智能检索中基于生成式模型和伪相关反馈的查询扩展方法 *

秦春秀 [1,2] 吕树月 [1,2] 王玉龙 [1,2] 马续补 [1,2] 李凡 [1,2]

[1] 西安电子科技大学经济与管理学院 西安 710071

[2] 陕西信息资源研究中心 西安 710071

摘 要：[目的 / 意义] 为改善检索系统中伪相关反馈对初检文档集过度依赖和生成式模型未考虑相关文档中潜在扩展项等问题，提出一种基于生成式模型和伪相关反馈的查询扩展模型。[方法 / 过程] 综合生成式模型和伪相关反馈两种方法的优势，分别采用查询生成模型和伪相关反馈生成候选扩展词集，将两种候选扩展词集合并得到最终扩展词集，实现查询扩展。最后，以 NQ 和 TriviaQA 两个标准开放域问答数据集为实验语料，基于密集检索验证所提出查询扩展方法的有效性。[结果 / 结论] 实验结果表明，所提出模型检索结果 Top-k 的检索准确率和 EM 均高于基准方法；另外，测试伪相关反馈查询词数量、生成式模型上下文类别以及问题类别对模型性能的影响，实验结果验证了所提出方法的有效性。所提出方法能够提高查询扩展词质量，改善信息检索性能。

关键词：查询扩展 文本生成 伪相关反馈 信息检索

分类号：G254.3

DOI：10.13266/j.issn.0252-3116.2024.15.010

引用本文：秦春秀，吕树月，王玉龙，等．智能检索中基于生成式模型和伪相关反馈的查询扩展方法 [J]. 图书情报工作，2024, 68(15): 117-127. (Citation: Qin Chunxiu, Lv Shuyue, Wang Yulong, et al. Query Expansion Method Based on Generative Model and Pseudo Relevance Feedback in Intelligent Retrieval[J]. Library and Information Service, 2024, 68(15): 117-127.)

1 引言 /Introduction

智能检索旨在借助人工智能程序理解用户查询意图，并以自然语言或非语言形式提供答案。信息检索在此过程中扮演着至关重要的角色，其核心任务是根据用户提出的查询问题检索相关文档并将其呈现给用户 [1]。然而，由于信息检索模型缺乏足够的认知和常识，在执行检索任务时往往难以准确识别用户查询意图，进而影响了信息检索的性能 [2]。查询扩展是解决这一问题的核心技术之一，该技术根据原始查询和相关信息添加查询扩展词或调整原始查询权重，以便获取更加全面、准确的查询扩展候选词项，丰富用户查询语义信息，改善信息检索性能 [3]。

伪相关反馈（pseudo relevance feedback，PRF）作为一种有效的查询扩展方法，它通过从文档抽取相关词项来获取扩展词，能够显著缩小查询词和文档之间的差距，但对于初检反馈文档集的质量要求较高。近年来，随着生成式预训练语言模型（如 GPT、BART[4] 等）在各类自然语言处理任务中的出色表现，查询生成逐渐用于增强信息检索性能 [5-7]。无需外部资源，预训练语言生成模型直接利用模型内部存储的知识，就能从相关上下文中学习并获得启发式线索，从而生成查询扩展候选词项。这在一定程度上解决了伪相关反馈对初检文档集的依赖性问题 [7]。然而，预训练语言生成模型仅利用数据集中问题对应的答案、标题、句子作为查询的上下文文档进行训练，未充分挖掘其他相关文档中潜在的查询扩展信息。

为改善伪相关反馈对初检文档集过度依赖和生成式模型未考虑相关文档中潜在查询扩展项的问题，进一步提高模型对用户查询意图识别的准确性，改善信息检索系统的信息检索性能，笔者综合利用生成式模型和伪相关反馈的优势，提出一种基于生成式模型和伪相关反馈的查询扩展模型（query expansion model based on generative model and pseudo relevance feedback, GPRF）。该模型分别采用查询生成模型和

* 本文系国家社会科学基金重点项目“场景驱动的我国关键核心领域文献资源精细组织与精准服务模式研究”（项目编号：22ATQ002）研究成果之一。

作者简介：秦春秀，教授，博士，博士生导师，E-mail: cxqin@xidian.edu.cn；吕树月，博士研究生；王玉龙，博士研究生；马续补，教授，博士，博士生导师；李凡，博士研究生。

收稿日期：2024-01-08 修回日期：2024-02-29 本文起止页码：117-127

伪相关反馈生成候选查询扩展词集，将两种候选扩展词集合并得到最终扩展词集，实现查询扩展。

2 相关研究 /Related work

2.1 伪相关反馈查询扩展

在信息检索领域，PRF 查询扩展技术被广泛视为提高信息检索性能的有效手段 [8]。该方法首先使用原始查询作为检索词来获取初检文档集。接着，假定初检文档集中排名前 n 篇文档包含符合用户查询意图的反馈信息，将前 n 篇文档视为伪相关反馈文档集，采用 TF-IDF 等方法从伪相关反馈文档集中提取查询扩展词。最后，基于查询扩展进行二次检索 [9]。PRF 查询扩展方法有助于补充原始查询检索结果中缺失的答案，通过两次检索实现信息检索性能的优化 [10]。

近年来，针对 PRF 查询扩展的研究主要集中在查询词处理技术和反馈文档质量两个方面 [11]。对于查询词处理的研究主要分为基于向量空间模型、距离和位置信息等多种类型。例如，1971 年 J. J. Rocchio 首次在检索系统中引入一种基于文本的查询扩展方法，按照伪相关反馈文档中词项权重抽取前 k 个词项作为查询扩展词以优化检索 [9]；Y. Lv 等提出位置相关性模型，通过增加邻近查询词权重来提升查询词的相关性 [12]；B. Aklouche 等利用术语和术语之间共现关系构建术语共现图，研究利用术语共现的全局统计量增强查询扩展项的选择和权重的方法 [13]。

在提高反馈文档集质量方面的研究主要通过计算反馈文档与查询词之间的相关性来提升查询扩展词的质量。例如，2004 年，A. Nasreen 等提出的相关语言模型 RM3（relevance-based language model，RM3）[14]，该模型作为一种基于语言模型的 PRF 方法，具备较强的鲁棒性，能够改善传统 PRF 初检结果质量不高的问题，因此被广泛用于查询扩展模型的改进，并被作为有力的对比模型 [8,11]。因此，笔者采用 RM3 模型进行伪相关反馈查询扩展，通过计算初检文档与查询之间的相关性来增强反馈文档质量，从而提升查询词质量。

