第33卷 第17期 计 算 机 工 程 2007年9月
Vol.33 No.17 Computer Engineering September 2007
·网络与通信· 文章编号：1000—3428(2007)17—0137—03 文献标识码：A 中图分类号：TP301.6

基于智能会话关联的腾讯语音流量识别算法

王 攀，金 婷，张顺颐，陈雪娇，李 薇
（南京邮电大学信息网络技术研究所，南京 210003）

摘 要： 通过实验分析了国内流行的即时通信软件——腾讯QQ的流量特征以及其语音会话的流量特征，应用净荷深度检测(DPI)和智能会话关联(ISA)技术来识别腾讯语音通话流量，设计了腾讯语音业务流量的识别模型和算法。模型和算法的准确性、可扩展性和健壮性在电信运营商IP骨干网中得到了验证。

关键词： 净荷深度检测；流量识别；VoIP识别；会话关联

Algorithm of Tencent's Voice Traffic Identification Based on Intelligent Session Association

WANG Pan, JIN Ting, ZHANG Shun-yi, CHEN Xue-jiao, LI Wei
（Research Center of Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003）

【Abstract】 This paper aims at voice traffic characteristics of Tencent QQ, which is the currently popular Internet instant communication software. It deeply focuses on how to effectively identify QQ voice traffic based on deep packet inspection(DPI) and intelligent session association(ISA) technology, then presents a model of identification and related algorithm. The voice identification system and the accuracy, scalibity, robustness of the model and algorithm are validated in core network of China Telecom.

【Key words】 deep packet inspection(DPI); traffic identification; VoIP identification; session association

随着即时通信软件和VoIP技术的发展，话音业务的格局发生了巨大的变化，即时通信软件低廉的资费和便利性吸引了大量的普通用户。在巨额利益的吸引下大量虚拟VoIP运营商充斥电信市场，不仅导致合法运营商话务量流失，更打破了原有电信市场的竞争格局，给传统的话音业务带来了巨大的冲击，电信运营商正遭受非法VoIP给其带来的巨大挑战，因此，必须将互联网上VoIP业务纳入良性控制的范畴。腾讯是国内最为流行的即时通信软件，其通话行为目前仍为PC2PC，也分流了不少传统话务用户，因此，对于电信运营商，了解和良性控制腾讯的语音业务具有重大意义。

1 腾讯业务的通信机制研究

1.1 腾讯语音业务的识别现状

针对腾讯QQ及其语音业务的识别具有一定的难度，原因如下：

(1)QQ的通信协议为私有协议，且其中部分信令采用加密算法；

(2)QQ的版本众多，升级比较频繁；

(3)现今大部分对QQ的研究集中在QQ登录退出过程以及文本聊天交互方式上，鲜有对其语音过程的分析，所以，可借鉴之处不多；

(4)腾讯QQ采用端口伪装技术，使用80端口；端口可随机配置；服务器有多个非固定IP地址，难以做到完全控制；

(5)QQ提供文本、数据、语音、视频等业务，各种业务的会话特征均不相同，因此对服务器IP地址的“野蛮”封堵并不是解决问题的根本办法，反而导致QQ的正常通信无法使用。

由此可见，采用传统的端口过滤、IP地址过滤以及协议分析等业务识别方法很难识别出QQ的语音过程，因此，必须另辟蹊径。

1.2 QQ会话及QQ语音业务会话的净荷特征分析

本文定义了2个术语：(1)QQ会话：泛指用户登录QQ之后所有的交互行为，包括登录认证、即时消息、语音、视频、退出等QQ交互过程；(2)QQ语音通话：特指QQ会话中的语音和视频通信过程。因此，一个QQ号码对应一个QQ会话，而QQ语音通话则特指一个QQ用户同另一个QQ用户的语音通信过程。

通过采用净荷深度检测(deep packet inspection, DPI)技术，测试QQ会话和语音通信的交互过程，发现二者均具有起始和末尾净荷为0x02/0x03的特征，而语音会话采用类似SIP协议的通信交互机制建立会话。因此，可以采用净荷深度检测机制和简单的协议分析技术来识别QQ的语音会话。净荷特征匹配串为“SIP/user-agent: Tencent-VQQ”、“SIP/reason=100”等。

