网络出版时间：2016-08-15 16:36:08
网络出版地址：http://www.cnki.net/kcms/detail/51.1196.TP.20160815.1636.098.html

一种面向文本分类的特征向量优化方法 *

郭正斌，张仰森，蒋玉茹

(北京信息科技大学智能信息处理研究所，北京 100192)

摘 要：文本分类属于文本挖掘的一项研究内容，存在着广阔的应用前景，近年来得到了广泛的关注和研究。对文本进行建模的普遍方法是使用向量空间模型构建文本向量，并利用权值调整和维度调整对文本向量进行优化。提出了一种面向文本分类的特征向量优化方法。首先提出利用剔除近义词方法优化文本向量中的特征项。然后提出贡献率因子的概念，并利用其优化特征值。实验表明，比朴素贝叶斯分类方法的效果提高了 0.96%。因此，通过去除近义词和对提取出的特征词调整权重，可以达到优化特征向量、提高文本分类效果的目的。

关键词：机器学习；Mahout；特征向量；向量优化；文本分类

中图分类号：TP391

Feature vector optimization method for text classification

Guo Zhengbin, Zhang Yangsen, Jiang Yuru

(Institute of Intelligent Information Processing, Beijing Information Science & Technology University, Beijing 100192)

Abstract: Text categorization, as a research branch of text data mining, has broad application prospects and has been widely studied in recent years. It is a general method that using vector space model to construct a vector to represent text. There are two methods to optimize the text vector: adjust weights or adjust dimensions. This paper proposed a novel feature vector optimization method for text classification. First optimized the features in text vector by removing the synonyms. Second proposed a novel concept -- contributor factor to optimize the feature value. Result shows that the text classification accuracy of this work is increased by 0.96 percent compared with the Naive Bayesian method. Therefore, by removing synonyms and adjusting the weight of the feature words, we can achieve the goal of optimizing the text vector and improving the accuracy of text classification.

Key Words: machine learning; Mahout; feature vector;vector optimization; text clustering

0 引言

文本分类作为数据挖掘的一项研究内容，要从文本中获取有价值的信息来处理，其任务是把文本划分到与它最相似的一类。文本分类，早期使用的是词匹配法、知识工程等方法，这些方法存在用时长、效率低的缺点。但随着互联网海量文本的出现，统计和机器学习方法开始适用于这一领域，并逐渐成为主流[1,2]。现在的分类方法，通用的策略是首先对已分类好的数据进行训练，生成分类模型，然后使用模型对未分类文本进行自动分类。

目前，很多研究者采用向量空间模型对文本进行向量化表示后，采用距离计算的方法实现文本的分类，还有一些研究者采用条件概率的方法（如朴素贝叶斯方法）实现文本分类。本文将向量空间模型与概率模型相结合，首先通过向量空间模型对文本进行表示，采用 TF-IDF 计算特征词的权重，对权重进行归一化处理转换成概率后，再采用朴素贝叶斯的概率分类模型，实现文本的分类。在利用向量空间对文本表示过程中，可能会出现向量高维稀疏的问题以及近义词干扰的问题。向量高维稀疏[3]会导致文本分类模型训练的不充分，从而影响分类器的性能，而近义词则会降低特征词的辨别力，进而也将影响文本分类器的效果。本文拟从权值调整、降维两方面对向量空间进行优化，利用贡献率因子 α 和 β 调整权值和去除近义词以实现向量空间模型降维优化，以提升文本分类器的分类效果。

1 相关的理论及其原理

1.1 向量空间模型

向量空间模型（vector space model，VSM）是一种文本表示模型，由 Salton 等人于 20 世纪 70 年代提出[4,5]，最初是为了

基金项目：国家自然科学基金资助项目（61370139）；北京市教委科研计划面上项目（KM201411232014）；北京市属高等学校创新团队建设与教师职业发展计划项目（IDHT20130519）；北京信息科技大学促进高校内涵发展专项（5111623403）

作者简介：郭正斌（1993-），男，山东潍坊人，硕士，主要研究方向为自然语言处理（guozhengbin11@163.com）；张仰森（1962-），男，山西临猗人，教授，博士（后），主要研究方向为自然语言处理、人工智能；蒋玉茹（1978-），女，辽宁沈阳人，讲师，博士研究生，主要研究方向为自然语言处理.

