****

**本科生毕业论文（设计）**

**Undergraduate Graduation Thesis（Design）**

题目Title：视频网站观看次数的特征分析及流量控制系

统设计

院 系

School (Department)：数据科学与计算机学院

专 业

Major： 网络工程

学生姓名

Student Name： 卢宇浩

学 号

Student No.： 14348089

指导教师(职称)

Supervisor（Title）：

时间： 2018 年 4 月 8 日

Date: April 8th, 2018

**说 明**

1. 毕业论文（设计）的写作格式要求请参照《中山大学本科生毕业论文的有关规定》和《中山大学本科生毕业论文（设计）写作与印制规范》。

2. 除完成毕业论文（设计）外，还须填写三份表格：

（1）表一 毕业论文（设计）开题报告；

（2）表二 毕业论文（设计）过程检查情况记录表；

（3）表三 毕业论文（设计）答辩情况登记表。

3. 上述表格均可从教务部主页的“下载中心”处下载，如表格篇幅不够，可另附纸。每份毕业论文（设计）定稿装订时应随同附上这三份表格。

4. 封三是毕业论文（设计）成绩评定的主要依据，请认真填写。

**Instruction**

1. Please refer to ‘*The Guidelines to Undergraduate Graduation Thesis (Design) at Sun Yat-sen University*’and ‘*The Writing and Printing Format of Undergraduate Graduation Thesis(Design) at Sun Yat-sen University* ’ for anything about the thesis format.

2. Three forms should be filled up before the submission of the thesis (design)：

（1）Form 1: Research Proposal of Graduation Thesis.

（2）Form 2: Process Check-up Form.

（3）Form 3: Thesis Defense Performance Form.

3. All the above forms could be downloaded on the website of the Office of Education Administration. If there is not enough space in the form, please add extra sheets. Each thesis (design) should be submitted together with the three forms.

4. The form on the inside back cover is the grading sheet. Please fill it up before submission.

毕业论文（设计）成绩评定记录

Grading Sheet of the Graduation Thesis (Design)

|  |
| --- |
| 指导教师评语  Comments of Supervisor：  成绩评定  Grade：    指导教师签名  Supervisor Signature ： Date: |
| 答辩小组意见  Comments of the Defense Committee：  成绩评定  Grade：    签名： Date:  Signatures of Committee Members |
| 院系负责人意见  Comments of the Academic Chief of School：  成绩评定  Grade：  签名 院系盖章  Signature： Stamp: Date: |

**表一：毕业论文（设计）开题报告**

**Form 1: Research Proposal of Graduation Thesis (Design)**

|  |
| --- |
| 论文（设计）题目  Thesis (Design) Title： |
| （简述选题的目的、思路、方法、相关支持条件及进度安排等）  （Please briefly state the research objective, research methodology, research procedure and research schedule in this part.）  Student Signature： Date: |
| 指导教师意见  Comments from Supervisor：    1.同意开题 2.修改后开题 3.重新开题  1.Approved( ) 2. Approved after Revision ( ) 3. Disapproved( )    Supervisor Signature： Date: |

**表二：毕业论文（设计）过程检查情况记录表**

**Form 2：Process Check-up Form**

|  |  |
| --- | --- |
| **指导教师分阶段检查论文的进展情况（要求过程检查记录不少于3次）**  **The supervisor should check up the working process for the thesis（design）and fill up the following check-up log. At least three times of the check-up should be done and kept on the log.**  **第1次检查（First Check-up）:**  学生总结  Student Self-summary：  指导教师意见  Comments of Supervisor：  **第2次检查（Second Check-up）：**  学生总结  Student Self-summary：  指导教师意见  Comments of Supervisor：  **第3次检查（Third Check-up）：**  学生总结  Student Self-summary：  指导教师意见  Comments of Supervisor：  **第4次检查**  **Fourth Check-up**  学生总结  Student Self-summary：  指导教师意见（Comments of Supervisor）：  学生签名（Student Signature）： 日期（Date）:  指导教师签名（Supervisor Signature）： 日期（Date）: | |
| **总体完成情况**  **（Overall Assessment）** | 指导教师意见Comments of Supervisor：  1、按计划完成，完成情况优（Excellent）： （ ）  2、按计划完成，完成情况良（Good）： （ ）  3、基本按计划完成，完成情况合格（Fair）：（ ）  4、完成情况不合格（Poor）： （ ）  指导教师签名（Supervisor Signature）：  日期（Date）: |

**表三：毕业论文（设计）答辩情况登记表**

**Form 3: Thesis Defense Performance Form**

|  |  |  |  |
| --- | --- | --- | --- |
| **答辩人 Student Name** |  | **专 业**  **Major** |  |
| **论文（设计）题目**  **Thesis（Design） Title** |  | | |
| **答辩小组成员**  **Committee Members** |  | | |
| 答辩记录  Records of Defense Performance：    记录人签名（Clerk Signature）： 日期（Date）: | | | |

**学术诚信声明**

本人所呈交的毕业论文，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料均真实可靠。除文中已经注明引用的内容外，本论文不包含任何其他人或集体已经发表或撰写过的作品或成果。对本论文的研究作出重要贡献的个人和集体，均已在文中以明确的方式标明。本毕业论文的知识产权归属于培养单位。本人完全意识到本声明的法律结果由本人承担。

本人签名： 日期：

**Statement of Academic Integrity**

I hereby acknowledge that the thesis submitted is a product of my own independent research under the supervision of my supervisor, and that all the data, statistics, pictures and materials are reliable and trustworthy, and that all the previous research and sources are appropriately marked in the thesis, and that the intellectual property of the thesis belongs to the school. I am fully aware of the legal effect of this statement.

