复杂网络环境下加密流量识别方法研究

Research on Encrypted Traffic Recognition Method in Complex Network Environment

童　博，施　俊，赵纯熙（中讯邮电咨询设计院有限公司，北京 100048）
Tong Bo，Shi Jun，Zhao Chunxi（China Information Technology Designing & Consulting Institute Co.，Ltd.，Beijing 100048，China）

摘　要：
互联网络快速发展，流量的变化日益多样和复杂，加密流量占比不断提高给网络安全带来挑战。通过对现有加密流量的主流识别方法进行分析对比，设计了一种适用于大型骨干网加密流量应用识别的系统，以对不同网络环境下的加密流量进行精准分类。

关键词：
DPI；加密流量；应用识别；流量分析
doi：10.12045/j.issn.1007-3043.2022.08.013
文章编号：1007-3043（2022）08-0070-05
中图分类号：TN919
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
The rapid development of the Internet，the increasingly diverse and complex changes in traffic，and the increasing proportion of encrypted traffic has brought challenges to network security. By analyzing and comparing the existing mainstream identification methods of encrypted traffic，a system suitable for the application identification of encrypted traffic in large backbone networks is designed to accurately classify the encrypted traffic in different network environments.

Keywords：
DPI；Encrypted traffic；Application recognition；Traffic analysis

引用格式：童博，施俊，赵纯熙. 复杂网络环境下加密流量识别方法研究[J]. 邮电设计技术，2022（8）：70-74.

1 概述

随着互联网络的发展，网络应用日益丰富，运营商骨干网出口流量也逐渐向复杂化和多样化方向演进。一方面，用户对数据的隐私保护需求不断增加，另一方面，网络应用软件使用私有协议进行数据传输，这样使得网络中加密流量的比例不断提高。根据谷歌的报告，2021年5月 Chrome 加载网页中启用加密的比例已经到了98%[1]。加密在保护隐私的同时也给网络安全带来了新的挑战。网络攻击者将加密作为隐藏攻击活动的工具，使攻击活动变得隐秘，不易识别。为了维护健康的网络环境，运营商需要提出有效的识别方法，提高网络监管技术手段，实现加密流量精细化管理，保障网络安全运行。

收稿日期：2022-06-24

2 加密流量应用类型识别方法

2.1 单包净荷特征识别

为了保护用户和网站数据的隐私，保护网络浏览和交易的安全，越来越多的互联网应用使用HTTPS加密传输，例如主流网站百度、淘宝、京东等。

HTTPS不是一种单独的协议，它是在会话层SSL/TLS之上的HTTP。SSL/TLS介于TCP和HTTP之间。HTTPS协议交互流程如图1所示。客户端和服务端TCP建链成功后，客户端验证完服务端的公钥证书后，生成正文密钥，使用公钥加密后发送给服务端。服务端使用私钥解密后，得到正文密钥。随后双方开始进入正文数据通信阶段，正文数据经过正文密钥加密后在线路中传输，保障了数据安全。

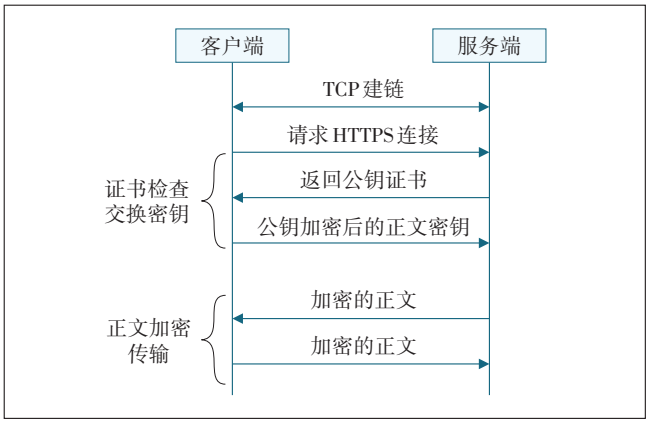

图1　HTTPS交互过程

QUIC（Quick UDP Internet Connection）是谷歌制定的一种基于UDP的低时延的互联网传输层协议。QUIC协议整合了TCP协议的可靠性和UDP协议的速度和效率，在UDP的上层，QUIC提供了可靠、有序、安全且快速的传输服务。目前，在Chrome中有85%以上关于谷歌自有业务的请求响应都是通过QUIC承载。

