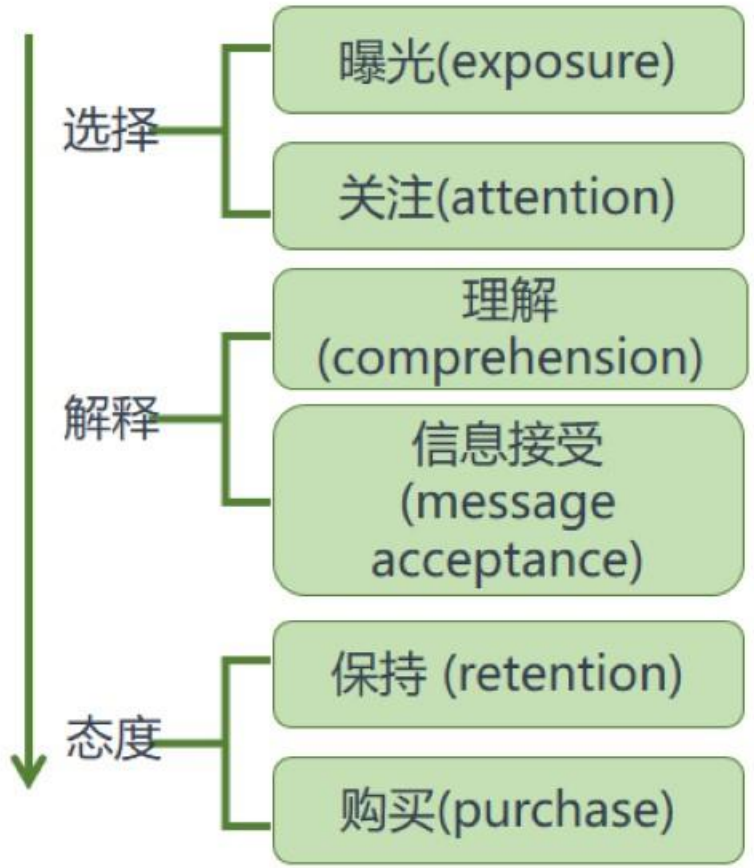
推荐细节

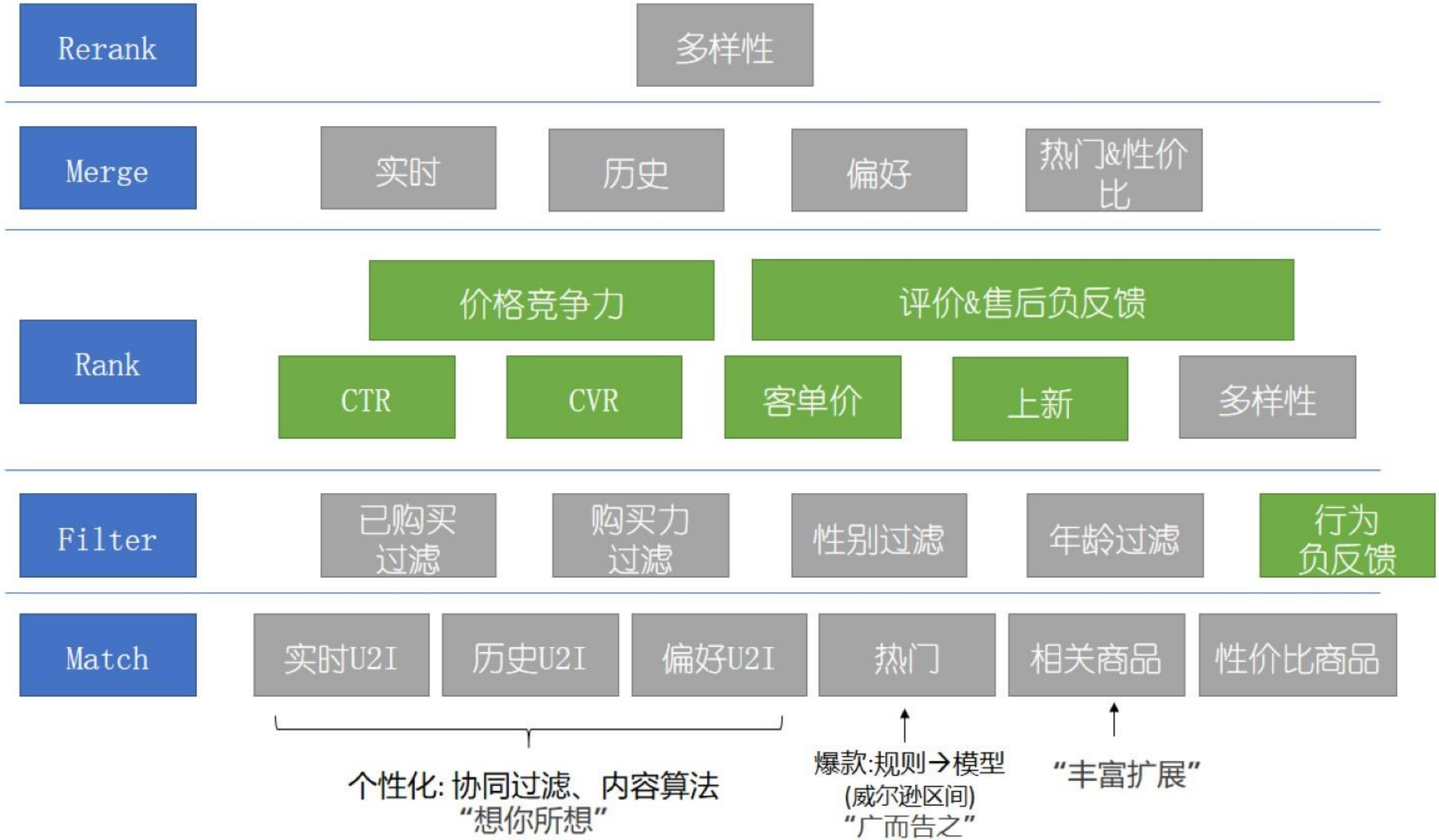
姚凯飞

2017.7.2

用户交互阶段模型


选择
曝光(exposure)
关注(attention)
解释
理解
(comprehension)
信息接受
(message
acceptance)
态度
保持 (retention)
购买(purchase)

推荐各层常用策略


Rerank
多样性
Merge
实时
历史
偏好
热门&性价比
Rank
价格竞争力
评价&售后负反馈
CTR
CVR
客单价
上新
多样性
Filter
已购买过滤
购买力过滤
性别过滤
年龄过滤
行为负反馈
Match
实时U2I
历史U2I
偏好U2I
热门
相关商品
性价比商品
个性化: 协同过滤、内容算法
“想你所想”
爆款:规则→模型
(威尔逊区间)
“广而告之”
“丰富扩展”

召回

召回模块生成的多种类型的推荐候选集:

- User profile标签索引列表
- ★ 相似列表(协同过滤、Content-Based、基于图论的算法、Knowledge-Based、Context-Aware、Hybrid-Based)
