短视频推荐方案

# 1、模型状态梳理

## 1.1 召回

排序：

1. 置顶视频，有置顶视频配置
   1. 全网置顶top5
2. 营销视频，有营销类型排序
   1. 除去置顶视频
   2. 筛选：用二级标签按5G业务＞流量＞权益＞会员＞新业务＞套餐＞终端＞宽带
   3. 用户标签ID（user\_label）和视频客群ID（labelId）有交集的情况排序优先级最高（高于5G业务）
   4. 排序：取前5个
      1. 地市>分省>全网
      2. 已浏览时间距现在时长；（video\_history\_d>video\_history\_h>video\_history\_k>video\_history\_r)
      3. label>5G业务＞流量＞权益＞会员＞新业务＞套餐＞终端＞宽带
3. 新视频，有发布时间排序
   1. 除去置顶、营销视频
   2. 筛选：发布时间是近7天
   3. 排序：取前10
      1. 地市>分省>全网
      2. 已浏览时间距现在时长；（video\_history\_d>video\_history\_h>video\_history\_k>video\_history\_r)
4. 其他视频
   1. 筛选：除去置顶、营销、新视频
   2. 排序：取前5
      1. 地市>分省>全网
      2. 已浏览时间距现在时长；（video\_history\_d>video\_history\_h>video\_history\_k>video\_history\_r)

召回25个视频:置顶视频:营销视频:新视频:其他视频=1 : 1 : 2 : 1

（5、20、40、20）---》（5、5、10、5）

## 1.2 样本

时间窗口：7天（20241015 20241021）

样本量：600万

目标：播放时长是否大于5s，正负样本比例：1 : 1

数据表：

1. feed流短视频播放维度统计数据日表

csap663.TB\_DW\_CT\_XDS\_LVCS\_VIDEO\_PLAY\_STATISTICS\_DAY

1. feed流短视频属性维度统计数据日表

csap663.TB\_DW\_CT\_XDS\_LVCS\_VDEO\_PROPERTY\_STATISTICS\_DAY

1. 短视频用户画像日表

casap662.TB\_RP\_CT\_RCM\_VIDEO\_USER\_PORT\_DAY

1. 视频直播用户轨迹曝光模型

csap662.TB\_DWS\_CT\_LVCS\_BH\_PAGE\_IMP\_DAY

1. 视频直播用户轨迹浏览点击离线型

csap662.TB\_DWS\_CT\_LVCS\_BH\_PAGE\_VEW\_CLK\_OFFLINE\_DAY

样本分析：统计播放时长时间分桶，类目播放时长分布

数据口径：

曝光：

## 1.3 特征

统计表格：[ksheet短视频特征](https://www.kdocs.cn/l/ckyKbvoHI8M6)

### 1.3.1 用户特征

#### 用户画像

用户画像原数据表，缺少用户基本信息。可扩充特征，年龄，性别，消费情况

[关联表]

### 1.3.2 物料特征

#### 短视频播放维度

短视频播放维度原数据表

[关联表]

#### 短视频属性维度

短视频属性维度原数据表，缺少基本信息，如标题

[关联表]

### 1.3.3 特征处理

#### 特征衍生

##### 用户历史有效播放序列

根据用户历史播放行为构造有效播放序列（播放时长大于5s）

[关联表]

