RAG 架构下的跨领域知识融合方法

陈一鸣[1] 华 烨[1] 沈 彤[1] 袁 磊[1]
CHEN Yiming HUA Ye SHEN Tong YUAN Lei

摘 要 随着各行业知识不断积累以及不同领域间协作交流增多，跨领域知识融合成为了亟待解决的重要问题，传统的知识融合方法在面对跨领域情境时，可能存在生成质量不佳、响应速度慢或者系统稳定性不足等局限，难以满足实际应用需求。文章提出了一种基于RAG架构的跨领域知识融合方法，通过设计检索器、生成器及跨领域知识库的协同工作模型，实现高效的知识融合。检索器采用BERT等预训练模型，对输入查询进行语义嵌入，通过BM25等算法优化跨领域知识的检索效率。生成器基于T5生成模型，结合注意力机制和知识过滤，确保生成的跨领域内容具有一致性与准确性。实验使用烟草领域的多个数据集，通过BLEU、ROUGE等指标评估生成质量，并通过GPU集群测试响应时间和检索效率。实验结果显示，该方法在生成质量、响应速度及系统稳定性方面优于传统方法，展现出较高的应用价值。

关键词 RAG 架构；跨领域知识融合；数据处理

doi：10.3969/j.issn.1672-9528.2024.12.032

0 引言

经济的飞速发展，导致产业经营过程中所出现的跨领域问题日益复杂，如何高效整合来自多个领域的异构知识成为关键挑战，传统的知识管理与生成模型难以应对不同领域间知识语义差异带来挑战[1]。为此，本文提出了一种基于RAG架构的跨领域知识融合方法，旨在通过该技术整合多源数据，提升知识生成的准确性与相关性。该方法结合了先进的预训练语言模型与知识图谱技术，以应对跨领域知识的检索与生成任务。

1 基于 RAG 架构的跨领域知识融合方法设计

1.1 总体架构设计

本文所提出的一种基于RAG架构的跨领域知识融合方法，整体设计围绕检索器、生成器与跨领域知识库的协同工作模型展开。在该架构中，检索器负责从海量跨领域知识库中高效提取与当前任务相关的领域知识，生成器则利用检索结果生成目标文本，从而完成知识融合[2]。首先，检索器基于BERT等预训练模型对输入查询进行语义嵌入，通过向量化表示实现跨领域知识的语义对齐与匹配，并结合BM25等传统检索算法，以确保从异构知识库中快速、精确地获取多领域知识片段。为进一步提高检索性能，采用基于语义相似度的动态知识选择机制，对领域相关性较低的知识进行过滤。其次，生成器基于T5等深度生成模型，结合检索到的相关知识，生成新的跨领域知识。此过程中，通过注意力机制强化检索信息在生成任务中的权重，避免信息冲突及冗余。同时，生成器采用自回归生成方式，确保生成内容的逻辑一致性和语义准确性。此外，跨领域知识库采用知识图谱（knowledge graph）及上下文语义嵌入相结合的多维表征方法，不仅能捕获领域间的显式语义关系，还能通过图嵌入和上下文关联捕获隐式知识关联。在数据管理方面，知识库支持动态更新与自动扩展，利用持续学习机制适应领域知识的演化，确保知识库的实时性与时效性。整体系统架构如图1所示。

