基于Flow的电子资源网络行为分析实践

何海涛，黎恩磊，赵琼，姚仁龙，韦雨君
（中山大学网络与信息中心，广东 广州 510275）

摘　要：在高校环境中，电子资源的访问情况分析对资源管理和优化至关重要。基于此，提出了一种基于流量长短流分类和聚合的方法，通过对流量进行精细化分类和统计，准确评估不同类型用户的使用行为。首先，对校园网访问数据库相关的流量进行了统计分析；接着，构造了访问scihub的流量数据并进行了统计分析；最后，根据电子资源访问的流量的统计分析，提出了流量长短流分类和聚合的方法，该方法可以有效对电子资源的访问行为进行分类。

关键词：流量分析；加密流量检测；电子资源；安全分析
中图分类号：TP393
文献标志码：A
DOI：10.11959/j.issn.1000−436x.2024242

Flow-based electronic resource network behavior analysis practice

HE Haitao, LI Enlei, ZHAO Qiong, YAO Renlong, WEI Yujun
Network and Information Center, Sun Yat-sen University, Guangzhou 510275, China

Abstract: In a university environment, analyzing the access patterns of electronic resource is crucial for resource management and optimization. A method based on the classification and aggregation of network flows into long and short flows was proposed. By finely classifying and statistically analyzing traffic, it accurately assessed the usage behaviors of different types of users. First, a statistical analysis of traffic related to electronic resource access was performed on the campus network.Then, traffic data related to accessing Sci-Hub was constructed and analyzed. Finally, based on the statistical analysis of electronic resource access traffic, a method for classifying and aggregating long and short flows was proposed. This method effectively classifies electronic resource access behaviors.

Keywords: traffic analysis, encrypted traffic detection, electronic resource, security analysis

0　引言

随着信息技术的不断发展，高校和科研院所对论文数据库等电子资源的使用越来越多[1]。电子资源的使用在方便了高校师生和科研人员的研究工作的同时，也给高校电子资源的管理带来了很大的挑战，近年来高校学生违规使用数据库现象时有发生[2]。

高校校园网通常只有固定数量的公网IP地址，校内用户通过网络地址转换（NAT）访问互联网。高校用户连接校园网后，可以访问学校图书馆已购买的电子资源。当电子资源提供商检测到用户存在电子资源过量下载行为后会封禁学校的公网IP，从而影响其他正常用户访问电子资源。因此有必要通过技术手段检测校园网中的电子资源过量访问行为并进行阻断。

传统的电子资源过量下载通常基于HTTP非加

收稿日期：2024−10−30
通信作者：黎恩磊，lienlei3@mail.sysu.edu.cn

密数据[3-4]或者使用正向代理来获取解密的HTTPS流量[5]，很少有直接针对加密流量本身进行电子资源访问情况的研究。当前大多数电子资源供应商都使用HTTPS来提供服务，因此需要通过对加密流量的分析来获取用户电子资源的访问情况。本文通过在校园网边界部署流量采集设备，并使用流量分析平台对流量数据进行采集和解析，提取TCP流量的统计信息。通过对加密的TCP流量进行统计分析，本文提出了长短流分类和聚合的方法，通过对流量的聚合，提高了流量分类和检测的速度，并且可以有效区分不同的电子资源访问行为。

1　数据采集方法

本文基于中山大学流量大数据平台进行流量的采集和解析[6]，通过对TCP流量解析后得到双向TCP Flow[7]。Flow由五元组唯一标识，Flow起始于TCP的3次握手，终止于TCP的FIN或RESET。TCP Flow主要包括TCP持续时间内收、发双向的连接时间、响应时间、分组数、字节数及持续时长等。

基于中山大学图书馆已订购的电子资源目录，参考中山大学图书馆提供的电子资源使用情况，选取了表1中的10个代表性的电子资源平台作为分析的目标，类别覆盖了综合类、人文社科类、理工类、医学类。以上述10个电子资源平台为目标，采集了2024年2月1日至2024年7月31日共半年的流量数据。

表1　　选取的10个电子资源平台

序号	电子资源平台域名
1	www.thieme-connect.com
2	onlinelibrary.wiley.com
3	www.jstor.org
4	mathscinet.ams.org
5	www.nature.com
6	www.sciencedirect.com
7	ieeexplore.ieee.org
8	www.clinicalkey.com
9	pubs.acs.org
10	dl.acm.org