应用于信息检索领域，后来被广泛应用于自然语言处理领域[6]。向量空间模型把文本映射成向量，特征词相当于维度，每个维度的权重可用数值的形式来表示。这样，一篇文档就可以映射成一个向量，文档之间语义的相似性就可以用向量之间的距离来度量。向量表现形式为：$d_j = (w_{1j}, w_{2j} \dots w_{ij})$ ，其中，d_j代表第 *j* 篇文档，向量分量w_{ij}表示第 *i* 个特征词（维度）t_i在文档d_j中的权重。

在中文信息处理方面，特征词通常是经过分词并且去除停用词后的结果，权值计算目前普遍使用 TF-IDF 权重计算方法[7,8]，如式（1）所示：

$$tfidf_{i,j} = tf_{i,j} * idf_i \tag{1}$$

其中，$tf_{i,j}$是特征词t_i在文档d_j中出现的频率，idf_i是特征词t_i的 idf（逆文档频率），idf 由总文档数目除以包含该特征词的文档数目，再将结果取对数计算得到[9]。

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2}$$

其中，$n_{i,j}$是特征词t_i在文档d_j中的出现次数，分母是在文档d_j中所有特征词的次数统计之和。

$$idf_i = \log \frac{|D|}{|\{j: t_i \in d_j\}|} \tag{3}$$

其中，$|D|$是语料库中的文档总和，$|\{j: t_i \in d_j\}|$是包含特征词t_i的文档数目。

1.2 文本分类

文本分类是机器对文本按照一定的分类体系自动标注类别的过程，属于数据挖掘的一项研究内容。随着统计和机器学习方法引入文本分类的领域，传统的方法被替代，分类的效果与效率得到明显提高。文本分类通用的步骤是：文本预处理→转换成为特征向量→训练特征向量生成分类模型（或者分类器）→用模型对文本进行分类并且评价分类效果[10]。因此，转换文本成为向量和训练分类模型，对于分类效果起着至关重要的作用。文本转换成向量，涉及特征词的选取及权重计算，特征词的选取有互信息、χ^2统计、信息增益等方法，权重计算有 TF、TF-IDF 等方法。目前有很多利用训练数据生成分类模型的算法，例如：朴素贝叶斯算法、KNN 算法[11]、支持向量机算法、决策树算法等。朴素贝叶斯算法由于简单、效率高[12,13]，被广泛使用。其两大基础是贝叶斯定理和特征项之间相互条件独立[14,15]。贝叶斯定理如式（4）所示：

$$P(b|a) = \frac{P(a|b) * P(b)}{P(a)} \tag{4}$$

其中，$P(b|a)$是在 a 发生的情况下 b 发生的可能性。

已知存在类别 $C_1, C_2 \dots C_m$，对 *n* 维文本向量 X={$x_1, x_2 \dots x_n$}分类。首先假设把向量 X 划分到类 $C_1, C_2 \dots C_m$ 中，计算每个类别对 X 的后验概率，然后把 X 分到后验概率最大的类。其后验概率公式[16] 如式（5）所示：

$$P(C_i \mid X) = \frac{P(X \mid C_i) * P(C_i)}{P(X)} \tag{5}$$

$P(X)$是一个常数，所以只需要$P(X \mid C_i) * P(C_i)$最大即可。$P(C_i)$是先验概率，它是每个类的样本数和所有类样本总数相除。$P(X \mid C_i)$是条件概率，根据特征项之间相互条件独立，计算公式如式（6）所示：

$$P(X|C_i) = P(X_1|C_i) * P(X_2|C_i) \cdots * P(X_n|C_i) = \prod_1^n P(X_k|C_i) \tag{6}$$

