# 互联网视频流量热点的分析与预测

摘要 本文通过网络爬虫的技术爬取了短视频数据统计平台卡思数据平台的数据，然后分析了全网各类型短视频的发展情况和网红经济对市场的影响，最后结合当前比较火的机器学习算法对数据中的规律进行一些探索分析。

Abstract In this paper, we use the technology of web crawler to crawl the data of CASS data platform, which is a short video data statistics platform. Then we analyze the development of all kinds of short video in the whole network and the influence of netred economy on the market. Finally, we combine the current popular machine learning algorithm to explore and analyze the rules in the data.

关键词 短视频；卡思数据；网红；B站；线性回归；决策树；聚类

## 获取数据

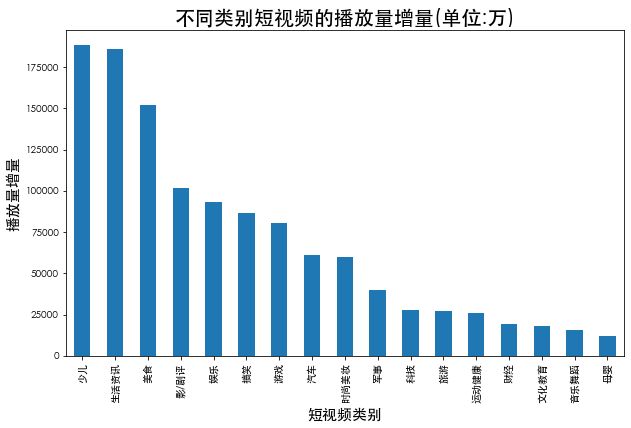
为了研究短视频的发展趋势，我们需要获取短视频的相关统计数据，我们主要通过网络爬虫的方式获取数据。网络爬虫是一个自动提取网页的程序，它为[搜索引擎](https://baike.baidu.com/item/%E6%90%9C%E7%B4%A2%E5%BC%95%E6%93%8E)从万维网上下载网页，是搜索引擎的重要组成。传统爬虫从一个或若干初始网页的URL开始，获得初始网页上的URL，在抓取网页的过程中，不断从当前页面上抽取新的URL放入队列,直到满足系统的一定停止条件。

我们主要通过python爬虫的方式来获取数据，数据来源主要为卡思数据平台。卡思数据平台是国内权威的视频全网大数据开放平台，依托专业的数据挖掘与分析能力，构建多种维度的数据算法模型，以卡思指数来体现红人、节目或创作团队的商业价值全貌。卡思数据平台提供全方位的数据查询、趋势分析、舆情分析、用户画像、视频监测、数据研究等服务，为内容创作团队在内容创作和用户运营方面提供数据支持，为广告主的广告投放提供数据参考，为内容投资提供全面客观的价值评估。我们主要爬取了网站上红人榜、PGC榜、网综榜和网剧榜的数据用于分析。除了卡思数据，我们还爬取了同花顺网页的部分股指数据用于论文实证。

