DataFunSummit

自然语言处理

在线峰会

信息抽取与检索论坛

2021.07.10（周六） 09：00~17：30

2021

飞猪技术 | DataFunSummit

2021

飞猪旅行酒店搜索相关性建设

林睿 阿里巴巴飞猪算法专家

目录

CONTENTS

01

题目

Subject

酒搜背景

酒店小搜

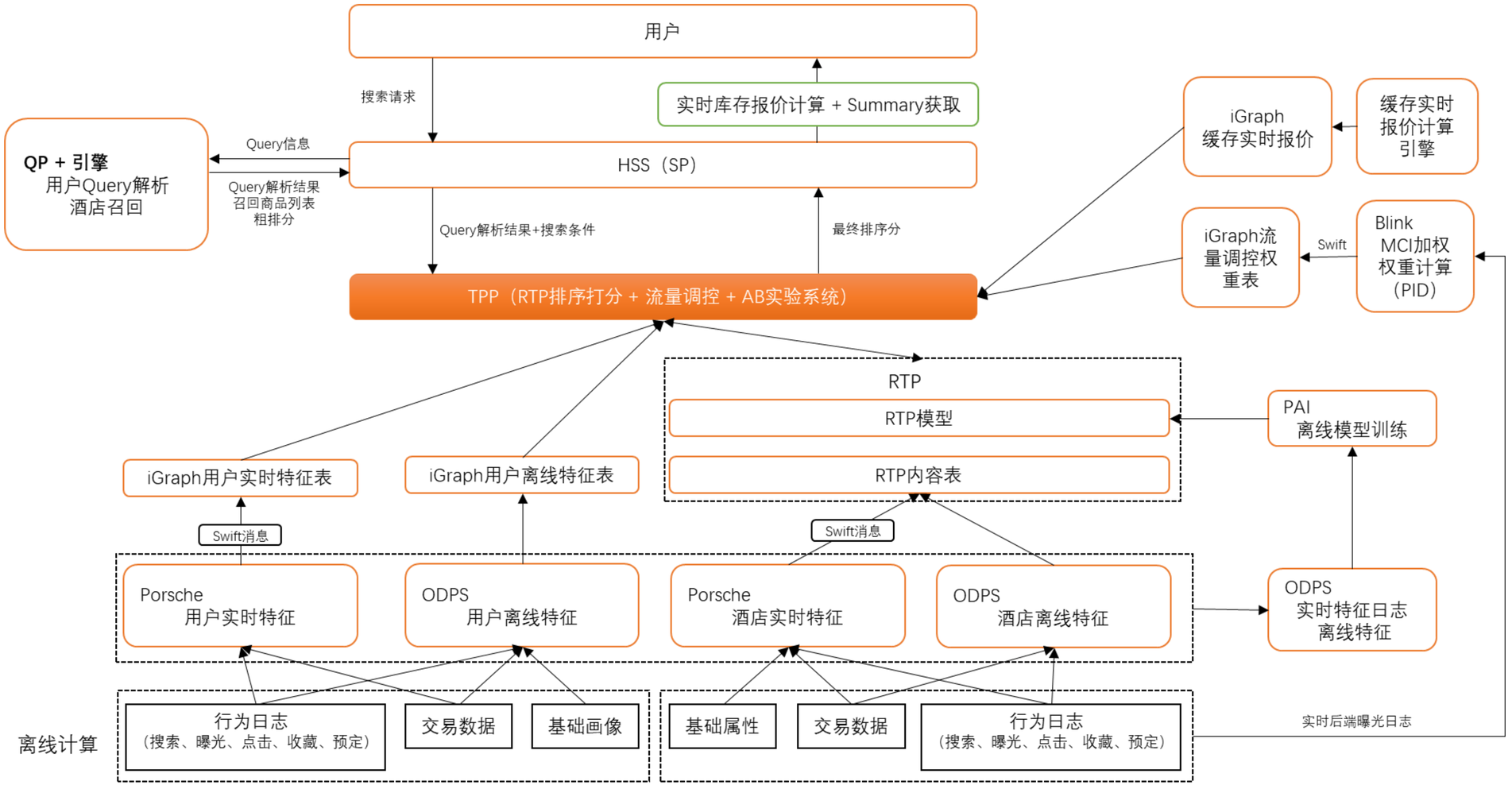
业务特点

- 多端多场景多意图
- 多元的搜索条件
- 决策周期长；用户行为稀疏
- 周期性需求
- 个性化的结果

带来挑战

- 多维的搜索query
  - 用户：距离、价格偏好
  - 关键词：POI、筛选条件
- 多维相关性需求
  - 空间
  - 价格
  - 文本

酒店搜索架构



02

题目

Subject

酒店相关性

场景与相关性

特点

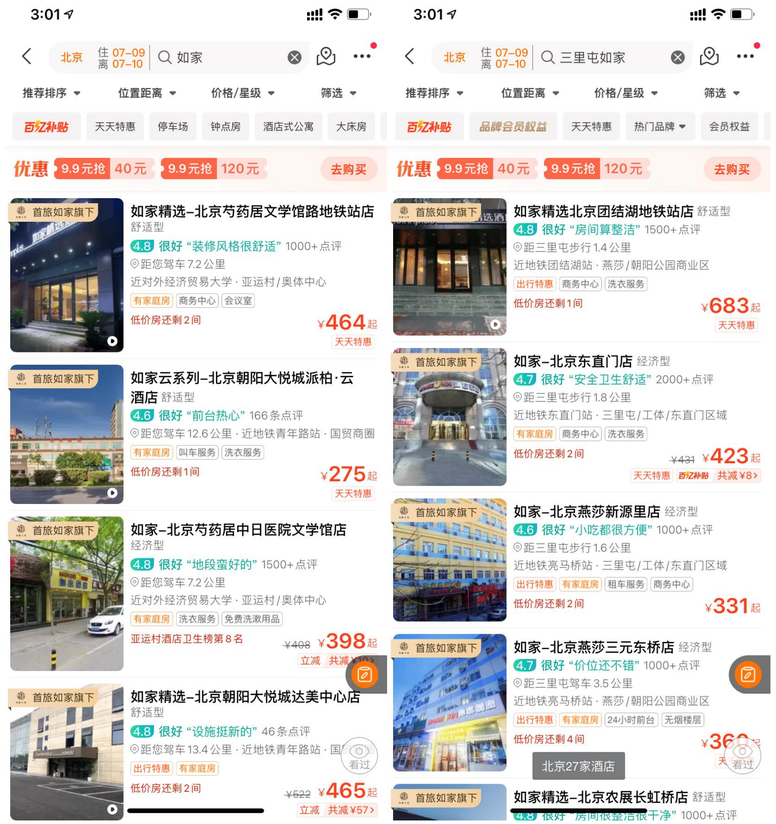
- 空搜/附近搜与景点/商圈搜都对距离相关性有比较强的需求，前者看重酒店与用户的距离，后者看重酒店与目标地点的距离。
