基于分布式词向量的文本分类探究

牛坤[1]，张欢欢[2]
NIU Kun[1], ZHANG Huanhuan[2]

华东理工大学，信息科学与工程学院，上海，200237
School of information Science and Engineering, East China University Of Science and Technology, Shanghai 200237, China

NIU Kun, ZHANG Huanhuan. Research on document classification based on distributed word representation. Computer Engineering and Applications

Abstract: Text classification is constrained by high dimension text representation and the missing of word context information, a text categorization algorithm named Distributed Representation and TFIDF Based Document Classification (DRTB-DC) based on word distributed representation is proposed. The algorithm using Word2vec to carry on the training of distributed representation model to the data set with context information retained, then using the tfidf algorithm to extract the keywords from the uncategorized text, Finally, getting the text categorization with keywords and classes set matrix calculated by the distributed representation model. The experimental results show that the proposed method can achieve effective performance in two classification and three classification experiments.

Key words: text classification; word2vec; distributed word representation

摘 要：针对文本分类方法中表示文本向量的维数较高，词的上下文信息丢失这一问题，提出基于分布式词向量的文本分类方法。该方法使用 word2vec 对收集到的大规模数据集进行处理得到分布式词向量模型，该模型可以保留训练语料中词的上下文信息，然后使用 tfidf 算法提取出待分类文本的关键词，使用分布式词向量模型计算得到关键词集合与类别集合的距离矩阵，从而判断得出待分类文本所属类别。实验证明，在二分类和三分类试验中，该方法取得了较高的准确率。

关键词：文本分类；Word2vec；分布式词向量

文献标志码：A　　中图分类号：TP391

1 引言

文本分类是对已有文本集合中的每一篇文档进行类别划分，使得每篇文档都尽可能正确的分到相应的类别中。国外文本分类工作可追溯到上世纪 50 年代，Luhn[1]首次提出将词频统计应用到文本分类当中。二十世纪 90 年代，基于机器学习的算法[2]相继被提出，并且很快成为文本分类领域的主流技术。常见的机器学习算法有朴素贝叶斯、K-近邻、决策树、支持向量机等，这些算法在文本分类领域取得很大进展。现有文本分类算法大多是以向量空间模型为基础，在此基础上的特征选择、权重计算等方面也取得较大进 展 ,克服了传统分类模型效率低、灵活性差的问题。从文本分类方法角度看，目前流行的文本分类技术有三种模式：基于规则的文

作者简介：牛坤(1992－),男,硕士研究生,主要研究方向为机器学习,神经网络,E-mail:niukun0813@163.com;张欢欢 (1969－),女,博士，副教授，硕士生导师，主要研究方向为形式化方法与验证技术。

收稿日期：2016-00-00；修回日期：2016-00-00

本分类、基于连接的文本分类和基于统计的文本分类[3]。

国内文本分类相关的研究起步较晚，二十世纪 80 年代汉清教授对文本分类做了相关介绍工作[4]，二十世纪 90 年代后期才有中文文本分类相关研究，这些研究主要集中在中科院计算所、清华大学、哈尔滨工业大学等高等学府[5]。

这些方法虽然能够有效的进行文本分类，但是仅保留词级别信息，词的上下文信息在分类过程中却被遗失掉，且对大量文本的分类效率较低。本文介绍一种基于分布式词向量和 tfidf 的文本分类算法（Distributed Representation and TFIDF Based Document Classification 简称: DRTB-DC），该方法能够有效利用训练语料中词的上下文信息。实验证明，此方法是可行的。

2 相关技术

2.1 词向量

2.1.1 One-hot Representation

在自然语言处理（Natural language processing，简称 NLP）中，最常见的一步是创建一个词库表并把每个词顺序编号。这就是词表示方法中的 One-hot Representation，这种方法把每个词顺序编号，每个词表示为一个很长的向量，向量的维度为所选语料对应词库表的大小。其中只有对应位置上的元素值为 1，其余为零。在实际应用中，一般采用稀疏编码存储，采用此的编号。

