基于图检索增强生成（GraphRAG）技术的

审计定性依据查询系统研究

【摘 要】生成式人工智能技术的迅速发展为内部审计效率和质量的提升带来了新的机遇。本文针对传统检索增强生成（Retrieval-Augmented Generation, RAG）方法构建的知识库问答系统存在的检索内容不完整、信息不够丰富等问题，提出了一种基于图检索增强生成（GraphRAG）技术的审计定性依据查询系统。该系统通过Neo4j图数据库构建审计知识图谱，并结合GraphRAG和大语言模型进行信息检索和内容生成，实现了基于源引用的查询响应和清晰的依据条款解释。实验结果表明，GraphRAG技术能够显著提高大语言模型在审计定性依据检索场景的关联性和准确性。本研究探索了内部审计数智化转型的实践方案，为人工智能技术在审计领域的深化应用提供了参考。

【关键词】审计定性依据 图检索增强生成 大语言模型 多实体语义嵌入 多路径检索

Research on Qualitative Audit Evidence Retrieval System Based on  
 Graph Retrieval-Augmented Generation (GraphRAG) Technology

**Abstract:** The rapid development of generative AI presents new opportunities to enhance the efficiency and quality of internal auditing. This paper introduces a qualitative audit basis query system leveraging Graph Retrieval-Augmented Generation (GraphRAG) technology, addressing issues such as incomplete retrieval and limited information richness in traditional Retrieval-Augmented Generation (RAG)-based knowledge systems. The system constructs an audit knowledge graph using the Neo4j graph database and integrates GraphRAG with large language models for information retrieval and content generation, delivering query responses with source citations and clear clause explanations. Experimental results demonstrate that GraphRAG significantly improves the relevance and accuracy of large language models in qualitative audit basis retrieval. This research explores a practical approach to the digital and intelligent transformation of internal auditing, providing valuable insights for the deeper application of AI in the audit domain.

**Keywords:** Audit qualitative basis, Graph Retrieval-Augmented Generation, Large Language Models, Multi-entity semantic embedding, Multi-path retrieval

一、引言

内部审计作为企业内部监督控制的重要抓手，必须不断提升履职能力，以适应数字化变革中出现的新需求。在网络强国、数字中国、大数据战略等国家政策的指导下，中央企业正在加快数字化转型，消除“数据孤岛”和“数据烟囱”，开展大数据与人工智能、物联网等技术融合应用。这一变革使信息载体从纸质变为电子文件，数据量剧增，数据关系复杂且标准不一，同时信息安全风险加大，这对内部审计方法和效率提出了新要求和新挑战。在推动内部审计工作数字化、智能化发展过程中，强化审计定性依据的规范性与适用性，是确保审计监督质量和效果的重要环节。作为审计工作的基本遵循，审计定性依据的科学分类和有效运用是提升审计效率和规范化水平的关键，而正确选择审计定性依据更是内部审计机构有效开展审计工作的基础。审计定性依据通常来源于具有法律效力的国家法律法规、监管规定，以及行业准则和企业内部规章制度等规范性文件（金昌铉等，2023），传统的依据查询主要依赖于人工翻阅资料或基于关键词的文本检索，存在效率低下、结果不准确等问题，难以满足现代内部审计的需求。为解决这一问题，学术界和业界开始探索将人工智能技术应用于审计知识检索领域，RAG技术在审计知识智能检索方面显示出潜力。然而，现有研究主要集中在传统RAG方法的应用，而这种方法对于复杂审计知识结构和多层次查询需求的处理存在较大的局限性。鉴于此，本文对业界现有审计知识智能检索方案进行优化创新，提出通过GraphRAG技术来构建审计定性依据查询系统，以提高审计依据查询的便捷性和准确性，帮助审计人员更好地理解审计问题及其定性依据，从而更好地服务于审计监督职能的有效履行。

二、研究概述

（一）业界研究概述

生成式人工智能（Generative Artificial Intelligence，GenAI）技术在内部审计领域的应用引发学术界和业界广泛关注。多位学者深入研究了人工智能在内部审计中的应用，探讨了GenAI技术在辅助编制审计计划、审计数据分析、自动化审计流程、审计知识检索等领域的应用价值（易冰心，2023；张春芝等，2024），并基于GenAI技术探索审计知识问答助手的构建和运用（张菲菲、韩磊，2024）。例如，通过GenAI技术自动化数据分析和模式识别，协助审计人员快速辨识潜在风险和问题；利用大语言模型（Large Language Model，LLM）根据预设规则和标准自动生成审计报告，对关键信息内容进行辅助校对。

在GenAI技术的企业应用落地方面，当前研究较为广泛的领域是利用RAG技术对企业沉淀数据、领域知识进行智能检索及智能问答（刘锦，2024；向小伟等，2024）。RAG技术由Lewis等人（2020）提出，是一种利用特定和私有数据集回答用户查询的技术，无需对大语言模型进行进一步训练。该技术结合向量检索和大语言模型，使系统能够从海量数据、知识库中迅速提取相关信息，并生成连贯且有原始数据支撑的回答（Zhao等，2023）。即使对传统搜索系统进行简单优化，也能显著提升企业知识的利用效率和问答准确性。在内部审计领域，部分研究学者通过FastGPT、Chato等AI开发框架对审计知识库智能检索进行研究和实践（刘锦，2024；陈亚盛等，2024），利用RAG开发引擎实现审计报告、法律法规、审计定性依据等知识问答助手的构建。以审计定性依据库为例，这个过程通常分为3个阶段：首先，将审计定性依据条目汇编为审计指南或参考文档，通过分块（Chunking）将长文本分割成较小的块（Chunk）以适应嵌入模型（Embedding Model）的上下文长度限制；其次，将每个块通过嵌入模型转换为用于检索的嵌入向量（Text Embedding），并在专有向量数据库或支持向量扩展的传统数据库中建立向量索引；在问答阶段，通过检索与查询语句向量表示最相关的块召回文本片段，将其作为上下文输入给LLM，利用LLM的总结归纳能力生成问答结果。基于长文本分块并进行文本块召回的RAG技术，以其实现相对简便、部署灵活等特点，在审计知识智能检索场景下展现出一定应用潜力。然而，随着审计工作复杂性提升及对知识检索精度要求日益提高，传统RAG方法在实际应用中逐渐暴露局限性。特别是在审计问题定性的认定中，往往需要可靠的依据来源以及准确的措辞表达，传统RAG方法在这一方面尚存在差距，难以满足高精度审计定性依据检索需求。

