一种改进的基于朴素贝叶斯算法的垃圾短信过滤技术

陈凯星，陈建英
（西南民族大学 计算机科学与技术学院 四川 成都 610041）

【摘 要】：针对朴素贝叶斯算法文本分词中过多的简化和条件独立性假设的缺陷，提出了一种改进的基于朴素贝叶斯算法的短信过滤算法。该算法通过改进概率算法将朴素贝叶斯算法更好地适用于垃圾短信的过滤器中，并且在求得的后验概率中加入了经过统计短信长度得到的不同长度所对应类别的一定概率值，从而降低正常短信被误判的概率。

【关键字】：垃圾短信过滤；Android；朴素贝叶斯算法

0 引 言

自动高效的过滤垃圾短信是当前手机应用中必须解决问题之一。目前使用较多的垃圾短信过滤技术普遍缺乏自适应学习能力，不能应对当前发展迅速、形式多样的垃圾短信。本文在规则过滤的基础上，结合基于文本分类以及朴素贝叶斯理论的过滤方法对垃圾短信进行分类，并提出了进一步的改善措施。

1 朴素贝叶斯算法

1.1 贝叶斯定理

贝叶斯算法是以著名数学家托马斯．贝叶斯(Thomas 贝叶斯)(1702-1761) 命名的一种基于概率分析的可能性推理理论，通过分析过去事件的知识，来预测未来的事件。贝叶斯过滤法对大量用户已经判定的垃圾短信和正常短信进行学习，根据垃圾短信和正常短信中相同词语出现的概率对比来确定垃圾短信的可能性。

贝叶斯定理描述如下：

设对于实验 E 的样本空间为 S，并且 $\{B_1,B_2,\cdots B\}$ 是 S 的一个划分。令{p(A)|A S}表示定义在 S 中所有事件上的一个概率分布，则对于 S 中的任意事件 A 和 B，有 p(A)>0，$p(B|A)=\frac{p(AB)}{p(A)}$ 表示条件概率，即在已知 A 发生的情况下 B 发生的概率。贝叶斯定理可以表示为：

$$p(B_i|A)=\frac{p(A|B_i)p(B_i)}{p(A)} \quad (i=1,2,\cdots,n) \quad 公式(1-1)$$

其中 p(A)>0，由全概率公式可得

$$p(A)=\sum_{j=1}^{n}p(A|B_j)p(B_j) \quad 公式(1-2)$$

在公式(1-2)中，$p(B_i|A)$ 为后验概率，$p(A|B)$ 为似然概率，$p(B)$ 为先验概率。

1.2 朴素贝叶斯定理

朴素贝叶斯过滤技术以贝叶斯定理为基础，并假定各个属性直接是独立地来进行预测，把从训练样本中计算出的各个属性值和类别频率比作为先验概率，然后利用贝叶斯公式计算出其后验概率，并选取具有最大后验概率的类别作为预测值，在垃圾短信过滤应用中有广泛应用。

设有 m 个样本空间 $\{c_1,c_2,\cdots c_n\}$，短信 d 中有 n 个特征项 $\{w_1,w_2,\cdots w_n\}$，对于给定的类 $c_k(k=1,2,\cdots,m)$，d 属于类 c_k 的概率为

$$p(c_k|d)=Max\{p(c_1|d),p(c_2|d),...,p(c_n|d)\} \quad 公式(1-3)$$

由贝叶斯概率公式可得：

$$p(c_k|d)=\frac{p(d|c_k)p(c_k)}{p(d)} \quad (k=1,2,\cdots,m) \quad 公式(1-4)$$

其中：

$$p(d|c_k)=p(w_1,w_2,...,w_n|c_k) \quad 公式(1-5)$$

公式(1-4)中，$p(c_k)$ 为先验概率，很容易计算，但 $p(d|c_k)$ 的计算比较困难。为了简化计算，引入了条件概率独立假设，即假定各特征项之间是相互独立的，这就是朴素贝叶斯过滤器，那么公式(1-5)就可以转换为：

经费支持：2013 年四川省大学生创新训练项目

$$p(d|c_k)=p(w_1,w_2,...,w_n|c_k)=\prod_{i=1}^{n}p(w_i|c_k)$$ 公式(1-6)

$$p(c_k)=\frac{|S_k|}{|S|}$$ 公式(1-7)