这种表示方法一个最大的缺点是无法捕捉词与词之间的相似度，就算是近义词也无法从词向量中看出任何关系，如："西红柿"表示为：$[0,\ 0,\ 0,\ldots 0,\ 1,\ ]^T$，"番茄"表示为：$[0,\ 1,\ 0,\ 0,\ldots\ldots 0,\ 0,\ 0]^T$，即使"西红柿"与"番茄"表示的是同一事物，但是从词向量无法看出两者的联系。此外这种表示方法还容易发生维数灾难，尤其是在深度学习相关的应用中。

2.1.2 Distributed Representation

Distributed Representation（分布式表示）最早是 Hinton 于 1986 年提出[6]，可以克服 One-hot Representation 的缺点。其基本思想是通过训练将语料中的每一个词映射成一个固定长度的词向量，将所有词对应的向量放在一起形成一个词向量空间，而每一个词向量则为该空间中的一个点，在这个空间中引入"距离"，则可以根据词对应向量之间的距离来衡量词与词之间的(词法、语义上的)相似性。

与 One-hot Representation 相比，词的分布式表示使用维度更低的稠密向量，维度的值可以人为指定，以几十或者几百为常见，一般远小于词库表的大小。不同的词之间的语义相似度可以使用词所对应的实数向量之间的距离来判断，可以使用传统的欧氏距离来衡量。

2.2 神经网络语言模型

Bengio 等[7]提出一个基于神经网络的语言模型，即神经网络语言模型(Neural Network Language Model，NNLM)，如图 1 所示

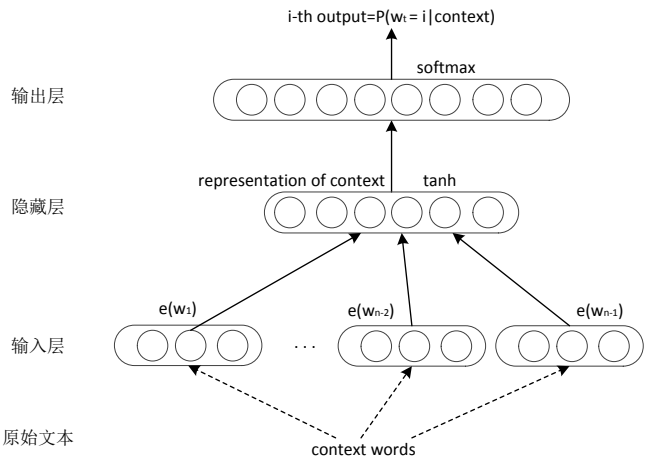

图 1 神经网络语言模型（NNLM）结构图

NNLM 可以计算一段上下文的下一个词为 w_i 的概率，即 $p(w_i|\ context(w_i))$，NNLM 采用 Distributed Representation 来表示词向量，即每个词被表示为一个浮点向量，并且词向量是 NNLM 训练得到的副产物。

词向量模型是具体的词与向量的集合，离线存储表现为一个向量文件。一般情况下向量的维数取值为几十到几百不等，需要根据实际情况来判断使用哪一数量级维度的向量模型更适合具体问题的解决，通常需要大量的实验进行选择。

2.3 CBOW 模型和 Skip-gram 模型

Word2vec 是 Mikolov 等[8][9][10]所提出的神经网络语言模型的一个实现，可以用来快速训练词向量。Word2vec 内含两种训练模型，分别是连续词袋模型(Continuous Bag-of-Words Model ,简称 CBOW)和 Skip-gram[11]模型，如图 2 所示。

从图 2 中可以看出，CBOW 模型和 Skip-gram 模型都包含输入层、隐含层和输出层。其中，CBOW 模型是通过上下文来预测当前词，而 Skip-gram 模型是使用当前词来预测其上下文。模型在训练过程中充分利用靠近词语的上下文，这使得语料中词的上下文信息得以保留，模型的输入是文本语料，输出是一个词汇表 ,其中每个词都有一个对应的向量，即本文 2.1.2 节描述的 Distributed Representation 分布式词向量。

