高技术通讯 2010 年 第 20 卷 第 1 期:38 ~ 44

doi:10.3772/j.issn.1002-0470.2010.01.007

基于流量统计实时识别 QQ 语音通信的方法①

李 冰② 金志刚③* 舒炎泰

(天津大学计算机科学与技术学院 天津 300072)

(*天津大学电子信息工程学院 天津 300072)

摘 要 针对腾讯 QQ 语音应用的识别问题,建立了客户源端模型,并对流量进行了离线分析。在此基础上,提出了基于流量统计的两种识别方法:Bayesian 识别法和心跳识别法。一方面,利用 QQ 语音应用的源端编码特点,得出语音数据流量的本质特征,并使用 Bayesian 理论进行检测。另一方面,针对 QQ 语音应用中一些非语音数据包的周期性特点,利用卡方检验,提出了心跳识别的方法。实验结果表明,这两种方法综合使用,可以很有效地实时检测 QQ 语音流量。

关键词 QQ 语音,流量统计识别,Bayesian 判别

0 引 言

随着互联网技术的发展,越来越多的即时通信软件涌入人们的生活,这些软件不仅能提供文本通信,而且随着网络语音(voice over Internet protocol, VoIP)[1]技术的发展,很多软件都已经提供了语音服务。

近年来,由于实时语音流量占全部网络流量的比重在增大,对其进行准确识别,已经成为一个很有研究价值的课题。首先,检测是网络流量分析的前提。作为一种流量分析工具,检测有助于管理人员管理网络,同时更有助于研究人员改善网络性能,为下一步的网络设计提供帮助。其次,任何一种业务的准确识别都将为入侵检测提供便利。将网络中大量的合法业务准确地识别出来,有助于入侵检测制定更可靠的规则集。最后,对语音应用所产生的流量进行识别,可更好地分析应用性能,为应用的改进提供理论指导。

腾讯 QQ 作为即时通信软件的代表,集成了文本、语音服务。目前,它已拥有 4.3 亿注册用户,2.9 亿活跃用户账号,是用户群最大的即时通信软件。在语音通讯方面,它使用 P2P (peer-to-peer)[2]技术,并不断尝试新的突破。从 2004 年,腾讯开始采用 GIPS (Global IP Sound)技术,它可根据网络环境自动调节编码码率,提供低码率高质量的语音。这使得腾讯 QQ 在市场中赢得了更大的业务量。它的流行性使得对它的识别具有很高的实用价值。

但是目前对 QQ 语音识别工作的研究面临较多的困难。首先,QQ 作为一种商业软件,通信协议不公开。其中的部分信令使用了加密算法,使得对协议作反向工程很困难。其次,QQ 的版本众多,在此基础上的插件也较多,识别工作要兼顾各种版本和插件来进行。再次,它采用了端口伪装技术,可以随机配置端口号;并且服务器有多个不固定的 IP 地址。这些使得难以通过端口号或 IP 地址的方式对其做到完全识别。以上原因,使得到目前为止并没有很成熟的方法可以对 QQ 语音进行准确检测。

本文针对 QQ 语音应用的特征,提出了一种基于流量统计特性的识别方案。一方面,通过对 QQ 语音应用数据包的产生过程进行分析,找出由源端编码技术所决定的数据包在网络流量上的特征,利用 Bayesian 方法对携带有语音数据的流量进行识别,另一方面,针对 QQ 语音应用中特定的控制信息包的周期性特点,将其作为识别依据,提出心跳识别的方法。该方法弥补了由于编码技术的相似性而使得在 Bayesian 方法中存在的误判。

1 相关工作

在网络通信流量分类识别方法中,为了避免端

① 863 计划(2007AA01Z220)和天津市自然科学基金(08JCYBJC14200)资助项目。
② 女,1982 年生,博士生;研究方向:计算机网络管理与安全;E-mail: libingice@tju.edu.cn
③ 通讯作者,E-mail: zgjin@tju.edu.cn
(收稿日期:2008-10-15)

