赵靓, 姚鑫, 朱鸿,等. 基于预训练语言模型的审计工作密集文本检索[J]. 智能计算机与应用,2026,16(3):100-108. DOI: 10.20169/j.issn.2095-2163.24051402

基于预训练语言模型的审计工作密集文本检索

赵 靓[1], 姚 鑫[2], 朱 鸿[1], 吕东辉[2], 孔菲菲[1], 王国涛[2]
(1 中国移动通信集团黑龙江有限公司, 哈尔滨 150000; 2 黑龙江大学 电子工程学院, 哈尔滨 150080)

摘 要: 审计工作作为一种监督机制,在企业中承担着重要职责。然而,传统的人工审计过程耗时耗力,工作量巨大,限制了审计效率的提升。为了解决这些问题,本文提出了一种结合对比学习与自注意力机制的密集文本检索方法,以更高效、智能地辅助审计工作。研究首先利用 BERT 预训练语言模型对审计文本进行语义编码,并通过对比学习增强模型的区分能力,使其能够更准确地识别与用户查询相关的条款。在此基础上,加入自注意力机制,使模型能够识别并聚焦于审计内容中的关键要素,从而优化文本检索的性能。该方法显著提高了模型在复杂审计语境下的精确性和适应性,为企业审计工作提供了智能化的高效支持。

关键词: 审计; BERT; 对比学习; 注意力融合
中图分类号: TP18 **文献标志码**: A **文章编号**: 2095-2163(2026)03-0100-09

Audit work-intensive text retrieval based on a pre-trained language model

ZHAO Liang[1], YAO Xin[2], ZHU Hong[1], LÜ Donghui[2], KONG Feifei[1], WANG Guotao[2]
(1 China Mobile Group Heilongjiang Company Limited, Harbin 150000, China;
2 College of Electronic Engineering, Heilongjiang University, Harbin 150080, China)

Abstract: As a supervisory mechanism, audit work holds significant responsibility within enterprises. However, traditional manual auditing is time-consuming and labor-intensive, posing limitations on auditing efficiency. To address these challenges, this paper introduces a dense text retrieval method that integrates contrastive learning and self-attention mechanisms, aimed at providing more efficient and intelligent support for audit tasks. Leveraging the BERT pre-trained language model, the paper encodes audit text semantically and enhances the model's discriminative capacity through contrastive learning, enabling more precise identification of clauses relevant to user queries. Additionally, the integration of self-attention enables the model to focus on key elements within audit content, thereby optimizing retrieval performance. This approach markedly improves the model's precision and adaptability in complex audit contexts, offering a high-efficiency solution for intelligent auditing of enterprises.

Key words: audit; BERT; contrastive learning; attention fusion

0 引 言

审计工作在企业、公司、各单位中都承担着监督管理的重要职责,其高效实施与企业健康发展紧密相连[1-2]。随着互联网信息技术的发展和企业数字化转型步伐的加快,审计工作必然要面对各式各样的复杂数据。所以,自然语言处理等技术与审计工作相融合有助于以快速、准确的方式定位和描述审计中发现的问题[3-4],从而完成审计业务中的大部分基础性工作,将此类技术与审计相结合已成大势所趋[5-6]。密集文本检索技术在审计工作中能够起到重要的作用。文本检索作为信息检索研究的核心领域之一,扮演着至关重要的角色。其主要任务是根据用户提供的查询条件,从庞大的知识库中精确地检索出与查询相关的文档,然后根据相关性对这些文档进行智能排序,并将排名最高的结果呈现给用户。文本检索的核心挑战在于如何准确、高效地度量查询与文档之间的相关性[7]。

基金项目: 黑龙江省重点研发计划(2022ZX03A06)。

作者简介: 赵 靓(1982—),男,硕士,技师,主要研究方向:数据审计与 AI 应用。Email:13936688663@139.com; 朱 鸿(1988—),女,硕士研究生,主要研究方向:大数据分析建模; 孔菲菲(1981—),女,学士,助理工程师,主要研究方向:移动通信技术研究。

收稿日期: 2024-05-14

密集文本检索在 Web 搜索、问答系统等许多领域取得了巨大的成功[8]。文本检索中的核心问题是如何衡量查询和文档的相关性。经典的密集文本检索模型使用双编码器结构对查询和文档进行编码，并将其映射到密集向量中[9-10]。基于此，该模型测量密集向量之间的距离，以表示查询与文档之间的相似度。通常，密集型文本检索存在文件类型不规范，章节格式标识不统一、不同文件的规章制度关系不清、专业概念内涵外延差异明显等问题。采用双编码器得到文本向量以达到样本间的差异化。密集检索任务首先通过负采样策略获得一定数量的负文档，然后使用双编码器获得密集向量，进一步优化信息噪声对比估计损失函数（InfoNCE）。InfoNCE 旨在尽可能地区分阳性和阴性样本。距离会影响文本检索的结果。理想情况下，嵌入空间中查询和正文本向量之间的距离较近，查询和负文本向量之间的距离较大，将提高检索性能[11-12]。

