doi :10.3969/j.issn.1671-1122.2017.01.009

基于 Word2vec 的微博短文本分类研究

张谦，高章敏，刘嘉勇

（ 四川大学电子信息学院，四川成都 610065 ）

摘　要：随着微博等社会化媒体的信息量急剧膨胀，人们迫切需要实现这些信息的自动分类处理，以帮助用户快速查找所需信息和过滤垃圾信息。针对传统文本分类模型存在的特征维数灾难、无语义特征等问题，文章基于 Word2vec 模型对微博短文本进行了分类研究。鉴于 Word2vec 模型无法区分文本中词汇的重要程度，进一步引入 TFIDF 对 Word2vec 词向量进行加权，实现加权的 Word2vec 分类模型。最后合并加权 Word2vec 和 TFIDF 两种模型，实验结果表明合并后模型分类准确率高于加权 Word2vec 模型和使用 TFIDF 的传统文本分类模型。

关键词：短文本分类；Word2vec；TFIDF；支持向量机

中图分类号：TP309.2　　**文献标识码：**A　　**文章编号：**1671-1122（2017）1-0057-06

中文引用格式：张谦，高章敏，刘嘉勇 . 基于 Word2vec 的微博短文本分类研究 [J]. 信息网络安全，2017（1）：57-62.

英文引用格式：ZHANG Qian，GAO Zhangmin，LIU Jiayong. Research of Weibo Short Text Classification Based on Word2vec[J]. Netinfo Security, 2017(1):57-62.

Research of Weibo Short Text Classification Based on Word2vec

ZHANG Qian, GAO Zhangmin, LIU Jiayong

(*College of Electronics and Information Engineering of Sichuan University*, *Chengdu Sichuan* 610065, *China*)

Abstract: With the rapid expansion of new available information on Microblogging and other social media. Text automatic classification becomes imperative in order to help people locate the information he inquires and filter spam. Based on the characteristics of curse of dimensionality and lack of semantic features in Traditional text classification model, put forward a short text classify based on Word2vec model.Since Word2vec can not distinguish the weight of words, we applied weights using tf-idf weighting with Word2vec, implemented weighted Word2vec. Then we concatenated tf-idf with our word2vec weighted by tf-idf. Our results show that the combination of Word2vec weighted by tf-idf without stop words and tf-idf without stop words can outperform either Word2vec weighted by tf-idf without stop words and tf-idf with or without stop word.

Key words: short text classification; Word2vec; TFIDF; SVM

收稿日期：2016-10-1

基金项目：国防保密通信重点实验室基金 [9140C110401140C11053]

作者简介：张谦（1987—），男，贵州，博士研究生，主要研究方向为网络信息安全、数据挖掘；高章敏（1991—），男，湖北，硕士研究生，主要研究方向为数据挖掘与机器学习；刘嘉勇（1962—），男，四川，教授，博士，主要研究方向为网络数据分析与信息安全。

通信作者：张谦 42297119@qq.com

0 引言

进入 Web 2.0 时代，微博等社交媒体开始兴起，微博是一个基于用户关系的信息分享、传播以及获的平台，人们可以通过微博发布、分享自己的看法。微博文本内容通常限制在 140 个字符之内，属于即时短消息，其特点有：稀疏性、用词不规范、网络用语较多等 [1]。将微博短文本进行分类研究，对挖掘用户兴趣、热点话题发现、个性化推荐系统都有较大研究价值 [2]。

文本分类是指按照预先定义的主题类别，为文档集合中的每个文档确定一个类别，是一个有监督的学习过程。随着数据时代的到来，互联网上电子文档的数量大幅增长，文本分类已经成为信息检索和管理的关键技术 [3]。文本分类首先需要将文本转换为计算机能处理的格式，传统的文档表示方法使用了信息检索领域技术，如连续词袋模型（Continuous bag-of -words, CBOW）、词频与逆文档频率（Term Frequency-inverse Document Frequency, TFIDF）等 [4]。在词袋模型中文档被看作是无序的词汇集合，忽略语法以及单词的顺序。TFIDF 是信息检索领域常用的加权技术，用以评估字词对于一个文件集或语料库中一份文档的重要程度。

