不可靠语料库的提纯及词权度量指标 IDF 的改进*

徐 山[1],杜卫锋[2]
(1.南京城市职业学院 教务处,江苏 南京 210038;
2.嘉兴学院 数理与信息工程学院,浙江 嘉兴 314001)

摘 要:不良短信的泛滥严重影响了社会风气,干扰了人们正常的生活秩序,研发不良短信过滤技术具有相当高的实用价值。研究了文本分类中的两个问题,可应用于不良短信过滤。其一是应用聚类方法进行不可靠语料集的提纯,实验表明,该方法对不可靠数据的提纯效果比较明显;其二是关于 IDF 词权度量指标的一点改进。

关键词:短信过滤;不可靠语料库;向量空间模型;IDF;聚类
中图分类号:TP181 **文献标识码**:A **文章编号**:1674-7720(2013)04-0061-03

The purification of unreliable corpus and the improvement of word weight index IDF

Xu Shan[1], Du Weifeng[2]
(1.Dean's Office, Nanjing City Vocational College, Nanjing 210038, China;
2.School of Mathematics, Physics and Information Engineering, Jiaxing University, Jiaxing 314001, China)

Abstract: The spread of bad message seriously affects the social ethos and disrupt the normal life order of people. It has considerable practical value to research and develop the filtering technology of bad short message. Two problems in text classification are studied in this paper, which can be used in the bad short message filtering. The first is the application of clustering method to purify unreliable corpus. Experiment shows that the method is quite obvious on purification effect of unreliable data; The second is about a little improvement of word weight index IDF.

Key words: message filtration; unreliable corpus; vector space model; IDF; clustering

随着我国移动通信业务的发展,短信业务因价格便宜、方便快捷,赢得了广大用户的青睐,短信开始被人们称为继报纸、广播、电视、互联网之后的"第五媒体"。然而伴随"拇指经济"爆发性增长的同时,无孔不入的不良短信骚扰让人不堪忍受。不良短信的泛滥,不仅影响我国的正常通信秩序,而且毒化社会风气,不利于社会发展进步。

扼制不良短信的传播,既需要监管部门制定相应的法律法规,也需要采取一定的技术对不良短信进行识别和过滤。本文研究了文本分类中的两个问题,可应用于不良短信过滤,分别是不可靠语料集的提纯和关于 TFIDF 词权度量指标的一点改进。

*基金项目:国家自然科学基金(61070213)

1 理论基础

1.1 类间相似度

设两类为 C、D,分别有 m、n 个样本,即:

$C=\{c_1, c_2, \cdots, c_m\}$

$D=\{d_1, d_2, \cdots, d_n\}$

类间相似度为[1]:

$$\mathrm{sim}(C, D)=\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{sim}(c_i, d_j) \tag{1}$$

其中,$\mathrm{sim}(c_i, d_j)$为向量 c_i、d_j 间的相似度,可以用两向量间的夹角余弦表示:

$$\mathrm{sim}(c_i, d_j)=\frac{c_i d_j}{|c_i||d_j|} \tag{2}$$

1.2 训练集和测试集的划分

对于有监督学习,将语料库分为训练集和测试集。训练集用于算法的学习,测试集用于评估算法的有效性。

对于训练集和测试集的划分，目前主要有两种方法[2]：保持(Holdout)方法和 k 折交叉验证(K-fold Cross Validation)方法。保持方法将已知数据随机划分为训练数据和测试数据两部分，一般训练数据占 2/3，测试数据占 1/3。使用训练数据导出分类模型，它在测试数据上的分类精度作为最终的分类精度。k 折交叉验证则将已知数据随机划分为 k 个大致相等的数据子集 S_1、S_2、…、S_k，训练和测试重复进行 k 次。在第 i 次过程中，S_i 作为测试数据，其余的子集则作为训练数据。最终分类器的分类精度取 k 次测试分类精度的平均值。这种方法适用于原始数据量较小的情况，这时不适合直接应用保持方法。

1.3 封闭测试和开放测试

用分类器对测试集进行分类，得到测试集的分类结果，从而可以对测试集的性能做出评价。测试有封闭测试和开放测试之分。封闭测试时，测试集是训练集的一部分，或者就是训练集本身；开放测试时，测试集是与训练集独立同分布的两个数据集[3]。一般来说，封闭测试的结果意义不大，分类中主要应用开放测试。

1.4 TFIDF

度量词权的加权体系中用某一权重值取代表示该词是否出现的布尔表示，通常具有更高的准确性。为了处理同样是高频词的专业词和通用词，挖掘领域一般采用词频反转文档频率 TFIDF(Term Frequency Inverse Document Frequency)这一指标衡量某个词的权重，公式如下：

$$\begin{cases}\text{TFIDF}=\text{TF}\times\text{IDF}\\ \text{IDF}=\lg\left(\dfrac{N}{n}\right)\end{cases}\tag{3}$$

