中图分类号：TP391.1　　单位代码：10231
学　　号：2018300604

硕士学位论文

基于机器人知识图谱的智能问答系统的设计与实现

学科专业：　计算机技术
研究方向：　人工智能
作者姓名：　王松磊
指导教师：　姜春茂　教授
　　黄春梅　副教授

哈尔滨师范大学
二〇二〇年六月

中图分类号：TP391.1 **单位代码：10231**

学　　号：2018300604

硕士学位论文

基于机器人知识图谱的智能问答系统的设计与实现

硕士研究生　：王松磊
导　　师　：姜春茂 教授
　　黄春梅 副教授
学科专业　：计算机技术
答辩日期　：2020 年 5 月
授予学位单位　：哈尔滨师范大学

A Thesis Submitted for the Degree of Master

Design and Implementation of Intelligent Question Answering System based on Robot Knowledge Graph

Candidate	: Wang Songlei
Supervisor	: Jiang Chunmao
	Huang Chunmei
Speciality	: Computer Technology
Date of Defence	: May, 2020
Degree-Conferring-Institution	: Harbin Normal University

目 录

摘 要

知识图谱（Knowledge Graph）作为大数据时代的重要设施基础，已经在下一代搜索引擎、智能问答系统等智能应用中有了广泛应用。知识图谱规范地定义了知识的存储，并且可以较为方便和高效的进行知识推理和决策。面向特定领域的知识图谱应用研究也越来越多。

当前，基于机器人领域的知识咨询热度持续升高，但配套的智能问答系统相关技术尚不成熟。本文立足于现实需要，深入研究探讨智能问答系统构建的核心技术，希望能为该系统的解决提供思路，加快系统实现进度，保证用户体验。

本文的研究分为三个步骤，包括：

⑴深入学习研究机器人领域相关知识语料，就实体与实体关系的抽取方式和实体属性的链接方法展开分析，基于此建立并完善知识语料组合构成的机器人知识图谱。

⑵深入研究问句解析技术。第一，分析基于规则匹配和朴素贝叶斯的问句类型识别算法作为确定智能问答系统中问句类型的算法，确定其可行性；第二，使用 TF-IDF 关键词提取算法来提取问句中的关键信息，并使用余弦相似度算法进行问句的匹配；第三，借助 Cypher 查询技术构建查询语句模板，生成具有可执行性的 Cypher 查询语句。结合问句类型识别算法和问句匹配算法，该语句可自动检索图形数据库（Neo4j）中的信息以生成答案，最后经由检索到的结果输出答案。

⑶根据上述确定的问答系统关键技术构建基于机器人领域的智能问答系统，并对系统和配套的算法展开测评，确保这些关键技术可推动智能问答系统的实现。

关键词：机器人；知识图谱；自然语言处理；机器学习；智能问答

Abstract

As an important infrastructure in the era of big data, Knowledge Graph has been widely used in the next generation of intelligent applications such as search engines and intelligent question answering systems. Knowledge Graph defines the storage of knowledge in a standard way and makes knowledge reasoning and decision more convenient and efficient. There are more and more applied researches on Knowledge Graph for specific fields.

At present, the popularity of knowledge consultation in the field of robotics continues to increase, but the related technologies of the supporting intelligent question answering system are still not mature. Based on the practical needs, this paper makes an in-depth study on the core technology of intelligent question answering system construction, hoping to provide ideas for the solution of the system, accelerate the implementation progress of the system, and ensure user experience.

The research in this paper is divided into three steps, including:

(1) Through in-depth study of the relevant knowledge corpus in the field of robotics, this paper analyzes the extraction method of the relation between entities and entities and the linking method of entity attributes, and based on this, establishes and improves robot knowledge graph constituted by the combination of knowledge corpus.

(2) In-depth study of question parsing technology. Firstly, this paper analyzes the problem type recognition algorithm based on rule matching and naive bayes to determine the problem type in the intelligent question answering system, and determines its feasibility; Secondly, TF-IDF keyword extraction algorithm is used to extract the key information in the questions, and cosine similarity algorithm is used to match the questions; Thirdly, the query statement template is constructed by using Cypher query technology, and generate executable Cypher queries. Combining the type recognition algorithm and the question matching algorithm, this statement can automatically retrieve the information in the graph database (Neo4j) to generate the answer, and finally output the answer through the retrieved result.

(3) According to the key technologies of the question answering system identified above, constructing the intelligent question answering system based on the field of robotics, and the evaluation of the system and supporting algorithms is carried out to ensure that these key technologies can promote the realization of the intelligent question answering system.

