文章编号:1674 - 2869(2018)02 - 0224 - 04

基于负载的协议识别技术在防火墙系统中的改进

邹连英,熊源远,高 宁
武汉工程大学电气信息学院,湖北 武汉 430205

摘 要:传统防火墙的应用识别技术存在无法识别具体的应用类型、难以对同一应用软件进行更细粒度的功能识别和控制等缺陷。通过对当前的各种协议识别技术进行深入研究,提出了一种综合运用基于端口及负载的协议识别方法。该方法增加了更详细的报文特征字段,对应用报文进行了更精确的分类和特征定义,并在防火墙设备中进行了应用验证。验证结果表明,该方法在提高识别效率的同时,保证了识别的准确率。

关键词:匹配规则; 应用识别; 防火墙
中图分类号:TP393.0 **文献标识码**:A **doi**:10. 3969/j. issn. 1674-2869. 2018. 02. 020

Improvement of Load-Based Protocol Identification Technology in Firewall System

ZOU Lianying, *XIONG Yuanyuan*, *GAO Ning*
School of Electrical and Information Engineering, Wuhan Institute of Technology, Wuhan 430205, China

Abstract: The application identification technology in traditional firewall can not identify the type of specific application, and is difficult to identify and control the function of the same application with finer granularity. This paper deeply analyzes the application identification technologies, and proposes a new application recognition method based on port and load, where more detailed message features fields were added, thus the classification and characterization of application message were conducted more precisely. This novel application identification technology was verified in firewall device, demonstrating that it improved the efficiency of recognition and ensured the accuracy of recognition.

Keywords: match rules; application identification; firewall

随着互联网的普及,网络应用越来越丰富,对网络应用进行有效的识别控制[1]也变得越来越复杂。传统的方法基于数据包的地址、端口等信息对网络应用予以识别控制已经不能有效识别网络应用[2]。目前许多网络防火墙产品采用简单的基于端口、测度、负载等方式,已经不能阻止应用层的复杂的攻击,更不能发现新的攻击。这种情况下,人们迫切需要采用新的防火墙技术来防范基于应用层的攻击和威胁。基于这样的背景,笔者对企业所面临的网络安全问题提出的一种全面的解决方案,设计并实现一个将端口及负载的识别方法进行综合运用的智能协议识别系统。

1 传统的协议识别

1.1 基于端口的协议识别

基于端口的协议识别使用IANA规定[3]的固定端口号来进行相应的应用层协议识别[4],是大家都非常熟悉的,也是常用的。但缺点是对于识别采用随机端口通信的应用层协议无能为力,其识别准确度也越来越低,无法满足需求。

1.2 基于测度的协议识别

基于测度的协议识别理论上可以识别所有的

收稿日期:2017-11-25
作者简介:邹连英,博士,副教授。E-mail: 55270072@qq.com
引文格式:邹连英,熊源远,高宁. 基于负载的协议识别技术在应用识别系统中的改进[J]. 武汉工程大学学报,2018,40(2):224-227.

协议，但是不同的协议其规范不同，目前所研究的算法只能将流分类，并不能精准识别。基于测度识别协议无需分析报文体的内容，其根据事先已有的标准集来判断当前流所属的分类[5]，其优点是不需要检测报文体的内容，消耗计算机资源相对较小，但是对于行为相似的应用协议分辨率较差，不适合在线处理的应用环境。

1.3　基于负载的协议识别

对每一个到来的数据包进行精准匹配[6]，当有新的协议到来时，必须对协议的每一负载值重新匹配一次。其优点是匹配精确，而缺点是工作量非常大，对特征码识别的算法要求高。

1.4　协议识别技术比较

三类协议识别技术各有优缺点，其性能的比较如表 1 所示。基于端口的协议识别技术虽然实现简单，但是随着网络的快速发展，其识别准确度也越来越低，无法满足需求；而基于测度的协议识别易于管理与维护[7]，但该算法实时性低，限制了其在具体环境中的应用；基于负载的协议识别技术在识别的准确性和实时性上较好，但开发和维护需要很大的工作量，且复杂度过高，不适宜在实际环境中使用。

