2010年8月 西北工业大学学报 Aug. 2010
第28卷第4期 Journal of Northwestern Polytechnical University Vol. 28 No. 4

基于改进朴素贝叶斯算法的垃圾邮件过滤器的研究

郑 炜[1]，沈 文[1]，张英鹏[2]
（1. 西北工业大学 软件与微电子学院，陕西 西安 710072；2. 西安财经学院 信息学院，陕西 西安 710072）

摘 要：基于朴素贝叶斯算法的垃圾邮件过滤器是目前比较高效、经济的垃圾邮件过滤技术之一，它已经广泛应用到垃圾邮件过滤领域。文章在对朴素贝叶斯过滤器分析的基础上，针对朴素贝叶斯算法的缺陷结合损失最小化的思想，并根据垃圾邮件的特性对朴素贝叶斯算法做了改进，提出了改进朴素贝叶斯算法，该算法能够通过调整 k 值，降低合法邮件被错判为垃圾邮件的概率，从而最大程度减少用户的损失。

关 键 词：概率，朴素贝叶斯，垃圾邮件过滤器
中图分类号：TP311 **文献标识码**：A **文章编号**：1000-2758(2010)04-0622-06

随着企业信息爆炸性增长，电子邮件已成为人们生活中较为普及的通信手段，为了能带给人们更多的方便，邮件系统的安全性和可靠性就成为了大家关注的焦点，尤其是垃圾邮件日趋泛滥的问题更值得我们去妥善地处理和解决。据中国互联网协会反垃圾邮件中心发布的《2009年第一季度中国反垃圾邮件状况调查报告》显示，中国垃圾邮件的总量依然在继续攀升，中国网民平均每周收到垃圾邮件的数量为17.68封，垃圾邮件每年给中国网民造成的经济损失达339.59亿元人民币。因此，研究有效的垃圾邮件过滤器有很重要的现实意义。

目前使用较多的垃圾邮件过滤技术有：黑名单、身份认证、关键字过滤、行为识别模式和白名单。但上述技术普遍缺乏自适应的学习能力，不能够应对当前发展迅速、形式多样的垃圾邮件。我们需要一种可以自适应的技术，这种技术必须能够适应垃圾邮件制造者不断变化的策略。无论垃圾邮件形式如何变化，其最终目的都是要发送到用户手中。利用这一特点，可以从邮件内容出发研究垃圾邮件过滤问题。朴素贝叶斯[1]（Naive Bayes）算法因其既具有自适应、统计智能等功能，又满足了个性化的要求，故而在商业和开源垃圾邮件过滤系统中得到广泛应用[2]。

1 朴素贝叶斯算法

1.1 贝叶斯原理

贝叶斯算法[3]是以英国数学家 Thomas Bayes（1702－1763）命名的一种基于概率分析的可能性推理理论。1763年，他在《论有关机遇问题的求解》中发表了贝叶斯统计理论，即根据已经发生的时间来预测试卷发生的可能性。

贝叶斯理论假设[4]：如果事件的结果不确定，那么量化它的惟一方法就是事件的概率。如果过去实验中事件的出现率已知，那么根据数学方法可以计算出未来实验中事件出现的概率。贝叶斯定理可以用一个数学公式表达，即贝叶斯公式。

贝叶斯公式（Bayes Formula）：设实验 E 的样本空间为 S，A 为 E 的事件，$B_1, B_2 \cdots, B_n$ 为 S 的一个划分，且 $P(A_i) > 0, P(B_i) > 0 (i = 1,2,\cdots,n)$，则

$$P(B_i \mid A) = \frac{P(B_i)P(A \mid B_i)}{P(A)} = \frac{P(B_i)P(A \mid B_i)}{\sum_{i=1}^{n} P(B_i)P(A \mid B_i)} \quad (i = 1,2,\cdots,n) \tag{1}$$

