文章编号:1673-0062(2016)04-0094-06

垃圾邮件行为过滤技术研究

赵治国,谭 邦,夏石莹

(南华大学 计算机科学与技术学院,湖南 衡阳 421001)

摘 要:分析目前基于邮件内容过滤技术存在的缺陷,根据垃圾邮件的大量发送和不请自来的行为特征,提出了一种垃圾邮件行为过滤技术.该技术将各邮件服务器组成一个垃圾邮件协作过滤网络,邮件服务器发送邮件时判断其发送行为,然后将发送行为信息加密;在 MTA 会话阶段,接收方先解密邮件发送行为信息,然后将不请自来的群发行为垃圾邮件进行过滤.实验结果表明,该技术在 MTA 会话通信阶段能过滤大量不请自来的垃圾邮件,具有较高的准确率和查全率,处理速度也较快,节省大量的网络资源,具有良好的过滤性能.

关键词:垃圾邮件;群发行为;不请自来;邮件服务器;MTA

中图分类号:TP393.08 **文献标志码**:B

Research of Spam Behavior Filtering Technology

ZHAO Zhi-guo, TAN Bang, XIA Shi-ying

(School of Computer Science and Technology, University of South China, Hengyang, Hunan 421001, China)

Abstract: Analysis of the shortcomings existed in the techniques of filtering based on the content of email, this paper proposed a spam behavior filtering technology (SBFT) according to the sending behavior characteristics of spam. A spam collaborative filtering network is structured between mail servers. Mail servers judge their sending behavior when they send messages, and encrypt the information of sending behavior. In the process of MTA communication, mail receivers first decrypt the information of sending behavior, then a large number of uninvited spam are filtered. The experiment result shows that the SBFT can filter a large number of uninvited spam in the process of MTA communication, have higher accuracy and recall, speed up the process, save a lot of network resources, and have well filtering capability.

key words: spam; mass sending behavior; uninvited; mail server; MTA

收稿日期:2016-09-21

基金项目:湖南省教育厅科学研究重点项目(14A121)

作者简介:赵治国(1977-),男,湖南邵东人,南华大学计算机科学与技术学院讲师,硕士.主要研究方向:计算机网络安全和 SDN.

0　引　言

电子邮件已成为现代社会人们生活、办公的主要交流工具之一,带来了很大的方便,但是由于电子邮件发送成本低,也成为非法分子的一种获利工具,他们通过电子邮件传播欺诈信息、商品广告、色情淫秽等信息,对正常用户造成了很大困扰.垃圾邮件的危害引起全球各国的高度重视,我国已经成为全球最大垃圾邮件受害国之一[1-2].目前国内外主要的反垃圾邮件系统,普遍采用的是基于黑名单、规则库关键字内容过滤技术,采取"截获样本、解析特征、生成规则、规则下发、内容过滤"这种类似传统杀病毒系统的原理[3],存在着许多难以克服的问题:1)面临着动态 IP(网际协议)跟踪难,缺乏实时性和有效控制;2)基于邮件内容过滤,由于信件内容多样化,依赖关键字规则判别垃圾邮件准确性较低,还需要很多匹配运算,资源消耗大,尤其是在遭受巨量邮件攻击时,可能导致系统崩溃;3)通过拆信检查内容的方式进行反垃圾邮件,侵犯了公民电子邮件通信自由权和隐私权,这种内容过滤技术将受到广泛的法律质疑[2,4].针对以上问题,研究者提出了垃圾邮件行为识别方法[5-10].根据垃圾邮件的定义:"不请自来"的"大量"邮件[4],本文提出一种基于邮件发送行为的垃圾邮件过滤技术,通过邮件发送服务器和邮件接收服务器的协作方式,在 MTA(会话通信阶段),根据邮件的发送行为判断出所接收的邮件是否为垃圾邮件而不需要检查邮件内容.

