安徽猜你喜欢详细设计

# 目标

根据不同场景推荐不同节目。

# 输入

媒资信息表：knowyou\_ott\_ods.dim\_pub\_video\_df （电视节目信息维度日表）

|  |  |  |
| --- | --- | --- |
| 字段名 | 字段类型 | 字段注释 |
| series | STRING |  |
| series\_action | STRING | 操作类型 发布－REGIST 更新－UPDATE 删除－DELETE |
| series\_name | STRING | 连续剧名称 |
| series\_ordernumber | STRING | 订购编号 |
| series\_originalname | STRING | 原名 |
| series\_sortname | STRING | 排序名称 |
| series\_searchname | STRING | 索引名称 |
| series\_orgairdate | STRING | 首播日期(YYYYMMDD) |
| series\_licensingwindowstart | STRING | 有效开始时间(YYYYMMDDHH24MiSS) |
| series\_licensingwindowend | STRING | 有效结束时间(YYYYMMDDHH24MiSS) |
| series\_displayasnew | STRING | 新到天数 |
| series\_displayaslastchance | STRING | 剩余天数 |
| series\_macrovision | STRING | 拷贝保护标志 0:无拷贝保护 1:有拷贝保护 |
| series\_price | STRING | 含税定价 |
| series\_volumncount | STRING | 总集数 |
| series\_status | STRING | 状态标志 0:失效 1:生效 |
| series\_description | STRING | 描述信息 |
| series\_type | STRING | 节目内容类型： 电视台|电影|电视剧|少儿|综艺|音乐|体育|教育|片花|纪录片|娱乐|时尚|搞笑|游戏|新闻|生活|动漫|其他 |
| series\_rmediacode | STRING | 关联内容唯一标 |
| series\_titlesc | STRING | 讯飞补全提供的节目名称 |
| series\_score | STRING | 讯飞提供豆瓣评分，取值范围0-10，精确到1位小数点，如：9.0 |
| series\_imagesc | STRING | 讯飞补充节目图片URL |
| series\_updatetimesc | STRING | 讯飞更新时间 格式：YYYY-MM-dd hh:mm:ss |
| series\_cmsid | STRING | 内容归属的CMSID |
| series\_result | STRING | 0: 成功 其他: 错误代码 |
| series\_errordescription | STRING | 错误描述 |
| series\_isfree | STRING |  |
| series\_director | STRING | 导演 |
| series\_actordisplay | STRING | 演员列表(只供显示) |
| series\_originalcountry | STRING | 国家地区 |
| series\_language | STRING | 语言 |
| series\_keywords | STRING | 关键字 |
| package\_series | STRING |  |
| createtime | STRING |  |
| updatetime | STRING |  |
| picture\_series\_1\_1\_fileurl | STRING | 媒体文件URL |
| series\_reserve1 | STRING | 保留字段 |
| series\_reserve2 | STRING | 保留字段 |
| series\_reserve3 | STRING | 保留字段 |
| series\_reserve4 | STRING | 保留字段 |
| series\_reserve5 | STRING | 保留字段 |
| video\_type | STRING | 手动维护节目分类 |
| picture\_series\_1\_1\_height | STRING |  |
| picture\_series\_1\_1\_rate | STRING |  |
| picture\_series\_1\_1\_width | STRING |  |
| dt | STRING | 日分区字段 |