对于HTTPS、QUIC等采用SSL/TLS标准加密的协议，如图2所示，可以利用协议交互过程中的明文信息进行应用类型识别。例如ClientHello报文的server-name字段和Server Certificate中的commonName均包含了域名等明文信息。此类加密的协议，可以根据域名信息，识别单包净荷特征字，并进一步确定加密流量的应用类型。

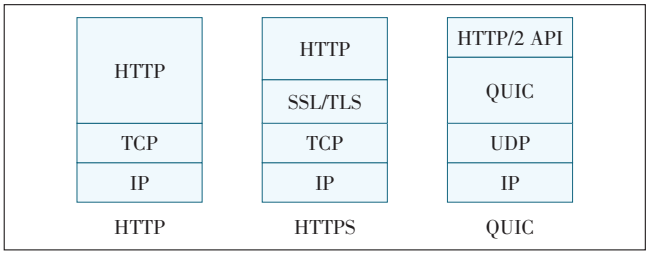

图2　HTTPS和QUIC协议栈

2.2 基于DNS域名回填

互联网流量采集系统针对HTTPS等协议生成相应协议话单，话单中包含识别出的应用类型。对于单包净荷特征不能判断出应用类型的HTTPS话单，可通过解析DNS报文，从中提取出域名信息，再回填到HTTPS话单，从而提高HTTPS应用识别率。

客户端通过域名HTTPS访问网站前，会给DNS解析服务器发送DNS报文，请求进行域名解析。DNS解析服务器返回的DNS应答报文中包含该域名对应的IP地址信息。客户端获得IP地址后，再访问相应网站。通过上述过程，运营商可以解析DNS应答报文，并缓存DNS报文里域名和IP地址映射关系，采用域名回填方案，获取话单中的应用类型。

但是，基于DNS域名回填的方法存在以下局限性和问题。

a）1个IP对应多个域名（例如一个服务器群中部署多个HTTPS网站或CDN环境）的情况不能使用DNS域名回填，否则会造成应用类型识别错误。

b）1个域名映射对应多个IP时，为提高关联率，需要记录所有IP用于关联匹配，这会增加本方案的复杂度。

2.3 基于多包关联的加密流量识别

网络中有些应用采用私有加密方式进行通信，根据单包特征，难以进行应用识别。但从数据流上来看，同一个流上多个报文之间存在着一定的指纹特征。基于多包关联识别方法有以下几种。

a）单包特征强度不够，容易误识别，可以通过多包特征匹配进行强化，降低误识别率。

b）单包特征强度不够，但与多包之间流量特征结合，即DPI+DFI进行强化，提高应用识别准确度。

c）没有任何担保特征，单纯依靠流量特征来识别，即纯DFI识别。例如多个报文长度之和等于固定值，且成组出现。DFI识别可以用于应用流量大类识别，很难精确识别到具体应用。

2.4 基于多流关联的加密流量识别

多流关联的加密流量识别有以下2种方式。

a）应用程序的控制面和数据面在2个数据流上传输。控制面报文中携带数据面的IP端口信息，数据面采用被动方式建立连接（如FTP业务），由服务器端动态分配IP端口，数据面上的数据流中只有数据内容，没有协议本身的指纹特征，难以识别。对于这类业务，需要在控制面数据流中解析出数据面的IP端口信息，并进行锁存。在数据面报文建立通信后，可以通过锁存的IP端口信息进行应用类型的识别。

b）应用程序在通信前协商加密算法和加密能力，会根据通信对端采用适当的加密算法进行通信。如果对端不支持加密，则会进行明文通信。此类应用可以通过抓取暴露的明文通信报文，进行IP端口的训练学习，建立IP端口与应用的映射库，通过IP端口信息识别出加密流量报文的应用类型。

2.5 基于多链路关联的加密流量识别

骨干出口等复杂网络环境存在非对称路由，即1条五元组流的上行和下行方向数据分散在不同链路或不同设备中传输。对于HTTPS等报文指纹特征主要出现在上行方向（请求方向）的应用，可以采用上下行链路话单关联的方法，通过上下方向识别出的应用类型回填到下行方向的话单中，提高下行方向应用识别效果。

多链路关联识别的网络部署可以采用以下2种方案。

a）在原始链路流量和DPI应用识别模块之间部署汇聚分流平台。全部链路接入汇聚分流平台。汇聚分流平台将全部流量进行同源同宿处理，进行上下行流量的关联后，再输出到后端DPI采集模块，保障数据报文上下行完整的采集和识别。

b）原始流量分别接入DPI采集模块，DPI采集模块后端部署话单合成系统。不同的DPI采集模块分别针对上行、下行流量独立产生数据话单，传送到话单合成系统，在话单合成系统，针对上行或下行的话单，按照五元组流进行上下行关联。