##### 用户播放历史衍生特征

根据历史序列衍生出统计特征。

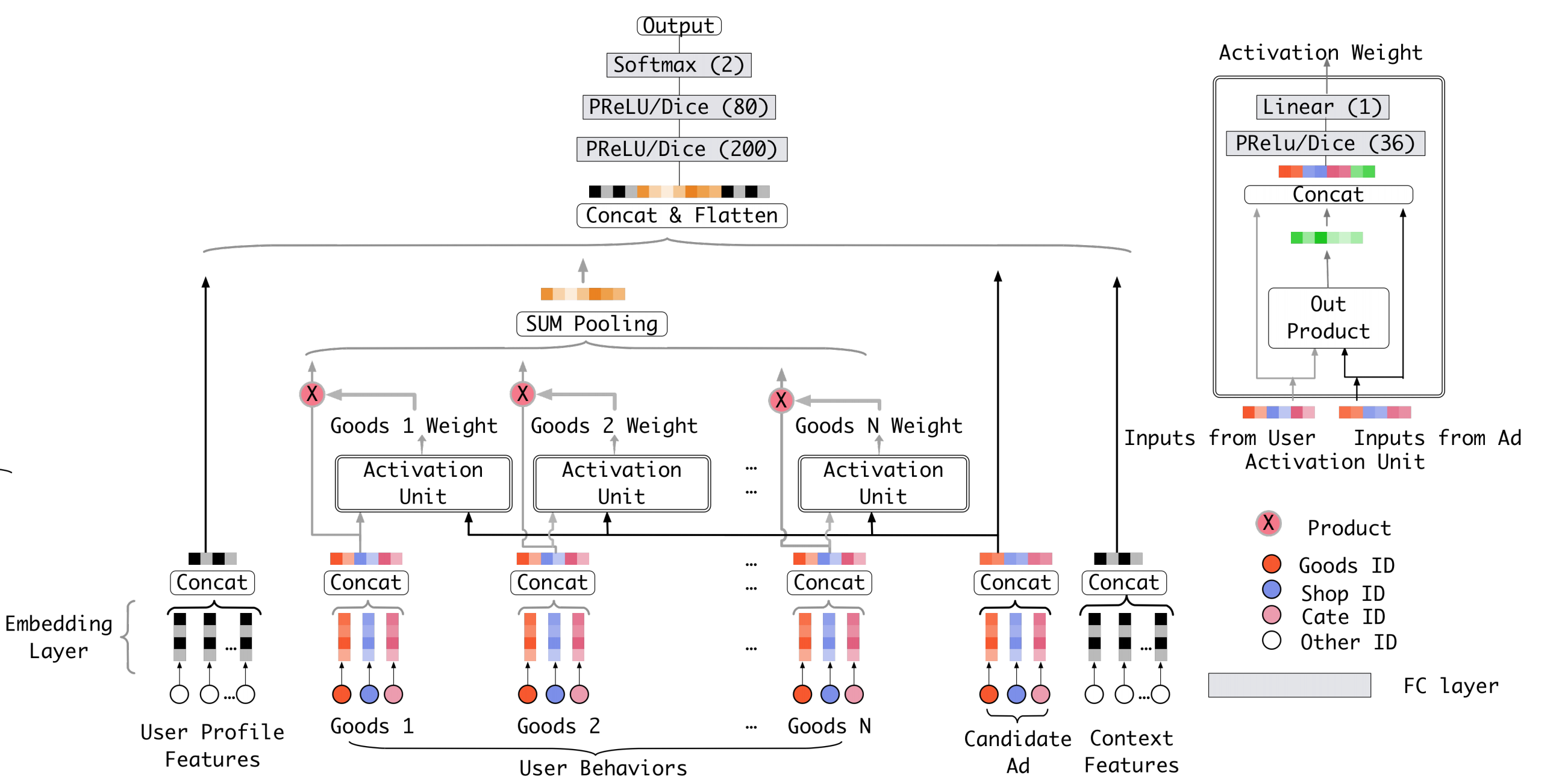
[关联表]

#### 特征预处理

一般特征无法直接输入模型，将特征进行归一化，标签化处理。

## 1.4 模型

深度兴趣网络（DIN）



# 2、模型优化

## 2.1 数据优化

目前数据存在一定缺失情况，期望增加的数据有：

1. 用户画像数据：
   1. 基本信息：年龄、性别、省份
   2. 资费特征：套餐价格、 【营销类视频、卖出去是哪些类的商品， 看可能跟哪些资费特征相关】
   3. 设备相关：使用设备类型、网络状况（是否5G、是否wifi）
   4. 用户行为特征扩展：类目1、 类目2 3天/7天/14天 统计特征扩展
2. 视频相关数据：发布日期， 3天内效率数据、7天内效率数据、14天内效率数据 【优先级高】