- 热门列表(分类热门/运营人工推荐列表)
- 召回列表将会作为推荐候选池

召回---trigger

召回四象限

- 流行
- 多样
- 新鲜
- 相关

- 上下文相关
  - 与上下文内容相关
  - 与上下文发布者相关
- 行为相关
  - 协同过滤
  - 实时反馈
- 热门
- 兴趣相关
  - 基于关系的兴趣
  - 基于消费行为的兴趣
  - 长期兴趣 & 短期兴趣
  - 人群兴趣

CF问题

$$w_{ij} = \frac{|N(i) \cap N(j)|}{|N(i)|^{1-\alpha} |N(j)|^{\alpha}}$$

$$w_{ij} = \frac{\sum_{u \in N(i) \cap N(j)} \frac{1}{\log 1 + |N(u)|}}{\sqrt{|N(i)||N(j)|}}$$

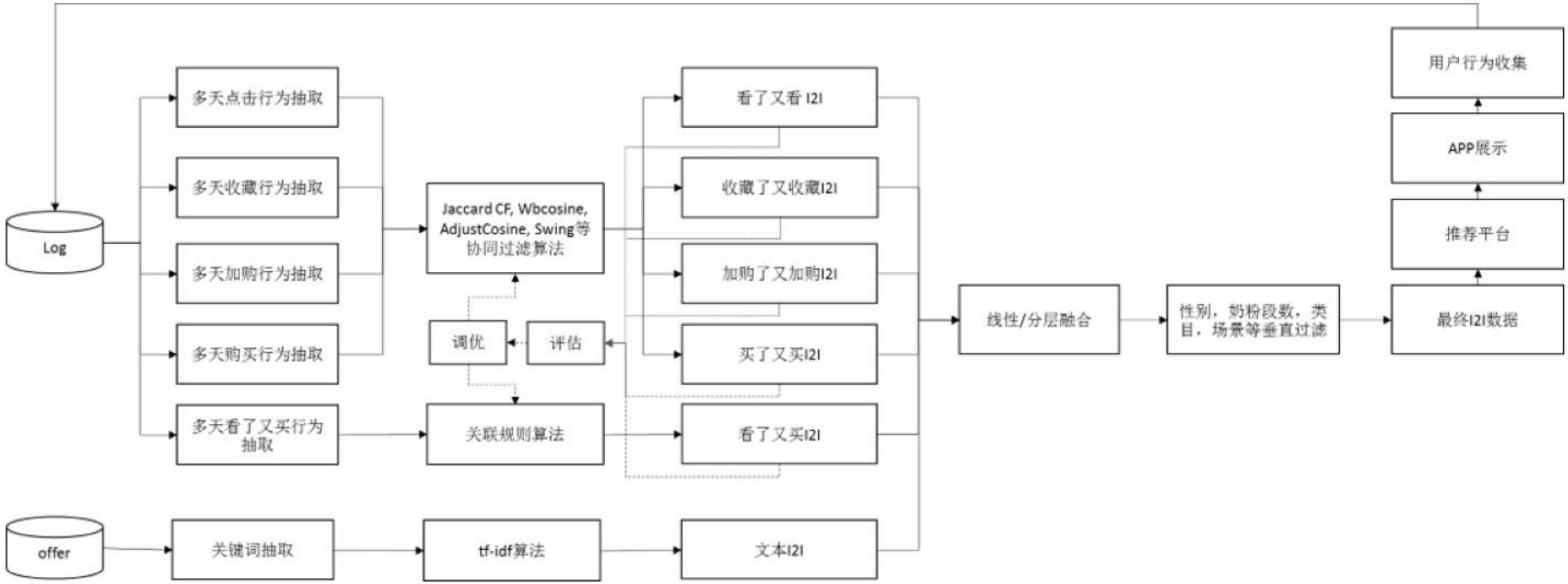
1.热门Item

2.过度活跃用户

3.User Base CF 的第一推动力问题(第一个用户从哪儿发现新的物品)

