第一章 绪论

1.1 研究背景

1.1.1 互联网发展现状

进入 21 世纪以来，互联网信息技术经历了日新月异的发展，并向政治、经济、文化等各个领域不断渗透，深刻的改变了人类社会的运作方式和创新模式。随着网络基础设施的建设、移动通信技术的发展以及智能终端设备的普及，用户可以愈发便捷、随时随地的接入互联网，来使用各种各样的互联网业务。

图 1-1(a)给出了从 2007 至 2016 最近十年间，全球互联网的用户数量变化情况[1]。从图中我们可以看出，全球互联网的普及率逐年快速增长。截止到 2016 年 7 月，全球共有 34.25 亿名互联网用户，占人口总数的 46.1%。对比于 2007 年的 13.73 亿名互联网用户，全球互联网用户数在十年间的涨幅高达 149%。图 1-1(b)则显示了目前全球互联网用户中，各大洲的用户比例。其中，亚洲用户占据的比例高达 48.4%，美洲和欧洲用户分别占 21.8%、19%，剩余约 10%的用户则为非洲和大洋洲用户。

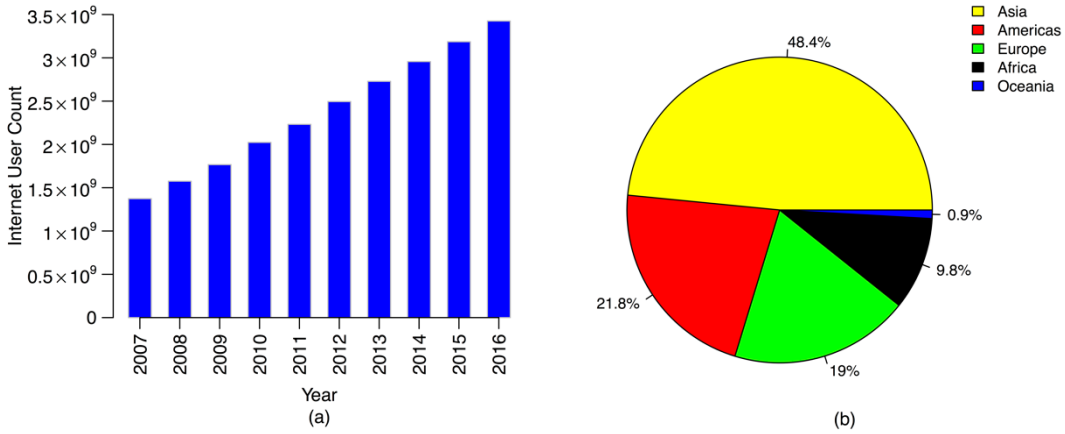

图 1-1 全球互联网用户：(a)近十年来的数量增长；和(b)目前各洲用户比例。

特别的，对于我国而言，虽然互联网技术起步较晚（于 1994 年正式接入互联网），但经过多年来的迅速发展，目前互联网用户规模已达到较大水平。根据中国互联网络信息中心（CNNIC）于 2017 年 1 月发布的第 39 次《中国互联网络发展状况统计报告》[2]，截止到 2016 年 12 月，我国互联网用户数量已达 7.31 亿人，占我国全国人口总数的 53%，互联网普及率已超过半数。表 1-1 给出了近十年来我国互联网用户规模的变化情况。从中我们可以看出，我国的互联网用户数量逐年保持增长，互联网普及率不断提高。不过，随着用户群体构成逐渐成熟，人口红利逐渐消退，互联网用户的增长率正趋于稳定。

表 1-1 中国互联网用户规模逐年概览

年份	互联网用户数（亿）	互联网用户年增长率（%）	互联网普及率（%）
2016	7.31	6.25	53.2
2015	6.88	6.09	50.3
2014	6.49	5.05	47.9
2013	6.18	9.50	45.8
2012	5.64	9.92	42.1
2011	5.13	12.20	38.3
2010	4.57	19.1	34.3
2009	3.84	28.9	28.9
2008	2.98	41.9	22.6
2007	2.10	53.3	16.0

此外，移动网络已成为目前互联网用户规模增长的首要因素。在我国 2016 年内新增的互联网用户中，使用手机上网的用户占据了 80.7%。截止至 2016 年 12 月，我国手机互联网用户的总规模达 6.95 亿。移动互联网的不断发展，有效的驱动了互联网业务类型向多元化、精细化发展，促进了线上线下应用场景的融合，并推动了服务范围的进一步扩展。

1.1.2 网络视频业务

伴随着互联网技术的进步与创新，多种多样的互联网业务大量涌现，如网络搜索、电子商务、网络社交、网络游戏、网络视频等。这些互联网业务已广泛渗透各个领域中，为用户的信息获取、生活服务、即时沟通和娱乐消遣提供了支持与便利。本文中，我们的研究工作主要关注于网络视频业务。网络视频业务指的是目前流行的基于互联网（HTTP 协议）、通过浏览器或专用 APP 向用户提供视频内容的服务。相较于有线电视、卫星电视等传统电视业务，网络视频业务使用互联网而非同轴电缆或卫星进行视频传输，具有更大的覆盖范围与价格优势。而相较于传统的网络流媒体业务，目前网络视频业务的播放形式更加灵活方便、操作更加用户友好、内容更加丰富多样，从而更受用户的欢迎。

网络视频业务是当今互联网中最为重要和最有价值的业务之一。从流量字节数的角度，网络视频业务是当今互联网最大的组成部分。根据思科公司于 2016 年 6 月发布的白皮书《White paper: Cisco VNI Forecast and Methodology, 2015-2020》[3]，在 2015 年全球的互联网流量中，网络视频业务的流量占据了 70%。

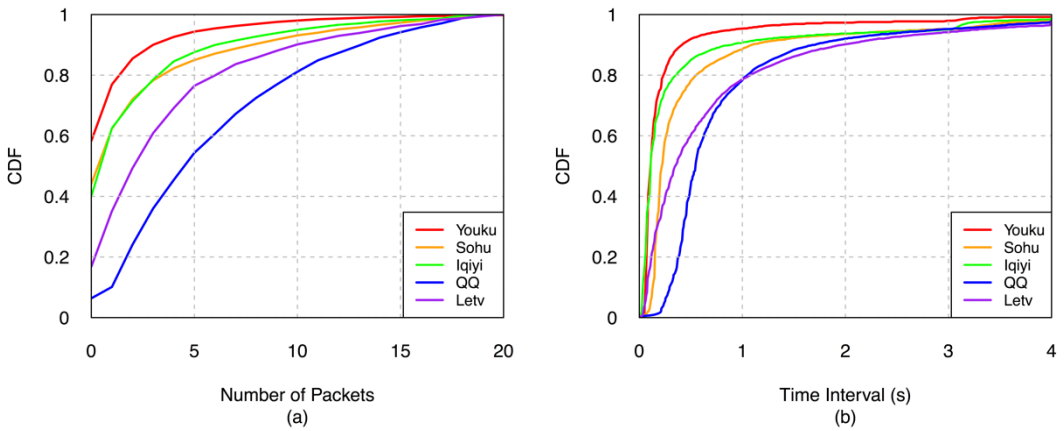

图 3-4 不同网络视频业务提供商的(a)PN(p_{ds}, p_{rs1})、(b)TI(p_{ds}, p_{rs1})的累积分布函数。

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，这两种分发服务器仅支持有限的几种特定的文件类型。图 3-5(a)和(b)分别显示了我们数据集中，各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”的文本类型实体内容。这些文本内容用以承载视频地址的动态信息。其中，”text/json”对应的 JSON 格式和“text/xml”对应的 XML 格式，是目前互联网中最为常见的用于动态信息传输的文件类型。除此之外，有些网络视频业务提供商会使用自定义格式文件，向视频播放器发送视频地址数据。这些自定义格式文件往往对应着 “text/html”或“text/plain”的 HTTP 实体内容类型。而对于资源服务器，其传输的 HTTP 实体内容类型主要是“video/flv”和“video/mp4”，分别对应着 FLV 与 MP4 两种文件格式。这两种格式是目前主流的网络视频封装格式。一般来说，低清晰度视频使用 FLV 格式，而高清晰度视频使用 MP4 格式。

