基于特征噪声加权的特征权重算法改进*

赵 航[1],杨天奇[1],赵小厦[2]
(1.暨南大学 信息科学技术学院,广东 广州 510632;
2.华南师范大学 计算机学院,广东 广州 510631)

摘 要:特征权重算法 TF-IDF 是文本分类的重要算法之一,该算法 IDF 值容易受特征噪声影响出现波动。提出一种基于特征噪声加权的特征权重改进算法,该算法通过分析噪声特征的分布特点,对不能准确表达文档真实意思的特征噪声进行加权,降低特征噪声对 IDF 的影响,最终有效地提高算法的精度和健壮性。

关键词:向量空间模型;文本分类;特征噪声;特征权重;健壮性
中图分类号:TP391 **文献标识码**:A **文章编号**:1674-7720(2012)03-0066-03

Feature weight algorithm based on feature noise weighting

Zhao Hang[1], Yang Tianqi[1], Zhao Xiaoxia[2]
(1.College of Information Science and Technology, Jinan University, Guangzhou 510632, China;
2.College of Computer, South China Normal University, Guangzhou 510631, China)

Abstract: The algorithm of term weighting TF-IDF is one of the most important algorithm, but it fluctuates greatly when affected by the term noises. The paper proposes a feature weight algorithm basing on feature noise weighting. This algorithm analyses the distribution features of the term noises and weights the term noise which can't express the true meaning of the author in the document. Thereby the influence on the IDF is reduced, which is caused by the term noise. Finally the precision and the robustness are improved obviously.

Key words: VSM; text classification; feature noise; feature weighting; robustness

随着信息技术的发展,信息极度膨胀,人们迫切希望找到一种信息自动处理技术。文本分类作为信息处理的技术之一,由于其能对信息进行分类,使得获取信息更加容易,因而得到广泛应用。在文本分类的算法中,空间向量法中的 TF-IDF 算法由于其以词频 TF 和逆文档频数 IDF 的乘积作为向量坐标系的值,具有简单、直观、处理速度快的优点,得到广泛应用。但在理论和实际应用中有很大局限性,以至于其精度在各种文本分类中一直是较低的[1]。

本文针对噪声特征对 TF-IDF 算法逆文档频率(IDF)影响进行分析,提出了一种 IDF 加权方法,调整对 IDF 产生影响的特征噪声权重,有效减少了算法对噪声的影响,提高了 TF-IDF 算法的精度和健壮性。虽然已有很多研究者对 TF-IDF 算法做了改进,从特征选择上减少噪声特征的选择,但特征噪声在分类中出现是不可避免的。

1 向量空间算法的分析

向量空间算法的基本思想是用词袋法表示文本,将文本看做特征空间的一个向量,用两个向量之间的夹角来衡量两个文本之间的相似度。用 TF-IDF 值表示向量空间的一个特征值权重。

词语权重计算唯一的准则就是要最大限度地区分不同的文档。所以针对词语权重的计算,主要考虑 3 个因素[2]:

(1)词语频率 tf(term frequency):该词语在此文档中出现的频率。常用的计算方法是 $tf(T_k)=\sqrt{TF(T_k)}$;其中 $TF(T_k)$ 表示特征 T_k 出现的频率。

(2)词语的倒排文档频率 idf(inverse document frequency):该词语在文档中分布情况的量化,常用计算方法[3]为 $idf(T_k)=\log_2(N/n_k+L)$。其中 N 为文档集合中的

*基金项目:澳门科学技术发展基金(046/2010/A)

文档数目；nk 为出现过特征 T_k 的文档数目；L 根据实验来确定。

(3)归一化因子(normalization factor)：对各分量进行标准化。

根据上述3个因素，可以得出：TF与IDF的联合公式如下(其中 i 表示类别号)：

$$W_{ik}=\frac{\sqrt{tf_{ik}}\log_2(N/n_k+L)}{\sqrt{\sum_{k=1}^{n}tf_{ik}[\log_2(N/n_k+L)]^2}} \quad (1)$$

式(1)的提出基于这样一种假设[2]：对区别文档最有意义的词语应该是在文档中出现频率足够高，但在整个文档中出现频率足够少的词语。所以向量空间模型的基础是词语的出现频率和出现的文档频率[2]，但同时一个文档中的特征在不管出现的频率多少与文档频率的计算无关，文档频率的计算只与该特征在文档中出现与否有关。而特征噪声在文档中出现一般是以较小的频率出现。当一个特征以特征噪声的形式在大量文档中出现时(该特征本不应该在这些文档中出现)，文档频率计算出的值伴随特征噪声出现文档数目的不同变化很大。由于没有考虑特征受噪声影响的程度，只是单纯的以特征是否在文档中出现为依据计算文档频率，文档频率就不能够很好地在分类时起作用。

TF-IDF算法的IDF函数本质是一种抑制噪声的加权[3]。IDF函数认为文档频数少的单词就重要，而文档频数多的单词就无用，这样也使 IDF 值容易受噪声的影响。文档中的用词本身带有很大的随意性，用与不用某个词都行。大量的文档本身就不规范，并含有很多不规范的用词，导致降低了 IDF 值对单词权重的区分。

