手机卫士个性化推荐系统实践

团队：手机卫士数据挖掘部
讲师：郭合苍

团队和个人介绍

郭合苍 手机卫士个性化推荐负责人

- 就职于奇虎360手机卫士事业部
- 手机卫士个性化推荐团队负责人
- 多年互联网和推荐系统开发经验
- 专注于个性化推荐算法的应用

课程提纲/内容

- Why - 为什么做推荐
- How - 如何做推荐
- Evaluate - 评价推荐效果
- Summary - 推荐系统总结

WHY - 业务场景

- 信息过载
- 用户体验
- 平台利益

- 大数据
- 低频访问用户
- 入口较深
- 复杂场景

亿级手机卫士用户
千万级日活跃用户

WHY - 挑战

WHY - 多场景


87% 10:46
优化完成
恭喜你,体检已完成
试试下面精彩推荐吧！
京东爆款特价
OPPO R11热力红 京东限量低价购
京东
超薄小蛮腰
上手更惊艳
大众点评
1.7亿人安装
下载
广告 大众点评--发现品质生活！
今日十大热点
155.8万人安装
下载
广告 聊天时没话题吗？今天的热点话题都在这儿！


86% 10:49
75%
游戏加速
恭喜你！内存已完全清理释放
热门游戏
妖精的尾巴-最强公会-高SS...
83.5万人安装
立即试玩
广告 《妖精的尾巴》官方正版：超高SSR爆率
思美人-新时装
122.1万人安装
立即试玩
广告 秦楚争霸燃斗来袭，新国风浪漫动作巨制邀您来战
胡莱三国2
83万人安装
立即试玩
广告 刘涛代言沙盘策略手游，胡莱三国原班人马打造
巅峰战舰-公平竞技问鼎巅峰
277.5万人安装
立即试玩
广告 亚洲霸主崛起，个人军港铸造战舰！
大话西游-小师妹杨紫送祝福
立即试玩
我的游戏
最新游戏


48% 09:57
清理加速
清理垃圾450.2MB
手机加速23%,延长待机时间20分钟
发现大量无用的安装包、照片和视频，占用手机空间，快来清理！
立即清理
看片卡慢？一键提速 快人一步
立即下载
手机剩余存储空间4%，快来释放更多空间！
立即清理
建议开启程序锁，防止微信记录和其它重

