文章编号:1671-9352(2026)03-0066-09 DOI:10.6040/j.issn.1671-9352.1.2024.061

融合关键概念和潜在概念的冗长查询缩略方法

朱铭洋[1,2], 黄于欣[1,2], 余正涛[1,2*]
(1. 昆明理工大学信息工程与自动化学院, 云南 昆明 650500; 2.昆明理工大学云南省人工智能重点实验室, 云南 昆明 650500)

摘要:查询缩略旨在通过简化和精炼冗长的查询输入,保留其中的关键信息来提升检索结果的召回率和准确率。然而,传统方法通常是基于统计或基于预训练模型来提取冗长查询中的关键词作为检索输入,难以应对查询的复杂性(如同义词和多义词),且在保留查询核心内容时容易丢失关键信息。针对以上问题,提出一种融合关键概念和潜在概念的冗长查询缩略方法,将代表查询核心内容的关键概念和对理解查询重要但未明确表达的潜在概念相结合,从而生成更完整和有效的查询。具体而言,首先利用预训练模型来生成简短有效的查询作为关键概念,然后使用伪相关反馈方法从原始查询的相关文档集中挖掘潜在概念,最后,将两者聚合作为最终的查询缩略结果,实现冗长查询检索。实验结果表明,在 Robust2004 数据集上使用密集检索模型评估时,相比基线模型,文中提出的方法在 R@1000 和 NDCG@10 两个指标上分别提高 2.1%和 3.6%。

关键词:信息检索;冗长查询;查询缩略;关键概念;潜在概念
中图分类号:TP391 **文献标志码**:A
引用格式:朱铭洋,黄于欣,余正涛. 融合关键概念和潜在概念的冗长查询缩略方法[J]. 山东大学学报(理学版),2026,61(3):66-74,85.

Method for verbose queries reduction by integrating key and latent concepts

ZHU Mingyang[1,2], HUANG Yuxin[1,2], YU Zhengtao[1,2*]
(1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, Yunnan, China; 2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500, Yunnan, China)

Abstract: Query reduction aims to enhance retrieval recall and precision by simplifying and condensing lengthy queries while retaining key information. Traditional methods often rely on statistical approaches or pre-trained models to extract keywords from lengthy queries for retrieval input. However, these methods struggle with query complexity (e.g., synonym and polyseme) and often lose crucial information. To address these issues, a method integrating key concepts and latent concepts for verbose query reduction is proposed. This approach integrates key concepts representing the core content of the query with latent concepts crucial for query understanding but not explicitly expressed to generate more comprehensive and effective queries. Specifically, pre-trained models generate concise and effective queries as key concepts, while pseudo-relevance feedback methods extract latent concepts from relevant document sets of the original query. Finally, both are combined to form the query reduction for improved retrieval. Experimental results on the Robust2004 dataset using a dense retrieval model show that the proposed method improves R@1000 and NDCG@10 by 2.1% and 3.6%, respectively, compared to baseline models.

Key words: information retrieval; verbose query; query reduction; key concept; latent concept

0 引言

查询缩略(query reduction)指的是将冗长且包含大量无关术语的查询缩减为简洁而准确的形式[1],

收稿日期:2024-09-15; **网络出版时间**:2025-09-04
基金项目:国家自然科学基金资助项目(62266027, U21B2027, U23A20388);云南省科技重大专项资助项目(202302AD080003, 202303AP140008);云南省基础研究重大专项资助项目(202401BC070021);昆明理工大学“双一流”创建联合专项资助项目(202201BE070001-021)
第一作者:朱铭洋(2001—),男,硕士研究生,研究方向为自然语言处理、信息检索. E-mail:969988932@qq.com
*通信作者**:余正涛(1970—),男,教授,博士生导师,博士,研究方向为自然语言处理、信息检索、机器翻译. E-mail:ztyu@hotmail.com

并基于缩减后的查询进行高效检索,从而提高检索结果的召回率和准确率。在实际应用中,查询缩略的过程通常涉及识别和去除查询中停用词和不必要的修饰语,提取能够准确表达查询核心内容的关键词和短语。

目前在查询缩略任务的研究中,主要有传统的统计方法和基于预训练模型提取关键词 2 类方法。1)基于传统的统计方法通常利用统计学方法从原始查询中提取关键特征,如词频、文档频率等来确定每个词的重要性。例如,Campos 等[2]结合了词的大小写形式、词频、词的位置、词与上下文的相关性以及词在不同句子中出现的频率 5 个特征来确定每个词在文本中的重要性。2) 基于预训练模型提取关键词的方法则是利用模型预先学习的语言表示能力,从文本中抽取出最具代表性和语境相关的关键词或短语。例如,Kim 等[1]利用 BERT 模型[3]的编码器计算查询中每个词的重要性得分,得分高于阈值的词汇视为核心术语并将其保留,其余予以剔除。

