

# 一种基于流特性描述的 P2P 流量模糊识别方法

孙知信<sup>1),3)</sup> 官 婧<sup>2)</sup>

<sup>1)</sup>(南京邮电大学计算机技术研究所 南京 210003)

<sup>2)</sup>(南京邮电大学数理学院 南京 210003)

<sup>3)</sup>(南京大学计算机软件新技术国家重点实验室 南京 210093)

**摘 要** 为了解决无法单纯地依靠关键字、端口+IP 的方式识别某些特征不明显的网络应用,该文首先将模糊数学的理论应用在 P2P 流量的识别中,提出了一种基于流特性描述的模糊识别方法 FCD. FCD 模糊识别方法首先对网络数据包进行捕获,然后根据数据包的特点进行分析,对其中关键性的流量进行统一描述,接着重点分析这些关键性流量中数据包的分布情况,得到它们的隶属度函数,作为评判时的评语集,最后用模糊评判方法判定它是否属于某种网络应用.文中以著名的网络游戏魔兽世界为例进行实验,实验结果说明,用 FCD 模糊识别方法可以识别出该种网络游戏,而且准确率较高.

**关键词** 对等网络;数据流;模糊评判;流特性;网络游戏

中图法分类号 TP309

## A Kind of P2P Fuzzy Recognition Method Based on Flow Characteristic Description

SUN Zhi-Xin<sup>1),3)</sup> GONG Jing<sup>2)</sup>

<sup>1)</sup>(College of Computer of Nanjing University of Posts and Telecommunications, Nanjing 210003)

<sup>2)</sup>(College of Applied Mathematics & Physics, Nanjing University of Posts and Telecommunications, Nanjing 210003)

<sup>3)</sup>(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093)

**Abstract** Some characteristics of network application flow are inconspicuous, and they can not be simply identified by keywords, ports+IP. The paper applies fuzzy mathematics theory to P2P traffic identification, and presents a kind of Fuzzy Recognition Method based on Flow Characteristic Description (FCD). FCD fuzzy recognition method analyzes captured network data first, and then describes uniformly key processes. Subsequently, the paper analyzes distributional state of those data in key process, and obtains their membership functions, which can be used as comment aggregates. Finally, fuzzy judgment method will judge whether the data are some kind of network application. The proposed recognition method is put to test through a popular network game. The experiment results verify that FCD fuzzy recognition method can identify network game flow and its accuracy is high.

**Keywords** peer to peer; data flow; fuzzy judgment; flow characteristic; network game

## 1 引 言

近几年来 P2P(peer to peer)作为一项全新的

Internet 技术飞速发展,其业务已悄然占据了互联网业务总量的 60%~80%,成为杀手级宽带互联网应用. P2P 业务应用围绕 IP 音频和视频文件共享快速发展,造成了网络带宽的巨大消耗,甚至引起网络

拥塞,降低其它业务的性能,失去平衡的网络进出口带宽资源分配,给企业网络管理和接入运营商带来极大的麻烦.另一方面,基于 P2P 的网络游戏,特别是一些大型网络游戏,越来越受到网络玩家的青睐.例如,魔兽世界、梦幻西游等,都在原先的 C/S 模式中加入了 P2P 的特点,这样一来,如果要识别这些大型的网络游戏流量,通过单一的关键字或者端口等方式来识别是行不通的.如果无法识别,就更加谈不上对其进行分类、标识和控制,所以如何识别是一个亟待解决的问题.

针对 P2P 流的各种特点,对于有固定服务端口的 P2P 应用的识别,可以采用基于检测默认的服务端口的 P2P 流识别技术<sup>[1-3]</sup>,然而如今很多 P2P 应用采用故意伪装端口技术<sup>[4]</sup>,这使得大约 30%~70% 的 P2P 流量无法使用已知的标准端口来进行识别.为此,一种基于 P2P 网络的连接模式在传输层识别 P2P 流的方式<sup>[5]</sup>被提出,它通过监控 {IP, port} 的连接模式,监控 TCP/UDP 端口.但是,由于 P2P 软件引入动态端口,通过分析 IP 和端口之间与 P2P 应用的关系来识别的方法,越来越表现出检测准确性不高的缺陷.为了提高检测准确性,有些学者提出了研究 P2P 协议的一部分已知的特征<sup>[6-7]</sup>,采用基于 P2P 流有效载荷进行识别的方法<sup>[8-9]</sup>. Karagiannis 等<sup>[10]</sup>提出从包格式和某些长度固定字节两方面来识别 P2P 流量的包签名技术,Sen 等<sup>[8]</sup>、Gummadi 等<sup>[11]</sup>提出了基于应用层的签名机制来检测 P2P 流的方式.然而这类技术往往需要有相关的研究机构提供所需检测的 P2P 流的应用特征,但是随着应用版本的更新,某些特征也将随之改变,文献 [12-14] 中介绍的几个最为典型的数据实例正是说明了这个现象.同时,这类技术要求在检测设备中数据包需要实时地进行重组,做到这一点是很困难的.还有些学者提出了基于网络动态性来刻画 P2P 流特性<sup>[15]</sup>,例如在 IP 层通过统计流量特征的方式<sup>[16]</sup>以及根据 P2P 网络的层次拓扑特性<sup>[17]</sup>来识别 P2P 流的方式,但是这类研究网络动态特性的方法对于那些特征不明确,或是协议不公开的 P2P 应用,则无法进行分析和识别.

为了解决无法单纯地依靠关键字、端口+IP 的方式识别某些特征不明显的网络应用,本文将模糊数学的理论运用到数据流的识别中,提出了一种基于流特征描述的模糊识别方法 FCD(Fuzzy Recognition Method based on Flow Characteristic Description).

