基于 DPI 技术的 IM 协议识别系统研究

王 凯,吴君钦
(江西理工大学 信息工程学院,江西 赣州 341000)

摘 要:针对不法分子利用 IM 协议通信软件泄露国家和企业机密以及传播反动言论的问题,文中在深入研究和分析多种即时通信软件的 IM 协议的基础上,总结以往 IM 协议识别系统的缺陷,配合 DPI 技术的应用设计了一个全新的 IM 协议检测系统,即基于 DPI 技术的 IM 协议识别系统,该系统能够有效地对多种即时通信软件进行识别和监控。通过实验对多种即时通信软件如 QQ,fetion,MSN,新浪微博桌面版,googletalk,yahoomsg 等的文本信息进行实时监控,验证了该系统对 IM 协议识别具备极高的识别率以及优越的稳定性。

关键词:即时通信;深度包检测;协议分析;识别系统
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2013)11-0120-04
doi:10.3969/j.issn.1673-629X.2013.11.030

Research on IM Protocol Recognition System Based on DPI Technology

WANG Kai, WU Jun-qin
(School of Information Engineering, Jiangxi University of Technology, Ganzhou 341000, China)

Abstract: Aiming at lawbreakers take advantage of the IM protocol communication software to divulge state and corporate secrets, as well as dissemination of reactionary remarks, in order to solve this problem, on the basis of studying and analyzing in-depth the IM protocol of a variety of instant messaging software, summarizing the previous IM protocol identification system's defects, combined with the application of DPI technology a novel IM protocol detection system is designed, which is the IM protocol recognition system based on DPI technology. This system is capable of effective identification and monitoring of a variety of real-time communication software. Through the real-time monitoring experiments on a variety of instant messaging software such as QQ, fetion, MSN, the Sina microblogging desktop edition, googletalk, yahoomsg etc, verify the system with superior recognition rate and perfect stability.

Key words: IM; DPI; protocol analysis; recognition system

0 引 言

近几年来随着科技的不断发展,即时通信(Instant Messenger,IM)已经发展成为最常使用的网络应用之一,我国流行的 IM 软件繁多,最受百姓追捧的有:腾讯公司的 QQ,中国移动的 fetion 以及新浪微博等,国外的有 MSN,ICQ 等即时通信软件。这些主流的即时通信软件拥有绝大部分的用户群,它们功能强大,服务完善,人们的生活越来越离不开它们[1]。然而,即时通信产品在企业内部的广泛使用使生产效率减低以及增加了公司内部重要资料泄密的风险,同时也方便了不良言论和反动内容的传播,因此对 IM 进行识别和检测是网络管理面临的迫切问题。回顾以往的识别研究[2-4],总结其不足主要体现在:(1)只能对少数知名的 IM 如 MSN、ICQ 进行监控;(2)难以对检测系统进行升级和加入新的 IM 监控。

文中提出的基于 DPI 技术的 IM 识别系统不仅能够对 QQ、fetion 等即时通信软件进行有效识别,而且系统中的升级规则模块可以及时地更新协议规则对新加入的 IM 协议进行监控。

1 IM 协议分析

即时通信软件种类繁多,国内外都没有统一的标准去定义所有在网络传输中的 IM 协议,公司为了其自身的发展利益都对 IM 软件定义一套独立的 IM 协

收稿日期:2013-01-26 **修回日期**:2013-05-09 **网络出版时间**:2013-08-28
基金项目:国家自然科学基金资助项目(11062002)
作者简介:王 凯(1988-),男,硕士研究生,研究方向为信息安全、嵌入式系统;吴君钦,副教授,研究方向为信号处理、嵌入式系统。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20130828.0759.006.html

议,这些协议都是不被公开的,所以有必要对每种 IM 协议进行分析。由于 fetion、ICQ、MSN、QQ 等众多 IM 应用层协议组织的数据在网络上采用明文传输(QQ 文本信息采用加密传输,文件、语音和视频采用明文传输),采用对比分析的方法,寻找各种协议所具备的不同特征(有规律的字段以及固定的字节),就能有效地识别出具体协议。表 1 整理出 4 种常用的基于 IM 协议应用软件所具备的特征。通过提取协议特征信息与系统的特征库比较可最大程度地提高协议识别率。

