基于内容过滤的内网防泄密系统的研究与实现

Research and Implement of Content – based Information Anti – Leakage in LAN

伍淳华　张鹏飞　左申正　（北京邮电大学 310 信箱 100876）

摘要：本文针对内网的主动泄密实现了一种基于内容过滤的内网防泄密系统 CIAL（Content – based Information Anti – Leakage in LAN），它以透明方式对进出网络的传输明文及电子文档进行监控，并运用中文信息处理技术对明文及电子文档的内容进行分析，一旦发现该信息涉密，立即阻止其传送，有效的阻止了内网的泄密同时也保证了网络的便捷性。并详细介绍了它的设计方案和实现技术。

关键词：内容过滤　电子文档　防泄密　中文信息处理

1　引言

随着信息技术的发展，各种先进的网络技术在给企事业单位带来了高效率的工作和管理方式的同时，也容易产生网内机密外泄。为防止信息外泄，各企事业单位往往不惜花巨资购进防火墙、入侵检测、漏洞扫描等各种网络安全产品，但这些产品仅仅只能防范外来者的侵入，对内部的主动泄密却无能为力。而据权威资料记载，大部份的机密、敏感数据，70% 以上都是被内部员工从内部网络系统的桌面终端计算机上通过各种传输、复制途径泄露出去的[2]。

为防范内部主动泄密，很多企事业单位已经对内部员工使用移动存储磁介质做了严格的规定，但是对于如何防范通过网络传输来进行的机密外泄却是扼住企业发展的一柄双刃剑。因为当前防范内部泄密的机制一般有两种：一种是直接封锁相应的端口，切断信息的流通，这种方式的确可以保证机密不外泄，但同时网络带来的便捷也大打折扣，极大的影响了工作效率；一种方式则是保持信息流通通道，但对信息的流通作监控，比如禁止用户上某些网站，或者对一些显示的流通文字作监控，且仅限于关键词匹配方式的监控，这种监控方式相对于第一种方式来说，对用户应用网络的影响较小，但监控力度也大大减弱，特别是对于一些以电子文档方式传输的机密无能为力[3]。

本文针对内网的主动泄密提出了一种基于内容过滤的内网防泄密系统 CIAL（Content – based Information Anti – Leakage in LAN），详细介绍了它的设计方案和实现技术。CIAL 以透明方式对进出网络的传输明文及电子文档进行监控，并运用中文信息处理技术对明文及电子文档的内容进行分析，一旦发现该信息涉密，立即阻止其传送，有效的阻止了内网的泄密同时也保证了网络的便捷性。

2　系统简介

2.1　系统的基本组成

系统主要由两大部分组成：一部分是 LINUX 平台的服务器端，另一部分是分布在内网中每台计算机上的客户端。服务器端根据内定的规则控制网内计算机与外网的连接，包括禁止/允许某些端口的开放，禁止/允许某台机器上网等等；驻守在局域网内每台机器上的客户端程序是网内计算机允许上网的通行证，局域网中的机器只有当其上的客户端程序是正常运行时，才会被允许上网，同时客户端程序一旦发现客户在访问机密文件时会切断其于外网的连接。

服务器端包括 A、B 两台设备，均串联在局域网的网关或者是防火墙后面，都工作在 LINUX 平台下。其中设备 B 主要作和数据相关的工作，包括一些网络访问日志和加密的电子文档的存储，以及分级查阅功能，根据用户的级别赋予其不同的访问权限。设备 B 与设备 A 独立连接，并且不能进行远程访问，这样可以防止存储在其上的数据被盗取。设备 A 则内网与

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

外网之间设起了一道安全屏障，重组内网数据包，采用中文信息处理技术对向外发送的信息进行分析，拦截可疑信息。

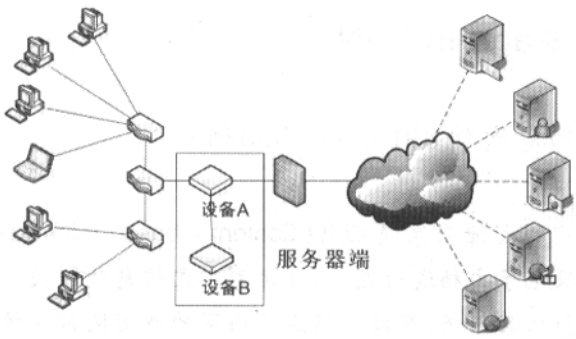

图 1　CIAL 网络拓扑图

2.2 系统的功能与架构

CIAL 主要是通过监控向外发送的信息来防止机密外泄，而信息的主要发泄方式就是通过电子邮件的方式向外发送。在 Internet 上，邮件的传送方式主要有两种，一种是利用 HTTP 协议，通过 webmail 的方式向外传送；还有一种就是利用 SMTP 协议，通过邮件服务器向外发送。CIAL 用防范和监控相结合的方式来防止机密信息的外泄：在内网的每台机器上均装有客户端程序，用于监控用户是否在操纵机密文档，一旦发现用户在操作机密文档，会立即通知服务器切断该用户与外网的连接；服务器端程序则通过对 80（HTTP）端口和 25（SMTP）端口进行监听，监控从内网发出的信息，尤其是电子文档信息，通过中文信息处理技术对其进行分析，一旦发现有涉密内容，就进行拦截。