Keywords: robot; knowledge graph; nlp; machine learning; intelligent question answering

第 1 章　绪论

1.1　课题研究背景

21 世纪是信息化的时代，互联网逐渐走入千家万户。互联网搜索引擎相关技术发展迅速，催生了百度、谷歌这类互联网搜索巨头，在一定程度上帮助人们高效捕捉有效信息。搜索引擎的基本技术思路是，用户输入所需查找的问题，引擎锁定关键词，在数据库中筛选与关键词存在较大关联性的信息，并将其排序反馈至用户端，用户再在此基础上筛选出实际所需的信息。整体来看，搜索引擎技术限制明显，它是一个傻瓜式的搜索系统，无法精准锁定客户需要，而仅仅是机械地捕捉相关信息，并将其提供给用户自行筛选，对于用户快速获得精准答案的需求无法得到满足。问答系统显著有别于搜索引擎，是当代最为重要的信息获取路径之一。问答系统能够精准捕捉用户自然语言问句中的重要信息，并就问题同用户完成信息的交互，实现精准解答。

问答系统实质上是一种信息服务系统，集数据挖掘、信息检索和自然语言处理等技术于一体。Weizenbaum 等[1]成功创建了全球首个机器人领域的问答系统 ELIZA，该系统主要用以辅助精神疾病的治疗。起初，问答系统智能化程度不高，只能机械地从数据库中检索固定答案并进行传递[2-3]。而到 20 世纪 70 年代，Lenat[4] 等借助 Cyc 技术充分利用结构化知识库，大大提高了问答系统的自然语言处理能力。到 2012 年，谷歌首次引用了“知识图谱”这一概念，知识图谱对自然语言的处理水平要显著优于传统的语义网，问答系统的智能化程度显著加深。

鉴于人工智能相关技术的突飞猛进，智能问答系统应用前景愈发广阔。智能问答系统集多种技术于一体，如自然语言处理和信息检索等，需要运用到多学科知识，技术含量极高。问答系统的关键在于其智能化程度，而智能化程度又主要取决于自然语言处理能力和信息检索能力，即系统对于问句语义的分析和相似信息的检索。自然语言显著有别于机器语言，语义复杂，表达多样，在不同的语境下传递的信息截然不同，而这正是自然语言处理所要解决的问题。自然语言处理实质上是通过自然语言技术帮助机器精准识别自然语义，促使机器能够在数据库中检索出相应的答案并传递给用户。

机器人教育是人工智能领域新兴起的一个研究热点，其需要将人工智能技术与信息技术有效结合起来。机器人教育相关研究经过多年的发展仍热度不断，其经历了多个发展阶段，包括初期的基础设计，到中期的教育框架搭建与实践活动落实，再到后期创新思维教育，国家对机器人教育的重视程度一再提高，并将其纳入了信息相关专业的课程之中。但即便如此，机器人教育的研究和实践并非一帆风顺，产业应用同教育之间的矛盾日益凸显。随着 STEAM 理念与创客教育理念的深入人心，机器人教育逐渐在基层教师中收获了优良口碑。机器人教育的教育核心理念渐渐明晰，即培养学生的创新思维能力和实践操作能力。在我国，机器人教育是一个新兴概念，发展态势尤为迅猛，但整体研究仍为数不多，普及程度还不够高。

在机器人知识咨询领域引入问答系统具有重要价值。基于机器人领域的智能问答系统可以智能化理解用户的问句语义，然后从知识图谱中抽取出语义精准的答案内容，并将其传递给用户实现机器与用户之间的交互。

1.2 国内外研究现状

在人工智能快速发展的环境下，智能问答系统的运用前景日益广阔，基于知识图谱的智能问答系统正在成为热门的研究方向。对于本文的国内外研究现状将分别从机器人教育、问答系统、知识图谱和基于知识图谱的问答系统的角度阐述相关研究进展。

1.2.1 机器人教育研究现状

随着《第一代人工智能发展规划》的公布和执行，我国人工智能技术将迈入一个新阶段，该规划极具前瞻性地预估了人工智能的未来发展，并明确了人工智能的发展方向和进展，其中着重强调了人工智能与教育的结合，鼓励在教育领域引入人工智能技术。随后，人工智能相关技术引入成为各行各业追逐的目标。在教育领域，机器人教育是人工智能技术在教育行业探索的先驱者和开拓者，随着 STEAM 理念和创客教育理念的流行，机器人教育的认可程度大大提高，获得广大中小学师生的青睐。机器人教育走进中小学课堂成为新的研究热点和发展趋势。

迄今为止，我国对“机器人教育”的定义尚未确定。总体而言，机器人教育相关课程为信息技术课，主要目的是培养学生的创新思维能力与实践操作能力。经过二十余年的风雨历程，机器人教育以其独特的实践模式逐渐在中小学教育中占据一席之地，但其进一步发展仍面临着重重困境。首先，机器人教育局限于一、二线