1. 行为相关数据：是否点赞、是否分享、操作 小时
2. 实时数据：【线上与离线数据整合，思考清楚如何使用】近1、5、10分钟点赞、近1、5、10分钟分享、近10条视频停留时长与视频时长占比、近10条视频停留时长、近10条视频类型

## 2.2 召回优化

目前的召回方案非常单一，主要依赖业务需求，增加多种召回方案能提高系统的覆盖率和召回率，同时保证推荐结果的多样性和新颖性，实施流程如下：

1. 召回池构建：将待召回的候选视频放入召回池。

2. 过滤和去重：根据业务规则，去掉与用户历史记录重复的内容、去掉业务侧要求不推荐的内容

3. 多样性调整：确保召回结果的主题和内容丰富度。

4. 候选集排序：将召回的候选视频进入排序模块，进行更细致的打分和排序。

**1. 基于协同过滤的召回**

（2）视频-视频协同过滤：根据视频之间的相似性，为用户推荐与他观看过的视频类似的内容。

具体的操作方法：通过视频共现矩阵计算相似性，或者使用ItemCF方法，对用户观看过的视频找到相似视频推荐。

**2. 基于内容的召回**

视频内容标签匹配：根据用户对某些内容标签（如搞笑、教育、体育等）的偏好，召回具有相同或相似标签的视频。

具体的操作方法：为视频和用户建立标签向量，基于标签相似度（如余弦相似度）进行推荐。

**3. 基于矩阵分解的召回**

矩阵分解（MF）：基于用户-视频交互行为（观看、点赞等）构建评分矩阵，使用矩阵分解技术将用户和视频映射到同一低维空间中，从而实现用户与视频的匹配。

具体的操作方法：使用SVD或ALS等矩阵分解方法将评分矩阵分解为用户和视频的隐向量矩阵，根据隐向量的相似度进行召回。

隐语义模型（LSI/LDA）：将视频内容和用户兴趣通过主题模型降维，生成用户和视频的隐语义空间向量，再通过相似度计算进行召回。

**4. 基于深度学习的召回**

深度召回模型（如DSSM、YouTube DNN）：通过深度神经网络学习用户和视频的特征向量，并在共享空间中进行匹配。

具体的操作方法：使用DSSM、YouTube DNN等模型分别编码用户和视频，将两者映射到相同的特征空间，计算向量相似度进行召回。

基于序列的召回（如RNN/LSTM/Transformer）：对于用户的观看行为序列，使用RNN或Transformer建模用户的兴趣变化，通过序列模型的隐层表示预测下一步感兴趣的内容。

具体的操作方法：通过Transformer等模型对用户观看历史进行编码，预测用户兴趣的演变，并推荐符合其兴趣的内容。

**6. 基于热点和新内容的召回**

热点内容：召回系统可以根据当前平台的热点内容推荐给用户，保证视频的时效性和关注度。

具体的操作方法：实时监控视频的播放、点赞和评论等行为，识别出平台的热点视频，向用户推荐。

冷启动推荐：对于新用户和新视频，通常召回一些整体流行度高的视频，帮助新用户和新视频快速建立初始曝光。

具体的操作方法：对新用户推荐全站热门内容或新用户群体普遍感兴趣的内容；对新视频设置特定曝光规则以观察用户反应。

**7. 混合召回**

多种召回方式结合：同时使用多种召回策略（如内容、协同过滤、深度学习等）进行召回，综合各策略的优势，保证召回内容的多样性。

具体的操作方法：可以设定不同召回策略的比例或优先级，例如在结果中按一定比例选取内容召回、协同过滤召回的结果进行综合推荐。

## 2.3 规则模型

基于热度的短视频推荐主要是通过对视频的观看、点赞、评论、分享等互动行为进行综合评分，评估视频的受欢迎程度，从而将最受欢迎的视频优先推荐给用户。以下是基于热度的推荐方案及具体实施步骤和计算方案：