近年来，人们探索了多种方法来构建用于密集文本检索的负训练实例以及负抽样方法。例如，BM25 是信息检索领域的经典算法，基于词频相关性计算查询和文档之间的相似度。BM25 可以根据文档的分数过滤出查询的负样本。但是这些负抽样方法更关注查询和文档之间的精确匹配信号。很可能会在语义层面上丢失查询的相关文档。一些研究者在小批量中选择局部硬负样本（与正样本更相似的负样本）进行密集文本检索。然而，该方法在单词学习和视觉表征方面是有效的，而在离散检索场景下效果不佳[13-14]。另一方面，由于查询和文档之间缺乏交互建模，双编码器体系结构在性能方面不如交叉编码器体系结构，在交叉编码器体系结构中，查询和文档是通过[SEP]符号连接后输入模型的。然而，双编码器结构能够独立高效地对查询和文档进行编码，这是检索任务中应研究考虑的一个方面。

在本文中，提出了一种结合对比学习和自注意力机制的密集文本检索模型。首先，模型使用双编码器结构将查询和文本段落映射到高维语义空间。随后，通过对比学习对文本向量进行优化，使模型能够更有效地分辨正样本和负样本，从而提高检索准确性。同时，模型引入自注意力机制，在对文本特征向量的处理上为不同部分的内容赋予不同的权重。这样一来，模型不仅能够捕捉文本间的语义相似性，还能更好地提炼出关键语义信息。对比学习与自注意力机制的有机结合，使模型在审计文本的语义检索中展现出更高的精准度和效率。

1　问题定义

审计工作处理的文本数据存在格式多样、表述不一致的问题，给信息检索带来了挑战。首先，审计文件格式不同，包含 Word、PDF 等多种文件类型，不同格式的章节标识和内容结构不统一，增加了信息提取的复杂度。其次，由于审计制度的描述不一致，术语的表述方式在不同文件中可能存在差异，导致语义匹配的难度增加。为解决这些问题，本文设计了融合对比学习和自注意力机制的文本检索方法。一方面，通过对比学习增强模型在无监督学习场景下的语义区分能力；另一方面，自注意力机制则帮助模型从复杂文本中提炼出关键信息。这一融合机制确保了在文本检索时不仅能实现精确的语义匹配，还能在不同格式和表述下保持高效、准确的审计条目定位。

在密集文本检索任务中，一个关键问题在于如何高效地度量查询与文档段落之间的语义相似度[15]。传统的稀疏检索方法（如 BM25）往往只能处理显式词汇匹配，而难以捕捉深层语义关系，特别是在处理审计领域专业术语时效果不佳。因此，在审计文档中进行语义匹配和内容检索，要求系统能够处理句子级别和段落级别的上下文信息，以更好地理解用户查询与文档内容之间的关系。

同时，审计文本中的信息量庞大且繁杂，这进一步增加了构建高效检索系统的复杂性。传统的检索方法在面对大量异构文档时，难以同时平衡检索效率和精度。因此，如何设计一种能够有效处理上下文信息、准确理解查询意图并高效进行语义匹配的文本检索模型，是当前审计工作中需要解决的问题。

2　算法流程

研究中使用的多视图对比学习模型的整体架构如图 1 所示。输入数据通过 Jieba 分词进行词频统计并提取关键词，再将关键词与输入通过模型进行分析。在分词和词频统计后，系统会从中提取出关键词。这些关键词将作为特征，输入到后续的模型中，用于更高效的文本比较和学习。在对比学习模型中，系统将不同文本的特征向量放置到一个向量空间中进行对比学习。在 BERT 模型后插入自注意力机制得到 2 组编码向量，再进行组内无监督对比学习与组间无监督对比学习。模型的目标是让语义相似的文本在空间中更接近，而语义差异较大的文本则距离较远[16]。然后，输入到注意力融合模块，

根据关键词和文本内容的重要性,对特征向量的不同部分赋予不同的权重。重要的信息会被赋予较高的权重,而无关信息则被弱化,进而输出一个更优化的特征向量,能够更好地反映文本的核心语义。

如果模型直接从预训练好的语言模型中获取句子向量进行文本检索,准确率会严重降低[17-18]。因此,研究中使用注意力机制与对比学习来生成更准确的编码向量来解决这一问题。其中对比学习由类内对比学习模块和类间对比学习模块组成。

