文本分类中词语权重计算方法的改进与应用

熊忠阳,黎 刚,陈小莉,陈 伟
XIONG Zhong-yang,LI Gang,CHEN Xiao-li,CHEN Wei

重庆大学 计算机学院,重庆 400030
College of Computer,Chongqing University,Chongqing 400030,China

XIONG Zhong-yang,LI Gang,CHEN Xiao-li,et al.Improvement and application to weighting terms based on text classification.Computer Engineering and Applications,2008,44(5):187-189.

Abstract: Text representation has been the fundamental problem in Information Retrieval.tf.idf (term frequency,inverse document frequency) as one of term weighting schemes in Vector Space Model is a good text representation,Which is popular and make good results in the field of Information Retrieval.The difference of the proportion of distribution of terms in text collection is one of the most important factors of expressing the content of text.But the calculation of IDF,don't consider the information of distribution about terms among classes,and don't consider the more term weighting for the terms of the relative distributed balance inner classes.The improved TFIDF are used to select feature,KNN algorithm and genetic algorithm are used to train the classifier.and proves that the improved TFIDF method is feasible.

Key words: text representation;Vector Space Model;feature selection;TFIDF

摘 要: 文本的形式化表示一直是信息检索领域关注的基础性问题。向量空间模型(Vector Space Model)中的 tf.idf 文本表示是该领域里得到广泛应用,并且取得较好效果的一种文本表示方法。词语在文本集合中的分布比例量上的差异是决定词语表达文本内容的重要因素之一。但是其 IDF 的计算,并没有考虑到特征项在类间的分布情况,也没有考虑到在类内分布相对均匀的特征项的权重应该比分布不均匀的要高,应该赋予其较高的权重。用改进的 TFIDF 选择特征词条、用 KNN 分类算法和遗传算法训练分类器来验证其有效性,实验表明改进的策略是可行的。

关键词: 文本表示;向量空间模型;特征选择;TFIDF
文章编号: 1002-8331(2008)05-0187-03 **文献标识码:** A **中图分类号:** TP391

1 引言

文档的向量化是进行文本分类的基础。只有文档向量很好地保存了原有文档的信息, 文本分类才可能有令人满意的结果。而文档向量的复杂程度(一般来说就是其维数)则很大程度上影响了文本分类的处理结果。在对文档向量化时,通常先使用 TF-IDF (Term Frequency-Inverse Document Frequency)公式将文档向量化。因此,向量空间模型文档表示的形式化方法是基于文档处理的各种应用得以形式化的基础和前提。本文通过研究发现传统的文本特征权值表示方法 TFIDF 的不足:其 IDF 的计算,并没有考虑到特征项在类间的分布情况,也没有考虑到在类内分布相对均匀的特征项的权重应该比分布不均匀的要高,应该赋予其较高的权重。本文对此进行了改进,结合特征项的类间,类内分布,以及类间特征项不完全分类对分类的影响,提出了一种新的特征权重算法。在为每个类训练分类器的时候,用到了 KNN 和遗传算法,最后通过实验证明改进的 TFIDF 方法是可行的。

2 传统的 TFIDF

传统的特征权重算法主要考虑特征项的频率信息 TF 以及反文档频率信息 IDF[2]。

2.1 特征项的频率信息[1]

特征项频率(TF)是指特征项在文档中出现的次数。特征项可以是字、词、短语,也可以是经过语义概念词典进行语义归并或概念特征提取后的语义单元。不同类别的文档,在某些特征项的出现频率上有很大差异,因此频率信息是文本分类的重要参考之一。在最初的文本自动分类中,文档向量就是用 TF 来构造的。

2.2 反文档频率信息

1972 年,Spark Jones 提出计算文献频率有助于计算词权重,从此,Inverse Document Frequency(IDF)公式在信息检索中占据重要地位[3]。反文档频率是特征项在文档集分布情况的量化。IDF 常用的计算方法为:

基金项目: 重庆市自然科学基金(the Natural Science Foundation of Chongqing City of China under Grant No.CSTC2006BB2021)。
作者简介: 熊忠阳(1964-),男,博士,教授,博士生导师,主要研究领域为数据挖掘、数据库、并行计算、网络信息处理;黎刚(1976-),男,硕士研究生,主要研究方向为数据挖掘,数据库在 Internet 上的应用、自然语言处理和 WEB 搜索;陈小莉(1979-),女,硕士研究生,主要研究方向为数据挖掘,自然语言处理;陈伟(1974-),男,硕士研究生,主要研究方向为数据挖掘,群集技术。
收稿日期: 2007-05-28 **修回日期:** 2007-07-25

$$idf(T_k)=\log(\frac{N}{n_k}+0.1) \quad (1)$$

其中 N 为文档集中的总文档数，n_k 为出现特征项 T_k 的文档数。

IDF 算法的核心思想是，在大多数文档中都出现的特征项不如只在小部分文档中出现的特征项重要。IDF 算法能够弱化一些在大多数文档中都出现的高频特征项的重要度，同时增强一些在小部分文档中出现的低频特征项的重要度。

一个有效的分类特征项应该既能体现所属类别的内容，又能将该类别同其它类别相区分。所以，在实际应用中 TF 与 IDF 通常是联合使用的。TF 与 IDF 的联合公式如下[4]（其中 i 代表类别号）：

$$Weight_{TF\_IDF}(T_k)=tf(T_{ik})\times idf(T_k) \quad (2)$$

在很多情况下还需要将向量归一化，TFIDF 的归一化计算公式[4]如下（其中 n 表示类别 i 中特征项的总个数）：

$$Weight_{TF\_IDF}(T_{ik})=\frac{tf(T_{ik})\times idf(T_k)}{\sqrt{\sum_{j=1}^{n}(tf(T_{ik})idf(T_k))^2}} \quad (3)$$

2.3 TFIDF 的不足

TFIDF 的不足，主要表现 TFIDF 没有考虑特征项在类间，类内和不完全分类的分布信息。

2.3.1 TFIDF 没有考虑特征项在类间的分布信息

如果某一类 C_i 中包含词条 T_k 的文档数为 m，而其它类包含 T_k 的文档总数为 k，显然所有包含 T_k 的文档数 $n=m+k$，当 m 大的时候，n 也大，按照 IDF 公式（1）得到的 IDF 的值会小，则表示该词条 T_k 类别区分能力不强。但是实际上，m 大，说明词条 T_k 在 C_i 类的文档中频繁出现，就说明 T_k 词条能够很好地代表 C_i 类的文本特征，应该赋予较高的权重并选作该类文本的特征词。这就是 IDF 没有考虑特征词在类间分布的一个方面；另一方面，虽然包含 T_k 的文档数 n 较小，但是如果其均匀分布在各个类间，这样的特征词不适合用来分类，应该赋予较小的权重，可按照传统的 TFIDF 算法计算其 IDF 值却很大。

2.3.2 TFIDF 没有考虑特征项在类内的分布信息

同样是集中分布于某一类别的不同特征项，类内分布相对均匀的特征项的权重应该比分布不均匀的要高。传统的 TFIDF 算法，也没有考虑这一情况。

2.3.3 TFIDF 没有考虑特征项不完全分类的情况

实际使用的已分类的训练文本集通常是不完全的分类。即有些类别的文档集还可以继续划分出更细的类别。如，计算机类一般来说至少可以再细分出计算机硬件、计算机软件两个子类。在这种不完全的分类条件下，各个子类文章所占的比重是不均衡的。可能在某个计算机类的文本集中，软件类的占了 80%，硬件类的只有 20%的比例。在这个训练集中，属于计算机硬件类的特征词也应该作为判别计算机类文章的特征词。如果某些词在一类文章中整体出现频率较低，但是在本类中一定数量的文章中出现较频繁，那么这些词也应该对分类来说具有较多的信息量。这就是不完全分类的情况。

3 改进的 TFIDF

特征项的分布信息，又称为特征项频率分布的离散度[1]。用特征词在类间和类内部的分布的离散度来描述特征词在类间和类内部的分布情况，用特征词在类间和类内部的分布的离散度和不完全分类的词频差异来修正 TFIDF 公式。

3.1 特征项的类间离散度[1]

