YYB搜索串讲

2019/12/4 nicoyxhuang

大家想知道啥

- 1、搜索是啥（只知道经常被爆badcase）
- 2、搜索产品只做策略吗，为啥搜索几百年体验没啥变化（产品是不是在摸鱼）
- 3、产品对于算法要懂到什么程度（我不信火哥会懂模型fm）
- 4、搜索和推荐的区别是什么（是不是会搜索就会推荐了）

- 搜索逻辑、垂直搜索/通用搜索之间的区别、新的动态搜索、

- More：基于市场上搜索情况介绍下，对搜索未实现的美好想象？头条的搜索是咋回事？新型的搜索类型，例如语音搜索？（你们随便问，反正我不会 ☺）

内容概览（why、事项、引申）

- 1、搜索是啥
  定义、发展历程、原理
- 2、应用宝搜索的框架、运作流程
- 3、数据源 wiki
  why数据源、针对数据源做了啥、其他思考
- 4、算法层 wiki
  why算法、针对算法做了啥、其他思考
- 5、体验层 wiki
  why体验、针对体验层做了啥、其他思考
- ~~6、数据层 wiki~~
  ~~准备来不及了，后面邀请宗良、飞哥给大家分享 嘻嘻~~
- ~~7、广告层 wiki~~
  ~~准备来不及了，后面邀请杨哥给大家分享 嘻嘻~~

1、搜索是啥——定义（1/3）

- **网络搜索引擎（英语：web search engine）**
  - ①自动从网站搜集特定的信息
  - ②提供给用户进行查询
  - ③将结果展现给用户的系统

- **展现形式：**
  - 搜索结果常以**行列式的链接**展示，称为搜索结果页

- **展现内容：**
  - 这些消息链接可能是连至网页、图像、影片、信息图表、文章或其他类型的文件

参考：https://en.wikipedia.org/wiki/Web_search_engine

1、搜索是啥——发展历程（2/3）

1、分类目录

导航时代，人工整理目录，将各个高质量网站分门别类，无技术手段（yahoo）

2、文本检索

经典的信息检索模型，计算关键词和网页文本内容相关度。但搜索质量不好（altavista,exctie）

3、链接分析

结合网页流行性、内容相似性改善搜索质量，但会结合链接分析，站长作弊（google）--yyb 状态

4、用户行为

个性化搜索，结合用户属性上文（地点、用户画像 etc）进行推荐搜索—yyb 将来的状态

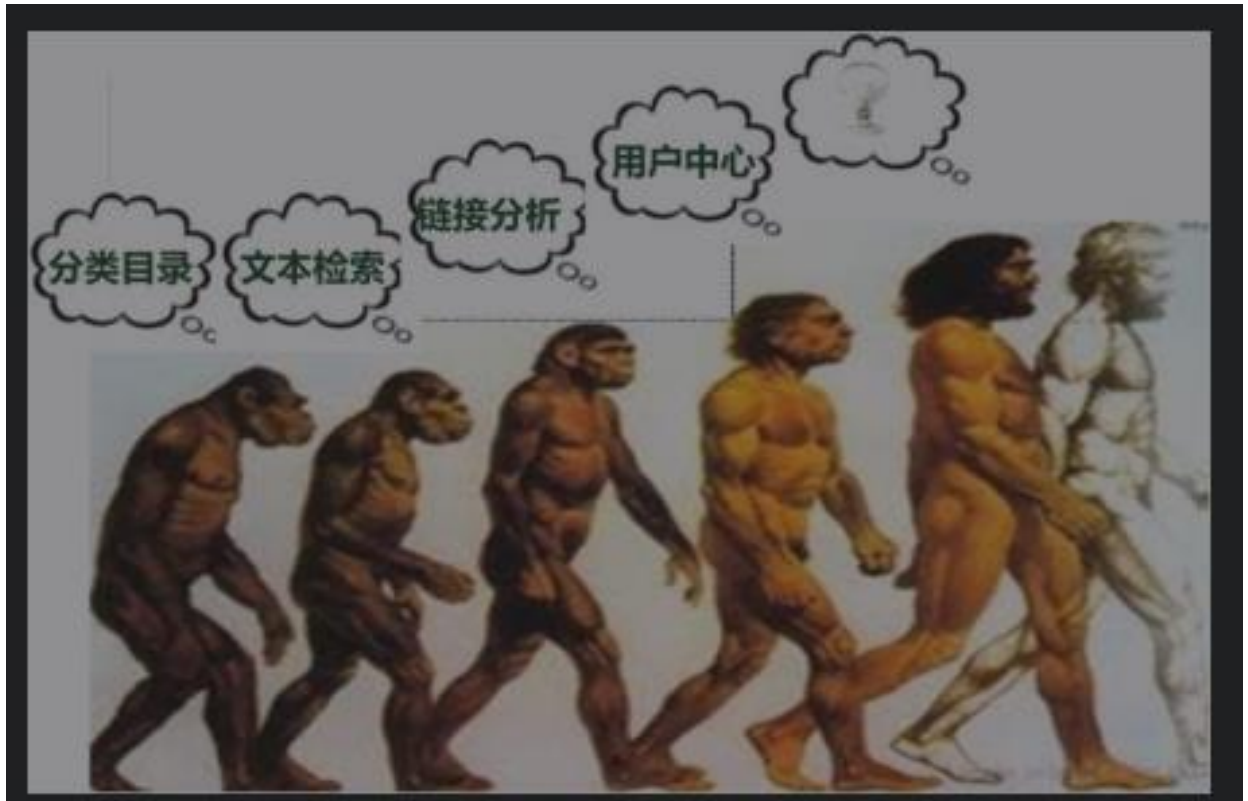
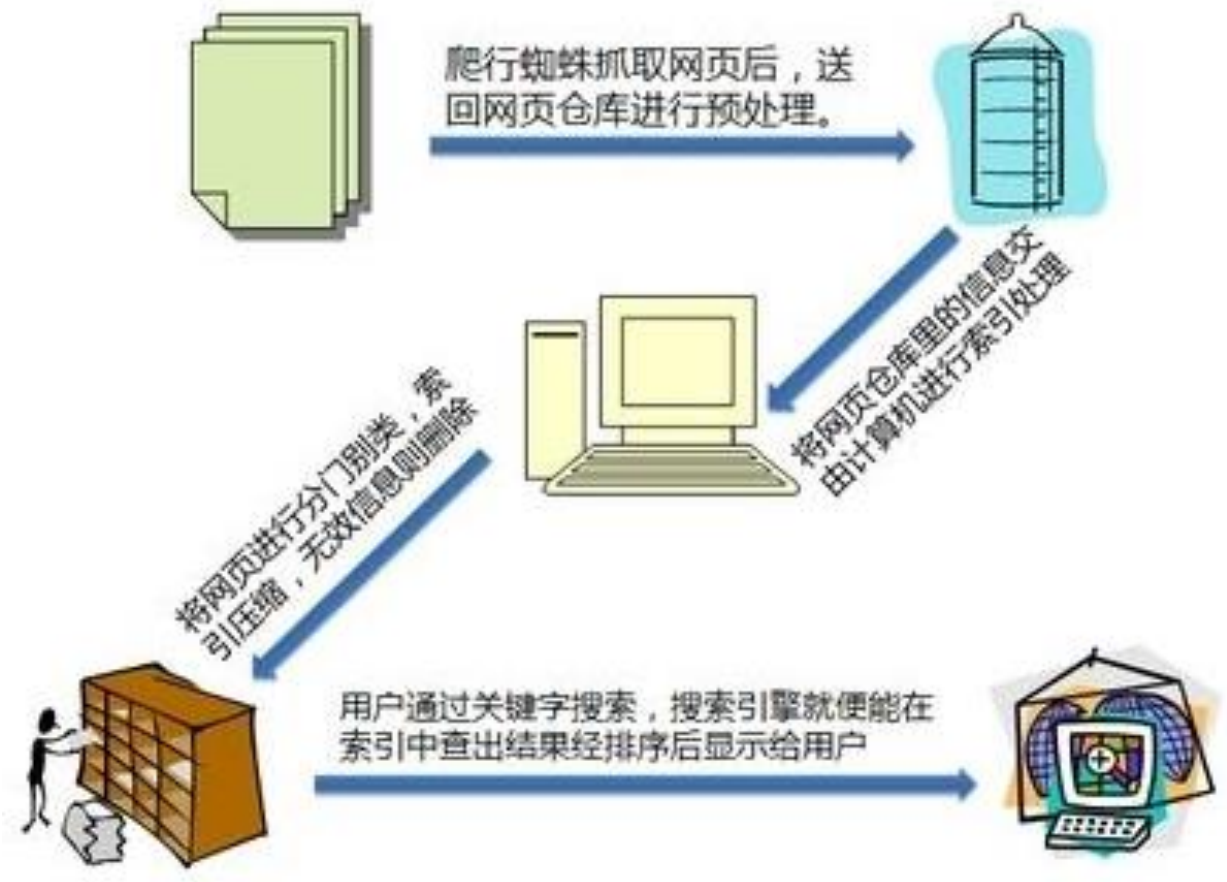


