基于大模型与主题树检索增强的防汛应急预案生成方法

刘会峰[1]，黄毅[2]，刘雪梅[3]，安新代[4]，董立新[2]，李海瑞[3]，王立虎[5]

（1. 华北水利水电大学 信息工程学院,河南 郑州 450046; 2.天津市水利科学研究院,天津 300061; 3.华北水利水电大学 数字孪生水利高等研究院,河南 郑州 450046; 4.黄河勘测规划设计研究院有限公司,河南 郑州 450003; 5.华北水利水电大学 管理与经济学院,河南 郑州 450046)

摘 要:【目的】针对防汛应急预案智能生成面临的领域专业知识支撑不足、场景泛化能力薄弱、生成结果可追溯性差等问题,研究领域知识库构建及可追溯的检索增强生成方法,以提升大语言模型生成防汛应急预案的专业质量。【方法】基于流域防洪资料构建防汛知识图谱,将其组织为主题树结构的知识库,提出基于主题树的混合检索增强生成方法(TH-RAG),利用 TH-RAG 对知识库检索得到的知识片段采用重排策略构造输入上下文,采用提示工程技术引导生成规范的应急预案文本。【结果】TH-RAG 在文档检索准确率与召回率均优于对比方法,分别达到 0.88 和 0.74,F1 分数综合表现最优;TH-RAG 生成的预案可操作性显著提升,在答案相关性和全面性上均取得最优值,分别为 0.95 和 0.93;构建的知识库具备高效的增量更新能力和泛化迁移潜力,增量更新机制相比全量重建快 7 倍以上,新增文档插入后性能保持稳定。【结论】TH-RAG 方法具备良好的跨模型迁移能力与动态可扩展性,可推广至水库大坝应急管理、气象灾害预警等相近领域的智能决策支持,为流域防洪减灾体系的智能化升级提供方法支撑。

关键词:大语言模型;检索增强生成;知识图谱;应急预案

中图分类号:TV87　　　　**文献标识码:**A

我国地理气候条件复杂,洪涝灾害频发,对人民生命财产安全和社会稳定构成严重威胁。科学、精准的防汛应急预案是有效组织抢险救灾、减少灾害损失的关键。然而,现实中防汛应急预案编制高度依赖专家经验,流程繁琐、周期长,探索科学高效的防汛应急预案智能生成方法十分必要。

目前主流的应急预案智能生成技术主要采用案例推理,通过匹配目标案例的相似性,修改相似度最高的历史案例应急方案来解决突发险情。郑静[1]、Yu X B[2]、Zhang K K[3]等分别采用剪枝算法、动态检索算法、设计差分进化算法等来优化案例检索,提高生成应急方案的准确性与实效性;刘雪梅[4]、LI B[5]、LIU X H[6]在领域应急管理中引进知识图谱,通过其强大的知识组织能力构建应急预案知识系统,为应急响应提供可靠信息。上述研究多依赖模板与规则优化,难以对未见场景与跨区域差异实现迁移,且泛化性与扩展性仍显不足。为了突破这些限制,需要从实际防汛业务需求出发,重新审视现有技术方案。

防汛应急预案生成涉及多部门联动,需要整合多维度知识。当前,大语言模型(Large Language Models,LLM)与检索增强生成(Retrieval Augmented Generation,RAG)等新兴技术为解决这一问题提供了新的可能。大语言模型在自然语言处理任务中展现出强大潜力,推动了基于生成式模型的知识服务系统在多个领域的应用探索[7-8]。检索增强生成框架因其结构简洁、可解释性强而被广泛采用[9-12]。张天鸿等人通过混合策略构建高质量数据集,实现了对灌浆通用知识的问答,展示了 LLM 在专业场景中的潜力[13];SARTHI P 等人构建层次化的树形检索系统,为大模型提供不同级别的抽象信息检索,提升复杂多步推理能力[14]。刘嘉琪等人构建基于大模型与多智能体协同的决策框架,支持复杂场景下的推理与决策[15]。这些探索表明,大模型与 RAG 结合具有显著应用潜力,为垂直领域的知识服务提供了新的思路。

大模型在水利领域已有初步应用,但将其应用

收稿日期/Received:2025-12-01　　　　**修回日期/Revised:**2026-04-08

基金项目:国家重点研发计划项目（2024YFC3210800）。

第一作者:刘会峰,男,研究生,从事大模型应用、智慧水利方面的研究。E-mail:x20231090820@stu.ncwu.edu.cn。

通信作者:刘雪梅(1965—),女,教授,博导,博士,从事数字孪生水利、大模型方面的研究。
E-mail:liuxuemei@ncwu.edu.cn。

于防汛应急预案生成时仍面临特定挑战。防汛应急管理涉及多部门,知识体系横跨多学科、多层次,大模型泛化能力虽强,但对防汛领域内在逻辑认知深度不足。另外,防汛应急预案的生成必须确保每项措施都能追溯至可信的知识源,大模型的黑盒生成机制难以满足这一要求。现有 RAG 方法各有局限,基于向量检索的增强生成(VectorRAG)利用语义相似度快速查询,但缺乏结构约束,难以应对防汛知识的逻辑关联。基于图的检索增强生成(GraphRAG)[16]具备清晰的结构表达能力,但语义理解能力相对较弱,且难以适应防汛应急场景多变的特点。两种方法都无法满足流域防洪业务对应急预案生成提出的可追溯、语义准确、适应灵活的综合要求。

基于上述分析,提出基于主题树的混合检索增强生成方法(Topic tree-based Hybrid Retrieval-Augmented Generation,TH-RAG)。研究采用三层技术架构,通过构建主题树知识库实现知识融合,设计混合检索策略构造上下文,规范预案生成方法,以提升大语言模型生成防汛应急预案的质量。

