文本分类TF-IDF算法的改进研究

叶雪梅[1,2],毛雪岷[1,2],夏锦春[1,2],王 波[1,2]
1.合肥工业大学 管理学院,合肥 230009
2.合肥工业大学 过程优化与智能决策教育部重点实验室,合肥 230009

摘 要:中国互联网环境的发展,让大量蕴含丰富信息的新词得以普及。而传统的特征词权重TF-IDF(Term Frequency and Inverted Document Frequency)算法主要考虑TF和IDF两个方面的因素,未考虑到新词这一新兴词类的优势。针对特征项中的新词对分类结果的影响,提出基于网络新词改进文本分类TF-IDF算法。在文本预处理中识别新词,并在向量空间模型表示中改变特征权重计算公式。实验结果表明把新词发现加入文本预处理,可以达到特征降维的目的,并且改进后的特征权重算法能优化文本分类的结果。

关键词:新词;词频-逆文档频率(TF-IDF);向量空间模型;文本分类
文献标志码:A **中图分类号:**TP391 doi:10.3778/j.issn.1002-8331.1805-0071

叶雪梅,毛雪岷,夏锦春,等.文本分类TF-IDF算法的改进研究.计算机工程与应用,2019,55(2):104-109.
YE Xuemei, MAO Xuemin, XIA Jinchun, et al. Improved approach to TF-IDF algorithm in text classification. Computer Engineering and Applications, 2019, 55(2):104-109.

Improved Approach to TF-IDF Algorithm in Text Classification

YE Xuemei[1,2], MAO Xuemin[1,2], XIA Jinchun[1,2], WANG Bo[1,2]
1.School of Management, Hefei University of Technology, Hefei 230009, China
2.Key Laboratory of Process Optimization and Intelligent Decision-Making(MoE), Hefei University of Technology, Hefei 230009, China

Abstract: With the development of Internet environment in China, a lot of new words with rich information have been popularized. The traditional term weight algorithm named TF-IDF(Term Frequency and Inverted Document Frequency) mainly considers two factors named TF and IDF without the advantage of new words. In view of the influence of new words in feature items on classification results, an improved TF-IDF algorithm based on new words of network is proposed in text classification. Research recognizes new words in the text preprocessing, and improves the weight calculation formula of them in the vector space model representation. Experimental results show that adding new word discovery process to text preprocessing can reduce feature dimension, meanwhile, the improved TF-IDF algorithm can optimize the result of text classification.

Key words: new words; Term Frequency and Inverted Document Frequency(TF-IDF); vector space model; text classification

1 引言

随着互联网的发展,网络成为用户获取信息的主要渠道,而信息的爆炸式增长使得用户难以从海量数据中获得需要的信息。为提升用户体验,对网络信息进行分类变得越来越重要。文本分类是指通过分类算法对未知类别的文档进行处理,判断它所属的预定义类别[1]。目前存在的分类算法主要有Bayes算法[2]、KNN算法[3]、支持向量机(SVM)算法[4]、神经网络[5]等。

分类算法主要建立在向量空间模型的基础上,特征项权重算法的优劣直接影响文本分类的精准度。TF-IDF算法是经典的特征项权重算法,众多学者就TF-IDF算法存在的问题提出不同改进方案。徐凤亚和罗振声考虑特征项在类间和类内的分布情况,对分布信息和低频高权特征信息进行联合加权[6]。Soucy等人基于对词语

基金项目:安徽省年度重点科研项目计划(No.JZ2016AKKG0825);国家自然科学基金创新群体项目(No.71521001)。
作者简介:叶雪梅(1994—),女,硕士,研究领域为数据挖掘,E-mail:1902873071@qq.com;毛雪岷(1974—),男,副教授。
收稿日期:2018-05-07 **修回日期:**2018-10-22 **文章编号:**1002-8331(2019)02-0104-06
CNKI**网络出版:**2018-11-05, http://kns.cnki.net/kcms/detail/11.2127.TP.20181101.1214.029.html

重要性的统计估计提出一种新的加权方法对TF-IDF进行改进[7]。熊忠阳等人考虑特征项在类间、类内和不完全分类的分布信息的不足,引入特征项在类间和类内分布的离散度来改进TF-IDF[8]。郭红钰综合考虑权重计算时特征项在各类别中的分布,提出了一种基于类别分布的权值计算的方法ETFIDF[9]。现有算法改进研究主要集中在算法的本身缺陷,忽视了文本表达方式变化带来的影响。