1　垃圾邮件协作过滤网络及结构设计

将所有邮件服务器组成一个垃圾邮件协作过滤网络,如图 1 所示,邮件发送服务器主要进行发送行为统计与判断以及发送行为加密,然后将加密信息添加在邮件头部 MailFrom(邮件来源)命令里一起发送到邮件接收服务器;邮件接收服务器主要根据邮件发送行为和通讯录来判断到来的邮件是否为垃圾邮件,然后对垃圾邮件进行处理.

设计了一个部署在邮件服务器端的垃圾邮件过滤模块,其结构如图 2 所示,主要有邮件发送行为判断、发送行为加/解密、通讯录和垃圾邮件处理组成,其中,实线箭头表示邮件服务器发送邮件时过滤模块对邮件的处理过程,虚线箭头表示邮件服务器接收邮件时过滤模块对邮件的处理过程.

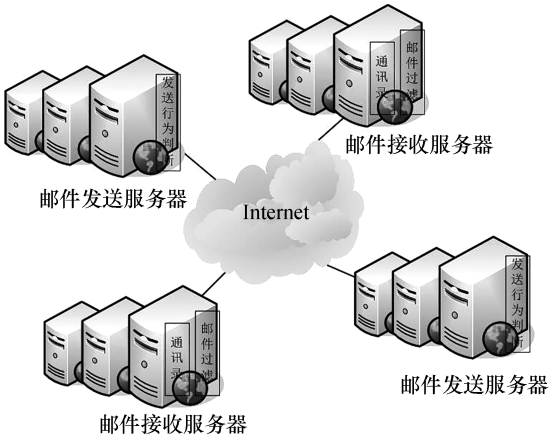

图 1　垃圾邮件协作过滤网络

Fig.1　Spam collaborative filtering network

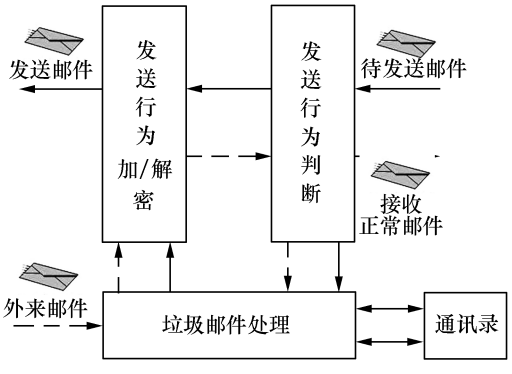

图 2　垃圾邮件发送行为过滤结构

Fig.2　The filtering structure of spam sending behavior

1)发送行为判断模块:统计邮件发送者在一定时间内所发送邮件的数量并判断其发送行为.对确定发送多少才算"大量"行为,可以参照行业标准和相关法律的规定.

2)发送行为加/解密模块:先统计邮件发送数量,然后判断其发送行为为正常行为还是群发行为,再利用一种多样性函数方法对邮件发行为进行加密并通过邮件会话通信发送给接收方,接收方对其进行解密并判断,然后采取相应的动作.

3)通讯录:用来记录邮件发送者常联系者或信任的邮件发送者.

4)垃圾邮件处理模块.外来邮件都经过垃圾邮件处理模块判断其是否为垃圾邮件,如果是,则进行相应的处理动作,如在会话阶段断开连接、设置为垃圾邮件标志或发出相应的警告代码等.另外,本邮件服务器用户发送的群发行为邮件,同样需要经过垃圾邮件处理模块检查,查询其接收者是否在本服务器的通讯录里,核对其是否是邮件

接收者主动请求的邮件,如果是,则为正常邮件;否则为垃圾邮件,则进行相应的处理动作.

2 垃圾邮件行为过滤技术的实现

2.1 通讯录存储与检索

2.1.1 通讯录存储结构

在邮件服务器为每个邮箱用户建立一个固定大小的通讯录,其结构如图 3 所示,包括联系人邮箱和联系时间.考虑到每个邮箱用户的联系人数不会很多,所以通讯录里的记录采用顺序方式存储.由于每个邮件服务器的邮箱用户非常多,而且是唯一的,为了实现快速查找,对邮箱用户采用 hash 算法[11]建立一个索引表,其结构包括关键码和指针,如图 3 所示.