1　主题树知识库构建

1.1　方法整体架构

提出的主题树知识库构建过程分为两步,防汛知识图谱构建和层级索引构建,整体架构如图 1 所示。防汛知识图谱构建过程包括知识建模、抽取、融合、存储四个阶段。层级索引构建阶段,在图侧执行社区划分,并在聚合过程中对每个社区节点生成面向规则的结构化摘要。迭代此过程以自下而上构建主题树 T,并通过增量更新机制维护索引的一致性与时效性。

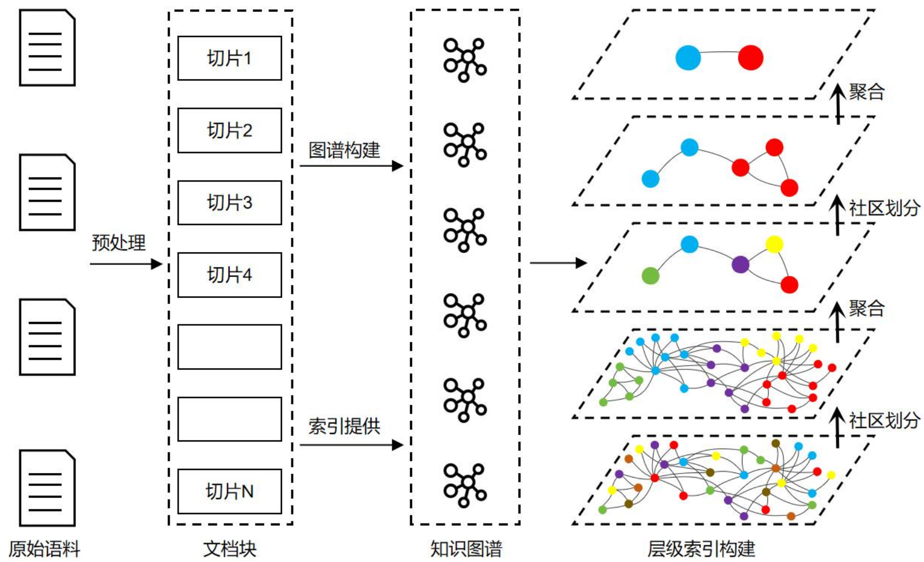

图 1　主题树知识库构建框架

Fig.1　Theme Tree Knowledge Base Construction Framework

1.2　防汛知识图谱构建

面向某流域防汛决策场景,构建涵盖关键知识要素的多源知识图谱。数据来源主要包括某流域及其控制的水利工程2019年至2024年五年内发布的防汛抢险预案文件、工程安全鉴定报告、河道与堤防工程资料、险工险段排查记录、一般险情抢护方法、抢险物资与装备清单、仓储设施分布信息等多个子类,覆盖了“人、物、地、事、时间”等应急管理的核心维度。这些数据来源从不同角度反映了防汛任务中涉及的对象关系、空间分布和时间演化特征。

1.2.1 本体建模

为了有效组织和管理多源防汛知识,基于业务需求分析,构建涵盖防汛应急全过程的本体模型。本体建模遵循自上而下的领域知识抽象与自底向上的样本归纳相结合的方法,逐步明确了本体类目体系及其层级结构。本体模型涵盖以下主要实体类:①工程类,例如水利工程、防洪地点等,用于表示防洪的基础设施与工程属性;②险情类,例如风险事件、险情表现,用于描述可能发生的险情及表征特征;③措施类:应急预案、工程措施等,用于表示对特定险情的处置方案;④资源类,物资、设备、调用人员等,用于管理应急所需的物资与装备资源;⑤动态类,防汛责任、响应级别等,用于表示风险的动态演化与响应决策。防汛本体模型为后续知识抽取统一的语义模式支撑,本体建模如图 2 所示。

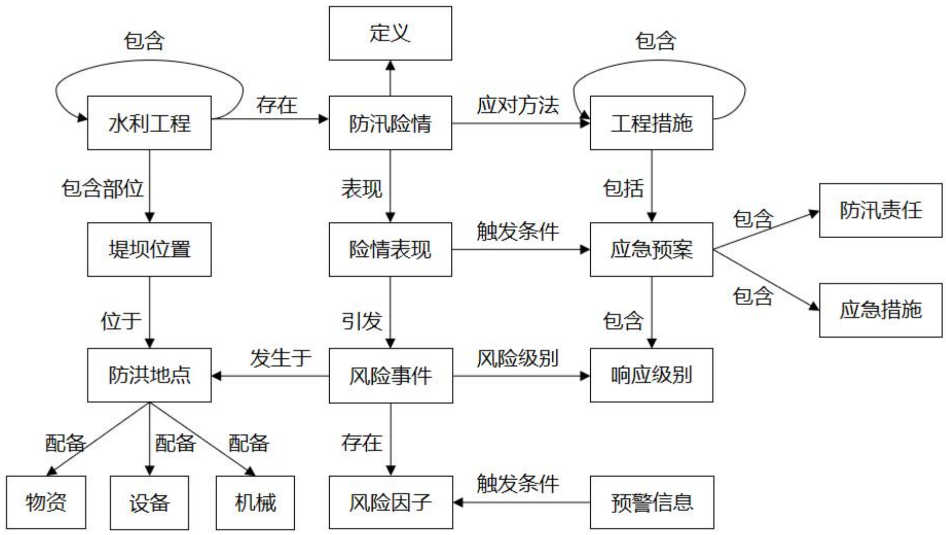

图 2 防汛领域本体模型

Fig.2 Flood Control Domain Ontology Model

1.2.2 知识抽取与融合

根据数据的异构性特点,采用多模式的知识抽取方式。针对某流域防汛数据高度异构、种类繁多的特点,构建了基于多智能代理(Agent)协同的知识抽取系统,以本体规则为基础,通过 Agent 协同实现实体识别、关系抽取、结果验证与知识融合的完整流程[17]。