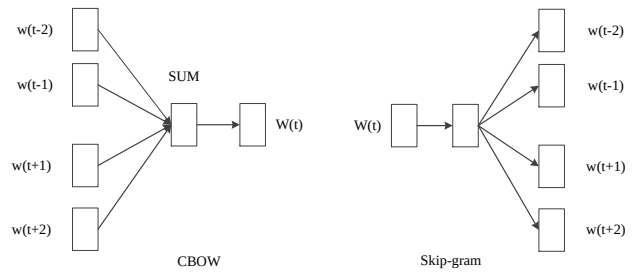

图 2 CBOW 模型和 Skip-gram 模型

2.4 语义相似度计算

每个词对应的向量都包含了该词在训练语料中所处的上下文信息。词与词之间的语义相似度可以通过词所对应的词向量之间的余弦值来衡量，词向量 d_i 和 d_j 之间的语义相似度计算如公式（1）:

$$\mathrm{sim}\left(d_i, d_j\right)=\frac{\sum_{k=1}^{n}\left(w_{ik} * w_{jk}\right)}{\sqrt{\left(\sum_{k=1}^{n} w_{ik}{}^{2} * \sum_{k=1}^{n} w_{jk}{}^{2}\right)}} \quad (1)$$

两个词之间无相似性表示为它们所对应分布式词向量之间的余弦值接近于 0，而相似度高表示为两个词之间的余弦值接近于 1。实验使用搜狗实验室提供[12]的搜狐新闻数据（SogouCS）和全网新闻数据(SogouCA)，使用默认配置训练 200 维词向量，如找到与词“法律”最相近的 10 个词语，输出它们与“法律”之间的余弦值，如表 1 所示:

表 1 与“法律”相近的单词排列

编号	单词	与单词“法律”之间的余弦值
1	法规	0.7666702270507812
2	规章	0.7154861092567444
3	条文	0.6712163686752319
4	合法	0.6702567934989929
5	相对人	0.66758131980896
6	违反	0.6529091596603394
7	规范性	0.6447259187698364
8	强制力	0.6442285776138306
9	规定	0.6384625434875488
10	诉讼法	0.6377309560775757
.....	.....	.....

3 基于分布式词向量的文本分类算法

DRTB-DC 算法以训练得到的分布式词向量模型为基础，使用 tfidf 算法抽取出待分类文本的 K 个关键词，算法中该 K 个关键词从语义上代表该待分类文本；使用分布式词向量模型计算得到关键词集合与类别集合的距离矩阵，得到各关键词与待分类别之间的语义相似度，找到 K 个关键词分别所属的类；某一关键词在分类活动中为文本分类产生的影响称为贡献（contribution），贡献的值为上一步中的语义相似度乘以该关键词的 tfidf 值，计算所得平均贡献最大的类，将此类别作为待分类文本所属类别并返回。

令 d 表示待分类文本，classes 表示类别集合，finalClass 表示最终分类结果，则 DRTB-DC 算法伪代码见算法 1：

算法 1 DRTB-DC 算法

输入 待分类文本 d，类别集合 classes

输出 文本所属分类 finalClass

Begin

1. keywords = getKeyWordsByTFIDF(d);//使用 TF-IDF 算法抽取出该文本的 K 个关键词；
2. foreach keyword[i] in keywords
3. keyClass = getNearestClass(keyword[i]);//关键词与每个类进行余弦值计算，得到与该关键词最近的类的距离
4. save(keyword[i]*tfidf, keyClass);//保存该关键词对文本分类所做的贡献和该关键词所属类别
5. foreach class[i] in classes

6. score[i] = getAverageScore(classes);// 计算每个类别所得平均贡献;

7. finalClass = Max(score); //选取出得到平均贡献最大的类作为最终分类结果

End

DRTB-DC 算法的关键是，使用合适的特征提取方法提取出待分类文本的关键词及确定关键词个数 K，以及使用高质量的词向量模型计算每个关键词所属的类别和该关键词所产生的贡献。关键词提取使用 tfidf 方法 ,选择待分类文本中 tfidf 值最高的前 K 个词返回。为获得高质量的词向量，本文参考文献[14]的结论，对语料进行一系列预处理，提高语料的纯度，并对适合文本分类的词向量模型配置进行了实验探究。

