计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201200082

基于高斯分布的改进词嵌入主题情感模型

李玉强[1] 张伟江[1] 黄 瑜[1] 李 琳[1] 刘爱华[2]
1 武汉理工大学计算机科学与技术学院 武汉 430063
2 武汉理工大学能源与动力工程学院 武汉 430063
(liyuqiang@whut.edu.cn)

摘 要 近年来,主题情感联合模型成为了无监督学习领域的一项重要研究内容,在文本主题挖掘和情感分析等方面均有实际应用。然而,在现实场景中,微博因其文字短小、结构不完整等特征,给主题情感联合模型带来了一定的挑战。因此,围绕微博主题情感模型展开相关的研究与改进工作,目前较为流行的主题情感模型——TSMMF模型(Topic Sentiment Model Based on Multi-feature Fusion)中引入了词向量技术,运用多元高斯分布从词向量空间中快速采样邻近词语,并替换掉原Dirichlet多项式分布产生的单词,从而将共现频率低、信息量少的单词转变成突出主题、信息明确的单词,同时使用最近邻搜索算法来进一步提升模型处理大型微博语料库的运行速度,进而提出了GWE-TSMMF模型。对比实验结果表明,GWE-TSMMF模型的平均$F1$值约为0.718,相比原模型和现有的主流词嵌入主题情感模型(WS-TSWE模型和HST-SCW模型),其微博情感极性的分析效果均有显著提升。

关键词:主题情感模型;高斯分布;词嵌入;微博情感极性分析
中图法分类号 TP391

Improved Topic Sentiment Model with Word Embedding Based on Gaussian Distribution

LI Yu-qiang[1],ZHANG Wei-jiang[1],HUANG Yu[1],LI Lin[1] and LIU Ai-hua[2]
1 School of Computer Science and Technology,Wuhan University of Technology,Wuhan 430063,China
2 School of Energy and Power Engineering,Wuhan University of Technology,Wuhan 430063,China

Abstract In recent years,the topic sentiment model as an important research in the field of unsupervised learning,has been used in text topic mining and sentiment analysis. However,Weibo has brought some challenges to the topic sentiment model because of its short text and in complete structure. Therefore,the related research and improvement work of this paper will be carried out around the topic sentiment model of Weibo. We introduce the word vector technology to the popular model-TSMMF(topic sentiment model based on multi-feature fusion),use multivariate Gaussian distribution to sample neighboring words fast from the word embedding space,and replace the words generated by the Dirichlet multinomial distribution. Thus,the words with low cooccurrence frequency and less information will be transformed into words with prominent topic and clear information. At the same time,the nearest neighbor search algorithm is used to further improve the running speed of the model when processing large-scale Weibo corpus,and then the GWE-TSMMF model is proposed. The experimental results show that the average $F1$ value of GWE-TSMMF model is about 0.718. The sentiment polarity analysis is better than the original model and the existing mainstream word embedding topic sentiment models (WS-TSWE and HST-SCW).

Keywords Topic sentiment model,Gaussian distribution,Word embedding,Weibo sentiment polarity analysis

1 引言

以微博为代表的新兴社交媒体以其独特的开放性、实时性、互动性为人们表达意见和交流看法提供了一个良好的媒介。越来越多的人愿意在社交媒体上表达自己对某些事情的看法,这些信息已经成为挖掘人们观点和情感的重要资源,在舆情监控、企业决策、智能推荐等方面有着重要作用。微博情感分析的巨大价值促进了多种微博情感分析方法的产生,传统的微博情感分析方法主要包括基于规则和基于机器学习两类[1-2]。基于规则的方法主要通过人工整理出情感词典和分类规则来计算文本的情感得分,从而判断文本的情感极性。这种方法虽然取得了一定成效,但是需要依赖复杂的人工特征干预,花费了高额的人工成本,增加了任务的复杂度,而且该方法无法直接迁移到其他领域的任务中,因此具有很大的

到稿日期:2020-12-08 返修日期:2021-03-13 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家社会科学基金项目(15BGL048)
This work was supported by the National Social Science Foundation of China(15BGL048).
通信作者:张伟江(913114863@qq.com)

局限性。基于机器学习的方法主要由计算机根据某种特定的算法对文本进行处理后输出情感分类。这类方法不需要依赖复杂的人工规则，在一定程度上降低了人工成本，但是也需要对训练文本进行人工标注，且对于特定的领域需要相关的专业知识，因此也投入了不少的人力物力。另外，微博情感极性与其讨论的主题是密切相关的，上述的微博情感分析方法始终没有将微博主题和微博情感进行协同分析，导致脱离主题的情感挖掘效果不够理想。2008 年，Pang 等[3]指出文档中的主题信息对于情感分类非常重要。随后，情感分类的研究重心逐渐转向考虑主题的情感生成模型。主题情感联合模型综合考虑了文本中的情感因素和主题因素，通过使用变量及分布定义文档的主题信息，并引入情感层来捕捉文档的情感信息，进而推断出文档的情感极性。之后，许多研究工作围绕着主题情感联合模型展开并取得了巨大的成就[4-5]。

但是，基于主题情感联合模型的微博情感分析方法仍存在一定的局限性。一方面，在实际的场景中，能用于解决问题的数据通常是有限的，其原因在于收集一个庞大的数据集的代价非常昂贵。而传统的主题情感联合模型仅通过微博中单词的共现来推断情感和主题分布，当微博的训练语料库规模很小时，这些主题情感模型不能有效挖掘低频词语的语义关系，导致推断出的情感-主题分布不尽如人意。另一方面，主题情感模型是基于词袋假设的，根据当前文档中的情感-主题-词语分布对下一个单词进行全局预测，但未考虑上下文语义信息，使主题的语义一致性不够，而微博的情感与主题又是紧密相关的，这些问题最终都会影响微博情感分析的准确性。

随着词向量技术的兴起，词嵌入模型能够学习基于单词的局部上下文的词向量，可以从大规模的语料中有效地学习词语细粒度的语义和句法规则，而主题情感模型利用跨文档的单词共现来识别与主题相关的单词。因此，考虑到两者的互补性，本文在目前较为先进的主题情感模型之一——TSMMF 模型中引入词向量技术，提出了基于词嵌入的改进 TSMMF 模型，并运用多元高斯分布从词向量空间中快速采样邻近词语，从而将共现频率低、信息量少的单词转变成突出主题、信息明确的单词。同时，使用最近邻搜索算法来提升词向量空间中邻近词的搜索速度，提高微博情感分析的准确性，同时也有助于促进相应研究成果的推广实施。

本文的主要贡献如下：

(1)针对 TSMMF 模型只考虑单词共现的情况，不能有效检测低频词语的语义信息，导致主题的语义一致性不足，进而影响情感分类的准确性问题。本文引入了词向量空间，将共现频率低、信息量少的单词转变成突出主题、信息明确的单词。

(2)为了进一步提升模型的运行速度，本文使用多元高斯分布来改进词嵌入主题情感模型的方法，并使用最近邻搜索算法来加快词向量空间中查找邻近单词的速度，进而提出了 GWE-TSMMF 模型。

