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（主动、range）、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 业务逻辑

播测

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

**HTTP协议属性字段**：用户与视频服务器交互的流量，大部分是基于应用层HTTP协议的。下面以优酷视频为例，我们对这些流量的协议头中各属性的值做概要性的统计分析。表x给出了我们播测数据集中HTTP请求报文的请求方法属性值的分布情况。我们发现，一共存在三种请求方法：GET、POST和HEAD。其中，如同我们预期的，GET方法占据了大部分的比例。这些GET请求主要用于下载网页及网页内嵌内容，以及从用户向服务器发送动态上报信息。POST方法也占据者不容忽视的报文比例。这些POST请求则主要用于向业务提供商的统计服务器自动上报视频播放进度、用户操作记录以及定时心跳等。此外，数据集中还存在少量的HEAD报文。在HTTP协议中，对于一个HEAD请求，服务器会返回一个只有协议头而无传输实体的HTTP应答报文。在网络视频业务中，该请求方法被用于在实际下载视频文件之前获取视频文件大小，以便视频播放器选择合适的文件范围（range）进行分片下载。

表x HTTP请求方法分布

|  |  |
| --- | --- |
| **请求方法** | **报文数比例** |
| GET | 69.63% |
| POST | 30.22% |
| HEAD | 0.15% |

对于HTTP应答报文，我们分析了其应答头部的状态码分布，如表x所示。

我们发现，尽管200状态码的HTTP应答在报文数上占据大多数，它们传输的流量仅占总字节数的四分之一。而对于206状态码的HTTP应答，它们占总报文数不到6%，却传输了四分之三的流量。这是由于，在优酷视频中，大多数的视频文件是通过HTTP协议的range头部属性来分段请求下载的。而“206 Partial Content”正是这种请求对应的应答状态码，所请求的文件范围分片可通过该HTTP应答的实体内容传输给用户。相较于其他类型的文件，视频文件的大小要大得多，因此这些206状态码的HTTP应答报文占据了很大比例的流量字节数。此外，从表中我们发现302状态码的HTTP应答报文占据了总报文数中不小的比例（约4.6%）。此种HTTP应答用于被请求资源的重定向，在网络视频业务中与视频分发技术密切相关。

表x HTTP应答状态码分布

|  |  |  |
| --- | --- | --- |
| **应答状态码** | **报文数比例** | **字节数比例** |
| 200 OK | 85.00% | 27.69% |
| 204 No Content | 0.14% | 0.00% |
| 206 Partial Content | 5.86% | 72.31% |
| 302 Found | 4.59% | 0.00% |
| 304 Not Modified | 1.15% | 0.00% |
| 4xx Client Error | 1.45% | 0.00% |
| 5xx Server Error | 1.81% | 0.00% |

我们还进一步分析了用户与服务器之间传输的实体内容的类型。表x列出了按请求报文数排名的前15位“content-type”属性字段的名称、报文数比例和字节数比例。我们从表中可以看出，约有35%的报文对应着图片类型（“image/xxx”）的实体内容。这说明视频页面中的很大一部分内嵌内容是各种图片。其中，“image/jpeg”对应的JPEG格式是使用最多的图片格式，报文数量比例超过了30%。而对于文本类型（“text/xxx”）的实体内容，约占总报文数的20%。此外，我们发现实体内容类型“application/json”占据了相当大比例的的报文（近30%）。其对应的JSON格式，是一种广泛用于用户与服务器间异步通信的数据交换格式。在优酷视频流量中，例如用户评论、视频描述、视频地址等动态加载内容是由JSON格式的文件从服务器传递给用户的。

然而，从传输的字节数角度来看，尽管上述的“image/xxx”、“text/xxx”和“application/json”三种实体内容类型占据超过85%的报文数量，其仅传输了不到5%的数据流量。而在总报文数中仅占5%左右的视频类型（“video/xxx”）报文，却传输了超过95%的流量字节数。从表中我们可以发现，传输的视频文件类型主要有两种：FLV（“video/flv”）和MP4（“video/mp4”）。其中，从报文数角度，FLV类型近5倍于MP4类型；然而从传输字节数角度，FLV类型仅是MP4类型的1.7倍。这是由于网络视频业务中，一个MP4视频的大小往往要远大于一个FLV视频的大小。虽然被观看的MP4视频数量上较小，但于流量消耗和带宽占据而言，其与FLV视频的差距并非如报文数所体现的那样大。

表x 实体内容类型分布

|  |  |  |
| --- | --- | --- |
| **实体内容类型** | **报文数比例** | **字节数比例** |
| image/jpeg | 30.13% | 2.04% |
| application/json | 28.62% | 0.04% |
| text/html | 14.44% | 0.02% |
| - | 6.99% | 0.00% |
| video/flv | 4.49% | 59.39% |
| text/javascript | 3.25% | 0.06% |
| image/png | 2.88% | 0.11% |
| application/vnd.apple.mpegurl | 1.74% | 0.02% |
| text/css | 1.59% | 0.02% |
| text/plain | 1.49% | 0.00% |
| image/webp | 1.44% | 0.06% |
| video/mp4 | 0.90% | 35.66% |
| text/xml | 0.81% | 0.00% |
| image/gif | 0.51% | 0.00% |
| application/x-javascript | 0.30% | 0.07% |