Student Signature： Date：

**摘 要**

**(\* 中文摘要不少于300字。语言力求精练、准确，以300—500字为宜。**

在摘要中简要描述你所解决问题的背景、问题的描述、难点与挑战性、研究方法、所提出的解决问题的具体方法、实验结果和最终结论。主要突出自己的工作，最后介绍论文的学术意义，与相关方法比较是否在某些方面有所提高。关键是让读者能从摘要中看出你在论文阶段完成的主要工作。

摘要应具有独立性和自含性，即不阅读论文的全文，就能获得必要的信息，摘要中有数据、有结论，是一篇完整的短文，可以独立使用，可以引用，可以用于推广。摘要的内容应包含与毕业论文同等量的主要信息，供读者确定有无必要阅读全文，也可供文摘等二次采用，摘要一般应说明研究工作的目的、实验方法、结果和最终结论等，而重点是结论和结果。

除非确实需要，否则在摘要中一般不要用图表、非公知公用的符号和术语。

摘要使用第三人称，被动语句；避免大量背景、必要性、意义的描述，可简要提及；不要过分使用形容词；\*)

**关键词：Youtube；用户生成内容；内容分发网络**

ABSTRACT

(\* 英文摘要以250—400个实词为宜，严格使用英文标点符号。中、英文摘要意思要基本相同。在撰写完英文摘要后，请执行一次自动拼写检查，以减少英文拼写错误的可能性。\*)

**Keywords: Youtube; UGC; CDN**

目 录

第一章 引言 1

1.1 研究的背景和意义 1

1.2 研究的问题 1

1.3 本文的工作 2

1.4 本文的结构 2

第二章 相关工作 3

2.1 UGC系统分析 3

2.2 视频点播中的用户行为分析 4

第三章 数据分析方法 5

3.1 分析内容 5

3.2 元数据解析 5

3.3 数据结构 5

3.4 特征提取 5

3.5 绘制图形 6

第四章 实验结果与分析 7

4.1 对YouTube网站视频的观看量分析 7

4.1.1 视频观看量的分布情况 7

4.1.2 视频时长的分布情况 11

4.1.3 视频观看量与视频长度的综合分析 13

4.1.4 视频观看量与视频评价数的关系 17

4.2 分析总结 20

第五章 总结与展望 22

5.1 本文的总结 22

5.2 未来的展望 22

参考文献 23

致 谢 24

附 录 25

# 引言

## 研究的背景和意义

目前，互联网中视频文件的传输占住着网络信息流的很大一部分，而用户生成内容(User Generated Content, UGC)系统中的视频占据着互联网中视频文件的主流。对UGC系统视频文件的观看次数特征进行分析，有助于互联网服务提供商更好地配置文件缓存，从而减少网络的拥塞状况，提高用户体验。

YouTube作为全球最大的UGC系统，具有代表性，它每时每刻都有大量的流媒体输出，产生大量的负载，而UGC系统对自身的稳定性要求高，当流媒体输出量达到一定程度时，单个或者少量的服务器远远无法达到负载的要求，因此内容分发的策略是十分必要的，而内容分发网络(Content Distribution/Delivery Networks, CDN)和对等网络(Peer-to-peer Networking)等架构已经被广泛应用于如今的流媒体系统乃至整个互联网。但是UGC系统拥有与传统的视频点播(Video-on-Demand, VoD)系统所不同的特性，由于大部分是个人生产的视频，数量和质量参差不齐，用户的随机性相比于传统的Vod系统(例如我国的腾讯视频、优酷土豆、爱奇艺等)更加显著。如何进行高效而且可靠的内容分发，需要从UGC系统的特性入手，通过分析它的特点来对CDN策略进行有效的指导。

充分分析和理解大规模VOD系统的用户行为，并以此为基础，为不同的用户群体设计相应的系统架构，可以有效地利用网络带宽和改进，从而提高系统的整体性能和用户体验。可以看出，了解和掌握大规模的视频点播系统的用户行为特征，将会有利于CDN服务提供商、VoD服务提供商和用户[1]。

## 研究的问题

对YouTube网站存储的文件进行分析，主要包括：1.分析视频文件观看次数的概率分布特征；2.分析视频文件时长的概率分布特征；3.分析视频的观看次数与视频文件时长之间的关系；4.分析视频文件观看次数与用户评价总次数之间的关系；

这个问题的难点在于如何利用有限的数据获得尽可能多的信息，合理选取的特征值，通过一系列的分析得出UGC系统的一些特性。

## 本文的工作

本文主要对YouTube网站视频文件进行分析，从而引出UGC系统的一些特性，有助于互联网服务提供商针对系统的特性更好地配置文件缓存，从而减少网络的拥塞状况，提高用户体验。

