一种互联网垃圾邮件综合过滤方案

侯立铭,彭 伟

(国防科技大学 计算机学院,湖南 长沙 410073)

摘 要:垃圾邮件是互联网上亟待解决的问题。介绍了几种典型的垃圾邮件过滤技术,提出了一种结合邮件过滤和病毒检测技术、可以个性化定制过滤需求的综合过滤方案。相比于已有的方案,文中提出的方案具有同时检测病毒、过滤垃圾邮件和个性化过滤的优点,可以更加有效地解决邮件安全和个性化过滤的问题。

关键词:垃圾邮件; 过滤; 病毒检测

中图分类号:TP393.098 **文献标识码**:A **文章编号**:1673-629X(2007)04-0117-03

An Integrated Spam-Filtering Approach for Internet

HOU Li-ming, PENG Wei

(Computer School, National University of Defense Technology, Changsha 410073, China)

Abstract: Spam is a very important problem on the Internet. In this paper, propose a novel spam-filtering approach which integrates e-mail filtering, virus detecting and personalized filtering technology on the basis of analysis of traditional spam-filtering approaches. Compared with other approaches, this approach can filter spam as well as detecting virus, and users may customize the filter rules expediently. Analysis shows that this approach can work effectively on encountering mail security and filtering customization.

Key words: spam; filter; virus detection

0 引 言

近年来垃圾邮件问题日益严重,根据《中国互联网协会反垃圾邮件规范》的定义,以下四种情况属于垃圾邮件:收件人事先没有提出要求或者同意接受的广告、电子刊物、各种形式的宣传品等宣传性的电子邮件;收件人无法拒收的电子邮件;隐藏发件人身份、地址、标题等信息的电子邮件;含有虚假的信息源、发件人、路由等信息的电子邮件。垃圾邮件占用大量网络传输、存储和计算资源,影响网络的正常运行,对用户的正常工作造成严重干扰,有的甚至危害国家利益。

中国互联网信息中心(CNNIC)的最新统计报告显示,全国互联网用户平均每周收到85.3封邮件,其中有57.5封是垃圾邮件[1]。垃圾邮件的内容也在不断发展变化,除了常见的广告、色情信息,还常常夹带有蠕虫和病毒。目前的垃圾邮件与病毒有不断融合的趋势,病毒发送者在被感染的计算机上开后门,而这些被感染的计算机则又被用来大量发送垃圾邮件。不同类别的邮件用户对垃圾邮件定义的理解、对邮件过滤要求也存在很大差异。因而对反垃圾邮件的技术提出了邮件内容过滤、反病毒和个性化过滤相结合的新要求。

1 常用的邮件过滤方法

目前,有多种算法可以用来过滤垃圾邮件,并且已经在实际应用中发挥一定的作用,但它们本身又各有优缺点。

1.1 白名单和黑名单

“白名单”是用户或MTA(Mail Transfer Agent,邮件传输代理)设置和维护一系列名单[2,3],上面记录着可以信任的IP地址和域名,从这些名单发送过来的邮件都被认为是合法邮件。“黑名单”与“白名单”相反,这是一个不受欢迎的IP地址和域名的列表,用户希望阻止它们发来的邮件。过滤系统在处理新到达的邮件时,首先查看邮件头部的发送方地址,对于地址处于白名单中的邮件将全盘接收,而对于处于黑名单中的邮件则直接拒收。优点是速度快、效率高。缺点是维护困难、对数据准确度要求很高。目前在黑名单技术上最流行的是实时黑名单(Realtime Blackhole List,简称RBL)技术,由一些著名的志愿者组织在他们的网站上维护着一系列的IP级的黑名单,它们或者是垃圾邮件

收稿日期:2006-07-05

作者简介:侯立铭(1974-),男,山西太原人,硕士研究生,研究方向为网络信息安全;彭 伟,副教授,研究方向为网络信息安全。

发送者的地址,或者是那些具有严重安全漏洞的邮件服务器地址。任何 ISP 都可以订阅这些服务,使这类邮件在到达之前就自动被拒绝。目前比较值得信任的组织有 Spamhaus 以及中国反垃圾邮件联盟维护的 RBL 服务。