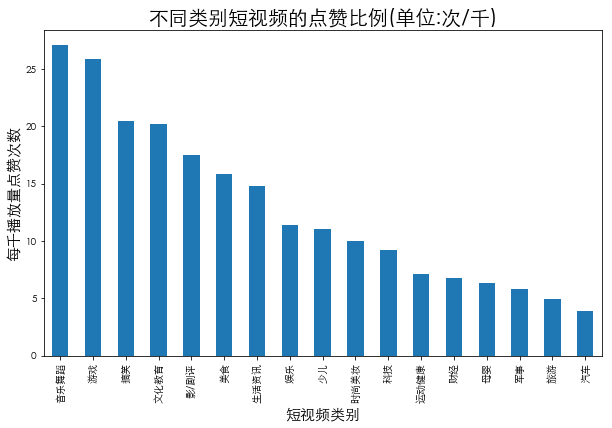
## 数据分析

### 短视频类别分析

为了查看各个类别短视频的客户群体大小，我们可以通过盘点全网短视频类别的统计期播放量来获知相应信息，如下图所示：

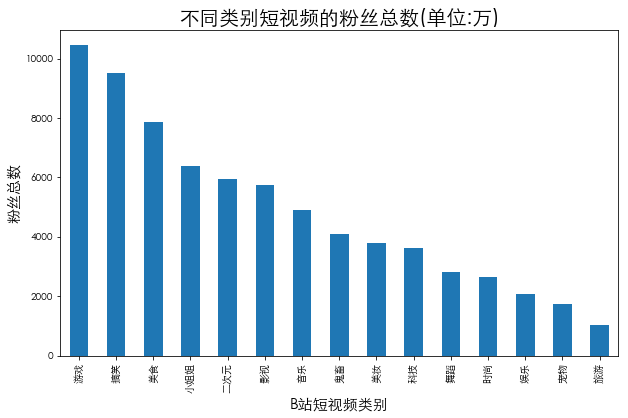


上图绘制出了全网短视频播放量前15位的类别，可以看出少儿、生活咨询、美食、影/剧评和娱乐等类别为播放量靠前的类别，运动健康、财经、文化教育、音乐舞蹈和母婴是相对考后的类别。此外，我们还绘制出了不同类别的点赞比例如下图：



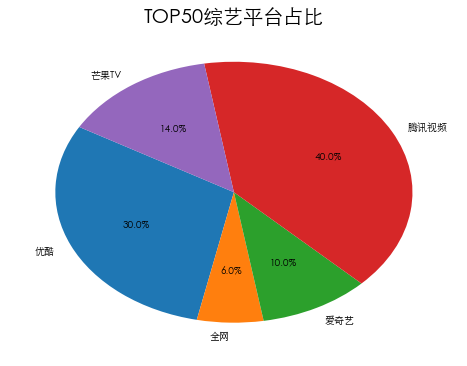
点赞比例某种程度上揭示了客户群体对该类别视频的客户粘性，结合上图可以发现音乐舞蹈、游戏和搞笑等三个类别的短视频的客户点赞比例是最高的。

此外，我们还专门对哔哩哔哩平台进行了分析。哔哩哔哩（英文名称：bilibili，简称B站）现为中国年轻世代高度聚集的文化社区和视频平台  ，该网站于2009年6月26日创建，被粉丝们亲切的称为“B站”。



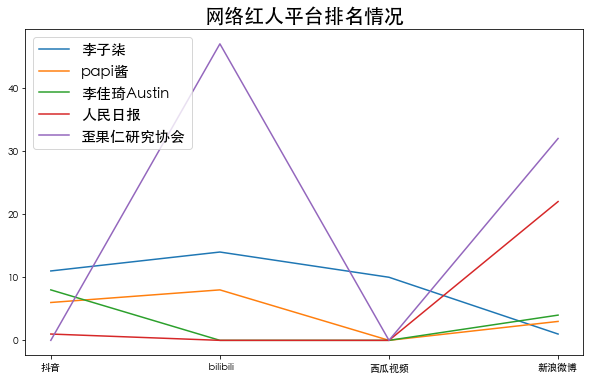
上图绘制除了B站不同类别短视频的粉丝总数，从上图可以看出B站的关注者分布情况，可以看出，B站的主要粉丝基本都聚集在游戏、搞笑、美食、小姐姐和二次元等短视频平台。

