一种基于 TF-IDF 的朴素贝叶斯算法改进

许甜华,吴明礼
(北方工业大学 信息学院,北京 100144)

摘 要:目前对以朴素贝叶斯算法为代表的文本分类算法,普遍存在特征权重一致,考虑指标单一等问题。为了解决这个问题,提出了一种基于 TF-IDF 的朴素贝叶斯改进算法 TF-IDF-DL 朴素贝叶斯算法。该算法以 TF-IDF 为基础,引入去中心化词频因子和特征词位置因子以加强特征权重的准确性。为了验证该算法的效果,采用了搜狗实验室的搜狗新闻数据集进行实验,实验结果表明,在朴素贝叶斯分类算法中引入 TF-IDF-DL 算法,能够使该算法在进行文本分类中的准确率、召回率和 F_1 值都有较好的表现,相比国内同类研究 TF-IDF-dist 贝叶斯方案,分类准确率提高 8.6%,召回率提高 11.7%,F_1 值提高 7.4%。因此该算法能较好地提高分类性能,并且对不易区分的类别也能在一定程度上达到良好的分类效果。

关键词:朴素贝叶斯;TF-IDF 算法;去中心化;位置信息;特征权重
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2020)02-0075-05
doi:10.3969/j.issn.1673-629X.2020.02.016

An Improved Naive Bayes Algorithm Based on TF-IDF

XU Tian-hua, WU Ming-li
(School of Informatics, North China University of Technology, Beijing 100144, China)

Abstract: At present, the text classification algorithm represented by the naive Bayes algorithm generally has the same feature weights and single index. In order to solve this problem, we propose an improved TF-IDF-based naive Bayes algorithm, TF-IDF-DL naive Bayes algorithm. Based on TF-IDF, this algorithm introduces decentralized word frequency factor and feature word position factor to enhance the accuracy of feature weights. In order to verify its effect, we use Sogou's Sogou news dataset to conduct experiments. The experiment shows that the TF-IDF-DL algorithm is introduced into the naive Bayesian classification algorithm, which can make the algorithm perform well in the accuracy, recall and F_1 value in text classification. Compared with the domestic similar research TF-IDF-dist Bayesian scheme, the classification accuracy rate is increased by 8.6%, the recall rate is increased by 11.7%, and the F_1 value is increased to 7.4%, so the proposed algorithm can improve the classification performance better and achieve a great classification effect to some extent for the indistinguishable categories.

Key words: naive Bayes; TF-IDF algorithm; decentralization; location information; feature weight

0 引 言

随着信息技术的发展,网络信息量急剧增加,其中文本信息是海量网络数据中的一大主体,但海量文本数据混乱存储,极大影响了信息获取的效率。如何快速准确地获取自己想要的信息便成为了一个重要问题。而现今广泛应用的分类技术可以帮助人们快速地筛选信息,并且从海量数据中提取信息进而构造高效的分类器,是数据挖掘领域中一个热门的研究方向。其中文本分类的过程一般分为以下步骤:数据预处理、数据特征提取、构建分类器、进行分类。

现今数据挖掘领域有多种分类算法,比如决策树、支持向量机、贝叶斯分类器和神经网络等。其中贝叶斯分类器通过某对象的先验概率,利用贝叶斯公式计算出其后验概率,然后选择多种分类中的最大后验概率作为该对象所属分类的分类器。其计算过程简单快速,在多分类问题上计算复杂度比较均衡,且在多分布独立的假设下,分类器效果很好,所需样本少。贝叶斯分类器以其上述优点在文本分类、垃圾文本过滤、情感判别、多分类实时预测、推荐系统等领域中被广泛应用。在贝叶斯分类器中,朴素贝叶斯分类假定各个特征相互独立,互不干扰,能够处理多分类任务,适合增量式训练,尤其在数据量超过一定程度时,可以进行批

收稿日期:2019-03-25 **修回日期**:2019-07-26 **网络出版时间**:2019-11-07
基金项目:国家自然科学基金(61672040)
作者简介:许甜华(1994-),女,研究生,研究方向为数据处理技术与软件服务;吴明礼,博士,讲师,研究方向为商业智能、数据仓库、文本挖掘。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20191107.0918.068.html

