破解 YouTube 的视频推荐算法

英文： Matt Gielen & Jeremy Rosen 翻译：刑无刀（他的知乎主页）

如果你是某个发行渠道（比如电影、戏剧、电视节目、网络视频）的内容工作者，那么内容的成败就取决于发行机制的运转逻辑。比如说，你制作了一档电视节目，你很想它能火起来，那么你就得知道该在哪里切入广告，怎么宣传节目，上哪个频道播放，所选的频道能被多少家庭收看，等等，诸如此类。

如果你的发行渠道是YouTube，那么你最应该搞清楚的是YouTube的算法是怎么工作的。然而，全天下所有由算法来运营的平台，要搞清楚这一点那不是一般的困难。

YouTube没有把他们算法用到的变量公之于众。要搞清楚其算法的运转原理，即使数据很有限，我们也得对这个大大的黑盒子一探究竟。有些算法倚重的变量，我们是一点数据也拿不到的（比如缩略图，标题印象，用户访问历史，用户行为，会话信息，等），如果能拿到这些数据，那等于就是把YouTube的算法脱光了让我们看，然而呢，呵呵哒，并没有。

看起来我们啥都没有，但还是想尽可能用手上这点数据大致搞清楚其算法逻辑。所以，我的前同事（为什么是“前”同事呢？因为我最近从Frederator离职啦，哇咔咔）Jeremy Rosen花了半年时间分析Frederator自己掌握和运营的频道数据，想搞清楚YouTube的算法。

开始之前，先明确一下：这篇文章内所指的算法包含多个YouTube增长类算法（为你推荐（Recommended），建议观看（Suggest），相关视频（Related），搜索（Search），原始评分（MetaScore），等等）。这些不同的算法产品，各有侧重，但有一个共同点，那就是它们的优化目标相同，都是观看时长（Watch Time）。

观看时长

先要说清楚的，“观看时长”并不是说观看过的分钟数。这个概念我们之前也讨论过[1]，观看时长由以下指标构成：

- 访问次数
- 访问停留
- 会话开始

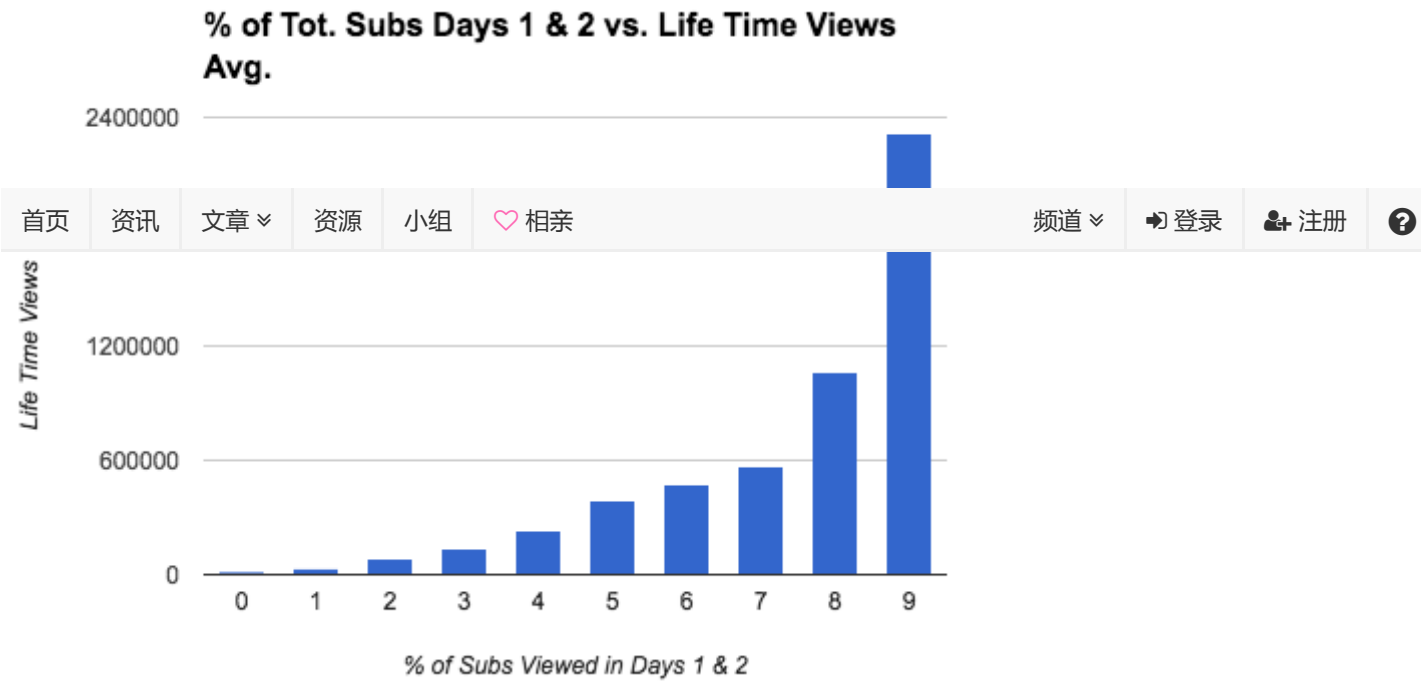
- 上传频率
- 会话时长
- 会话结束

本质上以上每一项都关系着频道以及频道的视频表现好坏，人们是不是经常来访问（开始一次页面访问的会话）以及是不是停留很长时间。

要在算法那里积累下任何变量的取值，你的频道和视频首先得有人来访问你才行。一个视频要成功（成功定义为订阅者中超过一半的人在前30天访问过）需要视频发布的前几分钟、前几小时、前几天内得到大量的访问，我们把这称之为访问速率（ View Velocity）