贝叶斯概率是通过先验知识和统计现有数据，使用概率的方法对某一事件未来可能发生的概率进行估计。

收稿日期：2009-06-21
作者简介：郑 炜（1975－），西北工业大学讲师，主要从事软件工程及软件测试的研究。

1.2　朴素贝叶斯算法及缺陷分析

1.2.1　基本原理

朴素贝叶斯算法(Naive Bayes,NB 算法)是目前公认的一种简单而且有效的概率分类方法[6]。它是在一般贝叶斯算法的基础上通过假定各因素之间不存在任何联系,即完全独立而得到的一种简化贝叶斯算法。朴素贝叶斯分类器是垃圾邮件内容过滤中广泛应用的分类方法。利用这种方法,可以根据训练集自动训练,训练的结果反映了训练集的性质。因此邮件用户可以提供一定数量的垃圾邮件和非垃圾邮件,训练自己的过滤器,从而反映了用户自己的个性需求。

贝叶斯分类器是一类常用的分类器,最基本的形式是朴素贝叶斯(也称为简单贝叶斯)分类器。其原理是计算文本 d_x 属于某个类别的概率 $P(c_j \mid d_x)$,将文本分到概率最大的类别中去。计算 $P(c_j \mid d_x)$ 时,利用了贝叶斯公式

$$P(c_j \mid d_x) = \frac{P(c_j)P(d_x \mid c_j)}{P(d_x)}, j = 1,2,\cdots,|C| \tag{2}$$

式中,根据全概率公式,有

$$P(d_x) = \sum_{j=1}^{|C|} P(c_j)P(d_x \mid c_j) \tag{3}$$

c_j 类的先验概率可以由训练集合很容易估计

$$P(c_j) = \frac{\text{训练集中属于 } c_j \text{ 类的文本数量}}{\text{训练集中的文本总量}} \tag{4}$$

文本的类条件概率 $P(d_x \mid c_j)$ 可以由文本中出现的特征类条件概率求得。

1.2.2　朴素贝叶斯算法的缺陷分析

根据 1.2.1 节到朴素贝叶斯文本分类算法

$$P(c_j \mid d_x) = \frac{P(c_j)P(d_x \mid c_j)}{P(d_x)} \propto P(c_j)P(d_x \mid c_j) \tag{5}$$

$P(c_j)$ 是类的先验概率,$P(d_x \mid c_j)$ 是类条件概率。对同一篇文本,$P(d_x)$ 不变。

$$P(d_x \mid c_j) = P(t_1 \mid c_j) * P(t_2 \mid c_j) * \cdots * P(t_n \mid c_j) = \prod_{i=1}^{n} P(t_i \mid c_j) \tag{6}$$

设 d_x 表示为特征集合$(t_1,t_2,\cdots,t_n)$,n 为特征个数,假设特征之间相互独立。$P(c_j)$ 和 $P(t_i \mid c_j)$ 都可以利用训练集估计。

通过研究和分析朴素贝叶斯算法的原理以及它在邮件过滤中的应用,我们发现朴素贝叶斯算法是文本分类领域中一种性能优越的分类算法,它具有很强的理论背景,易于实现,而且它的运算速度也非常快。这种算法计算起来相对简单,具有较高的精确度[6]。所以朴素贝叶斯算法被应用到相当广泛的领域。

从中文自然语言处理开发平台、中国教育和科研计算机网应急响应组(CCERT)、中国反垃圾邮件联盟提供的中文邮件语料库以及个人平时收集的邮件中随机选取了 4 000 封邮件作为测试集,并随机抽取 500 封作为训练样本的邮件进行分类(合法邮件为 200 封,垃圾邮件为 300 封)。其中基于朴素贝叶斯邮件过滤算法的实验结果(经过多次实验得出的结果取其平均值,这样就尽量避免了实验的偶然性)如表 1 所示。