1. 热度评分指标

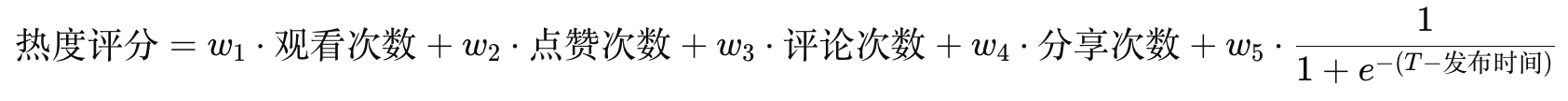
首先，定义哪些行为特征可以用来衡量视频的热度。通常使用以下几个指标：

* **观看次数**（views）：观看次数越多的视频一般热度较高。
* **点赞次数**（likes）：用户喜欢的视频更受欢迎。
* **评论次数**（comments）：评论数高的视频往往具有较高的互动性。
* **分享次数**（shares）：分享次数高的视频通常更具传播性。
* **视频发布时间**（time）：时间越接近当前的视频往往具备更多时效性。

2. 热度评分公式

采用线性加权模型，结合上述指标计算视频的热度评分：

线性加权模型是将不同行为特征按权重组合计算热度分数：



其中，w1,w2,w3,w4,w5 为每个特征的权重，需要根据业务需求进行调优，T为当前时间。

3. 计算视频热度评分

依据上述热度评分公式，逐条计算视频的热度分数，得到视频库中的所有视频的热度排序分数。

4. 定期更新热度评分

由于视频热度会随用户的互动行为和时间变化，通常定期（如每小时或每天）重新计算视频的热度评分，并对热度较高的视频进行缓存。

5. 筛选和排序

* **过滤低质量内容**：剔除违反规定或评分较低的视频。
* **排序**：对筛选后的候选视频按热度评分降序排列，将热度高的视频优先推荐给用户。

6. 多样性调整

为避免推荐列表过于单一，定期调整热门推荐池的视频内容，确保不同种类、不同主题的视频都能有机会展示。

## 2.4 冷启动

1、新用户冷启动策略

离线统计约20%用户没有有效播放序列，针对这部分用户做冷启动

**热门视频**

将当前平台上的热门视频推荐给新用户，利用大众喜好的视频来增加新用户的活跃度。

**实现**：通过点击率、观看时长、点赞率等指标定期筛选热门视频，作为新用户的冷启动推荐候选集。

**分类代表视频推荐**

从各类视频中选出代表性视频，为新用户提供多种内容选择。

**实现**：在推荐池中预设不同类型的代表视频，如娱乐、科技、教育等，随机展示以观察用户初始偏好。

**基于用户基本信息的推荐**：

根据用户的基本信息（年龄、性别、地区）提供适合特定人群的内容。

**实现**：根据用户的年龄段推荐符合其年龄层偏好的视频，例如年轻用户推荐流行音乐、游戏等内容。

2、新内容冷启动策略

新视频发布时缺少互动数据，导致推荐系统难以判断其质量和用户喜好。以下策略可以帮助新内容快速获得曝光：

**新内容优先曝光**：

将新发布的视频推送给更多用户以收集初始互动数据，并观察其受欢迎程度。

**实现**：在推荐池中加入新视频标签，优先在部分用户的推荐流中曝光一段时间，获得其初始观看和点赞数据。

**插入式曝光**：

在热门视频或推荐流中插入少量新内容，使其在流量大的页面获得更多点击和观看。

**实现**：随机选择一部分用户，在其推荐流中以一定频率展示新视频，观察其互动率。

**基于内容的相似性推荐**：

根据新视频的标签、关键词或多模态内容特征，将其推荐给观看过相似内容的用户。

**实现**：将新视频与已有视频的标签、分类特征匹配，推荐给该类内容活跃度较高的用户群体。

**利用创作者历史数据**：

若新视频来自平台上的活跃创作者，可根据创作者的历史视频表现预测新视频的推荐潜力。

**实现**：根据该创作者过往视频的表现（如观看率、完播率），给予新视频相应的曝光机会。

## 2.4 排序模型优化

**1、特征工程优化**

**行为序列增强**：

在DIN模型中，可以增加更多的行为特征，例如用户的观看停留时长、最近的互动频次、点赞、分享序列等，丰富用户的行为序列信息。

**上下文特征**：

将时间、地点等上下文特征加入模型中，使推荐结果更符合用户场景需求。例如，早上和夜间可以推荐不同类型的内容。

**交叉特征构造**

特征交互：通过特征交叉或者特征构造（如用户类别交叉项）生成更丰富的特征

用户特征和视频特征交叉：例如，将用户的偏好标签（如体育、音乐）和视频的标签交叉生成新特征。这样可以更精准地捕捉用户兴趣与视频类别的匹配度。

示例：用户偏好“音乐”和视频标签“流行音乐”交叉生成“音乐\_流行音乐”特征。

时间和用户行为交叉：在不同时间段（如早上、晚上）用户的观看偏好会有所不同，交叉生成新的时间行为特征可以更好地捕捉时段偏好。

示例：早上观看时长较短的视频生成“早晨短时观看”特征，晚间观看较长的视频生成“夜间长时观看”特征。

One-Hot编码交叉：适用于类别特征，如用户年龄段与视频类别的交叉。对特征进行One-Hot编码，生成交叉特征后，再将其进行组合编码。

Embedding向量交叉：将类别特征转化为Embedding向量，然后对Embedding向量进行加法、乘法或拼接操作，生成新的特征。这种方法可以有效利用低维特征表示，并避免维度爆炸。

操作：对用户兴趣和视频类别分别做Embedding，再对两个Embedding进行点乘或拼接，得到新的交互特征。

**2、模型融合及优化**

**多模型融合**：基于DIN和GBDT的现有框架，可以引入更多模型进行融合，例如DeepFM、Transformer或LightGBM。这些模型各有侧重，融合可以有效利用不同模型的优势。

**模型集成（Ensemble）**：通过模型集成技术，比如模型加权平均、Stacking或Boosting，结合DIN和GBDT模型的输出，提高排序的多样性和准确度。

**DIN结构改进**：可以尝试引入Attention机制来增强用户行为与目标视频之间的关联建模，或替换为DIEN（Deep Interest Evolution Network）以更好捕捉用户兴趣的变化。

**GBDT模型优化**：使用更轻量、高效的树模型，如LightGBM或CatBoost，并在调参上进一步提升模型的表现。可以通过Grid Search或Bayesian Optimization进行超参数调优。

**3、在线学习和实时反馈（如果有实时数据）**

增加在线学习模块，实时获取用户最新的行为数据（例如点击、点赞、评论），并对模型进行微调，以提升实时性和个性化效果。

**Bandit算法**：结合强化学习或Bandit算法，在用户反馈的基础上动态调整推荐策略，使推荐内容更具个性化。

# 附件：历史本版

[docs短视频历史效果统计](https://www.kdocs.cn/l/cjmC5PLLDPVS)

# 3、实施细则

## 2.1 数据优化实施（1）

目前数据存在一定缺失情况，期望增加的数据有：

1. 用户画像数据：
   1. 基本信息：年龄、性别、职业、地区、入网时间、套餐价格、当前剩余流量、上个月使用流量
   2. 设备相关：使用设备类型、网络状况（是否5G、是否wifi）
2. 视频相关数据：发布时间
3. 行为相关数据：是否点赞、是否分享、操作时间是否周末、操作时段（**凌晨**（00:00 - 06:00）；**早晨**（06:00 - 09:00）；**上午**（09:00 - 12:00）；**中午**（12:00 - 14:00）；**下午**（14:00 - 18:00）；**晚上**（18:00 - 21:00）；**深夜**（21:00 - 00:00）；）
4. 实时数据：近1、5、10分钟点赞、近1、5、10分钟分享、近10条视频停留时长与视频时长占比、近10条视频停留时长、近10条视频类型