图1-3 搜索引擎技术发展史

1、搜索是啥——逻辑原理（3/3）



1）搜集信息//爬虫 web crawling

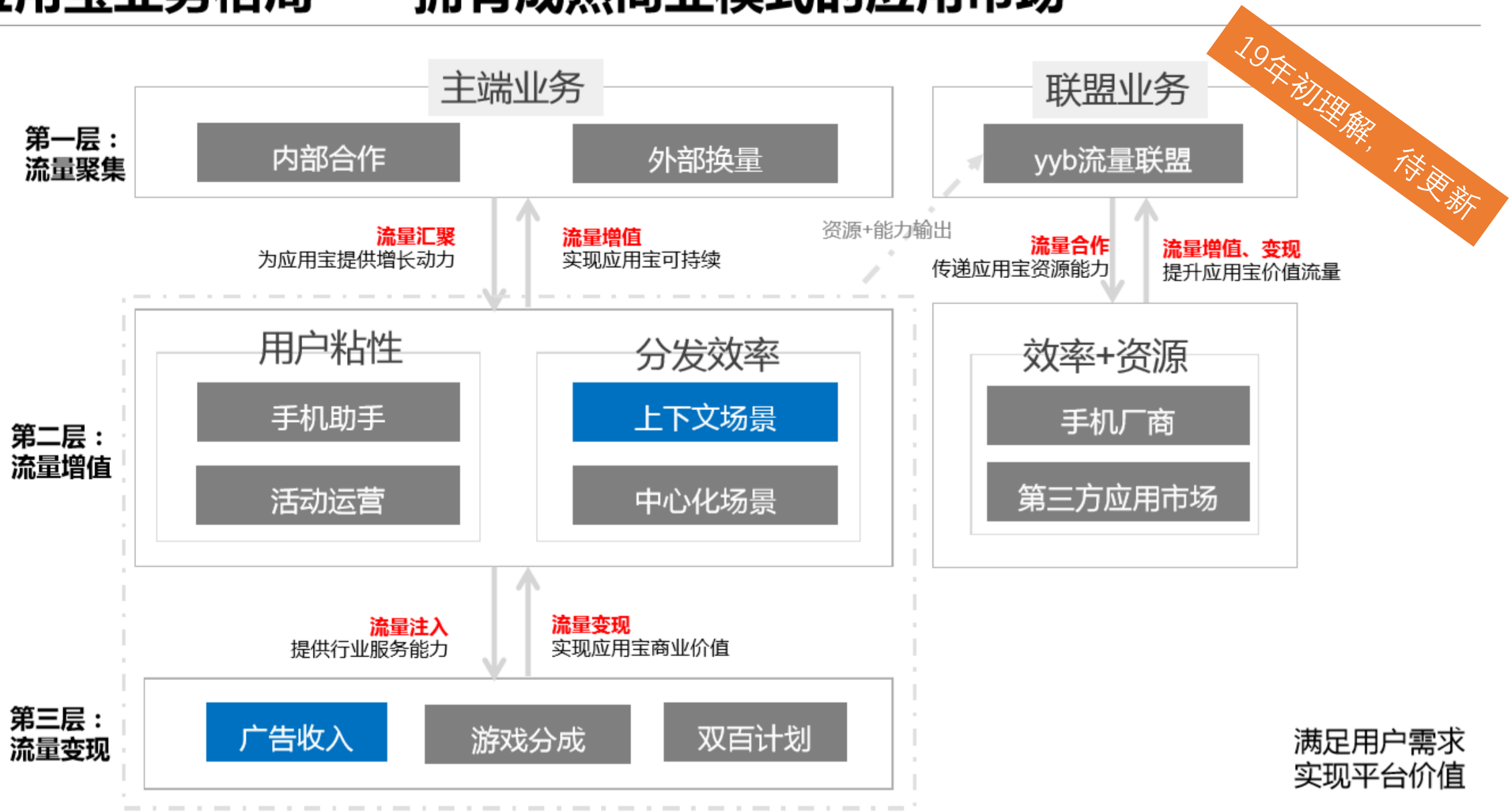
2）整理信息//索引 indexing

3）接受查询//搜索 searching

参考资料：《这就是搜索引擎：核心技术详解》/// NOTES：目前索引是盲区，别问 o(╯□╰)o

开始

应用宝业务格局——拥有成熟商业模式的应用市场



整体规划视图

19年4月视图未更

基础体验，做对，满足用户核心诉求

数据源

引入

- 竞品新热榜单监控
- query集搜索结果集差异引入
- 竞品 url穷举遍历

更新

- 头部app 下载url小时级监控更新
- 长尾app url天级监控更新

- 通用爬取能力建设
- 入库流程优化
- 系统架构、数据存储结构优化
- 场景使用负反馈系统建设

这个是什么?

算法（含广告逻辑）

排序

- 排序机制
- 广告CPM
- 冷启动逻辑

召回

- 文本召回
- 标签召回
- 语义召回
- 标签置信度
- 点击模型
- 文本相关性
- 语义相关性

Query理解

- 精准首位
- 分词/纠错/改写
- 切词权重
- 精准模糊判断
- 意图识别

数据

- 搜索行为
- 用户画像
- 领域专名、实体库
- 通用图谱数据
- 分类标签

能力提升，做好，满足用户潜在诉求

体验（含广告逻辑）

精准首位

- 已收录-激活体验
- 未上线-提示体验

通用推荐

- 精准词BOX
- 模糊词BOX

相关核心场景

- 直达区、结果页：分行业+广告逻辑优化
- 起始页：体验+推荐+广告逻辑优化等
- 联想词：体验+推荐+广告逻辑优化等
- One more App：体验+推荐+广告逻辑优化等
- 相关搜索：细化潜在、交叉搜索推荐等

内容是搜索的根本

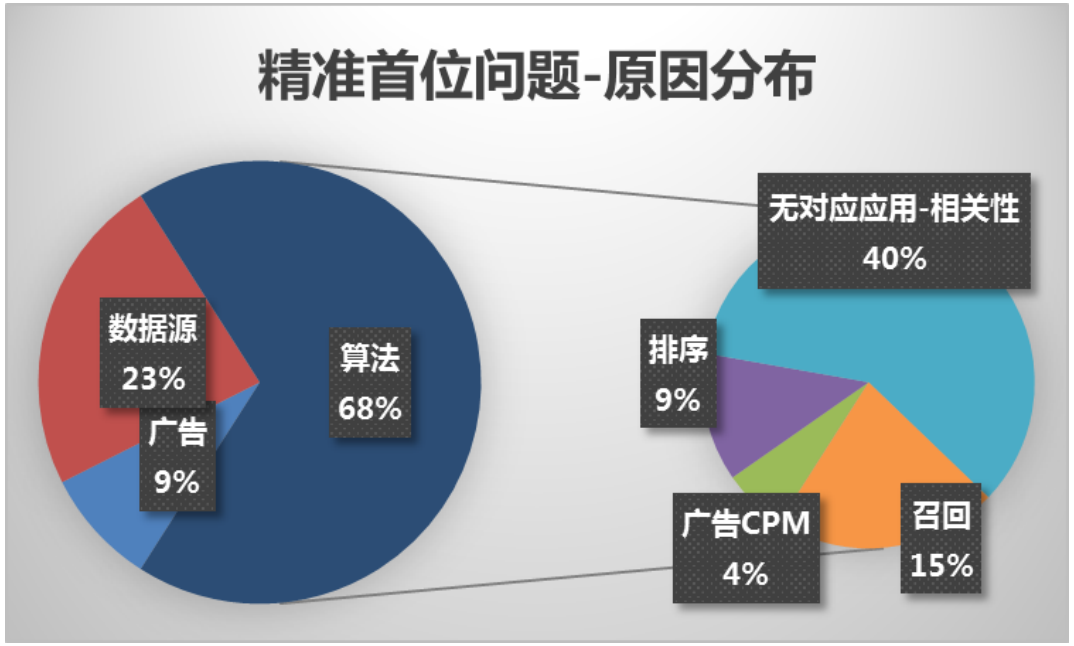
技术决定搜索的体验

初心：各场景定义/终局

附：《今日头条 CEO 朱文佳：做好搜索有三件事最关键》

2、数据源—why 数据源对搜索重要

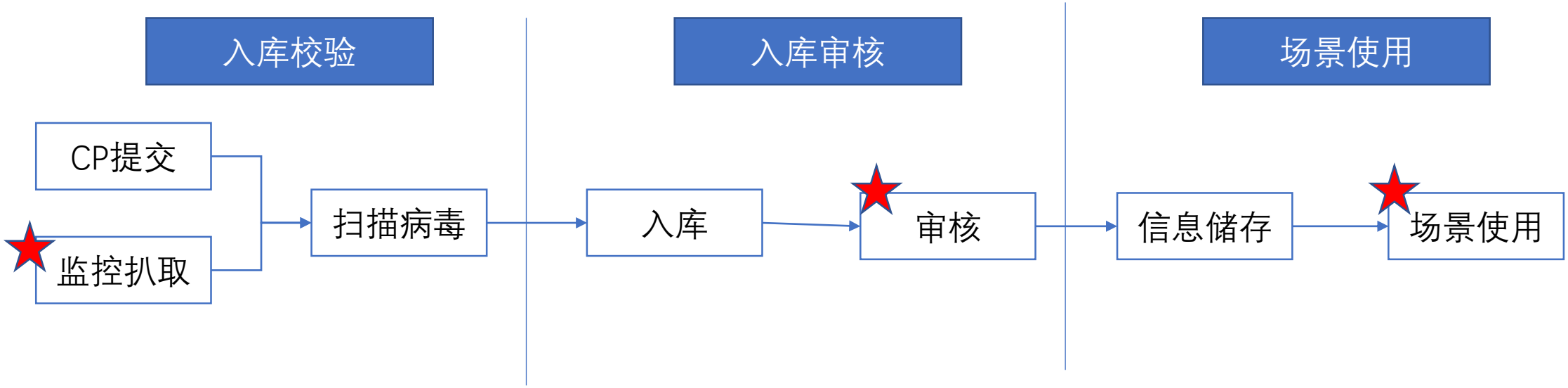
- 主启80%分发产生在搜索，搜索整体分发中有65%-75%产生在精准首位
- **精准首位评测**：badcase中27%算法相关性不足； badcase中23%数据源策略导致未展现



