第16卷 第2期
2006年2月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.16 No.2
Feb. 2006

一个基于 Naive Bayesian 垃圾邮件过滤器的改进

成宝国,冯宏伟
(西北大学 计算机科学系,陕西 西安 710069)

摘 要:近几年来,垃圾邮件成为互联网的公害之一。现有的反垃圾邮件技术中,基于统计方法的 Naive Bayesian 分类算法在垃圾邮件过滤中有很好的效果。文中简单介绍了 Naive Bayesian 分类算法,提出了一种旨在提高垃圾邮件过滤精确率的改进方案,并给出了实验结果。

关键词:垃圾邮件;Naive Bayesian 文本分类器;反垃圾邮件技术
中图分类号:TP393.098 **文献标识码**:A **文章编号**:1005-3751(2006)02-0098-02

Design of an Improved Spam Filter Based on Naive Bayesian Classifier

CHENG Bao-guo, FENG Hong-wei
(Computer Science & Technology Department, Northwest University, Xi'an 710069, China)

Abstract: In recent years, spam does much harm to Internet. In kinds of anti-spam technologies, Naive Bayesian classifier based on statistical method works effectively on spam filtration. In this paper, describes Bayesian filter in brief. Furthermore, presents an improved model based on Naive Bayesian spam filter.

Key words: spam; naive Bayesian filter; anti-spam technology

0 引 言

随着 Internet 的发展和普及,电子邮件(E-mail)以其方便、快捷、低成本的独特魅力成为人们日常生活中不可缺少的交流方式之一。但是电子邮件给人们带来极大便利的同时,也日益显示其负面影响,那就是人们每天收到的电子邮件中有很大一部分是那种"不请自来"的,它们或者是推销广告,或者是一些有害的不良信息,甚至还有病毒。据中国互联网信息中心(CNNIC)的统计报告显示,垃圾邮件的数量已经超过了正常邮件数量。作为垃圾邮件的发送方,其成本是极低的,而对电子邮件服务提供商和用户来说,垃圾邮件给他们带来了很大的危害和损失。

针对垃圾邮件泛滥的情况,到现在为止,国际上主要反垃圾邮件技术有如下6种:

1)IP地址、域名、邮件地址黑白名单方式。这种技术手段是最传统的方式,它通过黑名单技术对垃圾邮件进行屏蔽,通过白名单技术对允许的邮件进行放行。

2)基于信头、信体、附件的内容过滤方式。该项技术目前尚不成熟,因为现在的群发程序自动生成和发送的垃圾邮件对于发件人、收件人、邮件主题甚至邮件内容都是随机生成的,使得该种技术目前应用范围日趋狭窄。

3)基于统计分析的贝叶斯算法技术。基于统计的原则,采用标记权重的方式,根据对用户认为的垃圾邮件和非垃圾邮件进行统计计算,生成过滤规则,具有学习渐进的功能,可以逐渐取得好的效果。

4)基于连接频率的动态规则方式。由于一个正常用户发送邮件的数量和频率远远低于垃圾邮件发送者,因此可以根据垃圾邮件发送具有一定时间内邮件数量和邮件连接频率都非常大的情况,从频率和数量对垃圾发送者的连接行为进行控制。

5)电子邮票方案。因为垃圾邮件发送具有大规模发送成本很小的行为特征,微软公司提出了对发送邮件进行收费的解决方案。不过这种方式却是对广大的正常邮件发送者带来了新的负担,还需考虑。

6)Challenge-Response 方式。挑战-应答模式是从增加垃圾邮件发送者时间成本上入手,要求每发送一封邮件,就要求发件人回答一些问题的方式来增加发送时间。

众多的反垃圾邮件方法中,基于 Bayesian 的垃圾邮件过滤器效果突出,实验数据显示,准确率97%以上[1]。

1 Naive Bayesian 过滤器

1.1 Bayesian 方法简介

将文本分类抽象为一般描述:设类别总数为 $|C|$,c_j 表示第 $j(j=1,2,\cdots,|C|)$ 类,提供给分类器的有 $|D|$ 篇文本,特征空间 $(t_1,t_2,\cdots,t_n)$,n 为特征数量,每篇文本表示为 $d_i=(w_{i1},w_{i2},\cdots,w_{in})$,$i=1,2,\cdots,|D|$。一篇待分类的文本泛化表示为 $dx=(w_{x1},w_{x2},\cdots,w_{xn})$。

收稿日期:2005-05-27
作者简介:成宝国(1976—),男,陕西咸阳人,硕士研究生,研究方向为数据挖掘与人工智能;冯宏伟,副教授,研究方向为图形图像与多媒体技术。

Bayesian 分类算法[2,3] 应用于文本分类时,通过计算属于每个类别的概率 $p(c_j \mid d_x)$,将该文本归为概率最大的一类。Naive Bayesian 分类算法,它建立在"假定所有抽取的特征之间相互独立"基础之上。实际上,在生活中这种独立性很难存在,但从目前的实验结果来看,基于这个假设的 Bayesian 分类算法的效果很好[1],而且计算简单。

