

视频精准推荐系统实践

优酷土豆集团 吕红亮

lvhongliang@youku.com

个人介绍

- 2012年宝宝树 – 负责内容推荐系统
- 2014年360搜索 – 搜索结果页右侧推荐、新闻推送
- 2015年加入优酷土豆– 负责精准推荐系统

- 目前为优酷土豆数据智能团队高级技术经理，负责精准推荐系统和相关推荐系统

目录

- 背景介绍
- 精准推荐系统架构
- 精准推荐算法架构
- 计划与方向

背景介绍-需求场景



英语教学自频道长李雷勤奋专业有潜力，优土想扶持他成为“万万计划”的一员

流量运营

用户/场景定向

韩梅梅新做了一个餐馆评测，想尽快让更多中关村地区的小清新吃货看到~



联合出品《捉妖记》：哪种片花更能吸引大学生群体？

用户调查/反馈

背景介绍-展现形态1

The image shows a screenshot of the Youku website. At the top, there is a navigation bar with the Youku logo, a search bar, and user information. Below the navigation bar, there is a video grid with several video thumbnails. To the right of the video grid, there is a large advertisement for the ECOVACS robot vacuum. Below the advertisement, there are two smaller video recommendations.

YOUKU 优酷 首页 三 频道

搜索

上传 通知 谢达27203

精选 发现 订阅 记录 会员 订阅列表

搞笑频道 > 搞笑列表 > 恶搞短片

视频: 【哇哈哦哦】哈哈!拖拉机拔树,驾驶员被树连续拍打

笑死了!健身房里的奇葩!从头笑到尾啊!

【哇哈哦哦】笑死!女子不由自主拧了拧男骑士的油门...结

三少家族Robert

收藏 分享 顶 重播

张凤

暴力拖拉机:改装炸机发动机的拖拉机高清

【孽缘楼】太吓人了!油门和刹车一起踩下会发生什么事?

直升机没停好,直接翻进海,所有人都傻了!

【哇哈哦哦】实拍作死男脸被超级大猩猩“锤死”