## 本文的结构

综上所述，本文针对YouTube网站的视频文件分析得出UGC系统的特性。本章为全文的引言，主要简述研究的背景、意义以及研究的问题。第二章主要介绍与本文相关的前人工作和成果，包括Meeyoung Cha等人对UGC系统的分析，以及康石对视频点播中的用户行为的分析，第三章简要叙述实验的方法，第四章进行实验结果分析，得出UGC系统的特性并且提出相应CDN缓存的思路。第五章作为总结全文工作，并对未来的工作进行展望。

# 相关工作

本章主要介绍了与本文相关的工作，包括Meeyoung Cha等人对UGC系统的分析，以及康石对视频点播中的用户行为的分析。

## UGC系统分析

Meeyoung Cha在文献[2]中对YouTube视频网站整一个UGC系统进行了分析，并为解决UGC系统的拥塞问题，加快视频缓冲速度，提出并比较了几种基于UGC系统的CDN策略，并且提出了P2P在UGC系统中的潜在效益。

主要分析的方面有以下几点：一、比较了一些突出的UGC系统和其他标准的VoD系统诸如Netflix和Lovefilm，找出两种系统的主要不同点，主要集中在内容生产、消费和用户参与模式等属性；二、通过从多种类别的UGC服务中分析流行性分布, 通过追踪其时间的演变, 显示出了UGC流行性分布显示出的幂律；三、追踪在UGC上的内容盗版和非法上传，以及它们对不同系统特性的影响，内容盗版和非法上传是当今的UGC系统的负面问题，因为它们会降低UGC系统的效率并牵出高昂的诉讼成本。

Meeyoung Cha针对UGC系统提出了三种CDN的缓存策略，一是静态有限缓存，即对长期流行度高(观看量大)的视频进行缓存，缓存内容在一定时间内不改变，这种策略维护成本较低，但是UGC系统每天有大量新视频产生，这个策略针对这种情况的应对并不理想；二是动态无限缓存，即初始缓存是所有被用户访问过的视频，之后新产生的被访问视频全部都加到缓存中，这种策略是最理想化的策略，但是成本过于高昂且难以实现，只能作为一个参考；三是混合有限缓存：工作方式与静态有限缓存相似，但是会留出额外的空间去存储每天流行度最高的部分视频，这种策略可以说是前两种策略的折衷，既考虑到了UGC系统的特性也可以节约成本。

针对这三种策略，Meeyoung Cha对YouTube的视频数据进行了模拟，第二种策略是最优的，而第三种策略次之，并且比策略一的效率要高出30%左右。

文献[2]通过对UGC系统的分析，为UGC系统的内容分发提供了可行方法，不过方法中还存在着以下几点不足：其一是没有考虑视频大小对网络流量的影响，文献仅仅考虑以视频的观看量作为流行度去做CDN策略的分析，而每个视频的文件大小不一样，因此它们在CDN缓存中所造成的消耗也不同，在做CDN策略的时候应该将视频文件大小考虑进去；其二，没有解决CDN系统运维成本较高的问题，需要大量部署边缘缓存服务器，且某些边缘缓存服务器之间没有集中式的管理机制，各个节点之间的配置是相互独立的。

## 视频点播中的用户行为分析

文献[1]设计了一个针对VoD系统的日志分析系统，然后利用奥运会的视频点播数据系统进行分析，并且将分析得出的用户行为数据与其他VoD系统的用户行为相比较，得出各类视频点播服务的共通之处和区别。

根据对用户行为的分析，提出了一个基于CDN和P2P结合的内容分发系统，并且对该系统进行了性能分析和测试，新的混合架构相对于之前的单一CDN架构或单一的P2P架构提高了系统稳定性，在保证客户端播放的流畅性的同时减少了视频冗余，有效的提高了整个系统的负载，同时没有过多的增加服务器的负载。

# 数据分析方法

本章详细描述了本文研究所用到的方法。

## 分析内容

通过对YouTube的视频观看量、视频时长、用户评价等数据1进行特征提取分析，发掘UGC系统的特性。

## 元数据解析

原始数据保存在文本文档中，数据量大小为167万行，每一行代表一条视频信息，格式如下：“url | length | views | ratings | stars”代表一行数据，不同属性用竖线隔开，包括视频在YouTube主站的相对地址、视频时长、视频观看量、视频评论数、视频评分。例如“/watch?v=abc|01:30|100|5|4.0”表示视频地址为 [www.youtube.com/watch?v=abc](http://www.youtube.com/watch?v=abc)的一段视频，时长有一分三十秒，观看量为100，评论数为5，评分为4.0。

因此对原始数据的解析只需要根据“|”作为分隔符拿取每一行的每一段数据。

## 数据结构

Python语言提供了很多原生的基本数据结构，其中的列表list结构简单，很适合做这样的分析。

## 特征提取

原始数据本身的属性不多，所以特征提取的过程比较简单，只需要根据具体需要分析的一个或一组特征值进行提取即可，对于本次实验，一组特征值一般不会超过2个。

## 绘制图形

利用matplotlib对需要处理之后的数据绘制相应的图形，将数据直观的展现在图像上。

Matplotlib是Python中应用最多的2D图像的绘图工具包，使用matplotlib能够非常简单的可视化数据，它可以绘制多种图形的矢量图，包括线图、点图、柱状图等等。由于其强大的绘图能力和简单易用的特性，被很多开发人员应用于图像绘制当中。

