Computer Science and Application 计算机科学与应用, 2025, 15(11), 196-206
Published Online November 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.1511297

Hans汉斯

面向非结构化扫描PDF的大语言模型知识图谱生成框架

刘建民，肖维维

北方工业大学理学院，北京

收稿日期：2025年10月12日；录用日期：2025年11月12日；发布日期：2025年11月25日

摘　要

本文提出一种融合多模态大语言模型与图像增强技术的知识图谱生成框架，旨在解决非结构化扫描版数学教材处理中的三大核心难题：低质量文本识别、数学符号语义解析及知识体系结构化建模。针对扫描PDF图像模糊、公式识别困难等问题，设计多阶段图像增强流程，显著提升文本提取置信度。创新采用双模型协同架构：智谱清言模型负责增强图像的内容提取，DeepSeek模型完成实体关系抽取与三元组构建。通过章节节点动态容器化、游离节点三级消解策略，在Neo4j中实现数学概念的逻辑化存储与可视化关联网络。

关键词

大语言模型，低质量文本识别，知识图谱，Neo4j

A Framework for Generating Knowledge Graphs for Unstructured Scanned PDF Using Large Language Models

Jianmin Liu, Weiwei Xiao

School of Science, North China University of Technology, Beijing

Received: October 12, 2025; accepted: November 12, 2025; published: November 25, 2025

Abstract

This paper proposes a knowledge graph generation framework that integrates multimodal large language models with image enhancement techniques, aiming to address three core challenges in

文章引用: 刘建民, 肖维维. 面向非结构化扫描 PDF 的大语言模型知识图谱生成框架[J]. 计算机科学与应用, 2025, 15(11): 196-206. DOI: 10.12677/csa.2025.1511297

processing unstructured scanned mathematics textbooks: low-quality text recognition, semantic parsing of mathematical symbols, and structured modeling of the knowledge system. To tackle issues such as blurred PDF images and difficulties in formula recognition, a multi-stage image enhancement process is designed, significantly improving the confidence of text extraction. An innovative dual-model collaborative architecture is adopted: the Zhipu Qingyan model is responsible for enhancing content extraction from images, while the DeepSeek model completes entity relationship extraction and triple construction. Through dynamic containerization of chapter nodes and a three-level resolution strategy for free-floating nodes, the logical storage and visual association network of mathematical concepts are realized in Neo4j.

Keywords

Large Language Models, Low-Quality Text Recognition, Knowledge Graph, Neo4j

1. 引言

毕达哥拉斯曾言：“万物皆数”，深刻揭示了数学作为理解世界基石的本质。它不仅是所有学科的基础，更是人类对事物抽象结构与模式进行严谨描述、逻辑推导的通用语言。数学的魅力在于其普适性——其原理和方法能够渗透到现实世界的任何问题之中。在人类文明演进与社会生活的方方面面，数学都扮演着无可替代的角色，更是探索和驾驭现代科学技术不可或缺的核心工具。但值得注意的是，所有的数学对象，无论是数字、函数还是空间，本质上都是人类心智精心构建的抽象定义。

然而，数学因其高度的抽象性与严密的逻辑性，往往给学习者的理解带来一定的困难与挑战。因此，如何有效地将这些抽象、晦涩难懂的数学概念进行具象化呈现，降低认知门槛，提升学习效率与深度理解，成为数学教育与研究领域持续关注的核心议题。传统教学手段在应对这一挑战时，常显得心有余而力不足，亟需引入创新的技术与方法。

近年来，大语言模型(Large Language Model)作为一种新兴的技术受到广泛关注，由于出色的自然语言处理能力，其在各种任务例如文本生成、语言对话和实体识别中都能取得优秀的结果。相较于传统的光学字符识别(Optical Character Recognition)，部分 LLM 具有精度高、可以根据要求对内容进行选择性提取等特点，能够适应更复杂的环境。本文主要是通过 LLM 与知识图谱结合的方式处理非结构化扫描 PDF 的《数学分析》教材。

2. 知识图谱简介

知识图谱(Knowledge Graph)于 2012 年由谷歌公司提出并快速发展，自诞生起受到各行各业的广泛关注。它是一种以结构化形式组织和表达知识的语义网络，其核心逻辑是通过“实体－关系－实体”的三元组(如<爱因斯坦，提出，相对论>)构建关联网络：实体作为节点代表具体事物或抽象概念(人物、地点、事件、理论等)，关系作为边描述实体间的语义联系(属性、类别、作用、因果等)。这种图结构将碎片化信息转化为具有逻辑关联的知识体系，使机器能够理解数据背后的含义，并让原本孤立的知识点建立起联系，使得抽象复杂的关系可视化、清晰化。

