文章编号:1001 - 9081(2009)06 - 1662 - 03

加密网络流量类型识别研究

李 尧[1,2,3],郝 文[3]
(1. 内江师范学院 计算机科学学院,四川 内江 641110; 2. 内江师范学院 网络应用项目开发重点实验室,四川 内江 641110;
3. 国家信息安全工程技术研究中心,北京 100093)
(lycn_99@126.com)

摘 要:针对传统的检测方法无法识别加密网络流量类型的问题,对加密网络流量类型的识别进行了研究。从加密后保持不变的网络流量特征着手,提出了通过统计数据包长度、数据包到达时间间隔以及流的方向等方法可以正确地识别加密网络流量类型,最后设计出加密网络流量类型识别的模型和相应的加密网络流量实时识别方案,通过对该方案的验证表明采用该方案能够获得较好的效果。

关键词:数据包;网络流量;特征;识别
中图分类号: TP393 **文献标志码**:A

Research of encrypted network traffic type identification

LI Yao[1,2,3], HAO Wen[3]
(1. *College of Computer Science, Neijiang Normal University, Neijiang Sichuan* 641110, *China*;
2. *Network Applies Key Laboratory of Item Development, Neijiang Normal University, Neijiang Sichuan* 641110, *China*;
3. *National Information Security Engineering Technology Research Center, Beijing* 100093, *China*)

Abstract: Concerning the problem that the encrypted network traffic type cannot be identified using traditional inspection way, this paper studied its identification methods. Since the network traffic remains intact after being encrypted, the network traffic type could be identified correctly by calculating data packet size, timing and direction, etc. Finally, the identification model of the encrypted network traffic type and the corresponding real-time identification project were also designed. The result of the verification of the project shows that using the project can achieve good effect.

Key words: data packet; network traffic; features; identification

0 引言

随着 IP 网络逐步成为各种业务的承载网,其中的应用业务类型越来越多,这为业务的识别、监视、控制和管理方面带来了巨大的挑战。国内外对业务识别与控制技术已经进行了大量的研究,并且在网络的各个层面产生了许多相对独立的控制技术,典型的方案一般有在两个层次的检测结构:

第一层 根据数据包头部信息检测;

第二层 根据数据包载荷检测,对传输的明文数据进行检测。

这两个层次检测的综合应用能够提高方案的检测性能和检测效果。如图 1 所示。

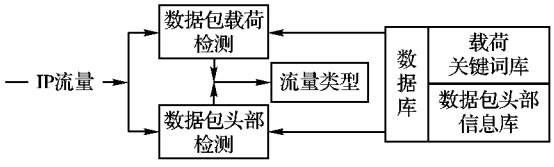

图 1 典型的 IP 网流量检测方案示意图

随着加密流量的增多,传统的检测方法很容易被躲避,目前,基于数据包头部和载荷的检测对于隐藏在加密流量中的数据已经不再有效,无法识别加密流量的类型。国内对加密流量分类研究的相关文献还比较少,本文试图起到抛砖引玉的作用。

1 加密流量类型识别方法

1.1 基于数据包长度的识别

不同业务类型应用的数据包长度是不同的。例如对于 UDP 传输的加密业务,可以很清晰的区分出业务类型。UDP 一般承载实时的业务,例如网络游戏、音频视频数据。UDP 传输协议是彻底的数据报传输协议,其流量特征和应用类型密切相关,根据不同应用的需要传输的数据包长度不同[1]。

以 VoIP 通信为例,表 1 列出了我们 Windows Netmeeting 在局域网通信实验获得的语音编码和数据包长度的对应关系。

表 1 部分 VoIP 语音编码对应的数据包长度

编码方式	速率/Kbps	数据包长度/位
G.723.1	5.3	144
G.723.1	6.3	192
G.729	8.0	20

在校园网出口采集的数据也证明,只要 VoIP 采用的语音编码确定,其数据包长度是基本固定的。由于不对流量的端口和载荷进行检测,因此无论流量是否加密,该方法都能够有效的检测 UDP 流量的类型。基于 UDP 应用的数据包长度只和业务类型有关,与网络拥塞程度和用户的设置没有关系,因

收稿日期:2008 - 12 - 11;**修回日期**:2009 - 03 - 06。
基金项目:四川省应用基础重点项目(07JY29-124);四川省教育厅重点项目(2006A145)。
作者简介:李尧(1965 -),男,副教授,CCF 高级会员,主要研究方向:网络信息安全、网络数据库、计算机信息系统; 郝文(1968 -),男,高级工程师,硕士,主要研究方向:信息安全。