# 实验结果与分析

本章展示了对原始数据的分析结果，针对YouTube网站视频观看量的分布情况，发掘了UGC系统的特性，并且分析了针对某些特性要进行文件缓存配置所需要注意的点。

## 对YouTube网站视频的观看量分析

接下来展示下面分别以不同的属性作为特征分析的结果。

### 视频观看量的分布情况

图4-1显示了视频的观看量和排名的关系，其中横坐标是视频按照观看量从高到低的排名，Rank = 1为观看量最高的视频；纵坐标是视频的观看量。图象在 时表现出Zipf分布的规律，Zipf分布是在研究自然语言时被发现的定律，可以表述为在自然语言的语料库里，一个单词出现的次数与它在频率表里的排名成反比，后来人们发现，不仅仅是自然语言的范畴，有很多现存的系统在某些属性上也都能表现出Zipf分布的规律，包括UGC系统在内，这说明UGC系统大部分的观看量集中在少部分视频中。

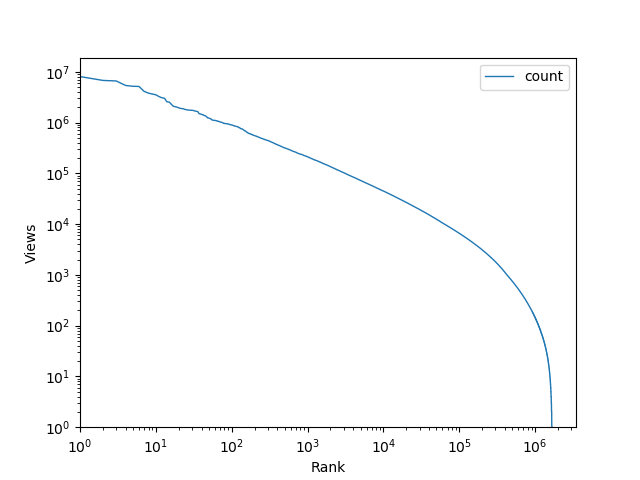


图4-1 视频观看量按照排名的分布

通过图4-2的累积分布可以得出一个结论：排名前10%的视频的观看量占据了超过所有样本观看量总和的90%，这个结果对UGC系统的CDN策略决策有着深刻的影响，考虑极端的一种情况，使用CDN缓存UGC系统排名前10%的视频，能够覆盖90%左右的观看量，这样的维护成本相比于大部分甚至是全部内容的缓存低得多，但这仅仅是理想情况，因为实际情况是UGC系统的视频流行度每时每刻都在变化，用户的随机性很强，实际的缓存覆盖的观看量远远没有90%那么多。在 之后这种Zipf分布的规律迅速衰减，出现了长长的“尾巴”。可能的原因是YouTube启用了某种过滤系统，导致观看量较少的视频更难出现在用户的视野中，排名靠后的视频观看量更难上升，再加上存在一定的“马太”效应因此曲线的尾部有一个迅速下降的趋势。

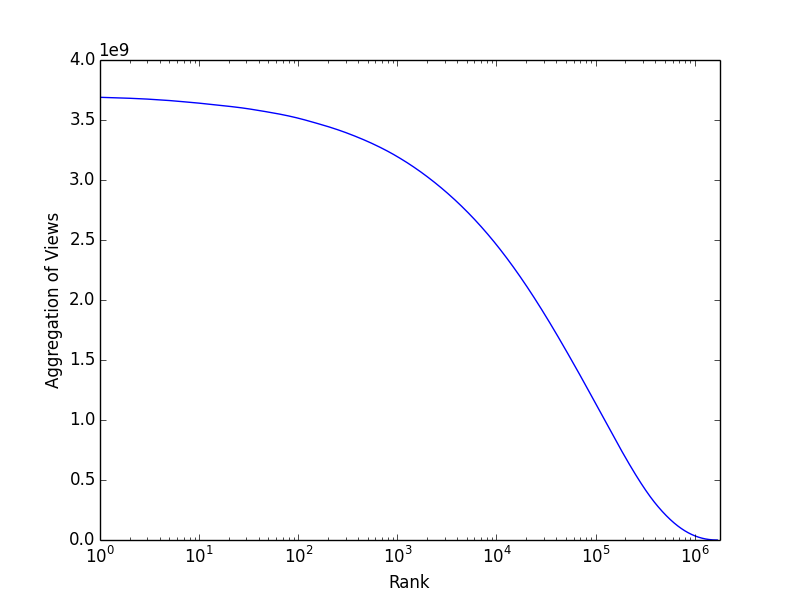


图 4-2 视频观看量按照排名的累积分布

进一步对视频观看量的分布进行分析，图4-3展示了视频观看量的数量分布图，横坐标是视频观看量，纵坐标是视频数量，视频的数量主要集中在观看量小于100的视频，这一部分的视频数量有着超过50%的占比，观看量较小的视频“垃圾”充斥着视频网站的内容，而在内容分发系统中，这些重要性不那么突出的视频却与其他观看量大的视频占据了同样的甚至更高的带宽，对内容分发系统的资源造成了浪费，这是需要避免发生的，所以视频分发策略应该保证这些观看量小但是基数巨大的视频不占用过多的缓存资源。

