第23卷第1期 计算机应用与软件 Vol. 23, No. 1
2006年1月 Computer Applications and Software Jan. 2006

基于贝叶斯理论的垃圾邮件过滤技术

戴劲松 白英彩
(上海交通大学信息安全工程学院 上海 200030)

摘 要 垃圾邮件已成为损耗生产力的问题,反垃圾邮件技术不断出现,基于贝叶斯理论的垃圾邮件过滤技术有其独特的优势,研究针对中文的贝叶斯垃圾邮件过滤技术具有理论和现实的意义。

关键词 垃圾邮件 贝叶斯过滤

ANTI-SPAM TECHNOLOGY BASED ON BAYESIAN THEORY

Bai Jingsong Bai Yingcai
(*College of Information Security Engineering, Shanghai Jiaotong University, Shanghai 200030, China*)

Abstract Spam has become wastage of productivity. Along with this, more and more antispam technology appear, Bayesian Filtering has some advantages among them. It is useful to do research on Bayesian Filtering for Chinese.

Keywords Spam mail Bayesian filtering

0 引 言

随着计算机网络的快速发展,电子政务、电子商务的普及,电子邮件成为越来越重要的一种通信方式。但随之而来产生的垃圾邮件问题,已经从仅仅让人觉得讨厌变成了对生产力的损耗。反垃圾邮件技术迅速成为一个热门的研究课题,贝叶斯关键元加权统计算法和它的变种是其中倍受关注的一类。

1 反垃圾邮件技术发展现状

垃圾邮件主要是广告邮件,就如同我们在现实社会中,不管愿与不愿,总无可避免的会接触到越来越多的广告一样,垃圾邮件也只会越来越多,而绝不会自动减少。

要做到很好的阻挡或过滤垃圾邮件,首先应分析垃圾邮件迅速增多的原因。垃圾邮件增多有两个相互关联的原因:一是发送成本低,具备一定条件可在短时间内大量发送,单封邮件发送成本几乎为零;二是对发送者有利可图,在发送者发出的大量垃圾邮件中,只要有极少数被人阅读,有了回应,发送者就可以获得利益。已出现的反垃圾邮件技术的思路针对以上特点主要有两类:一是增加发送者成本,使其无利可图,主要集中在协议的改进,应用前景集中在邮件服务提供方;二是针对垃圾邮件本身进行判别过滤,包括针对邮件头信息及邮件内容进行判别过滤两类,既可以应用在邮件服务提供方,也可以用在使用邮件服务的客户端。

目前影响较大的主流技术有以下三种:

1) 针对邮件头信息进行过滤的方法:如:黑白名单法反垃圾邮件技术,主机名反向验证技术;

2) 针对邮件内容进行过滤的方法,如各类贝叶斯关键元加权统计算法和它的变种;

3) 协议改进类的方法,如:①IRTF提出的在不放弃SMTP方式和许多其他这类协议的情况下的三种方法:校验电邮地址终端发送准许(SPF)、指定邮寄者协议(DMP)、保留邮件交换(RMX);②雅虎DomainKeys解决方案:通过验证邮件是否来自合法区域及是否带有正确的密码来识别合法邮件;③微软的黑便士邮票计划,微软研究组表示并非要求发Email的用户付费,而是让发邮件用户的电脑计算一个花费10到20秒时间的密码学谜语;④其他各类challenging response反应应答式反垃圾邮件技术和数字签名反垃圾邮件技术等。

第一类方法,相对简单而行之有效,已经被广泛应用,但有其局限性,并不能阻止所有的垃圾邮件,还需要其它方法作为补充。第三类方法,应用部署受到局限,很难在所有使用电子邮件的用户中推广。

上面两类方法难以解决如下两个问题。一是判别垃圾邮件标准不同。每个使用电子邮件系统进行通信的实体,如一家公司,或是个人,对于什么样的邮件是垃圾邮件有着不同的评判标准。例如,一个从事商业活动的用户不会认为跟本身业务有关的商业广告邮件是垃圾邮件。而一个不从事商业活动的用户可能会认为所有不是自己订阅的商业广告邮件都是垃圾邮件。二是邮件头信息及协议考查的参数总是存在可以伪造或欺骗的可能性。针对这些信息开发的反垃圾邮件技术可以说并没有把握垃圾邮件的本质特征,即垃圾邮件的内容依据某种标准来判断是"垃圾",对于一封垃圾邮件,这一点才是永远不会改变的。具有智能学习功能,针对内容进行过滤的反垃圾邮件方法,才能解决判定垃圾邮件标准不同的问题,并具有长久的适用性。贝叶斯过滤方法(Bayesian Filtering)正是一种具有智能学习功能,针对内容过滤的方法,国外已有人使用该方法进行实验,取得了