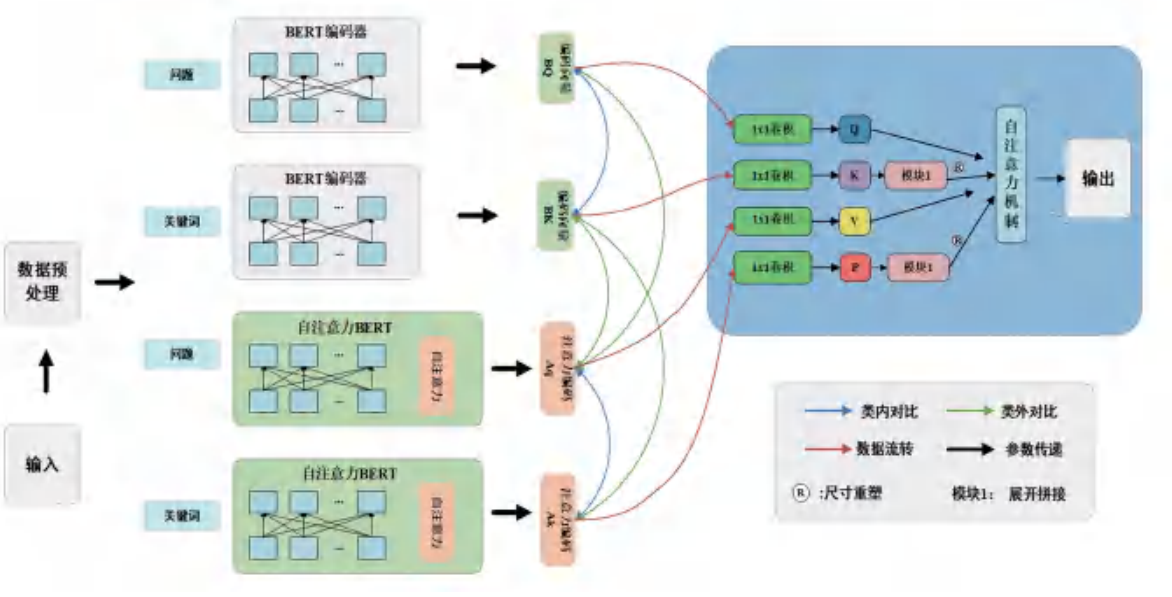

图 1 整体框架

Fig. 1 Overall framework

在类间对比学习模块中,通过架构优化,增强了查询与正向通道的相关性,同时有效降低了子查询与负面信息的相关性。

在类内对比学习模块中,采用无监督的方法对语言模型进行了修改,以提升查询(或段落)与查询(或段落)之间正视图的相似性,降低其与负视图的相关性。该方法让语言模型在空间中更好地捕捉不同句子之间的语义相似性,从而提升模型对同义表达和相关内容的检索能力。

3 模型的相关细节

模型的总体框架见图 1。首先,用 2 个预训练的 BERT 双编码器作为第一类编码器,然后引入自注意力机制形成 2 个 BERT _Attention 编码器作为第二类编码器。接下来,通过类内对比与类间对比,使用特定的损失函数来优化模型,使相似文本对的距离缩小,不相似文本对的距离拉大。模型处理后的输出是一个特征向量,每一个特征向量都代表文本的语义信息。这些向量为通过 1×1 卷积得到 Q, K, V, P 编码,并通过自注意力机制得到更优化的特征向量。

在问答过程中,系统将用户的查询表示为向量 q,并与知识库中的向量集合如 $K=\{k_1,\cdots,k_n\}$ 进行相似度匹配。对于未知知识 u_k,系统通过计算 q 与 K 中各向量的相似性,找到最接近的条目 k_i,作为对 u_k 的近似解,以此提供答案。相似度计算可以采用如下公式:

$$\mathrm{Sim}(q,k_i)=\frac{q\cdot k_i}{\|q\|\cdot\|k_i\|} \tag{1}$$

通过此方法,系统能够在知识库中高效检索出最相关的答案,满足用户查询的需求。在本文的算法中,使用相似度评分 score 来表示用户查询 q 和每个知识库条目的相似度,以此来衡量其与查询的相关性。此分数将用于排序和筛选最符合查询的知识条目。

在检索过程中,将知识库中的每个条目表示为文本段落 para,并通过预处理将其转化为向量形式 p。在执行查询时,系统首先将查询 p 与每个 para 的向量表示 p 进行相似度计算,并根据 score 筛选出最相关的几个 para。这些高分段落将进一步传递给后续模块以返回给用户。

3.1 数据预处理

在智能审计领域,识别并提取文档中的关键字对于高效分析和自动化审计至关重要[19]。为实现

这一目标，需要对审计数据进行深入的预处理，使其更适合智能审计系统的分析与关键词提取需求。具体来说，首先通过了 Jieba 分词将中文文本切分为词语序列，增大词汇粒度，从而更加准确地识别和处理特定的审计关键字。对于文档类型多样性的问题，包括 Word 和 PDF 等文件格式，可以利用正则表达式提取字符和段落，进行字符清洗，去除多余符号与格式干扰，进一步聚焦于有审计价值的关键信息。不同文件的切词和检索效果如图 2 所示。

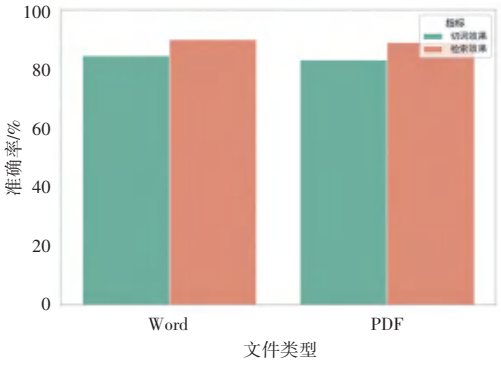

图 2　不同文件的切词和检索效果

Fig. 2　Tokenization and retrieval performance for different file types

完成初步文本清洗与切分处理后，使用 pandas 统计词频以识别文档中的高频关键词。pandas中的 value_counts( ) 函数能够快速、准确地统计词语的出现次数，从而有效获取文档中最具代表性的关键词。pandas 的这种方法，不仅提供了简单、灵活的数据处理流程，还便于在大规模文档数据中进行多维度分析（如按词汇出现频率分组，筛选出特定类别的关键词）。对于每个词，其词频（简单的频数）计算方法如下：