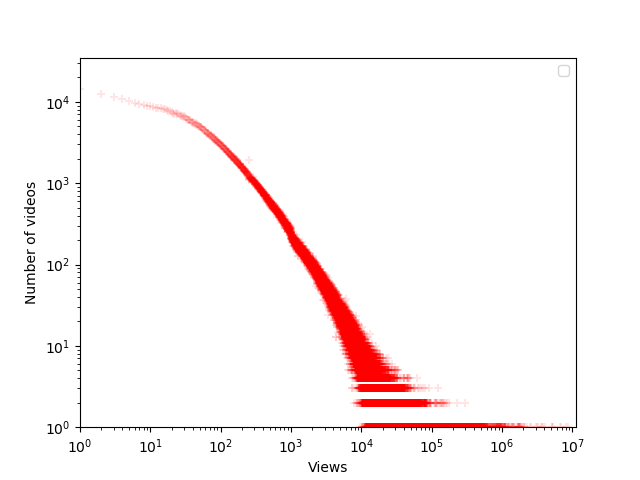


图4-3 视频观看量的数量分布

图4-4展示了视频观看量的累积曲线，横坐标是视频观看量，纵坐标是观看次数低于x的视频的观看量总和，从图4-4可以看出，中等观看量（介于3000次到100000次之间）的视频数量有着最大的占比，这一部分的视频是网站的主力，贡献了最多的观看量。因此，在CDN策略中，在考虑到缓存高观看量的视频的同时，应该把这部分中等观看量的视频纳入重点考虑的范围。

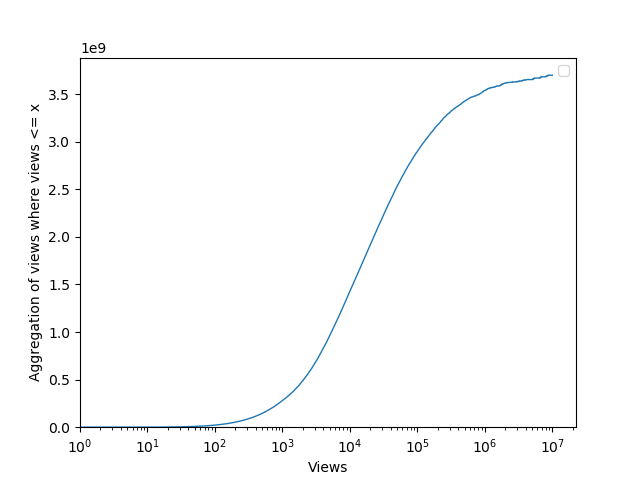


图4-4 视频观看量的累积曲线

### 视频时长的分布情况

图4-5展示了视频长度的概率密度分布，视频的长度集中分布在一分钟左右，3秒到100秒的视频居多，长度大于700秒的视频数量已经可以忽略不计。

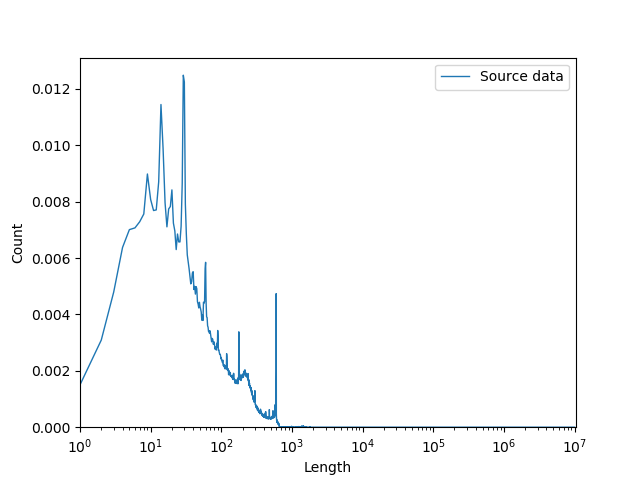


图4-5 视频长度的概率密度分布

从图4-6的累积分布也可以看出，长度在3到100秒的短视频占据了UGC网站的大部分。对于传统的Vod视频网站，比如提供电视剧、电影、综艺节目等内容的视频点播网站的视频长度都较长，少则十几二十分钟，多则两三个小时，相对来说，UGC系统提供的视频大多数为2分钟以内的短视频，这是因为在UGC系统发布视频内容几乎不需要成本，而且大部分用户发布自制视频并不是为了盈利而仅仅是娱乐需要。视频内容的长度是UGC系统相对于其他Vod系统的一个重要的区别，UGC系统以一分钟左右的短视频为主流。由于视频长度本质上可以对应视频的文件大小，所以长度越长的视频代表占用越大的带宽，因此在设计CDN缓存的时候应该考虑把视频的长度考虑在内。

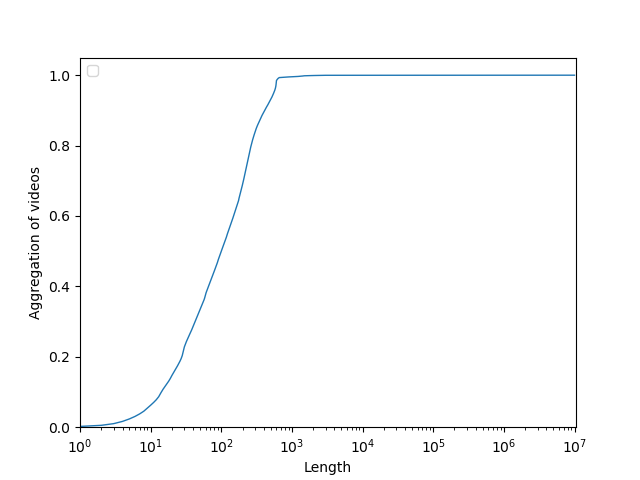


图4-6 视频长度的累积分布

### 视频观看量与视频长度的综合分析

既然视频观看量和视频长度都对CDN缓存的带宽有所影响，因此需要对两者进行综合的考虑。图4-7是视频观看量与视频长度的关系，从图中可以看出，视频长度大于700秒的视频观看量几乎为0，视频观看量较多集中在3秒到100秒的视频。