用户行为表：dws\_rec\_video\_playinfo\_di（推荐系统用户播放行为汇总表）

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名 | 字段类型 | 字段注释 | 备注 |
| data\_source | STRING | 数据源 | CP数据源及为具体CP名或者为空，探针数据填充为“探针” |
| deviceid | STRING | 设备id |  |
| actualtime | STRING | 数据发生时间 | yyyyMMddHHmmss |
| provincecode | STRING | 省份编码 |  |
| citycode | STRING | 地市编码 |  |
| urlvideoid | STRING | 播放url中的节目id |  |
| jsonvideoid | STRING | 节目json中的节目id |  |
| actionpath | STRING | 路径 |  |
| firstlevel | STRING | 一级栏目 | actionpath为2个内容时且第二个内容包含\_， |
| secondlevel | STRING | 二级栏目 | actionpath长度等于4且第二位内容包含\_，且最后3、4位不等于相关推荐、播放记录、大家都在看、专题、筛选、收藏、搜索 |
| threelevel | STRING | 三级栏目 | actionpath长度等于4且第二位内容包含\_，且最后3、4位不等于相关推荐、播放记录、大家都在看、专题、筛选、收藏、搜索 |
| entrytype | STRING | 入口类别 | 搜索、筛选、收藏、播放记录、推荐、专题，倒数第二位 为专题、主演、导演、筛选、播放记录，最后一位为相关推荐、大家都在看则为推荐 |
| pannelpos | STRING | 海报位置 | 1\_1 |
| videotype | STRING | 播放内容的视频类型信息 | 取值为：点播、直播、回看 |
| recommendedtype | STRING | 内容推荐的类型 | actionpath最后一位为相关推荐，大家都在看， |
| programname | STRING | 节目名称 | 以电视剧都挺好上报为例，“都挺好” |
| contentname | STRING | 内容名称 | 电视剧会包含集数信息“都挺好\_02”，电影名字与programName一般一样 |
| channelname | STRING | 频道信息 | 仅对直播和回看有效 |
| oldchannel | STRING | 原频道信息 | 仅在推荐节目时有效 |
| oldcontentname | STRING | 进入推荐节目的前一次观看节目名称 | 以电视剧都挺好上报为例：“都挺好”。仅在推荐节目时有效 |
| programurl | STRING | 节目url | 来自于播放器事件 |
| playtime | STRING | 播放时长 | 仅节目结束时候（包含正常结束和失败）上报字段，单位秒 |
| contenttype | STRING | 分类 |  |
| licensename | STRING | 牌照方名字 | 解析actioninfo中version字段获取 |
| pkg | STRING | 牌照方包名 | com.bestv.ott.baseservies |
| licenseversion | STRING | 牌照方版本全称 | BestTV\_LiteIU\_AHYD\_7.8.2110.2 |
| luaversion | STRING | lua脚本版本号 | 可用来判断脚本升级情况 |
| seriesheadcode | STRING | 剧集id |  |
| series\_action | STRING | 操作类型 发布－REGIST 更新－UPDATE 删除－DELETE |  |
| series\_name | STRING | 连续剧名称 |  |
| series\_searchname | STRING | 索引名称 |  |
| series\_status | STRING | 状态标志0:失效 1:生效 |  |
| series\_description | STRING | 描述信息 |  |
| series\_type | STRING | 节目内容类型： 电视台|电影|电视剧|少儿|综艺|音乐|体育|教育|片花|纪录片|娱乐|时尚|搞笑|游戏|新闻|生活|动漫|其他 |  |
| series\_cmsid | STRING | 内容归属的CMSID |  |
| series\_isfree | STRING | 免付费 |  |
| series\_director | STRING | 导演 |  |
| series\_actordisplay | STRING | 演员 |  |
| series\_originalcountry | STRING | 地区 |  |
| series\_language | STRING | 语言 |  |
| series\_keywords | STRING | 情节 |  |
| package\_series | STRING | 产品包 |  |
| picture\_series\_1\_1\_fileurl | STRING | 海报图片 |  |
| series\_reserve5 | STRING | 保留字段（发行年份） |  |
| dt | STRING | 日分区字段 |  |