## 2.2 召回优化实施（3）

**现状逻辑**

一、目前的召回视频必须是固定比例，置顶视频:营销视频:新视频:其他视频=1 : 1 : 2 : 1

二、根据时效性问题，现在找回的视频数量，保持在25个。

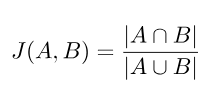
三、在每个视频类别下面，当前的召回排序业务逻辑是否必须保留

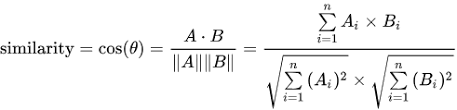
当前的排序逻辑：

* 1. 置顶视频，有置顶视频配置
     1. 全网置顶top5
  2. 营销视频，有营销类型排序
     1. 除去置顶视频
     2. 筛选：用二级标签按5G业务＞流量＞权益＞会员＞新业务＞套餐＞终端＞宽带
     3. 用户标签ID（user\_label）和视频客群ID（labelId）有交集的情况排序优先级最高（高于5G业务）
     4. 排序：取前5个
        1. 地市>分省>全网
        2. 已浏览时间距现在时长；（video\_history\_d>video\_history\_h>video\_history\_k>video\_history\_r)
        3. label>5G业务＞流量＞权益＞会员＞新业务＞套餐＞终端＞宽带
  3. 新视频，有发布时间排序
     1. 除去置顶、营销视频
     2. 筛选：发布时间是近7天
     3. 排序：取前10
        1. 地市>分省>全网
        2. 已浏览时间距现在时长；（video\_history\_d>video\_history\_h>video\_history\_k>video\_history\_r)
  4. 其他视频
     1. 筛选：除去置顶、营销、新视频
     2. 排序：取前5
        1. 地市>分省>全网
        2. 已浏览时间距现在时长；（video\_history\_d>video\_history\_h>video\_history\_k>video\_history\_r)

【**优化逻辑】**

* 1. 如果保留，那召回这边就没有可优化的余地了
  2. 如果可以不保留，就可以提供一个较大的带选集，开发召回算法
     1. 视频-视频协同过滤
        1. 创建用户- 视频有效播放矩阵
        2. 根据下图公式计算相似度
        3. 生成协同过滤矩阵