1.2 基于规则的过滤技术

基于规则的过滤技术指邮件服务器或用户可制订一些硬性规则,拒收或过滤符合规则的邮件。简单的规则包括拒绝主题中包含某个关键词的邮件,创建拒收邮件地址列表等;复杂的规则包括支持正则表达式匹配。它是通过与既定的规则相比较来判定是否为垃圾邮件,这些规则包括:特别的词语,如"快速致富"等;伪造的信件头,如不合理的日期等;大量的 URL 链接等。基于规则方法的优点是规则可以共享,因此它的推广性很强。存在的问题在于:一是采用规则匹配,尤其是正则表达式匹配,对用户要求较高,且易用性不好;二是规则的制订总是落后于垃圾邮件特征的变化时效性较差。

1.3 基于贝叶斯算法的内容过滤技术

贝叶斯算法是基于统计方法的垃圾邮件过滤技术[3~5]。基于统计的方法的优点就是分类器由程序自动学习出来,只要及时更新样本学习集就可以使分类机更新的速度跟得上垃圾邮件出现的速度,即它的时效性很强。贝叶斯算法的基本思想是通过对邮件头部和邮件信体中的单词进行概率计算,从整体上判断是否为垃圾邮件。单词的概率计算依赖于已知的垃圾邮件和正常邮件中单词出现的频率来完成,因此必须经过一段时间的学习之后才能开始为用户工作,它的工作流程包括两个阶段:

(1)学习阶段。学习阶段对收集到的不同类别的垃圾邮件集进行学习,统计出不同垃圾邮件集中的特征关键词的出现频率,并计算关键词在不同垃圾邮件集下的概率值。然后,根据统计出来的关键词和关键词的概率统计值。学习阶段不是在系统每次启动的时候进行,而是当数据库中垃圾邮件积累到一定数量或者系统运行一定时间的时候进行,也可以由管理维护终端触发进行学习。

(2)判别阶段。当一封新邮件到达时,系统需要对信件全部内容进行分词和选词,然后根据学习到的单词库中的信息,计算整个邮件或其中出现频率较高单词部分的概率,根据设定的阈值,最终判断该信件是否为垃圾邮件。贝叶斯算法是基于内容的,具有自学习、自适应的性质,即分类器的知识会随着垃圾邮件内容的变化而更新。贝叶斯算法的学习结果是依赖特定的学习集,学习集的内容对学习的结果有较大影响,因此,学习集的差异是达到不同用户过滤需求目的的关键。

2 垃圾邮件的综合过滤方案

随着携带病毒邮件增多和用户对垃圾邮件理解上的差异,上述的技术的单一应用显现出以下不足:一是单一技术的应用很难保证邮件的查全率和查准率;二是不能检查邮件是否携带病毒,对带有病毒的邮件无法进行查杀处理;三是过滤结果单一,很难满足各类用户的不同需求。比如,同样一封推销商品的邮件,对于从事网络营销的用户可能就不会认为是垃圾邮件,而对于其他用户而言可能就会认为是垃圾邮件。为了解决以上 3 个问题,在对邮件进行内容过滤的同时,一是将现有的过滤技术有机组合;二是加入了对邮件的病毒检测模块;三是加入了对用户个性化需求反馈的学习,并确保用户制定的策略具有较高优先级。采用多种过滤技术并不是将它们简单叠加,而是在邮件过滤的不同阶段发挥不同技术的优长;在病毒检测模块中主要依靠对已知病毒特征码的匹配,辅助以动态智能推理机对动态未知病毒的检测[6,7]。在个性化反馈的过程中,用户可以对过滤系统的各个阶段的规则进行个性化修订。若用户没有对过滤提出修改意见,则使用系统默认策略。综合过滤方案的目的是既保护正常邮件的收发,又可以高效地过滤垃圾邮件、查杀带有病毒的邮件,并且能使用户根据自身要求个性定制过滤规则,满足不同用户的特定需求。

