

语义特征空间模型在基于RAG的智能问答中的应用

黄红伟¹, 杜军¹, 卢云涛¹, 马继涛¹, 马健², 朱培虎³

- (1. 云南省科学技术情报研究院, 云南 昆明 650051;
2. 香港城市大学 信息系统系, 香港 999077;
3. 艾瑞思软件(深圳)有限公司, 广东 深圳 518063)

摘要: 为提升基于RAG架构的智能问答系统文本召回率, 研究分析了当前常用的文本向量化策略。针对各种策略存在的上下文语义不连贯及词嵌入后其向量中被引入噪声等问题, 提出一种语义特征空间模型以及利用文本要点进行语义检索的向量化策略。通过该模型分析并证明基于文本要点策略构造的语义特征空间能够更好地逼近领域知识空间, 并得到将文本向量投影到低秩语义特征空间进行语义检索以提高文本召回率的方法。整体应用该模型、策略、方法所形成的方案优化并改进了RAG架构, 实验结果显示, 其召回率较传统的RAG架构有显著提升, 以大语言模型为底座实现了科技政策法规智能问答。该方案进一步完善了RAG应用开发技术栈, 其语义特征空间可用于改进向量数据库的搜索算法。

关键词: 语义特征空间; 文本要点; 检索增强生成; 智能问答; 大语言模型

DOI: 10.11907/rjdk.242066

扫描二维码阅读全文:



中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(2026)002-0008-06

Application of Semantic Feature Space Model in Smart Q&A Based on RAG

HUANG Hongwei¹, DU Jun¹, LU Yuntao¹, MA Jitao¹, MA Jian², ZHU Peihu³

- (1. Yunnan Academy of Scientific & Technical Information, Kunming 650051, China;
2. Department of Information Systems, City University of Hong Kong, Hong Kong 999077, China;
3. IRISaas Limited, Shenzhen 518063, China)

Abstract: To enhance the text recall rate of smart Q&A systems based on the Retrieval Augmented Generation (RAG) architecture, research has analyzed commonly used text vectorization strategies. Addressing issues such as contextual semantic inconsistency and noise introduced into word embeddings after vectorization, a semantic feature space model and a vectorization strategy utilizing text key points for semantic retrieval are proposed. Through this model, it is analyzed and proven that the semantic feature space constructed based on the text key point strategy can better approximate the domain knowledge space. Furthermore, a method is derived for improving text recall rates by projecting text vectors into a low-rank semantic feature space for semantic retrieval. The overall application of this model, strategy, and method forms a scheme that optimizes and improves the RAG architecture. Experimental results demonstrate a significant increase in text recall rates compared to traditional RAG architectures, with the implementation of a smart Q&A system for science and technology policies and regulations using large language models as the foundation. This scheme further refines the RAG application development technology stack, and the semantic feature space can be applied to improve search algorithms for vector databases.

Key Words: semantic feature space; text key points; RAG; smart Q&A; LLM

收稿日期: 2025-02-20

基金项目: 云南省科技发展战略与政策研究专项(202404AL030008); 云南省公共科技服务平台专项(202305AH340004)

作者简介: 黄红伟(1977-), 男, 硕士, 云南省科学技术情报研究院高级工程师, 研究方向为人工智能、软件工程; 杜军(1978-), 男, 硕士, 云南省科学技术情报研究院高级工程师, 研究方向为计算机应用; 卢云涛(1983-), 男, 云南省科学技术情报研究院工程师, 研究方向为网络通信; 马继涛(1974-), 男, 硕士, 云南省科学技术情报研究院正高级工程师, 研究方向为计算机应用; 马健(1962-), 男, 博士, 香港城市大学信息系统学系教授, 研究方向为人工智能; 朱培虎(1991-), 男, 博士, 艾瑞思软件(深圳)有限公司数据科学家, 研究方向为人工智能。本文通讯作者: 马继涛、朱培虎。