- 空搜/附近搜需求较泛，排序偏用户个性化；景点/商圈搜用户有比较明确的需求，以满足用户需求为主。

空搜/附近搜　　景点/商圈搜

场景与相关性

特点

- 空搜/附近搜与景点/商圈搜都对距离相关性有比较强的需求，前者看重酒店与用户的距离，后者看重酒店与目标地点的距离。
- 空搜/附近搜需求较泛，排序偏用户个性化；景点/商圈搜用户有比较明确的需求，以满足用户需求为主。
- 名称搜注重文本相关性，以品牌、酒店名符合用户需求的酒店优先。
- 用户还有可能使用名称与商圈混合的搜索，需要综合文本相关性和空间相关性等，给出符合当前用户的最优排序



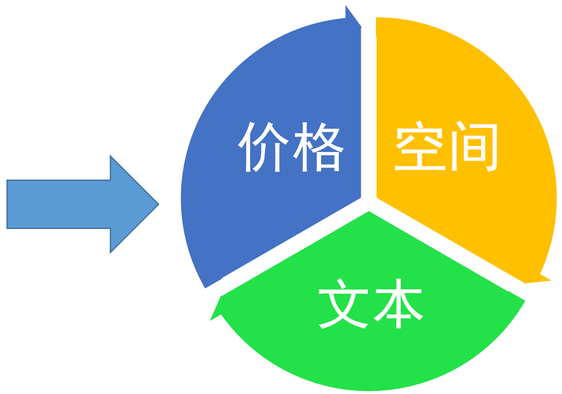
名称搜　　混合搜

酒店的相关性

总结

- 酒店相关性是一个由文本、空间、价格多元融合的相关性
- 受用户、场景、query等不同的条件影响，相关性的侧重点也会有所不同
- 相关性的多元化导致标注难度大，只能依赖点击和成交的label

用户	场景	query
• 年龄 • 性别 • 购买力 • 常住地 ……	• 附近搜 • 景点搜 • 商圈搜 • 名称搜 ……	• 关键词 • 筛选项 • 过滤条件 ……



