个性化推荐架构设计

技术创新，变革未来

为什么要做推荐系统

千万级视频资源

推荐引擎

月活亿级用户量

推荐系统是继搜索之后解决数据过载的重要方法

个性化推荐产品形态

产品形式:首页下拉个性化消费流

下载渠道:应用宝、百手等部分渠道下载

个性化推荐产品形态

产品形式:基于PGC/UGC的个性化短视频推荐APP

下载渠道:计划11月底上线iOS&Android

个性化推荐产品形态

搜狐视频

界面交互

- 入口
- 基础界面

搜狐视频推荐系统整体概况

- 整合全站视频资源，通过“推荐引擎”和“视频处理引擎”将个性化、新鲜的视频快速分发到以适合场景，以合适形式传递给适合用户

- 计算快速：2秒 ，分布式高可用实时计算，稳定/灵活/易扩展；

- 海量数据分析： Online 17亿+ Offline 170亿+

- 智能排序：实时特征工程、在线学习、多模型融合

- 基础组件：知识库、主题模型、用户/视频画像、实时反馈/统计、独立后台、推荐引擎、视频处理引擎等，保证产品功能完备；

推荐系统架构

推荐系统架构



点击日志

负载均衡
LVS

观影日志

负载均衡
Nginx

展示日志

应用容器
Je/y

前端日志接入

日志采集
Flume

实时传输
ka7a

实时传输
Rocketmq

数据采集

推荐引擎
storm

特征&画像引擎
Spark streming

视频处理引擎
Marathon+Mesos

缓存
redis

实时计算

存储
hbase

推荐结果

缓存
redis

存储
Hdfs

离线计算
MapReduce&Spark

hive

mllib

离线计算和分析

算法库

算法A

算法B

算法C

算法N

....

模型库

特征A_1

排序A_2

召回模型A

召回模型B

视频画像

视频评分

用户画像

模型N_2

....

推荐系统架构

推荐结果

视频画像

相关服务

视频处理引擎

点击日志处理

观影日志处理

展示日志处理

推荐引擎

知识库

实时反馈系统

排序

特征工程

召回模型

规则卡片封装

监控系统

主题模型服务

卡片类型BF

个性化配置

用户画像

...

主题模型

评分系统

自然语言处理系统

基础组件

推荐系统架构

存储系统

推荐结果

index	card
0	Card_0
1	Card_1
2	Card_2
3	Card_3
4	Card_4
..	...
..	...