4 实验结果与分析

4.1 数据准备

4.1.1 语料搜集

本实验使用的是数据堂[13]提供的面向文本分类研究的中文新闻语料库。该数据从凤凰、新浪、网易、腾讯等版面搜集，搜集时间段为 2009 年 12 月—2010 年 3 月。中文新闻语料分为 8 类：阅读，娱乐，历史，教育，社会，文化，科技，军事。数据规模如表 2 所示:

从数据堂下载的数据文件解压后得到 mdf 格式文件，首先附加到 SQL Server 数据库中，然后从数据库中取出所需要的中文文档，包括训练集和测试集，并将其作为原始数据进行预处理。

表 2 FinallyCorpus 数据集

类型	表单名称	文章 ID 范围	类别数目	是否为平衡语料
训练集	NewsTrainingCorpus	1- 13026	8	否
测试集	ReteursTestingCorpus	1-3254	8	否

数据库名称：FinallyCorpus.mdf，FinallyCorpus.ldf

数据格式：文本（MSSQL MDF 格式数据库）

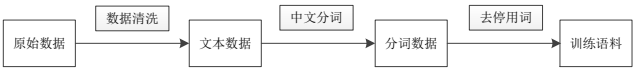

图 3 文本预处理流程图

4.1.2 数据预处理

要生成词向量，首先需要对获取的数据进行预处理操作，本实验中预处理操作流程如图 3 所示:

（1）数据清洗：此阶段针对数据库中中文新闻语料训练集和测试集

此步骤需要花费大量的时间，分析得到的文本数据，找出其中与分类无关的信息，如："凤凰社实习记者 张三报道"、"更多新闻请点击链接....."等类似无用信息，建立规则把这些无用信息过滤掉，初步完成对数据集的清理工作。常见的无用信息举例如表 3:

表 3 无用信息总结

类别	举例
新闻头部	**年**月**日 15:44 凤凰网文化专栏【大中小】【】0 位网友发表评论
作者信息	作者： 张三 编辑：李四
附加数据	(摘自加拿大《北美时报》作者:**法师）
声明信息	所有评论仅代表网友意见，凤凰网保持中立。
广告链接	更多文化内容请点击**
占位字符	[1]【】【】【】【】【】【大中小】【发表评论(0)】
固定格式	[责任编辑: 张三]
地址、电话信息等	[配送地址]**路**号**楼 1107，坐地铁到**站下
.....	.....

（2）中文分词

过滤文本中无用信息之后，实验采用张华平博士研发的 NLPIR 分词工具包对文本语料进行分词，不保留词性标记信息。

（3）消除停用词

本实验中使用哈工大停用词表 ,并在此基础上，根据语料特点，对停用词表进行了拓展。

表 4 描述中文新闻语料库经过预处理之后，各个类别文本数量的统计情况。

表 4 中文新闻语料库文本分布情况

类别	文化	教育	娱乐	历史	科技	军事	阅读	社会
数量	2410	80	913	1302	283	1149	4558	3145

最终得到适合用于词向量训练的文本，中文新闻语料库经过预处理后得到大小为 58.5M。文本中的每一行代表原数据集中一篇文档，这种表示方法为计算整个训练数据集计算 IDF 提供基础。

4.1.3 词向量维度选择

经过预处理的语料已经符合训练词向量的条件，本文选取不同的配置参数训练数据模型，得到不同的词向量文件。选取向量维度这一对词向量效果影响较大的参数进行配置。选取向量维数为 50~500 维，以 50、100、500 递增对语料进行训练，训练得到不同配置参数下的词向量模型。随机从娱乐、军事、社会每个类别中抽取 200 条数据进行实验，选取关键词个数为 5，以分类准确率为结果进行统计，结果如图:

由图 4 可知，维度选取偏大或者偏小都会对实验结果产生影响，本实验中维度为 200 的时候分类效果表现稳定，更适合用于分类算法当中。本文以下实验均采用 200 维的词向量进行计算。