（二）传统RAG方案的局限性

1. 关联信息分割，导致答案质量下降

为了适应嵌入模型的上下文长度限制，将大量有内部结构层次的问题定性依据条目分割成固定长度的片段可能导致相关内容断开连接，从而产生不完整的答案。例如，一个在开头描述问题、在结尾提供依据条款的问题定性依据条目可能在文本分割过程中被拆分，导致定性依据的关键部分被遗漏。尽管文本块重叠（Chunk Overlap）、滑动窗口分块（Sliding Window Chunking）等技术的应用能够增强文本块的语义完整性、召回更关联的上下文信息（Zhong等，2024），但受限于文本块的大小和计算资源的限制，这些技术仍然难以完全避免信息割裂带来的答案质量下降的问题。

2. 未构建实体关系，难以实现多跳查询

审计定性依据指引等参考文档具有复杂的内部信息结构，其中往往包含多个层级的实体关系，如“[审计问题C](定性依据是)[文件条款B]”“[文件条款B](属于)[依据文件A]”“[审计内容D](包含)[审计问题C]”等。传统RAG方案在将文档拆分为固定长度文本块时，未能有效建立这些实体之间的关联，导致对于多跳查询的支持能力不足（Tang等，2024）。例如，用户可能会提出类似“与工程项目应招未招相似的招投标问题还有哪些？”的问题，而传统RAG仅能从相似文本块中检索到部分信息，难以通过局部信息推理出完整答案。虽然可以通过多轮查询设计或链式推理部分解决多跳查询问题，但这类方案增加了系统复杂度，且每次查询的结果都需要进一步处理和重新检索，降低了系统的效率和准确性。

3. 缺乏动态查询规划，难以实现精准检索

尽管传统RAG方案可通过查询重写（Query Rewriting）和查询拆分（Multi-Query）等技术（Ma等，2023）拓展信息检索范围，但核心仍局限于基于文本语义相似度的静态检索机制，无法根据查询复杂度和语义结构动态调整检索策略。面对需要关系聚合或全局分析的查询场景，如“关于提前确认收入事项，有哪些表现形式？”“根据国家招投标法，招标代理机构应具备哪些条件？”等问题，传统RAG方案难以确定最优查询路径，因为无法像知识图谱那样利用结构化的实体关系网络，通过精确查询路由快速定位关键信息节点。这种局限性不仅影响了检索精准度，也限制系统在处理需要多维度信息综合查询时的灵活性。

