

美团点评旅游搜索的策略演进

郑刚

- 2015年至今 美团点评酒旅事业群

- 负责酒旅搜索排序推荐
- 负责酒旅数据仓库和数据产品建设

- 2014年之前 美团网技术部数据组

- 参与数据平台搭建
- 负责全平台数据仓库和数据产品建设

- 2011年 百度电子商务事业部

- 有啊商城的开发

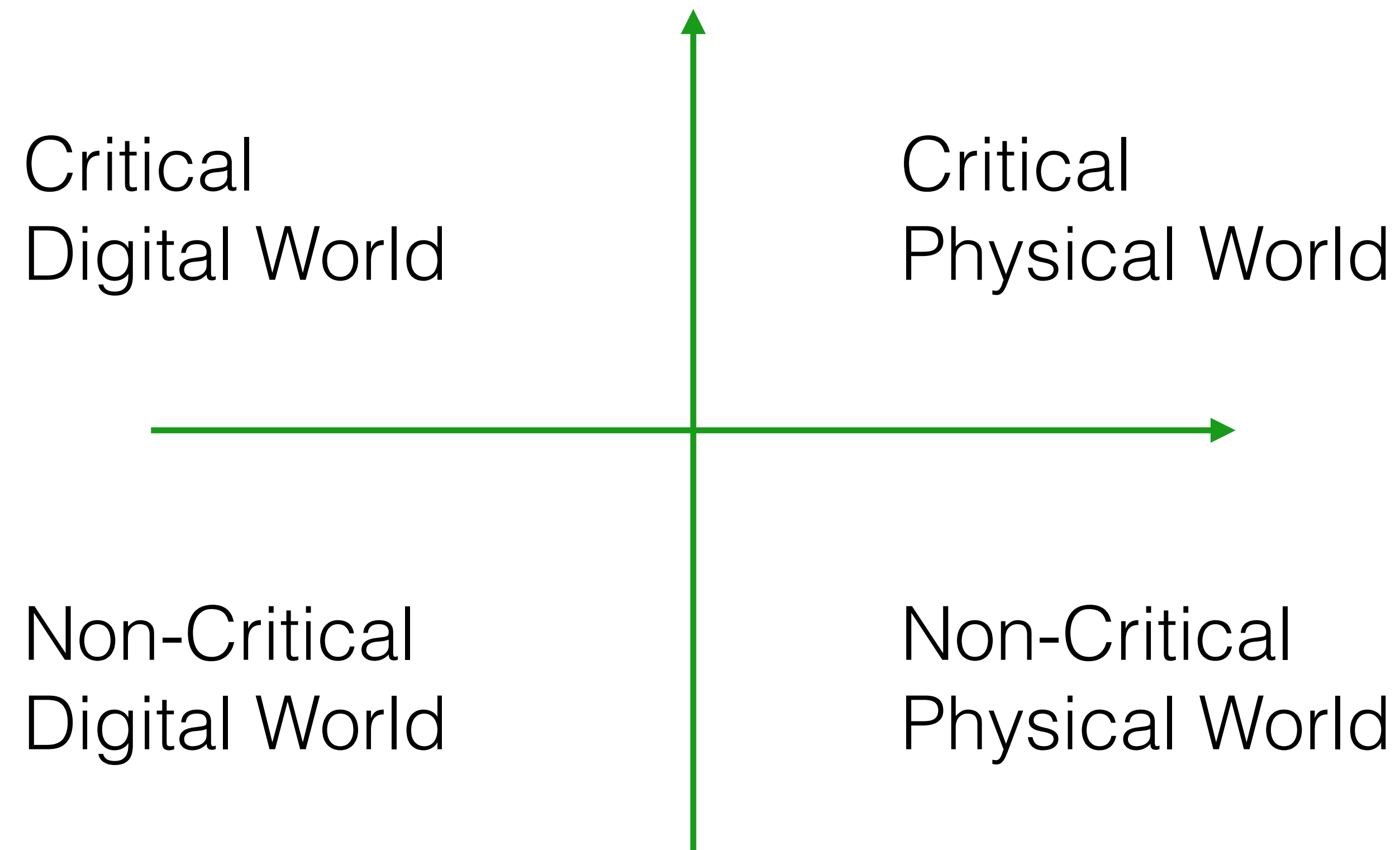
- 2010年毕业于中科院计算所

Outline



- 美团点评酒旅业务简介
- 旅游搜索召回策略演进
- 旅游搜索排序策略优化
- 机器学习在搜索排序中的应用

人工智能应用



新美大酒旅

国内发展最快的一站式综合住宿服务平台

国内最大的在线门票交易平台



酒旅搜索推荐

2015Q2 组建搜索团队

2015Q3 美团旅游搜索

2016Q1 平台搜索Merge Server

2016Q2 酒旅综合搜索

2017Q1 点评旅游搜索

2017Q2 美团点评门票筛选

酒店住宿

境内度假

境外度假

大交通

搜索/推荐

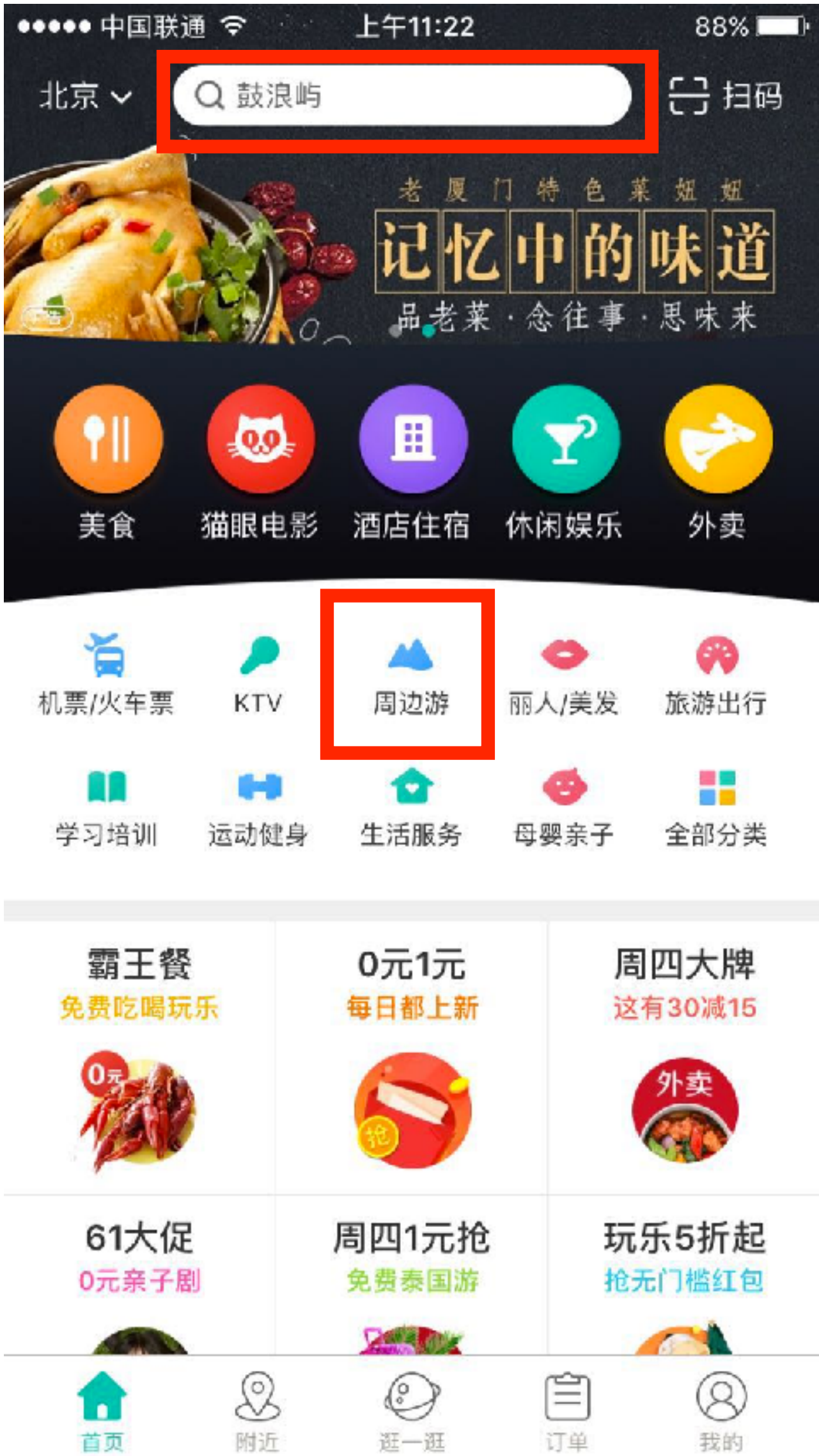
数据挖掘

数据产品

酒旅数据仓库

集团数据平台

旅游搜索产品形态



用户对酒旅品牌认知不强，
搜索是酒旅交易转化的重要入口

旅游搜索面临的问题



本异地差异大



搜索意图多样



季节性明显



底层脏数据较多

本异地差异大



东方明珠广播电视塔(5A)

好, 4.4分 | 270000+人消费

上海市, 距我>300km

端午节大促 随买随用

门票 观光票90元起, 观光票+游船票128元起, 多景点联票125元...

跟团 迪士尼小镇、东方明珠广播电视塔、石库门纯玩3日跟团...

套餐 北士大酒店1晚+东方明珠广播电视塔/东方明珠豪华浦江...

东方明珠相关景点



东方明珠豪华浦江游船东方明...

好, 4.3分 | 20000+人消费

端午节大促 随买随用

¥65起

上海市, 距我>300km



东方明珠公园

很好, 4.7分 | 0人消费

暂无售卖

上海市, 距我>300km

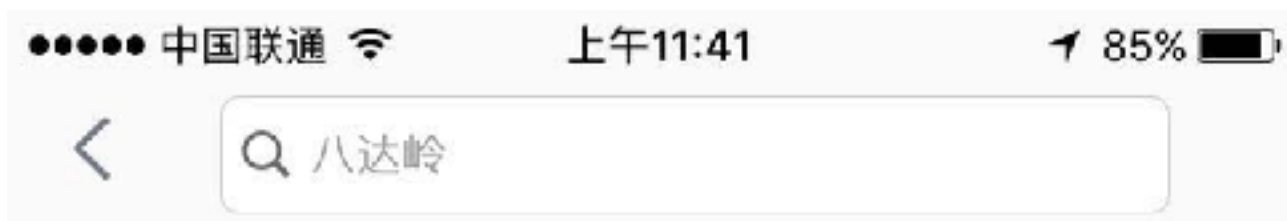


黄海明珠塔

暂无评分 | 7人消费

¥135起

丹东市, 距我>300km



八达岭长城(5A)

很好, 4.8分 | 60000+人消费

八达岭, 距我55.7km

端午节大促 随买随用

门票 成人票65元起, 学生票160元起, 多景点联票160元起

跟团 故宫博物院、八达岭长城、颐和园等纯玩4日跟团游*感...

套餐 北京国际饭店/北京西国贸大酒店/北京世纪金源大饭店3...