表 1　协议识别技术性能比较

Tab. 1　Performance comparison of protocol identification technologies

协议识别技术	准确性	实时性	吞吐量
基于端口	较低	高	高
基于测度	较高	较低	较高
基于负载	高	较高	较低

在对三种识别技术的综合应用进行研究后，将对各应用协议传输的端口、报文内容等特征进行识别，准确性、实时性、吞吐量都有所提高。

2　应用识别匹配规则设计

本文所采用的协议识别方法根据数据包的内容识别，并对数据包的端口等所有内容都要进行匹配。特征字段总是存在于负载的某一个区间内，在基于负载的协议识别的基础上，对数据包进行深度分析，添加减少匹配复杂度的字段[8]，来提高识别效率与准确度。传统的协议识别方法主要是研究会话的唯一性描述，但在对多种数据包进行分析时看出，除了按照已经被定义的方法，还可以总结出更多的规则添加到不同的方法中，来提高应用识别的效率与准确度。

首先可以对应用层不同功能及类型的应用进行分类，并对每种类型的应用进行编码。由于网络的快速发展，应用层软件层出不穷，且无时不刻不在更新。为了协议分析人员能够清晰明了的分辨某一应用的所属分类，可以简单地将功能不同，版本不同的应用进行归纳[9]，可通过应用名字和 ID 来唯一确定该规则适用的应用类型，如表 2 所示。

表 2　应用场景分类

Tab. 2　Classification of application scenarios

应用分类码	应用类型名	中文释义
0	General	杂项
1	Web mail	网页邮件
2	Game	游戏
3	Remotecontrol	远程控制
4	IM	及时通信
5	Nettv	网络电视
6	Stock	股票证券
7	P2P	P2P下载
8	Cloudestorage	云盘存储
9	EM	电子商务
a	Mobiles	移动社区
b	Living	生活服务
c	Update	软件更新
d	Telephone	网络电话
e	Bank	银行

然后，每一个应用特征码后加入相应的应用特征字段。特征字段是对不同协议的具体描述及归纳，由于每个应用所适应的规则差别大，对每个应用的分类可以根据应用的类型和功能来进行细分。如果应用涉及多条规则，则可以通过多个字段来进行控制，此外，开发人员在编辑或更新防火墙规则时，可以通过名字、ID 及字段规则进行一一对应来了解该规则的作用与意义[10]，如表 3 所示。

每一个类别中包含了不同的应用，而每一个应用在规则表中用不同的字段进行表示，各字段的内容限定了它们各自的特征，字段的表示如表 3 所示。

表 3 中最后一项 Sign_detail 内容由多种因素构成，其中包括：

1）IP 类型：表示为×××.×××.×××.×××的字符串格式，如果有多个 IP 用|分隔，如 192.168.1.3|192.168.1.4，最多可以同时表示的 IP 为 20 个；

2）端口类型[11]：S 关键字后跟着是源端口；D 关键字后跟着是目的端口，如 S777|D888；

表3 特征字段定义

Tab. 3 Definition of feature field

表项	说明
App_id	协议序号
App_name	协议名称
Datalen_check	是否需要检测数据长度,0不用,1需要
Datalen_mask	数据长度掩码,0x01表示等于,0x02表示大于,0x04表示小于(Datalen_check为1时有效)
Datalen	数据长度,十六位整型
Sign_seq	特征码匹配序列号,一般为1,表示有净荷的第一个报文
Weight	特征码匹配权重,表示匹配到该特征的权重,一个报文可能匹配到多个特征码,最后依据权重值大小决定是哪种应用(权重越大表示优先级越高)
Proto	特征码IP协议层,0表示TCP,1表示UDP,无符号char型
Sign_type	特征码类型,0x01表示目的IP地址匹配,0x02表示端口匹配,0x04表示是DNS过滤,0x08表示特征值匹配,可以多个类别混合,比如0x0F表示即有目的IP匹配,又有端口匹配,还有特征值匹配
Sign_detail	特征码内容,字符串

3)DNS类型[12]:字符串类型,不超过64个字节,不含通配符等,如www.baidu.com;

