

毕业论文（设计）开题报告

|  |  |
| --- | --- |
| 题 目 | 基于DPI技术的用户上网行为感知与分析 |

|  |  |
| --- | --- |
| 学生姓名： | 徐正 |
| 学 号： | 2002932221 |
| 学 院： | 计算机与软件学院 |
| 专 业： | 软件工程 |
| 班 级： | 软件2026 |
| 指导教师： | 岳兆新 |
| 开题日期： | 2021年 12 月 30日 |
|  | |

**填写说明**

1. 开题报告是毕业论文（设计）答辩委员会对学生答辩资格审查的依据材料之一。学生应当在毕业论文（设计）工作前期内完成，开题报告不合格者不得参加答辩。

2. 本报告中，由学生本人撰写的对课题和研究工作的分析及描述，没有经过整理归纳，缺乏个人见解仅仅从网上下载材料拼凑而成的开题报告按不合格论。

3. 开题报告的字数不少于2500字（艺术类专业不少于1500字），其中，文献综述字数不得少于1000字，开题报告的格式按学校《本科毕业论文（设计）撰写规范》的要求撰写。

4. 有关年月日等日期的填写，应当按照国标GBT 7408—2005《数据元和交换格式、信息交换、日期和时间表示法》规定的要求，一律用阿拉伯数字书写。如“2022年3月1日”或“2022-03-01”。