网络信息是一类特殊信息,形式自由多样,内容更新频繁,在中国互联网环境下容易产生大量新词。新词一般伴随着社会热点产生,具有不同于基础词汇的新形式与新用法[10],往往蕴含着丰富信息,以此改进TF-IDF算法将大大优化文本分类的结果。因此,本文针对中国互联网环境提出了基于新词发现改进的特征权重算法,在文本预处理中识别新词并对原始文本重新分词,在向量空间模型表示中根据新词信息量大的特点重新分配权重进而提高分类器的性能,即提高网络信息分类的正确率。

2 文本分类的实现

2.1 文本分类流程

2.1.1 分词处理

分词是使用分词算法把文本切割成单个字词、词语或短语的过程。对于英文文本来说,单词是用空格分割的,英文文本可以直接使用空格进行切分而不会产生歧义。但是,中文文本的字、词、短语之间没有间隔,它们是以连续的字符串形式呈现的。随着互联网和移动终端的普及,以及微博、微信、QQ等社交媒体的广泛使用,类似于“戏精”、“二次元”等新词大量出现并迅速传播。网络媒体中出现的新词能反映出社会的热点事件,当热点事件的热度降低时,新词却仍可以保留。新词的出现使得中文分词的准确率降低,由新词导致的分词错误率日益上升,因此有必要将新词加入用户词典,来提高分词的准确性。

2.1.2 特征词选择

当文档集中的中文文本数目过千时,每篇文档分词所得的词语数量将会大大增加,如果将这些词语全部作为特征项,在后期计算机处理时,将会大大增加程序运行的空间复杂度和时间复杂度。因此,对分词结果进行特征降维尤为重要。特征降维的方法有两种,一种是特征选择(Feature Selection),就是从原始全部的空间维度中选取最具表征信息的部分维度,这个新的特征集是原始特征的子集;另一种是特征提取(Feature Extraction),它通过一种映射(或变换)的方法将原始高维的特征向量映射到低维的特征空间[11]。本文使用特征提取来减少特征词的数量,将文本数据转化为结构化的数据,从而达到降维的目的。

2.1.3 文本表示

文本是非结构化数据,而分类算法所能处理的是结构化的数字数据。所以把文本从非结构化转化为结构化这一过程是整个文本分类工作的基石,这一转化过程的好坏将直接影响到最终的分类结果。目前常用的文本表示方法有:布尔模型(Boolean Model)[12]、概率模型(Probabilistic Model)、向量空间模型(Vector Space Model)[13],其中又以向量空间模型(简称VSM)的使用最为简单。

VSM模型中有三个重要概念:一是特征项(Term),特征项是经过特征选择所得的最能表达文本内容的词语;二是特征项权重(Weight),特征项权重是使用相应的特征权重算法对特征词赋值所得的权值;三是特征向量(Feature Vector),特征向量是用特征项和特征项权重共同表示的文本数字化向量。

VSM把每篇文档都表示为特征词-权重向量,把文本看作是一系列特征项 t 的集合,对每个特征项赋予对应的权值。特征项 $t_1, t_2, \cdots, t_n$ 可以看作是一个 n 维坐标系,而权值 $w_1, w_2, \cdots, w_n$ 表示其对应的坐标值,每篇文档 d_i 映射为该向量坐标空间中的一个特征向量 $V(d_i)=(t_1, w_{i1}; t_2, w_{i2}; \cdots; t_n, w_{in})$。文档集总体的VSM表示见表1。

表1 文本的向量空间模型

	t_1	t_2	t_3	$\cdots$	t_n
d_1	w_{11}	w_{12}	w_{13}	$\cdots$	w_{1n}
d_2	w_{21}	w_{22}	w_{23}	$\cdots$	w_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
d_m	w_{m1}	w_{m2}	w_{m3}	$\cdots$	w_{mn}

2.2 文本分类算法

文本分类算法是采用特定的训练方法训练类分类器,当类分类器训练完成之后,通过比较测试样本与类分类器的相似度,判断其所属的预定义类别。目前存在多种基于VSM模型的分类算法,常用的有SVM支持向量机算法,朴素贝叶斯算法,K 近邻算法(K-Nearest Neighbor,KNN)。在中文文本分类领域,SVM算法对核函数的选择缺乏指导,再加上核函数类型的限制,难以对具体的分类问题选择最佳的核函数。在朴素贝叶斯算法中,文本属于某个类别的概率等于该文本中每个词属于该类别概率乘积,而每个词所属的类别概率是用该词在该类别训练文本中出现的概率近似表示。但是组成文本的词语并不是相互独立的,而且该算法只有在训练样本数量非常多的情况下,才会取得比较好的效果。K 近邻算法是基于某种距离度量找出训练集中与给定测试样本最靠近的 k 个训练样本,然后综合这 k 个邻居的类别作为该测试样本所属类别。它能很好地适应分类标准的变化,训练的时间复杂度比SVM算法低;