八达岭相关景点



八达岭野生动物世界(3A)

好, 4.3分 | 40000+人消费

返券10元 明日可订

¥40起

八达岭, 距我56.7km



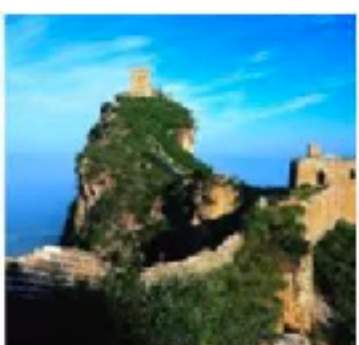
八达岭水关长城(4A)

很好, 4.6分 | 1000+人消费

端午节大促 随买随用

¥27起

八达岭, 距我53.2km



八达岭古长城(3A)

很好, 4.6分 | 600+人消费

返券5元 随买随用

¥20起

八达岭, 距我56.5km

超过30%订单来自异地请求

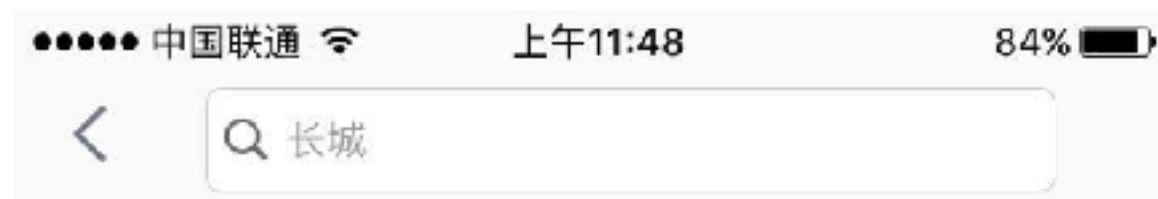
常驻城市!=浏览城市

搜索意图多样

美团点评
技术团队

IT大咖说
知识分享平台

景点

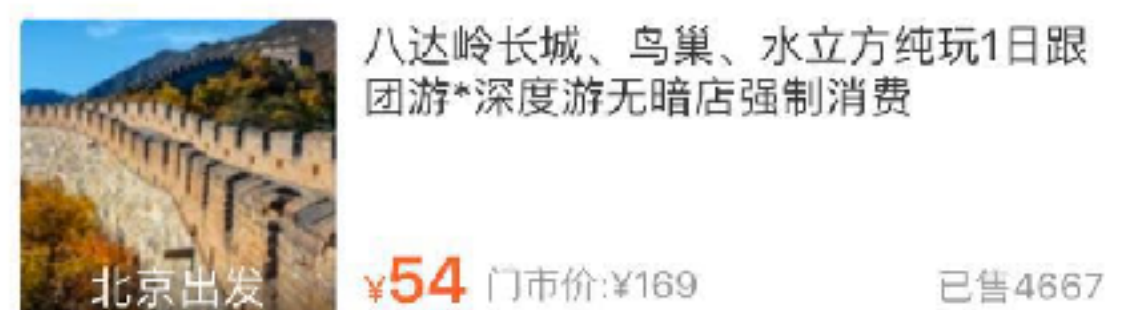


长城相关景点

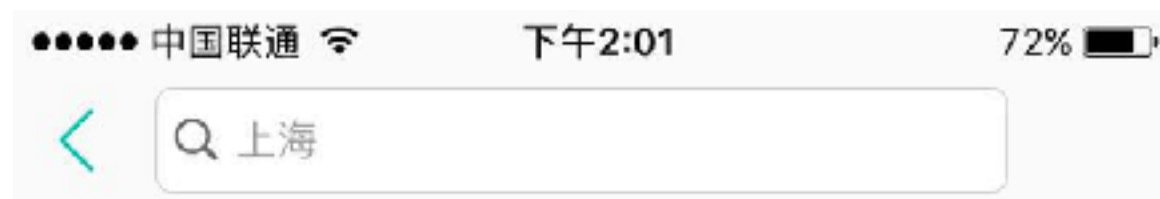


查看全部63个景点

长城相关度假产品



行政区



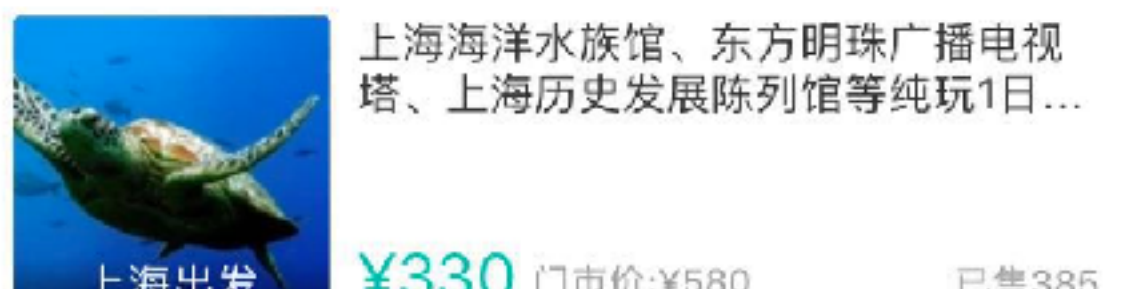
上海热门景点



北京-上海度假产品



上海当地度假产品



品类



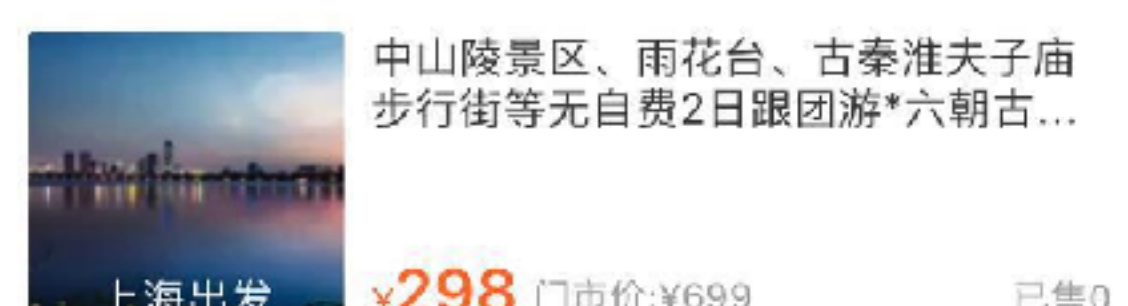
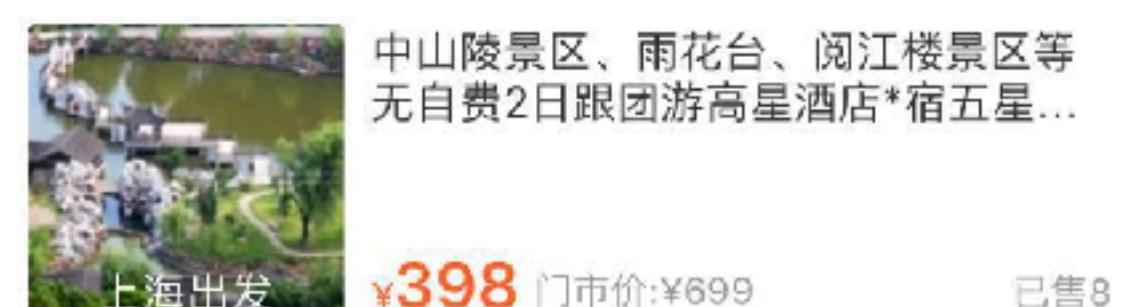
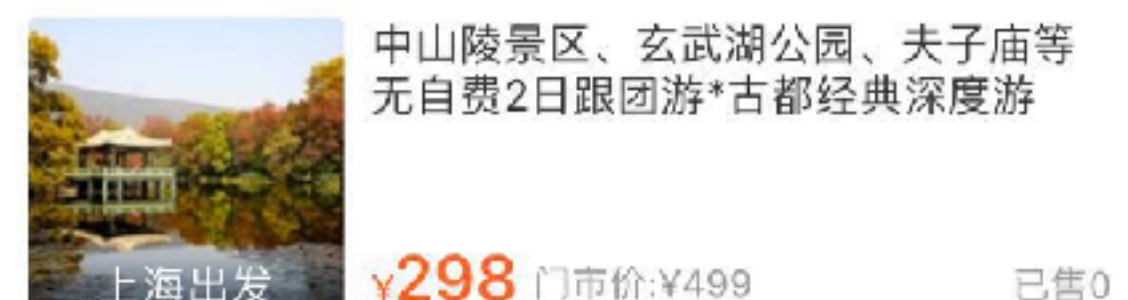
温泉相关景点



线路游

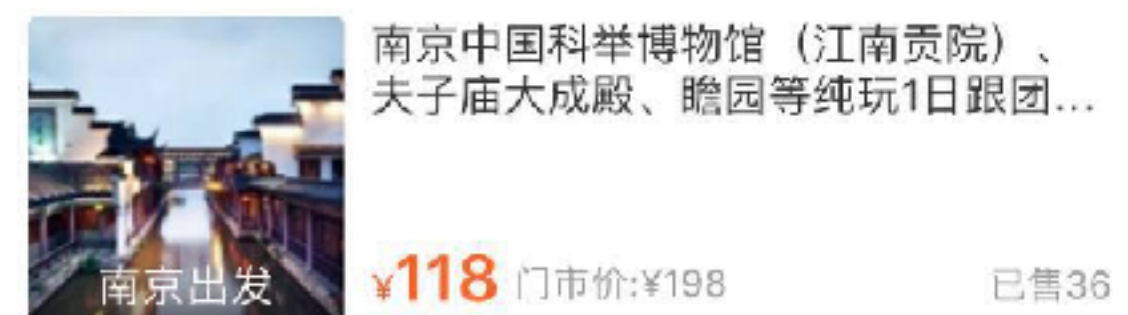


上海-南京度假产品



查看全部5个度假产品

南京当地度假产品



季节性明显



冬季温泉订单占比超过20%,
而夏季不到7%

底层脏数据多



【东方巴黎/凤城世家】**故宫博物院**
北江夜游1人

¥39 门店价 ¥89

已售 747 ★★★★★ 72人评价



【番禺区】**故宫博物院**
自助烧烤套餐1份, 提供免费WiFi

¥58 门店价 ¥168

已售 59 暂无评价



【番禺区】聚友烧烤场



【长兴县】陈武帝**故宫**成人票

●●●● 中国联通 下午1:55 71%

您有5元旅游抵用券即将过期>>

黄鹤楼农机相关商户

黄鹤楼农机

暂无评分 | 0人消费

暂无售卖

成都市, 距我>300km

你可能会喜欢

故宫博物院(5A)

很好, 4.8分 | 180000+人消费

返券5元 随买随用

¥5起

王府井/东单, 距我12.7km

颐和园(5A)

很好, 4.7分 | 70000+人消费

立减10元 随买随用

¥9.9起

颐和园, 距我18.4km

八达岭长城(5A)

很好, 4.8分 | 60000+人消费

端午节大促 随买随用

¥65起

八达岭, 距我55.7km

恭王府(5A)

•访购率为目标

- 只看点击率没有反映出交易属性
- 看最终收入
 - 消费受购买限制、退款条件等影响
 - 收入跟BD谈单毛利相关

• 点击率

- 无合作POI
- 有点行为比

•无结果率

·用户满意度

- ## ·人工评测

$$\text{每搜索用户收入} = \underbrace{\frac{\text{点击用户数}}{\text{搜索用户数}} \times \frac{\text{支付用户数}}{\text{点击用户数}}}_{\text{有点行为比}} \times \frac{\text{消费用户数}}{\text{支付用户数}} \times \text{每用户消费收入}$$