本文第 2 节介绍模糊评判规则的理论依据和设计思想;第 3 节介绍 FCD 模糊识别方法的具体实现;第 4 节描述 FCD 模糊识别方法在识别网络游戏中的应用,并对实验数据进行分析 and 讨论;最后是总结.

## 2 模糊评判规则

模糊集合用来描述一个模糊概念,它是内涵和外延都不明确的集合.本文利用模糊集合的概念,对于网络中的特征不明确、无法单纯地依据某一种因素进行识别的数据流量,由于该种数据流往往受到多种因素共同影响,因此本文首先对各种因素进行规范化描述,接着建立它们的特征隶属度函数,由模糊类集合中产生评语集,随后通过模糊评判,得到评语的可信程度,从而判断该数据流是哪种网络应用.

### 2.1 数据包集合的描述

**定义 1.** 登录过程的描述.

网络游戏登录过程中,一般都要经过一个与服务器交互的过程,在这个过程中,服务器的 IP 和端口,将可以作为识别何种网络应用的依据之一,同时在登录过程中,还会表现出一些特征,因此,对登录过程加以描述  $D(SI, SP, LU, LP, LK)$ , 其中:

$SI$ (Server IP)表示该种网络游戏所提供的服务器 IP,这里 IP 往往不是固定的一个,而是有多个;

$SP$ (Server Port)表示该种网络游戏所提供的服务器交互端口,某些网络游戏是有固定端口的,但也有些端口可变;

$LU$ (Login Update)表示网络游戏的登录过程描述,在用户进行登录的过程中,往往会出现版本更新的特征,例如,一些网络游戏,会在用户登录的时候,进行监测,或者是提示用户进行更新,或者采用强制更新的方式.而更新的方式,一般多为 P2P 的方式进行,这时就可以依据它更新的特征来判断出它是某种 P2P 游戏业务;

$LP$ (Login Password)表示游戏登录过程中,获取到客户端向登录服务器发送用户名及密码等信息;

$LK$ (Login Keyword)表示游戏的登录过程中所出现的关键字模式串及其特征,例如,登录中会发现有几个包主要是完成客户端与服务器交换数据以及进行账号认证的过程,由于数据段已经过加密,对于每个包的具体细节还不清楚,但是可以根据它们的特征进行模糊评判.

登录过程中,服务器的信息是比较好获得的,但

是它并不能作为确定某类网络应用的最终判断依据,必须继续对游戏的交互协议进行分析.因此,本文中对协议交互过程中的数据包进行规范化描述,可以进一步对某类网络应用的特征进行细致描述.

**定义 2.** 协议交互过程的描述.

当某种网络应用处于协议交互状态时,数据包的内容已经过加密,因此重点对协议中的关键特征(例如,网络游戏中的在离线、杀伤情况、地图变化等)动态的数据进行描述  $D(AC, AP)$ , 其中:

$AC$ (Alternation Characteristic)表示协议交互过程中出现特性数据包的描述,即虽然数据包被加密了,但仍然可以发现一些特征,例如:在游戏进行过程中,发现某组数据包中有 3 个数据包的内容完全一致这样的特征字符串;

$AP$ (Alternation Process)表示协议交互过程中出现的特殊过程,例如,游戏退出时,客户端会给服务器端发送一些数据包,以表明中断与它的连接.

**定义 3.** 在线更新过程的描述.

网络应用的版本是在不断更新的,例如游戏的玩家通常都希望用的是最新版本的游戏,同时有些游戏也是强制性更新的,而整个更新的过程,采用 P2P 方式为主,这就为识别网络数据包提供了另一个有效的方式.因此,通过分析,本文将在线更新过程描述为  $D(MU, AU)$ , 其中:

$MU$ (Manually Update)表示采用手动下载升级补丁(可执行文件),保存到游戏安装目录下,运行升级补丁即可完成升级;

$AU$ (Active Update)表示在登录的时候进行检查若有更新的版本则主动的进行在线升级.

根据定义 1、定义 2 和定义 3 的描述,本文对网络应用的描述主要分为:登录类描述、协议交互类描述、在线更新过程类等几类集合.为了实现识别网络应用数据流的目的,根据模糊数学的理论,就需要对各类集合建立相应的隶属函数,根据模糊集合间的关系以及所占的权重,才能进行计算得到评判结论.由于网络中的数据其实是一些随机现象,因此对隶属函数进行选取时,本文是根据数据某项特征的密度分布函数来确定其隶属度的函数,从而得到相应的评语集.下面就对各类集合定义其相应的隶属函数.

## 2.2 隶属度函数的定义

对于一个集合  $A$  可用隶属度函数  $\mu$  来表示,  $\mu$  的定义域是与该模糊概念相关的对象的集合,称为  $A$  的论域  $U$ ,而  $\mu$  表示了  $U$  中每个对象隶属于  $A$  的程度,称为隶属度,  $\mu \in [0, 1]$ . 模糊集合是模糊数学

的基本概念,通过它可以建立与普通集合的关系,继而用经典数学的方法来研究模糊现象.

定义 1~3 中对网络游戏从登录到退出时的特征都进行了描述,它们是进行评判时候的重要依据.为了将这种描述性的语义由模糊集合来表达,因此,本文中先将其数值化,之后再定义相应的隶属函数.

在确定隶属度函数时,本文采用了模糊统计的方法,通过一段时间的观察,得到函数的图形.

**定义 4.** 登录类的隶属度函数:

$$\mu(l) = \begin{cases} 1, & l \leq a \\ e^{-k(l-a)^2}, & l > a, k > 0 \end{cases} \quad (1)$$

式(1)中的  $f$  为捕获数据包的频率,  $l$  的含义是捕获到的登录数据包的个数占总体数据包个数的比值,这样,  $\mu(l)$  表示将登录类集合归为登录类模糊集.