口号检测的不准确性,目前研究较多的方法是使用流统计状态的识别。

文献[3,4]将很成熟的统计方法应用到流量识别的领域。其中文献[3]中使用了最近邻(nearest neighbor, NN)和线性判别分析(linear discriminant analysis, LDA)的统计方法。而文献[4]使用了 Bayesian 方法。但由于它们都使用了流的持续时间,流中所包含的数据包数等作为统计特征参量,使得这些方法只能作为一种离线的识别手段,不具备实时性。另外,由于它们不是针对具体应用的研究,所以在特征参量的选取过程中,也没有考虑流源端数据包形成过程中所具备的特性。

文献[5]中提出了对 VoIP 应用的总体识别方法。该方法是在考虑语音通信的机密性和实时性的基础上对流行为进行的研究。但研究仅局限于网络状况较好的情况,对于网络带宽受限的情况并未给出相应的识别方法。而且由于该方法是针对所有 VoIP 应用的,因而不具有针对 QQ 语音应用的专门性。

在 VoIP 的应用中,目前研究较多的是 Skype 应用。文献[6-8]中都提出了对 Skype 流量的识别方法,但是由于 QQ 语音使用的传输协议与 Skype 不同,所以这些方法也仅能作为参考。文献[6,7]中用到了 Skype 应用中存在的超级节点这个特点,对超级节点的 TCP 连接有一个默认端口号(33033),但这在 QQ 应用中并不存在。在文献[8]中,使用 Skype 的特定字段作为识别的重要依据,而这些特定字段和 QQ 应用是完全不同的。

对于 QQ 语音流量的识别,目前的研究尚处于不成熟阶段。文献[9]中提出了一个较完整的识别算法,它是在基于净荷深度检测和会话关联技术的基础上建立的模型。但是由于该模型是基于 QQ 语音的净荷特点提出来的,当由于升级等原因使得 QQ 协议稍有变化时,这种基于特定字段的匹配方案就需要重新寻找匹配字段,因而不具有版本的兼容性。更重要的是,如果用户使用第三方软件对协议进行加密,这种识别方式将完全失效。

本文提出的基于流量统计的 QQ 语音识别方法,在针对该应用特点的基础上,借鉴了流统计的理论基础,在提供实时识别方法的前提下也保证了较高的准确率。

2　QQ 语音源端模型剖析

由于 QQ 协议不公开,故只能通过对其流量的观测来提取特征建立模型,使用 Nistnet[10] 作为网络仿真器,仿真不同的网络带宽、延迟和丢包率。在不同的网络环境下,运行 QQ 语音应用,同时使用 Wireshark 捕获网络数据包,分析流量,进而建立模型,推测数据包的结构。

2.1　建立 QQ 语音源端模型

QQ 语音数据处理流程如图 1 所示。首先根据 QQ 语音应用所获知的当前的网络条件,对声音进行特定的采样、编码,形成编码块,然后附加上头部(即帧头)后形成语音帧。接着,一个或多个语音帧组合在一起附加上一个数据包头组成语音数据包。然后,数据包被发送到网络中。

