计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250100011

结合图检索与上下文排序的检索增强生成技术研究

薛晓楠
北京青年政治学院　北京 100102

摘　要　复杂问答任务需要模型能够从大规模异构知识中高效检索相关信息,同时支持生成高质量答案。然而,现有检索增强生成方法在知识检索、语意关联度和生成一致性上存在诸多挑战:1)知识检索模块的粒度和结构化信息不足;2)检索上下文相关性不足,排序能力有限,以及生成质量受限;3)生成模型难以准确整合检索到的知识并生成上下文一致的答案。为解决上述问题,提出了一种结合图检索增强生成与上下文排序的大语言模型生成框架 GraphRank-RAG。该框架通过引入基于图的检索机制,捕获上下文间的深层语义关联,优化上下文排序与答案生成过程。实验结果表明,该方法在多个开放域问答数据集上的表现优于现有方法,在检索准确率和生成质量上取得显著提升。

关键词:大语言模型;检索增强生成;图检索;上下文排序;检索技术
中图分类号　TP181

Research on Retrieval-augmented Generation Technology Combining Graph Retrieval and Contextual Ranking

XUE Xiaonan
Beijing Youth Political College,Beijing 100102,China

Abstract　Complex question-answering tasks require models to efficiently retrieve relevant information from large-scale heterogeneous knowledge sources while supporting the generation of high-quality answers. However,existing retrieval-augmented generation methods face numerous challenges in knowledge retrieval,semantic relevance,and generation consistency:(1) the granularity and structured information of the knowledge retrieval module are insufficient;(2) there is a lack of contextual relevance in retrieval,limited ranking capability,and constrained generation quality;(3) generative models struggle to accurately integrate retrieved knowledge and produce contextually consistent answers. This paper proposes a novel framework,GraphRank-RAG,which combines graph-based retrieval-augmented generation with contextual ranking to address the issues mentioned. By introducing a graph-based retrieval mechanism,the framework captures deep semantic relationships within contexts,optimizing both the ranking process and answer generation. Experimental results demonstrate that the proposed method outperforms existing approaches on multiple open-domain question-answering datasets,achieving significant improvements in retrieval accuracy and generation quality.

Keywords　Large language model,Retrieval augmented generation,Graph retrieval,Contextual rank,Retrieval technology

随着研究的不断进步,大语言模型(Large Language Model,LLM)在多种自然语言处理任务尤其是问答系统方面展现出了有效性和通用性,在开放领域问答(Open-Domain,QA)、多跳推理(Multi-Hop Reasoning)以及事实验证等任务中取得了显著进展。但在特定领域的精准回复方面,LLM 需要更丰富专业的知识来避免出现幻觉。

检索增强生成方法(Retrieval-Augmented Generation,RAG)是解决上述问题的可能技术之一,它通过检索技术连接外部知识数据库,让 LLM 无需基于专有数据库重新训练,就能够获取最新的信息,并产生可靠的输出。RAG 模型的核心功能包括索引(indexing)、检索(retrieve)、生成(generate);索引是对数据进行存储并建立索引,方便检索;检索根据用户的查询,基于 LLM 检索相关的外部知识数据,这是 RAG 最核心的功能;生成是根据检索匹配内容合并用户查询,返回生成结果,同时还需要注意合并的 prompt 不能超过上下文窗口限制。然而,传统的 RAG 框架在处理复杂任务时,仍面临以下挑战。

1)检索上下文的语义关联不足:当前的检索模型主要基于稠密嵌入或稀疏检索方法,无法有效挖掘跨文档的语义连接,而这些连接往往是多跳推理和复杂问答任务的关键。

2)上下文排序能力有限:在 RAG 框架中,检索模块通常生成大量候选上下文(Top-N),生成模块则需要从中选择 Top-K 上下文进行推理。然而,这种解耦流程往往导致上下文排序不准确,从而降低生成质量。

3)统一优化困难:检索、上下文排序与生成过程分离,导致各模块难以协同优化,尤其在多领域与多任务场景下表现出较低的泛化能力。

本文提出了一种 GraphRank-RAG 框架,结合了图检索增强生成(Graph)和上下文排序与生成能力(Rank),实现了检索、排序和生成的一体化优化。该方法不仅能够捕获跨文档的语义关联,还能通过统一指令微调(Instruction Tuning)提升模型的任务泛化能力。实验证明,本文提出的方法在

