

**编者按:**食品是人类赖以生存的物质基础,而传统的食品研发、生产与管理方式正面临着诸多挑战,存在创新性不足的问题。近年来,人工智能(AI)技术发展迅猛,正深刻地影响着人类的思维、生产模式以及生活方式。AI技术是通过计算机模拟人类智能行为的一种技术,它包括机器学习、深度学习、自然语言处理、计算机视觉等多个领域。AI技术也赋能食品行业,其应用不断深化与拓展,已成为全行业创新发展的新质生产力与驱动力。目前,AI技术已被用于食品安全与监管、食品分析与检测、食品加工与生产、食品研发与创新、食品营养与健康等诸多领域。本刊现开设“人工智能+食品”特约专栏,连续刊载相关文章。

(本刊编辑部。客座编辑:江南大学人工智能与计算机学院吴小俊教授。)

## 基于食品知识图谱的智能感知问答

宋晓宁, 蒋启一, 张文杰, 吴小俊

(江南大学人工智能与计算机学院 江苏无锡 214122)

**摘要** 食品安全对人们的生活至关重要,公众对可靠、有效信息的需求日益增强。然而,现有智能问答系统在处理食品安全相关问题时面临诸多挑战,包括知识推理、信息召回率以及问答泛化能力等不足。为此,本文提出一种基于大语言模型(Large language model, LLM)的创新性食品知识图谱智能感知问答系统。该系统通过大语言模型改善传统问题模板匹配方法的不足,结合文本向量库与知识图谱,并在知识图谱检索过程中引入高维向量语义检索技术,以显著提升知识图谱的检索召回率。试验结果表明,该系统在知识检索的准确性和效率方面均有显著提升,用户的使用体验得到明显改善。综上所述,所提出的食品知识图谱智能感知问答系统,不仅有效解决了现有系统的多项不足,还为食品安全领域的信息获取提供了高效、准确的方式,促进了社会治理的透明性和公众信任,具有广泛的应用前景与重要的社会价值。

**关键词** 人工智能; 食品安全; 知识图谱; 大语言模型; 智能问答

文章编号 1009-7848(2024)12-0001-12 DOI: 10.16429/j.1009-7848.2024.12.001

食品安全直接关系人们的生活质量和身体健康。随着公众对食品安全问题关注度的不断提高,健康饮食对身体健康的重要性也愈发突出。迫切需要一个智能问答系统来解决关于食品安全的问题。近年来,智能问答系统<sup>[1]</sup>得到迅速发展,其中基于知识图谱<sup>[2]</sup>的问答系统取得显著进展。通过知识图谱的构建并依赖于实体识别、关系匹配和模板匹配等技术,实现了与用户问题相关内容的有效检索。然而,现有智能问答系统在处理食品安全等特定领域问题时仍面临诸多挑战。首先,知识图谱构建过程复杂,构建成本过高,导致食品安全领域的知识图谱相对匮乏;其次,传统的知识图谱检索技术的检索召回率较低,无法充分利用构建好的

知识图谱信息;此外,大多数问答系统都是基于问题模板匹配<sup>[3]</sup>的方法,而此类方法在模板泛化、灵活性和上、下文理解方面表现不足,难以应对复杂、多变的问题情境。

为解决上述问题,本文提出一种基于大语言模型的食品安全智能感知问答系统。首先,利用爬虫技术广泛采集食品安全领域的公开数据,并通过光学字符识别(Optical character recognition, OCR)<sup>[4]</sup>、表格识别、数据清洗等技术搭建大规模食品安全知识库。接着,设计了基于信息抽取大模型的自动化知识图谱构建技术,准确构建并扩展食品安全领域的知识图谱,显著提升了图谱的覆盖面和准确性,以弥补当前知识图谱资源的不足。

同时,随着大语言模型在用户意图识别和用户交互表现上的突破,以及基于向量检索的检索增强生成(Retrieval-augmented generation, RAG)<sup>[5]</sup>技术的成熟,本文提出一种基于大语言模型的知识图谱问答范式,以提升系统的检索召回率和回

收稿日期: 2024-10-29

基金项目: 国家重点研发计划项目(2023YFF1105102, 2023YFF1105105)

第一作者: 宋晓宁,男,博士,教授

E-mail: x.song@jiangnan.edu.cn

答准确性。所构建的智能问答系统不仅为公众理解食品安全问题提供了重要的支持途径,而且是推动社会治理现代化和提升公共安全保障的重要举措。该系统能够为监管机构、企业和消费者提供精确、及时的信息服务,助力构建更安全、透明的食品供应链,为实现“健康中国”战略和社会治理智能化奠定坚实基础。

## 1 研究现状

谷歌于2012年首次推出基于知识图谱的搜索引擎,这一开创性举措极大推动了通用知识图谱的普及,例如在多领域应用广泛的CN-DBpedia<sup>[7]</sup>。然而,通用知识图谱在专业细分领域(如食品安全)难以满足更深层次的需求,因而激发了研究者对特定领域知识图谱的兴趣与探索<sup>[8]</sup>。近年来,食品安全领域的知识图谱研究逐步推进。比如,华中农业大学的翁永琳<sup>[9]</sup>构建了食品安全标准知识图谱,并设计了可视化查询系统,以便于用户查阅和展示相关信息。卜意磊等<sup>[10]</sup>采用BiLSTM-CRF模型识别实体,构建了面向食品安全监管的知识图谱。管铮懿等<sup>[11]</sup>基于Neo4j数据库,针对食品添加剂的重要性,构建了食品安全知识图谱。此外,周锴<sup>[12]</sup>基于时序动态知识图谱模型,实现了食品安全风险预测。

与此同时,文本向量库技术的应用,显著增强了智能问答系统在语义检索和信息匹配上的性能。通过将文本转化为向量表示,文本向量库有效提升了检索精度。例如,Chen等<sup>[13]</sup>指出,知识图谱已成为人工智能中的核心资源,结合深度学习进一步提高了系统性能。Qiao等<sup>[14]</sup>基于图神经网络编码知识图谱的结构和语义信息,将模型准确率提升至83.94%。然而,这些深度学习方法通常对大量标注数据和高性能硬件的需求较高,增加了系统的资源投入和时间成本。此外,在处理开放领域问题时,系统生成的答案往往冗长且不够简洁,影响了实际应用效果。