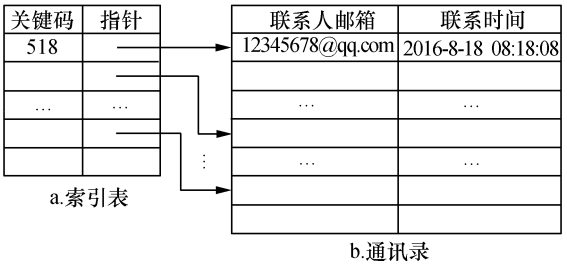

图 3 索引表与通讯录结构

Fig.3 The structure of index table and address book

2.1.2 索引表的建立与检索过程

索引表的建立过程,首先将邮件发送者的用户名的每个字符转换成十进制数字并求其和,然后以这个计算结果作为 hash 函数的关键码 key,计算 $h_1(\text{key})$ 值,以该值为地址到索引表中去查找.如果该地址对应的空间未被占用,就将该 key 存入这个存储单元里,并分配一个通讯录表并记录其联系者和联系时间.否则,比较关键码,如果相等,则更新其通讯录,否则要继续计算 $h_2(\text{key})$ 得到待查序列的值继续查找.如此反复到求出的某地址空间未被占用为止.经过一段时间后,每个邮箱用户的通讯录会越来越大,甚至会超过其最大值,这时就采用 LRU(最近最少用)调度算法把最久没有联系的联系人信息删掉,添加新的联系用户并记录其联系时间.

索引表的检索过程与建立过程相似,首先将邮件接收者的用户名的每个字符转换成十进制数字并求其和,然后以这个结果作为 hash 函数的关键码 key,计算 $h_1(\text{key})$ 值,以该值为地址到索引表中去查找.如果该地址对应的空间未被占用,则说明检索失败,表明邮件是不请自来者.否则比较关键码,如果相等则再用邮件发送者去匹配其主动通讯录的记录,若匹配则是主动联系者,并更新其联系时间;不然就是不请自来者.如果不相等则继续计算 $h_2(\text{key})$ 得到待查序列的值继续查找.如此反复到检索失败或成功为止.

2.2 加密/解密函数

主要对邮件发送行为进行加/解密,并把加密信息添加在邮件头部 MailFrom 命令里,然后与邮件接收服务器进行会话.

为了防止垃圾邮件发送者追踪邮件会话信息,之后通过推算就可得到函数库信息,从而伪造验证信息发送垃圾邮件,所以对加/解密函数的安全性有较高的要求.在这个函数模型中选用非线性变换,并且函数库尽量选用复杂的函数.另外,为了完成反函数的验证过程,且要求函数本身必须是双射的,即对于每一个 $f(x)$,有且仅有一个 x 与之对应[12],增加伪造验证信息的难度.

设置函数发生器 $f_1, f_2, f_3, \cdots, f_n$ 和邮件发送者的发送行为标识池 α_i 和 β_i(α_i 为邮件群发行为标识,β_i 为非群发行为标识).发送方过滤网关进行复合函数的运算,先在函数库中随机选择两个 f_i 函数和 f_j,然后根据邮件发送者的发送行为在相应的发送行为池里随机选择一个标识值如 a_1 并作为函数的初始值,计算 $b=f_i(\cdots f_j(a_1))$,并将 i 和 j 即函数在库中的序号、结果 b 和初始值 a_1 按一定顺序排列作为验证信息封装在 MailFrom 命令里,发送给接收方;接收方在收到验证信息后对其进行解密,对应发来的函数序号和结果 b,调用函数发生器,进行解密即反函数运算,求出函数初始值,并与发送过来的初始值比较,若等于则根据其值进行判断邮件发送者的发送行为并采取相应的处理动作,否则为伪造信息,采取相应的处理.

2.3 邮件发送数量统计与行为判断

合法邮件服务商发送的邮件,其发送行为判断都在发送方邮件服务器进行,对于通过非法软件发送的邮件,其发送行为判断就在接收方邮件服务器进行.