冯小刚

Key倒排

推荐内容倒排存储

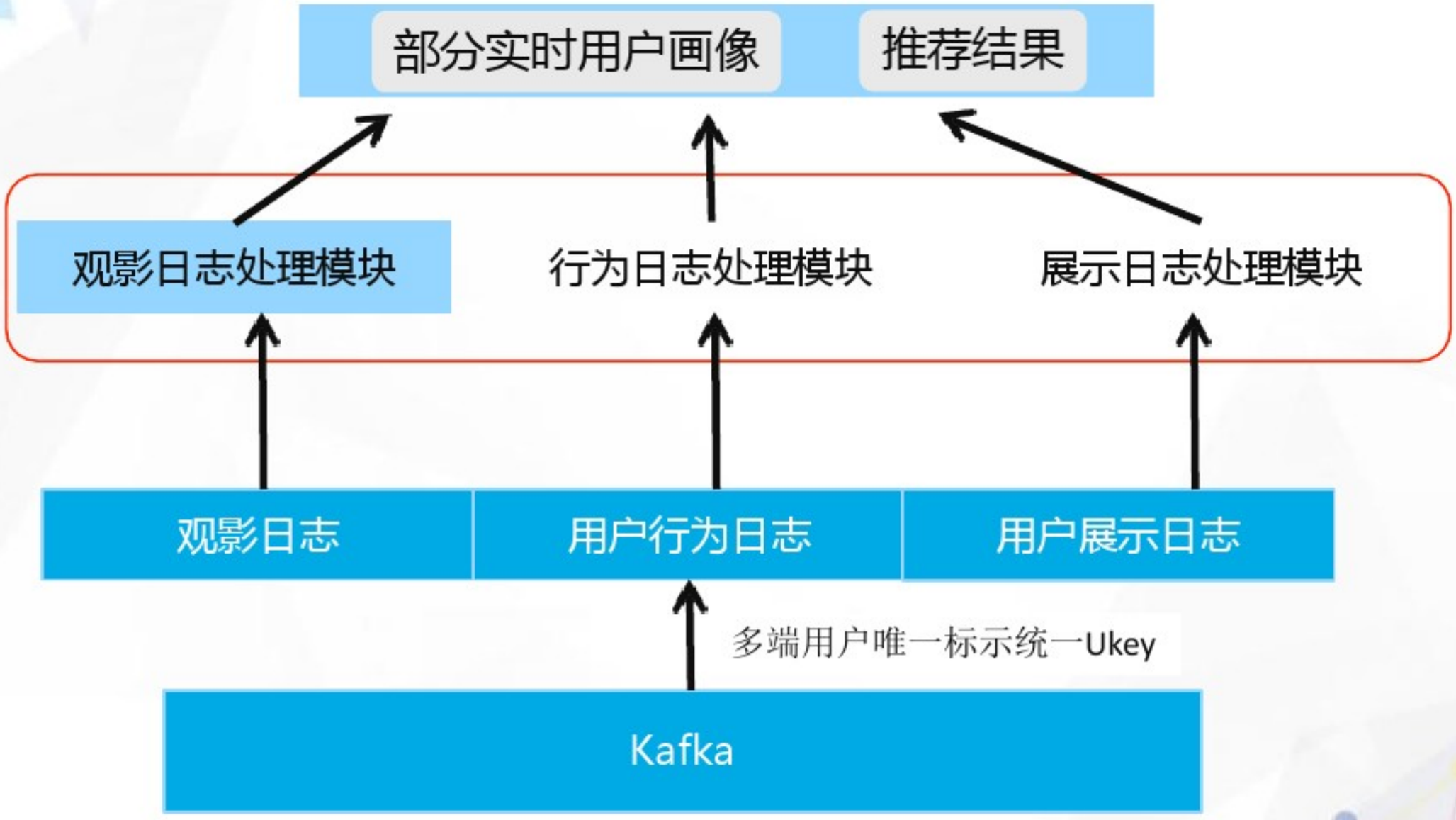
推荐引擎

- 召回模型
- 配比
- 排序
- 和谐性处理

视频处理引擎

推荐引擎

推荐系统架构-推荐引擎(storm)



- Bolt: localOrShuffleGrouping&fieldsGrouping
  - 基于Redis 一致性(CAS)

推荐系统架构-推荐引擎(storm)

行为处理模块

观影处理模块

展示处理模块

召回模型

配比

排序

和谐性处理

index	card
0	Card_0
1	Card_1
2	Card_2
3	Card_3
4	Card_4
..	...
..	...

推荐系统架构-推荐引擎

召回模型

- 海量的视频中选择用户感兴趣的候选集合的方法

配比

- 多角度看用户(多个召回模型结果融合)

排序

- 统一排序规则、多机器学习模型

和谐性处理

- 多样性、覆盖率

召回模型

推荐引擎-召回模型

海量的视频中选择用户感兴趣的候选集合的方法

方法

- 协同过滤：Item CF(Slope one),User CF,矩阵分解模型(SVD++、RSVD、ALS)、图模型(co-view图模型)
- 内容过滤：(Content-based Filtering)
- 基于人口统计学和社会化过滤(年龄、性别、工作、学历、居住地)
- 基于位置的过滤(场景和上下信息推荐方式)

- 离线：

1. SVD++、Slope one、ALS等矩阵分解模型为离线模型
2. 如基于图模型和内容推荐的融合:Item CF-KNN、User CF-KNN

- 在线：

1. 基于自然语言处理系统构建的分类体系、topic、keyword
2. 基于时间+地理位置的实时场景位置的构建。

配比

推荐引擎-配比

多角度看用户

推荐引擎-配比

假设有20个召回模型,我们用20维表示，每一维的值，代表配比数量，最大200

用户 ⟺ 召回模型

<X1,X2,X3,…,Xn-1Xn, Xn+1,Xn+2….,Xn+19, Xn+20 > 打开率 Score模型

ID	X1~X400	X401~X773	RM 对应召回模型个数
1	<1,2,…,0,3>	<0,4,0,2,…,0,3>	<10,2,4,5,6,3,6,6,6,7,8>
2	<1,2,…,0,3>	<0,4,0,2,…,0,3>	<9,2,4,5,6,3, 5,6,3,4,6>
3	<1,2,…,0,3>	<0,4,0,2,…,0,3>	<8,2,4, 6,6,6,7,8, 5,6,3>
4	<1,2,…,0,3>	<0,4,0,2,…,0,3>	<7,2,4,5,6,3,5,2, 3,4,6>

Id(配比编号)	Score(CTR)
1	0.7
2	0.65
3	0.71
4	0.74

排序

特征工程&排序模型

基础特征工程(Spark streaming)