2　流量分析

2.1　流量总体情况分析

本文基于2024年2月1日至2024年7月31日中山大学校园网访问电子资源的流量进行分析，共采集7 241 975条双向TCP Flow。表2统计了采集的数据流的统计信息，主要包括是发送流量和接收流量的详细数据。

表2　　采集的数据流的统计信息

类别	发送流量	接收流量
包数量/个	954 587 435	1 481 793 030
字节数/B	143 438 043 958	2 068 248 815 372
平均流长度/包	131.8	204.6
平均流大小/B	19 806.5	285 591.8
平均包大小/B	150.3	1 395.8

从表2中可以看出，对于电子资源的访问来说，发送和接收的字节数不对称，接收的字节数是发送字节数的14倍，符合电子资源访问的行为特征。发送和接收流量的平均流长度比较接近，分别是204.6和131.8，接收流量的平均流大小是发送流量的14倍，接收流量的平均流包大小是发送流量的9倍。

2.2　流量详细情况分析

本文统计了每天接收和发送字节数的统计信息，如图1所示。从图1可以看出，除4月17日外，其余日期的接收字节数均大于发送字节数。对于大多数时间，工作日的发送流量和接收流量大于节假日的流量，符合正常的作息规律。7月暑假的电子资源访问流量明显高于3月到5月的上课时间，但活跃的源IP数量更少，暑假期间通过VPN来访问电子资源的流量有所增加，且单个IP的电子资源访问量也有所增加。

2024年3月是寒假后的第一个月份，图2是2024年3月第一周中山大学校园网访问电子资源平台的接收流量按小时统计情况，其中0点—7点的休息时间接收字节数小于正常的工作时间，符合校园网流量的潮汐规律。

本文统计了流量接收字节数和接收包的分布情况，如图3所示。由图3可以看出，有90.399 6%的流量接收字节数不大于65 536 B，这部分流量的数据包数量仅占全部流量的3.429 3%，接收字节数仅占

全部流量的1.4597%；有96.046 0%的流量数据包不大于256个，这部分流量的数据包数量仅占全部流量的9.460 3%，接收字节数仅占全部流量的5.964 6%。由此，可以得出结论，在校园网访问电子资源的流量中，绝大多数流量的字节数和数据包数量比较小，少部分长流量占据了大部分接收字节数。