通信作者:薛晓楠(xuexiaonan@126.com)

多个开放域问答数据集上取得了显著的性能提升。

该方法虽然在许多任务中表现良好，但在复杂问答任务中依然存在以下主要挑战。

异构知识检索：传统检索方法（如文本检索）难以充分利用知识图谱中的结构化信息。

长文档处理：问题常涉及长文档中的多段上下文，生成模型难以处理超出其上下文长度限制的输入。

知识生成一致性：检索到的知识常包含冗余或噪声信息，生成模型难以有效整合，可能导致幻觉现象（生成虚假或不相关内容）。

1 检索增强生成

RAG 方法通过检索模块获取外部知识[1]，并将其作为大语言模型的输入以生成答案，能够有效弥补 LLM 的诸多缺陷，如幻觉[2]、信息陈旧[3]、长记忆[4]等。RAG 的核心包括检索器和生成器。通常，RAG 首先使用外部检索器来检索相关的文本来自一个特定知识源（如维基百科）的知识，然后将相关的文本知识视为外部知识生成基于知识的响应的上下文。传统 RAG 检索通常依赖基于稠密嵌入或稀疏检索（BM25）[5]的方法，稀疏方法对文本内容固有语义特征提取不足，对复杂问题的多跳推理能力有限。为此，研究人员提出了基于模型的文档编码密集检索方法，将查询作为密集向量，有效地表示语义文本内容的特征[6]。例如，DPR 使用两个预训练的语言模型对文档进行编码，单独查询，以更细致地理解内容。RECOMP 方法[7]通过压缩检索文档并进行选择性增强，降低推理成本，提高语言模型在各种任务中的性能。L 大型语言模型的零样本 Listwise Reranker 排序器[8]，无需使用任何特定任务的训练数据。

Yu 等[9]提出的 RankRAG，将上下文排序任务与生成任务统一到指令微调框架中，在推理阶段，LLM 先重新排序检索到的上下文，再根据精炼的前 $k(5)$ 生成答案。该框架可以广泛应用于各种知识密集型 NLP 任务，显著提升了检索上下文的相关性和生成质量。然而，RankRAG 依赖独立的初始检索模块，无法有效捕获跨文档的语义关联。

与使用向量数据库检索语义相似文本的基本 RAG 不同，Edge 等提出的 GraphRAG[10]通过结合知识图谱来增强 RAG。知识图谱是一种数据结构，它根据数据间的关系来存储和联系相关或不相关的数据。图检索通过构建节点（文档或段落）与边（语义关系）的图结构，捕获跨文档的隐式关联，如图 1 所示。该方法在多跳推理任务中表现出较好的效果，但与生成模型的结合仍处于探索阶段。

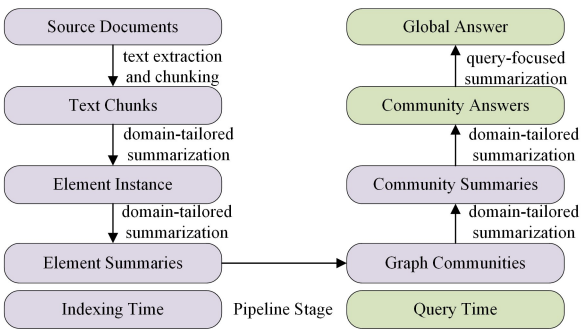

图 1 GraphRAG 结构

Fig. 1 Structure of GraphRAG

2 GraphRank-RAG

GraphRank-RAG 框架结合了图检索和上下文排序与生成的能力，充分利用图结构在捕获语义关联方面的优势，并通过上下文排序与生成进行上下文选择和答案生成，从而优化复杂问答任务中的检索、排序与生成过程。整体方法旨在解决复杂问答任务中检索与生成之间的脱节问题，通过统一优化检索、排序和生成步骤，实现更高效的知识利用和更精确的答案生成。具体来说，GraphRank-RAG 框架分为 3 个模块。