虽然食品安全领域的知识图谱和文本向量库技术已取得一定进展,但是仍面临若干挑战。食品安全知识图谱的规模相对较小,其信息检索需要更为专业的知识支持。现有基于知识图谱的智能问答系统在回答准确性和效率方面均有提升。然

而,这些系统在知识图谱的扩展、信息融合以及精准回答等方面仍有进一步优化的空间,以更好地满足食品安全领域的需求。

## 2 数据与方法

### 2.1 数据获取

为构建全面的食品安全知识库,作者团体从多个公开渠道采集丰富的专业数据,涵盖食品资讯、专业书籍、法规文件及标准规范等,确保信息来源的权威性和多样性。数据来源按内容类型和信息密度分为3个层次:第1层为实时资讯来源(食品资讯网站),用于提供最新的食品安全事件和行业趋势;第2层为专业参考数据(教科书、法规文件等),提供深层次的领域知识和政策支持;第3层为行业标准数据,帮助规范食品安全知识库的内容并确保合规性。以上数据来源在时间、内容和质量上均得到充分保证,通过多样化、结构化的数据构建食品安全领域的综合知识库。

1) 食品资讯网站 选择20余家权威食品资讯网站(如食品伙伴网、FoodTalks和Foodaily),从这些平台获取食品安全动态、行业趋势和科学研究等信息。食品资讯网站的数据量庞大,更新频率高,是获取最新食品安全资讯的重要来源。在筛选过程中,优先选择具有较高行业影响力和专业认可度的网站,以保证知识库的时效性和前沿性。

2) 食品专业书籍 为了涵盖食品科学基础知识和实践规范,选用70余本食品安全领域的教科书和专业参考书。这些书籍提供了食品安全的基础理论、关键概念及实操规范,为知识库的核心框架奠定了坚实的理论基础。

3) 法律法规、国家标准和行业标准等标准 基于食品行业合规和安全控制的需求,纳入国家层面的食品法律法规(如《食品安全法》)和20余项国家和行业标准等标准(如食品添加剂使用标准)。这些权威文件确保了知识库对食品安全的合规性和政策覆盖,具体数据分布如图1所示。

在网页数据采集集中,采用基于Python的网络爬虫技术来高效获取食品安全领域的核心内容。通过定向的爬虫配置和策略设置,批量采集食品资讯网页上的关键信息,包括食品安全新闻、产品公告、食品成分信息以及行业动态等内容。为确保

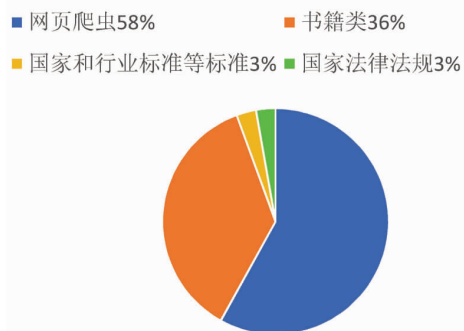


图 1 数据来源

Fig.1 Data sources

内容的准确性,设计了网页文档解析器,从网页结构中精确定位所需的文本元素,剔除 HTML 标签及其它非关键信息,使输出的文本内容清晰规范,便于后续的分析 and 知识库整合。

此外,通过适当的请求频率控制、请求头设置和代理使用等技术手段,降低了爬虫访问受限风险,同时保证了数据采集的连续性和覆盖广度。对于国家相关法律法规、行业标准及其它重要文档,通常直接从相关网站下载 PDF 文件。对于 PDF 格式的书籍和法律法规文件,尤其是非文本格式的文档,作者使用 OCR 技术<sup>[15]</sup>进行文本识别,将图像化的数据转化为可编辑的文本。OCR 识别的文本经过清洗与整理,去除了排版格式和图像噪点,确保了数据的规范性和结构性。

最终,PDF、HTML 等数据被合并至本地知识库。在数据收集完成后,所有食品相关的文本数据均经过严格的数据清洗与标准化处理,包括去除特殊符号、统一食品术语和单位格式等,以确保数据的规范性与一致性。这种处理不仅提高了数据质量,还为后续的知识提取和分析提供了更高的准确性与一致性,保障了食品安全知识库在应用中的科学性与实用性。

## 2.2 知识图谱构建

知识图谱在知识问答系统中具有重要作用。传统的知识图谱问答<sup>[16]</sup>虽然在信息检索上具备精准性,但是通常表现生硬,难以灵活地理解用户的真实意图。这是因为传统知识图谱问答依赖于预定义的问题模板和显式的关系匹配,所有在处理自然语言中的模糊表达、复杂上、下文和长尾问题

时常显得捉襟见肘。用户体验也因此受限,系统无法动态适应多样化的查询方式。

另一方面,单纯依靠大语言模型进行知识问答,虽然在对话流畅度和语义理解上有较大优势,但是模型的生成方式容易导致“幻觉”现象<sup>[17]</sup>,即模型生成的信息看似合理或有依据,但实际与事实不符或完全虚构。这种现象在生成自然语言文本时较为常见,表现为模型编造不存在的数据、事件、实体、引用,甚至提供错误的逻辑推导。此外,大语言模型缺乏结构化知识库支撑,对一些专业领域的准确性和一致性难以保障,导致回答质量的不确定性。

通过将知识图谱与大语言模型结合,这两者的优势得以互补。知识图谱提供了严谨的知识结构,确保信息的准确性和可靠性;大语言模型则提供自然语言理解能力,使问答更加灵活且富有智能性。结合知识图谱的大语言模型,不仅能够更准确地理解用户意图,避免模板生硬带来的限制,还能在回答过程中依托知识图谱的结构化数据源,有效减少“幻觉”现象的发生,提升问答的精准性和可信度。因此,结合知识图谱的智能问答系统在提升用户体验和信息可靠性方面具有显著优势。