指标内问题	指标外问题
查询意图理解 召回检索策略 个性化排序 页面改版 源数据清洗	广告运营 直签门票加权

策略迭代方法

以Case驱动为主



•质量评估

- 定义满意度标准和评估体系
- 定期评估搜索满意度
- 确定评估样本
- 了解Query分布、意图识别准召率、召回及排序情况

•问题分析

- 对问题梳理分类，明确优化方向

•项目开发

•实验迭代

旅游搜索召回策略演进

全国召回

美团点评
技术团队

IT大咖说
知识分享平台

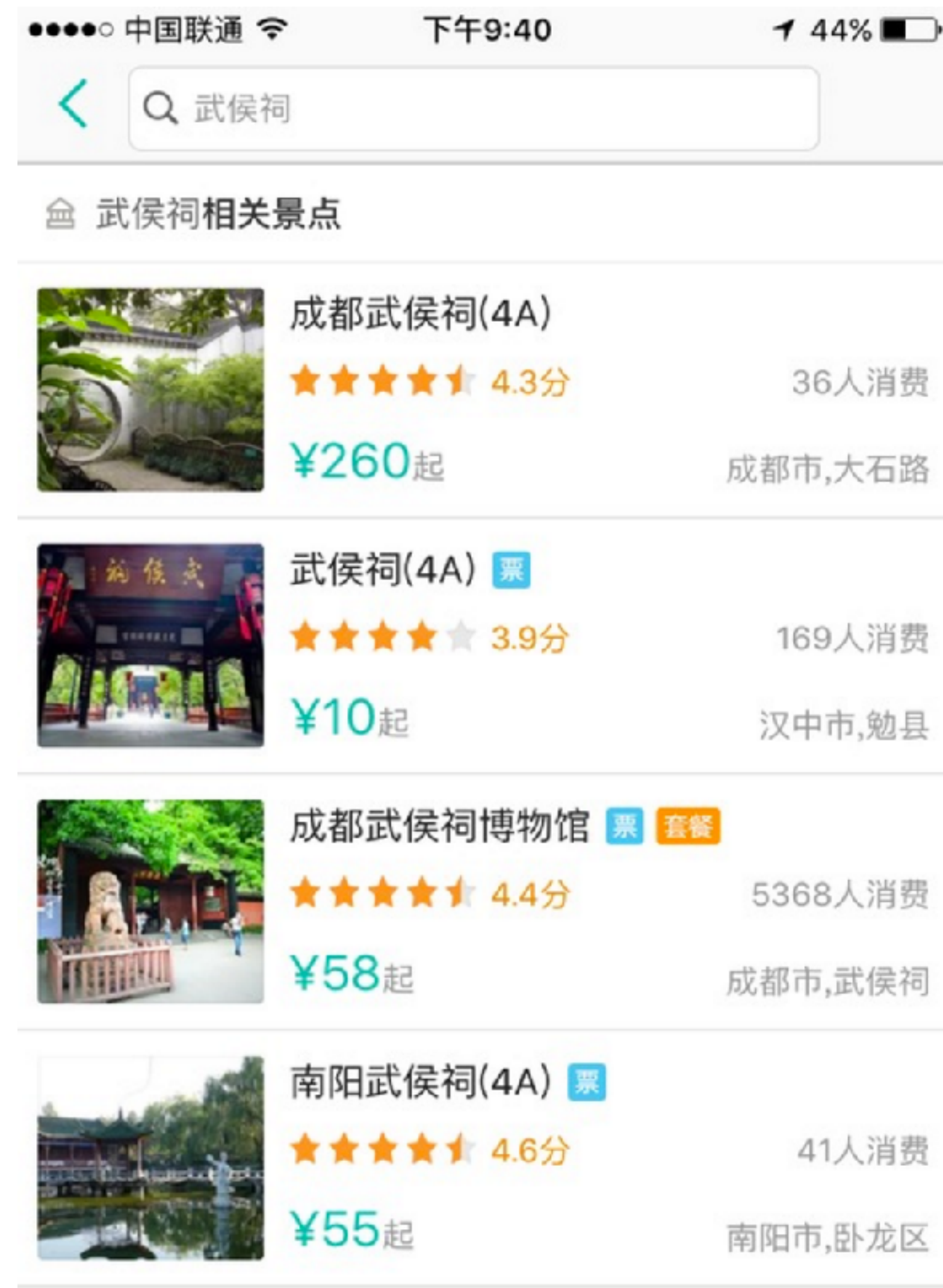
•搜索质量评估

- 无结果召回Query影响面为35%
- 除无供给问题外最重要原因是不支持异地搜索

•旅游超过30%订单来自异地请求

•放开上单城市限制

无结果率大幅下降
有本异地排序问题



模块化展示

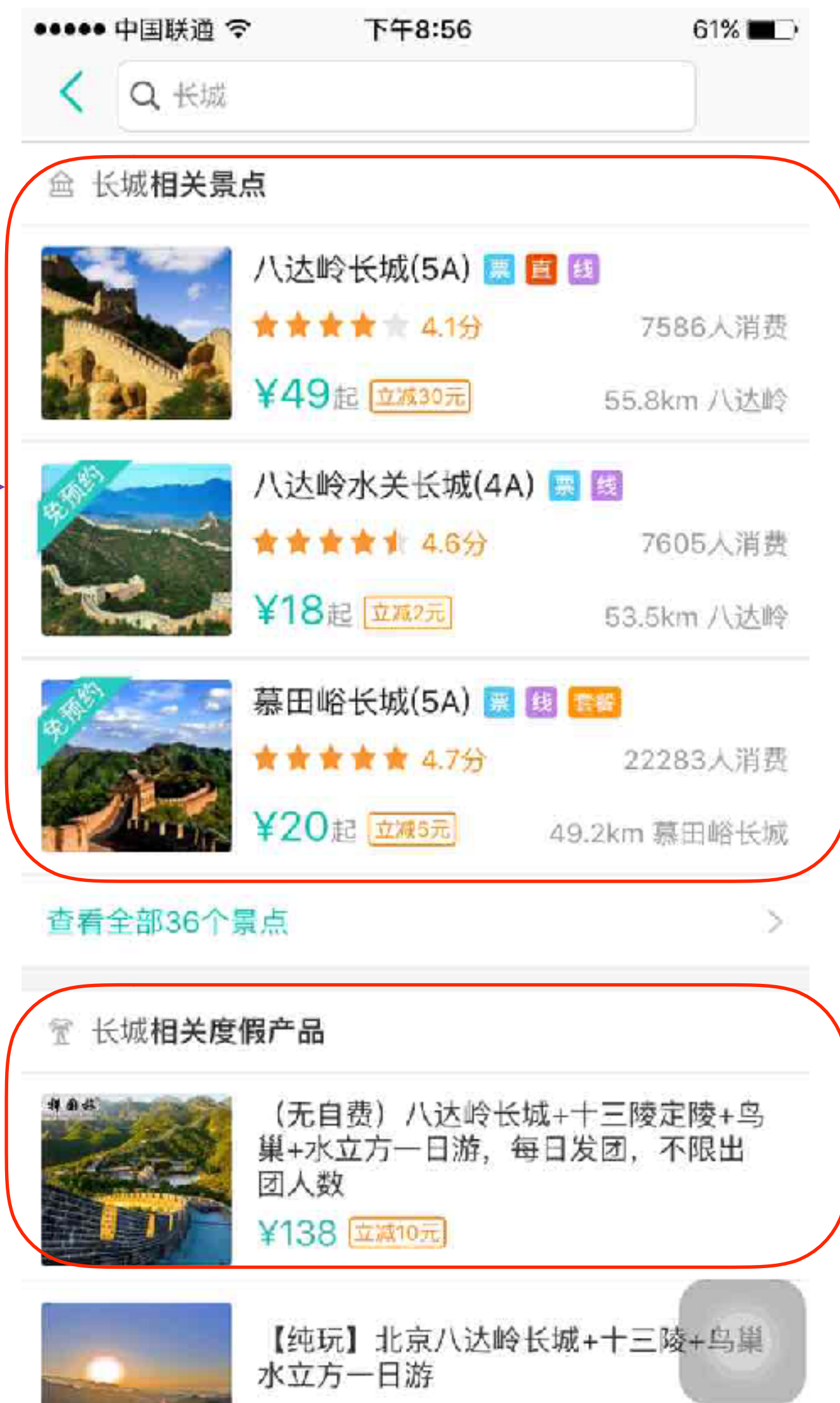
美团点评
技术团队

IT大咖说
知识分享平台

同时检索了POI名称和Deal标题字段，导致误召回



只检索POI相关字段 →



模块化展示

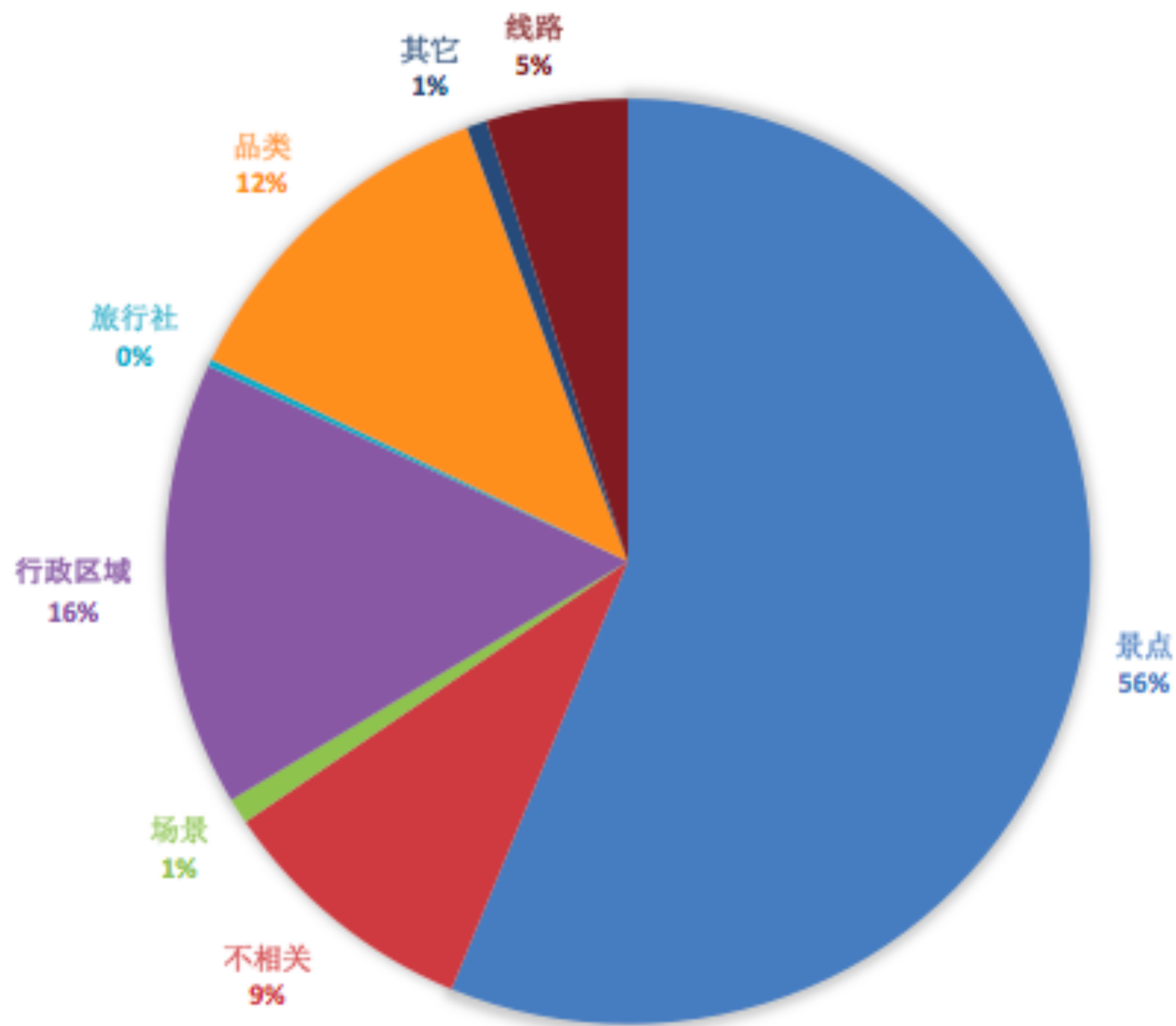
美团点评
技术团队

IT大咖说
知识分享平台

行政区

品类

QUERY分类



上海热门景点

更多>



北京-上海度假产品



周庄、乌镇、西湖等6日跟团游*梦幻迪士尼游梦幻水乡

北京出发

¥3280 门市价:¥45001

已售2

上海当地度假产品



上海海洋水族馆、东方明珠广播电视塔、上海历史发展陈列馆等纯玩1日...