城市，受众存在局限性，受区域因素的影响极大。其次，机器人教育的参与度不高，多集中于一些大型的省级、国家级和世界级竞赛，再辅以一些商业化程度较高的兴趣培训班。2012 年，我国新版基础教育新课标公布，其中明确规定了中小学需开设机器人教育相关课程，并正式强调要创建专业性的机器人教育教师团队，希望提高机器人教育教师的实践教育水平，有效培养学生的创新思维意识和实践操作能力[5]。虽然机器人教育的发展困难颇多，但其前景无疑是极为广阔的。

1.2.2 问答系统研究现状

早在 19 世纪 50 年代，Alan Turing（人工智能之父）就率先提出了“图灵测试”[6]，这就是问答系统的雏形。后续百多年来，问答系统相关研究一直是经久不衰的热点。初期，问答系统局限性非常高，仅在限定域内执行，当时的专家系统，其主要是由单一领域的专家建立并完善知识库，并创建一个自然语言处理程序的接口，从而问答一些预设好的专业问题。到 1961 年，Green 等成功创建了全球首个真正意义上的问答系统 BASEBALL[7]，该系统知识库存储了美国某棒球联盟某赛季相关信息，用户可通过该问答系统问询赛季相关的问题，如比赛时间和球员等。1973 年，Woods 也创建了问答系统 LUNAR[8]，它的知识库中存储了阿波罗从月球带回的岩石样本相关分析信息，能够用以回答样本相关问题。由于技术的限制，人工智能发展进入困难时期，这一阶段，人工智能领域的问答系统发展艰难，未曾取得明显进步。

此外，Androutsopoulos 等创建了以自然语言数据库为基础的问答系统 MASQUE[9]，其理论基础是用户以自然语言的方式映射至数据库中，具体而言就是，借助自然语言处理技术来分析问句的词法、语法和语义，将自然语句拆分整合成 SQL 查询语句，然后从数据库中抽取语义相关的答案并交互传递给用户，该方法是智能问答系统的一个重要发展方向。自 1999 年起，首届 TREC（文本检索会议）召开，此后每年一次，该会议为问答系统相关研究者提供了深度交流和探讨的空间，极大地促进了问答系统技术的进步和发展[10-11]。期间，将非结构化数据源引入问答系统引起广泛热议[12]。

21 世纪以来，互联网逐渐走入到大众的生活中，Stack Overflow、Quora 和知乎等社区性的问答系统（CQA）如雨后春笋般涌现，吸引了很多流量和关注，问答数据爆炸式增长。研究人员深入研究了社区型问答系统，开展了对检索识别[13-14]，专家发现[15]，排序规则[16-17]多个方面的分析。现阶段问答系统主要采取机器阅读理解(Machine Comprehension)的方式[18-19]，这种方式以庞大的文本信息量作为基

础，将文本信息作为一个文本库，当系统接收到用户的输入之后，以输入的语义信息为线索，在文本信息库中查找有关的背景知识和支持语料，进而找到答案。

1.2.3 知识图谱研究现状

开放域问答系统必须建立在海量背景知识的基础之上，而知识图谱(Knowledge Graph)这一结构化信息存储就很适合来处理庞大的文本信息。

2012 年 Google 公司首次提出了知识图谱(Knowledge Graph)这一概念[20]，知识图谱是一种知识库，通过利用其结构化的特点来优化搜索引擎性能。

早在 20 世纪 70 年代就已经出现了对知识图谱的研究，Cyc 常识知识库 [21]是这个阶段影响力较大的项目，Cyc 的本体知识库编码集成了生活中的各种常识知识。结构化知识库的发展在 2001 年得到了加速，这一年维基百科(Wikipedia)启动了国际百科全书协作计划，出现了 DBpedia[22]、Yago[23]等优秀的项目。除了维基百科，Freebase[24]这一平台也是通过用户构建的群体智能创作共享类网站，它的知识结构与维基百科不同，采用的是三元组形式。在深度学习与知识图谱的结合方面，词向量的提出[25]为其指明了思路方向，向量化研究成为了知识图谱研究的新热点[26-27]，在算法和工程层面知识图谱都得到了快速的发展。