0 引言

利用大语言模型 (large language Model, LLM), 通过 RAG (Retrieval Augmented Generation) 构建知识库是目前智能问答系统的主流方案^[1-5]。应用该方案, 将文本向量化后存入向量数据库, 需要将长文本分块, 于是有学者提出了按固定大小分块、段落分块、递归分块等向量化检索策略^[6-8]。不同策略对文本的召回效果不尽相同, 直接影响模型处理大型文本数据的效率和准确性。

对于选择什么策略, 目前没有明确的理论依据。本文通过研究基于 RAG 的科技政策智能问答, 提出了一种语义特征空间模型。通过该模型构建各不同向量化检索策略的语义特征空间, 同时优化并改进了 RAG 架构。本文有以下 3 点创新: ①提出了语义特征空间模型作为构建 RAG 应用的工具, 该模型可以较好地对各种向量化检索策略进行量化评价, 为文本向量化策略选择提供理论依据, 这为语义特征空间研究提供了新思路; ②提出了使用文本向量在低秩语义特征空间的投影向量进行语义检索以提高文本召回率的方法, 该方法适用于所有文本向量化检索策略, 利用该方法有望改进向量数据库中的搜索算法; ③提出了使用文本要点代替原文的向量化检索策略, 文本要点是从具有相对完整上下文的章节、段落等大文本块凝练出来的若干知识, 其表示的语义与碎片化原文片段所表示的微观语义相比, 携带着相对完整的语义信息, 且每一个要点长度适中, 解决了长文本向量化后噪声过多和稀疏性过高的问题。

1 语义特征空间模型

1.1 领域知识空间

在问答系统中, 问与答总是基于一个特定的知识域空间。比如: 关于科技计划项目管理, 问题与答案总是基于科技计划项目管理办法这个文件承载的知识空间; 关于科技奖励、重点实验室或是其他问题, 莫不如是。不同的领域构成不同的知识空间。将某个领域的知识记为 K , 该领域构成的知识空间记为 $S(K)$, 关于该领域的问题空间记为 $Q(K)$ 。承载这些领域知识的则是相关的管理办法、规章制度等文件, 将这些文件通过嵌入模型转化为向量集, 故本文讨论的空间是向量空间, 领域知识空间与问题空间是一个高维线性空间。

按照 RAG 检索→增强→生成的架构, 其检索过程本质上是将问题映射到知识点的过程, 即经过向量化后问题与对应的知识点本质上是位于同一向量空间中。由此得到: 问题空间本质上等同于知识空间, 即 $S(K) = Q(K)$ 。

1.2 语义特征空间建模

文献[2-4]中提到对承载知识的长文本分块后的向量

化实际上就是按不同的策略建立的知识空间。为了量化比较不同策略建立的空间, 需建立此类空间的数学模型。

本文使用相同的嵌入模型向量化文本, 因而所有向量具有相同的维度。将按同一策略分块后的文本经嵌入模型得到的向量按顺序作为行向量组成一个矩阵, 记为 M , 矩阵 M 的行向量所张成的线性空间称作语义特征空间, 记为 $S(M)$ 。

为了去除噪声与冗余, 对矩阵 M 进行 SVD 分解, 只取其前 h 个主特征, 得到式(1)。

$$(U, \Sigma, V) = \text{SVD}(M) \quad (1)$$

其中, $V = (v_1, \dots, v_h)^T$ 为右奇异向量矩阵, h 个 v_i 两两正交, 且其维度等于文本向量化后的维度。将以 h 个 v_i 作基张成的空间 $S(V)$ 作为语义特征空间 $S(M)$ 的近似表示, 有式(2)成立。

$$S(V) \approx S(M) \leq S(K) = Q(K) \quad (2)$$

1.3 基于文本要点的语义特征空间

常规分块向量化策略将分块后的原文片段直接向量化后存入向量数据库, 回答问题时直接用问题与原文片段进行语义相似检索。这种原文片段缺乏上下文关联, 仅包含文本片段本身的微观语义。本文提出了基于文本要点的向量化策略及其语义特征空间, 具体如下: ①将科技政策法规文档按章节进行切分并存入关系数据库, 考虑到大模型上下文窗口限制, 每个分块最多不超过 1 500 个字符。如果有章节超过 1 500 个字符, 再按段落切分; ②将文本分块逐个传给大语言模型, 按预先设计的提示词为每个分块生成若干文本要点, 每条要点不超过 200 个字符; ③对大模型生成的要点进行清洗 (去除重复或没有意义的要点), 将清洗后的文本要点存入关系数据库, 并建立与原文分块的关联; ④调用嵌入模型将文本要点向量化后存入向量数据库, 并按照上文方法构造基于文本要点的近似语义特征空间。