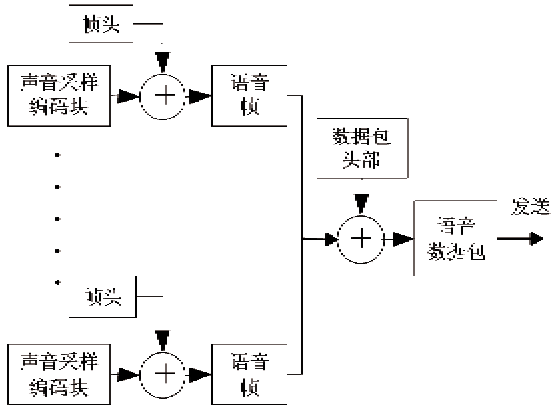


图 1　QQ 语音数据处理流程示意图

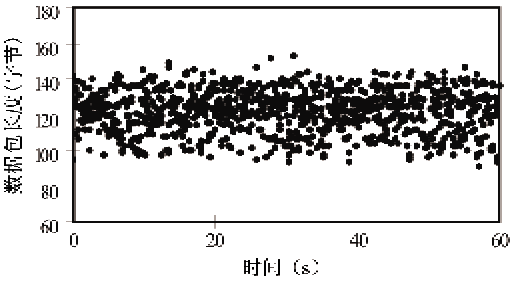
QQ 语音使用的是 GIPS 语音编码方案,主要编码方法如表 1 所示。在编码端有 3 个很重要的参数:(1)编码方法的比特率(Ratebit)——表示每秒钟编码(压缩)后的语音数据的比特数,用 R 表示;(2)编码方法所使用的帧大小(FrameSize)——相邻两个语音帧间隔的时间,用 FS 表示;(3)数据包中所包含的帧个数,用 n 来表示。由此可知 $FS \times n$ 将近似地等于两个相邻语音数据包的时间间隔大小 ΔT。

表 1　GIPS 的主要编码方案

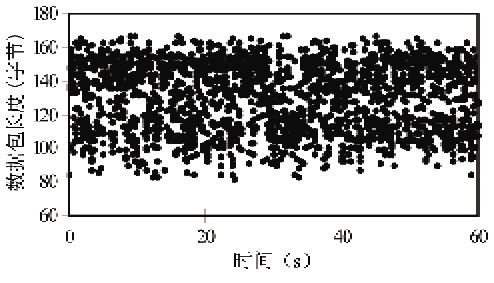
编码方法	帧大小(ms)	比特率(kbps)
IPCM-wb	10,20,30,40	80(平均)
ILBC	20,30	13.3,15.2
EG.711	10,20,30,40	48,56,64
ISAC	30-60	10-32

选取语音数据包长度和相邻语音数据包间隔作为观测参数。根据 GIPS 的编码方案,认为数据包长度小于 40 字节或相邻数据包间隔小于 1ms 的包均

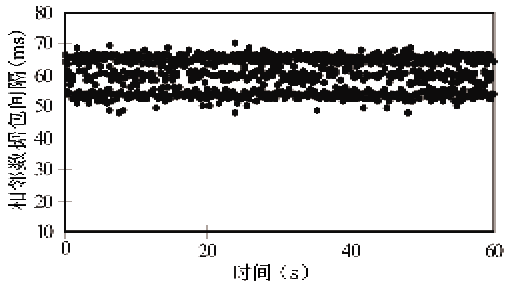
不携带语音数据。图 2 显示了不同网络带宽条件下,各参数的变化。可以看到,网络条件对数据包长度和相邻数据包间隔的影响是很明显的。在 32kbps 带宽条件下,数据包长度和相邻数据包间隔的趋势分别为 120 字节,60ms;而在 48kbps 带宽条件下,分别为 130 字节和 30ms。



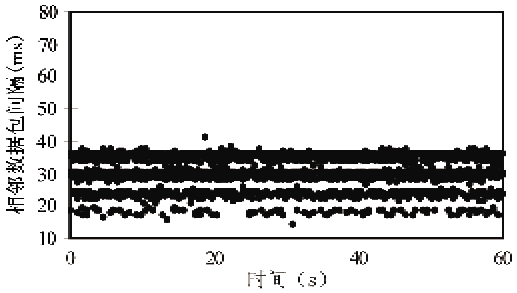
(a) 32kbps 下,数据包长度的散点图



(b) 48kbps 下,数据包长度的散点图



(c) 32kbps 下,相邻数据包间隔的散点图



(d) 48kbps 下,相邻数据包间隔的散点图

图 2 不同网络带宽条件下,QQ 语音 UDP 流量特征散点图

2.2 QQ 语音数据包

对捕获到的语音数据包进行分析。对于由用户数据报协议(UDP)建立的连接,数据包有 18 个字节的头部。格式如图 3 所示,下面给出几个关键字段的解释。

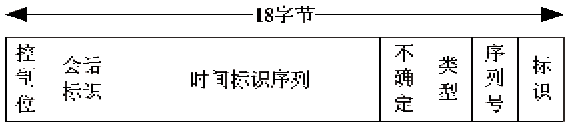


图 3 QQ 语音数据包头部格式

时间标识序列:9 个字节。可以精确到秒,即数据包每秒钟更新一次该序列。

类型:1 个字节。标识该数据包的类型。为了使语音帧的编码与网络环境相适应,发送端需要知道当前和接受端之间的网络状况。因此在发送语音数据包的同时,还要发送一些探测包和控制信息包。值 0X05 代表语音数据包,目前已知的探测包和控制信息包的值有 00X1,0X02,0X21,0X22,0X0f。

