垃圾邮件问题及其处理方法

罗　浩[1]，方滨兴[1,2]，唐剑琪[2]

(1. 哈尔滨工业大学计算机科学与工程系　哈尔滨 150001；2. 国家计算机网络应急技术处理协调中心　北京 100029)

摘 要　随着互联网的发展,电子邮件作为一种通信方式逐渐普及,而垃圾邮件问题严重干扰了电子邮件的有效使用并产生了极大的社会影响。本文从技术角度和社会角度分析了垃圾邮件问题的成因,对现有的垃圾邮件处理方法进行了系统的分析,并从使用范围、维护手段、效率等多方面对处理方法进行了比较。

关键词 垃圾邮件;成因;处理方法

1　前言

随着互联网的发展,电子邮件作为一种通信方式逐渐普及,已变成人们生活、学习和工作中不可或缺的一部分,甚至改变了部分人的生活方式，同时电子邮件也推动着国民经济和社会发展的信息化,成为国家和社会进步的重要支柱[1]。

随着用户的增多和使用范围的逐渐扩大，保证邮件本身的安全以及防止电子邮件对系统安全性造成不良影响越来越重要。除了电子邮件本身的安全性以外,电子邮件也带来了其他安全问题。其中一个问题就是垃圾邮件对电子邮件系统的影响。

垃圾邮件或者叫做 Spam[2]、UBE(unsolicited bulk E-mail,未经用户许可的大量电子邮件)、UCE (unsolicited commercial E-mail,未经用户许可的商业电子邮件)。这里可以简单地将垃圾邮件定义为:向未主动请求的用户发送的电子邮件广告、刊物、其

参考文献

1 JUNOS port mirror configuration, http://www2.juniper.net/techpubs/software/junos/junos71/swconfig71-ervices/html/flow-monitoring-config17.html

2 GenieATM 产品解决方案, http://www.genienrm.com/ sc/products/ products-atm2320.htm

3 Omnipeek 产品解决方案, http://www.wildpackets.com/products/omni/omnipeek/overview

Provide ISP Security by Sliced Network Monitoring System

Ji Ye

(China Telecom Group Beijing Corporation Network Support Response Centre, Beijing 100032, China)

Abstract How to ensure service provider's network security through network monitoring system, especially how to use centralized monitoring in an extensive Internet service provider's system has been a tough problem that always puzzle the administrators. Current monitoring measures all limited in a certain aspect that prevent them from solving the problem ultimately. Through integrating the SNMP sampling, NetFlow analysis, probe analysis, and the unique port mirroring function of Juniper router, this paper is going to present a 'sliced' monitoring system, which is more effective in both detailed and summarized analysis, so that to laid a stable foundation for further research and operation to Internet service provider.

Key words 'sliced' network monitoring system, ISP, network security

(收稿日期:2006-01-12)

他资料或者不良信息(如色情、反动言论等);没有明确的退信方法、发信人、回信地址等的邮件;利用网络从事违反其他 ISP 的安全策略或服务条款的行为;其他预计会导致投诉的邮件。

据统计在 2002 年初,垃圾邮件占整个互联网邮件发送量的 16%,2003 年初变成 42%,而到了 2004 年初这个数据则变成了 60%[3]。在我国,根据中国互联网络信息中心 2005 年 1 月 19 日公布的第 15 次中国互联网络发展状况报告的统计,我国网民平均每周收到 4.4 封电子邮件(不包括垃圾邮件),收到垃圾邮件 7.9 封,可见我国垃圾邮件问题的严重程度。流出和流入我国的垃圾邮件已经带来了一定的负面影响。从流出看,根据 Spamhaus 项目的统计结果,中国居于美国之后是全球第二大垃圾邮件发送国。2003 年,欧美许多国家公开封堵来自中国的邮件,主要封堵的有 263、SINA 和 SOHU 等后缀的邮箱发出的邮件。也就是说,从中国流出的垃圾邮件已经损害了中国互联网产业的整体形象。从流入看,垃圾邮件已经成为色情、网络犯罪、网络诈骗等行为的一种重要的传播手段。同时,大量的垃圾邮件也浪费了大量的金钱并且严重影响了邮件系统的正常使用。据中国互联网协会公布的数据,2003 年全年发向中国邮件服务器的垃圾邮件达到 1 500 亿封,垃圾邮件耗费国内 GDP 超过 48 亿元人民币。对用户来说,大量的垃圾邮件严重地影响了用户对正常邮件的阅读;对于服务商来说,大量的垃圾邮件严重地消耗了服务器的资源和存储系统的存储能力,更严重的是如果过多的邮件超过了有限的邮箱容量,会导致用户或系统不能正常接收邮件。