2.2 基于深度学习的查询扩展

随着自然语言处理技术的不断发展，深度学习方法在查询扩展的研究中得到广泛应用。已有研究使用深度学习模型精细化候选扩展词项的查询特征表示，以优化 PRF 查询扩展结果。例如，D. Roy 等使用词嵌入构建统计查询扩展方法，其考虑了语义关系和术语之间的组合性 [15]；H. Zamani 等提出利用词嵌入和 PRF 提高查询语言模型有效性的方法 [16]。一些学者还探索了利用神经网络文本处理能力，并结合特定的词项查找规则进行查询扩展研究。例如，余传明等提出一种基于深度学习的查询扩展模型，运用 LSTM 和 CNN 深度学习框架计算检索文档对应的概念相关簇，将原始查询与概念相关簇映射作为扩展词来源 [17]；刘高军等提出一种基于神经网络语义信息的查询扩展模型，利用神经网络深度挖掘能力，将局部可扩展词分布拟合为全局可扩展词分布，在冷门搜索数据等多组实验中验证了其有效性 [18]。这些基于神经网络模型表示和利用特定规则筛选扩展词的方法有效提高了查询词质量。但 PRF 对反馈文档集的质量和数量依赖性问题仍然存在，当这种依赖性对查询扩展过程产生负面影响时，可能导致模型无法正确识别用户查询意图，从而使查询扩展失效或者降低检索性能。

2.3 基于预语言模型的查询扩展

预训练语言模型以其强大的语言理解和生成能力，被认为是提高查询扩展词质量的有效途径 [19]。一方面，预训练语言模型的语言理解能力使其可以有效提取查询和反馈文档的深层文本特征，从而有助于识别用户的查询意图，进而改善反馈文档质量。J.Wang 等基于 BERT 预训练语言模型计算查询语句和反馈文档之间的相似度，提出了一种重排机制，通过缩短文本语义之间的距离以改善反馈文档质量 [3]；P. Singh 等则提出具有语义感知的 PRF 框架，使用 BERT 对原始查询变体进行语义表示，最大程度上保留用户查询意图，通过筛选和检查候选词项，从而提高 PRF 查询扩展词的质量 [20]；W. Zhu 等使用 BERT 作为 DPR[21]（dense passage retrieval，DPR）密集检索模型的阅读器，在 DPR 中增加基于 PRF 的上下文关联信息，利用 BERT 从反馈文档中提取用户查询特征和文档的文本特征，将选择的扩展术语添加到原始查询中，所提出的 QCER+DPR 检索模型在检索精度、重排检索精度以及问答性能上均获得提升 [2]。另一方面，生成式预训练语言模型无需二次检索，直接生成查询扩展词以增强检索效果。例如，V. Claveau 等提出使用 GPT-2 生成查询扩展词丰富查询内容 [22]，M. Lee 等仅依靠 CGAN 模型增强查询扩展 [23]；M. Huang 等将 PRF 结果作为 CGAN 的生成器和判别器的训练输入，得到的 CGAN 模型在基于 BM25 的初检结果和基于神经排序模型上均表现出性能提升 [24]；Y. Mao 等提出生成增强检索模型 GAR，在开放域问答数据集上使用多种上下文文本训练 BART 模型，

利用 BART 模型中存储的知识生成查询词项，从而提高检索模型性能 [7]。BART 模型具备双向性和自回归性质，以其独特的模型结构和训练方式同时学习语言表示和生成能力，在文本摘要、机器翻译和问题生成等序列到序列任务中表现出较高的性能 [25]。基于 BART 模型的查询扩展无需二次检索，同时凭借模型强大的语言理解能力和生成能力，通过大规模训练深入理解查询和文档语义特征，从而生成高质量扩展查询词。因此，笔者采用 BART 模型进行查询生成。

综上，生成式模型通过训练从相关上下文中获得启发式线索，无需外部资源和二次检索，直接利用存储在模型中的知识生成查询候选项来增强信息检索模型性能，从某种程度上可以解决伪相关反馈对初检文档集的依赖性问题 [7]。然而，生成式模型训练仅利用数据集中问题对应的答案、句子、标题作为查询的上下文文档进行训练，未充分考虑数据集中相关文档中潜在的查询扩展信息。

3 研究框架与方法 /Research framework and methodology

PRF 查询扩展依赖于伪相关反馈文档集的质量，这种依赖可能会导致查询主题漂移现象的产生，并增加信息检索结果改善的难度。BART 文本生成模型通过从答案、句子、段落等上下文文本中获得启发式查询线索，在一定程度上可以解决 PRF 查询扩展依赖于反馈文档质量的问题 [7]。然而，BART 文本生成模型训练仅利用数据集中问题对应的答案、句子、标题作为查询的上下文，并通过训练使模型从中获得启发式线索来生成查询扩展词，未充分考虑用户查询过程中相关文档中潜在的查询扩展信息。因此，为了进一步挖掘用户查询扩展信息，笔者综合 BART 查询生成模型和 PRF 查询扩展的优势，提出基于生成式模型和 PRF 的查询扩展模型 GPRF。

3.1 GPRF 模型结构

GPRF 模型的基本思想是利用伪相关反馈文档中的信息来填补查询扩展中忽略的相关文档内容，同时，利用预训练语言模型存储的知识来扩展查询，从而缓解 RPF 对初始文档集过于依赖的问题。具体而言，利用 BART 文本生成模型从问题相关的上下文中获取启发性线索，用于生成查询扩展词 [7]；利用 RM3 伪相关反馈方法来挖掘相关文档中潜在的查询扩展词 [14]。图 1 展示了 GPRF 模型的基本架构。在该架构中，首先，分别将查询式 Q 输入到 PRF 查询扩展模型 RM3 和查询生成模型 BART 中；其次，RM3 输出伪相关反馈候选扩展词集 CETS-PRF（CETS from pseudo relevance feedback，CETS-PRF），BART 输出生成候选扩展词集 CETS-GEN（CETS from generative model，CETS-GEN）；再次，合并 CETS-PRF 和 CETS-GEN 两个扩展词集获得最终候选扩展词集（candidate expansion term set，CETS）；最后，CETS 与 Q 合并形成新的查询 Q′ 进行二次检索。