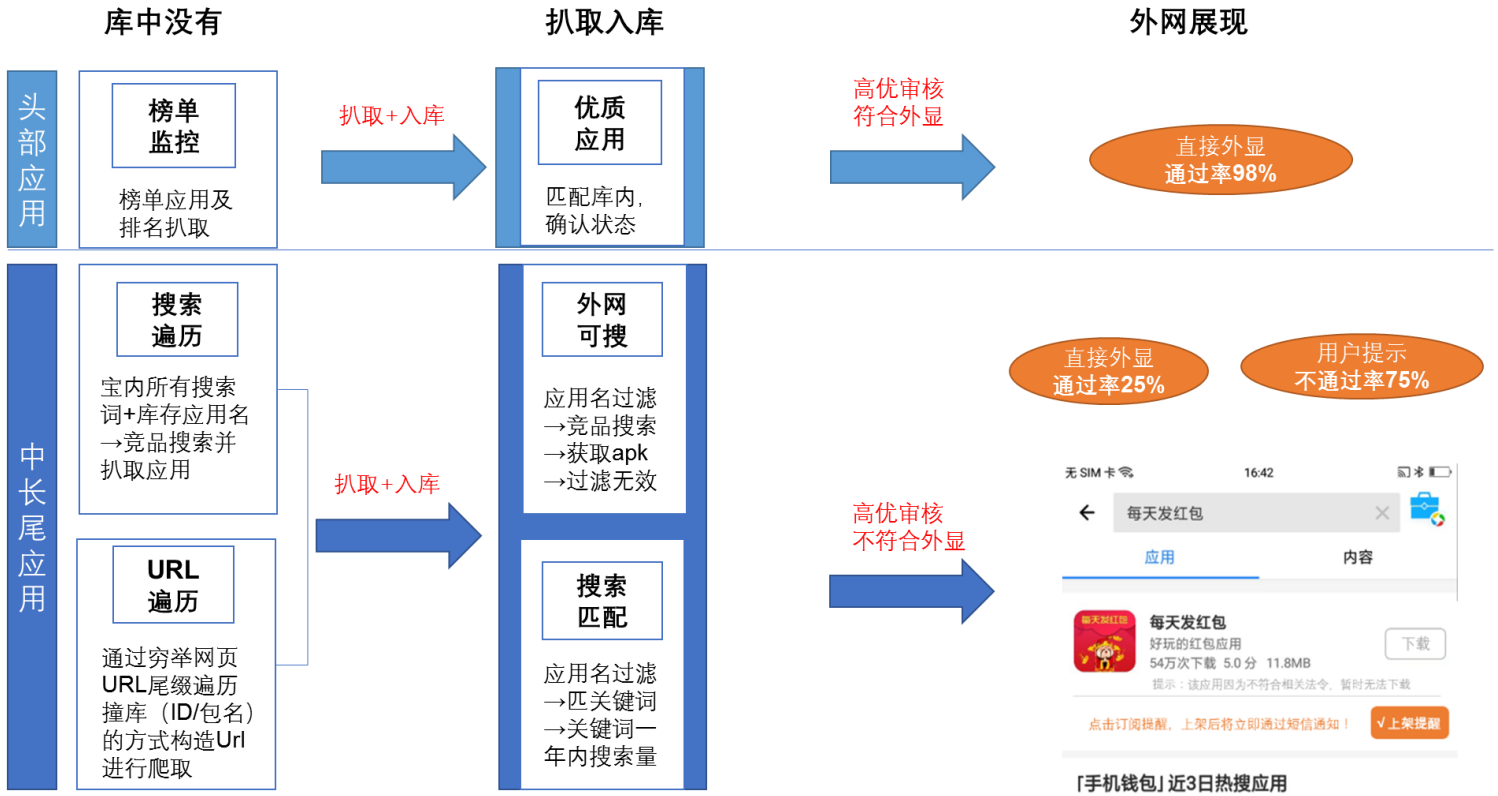
类型	问题模块	case占比	处理方式	时间节点
暂不优化	政策-应用市场竞品	8.33%	维持原状	/
可优化	已扒取未入库	8.33%	提高扒取入库效率，完成存量入库，并实现10分钟级别入库	4月底
	已入库未审核	16.67%	优先完成库中所有竞品可搜我们不可搜应用40w	4月底
	审核后未上线（运营商申请下架）	66.67%	优化外显提示，但不提供下载。同时ios/pc相关应用也会一并优化体验	5月中旬
	审核后未上线（政策-软著）			
	审核后未上线（政策-网信办）			
	审核后未上线（政策-其他问题）			

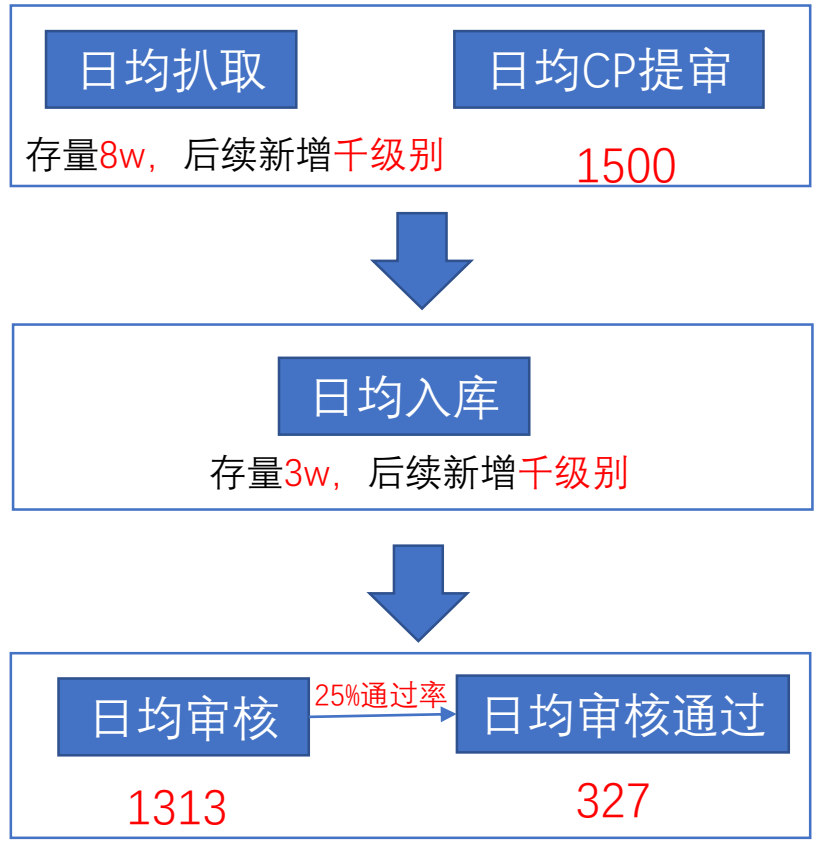
详细数据源逻辑wiki：http://iwiki.oa.com/pages/viewpage.action?pageId=31383725

2、数据源—入库校验、审核、场景使用



2、数据源——入库（监控、遍历、过滤）





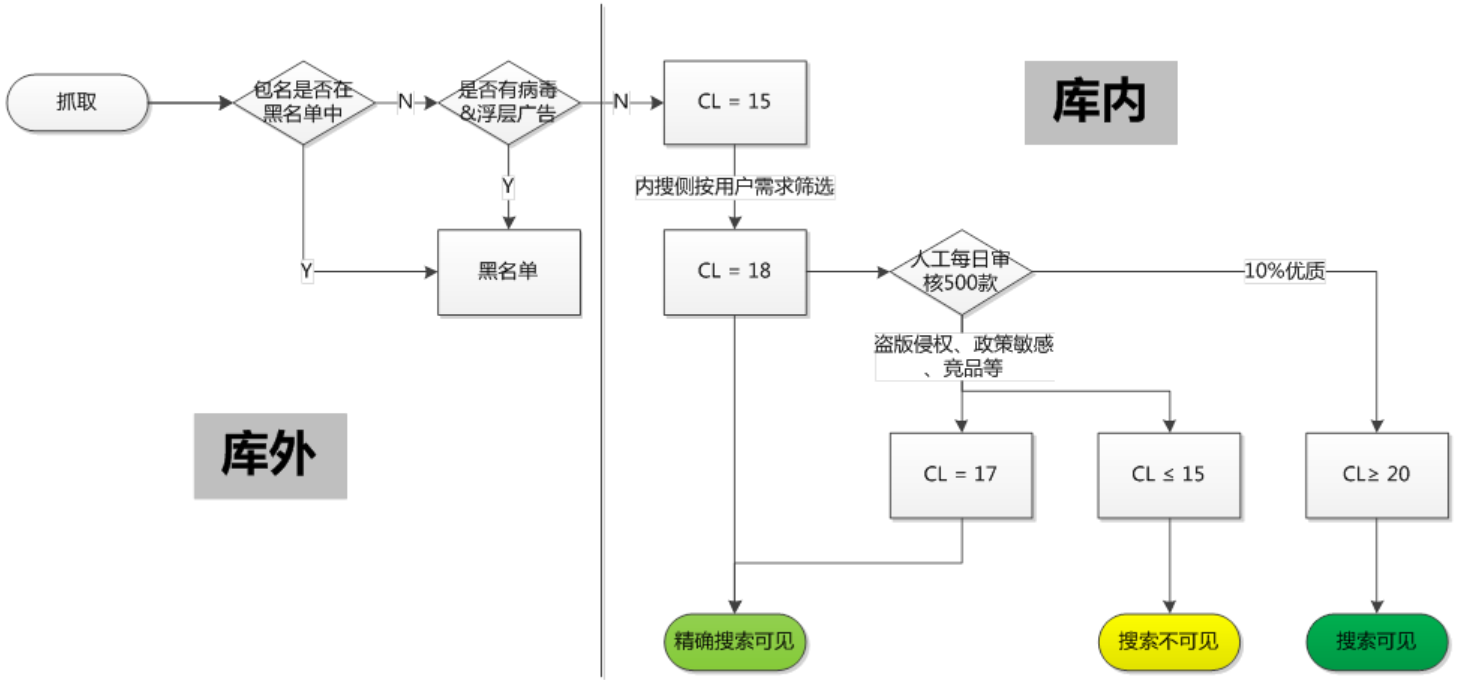
NOTES：

1、定义我们要的内容；2、要的内容从具体什么页面进行抓取；3、根据什么规则进行抓取（逻辑、更新周期）

4、抓取后如何过滤不必要的补充；5、过滤后怎么进行审核/识别优先级判断

https://www.wandoujia.com/apps/7899415 //　宽度优先遍历（抓取URL中包含的链接不断追加）

2、数据源——审核



是否严重病毒	政策/法务风险	竞品影响公司	影响平台权威	用户体验不好

NOTES：

分发重要，但是要考虑长期价值。从宏观层面到微观层面进行分级限制，保障长期利益。

短视的火哥回忆录
https://docs.qq.com/sheet/DYkNrcm1BT09wWUVJ?opendocxfrom=admin&tab=BB08J2&c=C1A0A0