* 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。

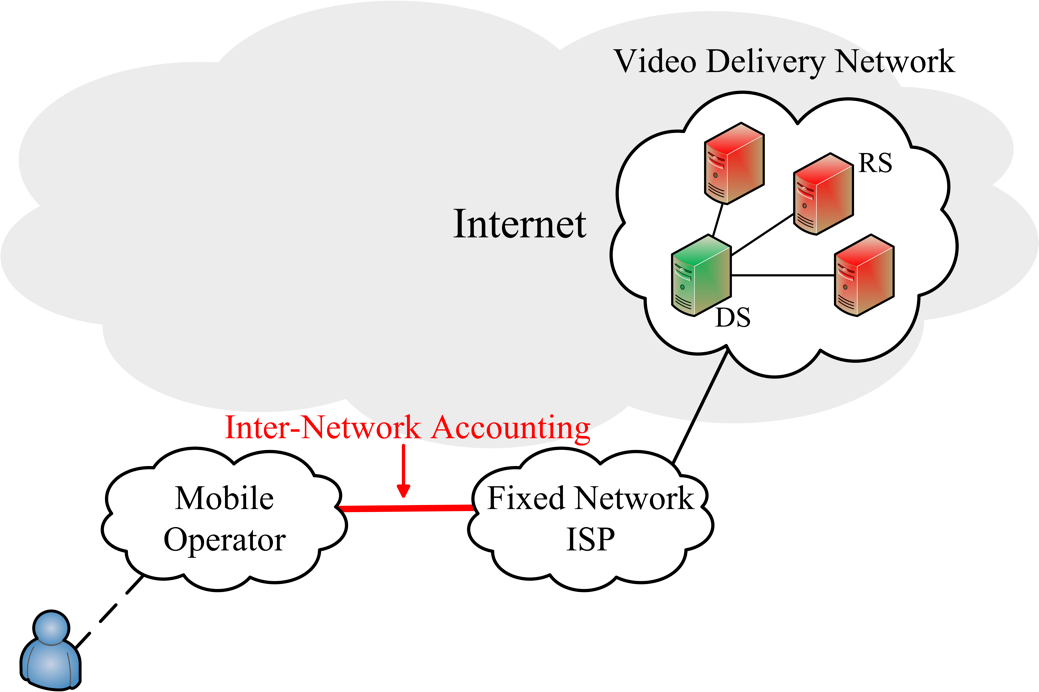


图3-1

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。

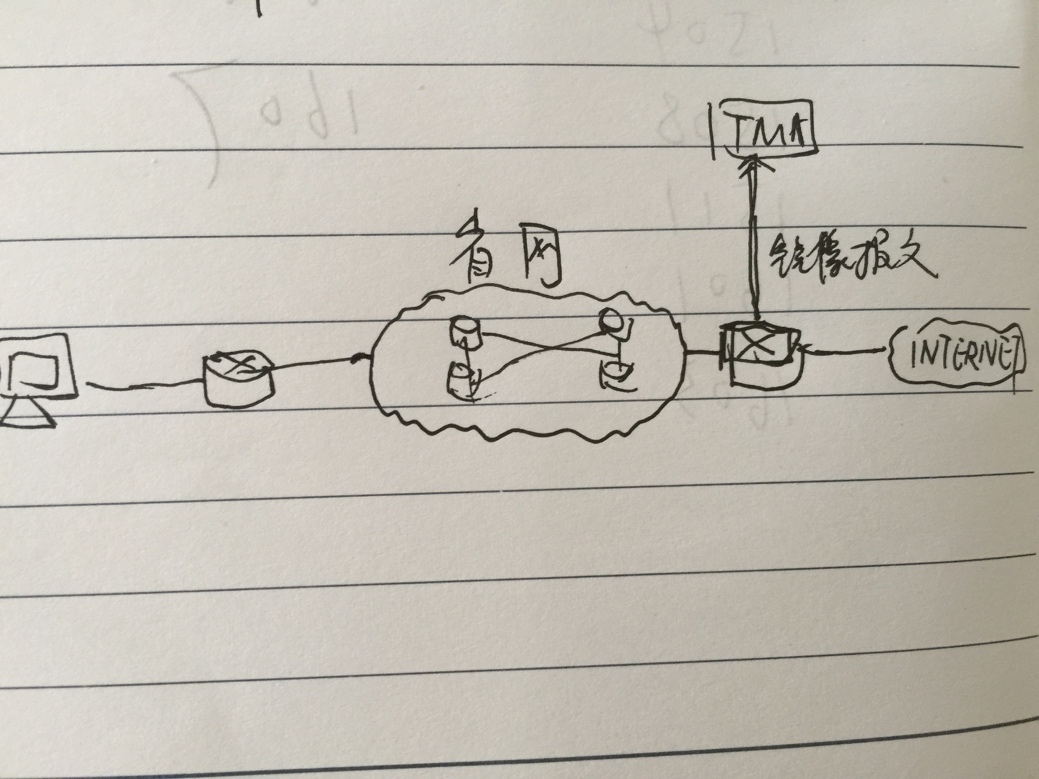


图3-2 数据采集部署示意

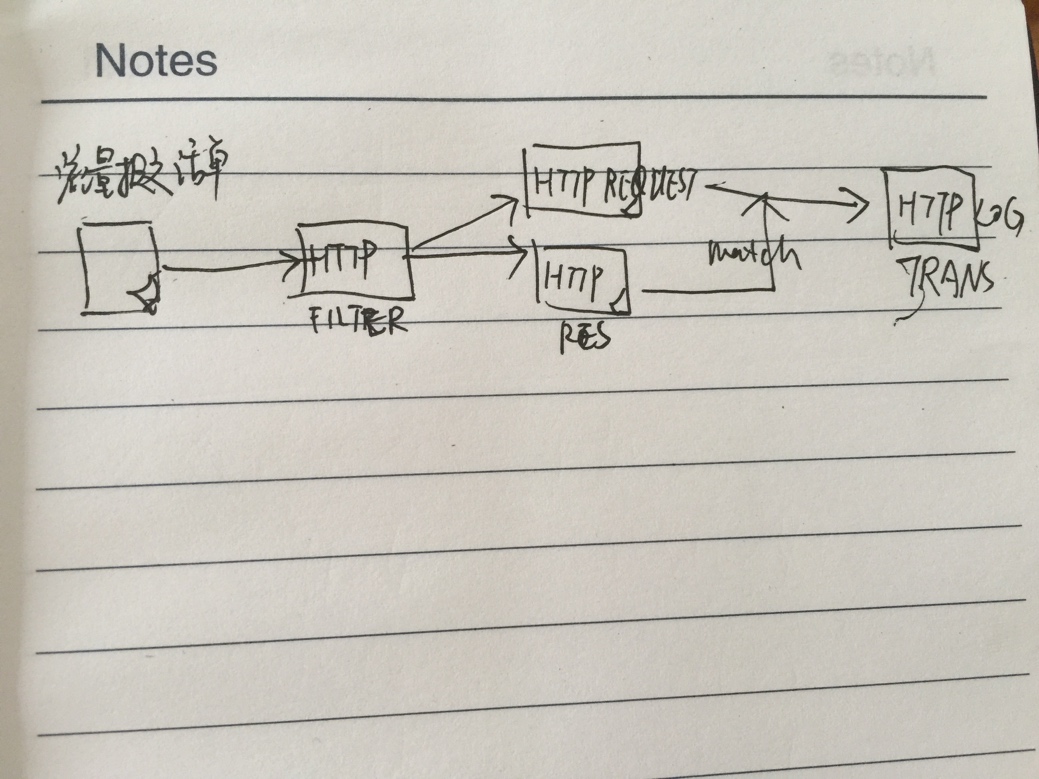


图3-3

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。

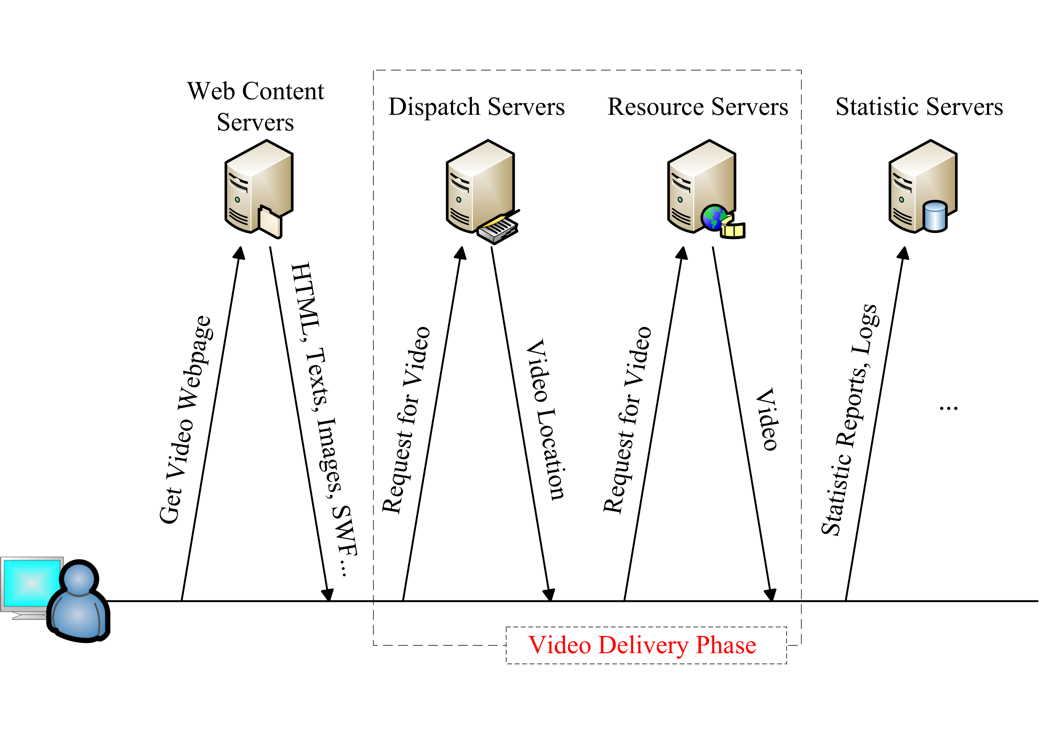


图3-4 通用的用户-服务器通信流程

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。

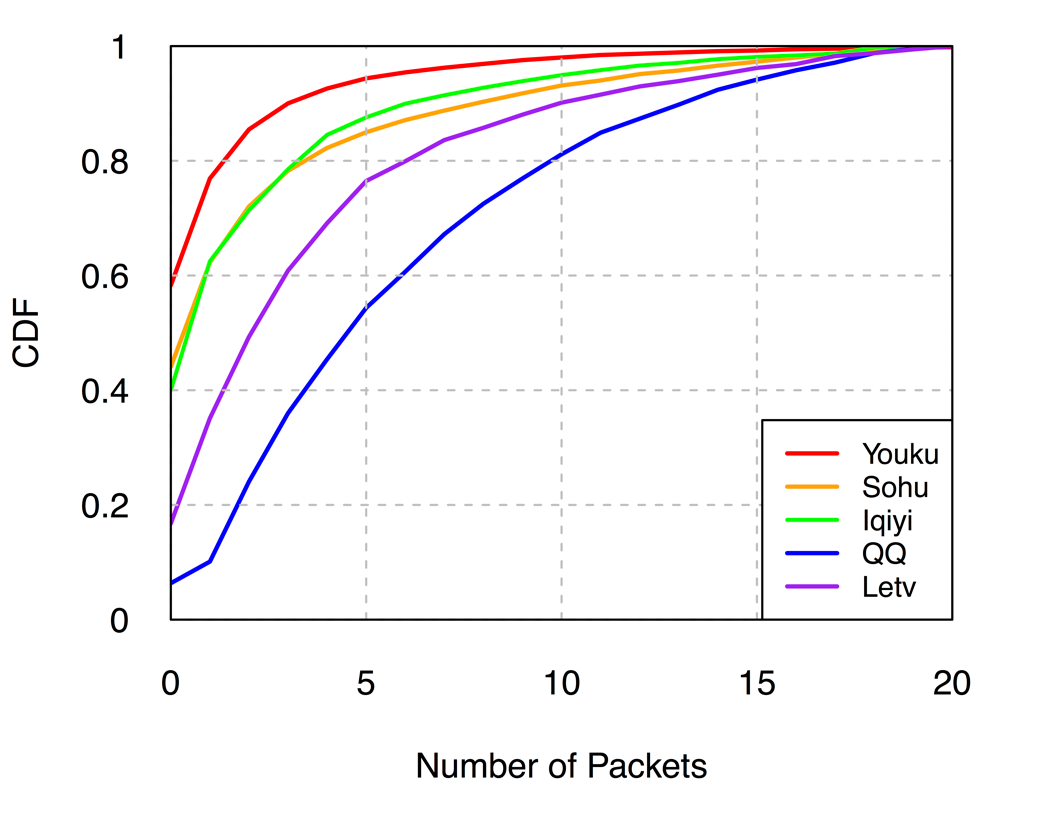


图3-5 不同网络视频业务提供商的的累积分布函数

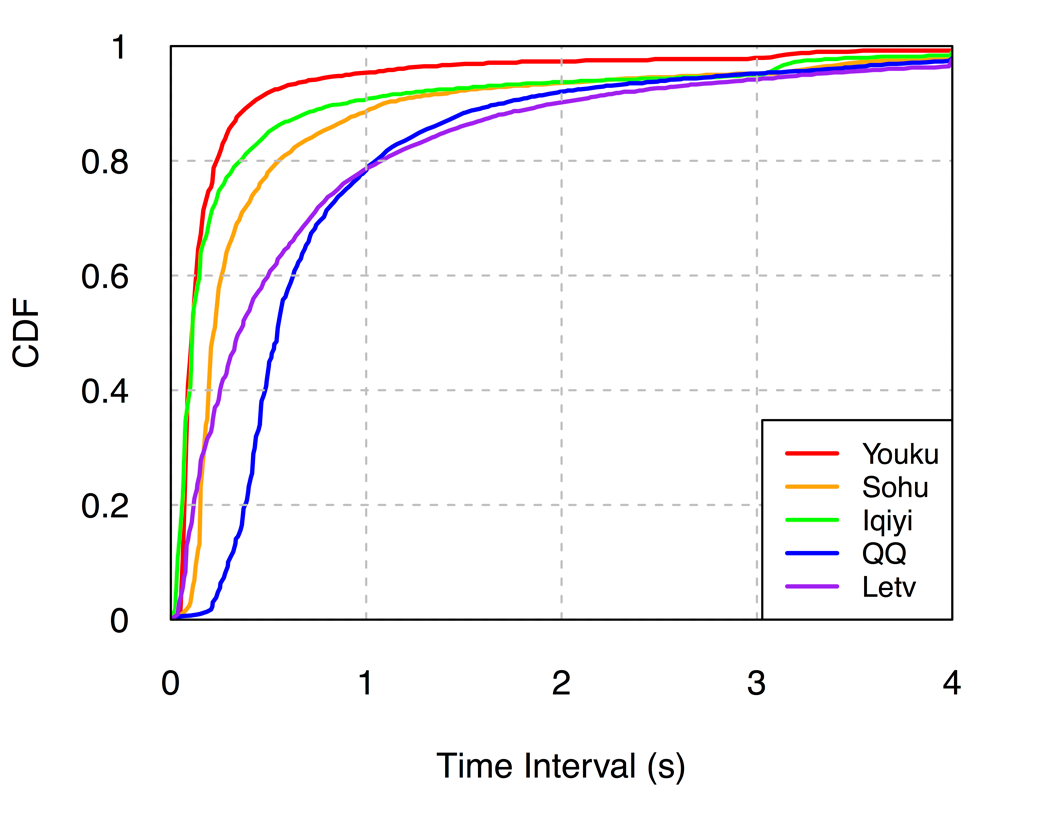


图3-6 不同网络视频业务提供商的的累积分布函数

**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。

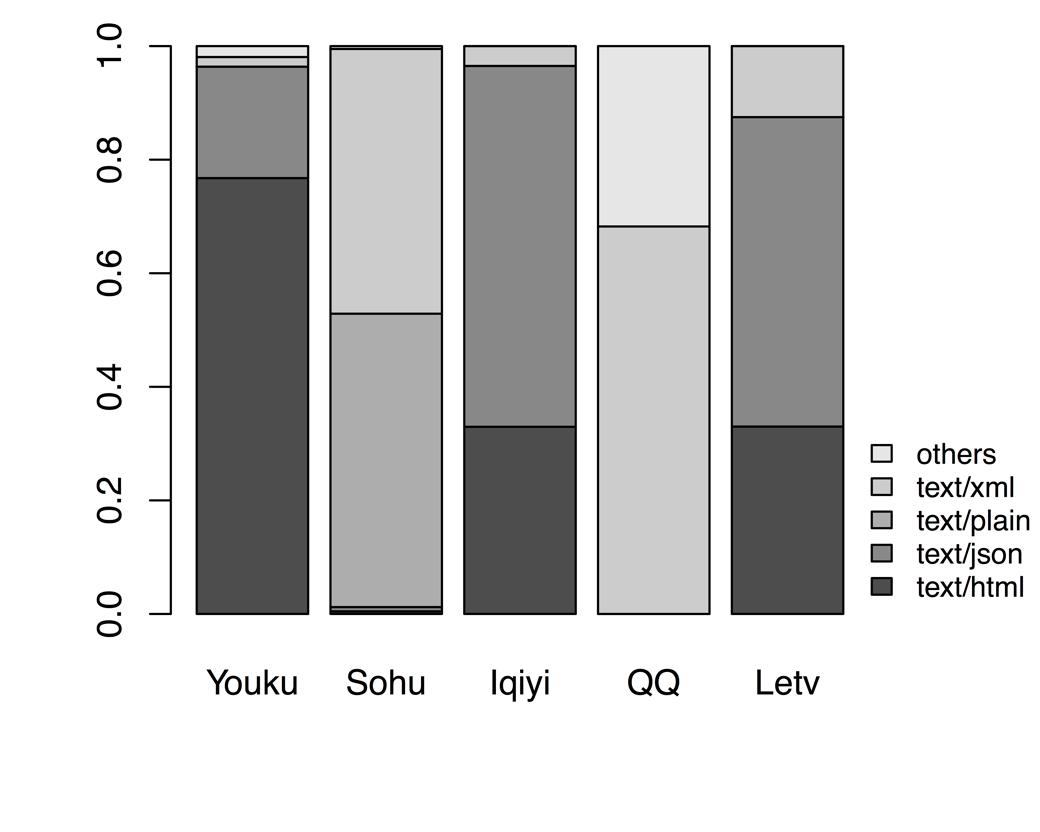


图3-7 不同网络视频业务提供商的调度服务器传输的HTTP实体内容类型分布

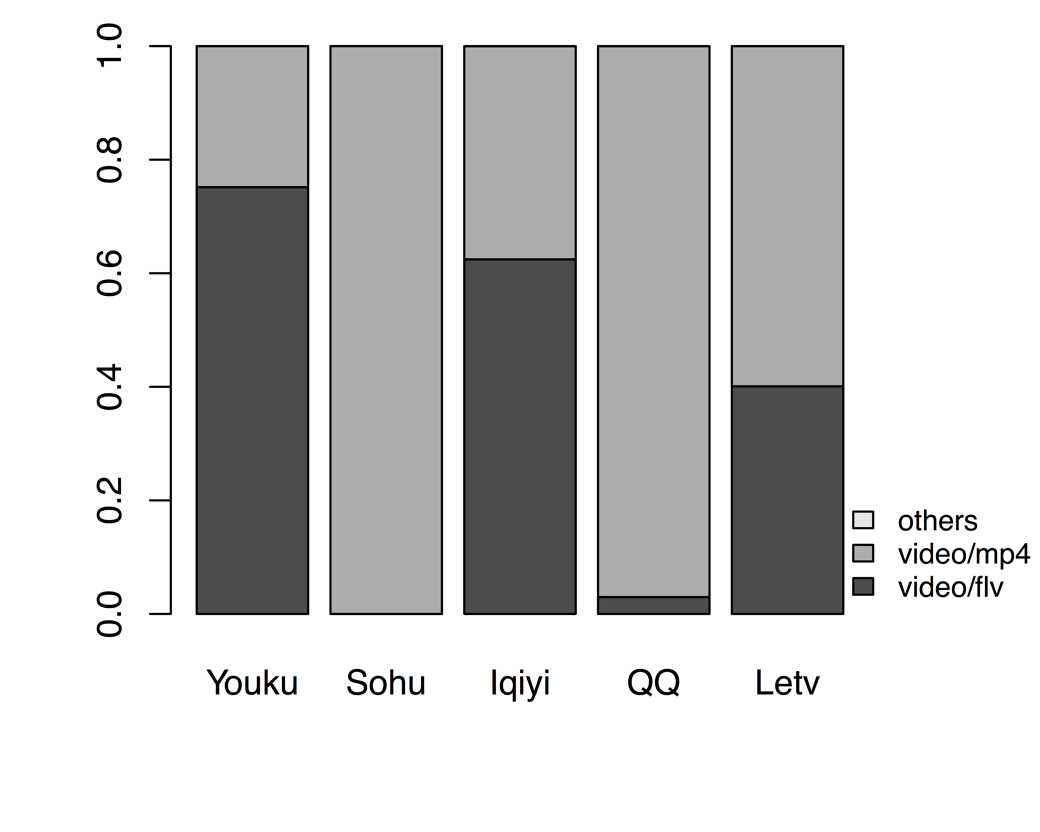


图3-8 不同网络视频业务提供商的资源服务器传输的HTTP实体内容类型分布

