

内容搜索

工虫

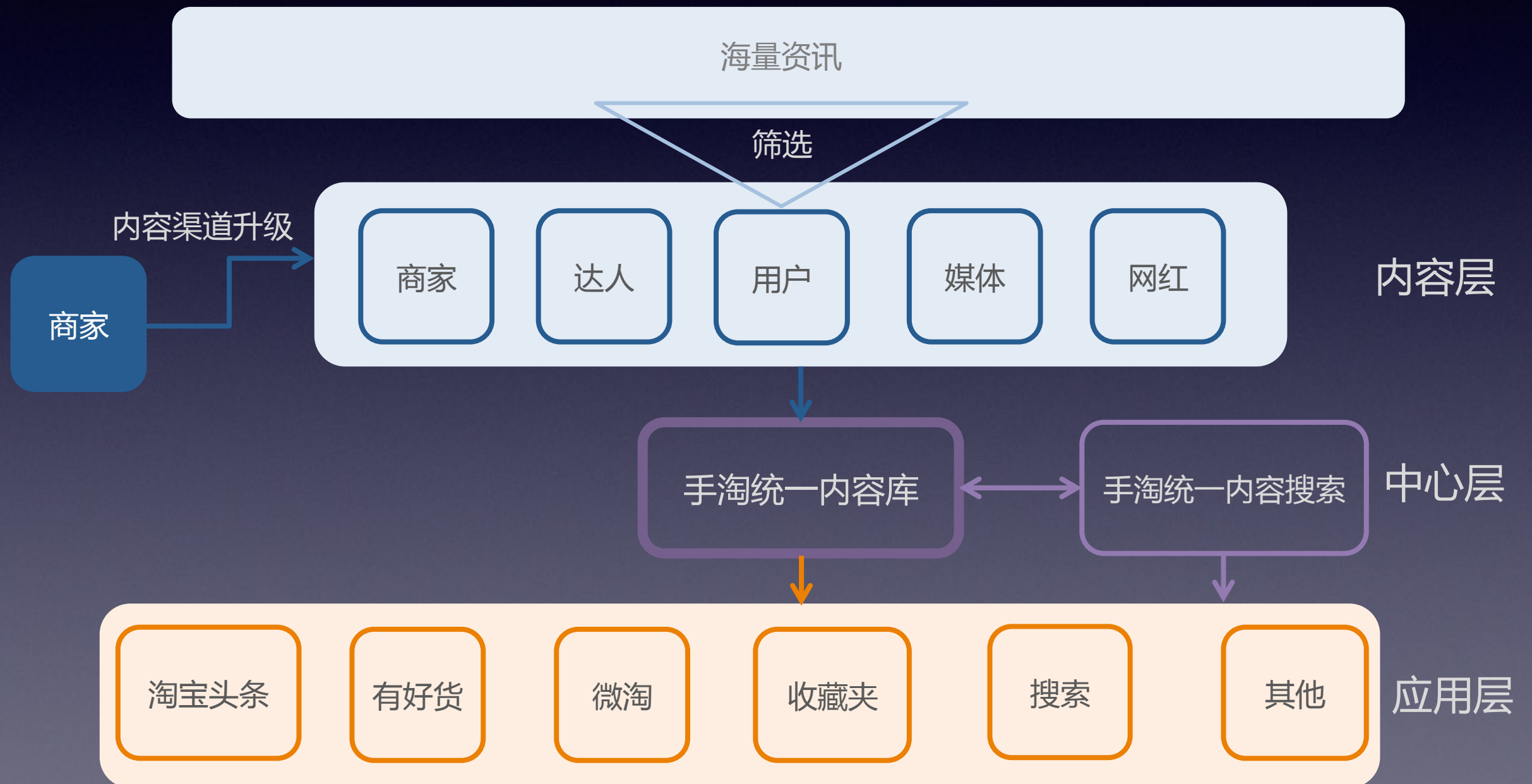
Outline

- 动机
- 目标
- 人员配置
- 项目规划
- 数据沉淀

动机

- 背景：
 - 9月手淘推出“内容开放计划”，优质内容创作者和机构可以共享市场佣金，预计规模20亿
 - 手淘趋势之一：内容生态社区
- 搜索技术解决的问题：
 - 内容信息越来越多，无法高效展现给用户所需内容
 - 目前缺乏完善内容生态，头条依然按照发帖时间展示
 - 没有定向推荐