方案：识别用户需求，构建多元相关性，根据用户需求得到最终的整体相关性

03

题目

Subject

基础建设

核心因子预估

背景

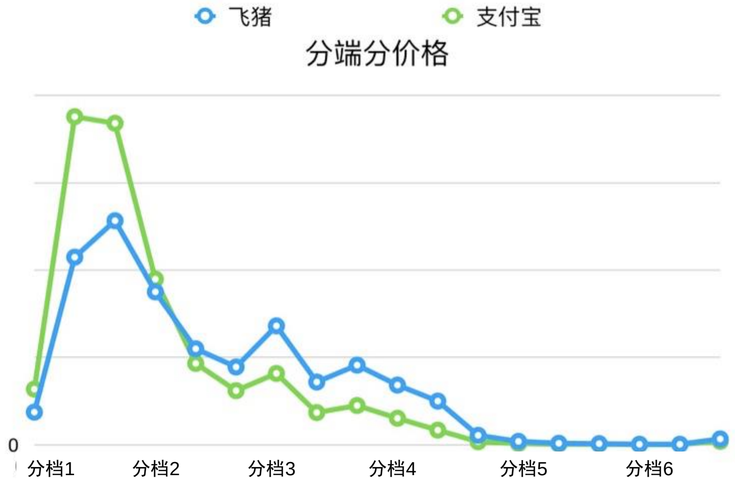
价格和距离是酒店购买决策中的重要组成部分，也是酒店相关性的重要一环，好的j需求预估能更好的构建酒店搜索相关性

面临问题

- 按实际物理意义划分，标签分布极度不均匀

距离因子分布

价格因子分布



$$\mathrm{argmax}_{y\in[L]}\exp(w_y^\top\Phi(x))/\pi_y^\tau = \mathrm{argmax}_{y\in[L]}\, f_y(x) - \tau\cdot\log\pi_y,$$

核心实体识别

背景

- LBS召回是酒店召回的极为重要一环，这就依赖于POI识别的准确性。同时，POI的准确识别也为空间相关性的计算提供了支撑
- 酒店名称、酒店品牌的识别能更好的辅助用户意图的判断。

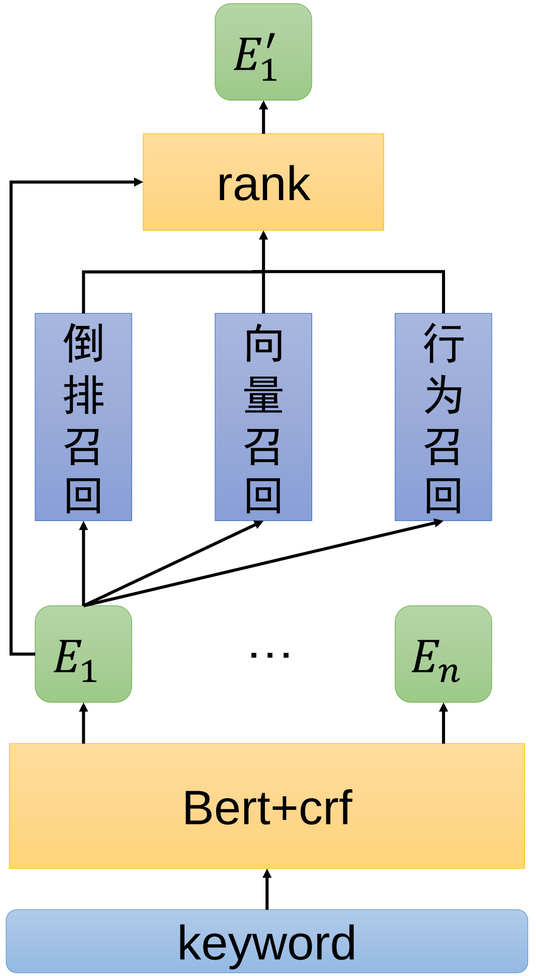
主要问题

- 实体词识别的准确率
- 实体词与业务中真实实体的映射

核心实体识别

方案

- Mention识别
  - 利用bert+crf的方式进行ner识别
  - 使用标注数据与实体库做数据增强
- 多路召回候选实体
  - 倒排精确召回
  - 向量召回
  - 用户行为召回
- 排序
  - 基于文本相似分、热度、点击、所在城市等特征构建了简单的排序模型