目前，针对文本分类的研究大多基于传统的向量空间模型（Vector Space Model，VSM）[5]，向量空间模型是文档表示最常用的模型，在处理传统长文本分类问题时获得了较好的效果 [6]。然而，与传统长文本相比，短文本具有稀疏性、实时性、不规范等特点，向量空间模型的稀疏性会降低短文本分类的准确率。针对向量空间模型的稀疏问题，一种办法是借助外部知识库（如维基百科、WordNet、HowNet[7-9] 等）对短文本进行扩展。王盛等人利用知网词语对的上下位关系对短文本进行扩展 [10]；范云杰、赵辉等人借助维基百科知识库对短文本特征进行扩展，以辅助短文本分类 [11,12]；翟延冬等人利用 WordNet 查询计算短文本相似度 [13]；宁亚辉等人从知网知识库中抽取领域高频词扩展短文本 [14]。由于外部知识库包含的领域和主题比较少，词汇更新速度慢，很难应用到互联网短文本分类。另一种方法是借助外部文本（如搜索引擎的结果）扩展短文本特征。BOLLEGALA[15] 等人、SAHAMI[16] 等人把短文本作为查询词输入搜索引擎，并根据返回的结果计算短文本相似度；王鹏等人利用依存关系抽取具有依存关系的词对，扩展短文本特征 [17]；ZELIKOVIT 借助未标记的外部背景语料度量文本相似度 [18]。

近年来，基于语义信息的短文本分类的研究也取得了一定的进展，BLEI 等人、PHAN 等人、CHEN 等人使用主题模型隐含狄利克雷分布（Latent Dirichlet Allocation, LDA）提取语料主题，扩展短文本特征，对 Web 短文本进行分类 [19-21]；方东使用 LDA 主题模型对微博进行分类 [22]；吕超镇等人在短文本原始特征的基础上，利用 LDA 主题模型提取文本主题，把主题中的词作为短文本的部分特征，扩展短文本特征 [23]。上述这些方法需要大量的外部语料，大大增加了计算的开销。

Word2vec 是 Google 在 2013 年推出的一款用于训练词向量的工具，Word2vec 提供了一种使用分布式向量对文本进行表示的方法 [24]。与传统文本向量空间模型相比，使用 Word2vec 模型表示文本，既能解决传统向量空间模型的高维稀疏特征问题，还能引入传统模型不具有的语义特征，有助于短文本分类 [25]。因此，本文提出了一种基于 Word2vec 模型的短文本分类方法，解决了使用传统向量空间模型处理短文本时的特征稀疏问题。针对 Word2vec 模型无法区分文本中词汇的重要程度，本文进一步引入 TFIDF 模型计算 Word2vec 词向量的权重，提出加权 Word2vec 模型。最后本文合并加权 Word2vec 和 TFIDF 两种模型，使用合并后的特征对文本进行表示。实验结果表明，合并后的模型分类准确率高于传统文本分类模型。

1 方法

文本分类领域常用的技术有朴素贝叶斯分类器（Navie Bayes Classifier）、支持向量机（Support Vector Machine, SVM）、TFIDF 等 [26-28]。本文提出的短文本分类算法结合 Word2vec 和 TFIDF 两种模型。

TFIDF（Term Frequency-inverse Document Frequency, 词频与逆文档频率）是一种统计方法，用以评估一个字词对于文件集或语料库中的一份文档或一个类别的重要程度。其主要思想是：如果某个词或短语在一个类别中出现的频率较高，并且在其他类别中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类 [29]。TFIDF 的