2.1 方案结构

1) 邮件过滤模块。

在邮件过滤模块中,系统对邮件进行多层过滤。主要集成黑白名单、规则过滤以及基于贝叶斯算法的内容过滤等。各种过滤算法过滤的侧重点不同:其中黑白名单检查邮件的信封部分;规则算法对邮件的信头信体进行简单的规则匹配;贝叶斯算法在上述过滤完成后,在预先学习的基础上,将对邮件的内容进行计算和判断。最后,系统将过滤的信件按照合法邮件、垃圾邮件分类存放,用户对系统过滤后的合法邮件和垃圾邮件分别审阅并将系统的判别结果反馈,反馈回的知识用于修订黑白名单、关键字、规则以及对贝叶斯过滤引擎的学习。

2) 病毒过滤模块。

病毒过滤模块的核心是运用智能推理机技术,分为静态和动态两种方式。静态方式是在系统虚拟机中展开的病毒文件中依据已知的病毒特征码对邮件进行扫描。动态方式主要针对未知病毒的检测,由于病毒在传染和破坏时都有许多行为特征,动态病毒检测部

分对未知的病毒进行行为特征分析,根据行为分析,对具有多种可疑行为的严重程度进行评估,按照系统设定的危害程度阈值判断,将可疑文件标记、隔离,并记入日志,同时提取特征码更新到病毒库。

3)个性化设置模块。

系统个性学习模块提供以下功能:一是可以个性定制过滤策略,用户既可以对整个邮件过滤过程制定过滤策略,也可以对特定过滤模块进行策略调整。比如:在新邮件到达后,用户可以设定为只依据黑白名单进行检查,而无需更多过滤检测;也可以选择是否将信体部分进行病毒检测和贝叶斯过滤等等,还可以选择各种策略以及策略组合。二是用户对黑白名单、关键字及规则的个性化设置。三是对贝叶斯过滤器的个性化学习。不同的用户可以将自己对垃圾邮件的不同理解反馈给贝叶斯过滤器学习,从而达到相同的过滤系统,相同的过滤流程,不同的过滤结果,满足用户个性需求。

4) 策略及配置管理模块。

策略管理模块提供两种策略的管理:一种是系统的管理策略,用于邮件系统管理者对邮件过滤策略的管理;二是用户的个性化过滤策略管理,用于对用户设置的各种策略的管理。策略管理模块根据系统负载、用户安全或过滤效率以及个性化需求,实时调整过滤策略。配置管理模块提供各个模块的运行参数设置以及系统日志、审计和流量监控等辅助管理功能。

2.2　工作流程

当一封新的邮件到达时,它需要经过黑白名单、基于规则的算法和贝叶斯算法以及病毒检测模块的层层过滤。根据用户的过滤策略,可以选择仅邮件过滤或是邮件与病毒过滤,系统首先查看邮件的具体发送方地址,如果是被列在白(或不在黑)名单中,则该信件会被直接送到用户邮箱,否则用规则算法对信体内容进行简单的过滤,如果不是垃圾邮件则直接送入用户邮箱,否则作为病毒检测的输入,进行病毒检测,如果没有病毒,则按照策略的配置对邮件进行贝叶斯算法进行判断,最后将系统过滤得到的合法邮件送用户邮箱,垃圾邮件送垃圾邮件文件夹供用户审阅。用户可以根据自己需求对黑白名单、关键字及规则进行添加或删除,在对收到的合法邮件和垃圾邮件审阅后,可将系统误判的邮件反馈到贝叶斯过滤引擎学习。用户还可以实时调整过滤策略,根据安全或效率需求,选择应用部分或全部模块。图 1 是按照用户选定所有的过滤模块的所有功能绘制,在实际使用中,过滤流程会因为用户不同设置而不同。