↓

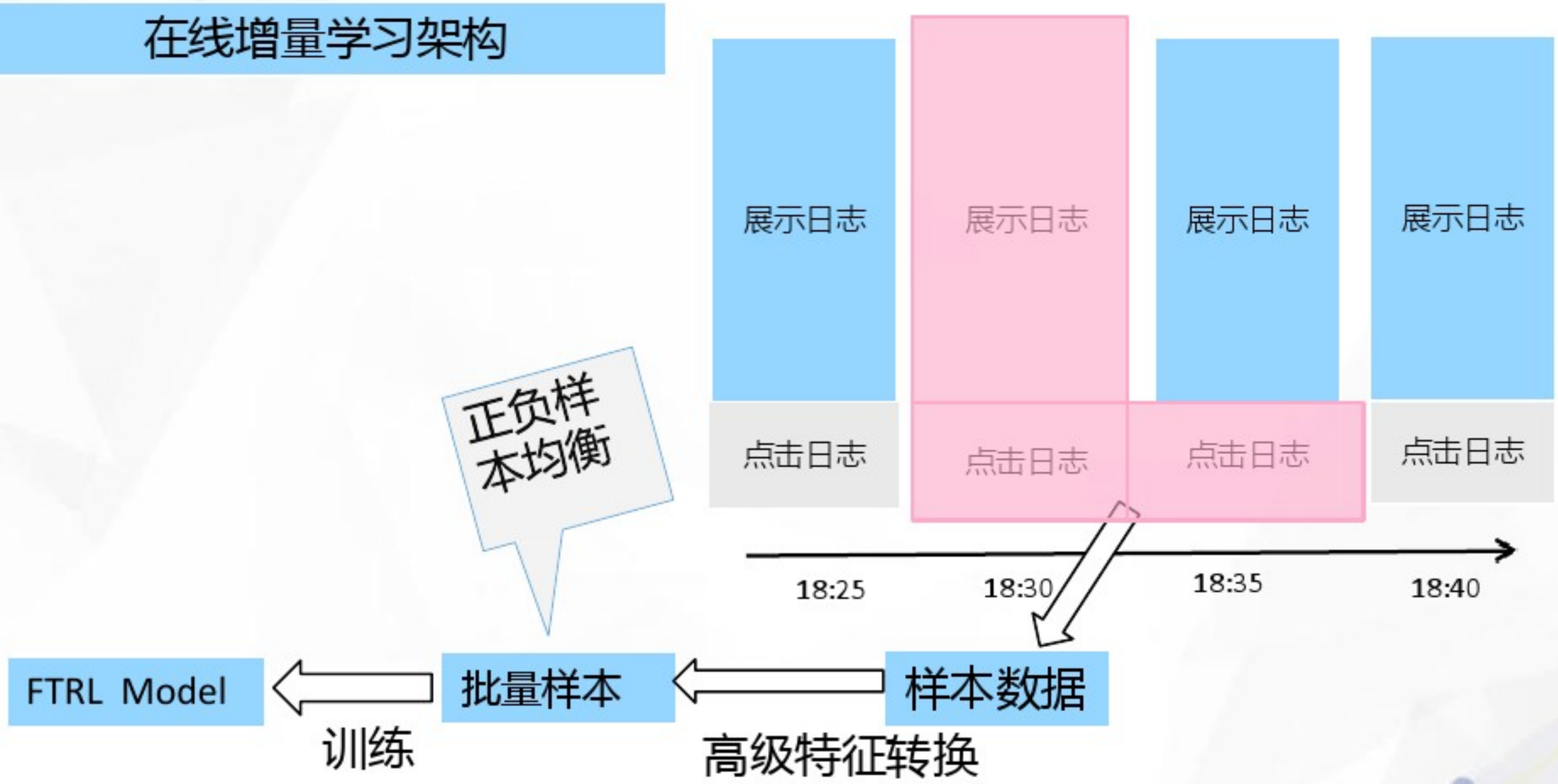
高级特征转化(GBDT、 DNN、 FM)

↓

算CTR[score模型](FTRL、 SGD、 L-BFGS、 FFM)