(3)为验证 GWE-TSMMF 模型的可行性和有效性，本文在真实微博语料库上进行实验，实验结果表明，GWE-TSMMF 模型在微博情感极性分析效果上优于其他词嵌入主题情感模型。

2 相关工作

随着最近对深度神经网络兴趣的激增，许多研究工作集中于学习连续空间中的实值单词嵌入，这种技术被称为词嵌入，它可以捕获单词之间的分布相似性(如词性、语义)。当前最为流行的词向量技术是 Word2Vec，由 Mikolov 等[6]于 2013 年提出，Word2Vec 主要分为两种模型：Skip-gram 和 CBOW(Continuous Bag-Of-Words)[7]。Zhang 等[8]通过检测不同的词向量模型在不同类型语料文本中的表现能力，验证了 CBOW 模型比 Skip-gram 模型在百度数据集上训练出来的词嵌入效果更好。一般来说，获取词向量的方式有两种，一种是通过对全网超大规模文本库进行训练得到开源的全局词向量库；另一种是根据自己收集的文本库训练得到一个局部词向量库[9]。本文采用 CBOW 模型训练中文维基百科语料库，通过引入外部数据的方式来丰富单词的语义。

与主题模型(如 LDA 模型)相比，词嵌入模型可以有效地在更大的语料库上进行训练，从而允许在主题模型推理过程中，使用更多的语义信息来加强主题的语义一致性。主题模型与词嵌入模型相结合这一领域的研究主要有两个方向：一种是通过结合主题模型的思想来改进词嵌入模型[10-11]，另一种是通过将词向量技术与先验知识相结合来改进主题模型[12-14]。本文侧重于后者，下文将详细讨论这方面的相关工作。Nguyen 等[15]将 LDA 模型与词向量技术相结合，提出了 LFLDA(Latent Feature LDA)模型，通过在巨大的外部语料库上训练得到潜在的特征向量，以改进在较小的语料库上学习主题-单词映射，在很大程度上提升了主题一致性。但是 LFLDA 模型在较大的数据集上运行太慢，因此不适用于大型语料库上的主题分析。与此同时，Das 等[16]提出了高斯 LDA 模型，该模型将 LDA 模型中的词语生成部分换成了高斯分布。虽然高斯 LDA 模型在性能方面有了一定改善，但是其推理速度仍较慢，难以适用于高维度的词向量。随后，Yang 等[17]提出了高斯主题模型 GenVector，该模型充分利用了词向量中包含的语义关联，但是忽视了利用传统共现模式获取的语义关系，造成主题聚类的效果不佳。针对高斯 LDA 模型难以使用高维词向量的缺陷，Stefan 等[18]提出了 WELDA(Word Embedding LDA)模型，该模型通过估计主题在词向量空间中的分布，并使用吉布斯采样从这个空间交换选定的主题词来实现与词向量技术的结合，然后使用主成分分析技术对词向量进行降维，以提高模型的运行速度。实验结果表明，WELDA 模型在主题一致性和解决单词入侵任务方面表现优异，但是在处理大型数据集时，该模型的性能还有待提高。随后，Hua 等[19]在 WELDA 模型的基础上提出了 WLDA(Word embedding LDA)模型，其中词嵌入模型运行在吉布斯采样的外层，因此相比 WELDA 模型，WLDA 模型的运行效率更高。在医院评论数据上的实验结果证明，与同类的主题模型相比，该模型具有更好的主题一致性。

上述研究从不同角度将词嵌入模型与主题模型相结合，在一定程度上改善了传统主题模型仅通过单词共现推断主题分布的问题，并通过训练外部语料库将单词的先验语义信息合并到主题模型中，增强了主题的一致性。然而，这些模型仅

完成了挖掘主题的任务，并未检测情感和主题的关联。截至目前，将词嵌入与主题情感联合模型结合的相关研究还较少。Fu 等[20]提出将词嵌入与主题情感联合模型 JST 相结合，提出了 WS-TSWE 模型(Weakly Supervised Topic and Sentiment Model with Word Embeddings)。该模型可以显著增强情感-主题-单词间的映射关系，并扩展单词的语义和句法信息。实验结果表明，主题情感模型与单词嵌入的结合改善了文档-情感分布，能有效地捕获积极和消极的情感。Xu[21]提出了一个新的情感主题模型 HST-SCW(A Hybrid Sentiment and Topic Model with Auxiliary Word Embedding for Sentiment Analysis)，该模型可以将词向量中词语之间的语义关系结合到 JST 模型中，以更好地挖掘文本数据的情感和主题。然而，无论是 WS-TSWE 模型还是 HST-SCW 模型都只是基于 JST 模型进行的改进工作，因此可以将其拓展到更好的主题情感模型上，从而进一步提升情感分类的准确性。另外，本文的相关研究是基于中文微博情感分析展开的，这些模型在中文微博数据集上的表现还有待验证，尤其是在处理大型微博数据集时，对 WS-TSWE 模型和 HST-SCW 模型的运行性能都是一个较大的考验。

综上所述，较多的研究将词向量技术融合到主题模型 LDA 以及主题情感模型 JST 中，提升了主题的挖掘质量。然而，随着微博用户量的增长，微博平台产生的数据量几乎呈指数级增长，造成主题情感模型处理大型微博数据集时的运行时间大幅增长，而词向量的引入会进一步延长模型的运行时间，严重影响模型的运行性能，降低用户体验。目前，在中文微博情感分析领域，TSMMF 模型作为最近已公开发表的主题情感建模方法之一，其微博情感极性分类的准确率远高于 JST 模型。因此，本文在 TSMMF 模型中引入词向量技术，并运用多元高斯分布从词向量空间中快速采样邻近词语，同时使用最近邻搜索算法来提升词向量空间中邻近词的搜索速度，进而提出了 GWE-TSMMF 模型(Topic Sentiment Model based on Multi-feature Fusion with Gaussian Word Embedding)。

3 GWE-TSMMF 模型

3.1 设计思路

WS-TSWE 模型和 HST-SCW 模型基于主题情感模型 JST 引入了词向量技术，增强了主题的一致性，从而提升了文本情感分类的准确性，因此词嵌入对主题情感模型来说是一个非常好的改进思路。鉴于此，本文希望将词嵌入技术与主题情感模型 TSMMF 相结合，并将其运用于大型微博语料库。但是，主题情感模型在处理大型数据集时的运行时间会大幅增长，而词向量的引入会进一步延长模型的运行时间。因此，寻找一种可靠简便的方法从大型词向量空间中快速采样单词，从而提升模型的运行速度，将是本文要解决的重要问题之一。本文首先想到了多元高斯分布，多元高斯分布模型因其广泛的适应性以及分析复杂统计数据的优势，在多元分析中占有重要地位。假设向量 $\boldsymbol{X}=[\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n]^{\mathrm{T}}$ 服从均值为 $\mu\in R_n$、协方差矩阵为 $\boldsymbol{\Sigma}\in S_n$ 的多元高斯分布，则概率密度函数的形式如下。

$$P(x,\mu,\boldsymbol{\Sigma})=\frac{1}{\left[(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}\right]}\exp\left[\frac{1}{2}(x+\mu)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(x-\mu)\right] \tag{1}$$

