郭辰阳（搜索、大模型算法工程师）

电话：13261539852 | 邮箱：13261539852@163.com | 微信：13261539852

教育背景

2014.07 - 2017.07	中国科学院大学 双一流	计算机系统结构 - 硕士
2010.09 - 2014.07	山东大学 双一流 985 211	软件工程 - 学士

工作经历

2024.06 - 今	北京虾皮信息科技有限公司	资深研发工程师
2017.07 - 2024.05	北京百度时代网络技术有限公司	资深研发工程师

项目经历

2023.09 - 2024.05　企业知识管理平台-百度智能云甄知　搜索和大模型任务调优

- **任务描述**：**甄知-企业知识管理平台**基于**文心大模型**能力，全面重构知识生产、加工、组织、分发、应用各个环节，为企业提供包括**对话式问答**、**对话式搜索**、**智能助手**等知识服务，进一步提升用户获取知识的效率。
- **本人工作**：跟进**搜索**、**RAG问答**、**摘要生成**、**标题生成**、**表头识别**、**表格语义转化**等相关算法和大模型任务。
- **主要方法**：
  - 【解析】基于**百度文心大模型 + paddle ocr**的基础能力，实现**文本 + 视觉 + 多模态**的融合解析方案。
  - 【分段 & 段落知识增强】基于**长度 + 句子分割符**对解析结果进行二次组装，生成检索和问答的基础单元；用大模型对表格数据进行**表名、表头识别**，并进行**表格语义转化**；针对文档段落进行**问答对、标签、摘要**挖掘，以提升召回和精排效果。
  - 【大模型摘要 & 标题生成】基于大模型进行**段落、文档的摘要生成**和**段落、问答session的标题生成**。针对**长文档摘要**场景，调研使用**map-reduce**和**refine**的范式进行优化。
  - 【召回 & 精排】**召回**：引入基于ES8的**hybrid search**，语义表示使用**bge embedding**，在解决语义召回类问题的同时，兼顾字面召回；**精排：**引入**bge-reranker**，提升垂类场景下精排的准确率。
  - 【prompt工程】**扩大上下文拼接范围**：采用拉链策略，将当前段落的后续n个段落拼接作为大模型回答的上下文。**扩大topN数量**：利用大模型更强的理解能力，给搜索更大的容错空间。**prompt调优**：明确场景、人设、思维链提示、prompt格式规范化等。**few-shot**：给大模型示例，明确输入和输出格式。
  - 【模型SFT】toB场景下常用的调优选型，通常使用**全参调优**或者**Lora**两种方法。
- **相关成果**：
  - 【召回 + 精排优化】**文档类型**数据采用bge召回和rerank策略之后**top3包含答案**的比例在**95%**左右，**端到端问答**准确率接近**90%**。
  - 【大模型任务效果】**表格语义转化：表格类问答**准确率提升**5%**（解决精排不准的问题）；**摘要 & 标题生成：**相关任务可用率均在**85%**以上，且长文档摘要结果的信息覆盖度提升较大，**ROUGE提升20%+**。

2020.01 - 今　知识中台-企业搜索　全文搜索负责人

- **任务描述**：**知识中台**是面向**企业知识**应用的全生命周期、一站式、智能解决方案。**企业搜索**是知识中台**应用层面的基础和核心部分**，提供通用的知识发现能力，目标为提升企业用户**获取知识的效率**。
- **本人工作**：负责企搜**整体的工程实现**和**部分策略实现**，提升搜索的**性能**和**效果**，助推百度搜索的私有化落地。

- **主要方法：**
  - 【query理解】完成**切词**、**词权重**、**紧密度**、**词典策略**（专名识别、同义泛化、纠错词、屏蔽词）、**归一**、**意图识别**等策略落地。
  - 【召回】基于ES的**字面召回** -> 落地query理解策略的**匹配树召回** -> 添加语义召回的**hybrid search**（语义引擎选型由faiss迭代为ES8.0）；
  - 【精排】基于**XGBoost**的基础LTR（覆盖字面CQR、CTR及simnet语义特征等）-> 基于**预训练模型Ernie-base**的语义精排。
  - 【重排】基于用户行**点击调权**、可配置的**多因子加权重排**框架。
  - 【片段检索】为了推进更细粒度的内容理解和检索，将**内容检索**升级为**片段检索**，产品层做文档折叠。
  - 【模型调优】基于**paddle sail框架+SimCSE**实现搜索飞轮调优工具，基于**标注数据驱动模型微调**。
  - 【大模型增强】大模型前置生成**query、短语、摘要**等数据，加入索引中，借助大模型强大的理解能力离线理解内容，提升召回和排序的效果。
  - 【性能优化】根据性能需求调整**缓存机制**（内存preload、term_vector预存储、agg全局基数预加载）、合理的**数据压缩**（http 304）、**IO**、**内存分配**、**线程池调整**等，形成系统的性能问题定位和调优方法。
- **相关成果：**
  - 【产品功能】从**0-1**构建整个全文检索通路，支持企搜**20+个版本**的迭代，完成全文检索基础模块、业务模块的功能设计和开发，支撑所有全文类知识的索引和检索，并支持可扩展。
  - 【效果提升】跟进召回和排序效果优化，**召回率提升：**优化后的策略较简单的BM25策略提升**20%+（Hit Ratio(topN)）**。**精排效果提升：**优化后的策略效果较无精排版本提升**30%+（NDCG）**。
  - 【项目收益】工作期间，直接或间接支持**30余个**项目的POC和交付调优工作，整体覆盖**可计收金额近亿元**，直接支持的金额为**5000w+**，多次支持KA项目的效果调优工作，效果得到用户的认可。
  - 【专利】期间撰写专利**9篇**，覆盖召回语义权重、排序、模糊检索等方法。

