一种快速的基于URL的垃圾邮件过滤系统*)

詹 川[1] 卢显良[2] 侯孟书[2] 邢 茜[2]

(重庆工商大学商务策划学院 重庆 400067)[1] (电子科技大学计算机科学与工程学院 成都 610054)[2]

摘 要 垃圾邮件是当前Internet上关注一个焦点问题,随着垃圾邮件的伪装技术的不断更新,以前主要的几种垃圾邮件过滤技术面临着新的挑战。针对目前垃圾邮件大都含有要宣传网站的URL地址的特点,我们构建了一个基于URL的垃圾邮件过滤系统,通过查询邮件中是否含有URL黑名单中的URL地址,来识别垃圾邮件。在URL地址查询过程中,采用Hplf散列函数来加速查询。通过试验测试,该系统能有效地减少垃圾邮件的数量。

关键词 URL,垃圾邮件,邮件过滤,散列函数

A Rapid URL-Based Spam Filtering System

ZHAN Chuan[1] LU Xian-Liang[2] HOU Meng-Shu[2] XING Qian[2]

(Strategical Planning College,CTBU,Chong Qing 400067)[1] (College of Computer Science and Engineering, UESTC of China,Chengdu 610054)[2]

Abstract The cat and mouse game between spammer and anti-spam vendors has continually evolved. As the latest generation of spam incorporates sophisticated tactics, previous spam filtering technologies face a new challenge. In the term of spammers using URL-based anti-filtering techniques, we design a rapid URL-based spam filtering system. The system identifies spams through URL queries and uses Hplf hash function to accelerate query speed. It is proved that the system is an effective method of anti-spam by experiment.

Keywords URL, Spam mail, Spam filtering, Hash function

1 前言

电子邮件由于快速、高效、低成本的特点使其在互联网中得到广泛的应用,大大方便了人们生活、工作和学习。近年来,一些公司或个人利用电子邮件的特点,为了商业或政治等目的,在互联网上大批量散发含有商业广告或不良信息的垃圾邮件。这种不请自来对邮件用户毫无用处的垃圾邮件浪费了大量的网络资源、个人的时间、精力,对个人和公司造成重大损失。据美国的Brightmail公司统计Internet中的电子邮件有60%是垃圾邮件,上海艾瑞咨询公司统计[1]2004年中国邮件用户平均每周收到31.9封邮件,其中19.3封为垃圾邮件。据估计,全世界的企业每年大概要花费80亿~100亿美元来解决垃圾邮件问题。

本文针对垃圾邮件问题,第2部分首先讨论了目前主要使用的几种垃圾邮件过滤技术,分析了它们的优缺点;然后在第3部分,先介绍了目前垃圾邮件伪装技术的趋势,针对目前垃圾邮件的新特点,构建了一个行之有效的基于URL垃圾邮件过滤系统;第4部分测试了基于URL垃圾邮件过滤系统性能;最后为结论。

2 主要的垃圾邮件过滤技术

2.1 基于关键字的过滤

这种技术搜索邮件头主题行或者邮件正文中是否含有预设的关键词,如挣钱、伟哥等,如含有,则认为此邮件为垃圾邮件并过滤。这种技术非常简单,现在的一般的邮件服务提供商为客户提供关键字过滤的功能。它需要客户自己预设关键字;随着垃圾邮件伪装技术的提高,垃圾邮件发送者避免使用或篡改常用的关键字来逃避基于关键字的过滤;同时,这种技术的误判率(false positive)非常高。

2.2 基于黑白名单的过滤

黑白名单是一个简单有效的过滤方法,把已知的垃圾邮件发送者列在黑名单中,把自己知道的正常发送者列在白名单中,当邮件到来时,检查邮件发送者,如果在白名单列表中就直接认为是正常邮件,如在黑名单中则认为是垃圾邮件。但这种方法存在以下问题:需要不断更新黑白名单数据;可在邮件头中的伪造发送者来逃避黑名单的过滤,这种发送者伪造技术在当前垃圾邮件中被普遍应用;当一个新的邮件发送者发送的邮件到来时,易被白名单误判。

2.3 基于规则的过滤

采用此种技术的垃圾邮件过滤器一般通过分析接收到的邮件是否匹配所设定的规则来判断该邮件的性质。若邮件匹配某种设定的规则,则该邮件的权值有可能增加或者减少。当邮件的权值超过了某一特定的阈值,则将其视为垃圾邮件而过滤它;否则认为是合法的。但因为垃圾邮件发送者采用新的伪造技术,所以需要不断更新过滤规则,对客户更难的是要懂得如何总结出规则以及如何设定和修改规则。

