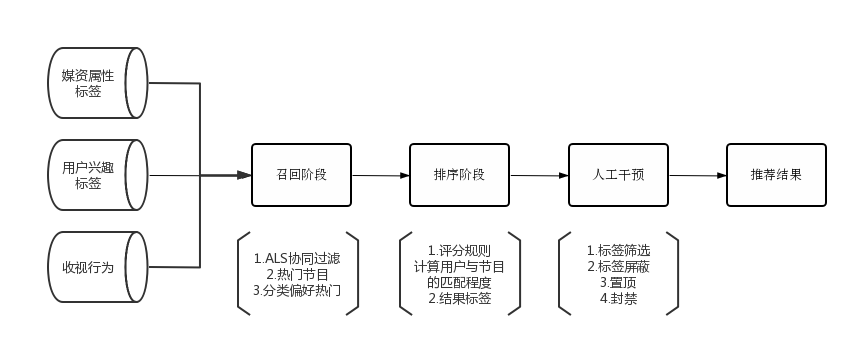
**影擎猜你喜欢处理流程**

**V2.0**

1. **整体流程**



1. **标签库**
2. 媒资属性标签

节目id

节目名

是否付费

内容类型：电影、电视剧、少儿、综艺、……

* 内容类型需要进行数据清洗，找出异常值 Select distinct content\_type

导演

演员

情节

* 如果有多个值，使用统一的分隔符（暂定为|）进行分割，对异常值、缺失值统一处理。**过滤掉出现次数较少的值**，在整个媒资库里出现过5次以下的导演、演员，出现过10次以下的情节，将其值排除。（或者只保留热度在前50%的导演、演员，具体视数据情况而定）

统计排序先看结果

地区

语言

* 需要进行数据清洗，注意同一地区或语言的不同表达，应处理成统一值，例如EN、【英】、英文、英语都处理成英语。

Input：具体省份的媒资信息表

从媒资信息表提取过滤格式化，对导演、演员、情节出现情况进行统计排序，取出50%的数据验证质量，根据数据质量决定保留内容。

Ouput: ads\_media\_attribute

与小飞或晓伟沟通，获取各媒资省数据来源以及数据判定规则，每日定时获取昨日数据。

|  |  |  |
| --- | --- | --- |
| field | desc |  |
| video\_id | 节目编号 | 唯一 |
| video\_name | 节目名 |  |
| is\_paid | 是否付费 | 0/1 |
| content\_type | 内容类型 | 格式化 |
| director | 导演 | 格式化，统计导演出现次数，取热度top |
| actor | 演员 | 格式化 |
| plot | 情节 | 格式化 |
| region | 地区 | 归一化 |
| language | 语言 | 归一化 |
| dt | 时间 | 分区 |

CREATE TABLE IF NOT EXISTS DB.ads\_media\_attribute

(

video\_id STRING,

video\_name STRING,

is\_paid STRING,

content\_type STRING,

director STRING,

actor STRING,

plot STRING,

region STRING,

language STRING

) PARTITIONED BY (dt STRING)

ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'

1. 用户兴趣标签

结合媒资属性标签和用户收视行为，生成用户兴趣标签库

长期兴趣标签：统计近三个月的收视行为（如果数据不足三个月，取最长时间）

短期兴趣标签：统计近一周的收视行为

生成方法：

对用户短期/长期观看过的节目，累加其媒资数据标签，并统计其**枚举值的出现次数占比**

举例说明：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 节目id | 节目名 | 是否付费 | 类型 | 导演 | 演员 | 地区 | 语言 | 情节 |
| 756454 | 绿野仙踪 | 免费 | 电影 | NULL | NULL | 美国 | NULL | NULL |
| 75599 | 悬崖之上 | 免费 | 电影 | 张艺谋 | 倪大红|张译|于和伟|秦海璐|刘浩存|余皑磊|雷佳音|沙溢|李乃文|朱亚文 | 中国大陆 | 国语 | 剧情|动作|悬疑 |
| 76199 | 少年派的奇幻漂流 | 免费 | 电影 | 李安 | 拉菲·斯波|伊尔凡·可汗|苏拉·沙玛|杰拉尔·德帕迪约|塔布|阿迪勒·侯赛因 | 美国 | 英语 | 奇幻|剧情|冒险 |
| 228114 | 安家 | 付费 | 电视剧 | 安建 | 孙俪|罗晋|张萌|海清 | 中国大陆 | 普通话 | NULL |