愈演愈烈的垃圾邮件问题已经引起国家相关部门的高度重视。2004 年 1 月 30 日,由公安部、教育部、信息产业部和国务院新闻办四家单位,联合发布了《关于开展垃圾邮件专项治理工作的通知》,真正开始了反垃圾邮件行动。本文主要从技术手段入手,分析了垃圾邮件问题的成因,并对当前主流的和下一代垃圾邮件处理方法进行了讨论,以便读者能够深入了解垃圾邮件问题及其应对措施。

2 垃圾邮件产生的原因

垃圾邮件的产生既有技术上的原因,也有商业、政治上的原因等,具体归纳起来如下。

2.1 技术缺陷

现有邮件传输协议 SMTP (simple mail transfer protocol)[4]由 1982 定义的 RFC821 和 2001 年提出的 RFC2821 组成。提出 RFC821 的时候,由于互联网的用户还很少,没有考虑到会有人滥用电子邮件服务的问题。SMTP 本身是一个简化的邮件递交协议,缺乏必要的身份认证,这是造成垃圾邮件泛滥的原因之一。由于 SMTP 中,允许发信人伪造绝大多数的发信人的特征信息,如:发信人、信件路由等,甚至通过匿名转发、开放转发和开放代理等手段,可以近乎完全地抹去垃圾邮件的发信人的特征,这对于发现并制止垃圾邮件的传播造成了很大的困难。当前互联网上已经有难以计数的邮件服务器,考虑从协议层进行限制涉及到已有服务器大规模升级的问题,具体操作起来就会比较困难。中国互联网协会副秘书长黄澄清曾指出,我国垃圾邮件的产生,很大一部分是由于邮件服务器缺省设置开放转发(open relay)功能所产生。该功能对于普通用户可有可无,但缺省却可使用户的邮件服务器成为垃圾邮件的中转站,导致发出的邮件被拒收。并且,我国许多企业防范意识淡薄,未采取有效措施,致使企业的邮件服务器充当了转发者角色。

2.2 相关法律法规尚需进一步完善

2004 年 1 月 1 日,美国联邦《反垃圾邮件法》开始实施,但由于其采用了选择退出机制(opt-out,即必须接收者申请退出垃圾邮件发送者才不能继续向用户发送垃圾邮件),事实上纵容了垃圾邮件发送者。2004 年 7 月,在日内瓦召开的国际电信联盟(ITU)反垃圾邮件的主题会议上,我国信息产业部官员宣布中国将采取对垃圾邮件较为严格的选择加入机制(opt-in,即发送者必须事先取得接收者的同意才可以向其发送广告邮件)。在 2004 年 9 月 2 日中国互联网大会国际反垃圾邮件高层论坛上,信息产业部官员再次确认如果接收者没有明确表示同意接收而继续发送广告电子邮件,将作为一种发送垃圾邮件的行为予以禁止。信息产业部于 2004 年 3 月也发布了一个相关的行业标准《互联网广告电子邮件格式要求》(YD/T1310-2004),主要规定了广告电子邮件的词法、头部字段和消息体的格式以及头部字段的语法。《互联网电子邮件服务器管理办法》的起草中,也要求发送广告邮件时必须在邮件标题中添加“广告”等提示字眼。但是,这些法律和法规或者约束力不够,或者可操作性较差,尚不足以对垃圾邮件发送者产生足够的约束。