目标



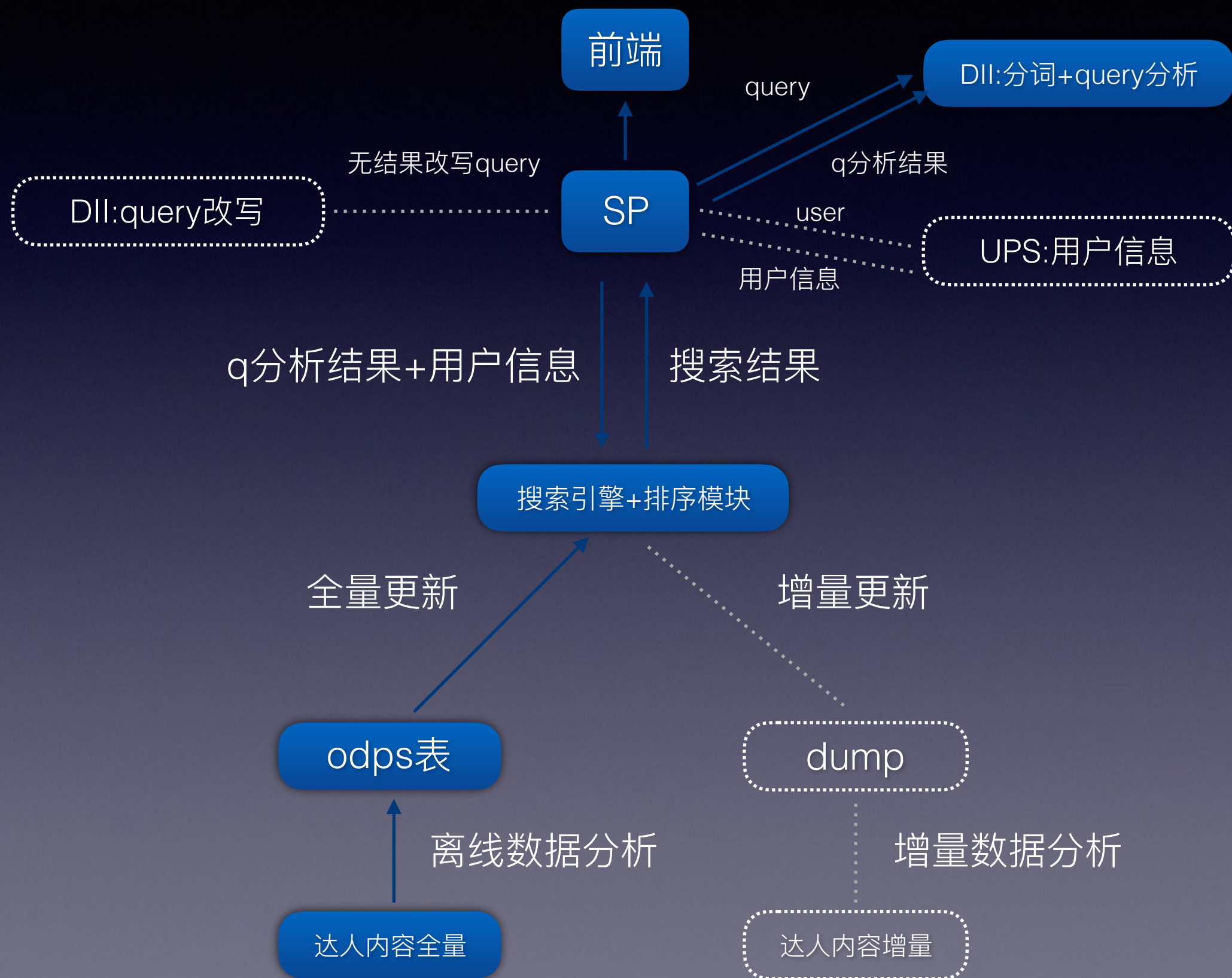
项目规划:内容搜索1.0

- 人员配置:
 - PD: 伽楠
 - PM: 工虫
 - 交互: 宁渊
 - 视觉: 幻心
 - 评测: 烟霞团队
 - 前端: 毋愁、东淮团队
 - 引擎: 梓羽
 - SP: 玄痴
 - DUMP: 德言
 - 算法-排序: 仁重、萱然
 - 算法-相关性 (文档/query): 工虫、旷谷、乐慷、玉昆