2 基于DPI和ISA的腾讯语音业务识别模型和算法

2.1 问题的提出

区分和识别腾讯的语音业务面临如下几个问题：

(1)腾讯QQ的端口可变，且易隐藏在80端口后面，此外服务器地址也都在域名的掩盖下不断变化，因此，简单的端口和IP地址匹配不适用；

基金项目： 国家“863”计划基金资助项目(2005AA121620)
作者简介： 王 攀(1979－)，男，助教、博士研究生，主研方向：NGN服务质量，VoIP和P2P安全；金 婷，硕士研究生；张顺颐，教授、博士生导师；陈雪娇，硕士；李 薇，硕士、副教授
收稿日期： 2007-03-19 **E-mail：** wangpan@njupt.edu.cn

(2)腾讯的业务有多种，如即时消息、语音、视频、数据传输，应如何无二义性地将语音业务从众多腾讯业务中区分和提取出来；

(3)腾讯升级频繁，应如何适应不断变化的应用特征；

(4)识别要求高准确性、低误报率和漏报率。

2.2 模型的总体设计

腾讯 QQ 是一个综合业务的即时通信系统，无论即时消息、语音、视频、数据传输都是腾讯的会话业务，对 QQ 语音业务的识别即：将 QQ 语音信令流和媒体流从 QQ 会话数据流中分拣出来，便于分析通话双方的主被叫 IP 地址、语音编解码类型、QQ 语音服务器地址等详细信息，甚至更深入地分析 QQ 语音的通话内容。

本文设计了一种基于净荷检测和会话关联技术的 QQ 语音业务识别模型和算法，其模型和技术框架如图 1 所示。从中可以看出，系统分为 4 个层面，从下往上依次是：数据采集层，协议分析层，流量识别（业务感知）层和 QQ 语音业务应用层以及表现层。

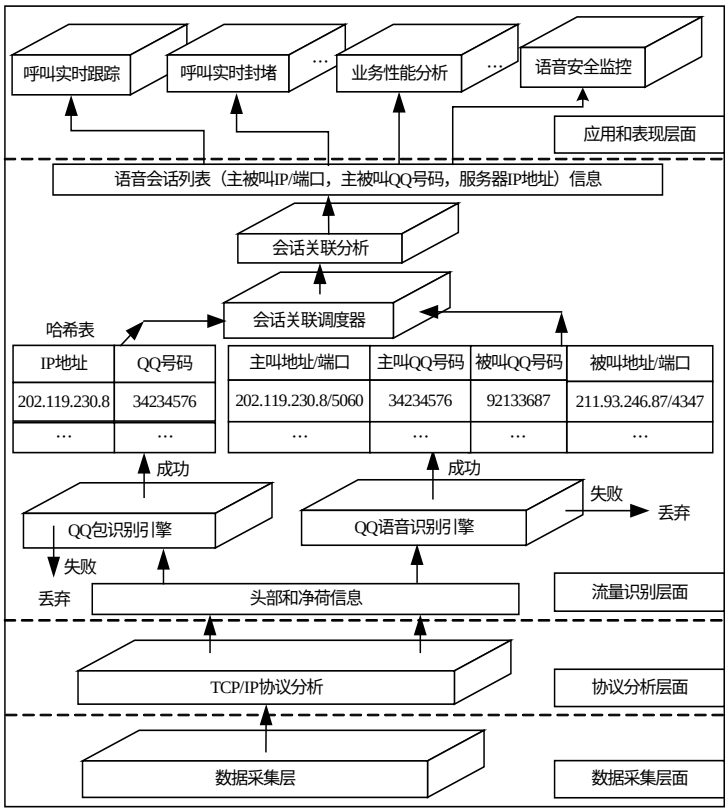

图 1　基于 ISA 的腾讯语音流量识别模型

分层模型如下：

(1)数据采集层

该层面提供针对不同链路的数据采集或复制技术，如 100/1000MFE、ATM、SDH 不同速率的采集或复制技术，以保障数据完整、可靠地传送至上一层面——协议分析层。

(2)协议分析层

该层面对数据进行 TCP/IP 协议解析，并且向上层提供足够的 IP 分组头部和 TCP/UDP 的头部信息及必要的分组净荷信息，以满足上一层面——流量识别层对业务的识别和感知。

(3)流量识别（业务感知）层

该层面是整个架构的核心层面，主要根据协议分析层提供的 IP 分组头部和 TCP/UDP 的头部信息及其净荷信息等特征有效识别出 QQ 业务，并丢弃匹配失败的分组。