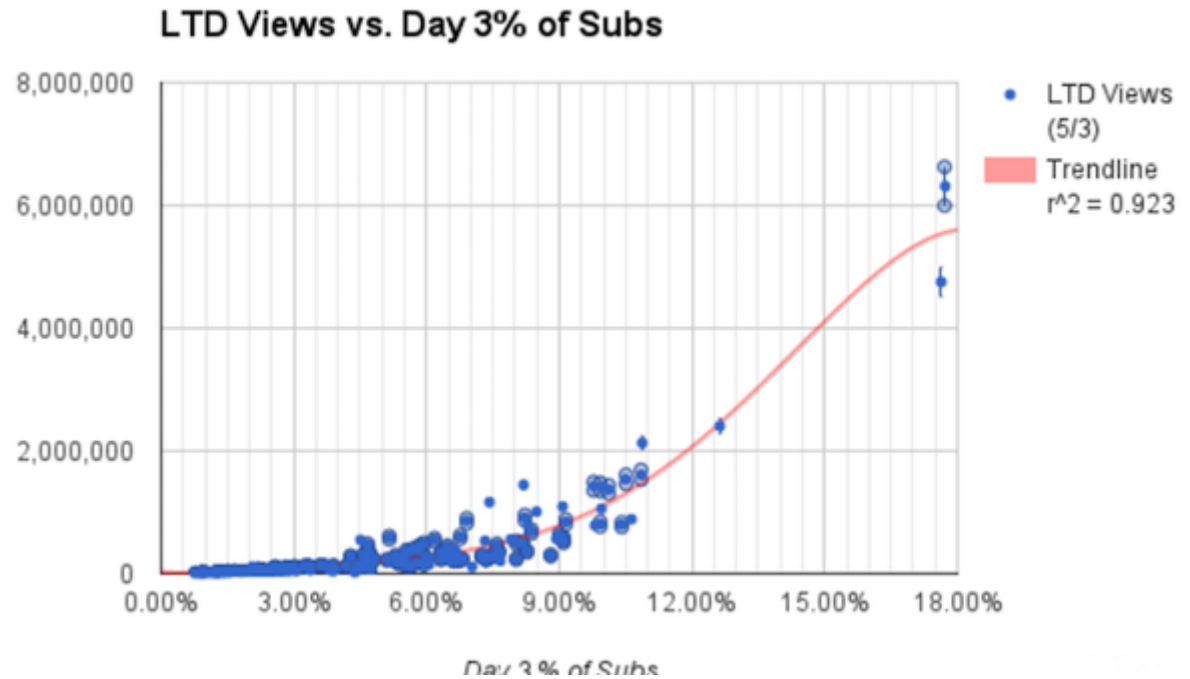
访问以及访问速率

我们分析Frederator的访问速率，发现整个生命周期内累计访问次数与前48小时内订阅用户访问百分比呈指数关系。



48小时内访问的订阅用户百分比与得到的平均访问次数

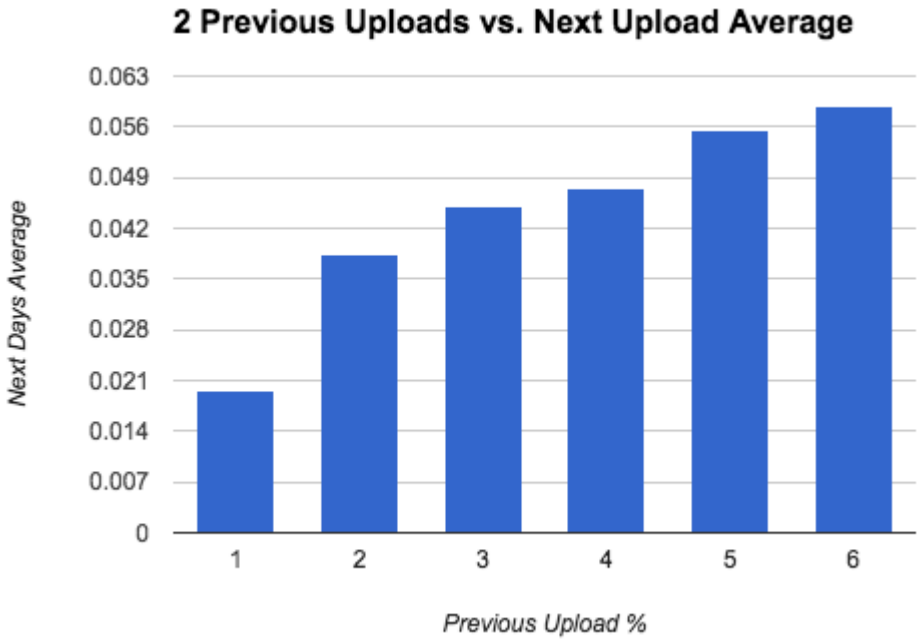
基于这个观察，我们稍微深挖了一下，发现用这个速率规律去预测一个视频是否会成功，可以做到92%的准确率。其实，还存在一个更直接的相关性：72小时内访问的订阅用户百分比，与视频整个生命周期的累计被访问次数之间。