项目规划

- 第一阶段： prototype及调优 (done)
- 第二阶段： 内容搜索1.0上线 (2015.1.20)
- 第三阶段： 内容搜索2.0

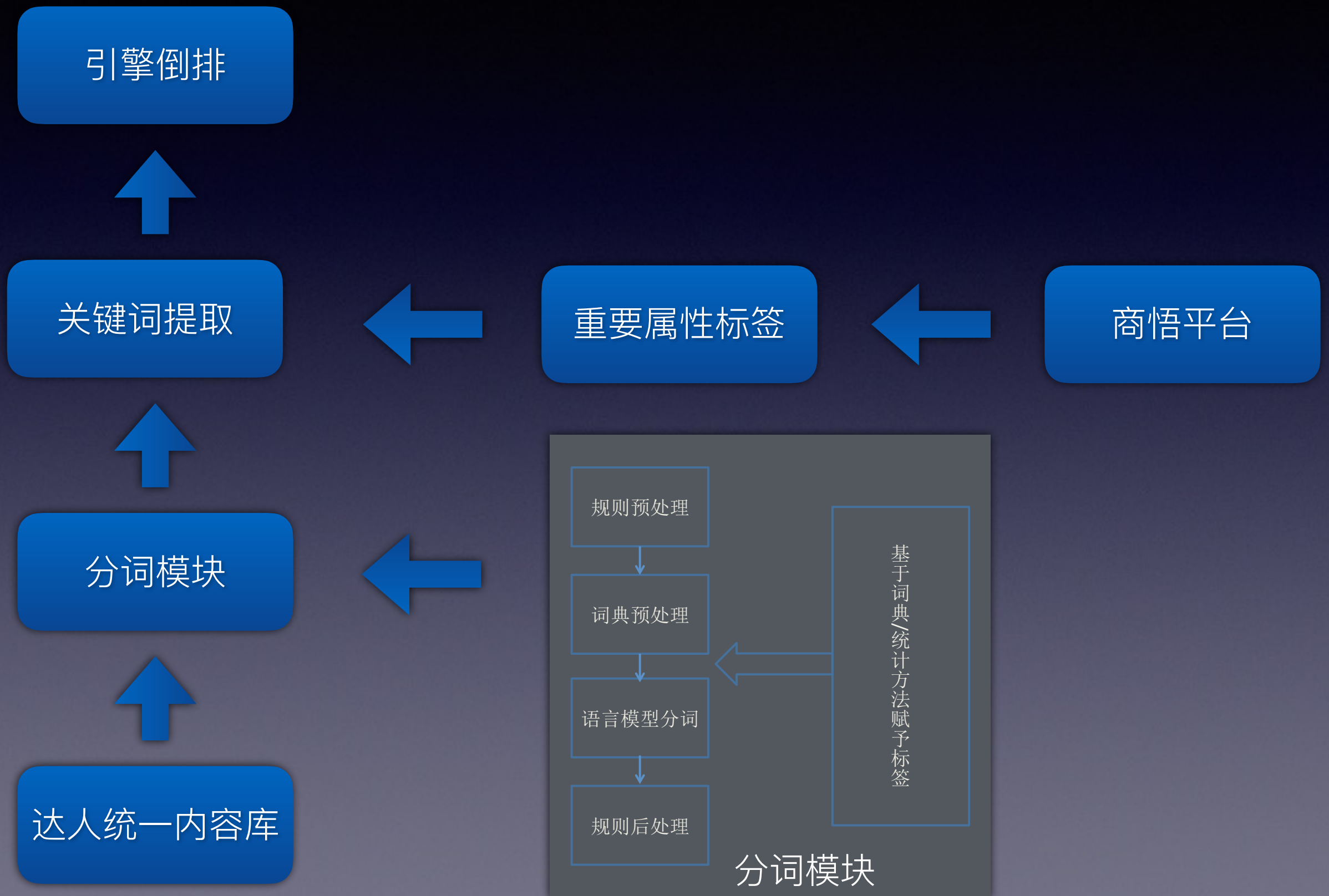
项目规划:Prototype



项目规划:Prototype

- 前端: DEMO
- 工程:
 - TISPLUS (HA3)
 - DII
- 算法
 - 文档分析
 - 分词
 - 基于标签的关键词提取
 - 相关类目提取
 - query分析: 分词
 - 排序: 关键词召回

