第 36 卷 第 9 期 计 算 机 工 程 2010 年 5 月
Vol.36 No.9 Computer Engineering May 2010
·人工智能及识别技术· 文章编号：1000—3428(2010)09—0197—03 文献标识码：A 中图分类号：TP391

文本分类特征权重改进算法

台德艺，王 俊
(合肥学院机器视觉与智能控制技术重点实验室，合肥 230601)

摘 要： TF-IDF 是一种在文本分类领域获得广泛应用的特征词权重算法，着重考虑了词频与逆文档频等因素，但无法把握特征词在类间与类内的分布情况。为提高在同类中频繁出现、类内均匀分布的具有代表性的特征词权重，引入特征词分布集中度系数改进 IDF 函数、用分散度系数进行加权，提出 TF-IIDF-DIC 权重函数。实验结果表明，基于 TF-IIDF-DIC 权重算法的 *K*-NN 文本分类宏平均 *F*1 值比 TF-IDF 算法提高了 6.79%。

关键词： 向量空间模型；文本分类；特征权重；特征分布

Improved Feature Weighting Algorithm for Text Categorization

TAI De-yi, WANG Jun
(Key Laboratory of Machine Vision and Intelligence Control Technology, Hefei University, Hefei 230601)

【Abstract】 TF-IDF as one of feature weighting schemes in Vector Space Model(VSM) is widely used and makes good results in the realm of text categorization. Although traditional algorithms consider about term frequency and inverse document frequency, Term Frequency/Inverse Document Frequency(TF-IDF) is oblivious to the term distribution information among and inside class. A new feature weighting algorithm based on the improved IDF and distribution coefficient is put forward to enhance the feature weighting of high frequency and homogeneous distribution in the same class. Experimental results show that compared with the conventional TF-IDF algorithm, f1 based on TF-IIDF-DIC raises by 6.79%.

【Key words】 Vector Space Model(VSM); text categorization; feature weighting; feature distribution

1 概述

文本分类是指按照预先定义的类别 $C=\{c_1,c_2,\cdots,c_l\}$，根据文档内容，将文档 d_j 归到一个或多个类别 c_l 的过程。随着各种电子资源的快速增加，文本自动分类在自动文摘、文本过滤、词义消歧、文档组织等信息技术领域的应用越来越广泛[1]。文本表示与文本分类算法的研究是文本分类领域中的 2 个基础性问题，本文仅对基于向量空间模型的文本表示中的特征词权重算法进行研究。

向量空间模型(Vector Space Model, VSM)因为将文档的内容形式化为多维向量空间中的一个点，将文档以向量的形式定义到实数域中，使模式识别中的各种成熟算法得以采用，提高了自然语言文档的可计算性与可操作性，所以成为最常用的文本表示模型[2]。VSM 的 2 个关键问题是特征的表征形式与权重计算。

特征的表征形式主要有基于字符、字、词、短语、句子及语义的表示。相关研究表明，基于词的文本表示具有简单、适用于处理大规模文本等优势而获得广泛应用，因此，基于词的权重算法将直接影响文本表示的准确性。

本文针对 TF-IDF(Term Frequency/Inverse Document Frequency)权重算法着重考虑特征词频(Term Frequency, TF)与逆文档频(Inverse Document Frequency, IDF)而没有考虑特征在各类别中分布比例情况，文献[3]引入信息论中信息增益的概念，以信息增益函数替换 IDF 来衡量特征词在文本集合中的分布差异，文献[4-5]利用特征选择函数代替 IDF 进行权值调整，取得了较好效果，文献[6]提出用于二元文本分类的 BNS 权重算法，文献[7]在置信度与支持度的基础上对 TF-IDF 权重算法进行了改进。这些算法主要以特征选择函数替代 IDF 函数，或者直接进行参数加权，在一定程度上改善了文本表示的准确性，提高了文本分类效果。为了突出在少量类别中大量出现且均匀分布的这样一类具有代表性的特征词的权重，本文提出一种基于分散度与集中度相结合的改进权重算法，利用集中度系数改进 IDF 函数，利用分散度参数进行加权，实验证明该方法简单、有效。

2 TF-IDF特征词权重算法

2.1 向量空间模型

VSM 的基本思想是用词袋法表示文本，将每个特征词作为向量空间坐标系的一维，文本被形式化为多维向量空间中的一个向量，文本之间的相似度用 2 个向量间的夹角衡量[8]。向量中每一维的值用特征词权重表示，权重体现了特征词在表征文档时的重要程度，特征词权重的计算主要使用文献[9]的 TF-IDF 算法。