¥330 门市价:¥580

已售385



温泉相关景点



顺景温泉

★★★★★ 4.4分 | 34298人消费

¥165起

王牌特权 立减3元 票 套餐

太阳宫 | 4.4km



华清温泉宾馆

★★★★★ 4.5分 | 7858人消费

¥35起

返券10元 票

天通苑 | 8.3km



京林温泉

★★★★★ 4.4分 | 1828人消费

¥58起

返券15元 票

首都机场 | 10.5km



裕龙温泉

★★★★★ 4.5分 | 9867人消费

¥149起

返券20元 票

航天桥 | 18.9km

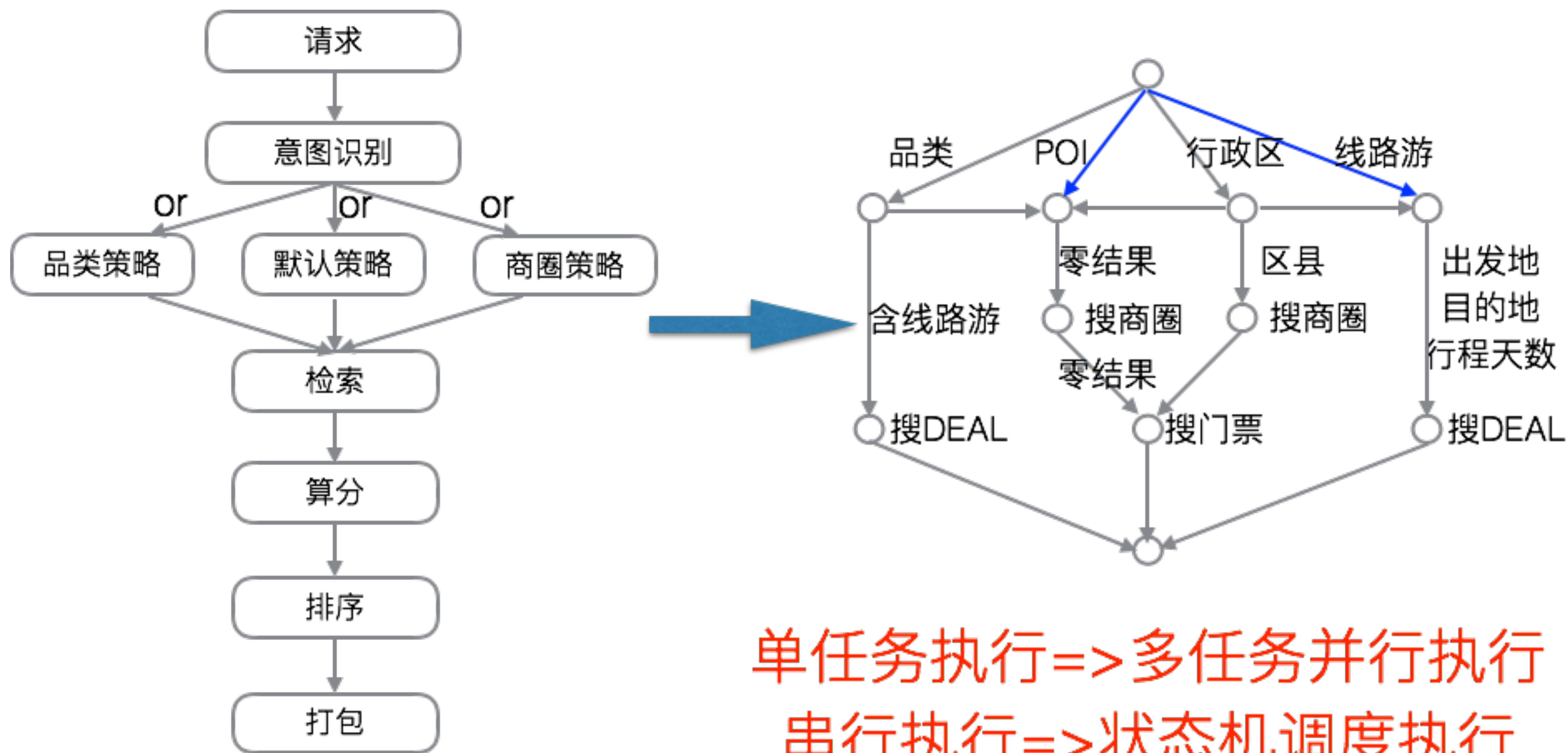


中家鑫园温泉酒店

★★★★★ 4.7分 | 6315人消费

¥105起

模块化展示



无结果优化

• 无结果率20%，其中30%是query理解问题

• 15%是Query包含冗余词

• 搜“东莞的隐贤山庄”

• 7%是Query含有错别字或同义词

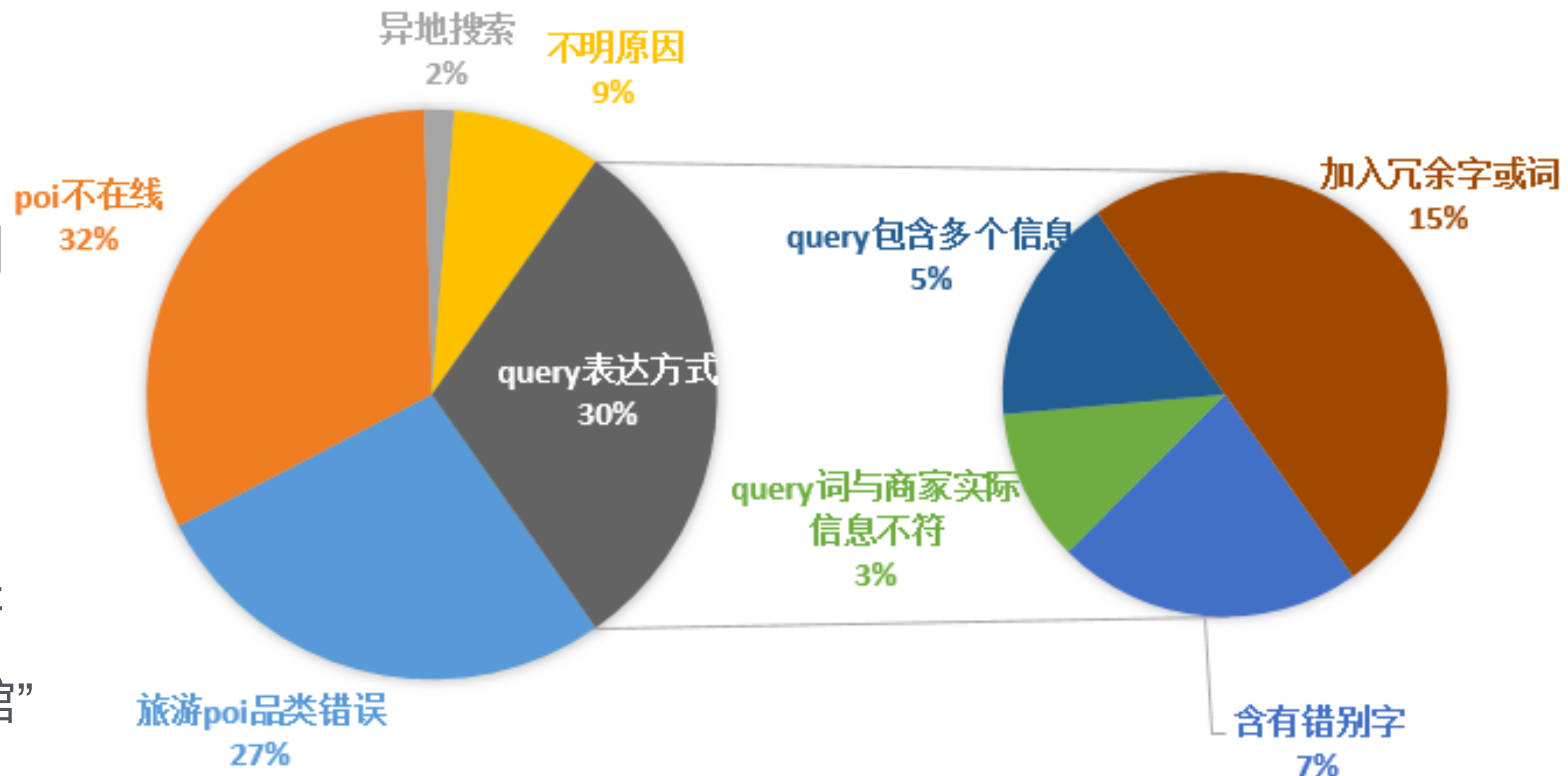
• “雁西湖”=>“雁栖湖”

• 5%是Query包含多个信息

• 搜“北京动物园海洋馆门票”

• 3%是Query与商家实际信息不符

• “798艺术3d体验馆”=>“活的3D博物馆”



A策略未召回原因分析

无结果优化-丢词&查询改



•丢词

- 高频停用词：的、价格、团购，直接丢弃

•同义词改写

- 同session内词频共现比较高的查询对，原词和同义词同时去检索

•二次召回

- 一次召回无结果时扩大检索字段和检索范围

•无结果推荐

无结果率进一步下降

无结果优化-无合作POI召回



·无法追回项目

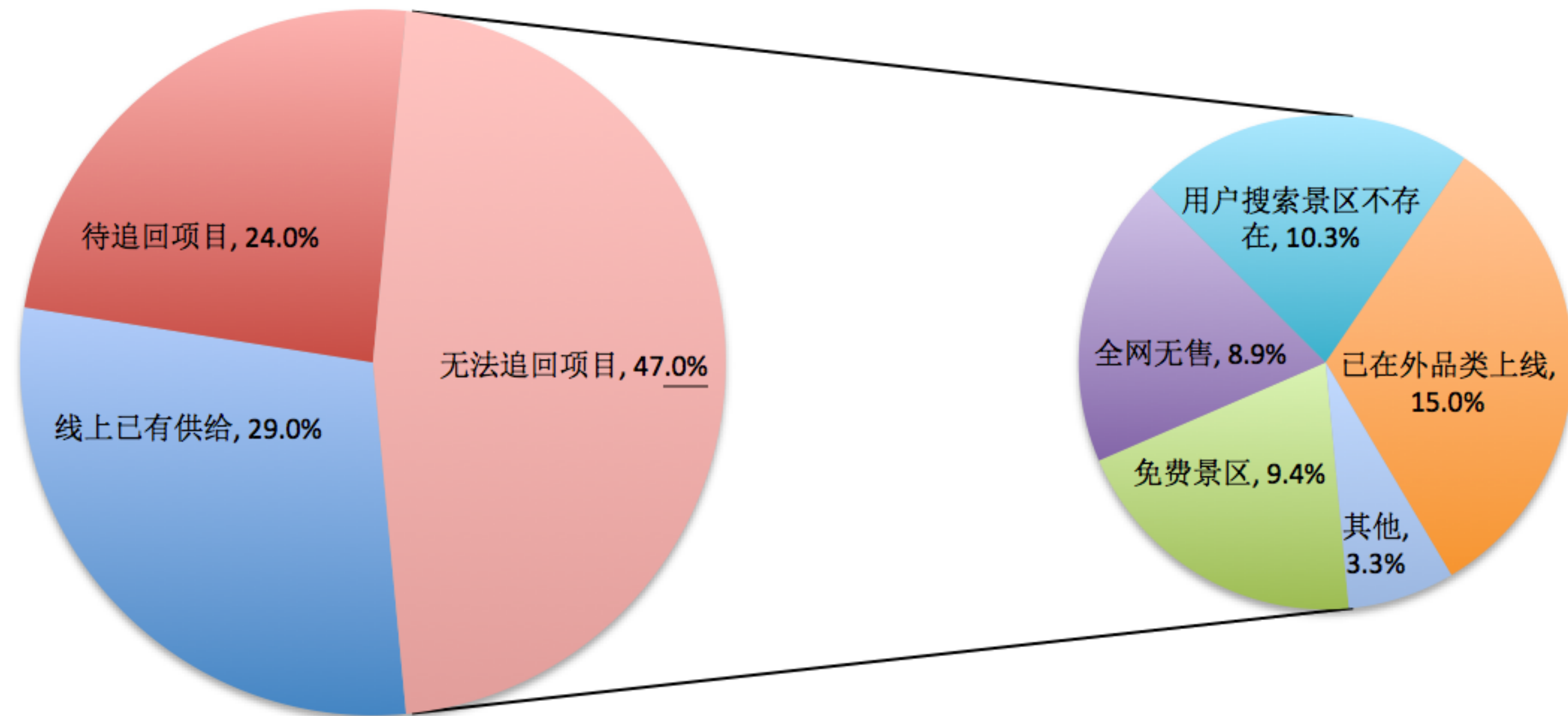
- 一次召回无结果时二次召回无合作POI
- 搜“潭柘寺”

·待追回项目

- 建立无结果反馈给BD的系统闭环

·线上已有供给

- 词表丢词->基于模型的二次丢词
- 基于Chunk Tag的丢词：停用词、线路游(从、到)、门票



无结果率下降到合理水平

分类意图识别-Chunk Tag体系

词表匹配

·精确匹配：“大理”、“云南大理”

·准确率较高，召回率不高：“大理旅游”、“去大理”

·模糊匹配：“北海公园”、“中山公园”