|  |  |
| --- | --- |
| **一、选题的背景和意义：**  随着网络服务的不断发展，因特网出现了许多网络新技术和新业务流量， 例如IPTV、网络聊天、视频网站、在线游戏、电子商务等。这些业务流量大都可被分为P2P(Peer-to-Peer)和VoIP(Voice over IP)两大类。由于拥有内容自由和呼叫便利等优点，这些新业务吸引了众多用户，且还在不断增长。由于网络运营商需要对其资源进行优先排序，并根据网络拥塞和业务流量来调整资源分配进度, 因此，不断增加业务给网络运营商带来了巨大的压力。寻找一种有效、简单快速的业务流量识别方法，并根据网络状态和业务流量采取不同的管理政策, 成为解决上述问题的关键。  深度包检测(DPI)是一种通过检测数据包来识别业务流的网络数据包检测技术。当IP数据包通过基于DPI技术的带宽管理系统时，该系统通过深入读取IP包载荷的内容来对应用层的信息进行重组，从而实现对业务流的识别和控制.然而, 业务检测任务仍然是具有挑战性的问题，目前还没有通用的标准算法和实现技术.一方面, 大都应用程序都没有公开其使用的协议, 要找出其底层协议需要大量的逆向工程。另一方面，许多程序员会采用端到端的加密, 混淆, 不按规则使用知名端口等方法，使得其应用越来越隐蔽。因此，要实时的识别这些业务，DPI识别方法必须快速、准确、并且容易升级。  DPI系统识别业务流量的传统方法有很多，包括:基于熟知端口识别、基于特殊字节系列识别、基于流量统计识别和基于流量行为识别等。上述方法都有各自的优势和局限性。基于端口的识别是最快又最简单的方法，但是由于很多新的业务会随机的改变端口，该方法的识别效率将大大降低，准确率也只有30%~70%。基于有效载荷签名的识别是通过发现独特的签名字节模式以识别各自的流量.该方法在纯文本流量识别中能取得很好的效果。但是当流量是加密或混淆时, 基于签名的识别方法也将不再适用。基于流量统计的识别是通过使用流量的统计参数来识别业务流，适用于纯文本流量和加密流量的识别。但是这种方法的精确度很大程度上依赖于训练数据集和有效的最近匹配算法。基于流量行为的识别方法通过检测按时间顺序到达的数据包组中会话数据包发送和接收行为来识别业务流.它需要进行以详细的时间为基础的流量分析，适用于纯文本流量和加密流量的识别，但大多只对VoIP流量分析有效。 | |
| **二、主要研究内容：**  本文主要基于DPI技术，首先对网络上的传输数据进行抓取，再对抓取的网络数据进行分析处理，最终实现用户上网行为感知与分析的目标。主要研究内容如下：  （1）网络通讯的真实内容分析  （2）网络故障分析  （3）程序网络接口分析  （4）木马通讯数据内容分析  抓包工具是拦截查看网络数据包内容的软件。通过对抓获的数据包进行分析，可以得到有用的信息。流行的抓包工具有很多，比较出名的有wireshark、sniffer、httpwatch、iptool等。这些抓包工具功能各异，但基本原理相同。我们的计算机通过向网络上传和从网络下载一些数据包来实现数据在网络中的传播。通常这些数据包会由发出或者接受的软件自行处理，普通用户并不过问，这些数据包一般也不会一直保存在用户的计算机上。抓包工具可以帮助我们将这些数据包保存下来，如果这些数据包是以明文形式进行传送或者我们能够知道其加密方法，那么我们就可以分析出这些数据包的内容以及它们的用途。  抓包工具更多的用于网络安全，比如查找感染病毒的计算机。有时也用于获取网页的源代码，以及了解攻击者所用方法、追查攻击者的ip地址等。 | |
| **三、方案（设计方案、或研究方案、研制方案）论证：**  近年来，伴随着互联网业的飞速发展，与互联网相关的产业呈多元化趋势，大量新生应用程序充斥着网络市场。因此，对网络上的数据流量分布和业务使用情况进行进一步的细分显得尤为重要。同时，一些企业也需要优化网络管理。例如，在高校，通常有互联网接入的高峰期，网络设备的CPU利用率高，导致网络拥塞。  应用层协议识别系统的设计需要考虑两种类型的人的需求，一种是一般用户，另一种是网络管理者。对于普通用户，该系统需要具有良好的交互式可视化显示功能，类似于科莱网络分析系统，让不具备网络专业技能的用户能够轻松了解网络的运行状态；对于网络管理员来说，除了使用基本的系统功能外，系统还应该具有高度的适应性，即能够适应网络环境的变化，识别新的网络协议，减少系统开销，并且要避免识别一些不必要的网络协议，这就要求系统需要具备维护特征库的功能。  1、功能需求分析  （1）首先该系统要能够成功调用网卡接口或者蓝牙接口以获取准确的源数据包，其次能够准确无误地将数据包里的码流逐一分析，并与应用协议类型相对应，然后要将这些信息清晰地展示在主界面上方便用户查看，最后还要保证识别率的精准以及用户能够对没有识别的特征进行手动添加。  （2）系统界面具有良好的交互性，适用于不同年龄段的用户，能够使用户快速上手。  （3）特征库必须可以适应复杂的网络环境，管理员可以根据需要，对特征库中已有的规则类型进行增删改查功能。  2、性能需求分析  （1）提高识别的准确率。  （2）要对系统进行适当的压力测试，来保证可以在严苛的条件下稳定运行。 | |
| **四、研究的总体安排和进度计划：**  1-2周：查看文献，与老师讨论，收集相关信息，完成开题报告  3-4周：进行需求分析和可行性分析，明确研究方向，完成需求报告  5-6周：明确所需数据，收集材料，准备下一阶段。  7-10周：进行总体方案设计，框架设计，模块设计，完成详细设计报告。  11-14周：进行多个应用数据采集，保留数据包进行数据流量识别研究。  15-16周：收集每个应用的规则特征，创建新的特征库。  16-17周：实现特征库动态更新策略。  18-20周：完善毕业论文的编写，修改等。  21周：毕业答辩。 | |