在土豪国度迪拜丢钱包路人什么反应?社会试验在迪拜

不信你不笑-09年10大雷人搞笑汽车大PK_标清

最强国产拖拉机!越野性能完爆各式SUV

会员免费 加勒比海空黑珍珠... 加勒比海上,惊险魂追迹 评分:★★★★★ 5.9 主演:约翰尼·德普/杰弗里·

会员用券 告别昨日 毕让新生运动励志片 评分:★★★★★ 7.0 主演:丹尼斯·克里斯托弗/

背景介绍-展现形态2

 **猜你喜欢** 这是根据您的口味给您推荐的视频哦~~

 换一换



郑多燕中文版第二部健身操02

 2.1万  3



屌丝在宿舍撸克罗地亚狂想曲
看到鼓手时我跪下尿了一

 10.9万  271



来疯直播 亲爱哒，能去你家卖

 1,490人正在观看



刘姝威:中国制造到中国创造需
要自主技术提升

 3,107  2



习近平参观考察美国农场 120
217 新闻联播

 1.2万  1



潜行吧，电脑！这个电脑坏
了，死机之歌

 3.4万  120



郑多燕中文版第二部健身操03

 2.8万  2



马克西姆港大表演克罗地亚狂
想曲

 25.0万  208

背景介绍-展现形态3



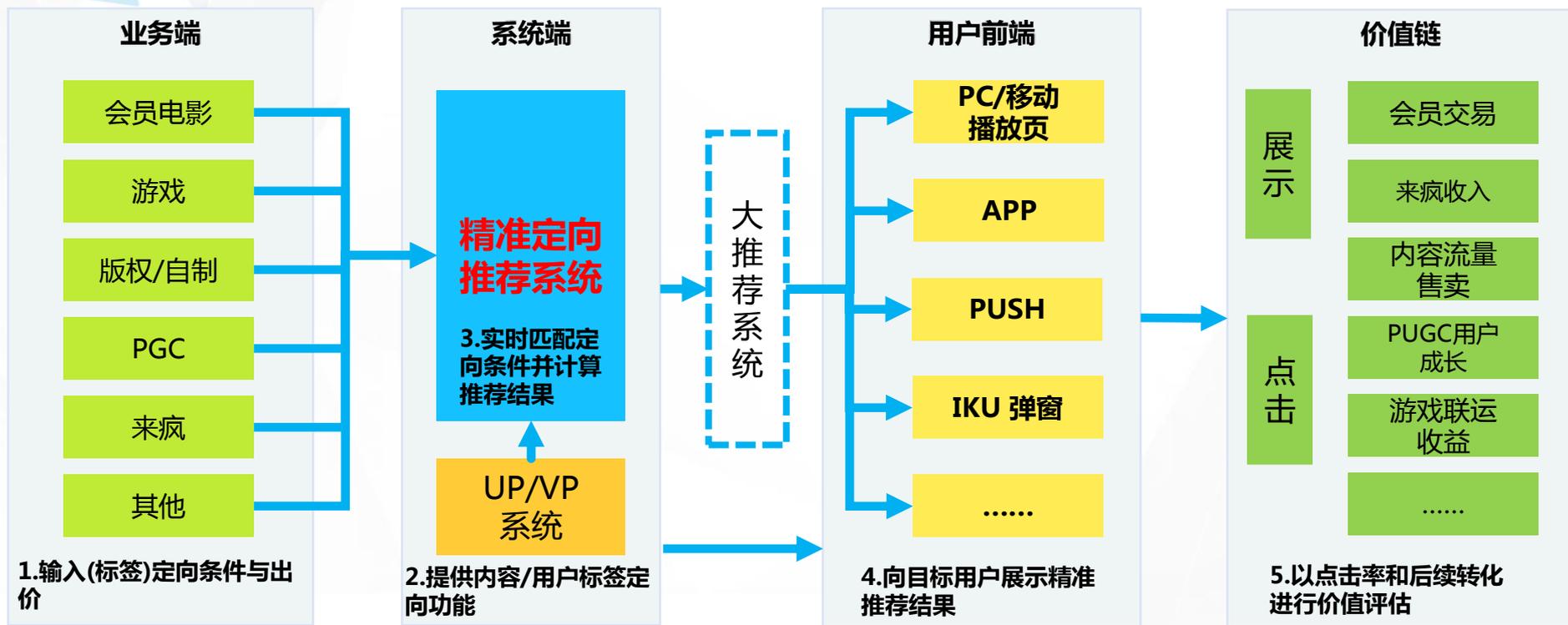
背景介绍

- 需求方
 - 内部业务方：来疯、会员、PGC、分成、生态、游戏、BD
 - 外部业务方：自频道主，PGC
- 精准推荐系统目的
 - 整合和优化推广位，尽量满足各方需求，使资源位的价值最大化
 - 打造视频生态：扶持自频道主、PGC，并引导他们的行为
 - 保证用户的体验：在合适的时机给合适的用户推荐合适的限定集合内的视频