1.2.3 知识存储

为同时满足结构化知识管理与语义检索的要求,采用图数据库与向量数据库的双存储策略。以 Neo4j 作为知识图谱库进行图谱管理,Milvus 作为高维向量索引库承载知识图谱与文档块的嵌入表示,二者通过索引联接保持一致联动。典型抽取示例见表 1。

表 1 渗水险情抢护方法三元组统计表

Tab.1 Triple Statistics of Seepage Emergency Response Methods

头实体	对应关系	尾实体	切片序号	文档序号
渗水险情	抢护原则	前截后导	D327	DOC003
前截后导	包含步骤	前截	D327	DOC003
前截后导	包含步骤	后导	D327	DOC003
前截	位置	堤防临水侧	D327	DOC003
前截	工程措施	外帮防渗	D327	DOC003
前截	使用材料	粘性土料	D327	DOC003
前截	功能	降低浸润线	D327	DOC003
后导	位置	堤背水坡	D328	DOC003
后导	工程措施	反滤排水设施	D328	DOC003
渗水险情	附加措施	修筑固脚平台	D328	DOC003

1.3 主题树层级索引构建

1.3.1 图谱社区划分

为实现多源知识融合,对知识图谱使用 Leiden 社区检测算法进行社区划分,以识别语义上紧密相关的实体子集[18-19]。Leiden 算法具备连通性保障与收敛性控制,能够有效避免传统划分算法中易出现的碎片化或孤立节点问题。

Leiden 算法的核心目标是最大化图结构中的模块度(Modularity,Q),为适应不同粒度的语义压缩需求,进一步引入了带分辨率参数的模块度优化机制,将模块度函数拓展为:

$$Q_r = \frac{1}{2m}\sum_{ij}\left[A_{ij} - r\frac{k_i k_j}{2m}\right]\delta(c_i, c_j), r > 0 \tag{1}$$

其中r为用户设定的分辨率参数,较小的r值将产生较大的社区,而较大的r值则导致更细粒度的社区。该机制允许系统根据知识图谱规模和应用场景灵活调整划分策略[20]。这里设定r的搜索网格,$r\in\{0.1,0.2,0.3....,1.0\}$。

1.3.2 层级摘要生成

完成知识图谱社区划分后,针对每个社区中实体及其关系网络预设一系列高频关键问题作为摘要生成的引导模板。问题种类包含定义、功能、适用条件、流程、约束、风险时效等。此类问题既源于人工总结的领域任务需求,也结合了现有语料中的高频指令表达。在此基础上,通过大语言模型对社区中的结构化三元组和原始文本片段进行统一处理,生成面向问答任务的主题摘要。生成的摘要不仅准确概括了社区的主题,也在一定程度上保留了其与其他社区的语义关联,为后续的主题树构建提供了高层语义锚点[21]。

1.3.3 主题树递归聚合

迭代重复 2.3.1 和 2.3.2 节的工作,知识库的构

建以图谱社区划分结果为基础,逐层聚合语义相近的子图,并对每一层划分对应的社区节点C。在整个递归聚合过程中,构建出多层次的主题树 T。树中每个节点同时保留以下三类信息:(1)本层社区摘要S_C;(2)对应的实体集合ξ_C;(3)原始文档片段索引I_C。

对于迭代过程终止的条件,采用带分辨率参数的模块度函数作为社区聚合判断依据[22]。任意两个或多个语义邻近的社区节点,系统尝试将其合并为更高层的父节点,并重新计算聚合后的模块度变化量ΔQ_r。

$$\Delta Q_r \quad C_u \oplus C_v) = Q_r(\mathcal{P} \setminus \{C_u, C_v\} \cup \{C_u \cup C_v\}) - Q_r \quad P) \tag{2}$$

C_u,C_v为当前划分中的两个社区节点,彼此不相交。$\mathcal{P}$为当前的社区划分,是若干个社区的集合。

设定阈值$\tau \in \{0,0.01,0.02....,0.03\}$,在满足最大化$J = Rel + Cov - \lambda \cdot Latency$同时,取最大的$\tau$。当$\Delta Q_r \geq \tau$时,执行社区合并操作;若$\Delta Q_r<\tau$,则停止当前层聚合。

1.3.4 增量索引更新

为适应知识体系的持续扩展需求,构建的知识库支持动态接入与融合。新文档按照上述流程生成一棵局部主题树T_{new}[22],系统依据其根节点与原有主题树T_{main}中各社区节点的匹配度,计算其插入候选节点集合[23]。匹配度函数公式如下:

$$S_{ins}(T_{new}, C_i) = \lambda_1 \cdot Sim_{sem}(T_{new}, C_i) + \lambda_2 \cdot Cov_{kw}(T_{new}, C_i) + \lambda_3 \cdot Align_{struct}(T_{new}, C_i) \tag{3}$$

式中,Sim_{sem}表示语义向量相似度,C_i表示主索引树一个候选已有社区,Cov_{kw}为关键词覆盖率,$Align_{struct}$为结构标签与关系对齐程度,λ_i为经验权重系数。

当匹配度S_{ins}高于预设阈值θ_{merge}时,执行消融插入,并将其子社区逐级挂载至该节点下;否则,执行独立插入,即将T_{new}作为一个新的子分支挂接至虚根节点,以便后续吸收与融合。

