基于ChatGPT的中国南海贝类知识智能服务

张蒋良[1, 2]，蒲秋梅[1, 2]，罗训[3]，李达[4]
（1. 中央民族大学信息工程学院，北京 100081；2. 民族语言智能分析与安全治理教育部重点实验室，北京 100081；
3. 天津理工大学计算机科学与工程学院，天津 300384；4. 国家计算机网络应急技术处理协调中心，北京 100029）

摘 要：南海地区的贝类资源十分丰富，但目前关于这些资源的信息分散在各种书籍和网站中。利用自行训练的深度学习模型进行知识关系的自动抽取可以减少人工整合信息的烦琐工作，但这一过程往往需要大量的数据标注和专家评估，且文本抽取效果的泛化性欠佳。利用 ChatGPT 构建中国南海贝类的知识服务体系，为上述问题提供了解决方案。通过将 ChatGPT 与贝类知识提问模板相结合，可以减少对数据集的依赖抽取文本，且效果较为理想；在此基础上，完成了贝类知识图谱的构建及可视化，并利用哈尔滨工业大学语言技术平台（LTP, language technology platform）4.0 技术开发了智能问答系统。该应用体系也为人工智能大模型在其他信息搜集和处理方面的应用提供了思路。

关键词：中国南海贝类；知识图谱；ChatGPT；智能问答
中图分类号：G203
文献标志码：A
doi：10.11959/j.issn.2096−3750.2025.00421

Research on intelligent services for knowledge about the South China Sea shellfish based on ChatGPT

ZHANG Jiangliang[1, 2], PU Qiumei[1, 2], LUO Xun[3], LI Da[4]
1. School of Information Engineering, Minzu University of China, Beijing 100081, China
2. Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance, Ministry of Education, Beijing 100081, China
3. School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China
4. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China

Abstract: The South China Sea region is home to abundant shellfish resources, yet the information about these resources is currently dispersed across various books and websites. To ease the manual integration of this information, knowledge extraction processes typically rely on self-trained deep learning models. However, these models often require extensive data labeling and expert evaluation, and their text extraction performance tends to have limited generalization. A solution by utilizing ChatGPT was proposed to build a knowledge service system for shellfish in the South China Sea region. By integrating ChatGPT with query templates for shellfish knowledge, the reliance on labeled datasets for text extraction was significantly reduced, yielding favorable results. On this basis, the construction and visualization of a shellfish knowledge graph was successfully completed, along with the development of an intelligent question-answering system using Harbin Institute of Technology's language technology platform (LTP) 4.0. This research provides valuable insights into the application of large-scale artificial intelligence models for information collection and processing.

Key words: the South China Sea shellfish, knowledge graph, ChatGPT, intelligent question answering

收稿日期：2024−04−07；**修回日期：**2024−09−13
通信作者：蒲秋梅，puqiumei@muc.edu.cn
基金项目：国家社会科学基金资助项目（No. 20BGL251）
Foundation Item: The National Social Science Fund of China (No. 20BGL251)

0　引言

除节肢动物门，软体动物是动物界中数量非常庞大的一类动物，其数量在海洋中居首。其中，许多种类的贝类与人们的生活息息相关[1]。罗训[2]及相关研究者大力推进南海生物数字孪生服务工作，并成立了智网互联实验室，已联合多家中国科研机构和哥伦比亚高校合作开展海洋生物多样性的保护工作。

目前，南海地区贝类的信息分布较为分散，其中部分知识描述存在碎片化和冗余的情况，不利于专业人员和用户系统清晰地了解相关知识。而且，当前大多文献的知识规格化和相关整理工作烦琐，对文本信息的自动抽取往往通过自行训练深度学习模型来实现。有研究基于双向编码器表征法（BERT, bidirectional encoder representations from transformers）、双向长短期记忆网络（BiLSTM, bidirectional long short-term memory）以及相关的网络结构模型实现了实体和关系的自动抽取分类，并取得了一定的效果[3]，然而这些深度学习文本提取方法需要大量的数据标注，在与数据集相关的文献来源中提取实体关系效果较为出色，但在很多文本描述风格不同的文献中，文本理解和抽取能力降低，泛化性不足导致结果欠佳。

ChatGPT是在GPT3的基础上经过微调而来的，大模型支持零样本学习且泛化性强[4]。微调过程引入了人类反馈强化学习（RLHF, reinforcement learningfrom human feedback）技术，通过将人类日常对话的语言习惯嵌入模型，并引入人类的价值偏好，使得模型输出与人类意图对齐。微调过程包括预训练、监督微调、设计奖励模型和反馈优化4个步骤[5]。由于ChatGPT的功能强大且具有良好的交互效果，社会各个领域都在积极探索其应用，将其出色的对话生成能力融入各种应用场景中[6]。

本研究利用ChatGPT大模型建立南海贝类知识图谱相关的知识服务体系。体系由4部分构成：

1）使用ChatGPT大模型自动抽取实体关系、建立南海贝类知识图谱相关的知识服务体系、通过图谱可视化展示关系以及利用智能问答系统提供贝类知识；

2）本文利用ChatGPT大模型的问答来实现搜集知识，并自动抽取文本中的实体和关系及规格化，在此过程中减少了对文本标注的依赖，提高了搜集信息的效率；

3）本文采用Neo4j图数据库来存储图谱关系，涵盖生态、地理、生物等多个维度的信息，实现了知识的规格化存储；

4）通过图谱的可视化展示，用户能够快速定位所需贝类信息，使知识分类更加清晰明了。最后，基于贝类图谱文本，本文利用哈尔滨工业大学分词技术将用户提问语句进行分词[7]，并与预设的问法模板进行匹配，实现了有关贝类知识的智能问答功能。

南海贝类知识服务体系不仅整理了中国南海贝类信息，也为其他领域的信息研究提供了一个基于大模型的应用实例。首先，本文利用ChatGPT搜集数据，克服了传统深度学习模型抽取文本需要进行大量数据标注的问题，相较于自行训练大模型，它具有更好的泛化性能，且ChatGPT能够及时搜集当前的知识并进行相关的文本规格化，以达到实时更新数据的效果。其次，本文能够利用知识图谱将分散烦琐的贝类知识进行分门别类的展示。相对于传统的文字知识库，知识图谱能够更清晰地呈现贝类知识的结构和关联性，使用户能够更加直观地了解贝类之间的关系和特点，提高了知识获取的效率。最后，结合知识图谱的智能问答系统，能够比传统搜索引擎更准确地满足用户的需求，结合知识图谱的推理和推荐功能，可以提供更加个性化、精准的答案，帮助用户更快地获取所需信息，有助于快速直观地掌握贝类知识。

1　相关研究

1.1　知识图谱构建与应用

知识图谱以其结构化的语义网络知识库，清晰展示了领域知识的概念和关联，被广泛应用于多领域[8]，涵盖智能搜索、自然语言处理、数据挖掘、金融投资、医疗健康等方面。知识图谱将专业领域信息整合为图结构化的实体关系[9]，在此基础上对图谱信息进行推理，高效解决了各个领域中的专家系统问题[10]。南海贝类知识图谱服务系统的构建旨在将贝类信息分门别类地整理，并在此基础上提供高效的智能问答及相关的知识服务。

1.2　文本关系抽取

早期的实体关系抽取，主要依赖传统的机器学习方法，其中Word2Vec在语言文本处理任务中取

得了良好的效果[11]。文献[12]在海洋中药知识多模态知识图（MMKG, multi-modal knowledge graph）基础上，运用Aho-Corasick、Word2Vec等传统机器学习算法，实现了细粒度海洋中药问题分类方法。但机器学习在处理一些上下文联系较强的语境任务时效果欠佳。随着深度学习技术兴起，大量研究以长短期记忆网络、循环神经网络、Transformer神经网络为基础，调整网络结构进行训练和应用，因此其结合上下文的特征提取文本能力大幅提高[13]。在渔业领域中，融合BERT与BiLSTM的实体识别模型[14]，成功提升了水产动物疾病命名实体识别准确率。但这些方法通常需要大量的数据标注及专家审核，而且在语境变化时仍面临识别分类准确率降低的问题。

大规模预训练语言模型的出现，使得自然语言处理的性能得到极大提升。在实体关系抽取任务中，这些大模型能够捕捉更丰富的语境信息，具有更好的泛化性能和准确率[15]。而且，基于大模型的微调或训练通常支持零样本学习。本次研究通过调用应用程序接口（API, application program interface）和ChatGPT问答的方式，搜集贝类文本并进行实体关系抽取，达到了获取结构化贝类数据的效果。

1.3　基于知识图谱的问答系统

在问答系统的发展历程中，涌现了多种类型的系统，分别在问题分析、信息检索、答案生成等关键技术上不断更新[16]。2010年后，随着知识图谱的发展，问答系统逐渐利用结构化数据进行更深入的问题回答，这一时期标志着问答系统逐渐从简单的问答模式发展为更加复杂、结构化的知识图谱时代[17]。

2022年后，基于大规模语言模型的相关应用ChatGPT出现，带来了更强大的自然语言处理能力[18]。本文尝试基于ChatGPT的API来实现问答系统，以期通过结合模型的强大语言理解能力实现对复杂问题的准确回答。但由于ChatGPT的回答具有不确定性，直接返回查询贝类知识数据库的语言结果准确率较低。在本文的贝类问答系统中，采用基于哈尔滨工业大学语言技术平台（LTP, language technology platform）4.0处理提问语句和提问模板进行匹配，从而实现对贝类图谱的知识查询和问答。

1.4　领域知识服务体系构建

近年来，随着深度学习技术的发展，越来越多的研究人员致力于将领域知识与文本数据相结合，通过自动化抽取构建知识图谱，从而更加清晰地梳理和呈现特定领域的知识脉络，以推动领域知识智能服务体系的构建。

文献[19]通过将领域特定的知识嵌入BiLSTM-CRF框架的词，引入矩阵初始化过程中，显著提高了命名实体识别的准确性。同时，他们提出了一种基于知识增强的文档级实体和关系抽取框架，用于工业领域的知识图谱构建。类似地，文献[20]提出了一种结合BERT与层次交叉注意力机制的知识图谱问答模型，用于桥梁领域的知识三元组关系提取，并基于此模型实现了智能问答服务。在地质领域，一些研究者定义了实体类型和概念间的关系，构建了地质语料库，并基于BERT预训练模型实现了命名实体识别和关系抽取，完成了地质知识图谱的三元组关系提取[21]。此外，文献[22]在网络安全领域提出了一个关键基础设施保护知识图谱，结合了BiLSTM-CRF和预训练的BERT模型进行命名实体识别，并通过KG-BERT模型进行关系预测，确保了知识图谱的动态更新和有效性。文献[23]构建了一个基于知识图谱的中文矿物问答系统，利用BERT模型实现了高效、准确的问答功能。而文献[24]提出了BERT-MCNN模型，专门用于从中国海事污染防治相关法规中提取多种关系和命名实体，从而有效构建海事法规知识图谱，并在Neo4j图数据库中实现了应用。最后，文献[25]提出的K-RET生物医学提取系统，通过引入改进的BERT模型，提高了生物医学关系提取的准确性，并支持更具解释性的生物医学关联预测。然而，这些领域知识图谱服务体系仍面临诸如数据标注工作量巨大、更新速度缓慢以及模型训练的泛化能力有限等挑战。因此，本研究旨在利用ChatGPT的高泛化能力和快速处理效率，避免繁重的文本标注工作，从而提升领域知识服务的智能化水平。

本文基于ChatGPT成功构建了南海贝类知识服务体系，通过大规模的语料库进行预训练，拥有更强大的语言理解和表示学习能力[26]。在实体关系抽取任务中，这些大模型能够捕捉更丰富的语境信息，具有更好的泛化性能和准确率。此外，基于大模型的微调或训练通常支持零样本学习，即在未见过的实体关系上也能表现出色[27]。这种进展为信息情报整理人员提供了直接适用于各领域抽取文本的方式，减少了对文本标注的依赖，提高了搜集信息

的效率。相较于贝类知识分布分散、部分内容碎片化和冗余，知识图谱更能够清晰地呈现贝类知识的结构和关联性，使用户能够更加直观地了解贝类之间的关系和特点。

2　研究框架

本文构建了一个基于 ChatGPT 的中国南海贝类知识图谱的智能服务框架，该框架主要包括本体建模、知识抽取、知识存储和知识服务。图 1 展示了中国南海贝类知识智能服务流程，首先进行本体建模，确立贝类实体的分类及其关系，然后基于 ChatGPT 进行文本关系提取，获取贝类知识的三元组关系，从而完成知识图谱的构建。此外，模型实现了基于贝类图谱的知识关系可视化和智能问答。通过这些方法和步骤，本文完成了中国南海贝类知识服务系统的构建。

2.1　本体建模

本体建模有助于梳理贝类领域中的基本概念，如不同分类层次的分类单位（门、纲、目、科、种），以及贝类的地域分布、形态特征等。贝类图谱的本体建模考虑了海洋生物领域专家的意见，明晰了相关概念和关系，以避免引发歧义的类和关系，进而更好地将贝类知识分门别类。模型在初始阶段通过爬虫从中国南海渔业网站和浙江大学鱼类数据集获取相关贝类文本，并将其规范化为指定类型的 CSV 格式文件。这些数据包含了贝类的种属关系、形态特征、分布、生活习性、经济价值和保护措施等重要信息，以这些贝类相关数据为基础，构建了初始的贝类图谱。后续该模型基于本体建模，从文本中抽取规格化贝类实体和关系构建知识图谱，清晰直观展示贝类之间的分类演化关系、地理分布规律和生态关联，为贝类生物学、生态学和保护学等领域的研究提供深入理解和分析的基础。

2.2　知识抽取

本文根据专业人员确立好的贝类领域知识的本体关系，利用 ChatGPT 进行知识抽取，形成规范的三元组关系。同时，该方式结合适当的提示文本，调用相关的 API 来进行贝类知识的问答，无须进行数据标注或监督学习，提高了知识抽取的效率和准确率。

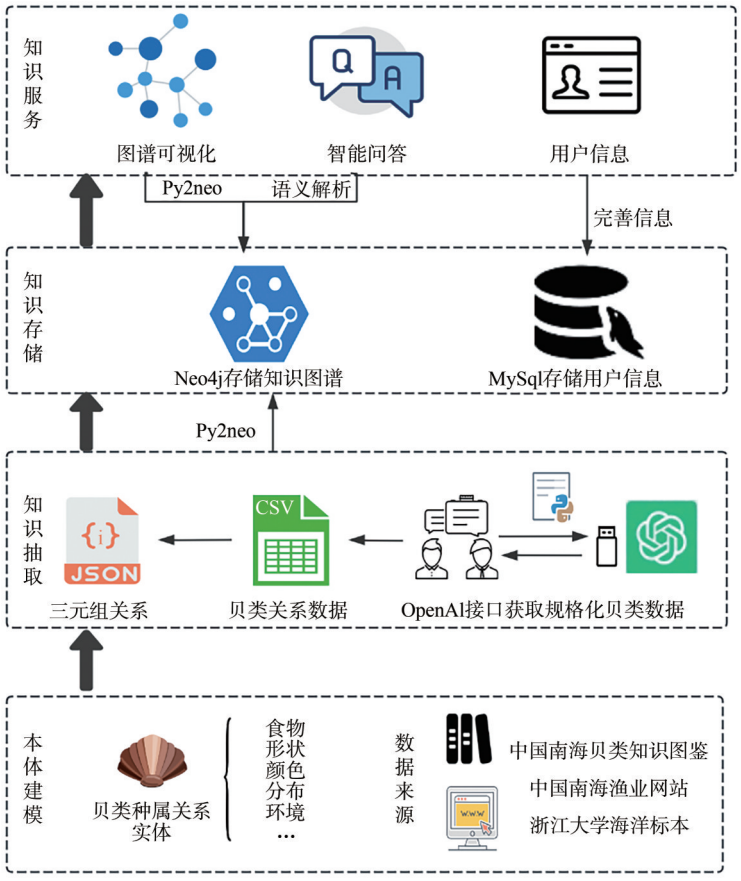


图 1　中国南海贝类知识智能服务流程

通过大量实验，模型确立了获取结构化贝类文本的提问模板，这些提问模板可以针对贝类领域的特点，引导 ChatGPT 生成符合本体建模定义的问题，从而有针对性地提取贝类文本中的实体、属性和关系。如针对不同的贝类特征，可以设计有针对性的提问模板——“鹦鹉螺的形状是什么？返回简短的词汇”“鹦鹉螺的壳色是什么?”。通过这一方案，本文对贝类文本中实体的识别和关系分类达到了较为理想的效果。ChatGPT 结合适当的提示文本，能够高效地进行知识抽取，将原始文本中的信息转化为结构化的知识，之后模型将三元组关系导入知识图谱，形成分门别类清晰的知识网络，从而基于贝类知识图谱进行可视化知识展示和智能问答。

2.3　知识存储

本文通过知识抽取得到了规格化的贝类三元组关系，如（平鲍，属于，鲍科）（牡蛎，分布，北太平洋）等，将三元组中的实体和关系创建导入图数据库完成知识存储。系统设计选择 Neo4j 作为知识图谱存储工具，Neo4j 可提供强大的查询语言（Cypher）检索知识，实现了对图类知识的高效组织和存储、信息的灵活检索与处理[28]。模型通过调用 Py2neo 的开发库，与 Neo4j 数据库连接可实现对贝类图关系数据的存储与查询。

本文设计了一个基于贝类知识图谱的网站系统，集成了知识图谱可视化和智能问答功能。为了更好地呈现贝类之间的复杂关系和特征，系统可视化模块引入了 ECharts 库中的力导向图进行前端渲染，展示贝类之间的分类层次、形态特征、生活习性等多方面的信息，以便明晰贝类之间的复杂关系。

2.4　知识服务

本文在智能问答系统的实现过程中，使用了 3 种不同方式将用户提问处理为贝类图谱查询语句。首先，尝试将用户提问通过人工预定的词典和模板匹配的方式实现。然而，这种方式的覆盖问答范围有限，且随着贝类图谱数据扩充，系统需要不断录入新增的贝类词汇和问法模板。其次，尝试利用 ChatGPT 将用户对贝类知识的提问转换为 Cypher 语句，但本次实验中调用开源大模型在理解和生成指定数据库的查询语句准确率较低。

最终，智能问答系统采用基于哈尔滨工业大学 LTP 4.0 的分词标注识别模型。模型通过分词和模式匹配的方式，将关于贝类关系提问的自然语言转换为 Cypher 语言查询图谱，避免了需要更新和维护庞大的南海贝类关系词典。

3　实例论证

3.1　贝类本体建模

南海贝类知识图谱的本体建模确立后，系统通过爬虫爬取中国南海渔业网站和浙江大学鱼类数据集，将贝类的实体和关系导入图数据库，完成了南海贝类知识图谱的初始化构建。表 1~表 4 分别展示了所搜集到的贝类知识图谱本体概念、贝类知识图谱本体属性、贝类实体的部分实例、贝类知识图谱本体关系。

表 1　　贝类知识图谱本体概念

本体概念名称	描述
门	贝类所属的门
纲	贝类的拉丁名
目	贝类所属的目
科	贝类所属的科
种	贝类的中文名
地域	贝类分布的海域
拉丁名	贝类的拉丁名

表 2　　贝类知识图谱本体属性

本体属性名称	描述
形状	贝类的形状，如对称
壳色	贝类的壳的颜色，如黄色
食性	贝类的摄食属性，如草食性
保护方法	贝类保护自己的方法，如毒液
生态系统	贝类的生活环境，如珊瑚礁
牙齿	贝类是否具备牙齿，分布的部位
足部	贝类是否具有足部，多少对

表 3　　贝类实体的部分实例

中文名	拉丁名	门	纲	目	科	地区
小刀蛏	Cultellus attenuatus	软体动物 Mollusca	双壳纲 Bivalvia	贫齿蛤目 Adapedonta	刀蛏科 Pharidae	南海
短蛸	Amphioctopus fangsiao	软体动物 Mollusca	头足纲 Cephalopoda	章鱼目 Octopoda	章鱼科 Octopodidae	中国沿海
绶贝	Mauritia mauritiana	软体动物 Mollusca	腹足纲 Gastropoda	玉黍螺目 Littorini morpha	宝贝科 Cypraeidae	西沙群岛

表 4　贝类知识图谱本体关系

本体关系名称	描述
近亲	贝类之间的关系，如平鲍和多变鲍是近亲
属于	分类系统种的关系，如多变鲍属于鲍科
分布	贝类与区域之间的关系，如平鲍分布在南海
拥有	地域与生态系统之间的关系，如南海拥有珊瑚礁生态系统
摄食	贝类与食性之间的关系
保护	贝类与保护方法之间的关系

本文通过构建南海贝类本体属性的三元组关系，揭示了贝类的种属关系、地理分布、颜色和形状等关键属性，这一本体的构建为知识图谱系统设计奠定了基础。

3.2　贝类知识抽取

ChatGPT 能够感知上下文信息，表现出更高的泛化性能和准确率。由于标注工作量大且传统深度学习模型训练结果泛化性不足，本文研究确立了基于 ChatGPT 提取相关贝类文本的提问模板。

模型利用 ChatGPT 这一预训练模型的强大能力，自动化地从自然语言描述中提取出结构化的贝类知识，为知识图谱的扩展提供支持。模型首先从官方科普网站获取了部分贝类数据，包含 186 个实体节点和 562 种实体关系，构建了初始的知识图谱，之后通过遍历初始图谱中的数据节点，与 ChatGPT 进行问答，获取相关本体关系的延展。初始数据图谱的部分实体关系如图 2 所示。

图 3 展示了基于 ChatGPT 搜集数据并进行规格化的流程，涵盖了模型选择、参数设置、数据问答及结果返回等关键步骤。通过 API 调用当前最先进的 GPT-4 版本，并使用 API 密钥进行通信，确保了请求和响应的安全性与准确性。此外，GPT-4 与互联网相连，保证了所获取信息的时效性和真实性，

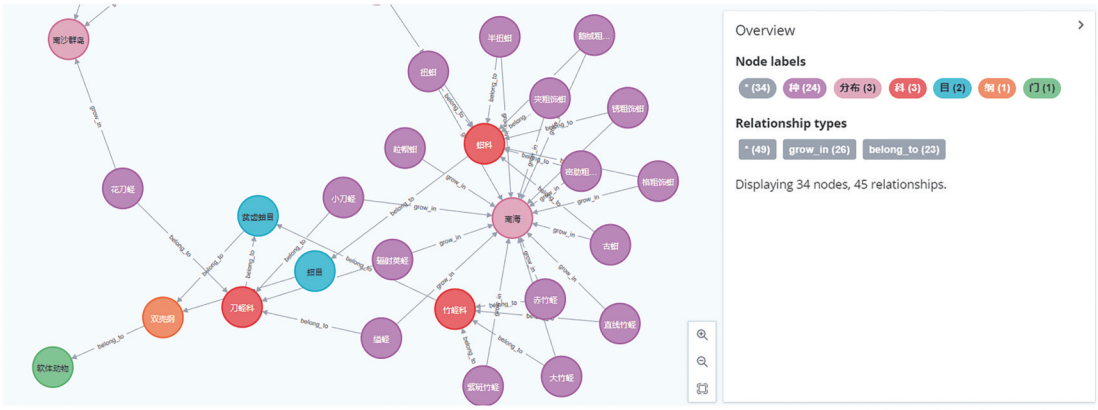


图 2　初始图数据库的部分实体关系

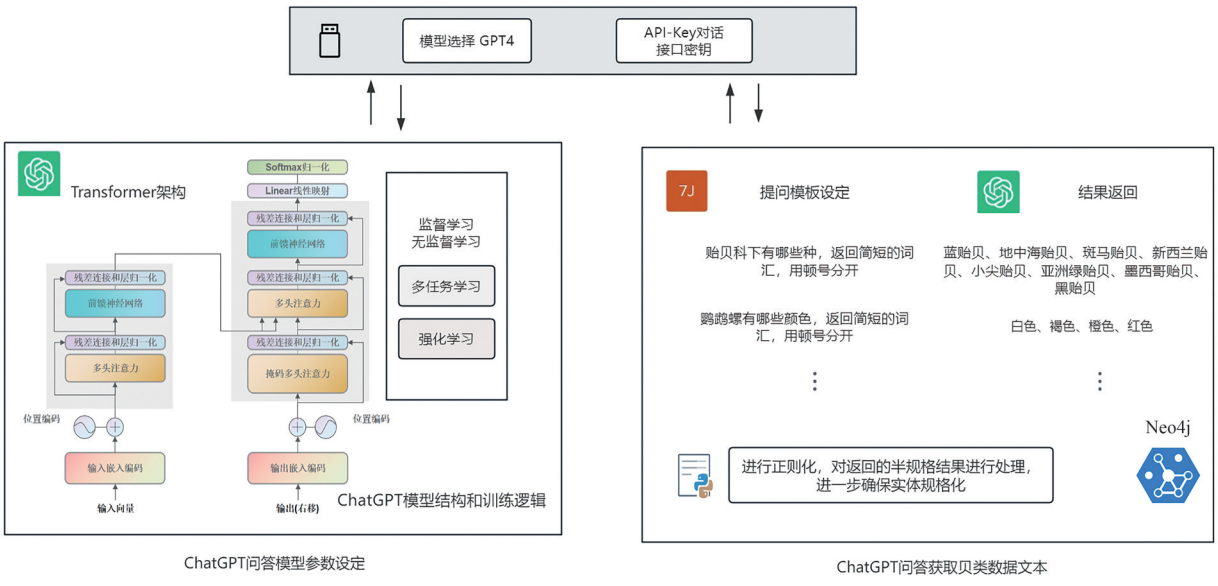


图 3　基于 ChatGPT 搜集数据并进行规格化的流程

而不仅是语法上的正确性。ChatGPT 的核心架构基于 Transformer 的多头注意力机制[4]，这使得模型能够在处理复杂的自然语言任务时，捕捉上下文中的细微差别。作为一种大语言模型，ChatGPT 经过大规模数据的训练，具有较强的语言泛化能力。模型还经过了多任务学习、监督学习、无监督学习以及强化学习的微调，展现出对各种语境的强大理解能力。

本研究采用了问答的方式，通过自动化的 Python 脚本与 ChatGPT 进行交互，遍历现有的贝类实体，自动扩展其三元组关系。为了更好地提取和扩充贝类知识，本文设计了多种问题模板，这些模板涵盖了贝类的分类、形态特征、颜色描述等信息，如对于贝类颜色信息的获取，设计了类似“某些贝类的颜色有哪些？请返回简短的词汇”这样的模板。

ChatGPT 根据所设定的模板返回结果，通常包括贝类的具体名称和属性信息。为了进一步处理这些返回的数据，本文使用正则化方法对结果进行过滤和格式化，使其更符合预期的结构要求。针对贝类颜色的描述，可能返回标准化的颜色词汇如“浅黄”“淡红”“白色”等，以三元组的形式构建贝类的知识关系（如{鹦鹉螺，颜色，淡红}）。处理后的贝类数据会被传递至 Neo4j 等知识图谱系统，用于构建和扩展贝类知识库。这一步骤确保了数据的系统化和结构化存储，从而支持后续在扩充知识图谱的过程中知识查询和分析。

以贝类颜色信息的获取为例，基于 ChatGPT 将贝类知识抽取进而扩充图谱实体关系的步骤如下。

1) 导入依赖包

导入必要的 Python 环境语言依赖包，其中包括访问 ChatGPT 接口的 OpenAI 包，用于正则化处理字符串的 re 包，以及与 Neo4j 数据库交互进行增删改查的 Py2neo 包。

2) ChatGPT 和图数据库参数配置

进行 OpenAI 密钥的设置，以确保能够向 ChatGPT 发出提问并获取相应的结果。同时，设计 Python 脚本建立与图数据库的连接，如图 4 所示。

3) ChatGPT 问答获取贝类结构化数据

本文通过调用 openai.Completion.create 函数向 ChatGPT 提出关于贝类颜色的问题，并得到返回的字符串文本，获取 ChatGPT 生成的半结构化的颜色信息，如图 5 所示。得益于 ChatGPT 的预训练模型，它能够基于海量语料库中的知识，从中提取出贝类颜色的相关特征信息，甚至在现有知识基础上进行合理的知识推测。

```
# Set up the OpenAI API client
openai.api_key = "sk-G7LAXsXBzVdB9axXbYnNT3BlbkFJsUgl

# Set up the model and prompt
model_engine = "text-davinci-003"

# 连接neo4j数据库，输入地址、用户名、密码
graph = Graph("http://localhost:7474", auth=("neo4j",
matcher=NodeMatcher(graph)
```

图 4　设计 Python 脚本建立与图数据库的连接

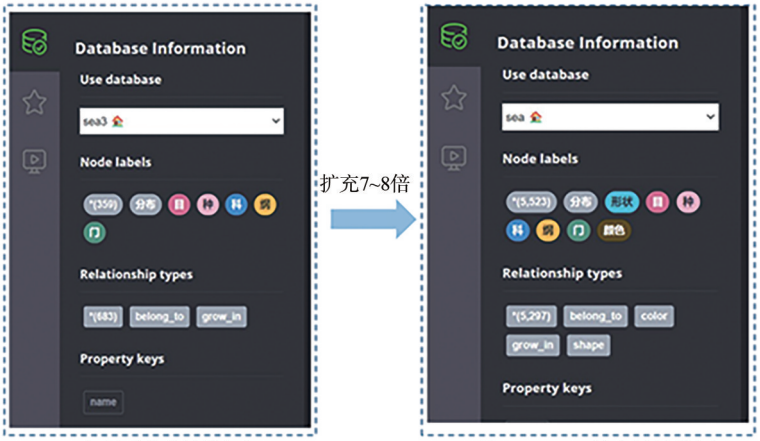
```
str1="的颜色有哪些？用简短的词汇描述"
for i in range(0, len(invoice_data)):
    node1.append(invoice_data['贝名'][i])

for i in range(0,len(invoice_data)):
    prompt=node1[i]+str1
    completion = openai.Completion.create(
        engine=model_engine,
        prompt=prompt,
        max_tokens=1024,
        n=1,
        stop=None,
        temperature=0.5,
    )
    words = completion.choices[0].text.strip()
```

图 5　获取 ChatGPT 生成的半结构化的颜色信息

4) 正则化规格贝类文本

模型对 ChatGPT 返回的贝类关系字符串进行正则化处理，去除多余的标点符号和一些助词如“了、的、等、和”等，从而提取出结构化的贝类文本关系。在此，ChatGPT 基于其在大规模语料库上的训练，能够生成高度相关且合理的回答，这些回答通过正则化处理后即可直接用于知识图谱的扩展。之后，模型与 Neo4j 数据库进行交互，将扩展的贝类实体和关系导入图数据库。

以上示例展示了通过 ChatGPT 自动扩充贝类知识图谱的流程。本文使用 Neo4j 图数据库来进行贝类知识的存储。经过知识抽取步骤，利用 Python 语言的第三方包 Py2neo 调用 Cypher 语句，将贝类知识的三元组数据存储到 Neo4j 图数据库中。最终，本文成功构建了由 5 523 个节点和 5 297 条边组成的贝类知识图谱。



(a) 初始采集的贝类知识
（实体数：359，关系数：683）

(b) 基于ChatGPT扩充的贝类知识
（实体数：5 523，关系数：5 297）

图 6　ChatGPT 扩充效果

扩充后的数据集具有超过 5 000 种的贝类实体和关系。ChatGPT 扩充效果如图 6 所示。

在实验中，向 ChatGPT 提问的部分尝试类型见表 5。如“鹦鹉螺有哪些颜色，请返回简短的词汇，并以顿号进行分割”，之后将 ChatGPT 返回的结果进行正则化处理，获得结构化的贝类实体和关系数据。

表 5　向 ChatGPT 提问的部分尝试类型

提问类型	提问语句 prompt
类型 1	鹦鹉螺的颜色有哪些
类型 2	鹦鹉螺有哪些颜色，请以中文形式回答
类型 3	鹦鹉螺有哪些颜色，请进行简短的描述
类型 4	鹦鹉螺有哪些颜色，请以列表的形式返回
类型 5	鹦鹉螺有哪些颜色，以 Python 的 list 形式返回
类型 6	鹦鹉螺有哪些颜色，以 JSON 文件形式返回
类型 7	鹦鹉螺有哪些颜色，以列表或字符数组形式返回

在实验过程中，有些 ChatGPT 的回答仍为非结构化数据，不能直接提取实体和关系。通过多次实验，本文确定了一种合适的提问机制，不断遍历图谱中贝类实体向 ChatGPT 提问，返回更多贝类相关知识并进行有关语句的正则化处理，向 ChatGPT 提问的部分标准方式见表 6。最终，系统通过 ChatGPT 成功扩充了南海贝类知识图谱，获得三元组实例结果如（平鲍，属于，鲍科）、（牡蛎，分布，北太平洋）等。

3.3　贝类知识存储

本文使用 Neo4j 图数据库进行知识的存储，Neo4j 作为一种图形式的存储模式，与传统的基于关系型数据库相比具有更好的知识关联、知识查询与知识推理的能力。Neo4j 数据库由标签、节点、关系及节点属性 4 类要素组成，其中类与标签、实例与节点、对象属性与关系、数据属性与节点属性一一对应，由此就完成了贝类知识的本体模式层到图谱数据层的匹配映射。如实体关系：（青螺，属于，鹦鹉螺科），标签分别为种与科，实例分别为“青螺”与“鹦鹉螺科”，关系为“属于”。

表 6　向 ChatGPT 提问的部分标准方式

实体问法	扩充的实体	扩充的关系
科下有哪些种，返回简短的词汇	种	种 属于 科
目下有哪些科，返回简短的词汇	科	科 属于 目
纲下有哪些目，返回简短的词汇	目	目 属于 纲
该贝有哪些颜色，返回简短的词汇	颜色	种 呈现 颜色
该贝有哪些形状，返回简短的词汇	形状	种 呈现 形状

本文采用 Neo4j 桌面版，图数据库版本 4.4，JDK 的依赖版本为 jdk-11.0。经过知识融合步骤完成了三元组数据的整理后，利用 Python 语言的第三方包 Py2neo 调用 Cypher 语句来将三元组数据存储到图数据库 Neo4j 中。

3.4　贝类知识服务

在传统搜索引擎中，贝类知识分布较为分散，用户获取的贝类文本通常较为冗长，使得对贝类关系的理解较为抽象和孤立。本文旨在改善此现状，基于贝类知识图谱设计了问答模块，并展示与之相关的贝类知识关系。贝类知识图谱搜索模块结构化展示了贝类知识，通过该系统可以清晰查阅贝类知识。本文设计的中国南海贝类知识服务系统框架如图 7 所示。

本系统进行了贝类图知识结构的可视化展示，设计了两种交互方式，帮助用户更直观地理解和探

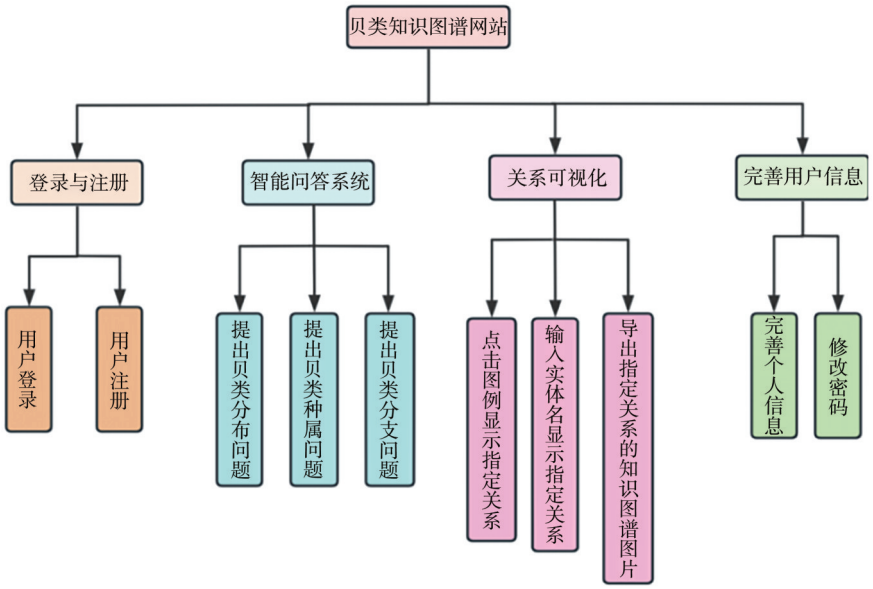


图 7　中国南海贝类知识服务系统框架

索复杂的贝类关系。首先，通过文本框输入搜索功能，用户可以轻松查找特定的贝类信息，系统会自动定位并显示相关的实体和关系。其次，系统还提供了点击图例的功能，用户可以直接在知识图谱的可视化界面上点击感兴趣的节点或边，动态展现贝类的分类层次、生态关系和地理分布等信息。这种可视化展示不仅让用户更直观地理解贝类之间的复杂关系，还能清晰地呈现贝类知识的整体结构和关联性，知识图谱搜索界面如图 8 所示，点击实体搜索界面如图 9 所示。

在智能问答系统中，本研究针对用户的自然语言提问设计了高效的处理流程，以确保能够准确地查询并返回相关的贝类知识。首先，系统使用哈尔滨工业大学的 LTP 4.0 中文分词模型，对用户输入的自然语言进行分词处理。这一步骤将输入的完整句子拆分为更易于处理的词语或短语，为后续的模板匹配奠定基础。系统将分词后的语句与预先设定的提问模板进行匹配，通过这种匹配机制，系统能够将自然语言提问自动转换为 Cypher 查询语言，从而执行对贝类知识的查询。此过程涵盖了用户对

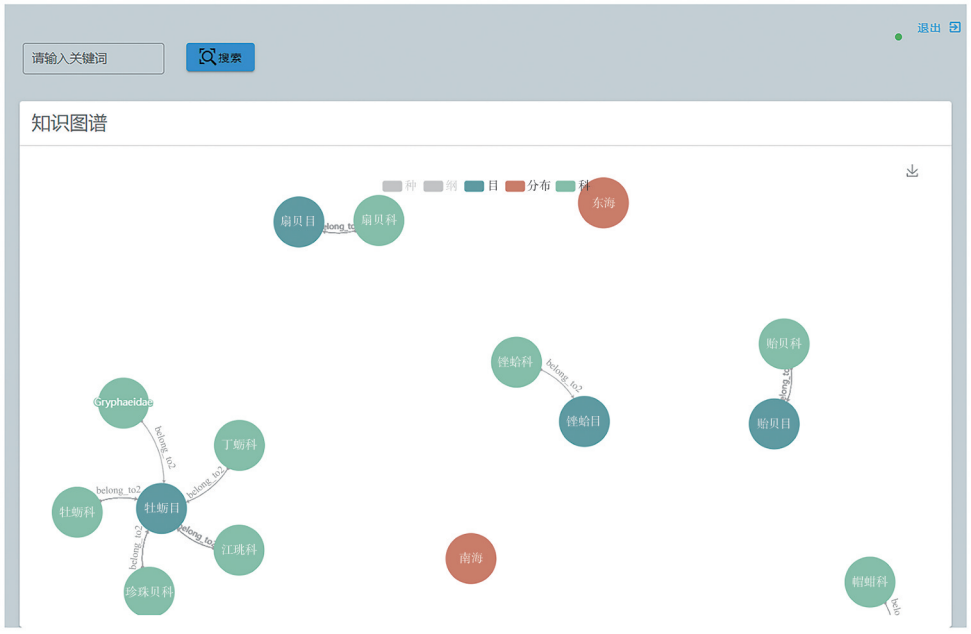
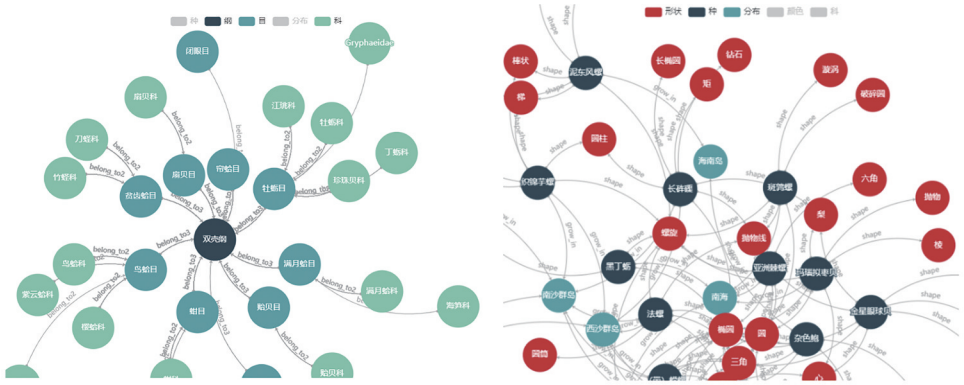


图 8　知识图谱搜索界面



(a) 可视化指定的科、纲、目关系　　(b) 可视化指定的形状、分布关系

图 9　点击实体搜索界面

图 10　智能问答模块界面

贝类的种属关系、地理分布、颜色特征等多种属性的查询，并为用户提供相应的答案。

如果系统在模板匹配过程中无法找到适合的匹配项，则会返回“对不起，这个问题我无法回答”的文本结果，确保用户能够及时了解系统的能力范围和限制。智能问答模块界面如图 10 所示。

4　结束语

本文开发了一个基于 ChatGPT 的中国南海贝类知识服务系统，展现了如何利用大语言模型在特定领域中进行高效知识抽取与整合。本研究不仅实现了从贝类文本中自动提取实体和关系，还通过知识图谱的构建与可视化，提供了一种便捷的方式来展示贝类之间复杂的生态与分类关系。与传统的深度学习算法相比，大模型表现出更高的准确性，减少了对标注数据集的依赖，同时也为其他领域的信息搜集和知识整理提供了一套模板方案。该系统还集成了智能问答功能，该功能可以根据知识图谱快速推断和定位所需的知识，解决了贝类知识文本分散分布和来自贝类科普网站和知识库的搜索结果冗长等问题。

在处理特定领域的文本时，尽管 ChatGPT 展现出强大的泛化能力，并且在大多数情况下实现了良好的文本提取性能，但仍然面临着不准确的实体识别和在特定专业领域提取关系时的偏差等挑战。因此，未来将考虑使用大型语言模型，适当地标注与贝类相关的文本数据，并通过提示模板微调模型[29]，以进一步提高在贝类领域文本提取方面的准确率。

参考文献：

[1] 杨文. 中国南海经济贝类原色图谱[D]. 湛江：广东海洋大学, 2014.
YANG W. Color atlas of economic mollusks in the South China Sea[D]. Zhanjiang：Guangdong Ocean University, 2014.

[2] 罗训. 新一代信息技术之虚拟现实助力一带一路[C]//2021 国际产学研用合作会议（南昌）报告摘要选集. 北京: 中国计算机学会(CCF), 2021: 35.
LUO X. Virtual reality empowering the belt and road initiative

with new generation information technology[C]//Proceedings of the 2021 International Conference on Industry-Academia-Research-Application Cooperation (Nanchang). Beijing: China Computer Federation (CCF), 2021: 35.

[3] DEVLIN J. Bert: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint, 2018, arXiv: 1810.04805.

[4] WU T Y, HE S Z, LIU J P, et al. A brief overview of ChatGPT: the history, status quo and potential future development[J]. IEEE/CAA Journal of Automatica Sinica, 2023, 10(5): 1122-1136.

[5] XU R X, LUO F L, ZHANG Z Y, et al. Raise a child in large language model: towards effective and generalizable fine-tuning[J]. arXiv preprint, 2021, arXiv: 2109.05687.

[6] 文森, 钱力, 胡懋地, 等. 基于大语言模型的问答技术研究进展综述[J]. 数据分析与知识发现, 2024, 8(6): 16-29.
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.

[7] CHE W X, FENG Y L, QIN L B, et al. N-LTP: an open-source neural language technology platform for Chinese[J]. arXiv preprint, 2020, arXiv: 2009.11616.

[8] 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, 2022, 33(2): 494-514.

[9] 张文豪, 徐贞顺, 刘纳, 等. 知识图谱补全方法研究综述[J]. 计算机工程与应用, 2024, 60(12): 61-73.
ZHANG W H, XU Z S, LIU N, et al. Overview of knowledge graph completion methods[J]. Computer Engineering and Applications, 2024, 60(12): 61-73.

[10] 蒋川宇, 韩翔宇, 杨文蕊, 等. 医学知识图谱研究与应用综述[J]. 计算机科学, 2023, 50(3): 83-93.
JIANG C Y, HAN X Y, YANG W R, et al. Survey of medical knowledge graph research and application[J]. Computer Science, 2023, 50(3): 83-93.

[11] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint, 2013, arXiv: 1301.3781.

[12] 洪海蓝, 李文林, 杨涛, 等. 基于知识图谱的海洋中药智能问答系统的设计与实现[J]. 世界科学技术—中医药现代化, 2023, 25(6): 1935-1941.
HONG H L, LI W L, YANG T, et al. Design and implementation of intelligent question answering system of marine traditional Chinese medicine based on knowledge graph[J]. Modernization of Traditional Chinese Medicine and Materia Medica-World Science and Technology, 2023, 25(6): 1935-1941.

[13] XU G X, MENG Y T, QIU X Y, et al. Sentiment analysis of comment texts based on BiLSTM[J]. IEEE Access, 2019, 7: 51522-51532.

[14] 刘巨升, 杨惠宁, 孙哲涛, 等. 面向知识图谱构建的水产动物疾病诊治命名实体识别[J]. 农业工程学报, 2022, 38(7): 210-217.
LIU J S, YANG H N, SUN Z T, et al. Named-entity recognition for the diagnosis and treatment of aquatic animal diseases using knowledge graph construction[J]. Transactions of the Chinese Society of Agricultural Engineering, 2022, 38(7): 210-217.

[15] 冯杨洋, 汪庆, 谢旻晖, 等. 从BERT到ChatGPT: 大模型训练中的存储系统挑战与技术发展[J]. 计算机研究与发展, 2024, 61(4): 809-823.
FENG Y Y, WANG Q, XIE M H, et al. From BERT to ChatGPT: challenges and technical development of storage systems for large model training[J]. Journal of Computer Research and Development, 2024, 61(4): 809-823.

[16] RAY S K, SHAALAN K. A review and future perspectives of Arabic question answering systems[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(12): 3169-3190.

[17] 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 & Applications, 2023, 32(8): 1-18.

[18] 王耀祖, 李擎, 戴张杰, 等. 大语言模型研究现状与趋势[J]. 工程科学学报, 2024, 46(8): 1411-1425.
WANG Y Z, LI Q, DAI Z J, et al. Current status and trends in large language modeling research[J]. Chinese Journal of Engineering, 2024, 46(8): 1411-1425.

[19] HAN Z L, WANG J. Knowledge enhanced graph inference network based entity-relation extraction and knowledge graph construction for industrial domain[J]. Frontiers of Engineering Management, 2024, 11(1): 143-158.

[20] YANG J X, YANG X X, LI R, et al. BERT and hierarchical cross attention-based question answering over bridge inspection knowledge graph[J]. Expert Systems with Applications, 2023, 233: 120896.

[21] MA K, TIAN M, TAN Y J, et al. Ontology-based BERT model for automated information extraction from geological hazard reports[J]. Journal of Earth Science, 2023, 34(5): 1390-1405.

[22] CHEN J R, LU Y Q, ZHANG Y, et al. A management knowledge graph approach for critical infrastructure protection: Ontology design, information extraction and relation prediction[J]. International Journal of Critical Infrastructure Protection, 2023, 43: 100634.

[23] LIU C J, JI X H, DONG Y H, et al. Chinese mineral question and answering system based on knowledge graph[J]. Expert Systems with Applications, 2023, 231: 120841.

[24] LIU C Y, ZHANG X Y, XU Y, et al. Knowledge graph for maritime pollution regulations based on deep learning methods[J]. Ocean & Coastal Management, 2023, 242: 106679.

[25] SOUSA D F, COUTO F M. K-RET: knowledgeable biomedical relation extraction system[J]. Bioinformatics, 2023, 39(4): btad174.

[26] MIN B N, ROSS H, SULEM E, et al. Recent advances in natural

language processing via large pre-trained language models: a survey[J]. ACM Computing Surveys, 2023, 56(2): 1-40.

[27] TAN Z, LI D W, WANG S, et al. Large language models for data annotation and synthesis: a survey[J]. arXiv preprint, 2024: arXiv: 2402.13446.

[28] 刘宇宁, 范冰冰. 图数据库发展综述[J]. 计算机系统应用, 2022, 31(8): 1-16.
LIU Y N, FAN B B. Survey on graph database development[J]. Computer Systems and Applications, 2022, 31(8): 1-16.

[29] 张思佳, 于红. 大模型在水产养殖病害防治中的创新应用与展望[J]. 大连海洋大学学报, 2024, 39(3): 369-382.
ZHANG S J, YU H. Innovative applications and prospects of large models in disease prevention and control for aquaculture: a review[J]. Journal of Dalian Ocean University, 2024, 39(3): 369-382.

[作者简介]

张蒋良（2000– ），男，中央民族大学信息工程学院硕士生，主要研究方向为知识图谱、自然语言处理和数字图像处理。

蒲秋梅（1976– ），女，博士，中央民族大学信息工程学院副教授、硕士生导师，主要研究方向为医学图像处理、自然语言处理和机器学习等。

罗训（1977– ），男，博士，天津理工大学计算机科学与工程学院教授、博士生导师，中国计算机学会（CCF）理事，CCF 虚拟现实与可视化专委会主任，智网互联实验室创始人，主要研究方向为数字孪生、虚拟现实、知识图谱和数字图像处理等。

李达（1998– ），男，国家计算机网络应急技术处理协调中心助理工程师，主要研究方向为知识图谱、深度学习、自然语言处理。