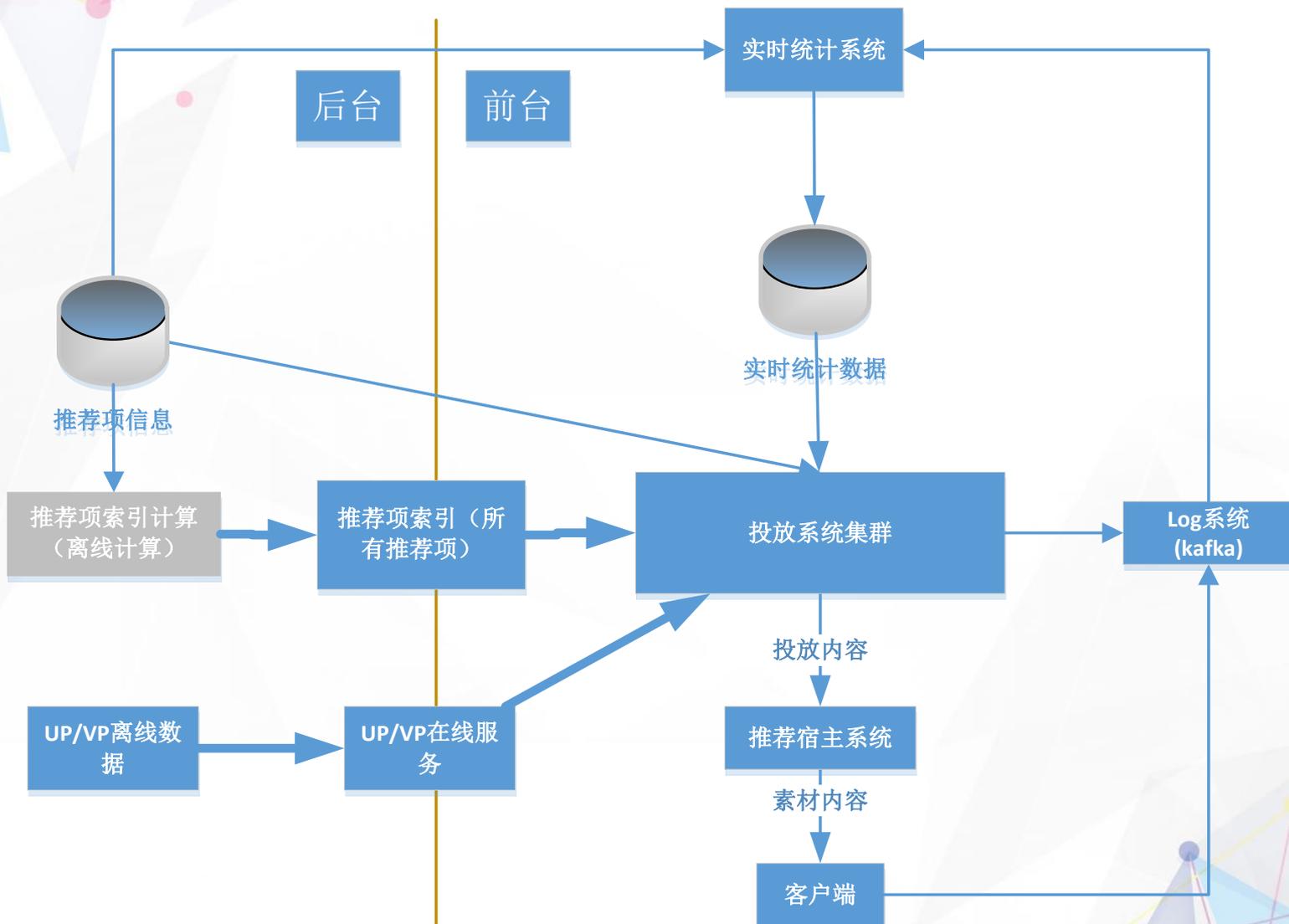
目录

- 背景介绍
- **精准推荐系统架构**
- 精准推荐算法架构
- 计划与方向

精准推荐系统-所处位置



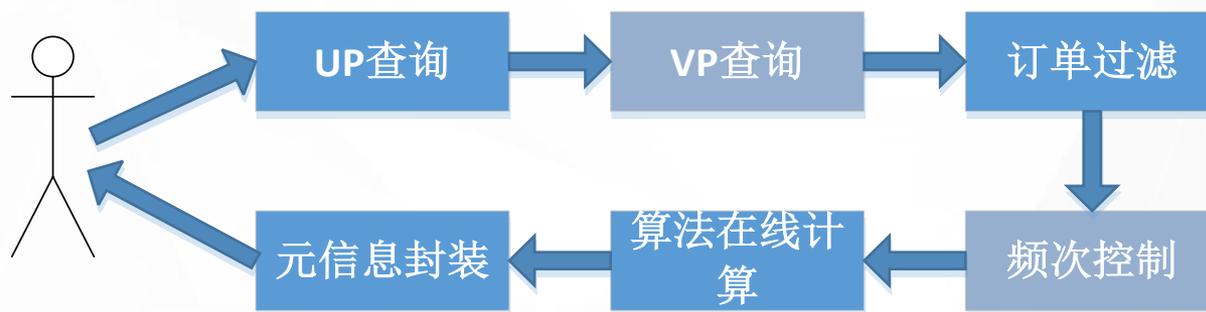
精准推荐系统架构-概览



精准推荐系统架构-链式结构

• 链式结构

- 责任链模式(Chain of Responsibility): 一系列有序的command能够按照顺序执行,并能够互相交换或者传递执行结果
- 解耦: 方便在线更换处理节点,通过节点组合实现不同的功能
- 自动降级: 方便跳过一些节点,实现自动降级功能