2.3 利益驱使

目前,数量众多的垃圾邮件中都含有大量的广告,通过电子邮件进行广告宣传,投入少、回报丰厚,使得许多广告商乐此不疲。比如在欧洲,发送一封电子邮件仅需 0.0005 欧元,比起庞大的广告费,这点费用实在微不足道。越来越多的商业需求以及电子邮件低廉的发送成本为垃圾邮件的滋生提供了土壤。

3 反垃圾邮件方法

反垃圾邮件方法从技术的角度来说可以分为邮件过滤方法、邮件验证方法和邮件协议增强方法。邮件过滤方法根据已知的垃圾邮件信息对接收邮件进行过滤,根据过滤结果判断邮件是否为垃圾邮件,是目前使用最为广泛的垃圾邮件问题解决方案。而邮件验证方法则不需要先验的垃圾邮件信息,直接根据接收到的邮件进行有效性验证,这些验证包括域名有效性验

证、发件人有效性验证等。邮件协议增强方法目前使用较少,一般是指在现有邮件协议的基础上增加额外的验证协议,来保证邮件信息的有效性。

3.1 邮件过滤方法

根据过滤对象的不同,垃圾邮件过滤方法可以分为基于邮件来源(即地址)的过滤方法和基于邮件内容的过滤方法。其中地址过滤依据发件人IP地址或电子邮件地址进行过滤,内容过滤则根据邮件的内容来过滤,主要判断邮件内容是否和已知的垃圾邮件内容相似。

(1) 邮件地址过滤方法

为了有效地拒绝来自垃圾邮件来源站点(包括被利用的垃圾邮件来源站点)所发来的垃圾邮件,最直接和有效的办法就是拒绝与该来源的连接。将确认后的垃圾邮件来源站点放入一个黑名单(black list),然后通过该名单来保护邮件服务器不受到黑名单中站点的侵扰是目前对抗日益严重的垃圾邮件的方法之一。

对于发现的发送垃圾邮件的IP地址进行屏蔽是一种消耗计算资源很少的技术手段,而且易于实施。从2003年8月开始,中国互联网协会反垃圾邮件协调小组开始不定期公布垃圾邮件黑名单。这种方法的缺点是需要不断维护IP地址清单。并且,因为垃圾邮件发送者经常修改他们的IP地址,并采用一个广泛的IP地址区间以逃避反垃圾邮件手段的检测,等到IP地址黑名单被公布时,该IP地址已经发送了大量的垃圾邮件。所以,该方法只是一个亡羊补牢的方法,在总体的垃圾邮件解决方案中仅起到补充作用。

为了解决定期公布黑名单的时效和滞后问题,实时黑名单(RBL)过滤技术应运而生。RBL也被称为DNSRBL,是对IP黑名单技术的一个改进。区别在于RBL是借助于第三方机构,由他们实时为用户提供黑名单的增加和删除,垃圾邮件的判断工作也是在Internet上进行的,不需要用户进行干涉和手动添加。目前国际上被广泛采用的RBL有ORDB(开放转发数据库)、DSBL(distributed sender black list)、NJABI、MAPS(邮件骚扰预防体系)、Spamhaus等。目前,中国反垃圾邮件联盟针对国内垃圾邮件的情况推出了实时黑名单服务,这是第一个面向国内用户的实时黑名单服务,主要面向中国国内的垃圾邮件情况,所甄选的黑名单地址也以国内的垃圾邮件反馈情况为主。可以说,其发布的垃圾邮件黑名单比国外的更适合中国国情。

RBL最重要的是黑名单IP列表的实时性和准确性。对于大型邮件系统来说,比较合理的做法是使用多个实时黑名单服务,综合实时黑名单的查询结果并与其他反垃圾邮件手段结合,最终得到此IP地址是否在发送垃圾邮件的判断。

(2) 邮件内容过滤方法

对邮件中的词语进行过滤是一个简单的阻断垃圾邮件的方法。词语过滤识别包含特定关键词的所有邮件,优点是实现简单,易构造、易实现,明显的缺点是词语过滤器需要经常升级维护,并且会产生较多误报情况。