图检索（Graph Retrieval）：将文档集合建模为图结构，通过语义相似度构建节点间的关系边，从中检索与问题相关的上下文子图。

上下文排序（Rank）：对子图中的上下文进行排序，筛选 Top-K 上下文。

答案生成阶段：将问题与筛选后的 Top-K 上下文输入到大语言模型中，生成最终答案。

框架整体流程如图 2 所示。

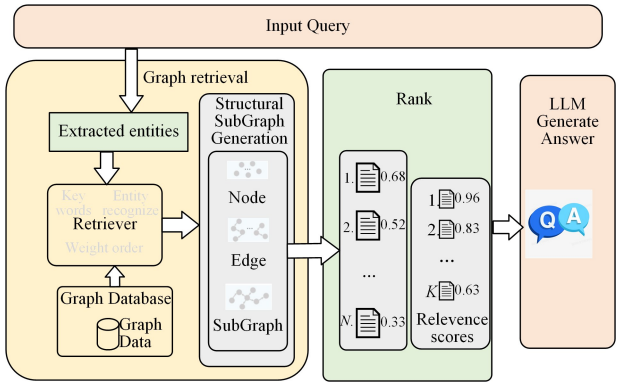

图 2 GraphRank-RAG 流程图

Fig. 2 Flow chart of GraphRank-RAG

2.1 图检索机制

2.1.1 图的构建

为了捕获文档集合中的语义关联，将文档分块后建模为语义图。

节点（Nodes）：每个节点对应一个文档分块（chunk），可以是段落、句子或预定义的文本片段。

边（Edges）：边的权重表示节点之间的语义相似性，计算方法包括以下 2 种。

1）语义嵌入相似性：利用预训练语言模型（如 BERT 或 Sentence-BERT）生成节点嵌入，计算余弦相似度 $\omega_{ij}=\cos(E_i,E_j)$。其中，$E_i$ 和 E_j 分别是节点 i 和 j 的嵌入表示。

2）共现关系：如果两个节点在相同的文档或上下文中频繁共现，则赋予其更高的权重。

2.1.2 动态子图生成

在检索阶段，给定问题 q，从全局语义图中提取与 q 相关的子图。

1）查询节点嵌入：将问题 q 映射到与图中节点相同的嵌入空间。

2）图搜索算法：使用 PageRank 算法衡量节点的重要性，根据节点与问题的相似性调整初始分数。使用最短路径搜索提取与问题相关的高关联节点。

3)子图提取:从全局图中提取 Top-N 个相关节点,构成候选上下文子图。

2.1.3 图的动态更新

为了适应知识的动态变化,我们设计了基于定期重建和增量更新的机制。

1)定期重建:定期重新计算全局图的节点和边。

2)增量更新:当新增文档或知识时,仅更新受影响的局部子图。

2.2 上下文排序与生成

在图检索阶段生成的候选上下文子图中包含大量无关或冗余的段落,我们引入 Rank 模块对上下文进行排序,并进一步生成答案。

2.2.1 上下文排序

上下文排序通过指令微调训练模型对上下文进行相关性评估,具体包括以下方面。

1)任务定义:让模型判断某个段落是否与问题相关,并生成 True 或 False 作为结果。

2)训练数据:使用真实 QA 数据(如 SQuAD 和 HotpotQA)标注段落相关性。

使用图检索生成的候选上下文构建"相关-不相关"对,进一步增强排序能力。

3)训练目标:模型收三元组(q,c,y),其中 q 为问题,c 为上下文,y 为目标输出(True 或 False)。

优化交叉熵损失函数:

$$\mathcal{L}=-\sum_{i=1}^{N}y_i\cdot\log(\hat{y}_i)+(1-\hat{y}_i)\log(1-\hat{y}_i)$$

其中,$\hat{y}_i$ 为模型对上下文的相关性预测。

2.2.2 筛选 Top-K 上下文

排序完成后,按照相关性分数对上下文进行排序,并选择 Top-K 段落用于生成阶段。选择的 K 值由任务需求和实验调优决定(典型值为 5 或 10)。

2.2.3 答案生成

在生成阶段,将问题 q 和排序后的 Top-K 上下文输入到大语言模型(如 Llama3 或 GPT-4)中进行答案生成。

1)输入格式化:将问题和上下文按照以下模板组织。

问题:{q}

上下文 1:{c1}

上下文 2:{c2}…

上下文 K:{cK}

答案:

2)生成目标:模型根据上下文生成答案,同时指明答案无法从上下文中得出时返回"无法回答"。

2.2.4 动态生成优化

在多轮对话场景中,本文允许生成阶段动态调整上下文权重或触发新的检索请求,以进一步提高答案质量。

2.3 统一指令微调

本文设计了一个两阶段指令微调框架,以同时提升模型的图检索、上下文排序与生成能力。

阶段一:基础任务微调

使用标准 QA、阅读理解(如 SQuAD 和 DROP)等数据集,提升模型的基础生成能力。

阶段二:GraphRank-RAG 微调

排序任务数据:通过图检索生成的上下文对,训练模型判断段落相关性。

生成任务数据:结合图检索和排序结果的 QA 数据(如 HotpotQA 和 TriviaQA),训练模型生成准确答案。

多任务联合训练:将排序和生成任务统一为三元组格式(q,c,y),通过共享参数实现知识迁移。

2.4 推理流程

在推理阶段,GraphRank-RAG 流程如下。

图检索:根据问题 q,从全局图中提取相关上下文子图。

排序:使用上下文排序对子图中的段落进行相关性评估和排序,筛选出 Top-K 段落。

生成:将问题和筛选的上下文输入到大语言模型中,生成最终答案。

推理流程的整体效率通过以下方式优化。

检索子图的边界约束:优先提取高中心性和高相关性的节点。

并行化排序与生成:排序阶段与生成阶段并行执行,减少推理时间。

3 实验与结果分析

3.1 实验设置与数据集

为了验证 GraphRank-RAG 在复杂任务的性能及有效性,使用如下数据集进行性能评估。数据集覆盖了开放问答、多跳推理和事实验证等知识密集型任务。

1)开放问答(Open-Domain QA)任务。

NQ(Natural Questions):一个针对真实用户问题的问答数据集。

TriviaQA:包含复杂事实性问题的问答数据集。

2)多跳推理(Multi-Hop Reasoning)任务

HotpotQA:需要基于多个文档进行多跳推理的问答任务。

3)事实验证(Fact Verification)

FEVER:基于事实验证的任务,需要验证给定陈述是否正确。

GraphRank-RAG 模型配置方面,图检索部分基于 BERT 嵌入计算节点间的语义相似度,通过 PageRank 算法提取相关上下文子图;排序与生成部分基于 RankRAG 框架,使用 Llama3(8B 参数)作为生成模型的基础。将本文提出的框架与传统检索生成框架 RAG、仅使用图检索模型 Gragph-RAG、仅使用上下文排序与生成模型 Rank-RAG 进行对比,评估指标如下。

EM(Exact Match):生成答案与真实答案的完全匹配度。

Accuracy:在事实验证任务中正确性判断的准确率。

Recall@10:检索到的 Top-10 上下文中包含正确答案的比例。

实验环境使用 2 块 NVIDIA A30 GPU 进行训练和推理,超参数设置方面,检索阶段 Top-N=100,排序阶段 Top-K=10;Rank 排序模块的学习率为 5×10^{-5},生成模块的学习率为 2×10^{-5}。

3.2 实验结果

如表 1 所列,GraphRank-RAG 在开放问答任务中显著优

于其他对比模型，尤其是在 PopQA 中提高了近 10 个百分点，表明其对复杂问题的语义建模能力更强。

表 1 开放问答任务的 *EM* 值

Table 1 *EM* scores for Open-domain QA tasks

(%)

模型	NQ	TriviaQA
RAG	44.2	69.5
RankRAG	47.8	73.1
Graph-RAG	49.5	74.8
GraphRank-RAG	52.6	78.4

多跳推理任务中，GraphRank-RAG 通过图检索捕获上下文间的语义关联，并结合排序进一步优化了上下文选择质量，达到了最优表现。

表 2 多跳推理任务的准确率

Table 2 Accuracy for multi-hop reasoning tasks

(%)

模型	FEVER(*Accuracy*)
RAG	79.1
RankRAG	82.5
Graph-RAG	83.4
GraphRank-RAG	87.2

GraphRank-RAG 在事实验证任务中的表现显著优于基线模型，如表 3 所列。

表 3 事实验证的 *Recall*@10

Table 3 *Recall*@10 for fact verification

(%)

