1. 绪论
   1. 研究背景

现状

重要性

* 1. 研究内容

内容与意义

困难

* 1. 研究方法
     1. 数据采集

被动监测

监测方法（参考李为民论文、董超论文、桂小林论文、郭敏杰论文有固网）

实验团队开发的互联网流量在线报文解析系统

主动查询

爬虫（郭敏杰论文）

* + 1. 数据处理

必要性（林文辉论文）

Hadoop（参考董超论文、郭敏杰论文介绍、乔媛媛论文）

Hadoop是由xx的xx。其核心由三个子项目组成： HDFS（Hadoop Distributed File System）[49]、MapReduce和Hadoop Common。

hdfs、MapReduce

* + 1. 数据分析

分布拟合

常见的重尾分布

PP图、QQ图

机器学习

分类

聚类

回归

* 1. 主要创新点
  2. 论文结构

1. 网络视频业务体系概要性研究
   1. 网络视频业务架构解析

示意图：并无统一标准。我们经过分析，工业界中各sp的业务体系架构的实现方式类似，如图

播放方式

接入方式

网络架构

文件分片、渐进式下载、动态分辨率调整

HTTP：Currently the mainstream technology of Internet video is web video, which delivers videos over the HTTP protocol and plays videos via web browsers [2] [3]. Advantages of such resolution include: reusing the existing web servers and software, utilizing web caches, seamlessly traversing firewalls, and requiring no client programs [4].

2. Begen A C, Akgul T, and Baugher M. Watching video over the web: Part 1: Streaming protocols. Internet Computing, 15(2):54–63, 2011

3. Erman J, Gerber A, Ramadrishnan K, Sen S, and Spatscheck O. Over the top video: the gorilla in cellular networks. In Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, 127–136, 2011

4. Summers J, Brecht T, Eager D, and Wong B. To chunk or not to chunk: Implications for http streaming video server performance. In Proceedings of the 22nd international workshop on Network and Operating System Support for Digital Audio and Video, 15–20, 2012

* 1. 用户-服务器通信流程

用户、网络、资源

不同行为的时序图

调度方式

调度通信特点

用户主动操作触发指标

* 1. 网络质量分析

报文分析：播放器自动上报触发

特征及检测方法（参考李为民论文）

1. 网络视频业务分发服务器检测
   1. 概述

近年来，网络视频业务发展迅猛。尤其从流量角度来看，网络视频业务的流量已成为了互联网流量的主要组成部分[1]。为了更好的服务在地理上广泛分布的用户群体，大多数网络视频业务提供商（视频网站）都使用了大规模的内容分发网络（Content Delivery Network，CDN）来支撑视频文件的传输。出于扩展性、安全性及便于管理等方面的原因，大多数的网络视频业务提供商都选择去建设自己的专属视频分发网络（Video Delivery Network），而非直接使用第三方的商用CDN。正如前文所分析，从功能上来看，这些专属的视频分发网络主要包含两种关键的功能性服务器：调度服务器（Dispatch Server，DS）与资源服务器（Resource Server，RS）。调度服务器接收用户的视频请求，并根据用户的地理位置将其引导至最合适的、往往也是地理位置上最近的资源服务器上；而资源服务器广泛部署于各个地点，用于存储视频文件并响应下载请求[2]。在本文中，我们统称这两种服务器为网络视频业务的分发服务器。

这些分发服务器往往产生大量的互联网流量，并占据较大比例的网络传输带宽。因此，对于网络运营商而言，视频分发服务器的信息在一系列网络管理、控制、优化的任务中是至关重要的。一个典型的例子是对过顶（Over The Top，OTT）业务流量的管控问题：在网络视频业务中，网络运营商并不是业务的提供商，只负责为业务流量提供传输管道。对于该业务产生的海量网络流量，运营商无法从业务逻辑上调整，以进行合理的网络资源管理与分配。此时，如果能够获取网络视频业务的分发服务器信息，就可以通过对服务器吞吐流量的管控，来实现对网络视频业务的管控，如动态带宽调整、多种服务质量（Quality of Service，QoS）提供、指定视频内容屏蔽等。此外，视频分发服务器的信息还有助于网络运营商减少非必要的开销，例如网间结算（Inter-Network Accounting）：有些移动运营商没有自己的互联网接入入口，当他们的手机用户使用互联网业务时，所产生的网络数据会被转发至一个具有合作关系的固网互联网提供商（Internet Service Provider，ISP）处。这些数据按流量大小计费，由移动运营商支付给固网ISP，被称为网间结算，如图3-1所示。由于相较于文本或图片文件，视频文件往往较大，故网络视频业务产生的流量往往远大于其它业务的流量，进而造成较高的网间结算费用。在此情形下，移动运营商往往会迫切的需要网络视频业务分发服务器的信息，以进行相应措施（如部署网内专用缓存）来降低费用。

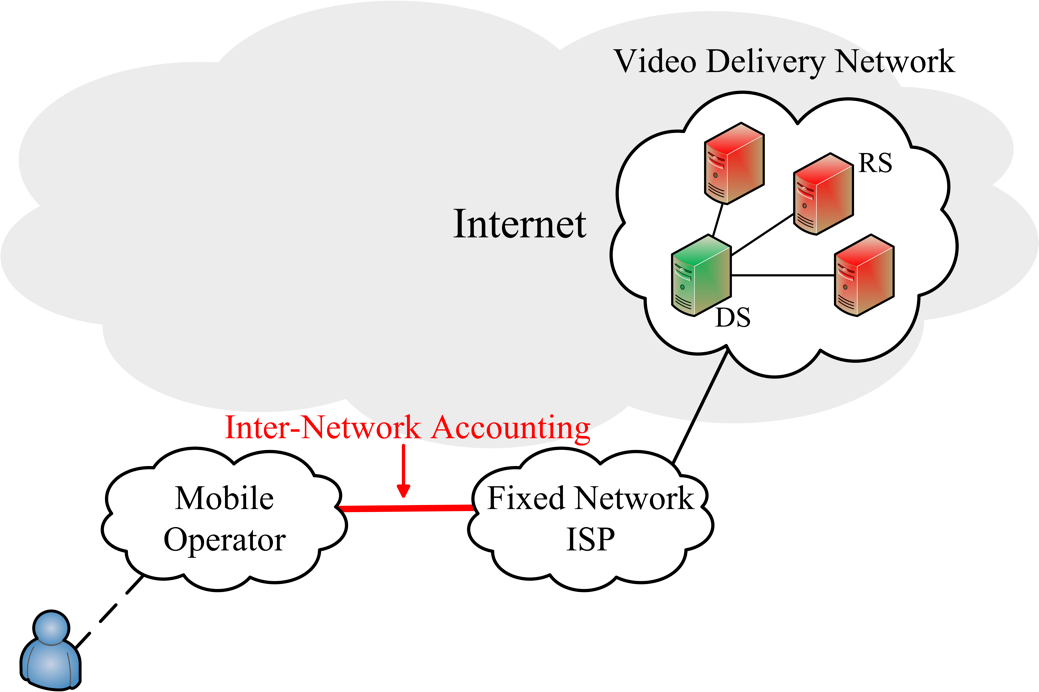


图3-1

然而，就算对于网络运营商而言，从互联网中众多的服务器中定位出各个网络视频业务提供商的分发服务器也并非易事。在视频分发网络中，分发服务器往往具有以下特点：1）大多数调度服务器与几乎所有的资源服务器并没有域名；2）这些服务器的IP地址不固定，会经常改变；3）新增的分发服务器随时可能被添加至视频分发网络中；现存的分发服务器也随时可能从视频分发网络中删除。这些特点使得分发网络可以进行灵活扩展以提升性能，但也对分发服务器的检测造成了较大的困难。例如，试图向各个业务提供商进行询问，使用简单的白名单匹配的方法，就无法适用于无域名、不固定IP地址、可动态增减的网络视频业务分发服务器的检测。

为解决此问题，在本章中，我们对网络视频业务中用户与服务器之间的通信进行详细的分析。基于分析结果，我们提出了一个高效的分发服务器检测方法，并使用真实的数据进行了验证。本章研究内容的主要贡献与创新点在于：1）新颖的研究问题。我们关注于网络视频业务分发服务器的检测。据我们所知，对这一具有重要实际意义的问题的研究尚属首次。2）深入的专项分析。我们基于对播测报文的分析，总结出了国内主流网络视频业务通用的用户-服务器通信流程，并进一步定义了若干衡量指标以揭示这些通信行为的特点。3）高效的检测方法。我们基于通信特征与机器学习算法，提出了网络视频业务分发服务器检测系统。该系统具有通用性，对已知和未知的网络视频业务提供商均可适用。实验结果显示该系统性能优秀：准确率可接近100%，同时召回率在85%以上。