基于关键词规则的评分系统是一个人工智能系统,对发现的每一个关键词赋予分数。分数越高,该邮件是垃圾邮件的可能性就越大。得分超过一定值时,该邮件将被分类为垃圾邮件。这样可以清除90%的垃圾邮件。这种方法和地址过滤面临同样的挑战,就是为使评分有效,规则必须经常更新,但是当前已经有第三方免费提供定义好的规则库,方便用户使用。

另一种针对邮件内容的过滤方法为邮件指纹过滤,即生成已知垃圾邮件的指纹,然后判断接收邮件的指纹是否与已知指纹相同或相似,并以此判断来件的性质。由于目前垃圾邮件发送工具大多具有自动对邮件体变形(morph)的能力,邮件指纹过滤的识别率很低,但是其精确度非常高,在样本准确的情况下几乎没有误报。

基于内容的垃圾邮件过滤技术与文本分类方法密切相关,垃圾邮件过滤本身相当于把接收到的邮件分成正常邮件和垃圾邮件两类,并且当前多种文本分类方法和机器学习理论也已经应用于垃圾邮件过滤。它们可以分为两类:基于规则的方法,从训练集中自动学习分类规则,如决策树、Boosting方法等[5];基于统计的方法,训练过程是一个统计学习过程,得到响应的分类器,如简单贝叶斯[6]、Memory-based方法[7]和支持向量机[8]等。这些邮件分类技术多采用一次性训练方法,不能动态学习用户的正常邮件的兴趣漂移和垃圾邮件变化,并且也不具备增量式在线学习能力。最为重要的是目前基于文本分类方法的垃圾邮件识别方法的识别率和精确度还有待提高。尤其是精确度对于垃圾邮件过滤问题极为重要,很小的错误率也会导致用户拒绝使用这种技术,这也是这种方法至今尚未大规模使用的原因之一;而另一个原因则是基于文本分类方法需要大量的计算资源来完成分词、匹配等操作,不适合在大规模邮件服务器上使用。基于上述两个原因,目前基于文本分类的垃圾邮件过滤方法大多应用于邮件客户端软件的垃圾邮件识别。

3.2 邮件验证方法

邮件验证方法作为无需垃圾邮件先验知识的识别手段,也是一种有效的解决垃圾邮件问题的方法,邮件验证技术主要包括发件地址验证和发件人有效性验证。

发件地址验证主要包括反向DNS验证和DNS MX记录验证。反向DNS验证是对邮件的来源IP地址采用反向DNS查找,如果反向DNS查找提供的域与邮件上的来源地址相符合,则该邮件被接受;如果不符合,则该邮件被拒绝。这种方法的优点是

开销比较小，但它有一个显著的缺点，就是目前很多反向 DNS 目录未被有效建立或无法正常建立，绝大多数情况下没有一个正确的反向 DNS 查找。在这种情况下，由这些域发送的邮件将被阻断，造成不可接受的高误报告率。同时，国内绝大部分邮件服务器也不能提供 DNS 反向解析，所以该方法也无法在国内广泛使用，只能作为反垃圾技术的参考。

DNS MX 记录验证是一项针对垃圾邮件发送者采用虚假的声明域地址或回复地址现象的有效阻断技术。系统在来源邮件地址的域上进行查找，如果该域没有一个有效的 DNS MX 记录，来源地址就是无效的，该邮件就被分类为垃圾邮件。这种方法的缺点是很多邮件列表服务器会被误判为发垃圾的服务器。

