DOI:10.16652/j.issn.1004-373x.2020.12.020

一种改进TF-IDF的中文邮件识别算法研究

吴小晴，万国金，李程文，林梦思，曹书强
（南昌大学 信息工程学院，江西 南昌 330031）

摘 要：传统的TF-IDF算法没有很好地分配分词的权重，对一些能代表邮件类别出现频率较大的词语计算的IDF值反而较小，IDF值小说明单词的区分能力弱而不符合实际情况。为了提升垃圾邮件识别的准确率，提出一种改进TF-IDF算法和类中心向量的中文垃圾邮件识别方法。通过改进传统的TF-IDF计算方式，在传统的TF-IDF算法里面加入卡方统计量CHI和位置影响因子能够很好地改善一些重要词汇的权重问题，并结合逆向最大匹配算法的邮件文本分词和类中心向量算法的特征选择进行垃圾邮件分类。实验结果表明，所提算法相较于传统的TF-IDF算法对垃圾邮件识别的准确率提升了约3.6%，具有一定的实际应用价值。

关键词：TF-IDF算法；邮件识别；卡方统计量；权重分配；邮件分类；仿真分析
中图分类号：TN911.23-34；TP181　　**文献标识码**：A　　**文章编号**：1004-373X(2020)12-0083-04

Research on improved TF-IDF Chinese mail recognition algorithm

WU Xiaoqing, WAN Guojin, LI Chengwen, LIN Mengsi, CAO Shuqiang
(School of Information Engineering, Nanchang University, Nanchang 330031, China)

Abstract: A Chinese spam recognition method with improved TF-IDF algorithm and class centre vector is proposed to improve the accuracy of spam recognition. The traditional TF-IDF algorithm does not assign the weight of word segmentation well, and the calculated IDF value for some words that can represent the mail category and has higher frequency of occurrence is relatively small. The small IDF value indicates that the capacity of distinguishing the words is weak and does not accord with the actual demand. In this paper, the traditional TF-IDF calculation pattern is improved. The traditional TF-IDF algorithm adding the chi-square statistic CHI and position influence factor can improve the weight of some important words, and the spam classification can be performed by combining it with the feature selection of class center vector algorithm and mail text segmentation of the reverse maximum matching algorithm. The experimental results show that, in comparison with the traditional TF-IDF algorithm, this algorithm can increase the accuracy of spam identification by about 3.6%, which has a certain practical application value.

Keywords: TF-IDF algorithm; mail recognition; CHI; weight allocation; mail classification; simulation analysis

0 引 言

如今邮件成为日常沟通通信的主要方式之一，而垃圾邮件的存在给用户带来困扰。在卡巴斯基实验室发布的2018年第二季度垃圾邮件和钓鱼邮件的数据里，来自中国的垃圾邮件数目占邮件总量的14.36%，在统计的国家中国排名第一[1]。邮件分类识别的常用方法有基于IP地址和域名的黑白名单拦截方法、朴素贝叶斯算法[2]、决策树算法[3]、支持向量机算法[4]、K近邻算法分类[5]、类中心向量算法。现存在的技术在垃圾邮件识别准确率并不是很高，故本文算法在传统的TF-IDF算法[6]上改进，采用更精确的分词算法与邮件特征词向量的转化算法，在保证正确提取邮件内容的前提下提高垃圾邮件识别的准确率。

收稿日期：2019-10-24　　**修回日期**：2019-12-05
基金项目：国家自然科学基金项目(61661030)

1 邮件分类的常用方法介绍

1）黑白名单拦截方法[7]。现有一些组织和机构专门管理邮件黑名单，处理垃圾邮件地址的问题。若一个IP地址被列入黑名单，ISP服务商就会屏蔽该IP地址，用户则不会收到该地址发送的邮件。但是如果对方设置动态IP或者在不知道对方邮箱的前提下就没办法拦截邮件。

2）朴素贝叶斯分类。朴素贝叶斯算法在邮件过滤中广泛运用，朴素贝叶斯算法是一个典型的统计学习方

法，主要理论基础是一个贝叶斯公式，贝叶斯公式的基本定义如下：

$$P(Y_k|X)=\frac{P(XY_k)}{P(X)}=\frac{P(Y_k)P(X|Y_j)}{\sum_j P(Y_j)P(X|Y_j)} \tag{1}$$

式中：Y表示类别；X表示特征；$P(Y_k|X)$是在已知特征X的情况下求Y_k类别的概率，而对$P(Y_k|X)$的计算又全部转化到类别Y_k的特征分布上来。朴素贝叶斯算法的关键点是构造朴素贝叶斯分类器[8]，训练集邮件经过邮件预处理、特征提取、向量化处理等步骤构成分类器。