* 1. 相关研究现状

对于网络视频业务中视频分发网络结构和分发策略的研究，学术界目前已有了一些工作。在文献[3]中，作者Saxena等人关注于三家不同网络视频业务提供商的分发网络，研究了其中服务器部署策略，并对比了各网络的服务质量。作者Adhikari等人使用被动测量技术[4]与主动测量技术[5][6]，对YouTube视频分发网络中的服务器位置与视频分发策略进行了探索。在文献[7]中，作者Torres等人在不同国家收集了YouTube的CDN流量，并对其网络结构与服务器选择策略进行了分析。在文献[8]中，作者Plissonneau等人研究了YouTube在不同ISP网络和不同国家中的视频分发流程以及对用户体验的影响。需要注意的是，现有的这些对网络视频业务分发网络的研究，大多数是面向业务提供商YouTube的。YouTube的分发服务器具有一组固定格式的域名，研究者可以直接从互联网流量中根据域名过滤来定位分发服务器。然而，正如上节中所提到的，对于其它的（尤其是国内主流的）网络视频业务提供商，其分发服务器可能并没有固定的域名或IP。当研究这些视频分发网络的结构和流量特性时，如何正确识别其分发服务器将会成为一个问题。

鉴于网络视频业务分发服务器检测这一问题的独特性与新颖性，据我们所知，目前还未有专门的研究工作提出过相应的检测方法。但是对于其它互联网业务的流量或服务器识别，学术界已有了一定的研究成果。例如，在文献[9]中，作者Korczynski等人提出了一个基于协议和报文分析的三阶段的混合分类方法，来对网络语音Skype业务的SSL加密流量进行检测。该方法的准确率及召回率可接近90%。在文献[10]中，作者Chu等人从网站URL的文本和域名结构方面提取了若干有效特征，并使用SVM分类器来检测针对热门网站的恶意钓鱼服务器。此方法的检测率可达98%，同时误判率在0.64%以下。作者Chaudhary等人在文献[11]中，基于语言特征、时间维度特征及流行度特征，提出了一个针对YouTube视频垃圾回复的检测方法。该方法在特定类别上的准确率可超过80%。上述这些研究中提出的方法，在解决各自的特定问题中都是十分有效的。然而，对于网络视频业务分发服务器检测这一问题，由于应用场景及可提取特征的不同，这些方法都已不再适用。如何针对我们的研究问题，提取合适特征并设计检测方法，仍有待于进一步的分析与解决。

* 1. 数据集

本章中的研究数据来自于某互联网提供商在我国东南某省的固定网络中。我们将实验室团队开发的网络流量采集器部署于该省际网络的出口处，如图3-2所示。传入、传出该省际网络的流量将会被实时镜像，传送到采集器中。如xx.xx节所介绍，采集器对这些报文流量进行高速解析，并生成话单。在本章的研究中，我们进一步过滤出了HTTP协议的报文数据，并为用户与服务器之间每一次的HTTP交互进行请求与应答匹配，最终生成HTTP记录。图3-3给出了数据集整理的具体流程示意。最终每条HTTP记录对应一次用户与服务器的HTTP交互，字段内容包括：时间戳、服务器IP地址、用户IP地址、HTTP请求方法、HTTP请求URL、HTTP应答状态码、HTTP请求头部字段、HTTP应答头部字段以及HTTP文本类型应答内容的前1000字节数据。

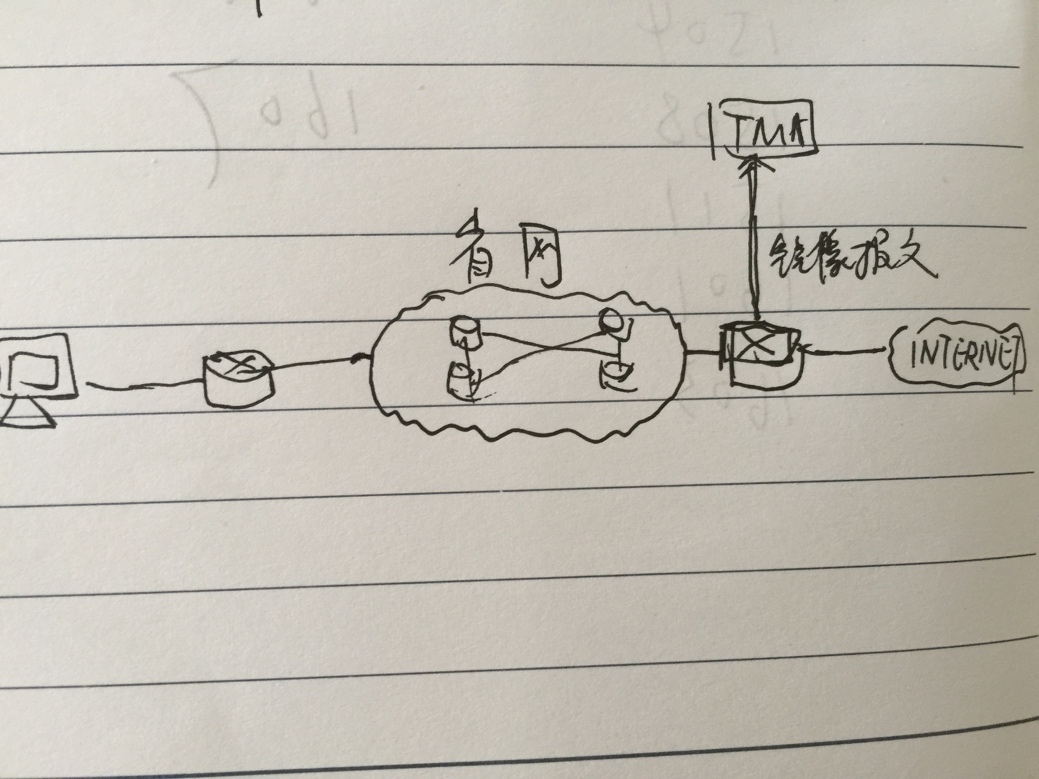


图3-2 数据采集部署示意

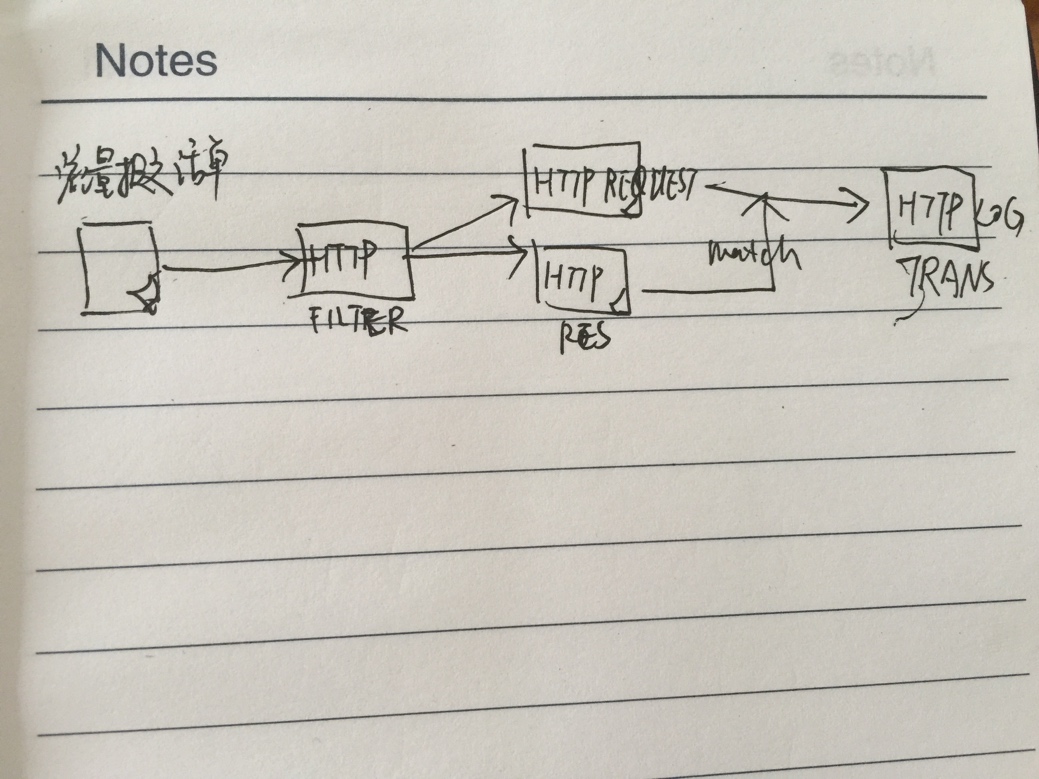


图3-3

整个数据采集阶段从2014年3月11日起至2014年3月15日止，共计5天。最终，我们一共采集到5,504,771条HTTP记录，覆盖3,027个用户与35,105台服务器。在本章的研究中，我们主要关注于5家目前国内最热门的网络视频业务提供商：1）Youku优酷视频，2）Sohu搜狐视频，3）Iqiyi爱奇艺，4）QQ腾讯视频，以及5）Letv乐视。