WHY - 冷启动

用户冷启动

- 每日新增用户
- 用户访问频率低
- 获取用户信息有限

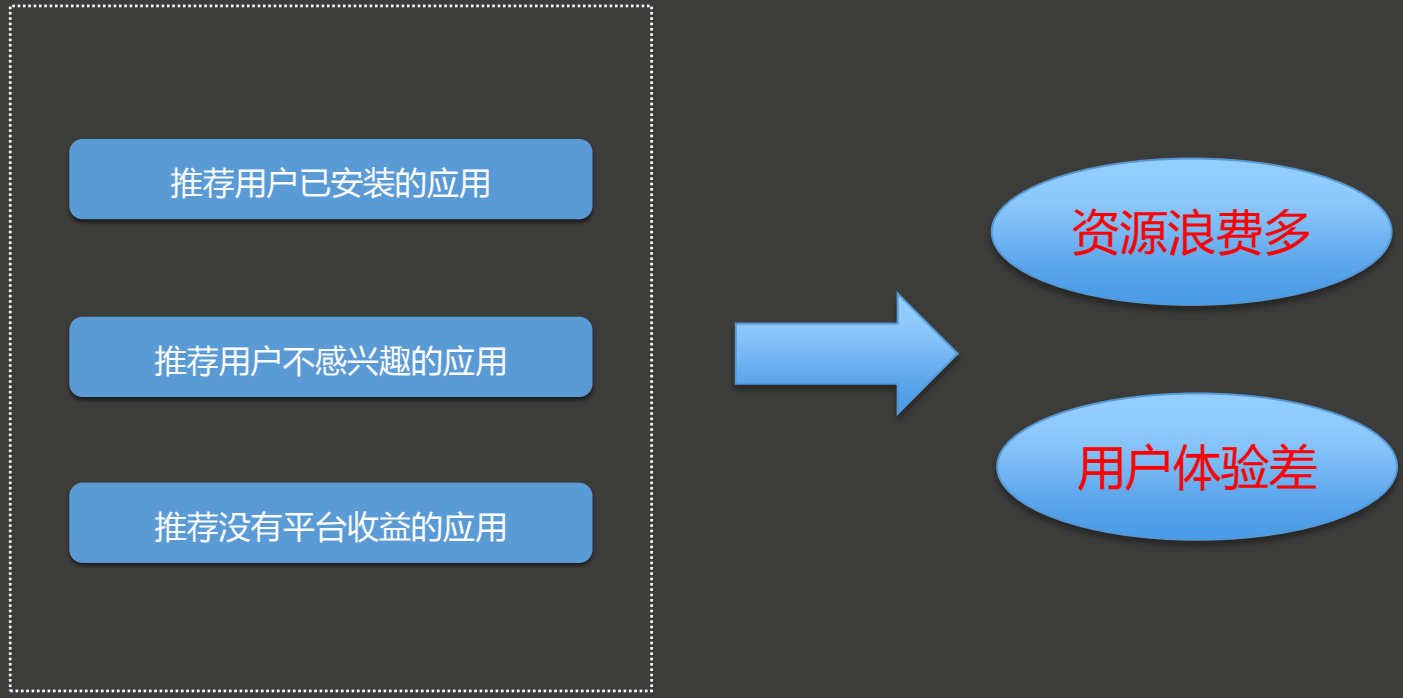
物品冷启动

- 每日新增应用
- 获取应用信息有限

WHY - 个性化需求

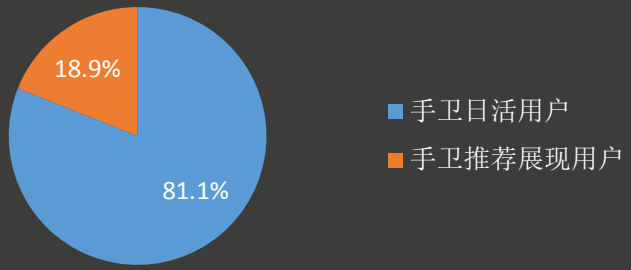
- 用户人群差异
  男性 女性 70后 80后 90后 app列表 app类别 用户标签 用户活跃度
- 手机设备差异
  品牌 型号 内存 sim卡 是否root 网络环境 客户端版本
- 地域性差异
  一线城市 二线城市 三线城市 南北差异 省份差异
- 时间性差异
  上午 下午 晚上 工作日 周末 节假日