# 输出

knowyou\_ott\_dmt.guess\_like\_recommend\_videos\_result

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名 | 字段类型 | 字段注释 | 备注 |
| Device\_id | STRING | 设备编号 |  |
| All\_rec\_list | STRING | 默认场景推荐 | [{  “video\_id”:video\_id1,  ”video\_name”:video\_name1,  “rating”:rating,  “rank”:rank  },  {  “video\_id”:video\_id2,  ”video\_name”:video\_name2,  “rating”:rating,  “rank”:rank  }, …] |
| All\_rec\_list\_free | STRING | 免费场景推荐 |  |
| All\_rec\_list\_paid | STRING | 付费场景推荐 |  |
| Film\_rec\_list | STRING | 电影类型推荐 |  |
| Film\_rec\_list\_free | STRING | 免费电影类型推荐 |  |
| Film\_rec\_list\_paid | STRING | 付费电影类型推荐 |  |
| 其他类型召回 | STRING |  |  |
| 其他类型免费召回 | STRING |  |  |
| 其他类型付费召回 | STRING |  |  |

# 实现流程

## 提取媒资属性

验证媒资信息表knowyou\_ott\_ods.dim\_pub\_video\_df中免付费，内容类型，导演，演员，情节，地区，语言数据格式，处理异常格式、脏数据等内容。

统计导演、演员、情节出现频次，与晓伟确定热度保留规则。

规则确定之后，对媒资属性字段提取，将免付费，内容类型，导演，演员，情节，地区，语言字段进行格式化，归一化处理，以‘|’分割多个内容。

免付费剔除非0/1的脏数据，内容类型重新分类后剔除脏数据；

导演，演员暂无处理；

情节，地区，语言脏数据替换成”未知”再根据分类规则进行格式化，分类参考《媒资分类规则》。

情节清洗存入表knowyou\_ott\_dmt.guess\_like\_media\_plot

提取结果存入knowyou\_ott\_dmt.guess\_like\_media\_attribute表中，并按照日期进行分区

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名 | 字段类型 | 字段注释 | 备注 |
| video\_id | STRING | 节目编号 | 唯一 |
| video\_name | STRING | 节目名 |  |
| is\_paid | STRING | 是否付费 | 0/1 |
| content\_type | STRING | 内容类型 | 格式化，确定所有类型，对异常数据进行合并或丢弃，确定所用情节 |
| director | STRING | 导演 | 格式化(、符号替换为|（\*\*\*）?符号删除)，多个导演按照‘|’的方式进行分割，导演为空的与晓伟确定数据后讨论处理方式。 |
| actor | STRING | 演员 | 格式化，多个演员按照‘|’的方式进行分割，与晓伟讨论为空的处理方式 |
| plot | STRING | 情节 | 格式化，多个情节按照‘|’的方式进行分割，排除出现频次少的情节 |
| region | STRING | 地区 | 格式化，多个地区按照‘|’的方式进行分割；归一化，相同含义的地区统一表达式，欧美日韩分类问题与晓伟进行讨论 |
| language | STRING | 语言 | 格式化，多个语言按照‘|’的方式进行分割；归一化，相同含义的语言统一表达 |
| dt | STRING | 日期 | 分区 |

## 提取用户兴趣标签

将用户兴趣标签分为两类：短期（一周）兴趣和长期（一月）兴趣，两者具有相同的标签，处理时逻辑也相同，唯一不同点在于时间维度。

短期兴趣标签存入：knowyou\_ott\_dmt.guess\_like\_short\_user\_interest\_preference

长期兴趣标签存入：knowyou\_ott\_dmt.guess\_like\_long\_user\_interest\_preference

付费偏好：分为付费和免费，分别统计当前用户在设定时间内两者观看数量占观看总数的占比。

类型偏好：主要类型为电视台|电影|电视剧|少儿|综艺|音乐|体育|教育|片花|纪录片|娱乐|时尚|搞笑|游戏|新闻|生活|动漫|其他，其他类型如何保留再与晓伟讨论，分别统计当前用户在设定时间内各类型观看数量占观看总数的占比。

