文章编号:1004-5694(2005)05-0636-04

基于贝叶斯概率模型的邮件过滤算法探讨*

刘明川,彭长生
(重庆邮电学院,重庆 400065)

摘　要:讨论了邮件过滤模块,通过分析研究该模块中垃圾邮件关键词的统计概率分布,提出了基于贝叶斯概率模型的邮件过滤算法,并对该算法的合理性和复杂度进行了分析。可以根据垃圾邮件内容的特征,建立贝叶斯概率模型,计算出一封邮件是垃圾邮件的概率,从而判断其是否为垃圾邮件。

关键词:误承认;过滤网;Hash 表
中图分类号:TP393.08　　文献标识码:A

0　引　言

在因特网发送的邮件中,75% 是垃圾邮件,用户的收件箱中常常充斥着不需要的、欺诈性的或者令人生厌的邮件,我们不得不为删除这些邮件浪费大量时间,即使我们有过滤功能的软件,也会担心重要的邮件被误投进垃圾箱。对于垃圾邮件的过滤目前所面临的主要问题是:邮件的接收者在收到邮件之前,如何判断是否是垃圾邮件,同时又要减少误判断的可能性。这一切必须由邮件防火墙的邮件过滤模块来实现,目前常用的基于结构化文本的过滤、基于规则的评定、分布式适应性黑名单的过滤方法已经远远不能满足我们的要求[1,2]。因此,需要我们的邮件过滤模块具有一定的智能性,使得它对垃圾邮件的判断率较高,同时误承认的概率较小。

我们对大量的垃圾邮件的单词进行了分析,认为垃圾邮件的单词是符合贝叶斯(Bayes)概率模型的。其基本思想是:在已知的垃圾邮件中,一些单词出现的频率较高,而在合法邮件中,另一些单词出现的频率较高。运用概率论与数理统计的数学知识,对每个单词找出可以生成一个“垃圾邮件指示性概率”。根据邮件中所包含的一组词,可以用一个简单的数学公式来确定邮件的“垃圾邮件概率”[3]。因此,我们可以根据垃圾邮件内容的特征(即所含有特殊单词和代码的概率)的多少,建立起贝叶斯概率模型,通过贝叶斯公式计算出一封邮件是垃圾邮件的概率,从而判断其是否为垃圾邮件。

1　基本原理

在设计邮件过滤代理模块的时候,面对的一个十分棘手的问题是如何尽可能地降低误承认率,理想情况下误承认率要尽量地小,甚至为 0,但是这是十分困难的。对于垃圾邮件的误承认就是指在对垃圾邮件过滤的时候,同时也将正常的合法的邮件过滤了,有时候发生这种情况是让用户不能容忍的。对于用户来说有时宁愿多收到一些垃圾邮件也不希望被邮件过滤器过滤掉一封正常邮件。经过对垃圾邮件所做的大量的数理统计分析,可以得出垃圾邮件过滤规则的严格性(filter rule stricter)与误承认的概率大小(false positive)间关系[4],如图 1 所示。

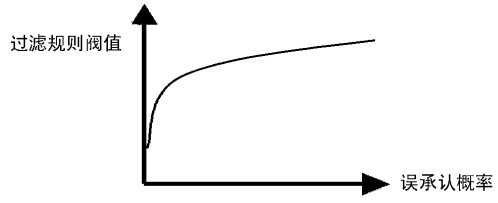


图 1　过滤规则与误承认率之间的关系

Fig. 1 Relationship between filter rule stricter and false positive

信息过滤是一个将用户感兴趣的文档从某个文档集中筛选出来的过程。有关信息过滤的研究有多种分类方法,其中对于符合贝叶斯概率模型的垃圾邮件过滤算法可以使用类似于基于特征的贝叶斯过滤网模型。

设事件 B 可以分解为两两互斥的 N 个小的事件 $A_1, A_2, \cdots, A_n$,$P(A_i) > 0 (i = 1, 2, \cdots, n)$,$P(B) >$

* 收稿日期:2005-01-11　　修订日期:2005-05-09
作者简介:刘明川(1974-),男,重庆合川人,讲师,研究方向为数据库技术和计算机安全。

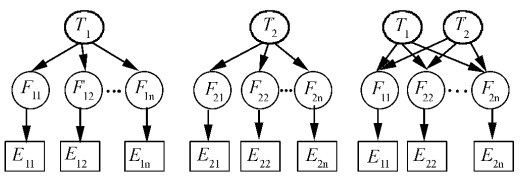
0,则有:

$$P(A_j \mid B) = \frac{P(A_j)P(B \mid A_j)}{\sum_{i=1}^{n} P(A_j)P(B \mid A_j)} \quad (1)$$