知识图谱的核心价值在于实现认知层面的推理与应用。它打破了传统数据库的表格局限，通过图遍历支持多跳查询(例如追溯“爱因斯坦的导师的学术影响”)，揭示隐含关联；同时为智能系统提供可解释

的结构化知识支撑，广泛应用于搜索引擎(直接呈现答案卡片)、推荐系统(基于知识关联扩展推荐路径)、金融风控(识别复杂股权网络)和问答系统(利用自然语言提问并快速获得准确回答)等领域。作为连接人类知识与机器智能的桥梁，知识图谱已成为推动语义理解与决策智能的关键基础设施。

在 LLM 大规模应用和普及前，对于知识图谱的构建，主要依赖人工标注数据，采用无监督方法(如 TF-IDF、TextRank 等)，或监督学习方法(如 CRF、LSTM 等)或半自动方法并结合外部知识库(如 DBpedia)进行实体对齐[1]；张蓉[2]提出自底向上的电力 PDF 表格知识图谱构建方法，结合 Tabula 抽取技术、单元格规则语言规范化表格，并通过 SPARQL 映射到本体，最终生成包含 13,400 个三元组的 RDF 图谱；Brian Walsh 等[3]提出 BioDBLinker 模块，从生物医学文本中构建知识网络。在 LLM 大规模普及和应用之后，也有学者尝试将 LLM 和知识图谱进行结合，如李念强等[4]结合 AI 虚拟实验(模拟电磁场分布)，降低实验成本；文怡[5]为解决大语言模型在船舶问答中的幻觉问题，提出多智能体知识图谱增强检索生成框架(MA-KGERAG)；吴金红等[6]聚焦专利 SAO 三元组提取，提出结合 DeepSeek-Chat 和知识图谱的方法；时宗彬等[7]针对有机电池材料，采用本地部署 LLM (WizardLM-70b)和提示工程实现无微调信息抽取，允许模型返回“None”减少幻觉；蔡子杰等[8]利用 ChatGPT 生成样本实例，通过指令设计和数据增强构建包含 5641 条指令的心理健康联合信息抽取数据集，覆盖命名实体识别、关系抽取和事件抽取三项任务；杨建梁等[9]通过 LLM 文本抽样人工复核的方式对 OCR 处理文本进行校对，采用无监督的实体识别方法，通过对大语言模型进行指令提示，指令模型从文献中抽取人物、时间、地点等实体及其关系，建立了基于 LLM 的红色档案资源交互系统；Chen 等人[10]将 LLM 与图神经网络结合进行知识补全。然而，这些方法主要针对特定的科学领域或者通用领域，大多数数据来源是结构化 PDF、开源数据等，而针对质量较差的非结构化扫描 PDF 数据源现主要采用的是 OCR 识别，对于大量数学公式与中文文本来说，此种方法会导致提取失败或乱码的情况发生。与现有技术相比，本文框架的特色在于：(1) 专门针对低质量扫描数学教材设计；(2) 采用双模型协同架构，分别优化内容提取和知识抽取；(3) 引入多级图像增强和游离节点处理机制。与传统方法相比，本文方法减少了人工特征工程的需求；与其他 LLM 方案相比，本方法更注重数学领域的特殊挑战，如公式处理和概念层次构建。

3. 大语言模型简介

近几年，ChatGPT、GPT-4 [11]等大语言模型横空出世，通过堆叠层数、扩大参数规模(从百万级到万亿级)，让大模型突然“开窍”，也让人工智能技术成为各国各领域关注和发展的焦点。智谱清言、豆包、DeepSeek 等国产大模型也相继问世，特别是 DeepSeek [12]采用了比 GPT 等大模型更轻量化且高效的设计，达到了 OpenAI 的水平。LLM 在多个任务中表现卓越，在军事、科研、学习等领域大放异彩。

3.1. DeepSeek-Chat 模型

DeepSeek 是由中国顶尖 AI 团队深度求索自主研发的通用大语言模型体系，其研发始于 2023 年，致力于突破认知智能的边界。作为国内首个全面对标 GPT-4 技术架构的 AI 大模型，DeepSeek 涵盖从 7B 到超千亿参数的完整模型系列，在数学推理、代码生成、多轮对话等核心能力上达到国际领先水平。目前已衍生出 DeepSeek-R1、DeepSeek-V2、DeepSeek-V3 等多个版本，广泛应用于智能客服、教育辅助、金融分析等领域。

