答疑系统开发计划方案

开发计划安排

第一周（4.14-4.20）：数据治理

这一阶段主要处理最新版PPT，结构化的重新构建数据库，保证治理后的数据是优质的，可被模型高效、高精度使用的。

主要工作：

- 识别所有PPT文档，转换为文本
- 尽可能精准的识别公式和理解图像
- 自动化拆解实体、关系
- 建立实体、关系数据库
- 实现基础的数据库增删改查以及查询功能

预期交付：

- 结构化文档集合，包含按页面切分的Markdown文档
- 提取的图像资源（统一存放在attachment文件夹）
- 查询示例

预期交付文件结构目录如下：

```
文件1(文件夹)/
├── 文件1_page_1.md
├── 文件1_page_2.md
├── ...
└── attachment/
    ├── 文件1_page_1_image_1.jpg
    ├── 文件1_page_1_image_2.jpg
    ├── ...
文件2/
├── 文件2_page_1.md
```

```
├── 文件2_page_2.md
├── ...
└── attachment/
    ├── 文件2_page_1_image_1.jpg
    ├── 文件2_page_1_image_2.jpg
    ├── ...
```

第二-三周（25.4.21-5.4）：学魁知识库建设

这两周的工作主要是初步实现知识库所有功能模块，开发知识增强算法并通过大量测试优化知识增强算法。

主要工作：

- 数据高级处理，结合母题、知识点提纲/思维导图构建高级索引表
- 实现知识库的增删改查
- 测试知识库增强问答效果
- 设计召回（引用参考文献）格式及实现
- 实现实体召回与母题召回双模式查询工作流

预期交付：

- 知识库
- 知识库API接口
- 可交互对话demo
- 测试性能

第四周（25.5.5-5.11）：答疑系统工作流开发

这一阶段基于前面实现的功能模块，深入嵌入学魁教学体系，构建多个工作流，基本实现APP功能，API对接前端。
（时间允许的情况下设计后端管理系统，方便教研组直接调试工作流）

主要工作：

- 提示词工程
    - 工作流设计

    - 科目专有提示词设计
    - 教学思维模型提示词合集
- 优化自动分流Agent：用户输入自动分配学科工作流并嵌入适合的思维模型
    - 提示词设计
    - 自动化工作流设计
- 前端APP
    - 设计文档召回展示UI
    - 设计生成教案报告方法（如PDF）及其展示组件（方便答疑老师使用）

