# 一种垃圾邮件过滤器的设计与实现
# Design and Implement of An Junkmail Filter

**刘红翼**

Liu Hongyi

(**广西师范大学数学与计算机科学学院,广西桂林** 541004)

(Coll. of Math. and Comp. Sci. ,Guangxi Normal Univ. ,Guilin,Guangxi,541004,China)

**摘要**:以朴素的贝叶斯过滤器为基础,采用二进制表示方法建立垃圾邮件特征表,设计并实现一种垃圾邮件过滤器。该过滤器适于客户端使用,当客户端接收方收到新邮件时,对邮件的内容进行扫描,通过与特征表的对比,计算出特征词出现的概率,从而判定一个邮件是否为垃圾邮件。

**关键词**:电子邮件 垃圾邮件 过滤器

**中图法分类号**:TP393.098 **文献标识码**:A **文章编号**:1002-7378(2005)04-0258-02

**Abstract**: A junkmail filter is designed and implemented on the basis of naive Bayes filter by tabulating junkmail features in a form of binary system. This filter is used in the client-side. The filter would scan the new mails which have come into the client-side . Through a contrast with the feature form, emergence probability of feature words, the filter can tell whether the new mails are junk ones or not.

**Key words**: E-mail, junkmail, filter

电子邮件(E-mail)已成为一种重要的联系手段,大量垃圾邮件和计算机病毒的广泛传播,占用了有限的网络资源并影响着网络的安全。垃圾邮件发送者发送垃圾邮件的方式有直接发送和通过第三方转发两种。直接传送方式由于邮件发送人的真实情况很容易被查出来,已很少有人使用,目前多数使用第三方服务器转发。目前对付垃圾邮件的方法,主要是以防为主,防治结合。防范措施可分为服务器端和用户端两部分进行,如设置垃圾邮件过滤器,安装防火墙。普遍采用的是过滤器方式,有基于数字签名的过滤器[1],有基于规则的过滤器,贝叶斯过滤器[2],基于遗传规则过滤器[3]等等。目前比较流行的过滤系统有 CRM114, ASK (Active Spam Killer), Bogofilter 等。而在邮件客户端使用的软件均具有反垃圾邮件功能,如 Outlook Express6、Foxmail 5.0 等,更可以使用第三方软件来进行防范,如 Norton AntiSpam、Mcafee SpamKiller 等。过滤器常用的过滤方法有:黑名单技术和白名单技术,针对标题、正文、附件进行邮件内容规则过滤,利用文本分类与统计算法对邮件进行分类的过滤技术,应付新型垃圾邮件的"智能过滤"技术[4]。朴素贝叶斯过滤、支持向量机、遗传算法等文本分类法已应用于垃圾邮件的过滤器设计中。本文利用朴素贝叶斯定理[5]设计过滤器,该过滤器适于客户端使用。

## 1 过滤器的设计方法

本过滤器以朴素的贝叶斯过滤器为基础,利用先验概率求出后验概率,并根据训练样本集构造过滤器,过滤器根据邮件的后验概率对样本进行分类。即先将邮件中的文字分解成特征词,再对邮件中所有的特征词出现的次数进行统计,将结果利用贝叶斯定理进行计算后生成过滤规则;在规则生成过程中使用二进制方法来生成特征表;当接收方收到新邮件时,对邮件的内容进行扫描,通过与特征表的对比,计算出特征词出现的概率,从而判定一个邮件是否为垃圾邮件。

## 2 过滤器的设计与实现

### 2.1 过滤器设计原理

每一条消息可以用一个矢量来表示,其中 $x_1$, $x_2, x_3, \cdots, x_n$ 是属性 $X_1, X_2, X_3, \cdots, X_n$ 的值。如果消息含有属性 $X_i$,则相应的值取 1,否则取 0。属性集对应于单词。根据Bayes 理论和全概率公式[5],当给定一篇文档 $d$ 的矢量 $\vec{x} = (x_1, x_2, x_3, \cdots, x_n)$, $d$ 属于类别 $c$ 的概率是:

**收稿日期**:2005-06-27

**作者简介**:刘红翼(1969-),女,广西桂林人,讲师,主要从事数据库技术,数据挖掘技术和网络教育研究。

$$P(C=c|\vec{X}=\vec{x})=\frac{P(C=c)\cdot P(\vec{X}=\vec{x}|C=c)}{\sum_{k\in\{spam\ legitimate\}}P(C=k)\cdot P(\vec{X}=\vec{x}|C=k)},$$

假设 $X_1,X_2,X_3,\cdots,X_n$ 是条件独立于 $C$ 的，则

$$P(C=c|\vec{X}=\vec{x})=\frac{P(C=c)\cdot\prod_{i=1}^{n}P(X_i=x_i|C=c)}{\sum_{k\in\{spam,legitimate\}}P(C=k)\cdot\prod_{i=1}^{n}P(X_i=x_i|C=c)}.$$

