姬小川, 应天和, 王枭雄, 等. 基于知识图谱的程序设计知识检索与智能问答系统研究[J]. 智能计算机与应用,2025,15(7):186-193. DOI:10.20169/j.issn.2095-2163.250728

基于知识图谱的程序设计知识检索与智能问答系统研究

姬小川[1], 应天和[1], 王枭雄[1], 胡建鹏[1], 邓洋洋[2]
(1 上海工程技术大学 电子电气工程学院, 上海 201620;
2 粒子跳动(苏州)科技有限公司, 江苏 苏州 215400)

摘 要: 随着知识图谱技术的日益成熟,知识驱动的教育模式在高等教育中的应用越来越广泛,本文旨在设计一个运用于编程学习中,知识获取和问题解决的平台。针对学生在编程实验中的语法错误,通过数据挖掘和自然语言处理技术,提供错误知识点链接,帮助学生更好地理解和解决问题。问答系统能够解析用户输入的问题,生成模板查询语句,并展示匹配结果,为学生提供详细的知识点和实践样例。系统采用领域驱动设计架构,主要功能模块包括用户管理、知识图谱管理、可视化查询、语法错误知识点链接以及知识问答。用户通过系统可视化界面进行知识检索,系统通过 SPARQL 查询语言实现对知识图谱的查询,并通过 Echart 组件呈现查询结果,该系统已在实际教学应用中取得了较好的效果。

关键词: 知识图谱; 知识抽取; 可视化查询; 智能问答; 语法错误链接
中图分类号: TP311 **文献标志码**: A **文章编号**: 2095-2163(2025)07-0186-08

Research on knowledge retrieval and intelligent question answering system based on knowledge graph

JI Xiaochuan[1], YING Tianhe[1], WANG Xiaoxiong[1], HU Jianpeng[1], DENG Yangyang[2]
(1 School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China;
2 Particle Jumping (Suzhou) Technology Co., Ltd., Suzhou 215400, Jiangsu, China)

Abstract: With the increasing maturity of knowledge graph technology, the knowledge-driven education model is more and more widely used in higher education, and this paper aims to design a platform for knowledge acquisition and problem solving in programming learning. In view of students' grammatical errors in programming experiments, the system provides links to knowledge points of syntax errors through data mining and natural language processing technology to help students better understand and solve problems. The Q&A system can parse the questions entered by users, generate template query statements, and display matching results, providing students with detailed knowledge points and practical examples. The system adopts the Domain-driven design architecture, and the main functional modules include user management, knowledge graph management, visual query, knowledge point linking of syntax errors, and knowledge question and answer. The system realizes the query of the knowledge graph through the SPARQL query language, and presents the query results through the Echart component, the system has achieved good performance in practical teaching.

Key words: knowledge graph; knowledge extraction; visual query; intelligent question answering; syntax error linking

0 引 言

人工智能是新一轮科技革命和产业变革的重要驱动力量,其技术的应用也正在改变着教育行业的面貌,传统教育行业注入了人工智能技术的"强心剂","在线智慧教学"为传统教学方式改革提供了新思路。国务院发布的《新一代人工智能发展规划》明确了人工智能技术和应用发展目标,同时提出了一系列政策措施,旨在加强人工智能与经济社会发展的深度融合。这一计划在全国范围内引起了广泛关注和热议,

基金项目: 上海市大学生创新训练计划项目(CS2302006)。
作者简介: 姬小川(2003—),男,本科生,主要研究方向:知识图谱; 应天和(2000—),男,硕士研究生,主要研究方向:知识图谱,实体对齐,知识抽取。
通信作者: 胡建鹏(1980—),男,博士,副教授,主要研究方向:软件工程,服务计算,云计算与物联网等。Email: mr@sues.edu.cn。
收稿日期: 2023-11-19

成为中国加速推动人工智能领域发展的重要指南。为响应这一政策，各大高校积极推进人工智能相关专业和课程的建设[1]，致力于培养更多的人工智能人才。这一举措旨在满足国家对人工智能人才日益增长的需求，推动学术界和产业界更好地协同发展，共同推动人工智能领域的创新和应用。