4.Item Base CF 物品冷启动

实际落地技术方案



- 提取**单天频次、历史频次、单天共现、历史共现**训练数据，并考虑时间衰减，强化近期行为
- 引入webcosine, adjust-cosine, swing等复杂模型进一步提升算法的准确率
- 设计了准确率，召回率，F1，覆盖率，零结果率，深度等多项离线指标，并同步开发了自动化评测工具和DEMO中心，离线评测更加高效和多样化
- 线性、分层等多种融合策略相结合，参数自动学习，最大化优化目标
- 针对行业特性，引入垂直过滤模块，比如性别，奶粉段数，类目限制等规则系统

混合的推荐机制

这里讲几种比较流行的组合方法。

1.加权的混合（Weighted Hybridization）

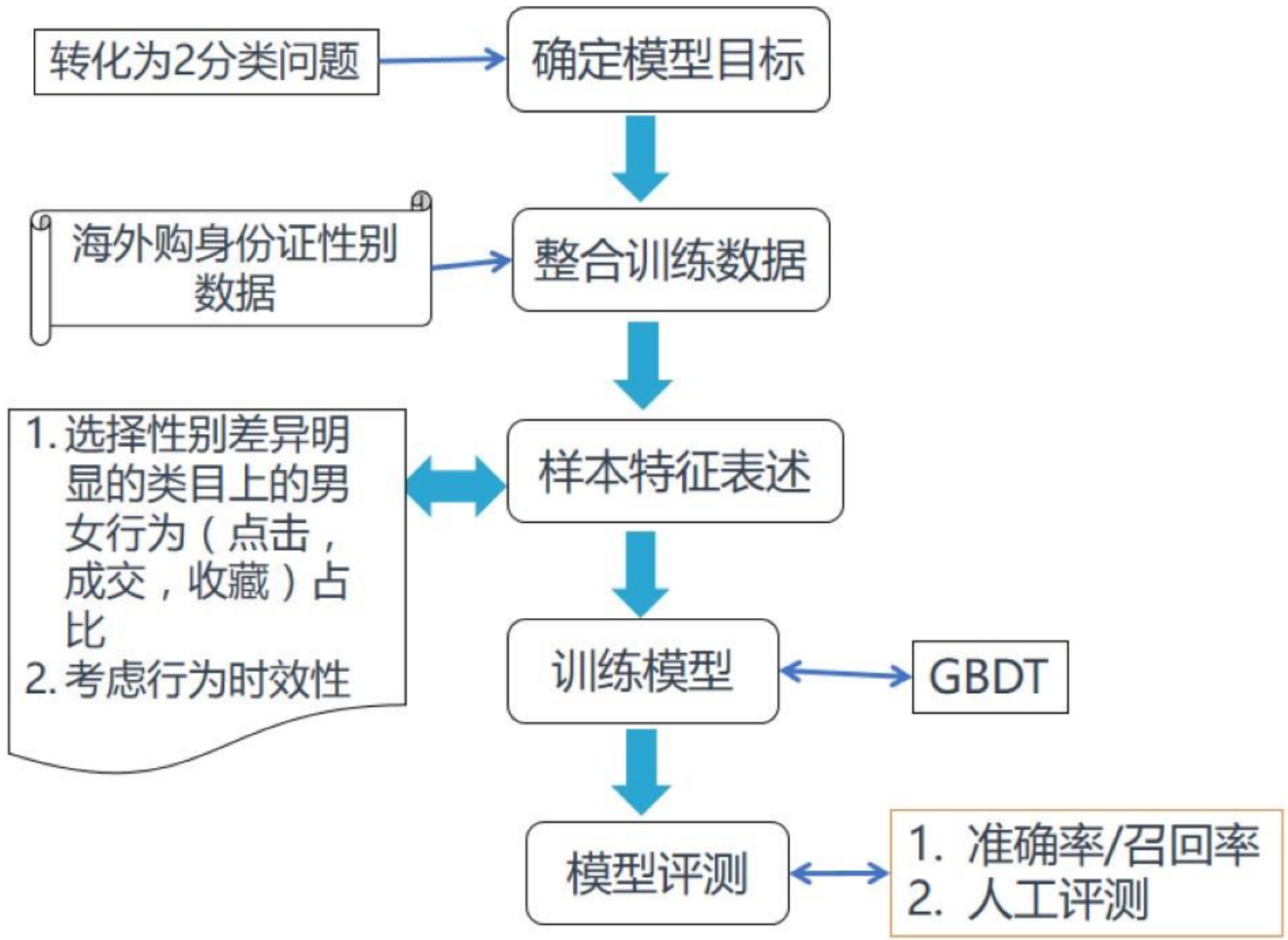
