匡盟盟

邮箱：mmkuang@connect.hku.hk | 电话：(+86)132-7015-3586

GitHub：https://github.com/kuangmeng | 主页：https://mkuang.nooa.tech

工作经历

2023.11 至今 | **推荐算法工程师**，DATA-抖音，字节跳动，上海

- **消息推荐全链路算法优化**：（1）建设用户画像、设备画像、内容属性、关闭行为相关的特征，构建 Uplift 模型在效率与关闭损失约束下决定每个 Device 推送的频次。结合高斯时钟建模推送时机与内容画风，提升时机 x 内容的推送体验。（2）从打破用户认知的角度思考，先后上线“抖音热点”、“笔记”、“静音图文”等创意以及 AIGC 文案，收获一致好评。（3）基础排序上（召回/精排），相继完成了多项关键优化，召回上涉及样本流、特征与多目标构建、召回模型结构升级等。精排上挖掘业务特有特征，多场景/多目标建模，加深加宽 CTR 模型等，引领推荐技术和框架升级。

- **结合大模型生成能力的主动消息推荐**：（1）结合大模型生成能力，设计并实施一套包含大模型实时个性化生成推送候选的主动消息推荐链路。（2）使用 Global DeviceID 打通抖音系与豆包系特征体系，利用跨端数据优化低活用户、长尾候选、新增候选的消息推荐效率尤其是新增候选的消息推荐冷启效率。

2021.03 - 2023.11 | **应用研究**，搜索技术中心，微信事业群，腾讯科技，广州

- **大模型在微信搜索系统中的应用**：（1）构建大模型 RLHF 的数据集，基于自研 self-instruction 构建出整套单/多轮对话数据集生成策略。为了使大模型结果更符合事实依据，我们研究了微信龙珠大卡以 Plugins 的形式接入大模型预测的可行性，同时提出 Memory 机制有效缓解了大模型输入上下文长度限制的问题。（2）利用大模型对评论、描述等文本进行关键词/短语提取，丰富文档特征，提升搜索系统效果。

- **账号类在线检索系统的通用 Doc 理解系统**：（1）从零开始搭建微信搜索的数据增强系统。通过对文档标题等数据进行增强，扩大召回范围并提高相关性计算的准确性。（2）对账号文档的描述、认证实体和子页面进行分析，以理解或挖掘出有利于召回和相关性计算的文本。同时，从账号菜单和外部 URL 中挖掘出可用的子服务以提升账号的召回率。

- **文本分类数据的样本优化**：数据质量一直是深度学习模型性能提升的瓶颈。为了提高现有人工标注数据的准确性，我们提出了一种两阶段噪声标签检测的样本优化策略。其作为微信搜索人工标注数据的质检与清洗模块，保证微信标注数据的准确性，为模型质量提供保障。

研究项目

结合大模型实时生成能力的主动消息推荐系统（豆包系 *App*）| *2024.12 - 至今*

豆包系 App 相比与其他内容类 App，有着候选供给不足的难点，同时天然比较易于与大模型交互来个性化生成候选。于是，我们设计并实施了一套大模型 x 消息推荐系统。该系统先由一个基于效率与关闭损失约束的 Uplift 模型来计算每天每人的合理推送频次，再由时机模型在各种信号的触发下确定每次主动消息生成的时机，时机到来时，通过召回用户之前的聊天历史 Bots/历史 Query，经过排序后选择 Top 50，结合用户特征、Bot 特征、Query 特征等由大模型生成并排序一条或多条主动消息/Suggestion，取生成并排序好的 Top 10 候选作为一路召回内容，进入站外通知推送链路，与其他路的召回内容进行混排，最终排出 Top 1 内容推送给用户。该链路在豆包系 App 均已实践，均能拿到 0.5% - 1% 的大盘 LT30 收益。

消息推荐基础排序链路优化（抖音热点、图文、音乐、文娱业务）| *2024.03 - 至今*

结合热点、图文、音乐、文娱业务特点，在召回策略上，我们相继完成了多项关键优化，包括为了解决分时段样本不均衡、Calibration 差异大的问题将召回样本流从 Fast Emit 升级至 Join Window 形式；改进负样本采样方式，如降采样、引入真实负样本；为了解决召回集中问题引入了级联目标召回，并与 CTR 目标召回联合组成 Multi-head 模型形式训练与部署，提高召回模型效率；实践突破双塔模式的向量召回来增加模型建模能力，提高复杂场景建模天花板；为了适配大候选量下召回而升级二向箔召回架构等。在精排迭代上，也结合业务做出不少突破，包括挖掘业务特有精排特征，提升业务在消息分发中的效率；为了更精细地挖掘各业务间的共性和差异性，各业务特征单独进塔（多场景建模）；结合各业务的不同需求，构建 CTCVR 多目标模型（热点小黄条点击目标、图文右滑目标等）；根据 Scaling Laws 指引，借助 TokenMixer 与 MMCN 结构加深加宽 CTR 模型等。在推荐基础排序框架上，结合消息推荐特点（兼顾内容与创意），迭代精排框架到内容 x 创意 2D 联合排序。改动累计 LT30 超 0.1%。

结合高斯时钟法的消息推荐时机模型 | *2023.11 - 2024.03*

时间是推荐系统中的重要影响因素。从后验来看一天内早中晚的大盘点击率不同，整体画风可能也有所差异，而且不同时间的行为可能也会影响内容模型的 Label（例如次留模型、时长模型等）。于是我们提出在抖音系推送时机建模过程中，将时间戳连续值转化成三角函数投影到一个单位圆上。具体实现上，我们把一天（86400 秒）转化为 0-2π 之间角度。考虑时机上 23:59 -› 00:01 有着天壤之别，我们另外将时间做成“线段”，即直接把一天的时机转化为 0-1 之间的数值，在此基础上做一些非线性变化，将这两者与神经网络模型在高层融合，强化内容推荐过程中的时机作用，最终不仅带来了不错的大盘 LT30 收益，而且在推送场景上首次验证出推送关闭率收益。

账号类在线检索系统的通用 *Doc* 理解系统（数据增强） | *2022.08 - 2023.11*

为了提高多数据源检索系统的通用性，以及对文档实体进行更全面、准确的描述，同时为了解决搜索改写难以解决的行业定制化表达的难题，我们提出一种通用数据增强系统。该系统能对多来源、异构数据进行统一增强和理解，并且输出结构统一的干净数据用于检索、排序。该系统采用分层多队列设计，从底往上分别由数据源、数据湖、低层增强策略、高层增强策略、策略融合、后验与知识储备等组成。该系统能综合使用向量检索、抽取模型、生成模型、回译、同义词、知识图谱等策略，从人工词典、搜索 Session、外部数据、文档自身数据等来源中获取高质量的增强数据，也能利用大模型对评论、描述等文本进行关键词/短语提取，丰富文档特征。该增强数据最终会以一定的格式作用于检索系统，提升整合系统的召回率。目前应用了该系统的微信账号类搜索的召回率达到 95% 以上。

文本分类数据样本优化（样本去噪/噪声适应性学习） | *2021.03 - 2023.11*

数据质量一直是深度学习模型性能提升的瓶颈。为了提高现有人工标注数据的准确性，我们提出了一种两阶段噪声标签检测的样本优化策略。首先，我们使用 BERT 在要去噪的数据上训练一个初步分类器，并生成噪声候选集。然后，我们使用剩余的干净候选集样本在该分类器上进行深度训练。接下来，我们通过该分类器对噪声候选集中样本各类别的预测概率，得出针对各类别的置信矩阵，并根据这些矩阵过滤掉每个类别中的不可信样本（即噪声样本）。实验结果表明，该方法能将标签净化率提升至 96% 以上。目前，这种样本优化方法已应用于微信搜索的各类样本清洗任务。

研究论文

[C1] Efficient two-stage label noise reduction for retrieval-based tasks. (WSDM 2022)

[C2] Multi-task learning based Keywords weighted Siamese Model for semantic retrieval. (PAKDD 2023)

[J1] MLProbs: A Data-centric Pipeline for better Multiple Sequence Alignment. (IEEE TCBB)

教育经历

2018.09 - 2020.12 **研究型硕士**，计算机科学，**香港大学**
导师：Hing-fung Ting 教授
毕设：以数据为中心的多序列比对算法研究
研究方向：多序列比对

2014.09 - 2018.06 **工学学士**，计算机科学与技术，**哈尔滨工业大学**
指导教师：赵铁军教授
毕设：面向领域快速移植的高精度汉语分词系统研究
成绩：89.6/100

奖励和证书

2024 Q1 字节跳动 Spot Bouns（个人）
2023.01 腾讯业务突破奖：搜一搜用户增长与商业化规模双突破项目
2021.03 李嘉诚奖学金（提名）& 港大优秀研究毕业生（提名）
2018.09 全额研究型硕士奖学金
2018.06 企业奖学金
2015.11, 2016.11 国家励志奖学金