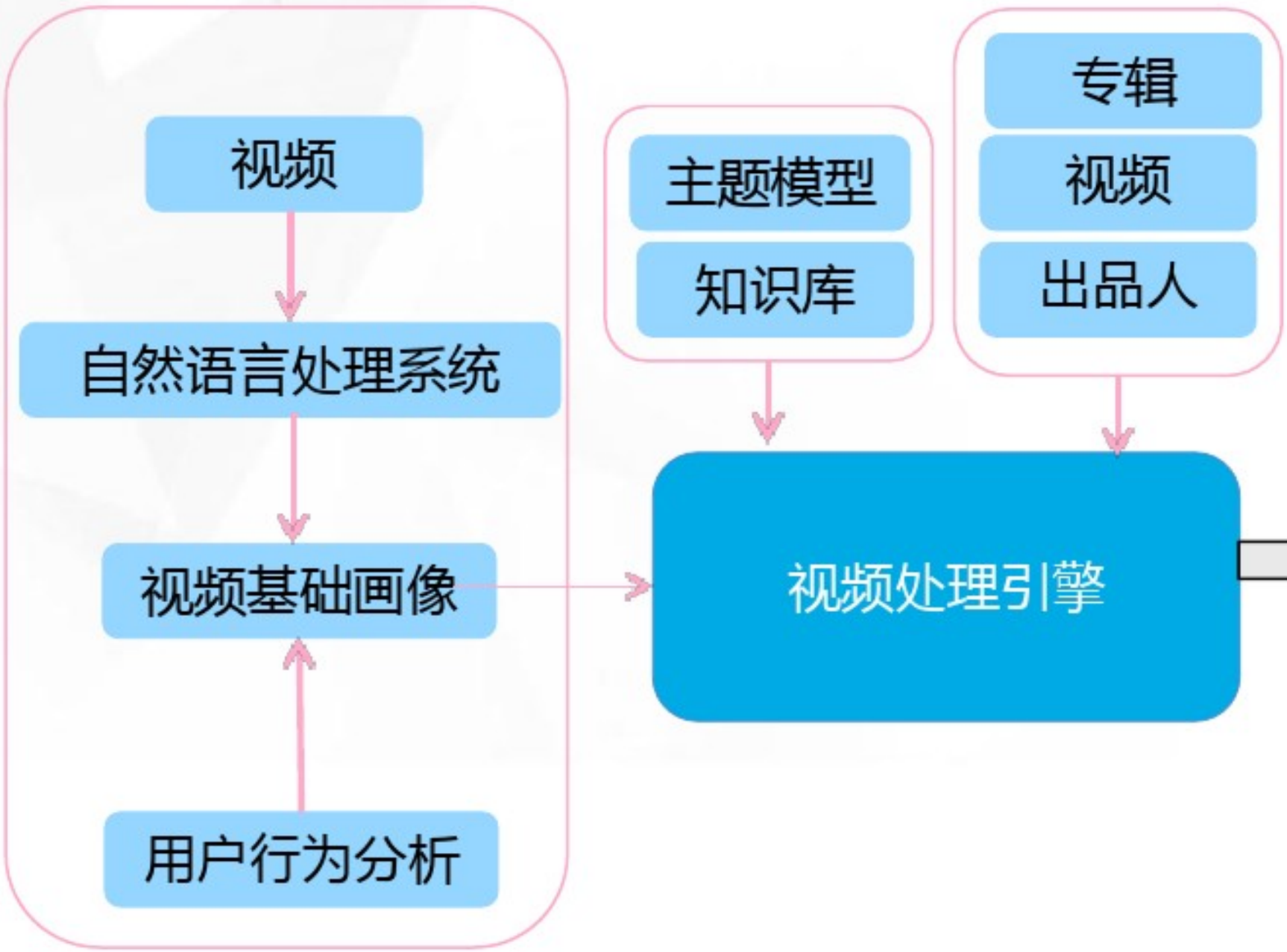
特征工程&排序模型

在线增量学习架构



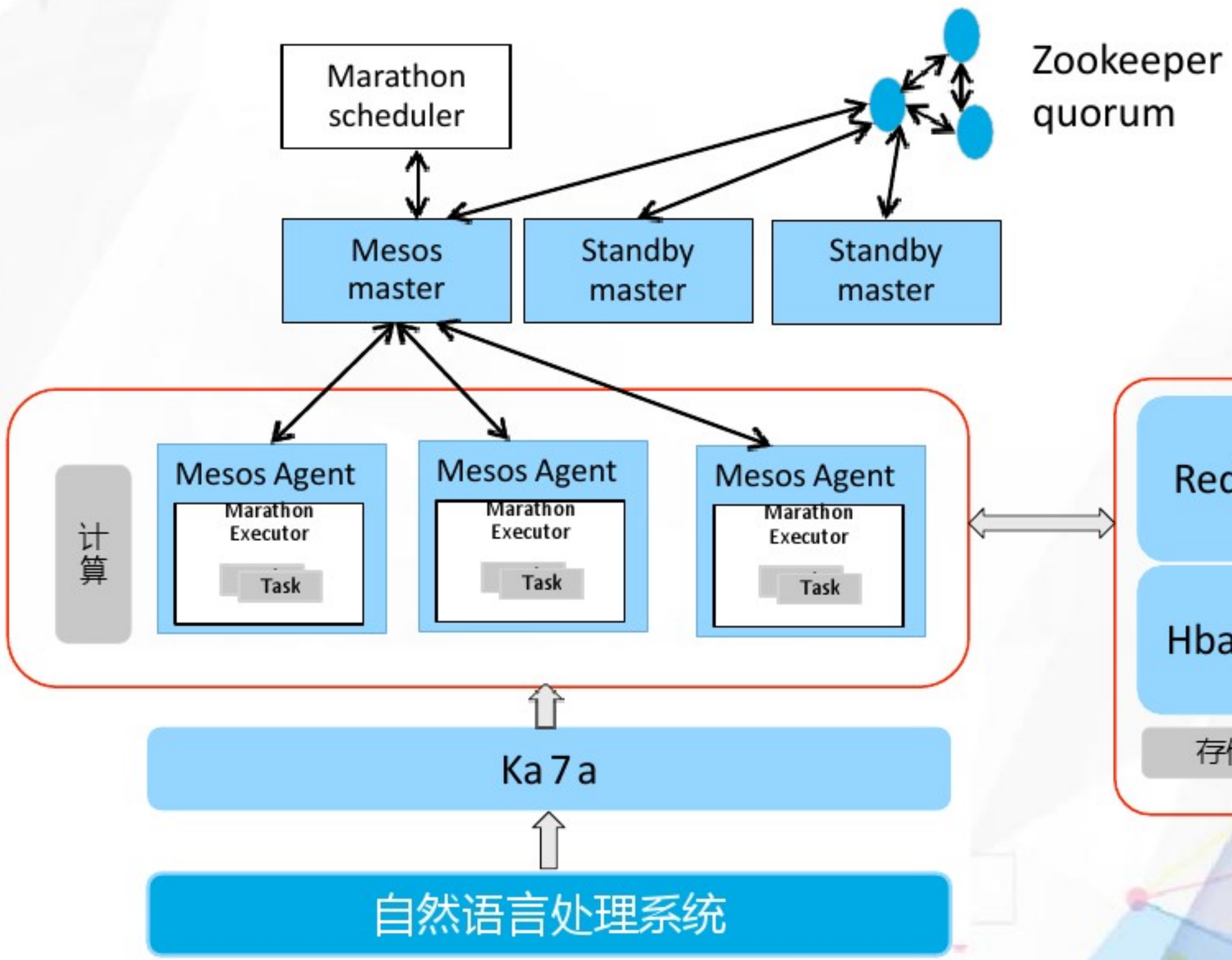
视频处理引擎

视频处理引擎



1、基于TAG倒排视频
2、基于明星、出品人、专辑倒排推荐视频 3、基于规则如：一级、 二级、Topic等卡片封装 和倒排
3、视频画像
4、卡片类型对应Bloomfilter
5、卡片主题分布
6、一词多意Blooomfilter
.....

视频处理引擎


Marathon scheduler
Zookeeper quorum
Mesos master
Standby master
Standby master
Mesos Agent
Marathon Executor
Task
Mesos Agent
Marathon Executor
Task
Mesos Agent
Marathon Executor
Task
计算
Redis
Hbase
存储
Ka 7 a
自然语言处理系统

用户画像&视频画像

视频画像&用户画像

业务数据：个性化推荐用户画像　个性化push画像

业务建模

模型预测：用户活跃度　用户价值　人群属性　潜在兴趣　...

机器学习建模

基础数据：人口属性　兴趣属性　行为属性　...

清洗、结构化、统计建模

原始数据：行为日志　观影日志　展示日志　...

用户画像

地域

一级类

二级类

长视频喜好

明星

短视频喜好

电视剧

用户Level

专辑

主题喜好

出品人

兴趣标签

…

搜索
Keyword

短期喜好

科技	20
IPHONE	5
.....	

长期喜好

CBA	10
新闻	20
汽车评价	5
综艺	4
.....	

场景:追剧、旅行、公交地铁

人群：上班族、学生等

召回模型反馈情况

性别

视频画像

一级类

二级类

关键词

...