Query意图体系

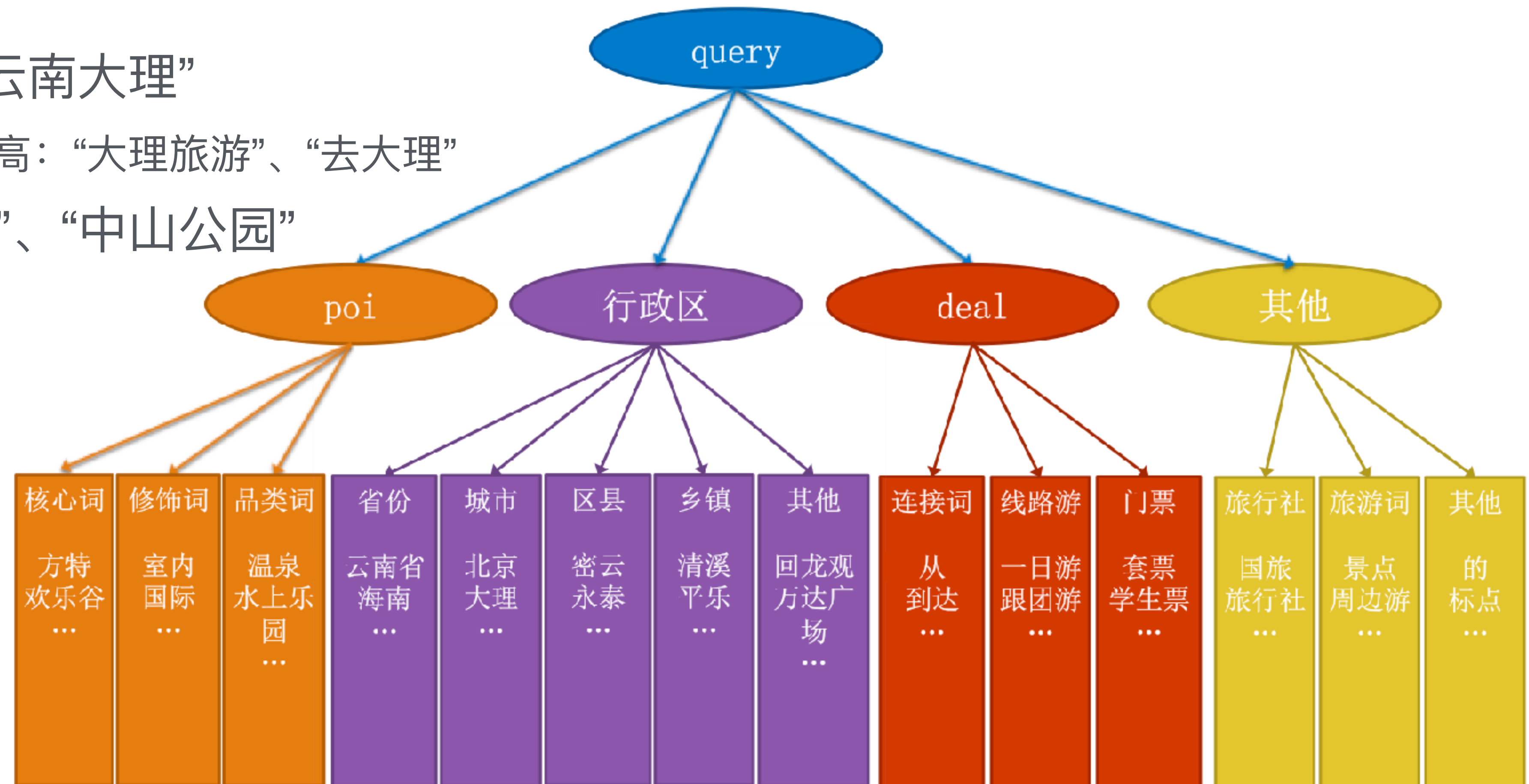
·POI

·行政区

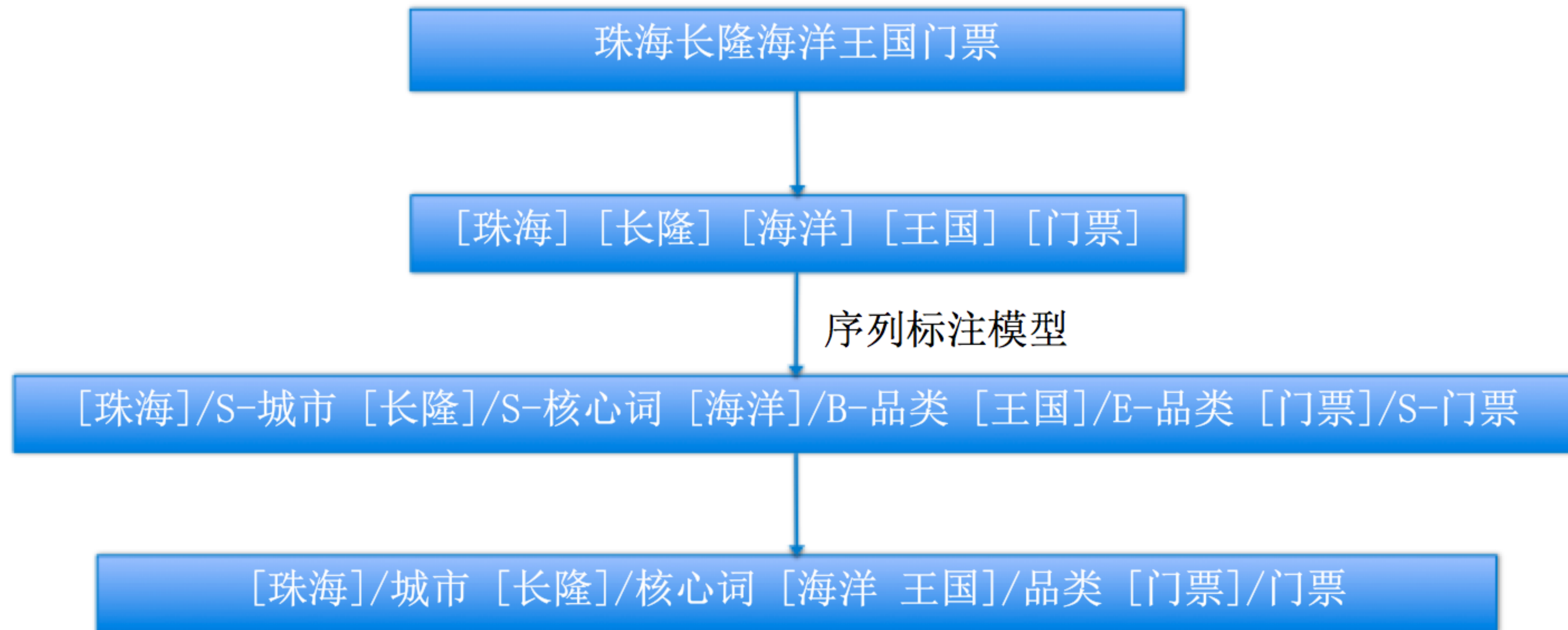
·品类

·线路游

·...



分类意图识别-Chunk分析



- 1) 识别Chunk边界: BMES
- 2) 为Chunk打Tag

分类意图识别-Chunk分析



• 算法模型：CRF

• 标记语料



- 搜索日志，词表+规则
- POI名称：核心词、品类词、修饰词

• 特征选取

- 边界特征：确定Chunk边界
- Tag特征：确定Chunk Tag
- 组合特征

term	term length	是否有城市Tag	是否有品类Tag		左熵	右熵	label
珠海	6	1	0	...	3	6	S-城市
长隆	6	0	0	...	3	4	S-核心词
海洋	6	0	1	...	5	3	B-品类词
王国	6	0	1	...	4	4	E-品类词
门票	6	0	0	...	7	4	S-门票

Chunk分析->意图识别



•意图识别

- 规则，优先级：线路游>POI>品类>门票
- “北京故宫一日游”：线路游，“北京故宫”：POI
- “北京动物园”：POI，“动物园”：品类

•意图归一

- 行政区归一
 - 知识推理：“北京朝阳”
 - 同义词
- 品类归一



粗排序改进-文本相关性



·问题

- 文本域长度影响过大：“庐山风景名胜区”、“庐山植物园”
- 多个域计算结果求和：品牌名为空，“天津欢乐谷”、“北京欢乐谷”
- 没有考虑域的动态权重，搜“动物园”：“苏州文化园”、“万鸟林”
- IDF不能体现词在Query中的重要程度

$$R_{Q,D} = \sum_{t \in Q} \left(\sum_{f \in H} \frac{tf_{t,f}}{l_f} * w_f \right) * idf_t$$

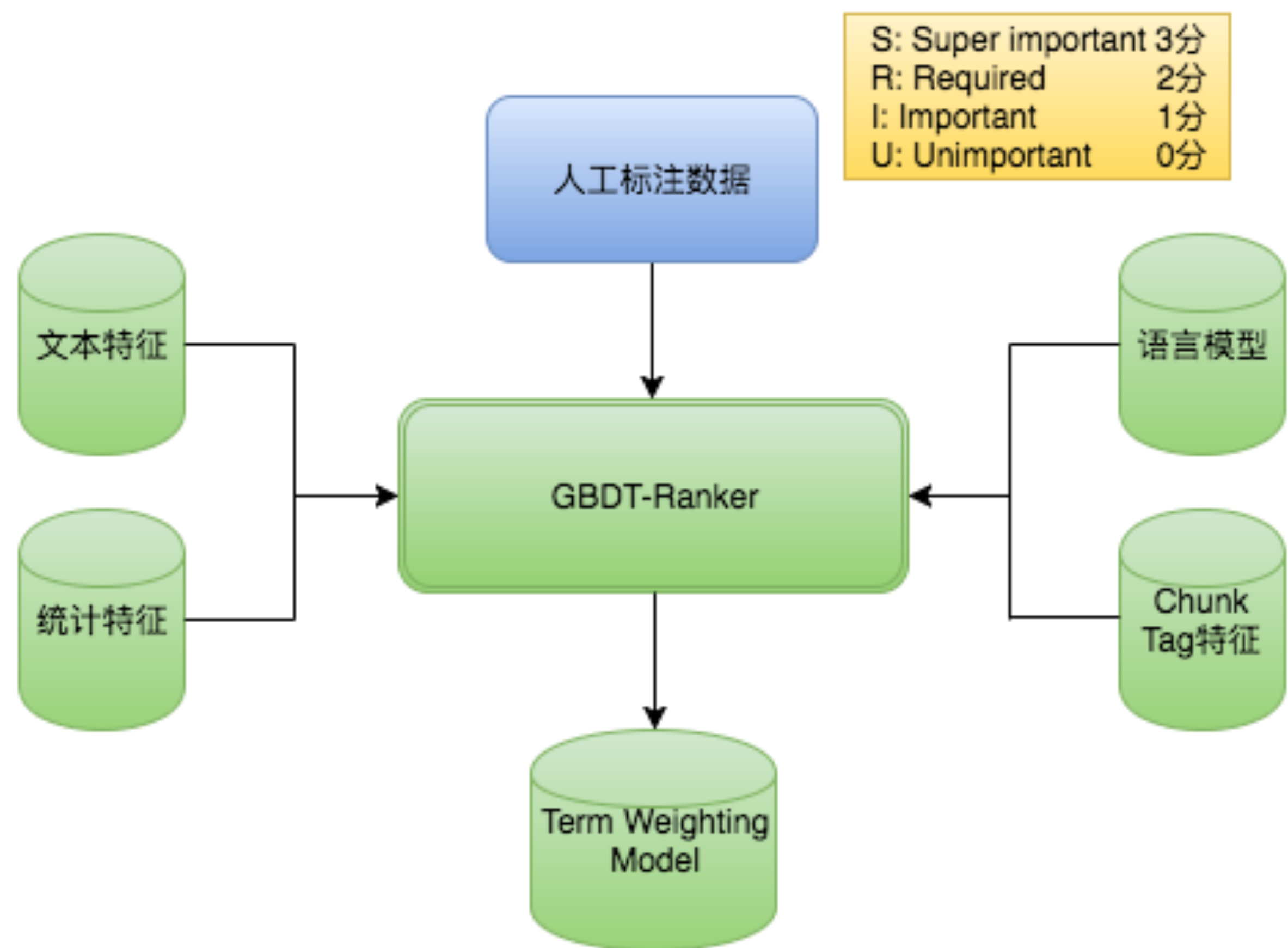
·文本相关性改进

- 降低文本域长度影响
- 对多个域求max
- 考虑域的动态权重
- Term重要度

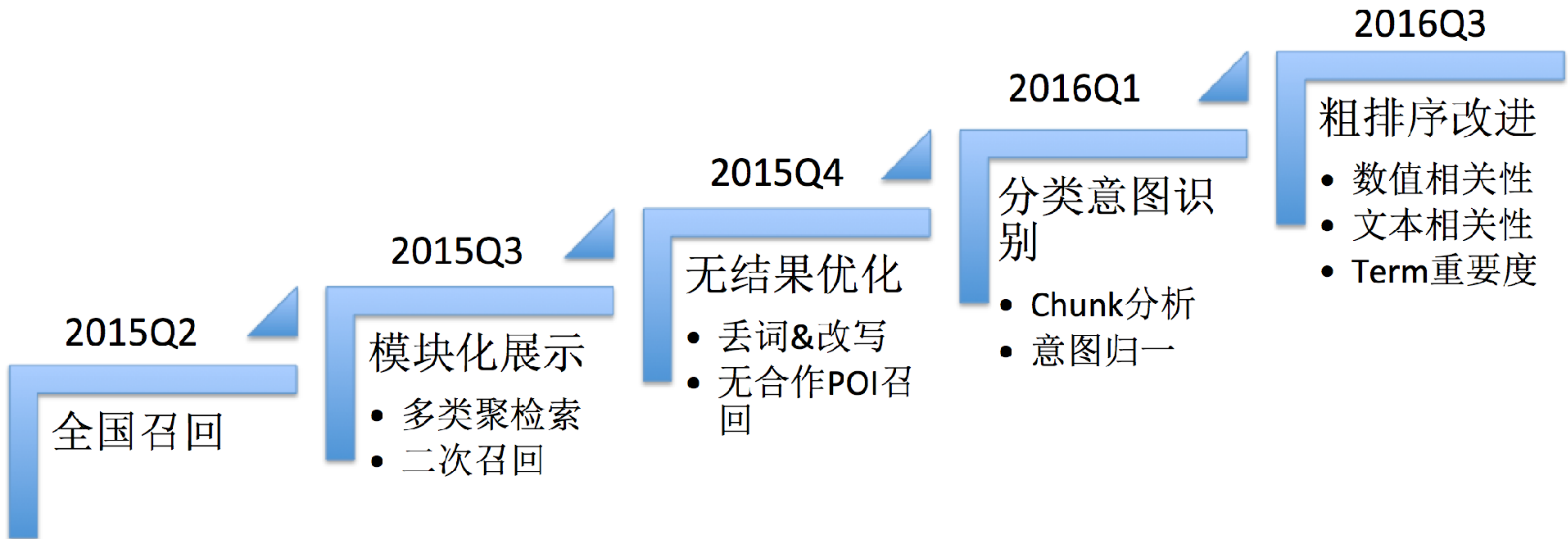
$$R_{Q,D} = \sum_{t \in Q} \max_{f \in H} \left\{ \frac{tf_{t,f} * (k_1 + 1)}{tf_{t,f} + K} * w_f * i_f \right\} * idf_t'$$
$$K = k_1 * (1 - b + b * \frac{l_f}{avgl_f})$$