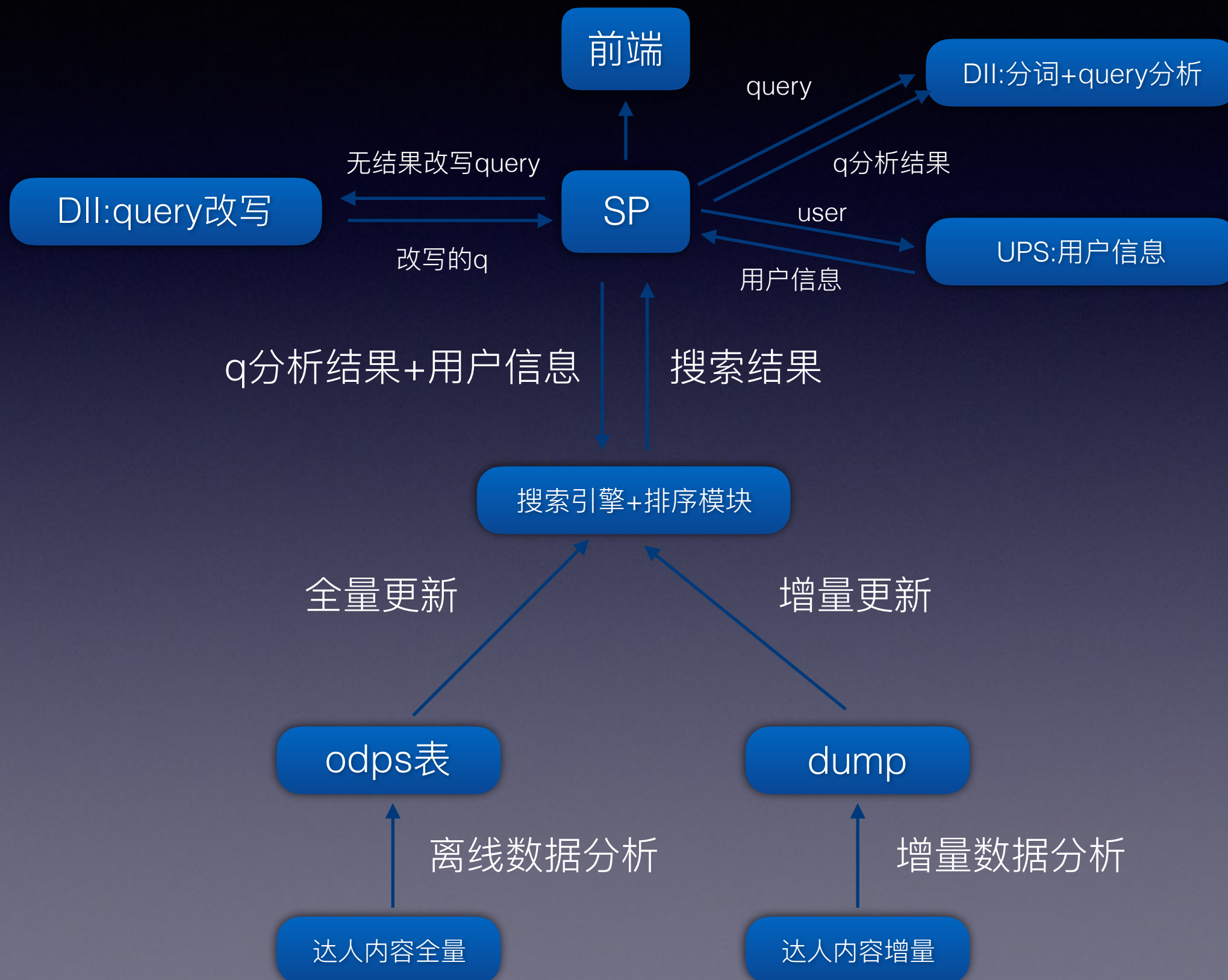
项目规划:Prototype-文档分析



项目规划:Prototype-问题

- TISPLUS:
 - 日常环境不够稳定
 - 缺少dump支持增量
- 相关性问题:
 - 缺少term weight
 - 缺少文档静态分
- 排序问题:
 - 没有开发排序插件
- 前端问题:
 - 还不是最终交互稿的样子
 - 前端数据展示无法调用达人库统一模板

项目规划:内容搜索1.0