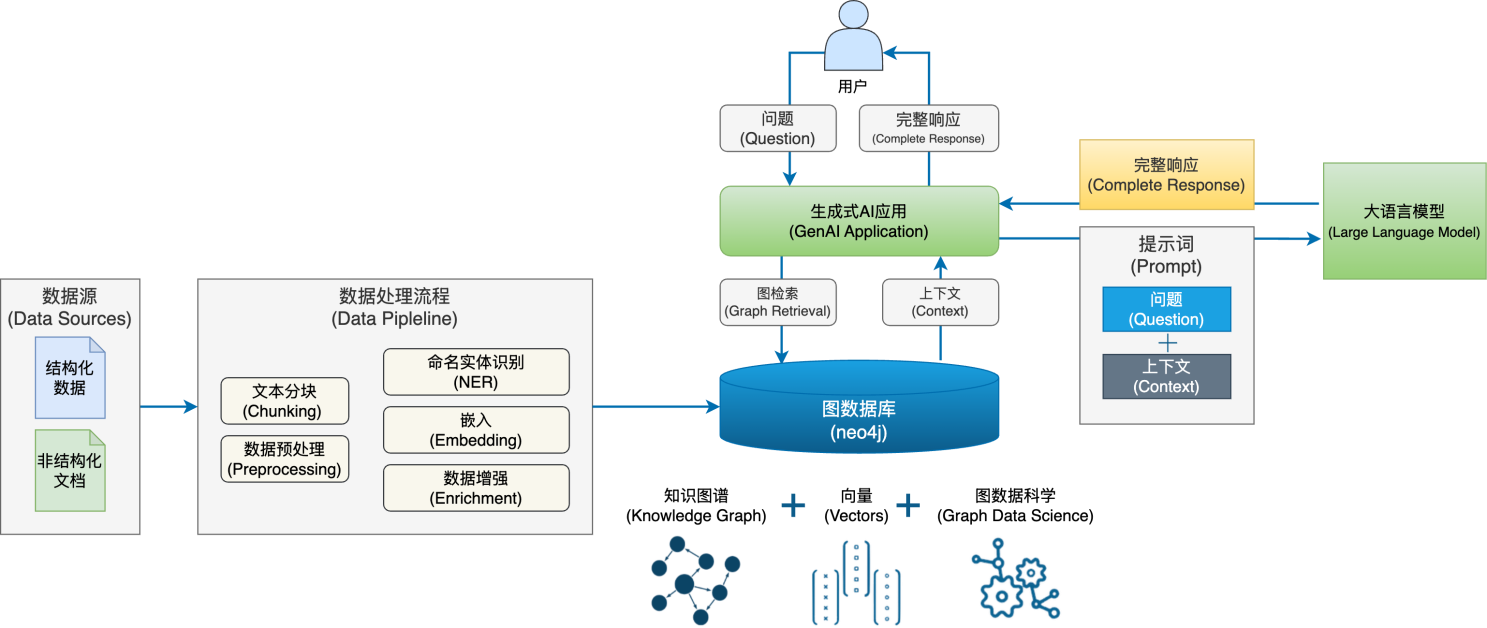
三、本文研究方案探讨

（一）GraphRAG技术介绍

由于传统RAG技术原本就是被设计用于在特定文本区域内找到必要答案，其难以从包含共同属性的关联信息片段中挖掘出关联信息，在对大型数据集或长文档中的总结性语义概念进行理解的任务表现也不尽如人意。为了解决这些局限性，Hu等人（2024）提出了GraphRAG，它是一个利用外部结构化知识图谱来提升语言模型（LM）的上下文理解能力并生成更丰富响应的技术框架。GraphRAG将知识图谱技术与传统的检索增强生成（RAG）系统相融合，从知识图谱中检索最相关的知识，从而增强下游任务的答案质量。GraphRAG可以被视为RAG的一个新的分支（Peng等，2024），其核心理念在于通过在RAG工作流程中引入知识图谱查询机制，充分利用图数据所蕴含的丰富语义关系和结构化信息，显著增强检索结果的质量和相关性。

GraphRAG系统的基本架构可以参考图1。

图1. GraphRAG系统基本架构



根据现有研究（Peng等，2024；Rathle，2024），GraphRAG的工作流程包含4个关键阶段：

1. 知识图谱构建：将结构化或非结构化知识数据转化为知识图谱，明确定义并建立实体、关系及其属性，形成一个相互连接且语义丰富的知识网络。

2. 图数据语义嵌入：针对知识图谱中的实体、关系或更大范围结构，利用专用文本嵌入模型将其语义信息进行向量嵌入并附加到图中，为后续的向量检索和语义分析提供数学基础。

3. 知识检索召回：结合向量空间的相似度搜索与图数据库的结构化查询，实现多维度、多路径的知识检索，以获取最相关、最全面的信息片段。

4. 上下文增强生成：基于检索到的丰富的上下文信息，利用大型语言模型进行深度语义理解和推理，生成准确、全面且逻辑连贯的回答，同时保持与原始查询的高度相关性。

GraphRAG通过将知识图谱与查询相结合，能够更好地识别和连接相关信息片段，将结构信息作为超越文本本身的额外知识，弥补了传统RAG技术在处理复杂语义关系时的不足，展现出更智能和更强大的综合检索能力（Xu等，2024）。通过将关联关系丰富的审计知识数据转化为结构清晰、关系明确的知识图谱，基于GraphRAG方案的审计定性依据查询系统能够精准地捕捉和利用各种概念实体（如审计问题、文件条款、依据文件等）之间的多维关联。这种方法可以显著提高信息检索的准确性和全面性，有效提升审计工作的质量和效率。

（二）审计定性依据查询系统构建思路

基于对GraphRAG技术原理的认识，结合审计定性依据查询场景的任务需求，本研究将从以下3个方面着手构建审计定性依据查询系统：

1. 结构化知识构建

利用图数据库技术构建审计定性依据指引的知识图谱，将审计问题、文件条款、依据文件等实体及其关系进行结构化存储。这种方法能够保留审计定性依据知识条目各部分的逻辑连贯性，有效解决关联信息分割的问题，确保提供完整且高质量的回答。

具体实现中，本文采用Neo4j图数据库来构建审计定性依据知识图谱。Neo4j是一种高性能的图数据库，使用属性图模型来存储和管理数据（Rathle，2024）。其核心元素包括节点（代表实体）、关系（表示节点间连接）和属性（附加于节点和关系的信息）。这种结构非常适合表示复杂的审计定性依据知识网络，能有效捕捉概念间的多维关系。Neo4j的图形结构直观表达实体间的网状关系，便于理解和分析。它支持Cypher查询语言，能高效执行复杂查询操作，如最短路径搜索和模式匹配，对审计定性依据的快速检索和关联分析至关重要。此外，Neo4j的图形处理引擎确保了数据的快速访问和处理能力，使系统能高效处理大规模审计数据和复杂查询请求。

2. 多实体语义嵌入

构建文本向量索引是RAG系统的关键步骤，它直接影响查询检索的性能和效果。在GraphRAG方案中，微软的研究提出了通过层次化构建实体关系图、社区识别和生成社区报告，创建多层次语义向量索引进行语义检索及图增强生成（Darren，2024）。“社区”被定义为知识图谱中语义或结构上紧密相连的实体和关系群组，构成相对独立的知识单元。大语言模型（LLM）被用于生成这些社区的综合性摘要，以捕获广泛的语义关系和主题结构信息。

这种方法能提供全面的背景信息，特别适合处理需要考虑广泛上下文的全局性问题。由于在索引过程中构建了多层次信息结构，它在处理大规模、多样化知识图谱时表现优异，能够提供全面、可靠且有据可查的回答。然而，这种方法也存在索引过程耗时长、计算成本高等问题，在数据集规模显著增大时，可能面临扩展性和稳定性的挑战。