2 特征权重算法的改进

针对传统算法没有考虑噪声影响，对特征特点进行分析提出了改进算法。

2.1 文档特征分析

该文选择了搜狗实验室—搜狐新闻数据900篇文档进行特征分析，选出一篇文档 D_i，首先对 D_i 进行分词预处理，进行特征提取，特征降维。统计 D_i 出现词频为 t ($t=1,2,3,\cdots,10$)的特征个数占该 D_i 所有特征数 D_{in} 的比例 r_i，且对所有文档取平均值；然后进行特征降维前后文档的对比。

经统计得出，在降维前出现词频为1的特征所占比例约80%；词频为1和2的特征共占约90%。通过降维后词频为1的特征所占比例有所降低，但仍然超过50%，词频为1和2的特征共超过60%。由此可见特征词频为1、2占特征总数的绝大部分，虽然可以通过各种算法降低特征数，但降维后特征词频为1、2的仍然占特征总数的绝大部分。如果特征词频为1、2的特征属于噪声特征，这些特征在文档中出现与否也许不会影响所在文档的分类，但由于训练库的文档数非常多，这样可能会造成文档频率DF出现较大波动，使得 IDF 值出现大的波动，扰乱TF-IDF算法的准确性。

2.2 改进方法

可以这样认为：当特征词频 TF 较小时(例如 $TF=1$)，并不能有效地代表此特征在文档中的重要性，有很大几率是作者偶然性使用该特征；当特征词 TF 较大时，很多次偶然使用同一特征词的几率不大，很可能是该文档不得不使用该特征。由于文档作者用词具有很大的随意性，可以很随意用其他特征词代替，从而很容易使 TF 较小的特征词频的 $TF=0$，这一变化对 IDF 产生影响，如果词频 TF 在很多文档中出现频数很低，IDF 值就更容易受文档作者用词的影响从而扰乱TF-ID特征值的计算，使TF-IDF特征值偏离代表分类权重的意义。

从上述分析可以得到文档中大部分特征的词频为1或2，因此，如何降低噪声特征影响对TF-IDF算法精度计算至关重要。

本文降低特征噪声对 IDF 值计算影响的方法主要是通过对统计文档频数进行加权。原算法文档频数计算值是统计特征在文档集中出现的文档数，而改进的算法是统计特征在文档集中出现的加权文档数。使噪声特征降低对 IDF 值的影响，从而降低 IDF 的波动，提高TF-IDF算法的精度和稳定性。

使用WIDF(加权反文档频率)代替 IDF，WIDF的计算公式如下

$$widf(T_k)=\log_2\left(\frac{N}{\sum_{i=1}^{N}w_{ik}TD_i}+L\right) \quad (2)$$

其中，L 的取值通过实验来确定。N 表示文档集总数文档数，TD_i 表示 T_k 在第 i 个文档中是否出现(出现为1，不出现为0)，w_i 表示 T_k 在文档 i 出现的权重(通过实验确定容易受噪声影响的设为较小权重，否则为较大权重)。

实验在确定式(2)中的 w_i 值时，对 T_k 出现1和2的词频进行处理，因为1和2的词频低，且在图1中可以看出占用比例很大的更容易受噪声影响。当 T_k 在文档中出现频率为1时，w_i 通过实验最终确定为0.5；频率为2时，通过实验最终确定为0.9；频率大于2的词频通过实验确定的 w_i 非常接近1，所以出现频率大于2时 w_i 取为1。

3 实验与分析

3.1 实验数据

实验所有语料来源于搜狗实验室—搜狐新闻数据(SogouC.reduced.20061127)选取财经、IT、健康、体育、旅游、教育、招聘、文化、军事9个大类，总共4 500篇文档，其中1 800篇作为训练集(每个类200篇)，余下的2 700篇(每个类300篇)文档作为测试集。

3.2 评价指标

实验采用分类精度来评估权重算法在不同类上的

分类性能。分类精度的定义如下：

$$P=\frac{\text{正确分到该类别的文档数}}{\text{分到某个类别中的文档数}} \tag{3}$$

3.3 实验分析

k 近邻(k Nearest Neighbor,k-NN)分类算法基于类比学习的非参数分类算法,在文本分类领域获得广泛应用,对于未知分布和非正态分布可以获得较高分类准确率。实验采用式(4)进行相似度计算,用式(5)进行类别判定[4]:

$$sim(d_i,d_j)=\cos t(d_i,d_j)=\frac{\sum_{k=1}^{N}W_{k_i}\times W_{k_j}}{\sqrt{\sum_{k=1}^{N}(W_{k_i})^2}\times\sqrt{\sum_{k=1}^{N}(W_{k_j})^2}} \tag{4}$$