预期交付：

- 整体工作流
- 完整版对话API
- 自动化答疑Agent
- 对接前端，实现完整功能

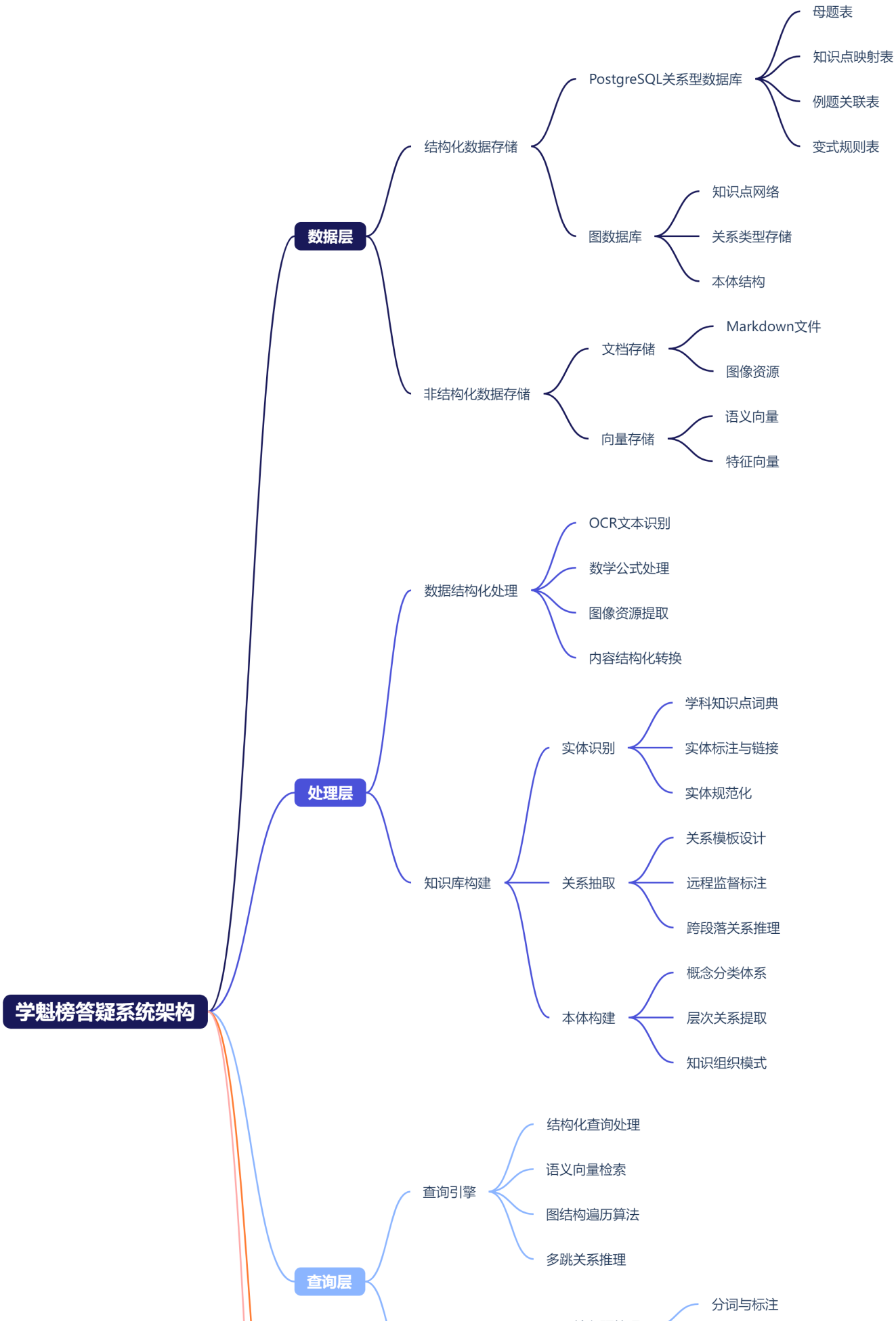
系统详细设计

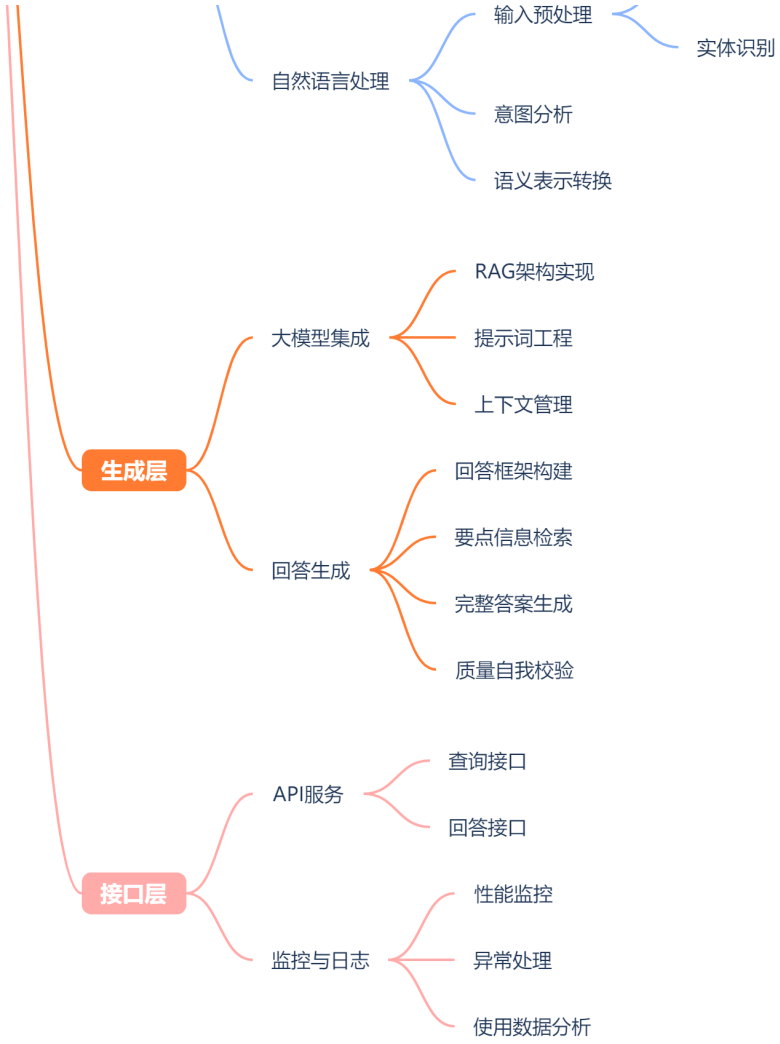
1 系统概述

1.1 系统预期目标

答疑系统旨在构建一个智能问答平台的后端算法和数据库集成系统，该系统的目标是能够对用户输入的问题进行知识点提取、关键词识别，并通过结构化的知识库快速定位到相关例题、母题及变式，同时结合PPT内容进行综合分析，最终提供准确、全面的解答。系统将利用OCR技术、数学公式识别、图/树数据库以及大型语言模型(LLM)等技术实现高效的知识检索与问题解答。

1.2 系统开发架构图


学魁榜答疑系统架构
数据层
结构化数据存储
PostgreSQL关系型数据库
母题表
知识点映射表
例题关联表
变式规则表
图数据库
知识点网络
关系类型存储
本体结构
非结构化数据存储
文档存储
Markdown文件
图像资源
向量存储
语义向量
特征向量
处理层
数据结构化处理
OCR文本识别
数学公式处理
图像资源提取
内容结构化转换
知识库构建
实体识别
学科知识点词典
实体标注与链接
实体规范化
关系抽取
关系模板设计
远程监督标注
跨段落关系推理
本体构建
概念分类体系
层次关系提取
知识组织模式
查询层
查询引擎
结构化查询处理
语义向量检索
图结构遍历算法
多跳关系推理
分词与标注



2 技术实现

2.1 数据结构化处理

本模块通过一套完整的数据处理流水线，将学魁榜现有PPT文件转化为计算机可读的结构化文本。首先通过文档解析器读取PPT文件并按页处理，保持原有逻辑结构；然后利用已完成的OCR模块识别文本内容和数学公式，并开发专用脚本将JSON格式公式转写为Markdown可读文本；同时并行处理PPT中的图表和图片资源，单独保存并建立与文本的关联；最终将识别结果按页生成Markdown文件，并建立多级键值表与提纲对应，形成有序的文件结构体系，为后续知识库构建提供基础数据支持。

