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

* 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。

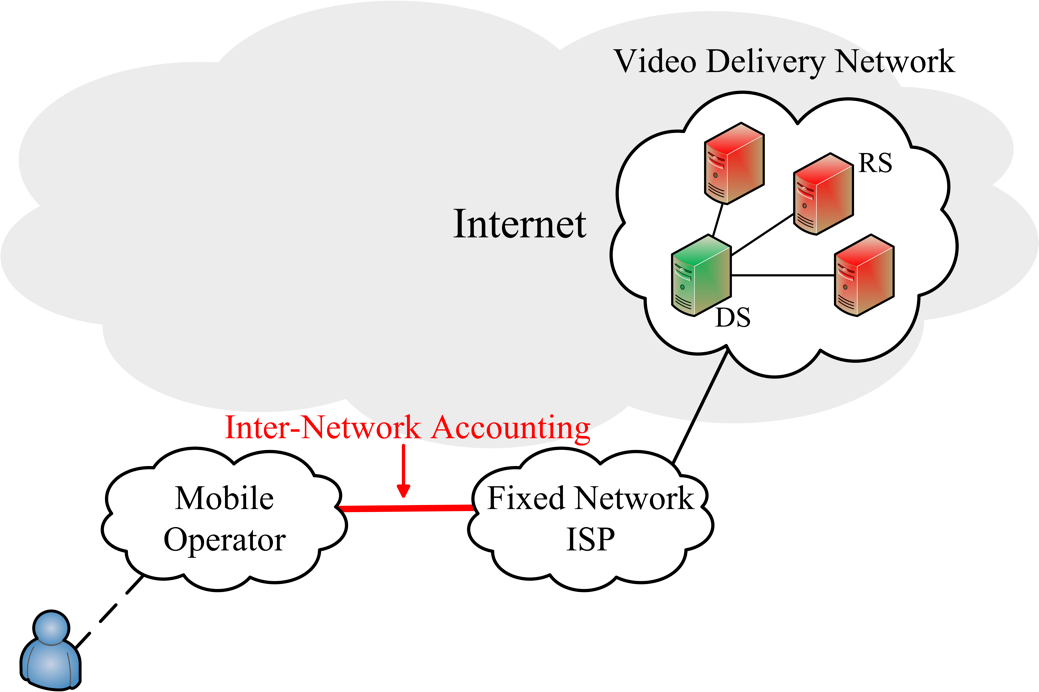


图3-1

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。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。对于采集到的流量报文话单，我们进一步过滤出HTTP协议报文，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，并最终生成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。