2.2.1 知识图谱模式层构建 本文采用三元组的方式对知识图谱  $G$  中的知识进行表示,即  $S=\{E, R, S\}$ ,其中  $E$  为所有实体的集合, $R$  为所有关系的集合, $S=\{e_1, r, e_2\}$  为知识库中所有三元组的集合。这里  $e_1$  和  $e_2$  分别为头实体和尾实体, $r \in R$  为两个实体间对应的关系。根据《食品安全国家标准 食品添加剂使用标准》(GB 2760-2024),设计了食品安全知识图谱的模式层。设计的实体类型共有 8 种,具体如表 1 所示。

对于食品各类实体,根据《食品安全国家标准 食品添加剂使用标准》(GB 2760-2024)中的附录 E(食品分类系统)将其划分为 5 个类别,分别为食品大类(一级)、食品亚类(二级)、食品品种(三级)、食品细类(四级)和食品亚细类(五级),相邻级别食品类别的实体之间为“划分”关系。各级食品类别实体与知识来源实体之间为“包含”关系。食品营养成分与各级食品类别实体之间为“含有”关系。食品营养成分和营养成分功能之间为“具有”关系。



表1 实体类型及对应属性

Table 1 Entity types and corresponding attributes

实体类型	对应属性
食品大类	食品名称、食品描述、Embedding 编码、节点编号
食品亚类	食品名称、食品描述、Embedding 编码、节点编号
食品品种	食品名称、食品描述、Embedding 编码、节点编号
食品细类	食品名称、食品描述、Embedding 编码、节点编号
食品亚细类	食品名称、食品描述、Embedding 编码、节点编号
食品营养成分	成分名称、成分描述、Embedding 编码、节点编号
营养成分功能	功能描述、Embedding 编码
知识来源	知识来源详细信息、Embedding 编码、节点编号

2.2.2 食品知识图谱构建 在信息抽取部分,采用 OneKE 信息抽取大模型。OneKE 信息抽取大模型相较于传统模型具备显著优势。它利用深度学习技术有效捕捉文本中的复杂语义关系,实现对实体、关系及属性的高准确度识别。此外,OneKE 通过预训练在海量数据上进行训练,具备较强的泛化能力和迁移学习能力,能够适应不同领域的

知识抽取任务,并支持多种数据源输入,包括结构化和非结构化文本,从而提高了实际应用的灵活性与效率。具体而言,设计了 Instruction 来指导 OneKE 信息抽取模型进行三元组的抽取。Instruction 的格式采用类 JSON 字符串的结构,实质上是一种字典类型的字符串,如表 2 所示。

表2 Instruction 组成字段

Table 2 Components of instruction

字段名称	字段作用
instruction	使用自然语言明确指出模型的角色和任务,告诉模型需要完成的工作
schema	待提取的标签列表,指明需抽取信息的关键字段,动态响应用户需求
input	信息抽取的源文本,即需要从中提取三元组的文本数据

在设计关系抽取规则时,本文从逻辑、效率和冗余 3 个方面制定严格的标准,以避免因标准不一致而导致的知识图谱差异性。逻辑:遵循知识图谱模式层的逻辑设计,确保构建的图谱符合实际业务需求,具备良好的可扩展性。效率:为了提升查询效率,系统只存储必要的信息。例如,对于图片属性,仅保存其链接地址,而不是图片的实际内容。冗余:对于多次重复出现的实体和关系,系统只存储 1 次,减少数据冗余,优化存储空间和查询性能。

通过这些策略,本文构建的知识图谱在逻辑严谨性、查询效率和数据精简性上均达到预期目标,为后续的智能问答提供了坚实的知识基础。

2.2.3 知识图谱数据库存储和可视化 为了高效存储和可视化构建的三元组数据,本文采用主流的开源图数据库 Neo4j 作为知识图谱的存储工

具。Neo4j 支持功能强大的 Cypher 查询语言,使用户能够便捷地查询三元组数据,并将实体与关系快速导入为图中的节点与边,便于进一步分析和展示。

传统知识图谱的索引和检索方法通常依赖于字符串匹配,即通过实体名称或属性字段的文本匹配进行查找。然而,这种方法在处理具有复杂语义关系的大规模图谱时检索效率低,往往无法准确识别和区分语义相似的节点,影响了系统的响应质量和灵活性。为克服这一局限,本文引入最新的向量索引技术来增强图谱中节点的检索能力。

通过向量索引,利用节点的嵌入向量表示数据对象的语义特征,将数据对象(如文本、图像、音频或文档)转换为高维向量,以捕捉其深层语义信息。这使得具有语义相似性的节点在向量空间中相邻,系统可以基于这些嵌入向量实现高效、准确

的查询。在本文的系统中,采用 Bge-embedding 和 Zhipu-embedding 模型作为主要的嵌入生成器,将文本数据转化为浮点数向量,借助 k-近邻(k-NN)查询来识别大规模数据集中语义相近的节点。具体可以表示为,给定一个文本数据集 $\{T_1, T_2, \dots, T_n\}$ ,使用 Bge-embedding 和 Zhipu-embedding 模型  $f$  将每个文本  $T_i$  映射到  $d$  维空间中的浮点数向量,如公式(1)所示:

$$f: T_i \rightarrow v_i \in R^d, i=1, 2, \dots, n \quad (1)$$

生成的嵌入集为 $\{v_1, v_2, \dots, v_n\}$ ,其中每个  $v_i$  都表示一个对应文本的向量表示。

要度量文本向量间的语义相似度,可以使用余弦相似度,如公式(2)所示:

$$\text{Cosine\_similarity}(v_i, v_j) = \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|} = \frac{\sum_{k=1}^d v_{i,k} \cdot v_{j,k}}{\sqrt{\sum_{k=1}^d v_{i,k}^2} \sqrt{\sum_{k=1}^d v_{j,k}^2}} \in (0, 1] \quad (2)$$

式中,  $v_{i,k}$  表示向量  $v_i$  的第  $k$  个维度的分量。

对于每一个向量  $v_i$ , 寻找其  $k$  个最近邻向量(即语义最相似的向量),具体计算方法如公式(3)所示:

$$\text{kNN}(v_i) = \arg \text{topk}_{\{v_j\}_{j \neq i}} \text{Cons\_similarity}(v_i, v_j) \quad (3)$$