72小时内访问的订阅用户百分比与整个生命周期内累计的访问次数

上一个和下一个视频。

下图说明如果Frederator上一个视频48小时内访问速率比较糟糕（少于5%的订阅用户访问），那么接下来上传的视频也会受其影响。

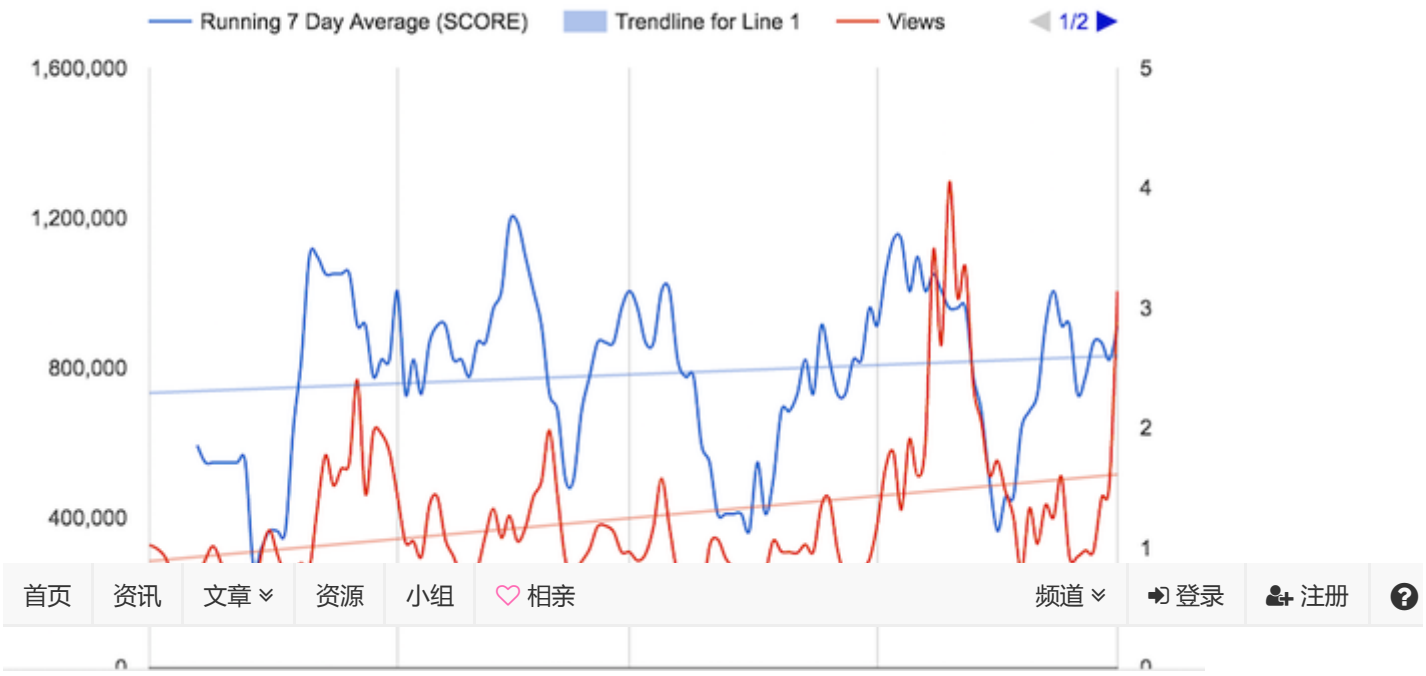


访问了下一个视频的订阅用户百分比与访问了前两个视频的订阅用户平均百分比之间的关系

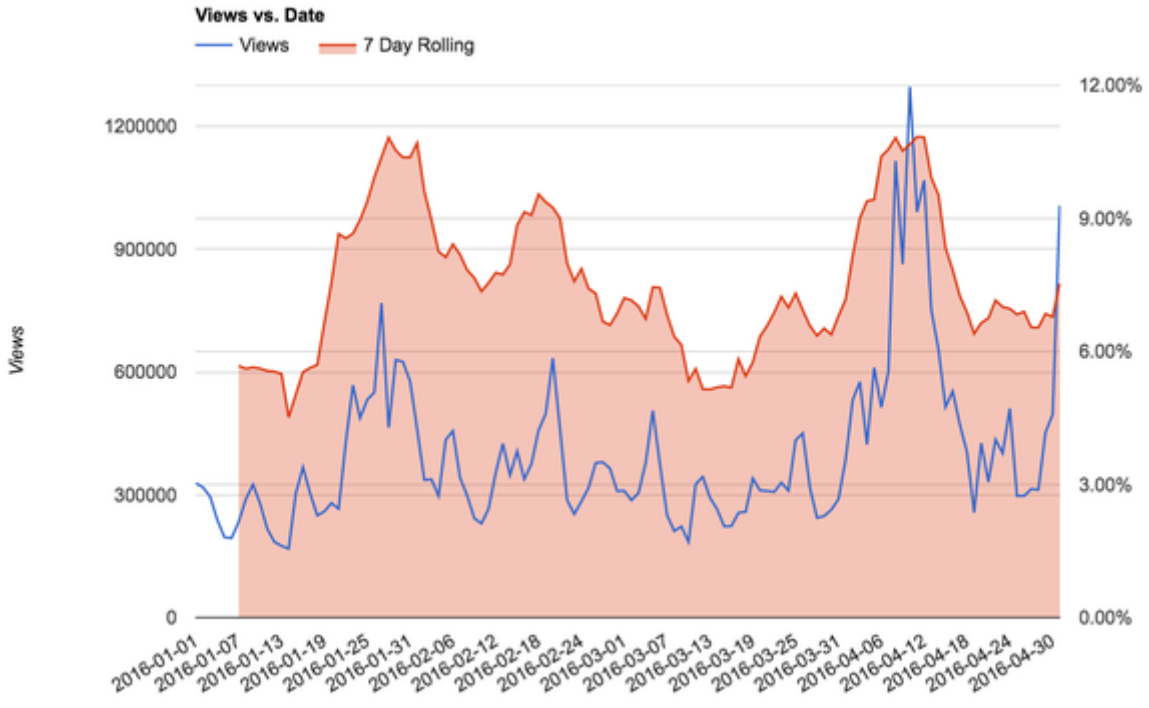
这个数据证实了Matthew Patrick的理论：如果某一个视频点击效果不好，那么你的下一次上传的视频，YouTube就不会给予太多权重让它被你的订阅用户看到。[2]

也可能是因为上一个视频表现糟糕，所以访问你的频道次数就会减少，自然地就导致更少的订阅用户以原生的方式访问到。不管到底“为什么”，结果反正就是酱紫。

另一个负速率对新上传视频的影响就是：有证据表明这还会伤害到你的整个视频库。下面的第一张图是视频上传48小时内就访问的订阅用户7天平均百分比（译者注：这7天上传了若干个视频，纪录每个视频上传后48小时就访问的订阅用户百分比，然后取这些百分比的平均值）与频道总访问次数（译者注：反应了整个视频库的效果）的关系。第二张图是某一天访问视频的总体订阅用户百分比与当日的总体访问次数之间的关系。



七天内的平均“48小时内访问视频的订阅用户百分比” 与 每日整个频道视频访问总数之间的关系



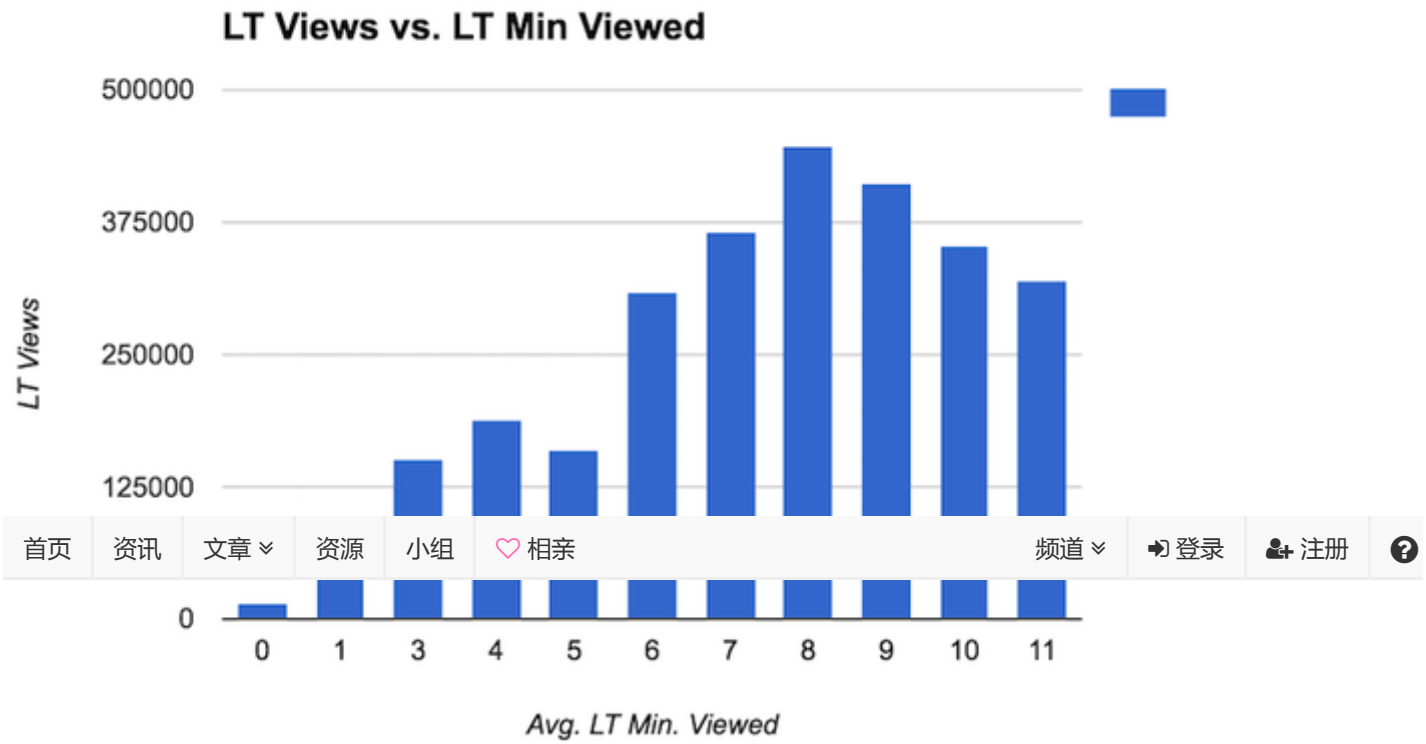
七天平均订阅用户访问人数 与 总体访问访问次数之间的关系

这些图标都说明一件事：一旦新上传视频和整个视频库的访问用户百分比走低，那么频道的总体访问次数也会走低。对于我们来说的启示是：YouTube算法更看重那些能够吸引到核心观众的频道，而惩罚那些不能吸引其核心观众的。

访问停留

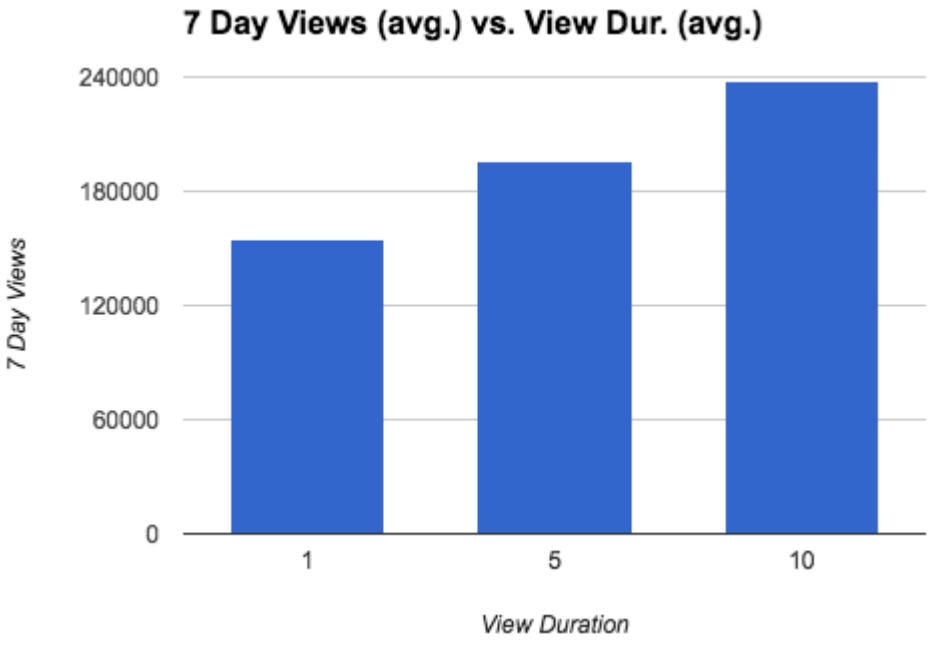
另一个算法非常看重的指标就是访问停留（View Duration）。

访问停留就是用户会花多长时间停留在单个视频页面。这个变量的权重很高，我们的数据中能看到一个明显的引爆点。Frederator其中一个频道，前30天内，平均访问时长8分钟的视频，比平均5分钟的要多350%的访问量。下图表明，Frederator的一个频道的视频访问量，与平均访问停留时长的关系。