为了获取真实的（ground truth）分发服务器类型，我们对这5家业务提供商的视频分发过程流量进行了过滤与分析。具体来讲，我们首先将HTTP记录按用户汇聚，并按时间排序。对于每个用户的记录队列，我们检查每一条记录的“HTTP请求URL”字段。如前文X.X所分析，各网络视频业务提供商的视频页面具有固定格式的URL。在我们的检查中，如果某条记录的URL匹配某一业务提供商的视频页面格式，则说明该用户开始使用该业务提供商的网络视频业务，即播放了一个视频。我们将这条记录及（该用户）后续的记录过滤出来。这些记录包含着至少一次的视频分发过程。然后，我们仔细分析这些记录中的HTTP应答头部字段及其文本内容的前1000字节数据。如果某条记录的HTTP应答头或文本内容中包含着一个视频URL，并且后续某条记录显示用户根据该视频URL下载视频文件，则我们认为第一条记录对应的服务器为一个调度服务器，而第二条记录对应的服务器为一个资源服务器。注意，这种分析方法无法归纳成一个检测方法，因为视频URL可能出现的位置、形式、格式是未知的，并且会随着网络视频业务提供商对服务器的配置而变化。在我们的研究中，对数据集中HTTP记录应答内容是否包含视频URL的判断由人工进行，并需要一定的领域知识。最终，我们成功对各网络视频业务提供商的分发服务器给出类型标签，具体统计如表3-1所示。

表3-1 数据集概要统计

|  |  |  |  |
| --- | --- | --- | --- |
| **业务提供商** | **用户数** | **调度服务器数** | **资源服务器数** |
| Youku | 1,991 | 21 | 737 |
| Sohu | 1,093 | 8 | 119 |
| Iqiyi | 1,295 | 19 | 46 |
| QQ | 1,338 | 3 | 321 |
| Letv | 1,367 | 5 | 408 |

* 1. 用户-服务器通信分析
     1. 通信流程

我们首先研究用户在使用网络视频业务时，是如何与服务器进行信息交互的。在工业界中，目前尚未有一个公认的明确标准来规定用户与网络视频业务服务器之间的通信流程。为了探寻不同业务提供商各自的自定义通信流程实现，我们使用实验环境进行播测分析，即：使用实验主机主动访问各视频网站的视频页面，播放视频的同时在实验主机上捕获传输报文，最后对报文队列进行分析。我们发现，对于不同对网络视频业务提供商，其用户与服务器之间的主要交互过程是非常相似的，可以用一个通用流程来概括，如图3-4所示。

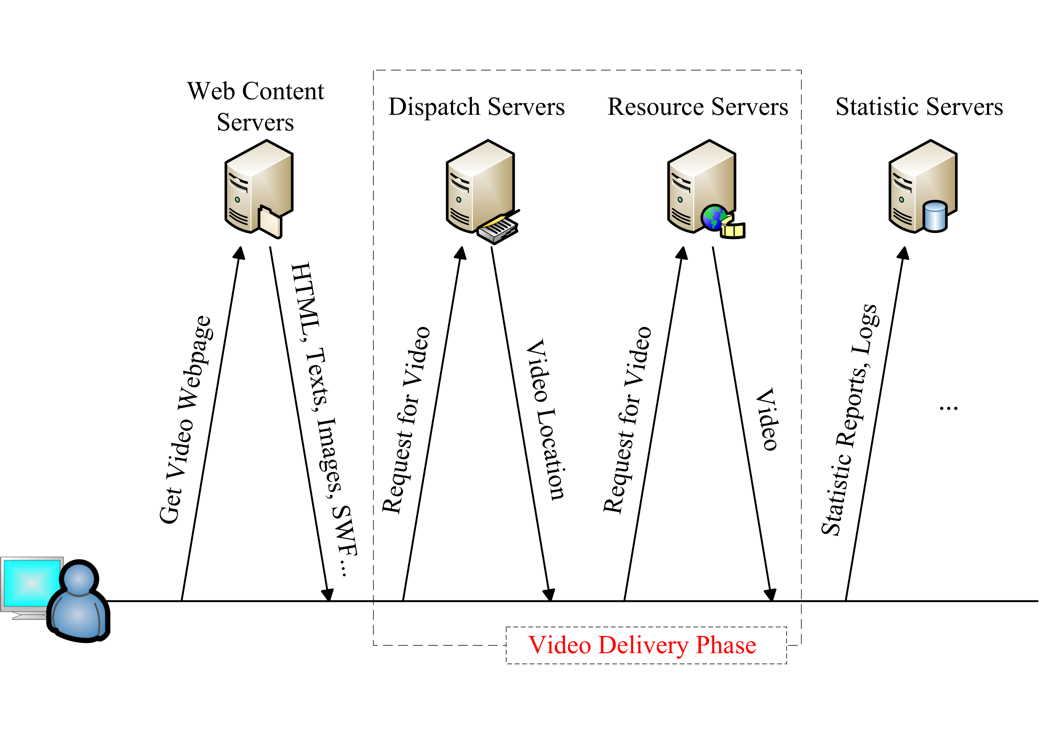


图3-4 通用的用户-服务器通信流程

1）首先，当用户打开一个网络视频时，会连接该网络视频业务提供商的内容服务器，下载视频页面的HTML文件，以及HTML的内嵌内容，如文字、图片、脚本等。值得注意的是，一个通常以SWF格式嵌入网页的视频播放器将会被下载。该播放器可运行于Adobe Flash Player插件中，为用户提供一个图形化界面，用于视频获取、播放控制、及信息上报等。

2）然后，视频播放器向调度服务器自动发送视频请求。调度服务器根据用户的IP地址，回复最合适的视频资源地址。通常，调度服务器会回复距离用户地理位置最近的资源服务器上的视频地址。但当网络环境较差时，其它较远资源服务器上视频地址也可能会被回复，以进行负载均衡。另外，经分析我们发现调度服务器的回复方式主要有两种：通过HTTP重定向和通过HTTP内容实体。前者十分简便，而后者可以在恢复中添加额外的信息。业务提供商可能会同时使用这两种方式以提高性能。

3）接下来，视频播放器自动连接资源服务器，下载视频文件。一旦播放器的缓存中有了足够多的数据，视频将自动开始播放。步骤2）和3）合称为视频分发阶段。由于大的视频文件可能会被分片存储，步骤2）和步骤3）在一次视频播放过程中可能会出现多次，以获取并下载同一视频文件的不同分片。

4）伴随着视频文件的下载与播放，视频播放器还会自动的向业务提供商的统计服务器上报反馈信息，例如用户操作记录、播放进度心跳、网络状况统计等。

另外，我们发现用户与网络视频业务的服务器之间的通行，全部都是基于HTTP协议的。这也是目前主流的网络视频业务与传统的基于流媒体的视频业务的区别所在。在我们的分析工作中，我们关注于用户与视频服务器的HTTP交互，即每次通信的HTTP请求应答对（request-response pair）。

* + 1. 通信特性

我们进一步对视频分发阶段中用户与分发服务器之间的HTTP请求-应答交互报文进行了分析，并发现了若干特性。这些特性有助于将网络视频业务的分发服务器从互联网海量的服务器中区分出来。

**较小的报文数与时间间隔**：我们将视频分发阶段用户与分发服务器之间的HTTP请求应答对（request-response pair）的序列表示为：

（3-1）

其中，表示用户与调度服务器之间的HTTP交互；表示用户与资源服务器之间的第一个HTTP交互；表示与之间与视频分发无关的用户-服务器HTTP交互。的产生是十分复杂与随机的。例如，对视频页面HTML内嵌内容的下载在发生时没有完成；或用户那那段时间内使用了其他的互联网业务。

我们定义为与之间的报文数，为与之间的时间间隔。我们对数据集中各视频分发过程的与进行了计算与统计，并发现这两个指标都比较小。图3-5与图3-6显示了我们数据集中不同网络视频业务提供商对应的与的累积分布函数（cumulative distribution function，CDF）。如图所示，无论哪个业务提供商，都有超过80%的都小于10；同时近80%的都小于1秒。这一现象是符合我们预期的：在用户-服务器通信流程中，当视频播放器收到了从调度服务器发来的回复时，会立即向资源服务器发起视频下载请求。所以与总是接连出现的。这一特性在对分发服务器进行检测识别时是非常有用的：如果我们能够定位用户与资源服务器的通信（即），则用户与调度服务器的通信（即）往往就在其前方不远处。