由表 1 知对 500 封邮件进行分类测试,实际共有 200 篇合法邮件和 300 篇垃圾邮件,而系统分类有 205.5 篇合法邮件和 294.5 篇垃圾邮件。其中 200 篇合法邮件中的 26.75 篇被系统误判为垃圾邮件,而 300 篇垃圾邮件中的 32.25 篇被系统误判为合法邮件。实验数据说明了朴素贝叶斯算法也存在着一定的缺陷。首先,它过多地简化使得很多对于分类很有用的信息丧失了,进而使得分类效果不很理想。它使得 26.75 篇真正的合法邮件被当作垃圾邮件过滤掉了,无法被用户阅读,这样也许对用户来说是非常重要的信息就丢失了。另一方面,32.25 篇真正的垃圾邮件被误判为合法邮件,也浪费了用户的精力给用户的工作学习带来麻烦。其次,为了简化计算,它假定各待分类文本特征量是相互独立的,即"贝叶斯假设"。实际上这种条件独立的假设在许多应用领域未必能很好满足甚至不成立。

在第 2 节中我们将针对朴素贝叶斯算法的不足,做出相应的改进,给出分类效果更好的改进朴素贝叶斯算法。

表 1　朴素贝叶斯邮件过滤算法的实验结果

	系统判别为合法邮件	系统判别为垃圾邮件	总数
实际为合法邮件	173.25	26.75	200
实际为垃圾邮件	32.25	267.75	300
总数	205.5	294.5	500

2 朴素贝叶斯算法的改进

在电子邮件的实际分类当中,有时对邮件的分类不仅要考虑到尽可能做出正确判断,而且还要考虑到做出错误判断时会带来什么后果。由于电子邮件只分为2类,即垃圾邮件和合法邮件,所以相应地会出现2种情况的分类错误:①将合法邮件分类到垃圾邮件中去;②将垃圾邮件分类到合法邮件当中去。如果把合法邮件判为垃圾邮件放到垃圾箱中可能会使用户的重要信件丢失带来严重后果;而如果本来就是垃圾邮件却判为正常,就会给用户带来许多麻烦。显然这2种不同的错误判断所造成损失的严重程度相差很大,错误地阻断一个合法邮件要比漏掉一个垃圾邮件的代价要大得多,这也就是很多用户不愿轻易使用垃圾邮件过滤设备的原因。这就类似于医生在诊断过程中,如果把身体健康的病人误诊为患病自然会给病人带来精神上的负担和折磨,而如果把原本患病的病人判断为正常,可能就会使其延误治疗的最佳时机。因此,我们需要一种可以使得损失尽量最小化的过滤算法。

利用前面章节的朴素贝叶斯概率公式可以分别计算出任一待分类邮件 d_x 属于合法邮件和垃圾邮件的概率:$P(c_1 \mid d_x)$(c_1 为垃圾邮件类)和 $P(c_0 \mid d_x)$(c_0 为合法邮件类)。在传统的方法中,一般当 $P(c_1 \mid d_x) > P(c_0 \mid d_x)$,就判定邮件 d_x 为垃圾邮件,否则就判为合法邮件。但是,这种判断方式并不精确,会产生较高的误判率和漏判率。在1.2.2节中分析了朴素贝叶斯算法存在的不足之处,它使分类过程丧失了很多有用的信息,从而很可能导致误判,造成严重的后果,所以直接应用它时分类偏差会比较大。

因此,为了能更加谨慎、准确地识别出垃圾邮件,减少由于把邮件误判而造成的损失,设当 $\frac{P(c_1 \mid d_x)}{P(c_0 \mid d_x)} > \theta$ 时,即当一封邮件 d_x 为垃圾邮件的概率是合法邮件概率的 θ 倍时,将其判定为垃圾邮件。当 θ 值越大,其为垃圾邮件的可能性就越大。