然而,使用上述方法在进行缩略查询时会面临信息丢失和语义理解的限制。1)该方法主要基于词汇的重要性得分进行缩略,而缩略结果中的语义不连贯破坏了原查询的整体语义结构,无法全面反映查询的核心内容。2)该方法也忽视了查询中的潜在概念(在查询语句背后的潜在主题、内容,而不是直接显露在查询中的具体关键词或短语),因此检索结果的覆盖面不足。

针对上述问题,本文提出了一种融合关键概念和潜在概念的冗长查询缩略方法。通过生成式方法提取原始查询的关键概念,同时关注原始查询的潜在概念。具体而言,首先本文提出基于文本到文本的迁移变换器(text-to-text transfer transformer, T5)的关键概念生成模块来生成简短有效的查询作为关键概念。然后本文提出基于 ColBERT 的潜在概念挖掘模块,采用伪相关反馈方法[4]来挖掘原始查询的词级潜在概念。最后,将两者聚合得到最终查询,将最终查询输入到检索器进行后续的信息检索。本文提出的查询缩略方法在语义理解和核心内容提取的全面性优于当前流行的基于 BERT 模型的关键词提取方法。同时,补充了原始查询中的潜在概念,有助于更全面和准确地理解查询内容。

本文在 Robust2004 数据集上进行多次实验,结果表明所提出的方法在下游任务中表现出色。使用本文方法,检索结果的大部分指标均优于其他基线模型。在 ColBERT 模型下,相比最佳基线模型,文中提出方法得到的检索结果在 MAP、MRR、R@1000 和 NDCG@10 这 4 个指标上分别提高 1.8%、0.6%、2.1%和 3.6%。

1　相关工作

1.1　传统的查询缩略方法

1) 删掉停用词:Huston 等[5]利用统计度量逆查询频率(inverse document frequency, IDF) 对术语进行排名,并决定哪些术语是停用词,通过从查询中删除停用词列表中的所有单词来实现查询缩略。

2) 使用词频或逆文档频率计算术语的权重来提取核心术语:Chaa 等[6]引入了一种新的度量词频-逆查询频率(term frequency-inverse query frequency, TF-IQF),该度量用于增加在查询集中出现较少的术语的权重,并减少在查询集中出现最多的术语的权重。

3) 选取子集:Kumaran 等[7]将缩略问题转化为一个学习问题,根据原始查询的预测质量对所有子集(子查询)进行排序,并选择最顶级的子查询。

但 3 类方法都存在着不同的问题:1) 停用词在特定上下文中是有意义的,将其删除可能会影响查询的准确性;2) 词频和逆文档频率仅仅基于词语在文档集合中的分布情况,而忽略了词语的语义信息,可能导致对于在特定上下文中具有重要意义的词语赋予了过低的权重;3) 确定哪些词语应该保留或删除以构成查询子集可能是一个困难的问题,特别是在处理长查询时,选择不当的子集可能会导致查询失去原本的意图和语义。

1.2　基于图的查询缩略方法

基于图的查询缩略方法是信息检索领域的一种新兴技术,通过将查询和数据表示为图结构,提高查询处理的效率和精度。Rousseau 等[8]提出了一种基于图的关键字提取方法 K-Core,它依赖于正在处理的文档的词图构造,分析此图以提取需要的关键字。Bougouin 等[9]提出了基于图形的关键短语提取方法 TopicRank,

它依赖于文档的主题表示。候选关键短语被聚类成主题,并用作完整图中的顶点。应用基于图的排序模型为每个主题分配显著性分数,然后通过从每个排名靠前的主题中选择一个候选词来生成关键短语。尽管基于图的查询缩略方法在提高查询精度和处理复杂关系方面具有优势,但其计算资源消耗大,且难以处理多义词。

1.3 基于预训练模型的查询缩略方法

近年来,基于神经网络深度学习的查询缩略方法,也得到了广泛关注。这类方法主要是利用预训练模型(如 BERT)来缩略查询,通过预训练模型计算原查询中每个词的重要性得分,将得分高于阈值的词视为核心术语并保留,其余词则删掉。Kim 等[1]提出了一个简单有效的查询缩略框架,使用预训练模型 BERT 来进行核心词提取和子查询选择,再将两者以集成的方式聚合。Podder 等[10]使用预训练的 BERT 模型生成整个查询和单个查询项的密集向量,再基于无监督马尔可夫随机场使模型更加关注上下文中心项。

以上方法在实现冗长查询的检索任务时取得了不错的性能。然而,仍存在以下问题:1) BERT 模型在移除或保留词语时会导致语义不连贯,且常规冗长查询数据集的格式(冗长查询-缩略后查询)不带数字标签,难以用于 BERT 模型的训练;2) 基于预训练模型的查询缩略方法忽视了查询中的潜在概念,导致语义损失、信息丢失从而使检索结果不准确。为此本文提出了融合关键概念和潜在概念的冗长查询缩略方法。