2.6 基于逆向技术破解加密数据包

网络中有些应用程序采用私有加密协议，包括常见的一些P2P下载、流媒体、游戏等应用程序等。通过对这些私有加密应用进行逆向分析，可破解一部分应用的加密协议。在实际应用中，以下3种加密方式可以通过逆向分析识别应用类型。

a）固定的密钥：应用代码中有固定的密钥，双方进行通信时，使用代码中固定的密钥进行加密和解密。针对解密后的内容，可以通过单包指纹特征匹配进行应用类型识别。

b）密钥在净荷中：在报文交互过程中，密钥存在净荷中，位置、长度不固定，但在净荷内有字段与之有关。可通过相关字段计算得到密钥的长度、位置，从而获得双方交互的密钥。针对解密后的内容，可以通过单包指纹特征匹配进行应用类型识别。

c）私有数据流传输格式：采用私有自定义数据流传输格式的应用程序，可以通过提取加密流或者加密数据流的特征字段，并在多字段的组合的基础上，进行加、减、与、或、异或等运算后，进行规则匹配，达到应用类型识别的效果。

2.7 基于机器学习的加密流量识别

采用强加密手段的应用无法通过逆向解析识别，可以通过机器学习，研究出这个应用的加密流量的数据模型。

不同应用类型的流量特征不同。如图3所示，语音流的流量特点是平均报文长度短，上下行流交替出现。网页浏览是上行方向发出请求后，下行方向传来对应网页，其特点是平均报文长，连续报文较少。对于视频流的特点是平均报文长，连续报文较多。

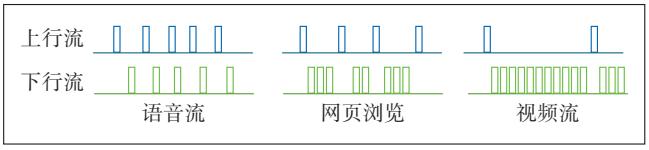

图3　不同应用类型流量特征对比

基于上述不同应用的流量特点，运营商可以针对流统计特征和时序特征，建立不同应用与数据流量特征对应关系的模型特征库，通过机器学习，积累大量的数据流和数据包特征信息作为训练集，根据决策树算法，对加密流量进行类型识别。流统计特征和时序特征包括：上下行数据包流量特征、数据包间隔、数据包的离散性、数据包长度分布、时间戳信息、私有加密协议的特殊端口等。

在机器学习中，应用流量识别算法可分为浅层学习和深度学习[2]。深度学习采用训练多个单层非线性网络，组合底层特征构成数据的抽象表示，从而表达数据的本质特征，在加密的流量识别中表现良好。目前主流的深度学习方法有自动编码器（SAE）、卷积神经网络（CNN）、循环神经网络（RNN）、深度置信网络（DBN）等。近年来，这些方法被广泛应用到加密流量识别中，并取得不错的效果。