1.4 语义特征空间逼近知识空间的度量

$S(K)$ 是真实的领域知识空间, $S(V)$ 是按照原文本分块后得到的近似语义特征空间。根据式(2), 按不同分块策略生成的 $S(V)$ 应能逼近 $S(K)$, 本文通过计算 $S(V)$ 与 $S(K)$ 之间的距离以度量 $S(V)$ 逼近 $S(K)$ 的程度。

定义函数 $\text{Dist}(S1, S2)$ 为空间 $S1$ 与 $S2$ 之间的距离, 根据上文的 $S(K) = Q(K)$, 可得到式(3)。

$$\text{Dist}(S(V), S(K)) = \text{Dist}(S(V), Q(K)) \quad (3)$$

计算语义特征空间与领域知识空间的距离, 并将其转化为计算语义特征空间与问题空间的距离。

通过收集日常的科技政策法规相关问题, 或将科技政策法规等文件传给大模型以生成与之相关的问题, 得到一个问集, 这些问题经向量化后构成了问题空间的子集。在 RAG 的检索阶段, 需要将问题转换为向量, 并在向量数据库中检索出语义最相近的向量。语义特征空间到问题子空间的距离可以通过统计问集中所有向量到语义特

征空间的距离加以分析与度量。

本文将问题向量 w 到语义特征空间 $S(V)$ 的距离定义为 w 在语义特征空间 $S(V)$ 中投影向量 w^p 与 w 的距离。 w^p 的计算公式如式(4)所示。

$$w^p = \sum_{i=1}^h \langle w, v_i \rangle v_i \quad (4)$$

其中, $v_1, \dots, v_h$ 为向量矩阵 M 进行 SVD 分解后得到的右奇异矩阵 V 中两两正交的行向量, 这 h 个向量是语义特

征空间 $S(V)$ 的基。 $\langle w, v_i \rangle$ 表示 w 与 v_i 作点积运算。 w 与 w^p 的距离使用 L2 范数, 公式如式(5)所示。

$$\text{Dist}(w, w^p) = \|w - w^p\|^2 \quad (5)$$

对 33 个科技政策法规文件, 分别使用文献[2-4]中提及的按固定大小分块、按符号递归分块以及本文提出的按文本要点分块等 3 个策略, 构建语义特征空间, 并计算问题集中所有问题到这 3 类语义特征空间的距离, 统计如表 1 所示。

Table 1 The distance of all problems in problem set to the semantic feature space

表 1 问题集中所有问题到语义特征空间的距离

语义空间				问题域到语义空间距离					样本数
政策文件	分块策略	最大	最小	众数	中位数	平均	标准差		
云南省技术创新中心建设运行管理办法	文本要点		0.677	0.279	0.4(13)	0.424	0.432	0.071	175
	按符号递归		0.863	0.507	0.647(11)	0.669	0.678	0.069	33
	固定长度、重叠大小	150,50	0.858	0.503	0.668(13)	0.67	0.678	0.068	44
		200,50	0.917	0.55	0.712(13)	0.728	0.727	0.061	29
		250,50	0.921	0.567	0.72(11)	0.742	0.746	0.064	22
问题域									624
云南省科技计划项目管理办法	文本要点		0.45	0.178	0.248(25)	0.272	0.279	0.043	417
	按符号递归		0.801	0.42	0.546(18)	0.594	0.603	0.066	66
	固定长度、重叠大小	150,50	0.786	0.412	0.573(27)	0.579	0.587	0.065	89
		200,50	0.836	0.439	0.689(22)	0.641	0.647	0.064	59
		250,50	0.866	0.477	0.625(17)	0.678	0.684	0.064	44
问题域									1 522
云南省重点实验室管理办法	文本要点		0.574	0.26	0.38(13)	0.371	0.377	0.055	251
	按符号递归		0.832	0.478	0.592(13)	0.608	0.619	0.062	57
	固定长度、重叠大小	150,50	0.809	0.458	0.572(12)	0.593	0.602	0.063	78
		200,50	0.886	0.487	0.593(10)	0.648	0.659	0.067	52
		250,50	0.878	0.549	0.656(10)	0.679	0.69	0.064	39
问题域									716