**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3-2 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3-3

|  |  |  |  |
| --- | --- | --- | --- |
| **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。

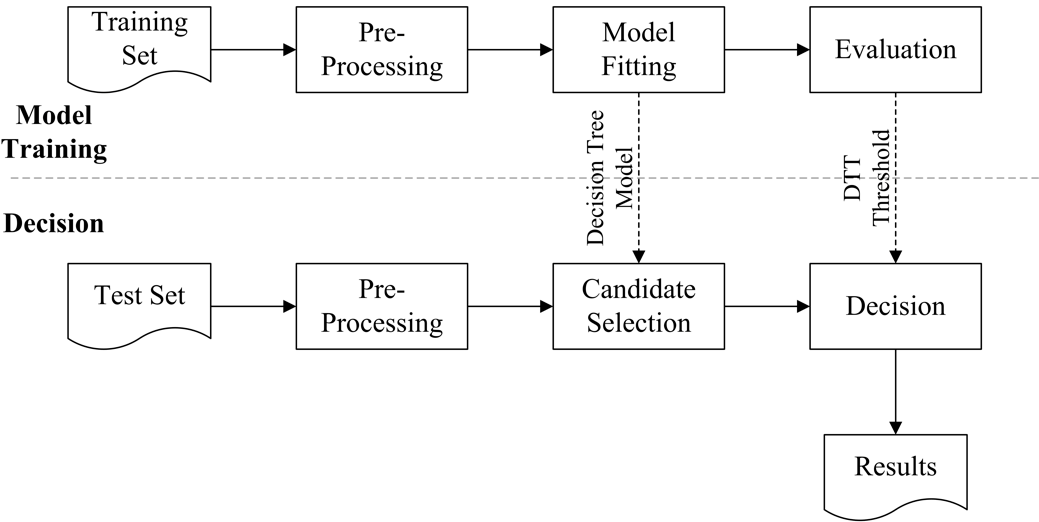


图3-9 网络视频业务分发服务器检测系统整体架构

* + 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3-4 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3-5 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：

(3-5)

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：

(3-6)

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3-6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **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3-7

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **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3-8 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% | - |