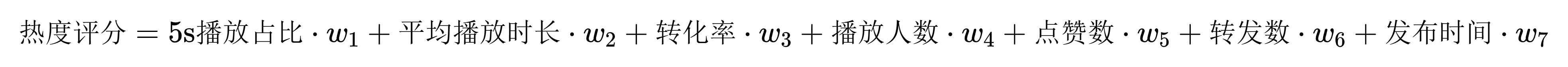
* + 1. 视频内容标签匹配（暂定）

视频内容标签匹配：根据用户对某些内容标签（如搞笑、教育、体育等）的偏好，召回具有相同或相似标签的视频。

具体的操作方法：为视频和用户建立标签向量，基于标签相似度（如余弦相似度）进行推荐。

## 2.3 规则优化实施（冷启动实施）（2）

热度计算考虑优化目标和热度因素



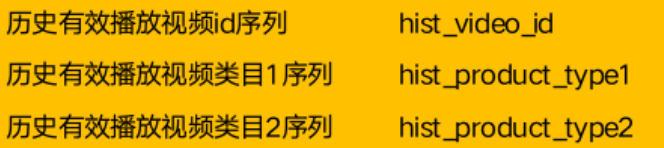
热度计算出来之后，在冷启动时使用，也就是说没有历史行为的用户，直接按照热度推荐

解析报文，判断用户行为序列是否为空。为空，以热度进行排序返回

## 2.5 排序模型优化实施（4）

一、**行为序列增强**

1、当前的行为序列构造情况：



2、增加行为需求构造数据：

1. 近一个月点赞的视频序列
2. 近一个月分享的视频序列
3. 用户播放的作者序列（经测试，作者数据空置率太高，不太适合做序列构造）

# 2024/11/20 工作讨论

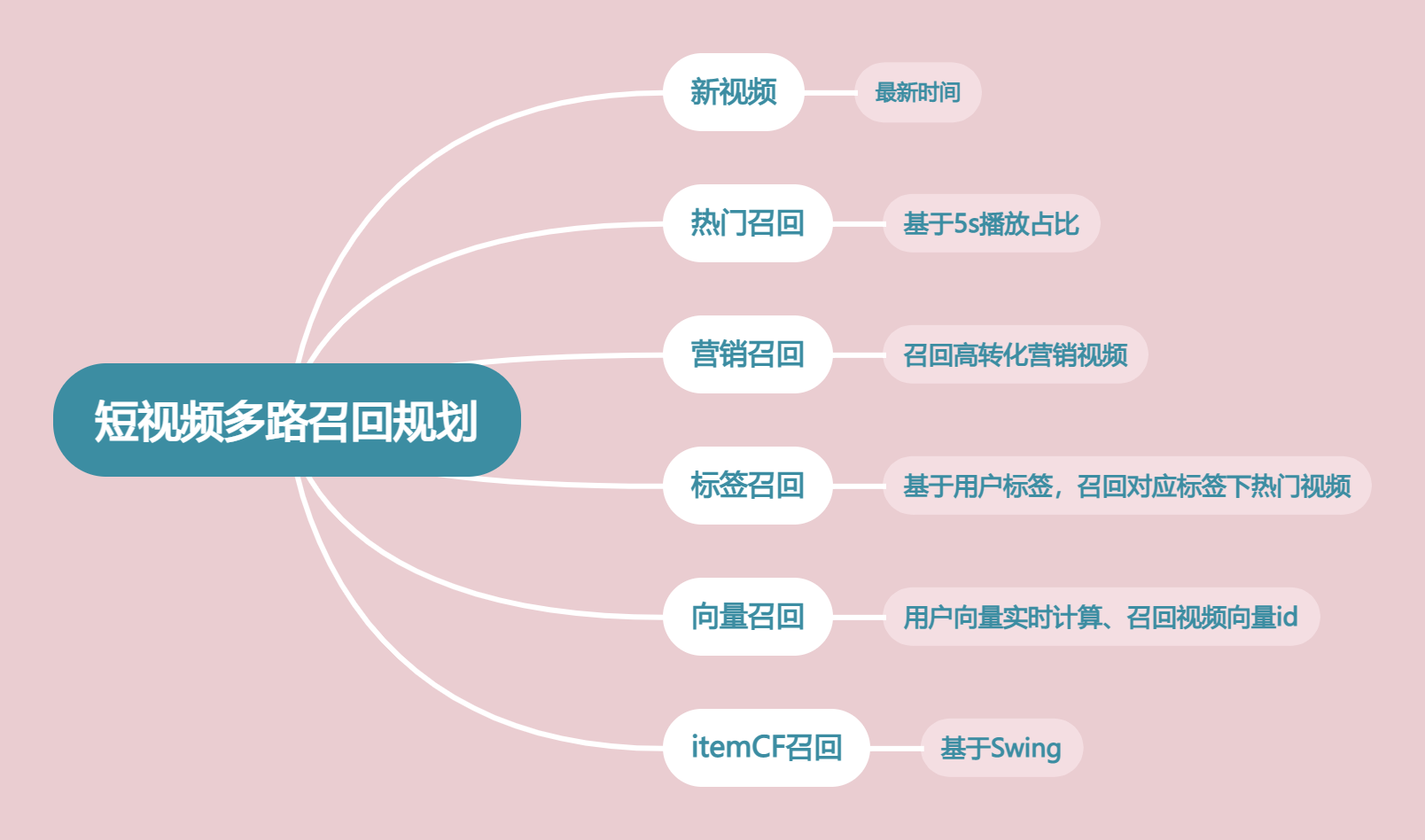
1. 模型耗时排查：给出排查问题方向、方案、日志添加、周四上线。 昆鹏 + 吴威龙
   1. 模型服务输出各模块耗时时间、pod IP，在工程服务统一打印
   2. 减少ES召回数据量，比如召回1000改为500或800
      1. 线上只召回营销视频(500个以下)比召回1000个耗时少80ms左右
   3. 增加ES召回条件限制（使召回量减少的条件）
      1. 多路召回测试缓存各省视频ID，带着ID召回可以减少ES召回耗时
   4. 模型服务传参，产品数据和用户数据分开不拼接
      1. 已区分用户和视频数据【排除
   5. 召回服务、预测服务合并，减少召回和预测之间的http请求
      1. http请求响应耗时8-10ms，【排除
   6. CPU资源不足
   7. 序列长度影响排除
   8. 报文解析耗时排除
   9. 数据标准化耗时排除
   10. 均衡负载正常
2. 召回优化：优化目的：（1）性能优化，目前召回耗时200ms 降低到100ms （2）候选集太少、候选集个性化程度低。 召回优化方向：多路召回、先从es召回视频、排序前从redis获取特征