算法链式处理

召回 (获取候选集)

排序

重排序

精准推荐系统架构-定向投放

- 通过UP、VP系统获取用户画像和视频画像
- 通过索引系统将订单索引起来



用户性别： 不限 自定义
 性别未知 男 女

用户年龄： 不限 自定义
 18岁及以下 18-24岁 25-29岁 30-34岁

兴趣偏好： 不限 自定义

请输入名称

- 娱乐
- 音乐
- 动漫
- 搞笑
- 游戏
- 生活
- 资讯
- 体育
- 电影
- 综艺
- 电视剧
- 拍客

相似人群： 不限 自定义

请输入名称

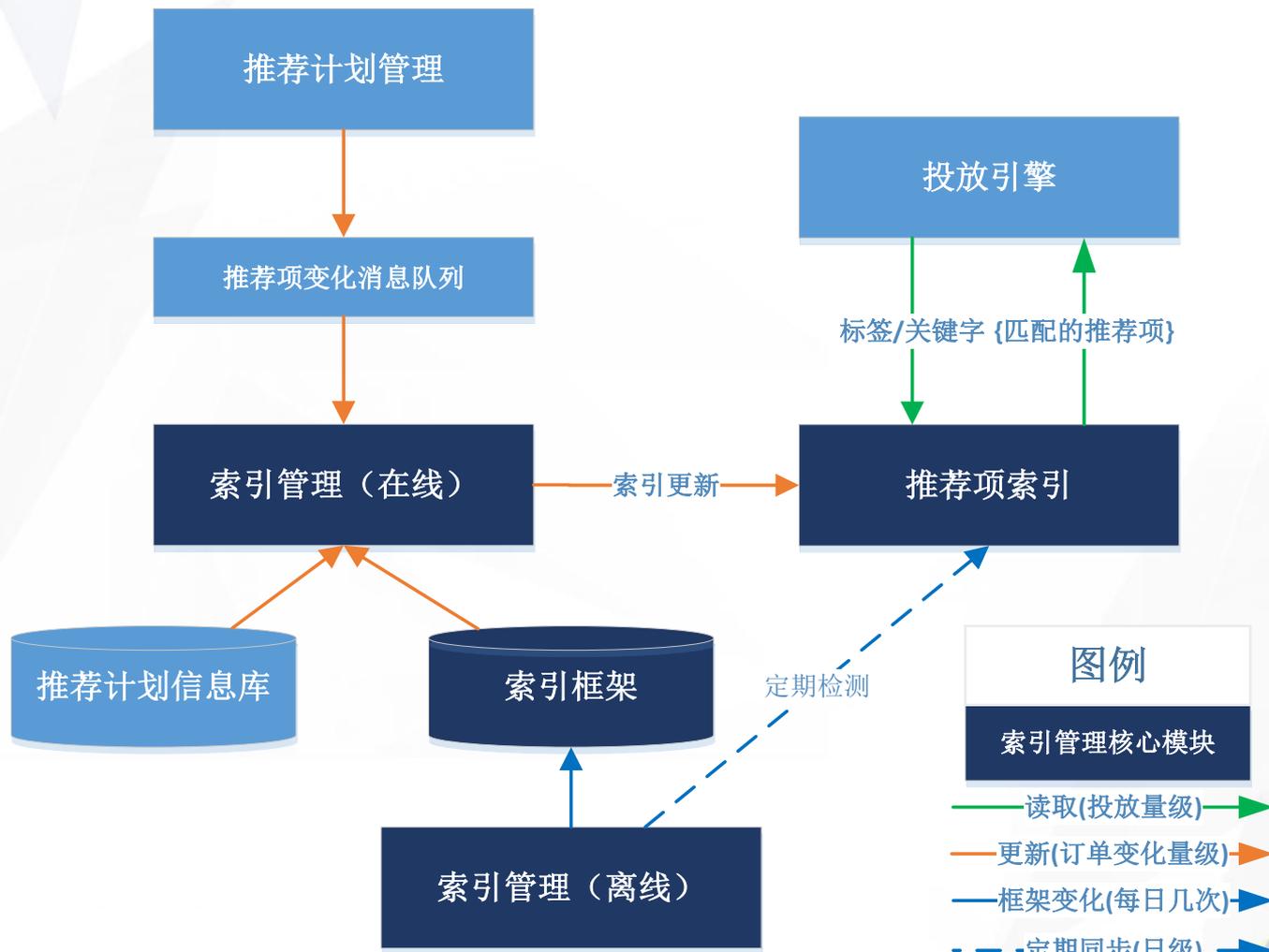
游戏迷

视频标签：

请输入名称

- 体育-覆盖视频数(
- 搞笑-覆盖视频数(
- 动漫-覆盖视频数(
- 娱乐-覆盖视频数(
- 游戏-覆盖视频数(
- 财经-覆盖视频数(
- 教育-覆盖视频数(
- 旅游-覆盖视频数(
- 时尚-覆盖视频数(
- 汽车-覆盖视频数(