$P(X_k|C_i)$也容易求得，至此，$P(C_i \mid X)$ 计算结束。

朴素贝叶斯分类算法的思想基础是这样的：在统计的基础上，对于待分类项，求其在各个类别出现的概率，最后划分待分类项到概率最大的类别。其算法流程如下：

Step 1：用 TF-IDF 方法构建特征向量，把数据集用向量表征出来，生成文本向量。

Step 2：把文本向量分离数据，生成训练集和测试集。

Step 3：计算训练集中类的先验概率和特征向量对应每一类的条件概率。

Step 4：计算测试集中待分类文本在每一类的后验概率，取最大值决定应该分入的类。

Step 5：与已知的文本分类标签比较，得到准确率、F1 值等评价结果。

Step 6：结束。

1.3 向量优化技术

一个中等规模的文本语料库通常具有几万个词，但其中的一篇文本，词语数目大约一千左右，因此向量空间模型带来的一个严重问题就是高维稀疏[17]。高维稀疏提高了向量处理的时间复杂度，也降低了文本分类的性能。除此之外，文本数据也有近义词这一语言现象。近义词指的是使用了不同的文字表达相同的内容，它丰富了表达形式，但这会对文本分类带来干扰，影响了分类的性能。例如，“中国”一词，可能有以下表达方式：中华人民共和国、内地、华夏等，它们共同分担了“中国”的权重，这会降低特征词的作用。因此，需要通过优化文本向量的表示来提高文本分类的性能。

向量优化包括权值调整、维度调整两个方面。权值调整，是通过调整特征词的权重，使其能够更加准确地反映特征词对文本的重要性。维度调整包括维度升高和维度降低，维度升高应用于短文本（例如微博等），维度降低应用于中长文本（例如新闻等）。本文研究的是新闻，属于中长文本，因此主要介绍维度降低（降维）。降维就是要降低向量空间的维度，使其能优化文本向量的表示。降维有特征选择和特征提取[18]两种方式，特征选择指从特征集中选取一部分特征词来表示整个特征空间；特征提取则是利用矩阵变化的方法，把 *m* 个特征变成 *n* 个特征（*n*<*m*），将原始特征空间映射成一个维度更低的特征空间。

2 面向文本分类的特征向量优化

2.1 调整权重

本文通过实验研究发现，经过 TF-IDF 计算后的文本向量中，特征项权值高的，并不一定对文本的正确分类有贡献。例

如，下面是一篇编号为 C000008_1532 的财经类新闻[1]（C000008 表示财经类，1532 表示文档的 id，去除停用词后共计 74 个词语），按照 TF-IDF 权重排序的前 10 个特征词及其权重如表 1 所示：

表 1 排名前 10 特征词、权重及其排名

特征词	TF-IDF 权值	TF-IDF 排名
板块	17.994	1
行情	13.557	2
正在	11.749	3
论坛	10.991	4
涨停	10.633	5
有色金属	10.331	6
大涨	9.69	7
航天	9.282	8
跌停板	9.183	9
军工	8.901	10

把这前 10 个特征词分成三组，第一组：行情、涨停、大涨、跌停板，第二组：板块、有色金属、航天、军工，第三组：正在、论坛。可以看到，第一组的特征词明显与财经类相关，第二组的特征词可能与财经类相关，也可能与其他类相关，第三组特征词明显与财经类不相关。第二、三组的特征词的 TF-IDF 权值很高，但它们对于分类作用很小，尤其是第三组，从人辅助分类的角度出发，不能根据这些词把新闻分到财经类。

