USvideos指标数据统计&榜单报告

19媒调 郑妍

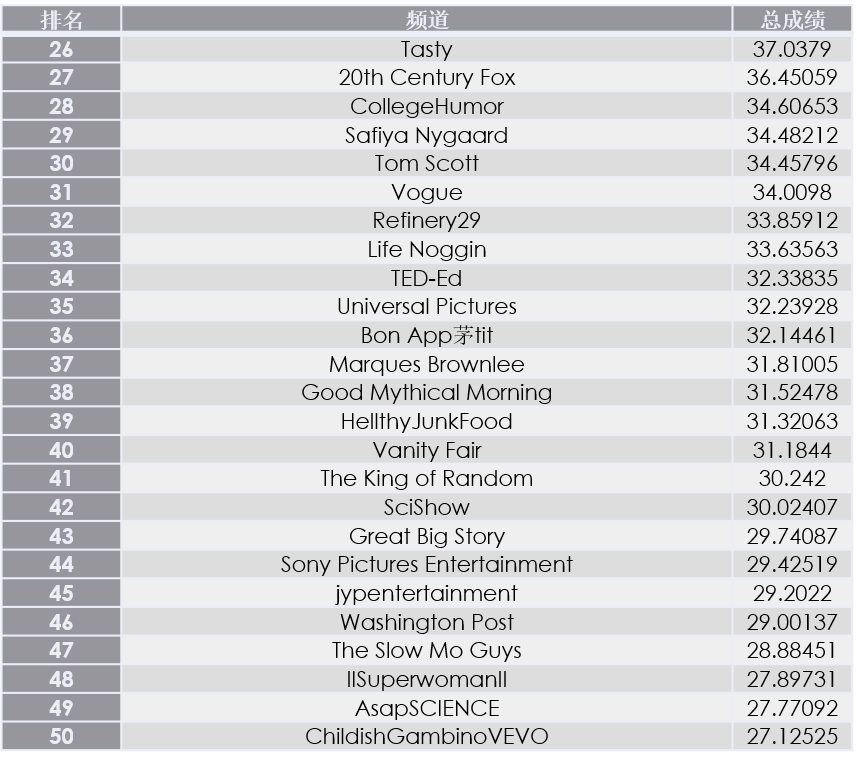
一、指标体系及说明

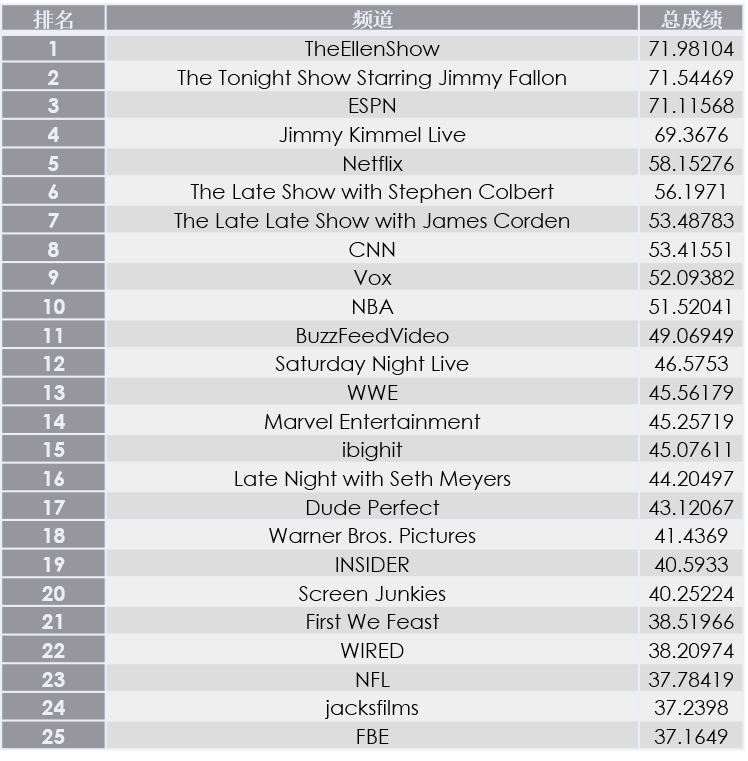
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **一级指标** | **权重** | **二级指标** | **权重** | **三级指标** | **权重** | **具体算法** |
| 频道信息 | 0.4 | - | | 上榜视频数 | - | 计数 |
| 视频信息 | 0.6 | 保有情况 | 0.4 | 是否可被观看 | - | video\_error\_or\_removed |
| 基本信息 | 0.25 | 至上榜天数（负） | 0.2 | -([trending\_date]-[publish\_time]) |
| 在榜天数 | 0.8 | trending\_date |
| 互动力 | 0.35 | 观看量 | 0.5 | views\*0.000001 |
| 喜欢占比 | 0.1 | [likes]/[views] |
| 不喜欢占比（负） | 0.15 | -[dislikes]/[views] |
| 评论占比 | 0.25 | [comment\_count]/[views] |

下面对指标进行几点说明。

由于采用了直接引用数据和运算后去比率等不同的数据，数量级不尽相同，所以对数量级较大的观看量进行了一定的缩小操作，使得数量级跟其余基本相同。同时，因为点赞、点踩和评论均基本与观看量成正相关，在此不列举具体数值作为指标而是直接取比例。

在进行了计算整理后，发现本指标体系仍存在些许问题。首先，尽管进行了相关操作，指标的数量级并不统一。此外，数量级相对较大的观看量其对应的权重较大，也较大影响了最后的结果。

 由于拿到数据的有限性，有些设想中的指标无法实现，在此对设想也进行一定的阐述。1.更新频率的统计及上榜视频的占比；2.视频第一次上榜当天的数据变化（尤其是对于较早发布的视频）；3.同时段内不同视频的质量差异；4.用户观看视频时长占视频总时长的比例。

二、最终榜单

这里展现的是榜单前50的信息。所有的数据清理、计算、整合工作全部由excel2016完成，所用到的功能主要为：分列、删除重复值、分类汇总、排序和筛选等。

三、其余发现

由于任何依据榜单得出的结论都依托于指标，而现在的指标还不是非常完善，所以其余结论的得出会倾向于在指标设计中发现的一些结论。

1.标签数并不适合表征视频观看量

在最开始的指标设定中，考虑到标签数可能会影响到浏览量，即标签数越多的视频浏览量越大。为了验证这一假设，取每个视频的浏览量与标签数生成如下的散点图：

可以看到，并不是标签数越多浏览量就一定越大，所以说，在这里，标签数并不适合作为指标出现。同样对类别进行相关处理，也可发现其与热度关联不大，故也不作为指标参考。

2.并不是发布越早的视频越容易缺失（video\_error\_or\_removed为TRUE）甚至情况可能相反

对于视频保有的情况，猜测越早发布的视频越容易丢失，故将保有情况与发布日期的关系生成散点图如下：

可以发现，先前的假设被否定——为数不多的已无法观看的视频甚至发布日期更接近出榜的日期。