Term重要度

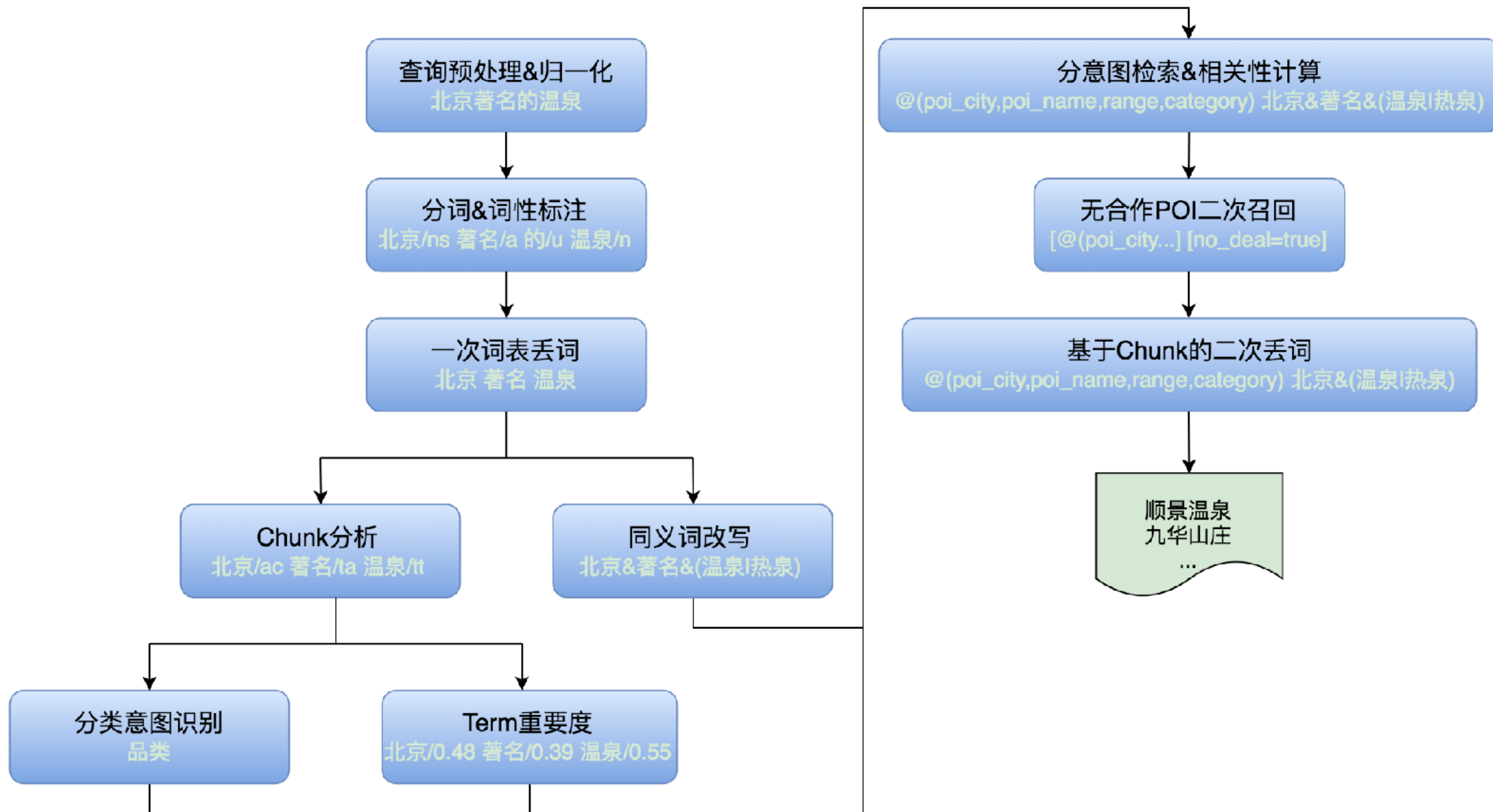
- Super important: POI核心词, “方特”、“欢乐谷”
- Required: 行政区、品类词, “北京 温泉”
- Important: 品类词、门票, “顺景 温泉”
- Unimportant: 线路游、泛需求词、停用词



搜索召回策略演进



搜索召回策略演进



旅游搜索排序策略优化

本异地排序

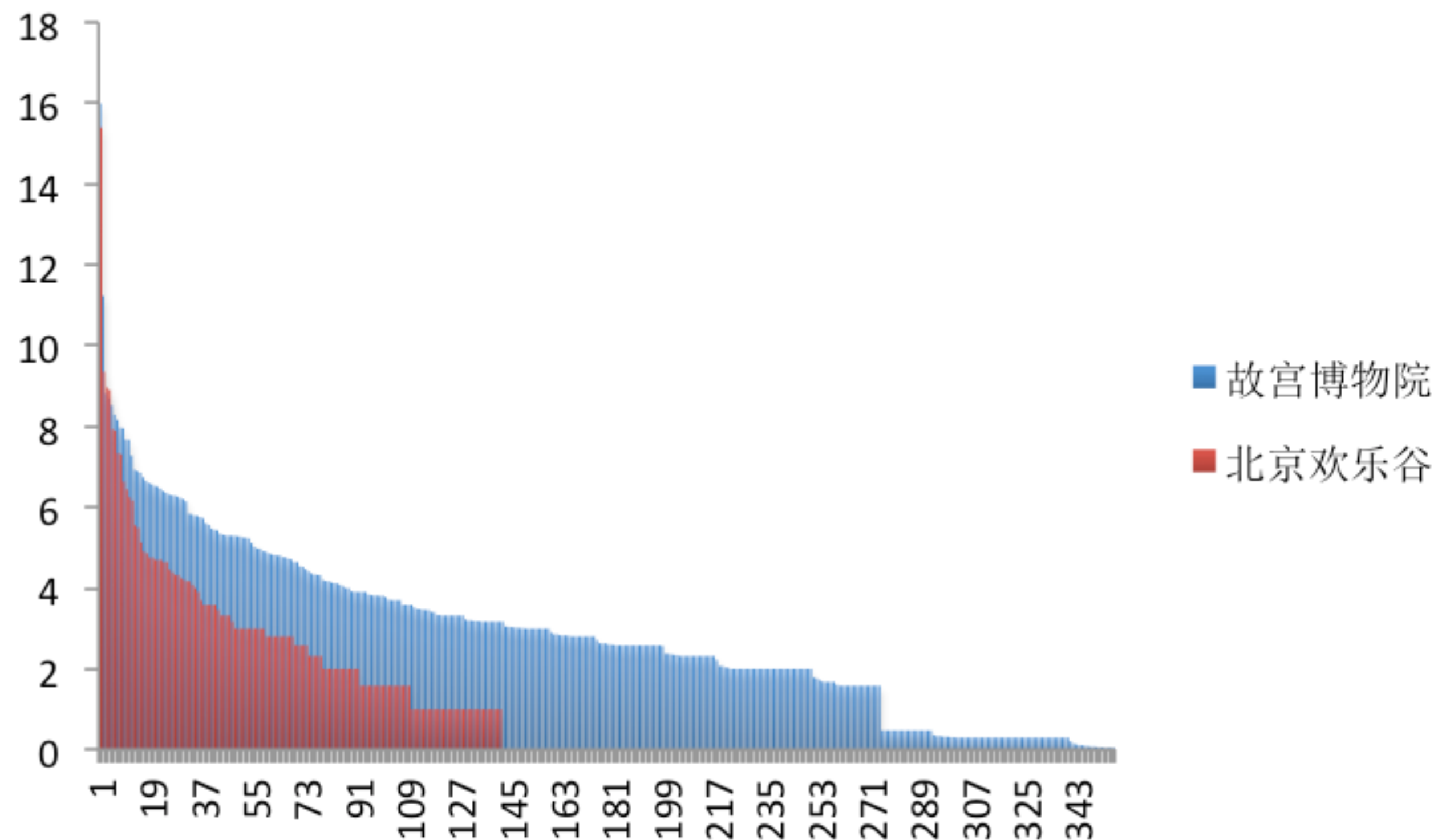
- 本地置顶

- 咸阳“乐华城”