为增强GraphRAG系统在审计定性依据等专业领域的检索可靠性，本文提出了多实体语义嵌入（Multi-Entity Semantic Embedding）方法，直接对多个实体的关键属性文本信息进行向量化，从而提供多个语义匹配候选目标。其主要优势是在前置查询路由确定检索实体后，能购实现更加精确的信息召回和关联分析。同时，该方法非常适用于处理语义信息密集、篇幅较短（通常在数百字符以内）的专业文本，如法律法规条款、制度文本条款、审计问题摘要等。

在审计定性依据查询系统中采用此方法，主要基于以下考虑：①审计定性依据知识图谱中的实体属性（如问题描述、条款内容、文件名称等）通常语义丰富且简洁；②直接对这些属性进行向量化可更精确地捕捉实体语义特征，提高检索准确性；③避免了传统RAG方案中文本块分割可能导致的上下文丢失问题；④无需传统RAG方案中复杂的分块及嵌入调优步骤，简化了语义向量索引过程。

文本嵌入模型选择上，本文使用阿里巴巴开源的Alibaba-NLP/gte-Qwen2-7B-instruct模型。该模型嵌入维度为3584，支持32k tokens的上下文，在多个中英文文本嵌入基准测试中取得最优效果（SOTA）（Li等，2023）。通过文本语义嵌入，将生成的向量作为一个新的属性附加到相应实体节点上，并在Neo4j中建立向量索引，实现多维度的语义向量检索。这种方法可以结合图数据库的关联节点检索机制，使系统能够实现多层次、复杂关系的查询，显著提升对复杂语义概念的捕捉和定位能力。

3. 多路径检索设计

设计多路径检索方案，实现复杂查询问题的高效准确匹配。这种方法通过定向选择查询实体和路径，提高检索的相关性。多路径检索的核心原理是利用LLM的函数调用（Function Calling）能力（Ge，2023）和提示工程（Prompt Engineering）技术（Chen等，2023）实现LLM与其他系统组件之间的结构化信息交互，使系统能够根据预定义的函数说明识别查询意图，调用和执行特定的查询函数。通过构造包含多个预定义查询实体、输出结果的函数集合，以及迭代调优后的提示词模板，系统能将用户的自然语言查询精确转化为知识图谱查询操作，根据问题类型动态选择最合适的查询路径，提供准确度高、关联性强的结果。这种方法不仅可以提高复杂语义查询的精确度和效率，还可以增强系统的灵活性和可扩展性。

在查询路径方面，系统通过LLM分析查询意图并定位目标实体、确定查询函数及参数，在知识图谱中进行关键属性的语义向量匹配和关联信息抽取。在查询模式方面，系统能自动判断查询语句是单跳查询（Single-hop Query）还是多跳查询（Multi-hop Query）类型。单跳查询直接针对特定实体进行检索，如通过审计问题描述快速获取定性依据信息；多跳查询用于处理需要在关联实体间进行跳跃式推理的复杂问题，如查询特定审计问题的关联问题等。通过查询路径和查询模式识别，系统能够准确调用相应函数执行查询，实现对多样化查询的准确响应。

例如，对于“国资委60号文对境外审计的要求有哪些？”这样的问题，系统处理流程如下：①LLM理解用户查询意图，匹配查询路径为“查询特定依据文件的内容详情”，确定目标实体、查询函数和参数。②调用相应查询函数，通过Neo4j的Cypher查询语言执行向量检索，寻找最匹配的依据文件信息、条款内容等。③LLM分析检索结果，根据用户查询语句形成流畅、清晰的回答，包括解释、总结或进一步分析建议。

通过结构化知识构建、多实体语义嵌入和多路径检索规划3个关键步骤，本文将构建一个能充分利用图数据结构优势、实现高效语义检索的GraphRAG系统。

四、系统构建与开发

（一）实验环境与工具

本研究使用Langchain等开发框架及工具进行审计定性依据查询系统的构建，表1列出了系统使用的主要工具和模型，以及它们的具体功能。

表1. GraphRAG系统开发工具

|  |  |  |
| --- | --- | --- |
| 系统组件 | 工具/模型 | 功能 |
| GraphRAG系统 开发框架 | Langchain | AI应用开发框架，可以将多个组件（如向量数据库、图数据库、大语言模型）链接起来，实现知识驱动的增强生成，提供多种链式结构，便于进行知识检索、查询和生成。 |
| RAG系统构建平台 (作为参考基准) | FastGPT | AI应用开发平台，可以快速构建和部署基于LLM的RAG系统，集成了数据导入、向量数据库、文本分割、文档索引等功能，快速搭建知识库问答应用。 |
| 图数据库 | Neo4j | 用于存储图数据结构，存储实体和它们的关系，适合关系密集型数据管理和复杂查询。 |
| LLM推理框架 | Ollama | 负责LLM的推理执行和计算高效性，通过该框架实现高效的LLM推理，提供计算资源管理。 |
| AI模型API后端 | One-API | 提供统一的人工智能模型推理接口，连接AI模型并进行API调用，为模型推理提供了一致的API接口，便于前后端交互。 |
| 函数调用模型 | Qwen/Qwen2.5- 32B-Instruct | 用于查询路径规划，通过分析用户查询意图，输出查询函数名及参数，确保任务能够合理分配和执行。 |
| 对话/评估模型 | Qwen/Qwen2.5- 32B-Instruct | 用于信息总结和任务评估，提取从Neo4j和其他数据源中获取的信息，并生成简洁明了的总结，可用于生成摘要、报告和其他总结性的信息。 |
| 嵌入模型 (Embedding Model) | Alibaba-NLP/gte-Qwen2- 7B-instruct | 将文本转换为向量表示，以便在向量数据库或Neo4j中进行检索，嵌入模型在知识检索、关系匹配等任务中扮演重要角色。 |