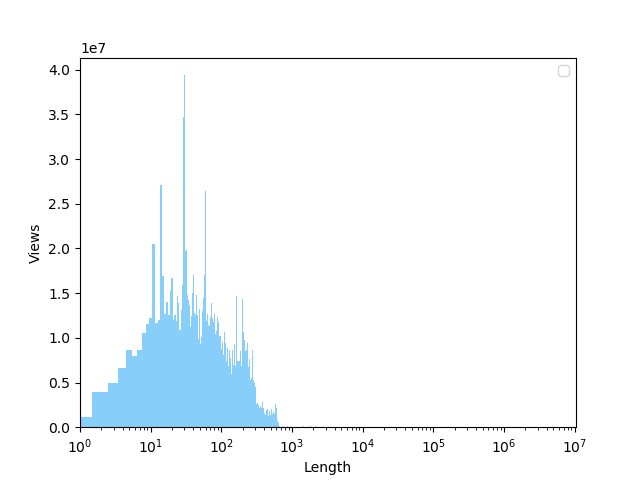


图4-7 视频观看量-视频长度分布

图4-8中视频观看率的累积分布也表明了这种分布情况。观看量与视频长度的概率密度分布展示了一定的规律，但是这种规律不能排除是视频长度的自身分布造成的，下面会验证这一点。

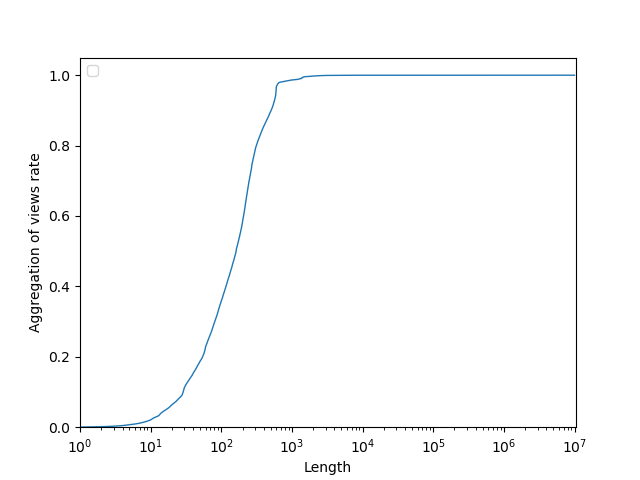


图4-8 视频观看量-视频长度累计分布

为了消除视频长度分布的影响，图4-9 对图4-7中的视频观看量取了平均值，展示了视频的平均观看量与视频长度的关系，可以发现它展示的规律与图4-7的分布不同，1000秒到10000秒的视频平均观看量较高，而小于1000秒长度的视频平均观看量几乎为0。

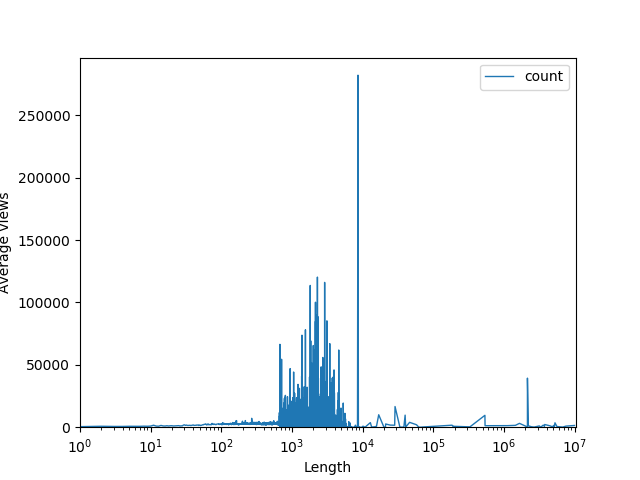


图4-9 视频平均观看量-视频长度分布

图4-10对应的累积分布也证明了这样的规律。这是意料之外的结果，但是经过一番思考之后也能想到其中的原因：虽然前面得出了视频观看量较多集中在3秒到100秒的结论，但是由于UGC系统中长度小于100秒的视频数量占据了绝大部分，这一部分的基数比较大，所以自然会累计较高的观看量，然而UGC系统的一大特性是少部分视频贡献了大部分的观看量（从图4-1可以看出），大部分的短视频还是无人问津，所以视频观看量在平均之后就会显得非常惨淡。