计算方法实际上是词频与逆文档频率的乘积，即 TF×IDF。词频（Term Frequency）是词 t 在文档 d 中出现的频率，而逆文档频率（Inverse Document Frequency）代表了词 t 的类别区分能力，包含词 t 的文档越少则 IDF 越大。TF 和 IDF 的计算公式分别如（1）和（2）所示。

$$\mathrm{tf}(t,d)=\frac{f(t,d)}{\sum_k f(w_k,d)} \quad \cdots\cdots(1)$$

$$\mathrm{idf}_t=\log\left(\frac{N}{1+df_t}\right) \quad \cdots\cdots(2)$$

其中 $f(t,d)$ 表示词条 t 在文档 d 出现的次数，df_t 表示语料库中包含词条 t 的文档数量，N 表示语料库中全部的文档数量。词条 t 的 TFIDF 权重为：$tfidf_t=tf(t,d)\times idft$。可以看出，词条 t 的权重随着它在文档中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。虽然 TFIDF 能够通过词汇在各类的重要程度实现文本分类，但是由于没有考虑到词汇的语义信息，在短文本分类时往往得不到更好的效果。

Word2vec 是 MIKOLOV 等人 [24] 提出的一种神经网络概率语言模型，可以用于计算单词的词向量。与传统的高维词向量 one-hot representation 相比，Word2vec 词向量的维度通常在 100~300 维之间，减少了计算的复杂度，也不会造成向量维数灾难。除此之外，Word2vec 词向量是根据词汇所在上下文计算出的，充分捕获了上下文的语义信息，很容易通过它计算两个词汇的相似程度。Word2vec 包含了两种训练模型，分别是 CBOW（Continuous Bag-of-words Model）和 Skip_gram。

CBOW 模型通过上下文预测给定词，CBOW 的数学表示如下：

$$P(W_t\mid\tau(W_{t-k},W_{t-k+1},\ldots,W_{t+k-1},W_{t+k}))\cdots\cdots(3)$$

其中 W_t 为语料词典中的一个词，即通过和 W_t 相邻上下文窗口大小为 K 的词来预测词 W_t 出现的概率。τ 运算符表示将上下文窗口相邻的词汇的词向量作相加运算。

Skip_gram 则是通过当前词预测其上下文，即通过词汇 W_t 去预测相邻窗口 k 内词汇的概率。Skip_gram 的数学表示如下：

$$P(W_{t-k},W_{t-k+1},\cdots,W_{t+k-1},W_{t+k}\mid W_t)\cdots\cdots(4)$$

与 CBOW 模型相比，Skip_gram 语义准确率高，代价是模型的计算复杂度高，模型训练耗时较长。CBOW 模型因为窗口大小的限制，导致无法预测与窗口以外词汇的关系。而 Skip_gram 模型会通过跳跃词汇来构建词组，避免了因窗口大小限制导致丢失语义信息的问题。本文实验中使用 Word2vec 的 Skip_gram 模型，其训练复杂度为：

$$Q=C\times(D+D\times\log2(V))\cdots\cdots(5)$$

其中，C 为 Word2vec 模型输入层的窗口大小，D 表示词向量维度，V 表示训练语料的词典大小。

设有训练语料词典 vocab 和文档 $d_i=<w_1,w_2,\cdots,w_j>$，N 是词向量维度。

$$\mathrm{vocab}=\{t_i\mid i\in 1\cdots N\}\cdots\cdots(6)$$

使用 Word2vec 模型训练语料，得到文本中单词词向量，将文本 d_i 中词向量累加得到文本 d_i 的向量表示 $R(d_i)$。其中，$word2vec(t)$ 表示词汇 t 的 Word2vec 词向量。

$$R(d_i)=\sum_t word2vec(t)\ \text{where}\ t\in d_i\cdots\cdots(7)$$

接下来，引入 TF-IDF 模型根据词汇重要程度计算 Word2vec 模型中词汇权重，将加权过的词向量累加得到文档 d_i 新的向量表示 $weight\_R(d_i)$。

$$weight\_R(d_i)=\sum_t word2vec(t)\times w_t\ \text{where}\ w_t=\mathrm{tfidf}_t\cdots\cdots(8)$$

最后，合并加权 word2vec 和 TFIDF 两种模型，得到新的文档向量表示 $C(d_i)$，其中 $tfidf(d_i)$ 表示文档 d_i 的 tfidf 向量表示。

$$C(d_i)=concatenate(\mathrm{tfidf}(d_i),\mathrm{weight\_R}(d_i))\cdots\cdots(9)$$