2.2 母题知识库构建

本模块构建一个以星型模式为核心的关系型数据库架构，以母题表为中心辐射连接知识点映射表、例题关联表和变式规则表，为每个母题分配唯一标识符并存储学科分类、适用年级和难度系数等元数据信息。通过特征向量化处理，将题目的语义特征转换为高维向量并存储在数据库中，同时记录结构化的解题步骤序列。系统实现基于

GIN索引的向量搜索功能和优化的联合索引，并按学科和年级实现水平分区策略，提升大规模数据查询性能，为大模型提供精准的题目理解和相似度计算基础。

2.3 数据治理与索引构建

本模块构建学魁榜大模型的核心知识库，首先通过构建学科知识点词典、设计正则表达式模式并微调预训练语言模型进行实体识别，同时建立同义词表和实体链接机制实现跨章节的实体共指消解；然后设计知识库关系模板，利用远程监督和依存路径进行关系抽取，应用BERT等模型进行关系分类，重点建立学习序列关系、难度层级关系和概念包含关系；最后通过分析知识点和课程大纲构建本体，提取核心概念分类，实现基于文本的概念层次提取和聚类分析，设计查询转换机制和本体知识注入流程，确保知识能有效支持模型的推理与应答。

2.4 查询与回答逻辑

本模块采用分层架构设计智能查询与回答系统，包括数据层、查询处理层、语义理解层和回答生成层，结合关系型数据库和图数据库的混合存储策略。查询引擎实现双模式查询策略，对结构化查询采用精确匹配，对自然语言输入进行语义向量化匹配，并通过深度优先搜索实现知识点关联遍历。自然语言处理管道包括输入预处理、意图分析和语义表示，将查询转换为标准化表示传递给查询引擎。大型语言模型集成采用RAG架构，设计专门的提示词模板引导模型生成结构化教学导向的回答，并通过多阶段生成策略确保回答质量。