注：文本生成和嵌入模型均由RTX4090 GPU推理计算

（二）检索系统设计

1. 知识图谱构建

定义审计问题、依据文件、文件条款、审计重点、审计内容等核心实体及其关系。通过实体间的映射关系，设计Cypher语句将预处理后的审计定性依据数据集导入Neo4j图数据库，构建审计知识图谱。

2. 知识语义嵌入

利用文本嵌入模型，将审计问题描述、文件条款内容、依据文件名称等关键文本属性转换为稠密向量表示。这些向量随后存储于Neo4j图数据库中，并建立相应的向量索引，为后续的相似性检索提供基础。

3. 检索增强生成

a) 查询意图解析：利用LLM的函数调用能力解析用户查询语句意图，确定目标实体和查询路径。

b) 查询语义向量化：使用与索引构建相同的嵌入模型将关键查询语句转换为向量表示。

c) 向量检索及召回：在Neo4j中执行向量检索，基于余弦相似度匹配最相关的实体及其关联信息。

d) 上下文融合与回复生成：将检索结果与用户查询结合作为上下文，输入LLM生成结构化且信息丰富的答案。

（三）系统构建

1. 系统功能框架

（1）知识图谱数据库：利用Neo4j构建审计定性依据知识图谱，包含审计问题、依据文件、文件条款、审计重点、审计内容等5个实体及其关联关系。为实体关键信息构建语义嵌入向量并建立索引。

（2）图向量检索服务：通过函数调用模型识别用户意图，执行预设Cypher查询语句进行向量查询，获取余弦相似度最高的k个实体（Top k）及其关联实体和属性数据。

（3）用户查询界面：支持自然语言输入，通过模板提示词将检索数据作为上下文，与用户查询一起提交给LLM生成回复。

2. 数据处理与知识图谱构建

（1）数据预处理：收集企业内部历史审计问题数据，按照业务域、审计内容、问题表现形式、定性依据、定性依据来源、审计意见等梳理数据，进行清洗、去重、信息补充、文件溯源、数据格式化等处理，使用LLM从定性依据中提取文件名称、文号、条款内容，经过人工复核清洗填充后，完成数据准备，形成审计定性依据数据集。

（2）实体及关系识别：基于预处理数据，定义实体属性（详见表2）及实体间关联关系（详见表3）。

表2. 实体及其属性

|  |  |  |  |
| --- | --- | --- | --- |
| 实体 | 属性 | 数据类型 | 示例数据 |
| 审计问题 | 问题定性 | 字符串 | 违规进行重大决策 |
| 问题描述 | 字符串 | 可行性研究在经济性、技术性等方面论证不充分，尽职调查走过场，存在重大疏漏等，导致实际投资效果未达到预期目标 |
| 是否金额类 | 布尔值 | false |
| 问题类别 | 字符串 | 履行重大经济决策 |
| 审计内容 | 字符串 | 重大投资（融资）项目 |
| 审计意见 | 字符串 | 严格执行“三重一大”集体决策程序规定，落实重大经济事项可行性研究论证程序，进行风险评估和合法合规性审查，确保决策合法合规。 |
| 问题描述 向量 | 数组 | [-0.029405003,0.0015058697,0.009459399,0.006633437,0.0021966973,0.007366527,0.0047365995,-0.016367959,-0.0048824917,-0.039434113,0.008279605,0.01467344… Show all] |
| 依据文件 | 文件名称 | 字符串 | 关于进一步推进国有企业贯彻落实“三重一大”决策制度的意见 |
| 文号 | 字符串 | 文号: 中办发〔2010〕17号 |
| 生效日期 | 日期型 | 2010-07-15 |
| 发布机构 | 字符串 | 中共中央办公厅 |
| 文件类别 | 字符串 | 政策文件 |
| 文件链接 | 字符串 | [https://www.gov.cn/jrzg/2010-07/15/content\_1655395.htm](https://www.gov.cn/jrzg/2010-07/15/content_1655395.htm" \o "https://www.gov.cn/jrzg/2010-07/15/content_1655395.htm) |
| 依据文件 向量 | 字符串 | [-0.01066704,-0.0061481055,0.00215083,0.01763684,-0.030529637,-0.022221943,-0.008346996,-0.0037971435,-0.012326519,0.017748537,-0.00248009,0.018583614,… Show all] |
| 文件条款 | 文件名称 | 字符串 | 文件名称: 关于进一步推进国有企业贯彻落实“三重一大”决策制度的意见 |
| 条款内容 | 字符串 | （七）“三重一大”事项提交会议集体决策前应当认真调查研究，经过必要的研究论证程序，充分吸收各方面意见。重大投资和工程建设项目，应当事先充分听取有关专家的意见。…… |
| 条款内容 向量 | 数组 | [-0.017118981,-0.008738745,-0.0015949388,0.027031474,-0.0150722815,-0.011608815,0.015934316,-0.009181158,-0.01136993,-0.009512999,0.00886081,0.01260081… Show all] |
| 审计重点 | 名称 | 字符串 | 履行重大经济决策 |
| 审计 内容 | 描述 | 字符串 | 重大投资（融资）项目 |
| 审计内容 向量 | 字符串 | [-0.008067916,-0.007763208,-0.006193832,0.015535411,0.0070490604,0.00008365559,0.0012068517,0.00071189745,-0.000047866914,-0.029735465,-0.0065630996,0.… Show all] |

