文章编号:1005-0523(2021)01-0122-09 DOI:10.16749/j.cnki.jecjtu.2021.01.019

# 基于TF-IDF-MP算法的新闻关键词提取研究

曹义亲,盛武平,周会祥

(华东交通大学软件学院,江西 南昌 330013)

**摘要**:TF-IDF算法使用词频和逆文档频率来判断文章中词语的重要性,但类别区分效果不是很好。为提高分类效果,提出TF-IDF-MP算法。首先对语料库中的文档进行段落标注,利用jieba分词工具分词并标注词性,然后根据特征词在单个文档中出现的次数与该特征词在语料库所有文档中出现的平均次数进行比较,采用改进后的Sigmoid函数调整特征词权值,同时根据相关文档的段落位置重要程度赋予不同的位置权重,根据特征词权重大小排序后用朴素贝叶斯分类器对文档进行分类。实验结果表明,TF-IDF-MP算法应用到新闻分类中,精确率、召回率和*F*1值等评价指标较TF-IDF及相关改进算法都得到较好的提升。

**关键词**:文本分类;关键词提取;TF-IDF;词频均值化;位置加权

**中图分类号**:TP391 **文献标志码**:A

**本文引用格式**:曹义亲,盛武平,周会祥. 基于TF-IDF-MP算法的新闻关键词提取研究[J]. 华东交通大学学报,2021,38(1):122-130.

# Research on News Keyword Extraction Based on TF-IDF-MP Algorithm

Cao Yiqin, Sheng Wuping, Zhou Huixiang

(School of Software, East China Jiaotong University, Nanchang 330013, China)

**Abstract**: The TF-IDF algorithm uses the word frequency and inverse document frequency to judge the importance of words, but the category discrimination effect is not very good. In order to improve the classification effect, a TF-IDF-MP algorithm is proposed. First, the documents in the corpus were marked with paragraphs. The word segmentation tool jieba was used to label and tag the parts of speech. Then, the number of times a feature word in a single document was compared with the average number of occurrences in the document, and the feature word weights were adjusted by the improved Sigmoid function. At the same time, different position weights were given according to the importance of the paragraph position of the relevant document. According to the weight of the feature words, Naive Bayes classifier was used to classify the documents. The experimental results show that the TF-IDF-MP algorithm is applied to the news classification, and the evaluation indicators such as accuracy, recall and *F*1 value are better than TF-IDF and related improved algorithms.

**Key words**: text classification; keywords extraction; TF-IDF; word frequency averaging; position weighting

**Citation format**: CAO Y Q, SHENG W P, ZHOU H X. Research on news keyword extraction based on TF-IDF-MP algorithm[J]. Journal of East China Jiaotong University, 2021, 38(1): 122-130.

**收稿日期**:2020-12-07

**基金项目**:国家自然科学基金项目(61967006)

**作者简介**:曹义亲(1964—),男,教授,研究方向为图像处理与模式识别。E-mail:yqcao@ecjtu.edu.cn。

文档关键词体现了文档主题与内容,是理解文档内容的最小单位。文档关键词抽取,也称关键词提取或关键词标注,是从文本中把与该文本所表达的意义最相关的一些词或短语抽取出来,文档的自动关键词抽取是识别或标注文档中具有这种功能的代表性的词或短语的自动化技术[1]。在文本分类中,文档通常使用向量空间模型(vector space model, VSM)[2]表示,然后通过有监督的机器学习方法将待分类文本划分到预定义的类别中。根据 VSM 模型可知,每个文档都被表示为一个特征向量,由文本语料库中提取的许多术语(词或特征)的权重组成。因此,如何给特征词赋予合适的权重是文本分类任务中的一个基本问题,直接影响到分类的准确性。

在文本分类过程中, 特征提取是一个关键步骤。首先采用某个特征评估函数计算每个特征的数值,然后根据数值对特征排序,最后选取若干个数值最高的作为特征词。它的主要作用是在不丢失文本关键信息的前提下尽量减少待处理词语数量,以此来降低向量空间维数,从而简化计算,提高分类的速度与效率。常用的特征提取的方式有 4 种:① 采用映射方法将高维的特征向量变换为低维特征向量; ② 从原始特征中挑选出一些最具代表性、分类性能好的特征;③ 根据专家知识选择最具有影响力的特征;④ 采用数学方法找出最能体现分类信息的特征。