2、数据源——各场景使用

2017年->for 分发
可以被下载，但需要提示用户信息

2018年->for 留存
不可以被下载，需要提示用户原因

2、数据源——其他胡思乱想

- 内容产品（甚至是评论），一定都涉及到审核
  - 审核定级（0或1）还是（n个等级）
  - 先审后发，还是先发后审
- 反作弊和算法的关系
  - 图片OCR（特征->tag->识别），声音识别能力（音频）
  - 召回率、准确率etc
- 内容型产品平台和开发者/内容创作者之间的关系
  - 各个厂商的开发者大会目的是什么？终端扒取 vs 生态合作关系建设
  - 怎么才能让平台内容更多、更优质 even not only apps
  - 扒取的内容是越多越好吗?
- 从事内容类产品的产品趋势，出路在哪里
  - 经济发展程度趋势，内容需求越来越旺盛。Pgc ugc，国家监管能力vs平台审核

3、算法层—why算法对搜索重要

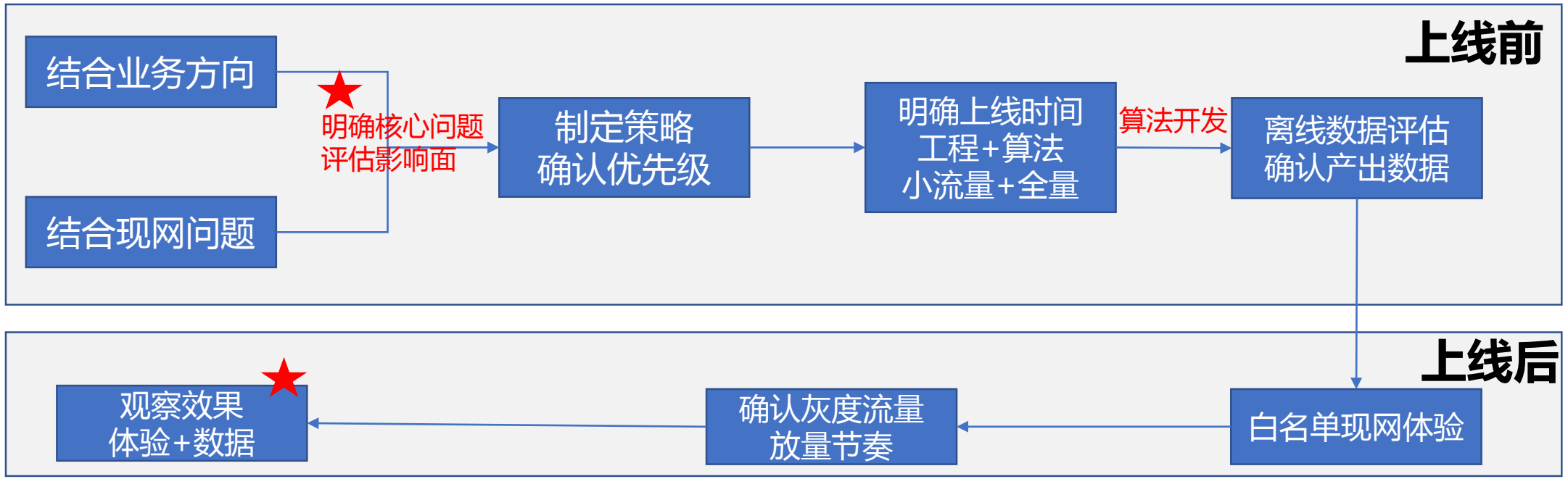
	query数分	query分布占比	qv数量	qv分布占比	下载量	下载分布占比	搜索人数分	搜索人数占比
头部	385	0.00%	357008772	78.29%	1715798	34.47%	381133	6.35%
腰部	16650	0.16%	71820894	15.75%	2145884	43.10%	855860	14.27%
长尾	281475	2.71%	20682477	4.54%	555314	11.15%	2751427	45.87%
超长尾	10088647	97.13%	6484445	1.42%	561299	11.27%	3414459	56.92%
总量	10387157	100.00%	455996588	100.00%	4978295	100.00%	5998549	100.00%

问题反馈时间	问题	问题等级	场景	反馈人	产品负责人
2019/10/20	搜Moo Diary出不来		搜索	mainding	nicholexu
2019/10/20	搜stock，股票类应用出不来		搜索	mainding	nicholexu
12	2019/10/28	facebook和face book搜索结果完全不一样		搜索	maining	nicoyxhuang
2019/11/3	热搜box 1、小红书tag出的不合理（tag不对） 2、tag内出的app不合		搜索	kinto	nicoyxhuang

序号	问题反馈时间	问题	问题等级	场景	反馈人	产品负责人
21	2019/11/6	“录屏”搜索结果一二位不准确 		搜索	maining、nealniu	nicholexu
2019/11/26	搜 演员请就位 和 吐槽大会第4季。直达区能出腾讯视频，结果页第一屏不出腾讯视频 				maining-丁飞	Nianhuaxie-谢年华

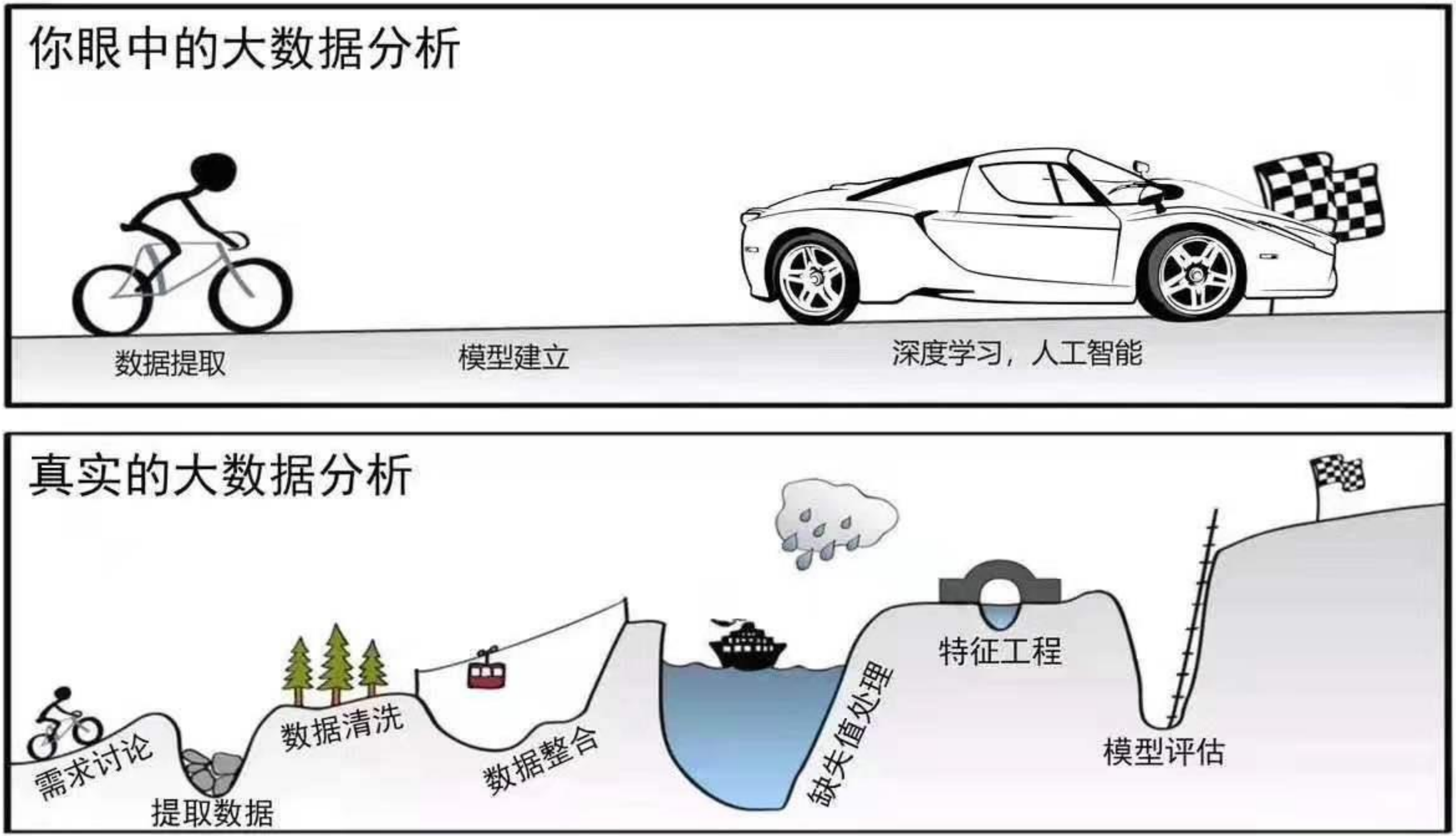
Moo diary	
stock	
Face book	
录屏	
演员请就位	
吐槽大会第4季	