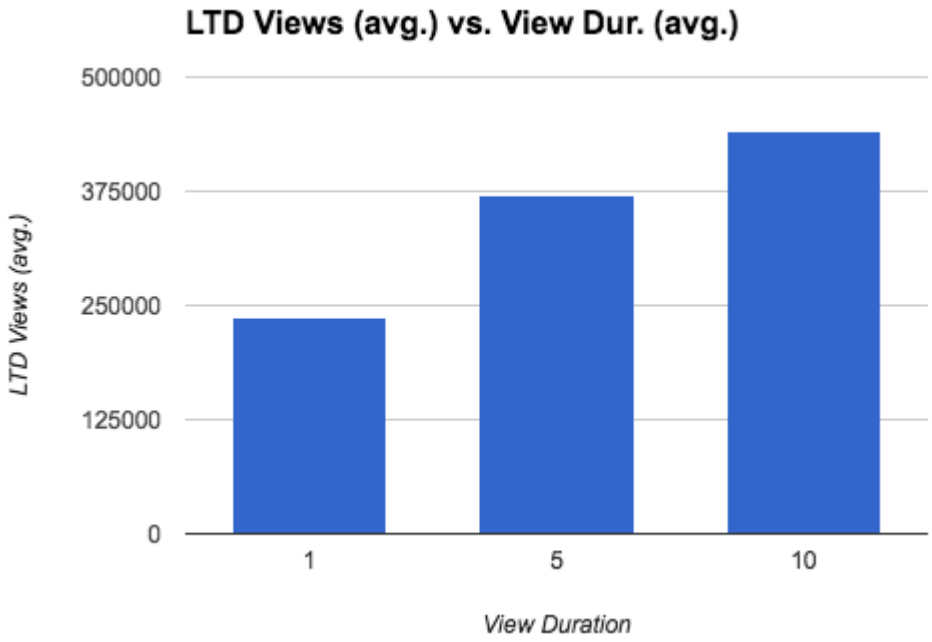
整个生命周期内，平均访问时长和平均访问量的关系 注意，这里没考虑访问时长在八分钟之上的数据。

我们还发现，访问停留时长越长，视频表现越好。下面这张图是七天内访问停留时长少于5分钟的视频（1），介于五分钟到十分钟的（5）， 十分钟以上的（10）分别与访问量的关系。



七天内平均访问量与平均访问停留时长的关系

下面这张图也是一个意思，不过从7天拉长到整个生命周期内了。



整个生命周期内平均访问量与平均访问停留时长的关系

基于这些发现，我们可以得出一个简单的结论：发布长视频可以提高访问效果。Frederator有一个

发一些炒剩饭的旧视频。除此之外，70分钟的视频和其他版本的视频有相同的平均访问停留时长。

于是，我们建议公司每周就只上传70分钟长度的视频就好了。就用了这个策略，频道日均访问量增长了50万，而过去6周里我们上传的视频个数却减少了75%。好了好了，我知道你受刺激了，不要崇拜哥。

会话开始，会话时长，会话结束

能做这篇研究，全都得益于我之前的一篇文章：《观看时长是个什么鬼》（WTF is WatchTime?）[1]

快速回顾一下，会话开始（Session Starts）就是指用户有多少次是从你的视频开始访问YouTube的。这其实说明了订阅用户能在前72小时访问你是多么重要。订阅用户是在视频发布后最早能看到的你人，他们也是最可能点击你频道图标的人，因为他们已经熟悉你的品牌了。

会话时长（Session Duration）就是你的内容让用户在YouTube平台上逗留了多久，他们访问你的视频，以及访问之后都算是在平台上逗留。除了用户平均访问时长（Average View Duration ）和独立访问数（ Unique Views），也没有更好的数据了。

