姓 名 ：杨新权

电 话 ：18811171651

邮 箱： starspringcloud@gmail.com

出生年月：1988.05

学 历 ：硕士

住 址：北京市

基本信息

教育经历

电子科技大学 2010/08-2013/06 硕士 / 模式识别

福建师范大学（一本） 2006/09-2010/07 本科 / 电子信息工程

工作经历

**2015.03-至今 阿里集团-高德-搜索 推荐算法专家**

1. 泛搜：高德本地生活（酒店、美食、生活服务、景点等）搜索，用户并没有明确的poi（Point of Interest）倾向，需要的是某一类的item的集合。 流量入口为高德地图框搜、高德附近页金刚位（频道位）、主图工具箱icon、poi周边搜、酒店portal页等。建模任务主要为ctr/cvr。建模侧重用户个性化和LBS空间相关性。工作内容涉及召回、粗排、精排、重排。
2. 精搜：高德地图基础搜索，用户有明确的poi意图。流量入口主要是高德的框搜。建模侧重相关性（文本相关性+空间相关性）。建模任务主要为ctr。负责过query的成分分析（NER）; query品牌识别（NLU）；深度语义相关性; 搜索词自动补全(suggest)。
   1. **– 2015.3 奇虎360 NLP算法工程师**
3. 传统NLP任务： 利用基于n-gram的分词算法对短信内容进行分词； Viterbi算法的同义词变换; 基于模版的对话系统。

项目经验

**2018.05-至今 高德 本地生活**

高德本地生活（酒店、美食、生活服务、景点等）搜索建模：

1. 召回：多路召回架构，包括文本召回（term倒排索引）、个性化深度向量召回、user trigger召回、冷启动保量召回。向量召回模型使用类似MIND模型的双塔序列结构。使用softmax loss，增加自监督学习作为辅助loss。最后多路结果按比例merge (100 : 100: 60: 40)。
2. 粗排：向量召回使用双塔cosine similarity排序，文本召回路使用gbrank模型，建模任务为ctcvr任务。特征分4个维度：user侧（profile + 行为序列）；item侧（商品属性）；query信息； LBS空间特征。
3. **精排**：样本选取有用户行为的展现日志落地表，正负样本比例约为1：4，特征体系和粗排类似。学习方法为point + pair-wise的方式。 模型从wide&deep到使用deepfm来学习特征的自动交互、再到使用multi-head attention对用户长期行为序列、实时行为序列学习LastN信息，并使用sim模型的hard-search思路筛选attention候选集。学习任务为ctr/cvr多任务（essm）。
4. **重排**：精排只注重item的排序指标，缺乏整个list的视野，做不到整体最优化。重排两个建模目的：a. 推荐系统锚定效应；b. 搜索结果的多样性。样本构造方法：对线上精排分进行落表，选取精排分top100。建模任务为ctcvr。特征体系复用精排并增加上下文context 特征。 模型结构为 PRM：把精排的输出作为transformer encoder模块的输入，softmax CE作为loss，同时增加辅助loss（多样性）
5. **统一模型**：多行业多任务建模：高德本地生活有多个业务（酒店、美食、生活服务、景点等），分别维护一套模型成本巨大，且部分场景数据稀少，不利于建模。此外同一个业务也有这多种任务（ctr、cvr）。在综合PLE、ESMM、HiNet等模型优点的基础上，实现了多行业多任务的统一精排模型。上线取得良好的正向收益
6. 多样性策略（DPP）, 用户冷启动（Airbnb KDD 2018）利用用户历史行为训练user embedding, 基于探索策略的商品冷启动。

专业技能

1. 熟悉垂类搜索/推荐系统：粗排、精排、重排序； 多行业多任务建模、多样性、冷启动
2. 搜索相关经验：learning to rank，Text Classification， Semantic matching，NER
3. 熟悉tensorflow框架和大数据开发。有良好编程功底Python、c++ 、c、java、hive
   * + 1. **高德 搜索词补全**

高德搜索建议（suggest）： 用户在输入框输入query的过程中，自动补全query。并推荐用户可能感兴趣的poi/keyword。

1. 前缀树召回候选结果，利用文本相关性（cqr、cpr等）进行粗排，获得候选poi/keyword。
2. 样本对齐：把中间输入过程的所有query，对齐到最终发起搜索的完整query上。
3. 基于用户历史行为、poi热度特征(点击率)、query特征，空间特征，构造2-3阶交叉特征。
4. pair-wise loss + gbrank建模，通过调loss function方法对模型引入先验知识
   * + 1. **高德 基础搜索**

高德基础搜索，用户希望检索某一指定的POI，需要关注文本和空间相关性。系统架构分为：NLP模块、倒排索引召回，相关性排序（粗排）、ctr预估排序（精排）。

1. 基于bert的深度相关性模型：

通过用户历史点击行为，挖掘query 和 doc title 对，有点击为正样本，负样本进行随机采样，正负样本为1:4。模型为双塔结构， 基塔是bert 中文预训练版。利用高德的数据进行fine-tuning，freeze底层embedding, 在CLS输出增加FC层。使用softmax ce作为分类loss。

1. 品牌搜：
   1. 品牌识别： 多分类建模，基于fasttext的query意图识别，样本增广：基于种子样本(人工标注 + 品牌知识库)， 利用种子样本 + 同义词 + 扩展词生成一个品牌识别模板，利用模版扫描query log 来构造新的样本，loss 为hierarchical softmax。
   2. 品牌排序：分层排序，优先考虑权威性，其次考虑距离。
2. 成分分析(chunk):

NER任务。针对地图场景，提炼20种成分标签，进行人工标注，作为CRF的训练样本，模型准召可达90%。亮点:采用2个CRF模型串行 (分别预测边界、标签)、标签归并等手段，提升4倍预测速度

项目经验