网络出版时间：2015-09-22 10:02:37
网络出版地址：http://www.cnki.net/kcms/detail/10.1108.TP.20150922.1002.444.html

反垃圾邮件技术分析和发展研究

文/杜猛

摘要

本文通过对反垃圾邮件技术的发展阶段和关键技术进行分析，同时对未来反垃圾邮件技术的研究和发展方向进行探讨。

【关键词】反垃圾邮件关键技术发展研究

1 垃圾邮件的特征和危害

垃圾邮件(spam)，又称UBE(Unsolicited Bulk E-mail)，即未经接受者同意而大量散发的电子邮件。垃圾邮件主要具备以下一个或者多个特征：一是收件人无法拒绝收取的邮件；二是含有虚假的信息源、发件人、路由等信息；三是邮件内隐藏有病毒、木马等破坏性程序，或者含有大量广告甚至色情图片、政治色彩的信息；四是隐藏发件人身份、地址等信息。

垃圾邮件具有以下五个明显的危害性：

（1）占用网络带宽，影响邮件服务器的正常工作，降低网络的运行效率。

（2）浪费用户的宝贵时间和上网费用。

（3）收件人隐私遭到侵犯，个人信息泄露。

（4）对网络安全形成威胁，邮箱遭受病毒或恶意攻击，成为黑客攻击他人的工具。

（5）不良信息泛滥，造成政治危害和社会危害，尤其对未成年人产生不良影响。

2 反垃圾邮件技术的发展历程

第一代反垃圾邮件技术以过滤技术为主，包括了规则过滤、统计过滤和地址列表过滤等。这种技术应用最为广泛，可以在不作任何协议修改的情况下直接使用。

第二代反垃圾邮件技术被称为行为识别模式，对垃圾邮件的频次、时间、数据包头格式、IP地址、发送标识、协议类型等各类特征通过概率统计模型进行统计分析。这些特征能够针对带有诸如“同一时段频繁发送、动态IP地址”等特点来判断垃圾邮件。

第三代反垃圾邮件技术是电子邮件认证技术。该技术可以有效阻断垃圾邮件制造者利用漏洞伪造邮件发送地址的行为。但目前由于部署电子认证系统需要投入较高的软硬件成本，并且受限于多种因素，尚不能广泛应用。

第四代反垃圾邮件技术是多技术整合分层过滤。该技术是上述三代垃圾邮件处理技术的综合利用，可以在最大程度上实现反垃圾邮件的最大威力。

3 反垃圾邮件的关键技术

3.1 过滤技术

如前文所述，过滤技术基于邮件样本检测和规则匹配的原理，可分为规则过滤、合作式过滤和地址列表过滤三类。规则过滤技术通过设定好规则的匹配来实现过滤，虽然能有效阻止垃圾邮件，但误判率较高、比较容易被干扰信息影响。统计过滤是规则过滤技术的升级，通过使用统计规律计算垃圾邮件附加特征出现的可能性，来区分邮件的合法性，这种方法误判率较低。地址列表过滤技术是指根据建立的黑名单（Black List）和白名单（White List），分别是已知的垃圾邮件发送者和可信任发送者IP地址或者邮件地址，来判断是否接收电子邮件。

作为最有效的过滤技术，这里我们着重介绍Bayesian(贝叶斯)过滤技术。它首先对正常邮件和垃圾邮件进行分类学习，分别提取它们的特征值，对每个特征值进行赋分。在收到邮件时，对其提取特征值（比如标题、地址、附件、路径等信息），用之前学习到的特征值和分数对其进行赋分。在邮件中出现正常邮件的特征串，就赋予一个正分数，如果在邮件中检测到了垃圾邮件的特征串，就赋予一个负分数，最后根据总分来判断其是正常邮件还是垃圾邮件。Bayesian过滤器是用户根据所接收到所有邮件的统计数据来创建的，这意味着垃圾邮件发送者无法猜测出过滤器的配置情况，从而有效阻止垃圾邮件。

由于垃圾邮件数量庞大，内容特征变化快，过滤技术面临规则维护工作量大、误判率高、网络开销大的技术瓶颈。但是由于较为成熟，且较易部署，所以过滤技术是应用最为广泛的反垃圾邮件技术。

3.2 行为分析技术

行为分析主要在一定范围内对邮件流量进行监测并分析其变化规律，进而为识别垃圾邮件提供依据。根据监测点所处的位置，分别在邮件发送阶段和接受阶段对网络流量进行分析。如根据某邮件蠕虫爆发期局域网内域名解析流量和失败的SMTP连接数目急剧增加的情况，可以判断出垃圾邮件的变化规律，研究邮件病毒的扩散趋势。

3.3 逆向查询技术

如果能够更高效地区分伪造的邮件和合法的邮件，那么就能从根本上解决垃圾邮件问题，验证查询技术应运而生。为了限制发送者的虚假地址，一些系统要求验证发送者邮件地址进行验证。上世纪九十年代初，出现了邮件交换纪录（MX），当发送邮件的时候，邮件服务器通过查询DNS的MX纪录来找到接收者的域名。逆向查询解决方案就是定义逆向的MX纪录（RMX），用来判断发送邮件的域名和IP地址是否对应。由于垃圾邮件的地址通常不会来自真实的RMX地址，因此可以判断是否非法。

4 最新技术与展望

4.1 意图分析技术

许多垃圾邮件标题和信体都与合法邮件一样，但是信体内有诱使接收者点击的URL地址，而URL地址链接的内容是其真正意图。意图检测技术就是对URL进行检查，根据链接的内容来判断是否为垃圾邮件，从而识破发送者真实意图，阻断邮件。

4.2 图片识别技术

针对图片垃圾邮件的技术有邮件指纹识别技术、ocr识别技术以及之后的第三代图像防御技术。图片垃圾邮件的发送者企图使用动态gif图像，或者用横线、符号和其他图像模糊图片内的文字。Ocr引擎则具备动态gif文件分析功能和模糊文本识别技术。

4.3 发件人特征识别技术

鉴于垃圾邮件制造者的伪装术越来越高，出现了针对“好人”身份欺骗的特征识别技术，首先要验证发信者身份并预测其行为，这其中包括列举垃圾邮件制造者的行为以及加强不依靠身份验证进行辨认的措施。

5 结语

当前，垃圾邮件已成为全球各国和互联网业界共同面临的严重问题，应当采用管理与技术并重方式，以先进的技术手段为基础，以完善的管理制度和法律法规为依托，不断加强国际合作，对垃圾邮件保持高压态势。未来反垃圾邮件的行动主要包括如下四个方面：

（1）加强互联网立法，制定严格法律严惩垃圾邮件制造者。

（2）设计更为安全和完善的邮件体系。

（3）加强技术研发和人才培养，不断提升核心技术能力。

（4）加强宣传和行业自律，净化网络空间。

参考文献

[1] 郑炜，沈文，张英鹏．基于改进朴素贝叶斯算法的垃圾邮件过滤器的研究[J]．西北工业大学学报，2010(03)．

[2] 王斌．潘文锋，基于内容的垃圾邮件过滤技术综述[J]．北京：中文信息学报，2005 19(5)．

作者单位

山西省通信管理局　山西省太原市　030006