会话结束（Session Ends）衡量用户是不是经常在看完你的视频后就离开了YouTube平台。这是算法利用的一个负面指标，但是我们根本拿不到数据。

一则算法理论

YouTube的算法设计时关注的是频道效果而不是单个视频效果。但是它要利用单个视频来提高频道效果。

算法结合了单个视频的特定数据和频道的聚合数据来决定推荐哪个视频。最终目标仍然是为频道聚拢其目标观众。

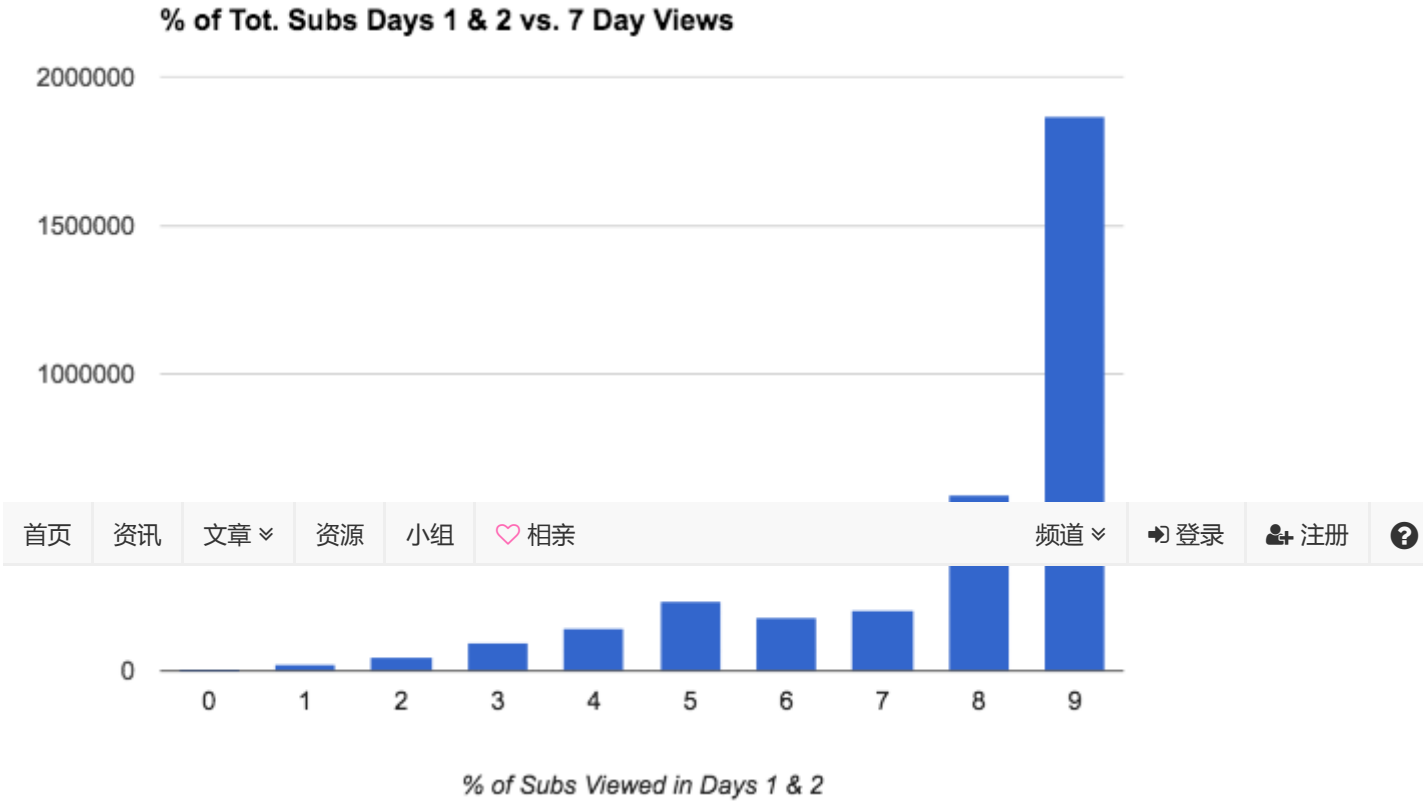
YouTube这么做是因为：

1. 让用户常常回访YouTube平台

2. 让用户在平台停留越久越好

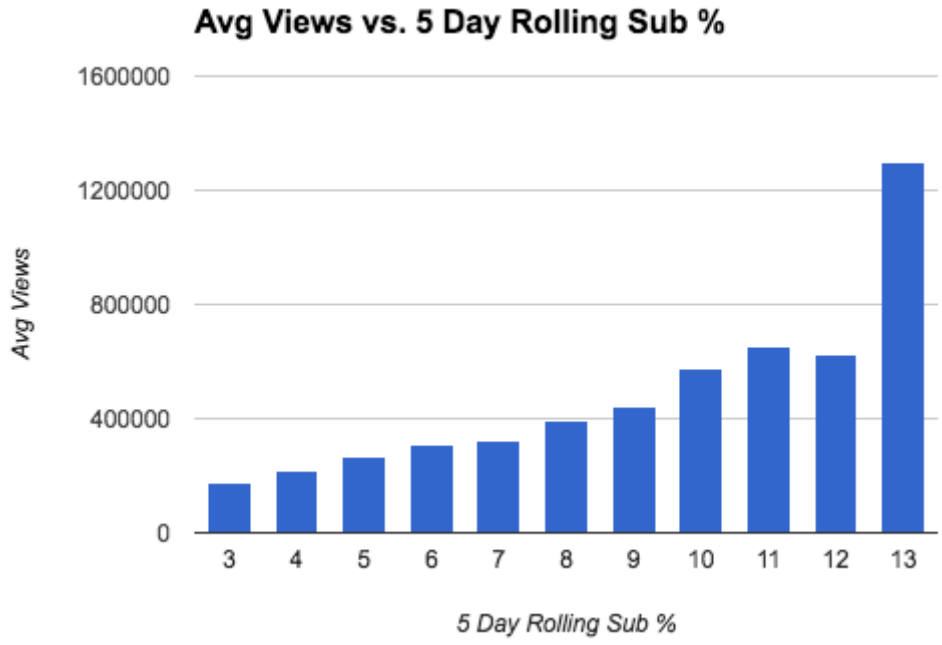
下面有三张图表来证明这则理论是成立的。

第一张图是48小时内访问的订阅者比例与7天内总访问量之间的关系。这张图说明，如果开始有大量用户从你的视频开始的平台会话，那么你的视频就会获得很大的访问量。到达一个阈值之后，就会呈指数级增长。



7日内总访问量与48小时内访问的订阅用户百分比

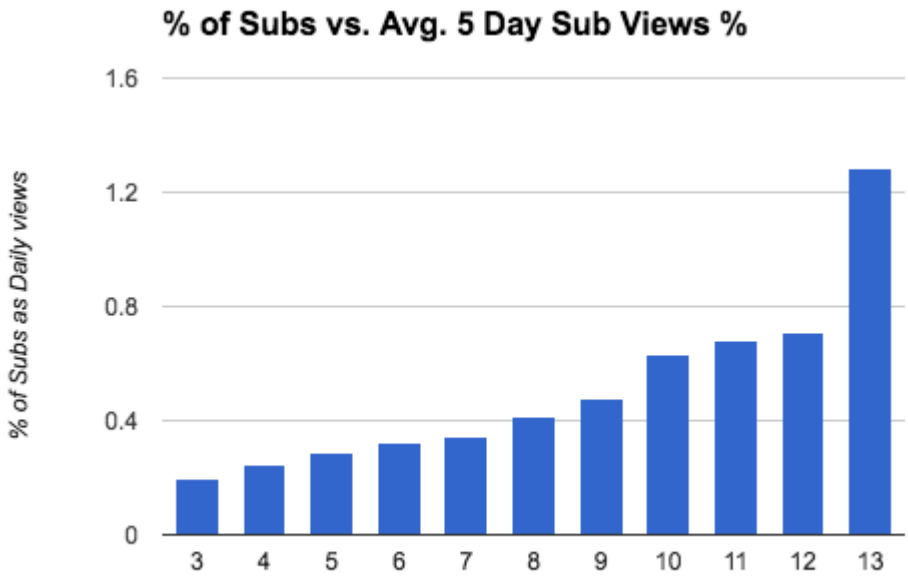
第二个图是频道内日均访问量与5日内访问的订阅用户百分比的关系。



日均访问量与5日内访问的订阅用户百分比的关系

这意味着如果能一直让大量用户从你开始访问YouTube（近5天内平均来看），那么算法就会将用户每日访问向你整个频道视频库倾斜。

最后一幅图是日均访问的订阅用户百分比与5天内访问的订阅用户百分比之间的关系。



日均访问的订阅用户百分比与5日内访问的订阅用户百分比之间的关系

我们相信这一切都表明，频道效果的连贯性与访问量之间存在相关性，访问量又表现在订阅用户访问百分比，YouTube就会因此把流量倾斜给你。

假如说你有一个游戏频道，10万个订阅用户，你每天上传6个视频，每个视频有5%的订阅用户访问。你的每个视频的平均访问订阅用户会稳定在区区5%。这意味你会每天产生30%的订阅用户访问次数（3万/天，60万/月）。现在假设你有1百万订阅用户，那么每日访问次数在30万，每月在600万。