发件人有效性验证是指通过技术手段验证邮件的发件人是否为自动垃圾邮件发送工具，目前较为成熟的为挑战—响应(challenge-response, CR) 验证方法。CR 方法维护已经确认的非垃圾邮件发送源的白名单，当接收到来源不在白名单的邮件时，CR 机制会自动激活，给邮件的来源发送验证消息，需要邮件的发送方进行手工验证，如果通过验证，则来源会被自动加入白名单，并且信件会被接收，如果在有效期内验证未被通过，则信件会被当作垃圾邮件处理。由于垃圾邮件发送者采用工具批量发送，无法对验证信息进行响应，或者干脆无法接收验证信息，因而会被有效地过滤。虽然这种方法稳定有效，但这种方法需要改变用户的使用习惯，会对邮件产生明显的延时，并且这种方法需要对邮件服务器进行适当的改造，因而其使用范围有限。

3.3 邮件协议增强方法

随着国际上垃圾邮件泛滥的情况越来越严重，很多公司和组织提出了各种各样的方案试图从根本上解决垃圾邮件问题。目前大家公认的比较有可能成为国际标准的技术有以下三种：SPF(sender policy framework)、Domain Key、Sender ID。

(1)SPF

SPF 是一种基于 DNS 技术的反垃圾邮件技术。DNS 的基本功能是把对域名的访问解析成对 IP 地址的访问。DNS 的 MX 记录指定该域用于接收邮件的服务器的 IP 地址。SPF 通过发布反向 MX 记录告诉查询者从某个域发送邮件的服务器的地址。当收到从某个域发来的邮件时，接收者可以检查这个反向 MX 记录确认这封邮件是否从发件地址而来。发布 SPF 记录，本质上就是在 DNS 记录中增加一条记录，使邮件接收者可以验证邮件是否从该服务器生成的域所属的服务器发送出来。

(2)Domain Key

Domain Key 是 Yahoo 提出的一项反垃圾邮件技术。它的基本原理是：域名的拥有者生成一个公钥/私钥对。私钥部署在域名拥有者的发信服务器上，用这个私钥签名所有从该域发出的邮件。公钥通过 DNS 发布。当域名拥有者的用户通过转发服务器发送邮件时，存储私钥的服务器为每封邮件产生一个数字签名附加在邮件头上。接收方服务器从邮件头中取出邮件的 FROM 域和数字签名，根据 FROM 域从 DNS 服务器上取回该域对应的公钥，校验该签名是否由对应的私钥产生的。根据校验结果来判断对方是否在冒用其他人的 FROM 域。Domain Key 技术对现有的邮件系统的改造比较大，推广的难度也比较大。目前采用 Domain Key 技术的邮件服务器还不多。

(3)Sender ID

Sender ID 是微软提出的一项反垃圾邮件技术，它的基本思想是通过鉴别电子邮件发送者的身份来防止垃圾邮件，是在 SMTP 通信过程中对邮件来源进行检查的一种技术，属于连接控制型技术的 DNS 信息检查类。其基本的工作流程如下：用户通过 SMTP 把邮件发送到接收邮件服务器，接收邮件服务器通过 Sender ID 技术对发信人所声称的身份进行检查(该检查通过 DNS 的特定查询进行)，如果通过检查，发现发信人所声称的身份和其发信地址相匹配，那么接收该邮件，否则对该邮件采取特定操作，比如直接拒收该邮件。

由于微软为 Sender ID 采用的两项技术申请了专利，IETF 认为，微软决定要为 Sender ID 保留技术秘密的做法是无法让人接受的，因此不准备接受微软的这项技术提案。

3.4 反垃圾邮件方法的比较

表 1 列出了当前使用较为广泛的垃圾邮件处理方法的比较结果，表中所列方法均为当前拥有一定用户并已应用于不同场合的方法，不包括仅发布理论研究结果而没有实际应用的方法。

从表 1 中结果可以看出，在邮件服务器端目前使用广泛的为 RBL 方法和基于关键词规则的内容过滤方法。这两种方法均有第三方免费发布资源或直接利用现有资源，并且维护代价低，最为重要的是，这两种方法的误报率都很低，并且运行效率较高，因而能够大规模使用。基于文本分类的内容过滤技术运行效率较低，但其漏报率和误报率低的特点使之能够在邮件客户端上大规模应用。