例如,对网络游戏魔兽世界的登录过程进行分析,分析所捕获的登录包占总体数据包的比例随着时间的变化情况,可以刻画出登录过程的隶属度函数.  $a$  是表示在登录过程基本稳定之后,包含有登录特征的数据包数目在总数据包中的比例开始以半正态倍的减少.

**定义 5.** 交互类的隶属度

$$\mu(a) = \begin{cases} 0, & a < \bar{f} - s \\ (s + a - \bar{f}) / (s - k), & \bar{f} - s < a \leq \bar{f} - k \\ 1, & \bar{f} - k < a \leq \bar{f} + s \\ (s - a + \bar{f}) / (s - k), & \bar{f} + s < a \leq \bar{f} + k \\ 0, & a \geq \bar{f} + k \end{cases} \quad (2)$$

式(1)中,  $a$  表示在协议交互过程中,符合定义 2 描述的数据包占总体数据包的比例,这样,  $\mu(a)$  表示将交互类集合归为交互类模糊集,其中,  $f$  是协议交互时,捕获数据包的频率,

$$\bar{f} = \frac{\sum_{i=1}^n f_i}{n}, s = \sqrt{\frac{\sum_{i=1}^n (f_i - \bar{f})^2}{n-1}}.$$

**定义 6.** 在线更新类隶属度:

$$\mu(o) = \begin{cases} 1, & o \leq n \\ e^{-m(o-n)}, & o > n, m > 0 \end{cases} \quad (3)$$

$\mu(o)$  表示将  $O$  类集合归为  $O$  类模糊集.例如,对魔兽世界的在线更新过程进行分析过程中,  $n$  点表示从该处在线更新特征的数据包逐渐稳定,开始以指数倍减少.

随着各种网络应用的日益增长,为了能够更好地对网络流量进行管理,需要能够有效地识别出各

种不同的网络应用. 由于网络应用的种类很多, 而特征不固定, 影响识别的因素也不唯一, 所以想要有效地识别出, 有一定的困难. 经过大量实验数据的分析, 我们发现, 很多网络应用都具有一系列的特有过程, 例如, 网络游戏在登录、交互以及在线更新时, 有明显的特征, 综合考虑这些特征, 可以判断网络中的数据流是否属于该种网络应用. 因此, 本文以大型网络游戏为例, 对登录类、交互类以及在线更新类函数进行定义, 之后根据所捕获数据包表现出的特征, 对它们的隶属度分别进行定义, 以便实现综合评判. 但是, 由于各种因素的影响作用不同, 因而不能简单地按照传统的评判方法进行评判, 所以本文提出一种基于流特征描述的模糊识别方法, 下面将具体阐述该识别方法的识别过程和评判规则.

3 基于特征描述的模糊识别方法 FCD

在实际应用中, 对某事物的评价常常受到多种因素影响, 其中模糊性是最主要的. 为了对各种影响进行合理的评判, 将模糊技术同传统的综合评判方法相结合, 将可以达到较为公平合理的评判结果.

采用模糊评判方法进行评判的方法是: 首先建立评判对象的因素集  $F = \{f_1, f_2, \dots, f_n\}$ , 接着建立合理的评语集合  $C = \{c_1, c_2, \dots, c_m\}$ , 生成评判矩阵  $\tilde{R} = (r_{ij})_{n \times m}$ , 并最终进行综合评判.

在本文中, FCD 模糊识别方法借鉴了模糊数学中的概念, 为了更加准确和公正地识别出网络中的不同应用流量, 本节制定了评价规则.

为了断定某个数据流属于哪种网络应用, 文中根据模糊关系的理论, 由模糊关系系数定义评价向量, 根据该评价向量, 进行综合评判. 所谓模糊关系系数是关系的推广, 在模糊关系中的事物间的关系不是仅用“有”或“无”来描述, 而是隶属度来表述模糊类之间的关联程度. 在本文中, 用模糊评判规则将比较复杂的事件或对事物的整体评估分成许多比较简单的小部分来分别评估, 最终得出对整个事件或事物的结论. 下面给出了进行模糊评判时的评判值的定义.

例如, 在对某种网络应用的评判中, 有很多因素影响评判的结果, 假设为因素 1、因素 2、因素 3 和因素 4.

下面说明模糊综合评判的过程.

1. 对网络中的数据流量进行检测, 特别关注其中的特

征性环节. 例如对网络游戏的过程进行捕获, 包括登录认证过程, 进入游戏过程, 游戏进行过程, 以及退出游戏过程.

2. 将上述捕获到的数据经过预处理设备进行预处理, 再规范化统一描述成  $D(SI, SP, LU, LP, LK), D(AC, AP), D(MU, AU)$  的格式.

3. 确定评判时的因素集:  $F = \{\text{因素 } 1(f_1), \text{因素 } 2(f_2), \text{因素 } 3(f_3), \text{因素 } 4(f_4), \text{因素 } 5(f_5)\}$ .

4. 根据式(1)~(4), 计算隶属度函数值, 从而得到模糊评判时的评语集.

5. 采用模糊评判的方法进行计算, 得到模糊综合评判结果.

6. 考虑到不同的因素对评判结果的侧重点不同, 因此对每个因素还有一个相应的权值.

7. 综合权值后再进行评判, 得到模糊评判的综合结果;

8. 根据评判结果判定所识别的流量是否是某种网络游戏的流量.