项目规划:内容搜索1.0

- 前端:
 - UED确定交互稿
 - 埋点与日志
 - 无结果问题

项目规划:内容搜索1.0

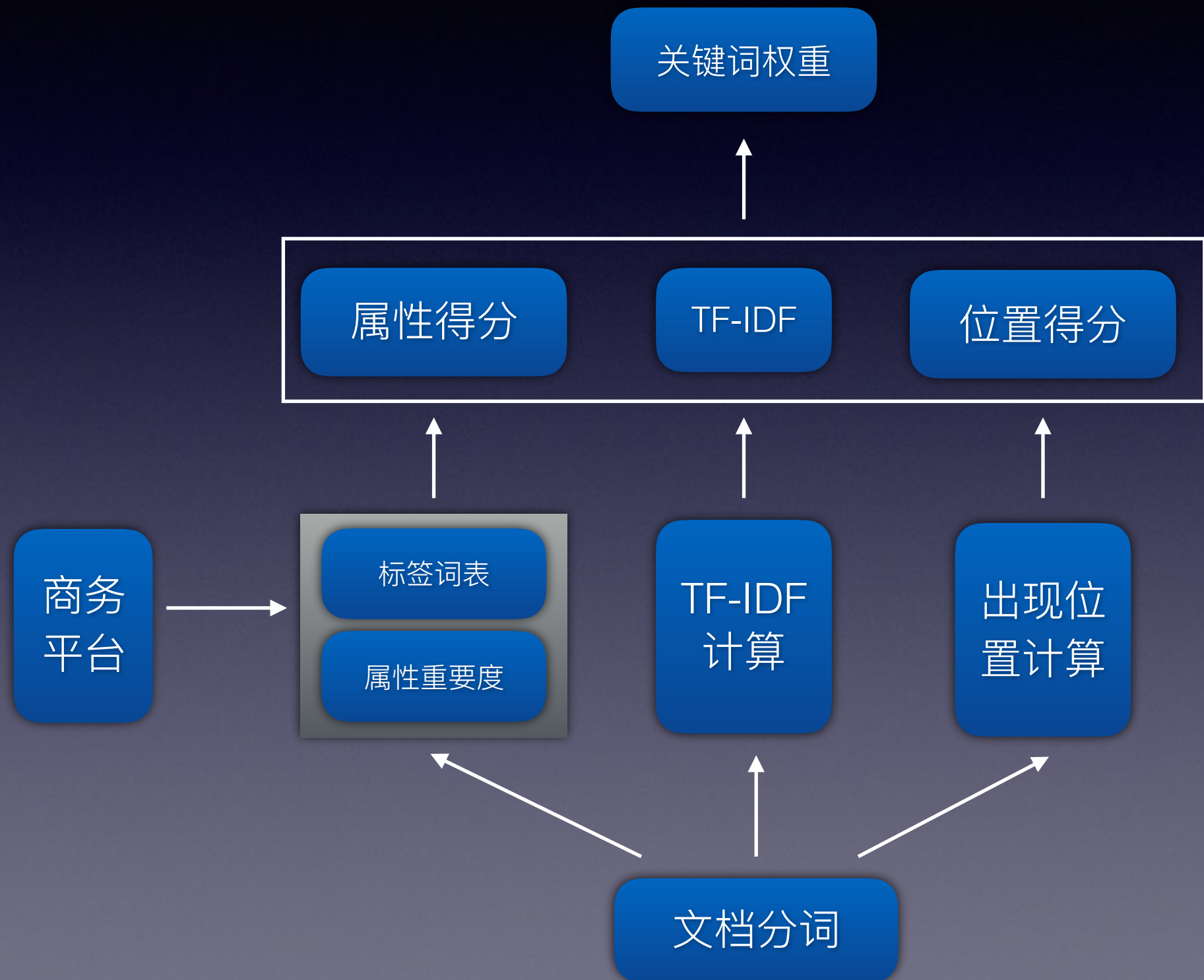
- 工程:
 - 引擎: TISPLUS
 - 需要协调机器与后续维护人
 - 排序插件
 - 相关性排序插件
 - 静态分排序（战马框架）
 - DUMP: 解决增量问题（主要是实时下线数据）
 - DII: 增加无结果改写query功能
 - SP: 增加无结果重搜逻辑

