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述
   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）

The records of Youku video service can be further extracted, according to the HTTP URL: 1) belonging to a Youku domain (‘‘youku.com’’ or ‘‘ykimg.com’’); or 2) in the form of a Youku CDN (Content Delivery Network) URL, which matches the regular expression ‘‘[0-9]{1,3}\\.[0-9]{1,3}\\.[0-9]{1,3}\\. [0-9]{1,3}/youku/[A-Z0-9]{26}/.\*’’. And video ID, a distinct 17-digit identifier for each video, can be extracted from the URL of a video request, which matches regular expressions ‘‘v\\.youku\\.com/v\ \_show/id\_[A-Za-z0-9=]{17}.\*’’ or ‘‘.\*api\\. (mobile|3g)\\.youku\\.com/videos/[A-Za-z 0-9=]{17}/.\*’’.

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}/.\*}}''.

各个运营商：域、视频页面

优酷：报文分析

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。

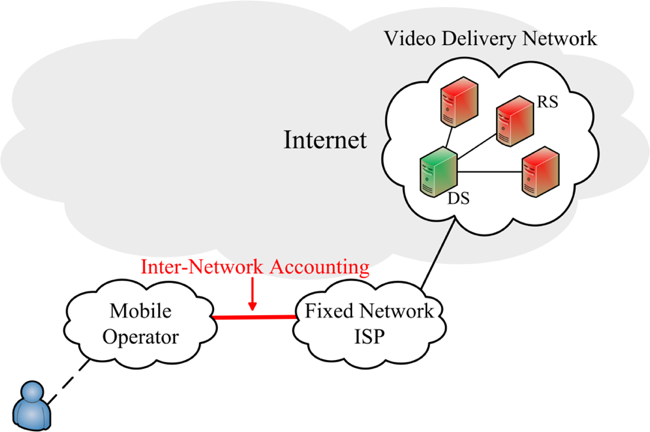


图3-1

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。

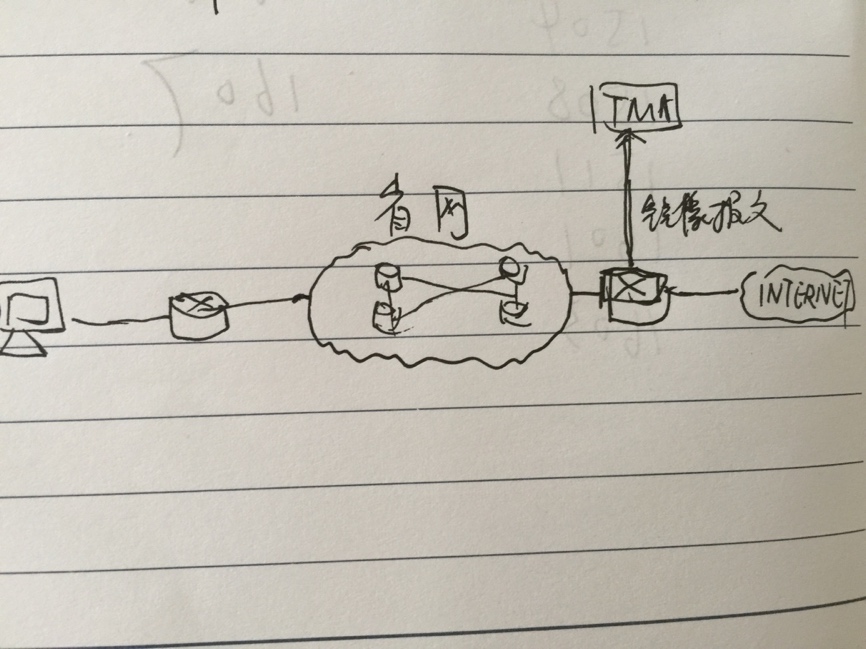


图3-2 数据采集部署示意

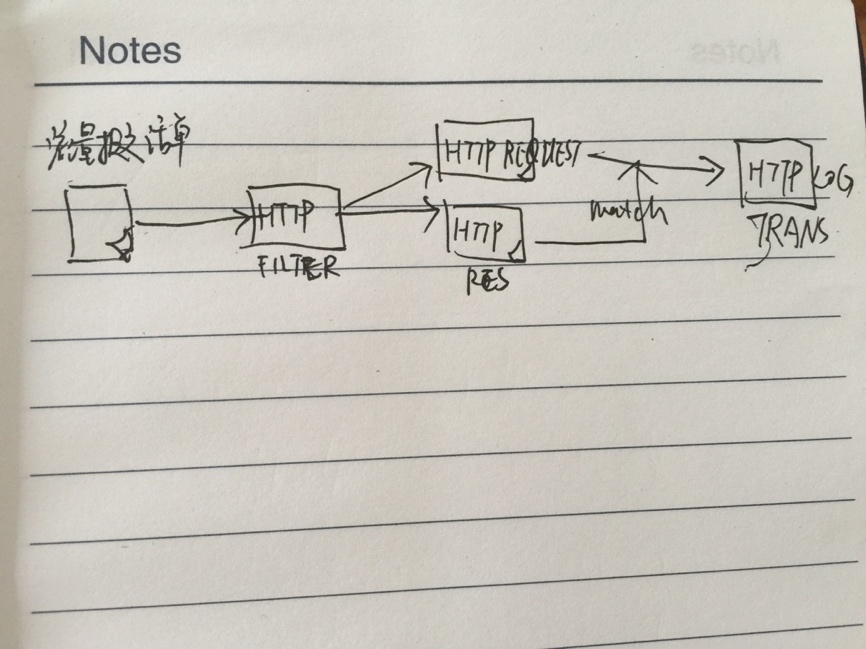


图3-3

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。

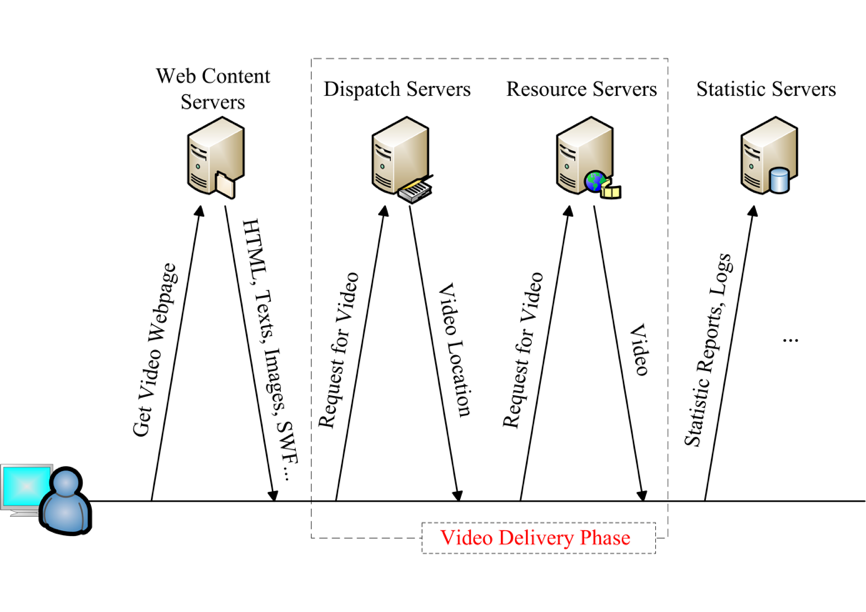


图3-4 通用的用户-服务器通信流程。

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(a)、(b)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 不同网络视频业务服务器传输的HTTP实体内容类型分布：

(a)调度服务器；(b)资源服务器。

**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；和分别表示和中的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。本章研究中，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% | - |

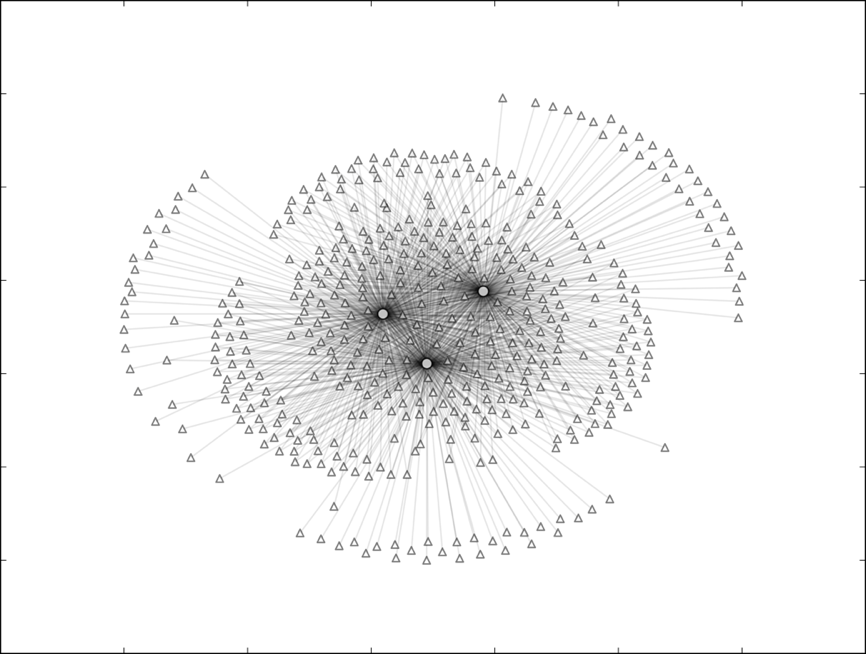


图3-10

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）[12]和决策树（decision tree）方法[13]。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[1]中指出，2015年全球的网络视频流量占据互联网流量的70%，并将于2020年突破82%。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[14]。2015年全球整体的IP流量中，固定网络流量占据52%，而移动网络流量占据48%。预计到2020年，移动网络流量的比例将超过三分之二[1]。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[15]。

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，从中提取了数据传输级别、无线接入级别和的业务级别的用户信息，并对Youku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数据消耗、位置移动和业务使用三个分析角度，分别对应着用户在核心网、无线接入网以及业务提供商的资源占用。尤其是，基于移动网络中的位置信息，我们对用户行为从空间的维度上展开了研究，这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。在对用户的行为强度分析中，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。

* 1. 研究现状

对于网络视频业务的用户行为的分析，目前学术界中已存在一些工作。其中，早期的研究工作主要是基于从固定网络环境中采集的数据，以分析网络视频业务是如何被固网用户所使用的，例如：作者Yu等人在文献[16]中从某固网运营商处获取数据，对该运营商旗下的某大型网络视频点播系统进行了深入研究，分析了视频用户的访问模式、到达率、会话长度以及播放数量等内容。在文献[17]中，作者Gill等人在某校园的网络出口处采集流量数据，并分析了校园用户使用YouTube视频业务时的访问模式、文件属性、引用关系以及传输行为的特性。在文献[18]中，作者Zink等人也从某高校的校园网中采集了HTTP流量数据，并对其中用户访问YouTube的文件属性、播放数量及访问模式进行了分析。作者Arvidsson等人在文献[19]中基于从瑞典某两个城市的网络出口处采集的网络流量数据，对YouTube视频业务的用户请求、视频调度、流量消耗、终端类型以及缓存部署进行了详细的分析。而在本章中，我们面向移动网络中视频业务的研究，是对上述这些工作很好的补充与拓展。

近年来，随着移动视频的快速发展，研究者们开始将注意力转向对移动网络中网络视频业务的流量分析，并取得了若干研究成果，例如：在文献[15]中，作者Ramos-Muñoz等人分析了用户在3G网络中YouTube流量的TCP连接与传输等方面的特性，并与固定网络中的YouTube流量特性进行了对比。在文献[20]中，作者Casas等人从移动网络和固定网络中同时采集了用户使用YouTube视频业务所产生的流量，并对服务器、流量及缓存特性进行了详细的分析。作者Li等人在文献[21]中基于从网络视频业务提供商PPTV处获取的用户访问日志，对使用移动设备和非移动设备的用户的播放行为和活跃度模式进行了对比性的分析。相较于上述这些研究工作，在本章中我们的研究覆盖了更广泛深入的分析角度。我们不仅分析了网络传输级别（Network Level）的特性，还从应用级别（Application Level）对用户的业务使用情况进行了分析。此外，我们的研究是基于大规模海量互联网流量的数据集，对上述使用小规模数据的研究工作形成了很好的补充与拓展。

* 1. 数据集

本章的大规模研究数据采集于某网络运营商在我国东北某省的2G/3G移动通信网络中。图4-1显示了用户使用移动视频业务时的整体网络结构示意图。从中我们可以看出，该网络中主要有四个关键组件：用户设备、无线接入网、核心网和业务提供商。当使用移动网络视频业务时，在无线接入网中用户的移动终端设备将直接与BTS（Base Transceiver Station）或Node B进行通信。网络流量信息进而被传输至BSC（Base Station Controller）或RNC（Radio Network Controller）处。这两种控制器将数据发送至核心网中的SGSN（Serving GPRS Support Node）节点。SGSN通过Gn接口与GGSN（Gateway GPRS Support Node）建立连接，并将流量数据传输至GGSN。通过GGSN，数据进入了互联网并最终到达互联网业务提供商处。



图4-1 示意图

我们的流量数据采集器部署于该移动通信网络运营商在东北某省的核心网Gn接口处。网络流量被实时镜像至采集器处。采集器高速解析报文，并生成话单。我们进一步匹配HTTP协议的请求应答对，并对每一次HTTP交互生成一条记录。我们使用Apache Hadoop框架[22]的HDFS（Hadoop Distributed File System）来存储大规模的话单记录数据。HTTP记录的字段包括：时间戳、HTTP请求URL、经过匿名处理的用户标识、移动通信网络的小区标识（LAC-CI）、对应TCP流的上下行字节数。除此之外，网络运营商还向我们提供了其移动通信网络中每个小区（cell）的信号发射塔经纬度。

整个数据采集阶段从2015年8月1日起至2015年8月10日止，共计10天。 最终，我们一共采集到17,570,755,031条 HTTP 记录。基于前文 X.X分析结果，我们通过HTTP请求URL进一步过滤出优酷视频对应的HTTP记录，共计37,570,167条。对于每条视频请求，我们可以从其URL中提取出视频ID。我们将视频ID作为优酷视频的唯一标识，计算去重后的视频数。最终，我们数据集的概览如表4-1所示。

表4-1 优酷视频HTTP记录数据集概要统计

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **日期** | **记录数** | **用户数** | **视频请求数** | **去重视频数** |
| 2015-08-01 | 3,904,472 | 8,952 | 28,209 | 10,278 |
| 2015-08-02 | 4,051,733 | 8,755 | 29,041 | 10,639 |
| 2015-08-03 | 3,958,286 | 9,673 | 30,127 | 10,240 |
| 2015-08-04 | 3,954,113 | 8,981 | 26,631 | 9,541 |
| 2015-08-05 | 3,705,146 | 9,267 | 28,203 | 9,639 |
| 2015-08-06 | 3,481,995 | 8,316 | 25,547 | 9,301 |
| 2015-08-07 | 3,539,899 | 8,553 | 27,034 | 9,976 |
| 2015-08-08 | 3,622,581 | 8,542 | 26,486 | 9,988 |
| 2015-08-09 | 3,542,081 | 8,775 | 28,097 | 10,379 |
| 2015-08-10 | 3,809,861 | 9,539 | 28,821 | 9,879 |

* 1. 用户数据消耗特性分析
     1. 流量字节数

首先，我们对用户的流量字节数进行分析。对于一个用户而言，其使用网络视频业务时所产生的流量大小显示了该用户的数据消耗强度。图4-2给出了我们数据集中用户流量字节数的累积分布函数。为了便于阅读，我们将图的横轴设置成对数刻度。从图中我们可以发现，用户流量消耗的分布是十分不均匀的。对于大多数用户而言，所消耗的数据流量是比较小的。80%的视频用户在10天内总共仅消耗了2MB左右的流量来观看视频。这些用户所产生的流量仅占数据集中总视频流量的6.3%。而与此同时，确实存在某些用户，能够消耗多达若干GB的移动数据流量来观看网络视频。相较而言，这些大流量用户会占据更多的网络传输资源。因此，在分配网络设施和调整业务设计时，网络运营商和业务提供商应对这些用户进行重点考虑。在此情形下，使用合适的方法检测出这些重度用户并对其进行专门的分析，将是十分重要而有实际意义的。



图4-2 xxx

* + 1. 重度用户检测

对于重度用户的检测问题，一个简单的方法是设置一个阈值，然后将所有考察指标超出该阈值的用户判定为重度用户。通常，阈值的具体数值是根据研究者经验来设置的，这显然存在着一定的独断性。另外，人工设置的阈值往往具有较弱的通用性，即对某个数据集合适的阈值可能并不适用于对其他的数据集。为解决这些问题，在我们的研究中，我们基于洛伦兹曲线（Lorenz curve）[23]提出了一个通用的非参数的重度用户检测方法。

洛伦兹曲线是对累积分布一种图形表示方法，其横轴为统计对象比例，纵轴为统计量累积分布占比。以用户的流量字节数为例，为了建立洛伦兹曲线，我们首先将用户按流量字节数进行升序排列。令为总用户数，为用户排序后序号，表示用户消耗的流量字节数，我们有。然后对于横坐标的点，确定其纵坐标，其中：

(4-1)

这些点形成的曲线即位洛伦兹曲线。根据定义可知，横轴与纵轴的范围都是。而我们的重度用户方法的思路，是希望在洛伦兹曲线横轴上找到一个合适的数值，来对应非重度用户的比例。则剩下占总用户数的用户为重度用户。对于的取值我们考虑两种极端情况：和。其中，是消耗流量字节数等于均值的用户所对应的横坐标值，即：

(4-2)

也就是说，只要用户消耗的流量字节数大于整体的平均值，即可被判为是重度用户。这种划分方式的限制非常的弱，是一个用户成为重度用户所需满足的最低标准。因此，我们将作为的取值的下界。而则是洛伦兹曲线在点处的切线与横轴的交点，如图4-3所示。一般来讲，洛伦兹曲线是凹函数，因为用户是按消耗流量字节数升序排列的，即的增长速度是越来越大的。当曲线上的点横坐标从开始增长，其对应的切线的斜率也在不断增长并在点处达到最大值。因此，我们将设定为的取值的上界。并且，用户的流量分布越不均匀，洛伦兹曲线就会越凹，点处切线的斜率就会越大，进而的数值就会越大。此外，如果用户的流量分布是指数形式的，其中为标度参数，则我们的方法可以得到。我们综合考虑上述的两种取值限制，最终定义。我们的检测方法受文献[24]启发，其中作者Thomas Louail等人使用的定义方法来检测地图中的热点区域。