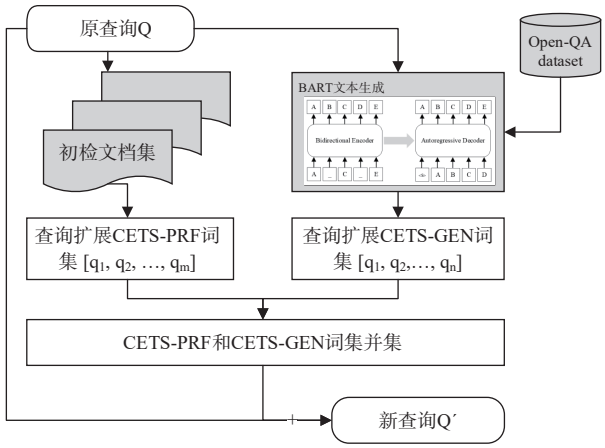

图 1 GPRF 查询扩展模型框架
Figure 1 GPRF query expansion model

3.2 伪相关反馈查询扩展模型

笔者采用 RM3 进行伪相关反馈查询扩展，该方法为基于语言模型的伪相关反馈技术，通过计算词项在伪相关反馈文档集中的概率分布来优化查询扩展 [14,26]。RM3 查询扩展的计算步骤为：

（1）初次检索与伪相关反馈。用户提交初始查询 Q，系统进行 BM25 初次检索并返回排名前 n 篇的文档作为伪相关反馈文档集。

（2）构建伪相关反馈文档集合的语言模型。针对伪相关反馈文档集合中的每个词语 t，建立语言模型以表示其词语分布。该模型中，词语 t 在伪相关反馈文档集合 D_{pseudo} 中的概率分布如公式（1）所示：

$$P\left(t \mid D_{pseudo}\right)=\frac{c\left(t, D_{pseudo}\right)+\mu \cdot P\left(t \mid C\right)}{\left|D_{pseudo}\right|+\mu} \quad \text{公式（1）}$$

其中，$c(t,D_{pseudo})$ 表示词语 t 在伪相关反馈文档集合中出现的次数，$|D_{pseudo}|$ 表示伪相关反馈文档集合的总词数，$P(t \mid C)$ 为词语 t 在整个语料库 C 中的全局概率分布，μ 是平滑参数。

（3）词语权重更新。针对初始查询中的每个词语 t，通过公式（2）计算其在伪相关反馈文档集合语言模型中的加权平均，并得到更新后的权重 $P(t \mid Q_{expanded})$。

$$P(t|Q_{expanded}) = \lambda \cdot P(t|Q_{original}) + (1-\lambda) \cdot P(t|D_{pseudo})$$

公式（2）

其中，$P(t \mid Q_{original})$ 表示词语 t 在初始查询 $Q_{original}$ 中的概率，$P(t \mid Q_{expanded})$ 表示更新后的词语 t 在扩展查询 $Q_{expanded}$ 中的概率，λ 用于平衡初始查询和伪相关反馈文档集合语言模型的权重参数。

（4）查询扩展。利用更新后的词语权重对初始查询进行扩展，汇集高权重但未出现在初始查询中的词语，形成 CETS-PRF。

3.3 基于 BART 的查询文本生成模型

BART 是一种基于 Transformer 结构的序列到序列（sequence to sequence，Seq2Seq）模型，通过自监督的方式进行预训练，利用 Seq2Seq 在预训练任务中通过噪声注入和自动编码器重构的方式进行。这种方法使得 BART 在摘要、对话生成、抽象问答、机器翻译、文本重构等自然语言生成任务上效果拔群[4]。

基于 BART 查询文本生成模型集成 BERT 和 GPT 的优点，建立在标准 Seq2Seq Transformers 模型的基础上。如图 2 所示，左侧编码器采用 BERT 双向特征表示编码器，BERT 是一种 Auto-Encoding 自编码语言模型，可以看作 Transformers 模型中的编码器部分，该模型基于 Masked Language Model 任务训练获得语言理解能力。右侧解码器采用 GPT 自左向右单向特征解码器，GPT 作为一种自回归语言模型，可以看作为 Transformers 模型的解码器部分，负责完成 Text Inifilling 文本重建任务，其优化目标是标准的语言模型目标，计算序列中所有 Token 的联合概率，采用自然序列中从左到右或者从右到左的因式分解。BART 相较 BERT 更适合文本生成的任务场景，相比 GPT 增加了双向上下文语境信息。

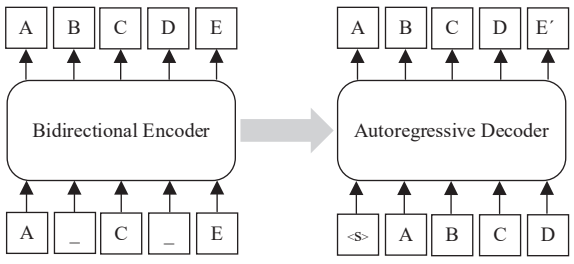

图 2 BART 查询生成模型结构

Figure 2 BART query generation model structure

模型训练过程中，由问题与开放问答数据集中对应的上下文组成序列，如问题—答案（question-answer）、问题—标题（question-title）、问题—句子（question-sentence），通过增加上下文文本来训练 BART 模型。使得查询在模型的解码过程中根据获得的启发式线索生成查询扩展词集 CETS-GEN，与 CETS-PRF 合并去重产生最终候选扩展词集 CETS，与原始查询 Q 共同构成新查询 Q′ 进而进行二次检索。如图 2 所示，以问题—答案序列对为例，在训练过程中，ABCD 代表输入模型的查询问题，E 为答案，答案以问题的上下文形式补充查询问题后，共同完成模型训练。在解码过程中，输入问题 ABCD，模型生成对应的答案 E′ 作为查询扩展词 CETS-GEN。

4 实验设计与结果分析 /Experiment design and results analysis

4.1 实验设计

笔者通过相关段落检索来评估所提出模型的有效性，分为 4 个实验：① GPRF 与基线模型的检索性能对比实验；②在 GPRF 模型中，RM3 查询扩展词数量对结果的影响；③在 GPRF 模型中，不同上下文情景 BART 生成查询词对检索结果的影响；④检验 GPRF 在不同问题类别上的检索性能。

4.1.1 实验数据

笔者选取 Natural Questions（NQ）[27] 和 TriviaQA[28] 两个标准开放域问答数据集进行查询扩展和问答检索实验。其中，Natural Questions 问答数据集中每条数据由一个谷歌查询和一个相应的维基百科查询结果示例组成，每个维基百科页面包含一个短答案和对应的注释长段落。TriviaQA 是基于文本的现实问答数据集，包含 662K 篇文档中的 950K 条问题—答案。表 1 展示了本文使用的数据集的训练集、开发集和测试集的数量统计以及问题—答案对示例数据。