WHY - 资源分配

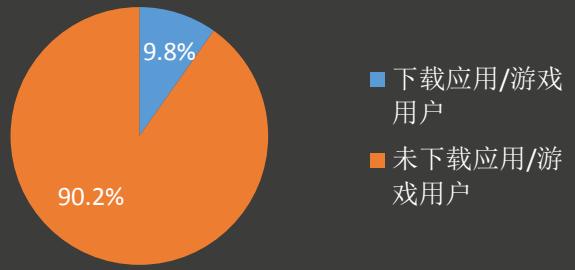


WHY - 用户粘性低

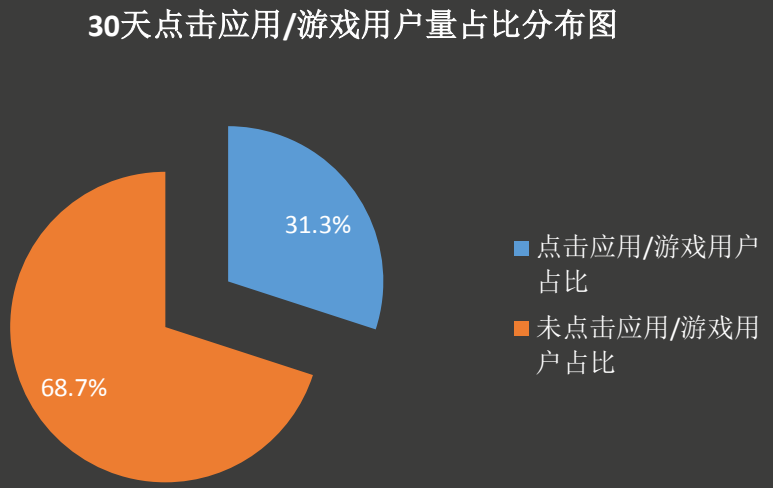
7天手卫推荐展现用户占比分布图



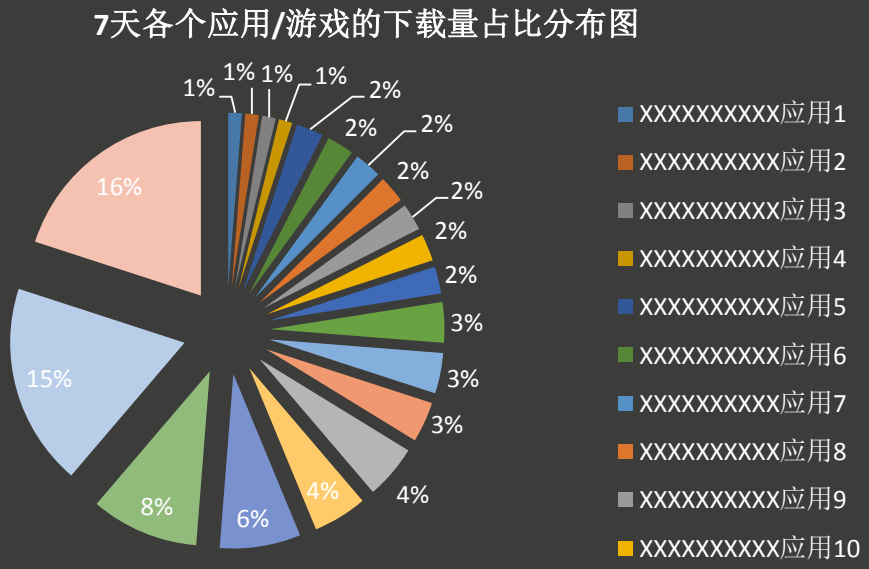
7天下载应用/游戏用户量占比



WHY - 用户质量不一



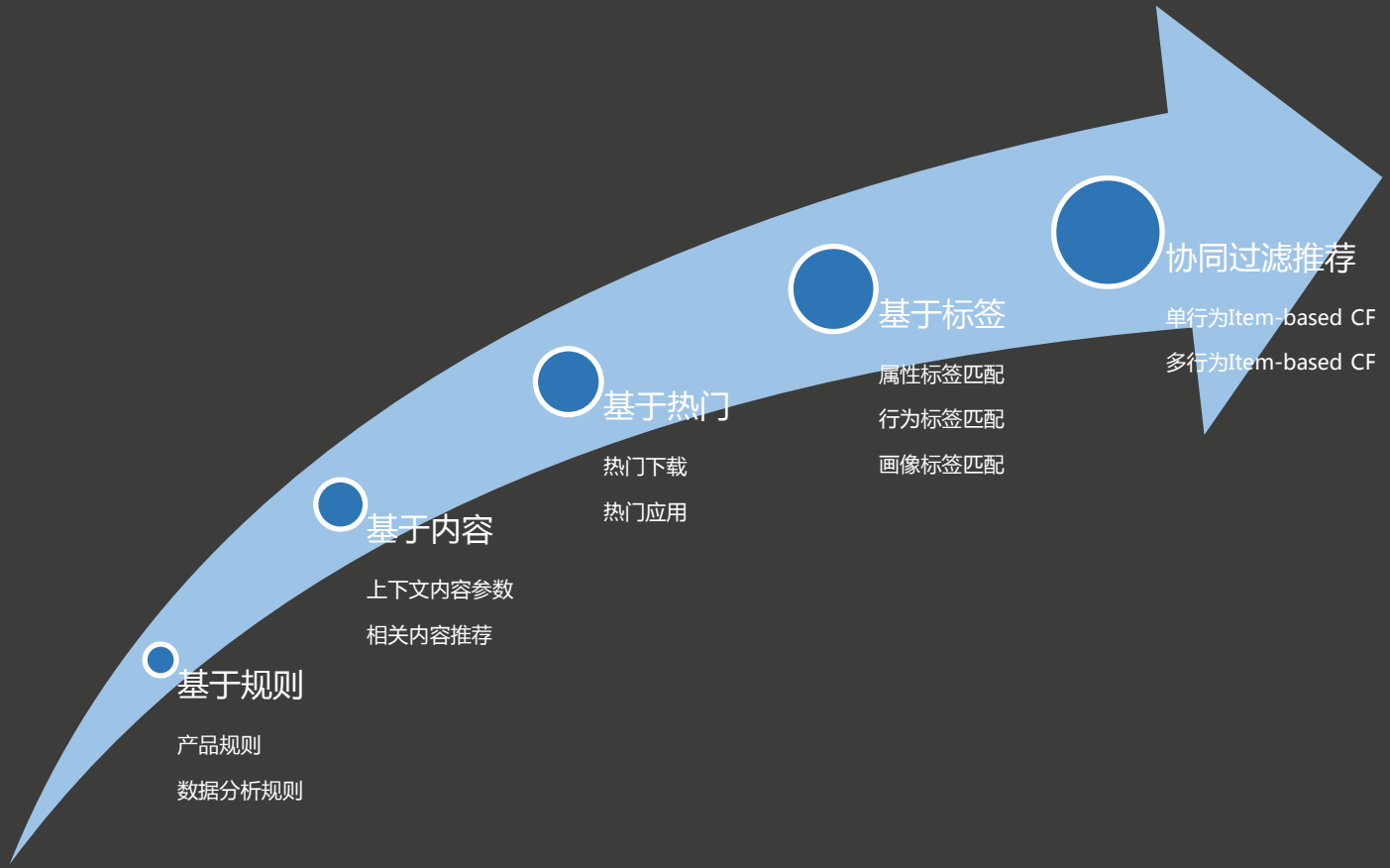
WHY - 应用质量不一


7天各个应用/游戏的下载量占比分布图
1%
1%
1%
1%
2%
2%
2%
2%
2%
2%
3%
3%
3%
4%
4%
6%
8%
15%
16%
XXXXXXXXXX应用1
XXXXXXXXXX应用2
XXXXXXXXXX应用3
XXXXXXXXXX应用4
XXXXXXXXXX应用5
XXXXXXXXXX应用6
XXXXXXXXXX应用7
XXXXXXXXXX应用8
XXXXXXXXXX应用9
XXXXXXXXXX应用10

HOW - 推荐系统框架


应用层
模型层
融合过滤层
召回层
数据层
应用
体检场景
清理场景
游戏加速场景
充电保护场景
红包场景
重排序
LR
FM
Xgboost
Xgboost+LR
Wide&Deep
DeepFM
候选集融合规则过滤
多策略融合
安装过滤
频次过滤
cpc值过滤
候选集触发
基于协同过滤策略
基于标签召回策略
基于内容召回策略
基于热门信息策略
基于相关规则策略
基于替补召回策略
数据存储
用户画像
协同过滤数据
热门数据
行为数据
数学模型
标签数据
数据生产
用户画像维护服务
协同过滤计算模块
热门信息计算模块
用户行为维护服务
数学建模
标签挖掘模块
离线 MR、HIVE、SPARK
实时 STORM、 SPARK STREAMING

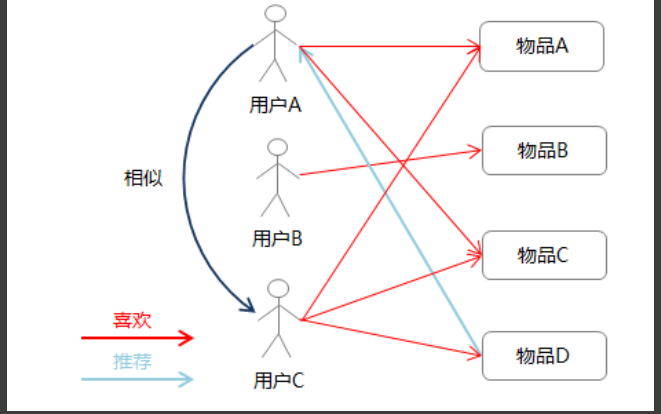
HOW - 召回策略演进


基于规则
产品规则
数据分析规则
基于内容
上下文内容参数
相关内容推荐
基于热门
热门下载
热门应用
基于标签
属性标签匹配
行为标签匹配
画像标签匹配
协同过滤推荐
单行为Item-based CF
多行为Item-based CF