2017.07 - 2019.12　　**知识图谱应用-百度汉语**　　**图谱数据负责人**

- **任务描述**：**百度汉语**是基于汉语知识图谱数据打造的产品矩阵，其产品覆盖**大搜阿拉丁卡片**、**百度汉语APP**、**小度智能音箱**、**多模少儿搜索** 等业务，同时与**小天才智能手表**等外部厂商合作，做数据和服务输出。
- **本人工作**：作为**汉语图谱数据方向的负责人**，提升汉语图谱的**权威性、全面性、丰富性**，提升应用竞争力。
- **主要方法**：
  - 【汉语图谱构建 & 优化】经过**知识获取**、**知识抽取**、**知识融合**、**知识扩充&校验**、**知识应用**等步骤，完成多个新类目的图谱数据构建，并优化各个环节，扩充和校验图谱数据。
  - 【知识自动扩充 & 校验】**实体、属性扩充：**基于业务发现的新类目扩充、基于大搜query和结果扩充标签（tags-seeker）、未登录词发现（自动化实体挖掘）、基于依存句法分析的自动化短语抽取、基于音频波形切割的诗句音频数据映射；**实体、属性校验：**诗词通用错别字校验，进行基于大搜结果统计的错别字发现方法，结合人工标注进行校验。
  - 【知识应用】基于**Lexparser**实现意图识别引擎，准确识别query意图，分发到汉语的不同资源中，并转化成图检索语句，基于图谱知识准确应答。

- **相关成果：**
  - 【业务支撑】**全面**负责汉语图谱数据工作，支撑多项上游业务，负责期间汉语整体的pv由**4700w扩充到6200w**、百度汉语APP**月活30w+**。
  - 【丰富性、全面性、权威性提升】提升了汉语图谱的**实体覆盖、事实覆盖**。汉语**实体种类扩充23%**，**实体数据量扩充15%**，**属性数扩充13.7%**；**多媒体数据量扩充6.9%**。K12数据**100%校验**，基于错别字发现工具纠正诗词句错误**万余处**。
  - 【自动收录系统】形成**未登录词挖掘**、**标签挖掘**、**短语搭配挖掘**等自动收录流程，天级贡献新实体、新属性**百级别**，热点词可天级别发现和更新。
  - 【应用创新】**教育、知识点图谱**（汉语图谱和语文教育关联）、**音频自动切割**（节省标注成本10w+）、**汉语feed**（基于汉语优质内容、多媒体资源进行知识推荐）
  - 【专利】期间撰写专利**11篇**，覆盖图谱构建方法、短语抽取、音频切割、标签挖掘等方法。

技术栈

- **工程：**
  - 常用编程语言**Java、python**，熟悉linux常见命令；
  - 熟练使用常见的**爬虫框架**（如Scrapy和PySpider），有爬虫使用经验；
  - 熟悉**大数据处理框架**hadoop，有**千万级用户**日志数据处理经历；
  - 熟悉**docker、k8s**等技术，有**容器化编程**和交付经验；
  - 熟悉**开源搜索引擎使用及原理（如Elasticsearch、Solr）**，有效果调优、性能调优经验。
- **策略算法：**
  - 熟悉**大模型理论**和**应用**知识，有**大模型微调**和**应用**经验。
  - 熟悉**搜索策略框架**和常见的搜索算法，有相关的**效果调优**经验；熟悉**向量检索（ANN）、语义表示、语义检索**的常见方法。
  - 熟悉**知识图谱**的构建过程，有**KBQA**及相关图谱应用全流程的应用经验。

奖项

- **获用户/客户至上奖（体系）**1次
- **创新个人**1次、**优秀个人**1次
- **百度小赞**1次、**团队项目小赞**1次
- 参与的某KA项目荣获**中国信通院“星河奖”**，**中国电子标准大数据创新联盟十大案例**