基于知识图谱的中文科技文献问答系统构建研究

李琳娜[1,2] 丁 楷[3] 韩红旗[1,2] 王 力[1,2] 李艾丹[1]
（1. 中国科学技术信息研究所，北京 100038；2. 富媒体数字出版内容组织与知识服务重点实验室，北京 100038；3. 中国航天科工集团六院情报信息研究中心，内蒙古呼和浩特 010000）

摘要：科技文献问答系统能以自然语言对话的方式为用户提供高水平的知识服务。针对语义解析型知识图谱问答系统存在跨领域适应性弱及现有基于深度学习、大模型的问答系统存在结果可解释性差且难以溯源的问题，提出基于句式特点的中文问题分类方法，并设计基于Pipeline方法的中文科技文献问答系统框架。实验结果表明，基于句式特点的问题分类具有不依赖于特定领域的特点且在效果上与基于意图的问题分类基本相当，基于Pipeline的问题解析方法能有效地将问题转化为知识图谱查询语句，从而满足用户对自动问答结果可解释、可溯源的基本需求。

关键词：中文科技文献问答系统；知识图谱；问题分类体系；集成学习
DOI：10.3772/j.issn.1674-1544.2024.04.007 **CSTR：**15994.14.issn.1674.1544.2024.04.007
中图分类号：G250 **文献标识码：**A

Research on the Construction of Question Answering System of Chinese Scientific and Technical Literature Based on Knowledge Graph

LI Linna[1,2], DING Kai[3], HAN Hongqi[1,2], WANG Li[1,2], LI Aidan[1]
(1. Institute of Scientific and Technical Information of China, Beijing 100038; 2. Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content, Beijing 100038; 3. Information Research Center, Sixth Academy of China Aerospace Science and Industry Group, Information Research Center, Huhhot 010000)

Abstract: The Q&A system of scientific and technical literature can provide high-level knowledge services for researchers with natural language. But the current semantic parsing-based knowledge graph Q&A system has poor cross-domain adaptability and Q&A systems based on deep learning or large language model suffer from poor interpretability and traceability of results. Aiming to address these issues, this article proposed a Chinese question categorical method based on sentence patterns and designed a Pipeline based framework for the Q&A system of Chinese scientific and technical literature. The experimental results show that question classification based on sentence patterns does not rely on specific domains and its effectiveness is basically comparable to the question classification based on intentions. The Pipeline-based question parsing method can effectively

作者简介：李琳娜（1981—）女，博士，中国科学技术信息研究所副研究员，研究方向为知识组织、自然语言处理；**丁楷**（1994—），男，硕士，中国航天科工集团六院情报信息研究中心助理工程师，研究方向为文本挖掘、自然语言处理；**韩红旗**（1971—）男，博士，中国科学技术信息研究所研究员，研究方向为文本挖掘、自然语言处理（通信作者）；**王力**（1993—）女，硕士，中国科学技术信息研究所助理研究员，研究方向为情报分析、文本挖掘；**李艾丹**（1984—），女，博士，中国科学技术信息研究所副研究员，研究方向为自然语言处理。

基金项目：中国科学技术信息研究所重点工作项目“智能情报融合创新体系建设研究与应用”（ZD2023-11）；国家重点研发计划项目“颠覆性技术识别理论、方法与专家预判系统”（2019YFA0707201）。
收稿时间：2023 年 2 月 23 日。

transform questions into knowledge graph query statements and effectively meets users' need for Q&A answers with interpretability and traceability of results.

Keywords: Q&A system of Chinese scientific and technical literature, knowledge graph, question categorical method, ensemble learning

0　引言

随着经济社会的发展和科学技术的进步，科研人员每年发表的科技文献数量也在快速增加。据统计，科技文献数量每9年就要翻倍[1]。这对收藏、储存科技文献及提供相关知识服务的图书馆提出了严峻的挑战。现有的信息检索难以满足用户对图书馆海量知识直接、间接的获取需求[2]。如何更好地服务科研人员查询文献、获取文献相关信息成为图书馆开展知识服务面临的重要问题。基于自然语言的问答系统相比于传统的检索引擎更加友好，也更加符合人类的习惯，成为解决这类问题的重要途径之一[3]。而近年兴起的知识图谱被应用到问答系统中能够较好地提升问答系统的性能[4-5]，用户不仅可以得到答案还可以得到与答案相关的其他内容[6]。

现有基于知识图谱的问答系统采用的方法主要有模板匹配、语义解析和深度学习3类[6]。模板匹配方法需要预先构建模板，不需对问题分析而将其转化为三元组形式，再根据三元组去匹配模板并到知识图谱中查询从而获得答案。这个方法虽然简单，但存在需要人工构建模板、模板难以复用的弊端。语义解析方法主要通过问题解析识别用户问题的意图类别，如关于人物的问题、关于地点的问题等类型。进一步抽取问题中蕴含的实体、关系，然后形成机器可理解的查询语言到知识图谱中查询并获得答案，但从意图角度对问题进行分类存在主观性强和领域依赖性强的问题。一方面，全面枚举问题的所有意图类型较为困难，若出现了之前未关注的意图，回答错误的可能性较大；另一方面，在一个领域构建的问答系统很难应用于另一个领域。深度学习方法主要是利用深度神经网络模型对问题与知识图谱三元组的高维抽象表达进行相似性计算，并利用相关评分机制获得最优答案，具有不需要人工定义特征、效率高等优势，但深度学习方法需要一定规模的标注数据、在实际应用中会受到语料不足的限制。另外，深度学习方法对问答结果缺乏可解释性[6]和相关信息的可溯源性。