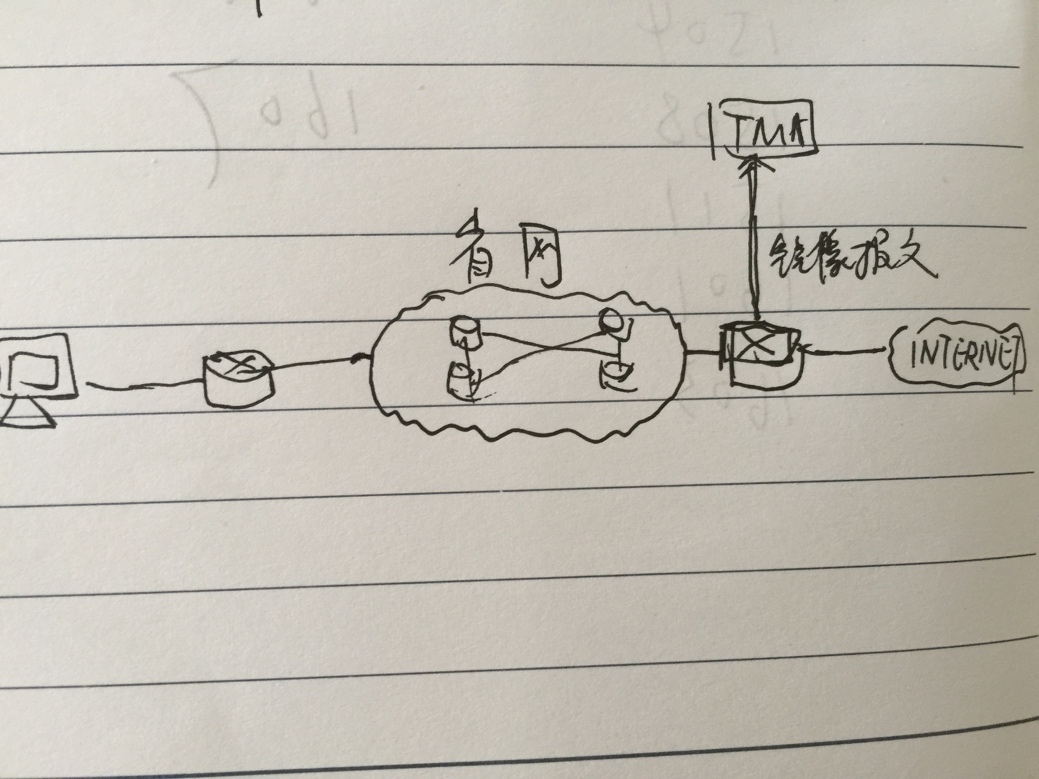


图3-2 数据采集部署示意

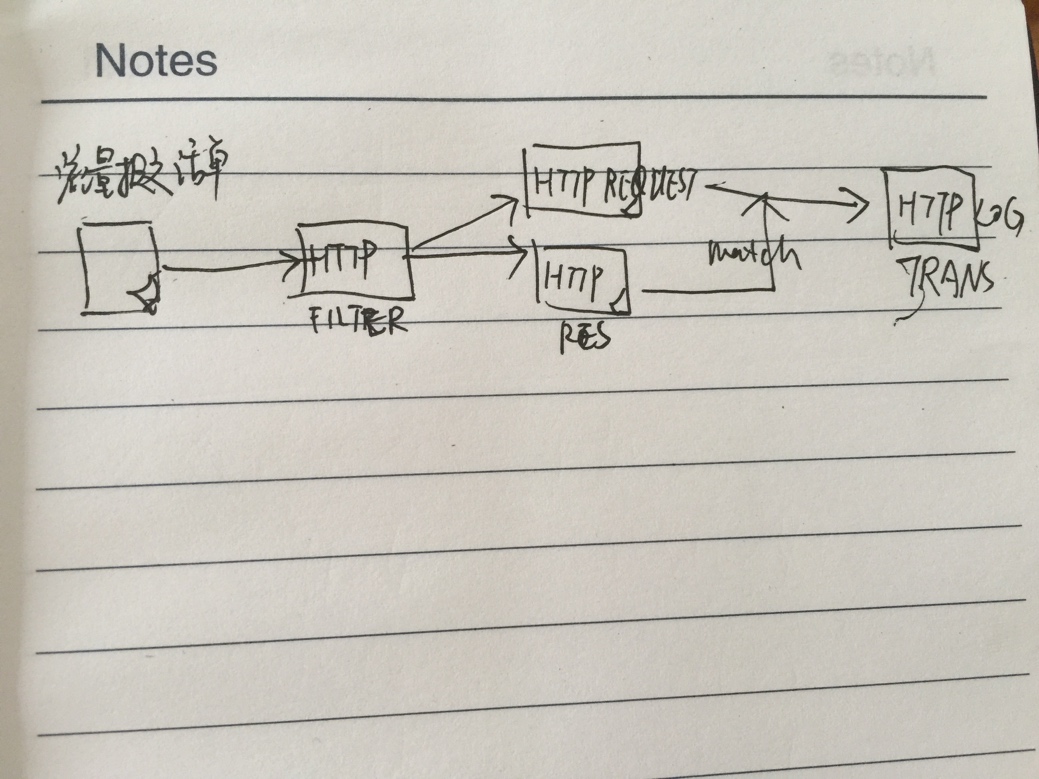


图3-3

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匹配一个视频页面的格式，则说明该用户开始使用网络视频业务，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。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的分析由人工进行，并需要一定的领域知识。最终，我们成功提取出3,410,383条对应分发服务器的HTTP记录，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** | **HTTP记录数** | **调度服务器数** | **资源服务器数** |
| Youku | 1,991 | 2,425,879 | 21 | 737 |
| Sohu | 1,093 | 266,028 | 8 | 119 |
| Iqiyi | 1,295 | 267,442 | 19 | 46 |
| QQ | 1,338 | 254,316 | 3 | 321 |
| Letv | 1,367 | 196,718 | 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，目前尚未有公认的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自定义的用户与服务器之间的通信流程，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。