图4-3 洛伦兹曲线上重度用户检测判决准则示意图。

使用上述方法，我们共从数据集的74,928位移动视频用户中检测出3,921位数据消耗的重度用户。这些重度用户仅占总用户数的5.23%，却消耗了79.61%的流量字节数。这进一步验证了前一小节的分析结果：移动视频用户的数据消耗是极不均匀的，大多数用户只产生了一小部分流量，而少量重度用户则产生了大部分的流量。

* + 1. 活跃时长

除了流量字节数这一指标，了解一个移动用户花费多长时间来使用网络视频业务，对于用户的数据消耗分析也是至关重要的。在本小节中，我们主要关注两个时间维度上的用户活跃指标：会话时长与业务总时长。会话时长指的是从用户在某一次使用网络视频业务时所消耗的时间长度。具体来讲，当用户首次与优酷的服务器产生通信，我们认为该用户的（第一个）会话开始；当该用户在较长的时间内不再与优酷服务器发生交互，则用户本次会话结束。我们将会话中（与优酷服务器通信的）首末报文的时间差作为本次会话时长。若一段时间之后用户再次与优酷的服务器产生了通信，我们则认为该用户发起了下一次会话。最后，我们将用户所有会话的时长总和定义为其业务总时长。为了选择一个合适的会话超时时限，我们使用从1分钟到30分钟的不同数值来计算我们数据集中的视频会话数量，如图4-X所示。可以看出，随着超时时限的增大，视频会话数不断减少。当超时时限数值较小时（小于8分钟），会话数的减少十分剧烈；而当超时时限在较大的区间变化时，会话数的变动幅度较小。因此，在我们的研究中，我们设置超时时限为10分钟，其在图中对应的会话数曲线比较稳定。



图4-X 不同超时时限下数据集中的视频会话数。

图4-4显示了我们数据集中所有用户和重度数据消耗用户对应的会话时长与业务总时长的累积分布函数。总体来看，大多数用户的会话时长较短。55%的用户会话时长少于5分钟，超过80%的用户会话时长少于20分钟。这符合我们的预期，因为多数用户使用移动设备通过移动网络观看视频的行为发生在他们的零散的空闲时间。此外，电池电量与流量计费等因素也会对移动用户观看网络视频的时长产生限制。而对于业务总时长，我们可以看出其分布曲线与用户会话时长对曲线非常类似。这是由于大多数用户的会话数往往较少。图4-5显示了用户的视频会话数累积分布函数。我们可以看到，82%的用户只发起了一个视频会话， 96%的用户会话数都小于等于3。而对于重度用户，我们发现其会话时长和会话数往往较大。超过40%的重度用户会话时长都在20分钟以上。超过一半的重度用户具有多个视频会话，20%的重度用户甚至具有4个以上的视频会话。因此，重度用户最终往往具有较大的业务总时长。50%的重度用户会花费超过40分钟使用移动网络视频业务。对于所有重度用户，平均业务总时长高达70.71分钟。



图4-4 所有用户与重度数据消耗用户的(a)会话时长与业务总时长累积分布函数、(b)会话数累积分布函数。

我们进一步对移动视频用户的活跃时长分布函数进行了研究。在比较了多种假设的分布形式，我们发现用户的会话时长分布与业务总时长分布都可以使用Pareto Type 2分布[25]来拟合。Pareto Type 2分布的概率密度函数形式如下：

(4-3)

其中形状参数，尺度参数。表4-2给出了对数据集拟合的分布模型参数。图4-6和图4-7给出了对四种情形的拟合结果Q-Q图。可以看出，各图中的散点都大体分布在直线上，这表明Pareto Type 2分布是对用户活跃时长分布的一个很好的近似。

表4-2 Pareto Type 2分布模型参数

|  |  |  |
| --- | --- | --- |
| **统计指标（用户类型）** |  |  |
| 会话时长（所有用户） | 2.50 | 1170.63 |
| 会话时长（重度用户） | 3.27 | 3839.94 |
| 业务总时长（所有用户） | 2.21 | 1295.16 |
| 业务总时长（重度用户） | 2.64 | 6959.81 |



图4-6 会话时长实际分布与Pareto Type 2分布Q-Q图：(a)所有用户，(b)重度用户。



图4-7业务总时长实际分布与Pareto Type 2分布Q-Q图：(a)所有用户，(b)重度用户。

* 1. 用户位置移动特性分析

本节中，我们从空间维度对移动用户的行为特性进行分析。由于用户精确的位置信息难以获取，在我们的研究中，我们根据移动通信网络中的小区信息对用户进行定位。具体来讲，我们在采集的HTTP话单数据记录了用户发起请求时所处的小区标识。同时，我们向移动网络运营商请求了各小区基站信号塔的地理位置经纬度信息。这样，我们将移动用户的位置信息转化为其所处的小区地理信息，如图4-X所示。



图4-X 移动网络中用户位置信息提取示意图。

* + 1. 访问小区数

我们首先对用户在移动网络中使用网络视频业务时访问的小区数里进行了分析。这一指标体现了用户的移动范围和移动频率。图4-X显示了我们数据集中所有用户的访问小区数累积分布函数。总体来看，多数用户访问小区数较小，分布呈现出重尾特性。约82%的用户在使用网络视频业务期间仅访问过1个小区，90%的用户的访问小区数都在3个以下。而与此同时，约3%的用户则访问过多达10个及以上的小区。基于4.4.2小节中提出的方法，我们从数据集中检测出了2,597名高移动性用户。这些用户占据总用户的3.47%，却对应着75.61%的访问小区数。此外，我们将数据集中访问小区数为1的用户定义为静止用户。而对于高移动性用户和静止用户之外的用户，则被定义为低移动性用户。一般来讲，用户的移动性越高，其在无线接入网中触发的小区切换等活动就越多，进而占据的无线资源就会越多。



图4-X

* + 1. 请求位置

接下来，我们关注于移动用户使用网络视频业务时的具体位置。首先，从用户整体的角度，我们分析了数据集中用户发起视频请求的地理分布强度，如图4-X的热度图所示。其中，横轴数值为维度，纵轴数值为精度。为保护用户隐私，我们对经纬度的真实数值做了匿名处理。整个地图被划分为个区域，各区域的视频请求数值大小由图像灰度表示。从图中我们可以发现，某些“热点”地区的视频请求数要明显高于其他区域。其中，一个主要的热点地区在中西部，而两个次要的热点地区分别在东北部和东南部。我们进一步调查了这些热点地区的周围环境，发现其主要是大型商圈和高校校园地带。可以推测，这些地区的顾客（比如在咖啡店休息时）和学生（比如在寝室娱乐时）更加的热衷于使用移动网络视频业务作为一项消遣。热点区域对应着大量的移动用户与视频请求，对其的发现具有十分重要的现实意义。例如，网络运营商可以优化网络设施部署和视频分发策略，以提升用户体验；而业务提供商可以设计基于地理位置的广告推送与视频推荐，以获取潜在利润。



图4-X 用户的视频请求整体分布热度图。

为了定量的衡量整体的用户请求在空间维度上分布的均匀程度，我们提出了标准化的请求位置熵的概念，具体如下：

(4-x)

(4-X)

其中，为总区域数，为区域中的用户视频请求数。为个区域所能达到的最大请求位置熵值，因此标准化后的数值在0到1之间。如果用户的请求位置在空间上分布的非常均匀，则标准化熵的值将接近1；否则，如果用户的请求位置仅集中在某几个区域内，则标准化熵的值将接近0。

接下来，我们分析了请求时刻与用户整体请求位置的关系。我们首先比较了在不同时刻用户的请求位置的集中程度，如图4-X所示。我们发现，一天之内标准化请求位置熵的差别很大。夜晚的熵值明显低于白天的熵值，这说明在夜晚时用户要更为集中。然而，在不同时刻各区域中的用户请求数量差别是很大的。图4-X显示了在19:00与2:00，数据集中各个区域的用户视频请求数量累积分布函数。可以发现，在请求位置熵数值很低的时刻（2:00），大多数区域内并没有用户的视频请求；而在熵值高的时刻（19:00），有相当一部分的区域内出现了用户的视频请求。由此，我们可以总结出请求位置熵在夜晚降低的原因：移动视频业务的用户活跃性在夜晚会出现下降。对于非热点区域，其中的用户请求数量本身就很小，此时迅速降低至0；而对于热点区域，在经历衰减后仍具有着可观的用户请求数量。因此，夜晚时用户的请求在空间维度上趋于集中，熵值下降。



图4-X (a)一天内不同时刻对应的标准化请求位置熵。(b)典型时刻各区域内视频请求数累积分布函数。

我们进一步对用户在不同时刻的具体请求位置进行了分析，并发现其呈现出了基本相同的热点地区特性。图4-X显示了我们数据集中请求位置熵最高时刻（19:00）和最低时刻（2:00），用户的请求位置热力图。为了简洁起见，我们略去了其他时刻相似的结果图。从图中可以看出，热点地区一直固定在地图的中西部、东北部和东南部。



图4-X 不同时刻用户请求位置热力图。上图时刻：19:00；下图时刻：2:00。

接下来，从用户单体的角度，我们关注于各个用户的请求位置历史信息。我们研究一个用户在不同天之间，是否会在同一位置使用移动视频业务。为此，我们提出了请求位置重合率的概念。具体来讲，对于在天中使用网络视频业务的用户，其各天之间的请求位置重合率定义如下：

(4-X)

其中，是用户在第天内使用网络视频业务时访问的小区集合；而表示集合中的元素个数。在图4-X中，我们对于数据集中在多天内使用移动视频业务的用户，给出了其请求位置重合率的累积分布函数曲线。从图中可以看出，重合率的分布极不均匀，呈现出了明显的两极分化现象。57.30%的用户每次发起会话时的位置都不相同（请求位置重合率为0）；而30.72%的用户在不同天中的请求位置完全相同（请求位置重合率为1）。在图中，我们也给出了高移动性用户的请求位置重合率的累积分布函数曲线。相较而言，高移动性用户的播放位置更为复杂。其请求位置重合率在极限情形（0或1）出现较少，且整体上分布更为均匀。



图4-X

* + 1. 移动模式

在本小节中，我们用户在移动网络中使用网络视频业务时的移动模式进行分析。移动模式指的是一种用来描述用户运动情况的空间-时间序列。在我们的研究中，我们将其定义为移动视频用户的访问小区-停留时间对的序列。具体来讲，如果用户在观看移动视频时改变了次接入小区，则其移动模式为：

(4-X)

其中，为用户第次改变后对应的接入小区标识，而为用户在该小区停留的时间。值得注意的是，不相邻的小区标识可以是相同的，那代表着用户的移动轨迹出现了环。通过对移动模式的定义与分析，我们可以获取用户在哪里、使用了多久移动视频业务的信息。

基于移动模式，我们首先对用户的移动轨迹进行了分析。表4-X列出了按用户数排名Top 6的用户移动轨迹模式。这些模式覆盖了我们数据集中93.82%的用户。从表中可以看出，用户观看移动视频时很少移动较长的距离。82.76%的用户一直停留在唯一的一个小区中（模式），即4.5.1中提到的静止用户。7.84%的用户移动到了另外一个小区（模式），占据着非静止用户的xx%。此外，还存在着一些序列长度大于2的更为复杂的轨迹模式。其中，我们发现了一个很有趣的现象：复杂移动轨迹模型中往往存在小环，例如模式和模式。这些复杂模式中用户的移动范围其实并不大。在我们的数据集中，最长的移动轨迹具有36个小区标识。考虑到移动通信网络中一个小区范围大概有1公里左右，该轨迹对应的情形非常可能是用户在乘坐交通工具时观看移动视频。

表4-X

|  |  |  |  |
| --- | --- | --- | --- |
| **排名** | **移动轨迹模式** | **所有用户中占比** | **非静止用户中占比** |
| 1 |  | 82.76% | - |
| 2 |  | 7.84% | 45.47% |
| 3 |  | 1.16% | 6.73% |
| 4 |  | 1.05% | 6.09% |
| 5 |  | 0.66% | 3.83% |
| 6 |  | 0.35% | 2.03% |

接下来，我们对用户移动模式中的停留时间进行分析。图4-X显示了所有用户以及非静止用户使用移动视频业务时，在各小区内停留时间长度的累积分布函数。我们发现大多数用户的停留时间较短。超过60%的用户在一个小区内仅停留不到5分钟的时间；约80%用户的小区停留时间少于15分钟。相较而言，非静止用户在各小区的停留时间会更短。约84%的非静止用户停留时间少于15分钟。在本小节的分析中，我们规定：若分钟，用户在小区的停留时间模式为short，；否则，若分钟，则用户在小区停留时间模式为long。表4-X列出了按用户数排名Top 6的用户停留时间模式。这些模式覆盖了我们数据集中94.21%的用户。从表中，我们发现long停留时间往往发生在静止用户身上。而对于具有长移动轨迹的非静止用户，其各小区的停留时间往往较短。可以推测，这些用户的移动速度平稳且较快，很可能是在乘坐交通工具的同时观看移动视频。



图4-X

表4-X

|  |  |  |  |
| --- | --- | --- | --- |
| **排名** | **停留时间模式** | **所有用户中占比** | **非静止用户中占比** |
| 1 |  | 61.67% | - |
| 2 |  | 21.09% | - |
| 3 |  | 4.19% | 24.30% |
| 4 |  | 2.95% | 17.13% |
| 5 |  | 2.94% | 17.07% |
| 6 |  | 1.37% | 7.95% |

* 1. 用户业务使用特性分析
     1. 观看视频数

在本小节中，我们对移动用户观看的网络视频数量进行分析。对于一个用户而言，其观看的（去重后的）总视频数量，体现着在应用级别上该用户对网络视频业务的使用强度。在图4-X中，我们给出了数据集中所有用户观看的视频数累积分布函数。整体来看，每个用户平均观看了3.71个视频，小于固网环境中用户平均观看视频数（一般在10个以上）[18]。超过40%的用户只观看过1个视频，而80%的用户观看视频数少于5。同时，确实存在近10%的用户，观看了多达10个及以上的视频。我们进一步将用户数与观看视频数画在了一个双对数坐标系中，如图4-X所示。其中，横轴是按升序排列视频数；而纵轴是具有该视频数的用户数。根据图中近乎直线的散点分布，我们可以认为用户数与用户所观看的视频数之间存在着幂律定律。我们进一步对数据进行了回归分析来确认这一性质，最终得到参数，同时决定系数。用户观看视频数的分布是偏斜的，这说明不同移动用户之间对于网络视频业务的使用强度差别很大。大多数的用户并不会观看很多视频，而少量用户则观看了绝大多数的被请求视频。基于在4.4.2小节中提出的方法，我们根据观看视频数从数据集中检测出了7,070名重度业务使用用户。这些重度用户占据总用户数的9.44%，却播放了总视频数的47.96%。



图4-X (a)用户观看视频数的累积分布函数。(b)双对数尺度下用户观看的视频数与用户数的关系。

* + 1. 观看时刻

接下来，我们还对数据集中用户的整体业务使用的昼夜节律模式进行了分析。图4-x显示了我们数据集中各个小时内用户的观看视频数分布。可以看出，视频数在一天内的变化十分剧烈，并且在白天数值较高而在深夜凌晨数值较低。具体来看，用户观看的视频数在早晨4:00至8:00时间段有一个明显的增长。在日间9:00至16:00时间段内，用户的业务使用情况相对比较平稳。其中在中午12:00，视频数出现了一个小高峰。接下来从傍晚17:00开始，视频数开始快速增长，并在晚上20:00达到最高值。之后，用户观看的视频数开始大幅度的衰减，直至凌晨3:00达到最低值。综上，我们可以看出，用户在移动网络中使用网络视频业务的高峰期主要出现在中午、傍晚和深夜之前。这与在固定网络中得到的研究结果不同。在固定网络中，用户使用网络视频业务的高峰期仅出现在中午，并在傍晚就开始衰减[17, 19]。这一差异表明，对于移动视频，用户往往当作一项消遣，在他们的非工作、闲暇零散的时间观看。



图4-X 用户观看的视频数在24小时内分布

* + 1. 重复播放行为

在一个用户所有的观看视频中，有些令该用户感兴趣的视频可能会被观看了多次。用户的这种重复播放某些视频的行为可以反应用户的播放习惯并影响用户的业务使用强度，因而是十分重要的。在我们的研究中，我们提出了重放率这一指标来衡量用户的重复播放行为。对于用户的重放率定义如下：

(4-x)

其中，为用户观看的所有的（去重）视频总数；而为用户观看了两次及以上的（去重）视频数。

图4-X给出了所有用户及重度业务使用用户的重放率累积分布函数。对于所有用户的整体分布，其在图中呈现出了明显的两极分化现象。52%的用户不存在重复播放行为（重放率为0）；而31%的用户重复播放过他们看过的所有视频（重放率为1）。总体来看，有35%的用户对应重放率超过0.5，即他们重复播放了多于一半的观看视频。而对于重度业务使用用户而言，其重复播放行为要更加频繁。从图中可以看出，重放率为0的重度用户比例很小，53%的重度用户都对应着超过0.5的重放率，而大约24%的重度用户对应重放率等于1。此外，重度用户的重放率分布更加均匀：随着重放率数值的增大，CDF的增长比较平稳。



图4-X

* 1. 多角度用户行为交叉比较

为了更好的理解用户行为对不同类型的资源消耗的影响，我们在本小节对三个分析角度（数据消耗、位置移动、业务使用）之间的用户的相关关系进行了分析。在上文的分析中，我们根据流量字节数定义了重度数据消耗用户；根据访问小区数定义了高移动性用户；根据观看视频数定义了重度业务使用用户。我们首先对三种重度用户之间的重叠关系进行分析。图4-X显示了三种重度用户集合的文氏图【】。集合间最大的重叠部分是1,888名既属于重度数据消耗又属于重度业务使用的用户。这些用户分别占据了48%的总重度消耗用户数，以及27%的总重度业务使用用户数。从图中可以看出，三个集合中的很大部分并没有相互重合。这表明在某一分析角度中重度用户并不一定是另一个分析角度中的重度用户



图4-X 三种重度用户集合的文氏图。

然而，经过对不同角度用户行为的交叉比较，我们发现某一分析角度中的重度用户，在另一个分析角度中，仍趋于产生比原分析角度中非重度用户要多的消耗。图4-X显示了不同移动性的用户所产生的流量字节数累积分布函数。从图中我们可以看出，高移动性的用户产生的流量往往也较大。高移动性、低移动性和静止用户所对应的平均流量字节数分别为36.41兆字节，11.87兆字节和2.96兆字节。而在图4-X中，我们给出了不同数据消耗的用户所访问的小区数累积分布函数。我们同样可以发现，数据消耗中的重度用户通常趋向于比非重度用户访问更多的小区数，进而消耗更多的无线接入资源。其他分析角度之间的交叉比较结果相似，为简洁起见，我们略去了具体的结果图。



