基于信息增益的特征词权重调整算法研究

张玉芳,陈小莉,熊忠阳
ZHANG Yu-fang,CHEN Xiao-li,XIONG Zhong-yang

重庆大学 计算机学院,重庆 400030
College of Computer,Chongqing University,Chongqing 400030,China
E-mail:snailisme@21cn.com

ZHANG Yu-fang,CHEN Xiao-li,XIONG Zhong-yang.Improved approach to weighting terms using information gain. Computer Engineering and Applications,2007,43(35):159-161.

Abstract: This paper applies the information gain to remedy the defect of TFIDF neglecting the proportion of distribution of terms in categories of the text collection.The information gain of terms as one factor for term-weighting can effectively weight the proportion of distribution of terms.

Key words: information gain;shannon entropy;distribution of terms;text classification

摘 要:传统权重公式 *TFIDF* 忽略了词语在集合中的分布比例,针对 *TFIDF* 的这个缺点,把信息增益公式引入文本集合中并提出 *IF*IDF*IG*,取得了较好的效果。在分析中发现单纯把信息增益引入文本集合并不能完全解决词语分布对词语权重的影响。从文档类别层次上考虑,把信息论中信息增益应用到文本集合的类别层次上,提出了一种改进的权重公式 $tf*idf*IG_c$,用改进的权重公式来衡量词语在文本集合的各个类别中分布比例上的差异,进一步弥补传统公式的不足。实验对比了改进的公式 $tf*idf*IG_c$ 和 *IF*IDF*IG* 的实验效果,实验证明 $tf*idf*IG_c$ 权重公式在表现词语权重时更有效。

关键词:信息增益;信息熵;词语分布比例;文本分类
文章编号:1002-8331(2007)35-0159-03 **文献标识码**:A **中图分类号**:TP391

文本分类是文本挖掘的一个重要组成部分,在提高信息检索的速度和质量方面有显著意义。文本分类指按预定义的类别将待分类文档进行归类,分类过程中的特征选择和特征提取是文本分类的首要任务和关键问题。

文本数据的半结构化甚至于无结构化的特点,使得表示文本数据的特征向量高达几万维甚至于几十万维。即使经过初始化筛选处理(使用停用词表、稀有词处理、单词归并),还会有很多高维数的特征向量留下。不是所有的高维特征对分类学习都是有用的,高维的特性还会增加机器学习的时间。因此,在进行文本分类中,特征选择就显得至关重要。

特征选择主要用于排除确定的特征空间中那些被认为无关的或是关联性不大的特性。在研究文本分类的过程中,特征提取是关键的环节之一,具有降低向量空间维数、简化计算、防止过分拟合以及去除噪声等作用,特征选择的好坏将直接影响着文本分类的准确率。

1 词语权重公式的改进

1.1 传统 tfidf 公式

目前最常用的文本特征描述方法是:加权关键词矢量的向量空间模型(VSM)。向量空间模型是由 Salton G 等人在 20 世纪 60 年代提出的[8],它把文档简化为以词语的权重为分量的向量表示,把分类过程简化为空间向量的运算,使得问题的复杂性大大减小。

VSM 将文本文档视为由一组词语$(t_1,t_2,\cdots,t_n)$构成,每一词语都赋以一定的权值 w,文档被映射为由一组词语矢量组成的向量空间中的一个向量。每个文档表示为特征向量:$\boldsymbol{d}_i=\{t_1,w_1;t_2,w_2;\cdots;t_k,w_k;\cdots;t_n,w_n\}$,其中 t_k 表示词语,w_k 表示词语 t_k 在文档 $\boldsymbol{d}_i$ 中的权值。

词语的权重是用来刻画词语在描述文档内容时所起作用的重要程度,权重的计算方法按其值类型分为两种:(1)是布尔型,其计算方法是将所有训练文档的词语作为全集,如果一个词语 t_k,在文档中出现就设其权值为 1,否则设为 0;(2)是实数型:将文档的词语按权重计算公式计算其权重。

通常,在计算词语权重所采用的特征词权重计算公式为 *tfidf* 公式:

$$Weight_{tfidf(t)}=tf(t)*idf(t) \quad (1)$$

词语频率 *tf*(Term Frequency):该词语在文档中出现的次数;反文档频率 *idf*(Inverse Document Frequency):该词语在文档集合众分布情况的量化。*idf* 常用的计算方法为:

$$idf(t)=\mathrm{lb}\frac{N}{n} \quad (2)$$

其中 N 为文档集中的总文档数,n 为出现特征项 t 的文档数。

1.2 tfidf 的不足

tfidf 主要从词语的频率 *tf* 和词语的逆文档频率 *idf* 两个方

基金项目:重庆市科委自然科学基金(No.CSTC2006BB2021)。

作者简介:张玉芳(1967-),女,硕士生导师,主要研究领域为数据挖掘、数据库、并行计算、网络信息处理;陈小莉(1978-),女,硕士研究生,主要研究领域为数据挖掘;熊忠阳(1964-),男,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。