即对于每个  $v_i$ , 计算其与其它所有向量 $\{v_j\}_{j \neq i}$ 的余弦相似度,并选择其中相似度排名前  $k$  的向量  $v_j$ 。综上所述,整个过程可以表示如公式(4)所示:

$$\text{kNN}(f(T_i)) = \{v_j \mid v_j = \arg \text{topk}_{\{v_1, v_2, \dots, v_n\} \setminus \{v_i\}} \frac{\sum_{k=1}^d v_{i,k} \cdot v_{j,k}}{\sqrt{\sum_{k=1}^d v_{i,k}^2} \sqrt{\sum_{k=1}^d v_{j,k}^2}}\} \quad (4)$$

这个公式表示,对于给定的嵌入向量  $v_i$ , 通过计算与其它所有向量的余弦相似度,可以获得与之最相似的  $k$  个向量,从而找到再语义上与  $T_i$  最接近的节点。

通过结合三元组存储与向量索引技术,本文构建的食品安全知识图谱不仅实现了结构化的实体和关系展示,还能基于语义相似性进行快速检索,显著提升了问答系统的智能性和响应效率。相较于传统字符串匹配方法,具有更高的准确性和适应性。知识图谱和知识库构建方法如图 2 所示。

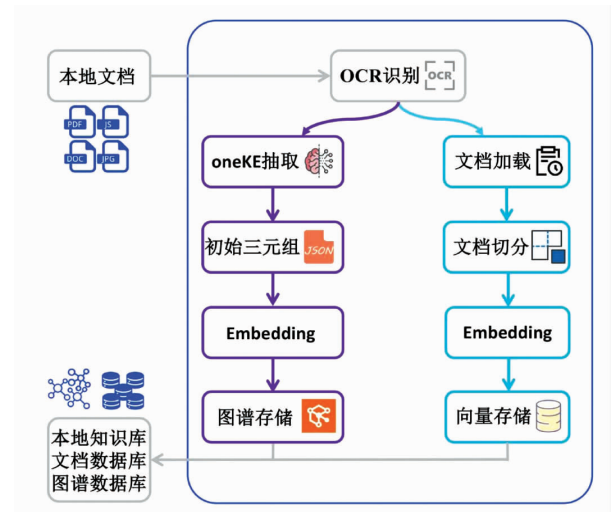


图 2 知识图谱和知识库构建方法

Fig.2 Construction methods of knowledge graph and knowledge base

### 2.3 知识库构建

在知识图谱之外,还额外构建了食品安全知识库作为食品知识图谱智能感知问答系统的一个关键部分。使用知识库的核心目的是为问答系统提供丰富且准确的领域信息支持,增强回答的准确性和覆盖面。为了实现高效的知识检索与生成,本文采用当前流行的 RAG 技术。RAG 技术将信息检索和生成模型结合,能够在回答复杂问题时动态地从知识库中获取相关内容,以弥补生成模型在领域知识方面的不足。

在本地语料库构建阶段,从预处理后的数据出发,进一步处理文本以支持向量化和检索操作,主要包括:

1) 文本数据加载 在预处理阶段,数据已被转化为非结构化文本。首先,使用适当的文件加载方法将这些非结构化文本数据导入系统,以确保数据一致、可用。

2) 数据分割 文本分割时需考虑两大因素:一是 Embedding 模型的 Tokens 限制,二是语义完整性对整体检索效果的影响。为此,提供了多种文本切分策略,可根据具体需求灵活选择。例如,ChineseRecursiveTextSplitter 策略可以递归地将中文文本分割成更小的部分,直到满足指定的 Tokens 限制或其它规则。此方法特别适用于处理长句或长段落,从而提升文本处理的效率和语义完

整性。

3) 向量数据库构建 本文采用 Bge-embedding<sup>[18]</sup>和 Zhipu-embedding 两种嵌入模型对分割后的数据生成语义向量。这些向量捕捉了文本的深层语义特征, 确保系统能够在语义空间中准确地定位相关信息。生成的向量数据存储于高性能向量数据库 Milvus<sup>[19]</sup>中, 以便系统通过向量检索高效地从知识库中提取相似信息。

## 2.4 方法

2.4.1 方法概述 在本研究中提出一种基于知识图谱与 RAG 技术相结合的方法, 致力于提升智能问答系统在食品安全领域的性能。RAG 技术结合了信息检索与生成式模型的优势, 允许问答系统在处理复杂查询时能够从知识库中动态提取相关内容, 从而弥补单纯生成模型在领域知识上的不足。

系统的问答流程包含 3 个核心步骤:“召回-排序-回答”。在召回阶段, 用户的查询内容会首先通过实体识别与嵌入向量技术在知识图谱和知识库中检索出相关内容, 确保涵盖尽可能多的有效信息。在排序阶段, 通过排序模型进一步优化召回内容的优先级, 从而过滤出最相关的知识片段, 确保输出的质量。最终, 生成模型基于召回的高相关度信息, 提供准确而权威的回答。

传统问答模型在处理自然语言的复杂性、上下文理解和长尾问题方面仍存在局限。本系统通过将大语言模型与知识图谱相结合, 一方面利用知识图谱的结构化数据来降低生成模型的幻觉现象, 确保回答的准确性; 另一方面借助 RAG 技术增强系统的灵活性和扩展性, 确保在食品法规、营养咨询等领域的问答质量与用户体验显著提升。本系统整体的知识图谱和知识库检索方法如图 3 所示。

2.4.2 知识图谱检索 命名实体识别(Named Entity Recognition, NER) 是从用户给定的自然语言中识别出具有特定意义的实体, 它是自然语言处理(Natural Language Processing, NLP)领域的一项基本任务。由于在汉语中词和短语的界限模糊, 如何使计算机准确识别出汉语中的实体, 一直是自然语言处理领域的一个难题。本研究设计了基于提示词工程指导大预言模型对用户提问进行