近期出现的大模型在通用领域的自动问答应用上取得了很好的表现，引起了工业界和学术界的广泛关注。针对ChatGPT对中文处理效果不太理想的问题，国内多个互联网公司及高校先后发布了不同的大模型，如百度的文心一言、阿里的通义千问等。但这些大模型都是在公开的数据上进行训练的，对一些较为专业领域的问答可能会出现胡言乱语的情况。研究人员提出了基于大模型的微调、基于Prompt的提问和将专业知识库与大模型相融合3种方法来解决此问题[7]。基于大模型的微调需要一定规模的高质量训练语料，并没有彻底解决答案的可靠性问题；基于Prompt的提问主要将特定领域的知识作为消息传递给大模型，此方法主要受到输入规模的限制，能够传入的领域知识有限；将专业知识库与大模型相融合是目前比较热门的研究方向，由大模型应用开发框架LangChain支撑其实现。但基于大模型的问答也存在可解释性差、无法对相关信息进行溯源的问题[8]。

科技文献问答系统旨在从海量的科技文献中搜索满足科研人员特定需求的答案，不同于一般的问答系统。一方面，科研人员通常需要科技文献问答系统对给出的答案进行解释，以确信答案的可靠性；另一方面，科研人员在得到答案后，经常需要对问题进行溯源以查找与答案相关的其他信息，即对科技文献用户来说，问答系统结果的可解释性和相关信息溯源是需要考虑的重要因素。

为了解决语义解析型知识图谱问答系统存在的领域适应性差问题，以及当前基于深度学习、大模型的科技文献问答系统可解释性差、难以对

问题结果溯源的问题，本文提出基于句式特点的中文问题分类方法以及基于知识图谱的科技文献问答系统框架，以实现一个能够满足科技文献用户基本需求、实用性强的问答系统。首先，采用Stacking集成学习方法融合多个基分类器提升问题分类的效果；然后，利用深度学习技术提出基于Pipeline方法的知识抽取模型；最后，制定不同的模板将不同句式的中文问题转化成Cypher查询语句以从知识图谱中抽取相应的答案，最终实现基于知识图谱的中文科技文献问答系统。

本文的主要贡献在以下几个方面。①提出了基于句式特点的中文问题分类体系。解决了基于意图分类方法的领域适用性问题。②提出了基于Pipline方法的中文科技文献问答系统构建框架。首先，采用集成学习方法进行问题分类、基于BiLSTM−CRF方法进行实体抽取；其次，利用制定的转化模板将中文问题转化为Cypher查询语句；最后，从知识图谱中获取问题的答案反馈给用户。在这个框架下构建的问答系统不仅可以对结果进行解释，还可以根据科研人员的需求对答案进行溯源。

1 相关工作

1.1 基于知识图谱的问答系统

问答系统旨在向用户给出用自然语言提出的问题的答案。近年来出现的知识图谱由于能表示实体及其之间的语义关系，融合知识图谱的问答系统能够通过挖掘、推理实体之间的潜在关系而较好地提高问答系统的性能，逐渐成为问答系统的研究热点。根据从知识图谱中获得答案的方式不同，可将基于知识图谱的问答系统分为模板匹配、语义解析和深度学习 3 种类型。基于模板匹配的知识图谱问答系统可以快速、准确地找到问题的答案，但是由于需要人工构建大量的模板，使得此方法的可移植性较差[6]。基于深度学习的知识图谱问答系统几乎不需要人工参与，但是这类方法的实现是一个黑盒子，无法对结果进行解释和对答案的相关信息进行溯源[9]。基于语义解析的知识图谱问答系统具有问题解答过程清晰、可解释性强等优点，比较适用于科技文献问答系统的构建。这类方法实现的关键在于将问题转换为知识图谱的查询语言，主要包括问题分类和问题解析两个步骤[10]。

问题分类旨在通过一定的方法将用户以自然语言方式提问的问题归类到预定义的类型。问题分类方法的本质是短文本分类，常用的方法有传统机器学习和深度学习两类[11]。传统机器学习方法主要分为特征提取和模型调优两个阶段，比较典型的分类算法有SVM、决策树等。深度学习方法不再需要专门的特征提取步骤，而是通过学习一组非线性变换直接将特征工程集成到模型拟合过程中[12]，但是这类方法通常需要一定数量的标注数据，比较常用的网络模型有CNN、RNN、Transformer等。

问题解析旨在将以自然语言表述的问题转换为知识图谱的查询语言，实现这一过程的方法主要包括关键词匹配、模板匹配、语义分析和翻译模型。关键词匹配方法首先对自然语言表述的问题进行分词、去停用词，然后筛选出问题的核心关键词，最后以关键词为核心将问题转换为知识图谱的查询语句。这类方法的关键在于对核心关键词的筛选，对于结构复杂或意图不明确的问题，其效果往往不够理想。模板匹配方法将用户问题依照模板库中预定义的模板转换为知识图谱的查询语句，模板库的构建质量直接决定了问题的转化效果。语义分析方法通过语义分析来捕捉问题中包含的语义信息，随后结合问题中包含的实体、关系信息将问题转换为知识图谱的查询语句。翻译模型方法是直接将用户问题转换为知识图谱查询语句的模型，在处理结构简单、意图明确的问题时，这种方法表现出了较好的性能，然而对于结构复杂的问题，其处理效果仍有待进一步提升。