首先,建立一个固定大小的链表缓存邮件发送者在一段时间内发送邮件数量,其结构如图 4 所示,包括邮件发送者,计数器、统计时间和链表指针.

邮件服务器发送邮件时发送行为判断算法描述如下:

1)接收邮件 SMTP 会话信息,提取 MailFrom

的 RCTP TO 命令里的邮件发送者并回复 250 OK 应答.

2)查询邮件发送数量统计链表,判断此邮件发送者是否在表里,若不在,则在链表的最后添加它的结点记录,第四步;否则,第三步.

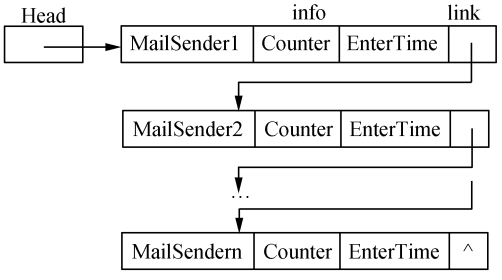

图 4　邮件发送数量的统计链表

Fig.4　The statistics list of the number of messages sent

3)获取 Systime(系统时间)和链表中相应结点的 EnterTime(开始统计时间),计算并判断 Systime-EnterTime ≥ T(T 根据行业标准规定),如果是,重新记录发送者链表信息;否则,Counter(计数器)累计邮件发送数量.

4)获取下一个命令信息,判断是否为 RCTP TO 命令,若是,则计数器 Counter 计数,然后发送 250 OK 应答,返回第四步;否则,下一步.

5)判断邮件数量,如果 Counter ≥ C(相关法律和行业标准规定的正常行为邮件数量最大值),则标记为群发行为,进行群发行为邮件处理;否则,标记为正常行为,加密发送行为信息,转发邮件.

2.4　垃圾邮件过滤算法

在 SMTP 会话阶段,邮件接收方服务器接收外来邮件,解密发送行为信息,对外来群发行为邮件和本邮件服务器产生的群发送行为邮件都进行进一步处理,查询邮件接收者通讯录,对不请自来的群发送行为邮件进行处理,比如在会话阶段断开连接阻止垃圾邮件发送,节省大量网络资源,或者设置为垃圾邮件标志等.另外,通过非法软件发出来的邮件,是没有发送行为验证信息,所以通过判断是否有发送行为验证信息,可以拦截掉非法软件发送的大量垃圾邮件.垃圾邮件过滤算法描述如图 5 所示.

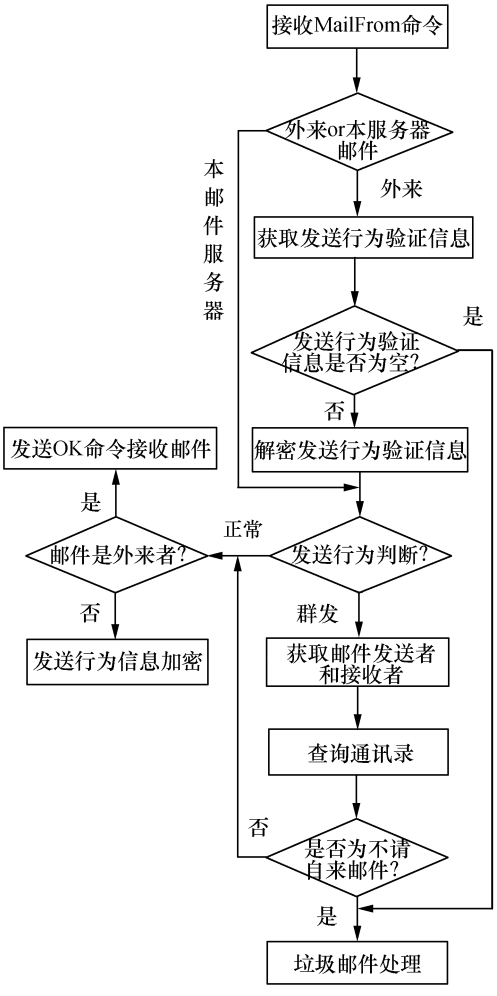

图 5　垃圾邮件过滤算法流程

Fig.5　Spam filtering algorithm flow

3　实验及结果分析

3.1　实验方案

在以太网网络环境下,配置五台邮件服务器组成垃圾邮件协作过滤网络和一台安装邮件群发软件的普通计算机,五台邮件服务器同时安装本文研究的垃圾邮件过滤软件(SBFT)和基于贝叶斯的邮件内容过滤软件 Foxmail 进行实验比较.