其中 TF 表示词条在文本中的权重，N 表示训练集总文本数，n 表示包含词条的文本数。对式(3)，有如下直观的解释：

(1)词在文本中出现的次数越多，则该词对文本内容越具代表性，其权值越大；

(2)词所出现的文本数越多，则该词区分文本类别属性的能力越低，其权值越小。

2 不可靠语料库的提纯

2.1 相似短信的删除

从网上搜集的语料来自不同的网站，其中有大量短信是相同或相似的，这些相似短信并不会给分类器增加信息，反而会干扰分类器的学习，增加分类器的空间和时间复杂度。本文提出了一种删除相似短信的方法。

如果采用文件比较的方法，只能比对出完全相同的短信。而通过对语料库的分析，发现其中有很多短信只是在措辞方面略有不同，而表述的内容几乎一致。

采用向量空间模型将短信向量化，应用式(2)给出的夹角余弦计算两个向量之间的相似度，设定某个阈值，如果相似度超过该阈值，则删除其中一个向量对应的短信。通过实验，发现阈值设置为 0.95 比较合适，可以基本删除相似短信，而不会造成误删。经过该步骤，总共删除了 330 条相似短信。

2.2 降低误分类短信的影响

另一种情况是短信分类错误。例如，有些短信明明是正常短信，却被错误地划分到了黄色短信的类中。但是，要对语料库中的所有短信逐一阅读，人工确定其类别，又是一项耗时耗力的工作。

如何对不纯的语料库进行提纯，本文提出了一种方法。假设最初语料库分为 n 类 C_1、C_2、…、C_n，应用聚类方法，将语料库聚合为 n 类。但是，聚类方法产生的类并没有类别标签，为找到聚类产生的类与原先分类之间的对应关系，假设聚类产生的类为 D_1、D_2、…、D_n，应用类间相似度确定聚类 D_i 与最初的哪个分类相对应，如式(4)所示：

$$J=\arg\max_j \text{sim}(D_i,C_j)\tag{4}$$

则聚类 D_i 与最初的分类 C_j 相对应。

经过实验，得到类间相似度结果如表 1 所示。

表 1 类间相似度

	C_1	C_2	C_3
D_1	0.218 2	0.023 1	0.017 8
D_2	0.016 4	0.252 0	0.021 9
D_3	0.014 5	0.017 4	0.268 4

其中，C_1、C_2、C_3 分别代表正常短信、黄色短信、反动短信。由表 1 可以得出，聚类 D_1、D_2、D_3 分别对应原先的分类 C_1、C_2、C_3。因此，实际上对语料库进行了两次划分，分别是：$\pi_1=\{C_1,C_2,C_3\}$，$\pi_2=\{D_1,D_2,D_3\}$。如图 1 所示。其中，细线表示划分 π_1，粗线表示划分 π_2。

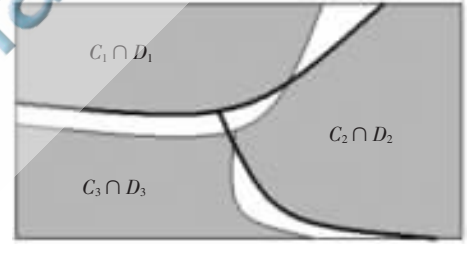

图 1 网站粗分类、聚类对语料库的两次划分及其交集

本文提纯方法如下，如果某条短信在原先分类中被归为一类，在聚类中还归为该类，则保留该文档，否则弃用。如图 1 所示，灰色区域对应的就是语料库中保留的部分，它们是分类和相应的聚类的交集 $C_i\cap D_i(i=1,2,3)$。

为验证此提纯方法的效果，对提纯前后的语料库分别进行了封闭测试和开放测试。由于提纯后语料库中只剩下不足 2 000 条短信，数据量较少，所以在测试时应用了 10 折交叉验证。

封闭测试下实验所得的查准率、查全率和微平均指标如表 2 所示。

开放测试下实验所得的查准率、查全率和微平均指标如表 3 所示。

由表 2、表 3 可见，本文方法对不可靠数据的提纯效果还是比较明显的。

表 2 封闭测试准确率对比

类别	提纯前			提纯后		
	查准率/%	查全率/%	微平均/%	查准率/%	查全率/%	微平均/%
正常短信	87.9	90.3		97.6	93.2	
黄色短信	90.2	81.3	88.58	95.5	94.4	95.46
反动短信	87.8	95.5		92.6	99.5	

表 3 开放测试准确率对比

类别	提纯前			提纯后		
	查准率/%	查全率/%	微平均/%	查准率/%	查全率/%	微平均/%
正常短信	87.5	83.1		91.4	84.6	
黄色短信	83.2	60.4	79.18	83.7	68.8	82.59
反动短信	65.9	95.9		68.2	96.6	