1.2 中文科技文献问答系统

近年来，针对特定领域和通用领域的问答系统取得了大量的研究成果，对于科技文献问答系统的研究也取得了一定的成果。如符泳淋[13]将数字图书馆资源服务的用户检索问题分为描述、机构&人物、地点、出版物、数字和实体 6 个大

类，通过对问题中心词抽取生成查询关键词；郭金刚[3]在对数字图书馆本体和数字资源元数据描述的基础上，采用自然语言解析技术对用户问题进行解析，通过模板匹配得到相应的答案；欧石燕等[14]针对图书馆关联数据自动问答，将用户自然语言问题分为涉及一个数据集的简单问题和涉及多个数据集的复杂问题，将复杂问题分解为只涉及单个数据集的简单问题，最后将简单问题转换为SPARQL查询语句；陆伟等[2]针对武汉大学图书馆的学术知识搜索需求，将用户问题进行分词、命名实体识别等预处理，并基于词典和CRF技术结合规则匹配和模板匹配实现学术知识自动问答；陈傅立等[15]基于科技文献知识图谱和自然语言处理技术设计了基于模板的科技文献问答式智能检索方法。这些方法在对问题的分类上，主要根据答案类型对问题进行分类，需要构建问题答案库作为支撑；在对答案的检索上，主要基于规则和模板两类方法，存在需要人工构建模板且模板不能复用的问题。

2　中文科技文献问答系统模型

2.1　科技文献知识图谱模式构建

知识图谱模式构建是知识图谱构建的第一步，也是知识图谱构建的关键[16]。本文设计的科技文献知识图谱模式如图1所示，主要包含文献、人物、机构、期刊、基金和研究主题6类[17]。考虑到问答系统的应用性，本文选择近年来广泛采用的Neo4j图数据库作为知识图谱的存储载体[18]。

2.2　中文科技文献问答系统实现框架

基于知识图谱的问答系统构建的核心是将用户以自然语言描述的问题转化为知识图谱的查询语句。本文中问题的Cypher查询生成框架基于Pipeline方法实现，如图2所示。主要包括3个部分：第一部分通过总结中文问题的句式特点，建立中文问题类型体系；第二部分采用集成学习Stacking方法融合多个分类模型对问题进行分类；第三部分基于分类结果对问题进行解析，通过实体识别、关系抽取和问题转化3个步骤将中文问题转化为Cypher查询语句，进而从构建的科技文献知识图谱中抽取问题的答案。

2.2.1　基于句式的中文问题分类

不同于现有研究从意图的角度定义问题的类别，本文根据中文问题的句式特点，设计了基于句式的中文问题类型体系，将中文问题分为简单问题、复杂问题和其他问题三大类。简单问题是