面进行考虑:如果某个词或短语,在一个文档中出现的频率 *tf* 高,并且在其它文档中很少出现,则认为此词或者短语具有很好的类别区分能力,适合用来进行分类;*idf* 的主要思想是:如果包含词语 t 的文档越少,也就是 n 越小,*idf* 越大,则说明词语 t 具有很好的类别区分能力。

进一步分析:如果某一类 C_i 中包含词语 t 的文档数为 m,而其它类包含 t 的文档总数为 k,显然所有包含 t 的文档数 $n=m+k$,当 m 大的时候,n 也大,按照 *idf* 公式得到的 *idf* 的值会小,则表示该词语 t 类别区分能力不强。但是实际上,m 大,说明词语 t 在 C_i 类的文档中频繁出现,就说明词语 t 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。另一方面,虽然包含 t 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。*tfidf* 公式存在上述这些缺点的原因主要是因为:*tfidf* 公式是将文档集合作为整体来考虑的,特别是其中 *idf* 的计算,并没有考虑到词语在各个类别中的分布情况。

1.3 引入信息增益

1.3.1 熵和信息增益的定义[6]

熵[6]是德国物理学家克劳修斯 1850 年提出的,表示一种能量在空间中分布的均匀程度,能量分布得越均匀,熵就越大。1948 年,Shannon 把熵应用于信息处理,提出了“信息熵”的概念。

信息熵(又称 Shnnon 熵)在随机事件发生之前,它是结果不确定性的量度;在随机事件发生之后,它是人们从该事件中所得到信息的量度(信息量)[6]。

信息论量度信息的基本出发点,是把获得的信息看作用以消除不确定性的东西,因此信息数量的大小,可以用被消除的不确定性的多少来表示。设随机事件 X 在获得信息 y 之前结果的不确定性为 $H(X)$,得到信息 y 之后为 $H(X/y)$,那么包含在消息 y 中的关于事件 X 的信息量:$I(X,y)=H(X)-H(X/y)$。

如信息概率空间为 $X=\{x_1:p_1,x_2:p_2,\cdots,x_n:p_n\}$其不确定程度可以表示为[6]:

$$H(X)=H(p_1,p_2,\cdots,p_n)=-\sum_{i=1}^{n}p_i*\mathrm{lb}\,p_i \tag{3}$$

$E(X)=H(X)$表示对 X 的不确定程度。

条件熵 $E(X/y)=H(X/y)$是观测信息 y 后信息空间 X 的不确定程度。

信息增益是信息熵的差,表示为:

$$IG(X,y)=H(X)-H(X/y) \tag{4}$$

$H(X)$表示在观测信息 y 前,系统的熵。就文本分类系统来讲,$H(X)$表示一个随机文档落入某个类的概率空间的熵,也就是对分类的不确定程度,即类别集合 X 所能提供的信息量。

$H(X/y)$表示观察到 y 后,文档落入某个类的概率的空间熵,即观察到 y 后对分类的不确定程度。这种不确定程度减少的量就是信息增益,即表示词语 y 对分类的作用,词语 y 所能提供的分类信息量。

1.3.2 使用信息增益调整词语权重

文献[2]从信息论的角度出发,把文档集合看作一个符合某种规律分布的信息源,依靠训练数据集合的信息熵和文档中词语的条件熵之间信息量的增益关系确定该词语在分本分类中所能提供的信息量,即词语在分类中的重要程度,并把这种重要程度反映到词语的权重中,提出了 *tf*idf*IG* 公式,提高了传统 *tfidf* 的效果。

分析发现这种单纯的把信息增益引入整个文档集合,并不能解决在 1.2 中分析的 *tfidf* 的不足。在 *tf*idf*IG* 公式的基础进一步分析,根据信息增益的定义,把信息增益公式引入到文档集合的类别间,即把文档集合看作一个符合某种规律分布的信息源,依靠训练数据集合的类别信息熵和文档类别中词语的条件熵之间信息量的增益关系来确定该词语在文本分类中所能提供的信息量,并把这个信息量反映到词语的权重中。公式如下:

$$IG(C,t)=E(C)-E(E/t) \tag{5}$$

其中 $E(C)$为:

$$E(C)=-\sum_{i=1}^{m}p(C_i)*\mathrm{lb}(p(C_i)) \tag{6}$$

$E(C/t)$为:

$$E(C/t)=-\sum_{i=1}^{m}p(C_i/t)*\mathrm{lb}(p(C_i/t)) \tag{7}$$

C 为文档的类别集合;$P(C_i)$表示类别 C_i 的概率,有多种计算方法,可基于文档统计来计算,也可以基于词频来计算;(C_i/t)表示词语 t 在类别 C_i 中出现的概率;