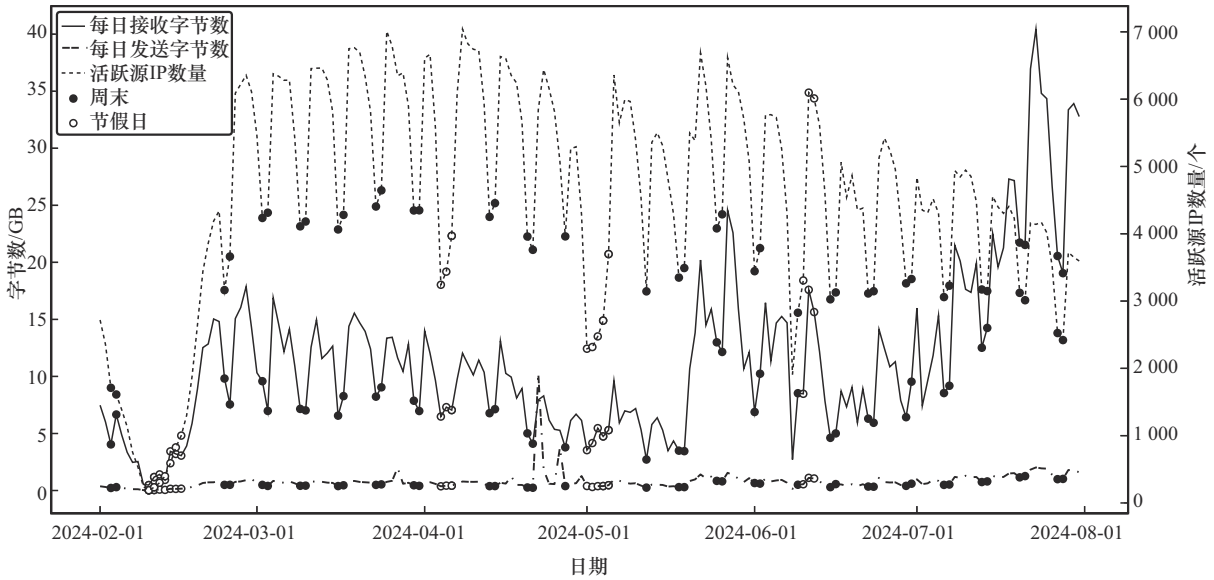


图1　接收和发送字节数的趋势

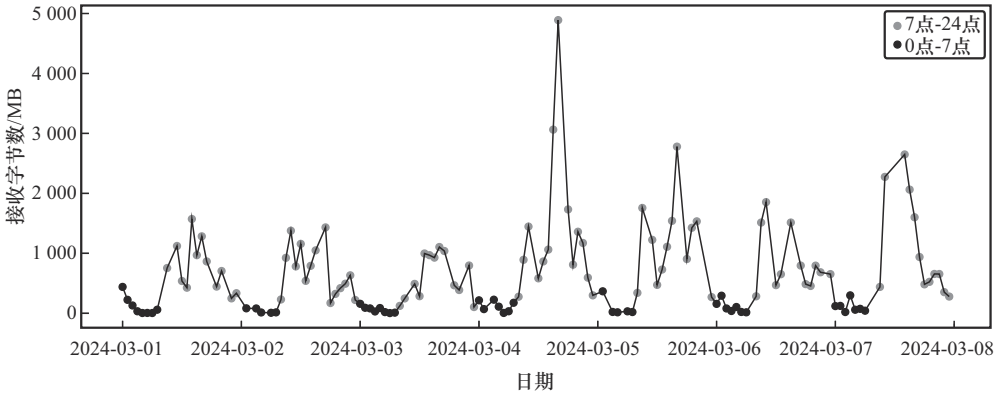
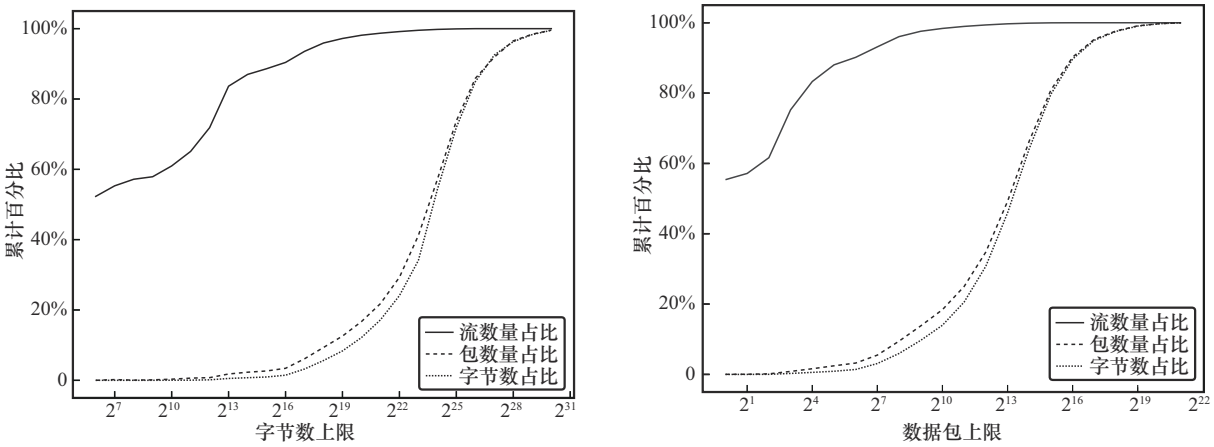