项目规划:内容搜索1.0

- 算法:
 - 文档分析:
 - 关键词权重
 - 类目预测
 - 主题标签
 - 静态分计算
 - query分析:
 - query关键词权重
 - 类目预测
 - 主题标签
 - 无结果改写query

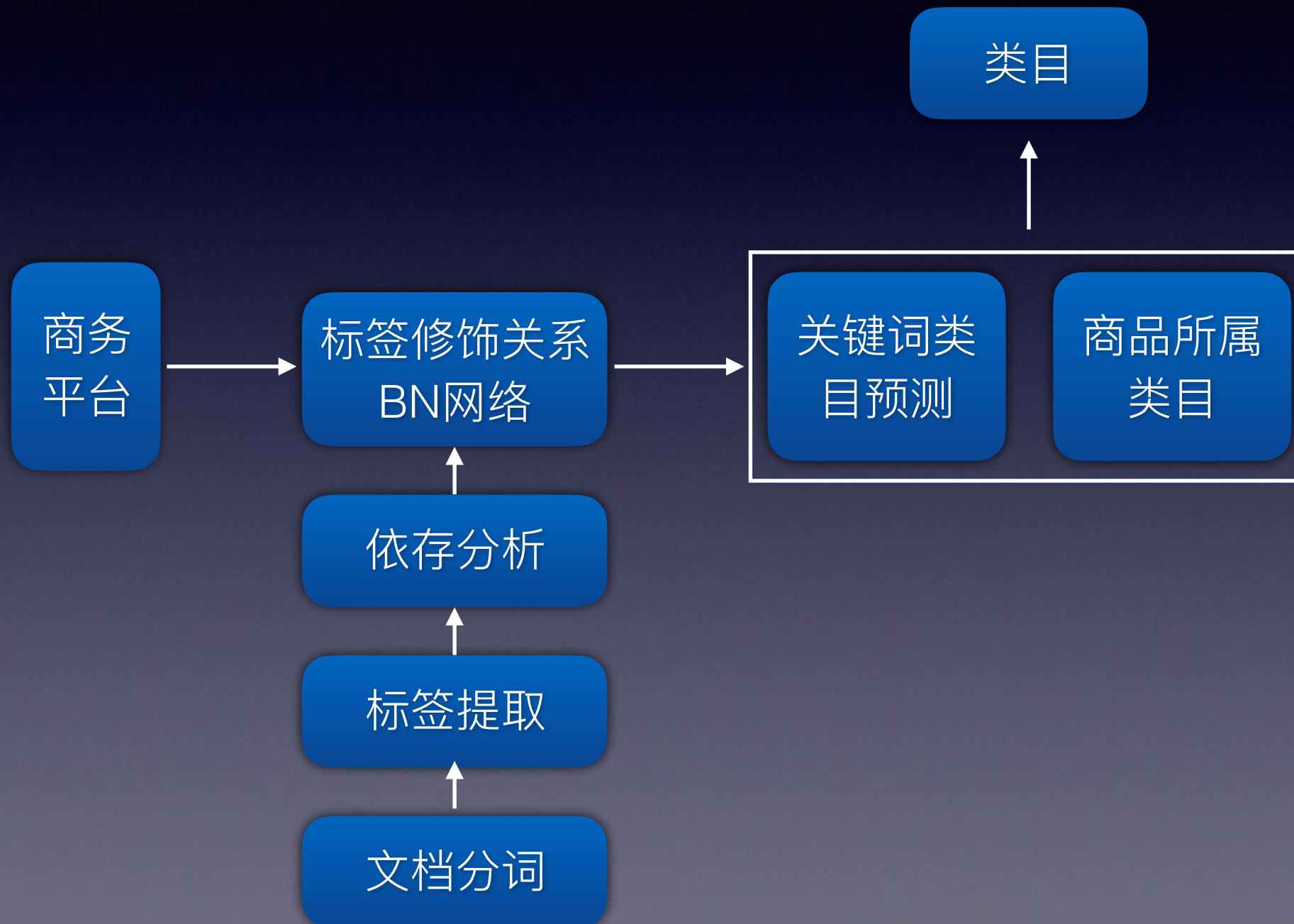
项目规划:内容搜索1.0

- 算法：文档分析-关键词权重



项目规划:内容搜索1.0

- 算法：文档分析-类目预测