与朴素贝叶斯相比,它对数据没有假设且对异常点不敏感,本文将采用 K 近邻算法来做文本分类。

在文本表示中距离度量使用余弦相似度来代替,余弦相似度的值越大表示越相似。待分类文本 d_i 与训练集文本 d_j 的相似度 $Sim(d_i,d_j)$ 计算公式如下:

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

其中,w_{ik} 表示待分类文本 d_i 的特征向量,w_{jk} 表示训练集文本 d_j 的特征向量,n 表示训练集中所有特征项的个数。

选取距离待分类文本 d_i 最近的 K 个训练文本,即余弦相似度最大的 K 个样本。这里的 K 是经验值,从几十到几千不等,一般可以根据样本的分布,通过交叉验证选取一个合适的 K 值。通过统计这 K 个训练文本中属于类 c_i 的文本权重 $P(d,c_i)$ 的大小,进一步判断它们所属的预定义类别,计算公式如下:

$$P(d,c_i)=\sum_{d_j\in KNN(d)}Sim(d,d_j)f(d_j,c_i) \tag{2}$$

$$f(d_j,c_i)=\begin{cases}1, d_j\in c_i\\0, d_j\notin c_i\end{cases} \tag{3}$$

公式(3)为类别判别函数,如果 d_j 属于 c_i 类,则该项取1,否则取0。比较各类权重 $P(d,c_i)$,将待分类文本 d_i 分到权重最大的类别中。

3 传统的TF-IDF算法

3.1 特征项频率TF

TF(Term Frequency)是词语在文档中出现的词频。由于不同文档的长度不同,这些频次差距较大,因此需要将其规范化,从而使这些频次可以在同等的环境下进行对比。为了实现规范化,通常的做法是取频次和其所在文档中所有单词总数的比值。

3.2 逆文档频率IDF

特征赋权的方式较多,主要分为“均权”和“非均权”两类。“均权”认为特征项在整个训练集中的重要程度相同,它们不会对分类结果产生任何实质性的影响,所以给所有的特征项赋予相同的权重。而“非均权”认为特征项的重要程度不同,可以通过赋权处理提升主要特征项的作用,降低次要特征项的作用。目前的研究接受度较高的是“非均权”类方式,其中最具代表性的就是“IDF”权。

逆文档频率IDF(Inverse Document Frequency)表示给定单词的重要性,其主要思想是如果某个特征项在一个文本中出现频率很高,同时在其他文本中出现的频率低,说明此特征项具有很好的类别区分能力,应该给予较高的权重[14]。当需要计算词频时,先假定所有单词是同等重要的。这时,为了抗衡那些经常出现的单词的频率,需要用一个系数将其权重变小,逆文档频率IDF就是对文档总数目和该单词出现的文档数目的比值取对数。

3.3 特征项的长度信息

特征项长度可以作为衡量该特征项重要程度的一个关键因素。中文分词后的统计结果表明:出现频率最高的是单字词,出现频率较低的是多字词。事实表明单字词所能表达的信息很少;而多字词却可以表达大量信息,相对的它们的重要度也较高。一般来说,较长的特征项可以表示专指的概念,比如“九华山风景区”专指“旅游”,因此需要给这样的多字段词以更高的权重。

3.4 传统TF-IDF算法的不足

传统的TF-IDF算法主要考虑特征项的TF和IDF两个方面的信息,而TF和IDF都未考虑到新兴词汇的特殊性。新词的发现不仅可以丰富中文语料库、帮助解决一些中文分词过程中出现的歧义切分的问题,提高中文分词的准确度。而且,根据特定情境产生的新词往往能更准确地表达相关概念,提高以词语为特征项的向量空间模型的表达能力,从而进一步提高中文文本特征向量的质量[15]。因此本文针对现有的特征权重算法未考虑到文本表达方式变化所产生的新词的影响,提出了基于新词发现特征权重算法的改进策略。

4 改进的TF-IDF算法

4.1 TF-IDF算法公式

TF-IDF表示的是TF×IDF。其表达式为:

$$w_{dt}=tf_{dt}\times\lg(N/n_t) \tag{4}$$

其中,w_{dt} 为特征项 t 在文本 d 中的权重,tf_{dt} 为特征项 t 在文本 d 中出现的频率,N 为文本语料库中文本的总数,n_t 为文本语料库中包含特征项 t 的文本数。