如图 1 所示,在对原查询“What adverse effects have people experienced while taking aspirin repeatedly?(反复服用阿司匹林有什么不良反应?)”使用 AdaptKeyBERT 方法[11]进行缩略时,缩略后的查询为“aspirin, effects, taking (阿司匹林,影响,服用)”,并没有全面地捕获查询的核心内容,没有涵盖“adverse(不良)”、“repeatedly(反复地)”这些在原查询中重要的术语,且缩略后的查询是一些不连贯的关键词。与图 1 中标准的相关文档相比较,返回的文档相关性欠佳。

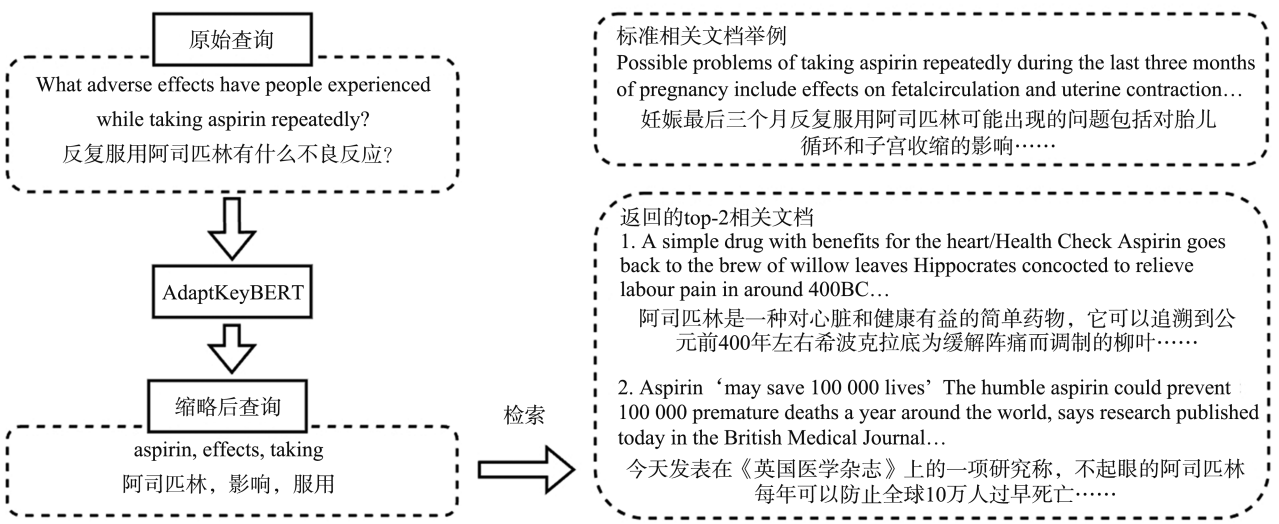

图 1 使用 AdaptKeyBERT 进行查询缩略后检索的示例

Fig.1 Example of using AdaptKeyBERT for query reduction retrieval

2 融合关键概念和潜在概念的冗长查询缩略方法

本文提出了一种融合关键概念和潜在概念的冗长查询缩略方法,该方法分为 2 个并行的步骤。1)针对关键概念:通过微调 T5 模型来生成简短有效的查询,这更倾向于一种生成式方法,不是简单地提取输入文本中的核心术语,而是根据输入文本的语义内容生成一个作为关键概念的简短查询。2)针对潜在概念:使用伪相关反馈方法挖掘原始查询的潜在概念,选择初始检索结果中排名靠前的 k 个文档中逆文档频率最高的 n 个术语作为原始查询的潜在概念。最后,将 2 个步骤得到的结果进行聚合,将每个查询的关键概念和潜在概念以相同的权重进行拼接串联处理,得到最终的缩略查询。图 2 详细展示了该方法的步骤及流程。