朴素贝叶斯算法根据概率统计来识别垃圾邮件但也存在弊端，假设属性之间是相互独立在一定程度上影响了分类的性能和准确性。

3）类中心向量算法。类中心向量算法是对邮件文本向量的处理，邮件文本在生成多维向量后，分别对训练集的正常邮件和垃圾邮件生成的向量求平均值，进行标准化处理，再计算同样操作的测试邮件向量分别与训练集邮件向量的相似度。设训练集的邮件为：

$$c=\{d_1,d_2,\cdots,d_k\}=\{<w_{1,1},w_{1,2},\cdots,w_{m,1}>\},\cdots,<w_{1,k},w_{2,k},\cdots,w_{m,k}>\} \tag{2}$$

式中，$w_{i,j}$表示特征词i在邮件d_k中的权重，则类别c_i的向量为：$v_{c_i}=\{<v_{w1,i},\cdots,v_{wm,i}>\}, v_{wj,i}=\frac{1}{k}\sum_{i=1,d\in c_i}^{k} w_{j,i}, j=1,2,\cdots,m$。

一个新邮件用改进的权重等一系列算法操作得到的向量，分别计算新邮件向量与正常邮件、垃圾邮件的相似程度。本文采用向量夹角的方式求近似度：

$$\cos(\boldsymbol{x},\boldsymbol{y})=\frac{\sum_{i=1}^{n}\boldsymbol{x}_i\times\boldsymbol{y}_i}{\sqrt{\sum_{i=1}^{n}\boldsymbol{x}_i^2}\times\sqrt{\sum_{i=1}^{n}\boldsymbol{y}_i^2}} \tag{3}$$

式中：$\boldsymbol{x},\boldsymbol{y}$为向量；$\boldsymbol{x}_i,\boldsymbol{y}_i$为向量分量。式(3)得到的值越大代表夹角越小，新邮件与此类别度越高，则新邮件是这个邮件类别的可能性越大。

2 TF-IDF算法介绍

2.1 TF的介绍

TF表示邮件中单词出现的次数。权重[9]是对一个单词重要程度的衡量，在一封邮件d_i中，单词t_i的所对应的权重为w_{ij}，TF的计算公式为：

$$\mathrm{TF}_i=\frac{n_{ij}}{\sum_k n_{kj}} \tag{4}$$

式中：n_{ij}表示词条t_i在邮件d_i出现的次数；$\sum_k n_{kj}$表示邮件d_i的词条总数。

2.2 IDF的介绍

IDF即逆向文件频率，IDF的计算公式为：

$$\mathrm{IDF}_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|} \tag{5}$$

式中：D为邮件总数；$|\{j:t_i\in d_j\}|$为包含词语t_i的邮件数目，即$n_{ij}\neq 0$的文件数目。如果该词语不在D中，就会导致被除数为0，因此一般情况下使用$|\{j:t_i\in d_j\}|+1$，即公式改为：

$$\mathrm{IDF}_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|+1} \tag{6}$$

TF-IDF的公式为：

$$\text{TF-IDF}=\mathrm{TF}_{ij}\cdot\mathrm{IDF}_i \tag{7}$$

为了避免TF-IDF的计算公式偏向于长邮件，则还要对TF-IDF进行归一化处理，归一化后的公式为：

$$\text{TF-IDF}=\frac{\mathrm{TF}_i\cdot\mathrm{IDF}_i\log\left(\frac{|N|}{|\{j:t_i\in d_j\}|+1}\right)}{\sqrt{\sum_{j=1}^{n}\left(\mathrm{TF}_i\cdot\log\left(\frac{|N|}{|\{j:t_i\in d_j\}|+1}\right)\right)^2}} \tag{8}$$

式中：N为邮件d_i中特征词的个数；$|\{j:t_i\in d_j\}|$为包含词语t_i的邮件数目。

传统的TF-IDF并不能很好地处理特征词的权重问题。传统的TF-IDF算法思想认为，如果垃圾邮件类别中包含词条t的邮件数为m，非垃圾邮件包含t的邮件数为k，则所有包含t的邮件数$n=m+k$，当m大的时候，n就会大，TF-IDF公式得到的IDF的值会小，说明单词t的区分能力不强。实际上，若某些词条多次出现在一个类别的邮件中，则该词条几乎能够代表这个类的邮件文本的特征，应给予较高的权重，并选来作为该类邮件的特征词。

3 改进后的TF-IDF与邮件分类的结合

3.1 改进的TF-IDF算法

利用改进后的TF-IDF和类中心向量方法对经过分词算法的邮件词组进行处理分类。一般情况下，邮件的主题和正文开头的第一句话对邮件的整体内容有很好的概括作用，因此邮件主题和正文第一句话的特征词应该赋予高一些的权重。在改进的TF-IDF算法中引入位置引用因子γ来表示词条权重值大小，邮件的主题部分一般会说明邮件的主旨与目的应赋予较大权重。本文规定邮件的主题部分位置引用因子$\gamma=5$，邮件正文第一句话的位置引用因子$\gamma=3$，引用卡方统计量[10]CHI来