3、算法层—搜索产品涉及工作内容

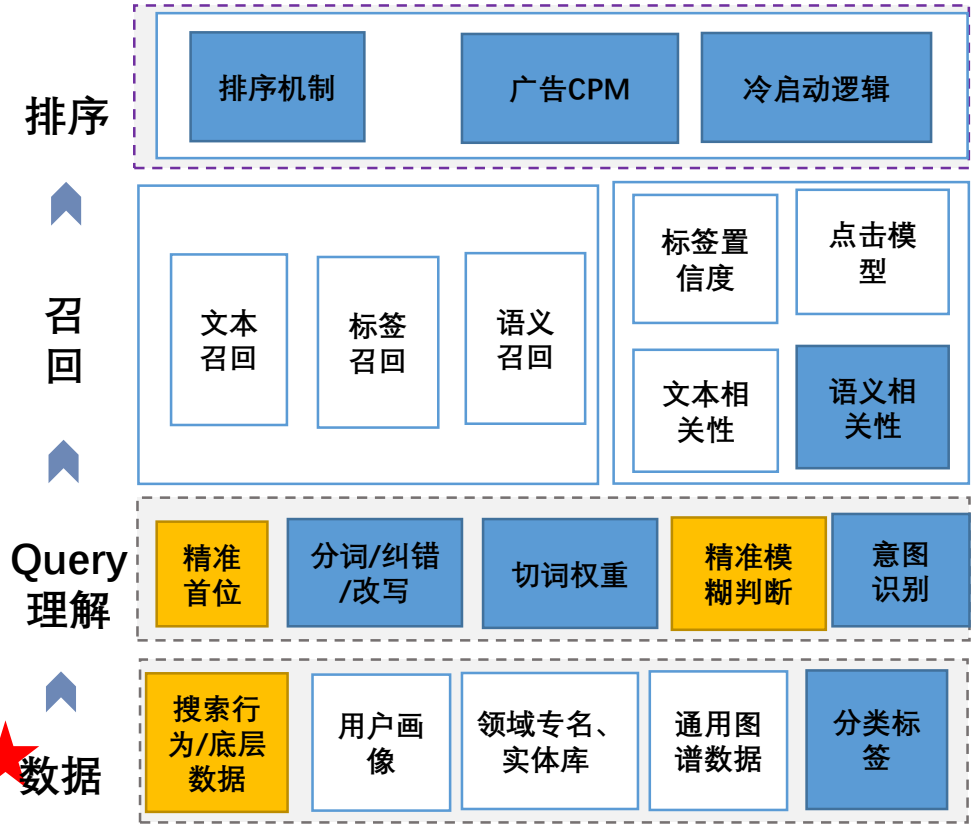


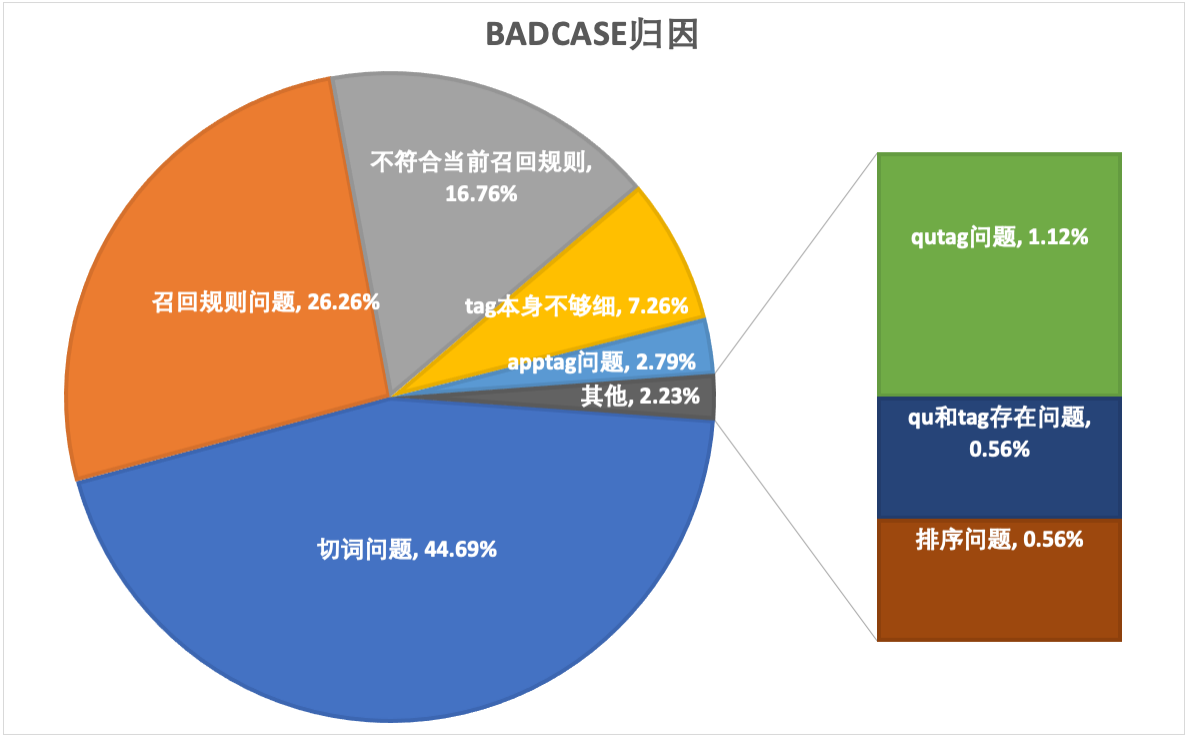
NOTES

1、定义问题；2、定位问题原因；3、明确问题后怎么做；3、怎么评估效果


你眼中的大数据分析
数据提取
模型建立
深度学习，人工智能
真实的大数据分析
需求讨论
提取数据
数据清洗
数据整合
缺失值处理
特征工程
模型评估

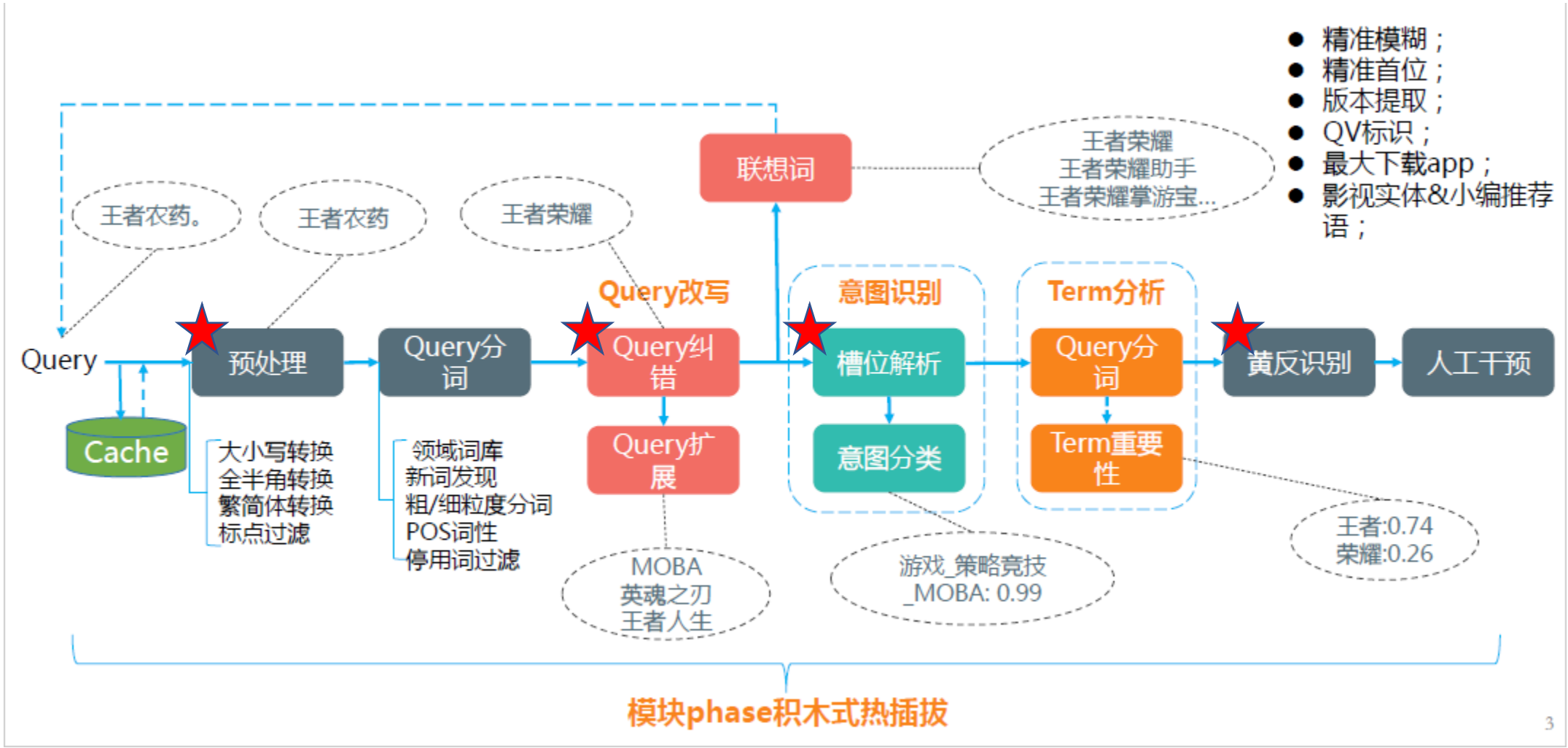
3、算法层—逻辑：底层数据、QU、召回、排序



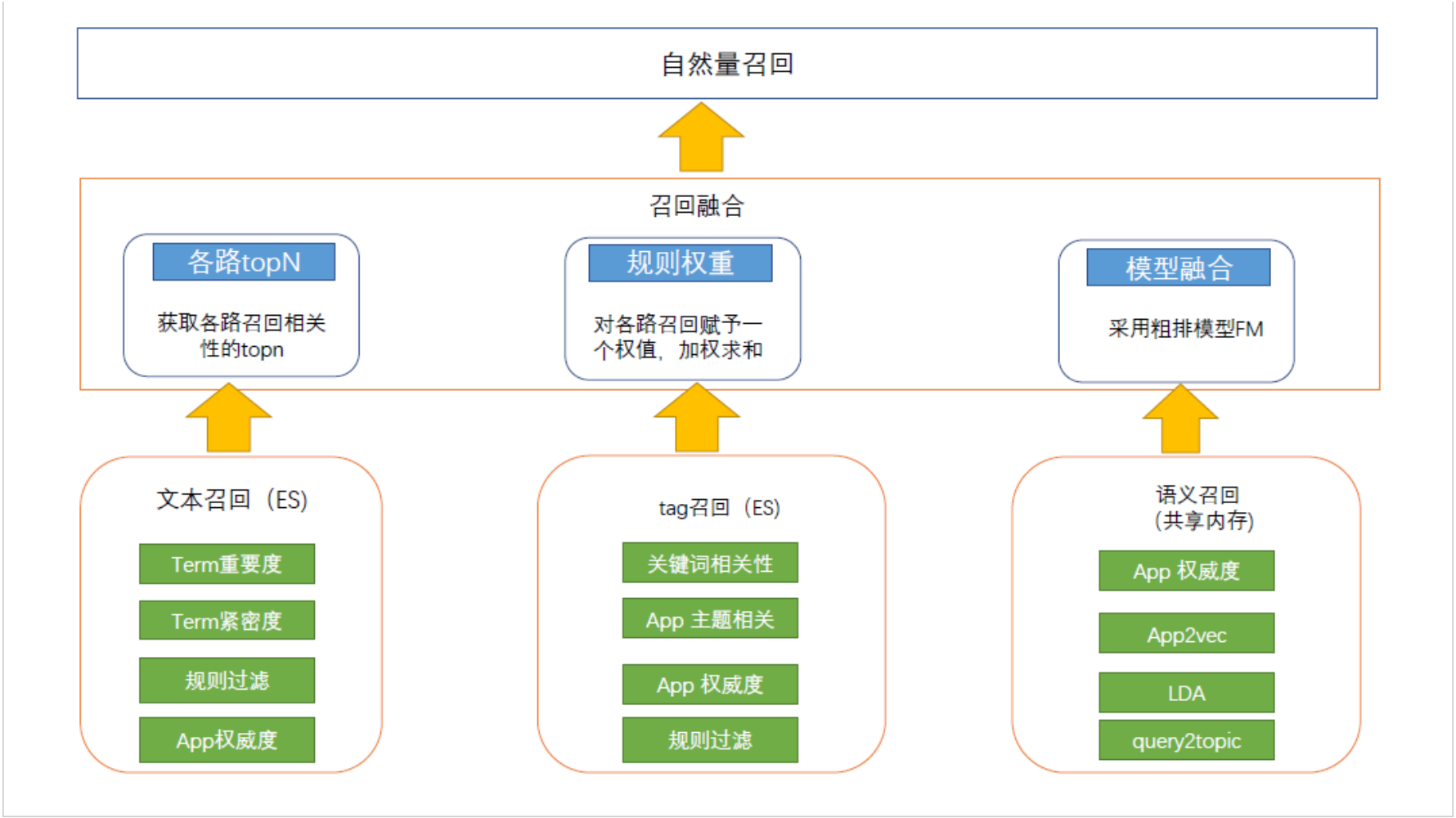


1、没过刷量 tdw；2、标签更新换代太快，叶子节点匹配，数据清洗，模型训练需要时间；3、appinfo索引召回失败

3、算法层—逻辑：底层数据、QU、召回、排序



3、算法层—逻辑：底层数据、QU、召回、排序



1、当前方式为规则
Y=ax+by+cz

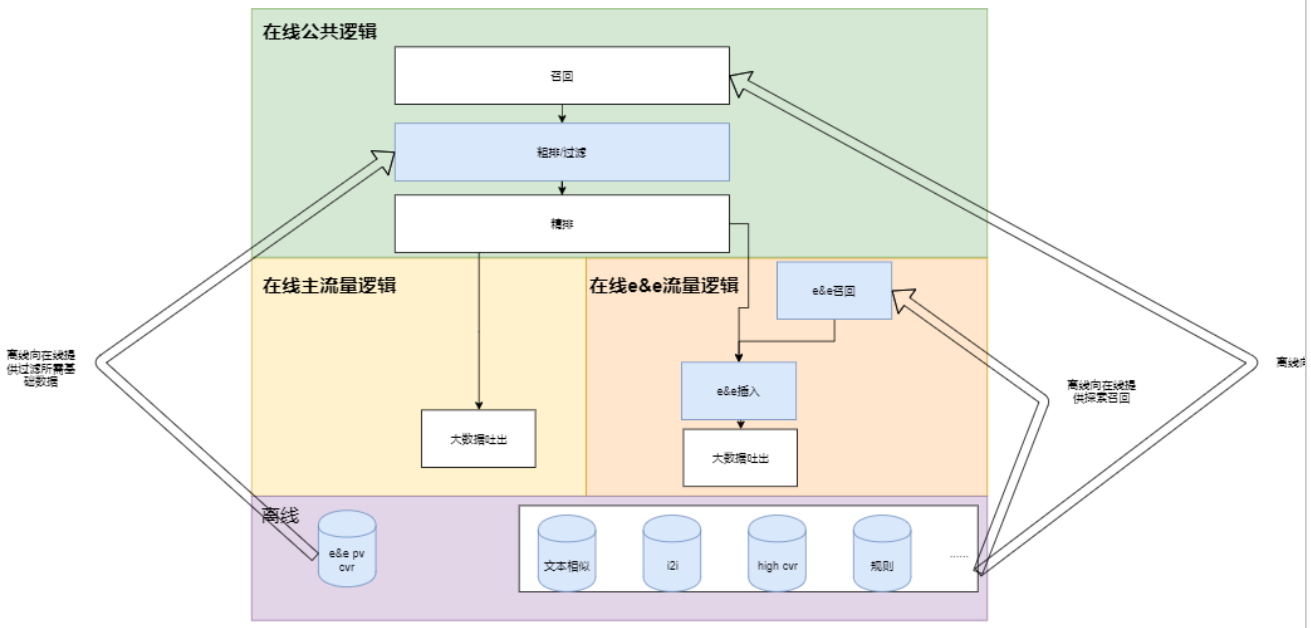
2、可适当了解LR、FM

资料

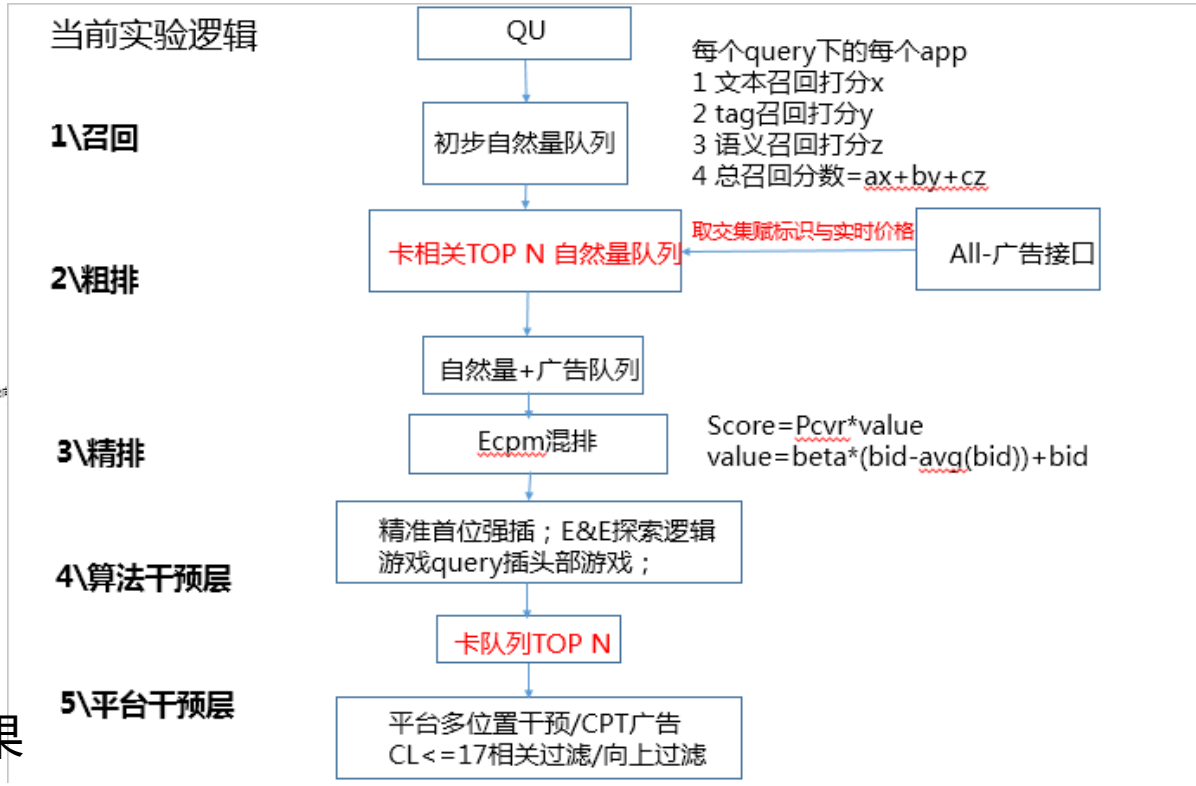
3、算法层—逻辑：底层数据、QU、召回、排序

- 现实情况：卡后验CVR；理想情况：用预估CVR（参考浩宇ppt）

直达区



结果页



针对直达区query-app的前30日cvr结果，剔除万分之6以下的结果

➢ 剔除的应用进入算法ee探索（按预期收益给予曝光机会，积累一定阈值的曝光后满足cvr条件会扩充至主流量召回中）

3、算法层——怎么定位问题

- 1、*定义问题；2、发现问题；3、问题归因；4、问题优化

精准query（包含语义完整app名称）

核心判断标准：1、应用核心功能是否相关；2、文本是否强相关；3、语义与功能是否相关

相关性召回位置(注：为召回位置)	1分——准确	0分——不准确	通俗理解
精准召回（<=1）	1.query与appname前缀匹配 例：**王者荣耀-王者荣耀** **王者荣-王者荣耀** **百度-百度** 2.query与app别名文本匹配，基于领域知识，能推断完全指向某app。 例：**b站-blibli**	不满足左侧标准	