其中 DeepSeek-Chat (即 DeepSeek-V3 [13])模型主要为用户提供友好便捷的文字交流界面，旨在通过自然语言处理技术实现人机对话交互功能。能够理解并回应用户的日常咨询或简单指令，提供更加流畅的人机对话体验，理解上下文并维持连贯的多轮次交流。

3.2. 智谱清言 GLM-4.1V-Thinking-FlashX 模型

智谱清言是由北京智谱华章科技有限公司推出的生成式 AI 助手，于 2023 年 8 月 31 日正式上线。其公司下面有多种不同模型，可以应对不同场景、需求，如文本生成、视频图片生成和识别、智能问答等。

其中 GLM-4.1V-Thinking [14]是一种视觉语言模型，主要用于推进通用多模态理解和推理。其主要采用带有课程抽样的强化学习，以增强在解决 STEM 问题、视频理解、内容识别、编码、基于 GUI 的代理和长文档理解等任务上的能力。其中作为开源模型的 GLM-4.1V-9B-Thinking，在长文档理解和 STEM 推理等具有挑战性的任务上比闭源模型 GPT-4.1o 等具有更好的效果。

4. 知识图谱的搭建

由于非结构化扫描 PDF 的页面属性是图像，且绝大多数图像上的文本信息是比较模糊或者低对比度字体，在进行图像增强的基础上，使用 PyMuPDF、OCR (EasyOCR, Tesseract OCR)等工具进行文本内容提取，效果很差，如在使用 Tesseract OCR 时置信度[15]为 69.93%~84.38%，平均置信度为 78.54%，甚至部分页面提取到的内容(尤其是公式)是乱码，而且对中文文本识别能力差，其中字符错误率(CER)为 84.43%，词错误率(WER)为 94.29%，对后续的三元关系提取和处理带来了极大的干扰。

因此，在图像增强后，本文利用智谱清言的 Glm-4.1V-Thinking-Flashx 模型提取文本内容，置信度为 75.52%~90.62%，平均置信为 85.19%，CER 为 7.63%，WER 为 9.8%。不使用增强置信度为 68.49%~88.56%，平均置信为 81.56%，CER 为 40.91%，WER 为 47.76%。并可以去除文本中的例题、习题、解、证等内容，以免无关内容对后续实体和关系的提取产生干扰，然后使用 DeepSeek 的 DeepSeek-Chat 模型进行实体和关系的提取以及知识图谱的构建，流程如图 1 所示。

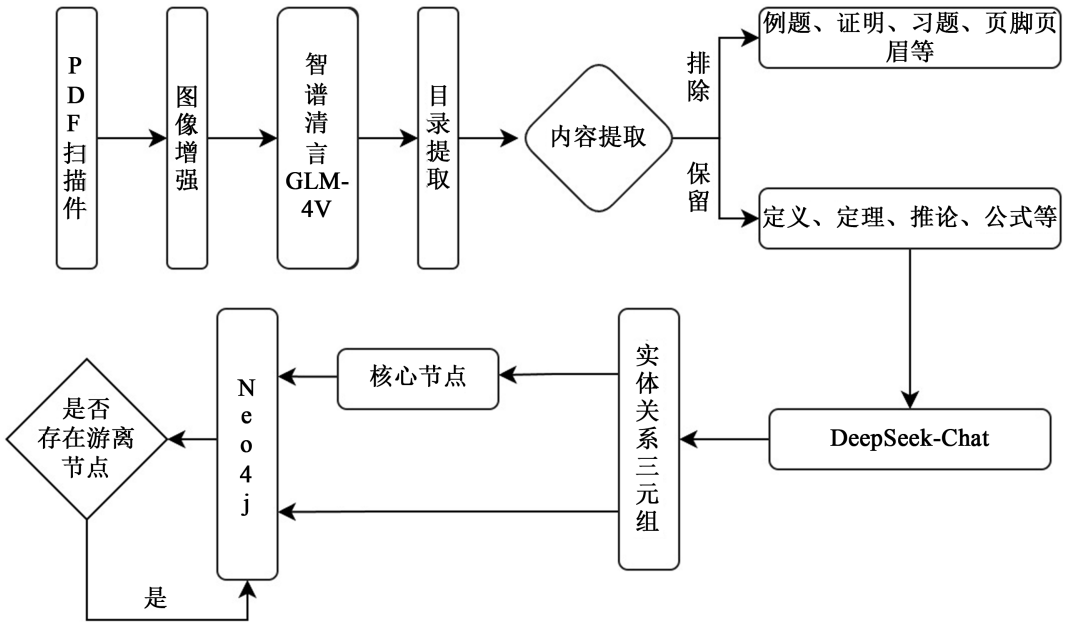

Figure 1. Knowledge graph construction structure
图 1. 知识图谱构建结构

4.1. 图像增强

针对扫描版 PDF 中普遍存在的图像模糊、字体对比度低等问题，本文设计了多阶段的图像增强流程。首先通过灰度转换消除色彩干扰，采用像素值反转技术，解决对比度低字体辨识困难的问题；接着应用