表示特征词t与邮件类别的相关程度，CHI的计算公式为：

$$\mathrm{CHI}=\frac{N\cdot(XQ-YM)^2}{(X+M)\cdot(Y+Q)\cdot(X+Y)\cdot(M+Q)} \tag{9}$$

式中：N为训练集中邮件总数。CHI计算的值越大，说明特征词t与类别c之间的相关程度越高，可以代表类别c，所计算得到的权重值越大。包含特证词t，X,Y,M,Q的解释如表1所示。

表1　类别特征表

属于垃圾邮件的数量	属于正常邮件的数量
X	Y
M	Q

因此，对改进后的TF-IDF归一化公式为：

$$\mathrm{TF\text{-}IDF}(w_{ij})=\gamma(t_j)\cdot\log(\mathrm{CHI})\cdot\mathrm{TF}(t_j)\times\log\left(\frac{\mathrm{TF}_i(t_j)\cdot\mathrm{IDF}_i\cdot\log\left(\frac{|N|}{|\{j:t_i\in d_j\}|+1}\right)}{\sqrt{\sum_{j=1}^{n}\left(\mathrm{TF}_i(t_j)\cdot\log\left(\frac{|N|}{|\{j:t_i\in d_j\}|+1}\right)\right)^2}}\right) \tag{10}$$

式中：$\gamma(t_j)$为位置引用因子；N为邮件d_i中特征词的个数；$|\{j:t_i\in d_j\}|+1$为包含词语t_i的邮件数目；$\log(\mathrm{CHI})$为对数化处理后的CHI；TF_i为特征词t_i邮件d_i中的词频。改进TF-IDF算法的处理流程图如图1所示。

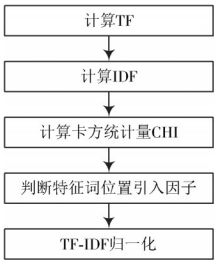

图1　改进TF-IDF算法的处理流程图

3.2　逆向最大匹配算法的介绍

最大匹配算法[11]原理是切分出单字，然后和词库进行比对，如果是一个词就记录下来，否则通过增加或者减少一个单字，再和词库进行比对，还剩下一个单字则终止。逆向最大匹配算法进行邮件文本分词的分词效率较高，图2为逆向最大匹配的算法流程。

3.3　邮件分类的步骤

使用的训练集为中文邮件，用逆向最大匹配对邮件进行预处理操作，预处理的分词经过改进后的TF-IDF算法属性加权与类中心向量算法结合组成分类器。图3为邮件识别分类步骤流程图。

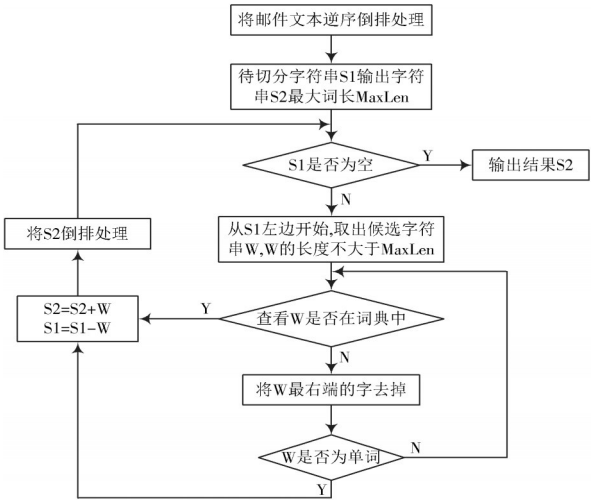

图2　逆向最大匹配算法流程

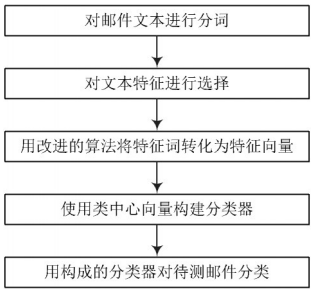

图3　邮件识别分类步骤

4　实验仿真结果

实验所用到的数据集来源于GitHub网站，使用2 000封邮件作为训练集提出特征词组成词典，用未经改进的TF-IDF类中心向量算法与改进的算法做对比，如表2所示，依次用100封、500封、1 000封、2 000封邮件做测试，仿真TF-IDF算法邮件分类的准确性。从图3的仿真结果可知，传统的算法平均准确率为82.55%，改进后的算法为准确率86.18%。因此，在其他条件相同时，本文改进的TF-IDF算法准确率更高，能够更好地应用于垃圾邮件分类上。