3 改进的 IDF

IDF 的主要思想是：如果包含某词条的文本越多，即 n 越大，IDF 越小，则说明该词条的类别区分能力越低。如果某一类 C 中包含该词条的文本数为 m，而其他类包含该词条的文本总数为 k，显然所有包含该词条的文本数 $n=m+k$，当 m 大时，n 也大，由式(3)得到的 IDF 的值会小，说明该词条类别区分能力不强。但实际上，如果一个词条在一个类的文本中频繁出现，则说明该词条能够很好地代表这个类的文本特征，这样的词条应该赋予较高的权重，并选来作为该类文本的特征词以区别于其他类文档。这就是 IDF 的不足之处。

针对 IDF 的不足，参考文献[4]提出了改进意见，设总文本数为 N，包含某词条的文本数为 n，其中某一类 C 中包含该词条的文本数为 m，则该词条在 C 类中的 IDF 表示如下：

$$\mathrm{IDF}=\lg\left(\frac{N}{n}\times m\right) \tag{5}$$

如果除 C 类外，包含该词条的文本数为 k，则式(5)可变形为：

$$\mathrm{IDF}=\lg\left(\frac{N}{m+k}\times m\right) \tag{6}$$

参考文献[4]还证明了该公式的性质：(1)IDF 是关于 m 的严格单调增函数；(2)IDF 是关于 k 的严格单调减函数。

上述性质实际上表达了如下含义：如果在某一类中包含某词条的文本数量大，而在其他类中包含该词条的文本数量小，则该词条能够代表 C 类的文本特征，具有很好的类别区分能力。

但是在某些情况下，某词条只在一个类中出现，即 $k=0$，则：

$$\mathrm{IDF}=\lg\left(\frac{N}{m+k}\times m\right)=\lg N$$

则不管在该类中包含该词条的文本数 m 为多少，值均为 $\lg N$，这与事实相违背，应该是该类中包含的文本数 m 越大，该值就越重要。因此本文对参考文献[4]的公式作了如下改进：

$$\mathrm{IDF}=\lg\left(\frac{N}{k+1}\times m\right) \tag{7}$$

IDF 的值和 m、k 的关系如下：

令：$f(m)=\dfrac{N}{k+1}\times m$，$m_1>m_2$

$$f(m_1)-f(m_2)=\frac{N}{k+1}\times m_1-\frac{N}{k+1}\times m_2$$

$$=\frac{N(m_1-m_2)}{k+1}$$

因为 $m_1>m_2>0$，$k\geqslant 0$，$N>0$，所以 $f(m_1)-f(m_2)>0$。改进后的 IDF 仍是关于 m 的严格单调增函数。

令：$f(k)=\dfrac{N}{k+1}\times m$，$k_1>k_2$

则：

$$f(k_1)-f(k_2)=\frac{N}{k_1+1}\times m-\frac{N}{k_2+1}\times m=\frac{Nm(k_2-k_1)}{(k_1+1)(k_2+1)}$$

因为 $k_1>k_2\geqslant 0$，$m\geqslant 0$，$N>0$，所以 $f(k_1)-f(k_2)<0$。改进后的 IDF 仍是关于 k 的严格单调减函数。

因此改进后的 IDF 仍能保持原来的两条性质，则该词条能够代表 C 类的文本特征，具有很好的类别区分能力。

4 结果分析与评估

用基于 KNN 的分类方法进行测试，得到了对 IDF 进行改进前后的数据，计算所得的查全率和查准率指标如表 4 所示。

表 4 对 IDF 改进前后准确率对比

类别	提纯前			提纯后		
	查准率/%	查全率/%	微平均/%	查准率/%	查全率/%	微平均/%
正常短信	91.4	84.6		92.3	86.2	
黄色短信	83.7	68.8	82.59	82.4	70.8	83.28
反动短信	68.2	96.6		71.7	95.3	

从表 4 可以看出，对 IDF 进行改进后，在查准率、查全率和微平均等指标上有了微小的提高。经过分析，在特征词限定为 1 000 个时，满足上面条件 $k=0$ 的特征词只有 8 个，占所有特征词的比例很小，因此效果不是太显著。

鉴于不良短信的泛滥和造成的重大危害，本文探讨了能应用于不良短信识别和过滤的文本分类中的两个技术问题。结合聚类分析方法，实现了对不可靠语料库的提纯，实验结果表明该方法是相当有效的。另外，对信息检索领域应用十分广泛的词权度量指标 TFIDF 的 IDF 提出了一点合理的改进，如果仅在一类中出现的特征词的比例稍大，该改进将会有一定的效果。

参考文献

[1] 李弼程，邵美珍，黄洁.模式识别原理与应用[M].西安：西安电子科技大学出版社，2008.

[2] HAN J，KAMBER M.Data Mining：Concepts and Techniques[M].北京：高等教育出版社，2001.

[3] 李家兵.交叉覆盖算法下文本分类的研究[D].合肥:安徽大学,2007.

[4] 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.

(收稿日期:2012-10-22)

作者简介:

徐山,男,1977 年生,硕士研究生,讲师,网络技师,主要研究方向:智能信息处理。

杜卫锋,男,1977 年生,博士,主要研究方向:粗糙集理论与应用,数据挖掘,智能信息处理。