其中，$|S_k|$表示类别 c_k 中的训练文本数量；$|S|$表示训练文本集总数量。

在短信训练集中只有两个样本空间，一个是正常短信，另一个是垃圾短信，所以对于样本空间 m 的值在实际应用中应为 2。

1.3 朴素贝叶斯算法应用于短信拦截的缺陷

朴素贝叶斯文本分类算法：

$$p(c_k|d)=\frac{p(d|c_k)p(c_k)}{p(d)}\quad (k=1,2,\cdots,m)$$

$p(c_k)$ 为先验概率，$p(d|c_k)$ 为类条件概率，$p(c_k|d)$ 为后验概率。对于短信训练集来说，朴素贝叶斯算法的样本空间只分为垃圾短信和非垃圾短信两部分。为了保证公平性，训练集中正常短信的文本个数和垃圾短信的文本个数是相同的，即 $p(c_{span})=p(c_{legi})$，对同一文本 $p(d)$ 不变，因此后验概率的大小主要是由类条件概率决定，但由于它过多的简化使用对于分类很有用的信息都丧失了，进而使得分类效果不是很明显，误判率难以降低。

2 朴素贝叶斯算法的改进

2.1 类条件概率

$$p(c_k|d)=\frac{p(d|c_k)p(c_k)}{p(d)}\quad (i=1,2,\cdots,n),(j=1,2)$$

$p(x_i|c_j)$ 为属性 x_i 在类别 c_j 出现的概率，其中，$N(X=x_i,C=c_j)$ 表示类别 c_j 中包含属性 x_i 的训练文本数量；$N(C=c_j)$ 表示类别中 c_j 的训练文本数量；M 值用于避免 $N(X=x_i,C=c_j)$ 过小所引发的问题；V 表示类别的总数。

2.2 长度特征提取

每条短信长度为 70 个中文字符，通过对短信长度统计，得到如下结果。

表 1 正常短信与垃圾短信长度差异比较表

短信长度	正常短信	垃圾短信
0~10	49%	0
10~20	26%	0
20~30	15%	1%
30~40	5%	5%
40~50	1%	9%
50~60	1%	31%
60~70	3%	53%
70 以上	0	1%

从上图可以看出，不同的短信长度所对应的为正常短信或垃圾短信的概率都有所不同，因此，垃圾和非垃圾短信在长度上有很明显的区别，正常短信的长度一般集中在 30 个字符以内，垃圾短信则集中在 40~70 个字符之间。对于 70 个字符以上的短信，可能是祝福类的短信或者是手机通信套餐之类的短信，这些都属于正常短信，所以放宽了它的概率值。

2.3 改进的朴素贝叶斯算法

首先利用求得的正常短信中最大的 $p(x_i|c_j)$ 并将此概率对应的词语与垃圾词库中的词语相匹配，如果匹配成功那么在得到的各个属性的总和的类条件概率的基础上再加上一定概率的值，此值必须通过多次的测验来给出。以此来扩大相应类别的概率值，避免出现大的错误率。经过调查发现不同长度的短信对应为垃圾短信或正常短信有一定的概率值，因此，当求得后验概率后还应加上相应短信长度所对应的概率来加大分类的精确度。

3 短信过滤处理过程及结果

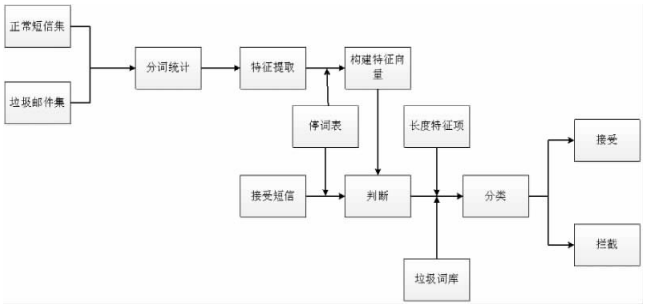

图 1 垃圾短信过滤处理流程示意图

算法改进后，短信误判率明显降低，测试短信数量及实际测试结果如表 2 所示。

表 2 过滤算法改进前后实验比对

误判率 / 短信类别	改进前	改进后	测试短信数
正常短信	33%	18%	100
垃圾短信	30.7%	9.3%	150

4 结束语

本文针对传统的朴素贝叶斯算法对垃圾短信的分类精准度与识别率不足的确定，提出了基于改进朴素贝叶斯算法的垃圾短信过滤技术，该算法通过改进概率算法将朴素贝叶斯算法更好地适用于垃圾短信过滤器，并且根据短信的长度，在求得的后验概率的基础上提高一定的概率值，降低正常短信被判为垃圾短信的概率，从而最大程度减少误判率。

(下转第 52 页)

从案例背景中可以了解到，该公司的所有员工都从仅有的两条线路上网和维护客户设备。当公司员工越来越多时，客户网络维护人员远程登录的速率也越来越慢，这很可能是员工上网的流量增多，挤占了两条线路的带宽所引起的。

2.2.3 案例解决与实施

在公司网络出口处的路由器上可以通过策略路由技术解决，在路由器的接口上实施基于应用的策略路由，让所有上网浏览网页的数据包从一条线路通过，而将另一条线路留给远程登录使用。

当我们了解了整个案例背景、需求和解决措施之后，我们可以搭建一个模拟的实验环境来对本案例的解决办法进行测试。

(1)测试实例的拓扑设计

该案例的拓扑结构可以按照图 1 进行设计。

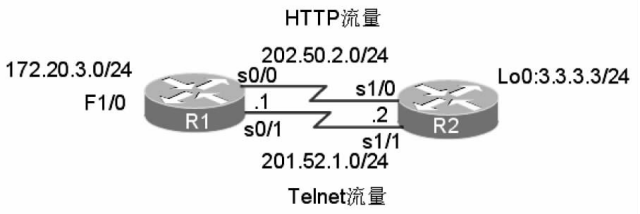

图 1 网络拓扑结构

(2)用 route-map 语句定义策略路由

在 route-map 语句中，应指定数据包的应用类型和对数据包的操作。

在路由器 R1 上进行配置，配置命令如下：

access-list 101 permit tcp any any eq 80

access-list 102 permit tcp any any eq 23

上面两条命令的作用是定义两个访问控制列表，分别对应于两种应用。

route-map lab permit 10

match ip address 101

set ip next-hop 202.50.2.2

route-map lab permit 20

match ip address 102

set ip next-hop 201.52.1.2

上面命令的作用是定义 route map 表 lab 的两条语句，第一条语句作用是网页访问应用的数据经过 s1/0 接口发送，第二条语句作用是远程登录应用的数据经过 s1/1 接口发送。

(3)在接口上应用 route map 语句

在路由器 R1 上进行如下配置：

interface f1/0

ip policy route-map lab

上面命令的作用是在接口上应用名字是 lab 的 route map 表。

(4) 使路由器本身产生的数据包也接受该策略路由的管理

在路由器 R1 的全局模式下进行如下配置：

ip local policy route-map lab

整个配置完成后，我们可以在路由器 R1 上使用 debug ip policy 命令来进行监测，从而验证实验配置的正确性。

3 结束语

本文在分析网络工程实验课程现状后，提出了基于工程案例教学的网络工程实验教学模式，通过基于策略的路由实验案例介绍了工程案例教学在网络工程实验课程的实施过程，这样可以增强学生学习实验课程的积极性，有助于学生理解所学知识在实际工程中的应用场景，达到理论知识和实际工程的紧密结合，从而真正掌握所学的网络技术。

参考文献：

[1]王燕,李华,卢慧. 网络工程课实验教学结构的改革探索[J]. 计算机教育,2013,14:29-32.

[2]石云辉. 计算机网络工程实验的设计与教学实践[J].电脑知识与技术,2010,6(14):3692-3694.

[3] 刘晓华, 郑更生, 赵卿松. 网络工程专业实践教学体系的研究[J].软件导刊.2011,10(5):197-199.

(上接第 43 页)

参考文献：

[1] 张东亮,董礼.基于改进的朴素贝叶斯算法在垃圾短信过滤中的研究[J]. 计算机测量与控制. 2012,20(02)

[2] 丁岳伟,潘涛.利用贝叶斯算法过滤报文内容分析系统中的垃圾短信[J]. 上海理工大学学报. 2008(01)

[3] 郑炜,沈文,张英鹏.基于改进朴素贝叶斯算法的垃圾邮件过滤器的研究[J]. 西北工业大学学报. 2010(08)

[4] 李钦.基于贝叶斯算法的短信过滤系统设计[J]. 中国科技论文在线.