|  |  |  |  |
| --- | --- | --- | --- |
| 用户ID | 播放开始时间 | 观看节目名 | 观看节目id |
| 9e861196d5a73bcd | 2022/3/1 17:00 | 绿野仙踪 | 756454 |
| 9e861196d5a73bcd | 2022/3/2 18:00 | 悬崖之上 | 75599 |
| 9e861196d5a73bcd | 2022/3/4 18:30 | 少年派的奇幻漂流 | 76199 |
| 9e861196d5a73bcd | 2022/3/6 17:00 | 安家 | 228114 |
| 9e861196d5a73bcd | 2022/3/7 17:00 | 安家 | 228114 |

用户兴趣标签

|  |  |
| --- | --- |
| 用户ID | 9e861196d5a73bcd |
| 短期付费偏好 | {付费:0.6,免费:0.4} |
| 短期类型偏好 | {电影:0.6,电视剧:0.4} |
| 短期导演偏好 | {张艺谋:0.25,李安:0.25,安建:0.5} |
| 短期演员偏好 | {倪大红:0.04,拉菲·斯波:0.04,孙俪:0.08,……} |
| 短期地区偏好 | {美国:0.4,中国大陆:0.6} |
| 短期语言偏好 | {国语:0.25,英语:0.25,普通话:0.5} |
| 短期情节偏好 | {剧情:0.33，动作:0.17,悬疑:0.17,……} |
| 长期付费偏好 | {……} |
| 长期类型偏好 | {……} |
| 长期导演偏好 | {……} |
| 长期演员偏好 | {……} |
| 长期地区偏好 | {……} |
| 长期语言偏好 | {……} |
| 长期情节偏好 | {……} |

宽表

Input：用户收视行为表，ads\_media\_attribute

Output：ads\_user\_interest\_preference

|  |  |  |
| --- | --- | --- |
| Field | desc |  |
| device\_id | 设备编号 |  |
| short\_paid\_pre | 短期付费偏好-收费 |  |
| short\_free\_pre | 短期付费偏好-免费 |  |
| short\_type\_film\_pre | 短期类型偏好-电影 |  |
| short\_type\_teleplay\_pre | 短期类型偏好-电视剧 |  |
| short\_type\_...\_pre | 短期类型偏好-查询到的所有类型 |  |
| short\_director\_zhangyimou\_pre | 短期导演偏好-张艺谋 |  |
| short\_director\_zhouxingchi\_pre | 短期导演偏好-周星驰 |  |
| short\_director\_...\_pre | 短期导演偏好-热度筛选留下的导演 |  |
| short\_actor\_zhourenfa\_pre | 短期演员偏好-周润发 |  |
| short\_actor\_zhoujielun\_pre | 短期演员偏好-周杰伦 |  |
| short\_actor\_...\_pre | 短期演员偏好-热度筛选留下的演员 |  |
| short\_region\_cn\_pre | 短期地区偏好-中国大陆 |  |
| short\_region\_en\_pre | 短期地区偏好-英国 |  |
| short\_region\_...\_pre | 短期地区偏好-查询到的大部分地区 |  |
| short\_language\_hanyu\_pre | 短期语言偏好-汉语 |  |
| short\_language\_minnanyu\_pre | 短期语言偏好-闽南语 |  |
| short\_language\_...\_pre | 短期语言偏好-其他主流语言 |  |
| short\_plot\_story\_pre | 短期情节偏好-剧情 |  |
| short\_plot\_action\_pre | 短期情节偏好-动作 |  |
| short\_plot\_...\_pre | 短期情节偏好-其他类型 |  |
| long\_paid\_pre | 长期付费偏好 |  |
| long\_type\_pre | 长期类型偏好 |  |
| long\_director\_pre | 长期导演偏好 |  |
| long\_actor\_pre | 长期演员偏好 |  |
| long\_region\_pre | 长期地区偏好 |  |
| long\_language\_pre | 长期语言偏好 |  |
| long\_plot\_pre | 长期情节偏好 |  |