Uysal A K 提出了一种改进的全局特征选择方法, 对通用特征选择方法的最后一步进行了修改,使用局部特征选择方法根据特征对类的区分能力来标记特征, 并在生成特征集时使用这些标记[3]。2018 年,他在原先研究的基础上,从不同的角度对文本分类的两阶段特征选择方法进行广泛的分析,研究基于滤波的局部特征选择方法与特征变换相结合的特征选择方法。实验结果表明,采用主成分分析方法获得的准确率相比较其他方法更高[4]。Wan C 等提出了一种基于文本结构的复合特征提取算法,既可以用于测量文本相关性又可以增加复合特征的值,并采用支持向量机和朴素贝叶斯分类器在 3 个数据集上进行实验,验证了该方法的有效性[5]。Agnihotri D 等采用关联评分法,它结合单词之间的相互信息与强联系来对文本进行分类,在 Webkb, 20Newsgroup, Ohsumed10 和 Ohsumed23 4 个标准文本数据集上分别进行实验, 实验结果表明 Macro_F1 值取得了显著的提高[6]。Zhang L G 等人基于朴素贝叶斯文本分类器提出了两种自适应特征加权方法,实验结果表明,该特征加权算法有效地提升了分类的准确率,保持了最终模型的简单性并缩短了执行时间, 但对输入数据的表达形式很敏感,分类决策存在一定的错误率[7]。Haj-Yahia Z 等提出一种无监督的方法,通过结合通用和特定领域的人类专业知识和语言模型来丰富类别标签,文本分类实验效果要比采取简单的监督方法更好些,但 model 不同的 trick 在不同数据集表现有差异性,而且采用贝叶斯算法作比较,充分性不太够[8]。Habibi M 等解决了从会话中提取关键字的问题,并使用关键字为每个简短的会话检索少量可能相关的文档,从而达到文档推荐的目的[9]。Wu Q W 等提出了一种新的基于随机森林的集成方法 ForesTexter,包括特征子空间选择和分割准则,将要素分为两组,并为要素生成有效的术语权重,实验结果证明了提出的 ForesTexter 方法的有效性[10]。

词频-逆文档频率 (Term Frequency-Inverse Document Frequency,TF-IDF)算法是一种经典的特征权重算法,在一定程度上,这个算法可以较好的反映出某个特征词在文本分类过程中区分文本属性的重要程度,但是其理论依据存在一些不足[11]。为此, 国内外许多学者针对 TF-IDF 算法中存在的问题进行了改进,有效地提升了特征权重算法的准确性和效率。

罗燕等采用齐普夫定律结合特征词在文档中的词频,推导出同频词的计算公式并计算出各频次词语的比例,结合 TF-IDF 算法提取文档关键词[12]。牛永洁等综合考虑特征词的位置、词性、词语关联性、词长和词跨度等因素, 结合 TF-IDF 算法提取关键词[13]。Ghosh S 等基于 TF-IDF 算法提出一种受监督的功能构建方法,结合不同灾难场景下发布的信息对推文进行分类[14]。Chen K 等比较研究许多不同的术语加权方案,利用了跨不同类别文本的细粒度术语分布,提出了一种新的术语加权算法 TF-IGM[15]。

张瑾提出基于TF-IDF、词位置和词跨度的关键词自动提取的方法，加入位置权值及词跨度权值，在情报关键词提取中有广泛的应用价值[16]。高楠等提出了一种融合语义特征的短文本关键词提取方法，该方法从统计信息和语义层面分析了词语的重要性，并结合特征词的词频、长度、位置和语言等特征提取出最相关的关键词集合[17]。

虽然这些文献对关键词提取算法都进行了有效改进，但是都没有同时考虑文档中特征词的位置信息与主题的关联程度以及该算法在样本不均衡的数据集上的差异。本文在TF-IDF算法的基础上，结合特征词词频均值化与特征词位置信息对权重算法进行改进，提出了TF-IDF-MP(Term Frequency-Inverse Document Frequency-Mean term frequency and Position weighting)算法。采用Sigmoid函数对词频与词频均值的差进行处理，同时，根据相关文档中某些位置的关键段落赋予一定的权重调节因子，最后结合TF-IDF计算特征词的权值。实验也证明了TF-IDF-MP算法有效地提高了分类精确率、召回率和$F1$值等评价指标。

## 1 相关算法

TF-IDF的基本思想来自语言建模理论，常用于信息检索与文本分类，同时也是一种统计方法，用来判定单个字词对一个文档集合或一个文档的重要程度。一个字词在文档中出现的频率越高，则其重要程度应成正比例增加，但若出现在语料库其他文档中的频率也很高，则其重要程度应成反比例下降。

TF-IDF的主要思想是：如果一个字词在一篇文档中出现的次数很多，而在语料库其他文档中出现的次数很少，那么就可以认为该字词具有良好的分类效果，适合当作分类关键词。

1）词频。TF表示词频，即某个词出现在文档中的次数，为了减少文档词数差异对结果造成的误差，通过对词频进行归一化处理(即用词频除以文章总词数)，如下

$$tf_i=\frac{N_{i,d}}{\sum_{i}^{n}=1^{N_{i,d}}} \tag{1}$$

式中：$tf_i$表示词$i$归一化处理后的值；$N_{i,d}$表示词$i$出现在文档$d$中的总次数；分母表示文档$d$中全部词语的总个数。

2）逆文档频率。IDF表示逆文档频率，如果包含词$i$的文档在语料库中比较少，则表明词$i$在区分文档类别时可以起到良好的效果。计算一个词的$idf_i$，可使用语料库中文档总数量去除以所有包含该词的文档数量，然后对结果取对数。如下

$$idf_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|+1} \tag{2}$$

式中：$|D|$为语料库中的文件总数；$|\{j:t_i\in d_j\}|$为包含词的文档数目(即$N_{i,d}$的文件数目)，如果该词不在语料库中，就会导致分母为零，一般情况下分母为$|\{j:t_i\in d_j\}|+1$。