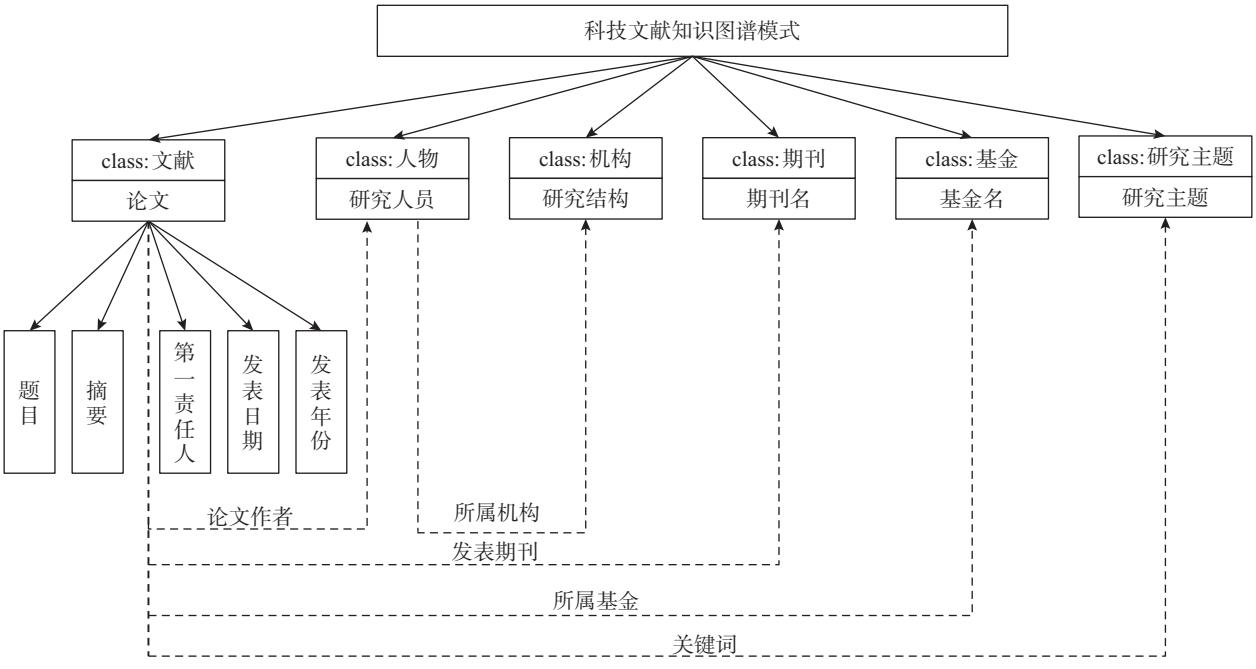

图1　科技文献知识图谱模式

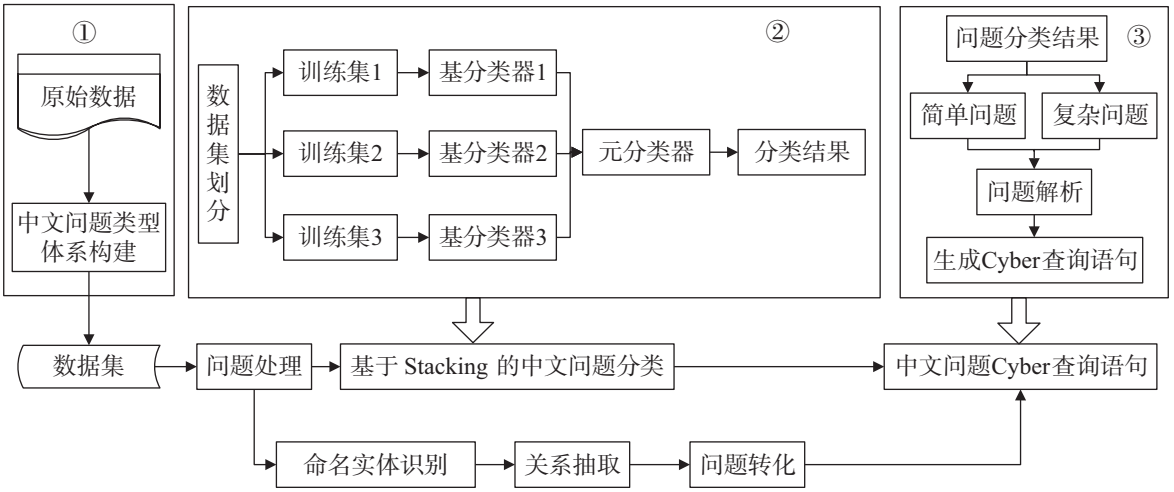

图 2　中文科技文献问答系统实现框架

只包含单一实体或单一关系的问题，即答案位于知识图谱中某个三元组中的主语、谓语或宾语部分；复杂问题的答案需对多个相关联的三元组通过一步或多步推理得到。复杂问题进一步细分为单跳问题、单重约束问题和多重约束问题 3 类。将知识图谱三元组表示为（头实体h，关系r，尾实体t），问题的答案为x，方括号[ ]的运算优先级高于圆括号（ ），基于句式的中文问题分类方法如表 1 所示。

单跳问题是从两个具有相同的头实体或者尾实体但关系不同的三元组中查找答案，即先从第一个三元组（h_1，r_1，t_1）中查找头实体或者尾实体，然后将此中间结果作为第二个三元组（h_2，r_2，t_2）的头实体或者尾实体从知识图谱中进行查找，最终得到问题的答案。

单重约束问题是从两个具有不同实体但关系相同的三元组中查找答案，即先分别从两个三元组（h_1，r，t_1）、（h_2，r，t_2）查找头实体或者尾实体，然后对其取交集得到问题的答案。

多重约束问题是从两个具有相同类型实体但是关系不同的三元组中查找答案，即先对三元组（h_1，r_1，t_1）、（h_2，r_2，t_2）查询得到相应的头实体或者尾实体，然后对其取交集得到问题的答案。

所有不能被归为上述类型的问题均为其他问题。通过对用户问题进行统计分析发现大部分问题都可归类为简单问题与复杂问题，仅有极少数的问题属于其他类别。鉴于这种情况，本文对此类问题暂不处理。

2.2.2　基于集成方法Stacking的中文问题分类方法

现有中文问答系统通常用单独的一个方法对问题分类，但单一方法容易受到数据分布、算法

表 1　基于句型的中文问题分类

总分类	细分类	答案位置	逻辑表述
简单问题	简单问题	查找头实体	（x, r, t）
		查找尾实体	（h, r, x）
复杂问题	单跳问题	查找头实体	（x_2, r_2, [h_1, r_1, x_1]） （x_2, r_2, [x_1, r_1, t_1]）
		查找尾实体	（[h_1, r_1, x_1], r_2, x_2） （[x_1, r_1, t_1], r_2, x_2）
	单重约束问题	—	（x_1, r, t_1）且（x_2, r, t_2） （h_1, r, x_1）且（h_2, r, x_2）
	多重约束问题	—	（x_1, r_1, t_1）且（x_2, r_2, t_2）
其他问题	—	—	一切不属于上述类型的问题

参数等因素影响，使得分类准确率不够理想[19]。因此，本文采用Stacking集成学习融合多个分类算法对中文问题分类。Stacking集成学习方法包括基分类器和元分类器两个部分。由于针对中文短文本分类决策树比其他机器学习方法有较好的效果[20]，本文以优化的决策树算法LightGBM、XGBoost和随机森林作为基分类器。LightGBM与XGBoost基于决策树模型使用Boosting的集成学习方式，随机森林基于决策树模型使用Bagging的集成学习方式。这3个模型在实践中都取得了较好的效果。元分类器既要纠正各算法的偏差，也要具有较好的泛化能力来防止过拟合，因此本文选择逻辑回归算法作为元分类器，具体如图3所示。

2.2.3　中文问题解析

问题解析模型旨在将分类后的问题转化为知识图谱Cypher查询语句，进而从构建的知识图谱中寻找问题的答案。问题解析过程主要包括命名实体识别、关系抽取和问题转化3个步骤，如图4所示。