HOW - 协同过滤算法

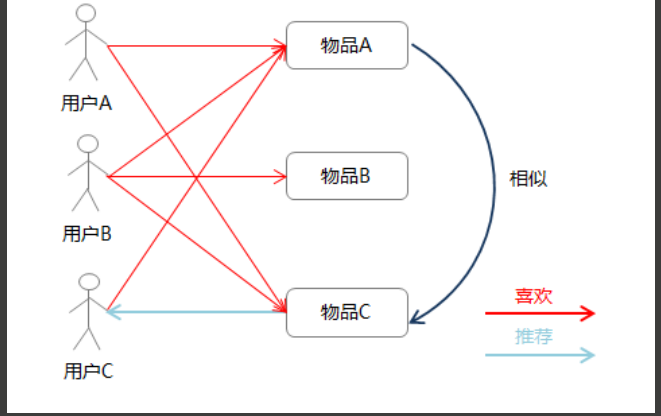
- User-based CF 算法

用户/物品	物品A	物品B	物品C	物品D
用户A	√		√	推荐
用户B		√		
用户C	√		√	√



- Item-based CF 算法

用户/物品	物品A	物品B	物品C
用户A	√		√
用户B	√	√	√
用户C	√		推荐



HOW - 多行为物品协同过滤

如何更精准的计算 Item 间相似度？

- 物品相似性、多行为融合、行为时效性、用户差异性

- Item 间相似度计算

$$D_{Item}(I_1, I_2) = \sum_{u \in U} d_u(I_1, I_2) \Leftarrow \sqrt{\sum_{u=1}^{U} (I_1^u - I_2^u)^2}$$

- 基于User的Item相似度计算

$$d_u(I_1, I_2) = \sum_{p \in P} d_p(I_1, I_2) \cdot d_B(I_1, I_2) \cdot d_T(I_1, I_2) \cdot d_{penalization}$$

- 物品相似性（物品标签）

$$d_p^t(I_1^t, I_2^t) = \frac{|I_1^t \cap I_2^t| + c}{\sqrt{|I_1^t| \cdot |I_2^t|}}$$