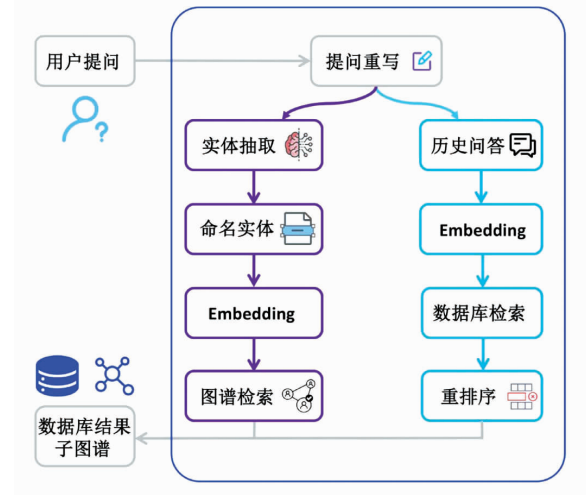


图 3 知识图谱和知识库检索方法

Fig.3 Retrieval methods of knowledge graph and knowledge base

命名实体识别的方法。对于抽取出来的实体, 采用和知识图谱相同的嵌入模型对其进行嵌入操作, 得到实体嵌入向量  $\vec{v}$ 。然后, 使用向量索引对实体进行相似度检索, 目前支持的相似度函数分别有 Euclidean (欧几里得) 相似和 Cosine (余弦) 相似度。

使用欧几里得距离计算相似度的方法如公式 (5) 所示:

$$\text{Euclidean\_similarity}(\vec{v}, \vec{u}) := \frac{1}{1 + \|\vec{v}, \vec{u}\|^2} \in (0, 1] \quad (5)$$

式中:  $\vec{v}$  表示用户提问中的实体嵌入向量;  $\vec{u}$  表示知识图谱数据库中的实体嵌入向量。

使用余弦距离计算相似度的方法如公式 (6) 所示:

$$\text{Cosine\_similarity}(\vec{v}, \vec{u}) = \frac{\vec{v} \cdot \vec{u}}{\|\vec{v}\| \|\vec{u}\|} \in (0, 1] \quad (6)$$

式中:  $\vec{v}$  表示用户提问中的实体嵌入向量;  $\vec{u}$  表示知识图谱数据库中的实体嵌入向量。

通过这样的检索方式可以找到知识图谱中与用户提问最相似的几个实体节点。为了保证模型回答的准确性, 通过设置 Cypher 语句中的关键字, 使得最后只返回相似度较高的节点的指定跳数以内的子图谱。

最后, 如果在语义相似度匹配中没有找到符

合要求的实体,并且本次提问不是用户的首次提问,系统会将本次对话视为多轮提问,继承上一轮对话中的实体集继续回答。

2.4.3 知识库检索 在知识库检索部分,首先利用大语言模型对用户的提问进行重写,以实现系统的连续问答功能。经过重写后,用户的提问会被转换为高维向量,便于后续的检索操作。这些查询向量将输入 Milvus 数据库进行检索,Milvus 通过预先构建的索引对存储的向量集合进行快速筛选,显著提高了检索效率。

在检索过程中,Milvus 对候选向量和用户查询向量进行相似度计算,采用的度量方法包括欧氏距离、余弦相似度和内积等。这些方法帮助系统有效评估向量间的相似性,确保返回的结果具有较高的相关性。根据设定的 Top-K 参数,Milvus 会返回与查询向量最相似的 K 个向量,用户还可以根据需求设定相似度阈值,过滤掉不相关的结果。最终,系统将返回最相似的向量及其对应的数据实体,为后续分析和查询结果展示提供支持。

此外,为了进一步优化检索结果,还调用 Bge-Rerank 模型,对初步检索结果进行二次排序,并设置阈值以筛选出更相关的结果,从而确保最终的知识库检索结果更加准确、高效。Bge-Rerank 模型工作过程表示为:

对于用户查询  $q$  和候选内容集合  $D=\{d_1, d_2, \dots, d_n\}$ ,使用 Bge-embedding 模型生成全嵌入向量:

$$v_q = f_{Bge}(q) = W_q \cdot \Phi(q) + b_q \quad (7)$$

式中: $f_{Bge}(q)$ 表示 Bge-embedding 模型; $W_q$ 为模型的权重矩阵; $\Phi(q)$ 为查询的词向量或特征表示; $b_q$ 为偏置项。

候选文档向量生成,对于每个候选片段  $d_i$ 生成嵌入向量  $v_{d_i}$ ,如公式(8)所示:

$$v_{d_i} = f_{Bge}(d_i) = W_q \cdot \Phi(d_i) + b_d \quad (8)$$

通过余弦相似度计算查询向量  $v_q$  和每个候选文档向量  $v_{d_i}$  的相似度得分,得到初步候选集合  $D'$ ,如公式(9)所示:

$$\text{sim}(v_q, v_{d_i}) = \text{Cosine\_similarity}(v_q, v_{d_i}) =$$

$$\frac{v_q \cdot v_{d_i}}{\|v_q\| \|v_{d_i}\|} \quad (9)$$

基于设定的相似度阈值  $\tau$ ,筛选出满足条件的候选项,如公式(10)所示,确保初步召回的准确性:

$$D' = \{d_i \in D \mid \text{sim}(v_q, v_{d_i}) \geq \tau\} \quad (10)$$

或采用  $k$ -近邻查询的方式,从相似度最高的  $k$  个候选项中筛选出满足阈值的项,如公式(11)所示:

$$D' = \{d_i \in D \mid \text{Topk}(\text{sim}(v_q, v_{d_i})) \text{ and } \text{sim}(v_q, v_{d_i}) \geq \tau\} \quad (11)$$

对初步拆选出的候选集合  $D'=\{d'_1, d'_2, \dots, d'_k\}$ ,使用排序模型重新计算每个候选项的相关性分数  $s_i$ ,如公式(12)所示:

$$s_i = g_{\text{Rerank}}(v_q, v'_{d_i}) = W_s \cdot \sigma(v_q \oplus v'_{d_i}) + b_s \quad (12)$$

式中: $g_{\text{Rerank}}$ 表示排序模型; $W_s$ 和  $b_s$ 分别为模型的权重和偏置; $\oplus$ 表示向量链接; $\sigma$ 为激活函数。

通过这些操作,系统能够为用户提供快速而精准的答案,显著提升响应速度和查询结果的相关性,从而极大地优化用户体验。

2.4.4 问答 该部分主要实现对用户提问的回答,如图 4 所示。系统首先利用大语言模型对用户的提问进行初步的意图判断。若提问无需检索,或在知识图谱和知识库中找不到相关内容,系统将

