分析大规模的UGC内容系统的视频流行度特性

四、流行度分布：

一般情况下，分布情况影响了生成它的潜在机制，针对流行度很高以及很低的部分分别讨论其分布的形状。

A. 二八定律：

图 3 ：从右往左看，不断将其观看数进行累加，设其曲线为y = f(x) 则化为cdf形式为 1-y。10%的视频的观看次数占到了总数的80%

点播系统曲线的弯曲程度要小于UGC系统，10%的视频的观看次数占全部的60%。

是由于这样的弯曲程度的分布是根植于UGC系统的特性或者一个更好地推荐系统能够把用户从流行度很高的视频带到流行度不高的视频

高强度的弯曲的一个直接应用是缓存。缓存存储一些应对高点击率的视频集就能让缓存变得很有效率。另一个应用是流行内容的p2p分布。

B. 统计特性：

研究UGC流行度的统计特性并检查用户的请求是如何分布于流行与非流行内容之间。使用两种不同的方式来表示流行度的分布：a. 一个能够表示视频的观看数和比其观看数多的视频的个数的频率图。b. 视频排名以及视频观看数的图。第一个图主要关注于流行度高的视频，第二个主要关注与流行度低的视频。

1. 流行内容分析：图4 a和b 表示了YouTuBe以及Daum的四个代表视频集的流行度分布。直线代表指数分布。相对于指数分布的直线来说，YouTube 的sci 和Daum的食物集的数据的流行度高的视频具有更快的衰减。 4 c 表示了Sci得各种不同的拟合。 最好的拟合是与平方指数分布有着相同的指数并有着指数中断的弯曲，但其尾部在log-normal之上。

拟合结果显示：尾部衰减的最佳拟合是平方律分布增加一个指数分界。然而真正的分布是依赖于视频分类的。

平方律分布的最有名的解释是yule过程（优先选择，富的更富）：有k个用户已经看了一个视频则其他用户看这个视频的速率是正比于k的。该过程并未解释分布的尾部的衰减。检查三个模型是否可以应用于我们的场景。

第一个模型 Amaral et al，时间的影响能够产生一个衰减。由于所有时间的视频集均呈现衰减所以时间的影响并不能应用与我们的案例。老的视频并不一定不活跃：一天中80%的请求是指向时间超过一个月的视频，一些老的视频甚至出现于最流行的列表中。

第二个模型 Mossa et al ，一个完全不同的模型来解释www的分布程度—— 信息过滤. UGC系统中一个用户不能收到所有可得的视频的信息，只能收到一部分视频的信息或者现有页面的固定数量。由于信息过滤，优先选择被阻碍、指数中断便出现了。该情况在UGC和视频点播系统中都有。然而，非常流行的视频会在点播系统中有显著的代表性并且来吸引更多的观看者，因此在我们的案例中视频过滤并不会引起衰减。

第三个模型 Gummadi et al 提供一个关于尾部曲线的更好解释。一项关于P2P下载中的文件流行度的研究表明，扭曲来源于最多一次获得的行为：P2P的用户仅获取一个文件一次。U代表用户数，V代表视频数，R代表每个用户的平均请求数。作者使用两种类型的用户群：平方及仅获取一次。两个用户群请求文件基于相同的初始zipf流行度。平方用户群能够请求一个视频多次，另一组仅能请求一个视频一次。结果流行度分布：平方用户群为直线，另一组出现曲线。（一般一个用户只会请求一个视频一次）

例如R和V的系统加上有限次获取的行为就会产生尾部衰减。

R增加则图线的衰减增加：R比较小时，有限获取的影响比较小，R增大时有限获取为主要影响因素，因为用户有很大的机会选择相同的流行的文件。增加U会使视频的观看数增加（x轴）,然而图像的整体形状并没有改变，推测用户数U对尾部形状的影响很小。同时增加U和R是图像的形状跟只增加R时的变化相似。

实际的UGC系统中，R和U同时增长，这代表不断有新用户加入系统同时旧用户发出的请求越来越多。实际的数据中，长时间的代表R跟U更大，一天内的流行度展现出一个明显的平方律形状，长时间的则表型出明显的尾部衰减。