2.4 基于贝叶斯的过滤

贝叶斯公式过滤[2]是利用数学统计方法,首先提取邮件的特征,选出最具代表性的特征单词,统计特征词在垃圾邮件中出现的概率,然后应用贝叶斯公式计算出这封邮件是垃圾邮件的概率。这种方法的好处是:不需要人为地制定规则,能自动智能地识别垃圾邮件;它从邮件内容整体上来进行判断,

*)信息产业部电子工业生产发展基金资助项目,编号:[2002]11006。詹 川 博士研究生,主要研究方向:反垃圾邮件过滤,信息安全,人工智能,数据挖掘。卢显良 教授,博士生导师,主要研究方向:计算机网络,分布式系统,信息安全。侯孟书 博士研究生。邢 茜 硕士研究生。

识别准确率较高。但需要收集足够多的邮件样本集和前期的学习，并且随着垃圾邮件采用的词越来越隐晦、在邮件中随机插入无关的句子，造成了这类方法的误判率的增高。

2.5 基于签名的过滤

这种方法[3]是每封新到的邮件通过特定的算法，生成独自的签名，检查其签名是否与获得的已知垃圾邮件库中签名相似，如果相似，则判定新邮件为垃圾邮件。这种方法可以准确高效地过滤大量重复发送的垃圾邮件，但是只能过滤已知的垃圾邮件，不能智能识别过滤新的垃圾邮件。

3 基于 URL 的垃圾邮件过滤系统

3.1 垃圾邮件伪装技术趋势

为了逃避过滤，垃圾邮件发送者增加了垃圾邮件的适应性和复杂性，比如通过设置虚假的域名和更改主题行内容来逃避过滤。使用 HTML 格式是目前垃圾邮件发送者最常用和最有效的抗过滤的方法。根据 Wall Street 杂志于 2003 年 6 月统计，超过 80%的关于成人内容的垃圾邮件应用的都是 HTML 格式。选择使用 HTML 格式主要有以下三个原因：

1. 具有丰富的表现力。可在 HTML 文件中嵌入图片、动画、声音及超链接使其宣传更加具有感染宣传力。

2. 具有反馈用户信息的能力。在 HTML 文件中嵌入反馈源代码，当用户下载邮件时就激活代码，给垃圾邮件发送者确认是否一个任意发送的邮箱为一个有效在用的邮箱。

3. 使垃圾邮件具有随机性和多样性。通过在 HTML 中插入无效的 HTML 标记，在白色背景下插入白色的文字，使用 HTML 表格形式等使 HTML 源代码具有多样性和随机性，但是显示给用户的内容却是相同的文本。

在 HTML 格式的垃圾邮件中常常会包含 URL 的超链接。垃圾邮件会诱惑用户在读了邮件后，进一步去点击它们包含的超链接。在链接的网页上，往往是他们提供的产品或服务的广告。据 MicroSoft 公司的 Geoff Hultent 统计分析 Hotmail 用户反馈的情况[4]，2003 年用户收到垃圾邮件中采用 URL 抗过滤技术的邮件占 22%，到 2004 年增加到 27%。使用 URL 抗过滤技术的垃圾邮件为了蒙骗邮件用户，发送者会采用在邮件中插入大量无关，似是合法邮件的文字，致使基于内容的过滤器误判为合法邮件或者邮件中就含有简短的一两句，使基于内容的过滤器难以判断；随机插入文字，使邮件貌似不同的邮件，只有超链接 URL 地址相同，致使基于签名的过滤器无能为力；显示为合法的网址，但实际链接的却是另一个网址。对这类特点的垃圾邮件用上面介绍的过滤方法都得不到满意的解决。

3.2 系统结构

针对目前垃圾邮件采用 URL 方式来抗过滤的特点，我们提出基于 URL 来过滤垃圾邮件的机制。因为垃圾邮件的伪装相对简单和低成本的，所以垃圾邮件是多变的；而获得一个 URL 地址和建立一个网站是需要成本和时间的，因此 URL 地址是相对固定不变的。我们通过捕获含有的 URL 信息能更容易、更准确地识别具有 3.1 节描述特点的垃圾邮件。系统基于的假设是认为如果新来的邮件中含有从已知垃圾邮件中获取的 URL，则认为此邮件为垃圾邮件。这种判断有点武断，但是在一定范围内相当有效。