设 $T_i(i = 1,2,\cdots,m)$ 为用户关注的某一命题或主题,$F_{ij}(j = 1,2,\cdots,n)$ 为主题 T_i 下被查询的文件特征,那么,贝叶斯公式(1)可以写成

$$P(T_i \mid F_{i1},F_{i2},\cdots,F_{in}) = \frac{P(F_{i1},F_{i2},\cdots,F_{in} \mid T_i)P(T_i)}{\sum_{i=1}^{n} P(F_{i1},F_{i2},\cdots,F_{in} \mid T_i)P(T_i)} \quad (2)$$

由式(2),可以构造如图2所示的贝叶斯过滤网。图2中的圆节点表示随机变量或命题,方节点表示证据(Evidence)或网的入口(Entrance),有向弧(箭头)表示节点间的依存关系,没有弧直接相连的节点互相独立。应该指出,常规人工智能的正向推理的指向为if(症状)→then(疾病);与此相反,贝叶斯网的弧的指向为原因(或条件)→结果,符合人的日常因果推理模式。图2a表示简单过滤网的一种树状结构模式:单个父(主题 T_i)节点多个子(特征 F_{ij})节点,各主题节点间和各特征节点间都互相独立。图2b的网结构就比较复杂些[5,6]。



a 树状结构 b 网结构

图2 贝叶斯过滤网

Fig. 2 Filter gateway of Bayes

邮件过滤代理模块的算法思想就是简单地基于贝叶斯概率模型下的条件概率[3],假设有 $A_i(i = 1,2,\cdots,n)$ 个特征可以判断一封邮件是垃圾邮件,$P(A_i)(i = 1,2,\cdots,n)$ 代表具备 A_i 特征的垃圾邮件的概率,那么则判断垃圾邮件的概率可以用式(3)表示。

$$P(\text{Spam}) = \frac{\prod_{i+1}^{n} P(A_i)}{\prod_{i=1}^{n} P(A_i) + \prod_{i=1}^{n} P(1 - A_i)} \quad (3)$$

当计算出 $P(\text{Spam})$ 后,我们可以根据预定的过滤规则的一个阈值 α 来判断是否是垃圾邮件,该阈值的大小决定了过滤规则的严格性,阈值 α 越大则过滤规则越严格,会导致误承认率增大(见图1);阈值 α 过小则使过滤规则严格性降低,导致过滤效率降低。为了使结果具备可比性,我们采用PU1语料测试贝叶斯分类算法应用于邮件过滤时的性能。该语料来自提供者在一段时间内收到的真实邮件。实验方法采用常用的"10次交叉证",结果取平均值。Androutsopoulos等人在PU1语料上实验贝叶斯,特征数量从50增长到700,每次增加50个。表1是他们得出的最好的结果[7]。从表1中可以看出,简单贝叶斯分类算法应用于垃圾邮件过滤的效果在实验中得到了验证。

表1 基于贝叶斯概率模型的实验结果

Tab. 1 Results of test based on the probability model of Bayes

语料	阈值	特征数量	Recall /%	Precision /%
PU1 bare	0.50	50	83.98	95.11
	0.90	100	78.77	96.65
	0.999	700	46.96	98.80
PU1 lemm	0.50	100	78.14	98.25
	0.90	100	75.86	98.50
	0.999	50	60.68	98.79
PU1 stop	0.50	50	84.19	97.34
	0.90	150	74.83	98.76
	0.999	700	47.17	97.96
PU1 lemm_stop	0.50	100	79.60	97.96
	0.90	100	75.86	97.91
	0.999	600	49.45	98.31

在对大量的垃圾邮件的单词进行了分析的基础上,我们设定阈值 $\alpha = 85\%$,当 $P(\text{Spam}) > \alpha$ 时,可认为邮件为垃圾邮件,从而将其滤除;当 $P(\text{Spam}) < \alpha$ 时,可认为是正常邮件,并将其转发到用户邮箱中[2]。

