基本信息

姓　　名：杨新权　　出生年月：1988.05

电　　话：18811171651　　学　　历：硕士

邮　　箱：starspringcloud@gmail.com　　住　　址：北京市

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)。

2013.07– 2015.3　　奇虎 360　　NLP 算法工程师

1. 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。 多路结果按比例 merge。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。
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 的输出之间计算欧式距离，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：
   a) 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。
   b) 多种行为类型建模：不同的用户行为（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行为类型（转化 < 到店 < 点击）排序。模型从序号生成 position embedding。同时为行为类型增加 type id embedding。

c) 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行为（2 天到 1 年），且实时行为对预测非常重要。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 train 的第一个 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。

4 **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，且小业务样本不足，且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，子网络的输入会进行特征选择。第二层是一个 CGC 多任务学习网络，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，交替训练可以改善训练波动。上线后所有业务均有正收益，其中美食 cvr+3.1%，酒店 cvr+2.3%。

5 **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 top20 作为样本。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）

项目经验

2017.03-2018.03　　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

2015.03-2017.03　　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。
2. 品牌搜：
   a) 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。
   b) 品牌排序：分层排序，优先考虑权威性，其次考虑距离。
3. 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

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