基于 URL 的垃圾邮件过滤系统如图 1 所示，首先需要收集已知垃圾邮件中链接的 URL 地址。系统中垃圾 URL 地址主要来源于三个方面：

1. 安装的蜜罐捕捉到的垃圾邮件中提取的 URL 地址；
2. 用户反馈的垃圾邮件中含有的 URL；
3. 可靠的第三方提供的 URL 黑名单列表。

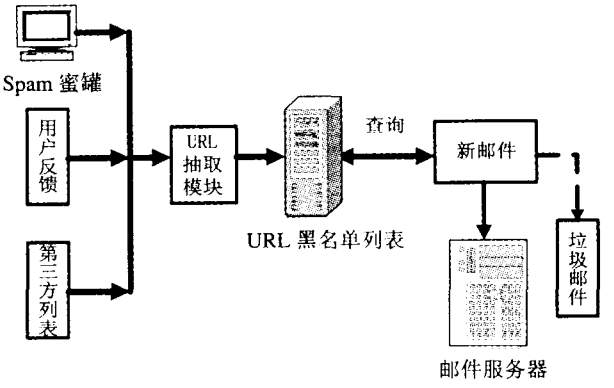

图 1 基于 URL 的垃圾邮件过滤系统结构

获得的垃圾邮件样本，需要通过 URL 抽取模块处理，首先分析该垃圾邮件头 content_type 是否为 html 格式的；如是，则进一步分析邮件主体中是否含有〈A HERF="。。。"〉字段；如有，则提取双引号中的字符串，获得垃圾邮件中真正指向的 URL 地址。

为了加快 URL 地址的比对和查询速度，我们系统应用了 Hflp 散列函数[5]，如图 2 所示，把字符串的匹配转换成数字大小的比较。Hflp 函数把不定长的 URL 地址散列成一个固定长度的数字，它具有很好的信息查询的可靠性和高速性。然后，把获得已知 URL 散列值按序的加入 URL 黑名单列表服务器中。

```
unsigned int Hflp( char * , int size )
{
unsigned int n = 0;
char * b = ( char * )&n;
for(int i=0; i<strlen(url); i++)
b[i%4]^ = url[i];
return n%size;
}
```

图 2 Hflp 散列函数

在每封新邮件进入邮件服务器前，分析邮件，进行 URL 地址提取。

如获得 URL 地址，则查询在 URL 黑名单列表中是否存在，如果匹配，则把这邮件作为垃圾邮件，否则存入邮件服务器。

4 测试

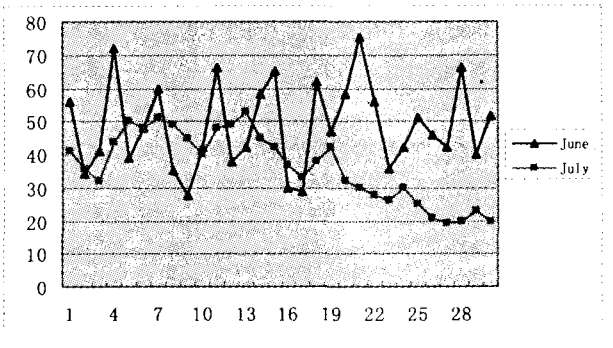

图 3 安装系统前后收到垃圾邮件数比较

（下转第 66 页）

本算法能很好地抗剪切攻击，这正是因为本文嵌入了多个水印版本的结果。

C. 缩放攻击。实验对嵌入水印后的图像分别进行先缩小到原来的 1/4 和 1/2 后再利用双三次插值放大到原图像大小；先放大到原图 2 倍、3 倍和 4 倍后，再缩至原图大小。实验结果如表 3 所示，可见，算法对缩放具有很强的鲁棒性。

表 3

缩放比例	1/4	1/2	2	3	4
NC	0.68957	0.98896	0.99877	0.99877	1

D. 叠加椒盐噪声。对嵌入水印后的图像叠加概率为 0.01 的椒盐噪声，盲检测出的水印，NC= 0.97178，恢复水印的能力强，这证实了本文的水印算法对椒盐噪声有很好的抵抗能力。

E. 滤波。对嵌入水印后的图像进行 3×3 的中值滤波处理，盲检测出水印，NC=0.98896。对嵌入水印后的图像进行高斯低通滤波后，NC=1。可见，本文的算法能很好地抵抗滤波操作。

结论 本文提出了一种新的数字水印算法，总结其特点如下：

(1) 选取 JPEG 量化取整后连续非零元素个数最多的区域来嵌入水印的，是自适应的，并且检测时不需要原始图像。

(2) 利用 JPEG 量化阶段的舍入误差来嵌入水印，因此能很好地抗 JPEG 压缩，并且对系数的修改很小，所以有很好的不可见性。

(3) 水印是经位扩展和调制产生的，并且嵌入了多个水印，所以能很好地抗剪切攻击。

参 考 文 献

1 Hartung F, Kutter M. Multimedia watermarking techniques. In: Proc. IEEE, 1999, 87(7):1079～1107

2 Schyndh R, Zrikh A, Osborne C. A digital watermark. In: Proc. IEEE Int. Conf. on Image Processing, 1994, 2:86～90

3 Barni M, Barolini F, Cappellini V. A DCT-domain system for robust image watermarking. Signal Processing, 1998, 66(3):357～372