网络视频业务分发服务器仅传输有限且特定类型的 HTTP 实体内容，这一特性有助于将其从其他互联网服务器中区分出来。

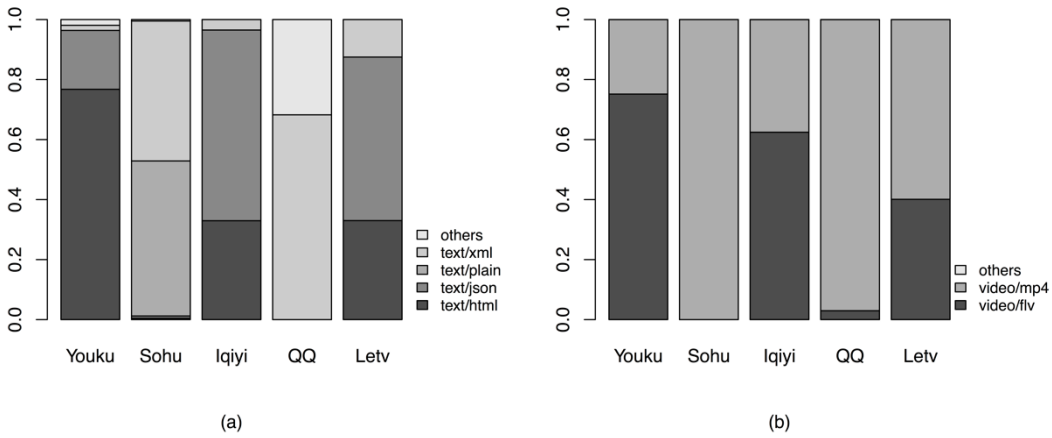

图 3-5 不同网络视频业务服务器传输的 HTTP 实体内容类型分布：

(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，并且新的 URL 由应答报文的“location”头给出。用户侧浏览器收到一个 HTTP 重定向应答时，会自动向“location”指定的 URL 重新发送 HTTP 请求。本文中，我们定义$\mathrm{HR}(p_i, p_j)$来表示 HTTP 重定向行为是否存在于两个 HTTP 交互中：

$$\mathrm{HR}(p_i, p_j) = \begin{cases} 1 & 300 \le \mathrm{SC}(p_i) \le 399, \mathrm{LOC}(p_i) == \mathrm{URL}(p_j) \\ 0 & \text{others} \end{cases} \tag{3-2}$$

其中，p_i和p_j表示两个 HTTP 交互，p_i在p_j之前；$\mathrm{SC}(p_i)$表示 p_i中 HTTP 应答状态码；$\mathrm{LOC}(p_i)$表示p_i中 HTTP 应答“location”头的值；$\mathrm{URL}(p_j)$表示p_j中 HTTP 请求的 URL。

一些网络视频业务提供商通过 HTTP 重定向来完成视频的调度分发。具体来讲，在视频分发阶段，调度服务器以 HTTP 重定向应答来响应用户的视频请求，并在应答报文的“location”头中给出视频在资源服务器上的 URL。表 3-2 给出了优酷视频中调度服务器重定向用户视频请求的 HTTP 交互示例。由于不需要向视频播放器中嵌入任何脚本或代码，HTTP 重定向可能是调度用户下载请求、实现视频分发的最简单方式。

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	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

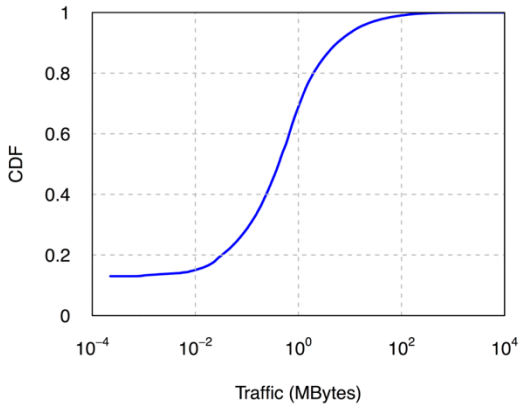

图 4-2 用户的流量字节数累积分布函数。

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，我们基于洛伦兹曲线（Lorenz curve）[74]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n为总用户数，$i = 1, 2, \dots, n$为用户排序后序号，β_i表示用户i消耗的流量字节数，我们有$\beta_1 < \beta_2 < \cdots < \beta_n$。然后对于横坐标$p = i / n, i = 1, 2, \dots, n$的点，确定其纵坐标$L(p)$，其中：

$$L(p) = \frac{\sum_{k=1}^{i} \beta_k}{\sum_{k=1}^{n} \beta_k} \tag{4-1}$$

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$[0, 1]$。而我们的重度用户检测方法的思路，是希望在洛伦兹曲线横轴上找到一个合适的数值P，来划分重度与非重度用户的各自比例。对于P的取值我们考虑两种极端情况：P_{lower}和P_{upper}。其中，P_{lower}是消耗流量字节数等于均值的用户所对应的横坐标值，即：

$$\sum_{k=1}^{P_{\text{lower}}} \beta_k = \frac{1}{n} \sum_{k=1}^{n} \beta_k \tag{4-2}$$

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P_{lower}作为P的取值的下界。而P_{upper}则是洛伦兹曲线在点$p = 1$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$L(p)$的增长速度是越来越大的。当曲线上的点横坐标从P_{lower}开始增长，其对应的切线的斜率也在不断增长，并在点$p = 1$处达到最大值。因此，我们将P_{upper}设定为P的取值的上界。可以看出，用户的

流量分布越不均匀，洛伦兹曲线就会越凹，点$p = 1$处切线的斜率就会越大，进而P_{upper}的数值就会越大。此外，如果用户的流量分布是指数形式的$e^{-(1-p)/a}$，其中a为标度参数，则我们的方法可以得到$1 - P_{\text{upper}} = a$。我们综合考虑上述的两种取值限制，最终定义$P = (P_{\text{lower}} + P_{\text{upper}})/2$。我们的检测方法受文献[75]启发，其中作者 Thomas Louail 等人使用P_{upper}的定义方法来检测地图中的热点区域。

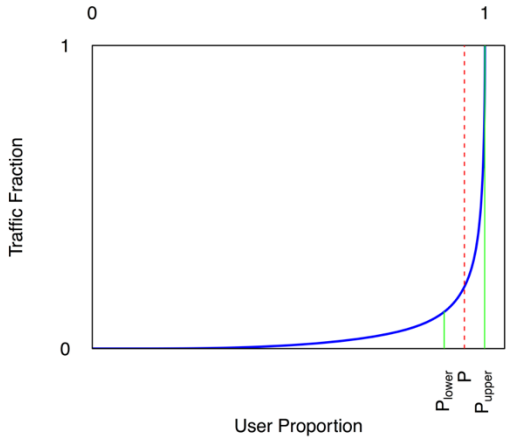

图 4-3 基于洛伦兹曲线的重度用户检测准则示意图。

通过上述方法，我们从数据集 74,928 位移动视频用户中共检测出 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。

4.4.3 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，我们关注于两种时间维度上的用户活跃指标：会话时长与业务总时长。会话时长指的是用户在某一次网络视频业务的使用中消耗的时间长度。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。我们将会话中用户与优酷服务器通信的首末报文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。并且，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 4-4 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数值的区间变化时，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。

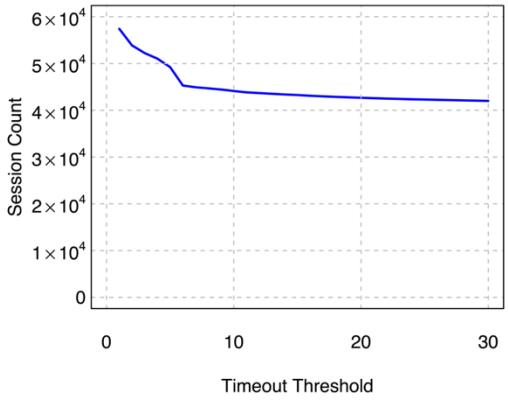

图 4-4 不同超时时限下数据集中的视频会话数。

图 4-5(a)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，因为多数用户观看移动视频的行为发生在他们的零散闲暇时间。此外，移动设备的电池电量与移动网络的流量计费等因素也会对移动用户观看视频的时长产生限制影响。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(b)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，而 96%的用户会话数都小于等于 3。对于重度用户，我们发现其会话时长和会话数则往往较大。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%的重度用户会总计花费超过 40 分钟的时间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。

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[76]来拟合。Pareto Type 2 分布的概率密度函数形式如下：

$$f(x) = \frac{\alpha\lambda^{\alpha}}{(x + \lambda)^{\alpha+1}} \tag{4-3}$$

其中形状参数$\alpha > 0$，尺度参数$\lambda > 0$。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$y = x$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。