4)特征值类型:根据每个报文的不同,通过正则表达式的表示方法将报文内容表示出来,如^H*6+[4:+30]*+\x30$。

3 QQ协议识别测试

3.1 特征字段表设计

腾讯采用OICQ协议对登陆过程进行保护[13]。客户端会先给服务器发送一个数据包请求登录,服务器会返回一个数据包包括了客户端的IP地址,版本信息等内容。在客户端收到此数据包后,紧接着会给服务器发送一个包含登录信息的登录请求。服务器会首先看看客户端的号码、IP址和版本是否合法,如果可以的话,就验证客户端的登录信息是否与服务器上保存的登录信息一致,如果匹配成功就向客户端返回一个登录成功的数据包,不匹配返回登录失败[14]。

QQ协议登录时首选的是UDP,如果UDP不可登陆,那么会再尝试使用TCP进行传输。UDP使用的端口是8000,TCP使用的端口是443,应用协议基本一样,只是在通过TCP进行传输时,前两个字节为协议内容的长度。

在保证网络条件正常的情况下,没有添加禁用应用的时候,对QQ客户端进行测试,QQ客户端都能正常登录,登陆过程中对其登录的数据包进行截取如图1所示。

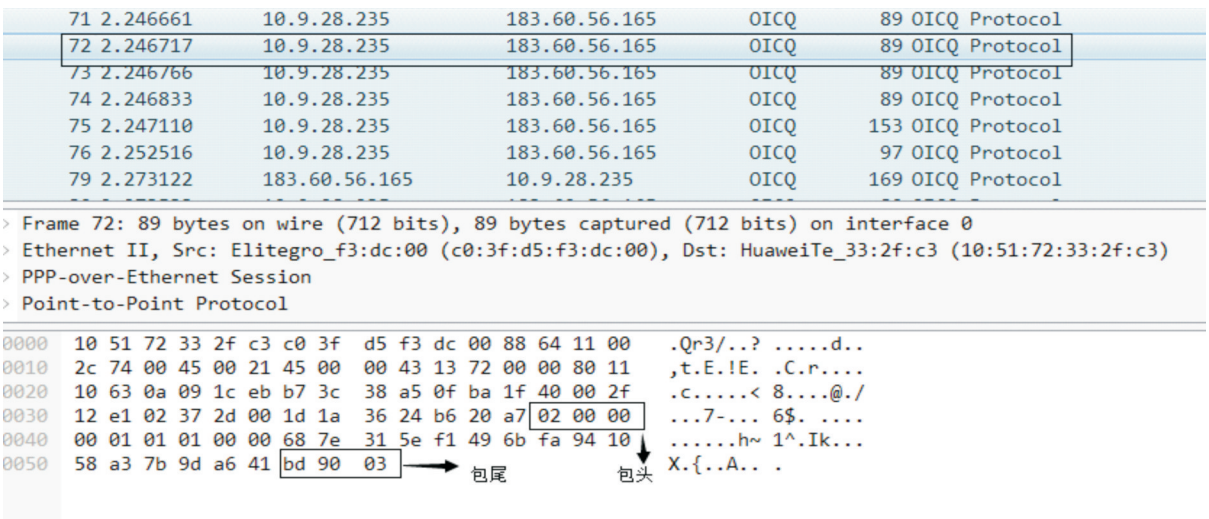

图1 QQ登陆报文

Fig. 1 Packets of QQ login

对QQ登陆过程中的某一个数据包进行分析,可以看见包头是十六进制0x02开始,以十六进制0x03结束。在规则表中可以用正则表达式[15]的语法添加,当特征内容为02 03 00 00 01 01 …… 5b 07 03时,可用正则表达式^0x02*+0x03 $表示。

根据表2和表3的设计,QQ软件属于即时通信类;分类码为4;App_id为0x401;App_name为QQ;Datalen_check为0,表示不需要检查报文长

度；Datalen_mask为0，表示报文长度在一个区间范围；Datalen为0，表示不判断报文长度；Sign_seq为1；Weight为1 000；Proto为0，表示tcp协议；Sign_type为0x08，表示特征值匹配；Sign_detail为^0x02*+0x03$，其中^\x02表示第一个字节为十六进制数0x02，*+表示匹配任何的字符，*+\x30$表示最后一个字节为数字0x30。（一行中间以空格符隔开，所以当特征码中间出现空格时需要用\x20代替。）