2.2 TF-IDF算法

TF-IDF 算法主要考虑特征词的词频、逆文档频、归一化等因素[10]。相关术语定义如下：

(1)词频。词频是指特征词 t_i 在文档 d_j 中出现的次数，用 $TF(t_{ij})$ 表示。在排除停用词及个别高频词的前提下，特征词在文档中出现的次数越多，其表征文档的能力越强。

基金项目： 安徽省高校省级自然科学基金资助项目(KJ2008B120)
作者简介： 台德艺(1974－)，男，讲师、硕士，主研方向：人工智能，管理信息系统；王 俊，教授、博士
收稿日期： 2009-10-15 **E-mail：** taideyi@hfuu.edu.cn

(2)文档频。文档频(Document Frequency, DF)是指文档集 C 中出现特征词 t_i 的文档数，用 $N(t_i,C)$ 表示，特征词 t_i 出现的文档数 $N(t_i,C)$ 越大， t_i 对文档 d_j 的代表性越弱。

(3)逆文档频。逆文档频是指特征词 t_i 对文档 d_j 的表示能力与其在文档中出现的次数 $N(t_i,C)$ 呈反比，用 $IDF(t_i)$ 表示：

$$IDF(t_i)=\mathrm{lb}(\frac{N(C)}{N(t_i,C)}) \tag{1}$$

其中， $N(C)$ 为训练集中的总文档数； $IDF(t_i)$ 随着 $N(t_i,C)$ 的增大而减小。文档集 C 中出现 t_i 的文档数 $N(t_i,C)$ 越小， t_i 对 d_j 越具有代表性。

(4)归一化。为降低个别高频特征词对低频特征词的抑制作用，对各分量进行归一化。

归一化后的 TF-IDF 计算如下：

$$weight_{\text{TF-IDF}}(t_{ij})=\frac{TF(t_{ij})\times\mathrm{lb}(\frac{N(C)}{N(t_i,C)}+L)}{\sqrt{\sum_{j=1}^{n}[TF(t_{ij})\times\mathrm{lb}(\frac{N(C)}{N(t_i,C)}+L)]^2}} \tag{2}$$

其中，L 为实验确定的参数，通常取 L=0.01。式(2)基于如下假设：区分文档能力强的词是那些在某个文档中出现次数足够多、在其他文档中出现次数足够少的词。

2.3 TF-IDF算法分析

TF-IDF 算法主要具有以下 2 个特点：

(1)逆文档频的概念是特征词 t_i 分布的文档数 $N(t_i,C)$ 越少就越具有代表性，没有考虑文档数在同一类别内增加的情况。如果特征词 t_i 在同类各文档中频繁出现，说明该特征能很好地代表这个类，这样的词条应该给予较高的权重。而在式(1)中，当 C_l 类中包含特征词 t_i 的文档数 $N(t_i,C_l)$ 增大时，IDF 反而减小。因此，在同类中大量出现而在其他类中较少出现的这样一类具有代表性的特征词，其权重在 TF-IDF 算法中不但没有得到加强，反而减弱。

(2)对于同一类别中的 2 个不同特征词 t_1， t_2， t_1 在各文档中出现次数比 t_2 平均，则 t_1 在表征类别的能力上比 t_2 更有代表性。如果 t_2 只是在类中的一二篇文档中大量出现，而在其他文档中出现很少，则不排除这一二篇文档是该类别中的特例，这样的特征不具有代表性，权重应较低，对于这种情况 TF-IDF 算法也无法区分。

3 特征词权重算法的改进

3.1 改进的思想

基于第 2 节的分析，权重除了与词频密切相关外，还应突出以下 2 类特征词：

(1)分布的类别越少，在同一类别中出现的文本越多，特征词的区分能力越强，权重应越大。

(2)在类内各文档中的分布越平均，特征词越具有代表性，权重应越大。

根据这一思想提出改进的特征权重算法 TF-IIDF-DIC：

$$weighting_{\text{TF-IIDF-DIC}}(t_{ij})=TF\times IIDF\times DIC \tag{3}$$

其中，TF 为词频，$TF=TF(t_{ij})$；$IIDF$ 为改进的逆文档频率；DIC 为类内分散度系数。

3.2 基于分布集中度的*IIDF*

考虑到第(1)类特征词的分布情况，对式(1)进行如下变形：

$$IDF=\mathrm{lb}(\frac{N(C)}{N(t_i,C)})=\mathrm{lb}(\frac{N(C)}{N(t_i,C_l)+N(w_j,\overline{C_l})}) \tag{4}$$