文本生成模型使用问题对应的上下文来训练查询生成模型，训练数据集中的答案（answer）、标题（title）和句子（sentence）均来自开放域问答数据集，表 1 中的上下文示例行展示了问题对应的答案、标题和句子示例。其中，答案代表数据集中问题对应的答案，标题代表检索到的包含问题答案的页面标题，句子代表与问题语义高度相关的真实段落。

4.1.2 评价指标

根据已有研究[2,7,21]，笔者使用 Top-k 检索准确率（Top-k Accuracy）和 Exact Match 指标评估模型检索性能。Top-k Accuracy 通过计算检索到的段落包含至少一个答案的问题的比例，量化检索结果中可回答问题的上限，计算方法如公式（3）所示：

表 1　开放域问答数据集基本介绍

Table 1　Basic introduction of open-domain question and answer datasets

数据集	NQ 数据集	TriviaQA 数据集
训练集 / 条	79 168	78 785
开发集 / 条	8 757	8 837
测试集 / 条	3 610	11 313
问题—答案对示例	**Question:** what color was john wilkes booth's hair? **Short answer:** jet-black **Long answer:** Some critics called Booth "the handsomest man in America" and a "natural genius", and noted his having an "astonishing memory"; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair , and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a "muscular, perfect man" with "curling hair, like a Corinthian capital"	**Question:** The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film? **Answer:** The Guns of Navarone **Excerpt:** The Dodecanese Campaign of World War II was an attempt by Allied forces to capture the Italian held Dodecanese islands in the Aegean Sea following the surrender of Italy in September 1943, and use them as bases against the German-controlled Balkans. The failed campaign, and in particular the Battle of Leros, inspired the 1957 novel The Guns of Navarone and the successful 1961 movie of the same name
上下文示例	**Answer:** jet-black **Title:** John Wilkes Booth **Sentence:** Some critics called Booth "the handsomest man in America" and a "natural genius", and noted his having an "astonishing memory"; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair , and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a "muscular, perfect man" with "curling hair, like a Corinthian capital"	**Answer:** Guns of Navarone (disambiguation)；Navarone **Title:** Dodecanese campaign **Sentence:** The Dodecanese campaign of World War II was an attempt by Allied forces to capture the Italian Dodecanese islands in the Aegean Sea following the Armistice with Italy in September 1943, and use them as bases against the German-controlled Balkans. Operating without air cover, the Allied effort was a costly failure, the whole of the Dodecanese falling to the Germans within two months. The Dodecanese campaign, lasting from 8 September to 22 November 1943, resulted in one of the last major German victories in the war

$$\text{Top-k Accuracy} = \frac{N_{CQ}}{N_Q} \qquad \text{公式（3）}$$

其中，Top-k Accuracy 即 Top-k 检索准确率，表示前 k 个检索结果中回答正确的样本比例，N_{CQ} 表示前 k 个检索结果中含有正确答案的查询问题数量，N_Q 表示查询总样本数量。

Exact Match (EM) 是用来衡量问答检索准确性的重要指标，计算经过字符串归一化（如去除冠词和标点符号）处理后系统预测的答案与真实答案完全匹配的比例，计算方法如公式（4）所示：

$$\text{EM} = \frac{N_{EM}}{N_Q} \qquad \text{公式（4）}$$

其中，N_{EM} 表示匹配问题的数量，即系统给出的答案和实际答案在经过预处理后完全相同的问题数量。

4.1.3　对比模型

为了验证笔者所提 GPRF 方法的有效性，选取以下方法作为对比模型。

（1）BM25[29]。该方法根据文档中词语与查询之间的匹配度计算文档相关性得分。实验中基于 Pyserini[30] 实现 BM25[29] 参数设置和检索模型。

（2）BM25+RM3。该模型采用 BM25 作为基础检索模型，利用 RM3 挖掘查询扩展词项 [14]，基于 Pyserini 工具进行实现。

（3）GAR。该模型为使用 BM25 的生成增强检索模型，通过启发式生成相关上下文来增强查询，而无需外部监督或耗时的下游反馈，一定程度上弥补了稀疏检索的语义缺陷，相较于稠密检索，在训练和推理方面更加轻量级且高效 [7]。

（4）DPR。DPR 模型 [20] 由文本编码器和匹配模块组成，聚焦于密集文本表示并进行文本间的相似度计算，相较于传统词袋模型或稀疏检索可以更好地捕捉语义信息，可以用于大规模文本语料库中高效信息检索。

（5）GAR+DPR。该模型结合 GAR 和 DPR 密集检索模型 [7]，其中，GAR 查询关联的启发式线索生成扩展词，而不是完美的答案，生成查询扩展基于密集检索模型问答，获得比 GAR 更好的检索结果。

（6）BM25+DPR。该模型结合了 BM25 稀疏检索和 DPR 密集检索的优势，使用 Pyserini 工具进行实现 [2]。

（7）QCER+DPR。该模型使用 BERT 作为 DPR 的阅读器，在 DPR 中增加基于 PRF 的上下文关联信息，利用 BERT 从 PRF 反馈文档中提取用户查询特征和文档特征，将选择的扩展术语添加到原始查询中 [2]。

4.1.4　实验环境及参数设置

笔者采用 PyTorch1.8.1，Python 3.8(ubuntu18.04)，Cuda 11.1 搭建实验环境，查询扩展部分采用 RTX 3080 Ti(12GB) * 1GPU，12 vCPU Intel(R) Xeon(R) Silver 4214R CPU @ 2.40GHz，密集检索采用 RTX 3080 Ti(12GB) * 2GPU，24 vCPU Intel(R) Xeon(R) Silver 4214R CPU @ 2.40GHz。

查询生成模块采用 BART-large 模型 [4]，分别在问题—上下文序列数据集上进一步训练得到查询生成模型。模型训练完全遵循 GAR[7] 中 BART 的实验部署和参数设置，最大编码长度为 1 024，嵌入向量维

度为 768。在推理过程中，batch_size 设置为 256，生成扩展词数量为 1，生成扩展词存储到 CETS-GEN。

伪相关反馈查询扩展模块采用 Pyserini 工具中的 BM25+RM3 进行初步检索和查询扩展，其中 k1 和 b 参考 Pyserini[30] 设置为 k1=0.9，b=0.4，RM3 平滑参数根据测试情况从 {500~2 000} 中选取，返回初检的前 10 篇文档构成初检文档集，扩展词数量设置为 1。最后，将返回的查询扩展词集 CETS-PRF 补充到 CETS-GEN 中，去重后与原始查询 Q 共同构成新查询 Q′。

二次检索采用 DPR 密集检索模型 [21]，采用 BERT-base 模型 [31] 作为阅读器，嵌入向量维度为 768，最大编码长度为 512，索引采用 DPR 提供的 Wikipedia 索引文件，检索过程中 batch_size 设置为 128。