次训练,所以在垃圾邮件过滤,文档分类中效果很好。

但是朴素贝叶斯网络在进行特征计算以及分类的过程中,默认所有特征的权重是一致的,这样的前提忽略了各文本特征的特性。而实际上,不同特征项在分类过程中起到的作用是不一样的,将特征的权重视为一致,会在一定程度上降低分类的准确率。比如:当一篇文章中多次出现了“雾霾”一词,便可认为文章主题和天气相关的概率是很大的,而当文章中只提到一次“雾霾”时,几乎是不能确定该文章主题和天气相关的。因此在使用朴素贝叶斯网络时,多与其他的特征加权算法共同使用,进行特征加权计算,以得到更好的分类效果。目前文本分类中常用的特征权重算法 TF-IDF(term frequency - inverse document frequency)是一种基于词频的特征权重算法[1],通过计算词频和逆文本频率来计算特征权重,在兼顾效率的同时也能得到较满意的效果。但是该算法没有体现特征词在文档类间和类内的分布信息。文献[2]中加入特征类间比重信息,使其对文档分布不敏感,从而对文档集有更好的适应性;文献[3]通过计算特征词间的相似度,选择最大相似度作为特征权重,提高分类效果;文献[4]提出新词发现特征权重算法,改进 TF-IDF 对网络新词的识别能力,优化文本分类效果;文献[5]通过改进特征选择算法和特征加权算法,增加位置选择信息来提高文本分类效果;文献[6-9]均对 TF-IDF 权重进行了类间改进优化。

虽然这些文献对权重进行了改进,但均未兼顾文档词频的分布位置和算法在正负样本不均衡的倾斜数据集上的不同。鉴于传统 TF-IDF 算法的不足,文中提出一种基于 TF-IDF 的朴素贝叶斯改进算法 TF-IDF-DL 朴素贝叶斯算法。相对于以上各种改进方法,文中拟打算从特征词词频及其位置与类别之间的关系出发,对词频进行去中心化处理并引入特征词位置影响因子,以达到分类算法对不同的文档有更强的分类适应性,并能够在分类结果的准确率、召回率和 F_1 值方面有所提高的目的。

1 相关研究

1.1 朴素贝叶斯算法

朴素贝叶斯算法假设各条件特征相互独立[10],计算文本中某些特征出现的情况下,该文本属于某分类的概率,最后通过对比各个分类概率的大小,找出最高概率值,从而得出当前文本所属分类。朴素贝叶斯的分类公式为:

$$C = \operatorname{argmax} P(C_n) \prod_m^j P(X_m \mid C_n) \tag{1}$$

其中,$P(C_n)$ 代表所要分类的文本属于类别 C_n 的概率,$P(X_m \mid C_n)$ 代表类别 C_n 中包含特征项 X_m 的概率。在朴素贝叶斯中,要求各特征独立,且将特征权重看作是一致的。但在实际应用中,各特征的权重是不一致的,为了让算法更加准确,使用特征加权算法进行特征权重的计算,从而提高分类性能[11]。

1.2 特征项频率 TF

TF(term frequency)是特征词在文档中出现的词频,其表达式为:

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2}$$

其中,分子 $n_{i,j}$ 表示该词在文件中的出现次数,分母为文件中所有字词出现的次数总和。但是由于文档长度不一,为了防止同一词语在较长文档中出现的频率比在较短文档中出现的频率高的现象,一般会对词频进行正规化处理的改进。

1.3 逆向文件频率 IDF

词频计算,在传统计算中是将所有特征词的权重看作是相同的。而特征词的权重在实际应用中并不一致,所以在文本分类中要提升主要特征项的作用,降低次要特征项的作用。

IDF(inverse document frequency)可以计算出给定词的重要性。某一特定词的 IDF,由文件总数目除以包含该词语文件的数目进行表示[6]。如果一个特征项在一个文本中出现的频率较高,而在其他文本中出现的频率较低,那么说明此特征项能够很好地区分此类文本和其他文本。公式如下:

$$\mathrm{IDF}_i = \lg \frac{N}{n_i} \tag{3}$$

但在计算过程中,会出现某一词并未在某一文本中出现的情况。为了防止出现这种分母为零的现象,最常用的方法是使用拉普拉斯平滑对上述公式进行处理,进行平滑处理后的公式为:

$$\mathrm{IDF}_i = \lg \frac{N}{n_i + 1} \tag{4}$$

最后,TF-IDF 传统计算公式为 TF * IDF,即:

$$w_{dt} = \mathrm{tf}_{dt} * \lg\left(\frac{N}{n_t + 1}\right) \tag{5}$$

其中,w_{dt} 为计算出的特征项 t 在文本 d 中的权重,tf_{dt} 为特征词在文本 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。