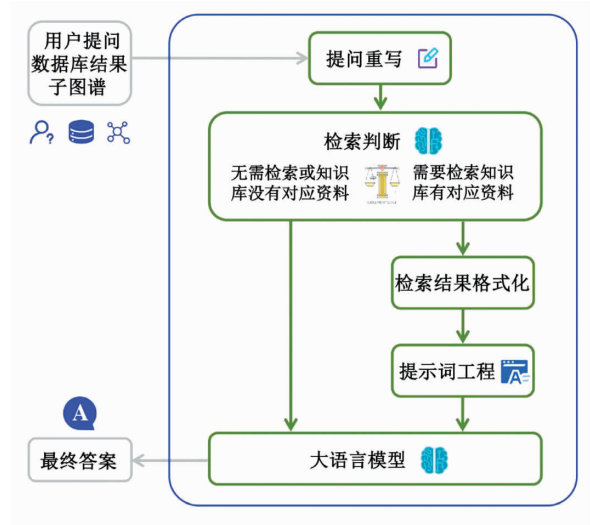


图 4 大语言模型问答方法

Fig.4 Large language model question answering method



直接生成回答;反之,如果提问涉及检索且知识图谱或知识库中存在相关信息,系统将提取相应的子图谱和知识片段,并对其进行格式化处理。格式化的目的在于将检索结果整理为更具可读性的连续文本。随后,使用提示词工程(Prompt engineering)<sup>[20]</sup>将这些格式化信息输入大语言模型,以生成最终的回答。这种基于大语言模型的问答方法相较于传统的问题模板匹配方法,显著提升了系统在泛化能力、上、下文理解能力、处理模糊与复杂问题以及多语言支持等方面的表现,从而改善了用户体验。

### 3 案例分析

以上研究表明,基于知识图谱的感知问答系统在食品知识领域展现明显的优势。为更直观地展示系统的优越性,接下来的案例分析将具体比较传统问答系统与本系统在处理用户需求时的表现。这些实例不仅突显了系统的创新能力,也为提升食品安全和社会治理提供了切实的支持。

#### 3.1 个性化营养与健康建议:开放性与长尾问题处理

传统问答系统在提供营养建议时,通常只能回答一些预设的、通用性较强的问题,如“什么是健康的早餐?”这类模板化问题显得生硬,难以满足用户在实际生活中的复杂需求,特别是涉及个性化需求或长尾问题时,这类系统的表现更显不足,难以提供深度定制化的建议或满足用户多样化的问答场景。

食品知识图谱智能感知问答系统的优势在于本系统基于大语言模型,能够理解自然语言中的复杂表达,并利用嵌入向量检索方法,从知识图谱中召回与用户查询相关的多维信息。例如,用户询问“我患有乳糖不耐症并想增肌,有哪些适合的食谱?”传统模板化系统难以处理这种开放性和长尾问题,而本系统可以准确理解用户需求,并在知识图谱中检索出与乳糖不耐症、增肌饮食相关的营养建议、食材替代方案等信息,提供个性化、精准的回答,如表3所示。

表3 开放性与长尾问题处理

Table 3 Handling of open-ended and long-tail questions

传统问答系统	知识图谱感知问答系统
用户:我最近被诊断出患有乳糖不耐症,想增肌,有什么适合我的食谱吗?	用户:我最近被诊断出患有乳糖不耐症,想增肌,有什么适合我的食谱吗?
系统:您可以尝试一些健康的早餐,如燕麦粥、鸡蛋和水果。这些食物通常被认为是健康的选择。	系统:根据您的需求,以下是几个适合乳糖不耐症患者的增肌食谱建议: 1) 燕麦早餐碗:将燕麦与无糖植物奶(如杏仁奶或燕麦奶)煮熟,加入一些坚果、切片香蕉和少量蜂蜜,提供丰富的碳水化合物和健康脂肪。 2) 鸡胸肉沙拉:用烤鸡胸肉、混合蔬菜(如生菜、菠菜、黄瓜)和橄榄油调味,提供高蛋白质和纤维素。 3) 豆腐炒蔬菜:将豆腐与西兰花、红椒和胡萝卜一起翻炒,使用酱油和姜蒜调味,富含植物蛋白和维生素。
用户:那我吃牛奶和奶酪呢?	用户:听起来不错!那我吃牛奶和奶酪呢?
系统:不建议。乳糖不耐症患者应该避免乳制品。	系统:不建议。乳糖不耐症患者应该避免乳制品。不过,您可以选择一些乳糖酶补充剂,帮助消化乳糖,或者尝试乳糖含量较低的奶酪,如切达奶酪。

#### 3.2 复杂食材替代与菜谱优化:用户友好性

在烹饪领域,用户常常会遇到多重限制条件下的食材替代问题,比如“在无坚果、低糖的情况下,如何替代蛋糕中的杏仁?”传统的基于模板的问答系统通常需要用户按照固定格式输入问题,若超出预设范围,系统就难以给出有效答案。这种生硬的交互方式不仅限制了用户的表达,还降低

了使用体验。

食品知识图谱智能感知问答系统的优势在于本系统通过自然语言交互,让用户可以任何形式提出问题,无需考虑固定的表达方式。系统利用大语言模型理解复杂的替代需求,并通过嵌入向量检索,全面检索知识图谱中的相关信息。例如,对于上述问题,系统不仅能提供可行的替代食材,还



会建议如何调整菜谱以确保成品的口感和营养。这种交互方式更加符合用户的语言习惯,显著提升了系统的用户友好性,同时避免了传统模板化问答的生硬性,如表 4 所示。

### 3.3 食品法规与合规咨询:高召回率与用户友好性结合

在食品行业中,合规咨询是一个重要环节。传统模板化系统只能回答一些固定的法规问题,如“食品添加剂的使用标准是什么?”一旦问题超出模板的范围,系统就难以给出准确的回答。此外,由于信息检索方式的限制,传统系统无法全面涵