文本分类之前一般都要经过去停用词等预处理技术，停用词主要包括英文字符、数字、数学字符、标点符号以及使用频率特别高的没有实际意义的汉字（如“的”、“在”、“和”、“以及”等），移除文本中的停用词能改善文本分类效果 [30,31]。

2 实验

2.1 实验数据

本文实验数据集来自于数据堂从新浪微博采集的微博内容，涵盖了 15 个主题下的共 110539 条微博，其中用于训练的数据共 88430 条微博，用于测试数据共 22109 条微博。15 个主题包括 IT、财经、动漫、房产、健康、教育、旅游、美食、女性、汽车、时尚、体育、游戏、娱乐、育儿。每个主题下的微博数量如表 1 所示。

表 1 各类别下微博数量

主题	数量	主题	数量	主题	数量
IT	6295	汽车	7878	美食	7598
财经	7338	时尚	7108	女性	7992
动漫	6960	体育	7452	教育	8006
房产	6451	游戏	7265		
健康	7553	娱乐	7442		
旅游	6756	育儿	8445		

2.2 分类性能评价指标

文本分类的评价指标采用精度（precision）、召回率（recall）、F-score 和准确率（accuracy）。

表 2 是两分类器混淆矩阵（Confusion Matrix），其中 TP（true positive）表示实际为正类、预测也为正类的样本数量；FN 表示实际为正类、预测为反类的样本数量；FP 表示实际为反类、预测为正类的样本数量；TN 表示实际为反类、预测也为反类的样本数量。

表 2 两分类混淆矩阵

	预测正例	预测反例
实际正例	TP	FN
实际反例	FP	TN

精度、召回率、准确率、F-score 定义如下：

$$precision = \frac{TP}{TP+FP} \quad (10)$$

$$recall = \frac{TP}{TP+FN} \quad (11)$$

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (12)$$

实际应用时，需要平衡精度和召回率，通常使用两者的调和平均数作为一个综合的评价指标，称之为 F-score：

$$F\text{-}score = \frac{2\times precision\times recall}{precision+recall} \quad (13)$$

2.3 Word2vec 参数调整

Word2vec 提供了许多参数来调整模型训练过程，不同参数的选择对生成的词向量的质量以及训练的速度都有影响。本文通过分析 Word2vec 参数窗口大小（windows）对分类准确率以及模型训练消耗时间的影响，根据 ARR（adjusted ratio of ratios）算法 [32,33] 调整窗口值的大小。ARR 算法能根据模型准确率和训练消耗时间评估模型的质量。公式（14）评估模型 a_p 相对模型 a_q 在数据集 d_i 上的优势，其中 SR 表示模型的准确率，T 表示模型消耗时间，$AccD$ 是衡量准确率和时间重要性的参数。本文实验中取 $AccD$ 值为 1%，表示选择模型过程中同等重要的考虑准确率和时间消耗两个标准。

$$ARR^{d_i}_{a_p,a_q} = \frac{SR^{d_i}_{a_p} / SR^{d_i}_{a_q}}{1+AccD*\log(T^{d_i}_{a_p} / T^{d_i}_{a_q})} \quad (14)$$

图 1 是使用 Word2vec 模型对 15 类微博进行分类，分类准确率随着 Word2vec 窗口大小变化的曲线。可以看出，模型在窗口大小等于 20 时，分类准确率达到最高。如果窗口设置的过小，可能会导致丢失上下文中一些重要词汇。而上下文窗口太大时，可能会引入太多与待预测词无关的上下文词汇，导致模型训练出的词向量质量差，降低分类准确率。

图 2 是在不同窗口大小的条件下，根据分类准确率与训练模型消耗时间计算的模型 ARR 值，从图 2 中可以看出窗口大小 20 对应模型 ARR 值最大。本文综合模型的分类准确率和训练消耗时间，将 Word2vec 参数窗口设置为 20。