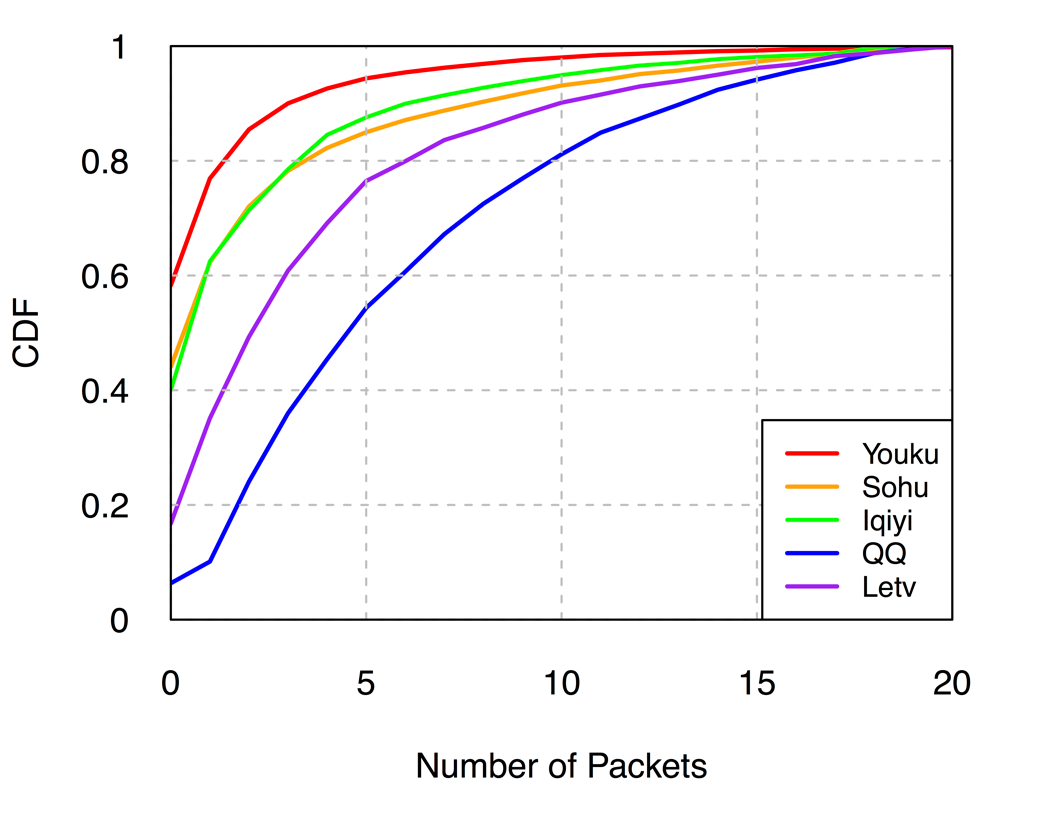


图3-5 不同网络视频业务提供商的的累积分布函数

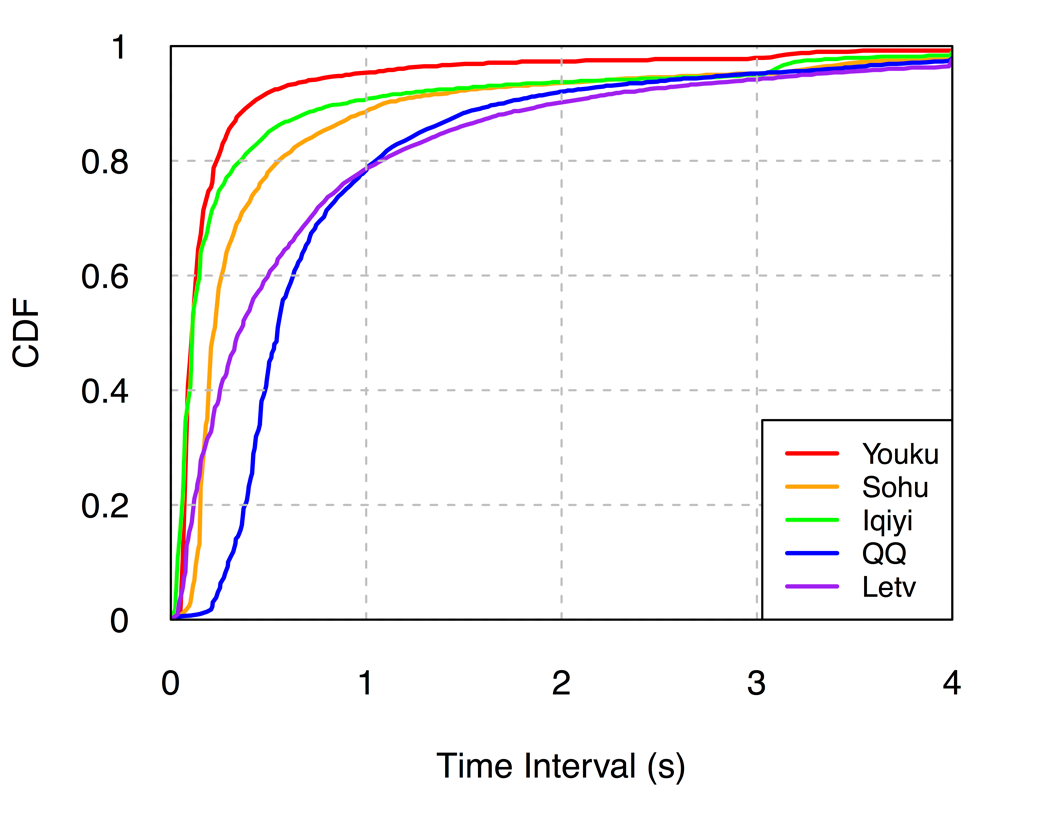


图3-6 不同网络视频业务提供商的的累积分布函数

**特定的实体内容类型**：在网络视频业务提供商建造的专属视频分发网络中，调度服务器仅用于向用户发送视频地址URL信息，而资源服务器仅用于向用户提供视频。因此，不同于互联网中常见的缓存服务器，其需要存储各种各样文件（如文本、图片、脚本、音频、视频等）来传送给用户，网络视频业务的分发服务器仅支持有限的几种文件类型。

图3-7和图3-8显示了我们数据集中各个网络视频业务提供商的调度服务器与资源服务器向用户传输的HTTP实体内容类型分布。从图中可以看出，调度服务器向用户发送的主要是“text/xxx”的文本类型内容，用以承载最合适的资源服务器上视频地址的动态信息。其中，”text/json”对应的JSON和“text/xml”对应的XML是目前互联网中较为常见的动态信息传输技术（如Ajax）所使用的文件类型。对于有些网络视频业务提供商，会使用自定义的纯文本格式，从服务器向视频播放器发送数据。这些纯文本文件相应的HTTP实体内容类型往往是“text/html”或“text/plain”。而对于资源服务器来说，其传输的HTTP实体内容类型主要是“video/flv”和“video/mp4”，对应着FLV与MP4两种文件格式。这两种文件格式是目前主流的网络视频封装格式，一般来说FLV对应着标准清晰度（standard definition，SD）视频，而MP4对应着高清晰度（high definition，HD）视频。