输入两表通过video\_id关联，分布统计用户收视各兴趣下每个类型观看数与该兴趣总数之比，按照短期/长期（1周/1月）统计后存入中间表中。

统计时是不去重用户收视数据

CREATE TABLE IF NOT EXISTS DB.ads\_user\_interest\_preference

(

device\_id STRING,

short\_paid\_pre STRING,

short\_type\_pre STRING,

short\_director\_pre STRING,

short\_actor\_pre STRING,

short\_region\_pre STRING,

short\_language\_pre STRING,

short\_plot\_pre STRING,

long\_paid\_pre STRING,

long\_type\_pre STRING,

long\_director\_pre STRING,

long\_actor\_pre STRING,

long\_region\_pre STRING,

long\_language\_pre STRING,

long\_plot\_pre STRING

) PARTITIONED BY (dt STRING)

ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'

分母：该用户的最近一周/一月的收视总数

分子：每个小项的总数

1. **召回阶段**

召回阶段的目的，是将每个用户的推荐节目数量从万级降低到百级，减少排序的计算量。

召回阶段使用多路召回策略，每一路使用不同的筛选逻辑，每路选出50~200个节目。

路线1：ALS协同过滤，使用一周~一个月的收视行为数据，每个用户输出50~200个候选节目

路线2：热门节目，统计一周~一个月的热门节目，取Top50~200

路线3：分类偏好热门，取每个用户排名前3的类型偏好，取分类热门节目共计50~200个

|  |  |
| --- | --- |
| 用户ID | 9e861196d5a73bcd |
| 召回路线1 | [节目id, 节目id, 节目id, 节目id, ……] |
| 召回路线2 | [节目id, 节目id, 节目id, 节目id, ……] |
| 召回路线3 | [节目id, 节目id, 节目id, 节目id, ……] |
| 总召回 | [节目id, 节目id, 节目id, 节目id, ……] |