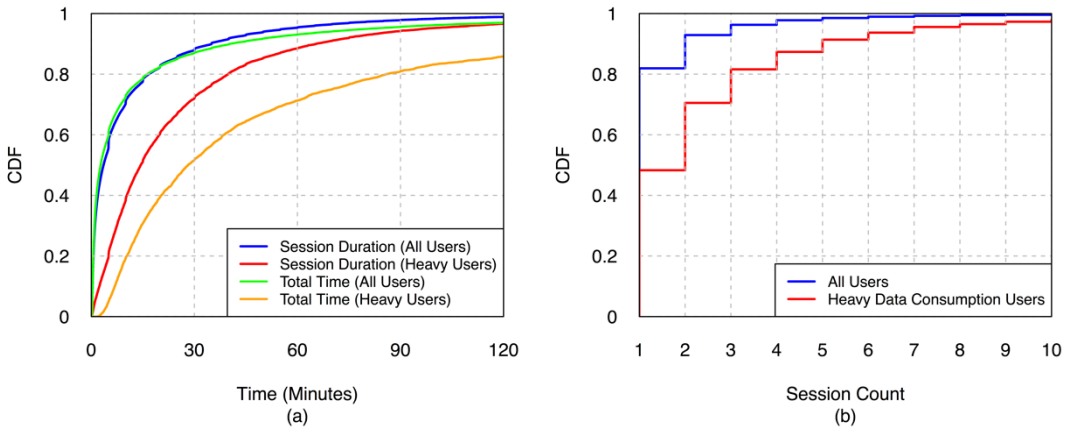

图 4-5 所有用户与重度数据消耗用户的：(a)会话时长与业务总时长累积分布函数；(b)会话数累积分布函数。

表 4-2 Pareto Type 2 分布模型参数

统计指标（用户类型）	α	λ
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

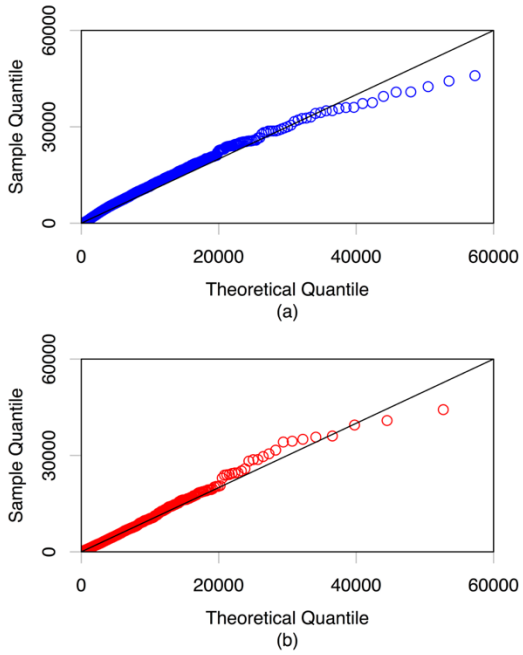

图 4-6 会话时长实际分布与 Pareto Type 2 分布 Q-Q 图：(a)所有用户；(b)重度用户。

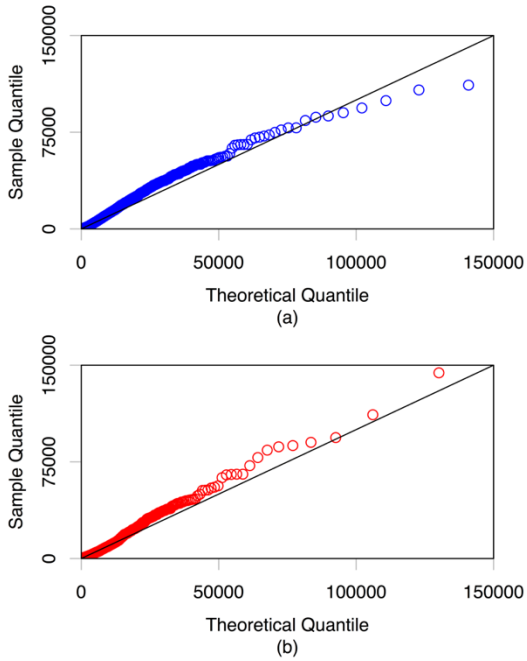

图 4-7 业务总时长实际分布与 Pareto Type 2 分布 Q-Q 图：(a)所有用户；(b)重度用户。

4.5 用户位置移动特性分析

本节中，我们在空间维度上对移动视频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 4-8 所示。

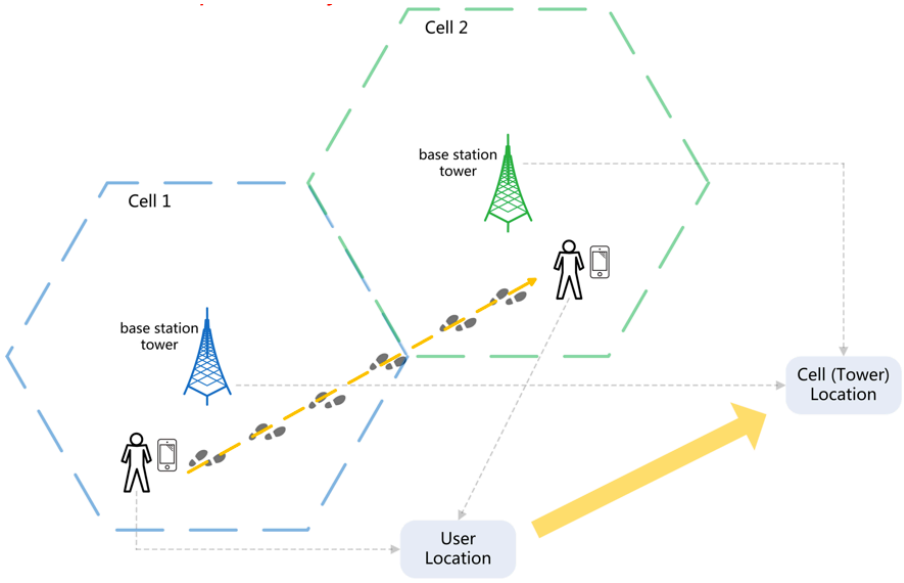

图 4-8 移动网络中用户位置信息提取示意图。

4.5.1 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 4-9 显示了我们数据集中所有视频用户的访问小区数累积分布函数。总体来看，大多数用户的访问小区数较小，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事件就越多，进而占据的无线资源就会越多。

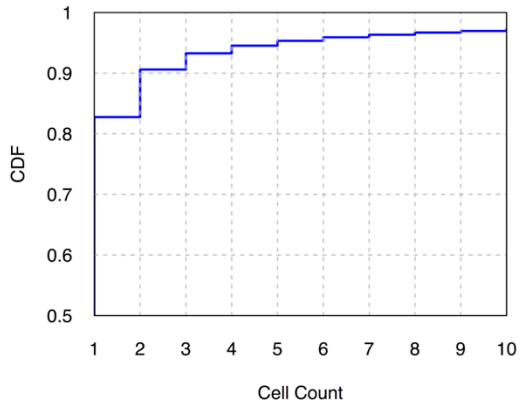

图 4-9 用户的访问小区数累积分布函数。

4.5.2 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，我们分析了用户发起视频请求的地理位置分布。图 4-10 给出了我们数据集中整个城市区域的视频请求热度图。其中，横轴数值为匿名化维度，纵轴数值为匿名化精度。整个地图被划分为100×50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城市的中西部，而两个次要的热点地区分别在城市东北部和城市东南部。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。移动网络视频业务的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

优化此处网络设施部署和视频分发策略，以提升用户体验；而业务提供商可以根据地理位置和周围环境进行广告推送与视频推荐，以获取潜在利润。

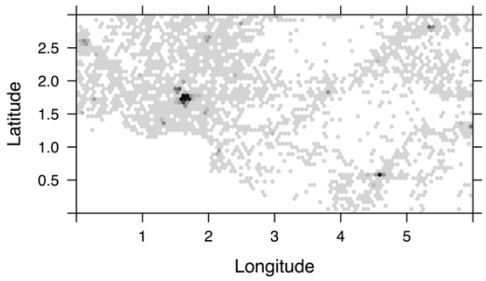

图 4-10 用户的视频请求整体分布热度图。

为了定量的衡量用户的视频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$H(R)$的概念，具体定义如下：

$$H(R) = -\frac{\sum_{r=1}^{N} p_r \log p_r}{\log N} \tag{4-4}$$

$$p_r = \frac{n_r}{\sum_{i=1}^{N} n_i} \tag{4-5}$$

其中，N为总区域数，n_r为区域r中的用户视频请求数。$\log N$为N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。

我们进一步分析了请求时刻与用户请求位置的关系。我们首先比较了在不同时刻，用户请求位置的集中程度，如图 4-11(a)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。此外，我们发现在不同时刻各区域中的用户请求数量差别是很大的。图 4-11(b)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，我们指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。接下来，我们分析了在不同时刻用户的请求位置分布。我们发现其呈现出了基本相同的热点地区特性。图 4-12 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）用户的视频请求热度图。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，热点地区一直保持在地图的中西部、东北部和东南部。

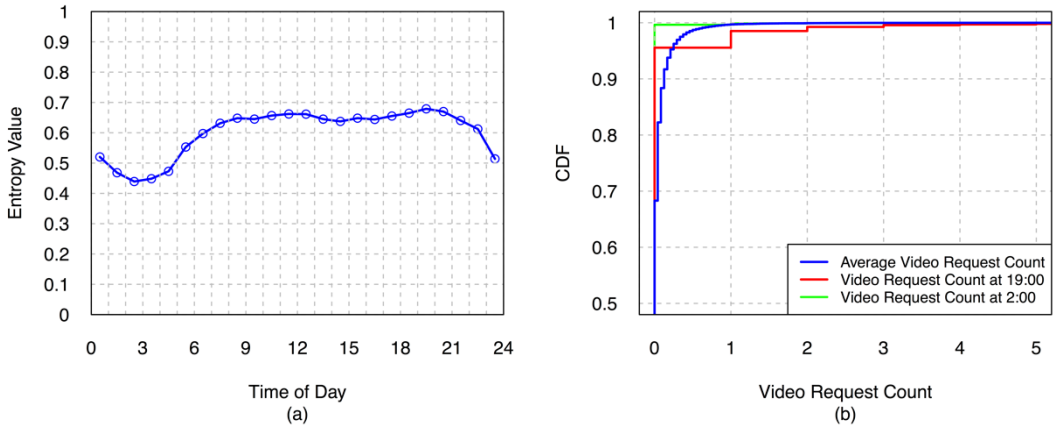

图 4-11 (a)一天内请求位置熵变化情况。(b)典型时刻各区域内视频请求数累积分布函数。

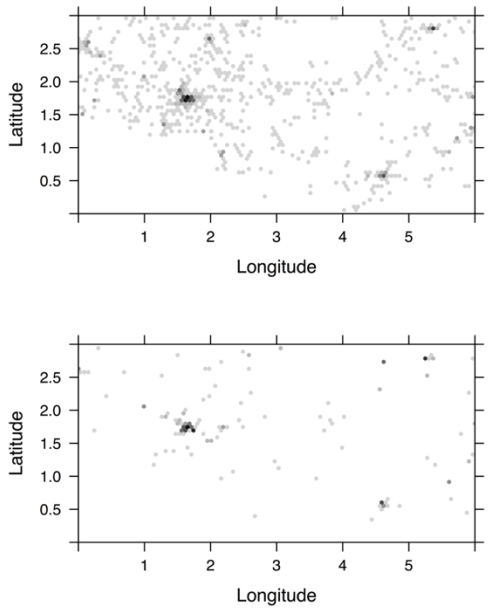

图 4-12 不同时刻用户请求位置热力图。上图时刻：19:00、下图时刻：2:00。

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的历史信息。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n天中使用网络视频业务的用户u，其各天之间的请求位置重合率定义$r_{(u,n)}$如下：

$$r_{(u,n)} = \frac{\mathrm{NUM}(C_{(u,1)} \cap C_{(u,2)} \cap \ldots C_{(u,n)})}{\mathrm{NUM}(C_{(u,1)} \cup C_{(u,2)} \cup \ldots C_{(u,n)})} \tag{4-6}$$