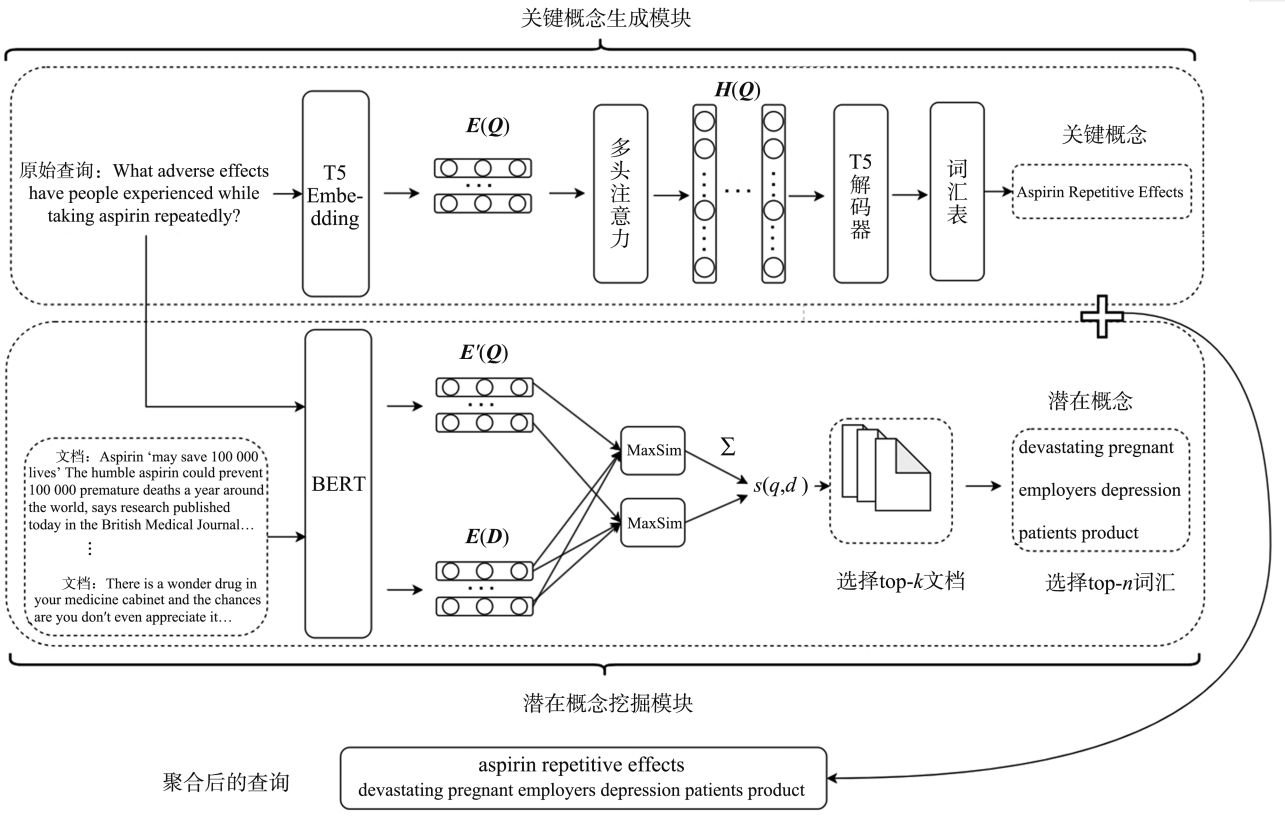

图 2　融合关键概念和潜在概念的冗长查询缩略方法

Fig.2　Method for verbose queries reduction by integrating key and latent concepts

2.1　关键概念生成模块

为了消除原冗长查询中的冗余信息,达到减少噪声干扰的目的,本文提出了基于 T5 的关键概念生成模块,如图 2 中的关键概念生成模块所示。

1)输入处理。首先将输入的原始查询 $\boldsymbol{Q}=(q_1,q_2,\cdots,q_{|q|})$(长度为 $|q|$)通过嵌入层转换成稠密向量,并加上位置编码,以保留位置信息。具体如式(1)所示:

$$\boldsymbol{E}(q_i)_{i=1}^{|q|}=\text{T5Embedding}(q_i)_{i=1}^{|q|}+\text{PostionalEncoding}(q_i)_{i=1}^{|q|},\tag{1}$$

其嵌入表示为

$$\boldsymbol{E}(\boldsymbol{Q})=(\boldsymbol{E}(q_1),\boldsymbol{E}(q_2),\cdots,\boldsymbol{E}(q_{|q|}))。$$

2)编码器。输入嵌入 $\boldsymbol{E}(\boldsymbol{Q})$ 经过 24 层的 Transformer 编码器[12]。每一层包含 2 个主要的组件:多头注意力机制和前馈神经网络。具体对输入嵌入 $\boldsymbol{E}(\boldsymbol{Q})$ 编码的过程如下:

$$\{\boldsymbol{h}_i\}_{i=1}^{|q|}=\text{TranformerEncoder}(\boldsymbol{E}(q_i))_{i=1}^{|q|}。\tag{2}$$

最后编码器的输出表示为

$$\boldsymbol{H}(\boldsymbol{Q})=(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_{|q|})。$$

3)解码器。解码器每层包含 3 个主要组件:掩码多头自注意力机制、编码器—解码器注意力机制、前馈神经网络。再通过 24 层的解码器进行解码,解码过程如下:

$$\{\boldsymbol{y}_i\}_{i=1}^{|m|}=\text{TranformerDecoder}(\boldsymbol{h}_i)_{i=1}^{|q|},\tag{3}$$

得到解码器的输出表示为

$$\boldsymbol{Y}=(\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_{|m|})。$$

4)输出生成。解码器的最终输出通过线性层映射到词汇表大小的向量,并通过 Softmax 层生成每个时间步的概率分布。具体如式(4)所示:

$$P(\boldsymbol{y}_i)=\text{Softmax}(W_0\cdot\boldsymbol{y}_i+b_0),\tag{4}$$

其中 W_0 为权重参数,b_0 为偏置。

5)训练及推理。在训练过程中,使用交叉熵损失函数来衡量模型生成的词与目标词之间的差异,如式(5)所示:

$$\mathscr{L}=-\sum_{t=1}^{|m|}\sum_{i=1}^{|v|}y_{(t,i)}\log P(y_{(t,i)}),\tag{5}$$