2 TF-IDF 的改进

2.1 去中心化词频因子

在 TF-IDF 的计算过程中,将特征词词频作为特征词权重大小的判断依据,以特征词出现的次数,以及

特征词文档比例来进行权重计算。但是各个特征词表达的意义并不相同,某些特征词出现频率较少,属于日常用词,对于文本分类的贡献并不大,但是在权重计算中被赋予较高的权重;某些特征词属于生僻词,能够代表某一类文本,出现次数较少,但是在权重计算中被赋予较低的权重。

针对以上不足,文中采用去中心化特征词频因子对特征词出现的次数进行去中心化处理。在计算特征词频时,根据特征词出现的相对次数对权重进行增加或者减少的处理,在这两个方面进行改进后对结果再进行正值化处理,最终去中心化特征词频因子(decentralization term frequency)公式如下:

$$\mathrm{DTF}_{d,t} = \mathrm{e}^{N_{d,t}-\overline{N_t}} \tag{6}$$

其中,$N_{d,t}$ 为特征词 t 在文档 d 中出现的次数,$\overline{N_t}$ 为特征词 t 在各文档中出现的平均次数。

将 DTF 添加到 TF-IDF 中,即分子变为:

$$w_{dt} = \mathrm{tf}_{dt} * \lg\left(\frac{N}{n_t + 1}\right) * \mathrm{e}^{N_{d,t}-\overline{N_t}} \tag{7}$$

若一个词在此文档中出现的频率低于该特征词出现的平均频率,那么 DTF 值小于 1,则最终权重降低;反之则权重增加。通过去中心化处理,可以降低常用词和生僻词在词频上的差异性。

2.2　特征词位置信息

在文档中,大多数文章都会在开始和结束包含文章的主题,所以从分类角度来看,文章的开始和结束部分的信息较为重要,应该给予更高的权重[12],所以文中将特征词所在位置增加为权重计算的一个因子[13]。

将文档中所有特征词第一次出现的位置排列成一个序列,以文章总词数为总长度,以 1 为单元刻度,取序列最中间的位置为原始坐标,计算其他词距离原始坐标的距离,距离越远,给予权重越大,说明该词对分类的影响越大。定义位置影响因子(location factor)如下:

$$\mathrm{LF}(d,t) = \begin{cases} \varepsilon, \mid p(x) - \bar{p} \mid \geqslant \delta \\ 1, \mid p(x) - \bar{p} \mid < \delta \end{cases} \tag{8}$$

其中,ε 为要增加的权重值倍数,δ 的范围在(0, $D/2$)之间,其中 D 为序列总长度。

将去中心化词频因子和特征词位置信息加入到传统的 TF-IDF 公式中,最终改进的 TF-IDF 公式(TF-IDF-DL)如下:

$$w_{dt} = \mathrm{TF} * \mathrm{IDF} * \mathrm{DIF} * \mathrm{LF} \tag{9}$$

最后将该公式与朴素贝叶斯算法相结合[14],改进后的朴素贝叶斯公式为:

$$C = \arg\max P(C_n) \prod_m^j P(x_m \mid C_n) * wnm \tag{10}$$

3　实验与分析

3.1　数据处理

该实验采用搜狗实验室的搜狗新闻精简数据集(SogouCS, 2012 版 http://www.sogou.com/labs/resource/cs.php),共 698 M,128 个新闻文档,完整新闻条数共 429 818 条,数据样式如下所示:

<doc>

<url>http://sports.sohu.com/20080612/n257439913.shtml?</url>

<docno>805e04a983c29427-71013306c0bb3300</docno>

<contenttitle>图文:凯尔特人备战总决赛　加内特与里弗斯</contenttitle>

<content>来源:搜狐体育▮搜狐体育讯▮北京时间 6 月 12 日消息 NBA 总决赛的第四战马上就要在洛杉矶打响了,比赛之前的头一天双方进行了适应场地的训练,并接受了媒体的采访。(责任编辑:张正)</content>