其中,W_{k_i}为特征词 tk 在文档 d_i 中的权重。

$$p(d_j,C_I)=\operatorname{argmax}\sum_{i=1}^{k}sim(d_i,d_j)P(d_i,C_I) \tag{5}$$

其中,k 为制定的最相似文本数量;$P(d_i,C_I)$在 d_i 属于 C_I 时取值为 1,否则为 0。分类判定时将待分类文本 d_j 的类别归为 $\sum_{i=1}^{k}sim(d_i,d_j)P(d_i,C_I)$ 最大时的类 C_I。

经过分词、去停用词、特征选择,表 1 和表 2 为 TF-IDF 与 TF-WIDF 两种特征算法在 k-NN 分类器上的实验结果,试验中取 k 为 50~75 中间的值,特征数为 3 000。在确定式(2)权重时,本实验只对出现词频为 1 或 2 的特征进行加权,词频为 1 的权重设为 0.5,词频为 2 的权重设为 0.9(即在计算特征文档频率时:当此特征在文档 D_i 中出现频率为 1 次时,在 D_i 中的文档频率为 0.5;当此特征在文档 D_i 中出现频率为 2 次时,在 D_i 中的文档频率为 0.9。其他保持不变)。

表 1 文档错误识别统计表(测试文档数 2 700)

K	TF-IDF 错误识别文档数	P	TF-WIDF 错误识别文档数	P
50	380	0.859 259 25	355	0.868 518 53
55	376	0.860 740 7	357	0.867 777 76
60	376	0.860 740 7	354	0.868 888 9
63	371	0.862 592 6	351	0.87
65	366	0.864 444 43	353	0.869 629 6
67	368	0.863 703 7	357	0.867 777 76
70	379	0.859 629 63	358	0.867 407 4
75	376	0.860 740 7	351	0.87

从表(1)可以看出在对 2 700 篇文档进行分类时,当 K 从 50~75 变化时:TF-IDF 算法错误识别文档数在 366~380 范围变化,波动范围为 14;TF-WIDF 算法错误识别文档数在 351~357 范围变化,波动范围为 6;由此得出当选不同 k 值时 TF-WIDF 算法比 TF-IDF 算法更加稳定。

从表(2)中可以看出 TF-WIDF 权重算法结合 k-NN 分类器在各类别上的精确度 P 除了在健康、财经有少许下降外大部分都有不同程度的提高,在所有类总体正确率提高 0.004~0.008。可以得出 TF-WIDF 算法比 TF-IDF 算法更加精确,与本文使用已经适当优化了传统 TF-IDF 算法有关,使其总体分类正确率高达 0.864 4,在如此高的正确率下再提高任何一点都是非常困难的,而本文正是在如此高的正确率基础上仍然使其提高 0.004~0.008,足可以证明本文的改进是有效的。从而说明 TF-WIDF 能有效地减少由于文档作者用词不规范、用词随意性造成文档特征噪声对 TF-IDF 算法的影响。尽管改进后的权重算法取得了一定效果,但文本分类问题设计文本表示、相似的计算、算法决策等多个方面改进权重算法并未使分类效果得到明显提高[4]。

表 2 各类正确率统计表

类别	TF-IDF K	TF-IDF P	TF-IDF K	TF-IDF P	提高值
军事	65	0.93	65	0.933 333 34	0.003
文化	65	0.816 666 66	65	0.836 666 64	0.020
招聘	65	0.893 333 3	65	0.926 666 7	0.033
教育	65	0.846 666 7	65	0.843 333 3	-0.003
旅游	65	0.833 333 3	65	0.833 333 3	0.000
体育	65	0.966 666 64	65	0.966 666 64	0.000
健康	65	0.833 333 3	65	0.823 333 3	-0.010
财经	65	0.85	65	0.83	-0.02
IT	665	0.81	65	0.833 333 3	0.023
所有类	665	0.864 444 43	65	0.869 629 6	0.005

通过加权减少了噪声特征对文本分类系统精度的影响。本文研究了传统的 TF-IDF 加权算法,在已适当优化算法基础之上提出噪声加权算法。实验证明,在传统算法基础上改进的加权算法及其他一些算法基础上的改进,都可有更好的表现。

参考文献

[1] 陆玉昌,鲁明羽.向量空间法中单词权重函数的分析和构造[J].计算机研究与发展,2002,39(10):1205-1210.

[2] 鲁松,李晓黎.文档中词语权重计算方法的改进[J].中文信息学报,2000,14(6):8-20.

[3] 李凯齐,刁兴春.基于信息增益的文本特征权重改进算法[J].计算机工程,2011,37(1):16-21.

[4] 台德艺,王俊.文本分类特征权重改进算法[J].计算机工程,2010,36(9):187-202.

(收稿日期:2011-08-12)

作者简介:

赵航,男,1985 年生,硕士,主要研究方向:数据挖掘,搜索引擎技术。

杨天奇,男,1961 年生,教授,主要研究方向:人工智能,数据挖掘,搜索引擎技术。

赵小厦,女,1982 年生,硕士,主要研究方向:云计算,移动互联网。