图4-X (a)不同移动性用户的数据消耗情况。(b)不同数据消耗用户的移动性情况。

* 1. 本章小结

在本章中，基于大规模网络数据，我们对移动网络中网络视频业务用户的行为特性进行了仔细深入的衡量分析。首先，从数据消耗的角度，我们关注于用户的流量字节数与活跃时长。同时，我们发现重度用户的存在，并提出了一个通用的非参数方法来检测重度用户。然后，从位置移动的角度，我们对用户的访问小区数、请求位置以及移动模式进行了衡量分析。接下来，我们研究了用户业务使用方面的观看视频数、观看时刻以及重复播放等行为特性。最后，我们对同一分析角度的重度与非重度用户进行了跨分析角度的行为特性比较分析。

综上分析，我们在本章的研究揭示了用户是如何在移动网络中使用网络视频业务的。我们得到的分析结果对于网络运营商和业务提供商具有十分重要的实际应用价值。例如：基于对用户观看时刻和请求位置的分析，网络运营商可以在高峰期和热点地区部署专门的缓存服务器，来减轻服务器传输压力并减少视频传输时延，以提升用户体验。而根据对用户重复播放行为的分析，业务提供商可以考虑将用户频繁请求的视频文件直接保存在用户设备上，以减少非必要的重复传输。另外，对重度用户对研究显示，这些用户消耗了更多的无线接入网、核心网以及业务提供商处的资源，因此也具有更强的商业价值。流量计费、带宽优化、广告投放、视频推荐等业务在设计时，应该对这些用户做重点考虑以获取潜在收益。

在未来工作中，我们准备在移动网络中采集其他网络视频业务提供商的数据源，分析其用户行为特性，并与本章中的优酷视频用户的分析结果进行对比。

1. 综合性网络视频业务的用户喜好特性对比分析
   1. 概述

在前一章的分析内容中，我们关注于网络视频业务用户的行为特性。在此基础上，本章中我们进一步对用户使用网络视频业务时所体现出的喜好特性进行了深入研究。更好的了解用户的喜好特性，在实际应用的多个方面中都具有非常重要的价值。例如：网络视频业务提供商可以根据这些信息来设计更有效的信息服务（如视频推荐、广告投放等），以获取更大利润。视频内容生产者可以根据用户喜好，有针对性的调整他们的发布内容，以吸引更多的用户并获得更大的影响力。而网络运营商则可以面向视频内容，主动分配、调整其网络的资源与设施部署，以优化视频文件的分发传输。在本章的分析中，我们对网络视频业务的用户使用了更广义的概念，即对应着两种不同基本类型的业务使用：视频上传与视频播放，如图5-1所示。视频上传用户，即上传者，将其录制的视频文件上传至网络视频业务提供商处进行发布。而视频播放用户，即播放者，访问视频网站或APP，对其中已发布的视频进行播放观看。在我们前文的研究工作中，狭义的网络视频业务用户指的就是播放者。



图4-1 两种网络视频业务用户示意图：上传者与播放者

本章中，基于从网络运营商处采集的超过120亿条大规模流量数据和从视频网站处爬取的连续30天的长时期视频元信息，我们对用户使用优酷视频时的喜好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。我们对优酷视频这家综合性网络视频业务提供商进行了分析。不同于传统的单一类型网络视频业务，优酷同时提供UGC分享、VoD点播以及在线直播等视频服务【】。这种一站式的网络视频业务解决方案为用户提供了更多的内容类别来进行选择，同时也对用户的喜好产生了相应影响。尽管目前越来越多的业务提供商开始将业务升级至综合类型，学术界中对该类型业务的分析工作尚不多见。2）多角度的分析内容。我们分别从用户活跃度、视频属性以及用户关系三个角度，对用户的业务使用特点进行了分析。我们进一步揭示了这些特点背后所对应的用户喜好特性。3）不同用户类型对比性分析。在我们的分析中，我们同时考虑了上传者和播放者，这两种对应着不同业务使用方式的用户。我们对上传者与播放者的喜好特性进行了对比性分析。据我们所知，这在网络视频业务用户喜好的研究中尚属首次。

* 1. 研究现状

一些现存的研究工作关注于网络视频业务中受到播放者偏好的视频的属性特性。在文献[26]中，作者Cha等人从YouTube和一家韩国网络视频业务提供商Daum处爬取数据，并对用户所观看视频的内容属性和播放量增长进行了分析。在文献[27]，作者Abhari等人从YouTube处爬取了长达5个月的数据，并分析了用户观看的视频长度、视频大小、视频播放量、视频评分等内容，并进一步揭示了这些属性所反映出的视频播放者访问模式。作者Cheng等人在文献[28]中同样对YouTube中用户播放的视频进行长期的元信息采集，并分析了视频文件属性、用户访问模式、视频活跃时长、视频播放量增长以及用户间社交网络等方面内容。而作者Abdesslem等人在文献[29]中从欧洲某蜂窝网络中采集了用户播放YouTube视频的请求数据，并分析了其中用户活跃度、视频属性以及视频播放量等方面的特性。相较于这些现存工作，我们的研究不仅关注于网络视频业务中的播放者的喜好，还同时考虑了上传者在业务中的喜好，并将两者进行了对比性分析。同时，我们的分析工作基于大规模的网络流量数据和长时期的视频元信息数据，是对现存使用单类型或小规模数据集的研究工作很好的补充与拓展。此外，我们的分析对象——优酷，是一个同时提供UGC、VoD、在线直播等服务的综合性网络视频业务。对于这种目前流行的网络视频业务，用户的喜好特性往往与传统的以UGC服务为主的网络视频业务（如YouTube）或以VoD服务为主的网络视频业务（如Netflix）有很大的不同。

另一方面，一些现存的研究工作则侧重于对互联网业务中用户之间的关系及社交网络进行分析。在文献[30]中，作者Mislove等人采集了Flickr、YouTube、LiveJournal和Orkut四个网站的用户数据，发现并验证了用户社交网络中的幂律特性、小世界特性以及无标度特性等内容。在文献[31]中，作者Wilson等人从Facebook爬取用户数据，并对社交链接是否是用户交互等有效指标进行了分析。作者基于用户之间的评论回复和图片回复建立了友谊图谱，发现Facebook中的用户与其超过一半的好友不存在互动，并趋于仅与其一小组的好友进行交互。在文献[32]中，作者Kwak从Twitter爬取了大量的用户数据，并基于用户之间的关注行为建立了用户关系图谱。作者研究了用户的关注者与被关注者数量分布以及与发文数的关系，并揭示了用户关系图谱在信息传播方面发挥的作用。作者Benevenuto等人在文献[33]中对YouTube用户之间的评论回复行为构建了关系网络，并研究了用户评论中视频交互的典型模型。作者Li等人在文献[34]中研究了优酷视频中用户之间关注关系构成的友谊图谱。他们发现用户的社交连通性较弱，且好友间喜好较为一致。基于此作者提出了两个好友推荐算法。在上述的这些研究工作中，用户之间的关系网络都是基于互助关系或评论行为构建的。而在我们的研究中，我们则从另一角度，基于用户的视频播放行为来构建用户关系网络。此种网络能够更好的反映出用户对于视频内容的不同偏好。由此，我们的工作对现存的关于互联网业务用户喜好的研究工作形成了很好的补充与拓展。

* 1. 数据集

本章中研究所使用的数据集包含两部分，分别是：1）用于用户播放喜好分析的，从网络运营商处采集的大规模网络流量数据；和2）用于用户上传喜好分析的，从业务提供商处爬取的长时期视频元信息数据。具体内容如下：

* + 1. 大规模网络流量

网络流量数据采集自某互联网提供商在我国东北某省的省际网络的出口处。采集器对每次HTTP交互生成一条话单，包含以下关键字段：时间戳、匿名用户标识、HTTP请求URL。我们根据优酷视频请求的特定URL格式，将其单独过滤出来。同时，对于每一条视频请求URL，我们可以提取出相应的视频ID。将视频ID通过预定义格式的HTTP GET报文发送至优酷提供的开放API[35]，我们可以获取相应视频的元信息数据。视频元信息为一条JSON格式数据，包含视频的静态属性信息和动态统计信息，例如：视频种类、视频时长、视频上传时刻、视频上传者信息、视频当前播放量等。

整个数据采集阶段从2015年12月12日起至2015年12月18日止，共计7天。我们使用Python语言开发了一个爬虫程序，自动的采集了我们数据集中所有视频的元信息。最终，我们一共采集到 12,106,121,482 条 HTTP 记录，包含155,991条优酷视频请求，覆盖了49,252名播放者与101,639个视频。

* + 1. 长时期视频元信息

为研究优酷中上传用户的喜好特性，我们爬取并追踪了一组新上传视频的元信息。具体来讲，整个采集过程分为两步：首先，我们在某一起始日期从优酷视频网站的“本日最新上传视频”部分爬取了所有新上传的视频ID列表。同时，我们通过优酷的开放API查询这些视频的起始元信息。接下来，我们追踪了这些视频在接下来30天中的视频播放量情况。我们每天查询一次这些视频的元信息，并提取当日的视频播放量。这样，对于每一个被追逐的视频，我们可以得到自其初始上传到一个月后，其每天的视频播放量序列。

对于第一步，我们从2015年12月12日起至2015年12月18日止，共采集了7天的当日最新上传视频列表。而在第二步对这7天视频的追踪过程中，共计29,916个视频被上传者删除或被优酷屏蔽。我们将这些不完整的数据从数据集中剔除。最终，我们共采集了144,457个视频的长时期元信息数据，来自于80,929名上传者。

* 1. 用户活跃度分析
     1. 每天活跃度

我们首先以天为时间粒度，分析上传者和播放者在使用优酷视频业务的时间偏好差异。图5-X和图5-X分别显示了我们数据集中每天的上传者数、上传视频数、播放者数以及视频请求数。首先，我们发现每天的上传视频数（视频请求数）要高于上传者（播放者）数。这表明一个用户往往会在一天之内上传（播放）多个视频。同时，我们发现在一周之内，工作日中的上传者（播放者）数量与周末时的上传者（播放者）数量相差不大。但对于上传视频数（视频请求数）而言，其在周末的数量要略大于在工作日中的数量。这意味着，用户在规模上虽然较为平稳，但却喜欢在周末使用更多的网络视频业务。对此，网络运营商和业务提供商应该考虑灵活调整其资源部署，来为用户活跃度较强的日期提供更多的支持，以提升用户体验。



图5-X 不同天内(a)上传者、(b)播放者的活跃度。红色日期为周末。

* + 1. 每小时活跃度

图5-X显示了我们数据集中以小时为时间粒度的上传者数、上传视频数、播放者数以及视频请求数。我们发现，一天之内上传者与播放者的活度程度变化十分剧烈。用户数在深夜及凌晨较低而在其他时段较高，而上传视频数（视频请求数）大体上与相应的上传者数（播放者数）成正比。对于上传者，其活跃度在早晨7:00至上午10:00时段内快速增长，并于11:00时达到最高值。在接下来11:00到22:00时段，上传者活跃度逐步降低。期间，两个显著的下降时刻出现在12:00和18:00，我们发现这正是午饭和晚饭时段。之后，从23:00到6:00，上传者活跃度剧烈下降并保持在较低水平。而对于播放者，其活跃度在早晨5:00到8:00时段也经历了大幅度的增长。在白天9:00到16:00内，播放者活跃度较为平稳，并在12:00时经历了一个小高潮。接下来从17:00活跃度开始快速增长，在20:00达到最高值，并保持较高水平直至22:00。最终，播放者活跃度在深夜大幅下降，并在凌晨3:00达到最低值。



图5-X 不同小时内用户的活跃度。

将上传者和播放者的每小时活跃度进行对比，我们发现了若干不同之处。首先，两种用户在早晨都出现了活跃度的快速增长，但播放者的增长时段要比上传者的增长时段整体提前了近2个小时。另外，在午饭时刻、晚饭时刻以及傍晚时段内，播放者的活跃度保持增长，而上传者的活跃度则出现了下降。尤其是晚上17:00到22:00时段，对于播放者而言这是一天之内的高峰期，而对于上传者而言，用户活跃度在大幅度衰减。经过仔细的分析，我们认为上述业务使用时间上的偏好差异可由不同类型用户使用网络视频业务的目的不同来解释。播放者观看网络视频更多的是为了消遣。他们可以在想要的任意时刻（例如在清早刚刚醒来时）发起视频请求。并且，他们倾向于在零散的空闲时段内（例如午饭和晚饭休息时）观看网络视频。而对于上传者来说，他们在将视频文件上传至网站前通常要做些准备工作，并且往往将视频上传行为当作一项任务而在工作时间完成。因此，如图5-X所示，视频上传和视频播放的高峰期基本上是不重合的。根据此分析结果，我们建议上传者根据播放者在不同时段的数量，来调整他们的视频上传时间。也就是说，上传者最好在播放者的活跃高峰期附近发布他们的视频，以便吸引更多的潜在用户播放该视频来扩大影响力。

* + 1. 用户业务使用

最后，在本小节中我们关注于用户的业务使用强度分布。图5-X显示了我们数据集中上传者一天内的上传视频数和播放者一天内的视频请求数的累积分布函数。对于两种用户，我们从图中观察到了十分相似的重尾分布特性。具体来讲，76.93%的上传者仅上传了1个视频，92.62 %的上传者对应上传视频数小于4；而68.31%的播放者仅观看了1个视频，88.72%的播放者对应视频请求数小于4。与此同时，分别存在着近2%的上传者（播放者），在一天之内上传（播放）了10个以上的视频。



图5-X

我们进一步给出了用户比例与业务使用累积占比的关系，如图5-X所示。其中，横轴为按上传视频（视频请求）数降序排列的用户比例，纵轴是相应用户的上传视频（视频请求）数量累计占比。从图中我们发现两种用户的曲线十分相似。前20%的用户都占据了约55%的业务使用。由此，我们可以得到结论：著名的帕累托定律[36]（Pareto principle，即约80%的效果来自20%的原因）对优酷视频业务中的上传者（播放者）的业务使用强度并不适用。

* 1. 视频属性分析
     1. 视频类型

在本小节中，我们关注于用户在使用网络视频业务时，都在上传或观看什么样的视频。在优酷视频中，业务提供商预定义了25种视频类型来描述视频内容。而每一个优酷视频都会被指定一个视频类型。由此，我们对数据集中用户上传或观看的视频所对应的视频类型进行了对比性分析。表5-X分别列出了上传视频和播放视频前十名的视频种类及占比。这些视频种类覆盖了82%的总上传视频和80%的总播放视频。从图中我们可以发现，不同种类用户的视频类型喜好差距很大。对于上传者而言，其上传最多的视频种类是“生活（Life）”，接下来是“教育（HowTo）”、“音乐（Music）”、“游戏（Game）”和“家庭（Family）”。这些类型中的大多数视频都对应着用户自己拍摄的视频内容（UGC），例如：烹饪教程、化妆教程、游戏解说录像、私人K歌录像等。而对于播放者而言，其观看的视频中，超过四分之一的视频类型为“电视剧（TV Episode）”。被观看的视频种类前三名是“电视剧（TV Episode）”、“动漫（Animation）”和“电影（Movie）”，全部都对应着版权点播内容（copyrighted VoD content）。由此，我们可以得出，虽然优酷中每天被用户上传的大多数视频都属于UGC视频，能够吸引用户来播放的关键视频却都属于VoD视频。此外，从不同用户的业务使用目的来看，很大一部分的上传着使用优酷视频来保存并分享他们的私人视频；而大多数的播放者观看优酷视频则主要是为了休闲娱乐。

表5-X

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **上传者** | **视频种类** | | **视频时长（分钟）** | |
| **排名** | **种类名称** | **视频数比例** | **中位数** | **均值** |
| 1 | Life | 16.45 | 4.28 | 9.49 |
| 2 | HowTo | 11.68 | 8.15 | 17.07 |
| 3 | Music | 10.86 | 3.71 | 7.40 |
| 4 | Game | 9.32 | 20.27 | 25.15 |
| 5 | Family | 7.14 | 1.93 | 3.84 |
| 6 | TV Episode | 7.00 | 17.49 | 20.27 |
| 7 | News | 6.02 | 2.26 | 6.13 |
| 8 | Sport | 5.44 | 3.97 | 11.40 |
| 9 | Creativity | 4.12 | 3.80 | 5.38 |
| 10 | Amateur | 4.02 | 1.85 | 5.63 |
| **播放者** | **视频种类** | | **视频时长（分钟）** | |
| **排名** | **种类名称** | **视频数比例** | **中位数** | **均值** |
| 1 | TV Episode | 26.45 | 44.85 | 43.50 |
| 2 | Animation | 10.20 | 13.00 | 15.59 |
| 3 | Movie | 8.07 | 87.21 | 68.60 |
| 4 | Life | 6.89 | 4.39 | 9.00 |
| 5 | Entertainment | 6.04 | 5.04 | 9.04 |
| 6 | Variety Show | 5.94 | 63.46 | 53.73 |
| 7 | Music | 5.64 | 4.03 | 6.23 |
| 8 | Amateur | 3.84 | 4.31 | 7.06 |
| 9 | HowTo | 3.57 | 4.67 | 16.50 |
| 10 | Humor | 3.45 | 3.65 | 5.64 |

我们的分析结果具有着十分重要的现实应用价值。例如，网络视频业务提供商可以根据不同的视频类型分布，来进一步调节版权内容的购买选择和广告内容的部署策略等。而网络运营商则可以基于视频类型，对某些热门的VoD视频做单独的缓冲部署，以提升用户体验。

对于视频种类，我们发现的另一个值得注意的现象是：有些用户只上传（播放）某些特定类型的视频，而另一些用户的兴趣则非常广泛。为了定量的衡量用户在不同视频种类上的偏好程度，我们提出了标准化的视频种类熵的概念。具体来说，令为用户上传（播放）的视频中属于类型的视频数，为总视频类型数，则用户的标准化视频种类熵为：

(5-X)

(5-X)