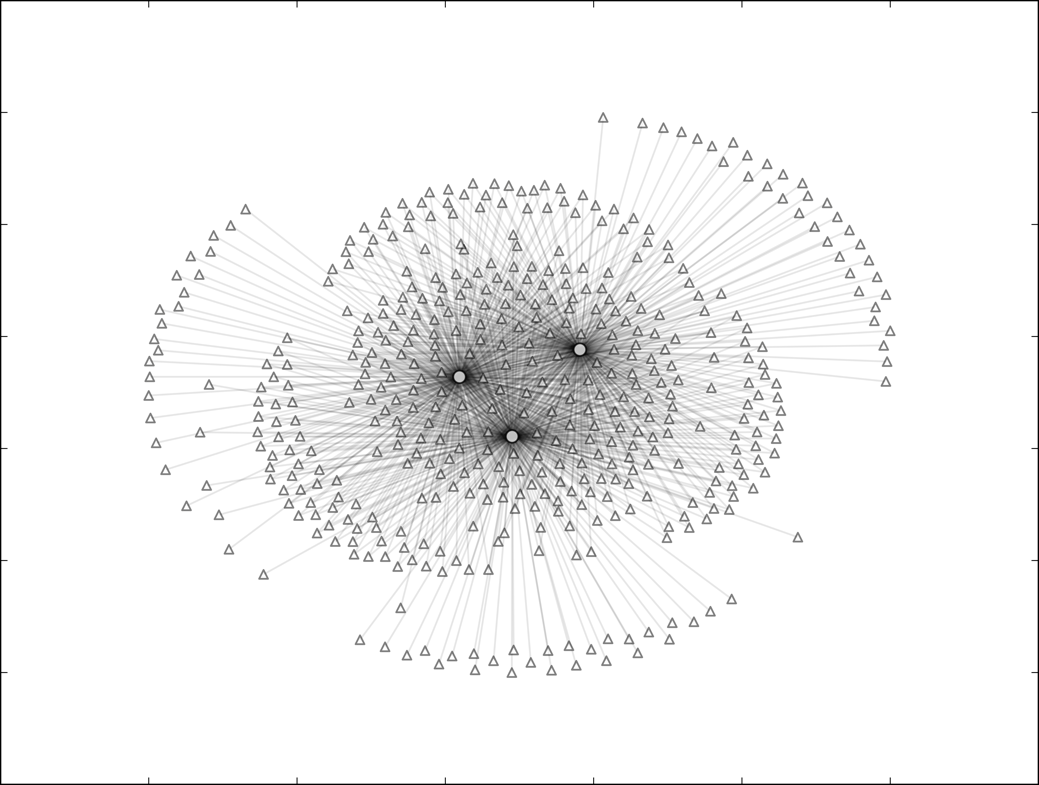


图3-10

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3-9 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3-10 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

基于大规模数据的互联网业务用户行为特性分析是一个十分重要的研究领域。其有助于网络运营商和业务提供商更好的部署网络设施和调整业务设计，从而进一步提高用户体验。目前，互联网中主流的业务主要包括：网络视频业务、网络社交业务、电子商务业务、网络游戏业务、网络音乐业务等等。其中，从流量的角度来看，网络视频业务是现今的互联网中最大的组成部分，并将在未来占据更大的比例。如思科公司的白皮书【】中指出，201X年全球的网络视频流量占据互联网流量的x%，并将于201X年突破x%。因此，对网络视频业务的用户行为进行专门的深入的分析研究，是十分必要的。

通常来讲，用户有两种方式接入互联网来使用视频业务：通过固定网络、和通过移动网络。目前学术界对于固定网络中网路视频业务用户行为特性已经有了一定的研究。但近年来，由于大屏幕智能手机的普及、专用视频业务APP的出现、以及更快速的无线通信标准的实现，越来越多的用户开始转向通过移动网络来使用网络视频业务【1504】。201X年全球整体的IP流量中，移动网络流量就占据了x%。并且，预计于201X年，移动网络流量的比例将首次超过固定网络流量的比例【】。因此，对于网络视频业务用户行为的研究，应该转向之前少有分析的移动网络环境。当用户使用移动终端、通过移动网络来使用网络视频业务时，其会受到无线信号强度、终端电池电量、数据流量计费等新因素等影响。此时，用户的行为将与其在固定网络中的行为体现出很大的不同【1508-3】。

在本章中，我们从中国东北某省的移动网络中采集了超过170亿条流量话单数据，从中提取了业务级别、无线接入级别和数据传输级别的用户信息，并对优酷视频用户的行为特性进行了深入的分析。本章研究内容的主要贡献与创新点在于：1）新颖的研究对象。我们关注于移动网络中网络视频业务的用户行为特性。对于视频用户的行为分析，现存的研究工作多集中于固定网络环境中，我们的工作很好的弥补了在这新兴的移动网络环境中的研究空白。2）独特的分析角度。基于大规模的移动流量数据，我们对视频用户行为的分析着眼于业务使用强度、请求时刻、请求位置、数据消耗等角度，反映了用户在业务提供商、无线接入网以及核心网的资源消耗。尤其是，我们基于移动网络中的位置信息，对用户从空间的维度上展开了研究，在网络视频业务用户行为分析中上尚属首次。3）对重度用户的关注。在对用户的资源消耗分析中，我们发现了某些重度用户的存在。我们提出了一个非参数的重度用户自动检测方法，有效的避免了参数方法所造成的独断性的偏差。我们进一步分析了重度用户与非重度用户、以及不同类别的重度用户之间的行为特性区别。

We demonstrate the strong correlations between users, which is reflected by the small-world phenomenon.

%so that they form a small-world network.

Eventually, we study the distribution of view count and find out four growth patterns for the new published videos.

Our analysis result provides an insight into how Youku is utilized by mobile users and demonstrates promising practical applications for both network operators and service providers.% how

%With results we imp and try to answer the questions such as:

%When and where should proxy servers be utilized to cache videos?

%Which videos are more valuable to cache?

%And how long should the videos be cached?

%Our results provide a comprehensive insight into the way people use YouTube on mobile devices, and show a very high potential for video cacheability on the cellular network.

These features above make the video traffic in mobile network in China quite different from that of other previously analyzed cases.

Therefore, understanding the characteristics of this specific service is of great importance for Chinese network operators and service providers to tune their infrastructures and services.

In this paper, we present an in-depth characterization study on the mobile traffic of a leading video service provider in China, namely Youku.

The work is based on a large scale dataset collected from a cellular network in a southeast province in China on December 24th 2014 and January 24th 2015.

We filter the traffic records of Youku from the collected dataset, and study how the service is used by mobile users based on the filtered traffic records of Youku.

%We filter the traffic records of Youku and study how the service is used by mobile users.

For network level characteristics, we analyze the protocol fields of mobile video service.

And for application level characteristics, we look through the user activity and the video properties.

%We provide a detailed and in-depth analysis work on characteristics of the Youku traffic in mobile network.

The main contributions of this paper are summarized as followed:

1) We thoroughly analyze the characteristics of Youku traffic in mobile network from multiple aspects;

2) We find the usage patterns and specific features of Youku workload and explain their reasons;

3) We reveal the application level diversities of the requested videos, including category, duration and age.

* 1. 研究现状

对的早期研究集中于固定网络中：\noindent \textbf{Video Traffic Characterization in wired network: }Early studies mainly utilized the data collected from fixed networks.

Based on traffic collected in a campus network, Gill \textit{et al.} studied the usage pattern, video properties, referencing characteristics and transfer behaviors of YouTube \cite{gill2007youtube}.

Zink \textit{et al.} collected HTTP traffic of YouTube also in a campus network and analyzed the duration, data rate, population and access patterns \cite{zink2009characteristics}.

Arvidsson \textit{et al.} analyzed YouTube user demand patterns and locality properties based on the data collected from two municipal networks in Sweden \cite{arvidsson2013analysis}.

In \cite{abhari2010workload} and \cite{cheng2013understanding}, authors performed long-term data crawling from the YouTube website, and analyzed the video properties, popularity distribution and access pattern of the service.

\noindent \textbf{Video Traffic Characterization in mobile network: }Ramos-Munoz \textit{et al.} in \cite{ramos2014characteristics} analyzed the network level characteristics of YouTube traffic captured from a 3G networks and compared the results with the video traffic in wired networks .

In \cite{casas2014youtube}, Casas \textit{et al.} collected YouTube traffic from both mobile and fixed networks, and studied the server infrastructure, traffic characteristics and caching usage.

Abdesslem \textit{et al.} collected YouTube requests from a nationwide cellular network, and analyzed the user activity, video properties and content popularity \cite{ben2014large}.

All these existing works are either focusing on video traffic in wired network or the video service provider Youtube.

Our study complements these existing works by analyzing the video traffic in mobile network in China.

And providing a detailed mobile video traffic characterization based on Youku, which provides multiple videos service by one-stop solution, including UGC, VoD, live streaming and so on.

Meanwhile, since the specific billing plans of cellular data network in China may also effect the user behavior, they are also taken into account in the analysis.

With the increasing popularity of online video service, more and more attention has been attracted from the researchers.

Many characterization works on video service are published, especially for the UGC service provided by YouTube.

%In this section we outline some previous related works and how our study complements them.Some studies utilized the data collected in fixed local-area networks or crawled from video websites.

\noindent \textbf{Video service characterization works based on wired local network or crawled data:}

Gill \textit{et al.} \cite{gill2007youtube} examined the usage patterns, file properties, popularity, referencing characteristics and transfer behaviors of YouTube in a campus network.

Zink \textit{et al.} \cite{zink2009characteristics} captured HTTP headers in a university campus network and analyzed the duration, data rate, population and access patterns of YouTube videos.

Abhari \textit{et al.} \cite{abhari2010workload} investigated the data crawled from YouTube website in five months, and analyzed the popularity distribution and access pattern.