2.切换的混合（Switching Hybridization）

3.分区的混合（Mixed Hybridization）

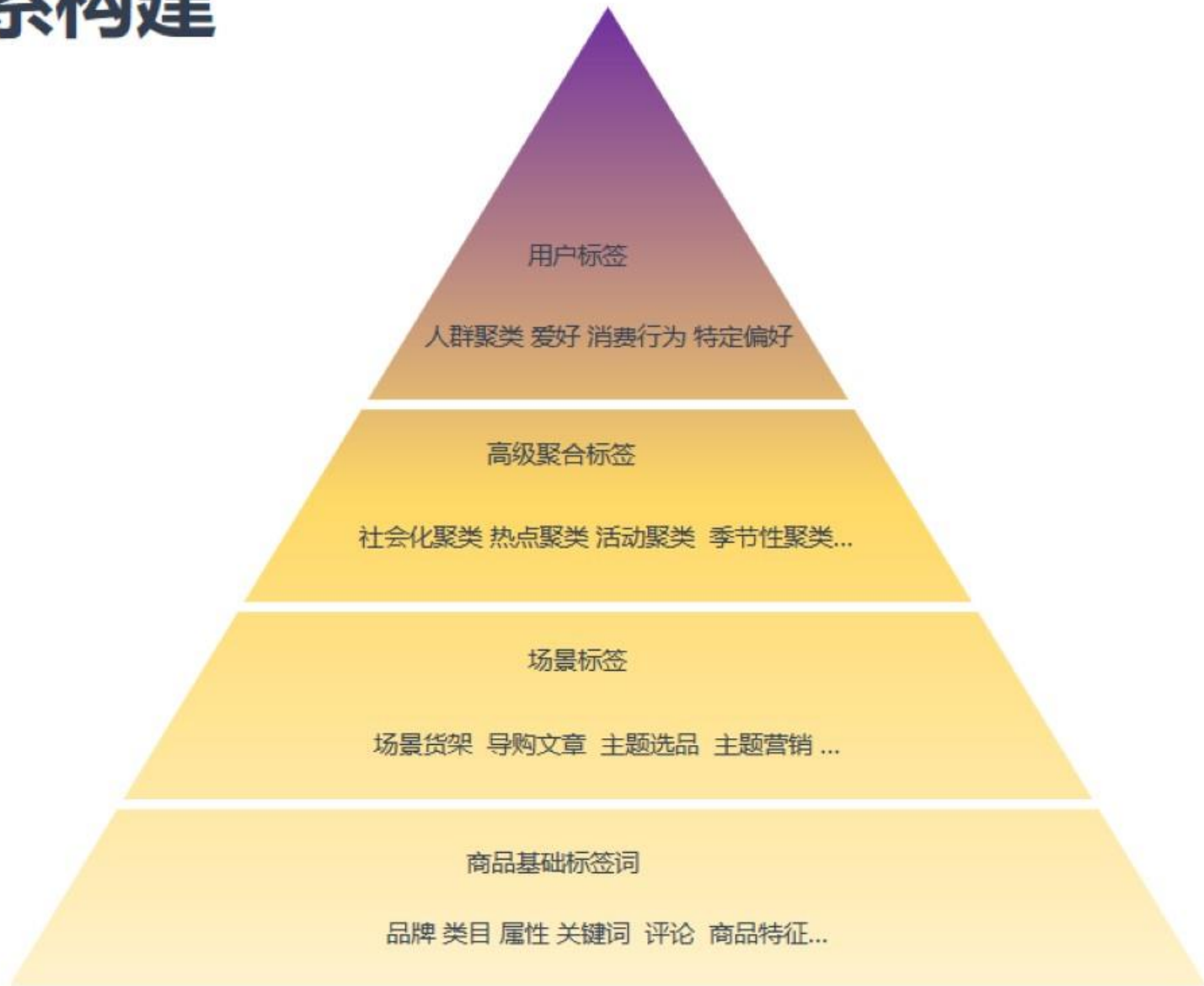
4.分层的混合（Meta-Level Hybridization）

用户画像

性别模型


转化为2分类问题
确定模型目标
海外购身份证性别数据
整合训练数据
1. 选择性别差异明显的类目上的男女行为（点击，成交，收藏）占比
2. 考虑行为时效性
样本特征表述
训练模型
GBDT
模型评测
1. 准确率/召回率
2. 人工评测

标签体系构建


用户标签
人群聚类 爱好 消费行为 特定偏好
高级聚合标签
社会化聚类 热点聚类 活动聚类 季节性聚类...
场景标签
场景货架 导购文章 主题选品 主题营销 ...
商品基础标签词
品牌 类目 属性 关键词 评论 商品特征...