若一个词区分类别效果比较好，则这个词应该赋予较大权值，反之就赋予较小权值，一个词的$tf$-$idf$值就是

$$tf\text{-}idf=tf\times idf \tag{3}$$

3）朴素贝叶斯算法。朴素贝叶斯分类器是一种基于贝叶斯定理的简单概率分类器，对条件概率分布做了独立性假设，通过将条件概率彼此相乘来计算最大后验概率从而对文档进行分类[18]。朴素贝叶斯算法的流程可以描述如下：由多个特征词组成的文档$d$表示为式(4)，并根据贝叶斯规则其对应的类别标签为式(5)[19]

$$d=W_1,W_2,W_3,\cdots,W_n \tag{4}$$

$$label(d)=\arg_c\max(P(Y=c)\prod_{i=1}^{n}P(w_i|Y=c)) \tag{5}$$

在这种情况下，对于给定的类别$c$，$P(Y=c)$是类别$c$的概率，而$P(w_i|Y=c)$是特征词$w_i$的概率。多项式模型和多元伯努利模型在式(5)中的$P(w_i|\ Y=c)$的计算上有所不同。根据多项式和多元伯努利事件模型，概率计算分别为式(6)和式(7)

$$P(w_i|Y=c)=\frac{tf_{w_i,c}}{|c|} \tag{6}$$

$$P(w_i|Y=c)=\frac{df_{w_i,c}}{N_c} \tag{7}$$

式中：$tf_{w_i,c}$是类别$c$中$w_i$的词频；$|c|$是类别$c$中词频的总和；$df_{w_i,c}$是类别$c$中$w_i$的文档频率；$N_c$是类别$c$中的文档总数。如果文档$d$中不存在单词$w_i$，则概率公式对于特征词$w_i$变为式(8)

$$P(w_i|Y=c)=1-P(w_i|Y=c) \tag{8}$$

在本文中，将多元伯努利事件模型用于朴素贝叶斯分类。

## 2　TF-IDF-MP 算法

### 2.1　均值化词频

传统的 TF-IDF 算法根据特征词词频和特征词的逆文档频率的乘积来进行权重计算,简单的认为词频高的特征词应该赋予较高权值。但一些日常用词,如“的”、“虽然”、“一些”等,在文档中出现的次数比较多,对分类会产生负效果,赋值较大是不合理的。

首先根据特征词在单个文档中出现的次数与该特征词在语料库所有文档中出现的平均次数进行比较,若某个特征词在单个文档中出现的次数大于出现在语料库文档的平均次数,则说明该特征词对这个文档的重要程度要比其他文档更高,应该赋予更大权重,反之赋予较小权重;然后采用 Sigmoid 函数对两者的差值进行处理。

Sigmoid 函数的图像是一条单调递增平滑曲线,易于求导,值域在 0 和 1 之间,可以用来做二分类,在特征相差不是很大时结果比较好。Sigmoid 函数的公式如下

$$S_X=\frac{1}{1+e^{-X}} \tag{9}$$

图像如图 1 所示。

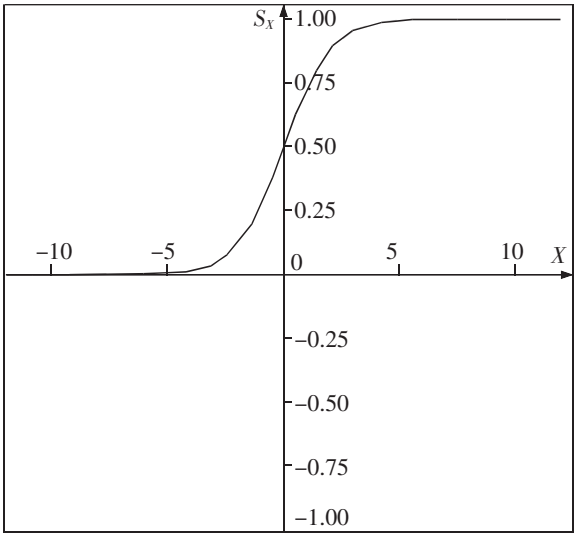


**图 1　Sigmoid 函数图像**
**Fig.1　Sigmoid function image**

从图 1 可以看出,当横坐标为 0 时,纵坐标为 0.5。在本算法中,若直接将 Sigmoid 函数中的 $X$ 替换为上述两者的差值,可发现当两者的差值相等时,即横坐标为 0,特征词词频缩小为原先的二分之一,特征词的权重也缩小为原先的二分之一,这是不合理的,根据前面的描述,此时该特征词对这个文档的重要程度应与其他文档一致。