那么哪些词才对文本分类有贡献，怎样进行特征向量优化呢？

本文使用 TF-IDF 排序后的特征词调整权值，其核心思想为：

a) 对每个类的训练数据计算 TF-IDF，然后对 TF-IDF 排序。

b)取每个类 TF-IDF 排序后特征词权值最高的前 K 项（下面把每个类权值最高的前 K 项特征词作为一个集合，称为 Top-K）。

c)使用 Top-K，调节特征词权重值，达到优化特征向量、提高分类效果的目的。

下面来讨论当把文本分类为一个类别和多个类别时 Top-K 的作用。

当把文本分类为一个类别时，分类的目的是筛选出符合这个类别特征的文本，很容易根据 Top-K 作出选择。例如，还是财经新闻 C000008_1532，剩余的部分特征词（如表 2 所示），这些词属于财经类的 Top-K（*K* 取 5000 时）中，会对新闻分入财经类有贡献。尽管经过计算 TF-IDF 后，这些特征词的权重以及排名并不高。

表 2 部分特征词、权值及其权值排名

特征词	TF-IDF 权值	TF-IDF 排名
涨停板	7.104	23
盘面	6.293	32
纵深	6.249	33
ST	6.239	34
停牌	6.228	35
投资者	5.877	40
复牌	5.699	42
个股	5.244	46
下跌	5.126	49
大盘	4.997	51

所以，本文引入单贡献率因子α（$\alpha > 1$），用 α 乘以计算后的 TF-IDF 的权值，来提高对分类有贡献的特征词的权重，如式（7）所示：

$$w\alpha = w1 * \alpha \tag{7}$$

其中，$w\alpha$为优化后的权重值，$w1$为优化前的权重值，α为单贡献率因子。

定义 1 负作用特征词：负作用特征词为对分类有干扰并且出现在多个类别的 Top-K 中的特征词。

当把文本分类为多个类别时，每个类别的 Top-K 并不是完全不一样，有一部分是相同的，并且有的特征词在多个类别的 Top-K 中都会出现。例如，财经新闻 C000008_1532 中，板块、有色金属、航天、军工、投资者等词也会出现在其他类别的 Top-K 中。那么对于这些在多个类别 Top-K 中都出现的特征词，本文引入多贡献率因子β（$\beta < 1$），用β乘以计算后的 TF-IDF 的权值，来调整有负作用特征词的权值，降低这些词对分类的贡献率，如式（8）所示。

$$w\beta = w1 * \beta^m \tag{8}$$

其中，$w\beta$为优化后的权重值，*w1* 为优化前的权重值，β为多贡献率因子，*m* 表示特征词在 *m* 个类别的 Top-K 中出现。

2.2 特征选择——去除近义词

本文通过去除近义词，构建一个特征词更少的子集实现特征选择。

为了识别文本中的近义词，本文使用了《同义词词林（扩展版）》[2]作为语义词典。此词典共分为五层结构，每层结构赋予一个语义编码，五层结构构成的完整编码表示一个原子词群。原子词群使用标记符号＝、＃、@分别表示同义词、相关词、独立词，其中，标记为“＝”的原子词群，表示该原子词群里面的所有词都是同义词。例如，该词典的一个原子词群如下表 3 所示，它表示词群编码 Di02A03，其词群内容“中原”、“华夏”、“中华”等词都是近义词。

表 3 Di02A03 原子词群

词群编码	标记符号	词群内容
Di02A03	=	中原 华夏 中华 华 赤县 神州 九州 赤县神州 炎黄 中国 礼仪之邦

本文需要获取近义词，所以只选取标记为“＝”的原子词群。去除近义词的方法是：对文本信息处理时，把每一个原子

[1] 该新闻来自搜狗实验室提供的文本分类语料库（http://www.sogou.com/labs/dl/c.html）

[2]《哈工大信息检索研究室同义词词林扩展版》

词群的第一个词当作该词群的代表词，将文本中的近义词全部替换成代表词。

3　实验及结果分析

本文使用的数据集——搜狗新闻分类语料（共 9 个类，每个类 1990 个文件），是一个已经分类好的新闻数据集，可以从搜狗实验室网站[3]下载。进行朴素贝叶斯分类时，本文把语料随机分为训练语料（80%）和测试语料（20%）。

