建表思路与分析

1. 数据源
2. 使用了192.168.0.194服务器上数据
3. 全量样本：databank.freq\_info 中 op='liantong' and m=201909 分区的数据。
   1. 全量样本数量: 440466597
   2. 男性样本数量(sex=0): 170773364 占全量比重约为: 38.77%
   3. 女性样本数据(sex=1): 93600024 占全量比重约为: 21.25%
   4. 未知性别样本数量(sex=-1): 176093209 占全量比重约为: 39.98%
4. 正样本：default.positive\_sample 中 bsn='acadsoc' 分区的数据。
   1. 正样本数量: 349
   2. 男性样本数量(sex=0): 77 占全量比重约为: 22.06%
   3. 女性样本数量(sex=1): 269 占全量比重约为: 77.08%
   4. 未知性别样本数量(sex=-1): 3 占全量比重约为: 0.86%
5. 基于全量和正样本，分别选取“男”，“女” 样本建表分析。
6. 见标签组合树
7. 按照性别标签，首先分为男，女
8. 在男女群体中分别计算 年龄，文章分类，app行为，兴趣定向 四个类目的TGI值
9. 选取年龄段TGI大于1的标签作为标签组合分类依据
   1. 男：
      1. 41—49
      2. 18—23
      3. 31—40
   2. 女：
      1. 41—49
      2. 18—23
10. 因为“文章分类”标签不影响头条人数， 取TGI大于1的全体，不做具体划分
11. 选取“app行为”，“兴趣定向” 类目中TGI大于1的项目，按照顺序从大到小排序，然后划分为两类：TGI较大的一组，TGI较小的一组。建立标签组合。
12. 实现标签组合间互斥
13. 兴趣定向标签按照“一级&二级”联合标签
14. 地域标签“不限”