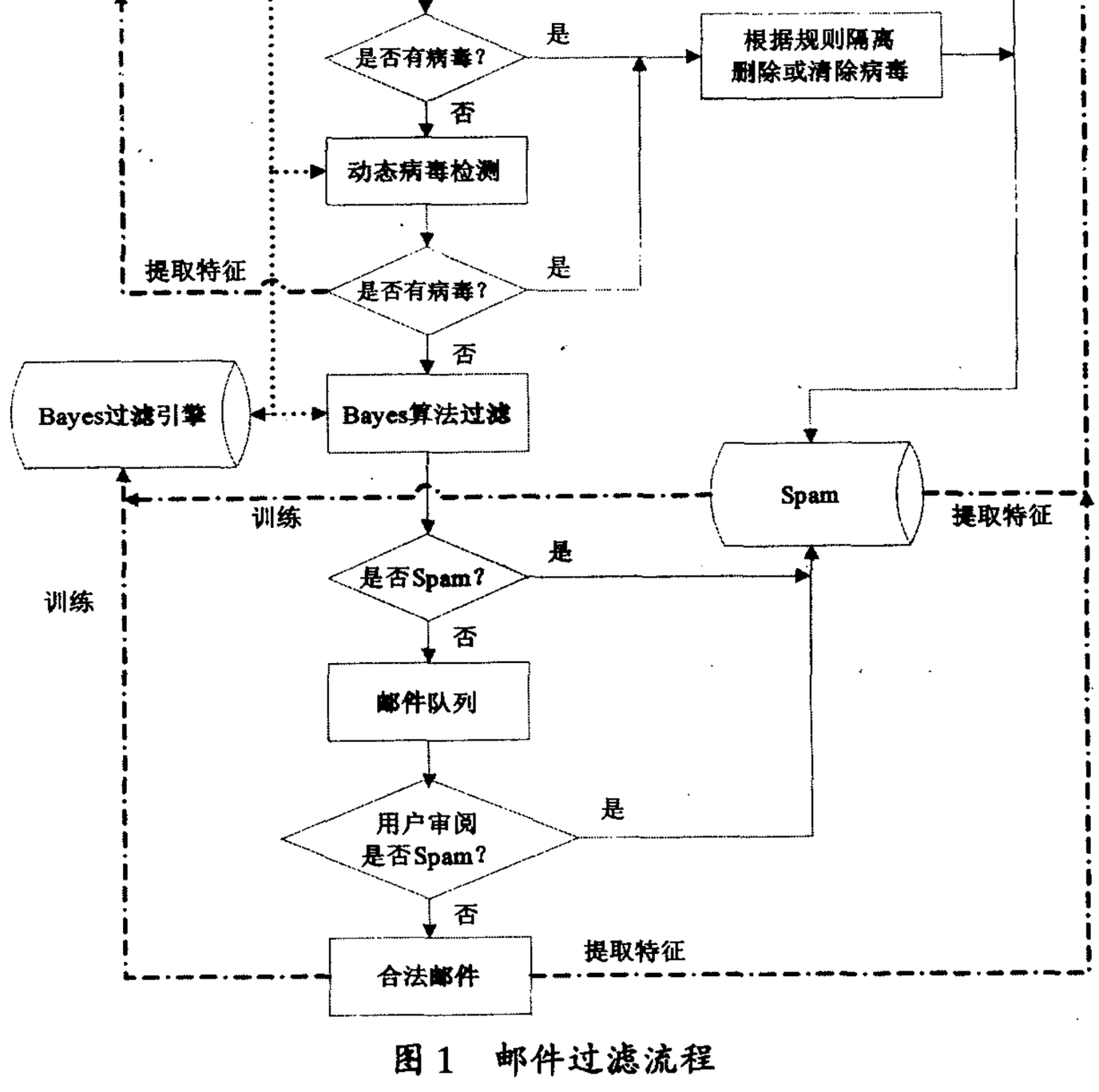

图 1　邮件过滤流程

2.3　方案分析

本方案存在如下的优点:

1) 采取多种过滤方式,使得过滤的标准更为严格,过滤结果也相对更加准确。

2) 能够对于利用邮件传播的已知病毒进行有效阻止,对各种变种及未知病毒,也能及时分析、判断并加以处置,最大程度减少对用户的威胁。

3) 能够及时接收用户个性化反馈意见,调整系统

(下转第 124 页)

NY:

取出该记录中的客户端套接字,通过它向客户机发拒绝信息;

取出服务器套接字,关闭它;

否则:

如果该会话记录中的客户端套接字不为 NULL:

通过该客户端套接字转发应答信息到客户机;

否则:

关闭该客户端套接字;

⑥释放互斥量;

⑦存储成功转发的应答信息到代理服务器硬盘,以便实现内容重现。

对响应信息进行关键字和基于语义的内容过滤判断,如果需过滤,就通过 m_pClient 向客户机发拒绝信息,否则,就通过 m_pClient 向客户机转发此响应信息。