1.2 Bayesian 分类算法的原理

对于待分文档 d_x,计算 d_x 属于某个类别的概率 $P(c_j \mid d_x)$,这时要用到 Bayesian 公式计算 $p(c_j \mid d_x)$。

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

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

根据全概率公式:$P(d_x) = \sum_{j=1}^{\mid C \mid} P(c_j)P(d_x \mid c_j)$

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,则有:

$$P(d_x \mid c_j) = P(t_1 \mid c_j)P(t_2 \mid c_j)\cdots P(t_n \mid c_j)$$

$$= \prod_{i=1}^{n} P(t_1 \mid c_j)$$

$P(c_j)$ 和 $P(t_i \mid c_j)$ 都可以利用训练集估计。

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

Naive Bayesian 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。

2 提出的改进方案

2.1 改进方案

基于传统方法的 Naive Bayesian 垃圾邮件过滤模型如图 1 所示。

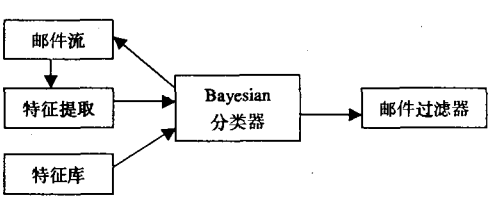

图 1　Naive Bayesian 垃圾邮件过滤模型

对于基于内容过滤的垃圾邮件过滤器来说,若把垃圾邮件里的关键的词随机让它变化形式,邮件过滤器就很难检测出来。例如,'adult'就可以变化为'a * dul \ t','a + d * ult','adul - t!'等等。但是这些对邮件用户来说,很容易明白变化后的词的意思。针对上面的情形,提出一种基于规则的词干化(word stemming)技术[4]。

步骤如下(仅以英文为例)。对于一个单词:

1)删除单词中所有非字母的字符(比如'*','-','?'等);

2)删除单词中的所有数字;

3)把单词中持续重复的字符用单个相应字符替换;

4)再用 stemming 算法[5]处理。

改进后的邮件过滤模型如图 2 所示。

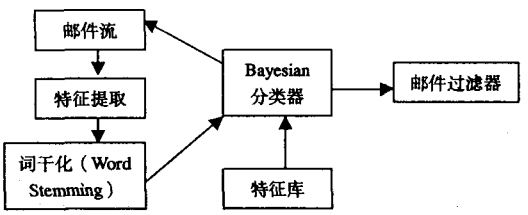

图 2　改进后的邮件过滤模型

2.2 实验结果

按图 2 的改进方案,笔者收集了两组电子邮件进行实验。第一组是 2003 年的 6 月至 10 月间收集的 2933 个电子邮件;第二组是 2004 年的 1 月至 3 月间的 3954 个电子邮件。从图 3 清楚地看到垃圾邮件过滤器的效率在下降,这说明了垃圾邮件的发送者在改变方式以欺骗过滤器。

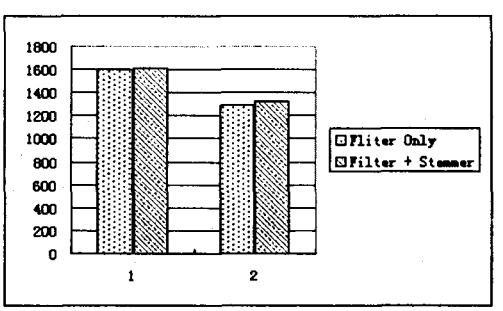

图 3　改进后的过滤器和没有改进的过滤器的工作效果比较

3 结　论

随着反垃圾技术的发展,垃圾邮件的发送者也在改变邮件文本的表现方式来逃避过滤器的检测。从实验结果来看,文中所提出的方法提高了 Bayesian 垃圾邮件过滤器的精确率,达到了预期的结果。

参考文献:

[1] Androutsopoulos I, Koutsias J, Chandrinos K V, et al. An Evaluation of Naive Bayesian Anti - Spam Filtering[A]. Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (ECML 2000)[C]. Barcelona, Spain: [s. n.], 2000. 9 - 17.

[2] Graham. A Plan for Spam[EB/OL]. http://www.paulgraham.com/spam.html, 2005 - 04 - 05.

[3] 潘文峰. 基于内容的垃圾邮件过滤研究[D]. 北京:中国科学院, 2004.

[4] Ahmed S, Mithun F. Word Stemming to Enhance Spam Filter [A]. Proceedings of the 1st Conference on Email and Anti - Spam (CEAS 2004)[C]. CA, USA: Mountain View, 2004.

[5] Porter M. Stemming Algorithm[EB/OL]. http://www.tartarus.org/martin/PorterStemmer/, 2005 - 05 - 17.