4.2 实验结果

4.2.1 对比实验结果

为了验证所提出模型的有效性，将本文模型与基线模型进行比较。对比结果见表 2。结果表明：①与基于稀疏检索的模型 BM25、BM25+RM3 和 GAR 相比，GPRF+DPR 采用密集检索的语义表示更加丰富，其在 NQ 数据集上 top-5 至 top-1 000 检索准确率分别为 77.1%、87.7%、93.1%、96.6% 和 97.6%，在 TriviaQA 数据集上 top-5 至 top-1 000 检索准确率分别为 82.9%、91.4%、96.5%、98.8% 和 99.2%，均高于基线模型的检索结果。②与基于密集检索的模型 DPR、BM25+DPR、GAR+DPR 和 QCER+DPR 相比，所提出模型的 top-5 至 top-1 000 检索准确率在两个数据集上同样高于所有基线模型。同当前开放域问答检索性能最好的 SOTA 方法 QCER+DPR 相比，本文模型在 NQ 数据集上 Top-5、Top-20、Top-100 检索准确率分别提升 4.5%、5.5% 和 4.8%，在 TriviaQA 数据集上 Top-5、Top-20、Top-100 检索准确率分别提升 4.8%、8% 和 9.5%。

表 2 本文算法与对比模型的检索性能 Top-k Accuracy 值

Table 2 Top-k Accuracy of our model and baseline models on the test sets

模型		NQ 数据集					TriviaQA 数据集				
		Top-5	Top-20	Top-100	Top-500	Top-1000	Top-5	Top-20	Top-100	Top-500	Top-1000
sparse	BM25	43.6	62.9	78.1	85.5	87.8	67.7	77.3	83.9	87.9	88.9
	BM25+RM3	44.6	64.2	79.6	86.8	88.9	67.0	77.1	83.8	87.7	88.9
	GAR	**60.9**	**74.4**	**85.3**	**90.3**	**91.7**	**73.1**	**80.4**	**85.7**	**88.9**	**89.7**
hybrid	DPR	66.5	79.6	86.2	90.0	91.4	61.6	72.8	81.4	86.6	88.0
	GAR+DPR	70.7	81.6	**88.9**	**92.0**	**93.2**	76.0	82.1	86.6	-	-
	BM25+DPR	71.8	**82.6**	88.6	91.9	92.3	76.0	82.6	86.6	89.1	89.9
	QCER+DPR	**72.6**	82.2	88.3	91.5	92.5	**78.1**	**83.4**	**87.0**	**89.7**	**90.4**
	GPRF+DPR	**77.1**	**87.7**	**93.1**	**96.6**	**97.6**	**82.9**	**91.4**	**96.5**	**98.8**	**99.2**

如表 3 所示，对模型在问答检索任务中的 EM 进行计算和比较。所提出模型在 NQ 数据集上的 EM 分数为 44.8%，在 TriviaQA 数据集上的 EM 分数为 65.6%，相较 QCER+DPR 分别提高了 2.2% 和 4.4%，该结果进一步验证了模型在问答检索任务中的有效性。

表 3 本文算法与对比模型的检索结果 EM 值

Table 3 EM metrics of our model and baseline models on the test sets

模型	NQ 数据集	TriviaQA 数据集
BM25	37.7	60.1
GAR	41.8	62.7
DPR	41.5	57.9
GAR+DPR	43.8	-
BM25+DPR	41.9	59.7
QCER+DPR	42.6	61.2
GPRF+DPR	**44.8**	**65.6**

结合图 3 和图 4 中模型在两个数据集上的检索性能可视化结果可以看出，GPRF+DPR 模型在 TriviaQA 数据集上的提升效果更加显著。该结论同 Y. Mao 等 [7] 的实验结果相似，为了进一步了解模型变化特点，4.2.2 至 4.2.5 实验主要在 NQ 数据集上进行验证。

4.2.2 查询扩展示例

本节对 GPRF 模型的查询扩展实际效果进行示例分析，对比 GAR+DPR 模型和 GPRF+DPR 模型在 NQ 数据集中两个具体问题上的查询扩展及检索结果，验证所提出查询扩展模型的有效性。如表 4 所示，Case1 的 Passage1 中包含正确的答案，Passage2 中新增了除正确答案之外的相关答案信息。Case2 的 Passage1 中不包含正确答案，且段落无法回答 Query，Passage2 中不包含正确答案，但可以回答 Query 且符合实际情况，这表明 GPRF 能够一定程度提升查询扩展词质量和问答检索的性能。

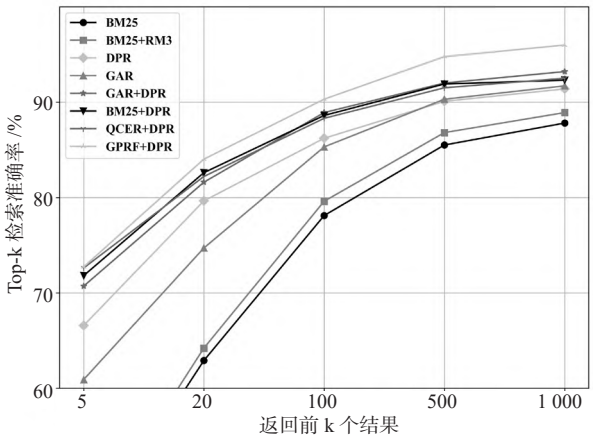

图 3 NQ 数据集上模型检索性能对比

Figure 3 Top-k retrieval accuracy of our model and baseline models on the test set of NQ

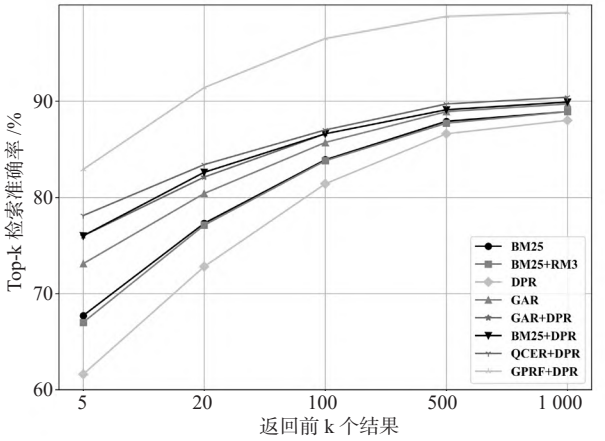

图 4 TriviaQA 数据集上模型检索性能对比

Figure 4 Top-k retrieval accuracy of our model and baseline models on the test set of TriviaQA