其中， $N(t_i,C_l)$ 表示 C_l 类中出现特征词 t_i 的文档数； $N(t_i,\overline{C_l})$ 表示非 C_l 类中出现特征词 t_i 的文档数。

特征词分布的类别越少，在同一类别中出现的文本越多，特征词的区分能力和权重应越大，因此，加入集中度系数 λ 对 IDF 进行改进。

$$\lambda=\frac{N(t_i,C_l)}{N(C_l)} \tag{5}$$

其中， $N(C_l)$ 为训练集中 C_l 类的总文档数； $\lambda\leqslant 1$。

加入集中度系数 λ 的 $IIDF$ 如下：

$$\begin{aligned}IIDF&=\mathrm{lb}[\frac{N(C)}{N(t_i,C_l)+N(t_i,\overline{C_l})}\times\lambda]=\\&\mathrm{lb}[\frac{N(C)}{N(t_i,C_l)+N(t_i,\overline{C_l})}\times\frac{N(t_i,C_l)}{N(C_l)}]=\\&\mathrm{lb}[\frac{N(t_i,C_l)}{N(t_i,C_l)+N(t_i,\overline{C_l})}\times\frac{N(C)}{N(C_l)}]\end{aligned} \tag{6}$$

其中， $N(C)$, $N(C_l)$ 为常数，可以证明 $\frac{N(t_i,C_l)}{N(t_i,C_l)+N(t_i,\overline{C_l})}$ 是 $N(t_i,C_l)$ 的单调递增函数，因此，$IIDF$ 是 $N(t_i,C_l)$ 的单调递增函数。

$\frac{N(t_i,C_l)}{N(t_i,C_l)+N(t_i,\overline{C_l})}$ 是 $N(t_i,C_l)$ 的单调递增函数证明如下：

令 $f(N(t_i,C_l))=\frac{N(t_i,C_l)}{N(t_i,C_l)+N(t_i,\overline{C_l})}$，设 t_i 在 C_l 类中出现的文档数增多为 $N_2(t_i,C_l)$，而在 $\overline{C_l}$ 类中出现的文档数没有变化，那么，

$$\begin{aligned}&N_2(t_i,C_l)>N(t_i,C_l)\text{ , }N_2(t_i,\overline{C_l})=N(t_i,\overline{C_l})\\&f(N_2(t_i,C_l))-f(N(t_i,C_l))=\\&\quad\frac{N_2(t_i,C_l)}{N_2(t_i,C_l)+N_2(t_i,\overline{C_l})}-\frac{N(t_i,C_l)}{N(t_i,C_l)+N(t_i,\overline{C_l})}=\\&\quad\frac{N_2(t_i,C_l)N(t_i,\overline{C_l})-N(t_i,C_l)N_2(t_i,\overline{C_l})}{[N_2(t_i,C_l)+N_2(t_i,\overline{C_l})][N(t_i,C_l)+N(t_i,\overline{C_l})]}=\\&\quad\frac{N_2(t_i,C_l)N(t_i,\overline{C_l})-N(t_i,C_l)N(t_i,\overline{C_l})}{[N_2(t_i,C_l)+N_2(t_i,\overline{C_l})][N(t_i,C_l)+N(t_i,\overline{C_l})]}=\\&\quad\frac{N(t_i,\overline{C_l})[N_2(t_i,C_l)-N(t_i,C_l)]}{[N(t_i,C_l)+N(t_i,\overline{C_l})][N(t_k,C_l)+N(t_k,\overline{C_l})]}>0\end{aligned}$$

当 $N_2(t_i,C_l)>N(t_i,C_l)$ 时， $f(N_2(t_i,C_l))>f(N(t_i,C_l))$ ，$f(N(t_i,C_l))$ 是 $N(t_i,C_l)$ 的单调递增函数得证。