自适应直方图均衡化(CLAHE)增强局部对比度；实施伽马校正优化亮度分布；基于拉普拉斯方差检测智能判断模糊区域，仅对需要区域应用锐化卷积核；最后通过中值滤波降噪(3 × 3 核)和三级微调完成优化，使得字体更清楚，与背景对比度更强，使得文本提取、公式识别更加准确。其具体算法如下：

```
# 图像增强方法
函数 类增强图像(输入图像):
    尝试:
        # 基础转换
        将图像转为灰度图
        转换为 numpy 数组
        # 颜色反转(针对浅色文字)
        反转图像颜色值(255 - 当前值)
        # CLAHE 对比度增强
        创建 CLAHE 增强器(参数: clipLimit=3.0, tileGridSize=8x8)
        应用 CLAHE 增强
        # Gamma 校正
        Gamma 值 = 1.8
        应用 Gamma 校正公式: 输出 = (输入/255)^Gamma × 255
        # 降噪处理
        应用中值滤波(核大小=3)
        # PIL 图像增强
        转换为 PIL 图像
        亮度增强(增强因子 1.8)
        对比度增强(增强因子 3.0)
        锐度增强(增强因子 2.5)
        返回增强后的图像
    异常处理:
        打印错误信息
        返回原始图像
# 模糊检测函数
函数 图像模糊检测(图像数组, 阈值=100):
        计算图像的拉普拉斯变换
        计算拉普拉斯值的方差
        返回 方差 < 阈值  # 方差越小越模糊
```

4.2. 目录和文本内容提取

在程序运行后，将手动输入正文起始页码(正文起始页在 PDF 中的页码)和目录起始页码，并自动计算偏移量(在 PDF 中的页码和实际页面标注的页码之间的差值)，智谱清言大模型提取目录页范围后，自动识别并提取章节信息(章节编号、标题、起始页码)，处理非章节条目(如附录、索引等)，通过提取到的章节信息和偏移量计算每一章的起止页码，并根据第一个非章节条目的页码或总页数来确定最后一章的结束页码。

根据每一章的起止页码，使用 PyMuPDF 提取相应的页面，进行图像增强处理，然后智谱清言大模型进行图像识别和文本提取的处理，在这过程中将严格排除例题、习题、解、证、页脚等内容，从而来提高文本质量，其中提示词如图 2 所示，并将抽取的内容生成为 JSON 文件，方便后续的内容提取和处理。

```
prompt = """
        你是一位数学教材分析专家，正在分析数学教材中的内容。

        ## 特别注意：
        1. 提取范围：全页除以下内容外的所有内容：
           - 以"例"开头的段落（包括例1、例2等）
           - 以"证"开头的内容（最后以符号□结尾）
           - 以"解"开头的内容（最后以符号□结尾）
           - 习题中的内容（以习题开头，题的开头是（1.，2.，……）的形式，以下一节标题结束
           - 以"分析"开头的段落
           - 以"注"开头的段落（包括注1、注2、注3等）
           - 第X章标题（如"第9章 定积分"）
           - 小节标题中带*号的内容（如果用户选择忽略）
           - 页脚页眉
           - 以"规定"开头的内容（包括规定1、规定2等）
        2. 包含所有数学公式，使用LaTeX格式
        3. 保留原始文本的完整数学表述
        4. 联系上一页结尾的内容检查，确保没有误忽略或忽略不完全
        5. 如果当前页包含章标题，则不用联系上一页结尾的内容检查
        6. 必须保留定理、性质、定义、推论
```

Figure 2. Page content extraction prompt words
图 2. 页面内容提取提示词

4.3. 提取实体和关系内容

由于经过 API 时文本容量有限制，首先需要将每一章的提取内容进行分块，本文采取的是根据提取到的文本内容按 4 页分块，并且除本章的第一块外，后续分块都包含上一块的最后一页，以保持语义文本的连贯性。然后用 DeepSeek 根据提示词要求对每一章的文本进行实体提取和关系生成，为了避免出现某一概念在不同章节出现而创建出同一名称的实体，导致知识图谱结构混乱，本文采取只将第一次遇到某一概念创建为实体(如确界定理在“第一章 实数集与函数”和“第七章 实数的完备性”中都有提及)，后续再出现这一概念时，将页码等信息重新整理到创建的节点属性中。