收稿日期:2004-07-05。戴劲松,硕士生,主研领域:信息网络安全。

良好的效果。

2　贝叶斯过滤方法发展历史及研究现状

贝叶斯理论由托马斯·贝叶斯(1702 - 1761)提出,在他身后于1763年发表在伦敦皇家学会哲学学报上的一篇名为《论有关机遇问题的求解》的论文中。贝叶斯定理是计算概率的一种方法,即认为一个事件会不会发生取决于该事件在先验分布中已经发生过的次数。该理论在许多需要具备自学能力的智能系统中得到广泛的应用。

使用贝叶斯方法过滤垃圾邮件的基本原理简述如下[1]:

1) 收集一定数量的邮件,建立垃圾邮件集和非垃圾邮件集。

2) 提取邮件主题和邮件正文中的独立字串作为TOKEN串,并统计提取出的TOKEN串出现的次数,即字频。

3) 每一个邮件集对应一个哈希表,Hashtable_good对应非垃圾邮件集,而Hashtable_bad对应垃圾邮件集。表中存储TOKEN串到字频的映射关系。

4) 计算每个哈希表中某一TOKEN串出现的概率 $p = \frac{\text{某 TOKEN 串的字频}}{\text{对应哈希表的长度}}$。

5) 综合考虑Hashtable_good和Hashtable_bad,推断出当新来的邮件中出现某个TOKEN串时,该新邮件为垃圾邮件的概率。数学表达式为:

A 事件——邮件为垃圾邮件;

$t_1, t_2, \cdots, t_n$ 代表TOKEN串,则 $p(A/ti)$ 表示在邮件中出现TOKEN串 t_i 时,该邮件为垃圾邮件的概率。

设 $p_1(t_i)$ = (t_i 在 Hashtable good 中的值),$p_2(t_i)$ = (t_i 在 Hashtable bad 中的值),则 $p(A/t_i) = \frac{P_2(t_i)}{p_1(t_i) + p_2(t_i)}$。

6) 建立新的哈希表Hashtable_probability存储TOKEN串 t_i 到 $p(A/t_i)$ 的映射。

7) 至此,垃圾邮件集和非垃圾邮件集的学习过程结束。根据建立的哈希表Hashtable_probability可以估计一封新到的邮件为垃圾邮件的可能性。

当新到一封邮件时,按照步骤2)提取TOKEN串。查询Hashtable_probability得到该TOKEN串的键值。

假设由该邮件共得到 N 个TOKEN串,$t_1, t_2, \cdots, t_n$,Hashtable_probability中对应的值为 $p_1, p_2, \cdots, p_N$,$P(A/t_1, t_2, t_3, \cdots, t_n)$ 表示在邮件中同时出现TOKEN串 $t_1, t_2, \cdots, t_n$ 时,该邮件为垃圾邮件的概率。

由复合概率公式可得:

$$p(A/t_1,t_2,t_3,\cdots,t_n) = \frac{p_1 * p_2 * \cdots * p_N}{(p_1 * p_2 * \cdots * p_N + (1-p_1) * (1-p_2) * \cdots * (1-p_N))}$$

当 $P(A/t_1, t_2, t_3, \cdots, t_n)$ 超过预定阈值时,就可以判断邮件为垃圾邮件。

在使用过程中,贝叶斯反垃圾邮件过滤系统一般不采取直接阻断垃圾邮件的方式,而是在判定为垃圾的邮件上做上标识,让用户看到所有的邮件。并且用户可以人为地更正过滤系统的判断结果,系统记录修正的情况,调整新收到的邮件中TOKEN串在不同表中的位置,作为以后过滤的依据。这也是贝叶斯过滤法进行学习,以更准确的根据使用者自身的标准进行垃圾邮件判断和过滤的原理。