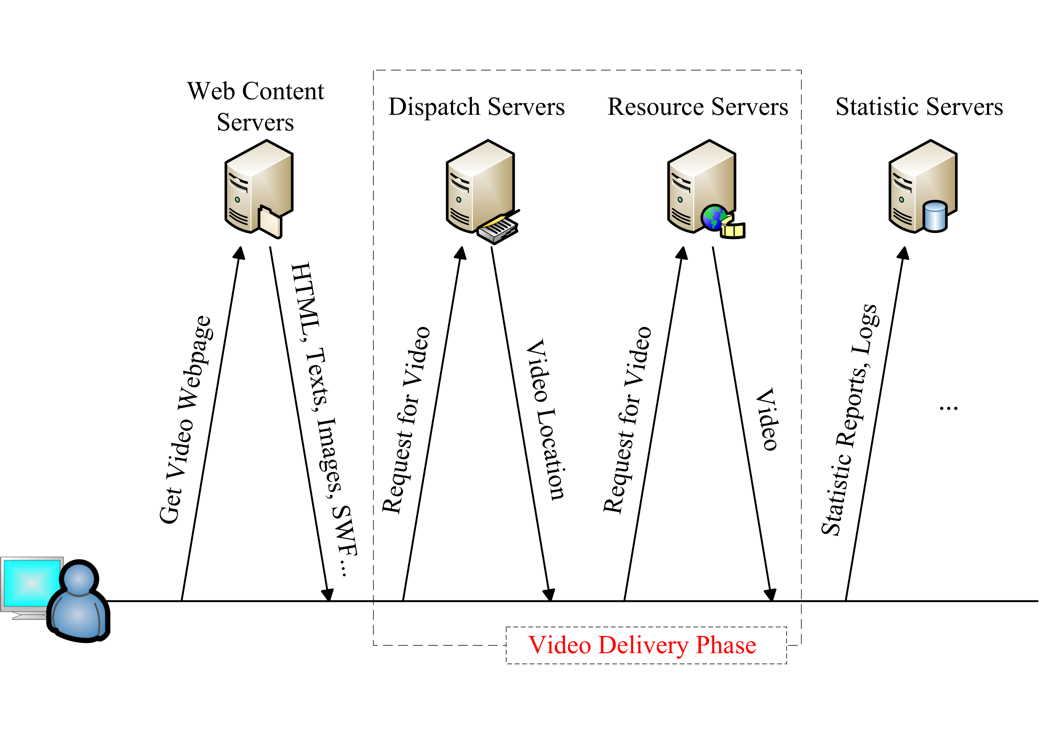


图3-4

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。该播放器用于视频获取、用户操作以及信息交互。

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。

* + 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，大多数都小于10，同时大多数都小于1秒。这一现象符合我们的预期：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。

**特性的实体内容类型**：

* 1. 系统架构

模块介绍

* 1. 实验及结果
  2. 本章小结

The delivery servers in those networks are crucial for network operators to 过顶业务的流量管控optimize the infrastructure deployment and reduce unnecessary expenses.

Our study is based on a large-scale dataset containing over 5 millions of HTTP traces, collected by a major cellular network operator in a northeastern province of China.

We analyze the delivery traffic and detect the delivery servers for five leading online video service providers in China.

In this paper, based on over 5 million network traces collected from a cellular network operator in China, we study in-depth the user-server commutations of the online video service, and propose an effective method to tackle the problem of delivery server detection.

We further evaluate our approach on the real-world traffic data and achieve outstanding performance: the precisions of the prediction can reach as high as 100\%, and the recalls are over 85\%.

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]