### 网络综艺分析



上图绘制出了TOP50的网络综艺占比，可以发现腾讯视频的占比高达40%，优酷的占比高达30%，芒果TV占比为14%。

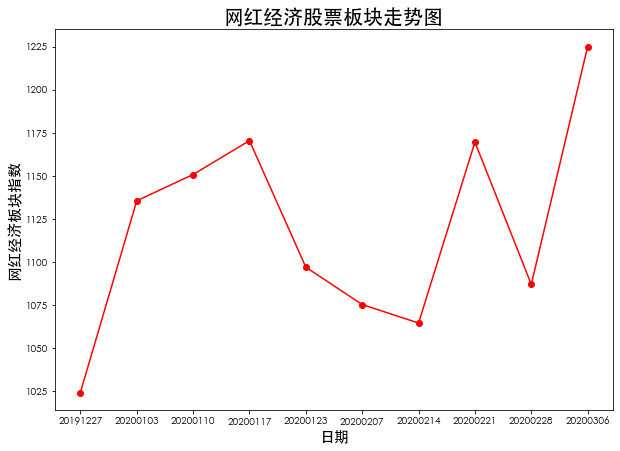
### 网络红人分析



卡思数据会综合粉丝数、播放量和活跃情况等多个维度的数据对不同的网红在不同平台上进行排名，我们选择了5个网红账号：李子柒、papi酱、李佳琦Austin、人民日报和歪果仁研究协会进行分析，分别分析他们在四个平台（抖音、blibli、西瓜视频和新浪微博）的排名情况。可以发现，不同的平台的客户群体存在较大的差异，因此不同的类别的网络红人在不同平台上的受关注程度往往有较大的差别，最终体现在排名上有较大的差异。

### 网红经济分析

网红经济是19年底流行起来的词汇，百度百科给出的定义为网红经济是以年轻貌美的时尚达人为形象代表，以红人的品味和眼光为主导，进行选款和[视觉推广](https://baike.baidu.com/item/%E8%A7%86%E8%A7%89%E6%8E%A8%E5%B9%BF)，在社交媒体上聚集[人气](https://baike.baidu.com/item/%E4%BA%BA%E6%B0%94/33565)，依托庞大的粉丝群体进行定向营销，从而将[粉丝](https://baike.baidu.com/item/%E7%B2%89%E4%B8%9D/6368045)转化为购买力的一个过程。



上图绘制出了同花顺网红经济板块的股票指数，可以发现自19年底同花顺添加该板块指数之后，板块股指基本一直在震荡上行，从侧面反馈出网红经济相关的词汇在市场的热度是比较高的，