1.2.4 基于知识图谱问答系统的研究现状

知识图谱与纯文本数据源相比，其优势主要体现在数据的结构化、精度、关联度等方面。知识库问答（结合知识图谱的问答）引起了大量研究人员的关注。

基于知识图谱的问答可以被分成两种，一种是借助符号表示的方法[28]，知识库在解析问句语义时采用一些语义解析(Semantic Parsing)的手段，通过识别实体、关系区分、实体消歧的过程理解问句，并将问句转化为可以在知识图谱中查询的逻辑表达式，最后在知识图谱中找到查询问题的相应答案。第二种是在问答匹配时语义表示表现为分布式[29-30]，用分布式的向量来分析输入问题，再将知识图谱中的信息也向量化，通过对比匹配向量的相似度，在知识图谱中找到最合适的答案，某种意义上这可以视为一种新型的搜索引擎技术。在深度学习的逐渐发展和词向量的引入扩大过程中[31]，可以端到端训练的基于分布式语义表示的知识库问答研究成为了知识库问答的主流研究方法，这一点在目前自然语言处理和全球人工智能顶级会议中可以看出，基于分布式语义表达的方法是目前研究的热点。

知识库问答从问答的复杂程度来看可以分为两类，分别是基于单条知识库三元组的单关系问答（Simple QA）和基于多条三元组的多关系问答（Multi-relation

QA）。在单关系问答中，典型的问题比如“北京是哪个国家的首都？”，单条知识库的三元组即“北京，capital of，中国”就可以与问题匹配。对于多条三元组的多关系问答，比如“中国的首都的地理位置在哪里?”，显然单条数据库的三元组已不能满足需求，需要两条来实现。尽管两种问答难度有所不同，但两者目前都处于探索阶段，并未找到合适的解决方案，有关单关系问答和多关系问答的研究在公开数据集的促进下开始走上正轨。

2014 年 Bordes 等学者围绕多关系问答提出了建立在原有知识图谱基础之上的子图嵌入问答系统[32]，这种问答系统的思路主要从问题出发，分析问题的主体实体，在知识图谱中根据实体找到候选答案，建立主体实体与候选答案的子图，嵌入之后再引入向量研究计算相似度，找到最合适的答案。2015 年 Dong 等人 [33]使用多列卷积神经网络分别对答案的类型、路径、上下文信息进行卷积，提取表征向量，然后与问题向量进行相似度计算。Hao 等人在 2017 年在提取问题特征时引入了注意力机制，在其知识图谱问答设计中采用了双向 LSTM 网络[34]。

学者们对于单关系的问答也做了大量的研究，同样地，主要思路还是从问题出发，分析问题的主体实体，在知识图谱中根据实体找到候选答案，再比较相似度进而找到答案。2015 年 Bordes 等人设计了联系记忆力网络形成的知识库问答[35]。Yin 等人在 2016 年提出了结合注意力机制针对单关系问答的卷积神经网络模型[36], Dai 等人在解决知识库问答问题时建立了神经网络模型，在条件概率框架下解决问题[37]。Golub 等人[38]与 Lukovnikov 等人 [39]分别设计了神经网络匹配模型和注意力神经网络模型，前者基于字符级别，后者基于单词与字符两种级别，均有成效。

国内对于问答系统的研究相比国外较为落后，但其中也有不错的成果展示。杜泽宇等人[40]对问答系统的研究是以电商领域为主要对象，以 JIMI（京东客服机器人）为切入点展开了探究，在其算法上提出了改进建议，发布了其为电商领域研究设计的专业化智能问答系统。钱宏泽[41]针对中草药相关知识设计了智能问答系统，系统利用了本体技术构建了语义网，其三元组容量达到了 300 万，中草药实体来源于中草药数据库，通过制定的模板来自动提取，对于问题的查询，其采用了 SPARQL 过滤式查询策略。在医学领域，张巍[42]等人构建了医疗智能问答系统，这个系统的基础是医学方面的常规问答，通过对知识语料收集并分类整理之后，系统可以提取用户问题与存储问题相比对，这一过程由浅层语义分析技术抽取出问句中的关键词并通过向量计算问句相似度值来实现，最终找到合适的答案。在航空航天领域，张克亮等人[43]构建了专业化智能问答系统，其根据航空领域所涉猎的问题收集存储了大量的三元组集合，系统在识别问题时会识别问题模板并进行对问

题进行分类，大大提高了问答匹配的效率。

1.3 课题研究的目标及意义

现今许多学者都对问答系统方向的研究十分重视，并将其作为研究对象，作为自然语言处理和人工智能领域非常火热的研究方向，问答系统具有巨大的发展潜力。现今科技正飞速发展，超大规模知识图谱技术的研发取得了较大成果，为该课题的研究创造了优良的研究环境，大量的知识和数据能够借助该技术实现结构化的表述并将知识图谱作为存储环境进行保存。工业界和学术界都在基于知识图谱的问答系统(KBQA)领域的研究上花费了大量的精力和资源投入。

