# 行为作弊聚类算法

## 行为作弊聚类算法

### 项目背景

社区内很多笔记或个人存在请第三方个人或公司**有偿**地通过认为干预的方式提高笔记的点击率、转化率、互动量等行为，这些行为严重地影响了平台用户的使用体验、统计数据的分析和策略的制定
**现有模型及难点**
目前我们可以通过检测模型定位到存在作弊倾向的笔记，但是定位到具体作弊的行为还是存在难度，原因是：

- 清理作弊行为容错率低，否则会带来客诉
- 检测模型并不能定位到所有的作弊行为

### 特征分析

分析每篇作弊笔记，我们发现同笔记下作弊用户的特征有聚集现象：

- 作弊行为的用户注册渠道通常是微博和QQ
- 作弊的笔记的安卓用户设备占比明显高于正常笔记
- 作弊笔记的用户app版本较集中

下图中分别统计出了正常笔记和作弊笔记的特征分布情况

**正常笔记**


**作弊笔记**


这样的特征分布非常适合使用无监督聚类模型，目前使用的特征有

- 行为特征：
  - 行为时间
  - 行为ip

        - 。。。
- 用户特征：
    - 注册时间
    - 注册渠道
    - 常用城市
    - 用户性别
    -
- 设备信息：
    - app版本号
    - 设备平台
    - 设备机型
    - 。。。

# 模型介绍

无监督聚类模型实现流程如下：

1. 检测模型定位到疑似作弊笔记
2. 针对每篇笔记取出单天增量的所有点赞收藏及相关用户信息
3. 对于不同的特征类型，进行预处理
    a. 数据缺失的补全
    b. 类别特征，长尾处理+one-hot预处理
    c. 数值相关信息，转化为数值特征，归一化
4. 每篇笔记下的行为，做kmeans聚类
5. 对于聚类得出的每个簇，进行打分，下面的信息会影响簇的分数
    a. 簇越大，分数越高
    b. 簇占比越大，分数越高
    c. 针对类别特征，信息熵越小，分数越高
    d. 对于数值类特征，均方差越小，分数越高
6. 高分簇下的行为被识别为作弊行为

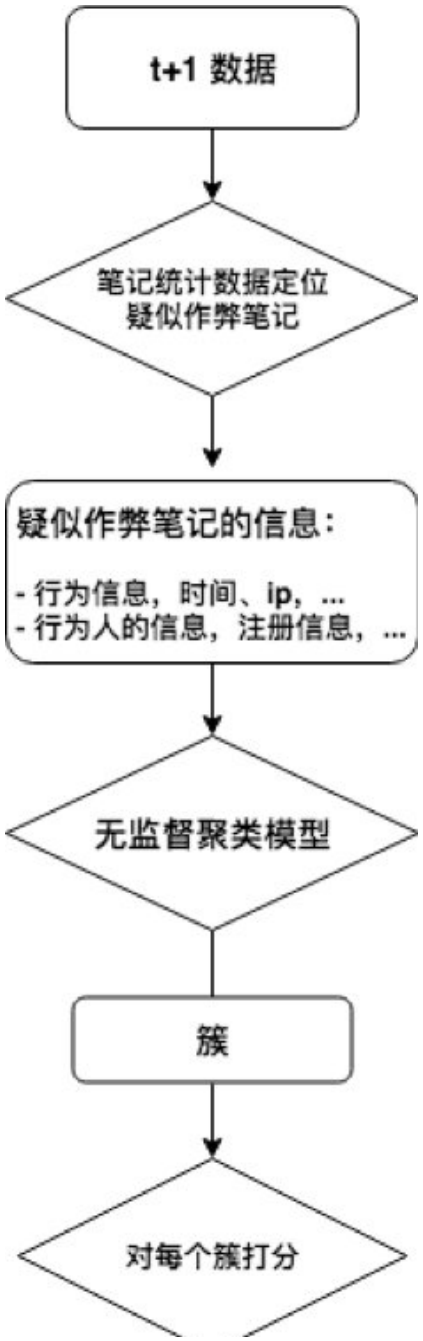
# 模型评估及思考

- 对模型结果用打点及渠道数据进行验证，模型准确率99.9%，模型上线时召回率为60%

## 思考

模型上线近半年，召回率由60%下降至10%，分析原因如下：

- 该模型被用户打击，和黑产产生了互动。黑产在进行不同的测试后找到了绕开模型的方法，如：
    - 切换不同的ip
    - 使用不同的渠道注册账号
    - 。。。
- 算法依赖的特征易被篡改，如ip，行为时间等。


t+1 数据
笔记统计数据定位
疑似作弊笔记
疑似作弊笔记的信息：
- 行为信息，时间、ip，...
- 行为人的信息，注册信息，...
无监督聚类模型
簇
对每个簇打分


作弊行为