2 基于混合检索的预案生成方法

知识库的建立为防汛应急预案生成提供了知识基础,本节聚焦于如何利用上述构建的知识库,进行大模型驱动的智能化预案生成。

2.1 混合检索策略

在检索阶段,采用级联双约束检索策略。基于知识库的特点进行树遍历,沿已构建的主题树进行主题检索。随后在定位到的候选范围内,执行语义检索。算法流程如下:

输入:查询q,主题树T,知识图谱与文档联接产生的文本块集合$\mathcal{I}(C)$,阈值θ,θ_high,结构打分权重α,束宽b,最大层高H。相关参数在验证集上调优得到。

阶段 A:树遍历

自根节点起按层遍历/束搜索,对每个社区节点C计算路径相关性分数

$$S_{path}(q, C) = \alpha\, sim_{scm}(q, S_C) + (1 - \alpha)\, cov_{kw}(q, S_C) \tag{4}$$

式中sim_{scm}为余弦语义相似度(归一化到[0,1]),cov_{kw}为关键词覆盖度。仅保留$S_{path} \geq \theta$的分支进入下一层,并在每层对候选子节点按S_{path}排序取前b。若某节点满足$S_{path} \geq \theta_high$,将其标记为潜在答案节点。阶段 A 的输出为社区集合:

$$\mathcal{P} = \{C | C\text{被保留/标记}\} \tag{5}$$

以及其聚合文本块候选域:

$$D_{\mathcal{P}} = \bigcup_{C \in \mathcal{P}} \mathcal{I}(C) \tag{6}$$

阶段 B:语义检索

在受限语料域$D_{\mathcal{P}}$上执行向量召回,获得细粒度事实证据。对每个文本块$d \in D_{\mathcal{P}}$计算向量相似度$s_{vec} = cos(e(q), e(d)$。为体现结构先验,引入路径感知权重:

$$\mathrm{w}_{\mathrm{struct}}(\mathrm{d}) = exp\,(-\delta \cdot dist(\mathrm{d}, \mathcal{P})) \tag{7}$$

式中 *dist* 为文本所属社区到选中路径的层距,得到阶段性分数:

$$s_0(d) = \lambda\, s_{vec} + (1 - \lambda)\, w_{struct}(d) \tag{8}$$

据此选取 top—N 文本块V,若$|D_{\mathcal{P}}|$过小,可按比例ρ引入全库全局 $Top—N_g$作为回退补充,防止漏检。

阶段 B 的输出:路径摘要节点$\mathcal{P}$与文本块V。

2.2 上下文构造策略

本节在输出的候选上完成统一打分与上下文选择。设定候选池:

$$C_{cand} = \mathcal{P} \cup V \tag{9}$$

为兼顾"路径一致性、语义覆盖、片段连贯、实体约束",对任一候选 $c \in C_{cand}$定义四个可计算指标[24],并归一化到[0,1]。最终融合分数:

$$S(c) = w_1 \hat{f}_{cons}(c) + w_2 \hat{f}_{cov}(q, c) + w_3 \hat{f}_{dens}(c) + w_4 \hat{f}_{ent}(q, c) \quad \sum_{i=1}^{4} w_i = 1 \tag{10}$$

为稳定全局主题与局部证据的比例,设摘要节

点与文本块的配额分别为K_s,K_d总预算$B=K_s+K_d$,分别在摘要集合$\mathcal{P}$与文本集合V内按$S(c)$降序取前K_s,K_d。流程中将摘要节点置于前,随后按文档段落顺序聚合相邻文本块,输出上下文 Context(q),保留来源 ID 以支持可回溯。

2.3 预案生成引导

为引导大模型生成专业、规范的防汛应急预案,基于提示工程设计了一套结构化的防汛应急预案生成引导流程。流程采用多层次架构设计,主要包含四个关键模块:系统角色定义模块、知识上下文注入模块、结构化输出引导模块和合规性约束模块。

系统角色定义模块通过精心设计的提示词为大语言模型赋予特定的专业身份和角色定位。该模块明确定义角色的职责边界和专业领域,使模型输出内容更贴近专业人士的思维逻辑和表达方式。

知识上下文注入模块负责将混合检索获得的多维度知识有效融合到生成过程中。该模块通过保留片段级引用信息构建可追溯的知识上下文,根据当前生成内容的主题和逻辑需求,动态选择相应的背景知识,确保生成内容具备专业知识支撑。

结构化输出引导模块基于专业领域文本的标准框架和规范要求,设计了详细的章节模板和逻辑关系图。该模块提供格式规范模板,指导模型按照标准格式输出相应内容。

合规性约束模块实施双重验证机制,确保生成内容符合相关标准和规范要求。建立基于防汛应急行业标准,通过关键词匹配、格式验证确保生成内容的规范性。

2.4 评估指标

为评估不同检索方法在防汛知识问答任务中的表现,检索采用通用评估指标。如下:

准确率(Precision):检索结果中相关文档的比例。公式如下:

$$P_k=\frac{TP}{TP+FP} \tag{11}$$

P_k为前K个结果的准确率,TP为检索到的相关文档;FP为检索到的不相关文档。

召回率(Recall):成功检索到的相关文档占所有相关文档的比例。公式如下:

$$R_k=\frac{TP}{TP+FN} \tag{12}$$

R_k为前K结果的召回率,FN为未被检索到的相关文档。

F_1分数是准确率和召回率的调和平均值,平衡两者的权重。公式如下:

$$F_1=\frac{2TP}{2\times TP+FP+FN} \tag{13}$$

采用 RAGAS[25]框架中的多项指标来评估生成文本的质量,包括忠实度(Faithfulness,F)、答案相关性(AnswerRelevance,AR)、上下文精度(ContextPrecision,CP)、全面性(AnswerCoverage,COV)。同时引入防汛预案专属指标:

可操作性(Operability,OP):衡量生成的防汛预案是否包含清晰、具体、可直接执行的操作步骤和措施。

$$OP=\frac{|\{i|instruction_i\}|}{|all\_instruction|} \tag{14}$$

其中,OP 表示可操作性评分,$instruction_i$表示预案中的第 i 条指令。

3 实验结果分析

3.1 实验设置

为系统性评估提出方法的效果,引入三种主流对比方法作为参考基线:VectorRAG、GraphRAG 和 LightRAG[26]。

为统一检索侧配置并降低实现差异带来的干扰,除特别说明外,所有方法均共享同一嵌入与索引参数以确保公平比较。研究使用三种不同的大模型后端 DeepSeek-R1、Kimi2、GLM4.5 进行测试,每组问题重复三次取平均值。

实验评估构建验证集,旨在检验所提出的方法有效性。将子集问题划分为两类,基础检索型问题和指令生成型问题,见表 2。

表 2 验证集统计

Tab.2 Validation Set Statistics

问题类型	数量/条	生成方式	示例问题	关注点
基础检索型问题	70	模板化规则+人工标注	"洪涝灾害IV级应急响应条件有哪些?"	实体层级信息定位精度
指令生成型问题	30	人工标注+大模型审核	"当津汉改线桥至南环桥段发生溃坝险情时,当前物资调配策略是否存在短板?"	全局知识组织与社区语义理解

总计	100	—	—	—

3.2　知识库构建结果分析

构建主题树知识库,具体情况如表 3 所示。主题树的深度由社区规模 N 和树的平均分支数 b 确定,根据实际的 Leiden 社区划分结果,计算得出该主题树深度为 5。深度 5 在检索延迟与语义覆盖度间达到最优平衡,生成的预案质量在后续实验中表现最优。按照从低到高的顺序,最底层为叶节点。叶节点为图谱实体和文档块的联接,树形检索体系的构造以此为基础。四级节点索引连接若干区域的图谱实体和文档块,如节点“工程险情抢险方法”“工程险情抢险原则”等。每升一层,节点的概括范围更广,抽象程度更高,逐步从具体技术细节向领域概念、主题类别过渡,最终形成完整的防汛知识体系框架。

表 3　层级主题数构建统计
Tab.3　Hierarchical Topic Quantity Construction Statistics

数据类型	数量	示例	备注
一级社区节点	7	"险情处置"	摘要小于 350 字
二级社区节点	25	"堤防险情"	摘要小于 350 字
三级社区节点	128	"管涌""漫溢"	摘要小于 350 字
四级社区节点	721	"管涌险情抢护原则"	摘要小于 350 字
图谱文档联接	10479	"抢护原则应是前截后导..."	与文档块数量一致

3.3　检索结果分析

3.3.1 检索案例分析

本节通过一个具体案例来分析其检索过程及与其他 RAG 方法的对比。设置查询问题:“当北三河险工险段王善庄段发生渗水险情时,应采用什么技术措施?”该问题涉及特定地理位置、特定险情类型以及相应的抢护方法,是典型的复合查询。

TH-RAG 检索过程分析:

阶段 A:TH-RAG 首先沿主题树自顶向下进行树遍历,系统计算查询与一级社区节点的相关性分数,“险情处置”社区获得最高匹配分数,成功定位到相关主题域。在“险情处置”下,“堤防险情”子社区因包含渗水、管涌等险情类型,获得更精确匹配。继续向下,“渗水”社区节点因直接对应查询中的险情类型,触发高相关性阈值,被标记为潜在答案节点。进入四级社区,渗水险情抢护原则"等具体技术节点被纳入候选集合 P。通过树遍历,系统将搜索范围从全库 10479 个文档块压缩至 237 个高相关文档块,搜索效率提升至 97.7%。

阶段 B:在受限语料域 DP 上执行语义检索,检索到的 top-10 文档块中,8 个直接来自“渗水险情抢护原则”社区。路径感知权重机制确保了来自相关路径的文档获得更高排序,最终选出的文档块包含“前截后导”抢护原则、“外帮防渗+反滤排水”技术措施等核心信息。

对比其他方法,VectorRAG 直接进行向量检索,虽然能检索到部分相关文档,但存在以下问题。其检索结果中混入了“滑坡险情”、“冲刷险情”等语义相近但类型不符的内容,致使语义漂移,且无法区分地理位置的针对性,检索到了其他流域的处置方法。GraphRAG 基于知识图谱进行检索,严格按照图谱边进行遍历,与 TH-RAG 总检索时长接近。但是对于“北三河险工险段”这样的复合地理概念,存在对于实体识别不准确而漏检的情况,比如遗漏了王善庄段防抢措施中的防抢物资和机具数量。

3.3.2 检索结果讨论

基于前述评估指标,使用验证集对 TH-RAG 及三种基准方法进行了系统评估。表 4 展示了不同方法在防汛知识检索任务中的表现对比。在文档检索任务中,TH-RAG 与 VectorRAG 在准确率上表现相当,但 TH-RAG 在召回率上提升了 4.2%。在实体检索任务中,GraphRAG 展现出最佳性能,其准确率和 F1 分数均高于其他方法,但在召回率上与 TH-RAG 相当,表明其在实体检索方面也具有较强竞争力。为了量化验证 TH-RAG 中的双约束策略的有效性,进行消融实验。实验结果证明移除任一模块都会导致性能下降,树遍历的加入提升了 F1 约 2.5%(文档),这说明主题树的层级结构能够有效地缩小检索范围,避免在庞大的知识库中“迷失”。语义检索的加入提升了 F1 约 10.8%(实体),说明语义检索能够补充树遍历的不足,通过向量相似度捕捉语义关联,并处理树结构中难以表达的跨社区关系。