- 汽车评测-覆盖
- 汽车知识-覆盖
- 汽车资讯-覆盖

精准推荐系统架构-索引管理



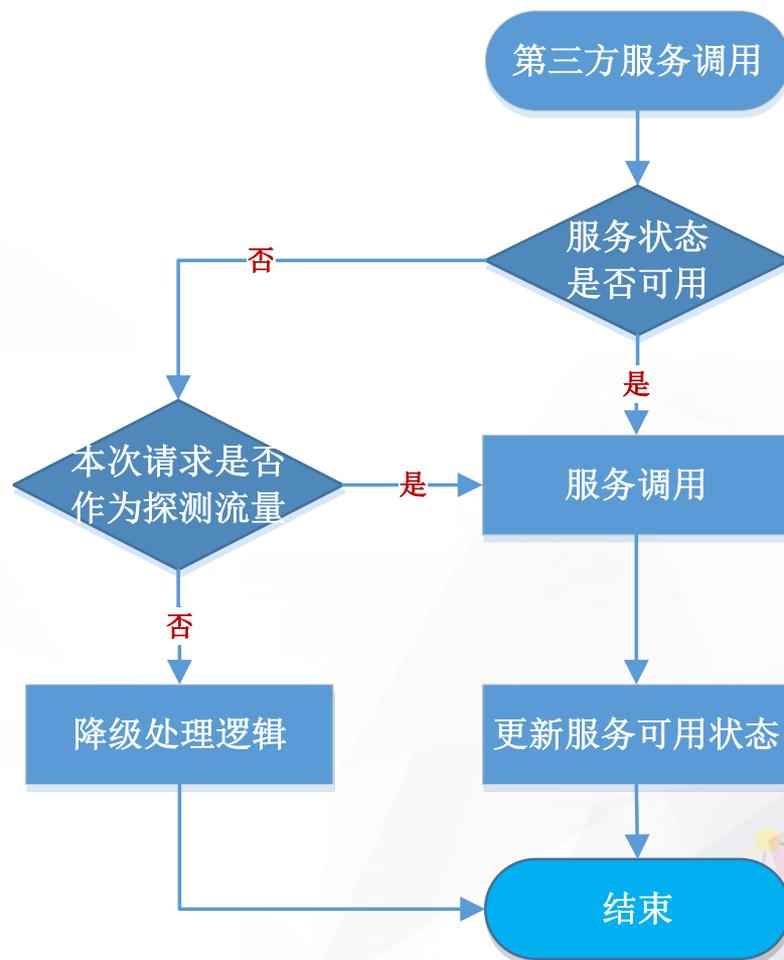
系统架构经验1-服务稳定性保障

• 服务降级

- 当服务器负载过高时，简化运算逻辑
- 外部依赖的外部服务出现异常，及时熔断防止主服务被拖垮

• 降级顺序

- 算法降级：非线性模型（GBDT）->线性模型（LR）
- 分流控制模块关闭
- up、tp模块功能关闭
- 投放缓存内容



系统架构经验2-在线debug系统

- 快速定位问题
- 发现潜在问题
- 便于技术和产品调研->找到新的优化方向

系统架构经验3-监控报警系统

- 系统状态监控
 - CPU、IO、内存
 - 接口响应时间、外部依赖接口响应时间
 - 线程数、jvm状态
- 自定义sql监控数据库
 - 异常订单报警
 - 运营数据监控
- 自定义接口监控
 - 支持自定义接口实现复杂的监控功能