表 4 查询扩展示例

Table 4 Query expansion cases from NQ

Case 1		Query	**when is the next deadpool movie being released**
		Answer	“May 18，2018”
	GAR	Expansion term 1	**“Deadpool 2”**
		Passage 1	...\”Deadpool 2\”was released in the United States on May 18, 2018…
	GPRF	Expansion term 2	**“Deadpool 2”，“introduced”**
		Passage 2	...score was released by Columbia Records on May 18, 2018, coinciding with the film’s release. A soundtrack album covering... …A sequel, \”Deadpool 2\”, was released in **May 2018**…
Case2		Query	**who was the first lady nominated member of the rajya sabha**
		Answer	“Mary Kom”
	GAR	Expansion term 1	**“Violet Hari Alva”**
		Passage 1	…in 1948 she was appointed as a member of the senate of ceylon a position she retained until 1952. cooray died on 6 november 1965, at the age of 76. cissy cooray cissy cooray, obe …
	GPRF	Expansion term 2	**“Violet Hari Alva”，“societies”**
		Passage 2	…**vasisht** was elected to the **rajya sabha** (the upper house of the parliament of india) in 1960. her term lasted from 3 april 1960 to 2 april 1966, under the governments of jawaharlal nehru, lal bahadur shastri and indira gandhi. in 2008 vasisht released…

4.2.3 RM3 查询词数量比较实验结果

为了检验不同数量查询扩展词对检索结果的影响，本实验以 NQ 数据集中 nq-test 为测试集，控制 RM3 查询扩展词数目为 1—5，采用 NQ 数据集中问题对应的答案作为上下文信息训练 BART 模型，生成查询扩展词，以测试所提出模型的 top-1 至 top-1 000 检索准确率。表 5 展示了该实验的测试结果，其中 q 表示 RM3 查询扩展词数量。如图 5 所示，根据该实验可视化结果，可以看出 GPRF 模型在 NQ 测试集的检索性能与 RM3 输出查询扩展词数量成正比。另外，随着 k 值增加，检索准确率增加趋势呈现“长尾效应”。如表 5 所示，当 q=5 时，Top-1 000 检索准确率为 99.39%，趋近 100%。可以看出随着扩展词增加，q 增加带来的模型检索性能提升空间逐渐变小。基于上述结果，本文实验设置 q=1，利用伪相关反馈查询扩展带来的检索性能提升的同时可以尽可能减少计算资源消耗。

表 5 不同查询词数量检索结果

Table 5 Retrieval results on different number of query expansion term

q	Top-k 检索准确率 /%									
	1	5	10	20	50	100	200	300	500	1 000
1	50.00	72.71	78.92	84.02	88.09	90.30	92.55	93.57	94.76	95.98
2	50.44	73.35	79.53	85.07	88.92	91.55	94.02%	95.15	96.34	97.48
3	50.94	74.10	80.58	86.26	90.42	93.49	95.84%	96.73	97.53	98.64
4	51.22	74.65	81.36	86.95	91.39	94.43	96.93%	97.65	98.39	99.14
5	51.83	75.54	82.49	88.28	92.77	95.46	97.59%	98.31	98.89	99.39

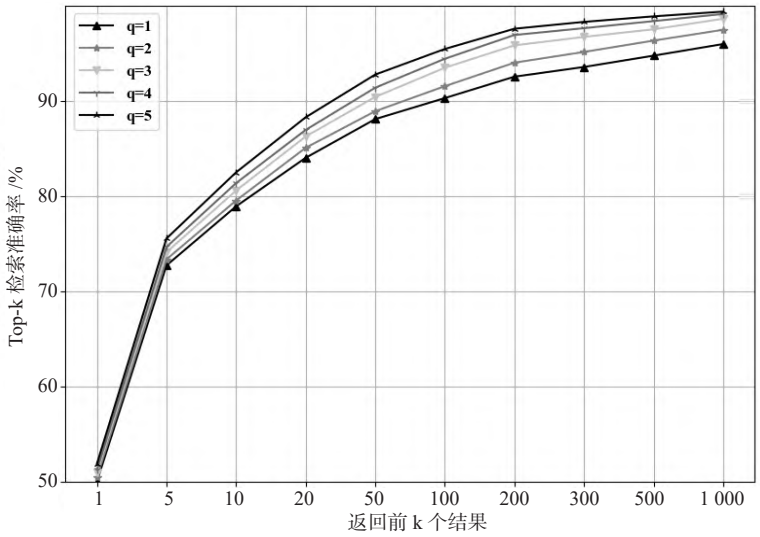

图 5　不同 RM3 查询扩展词数量的测试结果

Figure 5　Different RM3 query expansion words on the test set of NQ

4.2.4　不同上下文情景下模型检索性能

在 GPRF 模型中，BART 查询生成模型可以从不同类型的查询上下文中训练获得启发式线索，生成查询扩展词。本实验将上下文类型分为 answer、title、sentence、answer+title、answer+sentence 和 answer+title+sentence 共 6 类，answer+title、answer+sentence 和 answer+title+sentence 表示将不同上下文类型的扩展词合并。与 4.2.1 实验设置相同，本实验将 RM3 查询扩展词数量控制为 1，测试模型在不同上下文类型下生成查询词对 top-1 至 top-1 000 检索准确率的影响。

结果见图 6，整体上看，answer+title+sentence 上下文类型的检索结果准确率最高。随着 k 值升高，不同类型上下文的检索准确率均呈上升趋势，且差距逐渐缩小。从图 6 和表 6 可以看出，基于单类型上下文生成查询词的检索结果中，title 整体检索准确率高于 answer 和 sentence。k=100 时，title 类型检索准确率为 93.13%，比 answer 和 sentence 类型下检索准确率分别高出 2.83% 和 5.78%。基于多类型查询扩展词融合的检索结果中，answer+title+sentence 拥有较高性能，answer+title 与 answer+title+sentence 类型下模型具备相当的检索能力，当 k ≥20 时，answer+title 与 answer+title+sentence 模型检索准确率基本相同。answer+sentence 类型在 3 种类别中检索准确率从最低，结果与 answer 单类型检索性能基本一致。综合单类型上下文与多类型上下文融合的检索结果来看，title 类型对模型检索能力提升效果最为显著，sentence 类型对模型检索能力的提升效果最小。