多元高斯分布在分析复杂统计数据方面的优势能够为高效地引入词向量空间提供一定的参考。充分利用词向量空间的关键步骤是，根据学习的词嵌入概率分布中主题性更相似的单词来取代原模型中利用情感-主题-词语 Dirichlet 多项式分布产生的单词。具体操作遵循以下过程：通常，训练词嵌入模型，以预测准确位置上的单词，这意味着，想要获取给定单词的相似单词，词嵌入模型会预测与给定单词处于同一位置的单词。因此，本文可以利用相似单词在词向量空间中的编码相近这一规律，通过重新采样词向量空间中与热门情感主题单词邻近的单词来替换模型中共现频率低且信息量少的单词。为此，本文需要为每个情感-主题对学习词向量空间中的多元高斯分布 $\Omega_{s,t}$(也称为嵌入式情感-主题分布)，而该多元高斯分布 $\Omega_{s,t}$ 是通过情感-主题-词语分布 φ 下的多个热门单词的向量表示统计得到的：借助情感-主题-词语分布 φ 可以获取每个情感-主题对(s,t)中的前 N_{top} 个热门单词，且每个单词 w 的向量表示为$[\boldsymbol{d}_1,\boldsymbol{d}_2,\boldsymbol{d}_3,\cdots,\boldsymbol{d}_{\mathrm{dim}}]$(其中，dim 表示向量的维数)，然后计算出前 N_{top} 个热门单词向量的均值 $\mu_{s,t}$ 和协方差 $\Sigma_{s,t}$，进而可以表示出多元高斯分布 $\Omega_{s,t}$。这样，就可以在 GWE-TSMMF 模型的训练迭代期间，选择性地将微博中的单词替换为嵌入式情感-主题分布产生的单词。

由上文可知，在通过多元高斯分布 $\Omega_{s,t}$ 采样一个样本后，需要在词向量空间中寻找与该样本最邻近的词语。如果采用线性搜索，即对词向量空间内的所有单词进行求值计算，由于是在吉布斯采样内层进行搜索，时间复杂度为 $O(N_{iter}*M*V)$，在处理大型微博语料库时模型的运行速度会变得十分缓慢，因此，如何快速地从大型词向量空间中搜索出邻近单词也是本文需要解决的一个关键问题。

KD(K-Dimensional)树算法是一种基于树结构的近邻检索方法，该算法在低维空间中性能较高，但在高维空间中其性能会迅速下降[22]。近年来，哈希算法因其在存储空间和计算时间上的优势得到了广泛应用，如关键点检测[23]、图像和视频检索[24]等。基于哈希的检索方法最早起源于局部敏感哈希 LSH(Locality Sensitive Hashing)算法[25]，LSH 算法基于空间中的随机超平面，使用 Hash 函数对所有向量进行降维，将相近的对象散列到同一个桶中，能够对高维的文本数据进行分类或聚类，因此被广泛用于处理海量高维数据中的最近邻搜索问题。然而，对于维数特别低的空间，LSH 算法的执行效果不理想[26]。

KD 树算法在低维空间中能保证找到最近的邻居，且其平均检索时间复杂度为 $O(\log k)$。但是，对于超过 10 维以上的数据，KD 树算法的性能下降为线性搜索，此时可以使用局部敏感哈希算法 LSH 来搜索更高维度空间的近似最近邻居。本文的研究重点之一是在大型词向量空间中快速搜索邻近单词，因此 KD 树和 LSH 算法刚好适用。为了综合评估 KD 树和 LSH 算法的性能，本文将在不同的维数下比较两种算法的搜索时间，以找到适合两种算法的维数空间。

3.2 GWE-TSMMF 模型描述与推导

为了将词向量技术与主题情感模型 TSMMF 结合，并提升其在大型微博语料库上的运行性能，本文引入了高斯分布，在此基础上提出了基于高斯分布的改进词嵌入主题情感模型 GWE-TSMMF，其结构如图 1 所示。

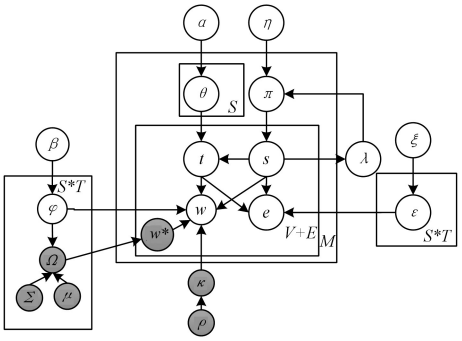

图 1 GWE-TSMMF 概率图模型

Fig. 1 GWE-TSMMF probabilistic graphical model

GWE-TSMMF 模型在原 TSMMF 模型的基础上引入多元高斯分布，其目的是从大型词向量空间中快速采样单词，从而提升模型的运行速度。除此之外，本文还引入了伯努利分布 $\kappa \sim Ber(\rho)$ 来决定该单词是由 Dirichlet 多项式分布产生还是由词向量空间生成，ρ 表示单词由词向量空间产生的概率，κ 的取值为 0 或 1。图 1 中，该模型与原 TSMMF 模型的不同之处以黑色阴影加粗标明。GWE-TSMMF 模型中涉及到的符号及含义如表 1 所列。

表 1 GWE-TSMMF 模型的符号及含义

Table 1 Meanings of symbols in GWE-TSMMF Model

符号	含义	符号	含义
M	微博个数	e	表情符号
V	语料库词汇数	w	词语
S	情感标签个数	s	情感标签
T	主题标签个数	t	主题标签
E	表情符号个数	λ	性格情绪参数
θ	主题的概率分布	α	分布 θ 的先验参数
φ	词语的概率分布	β	分布 φ 的先验参数
π	情感的概率分布	η	分布 π 的先验参数
ε	表情符号的概率分布	ξ	分布 ε 的先验参数
Ω	多元高斯分布	ρ	词语由词向量空间产生的概率
μ	词向量均值	κ	二元指示变量
Σ	词向量协方差	w^*	词向量空间采样词语

GWE-TSMMF 模型的生成过程如下：

(1)为每篇微博 m 生成情感分布 $\pi_m \sim Dir(\eta)$；接着为每种情感标签 s 生成情感-主题分布 $\theta_{m,s} \sim Dir(\alpha)$。

(2)为每个主题情感对 (t,s) 分别生成情感-主题-词语分布 $\varphi_{t,s} \sim Dir(\beta)$ 以及情感-主题-表情符号分布 $\varepsilon_{t,s} \sim Dir(\xi)$。