其中知识实体节点通过幂等创建确保唯一性，采取标准化对节点名称进行处理(如某些节点名称识别为“定理 1.1”等)，节点核心属性包括：(1) 标准化名称，(2) 类型标记(如概念、定理、定义、公式、性质、推论等)，(3) 核心标识(is_core)，(4) 颜色编码(按类型分配)，(5) 页码列表，(6) 游离节点标识(is_orphan)。每生成一个实体节点将会给出相应节点的详细描述，并基于重要性、在标题中的出现频率、数学体系中的地位等标准筛选出本章中的 3~5 个核心节点。实体关系构建采用名称精确匹配策略来匹配源和目标实体，通过预定义关系类型转换表来保证语义一致性，根据提示词中预先设定的关系(如定义、推导、术语、公式等)来确定两两节点之间的关系，并给出关系的详细说明。将以上内容汇总整理后，最终输出三元组数据的 JSON 文件，包含实体描述、关系语义及页码定位信息，形成完整的知识单元，以便后续的操作处理。其中实体与关系的精确率为 74.45%，召回率为 84.68%，F1 分数为 79.23%。图 3 为提取实体与关系时的提示词。

```
prompt = f"""
        你是一名数学分析专家，正在构建精确的数学知识图谱。请严格遵循以下规则从文本中提取实体和关系：

        ## 实体提取规则：
        1. 每个实体必须有明确的数学定义或描述
        2. 实体名称必须使用标准数学术语（例如，不要使用"定理1"，而应使用"确界原理"这样的标准名称）
        3. 如果遇到编号定理（如定理1.1），请根据内容赋予其标准名称
        4. 包括以下类型的实体：
           - 核心概念（如"实数"、"函数"）
           - 定理（如"确界原理"）
           - 定义（如"上确界定义"）
           - 公式（如"三角不等式"）
           - 性质（如"有界性"）
           - 推论（如"推论1.1"）
        5. 严格忽略以下内容：
           - 未定义的非标准术语
           - 次要的、一次性的概念
           - 类似于函数的表达形式（如解析法、列表法等）的概念
           - 特定问题的解决方法
        6. 优先提取在多个上下文中出现的实体
        7. 为每个实体提供清晰完整的描述（20-50字）
        8. 每个实体必须至少参与一个关系

        ## 关系提取规则：
        1. 只提取有明确数学逻辑的关系
        2. 关系必须连接两个重要实体
        3. 优先提取核心概念之间的关系
        4. 为每个关系提供清晰完整的描述（20-50字）
```

Figure 3. Entity and relationship prompt words
图 3. 实体与关系提示词

4.4. 知识图谱构建

本文基于 Neo4j 图数据库实现知识结构化存储。首先创建章节点作为逻辑容器，包含编号、标题、教材版本属性，核心节点直接与章节点相连，搭建起本章知识图谱的基本框架，使其更加方便查找、修改、读取以及理清某一章的整体逻辑，其余节点则作为核心节点的子节点、孙子节点，来进一步丰富知识图谱的完整性。

针对游离节点问题，本文主要采取三级解决方案：(1) 关系创建时动态更新节点状态；(2) 基于文本嵌入的余弦相似度计算关联核心概念；(3) 建立章节点兜底机制。最终构建的知识图谱包含完整的概念层级体系和语义关系网络，为数学知识推理奠定基础。其兜底机制代码如下：

```
# 游离节点兜底机制
函数 处理游离节点兜底():
    # 获取所有游离节点
    游离节点 = [实体 for 实体 in 所有实体 if 实体.is_orphan == True]
    1)相似度匹配
      函数 相似度匹配(节点列表):
          成功节点 = []
```

```
        核心概念 = 获取核心概念()
        # 预计算核心概念嵌入向量
        核心嵌入映射 = {}
        for 核心 in 核心概念:
            核心嵌入映射[核心.名称] = 获取文本嵌入(核心.名称 + 核心.描述)
        for 节点 in 节点列表:
            if 节点.是核心概念: continue
            节点嵌入 = 获取文本嵌入(节点.名称 + 节点.描述)
            最佳匹配, 相似度 = 计算最佳匹配(节点嵌入, 核心嵌入映射)
            if 相似度 >= 0.45:  # 相似度阈值
                创建关系(节点.名称, 最佳匹配, "属于")
                节点.is_orphan = False
                成功节点.append(节点)
        return 节点列表 - 成功节点
    剩余节点 = 相似度匹配(游离节点)
    2)章节点连接
      函数 章节点连接(节点列表):
        章节点 = f"第{当前章节号}章"
        for 节点 in 节点列表:
            创建关系(节点.名称, 章节点, "属于")
            节点.is_orphan = False
        return []
    3)特殊处理：定理节点优先
      函数 定理节点处理(节点列表):
        定理节点 = [节点 for 节点 in 节点列表 if "定理" in 节点.名称]
        for 定理 in 定理节点:
                相关核心 = 关键词匹配(定理.描述, 获取核心概念())
            if 相关核心:
                创建关系(定理.名称, 相关核心, "属于")
            else:
                创建关系(定理.名称, f"第{当前章节号}章", "属于")
            定理.is_orphan = False
```