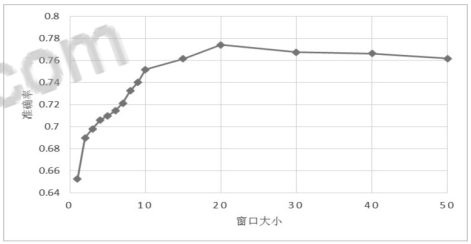

图 1 窗口大小 - 准确率曲线

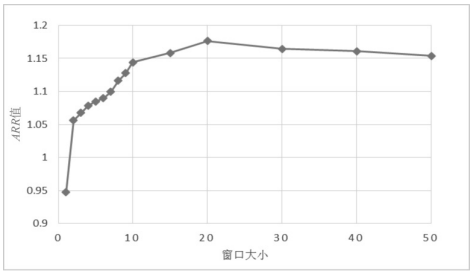

图 2 不同窗口大小 Word2vec 模型 ARR 值

2.4 微博分类

本文将分别使用 TFIDF、Word2vec、加权 Word2vec 以及合并加权 Word2vec 和 TFIDF 这四种模型对实验数据集中的微博进行分类，并分析停用词对分类准确率的影响。

首先使用中科院分词工具 ICTCLSA 进行分词，然后使用语料训练 Word2vec 模型，累加词向量得到文档的向量化表示。使用 Scikit-learn 提供的 TfidfTransformer 模块计算词汇的 TFIDF 权重，将词向量和对应词汇的 TFIDF 权重相乘得到加权 Word2vec 词向量，累加加权词向量得到加权文本向量化表示。对于 TFIDF 分类模型，使用

TfidfVectorizer 模块提取文本特征并将文本向量化。实验最后结合加权 Word2vec 与 TFIDF 两种模型，合并加权 Word2vec 和 TFIDF 向量来表示文本。实验中分类算法使用 Scikit-learn 提供的 LinearSVM 算法。

表 3 是上述四种模型的不同停用词策略在两类微博上分类的实验结果。值得注意的是文本所属类别对结果有很大的影响，例如，运动和农业的分类容易获得较高的分类准确率，因为这两种类别的共现词较少。从表 3 可以看出，单独使用 Word2vec 的分类效果并不理想，而采用 TFIDF 加权后的 Word2vec 模型分类准确率高于 Word2vec 模型，仍略低于 TFIDF 文本分类模型。合并加权 Word2vec 和 TFIDF 两种模型，使用合并后的向量对文本进行表示的方法取得了最高的准确率。此外，去掉文本中停用词能极大改善分类准确率。

表 3 两类别微博分类结果

类别	accuracy	precision	recall	F-score
带停用词 Word2vec	0.9038	0.8746	0.9407	0.9065
去停用词 Word2vec	0.9446	0.9239	0.9680	0.9454
去停用词 TFIDF	0.9797	0.9896	0.9693	0.9794
带停用词 TFIDF	0.9797	0.9889	0.9700	0.9794
加权 Word2vec	0.9706	0.9772	0.9632	0.9702
加权 Word2vec+TFIDF	0.9828	0.9876	0.9775	0.9825

表 3 中的两类别微博分类是最简单的二分类问题，所有待分类的测试样本结果非正即负，因此其中一类的特征好坏可以间接影响另一类的分类结果。在微博短文本分类任务中，不仅需要处理二分类问题，更重要的是处理多分类问题。这就要求分类方法必须有效提取每一类的特征，才能达到不错的分类效果。

本文进一步在 15 类、10 类和 2 类的微博上进行了分类实验。图 3 是 15 类、10 类、2 类微博分类准确率，横坐标分别代表了表 3 中的 6 种分类方法。具体的，1 表示使用带停用词 Word2vec 模型；2 表示去停用词 Word2vec 模型；3 表示去停用词 TFIDF 加权 Word2vec 模型；4 表示去停用词 TFIDF 模型；5 表示带停用词 TFIDF 模型；6 表示合并加权 Word2vec 和带停用词 TFIDF 后的模型。从图 3 中可以看出，类别数量越少，Word2vec 和 TFIDF 分类的准确率越接近；同时类别数量越少，分类的准确率越高。