(3)为了生成微博中的每个词语 w：

1)根据 $s \sim Mul(\pi_m)$ 随机选择一个情感标签；

2)根据 $t \sim Mul(\theta_{m,s})$ 随机选择一个主题；

3)选择二元指示变量 $\kappa \sim Ber(\rho)$；

4)如果 $\kappa=1$，从词向量空间中重新采样单词 $w^* \sim N(\mu_{s,t}, \Sigma_{s,t})$ 来替换单词 w；如果 $\kappa=0$，则选择单词 $w \sim Mul(\varphi_{t,s})$。

(4)而对于微博中的每个情感符号 e，通过 $e_{t,s} \sim Mul(\varepsilon_{t,s})$ 获取。

本文采用吉布斯采样方法对 GWE-TSMMF 模型进行推导，需要先对条件概率分布进行估计：

$$P(t_i=t, s_i=s \mid t_{-i}, s_{-i}, w, e, \alpha, \beta, \eta, \xi, \lambda) \tag{2}$$

若想估计上述条件概率公式，则需先计算联合概率公式 $P(w,e,t,s)$，根据条件概率公式的基本定理可知：

$$P(w,e,t,s)=P(w \mid t,s)P(e \mid t,s)P(t \mid s)P(s) \tag{3}$$

若想求解 $P(w \mid t,s)$，$P(e \mid t,s)$，$P(t \mid s)$，$P(s)$ 等因子的值，则需对 φ 分布、ε 分布、θ 分布和 π 分布分别进行积分运算，具体如式(4)－式(7)所示。

$$\begin{aligned} P(w \mid t,s) &= \int P(w \mid t,s,\varphi)P(\varphi \mid \beta)\mathrm{d}\varphi \\ &= \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^{S*T} \prod_{s=1}^{S}\prod_{t=1}^{T}\frac{\prod_{w=1}^{V}\Gamma(N_{w,t,s}+\beta)}{\Gamma(N_{t,s}+V\beta)} \end{aligned} \tag{4}$$

其中，$N_{w,t,s}$ 表示词语 w 在情感标签 s 和主题标签 t 下出现的次数，$N_{t,s}$ 表示所有词语在情感标签 s 和主题标签 t 下出现的总数，$\Gamma()$ 表示伽马函数。

$$\begin{aligned} P(e \mid t,s) &= \int P(e \mid t,s,\varepsilon)P(\varepsilon \mid \xi)\mathrm{d}\varepsilon \\ &= \left(\frac{\Gamma(E\zeta)}{\Gamma(\zeta)^E}\right)^{S*T} \prod_{s=1}^{S}\prod_{t=1}^{T}\frac{\prod_{e=1}^{E}\Gamma(M_{e,t,s}+\xi)}{\Gamma(M_{t,s}+E\xi)} \end{aligned} \tag{5}$$

其中，$M_{e,t,s}$ 表示表情符号 e 在情感标签 s 和主题标签 t 下出现的次数，$M_{t,s}$ 表示所有表情符号在情感标签 s 和主题标签 t 下出现的总次数。

$$\begin{aligned} P(t \mid s) &= \int P(t \mid s,\theta)P(\theta \mid \alpha)\mathrm{d}\theta = \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^{S*M} \\ &\prod_{m=1}^{M}\prod_{s=1}^{S}\frac{\prod_{t=1}^{T}\Gamma(N_{m,t,s}+\alpha)}{\Gamma(N_{m,s}+T\alpha)} \end{aligned} \tag{6}$$

其中，$N_{m,t,s}$ 表示微博消息 m 中的当前元素同时出现在情感标签 s 和主题标签 t 下的次数，$N_{m,s}$ 表示微博消息 m 中的所有元素出现在情感标签 s 下的次数。

$$\begin{aligned} P(s) &= \int P(s \mid \pi)P(\pi \mid \eta)\mathrm{d}\pi \\ &= \left(\left(\frac{\Gamma(S(\eta+\lambda))}{[\Gamma(\eta+\lambda)]^S}\right)\right)^M \prod_{m=1}^{M}\frac{\prod_{s=1}^{S}\Gamma(N_{m,s}+\eta+\lambda)}{\Gamma(N_m+S(\eta+\lambda))} \end{aligned} \tag{7}$$

其中，N_m 表示微博消息 m 中元素(词语或者表情符号)的总数，λ 表示用户的性格情绪参数。

根据式(4)－式(7)可以计算联合概率 $P(w,e,t,s)$，从而进一步得到吉布斯采样所需的条件概率公式：

$$\begin{aligned} P(t_i=t, s_i=s \mid t_{-i}, s_{-i}, w, e, \alpha, \beta, \eta, \xi, \lambda) \propto & \frac{(N_{m,t,s})_{-i}+\alpha}{(N_{m,s})_{-i}+T\alpha} \times \\ & \frac{(N_{w,t,s})_{-i}+\beta}{(N_{t,s})_{-i}+V\beta} \times \frac{(M_{e,t,s})_{-i}+\xi}{(M_{t,s})_{-i}+E\xi} \times \frac{(N_{m,s})_{-i}+\eta_s+\lambda_s}{(N_m)_{-i}+\sum_{s=1}^{S}(\eta_s+\lambda_s)} \end{aligned} \tag{8}$$

获得上述后验分布后，可以计算出各项 Dirichlet 先验分布如式(9)－式(12)所示。

$$\theta=\frac{N_{m,t,s}+\alpha}{N_{m,s}+T\alpha} \tag{9}$$

$$\varphi=\frac{(N_{w,t,s})_{-i}+\beta}{(N_{t,s})_{-i}+V\beta} \tag{10}$$

$$\varepsilon=\frac{M_{e,t,s}+\xi}{M_{t,s}+E\xi} \tag{11}$$

$$\pi=\frac{N_{m,s}+\eta_s+\lambda_s}{N_{m+s}(\eta+\lambda_s)} \tag{12}$$

此时，通过微博的情感分布 π 可以得到微博在情感标签 s 下的概率，从而获得微博的情感极性。

3.3 GWE-TSMMF 模型的描述

根据 3.2 节中 GWE-TSMMF 模型的描述和推导过程，本文给出 GWE-TSMMF 模型进行中文微博情感极性分析的流程，采样过程使用的是吉布斯采样方法，具体步骤如算法 1 所示。

算法 1 基于 GWE-TSMMF 模型的微博情感极性分析

输入：中文微博数据集 Corpus=$\{m_1, m_2, \cdots, m_M\}$，词向量矩阵 $\boldsymbol{\omega}$

输出：每条微博 m 的情感极性

1. 对微博进行预处理，然后提取所有的单词和表情符号
2. 初始化先验参数 α, β, η, ξ 和分布 $\theta, \pi, \varphi, \varepsilon$，二项分布概率 ρ 值，初始迭代次数 $N_{initIter}$，情感标签数目 S，主题标签数目 T，热门词汇数目 N_{top} 以及词向量的降维数 N_{pca}
3. 初始化迭代操作
4. FOR n=1 to $N_{initIter}$
5. FOR m=1 to M(微博个数)
6. FOR v=1 toV(语料库词汇数)
7. **根据二元指示变量 $\kappa \sim Ber(\rho)$，如果 $\kappa=1$，则从词向量空间中重新采样单词 $w * \sim N(\mu_{s,t}, \sum_{s,t})$ 来替换当前词语 w**
8. 排除具有情感 s 和主题 t 的当前词语
9. 更新计数变量
10. 根据式(8)为单词计算新的情感标签和主题标签
11. END FOR
12. END FOR
13. END FOR
14. 运用采样结果更新分布 $\theta, \pi, \varphi, \varepsilon$
15. 根据微博情感分布 π 计算微博情感极性，if ($P_{pos} > P_{neg}$)，则该篇微博的情感极性为积极；否则其情感极性为消极