表 1　常用 IM 通信应用的特征分析

协议名称	网络传输协议	端口号	文本传输形式	特征串
QQ_IM	UDP	8000	密文	偏移量尾部 key0=03 偏移量首部 key3=00 偏移量首部 key0=02
MSN_IM	TCP	1863	明文	MSG、UBX、NLN
Fetion_IM	TCP	8080/80	明文	fetion. com. cn
Microblog_Sina	TCP	443	明文	ww4. sinaimg. cn

国内外许多即时通信软件大多工作在 TCP/IP 层,为了绕过防火墙它们通常会在 TCP/IP 协议层之上添加一些其他协议,如 FTP、HTTP 等,再构建 IM 协议层[5]。然而 FTP、HTTP 协议毕竟只是占据少数网络流量,且 IM 协议不变,基于这个特点就可有效地提取出各个即时通讯软件的 IM 协议所具备的特征规则,从而在数据流绕过防火墙后系统依然可有效的识别。

2　DPI 技术

DPI(Deep Packet Inspection),中文译名为"深度包检测",和普通的报文仅仅分析 IP 包的 4 层内容,包括源地址、目的地址、源端口、目的端口以及协议类型相比,深度包检测还增加了应用层分析,可以识别各种应用以及其内容,因此基于 DPI 技术的 IM 协议识别系统可以有效地防止不良言论和反动内容的传播,同时也能减低公司机密文档泄密的风险。DPI 识别技术可以分为以下三种。

2.1　端口检测技术

所谓的端口检测技术就是从提取到的数据包中读取应用层端口,通过端口数值去判断应用协议,端口检测技术是一种最为简单的识别技术,随着 IM 技术的不断更新发展,很多 IM 协议都能够动态地变更端口例如 ICQ,端口识别技术逐渐失效。

2.2　基于"特征字"的识别技术

在网络中传输的每一条数据流都会有其独一的特征,这些特征被称之为协议的规则(有些文献称之为"指纹")[6-10]。根据数据流中的这些规则就能够确定协议的种类。在系统的设计中,数据库的更新就是指当有新的特征出现时,将其填充进升级规则库模块中。图 1 中采用的是 wireshark 软件对 fetion 进行数据报文的捕获,从图中可以显示出 fetion 在应用层中的一个规则 fetion. com. cn,也是通过这个规则去确定 fetion_IM 协议。

基于"特征字"识别的关键技术是指从形式各样的数据包中判断出协议的规则,并建立规则库,从而把捕获到的报文送到规则库,判断是否匹配。若匹配,即可确定协议所承载的应用[11]。

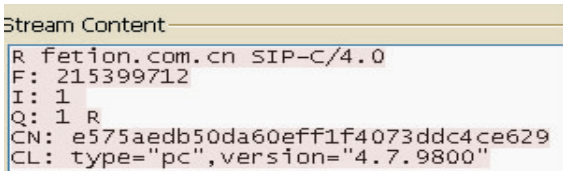

图 1　fetion 数据报文特征

2.3　应用层网关识别技术

某些应用层业务的控制流和业务流是分离的,在业务流没有任何特征[12]。这种情况下,就要使用应用层网关识别技术来识别。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业务流[13]。

3　系统设计

3.1　系统结构

由于系统是用于 IM 协议识别,因此系统的核心部分是识别模块。另外由于 DPI 技术需要不断更新特征库,所以设计一个检测更新的模块是必不可少的[14]。其框架图如图 2 所示。