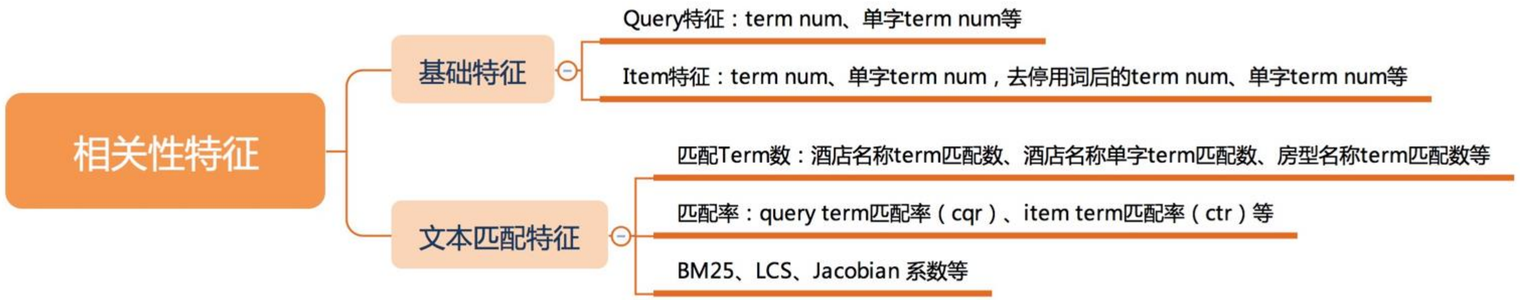
04

题目

Subject

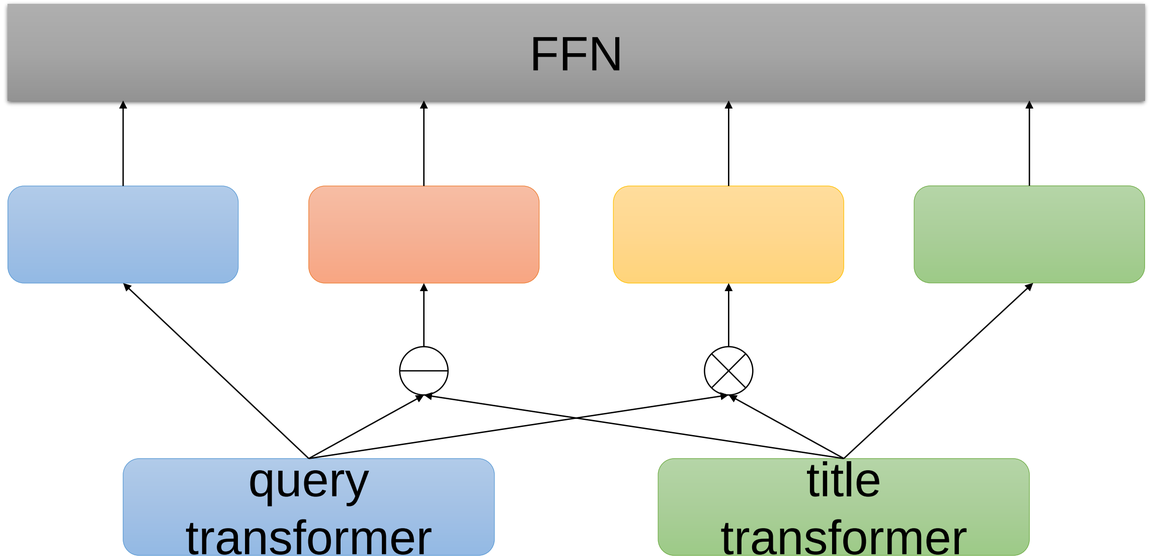
相关性建模

文本相关性



方案

- 粗排
  - 利用计算好的文本匹配特征排序
- 精排
  - 利用文本匹配特征分档作为精排相关性特征
  - 利用原始的term和实体特征构建文本相关性网络



空间相关性

现状

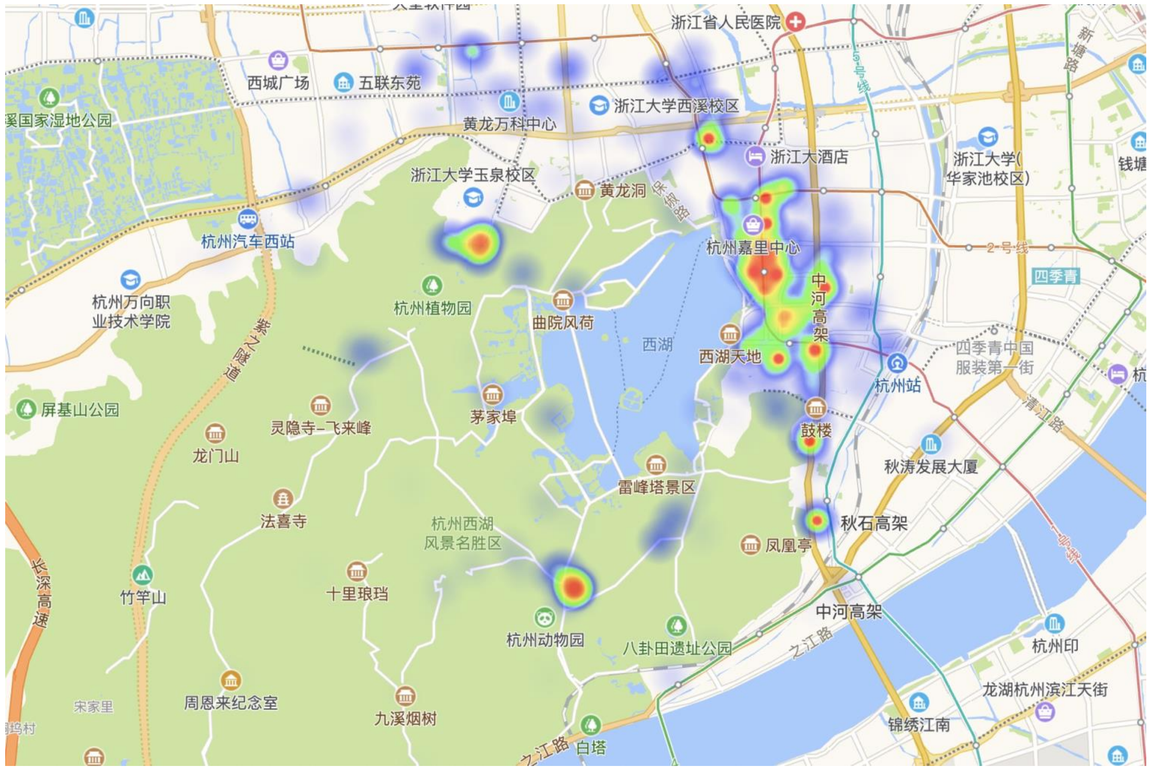
- 用户酒店距离 & POI酒店距离特征与距离预估因子交叉
- 酒店、POI及用户的geohash id特征