导演偏好：根据讨论结果选定导演列表，统计当前用户在设定时间内热门导演节目观看数量占观看总数的占比。

演员偏好：根据讨论结果选定演员列表，统计当前用户在设定时间内热门演员节目观看数量占观看总数的占比。

地区偏好：根据讨论结果选定地区列表，主要考虑国外节目是否按照欧美、日韩等分类，统计当前用户在设定时间内各地区节目观看数量占观看总数的占比。

语言偏好：考虑中文方言情形是否需要合并，统计当前用户在设定时间内各语言节目观看数量占观看总数的占比。

情节偏好：需要对数据库中的情节做筛选后，统计当前用户在设定时间内各情节节目观看数量占观看总数的占比。

以上情景都需要考虑‘无’，‘NULL’，空值等特殊值的处理逻辑。

## 多路召回，获取推荐列表

召回阶段以三条路线为主，分别为(1)ALS协同过滤；(2)热门推荐；(3)分类下的热门推荐：

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

路线2：热门节目，统计一周的热门节目（以每个节目的观看用户数为依据，分总热门和各个类型的热门），取Top50~200。

路线3：分类偏好热门，根据用户的短期类型兴趣偏好取出每个用户排名前3的类型偏好，再根据类型取出分类热门节目共计50~200个。

同时为了照顾各种不同场景，每种场景需要额外做精细化处理，针对每个用户具体可分为以下情景：

总召回：路线1所有节目带入ALS模型训练后得到的推荐清单 + 路线2所有推荐的热门节目清单 + 路线3当前用户偏好前三的节目类型中每个类型的热门推荐清单；

总召回-付费：路线1过滤掉免费节目后带入ALS模型训练后得到的推荐清单 + 路线2所有推荐的热门付费节目清单 + 路线3当前用户偏好前三的节目类型中每个类型的热门付费推荐清单；

总召回-免费; 路线1过滤掉付费节目后带入ALS模型训练后得到的推荐清单 + 路线2所有推荐的热门免费节目清单 + 路线3当前用户偏好前三的节目类型中每个类型的热门免费推荐清单；

节目类型召回-暂定类型为电视台|电影|电视剧|少儿|综艺|音乐|体育|教育|片花|纪录片|娱乐|时尚|搞笑|游戏|新闻|生活|动漫|其他，以电影类型为例分：

电影召回：路线1所有电影类型节目带入ALS模型训练后得到的推荐清单 + 路线2所有推荐的热门电影节目清单；

电影召回-付费：路线1所有电影类型过滤掉免费节目后带入ALS模型训练后得到的推荐清单 + 路线2所有推荐的热门付费电影节目清单；

电影召回-免费：路线1所有电影类型过滤掉付费节目后带入ALS模型训练后得到的推荐清单 + 路线2所有推荐的热门免费电影节目清单；

总计得到（类型数 + 1）\* 3种召回路线，每种代表各自的场景。

召回时还需考虑封禁情况，暂定从Mysql中获取需要屏蔽的节目，在最终召回时将这类节目从召回列表中移除。

最后将，相同场景中所有的召回路线去重合并用于下一步排序。

结果存储时设计四张表，一张总召回表，一张付费召回表，一张免费召回表，以及总召回表：

总召回表：【 device\_id, 路线1总， 路线2总， 路线3总，ALL；

device\_id, 电影路线1总，电影路线2，null， 电影；

其他类型，路线1总， 路线2， null， 类型 】

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，ALL；

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

其他类型同理… 】

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，ALL；

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

其他类型同理… 】

最终召回表：

【device\_id，总召回，付费总召回，免费总召回，电影总召回，付费电影总召回，免费电影总召回，其他场景召回】

## 评分计算

每个场景都得到一条召回路线，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：

（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 +

最终得到每个用户在每个场景下的推荐节目以及该节目的具体评分，根据评分对节目进行排序，得到最终的推荐结果，并将结果存入knowyou\_ott\_dmt.guess\_like\_recommend\_videos\_result