$$\mathrm{TF}_t = N_T \tag{2}$$

其中，N_T 表示词 t 在文档中出现的总次数。在对文本进行预处理的过程中，还将其转化为适用于计算的知识向量。研究中设输入文本 T 由词语集合 $\{w_1, \cdots, w_n\}$ 组成，每个词 w_i 被映射为一个词向量 $V(w_i)$。为了构建文本的整体表示，对每个词向量进行平均池化，已得到文本的表示向量 $V(T)$。其数学定义公式为：

$$V(T) = \frac{1}{n}\sum_{i=1}^{n} V(w_i) \tag{3}$$

该向量表示包含了文本的主要语义信息，便于后续查询的相似度计算和检索。最后，将统计出的高频词作为模型输入的关键词进行编码分析[20-21]。

同时，在审计制度库中，构建了一种关系模式，分别为制度文件表和预处理句子表两个核心数据表，并通过实体关系图如图 3 所示。

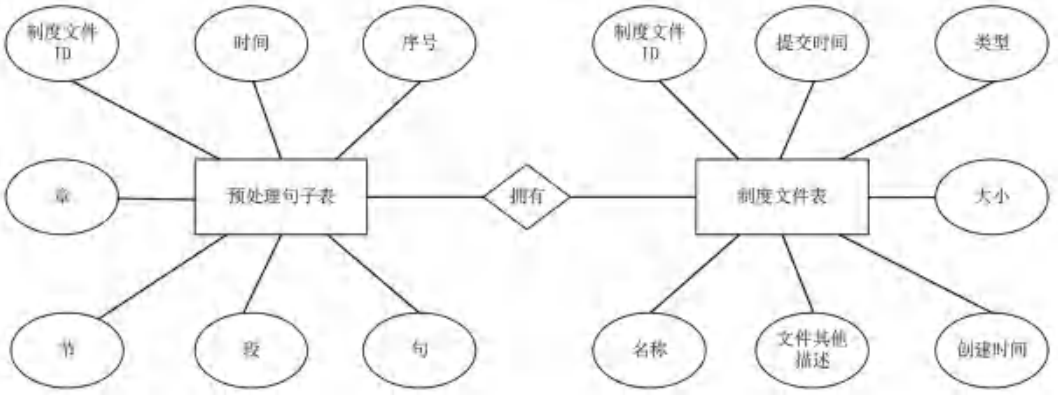

图 3　审计制度库中的关系模式

Fig. 3　Relationship schema in the audit regulation database

制度文件表主要记录了与审计相关的制度文件的元数据，包括制度文件 ID（作为主键）、提交时间、文件名称、文件类型、文件大小、创建时间、文件描述以及提交人信息。而预处理句子表记录了对制度文件进行预处理后所生成的细粒度文本信息，包括 序号（主键）、时间、处理方法、制度文件 ID（外键，用于关联制度文件表）、以及预处理过程中划分的章、节、小节、段落、句和页号。

制度文件表和预处理句子表通过制度文件 ID 建立关联，表示制度文件经过预处理步骤生成多个预处理句子。这一设计确保了制度文件与其预处理结果之间的有效追溯，并为后续的智能检索和审计分析提供了数据基础。

通过上述表结构及其关系的设计，确保了审计制度库的灵活性和扩展性，使得智能检索算法能够根据用户输入的查询段落高效获取相应的制度条目

或案例,从而更好地支持自动化审计工作流程。

3.2 对比学习

在本文中,主要考虑一种用于密集文本检索的对比学习范式。给定一个查询 q 和通道 C,其中查询 q 为包含上下文的句向量,而通道 C 表示为分出层次的段落向量。通过对比学习训练后,这些向量将生成表示查询和知识条目之间相似度的特征,用于提升文本检索的准确性。

典型的密集文本检索模型是双编码器结构,其中 2 个独立的密集编码器(EncoderA 和 EncoderB)用于将段落和查询映射到 d 维密集向量。具体来说,采用 2 个预训练的语言模型,如 BERT 初始化 EncoderA 和 EncoderB,并使用其最后一层输出中的[CLS]标记表示作为查询和段落的编码。接着在每个查询和消息的开头添加[CLS]起始符号作为输入,然后将 BERT 最后一层[CLS]位置的输出向量作为整个句子表示,再用数据库对向量进行索引并进行高效检索。查询 q 与通道 C 计算可以采用基于向量的加权余弦相似度。具体公式为:

$$\mathrm{Sim}(q,C)=\frac{q\cdot C}{\|q\|\cdot\|C\|} \tag{4}$$

密集文本检索模型是通过传统的对比学习损失来训练的,其本质是学习查询和段落的密集表示,使得查询和正段落比查询和负段落具有更高的相似度。给定查询集合 $Q=\{q_1,\cdots,q_n\}$ 和通道集合 $C=\{c_1,\cdots,c_n\}$,则对比损失函数 L 为:

$$L=\sum_{i=1}^{n}\left(1-\frac{\mathrm{sim}(q_i,C_i^{+})}{\mathrm{sim}(q_i,C_i^{-})+\delta}\right)^2 \tag{5}$$