网络视频业务分发服务器这种仅传输有限且特定类型文件等特性，有助于将其与互联网中其他的服务器区分出来。

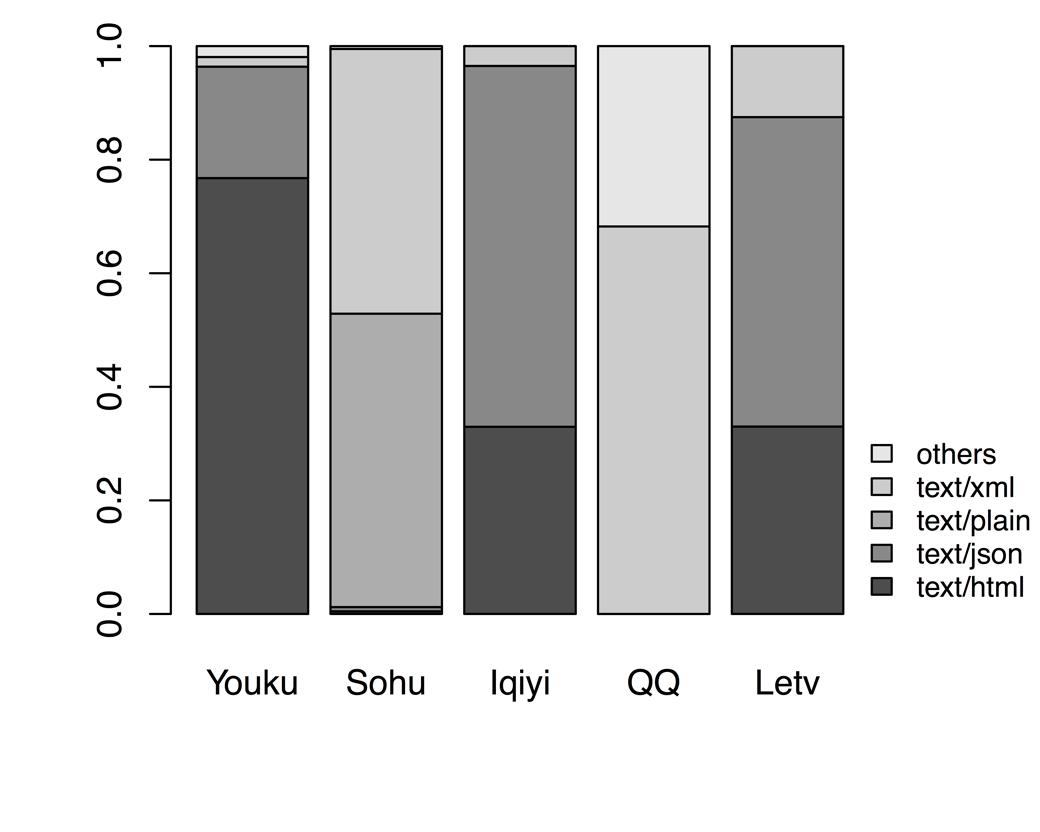


图3-7 不同网络视频业务提供商的调度服务器传输的HTTP实体内容类型分布

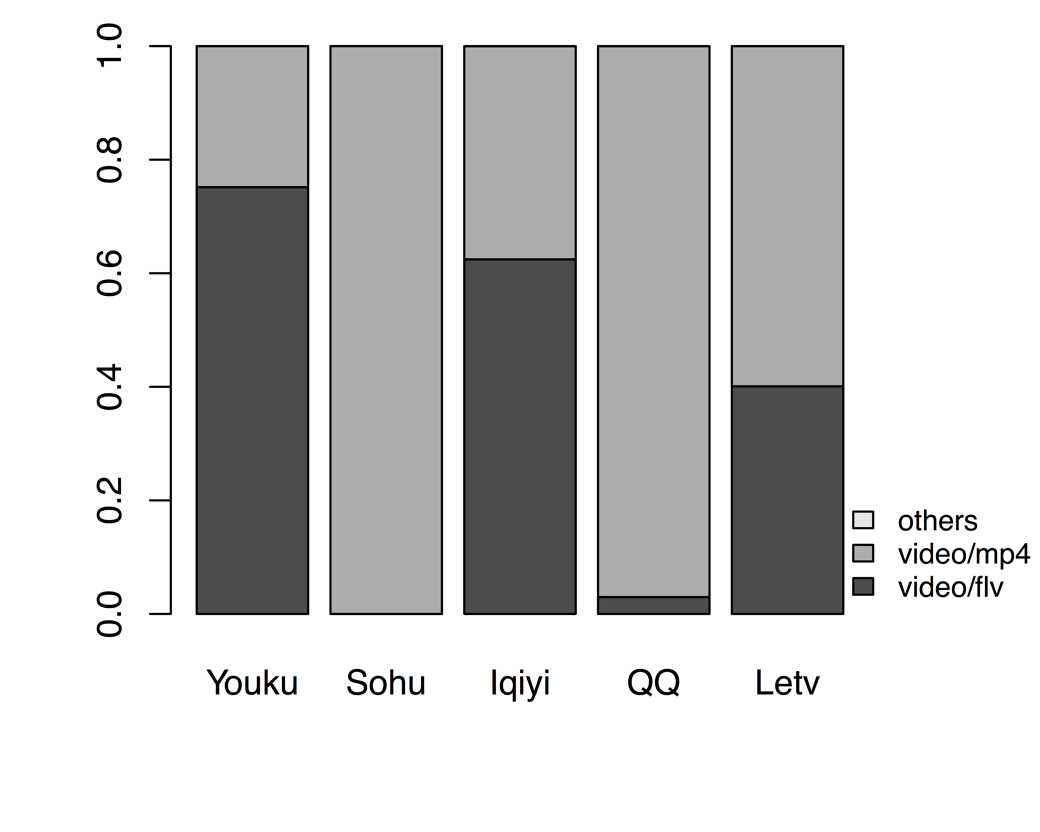


图3-8 不同网络视频业务提供商的资源服务器传输的HTTP实体内容类型分布

**HTTP重定向行为**：HTTP协议允许服务器将用户的请求重定向到另一个服务器上。此时，HTTP应答报文的状态码取值在300～399范围内，并且重定向的目的URL由应答报文的“location”头给出。用户侧的浏览器收到一个HTTP重定向应答时，会自动向“location”指定的重定向URL重新发送HTTP请求。本文中，我们定义来表示HTTP重定向行为是否存在于两个HTTP交互中：

(3-2)

其中，和两个HTTP交互，在之前；表示中HTTP应答状态码；表示中HTTP应答“location”头的值；表示中HTTP请求的URL。

在视频分发阶段，有些网络视频业务提供商的调度服务器使用HTTP重定向来将用户引领至资源服务器。具体来讲，调度服务器返回一个状态码以3开头的HTTP重定向报文，并将“location”头部赋值为最合适的资源服务器对应的视频地址URL。表3-X给出了优酷视频的一对“用户请求-调度服务器重定向”HTTP交互报文例子。由于不需要向视频播放器中嵌入任何脚本或代码，直接使用HTTP重定向可能是实现视频分发调度最简单的方式。

考虑到大多数互联网服务器是用来向用户提供文件内容或传输数据的，HTTP重定向对于一般的服务器来说是一个不常见行为。因此，如果一个服务器频繁的将用户的（视频）请求重定向到其它的服务器上，那么这个服务器很有可能就是网路视频业务中的调度服务器。

表3-X 优酷视频中用户进行视频请求及调度服务器重定向实例

|  |  |  |
| --- | --- | --- |
| GET /player/getFlvPath/sid/139830840964516453625\_00/st/flv/fileid/0300020100533AA654FCF1003E880381465D99-B4F1-45C1-560C-3067B764ABF6 HTTP/1.1 |  | HTTP/1.1 302 Found |
| Accept: \*/\*  Accept-Language: zh-CN  Referer: http://static.youku.com/v1.0.0426/v/swf/player.swf  x-flash-version: 12,0,0,70  Accept-Encoding: gzip, deflate  User-Agent: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)  Host: k.youku.com  Connection: Keep-Alive  Cookie: \_\_ysuid=13956529867728JC; xreferrer=; ykss=738258531c81dd4b6fec24c5; u=\_\_LOGOUT\_\_; P\_F=1; P\_T=1398316696 |  | X-Powered-By: PHP/5.3.3  Expires: -1  Cache-Control: private, max-age=0  Pragma: no-cache  Location: http://118.228.18.36/youku/67721102DBE3482E9DE1942F42/0300020100533AA654FCF1003E880381465D99-B4F1-45C1-560C-3067B764ABF6.flv  Content-type: text/html  Content-Length: 0  Connection: close  Date: Thu, 24 Apr 2014 03:18:17 GMT  Server: F\_LIGHTY\_BJ\_EDU02 |

**相似的HTTP请求URI**：在HTTP协议中，URL由主机域名和URI两部分构成。例如，“www.example.com/logo.gif”这个URL中，“www.example.com”是主机域名，而“/logo.gif”是URI。在网络视频业务中，我们发现对于很多业务提供商，在其用户向调度服务器和资源服务器发送的两个HTTP请求报文的 URI中，存在着一组很长的公共的子字符串。通常，这部分公共子字符串是一个文件路径，包含一个多级的目录和一个较长的文件名，例如：“/videos/comic/20130717/c67234f99b3fe2011a373bcf77593403.flv”。在某些实例中，这两个请求报文的URI甚至是完全一样的。我们经分析得出，造成此现象的原因可能是该网络视频业务提供商在调度服务器与资源服务器上使用的相同的文件目录结构，来散列文件地址信息与保存视频文件。为了衡量这一特性，我们定义两个HTTP请求之间的URI相似度，如下：

(3-3)

其中，和分别表示两个HTTP请求应答交互，在之前；和分别表示和中HTTP请求URI；表示的字符串长度，而表示和中包含的最大公共子字符串的长度。举例来讲，对于给定两个URI： = “/dir123/video123”， = “/dirabc/video123”。其最长的公共子字符串是“/video123”，则，最终两个URI的相似度为。

表3-X给出了网络视频业务提供商优酷、搜狐和爱奇艺中，用户向调度服务器、资源服务器发送的HTTP请求URI实例及相似度。可以看出，URI相似度都在60%以上。相似的HTTP请求URI这一特性，在资源服务器与其它的非调度服务器之间是很少存在的。因此，如果某个服务器与资源服务器的URI相似度非常高，该服务器很可能就是一个调度服务器。