此无论数据包是否加密,都可以用数据包长度来区分 UDP 业务的类型。

对于视频类 UDP 应用而言,由于受流媒体编码和播放速率的限制,这些数据在传输过程中数据包长度和流速需要有一定的规律才能满足要求。对游戏等应用而言,同样由于软件在游戏交互过程的需要,其数据包长度和流速是基本固定的,保证了接收方对数据到来的正确解码。目前各种基于实时业务的编码速率都是固定的,所以只要能够得到数据包的长度,我们就能够识别出其业务类型。

1.2　基于数据包到达时间间隔和双向包长度的识别

对于加密网络流量来说,通信流量模式有些是无法隐藏的,例如,数据包长度和到达的时间间隔等流量特征可以很容易获得。我们对从校园网出口采集的数据中分析得出,不同的流量类型在通信模式有着明显的差别(如图 2 所示)。

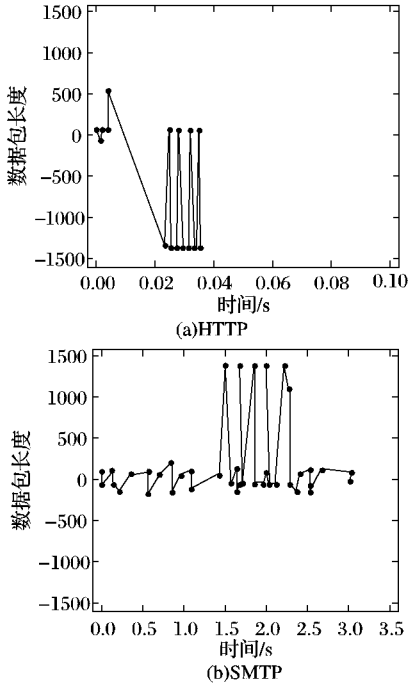

图 2　典型的 HTTP 连接和 SMTP 连接

图2中,t = 0时刻一个连接开始建立,即第一个数据包到达,一个点代表一个数据包,x 轴代表数据包到达时间,y 轴代表数据包长度(y 为正值代表数据包从客户端向服务器传输,为负值代表服务器向客户端发送数据)。HTTP 和 SMTP 的流量在数据包到达时间间隔和数据包长度以及方向上明显不同。图 2(a)中,开始的三个短包是 TCP 建链的“三次握手”信号,建链后大约 5 ms 从客户端向服务器请求 HTTP 连接开始,大约在 21 ms 处服务器开始向客户端发送数据。图 2(b)显示的 SMTP 传输过程完全不同于同图 2(a)的 HTTP 传输,HTTP 主要是服务器向客户端传输数据,而 SMTP 则主要是客户端向服务器发送数据。SMTP 与 HTTP 在流量特征上还有一个非常明显的区别,那就是 SMTP 在传输数据以前,客户端和服务器建立连接需要多次会话发件箱和收件箱,如图 2(b)所示,在 TCP 建链的“三次握手”信号之后,在客户端和服务器之间还有一系列的交互过程,一般维持在 0.1 s 至1.1 s;在 1.4 s 左右,包含较大附件的邮件开始从客户端向服务器发送。

1.3　基于其他流量特征的识别

很多研究对于加密网络流量的识别都有一个假设,即如果可以知道一个流的通信过程,则对一个流中的全部数据包都能够归到一起并提取出一些流量特征。例如对于 SSL 或者它的改进型 TSL 加密,能够保持网络层或者传输层的头部,因此可以使用传统的工具来估计 RTT、丢包等有些 TCP 特征。但是如果用户采取措施隐藏自己访问对象的流量模式,例如在 IPSec 的流量中,除了 IP 头部信息有所保留外都加密了,并被封装了新的 IP 头部,用来跟踪获取传输层的信息(如 TCP 标识等)就都不可见。

对于网络层加密的流量,至少我们可以获得一些流量特征信息,例如数据包持续时间,数据包长度和 IP 地址。经过对 TCP 协议深入研究分析可以发现,每一次 TCP 连接都有一个“三次握手”过程,并且其载荷很少,可以据此识别 TCP 连接的开始。同样,不同的应用 TCP 连接平均时间不同。根据研究,HTTPS 的持续时间为 10.1 s,WWW 的持续时间为 22.7 s,FTP 的持续时间为 89.8 s[2]等,文献[3]提出了通过加密后保持不变的特征(数据包字节数、持续时间和流方向)来研究识别应用协议。通过提取这些特征,建立加密流量识别模型,并对各种情况的加密流量进行识别判断。