4 FCD 模糊识别方法在识别网络游戏中的应用和分析

4.1 用 FCD 模式识别方法识别魔兽世界

“魔兽世界”由世界著名游戏公司暴雪出品, 游戏为不同的角色定制了几千种武器、魔法和技能, 让玩家有机会体验到完全不同角度的“魔兽世界”. 全真实的 3D 构筑以及世界顶级美工大师们的精益求精更让整个游戏中的影音效果几乎发挥到了极限, 甚至游戏内的每一棵树都是由暴雪的美工手绘, 给玩家带来全方面的视觉震撼! 因此, 无论对网游高手还是新进玩家来说, “魔兽世界”都具有同样巨大的吸引力. 正是因为该游戏复杂的设计和游戏过程, 单纯的分析某一个数据包或端口, 是不可能确定所捕获的数据流的性质的, 对于此类问题也不能以是或者不是来简单地加以判断, 因此, 本文提出了对游戏中的特征数据包进行综合描述, 之后采用模糊判断的方法来进行判定, 具有较高的可信度.

本文以魔兽世界为例, 对其整个游戏过程进行捕获、分析、识别, 对数据包捕获以及简单的分析, 借助 Win Ethereal 软件完成.

4.1.1 登录过程的描述

(1) 启动魔兽客户端后, 首先出来一个选择登录界面, 有 6 个区选择:

- 一区(上海), cn1. grunt. wowchina. com
- 二区(北京), cn2. grunt. wowchina. com
- 三区(四川), cn3. grunt. wowchina. com
- 四区(广东), cn4. grunt. wowchina. com

五区(上海),cn5.grunt.wowchina.com  
六区(北京),cn6.grunt.wowchina.com  
这就是选择“登录/账号服务器”(Login Server). 这里抓包分析选择的是四区(广东),cn4.

85	40.837639	10.10.80.163	219.133.56.109	TCP	1323	>	3724	[ACK]	Seq=1	Ack=1	win=17520	Len=0
86	40.911803	10.10.80.163	219.133.56.109	TCP	1323	>	3724	[PSH, ACK]	Seq=1	Ack=1	win=17520	Len=41
87	40.938031	219.133.56.109	10.10.80.163	TCP	3724	>	1323	[ACK]	Seq=1	Ack=42	win=5840	Len=0

图 1 登录过程数据包 1

(2)由 Client 向登录/账号服务器(Login Server)发送用户名及密码等信息.

根据所捕获到的数据包,对登录过程加以描述  $D(SI,SP,LU,LP,LK)$ :

$SI$ (Server IP)表示该种网络游戏的所提供的服务器 IP,这里服务器的 IP 地址 IP 有多个,在本次实验中,服务器 IP 为 219.133.56.109.

$SP$ (Server Port)表示该种网络游戏所提供的服务器交互端口,经过测试得到:魔兽的服务器端口是 3724.

$LU$ (Login Update)表示网络游戏的登录过程描述,在用户进行登录的过程中,往往会出现版本更新的特征,检查登录界面是否有更新.客户端一登录

grunt.wowchina.com 是 login server 的 DNS.  
从所捕获的数据包,如图 1 所示,得知这里的 Login Server 的 IP 是 219.133.56.109.

便检查登录的界面是否有更新,有的话便进行下载并保存到游戏目录下并覆盖原有的登录界面.登录界面主要包括主界面、浮动图标、控件图标、背景图片等.

$LP$ (Login Password)表示游戏登录过程中,获取到客户端向登录服务器发送用户名及密码等信息,这里用户名是 gumbour;

$LK$ (Login Keyword)表示游戏的登录过程中,所出现的关键字模式串及其特征;图 2 中,黑线部分对所有客户端来说都是固定的,这部分数据可作为关键字;而虚线框部分(25,00)部分表示后面数据长度,对于同一客户端是固定的.

0030	44 70 ac 1e 00 00 00 02 25 00 57 6f 57 00 01 0a	Dp..... %.wow...
0040	02 b6 14 36 38 78 00 6e 69 57 00 4e 43 68 7a e0	...68x.n iw.Nchz.
0050	01 00 00 0a 0a 50 a3 07 47 55 4d 42 4f 55 52	.....P.. GUMBOUR

图 2 登录过程数据包 2

4.1.2 游戏交互过程的描述

经多次抓包发现,游戏阶段本机主要和两个服务器端进行交互:61.135.177.59 和 61.139.105.153.

当玩家处于游戏状态时,服务器端(61.135.177.59)不断以 TCP 方式向客户端发送数据包,客户端收到之后返回一个确认信号.数据包的内容已经过加密,可以猜测主要是关于玩家信息(在离线、

杀伤情况等)、地图变化等一些动态的数据.当然,客户端也会给服务器发送数据包,但相对要少很多,猜测主要是把玩家在客户端的一些操作信息反馈给服务器.下面分析游戏过程中两个主要过程的数据包特征:

(1)角色在走动时

取部分数据包如图 3 所示.

72	3.792082	61.135.177.59	10.10.80.99	TCP	3724	>	1169	[PSH, ACK]	Seq=1369	Ack=102	win=4143	Len=186
74	3.931485	10.10.80.99	61.135.177.59	TCP	1169	>	3724	[ACK]	Seq=102	Ack=1555	win=64857	Len=0
75	4.104301	61.135.177.59	10.10.80.99	TCP	3724	>	1169	[PSH, ACK]	Seq=1555	Ack=102	win=4143	Len=37
76	4.233231	10.10.80.99	61.135.177.59	TCP	1169	>	3724	[ACK]	Seq=102	Ack=1592	win=64820	Len=0
82	4.374473	61.135.177.59	10.10.80.99	TCP	3724	>	1169	[PSH, ACK]	Seq=1592	Ack=102	win=4143	Len=185
84	4.534996	10.10.80.99	61.135.177.59	TCP	1169	>	3724	[ACK]	Seq=102	Ack=1777	win=64635	Len=0
85	4.675773	61.135.177.59	10.10.80.99	TCP	3724	>	1169	[PSH, ACK]	Seq=1777	Ack=102	win=4143	Len=74
86	4.836738	10.10.80.99	61.135.177.59	TCP	1169	>	3724	[ACK]	Seq=102	Ack=1851	win=64561	Len=0
87	4.893186	10.10.80.99	61.135.177.59	TCP	1169	>	3724	[PSH, ACK]	Seq=102	Ack=1851	win=64561	Len=34

图 3 游戏交互过程数据包 1

数据方向(包含数据段的包,图中[PSH,ACK]包)主要是:服务器→客户端(61.135.177.59→10.10.80.99),如图中 72,75,82,85 包,这些数据包长度不固定,数据内容也没有规律.客户端收到后返回

一个 ACK 信号.同时,也有少量的数据:客户端→服务器(10.10.80.99→61.135.177.59),如 87 包,这类数据包很少,但是有一定的规律,数据包内容如图 4 所示.