（1）命名实体识别。命名实体是中文问题的核心词汇，其识别效果直接影响着问题的转化效果，本文对构建的科技文献知识图谱中的实体分别采用不同的方法进行识别。其中，由于论文题目和基金项目名相对较长且不固定，故采用长实体识别效果好的AC自动机[21]对其进行识别。AC自动机不仅具有较高的识别效率、不需要标注数据来训练模型，而且能够进行快速部署。人名、机构名和研究主题由于长度较短，故采用在多个研究中被证实有效[22]的成熟技术BiLSTM-CRF[23]对其进行识别。由于期刊名较少变化、期刊出版时间具有固定的格式，因此对期刊名和时间实体采用基于规则的方式进行识别。

（2）关系抽取。本文构建的知识图谱中主要有论文作者、发表期刊等8种关系。由于BERT模型[24]在多项研究中被证实是非常有效的关系抽取方法[25]，因此采用BERT模型进行关系抽取。将问题与既定的关系拼接输入BERT模型，从而将关系抽取转化为二分类问题。

（3）问题转化。本文采用启发式规则将中文

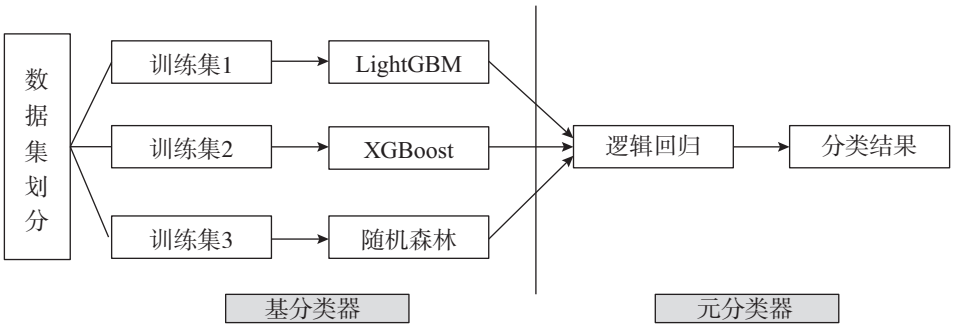

图3　基于集成方法Stacking的中文问题分类模型

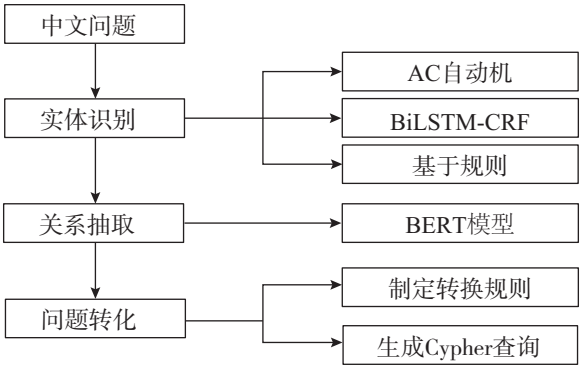

图4　中文问题解析流程

问题转化为知识图谱的Cypher查询语句，每类问题的转化规则如下。

第一，由于简单问题的查询只涉及一个三元组，在问题转化时主要任务是判断所要查询的答案是头实体或尾实体，结合抽取的实体和关系信息制定生成Cypher语句的模板，从而将简单问题转为知识图谱的Cypher查询语句。转化示例如图5所示。

第二，单跳问题的查询涉及两个三元组，在问题转化时需要先从问题中抽取出1个实体和2个关系，然后判断实体和关系的组合方式，进而确定中间答案和另一个关系的组合方式，根据“实体类型1 <关系1> 实体类型2<关系> 实体类型3”的格式，将单跳问题转为知识图谱的Cypher查询语句。转化示例如图6所示。

第三，单重约束问题的查询涉及两个三元组，在问题转化时需要先从问题中抽取出2个命名实体（属于同一类）和1个关系，然后判断每个实体和关系的组合方式，根据“[Class1_entity1] <Relation1> [Class2] And [Class1_entity1] <Relation1> [Class2]”的格式，实现将单重约束问题转为Cypher查询语句。转化示例如图7所示。

第四，多重约束问题的查询涉及两个三元组，在问题转化时需要先从问题中抽取2个实体（属于不同的类）和2个关系，然后判断2个实体与2个关系的组合方式，根据“[Class1] <Relation1> [Class2] And [Class3] <Relation2> [Class4]”的格式，将多重约束问题转为Cypher查询语句。转化示例如图8所示。

3 实验结果与分析

3.1 评价指标

本文构建的科技文献问答系统由问题分类和

示例问句：研究人员包昌火所撰写的论文有哪些?
句型类型：简单问句
命名实体：Class研究人员_包昌火
关系信息：<论文作者>
对应规则：Class文献<论文作者> Class研究人员
规则解释：确定命名实体“Class研究人员”的位置在关系“论文作者”之后，查找的答案类为“Class文献”
Cypher：
```
match (n:Paper)-[:'论文作者']->(p:People)
where p.name="包昌火"
return n
```