我们认为这一段数学运算是不会骗人的。这意味YouTube在根据一些指标选择一些频道进行推荐，然后只要算法帮这个频道提高访问量。

但，壮士请留步，以上还仅仅是理论上的分析！

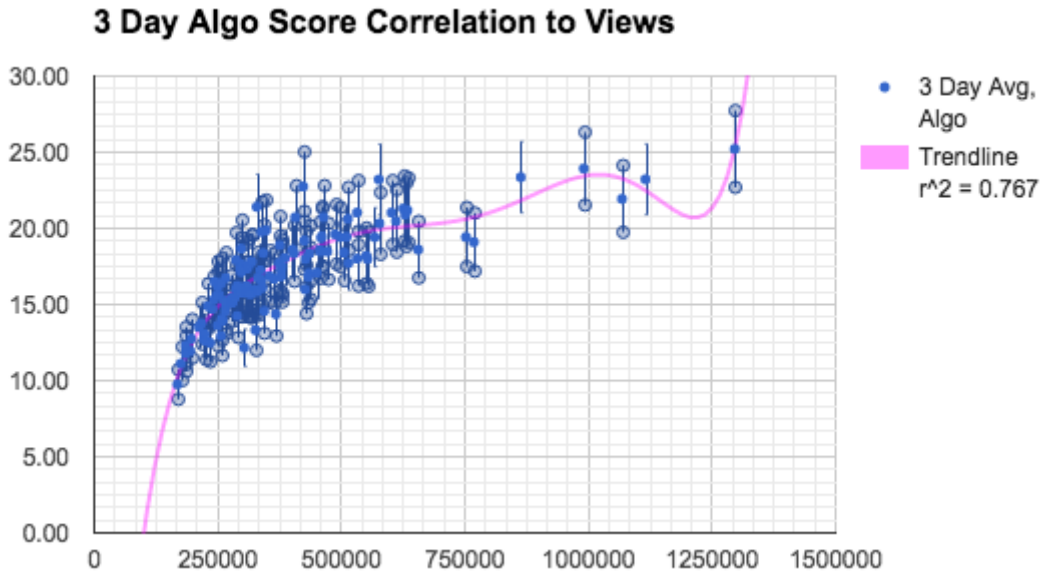
一种打分算法

这里我们打算破解YouTube的算法，然后重建一个。用了15个信号量，以及我们估计的权重，来重新构建打分算法。信号量列举如下：

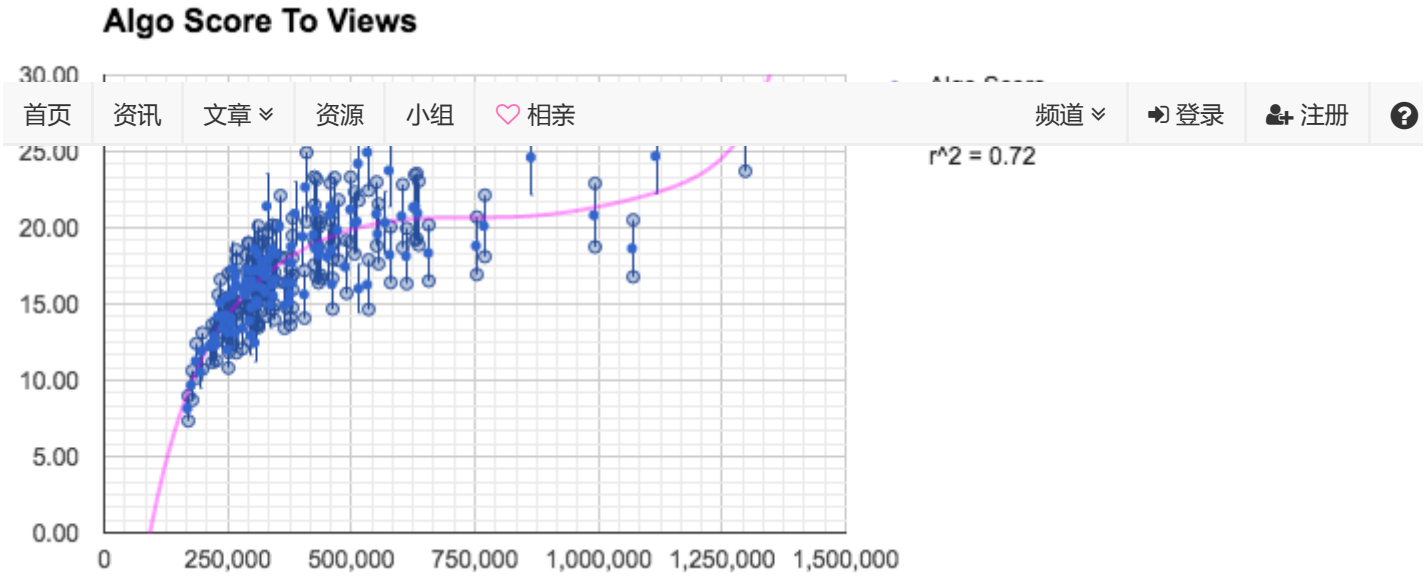
Session Start Score
Session Dur Score
View dur
Consistency Score
Engagement Rate
Publish Boost
Relevancy
Rolling Relevancy
Rolling Subs 5 Day
7 Day Avg. View Dur Score