设总共有 n 个类，$tf_i(T_k)$ 代表词条 T_k 在 C_i 类的出现频率，$\overline{tf(T_k)}$ 代表词条 T_k 的在各个类的平均词频，计算公式为

$$\overline{tf(T_k)}=\frac{1}{n}\sum_{i=1}^{n}tf_i(T_k) \quad (4)$$

则类间的离散度 D_{ac} 为：

$$D_{ac}=\frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(tf_i(T_k)-\overline{tf(T_k)})^2}}{\overline{tf_{(T_k)}}} \quad (5)$$

3.2 特征项的类内离散度

词条 T_k 在各个类内部分布情况，设 C_i 类中总的文档数为 m，将 T_k 在各个文档的词频看作是 T_k 在各个文档中的取值，$\overline{tf_i(T_k)}$ 表示 T_k 在类 C_i 文档中的平均词频，其计算公式为：

$$\overline{tf_i(T_k)}=\frac{1}{m}\sum_{j=1}^{m}tf_{ij}(T_k) \quad (6)$$

用 D_{ii} 表示 T_k 在类 C_i 中的文档中无偏方差，则：

$$D_{ii}=\frac{1}{m}\sum_{j=1}^{m}(tf_{ij}(T_k)-\overline{tf_i(T_k)})^2 \quad (7)$$

则类内离散度 D_{ic} 为：

$$D_{ic}=\frac{\sqrt{D_{ii}}}{\frac{m}{\sqrt{m-1}}\overline{tf_i(T_k)}} \quad (8)$$

可以证明，$D_{ic}<1$。

3.3 特征项的不完全分类的词频差异

考虑到特征项在类中的不完全分类，引入了一个权重参数：词频差异（Word Frequency Differentia Based，WFDB）：

$$WFDB=W(t_{ik})=1+\lambda\frac{\sqrt{D_{ii}}}{tf_i(T_k)} \quad (9)$$

λ 为比例系数，由实际的情况进行调整。

结合以上三个方面的考虑，加上归一化处理后，TF-IDF 变为了 TF-IDF-DI-WFDB：

$$Weight_{TF-IDF-DI-WFDB}(T_{ik})=\frac{tf(T_{ik})\times idf(T_k)\times D_{ac}}{\sqrt{\sum_{k=1}^{n}(tf(T_{ik}))^2[idf(T_k)]^2}}\times(1-D_{ic})\times WFDB \quad (10)$$

4 分类模型及策略

本文采用的分类模型如图1[10]所示。

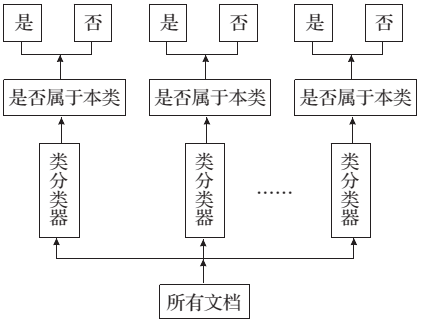

图 1 分类模型

当类分类器训练好后,要看测试文档是否属于该类,比较测试文档和类分类器的相似程度,将其分类到与之相似度最大的分类器所对应的类中。

本文采用 KNN 和遗传算法进行分类，以验证改进的 TFIDF 的有效性和可行性。

5　实验及其结果分析

5.1　评价指标

对分类器性能评价的主要指标有召回率(Recall,亦称查全率)、精确率(Precision,亦称查对率)。假设 TP_i 表示测试文档集中本来属于类别 C_i 而且被分类器分类到类别 C_i 的文档数,FP_i 表示测试文档集中本来不属于类别 C_i 但却被分类器错误分类到类别 C_i 的文档数,FN_i 表示本来应该属于类别 C_i 但被分类器分类到别的类别的文档数,而 TN_i 表示本来不属于类别 C_i 也没有被分类器分类到类别 C_i 的文档数。那么,分类器在类别 C_i 上的查全率(Recall)定义[5]如下:

$$Recall_i=\frac{TP_i}{TP_i+FN_i} \tag{11}$$

类器在类别 C_i 上的查对率定义[5]如下:

$$Precision_i=\frac{TP_i}{TP_i+FP_i} \tag{12}$$

对于类别 C_i,其 F_1[6]定义为:

$$F_1=\frac{2*Recall*Precision}{Precision+Recall} \tag{13}$$

5.2　实验结果分析

实验采用数据集来源于复旦大学,其中训练样本和测试样本分别都有 10 个类,训练样本共有 1 882 个文档,而测试样本有 2 816 个文档。

实验结果的 KNN 算法的 K=8,相似度阈值为 0.8;遗传算法的初始种群为 100，交叉概率取为 0.75，变异概率取值为 0.05,相似度阈值为 0.78。