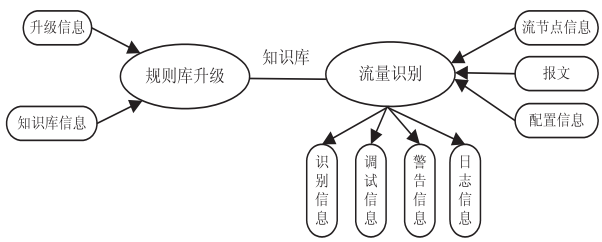

图 2　系统框架

由图可以很清晰地了解到系统的识别流程:首先,在规则库升级模块中配置升级信息并植入知识库信息,运行规则升级模块在服务器中升级规则库,规则升级模块为流量识别模块提供知识库信息。之后把捕获到的数据信息和配置信息送入流量识别模块,并在识别模块中输出结果信息,其中包括有日志、警告、调试和识别信息,若输入的报文信息与知识库定义的规则不匹配,则把日志信息反馈给规则库升级模块。

3.2　DPI 识别原型系统设计

图 3 所示是系统的具体流程,在图中可以看出,系统从数据流中提取出数据包,通过预处理获取纯净流,并判断是否为新的节点,是否需要创建一个流节点,系

统提供一个储存空间去存放流节点。若是与流表相同并且需要对流节点继续识别则将报文传输给识别引擎进行识别,若是成功那么确认该协议释放空间;若是新创建的节点需要识别,送到识别引擎中,若识别成功确认协议并释放空间,要是不成功则识别为 UNKNOW 并且释放空间同时告诉规则库,需要更新规则。

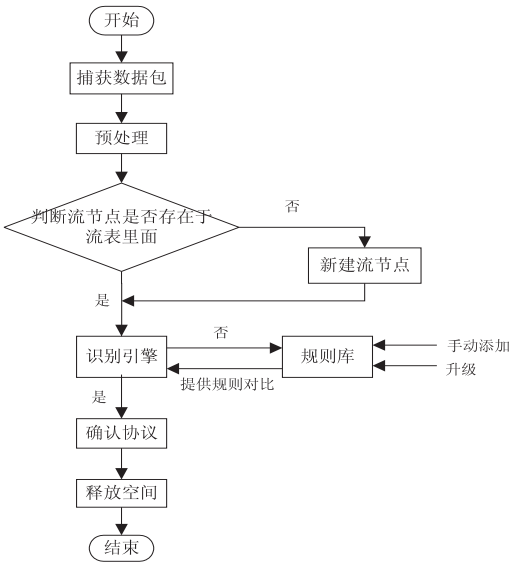

图 3 识别流程图

识别引擎是整个系统中最核心的部分,该部分实现程度的好坏决定了协议识别率的高低。其算法描述如下:

```
PACKET packet //捕获的流
Define PACKET_CONNECT( ) //关联识别函数
Define PACKET_FEATURE( ) //特征字识别函数
Define PACKET_PORT( ) //端口识别函数
Define PACKET_STATISTICS( ) //统计识别函数
for( )
{
Wait for a packet;
switch(packet)
{
case (PACKET_CONNECT(packet))= =1:
CONFIRM_RELEASE(Protocol,&Space);
//确认协议并释放空间
break;
case (PACKET_FEATURE(packet))= =1:
CONFIRM_RELEASE(Protocol,&Space);
//确认协议并释放空间
break;
case (PACKET_PORT(packet))= =1:
CONFIRM_RELEASE(Protocol,&Space);
//确认协议并释放空间;
break;
case (PACKET_STATISTICS(packet))= =1:
CONFIRM_RELEASE(Protocol,&Space);
//确认协议并释放空间
break;
default:
packet = UNKNOW;
//不能识别,返回 unknown
break;
}
}
```

3.2.1 关联识别函数 PACKET_CONNECT( )

报文先会通过关联识别,关联识别只需遍历流表可筛选出已识别的协议,这样就可以降低规则匹配所消耗的系统时间及资源。若是关联识别成功,则确认协议;否则,把节点信息传下一层进行特征串识别。

3.2.2 特征字识别函数 PACKET_FEATURE( )