本算法将 Sigmoid 函数进行了改进,修改后的公式如下

$$S_X=\frac{2}{1+e^{-X}} \tag{10}$$

式(10)的图像如图 2 所示。

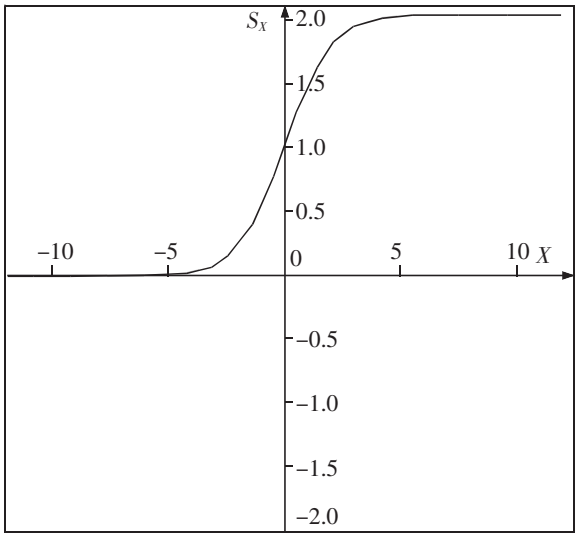


**图 2　式(10)对应的函数图像**
**Fig.2　Function image corresponding to Formula 10**

当上述两者的差值相等时,此时纵坐标的值为 1,表示特征词的权重与根据 TF-IDF 计算出的权重一致;当两者差值大于 0 或小于 0 时,此时纵坐标的值相应的大于 1 或小于 1,符合本实验的要求。若将 Sigmoid 函数的分子改为 3 或者更大,可发现特征词的权重被放大很多倍,实验误差较大。为此,本算法中将 Sigmoid 函数的分子改为 2,可以有效地缩小特征词词频之间差异,使得关键词提取算法更加准确。

为此,均值化词频(Mean Term Frequency,M)公式如下

$$M=\frac{2}{1+e^{-(N_{i,d}-\overline{N_i})}} \tag{11}$$

式中:$N_{i,d}$ 为在文档 $d$ 中特征词 $i$ 出现的次数;$\overline{N_i}$ 为特征词 $i$ 在语料库文档中平均出现的次数。

若特征词出现单个文档中的次数低于该特征词出现在语料库文档中的平均次数,那么 $M$ 值小

于1,则最终权重降低,反之则权重增加。通过对词频均值化处理，可以降低常用词在词频上造成的影响。

### 2.2 特征词位置加权

特征词位置信息的权重赋值法是将特征词在文档中的位置信息作为位置权重因子，并结合词频-逆文档频率计算特征词最后的权重。TF-IDF算法并未将特征词位置信息作为权重影响因素加入公式中计算，但事实上特征词在文档中位置的不同,对整个文档内容的重要性也有较大差异的。

在新闻网站中,基本上文章的主题都会在第一段和最后一段表现出来，所以从分类角度来看,文章的开始和结束部分一般都会出现关键词,比较重要，所以应该赋予这两部分的特征词更高的权值。为此,本文采用jieba分词并进行词性标注,将文章第一段和最后一段出现的名词的位置权重因子设为$P$,其余特征词位置权重因子为1,定义位置权重因子$P_i$如下

$$P_i=\begin{cases}P,\text{特征词非第一段或最后一段出现的名词}\\1,\text{特征词非第一段或最后一段出现的名词}\end{cases}\tag{12}$$

### 2.3 均值化词频-特征词位置加权

本文在TF-IDF算法的基础上，考虑文档中特征词的位置信息与主题的关联程度以及样本不均衡数据集上的差异,加入均值化词频和特征词位置信息等参数，最终计算特征词权重的TF-IDF-MP公式如下

$$W_{i,d}=tf\times idf\times M\times P\tag{13}$$

将式(1),式(2),式(11),式(12)代入式(13),得到

$$W_{i,d}=\frac{N_{i,d}}{\sum_{i}^{n}=1^{N_{i,d}}}\times\log\frac{|D|}{|\{j:t_i\in d_j\}|+1}\tag{14}$$

## 3 TF-IDF-MP算法在新闻分类中的应用

### 3.1 实验设计

实验步骤示意图如图3所示。

1）数据集选择。本实验采用的是搜狗新闻数据集，包含health,house,news,business等14个类别的新闻,不同类别的新闻数量差异较大,存在样本不均衡特性。数据格式如下：

<doc>

<url>http://news.sohu.com/20120612/n345428229.shtml</url>

<docno>c172394d49da2142-69713306c0bb3300</docno>

<contenttitle>公安机关销毁10余万非法枪支跨国武器走私渐起</contenttitle>

<content>中广网唐山6月1日涉黑、涉恶的团伙犯罪、毒品犯罪，还有从境外非法走私的枪支爆炸物.....</content>

</doc>

然后根据<url>标签中网址的二级域名进行分类,上述例子中的新闻类别为news类,根据这种方式提取所有文档新闻类别，并提取出相应的<content>标签中的新闻内容信息。分类后的文件列表如图4所示。