图5 简单问题转化示例

示例问句：研究人员苏新宁所撰写的论文都有哪些研究主题?
句型类型：单跳问题
命名实体：Class研究人员_苏新宁
关系信息：<论文作者>，<关键词>
对应规则：Class研究人员<论文作者> [Class文献] <关键词> Class研究主题
规则解释：确定实体“Class文献”在两个关系中的位置，并确定查找的答案实体类为“Class研究主题”
Cypher：
```
match (n:People)<-[:'论文作者']-(p:Paper)-[:'关键词']->(q:key)
where p.name="苏新宁"
return q
```

图6 单跳问题转化示例

示例问句：研究人员苏新宁和许鑫所撰写的论文都有哪些？
句型类型：单重约束问题
命名实体：Class研究人员_苏新宁
Class研究人员_许鑫
关系信息：<论文作者>
对应规则：Class研究人员<论文作者> [Class文献] AND<论文作者> Class研究人员[Class文献]
规则解释：两个命名实体“Class研究人员”的位置位于关系“论文作者”的两侧之后，查找的答案类为Class文献”
Cypher：match (n:Paper)-[:'论文作者']->(p:People)where p.name="苏新宁" AND p.name="许鑫"
return n

图 7　单重约束问题转化示例

示例问句：研究人员苏新宁所撰写的科技情报主题的论文有哪些？
句型类型：多重限定问句
命名实体：Class研究人员_苏新宁
Class研究主题_科技情报
关系信息：<论文作者>
<关键词>
对应规则：Class研究人员<论文作者> [Class文献] AND[Class文献] <关键词> Class研究主题
规则解释：确定两个实体和两个关系的位置，并确定查找的答案类为“Class文献”
Cypher：match (n:People)<-[:'论文作者']-(p:Paper)-[:'关键词']->(q:key)
where p.name="苏新宁" AND q.name="科技情报"
return p

图 8　多重约束问题转化示例

问题转化两部分组成，对其的评价也包括两个部分。一是对问题分类、命名实体识别和关系抽取每一个模块效果的评价；二是问题转化整体效果的评价，检测系统将自然语言问题转化为Cypher查询语句的准确率，进而判断问答系统向用户提供的答案的准确率。主要采用精确率（P）、召回率（R）和$F1$值3种评价指标[26]。其定义分别为：

$$P=\frac{\text{正例预测正确的样本数}}{\text{预测为正例的样本数}} \quad (1)$$

$$R=\frac{\text{正例正确的样本数}}{\text{正例样本总数}} \quad (2)$$

$$F1=\frac{2\times P\times R}{P+R} \quad (3)$$

3.2　实验数据集

由于图书馆领域的中文科技文献问答系统研究没有公开的数据集可用，本文选取了“竞争情报”作为研究领域，从中国知网（CNKI）采集了6 544条论文数据（检索时间为“2021年5月25日”）。首先对采集的论文数据中的著者姓名、机构、基金进行规范化，然后基于设计的知识图谱模式完成实体和关系抽取，共抽取26 563个实体、39 908个三元组。为了构建问句数据集，本文定义<e><r><e>、<h><a><v>（e代表实体，a代表属性名，v代表属性值）为一阶知识元组，$<e_1><r><e_2><a><v>$、$<e_1><r_1><e_2><r_2><e_3>$、$<e_1><r><e_2><r><e_3>$为二阶知识元组，高于二阶的知识元组以此类推。基于抽取的知识元组，确定空缺的知识单元，从而为每一类知识元组分别制定问题预定义模板，进而生成每条格式为“问题、问题的句式类别、基于意图的问题类别、问题中包含的实体及类名、问题中包含的关系”的数据，最终得到52 5542条数据。进一步删除含

有空值的问题，并为了保持类之间的数据分布均衡，以数量最少（17 816 条）的单重约束问题为基础，随机抽取相应数量的其他类型问题，最终得到包含 71 264 条数据的问题数据集。在实验中，按照 8:2 的比例划分训练集和测试集。

3.3 基于Stacking集成方法的中文问题分类实验

为了验证本文提出的基于句式的中文问题分类方法的有效性，基于支持向量机、决策树、随机森林和朴素贝叶斯分类模型分别在构建的数据集上进行基于句式的问题分类和基于意图的问题分类。考虑到当前研究对复杂问句的处理都不太理想，故单独对复杂问题进行分类实验。其实验结果如表 2 所示。

从表 2 可以看出：针对完整数据集，采用决策树和随机森林分类方法，基于句式的问题分类方法略优于基于意图的问题分类方法，但采用支持向量和朴素贝叶斯分类方法，基于意图的问题分类方法优于基于句式的问题分类方法；针对复杂问题，两种分类方法的精确率均有所提升。本文所提出的基于句式的中文问题分类方法在部分分类器中的表现优于基于意图的分类方法，这在一定程度上说明本文所提出的基于句式的中文问题分类方法的有效性。但本文所提出的基于句式的中文问题分类方法能够有效利用中文句式的特点，更加全面、高效地构建问题数据集，在构建问答系统的过程中，这个方法具有更强的可操作性和实用性。

为了验证集成方法Stacking对问题分类的效果，本文采用如下训练方式：首先将中文问题的训练数据均匀地分为 3 份，每份数据分别作为基分类器的输入，然后将每个单分类模型的结果组合起来，作为新的训练数据输入到元分类器中。基于TF−IDF方法进行文本特征提取，通过交叉网络验证来优化参数，从而确保每个分类模型都能达到最佳性能。将集成方法Stacking与 4 个单分类模型、基于BiLSTM的分类模型和基于BERT的分类模型进行实验对比，对比结果如表 3 所示。