路线1：

输入：用户收视数据

将用户和媒资符号化带入ALS模型进行训练，得到特征矩阵后，对每个用户进行推荐，取出top节目列表，要做总数据的模型以及各个类型的模型

输出：ads\_als\_recommend\_videos【用户，推荐节目列表】

路线2：

输入：用户收视数据

一周分类热门，总热门，统计每个节目的观看用户数，排序取出数量最多的top个，总的免费的付费的

输出：ads\_hot\_recommend\_videos【热门节目列表】

路线3：

输入：ads\_user\_interest\_preference

取出用户短期类型偏好，对偏好列表排序，取出用户排名前3的类型偏好，统计每个类别下各个媒资的观看数，每个类别取出top个作为该类型的热门节目列表。

输出：ads\_type\_hot\_recommend\_videos【用户，类型1热门列表，类型2热门列表，类型3热门列表】 总的免费的付费的

路线合并去重，一周做一次

总召回：

路线1总的，路线2总的，路线3总的

路线1过滤掉付费的，路线2免费的，路线3免费的

路线1过滤掉免费的，路线2付费的，路线3付费的

其余各类型也是分为总的，免费的，付费的

电影类型路线1总的，电影热门路线2总的

电影类型路线1过滤掉付费的，电影热门推荐路线2免费的

电影类型路线1过滤掉免费的，电影热门推荐路线2付费的

其他类型同样

设计三张表，一张总召回表，一张付费召回表，一张免费召回表

总召回表：【device\_id, 路线1总，路线2总，路线3总；

device\_id, 电影路线1总，电影路线2，null；…】

付费召回表：【device\_id, 路线1总 付费，路线2总 付费，路线3总 付费；

device\_id, 电影路线1总 付费，电影路线2 付费，null；…】

免费召回表：【device\_id, 路线1总 免费，路线2总 免费，路线3总 免费；

device\_id, 电影路线1总 免费，电影路线2 免费，null；…】

1. **排序阶段**

计算每个用户，对每个召回候选节目的评分。

计算公式：

（用户，节目）短期评分 =

W1 \* 用户短期付费偏好 · 节目是否付费 +

W2 \* 用户短期类型偏好 · 节目类型 +

W3 \* 用户短期导演偏好 · 节目导演 +

W4 \* 用户短期演员偏好 · 节目演员 +

W5 \* 用户短期地区偏好 · 节目地区 +

W6 \* 用户短期语言偏好 · 节目语言 +

W7 \* 用户短期情节偏好 · 节目情节 +

· 定义为内积运算，将两边都出现的项对应的值相乘，然后求和。

举例说明1：

用户短期类型偏好={电影:0.6,电视剧:0.4}

节目类型=电影={电影:1}

用户短期类型偏好 · 节目类型=0.6\*1=0.6

举例说明2：

用户短期演员偏好={倪大红:0.04,拉菲·斯波:0.04,孙俪:0.08,……}

节目演员=孙俪|罗晋|张萌|海清={孙俪:0.25, 罗晋:0.25, 张萌:0.25, 海清:0.25}

用户短期演员偏好 · 节目演员=0.08\*0.25+0.08\*0.25+0.08\*0.25+0.08\*0.25=0.08

W1~W7为权重，是大于等于0的实数，具体数值可以根据实际效果调整，或者通过逻辑回归模型训练。

（用户，节目）长期评分与（用户，节目）短期评分计算方法相类似。

评分计算的本质就是求用户偏好与节目属性的匹配度，神经网络、XGBoost、LightGBM等复杂模型是为了自动寻找交叉特征、特征与特征之间的运算关系，但在我们的场景中特征之间的对应关系非常简单明了，就是付费偏好对应节目是否付费，类型偏好对应节目类型，所以可以直接写出公式，无需使用复杂模型。

|  |  |
| --- | --- |
| 用户ID | 9e861196d5a73bcd |
| 召回路线1 | [节目id, 节目id, 节目id, 节目id, ……] |
| 召回路线2 | [节目id, 节目id, 节目id, 节目id, ……] |
| 召回路线3 | [节目id, 节目id, 节目id, 节目id, ……] |
| 候选节目 | [节目id, 节目id, 节目id, 节目id, ……] |
| 节目评分 | [ {节目id:xxx, 评分:xxx, 排名:xxx}, {节目id:xxx, 评分:xxx, 排名:xxx}, {节目id:xxx, 评分:xxx, 排名:xxx}, ……(根据场景需求约有10~20个推荐节目) ] |

输入：ads\_user\_interest\_preference，

ads\_media\_attribute,

ads\_als\_recommend\_videos,

ads\_hot\_recommend\_videos,

ads\_type\_hot\_recommend\_videos

对于每一个用户，从三路召回中合并去重所有媒资，得到不同类型的各种召回路线。通过关联用户偏好表获取用户的短期偏好，通过关联媒资属性表获取媒资属性。根据公式计算每个用户每个节目的评分。权重根据具体数据来调整或通过逻辑回归训练。

输出：ads\_recommend\_videos\_result

|  |  |  |
| --- | --- | --- |
| field |  |  |
| device\_id |  |  |
| rec\_res | {“video\_id”:video\_id, ”video\_name”:video\_name, “rating”:rating, “rank”:rank} |  |

三张召回表同类型（同一行）的所有召回路线去重合并，并依次计算每种路线的节目评分

1. **人工干预**

7

人工干预根据需求不同，在不同的阶段进行，不一定要在排序阶段过后再去做筛选。

封禁内容：在节目标签库新增封禁字段，封禁节目不参与召回

单开一张Mysql封禁表，每次程序读取并关联获取封禁内容

置顶内容：在结果输出阶段插入要置顶的节目

不需要考虑

标签筛选：在召回阶段，只选择指定标签的节目，如只要电影类型的节目

总类型

各个类型的ALS模型

召回设置多分支路线

标签屏蔽：在召回阶段，屏蔽择指定标签的节目

混合编排：例如推荐列表固定为1置顶+2热门+3个性化

1. **结果输出**

具体数据库类型、表名、字段，由宋坤负责设计并输出文档。