图 4 为生成的数学分析知识图谱的部分展示。

4.5. 案例

为了具体说明本框架的处理能力，在此以《数学分析》中“确界原理”为例，展示从原始扫描图像到知识图谱节点的完整转化过程。

(1) 如图 5 所示是包含“定理 1.1 (确界原理)”的模糊扫描页面，在通过图像增强后，由智谱清言模型完整提取出文本内容；

(2) 通过 DeepSeek 模型，识别出“确界原理”这一实体，并标记为核心节点，记录第一次及后续出

现的页码，并生成“任何非空有上界的数集必有上确界，任何非空有下界的数集必有下确界”进行描述，并与最近章节出现的“上确界”、“下确界”建立起“上/下确界由确界原理定义”的关系；

(3) 后续章节中，在遇到“柯西收敛准则”、“单调有界定理”、“区间套定理”、“有限覆盖定理”、“聚点定理”的实体后，则构建出“柯西收敛准则证明确界定理”、“确界定理证明单调有界定理”、“单调有界定理证明区间套定理”、“区间套定理证明有限覆盖定理”、“有限覆盖定理证明聚点定理”、“聚点定理证明柯西收敛准则”的闭环关系，符合实数完备性六个定理互相证明的逻辑。

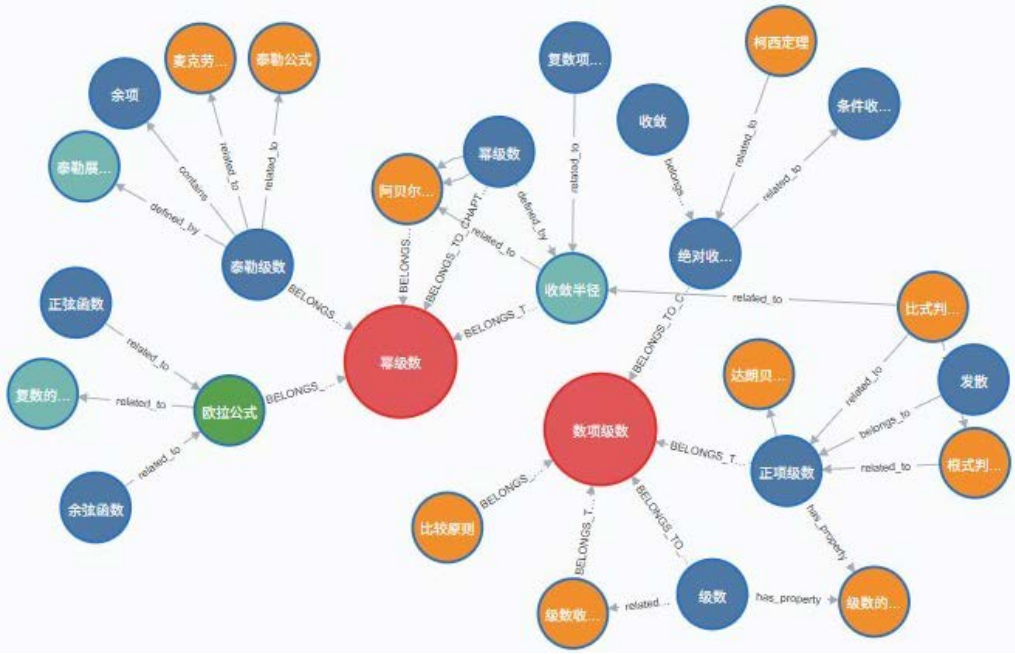

Figure 4. Mathematical analysis knowledge graph
图 4. 数学分析知识图谱

定理 1.1(确界原理) 设 S 为非空数集.若 S 有上界,则 S 必有上确界;若 S 有下界,则 S 必有下确界.

证 我们只证明关于上确界的结论,后一结论可类似地证明.

为叙述方便,不妨设 S 含有非负数.由于 S 有上界,故可找到非负整数 n,使得

1) 对于任何 $x \in S$,有 $x<n+1$;

2) 存在 $a_0 \in S$,使 $a_0 \geqslant n$.