表 1 中, 以技术创新中心管理办法为例, 以文本要点、按符号递归、固定长度为向量化策略, 其中固定长度又细分为 250、200、150 这 3 种长度。最后一列样本数表示该管理办法按对应策略生成的分块数, 按文本要点, 该办法生成了 175 个要点, 按符号递归, 该办法被划分为 33 个文本块。问题域的样本数表示为该办法收集了 624 个问题。最大值表示这 624 个问题向量与其投影向量的距离最大值, 平均值表示这 624 个问题向量与其投影向量的距离平均值。

可以看出, 基于文本要点策略建立的语义特征空间相较于另 2 种策略建立的语义特征空间, 其与问题空间的距离明显小得多, 较好地逼近了领域知识空间。分析其原因如下:

(1) 文本要点是大模型对原文进行凝练得到的总结性文字, 进行词嵌入后其向量蕴含着一定的上下文信息。而其他两种策略只是对原文简单分块, 上下文不完整, 进行词嵌入后其向量也仅包含文本片段本身的微观语义信息。

(2) 固定分块及按符号递归这两种策略, 为了确保一定的上下文, 采用了文本片段头与前一片段尾有文本重叠的做法。有时, 重叠部分与下文并没有直接的上下文联系, 进行词嵌入后其向量中被引入了噪声。

观察发现, 对于固定分块, 分块长度越小, 其到语义空间的距离也越小。当分块长度为 50, 重叠为 10 时, 其到语义空间的距离各项统计甚至远远小于使用文本要点的距离。但在实际检索阶段的文本召回过程中, 其召回率并没有基于文本要点的召回率高, 其原因有待进一步研究。

2 语义特征空间的应用

众所周知, 向量空间 S 外的任意一个向量 v 到 S 中的任意向量 v^* 的距离总是大于或等于 v 在空间 S 中的投影 v^p 到 v^* 的距离, 如式(6)所示。

$$\text{Dist}(v, v^*) \geq \text{Dist}(v^p, v^*) \quad (6)$$

其中, $\forall v^* \in S$ 且 v^p 为 v 在 S 中的投影。

RAG 检索阶段是在语义特征空间中检索与问题最相近的文本, 根据式(6)可知, 问题向量 v 与语义特征空间中最相近的文本向量 v^a 的距离总是大于或等于其在特征空间中的投影向量 v^p 与 v^a 的距离。

以下是“通过验收的条件是什么”这个问题直接在语义特征空间中检索以及用其投影向量检索得到的部分结果。

如表 2 所示, 通过问题向量的投影检索, 其 $\cos$ 距离明

Table 2 Comparison of text recall effect using projection-based search
表 2 使用投影检索召回文本效果比较

直接使用问题向量检索		使用问题向量在语义特征空间的投影检索	
语义相似 top-5 文本	距离	语义相似 top-5 文本	距离
通过验收。项目各项考核指标和任务完成度达到 90% 以上项目经费到位且使用合理合规通过验收。不通过验收。凡具有下列情况之一为不通过验收:……	0.319	通过验收。项目各项考核指标和任务完成度达到 90% 以上项目经费到位且使用合理合规通过验收。不通过验收。凡具有下列情况之一为不通过验收:……	0.225
不通过验收。存在下列情形之一的不予通过验收:任务指标总体完成程度低于 50% 或者约束性指标任意一项未完成的	0.324	不通过验收。存在下列情形之一的不予通过验收:任务指标总体完成程度低于 50% 或者约束性指标任意一项未完成的	0.234
验收专家必须严格遵守国家法律、法规, 具有良好职业道德, 作风正派、坚持原则、责任心强。专家组由相关专业学科技术、经济、财务、管理等方面 5 ~ 7 人组成, 其中财务专家 1 ~ 2 名	0.344	不通过验收。凡具有下列情况之一为不通过验收:项目指标任务完成程度整体低于 60% 或者约束性指标任意一项未完成的	0.243
不通过验收。凡具有下列情况之一为不通过验收:项目指标任务完成程度整体低于 60% 或者约束性指标任意一项未完成的	0.355	第三十四条 项目验收以项目合同书、实施过程中经批准调整的内容为依据对项目实施的组织管理、目标任务完成和项目资金使用情况等进行一次性综合绩效评价。验收评价的主要内容有:……	0.252