从表 3 可以看出，集成方法Stacking的分类效果优于各个单分类模型，其*F*1 值达到 97.80%，比单个模型的*F*1 均值提升了 10.81%，略优于在BiLSTM模型和BERT模型上的效果。推测这是由于中文问题句子较短且训练数据数量少，在BiSLTM和BERT这两个深度学习模型上的效

表 2　基于和意图的中文问题分类实验结果

单位：%

序号	分类模型	完整数据集		复杂问题	
		句式	意图	句式	意图
1	支持向量机	91.45	96.28	97.63	98.20
2	决策树	94.08	94.05	96.76	94.82
3	随机森林	94.84	94.04	96.75	95.00
4	朴素贝叶斯	78.36	84.75	86.71	90.00

表 3　基于Stacking集成方法的中文问题分类实验结果

单位：%

模型	*P*	*R*	*F*1
LightGBM	95.56	96.52	96.03
XGBoost	97.88	96.92	95.92
Random Forest	94.84	95.12	94.98
Logistic Regression	63.56	58.68	61.02
Stacking模型	98.25	97.35	97.80
BiLSTM模型	95.92	96. 27	96.09
BERT模型	97.71	96.70	97.20

果略差于Stacking模型。这也验证了集成方法Stacking对短文本类型的中文问题分类的有效性，为问答系统的实现提供了良好的基础。

3.4　中文问题解析实验

中文问题解析基于Pipeline实现，包括实体识别、关系抽取和问题转化3个子实验。

3.4.1　实体识别

（1）实验数据。根据BiLSTM−CRF模型对输入的要求，对构建的数据集利用BIO标注体系进行人名、机构名和研究主题3类命名实体的标注。最终形成38 560条数据，并按照6:2:2比例划分训练集、验证集和测试集。其数据分布如表4所示。

（2）实验设置。在基于规则的命名实体识别中，由于是采用规则匹配的方式，因此将精确率作为评价指标。在基于深度学习的命名实体识别中，使用Tensorflow深度学习框架实现BiLSTM−CRF模型，利用Word2Vec基于中文维基百科语料训练词向量，其余参数如表5所示。

（3）实验结果与分析。基于AC自动机的命名实体识别精确率达到98%。BiLSTM−CRF方法对人名、机构名和研究主题的识别效果如表6所示。实验结果表明BiLSTM−CRF方法能够有效识别中文问题中的这三类实体。

3.4.2　关系抽取

（1）实验数据。基于构建的数据集将问题与候选关系按照“[*CLS*]问题[*SEP*]关系名[*SEP*]”的形式逐一组合。当问题中包含候选关系时，组合的标签为1，记为正样本；否则，相应的标签为0，记为负样本。最终得到672 507条数据，并按照6:2:2的比例划分训练集、验证集和测试集。

（2）实验设置。本文使用的BERT预训练语言模型为BERT− BaseChinese[27]，模型参数如表7所示。

（3）实验结果与分析。关系抽取的实验结果如表8所示。BERT模型在正样本数据上的精确率达到99.06%，但召回率相对较低，为77.68%，*F*1值为87.08%；在负样本数据上的精确率达到95.95%，召回率为99.86%，*F*1值为97.87%。在

表4　实体识别数据集

单位：条

数据集	人名	机构名	研究主题	总计
训练集	5 509	6 365	11 263	23 136
验证集	1 745	2 256	3 711	7 712
测试集	1 835	2 155	3 722	7 712
总计	9 089	10 776	18 696	38 560

表5　BiLSTM-CRF模型参数

参数	设置
batch−size	16
Word emmbeding	300
learning rate	10−3
dropout	0.5
LSTM−size	128

表6　实体识别实验结果

单位：%

类别	*P*	*R*	*F*1
人名	99.42	99.22	99.32
机构名	92.33	92.95	92.64
研究主题	99.10	98.08	98.58
总计	98.87	98.28	98.58

整个测试数据集上的精确率为97.51%，召回率为88.77%，*F*1值为92.47%，因此这个模型能满足从中文问题抽取关系的需求。

3.4.3 问题转化

由于本文所构建的中文科技文献问答系统框架是先对中文问题进行分类，然后抽取问题中的实体和关系，再依据一定的转化规则将问题转化为知识图谱的cypher查询语句，最终从知识图谱中抽取得到问题的答案，所以问题转化是否准确直接决定了最终的答案是否正确，故本文用问题转化的精确率作为构建问答系统的精确率。将3.2小节构建的71 264条问句数据集按照8:2比例划分为训练集和测试集，用以检验问题解析的效果，评价指标采用精确率。其结果如表9所示。

从表9可以看出，本文所提出的基于Pipeline方法的中文科技文献问答系统对简单问题的回答精确率为90.34%，对复杂问题的回答精确率为80.36%，综合精确率为83.93%。实验表明，本文所设计提出的中文科技文献问答系统在具有答案可溯源、结果可解释性优点的情况下，能够为用户提供精确率良好的结果。