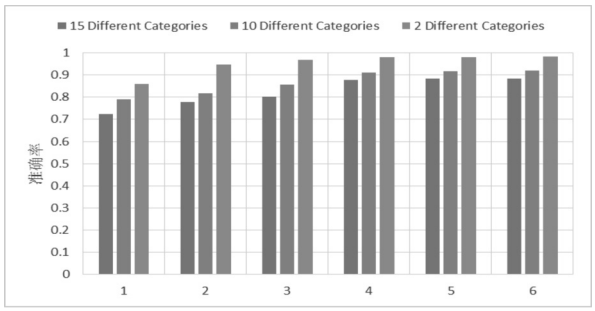

图 3 多类别微博分类准确率

3 结束语

与传统文本向量空间模型相比，Word2vec 模型既能解决传统向量空间模型的高维稀疏特征问题，还能引入传统模型不具有的语义特征，有助于短文本分类。本文基于 Word2vec 模型对微博短文本分类进行研究，鉴于 Word2vec 模型无法识别文本中词汇的权重，本文用 TFIDF 计算 Word2vec 词向量的权重，提出了加权 word2vec 模型，最后合并加权 Word2vec 和 TFIDF 两种模型，基于合并后的模型对微博进行分类。实验结果表明，合并后的模型分类准确度高于传统的文本分类模型。从实验结果还可以得出，Word2vec 分类模型准确度与分类类别、类别数量等因素有关，类别之间文本中的共现词越少，模型分类准确度越高。

在接下来的研究中，将考虑用词向量累加的方式表示短文本可能会丢失词向量中的部分语义信息，利用基于词向量距离的短文本分类算法，通过计算短文本中词汇 Word2vec 词向量距离间接得到短文本相似度并进行分类。（责编 吴晶）

参考文献：

[1] 刘丽清 . 微博虽“微”足值道尔——微博特性之浅析 [J]. 东南传播 , 2009(11):153-154.

[2] 崔争艳 , CUIZheng-yan. 基于语义的微博短信息分类 [J]. 现代计算机 : 专业版 , 2010(8):18-20.

[3] 代六玲 , 黄河燕 , 陈肇雄 . 中文文本分类中特征抽取方法的比较研究 [J]. 中文信息学报 , 2004, 18(1):26-32.

[4] JOACHIMS T. Text Categorization with Support Vector Machines: Learning with many Relevant Features[M]. Berlin Heidelberg: Springer, 1998.

[5] SALTON G, WONG A, YANG C S. A Vector Space Model for Automatic Indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[6] BERRY M W. Survey of Text Mining[J]. Computing Reviews, 2004,