算法 1 中，字体加粗部分为该算法与原 TSMMF 模型的不同之处。步骤 1、步骤 2 主要是对微博文本进行预处理和参数的初始化，为了加快模型的学习速率，算法 1 使用了主成分分析技术(PCA)，因此该算法还需要初始化词向量的降维数 N_{pca}。步骤 3 为初始化迭代操作，此处参照的是原 TSMMF 模型的迭代过程。通过前面的步骤可以得到每个情感-主题对 (s,t) 下的热门词汇，然后根据热门词汇的向量表示统计得到多元高斯分布 $\Omega_{s,t}$。步骤 4—步骤 13 通过二项分布来决定是否使用多元高斯分布 $\Omega_{s,t}$，从词向量空间中采样邻近词语替换当前词语，然后对单词重新采样新的情感标签和主题标签并进行进一步的推断。最后两步是计算微博的情感极性。

3.4 GWE-TSMMF 模型的复杂度分析

从算法 1 中可以看出，GWE-TSMMF 模型的时间复杂度与迭代次数 $N_{initIter}$、情感标签数目 S、主题标签数目 T、热门词汇数目 N_{top}、微博总数 M 以及语料库词汇数目 V 相关，其时间复杂度为 $O(N_{initIter} * M * V * (S * T + N_{top}))$。其空间复杂度与微博总数 M、情感标签数目 S、主题标签数目 T、热门词汇数目 N_{top}、语料库词汇数目 V、词向量维度 E 以及文档中所有单词的平均长度 C 相关，其空间复杂度为 $O(M * S * T + M * S + S * T * (V + N_{top}) + E * N_{top} + M * C)$。

4 实验与分析

4.1 实验数据集

(1)数据集

本文实验所用数据集融合了 NLP&CC 2013 和 NLP&CC 2014 提供的中文微博情感标注语料集，该数据集将微博情感分为 7 类，包括愤怒、厌恶、恐惧、悲伤、高兴、喜好、惊讶。由于本文研究的是微博情感极性分析，因此将前 4 种情感归为负向情感类，将后 3 种情感归为正向情感类。本文收集到了 19664 条标注微博(10054 条为正向微博，9610 条为负向微博)，然后从标注微博语料集中选取 15000 条微博作为训练集，选择 4000 条微博作为测试集，最后利用上述数据集在 GWE-TSMMF 模型上进行相应的实验，具体实验结果如表 2 所列。

表 2 实验数据集

Table 2 Dataset of experiment

(单位：条)

类别	训练集	测试集
积极微博数目	7489	1974
消极微博数目	7511	2026

(2)表情符号库

GWE-TSMMF 模型使用了表情符号作为情感先验知识，在一定程度上增强了微博的情感极性。本文根据表情符号使用频率构建了表情符号库，其中积极和消极情感极性的表情符号各 20 个，由于篇幅有限，这里仅给出部分实例，如表 3 所列。

表 3 微博表情符号库

Table 3 Emoji library of Weibo

符号	文字表示	符号	文字表示
	[微笑]		[怒骂]
	[嘻嘻]		[委屈]
	[笑哭]		[开心]
	[费解]		[可爱]
	[黑线]		[舔屏]
	[打脸]		[泪]

4.2 评价指标

本文选取 F1 值来评价主题情感模型进行微博情感分类的效果，通过测量运行时间(单位为 s)来比较不同的 PCA 降维数下引入 KD 树和 LSH 算法的搜索性能，以及 GWE-TSMMF 模型与对比模型的运行性能，采用 Mimno 等[27]提出的主题一致性得分来评估主题质量。

4.3 实验设计

本文在中文维基百科语料库上通过 CBOW 模型来训练得到词向量空间，且词向量初始维度为 200 维。GWE-TSMMF 模型中的先验参数 α, β, η 和 ξ 参考 TSMMF 模型来设置，设对称超参数 $\alpha=0.1, \beta=0.01, \xi=0.01$；对于不对称超参数，若是积极情感，则 $\eta=0.01$，若是消极情感，则 $\eta=5$。为了加快 GWE-TSMMF 模型的运行速度，本文首先划分适合两种搜索算法的维数空间，以供模型根据 PCA 降维数来选择相应的算法；然后探寻 PCA 降维数对情感分类的影响，因此

需要调整最优 N_{pca} 值。由文献[28]可知，主题数在不同数据集上对模型的情感分类能力的影响不同，因此本文需要进行对比实验来获取最优主题数目。由于本文引入了伯努利分布 $\kappa \sim Ber(\rho)$ 来控制单词的生成方式，因此二项分布概率 ρ 的值会对最终的分类结果产生一定的影响。为了评估这种影响，随后本文将 ρ 作为重点调整的参数。最后设置对比实验来考察引入基于高斯分布的词向量空间对模型运行性能及情感分类准确性的影响。实验中，每组实验进行 10 轮，取 10 次实验结果的平均值，以增加实验结果的可靠性。

4.4 实验结果与分析

在实验部分，本文首先验证了 GWE-TSMMF 模型引入词向量技术的优势；然后通过实验确定了对情感分类结果有影响的 PCA 降维数、主题数目、二项分布概率值等因素的最优值；最后设置对比实验考察引入高斯分布词向量技术对模型运行性能及情感分类准确性的影响。

(1)GWE-TSMMF 模型主题质量评估

本文引入词向量技术，以充分利用上下文语义信息，改善主题一致性，进而调整微博-情感-主题分布，最终达到提升情感分类准确性的目的。因此，对主题质量的评估是非常有必要的，它能更直观地展现词向量技术的引进对主题情感模型的影响。为了量化主题质量的评估，本文使用主题一致性得分这一评价指标，然后使用 GWE-TSMMF 模型、TSMMF 模型以及 JST 模型在数据集上提取情感相关的主题并计算主题一致性得分。表 4 列出了各模型在最优主题数目下前 N (N 分别为 5,10,15,20)个单词的平均一致性得分。

表 4 模型在最优主题数目下前 N 个单词的平均一致性得分

Table 4 Average topic coherence score of the first N words of the model under the optimal number of topics

N	GWE-TSMMF	TSMMF	JST
5	−36.11	−39.82	−41.42
10	−238.45	−246.14	−250.27
15	−532.34	−545.98	−553.31
20	−1013.71	−1125.23	−1128.16