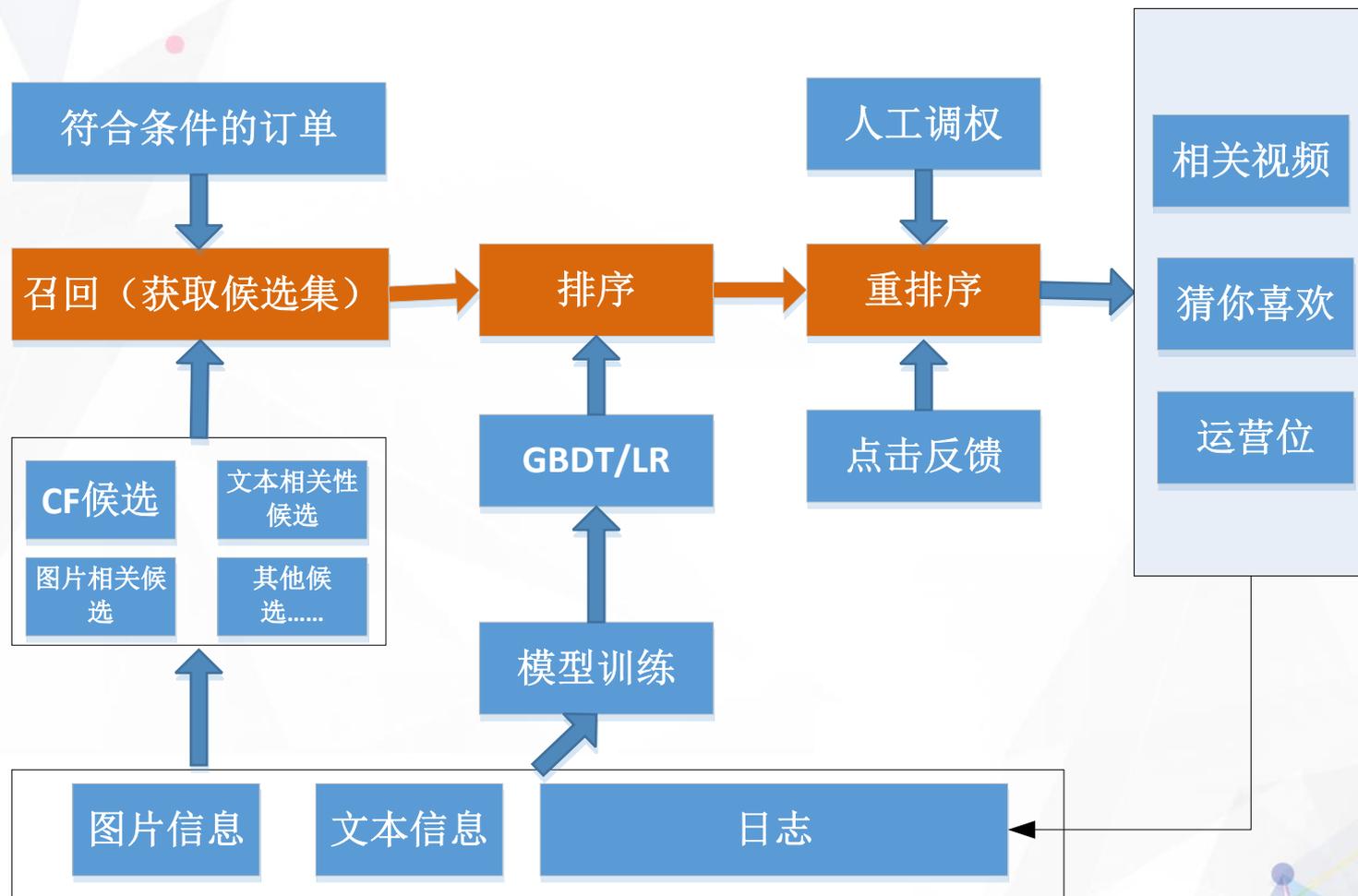
精准推荐系统架构-关键数据

- 每天收到的请求量5亿+
- 峰值QPS8000+, 平均响应时间20ms以下
- 每天的曝光4亿+, 点击率0.5%以上
- 满足优士集团十几个事业部的内部推广需求