序列号:2 个字节。数据包序列号。每种类型的数据包序列号单独编排。每发送一个该类型的数据包,则其序列号加 1。最大值为 0Xff ff,当数据包的数目超过这个值时,该序列号可以循环使用。

标识:1 个字节。用来表示该数据包是否为语音数据包。

对于由传输控制协议(TCP)建立的连接,其数据包头部相对简单一些,由 11 个字节组成。相比 UDP 省略了控制位,会话标识和时间标识序列,同时增加了一个字节表示数据包的长度。

当类型位的值为 0X05 时,数据包头部之后将是一个或者多个语音帧。每一个语音帧由帧头和数据两部分组成。语音帧的帧头为 14 个字节。其格式如图 4 所示。

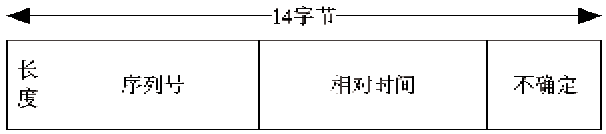


图 4 语音帧帧头格式

长度:1 个字节。指不包括头部的语音帧长度。

序列号:4 个字节。语音帧的序列号。本次会话的第一帧的序列号是由系统时间生成的,以后每个语音帧的序列号顺次加 1。

相对时间:语音帧编码的相对时间,以毫秒为单

位。对这个字段的观测,可计算相邻数据包在源端的间隔值。这个间隔值在后面的识别中会用到。

2.3　探测包和控制包

应用程序除了发送携带有实际语音编码数据信息的数据包之外,还会有规则的发送探测包和控制包,相对于携带有语音信息的数据包,这些数据包的长度明显较短。在 UDP 连接中,应用层数据长度一般是 18B,19B,22B,69B。其中 18B 的包仅有包头部分,其作用仅仅是应答,相当于 TCP 连接中的 ACK 包。在 TCP 连接中,一般是 12B,15B,62B。

3　QQ 语音的识别

本节提出了两种基于统计的 QQ 语音流量识别方法:Bayesian 识别法和心跳识别法。在 Bayesian 识别法中,以携带语音编码数据的流量作为统计对象,根据对 QQ 语音应用的源端模型分析,计算衍生信念值,从而对流量进行判别。在心跳识别法中,以非语音数据的流量作为统计对象,定义了一种叫做心跳的数据包,使用卡方作为统计量,对心跳包进行识别,从而识别出 QQ 语音流量。整体的设计流程如图 5 所示。首先对流量进行 Bayesian 识别,如果未通过,则表明该流量未表现出语音流量的特征,故直接将其判定为非 QQ 语音流量。否则继续进行心跳识别,考虑到网络中各种因素可能对心跳识别法造成较大的干扰,所以在流第一次未通过心跳法识别时,仅是将流标记为疑似流,然后重新判定。

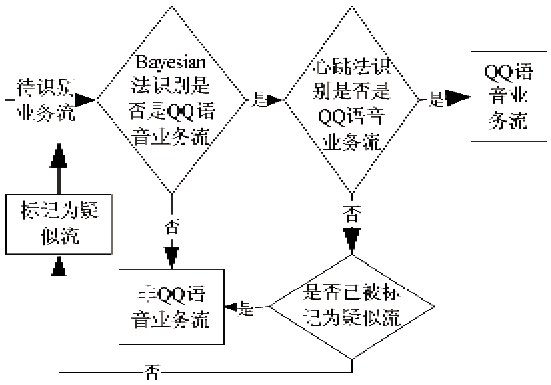


图 5　QQ 语音流量识别方案的整体构架

3.1　Bayesian 识别法

QQ 语音流量在不同的网络条件下显示出不同的统计特点。选取数据包长度(S)和包间隔(T)作为主要的随机变量参数,应用 Bayesian 方法进行识别。由于这两个变量受到网络环境和源端编码环境中大量相互独立因素的影响,故将这两个变量建模为二元正态分布且认为两个变量之间相互独立。