如果V很小时，有限获取的影响将被放大因为只有一小部分视频能够供用户选择。图5C对比于具有更少的v的图4 a,b 可以看出图4中的衰减更加明显。

综上，流行度高的视频主要影响因素为R和V

2. 长尾分析：非流行项目的潜在分布是什么？是什么形成了这种分布？UGC服务能从长尾中得到何种程度的利益？

6 a为view-rank的对数坐标图，展示了一个直线腰部并有着衰减尾部的图像。进行图像拟合时，该图更加贴近于平方律并有着指数中断的曲线。第二好的拟合是log-normal的图像，虽然它不能非常好地拟合图像的尾部。

下面讨论两种尾部衰减的原因：

首先，Zipf流行度分布在现实世界是非常流行的。UGC的内容在质量上跨度很大，它的很显著的一部分对于大部分用户是没兴趣的。

其次，分布的本质形状是Zipf，衰减的尾部可能是由于系统的瓶颈（信息过滤或推送过滤）。推荐或者搜索引擎一般会返回或热衷于一小部分的流行项目，驱使用户远离不流行的项目从而造成了尾部衰减的现象。这种衰减会随着时间的发展变得更加明显，因为老的不流行的视频会长久地暴露于推送过滤之中。 6 b展示了不同时间的流行度分布。仅一天的视频相比其他的视频来说弯曲程度明显更小。

如果ZIpf是本质的分布并且尾部衰减时由于不可忽略的瓶颈造成的。那么在一个没有瓶颈的系统中弯曲区域的视频会得到应有的观看。这给用户提供了一个更好的机会来发现*不流行的视频。同时会给版权公司提供潜在的商业机会。为了评估去除这样的瓶颈所带来的潜在利益，我们计算了在最佳拟合zipf曲线中的额外观看数总数在现有观看数总数中的比例。 表三展示了对UGC视频分类来说估计的利益。YouTube在长尾经济中具有更大的机遇，因为它有相当大数量的视频可以获益。Daum来说视频总数并不多，收益自然也会减少。当视频数比较少时系统的影响就比较小，因为信息能够被更加容易地找到。*

*对于大部分的UGC数据来说最好的拟合是Zipf加上指数中断。zipf在本质上是尺度无关的，指数分布是尺度相关的。因此这两种分布会很少在一种机制中连贯并自然地出现。一种更加可能的解释是：潜在的机制是zipf、指数中断揭露了过滤的影响造成了尾部的衰减。揭露真正造成尾部衰减的机制需要后续更加深入的研究。*

*五、流行度随着时间的演变：*

*首先分析请求是否集中于老的或者新的视频。其次调查对于不同时间的视频来说，流行度排名的改变有多快，之后进一步测试一个视频的未来流行度是否能够被预测。*

*A.流行度分布VS时间：*

*为了研究被请求视频的时间分布，首先将视频按照时间进行分组（5天一组），然后对每个时间组的总请求容量进行计数。时间短的视频组的视频数量要多于时间长的视频组的视频数量。7a展示了每个时间组的平均请求的最大值、中位数以及平均数。纵轴为对数坐标。对于时间不超过一个月的视频来说，我们看到平均请求有个微小的增长，这意味着观看者对新视频有更大的兴趣。然而这个趋势在最大值上却并不十分显著。一些老的视频也得到相当多的请求。实际上，结果表明给定的一天中80%的被请求视频是时间长于一个月的并且这些流量占到总请求流量的72%。一年以上的视频曲线变得混乱，由于视频的数目很少。总而言之，如果我们排除了非常新的视频，用户的偏好看起来变得对视频的时间变得不敏感。同时新视频比了老视频得到的观看数更多因为新视频组包含的视频数目更多。*

*虽然用户的兴趣在一定程度上是对视频时间不敏感的，但在任何一个给定的一天中请求最多的视频是时间短的视频。为了验证这点，我们研究了20个请求最多的视频的分布。7b展示了四种不同的时间窗口：一天、一周、一个月、所有时间。一天之内，20个最热视频中大约50%的视频时间少于三个月。然而，随着时间窗的增加，时间的中位数向老的视频转移。这展示了新视频的流行度的短暂的特性。*

*对于视频点播系统进行重复试验。租借的中位数是对视频时间不敏感的。相对于YouTube，Netflix显示出一些独特的模式：a. 租借数目的最大值是严格的时间不敏感的b.老视频的最小值比新视频的最小值要大。结果显示：18%的租借是由2003年之后的视频（视频集的10%）所创造的，72%的租借是由2003之前的视频创造的。*