适合人群

明星

评分

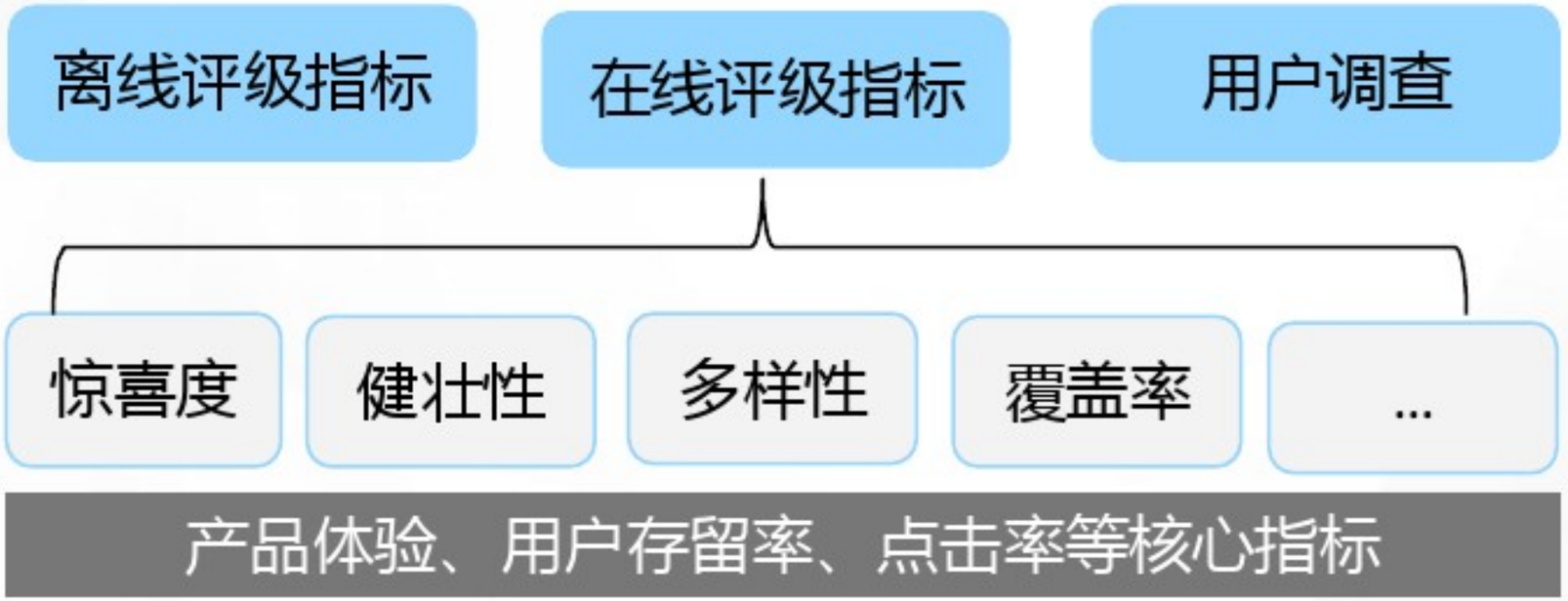
主题

专辑

出品人

如何评价推荐系统

如何评价推荐系统



如何评价推荐系统

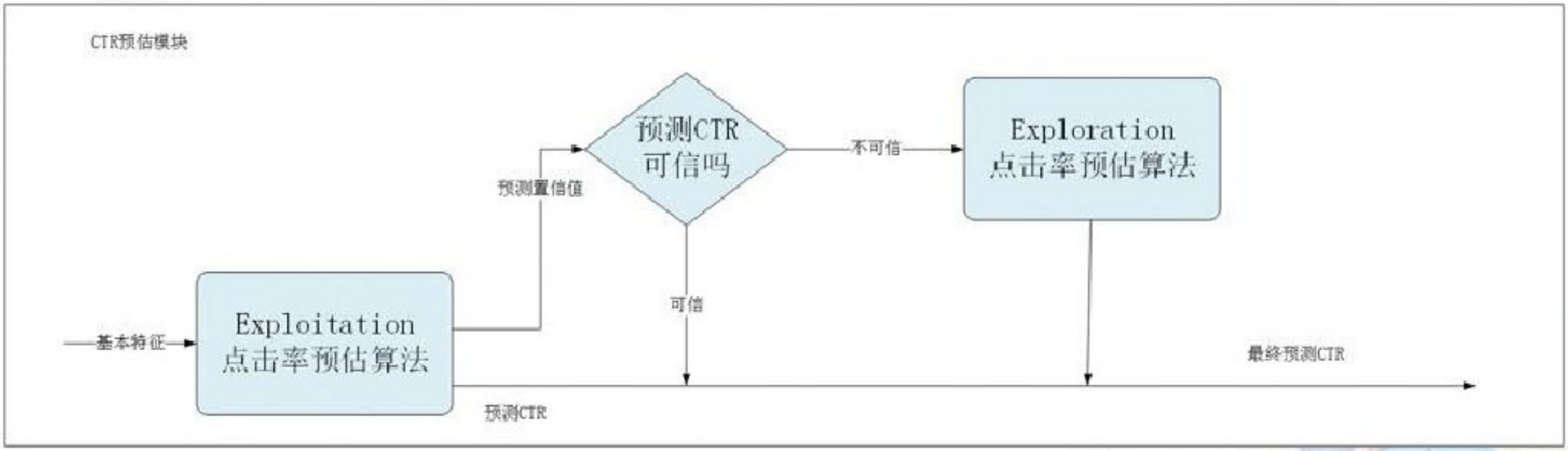
预测CTR可信吗？

- 机器学习是典型data driven的，当训练数据中某种情况的数据不足时，这种情况下的预测值很有可能被其他数据拉偏。

第i维feature非零的训练向量的个数