因此，QQ应用软件的特征字段表值为：

4 0x401 QQ 0 0 0 1 1000 0 0x08 ^\x02*+\x03$

将此特征字段表项值保存到配置文件，执行命令，对进行的配置下发，添加禁用应用的规则。

3.2　QQ协议禁用规则测试

网络应用识别控制系统的流程分为两大步：首先开发人员需要打开防火墙Web界面，输入正确的账号与密码，登录到防火墙配置界面；然后防火墙界面进行配置，对相应的规则进行增加与删除，并确认相应的网络连接是否正确，最后再进行测试。

在配置了上述QQ应用特征禁用规则后，登录QQ。对QQ客户端进行测试，观察是否能够使用软件，经过测试客户端不能登录。

经过测试网络应用识别系统，可以正确识别QQ应用，使其不能正常使用，可以实现较为精细、准确的控制，对软件的登录进行封堵。

3.3　与传统协议识别技术测试结果的比较

对QQ应用软件分别用传统的协议识别方法进行测试，其结果均不能达到预期的目标。在对端口进行封堵后，并不能有效的控制QQ的登录；而对负载的控制会使识别时间变长，最后也不能使QQ得到控制。综上所述，对传统应用识别技术的改进已取得了良好的效果，在有效性及实时性方面有很大的改善。

4　结　语

通过对基于端口、测度、负载三种识别方法的研究，发现这三种方式均不能满足当下的网络需求。基于端口的应用识别方式虽然实现简单，但是随着网络的快速发展，其识别准确度也越来越低，无法满足需求；而基于测度的协议识别算法实时性低，限制了其在具体环境中的应用；基于负载的协议识别技术在识别的准确性和实时性上较好，但开发和维护需要很大的工作量，且复杂度过高。通过对这三种最常用的应用识别方法的算法进行综合运用，研究出了一套高效率及高准确率的应用识别方案。论文通过实例的QQ登录过程，对该系统设计进行具体的测试验证。实验证明本系统实际运行效果显著，具有良好的实用和推广价值。

参考文献：

[1] 吴鹏冲. 非默认端口网络协议识别系统的研究与实现[D]. 北京：北京邮电大学，2009.

[2] 陈亮，龚俭，徐选. 基于特征串的应用层协议识别[J]. 计算机工程与应用，2006(24)：16-19.

[3] 朱迪. 互联网号码分配机构(IANA)管理权移交进展[J]. 中国教育网络，2016(7)：31-33.

[4] 刘萌. 基于下一代防火墙技术的网络应用识别控制系统设计与实现[D]. 北京：中国科学院大学，2014.

[5] 张洛什，王大伟，薛一波. 基于流感知的复杂网络应用识别模型[J]. 通信学报，2015，36(3)：192-200.

[6] 杜大跃. 基于深度包数据检测技术的应用特征识别系统的设计与实现[D]. 北京：北京交通大学，2012.

[7] 陈琳，孔华锋，沈开心. P2P应用多层次识别方法研究[J]. 华中科技大学学报(自然科学版)，2014，42(11)：117-120.

[8] 陈佳. 应用层协议快速识别的研究与实现[D]. 北京：北京邮电大学，2010.

[9] 郑伟. 基于防火墙的网络安全技术的研究[D]. 长春：吉林大学，2012.

[10] 范慧萍，宣蕾，陈曙晖，等. 基于正则表达式的应用层协议识别加速[J]. 计算机研究与发展，2008(增刊1)：438-443.

[11] 周亚建，薛超，平源. 基于端口特征的P2P应用识别方案[J]. 北京工业大学学报，2013，39(11)：1667-1672.

[12] 赵雷. 基于DNS的恶意域名识别系统的设计与开发[D]. 济南：山东大学，2013.

[13] 李梅，杨传斌. OICQ的攻击与预防[J]. 微机发展，2002(2)：29-31.

[14] LEE S H, SEOK S J, KANG C G, et al. The two markers system for TCP and UDP flows in a differentiated services network [J]. Computer Communications, 2003, 26(4): 338-350.

[15] 方爽. 基于特征匹配的WEB应用防火墙的研究与实现[D]. 合肥：安徽大学，2014.

本文编辑：陈小平