表3-X

|  |  |  |  |
| --- | --- | --- | --- |
| **网络视频**  **业务提供商** | **交互**  **服务器** | **URI** | **URI**  **相似度** |
| **优酷** |  | /player/getFlvPath/sid/939830949540410dfaa10\_00/st/flv/fileid/03000201005357C9FDED2D14AB15D13259D4ED-BEC6-33D2-BD8F-461A6050F9D5 | 64% |
|  | /youku/677471A88734D84577DB932778/03000201005357C9FDED2D14AB15D13259D4ED-BEC6-33D2-BD8F-461A6050F9D5.flv |
| **搜狐** |  | /prot=2&file=/tv/20140416/1719506-260cd73f-dec6-418b-82c7-226ba1f86497.mp4&new=/142/209/uQVYly2MekYkbwJH2stZ24.mp4 | 81% |
|  | /sohu/8/142/209/uQVYly2MekYkbwJH2stZ24.mp4 |
| **爱奇艺** |  | /videos/other/20140411/a9/3a/b1/b7edd2addbecbc60205edef6f7657929.f4v?pv=0.1 | 100% |
|  | /videos/other/20140411/a9/3a/b1/b7edd2addbecbc60205edef6f7657929.f4v?pv=0.1 |

**同一SWF的referer头**：在HTTP协议中，请求报文中的“referer”头部指定该HTTP请求是从哪里产生的。通常情形下，HTTP请求由网页的HTML文件或JavaScript脚本触发，进行内容下载或信息提交。因此，互联网HTTP流量报文中大多的“referer”头所对应的值往往以“.html”或“.js”结尾，例如“Referer: http://v.youku.com/v\_show/id\_XNzAyNzQ5NjM2.html”。

然而，如3.4.1小节所分析，在网络视频分发阶段，发往调度服务器与资源服务器的HTTP请求是由视频页面中内嵌的SWF格式的视频播放器自动生成的。这使得这些HTTP请求中的“referer”协议头的值指向一个SWF格式的播放器，例如“Referer: http://static.youku.com/v1.0.0426/v/swf/player.swf”，而非一个HTML或JavaScript文件。并且，由于这些请求是由同一个SWF格式播放器产生的，其“referer”头部的值是完全相同的。本文中，我们定义来表示两个HTTP交互的请求中是否具有相同的以SWF为后缀的“referer”头：

(3-4)

其中，和两个HTTP交互，在之前；与分别表示与中HTTP请求的“referer”头。

一般而言，对于其它没有用到Flash技术的互联网业务的HTTP请求，以SWF结尾的“referer”头是极少出现的。因此，发往视频分发服务器的HTTP请求包含相同的以SWF结尾的“refer”头这一特性，有助于将分发服务器从其它的服务器中区分出来。

* 1. 系统架构

基于以上分析结果，我们定义了一系列有区分度的特征，并提出了一个网络视频业务分发服务器检测系统。该系统由3个关键模块组成：预处理（pre-processing）、预选择（candidate selection）与综合判决（decision）。整体检测系统架构如图3-9所示。下文将对各系统模块进行具体的介绍。

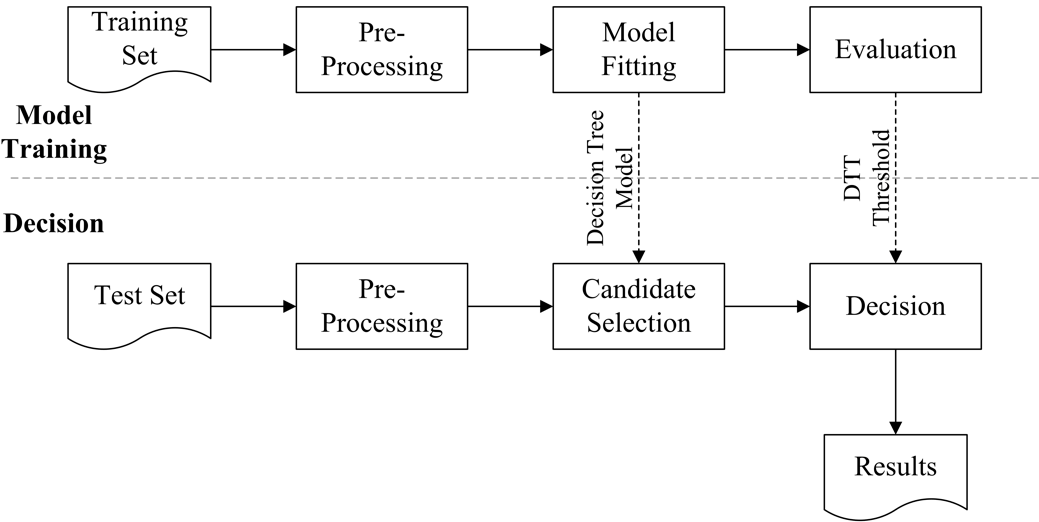


图3-9 网络视频业务分发服务器检测系统整体架构

* + 1. 预处理

在预处理模块中，我们尝试为每一次的视频分发过程提取出一组包含分发服务器的HTTP记录。首先，我们将HTTP记录按用户汇聚。对于每一个用户的HTTP记录，我们按请求时间对记录进行排序。然后，我们对各用户的记录序列，逐条检查HTTP头部“content-type”的值。如果某条记录的“content-type”值为“video/flv”或“video/mp4”，根据3.4.2中的“特定的实体内容类型”特性，该记录对应的服务器很可能是网络视频业务中的调度服务器，我们将其定义为疑似资源服务器。又根据3.4.2中分析出的“较小的报文数与时间间隔”特性，我们可以推断本次视频分发过程中，调度服务器对应的记录应该就在这条疑似资源服务器对应的记录前面不远处。

因此，我们将疑似资源服务器对应的记录与其之前的10条记录提取出来，共计11条HTTP记录，做为一个HTTP记录组。最终，预处理模块的输出是一系列的HTTP记录组，每个记录组都包含着一条疑似资源服务器对应的记录和该记录之前的10条记录。

* + 1. 预选择

在前小节预处理模块中，我们过滤出了疑似资源服务器；而在本小节预选择模块中，我们尝试在预处理给出的HTTP记录组中定位可能的调度服务器。基于3.4节中对用户与服务器在视频分发阶段的通信流程与通信特点的分析，我们定义了9个特征来用于调度服务器的检测。具体的特征列表如表3-2所示。

表3-2 疑似调度服务器检测特征概要

|  |  |  |
| --- | --- | --- |
| **序号** | **特征** | **说明** |
| 1 | Request Method | HTTP交互中请求报文的请求方法字段。常见的取值包括：GET，POST，HEAD等。 |
| 2 | HTTP Redirection | 本次HTTP交互的请求是否被重定向至疑似资源服务器上，。 |
| 3 | URI Similarity | 本次HTTP交互记录与HTTP记录组中最后的疑似资源服务器记录的URI相似度。 |
| 4 | Same SWF Referer | 本次HTTP交互记录是否拥有与HTTP记录组中最后的疑似资源服务器记录相同的以SWF结尾的HTTP“referer”头，。 |
| 5 | Content Type | 本次HTTP交互中应答报文的实体内容类型字段，即“content-type”头部的值。 |
| 6 | Content Length | 本次HTTP交互中应答报文的实体内容长度字段，即“content-length”头部的值。 |
| 7 | Transfer Enconding | 本次HTTP交互中应答报文“transfer-encoding”头部的值是否为“chunked”。 |
| 8 | Number of Packets | 本次HTTP交互记录相距HTTP记录组中最后的疑似资源服务器记录的报文数。 |
| 9 | Time Interval | 本次HTTP交互记录与HTTP记录组中最后的疑似资源服务器记录的时间间隔。 |

其中，特征1、5、6、7描述了一次单独的HTTP交互的特性，而特征2、3、4、8、9则描述了本次HTTP交互与记录组最后的与疑似资源服务器交互的关系。基于这些特征，我们使用决策树（decision tree）分类算法[]作为检测器，来定位疑似调度服务器。具体来讲，对于每个HTTP记录组，我们对前10条HTTP交互记录提取上述9种特征，然后输入至决策树分类模型中。对于检测结果，我们将阳性（positive）记录所对应的服务器作为疑似调度服务器；而阴性（negative）记录所对应的服务器则被认为是与视频分发无关的其他服务器。

最终，预选择模块的输出是一系列的HTTP记录对（record pair），包含着疑似调度服务器的记录和与之相应的疑似资源服务器记录。注意，来自不同用户记录的HTTP记录对中，疑似调度服务器可能是相同的。这是由于不同用户都与某同一个服务器发生了交互，并且该服务器被本模块中的分类器检测了出来。

* + 1. 综合判决

在综合判决模块中，我们从预选择的输出中过滤出误判结果，并给出对网络视频业务分发服务器的最终检测结果。通常情形下，疑似调度服务器的误判实例，来自于用户在视频分发阶段使用的其他互联网业务的HTTP交互。这些业务所产生的无关的HTTP记录在预选择模块中被错误的认为是调度服务器产生的HTTP记录。这些误判记录与该用户的疑似资源服务器记录一起组成记录对，被输入到了综合判决模块中。