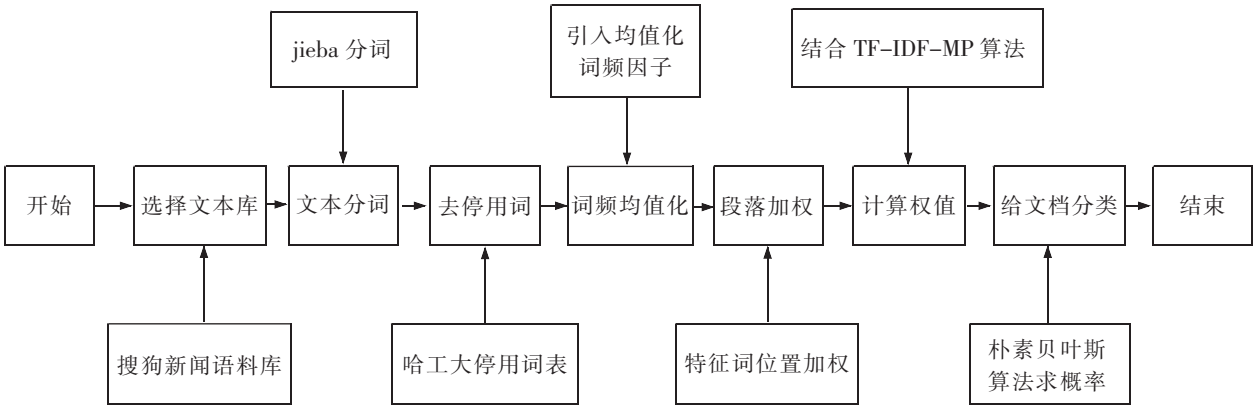


**图3 实验步骤示意图**

**Fig.3 Schematic diagram of experimental steps**

| 名称 | 修改日期 | 类型 | 大小 |
|---|---|---|---|
| auto.sohu.com.txt | 2020/7/21 8:44 | 文本文档 | 29,540 KB |
| business.sohu.com.txt | 2020/7/21 8:44 | 文本文档 | 243,413 KB |
| career.sohu.com.txt | 2020/7/21 8:44 | 文本文档 | 113 KB |
| cul.sohu.com.txt | 2020/7/21 8:44 | 文本文档 | 12,843 KB |
| health.sohu.com.txt | 2020/7/21 8:45 | 文本文档 | 21,122 KB |
| house.sohu.com.txt | 2020/7/21 8:46 | 文本文档 | 205,926 KB |
| it.sohu.com.txt | 2020/7/21 8:46 | 文本文档 | 39,996 KB |
| learning.sohu.com.txt | 2020/7/21 8:46 | 文本文档 | 30,264 KB |
| mil.news.sohu.com.txt | 2020/7/21 8:46 | 文本文档 | 11,708 KB |
| news.sohu.com.txt | 2020/7/21 8:45 | 文本文档 | 291,232 KB |
| sports.sohu.com.txt | 2020/7/21 8:45 | 文本文档 | 168,768 KB |
| travel.sohu.com.txt | 2020/7/21 8:45 | 文本文档 | 22,331 KB |
| women.sohu.com.txt | 2020/7/21 8:45 | 文本文档 | 40,165 KB |
| yule.sohu.com.txt | 2020/7/21 8:45 | 文本文档 | 73,285 KB |

**图 4　分类后的新闻文件列表**

**Fig.4　List of classified news files**

接下来选取每篇字数不低于 200 字的新闻文档,每个新闻类别选 400 篇,选 10 个类别一共 4 000 篇文章进行实验,其中选择 10 个类别文档各 300 篇共 3 000 篇为实验训练集，剩下的 1 000篇为实验测试集。

2）文本分词。采用 jieba 分词工具对每篇文档内容分词后再标注词性。

3）去停用词。使用哈工大停用词表对数据集中的文档去除停用词。

4）词频均值化。根据特征词在单个文档中出现的次数与该特征词在语料库文档中出现的平均次数进行比较，然后采用 Sigmoid 函数对特征词权重进行增加或者减少处理。

5）段落加权。在初始范围内分类的精确率随段落中名词位置权重因子的增加而提高,但当位置权重因子达到一定数值时,该名词对文章实际的作用效果被夸大,降低分类精确率,因此位置权重因子存在一个精确率峰值。为此,选取 100 篇新闻按照本文实验步骤进行实验,给文档第一段和最后一段出现的名词设置不同的权重因子 $P_i$,并使用精确率为评价指标寻求最合适的权重因子,计算不同 $P$ 值测试得到的精确率的平均值。实验中，权重因子 $P$ 在 1 到 2 之间递增选取,取 0.05 为步长,依次进行实验,将实验结果整理绘制成图 5。根据图 5 可知该数据集的 $P_i$ 最优取值为 1.2，因此将文档第一段和最后一段出现的名词的位置权重因子设为 1.2,其余特征词位置权重因子为 1。