这样词语的权重公式修改成如下形式(把改进后的公式简称为 $tf*idf*IG_c$):

$$Weight_{tfidf}(t_i)=TF*IDF*IG(C,t_i) \tag{8}$$

当词语 t 在文档集合的类别中分布不均匀时,即在某个类别中分布较高,其他类别中分布较少,词语带有较大的类别信息,应用信息增益公式计算可得到较高的信息增益值,用公式(8)计算所的的权重值也会较高,从而提高词语 t 的权重;另一方面当词语 t 在文档集合中的数量虽小,但是如果其均匀分布在各个类别间,则其带有的类别信息少,对系统得不确定性程度影响小,则由信息增益公式计算得到的信息增益值较小,用公式(8)计算词语 t 的权重也相对较低。

因而改进的权重公式能很好的反映词语在类别间的分布情况。

下面通过一个简单的例子来说明上面的问题,假设有三个类别,每个类别各 5 篇文档,只考虑 3 个特征项 t_1、t_2、t_3(如表 1 所示)。

表 1 文档词语频率表

特征项	文档														
	类 1					类 2					类 3				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
t_1	5	5	5	0	0	0	0	0	0	0	0	0	0	0	0
t_2	16	0	0	0	0	0	0	16	0	0	0	16	0	0	0
t_3	0	0	0	0	0	0	8	0	0	0	0	12	0	0	0

从表 1 中可以看出:t_1 只是在类 1 中的文档出现,分类能力应该最强,而 t_2 在三个类中都出现了,对分类基本没有信息贡献,因此其分类能力应该较弱。但从表 2 的结果来看,t_1 对类 1 的权重为 10.485,t_2 的权重值为 11.183 5 非常高。这是因为根据 tfidf 算法的定义,特征项的权重由 *tf* 和 *idf* 决定。当文档集中包含特征项 t_1、t_2 的文档数相同时,这些特征项的 *idf* 相同,特征项的权重由 tf 唯一确定所以导致表 2 得到一个极不合理的结果,几乎没有分类能力的被赋予了很高的权值。由此可见,在没有考虑到词语在各个类别间分布的比例时,单纯使用 *tfidf* 算法会导致很大误差。

表 2　TFIDF 的权重结果

特征项	文档		
	类$_1$	类$_2$	类$_3$
t_1	10.485	0	0
t_2	11.183 5	11.183 5	11.183 5
t_3	0	7.0	10.5

下面用改进的权重公式 $tf*idf*IG_c$ 来计算词语的权重，同时和文献[2]提出的公式比较如表 3、表 4、表 5 所示：

表 3　IGc 信息增益计算表

$H(C)$	0.477		
$H(C/t_1)$	0	$IG(C,t_1)$	0.477
$H(C/t_2)$	0.477	$IG(C,t_2)$	0
$H(C/t_3)$	0.276 4	$IG(C,t_3)$	0.200 3

表 4　IG[2]信息增益计算表

$H(D)$	0.752 57		
$H(D/t_1)$	0.477	$IG(D,t_1)$	0.275 57
$H(D/t_2)$	0.477	$IG(D,t_2)$	0.275 57
$H(D/t_3)$	0.276 4	$IG(D,t_3)$	0.476 17

表 5　$tf*idf*IG$[2]和 $tf*idf*IG_C$ 权重结果比较

权重公式		特征项		
		t_1	t_2	t_3
类$_1$	$tf*idf*IG_c$	5.001 35	0	0
	$tf*idf*IG$	2.889 35	3.081 8	
类$_2$	$tf*idf*IG_c$	0	0	1.402 100
	$tf*idf*IG$	0	3.081 8	3.333 190
类$_3$	$tf*idf*IG_c$	0	0	2.103 500
	$tf*idf*IG$	0	3.081 8	4.999 785

表 3 是本文提出的把信息增益公式引入到文档类别集合中，词语的信息增益值；表 4 是文献[2]提出的把信息增益公式引入到整个文本集合，不考虑层次类别的信息增益值。

从表 3 可以看到在类 1 中分布较高，其他类别中分布较低的词语 t_1 信息量较高，而在类别中均匀分布的词语 t_2 不带有分类信息，是符合上述推理的。而从表 4 可以看到词语 t_1 和 t_2 带有相等量的分量信息，这是和本文的分析不一致的。