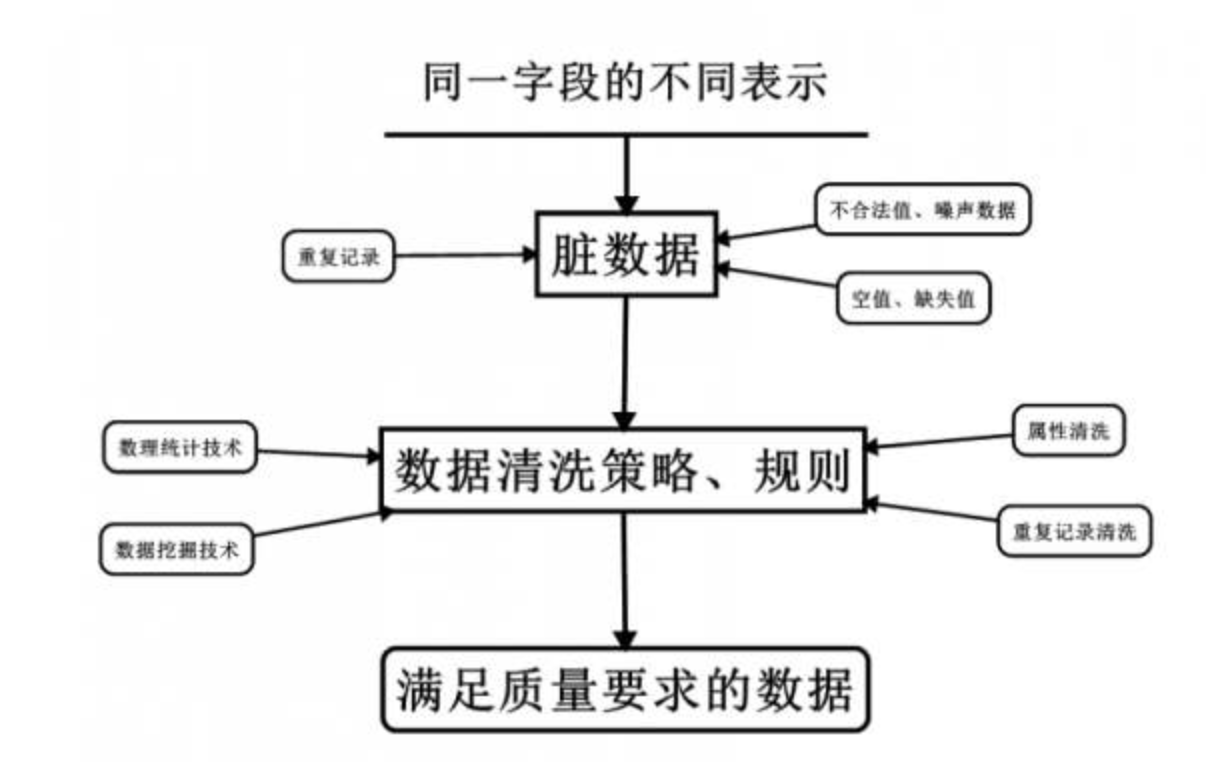
## 算法分析

### 数据建模

数据建模是算法建模前的数据处理阶段，主要过程为数据清洗，数据清洗是检测和去除数据集中的噪声数据和无关数据，处理遗漏数据，去除空白数据域和知识背景下的白噪声。数据清洗分为有监督清洗和无监督清洗两类：

1. 有监督清洗：在对应领域专家的指导下，收集分析数据，手工去除明显的噪声数据和重复记录，填补缺值数据等清洗动作
2. 无监督清洗：根据一定的业务规则，预先定义好数据清洗算法，由计算机自动执行算法，对数据集进行清洗，然后产生清洗报告

在实际项目中，完全有监督的清洗估计不太可能，因此，一般是先进行无监督清洗并产生相应的清洗报告，然后再让专家根据清晰报告对清洗的结果进行人工整理。在数据清理前，我们需要分析数据的特点并定义数据清洗规则，清洗结束后为了保证清洗的质量，我们还需要验证清洗结果。数据的分析是根据相关的业务知识和相应的技术，如统计学，数据挖掘的方法，分析出数据源中数据的特点，为定义数据清洗规则奠定基础。常用的清洗规则主要包括：空值的检查和处理；非法值的检测和处理；不一致数据的检测和处理；相似重复记录的检测和处理。执行数据清洗规则时需要检查拼写错误，去掉重复的（duplicate）记录，补上不完全的（incomplete）记录，解决不一致的（inconsistent）记录，用测试查询来验证数据，最后需要生成数据清晰报告。在清洗结果验证中，需要对定义的清洗转换规则的正确性和效率进行验证和评估，当不满足清洗要求时要对清洗规则或系统参数进行调整和改进，数据清洗过程中往往需要多次迭代的进行分析，设计和验证。



上述图片大致描述了数据处理的整体环节，首先进行脏数据的处理，主要包括去除重复记录，剔除不合法数值和噪声数据，和处理空值缺失值，然后进行数据清洗，数据清洗的部分主要是处理数据中的量纲，由于播放量存在万、百万、亿等多个单位，因此需要处理成同一单位，还会进行一些属性清洗的操作。

### 算法原理

本项目主要尝试了三种算法，线性回归算法，决策树算法和kmeans聚类算法，下面分别介绍下三种算法的基本原理：

1. 线性回归

在统计学中，线性回归(Linear Regression)是利用称为线性回归方程的最小平方函数对一个或多个[自变量](https://baike.baidu.com/item/%E8%87%AA%E5%8F%98%E9%87%8F)和[因变量](https://baike.baidu.com/item/%E5%9B%A0%E5%8F%98%E9%87%8F)之间关系进行建模的一种[回归分析](https://baike.baidu.com/item/%E5%9B%9E%E5%BD%92%E5%88%86%E6%9E%90)。这种函数是一个或多个称为回归系数的模型参数的线性组合。只有一个自变量的情况称为简单回归,大于一个自变量情况的叫做多元回归。（这反过来又应当由多个相关的因变量预测的多元线性回归区别，而不是一个单一的标量变量。）

在线性回归中，数据使用线性预测函数来建模，并且未知的模型参数也是通过数据来估计。这些模型被叫做线性模型。最常用的线性回归建模是给定X值的y的条件均值是X的仿射函数。不太一般的情况，线性回归模型可以是一个中位数或一些其他的给定X的条件下y的条件分布的分位数作为X的线性函数表示。像所有形式的回归分析一样，线性回归也把焦点放在给定X值的y的条件概率分布，而不是X和y的联合概率分布（多元分析领域）。

线性回归是回归分析中第一种经过严格研究并在实际应用中广泛使用的类型。这是因为线性依赖于其未知参数的模型比非线性依赖于其位置参数的模型更容易拟合，而且产生的估计的统计特性也更容易确定。