本课题旨在研究并构建一个基于机器人知识图谱的智能问答系统。以下内容是本课题研究的主要目标和意义：

(1)对机器人领域知识的表达方式进行深入分析，并将此类表示方式作为基础的问答技术进行探讨，同时给不同受限领域的智能问答系统的创建提供一定的参考基础。

(2)针对机器人领域的智能问答系统算法框架进行设计与构造，对搜索引擎不能很好地理解语义这一难题和网络知识碎片化的问题进行解决。

(3)针对机器人领域设计并实现相关的智能问答系统，通过基于规则匹配和机器学习的相关知识，为用户提供更加便捷、智能化的咨询服务。

1.4 本文目录结构

本文共有四个章节，以下是详细的目录结构：

第一章 绪论。本章先对该课题的研究背景进行分析。并对问答系统、机器人教育、知识图谱和基于知识图谱的问答系统的国内外研究现状进行研究综述，最后对本课题研究的目标和意义进行了介绍。

第二章 课题相关理论及技术分析。本章对涉及到的技术基础和理论知识进行简要的讨论。第一对知识图谱的知识表示方式和基本概念进行介绍，第二对问答系统的技术基础进行概述，并对词向量、命名实体识别、相关机器学习算法和自然语言处理技术进行了介绍与研究。

第三章 智能问答系统的设计及关键技术。本章对智能问答系统的设计及关键技术进行了介绍。主要对系统需求分析和系统整体设计，主要是系统流程设计进行

了介绍，并对本文设计问答系统相关算法，包括基于朴素贝叶斯问句类型识别算法、基于 TF-IDF 和余弦相似度问句匹配算法和 Cypher 查询语句模板生成进行了详细阐述。最后对领域知识的来源和存储进行介绍。

第四章 基于知识图谱的问答系统的实现。本章介绍了智能问答系统的开发环境和整体架构。结合前面几章的研究，编码实现一个基于机器人知识图谱的智能问答系统。最后对智能问答系统进行了整体展示，对本文提出的算法和问答效果进行了实验和分析。

最后是全篇的总结与展望。对本文设计和实现的智能问答系统进行了总结，并且指出了其中尚存的缺点以及下一步的工作方向。

第 2 章　课题相关理论及技术分析

本章对知识谱图所涉及的概念和知识点进行分析，并简单描述本课题所应用的机器学习和自然语言处理的模型结构以及基础理论。

2. 1　知识图谱基础

本节基本概述了知识图谱的知识表示方式，并将包括结构和组成成分等在内的知识图谱的理论部分进行讲解。

2.1.1　知识图谱基本概念

知识图谱在对现实世界的相互关系和概念进行描述时主要利用符号进行，这种知识数据库通常建立在图的结构化的基础上。边将各个类型的实体连接在一起，是概念和实体间相互关系的具象化，而现实世界的语义本体所对应的概念（concept）或实体（entity）用节点表示。知识图谱的基本组成单元是三元组（triple），即{实体，关系，实体}（{subject，relation，object}），且对实体属性的描述也在知识图谱的囊括范围之内，图 2-1 是知识图谱的基本结构。

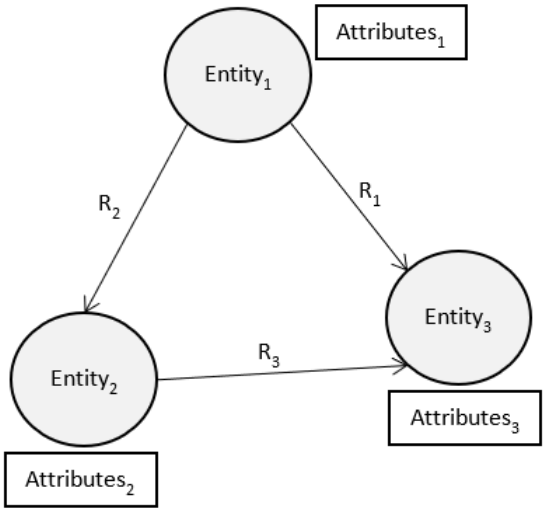

图 2-1 知识图谱基本结构

Figure 2-1 Basic structure of Knowledge Graph

知识图谱选择三元组形式作为知识的描述方式，有助于计算机将实际生活中的知识和信息进行图像化表述，并提高计算机对相关知识和信息的处理和理解。图 2-2 中所示，“成龙居住在香港”这条知识就被为{成龙, lives in, 香港}的三元组形式存储在知识图谱中，其中“成龙”与“香港”是两个实体，“lives in”是两个实体间的联系（relation）。属性（Property）是知识图谱所具有的另一个类型的实体关系，如三元组（成龙, gender, 男人）便是属性三元组。