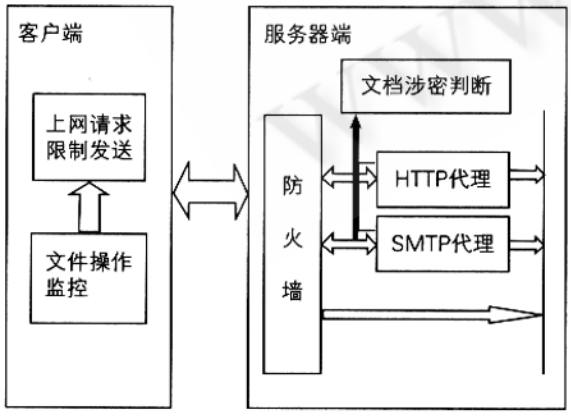

图 2　CIAL 架构

客户端程序运行于局域网内计算机，其主要有两方面的功能，一方面利用文件系统驱动截获文件访问，并利用中文信息处理方法来判断是否为涉密文件，若是涉密文件则通知服务器端，限制该用户上网。同时客户端还定时与服务器端通讯，客户端应用程序必须每隔固定时间就向服务器端发送上网的请求，若经过两个时间间隙服务器端并没有收到来自局域网内某台计算机的上网请求，则服务器端会自动切断该机的上网通路，这样就保证了只有在客户端程序正常运行的情况下用户才能上网，防止了用户恶意绕过客户端。

服务器端一方面要对局域网内的用户上网进行必要的控制，一方面又要最大限度的给予用户便捷快速的上网方式。服务器端运用防火墙来对局域网内的用户上网进行必要的控制，包括对端口的开放或封锁，以及根据客户端的的请求发送情况来允许或限制局域网内某一台计算机的上网。同时服务器上还运行着两个代理程序，在尽量不影响用户上网的前提下对发往外网的文件作必要的监控。由于 SMTP 协议和 HTTP 协议在实时性上的要求不同，所以这两个代理在功能的实现上也有所不同。SMTP 协议用来发送邮件的协议，这种协议对实时性要求不是很高，系统采取“邮件落地”的形式进行处理，即将发送的邮件内容以及该邮件的相关信息（用户名，密码）等记录下来，并在对邮件内容进行检查确认后再决定是否发送该邮件。而 HTTP 协议是互联网 Web 访问协议，这种协议的实时性要求非常高，不能像 SMTP 的“邮件落地”一样处理，因此 CIAL 采取一种“单向代理”的方式完成——只关心外发的数据，对返回的数据实施高速路由返回客户端，这样做的出发点是只有向外发送的数据才有可能将涉密文档发送出去，而从服务器返回的数据不会有此问题。

3 关键技术

3.1 HTTP 代理

HTTP 代理具有量大且实时性很高的特点，为了尽量不影响上网的速度，对这种协议的内容监控采用“应用层代理 + web 缓冲”的方式完成，且仅监控外发的数据，对返回的数据不做检查。采用应用层代理的方式实现较为简单，缺点是面临大量的服务请求时有可能

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

造成瓶颈，因此采用与 Squid 协同工作的方式，利用 Squid 的高速 web 缓冲解决瓶颈问题。

应用层代理和 Squid 代理共同组成了一个具有信息内容过滤功能的 HTTP 透明代理服务器。其中，应用层代理主要用于截获用户的上网请求，并判断该请求是否有外带数据，如果有则获取该信息并交由中文信息处理程序来判断该数据是否涉密，如若涉密，则切断此次上网请求，否则，则在客户端浏览器和 Squid 间建立起一个网络通路来完成一次上网请求。

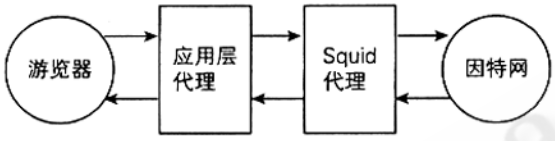

图 3　HTTP 代理体系结构

3.2 SMTP 代理

SMTP 代理对实时性要求不是很高，CIAL 获取局域网内向外发送的邮件，记录相关信息并对该邮件的内容进行核实，如若合法，则向外发送，否则拦截该邮件并进行相关的日志记录。这种代理方式是一种透明代理的处理方式，为此，必须在 LINUX 系统中插入一个内核模块，将所有内网的 SMTP 协议转向到本系统的 SMTP 服务器，并且将原始目的地址记录下来，已备后面真正发送邮件时使用。