问题

- 距离这个特征不能很好的衡量空间相关性
- geohash作为id特征过于稀疏，丢失了geohash中包含的地理信息

优化

- 将geohash转化为原始的二进制序列，用一个token list表示，保留了原有的地理信息。

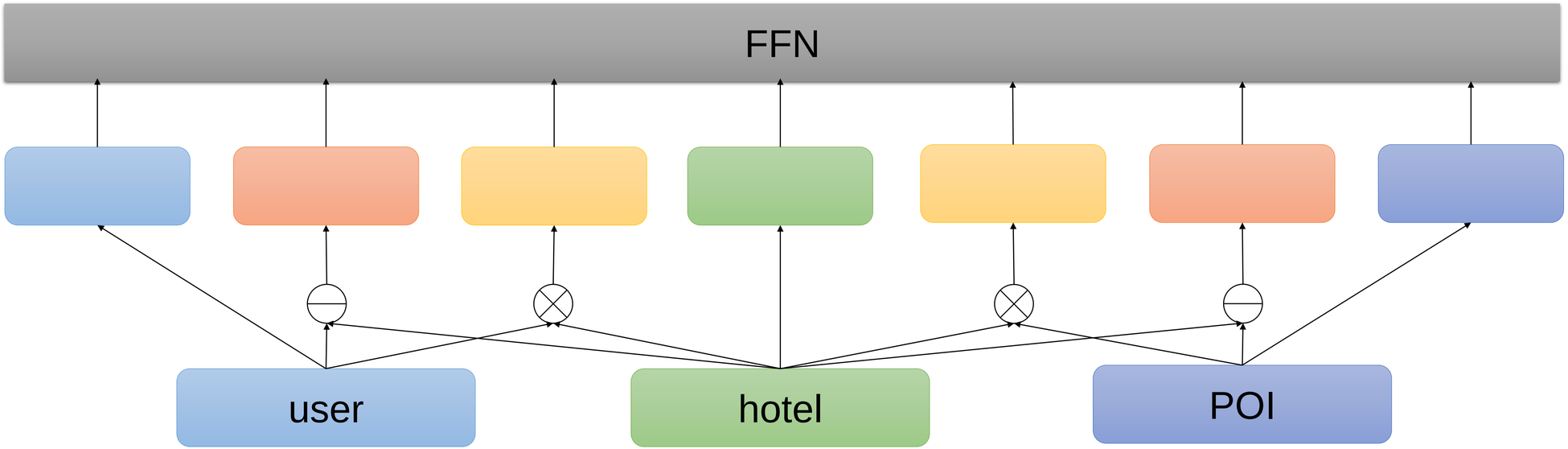


空间相关性

W					M					7					0					N					T				
1	1	1	0	0	1	0	0	1	1	0	0	1	1	1	0	0	0	0	0	1	0	1	0	0	1	1	0	0	1

优点