0030	fa a3 2b d1 00 00 3c bb e2 8f 03 2b 01 00 00 00	...+...<. ...+...
0040	a7 b5 20 00 46 e5 3e c5 49 07 fa c3 54 be d2 41	.. .F>. I...T..A
0050	75 2c 93 3f 73 00 00 00	u...?s...

图 4 游戏交互过程数据包 2

AC(Alternation Characteristic)这类数据包数据部分长度固定为 34bytes,且有固定字段,见图 4 中黑线部分,并另有局部固定字段,如图 4 中白线部分,这部分字段在一个区域内(例如提瑞斯法林地:

24	4.189152	61.135.177.59	10.10.80.99	TCP	3724	>	1169	[PSH, ACK]	Seq=305 Ack=0 win=4143 Len=220
25	4.385540	10.10.80.99	61.135.177.59	TCP	1169	>	3724	[ACK]	Seq=0 Ack=525 win=64627 Len=0
26	4.770164	61.135.177.59	10.10.80.99	TCP	3724	>	1169	[PSH, ACK]	Seq=525 Ack=0 win=4143 Len=36
27	4.888462	10.10.80.99	61.135.177.59	TCP	1169	>	3724	[ACK]	Seq=0 Ack=561 win=64591 Len=0
28	5.106481	10.10.80.99	61.135.177.59	TCP	1169	>	3724	[PSH, ACK]	Seq=0 Ack=561 win=64591 Len=20
29	5.246957	61.135.177.59	10.10.80.99	TCP	3724	>	1169	[ACK]	Seq=561 Ack=20 win=4143 Len=0
30	5.316674	61.135.177.59	10.10.80.99	TCP	3724	>	1169	[PSH, ACK]	Seq=561 Ack=20 win=4143 Len=34
31	5.491973	10.10.80.99	61.135.177.59	TCP	1169	>	3724	[ACK]	Seq=20 Ack=595 win=64557 Len=0
32	5.632818	61.135.177.59	10.10.80.99	TCP	3724	>	1169	[PSH, ACK]	Seq=595 Ack=20 win=4143 Len=192
34	5.793721	10.10.80.99	61.135.177.59	TCP	1169	>	3724	[ACK]	Seq=20 Ack=787 win=64365 Len=0

图 5 游戏交互过程数据包 3

和前面一样,数据方向主要是:服务器→客户端,如 24,26,30,32 包;少部分数据包方向:服务器

→客户端,如图中 28 包,这些数据包也有一些规律,数据包内容如图 6 所示。

0030	fc 4f ce 59 00 00	af e9 95 95 63 da 4e 00 00 00	.O.Y... ..C.N...
0040	02 00 f7 69 9f 14 91 0b	07 f0	...f... ..

图 6 游戏交互过程数据包 4

AC(Alternation Characteristic)这类数据包数据部分长度固定为 20bytes,并有固定字段,见图 6 中黑线部分;

在游戏中的聊天输入栏中敲入:aaaaaaaaaaaa  
aaaaaaaaaaaa,然后找寻有无与此对应的数据包.在截获的数据包中找到 Client 发向 Game Server 的包(如图 7 所示)。

(3)聊天过程的分析

0000	00 03 e3 84 t1 48 04 04	04 04 04 04 08 00 45 00	.....H.. .....E.
0010	00 52 74 f4 40 00 80 06	16 4f 0a 0a 50 a3 db 85	.Rt.@... ..O..P...
0020	39 30 05 2c 0e 8c e5 f3	01 b7 d0 3a c9 66 50 18	90..... ..fP.
0030	41 d9 a9 26 00 00 26 02	f3 c3 49 83 03 00 00 00	A.&.&. ..I.....
0040	07 00 00 00 61 61 61 61	61 61 61 61 61 61 61 61	....aaaa aaaaaaaa
0050	61 61 61 61 61 61 61 61	61 61 61 61 61 61 61 61	aaaaaaaa aaaaaaaa

图 7 聊天过程的数据包

可以发现,魔兽世界的聊天封包的说话内容是明文的,每个聊天封包的结尾字节都是 0x00.并且每次聊天时客户端都是直接把聊天内容发送给服务

器,可见聊天过程是 C/S 模式,而非 P2P 模式。

(4)退出过程的分析

游戏在退出时产生的数据包如图 8 所示。

18284	23.178277	10.10.80.180	61.135.177.59	TCP	1881	>	3724	[FIN, ACK]	Seq=392 Ack=15284 win=65535 Le
18421	23.321627	61.135.177.59	10.10.80.180	TCP	3724	>	1881	[PSH, ACK]	Seq=15284 Ack=393 win=2264 Len
18422	23.321678	61.135.177.59	10.10.80.180	TCP	3724	>	1881	[FIN, ACK]	Seq=16098 Ack=393 win=2264 Len
18423	23.321903	10.10.80.180	61.135.177.59	TCP	1881	>	3724	[RST, ACK]	Seq=393 Ack=16098 win=0 Len=0