表3. 实体间关联关系

|  |  |  |
| --- | --- | --- |
| 实体A | 关联关系 | 实体B |
| 审计问题 | 定性依据 | 文件条款 |
| 文件条款 | 属于 | 依据文件 |
| 审计内容 | 包含 | 审计问题 |
| 审计重点 | 包含 | 审计内容 |
| 审计重点 | 包含 | 审计问题 |

（3）图数据导入：设计Cypher查询语句，将处理后的数据导入Neo4j数据库，构建完整的审计知识图谱。

3. 向量索引构建与查询逻辑设计

（1）向量索引构建：为审计问题、依据文件、文件条款、审计内容等实体创建向量索引。使用嵌入模型生成文本嵌入向量，并将其存储为相应实体的属性。

（2）查询处理流程：a) 用户输入自然语言查询；b) 将查询转换为向量形式；c) 在知识图谱中检索相似实体及关联信息；d) 将检索结果与原始查询作为上下文提交给LLM；e) LLM生成结构化回复，包括问题定性、描述、依据条款、所属文件和审计建议。

五、查询验证与结果讨论

为了评估信息检索能力和回答质量，本文设计了一个综合查询方案来验证GraphRAG系统的效果，这个方案涵盖了简单查询、多跳查询及涉及不同实体的复杂查询。在评估方法上，本文采用机器评估结合人类专家评估的比较式评估方法（Pair-wise Comparison），从内容相关性、回复完整性和答案逻辑性3个维度全方位对比GraphRAG方案和基于FastGPT的传统RAG方案的检索效果。

（一）查询设计

本研究从知识库中随机抽取工程项目管理、招投标、财务管理等业务领域的10条审计定性依据数据，利用LLM生成语义相近的10个查询语句（详见表4）。

为了确保实验可比性，GraphRAG方案和基于FastGPT的传统RAG方案（作为参考基准）采用相同的文本嵌入模型（Alibaba-NLP/gte-Qwen2-7B-instruct）和总结回复模型（Qwen/Qwen2.5-32B-Instruct）。

表4. 10个查询语句

|  |  |  |
| --- | --- | --- |
| 查询 编号 | 查询语句 | 查询类型 |
| 1 | 大额物资采购项目应招未招，有哪些参考定性依据？ | 简单查询 |
| 2 | 预算管理方面的有哪些问题？具体描述其详情。 | 简单查询 |
| 3 | 请帮我查询经营性结余资金处置的相关规定。 | 简单查询 |
| 4 | 与供应商评价事项相关的审计问题有哪些？ | 多跳查询 |
| 5 | 工程监理单位违规转包其承包的监理业务，类似的问题还有什么？ | 多跳查询 |
| 6 | 未按项目批准范围及内容实施，例如超规模或超预算建设，与其相关的问题有哪些？ | 多跳查询 |
| 7 | 投资项目可行性研究、尽职调查方面有哪些要求，请提供依据文件。 | 复杂查询 |
| 8 | 境外投资财务管理办法中针对境外投资项目连续多年亏损事项处理是如何规定的？ | 复杂查询 |
| 9 | 请帮我查询国家招标投标法对于招标代理机构资格条件的相关规定。 | 复杂查询 |
| 10 | 金融衍生业务审计中，容易出现哪些典型问题，具体表现形式是什么？ | 复杂查询 |