</doc>

从上述样式的<url>标签得出此条信息的新闻类别为 sports 类,以此方式进行所有文档新闻类别的提取,并提取对应的<content>标签中的新闻内容信息。

同时,还需对得到的数据集进行进一步的处理。首先,将常用的停用词(的,并不,而且等)进行过滤,其次将新闻内容短于 50 字符的新闻视为垃圾新闻并进行剔除。最终数据集将分为 12 类,该实验选择其中 5 类进行分析,分别为:women, entertainment, travel, health, sports。为保证数据均匀分布,各类新闻各取 5 000条作为训练集,取 1 000 条作为测试集,如表 1 所示。

表 1　数据集

Category	Number of train dataset	Number of test dataset
women	5 000	1 000
entertainment	5 000	1 000
health	5 000	1 000
travel	5 000	1 000
sports	5 000	1 000

3.2　实验步骤

文中分别采用传统的 TF-IDF 算法、文献[2]中的 TF-IDF-dist 算法以及 TF-IDF-DL 算法进行特征权重计算并将其应用于朴素贝叶斯分类器中进行文本分类,对比实验结果并进行分析,具体实验步骤如下:

(1)输入文档转化为特征词后的词频向量;

(2)进行文本的特征词提取,并使用卡方检验(CHI-Squre)方法计算特征值的卡方,并按照卡方值从大到小进行排序,选取 Top N 的特征词;

(3)分别使用 TF-IDF 算法,TF-IDF-dist 算法及 TF-IDF-DL 算法计算各特征词的权重值;

(4)将各个特征词的权重值加入到朴素贝叶斯算

法中,计算得出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,输出对应分类信息;

(5)对比分析实验结果。

3.3 实验评估指标

文中使用准确率、召回率、F_1 值三个指标来评估算法效果。

(1)分类准确率 precision。

对于类别 C_i 的分类准确率定义为:分类结果中正确分类为 C_i 的样本数占分类结果中所有分为 C_i 类别的样本数(包含正确结果和错误结果)的比例。

$$P(C_i) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{11}$$

(2)召回率 recall。

对于类别 C_i 的召回率定义为:分类结果中正确分类为 C_i 的样本数占实际情形中分类为 C_i 的比例。

$$R(C_i) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{12}$$

(3) F_1 值

F_1 值其实是准确率和召回率的调和平均值,它的最大值是 1,最小值是 0。

$$F_1 = 2 * \frac{\mathrm{precison} * \mathrm{recall}}{\mathrm{precison} + \mathrm{recall}} \tag{13}$$

3.4 实验结果分析

在文本分类中少量的特征词不能对文本进行准确的分类预测,但特征词数量过大也会对实验有一定的消极影响。因此需要在分类前,找出最合适的特征词数量,由于特征词个数对所有权重值算法均适用,所以选择以 TF-IDF 算法为基准进行分类实验。由图 1 可得,随着特征词数量增加,precision 值逐渐提高,但当特征词数量过大时,文本分类时间将会大幅增加。针对选取的数据集,在选择特征词数量为 125 左右时,precision 增加速度开始减缓,且特征词数量在 160 左右时,分类时间开始变长。为了兼顾准确率和效率,该实验选取中间值 143 作为分类的特征词数量。