图 8 游戏退出时的数据包

AP(Alternation Process):本地游戏客户端请求退出,首先发出 FIN=1 TCP 报文段,接着服务器收到客户端的释放请求后,如果发送给客户端的数据尚未传送结束,则允许继续发送;随后,发完数据,然后服务器会再给客户端发送 FIN=1 的 TCP 段,请求释放;最后,客户端在收到服务器端的释放请求后,尽管服务器端已经释放成功,仍允许客户端给服务器端发出一应答加以确认。

4.1.3 在线更新过程

《魔兽世界》的更新是强制性的,玩家每次玩时都要保证所用的版本是最新的,否则玩不了,可见该游戏的新旧版本是不兼容的.《魔兽世界》有两种更新方式,一种是采用手动下载升级补丁(可执行文件),保存到游戏安装目录下,运行升级补丁即可完

成升级;另外一种就是在登陆的时候进行检查,若有更新的版本则进行在线升级;当 Client 向登录/账号服务器(Login Server)发送用户名及密码后,若无更新,直接连接 Game Server 进入游戏;若有更新,则下载并安装更新,然后再进入游戏.《魔兽世界》在线更新时用 P2P 方式来下载补丁包。

AU(Active Update)在这里表示在登录的时候进行检查若有更新的版本则主动地进行在线升级.魔兽世界是强制更新的,现在一般用“暴风下载器”更新,该下载器已经被嵌入到《魔兽世界》的安装程序中,一旦有新版本推出,并且玩家没有手动下载补丁包升级,那么一运行《魔兽世界》,游戏便会自动进行在线升级。

4.2 隶属度函数分析

本文以魔兽世界为例,通过一段时间的数据捕获、分析,得到登录类、在线交互类以及在现更新类的隶属度函数,如图 9 所示.

(1) 登录类隶属度函数

如图 9 所示,登录开始的时候,登录包占据了主导地位,随着登录交互过程的进行,登录过程在  $a$  点处稳定下来,并以半正态的方式开始递减,直至登录过程结束,几乎递减至零.

(2) 游戏交互类隶属度函数

如图 10(a)~10(c)所示,在游戏交互中,角色走动、角色对抗以及退出游戏的过程中,隶属度函数的图形是类似的.

$f_i (i=1,2,3)$  对应角色走动、角色对抗以及退

出游戏三种不同阶段捕获数据包频率,其中,  $f=$

$$\frac{\sum_{i=1}^n f_{ij}}{n}, s=\sqrt{\frac{\sum_{i=1}^n (f_{ij}-f)^2}{n-1}}, \text{ 而 } k \text{ 值反应了当该类数}$$

据包处于最高峰阶段持续的时间.

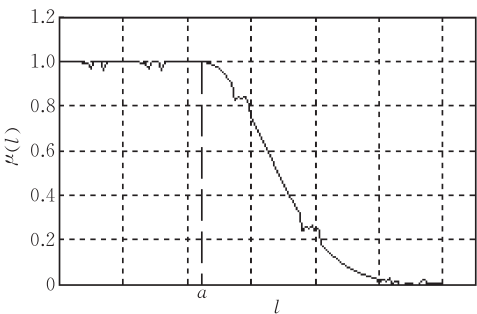
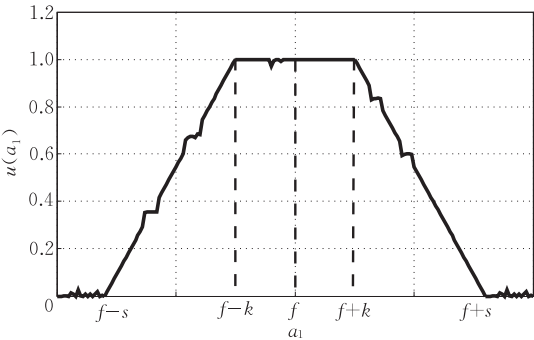
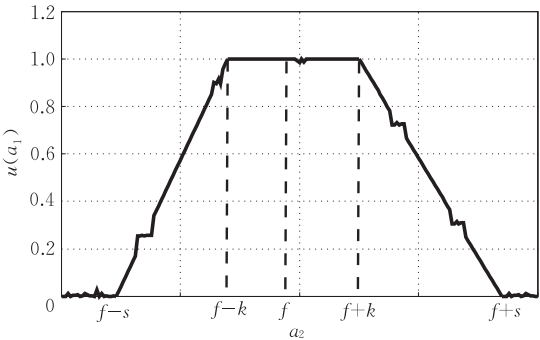


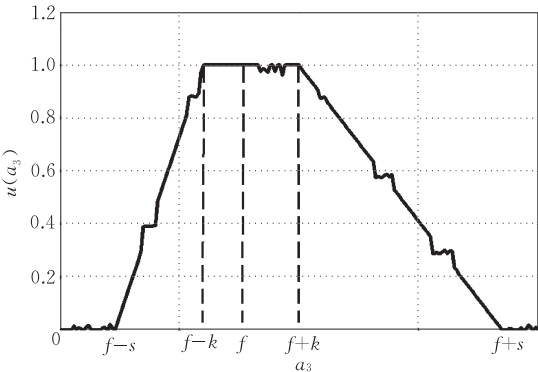
图 9 登录类隶属度函数



(a) 函数 1



(b) 函数 2



(c) 函数 3

图 10 游戏交互类隶属度函数

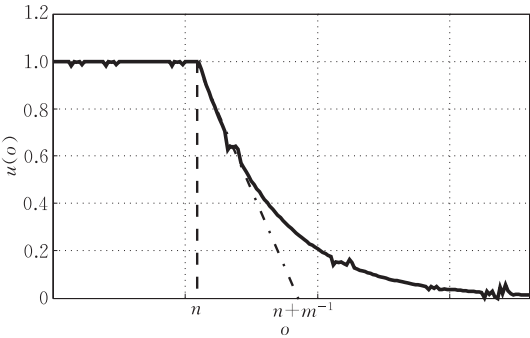


图 11 在线更新类隶属度函数

(3) 在线更新过程

如图 11 所示,在线更新的过程中,开始一段时间的在线更新数据包占主导,在更新下载稳定之后,从图中  $n$  点开始,数据包成指数倍下降,下降的趋势为  $n+m^{-1}$ ,最终更新过程结束时,数据包的数量也近乎为零.