显小于直接用问题向量检索得到的距离,更为重要的是其召回的部分文本排序改变了。于是可以作以下猜想:使用文本向量在语义特征空间中的投影到向量数据库中检索可以提高召回率。

为验证以上猜想,分别使用文本要点向量化策略及其他分块策略按上文模型构建各自的语义特征空间,测试结

果如表 3 所示。

原生检索表示问题直接在向量数据库中检索。百分数表示对按相应分块策略生成的矩阵(行向量构成的矩阵)进行 SVD 分解,按百分数取最前面的右奇异向量为基构造语义特征空间。中间数值表示 1 080 个问题在相应语义特征空间中投影后检索能够正确召回原文的问题数。

Table 3 Text recall performance when projected into different ranks of semantic feature spaces
表 3 在不同秩的语义特征空间中投影的文本召回情况

分块向量化策略	测试问题总数	原生检索	SVD 分解后按比例取主特征构造空间					
			100%	80%	60%	50%	30%	15%
文本要点	1 080	909	935	938	939	946	956	955
字符递归(100)	1 080	880	929	936	929	916	866	754
字符递归(200)	1 080	852	908	906	879	872	841	660
固定长度(50,20)	1 080	881	917	907	920	920	907	861
固定长度(100,50)	1 080	862	892	897	903	896	863	772
固定长度(200,50)	1 080	792	859	863	830	818	757	604

可以看出,无论是否使用投影进行检索,基于文本要点策略的召回均优于其他策略。投影到低秩特征空间后,基于文本要点策略的召回在取前 30% 右奇异向量张成的语义特征空间中达到最优。其他策略也分别在不同秩的特征空间中达到最优,如图 1 所示。

这些文本要点多数可能语义相似,对这些文本要点向量化后构成的矩阵进行 SVD 分解,即可取较低的秩以构造其最优语义特征空间。于是可以得出结论:使用文本向量在低秩语义特征空间中的投影到向量数据库中检索可以提高召回率。

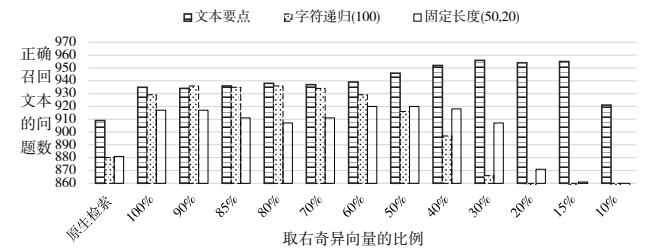


Fig. 1 Text recall performance when projected into different ranks of semantic feature spaces

图 1 在不同秩的语义特征空间中投影的文本召回

基于文本要点策略之所以能取低至 30% 的右奇异向量达到最优,原因是其他策略分块长度相对固定,对于同一段文本而言,其分块数也相对固定,而通过大模型对同一段文本可以生成数倍于其他策略分块数的文本要点,但

3 基于语义特征空间的 RAG 架构

3.1 RAG 架构设计

为了增强 RAG 的性能, Ma^[9]、王海涛^[10]等提出将用户问题提交到大模型之前对问题进行改写的方法,通过改写用户输入的问题,缩小查询与检索对象之间的差距;江双五等^[11]提出通过查询路由对用户问题进行查询通道分配的方案;王合庆等^[12]提出对检索出的文本重排序的方案。