从上文的描述中可以看出,基本的贝叶斯过滤方法思路清晰而简单。当然在实现过程中,对于某些环节可以做出调整和改进,比如人为增加代表垃圾邮件的TOKEN串在概率表(Hashtable_probability)中的权值[1,2],以提高过滤算法的效率和准确性等等。总的来说,针对字母语言的贝叶斯过滤方法已有很多公开的研究成果及实验结论,但目前还没有看到针对中文的研究成果。这是因为,中文有着跟字母语言截然不同的特点,研究实现针对中文特点的贝叶斯过滤方法对抗垃圾邮件,具有理论和现实的意义。

3　针对中文的贝叶斯垃圾邮件过滤要考虑的特殊问题

3.1　中文词法分析

决定一篇邮件内容的关键是邮件中包含的实意词,如果不能很好的提取邮件中的实意词,就不能很好的对邮件的内容进行判断。字母语言由明显的单词构成,单词中间一般都有明显分隔符,便于提取和处理。针对中文的贝叶斯垃圾邮件过滤算法,关键在于很好的分析中文语言的特点,实现对中文邮件中词语的合理提取。汉语在语法上有一些特点,仅仅从形式上看,这种特点主要体现在以下几个方面[4]:

1) 汉语的基本构成单位是汉字而不是字母。常用汉字就有3000多个(GB2312一级汉字),全部汉字达数万之多(UNICODE编码收录汉字20000多);

2) 汉语的词与词之间没有空格分开,也可以说,从形式上看,汉语中没有"词"这个单位;

3) 汉语词没有形态上的变化(或者说形态变化非常弱),同一个词在句子中充当不同语法功能时,形式是完全相同的;

4) 汉语句子没有形式上唯一的谓语中心词。

在贝叶斯垃圾邮件过滤算法中要用到的中文词法分析应能完成下面几项任务:

1) 查词典识别标准词;

2) 处理重叠词、离合词、前后缀;

3) 未定义词识别;

a) 时间词、数词处理;

b) 中国人名识别;

c) 中国地名识别;

d) 译名识别;

e) 其他专名识别。

4) 某些有特殊意义的非汉字字符的识别提取,如¥、$等。

在中文词法分析方面,国内开展研究比较早,已有很多成熟有效的技术,经过修改,可以很好的应用到贝叶斯方法的反垃圾邮件系统中。

3.2　抵抗反过滤措施

针对内容过滤法,垃圾邮件发送者已经想出了一些对抗的办法,典型的一种是:在某些敏感的词语中间加入不等个数的无意义字符。例如我们想过滤掉含有"免费"两个字的邮件,发送者在"免"、"费"这两个字中间加上数量不等的空格,星号等字符,以对抗过滤措施。

对这一问题,可考虑用如下方法解决:设定一个默认词库,包含需要过滤的关键字,以及经过测试典型的代表垃圾信息的词,在针对邮件全文提取词时,一旦发现某个字是默认词库中某

(下转第124页)

攻击者将它重定向到 1139 端口。在 1139 端口上,攻击者的代理程序正在监听。

negprot request [client]——>[attacker][server]

攻击者接收到了 negprot 请求数据报。

negprot request [client][attacker]——>[server]

攻击者重定向 negprot 请求数据报到服务器。

negprot reply [client][attacker]<——[server]

(加密属性位设置为要求对密码进行加密)

服务器发送了一个将加密属性设置为 1 的 negprot 应答数据报,要求客户端对密码进行加密传送。攻击者并不会重定向发送这个数据报。他改变了要求加密的那一位,而告诉客户端不用对口令进行加密。

negprot reply [client]<——[attacker][server]

(加密属性位被设置为无需加密口令)

攻击者将修改过的无需加密口令的 negprot 应答报文发送到客户端。

SesssetupX request [client]——>[attacker][server]

(密码以明文形式传送)

客户端将密码以明文形式传送,这时攻击者得到了服务器的密码了。

SesssetupX request [client][attacker]——>[server]

(密码被加密后传送)

攻击者将密码加密后通过 SesssetupX 请求数据报发送到服务器。

sesssetupX reply [client]<——[attacker]<——[server]

服务器发送 SesssetupX 应答报文,攻击者只是将它重定向到客户端。

[client]<—>[attacker]<—>[server]

攻击者继续在服务器与客户端之间重定向数据报,直到 SMB 会话结束。

5 安全认证的改进

Samba 服务器采用 Windows 工作站中的不可逆加密算法对用户密码进行加密,对不同的 Windows 工作站采用不同的加密算法。一般来说有:(1)LM 混编(LAN Manager 采用的混编技术),采用 DES 加密算法;(2)NT 混编(WindowsNT 采用的混编技术),采用 MD4 加密算法。改进的 Samba 安全认证模型如图 5 所示。