为了能够将腾讯语音业务从整个腾讯会话业务中区分和识别出来，且保证识别的准确性，必须先根据一定的机制识别出语音业务，再通过其与整体会话的共性来验证识别准确性。因此，需要在初步识别出语音业务后通过关联算法与活动的 QQ 会话进行关联以验证其会话是否存在。该层主要包含 2 个算法：QQ 会话识别算法和 QQ 语音智能会话关联算法(intelligent session association, ISA)。此外，还须采用合适的机制保障识别算法灵活适应 QQ 业务特征的变化。

1)QQ 会话识别算法

算法处理过程（图 2）如下：

①初始化 QQ 会话哈希表。该哈希表用于存储 QQ 会话标识，即 QQ SessionID。该标识用 QQ 号码及其 IP 地址二元组来表示，因为一个 QQ 号码只能对应一个 IP 地址，所以从存储和查找的效率来看，用哈希表存储最合适。哈希表中所有的元素初始化为 0，即所有 QQ 号码对应的 IP 地址初始化为 0。

②接收分组。

③根据 QQ 会话净荷特征进行 DPI 检测，以判断该分组是否为 QQ 分组，再判断该分组是否为 QQ 会话的请求登录令牌分组，如是，获取 QQ 号码，转④；如匹配失败，丢弃分组，转③。

④判断该会话是否已经存在于哈希表中，如果是，丢弃分组，转④；如果不是，转⑤。

⑤保存 QQ 会话标识。Key 为 QQ 号码，value 为该 QQ 号码的登录 IP 地址。

⑥转②。

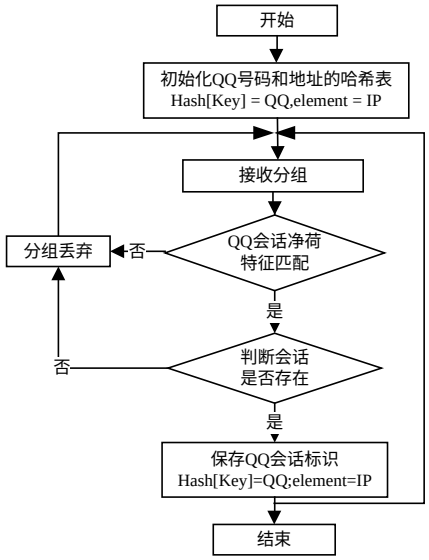

图 2　QQ 会话识别流程

2)QQ 语音业务识别和 ISA 算法

算法流程（图 3）如下：

①接收分组。接收过程同 QQ 会话是同一过程，只是同一分组复制之后用于不同之处。

②根据 QQ 语音净荷的类 SIP 特性进行净荷深度检测。如匹配成功，则转③；否则，丢弃分组，转①。

③将语音会话同 QQ 会话进行关联识别。由于仅通过 QQ 语音净荷的特征分析并无法确定该分组就是 QQ 语音会话分组，因此必须将该分组同已有的 QQ 会话进行关联检测，如

该 QQ 会话存在，则分组的判断很可能是正确的。具体的关联过程即用该语音分组中获取的 QQ 主叫号码作为 key，到 QQ 会话哈希表中查询，如查询出的元素为一个 IP 地址，那么该 QQ 会话存在，转④；如果查询出是 0，则该 QQ 会话不存在，丢弃分组，转①。

④保存和更新 QQ 语音会话信息。保存 QQ 语音会话的关键信息，如主被叫 QQ 地址和端口、主被叫 QQ 号码、语音编解码类型、呼叫发起时间、呼叫结束时间。当有其他语音会话的分组到来，相应地更新相关信息，形成 QQ 呼叫的详细记录 CDR。

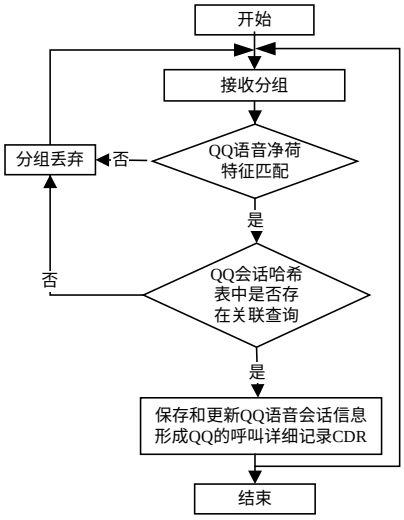

图 3　算法流程

3)ISA 算法的具体过程

在 QQ 语音业务的识别算法中，ISA 是个很重要的过程，该过程采用的关联元素，即关联会话标识为 Association Session ID=(主叫 IP 地址，主叫 QQ 号码)。采用会话调度器机制智能地确保 QQ 语音会话的识别准确率。调度器可以采用定时轮询和消息通知 2 种机制。定时轮询指调度器每隔一定时间查询是否有新增的 QQ 语音会话标识；消息通信机制是有新增的 QQ 语音会话时，智能地发消息通知调度器。前者效率较低、开销大、实时性不够，但实现简单；后者效率高、实时性和智能性好、实现较为复杂。通常采用第 2 种消息机制，以保证实时地进行真正语音的比较。