强相关召回(<=4)	1.文本部分匹配，且app为头部 例：**百度-百度输入法、百度网盘** 2.若query是游戏query，app tags 包含query意图分数最高的tag 。 例：**三国杀-英雄杀** 3.60天去刷量cvr>2%(pv>1000)	不满足左侧标准	1、文本部分匹配且是头部app 2、query指向的app核心玩法与本app一致。 3、后验cvr大于2%
弱相关召回(>4)	1.app tags 与 query tags有交集。 例：**百度-QQ浏览器** 2. 文本部分匹配但app非头部 3. 60天去刷量cvr>0.06%(pv>500)	不满足左侧标准	1、文本部分匹配。 2、query的tag和apptag有交集即可。 3、后验cvr大于0.06%

模糊query（非appname、单字、tag词、标点等）

核心判断标准：1、应用核心功能是否相关；2、文本是否强相关；3、语义与功能是否相关

相关性召回位置(注：为召回位置)	1分——准确	0分——不准确	
强相关召回	1.app名包含query文本，且app为头部（头部：历史30天，在应用宝搜索分发大于2000 or 广告主 or 历史下载量大于100w） 例：**小-小红书、小猿搜题** **b-blibli、百度、boss直聘** **枪战-穿越火线、枪战王者** 2.tag词-指向以该功能（游戏则是该玩法/主题）作为核心功能（玩法/主题）的头部app 例： **吃饭-饿了么、大众点评、美团** **打车-滴滴出行、曹操专车** 3.60天去刷量cvr>1%(pv>1000)	不满足左侧标准	1、文本部分匹配且是头部app 2、query指向的app核心玩法与本app一致。 3、后验cvr大于1%
弱相关召回	1.app tags 与 query tags有交集 例：**小-作业帮** 2.app 名包含query文本，不限头部 例：**吃饭-吃饭饭** 3. 60天去刷量cvr>0.06%(pv>500)	不满足左侧标准	1、文本部分匹配。 2、query的tag和apptag有交集即可 3、后验cvr大于0.06%