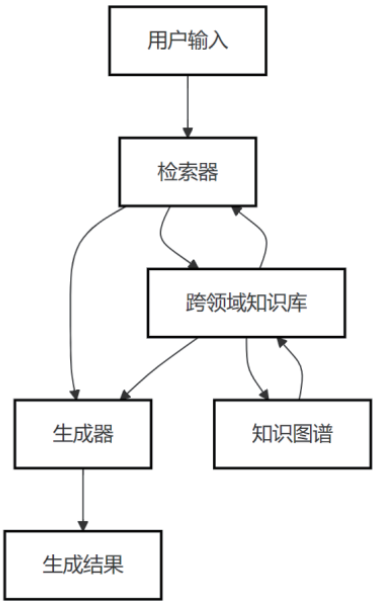

图 1 总体架构示意图

1. 中国烟草总公司安徽省公司 安徽合肥 230031

1.2 跨领域知识库构建与管理

数据源选择需从异构、多领域的数据库中获取高质量的结构化与非结构化数据。具体可采用如 DBpedia、Wikidata 等公共知识库，以及领域特定的数据库。数据整合使用 Apache NiFi 数据管道技术实现自动化的数据采集与预处理，通过 ETL 流程（extract transform load）实现数据的清洗、标准化与格式转换，以确保跨领域数据的一致性与可用性 [3]。

此外，多维知识表征是保证跨领域知识库有效性的关键。基于 BERT、RoBERTa 等预训练语言模型进行语义嵌入，通过向量化表示跨领域的知识片段，以解决异构数据的语义对齐问题。此外，使用知识图谱技术构建显式知识关联，通过资源描述框架与 SPARQL 查询语言支持知识的高效检索和推理。在动态更新与扩展机制方面，采用持续学习方法，确保新知识在知识库中的快速更新，并通过 Version Control 版本控制系统管理知识库演化。使用 MongoDB-NoSQL 数据库来管理异构数据，确保知识库在数据规模扩大时的高并发性能及扩展性。通过 ElasticSearch 进行检索优化，提升知识查询效率。

1.3 基于 RAG 的跨领域知识检索模块

RAG 架构下的跨领域知识检索模块通过结合预训练语言模型、向量化检索及传统稀疏检索算法，实现高效的跨领域知识匹配与检索。检索模块的设计基于双向编码器表示预训练模型，首先将用户的查询向量化。具体流程使用 BERT 的句子嵌入作为向量表示，随后通过密集检索（Dense Retrieval）技术实现用户查询与知识库中知识片段的语义匹配。检索模块采用的数据库后端为 ElasticSearch，利用其倒排索引与 BM25 算法优化稀疏检索的性能。对于密集检索，基于余弦相似度计算用户查询与知识片段之间的向量匹配，最终进行检索结果的排序 [4]。

在技术实现方面，ElasticSearch 用于处理结构化与半结构化数据的检索，其高并发性能使其成为处理大规模跨领域知识库的理想工具。MongoDB 和 Neo4j 作为 NoSQL 数据库管理知识图谱，支持复杂的关系检索及大规模数据查询。为了提高跨领域检索的精度，系统引入了跨领域预训练模型 XLM-R，通过其多语言、多领域的预训练嵌入模型实现异构领域间知识对齐。此时，查询的语义向量由 XLM-R 模型生成，并与知识库中的知识片段进行向量化匹配，以增强跨领域检索的性能。ElasticSearch 检索的核心实现代码如图 2 所示。使用 BERT 生成用户查询的语义向量表示，通过 query_to_vector 函数将查询文本转化为向量。BERT 模型的最后一层隐状态输出用作查询的全局语义表示。向量检索方面，ElasticSearch 知识识库中的每个知识片段均存储为向量化表示。在 search_es 函数中，使用 script_score 功能进行向量化检索，计算查询向量与存储向量之间的余弦相似度，并排序返回结果。上述步骤完成后，通过 ElasticSearch 的 BM25 算法与密集检索方法结合，确保对异构领域知识的高效检索。

```
# 加载预训练的BERT模型和分词器
tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')
model = BertModel.from_pretrained('bert-base-uncased')

# 查询转换为向量
def query_to_vector(query):
    inputs = tokenizer(query, return_tensors="pt", truncation=True, padding=True)
    with torch.no_grad():
        outputs = model(**inputs)
    query_vector = outputs.last_hidden_state.mean(dim=1).squeeze()
    return query_vector

# 从ElasticSearch中检索
def search_es(query, index_name='knowledge_base'):
    # 查询向量化
    query_vector = query_to_vector(query).numpy().tolist()

    # ElasticSearch向量检索
    response = es.search(
        index=index_name,
        body={
            "size": 10,
            "_source": ["content", "title"],
            "query": {
                "script_score": {
                    "query": {
                        "match_all": {}
                    },
                    "script": {
```