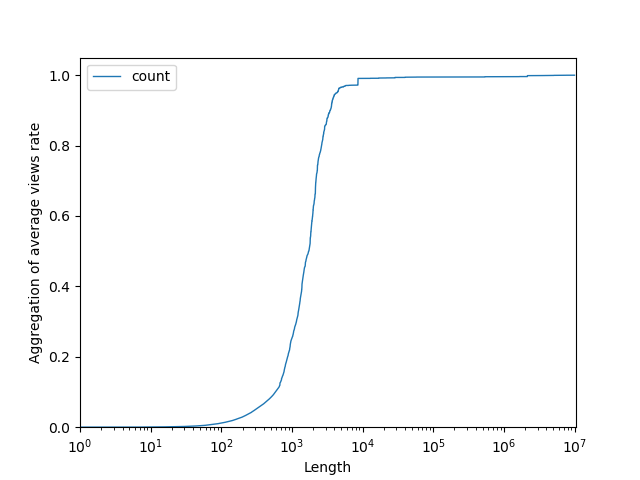


图4-10 视频平均观看量-视频长度累积分布

至于为什么1000秒到10000秒的视频平均观看量高，可以发现这个区间的视频长度刚好是40分钟左右电视剧、或者两三个小时电影等内容的长度，这种较长的视频制作成本一般较高，而且多由比较专业的个人或团队制作，其中也有很多由用户从其他视频网站上搬运过来的电视剧、电影等等，这种视频的受众都比较广泛、稳定，所以都拥有了较高的观看量。另外，这种播放时长较长的视频相对于100秒内的短视频基数小得多，不容易被平均。因为1000秒到10000秒长度的视频在整个UGC系统中只占有了极小的一部分，但是它们有着很高的平均观看量，也就是说这部分的视频可能会被频繁访问，而且这部分长视频文件大小较大，对带宽的要求更高，所以在设计CDN缓存的时候，应当把这一部分的高观看量视频纳入考虑。

### 视频观看量与视频评价数的关系

一般来说，用户对视频的评价表明了用户对这个视频有较深的兴趣，视频观看量只能表明用户点击这个视频的次数，评价数比视频观看量更能表示视频的受欢迎度。图4-11 展示了视频观看量与视频评价数的关系，剔除了评价数为0的视频。评价数为2的视频观看量最高，然后观看量随评价数升高而递减，评价数高于100后观看量降低到了较低的水平。

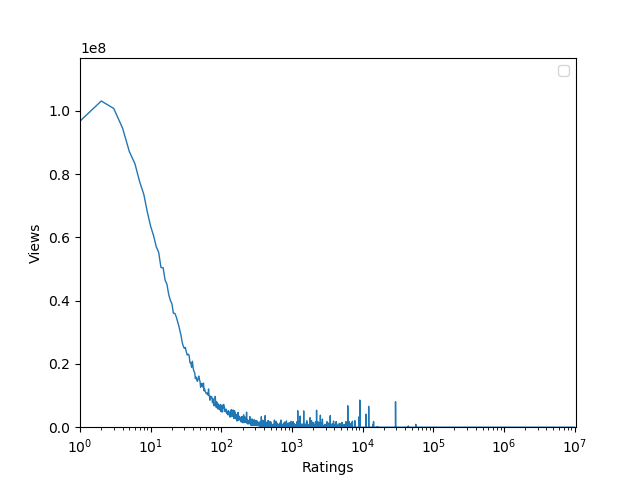


图4-11 视频观看量-评价数分布

图4-12是相应的累积分布，评价数小于200的视频占据了80%的观看量，评价数低的视频基数较大造成了这种结果。

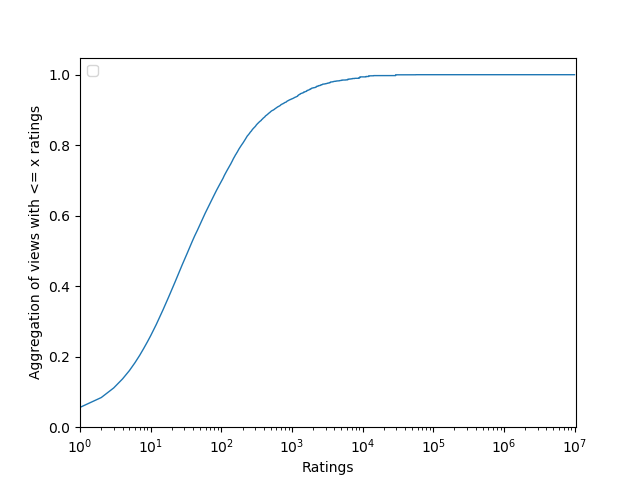


图4-12 视频观看量-评价数累积分布

接下来我对不同评价数对应的观看量取平均值，下面展示了结果。

图 4-13是视频评价数对应的平均观看量曲线，评价数在1000到10000之间的视频观看量较高，这一部分的分布比较集中，但是没有展现出明显的规律，进一步观察图4-14的累积分布，评价数在1000到10000之间的视频平均观看量远远高于其他区间的平均观看量，而评价数低于200的视频平均观看量因为过大的基数而被稀释，评价数大于10000的视频则因为数量太少没有对分布产生显著的影响。对于观看量相同的视频，评价数较高的视频往往拥有更大的用户粘性，用户必须对某个视频有足够高的关注度才会去对其作出评价，所以用户更有可能再次观看这个视频甚至把它推荐给其他用户。在考虑CDN缓存设计的时候，也不能忽视这一部分的影响：评价数在1000到10000之间的视频，这一部分的视频基数小，但是观看量高，对网络流量拥塞有着一定的影响。