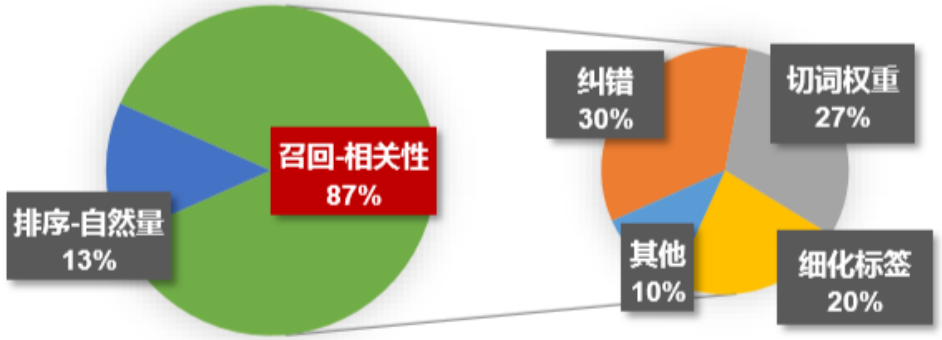
NOTES：问题的本质是什么？是排序不合理，还是召回不合理；引申话题，不想推荐的应用是降权还是走checklevel

*1）满足规则且规则合理则无视；2）不满足规则且规则合理，则case；3）不满足规则且规则不合理，则该规则；

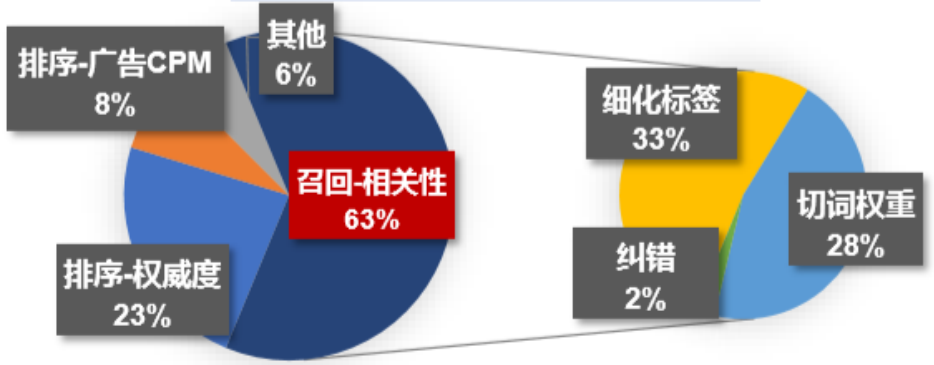
3、算法层——怎么定位问题

- 1、定义问题；2、发现问题；**3、问题归因；4、问题优化**

精准首位算法问题分布



2-10位相关性算法问题分布



更新进展

Case归类	优化模块	核心进展
5月底 **意图识别（关键词理解）** GSB长尾badcase率降低到3%	关键词意图优化（启动两轮优化）	- 体验: 随机抽样200条，G:B= 36:18 - 数据：相对提升1.15%，分发提4.07w/d
	关键词纠错（启动两轮优化）	- 体验：明显提升。随机250条,G:B=168:34； - 数据效果：待周一更新，0517小流量实验
	关键词分词权重	- 数据实验分析中，预期523小流量实验
6月中 **召回相关性** GSB长尾badcase率降低到5%	优化标签体系	- 完成外网标签挖去及app-tag的关系映射 - 预期5月底前完成 app-keyword的优化
	【相关性模型】	- 已上实验 - 体验及数据需再实验，依赖意图和纠错优化
6月底 **排序** 整体分发提升2%	冷启动/中长尾排序	-待启动
	商业化排序	-待启动

3、算法层——效果评估（流程）

评估流程

优先级

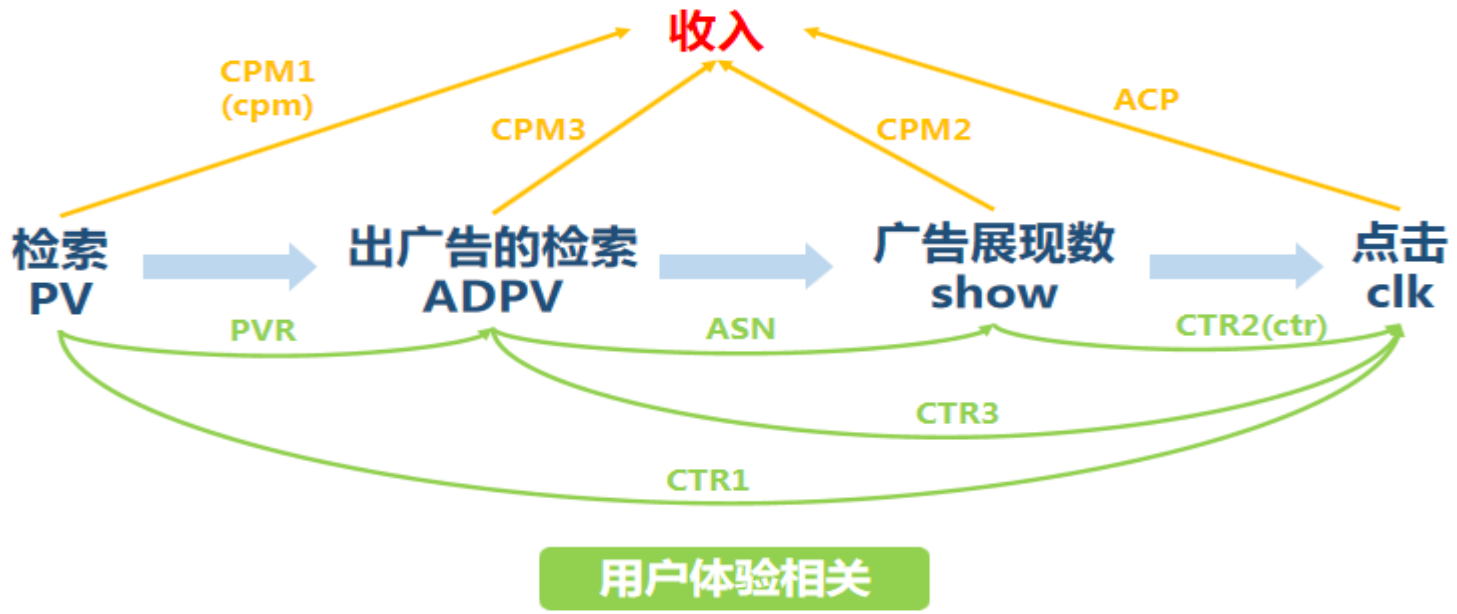
需求目的	示例实验	预期效果	重点关注	辅助验证
体验提升	去刷量	保证分发和收入持平/提升，提用户体验	GSB体验	分发+收入
分发提升	联想词重构	保证体验和收入持平/提升，提升分发量	分发数据	GSB+收入
平台价值	切换标签	保证体验分发和收入持平/提升，提用户体验	GSB	分发收入体验
收入提升	广告保序	保证体验和分发持平/提升，提商业价值	广告收入	GSB+分发