用来开发打分算法的信号量 / 因素

下面这些图是这些信号量实际产生的效果。

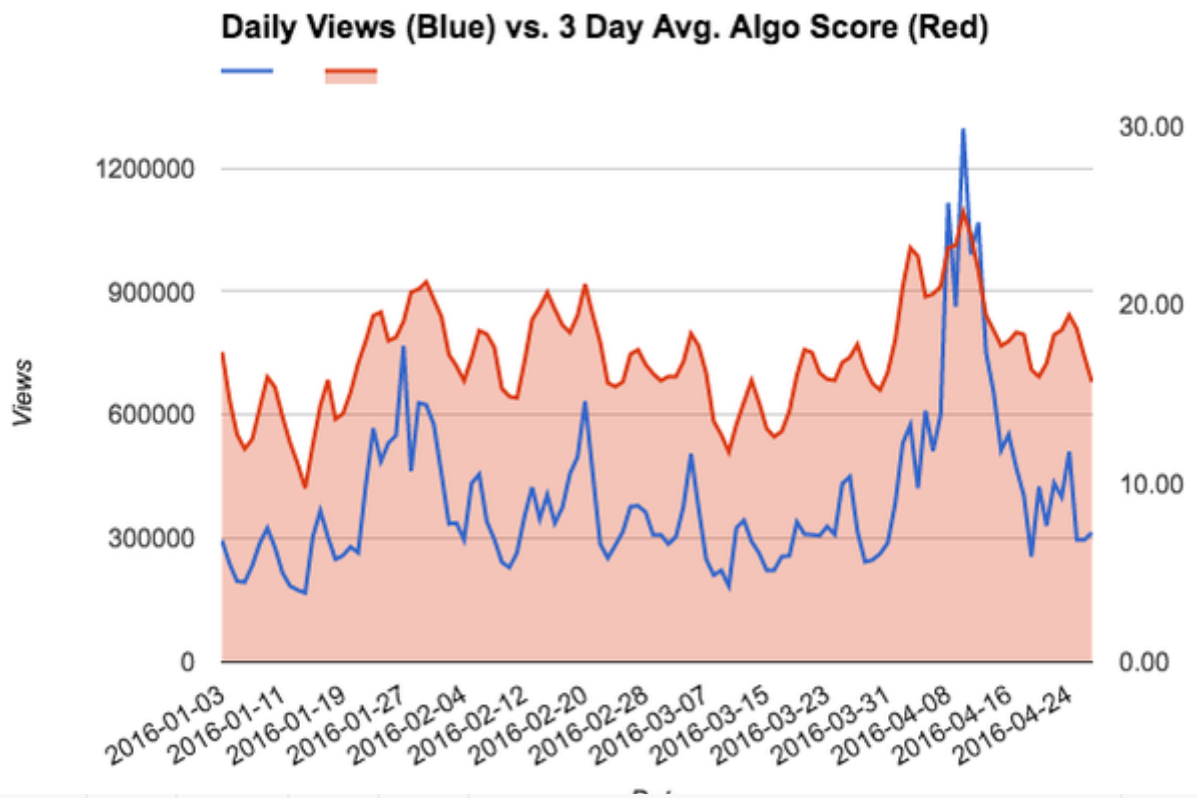


三天的算法平均分与访问量的相关趋势



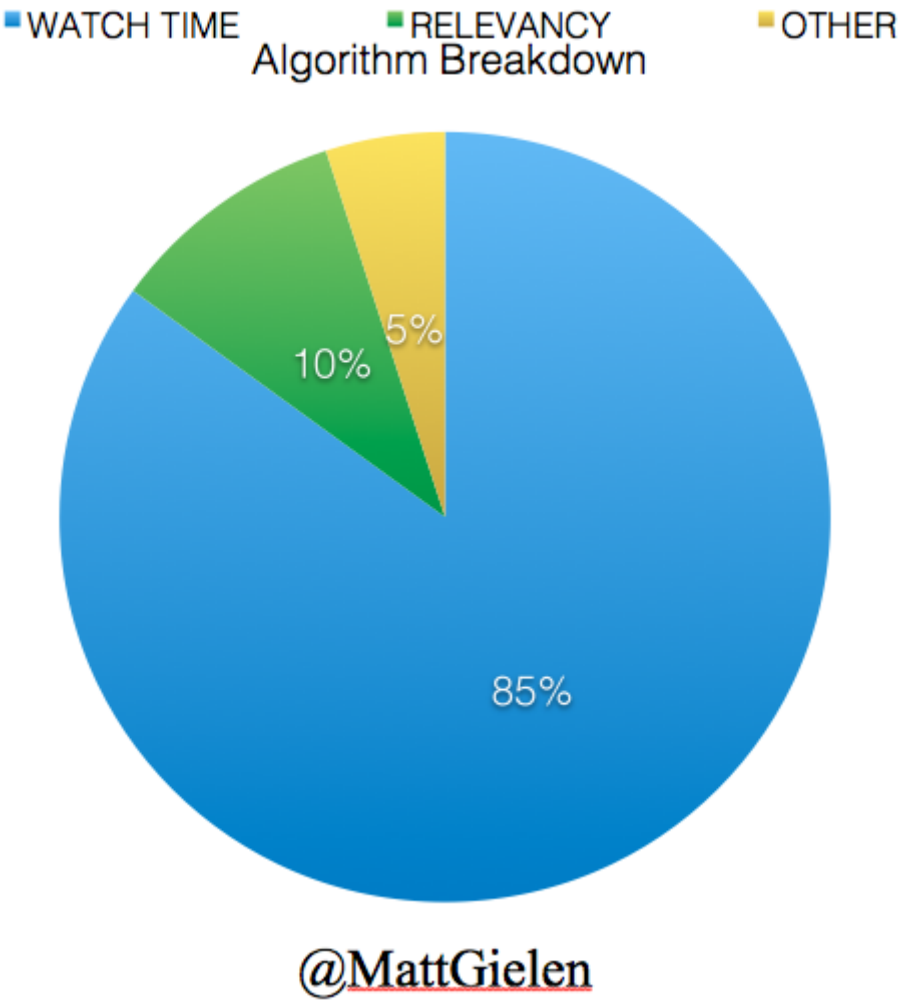
算法打分与访问量的相关性趋势

下面这张图更详细一些。

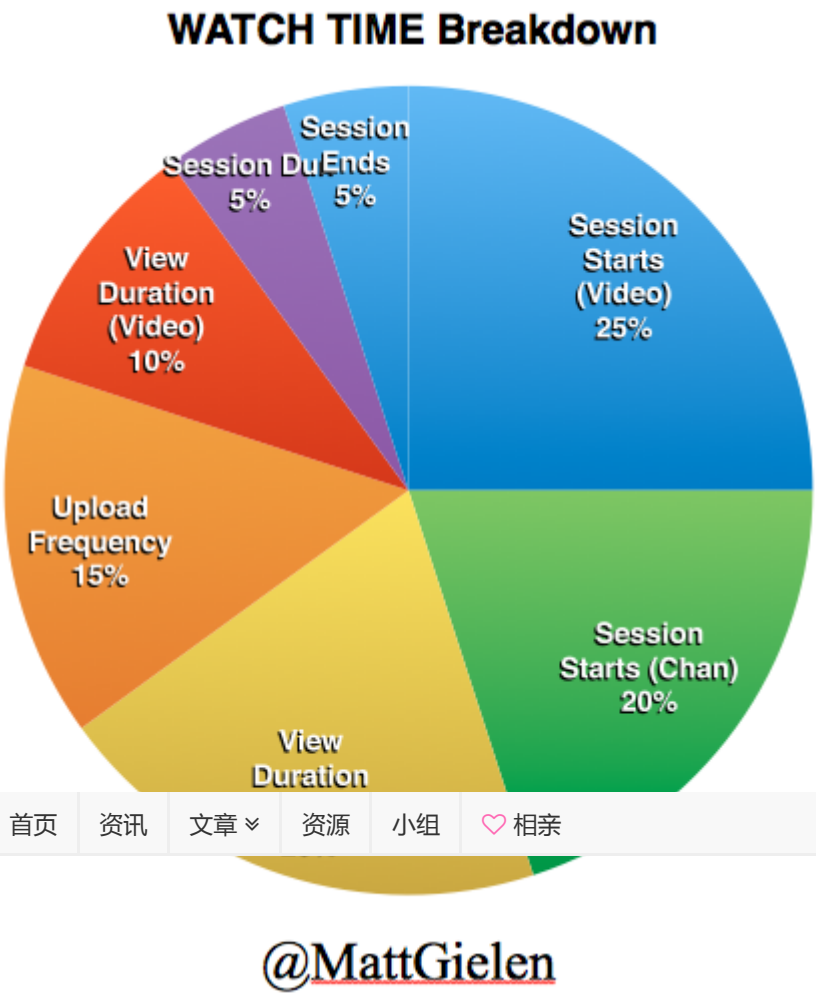


三天的算法打分均值与每日访问量

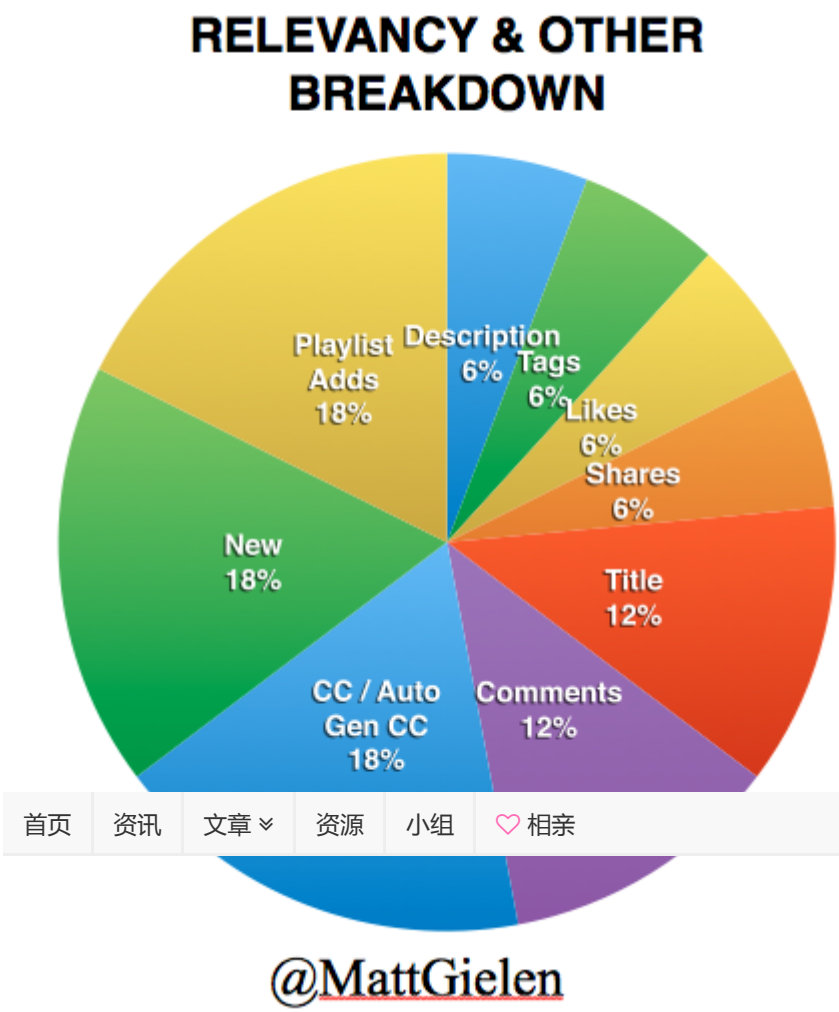
知道你还是很好奇，那下面就揭晓我们模拟出来的各种权重：



各种算法的权重分布模拟



观看时长优化算法的各信号量权重分布模拟



相关推荐及其他算法的各信号量权重分布

然而但是but，我们也没有其他数据了，所以我们也不敢肯定在计算相关性时该用哪种回归方式，也只敢说大多数信号和算法之间很相关，而已。也正因为如此，我们对YouTube算法一直热情不减。