表 4　检索结果评估分数展示
Tab.4　Retrieval Result Assessment Scores

指标类型	检索方法	P_k	R_k	F_1分数
文档检索	VectorRAG	0.88	0.71	0.78
文档检索	LightRAG	0.85	0.69	0.76
文档检索	TH-RAG	0.88	0.74	0.80
文档检索	TH-RAG(仅语义检索)	0.88	0.71	0.78
实体检索	GraphRAG	0.80	0.68	0.76
实体检索	LightRAG	0.71	0.58	0.64
实体检索	TH-RAG	0.78	0.68	0.74
实体检索	TH-RAG(仅树遍历)	0.71	0.61	0.66

3.4　生成结果分析

3.4.1 生成案例分析

为展示 TH-RAG 在预案生成环节的优势,本节

通过具体案例分析其生成过程和结果质量。基于前述检索案例的查询:“当某流域险工险段王善庄段发生渗水险情时,应采用什么技术措施?请为我生成完整的应急预案”。

TH-RAG 生成过程分析:

上下文构造:经过检索阶段,TH-RAG 获得了丰富的上下文信息,主题包括来自“渗水险情抢护”社区的结构化摘要,提供宏观处置框架。具体信息包含“前截后导”抢护原则、“外帮防渗”工程措施、“粘性土料”材料选择等详细信息,并涵盖了某流域的工程特点和历史处置经验。同时,每个知识片段都标注了来源文档 ID,确保可追溯性。

预案生成引导:系统使用预设模板,包含以下关键章节(1)险情判断与分析;(2)抢护原则确定;(3)技术措施选择;(4)物资设备调配;(5)组织协调机制;(6)应急保障措施。

TH-RAG 生成过程中,通过主题树的层级结构从宏观摘要中获取处置框架和原则指导,从具体文档块中提取详细的技术参数和操作要求。主题树的层级关系天然反映了知识的逻辑结构,整合多源信息形成完整的预案体系,构造逐层递进的关系。

从防汛应急预案的实用性角度,TH-RAG 生成的预案表现出较强的可操作性。与其他方法生成结果对比,VectorRAG 生成的预案存在以下问题,由于其检索到的文档缺乏结构化组织,生成的预案逻辑性较差,且相似内容的重复表述,导致预案冗长重点不突出。生成的预案缺乏针对性,混入了其他类型的险情的处置方法,专业性不足。与 GraphRAG 生成预案对比,GraphRAG 虽然保持了较好的结构性,但其生成内容有限。生成的预案严格依赖图谱关系,遗漏了一些重要的技术细节,且生成内容过于依赖原文表述,缺乏自然语言的灵活性。

3.4.2 生成结果讨论

表 5 展示了各方法在忠实度、答案相关性、上下文精度、全面性、可操作性五个指标上的表现。GraphRAG 通过图结构获得了最高的上下文精度和较高的全面性,但缺乏细节补充。VectorRAG 在文档检索准确率上表现良好,但缺乏结构化组织导致上下文精度较低。LightRAG 在各项指标上表现均衡,在各项指标都接近领先水平。TH-RAG 通过主题树结构成功平衡了各项指标,在文档检索 F1 分数上领先,在答案相关性和全面性上达到最高,分别为 0.95 和 0.93,可操作性也达到最高。研究进行了生成任务的消融实验。该实验通过移除树遍历或语义检索模块,评估每个模块对生成质量四个方面的贡献。实验数据显示,移除树遍历后,答案相关性和重视度下降幅度最小,说明树结构在确保预案的针对性和覆盖度方面的作用显著。移除语义搜索,系统只能获得树中显式组织的知识,容易遗漏那些语义上相关但不在树的搜索路径的知识。语义检索补充了这些隐性相关的知识,使预案更加准确和全面。树遍历和语义检索形成了一个互补的约束框架,既有结构性的约束,又有灵活性的补充,这正是处理防汛这类复杂、多维度问题的必然需求。实验结果验证了主题树结构的有效性,通过主题树结构凝练主题加强答案相关性,又提供足够宽广的知识覆盖增强答案全面性。

表 5　生成结果评估分数展示
Tab.5　Generation Result Assessment Scores

方法	F	AR	CP	COV	OP
VectorRAG	0.93	0.91	0.83	0.82	0.68
GraphRAG	0.92	0.88	0.95	0.91	0.71
LightRAG	0.95	0.93	0.90	0.90	0.75
TH-RAG	0.96	0.95	0.91	0.93	0.88
TH-RAG(仅树遍历)	0.93	0.89	0.88	0.88	0.75
TH-RAG(仅语义检索)	0.94	0.92	0.83	0.82	0.72

3.5　增量更新性能评估

3.5.1 增量更新实验设计

基于增量更新的动态索引维护机制,用于高效地处理新增防汛文档的接入与融合。本节通过模拟新文档插入的过程,验证该机制在不同规模数据增量下的效率表现和检索准确率的稳定性。实验目标为评估系统在面临新增防汛文档时的以下性能指标:①索引维护效率,即单个文档的平均插入时间、批量插入的吞吐量;②检索稳定性,即新文档插入前后的检索准确率变化。新文档从实际防汛预案库中抽取,覆盖各类防汛场景,文档内容的语言风格与预处理方式与原库保持一致。增量插入模拟 4 个批次的新文档插入,每个批次包含不同数量的文档。评估基准控制组每次插入后执行全量重建,重新运行 Leiden 社区划分和摘要生成,实验组采用研究提出的增量更新机制。