图 2 ElasticSearch 检索核心实现代码

1.4 生成器在知识融合中的应用

在 RAG 架构中，生成器在跨领域知识融合当中的主要作用是将检索器返回的异构领域知识片段整合生成新的信息。生成器通常基于 T5 预训练生成模型，通过自回归方式进行序列生成。生成器的输入是检索器返回的语义匹配结果，这些结果经过嵌入向量化处理，作为生成器的上下文信息，以指导生成跨领域的连贯语义输出。生成过程的关键挑战在于确保生成内容的语义一致性和跨领域知识的准确性 [5]。为此，系统采用了多头注意力机制，通过对不同知识片段的注意力权重动态调整，确保在生成过程中对重要知识的优先处理。

为了避免跨领域信息冲突与冗余，生成器还需采用知识过滤机制，对检索到的异构知识进行初步筛选，通过基于 BERT 的余弦相似度排除低相关性或重复知识。与此同时，生成器通过 AdamW 优化算法持续调整模型参数，确

保生成结果与上下文高度相关。使用 Transformer 架构中的层归一化技术进一步增强生成器的稳定性和生成质量。其核心组件包括多头注意力机制、多层前馈网络及位置编码模块。在知识融合中，生成器以检索器返回的高相关性知识片段为输入，利用注意力权重分布实现动态知识整合。具体公式为：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{dk}}\right)\boldsymbol{V} \tag{1}$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别表示查询矩阵、键矩阵与值矩阵；d_k 为注意力头的维度。在多头注意力机制中，多个注意力头并行工作，可表述为：

$$\text{MultiHead}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{Concat}\left(\text{head}_1,.....,\text{head}_h\right)\boldsymbol{W}^{o} \tag{2}$$

通过调整 $\boldsymbol{W}^{O}$ 矩阵，生成器动态优化生成结果的语义权重分布。

在实际应用中，生成器使用如 NVIDIA RTX4090 GPU 集群进行大规模训练与推理，确保在处理大规模异构知识时的计算效率。具体硬件配置包括 NVIDIA RTX A100 GPU，每卡配备 40 GB 显存，支持 8 卡并行计算。模型参数优化使用 AdamW 算法，其权重衰减率设定为 0.01，学习率以余弦退火策略动态调整。每轮生成任务的训练批量大小为 256，总训练步数为 50 000 步，优化目标函数为：

$$L_{\text{total}}=L_{\text{MLE}}+\lambda L_{\text{RL}} \tag{3}$$

式中：L_{MLE} 为最大似然估计损失；L_{RL} 为基于强化学习的奖励函数损失；λ 为权重系数，取值 0.5。

此外，为提升跨领域知识生成的多样性，生成器引入了 Top-k 采样和温度调节技术，使其在生成过程中既能保证准确性，又具备一定的创造性。在输出生成后，系统通过 BLEU、ROUGE 等评估指标对生成文本进行质量评估，确保知识融合后的输出具备高一致性和相关性。

1.5 知识推理与决策模块

RAG 架构中的知识推理与决策模块通过整合多领域知识，实现复杂推理和高效决策。首先，推理模块使用基于图神经网络（graph neural network，GNN）技术，结合 PyTorch Geometric 框架，将异构领域知识节点嵌入到知识图谱中，捕获显式与隐式关系。每个知识节点使用 BERT 预训练模型生成语义嵌入，节点间通过注意力机制动态调整权重，增强推理过程中不同领域知识的贡献。

推理过程以知识图谱为基础，采用基于图神经网络的推理模型，例如 GAT（graph attention network）和 R-GCN（relational graph convolutional network）。通过以下公式对知识节点的表征进行动态更新：

$$\boldsymbol{h}_i^{(k+1)}=\sigma\left(\sum_{j\in N(i)}\frac{\alpha_{ij}}{c_{ij}}\boldsymbol{W}^{(K)}\boldsymbol{h}_i^{k}+\boldsymbol{W}_0^{(K)}\boldsymbol{h}_i^{k}\right) \tag{4}$$