- 保留了空间信息，相邻的geohash最终在网络中的embedding相近
- 转化成了一个token序列，可以用文本的方式计算相关性



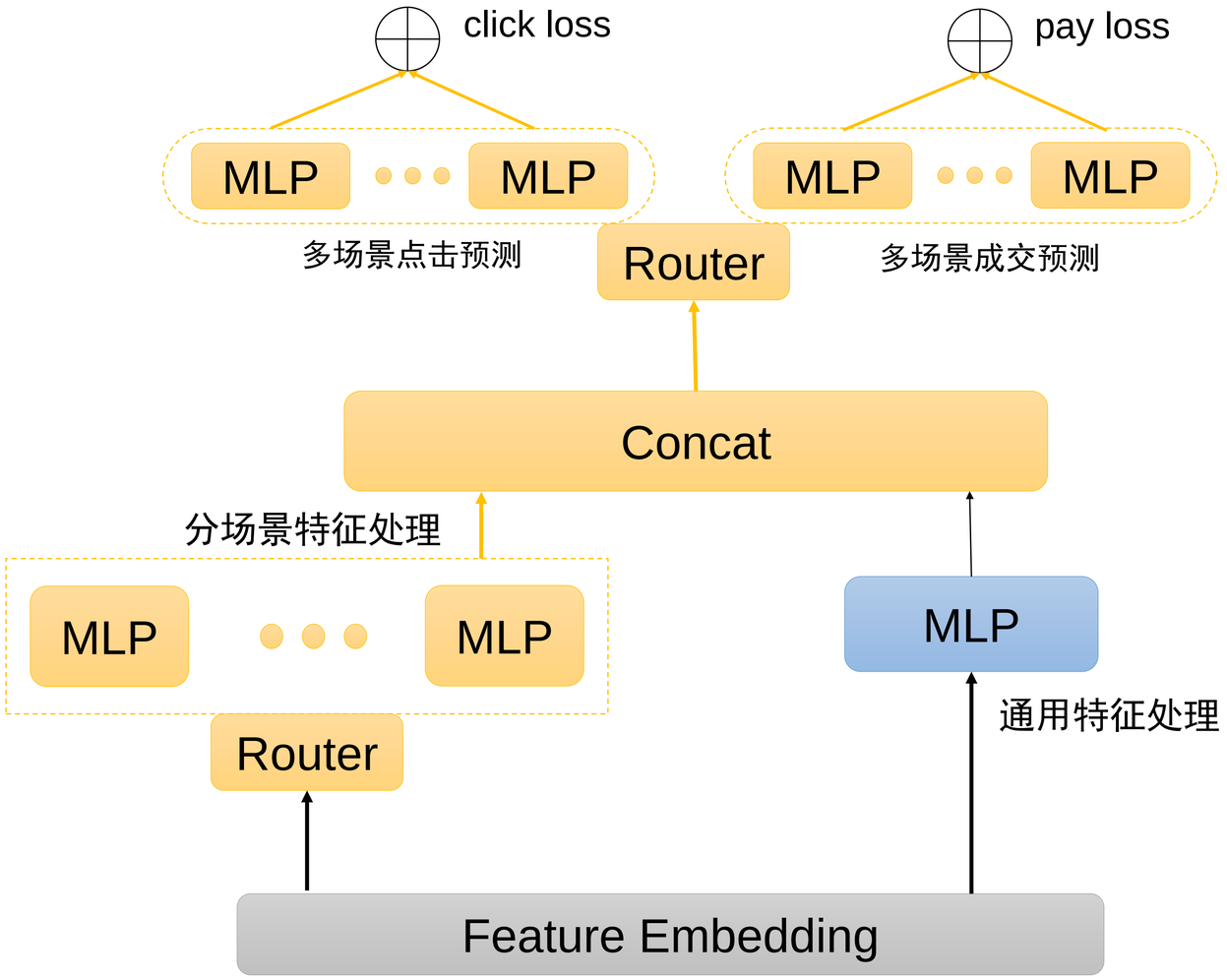
多场景相关性

问题

- 酒店搜索的多端多场景，多元的相关性
- 不同场景对相关性侧重有异同
- 场景存在融合情况

方案

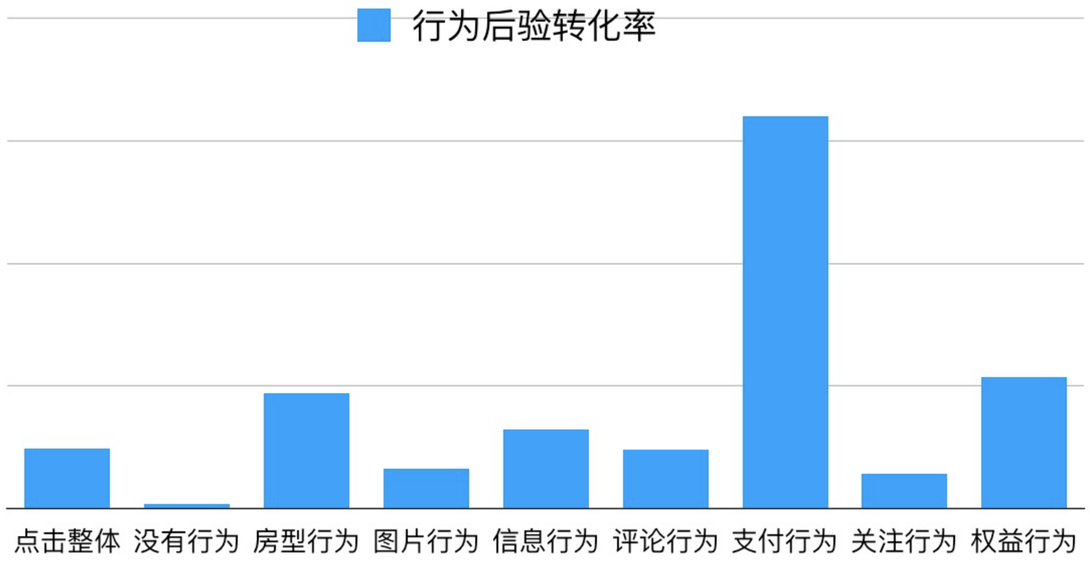
- 同时训练通用特征提取和不同场景的子任务特征提取，让不同场景关注不同的特征
- 在点击和成交预测上也是用分场景的分类器。