3.5.2 增量更新结果分析

针对知识库的更新机制进行了实验,其中索引维护效率如表 6 所示,检索生成效率如表 7 所示。由表中数据可以看出,增量更新的速度始终快 7 倍以上,随着数据规模增大,优势愈发明显。检索效率比较稳定,即使增加大量级的文档,F1 值仅略微提升 1.3%,说明新增文档中包含了与原库高度相似的知识,增量更新机制有效地将其融入了现有索引。准确率的小幅提升证明了增量更新的有效性,新增文档使得某些知识的覆盖更加完整,但不会因为过度

调整而破坏原有的检索性能。

表 6　不同数据增量的索引维护时间对比

Tab.6　Comparison of Index Maintenance Time for Different Data Increments

数据增量	新文档数	增量更新	全量重建	检索延迟增长
Batch 1	10	12.3	87.5	+0.8ms
Batch 2	20	24.1	175.2	+1.2ms
Batch 3	50	58.6	438.7	+2.1ms
Batch 4	100	115.2	877.4	+3.5ms

表 7　新文档插入对检索性能的影响

Tab.7　Impact of New Document Insertion on Retrieval Performance

批次	新文档数	插入前 F1	插入后 F1	平均检索延迟
基准	0	0.857	-	78ms
1	10	0.857	0.859	79ms
2	20	0.859	0.862	80ms
3	50	0.862	0.864	82ms
4	100	0.864	0.868	85ms

4　结论

面向某流域知识数据进行梳理,提出基于主题树的混合检索增强防汛应急预案生成方法,总结如下:

1)提出基于主题树的知识库构建方法,通过防汛知识图谱构建与 Leiden 社区划分、递归聚合相结合,将多源异构知识融合为具有层次语义结构的主题树,有效解决了大模型在防汛专业场景中知识匮乏、针对性不足的问题。该知识库具备高效的增量更新能力,相比全量重建快 7 倍以上,且新增文档插入后检索性能保持稳定。

2)设计了级联双约束混合检索策略,通过树遍历定位相关主题域,再经语义检索获取细粒度事实证据,并通过路径感知权重与多维融合打分机制构造可回溯上下文,有效解决了大模型生成预案可追溯性差的问题。

3)TH-RAG 方法具有良好的跨模型迁移能力和动态可扩展性,可推广至水库大坝应急管理、气象灾害预警等相近领域的智能决策支持,为流域综合防洪减灾体系的现代化升级提供有力支撑。

目前该方法仍存在一些待完善之处,未来研究将重点探索基于变更感知的图谱动态更新机制以及更为紧耦合联合索引框架,提升系统在大规模动态环境下的稳定性与泛化能力。

参 考 文 献

[1] 郑静,王应明,陈胜群.基于主观偏好和客观信息的应急决策动态案例检索方法[J].IEEE/CAA 自动化学报,2018,5(3):749-757.[ZHENG Jing,WANG Yingming ,CHEN Shengqun .Dynamic case retrieval method with subjective preferences and objective information for emergency decision making[J].IEEE/CAA Journal of Automatica Sinica,2018,5(3):749-757.]

[2] YU X B,LI C L,ZHAO W X,et al.A novel case adaptation method based on differential evolution algorithm for disaster emergency[J].Applied Soft Computing,2020,92:106306.

[3] ZHANG K K,LUO N X,LI Y B.STGA-CBR:A case-based reasoning method based on spatiotemporal trajectory similarity assessment[J].IEEE Access,2020,8:22378-22385.

[4] 刘雪梅,卢汉康,李海瑞,等.知识驱动的水利工程应急方案智能生成方法:以南水北调中线工程为例[J].水利学报,2023,54(6):666-676.[LIU Xuemei M,LU Hangkang,LI Hairui ,et al.Knowledge-driven intelligent generation method for water conservancy engineering emergency plan:a case study of the middle route of South-to-North Water Diversion Project[J].Journal of Hydraulic Engineering,2023,54(6):666-676.]

[5] LI B,LI T T,JIANG Q,et al.A knowledge-based system for disaster emergency relief[J].International Journal of Pattern Recognition and Artificial Intelligence,2020,34(11):2059038.

[6] LIU X H,LIU Z,TIAN J K,et al.Integration of a geo-ontology-based knowledge model and spatial analysis into emergency response for geologic hazards[J].Natural Hazards,2021,108(2):1489-1514.

[7] PATIL R, G V.A review of current trends, techniques, and challenges in large language models (llms)[J].Applied Sciences,2024,14(5):2074.

[8] Zhao Y,He J W,ZHU S C,et al.Security of Large Language Models:Current Status and Challenges[J].Computer Science, 2024, 51(01):68-71.

[9] LEWIS P,PEREZ E,PIKTUS A,et al.Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Advances in Neural Information Processing Systems.Red Hook:Curran Associates Inc,2020,33:9459-9474.

[10]IZACARD G,GRAVE E.Leveraging passage retrieval with generative models for open domain question answering [C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume.Stroudsburg:Association for Computational Linguistics,2021:674-687.

[11]GUU K,LEE K,TUNG Z,et al.REALM:retrieval-augmented language model pre-training[C]//Proceedings of the 37th International Conference on Machine Learning.Cambridge:MIT Press,2020:3929-3938.

[12]TANG Y X,YANG Y.MultiHop-RAG:benchmarking retrieval-augmented generation for multi-hop queries[J/OL].arXiv preprint,2024:arXiv:2401.15391.