表7 BERT模型参数

参数	设置
编码器	12层
隐藏状态维度	768
Learning rate	2e−5
Batch size	32
dropout	0.1
迭代次数	100

表8 关系抽取实验结果

单位：%

类别	*P*	*R*	*F*1
正样本	99.06	77.68	87.08
负样本	95.95	99.86	97.87
总计	97.51	88.77	92.47

表9 问句转化实验结果

模型	精确率/%
简单问题	90.3
复杂问题	80.36
完整数据集	83.96

4 结语

本文针对当前科技文献问答系统从意图角度对问题进行分类存在跨领域适应性弱问题，以及基于深度学习方法的知识图谱问答系统存在的可解释性差、难以对问题结果溯源等问题，提出基于句型特点进行问题分类和解析、采用知识图谱技术实现的中文科技文献问答系统。基于设计的问题类型，采用Stacking集成学习框架实现中文问题的分类，保证了自然语言问题分类的准确率。针对中文科技文献问答系统对结果可解释性和相关信息可溯源性需求，采用Pipeline方式将自然语言表示的问题转化为知识图谱的Cypher查询语句，进而从知识图谱中查询答案。实验表明，本文所提出的基于句式特点的问题分类方法能够较好地满足将中文问题转为Cypher查询语句的需求，通过知识图谱为用户返回精确度良好的答案。需要说明的是，由于图书馆领域的中文科

技文献问答系统研究还没有公开的数据集可用，本文自建了一个实验数据集，目前仅在此数据上验证了所提方法的有效性。在后续工作中，将在更多的数据集上对本文所提方法进行评估。另外，由于大模型的卓越表现，如何更好地利用大模型提高科技文献问答系统的性能，并解决用户对文献信息溯源的需要，也是今后进一步的研究方向。

参考文献

[1] BORNMANN L, MUTZ R. Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references[J]. Journal of the association for information science & technology, 2015, 66(11): 2215−2222.

[2] 陆伟, 戚越, 胡潇戈, 等. 图书馆自动问答系统的设计与实现[J]. 情报工程, 2019, 5(2): 5−16.

[3] 郭金刚. 面向数字图书馆中文自动问答系统的设计与实现[D]. 成都: 电子科技大学, 2010.

[4] 王寅秋, 虞为, 陈俊鹏. 融合知识图谱的中文医疗问答社区自动问答研究[J]. 数据分析与知识发现, 2023, 7(3): 97−109.

[5] 陈金菊, 王义真, 欧石燕. 基于道路法规知识图谱的多轮自动问答研究[J]. 现代情报, 2020, 40(8): 98−110.

[6] 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.

[7] 张鹤译, 王鑫, 韩立帆, 等. 大语言模型融合知识图谱的问答系统研究[J]. 计算机科学与探索, 2023, 17(10): 2377−2388.

[8] 桑基韬, 于剑. 从ChatGPT看AI未来趋势和挑战[J]. 计算机研究与发展, 2023, 60(6): 1191−1201.

[9] 郑泳智, 朱定局, 吴惠彘, 等. 知识图谱问答领域综述[J]. 计算机系统应用, 2022, 31(4): 1−13.

[10] 张宁, 朱礼军. 中文问答系统问句分析研究综述[J]. 情报工程, 2016, 2(1): 32−42.

[11] 檀莹莹, 王俊丽, 张超波. 基于图卷积神经网络的文本分类方法研究综述[J]. 计算机科学, 2022, 49(8): 205−216.

[12] 唐望径, 许斌, 仝美涵, 等. 知识图谱增强的科普文本分类模型[J]. 计算机应用, 2022, 42(4): 1072−1078.

[13] 符泳淋. 基于数据驱动的数字图书馆中文自动问答系统的研究与实现[D]. 北京: 北京大学, 2014.

[14] 欧石燕, 唐振贵. 面向图书馆关联数据的自动问答技术研究[J]. 中国图书馆学报, 2015, 41(6): 44−60.

[15] 陈博立, 鲜国建, 赵瑞雪, 等. 科技文献问答式智能检索总体设计与关键技术探析[J]. 中国图书馆学报, 2023, 49(3): 92−106.

[16] 杜悦, 常志军, 董美, 等. 一种面向海量科技文献数据的大规模知识图谱构建方法[J]. 数据分析与知识发现, 2022, 6(7): 1−14.

[17] 丁楷. 中文问答系统问句的Cypher查询生成方法研究[D]. 北京: 中国科学技术信息研究所, 2021.

[18] 王力, 韩红旗, 高雄, 等. 关系数据库向Neo4j图数据库转化的应用研究: 以工程科技词系统为例[J]. 中国科技资源导刊, 2021, 53(5): 55−65.

[19] 冉亚鑫, 韩红旗, 张运良, 等. 基于Stacking集成学习的大规模文本层次分类方法[J]. 情报理论与实践, 2020, 43(10): 171−176, 182.

[20] 苑擎飏. 基于决策树中文文本分类技术的研究与实现[D]. 沈阳: 东北大学, 2008: 56−60.

[21] DORI S, LANDAU G M. Construction of aho Corasick automaton in linear time for integer alphabets[J]. Information processing letters, 2006, 98(2): 66−72.

[22] 石教祥, 朱礼军, 望俊成, 等. 面向少量标注数据的命名实体识别研究[J]. 情报工程, 2020, 6(4): 37−50.

[23] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C]//Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human Language Technologies. 2016.

[24] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre−training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2019: 4171−4186.

[25] 葛世奇, 孙新, 寇桓锦, 等. 基于预训练模型的政务领域实体关系抽取[J]. 情报工程, 2022, 8(4): 3−13.

[26] 刘琼昕, 宋祥, 王鹏. 面向出版社富媒体知识的文本分类研究[J]. 情报工程, 2019, 5(2): 40−48.

[27] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre−training of deep bidirectional transformers for language understanding[C]// Proceedings of NAACL−HLT 2019. New York: Association for Computing Machinery, 2018: 4171−4186.