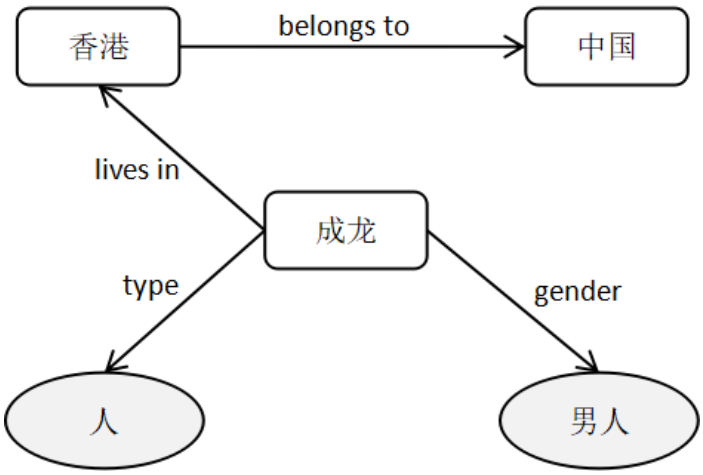

图 2-2 知识图谱三元组样例图

Figure 2-2 Triplet sample graph for Knowledge Graph

2.1.2 知识图谱的知识表示方法

知识表示的概念是人们在人工系统中对于各种心理活动如偏好、渴求、目标、感觉、行为、信仰和知识等进行编码时选择的编码方式。对知识表示方式进行评价时多从自然性、准确性、清晰性三个维度层面进行评价。一个好的知识表示方式应当满足以下要求：第一易于人类理解，第二满足细节要求，第三需不存在歧义。

作为知识表示的重要方式之一，语义网具有关键性的地位。语义网络在对客体、动作、状况、概念和事件之间的关系进行描述时通常利用有向图完成相关工作，该有向图由语义网中的弧（带标记的边）和节点构成。这种描述方式使得语义网络具有灵活且强大的表达能力。在对各个客体之间存在的联系予以描述时这种具备各种标记的有向图能够使其更加顺畅自然[44]。在语义网络中各个节点都可

以有多个属性，并能够对各类动作状态、属性、情况、概念、事物等进行表示。同时，节点还可以作为语义子网络构成嵌套结构，该结构还具有多层次的特点。而带标记的边能够将各个节点间的某种语义关系进行描述，从而对各类语义联系进行阐述。为了对同样对象的不同语义联系和不同对象之间的语义联系进行分辨，弧和节点应当具备相应的标识。一个（节点 1，弧，节点 2）的三元组是最简单的语义网络类型。

以下是语义网具备的优势：

⑴结构性：作为结构化的知识表示方式，语义网络可以形象化的描述事物间的语义联系及其属性。

⑵联想性：语义网的雏形是人类联想记忆模型。

⑶自然性：该表示法能够迅速转化语义网络和自然语言且操作简单，同时还能够使人更好的进行理解语义关系，在表示事物语义联系和属性时具有直观性。

2.2 问答系统技术基础

本节主要介绍机器学习和自然语言处理的经典工作，提供本文相关工作的预备知识，即问答系统常用技术基础。

2.2.1 词向量

机器语言只能识别数字形式，因此对存在于问答过程中的自然语言问句而言只有将其转化成数字形式，才能被机器识别。在对篇章、句子或者词进行描述时需要借助数值化向量。此类表示方式分为两种，包括分布式表示（Distributed Representation）和离散表示（Discrete Representation）。第一种又被称作 word embedding，是利用一个稠密、低维、压缩向量来描述词，第二种表示方式的别称是 one-hot 表示，即把词语表示成一个原子符号，其过程建立在基于规则或者统计方式的基础上。由于本文中使用了 one-hot 词语向量化的表示方法，下面对该表示方法进行介绍。

在自然语言领域的发展早期，One-Hot 是使用频率最高的文本分析方式，也被称作词袋模型，该模式向量的维度和词典大小有关，向量维度会随着词典大小的变化而变化，因此该表示方式具有稀疏性和高维性。单词在词袋模型中用数值向量表示，该向量的长度固定且为词典维度。向量中大部分维度为 0，维度为 1 的只有对该词位置的表述。

“高中”用[0 0 0 0 0 1 …… 0]进行描述

“初中”用[0 0 1 0 0 0 …… 0]进行描述

举例说明，对于“高中”、“初中”而言，在其 One-Hot 向量表述中能够用 1 表示的只有一位。查询词典，两个词在其中的位置分别是 6、3，因此在 One-Hot 向量进行标记时，只有 6、3 位置的数值是 1。One-Hot 具有高度简洁的表达特点，在解决自然语言处理领域的问题时能够和条件随机场（CRF）、支持向量机（SVM）等其他模型共同完成自然语言处理领域的工作。

2.2.2 命名实体识别

命名实体识别是实体抽取技术，也可以称之为关键词提取，其能够在原始的数据语料库当中自动识别出命名实体。

早期实体抽取的重点目标是原始数据当中的时间、地点、人名和专业名称等，并且面向特定行业、业务及领域。1991 年，Rau[45]把启发式算法与人工定制规则融合到一起，实现在文本当中自动抽取并识别公司名称系统的操作。可是，由于建立规则需要耗费大量人力，且延展性差，而且不能在不同领域当中进行随机数据抽取，一定程度上，这种方法仍具有一定的限制。此后，人们开始通过条件随机场（Conditional Random Field，CRF）模型[46]与 K 最近邻（K-Nearest Neighbors，KNN）算法[47]等方式利用统计机器学习方法来对相应的问题制定解决方案，实现将实体部分从原始文本数据当中提取出来。

当下，基于统计的方法、基于词典和规则的方法是实体识别方法中的主要内容。

⑴基于统计的方法

人工标注的语料训练模型是基于统计方法的主要因素，随后依据训练后的模型展开识别工作。在此方法中，条件随机场(Conditional Random Field，CRF)、隐马尔可夫模型(Hidden Markov Model，HMM)、最大熵(Maximum Entropy，ME)以及支持向量机(Support Vector Machine，SVM)是其重要组成部分。但是，这种方法需要准备大量的人工标记语料。

⑵基于词典和规则的方法

起初，命名实体主要是通过词典搭配的规则进行识别的。在这个过程中，词典的构造与规则库是十分重要的，因为专家在工作开始时要先打造出一批规则模板，进而以模板为模型进行识别匹配。可是，数据的数量随着时间的增长日益增多，且各种领域需要对应不同的规则集，凭人工操作已经不容易实现。因此，该

方法在数据量较小时较为适用。

2.2.3 朴素贝叶斯分类

在贝叶斯理论当中，是通过一个已发生事件的概率，计算另一个事件发生的概率。贝叶斯理论从数学上的表示可以写成公式(2-1)这样[48]:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)} \tag{2-1}$$

在这当中，P(A|B)叫作“后验概率”，指的是在 B 事件已发生的情况下，对 A 事件发生概率的预测。P(A)叫作“先验概率”，指的是在 B 事件未发生的情况下对 A 事件发生概率的预测。P(B|A)/P(B)叫作“调整因子”，其中，P(B|A)/P(B)可以小于或者大于 1，指的是在 B 事件发生的情况下，A 事件概率的变化情况。即贝叶斯公式也可表示公式(2-2)所示：

后验概率=调整因子∗先验概率 (2-2)

贝叶斯分类技术学习归纳出分类函数，利用分好类的样本子集展开训练，并通过训练后获得的分类器对尚未分类的数据展开分类工作。其实，贝叶斯分类算法属于非规则的分类方法，并且是一种统计学分类，在众多分类算法中处于重要地位。

朴素贝叶斯分类名字的由来是因为其思想算法较为简单，其属于贝叶斯分类算法的一种。其具体思想如下所示：对给出的待分类项进行分析，研究在该项出现的情况下，每个类别出现的事件概率，找出概率最大的一项，规定该项就是待分类项的具体类别。朴素贝叶斯的思想基础可以概括为：在没有其他能够利用的信息的时候，人们会选择条件符合率最高的类别。

以下为朴素贝叶斯分类的主要定义：

⑴ 设 $x=\{a_1,a_2,...,a_m\}$ 为一个待分类项，而每个 a 为 x 的一个特征属性

⑵ 有类别集合 $C=\{y_1,y_2,...,y_n\}$

⑶ 计算 $P(y_k|x)$ ， $P(y_2|x),...P(y_n|x)$

⑷ 如果 $P(y_k|x)=\max\{P(y_k|x)，P(y_2|x),...P(y_n|x)\}$ ，则 $x\in y_k$

2.2.4 TF-IDF 模型

TF-IDF 模型是一类应用广泛的加权技术，经常被用来进行信息检索和数据挖

掘。TF 是词频(英文全称为 Term Frequency)的简称，可理解为文本内词汇出现的频率，逆文本频率的缩写为 IDF，即一个词语普遍关键性的度量。

此模型的核心思想为：若某短语（或词）于一篇文章内多次出现，即 TF 较高，同时甚少出现于其它文章内，那么判定该短语（或词）具备良好类别区分性能，在分类方面具备适用性。实际上，TF-IDF 为 TF ＊ IDF。其中，TF 代表文档内词条出现的频率。后者 IDF 的核心思想为：若包含词条 t 的文档愈少，即 n 愈小，IDF 则愈大，那么表示，词条 t 在分类区分方面能力突出。若某一类文档 C 内有 m 个文档均内含词条 t，而非此类文档内所含 t 的文档量合计是 k，很明显，n（包含 t 的全部文档量）为上述 m、k 之和。当词条 t 类别区分能力不强的时候，通常表现为，m 大时，n 也大，由公式 IDF 可得 IDF 的值反而小。但是实际上，当一个类的文档内屡次出现一个词条时，其实意味着这个词条能够丰富完整地体现出这一类的文本属性，对于此类词条，应赋予其较高权重，同时可将其当做此类文本的特征词，用来和其它类文档作鉴别[49]。