目录

- 背景介绍
- 精准推荐系统架构
- **精准推荐算法架构**
- 计划与方向

精准推荐算法架构-概览



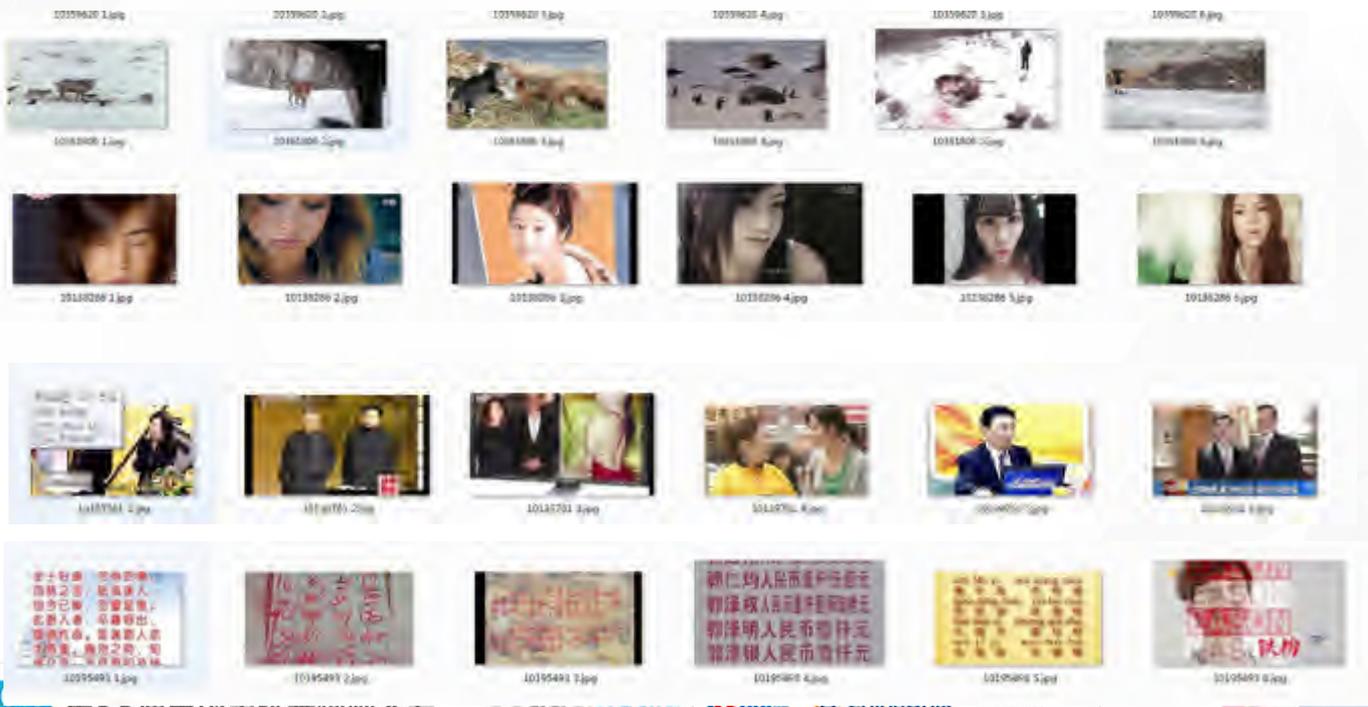
精准推荐算法-召回

- 用户行为：（Item-based CF）
 - 看了又看.....
 - 收藏了还收藏了.....
 - 评论了还评论了.....
- 内容：
 - 同主演、同导演、同嘉宾
 - 同一题材：校园偶像剧、时空穿越类科幻
 - 标题相似：LSH、word2vec
 - 视频内容相似：截图、关键帧
- 热门
 - 夏季热播
 - IMDB Top 200、2016奥斯卡获奖电影

