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）搜索建模：

1. 召回：多路召回架构，包括文本倒排召回、个性化深度向量召回、user trigger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**：样本选为用户行为的展现日志落地表，正负样本比例约为1：4。学习方法为point + pair-wise。使用MOE结构学习多场景的不同需求（本/异地，周边/全城）， 在expert的输出间计算欧式距离，取负后作为辅助loss，鼓励不同expert学习到不同的分布。使用attention分别对长短周期用户序列建模，包括以下几个方面：
4. 多业务行为建模：高德行为数据可以分为多业务（酒店、美食等）。酒店业务的行为在线计算target attention，高德其他域的用户行为离线计算self-attention后存入feature server，推理时候直接读取即可，用以解决序列超长问题,。
5. 多种行为类型建模：不同的行为（转化、点击等）含义不同，拆开成不同序列会造成数据稀疏。因此采用多行为融合序列，并人工进行序号优化：先按时间窗口（近一周 < 近一月< 近半年 < other）排序，同一时间窗口的行为按（转化 < 到店 < 点击）排序。从序号生成position embedding。同时为行为类型增加type id embedding。
   1. 长周期和实时序列分开建模：实时行为（最近2天）的数量远远小于长周期（2天到1年）行为，实验表明在同一个序列计算attention用造成实时行为被覆盖。因此分成两个序列，并设计了gate net对二者的output进行加权。为防止gate出现极化，对gate的输出增加L2约束。同时在第一个epoch对实时序列的权重增加一个正的bias鼓励模型对实时序列的依赖。
6. **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），分开建模成本大，且小业务效果差。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，设计了分层式的统一模型，第一层包括各个业务子网络（稀疏激活）和一个共享网络，用于学习业务区分和知识迁移，子网络的输入进行了特征选择。第二层是一个CGC多任务网络，防止gate极化，增加 dropout约束。并使用GradNorm平衡多任务loss。样本流拆分为多个场景，交替训练。上线后所有业务均有正收益，美食cvr+3.1%，酒店cvr+2.3%。
7. **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）

专业技能

项目经验

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