图4是一个基于卷积神经网络的应用流量识别的机器学习模型。在该模型中，原始数据包输入卷积神经网络模型经过重复训练，提取应用统计模型。目前主要有以下2种训练模式。

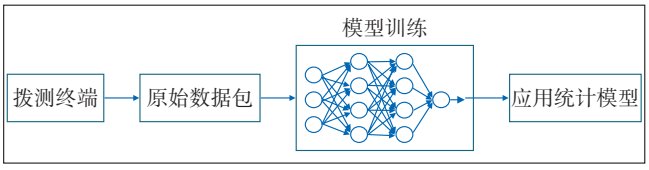

图4　卷积神经网络结构

a）基于数据报文流特征进行训练，包括报文间隔、数据报文长度、窗口大小等统计级别数据。

b）基于数据报文的内容进行训练，主要基于数据包净荷数据前段中的内容进行训练。

3 复杂网络环境下加密流量应用识别系统

基于上述研究的结果，本文设计了一种适用于骨干出口网络加密流量的应用识别系统。加密流量应用识别系统如图5所示。

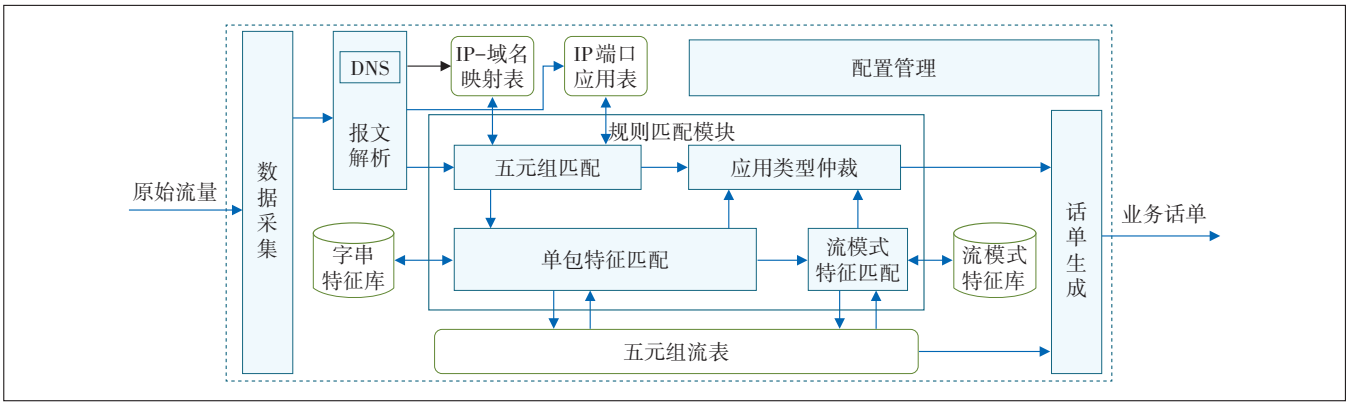

图5　加密流量应用识别系统框图

加密流量应用识别系统主要由数据采集、报文解析、特征库、五元组流表、规则匹配模块、话单生成模块和配置管理模块组成。

3.1 数据采集

数据采集模块使用DPDK等高性能报文处理技术，针对链路上的实时数据流量进行高速采集和预处理。预处理过程中，首先将非对称流量进行同源同宿处理，然后通过哈希算法将流量负载分担至多个CPU核心，以此提高数据处理性能和准确性。

3.2 报文解析

报文解析模块按不同协议对报文进行解析，提取出五元组信息、元数据信息以及针对不同协议的关键特征字段进行应用识别。针对数据流中只有数据内容，没有协议特征字段的协议，采用如下方式，进行加密流量的多流关联识别。

a）对于DNS报文，提取出IP和域名的映射关系后，记录到IP—域名映射表中，待后续域名回填应用类型时使用。

b）对于FTP等在控制面中携带数据面端口信息的应用，提取出IP端口信息，记录到IP端口应用表中，待数据面报文到达后，可以进行跨流的多流关联识别。

c）对于同时支持加密和不加密的应用，在其明文传输数据时，锁定相应端点的IP端口信息，待相同IP端口加密流量到达后，可以进行跨流的多流关联识别。

3.3 特征库

特征库包括字串特征库和流模式特征库2种。

字串特征库中存储的是各类应用报文的指纹特征字串，例如域、URL、16进制数串等，主要用于单包报文的匹配。字串匹配方式包括固定偏移位置匹配、浮动位置匹配和域名等协议字段匹配。字串特征支持多个字串相与运算的操作，组合成复杂字串特征。

流模式特征库中存储的是流统计特征（通过机器学习获得的各类应用的多维度流量特征），包括报文间隔、报文长度分布、报文离散度、跨包流特征组合等信息。

3.4 五元组流表

五元组流表用于记录一个五元组流的流信息，这些信息分为通用信息和协议信息2类。通用信息是数据传输过程中的通用流量特征，例如TCP建链时延、服务访问应答时延、平均报文长度、报文长度分布、时间间隔分布、序列号增长、丢包重传计数、报文乱序计数等信息。协议信息是根据应用协议类型存储与协议有关的相应字段。五元组流表中存储了同一个五元组流上的相关流信息，对于同一流上先后到达的报文，可以通过五元组流表进行关联识别。

3.5 规则匹配模块

规则匹配模块负责完成采集的数据与系统内静态或动态的特征库匹配操作，从而完成流量数据的分类和智能分析。规则匹配模块是整个系统的核心模块，规则匹配算法的优劣决定了整个系统的处理性能的高低。