结合以上方案,本文应用语义特征空间模型,选择基于文本要点的向量化策略构建 RAG 检索向量数据库,进一步改进了 RAG 架构。

图 2(a) 是原生 RAG 架构,图 2(b) 是引入语义特征空间模型并基于文本要点向量化检索策略改进的 RAG 架

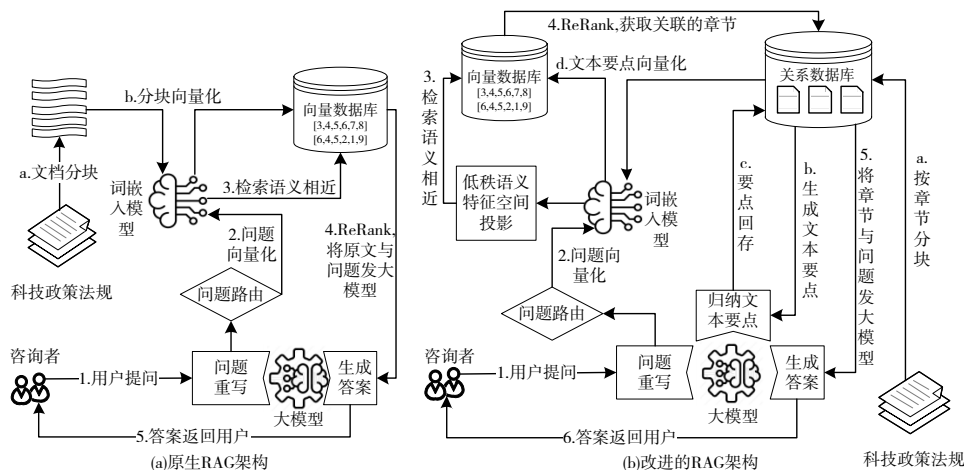


Fig. 2 Comparison of RAG architecture

图2 RAG 架构比较

构。对比两个架构可以看到不同之处:①改进的架构中引入了关系数据库以存贮按章节划分的原文分块,并调用LLM生成分块的若干要点,之后将要点向量化后存入向量数据库,原生RAG则是直接将原文分块向量化后存入向量数据库;②文本要点是对原文信息的浓缩与凝练,不可避免地会丢失原文中的一些关键信息,因而在改进的RAG架构中,文本要点在RAG检索过程中只是起到索引作用,最终传给大模型的是与该要点关联的原文,这也是基于文本要点的向量化检索策略明显区别于其他策略的一个关键之处;③原生RAG中使用问题向量在向量数据库检索与问题文本语义相近的向量,改进架构中则应用上文结论,构造一个最优的低秩语义特征空间并计算问题向量在此空间中的投影,最后通过投影检索语义相近的向量。

从以上不同之处可以看到,改进的RAG架构具有如下优势:①文本要点具有相对独立的主题与完整的上下文,

向量化后不会有过多的噪声与冗余,结合在低秩语义特征空间投影,在RAG检索阶段较原生RAG有更高的召回率;②在RAG增强生成阶段将完整的章节发给大模型与原生RAG将文本片段发给大模型相比,完整的章节具有更完整的上下文,有助于大模型推理,更好地回答问题;③长度适中的文本要点与完整的章节分别作用于检索与增强生成两个阶段,有效解决了原生RAG中较大分块能保持上下文完整性但却带来噪声过大的问题。

3.2 RAG 架构测试

为验证改进的RAG架构与原生RAG架构在文本召回率上的优势,分别选取表3中召回率最高的固定长度(50,20)及字符递归(100)两种向量化策略作为对照组,以文本要点向量化后SVD分解并取30%的主特征向量构造语义特征空间。通过大模型生成的6464个问题分别在改进的RAG架构及原生RAG架构上测试召回率,结果如表4所示。