表2　测试邮件种类数量选取表

实验次数	正常邮件	垃圾邮件	邮件总数
1	50	50	100
2	200	300	500
3	500	500	1 000
4	400	1 600	2 000

图4为改进的TF-IDF算法与传统的TF-IDF在准确率上的对比。图5为朴素贝叶斯分类器在Python平台

下用SKlearn库朴素贝叶斯分类器与本文改进算法的测试结果对比，仿真结果显示改进的TF-IDF算法准确率要高于朴素贝叶斯分类器。

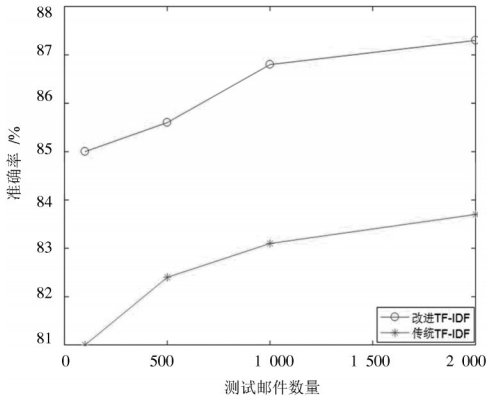

图4 准确率实验仿真结果对比图(一)

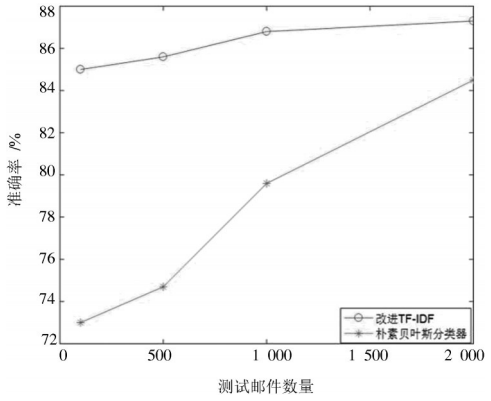

图5 准确率实验仿真结果对比图(二)

5 结 语

本文提出一种改进TF-IDF算法和类中心向量的中文垃圾邮件识别方法。改进的TF-IDF算法对邮件中不同位置的特征词计算出相对应的权值，邮件中的主题和邮件正文首句可以代表主体邮件信息，所以直接给出固定的权值，邮件的其他部分内容用卡方统计量改进传统TF-IDF的不足，提高垃圾邮件识别准确率。逆向最大匹配算法在分词效率与准确性上都高于正向最大匹配，类中心向量算法对邮件向量有着较高的处理效率。从实验结果可知，改进的TF-IDF算法提高了垃圾邮件的识别在准确率，而且测试文本集越大得到的准确率越高。

注：本文通讯作者为万国金。

参 考 文 献

[1] 王蕊.卡巴斯基发布第三季度垃圾邮件和钓鱼攻击报告[J].计算机与网络，2015，41(24)：45.

[2] 黄勇，罗文辉，张瑞舒.改进朴素贝叶斯算法在文本分类中的应用[J].科技创新与应用，2019(5)：24.

[3] 雷飞.基于神经网络和决策树的文本分类及其应用研究[D].成都：电子科技大学，2018.

[4] 郭太勇.一种基于改进的TF-IDF和支持向量机的中文文本分类研究[J].软件，2016，37(12)：141-145.

[5] 刘发升，董清龙，李文静.变精度粗糙集的加权KNN文本分类算法[J].计算机工程与设计，2019(5)：1339-1342.

[6] 叶雪梅，毛雪岷，夏锦春.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.

[7] 杨雷，曹翠玲，孙建国，等.改进的朴素贝叶斯算法在垃圾邮件过滤中的研究[J].通信学报，2017，38(4)：140-148.

[8] 高晓利，王维，赵火军.几种改进的朴素贝叶斯分类器模型[J].电子世界，2018(21)：40-41.

[9] 陈奕辰.基于句子权重和篇章结构的自动文摘算法的研究与实现[D].长沙：湖南师范大学，2015.

[10] 石俊涛.中文文本分类中卡方特征提取和对TF-IDF权重改进[D].成都：西华大学，2017.

[11] 杨贵军，徐雪，凤丽洲，等.基于最大匹配算法的似然导向中文分词方法[J].统计与信息论坛，2019，34(3)：18-23.

作者简介：吴小晴(1994—)，女，安徽人，硕士，主要研究方向为短波通信。
万国金(1955—)，男，江西人，教授，主要研究方向为信号处理、通信与通信对抗。
李程文(1995—)，男，江西人，硕士，主要研究方向为短波通信。
林梦思(1994—)，女， 江西人，硕士，主要研究方向为短波通信。
曹书强(1995—)，男，江西人，硕士，主要研究方向为短波通信。