[13]张天鸿,王晓玲,余红玲,等.基于大语言模型的灌浆工程知识服务系统[J].水利学报,2025,56(1):130-142.[ZHANG Tianhong ,WANG Xiaoling,YU Hongling,et al.Grouting works k-nowledge service system based on large language model[J].Journal of Hydraulic Engineering,2025,56(1):130-142.]

[14]SARTHI P,ABDULLAH S,TULI A,et al.RAPTOR:recursive abstractive processing for tree-organized retrieval[C]//The Twelfth International Conference on Learning Representations.Vienna:OpenReview,2024.

[15]刘嘉琪,刘雪梅,孙维亚,等.南水北调中线工程防汛应急智能体研发及应用[J/OL].华北水利水电大学学报(自然科学版),20260313.1046.002].[LIU Jiaqi ,LIU Xuemei ,SUN Yawei ,et al.Research and Application of Flood Control Emergency Agent for the Middle Route of South-to-North Water Diversion Project[J/OL]. Journal of North China University of Water Resources and Electric Power (Natural Science Edition),20260313.1046.002.]

[16]EDGE D,TRINH H,CHENG N,et al.From local to global:a graph RAG approach to query-focused summarization[J/OL].arXiv preprint,2024:arXiv:2404.16130.

[17]SINGH A,EHTESHAM A,KUMAR S,et al.Agentic retrieval-augmented generation:a survey on agentic RAG[J/OL].arXiv preprint,2025:arXiv:2501.09136.

[18]SAHU S,KOTHAPALLI K,BANERJEE D S.Fast Leiden algorithm for community detection in shared memory setting[C]//Proceedings of the 53rd International Conference on Parallel Processing.New York:ACM,2024:11-20.

[19]Verweij G. Faster community detection without loss of quality: parallelizing the Leiden algorithm[D]. Leiden: Leiden Institute of Advanced Computer Science, Leiden University, 2019.

[20]NAVLAKHA S,FALOUTSOS C,LESKOVEC J.Community-aware graph summarization for fast query processing[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.New York:ACM,2014:401-410.

[21]ANAGNOSTOPOULOS A,LILLIS D,QIU J,et al.General-purpose query processing on summary graphs[J].Social N-etwork Analysis and Mining,2024,14(1):34.

[22]SHERVASHIDZE N,SCHWEITZER P,VAN LEEUWEN E J,et al.Weisfeiler-Lehman graph kernels[J].Journal of Machine Learning Research,2011,12:2539-2561.

[23]SUN Z,HU W,ZHANG Q,et al.Bootstrapping entity alignment with knowledge graph embedding[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence.Stockholm:IJCAI,2018:4396-4402.

[24]兰天,马梓奥,周杨浩,徐晨,等.生成式文本质量的自动评估方法综述[J].计算语言学,2024,2:169-196.[Lan Tian., Ma Ziao, Zhou Yanghao, Xu Chen,et al.A survey of automatic evaluation methods for generative text quality[J].Computational Linguistics,2024,2:169-196.]

[25]ES S,JANES J,ESPINOSA-ANKE L,et al.RAGAs:automated evaluation of retrieval augmented generation[C]//Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics.Stroudsburg:ACL,2024:150-158.

[26]GUO Z R,XIA L H,YU Y H,et al.LightRAG:simple and fast retrieval-augmented generation[J/OL].arXiv preprint,2024:arXiv:2410.05779.

A Flood Emergency Response Plan Generation Method Based on Large Language Models and Topic Tree Retrieval Augmentation

LUI Hui Feng[1], HUANG Yi[2], LUI Xue Mei[3], AN Xin Dai[4], DONG Li Xin[2], LI Hai Rui[3], WANG Li Hu[5]

(1 School of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450046, China;2 Tianjin Research Institute of Water Resources Science, TianJin 300061, China;3 Advanced Research Institute for Digital Twin Water Conservancy, North China University of Water Resources and Electric Power, Zhengzhou 450046, China;4 Yellow River Engineering Consulting Co., Ltd. , North China University of Water Resources and Electric Power, Zhengzhou 450003, China;5 School of Management and Economics, North China University of Water Resources and Electric Power, Zhengzhou 450046, China)

Abstract:【**Objective**】To address the problems of insufficient domain knowledge support, weak scenario generalization ability, and poor traceability of generated results in the intelligent generation of flood control emergency plans, this study investigates the construction of a domain knowledge base and a traceable retrieval-augmented generation method, aiming to improve the professional quality of flood control emergency plans generated by large language models.【**Methods**】Based on basin flood control data, a flood control knowledge graph is constructed and organized into a knowledge base with a topic tree structure. A topic tree-based hybrid retrieval-augmented generation method (TH-RAG) is proposed. TH-RAG retrieves knowledge snippets from the knowledge base, employs a reranking strategy to construct the input context, and uses prompt engineering techniques to guide the generation of standardized emergency plan texts.【**Results**】In document retrieval, TH-RAG achieves a precision of 0.88 and a recall of 0.74, both outperforming the compared methods, with the best overall F1 score. The emergency plans generated by TH-RAG show significantly enhanced operability, attaining optimal scores in answer relevance (0.95) and comprehensiveness (0.93). The constructed knowledge base possesses efficient incremental updating capability and generalization transfer potential; the incremental update mechanism is over seven times faster than full reconstruction, and the performance remains stable after inserting new documents.【**Conclusion**】The TH-RAG method exhibits good cross-model transferability and dynamic scalability, and can be extended to intelligent decision support in related fields such as reservoir and dam emergency management and meteorological disaster warning, thereby providing methodological support for the intelligent upgrading of basin flood control and disaster reduction systems.

Keywords: Large Language Model; Retrieval-Augmented Generation; Knowledge Graph; Emergency Plan