   1. 验证工作：es 多路召回性能验证、 redis 获取特征性能验证 昆鹏
      1. es 多路召回性能验证
         1. 新视频、热门视频、标签召回三种耗时最低115ms，最高200ms以上【云桌面测试】
   2. 召回方案的细则完善：昆鹏 + 吴威龙
      1. 新视频：字段is\_7d\_update 近7天是否更新，召回is\_7d\_update=1一周内视频数据100条
      2. 热门视频：字段cur\_5s\_video\_ratio 5秒播放比例，统计cur\_5s\_video\_ratio中位数，取高于中位数视频数据100条
      3. 营销视频：字段is\_market\_video 是否营销视频，取is\_market\_video=1，统计cur\_play\_video\_cnt播放次数中位数，取高于中位数视频数据100条
      4. 标签召回：字段firstLevalTagId 一级标签、tagIds二级标签、cur\_5s\_video\_ratio高于中位数，100条
      5. 协同过滤召回：算法基于swing生产协同过滤矩阵，每个种子（主商品）保留最多10个候选相似商品。生产后数据格式 A |“B:0.6,C:0.4,D:0.3“【按照分数排序完成后，字符串格式】 【武威龙 11.22号完成】 A为key 研发协助将数据写入 redis 线上架构处理用户行为特征时候，直接去redis获取相关视频传给模型侧，模型侧拿到id 去Es取标签特征。
         1. 架构具体召回逻辑：获取用户行为序列【字段名: vdu\_feature\_info】，按照时间排序，取最近的N次【可设置】有效播放行为（大于5s），召回行为对应的相似物品，传输给模型侧数据格式 字段名称【behavior\_cov\_recall】[{oper\_time（行为时间）, main\_item(用户行为的id，上例中A), citems(相似的物料集合，上例中B:0.6,C:0.4,D:0.3 )}， {}] 【jsonArray】 如果没有召回结果返回 []