标签建模

关键词偏好

- 通过用户点击，购买，成交，收藏的商品的标题挖掘用户所关注的语意单元的信息
- 方法：
  - 基于历史商品标题分词粒度的 TF-IDF 统计模式
    - 问题：
      - 粒度太细，用户在单个词上难有长期偏好
      - 词太多，存储空间大
      - 页面展示效果较凌乱
  - 建立基于<user-商品>原始统计为基础的topic model 的解决方案
    - PLSA
    - LDA
    - 人工review topic下的词

特征工程

- 特征预处理:
  - 归一化
  - one-hot
  - 缺失值补充
  - 异常值去除
  - 数据变化

- 特征类型:连续/离散/周期型/二维联合
- high/low level 特征
  - 候选集自带的特征，cf相似度，文本相关性分数
  - 用户类:
    - 人口统计学特征:年龄/性别/收入
    - 类目/品牌偏好
    - 兴趣标签
    - 终端类别
  - 商品类
    - 类目/品牌
    - 标题/描述
  - 上下文
    - 时间
    - 位置
  - 组合特征候选类:2^n

特征处理

特征预处理、数据清洗是很关键的步骤，往往能够使得算法的效果和性能得到显著提高。

- 特征离散化(性别年龄等)、ID类特征
  - 加快处理速度
  - 非线性
- 特征平滑
  - 威尔逊区间
  - PV越小，CTR的置信度越小
  - 防止低pv的商品占优势
- 特征组合
  - 非线性
  - PV+IPV组合，比CTR的信息更多

特征聚合

- 特征降维：特征维度高、稀疏、误差大
- 相似特征有相似的权重
- 特征的权重近似于后验概率

Low Level特征，High Level特征：

- Low Level 比较有针对性，单个特征覆盖面小（含有这个特征的数据不多），特征数量（维度）很大。
- High Level比较泛化，单个特征覆盖面大（含有这个特征的数据很多），特征数量（维度）不大。长尾样本的预测值主要受High Level特征影响。
- 高频样本的预测值主要受Low Level特征影响。
- 对于访购率问题，有大量的High Level或Low Level的特征

Item2vector

- 类似word2vector思想。
- Word2vec利用文本中每个词的前后依存关系来建模，并将word映射为一个向量。
- Item2vec前后item的依存关系是由每个用户在网站的交互item列表构建。

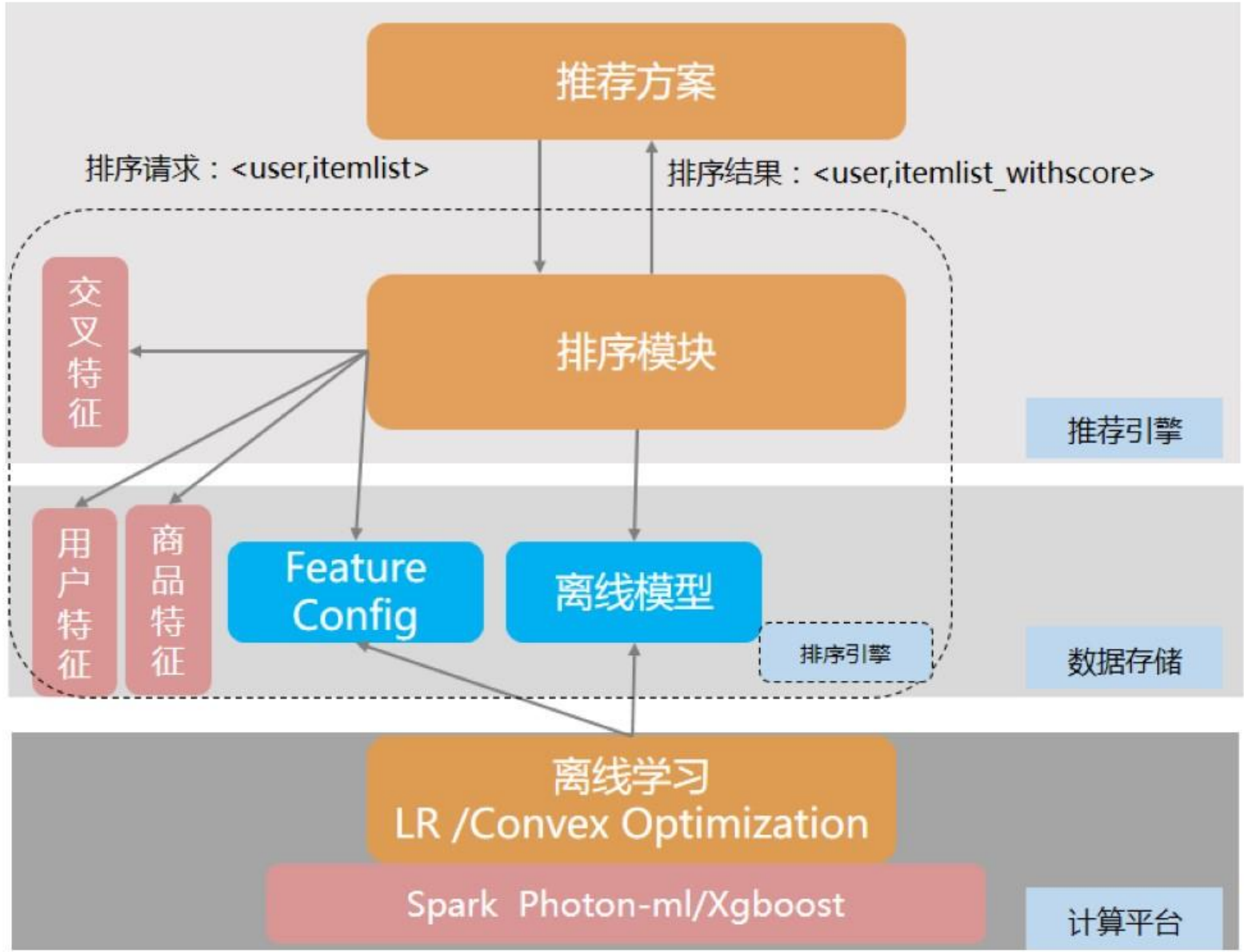
点击率预估问题

- 模型校正
- 静态/动态特征，静态/动态模型
- 位置偏置
- 平滑(浏览器/时间。。。)
- E&E（无过多反馈数据—新商品，新标签，新商家）
- 由于数据的稀疏性，转化预估一般比点击预估更难