标准化后的熵值在0至1之间。当用户的视频种类越分散，其熵值越接近1；否则，当用户的视频种类越集中，其熵值则越接近0。

图5-X显示了我们数据集中具有多个视频的上传者与播放者的标准化视频种类熵的累积分布函数。由图可知，种类熵值往往较低：70%的上传者和73%的播放者，对应的熵值都小于0.2。这意味着上传者和播放者都趋于仅关注于一小部分特性的视频类型。尤其是，对于多达62%的上传者和67%的播放者，其熵值甚至为0。这表明这些用户仅对某一种特定的视频类型感兴趣。此外，在0.22熵值附近，上传者和播放者的分布都出现了一个明显的增长。这是对应着具有两个不同种类的视频的用户。根据5.4.3小节的分析结果，上传（播放）了两个视频的用户占比较大，从而导致了此时视频种类熵分布的波动。



图5-X

* + 1. 视频时长

接下来，我们对用户所上传的和所播放的视频，在时间长度上的特性进行分析。在图5-X中，我们分别给出了数据集中上传视频和播放视频的时长累积分布函数。而表5-X也列出了各个视频类型所对应的视频时长中位数和均值。我们可以看出，总体来说用户上传视频的时长往往较短。76%的上传视频对应时长少于10分钟，而超过90%的上传视频对应时长都在半小时之内。相比之下，用户的观看视频短于10分钟和半小时的比例仅为43%和62%。也就是说，用户的观看视频由更多的长视频组成。这是符合我们预期的，因为观看视频中的很大一部分是版权VoD视频内容，例如电视剧、电影、综艺、动漫等。从图5-X中可知，这些种类的视频往往都是长视频。此外，在图5-X中我们发现，对于被观看视频的时长分布，在45分钟附近有一个明显的增长。此处分布增长对应着大概20%的视频，由表5-X可知，主要是大量的电视剧视频和综艺视频。

对不同用户所偏好的视频的时长特性分析，具有着重要的实际应用价值。例如，可以根据用户种类灵活的调整网络视频业务占据的上下行带宽，并为视频流量传输设计更好的缓存机制。



图5-X

* + 1. 视频播放量

在本小节中，我们对用户的上传视频与播放视频，在动态流行度方面特性进行分析。我们使用播放量这一指标来衡量视频的流行度。视频的播放量可以从一定程度上反映出用户对于视频的访问模式。对于用户的上传视频，我们检查了自其发布一个月后的视频播放量。而对于用户的播放视频，我们则关注于播放者请求时该视频已有的播放量。在大量的现存工作中[29, 37, 38]，研究者发现网络媒体内容的流行度分布服从奇普夫定律[39]（Zipf’s law）。具体来讲，对于排名为的内容，其被访问频率为：

(5-X)

其中为内容总数，为分布模型参数。该函数在双对数坐标尺度下呈现为一条直线：

(5-X)

为检查我们数据集中用户的上传视频与播放视频的流行度是否同样符合奇普夫定律，在图5-X和图5-X中我们分别给出了双对数坐标系中视频播放量排名和数值的关系。



图5-X

对于被上传的视频，从图5-X中我们可以观察到播放量排名和数值间存在着近似线性的关系。我们进一步对数据运行了线性回归来进行验证。我们得到回归参数，。具体回归曲线在图中用红色标出。可以看出，和严格的奇普夫分布相比，被上传视频的实际数据中，极为流行的和极不流行的视频数量较少。

而对于被观看的视频，如图5-X所示，我们发现其曲线是强烈偏斜的，在尾部显著降低。这意味着，在被观看视频的实际数据中，较为不流行的视频数量要远远小于奇普夫模型所预计的数量。显而易见，一条直线是无法很好的拟合此情形的分布的。在我们的分析中，我们尝试使用一个分段函数来进行拟合：在视频播放量较大的分布开始部分，我们保留使用线性的奇普夫模型；而对于分布的重尾部分，我们引入了一个指数模型。

(5-X)

我们对实际数据运行了线性回归来进行验证。我们得到回归参数，，，，。具体回归曲线在图中用红色标出。可以看出，我们的模型可以很好的拟合实际数据。

综上分析，我们发现对于用户新上传的视频，其一个月后的播放量分布近似符合奇普夫定律；而对于用户观看的视频，其播放量分布可由一个线性加指数的分段函数来拟合。

* 1. 用户关系分析

在本节中，我们对网络视频业务中基于用户喜好的用户间关系进行分析。由于业务使用的情形不同，我们对上传者与播放者进行了分别的研究。在视频上传方面，由于不同的上传者之间相互较为独立，我们转而研究了各上传者与其粉丝之间的关系。上传者的粉丝数量能够体现其发布内容的吸引力与影响力。而在视频播放方面，我们则关注于播放者对于视频的选择。我们基于用户对某些视频的共同喜好，构建了关系网络并进行分析。

* + 1. 上传者粉丝数

图5-X显示了我们数据集中上传者具有的粉丝数的累积分布函数。从图中我们发现，大多数上传者的粉丝数量较少，而少数上传者具有极大的粉丝数。具体来讲，52.92%的上传者仅有不到10名粉丝，而31.50%的上传者甚至根本没有粉丝。与此同时，却存在着12.21%的上传者，其具有粉丝数超过5,000。我们进一步研究了上传者的粉丝数整体分布，并发现其大体上服从Weibull分布[40]。Weibull分布的概率密度函数如下：

(5-X)

其中为形状参数，为尺度参数。我们使用数据集进行回归，得到回归参数，。为显示拟合效果，我们在图5-X中也给出了拟合的累积分布函数曲线。拟合曲线与真实数据的曲线大体上重合，表明Weibull分布是对上传者粉丝数整体分布一个很好的近似。



图5-X

接下来，我们分析了上传者的粉丝数是如何影响其上传视频的流行度的。在优酷中，每当上传者新发布一个视频时，其所有粉丝都将收到一个包含该视频题目、简介和链接的系统通知。通过这一机制，粉丝可以密切关注上传者动态，并且被鼓励去播放其感兴趣的新上传视频。由此，上传者的粉丝数越大，其发布视频的初始潜在播放量也应该越大。图5-X显示了我们数据集中上传者的粉丝数和其发布视频的播放量之间的关系。为了便于阅读，我们将图中坐标轴设置成对数刻度。由图所示，我们发现当上传者的粉丝数较大时，其发布视频的总播放量往往也很大。然而，当上传者的粉丝数较小时，不同上传视频的播放量间差异很大，并且也存在着一些视频具有较大的播放量。为弄清这一现象的原因，我们进一步从上传者粉丝数的角度，对大播放量视频的类别构成进行了分析。在表5-X中，对于数据集中播放量超过10,000的视频，我们分别列出了来自低粉丝数（少于100）上传者和来自高粉丝数（大于10,000）上传者，这两部分视频的类型排名前五名。可以看出，这两部分视频的类型构成相差很大。对于来自低粉丝数的流行视频，各类别的视频数量占比十分接近（6%左右）。我们进一步检查了这些视频的内容，并未发现任何明显的特点。这意味着，只要视频内容本身足够吸引人，就算其发布者的粉丝数较小，该视频一样能够非常流行。而对于来自高粉丝数的流行视频，前三位视频类型“游戏（Game）”、“新闻（News）”、“娱乐（Entertainment）”占据了近一半的视频数。经过进一步检查，我们发现这些视频大多为游戏解说、当日新闻和娱乐资讯等内容。这些视频内容往往由某些官方发布者周期性的在优酷上发布，并且能都吸引大量的忠实观众。因此，每当这些发布者发布了新的一期节目，往往会吸引大量的粉丝进行观看。



图5-X

* + 1. 播放者喜好网络

不同的播放者可能拥有相似的喜好，进而观看了同一个视频。在本小节中，我们对由播放者的共同视频喜好所组成的网络进行分析。具体来讲，令播放者集合为图的顶点集合，若播放者与播放者观看了同一个视频，则将与通过一条边连接。最终，所有播放者顶点和其间的边构成了播放者的喜好网络。与传统的基于关注关系的用户社交网络不同，我们的用户网络基于播放行为建立，能够更为直接的反映用户喜好的整体关联关系。由于我们数据集中并非所有的播放者都是相互连接的，为了简单起见，我们的研究仅关注于喜好网络的最大联通子网络。该子网络包含23,057个顶点，2,643,381条边。

我们首先对播放者喜好网络的平均路径长度进行分析。平均路径长度衡量了图的信息传递有效性，具体定义如下：

(5-X)

其中，为中顶点数，为与之间的最短路径长度。在现实世界中的人际社交网络往往具有较小的平均路径长度，大概在6左右[41, 42]。而优酷播放者喜好网络的平均路径长度要更小，仅有3.0913。这表明喜好网络的边十分的密集：播放者之间具有密切的强相关程度的视频选择喜好。

接下来，我们衡量了播放者喜好网络的聚类系数[43, 44]。图的聚类系数用于衡量相邻节点形成高度连接群集的趋势，具体定义如下：

(5-X)

其中，为中顶点数，为图中以为顶点的三角形数，为图中以为顶点的三元组数。经计算，我们数据集中播放者喜好网络的聚类系数为0.8583。

在复杂网络理论中，小世界网络[45]是一种数学图类型，在这种图中大多数的非邻居节点可以经由少数几跳而到达彼此。如果我们的播放者喜好网络属于小世界网络，则可断定播放者在视频选择的喜好上具有非常强的相关性。为检验播放者喜好网络是否是一个小世界网络，我们进一步构造了一个具有相同节点数和边数的等效随机图。经计算，此随机图的平均路径长度为2.9147，而聚类系数仅有0.0099。综上我们可以得出，在聚类系数上，播放者喜好网络的数值远大于其等效随机图的数值；而在平均路径长度上，两者相差不大。由文献[45]中的结论，我们可以判定播放者喜好网络确实是一个小世界网络。

这是一个十分有意义的发现。在小世界网络中，顶点趋向于聚集成集群。对应到我们的数据集，这代表了具有相似视频喜好的播放者群组大量存在。这一现象为业务提供商创造了重要机会来提升其视频推荐算法。例如，同播放者群组中用户与用户观看过相同的一些视频，用户与用户观看过相同的一些视频，则用户很可能会对用户看过的视频感兴趣。因此，业务提供商应该向用户推荐用户看过的某些热点视频。图5-X显示出了一部分播放者喜好网络（1000个节点）的图例。从图中我们可以可以清楚的观察到小世界网络的特性（大量的顶点三角形和顶点团）。



图5-X

* 1. 本章小结

在本章中，基于大规模的网络流量数据和长时期的视频元信息数据，我们对网络视频业务中两种主要用户，即上传者与播放者，的喜好特性进行了深入的对比性分析。首先，我们研究了用户在不同时间粒度中体现出的活跃性特性，并对比了两种用户的整体业务使用情况。然后，我们对比的分析了上传者和播放者所偏好的视频的属性特点，包括视频类型、视频时长和视频播放量等方面内容。最后，我们关注于用户喜好特性所体现出的用户关系，并对上传者的粉丝数与播放者间的喜好网络分别进行了研究。

综上，我们的研究揭示了网络视频业务中用户行为背后的喜好特性。我们得出的分析结果，对网络视频业务的各利益相关方而言，是十分重要而可靠的。例如，内容生产者可以根据播放者的活跃时期和内容类型偏好，来调整其视频发布时间和视频类型，以便吸引更多用户、获取潜在利益。业务提供商和在线广告商可以基于用户的活跃时期和喜好网络，来进行合理的信息推送。而网络运营商则可以利用用户的上传视频与播放视频的时长、播放量等方面的特性，灵活的配置上下行带宽、选择缓存内容及时长，来优化对网络视频数据的传输。

在未来工作中，我们则希望能够与网络视频业务提供商进行深度合作，以获取更为详细的用户个人信息、访问日志、操作记录等数据，来对用户喜好更细粒度的进一步分析。

1. 网络视频流行度分析及预测
   1. 概述

随着网络视频业务的快速发展，越来越多的视频被带进互联网中。每秒内，全球互联网中就有近百万个视频被传输。如果一个用户想要看完一个月内互联网中所有的新上传视频，需要花费其超过500万年的时间[1]。如此数量巨大的网络视频，彼此间竞争着用户有限的时间和注意力。因此，网络视频的流行度以相当不对称的方式分布[46, 47]并不令人奇怪：一小部分视频获取了绝大多数的用户播放，而大多数的视频基本上没有被用户注意到。

由于网络视频的数量巨大并且用户关注度极不均匀，对其流行度特性的理解及进一步对未来流行度的预测，具有多方面的实际应用价值。例如，业务提供商可以设计更精准有效的视频推荐、视频搜索等信息服务算法[26]；在线广告商可以更有针对性的调整其广告部署策略，通过加大对潜在的热门视频的广告投放来获取利益[48]；网络运营商可以提前调整网络设施的配置，消除潜在的传输瓶颈以保障业务质量[49, 50]。并且，在移动终端间的机会网络通信中，网络视频的未来流行度信息也是至关重要的[51]。在此情形中，终端的传输带宽和存储能力十分有限，如果能对流行视频进行预测，则可以提前调整视频传输复制策略。此外，从社会学的角度，对视频流行度增长的深入分析，将有助于理解人类集体行为的触发规律，如随机产生、内部原因、外部原因等[52]。

本章中，基于从优酷视频网站爬取的视频元信息数据，我们首先对网络视频流行度的群体特性和单体特性进行了深入的分析。我们使用播放量作为网络视频流行度的衡量指标。播放量这一指标被广泛的应用于视频流行度的研究中，而评论量、收藏量等其他指标与其是强相关的[53, 54]。基于分析结果，我们进一步对网络视频的未来流行度预测问题进行了研究。我们分别提出了面向流行度级别和面向流行度数值的多种高效预测方法，并使用真实数据进行了验证。我们研究工作的主要内容及贡献如下：1）我们揭示了对于同天发布的视频群体，其长期播放量的整体分布模型及影响因子。2）我们分析了各视频在观察期各天内的播放量获取情况，并定义了若干播放量增长模式，来描述视频的流行度演化趋势。特别的，我们提出的模式首次运用了播放量激增的概念，并够能够覆盖视频整个活跃期。3）我们提出了一系列与视频流行度相关的特征，并结合高效的分类算法，探寻了不同情形中，对视频未来流行度级别预测的可行性与有效性。我们在视频发布时，立即预测其未来的流行度级别，精度可达74%而召回率约为60%。而如果引入初始观察期，我们的预测能力可大幅提升，至约95%的精度和约91%的召回率。4）我们分别考虑视频播放量的增长模式和级别转换，进一步提出了两种分组专用模型，对网络视频的未来流行度数值进行预测。实验结果显示，在相对误差指标上，我们的方法可优于作为基线的传统方法超过30%。

* 1. 研究现状

目前学术界中已存在一些面向互联网内容流行度的研究工作。在文献[26, 55]中，作者Cha等人分析了网络视频业务提供商Youtube和Daum的视频流行度特性。他们检查了两种数据集中的视频播放量随时间的增长及分布情况，并根据分析结果来改善视频传输的机制。在文献[56]中，作者Tan等人从国内网络视频业务提供商PPTV处收集了大量的用户观看行为记录，并研究了用户对于视频播放的集中趋势，分散趋势和对称程度。在文献[57]中，作者Liu等人从四个移动应用商店处采集了长时期的每日用户下载记录，并分析了Android商店和iOS商店之间，移动应用流行度特性的相同之处和不同之处。上述这些研究工作，为互联网内容的流行度特性提供了宝贵的见解。然而，这些工作主要关注于一组互联网内容的整体流行度分布，并且仅分析了互联网内容在一天或少数几天抽样中的流行度特性。而在我们的研究工作中，我们追踪并分析了网络视频在（自初始上传的）整个观察期内的流行度特性。并且，我们的研究角度更加的全面，同时涵盖了对视频群体与视频单体的研究。

对网络视频单体的流行度演化趋势的分析，是学术界另一个热门的研究方向。在文献[52]中，作者Crane等人引入传染病模型来对网络视频流行度进行研究。他们发现视频播放量增量的松弛过程遵循幂律，并根据高峰日的播放量增量占比将视频流行度演化趋势划分成四类。在文献[58]中，作者Cheng等人从YouTube处爬取数据，并分析了YouTube视频的属性、用户访问模式、播放量分布及播放量增长模式。作者使用了一个简单的对数模型来描述视频流行度的演化趋势。在文献[54, 59, 60]中，作者Figueiredo等人研究了YouTube中视频的播放量分布特性、增长模型以及来源等方面内容。基于时间序列的聚类算法[61, 62]，作者提出了四种视频播放量增长模式，并分析了不同数据集中各模式的出现情况。而在文献[63]中，作者Bao等人使用自激霍克斯过程（self-excited Hawkes processes），来对的微博内容的流行度演化进行了建模。相较于这些现存工作，我们研究中提出的视频播放量增长模式适用于整个观察期，而非仅针对于高峰日。此外，我们首次提出了视频激增的概念，并在播放量增长模式中考虑了视频激增的数量和时间位置。

此外，许多研究者致力于对网络视频的未来流行度进行预测。其中，有些工作的预测对象为互联网内容的流行度级别。作者Jamali等人在文献[64]中，基于评论特征、用户特征和机器学习分类算法，对网络新闻业务Digg的内容流行度级别进行了预测。作者Tsagkias等人在文献[65]中使用了两个连续的分类任务，来解决网络新闻评论量的级别预测问题。他们首先预测一条新闻是否会收到评论，然后根据结果进一步预测评论量是低级别还是高级别。作者Lee等人在文献[66]中运用生存分析方法中的Cox比例风险回归模型[67]，来对互联网内容在未来某时间点是否还会流行的问题，进行二分类预测。作者Vallet等人在文献[68]中从Twitter的内容话题中提取了关于YouTube视频的跨域特征，并利用梯度提升树分类算法对视频的流行性种类进行预测。作者Roy等人在文献[69]中同样从Twitter中提取了跨域特征。他们提出了一种算法来描述社交信息的转换，并进一步预测了YouTube视频是否会在未来经历播放量的激增。作者Vasconcelos等人在文献[70]中同时利用分类与回归方法，对一家基于地理位置的网络社交业务Foursquare内容流行度级别进行预测。除此之外，另一些研究工作的预测对象为互联网内容的流行度数值。在文献[47]中，作者Szabo等人研究了互联网内容的前期流行度与后期流行度之间的关系，并提出了一个对数线性模型来预测互联网内容的未来流行度数值。该模型被其他研究者在多种数据集上进行了验证，如：YouTube网络视频、Digg网络新闻、荷兰各个在线新闻门户[71]、法国各个在线新闻门户[72, 73]、iFeng网络新闻[74]等。基于对数线性模型，作者Pinto等人在文献[75]进一步考虑初始观察期内各天的播放量增量的权重，提出了一个多元线性回归模型来预测网络视频的未来播放量。除了基于回归的方法，研究者们还积极尝试了其他的预测手段，包括：储备池计算[76]、时间序列分析[77]、隐式马尔可夫模型[78]等。相较于这些现存的流行度预测研究工作，我们在流行度的级别预测中，从个多角度提出了高效预测特征，并考虑了多种预测情形；而在流行度数值预测中，我们使用了分组专用模型来替代简单的通用模型，极大的提升了预测性能。