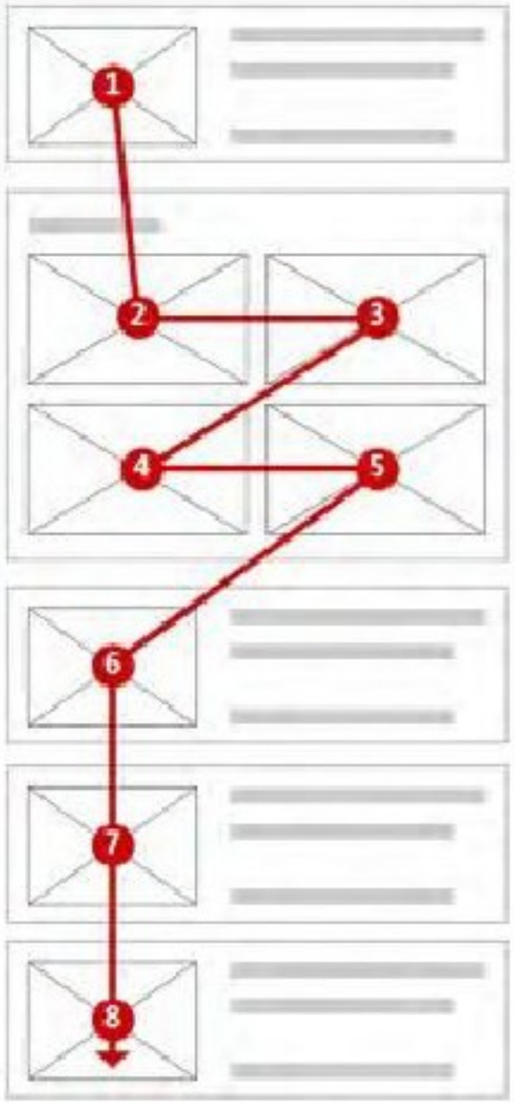
$$\text{confidence} \propto n_i$$

- 训练数据越多则可信度越高



如何评价推荐系统

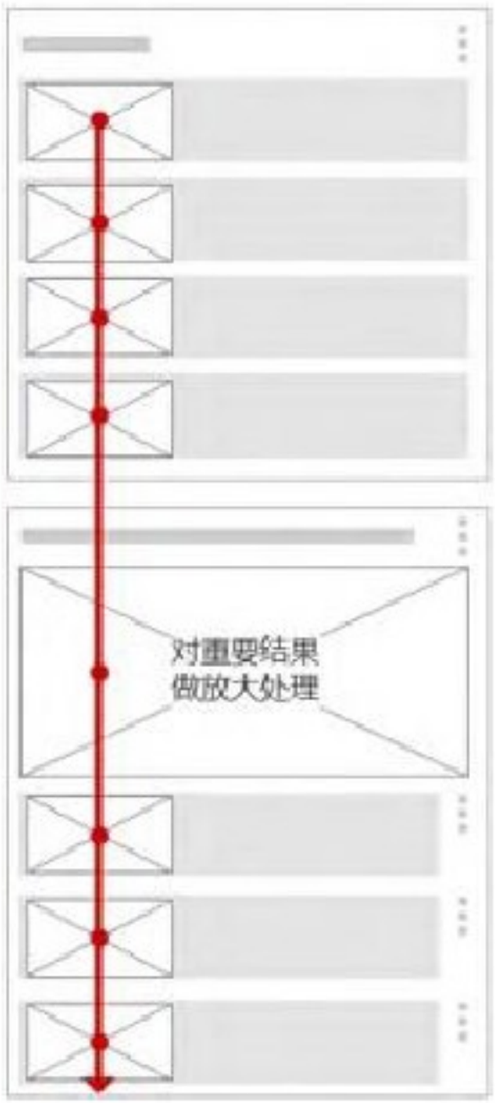
V1



一次用研

- 视线流畅；
- 入口强化；
- 修正回首页误操作；

V2

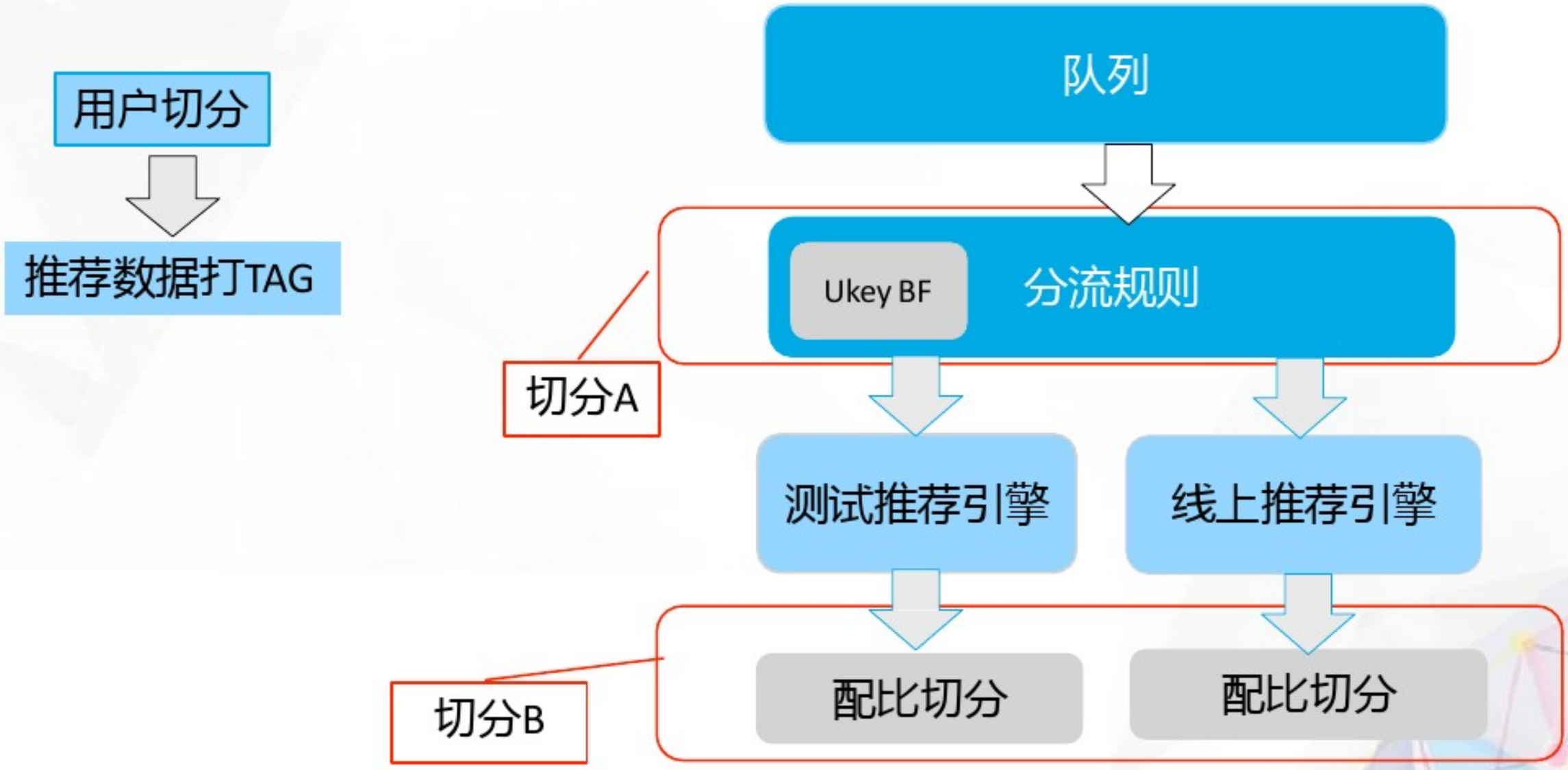


二次用研

- 主题卡片可查看更多；
- 对主题添加参考置出：
  - 头像、推荐理由；
- 其他细节优化：
  - 下加载；
  - 负反馈；
  - 图文宽度占比；
  - 定制模板，非通用模板；

V3

如何测试推荐系统


用户切分
推荐数据打TAG
队列
Ukey BF
分流规则
切分A
测试推荐引擎
线上推荐引擎
切分B
配比切分
配比切分

如何测试推荐系统

PDNA:G;N;T;P;pointer;vid:site;.....

配比编号

P:编码

000,000,000

排序编号

召回模型编号