（二）效果评估

1. 评估流程

首先，使用Qwen/Qwen2.5-32B-Instruct模型作为机器评估基准模型，评估两种方案的查询效果。具体评估流程如下：

（1）分别通过GraphRAG方案和传统RAG方案对预设问题集进行回答；

（2）将查询语句、原始参考数据以及2种方案的回答结果同时提交给评估模型。为确保评估结果的可重复性和一致性，Temperature参数设置为0，以消除随机性影响；

（3）指示LLM从内容相关性、回复完整性、答案逻辑性三个维度进行定性评估，判定优势方案。

然后，邀请3位内部审计业务专家参与评估，弥补机器评估可能存在的局限性。专家评估采用与LLM评估相同的多维度评判标准，判定优势方案。

最后，将LLM评估结果与业务专家评估结果进行综合，得出每个维度的优势方案。

2. 评估示例

图2至图4分别展示了机器模型和人类专家对3个综合查询结果的评估示例。

图2. 简单查询



图3. 多跳查询



图4. 复杂查询



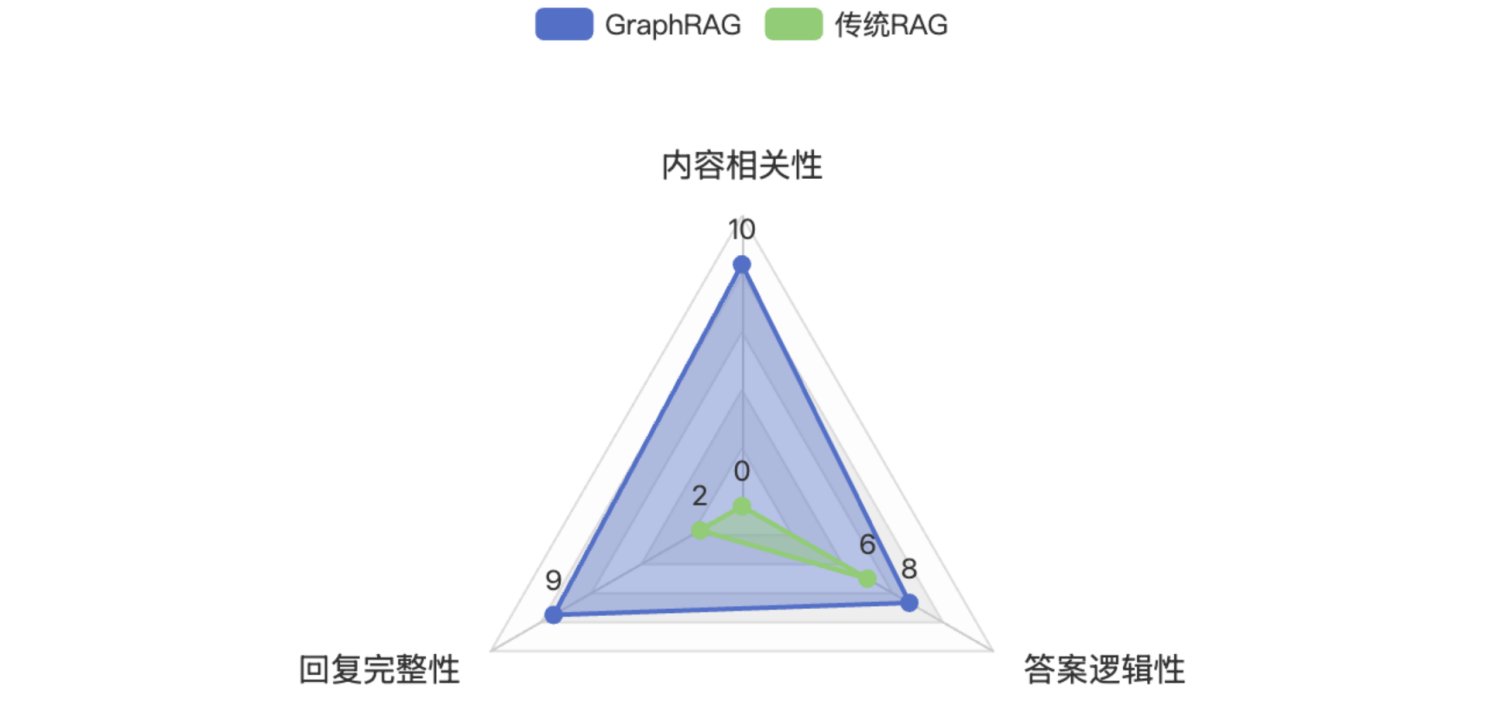
3. 评估结果

在对2种方案的查询问答进行多维度综合评估后，得出评估统计结果（详见表5）和查询性能对比情况（详见图5）。

表5. 评估统计结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 查询编号 | 内容相关性 | 回复完整性 | 答案逻辑性 | 查询类型 |
| 1 | GraphRAG | 传统RAG | 传统RAG | 简单查询 |
| 2 | GraphRAG | GraphRAG | 表现相近 | 简单查询 |
| 3 | GraphRAG | GraphRAG | GraphRAG | 简单查询 |
| 4 | GraphRAG | GraphRAG | GraphRAG | 多跳查询 |
| 5 | GraphRAG | 表现相近 | 表现相近 | 多跳查询 |
| 6 | GraphRAG | GraphRAG | 表现相近 | 多跳查询 |
| 7 | GraphRAG | GraphRAG | GraphRAG | 复杂查询 |
| 8 | GraphRAG | GraphRAG | 表现相近 | 复杂查询 |
| 9 | GraphRAG | GraphRAG | GraphRAG | 复杂查询 |
| 10 | GraphRAG | GraphRAG | 传统RAG | 复杂查询 |

图5. 查询性能对比



评估结果显示，在处理综合查询任务时，GraphRAG方案相比传统方案具有明显的优越性：

（1）内容相关性：GraphRAG方案在所有查询中都表现出明显优势，无论是简单、多跳还是复杂查询。这说明GraphRAG系统能够更准确地定位相关信息，特别是在处理复杂的多实体关系时。

（2）回复完整性：在10个查询中，GraphRAG方案在8个查询中表现更好，1个查询无明显差距，仅在1个简单查询中表现不如传统RAG。这显示出GraphRAG系统通常能够提供更全面的回答，尤其是在需要整合多个信息源的多跳和复杂查询中。

（3）答案逻辑性：在10个查询中，GraphRAG方案在4个查询中表现更好，4个查询中与传统RAG表现相近，仅在2个查询（1个简单查询和1个复杂查询）中表现不如传统RAG。这表明GraphRAG系统在处理需要多步推理的复杂问题时，通常能够提供更有逻辑性的回答，但在某些简单查询中可能不如传统RAG方案简洁明了。

此外，业务专家也认同GraphRAG方案在处理审计领域多实体和复杂问题查询时表现更优异的结论。

评估结果表明，GraphRAG方案在处理审计定性依据综合查询任务时具有显著优势，在处理多跳查询和需要整合实体关系数据进行复杂推理的查询场景中能够提供相关性高、完整性好、逻辑性强的结果。

六、总结与展望

本文提出了基于GraphRAG技术的审计定性依据查询系统，通过实验验证了其有效性以及相对于传统RAG方案的优势。GraphRAG技术能够显著提高审计定性依据检索的效率和准确性，为内部审计工作提供有力支持。

（一）主要贡献

1. 创新性地应用GraphRAG技术

本文将GraphRAG技术应用于审计知识检索领域，通过结合图数据库、大模型函数调用和向量检索技术，有效解决了传统RAG方法在处理复杂审计知识结构时的局限性，提升了查询的相关性和准确性，也为审计知识管理探索出一条新路径。相较于传统RAG方法，GraphRAG能够更全面地捕捉文本中实体间的关联关系，提供了更完整和高质量的查询结果。