2　加密网络流量实时识别方案的设计与验证

2.1　加密网络流量识别模型

识别加密网络流量的业务类型,是基于这样一个事实:在流量类型识别以前,建立一个包含大量加密网络流量特征的数据库,和要检测的加密流量的特征进行比较分类,可以在一定程度上确定具体流量类型。如图 3 所示。

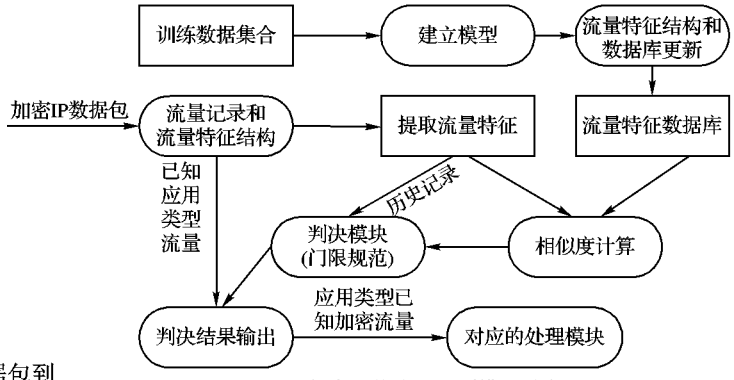

图 3　加密网络流量识别模型示意图

首先,我们要找到一个比较完善的训练数据集合,能够提取出加密流量类型的流量特征。针对这个训练数据集合做了以下一些工作:实时地从加密网络流量中提取了一些网络流量的基本特征数据,比如包长度的信息、包到达时间间隔(持续时间)、流方向等,这些基本特征比较详细地描述了加密网络流量的类型。

然后将提取到的流量特征结构存储在一个数据库中,由于该数据库是整个加密网络流量类型识别的基础,为了保证该识别的可靠性和比较强的可扩展性,就要求该数据库需要不断根据流量特征的改变而更新。但同时由于流量特征的提取有实时性的要求,以及考虑到现有计算机的计算能力等问题,特征的提取不可能对所有加密网络流量信息全部进行提取,而必须对之进行选择。为此在数据库中预留了大约 80 个保留项,以便将来的扩展之需。由这些保留项以及上述提取的内容共同组成了一个有 256 项的数据库。该数据库具有:

1)比较详细地涵盖了现有加密网络中主要流量的各种统计信息。

2)不含敏感信息,比如包内容信息等。

3)其存储空间完全有限,如果每隔60 s统计一次,一个月大约有30×24×60=43 200条记录,每条记录由256个数字组成,按照文本格式保存大约是2 048 B。因而按照这种方式保存一年的数据所需空间是(约是1 GB):

2 048×43 200×12=1 061 683 200

当判决一个加密数据包的类型的时候,我们需要提取出其相同的流量特征结构,和数据库中的结构比较,根据结果判断流量的业务类型,如果已经有历史记录,则根据历史记录直接判断流量类型。

2.2 加密网络流量类型实时识别方案设计

对于加密网络流量类型的实时识别,我们需要根据不同的流量特征区分不同的识别方案,我们设计的加密流量实时识别方案如图4所示。由于UDP传输的流量类型相对较少,可以准确的应用判断,我们单独把它作为一个模块判决。

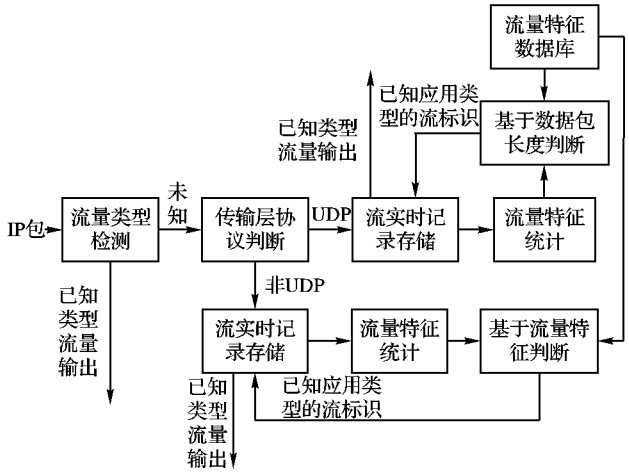

图4 加密网络流量实时识别方案

首先对到来的IP数据包基于目前常用的流量类型检测,区分出已知和未知两种类型的流量,把已知类型的流量识别出来;对未知类型的流量判断传输层协议并实时记录流量的数据,如果是UDP则统计数据包长度,并根据数据库中的流量特征准确的判断业务应用类型,把流标识符送到存储单元并调度输出数据;如果不是则进入非UDP流量判决,对流量特征进行统计,再根据流量特征数据库判断具体应用,把流标识符送到存储单元调度输出数据。对于流标识由IP地址经过哈希运算确定[4]。