在当前人工智能等高新技术产业飞速发展的背景下，编程学习已经成为许多学科中不可或缺的一环，区别于传统线下教育模式，线上教育平台显然在编程教育中更具有优势。薛成龙等[2]提出线上教育平台的出现，克服了学生长时间被束缚在传统结构化课程学习中的不足，拓宽了学生在学习中获取的知识广度。李振等[3]提出知识图谱作为推动人工智能发展的核心驱动力，为教育信息化 2.0 时代的教育教学提供了新的赋能力量。将现有的线上教育资源与知识图谱技术结合，将使线上教育平台的知识储备量增加[4]，知识可靠性增强。郎亚坤等[5]提出了基于 Neo4j 构建 C++知识图谱的理论依据，可基于知识图谱进行推理，提供了构建基于知识图谱的程序设计知识检索与智能问答系统的知识基础。

知识图谱是通过图的形式来表现客观世界中的概念和实体及概念和实体之间关系的知识库[6]，知识图谱的概念由谷歌于 2012 年正式提出[7]，2013 年以后开始在学术界和业界普及，并在智能问答、情报分析等应用中发挥着重要作用。国内知识图谱研究文献最早起源于 2006 年大连理工大学侯海燕[8]和刘则渊[9]分别在《情报杂志》和《科学学研究》上发表的两篇文章，之后国内知识图谱研究自 2006 年奠基后，在 2012 年前主要聚焦科学计量与文献可视化，服务于科研评价。2012 至 2016 年受谷歌知识图谱推动，与自然语言处理、深度学习技术深度融合，2017 年至今延伸至智能搜索、金融风控、医疗诊断及工业互联网领域。阿里巴巴等企业构建大规模行业知识图谱，高校则深耕认知推理、多模态融合等前沿方向。当前研究正向动态更新、因果推理和可信知识计算等方向突破。

C 语言学习是计算机专业以及众多非计算机专业高校学生的必修课，有助于学生在学习编程初期构建完整的编程思想。基于多年的教学经验发现，C 语言的语法结构，编程逻辑等知识体系，很难通过传统的书面教学方式进行教学，所以大部分设计编程语言的课程都有机房实操环节。基于这种传统书面教学+机房实验的现有教学模式。本文提出了一种基于知识图谱的程序设计知识检索与智能问答系统，有助于在线上教育发展如雨后春笋的后疫情时代延续线上教育的优势[10]，同时结合知识图谱和机器学习的技术创新，完善系统的不足之处。

1　C 语言领域知识图谱构建方法

1.1　C 语言知识图谱特点

1.1.1　普适性

C 语言知识图谱与其他领域的知识图谱在两个主要方面有显著区别。对于知识体系的关联性，C 语言课程的知识点是相互依存的。例如，由于 C++是 C 语言的高级版本，这两门课程存在相通的知识点，各种语言之间的数据类型也有共同的概念。因此，C 语言知识图谱应具有一定的普适性。

1.1.2　知识图谱的完整性

为了有效地学习 C 语言课程，学生不仅需要掌握其语言特点和语法知识，还需要获得与之相关的专业知识。例如，计算机中的数据存储和数据结构[11]。这些额外的知识有助于理解计算机语言的逻辑，有利于更好地学习编程语言。然而，许多教科书未包含这些重要内容，导致学生在编程思想方面存在欠缺。因此，C 语言知识图谱应具有一定的完整性。

1.2　构建 C 语言知识图谱

1.2.1　数据源与相关技术

知识图谱的构建首先需要构建语料库[12]，再进行知识抽取和扩展，最后进行知识加工。语料库构建的主要数据来源不仅包括教科书等教学资源，本文还通过 Python 爬虫技术爬取编程网站的内容，再邀请 C 语言授课老师以及相关领域专家对网络知识进行审核，前者保证了语料库的准确性，后者保证了语料库的全面性。

同时，本文使用 Text-Rank 算法进行专业术语的提取，然后进行数据预处理。本方法采用远程监督[13]的思想，通过知识网站搜索与主实体相关页面中的 Infobox 列表，抽取得到相关三元组集合，加入到候选语料库中；然后利用启发式算法[14]，通过引入外部知识库自动回标和三元组实体对相关的句子，从而构建训练语料；最后，通过人工审核来保证数据集的质量。

1.2.2　知识扩展

目前，知识抽取仅涵盖了 C 语言知识点的一部分。为了让 C 语言知识图谱更加完整，在未来还需要通过迭代不断增加新的知识点及其关系。具体算法流程如图 1 所示。

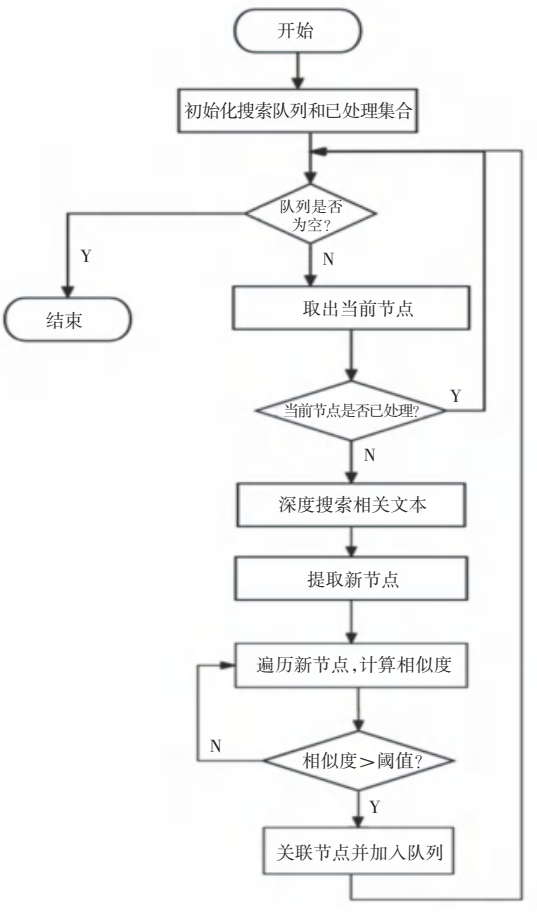

图 1 算法流程

Fig. 1 Algorithmic flow

算法流程主要步骤:

1)在现有知识图谱的基础上,持续搜索与已有知识点相关的文本数据,识别其中的新知识点,并评估新知识点与已有知识点之间的关联度。

2)根据计算结果,如果某个新的文本数据与某一知识点之间的关联度高于特定阈值,则将该文本数据添加为新的知识点。

3)重复上述步骤,直到无法搜索到新的知识点为止,逐步完善 C 语言知识图谱的覆盖范围。

知识点的相关性通常通过语义相似度和语义相关度两种方式进行判断。语义相似度根据知识点的上下文来评估,即两个知识点的上下文越接近,其语义相似度越高;语义相关度通过两个知识点在同一文本中同时出现的概率来评估,如果这两个知识点经常一起出现,则这两个知识点之间可能存在某种语义关联。将语义相似度和语义相关度结合起来,能更好地量化知识之间的关联度,从而通过添加关联度较高的知识点来扩展 C 语言知识图谱。

1.3 C 语言知识图谱可视化

本系统使用基于 Neo4j 的知识图谱数据库为核心,实现知识图谱可视化以及数据库管理。如图 2 所示,作为一款成熟的图数据库,Neo4j 的图形化界面有助于教师与管理员对库数据进行修改与管理,也有助于学生在使用系统时更直观的感受 C 语言知识点之间的逻辑关系。

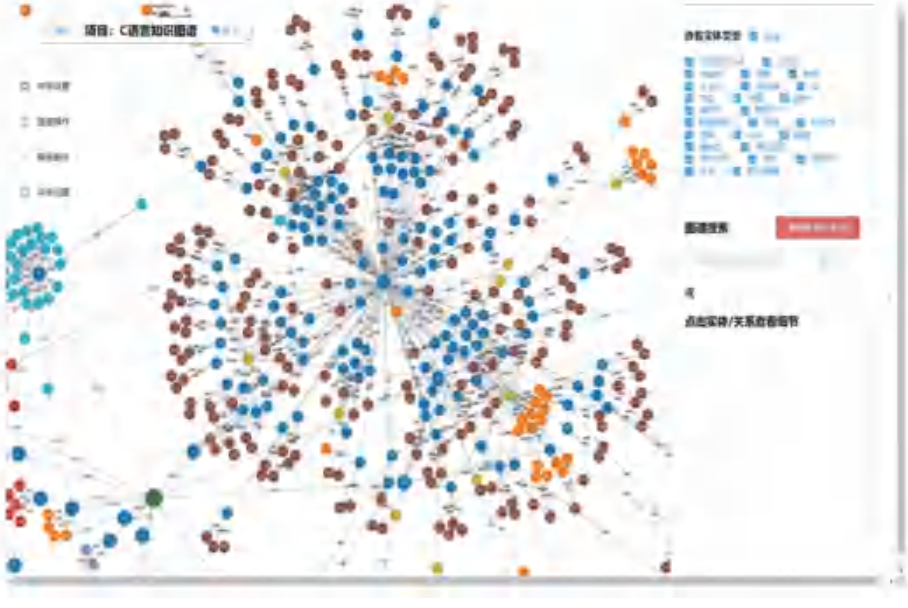

图 2 基于 Neo4j 实现的图谱可视化以及部分操作界面

Fig. 2 Visualization of the graph based on Neo4j and some of the operation interfaces

2　知识检索与智能问答系统的设计与实现

2.1　系统模块设计

本系统实现了用户管理、可视化查询、错误知识点链接、知识问答和知识图谱管理 5 大功能模块，每一个独立的模块都被设计成小的微服务，以便后续业务扩展开发。如图 3 所示，微服务之间的调用统一使用 RestTemplate 进行转发，使用心跳机制确保 Java 与 Python 间的微服务通信。在 Flask 中，提供心跳接口返回服务状态。在 Spring Boot 服务中，启动一个线程对该接口进行监听，如果发现微服务异常，则对所有将转发到 Flask 服务的请求进行熔断。

图 3　微服务通信流程

Fig. 3　Micro-service communication process

2.1.1　语法错误信息知识点链接模块设计

该模块是基于数据挖掘技术和自然语言处理技术[15]实现的。通过分析学生在编程实验中经常出现的错误类型，并将这些错误类型与相关知识点连接起来，为学生提供可能的知识点遗漏补全。系统实现对外提供 API 接口，可以将已有的实验平台系统接入此模块，学生在编译出错时可以根据报错信息去了解相关的知识点。同时，学生可以通过输入编译器返回的错误信息来获取相关的知识点，也可以根据提示信息去了解相关知识点，弥补知识盲点，形成个性化学习闭环。具体流程如图 4 所示。

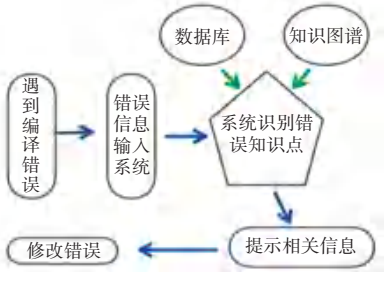

图 4　代码语法错误知识点链接流程

Fig. 4　Code syntax errors, knowledge point linking process

此模块旨在帮助学生更好地理解编程实验中的错误，并提供相关的知识点(如语法规则、实践和编程技术等)，帮助学生更快地解决问题。通过提供上下文敏感信息，学生可以更容易地理解错误的根本原因，并对整个实验过程有更深入的了解。

2.1.2　知识问答模块设计

问答系统可以解析用户输入的问句(见表 1)生成模板查询语句，然后执行得到匹配结果。除了展示详细知识点外，该模块可以通过查询具体的知识点以及关系，帮助学生了解 C 语言编程的实践样例，使其成为初学者的重要参考资源。

表 1　常见问题类别

Table 1　Frequently asked questions

序号	问题类型
1	什么是 XX？/ XX 是什么？
2	XX 的定义/声明/描述/类型/返回值是什么？
3	XX 和 XX 的区别是什么？
4	XX 函数如何调用？/ XX 函数的参数有哪些？/ XX 函数的功能是？
5	XX 代表的含义是？
6	XX 包括什么？/ XX 有哪些？

系统支持的查询类型见表 2，本章为这 3 类查询预定义逻辑模板，最终填充模板生成查询语句。

表 2　查询模板样例

Table 2　Sample query templates

查询类型	自然语言查询	逻辑模板	样例
实体检索	printf 是什么	S	printf
属性检索	printf 的声明是	S:P	printf:声明
多跳检索	printf 如何打印整数	S:P1:P2	printf:参数:格式符

2.2　关键功能实现以及算法设计

本文设计的问答系统可以解析用户输入(见表 2)的问句生成模板查询语句，基于模板执行得到匹配的结果。

系统对于用户输入的问句，识别出其包含的实体、属性以及属性值等。实体识别过程如图 5 所示，该算法使用正则表达式将查询分成若干部分，并使用 jieba 工具对每个部分进行分词。对于每个分词结果，调用 generate_ngram_word 函数进行查找。如果查找到的词组在字典 ent_dict 中，则将其添加到列表 E 中，最终返回包含知识库中实体名的列表。算法还可以将字典匹配替换为搜索引擎(Elasticsearch)，以适应大数据量的场景。

属性识别如图 6 所示，其中函数 map_attr 用于将同义属性映射到知识库中的属性。算法利用 AC 自动机在自然语言查询中查找属性名，并返回匹配的属性名列表。如果需要进行同义属性映射，则调用 map_attr 函数进行映射，最终返回包含知识库中属性名的列表。在识别出所包含的实体、属性等，根据种类确定查询的类型，以便映射到对应的逻辑模

版中。

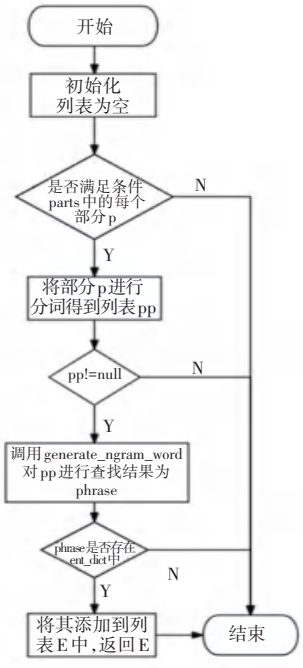

图 5 实体识别

Fig. 5 Entity Recognition

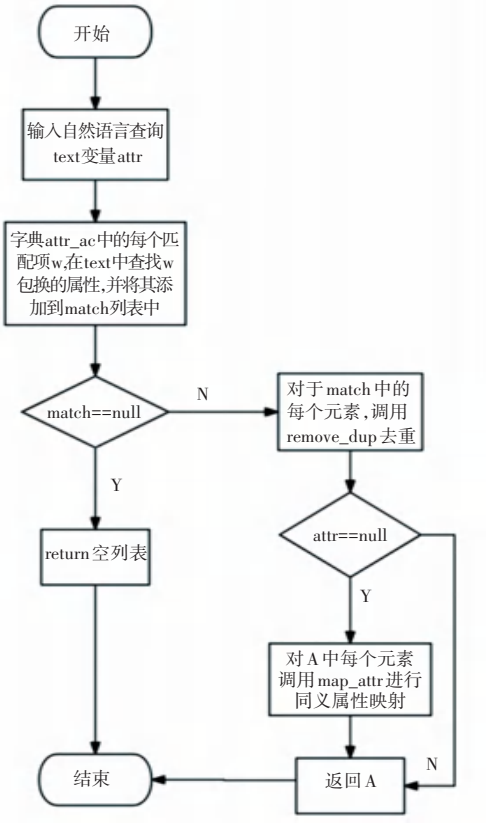

图 6 属性识别

Fig. 6 Attribute identification

属性分类分为以下两种情况:

1)如果出现多个属性,即判定为属性值的多跳查询;如果只有单个属性,则判断实体和属性的位置以及中间的连接词。如果实体的位置出现在属性前,则属于实体的属性查询。

2)如果没有出现实体名称,即判定为根据属性名查询相关的实体类型。对于缺省属性名的属性值,为其补全属性名。

经过上述操作生成逻辑模板之后,根据用户输入的内容需要查询的实体、属性以及类型都将被系统格式化表达,可以直接匹配所属的模板,将逻辑模板转化成系统查询语句(见表 3)。

表 3 部分查询翻译模板

Table 3 Partially queries translation templates

逻辑模板	系统查询语句
S	“query”:{“bool”:{“must”:{ “term”:{“subj”:S} }}}
P:O	“query”:{“bool”:{“must”:{ “term”:{“po. obj”:O} “term”:{“po. pred”:P}}}}

2.3 关键数据结构

在系统可视化查询模块中,使用了知识图谱技术和 Echart 组件,用来展示知识图谱中实体和关系的关联性。为了让用户更加方便地查询知识图谱中的实体和关系,系统采用了基于 RDF 数据模型的查询语言 SPARQL[16],RDF 是一种用于表示 Web 资源的框架和语言的数据模型。其通过一系列的三元组来描述资源的属性和关系,可以方便地在不同的应用和平台之间共享和交换数据。而查询语言 SPARQL 可以用来查询知识图谱中的实体和关系,同时支持对实体和关系之间的关联性进行深入的分析。用户只需要输入实体关键词或关系关键词,系统就可以自动匹配知识图谱中的相关实体和关系,并将查询结果以可视化的形式呈现出来。

3 系统应用效果

在学生使用本平台时,学生通过 Vue3 构建的 html 网页进行操作,可以直接搜索相关知识点,并在页面中查看和知识点有关的相应图谱的可视化结构,也可以在界面上进行基于知识库的问答。网页内还包含编译运行 c 语言代码的功能模块,方便学生在学习知识点的同时进行实操。结合在线编辑运行和错误代码知识点连接功能,可以直接将报错的代码部分连接到知识库中的相应知识节点,效果直观。

问答系统主体基于模板来匹配学生问句,以达到对问句中实体与提问意图的识别。本设计基于完

整 C 语言知识库中的实体节点，结合自然语言大模型生成批量问句生成数据集。基于这个数据集与相应的实体标签进行训练，来提高对问句中实体抽取的准确率。针对模板列表中不存在的问句类型，采取让大模型生成模板外的问句，来提高对学生问句的覆盖率。然而，由于大模型的数据源复杂，不乏劣质数据，生成的模板外问句可能并不会有实际意义，会对数据集的质量造成影响，但大部分生成的问句还是符合要求的。在大模型和机器学习的结合作用下，问答模块对用户的问句意图识别接近 90%。由于后台有完备的知识图谱支持，模块在槽位填充方面的准确率能达到 90%以上。基于上述步骤，本设计中问答系统的问句意图识别率能到达 80%。在生成不唯一的答案的同时还能连接到知识图谱中的相关节点，显示出相关上下文逻辑关系，效果如图 7 所示。

图 7　图谱问答系统

Fig. 7　Graph question and answer system

错误代码链接知识图谱模块，可以将学生输入的错误代码部分与知识图谱中相关知识节点连接，可能涉及多个实体。基于学生输入的历史错误信息和现有 C 语言知识图谱，错误代码链接知识图谱模块的准确率能达到 85%。与问答模块相似，此模块也可以通过图谱可视化的方式展示链接结果，具体使用过程如下：

首先，将编译器中编译过程中得到的错误提示复制到剪切板，如图 8 所示。

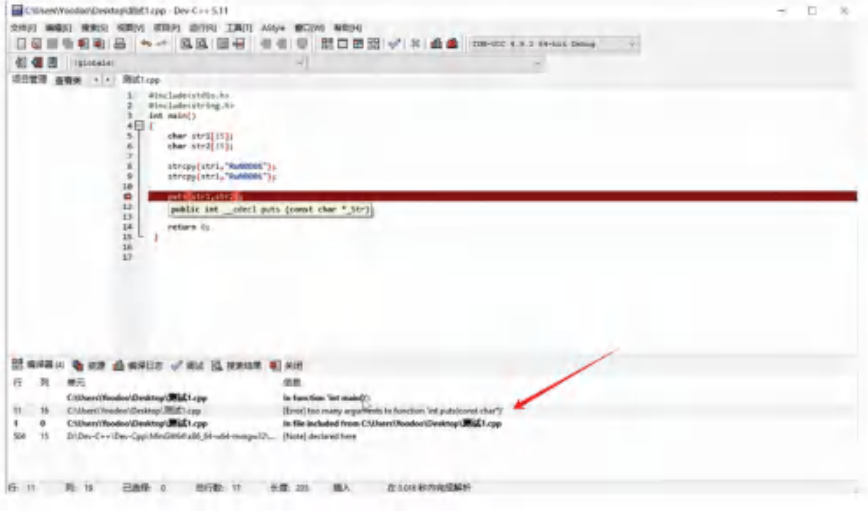

图 8　编译出错

Fig. 8　Compilation error

之后,将错误信息复制到错误知识点链接的搜索栏中,点击右侧进行,搜索如图 9 所示。

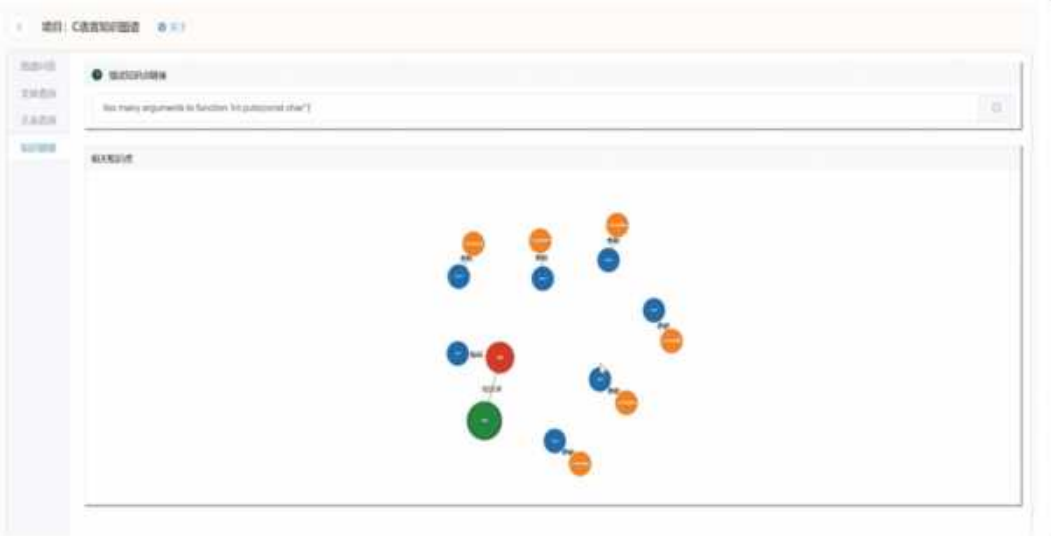

图 9 错误知识点链接

Fig. 9 Error Knowledge Point Link

最后,点击想要查看的知识点可以查看详细信息,如图 10 所示。

图 10 知识点详细界面

Fig. 10 Detailed interface of knowledge points

不同于学生在输入错误代码时直接纠正,通过这种方式,能够从侧面辅助学生学习编程。直接给出答案往往会让学生跳过思考阶段,而通过知识点启发学生,能够加强学生的逻辑思考与自主学习能力。同时发挥线上学习平台的优势,一定程度上减轻教师在授课中的答疑压力。

4 结束语

本研究基于知识图谱技术,结合数据挖掘与自然语言处理技术,设计了一套面向 C 语言教学的知识检索与智能问答系统。通过构建覆盖语法规则、编程逻辑及计算机底层知识的 C 语言知识图谱,实现了知识点关联性表达与语义推理。实验证明,系统在问答意图识别准确率(80%)、语法错误知识点链接准确率(85%)及学生自主学习效率提升方面表现显著,实际教学中有效降低了学生对教师的依赖,强化了编程逻辑思维的培养。该研究验证了知识图谱技术在教育领域的实用性与创新性,为人工智能驱动的新型教育模式提供了可落地的解决方案。

本系统可快速适配其他编程课程,为高校构建智能化教学平台提供技术参考;同时,基于 SPARQL 查询与 Neo4j 可视化的框架可复用于金融、医疗等领域的知识管理;结合大语言模型增强复杂问题处

理能力，融入多模态教学资源（代码案例、视频讲解），有望推动编程教育向“个性化+智能化”方向深度转型，具有较高的推广价值。

参考文献

[1] 楼旭明. 建设“四融合”智学空间，赋能数字时代人才培养[J]. 陕西教育（综合版），2025，646（3）：17-18.

[2] 薛成龙，郭瀛霞. 高校线上教学改革转向及应对策略[J]. 华东师范大学学报（教育科学版），2020，38（7）：65-74.

[3] 李振，周东岱，王勇. “人工智能+”视域下的教育知识图谱：内涵、技术框架与应用研究[J]. 远程教育杂志，2019，37（4）：42-53.

[4] 钟卓，唐烨伟，钟绍春，等. 人工智能支持下教育知识图谱模型构建研究[J]. 电化教育研究，2020，41（4）：62-70.

[5] 郎亚坤，苏超，王国中，等. 基于 Neo4j 的 C++课程知识图谱的构建和推理[J]. 智能计算机与应用，2021，11（7）：144-150.

[6] 黄恒琪，于娟，廖晓，等. 知识图谱研究综述[J]. 计算机系统应用，2019，28（6）：1-12.

[7] 漆桂林，高桓，吴天星. 知识图谱研究进展[J]. 情报工程，2017，3（1）：4-25.

[8] 侯海燕. 权威科学计量学家对科学的关注及贡献[J]. 情报杂志，2006，25（4）：118-120.

[9] 刘则渊. 科学学理论体系建构的思考——基于科学计量学的中外科学学进展研究报告[J]. 科学学研究，2006，24（1）：1-11.

[10] 谭永平. 混合式教学模式的基本特征及实施策略[J]. 中国职业技术教育，2018，684（32）：5-9.

[11] 赵腾飞. C 语言程序设计课程项目化教学改革探索与实践[J]. 知识窗（教师版），2025（1）：114-116.

[12] 范厚龙，房爱莲，林欣. 文本信息与图结构信息相融合的知识图谱补全[J]. 华东师范大学学报（自然科学版），2025，239（1）：111-123.

[13] 赵明，刘胜全，岳柳. 基于 SENT 改进的远程监督关系抽取方法[J]. 现代电子技术，2024，47（16）：51-57.

[14] 桂梁，徐遥，何世柱，等. 基于动态邻居选择的知识图谱事实错误检测方法[J]. 山东大学学报（理学版），2024，59（7）：76-84.

[15] 李爽，陈丽. 国内外网上智能答疑系统比较研究[J]. 中国电化教育，2003（1）：80-83.

[16] 黄恒琪，于娟，廖晓，等. 知识图谱研究综述[J]. 计算机系统应用，2019，28（6）：1-12.