- 多行为融合（权重计算）

$$d_B(I_1, I_2) = \sum_{b \in B} |I_1^b \cap I_2^b| \cdot w^b$$

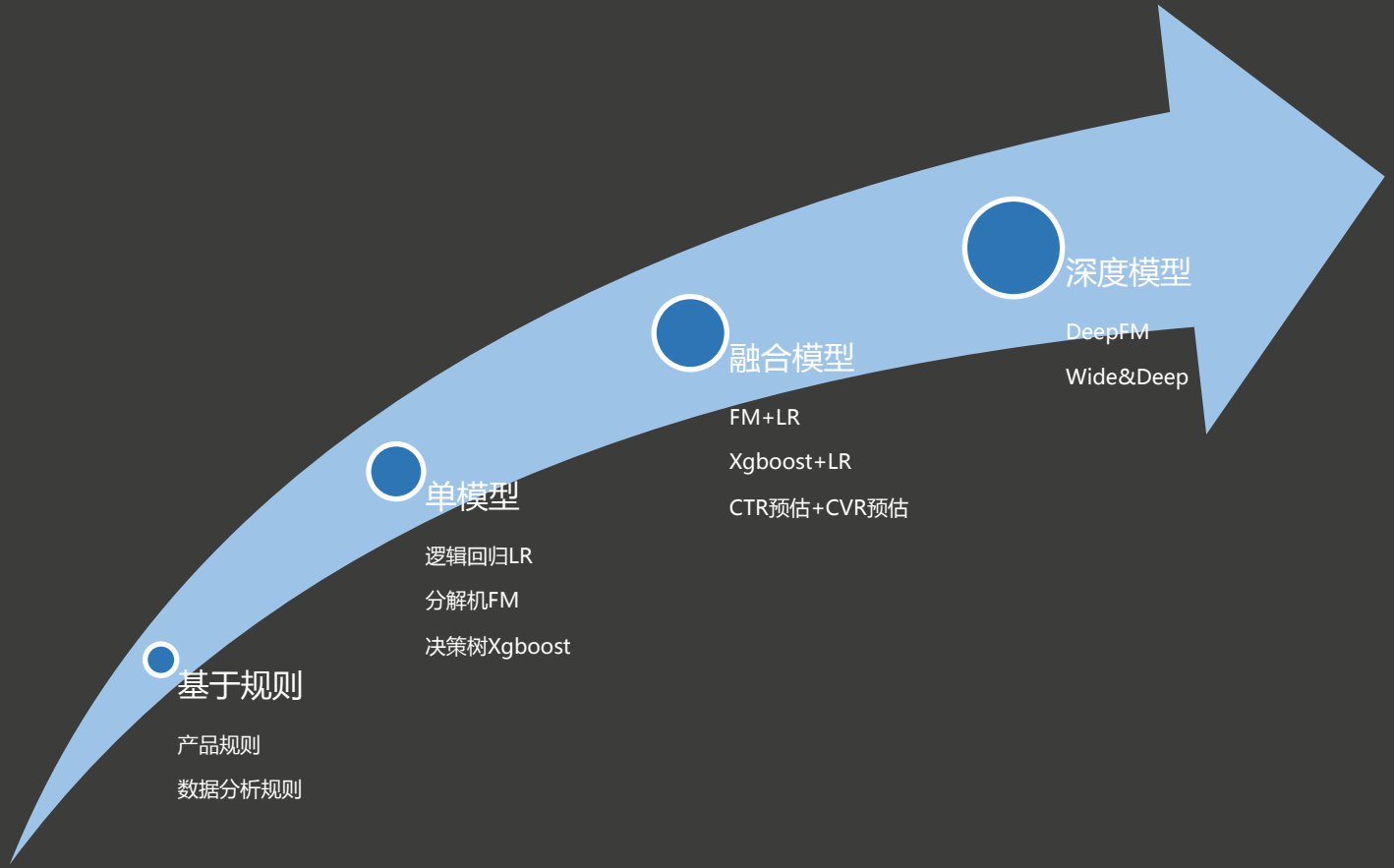
- 时效性（用户行为时间衰减）

$$d_T(I_1, I_2) = N_0 e^{-\lambda(t_{now} - t_{I_2})}$$

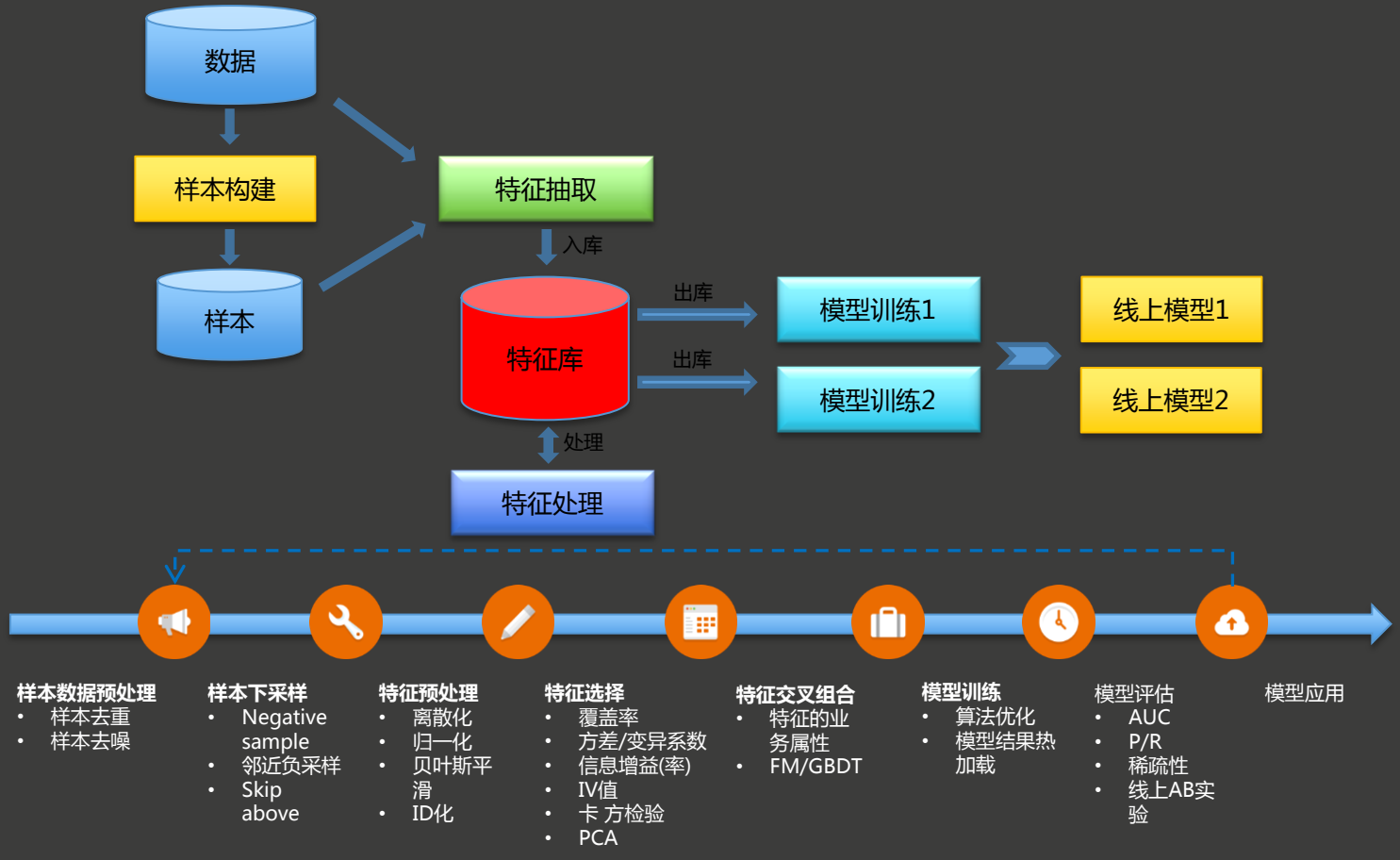
- 用户差异性

$$d_{penalization}^u = \frac{1}{log_a(N_u + c)}$$

HOW - 排序模型演进


基于规则
产品规则
数据分析规则
单模型
逻辑回归LR
分解机FM
决策树Xgboost
融合模型
FM+LR
Xgboost+LR
CTR预估+CVR预估
深度模型
DeepFM
Wide&Deep