其中，$C_{(u,i)}$是用户在第i天内使用网络视频业务时访问的小区集合；而$\mathrm{NUM}(C)$表示集合C中的元素个数。在图 4-13 中，我们对于数据集中多天使用移动视频业务

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。

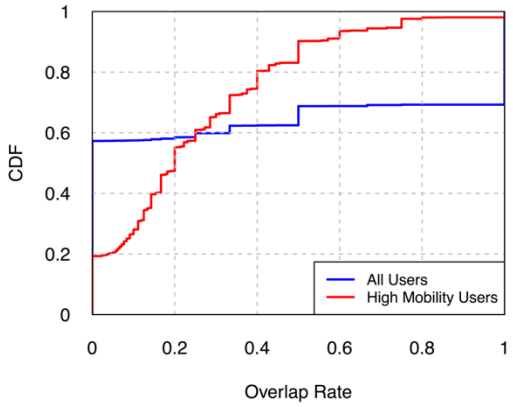

图 4-13 用户的请求位置重合率累积分布函数。

4.5.3 移动模式

在本小节中，我们对用户使用网络视频业务时的移动模式进行分析。移动模式指的是用以描述用户运动情况的空时序列。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”指标对的序列。具体来讲，如果用户u在观看移动视频时改变了n次接入小区，则其移动模式P_u为：

$$P_u = [(c_{(u,0)}, t_{(u,0)}), (c_{(u,1)}, t_{(u,1)}), \dots, (c_{(u,i)}, t_{(u,i)}), \dots, (c_{(u,n)}, t_{(u,n)})] \quad (4\text{-}7)$$

其中，$c_{(u,i)}$为用户第i次改变后对应的接入小区标识，而$t_{(u,i)}$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 4-3 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（c_0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（$c_0 \rightarrow c_1$模式），占据着非静止用户的 45.47%。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$c_0 \rightarrow c_1 \rightarrow c_0$模式和

$c_0 \rightarrow c_1 \rightarrow c_2 \rightarrow c_1$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 4-3 用户移动轨迹模式概览

排名	移动轨迹模式	所有用户中占比	非静止用户中占比
1	c_0	82.76%	-
2	$c_0 \rightarrow c_1$	7.84%	45.47%
3	$c_0 \rightarrow c_1 \rightarrow c_0$	1.16%	6.73%
4	$c_0 \rightarrow c_1 \rightarrow c_2$	1.05%	6.09%
5	$c_0 \rightarrow c_1 \rightarrow c_2 \rightarrow c_1$	0.66%	3.83%
6	$c_0 \rightarrow c_1 \rightarrow c_2 \rightarrow c_3$	0.35%	2.03%

接下来，我们对用户移动模式的停留时间进行分析。图 4-14 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。由此，在本小节的分析中我们规定：若$t_{(u,i)} < 15$分钟，用户u在小区i的停留时间模式为 short,；否则，若$t_{(u,i)} \geq 15$分钟，则用户u在小区i停留时间模式为 long。表 4-4 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。

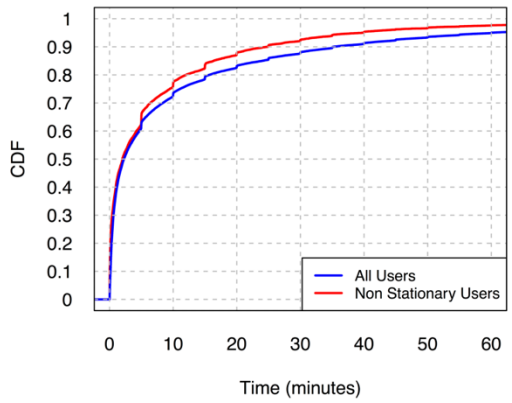

图 4-14 用户在各小区内停留时间累积分布函数

表 4-4 用户停留时间模式概览

排名	停留时间模式	所有用户占比	非静止用户占比
1	short	61.67%	-
2	long	21.09%	-
3	short → short	4.19%	24.30%
4	short → short → short	2.95%	17.13%
5	short → long	2.94%	17.07%
6	short → short → short → short	1.37%	7.95%

4.6 用户业务使用特性分析

4.6.1 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 4-15(a)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[29] ）。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 4-15(b)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$f(k) \propto k^{-\alpha}$。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$\alpha = 2.25$，同时决定系数$R^2 = 0.9689$。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%。

4.6.2 观看时刻

接下来，我们对用户业务使用的整体昼夜节律模式进行了分析。图 4-16 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，在白天的数值较高而在深夜与凌晨的数值较低。具体来讲，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。在白天 9:00 至

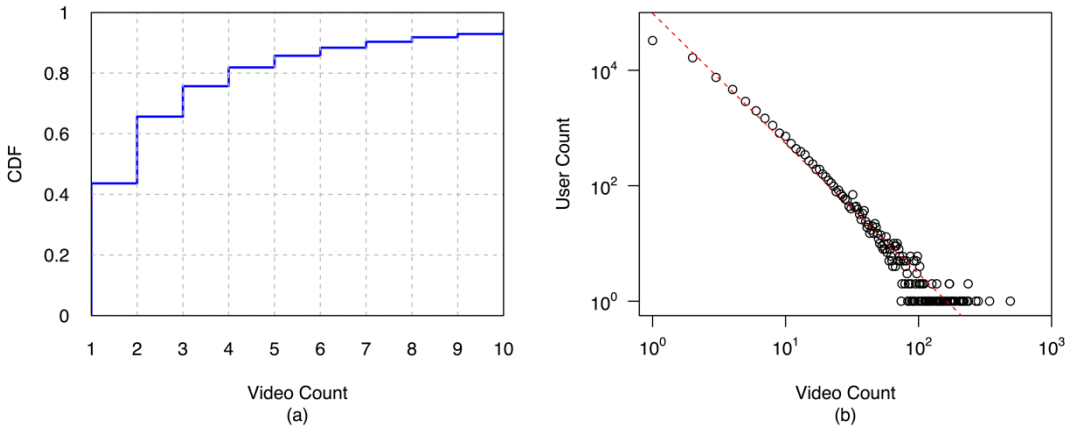

图 4-15 (a)用户的观看视频数累积分布函数。(b)双对数坐标尺度下，观看视频数与用户数的关系。

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 时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，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。在固定网络环境中，用户使用网络视频业务的高峰期仅出现在中午，并在傍晚开始衰减[25, 28]。这一差异表明，用户往往将移动视频的观看当作一项消遣，在他们的非工作、闲暇零散的时间内进行。

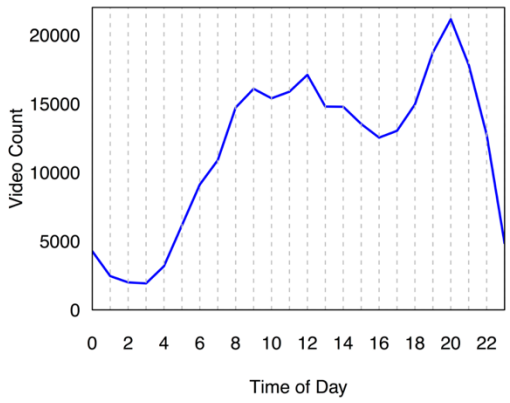

图 4-16 用户观看的视频数在 24 小时内分布。

4.6.3 重复播放行为

用户在使用网络视频业务时，可能会对某些视频特别的感兴趣，进而重复多次的观看这些视频。这种重复播放行为反应了用户的兴趣习惯，并且影响着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u的重放率γ_u定义如下：

$$\gamma_u = \frac{R_u}{N_u} \tag{4-8}$$

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

图 4-17 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。对于全体用户，重放率的分布呈现出了明显的两极分化现象。52%的用户不存在任何重复播放行为（重放率为 0）；而 31%的用户重复播放了所有观看过的视频（重放率为 1）。总体来看，有 35%的用户对应重放率超过 0.5，即他们重复播放了超过一半的观看过的视频。对于重度业务使用用户而言，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。