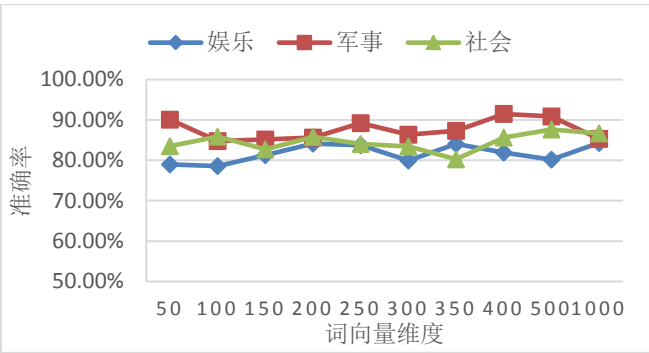

图 4 三个类别数据在不同维度下分类准确率

4.1.4 评价标准

为了对文本分类效果进行分析评估以及分类精度的对比，采用准确率(precision)、召回率(recall)、综合准确率与召回率考虑的 F 值 (F-measure) 和宏平均来进行测量。

表 5 分类结果混合矩阵

	真正属于该类别文档	真正不属于该类别文档
判断属于该类别文档	A	B
判断不属于该类别文档	C	D

准确率(P)=A/(A + B)

召回率(R)= A/(A + C)

F 值(F)=2PR/(P+R)

宏平均指所有类别的测试结果的算数平均值

4.2 二分类实验

本实验探究在其他条件相同的情况下，关键词个数对二分类实验结果的影响。实验采用的是中文新闻语料测试集中娱乐、军事两个类别的数据，使用中文新闻语料的训练集训练词向量，词向量的维数选取最常用的 200 维，使用上一章节中处理得到的测试集，每个类别随机抽取 200 条数据进行实验。

最终得到分类结果的准确率按照类别统计，如图 5 所示:

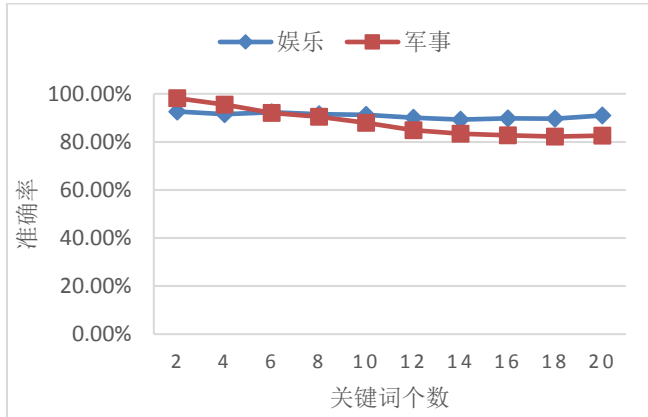

图 5 关键词数量不同条件下的二分类准确率

由此结果可以看出：(1) 在关键词数较少的时候，比如关键词小于 7 的时候，两个类别的测试结果准确率均在 90.0%以上；(2) 但是随着关键词个数的增加，准确率却逐渐递减，经分析，出现这种趋势的原因是：

(1)使用 DRTB-DC 算法提取出的关键词能够很好的表征待分类文本的某些上下文信息，这是因为在使用关键词与类别判断的基础上，引入了 tfidf 权值，这样可以根据该关键词对文本的重要程度对做出相应的调整，使对文本比较重要的词获得较大分数，而对文本影响不大的词获得相对较小的分数；

(2) 关键词数量增多，分类准确率反而减少。由上一部分的分析，由于本文提出的算法设置，后面新添加的关键词的 tfidf 相对较小，在分类文档过程中的取得的分数越小，对分类结果的影响就越弱，因此关键词数越多反而会影响文本分类的结果。

4.3 三分类实验

与二分类实验设置相同，本实验对娱乐、军事、社会三个类别的测试数据进行分类。

最终得到分类结果的准确率按照类别统计，如图 6 所示:

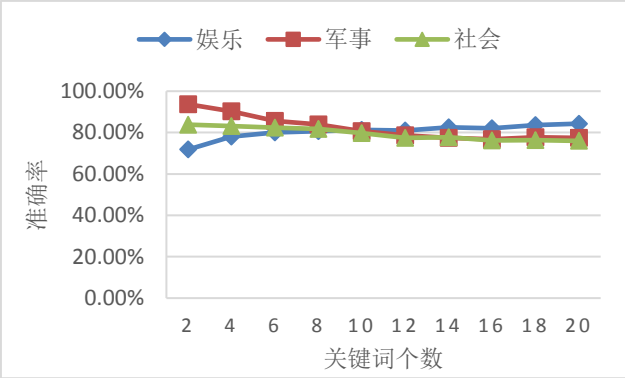

图 6 关键词数量不同条件下的三分类准确率

由图 4.6 可以看到三分类的结果趋势与二分类基本一样，但是整体准确率比二分类实验偏低。产生这一现象的原因是：随着待分类别的增加，不同类别之间会产生影响，如在二分类中某个关键词分类类别 1 的概率为 70% ,分类类别 2 的概率为 50%，则会将此关键词分到类别 1 下；在三分类中，如果算法判别该关键词属于类别 3 概率为 80%，那么在三分类中就会把该词分到类别 3 中。

对比图 4.5 和图 4.6 可以发现 ,二分类和三分类的实验结果趋势整体是一致的。由图 4.5 可知，二分类结果准确率在关键词数量不超过 7 的时候，分类准确率均在 91%以上；由图 4.6 可知，二分类结果准确率在关键词数量不超过 7 的时候，分类准确率均在 80%以上。

同时也应注意到不管是二分类实验还是三分类实验，关键词数量为 1 或 2 的时候，分类准确率非常高，然而高准确率并不意味着分类效果非常好。在只有极少数关键词个数的时候，一般情况下会选择待分类文档中得分最高的关键词所属的分类作为最终结果，但是因为中文新闻语料规模的限制，不能把所有的词汇都包含在内，有可能会出现一个从待分类文本选出一个关键词而这个关键词在训练的模型中没有对应的表示，这种情况的存在使得分类过程会过滤掉一部分文本，这些文本不能正常的参与到分类活动中。出现这一限制的根本原因是语料没能充分包含所有词汇。从另外一方面去分析，仅仅取出一个关键词或者两个关键词来代表一篇文本，这不符合本文所提算法的思想，即从文本中抽取的信息应尽可能的保留原文本词的上下文信息，显然仅使用一个或者两个关键词做法本身也是不合理的。

综上分析，试验中最佳关键词选取个数可为 3 至 7 中任意整数。

4.4 中文新闻数据集测试及算法对比

本节使用中文新闻语料中的娱乐、军事、社会三个类别的数据进行实验，使用本文提出的 DRTB-DC 算法进行分类实验，每个类别随机选取 500 条数据参与分类，词向量模型训练采用默认配置，向量维度设置为 200 维，关键词个数选择 5，即 K=5，结果如表 6：

表 6 分类结果

	Precision/%	Recall/%	F-measure/%
娱乐	87.47	89.57	88.51
军事	88.26	81.34	84.65
社会	83.20	87.47	85.29

将本文提出的算法与文献[15]中的 TFIDF+SVM 和 LDA+SVM 方法在中文新闻语料上进行对比，使用文献中推荐的，训练数据集与测试数据集数据比例为 9:1 ,对比结果采用宏平均值 ,如图 7 所示。