盖法规的各个方面,导致召回率较低。

食品知识图谱智能感知问答系统的优势在于本系统利用嵌入向量检索方法,将法规条文、标准要求等信息转化为向量形式进行检索,提高了召回率。例如,当企业咨询“出口到某国的食品需要满足哪些法规要求?”系统可以在知识图谱中检索到与该国外所有法规信息,提供完整且详细的合规建议。同时,用户可以自然语言形式随意提问,无需考虑特定格式,这种交互方式使得咨询过程更加流畅、自然,体现出系统的用户友好性,如表 5 所示。

表 4 用户友好性

Table 4 User friendliness

传统问答系统	知识图谱感知问答系统
<p>用户:在无坚果、低糖的情况下,如何替代蛋糕中的杏仁?</p> <p>系统:请提供更具体的问题格式,您可以问“如何替代杏仁?”</p> <p>用户:我说的是无坚果和低糖的情况下替代蛋糕中的杏仁。</p> <p>系统:对不起,系统无法处理该请求。</p>	<p>用户:在无坚果、低糖的情况下,如何替代蛋糕中的杏仁?</p> <p>系统:针对您的需求,以下是几个替代杏仁的建议:</p> <ol style="list-style-type: none"> <li>1) 椰子粉:可以用椰子粉替代杏仁粉,提供类似的质地和香气,且含糖量低。</li> <li>2) 亚麻籽粉:亚麻籽粉是一个很好的替代品,富含纤维和健康脂肪,可以增加蛋糕的营养价值。</li> <li>3) 南瓜籽粉:这种粉末不仅低糖,还能提供丰富的蛋白质和矿物质,适合无坚果的饮食。</li> </ol> <p>用户:那在配方中,我应该如何调整以保持口感?</p> <p>系统:您可以在使用椰子粉或亚麻籽粉时,适当增加液体成分(如水或植物奶),以确保蛋糕的湿润度和口感。另外,如果您使用南瓜籽粉,建议加入一点点香精以提升风味。</p>

表 5 高召回率与用户友好性结合

Table 5 Combining high recall rate and user friendliness

传统问答系统	知识图谱感知问答系统
<p>用户:出口到某国的食品需要满足哪些法规要求?</p> <p>系统:对不起,我无法提供该国的具体法规信息。您可以查阅相关的法律文件。</p> <p>用户:那食品添加剂的使用标准是什么?</p> <p>系统:食品添加剂的使用标准是……(提供固定答案)</p>	<p>用户:出口到某国的食品需要满足哪些法规要求?</p> <p>系统:根据您的提问的国家,以下是相关的法规要求:</p> <ol style="list-style-type: none"> <li>1) 标签法规:所有出口食品必须标明成分、过敏原和营养信息。</li> <li>2) 卫生标准:食品需符合特定的卫生标准,包括微生物和化学污染物的限制。</li> <li>3) 包装要求:包装材料必须符合该国的安全标准,确保食品在运输过程中的安全。</li> <li>4) 检验和检疫:进入该国的食品需经过检验,确保符合当地的检疫要求。</li> </ol> <p>此外,您可以随时咨询具体的条款或其他相关问题,我将为您提供更详细的信息。</p>

### 3.4 大语言模型幻觉的改善

在食品行业,卫生问题的处理至关重要。当用户询问餐厅卫生问题的赔偿标准时,传统的大语

言模型由于幻觉问题,可能仅回答“您可能会获得赔偿”,无法提供具体的赔偿金额,导致信息模糊,用户难以作出决策。

相比之下,食品知识图谱智能感知问答系统能够清晰地识别问题,并根据知识库中的法规提供准确的信息。例如,当用户询问“如果饭店出现卫生问题,我应该得到多少赔偿?”时,系统能直接回答“根据相关法律,您可以获得十倍赔偿。”这种准确性消除了大语言模型的幻觉问题,为用户提供了专业可靠的信息支持,增强了信任感和满意度,如表6所示。

#### 4 结论与讨论

本研究提出的食品知识图谱智能感知问答系

统基于大语言模型,旨在提升食品安全领域的信息检索和用户交互体验。该系统通过整合知识图谱与嵌入向量检索技术,能够有效应对开放性和长尾问题,提供个性化的营养建议及合规咨询,从而克服了传统问答系统在复杂问题处理上的局限性。研究表明,该系统在信息召回率和回答准确性方面显著优于现有方案,特别是在用户提问场景中表现出色,系统整体架构如图5所示。

虽然本方法在多方面取得显著进展,但是在处理某些专业术语时仍存在一定局限。未来的研究可以着重扩充数据源,以涵盖更广泛、深入的食

表6 大语言模型幻觉问题

Table 6 Illusions of large language models

大语言模型幻觉问题	知识图谱感知问答系统
用户:如果我在饭店吃了不干净的食物,出现了卫生问题,我该怎么办?	用户:如果我在饭店吃了不干净的食物,出现了卫生问题,我该怎么办?
大模型:您可以联系饭店,并请求赔偿。具体赔偿金额可以根据情况而定。	系统:如果您在饭店出现了卫生问题,建议您立即采取以下措施:
建议您立即采取以下措施:	保留证据:拍摄相关食品的照片,并保存消费凭证,以便后续投诉时使用。
保留证据:拍摄相关食品的照片,并保存消费凭证,以便后续投诉时使用。	联系饭店:直接向饭店经理反映情况,说明您的遭遇并要求其采取措施。
联系饭店:直接向饭店经理反映情况,说明您的遭遇并要求其采取措施。	投诉到相关部门:如当地的食品药品监督管理局,进行正式投诉。您有权利要求饭店赔偿您的损失。
投诉到相关部门:如当地的食品药品监督管理局,进行正式投诉。您有权利要求饭店赔偿您的损失。	用户:我可以要求十倍的赔偿吗?
用户:我可以要求十倍的赔偿吗?	系统:是的,根据相关法律法规,您有权要求十倍的赔偿。确保您有充分的证据来支持您的索赔请求。
大模型:一般来说,赔偿金额会根据损失情况而定。	



图5 食品知识图谱智能感知问答系统的整体架构

Fig.5 Overall architecture of the knowledge graph-enabled intelligent QA system

品安全领域知识。如在线搜索引擎,以实现信息的实时更新并提升系统的动态响应能力。同时提升知识图谱构建的精度,以准确呈现领域特定的复杂关系,提升应对食品安全领域的专业需求的能力。