6）计算权值。结合 TF-IDF-MP 算法计算权值并按照权值大小从大到小排序。

7）分类。选取每篇文档中权值最大的 5 个特征

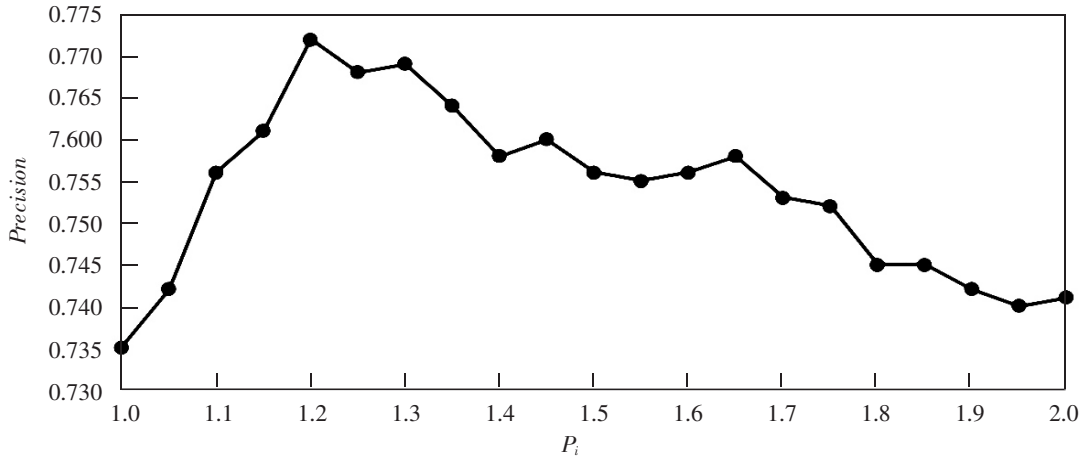


**图 5　不同 $P_i$ 值对 *Precision* 值的影响**

**Fig.5　The effect of different $P_i$ values on the *Precision* value**

词,将其权重值添加到朴素贝叶斯算法中,计算出每篇文档属于各分类的概率,选择分类概率中的最大值作为最终类别。

8)对比分析实验结果。

### 3.2 评价指标

为验证新算法的有效性,本实验选取health,house,news,business等10个类别不同的文档各100篇作为测试集,使用TF-IDF、文献[13]中算法、文献[16]中算法和本文算法进行对比实验。采用精确率、召回率和*F*1值来评价函数性能,其定义如下。

1)精确率(Precision)。表示分类结果全部预测为正的文档中正确的数量在总数的占比,计算公式如下

$$Precision = \frac{TP}{TP+FP} \tag{15}$$

2)召回率(Recall)。表示分类结果全部预测为正的文档中正确的数量占实际为正总数的比例,计算公式如下

$$Recall = \frac{TP}{TP+FN} \tag{16}$$

3)综合评价指标(F1)是精确率和召回率的调和均值,相当于精确率和召回率的综合评价指标,计算公式如下

$$F1 = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{17}$$

上述3个公式中,*TP*代表将实际为正类样本分类成正类样本的个数,*TN*代表将实际成负类样本分类成负类样本的个数,*FP*代表将实际为负类样本分类成正类样本的个数,*FN*代表将实际为正类样本分类成负类样本的个数。

### 3.3 实验结果和分析

通过精确率、召回率和*F*1值这3个评价指标对TF-IDF算法、文献[13]改进算法、文献[16]改进算法与本文改进算法对提取关键词进行比较分析,结果如表1所示。