TF指特征项在文档中出现的频次,在这种定义下会出现某个词语在长文本中出现的频次比短文本中出现频次高,和小数据集lg函数为零的情形。所以在实际处理过程中应避免TF对长文本的偏袒,通过取频次和文档中所有单词个数的比值对TF规范化,基于此将公式(4)改进为:

$$w_{dt}=(m_{dt}/M_t)\times\lg(N/n_t+0.01) \tag{5}$$

其中,m_{dt} 表示特征项 t 在文本 d 中出现的次数,M_t 表示文本 d 中的词语总数,其他参数与公式(4)中相同。

目前TF-IDF最常用的是一种名为TFC[16]的改进算法,该算法在避免对长文本偏袒的同时,将特征词的权重归一化。众多学者的研究中提到的TF-IDF算法都是指改良之后的计算公式,其中TFC的应用最为广泛,其表达式为:

$$w_{dt}=\frac{tf_{dt}\times\lg(N/n_t+0.01)}{\sqrt{\sum_{P=1}^{K}[tf_{dt}\times\lg(N/n_t+0.01)]^2}} \tag{6}$$

其中，K 为文本 d 中特征项的个数，其他参数与公式(4)中相同，本文基于新词的独特性在公式(6)的基础上提出改进。

4.2 新词集合的构建

新词的发现主要有两种方法：一种是使用统计方法通过对词共现的概率进行统计而得，另一种是基于规则的方法使用标注字典以及组词规则来识别新词[17]。本文综合两种方法，使用NLPIR PARSE中文分词工具完成新词发现与自适应分词功能。

在新词集合构建中需要考虑计算新词与类别标识名之间的相似度。词语相似度计算方法主要可分为两大类：一类是基于词典，多是利用词典中现有的层次关系来计算词语之间的相似度；另一类是基于大规模语料库，多是利用上下文的特征，用向量表示词语来进一步计算词语之间的相似度[18]。但是新词暂未加入词典，也不在已有的中文语料库中，因此无法使用现有的词语相似度计算方法计算新词与类别标识名之间的相似度。基于此，本文在新词的构建过程中，对整体实验语料进行处理，把所有新词存放在一个文档中，然后剔除与分类主题无关的新词，如：新闻来源、发表时间等。

4.3 特征权重算法的改进

特定情境下的新词往往更能表达出文本信息，对新词赋予更高的权重将会优化分类结果。加大网络新词的权重是特征提取的进一步工作，它是在特征提取之后，对信息含量大的特征项赋予更高的权重，目的是提高分类的准确率但是没有减少特征向量的维度。而主成分分析(PCA)法，是特征提取的经典方法之一，该方法将原始的 n 维特征映射到 k 维空间上 $(k<n)$，这 k 维特征是利用协方差矩阵对特征值分解，得到前 k 个大的特征值所对应的特征向量，是全新的正交特征，目的是使信息丢失量小的同时较好地表达了原始样本的信息并且减少了特征向量的维度。本文基于新词的重要性及特殊性在公式(6)的基础上进行改进，提出NewTFIDF公式，见公式(7)：

$$w'_{dt}=\frac{tf_{dt}\times\lg(N/n_t+0.01)+len(t)}{\sqrt{\sum_{P=1}^{K}[tf_{dt}\times\lg(N/n_t+0.01)]^2+len(t)}} \tag{7}$$

其中 $len(t)$ 指的是新词的长度，w'_{dt} 指采用NewTFIDF计算出的特征项 t 在文本 d 中的权重，其他参数与公式(4)中相同。将分子，分母同时加上 $len(t)$ 将会提高特征项 t 在文本 d 中的权重。

特征词集合(*Term*)，是新词集合(*NewTerm*)和普通特征词集合的并集。普通特征词和新词将会选用不同的权重计算方法，当特征项属于 *NewTerm* 时，使用公式(7)计算特征项的权重；否则，使用公式(6)计算特征项的权重，得到公式(8)：

$$w_{dt}(t\in Term)=\begin{cases}w'_{dt}, t\in NewTerm\\ w_{dt}, t\notin NewTerm\end{cases} \tag{8}$$

改进的特征权重算法对特征项的赋值过程如下：

(1)把文档 d 转化为词频向量。

(2)导入特征提取器，提取特征并构建特征词典TermDict。

(3)对于文档 d 中的特征项 t，判断它属于新词集合还是普通特征词集合。

(4)若 t 属于新词集合，使用公式(6)计算特征权重；否则使用公式(7)计算特征权重。

(5)根据(4)的结果，将词频向量转化为特征向量。

基于新词改进的TFIDF算法的文本分类流程图如图1。

5 实验分析

5.1 实验语料

本文使用网络爬虫从新浪网新闻中心上抓取新闻数据构成实验语料，该平台提供全面及时的新闻资讯，内容覆盖国内外突发新闻事件和重大新闻事件。实验语料包含4个话题，分别为“暴恐事件”，“踩踏事故”，“火灾”，“九华山”，共4 332篇新闻。网页数据是由HTML