为了滤出误判结果，我们对数据集中预选择模块的伪阳性（false positive）结果进行了分析。我们发现，由于在视频分发阶段不同用户使用的无关的互联网业务是是随机且各不相同的，不同用户的伪阳性结果中对应的疑似调度服务器往往也是不同的。但对于某一真正的调度服务器，从不同用户处检测出来的真阳性（true positive）结果应该是相同的。因此，从整体上来看，某一个误判的调度服务器，其从数据集中所有用户记录中被检测出来的次数会很小，远远小于一个真正的调度服务器的总共被检出次数。我们利用这一特性，在综合判决模块中计算每一个疑似调度服务器在预选择模块中的被检出次数，并据此做出最终的判决结果：只有当某一个疑似调度服务器的被检出次数大于一个预先设定的阈值（detected times threshold，DTT），该服务器才会被认定为一个网络视频业务的调度服务器，而其后相应的疑似资源服务器才会被认定为一个网络视频业务的资源服务器。

我们选择使用这一判决准则还造成了一个的有意思的影响：大规模的输入数据对于我们的检测方法实际上是有利的。被用来进行检测的用户数越大，在预选择模块中给出的真阳性与伪阳性结果的检出次数差距就会越大，进而在综合判决模块中的过滤效果就会越好。

* 1. 实验及结果
     1. 实验设置

我们使用现实网络环境的联网流量数据对我们的检测方法进行验证。实验数据从我们的数据集中随机抽出，包含574,651条HTTP记录。我们将实验数据随机分成两部分，作为训练集和测试集，具体统计概要如表所示。训练集用于模型的学习和阈值参数的选择，而测试集用于评价检测方法性能。

表3-3 实验数据集统计概要

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **HTTP记录数** | | **服务器数** | | |
|  | **分发服务器相关** | **分发服务器无关** | **调度服务器** | **资源服务器** | **其他服务器** |
| **训练集** | 52,220 | 234,990 | 21 | 890 | 8,821 |
| **测试集** | 52,262 | 235,179 | 22 | 883 | 8,856 |

为了确定在综合判决模块中阈值DTT的选择，我们在训练集上使用了5折交叉验证（5-fold cross validation）。结果显示，当DTT的取值在37至677之间时，检测结果保持不变。这表明给定数据集时，我们的检测系统对于DTT的取值并不是特别敏感，合适的的取值区间较大。在我们的实验中，我们设定DTT=40。需要注意的是，虽然DTT本身具体一定健壮性，但DTT的合适取值区间是受数据集大小的影响的。在实际应用中，我们应该保证建模所用数据与检测数据的大小相近。例如，对于以天为时间粒度采集的数据，各天的流量数据大小相差不大，我们可以使用某天的数据来发现DTT的合适的取值区间，并选取DTT具体取值，然后将该DTT应用到之后的各天新数据的检测上。

在我们的实验中，对于测试集产生的检测结果，我们使用每种服务器类型的检测精度（precision）和召回率（recall）作为系统性能的评价指标。其中，精度定义如下：

(3-3)

为相关检出记录数与检出记录数的比值。而召回率定义如下：

(3-4)

为相关检出记录数与相关记录数的比值。具体来讲，在我们的实验中，某一记录或服务器类别的精度为“正确识别的该类别记录或服务器数量”与“检测结果中该类别的记录或服务器数量”的比值。而某一记录或服务器类别的召回率则是“正确识别的该类别记录或服务器数量”与“该类别真实的记录或服务器数量”的比值。

* + 1. 检测性能

**预选择模块检测性能**：我们对预选择模块中疑似调度服务器的检测结果进行分析。首先，表3-X列出了对预处理模块输出的各HTTP记录组中，各HTTP记录是否对应调度服务器的检测结果。从表中可以看出，我们的方法对于调度服务器HTTP记录的识别力很强，大多数HTTP记录都被准确的分类，各类别的检测精度和召回率都在97%以上。这为后面的检测步骤打下了良好的基础。基于HTTP记录的检测结果，我们进一步从服务器的角度分析预选择模块的检测性能，如表3-X所示。我们发现，虽然对于其它服务器检测精度和召回率很高，调度服务器的检测精度仅有不到35%。正如3.5.3小节所讨论的，这是由于一小部分来自不同用户的伪阳性记录（约3%）覆盖了大量的与视频分发无关的服务器（超过65%）。因此，预选择模块本身不能很好的对调度服务器进行检测，我们加入后续的综合判决模块是非常有必要的。

表3-X

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **真实记录类型** | | **精度** |
|  |  | **调度**  **服务器记录** | **其他**  **服务器记录** |
| **判决结果** | **疑似调度**  **服务器记录** | 25,905 | 728 | 97.27% |
| **其它**  **服务器记录** | 226 | 23,4451 | 99.90% |
| **召回率** | | 99.14% | 99.69% | - |

表3-X

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **真实服务器类型** | | **精度** |
|  |  | **调度服务器** | **其他服务器** |
| **判决结果** | **疑似调度服务器** | 22 | 41 | 34.92% |
| **其它服务器** | 0 | 8815 | 100% |
| **召回率** | | 100% | 99.54% | - |

**系统整体检测性能**：表3-5列出了系统对于测试集中服务器类型的最终检测结果性能。由表可知，我们的系统用于检测无论是调度服务器、资源服务器还是其他类型无关的服务器，在性能指标（精度和召回率）上都同时十分优秀。其中检测精度都接近100%，而召回率都在85%以上。图3-X给出了网络视频业务提供商优酷几个被检测出来的调度服务器与资源服务器之间的拓扑关系。其中，圆圈代表调度服务器，三角代表资源服务器。圆圈与三角之间的边表示调度服务器与资源服务器的对应关系，即该调度服务器发送了该资源服务器上的视频地址URL给用户。从图中，我们可以清晰的看到两种服务器间的多对多的关系：大多数的资源服务器都对应着多个调度服务器。这样，即使某个调度服务器未被系统检测出来，其对应的资源服务器也能通过其它的被检出调度服务器而被识别出来。因此，在最终结果中，未检出的资源服务器比例要远小于未检出的调度服务器比例。

表3-5 系统对分发服务器的最终检测性能

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **真实类型** | | | **精度** |
| **调度服务器** | **资源服务器** | **其他服务器** |
| **判决结果** | **调度服务器** | 19 | 0 | 0 | 100% |
| **资源服务器** | 0 | 840 | 0 | 100% |
| **其他服务器** | 3 | 43 | 8,856 | 99.48% |
| **召回率** | | 86.36% | 95.13% | 100% | - |

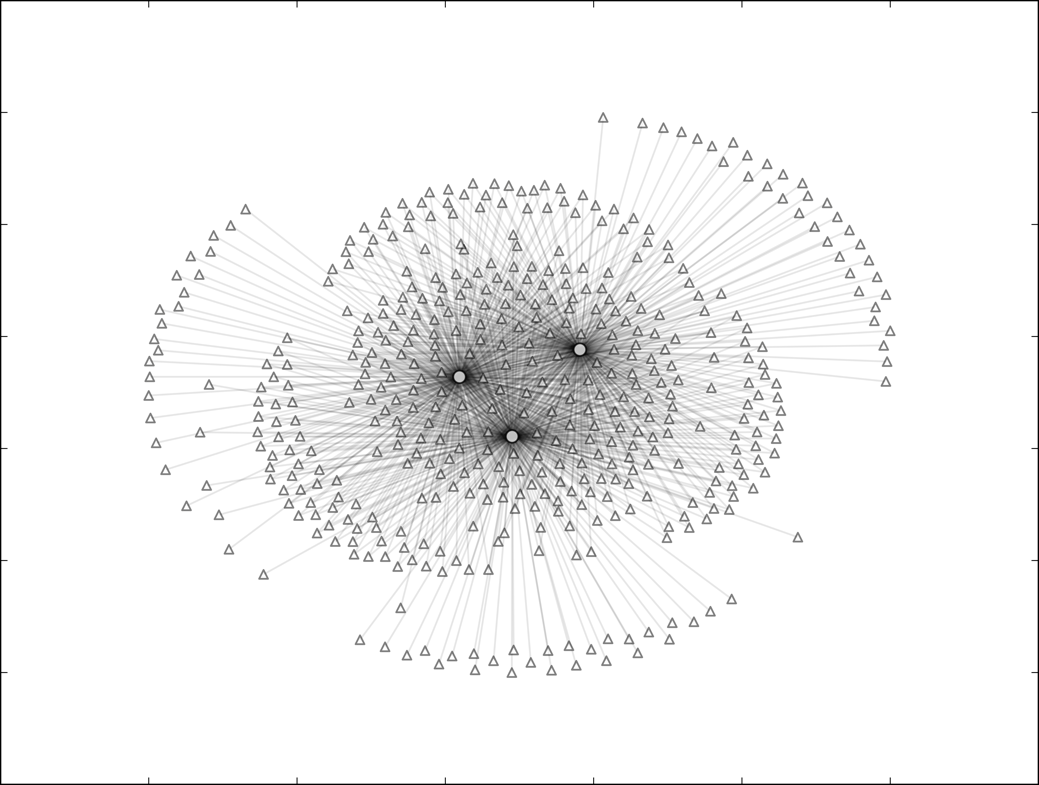


图3-X

更进一步，我们对于检测结果中的错误实例进行了分析。我们发现错误实例主要包含两种：1）调度服务器被判断为其他服务器；和2）资源服务器被判断为其他服务器。对于情形1），我们发现其出现的原因是有些调度服务器在我们数据集中出现的次数过小，有的甚至只出现过1、2次。这样，在预选择阶段中，这些调度服务器虽然能被成功检出，但其检出次数无法超过阈值DTT，故而在综合判决阶段被判决成了其他服务器。而有些资源服务器仅仅和这些未检出的调服服务器有对应关系。根据我们的检测机制，资源服务器的检测要依靠已检出的调度服务器，故这些资源服务器也同时无法被检测出来，从而造成了情形2）。