从表 4 中主题一致性得分的结果来看，随着 N 的增大，各模型的主题一致性得分减小。通过进一步观察可知，GWE-TSMMF 模型的一致性得分始终高于 TSMMF 模型和 JST 模型，说明 GWE-TSMMF 模型相比原模型，能够改善主题一致性，加强主题质量，这一结论符合引入词嵌入技术的预期。

(2)不同 PCA 降维数下 KD 树和 LSH 算法的搜索性能比较

为了综合评估 KD 树和 LSH 算法的性能，需要在不同的维数 N_{pca} 下比较两种算法的搜索时间。因此，本文在 2～10 之间选取 9 个维度(抛开维度为 1 的情况)，设定主题数目为 20，ρ 值为一个中间值为 0.5，然后在不同的维数下模型分别使用 KD 树和 LSH 算法迭代至收敛状态，取模型的运行时间为最终结果，实验结果如图 2 所示。由图 2 可知，KD 树算法的运行时间随着维度的增加逐渐延长，验证了 KD 树算法面对高维数据时性能会下降；而 LSH 算法的运行时间随着维度的增加逐渐缩短，说明数据维数越低，LSH 算法的耗时就越长。通过观察发现，两种算法的交叉点在维度为 5 和 6 之间的位置，因此以 $N_{pca}=5$ 为两种算法的划分区间，当 $N_{pca}<=5$ 时，GWE-TSMMF 模型采用 KD 树算法；当 $N_{pca}>5$ 时，GWE-TSMMF 模型采用 LSH 算法。

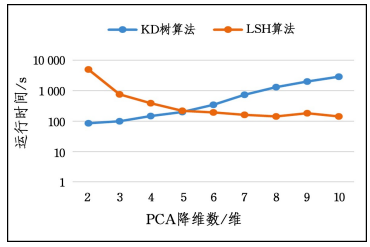

图 2 不同 PCA 降维数下 KD 树和 LSH 算法的搜索性能比较
(电子版为彩色)

Fig. 2 Performance comparison of KD tree and LSH for the nearest neighbour word search based on the number of PCA-dimension

(3)PCA 降维数及热门词汇数目对情感分类的影响

在计算多元高斯分布之前，本文使用 PCA 技术将高维的词向量空间降维到较低的维数。而参数 N_{top} 和 N_{pca} 之间存在着一定的折衷：一方面，词向量维度越低，用于计算多元高斯分布的词向量越少，意味着可以充分利用一个主题下最具有代表性的前 N_{top} 个热门单词，另一方面，当词向量被投影到较低的维度时，意味着丢失了词向量空间中的部分语义信息，使得词向量之间的相似距离变得不太可靠。为了探究 PCA 降维数及热门词汇数目对情感分类的影响，本文选取 2,5,10,30,50 这 5 个维度，其中维度为 2 和 5 时，选用 KD 树算法，剩下的选用 LSH 算法。此处暂时设定主题数目为 20，ρ 值为一个中间值 0.5，然后在 0～400 之间以 50 为间隔选择热门词汇数目 N_{top}，每组实验进行 10 轮，取 10 次实验结果的平均 F1 值作为最终的结果，实验结果如图 3 所示。

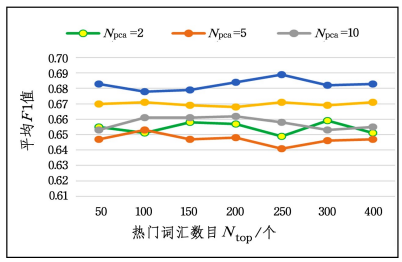

图 3 PCA 降维数和热门词汇数目对情感分类的影响
(电子版为彩色)

Fig. 3 Influence of the number of top words and the number of PCA-dimension on sentiment analysis

由图 3 可知，随着热门词汇数目 N_{top} 的增加，情感分类效果几乎不受影响，说明实验的重点在于调整参数而不在于热门词汇数目。通过进一步观察可知，当 PCA 降维数 $N_{pca}=50$ 时，相比其他几个低维度值取得了最好的情感分类效果，说明只有选择适当的词向量维数才能充分利用词向量空间中的语义信息。因此，对于设定的这几个维度值来说，本文选择 $N_{pca}=50$ 为最优 PCA 降维数。结合上述可知，当 $N_{pca}=50$ 时，GWE-TSMMF 模型采用 LSH 算法。

(4)主题数目对情感分类的影响

为了获取最优主题数目，本组实验选取 10,20,30,40,50,60,70,80,90,100 这 10 个主题数目，设定 PCA 降维数值为 50，ρ 值为一个中间值 0.5，然后在不同的主题数目下模型迭代至收敛状态，每组实验进行 10 轮，取 10 次实验结果的平均 F1 值作为最终的结果，实验结果如图 4 所示。

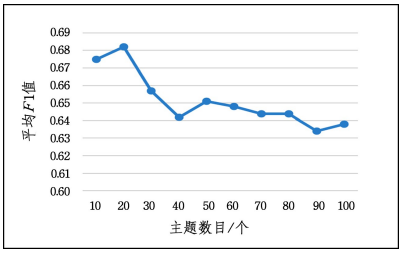

图 4 主题数目对情感分类的影响

Fig. 4 Influence of topic number on sentiment analysis

由图 4 可知，当主题数目为 20 时，情感分类的效果最佳。而当主题数目大于 20 时，随着主题数目的增加，情感分类的效果反而逐渐变差。这是因为过多的主题数目容易引入噪声主题，从而影响情感分类的效果。

(5)二项分布的概率 ρ 值对情感分类的影响

为了考察二项分布的概率 ρ 值对情感分类性能的影响，本文初步在 0～1 之间以 0.1 为间隔选取 10 个点，然后在最优主题数目 20 以及 PCA 降维数为 50 的条件下，选取不同的 ρ 值运行迭代至收敛状态，每组实验进行 10 轮，取 10 实验结果的平均 F1 值作为最终的结果，实验结果如图 5 所示。

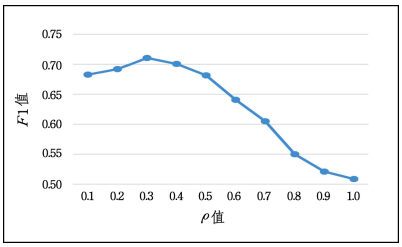

图 5 二项分布概率 ρ 值对情感分类的影响

Fig. 5 Influence of hyperparameter ρ on sentiment analysis

由图 5 可知，随着 ρ 值从 0.1 到 0.3 逐渐增大，F1 值也逐渐变大，当 ρ 值大于 0.3 时，F1 值呈现下降趋势。通过分析可知，当 ρ 值过小时，GWE-TSMMF 模型退化为 TSMMF 模型；而当 ρ 值过大时，GWE-TSMMF 模型过度依赖词向量空间来采样单词，而忽视了原有的情感-主题-词语 Dirichlet 多项式分布，这在一定程度上拉低了情感分类的准确性。为了进一步优化 ρ 值，需要继续在 0.3～0.4 之间以 0.01 为间隔选取 10 个点进行对比，实验结果如图 6 所示。