* 1. 数据集

为了分析和预测网络视频的流行度，我们需要获取一组视频在一段观察期内的播放量。本章中，我们的研究数据采集自优酷视频。优酷是国内领先的网络视频业务提供商，其月活跃用户超过5亿人，平均的日播放量在8亿次以上[79]。通过优酷提供的开放API[35]，我们爬取了大量新上传视频的视频元信息和上传者元信息，并对各视频在发布后一个月内的播放量进行了追踪。其中，视频元信息包含视频的静态属性信息和动态统计信息，例如：视频标题、视频内容描述、视频内容标签、视频类型、视频时长、视频清晰度、视频文件格式、视频版权类型、视频公开类型、视频上传时间、视频上传者ID、视频上传客户端类型、视频发布状态、视频当前播放量、视频当前评论量、视频当前收藏量、视频被顶次数、视频被踩次数等等。而上传者元信息则包含上传者的个人信息和统计信息，例如：用户性别、注册时间、是否为会员账户、是否为共享账户、是否为认证账户、发布的视频数、发布的视频专辑数、发布的状态数、关注的用户数、粉丝数、获取的总播放量、获取的总收藏量等等。

与5.3.2小节中的介绍相似，本章中的数据采集过程分为两步：1）首先，在某一初始日期，我们从优酷网站的“本日最新上传视频”处爬取视频ID列表。我们将这些视频ID发送至优酷开放API，可获取视频元信息。我们再将各视频元信息中的上传者ID发送至优酷开放API，则可获取相应的上传者元信息。2）接下来，我们对列表中的各个视频，追踪接下来30天的播放量情况。我们每天查询一次各视频的元信息，并提取该视频当日的播放量。如此，对于采集的每一个新上传视频，我们可以获取其视频元信息、上传者元信息以及最初一个月的播放量序列。

对于采集过程第一步的列表爬取，我们自2015年10月11日起至2015年10月20日止，共采集了10组的每日新上传视频的ID列表。需要指出的是，我们固定的在深夜11:50时运行爬虫程序，以获取尽可能完整的该日所有新上传视频。而在第二步对各视频播放量的追踪过程中，共计47,621个视频被上传者删除或被优酷屏蔽。我们将这些不完整的数据从数据集中剔除。最终，我们采集了来自于130,412名上传者的共计200,773个视频的元信息及播放量序列数据。这些视频总共被播放者观看了2,024,986,133次。

在优酷网站中，除了“本日最新上传视频”列表外，还提供了其他的视频列表，如“本日最多播放视频”、“本日最多评论视频”等。在我们的研究中，我们关注于用户最新发布的视频。这样选择有助于避免抽样偏差[47, 80]，并可以获取视频自发布日起的完整播放量变化情况。

* 1. 视频群体流行度分析

首先，我们从群体的角度对网络视频的流行度进行研究。对于同一天发布的一组视频，我们分析了其在发布后一段时间内的播放量获取情况。我们根据视频播放量的整体分布，将视频的长期流行度进一步划分为不同的级别。此外，我们还对能够影响视频流行度的潜在因子进行了分析。

* + 1. 整体播放量分布

首先，对于我们数据集中（在10天中采集）的10组视频，我们给出了各组视频在发布日期30天后的整体播放量0.25、0.5和0.75分位数，如图6-1所示。全体数据的0.25、0.5和0.75分位数也在图中由虚线给出。从图中我们可以看到，视频长期播放量的分为数曲线在不同的视频发布日期之间保持稳定。这表明群体视频的长期播放量分布对发布日期不敏感，即对于不同天发布的各组视频，其长期播放量的分布是类似的。因此，我们可以合并考虑数据集中的不同天发布的视频，对其长期播放量分布进行通用的分析。



图6-1

许多现存的研究工作已经观察到，互联网内容的长期流行度分布通常是不均匀的，呈现出重尾特性[46, 81, 82]。我们发现该特点也同样适用于优酷视频数据。图6-X显示了我们数据集中所有视频在发布日期30天后的播放量累计分布函数。为了便于阅读，我们将图中横坐标设置成了对数刻度。图中横坐标跨越了6个数量级，说明不同视频的播放量差别很大。并且从图中我们可以看出，视频长期播放量的整体分布是十分不均匀的。大多数视频几乎不被用户留意到，而少数一些视频却获取了绝大多数的用户播放。具体来讲，在全体视频中超过一半的视频被播放了不到100次。有794个视频在最初30天内甚至仅仅被播放了一次。相比之下，4.53%的视频则获取了超过10,000次的用户播放。而我们数据集中最热门的视频甚至被播放了38,461,773次。综上分析，我们发现尽管每天中有大量的视频被上传至视频网站，但只有其中的一小部分能够真正的流行起来。如此偏斜的视频播放量分布体现了用户喜好的高不对称性。此外，我们发现网络视频业务提供商的推荐机制，也在视频的播放量差异中也起到了重要的作用[83]。在优酷中，某些类型的新发布视频（通常是电视剧、综艺等版权VoD内容），往往会被列在网站首页上来突出显示一段时间。这些视频对于用户更加可见，由此产生了所谓的富者更富效应（rich-get-richer effect）[84]，极易获取数量巨大的播放量。



图4-X

接下来，我们进一步探讨能否用一个数学模型来描述视频长期播放量的整体分布。在现存工作中[58, 80]，有研究者使用使用Weibull分布[40]和Log-Normal分布[85]来拟合不同数据源的网络视频播放量分布。在我们的研究中，我们则使用了这两个分布以及Pareto Type 2分布[25]进行尝试。三种分布的概率密度函数具体如下：

(6-1)

(6-2)

(6-3)

其中为形状参数，为尺度参数，和分别为均值和标准差。基于我们的数据集，我们对三个模型分别进行回归，并在图6-X中给出了相应的P-P图。从图中可以看出，Pareto Type 2分布的拟合效果最好：其P-P图中的大多数散点都分散在直线附近。故对于优酷视频的长时期播放量分布，Pareto Type 2模型是上述三个模型中最好的近似。



图6-X

* + 1. 流行度级别

根据视频的长期播放量，我们进一步划分出不同的流行度级别，来描述数据集中新发布的视频在一段时间之后受欢迎的程度。由于视频长期播放量的分布并不均匀，不同流行度级别的划分大小也不应相等。事实上，低流行级别应该覆盖大部分的视频，而高流行级别应该能够突出具有极大播放量的视频。在我们的研究中，我们将视频长期流行度划分为4个不重合的范围，如表6-1所示。其中，每个流行度级别的范围都要比前一个等级的范围大将近一个数量级。各个流行度级别分别约占总视频数的50%、35%、10%和5%。不同类型的视频，在网络运营商和业务提供商处，具有着不同重要性和商业价值。相较而言，流行视频（Level 4）虽然数量占比很小，却比非流行视频（Level 1、Level 2、Level 3）要重要的多。

表6-1 视频长期流行度级别划分概要

|  |  |  |  |
| --- | --- | --- | --- |
| **流行度级别** | **播放量范围** | **视频数比例** | **说明** |
| Level 1 |  | 51.03% | 极其不流行视频 |
| Level 2 |  | 32.67% | 不流行视频 |
| Level 3 |  | 11.64% | 边界视频 |
| Level 4 |  | 4.66% | 流行视频 |

需要指出的是，对网络视频长期流行度级别的划分是受数据集影响的。即我们的研究中的对优酷数据集的各等级划分范围，可能并不适用于其他网络视频业务提供商的数据。但根据现存的研究工作和我们在前一小节的分析结果，互联网内容的长期流行度分布都呈现出重尾特性。因此，不考虑具体的范围数值，我们所提出的划分思路在其他的数据集中是可以通用的。即对互联网内容的长期流行度进行划分时，各流行度级别的大小应该不同。低流行度的级别范围要大，并覆盖大量网络内容；而高流行度的级别范围应很小，以突出流行度数值极大的少数内容。

* + 1. 流行度影响因子

接下来，我们将视频的静态属性作为其长期播放量的潜在影响因子进行探寻。我们希望揭示具有哪些特点的视频更可能成为热门视频。我们衡量了发布后30天的视频播放量与一些数值类型的视频属性之间的Pearson相关系数，具体如下：视频标题长度（0.002）、视频内容描述长度（0.015）、视频内容标签数（0.003）、视频时长（0.034）、视频清晰度（0.137）、上传者发布的视频数（0.003）等。这些视频属性和视频长期播放量之间并没有观察到显着的相关性。但是，我们发现视频类型这一类别属性对视频流行度有很大影响。不同类型的视频，在获取播放量的能力上差异可能很大。举例来说，图6-X显示了三个具有代表性的视频类型：“家庭（Family）”、“娱乐（Entertainment）”和“游戏（Game）”，所对应的视频长期播放量累积分布函数。我们可以观察到，在不同类别的视频之间显着的流行度差异。平均而言，三个类别的视频分别被观看了865，9786和21780次。“家庭”类别视频的观看次数通常非常少。其中，93％的“家庭”视频被观看次数少于200。将较而言，播放量少于200的“娱乐”、“游戏”视频仅占各自视频比例的约30%和15%。“娱乐”类别视频的播放量分布相对比较平均。而“游戏”类别视频的播放量往往很大。超过71%的“游戏”视频被观看了超过1000次。视频播放量的差异反映了用户对不同类型视频的兴趣差异。“家庭”类型的视频往往都是一些用户自行录制的记录视频，如婴幼儿的日常生活录像等。这些视频被上传至网站，以便长期保存并与亲友分享。但陌生用户则很少会对这些私人视频感兴趣。而对于“娱乐”类型视频，其内容更加多样化并且受众较广，因此吸引用户的能力各不相同。至于“游戏”视频，其中一大部分是电子游戏的录像与评论。这些内容在年轻用户中非常受欢迎，因此往往能够保证较大的播放量。



图6-X

此外，我们还分析了视频内容标签对于视频长期播放量的影响。我们首先提取了出现在我们的数据集中超过10个视频的“热门”标签。我们共计提取到563个热门标签，仅占据总标签数的2.51%。由此可以看出，大多数的视频内容标签并不会频繁的出现。然后我们进一步分析了流行视频（Level 4）与这些热门标签之间的关系。我们发现，尽管多数视频标签在不同流行度视频中都出现过，有97种热门标签只存在于流行视频中。例如，对于热门标签“幸存者游戏”，其出现在我们数据集中28个视频中，而这些视频全部都是流行视频，其播放量平均值为50833，最大值为123943。由此可见，视频的内容标签可以在一定程度上影响其长期播放量。

* 1. 视频单体流行度分析

在本节中，我们从视频单体的角度，分析了各个视频在观察期内播放量获取情况。首先，我们通过考察视频每天获取的播放量，来定义活跃天的概念。接下来，基于活跃天出现的位置，我们进一步探寻了各视频的活跃期。此外，我们对视频在观察期各天内获取播放量的均匀程度进行了衡量。最后，我们提出了播放量增长模式的概念，来描述各个视频的流行度演化趋势。

* + 1. 活跃天

以天为时间粒度，我们发现视频在某些天中，会被用户非常广泛的观看；而在另外一些天中，则仅能获取极为有限的播放量。为了衡量一个视频单体在不同天中获取播放量的能力，我们提出了活跃天的概念。具体来讲，如果视频在某天能够获取足够的播放量，即超过预定义的阈值，则我们认为该天是该视频的一个活跃天。否则，如果视频在某天获取的播放量小于，则该天为该视频的一个非活跃天。可以看出，对于活跃天的概念，一个关键的问题是如何为每个视频单体定义一个合适的。对此，我们首先定义一个绝对数量参数，代表一个视频在活跃天中应获取的播放量的最低标准。我们改变的取值从0至2000，来计算数据集中所有视频的活跃天数量，如图6-X所示。从图中可以看出，当取值于较小的数值区间时（约在500以下），随着的增加，总活跃天数的减少十分剧烈。而当的取值较大时，总活跃天数的变化则较为平稳。因此，在我们的研究中，我们设置。



图6-X 不同数值对应的总活跃天数

然而，如6.4.1小节中的分析显示，超过80%的视频在发布30天后播放量少于1000，即平均日播放量不到34，远小于阈值。这表明仅使用绝对数量参数来定义活跃天，对于不那么流行的视频来说可能过于严格了。为解决这一问题，我们同时从绝对数值和相对占比的角度，进一步引入了两个参数和，来衡量不那么流行的视频的活跃天。具体来讲，如果一个视频的日播放量满足：1）超过次；2）超过其平均日播放量倍，我们则也可以将该天作为该视频的活跃天。图6-X显示了不同下，与总活跃天数之间的关系。从图中可以看出，不同对应的曲线，形状是非常相似的。当取值于较小的数值范围时，随着的增加，总活跃天数的减少十分剧烈。而当取值相对较大时（约40以上），总活跃天数的变化较为平稳。因此，在我们的研究中，我们设置。而对于的选择，我们发现当数值较小时，图中两条相邻曲线的间距较大。而当数值较大（约1.6以上）时，相邻的曲线近乎重合。由此，在我们的研究中，我们设置，以获取较为平稳的总活跃天数。综上，我们将一个视频单体在其活跃天中应获取的播放量阈值定义为：

(6-X)

其中，为观察期的总天数；为视频在第天时获取的总播放量；，，。



图6-X 不同与数值对应的总活跃天数

基于上述定义，我们对数据集中各个视频的活跃天进行了识别。我们发现有17076个视频在观察期内不存在活跃天。我们将这些视频视为非活跃视频，而其余视频则被视为活跃视频。显然，非活跃视频具有较差的吸引用户的能力。对于业务提供商和网络运营商而言，这些视频在实际应用中的价值是非常少的。因此，我们的分析主要关注于活跃视频。图6-X显示了我们数据集中活跃视频的活跃天数累积分布函数。从图中可知，大多数视频等活跃天数较小。具体来讲，约25%的视频在观察期内仅有1个活跃天，而近80%的视频总活跃天数短于7天。我们在图中还同时给出了流行视频（Level 4）和非流行视频（Level 1、Level 2、Level 3）的活跃天数累积分布函数。我们可以看出，非流行视频通常具有较少的活动天数，而流行视频的活动天数则往往较大。多达92.85%的非流行视频具有活跃天数不超过5。相比之下，流行视频对应的比例则只有35.41％。同时，约40%的流行视频则具有超过10天的活跃天数。这些额外的活跃天给视频带来了更多的播放量，从而使视频变得流行。



图6-X 视频活跃天数累积分布函数

* + 1. 活跃期

接下来，我们进一步对各视频的活跃天在观察期内出现的时间位置进行分析。图6-X显示了整个观察期中，各天作为活跃天时对应的视频总数、流行视频数以及非流行视频数的直方图。从图中我们可以清楚的看出，对于非流行视频，大多数的活跃天集中于视频刚刚发布的一段时期；而对于流行视频，其活跃天的分布在观察期内较为均匀，在早期的计数略高于其他时期。



图6-X 不同位置活跃天对应的视频数直方图

基于视频各活跃天的出现位置，我们提出了视频活跃期的概念。视频活跃期用于描述一个新上传的视频能够在多长时间上积极的吸引用户的关注。具体来讲，如果一个视频从发布后第天起，连续超过天未能出现一个活跃天，则我们认为其活跃期在第天结束，即该视频的活跃期长度为天。在我们的研究中，我们令。图6-X显示了我们数据集中视频的活跃期长度的累积分布函数。从图中，我们可以发现视频在短活跃期和长活跃期两极分化现象。具体来讲，52.47%的视频对应活跃期不超过5天；而24.76%的视频对应活跃期在25天以上。其余视频的活跃期介于两者之间，仅占22.76％。全体视频的平均活跃期为12.11天。在图6-X中，我们同时也给出了流行视频和非流行视频的活跃期累积分布函数。整体而言，非流行视频的活跃期要短于流行视频。53.31%的非流行视频对应活跃期仅为1天，65.47%的非流行视频活跃期不超过5天。相比之下，流行视频在相应情形下的占比仅为6.23%和43.85%。这表明，非流行视频仅能在发布后的几天之内吸引到用户的注意，而流行视频则可以保持对用户的吸引力很长一段时间。



图6-X

对视频活跃期的深入理解，有助于业务提供商和网络运营商更好的进行业务调整和资源分配。例如，基于上文分析结果，我们可以得出：对新上传视频进行广告投放或缓存部署的黄金时期是其刚发布后的几天之内。仅仅在5天之后，超过一半的视频将不再被用户活跃观看。此时，这些视频的商业应用价值将大大降低。

* + 1. 日播放量分布

在本小节中，我们关注于视频获取的总播放量在观察期各天中的分布情况。我们发现有的视频在某几天中会获得的比其余天大多的的播放量；而有的视频在整个观察期的各天中获得的播放量都相差不大。为了定量的分析视频的日播放量分布的均匀性，我们提出了视频的播放量熵的概念。具体来讲，令表示视频在发布后第天获取的播放量（增量），表示视频在发布天后总共获取的（累积）播放量。显然，我们有。对于整个观察期天中，我们定义视频在发布后第天的播放量增长率。则视频的（标准化）播放量熵为：

(6-X)

标准化后的熵值在0至1之间。如果视频的播放量在不同天之间均匀分布，则熵值接近于1；否则，如果视频的播放量集中在少数几天内，熵值将接近于0。图6-X显示了我们数据集中所有视频在观察期内的标准化播放量熵的直方图和累积分布函数。可以注意到，熵的分布整体上趋于向小的数值偏斜。具体来讲，65.25%的视频的熵值都小于0.5，而8.13%的视频的熵值甚至为0。这表明很大一部分视频是在少数几天之中获取了其观察期内的绝大多数播放量。



图6-X

为了描述视频在某些天中播放量的爆炸式增长，我们提出了播放量激增这一概念。具体来讲，对于视频在发布后第天获取的播放量，若其大于一个预定义阈值，则认为该视频该天的播放量增长趋势为激增（burst）；否则，则认为该视频该天的播放量增长趋势为缓慢（slow）。阈值定义如下：

(6-X)

其中，为观察期的总天数；为视频在第天时获取的总播放量；为一个相对于平均日播放量（）的尺度参数。为获取合适的数值，我们变换的取值从1.6到8.0，并计算了数据集中所有视频的激增趋势数量，如图6-X所示。从图中可以看出，激增趋势的数量随着取值的增加而减少。当取值较小时（4.0以下），激增数降低幅度剧烈；而当取值较大时，激增数降低较为平稳。在我们的研究中，我们设定。