精准推荐算法-截图召回

- 截图召回

- 图片理解：通过CNN embedding到高维空间
- 计算图片相似性



精准推荐算法-排序

- 模型：
 - LR、FTRL
 - GBDT
- 特征工程
 - 特征预处理：归一化，one-hot编码，缺失值补充，异常值去除，数据变化
 - 特征选择
 - Filter：计算特征和目标的相关性，比如方差、卡方检验、相关系数
 - Wrapper：根据损失函数，递归增加或者消除若干特征
 - 模型的方法：L1正则，基于树模型
 - 实践比较有效的特征选择方法
 - 专家经验：和产品运营的同学多聊
 - 模型选择方法：L1正则，GBDT给出feature重要性排序
 - 用非线性模型取代线性模型来减少对组合特征发现的依赖
 - 快速迭代，快速上线测试

精准推荐算法-特征

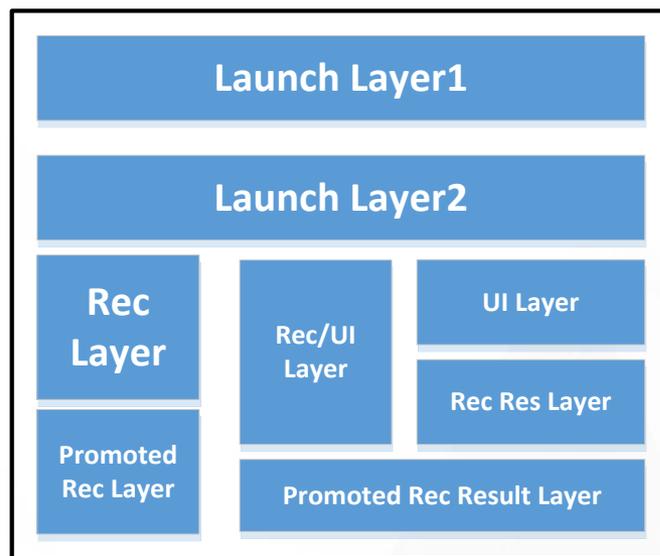
- 模型特征：候选集自带的特征，比如CF相似度，文本相关性相似度
- 用户类：
 - 人口统计学特征，比如年龄、性别、收入
 - 频道偏好
 - 兴趣标签
 - 使用的终端类别
- 源视频
 - 所属频道、标签
 - 标题、描述
 - 视频时长
- 目标视频
 - 视频质量、播放完成率
 - 所属频道、标签
 - 视频时长
- 上下文（Contextual Info）
 - 时间
 - 位置

精准推荐算法-重排序

- 点击反馈
- 多样性
- 根据线索分组
- 人工调权
- 节奏控制

精准推荐算法架构-A/B Test系统

- 分层实验，促进流量的最大化利用
- 保证用户真正随机
 - 为每次实验构建一个实验层



*参考paper: **Overlapping Experiment Infrastructure:More, Better, Faster Experimentation**

目录

- 背景介绍
- 精准推荐系统架构
- 精准推荐算法架构
- **计划与方向**

计划与方向-技术

- 实时化
 - 模型实时更新
 - 使用更多的实时化特征：比如用户短时兴趣
- 增强推荐解释
 - 同标签、同主演、同嘉宾、同类top10、奥斯卡、本站最热.....
- 用户显性反馈
- 深度学习
 - 观看序列分析：RNN、LSTM
 - 视频截图和视频内容分析：CNN

计划与方向-产品

- 让整体的流量智能可控
 - 让推荐流量和运营位的流量可以比较智能可控地支持视频生态建设
- 引导PGC的行为
 - 给PGC一些运营上的指导
 - 多上传原创视频
 - 持续上传、保持视频质量稳定
 - 视频主题尽量集中
 - 给PGC一些视频制作上建议
 - 视频具体的题材方向
 - 视频标签、视频长度
- 将这些能力开放给更多的PGC



Q&A

优酷土豆推荐团队氛围非常好，成长空间大，现在求贤若渴，机器学习/数据挖掘方向，请赐简历至：lvhongliang@youku.com，非常感谢！

THANKS

SequeMedia
盛拓传媒

IT168.com
1999-2010

ChinaUnix

ITPUB
www.itpub.net