把传统的 TFIDF 和 KNN 的分类效果与改进的 TFIDF 和 KNN 的分类效果，改进的 TFIDF 和遗传算法的分类效果进行比较,如表 1,2 所示。

表 1　宏平均、微平均分类效果比较

		TFIDF 与 KNN	New TFIDF 与 KNN	New TFIDF 与 GA
查全率/%	宏平均	90.119	91.466	90.639
	微平均	91.264	92.259	91.832
查对率/%	宏平均	92.307	93.296	93.149
	微平均	91.264	92.259	91.832
F_1/%	宏平均	88.261	90.126	88.204
	微平均	45.043	45.716	45.273

TFIDF 与 KNN 指传统的 TFIDF 和 KNN 的分类效果,New TFIDF 与 KNN 指改进的 TFIDF 和 KNN 的分类效果,New TFIDF 与 GA 指改进的 TFIDF 和遗传算法的分类效果,R、P、F 分别指查全率、查准率和 $F1$ 评估值。

表 2　各个类的分类效果比较

	TFIDF 与 KNN			New TFIDF 与 KNN			New TFIDF 与 GA		
	R/%	P/%	F/%	R/%	P/%	F/%	R/%	P/%	F/%
环境	85.075	90.476	87.692	88.557	91.753	90.126	85.572	91.005	88.204
医药	83.824	94.475	88.831	85.784	98.315	91.623	83.333	98.266	90.185
军事	84.337	92.920	88.420	86.345	89.958	88.114	85.141	92.982	88.888
经济	92.923	87.791	90.284	93.538	88.372	90.881	94.769	83.469	88.760
教育	91.364	92.202	91.781	94.091	93.243	93.665	92.727	93.578	93.150
体育	96.889	94.168	95.509	97.111	92.979	95.000	97.333	95.425	96.369
艺术	88.710	90.909	89.796	89.516	95.279	92.307	89.113	94.850	91.892
政治	95.050	85.106	89.803	94.257	87.823	90.926	95.446	86.225	90.601
计算机	90.500	98.907	94.516	92.000	100.000	95.833	89.500	100.000	94.459
交通	92.523	96.117	94.285	93.458	95.238	94.339	93.458	95.694	94.562

6　结束语

本文从类间、类内和类内的不完全分类的角度,对 TFIDF 进行了改进,并采用 KNN 和遗传算法来为每个类训练分类器,把传统 TFIDF 结合 KNN 的分类效果分别和改进的 TFIDF 结合遗传算法的分类效果,改进的 TFIDF 结合 KNN 的分类效果作了比较。实验结果表明,改进的 TFIDF 结合遗传算法的分类效果和改进的 TFIDF 结合 KNN 的分类效果,从总体上都要比传统 TFIDF 结合 KNN 的分类效果好，因此改进的 TFIDF 是有效的且可行的。

参考文献:

[1] 徐凤亚,罗振声.文本自动分类中特征权重算法的改进研究[J].计算机工程与应用,2005,41(1):181-183.

[2] Auen J.Natural language understanding[M].[S.l.]:The Benjamin/Cummings Publishing Company,1991.

[3] 代六玲,黄河燕.中文文本分类中特征抽取方法的比较研究[J].中文信息学报,18(1).

[4] Salton G,Buckley B.Term-weighting approaches in automatic text retrieval[J].Information Processing and Management,1998,24(5):513-523.

[5] Mitchell T.Machine learning[M].[S.l.]:McCraw Hill,1996.

[6] Van Rijsbergen C J.Information retrieval [M].2nd ed.London:Butterworths,1979.

[7] 宋枫溪,郑如冰.自动文本分类中两中文本表示方式的比较[J].计算机工程,30(18):124-126.

[8] 张文进.文本信息检索中的概率模型[J].情报杂志,2005(3).

[9] 鲁松,李晓黎,白硕,等.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.

[10] 张玉芳,彭时名,吕佳.基于文本分类 TFIDF 方法的改进与应用[J].计算机工程,2006,32(19):76-78.