2 算法分析及实现

邮件防火墙的关键模块是邮件过滤模块。Sieve是一种Internet邮件过滤的标准性通用语言,它可以基于各种软件平台进行过滤操作,不依赖任何操作系统或者邮件框架。我们选择在QMAIL上利用Sieve和C语言进行二次开发,在服务器端的MTA代理qmail邮件队列程序(qmail-queue)前加入过滤器的方案,从而实现邮件的过滤。

Qmail的投递机制为:当Qmail接收到来自本地和远程邮件系统发来的邮件后,经Qmail-inject和Qmail-smtpd分别送到Qmail-queue程序再发送到邮件队列,然后调用Qmail-send将邮件发送到Qmail-local和Qmail-Remote来实现本地邮件和远程邮件的投递[3],其过程如图3所示。

由以上Qmail的投递模型可知,我们只要在Qmail-queue的前面加上一个Filter模块即可以实现

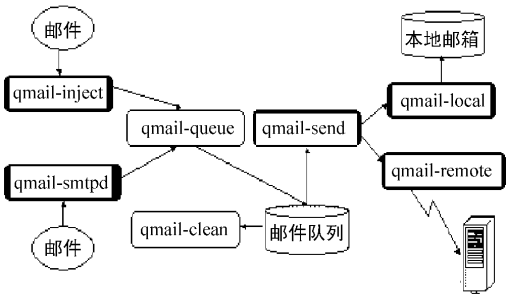


图 3　Qmail 投递模型

Fig. 3 Model of Qmail's sending

邮件的过滤。

2.1　构建关键词 Hash 表

对所收到的邮件进行分类归档,分为垃圾邮件集和正常邮件集。这是一个数据积累的过程,该步的时间复杂度为 $O(1)$ [2]。

对于所收集到的垃圾邮件作为一个整体的信息库,首先判断它是否是中文邮件,如果是中文邮件,则删除空格后对他进行关键词扫描,并统计在垃圾邮件信息库中关键词所出现的概率;如果是英文邮件,则直接进行关键词扫描,并统计在垃圾邮件信息库中关键词所出现的概率,在这一步中各种标点符号、HTML 标记、数字等信息都被忽略。时间复杂度为 $O(n)$。

在对关键词的概率统计的时候特别注意邮件头中的'from:'字段中的用户名、IP 地址、主机名,以及正文中的敏感单词如:dear,Sir,madam,sex,法轮功等的统计,统计过后计算它们的概率分布,这里可以使用专门的数理统计软件(如 NoSA V2.30),然后取每个关键词的数据结构(如图 4 所示)。

Key	Perc	nex
sex	0.99	Nul

图 4　关键词数据结构

Fig. 4 Data structure of Key words

提取关键词并对关键词进行统计分析以后,需要建立多张 Hash 表,这主要是因为垃圾邮件的多样性,以及不同的用户对垃圾邮件的定义也不同。为了根据不同的用户需要进行垃圾邮件的分类过滤,必须依据垃圾邮件不同的类型建立多个 Hash 表。

该 Hash 表的散列函数使用的是 ELFhash 函数,它把字符串的绝对长度作为输入,可以使关键字字符串产生平均分布,从而减少散列函数的冲突。把散列表中的每一个槽定义成一个链表的开头,散列到一个特定的槽内的所有记录都放到这个槽的链表中,用开散列的方法进行冲突避免,如图 5 所示,其时间复杂度为 $O(n)$。

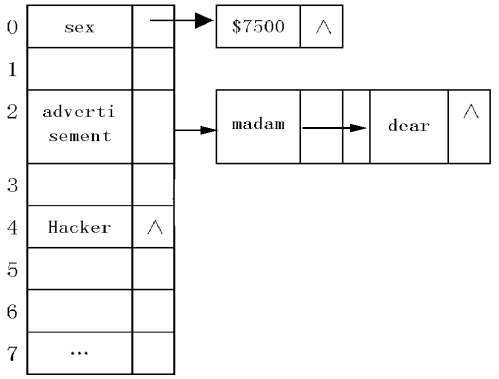


图 5　关键词散列冲突避免

Fig. 5 Avoidance of Key words collision

2.2　垃圾邮件概率的计算

当新的邮件到达邮件防火墙的邮件过滤代理的时候,它对邮件关键词进行扫描,提取其中有兴趣度的关键词(以该关键词在垃圾邮件中出现的概率来判断),最高的 15 个关键词。其过滤算法如图 6 所示,其时间复杂度为 $O(n)$。