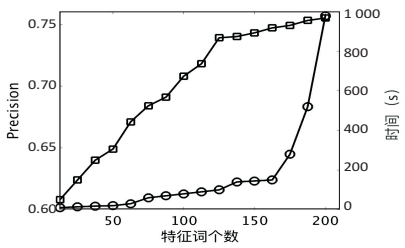

图 1 特征词个数对 precision 和时间的影响

在采用 TF-IDF-DL 算法计算贝叶斯特征权重时,需要计算出位置信息的影响因子:ε 和 δ。当 δ 值一定时,在初始范围内分类的准确率随词频位置影响度的增加而提高,但当词频位置影响力度达到一定程度时,会超出该词频实际的作用效果,从而夸大其影响力,对分类效果产生负面影响,因此词频位置信息的影响度会存在一个准确率峰值,当 ε 值小于这个峰值时,分类准确率会随着 ε 值的增大而提高,当 ε 值大于该峰值时,准确率会随之下降。同理,当 ε 值一定时,对分类影响大的词频会分布在近首尾处,但是与中心位置坐标距离太小和太大都会对准确率造成一定的不良影响,因此最优的 δ 值也存在一个准确率峰值。通过图 2(多个不同的 δ 值进行测试取得准确率的平均值)和图 3(多个不同的 ε 值进行测试取得准确率的平均值)可知,对于该数据集的 ε 和 δ 的最优取值分别为 1.5 和 $D/6$。

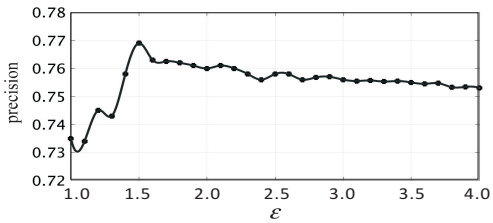

图 2 不同 ε 对 precision 值的影响

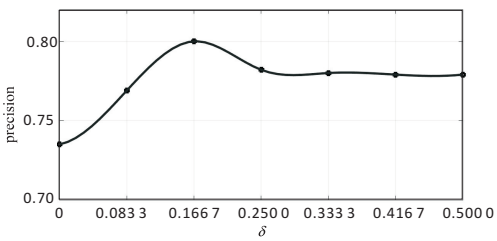

图 3 不同 ε 对 precision 值的影响

在对上述未知参数进行最优值求解后,进行 TD-IDF,TD-IDF-dist 以及 TF-IDF-DL 的权重值求解,并分别将求解权重值应用到贝叶斯文本分类中,得出相应的朴素贝叶斯分类器。并对选取的五种数据类别进行测试,记录每个类别对应测试结果的 precision、recall、F_1 值[15],如图 4 ~ 图 6 所示。

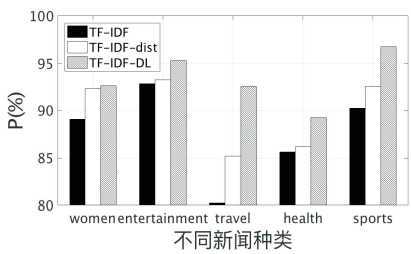

图 4 在不同新闻种类下不同算法对 P 的影响

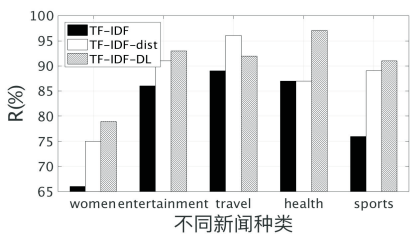

图 5 在不同新闻种类下不同算法对 R 的影响

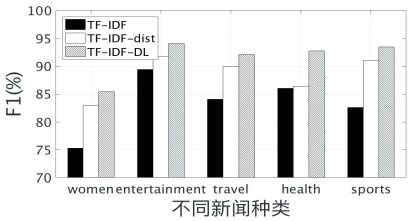

图 6　在不同新闻种类下不同算法对 F_1 的影响

通过结果可以看出,在特征词词频差异不明显且特征词位置没有明显规律的 women 类别上,应用 TF-IDF-DL 算法的朴素贝叶斯分类准确率没有特别明显的提高。

在特征词词频差异性和特征词位置规律性明显的类别上,基于 TF-IDF-DL 的贝叶斯文本分类表现出明显的优势。以 travel 类别为例(travel 类别文本中近首尾处多出现“游客”、“景点”等词汇),应用传统 TF-IDF 和 TF-IDF-dist 算法的朴素贝叶斯分类效果表现都不是很好,而应用 TF-IDF-DL 算法进行贝叶斯分类时在 travel 分类上表现依然良好。在研究以 TF-IDF-dist 计算权重的分类结果后,发现平均有近 10% 的 travel 新闻被分类到 entertainment 类别中,有 3.46% 的 travel 新闻被分类到 health 中。统计分类错误的新闻特征词发现,其中明显为 entertainment 分类的特征词占统计特征词的 31.67%,明显为 health 分类的特征词占统计特征词的 9.82%。这是由于 TF-IDF-dist 算法仅仅考虑了特征词在类内和类间的分布关系,却忽略了特征词在词频上的差异性和特征词位置信息规律这两个因素。而 TF-IDF-DL 算法在去除了此类文章中 entertainment 和 health 类别所属特征词的中心词频,且加入了特征词频的位置信息影响因子。