4.3 结果分析

经过上述分析,可以总结出影响评判是否是魔兽世界的因素集为

$F = \{\text{登录过程}(f_1), \text{游戏交互中的角色走动信息}(f_2), \text{游戏交互中的角色对抗信息}(f_3), \text{退出游戏过程}(f_4), \text{在线更新过程}(f_5)\}$ .

经过一段时间的捕获,通过各个因素所表现出的隶属度图形,得到多组对每种因素进行评判时的评判值.

评语集:  $C = \{\text{很正确}(c_1), \text{正确}(c_2), \text{一般}(c_3), \text{不正确}(c_4)\}$ .

对  $F$  中的每一因素进行评判,设  $f_i (i = 1, 2, \dots, 5)$  的评判为  $\tilde{C}_i$ , 例如:

$\tilde{C}_1 = (0.2, 0.5, 0.3, 0), \tilde{C}_2 = (0.1, 0.3, 0.5, 0.1), \tilde{C}_3 = (0, 0.4, 0.5, 0.1), \tilde{C}_4 = (0, 0.1, 0.6, 0.3), \tilde{C}_5 = (0.5, 0.3, 0.2, 0)$ .

由此可以构成模糊评判矩阵  $\tilde{R}$ :

$$\tilde{R} = (\tilde{C}_1, \tilde{C}_2, \tilde{C}_3, \tilde{C}_4, \tilde{C}_5)^T$$
$$= \begin{bmatrix} 0.2 & 0.5 & 0.3 & 0 \\ 0.1 & 0 & 0 & 0.5 \\ 0.3 & 0.4 & 0.1 & 0.3 \\ 0.5 & 0.5 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.3 & 0 \end{bmatrix},$$

考虑到不同的因素对评判结果的侧重点不同,因此对每个因素还有一个相应的权值.

因素 1~5 的权值依次为: 0.1, 0.1, 0.3, 0.15, 0.35, 表示成模糊集  $\tilde{A} = (0.1, 0.1, 0.3, 0.15, 0.35)$ , 这样, 评判的结果为  $\tilde{B} = \tilde{A} \circ \tilde{R} = (0.35, 0.3, 0.3, 0.15)$ , 再进行归一化得

$$\left(\frac{0.35}{1.1}, \frac{0.3}{1.1}, \frac{0.3}{1.1}, \frac{0.15}{1.1}\right) \approx (0.32, 0.27, 0.27, 0.14).$$

从评判的结果来看, 只有 14% 的可能认为评判所捕获的数据流不是魔兽世界世界的. 所以, 我们可以断定所分析的流量是魔兽世界的数据流.

4.4 FCD 模糊识别方法用于识别其它的网络应用

Skype 的协议是经过加密的一种 P2P 应用协议, 它是一种特征不明确、协议尚未公开的 P2P 流, 无法通过关键字检测或端口检测的方式判断出来. 因此对该类应用的识别与分析比较困难.

通过对 Skype 的详细分析, 我们发现和前述网络游戏的魔兽世界相似, 在整个应用工作中它也有几个关键性的环节, 可以作为识别该种应用的切入点.

用 FCD 模糊识别方法对 Skype 进行识别的过程是:

(1) 捕获数据进行分析, 发现 Skype 的关键性的几个数据包: 超级节点 (Super Node, SN) 回复给客户端 (Skype client, SC) 的数据包、包含的“02”特征字的数据包、TCP 流量的特征包 1、TCP 流量的

特征包 2.

(2) 确定因素集为  $\{F = \{\text{SN 回复 SC 的数据包}(f_1), \text{包含的“02”特征字的数据包}(f_2), \text{TCP 流量的特征包 1}(f_3), \text{TCP 流量的特征包 2}(f_4)\}\}$ ;

(3) 对这 4 个因素的统一描述:

$D1(\text{长度}, \text{UDP/TCP}, \text{频率}, \text{特性}) = D1(11, \text{UDP}, \text{登录时出现频率 } 90\%, \text{第 4 个到第 7 个字节为 SN 的 IP 地址})$ ;

$D2(\text{关键字模式串所在的信令数据包的包长}, \text{关键字模式串的字节数}, \text{关键字模式串在报文中的位置}, \text{关键字模式串所代表的含义}) = D2(18, 1, 3, \text{连接时的一次呼叫})$ ;

$D3(\text{长度}, \text{UDP/TCP}, \text{频率}, \text{特性}) = D3(14, \text{TCP}, \text{登录时出现频率 } 80\%, \text{客户端向 ui.skype.com 服务器发送一个 HTTP 请求})$ ;

$D4(\text{关键字模式串所在的信令数据包的包长}, \text{关键字模式串的字节数}, \text{关键字模式串在报文中的位置}, \text{关键字模式串所代表的含义}) = D4(72, 16, 1, \text{与对方的 443 端口建立连接})$ .

$D5(\text{关键字模式串所在的信令数据包的包长}, \text{关键字模式串的字节数}, \text{关键字模式串在报文中的位置}, \text{关键字模式串所代表的含义}) = D5(93, 16, 1, \text{对方回复的 443 端口连接})$

(4) 对数据包进行一段时间的观察, 得出各个因素的隶属度函数, 确定对每个因素的评语集;

(5) 采用 4.3 小节中描述的计算方法, 计算出评判的结果;

(6) 根据评判结果可以较为准确地判定所识别的流量是否是 Skype 的流量.