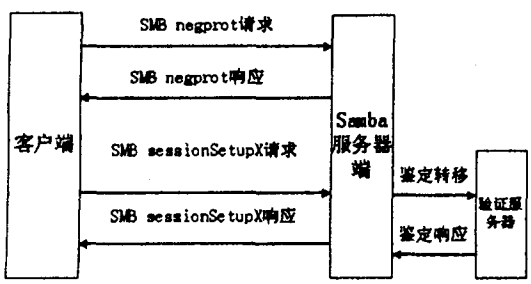

图 5 安全认证模型的改进

在带有验证服务器的基本安全模型中,客户端先向 Samba 服务器发送 SMB negprot(协议商定)请求,在 Samba 服务器返回的 SMB negprot 响应中获得 8 字节长的随机询问信息,然后,客户端通过填充 8 个空字节,将其 16 字节加密密码扩展到 21 字节,并将该 21 字节分成 3 个 56 比特长的 DES 密钥分别对询问信息加密,并将 3 个 8 字节的加密结果连接起来,在 SMBsessionSetupX 请求中返回给 Samba 服务器。最后,Samba 服务器把鉴定请求转移给验证服务器,验证服务器通过相同的计算验证加密结果。若验证成功,则返回 SMBsesssionSetupX 响应给客户端,若不能通过验证,则拒绝访问。由于这种鉴定认证过程避免了用户的明文密码或者加密密码直接在网络中传送,因此,可以防止 MITM 攻击,认为是安全的。

6 结束语

本文从 Samba 软件包所采用的 SMB/CIFS 协议出发,介绍了其实现文件共享的基本结构,探讨了其安全认证的模型。在异构网络环境下,Samba 很好地解决了不同平台下的网络共享问题,是个优秀的软件包,由于采取源码开放,具有重要的研究价值,对于网络计算机(NC)规范的制定有参考价值,对于研究 NAS(Network Attached Storage)网络附属存储的软件平台也有很好的借鉴作用。

参 考 文 献

[1] Christopher R. Hertel, Implementing CIFS the Common Internet File System, http://www.ubiqx.org/cifs.

[2] SAMBA Developers Guide, 14th November 2003, http://www.samba.org.

[3] SNIA CIFS Documentation Work Group CIFS Protocol version CIFS-Spec 0.9(Z).2001.3.

[4] (英)贝思(Baines,D.)著,沈立等译,Samba 技术内幕/Samba Block Book 北京:机械工业出版版,2000.10.

(上接第 111 页)

个词语的首字,即对后文 5 个以内字进行检查,如在 5 个字以内找到与该词语第二个字匹配的字,再进行对下文类似的检查,直到整个词完成匹配为此。如果用这种方法匹配出了默认词库中的词,即认为这封邮件包含这个词,然后再用正常的贝叶斯过滤算法进行处理。

还有一种常见的对抗内容过滤的方法是使用谐音字、近形字等,这种方法对贝叶斯过滤算法难以起到作用[3]。因为使用谐音字、近形字组成的词,经过词法分析后,会作为一个特殊的词提取出来,如果包含这种词的垃圾邮件第一次没有被贝叶斯过滤系统正确识别出来,经过使用者调整判断结果,下一次系统就能做出更准确的判断。

通过分析讨论可以看出,基于贝叶斯过滤方法的反垃圾邮件技术,针对垃圾邮件含有垃圾信息这一本质特征,并且可以解决不同用户对垃圾邮件判定标准不同的问题,是一种有较好发展前景的反垃圾邮件技术。

参 考 文 献

[1] Paul Graham, A Plan For Spam, 2002.08; Better Bayesian Filtering, 2003,01; Filters That Fight Back, 2003.08; Stopping Spam, 2003.08.

[2] Gary Robinson, Gary Robinson's Spam Rants, 2003.06.27 ~ 2003.10.02.

[3] William S. Yerazunis. The Spam-Filtering Accuracy Plateau at 99.9% Accuracy and How to Get Past It, Presented at the 2004 MIT Spam Conferenct, January 18, 2004.

[4] 刘群,汉语词法分析和句法分析技术综述,2002.08.