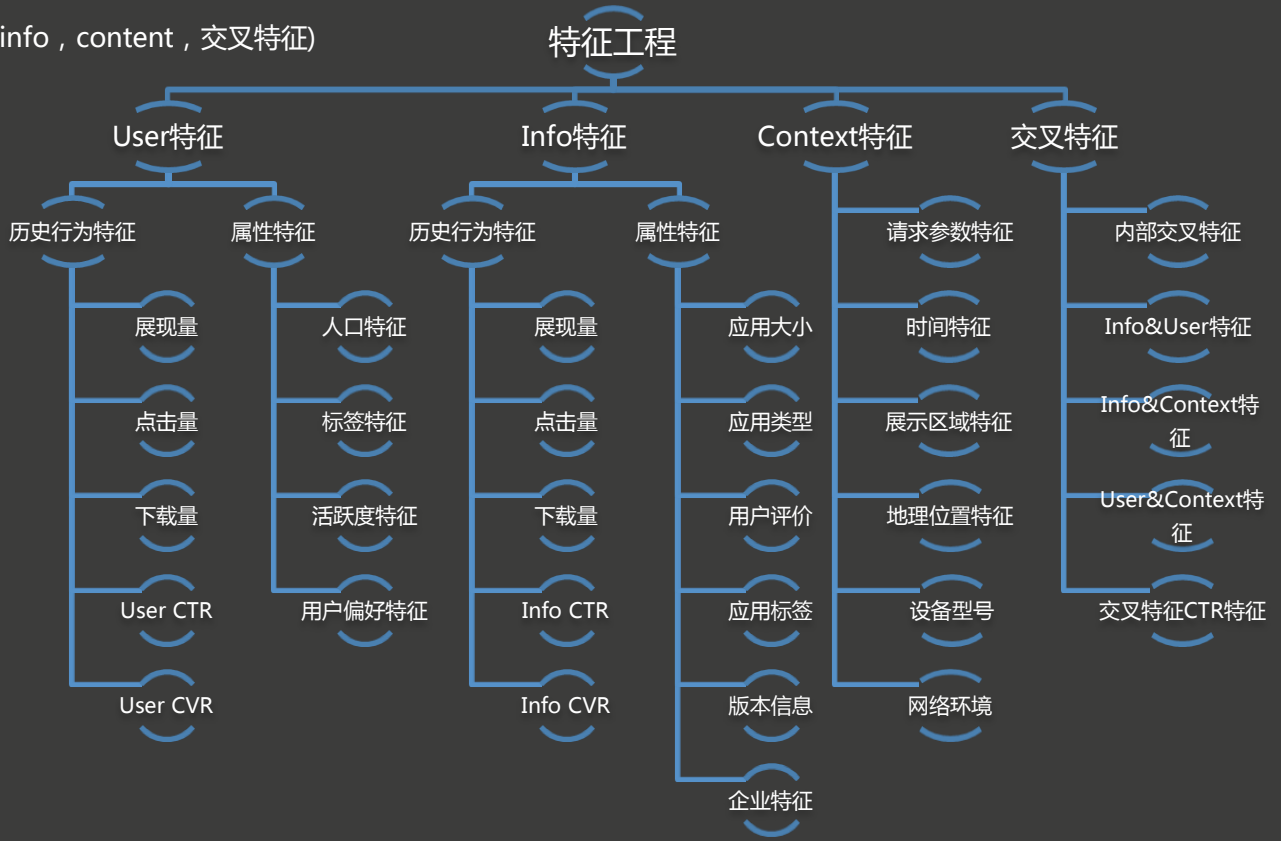
HOW - 排序模型系统流程



HOW - 排序模型特征工程

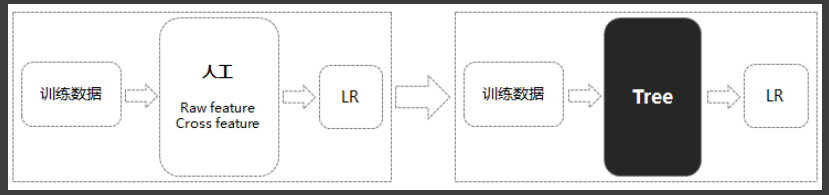
排序模型

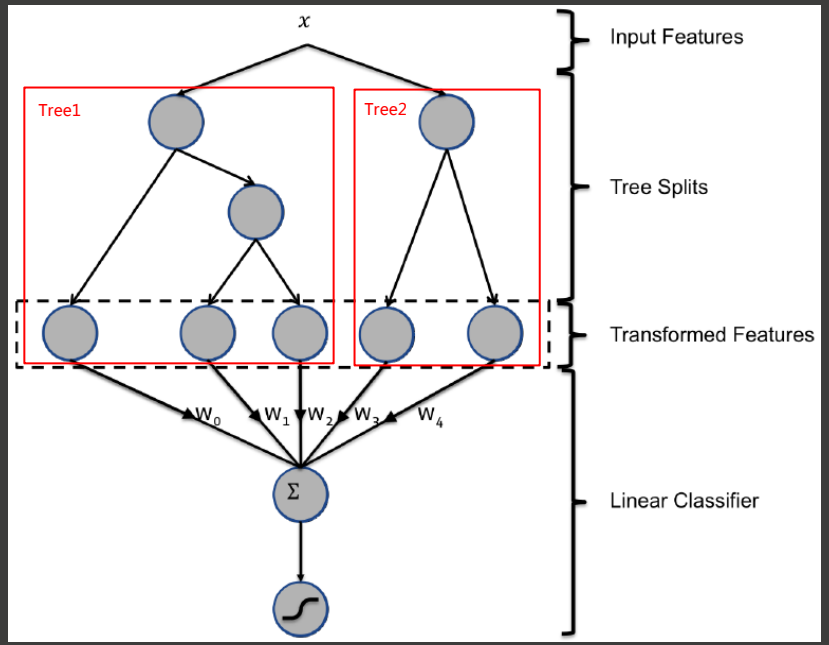
score = func(user，info，content，交叉特征)