Cheng \textit{et al.} \cite{cheng2013understanding} performed a long-term crawling and studied the length, access pattern, active life span, growth trend and social networking of YouTube.

Our study complements these existing works by analyzing the comprehensive type video service, based on a large-scale dataset collected in mobile network.

%\onote{None of them has referred to the data in mobile network.}

% \onote{Those researchers either utilize the data collected in fixed local-area networks or data crawled from video websites.}

%Arvidsson \textit{et al.} \cite{arvidsson2013analysis} analyzed YouTube user demand patterns and locality properties based on the data collected from two municipal networks in Sweden.

%

%Some studies focus on the network level characteristics of video traffic in mobile networks.

\noindent \textbf{Network level characterization of video traffic in mobile networks: }

Ramos-Munoz \textit{et al.} \cite{ramos2014characteristics} characterized the TCP connections and video streaming of YouTube in a 3G mobile networks, and compared the results with the cases in wired or WiFi networks.

Casas \textit{et al.} \cite{casas2014youtube} collected YouTube traffic in both mobile and fixed networks, and focused on the characteristics of server infrastructure, flow and caching usage. Instead of network level characteristics, we focus on the application level (the users and the videos) characterization of comprehensive online video service in this paper.

%Our work complements these existing works by a provincial-scale dataset \onote{collected} in mobile network and a detailed analysis on the application level of comprehensive online video service.

The most relevant work is \cite{ben2014large}, in which Abdesslem \textit{et al.} studied the characteristics of users, videos and trends of popularity in YouTube in a cellular network.

Compared with their work, our study focuses on a comprehensive online video service instead of the UGC-only service.

We also perform initial analysis in many new facets, such as user behavior in the spacial dimension, small-world phenomenon in user network and growth patterns for new published videos.

* 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

We collect a large-scale traffic trace database for the analysis in this paper.

It's gathered from a major telecom operator at provincial network exports in southeast China on December 24th, 2014 and January 24th, 2015.

%The analysis performed in this paper is based on a large-scale traffic traces dataset collected from a major telecom operator at provincial network exports in southeast China.

The traces consist of the records for each HTTP transaction, including HTTP/TCP protocol headers, timestamp, anonymized user identifier and traffic bytes of the TCP flow.

We further identify the traffic traces related to the Youku video service.

Youku traffic can be identified by the HTTP \textit{host} and \textit{referer} fields ending with ``\url{youku.com}" or ``\url{ykimg.com}'', both of which are Youku server domain names.

And in particular, the requests for videos of Youku have specific URLs, which match the regular expressions ``\textit{\url{v\.youku\.com/v\_show/id\_[A-Za-z0-9=]{13}.\*}}'' or ``\textit{\url{.\*api\.(mobile|3g)\.you\.com/videos/[A-Za-z0-9=]{13}/.\*}}''.

Moreover, the CDN traffic of Youku can be detected by checking whether the URI matches the regular expression ``\url{/youku/A-Z0-9]{26}/.\*}'' or not.

%We also include the CDN (Content Delivery Network) traffic of Youku,

%And video requests of Youku also have specific URLs which match the regular expressions ``\textit{\url{v\.youku\.com/v\_show/id\_[A-Za-z0-9=]{13}.\*}}'' or ``\textit{\url{.\*api\.(mobile|3g)\.you\.com/videos/[A-Za-z0-9=]{13}/.\*}}''.

We converge the Youku traffic according to user identifiers, and filter out the video playback sessions in which users send at least one video request during the measurement period.

Youku assigns each video with a distinct 13-digit ID composed of 0-9, a-z, A-Z and =.

We can extract these video IDs from the URI of a video request.

By sending a video ID to the public API provided by Youku for developers \cite{ykapi}, we can collect the meta-data of the video, including its category, duration, age and etc.

%We collected the traffic traces

Overall, we got over 64 million HTTP transaction records, which contained roughly 3 million Youku HTTP transactions records with 32,516 video requests, 11,065 unique users and 12,193 unique videos.

The data used in this paper consists of three parts: traffic traces, meta-data of the videos in those traffic traces and meta-data of the videos long-term crawled from Youku website.

\noindent \textbf{Traffic traces} were collected by a major telecom operator, at provincial cellular network's exports in northeast China from April 23rd to April 29th, 2015.

These traces contain the URL of around 14 billion HTTP transactions, with time-stamp, anonymized user identifier, cell identifier (LAC-CI code) and user geographic latitude-longitude.

Video request of Youku can be identified by its specific URL, which matches the regular expression ``\textit{\url{v\.youku\.com/v\_show/id\_[A-Za-z0-9=]{13}.\*}}'' or ``\textit{\url{.\*api\.(mobile|3g)\.youku\.com/videos/[A-Za-z0-9=]{13}/.\*}}'', from those HTTP transactions.

Then, video ID, a distinct 13-digit identifier for each video, is extracted from each video URL.

Overall, 362,192 video requests are filtered in total, covering 86,205 unique users and 72,863 unique videos.

%The first part is traffic traces collected by a major telecom operator at provincial cellular network exports in northeast China from April 23rd to April 29th, 2015.

\noindent \textbf{Meta-data of the videos in the traffic traces:}

These data were collected by utilizing the API provided by Youku for developers \cite{youkuapi}.

By sending video ID in a specific formatted HTTP GET request to the API, the video meta-data information will be replied with a JSON object, including duration, category, published date, view count and etc.

We developed a crawler in Python language to automatically retrieve those data.

Since some videos have been deleted by the uploaders at the time of query, meta-data for 71,898 videos are collected eventually.

\noindent \textbf{Long term crawled meta-data for a set of videos:}

To study the growth trend of view count, we first collected all the 13,123 video IDs of the videos published on August 1st, 2015.

Then, we had used our crawler to track and grab the meta-data for these videos every day, until September 1st, 2015.

The whole collection procedure lasted for 32 days.

In this way, we get the view counts of videos for 32 days from their release date.

* 1. 用户行为特性分析
     1. 业务使用强度

从数据集中，我们可以获取每次视频请求的用户标识和视频ID。对于各个用户，我们将其总共发送的视频请求数，以及按视频ID去重后的播放视频数，作为其使用网络视频业务的强度指标。

%the video number is generally larger than the user number.

%This is as expected, due to one user may watch multiple videos during the 1 hour period.

%To further distinguish the duplicated requests caused by different reasons, we plot the distribution of time interval between duplicated requests sent by the same user in Figure.~\ref{fig\_intvl\_replay} (a).

%As the figure shows, the time interval within 10 seconds accounts for about 55 \%, an order of magnitude higher than the fractions of other interval ranges.

%We find most of these short-interval duplicated requests are retransmissions.

%We regard requests with more than 60 seconds as replay requests, and plot the cumulative distribution function of users against replay rate in Figure.~\ref{fig\_intvl\_replay} (b).

%\noindent \textbf{Duplicated requests and replaying behavior:}

As shown in Figure.~\ref{fig\_time\_of\_day}, the number of requests is much larger than videos or users.

This is largely caused by the duplicated requests for the same video sent by some users.

%It can be noticed from the Figure.~\ref{fig\_time\_of\_day} that the request number is much larger than the video or user number.

%Among these requests we find some users send duplicated requests for the same video.

Those duplicated requests are caused by two key factors: 1) retransmitting packets in bad network conditions, and 2) replaying videos.

The replaying behavior is crucial in reflecting user's preferences.

Figure.~\ref{fig\_replay\_reqvid} (a) illustrates the cumulative distribution of the users over the replay rate.

The replay rate is calculated as the ratio of the number of replayed videos to the number of total videos.

It can be noticed that the replaying behavior lies in about 30\% users.

%We find about 30\% users have the replaying behavior.

And for the ``heavy'' users who watch more than 10 videos, around 86\% of them will replay at least one video.

Over 40\% of the heavy users replay more than 20\% of the videos they watched before.

%We then analyze the request number and video number according to users.

%\noindent \textbf{The distribution of request/video number per-user:}

Figure.~\ref{fig\_replay\_reqvid} (b) shows the distribution of request/video number per-user, and a typical heavy tail scenario can be observed.

Over 70\% users only request 1 video, and

90\% users request less than 4 videos and initiate less than 8 requests.

Meanwhile about 2\% ``heavy users'' watch more than 10 videos.

In Figure.~\ref{fig\_replay\_reqvid} (b) we also plot the proportion of the number of videos/requests over the proportion of users.

It illustrates a slow growth in users for the number of videos/requests.

Top 10\% users account for 40\% of the videos/requests while the top 20\% account for 50\%. Therefore, the Pareto principle \cite{juran1954universals}, which states that roughly 80\% of the effects come from 20\% of the causes, can not be adopted into this situation.

%That is, the famous Pareto principle \cite{juran1954universals}, which states that roughly 80\% of the effects come from 20\% of the causes, does not hold for the requests and videos of Youku in mobile network.

%Overall, the video/request numbers in our Youku workload is quite small, less than video service access via fixed network [][].

%This may be because the traffic in mobile network is charged more expensive than in fixed network.

%Another interesting fact can be noticed is that, after the user proportion larger than a certain value, the distribution curves become straight lines.

%This indicates a linear relationship between users and video/request count when the video/request number is small, as the x-axis is sorted in descending order.

%This is due to the fact that the users with one or two videos/requests account for a large proportion of the total users.

%They contribute the same growth speed to the total videos/requests, hence forming a straight line.

We analyze the distribution of unique videos and video requests according to users, to get the access pattern of Youku.

Figure.~\ref{user\_activity2} (a) and (c) show the cumulative distribution function of the video numbers and request numbers per user.

Figure.~\ref{user\_activity2} (b) and (d) show the cumulative fraction of videos and requests over user proportion sorted in descending order.