对YouTube算法的看法

根据我们的数据，至少可以得到6个粗浅结论：

1. YouTube用算法决定了我们的视频和频道能得到多少访问量。

2. 成功的频道都是专注在特定类型的内容或创意上。

3. 频道自己一旦明确了哪种类型的内容成功之后，就不要再摇摆了。

4. 内容制作者光靠钱在YouTube平台上绝无可能成功，因此土豪型的制作者不太会全身心拥抱YouTube。

5. 个性化的节目/频道会一直是YouTube上面占统治地位的内容类型，因为这就是人们要找的“特定类型的内容”。

6. 新建的频道，如果不能在YouTube站外导流进去的话，相当长时间内增长都会比较困难。

前面说到，YouTube更注重于提高频道的访问效果，这个观点只是我们推测得到的。频道能够上传很多视频，从而获得和留住大量的目标观众。如果你想在YouTube上成功，我们能给的建议就是：瞄准一个非常垂直的兴趣类型，然后持续去制作10分钟以上的视频，一定得是你选定的这个兴趣类型的视频。

我这里是私人博客，需要提醒一下，YouTube可是储备了大量的算法弹药啊，也希望他们不把本文视为对算法的负面消息。通过这篇研究，我更加感谢YouTube及其算法工程师们，有预见性地设计了这些算法。毕竟，他们还是想努力让这个世界上的十亿用户能在一个月内不重样地观看视频。如果你能停下来回头再整体上审视一下这一切，你会惊叹于YouTube算法设计如此优雅，在实现商业目标上和保护平台健康发展上做得难以置信的好。为他们点32个赞！

作者简介：

Matt Gielen是Frederator Networks的前副总裁， 主管编程和观众开发。

Matt所管的团队是世界上最大的动画制作网络公司，Frederator网络频道。

他还带领团队制作和编程了Frederator Networks自己的YouTube运营频道：Channel Frederator，The Leaderboard，Cinematica。

你还可以在twitter上关注他@mattgielen。

译后记：

最初看到这篇文章是@fengyoung 在Facebook上分享的，觉得题目很有意思就看了一遍，看完后感觉很有启发，遂决定翻译一下让更多人看到。

1. 从YouTube平台的算法设计人员角度，设计繁多的推荐算法，是为了提高频道的观看时长，而提高频道的观看时长又是为了让用户能够经常访问平台。这是一种双赢的思维，说白了：谁能帮平台留住用户，平台就重点扶持他。

2. 文章得出结论，要做垂直内容才能在YouTube上活下去。平台上内容越多样，平台越健康，这是毋庸置疑的，尽管我赞同这个结论，但是我没有在本文中看到作者是如何得到这个结论的。这一点就是YouTube和国内视频平台最大的差别，国内的视频平台严重趋同，花高价购买独家版权似乎是国内视频平台的唯一出路，也是一个妖魔化的出路，反观YouTube，他们利用算法驱使了各个频道专耕某一个垂直内容，然后把最适合的用户给你匹配上，这才是更宏大的一盘内容棋。

3. 本文作者给我们了一个启示，算法并不是黑盒子，是可以hack的，尽管这个也只能hack到冰山一角，但是也比我们盲目地运营要明亮很多了。作者的研究方式，首先是明确了一个平台的算法目标是什么，YouTube是watch time，那么就去观察这个目标和哪些指标有关，进一步看到每个指标又能怎么提高。

[1] http://www.tubefilter.com/2016/05/12/youtube-watch-time-metric-algorithm-statistics/

[2] https://www.youtube.com/watch?v=HLJQ0gFHM8s

- 大家都是怎么管理时间的？
- 迅雷2016研发工程师笔试题 • 1
- 这一年都做了什么，回顾我的2016 • 3
- 关于apache用户认证配置的问题
- Java程序员两年，感觉自己还是菜鸟，往前端转如何？ • 3
- 在厦门的程序猿，刚毕业几年是什么状态？ • 7

登录后评论 新用户注册

直接登录

最新评论

浪子哥 （ 1 ） 11/29

youtube这些网站相关算法，看不懂。。

- 本周热门文章
- 本月热门文章
- 热门标签

0 单点登录原理与简单实现

1 从 Hello World 说程序运行机制

2 程序员会喜欢的 12 款键盘

3 Git 王者超神之路

4 SVN、GIT日常看我就够了

5 程序员，你为什么值这么多钱？

6 FB前工程主管：发布代码的正确方式

7 Linux 内核数据结构：位图（Bitma...

8 程序运行时的内存空间分布

9 MySQL误操作后如何快速恢复数据

业界热点资讯 更多 »

[2017 年该学习的编程语言、框架和工具]
22 小时前 • 37

[谷歌员工吐槽：人生不如意十之八九，在谷歌工作也不容易]
1 天前 • 32 • 1

[每天都离不开的 WiFi 网络，却没几个人了解它]
1 天前 • 10

[谷歌物联网操作系统 Android Things 揭开面纱]
1 天前 • 6

[Google开源了“高维数据可视化工具”Embedding Projector...]
1 天前 • 6

精选工具资源 **更多资源 »**

[PyMC：马尔科夫链蒙特卡洛采样工具]
[科学计算与分析]

[statsmodels：统计建模和计量经济学]
[科学计算与分析]

[Clang Static Analyzer：源代码分析工具（C、C++和Obje...](#)
[静态代码分析](#)

[Pylearn2：一个基于Theano的机器学习库](#)
[机器学习](#) • 1

[Electron：用Html、CSS和JavaScript构建跨平台的客...](#)
[MVC框架和库](#)

关于伯乐在线博客

在这个信息爆炸的时代，人们已然被大量、快速并且简短的信息所包围。然而，我们相信：过多“快餐”式的阅读只会令人“虚胖”，缺乏实质的内涵。伯乐在线内容团队正试图以我们微薄的力量，把优秀的原创文章和译文分享给读者，为“快餐”添加一些“营养”元素。

快速链接
网站使用指南 »
问题反馈与求助 »
加入我们 »
网站积分规则 »
网站声望规则 »

关注我们

新浪微博：@伯乐在线官方微博
RSS：订阅地址
推荐微信号

合作联系
Email：bd@Jobbole.com
QQ： 2302462408 （加好友请注明来意）

更多频道

小组 - 好的话题、有启发的回复、值得信赖的圈子
头条 - 分享和发现有价值的内容与观点
相亲 - 为IT单身男女服务的征婚传播平台
资源 - 优秀的工具资源导航
翻译 - 翻译传播优秀的外文文章
文章 - 国内外的精选文章
设计 - UI,网页，交互和用户体验
iOS - 专注iOS技术分享
安卓 - 专注Android技术分享
前端 - JavaScript, HTML5, CSS
Java - 专注Java技术分享
Python - 专注Python技术分享

 文章 小组 相亲 加入我们 反馈

首页 资讯 文章 资源 小组 相亲 频道 登录 注册