2.3 方案验证

本文设计所采用的逻辑电路由Xilinx Virtex-4 LX160 FPGA完成,FPGA所具有的大规模并行处理能力和可编程的灵活性使得该模块能获得极高的并行处理性能。对于逻辑电路,FPGA的工作时钟在125 MHz,同时处理128比特的数据,这样可以得出引擎的吞吐率:128×125 MHz=16 Gbps。流量统计特征数据在IDT 71T75602 SRAM中存取,设定工作频率为125 MHz,SRAM的读写周期为8 ns,接口位宽32比特,每到一个数据包,需要读写SRAM各一次,共需16 ns。设链路数据率为10 Gbps,最小包长为64 B,则每个数据包的传输时间为51.2 ns。满足于我们所选用的FPGA要求,能够实现10 Gbps的线速处理。

为了验证本文方案的可行性,简单的以数据包长度和流持续时间为流量特征参数。在测试时,使用Spirent AX4000测试仪发送包含各种加密流量类型的数据包,将接收到的分类后数据与测试仪发送数据比较,其识别准确度在80%以上。如表2所示。

表2 测试结果

加密流量类型	发送数据包数	分类后数据包数	识别准确度/%
HTTPS	3 068 517	2 700 295	88
VoIP	420 568	353 277	84
SMTP	695 342	597 994	86

本文设计的方案是在国家863重大专项"大规模接入汇聚路由器(ACR)系统性能和关键技术研究"平台上验证的,完全能够满足方案设计的硬件要求。

3 结语

本文主要对加密流量类型的识别进行了分析和研究。首先从加密后保持不变的流量特征着手,提出了通过统计数据包长度、数据包到达时间间隔以及流的方向等方法可以正确地识别加密网络流量类型,然后给出了加密流量识别的模型和相应的加密流量实时识别方案,并对该方案的可行性进行了验证。结果表明采用该方案能够获得较好的效果。

参考文献:

[1] PARISH D J, BHARADIA K, LARKUM A, *et al*. Using packet size distribution to identify real-time networked applications [J]. IEE Proceedings-Communications, 2003, 150(4): 221-227.

[2] ROUGHAN M, SEN S, SPATSCHECK O, *et al*. Class-of-service mapping for QoS: A statistical signature-based approach to IP traffic classification [C]// Proceedings of ACM SIGCOMM IMC 2004. Taormina, Italy: ACM Press, 2004: 5-27.

[3] CHARLES V W, FABIAN M, GERALD M M. On inferring application protocol behaviors in encrypted network traffic [J]. Journal of Machine Learning Research, 2006, 7(12): 2745-2769.

[4] 程光,龚俭,丁伟,等. 面向IP流测量的哈希算法研究[J]. 软件学报, 2005, 16(5): 652-658.

(上接第1631页)

果表明:在测距存在误差的情况下,利用该方法捕获sybil节点仍有很高的准确性。它非常适合于无人值守恶劣环境下的无线传感器网络。

参考文献:

[1] NEWSOME J, SHI E, SONG D, *et al*. The sybil attack in sensor networks-analysis-defenses [C]// Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks. New York: ACM Press, 2004: 259-268.

[2] 邱慧敏. sybil攻击原理和防御措施[J]. 计算机安全, 2005(10): 63-65.

[3] LEVINE B N, SHIELDS C, MARHOLIN N B. A survey of solutions to the sybil attack [R]. Amherst: University of Massachusetts Amherst, 2006.

[4] 钟志光. 一种基于相位差测量的WSN节点测距方法[J]. 传感技术学报, 2007, 20(12): 2728-2732.

[5] 张建明,于群,王良民. 基于地理信息的传感器网络Sybil攻击检测方法[J]. 系统仿真学报, 2008, 20(1): 259-263.

[6] 于群,张建明. 无线传感器网络中的Sybil攻击检测[J]. 计算机应用, 2006, 26(12): 2897-2902.

[7] WANG A, HEINZELMAN W B, SINHA A, *et al*. Energy-scalable protocols for battery-operated microsensor networks [J]. Journal of VLSI Signal Processing, 2001, 29(3): 223-237.