这样从训练文档集中就可以很容易将属于类别和不属于类别的概率预测出来，从而可以用来进行是否为垃圾邮件的预测。

### 2.2 特征表的建立

一封邮件是否定为垃圾邮件是通过对其所包含的词进行分类来确定的。本系统采用特征向量进行表示。一篇文档用 1 个特征向量 $X$ 来表示，1 个词对应着 1 个特征，特征向量可以用从文档中提取出的所有词的各种组合来表示。构造特征向量的方法可以通过词频方式，即 $X$ 的第 $i$ 个元素是第 $i$ 个词在 $X$ 对应的文档中的出现次数。1 个词只有在至少 3 篇文档中出现过才作为 1 个特征。也可以采用二进制表示 1 个特定的词是否出现在某一篇文档中。这种方法排除在少于 3 篇文档中出现过的词。本系统采用二进制表示方法建立特征表，对前期搜集到的邮件进行扫描，包括邮件头、嵌入的 HTML 和 JAVAScript 语句等所有内容，对于那些同时包含有文字和数字的词，可以作为特征词，其他的看成分隔符；对于那些网页注释、纯数字信息可忽略不计。

### 2.3 过滤器实现

统计出所有的特征词在合法邮件和垃圾邮件中出现的次数，可以得到两份用来标识每个特征词出现次数的哈希表。然后建立第 3 个哈希表，在这个表中，每个特征词对应的是其所属邮件是垃圾邮件的概率。算法如下：

```
t=gethash word from mail
s=gethash word from spam
If  t+s>=3 then
  tt=t/nmail
  ss=s/nspam
  temp=min(1,ss)/(min(1,ss)+min(1,tt))
  probability=max(0.01,min(0.99,temp))
Endif
```

此处的 word 是指我们正在统计的特征词，而 mail 和 spam 是创建的 2 个哈希表，nmail 和 nspam 是该词分别出现在非垃圾邮件和垃圾邮件中的次数，就是 word 的先验概率。对只出现 1 次的词的概率，这里使用了 0.01 和 0.99，本过滤器只考虑那些在整个过程只出现过 3 次以上的词。这些数据可根据需要进行修改。每当一个新邮件到来，首先将其扫描成特征词，再根据其中特别关键的 15 个特征词进行计算，得出此邮件是否为垃圾邮件的概率。对于那些几乎从不出现的词或偶尔出现在合法邮件中的词，可以通过对 mail 中的数值进行加倍实现偏差处理，即增加一条语句 t=t * 2，从而可以避免错误判定。

在实际应用中，那些第 1 次出现的词，在哈希表中是不存在的。对于这样的词，我们一般认为其是无害的，其概率取值可定为 0.4。采用以上算法，当得出的概率超过 0.9 时，就可以认定邮件为垃圾邮件。

## 3 结束语

采用贝叶斯过滤器可计算性强，采用的特征词可以完全由用户根据自己所接收的垃圾邮件和非垃圾邮件来创建，可以设计成对用户来说独一无二的过滤器。但这种方法只能将邮件进行是与非的划分，分类精确度不可能达到 100%。宗平等[6]已经提出了改进方案，把经过朴素 Bayes 分类后的邮件，再用相同方法按照重要程度，例如一级、二级、三级等进行二次分类，这样可以提高分类的精确程度。此外，还可以有很多的方法对朴素贝叶斯方法进行改进，如采用信息增益、特征长度和文档频率范围[7]等。还可以将过滤器与邮件服务器集成，以提高分类的准确度。

参考文献：

[1] Symantec corporation. Brightmail Acquired by symantec[EB/OL]. http://brightmail.com, 2005-08-30.

[2] Mehran Sahami, Susan Dumais, David Hecherman, et al. A bayesian approach to filtering junk E-mail[M]. In: Learning for Text Categorization: Papers from the 1998 workshop, Madison, Wisconsin. AAAI Technical Report WS-98-05, 1998.

[3] Hooman Katirai. Filtering Junk E-mail: A Performance Comparison between Genetic Programming & Native Bayes[A]. Technical report, University of Waterloo, 1999.

[4] 李 建，刘克胜，揭 摄. 垃圾邮件与病毒防治[J]. 计算机安全，2004，(10)：53-54.

[5] 魏宗舒. 概率论与数理统计教程[M]. 北京：高等教育出版社，1983.

[6] 宗 平，田震生. 基于朴素贝叶斯分类器邮件分类系统的改进[J]. 计算机与现代化，2004，12：48-49.

[7] 李惠娟，高 峰，管晓宏，等. 基于贝叶斯神经网络的垃圾邮件过滤方法[J]. 微电子学与计算机，2005，22(4)：107-111.

（责任编辑：邓大玉）