对半开区间 $[n,n+1)$ 作 10 等分,分点为 $n.1, n.2, \cdots, n.9$,则存在 $0,1,2,\cdots,9$ 中的一个数 n_1,使得

Figure 5. The initial text of the supremum and infimum principle
图 5. 确界原理的初始文本

5. 错误分析

(1) 在目录提取中，有时会出现将小节标题识别为章标题，或者将非章节标题识别为章标题，可能原因有“第 x 章”与小节标题的序号都是使用中文数字一、二、三等，导致模型识别时出现混乱，图像增强无法完全解决字体大小、排版相似的识别问题，智谱清言模型对二级标题的理解不准确，未完全排除非章节部分，特别是在目录换页后未识别出非章节标题；

(2) 在实体过程中，会出现如“定理 1.1”的名称，但在经过标准化后有时出现没有重新命名的情况，

成为游离节点且无法被连接，可能原因有正则化后 DeepSeek 未能从节点描述中总结出该节点名称；

(3) 有的节点及关系描述很短，起不到说明作用，可能原因有网络波动导致提取的文本不够完整，DeepSeek API 的 max_tokens = 4000 可能限制了详细描述的生成，温度参数 temperature = 0.2 设置较低，可能导致描述过于保守和简短，缺乏重新生成或补充的机制；

(4) 有些游离节点在经过兜底处理后会与其他游离节点相连接，但不与章节点、核心节点及其子节点相连接，可能原因有游离节点之间可能因为描述中的常见词而产生虚假相似度，关键词的匹配无法理解深层的数学逻辑关系，嵌入模型可能无法准确捕捉数学概念间的语义关系。

6. LLM 在知识图谱搭建中的优势和劣势

6.1. 优势

(1) 相较于传统的对 PDF 文件处理和图像识别(特别是 OCR)的方法，LLM 对文本的提取具有速度快、精度高的特点，对于质量较差的非结构化扫描 PDF，在经过图像增强之后仍有较高的精度，特别是对公式和数学符号上的识别，可以将识别到的公式转化为 LaTex 格式，使其更加准确和易读。通过提示词工程，系统能够有效过滤例题、习题等干扰内容，提升后续实体关系抽取的质量；

(2) 在处理效率上，本框架能够在数小时内完成整本教材的初步知识提取，相比完全人工标注显著提升了处理速度，且避免了因人工疲劳导致的系统性遗漏；

(3) 数学的知识网络具有强依赖性，人工标注易受线性阅读限制，难以全局追踪跨章节概念关联，而且容易出现知识点和关系遗漏疏忽的情况。而系统可以通过自注意力机制能够识别部分跨章节的概念联系，如自动建立实数完备性不同表述形式间的关联网络，辅助构建相对完整的知识体系；

(4) 系统展现出一定的层次化结构建模能力，能够区分核心概念与次要概念，为后续知识推理提供基础；

(5) 框架设计具备一定的通用性，可扩展至不同数学教材的处理，为跨教材知识融合提供了技术基础。

6.2. 劣势

(1) 大型语言模型 API 的稳定性高度依赖网络传输质量，网络波动可能导致关键处理环节的功能性失效。在文本内容解析、实体抽取及关系生成的执行过程中，网络中断或延迟易引发数据包丢失，进而造成输出结果的信息缺失、冗余噪声干扰或结构化格式崩坏，最终破坏知识图谱构建流程的可靠性；

(2) 由于大语言模型的幻视等问题的存在，在识别文本内容中，有时会将需要排除的内容也进行提取，导致文本质量降低，使得提取目录、实体、关系时出现问题，例如将目录中非章节标题的“积分表”中内容识别为章节标题，导致提取的目录存在偏差和误导。有时会提取一些无关痛痒的知识甚至是乱码作为实体，或者是当前章节没有的内容识别为实体，例如在第一章时将目录上的章节标题当作是第一章的实体进行提取，导致其他大量的节点无法匹配成为游离节点，从而污染整个知识图谱的构建；

(3) 由于在上传到 API 时文本容量的限制，只能通过分块处理，在这过程中会损失一部分上下文文本的逻辑信息，特别是数学的知识体系具有严密的连续性与依赖性，当文本被机械分割为独立块时，跨块的关键逻辑链条将被强制截断，导致 LLM 无法获取完整的演绎语境，使生成的三元组面临语义完整性风险；

(4) 对于 LLM 的 API 提示词需要格外谨慎。由于在通过 API 调用 LLM 时需要提示词来告诉 LLM 当前需求、注意事项等信息，若提示词宽松，将出现输出非标准化实体、混淆数学概念、提取大量无用内容。若约束过强，又可能引发模型认知僵化，无法把握所有需求，导致关键实体漏提取，甚至是一个实体都无法提取，使得知识图谱的构建出现严重问题。