模型	*Recall*@10
RAG	72.8
RankRAG	76.2
Graph-RAG	78.5
GraphRank-RAG	84.7

实验结果表明，GraphRank-RAG 通过结合图检索与上下文排序，显著提升了开放问答、多跳推理和事实验证任务的性能。相较于传统 RAG，GraphRank-RAG 在所有任务中均取得了显著优势。图检索与上下文排序的协同作用在多跳推理和复杂任务中尤为明显。检索召回率的提升，证明了图检索对高质量上下文筛选的重要性。

结束语 本文提出了一种结合图检索与上下文排序与生成的统一框架 GraphRank-RAG，旨在解决复杂问答任务中检索相关性不足、上下文排序能力有限和生成质量受限等问题。通过将文档集合建模为语义图，GraphRank-RAG 能够捕获跨文档的隐式语义关系，并提升上下文排序与生成的质量。实验结果表明，GraphRank-RAG 在多个知识密集型任务中均表现出卓越的性能，该方法在开放问答、多跳推理和事实验证等任务上表现优异，为知识密集型任务中的大语言模型提供了新思路。

GraphRank-RAG 框架尽管在多个任务中取得了优异的表现，但仍有一些问题和方向值得进一步探索。

1）动态图检索的扩展：现有的图检索机制主要基于静态文档集合构建图。未来可以探索动态图检索，允许在用户交互过程中动态更新图结构，从而适应实时信息变化或连续对话场景。

2）与多轮交互生成的结合：当前的生成模块主要基于单轮问答。未来可以将 GraphRank-RAG 扩展至多轮对话任务，结合多轮上下文检索与排序策略，提升生成的连贯性和上下文理解能力。

3）跨领域任务的泛化能力：本文的实验主要集中在开放领域问答与事实验证任务。未来可以探索 GraphRank-RAG 在垂直领域（如医疗、法律）中的泛化能力，并通过领域特定的图构建和排序优化提升性能。

4）与知识库融合：当前的框架主要依赖非结构化文档的检索与生成。未来可以将 GraphRank-RAG 与结构化知识库（如 Wikidata）结合，增强模型在结构化信息处理和生成中的表现。

尽管如此，与传统 RAG 框架相比，GraphRank-RAG 在检索、排序与生成的协同优化方面展现出了显著的优势，特别是在需要复杂推理和多文档整合的任务中。通过进一步扩展动态图检索、多轮交互生成和跨领域任务应用，GraphRank-RAG 有望在实际场景中发挥更大的潜力，推动检索增强生成技术的发展。

参 考 文 献

[1] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks[J]. arXiv: 2005.11401, 2020.

[2] TONMOY S M T I, ZAMAN S M M, JAIN V, et al. A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models[J]. arXiv: 2401.01313, 2024.

[3] KASAI J, SAKAGUCHI K, YOICHI T, et al. Realtime QA: What's the answer right now? [C]// NeurIPS. 2023.

[4] XU P, PING W, WU X C, et al. Retrieval meets long context large language models[C]// Proceedings of the International Conference on Learning Representations(ICLR). 2024.

[5] ROBERTSON S, ZARAGOZA H. The probabilistic relevance framework: BM25 and beyond[J]. Foundations and Trends © in Information Retrieval, 2009, 3(4): 333-389.

[6] KARPUKHIN V, OĞUZ B, MIN S, et al. Dense passage retrieval for open-domain question answering[J]. arXiv: 2004.04906, 2020.

[7] XU F, SHI W, CHOI E. Recomp: Improving retrieval-augmented lms with compression and selective augmentation[J]. arXiv: 2310.04408, 2023.

[8] MA X, ZHANG X, PRADEEP R, et al. Zero-shot listwise document reranking with a large language model[J]. arXiv: 2305.02156, 2023.

[9] YU Y, PING W, LIU Z, et al. Rankrag: Unifying context ranking with retrieval-augmented generation in llms[J]. arXiv: 2407.02485, 2024

[10] EDGE D, TRINH H, CHENG N, et al. From local to global: A graph rag approach to query-focused summarization[J]. arXiv: 2404.16130, 2024.

XUE Xiaonan, born in 1989. His main research interests include AI and cybersecurity.