X, Sun J, Wang X. Dynamic mixture models for multiple time series. Proceedings of the 20th International Joint Conference on Artifical Intelligence. Hyderabad, India. 2007. 2909–2914.

9 夏友青. 基于 LDA 的在线主题演化模型研究与优化[硕士学位论文]. 长沙: 国防科学技术大学, 2012.