HOW - 排序模型选型

- 机器学习模型
- LR
- FM
- Xgboost
- FM+LR
- Xgboost+LR





HOW - 排序模型选型

- 深度学习模型
- Wide & Deep

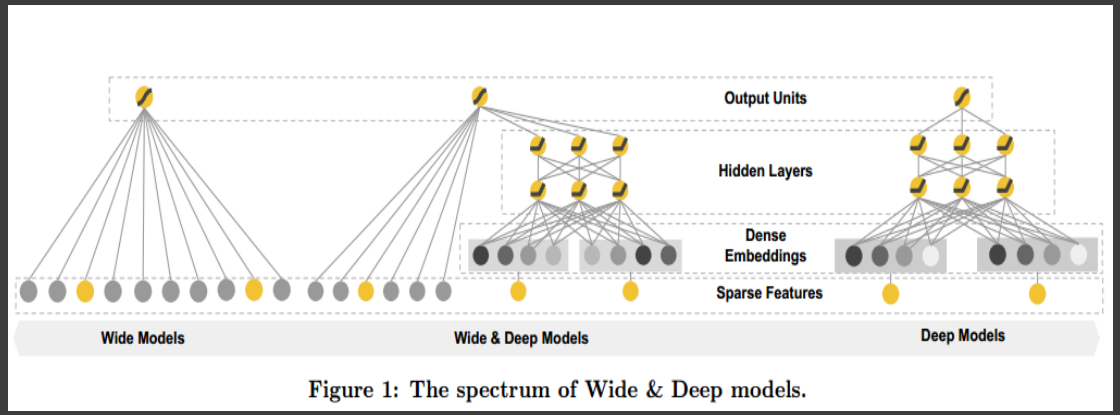


Figure 1: The spectrum of Wide & Deep models.

Wide Models为线性模型

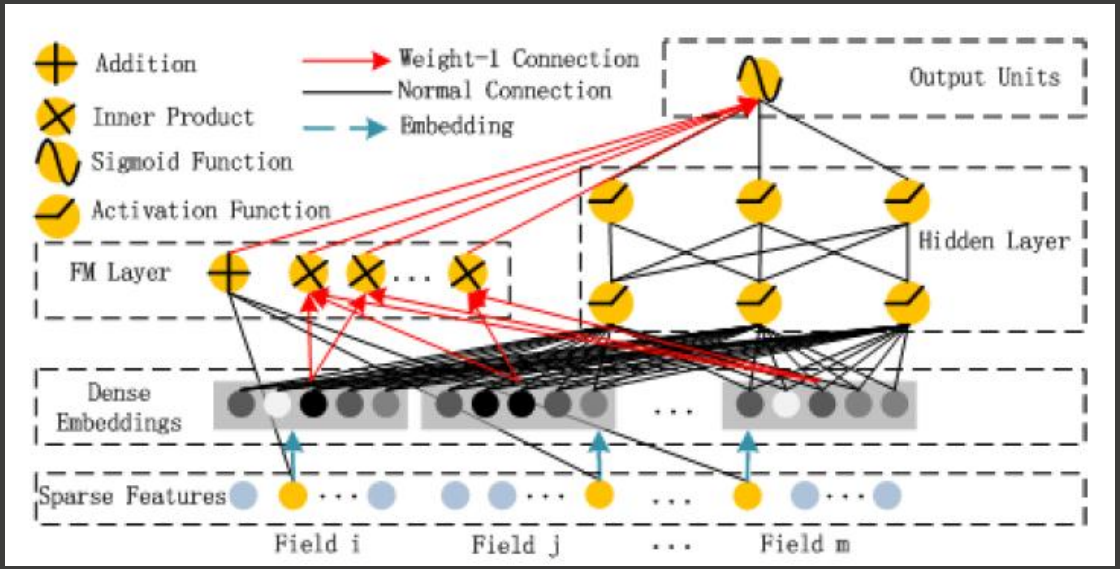
$$y = w^T x + b$$

Deep Models为DNN模型

$$a^{(l+1)} = f(W^{(l)} a^{(l)} + b^{(l)})$$

HOW - 排序模型选型

- 深度学习模型
- DeepFM



DeepFM的模型预测结果

$$\hat{y} = sigmoid(yFM + yDNN)$$

HOW - 排序模型选型

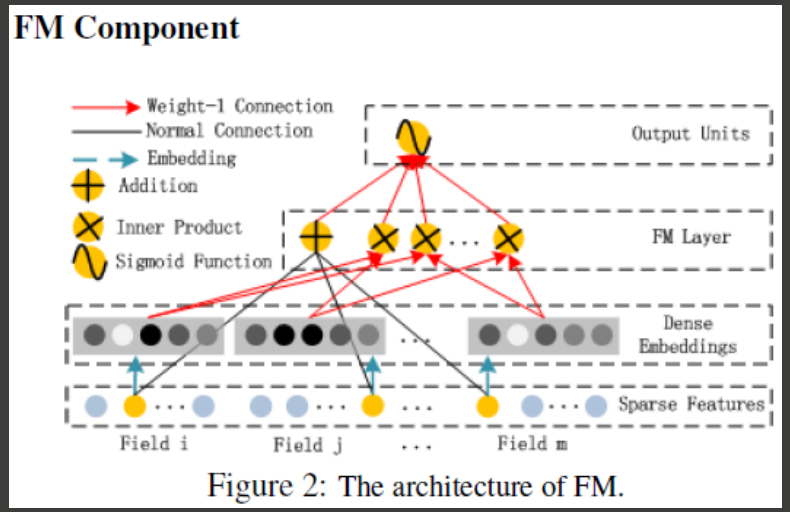


Figure 2: The architecture of FM.