%The x-axis shows the proportion of the users sorted in descending order according to the frequency of watching videos or sending requests, and the y-axis shows the cumulative fraction of total videos or requests.

We find the distributions of the two day data are quite similar.

It indicates that there lies a relatively stable access pattern for the videos and the requests.

%And the distribution of requests is proportional to the distribution of videos, which is consistent with the analysis in the subsection 24-hour service usage.

A typical heavy tail scenario can be noticed in Figure.~\ref{user\_activity2} (a) and (c).

Over 60\% of the total users request only 1 video, and over 50\% of the users generate only 1 video request during the measurement period.

Most users (about 80 \%) request no more than 2 videos, and initiate less than 4 requests.

Meanwhile, there are also some ``heavy users'' request more than 10 videos a day.

Compared with YouTube \cite{zink2009characteristics}, the video/request number per user in our Youku workload is much smaller.

This might be because, as mentioned before, in China mobile traffic is charged by byte and quite expensive.

Users will not watch videos very frequently due to the cost.

For the video/request numbers over users in relative scale, as shown in Figure.~\ref{user\_activity2} (b) and (d), the top 10\% users only account for 40\% of the videos/requests while the top 20\% account for a mere 50\%.

The famous Pareto principle, which states that roughly 80\% of the effects come from 20\% of the causes, is generally not applicable for the Youku user access situations.

We infer this phenomenon is caused by the large number of videos and the differences of user interests.

Since there are too many videos in the comprehensive video service, users have more various choices on the videos they watch, which leads to less requests per video.

%As too many videos are provided, the diversities of user interests are magnified, and the choices of videos are translated into fewer specialized requests per video.

Another interesting fact can be noticed is that, after the user proportion larger than a certain value, the cumulative distribution curves become straight lines.

This indicates a linear relationship between users and videos/requests when the video/request number is small, as the x-axis is sorted in descending order.

This is due to the fact that the users with one or two videos/requests account for a large proportion of the total users.

They contribute the same growth speed to the total videos/requests, hence forming a straight line.

In fact, the threshold x value at where the distribution curve becomes a straight line is equal to proportion of the users with one or two videos/requests.

* + 1. 请求时刻

时间维度

%And for the billing plan of cellular data network in China,

Meanwhile, the network operators in China provide monthly limited Internet access plans to users, and charge for the traffic beyond the plans by bytes.

Furthermore, for the non-busy hours, from 11 pm to 7 am, there are low-price plans or billing discounts.

It leads to the situation in which the user will care about the volume and time of traffic consumption while watching mobile videos, thus consequently influences the playback behaviors.

Figure.~\ref{fig\_total\_bar} shows the total number of videos requests, unique users and unique videos for each day in our dataset.

It can be noticed that there is no apparent difference between weekdays and weekends.

%It can be seen that these isn't great difference between weekdays and weekends.

While, the numbers vary dramatically within a day.

The distributions of the hourly average for requests, users and videos of the day are shown in Figure.~\ref{fig\_time\_of\_day}.

There is a major drop in user activity between 22:00 and 4:00, and an obvious increase in the morning between 5:00 and 8:00.

During the daytime, the user activity is relatively steady, with a peak appears at 12:00.

An increase starts from 17:00 and touches the top at 20:00 in the evening.

It can be inferred that numerous activities happen in the evening.

Meanwhile, the highest request number is as much as 8 times of that in the bottom.

The peak time appears at 20:00 while the bottom is reached at 3:00.

%We notice that more activities happen during the evening, and the request number during peak period (20:00) can be as much as 8 time more than that during nadir period (3:00).

Rush hours happen at noon and in the evening, indicating individuals watch mobile videos at their spare time as a recreational activity.

（时间粒度分布）

The total number of video requests varies significantly during both 24 hours.

There is a sharp rise in the morning from 5 am to 7 am, when people wake up and begin to watch videos using their cellphones.

The request number keeps high and steady during the daytime from 7 am to 6 pm, and subsequently followed by a rise with fluctuations at night from 6 pm to 0 am.

The curve reaches the peak at midnight between 11 pm and 0 am.

%The increase beginning from 11 pm is rather sharp.

Then there is a major reduction after midnight from 0 am to 5 am as expected, since most people are sleeping.

The usage of mobile Youku service is quite different from the observations of video services in fixed networks \cite{gill2007youtube},\cite{arvidsson2013analysis} and \cite{ben2014large}, where reduction happens at night and the peak appears at afternoon.

We conjecture this phenomenon is caused by the ``on bed watching'' effect:

At night especially before falling asleep, people tend to shut down their wired devices and watch several videos on bed using cellphones via mobile network.

What is more, the mobile traffic billing plans offer discounts for using cellular data during this period of time.

Hence, users prefer to watch videos with cellphones from the time 11 pm.

Note that the requests for videos always exist during 24 hours a day, even at late night and before dawn, from 0 am to 5 am, there are non-negligible requests observed.

The numbers of unique users and unique videos are roughly proportional to the number of video requests.

The frequency values of unique users are higher than those of unique videos, while the frequency values of video requests are the highest of all.

This is because of the fact that for a given period of time (one hour in our case), one user will watch multiple videos, and send even more video requests.

We further analyze the reasons for the duplicated requests sent by the same user for the same videos.

It can be noticed that most of these requests are caused by: 1) refreshing the browser within a short time, or 2) replaying the same video.

* + 1. 请求位置

空间维度请求特性

（位置分布、小区数、移动模式）

The geographic map of user access activities for different time periods of the day is shown in Figure.~\ref{fig\_geo\_location}.

%With the user geographic information in our dataset, we examine the distribution of user numbers over locations during different periods of day, as shown in

The x-axis is longitude and the y-axis is latitude, in the range of a city in north-east China.

%The x axis and y axis are longitudes and latitudes, in the range of a city in northeast China.

To protect user privacy, we anonymize the real values of the longitude and latitude.

The whole city area is divided into $200\times100$ regions.

User number of each region is monitored every 4 hours.

%We divide the whole city area into $200\times100$ regions, and count the cumulative user number of each region in every 4 hours.

Some ``hot spots'' regions are exposed.

Those regions are with significantly more users than other regions.

The number of users is degrading with the locations going further from the ``hot spots''.

%and the user population radiates out from the hot spots with depletion.

The positions of ``hot spots'' are relatively fixed through the whole day: in the mid-west, north-east and south-east parts of the city.

The concept of access entropy is proposed to measure the heterogeneity of user access locations in the mobile network.

%\onote{To} measure the heterogeneity of user access locations in mobile network, here we propose the concept of access entropy.

For a period of time $t$, the global access entropy $H(t)$ is defined as:

\begin{equation}

\label{eqn\_loc\_entropy}

H(t) = -\sum\_{i=1}^{n}p\_{i}(t)\log{p\_{i}(t)}

\end{equation}

\begin{equation}

\label{eqn\_loc\_p}

p\_{i}(t) = \frac{u\_i(t)}{\sum\_{j=1}^{n}u\_{j}(t)}

\end{equation}

where $n$ is the number of regions and $u\_i(t)$ is the number of users in region $i$ during $t$.

Figure.~\ref{fig\_geo\_location} also shows the global access entropy $H$ of each period for the monitored city.

The entropy is relatively stable between different time periods, from 4:00 to 24:00.

%We find from 4:00 to 24:00 the numerical differences of access entropies are not big.

This further verifies the conclusion that the geographic distribution of users is relatively fixed.

The access entropy between 0:00 to 4:00 is smaller than others.

This is because the user reduction in late night affects the hot spots less than the less-hot spots, thus increases the inhomogeneity.

To measure the mobility of users, we check the switch times of LAC-CI code during the time user watching videos.

LAC-CI code is the identifier of a cell in mobile access network, which will change if the user moves from one cell to another.

%If the LAC-CI code of a user changes, it means \onote{that} the user moves from one cell to another.

%Hence the number of LAC-CI code switching indicates the activity of user moving.

Figure.~\ref{fig\_switch\_count} (a) shows the cumulative distribution of the LAC-CI switch times for all users, switch users (users with at least 1 switch), switch users with more than 5 videos and switch users with more than 10 videos.

It can be noticed that over 90\% of users' LAC-CI codes do not switch, and around 80\% of the switch users' LAC-CI codes just switch once.

This indicates that users barely move long distance while watching mobile videos.

And for the switch users, more videos they watch, more often their LAC-CI codes may switch.

It also shows that about 30\% of the switch users with 10+ videos switch their LAC-CI codes more than 3 times.

Considering that the range of a cell is usually about 1 kilometer, these users have moved several kilometers while watching mobile videos.

It's highly possible that those users have watched those videos in vehicles.

%This is very likely that these users watch videos in vehicles.

Then in Figure.~\ref{fig\_switch\_count} (b), we plot the cumulative distribution of switch count in different periods of day.

The differences between different time periods are quite minimal, indicating that user mobility has little correlation with time.

%We find the difference between different periods are quite small, indicating user mobility has little correlation with time.

%The analysis results of user activity above in both temporal and spacial dimensions have great values to the practical applications.

According to the analysis results of user activity in both temporal and spacial dimensions, some suggestions of cache deploying are given.

First, with the knowledge of user activity at different time of day, cache servers can be introduced for rush hours dynamically and flexibly.

Second, replayed videos should be stored at cache servers or in users' local devices, to avoid transmitting unnecessary repeated video files.

Our analysis results show that caching 20\% of the replayed videos can eliminate over 60\% of the duplicated requests.

Third, since many users only watch a few videos, it's more necessary to provide caches to the ``heavy'' users, to gain the most bandwidth saving possibility.

Finally, as the locations of hot spots are relatively fixed and most users do not move long distance while watching mobile videos, it's effective to deploy proxy cache servers at relevant locations for the hot spots.

% to save the backbone bandwidth and prevent users from access delays.