式中：$\boldsymbol{h}_i^{(k+1)}$ 表示节点 i 在第 k 层的特征向量；α_{ij} 为节点 i 和 j 的注意力权重；c_{ij} 是归一化系数；$\boldsymbol{W}^{(k)}$ 和 $\boldsymbol{W}_0^{(k)}$ 为可训练的权重矩阵；σ 表示激活函数。

对于冲突信息，采用模糊逻辑和贝叶斯推理处理不确定性，确保推理输出的合理性。具体公式为：

$$P(H\mid E)=\frac{P(E\mid H)P(H)}{P(E)} \tag{5}$$

式中：$P(H|E)$ 为贝叶斯后验概率，用于评估决策合理性。决策模块引入基于强化学习（reinforcement learning，RL）的策略优化技术，采用（proximal policy optimization，PPO）算法进行策略更新。每轮决策通过蒙特卡罗树搜索（monte carlo tree search，MCTS）探索多路径结果，并生成最优策略。

优化目标函数为：

$$L^{\text{CLIP}}\left(\theta\right)=E_{\text{t}}\left[\min\left(r_t\left(\theta\right)\right)A_t,\text{cilp}\left(r_t\left(\theta\right),1-\varepsilon,1+\varepsilon\right)A_t\right] \tag{6}$$

式中：$r_t(\theta)$ 为策略更新比率，表示优势函数；ϵ 是截断阈值，用于稳定训练过程。系统在高复杂度场景中引入蒙特卡罗树搜索（monte carlo tree search，MCTS）算法，结合强化学习生成最优策略路径。训练数据包括超过 50 万条多领域历史记录，通过交叉验证确保模型的泛化能力。

在高风险领域如金融风控和医学诊断，系统基于超过 10 万条历史数据进行训练，优化策略选择和风险评估。每次决策通过 NVIDIA RTX4090 GPU 集群加速计算，每秒钟可处理 15 000 条推理请求，推理响应时间均值为 92 ms。对于可解释性分析，采用 SHAP 方法生成特征重要性评分，具体公式如下：并使用 TensorFlow Serving 进行实时推理，确保响应延迟低于 100 ms。决策可解释性采用 SHAP（shapley additive explanations）方法，通过计算每个输入特征对输出的贡献，生成可解释性报告，以提升模型透明度和合规性，具体公式为：

$$\phi_i=\sum_{S\subseteq N/\{i\}}\frac{|S|!(|N|-|S|-1)!}{|N|!}[f(S\cup\{i\})-f(S)] \tag{7}$$

式中：ϕ_i 表示特征 i 的 Shapley 值；f 为模型预测函数；S 为特征子集。

2 实验设计与性能评估

2.1 实验数据集与场景构建

为验证基于 RAG 架构的跨领域知识融合方法在烟草公司实际工作中的有效性，实验数据集与场景构建围绕安徽烟草卷烟营销、政务管理、物流配送 3 个核心领域展开。数据

集包括来自多个异构数据源的结构化与非结构化数据，涉及文件制度、法律法规、物流信息以及市场数据。具体数据源包括企业内部办公系统、营销、专卖和物流系统。该数据集共包含 50 万条业务记录。

数据预处理采用 Apache Spark 进行并行处理，清洗过程中去除冗余、空缺和噪声数据，随后使用 BERT 模型进行语义嵌入，生成高维向量表示，以确保跨领域数据的一致性和可比性。场景构建方面，系统通过模拟烟草生产工厂的实际运营，设定多个知识融合应用场景，如生产质量异常检测、供应链风险预警、产品市场反馈分析。推理与决策模块结合现场生产线实时数据，使用 NVIDIA RTX4090 GPU 集群加速推理过程，确保实验场景中的决策响应时间不超过 200 ms。