其中,δ 表示平滑参数,用于避免分母为零的情况。这个自定义损失函数通过最小化查询与正样本之间的距离差平方,实现正负样本的高效区分。研究的目标是用 n 个负文档训练实例来优化负对数似然。由于计算成本的限制,将 n 设置为远小于 c 的值。在本文中,用于训练密集通道检索的负样本是通过数据增强技术生成的。最终,将此方法用于审计工作并验证其可行性。

在对比学习中,将通过 BERT _Attention 编码器得到编码向量 AQ 和 AK。其中,BERT 的最后一层 CLS 向量进行自注意力编码得到 AQ 和 AK,定义如下:

$$\mathrm{AQ}=\mathrm{SelfAttention}(\mathrm{BERT}(\mathrm{question})) \tag{6}$$

$$\mathrm{AK}=\mathrm{SelfAttention}(\mathrm{BERT}(\mathrm{question})) \tag{7}$$

为了使查询更靠近正通道,远离负通道。形式上,将其类内对比损失定义如下:

$$I\_\mathrm{Loss}_1=-\log\frac{e^{\sin(\mathrm{AQ},\mathrm{AK})}}{e^{\sin(\mathrm{AQ},\mathrm{AK})}+\sum_{i=1}^{n}e^{\sin(\mathrm{AQ},\mathrm{AK})}} \tag{8}$$

考虑到计算效率,在训练时使用 BERT 预训练模型,以提供查询和通道的正视角表示。研究希望尽可能保留预训练模型的强大推理能力,所以将输入问题与提取出的关键词分别输入到 BERT 双编码器模型中,计算得到编码向量 BQ 和 BK。其定义如下:

$$\mathrm{BQ}=\mathrm{BERT}(\mathrm{question}) \tag{9}$$

$$\mathrm{BK}=\mathrm{BERT}(\mathrm{question}) \tag{10}$$

此外,使用平均池化对 BERT 的最后一层输出进行处理,然后采用无监督对比学习训练。研究中将其类内对比损失定义如下:

$$I\_\mathrm{Loss}_2=-\log\frac{\mathrm{esim}(\mathrm{BQ},\mathrm{BK})}{\sum_{i=1}^{N}e^{\mathrm{sim}(\mathrm{BQ},\mathrm{BK})}} \tag{11}$$

其中,点积函数公式为:

$$\mathrm{Sim}(q,p)=E_Q(q)^{\mathrm{T}}\cdot E_p(p) \tag{12}$$

如上所述,文中根据编码器的组间对比得到了对比学习后的 2 组编码向量。为了尽可能多地提高模型的表示能力,将引入组件对比,即将是否通过自注意力机制学习的编码结果分为 2 组,并分别与另一组中的 2 个编码向量进行对比学习。直观上,交叉型对比学习可以看作是另一种数据增强方法,但并不依赖于额外的数据工程,可以构建高度判别的对比样本。其损失函数定义如下:

$$C\_\mathrm{Loss}_2=-\log\frac{e^{\mathrm{sim}(A,B)}}{e^{\mathrm{sim}(A,B)}+\sum_{i=1}^{n}e^{\mathrm{sim}(A,B)}}-\log\frac{e^{\mathrm{sim}(A,B)}}{e^{\mathrm{sim}(A,B)}+\sum_{i=1}^{n}e^{\sim(A,B)}} \tag{13}$$

其中,A 与 B 分别表示 BERT 编码器与 BERT _Attention 编码器得到的编码向量。这里,$A\in(\mathrm{AQ},\mathrm{AK})$;$B\in(\mathrm{BQ},\mathrm{BK})$。

3.3 注意力融合

研究中将对比学习得到的 4 组特征向量通过多个 1×1 的卷积层,生成查询(Query)、键(Key)和值(Value)三组特征表示。具体公式如下:

$$Q,K,V,P=\mathrm{Conv}_{1\times1}(\mathrm{Vector}) \tag{14}$$

为了捕捉更多的语义信息,使用语义编码添加到查询和键特征中。对此可以表示为:

$$Q_{\mathrm{pos}}=Q+\mathrm{PE} \tag{15}$$

$$K_{\mathrm{pos}}=K+\mathrm{PE} \tag{16}$$

对查询和键特征进行展开和填充（Unfold and Pad，U&P）操作，以适应自注意力机制的输入要求。并计算查询和键的点积，得到注意力权重，应用于值特征进行加权求和。对此可以表示为：

$$\text{Attention}(Q_{\text{pos}},K_{\text{pos}},V)=\text{Softmax}\left(\frac{Q_{\text{pos}},K_{\text{pos}},V}{\sqrt{d_k}}\right)V \quad (17)$$

4　检索系统流程

该流程采用 Flask 作为后端框架，并连接到审计制度库以管理数据，实现了一个简洁高效的问答检索流程。检索流程如图 4 所示。首先，用户通过网页前端输入查询问题，前端将问题通过 HTTP 请求发送到 Flask 后端。Flask 接收到请求后，将问题传递给一个检索模型进行处理。模型对用户的问题进行语义分析，将问题转化为特征向量表示，以便与审计制度库中的答案进行相似度匹配。审计制度库的结构为了实现高效的审计条款检索和智能审计问答系统，设计了一套结构化的审计制度库。该制度库包含以下 3 个主要关系模式：