又由1.2.1节的公式得到 $\frac{P(c_1 \mid d_x)}{P(c_0 \mid d_x)} > \theta$ 可以表示为

$$\frac{P(c_1 \mid d_x)}{1 - P(c_1 \mid d_x)} > \theta \tag{7}$$

$$P(c_1 \mid d_x) > \frac{\theta}{1 + \theta} = k \tag{8}$$

也就是当 $P(c_1 \mid d_x) > k$ 时,将 d_x 判为垃圾邮件。

3 实验及评价

由公式(8)的推导得到一个 k 值,k 的取值范围在0到1之间即 $0 < k < 10$,这样就只需通过设定 k 值来约束对邮件的判断,减少了计算量。可以根据用户的要求,通过调整 k 值的大小,来采取不同的策略,从而最终获得相对满意的结果。实际应用中要想确定较为合适的 k 值需要有一定的经验和通过大量的实验,往往要根据所研究的具体问题,分析误判决策造所成损失的严重程度等等。

采用改进朴素贝叶斯算法进行邮件过滤,根据提供 k 值的不同产生不同的结果。在此,只列出所做的多组实验中当 k 分别取0.09、0.16、0.29、0.39、0.42、0.46、0.53、0.5、0.57、0.6、0.75、0.9 和 0.99 时的情况,其中当 $k = 0.5$(即 $\theta = 1$)时的实验数据就是采用朴素贝叶斯算法进行邮件过滤时的实验结果,参见表1。

每次测试过程中改变 k 值的大小,对邮件进行分类得到结果如表2所示。

表2 改进朴素贝叶斯过滤算法的实验结果

$k=0.09$			$k=0.16$		
	系统判别为合法邮件	系统判别为垃圾邮件		系统判别为合法邮件	系统判别为垃圾邮件
实际为合法邮件	173.2	26.5	实际为合法邮件	174.6	25.4
实际为垃圾邮件	0	300	实际为垃圾邮件	0	300
$k=0.29$			$k=0.39$		
	系统判别为合法邮件	系统判别为垃圾邮件		系统判别为合法邮件	系统判别为垃圾邮件
实际为合法邮件	176.5	23.5	实际为合法邮件	178.3	21.7
实际为垃圾邮件	0	300	实际为垃圾邮件	0	300

$k=0.42$			$k=0.46$		
	系统判别为合法邮件	系统判别为垃圾邮件		系统判别为合法邮件	系统判别为垃圾邮件
实际为合法邮件	182.1	17.9	实际为合法邮件	185.3	14.7
实际为垃圾邮件	3	297	实际为垃圾邮件	6.5	293.5

$k=0.53$			$k=0.57$		
	系统判别为合法邮件	系统判别为垃圾邮件		系统判别为合法邮件	系统判别为垃圾邮件
实际为合法邮件	190	10	实际为合法邮件	193.5	6.5
实际为垃圾邮件	18.2	281.8	实际为垃圾邮件	22.1	277.9

$k=0.6$			$k=0.75$		
	系统判别为合法邮件	系统判别为垃圾邮件		系统判别为合法邮件	系统判别为垃圾邮件
实际为合法邮件	200	0	实际为合法邮件	200	0
实际为垃圾邮件	24.3	275.7	实际为垃圾邮件	26.1	273.9

$k=0.9$			$k=0.99$		
	系统判别为合法邮件	系统判别为垃圾邮件		系统判别为合法邮件	系统判别为垃圾邮件
实际为合法邮件	100	0	实际为合法邮件	100	0
实际为垃圾邮件	27.3	272.7	实际为垃圾邮件	27.8	272.2

由表2可知,从语料库的测试集中选取500封邮件进行分类测试(实际共有200篇合法邮件和300篇垃圾邮件)。当k值为0.09时,有26.8篇合法邮件被系统误判为垃圾邮件,没有垃圾邮件被系统误判为合法邮件。当k的值为0.16时,有25.4篇合法邮件被系统误判为垃圾邮件,没有垃圾邮件被系统误判为合法邮件。当k值为0.29时,有23.5篇合法邮件被系统误判为垃圾邮件,没有垃圾邮件被系统误判为合法邮件。当k值为0.39时,有21.7篇合法邮件被系统误判为垃圾邮件,同样没有垃圾邮件被系统误判为合法邮件的情况。