图6-X 不同数值对应的总激增趋势数

图6-X显示了我们数据集中所有视频具有激增趋势的天数累积分布函数。我们发现，大多数的视频（92.37%）在观察期内都会经历播放量的激增。其中，42.57%的视频仅经历1次激增；而39.09%的视频则经历了2次激增。而经历了3次及以上播放量激增的视频则仅占总视频数的10.71%。



图6-X

* + 1. 播放量增长模式

在本小节中，我们关注于视频单体的流行度是如何随时间而进行演化的。我们希望归纳出若干主要模式，来对视频播放量的演化趋势进行描述。受前一小节的分析结果启发，基于播放量激增的数量与时间位置，我们定义了视频单体的播放量增长模式。具体来讲，对于一个视频单体，我们计算其在观察期内各天的播放量增长趋势序列，记为。其中，若视频在第天的播放量增长趋势为burst，则；若视频在第天的播放量增长趋势为slow，则。然后，我们合并相邻天中的相同趋势，以生成最终的播放量增长模式。例如，对于趋势序列，我们合并连续出现的1和0后，得到最终的播放量增长模式“10”，即burst-slow。此模式表明，该视频的播放量在其发布初期经历了一个爆炸式增长，并在之后增长放缓。对于播放量增长模式为“0”，即在观察期内未经历激增的视频，我们定义其播放量的增长模式为steady。

然而，我们发现在视频的播放量增长趋势序列中，往往存在一些随机扰动。例如，一个激增后平稳的播放量增长模式（burst-slow），可能会对应趋势序列，即序列中存在一天的“噪声”趋势“0”（slow）。如果我们直接合并此序列来生成播放量增长模式，最终的结尾为“1010”（burst-slow-busrt-slow），而非我们所期望的“10”（burst-slow）。为解决这一问题，我们设计了一个平滑算法，在合并步骤之前应用在趋势序列上，以滤除序列中的扰动。该算法使用长度为的窗口来遍历趋势序列，并根据窗口中的各状态的值和尺度参数，来重新确定窗口中最后一个趋势的输出状态。算法的伪代码见代码6-1所示。值得注意的是，我们对视频播放量增长模式的定义方式，与现存的研究工作[52, 54, 59, 60]有所不同。在那些工作中，作者主要关注于视频播放量在（单一的）高峰日期间的增长形状；而在我们的研究中，我们考虑了视频在整个观察期内的播放量演化趋势。

代码6-1 播放量增长趋势序列的平滑算法

|  |  |
| --- | --- |
| 1: | **Input**: |
| 2: | **Output**: |
| 3: | **procedure**: SmoothFiltering() |
| 4: |  |
| 5: | **for** **to** **do** |
| 6: | **if** **then** |
| 7: |  |
| 8: | **else** |
| 9: |  |
| 10: | **end if** |
| 11: | **end for** |
| 12: | **for** **to** **do** |
| 13: | **if** **then** |
| 14: |  |
| 15: | **else** |
| 16: |  |
| 17: | **end if** |
| 18: | **end for** |
| 19: | **end** **procedure** |

表6-X列出了我们数据集中按视频数排名top 5的播放量增长模式，覆盖了96.51%的视频。而在图6-X中，我们给出了各模式的视频（累积）播放量随时间变化的典型曲线。从表中我们可以看出，排名前两位的burst-slow和burst-slow-burst-slow是主导模式，其覆盖了超过88%的视频。我们注意到这两种模式都是以burst开头，即大多数视频在发布初期都是经历一个播放量的激增过程。经分析，我们认为这种现象的两个主要原因如下：1）在优酷中，视频在发布后的最初几天中，将会被展示在网站的“最近发布视频”处。此时，视频更可能被用户注意到，从而快速的累积播放量。2）在视频刚刚发布时，上传者进行分享与推广行为的积极性更高。由此，上传者的社交影响力可以为视频带来大量的初始播放量。在表6-X中，我们还发现了数量不容忽视的steady视频。这些视频在整个观察期内稳步获取播放量，其内容往往是能够持续吸引用户的热门话题。此外，有的播放量增长模式在视频的活跃期中间出现了激增。尤其是slow-burst-slow模式，其并未以burst开头。这种中期出现的播放量激增通常与外部影响因素有关，如视频被某些“网络大V”转发到了微博等在线社交网络中。这些影响能够引起播放量的反弹，使得一个不再流行的视频重新变得流行起来[68, 69]。

表6-X 视频播放量增长模式概要

|  |  |  |  |
| --- | --- | --- | --- |
| **排名** | **合并后状态序列** | **播放量增长模式** | **视频数占比** |
| 1 | 10 | burst-slow | 72.51% |
| 2 | 1010 | burst-slow-burst-slow | 15.64% |
| 3 | 010 | slow-burst-slow | 4.02% |
| 4 | 101010 | burst-slow-burst-slow-burst-slow | 2.81% |
| 5 | 0 | steady | 1.54% |



图6-X

接下来，我们进一步分析了不同流行度级别之中的视频播放量增长模式。表6-X给出了我们数据集中各流行度级别的视频中，播放量增长模式的前5名。从表中可以看出，播放量增长模式的分布受到视频流行的级别的影响。对于极不流行的Level 1视频，其播放量增长的波动性很强。整体而言，其增长模式中的激增数量要比其他的流行度级别要多。这是因为Level 1视频的总播放量较小（100以下），在日播放量的获取上，一个较小的绝对数值往往对应着一个较大的相对比例，从而体现为一次激增。而对于Level 2、Level 3和Level 4视频，我们发现其top 5的播放量增长模式类型相同，仅在比例上有所不同。对于相对不流行的Level 2和Level 3视频，播放量增长模式的分布向top 1集中；而对于Level 4流行视频，其分布则向各个模式扩展较多。这表明视频流行度越大，其播放量的增长模式就越复杂。此外，我们还发现随着视频流行度的增高，steady模式的视频占比也在不断增大。这是符合预期的，因为连续稳定的日播放量的获取，往往可以导致视频最终的累积播放量很大。

表6-X 各流行度级别的视频播放量增长模式概要

|  |  |  |  |
| --- | --- | --- | --- |
| **排名** | **合并后状态序列** | **播放量增长模式** | **视频数占比** |
| **Level 1视频** | | | |
| 1 | 10 | burst-slow | 68.54% |
| 2 | 1010 | burst-slow-burst-slow | 24.12% |
| 3 | 101010 | burst-slow-burst-slow-burst-slow | 3.36% |
| 4 | 10101010 | burst-slow-burst-slow-burst-slow-  burst-slow | 1.93% |
| 5 | 010 | slow-burst-slow | 0.85% |
| **Level 2视频** | | | |
| 1 | 10 | burst-slow | 81.66% |
| 2 | 1010 | burst-slow-burst-slow | 10.29% |
| 3 | 010 | slow-burst-slow | 3.67% |
| 4 | 0 | steady | 1.07% |
| 5 | 01010 | slow-burst-slow-burst-slow | 0.82% |
| **Level 3视频** | | | |
| 1 | 10 | burst-slow | 73.20% |
| 2 | 010 | slow-burst-slow | 9.57% |
| 3 | 1010 | burst-slow-burst-slow | 8.13% |
| 4 | 0 | steady | 4.32% |
| 5 | 01010 | slow-burst-slow-burst-slow | 1.17% |
| **Level 4视频** | | | |
| 1 | 10 | burst-slow | 46.56% |
| 2 | 010 | slow-burst-slow | 30.82% |
| 3 | 0 | steady | 12.99% |
| 4 | 01010 | slow-burst-slow-burst-slow | 4.45% |
| 5 | 1010 | burst-slow-burst-slow | 4.13% |

* 1. 未来流行度级别预测
     1. 问题定义

在本节中，我们关注于视频单体的未来流行度级别预测这一研究问题。对于一个新发布的视频，我们希望能够估计其在不远将来（near future）的大体流行度程度。由此，基于我们在6.4.2小节的分析，我们将视频在发布后第30天时的播放量划分为4个不重合的范围，定义成4个流行度级别（Leve 1～4）并将此作为预测对象。可以看出，对网络视频的未来流行度级别预测本质上是一个分类问题。在后文中，我们将交互使用预测与分类这两个术语。

我们对于目标问题，分别考虑了两种不同的预测情形：1）在视频的发布时刻进行预测；和2）在初始观察期后进行预测。在情形1中，我们可获取的用于预测的数据仅有视频发布时给出的视频元信息和上传者元信息。而对于情形2，当视频发布后，我们先对其观察一段短暂的时期，然后再进行预测。此时，我们能够获取的用于预测的数据，除了视频元信息与上传者元信息外，还包括在初始观察期内追踪的该视频的动态属性信息。

在对预测模型的构造中，我们使用了经典的机器学习分类算法作为分类器，而非提出一种新的分类技术。这是因为，根据实验结果，这些现有方法的性能已经足够好，可以满足大多数应用需求[70]。此外，我们的研究兴趣并不在于进一步优化分类算法结果。与此相反，我们的目标是评估不同的分类方法，在流行度级别预测这一新领域问题上的相对性能差异。并且，我们关注于探寻不同特征和特征组在此分类问题上的效用与重要性，以揭示网络视频流行度的影响因子。这些分析结果将对业务设计、内容生产、广告投放等实际应用提供有意义的决策指导与帮助。

* + 1. 特征介绍

在本小节中，我们对流行度级别预测中所使用的各特征进行介绍。在现实世界中，网络视频的流行度会受到各种各样的因素所影响。而我们研究的一个关键贡献，就是鉴别出了多种在视频流行度预测问题中需要考虑的因素。在我们的预测模型中，这些影响因素通过分类器所使用的特征来体现。我们从5个不同的角度，共提出192个特征。完整的特征列表见表6-X所示。对各组特征的介绍如下。

**视频属性特征**：这些特征用于捕获视频的内容属性（例如视频类型、视频时长等）和发布信息（例如视频上传时间、视频上传客户端类型等）。

**上传者属性特征**：这些特征用于捕获上传者的个人信息（例如注册时间、是否为认证账户等）、发布历史统计（例如发布的视频数、发布的视频专辑数等）和社交影响统计（例如关注的用户数、粉丝数等）。

**内容话题特征**：视频本身的内容与其流行度是密切相关的[86]。与热门话题相关的视频往往对用户更具吸引力，从而更容易获取较大的播放量。在本组特征中，我们通过检查视频标题（或标签）中的某些关键词，来捕获视频的内容话题信息。首先，我们将数据集中的视频按流行度级别进行汇聚，从而得到4组视频（Level 1～Level 4）。然后，对于每组中所有视频，我们统计视频标题（或标签）中各短语的出现次数。我们将各组中出现频次最高的前500条短语提取出来，从而得到4组短语列表。我们进一步检查各列表中的短语，并剔除1）没有实际含义的短语，如“的”、“非常”等；2）在其他列表中也出现的短语。最终，我们为每个流行度级别保留了一个长度为100的关键短语列表。各关键列表中的每个短语都只在具有相应流行度的视频的标题（或标签）中大量出现，即这些短语描述了具有特定流行度的视频内容话题。对于属于流行度级别的关键短语，我们定义其权重为：

()

其中为流行度级别的关键短语列表长度，为关键短语在数据集中的出现频次。最后，我们提出了视频在流行度级别上的关键短语分数的概念。

(6-X)

其中，为流行度级别的关键短语列表，而为视频的标题（或标签）中的短语，为该短语在标题（或标签）中出现次数。

**文本语言特征**：这些特征从语言学的角度，捕获与视频相关的文本内容的情感、结构和语法信息。优酷中的文本内容主要由中文书写。在我们的研究中，我们使用了专门面向中文的自然语言处理工具包*snownlp*[87]，来提取视频标题、内容描述和内容标签中的情感度。此外，我们还使用了中文分词工具包*jieba*[88]，来提取视频标题和内容描述中各短语的词性信息。

**历史动态特征**：这些特征用于捕获视频在初始观察期的各流行度指标动态变化情况（例如每日播放量增长的数值与比例、视频获取的评论数、收藏数等）。此外，我们还引入了视频在初始观察期的播放量增长模式作为特征。对于初始播放量增长模式的提取，我们没有合并相邻的相同增长趋势。例如，如果视频前2天的播放量增长趋势为激增burst，之后为缓慢slow，则其初始播放量增长模式为“1100…0”。这样做的原因是初始观察期十分短暂，我们需要突出每一天演化趋势，以更好的进行区分。

表6-X 网络视频流行度级别预测特征列表

|  |  |  |
| --- | --- | --- |
| **序号** | **名称** | **说明** |
| **视频属性特征** | | |
| 1 | 视频类型 | 类别类型，优酷预定义的24种类型之一。 |
| 2 | 视频时长 | 数值类型，以秒为单位的视频总时间长度。 |
| 3 | 视频内容标签数 | 数值类型，视频内容标签的个数。 |
| 4 | 视频清晰度 | 类别类型，表明视频分辨率为“标清”、“高清”还是“超清”。 |
| 5 | 视频文件格式 | 类别类型，优酷支持的10种文件格式之一。 |
| 6 | 视频版权类型 | 类别类型，表明视频是由上传者“原创”还是“转载”的。 |
| 7 | 视频公开类型 | 类别类型，表明视频是“完全公开”、“仅向上传者好友公开”还是“输入密码访问”的。 |
| 8 | 视频上传时间 | 时间类型，视频被上传至优酷的UTC时间。 |
| 9 | 视频上传客户端类型 | 类别类型，优酷支持的10种常见客户端类型。 |
| 10 | 视频发布状态 | 类别类型，表明视频当前状态是“正常”、“转码中”、“失败”还是“被屏蔽”。 |
| **上传者属性特征** | | |
| 11 | 用户性别 | 类别类型，表明上传者是“男性”、“女性”还是“未填写”。 |
| 12 | 注册时间 | 时间类型，上传者的账号注册时间。 |
| 13 | 是否为会员账户 | 布尔类型，表明上传者的账号是否属于优酷会员帐户。 |
| 14 | 是否为共享账户 | 布尔类型，表明上传者的账号是否为多机构公用帐户。 |
| 15 | 是否为认证账户 | 布尔类型，表明上传者的账号是否为优酷认证的官方视频内容发布者。 |
| 16 | 发布的视频数 | 数值类型，上传者发布过的总视频数。 |
| 17 | 发布的视频专辑数 | 数值类型，上传者发布过的总视频专辑数。 |
| 18 | 发布的状态数 | 数值类型，上传者发布过的总社交状态数。 |
| 19 | 关注的用户数 | 数值类型，上传者主动关注的用户数。 |
| 20 | 粉丝数 | 数值类型，关注了上传者的用户数。 |
| 21 | 获取的总播放量 | 数值类型，上传者发布过的所有视频所获取的播放量总量。 |
| 22 | 获取的总收藏量 | 数值类型，上传者发布过的所有视频所获取的收藏量总量。 |
| **内容话题特征** | | |
| 23-26 | 视频标题话题 | 数值类型，视频标题在4个流行度级别上获取的关键短语分数。 |
| 27-30 | 视频内容标签话题 | 数值类型，视频内容标签在4个流行度级别上获取的关键短语分数。 |
| **文本语言特征** | | |
| 31 | 视频标题情感度 | 数值类型，视频标题文本的情感度分数。取值0到1之间，0为极负面情况，1为极正面情感。 |
| 32 | 视频内容描述情感度 | 数值类型，视频内容描述文本的情感度分数。 |
| 33 | 视频内容标签情感度 | 数值类型，视频内容标签文本的情感度分数。 |
| 34 | 视频标题长度 | 数值类型，视频标题的字符数。 |
| 35, 36 | 视频标题中文字符 | 数值类型，视频标题中，中文字符的数量和比例。 |
| 37, 38 | 视频标题非中文字符 | 数值类型，视频标题中，非中文字符的数量和比例。 |
| 39, 40 | 视频标题字母 | 数值类型，视频标题中，字母字符的数量和比例。 |
| 41, 42 | 视频标题数字 | 数值类型，视频标题中，数字字符的数量和比例。 |
| 43, 44 | 视频标题标点符号 | 数值类型，视频标题中，标点符号字符的数量和比例。 |
| 45 | 视频内容描述长度 | 数值类型，视频内容描述的字符数。 |
| 46 | 上传者用户名长度 | 数值类型，上传者用户名的字符数。 |
| 47, 48 | 上传者用户名中文字符 | 数值类型，上传者用户名中，中文字符的数量和比例。 |
| 49, 50 | 上传者用户名非中文字符 | 数值类型，上传者用户名中，非中文字符的数量和比例。 |
| 51, 52 | 上传者用户名字母 | 数值类型，上传者用户名中，字母字符的数量和比例。 |
| 53, 54 | 上传者用户名数字 | 数值类型，上传者用户名中，数字字符的数量和比例。 |
| 55, 56 | 上传者用户名标点符号 | 数值类型，上传者用户名中，标点符号字符的数量和比例。 |
| 57-114 | 视频标题词性 | 数值类型，视频标题文本中的短语对应58种常见词性的各类别计数。 |
| 115-172 | 视频内容描述词性 | 数值类型，视频内容描述文本中的短语对应58种常见词性的各类别计数。 |
| **历史动态特征** | | |
| 173-179 | 视频日播放量（增量） | 数值类型，视频在初始观察期（7天）中各天获取的播放量（增量）。 |
| 180-186 | 视频日播放量增长率 | 数值类型，视频在初始观察期中各天的播放量增长率。 |
| 187 | 视频初始播放量增长模式 | 类别类型，视频在初始观察期的播放量增长模式。 |
| 188 | 视频播放量 | 数值类型，视频在初始观察期内获取的播放量。 |
| 189 | 视频评论量 | 数值类型，视频在初始观察期内获取的评论量。 |
| 190 | 视频收藏量 | 数值类型，视频在初始观察期内获取的收藏量。 |
| 191 | 视频被顶次数 | 数值类型，视频在初始观察期内获取的被顶次数。 |
| 192 | 视频被踩次数 | 数值类型，视频在初始观察期内获取的被踩次数。 |

* + 1. 实验分析

基于上述特征，我们使用多种高效分类算法，在不同情形下对我们数据集中视频的未来流行度级别进行预测分析。

**实验设置**：我们使用6.3节介绍的优酷数据集进行实验，包含200,773个视频的元信息及播放量序列。我们将数据集随机的划分为训练集和测试集两部分，各占总数据量的50%。训练集用于对分类模型的构建；而测试集用于对各模型预测性能的评估。

我们将各视频在发布后第30天的流行度级别作为预测目标。我们从视频的元信息中可以直接提取视频属性特征、上传者属性特征和文本语言特征；而从视频播放量序列中，我们可以进一步提取历史动态信息。对于内容话题特征，我们首先从训练集中，为各个流行度级别统计其视频标题（或标签）的关键短语列表；然后为测试集中的各个视频，我们计算其关键短语分数。