* + 1. 实验讨论

**检测器选择**：在检测系统的预选择模块中，我们尝试使用了不同的机器学习分类算法作为检测器，来识别网络视频业务中的调度服务器。我们分别比较了朴素贝叶斯（naive Bayes）[]、线性核（linear kernel）支持向量机（support vector machine, SVM）[]和决策树方法。对于每种方法，我们在训练集上使用5折交叉验证来进行检测性能评价。各方法的平均检测精度与平均召回率如表3-6所示。由表可知，各个分类方法对于资源服务器和其它服务器的检测性能相差不大。但对于调度服务器等检测来说，朴素贝叶斯方法的精度仅有64.5%，而支持向量机方法的召回率仅有68.2%。综合来看，决策树方法能够将检测精度和召回率同时平衡在较高的水平上。除此之外，决策树方法还具有不要求特征之间的独立性，及树的结构可以体现各特征的重要程度等优点。

因此，在我们的方法中，我们选择了决策树分类方法作为预选择模块中的分类器。

表3-6 预选择模块中不同分类器性能比较

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **调度服务器** | | **资源服务器** | | **其它服务器** | |
|  | **精度** | **召回率** | **精度** | **召回率** | **精度** | **召回率** |
| **朴素贝叶斯** | 64.5% | 90.9% | 100% | 95.5% | 99.5% | 99.9% |
| **支持向量机** | 83.3% | 68.2% | 100% | 88.7% | 98.8% | 100% |
| **决策树** | 100% | 86.4% | 100% | 95.1% | 99.5% | 100% |

**特征重要性**：我们进一步分析了调度服务器检测中各特征的相关重要程度。我们基于构造决策树模型时，计算节点分裂时使用等基尼系数（Gini index）来衡量特征重要性，如表3-7所示。其中，“URI相似性”是最具有区分度的特征。接下来是“HTTP重定向行为”和“相同的SWF后缀referer头”两个特征。当一个服务器与一个传输视频文件的服务器出现了上述这些通信行为，很大可能性上这两个服务器就是网络视频业务的调度服务器和资源服务器。但是，并非所有的调度服务器和资源服务器之间的通信都会出现上述特征。所以，后面的6种特征仍具有很高的重要性。我们的方法将这些特征综合考虑，能够很好的从无关的其他服务器中区分出调度服务器。

表3-7 特征重要性排名

|  |  |  |
| --- | --- | --- |
| **排名** | **特征** | **重要性** |
| 1 | URI Similarity | 829.10 |
| 2 | HTTP Redirection | 700.58 |
| 3 | Same SWF Referer | 456.89 |
| 4 | Transfer Encoding | 290.12 |
| 5 | Content Type | 108.61 |
| 6 | Request Method | 53.72 |
| 7 | Content Length | 44.02 |
| 8 | Number of Packets | 42.54 |
| 9 | Time Interval | 41.38 |

* 1. 本章小结

在本章中，我们关注于网络视频业务中分发服务器的检测问题。能够从互联网中海里的服务器中自动、准确、及时的定位网络视频业务中的调度服务器与资源服务器，将有助于网络运营商解决对于过顶业务流量的管控、管理优化网络设施部署、减少非必要开销等问题。首先，我们深入的分析了用户与视频服务器之间是如何进行通信的。我们利用实验环境，对不同的网络视频业务提供商进行了主动播测，并捕获了传输报文。基于报文分析，我们发现不同业务提供商之间存在着一个通用的用户-服务器通信流程。我们对该通用流程进行了解析与总结。然后，基于从我国东南某省固网中采集到的超过五百万条互联网流量数据，我们对视频分发阶段中用户与服务器之间的通信行为特性进行了分析。根据通信特性，我们定义了一系列具有区分度的特征，并结合机器学习技术分类算法，提出了一个网络视频业务分发服务器检测系统。最后，我们使用真实数据对该检测系统进行了验证。实验结果显示，我们提出的系统检测性能优异：检测精度接近100%，同时召回率超过85%。此外，我们还对系统中使用的分类器的选择，以及分类问题的特种重要性进行了讨论。

在未来工作中，我们希望采集不同种类的数据源，如移动网络数据，其他网络视频业务提供商数据，其他地域或国家用户数据，来进一步验证并提高我们的检测方法。

1. 移动网络中网络视频业务用户行为特性分析
   1. 概述

绪论及每一章开头，一定要说清楚跟已有研究不一样的地方。

意义、创新点

1）新颖的研究对象。我们关注于网络视频业务分发服务器的检测。据我们所知，对这一具有重要实际意义问题的研究尚属首次。2）全面的分析角度。3）高效的检测

* 1. 研究现状

Related work

* 1. 数据集

观看者产生的流量、上传者的网站记录

从哪 如何 格式 统计

* 1. 业务使用特性

时间维度用户活跃性

空间维度用户分布

用户移动模式

重度用户检测

用户资源消耗模型

* 1. 本章小结

1. 网络视频业务用户喜好对比性特性分析
   1. 概述

绪论及每一章开头，一定要说清楚跟已有研究不一样的地方。

意义、创新点

* 1. 研究现状

Related work

* 1. 数据集

观看者产生的流量、上传者的网站记录

* 1. 内容喜好特性

视频类型分布

用户兴趣聚类

上传者影响力

用户喜好网络

用户播放量分布

* 1. 本章小结

1. 网络视频业务内容流行度分析及预测
   1. 概述

绪论及每一章开头，一定要说清楚跟已有研究不一样的地方。

意义、创新点

* 1. 研究现状
  2. 数据集

新发布视频30天中的播放量

* 1. 流行度分析

群体角度：分布

单体角度：增长趋势

* 1. 流行度预测

有观察：数值

有观察：级别

无观察：由回归问题退化成分类问题

* 1. 本章小结

1. 总结与展望

参考文献

[1] 思科白皮书

[2] Iraj Sodagar, “The mpeg-dash standard for multimedia streaming over the internet,” IEEE MultiMedia, 2011.

[3] Mohit Saxena, Umang Sharan, and Sonia Fahmy, “Ana- lyzing video services in web 2.0: a global perspective,” in International Workshop on Network and Operating Systems Support for Digital Audio and Video, 2008.

[4] Vijay Kumar Adhikari, Sourabh Jain, and Zhi-Li Zhang, “Youtube traffic dynamics and its interplay with a tier-1 isp: an isp perspective,” in ACM SIGCOMM conference on Internet measurement, 2010.

[5] Vijay Kumar Adhikari, Sourabh Jain, Yingying Chen, and Zhi-Li Zhang, “Vivisecting youtube: An active measurement study,” in IEEE INFOCOM, 2012.

[6] VijayKumarAdhikari,SourabhJain,andZhi-LiZhang, “Where do you” tube”? uncovering youtube server se- lection strategy,” in International Conference on Com- puter Communications and Networks, 2011.

[7] Ruben Torres, Alessandro Finamore, Jin Ryong Kim, Marco Mellia, Maurizio M Munafo, and Sanjay Rao, “Dissecting video server selection strategies in the youtube cdn,” in International Conference on Dis- tributed Computing Systems, 2011.

[8] Louis Plissonneau, Ernst Biersack, and Parikshit Juluri, “Analyzing the impact of youtube delivery policies on user experience,” in International Teletraffic Congress, 2012.

[9] Maciej Korczynski and Andrzej Duda, “Classifying ser- vice flows in the encrypted skype traffic,” in IEEE ICC, 2012.

[10] Weibo Chu, Bin B Zhu, Feng Xue, Xiaohong Guan, and Zhongmin Cai, “Protect sensitive sites from phishing at- tacks using features extractable from inaccessible phish- ing urls,” in IEEE ICC, 2013.

[11] Varun Chaudhary and Ashish Sureka, “Contextual fea- ture based one-class classifier approach for detecting video response spam on youtube,” in International Con- ference on Privacy, Security and Trust, 2013.

[8]

[8]

[8]

[8]

[8]

[8]