**表1 TF-IDF算法、文献[13]算法、文献[16]算法和本文改进算法提取关键词的实验结果**

**Tab.1 TF-IDF algorithm, Ref[13] algorithm, Ref[16] algorithm and the experimental results of this algorithm**

%

| 评价指标<br>实验数据 | TF-IDF | | | 文献[13]算法 | | | 文献[16]算法 | | | 本文算法 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Precision* | *R* | *F*1 | *Precision* | *R* | *F*1 | *Precision* | *R* | *F*1 | *Precision* | *R* | *F*1 |
| business(50) | 81.42 | 79.64 | 80.31 | 83.14 | 81.36 | 80.01 | 86.24 | 81.07 | 80.75 | 87.21 | 82.81 | 80.90 |
| health(50) | 80.01 | 79.47 | 80.78 | 80.56 | 82.33 | 82.12 | 81.25 | 82.46 | 83. 09 | 81.23 | 82.57 | 83.06 |
| house(50) | 87.25 | 85.16 | 86.67 | 89.57 | 87.04 | 87.48 | 90.53 | 86.74 | 87.35 | 92.36 | 89.50 | 91.41 |
| news(50) | 83.42 | 84.06 | 81.27 | 87.46 | 86.01 | 85.79 | 88.12 | 85.09 | 85.47 | 90.85 | 87.43 | 88.27 |
| it(50) | 79.63 | 77.81 | 82.39 | 82.57 | 79.67 | 82.65 | 83.58 | 79.43 | 83.51 | 85.49 | 82.90 | 85.03 |
| learning(50) | 80.69 | 83.02 | 82.17 | 83.71 | 84.63 | 85.01 | 84.25 | 83.96 | 82.14 | 85.74 | 85.36 | 87.43 |
| sports(50) | 85.48 | 84.36 | 86.62 | 89.93 | 87.12 | 89.81 | 90.76 | 88.13 | 90.31 | 92.85 | 89.72 | 92.36 |
| yule(50) | 87.91 | 84.31 | 85.42 | 90.85 | 86.13 | 87.25 | 91.30 | 87.17 | 86.75 | 93.99 | 89.42 | 90.20 |
| travel(50) | 81.60 | 81.14 | 82.09 | 82.75 | 81.39 | 83.85 | 82.68 | 83.30 | 83.94 | 84.29 | 83.27 | 85.03 |
| mil(50) | 84.23 | 80.89 | 84.51 | 85.13 | 82.36 | 86.28 | 85.49 | 81.54 | 86.69 | 87.52 | 84.11 | 88.74 |

通过精确率、召回率和*F*1值这3个评价指标对TF-IDF算法、文献[13]改进算法、文献[16]改进算法与本文改进算法采用朴素贝叶斯算法分类后进行比较分析,结果如表2所示。

**表 2 TF-IDF 算法、文献[13]算法、文献[16]算法和本文改进算法采用朴素贝叶斯分类后的实验结果**

**Tab.2 TF-IDF algorithm, Ref [3] algorithm, Ref [16] algorithm and experimental results of the improved algorithm in this paper after using Naive Bayes classification**

%

| 评价指标<br>实验数据 | TF-IDF | | | 文献[13]算法 | | | 文献[16]算法 | | | 本文算法 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Precision* | *R* | *F*1 | *Precision* | *R* | *F*1 | *Precision* | *R* | *F*1 | *Precision* | *R* | *F*1 |
| business(50) | 82.31 | 80.63 | 76.95 | 85.92 | 82.25 | 80.13 | 87.31 | 81.49 | 81.32 | 88.36 | 83.60 | 81.21 |
| health(50) | 79.27 | 80.29 | 80.78 | 80.44 | 82.14 | 81.91 | 80.11 | 81.83 | 81.20 | 81.12 | 82.91 | 83.15 |
| house(50) | 86.59 | 84.57 | 86.28 | 89.23 | 86.34 | 87.07 | 88.36 | 86.10 | 86.51 | 92.47 | 89.43 | 91.34 |
| news(50) | 87.54 | 84.34 | 84.19 | 89.31 | 86.07 | 86.25 | 89.03 | 85.32 | 86.20 | 91.21 | 87.53 | 88.64 |
| it(50) | 78.26 | 76.37 | 81.13 | 82.34 | 79.59 | 82.37 | 83.10 | 78.62 | 82.13 | 83.32 | 81.18 | 83.78 |
| learning(50) | 80.43 | 82.10 | 81.57 | 83.48 | 84.31 | 84.26 | 84.36 | 83.37 | 82.36 | 85.17 | 84.94 | 86.68 |
| sports(50) | 85.69 | 84.13 | 87.47 | 89.34 | 86.43 | 90.23 | 90.37 | 87.24 | 89.74 | 92.17 | 88.38 | 91.74 |
| yule(50) | 87.74 | 85.47 | 85.41 | 91.10 | 86.23 | 87.33 | 92.14 | 87.35 | 88.16 | 94.30 | 89.38 | 90.11 |
| travel(50) | 81.25 | 80.31 | 82.04 | 82.73 | 81.40 | 83.46 | 82.31 | 83.04 | 83.72 | 83.74 | 82.43 | 84.16 |
| mil(50) | 83.41 | 80.52 | 84.17 | 85.66 | 82.47 | 86.62 | 84.76 | 81.49 | 86.57 | 86.93 | 83.33 | 88.14 |

注:表 1 和表 2 中,*R* 表示召回率。

通过表 1 可以发现，本文提出的 TF-IDF-MP 算法在提取关键词时要比 TF-IDF 算法、文献[13]中的算法和文献[16]中的算法性能更优,3 个评价指标都有了明显的提高，从而也验证了本文算法的合理性。

通过表 2 可以发现,采用朴素贝叶斯算法对提取的文档关键词进行分类后,精确率、召回率和 *F*1 评价指标值整体有一定提升。这是因为,本文的文档数量虽然比较多,但只是对每篇文档中 5 个权值较大的特征词进行分类,数据规模比较小,分类效率稳定,符合朴素贝叶斯的应用场景。