3.1.1　Bayesian 判别原理

若事物特征参量 X 的抽样值为 x,则事物属于类别 C 的概率 $P(C|x)$可以表示为

$$P(C \mid x) = \frac{P(x \mid C)}{P(x)} P(C) \tag{1}$$

这就是 Bayesian 判别,它给出了一种用先验概率 $P(x \mid C)$ 求后验概率 $P(C \mid x)$ 的方法。

但在将 Bayesian 定理应用于实际判别时,往往更习惯采用最大似然准则,此时先验概率被称作信念(belief)。信念 $P(x \mid C)$ 越大,则事物属于类别 C 的概率就越大。

3.1.2　参数的选取

选择少数有代表意义的流量特征作为参数进行识别。由于其实时语音通信的本质,使得 QQ 语音流量具有持续低比特率的特征:整个数据流是由一系列长度较小的数据包组成,其具体的数据包长度和相邻数据包间隔与具体的编码方案相关。

首先考虑数据包长度。根据 QQ 语音消息处理流程(图 1),数据包长度的均值可以表示为

$$\mu_1 = (R \cdot FS + H_1) \times n + H_2 \tag{2}$$

这里 H_1 代表语音帧帧头的长度,H_2 代表数据包包头的长度。包长度的标准方差应与具体的编码方式相关,经过实验观测,建议使用 $\sigma_1 = 7$。对于标准方差的大小,在实际实验中对 μ_1 的选取间隔以 10 字节为单位。

对于相邻数据包间隔,均值可简单的表示为

$$\mu_2 = FS \cdot n \tag{3}$$

它的方差与编码方式无关,建议值为 $\sigma_2 = 1\text{ms}$。

尽管在 GIPS 中,提供的语音帧大小的值从 10ms 到 60ms 不等,但是经过对 QQ 语音流量的长期观测发现,常用的语音帧大小仅为 30ms 和 60ms。由此,相邻数据包间隔一般应该为 30ms,60ms,90ms,120ms。

3.1.3　衍生信念值

在识别过程中,特征参量 $X = [S, T]$ 的某一次观测值 x_i 可能具有片面性。为此,将一小段时间 t 内的概率均值作为判定的依据。为了避免得到的信念值 $B(C)$ 接近 0 而不方便比较,这里对概率值取对数之后再进行比较。于是,QQ 流量的衍生信念 $B(C)$ 可以表示为

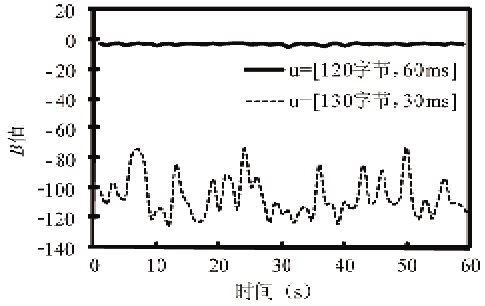
$$B(C) = \log\left[\frac{1}{w}\sum_{i=1}^{w} P(x_i \mid C)\right] \tag{4}$$