7. 结论与展望

本文构建的数学知识图谱系统，通过创新性融合多模态大模型与图像增强技术，主要解决了扫描版数学教材处理的三大难题：(1) 低质量图像的字符识别，(2) 数学符号的语义解析，(3) 知识体系的结构化建模。系统采用的双阶段处理架构(GLM-4V 内容提取 + DeepSeek 知识抽取)，其性能和效率相较于传统 OCR 方案和人工标注方案具有一定的优势，其层次化实体关系模型有效捕捉了数学概念的逻辑关联，为智能教育应用提供了帮助。

然而，本文模型由于采用无监督方式，因此在建立完知识图谱之后仍需要人工校验，在未来可以构建基于知识一致性的动态校验系统，通过数学公理库等方式实时检测实体矛盾，结合上下文感知技术自动修正描述偏差，提升知识图谱的逻辑严谨性。开发多模态融合算法，集成版面分析、字体特征聚类和语义理解，实现全自动目录解析，消除当前人工输入起始页的交互瓶颈。使用更智能的模型，研发数学专业领域微调模型，强化对复杂推理链的语义解析，进一步提高知识图谱建立的质量。

基金项目

北京市高等教育学会 2024 年立项面上课题(课题编号 MS2024219)。

参考文献

[1] 王俊彦, 罗剑. 课程知识图谱技术及应用综述[J]. 计算机时代, 2025(3): 30-35. https://doi.org/10.16644/j.cnki.cn33-1094/tp.2025.03.007

[2] 张蓉. 面向电力 PDF 文档中表格的知识图谱构建技术研究[D]: [硕士学位论文]. 兰州: 西北师范大学, 2023.

[3] Walsh, B., Mohamed, S.K. and Nováček, V. (2020) BioKG: A Knowledge Graph for Relational Learning on Biological Data. *Proceedings of the* 29*th ACM International Conference on Information & Knowledge Management*, 19-23 October 2020, 3173-3180. https://doi.org/10.1145/3340531.3412776

[4] 李念强, 周沛, 黄于, 等. 结合人工智能与知识图谱的智慧创新课程探索——以电磁场与电磁波课程为例[J]. 高教学刊, 2025, 11(12): 76-79. https://doi.org/10.19980/j.CN23-1593/G4.2025.12.019

[5] 文怡. 基于知识图谱与大模型的船舶问答系统研究与应用[D]: [硕士学位论文]. 成都: 电子科技大学, 2025. https://doi.org/10.27005/d.cnki.gdzku.2025.002224

[6] 吴金红, 任晓露, 张欣妍, 等. 基于大语言模型与知识图谱的专利 SAO 三元组提取方法研究[J/OL]. 情报杂志, 1-9. https://link.cnki.net/urlid/61.1167.G3.20250728.1530.004, 2025-08-07.

[7] 时宗彬, 朱丽雅, 乐小虬. 基于本地大语言模型和提示工程的材料信息抽取方法研究[J]. 数据分析与知识发现, 2024, 8(7): 23-31.

[8] 蔡子杰, 方荟, 刘建华, 等. 基于大型语言模型指令微调的心理健康领域联合信息抽取[J]. 中文信息学报, 2024, 38(8): 112-127.

[9] 杨建梁, 王一多, 黄美雯, 等. 基于大语言模型的红色档案资源交互式知识发现研究——以《南方局党史资料大事记》为例[J]. 图书情报工作, 2025, 69(15): 112-123. https://doi.org/10.13266/j.issn.0252-3116.2025.15.010

[10] Liu, Y., Cao, Y., Lin, X., *et al.* (2025) Enhancing Large Language Model for Knowledge Graph Completion via Structure-Aware Alignment-Tuning. arXiv: 2509.01166

[11] Achiam, J., Adler, S., Agarwal, S., *et al.* (2023) Gpt-4 Technical Report. arXiv: 2303.08774

[12] Guo, D., Yang, D., Zhang, H., *et al.* (2025) Deepseek-r1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arxiv: 2501.12948

[13] Liu, A., Feng, B., Xue, B., *et al.* (2024) Deepseek-v3 Technical Report. arXiv: 2412.19437

[14] Hong, W., Yu, W,. Gu, X., *et al.* (2025) Glm-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning. arXiv: 2507.01006

[15] 伍凌辉, 马聪, 周玉, 等. 融入置信度的文本图像翻译研究[J]. 中文信息学报, 2024, 38(12): 64-73.