本文使用了 Mahout 机器学习工具[4]，特征权值计算采用的是 TF-IDF，分类算法采用的是朴素贝叶斯（Naive Bayes）分类算法。

本实验获取 Top-K 采取的实验方法：

a)分别对每个类的新闻文本（训练数据）计算 TF-IDF；

b)对每个类的 TF-IDF 排序；

c)得到每个类的特征词权值最高的前 K 项（Top-K）。

本实验分类阶段采取的实验方法：

a)计算 TF-IDF, 利用 Top-K，训练单贡献率因子 α 和多贡献率因子 β，重新生成文本向量；

b)根据设定的比例分割数据，随机生成训练数据(80%)和测试数据(20%)；

c)对训练数据进行训练，生成分类模型；

d) 使用分类模型测试数据，并且评价分类效果；

e)重复步骤 a)~d)，共 100 次，取 100 次结果的平均值作为最终分类结果。

本文调整权重的目的是得到最优的参数，为了避免出现过拟合现象，采用了随机子抽样法[19]。具体来讲，将原始数据随机生成训练数据和测试数据，然后将这个过程重复 F 次，总的评价结果取 F 次实验结果的平均值。

目前分类的评价标准有准确率（accuracy）、召回率（recall）、F 值（F-Score）等等，本文中实验采用 F1 值作为评价标准。

本文实验的系统流程图如图 1 所示。

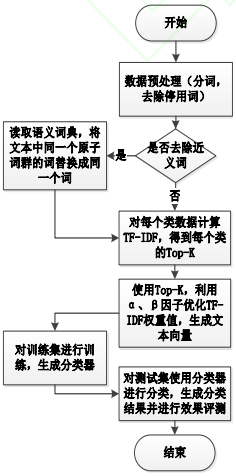

图 1　程序处理流程图

[3] 文本分类语料库　（http://www.sogou.com/labs/dl/c.html）

[4] Mahout　（https://mahout.apache.org/）

3.1　Top-K 中 K 对分类结果的影响

本文首先测试聚类结果 Top-K 中 *K* 取不同值对分类结果的影响，*K* 从 0-15000，评价标准取 100 次实验数据 F1 值的平均值。*K*=0 时的分类结果作为基准值，因为此时 *K*=0，没有需要优化的特征词，未使用α、β来调整特征词的权重，原始 TF-IDF 的计算结果没有任何改变的保留，实验结果表明基准分类的 F1 值为 0.880151。因为语料库的大小限制，本文实验时 K 的最大值只取到了 15000。

图 2 为当$\alpha = 1.5$，$\beta = 0.9$时贝叶斯分类结果，分类结果 F1 值随着选取的 *n* 变化曲线（图 2 只标注了部分的数据，详细数据如表 4 所示）：

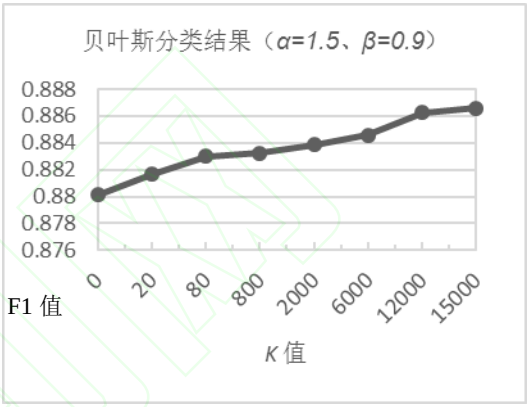

图 2　F1 值随着 *K* 变化的图

表 4　*K* 取值结果集

K 值	F1 值	*K* 值	F1 值	*K* 值	F1 值
0	0.880151	800	0.883225	8000	0.884845
10	0.880711	1000	0.883645	9000	0.88555
20	0.881652	2000	0.883869	10000	0.885206
40	0.882203	3000	0.883892	11000	0.885996
80	0.882984	4000	0.883549	12000	0.886253
100	0.883036	5000	0.88379	13000	0.886483
200	0.883098	6000	0.884593	14000	0.886492
400	0.883301	7000	0.884269	15000	0.886607