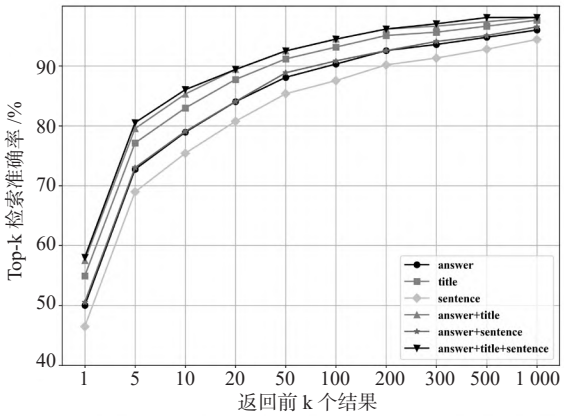

图 6　NQ-test 数据集上生成增强上下文文本类型对检索结果影响结果

Figure 6　Top-k retrieval accuracy on the test set of NQ when fusing retrieval results of different generation-augmented queries

表 6　不同上下文类别检索结果

Table 6　Retrieval results on different context type

上下文类型（context）	Top-k Accuracy/%									
	1	5	10	20	50	100	200	**300**	**500**	**1 000**
answer	50.00	72.71	78.92	84.02	88.09	90.30	92.55	93.57	94.76	95.98
title	54.93	77.12	82.96	87.73	91.16	93.13	95.07	95.62	96.62	97.62
context	46.48	68.95	75.40	80.75	85.35	87.53	90.17	91.30	92.77	94.40
answer+title	57.51	79.53	85.29	89.39	92.49	94.46	96.15	96.62	97.37	98.09
answer+sentence	50.14	72.91	79.06	84.04	88.24	90.43	92.55	93.61	94.76	96.01
answer+title+sentence	57.53	79.61	85.32	89.39	92.49	94.46	96.15	96.70	97.50	98.09

4.2.5 不同问题类型下模型检索性能

为了验证所提出模型在不同问题类型上的检索性能，本实验在 NQ-test 数据集上测试 Top-100 检索准确率，结果见表 7。Type 表示不同查询问题的类别，percentage 表示该问题类型在测试集所有问题中所占数据比例。与两种基础检索模型 BM25、DPR 和生成增强检索模型 GAR、GAR+DPR 相比较可以看出，除 which 这一问题类型外，所提出模型在其他所有类型上均高于对比模型。

表 7 NQ 数据集上不同问题类型下模型 Top-100 检索准确率
Table 7 Top-100 retrieval accuracy of different types of question on the test set of NQ

问题类型（Type）	百分比（Percentage）/%	BM25	DPR	GAR	GAR+DPR	**GPRF+DPR**
who	37.50	82.1	88.0	87.5	90.8	**91.9**
when	19.00	73.1	86.9	83.8	88.6	**90.4**
what	15.00	76.5	82.6	81.5	86.0	**87.5**
where	10.90	77.4	89.1	87.0	90.8	**93.1**
other	9.10	79.3	78.1	81.8	84.2	**87.1**
how	5.00	78.2	83.8	83.2	85.5	**90.2**
which	3.30	89.0	90.7	94.1	94.9	87.0
why	0.30	90.0	90.0	90.0	90.0	**90.9**

5 结语 /Conclusion

本文提出了一种基于生成式模型和伪相关反馈查询扩展的查询扩展方法 GPRF。该模型将原始查询输入 BART 文本生成模型和 RM3 查询扩展模型，分别得到生成候选扩展词集和伪相关反馈候选扩展词集，最后将上述两种候选扩展词集合并得到最终候选扩展词，实现查询扩展。接着，基于 DPR 密集检索提出 GPRF+DPR 模型，设置实验以验证所提出模型有效性。以两个标准开放域问答数据集 NQ 和 TriviaQA 为实验数据，实验结果表明，本文所提出的查询扩展模型是有效的，其在两个数据集上的实验结果 Top5 至 Top1 000 检索准确率和 EM 值均超过基线模型，能够改善信息检索模型的信息检索性能。同时，其在 TriviaQA 数据集上的检索性能提升效果更加显著。

本文所提出查询扩展方法可应用于开放域问答、社交媒体平台智能检索、学术搜索等信息检索任务。不过，本文提出的模型仅在开放域问答数据集上进行训练和验证，未在其他类型信息检索场景中进行实验，实际应用时需要在更多信息检索数据集上进行测试。未来的研究工作中，可以面向具体的使用场景，利用知识背景更丰富的语言模型设计查询扩展模型，从而进一步提高模型的泛化能力和实际应用价值。

参考文献 /References：

[1] LI S, GONG C, ZHU Y, et al. Context-aware multi-level question embedding fusion for visual question answering[J]. Information fusion, 2024, 102: 102000.

[2] ZHU W, ZHANG X, YE L, et al. Query context expansion for open-domain question answering[J]. ACM transactions on Asian and low-resource language information processing, 2023, 22(8): 1-21.

[3] WANG J, PAN M, HE T, et al. A pseudo-relevance feedback framework combining relevance matching and semantic matching for information retrieval[J]. Information processing & management, 2020, 57(6): 102342.

[4] LEWIS M, LIU Y, GOYAL N, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proceedings of the 58th annual meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 7871-7880.

[5] YU S, LIU J, YANG J, et al. Few-shot generative conversational query rewriting[C]//Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval. New York: Association for Computing Machinery, 2020: 1933-1936.

[6] LIN S, YANG J, NOGUEIRA R F, et al. Query reformulation using query history for passage retrieval in conversational search[EB/OL]. [2024-04-05]. https://arxiv.org/pdf/2005.02230v1.pdf.

[7] MAO Y, HE P, LIU X, et al. Generation-augmented retrieval for open-domain question answering[C]//Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing. Stroudsburg: Association for Computational Linguistics, 2021: 4089-4100.

[8] LV Y, ZHAI C. A comparative study of methods for estimating query language models with pseudo feedback[C]//Proceedings of the 18th ACM conference on information and knowledge management. New York: Association for Computing Machinery, 2009: 1895-1898.

[9] ROCCHIO J J. Relevance feedback in information retrieval[M]// SALTON G. The smart retrieval system-experiments in automatic document processing. Englewood Cliffs: Prentice-Hall, 1971:

313-323.

[10] PAN M, PEI Q L, LIU Y, et al. SPRF: a semantic pseudo-relevance feedback enhancement for information retrieval via ConceptNet[J]. Knowledge-based systems, 2023, 274: 110602.

[11] PAN M, HUANG J X, HE T, et al. A simple kernel co-occurrence-based enhancement for pseudo-relevance feedback[J]. Journal of Association for Information Science and Technology, 2020, 71(3): 264-281.