*7d 展示了一个月、一年、所有时间中Netflix的租借最热100的视频以及所有视频的时间分布。当我们增加时间从一个月到所有时间时最热视频的平均时间不断增加。这表明在给定的一个月内最流行的视频稍微偏向新的视频。视频集的前100的视频的时间跨度为1年到将近9年，所有视频的时间跨度为1个月到109年。为了更好地理解UGC视频的动态流行度特性，我们将讨论视频流行度将如何随着时间演变。*

*B. 短暂的聚焦*

*现在研究个别的UGC视频的流行度如何随着时间进行演变，流行度的变化有多快或者多慢，以及一个视频的流行度是否可预测。*

*1. 视频随着时间被观看的概率：图8 ：到了X天时，观看数少于V（0-10000）的视频的比例。图像显示：第一天后，90%的视频至少被看了一次，将近60%的视频被看了超过十次。横轴的对数表示一个给定的视频的请求将会随着时间显著减少。当我们考虑V=10时，一个给定的视频得到10次以上的请求在24小时、6天、3周以及11个月的概率分为为0.43、0.18、0.17以及0.14.这表明如果一个视频在它的第一天没有得到多次的请求，它将不太可能会在将来得到很多请求。*

*2. 预测不远的将来的流行度：预测未来的流行度的能力在很多方面是很有用的，因为服务提供者可以使用多重代理或者缓存来预处理那些将流行的视频，同时内容所有者可以使用快速的反馈来更好地管理他们的内容。为了探索利用之前的浏览记录来预测未来的流行度的可能性，我们将前几天的视频和后来的视频进行对比。如果两个视频集有100%的关联就意味着是完美的预测。然而，如果两者的关联很小（80%以下），这就意味着我们不能基于先前的浏览来预测后来的视频流行度。*

*7天后的视频分别跟24小时、1天、2天的相关系数分别是0.5885、0.8793以及0.9367.第一天的观看是不稳定的，因为一些视频尽在系统中暴露很短的时间。我们的结果表明第二天的记录有着相当高准确性的估计（相关性接近90%）。使用第三天的记录仅能少量得提升预测的准确性。当与上传90天后的视频进行对比时，第二天、第三天的相关系数分别为0.8425、0.8525，这表明跟很久远的流行度也有着很高的相关性。*

*3. 流行度变化：现在将新视频和老视频变得流行的可能性作为时间的函数。首先观察视频排名如何随着时间进行改变。图9a中我们使用流量六天数据中的两个（第0和第5）并且只考虑出现在其中的视频。我们通过时间（10天）将视频进行分组然后绘制它们的排名随着视频时间的变化。对于每一组取排名变化的最大值，99%的值以及平均值。纵轴从-4059到235,132说明一些视频排名在采集期间降低了4059，另一些增长了235,132。*

*年轻的视频能够很快改变排名位置然而老的视频的排名波动很小。一些老的视频仍然能够增加显著地增加排名，这意味着由于长尾效应以及好的推荐机制老的视频也能在很长一段时间后变得流行。然而很难从图中得出这个结论，因为很少的视频能够有大幅的排名变化。*

*最大值和99%之间的间隙表示只有一小部分的年轻视频（少于1%）能够产生比较大的排名改变。同时所有时间组的连续最小值曲线在-4000附近，仔细观察之后发现在采集期间他们并没有得到请求因此他们的排名由于别的视频至少有1个请求而变得靠后。这表明没有收到任何请求的非流行视频将会在排名表中死亡。*

*仅仅关注观看次数或者排名的真实变化对于研究流行度分布的主要变化是不够的，视频可能收到很多请求却有很小的排名变化，反之亦然。为了辨别产生巨大排名变化同时收到很多请求的视频，我们提议使用排名变化以及观看数增加的乘积（图9b纵轴对数）。图中并没有500天的数据点，因为视频组里的所有视频都有排名的下降以至于排名的改变为负值。图中我们看到新视频在流行度上的更大变化，几乎没有一个老视频能够得到足够多的请求来在流行度图表中产生一个显著的上升。总而言之，老视频复苏的现象在我们的数据中很少出现。*

*C. 最流行视频的时间演化：*