在文献[13]中,综合考虑了特征词的位置、词性、词语关联性、词长和词跨度等因素,但并没有考虑因词频差异带来的问题,没有去掉文档中的停用词,不同位置的权重设置也不太合理,一篇文章中首段和尾段的位置权重应该设为一致,而且最后的权重计算应该是各个影响因素相乘，而不是相加,权重相乘更能减少特征词权重的差异,提高实验准确率。在文献[16]中,综合考虑了位置权值及词跨度权值,但不同位置设置的权重值相差过大,也没有考虑特征词词频因素,容易增大实验误差。

TF-IDF-MP 算法结合特征词在语料库中词频的分布情况和在特征词文档中的位置信息,对那些在文档中出现高于特征词词频均值的特征词和更能体现文档主题的文档第一段以及最后一段的名词赋予较高的权重,而对那些低于特征词词频均值的特征词降低权重，使得 TF-IDF-MP 算法在提高关键词提取效果与文本分类方面起到了积极作用。

## 4　结论

1）TF-IDF-MP 算法在 TF-IDF 算法中加入均值化词频与特征词位置权重因子等参数来调节特征词权重以提取文档关键词。

2）新算法根据特征词在单个文档中出现的次数与该特征词在语料库所有文档中出现的平均次数进行比较，采用 Sigmoid 函数调整特征词权值大小,然后根据标注好词性的特征词,将文章第一段和最后一段出现的名词的位置权重因子设为 1.2,据此对 TF-IDF 算法进行改进。实验结果验证了本文提出的改进算法的合理性和可靠性，较相关算法,精确率、召回率和 *F*1 值均得到较好的提升。

3）该算法还有一些待进一步深入研究的问题。在设置特征词位置权重因子时，应该做进一步深入的研究分析，以期得到更合理更全面的权重因子，进一步提高实验结果的可靠性。在接下来的研究过程中，笔者将不断进行研究实验来寻找最适合本算法的权值因子，并结合特征词类内间分布和根据词语相似度合并同类词语来增加文本分类的精确率。

**参考文献：**


[1] 赵京胜，朱巧明，周国栋，等. 自动关键词抽取研究综述[J]. 软件学报，2017，28(9)：2431-2449.

[2] 叶雪梅，毛雪岷，夏锦春，等. 文本分类TF-IDF算法的改进研究[J].计算机工程与应用，2019，55(2)：104-109.

[3] UYSAL A K. An improved global feature selection scheme for text classification[J]. Expert Systems with Application，2016，43(1)：82-92.

[4] UYSAL A K. On two-stage feature selection methods for text classification[J]. IEEE Access，2018，6：43233-43251.

[5] WAN C，WANG Y，LIU Y，et al. Composite feature extraction and selection for text classification[J]. IEEE Access，2019，7：35208-35219.

[6] AGNIHOTRI D，VERMA K，TRIPATHI P. An automatic classification of text documents based on correlative association of words[J]. Journal of Intelligent Information Systems，2018，50(3)：549-572.

[7] ZHANG L G，JIANG L X，LI C Q，et al. Two feature weighting approaches for naive bayes text classifiers [J]. Knowledge-Based Systems，2016，100(3)：137-144.

[8] HAJ-YAHIA Z，SIEG A，DELERIS L A. Towards unsupervised text classification leveraging experts and word embeddings[C]//ACL，2019：371-379.

[9] HABIBI M，POPESCU-BELIS A.Keyword extraction and clustering for document recommendation in conversations[J]. IEEE/ACM Transactions on Audio Speech and Language Processing，2015，23(4)：746-759.

[10] WU Q W，YE Y M，ZHANG H J，et al. Fores Texter：An efficient random forest algorithm for imbalanced text categorization[J]. Knowledge-Based Systems. 2014，67(9)：105-116.

[11] 马慧芳，王双，李苗，等. 融合图结构与节点关联的关键词提取方法[J]. 中文信息学报，2019，33(9)：69-78.

[12] 罗燕，赵书良，李晓超，等. 基于词频统计的文本关键词提取方法[J]. 计算机应用，2016，36(3)：718-725.

[13] 牛永洁，田成龙. 融合多因素的TF-IDF关键词提取算法研究[J]. 计算机技术与发展，2019，29(7)：80-83.

[14] GHOSH S，DESARKAR M S. Class specific TF-IDF boosting for short-text classification：application to short-texts generated during disasters[C]//Companion of the The Web Conference，2018：1629-1637.

[15] CHEN K，ZHANG Z，LONG J，et al. Turning from TF-IDF to TF-IGM for term weighting in text classification[J]. Expert Systems with Application，2016，66：245-260.

[16] 张瑾. 基于改进TF-IDF算法的情报关键词提取方法[J]. 情报杂志，2014，33(4)：153-155.

[17] 高楠，李利娟，李伟，等. 融合语义特征的关键词提取方法[J]. 计算机科学，2020，47(3)：110-115.

[18] 贺鸣，孙建军，成颖. 基于朴素贝叶斯的文本分类研究综述[J]. 情报科学，2016，34(7)：147-154.

[19] 丁月，汪学明. 基于改进特征加权的朴素贝叶斯分类算法[J]. 计算机应用研究，2019，36(12)：3597-3600.