其中,$|m|$为目标序列的长度,$|v|$为词汇表的大小,$y_{(t,i)}$为目标序列在时间步 t、词 i 的真实标签,$P(y_{(t,i)})$为模型在时间步 t 预测词 i 的概率。

在推理过程中,解码器从起始标记〈s〉开始,通过逐步生成每个词,形成最终的输出序列。在每个时间步,解码器接收当前时间步生成的词及编码器的输出,并生成下一个词的概率分布。直到生成结束标记〈/s〉或达到最大长度。

2.2 潜在概念挖掘模块

全面的查询语义信息是实现查询缩略的关键。为了提升检索结果的全面性,本文提出了潜在概念挖掘模块,通过伪相关反馈方法挖掘查询中未显式表达但高度相关的信息。如图 2 中的潜在概念挖掘模块所示,该模块使用 ColBERT 模型[13]对查询和所有的文档进行编码,并计算查询和每个文档的相似度。选取相似度分数最高的前 k 个文档作为原始查询的相关文档集,并从该文档集中提取最具重要性的 n 个术语作为原始查询的潜在概念。

1)查询和文档编码。给定原始输入查询 $\boldsymbol{Q}=(q_1,q_2,\cdots,q_{|q|})$ 和文档集中的其中一个文档 $\boldsymbol{D}=(d_1,d_2,\cdots,d_{|d|})$。查询和文档的长度分别为 $|q|$ 和 $|d|$。使用预训练的 BERT 模型分别对查询和文档进行编码,每个查询和文档都会被编码成一个高维向量表示。两者的向量分别表示为 $\boldsymbol{E}'(\boldsymbol{Q})$ 和 $\boldsymbol{E}(\boldsymbol{D})$。编码过程如式(6)、(7)所示:

$$\boldsymbol{E}'(\boldsymbol{Q})=\mathrm{BERT}(\boldsymbol{Q}),\tag{6}$$

$$\boldsymbol{E}(\boldsymbol{D})=\mathrm{BERT}(\boldsymbol{D})。\tag{7}$$

2)计算相似度得分矩阵。在查询和文档的向量空间中,通过计算查询和文档每个词的内积来度量两者的相似度。相似度得分矩阵 $\boldsymbol{S}$ 的元素 S_{ij} 表示查询中的第 i 个词 q_i 和文档中的第 j 个词 d_j 的相似度。具体如式(8)所示:

$$S_{ij}=\boldsymbol{E}_{q_i}\cdot\boldsymbol{E}_{d_j}^{\mathrm{T}}。\tag{8}$$

3)聚合得分。为了得到查询和文档的最终相关性得分,本文通过最大池化(max-pooling)对相似度得分矩阵 $\boldsymbol{S}$ 进行聚合。具体如式(9)所示:

$$\boldsymbol{S}_{q,d}=\sum_{i\in|q|}\max_{j\in|d|}(\boldsymbol{E}_{q_i}\cdot\boldsymbol{E}_{d_j}^{\mathrm{T}})。\tag{9}$$

4)挖掘潜在概念。经过第一次密集检索后,选取每个查询对应的相关性得分 $\boldsymbol{S}_{q,d}$ 较高的 top-k 文档(本文方法中 $k=100$)作为相关文档集,从中提取潜在概念。为了确保术语提取的有效性,本文使用了 IDF 作为筛选标准。IDF 通常用于衡量一个词在整个文档集合中的重要性,其核心思想是,如果一个词在较少的文档中出现,那么它更具辨别力;而如果一个词在大多数文档中频繁出现,则它的信息量较低。具体如式(10)所示:

$$\mathrm{IDF}(t)=\log\frac{N}{\mathrm{df}(t)},\tag{10}$$

其中,N 表示文档集合中的总文档数,$\mathrm{df}(t)$ 表示包含术语 t 的文档数。通过提取 IDF 最高的 n 个术语(本文方法中 $n=6$)作为潜在概念。

3 实验

3.1 数据集

训练集查询数量为 1 000 条。〈desc〉标签下的内容为描述性的冗长查询,〈title〉标签下的内容为简短查询。测试集则是公共数据集 Robust2004[14]查询集中〈desc〉标签下的 250 条查询。具体格式如图 3 所示。

```
<num> Number: 301
<title> International Organized Crime

<desc> Description:
Identify organizations that participate in international criminal
activity, the activity, and, if possible, collaborating organizations
and the countries involved.
```

图 3　Robust2004 数据集格式示例
Fig.3　Example format of the Robust2004 dataset

3.2　基线模型

本文采用 6 种现有方法作为基线。(1)原始查询。(2)统计方法:Yake。(3)基于 BERT 模型提取关键词的查询缩略方法:AdaptKeyBERT。(4)基于图的关键字提取方法 :TopicRank。(5)基于词频的反馈方法:RM3。(6)基于预训练模型的反馈方法:ANCE-PRF。