其中，w 为小段时间 t 内 X 的抽样值个数(也就是数据包的个数)。建议使用的 t 值为 1s，在 1s 内，大约有 15 到 40 个抽样值。

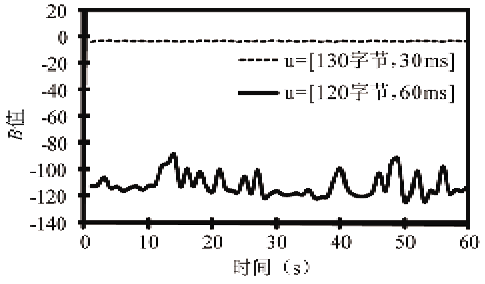
由于参数的选取与具体编码方式相关，故计算得出的信念值也与编码参数相关，如图 6 所示。取各种编码方式中最大值

$$B = \max(B(C)) \tag{5}$$

作为最终判定流量是否是 QQ 语音的依据。



(a) 32kbps 下，不同参数下的信念值



(b) 48kbps 下，不同参数的信念值

图 6 不同网络带宽条件下的信念值

当我们选取合理的均值和方差时，由 QQ 语音流量计算得到的 B 值会很大。选取一个合理的阈值 B_{min}，如果 $B > B_{min}$，判定这个流量可能为 QQ 语音；否则认为它不是由 QQ 语音产生的。本文在实验过程中所使用的 B_{min} 值在 UDP 流时为 −5，在 TCP 流时为 −7。

3.2 心跳识别法

在语音应用发送的探测包和控制包中，有一类包很有规律，可用来识别 QQ 语音应用。在 UDP 流量中，应用每隔 5s 发送一个探测包对。这个包对中的两个包长度均为 69B，内容也完全相同且同时发送(在 100Mb 的主流接入带宽中其发送间隔不超过 0.1ms)。这个包对将帮助应用、了解当前网络状况，应用根据当前的网络状况来决定具体采用何种编码方式。在 TCP 流量中，应用同样每隔 5s 发送一个探测包对。但包的长度由于包头的长度不同而变成了 62B。我们将这种具有发送周期的包，形象地称之为心跳。

尽管心跳具有很准确的发送周期，但是考虑到网络环境中多种复杂因素的影响，在检测点处检测到的心跳包对的间隔和心跳周期应该会有一定的抖动。在 UDP 流中，如果两个长度为 69B 的包的间隔不超过 1ms，即认为是心跳包对。用高斯分布来建模心跳周期。对流量持续观测一小段时间(建议 30s)，则可以利用公式

$$\chi^2(n) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(T_i - t)^2 \tag{6}$$

来检测观测到的心跳周期进行判别。其中 T_i 是观测的心跳周期，n 是观测的周期次数，t 是标准的心跳周期，其值为 5s，考虑到网络诸多因素的影响并参考大量实验观测，建议将心跳标准差 σ 的值取为 0.2s。如果 $\chi^2(n) < \chi_{min}$，则认为该流量属于 QQ 语音，否则认为该流量未通过心跳检测。χ_{min} 的值可以根据具体环境的需要，选取不同的分位点。

4 实验与算法性能评价

为了验证本文提出的设计方案的性能，设计了如下的实验场景，如图 7 所示。主机 A 上运行各种常见的网络应用，这里主要包括：Web 浏览器客户端，文件传输协议(FTP)客户端，BitComet 下载软件，QQ 聊天软件(包括语音和文本两部分)，Skype 网络电话。在主机 C 上安装 LINUX 操作系统，打开路由功能，运行 Nistnet 软件模拟各种网络环境，主机 B 上运行该识别方案。对不同网络环境下主机 A 的流量行为进行监测。

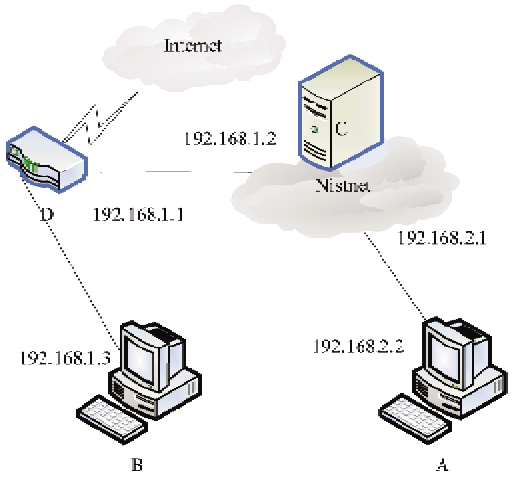


图 7 实验场景图

实验结果显示,在 Bayesian 判别法中,QQ 语音应用的流量数据所对应的信念值很大,而大多数的非 QQ 语音应用所对应的信念值都很小,例如,FTP 的下行流量和 QQ 应用文本传输流量数据的信念值趋于负无穷。因此 Bayesian 判别法可以将 QQ 语音流量与大多数流量区分开,尤其是与非流媒体流量区分开。