图2　2024年3月第一周接收字节数情况统计



(a) 流量接收字节数累计分布　　(b) 流量数据包累计分布

图3　接收流量字节数和数据包累计分布情况统计

3　长短流聚合算法

3.1　流划分和聚合算法

3.1.1　流划分方法

Thompson 等[8]提出了广域网流量模式和特性分析方法。Brownlee 等[9]提出了网络流的概念，并提出了可以根据流的持续时间和字节数 2 个维度对流进行分类。Fu 等[10]提出了基于交互图的恶意流量检测方法，通过对流量的分类和聚合来实现恶意加密流量的检测。本文使用 Python 爬虫从 sci-hub 下载了 10 239 篇文献，并详细分析了下载的论文的 pdf 文件大小的分布情况，如图 4 所示。从图 4 中可以看出，只有 11.48% 的文件小于 100 KB，对应的文件大小总和占所有文件的 0.77%。结合以上数据分析结果和中山大学校园网访问情况的分析，可以根据流的接收字节数将 TCP Flow 为长流和短流。

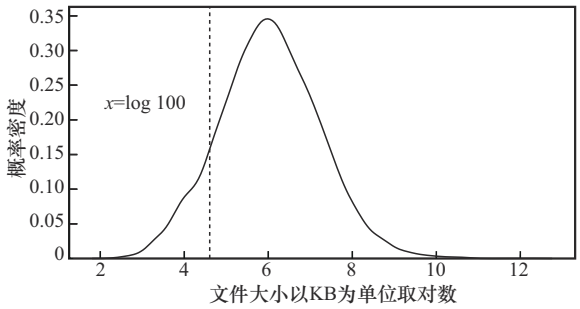


图 4　scihub 论文 pdf 文件大小分布情况

3.1.2　流聚合方法

当用户访问电子资源平台时，通常会访问服务器的 443 端口，源端口则会根据访问过程发生变化。将源 IP 和目的 IP 将相同的流进行聚合处理，可以得到一条聚合后的新流，通过统计分析聚合后新流的特征，可以进一步对用户的电子资源访问行为进行分类。同时，对于电子资源的访问，本文不关注短流。因此，可以定义 FLOW_LINE 为长短流分类的阈值，将接收字节数大于 FLOW_LINE（默认为 102 400）的流定义为长流，反之为短流，该值可以根据不同电子资源平台的资源特性进行调整。对于流量中占比超过 90% 的短流，直接做丢弃处理；对于长流，则根据流量聚合算法做流量聚合。算法 1 描述了详细的长短流分类和聚合算法，其将 TCP Flow 中源 IP 和目标 IP 相同的长流进行聚合处理，将时间间隔小于 JUDGE_INTERVAL（默认间隔为 10 min）的多条长流合并为一条长流，并提取流的总接收字节数和长流的数量作为聚合后流的数量。

算法 1　长短流分类和聚合算法

输入　TCP 双向流 Flows

输出　聚合的长流特征 FlowFeatures

```
1) for flow in Flows do
2)   if flow.rxbytes < FlOW_LINE then
3)     continue
4)   根据源和目的IP把流加到SrcDstTable
5)   // 长流聚合
6)   for flows in SrcDstTable do
7)     // SrcDstFlowList中源和目标IP相同
8)     if time.now-flows[-1].endtime >间隔 then
9)       添加特征到FlowFeatures
10)      clear SrcDstFlowList
11) return FlowFeatures
```

3.2　流量聚合效果

本文对中山大学 2024 年 2 月到 7 月的 7 241 975 条 tcpflow 做了长短流分类和聚合处理，其中包括 957 257 条长流，占比 13.2%，按照 FLOW_LINE=102400 和 JUDGE_INTERVAL=10 分钟的参数配置对流量做分类和聚合处理，结果得到 774 520 条聚合后的流，聚合后的流只占原始流量的 10%。从图 5 可以看出，大多数的聚合后的流的接收字节数和流的持续时间都不大，部分异常的流占据了大多数流量。

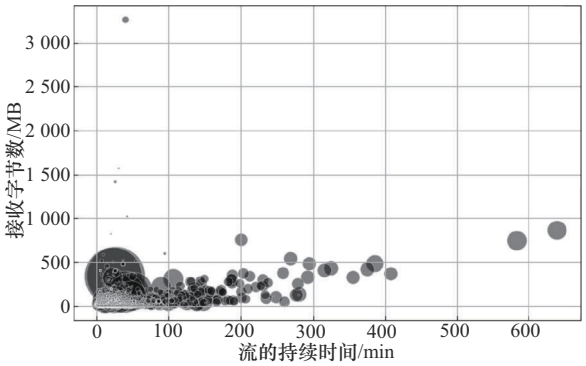


图 5　电子资源访问流量聚合后的分布情况

假设聚合后的流中子流数目大于 25 或者接收字节数大于 25 MB 的聚合流是异常流，异常的聚合流有 12 736 条，接收的总字节数是 669 546 MB，占据了总接收字节数的 35.84%，仅占所有流总数