1)原始查询。Robust2004 查询集的〈desc〉标签下的查询[14],例如 Identify organizations that participate in international criminal activity, the activity, and, if possible collaborating organizations and the countries involved。

2)Yake[2]是一种完全统计的方法,最近被引入用于关键字提取。它依赖于捕获文档中术语的 5 个统计特征,并将这些特征得分集成到关键短语确定的最终得分中。

3)AdaptKeyBERT[4]是一种基于 BERT 的关键词提取技术,由 Priyanshu 等[11]于 2022 年提出。该技术旨在根据输入内容自动提取与其最相似的关键词和关键短语。AdaptKeyBERT 利用 BERT 模型的强大语义理解能力,能够更好地理解和抽取文本的语义信息,从而提供更准确和相关的关键词提取结果。

4)TopicRank[10]是另一种基于图的关键字提取算法。在 TopicRank 中,图节点不仅仅是单个单词。相反,文档被表示为主题和这些主题之间关系的完整图。主题被定义为一组相似的单词和多词短语。

5)RM3[15](relevance model with pseudo-relevance feedback)是一种用于信息检索的经典反馈算法,由 Zhai 等[15]于 2001 年提出。该算法通过利用检索到的相关文档来动态调整查询模型,以提高检索结果的质量和准确性。在 RM3 中,首先使用初始查询从文档集中检索出一批相关文档,然后根据这些反馈文档中的信息重新建模查询,进而改进原始查询。

6)ANCE-PRF[16]是一种新的查询编码器,该编码器利用伪相关反馈(pseudo-relevance feedback, PRF)改进查询表示,用于密集检索。ANCE-PRF 使用 BERT 编码器,该编码器使用从密集检索模型 ANCE 获取的查询和顶部检索文档,并学习直接从相关标签生成更好的查询嵌入。

3.3　实验参数设置

1) 生成关键概念的 T5-Large 模型

使用 AdamW 作为该模型的优化器,AdamW 的学习率设置为 5e-05,batch_size 设置为 8,epoch 设置为 5。模型中编码器和解码器的隐藏层数设置为 24,每层注意力头数设置为 16,隐藏状态向量维度设置为1 024。

2) 挖掘潜在概念的 ColBERT 模型

使用 Adam[17]作为该模型的优化器,Adam 的学习率设置为 3e-06,dropout 值设置为 0.1。模型中编码器隐藏层数设置为 12,每层注意力头数设置为 12,隐藏状态向量维度设置为 768。查询和文档的最大长度分别设置为 64 和 512。

3.4　实验指标

本文使用 NDCG(normalized discounted cumulative gain)[17]、MAP(mean average precision)[18]、Recall[19]和 MRR (mean reciprocal rank)4 种评价指标来评估模型的性能。

1) NDCG 是一种用于评估搜索结果排序质量的指标,它考虑了搜索结果的相关性以及它们在排序列表中的位置。NDCG 的值介于 0 和 1 之间,1 表示最佳排序,0 表示最差排序。

2) MAP 是检索任务中的另一个重要指标,它用于评估检索系统在不同查询上的平均精度。MAP 计

算了每个查询的平均精度(average precision),然后取所有查询的平均值作为最终的 MAP 值。

3) Recall 是用于评估检索系统在给定查询下检索到的相关文档数量与全部相关文档数量之比。Recall 表示了检索系统能够检索到多少相关文档,是一个召回率指标。

4) MRR 是用于评估检索系统在多个查询上的平均排名的倒数,它考虑了检索系统首次返回相关文档的位置,越靠前的相关文档排名越高。

3.5 基线模型对比实验

为了验证本文提出方法的有效性,在 Robust2004 数据集上进行了实验。具体来说,对原始查询、Yake、AdaptKeyBERT、TopicRank、RM3、ANCE-PRF 以及本文提出的融合关键概念和潜在概念的冗长查询缩略方法进行了查询缩略。然后,利用稀疏检索模型(BM25)对缩略后的查询进行了检索,并评估了检索结果。同时还利用 ColBERT 对 Yake、AdaptKeyBERT、TopicRank、ANCE-PRF 以及本文提出的方法对缩略后的查询进行了评估。实验结果分别如表 1、2 所示。

表 1 不同模型实验结果对比(BM25 模型)

Table 1 Comparison of experimental results of different models (BM25)

模型	MAP	MRR	R@100	R@500	R@1000	NDCG@10	NDCG@50	NDCG@100
原始查询	0.200	0.606	0.355	0.527	0.604	0.393	0.384	0.423
Yake	0.212	0.630	0.361	0.550	0.627	0.408	0.391	0.429
AdaptKeyBERT	0.213	0.636	0.361	0.549	0.627	0.407	0.390	0.397
TopicRank	0.198	0.565	0.337	0.520	0.593	0.372	0.363	0.400
RM3	**0.216**	0.559	0.354	0.545	0.613	0.392	0.366	0.399
ANCE-PRF	0.200	0.600	0.35	0.557	0.643	0.402	0.384	0.419
Ours	0.175	**0.714**	**0.479**	**0.623**	**0.710**	**0.467**	**0.433**	**0.489**

