构建训练集、预测集、离线测试集 I --- 操作记录抽取

> *训练集和预测集的构建包括了操作记录抽取、特征提取、正负样本标注、正负样本比例调整等步骤*
> *本文档主要包含`操作记录的抽取`部分*

清洗爬虫

> *爬虫一般点击大于500但不买或者没有点击却有其它操作*
> *前者可以在未买用户的表中按点击次数去除，后者应先对`tianchi_p`过滤后产生新`tianchi_p`表，这样`才可以进行之后的步骤！！！`*

涉及表的含义

1.之前存在的表

- `tianchi_p` 所有对商品子集P中商品的操作记录
- `tianchi_p_10` 所有在11月18日到12月18日间最后10天有操作的用户对商品子集的操作记录（not used）
- `tianchi_p_10_buy` 所有在11月18日到12月18日间最后10天有操作且曾购买过的用户对商品子集的操作记录 （not used）

2.新创建的表

- `tianchi_p_1_30` 11月18日到12月17日对商品子集的操作记录
- `tianchi_p_ten_1_30` 11月18日到12月17日间最后10天有操作的用户对商品子集的操作记录
- `tianchi_p_2_31` 11月19日到12月18日对商品子集的操作记录
- `tianchi_p_ten_2_31` 11月19日到12月18日间最后10天有操作的用户对商品子集的操作记录
- `tianchi_p_31_buy` 12月18日当天的购买情况
- `tianchi_p_ten_buy_1_30` 11月18日到12月17日间有购买过且最后10天有操作的用户对商品子集的操作记录
- `tianchi_p_ten_buy_2_31` 11月19日到12月18日间有购买过且最后10天有操作的用

户对商品子集的操作记录
- `tianchi_p_ten_nobuy_1_30` 11月18日到12月17日间没有购买过、最后10天有操作且最后5天操作次数大于10小于500（有待商榷）的用户对商品子集的操作记录
- `tianchi_p_ten_nobuy_2_31` 11月19日到12月18日间没有购买过、最后10天有操作且最后5天操作次数大于10小于500（有待商榷）的用户对商品子集的操作记录

操作记录抽取

> *训练集：从11月18日到12月17日共30天的记录中抽取特征，用于训练模型及离线成绩测试*
> *预测集：从11月19日到12月18日共30天的记录中抽取特征，用于预测12月19日的购买情况*
> *离线测试集：12月18日当天有购买的（user_id,item_id）对，用于离线成绩测试*
> *对`购买过`的用户和`未购买但最后5天交互频繁`的用户分别抽取操作记录*

0. 通用步骤

- 从表`tianchi_p`中选出11月18日到12月17日对商品子集的操作记录作为表`tianchi_p_1_30`
- 从表`tianchi_p`中选出11月19日到12月18日对商品子集的操作记录作为表`tianchi_p_2_31`
- 从表`tianchi_p_1_30`中选出在11月18日到12月17日间最后10天有操作的用户对商品子集的操作记录作为表`tianchi_p_ten_1_30`
- 从表`tianchi_p_2_31`中选出在11月19日到12月18日间最后10天有操作的用户对商品子集的操作记录作为表`tianchi_p_ten_2_31`
- 从表`tianchi_p`中选出12月18日当天的所有*behavior_type*为4的用户商品对（*user_id,item_id*） 作为表`tianchi_p_31_buy`

1. 购买过用户的操作记录抽取

- 从表`tianchi_p_ten_1_30`中选出在11月18日到12月17日间购买过的用户的操作记录作为表`tianchi_p_ten_buy_1_30`
- 从表`tianchi_p_ten_2_31`中选出在11月19日到12月18日间购买过的用户的操作记录作为表`tianchi_p_ten_buy_2_31`
- 之后会从表`tianchi_p_ten_buy_1_30`和表`tianchi_p_ten_buy_2_31`中抽取各种特征

2.未购买用户的训练集构建

- 未购买的用户如果最后5天交互次数较多（大于一定次数但不应过多，否则就是爬虫），有较大可能在接下来的一天有购买行为，故应考虑在内
- 从表`tianchi_p_ten_1_30`中选出在11月18日到12月17日间未购买过但最后5天操作次数超过10次但小于500次（有待商榷）的用户的操作记录作为表`tianchi_p_ten_nobuy_1_30`
- 从表`tianchi_p_ten_2_31`中选出在11月19日到12月18日间未购买过但最后5天操作次数超过10次但小于500次（有待商榷）的用户的操作记录作为表`tianchi_p_ten_nobuy_2_31`
- 之后会从表`tianchi_p_ten_nobuy_1_30`和表`tianchi_p_ten_nobuy_2_31`中抽取各种特征