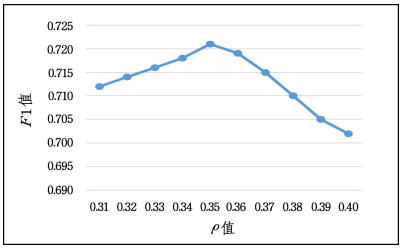

图 6 优化二项分布概率 ρ 值

Fig. 6 Optimize this hyperparameter ρ

由图 6 可知，当 ρ 值为 0.35 时，情感分类的效果达到最佳，因此本文选择 0.35 为 GWE-TSMMF 模型的最优 ρ 值。

(6)引入基于高斯分布的词向量空间对模型运行性能的影响

为了考察基于高斯分布的词向量空间对模型运行性能的影响，将对比 TSMMF 模型、JST 模型、不使用 LSH 算法的 GWE-TSMMF 模型以及使用 LSH 算法的 GWE-TSMMF 模型的运行时间(单位为 s)。结合前面的调优结果，各个模型都在最优主题数目为 20 的条件下进行，其中 GWE-TSMMF 模型中设置 ρ 最优值为 0.35，PCA 降维数为 50，并选用 LSH 算法。所有模型各进行 10 轮，10 次实验的结果如图 7 所示。

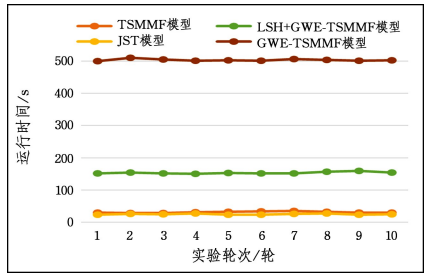

图 7 不同模型的运行时间比较(电子版为彩色)

Fig. 7 Comparison of runtime among different models

由图 7 可知，TSMMF 模型和 JST 模型的平均运行时间相近且用时最短，约为 30 s，未使用 LSH 算法的 GWE-TSMMF 模型的平均运行时间耗时最长，约为 500 s，而使用了 LSH 算法的 GWE-TSMMF 模型的运行时间有了明显的缩短，约为 150 s。综上所述，本文使用 LSH 算法来加速搜索词向量空间中邻近单词的思路是合理有效的。纵观使用 LSH 算法的 GWE-TSMMF 模型与 TSMMF 模型、JST 模型可以发现，前者与后两者运行时间的差距在 150 s 以内，在一个可以接受的范围内，这是由于词向量技术的引入势必会增长模型的运行时间，但是基于高斯分布的词向量空间使得模型前后的运行时间的差距在一个数量级内，说明本文提出的模型运行于大型微博数据集上是可行的。本文引入基于高斯分布的词向量空间的目的是尽可能地提升模型的运行效率，虽然整体的运行时间仍高于原模型，但是应该考虑其对模型情感分类准确性的影响。

(7)引入基于高斯分布的词向量空间对情感分类准确性的影响

为了考察基于高斯分布的词向量空间对情感分类准确性的影响，对 TSMMF 模型、JST 模型、不使用 LSH 算法的 GWE-TSMMF 模型以及使用 LSH 算法的 GWE-TSMMF 模型进行对比。结合上文的调优结果，各个模型都在最优主题数目为 20 的条件下进行，其中 GWE-TSMMF 模型中设置 ρ 最优值为 0.35，PCA 降维数为 50。两组实验各进行 10 轮，10 次实验的结果如图 8 所示。

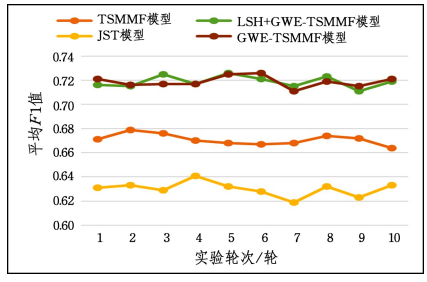

图 8 不同模型的情感分类效果比较(电子版为彩色)

Fig. 8 Comparison of sentiment classification effect among different models

由图 8 可知，JST 模型的平均 $F1$ 值最低，约为 0.63，基于 JST 模型改进的 TSMMF 模型的平均 $F1$ 值显著提升，达到了 0.67，而 GWE-TSMMF 模型和使用了 LSH 算法的 GWE-TSMMF 模型的平均 $F1$ 值整体一致，达到了 0.72，明显高于 TSMMF 模型的平均 $F1$ 值，为 0.67，两者的差距接近 10%。上述结果足以说明，引入基于高斯分布的词向量空间提升了微博情感极性分析的性能，并进一步验证了 GWE-TSMMF 模型在大型微博数据集上的分类效果。

结合实验(6)和实验(7)可知，GWE-TSMMF 模型引入了基于高斯分布的词向量空间，虽然使模型的运行时间延长，但在大型微博数据集上该差距在可承受的范围内。更重要的是，GWE-TSMMF 模型的情感分类准确率相比原 TSMMF 模型显著提升，进一步说明了本文的改进工作是可行且有效的。

(8)GWE-TSMMF 模型与主流词嵌入主题情感模型的对比实验

为了更好地比较 GWE-TSMMF 模型、WS-TSWE 模型和 HST-SCW 模型的情感分类效果，同样需要预先进行对比调优实验，以获取 WS-TSWE 和 HST-SCW 模型在所用微博数据集上的最优主题数目以及相关的实验参数，而 GWE-TSMMF 模型的最优主题数目以及其他参数设置可以参照前文。然后本文分别进行 10 组实验并统计 10 次实验结果的平均 F1 值，实验结果如图 9 所示。

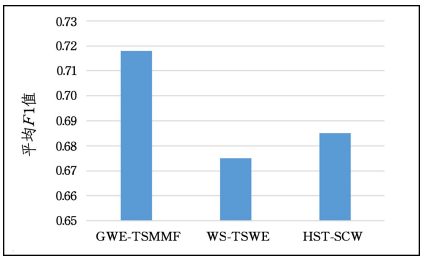

图 9 GWE-TSMMF 模型、WS-TSWE 模型和 HST-SCW 模型的情感分类效果对比

Fig. 9 Comparison of sentiment classification effect GWE-TSMMF model, WS-TSWE model and HST-SCW model

由图 9 可知，GWE-TSMMF 模型的平均 F1 值高于 WS-TSWE 模型和 HST-SCW 模型。由于 GWE-TSMMF 模型是在 TSMMF 模型上进行的词嵌入改造，而 TSMMF 模型在大型微博语料库上的情感分类效果优于 JST 模型，因此本文提出的 GWE-TSMMF 模型在微博语料库上的情感分类效果优于目前主流的词嵌入主题情感模型。

结束语 本文在 TSMMF 模型中引入词向量技术，以丰富单词的语义信息，考虑到面临大型微博语料库时模型会出现运行速度缓慢的问题，提出了基于高斯分布的改进词嵌入主题情感模型——GWE-TSMMF 模型。然后通过对比实验调整了 GWE-TSMMF 模型的各项最优参数，并且验证了该模型相比原模型和主流词嵌入主题情感模型，具有更优的微博情感极性分析性能。