[12] LV Y, ZHAI C X. Positional relevance model for pseudo-relevance feedback[C]// Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval. New York: Association for Computing Machinery, 2010: 579-586.

[13] AKLOUCHE B, BOUNHAS I, LIMANI Y. A discriminative method for global query expansion and term reweighting using co-occurrence graphs[J]. Journal of information science, 2023, 49(1): 183-206.

[14] NASREEN A J, JAMES A, W B C, et al. Umass at TREC 2004: novelty and HARD [C]//Proceedings of the thirteenth text retrieval conference. Amherst: University of Massachusetts Press, 2004: 189.

[15] ROY D, GANGULY D, MITRA M, et al. Word vector compositionality based relevance feedback using kernel density estimation[C]//Proceedings of the 25th ACM international on conference on information and knowledge management. New York: Association for Computing Machinery, 2016: 1281-1290.

[16] ZAMANI H, BRUCE CROFT W. Embedding-based query language models [C]//Proceedings of the 2016 ACM international conference on the theory of information retrieval. New York: Association for Computing Machinery, 2016: 147-156.

[17] 余传明 , 蔡林 , 胡莎莎 , 等 . 基于深度学习的查询扩展研究 [J]. 情报学报 , 2019, 38(10): 1066-1077. (YU C M, CAI L, HU S S, et al. Research on query expansion based on deep learning[J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(10): 1066-1077.)

[18] 刘高军 , 方晓 , 段建勇 . 基于深度语义信息的查询扩展 [J]. 计算机应用 , 2020, 40(11): 3192-3197. (LIU G J, FANG X, DUAN J Y. Query extension based on deep semantic information[J]. Journal of computer applications, 2020, 40(11): 3192-3197.)

[19] CHUANG Y, FANG W, LI S, et al. Expand, rerank, and retrieve: query reranking for open-domain question answering[C]// Findings of the Association for Computational Linguistics: ACL 2023. Stroudsburg: Association for Computational Linguistics, 2023: 12131-12147.

[20] SINGH P, BHOWMICK P K. Semantics-aware query expansion using pseudo-relevance feedback[EB/OL]. [2024-04-05]. https://doi.org/10.1177/01655515231184831.

[21] VLADIMIR K, BARLAS O, SEWON M, et al. Dense passage retrieval for open-domain question answering[C]//Proceedings of the 2020 conference on empirical methods in natural language processing. Stroudsburg: Association for Computational Linguistics, 2020: 6769-6781.

[22] CLAVEAU V. Query expansion with artificially generated texts[EB/OL]. [2024-04-05]. https: //arxiv.org/abs/2012.08787.

[23] LEE M, GAO B, ZHANG R. Rare query expansion through generative adversarial networks in search advertising[C]// Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. New York: Association for Computing Machinery, 2018: 500508.

[24] HUANG M, WANG D, LIU S, et al. GQE-PRF: generative query expansion with pseudo-relevance feedback[EB/OL]. [2024-04-05]. https: //arxiv.org/abs/2108.06010.

[25] CHEN H, DOU Z, ZHU Y, et al. 2022. Enhancing user behavior sequence modeling by generative tasks for session search[C]// Proceedings of the 31st ACM international conference on information and knowledge management. New York: Association for Computing Machinery, 2022: 180-190.

[26] 潘敏 . 基于潜在语义关系的伪相关反馈查询扩展技术研究 [D]. 武汉 : 华中师范大学 , 2019. (PAN M. Research on pseudo relevance feedback query expansion technology based on latent semantic relation[D]. Wuhan: Central China Normal University, 2019.)

[27] TOM K, JENNIMARIA P, OLIVIA R, et al. Natural questions: a benchmark for question answering research[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 452-466.

[28] MANDAR J, EUNSOL C, DANIEL W, et al. TriviaQA: a large scale distantly supervised challenge dataset for reading comprehension[C]//Proceedings of the 55th annual meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2017: 1601-1611.

[29] ROBERTSON S, ZARAGOZA H. The probabilistic relevance framework: BM25 and beyond[J]. Foundations & trends in information retrieval, 2009, 3(4): 333-389.

[30] LIN J, MA X, LIN S, et al. Pyserini: a python toolkit for

reproducible information retrieval research with sparse and dense representations[C]//Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. New York: Association for Computing Machinery, 2021: 2356-2362.

[31] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 4171-4186.

作者贡献说明 /Author contributions：

秦春秀：提出研究问题和研究思路，指导论文修改；
吕树月：设计研究方案，完成实验和论文撰写；
王玉龙：论文修改；
马续补：提出修改意见；
李凡：论文修改。

Query Expansion Method Based on Generative Model and Pseudo Relevance Feedback in Intelligent Retrieval*

Qin Chunxiu[1,2] Lv Shuyue[1,2] Wang Yulong[1,2] Ma Xubu[1,2] Li Fan[1,2]

[1]School of Economics and Management, XIDIAN University, Xi'an 710071

[2]Shaanxi Information Resources Research Center, Xi'an 710071

Abstract: [Purpose/Significance] To address the issues of over-reliance on the original retrieved document collection in pseudo-relevance feedback and the neglect of potential expansion elements in relevant documents by generation-augmented retrieval model in retrieval systems, this paper proposes a query expansion model based on generative model and pseudo-relevance feedback. **[Method/Process]** According to the advantages of both generative models and pseudo relevance feedback, it generated candidate extended word sets using query generative models and pseudo-relevance feedback, respectively. Then, it combined the two sets to obtain the final extended word set, achieving query expansion. Finally, taking NQ and TriviaQA as experimental data, it confirmed the efficiency of the proposed query expansion model using dense passage retrieval. **[Result/Conclusion]** The experimental results demonstrates that the Top-k retrieval accuracy and EM of the proposed model is higher than the baseline ones. In addition, the effects of the number of pseudo-relevance feedback query words, the context category of the generative model, and the question category on the model performance are tested, and the experimental results verify the effectiveness of the proposed method. The proposed model can improve the quality of query expansion words and information retrieval performance.

Keywords: query expansion text generation pseudo relevance feedback IR

*This work is supported by National Social Science Fund of China project "Research on the Fine Organization and Precise Service Mode of Literature Resources in Key Core Areas of China Driven by Scenarios" (Grant No. 22ATQ002).

Author(s): Qin Chunxiu, professor, PhD, doctoral supervisor, E-mail: cxqin@xidian.edu.cn; Lv Shuyue, doctoral candidate, Wang Yulong, doctoral candidate; Ma Xubu, professor, PhD, doctoral supervisor; Li Fan, doctoral candidate.

Received: 2024-01-08 Revised: 2024-02-29 Pages: 117-127