实验测试样本共 1 000 封邮件,其中正常邮件 650 封,垃圾邮件 350 封.选用 Network Associates 的 Justin Mason 提供的 Spam Assassin 开放语料集作为英文类邮件测试[13],其中垃圾邮件 80 封、正常邮件 270 封,由于中文和图片类垃圾邮件没有研究机构提供开放语料,通过收集和特意创造这类邮件,其中,中文类 570 封,图片类 80 封.对测试邮件样本分四组进行实验,具体实验样本数据如表 1 所示.

表 1 实验数据

Table 1 Experiment data

单位:封

分组	垃圾邮件			正常邮件		合计
	英文类	中文类	图片类	英文类	中文类	
G1	15	16	16	55	100	202
G2	18	22	18	60	110	228
G3	20	28	22	73	120	263
G4	27	34	24	82	140	307

3.2 实验结果分析

五台邮件服务器按着实验要求提供的邮件互相发送正常邮件和群发垃圾邮件,邮件群发主机也按着实验要求提供的垃圾邮件向五台邮件服务器随机发送垃圾邮件.采用同样的发送方式分别对 Foxmail 软件和本文研究的 SBFT 软件进行实验,SBFT 中统计时间参数 T 值设置为 60S,一次发生正常邮件数量参数 C 设置为 10 封,最后得出的实验结果如表 2 所示.

表 2 垃圾邮件过滤结果

Table 2 Spam filtering results

单位:封

	SBFT				Foxmail			
	垃圾邮件		正常邮件		垃圾邮件		正常邮件	
	识别为垃圾邮件	误判为正常邮件	误判为垃圾邮件	识别为正常邮件	识别为垃圾邮件	误判为正常邮件	误判为垃圾邮件	识别为正常邮件
G1	43	4	0	155	41	6	0	154
G2	52	6	0	170	49	9	1	169
G3	63	7	0	193	59	11	1	192
G4	76	9	0	222	72	13	2	220

从表 2 实验结果看出,本文研究的邮件发送行为过滤技术 SBFT 在本次实验中对正常邮件识别的准确率非常高,没有一封正常邮件被误判为垃圾邮件,而基于内容过滤技术 Foxmail 在本次实验中共有 4 封正常邮件被误判为垃圾邮件.在本次实验中,SBFT 正确查出邮件共 974 封,Foxmail 正确查出邮件共 957 封,邮件查全率比较如图 6 所示.

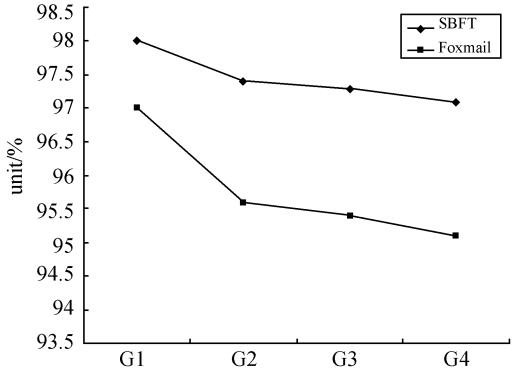

图 6 邮件查全率

Fig.6 Message recall

SBFT 的垃圾邮件查全率平均为 97.4%,Foxmail 的垃圾邮件查全率平均为 95.7%,SBFT 比 Foxmail 略高一点.在垃圾邮件的误判方面,SBFT 有 26 封误判为正常邮件,Foxmail 有 39 封误判为正常邮件,其误判率比较如图 7 所示,SBFT 的垃圾邮件误判率平均为 9.8%,Foxmail 的垃圾邮件误判率平均为 14.8%,SBFT 垃圾邮件误判主要原因是不能识别非群发方式的垃圾邮件,而 Foxmail 对图片类型垃圾邮件存在误判.