7 性能分析

在 Windows XP SP2 环境下,用 Visual C++ 6.0 实现了上述算法。在 10 台客户机和一台代理服务器(均为 P4/2G/256M)搭建的子网环境中试验。取 100 篇 Web 页(文章),其中设置对某问题的赞成、反对、中立观点各 25 篇。应用简单的关键字过滤,查全率达到 98%,应用基于语义的过滤,根据设置的过滤语句模板,可以较好地甄别出 3 种观点,查准率达 93.5%。在内容重现的基础上对代理技术改进之后,对常用站点的访问速度明显加快。

参考文献:

[1] 马文斌,王 庆. Web 内容过滤实现方法的研究[J]. 计算机工程,2004(12):588-589.

[2] 周文刚,王景中. 基于语义的信息过滤算法的设计和实现[J]. 周口师范学院学报,2006,23(2):99-100.

[3] 张明武,陈启祥,楚惟善,等. HTTP 代理服务系统的实现与分析[J]. 计算机工程,2001,27(3):145-147.

[4] Policová G, Návrat P. Semantic Similarity in Content-Based Filtering[C]//In: Advances in Databases and Information Systems: 6th East European Conference, ADBIS 2002. Bratislava, Slovakia: [s. n.], 2002.

[5] Jorge J, Flores G. Semantic Filtering of Textual Requirements Descriptions[C]//In: 9th International Conference on Applications of Natural Language to Information Systems, NLDB 2004. Salford, UK: [s. n.], 2004.

[6] YU Fei, SHEN Yue, AN Ji-yao, et al. Information Audit Based on Image Content Filtering[J]. Wuhan University Journal of Natural Sciences, 2006, 11(1): 234-238.

[7] 李家欣,倪 亮,王 乘. 具备高速缓存的 HTTP 代理防火墙的设计与实现[J]. 计算机工程,2003,29(3):85-86.

(上接第 119 页)

的过滤策略,增加邮件过滤的针对性。

4) 便于功能扩充,系统管理者可自主增加、停用或修改过滤引擎,进行自定义的功能扩展。

5) 配置方便,各个模块分工明确,系统管理者可根据邮件用户需求定制使用或是不使用各个过滤模块。

在实现上,方案中存在的关键问题有:

(1) 对邮件内容进行规则匹配、贝叶斯运算、检测病毒都会占用大量的系统资源。

(2) 由于邮件过滤模块的层次较多,各过滤方式之间过滤内容上有重复、叠加的情况,导致过滤效率降低。

对问题(1),可以将病毒过滤模块和邮件过滤模块分别部署在不同主机上,将各模块采取分布式结构加以解决。对问题(2),可以不断优化策略管理来减少各个过滤方式在过滤内容上的重复过滤。

3 结束语

在计算机信息安全领域,垃圾邮件的泛滥成为严峻的现实,带毒邮件的检测过滤更是成为垃圾邮件过滤的新挑战,不同用户的不同需求也应得到个性体现和满足。基于综合方法的邮件过滤方案的提出,在一定程度上反映了反垃圾邮件的安全和个性化需求,也是对现有过滤方法的有益补充。

参考文献:

[1] 中国互联网络信息中心. 中国互联网络发展状况统计报告[EB/OL]. 2006-01. http://www.cnnic.net.cn.

[2] RFC706. On the junk mail problem[S]. 1975.

[3] 潘文峰. 基于内容的垃圾邮件过滤研究[D]. 北京:中国科学院,2004.

[4] Graham. A Plan for Spam[EB/OL]. 2002-08. http://www.paulgraham.com/spam.html.

[5] Androutsopoulos I, Koutsias J, Chandrinos K V, et al. An Evaluation of Naive Bayesian Anti-Spam Filtering[C]//Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (ECML 2000). Barcelona, Spain: [s. n.], 2000.

[6] Symantec. Understanding Polymorphic Viruses[EB/OL]. 2005-08. http://www.symantec.com.

[7] 刘尊全. 计算机病毒防范与信息对抗技术[M]. 北京:清华大学出版社,1991.