项目规划:内容搜索1.0

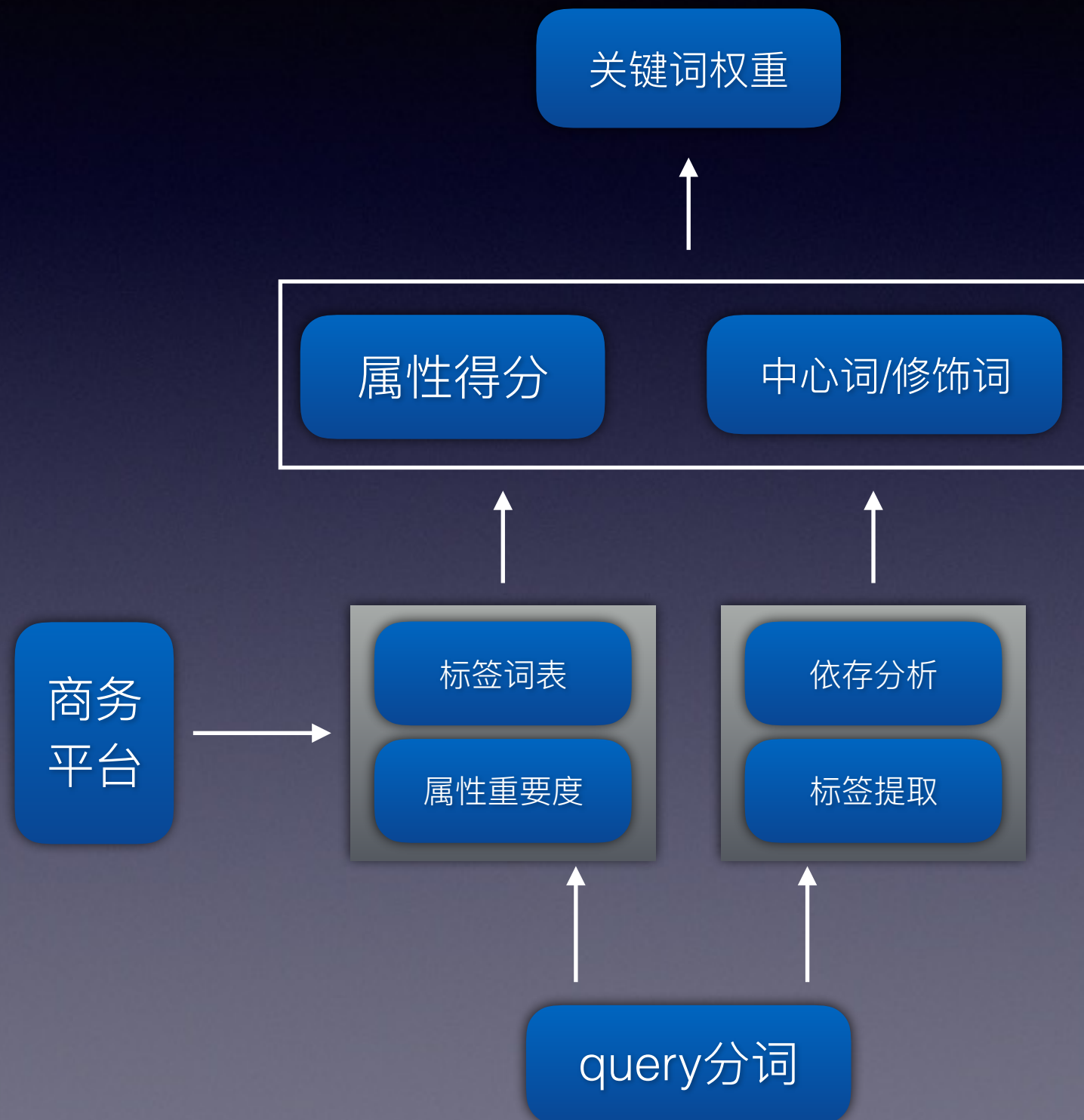
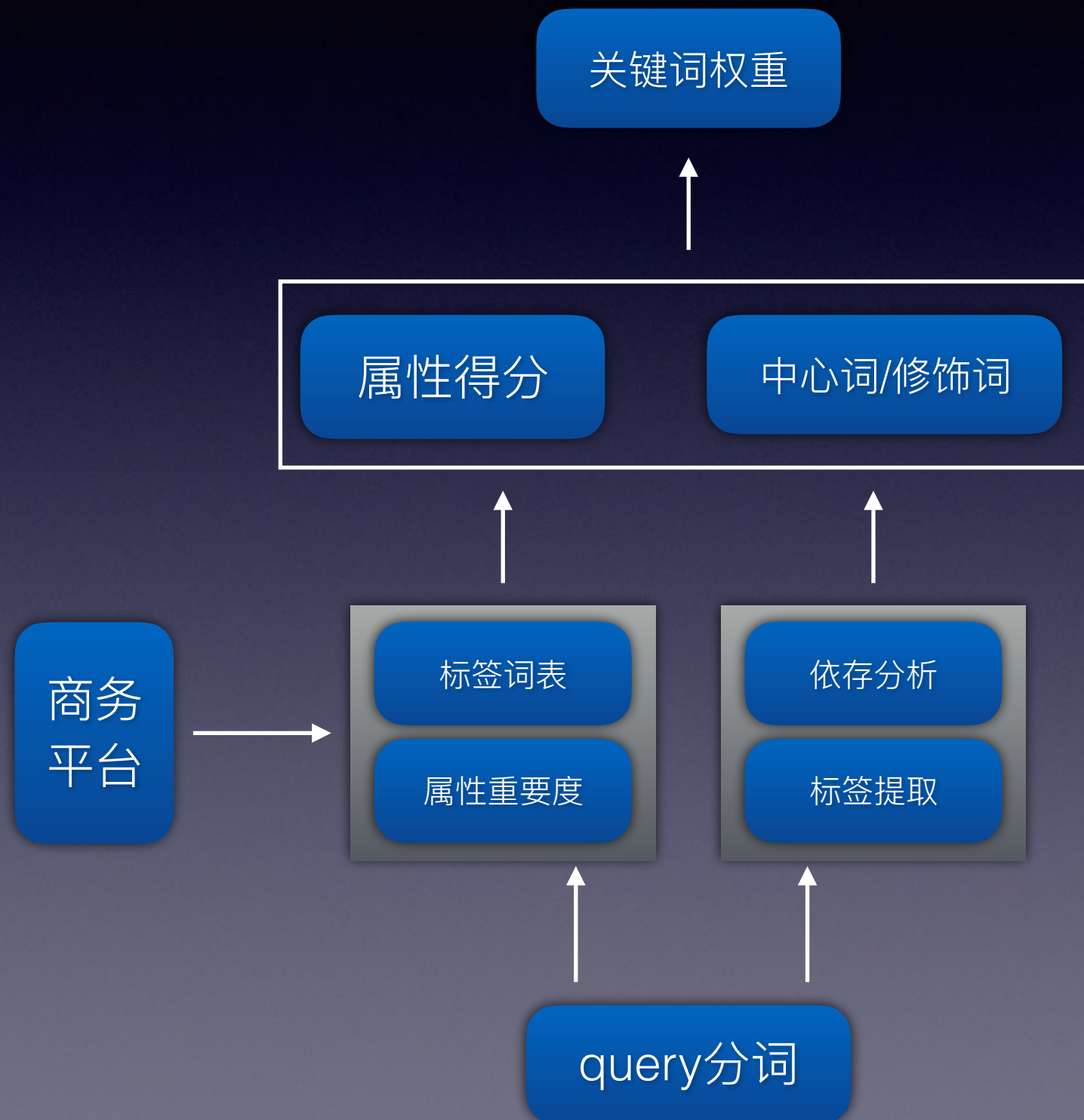
- 算法：主题标签
 - 基于pattern/关键词的query意图识别
 - “怎么样” -> 品论类
 - “怎么做” -> 经验分享类
 - 基于监督学习的文档主题分类
 - 人工标注训练数据
 - 特征选择：以标题中的词为主要特征