（1）制度文件管理。审计制度库中包含不同类型的制度文件，包括政策法规、公司内部规章、审计手册等。这些文件按照类别和版本进行管理，以确保检索到最新、最相关的条款。

（2）文件类型信息。为每一类文件分配特定的标签，以便在检索时根据文件类型快速筛选相关内容。例如，法规文件和内部审计手册在处理上会有所区分。

（3）切词后的字词 token。在进行文本预处理时，所有制度文件会通过 Jieba 分词工具切分成字词，并生成相应的 token。此后，这些 token 会存储在数据库中，用于后续的语义匹配和检索。

审计制度库中存储了大量的答案及其对应的向量表示，这些向量用于快速计算问题和答案之间的相似度。

在相似度计算阶段，系统对问题向量与数据库中各答案向量之间的相似性进行计算，可能采用余弦相似度或其他相似度度量方法，从而找到最匹配的答案。在密集文本检索阶段，将输入的查询 q 与审计制度库中的知识向量 K 进行对比。知识向量 K 由字词 token 和文件类型信息共同组成，通过对比学习算法学习到的语义表示，使得模型能够更准确地区分正负样本。在注意力机制中，引入了基于文件类型的信息权重，以确保在检索时优先匹配更相关的条款。系统从中选择 TOP－K 个最相关的答案，并将这些候选答案传回 Flask 后端。最终，Flask 后端将这些答案返回给网页前端，前端将其展示给用户，用户可以在页面上查看到与其查询问题最相关的几个答案。该系统利用 Flask 和制度库的灵活性，实现了高效的文本检索能力，能够快速响应用户查询并提供精准的答案。

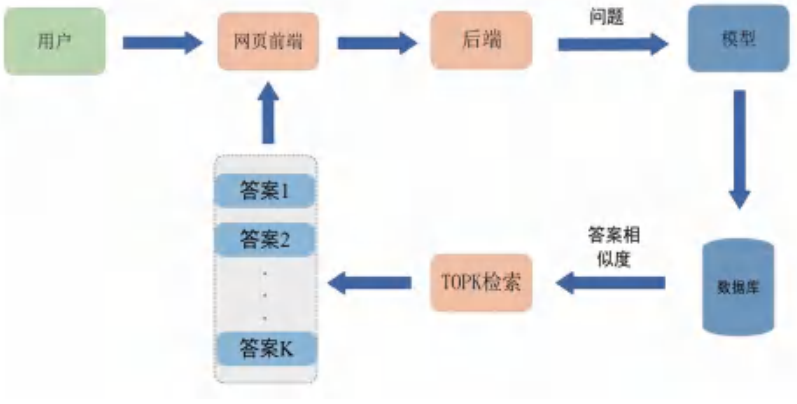

图 4　检索系统流程

Fig. 4　Flowchart of retrieval system process

5　系统测试

5.1　数据集

在本研究中，选用了 2 个公开数据集：Natural Questions（NQ）和 MSMARCO，用于评估所提出模型的性能。选择这 2 个数据集的原因在于其在自然语言处理领域的广泛使用和高度权威性，且适合密集文本检索任务的验证。表 1 展示了这 2 个数据集的详细信息。这些数据集不仅覆盖了不同类型的查询场景，同时也能够有效评估模型在处理开放域问答和信息检索任务中的泛化能力。

表 1 数据集

Table 1 Statistics of datasets

数据集	文档数量	测试集中查询数量	开发集查询数量	训练集查询数量
MSMARCO	8 841 323	6 837	6 980	502 939
NQ	21 015 324	3 610	6 515	58 812

5.2 实验设置

本文使用深度学习框架 Pytorch，并且基于 Hugging Face 库在 GPU 为 RTX 4060 下进行实验，利用 APEX 和梯度累计降低 GPU 内存消耗。双编码器用 BERT-Base 进行初始化，其它的、本文也尝试使用 Condensor 初始化，这是一个针对稠密文本检索任务的预训练模型。对于实验流程，具体如下：

（1）本文遵循 DPR 的实验设置，首先使用 BM25 负样本训练一个双编码器，并检索前 200 文档，然后本文加载了一个训练有素的交叉编码器模型，例如 ERNIE-2.0-Large，对被检索到的文档去噪，并且利用异质数据增强策略生成多样性的训练实例。

（2）本文以 Condensor 初始化另外一个双编码器，并且在具有多样性属性的 NQ、MSMARCO 数据集上微调双编码器。

5.3 模型性能比较

本文将该模型与传统模型进行比较。RocketQA 和 PAIR，Condensor、coCondensor 通过构造高质量“硬负样本”、多阶段训练、数据增强、知识蒸馏、对比预训练等方式改进稠密检索。图 5 展示了本文方法在稠密文本检索方面展示了最新的结果。