特征字识别是识别引擎中最重要的识别方式,它通过从更新规则库模块传输过来的知识库规则与流节点进行规则匹配。若匹配成功,则输出协议。其中特征串匹配模式分为单包匹配和多包匹配,所谓单包匹配是指通过规则库一条已知报文对流节点进行预测判断。而多包匹配是指对同一种应用的多个报文带有多个报文规则,按照顺序依次对比匹配,建立特征索引。若特征字识别失败,则将报文传到下一步的端口识别中进行识别。

3.2.3 端口识别函数 PACKET_PORT( )

端口识别中,识别引擎通过一些特定的协议标准端口或固定端口值进行匹配。若匹配成功,则确认协议;若匹配不成功,则传到下一步进行统计识别。

3.2.4 统计识别函数 PACKET_STATISTICS( )

统计识别主要针对没有明显特征字的加密协议,通过分析其流的统计特征进行识别,统计识别模式最多对 32 个报文进行统计识别,分两个阶段进行。第一阶段是入口条件判断,对于流中的第 1 个至第 32 个报文,每一个都进行入口条件的判断,入口条件包括:命中关键字、命中关键表、命中规则、命中 IP 关联表、命中包长、命中端口。第二阶段为统计分析,统计识别可以进行如下统计分析:同向固定包长序列、双向固定包长序列、同向固定包长集合、双向固定包长集合、同向连续包长范围、双向连续包长范围、同向连续包长平均值、双向连续包长平均值、同向连续包长求和、双向连续包长求和、指定包长复现、同向连续包数统计、同向包数统计。若匹配成功,则确认识别协议;若不能够识别,则定义该协议为 unknown,并把信息反馈给升级规则库模块。

4 实验结果

为了对系统进一步的分析,实验通过对 QQ,fe-

tion,MSN,Sinaweibo 等一些占据国内大部分市场的 IM 应用软件在现有规则库下进行文本协议识别分析。

实验配置:主机 CPU 频率 2.69 GHz;1.99 GB 内存;操作系统是 Windows XP 系统;网络配置器是 Realtek RTL8169/8110 Family Gigabit Ethernet NIC。

实验步骤:

第一步,抓包软件 wireshark 的配置,因为所选软件是 Windows 应用所以不选择捕抓杂项,再匹配好本机的 IP;

第二步,采用 wireshark 抓包软件对 fetion、MSN 和 Sinaweibo 等进行即时文字聊天场景报文的获取,每个软件最少捕获 600 个数据流,遵循抓包行为规范化;

第三步,捕获到的报文信息送入到识别系统中,然后进行系统配置;

第四步,开始识别。

实验结果分析:实验中把捕获到国内外常用的即时通讯软件的数据报文信息送入识别系统,表 2 显示了具体细节。

表 2　协议识别率情况表

应用软件	软件版本	识别协议	识别率/%
MSN	4.7.3001	MSN_IM	99.17
Gadu-gadu	2.10.5	Gadu_Gadu	98.65
Google-talk	1.0.0.1	Googletalk_IM	98.59
YahooMsg	11.5.0.228-us	YahooMsg_IM	97.34
ICQ	8.0.598	ICQ_IM	96.59
飞信 2012	4.7.9800	Fetion_IM	96.45
QQ	2012 正式版(5119)	QQ_IM	96.35
ICALL	7.1.524	Icall_IM	89.28
微博桌面 2012	Build2.0.10.30196	Microblog_Sina	88.22

从表中的实验结果可以看出多部分协议识别率都占据 96% 以上,基于 DPI 技术的识别系统设计具有高识别率和高精确度的特点,图 4 所示是识别实验结果。

```
protoId =   80  protoname =       ICQ_IM  hitsize =    4495.00  hitratio =96.59%
protoId =   82  protoname =        QQ_IM  hitsize = 82207240.00  hitratio =96.35%
protoId =   83  protoname =       MSN_IM  hitsize =   26113.00  hitratio =100.00%
protoId =   85  protoname = YahooMsg_IM  hitsize =  631481.00  hitratio =97.34%
protoId =   86  protoname = GoogleTalk_IM  hitsize =  180094.00  hitratio =98.59%
protoId =   90  protoname =    Fetion_IM  hitsize =  178977.00  hitratio =96.45%
protoId =  313  protoname =    Gadu_Gadu  hitsize = 4203942.00  hitratio =98.65%
protoId =  949  protoname = Microblog_Sina  hitsize =  647360.00  hitratio =88.22%
protoId = 1310  protoname =     iCall_IM  hitsize =  320191.00  hitratio =89.28%
```