4 黄继武，SHI Y，程卫东. DCT 域图像水印：嵌入对策和算法. 电子学报，2000，28(4)：57～60

5 Hsu C, Wu J. Hiding digital watermarks in image. IEEE Trans. on Image Processing, 1999, (1):58～68

6 Lin S, Chen C. A robust DCT-based watermarking for copyright protection. IEEE Trans. on Consumer Electronics, 2000, 46:415～421

7 Kunder D, Hatzinkos D. A robust digital image watermarking method using wavelet-based fusion. in Int. Conf. on Image Processing, 1997, 3:544～547

8 Deng F, Wang B. A novel technique for robust image watermarking in the DCT domain. IEEE Int. Conf. Neural Networks &Signal Processing, 2003, 2: 1525～1528

9 Wu J, Xie J. Adaptive image watermarking scheme based on HVS and fuzzy clustering theory. IEEE Int. Conf. Neural Networks &Signal Processing, 2003,2:1493～1496

10 Eyadat M. Factors that affect the performance of the DCT-block based image watermarking algorithms. In: Proc. IEEE Int. Conf. on information technology: Coding and Computing, 2004, 1:650～654

11 Chen L, Lin J. Mean quantization based image watermarking. Image and Vision Computing, 2003, 21(8):717～727

12 Karybali I, Berberidis K. Blind image-adaptive watermarking. In: Proc. IEEE Int. Conf. on, Electronics, Circuits and Systems, 2003,2:894～897

13 Yu D, Sattar F, Razul S G. Transparent robust information hiding for ownership verification. In: Proc. IEEE Int. Conf. on, Acoustics, Speech and Signal Processing, 2004, 3:401～404

14 Bertran M, Delaigle J F, Macq B. Some improvements to HVS models for fingerprinting in perceptual decompression. In: Proc. 2001 Int. Conf. on Image Processing [C], 2001,2:1039～1042

15 Cox I J, Kilian J, Kilian J, Leighton F T, Shamoon T. Secure spread spectrum watermarking for multimedia. IEEE Trans. on Image Processing, 1997,(12):1673～1686

(上接第 56 页)

我们把基于 URL 垃圾邮件过滤系统，安装在西南教育网信息中心的邮件服务器上，统计了在安装系统前的一个月(6 月)和安装系统后的一个月(7 月)5 个邮箱每天收到的垃圾邮件数，如图 3 所示。6 月 5 个邮箱总共收到 1456 封垃圾邮件，平均每天 48.5 封，7 月总共收到 1097 封，平均每天 36.6 封。安装基于 URL 垃圾邮件系统后，平均每天比安装前减少了 24.5%的垃圾邮件，随着使用时间的增加，URL 列表服务器中收集的垃圾 URL 地址也增多，系统的过滤效果更好，收到的垃圾邮件逐渐减少，在 7 月底的几天内每天收到 20 封左右的垃圾邮件。

结论 随着垃圾邮件发送者的伪装技术增强，使用 HTML 格式含有 URL 的垃圾邮件不断增多。针对其特点，我们构建的基于 URL 的垃圾邮件过滤系统，对此类垃圾邮件具有很好的过滤效果，特别是对于短的、含有 URL 超链接的垃圾邮件；在邮件中大量随机插入无关，像是合法邮件，含有 URL 超链接的垃圾邮件；以上两类垃圾邮件过滤方法难以准确、高效地过滤。本文采用的 URL 地址字符串散列函数，加快了垃圾邮件的查询速度，减少了邮件的处理时间。试验测试显示出基于 URL 的垃圾邮件过滤系统是种有效的垃圾邮件过滤方法。

参 考 文 献

1 上海艾瑞市场咨询公司. 2004 年中国反垃圾邮件研究报告，2004，3

2 Sahami M, Dumais S, et al. A Bayesian Approach to Filtering Junk E-Mail. Learing for Text Categorization -Papers from the AAAI Workshop. Madison Wisconsin, 1998

3 Levitt M, Burke B E. Choosing the Best Technology to Fight Spam. White Paper. Commtouch Company, April 2004

4 Hulten G, Penta A, Opalaks G, Anavm M. Trends in spam products and methods. Http://www.ceas.cc/papers-2004/165.pdf

5 李晓明，凤旺森. 两种对 URL 的散列效果很好的函数. 软件学报，2004,15(2)