的1.33%。使用聚合流的特征中的接收字节数和子流数量可以有效区分电子资源访问中的异常行为。

4 结束语

本文首先分析了中山大学校园网的电子资源访问情况，发现大多数流量的接收字节数比较小，访问的流量符合正常的潮汐规律；接着分析了scihub的论文pdf文件的大小分布规律，发现可以根据字节数将流量分为长流和短流；最后提出了长短流分类和聚合的算法，并对校园网电子资源访问流量进行聚合处理，发现聚合后的长流特征可以有效区分流量中的异常行为。

在后续的研究中，笔者将统计更多的聚合流量特征，并应用机器学习异常检测算法，实现电子资源访问情况的实时分析和告警。

参考文献：

[1] 吴汉华, 王波. 2022年中国高校图书馆基本统计数据分析[J]. 大学图书馆学报, 2023, 41(6): 63-72.
WU H H, WANG B. An analysis of basic statistical data of Chinese academic libraries in 2022[J]. Journal of Academic Libraries, 2023, 41(6): 63-72.

[2] 冀春雨. 近年来高校学生违规使用数据库现象多发[R]. 2023.

[3] 刘莉, 冯骐, 汪志莉. 电子资源防恶意下载系统研究: 以华东师范大学为例[J]. 图书馆学刊, 2015, 37(1): 99-102.
LIU L, FENG Q, WANG Z L. Research on anti-malicious download system of electronic resources—taking East China normal university as an example[J]. Journal of Library Science, 2015, 37(1): 99-102.

[4] 沈奎林, 邵波, 杜瑾. 基于网络日志分析的数字资源监测系统的实现[J]. 图书馆学研究, 2015(16): 21-25.
SHEN K L, SHAO B, DU J. The realization of digital resource monitoring system based on network log analysis[J]. Research on Library Science, 2015(16): 21-25.

[5] 陈广. 基于Fiddler代理程序的电子资源使用统计分析系统的设计与应用[J]. 图书情报工作, 2018, 62(13): 30-36.
CHEN G. Design and application of electronic resource use statistical analysis system based on fiddler agent[J]. Library and Information Service, 2018, 62(13): 30-36.

[6] 杨敏, 何海涛, 赵琼. 流量大数据安全分析平台的设计与实现[J]. 通信学报, 2018, 39(S1): 104-109.
YANG M, HE H T, ZHAO Q. Design and implementation of traffic big data security analysis platform[J]. Journal on Communications, 2018, 39 (S1): 104-109.

[7] IN C S. A survey of network traffic monitoring and analysis tools[R]. 2009.

[8] THOMPSON K, MILLER G J, WILDER R. Wide-area Internet traffic patterns and characteristics[J]. IEEE Network, 1997, 11(6): 10-23.

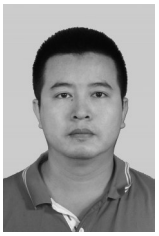
[9] BROWNLEE N, CLAFFY K C. Understanding Internet traffic streams: dragonflies and tortoises[J]. IEEE Communications Magazine, 2002, 40 (10): 110-117.

[10] FU C P, LI Q, XU K. Flow interaction graph analysis: unknown encrypted malicious traffic detection[J]. IEEE/ACM Transactions on Networking, 2024, 32(4): 2972-2987.

[作者简介]

何海涛（1975– ），男，安徽淮北人，博士，中山大学高级工程师，主要研究方向为因特网流量行为、大数据等。

黎恩磊（1995– ），男，河南漯河人，中山大学助理工程师，主要研究方向为数字取证、流量分析。



赵琼（1983– ），男，湖北黄冈人，中山大学工程师，主要研究方向为计算机网络技术。

姚仁龙（1992– ），男，安徽安庆人，中山大学助理工程师，主要研究方向为人工智能、流量分析。

韦雨君（1997– ），女，贵州贵阳人，中山大学助理工程师，主要研究方向为网络安全、流量分析。