线性回归模型经常用最小二乘逼近来拟合，但他们也可能用别的方法来拟合，比如用最小化“拟合缺陷”在一些其他规范里（比如最小绝对误差回归），或者在桥回归中最小化最小二乘损失函数的惩罚.相反,最小二乘逼近可以用来拟合那些非线性的模型.因此，尽管“最小二乘法”和“线性模型”是紧密相连的，但他们是不能划等号的。

1. 决策树

决策树(Decision Tree）是在已知各种情况发生概率的[基础](https://baike.baidu.com/item/%E5%9F%BA%E7%A1%80/32794)上，通过构成决策树来求取净现值的[期望](https://baike.baidu.com/item/%E6%9C%9F%E6%9C%9B/35704)值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。由于这种决策分支画成图形很像一棵树的枝干，故称决策树。在机器学习中，决策树是一个预测模型，他代表的是对象属性与对象值之间的一种映射关系。Entropy = 系统的凌乱程度，使用算法[ID3](https://baike.baidu.com/item/ID3), [C4.5](https://baike.baidu.com/item/C4.5)和C5.0生成树算法使用熵。这一度量是基于信息学理论中熵的概念。

决策树是一种树形结构，其中每个内部节点表示一个属性上的测试，每个分支代表一个测试输出，每个叶节点代表一种类别。

分类树（决策树）是一种十分常用的分类方法。他是一种监管学习，所谓监管学习就是给定一堆样本，每个样本都有一组属性和一个类别，这些类别是事先确定的，那么通过学习得到一个分类器，这个分类器能够对新出现的对象给出正确的分类。这样的机器学习就被称之为监督学习。

1. kmeans聚类

聚类是一个将数据集中在某些方面相似的数据成员进行分类组织的过程，聚类就是一种发现这种内在结构的技术，聚类技术经常被称为无监督学习。

k均值聚类是最著名的划分聚类算法，由于简洁和效率使得他成为所有聚类算法中最广泛使用的。给定一个数据点集合和需要的聚类数目k，k由用户指定，k均值算法根据某个距离函数反复把数据分入k个聚类中。

k均值聚类算法（k-means clustering algorithm）是一种迭代求解的[聚类分析](https://baike.baidu.com/item/%E8%81%9A%E7%B1%BB%E5%88%86%E6%9E%90/3450227)算法，其步骤是，预将数据分为K组，则随机选取K个对象作为初始的聚类中心，然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，误差平方和局部最小。

### 算法建模

算法建模的目标为在获取数据建模处理后的数据进行算法实现，输出算法结果和进行算法评估，本项目主要使用python进行实现。

算法评估时我们采用了交叉验证的方法，在使用训练集对参数进行训练的时候，经常会发现人们通常会将一整个训练集分为三个部分（比如mnist手写训练集）。一般分为：训练集（train\_set），评估集（valid\_set），测试集（test\_set）这三个部分。这其实是为了保证训练效果而特意设置的。其中测试集很好理解，其实就是完全不参与训练的数据，仅仅用来观测测试效果的数据。而训练集和评估集则牵涉到下面的知识了。

因为在实际的训练中，训练的结果对于训练集的拟合程度通常还是挺好的（初始条件敏感），但是对于训练集之外的数据的拟合程度通常就不那么令人满意了。因此我们通常并不会把所有的数据集都拿来训练，而是分出一部分来（这一部分不参加训练）对训练集生成的参数进行测试，相对客观的判断这些参数对训练集之外的数据的符合程度。这种思想就称为交叉验证。

交叉验证（Cross Validation），有的时候也称作循环估计（Rotation Estimation），是一种统计学上将数据样本切割成较小子集的实用方法，该理论是由Seymour Geisser提出的。

在给定的建模样本中，拿出大部分样本进行建模型，留小部分样本用刚建立的模型进行预报，并求这小部分样本的预报误差，记录它们的平方加和。这个过程一直进行，直到所有的样本都被预报了一次而且仅被预报一次。把每个样本的预报误差平方加和，称为PRESS(predicted Error Sum of Squares)。

### 算法结果