结果分析：实验表明，随着K的逐渐增大，尽管个别数据点有波动，但 F1 值总体趋势是上升的。当$K = 15000$时，实验效果最佳，F1 值为 0.886607，可以看到实验结果比最初的基准值（K=0 时，F1 值 0.880151）提高了 0.6456%，这证明提出的方法对提高朴素贝叶斯分类是有效的。接着本文探索α、β对分类结果的影响，*K* 则取分类结果最好时的最大值 15000 不变。

3.2　多贡献率因子 β 对分类结果的影响

探索多贡献率因子β对分类结果的影响，采取的方法是当$\alpha = 1.5$ 不变时，β从 0.95-0.86 递减变化，共 10 个测试数据点，每个β值都实验 100 次，取其平均值。图 3 为当$\alpha = 1.5$ 时，分类结果 F1 值随着β变化曲线：

结果分析：实验表明，随着β的由大到小，可以看到实验结果有一定的波动，这说明β值在 0.95-086 区间时，对分类影响

较小。而且当n=15000、β=0.9 时，实验结果加权 F1 值 0.886607 最高。

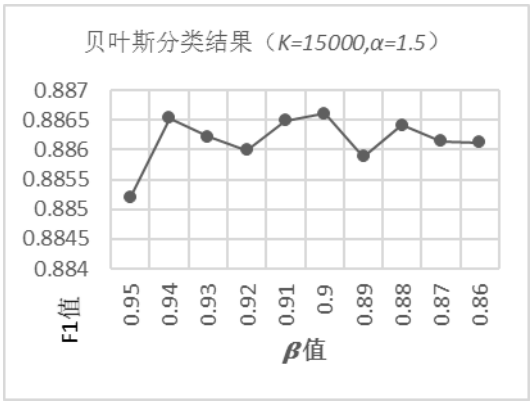

图 3 F1 值随着β变化图

3.3　单贡献率因子 α 对分类结果的影响

探索单贡献率因子α对分类结果的影响，采取的方法是取$\beta = 0.9$(由 4.2 实验可知，β取 0.9 时效果最佳)不变，α从 1.1-2.0 递增变化，共 10 个测试数据点，每个α值都实验 100 次，取其平均值。图 4 为当$\beta = 0.9$时，分类结果 F1 值随着α变化曲线：

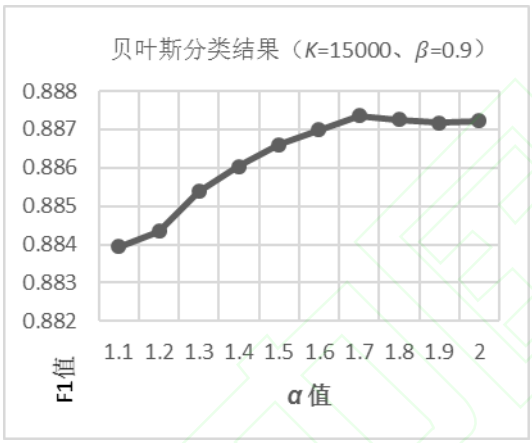

图 4　F1 值随着α变化图

结果分析：实验表明，随着α的由 1.1 逐渐增加到 1.7，可以看到 F1 值曲线逐渐上升，而从 1.7 到 2.0，F1 值的曲线逐渐下降，类似抛物线的曲线。由此表明，当α=1.7 时，实验结果取得最大值 0.887354，实验结果比最初的基准值（K=0 时，F1 值 0.880151）提高了 0.7203%。