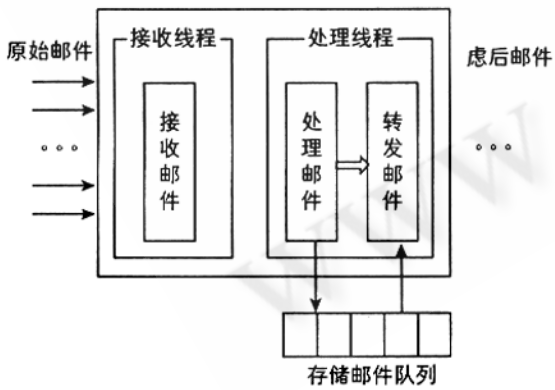

图 4　SMTP 代理体系结构

为了处理局域网内大量请求到来的情况，系统的调度方案采取了“小马拉大车”的机制。系统中只创建了两种线程，一种是接收线程，一种是处理线程。前者只管接收客户邮件，然后压入队列，由后者负责从队列中读取邮件并过滤转发。这种机制执行效率高，可以缓解过多用户请求到来的压力。

3.3 中文信息处理

中文信息处理主要包括了对各种信息的格式识别及文本化，并对文本文档进行涉密判断。格式识别和文本化是将网络中各种各样的编码方式进行统一，并识别其中电子文档的格式，包括 DOC 文档、PDF 文档、ZIP 文档等，将其统一转化为文本文档。这些文档经过特征抽取模块抽取特征之后再与语料库中的涉密文档进行比对来判断输入文档是否涉密，以采取相应的措施。

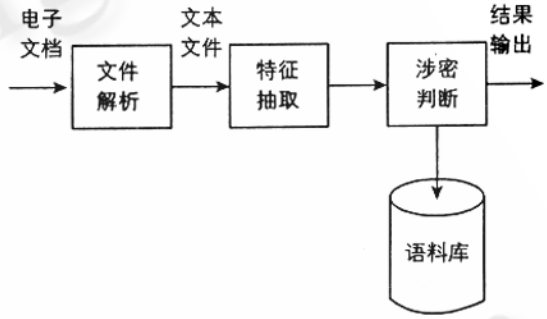

图 5　中文信息处理模块体系结构

目前的中文信息处理方法多是从语法的层次上，通过机械的词语、词频的比对来进行文档的分析。这种方法得出的结果往往是非常粗糙的，常常会产生误判或者是漏判。CIAL 以全信息理论[6]为指导，充分考虑上下文，通过语义和语境的标注，达到从语用的层次上来理解输入文档的实际内容，从而能较准确的判断出文档是否涉密。

3.4 数据库开发

CIAL 中的日志以及一些涉密文档都存储在设备 B 上的数据库中，数据库的类型为 mysql。数据库的开发主要有两方面的内容，一是数据的加密存储，一是数据的分级展示。

CIAL 的数据传输和存储都需要经过加密处理，以防被不法分子获取。加密分成两种情况，一种是不需要进行解密的数据，象用户的密码等信息；另一种是需要进行解密以向用户展示原始内容的数据。经过对各种加密算法的功能分析，第一种情况可以用 MD5 加密

（下转第 115 页）

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

(上接第69页)

算法完成,第二种情况则采用DES对称加密解密方式完成,密钥为固定密钥,由服务器和客户端私下协商决定,不需要在网络内进行传送,从而降低了风险性。

数据的展示则需要开发PHP Web应用程序,展示的主要内容包括

- 系统的状态
- 当前在线用户,以及每个用户的状态
- 各种协议的日志记录
- 加密文档的内容

在进行数据展示时,考虑到涉密文档的特殊性,对用户进行了分级管理,只对权限范围内的用户展示相关内容。

4 结论

本文提出了一种基于内容过滤的内网防泄密系统(CIAL),采用了中文信息处理方式,以透明的方式对从内网发出的消息进行监控,在保证了网络便捷性的同时,也有效的降低了内网泄密的危险性。该系统已经在企业内部网络中运行了一段时间,在网络的性能和防泄密功能上表现良好。

参考文献

1 Maximum security. 2nd ed. Sams Publishing, 1998.

2 李培修、敖勇、贾永强,内网涉密信息泄露途径及防范,计算机安全,2005(7):75-76.

3 张秋江,涉密网的安全构建,信息安全与通信保密,2006,3:27-29.

4 吴晓昶、李名世,办公业务网信息临控系统设计,厦门大学学报,2004(43):332-335.

5 唐正军、田仲、王兵等,网络入侵检测系统的设计与实现[M],北京:电子工业出版社,2002.

6 钟义信,信息科学原理[M],第三版.北京:北京邮电大学出版社,2002.1220.

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn