# 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

- **抖音系 App 消息推荐全链路算法优化**：(1) 挖掘与用户、设备、推送内容属性、关闭行为有关的特征、构建 Uplift 模型预测抖音系 App 推送的频次与关闭敏感度。结合高斯时钟建模推送时机与推送内容类目，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) 在召回策略上，我们相继完成了多项关键优化，包括：将样本流从天级 Fast Emit 升级至小时级 Join；改进负样本采样，如降采样与引入真实负样本；同时，引入了级联目标召回、突破双塔模式的向量召回、二向箔召回以及多头召回模型结构等。(4) 聚焦热点及图文业务，精排迭代上：挖掘业务特有精排特征、业务特征单独进塔、构建 CTCVR 多目标模型、以及借助 TokenMixer 与 MMCN 结构加深加宽 CTR 模型。

- **豆包系 App 结合大模型的主动消息推荐**：(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/推送链路。(2) 使用 global did 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。为了为大模型提供事实依据，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。

- **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模块计算的准确性。(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。

- **账号类搜索项目的关键词加权向量检索**：为了使用户查询能匹配更合适的文档，我们在原有的文本检索基础上引入了一种关键词加权的双塔模型，训练出高质量的向量检索队列，以此扩充召回范围，同时从效率考虑我们采用聚类量化方式线上部署。

## 研究项目

### 豆包系 *App* 结合大模型实时生成的消息推荐系统 | *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，但也有比较容易与大模型交互来个性化生成候选的优势。基于此，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 模型的因果建模模块确定每天每人应该推送的频次，再由时机模型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/聊天历史 Query，经过精排后选择 Top 50 候选，结合用户特征、Bot 特征、Query 特征等由大模型生成一条或多条主动消息/Suggestion（推送文案），这些生成的候选进而作为一路召回内容，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，均能拿到差不多 0.5% - 1% 的大盘 LT30 收益。

### 抖音系 *App* 结合高斯时钟法的消息推荐时机建模 | *2023.11 - 2024.05*

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 0-2$\pi$ 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，最终不仅提升了大盘 LT30，而且在推送上首次验证出关闭率收益。

**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。

**账号类搜索项目的关键词加权向量检索 | *2021.03 - 2021.07***

为了更准确地检索匹配度较高的文档，我们需要精确识别查询和文档中的关键字。为此，我们设计了一种领域自适应的多任务学习模型，通过联合训练一个双塔语义匹配模型和一个关键字识别模型，以精确获取查询与文档间的关联性。双塔语义模型将查询与文档分别编码为语义向量，并在相似度计算阶段将它们结合起来。此外，我们引入了一个关键字识别模型，用于自动检测查询与文档中的关键字权重。我们在公开的 MS MACRO 数据集和微信搜索数据集上进行了实证研究，结果表明，我们的方法优于其他语义检索模型，同时从效率考虑我们采用聚类量化方式线上部署，此框架已成功应用于微信搜索的服务搜索场景。

## 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、机器学习、深度学习

*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