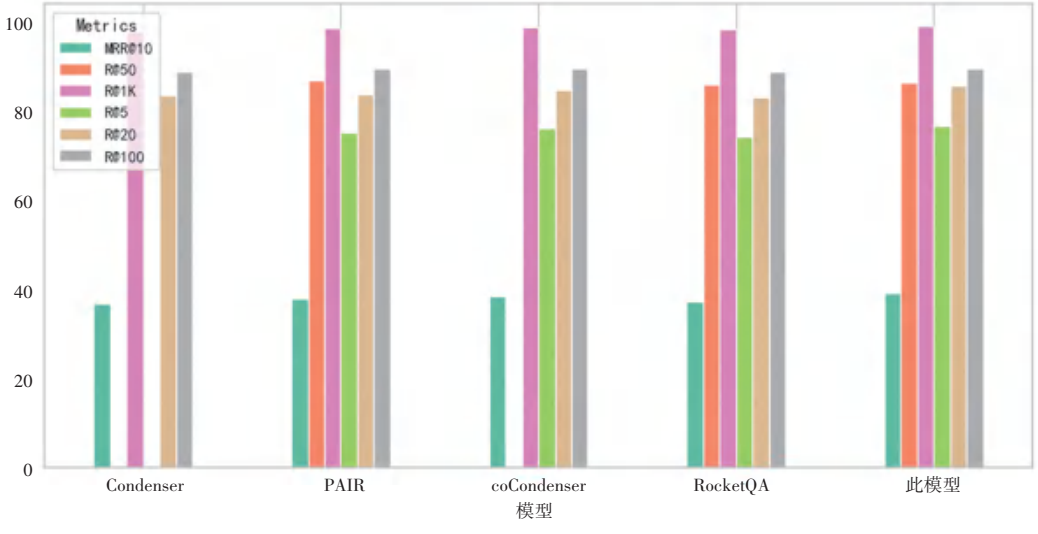

图 5 不同模型在数据集上的文本检索结果

Fig. 5 Text retrieval results of different models on the datasets

5.4 实验结果分析

由图 5 可以看到该模型在 MS-MARCO 和 NQ 数据集上都有显著的效果。RocketQA 通过跨批负、结合 Re-Ranker 去噪的“硬负样本”训练双编码器。PAIR 利用以查询为中心和以文档为中心的对比关系进行稠密的文档检索，也同样采用了 RocketQA 的数据增强方法。本文观察到此模型在 MSMARCO 以及 NQ 数据集上的 MRR，Recall@50，Recall@1K 都优于现阶段最新的 RocketQA 和 PAIR。本文认为该模型的多重对比结构以及异质数据增强改进了检索器，使得其更适用于稠密文本检索。

5.5 消融实验

本文对审计工作所需的待审文件进行文本检索来验证该方法是否可以应用到审计工作中。为了检验多视角对比学习框架对稠密文本检索的效果，除了传统的对比训练损失（CrossLoss_1）外，本文删除了其余的对比损失，例如 Inner-type 对比学习模块中的 InnerLoss_1 和 InnerLoss_2，Cross-type 对比学习模块中的 CrossLoss_2 和 CrossLoss_3。此外，本文验证了异质数据增强策略对该模型性能的影响，本文通过使用 BM25 负样本而不是数据增强处理后的数据实现。

图 6 显示了消融的实验结果。从图 6 中可以看到，多视角对比学习框架对稠密文档检索有显著改进。其中，Cross-type 对比学习模块的影响最大，本文将 Cross-type 对比模块视作一种没有数据工程的

数据增强方式，可以模拟大量未标记样本，从而有效地区分正向和负向样本，Inner-type 对比学习则进一步的改进了模型的表示能力。去掉数据增强模块导致此模型的性能下降，也直观地说明了大规模训练数据对稠密检索任务的有效性和必要性。