2. 提出多实体语义嵌入方法

本研究基于Neo4j图数据库构建审计依据知识图谱，以GraphRAG技术框架为基础，创新性地提出了多实体语义嵌入方法，通过建立多个关键实体的语义向量索引，实现了更加细粒度的检索匹配。对于结构化程度高、语义密集的领域专业文本数据，该方法解决了传统文本分块方法可能导致的语义割裂、关键信息被掩埋的问题，大幅提高了知识检索精度。

3. 应用多路径检索优化算法

本研究设计了一种基于大模型函数调用能力的检索优化算法，能够准确地识别用户查询意图，并据此选择最优检索路径和查询参数，显著增强系统处理复杂查询的能力，在信息召回关联性方面表现优异。该算法灵活度高，除审计定性依据外，亦可对不同类型审计知识进行动态扩展，为构建智能、高效的综合审计知识检索系统奠定了技术基础。

（二）未来研究方向

1. 自动化图谱数据处理流程

现阶段审计定性依据数据准备仍依赖于审计人员人工收集和整理，数据专家进行数据清洗、知识图谱构建。为了提高系统的构建效率和可扩展性，未来可以进一步研究如何利用LLM实现审计知识数据的自动化处理和知识图谱的自动构建。例如，可以利用LLM对非结构化文本数据进行自动分类、关键信息提取和结构化转换，自动识别和纠正数据错误、异常值和不一致性，根据数据结构和关系信息自动生成图查询语句从而实现自动化图数据写入等。

2. 优化查询预处理和召回排序

采用更先进的查询预处理和召回排序技术，进一步提升信息检索的完整性和可靠性。例如利用查询重写技术拆分查询语句为多个关键词，将其映射到图谱中对应节点，并以这些节点为起点进行知识图谱搜索，发现直接和间接相关的信息，从而提高参考知识数据的丰富程度。此外，可以研究利用重排序（Rerank）模型对召回结果进行进一步筛选和排序、利用向量相似度数值动态确定引用数据量等方法以提升最终回复的匹配性和相关性。

3. 探索非结构化知识的GraphRAG方案

本文的GraphRAG研究方案对结构化知识图谱构建进行了优化适配。然而，很多审计知识检索场景中的数据是完全非结构化的，如审计报告、访谈记录、工作总结等。因此，未来可以继续探索针对非结构化文本数据的GraphRAG方案。例如，研究基于语义理解的方法，利用深度学习模型进行文本语义分析，识别隐含的概念和关系，构建语义网络。这将有助于处理更复杂和多样化的数据类型，进一步扩展其应用场景和实用性。

主要参考文献：

易冰心，刘思琦，王志勇，2023. AIGC在审计领域的运用[J]. 审计与鉴证(11): 69-73.

刘锦，2024. AIGC技术在国家审计中的应用[J]. 审计研究(4): 18-29.

金昌铉，孙丽丽，郑石桥等，2023. 内部审计的合规审计定性依据探讨[J]. 商业会计(7): 78-80.

向小伟，申艳光，胡明昊等，2024. 大模型驱动的科技政策法规问答系统研究[J]. 计算机科学与探索(9): 2349-2360.

陈亚盛，蒋礼蔚，单敏等，2024. 审计大模型的构建及应用研究——以员工违规经商办企业专项审计为例[J]. 审计研究(4): 139-149.

张春芝，张茜，党彦威，2024. 智能审计发展与应用分析[J]. 中国注册会计师(7): 104-107.

Peng B, Zhu Y, Liu Y, et al., 2024. Graph retrieval-augmented generation: a survey[R/OL]. arXiv: 2408.08921.

Darren E, Ha T, Newman C, et al., 2024. From local to global: a graph RAG approach to query-focused summarization[R/OL]. arXiv: 2404.16130.

Chen B, Zhang Z, Langrené N, et al., 2023. Unleashing the potential of prompt engineering in large language models: a comprehensive review[R/OL]. arXiv: 2310.14735.

Hu Y, Lei Z, Zhang Z, et al., 2024. GRAG: graph retrieval-augmented generation[R/OL]. arXiv: 2405.16506.

Lewis P, Perez E, Piktus A, et al., 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Advances in Neural Information Processing Systems. Cambridge: MIT Press: 9459-9474.

Ma X, Gong Y, He P, et al., 2023. Query rewriting for retrieval-augmented large language models[R/OL]. arXiv: 2305.14283.

Rathle P, 2024. The GraphRAG manifesto: adding knowledge to GenAI[EB/OL]. [2024-07-11]. https://neo4j.com/blog/graphrag-manifesto/.

Tang Y, Yang Y, 2024. MultiHop-RAG: benchmarking retrieval-augmented generation for multi-hop queries[R/OL]. arXiv: 2401.15391.

Xu Z, Cruz M J, Guevara M, et al., 2024. Retrieval-augmented generation with knowledge graphs for customer service question answering[R/OL]. arXiv: 2404.17723.

Ge Y, Hua W, Mei K, et al., 2023. OpenAGI: when LLM meets domain experts[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. New York: ACM: 5539-5555.

Yang A, Yang B, Hui B, et al., 2024. Qwen2 technical report[R/OL]. arXiv: 2407.10671.

Zhao P, Zhang H, Yu Q, et al., 2024. Retrieval-augmented generation for AI-generated content: a survey[R/OL]. arXiv: 2402.19473.

Zhong Z, Liu H, Cui X, et al., 2024. Mix-of-granularity: optimize the chunking granularity for retrieval-augmented generation[R/OL]. arXiv: 2406.00456.

Li Z, Zhang X, Zhang Y, et al., 2023. Towards general text embeddings with multi-stage contrastive learning[R/OL]. arXiv: 2308.03281.