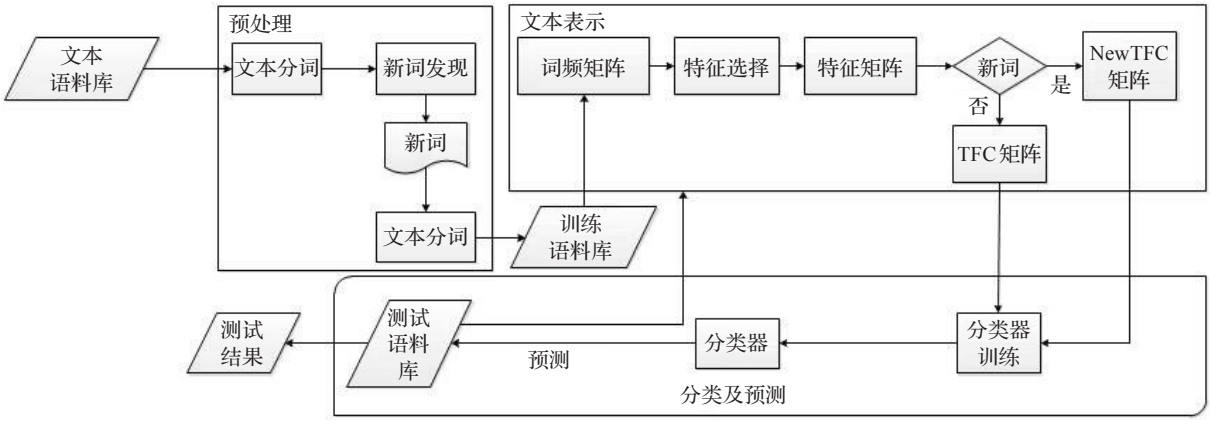

图1　基于新词改进的TFIDF算法的文本分类流程图

标签和文字段落组成的非结构数据,实验中对网页标记和特殊符号进行处理,提取出其中的文字信息,除去图集新闻、视频新闻和重复报道后,共得到4 000篇只包含搜索关键字的新闻片段文章作为本实验的数据集。实验采用随机取样的方式,75%的数据用于训练分类器,25%的数据用于测试分类结果(即训练集样本数量为3 000,测试集样本数量为1 000)。

5.2　实验流程

基于新词改进的TFIDF算法的文本分类流程:

(1)预处理

① 文本分词,去除停用词、无意义的虚词还有标点符号,如“的”,“了”,“吗”,“?”,“!”,“,”等。本文采用NLPIR PARSER中文分词工具,它分词速度快,准确度高,具有命名实体识别和关键词提取功能,最重要的是该工具具有新词发现功能。

② 采用基于信息交叉熵算法发现新词,经过筛选后,把新词加入用户词典。本文对整体实验语料进行处理,把所有新词存放在一个文档中,然后把诸如“央行网站”,“发布消息”,“中新网9月”,“新华社快讯”等与分类主题无关的新词剔除,这样便完成了新词集合的构建。

③ 导入用户词典重新分词。这时便可识别诸如“朝觐者踩踏事故”,“打击暴恐”,“四大佛教名山”,“外滩踩踏事件”等能看出包含文本信息的新词。

(2)文本表示

① 把文本转换为对应的词频矩阵。

② 使用特征提取方法提取重要特征来完成特征选择。

③ 定义TF-IDF变换器把词频矩阵转换为TFIDF矩阵,针对特征词的种类不同,使用不同的特征权重算法。

④ 完成文本的向量空间表达,得到其特征向量。

(3)分类及预测

① 针对实验语料,分析选择KNN算法最佳的 K 值和投票策略。

② 使用特征向量训练分类器。

③ 用训练所得的分类器预测测试语料所属的类别。

本实验采用KNN分类算法做了三次实验。第一次实验,用NLPIR PARSER中文分词工具对原始文档集进行分词处理,然后采用传统的TF-IDF算法为特征权重赋值;第二次实验仍然使用NLPIR PARSER中文分词工具对原始文档进行处理,不同的是在处理的过程中发现新词,然后将筛选后的新词加入用户词典中,导入用户词典重新对原始文档集进行分词处理,对特征权重赋值依然采用传统的TF-IDF算法;第三次实验在实验二的基础上,增加对特征词的词类判断步骤,对普通特征项权重赋值使用传统的TF-IDF算法,对新词使用改进的NewTFIDF算法赋权值。