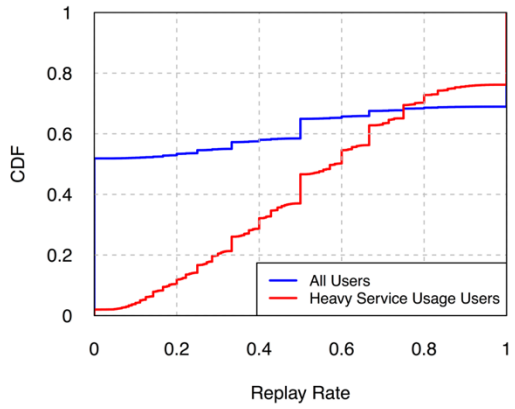

图 4-17 用户的视频重放率累积分布函数。

4.7 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资源的消耗影响，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）之间的用户进行关联分析。在上文中，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。接下来，我们对三种重度用户类型的重叠关系进行分析。图 4-18 给出了我们数据集中三种重度用户集合的文氏图[77]。用户集间最大的重叠部分是 1,888 名同时为重度数据消耗和重度业务使用的用户。这些用户占据了 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，各集合中都存在很大一部分不与其他集合重合。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。

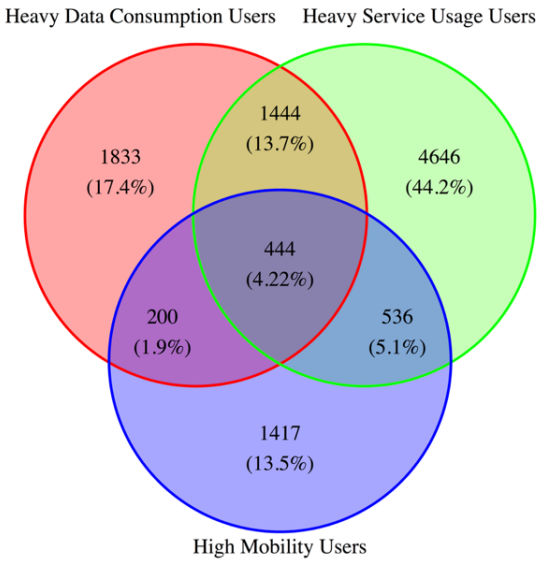

图 4-18 三种类型重度用户集合的文氏图。

我们进一步对不同分析角度的用户行为进行交叉比较，并发现某一分析角度的重度用户，在其他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 4-19(a)显示了不同移动性用户产生的流量字节数累积分布函数。可以看出，高移动性用户产生的流量也往往较大。高移动性、低移动性和静止用户对应的平均流量字节数分别为 36.41MB，11.87MB 和 2.96MB。而在图 4-19(b)中，我们给出了不同数据消耗用户的访问小区数累积分布函数。类似的，我们发现数据消耗中的重度用户比非重度用户趋于访问更多的小区数，进而消耗更多的无线接入资源。对其他分析角度之间交叉比较，我们都得到了相似的结论。为简洁起见，我们略去了具体的结果图。

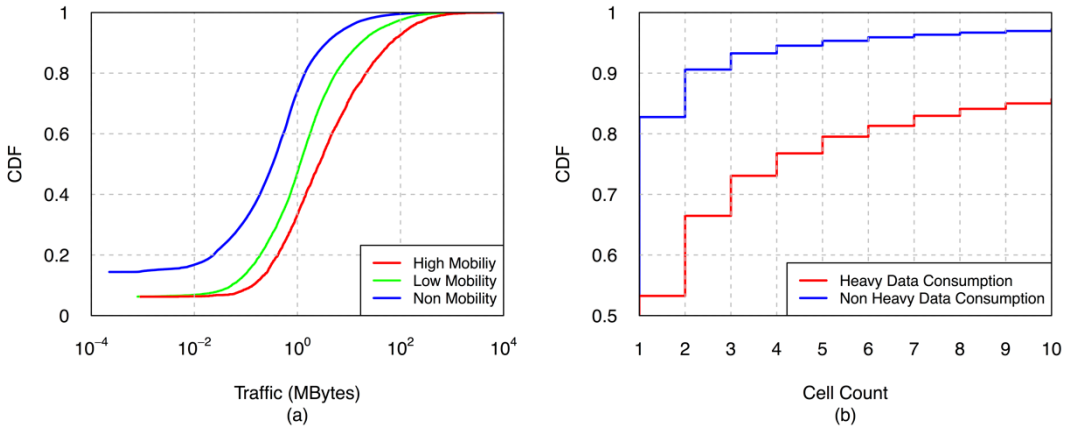

图 4-19 (a)不同移动性用户的数据消耗情况。(b)不同数据消耗用户的移动性情况。

4.8 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

5.4.2 每小时活跃度

图 5-3 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。两种用户的数量都在深夜及凌晨较低而在其他时段较高。而视频数（请求数）则大体上与上传者数（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 时段内，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 时段内也经历了大幅度的增长。在白天 9:00 到 16:00 间，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。

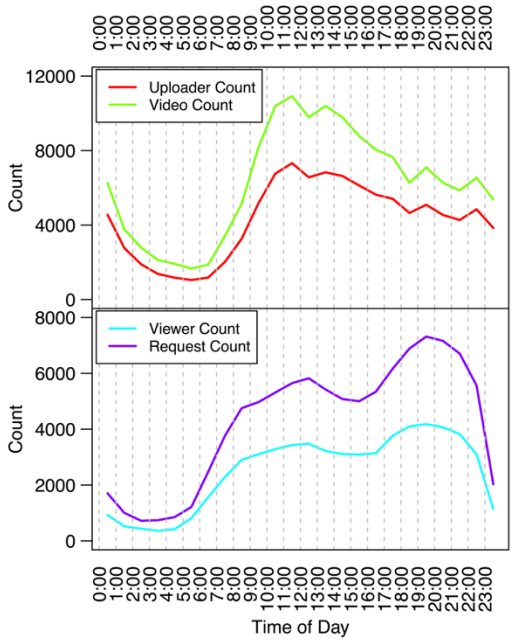

图 5-3 一天中不同小时内的用户活跃度。

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。

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 5-3 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，上传者可以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，上传者应该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。

5.4.3 用户业务使用

本小节中，我们关注于用户的业务使用强度整体分布。图 5-4 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。

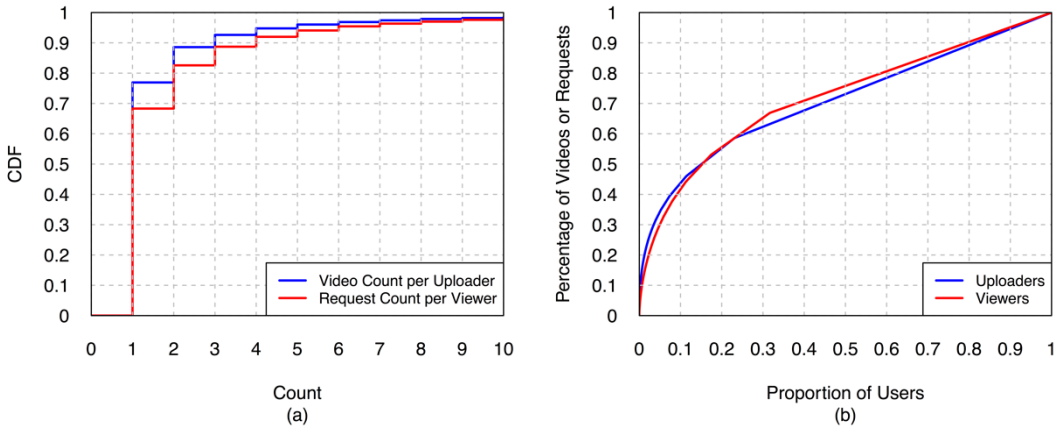

图 5-4 (a)用户的业务使用强度累积分布函数。(b)经排序的用户比例与业务使用强度累积占比的关系。

我们进一步给出了用户比例与业务使用强度累积占比的关系，如图 5-4(b)所示。其中，横轴为按上传视频（视频请求）数量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[80]（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。

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。

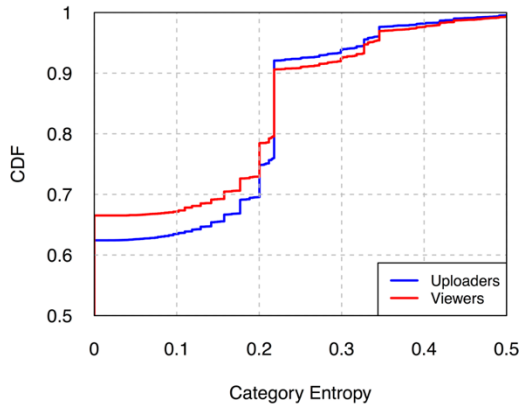

图 5-5 用户的视频种类熵累积分布函数

5.5.2 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 5-6 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 5-1 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，用户的观看视频中存在着更多的长视频。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购买内容，例如电视剧、电影、综艺、动漫等。从表 5-1 中可知，这些种类的视频往往都是长视频。此外，在图 5-6 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 5-1 可知，这些视频主要是电视剧视频和综艺视频。

对不同类型视频的时长特性分析具有着重要的实际应用价值。例如，网络运营商和业务提供商可以根据用户类型、视频类型及视频时长，灵活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。

5.5.3 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，在动态播放量方面的特性进行分析。视频的播放量在一定程度上反映了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