但是在实验过程中发现,由于 Skype 应用也使用了 ILBC 和 ISAC 编码方案,尽管它与 QQ 语音使用的封包方法可能不同,但是对 Skype 语音产生的流量使用 Bayesian 判别之后,也会有较大的信念值,在部分情况下会将它误判为 QQ 语音流,图 8 显示的是实验结果的很小一部分。可以看到有一部分 Skype 流的 B 值是大于 -5,因此它们能通过 Bayesian 识别法,被暂时误判为 QQ 语音流。但由于在 Skype 应用中,没有和 QQ 语音相似的心跳,所以在心跳判别时可以将二者区分开。

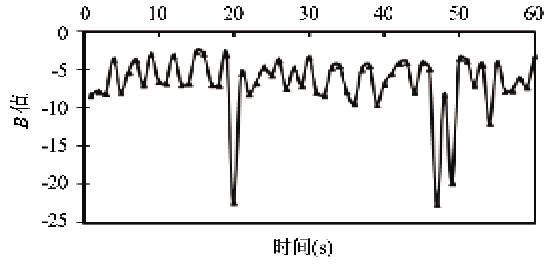


图 8　在 32kbps 下,Skype 语音流量得到的 B 值(UDP 传输)

使用漏判率($FN\%$)和误判率($FP\%$)作为对识别结果准确度的评判参数。这里漏判(FN)指的是错误地将 QQ 语音流量判断为非 QQ 语音流;误判(FP)指的是错误地将非 QQ 语音流量判断为 QQ 语音流量。若流的总数为 N,其中有 N_Q 条流为 QQ 语音流,同时,在我们的判别方案中有 N_I 条流被识别为 QQ 语音流,而其中有 N_{IC} 条流是被正确识别为 QQ 语音流的,则漏判率和误判率可以通过以下两个公式求出:

$$FN\% = \frac{N_Q - N_{IC}}{N_Q} \times 100\% \tag{7}$$

$$FP\% = \frac{N_I - N_{IC}}{N - N_Q} \times 100\% \tag{8}$$

在网络条件较好的情况下,QQ 语音应用会建立 UDP 连接,这时 QQ 语音流的特征会很明显,所以将 B_{min} 的值设置的较大(这里使用的是 -5),而当网络状况差一些的时候,会建立 TCP 连接,这时 TCP 会不断增减拥塞窗口,使得语音流的特征不断变化,所以建议将 B_{min} 的值设置的较小一些(这里使用的是 -7),以免产生太大的漏判率。

实验结果如图 9 所示,综合使用 Bayesian 判别法和心跳判别法得到的漏判率小于 1.98%,而误判率则更小,不足 0.3%。这里错误地将 QQ 语音流量判断为非 QQ 语音流的主要原因是当网络状况波动极大时,QQ 语音应用将会不断地根据网络状况调整源端的编码方式,这时可能会出现一些异常流量,这样使得有时 Bayesian 判别法错误地将其归并到非 QQ 语音应用流量中。