当k值为0.42时,有17.9篇合法邮件被系统误判为垃圾邮件,有3篇垃圾邮件被系统误判为合法邮件。当k值为0.46时,有14.7篇合法邮件被系统误判为垃圾邮件,有6.5篇垃圾邮件被系统误判为合法邮件。当k值为0.53时,有10篇合法邮件被系统误判为垃圾邮件,有18.2篇垃圾邮件被系统误判为合法邮件。当k值为0.57时,有6.5篇合法邮件被系统误判为垃圾邮件,有22.1篇垃圾邮件被系统误判为合法邮件。

当k值为0.6时,没有合法邮件被系统误判为垃圾邮件,有24.3篇垃圾邮件被系统误判为合法邮件。当k值为0.75时,没有合法邮件被系统误判为垃圾邮件,有26.1篇垃圾邮件被系统误判为合法邮件。当k值为0.9时,没有合法邮件被系统误判为垃圾邮件,有27.3篇垃圾邮件被系统误判为合法邮件。当k值为0.99时,同样没有合法邮件被系统误判为垃圾邮件的情况,有27.8篇垃圾邮件被系统误判为合法邮件。

经过大量实验数据,发现当$0 < k \leqslant 0.39$时系统的召回率为100%,即能保证系统判定为合法的邮件都是真正的合法邮件,不存在把垃圾邮件误判为合法邮件的错误。当$0.6 \leqslant k < 1$时系统的正确率为100%,即保证了系统判断为垃圾的邮件都是真正的垃圾邮件,不存在把合法邮件误判为垃圾邮件的情形。当k在0.39到0.6这个范围中时,系统判定的合法邮件和垃圾邮件都没有百分之百的把握。当$0.39 < k < 0.5$时,k值越接近0.39,垃圾邮件被误判为合法邮件的数量越少;相应地,当$0.5 < k < 0.6$时,k值越接近0.6,合法邮件被系统误判为垃圾邮件的数量越少。由此可见,当$k = 0.6$时,可使合法邮件最小程度下被误判为垃圾邮件。