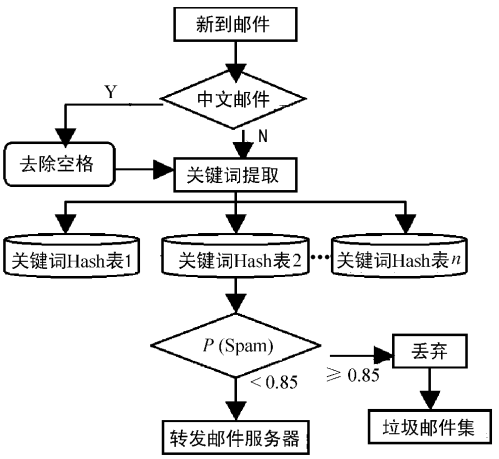


图 6　基于贝叶斯概率模型的邮件过滤算法

Fig. 6 Filter algorithm based on probability model of Bayes

当 15 个关键词被提取出来以后,对这几个关键词查找关键词相应的概率分布 Hash 表,得到每个关键词的贝叶斯统计概率,如果 Hash 表查找失败,则根据选取关键词的贝叶斯统计概率为 0.4 [4],得到所有关键词的贝叶斯统计概率值后,再利用式(3)进行计算,得到垃圾邮件的概率值 $P(\text{Spam})$ [5]。

①假如 $P(\text{Spam}) \geqslant \alpha(\alpha = 85\%)$,则判断该邮件为垃圾邮件,将其滤除并存入垃圾邮件集合中。如果垃圾邮件集合中的邮件超过 4000 则重新构建 Hash 表。否则,退出邮件过滤代理。

②假如 $P(\text{Spam}) < \alpha(\alpha = 85\%)$,则判断该邮件

为合法邮件将其转发到邮件服务器,并退出邮件过滤代理。该步的时间复杂度为 $O(n^2)$。从以上的分析可以看出整个算法的时间复杂度为:

$$O(1) + O(n) + O(n) + O(n) + O(n^2) = O(1 + 3n + n^2) = O(n^2) \quad (4)$$

所以该算法是多项式时间复杂度的算法,算法的性能满足实际情况的要求。

3 结束语

针对目前常见的邮件防火墙系统的特点及不足,我们提出了一个基于模块化的邮件防火墙系统方案,重点讨论了邮件过滤模块,通过分析研究该模块中垃圾邮件关键词的统计概率分布,提出了基于贝叶斯概率模型的邮件过滤算法,并对该算法的合理性和复杂度进行了分析。将此算法应用到邮件防火墙系统中,整个系统可根据用户的需求建立不同类型的过滤器,从而根据邮件关键词概率分布推断出收到的垃圾邮件的概率,由此判断出是否为垃圾邮件,提高了系统对垃圾邮件过滤的自适应性。

参考文献:

[1] 林晓东,杨义先. 网络防火墙技术[J]. 电信科学. 1997,13(3):41-43.

[2] 熊安萍. 基于邮政三网的电子邮政[J]. 重庆邮电学院学报(自然科学版),2004,16(2):101-104.

[3] PAUL Graham. A better Bayesian filtering [A]. 2003 Spam Conference [C]. January 2003,156-188.

[4] 李宝林. 阻止不良信息过滤器的研究与设计[J]. 计算机应用研究,2004,21(11):142-144.

[5] PAUL Graham. A plan for spam [A]. 2002 Spam conference [C]. August 2002,73-76.

[6] CHANG K C, FUNG R M. Target identification with Bayesian networks in a multiple hypothesis tracking system [J]. IEEE Trans. Optical Engineering, 1997,36(3):684-691.

[7] 潘文峰. 基于内容的垃圾邮件过滤的研究[EB/OL]. http://www.nosounds.com/meteor/01.pdf2004-05:27-29. (责任编辑 郭继笃)

Research on mail filter algorithm based on bayes probability model

LIU Ming-chuan, PENG Chang-sheng

(*Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *P. R. China*)

Abstract: Mail filter model is discussed. Under this model, The authors analyze statistics of junk mail keyword in this paper. The authors put forth a mail filter algorithm based on Bayes probability model, and analyze rationality and complexity, and estabtish the probability model of Bayes according to the characteristic of the keywords, And then the mail can be filtered through the model.

Key words: mis-admit; filter gateway; Hash list

下 期 要 目