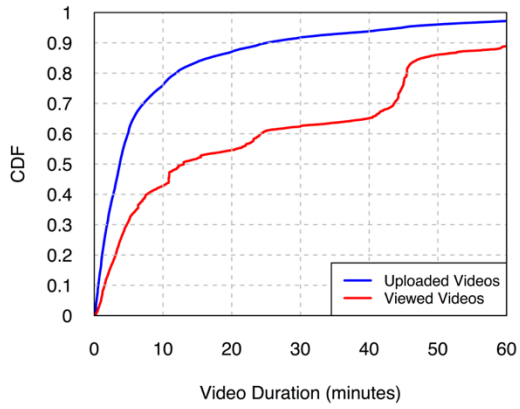

图 5-6 视频时长累积分布函数

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[31, 81, 82]，研究者发现网络媒体内容的流行度整体分布服从奇普夫定律[83]（Zipf's law）。具体来讲，对于排名为k的内容，其被访问频率f_k为：

$$f_k = \frac{1 / k^s}{\sum_{n=1}^{N}(1 / n^s)} \tag{5-3}$$

其中N为内容总数，s为分布模型参数。该函数在双对数坐标下呈现为一条直线：

$$y = \alpha \cdot x + \beta \tag{5-4}$$

为检查我们数据集中用户的上传视频与播放视频的整体流行度分布是否同样符合奇普夫定律，在图 5-7(a)和(b)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。

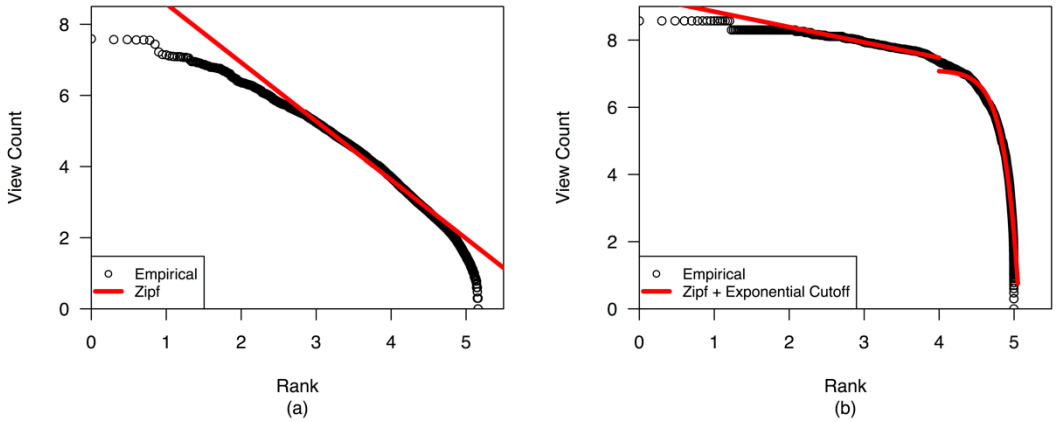

图 5-7 双对数坐标系下视频播放量排名与播放量数值的关系：(a)上传视频；(b)播放视频。

对于被上传的视频，从图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，得到回归参数$\alpha = -1.6510$，$\beta = 10.2300$。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 5-7(b)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。具体模型函数如下。

$$y = \begin{cases} \alpha \cdot x + \beta & x < r \\ a \cdot e^{b \cdot x} + c & x \geq r \end{cases} \tag{5-5}$$

我们对实际数据运行了线性回归来进行验证，得到回归参数$\alpha = -0.4570$，$\beta = 9.3037$，$a = -3.1225$，$b = 5.1560$，$c = 7.1043$。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。

5.6 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构成的用户关系进行分析。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，不同的上传者之间相互较为独立，没有明显关系。因此，我们转而研究了各上传者与其粉丝之间的关系。上传者的粉丝数体现了其发布内容的吸引力与影响力。而在视频播放方面，我们则关注于各播放者对于观看视频的选择。我们基于用户对视频的共同喜好，构建了关系网络并进行分析。

5.6.1 上传者粉丝数

图 5-8 显示了我们数据集中上传者具有的粉丝数累积分布函数。从图中我们可以发现，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%的上传者拥有超过 5,000 名的粉丝。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[84]。Weibull 分布的概率密度函数如下：

$$f(x) = \frac{\alpha}{\lambda}(\frac{x}{\lambda})^{\alpha-1}\exp\left[-(\frac{x}{\lambda})^{\alpha}\right] \tag{5-6}$$

其中α为形状参数，λ为尺度参数。我们对数据集进行回归分析，得到回归参数$\alpha = 0.2272$，$\lambda = 127.6571$。为显示拟合效果，我们在图 5-8 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曲线大体上重合，这表明 Weibull 分布是对上传者粉丝数整体分布的一个很好近似。

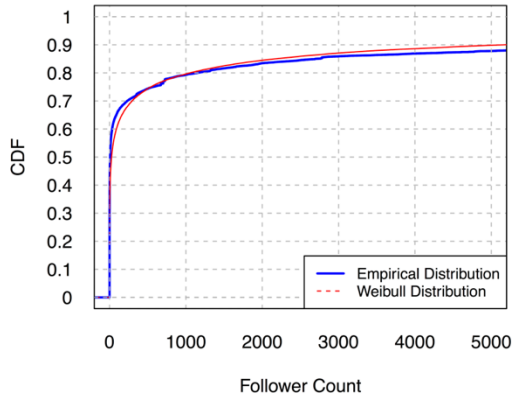

图 5-8 上传者的粉丝数累积分布函数。

接下来，我们进一步分析上传者的粉丝数与视频流行度的关系。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 5-9 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，当上传者的粉丝数较大时，其发布视频的播放量往往也很大。而当上传者的粉丝数较小时，不同上传者的视频播放量之间差异很大。一些上传者的粉丝数量较少，但其发布的视频却仍能够获取较大的播放量。为弄清这一现象的原因，我们进一步分析了对于具有不同粉丝数量的上传者，分别发布什么类型的视频才能获取大播放量。具体来讲，我们将数据集中播放量超过 10,000 的视频划分为两组，分别来自于低粉丝数（少于 100）上传者和高粉丝数（大于 10,000）上传者。我们在表 5-2 中给出了两组视频的视频类型排名前五名。从表中可以看出，这两组视频的类型构成相差很大。对于来自低粉丝数上传者的流行视频，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上传者的流行视频，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。这些视频内容往往是由某些官方发布者周期性的在优酷上进行发布的，能够吸引大量的忠实观众。因此，每当这些发布者发布了新

还对能够影响视频流行度的潜在因子进行了分析。

6.4.1 整体播放量分布

首先，对于数据集中（分别在 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，不同发布日期的视频组之间，视频长期播放量的分位数保持稳定。这表明群体视频的长期播放量整体分布对发布日期是不敏感，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。

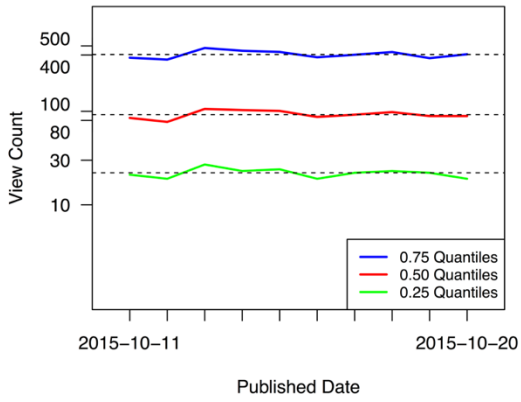

图 6-1 发布于不同日期视频组的长期播放量分位数分布

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是不均匀的，呈现出重尾特性[90, 114, 115]。我们发现该特点也同样适用于优酷视频数据。图 6-2 显示了我们数据集中所有视频在发布日期 30 天后的播放量累积分布函数。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。从图中可以清楚看出，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 1 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，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[116]。在优酷中，某些类型的新发布视频（通常是电视剧、综艺等版权购买内容），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）[117]，极易获取数量巨大的播放量。

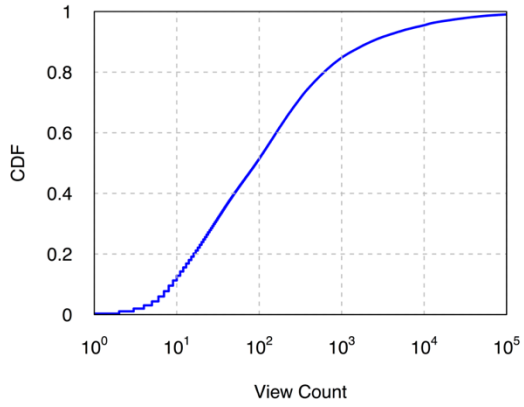

图 6-2 视频长期播放量累积分布函数。

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[97, 113]，有研究者使用使用 Weibull 分布[84]和 Log-Normal 分布[118]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，我们对这两个分布以及 Pareto Type 2 分布[76]进行尝试。三种分布的概率密度函数具体如下：

$$f_{\text{Weibull}}(x) = \frac{\alpha}{\lambda}(\frac{x}{\lambda})^{\alpha-1}\exp\left[-(\frac{x}{\lambda})^{\alpha}\right] \tag{6-1}$$

$$f_{\text{Log-Normal}}(x) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right] \tag{6-2}$$

$$f_{\text{Pareto}}(x) = \frac{\alpha\lambda^{\alpha}}{(x+\lambda)^{\alpha+1}} \tag{6-3}$$