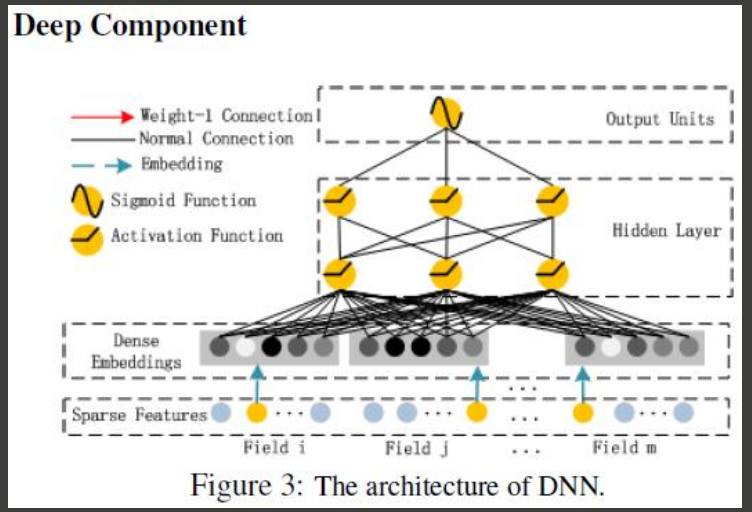


Figure 3: The architecture of DNN.

FM部分是一个因子分解机

Deep部分是一个前馈神经网络

$$y_{\text{FM}} = \langle w, x \rangle + \sum_{j_1=1}^{d} \sum_{j_2=j_1+1}^{d} \langle V_i, V_j \rangle x_{j_1} \cdot x_{j_2}$$

$$a^{(l+1)} = \sigma(W^{(l)} a^{(l)} + b^{(l)})$$

HOW - 排序模型选型

- 模型选型

模型	优点	缺点
LR	模型简单，训练速度快，适合处理线性问题，可解释性强	特征工程工作量大，人工选择特征和交叉特征
FM	与LR对比，模型中加入了二阶特征，通过embedding的内积来表达交叉特征权重，有更好的泛化能力。	无法学习三个及以上的特征间的关系，仍然需要人工选择特征，工作量大
GBDT	很好的非线性拟合能力，尤其对连续性特征，能够筛选信息量较大的特征	模型计算复杂度高，容易过拟合，计算速度慢
Xgboost	与GBDT对比，还支持线性分类器，加入正则项控制模型的复杂度，防止过拟合，并行计算速度快	模型参数多，调参优化复杂
DNN	自动特征组合学习和强泛化能力，直接输入原始的特征，减少了交叉特征的选择工作，并且可以支持大量的特征输入	模型参数多，调参优化复杂，工程化和实时化难度大，可解释性差

HOW - 排序模型总结

- 模型总结

样本

- 根据计算资源、存储资源、整体方案确定样本量

特征

- 与时间有关的特征要注意，防止穿越
- 线上和线下使用同一份特征工程，以保证特征一致性

模型

- 优先使用业界常用的模型、自己熟悉的工具
- 单模型：LR/FM/Xgboost
- 融合模型：FM+LR/Xgboost+LR

系统

- 建立高效的流程、分工，快速迭代
- 做好迭代记录，方便回滚、复用

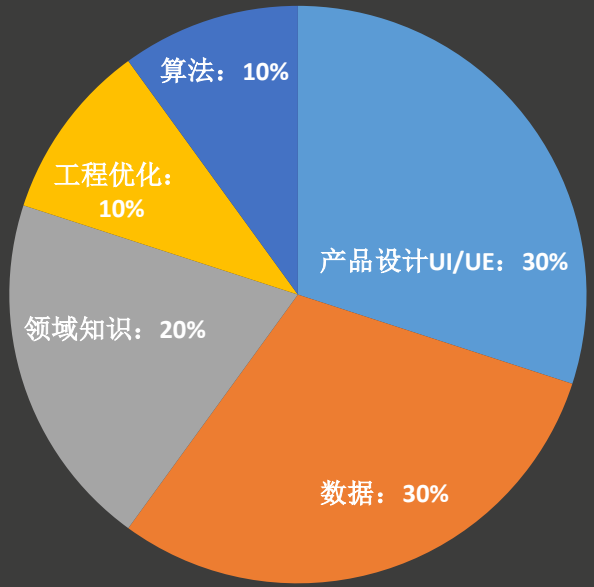
EVALUATE - 评价推荐系统效果

- 评价推荐效果

SUMMARY - 推荐系统总结

- 产品设计：好的产品设计能够引流到推荐，大大提高用户体验性
- 数据：最基础，最重要
- 领域知识：业务知识
- 工程优化：模块化，可扩展性，高性能，A/B实验框架非常重要
- 算法：个性化推荐和快速迭代，增加显著收益

影响推荐系统效果的因素



谢 谢!

技术交流：奇卓社（手卫技术微信公众号）