实验过程中对 K 近邻分类模型中的最近邻的数量 K 值和投票策略weights这两个参数进行分析选择。图2为实验效果随 K 值及不同投票策略变化的情况。

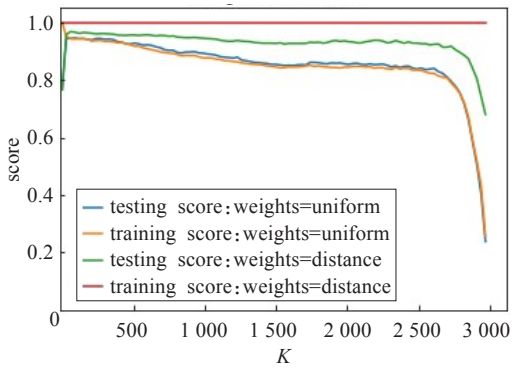

图2　实验效果随K值及不同投票策略变化的情况

从图2可以看出,使用uniform投票策略(即投票权重都相同)的情况下,分类器随着 K 的增长,预测性能先是缓慢下降,然而当 K 在2 700左右的位置,预测性能是急速下降。这是因为当 K 增大时,输入实例较远的训练实例也会起到预测作用,从而使预测发生错误。在使用distance投票策略下(即投票权重与距离成反比),分类器随着 K 的增长,对测试集的预测性能相对比较稳定。这是因为虽然 K 增大时,输入实例较远的训练实例对预测起作用,但因为距离较远,其影响会小很多。

当判断文章所属类别时, K 越大,计算代价越高。经过分析比较,本实验设定参数K=20,weights=distance。在参数设定的情况下,用KNN分类算法分别执行上述的三个实验过程。

5.3　评价标准

对于文本分类的性能评估测试,国际上通用的评价指标为精度(Precision)、召回率(Recall)和F1得分(F1 score)。

精度(又称查准率)是指被分类器正确分类的样本数量占分类器总分类样本数量的百分比,分类器在类 c_i 上的精度定义如下:

$$P_i=\frac{N_{cpi}}{N_{pi}} \tag{9}$$

其中, N_{cpi} 是分类器正确分类的文档数, N_{pi} 是分类器预测为 c_i 类的文档数。

召回率(又称查全率)是指应被正确分类的样本数量占某分类总样本数量的百分比,分类器在类 c_i 上的召回率定义如下:

$$R_i=\frac{N_{cpi}}{N_{ci}} \tag{10}$$

其中, N_{ci} 是实际属于 c_i 类的文档数。查准率和查全率是衡量分类效果的两个不同指标,它们是二律背反的,为使这两个指标相对均衡,引入 $F1$ 值。对类别 c_i ,其 $F1$ 值为:

表3　TFIDF和NewTFIDF的KNN评价指标对比　%

类别标号	TFIDF			TFIDF&NewTerm			NewTFIDF&NewTerm		
	P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
1	99.12	89.29	93.95	97.19	96.03	96.61	99.19	97.62	98.40
2	97.50	92.86	95.12	94.21	96.83	95.50	98.02	98.41	98.22
3	95.02	87.40	91.05	95.44	95.80	95.62	96.99	98.47	97.73
4	79.45	99.15	88.21	99.13	97.01	98.06	99.57	99.15	99.36
avg/total	93.04	92.00	92.14	96.43	96.40	96.41	98.41	98.40	98.40

$$F1=\frac{2R_iP_i}{R_i+P_i} \tag{11}$$

此外还有宏平均，微平均两种来计算 $P,R,F1$ 的方法。本实验记录KNN分类器的查准率、查全率、$F1$ 值和宏平均值，最终采用 $F1$ 值和宏平均值作为评价标准。

5.4　实验结果及分析

5.4.1　实验结果

根据实验结果随机给出8个新词，以及它们在测试集中的使用传统的TFIDF算法和基于新词改进的NewTFIDF算法的值以示参考，见表2。

表2　传统的TFIDF算法和基于新词改进的NewTFIDF算法对应的新词权重对比表

NewTerm	TFIDF	NewTFIDF
踩踏事故	0.161 09	0.400 79
暴恐分子	0.130 05	0.378 61
朝觐者踩踏事故	0.061 29	0.447 82
暴恐音视频	0.230 20	0.486 80
中国佛教四大名山	0.231 26	0.572 92
网上暴恐	0.151 43	0.393 87
风景名胜区	0.162 97	0.441 98
消防人员	0.084 52	0.346 09