式(6)表明，当特征词 t_i 在同类中出现的文档数 $N(t_i,C_l)$ 增大时，$IIDF$ 增大，在其他类别出现的文档数 $N(t_i,\overline{C_l})$ 增大时，$IIDF$ 减小。当 t_i 只在一个类别中出现时，$IIDF$ 取得最大值 $\mathrm{lb}\frac{N(C)}{N(C_l)}$，此时 t_i 对 C_l 类的代表性最强。当 t_i 在文档集的每个文档中都出现时， $N(t_i,C_l)=N(C_l)$, $N(t_i,C_l)+N(t_i,\overline{C_l})=N(C)$，$IIDF$ 为 0，这样的 t_i 对 C_l 类不具有代表性。

3.3 基于分布分散度的*DIC*

特征词 t_i 在类内各文档中分布越平均越具有代表性，可以用式(7)的分散度系数 DIC 表示。

$$DIC=1-\frac{\frac{1}{N(C_l)-1}\sum_{j=1}^{N(C_l)}[TF(t_{ij})-\overline{TF}(t_i,C_l)]^2}{N(C_l)\times\overline{TF}(t_i,C_l)^2} \tag{7}$$

其中，$\overline{TF}(t_i,C_l)$ 为特征词 t_i 在 C_l 类各文档中出现的平均次数，$\overline{TF}(t_i,C_l)=\frac{1}{N(C_l)}\sum_{j=1}^{N(C_l)}TF(t_{ij})$。当 t_i 在 C_l 类文档中均匀出现，$TF(t_{ij})=\overline{TF}(t_i,C_l)$，此时 *DIC* 取得最大值 1，特征词 t_i 对类 C_l 的表示能力最强。

当 t_i 仅在 C_l 类中的一个文档中出现时，*DIC* 取最小值 0，这样的特征词 t_i 对类 C_l 不具有代表性。证明如下：

当 t_i 仅在一个文档中出现时，

$$\sum_{j=1}^{N(C_l)}[TF(t_{ij})-\overline{TF}(t_i,C_l)]^2=$$
$$(N(C_l)-1)\overline{TF}(t_i,C_l)^2+[TF(t_{ij})-\overline{TF}(t_{ij},C_l)]^2 \tag{8}$$

代入式(7)：

$$DIC=1-\frac{\frac{1}{N(C_l)-1}\sum_{j=1}^{N(C_l)}[TF(t_{ij})-\overline{TF}(t_i,C_l)]^2}{N(C_l)\times\overline{TF}(t_i,C_l)^2}=$$
$$1-\frac{\frac{1}{N(C_l)-1}[(N(C_l)-1)\overline{TF}(t_i,C_l)^2+[TF(t_{ij})-\overline{TF}(t_{ij},C_l)]^2]}{N(C_l)\times\overline{TF}(t_i,C_l)^2}=$$
$$1-\frac{\frac{1}{N(C_l)-1}[(N(C_l)-1)\overline{TF}(t_i,C_l)^2+[N(C_l)\times\overline{TF}(t_i,C_l)-\overline{TF}(t_i,C_l)]^2]}{N(C_l)\times\overline{TF}(t_i,C_l)^2}=$$
$$1-\frac{\frac{1}{N(C_l)-1}[(N(C_l)-1)+[N(C_l)-1]^2]}{N(C_l)}=1-\frac{1+[N(C_l)-1]}{N(C_l)}=0$$

3.4 TF-IIDF-DIC 归一化

为了降低个别高频特征对低频特征的抑制作用，对各分量进行归一化。归一化后的 TF-IIDF-DIC 权重公式如下：

$$weighting_{\text{TF-IIDF-DIC}}(t_{ij})=\frac{TF\times IIDF\times DIC}{\sqrt{\sum_{i=1}^{n}(TF\times IIDF\times DIC)^2}} \tag{9}$$

4 实验与分析

4.1 实验数据

实验数据来源于中文自然语言处理开放平台上由李荣陆提供的语料库，训练语料采用人工标注的 9 804 篇文档，分 20 个类别，测试语料共 9 833 篇文档，训练语料和测试语料基本按照 1:1 划分。由于语料库分布极不均衡，训练集中文档数最多的经济类达到 1 600 篇，最少的通信类仅 25 篇，且存在部分重复，因此，实验中选择文档较多的艺术、航空、计算机、环境、农业、经济、政治、体育 8 个类别，并分别从训练和测试集中随机选出 500 篇文档进行实验。