表 2 不同模型实验结果对比(ColBERT 模型)

Table 2 Comparison of experimental results of different models (ColBERT)

模型	MAP	MRR	R@100	R@500	R@1000	NDCG@10	NDCG@50	NDCG@100
Yake	0.184	0.593	0.311	0.469	0.547	0.368	0.321	0.321
AdaptKeyBERT	0.135	0.487	0.262	0.413	0.479	0.285	0.252	0.264
TopicRank	0.177	0.560	0.294	0.437	0.502	0.353	0.303	0.307
ANCE-PRF	0.201	0.675	0.341	0.511	0.578	0.404	0.350	0.358
Ours	**0.219**	**0.681**	**0.350**	**0.526**	**0.599**	**0.440**	**0.377**	**0.380**

实验结果表明,1)在 BM25 模型的 MAP 指标上,尽管本文方法未能达到最佳,但在其他主要评估指标上表现优异,超过了 BM25 模型中最优的基线模型(AdaptKeyBERT),这证明了本文方法在稀疏检索模型上的有效性。

2)在密集检索模型 ColBERT 上,本文方法在所有评估指标上均优于基线模型,在 MAP、MRR、R@1000和 NDCG@10 这 4 个指标上分别提高 1.8%、0.6%、2.1%和 3.6%。密集模型本身依赖于更复杂的语义表征,而本文的方法生成了更加丰富的查询语义表达。实验结果进一步证明了该方法在密集检索场景中的有效性。

3)与只提取关键概念的方法(TopicRank、Yake、AdaptKeyBERT)相比,本文提出的方法更有效。虽然这些方法能够提取出查询中的核心词汇,但它们忽略了潜在概念,导致对长查询或复杂查询的理解不够全面。实验表明,潜在概念作为关键概念的补充,对于提高查询质量具有重要作用。

4)与只挖掘潜在概念的方法(RM3、ANCE-PRF)相比,本文提出的方法更有效。实验结果证明,纯粹依赖潜在概念的方法可能会导致原始查询中的重要信息丢失,特别是在信息密集的复杂查询中。

3.6 消融实验

为了验证本文所提出的 2 个模块的有效性,本文使用 Robust2004 数据集在 ColBERT 模型下进行消融实验。1) W/O key_concept:移除基于 T5 的关键概念生成模块。2) W/O latent_concept:移除基于 ColBERT 的潜在概念挖掘模块。具体消融实验结果如表 3 所示。

表 3　在 ColBERT 下进行的消融实验结果
Table 3　Ablation study results conducted under ColBERT

模型	MAP	MRR	R@1000	NDCG@10
W/O key_concept	0.199	0.678	0.575	0.409
W/O latent_concept	0.206	0.675	0.559	0.429
Full model	**0.219**	**0.681**	**0.599**	**0.440**

3.7　不同数据集下的实验对比

为了验证方法的鲁棒性,本文在 Vaswani 数据集上进行了实验。Vaswani 数据集包含约 11 000 篇科学摘要,主要用于文本检索任务。它包括 93 个自然语言查询(主题)和 2 083 个相关性评估(query relevance, qrels)。利用 ColBERT 对 Yake、AdaptKeyBERT、TopicRank、ANCE-PRF 以及本文提出的方法对缩略后的查询进行了评估。实验结果如表 4 所示。

表 4　在 vaswani 数据集下进行 ColBERT 的实验结果对比
Table 4　Comparison of experimental results for ColBERT on the vaswani dataset

模型	MAP	MRR	R@100	R@500	NDCG@10	NDCG@50
Yake	0.237	0.669	0.380	0.390	0.576	0.832
AdaptKeyBERT	0.250	0.605	0.389	0.396	0.574	**0.841**
TopicRank	0.235	0.628	0.370	0.382	0.540	0.800
ANCE-PRF	0.277	0.69	0.423	0.427	0.588	0.838
Ours	**0.279**	**0.710**	**0.427**	**0.429**	**0.601**	0.835

实验结果表明:1)本文方法在除了 NDCG@ 50 指标的其他指标上取得了最好的效果; 2)本文方法在不同类别的查询中展现了较好的适应性,证明了其在处理多样化查询时的鲁棒性。vaswani 数据集的实验结果进一步证明了本文方法在复杂检索场景中的有效性和泛化能力,在处理科学文献检索任务时也能够实现更高的准确率和召回率。

3.8　实例分析

本节对比了基线模型(AdaptKeyBERT)和本文提出的融合关键概念和潜在概念的冗长查询缩略方法在查询缩略后进行信息检索任务的表现,结果如下。

原始查询　What adverse effects have people experienced while taking aspirin repeatedly?
(反复服用阿司匹林有什么不良反应?)