将“暴恐事件”，“踩踏事故”，“火灾”，“九华山”分别记为“类别1”，“类别2”，“类别3”，“类别4”。表3和图3中TFIDF代表实验一的过程，即直接对文档集分词并采用传统的TF-IDF特征权重算法；TFIDF&NewTerm代表实验二的过程，即在分词过程中加入新词发现的过程并采用传统的TFIDF特征权重算法；NewTFIDF&NewTerm代表实验三的过程，即在分词过程中加入新词发现的过程，但是采用的是基于新词发现改进的TFIDF算法。为了避免实验的偶然性，每组实验独立重复进行5次，以其算数平均值为最终的性能指标，分别记录每组实验中这四个类别对应的 P 值，R 值，$F1$ 值和宏平均值（记为avg/total），见表3。图3在表3的基础上通过一个对比图来直观表示分类效果。

5.4.2　结果分析

（1）在实验中发现未加入新词发现过程时，特征项的总数量是10 654；而加入新词发现过程后，特征项的总数量是9 759，实验表明把新词发现加入文本预处理的过程中能减少特征词的数量，达到特征降维的目的。

（2）从表3中可得实验一 $F1$ 的宏平均值为92.14%，实验二 $F1$ 的宏平均值为96.41%，实验三 $F1$ 的宏平均值为98.40%，对比发现实验一的分类效果相对较差，实验三的分类效果相对较好。其中实验二的 $F1$ 值比起实验一的 $F1$ 值均有明显的提升，其中“暴恐事件”提升了2.66%，“踩踏事故”提升了0.38%，“火灾”提升了4.57%，“九华山”提升了9.85%；同时实验三的 $F1$ 值比实验二也有所提升，其中“暴恐事件”提升了1.79%，“踩踏事故”提升了2.72%，“火灾”提升了2.11%，“九华山”提升了1.30%。从图3中也可以明确看出TFIDF和NewTerm相结合的分类能力优于TFIDF，同时NewTFIDF和NewTerm相结合的分类能力又优于TFIDF和NewTerm相结合的。通过实验对比可得，发现新词并把它们加入分词过程中能使分类结果有所提升，而且在发现新词的同时给它们赋予更高的权重又能进一步优化分类结果。

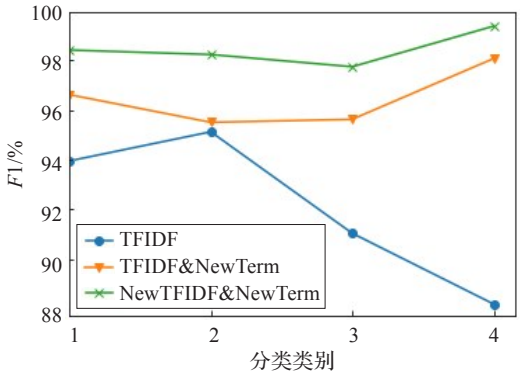

图3　实验结果对比图

6　结束语

本文通过对搜索关键字按相关度排序，抓取包含搜索关键字的新闻片段作为实验语料库。在文本预处理中识别新词并对其进行筛选，进而在文本表示中，针对新词和关键词采用不同的权重赋值公式。最后对KNN分类模型中的重要参数进行分析比较，选择恰当的参数去完成文本分类实验。实验结果表明，将新词发现加入文本预处理过程能降低特征空间的维度，同时基于新词发现改进的TF-IDF特征权重算法能有效提高分类器性能，优化分类结果。

参考文献：

[1] 郑霖，徐德华.基于改进TFIDF算法的文本分类研究[J].计算机与现代化，2014(9)：6-9.