预估部署

预估模型中的平滑---类似利用置信度的平滑

CTR/CVR预估框架



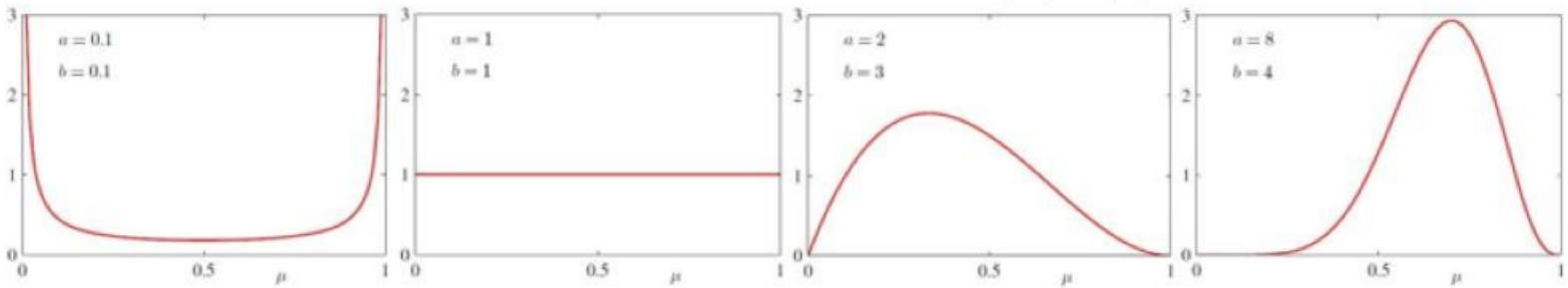
点击反馈的平滑

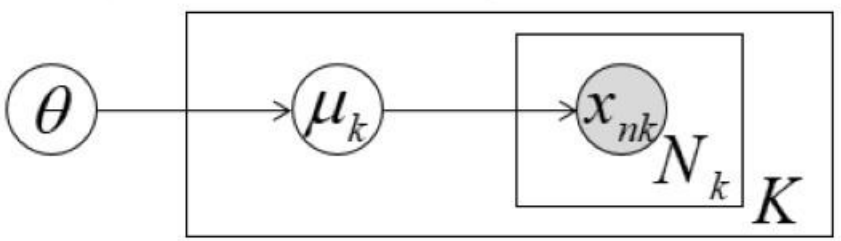
- 问题: 在数据稀疏的情况下较稳健地估计CTR或COEC
- 经验贝叶斯方案
  - 点击产生概率模型(Binomial分布, 其中μ为点击率):

$$p(x \mid \mu) = \mu^x (1-\mu)^{1-x}$$

  - 视μ为随机变量, 采用Beta分布共轭先验进行regularization:

$$p(\mu \mid \theta = \{a,b\}) = \mathrm{Beta}(\mu; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1}$$





排序几点考虑：

1） 模型融合：

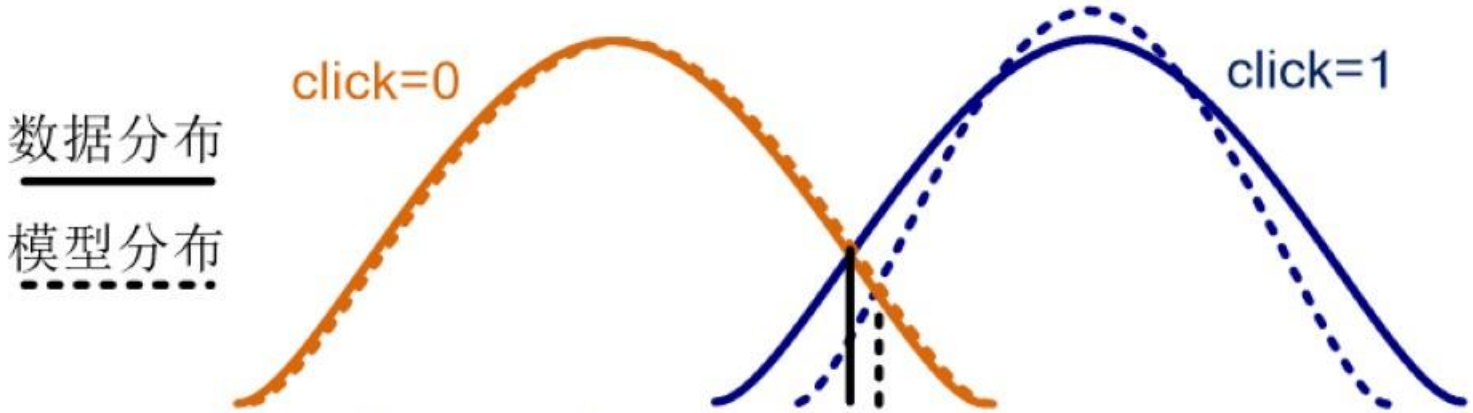
针对不同的召回策略，主要秉持“短期行为优先，兼顾长期行为”，“保证多样性，避免过度个性化”原则。算法上主要通过**埋点模型训练学习**，**长短期用户行为密度建模**，**正负反馈反向调节**等手段来实现。