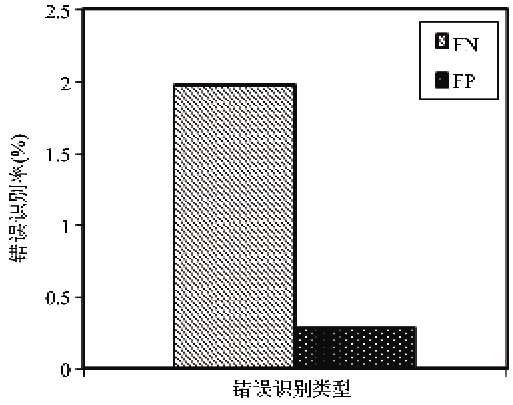


图 9　实验结果的错误识别率

在进行对 QQ 语音流量高准确度的识别的同时,我们也对其他的 VoIP 应用进行了一定程度的研究。以微软的 MSN 语音为例,在对它的 trace 文件分析过程中,我们也发现了一些几乎同时在源端发送的非数据包对,但是这些包对与 QQ 中的不同,它们不具备很严格的周期,因而不能称之为心跳。这一方面说明心跳方法可以有效地将 QQ 应用的流量和其他相似应用区分开来。另一方面,心跳方法不可以简单地直接扩展到对其他同类应用的识别中。但是我们对非数据流量进行研究,从而作为识别标准的方法,是可以扩展到其他相似应用流量的识别中的。

5　结　论

本文对 QQ 语音应用的客户源端建立模型,对其流量进行离线统计,分析编码特点和数据包结构,进而提出了两种基于流量统计的识别方法:Bayesian 识别法和心跳识别法。这两种方法分别以语音数据流量和非语音数据流量作为检测对象,从不同角度进行识别,可以实现优势互补。实验结果表明,这两

种方法综合使用,能够很有效地实时检测出QQ语音流量。同时我们的方法也为今后进行其他相似应用的识别,提供了可以借鉴的理论。

参考文献

[1] Goode B. Voice over Internet protocol (VoIP). *Proceedings of the IEEE*, 2002, 90(9):1495-1517

[2] 关志涛,曹大元,祝烈煌等. 混合P2P环境下基于信度模型的激励策略. 北京理工大学学报,2007,27(7):599-603

[3] Roughan M, Sen S, Spatscheck O, et al. Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification. In: Proceeedings of the 4th ACM SIGCOMM conference on Internet measurement, Taormina, Sicily, Italy, 2004. 135-148

[4] Moore A W, Zuev D. Internet traffic classification using Bayesian analysis techniques. In: Proceedings of ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, Banff, Canada, 2005. 50-60

[5] Okabe T, Kitamura T, Shizuno T. Statistical traffic identification method based on flow-level behavior for fair VoIP sevice. In: Proceedings of the 1st IEEE Workshop on VoIP Management and Security, Vancouver, Canada, 2006. 35-40

[6] Perenyi M, Gefferth A, Dinh Dang T, et al. Skype traffic identification. In: Proceedings of Global Telecommunications Conference, Washington, D.C., USA, 2007. 399-404

[7] Perenyi M, Molnar S. Enhanced Skype traffic identification. In: Proceedings of the 2nd International Conference on Performance Evaluation Methodologies and Tools, Nantes, France, 2007

[8] Yu Y F, Liu D D, Li J, et al. Traffic identification and overlay measurement of Skype. In: Proceedings of 2006 International Conference on Computational Intelligence and Security, Guangzhou, China, 2006. 1043-1048

[9] 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. 重庆邮电学院学报,2006,18(6): 789-792

[10] Carson M, Santay D. NIST Net: a Linux-based network emulation tool. *ACM SIGCOMM Computer Communication Review*, 2003,33(3):111-126

Real-time identification of QQ voice communication based on traffic statistics

Li Bing, Jin Zhigang*, Shu Yantai

(School of Computer Science and Technology, Tianjin University, Tianjin 300072)

(*School of Electronics and Information Engineering, Tianjin University, Tianjin 300072)

Abstract

This paper establishes the source model of QQ voice and analyzes its traffic offline in order to identify the QQ voice traffic. Based on that, it proposes two identification methods: the Bayesian method and the heartbeat method. On the one hand, the theory of Bayesian is employed to identify the essential character of voice traffic, which is extracted according to the encoding at source. On the other hand, the heartbeat method is presented based on the Chi-square test by identification of the periodic feature of non-voice traffic. The experiment results show that the integrated deployment of the two methods is very effective for real-time detection of QQ voice traffic flow.

Key words: QQ voice, statistic traffic identification, Bayesian class