| **五、文献综述（或调研报告）：**  DPI 全称为“Deep Packet Inspection”，称为“深度包检测”。所谓“深度”是和普通的报文分析层次相比较而言的，“普通报文检测”仅分析IP包的层4 以下的内容，包括源地址、目的地址、源端口、目的端口以及协议类型，而DPI 除了对前面的层次分析外，还增加了应用层分析，识别各种应用及其内容。  **DPI技术原理**  DPI 的技术关键是高效的识别出网络上的各种应用。普通报文检测是通过端口号来识别应用类型的。如检测到端口号为80时，则认为该应用代表着普通上网应用。而当前网络上的一些非法应用会采用隐藏或假冒端口号的方式躲避检测和监管，造成仿冒合法报文的数据流侵蚀着网络。此时采用L2~L4层的传统检测方法已无能为力了。  DPI 技术就是通过对应用流中的数据报文内容进行探测，从而确定数据报文的真正应用。因为非法应用可以隐藏端口号，但目前较难以隐藏应用层的协议特征。 DPI的识别技术可以分为以下几大类： **（1）基于“特征字”的识别技术**  不同的应用通常依赖于不同的协议，而不同的协议都有其特殊的指纹，这些指纹可能是特定的端口、特定的字符串或者特定的Bit 序列。基于“特征字”的识别技术通过对业务流中特定数据报文中的“指纹”信息的检测以确定业务流承载的应用。  根据具体检测方式的不同，基于“特征字”的识别技术又可以被分为固定位置特征字匹配、变动位置的特征匹配以及状态特征匹配三种技术。 通过对“指纹”信息的升级，基于特征的识别技术可以很方便的进行功能扩展，实现对新协议的检测。  DPI的关键在于，它要不断地在格式不定的数据包中判断出各种特征字，实现这一过程的基础技术就是模式匹配（Pattern-Matching）即字符串匹配，就是从数据中搜索是否存在目标字符串。通常，目标字符串采用正则表达式（Regular Expression）标准语法来描述。  **（2）应用层网关识别技术**  某些业务的控制流和业务流是分离的，业务流没有任何特征。这种情况就需要采用应用层网关识别技术。  应用层网关需要先识别出控制流，并根据控制流的协议通过特定的应用层网关对其进行解析，从协议内容中识别出相应的业务流。对于每一个协议，需要有不同的应用层网关对其进行分析。如SIP、H323协议都属于这种类型。SIP/H323通过信令交互过程，协商得到其数据通道，一般是RTP格式封装的语音流。也就是说，纯粹检测RTP流并不能得出这条RTP流是那通过那种协议建立的。只有通过检测SIP/H323的协议交互，才能得到其完整的分析。 **（3）行为模式识别技术**  行为模式识别技术基于对终端已经实施的行为的分析，判断出用户正在进行的动作或者即将实施的动作。行为模式识别技术通常用于无法根据协议判断的业务的识别。例如：SPAM（垃圾邮件）业务流和普通的Email业务流从Email的内容上看是完全一致的，只有通过对用户行为的分析，才能够准确的识别出SPAM业务。  以上三种识别技术分别用于不同类型协议的识别，无法相互替代。而目前华为公司在应用DPI 技术部署DPI 系统时采用了多业务控制网关MSCG分层DPI 解决方案，综合运用了这三种技术，在检测效率和灵活性方面均达到最优。 **DPI 技术的应用**  利用DPI 技术在IP网络中部署DPI 系统可实现网络运营中的业务识别、业务控制和业务统计三大功能。  **业务识别**  一般而言，对于业务识别有两种方法，一种是对运营商开通的合法业务，另  外一种是运营商需要进行监管的业务。  前者可以通过业务流的五元组来标识，如VOD业务，其业务流的地址是属于VOD  服务器网段的地址，其端口是一个固定的端口。系统一般采用ACL的方式，识别  出该类业务。  后者需求DPI技术，通过前述的业务识别方法，通过对IP数据包的内容进  行分析，通过特征字的查找或者业务的行为统计，得到业务流的类型。  **业务控制**  通过DPI 技术识别出各类业务流之后，根据网络配置的组合条件，如用户、  时间、带宽、历史流量等，对业务流进行控制。控制方法包括：正常转发、阻塞、限制带宽、整形、重标记优先级等。  为了便于业务的运营，业务控制策略一般集中配置在策略服务器中，用户上  线后动态下发。  **业务统计**  DPI 的业务统计功能是为了直观的统计网络的业务流量分布和用户的各种业  务使用情况，从而更好的发现促进业务发展和影响网络正常运营的因素，为网络和业务优化提供依据。如：发掘对用户有吸引力的业务、验证业务提供水平是否达到了用户的服务等级协议SLA、统计分析出网络中的攻击流量占多少比例、多少用户正在使用某种游戏业务、哪几种业务最消耗网络的带宽和哪些用户使用了非法VOIP等等。  **DPI 技术的发展**  可以看出，DPI 的检测技术和网络上非正常应用的反检测是矛和盾的关系。  前面谈到的DPI技术不是静止不变的，随着检测技术的发展，非正常应用的隐藏技术也在演进。如对数据部分加密、隐藏特征字和通过隧道技术躲避检测等  等。DPI 技术在发展中将不断调整上述的检测方法，从而达到比较高的检测精  度。总之，DPI 技术将逐渐在安全、业务控制等方面广泛应用，为运营商精细控  制和运营网络提供一种利器。  **参考文献**   1. 王凯,吴君钦.基于DPI技术的IM协议识别系统研究[J].计算机技术与发展,2013,23(11):120-123. 2. 陈亦余,陈爱民.一种基于自主学习的DPI业务识别方法[J].福建师范大学学报(自然科学版),2017,33(03):19-26. 3. 田良,徐孟江.基于DPI的移动网数据深度检测分析实现对恶意程序的监测及处置[J].科学技术创新,2021(31):167-169. 4. 王浩东,成实,张锐,李磊,张能.基于深度数据包检测(DPI)及R平方值的数据承载网规划方法的研究[J].信息通信,2019(05):182-184. 5. 张勇,李泽凯,常有宝.一种基于DPI技术的用户上网行为管理方法[J].电信快报,2018(10):34-36+40. 6. 饶瑾.深度包检测(DPI)技术浅谈及应用[J].信息通信,2014(11):245-246. 7. 申进. 基于DPI和DFI的网络流量分类方法研究与应用[D].西南科技大学,2020.DOI:10.27415/d.cnki.gxngc.2020.000345. 8. 陈贞贞.基于DPI和机器学习的加密流量类型识别研究[J].信息通信,2018(04):258-260. | |
| **指导教师意见：**  指导教师签字：  年 月 日 | |
| **所在专业审核意见：**  专业负责人签名：  年 月 日 | **学院意见：**  教学院长签名：  年 月 日 |