- 分本异地训练模型

- 数据更稀疏

- 分城市统计销量、转化率



- 距离特征优化：用户定位城市、Query城市、选择城市、POI城市

多意图排序

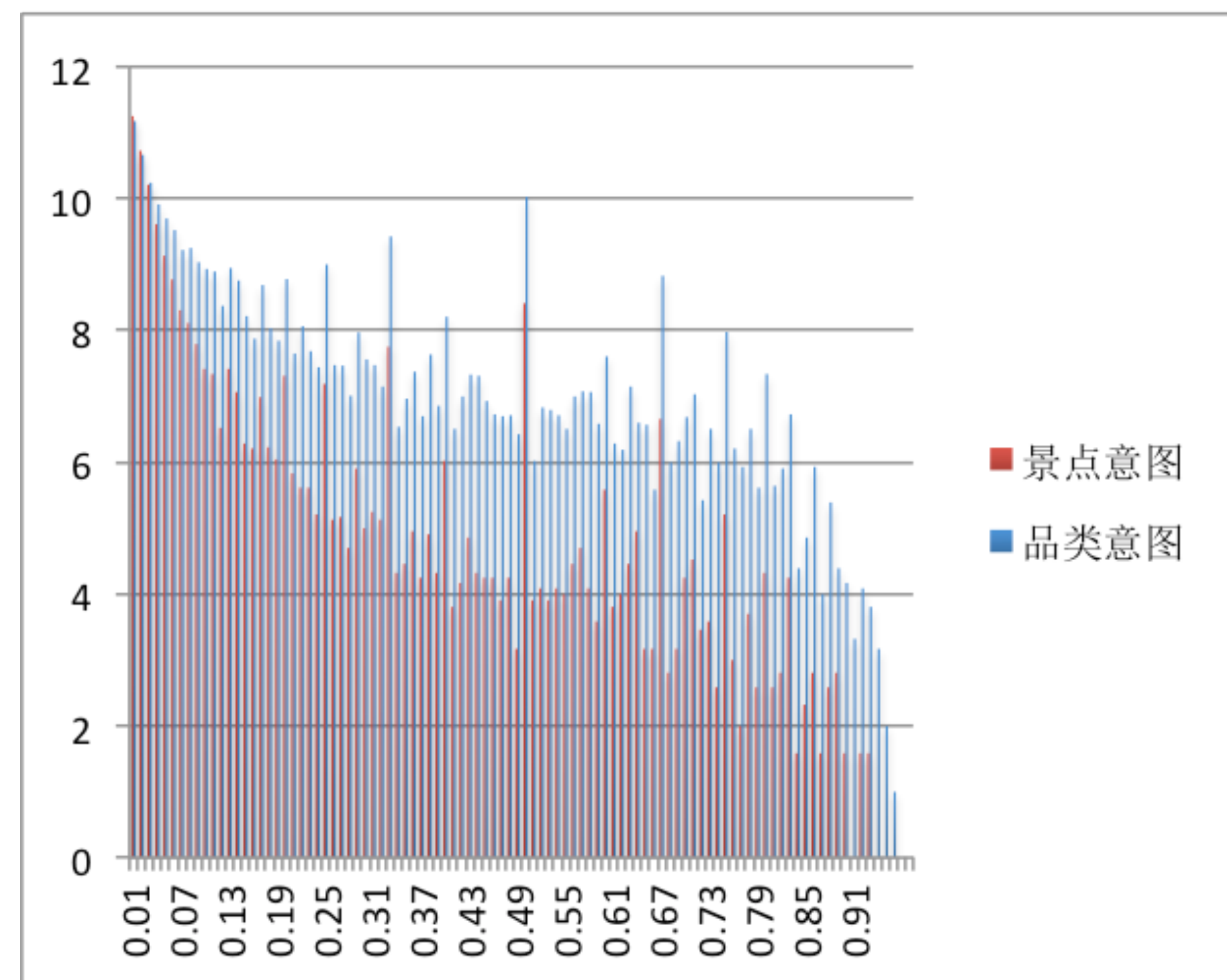
- 同一POI在不同Query下的排序

- 搜“顺景温泉”：景点POI意图
- 搜“温泉”：品类意图，距离因素很重要

- 分意图拆分销量、转化率

- 增加搜索意图特征

- 线路游排序：相似线路游Deal排序离散化



未合作POI排序

·访购主模型和文本小模型Bagging

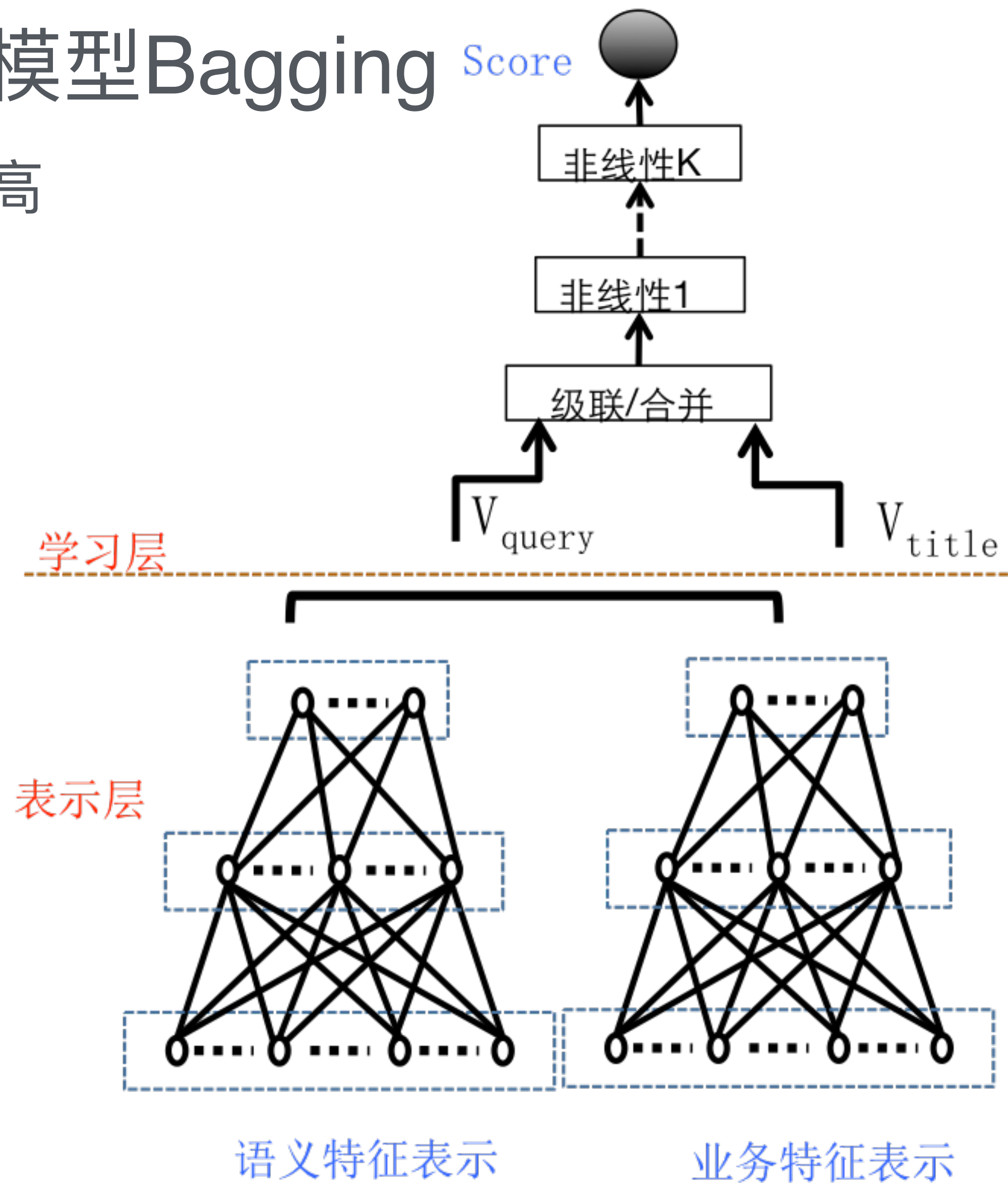
·文本/距离特征权重相对不高

·文本小模型

·文本

·距离

·ClickNet语义特征



南山滑雪场相关景点



南山滑雪场
★★★★★ 4.6分
暂无售卖

北京市 | 南山滑雪场



右玉南山滑雪场
★★★★★ 4.5分
暂无售卖

朔州市, 右玉县

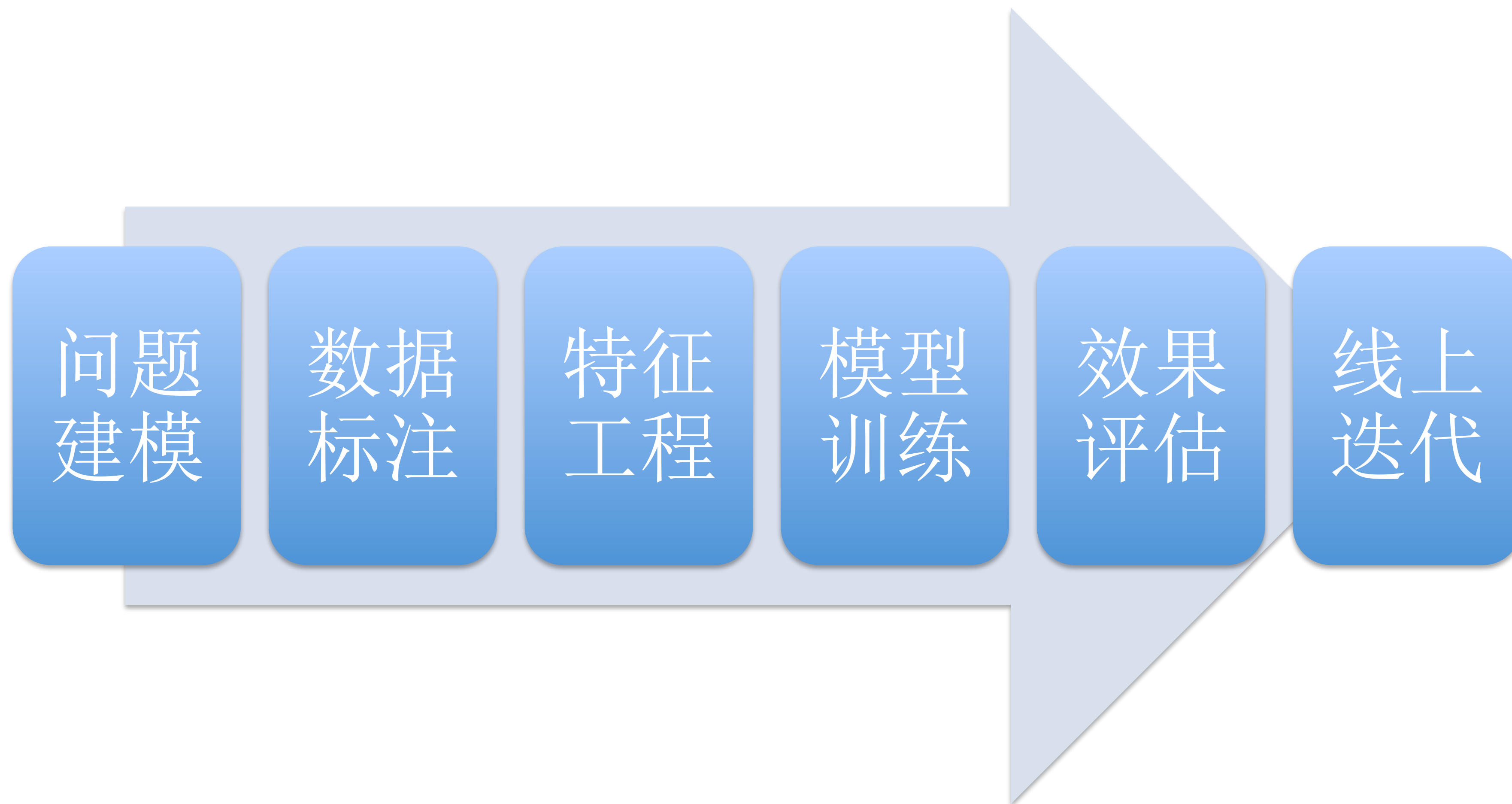


世界之窗阿尔卑斯冰雪世界
★★★★★ 4.3分 | 44771人消费
¥40.3起

深圳市, 世界之窗

机器学习在搜索排序中的应用

机器学习流程



问题建模

•GBDT

- 非线性
- High Level特征多

•XGBoost

- 泰勒展开，利用了二阶导数信息
- 对数据预排序，性能更高

•Pointwise L2R

- Pairwise性能问题

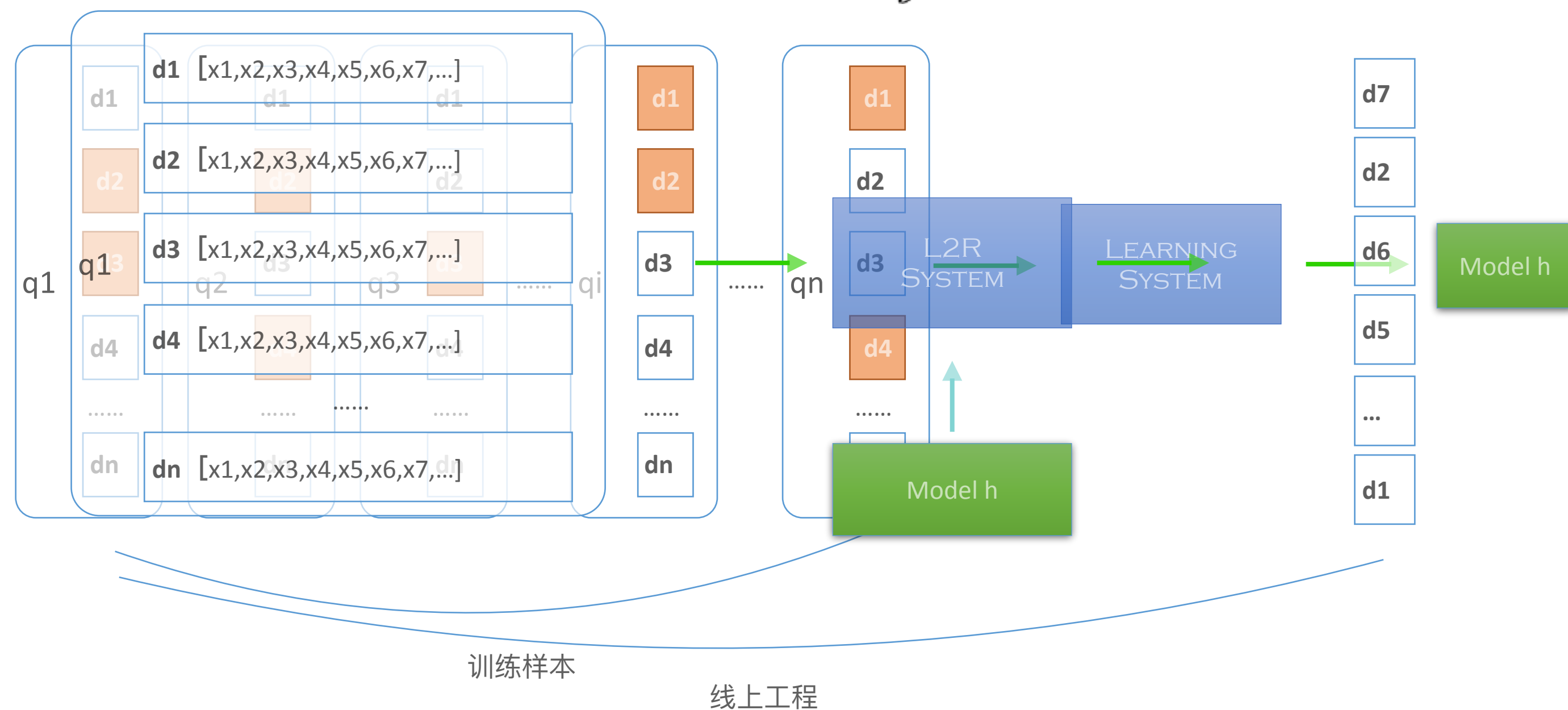
Hypothesis:
$$h_M(x) = \sum_{m=1}^M \beta_m T(x; \Theta_m)$$