4)采用正则表达式动态更新 QQ 业务和语音业务特征库

QQ 版本的改动或者协议的改动均会带来 QQ 净荷特征的变化，因此，上述识别算法必然会发生一定的变化。如何不改动系统而通过简单的配置就完成对 QQ 新业务特征的适应是算法的一大挑战。正则表达式是一个非常好的解决方案。本算法采用正则表达式来表现 QQ 的会话特征和语音会话的特征，因此，当 QQ 版本或者特征发生变化，本算法只需要简单地修改正则表达式的特征配置文件，无须重新修改代码和算法，更新快速高效。

综上所述，本算法能够保证系统的可行性，具有较高的 QQ 识别及 QQ 语音识别的准确率。

(4) QQ 话音业务应用层以及表现层

对 QQ 语音业务的识别可以应用在诸如 QQ 语音业务性能分析、QQ 语音流量控制和呼叫跟踪、QQ 资费影响因子估算等方面。

3　算法验证

腾讯语音呼叫的识别准确率的计算包含如下 2 个因素：识别次数(I)和试呼次数(C)。识别准确率用Δ来表示。

$$\text{识别准确率}\Delta=(\text{识别次数}\ I\ /\text{试呼次数}\ C\ )\times 100\%$$

其中，识别次数指识别出该会话为腾讯 QQ 语音呼叫业务的次数；试呼次数表示腾讯 QQ 语音呼叫的实际次数，$C > 100$ 时统计结果才较有意义，否则样本空间中的样本点过少，导致结果失去统计意义。测试可以采用被动采集并同其他相关系统作对比和主动拨打测试 2 种方法。

根据本算法开发出的腾讯语音呼叫业务检测系统在中国电信广西分公司的 10G 骨干网上得到了具体的验证。系统采用分光方式将 10G 流量负载均衡分流至若干台业务识别处理机上，业务识别处理机完成核心算法的实现，从纷繁复杂的分组中提取、分析、识别和关联出 QQ 的语音会话。通过在骨干网的实际运行和拨打测试，针对 QQ 语音业务的识别准确率均在 90%以上，很好地体现了算法的实施效果，验证了算法的准确性，如表 1。

表 1　识别准确率对比表

骨干网采集点	呼叫发起次数/次	基于 QQ 的 8000 端口和 IP 地址识别准确率/%	基于智能会话关联技术识别方法准确率/%
南宁	887（被动采集）	72.3	92.1
南宁	134（主动拨测）	77.8	94.5
柳州	647	74.8	94.4
柳州	119	76.9	93.8

4　结束语

本文介绍了基于 DPI 和 ISA 的 QQ 语音业务识别系统模型，主要阐述了基于 DPI 的 QQ 会话识别算法和语音会话关联识别算法。本文的模型和算法具有良好的可扩展性和准确性，易于与运营商相关的应用接口对接，通过各种应用开发，方便电信运营商对 VoIP 业务进行监管。本文所有的算法和模型均在实际的 IP 骨干网上得到了验证。

参考文献

1 Fathi H, Chakraborty S S, Prasad R. On SIP Session Setup Delay for VoIP Services over Correlated Fading Channels[J]. IEEE Transactions on Vehicular Technology, 2006, 55(1).

2 Levy H, Zlatokrilov H. The Effect of Packet Dispersion on Voice Applications in IP Networks[J]. IEEE/ACM Transactions on Networking, 2006, 14(2).

3 Gao Lisha, Luo Junzhou. Performance Analysis of a P2P-based VoIP Software[C]//Proc. of the Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services. 2006.

4 Cao Feng, Bryan D A, Lowekamp B B. Providing Secure Services in Peer-to-peer Communications Networks with Central Security Servers[C]//Proceedings of the Advanced Int'l Conference on Telecommunications and Int'l Conference on Internet and Web Applications and Services. 2006.

5 万　敏, 万晓榆. 基于 SIP 的 VoIP 在下一代网络中的应用[J]. 重庆邮电大学学报(自然科学版), 2003, 15(4).