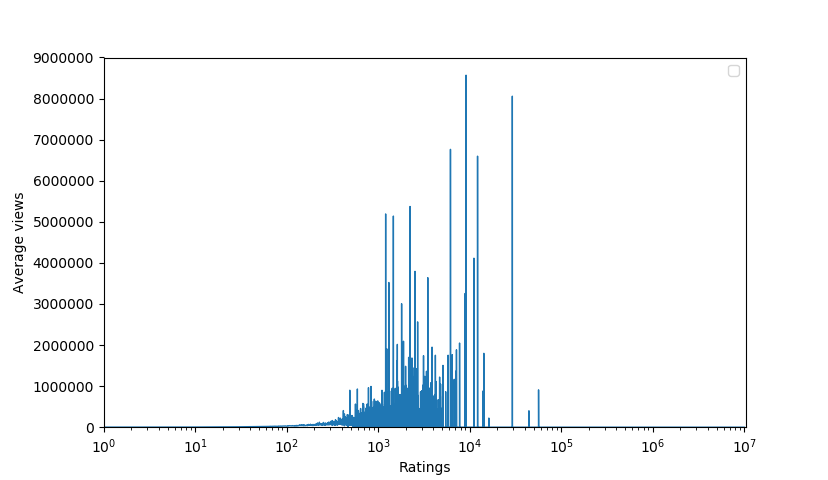


图4-13 视频平均观看量-评价数分布

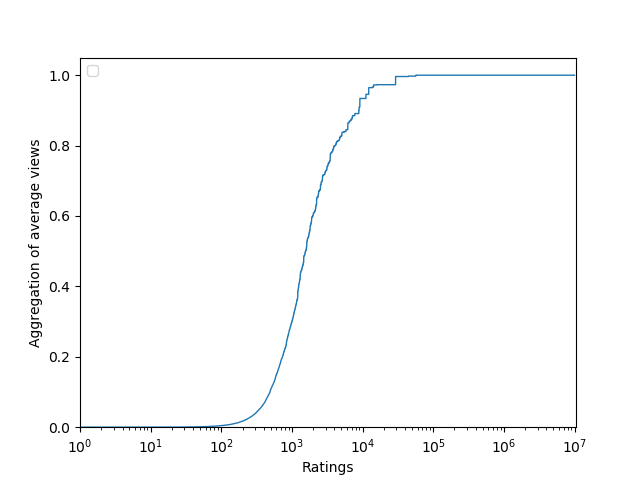


图4-14 视频平均观看量-评价数累积分布

## 分析总结

针对YouTube网站视频观看量的分布情况，本章分析了UGC系统的一些特性，并且分析了针对某些特性要进行文件缓存配置所需要注意的点。首先，大部分的观看量集中在小部分的视频中，视频观看量的排名规律符合Zipf定律，这说明了针对小部分视频的CDN缓存可以显著缓解流媒体服务器的压力；其次，在视频长度方面，UGC系统的短视频数量占据了大部分，因此短视频的观看量也很大，然而从平均观看量来看，长视频的观看量更加稳定，且结合其文件大的特点，更加适合配置CDN缓存；最后，视频的评价次数也可以作为配置CDN缓存的参考，由于评价数低的视频基数太大而平均观看量被稀释，评价数高的视频基数过小没有对系统产生显著影响，评价数居中的视频往往有着更加稳定的观看量分布，配置缓存的时候可以从这个角度去入手。

分析UGC系统的特性有助于流媒体提供商针对自身的系统特点实现更加精确的缓存配置，在提高系统稳定性、节约成本等方面都能起到很好的效果。

# 总结与展望

## 本文的总结

本文的主要贡献是通过分析YouTube网站视频观看量的分布情况，得到了UGC系统的一些特性，并且提出了针对某些特性要进行文件缓存配置的思路，有助于互联网服务提供商更好地配置文件缓存，从而减少网络的拥塞状况，提高用户体验。

具体来说，分别从视频观看量、视频时长、用户评价等角度分析了YouTube的视频观看量分布，总结出UGC系统的流行度在不同观看量、时长、评价区间的演变规律，并且利用这种规律提出针对性的CDN缓存的思路，为潜在的提高CDN系统效率实现了理论基础。

## 未来的展望

基于现有的工作基础，在未来可以对以下方面进行深入研究：寻找更多的不同的UGC系统的数据进行分析，比较它们之间的不同；完善文中所提出的CDN缓存的思路，并且形成一个架构的实现并对其性能进行检测。

# 参考文献

1. 康石. CDN-P2P混合架构视频点播系统的实现与用户行为分析[D].清华大学,2010.
2. M.Cha, H. Kwak, P. Rodriguez, Y. Ahn, S. Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World’s Largest User Generated Content Video System.Proceedings of the 7th ACM SIGCOMM conference on Internet measurement,1-14.

# 致 谢

诚挚感谢指导老师范正平副教授，尹老师悉心的教导使我得以了解专业领域更为深奥的内容，不时的讨论并为我指点迷津，帮助我开拓研究思路和方法，精心点拨、热忱鼓励，使我在研究设计中受益良多。

# 附 录

毕业论文成绩评定记录

|  |
| --- |
| 指导教师评语：  成绩评定：    指导教师签名： 年 月 日 |
| 答辩小组或专业负责人意见：  成绩评定：    签名（章）： 年 月 日 |
| 院系负责人意见：  成绩评定：  签名（章）： 年 月 日 |