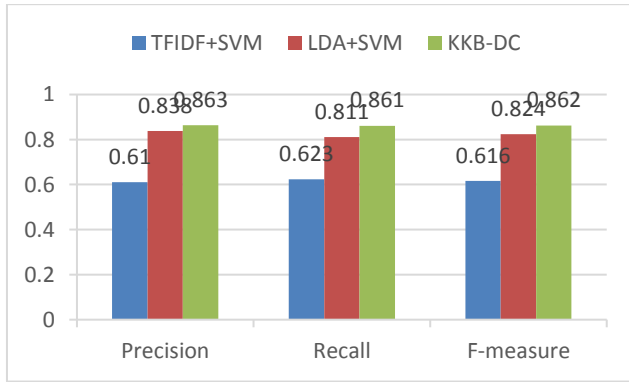

图 7 三种方法的准确率、召回率和 F 值的对比

从图 7 中结果可以看出，在使用中文新闻数据集进行的文本分类中，本文提出的 DRTB-DC 算法的分类效果好于文献[15]中基于 TFIDF 的 SVM 分类算法和基于 LDA 的 SVM 分类算法，本算法无论在准确率、召回率还是 F 值均有一定优势。

4.5 结果分析

综上实验分析，本文提出的 DRTB-DC 算法能够比较稳定的把属于不同类别的文本在二分类和三分类的条件下取得较好的分类结果。尤其是二分类结果，部分情况下准确率可以达到 94.64%。通过对

本节实验结果的分析，可总结出如下几点：

（1）在对语料进行充分预处理并且得到高质量词向量的情况下，在关键词较少时下分类效果相对较好，这表明算法能够较好的提取出能够表示文本信息的关键词，而且这些关键词也能够基本在语义方面代表待分类文本；

（2）在关键词数量极少的情况下，虽然准确率比较高，但是可信度较低，原因是受语料限制，不能充分覆盖所有的字、词，在关键词个数极少的时候能够参与准确分类活动的并非所有待测文本，因此本文对于关键词数量极少的情况仍有较高准确率这一现象暂持保留态度；

（3）在本实验所用语料条件下，实验得出最适合参与算法的词向量的维数是 200 维。

5 结束语

本文针对文本分类中的文本表示维度较高和词的上下文信息丢失这一问题，提出了基于分布式词向量的文本分类算法，该算法利用已有工具 word2vec 训练得到分布式词向量模型，通过 tfidf 算法得到待分类文本的关键词，并通过对这些关键词与类别的语义相似度计算得出该文档的分类。实验证明，本算法可以有效保留训练数据集中词的上下文信息，通过与现有分类模型进行对比实验，证明所提算法在二分类、三分类的有效性。将来的工作重点将放在使用大规模相关语料训练分布式词向量模型，探讨算法在多分类实验的分类效果。

参考文献

[1] Luhn H P. Auto-encoding of documents for information retrieval systems[M]. IBM Research Center, 1958.

[2] 张振峰 .基于向量空间模型的文本分类算法研究[D] .杭州电子科技大学，2011

[3] 祝晓鲁，白振兴，贾海燕. 自动文本分类技术研究[J]. 现代电子技术, 2007, 30(3):121-124.

[4] 侯汉清. 分类法的发展趋势简论[J]. 情报科学, 1981(1):58-63.

[5] 王明亚.基于词向量的文本分类算法研究与改进[D].华东师范大学，2016.

[6] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088):533-536.

[7] Bengio Y, Ducharme R, jean, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3(6):1137-1155.

[8] Mikolov T, Yih W T, Zweig G. Linguistic regularities in continuous space word representations[J]. In HLT-NAACL, 2013.

[9] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[10] 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.

[11] Guthrie D, Allison B, Liu W, et al. A Closer Look at Skip-gram Modelling[C]// 2006:1222--1225.

[12] 搜狗全网新闻数据：https://www.sogou.com/labs/resource/ca.php

[13] 数据堂面向文本分类研究的中文新闻语料库 http://more.datatang.com/data/13484

[14] Lai S, Liu K, He S, et al. How to generate a good word embedding[J]. IEEE Intelligent Systems, 2016, 31(6): 5-14.

[15] Wu X, Fang L, Wang P, et al. Performance of using LDA for Chinese news text classification[C]//Electrical and Computer Engineering (CCECE), 2015 IEEE 28th Canadian Conference on. IEEE, 2015: 1260-1264.

联系人：牛坤
通讯地址：上海市徐汇区梅陇路 130 号华东理工大学
邮编：200237
邮箱：niukun0813@163.com
电话：15021597239