本节中,从测试数据集中选取3组邮件作为测试样本。第1组测试总共选取邮件500封,其中包括合法邮件300封,垃圾邮件200封;第2组测试共

选取邮件1 000封,其中合法邮件与垃圾邮件各500封;第3组测试一共选取了3 000封,其中合法邮件1 000封,垃圾邮件2 000封。

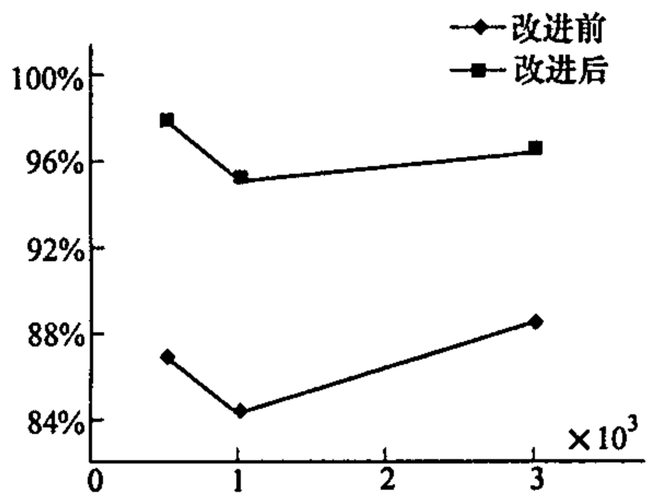

图1 召回率对比图

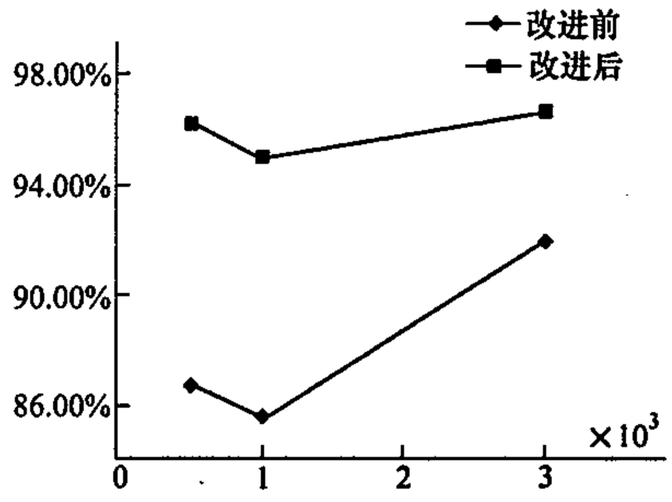

图2 精确率对比图

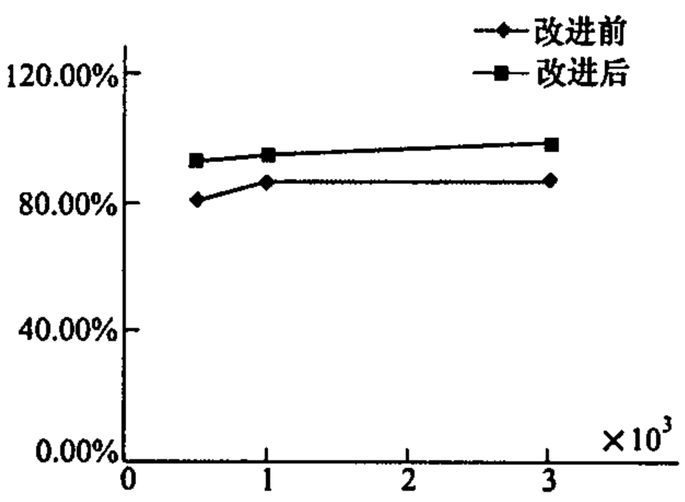

图3 正确率对比图

分别对改进前朴素贝叶斯过滤器和改进后朴素贝叶斯过滤器的召回率(Recall)、精确率(Precision)和正确率(Accuracy)进行对比,实验结果显示,对500封邮件进行分类测试,其中300篇合法邮件中的40篇被系统误判为垃圾邮件,而200篇垃圾邮件中的26篇被系统误判为合法邮件,系统召回率为87%,精确率为86.8%,正确率为81.31%。对1 000封邮件进行分类测试,其中500篇合法邮件中的66篇被系统误判为垃圾邮件,而500篇垃圾邮件中的78篇被系统误判为合法邮件,系统召回率为84.4,精确率为85.6%,正确率为86.48%。对3 000封邮件进行分类测试,其中1 000篇合法邮件中的155篇被系统误判为垃圾邮件,而2 000篇垃圾邮件中的228篇被系统误判为合法邮件,系统召回率为88.6%,精确率为91.96%,正确率为87.23%。由此可见,未改进的朴素贝叶斯过滤系统的错判率较高。

采用改进朴素贝叶斯过滤器,并设置 $k=0.6$,对500封邮件进行分类测试,其中300篇合法邮件中的15篇被系统误判为垃圾邮件,而200篇垃圾邮件中的4篇被系统误判为合法邮件,系统召回率为98%,精确率为96.2%,正确率为92.89%。对1 000封邮件进行分类测试,其中500篇合法邮件中的26篇被系统误判为垃圾邮件,而500篇垃圾邮件中的24篇被系统误判为合法邮件,系统召回率为95.2%,精确率为95%,正确率为94.82%。对3 000封邮件进行分类测试,其中1 000篇合法邮件中的33篇被系统误判为垃圾邮件,而2 000篇垃圾邮件中的68篇被系统误判为合法邮件,系统召回率为96.6%,精确率为96.63%,正确率为98.32%。由实验结果,可以看出本文的邮件过滤系统有较高的召回率、精确率和正确率,系统性能较为稳定,整体过滤效果较好,从而证明了改进算法的有效性及高效性。如图1~图3所示,改进后的邮件过滤系统的召回率、精确率比未改进前有10%左右的提升,正确率比未改进前有5%的改进。