同样, 用类似的方法, 采用 FCD 模糊识别方法, 通过客观的描述网络流量, 并对识别出的结果给出合理的评价, 可以识别出大多数网络应用数据流.

5 总 结

迄今为止, 对于如何识别某些特征不明显的网络应用, 仍然处在研究探讨阶段. 本文提出的 FCD 模糊识别方法, 能够较好地识别网络流量中的某些网络应用流, 对于其它的网络应用流量识别, 同样适用, 具有较好的准确性和可扩展性.

参 考 文 献

[1] Sears W, Yu Z, Guan Y. An adaptive reputation based trust framework for Peer-to-Peer application//Proceedings of the International Symposium on Network Computing and Appli-



- cations. Washington DC, USA. 2005; 13-20
- [2] Sen S, Wang J. Analyzing peer-to-peer traffic across large networks. *IEEE/ACM Transactions on Networking*, 2004, 12(2): 219-232
- [3] Ripeanu M, Foster I, Iamnitchi A. Mapping the Gnutella network; Properties of large-scale peer-to-peer systems and implications for system design. *IEEE Internet Computing Journal Special Issue on Peer-to-Peer Networking*, 2002, 6(1): 50-57
- [4] Leibowitz N, Ripeanu M, Wierzbicki A. Deconstructing the Kazaa Network//*Proceedings of the 3rd IEEE Workshop on Internet Applications (WIAPP'03)*. USA: The Printing House, 2003; 112-120
- [5] Karagiannis T, Broido A, Faloutsos M, Claffy K. Transport layer identification of P2P traffic//*Proceedings of the IMC'04*. Taormina, Sicily, Italy, 2004; 121-134
- [6] Izal M, Urvoy-Keller G, Biersack E W, Felber P A, Hamra A A, Garc'es-Erice L. Dissecting BitTorrent: Five months in a torrent's lifetime//*Proceedings of the 5th Passive and Active Measurement*. Franc, 2004; 1013-1017
- [7] Karbhari P, Ammar M, Dhamdhere A, Raj H, Riley G, Zengura E. Bootstrapping in Gnutella: A measurement study//*Proceedings of the 5th Passive and Active Measurement*. Franc, 2004. Heidelberg, Germany; LNCS, 2004; 1002-1009
- [8] Sen S, Spatscheck O, Wang D. Accurate, scalable In-Network Identification of P2P traffic using application signatures//*Proceedings of the 13th International Conference on World Wide Web (WWW2004)*. New York, USA, 2004; 512-521
- [9] Karagiannis T, Broido A, Brownlee N, Claffy K, Faloutsos M. Is P2P dying or just hiding?//*Proceedings of the Global Telecommunications Conference (GLOBECOM'04 IEEE)*. Dallas, USA, 2004, 3; 1532-1538
- [10] Manolis P, Maria P, Thomas K. Multilevel application-based traffic characterization in a large-scale wireless network//*Proceedings of the WOWMOM 2007*. Helsinki, Finland, 2007; 1-9
- [11] Gummadi K P, Dunn R J, Saroiu S, Gribble S D, Levy H M, Zahorjan J. Measurement, modeling, and analysis of a peer-to-peer file-sharing workload//*Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP-19)*. New York, 2003; 314-329
- [12] Saroiu S, Gummadi P K, Gribble S D. A measurement study of peer-to-peer file sharing systems//*Proceedings of the MMCN*. San Jose, 2002; 1008-1012
- [13] Tutschku K. A measurement-based traffic profile of the eDonkey filesharing service//*Proceedings of the 5th Passive and Active Measurement*, Franc, 2004; 988-995
- [14] Pagiamtzis K, Sheikholeslami A. Content address-able memory (CAM) circuits and architectures: A tutorial and survey. *IEEE Journal of Solid-State Circuits*, 2006, 41(3): 712-727
- [15] Li Jiang-Tao, Jiang Yong-Ling. Survey of P2P traffic identification and engineering technology. *Telecommunications Science*, 2005, 15(3): 57-61(in Chinese)  
(李江涛, 姜永玲. P2P 流量识别与管理技术. *电信科学*, 2005, 15(3): 57-61)
- [16] Soldani C. Peer-to-peer behaviour detection by TCP flows analysis. University of Liege Faculty of Applied Sciences, 2004, 5
- [17] Sen S, Wang J. Analyzing peer-to-peer traffic across large networks. *IEEE/ACM Transactions on Networking*, 2004, 12(2): 219-232



**SUN Zhi-Xin**, born in 1964, Ph.D., professor. His current research interests include network security, P2P computing and software engineering.

**GONG Jing**, born in 1978, lecturer. Her current research interests focus on P2P computing.

## Background

At present, common P2P flow identification ways include the methods based on detecting default service ports, transport layer, payload, network dynamic characteristic, etc. However, these methods can't analyze and recognize P2P applications which have inconspicuous characters, or unopened protocols. This paper applies the theory of fuzzy mathematics to P2P flow identification, and proposes a fuzzy identification method based on flow characteristic description (Flow Characteristic Description, FCD). FCD fuzzy identification method first analyzes captured P2P flows, and then

describes key flows uniformly. Subsequently, the paper analyzes distributional situation of data in these key flows, and obtains their membership functions, which can be used as comment aggregates. Finally, fuzzy judgment method determine whether the data are some kinds of network applications. FCD fuzzy identification can more accurately recognize P2P applications with ambiguous characters, or unopened protocols, such as some large-scale network games.

The research group has many publications of high-quality and patents in this domain.