摘要：通过对视频网站观看量数据的分析，为视频网站cdn策略提供一种思路。

我从 #todo 拿到了 #todo 数据，并对其进行整理分析。

xFig0

1. 如果符合Zipf分布（#todo简单介绍一下zipf分布），那么图象应该是一条直线。图 #fig0 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x<=5时表现出Zipf分布的规律，（#todo这里说明一下有什么意义）但是x>5之后迅速衰减。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，因此曲线的尾部有一个迅速下降的趋势。

xFig1

fig1\_1

1. #fig1\_1展示了视频观看量的分布图，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，这一部分的视频数量有着极大的占比，y = -klogx + b，因此，观看量较小的视频充斥着视频网站的内容，而在视频分发中，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，这是不应当发生的，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满 #高速缓存。

fig1\_2

1. #fig1\_2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#fig1\_2可以看出，中等观看量（介于3000次到100000次之间）的视频观看量有着最大的占比，这一部分的视频是 #UGC 网站的主力，贡献了最多的观看量。（#有空算一下概率密度函数）

xFig2

fig2\_1，2\_2

1. 图#fig2\_1展示了视频长度的概率密度分布，视频的长度集中在10秒左右，3秒到100秒的视频居多，长度大于700秒的视频几乎不存在。从 #fig2\_2 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。对于一般的视频提供网站，比如提供电视剧、电影、综艺节目等内容的视频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 视频网站提供的视频大多数为2分钟以内的短视频，这与视频生产者和盈利模式有关，这里不做赘述。

xFig3

1. 图#fig3\_1是视频观看量与视频长度的关系，从该图可以看出，视频长度大于1000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#fig3\_2 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关系。
2. 为了消除视频长度分布的影响，图#fig3\_3 对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，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#fig3\_4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，而 #UGC 系统的特性是少部分视频贡献了大部分的观看量（从#fig0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。

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的视频相对于100秒内的短视频基数小得多，不容易被平均。

xFig4

1. 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#fig4\_1 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#fig4\_2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，这可能是评价数低的视频基数较大造成的，接下来我对 #评价数 取平均值，下面展示了结果。
2. 图 #fig4\_3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 #fig4\_4 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，这说明了 #todo啥 ，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。