例：[{"oper\_time":"2024XXXX", "main\_item":"15165", "citems":"151:0.5,1567:0.15"}]

* + - 1. sftp 42 /download/video/
    1. 召回架构注意：每一路召回用一个recallTag =1/2/3/4 字段表示，可以设置优先级 优先级表示merge合并时候优先保留该路召回。recallTag字段作为商品一个属性字段添加。

|  |  |  |
| --- | --- | --- |
|  | recallTag | recall-number |
| 营销 | 5 | 10 |
| 协同过滤召回 | 1 | 40 |
| 新视频 | 2 | 20 |
| 标签 | 4 | 10 |
| 热门视频 | 3 | 20 |

* + 1. 可设置项：每一路召回的召回个数、整体召回保留个数、召回merge优先级 merge\_tag\_sort=[5,1,2,,4，3】、协同过滤每个主商品可召回的相似商品个数。
  1. 召回要保证类目多样性 方案】



1. 模型效果不佳分析：
   1. 亚信、轻渠道的视频5s 播放比例领先较大，分析亚信、轻渠道 用户播放视频的热门视频、我方是否有正确召回 是否能正常进入候选集
   2. 对方营销视频比例的播放比例是否比我们少很多
   3. 置顶视频的播放比例
   4. 人均播放类目数对比

[关联表]