AdaptKeyBERT 缩略结果

aspirin, effects,taking(阿司匹林,影响,服用)

检索到的 top2 文档:

1)A simple drug with benefits for the heart / Health Check Aspirin goes back to the brew of willow leaves Hippocrates concocted to relieve labour pains in around 400 BC…

(阿司匹林是一种对心脏和健康有益的简单药物,它可以追溯到公元前 400 年左右希波克拉底为缓解阵痛而调制的柳叶……)

2)Aspirin ‘may save 100 000 lives’ The humble aspirin could prevent 100 000 premature deaths a year around the world, says research published today in the British Medical Journal…

(今天发表在《英国医学杂志》上的一项研究称,不起眼的阿司匹林每年可以防止全球 10 万人过早死亡……)

融合关键概念和潜在概念的冗长查询缩略方法缩略结果

关键概念:aspirin repetitive effects(阿司匹林的重复作用)

潜在概念:devastating, pregnant, employers, depression, patients, product (毁灭性的,怀孕的,雇主,抑郁症,病人,产品)

检索到的 top2 文档:

1)The FDA said it will require all aspirin and aspirin-containing products to bear a warning on their labels

alerting pregnant women to the dangers of taking the drug during their final trimester. Possible problems of taking aspirin repeatly during the last three months of pregnancy include effects on fetal circulation and uterine contraction…

(美国食品和药物管理局表示,将要求所有阿司匹林和含阿司匹林的产品在标签上注明警告,提醒孕妇在妊娠后期服用该药的危险。妊娠最后三个月反复服用阿司匹林可能出现的问题包括对胎儿循环和子宫收缩的影响……)

2) The agency said the stronger warning on labels is being required because aspirin can affect fetal circulation and uterine contraction…

(该机构表示,由于阿司匹林会影响胎儿循环和子宫收缩,因此需要在标签上发出更强烈的警告……)

4 结语

本文提出了一种融合关键概念和潜在概念的冗长查询缩略方法,旨在缓解冗长查询在信息检索中的语义鸿沟,提升复杂查询处理的能力从而提高检索性能。通过实验和分析发现,与现有的各类基线模型相比,在 MAP、MRR、R@1000、NDCG@10 这 4 类评价指标上取得了有竞争力的结果,证明了该方法能提高冗长查询缩略的质量。通过进行消融实验,验证了本文提出的 2 个核心模块用于冗长查询缩略任务的有效性。在未来的工作中,会重点探索如何使用该方法进行自监督迭代的训练,以进一步提升冗长查询信息检索任务的性能。

参考文献:

[1] KIM H, CHOI M, LEE S, et al. ConQueR: contextualized query reduction using search logs[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. Taipei: ACM, 2023:1899-1903.

[2] CAMPOS R, MANGARAVITE V, PASQUALI A, et al. YAKE! Keyword extraction from single documents using multiple local features[J]. Information Sciences, 2020, 509:257-289.

[3] DEVLIN J, CHANG M-W, 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: Human Language Technologies. Minneapolis: ACL, 2019:4171-4186.

[4] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. Journal of Machine Learning Research, 2020, 21(140):1-67.

[5] HUSTON S, CROFT W B. Evaluating verbose query processing techniques[C]//Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. Shanghai: ACM, 2010:291-298.

[6] CHAA M, NOUALI O, BELLOT P. New technique to deal with verbose queries in social book search[C]//Proceedings of the International Conference on Web Intelligence. Jinan: IEEE, 2017:799-806.

[7] KUMARAN G, CARVALHO V R. Reducing long queries using query quality predictors[C]//Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Fuji: ACM, 2009:564-571.

[8] ROUSSEAU F, VAZIRGIANNIS M. Main core retention on graph-of-words for single-document keyword extraction[C]// Advances in Information Retrieval: 37th European Conference on IR Research. Vienna: Springer, 2015:382-393.

[9] BOUGOUIN A, BOUDIN F, DAILLE B. Topicrank: graph-based topic ranking for keyphrase extraction[C]//International Joint Conference on Natural Language Processing (IJCNLP). Nagoya: ACL, 2013:543-551.

[10] PODDER D, PAIK J H, MITRA P. Neural language model based attentive term dependence model for verbose query (student abstract)[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington: AAAI, 2023:16300-16301.

[11] PRIYANSHU A, VIJAY S. AdaptKeyBERT: an attention-based approach towards few-shot & zero-shot domain adaptation of keybert[EB/OL]. (2022-11-16)[2024-09-15]. https://arxiv.org/abs/2211.07499.

[12] VASWANI A. Attention is all you need[C]//Advances in Neural Information Processing Systems. Long Beach: NIPS, 2017:1-15.

[13] KHATTAB O, ZAHARIA M. Colbert: efficient and effective passage search via contextualized late interaction over BERT [C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle: ACM, 2020:39-48.

[14] VOORHEES E M. Overview of the TREC 2004 robust track[C]// Text Retrieval Conference. Washington: NIST, 2004:1-12.

(下转第 85 页)