2） 转化能力：

通过对用户，商品，会场等多维度采集特征数据，训练点击转化和交易转化模型。大促期间根据不同的阶段，采取差异化的处理策略。比如**预热期强化点击转化**，**正式售卖期则强调交易转化。**

正负例不均衡情况下的修正

- 正负例不均衡时，逻辑回归最大似然解是有偏的



- 参数w的偏差可以估计如下并加以修正

$$\tilde{w} = \hat{w} + \left(X^{\mathrm{T}} W X\right)^{-1} X^{\mathrm{T}} W \xi$$

- 进一步考虑 $\mathrm{var}(\hat{w})$ 可得ctr的Bayesian预测

$$p(\mathrm{click}=1) \approx (0.5 - \tilde{\pi}_1)\tilde{\pi}_1(1-\tilde{\pi}_1)x_0 \,\mathrm{var}(\tilde{w})x_0^{\mathrm{T}}, \tilde{\pi}_1 = \sigma\left(\tilde{w}^{\mathrm{T}} x_0\right)$$

点击价值估计

- 挑战:
  - 非常稀疏的训练数据
  - 与商品类型强烈相关的行为模式
- 点击价值估计若干原则
  - 模型估计时, 用较大的bias换较小的variance, 已达到稳健估计的目的.
  - 充分利用广告商类型的层级结构，以及转化流程上的特征

探索与利用 (E&E)

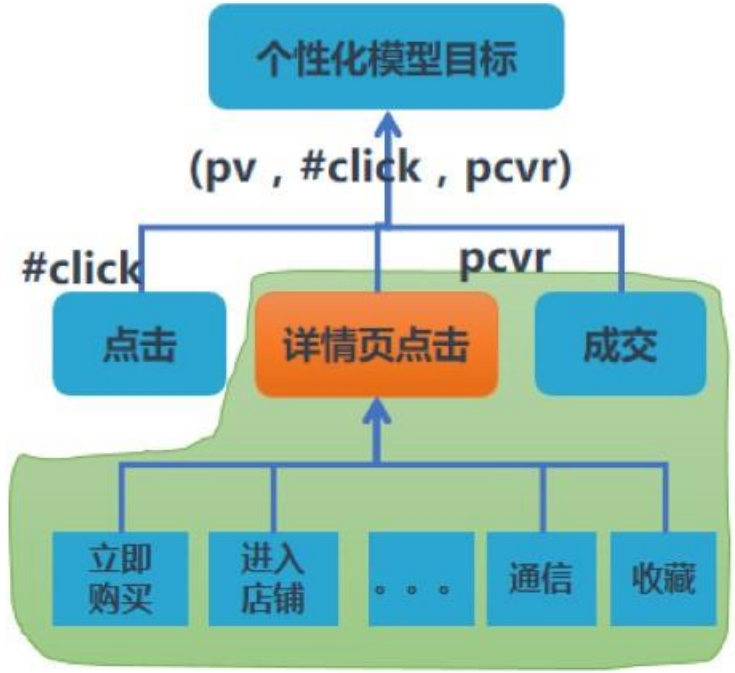
- 问题
  - 为长尾的(a, u, c)组合创造合适的展示机会以积累统计量，从而更准确地估计其CTR
  - 提升整体的收入，即需要严格控制探索的量和有效性
- 方法思路
  - 通常描述为Multi-arm Bandit (MAB) 问题
    - 有限个arms(或称收益提供者)a, 每个有确定有限的期望收益$E(r_{t,a})$
    - 在每个时刻t, 我们必须从arms中选择一个, 最终目标是优化整体收益
  - 基本方法为ε–greedy: 将ε比例的小部分流量用于随机探索
- 主要挑战
  - 海量的组合空间需要被探索
  - 各个arm的期望收益是动态变化的

E&E 算法 - UCB

- 方法思路
  - 在时间t, 通过以往的观测值以及某种概率模型, 计算每个arm的期望收益的upper confidence bound (UCB)，并选择UCB最大的arm
  - 我们不可能一直选择非最优的arm, 原因是我们选择的此arm次数越多, 其UCB就越接近于其期望收益
- 具体UCB策略
  - β-UCB策略: 依一个很大的概率, 我们选择非最优arms的次数存在着一个上界, 该上界与总的选择次数无关
  - UCB-tuned策略: 我们已选择的次数越多, 就越可以自信地抛弃不太有前途(但仍有可能最优)的arm.