（下转第161页）

scopic image quality assessment based on the binocular energy[J].Multidimensional Systems & Signal Processing, 2013,24(2):281-316.
[7] Shao F,Lin W,Gu S,et al.Perceptual full-reference quality assessment of stereoscopic images by considering binocular visual characteristics[J].IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society,2013,22(5):1940-1953.
[8] Zhang W,Ma L,Ma L,et al.Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network[J].Pattern Recognition,2016, 59(C):176-187.
[9] Shao F,Lin W,Wang S,et al.Learning receptive fields and quality lookups for blind quality assessment of stereoscopic images[J].IEEE Transactions on Cybernetics, 2016,46(3):730.
[10] Zhou W,Yu L,Qiu W,et al.Utilizing binocular vision to facilitate completely blind 3D image quality measurement[J].Signal Processing,2016,129:130-136.
[11] Ryu S,Sohn K.No-reference quality assessment for stereoscopic images based on binocular quality perception[J]. IEEE Transactions on Circuits & Systems for Video Technology,2014,24(4):591-602.
[12] Zhong X,Li C,Zhang W,et al.No-reference image quality assessment using dual-tree complex wavelet transform[C]//International Congress on Image and Signal Processing,2015:596-601.
[13] 田维军,邵枫,蒋刚毅,等.基于深度学习的无参考立体图像质量评价[J].计算机辅助设计与图形学学报,2016,28(6):968-975.
[14] Wang Z,Simoncelli E P,Bovik A C.Multiscale structural similarity for image quality assessment[C]//Conference Record of the Thirty-Seventh Asilomar Conference on Signals,Systems and Computers,2004:1398-1402.
[15] Kingsbury N.A dual-tree complex wavelet transform with improved orthogonality and symmetry properties[C]// International Conference on Image Processing,2000: 375-378.
[16] Roth J P,Scott Lovell † A,Mayer J M.Intrinsic barriers for electron and hydrogen atom transfer reactions of biomimetic iron complexes[J].Journal of the American Chemical Society,2000,122(23):5486-5498.
[17] 吴一全,尹丹艳,纪守新.基于双树复数小波和SVR的红外小目标检测[J].仪器仪表学报,2010,31(8):1834-1839.
[18] 周俊明,郁梅,蒋刚毅,等.利用奇异值分解法的立体图像客观质量评价模型[J].计算机辅助设计与图形学学报, 2011,23(5):870-877.
[19] Wang L,Duan F Q,Liang C.A global optimal algorithm for camera calibration with one-dimensional objects[C]// International Conference on Human-Computer Interaction.Berlin,Heidelberg:Springer,2011:660-669.
[20] Lasmar N E,Stitou Y,Berthoumieu Y.Multiscale skewed heavy tailed model for texture analysis[C]//IEEE International Conference on Image Processing,2009:2281-2284.
[21] Heidary K,Caulfield H J.Margin-setting with hyperellipsoidal surfaces[J].Optical Memory & Neural Networks, 2010,19(2):97-109.

(上接第109页)

[2] 史瑞芳.贝叶斯文本分类器的研究与改进[J].计算机工程与应用,2009,45(12):147-148.
[3] Ravitz J.An ISD model for building online communities: furthering the dialogue[J].Computer Mediated Communication,1997(1):12.
[4] 陈海红.多核SVM文本分类研究[J].软件,2015(5):7-10.
[5] 朱云霞.结合聚类思想神经网络文本分类技术研究[J].计算机应用研究,2012,29(1):155-157.
[6] 徐凤亚,罗振声.文本自动分类中特征权重算法的改进研究[J].计算机工程与应用,2005,41(1):181-184.
[7] Soucy P,Mineau G W.Beyond TFIDF weighting for text categorization in the vector space model[C]//International Joint Conference on Artificial Intelligence,2005:1130-1135.
[8] 熊忠阳,黎刚,陈小莉,等.文本分类中词语权重计算方法的改进与应用[J].计算机工程与应用,2008,44(5):187-189.
[9] 郭红钰.基于信息熵理论的特征权重算法研究[J].计算机工程与应用,2013,49(10):140-146.
[10] 刘哲,黄永峰,罗芳,等.网络新词识别算法研究[J].计算机工程与科学,2013,35(9):141-145.
[11] 奉国和,郑伟.文本分类特征降维研究综述[J].图书情报工作,2011,55(9):109-113.
[12] Cooper W S.Getting beyond Boole[J].Information Processing & Management,1988,24(3):243-248.
[13] Salton G,Wong A,Yang C S.A vector space model for automatic indexing[J].Communications of the ACM, 1995,18(11):613-620.
[14] 陈朔鹰,金镇晟.基于改进的TF-IDF算法的微博话题检测[J].科技导报,2016,34(2):282-286.
[15] 杜丽萍,李晓戈,于根,等.基于互信息改进算法的新词发现对中文分词系统改进[J].北京大学学报(自然科学版), 2016,52(1):35-40.
[16] Salton G,Buckley C.Term-weighting approaches in automatic text retrieval[J].Information Processing & Management,1988,24(5):513-523.
[17] 陈飞,刘奕群,魏超,等.基于条件随机场方法的开放领域新词发现[J].软件学报,2013(5):1051-1060.
[18] 石静,吴云芳,邱立坤,等.基于大规模语料库的汉语词义相似度计算方法[J].中文信息学报,2013,27(1):1-7.