规则匹配模块由五元组匹配、单包特征匹配和流模式特征匹配组成。

a）五元组匹配支持按5个元组的组合来进行匹配。本系统内容主要是针对IP、IP端口进行匹配。通过IP查询IP—域名映射表，再用查询到的域名，送入单包特征匹配单元。通过IP端口表的查询，进行多流关联分析，提高加密流量识别率。

b）单包特征匹配支持按报文净荷内容匹配字串特征库预定义的特征字串。字串匹配方式包括固定偏移位置匹配、浮动位置匹配和域名等协议字段匹配。字串特征支持多个字串相与运算的操作，组合成复杂字串特征。

c）流模式特征匹配支持按五元组流表记录的流统计信息匹配流模式特征库中流量模型。流模式特征匹配支持基于五元组流表的多包关联识别，同时也支持流特征的多维组合智能分析，根据五元组流表中的流量模型，识别出相应的应用类型，提高加密流量的识别率。

规则匹配模块中各个匹配单元均根据各类协议应用的规则库，进行流量识别和分类。为了提高应用识别的准确度，在规则匹配模块中，针对某些应用类型有多重匹配规则。模块中设计应用类型仲裁单元，针对不同匹配规则制定优先级，采用按优先级加权重的优化算法，智能进行应用识别和分类，并将结果填写到五元组流表中，最终进一步提升应用识别的准确性。

3.6 话单生成

原始数据流量在五元组流表中完成相关流信息的关联识别后，五元组流信息将从流表中删除，并将结果输出到话单生成模块，进行五元组表项格式化整理，再送给话单大数据平台。五元组流话单的内容包含了流的完整关键信息，全量完整的话单后期可以在大数据平台进行多链路、多场景关联分析，可以进一步提升加密流量识别率。

3.7 配置管理

配置管理模块实现对系统的配置管理和对外接口功能，包括系统运行状态监控、系统告警、系统参数配置、系统升级、规则库在线升级等功能。

3.8 系统部署

复杂网络环境下加密流量识别系统的网络部署如图6所示。

系统在运营商网络中部署应用时，考虑运营商网络设备多为双局址部署，涉及多地多链路环境下的系统应用，会导致非对称流量没有完整识别，影响应用识别的准确率，为保证最终的识别效果应考虑以下情况。

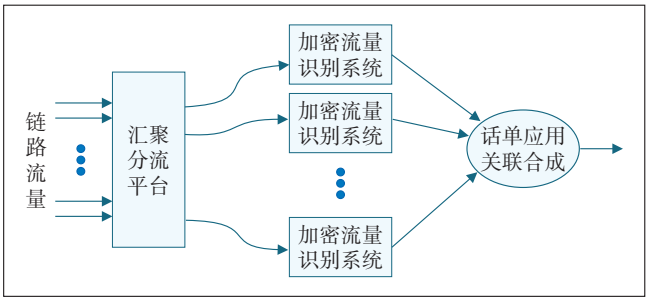

图6　多链路环境加密流量识别系统部署图

a）所有链路到同一个机房：系统采用集中部署，将多地的链路通过传输电路，汇聚至同一节点，接入汇聚分流平台，在进行应用识别前，进行上下行流量的关联。

b）同一个出口有多个机房：所有机房的加密流量识别系统输出话单到省级中心数据平台。针对跨机房产生的非对称流量，由数据分析平台对话单进行关联合成，提升加密流量识别效果。

4 结束语

本文针对加密流量应用类型识别方法进行了研究和探讨，采用机器学习的方式，对明文流量和加密流量进行特征提取和行文分析，提炼出流量特征知识库，设计了一种复杂网络环境下加密流量应用识别系统，利用字串特征库和流量特征库，在非对称网络环境中，对加密流量进行精准分类。随着网络和互联网业务的发展，加密流量行为分析的算法、流量模型建立和学习还需要进一步深化研究和测试。

参考文献：

[1] 陈雪娇，王攀，俞家辉. 基于卷积神经网络的加密流量识别方法[J]. 南京邮电大学学报(自然科学版)，2018，38(6)：36-41.
[2] 苏钰. HTTPs加密流量识别方法研究[J]. 信息通信，2019(6)：81-82.
[3] 曾勇，吴正远，董丽华，等. 加密流量中的恶意流量识别技术[J]. 西安电子科技大学学报(自然科学版)，2021，48(3)：170-187.
[4] 程永新，张德治，廖竣锴，等. 一种加密流量行为分析系统的设计研究[J]. 通信技术，2020，53(4)：976-980.

作者简介：
童博，高级工程师，硕士，主要从事新型IP网络技术、SDN及网络创新产品的研究、研发工作；施俊，工程师，硕士，主要从事新型IP网络技术研究、智能网络运营相关工作；赵纯熙，助理工程师，硕士，主要从事网络创新产品研发设计工作。