其中α为形状参数，λ为尺度参数，μ和σ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 6-3 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$y = x$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。

6.4.2 流行度级别

根据视频的长期播放量，我们进一步划分了不同的流行度级别，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

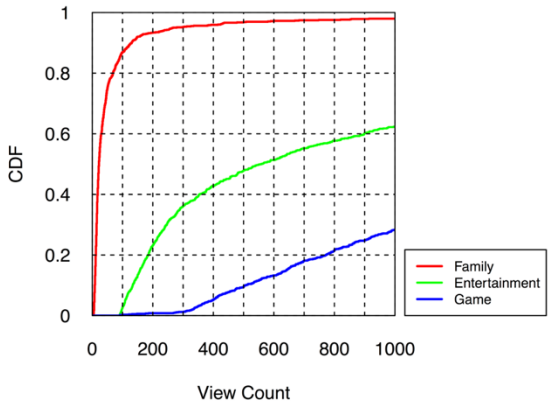

图 6-4 三种典型类型视频的长期播放量累积分布函数。

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 50,833，最大值为 123,943。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。

6.5 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，我们通过考察各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。

6.5.1 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，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V_{active}，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V_{active}，则该天为该视频的一个非活跃天。可以看出，对于活跃天概念，一个关键的问题是如何为每个视频单体定义一个合适的V_{active}。对此，我们首先定义一个绝对数量参数η，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η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 6-5 所示。从图中可以看出，当η取值于较小的数值区间时（约在 500 以下），随着η的增加，总活跃天数的减少十分剧烈。而当η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$\eta = 500$。

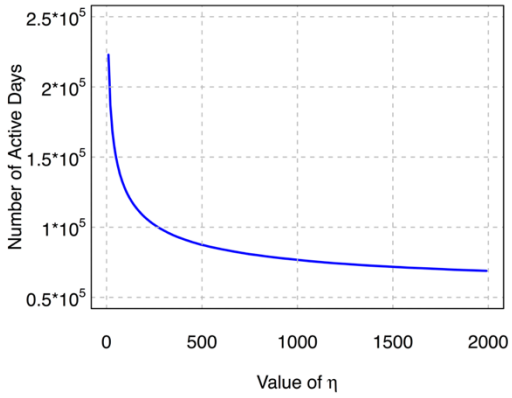

图 6-5 不同η数值对应的总活跃天数。

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$\eta = 500$。这表明仅使用绝对数量参数η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δ和α，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δ次；2）超过其平均日播放量α倍，我们则也可以将该天作为该视频的活跃天。图 6-6 显示了不同α下，δ与总活跃天数之间的关系。从图中可以看出，不同α对应的曲线，形状是非常相似的。当δ取值于较小的数值范围时，随着δ的增加，总活跃天数的减少十分剧烈。而当δ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$\delta = 40$。而对于α的选择，我们发现当α数值较小时，图中两条相邻曲线的间距较大。而当α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$\alpha = 1.6$，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V_{active}定义为：

$$V_{\text{active}} = \min\left(\eta, \max\left(\delta, \alpha \frac{N_v(n)}{n}\right)\right) \tag{6-4}$$

其中，n为观察期的总天数；$N_v(n)$为视频v在第n天时获取的总播放量；$\eta = 500$，$\delta = 40$，$\alpha = 1.6$。

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 17,076 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 6-7 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%

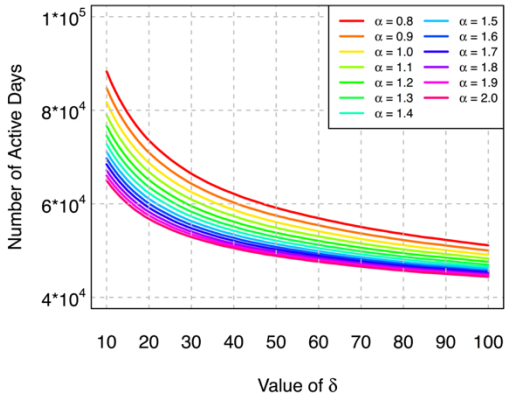

图 6-6 不同δ与α数值对应的总活跃天数。

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。

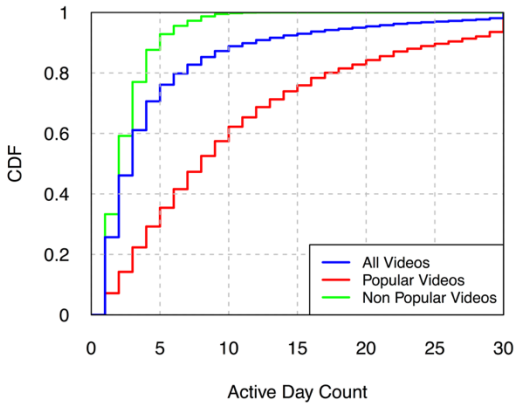

图 6-7 视频活跃天数累积分布函数。

6.5.2 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 6-8 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。

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

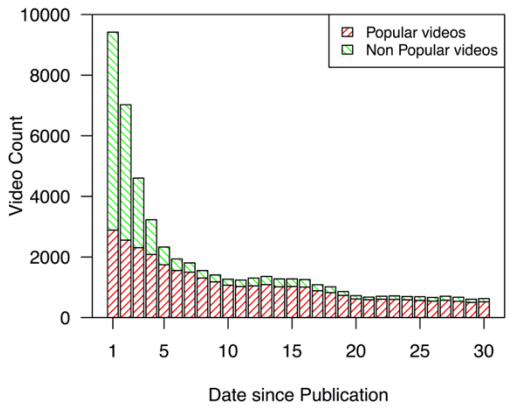

图 6-8 活跃天出现的不同位置对应的视频数直方图。

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k天起，连续超过T天未能出现一个活跃天，则我们认为其活跃期在第k天结束，即该视频的活跃期长度为$k-1$天。在我们的研究中，我们令$T=5$。图 6-9 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 6-9 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。

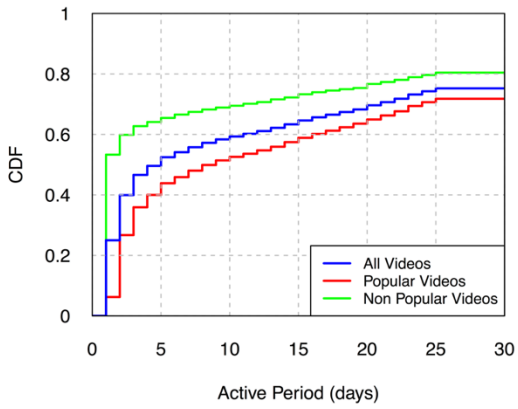

图 6-9 视频活跃期长度累积分布函数。

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发现：对新上传视频进行广告投

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[105]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 6-14 显示了使用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。

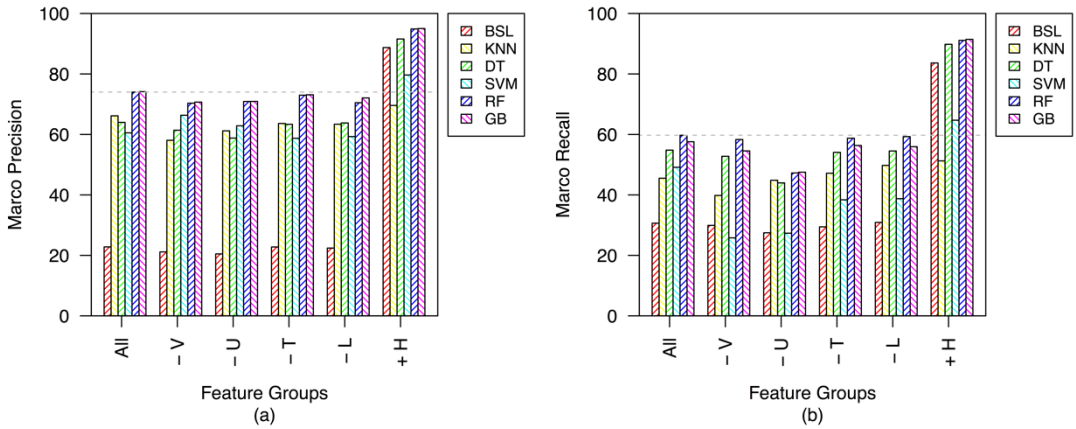

图 6-14 使用不同特征组的各分类模型性能：(a)宏平均精度；(b)宏平均召回率。

接下来，我们关注于各特征组在预测中的有效性。我们进行多次的分类实验，

轮流移除数据集中的一组特征。图 6-14 中也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 分类器时各节点的 Gini 系数平均下降，来对特征进重要性行排序。表 6-5 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 6-5 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	517.83
2	上传者发布的视频数	336.02
3	上传者粉丝数	240.28
4	视频种类	215.79
5	视频时长	206.43
6	上传者获取的总收藏量	198.75
7	上传者注册时间	192.44
8	视频标题中文字符数	163.01
9	视频标题情感度	151.79
10	视频上传时间	150.43

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[45]被用来预测 YouTube 视频流行度。

图 6-14 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。此外，我们发现 KNN 分类器的性能提升较小，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，我们分析了初始观察期的长度对预测性能的影响，并探寻了平衡及时性与准确性的合适观察期长度。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 6-15 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 6-14，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，在我们的研究中，我们选择 7 天作为初始观察期的长度。