3.4　去除近义词并且使用 α 、β 对分类结果的影响

3.4.1 去除近义词

在文本预处理阶段，直接合并近义词，使其用一个词表示。例如原子词群 Di02A03，中原、华夏、中华、华、赤县、神州等词统一用中原表示。合并近义词后，计算 TF-IDF 并且使用朴素贝叶斯算法分类。实验表明，去除近义词后的分类效果得到提升，实验结果 F1 值为 0.883158，比基准值提高了 0.3007%，这证明去除近义词对特征选择起到了不错的效果。

3.4.2 去除近义词并且使用 α 、β

本文通过结合去除近义词并且使用α、β，希望达到最佳分类效果，理论上 F1 值应该提升 0.7203%+0.3007%=1.021%。在去除近义词的基础上，使用$\alpha = 1.7$，$\beta = 0.9$ 优化权重，实验结果 F1 值为 0.889217，提高了 0.9066%，这说明还有提升空间。重新训练α、β，实验 100 次取其平均值，实验结果如表 5 所示：

表 5 综合方法实验结果

α	β				
	0.88	0.89	0.90	0.91	0.92
1.62	0.888262	0.888251	0.888286	0.888289	0.888392
1.63	0.888986	0.888820	0.888544	0.888710	0.890156
1.64	0.888944	0.889023	0.889294	0.888843	0.888986
1.65	0.889166	0.889377	0.889758	0.889467	0.889361
1.66	0.889156	0.889349	0.889475	0.889071	0.889134
1.67	0.889287	0.889315	0.889525	0.889286	0.889021
1.68	0.889554	0.889118	0.889265	0.889071	0.889436
1.69	0.889211	0.889358	0.889236	0.889088	0.889173
1.70	0.888930	0.888892	0.889217	0.888910	0.888822
1.71	0.888748	0.888729	0.889016	0.889092	0.889268
1.72	0.888970	0.888912	0.888858	0.888860	0.889166

结果分析：实验表明，随着α的由小变大，可以看到 F1 值大体趋势是先增大后减小；随着β的由小变大，可以看到 F1 值没有明确的变化规律。当$\alpha = 1.65$，$\beta = 0.9$时，实验结果取得最大值 0.889758，实验结果比最初的基准值（K=0 时，F1 值 0.880151）提高了 0.9607%。实验结果没有达到最大期望值，因为是多次随机分割数据然后取其平均值，所以实验结果有所波动。

从实验结果发现，单贡献率因子α的优化作用明显，而多贡献率因子β的优化作用微弱。本文β值的目的是降低出现在多个类别 Top-K 的特征词的作用，这与 TF-IDF 方法的作用在一定程度是相同的。因为，对于出现在多个 Top-K 的特征词，说明其在多个类别经常出现，这可能导致了特征词在所有文本中的高文件频率，因此，这些特征词的 idf、TF-IDF 值也是比较小的。所以，β值降低权值的优化作用不明显，对分类的影响较小。

通过上述实验，采用本文提出的优化方法，当$\alpha = 1.65$，$\beta = 0.9$时，实验结果取得 F1 值最大值为 0.889758，分类效果最好。

4　结束语

本文采用调整权值和特征选择优化特征向量，实现了对新闻文本分类过程的优化。调整权重方面，本文首先对每个类计算 TF-IDF 并排序，求得每个类的 Top-K。然后根据这 Top-K 中的特征词,对计算后的 TF-IDF，使用单贡献率因子α提高权重值和多贡献率因子β降低权重值进行权重优化。特征选择方面，使用去除近义词的方法。本文使用朴素贝叶斯分类算法检验优化效果。实验证明，本文提出的实验方法取得了不错的结果，比基准的分类结果提高了 0.96%。