4.2 评价指标

实验采用准确率 *P*、召回率 *R*, *F*1 测试值评估权重算法在不同类别上的分类性能，用宏平均 *MacF*1 评价特征权重算法在整个数据集上的分类性能。

$$P=\frac{\text{正确分类的文本数}}{\text{实际分类输出的文本数}} \tag{10}$$

$$R=\frac{\text{正确分类的文本数}}{\text{分类中应用的文本数}} \tag{11}$$

$$F1=\frac{2\times P\times R}{P+R} \tag{12}$$

$$MacF1=\frac{1}{n}\sum_{i}^{n}F1_i \tag{13}$$

4.3 实验分析

k 近邻(*k* Nearest Neighbor, *k*-NN)分类算法是一种基于类比学习的非参数分类方法，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的分类准确率。实验中 *k*-NN 分类算法采用式(14)进行相似度计算，用式(15)进行类别判定。

$$sim(d_i,d_j)=\cos(d_i,d_j)=\frac{\sum_{k=1}^{n}w_{ki}\cdot w_{kj}}{\sqrt{\sum_{k=1}^{n}w_{ki}^{\ 2}}\times\sqrt{\sum_{k=1}^{n}w_{kj}^{\ 2}}} \tag{14}$$

其中，w_{ki} 为特征词 t_k 在文档 d_i 中的权重。

$$p(d_j,C_l)=\arg\max\sum_{i=1}^{k}sim(d_j,d_i)P(d_i,C_l) \tag{15}$$

其中，k 为指定的最相似文本数量；$P(d_i,C_l)$ 在 d_i 属于 C_l 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 $\sum_{i=1}^{k}sim(d_j,d_i)P(d_i,C_l)$ 最大时的类 C_l。

经过分词、去除停用词、特征选择，表 1 为式(2)与式(9) 2 种特征权重算法在 *k*-NN 分类器上的实验结果，实验中选取 *k* 为 15，特征维数为 4 000。

表 1　2 种算法在 *k*-NN 分类器上的结果比较　(%)

类别	TF-IDF			TF-IIDF-DIC		
	P	*R*	*F*1	*P*	*R*	*F*1
艺术	77.23	75.83	76.52	82.46	80.11	81.27
航空	78.86	75.32	77.05	87.71	79.83	83.58
计算机	78.05	76.45	77.24	87.44	83.87	85.62
环境	70.02	67.81	68.90	85.65	70.04	77.06
农业	78.87	73.45	76.06	83.42	80.26	81.81
经济	80.12	55.86	65.83	80.78	70.08	75.05
政治	78.75	67.43	72.65	86.28	78.06	81.96
体育	86.46	78.83	82.47	88.23	81.42	84.69

从表 1 可以看出，TF-IIDF-DIC 权重算法结合 *k*-NN 分类器在各类别上的查全率、查准率及 *F*1 值均高于 TF-IDF 算法。其中，经济类的 *F*1 值提高最多，由原来的 65.82%提高到 75.05%，增长了 9.22 个百分点，增长最小的体育类也达到 2.22 个百分点。经济类与环境类分类效果差，可能与这 2 类文档中的特征词有一定的交叉有关。实验表明，改进的权重算法提高了具有代表性的特征词的权重，弱化了表征能力不强的特征以及对分类没有帮助的噪声特征，使向量在特征空间中向有用特征代表的维度旋转了一个角度，旋转后无用特征的词频差异对向量夹角的影响被减小，而有用特征词频差异对向量夹角的影响被加强。

在总体性能的评价上，TF-IIDF-DIC 权重算法结合 *k*-NN 文本分类器的宏 *F*1 值达到了 81.38%，比 TF-IDF 算法的 74.59%高出 6.79 个百分点，这说明改进的权重算法通过对特征词进行权值调整，突出了重要特征，抑制了次要特征，使文本表示在总体上更合理。同时应该指出，尽管改进后的权重算法取得了一定效果，但文本分类问题涉及到文本表示、相似度计算和算法决策等多个方面，改进的权重算法并未使分类效果得到明显的提高。

5 结束语

为了解决 TF-IDF 算法在处理分布类别少、出现文本多且平均分布的这类具有代表性的特征词的权重方面存在的不足，本文通过分布集中度系数改进 *IDF*，用分散度系数进行加权，提出了 TF-IIDF-DIC 算法。实验表明，TF-IIDF-DIC 算法性能优于经典 TF-IDF 算法，使文本表示更合理，有效提高了文本分类精度。

（下转第 202 页）