* + 1. 数据消耗

（字节数、时长）

We make a further analysis of the time and traffic volume that users spend on watching videos, to better understand the characteristics of user consumption.

Figure.~\ref{user\_activity3} (a) shows the cumulative distribution function of the video session duration.

Over 80\% of the users watch videos for less than 20 minutes, while approximate 5\% of the users watch videos for more than 1 hour.

Three reasons cause this situation: 1) most of video sessions are not too long; 2) people mostly watch videos using cellphones in short spare time; 3) the high cost of by-bytes charging policy is another limitation of the session duration.

An obvious sharp increase at 5 minutes of the session duration can be noticed.

This is because that the video length of most short videos in Youku is about 5 minutes.

It can be inferred most short video tend to be watched completely by users.

Figure.~\ref{user\_activity3} (b) shows the total traffic over users in relative scale.

The x-axis shows the proportion of the users sorted in descending order according to the amount of generated traffic, and the y-axis shows the cumulative fraction of total traffic.

There is a large variation for users on the traffic consumption.

A small fraction of the users (about 10\%) contribute to the vast majority of the total traffic (over 80\%).

This result can be generally fitted by the Pareto distribution.

* + 1. 重度用户
  1. 本章小结

We further discuss the implications of our observations. We find potential cache-ability of online video service in mobile network, and novel schemes (such as recommenda- tion) can also be designed according to our analysis results.

(iv) We point out high potential cache-ability of comprehensive online video service in mobile network and novel schemes (such as recommendation) can also be designed according to our analysis results.

%Finally, together with the analysis we discuss the practical applications of our measurement results.

%We find the comprehensive online video service has a high potential cache-ability in mobile network, and novel schemes (such as recommendation) can be designed according to our findings.

Online video service has become prevalent in recent years. A better understanding of the user behavior of such service is crucial for allocating the network resources and adjusting the service design. While there are some measurement studies on the non-mobile video services in fixed network, the usage of video service in mobile network is yet to be explored. In this paper, we present a detailed analysis of the user behavior characteristics of a leading comprehensive online video service, namely Youku, in cellular network. This paper is based on a large-scale data set containing over 17 billion traffic traces, collected from a major cellular network in Northeastern China. We analyze the user behavior from three key aspects: data consumption, service usage, and mobility. We provide an insight into how the mobile video service is utilized by users (especially the heavy users), by measuring the user intensities and various representative behavior features in each analysis aspect, such as active time, replay rate, video category, access location, and residence time. We reveal the patterns of different user behaviors, and discuss the implications for practical application. The findings of this paper can provide direct help for network operators and service providers to improve the network performance and user experience.

This paper has presented the first detailed characterization study of the comprehensive online video service of Youku, a leading comprehensive online video service in China, in mobile network.

Based on large-scale datasets containing about 14 billion traffic traces and over 30 days long-term crawling data, four key aspects have been analyzed: user activity, small-world phenomenon of user network, static video properties and view count of videos.

Our measurements shed light on the characteristics of Youku service usage in mobile network.

We found the promising scalability and cache-ability of the service.

The results presented in this paper are crucial and reliable for both network operators and service providers, to improve the network performance and user experience.

Understanding the traffic characteristics of mobile video service is crucial for both network operators and service providers to allocate the resources and adjust the service.

%Mobile video services in China have some specific features, due to the service type and billing plan.

In this paper we present an in-depth characterisation of the mobile video traffic of a leading service provider in China, namely Youku.

%which is filtered from 64 million traffic traces collected from a cellular network in southeast China.}

Our study is based on over 64 million traffic traces collected from a cellular network in southeast China.

%For both network level and service level characteristics, we analyse three key aspects of the video traffic: protocol fields, user activity and video properties.

We analyse three key aspects of the mobile video traffic: protocol fields, user activity and video properties, in both network level and service level.

We find that many specific features of the mobile video traffic in China are caused by the service type and billing plan.

%We look through the compositions of mobile video traffic through the protocol statistics and explore the causes of the distributions.

We reveal the relatively stable patterns of workload protocol fields and user activities.

%, and investigate the diversity of the requested videos.

%user activities over time, videos, requests and traffic volume in the Youku workload.

We also investigate the diversity of the requested videos according to video category, duration and age.

Our analysis result provides a insight into the usage of Youku mobile video traffic, and also has promising practical applications in network engineering and service design.

In this paper, we have presented a detailed characterization study of Youku workload.

Three key aspects have been analyzed including the protocol fields, the users and the videos.

We first examined the protocol statistics and explore the causes of the distributions.

Then we provided insights into the user activity over time, videos, requests and traffic.

At last, we investigated the diversity of requested videos in category, duration and age.

We reveal the characteristics of video traffic in mobile network in China, and explain the reasons for these characteristics.

The results presented in this paper are important for both network operators and service providers, and we believe that the provided insights can help them improve the network performance and user experience.

1. 综合性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

用户喜好分析重要性：业务、推荐、广告盈利

综合性视频业务

网络视频业务往往产生大量的网络流量并占据大量的网络带宽

Youku (www.youku.com) is a lead- ing online video service in China, with around 500 million monthly active users and around 800 million daily video views [1508-4].

Unlike the traditional single-type video services (e.g. YouTube and Netflix), Youku offers a comprehensive type of service, providing both UGC (User Generated Content) and copyrighted VoD (Video on Demand) videos at the same time, to give the users more choices in video categories.

Nowadays, more and more video portals have upgraded their service to a comprehensive one-stop solution, providing UGC, VoD and Live streaming at the same time.

This one-stop solution offers much more video categories to users and attracts users with various of interests, thus effects the usage of the service.

In the past, most of the online video portals provide only one type of service: either UGC (User Generated Content) such as YouTube, or VoD (Video on Demand) such as Netflix.

With the rapid growth of online video service, comprehensive online video service emerges and becomes more and more popular, to meet the various interests of users.

This kind of video service uses the similar HTTP video delivery technology as YouTube or Netflix, but provides multiple types of video services including UGC, VoD, Live streaming and etc.

Users can upload and share videos, follow the latest TV episodes, or watch live news broadcast using just one comprehensive online video service.

%Now more and more video service providers have switched their service from a single type to a comprehensive type.

Our study is based on large-scale datasets containing 16 billion network traffic traces and long-term crawling data (over 1 month).

In this paper, we characterize the usage of a leading comprehensive online video service in mobile network in China, Youku.

Our study is based on around 16 billion traffic traces and over 1 month long crawled video meta-data.

The traffic traces are collected by a major mobile network operator in a northeast province in China.

The video meta-data are retrieved by our web crawler from the website of Youku.

With these data we provide a detailed and in-depth analysis work on how Youku service is utilized by the mobile users.

%analyze Youku video requests in the traffic traces and the long-term crawled data

To the best of our knowledge, this is the first large-scale study of characterizing comprehensive online video service in mobile network.

1. 新颖的研究对象。我们关注于综合性的网络视频业务中用户的喜好特性。

尚属首次In this paper we present the first characterization work in mobile network on a leading comprehensive online video service in China, Youku.

It complements the previous studies on video services with small-scale, wired network access or single-type (UGC or VoD only) datasets.

It complements previous studies on wired Internet access and single type video service, utilizing small scale datasets.

(iii) We also look through the interests of users, according to their service level behaviors. On an individual level, we study the interest clusters of users; and on a global level, we analyze the interest network of users. We reveal the impacts of user interests on the resource consumptions.

尤其是，(ii) We demonstrate a small-world phenomenon in the user network by studying the relationship between users.

%Second, we study the relationship between users and find a small-world phenomenon in the user network.

* 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. 用户喜好特性分析
     1. 用户活跃度
     2. 视频属性

视频类型

We also investigate the static properties of Youku mobile videos and observe noticeable differences from other UGC videos. %compared with the UGC videos.

To figure out the diversities of the videos watched by users, we further investigate the categories of the videos.

%Youku categorizes the videos on its website into 25 categories.

By crawling the video meta-data we get the category information of each requested video in our dataset.

The percentage of video number for the 12 most popular categories are shown in Figure~\ref{pie}.

% from the pie chart that the category account for the largest fraction , followed by , .

%This seems a little counter intuitive, as TV Episodes are usually long and large video files.

%We didn't expect users choose to watch so many TV Episodes using cellphones via mobile network.

The top 2 categories, `TV Episode'' (28.46\%) and ``Animation'' (9.14\%), account for over 30\% of the total requests, which indicates that following the up-to-date episodes of TV or animation series is an very important reason for users to watch mobile videos.

``Music'' (7.93\%) and ``Film'' (7.54\%) hit the third and fourth most popular respectively.

It can be noticed that all the categories above correspond to the VoD videos.

Hence, service providers should keep purchasing the copyrights of such contents and provide them free of charge in the service to attract users.

And the distribution of video numbers in other categories is quite uniform.

We also notice among the top 12 categories, only 1 category (News) is for obtaining information.

It indicates that most users watching mobile videos are just for pleasure.

用户兴趣聚类

视频长度、年龄

With the API provided by Youku, we collect the meta-date for all videos in our traffic traces dataset.

In order to figure out what kind of videos are widely watched by mobile users, we examine the distribution of video category, duration and age.

%Some of the results are rather contrary to intuition.

Youku categorizes its videos into 25 categories.

We list the top 12 most viewed categories and their proportions in Table.~\ref{table\_category}.

These categories cover 85\% of the total requested videos.

As shown in the table, the most popular category is ``TV Episode'' (19.30\%), which is requested more than two times of the second most popular category ``Music'' (9.14\%).

This is very interesting and unexpected as TV Episodes are usually long and large video files.

Even in mobile network using cellphones, users still choose to watch so many TV Episodes.

This proves {Youku's strategy} to attract users by providing copyright contents is quite successful.