### 参 考 文 献

- [1] 闫悦, 郭晓然, 王铁君, 等. 问答系统研究综述[J]. 计算机系统应用, 2023, 32(8): 1-18.  
YAN Y, GUO X R, WANG T J, et al. Survey on question answering system research[J]. Computer Systems and Applications, 2023, 32(8): 1-18.
- [2] 陈海红. 基于知识图谱的问答系统研究[J]. 信息与电脑, 2024, 36(6): 104-107  
CHEN H H. Research on knowledge graph based question answering system[J]. Information and Computer, 2024, 36(6): 104-107.
- [3] 王迪, 李海生, 李勇, 等. 基于知识图谱的食品安全问答方法[J]. 计算机技术与发展, 2024, 34(6): 118-124.  
WANG D, LI H H, LI Y, et al. A Question answering method for food safety based on knowledge graph [J]. Computer Technology and Development, 2024, 34(6): 118-124.
- [4] 王珂, 杨芳, 姜杉. 光学字符识别综述[J]. 计算机应用研究, 2020, 37(S2): 22-24.  
WANG K, YANG F, JIANG S. Review of optical character recognition [J]. Application Research of Computers, 2020, 37(S2): 22-24.
- [5] GAO Y F, XIONG Y, GAO X Y, et al. Retrieval-augmented generation for large language models: A survey[DB/OL]. (2023-12-18) [2024-10-15]. <https://arxiv.org/abs/2312.10997>.
- [6] CHANG C H, KAYED M, GIRGIS M R, et al. A survey of web information extraction systems[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(10): 1411-1428.
- [7] JI S X, PAN S R, CAMBRIA E, et al. A survey on knowledge graphs: Representation, acquisition, and applications [J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(2): 494-514.
- [8] 袁刚, 郭爽, 唐琦, 等. 基于深度学习的食品安全风险知识图谱构建方法[J]. 质量安全与检验检测, 2023, 33(5): 60-65.
- [9] 翁永琳. 基于神经网络的食品安全知识图谱构建研究[D]. 武汉: 华中农业大学, 2022.  
WENG Y L. Studies on the food safety knowledge graph based on graph neural networks [D]. Wuhan: Huazhong Agricultural University, 2022.
- [10] 卜意磊, 庞文迪, 吴甜甜, 等. 面向食品监管领域的知识图谱构建研究[J]. 计算机技术与发展, 2023, 33(6): 202-207, 214.  
BU Y L, PANG W D, WU T T, et al. Research on knowledge graph construction for food supervision [J]. Computer Technology and Development, 2023, 33(6): 202-207, 214.
- [11] 管铮懿, 周淑娟, 雷永琪. 基于 Neo4j 的食品安全知识图谱构建研究[J]. 信息与电脑(理论版), 2021, 33(6): 45-48.  
GUAN Z Y, ZHOU S J, LEI Y Q. Research on the construction of food safety knowledge graph based on Neo4j[J]. Information and Computer (Theoretical Edition), 2021, 33(6): 45-48.
- [12] 周锴. 基于时序动态知识图谱的食品安全风险预测技术研究[D]. 北京: 北方工业大学, 2023.  
ZHOU K. Research on food safety risk prediction technology based on temporal knowledge graph [D]. Beijing: North China University of Technology, 2023.
- [13] CHEN H, ZHANG C W, LI J, et al. Kggen: A generative approach for incipient knowledge graph population[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(5): 2254-2267.
- [14] QIAO C, HU X. A neural knowledge graph evaluator: Combining structural and semantic evidence of knowledge graphs for predicting supportive knowledge in scientific QA [J]. Information Processing & Management, 2020, 57(6): 102309.
- [15] DU Y N, LI C X, GUO R, et al. Pp-ocr: A practical ultra lightweight ocr system[DB/OL]. (2020-9-21) [2024-10-15]. <https://arxiv.org/abs/2009.09941>.
- [16] 赵同明, 钱佳琛, 王翔, 等. 健康饮食知识图谱和问答系统构建研究[J]. 现代信息科技, 2024, 8(6): 7-10.  
ZHAO T M, QIAN J C, WANG X, et al. Research



- on the construction of healthy diet knowledge graph and Q & A system[J]. *Modern Information Technology*, 2024, 8(6): 7–10.
- [17] JI Z W, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[J]. *ACM Computing Surveys*, 2023, 55(12): 1–38.
- [18] XIAO S T, LIU Z, ZHANG P T, et al. C-pack: Packaged resources to advance general Chinese embedding[DB/OL]. (2023–09–14) [2024–10–18]. <https://arxiv.org/abs/2009.09941>.
- [19] WANG J, YI X, GUO R, et al. Milvus: A purpose-built vector data management system[C]. New York: Proceedings of the 2021 International Conference on Management of Data, 2021: 2614–2627.
- [20] SAHOO P, SINGH A K, SAHA S, et al. A systematic survey of prompt engineering in large language models: Techniques and applications[DB/OL]. (2024–02–05) [2024–10–18]. <https://arxiv.org/abs/2402.07927>.

## Food Knowledge Graph-based Intelligent Perception Question Answering

Song Xiaoning, Jiang Qiyi, Zhang Wenjie, Wu Xiaojun

(School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, Jiangsu)

**Abstract** Food safety is of vital importance to people's lives. With the improvement of living standards, public concern about food safety has also increased, and the urgent need for reliable and effective information response has arisen. Existing intelligent question-answering systems, such as those based on text vector libraries, have made some progress in knowledge reasoning, problem handling, and semantic relationship recognition, but still have some shortcomings; question-answering systems based on knowledge graphs face inherent challenges in terms of graph construction cost and information recall rate; while question-answering systems based on template matching are poor in question generalization ability and context understanding. This paper proposes an innovative knowledge graph-based intelligent perception question-answering system based on large language models (LLM). Using large language models to improve the shortcomings of the template matching method, combining text vector libraries and knowledge graphs at the same time, and introducing high-dimensional vector semantic search technology into the knowledge graph retrieval process to improve the knowledge graph retrieval recall rate. This innovative approach significantly improves the overall performance of the question-answering system and the user's experience.

**Keywords** artificial intelligence; food safety; knowledge graph; large language models; intelligent question answering