本文下一步研究计划初步包括两个方面：调整权重方面，优化多贡献率因子；特征选择方面，使用语义相似性去除近义

词。

本文只是简单的使用了多贡献率因子β，对于出现在多个类（>=2）的 Top-K,进行了降低权重处理。但实际上，对于出现在多少个类中的 Top-K 才应该进行降低权重，这需要进一步的实验来做改进。下一步研究计划之一就是确定这个类值，来获取最佳分类结果。

除此之外，考虑语义相似性词语对特征向量的优化。例如，使用训练好的 word2vec 获取“中国”语义相似的词语，可以发现其中包括“我国”、“国内”、“内地”等词。尽管这些词不是“中国”的近义词，但是在上下文环境中，它们通常代指“中国”。如果能识别语义相似性词语，也会对优化特征向量起作用。

参考文献

[1] 刘赫, 刘大有, 裴志利, 等. 一种基于特征重要度的文本分类特征加权方法[J]. 计算机研究与发展, 2009, 46(10): 1693-1703.

[2] 张玉芳, 万斌候, 熊忠阳. 文本分类中的特征降维方法研究[J]. 计算机应用研究, 2012, 29(7): 2541-2543.

[3] Ljpvd M, Eo P, Hjvd H . Dimensionality reduction : a comparative review[J]. Journal of Machine Learning Research, 2007, 10(1): 1-35.

[4] Fu Ruiji , Qin Bing , Liu Ting. Open-categorical text classification based on multi-LDA models[J]. Soft Computing, 2015, 19(1): 29-38.

[5] 王子慕. 一种利用 TF_IDF 方法结合词汇语义信息的文本相似度量方法研究[D]. 吉林: 吉林大学, 2015.

[6] 姚清耘, 刘功申, 李翔. 基于向量空间模型的文本聚类算法[J]. 计算机工程, 2008, 34 (18): 39-41.

[7] 施聪莺, 徐朝军, 杨晓江. TFIDF 算法研究综述[J]. 计算机应用, 2009, 29(S1): 167-170.

[8] 陈治纲, 何丕廉, 孙越恒. 基于向量空间模型的文本分类系统的研究与实现[J]. 中文信息学报, 2008, 19(1): 36-41.

[9] 张俊丽. 文本分类中的关键技术研究[D]. 武汉: 华中师范大学, 2008.

[10] 庞观松, 蒋盛益. 文本自动分类技术研究综述[J]. 情报理论与实践, 2012, 35(2): 123-128.

[11] 耿丽娟, 李星毅. 用于大数据分类的 KNN 算法研究[J]. 计算机应用研究, 2014, 31(5): 1342-1344.

[12] Jiang L, Li C, Wang S, *et al*. Deep feature weighting for naive Bayes and its application to text classification[J]. Engineering Applications of Artificial Intelligence, 2016, 52: 26-39.

[13] Taheri S, Yearwood J, Mammadov M, *et al*. Attribute weighted naive Bayes classifier using a local optimization[J]. Neural Computing and Applications, 2014, 24(5): 995-1002.

[14] 曹洋, 成颖, 裴雷. 基于机器学习的自动文摘研究综[J]. 图书情报工作, 2014, 58(18): 122-130.

[15] 陈祎荻, 秦玉平. 基于机器学习的文本分类方法综述[J]. 渤海大学学报: 自然科学版, 2010, 31(2): 201-205.

[16] 马宾, 殷立峰. 一种基于 Hadoop 平台的并行朴素贝叶斯网络舆情快速分类算法[J]. 现代图书情报技术, 2015, 255(2): 78-83.

[17] 刘海峰, 王元元, 张学仁, 等. 文本分类中基于位置和类别信息的一种特征降维方法[J]. 计算机应用研究, 2008, 25(8): 2292-2294.

[18] 盛秋艳, 何文广. 一种改进的向量空间降维方法[J]. 黑龙江工程学院学报, 2011, 25(1): 60-62.

[19] 武亚昆, 段富, 尹雪梅. 分类器准确率评估的研究[J]. 电脑开发与应用, 2011, 24(4): 10-12.