45(9): 548.
[7] BANERJEE S, RAMANATHAN K, GUPTA A. Clustering Short Texts Using Wikipedia[C] // ACM. SIGIR 2007: Proceedings of the, International ACM SIGIR Conference on Research and Development in Information Retrieval, July 23 –27, 2007, Amsterdam, the Netherlands. New York: ACM, 2007:787–788.
[8] HU X, SUN N, ZHANG C, et al. Exploiting Internal and External Semantics For the Clustering Of Short Texts Using World Knowledge[C] // ACM. ACM Conference on Information and Knowledge Management, CIKM 2009, November 2–6, 2009, Hong Kong, China. New York: ACM, November. 2009: 919–928.
[9] LIU Z, YU W, CHEN W, et al. Short Text Feature Selection for Micro–Blog Mining[C] // IEEE. International Conference on Computational Intelligence and Software Engineering, December 10–12, 2010, New York. New York: IEEE, 2010: 1–4.
[10] 王盛，樊兴华，陈现麟 . 利用上下位关系的中文短文本分类 [J]. 计算机应用，2010，30（3）: 603–606.
[11] 范云杰，刘怀亮 . 基于维基百科的中文短文本分类研究 [J]. 现代图书情报技术，2012（3）：47–52.
[12] 赵辉 . 一种基于维基百科的中文短文本分类算法 [J]. 图书情报工作，2013，57（11）：120–124.
[13] 翟延冬，王康平，张东娜，等 . 一种基于 WordNet 的短文本语义相似性算法 [J]. 电子学报，2012，40（3）：617–620.
[14] 宁亚辉，樊兴华，吴渝 . 基于领域词语本体的短文本分类 [J]. 计算机科学，2009，36（3）：142–145.
[15] BOLLEGALA D, MATSUO Y, ISHIZUKA M. Measuring Semantic Similarity between Words Using Web Search Engines[EB/OL]. http://ymatsuo.com/papers/jws08danu.pdf, 2016–12–2.
[16] SAHAMI M, HEILMAN T D. A Web–based Kernel Function for Measuring the Similarity of Short Text Snippets[C] //IEEE. International Conference on World Wide Web, WWW 2006, May 23–26, 2006, Edinburgh, Scotland, UK. NewYork: IEEE, 2006: 377–386.
[17] 王鹏，樊兴华 . 中文文本分类中利用依存关系的实验研究 [J]. 计算机工程与应用，2010, 46（3）: 131–133.
[18] ZELIKOVITZ S, HIRSHi H. Improving Short Text Classification Using Unlabeled Background Knowledge to Assess Document Similarity[C] // IEEE. Proceedings of the seventeenth international conference on machine learning, June 29 – July 2, 2000, San Francisco, USA. San Francisco: Morgan Kaufmann Publishers Inc, 2000: 1183–1190.
[19] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation[J]. the Journal of Machine Learning Research, 2003(3): 993–1022.
[20] PHAN X H, NGUYEN L M, HORIGUCHI S. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large–scale Data Collections[C] // 北京航空航天大学，国际万维网会议委员会 . The 17th International Conference on World Wide Web，April 21–25, 2008, Beijing, China. Beijing：国际万维网会议委员会，2008:91–100.
[21] CHEN M, JIN X, SHEN D. Short Text Classification Improved by Learning Multi–granularity Topics[C] // International Joint Conference on Artificial Intelligence, July 16–22, 2011, Barcelona, Catalonia, Spain. New York: AAAI Press, 2011:1776–1781.
[22] 方东昊 . 基于 LDA 的微博短文本分类技术的研究与实现 [D]. 沈阳：东北大学，2011.
[23] 吕超镇，姬东鸿，吴飞飞 . 基于 LDA 特征扩展的短文本分类 [J]. Computer Engineering and Applications, 2015, 51（4）: 6–7.
[24] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013（1）: 28–36.
[25] LILLEBERG J, ZHU Y, ZHANG Y. Support Vector Machines and Word2vec for Text Classification with Semantic Features[C]// IEEE, International Conference on Cognitive Informatics & Cognitive Computing, July 6–8, 2015, Beijing, China. New York: IEEE, 2015:136–140.
[26] 李静梅，孙丽华 . 一种文本处理中的朴素贝叶斯分类器 [J]. 哈尔滨工程大学学报，2003, 24（1）: 71–74.
[27] 张士豪，顾益军，张俊豪 . 基于用户聚类的热门微博分类研究 [J]. 信息网络安全，2015（7）：84–89.
[28] 施聪莺，徐朝军，杨晓江 . TFIDF 算法研究综述 [J]. 计算机应用，2009, 29(B06): 167–170.
[29] 张玉芳，彭时名，吕佳 . 基于文本分类 TFIDF 方法的改进与应用 [J]. 计算机工程，2006, 32（19）: 76–78.
[30] 张越今，丁丁 . 敏感话题发现中的增量型文本聚类模型 [J]. 信息网络安全，2015（9）：170–174.
[31] PATEL B, SHAH D. Significance of Stop Word Elimination in Meta Search Engine[C]//IEEE. International Conference on Intelligent Systems and Signal Processing, March 1–2, 2013, G H Patel College of Engineering and Technology, Vallabh Vidyanagar, Gujarat, India. New York: IEEE, 2013:52–55.
[32] BRAZDIL P B, SOARES C, DA COSTA J P. Ranking Learning Algorithms: Using IBL and Meta–learning on Accuracy and Time Results[J]. Machine Learning, 2003, 50(3): 251–277.
[33] WOLF L, HANANI Y, BAR K, et al. Joint word2vec Networks for Bilingual Semantic Representations[J]. International Journal of Computational Linguistics and Applications, 2014, 5(1): 27–44.