因此，可以得出：由改进后的 $tf*idf*IG_c$ 公式计算的词语权重更符合词语本身带有的分类信息的规律的。

2　实验与分析

实验分别采用文献[2]提出的 $tf*idf*IG$ 和本文提出的 $tf*idf*IG_c$ 公式来计算词语的权重，采用 *KNN* 算法作为分类算法，对两种权重计算方法的分类效果进行了比较。

在实验中所使用的实验数据的训练语料由人工标注的 1 882 篇文档，分 10 个类别；测试语料共 934 篇文档，实验结果通过召回率(recall)和正确率(precision)两个指标加以衡量和比较。

经过多次实验比较发现 *KNN* 分类算法的 K 取 14 和 15 时改进的公式 $tf*idf*IG_c$ 分类效果最好，原 $tf*idf*IG$ 公式在 K 取 14 时分类效果最好，在表 6 中是 K=14 时两个公式的分类结果的比较如表 6 所示：

从表 6 中可以看到改进后的 $tf*idf*IG_c$ 只是在对少数某些类别进行分类时召回率和准确率比 $tf*idf*IG$ 稍差(计算机和政治)，但从总体效果来看 $tf*idf*IG_c$ 公式无论在召回率还是在正确率上都要好于 $tf*idf*IG$。

表 6　词语权重公式 $tf*idf*IG$ 与 $tf*idf*IG_c$ 结果比较表(R：召回率 P：正确率)

权重公式	评价指标/%	交通	体育	军事	医药	政治	教育	环境	经济	艺术	计算机
$tf*idf*IG$	R	92.96	97.99	80.72	88.24	94.01	89.04	79.10	94.44	90.24	87.88
	P	91.67	92.99	94.37	96.78	93.51	92.86	91.38	86.44	93.67	98.30
$tf*idf*IG_c$	R	94.37	98.67	81.93	98.76	93.41	93.15	82.09	97.22	91.46	86.36
	P	95.71	93.03	94.44	98.39	84.78	93.15	94.83	87.40	97.40	98.28

以上都是针对某一个类别，这些指标只能代表局部意义。为了在全局意义上评价分类器，就必须考虑所有的类别。有两种方法用来综合所有类别的查全率和查对率，即宏平均(Macro-Averaging)和微平均(Micro-Averaging)。表 7 就是用宏平均来评估分类器。

表 7　宏平均评估分类器

公式	评价指标/%	K=8	K=10	K=13	K=14	K=15	K=16	K=18	K=22	K=35
$tf*idf*IG$	宏平均查全率	88.70	89.03	88.82	89.46	89.27	88.95	88.85	88.28	87.96
	宏平均查准率	90.87	91.25	91.52	92.20	90.05	91.80	91.83	91.76	91.63
$tf*idf*IG_c$	宏平均查全率	88.18	89.55	90.53	90.69	90.74	90.15	89.20	88.48	88.11
	宏平均查准率	90.76	91.94	93.02	93.39	93.44	92.98	92.43	92.41	92.17

从表 7 的实验结果来看，除了 k 取 8 时，在 k 取不同的值时，采用 $tf*idf*IG_c$ 权重公式的宏平均查全率和查准率都比原来的 $tf*idf*IG$ 权重公式效果好。

3　结束语

本文从信息论的观点出发，在词语权重公式中引入信息增益的概念，提出一种基于信息增益的词语权重公式 $tf*idf*IG_c$，该公式考虑了词语在文档集合的各个类别中的分布对词语权重的影响，从而提高了文本的分类效果。在下一步的工作中，我们将进一步完善 $tf*idf*IG_c$ 公式，把它应用到文档集合的各个类别内部，提高在某个类别内部分布均匀的词语的权重，使文本分类效果进一步提高。(收稿日期：2007 年 6 月)

参考文献：

[1] Jiawei Han Micheline Kamber 范明，孟小峰，译.Data Mining Concepts and Techniques[M].机械工业出版社，2001.

[2] 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).

[3] Yang Y，Pedersen J O.A comparative study on feature selection in text categorization[C]//Proceedings of the Fourteenth International Conference on Machine Learning，ICML97 Morgan Kaufmann Publishers Inc，San Francisco，CA，USA.

[4] 李文斌，刘椿年，陈嶷英.基于特征信息增益权重的文本分类算法[J].北京工业大学学报，2006，32(5).

[5] 陆玉昌，鲁明羽，李凡，等.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.

[6] 周荫清.信息理论基础[M].北京航空航天大学出版社，1993.

[7] 刘立柱.概率与模糊信息论及其应用[M].北京：国防工业出版社，2004.

[8] 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)：125-127.

[9] 秦进，陈笑蓉.文本分类中的特征抽取[J].计算机应用，2003，23(2).