项目规划:内容搜索1.0

- 算法：排序
 - 相关性分档
 - 文本相关性：query与标题、摘要、详情、商品的相关性
 - 类目相关性
 - 热度分
 - PV/生命周期
 - 质量分
 - 内容本身质量：作者等级、信息量（选购类资讯）
 - 反馈：ctr、收藏、点赞等
- query质量分

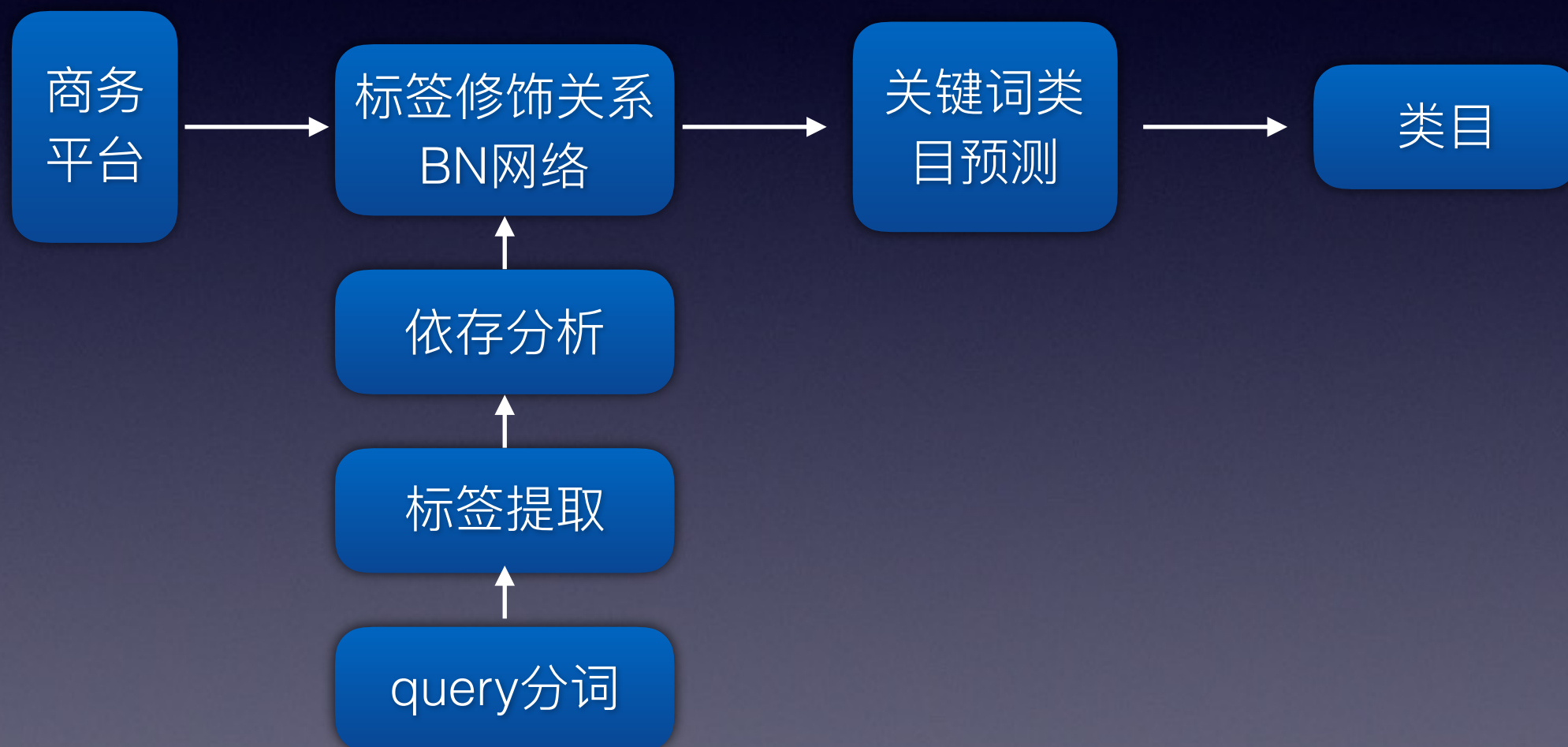
项目规划:内容搜索1.0

- 算法: query分析-关键词权重



项目规划:内容搜索1.0

- 算法: query分析-类目预测



项目规划:内容搜索1.0

- 算法: query分析-主题标签
 - 离线文档训练模型
 - 在线DII模块计算

项目规划:内容搜索1.0

- 算法: query分析-无结果改写query
 - 策略: 根据静态分和类目, 推荐热门内容贴
 - 算法: 根据行业 (类目), 推荐热门内容贴

项目规划:内容搜索1.0

- 上线场景
 - 文字锦囊、主题搜导流至于landing page
 - SRP页新增tab（需要跟随SRP改造项目）
 - 系统+landing page上线点： 1.20

项目规划:内容搜索1.0

- 风险点:
 - 进度问题: 涉及team比较多, 需要良好得沟通与信息同步
 - 技术问题:
 - TISPLUS上的插件开发目前还没有人有经验, 后续可能需要引擎插件团队支持
 - 语义标签是一个新尝试, 需要摸索
 - 知识图谱团队大部分成员还未接触过整个搜索系统, 作为其第一次开发相关系统可能存在一些困难, 会导致进度延迟

项目规划:内容搜索2.0

- 前端:
 - 搜索框需要: 底纹/下拉
 - 搜索结果详情页改造: 可能增加商品推荐、收藏等一系列功能
- DUMP与引擎: 需要支持实时增量

项目规划:内容搜索2.0

- 算法:
 - 底纹/下拉推荐query计算
 - 文档/query 语义标签2.0 (场景与意图识别)
 - 相关性模型
 - 动态排序计算
 - query-内容的ctr预估
 - bandit热度测试
 - 新内容的冷启动
 - 热点的透出
 - 个性化计算
 - 实时增量流程计算

项目规划:内容搜索2.0

- 应用场景：内容库内容展现的地方，譬如淘宝头条、微淘、我的搜索、收藏，等等
- 上线时间点：需要和透出场景业务方紧密协作

内容搜索数据沉淀

- 文档分析数据
 - 标签
 - 主题
 - 静态分
- query分析数据
 - 长尾、知识型query积累
 - query偏好分析（什么样的query适合透出内容）
- 用户个性化数据
 - 用户行为积累
 - 用户偏好分析

Thanks