4 结论

随着互联网的发展，电子邮件作为一种通信方式逐渐普及，而垃圾邮件问题严重干扰了电子邮件的有效使用并产生了极大的社会影响。垃圾邮件问题产生的主要原因是其存在巨大的商业价值和电子邮件协议本身的技术缺陷，解决垃圾邮件问题的关键在于技术的完善和相关法律的进一步到位。从电子邮件使

表 1　反垃圾邮件方法的比较

技术名称	使用范围	维护手段	使用手段	效果	应用状态
IP/E-mail 黑名单	适合大规模邮件服务器使用	纯手工维护,维护代价高	以服务器插件形式使用	漏报高/误报低,运行效率高	小规模使用
RBL	适合大规模邮件服务器使用	自动更新,第三方免费提供,维护代价低	以服务器插件形式使用	漏报中/误报低,运行效率高	大规模使用
基于关键词规则的内容过滤	适合中等规模邮件服务器使用	手工更新知识库(第三方提供)	以服务器插件形式使用	漏报低/误报低,运行效率中	大规模使用
基于文本分类的内容过滤	多用于客户端	手工更新知识库	以插件形式使用	漏报低/误报低,运行效率低	客户端大规模使用
邮件验证方法(MX 验证)	适合大规模邮件服务器使用	无需更新,利用现有资源	以服务器插件形式使用	漏报低/误报中,运行效率高	小规模使用
邮件验证方法(DNS 验证)	适合大规模邮件服务器使用	无需更新,利用现有资源	以服务器插件形式使用	漏报低/误报低,运行效率高	小规模使用
邮件协议增强技术	适合大规模邮件服务器使用	自动发布	需要更改邮件协议和服务器程序	——	小规模使用

用者和电子邮件服务提供者的角度来看,加强自律和自身管理是减轻垃圾邮件危害的有效手段。

目前减轻垃圾邮件影响的方法很多,从邮件服务提供者的角度来看,RBL 方法和基于关键词规则的内容过滤方法是切实可行的方案,这两种方法均有第三方免费发布资源或直接利用现有资源,维护代价低,误报率低,并且运行效率较高。而从邮件使用者的角度使用具有低漏报率和误报率的基于文本分类的内容过滤技术也是可以采用的有效技术手段。

参考文献

1 中国互联网络信息中心. 第 15 次中国互联网发展状况报告,http://www.cnnic.net.cn/

2 Search Mobile Computing. SPAM definition, http://searchmobilecomputing.techtarget.com/sDefinition/0, sid40_gci213031.00.html

3 Brightmail. Spam statistics, http://nospam-pl.net/pub/brightmail.com/spam-stats_March 2004.html

4 Postel J. Simple mail transfer protocol. RFC 821,1982

5 Carreras X, Marquez L. Boosting trees for anti-spam E-mail filtering. In: Proc Euro Conference on Recent Advances in Natural Language Processing (RANLP 2001), Bulgaria, Sep 2001

6 Sahami M, Dumais S, Heckerman D, *et al*. A bayesian approach to filtering junk E-mail. In: Proc of the AAAI Workshop on Learning for Text Categorization, Madison, Wisconsin, US, July 1998

7 Sakkis G, Androutsopoulos I, Paliouras G, *et al*. A memory-based approach to anti-spam filtering. Techreport DEMO 2001, National Centre for Scientific Research Demokritos, 2001

8 Drucker H, Wu D, Vapnik V N. Support vector machines for spam categorization. IEEE Transactions on Neural Networks,1999,20(5): 1048~1054

Spam Mail and Process Method

Luo Hao[1], Fang Binxing[1,2], Tang Jianqi[2]

(1. School of Computer Science and Technology, HIT, Harbin 150001, China;

2. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China)

Abstract The electronic mail is the most popular Internet application and the spam E-mail disturbs the usage of the E-mail gravely. This paper discusses the reason of spam mail problem and analysis the process methods of spam mail and compares the methods from the application range, the maintenance, efficiency and other points.

Key words spam mail, reason, process method

(**收稿日期**:2006-01-09)