Table 4 Recall of different RAG architectures

表4 不同RAG架构的召回率

向量化策略及RAG架构	测试问题数	原文本在检索结果中排名							
		top 10		top 5		top 3		top 1	
		召回数	占比/%	召回数	占比/%	召回数	占比/%	召回数	占比/%
固定长度(50,20)+RAG原生架构	6 464	6 430	99.47	6 266	96.94	5 940	91.89	4 718	72.99
字符递归(100)+RAG原生架构	6 464	6 409	99.15	6 156	95.24	5 870	90.81	4 622	71.50
文本要点 RAG原生架构	6 464	6 448	99.75	6 355	98.31	6 190	95.76	5 192	80.32
RAG原生架构+语义特征空间	6 464	6 448	99.75	6 371	98.56	6 232	96.41	5 295	81.92

可以看出,对于3种不同的向量化策略,99%的问题其召回的原文本都能位列前10,排名越靠前,基于文本要点的向量化策略优势越明显,尤其在排名第1的召回上区别更大,传统的固定大小与字符递归只达到72%左右,但文本要点达到80%以上。此外,当在RAG架构中应用本文提出的语义特征空间后,对于排名前5的召回,相较于原生RAG架构,排名越靠前,召回率提升越高,排名第1的召回率甚至提升了1.6%,充分证明了本文提出的文本要点向量化策略和语义特征空间在基于RAG的智能问答应用中的先进性。

4 关于测试的说明

本文聚焦于语义特征空间模型在RAG中的应用,不对大模型的推理进行研究,故测试只针对前文所述的多种分块向量化策略在各自语义特征空间中的多项性能,测试结果如前文所述。测试涉及的环境、数据及测评标准如下:

(1)测试环境。本文使用的向量数据库是Chroma,词嵌入模型是bge-large-zh-v1.5,大语言模型分别是gpt-3.5-turbo、Qwen-plus、SparkLLM-3.0,矩阵SVD分解采用

numpy.linalg。所有代码编写与运行环境是 jupyter notebook。

(2)测试数据。将收集到的33个科技政策与法规文件按章节切分,每一个分块按固定大小、符号递归以及文本要点3种策略二次切分之后用嵌入模型向量化并分别存入向量数据库。对按章节切分的分块调用大模型分别生成若干问题,再使用随机与人工的方式从这些问题中挑选出1080个常见问题用于测试。

(3)测评标准。每一个问题向量化后,如果在向量数据库中的检索结果关联的章节ID与问题关联的章节ID相同,则认为成功召回,否则视为召回失败。

5 结语

本文从优化RAG中的分块向量化策略入手,利用分块向量形成的矩阵通过SVD分解设计了一个语义特征空间模型,并提出了基于文本要点的RAG向量化策略。对上述语义特征空间作进一步研究,得到了将查询文本向量投影到低秩语义特征空间以显著提高文本召回率的方法。通过在RAG的智能问答系统中应用基于文本要点的向量化策略及在此基础上构造的语义特征空间,改进了RAG架构,提高了文本召回率,实现了云南省科技政策智能问答。

此外,本文方法进一步丰富了RAG技术栈。并且,语义特征空间还可用于向量数据库的搜索算法改进。目前,基于知识图谱的RAG研究是当下热点,结合知识图谱,对语义特征空间模型作深入研究,有望进一步优化智能问答的整体解决方案^[13-20]。

参考文献:

- [1] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks [DB/OL]. <http://arxiv.org/abs/2005.11401>.
- [2] GUU K, LEE K, TUNG Z, et al. REALM: retrieval-augmented language model pre-training[C]//Proceedings of the 37th International Conference on Machine Learning, 2020:3929-3938.
- [3] MAO Y, HE P C, LIU X D, 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, 2020:4089-4100.
- [4] JIANG W, WANG Y H, ZHU L, et al. Research on intelligent generation technology of industrial internet thing model based on large language model [J]. Designing Techniques of Posts and Telecommunications, 2024(11): 19-24.
蒋维,王延红,朱亮,等.基于大模型的工业互联网物模型智能生成技术研究[J].邮电设计技术,2024(11):19-24.
- [5] JU W G, WANG P, WANG J. The continuous delivery intelligent question-answering system based on large language models and RAG[J]. Computer Technology and Development, 2025, 35(2): 107-114.
鞠伟刚,汪鹏,王佳.基于大语言模型和RAG的持续交付智能问答系统[J].计算机技术与发展,2025,35(2):107-114.
- [6] HONG L, GUO Y, LIU X L, et al. Intelligent Q&A model of coal mine safety based on RAG [J]. Journal of Heilongjiang University of Science and Technology, 2024, 34(3): 487-492.
洪亮,郭瑶,刘兴丽,等.基于RAG的煤矿安全智能问答模型[J].黑龙江科技大学学报,2024,34(3):487-492.
- [7] CHENG Y, LYU S, CHEN G X. Research on intelligent question answering for standard literature based on large-scale models [J]. Information Technology & Standardization, 2024(8): 38-43.
程云,吕爽,陈国祥.基于大模型的标准文献智能问答技术研究[J].信息技术与标准化,2024(8):38-43.
- [8] HE Z. Application of intensive intelligent knowledge management based on language large model [J]. Wireless Internet Science and Technology, 2024, 21(9): 46-48.
何钊.一种基于语言大模型的强度智能化知识管理的应用[J].无线互联科技,2024,21(9):46-48.
- [9] MA X, GONG Y, HE P, et al. Query rewriting for retrieval-augmented large language models[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023:5303-5315.
- [10] WANG H T, SHI Y K. Research on query extension method based on large language model [J]. Computer Technology and Development, 2025, 35(3): 148-155.
王海涛,师杨坤.基于大语言模型的查询扩展方法研究[J].计算机技术与发展,2025,35(3):148-155.
- [11] JIANG S W, ZHANG J W, HUANG L S, et al. Implementation of meteorological database question-answering based on large-scale model retrieval-augmentation generation [J]. Computer Engineering and Applications, 2025, 61(5): 113-121.
江双五,张嘉玮,华连生,等.基于大模型检索增强生成的气象数据库问答模型实现[J].计算机工程与应用,2025,61(5):113-121.
- [12] WANG H Q, WEI J, JING H Y, et al. Meta-RAG: a metadata-driven retrieval augmented generation framework for the power industry [J/OL]. Computer Engineering. 2024-12-25. <https://doi.org/10.19678/j.issn.1000-3428.0070415>.
王合庆,魏杰,景红雨,等.Meta-RAG:基于元数据驱动的电力领域检索增强生成框架[J/OL].计算机工程.2024-12-25. <https://doi.org/10.19678/j.issn.1000-3428.0070415>.
- [13] EDGE D, TRINH H, CHENG N, et al. From local to global: a graph RAG approach to query-focused summarization [DB/OL]. <https://arxiv.org/abs/2404.16130>.
- [14] JI X Y, WANG X, ZHANG H Y, et al. Knowledge augmentation on traditional Chinese medicine language model [J]. Journal of Frontiers of Computer Science and Technology, 2024, 18(10): 2616-2629.
- [15] GAO Y J, LI R Z, CASKEY J, et al. Leveraging a medical knowledge graph into large language models for diagnosis prediction [DB/OL]. <https://arxiv.org/abs/2308.14321>.
- [16] PENG B, ZHU Y, LIU Y C, et al. Graph retrieval-augmented generation: a survey [DB/OL]. <https://arxiv.org/abs/2408.08921>.
- [17] YE Y, WU J F, LI X N. Knowledge graph construction for cityscape evaluation theories empowered by artificial intelligence [J]. World Architecture, 2024(10): 65-71.
- [18] REN H Y, LIU J P, WANG J, et al. Research on intelligent question answering system based on large language model: a survey [J/OL]. Computer Engineering and Applications. 2024-12-30. <http://kns.cnki.net/kcms/detail/11.2127.TP.20241227.1952.011.html>.
- [19] WEN S, QIAN L, HU M D, et al. Review of research progress on question-answering techniques based on large language models [J]. Data Analysis and Knowledge Discovery, 2024, 8(6): 16-29.
- [20] WANG F, BAO R, WANG S, et al. InfuserKI: enhancing large language models with knowledge graphs via infuser-guided knowledge integration [DB/OL]. <https://arxiv.org/html/2402.11441v2>, 2024.

(责任编辑:孙娟)