目标变形

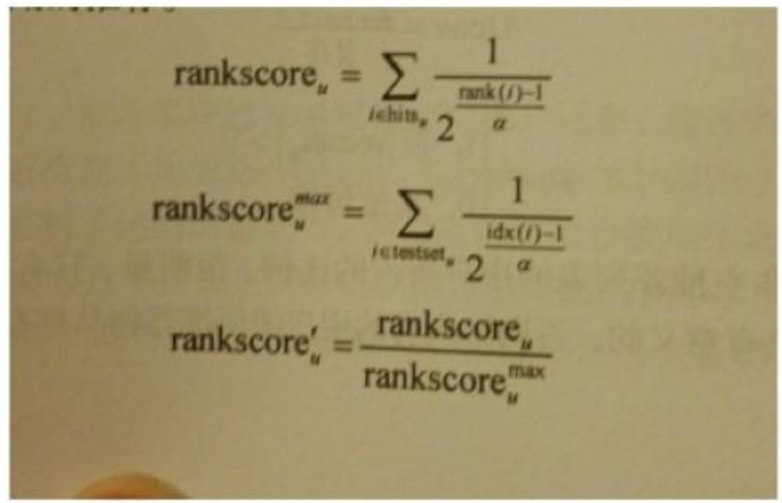
- **购物链条上的丰富用户行为**
  - 推荐页上的点击
  - 商品详情页上的行为
    - 显示： 发生的各种各样的点击行为
    - 隐示：停留时间
  - 最终的成交行为



- **排序的优化目标**
  - 量化购物链条的pre-click和post-click行为

<user-商品>对目标 ：

$$\frac{\left(\#click+\frac{1}{\overline{pcvr}}\times\#click\times pcvr\right)}{pv+\left(\#click+\frac{1}{\overline{pcvr}}\times\#click\times pcvr\right)\times\beta}$$



$$rankscore_u=\sum_{i\in hits_u}\frac{1}{2^{\frac{rank(i)-1}{\alpha}}}$$

$$rankscore_u^{max}=\sum_{i\in testset_u}\frac{1}{2^{\frac{idx(i)-1}{\alpha}}}$$

$$rankscore'_u=\frac{rankscore_u}{rankscore_u^{max}}$$

pcvr： $<user-商品>$ 对的预估转化率　$\overline{pcvr}$： $<user-商品>$ 平均预估转化率

Pcvr 的计算通过建立基于详情页点击的行为为特征转化率预估模型求得

评估指标

- 线下
  - 特征有效性分析(相关系数、卡方检验、平均互信息、条件熵、后验概率、逻辑回归权重)
  - 特征权重分析
  - AUC (area under curve)
  - NDCG (Normalized Discounted Cumulative Gain)
  - 分query AUC
  - 分用户+类目AUC
- 线上：A/B test
  - 成交转化率，CTR业务指标
  - Case 查看

数据标注

- 样本关联: 串联展示、点击、下单
- 样本选择:时间窗口(近N月)、过滤(时长/访问次数限制)、样本迁移、skip above(点击模型 +2)
- 样本采样:访购模型:减少负样本(非必要不做采样，采样常常可能使实际数据分布发生变化，但是如果数据太大无法训练或者正负比例严重失调（如超过100:1）,则需要采样解决)
- 样本权重:支付>下单>点击
- 负样本: 热门但未被点击、主动删除的显示负反馈数据(高质量)

推荐1.0—统计方法推动

- 规则
- 热销，爆款模型

推荐2.0--算法推动

- 关联规则
- 协同过滤

推荐3.0—架构通用化

- 引擎(后台生成猜你喜欢)意图
- 画像
- 千人千面
- cf
- 情境(节日/季节)
- 反向(复购时间)
- 主题(商品聚合)

- 自适应召回context/profile/parameter
- search/recom server(federation layer)
  - 意图分析
  - query builder
  - 通信/存储模块
- ranking/predication server

推荐其它资源

- 图像
- 文字

推荐问题

- 多样性
- 稀疏性
- 噪音
- 推荐需要解决的问题
  - 商品更新
  - 商品质量
  - 商品与买家的匹配程度

- 马太效应有者愈有，强者愈强
- 热门商品越热门，新商品很难发现
- 解决
  - 补偿不利位置商品
  - 随机
  - 个性化推荐相关冷启动方法，找到相似的老商品作为参考

- **人类视觉窄化**

推荐的可视化

- demo仿真系统
  - 精确模拟真实的推荐
  - 有效评估算法的改进效果
  - 结果的可解释性
  - 业务、产品的协同
- 实时监控
  - 实时获取调度监控指标
    - 请求量
    - 失败量
    - 平均延时
    - 历史对比
    - 波动告警
  - 及时预警引入人工干预
  - 横向，纵向指标发现异常

推荐实时化

- 实时个性化几个值
  - 时间窗口
  - 观测值的累积效应
  - 新热度评估策略
- 个性化的时效性(当前点击xxx，未来点击xxx的概率)
- 变化是时刻进行的
  - 商品在变化
  - 用户个体在变化
  - 群体/环境在变化
  - 个体和群体的隶属关系也在动态变化
- 实时特征和模型一般使用离线数据(item+user特征)+实时用户数据(user特征)
- 推荐时机
  - 兴趣发现和收敛速度
  - 对于智能程度的感知

运营体系自有类目与品牌的细粒度挖掘

- 以细化类目粒度为例