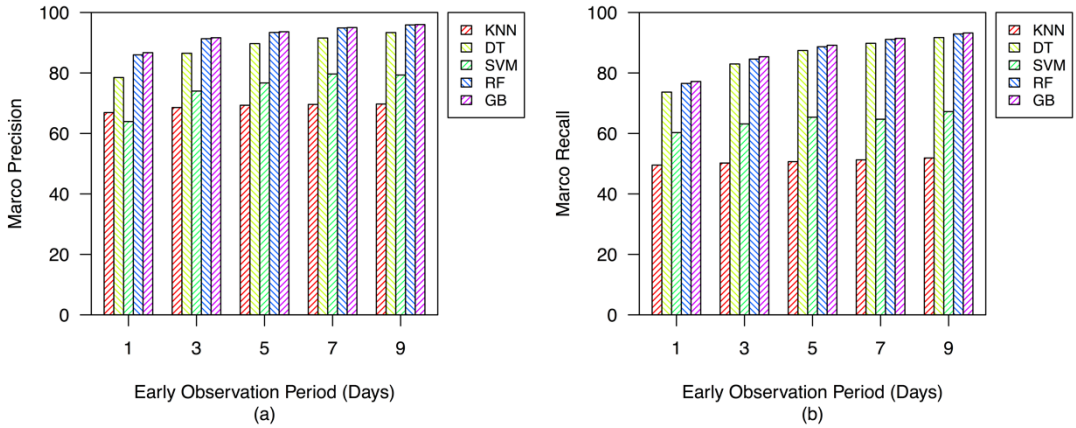

图 6-15 使用不同初始观察期长度的各分类模型性能：(a)宏平均精度；(b)宏平均召回率。

6.7 基于播放量增长模式的未来流行度数值预测

6.7.1 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v，经过一段长度为k的初始观察期后，在其发布后第k天时，我们希望能够对其未来第r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。

6.7.2 早期-长期播放量关系

在之前工作中[44] [45]，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关注的内容，很可能成为未来的流行内容。反之，较小的早期流行度则往往对应不受欢迎的内容。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 6-16 给出了数据集中视频在第 7 天的播放量$N_v(7)$（作为早期流行度）与其第 30 天的播放量$N_v(30)$（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$N_v(30) = \alpha N_v(7) + \beta$回归，我们进一步得到了参数$\alpha = 1.2849$和$\beta = 52.4053$。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。

根据前文 6.5.4 小节中的分析内容，我们推测网络视频早期-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 6-17 中我们给出了数据集中三个典型视频，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，三个视频的播放量几乎相同：$N_{\text{video1}}(7) = 131$，$N_{\text{video2}}(7) = 131$，$N_{\text{video3}}(7) = 133$。然而，在一个月后的第 30 天，各视频的播放量差距很大：$N_{\text{video1}}(30) = 428$，$N_{\text{video2}}(30) = 135$，$N_{\text{video3}}(30) = 291$。其中，最大播放量是最小播放量的 3 倍以上。

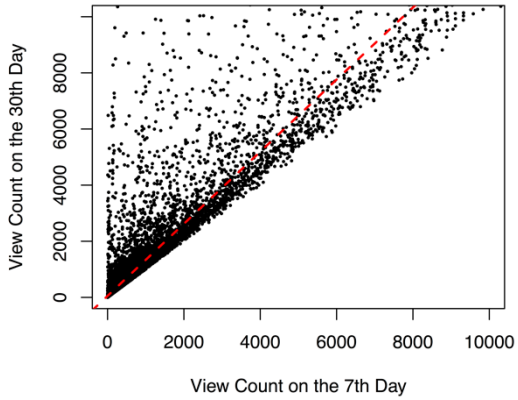

图 6-16 视频早期流行度与长期流行度的关系示意图。

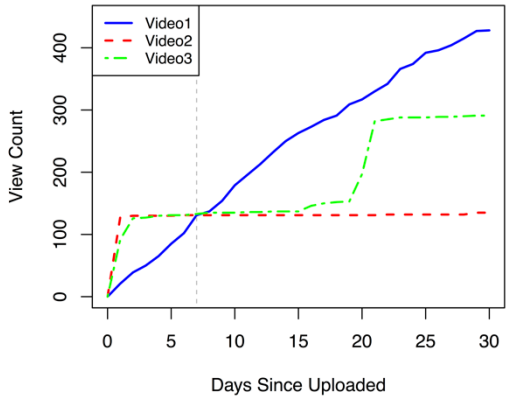

图 6-17 不同播放量增长模式对于早期-长期播放量关系的影响。

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。

6.7.3 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 m及 others，共$m+1$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v在未来是否会经历播放量激增进行了二值预测$b_v=1$或$b_v=0$。对于$b_v=1$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v在发布后第r天的播放量，我们的预测结果为$\widehat{N_v}(r)$：

$$\widehat{N_v}(r)=\sum_{i=1}^{k}\omega_{(p,i)}I_v(i)+\beta b_v N_v(k) \tag{6-12}$$

其中，k为初始观察期的长度；$I_v(i)$为视频v在第i天的播放量（增量）；p为该视频的早期播放量增长模式；$\omega_{(p,i)}$为取决于p与i的回归系数；$N_v(k)$为视频v在第k天的（累积）播放量；β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$\boldsymbol{F}_v$为：

$$\boldsymbol{F}_v=(I_v(1),I_v(2),\dots,I_v(k),N_v(k)) \tag{6-13}$$

而对于早期播放量增长模式p，我们定义模型的参数向量$\boldsymbol{P}_p$为：

$$\boldsymbol{P}_p=(\omega_{(p,1)},\omega_{(p,2)},\dots,\omega_{(p,k)},\beta b_v) \tag{6-14}$$

则模型可进一步表示为：

$$\widehat{N_v}(r)=\boldsymbol{P}_p\boldsymbol{F}_v \tag{6-15}$$

给定一组训练集T，模型参数的最优值可通过最小化T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V，预测结果的 MRSE 定义为：

$$\text{MRSE}=\frac{1}{|V|}\sum_{v\in V}\left(\frac{N_v(r)-\widehat{N_v}(r)}{N_v(r)}\right)^2 \tag{6-16}$$

由此，模型参数可获取如下：

$$\arg\min_{\boldsymbol{P}_p}\frac{1}{|V|}\sum_{v\in V}\left(\frac{N_v(r)-\boldsymbol{P}_p\boldsymbol{F}_v}{N_v(r)}\right)^2 \tag{6-17}$$

令$W_v=\left(\frac{1}{N_v(r)}\right)^2$，则该优化问题可被表示为：

$$\arg\min_{\boldsymbol{P}_p}\frac{1}{|V|}\sum_{v\in V}W_v(N_v(r)-\boldsymbol{P}_p\boldsymbol{F}_v)^2 \tag{6-18}$$

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。

表 6-7 VCGP_ML 模型预测性能概览

视频组类型	MRSE（%）		
	Log-Linear	**Multi-Linear**	**VCGP_ML**
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

6.8 基于流行度级别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，仅达到 77.76%。对视频中期播放量激增精准预测的困难，会成为进一步提升 VCGP_ML 模型性能的瓶颈。然而，在另一问题上，我们对视频未来流行度级别的预测却是很成功的，如上文 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。由此，在本节中，我们利用对视频未来流行度级别的预测结果，提出了一个基于流行度级别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。

6.8.1 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，我们可以直接利用 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流行度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级别保持不变；和 2）未来较早期的流行度级别发生改变。

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 6-19 所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 6-19 中的视频 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[43, 104]，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。

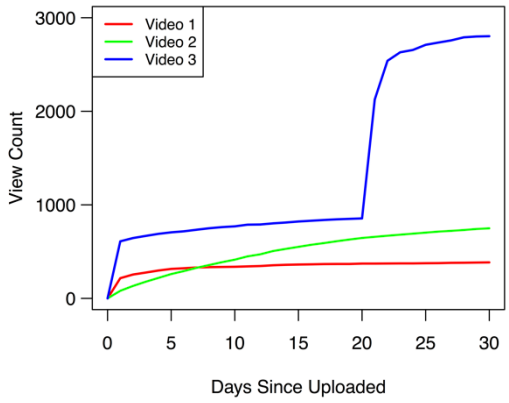

图 6-19 不同流行度级别转换对于早期-长期播放量关系的影响。

综上，我们的 PLT_ML 预测模型具体定义如下。令二元组$t = (L_v(k), \widehat{L_v}(r))$代表视频$v$的流行度级别转换，其中$L_v(k)$为$v$在初始观察期后预测时刻的流行度级别，$\widehat{L_v}(r)$为$v$在未来第$r$天的预计流行度级别。注意，由于$\widehat{L_v}(r)$是由累计播放量衡量，所以一定有$\widehat{L_v}(r) \geq L_v(k)$。对于流行度级别转换$t$的视频$v$，我们预测其在未来第$r$天的播放量为$\widehat{N_v}(r)$：

$$N_v = \begin{cases} \sum_{i=1}^{k} \omega_{(t,i)} I_v(i) & \widehat{L_v}(r) = L_v(k) \\ \alpha_t N_v(k) + \beta_t & \widehat{L_v}(r) > L_v(k) \end{cases} \tag{6-21}$$

其中，$\omega_{(t,i)}$、α_t、β_t为取决于t的模型参数。这些模型参数的最优取值，可通过最小化训练集上的 MRSE 值来获取。