4 结 论

基于朴素贝叶斯算法的垃圾邮件过滤器是目前比较高效、经济的垃圾邮件过滤技术之一,它已经广泛应用到垃圾邮件过滤领域。本文在对朴素贝叶斯过滤器分析的基础上,针对朴素贝叶斯算法的缺陷结合损失最小化的思想,并根据垃圾邮件的特性对朴素贝叶斯算法做了改进,提出了改进朴素贝叶斯算法,该算法能够通过调整 k 值,来降低合法邮件被错判为垃圾邮件的概率,从而最大程度上减少用户的损失。

如果用户要求过滤后的合法邮件中全部都是真正的合法邮件,不存在被误判的垃圾邮件,可选用第2节提出的改进朴素贝叶斯算法;但如果用户想要邮件过滤的整体性能和分类精度得到提高,即垃圾邮件与合法邮件的误判数量都减少,本文还没有提出更为有效的解决方案。所以使邮件过滤系统更加人性化和个性化,还需做出进一步的努力。

参考文献:

[1]Zhang H. Exploring Conditions for the Optimality of Naive Bayes. International Journal of Pattern Recognition and Artificial Intelligence, 2005, 19(2): 183 ~ 198

[2]Vangelis Metsis, Ion Androutsopoulos, Georgios Paliouras. Spam Filtering with Naive Bayes——Which Naive Bayes? CEAS 2006 Third Conference on Email and AntiSpam, 2006

[3]Mehran Sahami, Susan Dumais, David Heckerman, Eric Horvitz. A Bayesian Approach to Filtering Junk E-Mail. AAAI Workshop, Madison, Wisconsin. 1998:55 ~ 62

[4]Johan Hovold. Naive Bayes Spam Filtering Using Word-Position-Based Attributes. 2nd Conference on Email and Anti-Spam, Stanford, CA, 2005

[5]Zhang I E, Zhu Jingbao, Yao Tianshun. An Evaluation of Statistical Spam Filtering Techniques. ACM Trans on Asian Language Information Processing, 2004, 3(4): 243 ~ 269

[6]Aris Kosmopoulos, Georgios Paliouras, Ion Androutsopoulos. Adaptive Spam Filtering Using Only Naive Bayes Text Classifiers. CEAS 2008 Fifth Conference on Email and AntiSpam, 2008, Mountain View, California USA

Implementing Spam Filter by Improving Naive Bayesian Algorithm

Zheng Wei[1], Shen Wen[1], Zhang Yingpeng[2]

(1. Department of Software Engineering, Northwestern Polytechnical University, Xi'an 710072, China
2. School of Information, Xi'an University of Finance and Economics, Xi'an 710061, China)

Abstract: Our aim is to decrease the probability under which the spam filter misjudges legal e-mail as spam by adjusting the k value of the naive Bayesian algorithm, thus minimizing Internet users' economic loss. Section 1 of the full paper analyzes the classification deficiencies of the naive Bayesian algorithm. Section 2 implements the spam filter by improving the naive Bayesian algorithm through obtaining the k value as shown in eq. (8). Section 3 tested the spam filter by adjusting the k value of our improved Bayesian algorithm; the test results, presented in Table 2, and their comparison, given in Figs. 1, 2 and 3, show preliminarily that the spam filter that uses our improved Bayesian algorithm can increase the recall rate by 10% and the accuracy by 5%, thus effectively decreasing the probability of misjudging legal e-mails as spams.

Key words: algorithms, probability, naive Bayesian algorithm, spam filter