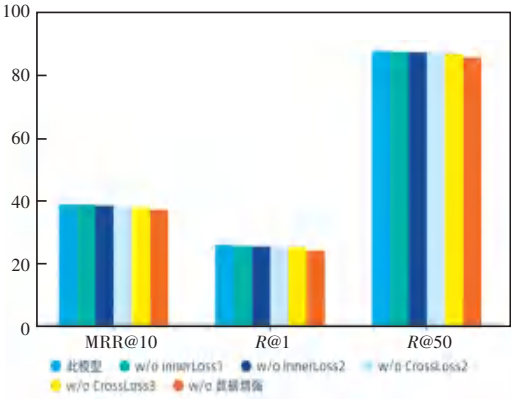

图 6　消融实验

Fig. 6　Ablation research

5.6　审计工作检索结果

本文对审计工作所需的待审文件进行文本检索来验证该方法是否可以应用到审计工作中。审计工作检索结果如图 7 所示。

由图 7 可知，该模型在审计工作文本检索的各项任务都有着不错的表现，由此可知该方式应用到实际工作中是有一定价值的。

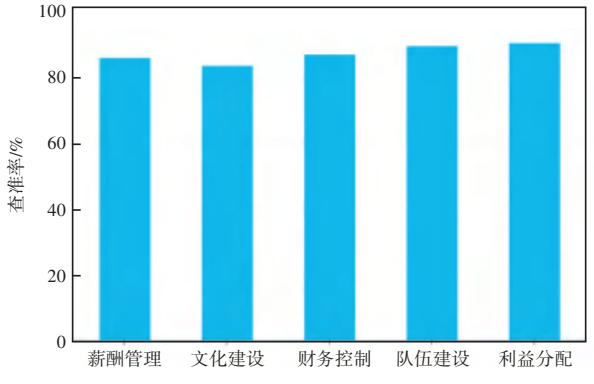

图 7　审计工作检索结果

Fig. 7　Audit work retrieval results

智能检索算法可以根据用户输入的相关段落，来获取在审计过程中想要的制度或案例，由此实现审计的目的。智能检索算法代码具体如下。

算法 1　智能检索算法

输入　查询语句 q，topk

输出　与输入相关的段落文本

```
1. begin:
2.     v = vector(q)
3.     for item in databases do
4.         item_v = index(item)
5.         score = relative (v, item_v)
6.         scores. add(score), para. add(item)
7.     rank(scores,para)
8.     result = extract(scores,para,topk)
9. End
```

6　结束语

在密集文本检索研究中，针对审计工作中的应用实现问题，本文在 BERT 模型的基础上加入了对比学习与数据增强，提高了密集文本检索的性能，显著提高了模型区分正样本和负样本的能力。该模型可以使审计工作更加便捷，同时减少了大量的人力资源。在后续审计工作中将着重研究工作的智能化。

参考文献

[1] YANG Y, JUN Z, LIN Z, et al. Multi-views contrastive learning for dense text retrieval[J]. Knowledge-Based-Systems, 2023, 274:110624 .

[2] 赵艳霞. 人工智能在交通运输部门内部审计中的应用研究[J]. 会计之友, 2023(15):122-127.

[3] 朱昊, 孙宇. 基于人工智能下智慧审计的探究[J]. 国际商务财会, 2023(8):33-37.

[4] PAN Jiahui, YU Yangzuyi, LI Man, et al. A multimodal consistency-based self-supervised contrastive learning framework for automated sleep staging in patients with disorders of consciousness [J]. IEEE Journal of Biomedical and Health Informatics, 2025, 29(2):1320-1332.

[5] 高远. 人工智能视域下高校财务收支审计研究[J]. 上海商业, 2023(1):122-124.

[6]范琳琳, 孟锦, 董坤, 等. 基于人工智能的高校财务收支审计研究[J]. 会计之友, 2022(19):18-23.

[7] KALYANATHAYA K P, AKILA D, SUSEENDREN G. A fuzzy approach to approximate string matching for text retrieval in NLP [J]. Journal of Computer Information Systems, 2019, 15(3): 26-32.

[8] XIONG L, XIONG Chenyan, LI Ye, et al. Approximate nearest neighbor negative contrastive learning for dense text retrieval[J]. arXiv preprint arXiv, 2007. 00808, 2020.

[9] ZHAO W X, LIU Jing, REN Ruiyang, et al. Dense text retrieval based on pretrained language models: A survey[J]. arXiv preprint arXiv,2211. 14876, 2022.

[10]YAO T, PENG S, WANG L, et al. Cross- modality interaction reasoning for enhancing vision-language pre-training in image-text retrieval[J]. Applied Intelligence, 2024, 54(23):12230-12245.

[11]KUMAR R, SHARMA S C. Hybrid optimization and ontology-based semantic model for efficient text-based information retrieval [J]. The Journal of Supercomputing, 2023, 79(2): 2251-2280.

[12] LU Shuqi, HE Di, XIONG Chenyan, et al. Less is more: Pretrain a strong Siamese encoder for dense text retrieval using a

weak decoder[J]. arXiv preprint arXiv, 2102. 09206, 2021.
[13]MA Xueguang, WANG Liang, YANG Nan, et al. Fine-tuning LLaMA for multi-stage text retrieval[J]. arXiv preprint arXiv, 2310. 08319, 2023.
[14] REN Ruiyang, LV Shangwen, QU Yingqi, et al. PAIR: Leveraging passage-centric similarity relation for improving dense passage retrieval[J]. arXiv preprint arXiv, 2108. 06027, 2021.
[15]NGUYEN D K B, DANG K T, BINH T N , et al. A High-performance non-indexed text search system[J]. Electronics, 2024, 13(11):2125.
[16]KHANG G H N , NHAT M N , QUOC N T , et al. Vietnamese legal text retrieval based on sparse and dense retrieval approaches [J]. Procedia Computer Science, 2024, 234:196-203.
[17]LI Chengyu, CHENG Debo, ZHANG Guixian, et al. Contrastive learning for fair graph representations via counterfactual graph augmentation[J]. Knowledge-Based Systems, 2024, 305: 112635.
[18]GAO L, CALLAN J. Unsupervised corpus aware language model pre-training for dense passage retrieval[J]. arXiv preprint arXiv, 2108. 05540, 2021.
[19]KOU H, YANG Y, HUA Y. KnowER: Knowledge enhancement for efficient text-video retrieval[J]. Intelligent and Converged Networks, 2023, 4(2): 93-105.
[20]ZHANG J , WANG L , ZHENG F , et al. An enhanced feature extraction framework for cross-modal image - text retrieval[J]. Remote Sensing, 2024, 16(12): 2201.
[21]FANG Hongchao, WANG Sicheng, ZHOU Meng, et al. Cert: Contrastive self-supervised learning for language understanding [J]. arXiv preprint arXiv, 2005. 12766, 2020.