虽然 GWE-TSMMF 模型表现出了一定的性能提升效果，但是仍存在一些需要在未来继续深入研究的工作：

(1)本文结合基于高斯分布的词向量空间虽然能够提升主题情感模型的情感分类准确性，但是运行时间明显高于原模型，而多余的运行时间开销主要来源于从词向量空间搜索单词并替换掉原来的单词的过程，因此如何寻找更快速的搜索算法，尽可能地加快模型的运行速度将是未来的一个研究重点。

(2)本文模型中结合的词向量技术只针对微博的词语，而作为微博中的另一种元素——表情符号，由于只利用其情感先验值来捕获微博的隐含情感，并没有考察其上下文语义关系。He 等[29]借助词向量表示技术为常用表情符号构建情感空间的特殊表示矩阵，有效地提升了微博情感分类效果。因此，将本文模型中出现的表情符号元素进行词向量化也可以作为未来的一个研究方向。

参 考 文 献

[1] ZHANG S, WEI Z, WANG Y, et al. Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary [J]. Future Generation Computer Systems, 2018, 81: 395-403.

[2] WANG Y. Iteration-based naive Bayes sentiment classification of microblog multimedia posts considering emoticon attributes [J]. Multimedia Tools and Applications, 2020, 79: 19151-19166.

[3] PANG B, LEE L. Opinion mining and sentiment analysis[J]. Foundations and Trends in Information Retrieval, 2008, 2(1/2): 1-135.

[4] DERMOUCHE M, KOUAS L, VELCIN J, et al. A joint model for topic-sentiment modeling from text[C]// Proceedings of the 30th Annual ACM Symposium on Applied Computing. Salamanca: ACM, 2015: 819-824.

[5] HUANG F L, YU G, ZHANG J L, et al. Weibo Topic Sentiment Mining Based on Social Relationship[J]. Journal of Software, 2017, 28(3): 694-707.

[6] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

[7] YUAN T T, YANG W Z, ZHONG L J, et al. PLSTM, a personality-based sentiment analysis model for microblogs[J]. Computer Application Research, 2019, 37(2): 1-6.

[8] ZHANG X J, LU X Q, ZHOU Q. Research on multi-level differences in written texts based on word embedding [J]. Computer Engineering and Applications, 2019, 23(55): 142-149.

[9] GAO M X, JING W. Chinese short text classification method based on Word2Vec word model[J]. Journal of Shandong University (Engineering Science Edition), 2019, 49(2): 34-41.

[10] CHENG J P, WANG Z Y, WEN J R, et al. Contextual Text Understanding in Distributional Semantic Space[C]// Proceedings of the Conference on Information and Knowledge Management. New York: ACM, 2015: 133-142.

[11] SUN F, GUO J F, LAN Y Y, et al. Learning Word Representations by Jointly Modeling Syntagmatic and Paradigmatic Relations[C]// Proceedings of the Meeting of the Association for Computational Linguistics. Beijing: ACL, 2015: 136-145.

[12] LIU Y, LIU Z, CHUA T S, et al. Topical word embeddings [C]// Proceedings of the Twenty-ninth AAAI Conference on Artificial Intelligence. San Francisco: AAAI Press, 2015: 2418-2424.

[13] LI S H, CHUA T S, ZHU J, et al. Generative Topic Embedding: a Continuous Representation of Documents[C]// Proceedings of the Meeting of the Association for Computational Linguistics. Berlin: ACL, 2016: 666-675.

[14] QIANG J, CHEN P, WANG T, et al. Topic Modeling over Short Texts by Incorporating Word Embeddings[J]. PAKDD, 2017, 10235: 363-374.

[15] NGUYEN D Q, BILLINGSLEY R, DU L, et al. Improving topic models with latent feature word representations[J]. Transactions of the Association for Computational Linguistics, 2015, 3: 299-313.

[16] DAS R, ZAHEER M, DYER C. Gaussian LDA for Topic Models with Word Embeddings[C]// Proceedings of the Meeting of the Association for Computational Linguistics. Beijing: ACL, 2015: 795-804.

[17] YANG Z, TANG J, COHEN W. Multi-Modal Bayesian Embeddings for Learning Social Knowledge Graphs[C]// Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. New York: IJCAI, 2016: 2287-2293.

[18] STEFAN B, KRESTEL R. WELDA: Enhancing topic models by incorporating local word context[C]// Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries. New York: JCDL, 2018: 293-302.

[19] HUA S W, ZHANG Y H. Short Text Comment Sentiment Analysis of Improved Topic Models[J]. Computer Systems & Applications, 2019, 28(3): 255-259.

[20] FU X, SUN X, WU H, et al. Weakly supervised topic sentiment joint model with word embeddings[J]. Knowledge-Based Systems, 2018, 147: 43-54.

[21] XU K. Research of topic model-based approaches for sentiment and topic modeling on texts[D]. Nanjing: Southeast University, 2017.

[22] SILPA-ANAN C, HARTLEY R. Optimised KD-trees for fast image descriptor matching[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Anchorage: IEEE, 2008: 1-8.

[23] WU C, ZHU J, ZHANG J, et al. A Convolutional Treelets Binary Feature Approach to Fast Keypoint Recognition[C]// Proceedings of European Conference on Computer Vision. Berlin: Springer, 2012: 368-382.

[24] HU L J, NOOSHABADI S. High-dimensional image descriptor matching using highly parallel KD-tree construction and approximate nearest neighbor search[J]. Journal of Parallel Distributed Computing, 2019, 132: 127-140.

[25] ADITYA B, MAHESHAKYA W. Distributed Clustering via LSH Based Data Partitioning[C]// Proceedings of the 35th International Conference on Machine Learing. Stockholm: PMLR, 2018: 569-578.

[26] FENG X K, CUI J T, LI H, et al. An efficient LSH indexing on discriminative short codes for high-dimensional nearest neighbors[J]. Multimedia Tools and Applications, 2019, 78(17): 24407-24429.

[27] MIMNO D, WALLACH H M, TALLEY E, et al. Optimizing semantic coherence in topic models[C]// Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh: EMNLP, 2011: 262-272.

[28] HUANG F L, FENG S, WANG D L, et al. Mining Topic Sentiment in Microblogging Based on Multi-feature Fusion[J]. Chinese Journal of Computers, 2017, 40(4): 872-888.

[29] HE Y X, SUN S T, NIU F F, et al. A deep learning model enhanced with emotion semantics for microblog sentiment analysis[J]. Chinese Journal of computers, 2017, 40(4): 773-790.

LI Yu-qiang, born in 1977, Ph.D, associate professor, master tutor. His main research interests include machine learning and big data analysis.

ZHANG Wei-jiang, born in 1994, postgraduate. His main research interests include machine learning and big data analysis.

(责任编辑:喻藜)