图 4　多协议识别结果

综上基于 DPI 的 IM 协议识别系统是稳定的并且具备高识别率,检测系统能够及时地针对未知协议进行升级,实现新加入 IM 协议的有效监控。该系统还可以识别出多种明文通信软件的内容如 fetion,可以有效地防止不良言论和反动内容的传播,同时也减低了公司机密文档泄密的风险,这满足了设计的目的。

5　结束语

在深入学习和分析 IM 协议,并且探讨了一些 IM 协议检测系统的基础上,结合 DPI 技术提出了一种基于 DPI 技术的 IM 协议识别系统,系统通过对现有国内外最为常用的即时通信应用软件进行实验验证分析,可以看出该系统具备高识别率和高精确度的特点。而且也能够通过更新协议规则库实现了对新加入的 IM 协议进行监控,随着 IM 技术的不断更新与发展 DPI 识别技术将会得到更为重要的应用。

参考文献:

[1] 杨　阳. 即时通讯流量识别还原技术研究[D]. 长沙:湖南大学,2008.

[2] 李　鑫. 基于 DPI 的网络流量识别系统的设计与实现[D]. 武汉:武汉理工大学,2010.

[3] 付安民,张玉清. 即时通实时监控系统的设计与实现[J]. 通信学报,2008,29(10):165-172.

[4] 李远杰,刘渭锋,张玉清,等. 主流即时通软件通信协议分析[J]. 计算机应用研究,2005(7):243-245.

[5] 胡振宇,刘在强,苏璞睿,等. 基于协议分析的 IM 阻断策略及算法分析[J]. 电子学报,2005,33(10):1830-1834.

[6] Xiao Zhen, Guo Lei, Tracey J. Understanding instant messaging traffic characteristics[C]//Proc of 27th international conference on distributed computing systems. Toronto:[s. n.],2007.

[7] Kawano S, Okugawa T, Yamamoto T, et al. High-speed DPI method using multi-stage packet flow analysis[C]//Proc of 2012 9th Asia-Pacific symposium on information and telecommunication technologies. Santiago and Valparaiso:[s. n.], 2012.

[8] Antonello R, Fernandes S, Sadok D, et al. Characterizing signature sets for testing DPI systems[C]//Proc of GLOBECOM Workshops. Houston, TX:IEEE,2011.

[9] Shen Zihao, Wang Hui. Network data packet capture and protocol analysis on Jpcap-based[C]//Proc of 2009 international conference on information management, innovation management and industrial engineering. Xi'an, China:[s. n.],2009.

[10] Yang Chu-Sing, Liao Ming-Yi, Luo Mon-Yen, et al. A network management system based on DPI[C]//Proc of 13th international conference on network-based information systems. Takayama, Japan:[s. n.],2010.

[11] 黄晓武. 基于 DPI 技术的网络流控策略[J]. 电脑知识与技术,2011,7(6):1260-1261.

[12] 陈朝晖. 一种基于 DPI 和 DFI 技术的应用识别系统[J]. 中国高新技术企业,2011(16):77-80.

[13] 华为数通. DPI 深度包检测技术及其作用[EB/OL]. 2007-04. http://www.huawei.com/.

[14] 金　婷,王　攀,张顺颐,等. 基于 DPI 和会话关联技术的 QQ 语音业务识别模型和算法[J]. 重庆邮电学院学报(自然科学版),2006,18(6):789-792.

基于DPI技术的IM协议识别系统研究

作者： 王凯， 吴君钦， WANG Kai， WU Jun-qin
作者单位： 江西理工大学 信息工程学院,江西 赣州,341000

刊名： 计算机技术与发展

ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2013(11)