为解决预测问题并比较性能，我们在实验中使用了不同的机器学习分类算法作为分类器，包括：K最近邻（k nearest neighbor，KNN）[89]、决策树（decision tree，DT）[13]、支持向量机（support vector machine, SVM）[12]、随机森林（random forest，RF）[90]以及梯度提升（gradient boosting，GB）[91]方法。

在我们的实验中，我们使用精度（precision）和召回率（recall）作为各类别（即视频未来流行度级别）的预测性能评价指标。其中，精度为“正确预测属于的某流行度级别的视频数量”与“检测结果中该流行度级别中视频的总数量”的比值。而召回率则是“正确预测属于的该流行度级别的视频数量”占“该流行度级别中视频总数”的比例。然而，根据6.4节的分析结果，我们知道视频播放量的整体分布是十分不均匀的，各流行度级别中的视频数量差距很大。Level 1视频的比例超过50%，而Level 5视频则仅占不到5%。对于业务提供商和网络运营商而言，流行度高的视频往往具有更大实际应用价值。因此，在对模型进行整体的性能评价时，我们使用了精度和召回率在各流行度级别上的宏平均值（macro-averaged values），而非对全体视频的平均值。按预测类别进行宏平均的精度与召回率具体计算方式如下：

(6-X)

(6-X)

此外，在计算模型超参数时，我们同时考虑精度与召回率两个指标，并使用（宏平均）值作为评价指标。值是对精度和召回率的加权平均，具体计算方式如下：

(6-X)

我们使用网格搜索和3折交叉验证，通过极小化宏平均值来确定各模型中的超参数。

**在发布时预测**：我们首先探寻能否在视频刚刚发布时立即预测其未来的流行度级别。这是一个非常有挑战性但也非常有实用价值的情形。此时，我们仅能利用视频元信息和上传者元信息。为解决此问题，我们使用视频属性特征（V）、上传者属性特征（U）、内容话题特征（T）和文本语言特征（L）构建各分类模型，并评估它们的预测性能。我们使用普通最小二乘法（OLS）回归模型作为基线方法（baseline，BSL）。该方法在文献[70]中被用来对Foursquare中的内容流行度级别进行预测。对于分类任务，OLS模型的数值输出被映射到相应的流行度级别中。

图6-X显示了使用了视频发布时所有特征的各分类模型的宏平均精度和宏平均召回率。对于SVM分类器，我们分布尝试了线性核函数和径向基核函数。为简洁起见，我们仅给出了二者中的最佳结果（线性核函数）。从图中可以看出，所有的分类方法都优于简单的基线方法。这表明我们提出的特征对于流行度的预测是非常有效的。RF分类器的预测性能最好，其宏平均精度达74.02％，而宏平均召回率达59.74％。GB分类器在宏平均精度上与RF分类器相差不大，然而在宏平均召回率上其性能要略差。KNN分类器和DT分类器都可以达到约65％的宏平均精度，而在宏平均召回率上DT分类器优于KNN分类器9.30％（54.82％ vs 45.52％）。在5种分类器中，SVM的性能最差，其宏平均精度为55.51%，宏平均召回率为44.13%。此外，我们注意到分类器的宏平均精度往往要大于其宏平均召回率。经分析，我们发现这是因为高流行度级别的召回率过低。以RF分类器为例，其在Level 1和Level 2流行度级别上的召回率都超过了60%，而在Level 3和Level 4上的召回率却都不到40%。这种差异可以结合视频的播放量增长模式来解释。如6.5.4小节分析，在高流行度级别中，视频的播放量增长模式更为丰富，并且趋向于在活跃期中间产生播放量激增。这增加了这些视频的未来流行度的不确定性，使得对高流行度级别的预测更加困难。



图6-X 使用各特征组的各分类模型性能：（a）宏平均精度；（b）宏平均召回率。

接下来，我们关注于各特征组在预测中的有效性。我们进行多次分类实验，轮流移除数据集中的一组特征。图6-X中显示了在分别移除视频属性特征（-V）、上传者属性特征（-U）、内容话题特征（-T）和文本语言特征（-L）时，各分类器相应的预测性能。在宏平均精度上，当有特征组被移除时，大多分类器的性能都出现了轻微的降低（约3％）。不过，SVM分类器是个例外，其宏平均精度反而略有上升。但在宏平均召回率上，SVM分类器的性能则出现了大幅度下降。并且，当移除不同组特征时，各分类器宏平均召回率的下降程度不同。我们发现上传者属性特征是最具区分度的特征组。当其被移除时，宏平均召回率的下降最大，约10%。而移除其余的特征组时，都会出现约3％的宏平均召回率降低。这意味着这些特征组都对流行度级别的预测做出了贡献。我们进一步分析了各特征在预测任务中的相对重要性。我们根据构建RF分类器时的节点纯度（由Gini系数衡量），来对特征进行排序。表6-X列出了前十个最具有区分度的特征。我们发现排名靠前的特征，如“上传者获取的总播放量”、“上传者发布的视频数”、“上传者粉丝数”都与上传者的发布历史信息和社交影响力有关。这也与上文对特征组重要性的分析相符。

表6-X 特征重要性Top 10概览

|  |  |  |
| --- | --- | --- |
| **排名** | **特征** | **重要性** |
| 1 | 上传者获取的总播放量 | 0.1173 |
| 2 | 上传者发布的视频数 | 0.1062 |
| 3 | 上传者粉丝数 | 0.0628 |
| 4 | 视频种类 | 0.0579 |
| 5 | 视频时长 | 0.0543 |
| 6 | 上传者获取的总收藏量 | 0.0375 |
| 7 | 上传者注册时间 | 0.0244 |
| 8 | 视频标题中文字符数 | 0.0201 |
| 9 | 视频标题情感度 | 0.0179 |
| 10 | 视频上传时间 | 0.0143 |

**在初始观察期后预测**：在前文研究中，我们关注于最困难的预测情形，即在视频发布时进行预测。理想情况下，人们希望在视频发布后立即确定其未来播放量情况。然而，如6.5.4小节分析，不同视频可能会经历不同的播放量增长模式，进而达到完全不同的流行度级别。通过对视频单体在发布后进行短时期的播放量追踪，我们尝试捕获视频的播放量增长模式信息，并将之应用在未来流行度级别预测中。为此，我们向分类任务中引入了历史动态特征组（H），并分析了各分类器的预测性能提高情况。我们选择了一个简单的仅利用历史动态特征的多元线性回归模型作为基线模型。该模型在文献[75]被用来预测YouTube视频流行度。

图6-X同样给出了在初始观察期后的预测情形下（+H），各分类器的宏平均精度和宏平均召回率。可以看出，相较于在视频发布时预测的情形，各分类器的预测性能有了大幅度的提高。其中，RF分类器和GB分类器的性能近似，并优于其他方法。二者都取得了约95％的宏平均精度和约91％的宏平均召回率。DT分类器也表现良好，其宏平均精度和宏平均召回率分别为90％和89％。而SVM分类器的性能仍然较差，这表明线性核函数和径向基核函数都并不适合我们的分类任务。在未来的工作中，应尝试使用其他的核函数。另外，我们发现KNN分类器的性能提升幅度较小，其在所有方法中表现最差。这是由于视频的播放量增长模式是十分复杂的。如6.5.4小节中分析，一些视频会在活跃中期经历播放量激增。在初始观察期中流行度演化趋势近似的视频，可能会在之后经历不同的播放量增长，从而对应不同的未来流行度级别。然而，由于KNN分类器是基于特征域中的相似性进行分类的，其并不能很好的适用于这种情况。此外，我们发现对于仅使用了历史动态信息的基线方法，其在预测中表现已经相当好，并优于前一预测情形中的所有模型性能。这表明历史动态信息是预测视频未来流行度的主要特征。

接下来，我们分析初始观察期的长度对预测性能的影响，并探寻平衡及时性与准确性的合适长度。我们考虑初始观察期为1、3、5、7和9天的情况，并评估各分类器的宏平均精度和宏平均召回率，如图6-X所示。可以看出，RF分类器和GB分类器在所有情况下都达到性能最佳，其次是DT分类器。SVM分类器和KNN分类器则表现不佳，在不同观察期内的性能提升有限。对比前一预测情形的图6-X，我们发现通过追踪视频的流行度历史动态仅1天的时间，各分类器的预测宏平均精度和宏平均召回率就可分别提升近10%与20%。并且，延长初始观察期将提升两个预测性能指标。不过，我们发现自第七天开始，性能指标的提升开始放缓，对观察期长度的敏感性开始降低。考虑到及时性对于一项预测任务的重要性，在我们的分析中，我们选择7天作为初始观察期的长度。



图6-X 不同初始观察期长度的各分类模型性能：（a）宏平均精度；（b）宏平均召回率。

* 1. 基于播放量增长模式的未来流行度数值预测
     1. 问题定义

在本节中，我们关注于视频单体的未来流行度数值预测这一研究问题。对于一个新发布的视频，我们希望能够估计其在不远将来（near future）的精准流行度数值。具体来讲，对于一个视频，经过一段长度为的初始观察期后，在其发布后第天时，我们希望能够对其未来第天的播放量进行预测。此时，我们能够获取的用于预测的数据，既包括视频元信息和上传者元信息，又包括初始观察期内的视频流行度动态。在预测模型中，我们主要使用视频在初始观察期内的播放量作为预测变量。同时，我们还希望能够利用视频属性、上传者属性、视频内容话题与视频文本语言中的信息，并结合前文中对视频流行度增长模式的分析内容，来进一步提升预测方法的性能。可以看出，网络视频的未来流行度数值预测本质上是一个回归问题。在后文中，我们将交互使用预测与回归这两个术语。

* + 1. 早期-长期播放量关系

在之前的工作中[47] [75]，研究者发现互联网内容的早期流行度在一定程度上可以反映其长期流行度。一般来说，发布后立即获取了大量用户注意力的内容，很可能在未来成为流行内容。反之，较小的早期流行度通常对应于不受欢迎的内容。在我们的研究中，我们对优酷视频数据集是否具有这一性质进行了检查。图6-X给出了优酷中新上传视频在第7天的播放量（作为早期流行度）与其第30天的播放量（作为长期流行度）之间的关系。从图中我们可以观察到粗略的线性关系。通过对数据进行回归，我们可以进一步得到参数和。具体的回归直线在图中由红色虚线标出。但是，从在图中我们还发现了大量的点，其并未分散在回归直线附近。即，对许多早期流行度并不高的视频，其长期流行度却相当的高。由此，我们可以断定，基于早期-长期播放量的简单线性模型，并不能很好的胜任对网络视频未来播放量预测的任务。



图6-X

根据前文6.5.4小节的分析内容，我们可以推测优酷视频中早期-长期播放量的复杂关系是由播放量的多种增长模式引起的。不同播放量增长模式的视频，可以在早期具有相似的播放量，但在之后显示出完全不一样的流行度演化过程。例如，在图6-X中我们给出了数据集中的三个视频，其播放量增长模式分别为steady、burst-slow和burst-slow- burst-slow。在发布后第7天，三视频的播放量几乎相同：，，。然而，在一个月后的第30天，各视频的播放量差距很大：，，。其中，最大播放量是最小播放量的3倍以上。



图6-X

综上，我们得出如下结论：视频的早期播放量可以通过一个粗略的线性关系来反映其长期播放量。但此关系受视频的播放量增长模式影响较大。因此，当基于早期播放量来回归未来播放量时，对不同类型的视频使用同一个通用模型将存在固有缺陷，无法获取良好的预测效果。

* + 1. 预测模型

为了对视频单体未来播放量进行预测，我们提出了一个基于播放量增长模式的分组多元线性回归模型（View Count Growth Pattern based Multivariate Linear regression model，VCGP\_ML）。该模型利用以下视频流行度特性作为理论基础：1）视频早期播放量与长期播放量之间的关系；2）视频播放量增长模式对于长期播放量的影响；3）视频在活跃中期经历播放量激增的可能性。

具体来讲，我们向回归模型中引入视频的流行度演化趋势：我们根据视频的早期播放量增长模式，建立了不同的专用分组回归模型。如6.6.2小节历史动态特征中的介绍，我们从初始观察期中提取视频的早期播放量增长模式。在模型中，我们考虑top 及others，共种增长模式。对于各个增长模式，我们对预测变量（即视频在初始观察期内各天的播放量）使用单独的一组回归系数（即权重）。此外，我们在模型中添加了视频可能在中期激增的播放量。我们基于6.6.2小节中介绍的视频属性特征、上传者属性特征、内容话题特征、文本语言特征和历史动态特征，并以随机森林作为分类器，对一个视频在未来是否会经历播放量激增进行了二值预测或。对于的视频，我们为其累加一定比例的播放量，来体现其中期播放量激增。

综上，对于视频在发布后第天的播放量，我们的预测结果为：

(6-X)

其中，为初始观察期的长度；为视频在第天的播放量（增量）；为该视频的早期播放量增长模式；为取决于与的回归系数；为视频在第天的（累积）播放量；为基于已获取播放量对未来激增进行描述的模型参数。

模型的最优参数可以从训练数据集中学习得到。具体来讲，我们定义模型的特征向量为：

(6-X)

而对于早期播放量增长模式，我们定义模型的参数向量为：

(6-X)

则模型可进一步表示为：

(6-X)

给定一组训练集，模型参数的最优值可通过最小化上的预测误差来计算得到。在我们的研究中，我们使用平均相对平方误差（mean relative squared error，MRSE）作为评价模型预测性能的指标。对于一组视频，预测结果的MRSE定义为：

(6-X)

由此，模型参数可获取如下：

(6-X)

令，则该优化问题可被表示为：

(6-X)

最终，最优模型参数可以简单的通过求解该加权最小二乘问题来得到。

* + 1. 实验分析

**实验设置**：我们使用6.3节介绍的优酷数据集进行实验，包含200,773个视频的元信息及播放量序列。整个检测系统的概览如图6-X所示。我们随机提取50％的数据作为训练集，并将剩余的50％作为测试集。训练集用于获取早期播放量的top 增长模式、中期播放量的激增预测模型、以及最终回归模型的最优参数；而测试集用于对预测性能进行评估（MRSE）。



图6-X 检测系统概览示意图

当对一个视频的未来播放量进行预测时，我们首先提取其早期播放量增长模式。然后，我们预测该视频是否会出现中期播放量激增。最后，根据前两步的结果，我们选用相应参数的回归模型来预测该视频的未来播放量。

实验中，我们将VCGP\_ML模型与两个目前先进（state-of-the-art）的基线模型进行比较。一个基线模型是由Szabo和Huberman在文献[47]中首次提出的对数线性模型（Log-Linear）。该模型为一个通用的一元线性回归模型，使用对数变换的早期播放量来预测对数变换的未来播放量如下：

(6-X)

其中，为模型参数由训练集得到。另一个基线模型是一个通用的多元线性回归模型（Multi-Linear），由Pinto等人在文献[75]中提出，其使用视频在观察期各天内的播放量（增量）作为预测变量，对未来播放量预测如下：

(6-X)

其中，为模型参数由训练集得到。我们使用与[47, 75]相同的场景来评估模型，即、。而对于我们模型中考虑的早期播放量增长模式的数量，我们设置。

**预测结果**：我们首先对视频中期播放量激增的预测结果进行分析。为了选择合适的预测模型，我们对比了几种高效分类器在相同训练集和测试集下的预测性能，包括：K最近邻分类器[89]、朴素贝叶斯naive Bayes分类器[92]、决策树DT分类器[13]、支持向量机SVM分类器[12]、随机森林RF分类器[90]。对于KNN分类器，我们变换值从1至9。对于SVM分类器，我们尝试了线性核函数和径向基核函数。我们在训练集上使用网格搜索和10折交叉验证，来为每个分类器选择最佳超参数。在表6-X中，我们给出了预测结果对精度、真阳性和真阴性比例。为简洁起见，我们对KNN分类器和SVM分类器仅列出了最优结果，分别对应和线性核情况。从表中可以看出，RF分类器具有最大的精度及最佳的总体预测性能。KNN、DT、SVM及RF分类器都可获取较高的真阴性比例（超过80%），但在真阳性比例上RF分类的数值则要比其他分类器高得多。而对于naive Bayes分类器，其真阳性比例虽然很高（超过90%），真阴性比例仅有约7%。因此，在我们的模型中，我们使用随机森林RF分类器来预测视频是否会经历中期播放量激增。

表6-X

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | KNN | naive Bayes | DT | SVM | RF |
| **精度** | 67.19% | 36.24% | 72.58% | 70.72% | 77.76% |
| **真阳性比例** | 32.33% | 93.69% | 40.42% | 28.96% | 47.59% |
| **真阴性比例** | 83.68% | 7.12% | 88.02% | 90.76% | 92.02% |

接下来，我们对模型的最终播放量预测结果进行分析，如表6-X所示。为了更好的理解模型性能，除了预测结果的整体MRSE，我们还分别列出了不同早期播放量增长模式的视频组所对应的MRSE值。从表中可以看出，VCGP\_ML模型要优于两个基线模型。对于整体MRSE，我们的VCGP\_ML模型在Log-Linear模型和Multi-Linear模型的基础上，分别产生了高达33.74％和11.90％的降低。相较于Log-Linear模型，VCGP\_ML模型在各个播放量增长模式视频组中都能够获得较小的MRSE。这造成了二者在整体MRSE上的巨大差异。而对比VCGP\_ML模型和Multi-Linear模型，虽然二者在前两种模式视频组中的MRSE相近，VCGP\_ML模型在其他视频组的预测性能要远胜于Multi-Linear模型。这表明Multi-Linear模型向具有某些早期播放量增长模式的视频产生了偏斜，而我们的VCGP\_ML模型则可胜任所有类型的视频。

表6-X

|  |  |  |  |
| --- | --- | --- | --- |
| **视频组类型** | **MRSE（%）** | | |
| **Log-Linear** | **Multi-Linear** | **Proposed** |
| 整体 | 9.0192 | 6.7833 | 5.9764 |
| 1000000 | 8.5836 | 5.9430 | 5.4247 |
| 1100000 | 7.5627 | 5.3927 | 4.9703 |
| 0000000 | 14.2549 | 13.9902 | 12.7569 |
| others | 13.2417 | 12.7563 | 9.8943 |