For the rest 10 categories, the proportion differences are not so big.

Another interesting fact can be noticed is that among the top 12 categories, only 1 category (News) is for obtaining information.

It can be inferred that users watch mobile videos are just for entertainment and relaxation.

Figure.~\ref{fig\_duration\_age} shows the histogram and cumulative distribution of video duration and video age.

Video age is defined as the number of days between the release date and the viewing date.

The mean and median of video duration and age in each category are also shown in Table.~\ref{table\_category}.

It can be noticed that Youku videos are comprised of both short videos and long videos, and videos with duration less than 1 hour account for about 90\%.

We observe three obvious peaks in the video duration distribution in Figure.~\ref{fig\_duration\_age} (a).

The first peak appears within 5 minutes and contributes about 50\% of the videos, which are formed by a large number of short videos across different categories.

The second peak is reached close to 25 minutes and contributes about 10\% of the videos, which is mainly formed by videos from the Animation category.

And the third peak is near 45 minutes and accounts for around 20\% videos, corresponding to a large number of TV Episodes and Show videos.

Compared with UGC video service (such as YouTube) \cite{gill2007youtube, abhari2010workload, cheng2013understanding}, more long videos are watched by users in Youku.

Most of those videos are contents with copyrights for the VoD service such as TV Episodes, Movies, Variety Shows and Animations.

It can be noticed that ``young'' videos account for a large fraction of the total videos, as shown in Figure.~\ref{fig\_duration\_age} (b).

%For the video age as shown in Figure.~\ref{fig\_duration\_age} (b), we find the ``young'' videos account for a large fraction of the total videos.

About 35\% of the videos are less than 30 days old, and half of the videos are less than 5 months.

%And from Table.~\ref{table\_category}, we learn

The median video ages in Game, Show, and Entertainment categories are much smaller than those in other categories (several days v.s. more than a year), as shown in Table.~\ref{table\_category}.

This indicates that the videos in these categories update in a high frequency and keep attractive only for a short period of time.

We check some ``young'' videos in the Game category and find they are weekly released video commentaries for the computer game DOTA.

%We carefully look through the distributions and relationships of the static videos properties above.

From the analysis of static video properties above, we get results which are useful in practical applications.

%The results we get have practical significance.

It's crucial for service providers to learn the categories and age of the videos that are watched by user, in order to adjust the contents they provide to attract more users.

%For instance, it is of great importance for service providers to learn what categories and how old the videos watched by users are, in order to adjust the contents they provide and attract more users.

For network operators, a better understanding of the video duration can provide help to the design of cache mechanism.

In addition, we analyze the duration and the age of the videos that users watch.

The histogram and CDF of video duration and age are shown in Figure.~\ref{video\_info2}.

%Table.~\ref{video\_info} also shows the mean and median duration and age of each category videos.

Figure.~\ref{video\_info2} (a) reveals about 90\% videos are within 1 hour long.

There are two obvious peaks for the video duration distribution.

The first peak is within 5 minutes, because of the large number of short UGC videos.

The second peak is close to 45 minutes, which is mostly caused by the VoD videos in the TV Episode and Show categories.

It also depicts that there are many videos with durations close to 25 minutes.

Most of them are animations.

Hence, we find the comprehensive service type has a great impact on the distribution of video durations.

Figure.~\ref{video\_info2} (b) shows the distribution of video age.

Video age is the number of days between the publication date and the view date.

It can be seen that 25\% of the videos are less than 30 days old at the time of request.

Half of the videos are younger than 9 months.

And most of the videos (about 80\%) are no more than 3 years old.

We find the video age in mobile network is younger than that in fixed network, comparing with the results of \cite{gill2007youtube}.

%same {large scale}

This indicates, when watching videos in mobile network, users prefer the up-to-date videos.

These videos keep attractive only for a short period, and replaced by the later published videos.

* + 1. 视频流行度

View count is an important dynamic property, as it reflects video's popularity and access pattern.%the popularity and access pattern of a video.

The popularity of many kinds of web content follows the Zipf's law \cite{adamic2002zipf,ben2014large}.

The Zipf distribution is shown in Eq.~\ref{eqn\_zipf}.

%Zipf distribution, the frequency of content against the rank in decreasing order should follows:

\begin{equation}

\label{eqn\_zipf}

f(k;s,N)=\frac{1/k^s}{\sum\_{n=1}^N (1/n^s)}

\end{equation}

Where $f$ is the access frequency of content; $N$ is the total number of the content; $k$ is the rank; and $s$ is the value of the exponent characterizing the distribution.

This function follows a straight line when plotted on a log-log scale axis.

In this subsection, we analysis the view count of Youku videos in our dataset, and check if it follows a Zipf distribution.

Figure.~\ref{fig\_view\_count} shows the view count against rank on a log-log scale.

It can be noticed that the curve is strongly skewed: the beginning is dropping linearly while the heavy tail decreases dramatically.

We run a regression using the Zipf model, as shown with the red line in Figure.~\ref{fig\_view\_count}.

The regression curve greatly deviates from the actual distribution situation, indicating the linear Zipf model does not fit the distribution well.

This is because in the case of Youku, unpopular videos are not watched as many times as the Zipf model predicts.

%in the case of Youku, there are not so many unpopular videos watched by users as the Zipf law predicts.

We try to model the distribution of view count $f(k)$ with a piecewise function:

\begin{subnumcases}{f(k)=}

\label{equ\_hr}

\alpha \cdot k + \beta \quad \quad \quad f \geq V\\

a \cdot e^{b \cdot k} + c \quad \quad f < V

\end{subnumcases}

where $k$ is the rank; $\alpha, \beta, a, b, c$ are coefficients; and $V$ is the point of demarcation.

The linear Zipf model is kept for the beginning part when the view count is large, and an exponential distribution is used to fit the heavy tail.

The regression curve using our model is shown in Figure.~\ref{fig\_view\_count} with the green line.

It's obvious that our model can fit the practical distribution more favorably and properly than the Zipf model.

* + 1. 用户关系

上传者影响力

用户喜好网络

We analysis the network formed by video-watching users.

%We next examine the network among users when watching videos.

Given all the users as vertices, if user $a$ and user $b$ watch the same video, then there is an edge connecting $a$ and $b$.

Eventually, all the users and the links between them form a network.

Noticeable small-world characteristics are lying in this network.

%We find an interesting fact that this network has noticeable small-world characteristics.

A small-world network is a type of mathematical graph in which most nodes, who are not neighbors, can be reached from other nodes by a small number of hops.

Small-world networks can be identified according to two independent graph structural features: clustering coefficient and characteristic path length \cite{watts1998collective}.

A graph $G=(V,E)$ consists of a set of vertices $V$ and a set of edges $E$.

The average clustering coefficient $C\_G$ and the average characteristic path length $L\_G$ are defined as:

\begin{equation}

\label{eqn\_c}

C\_G = \frac{1}{n}\sum\_{i=1}^{n} \frac{\lambda\_G(v\_i)}{\tau\_G(v\_i)}

\end{equation}

\begin{equation}

\label{eqn\_l}

L\_G = \frac{1}{n(n - 1)}\sum\_{i \ne j} d(v\_i, v\_j)

\end{equation}

where $n$ is the number of vertices in $G$; $v\_i$ and $v\_j$ are vertices in $V$.

$\lambda\_G(v\_i)$ denotes the number of triangles on $v\_i$ in $G$.

That is, the number of subgraphs of G with 3 edges and 3 vertices, one of which is $v\_i$.

$\tau\_G(v)$ denotes the number of triples on $v\_i$ in $G$.

That is, the number of subgraphs with 2 edges and 3 vertices, one of which is $v\_i$.

$v\_i$ is incident to both edges.

$d(v\_i, v\_j)$ denotes the shortest distance between $v\_i$ and $v\_j$.

If $v\_j$ cannot be reached from $v\_i$, then set d$(v\_i, v\_j) = 0$.

The existing of small-world phenomenon can be measured by comparing its clustering coefficient $C\_G$ and characteristic path length $L\_G$ with an equivalent random network with same degree distribution:

if the $C\_G$ is significantly higher than that of the random graph while the $L\_G$ is approximately the same with the random graph, then the small-world phenomenon indeed lies in this network.

We measure the user network with 7 randomly selected subsets, consisting of different numbers of videos.

Figure.~\ref{fig\_smallworld\_cl} shows the clustering coefficient and the characteristic path length of the user network and its equivalent random graphs over the size of dataset.

As shown in the figure, the clustering coefficient of Youku user network is much higher than that of the equivalent random graph, while the characteristic path lengths of the two networks are approximately the same.

It can also be noticed that with more users in the dataset, the clustering coefficient of the user network increases, meanwhile the small-world phenomenon is more obvious.

A visual illustration for a part of the user network (3000 nodes) is shown in Figure.~\ref{fig\_smallworld\_eg}.

The small-world characteristics can be clearly observed: many triangles and cliques are lying in the graph.

The small-world phenomenon suggests Youku users have strong correlations with each other on video watching.%while watching videos.

%\onote{It's reliable and efficient to improve the caching efficiency and develop novel recommendation schemes according to this.}

This creates great opportunities for improving cache efficiency and developing novel recommendation schemes.

For instance, if user $a$ and user $b$ watch some same videos, and user $b$ and user $c$ watch some same videos, then the service provider can recommend hot videos watched by user $c$ to user $a$.

Since according to the small-world phenomenon, user $a$ and user $c$ are highly possible to have same preference on videos.

* 1. 本章小结

1. 网络视频播放量动态分析及预测
   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. 播放量动态分析

群体角度：分布

单体角度：增长趋势

* 1. 未来播放量预测

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. 未来播放量级别预测
    2. 基于播放量级别转移的未来播放量数值预测
    3. 基于播放量增长模式的未来播放量数值预测
  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]