- 针对本实验，是否可放量
- 可放量，是否还可有优化空间及下一步//不可放量，原因及迭代下一步是什么

3、算法层——效果评估（debug）

1、 结果页分桶数据，重点在同步和下一步

	曝光用户	App曝光次数	CVR	分发系数	广告曝光占比	广告cpm	广告cvr	广告Arpu
对照组	21577	870184	1.39%	55.96%	12.50%	45.30	0.016	22.82
实验组	104985	3304646	1.80%	56.75%	9.43%	77.50	0.028	23.02
绝对值			0.42%	0.78%	-3.07%	32.20	0.01	0.20
对比值		-28.12%	29.91%	1.40%	-24.56%	71.08%	75.31%	0.88%



用户体验相关

批注数(3)

房ny 编辑 删除

1、是分发提升涨的多了，还是无效曝光降的多了

2、为什么广告cpm涨这么多，cvr持平没有变化，按理cpm和cvr之间的关系应该比较线性，还是说单价跑的更快。明显的query是哪些，有什么共性

3、分发系数、cvr持平略涨是否符合预期？置信涨还是正常波动。

房ny 12-11 15:51

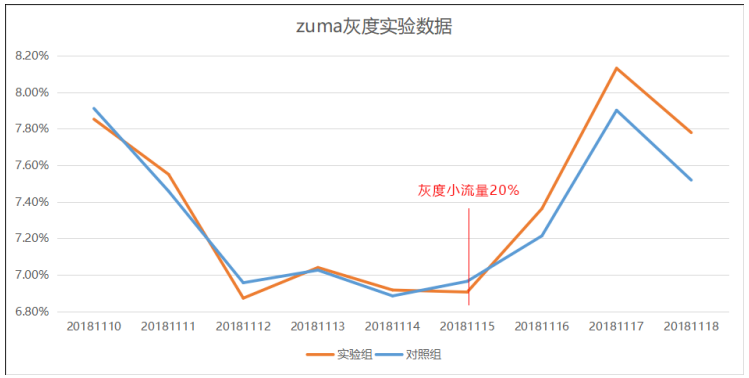
4、广告曝光占比为啥会下降？广告曝光减少和位次曝光占比变化的趋势是否符合预期。从哪些位置开始广告曝光降幅比较多？新的规则是在哪方面对广告产生了影响

3、算法层——效果评估（汇报）

1、数据侧，自上线以来

- checkpoint1：分发系数相较主流量提升XX（相对值），预期全量可提升XX（绝对值）
- checkpoint2：广告ARPU相较主流量提升XX（相对值），预期全量可提升XX（绝对值）
- checkpoint3：头部游戏相较主流量提升XX（相对值），预期全量可提升XX（绝对值）
- LTV xxxx

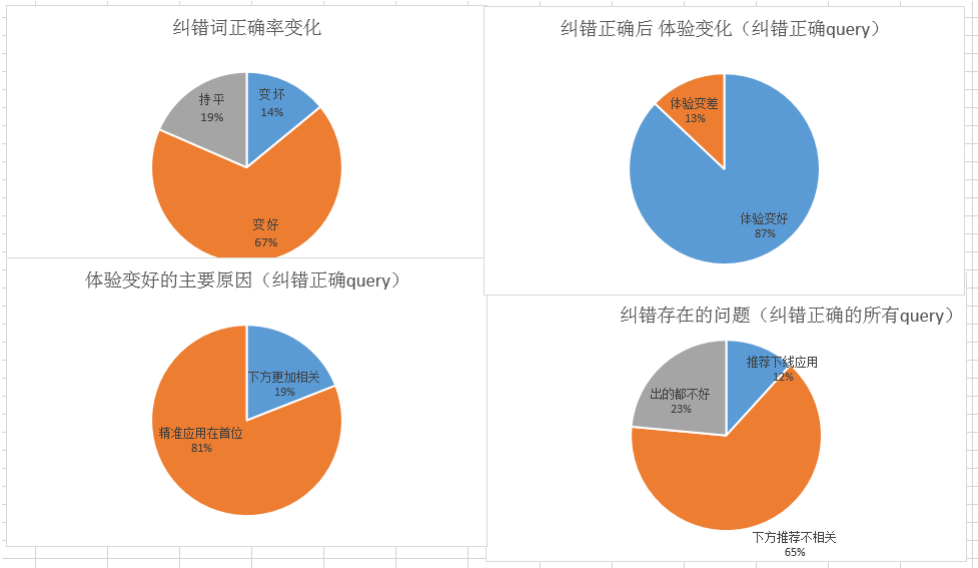
*趋势图包括：上线时间节点，实验组对照组持续至少3天的效果对比，节点明显



2、体验侧，自上线以来 示例

- checkpoint4：相较原来流量/竞品good case 占比，soso case、badcase占比
- checkpoint5：核心要提升效果效果如何，XXX

*数据包括：上线时间节点，取样样本，取样标准



3、TODO

- 针对本实验，是否可放量
- 可放量，是否还可有优化空间及下一步//不可放量，原因及迭代下一步是什么

同时根据前期的预期目的，关注的效果重点（数据呈现+分析人力资源）要有倾斜

3、算法层——怎么评估效果（工具篇）

- 打破“黑盒”，每路结果都可以溯源。
- 确保模拟结果和真实结果一致，从后台链路发现实际问题（cvr、testid ruleid）
- 评测也可以在该工具中进行

召回结果

真实结果

http://zuma.oa.com/zumasearch_v3/

3、算法层——其他胡思乱想

- **产品针对算法需要了解到什么程度**
  - 区分规则和模型
  - 模型并非难懂，结合数理统计、机器学习、在线材料可大概掌握
- **搜索和推荐真的有很大的区别吗**
  - 上文≠query
  - 当前的上文还有很大的挖掘空间
- **哪些逻辑在召回做，哪些逻辑在排序做**
  - 学习算法同学对于召回的理解，提高相关性、降低能耗（且不说正确与否）
  - 避免惯性思维，即使是最常见的定义/流程，也要思考其存在的意义、必要性、具体使用方式
- **推荐学习内容**
  - 公众号：深度传送门、机器之心、datafuntalk
  - 公司K吧：WXG技术能力提升

4、体验层—思考方式3圈+gap

Link：how_great_leaders_inspire_action

NOTES

1/ 所有自认为想清楚的why都是阶段性的想清楚，需要不断的提升认知，螺旋式

2/ 不敢怼老板的产品不是好产品 (*^▽^*)

4、体验层—思考方式3圈+gap

VISON CURRENT GAP（期望-现状-差距）

- FOR自己：怎么做，做什么，怎么调研，排定优先级
- FOR 研发：明确产品具体诉求，确认细节，提升开发效率

NOTES

1/ 明确一定要做，不做会死的功能/项目点，其他的为了正常进度可以适当妥协

4、体验层—实例1，起始页热词

（2） 提升全局思考能力，避免切片思维（包含关系、上下游关系）

- 起源对于起始页和首页二者若均为缩短用户路径，其核心区别是什么
- 之前存在两个大的全局思考不足。①包含关系，起始页是搜索的一部分，其定义将不应脱离搜索本身，缩短用户路径，结合场景本身，不期望用户发起搜索；②上下游关系，搜索是应用宝的一部分，且必经之路为各类上游场景，起始页虽和其他推荐页面可能部分逻辑一致，但需整体考虑上文的行为均会对下游的推荐都是会产生影响的

典型的自以为想清楚了，可是每个阶段都其实是当时清楚。实际不够清楚的

4、体验层—实例2，内容/分行业搜索

2017年

通用query
卡队列

2019年8月

拆解行业
内容形态

2019年12月

拆解行业
卡队列+内容形态

- **大方向**
  - why 导流
- **正确的方法**
  - 卡队列
  - 内容形态
  - 卡队列+内容形态
- **研发的支持**
  - 研发技术能力
  - ABTEST能力
  - 数据拆解能力
- **战略决心**

3、体验层——其他胡思乱想

- **产品和技术的win win**
  - AB/数据/BUG/实现能力…
  - 产品要懂技术，技术要求了解产品，双方之间要加强协作

- **提升全局思考能力，避免切片思维**
  - ①包含关系，起始页是搜索的一部分
  - ②上下游关系，搜索是应用宝的一部分，且必经之路为各类上游场景

- **商业化和平台之间的关系**
  - 用户来 （<=>）用户留存（<=>）用户增值
  - 当下的ctr cvr arpu vs ltv 留存（尤其是游戏，新游）

- **什么是站在用户的角度思考（参考）**
  - 做产品=汇报=发会议邀请=报case->所有要和人打交道的事情
  - 核心本质是站在“对方”的角度思考，用户=老板=受邀人=解决bug的人

- 感恩，鞠躬~ 谢谢大家
- 欢迎大家加入搜索小分队，嘻嘻 (*^▽^*)