Loss Function:
$$\min_{h \in H} L(h) \equiv \min_{h \in H} \frac{1}{2} \sum_{i=1}^N (y_i - h(x_i))^2$$

Update Function:
$$\{\hat{\Theta}_m, \beta_m\} = \arg \min_{\Theta_m, \beta_m} \sum_{i=1}^N L(y_i, h_{m-1}(x) + \beta_m T(x; \Theta_m))$$

$$G_j = \sum_{i \in I_j} g_i \quad H_j = \sum_{i \in I_j} h_i$$

$$Obj = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T$$



数据标注

样本关联

- global id: 串联展示、点击、下单

样本选择

- 时间窗口
- 过滤
- 样本迁移
- Skip above

样本采样

- 访购模型: 减少负样本
- 点击模型: skip above+2

样本权重

- 支付>下单>点击

特征工程

USER-POI特征

- 浏览过
- 购买过
- 距离

POI特征

- 评分/星级
- 分城市CTR/CVR
- 分城市分意图CTR/CVR
- 销量
- 地理位置

POI-QUERY特征

- 分城市CTR/CVR
- 文本相似度
- 语义相似度

USER特征

- 常住地
- 购买倾向
- 品类/价格偏好

QUERY特征

- 结果数
- 查询意图
- 粗排位置

上下文特征

- 周几/小时
- 本异地
- 天气

特征工程



•特征预处理

- missing value：不需要处理
- position bias：COEC
- One-Hot Encoding？ 周几/小时/city id
- Normalize？

•特征选择

- 特征在每棵树每个节点的信息增益之和

•特征监控

- 覆盖率
- 值域范围
- 分布异常

特征	coverRatio			numValid		
	参考值	监控值	波动率	参考值	监控值	波动率
CLICKNET	0.931341	0.481258	0.483263	3271405	1745754	0.46636
DISTANCE	0.651660	0.033357	0.948812	13447	662	0.950770

模型训练

•模型训练

- 单机VS分布式
- 目标函数：binary:logistic
- 过拟合VS欠拟合
 - 样本大小&树的棵数
 - 样本和特征随机采样
 - 模型复杂度：max_depth, min_child_weight

•通用离线训练工具

- 流程抽象化、组件化
- 提供公共组件，支持定制组件



效果评估&线上迭代

• 离线评估

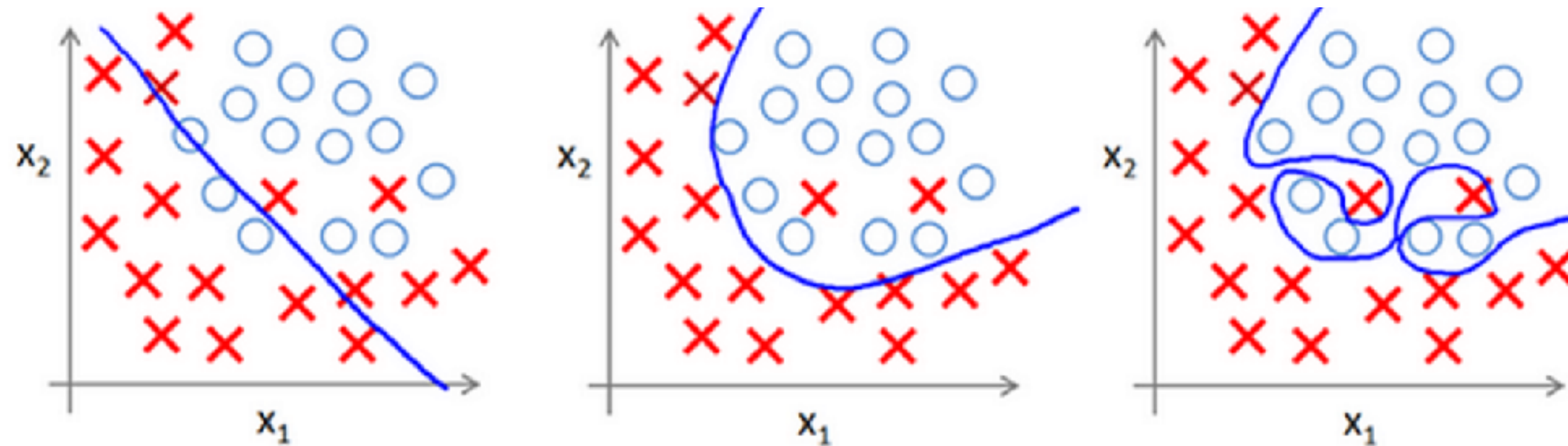
- AUC
- logloss

• 在线评估

- ABTest: 按UUID分流

• 线上迭代

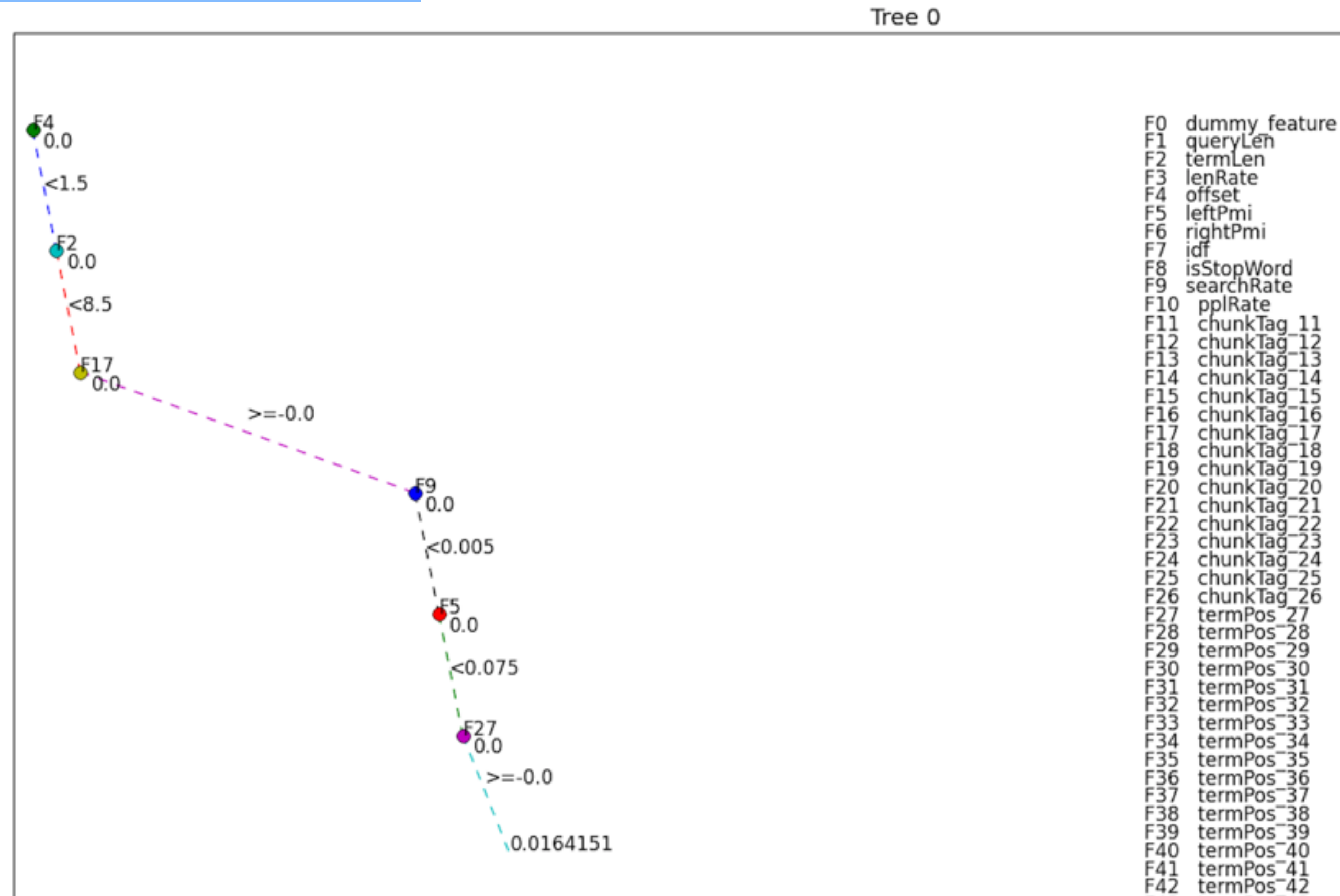
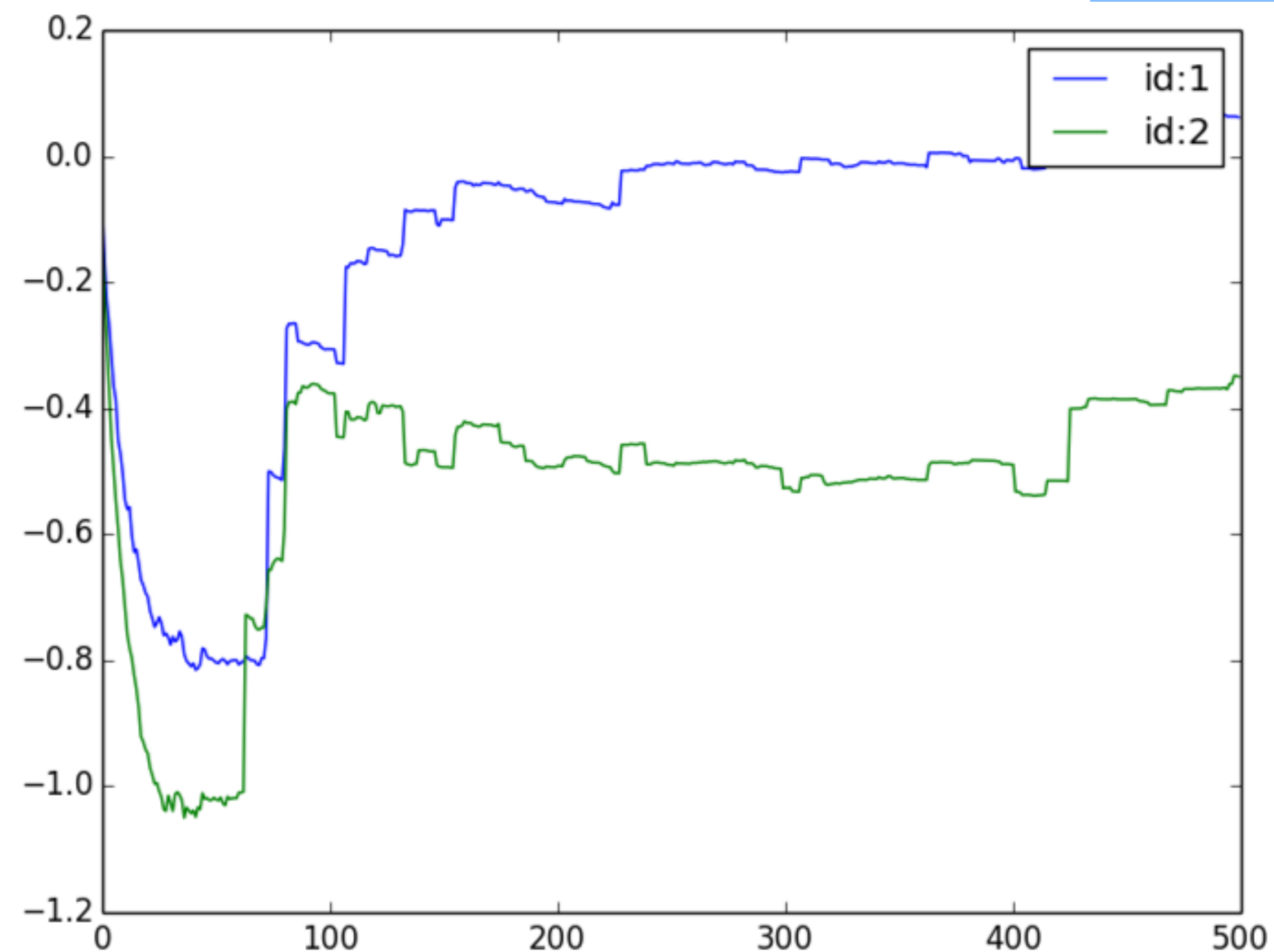
- 模型更新
- 特征漂移: 更新延迟



训练集表现	测试集表现	问题
<期望目标值	<期望目标值	Underfitting
>期望目标值	接近或略逊于训练集	合适
>期望目标值	远差于训练集	Overfitting

模型调试

模型Debug工具



Q&A

Thanks

Eat better, Live better.