综上，我们根据播放量增长模式将视频分组，并为各视频组建立了专用的回归模型来预测视频未来播放量。我们的方法在预测性能上有较大的提高。这是因为，通过区分不同播放量增长模式的视频，我们可以在模型训练阶段减少“噪声”数据，从而为各分组专用模型得到更合适的模型参数。而对于通用的预测模型，如基线方法，所有的视频都会被共同处理，从而造成播放量增长模式对视频流行度的影响被掩盖。

* 1. 基于流行度等级转换的未来流行度数值预测

在前一节中，基于播放量增长模式，我们对视频的未来播放量进行了预测，并取得了较好的预测结果。然而，我们注意到在VCGP\_ML模型中的中期播放量激增预测部分，我们能达到的预测精度并不高，仅77.76%。对视频中期播放量激增精准预测的困难，将成为进一步提升VCGP\_ML模型性能的瓶颈。然而，在另一角度上，我们对视频未来流行度级别的预测却是很成功的，如6.6节所示。而且，视频流行度级别的变化，本质上能够涵盖播放量激增所造成的影响。因此，在本节中，我们利用对视频未来流行度级别的预测，提出了一个基于流行度等级变换的分组多元线性回归模型（Popularity Level Transition based Multivariate Linear regression model，PLT\_ML），来预测视频的未来播放量。对模型的具体介绍如下。

* + 1. 预测模型

我们的预测模型共包括两个阶段。首先，我们基于丰富特征和高效分类算法来估计视频的未来流行度级别。然后，根据从早期流行度级别到未来流行度级别的转换，我们构建分组专用的回归模型来预测视频的未来播放量。

对于第一阶段的未来流行度级别的预测，我们直接使用了6.6节中的研究成果。基于多角度特征和高效分类算法，模型的预测性能可以达到约95％的宏平均精度和约91％的宏平均召回率。

而对于第二阶段的基于播放量级别转换的未来播放量预测，我们考虑了两种情形：1）未来和早期的流行度等级保持不变；和2）未来较早期的流行度等级发生改变。

**情形1**：如果视频被预测的未来流行度级别与其在观察期（最后一天）的流行度级别相同，则其播放量的积累主要受观察期内的增长趋势所影响。以图6-X显示的我们的数据集中的三个视频为例。对于视频1和视频2，其早期和长期的流行度级别都是Level 1，但播放量的增长模式不同。视频1的播放量增长以一个激增开始，然后逐渐放缓；而视频2的播放量则保持持续增长。因此，尽管在第7天二者的播放量近似相同，在第30天视频2的播放量要远大于视频1的播放量。对此情形，我们使用一个多元线性回归模型进行视频未来播放量的预测。该模型可以体现不同增长模式时，视频在早期各天获取的播放量对于长期播放量的不同影响。

**情形2**：而对于未来流行度转换到了一个更高级别的视频（如图中视频3），我们发现其流行度的关键影响因素变成了中期播放量激增。这些激增大多对应着外部影响[68, 69]，与视频在早期的播放量增长趋势关系不大。因此，对此情形，我们在模型中使用视频在观察期获取的总播放量，而非各天播放量，作为预测变量。此外，我们向模型中添加一个截距，以描述导致流行度级别发生转变的播放量激增。



图6-X

综上，我们的基于流行度等级转换的未来播放量预测模型（PLT\_ML）具体定义如下。

We test our method on a real-world dataset and compare the prediction performance with two state-of-the-art baseline models of online video

popularity prediction. The experimental results show that our method leads to significant reductions in the prediction errors, reaching up to 32.25% and 19.82% over the baseline models, respectively. We further analyze the potentials and limita- tions of different detectors in the popularity level prediction. In addition, we investigate the importance and utility of each feature and feature group, and shed light on the key impact factors related to the video popularity. The main contributions of our work are summarized as follows:

We evaluate our model on a real-world dataset and compare the prediction performance with two state- of-the-art online video popularity prediction models. Our approach leads to significant reductions in prediction errors of 32.73% and 11.28% over the baseline models respectively.

We analyze the potentials and limitations of different detectors and model parameters in the prediction. We shed light on the importance of each feature and each feature group used in the burst prediction.

* 1. 本章小结

1. 总结与展望

参考文献

1. Cisco. *White paper: Cisco VNI Forecast and Methodology, 2015-2020*. 2016; Available from: <http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html>.

2. Sodagar, I., *The mpeg-dash standard for multimedia streaming over the internet.* IEEE MultiMedia, 2011. **18**(4): p. 62-67.

3. Saxena, M., U. Sharan, and S. Fahmy. *Analyzing video services in web 2.0: a global perspective*. in *Proceedings of the 18th International Workshop on Network and Operating Systems Support for Digital Audio and Video*. 2008. ACM.

4. Adhikari, V.K., S. Jain, and Z.-L. Zhang. *YouTube traffic dynamics and its interplay with a tier-1 ISP: an ISP perspective*. in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. 2010. ACM.

5. Adhikari, V.K., et al. *Vivisecting youtube: An active measurement study*. in *INFOCOM, 2012 Proceedings IEEE*. 2012. IEEE.

6. Adhikari, V.K., S. Jain, and Z.-L. Zhang. *Where do you" tube"? uncovering youtube server selection strategy*. in *Computer Communications and Networks (ICCCN), 2011 Proceedings of 20th International Conference on*. 2011. IEEE.

7. Torres, R., et al. *Dissecting video server selection strategies in the youtube cdn*. in *Distributed Computing Systems (ICDCS), 2011 31st International Conference on*. 2011. IEEE.

8. Plissonneau, L., E. Biersack, and P. Juluri. *Analyzing the impact of YouTube delivery policies on user experience*. in *Proceedings of the 24th International Teletraffic Congress*. 2012. International Teletraffic Congress.

9. Korczyński, M. and A. Duda. *Classifying service flows in the encrypted skype traffic*. in *Communications (ICC), 2012 IEEE International Conference on*. 2012. IEEE.

10. Chu, W., et al. *Protect sensitive sites from phishing attacks using features extractable from inaccessible phishing URLs*. in *Communications (ICC), 2013 IEEE International Conference on*. 2013. IEEE.

11. Chaudhary, V. and A. Sureka. *Contextual feature based one-class classifier approach for detecting video response spam on youtube*. in *Privacy, Security and Trust (PST), 2013 Eleventh Annual International Conference on*. 2013. IEEE.

12. Cortes, C. and V. Vapnik, *Support-vector networks.* Machine learning, 1995. **20**(3): p. 273-297.

13. Breiman, L., et al., *Classification and regression trees*. 1984: CRC press.

14. Erman, J., et al. *Over the top video: the gorilla in cellular networks*. in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. 2011. ACM.

15. Ramos-Muñoz, J.J., et al., *Characteristics of mobile youtube traffic.* IEEE Wireless Communications, 2014. **21**(1): p. 18-25.

16. Yu, H., et al. *Understanding user behavior in large-scale video-on-demand systems*. in *ACM SIGOPS Operating Systems Review*. 2006. ACM.

17. Gill, P., et al. *Youtube traffic characterization: a view from the edge*. in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. 2007. ACM.

18. Zink, M., et al., *Characteristics of YouTube network traffic at a campus network–measurements, models, and implications.* Computer networks, 2009. **53**(4): p. 501-514.

19. Arvidsson, A., et al. *Analysis of user demand patterns and locality for youtube traffic*. in *Teletraffic Congress (ITC), 2013 25th International*. 2013. IEEE.

20. Casas, P., et al. *YouTube all around: Characterizing YouTube from mobile and fixed-line network vantage points*. in *Networks and Communications (EuCNC), 2014 European Conference on*. 2014. IEEE.

21. Li, Z., et al. *User behavior characterization of a large-scale mobile live streaming system*. in *Proceedings of the 24th International Conference on World Wide Web*. 2015. ACM.

22. Apache. *Apache Hadoop*. Available from: <https://hadoop.apache.org/>.

23. Lorenz, M.O., *Methods of measuring the concentration of wealth.* Publications of the American statistical association, 1905. **9**(70): p. 209-219.

24. Louail, T., et al., *From mobile phone data to the spatial structure of cities.* Scientific Reports, 2014. **4**: p. 5276.

25. Lomax, K., *Business failures: Another example of the analysis of failure data.* Journal of the American Statistical Association, 1954. **49**(268): p. 847-852.

26. Cha, M., et al., *Analyzing the video popularity characteristics of large-scale user generated content systems.* IEEE/ACM Transactions on Networking (TON), 2009. **17**(5): p. 1357-1370.

27. Abhari, A. and M. Soraya, *Workload generation for YouTube.* Multimedia Tools and Applications, 2010. **46**(1): p. 91.

28. Cheng, X., J. Liu, and C. Dale, *Understanding the characteristics of internet short video sharing: A YouTube-based measurement study.* IEEE Transactions on Multimedia, 2013. **15**(5): p. 1184-1194.

29. Abdesslem, F.B. and A. Lindgren. *Large scale characterisation of YouTube requests in a cellular network*. in *A World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2014 IEEE 15th International Symposium on*. 2014. IEEE.

30. Mislove, A., et al. *Measurement and analysis of online social networks*. in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. 2007. ACM.

31. Wilson, C., et al. *User interactions in social networks and their implications*. in *Proceedings of the 4th ACM European conference on Computer systems*. 2009. Acm.

32. Kwak, H., et al. *What is Twitter, a social network or a news media?* in *Proceedings of the 19th international conference on World wide web*. 2010. ACM.

33. Benevenuto, F., et al., *Video interactions in online video social networks.* ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2009. **5**(4): p. 30.

34. Li, Z., et al., *Social connections in user-generated content video systems: Analysis and recommendation.* IEEE Transactions on network and service management, 2013. **10**(1): p. 70-83.

35. Youku. *Youku Open API*. Available from: <http://cloud.youku.com/docs>.

36. Juran, J.M., *Universals in management planning and controlling.* Management Review, 1954. **43**(11): p. 748-761.

37. Adamic, L.A. and B.A. Huberman, *Zipf’s law and the Internet.* Glottometrics, 2002. **3**(1): p. 143-150.

38. Li, W., *Zipf’s law everywhere.* Glottometrics, 2002. **5**: p. 14-21.

39. Powers, D.M. *Applications and explanations of Zipf's law*. in *Proceedings of the joint conferences on new methods in language processing and computational natural language learning*. 1998. Association for Computational Linguistics.

40. Papoulis, A. and S.U. Pillai, *Probability, random variables, and stochastic processes*. 2002: Tata McGraw-Hill Education.

41. Travers, J. and S. Milgram, *The small world problem.* Phychology Today, 1967. **1**: p. 61-67.

42. Travers, J. and S. Milgram, *An experimental study of the small world problem.* Sociometry, 1969: p. 425-443.

43. Luce, R.D. and A.D. Perry, *A method of matrix analysis of group structure.* Psychometrika, 1949. **14**(2): p. 95-116.

44. Wasserman, S. and K. Faust, *Social network analysis: Methods and applications*. Vol. 8. 1994: Cambridge university press.

45. Watts, D.J. and S.H. Strogatz, *Collective dynamics of ‘small-world’networks.* nature, 1998. **393**(6684): p. 440-442.

46. Wu, F. and B.A. Huberman, *Novelty and collective attention.* Proceedings of the National Academy of Sciences, 2007. **104**(45): p. 17599-17601.

47. Szabo, G. and B.A. Huberman, *Predicting the popularity of online content.* Communications of the ACM, 2010. **53**(8): p. 80-88.

48. Ghose, A. and S. Yang, *An empirical analysis of search engine advertising: Sponsored search in electronic markets.* Management Science, 2009. **55**(10): p. 1605-1622.

49. Famaey, J., T. Wauters, and F. De Turck. *On the merits of popularity prediction in multimedia content caching*. in *Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on*. 2011. IEEE.

50. Zhou, Y., et al., *Video popularity dynamics and its implication for replication.* IEEE transactions on multimedia, 2015. **17**(8): p. 1273-1285.

51. Han, B., et al., *Mobile data offloading through opportunistic communications and social participation.* IEEE Transactions on Mobile Computing, 2012. **11**(5): p. 821-834.

52. Crane, R. and D. Sornette, *Robust dynamic classes revealed by measuring the response function of a social system.* Proceedings of the National Academy of Sciences, 2008. **105**(41): p. 15649-15653.

53. Chatzopoulou, G., C. Sheng, and M. Faloutsos. *A first step towards understanding popularity in YouTube*. in *INFOCOM IEEE Conference on Computer Communications Workshops, 2010*. 2010. IEEE.

54. Figueiredo, F., F. Benevenuto, and J.M. Almeida. *The tube over time: characterizing popularity growth of youtube videos*. in *Proceedings of the fourth ACM international conference on Web search and data mining*. 2011. ACM.

55. Cha, M., et al. *I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system*. in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. 2007. ACM.

56. Tan, X., et al. *Characterizing user Popularity Preference in a large-scale online video streaming system*. in *Wireless, Mobile and Multi-Media (ICWMMN 2015), 6th International Conference on*. 2015. IET.

57. Liu, W., et al. *A measurement-based study on application popularity in android and iOS app stores*. in *Proceedings of the 2015 Workshop on Mobile Big Data*. 2015. ACM.

58. Cheng, X., C. Dale, and J. Liu. *Statistics and social network of youtube videos*. in *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*. 2008. IEEE.

59. Figueiredo, F. *On the prediction of popularity of trends and hits for user generated videos*. in *Proceedings of the sixth ACM international conference on Web search and data mining*. 2013. ACM.

60. Figueiredo, F., et al., *On the dynamics of social media popularity: a YouTube case study.* ACM Transactions on Internet Technology (TOIT), 2014. **14**(4): p. 24.

61. Yang, J. and J. Leskovec. *Modeling information diffusion in implicit networks*. in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. 2010. IEEE.

62. Yang, J. and J. Leskovec. *Patterns of temporal variation in online media*. in *Proceedings of the fourth ACM international conference on Web search and data mining*. 2011. ACM.

63. Bao, P., et al. *Modeling and predicting popularity dynamics of microblogs using self-excited hawkes processes*. in *Proceedings of the 24th International Conference on World Wide Web*. 2015. ACM.

64. Jamali, S. and H. Rangwala. *Digging digg: Comment mining, popularity prediction, and social network analysis*. in *Web Information Systems and Mining, 2009. WISM 2009. International Conference on*. 2009. IEEE.

65. Tsagkias, M., W. Weerkamp, and M. De Rijke. *Predicting the volume of comments on online news stories*. in *Proceedings of the 18th ACM conference on Information and knowledge management*. 2009. ACM.

66. Lee, J.G., S. Moon, and K. Salamatian, *Modeling and predicting the popularity of online contents with Cox proportional hazard regression model.* Neurocomputing, 2012. **76**(1): p. 134-145.

67. Cox, D.R., *Regression models and life-tables*, in *Breakthroughs in statistics*. 1992, Springer. p. 527-541.

68. Vallet, D., et al. *Characterizing and predicting viral-and-popular video content*. in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. 2015. ACM.

69. Roy, S.D., et al., *Towards cross-domain learning for social video popularity prediction.* IEEE Transactions on multimedia, 2013. **15**(6): p. 1255-1267.

70. Vasconcelos, M., J.M. Almeida, and M.A. Gonçalves, *Predicting the popularity of micro-reviews: A Foursquare case study.* Information Sciences, 2015. **325**: p. 355-374.

71. Tsagkias, M., W. Weerkamp, and M. De Rijke. *News comments: Exploring, modeling, and online prediction*. in *European Conference on Information Retrieval*. 2010. Springer.

72. Tatar, A., et al. *Ranking news articles based on popularity prediction*. in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. 2012. IEEE Computer Society.

73. Tatar, A., et al., *From popularity prediction to ranking online news.* Social Network Analysis and Mining, 2014. **4**(1): p. 1-12.

74. Chen, H., et al. *Online prediction algorithm of the news' popularity for wireless cellular pushing*. in *Communications in China (ICCC), 2015 IEEE/CIC International Conference on*. 2015. IEEE.

75. Pinto, H., J.M. Almeida, and M.A. Gonçalves. *Using early view patterns to predict the popularity of youtube videos*. in *Proceedings of the sixth ACM international conference on Web search and data mining*. 2013. ACM.

76. Wu, T., et al. *On the use of reservoir computing in popularity prediction*. in *Evolving Internet (INTERNET), 2010 Second International Conference on*. 2010. IEEE.

77. Gürsun, G., M. Crovella, and I. Matta. *Describing and forecasting video access patterns*. in *INFOCOM, 2011 Proceedings IEEE*. 2011. IEEE.

78. Jiang, L., et al. *Viral video style: A closer look at viral videos on youtube*. in *Proceedings of International Conference on Multimedia Retrieval*. 2014. ACM.

79. Youku. *Youku Tudou Partners With Xiaomi to Accelerate Multi-Screen Ecosystem Development*. Available from: <http://ir.youku.com/phoenix.zhtml?c=241246&p=irol-newsArticle_print&ID=1988630>.

80. Borghol, Y., et al., *Characterizing and modelling popularity of user-generated videos.* Performance Evaluation, 2011. **68**(11): p. 1037-1055.

81. Breslau, L., et al. *Web caching and Zipf-like distributions: Evidence and implications*. in *INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*. 1999. IEEE.

82. Ratkiewicz, J., et al., *Characterizing and modeling the dynamics of online popularity.* Physical review letters, 2010. **105**(15): p. 158701.

83. Zhou, R., S. Khemmarat, and L. Gao. *The impact of YouTube recommendation system on video views*. in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. 2010. ACM.

84. Merton, R.K., *The Matthew effect in science.* Science, 1968. **159**(3810): p. 56-63.

85. Johnson, N., S. Kotz, and N. Balakrishnan, *Continuous Univariate Probability Distributions,(Vol. 1)*. 1994, John Wiley & Sons Inc., NY.

86. Figueiredo, F., et al. *Does content determine information popularity in social media?: A case study of youtube videos' content and their popularity*. in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. 2014. ACM.

87. snownlp. *Snownlp simplified Chinese text processing*. Available from: <https://github.com/isnowfy/snownlp>.

88. jieba. *Jieba Chinese text segmentation*. Available from: <https://github.com/fxsjy/jieba>.

89. Altman, N.S., *An introduction to kernel and nearest-neighbor nonparametric regression.* The American Statistician, 1992. **46**(3): p. 175-185.

90. Ho, T.K. *Random decision forests*. in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*. 1995. IEEE.

91. Breiman, L., *Arcing the edge*. 1997, Technical Report 486, Statistics Department, University of California at Berkeley.

92. Russell, S., P. Norvig, and A. Intelligence, *A modern approach.* Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs, 1995. **25**: p. 27.