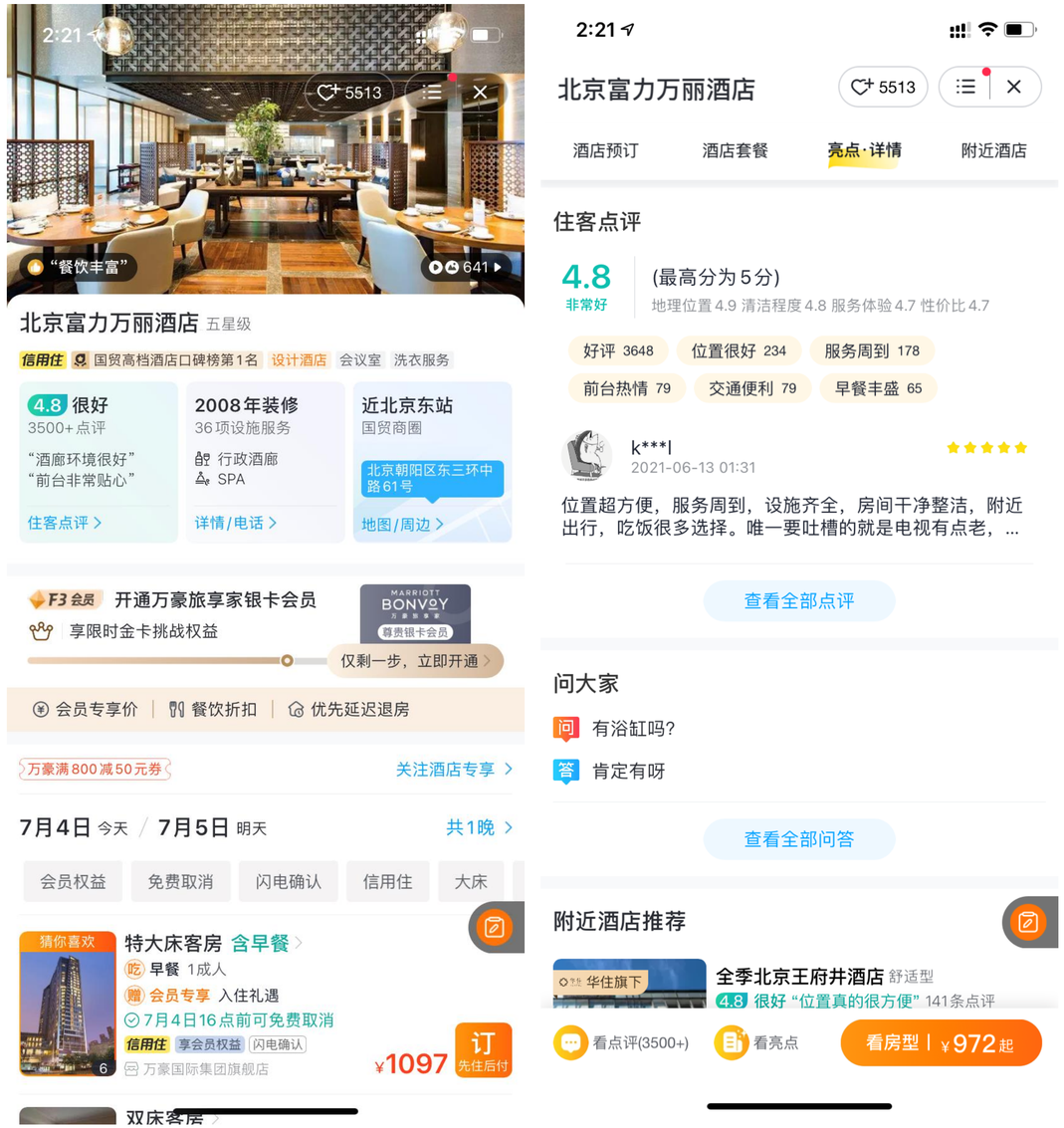


详情页特征

优化

- 在成交和点击之间，用户在酒店详情页上也会有丰富的行为，但是传统的模型忽略了这一点。仅使用了点击与成交的label





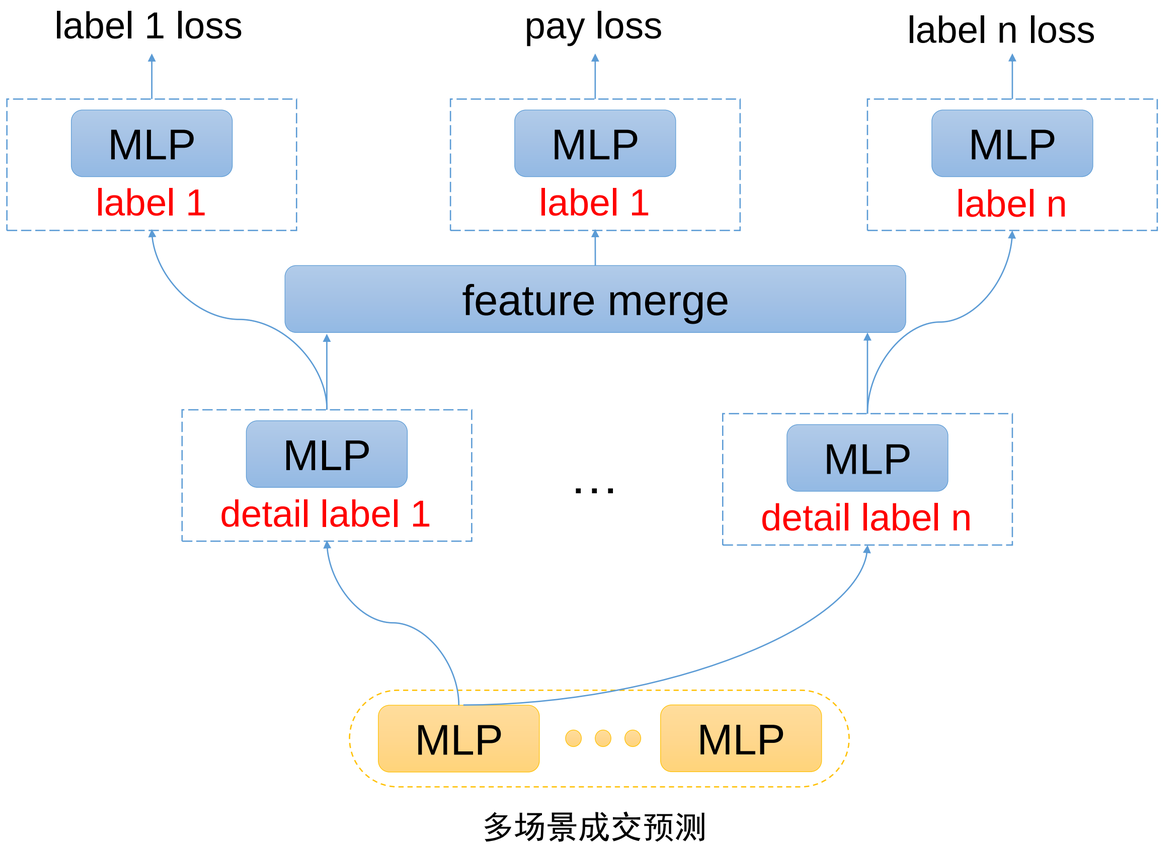
详情页特征

优化

- 在成交和点击之间，用户在酒店详情页上也会有丰富的行为，但是传统的模型忽略了这一点。仅使用了点击与成交的label
- 我们引入详情页行为预测任务作为辅助任务，期望其为相关性任务优化带来增益
- 最终辅助任务和点击成交预估任务按照一定的人工权重融合计算最终loss

$$loss = w_{click} \cdot loss_{click} + w_{pay} \cdot loss_{pay} + w_{detail} \sum_{1}^{N} loss_{label_n}$$

- 考虑到酒店购买决策周期较长，我们还引入了全域的成交来优化酒店搜索的成交label



Future Work

Query分析

- 更精准的空间价格预估
  - 用空间分布预估替代距离分布预估
  - 构建价格比例分布预估

相关性

- 空间和文本相关性模型结构升级,价格相关性模型构建
- 引入历史搜索序列计算上下文的相关性

飞猪技术 | DataFunSummit

THANKS!

今天的分享就到这里...

Ending