通过实验对比,基于 TF-IDF-DL 的贝叶斯算法在分类准确率、召回率和 F_1 值这三方面最高可比基于 TF-IDF-dis 的贝叶斯分类提高 8.6%、11.7% 和 7.4%。说明文中提出的基于 TF-IDF-DL 的贝叶斯分类算法在特征词词频有差异、特征词位置信息有规律的数据集上分类效果较好,是一种良好的分类算法。

4　结束语

通过研究词频出现规律以及文档中特征词的出现位置,提出加入去中心化词频因子和特征词距离因子来改进 TF-IDF 算法,并将改进后的 TF-IDF-DL 算法应用到朴素贝叶斯算法中。该算法能够解决在文本分类过程中存在特征属性权重一致及考虑指标单一的问题。通过使用搜狗实验室新闻数据作为数据集进行实验验证,并对实验结果进行分析。结果表明,该算法能够较好地提高分类性能,并对于不易区分的类别也能达到良好的分类效果,与国内最新研究的 TF-IDF-dis 相比,在分类准确率、召回率和 F_1 值这三方面最高可比其高 8.6%、11.7% 和 7.4%。但是该算法也存在一定的局限性,对于特征词词频差异小且词频位置不规律的数据分类效果没有明显提高,还需进一步完善。

参考文献:

[1] 杨　彬,韩庆文,雷　敏,等. 基于改进的 TF-IDF 权重的短文本分类算法[J]. 重庆理工大学学报:自然科学版,2016,30(12):108-113.

[2] 李鹏鹏,范会敏. 文本分类中特征权重算法改进研究[J]. 计算机与现代化,2018(2):66-70.

[3] 周丽杰,于伟海,郭　成. 基于改进的 TF-IDF 方法的文本相似度算法研究[J]. 泰山学院学报,2015,37(3):18-22.

[4] 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.

[5] 付　鑫. 基于改进型特征选择算法的文本分类方法研究[D]. 济南:山东师范大学,2018.

[6] YANG Y. Research and realization of internet public opinion analysis based on improved TF-IDF algorithm[C]//International symposium on distributed computing & applications to business. Anyang:IEEE,2017:80-83.

[7] DOMENICONI G,MORO G,PASOLINI R,et al. A comparison of term weighting schemes for text classification and sentiment analysis with a supervised variant of tf. idf[C]//International conference on data management technologies and applications. Colmar,France:Springer,2016:39-58.

[8] ALBITAR S, FOURNIER S, ESPINASSE B. An effective TF/IDF-based text-to-text semantic similarity measure for text classification[C]//International conference on web information systems engineering. Thessaloniki, Greece: Spinger,2014:105-114.

[9] CALVOC H. Simple TF · IDF is not the best you can get for regionalism classification[C]//Proceedings of the 15th international conference on computational linguistics and intelligent text processing. Kathmandu,Nepal:Springer,2014:92-101.

[10] LEE C H, GUTIERREZ F, DOU D. Calculating Feature weights in naive Bayes with Kullback-Leibler measure[C]//2011 IEEE 11th international conference on data mining. Vancouver,BC:IEEE,2011:1146-1151.

[11] 蔡银珊,黄英铭. 基于改进的 TF-IDF 特征权重算法的网页自动分类[J]. 绵阳师范学院学报,2010,29(8):106-109.

[12] 龚　静,胡平霞,胡　灿. 用于文本分类的特征项权重算法改进[J]. 计算机技术与发展,2014,24(9):128-132.

[13] 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.

[14] 陈　凯,黄英来,高文韬,等. 一种基于属性加权补集的朴素贝叶斯文本分类算法[J]. 哈尔滨理工大学学报,2018,23(4):69-74.

[15] 隗中杰. 文本分类中 TF-IDF 权重计算方法改进[J]. 软件导刊,2018,17(12):39-42.