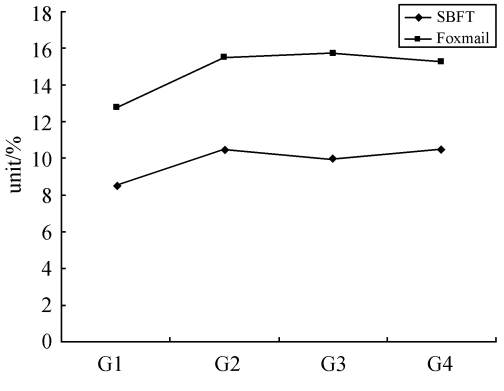

图 7 垃圾邮件误判率

Fig.7 Spam false

4 结 论

垃圾邮件的日益泛滥已是全球性的问题,不

仅占用大量网络资源,其内容还带有欺诈信息,困扰邮件服务商和邮件用户,因此,垃圾邮件过滤技术一直是研究重点.本文根据垃圾邮件大量发送行为和不请自来的特征,研究了一种垃圾邮件发送行为过滤技术,将各邮件服务器组成一个垃圾邮件协作过滤网络,邮件发送方进行邮件发送行为判断,在 MTA 会话阶段,接收方对不请自来的群发行为邮件进行垃圾邮件识别.实验结果表明,本文研究的垃圾邮件过滤技术 SBFT 在 MTA 会话通信阶段能过滤大量不请自来的垃圾邮件,具有较高的准确率和查全率、处理速度也较快,节省大量的网络资源,相比基于邮件内容过滤的 Foxmail 具有更好的性能.

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[R].北京:中国互联网信息中心,2016.

[2] 信息产业部.互联网电子邮件服务管理办法[EB/OL].(2010-03-02)[2016-09-16]http://www.mii.gov.cn/art/2010/03/02/art_521_7342.html,2010-03-02.

[3] 翟军昌,秦玉平,车伟伟.垃圾邮件过滤中信息增益的改进研究[J].计算机科学,2014,41(6):214-224.

[4] PAUL H.Internet Mail Consortium Internet Mail Consortium Report:Unsolicited Bulk Email:Definitions and Problems. IMCR-004,October 5,1997[EB/OL].[2016-09-16]http://www.imc.org/ube-def.html,2012-2-15.

[5] CHOUHAN S. Behavior analysis of SVM based spam filter using various kernel functions and data representations[J]. International journal of engineering research and technology,2013,2(11):3029-3036.

[6] ALMEIDA T A,YAMAKAMI A.Advances in spam filtering techniqueso[M]//Computational Intelligence for Privacy and Security.Berlin:Springer,2012:199-214.

[7] WIN Z M,AYE N.Identification of image spam by using histogram and hough transform[J]. International journal of Science and Research,2013,2(11):310-314.

[8] LIU W Y,WANG T.Online active multi-field learning for efficient email spam filtering[J]. Knowledge and information systems,2012,33(1):117-136.

[9] COSTA J,SILVA C,ANTUNES M,et al.Customized crowds and active learning to improve classification[J].Expert system with applications,2013,40(18):7212-7219.

[10] BERTINI J R,ZHAO L,LOPPES A A.An incremental learning algorithm based on the K-associated graph for non-stationary data classification[J]. Information sciences,2013,246:52-68.

[11] 严蔚敏,吴伟民.数据结构[M].北京:清华大学出版社,2011.

[12] 同济大学应用数学系.高等数学:上册[M].7 版.北京:高等教育出版社,2014.

[13] SpamAssassin. The Apache SpamAssassin Project[EB/OL].(2015-04-02)[2016-09-16]http://spamassassin.apache.org/downloads.cgi? update=201504002400.