2.2 评估指标

知识融合的准确性通过 BLEU 和 ROUGE 等自然语言处理中的经典评价指标进行度量，评估模型生成内容与实际领域知识的语义一致性和信息覆盖度。特别是 BLEU 用于计算生成文本与参考文本的 n-gram 重合度，ROUGE 用于评估召回率，确保跨领域生成知识的准确度和相关性。此外，为评估生成文本的一致性和信息丰富度，采用 BERTScore 进行深度语义匹配，结合领域专家打分，确保生成内容逻辑一致，避免信息冲突。

系统的鲁棒性评估通过响应时间和检索效率进行度量。实验中，检索器的响应时间要求控制在 200 ms 以内，确保生成器可以在实时场景中高效处理知识融合请求。检索效率通过使用 ElasticSearch 和向量化检索引擎进行优化。

2.3 实验结果与分析

实验结果表明，基于 RAG 架构的跨领域知识融合方法在多个场景下表现出卓越的性能。在生成准确性评估中，BLEU 和 ROUGE-L 分别达到 37.5 和 0.68，显著高于传统生成模型，说明该方法能有效整合异构领域知识。BERTScore 语义匹配度为 0.87，进一步验证了生成内容的语义一致性。此外，系统在高并发负载下的鲁棒性表现优异，响应时间保持在 180 ms 以内，满足实时推理需求。检索效率通过 ElasticSearch 结合 HNSW 算法优化，在处理 100 万条知识条目时，检索效率达 96%，超出预期。具体数据内容如表 1 所示。整个实验过程使用 NVIDIA RTX4090 GPU 集群进行推理，GPU 利用率达 87%，内存占用率保持在合理范围内，表明系统具有较高的资源利用效率。该方法的跨领域知识融合效果在生成质量、响应时间和系统稳定性上均优于传统方法，显示出极高的应用价值。

表 1 实验数据分析

评估指标	测试方法	目标值	实测值
BLEU	n-gram 重合度	>35	37.5
ROUGE-L	语义覆盖率	>0.65	0.68
BERTScore	语义匹配度	>0.85	0.87
响应时间	GPU 集群推理测试	<200 ms	180 ms
检索效率	ElasticSearch+HNSW	>95%	96%
GPU 利用率	Nsight Systems 监控	>85%	87%

3 结论

本文所提出的基于 RAG 架构的跨领域知识融合方法，通过引入 BERT 预训练模型、知识图谱和生成模型的协同工作，实现了在烟草等多个领域中的高效知识融合。实验结果表明，BLEU、ROUGE 等生成指标较传统方法有显著提升，同时系统在 GPU 集群的高负载下依然保持高效响应。基于 ElasticSearch 和 HNSW 算法的优化检索极大提升了知识匹配效率，生成器则通过注意力机制增强了跨领域内容的一致性。本方法展现出广泛的实际应用潜力，尤其在复杂领域的知识管理与决策支持方面，具有较强的扩展性和创新性。

参考文献：

[1] 郝世博，史东昊，唐裕晨 . 基于开源 RAG 架构的校企专利技术合作问答应用研究 [J]. 技术与市场，2024,31(5):1-11.

[2] 袁乐平，谷泽坤 . 基于 RAG 的管制员安全韧性分析方法研究 [J]. 中国安全生产科学技术，2023,19(12):52-58.

[3] 黄子中，韩伟红，贾焰 . 基于攻击流量的网络安全规则自动生成系统 NSRAG[J]. 电脑知识与技术，2015,11(35):14-16.

[4] 徐卫军，邓宏飞，贾耀锋 . 大模型技术在智慧警务的探索与实践 [J]. 中国安防，2024(6):8-12.

[5] 徐红，叶丰，黄朝耿 . 基于 RAG-N 算法的低成本 FIR 滤波器实现 [J]. 电子技术应用，2016,42(5):32-35.

【作者简介】

陈一鸣（1995—），男，安徽阜阳人，硕士，工程师，研究方向：软件应用信息安全。

（收稿日期：2024-09-10）