在某指定文件内，可将 TF 理解为某给定词语于此文件内出现的频率。此数值为词数(term count)归一化处理的结果，即对向量长度实施缩放处理，全部元素的合计值等于 1，由此避免它偏向长文件（相较短文件，同一词语在长文件内的词数可能更高，而与此词语是否重要无关）。就在某一特定文件内的词语 t_i 而言，可通过下式(2-3)来体现其重要性：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2-3}$$

在上式内，符号 $n_{i,j}$ 代表 t_i 此词在 d_j 此文件内出现的次数，那么上式分母表示 d_j 内全部字词出现次数的合计值。

IDF 为逆文本频率的简称，将其作为度量来评估一个词语是否具有普遍重要性。计算词频时，假定所有的单词都是同等重要的。但是不能只依赖每个单词出现的频率，因为像 and 和 the 这样的词出现很多次。为了平衡这些常见词语的频率，需要减少它们的权重并衡量这些罕见词汇。这有助于识别出对每个文档都独一无二的单词，从而制定一个独特的特征向量。某一特定词语 IDF 的求解途径为，总文件数量÷包含此词语的文件数量，然后把两者相除所得值取对数即为 IDF，具体如公式(2-4)所示：

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|} \tag{2-4}$$

此公式中：

- $|D|$代表语料库内文件总量。
- $|\{ j : t_i \in d_j \}|$代表包含 t_i 此词语的文件数量（也可理解为 $n_{i,}$ 不等于零的文件数量）。若此词语 t_i 未在语料库内，就会出现除数等于零的结果，所以，通常使用 $1+|\{ j : t_i \in d_j \}|$。

最后，TF-IDF 的值是这两个值的乘积值, 如公式(2-5)所示：

$$tfidf_{i,j} = tf_{i,j} \times idf_i \tag{2-5}$$

高权重的 TF-IDF 常用来保留重要词语、过滤常用词语。只需要将某文件中出现的高词语频率，以及在文件集合范围内，该词语出现的低文件频率结合，即可生成 TF-IDF。

2.3 本章小结

本章主要是对技术基础和理论知识的概括。主要对知识图谱的基本该概念和知识表示方法，以及包含了词向量、命名实体识别、朴素贝叶斯分类和 TF-IDF 模型在内的对于机器学习技术、关于问答系统的自然语言处理技术的基本介绍。

第 3 章　智能问答系统的设计及关键技术

本章介绍智能问答系统的设计及关键技术。本文设计的问答系统，主要由问题理解、问题求解和答案生成三个模块构成。问句理解主要指分析用户自然语言问句包含的语义信息，常结合分词、词性标注和实体识别技术进行分析。在问题理解后，在知识图谱中搜索查找与其相关的资料和信息，并由此得出答案就是问题求解和答案生成的主要过程。其中问题理解是问答系统研究的重点，也是本文在设计智能问答算法时探讨的重点，本章将对本文设计的智能问答算法进行详细介绍。

3.1　需求分析

由于此问答系统主要是针对机器人领域，为了优化用户体验，提供比搜索引擎更精确的答案，以及简明方便的交互界面，其功能范围和服务需求主要包括以下几点：

⑴在问答算法框架的基础上，对用户输入的问题进行理解，并转化为 Cypher 查找语句，同时返回答案；

⑵及时更新与机器人相关联的知识，完善查询服务，提高其管理的高效性；

⑶通过问答处理的日志，发现并优化该系统在问题的理解方面的偏误与差距，完善知识图谱中存在的纰漏与不足。

3.2　系统整体设计

3.2.1　系统流程设计

本智能问答系统的组成分为问题理解、问题求解和答案生成三个模块。首先由问题理解模块理解剖析收到的用户问题，从中提取出有效信息传达给下个模块。问题求解模块在图形数据库中查询与问题理解模块输出信息相关联的内容。最终答案生成模块以自然语言呈现的结果反馈给用户。因为整个智能问答过程与知识图谱紧密联系，问题理解模块的主要目的就是将用户提供的问题映射到知识图谱中的对应实体上，并由此表达出用户的真正意图。

系统流程如图 3-1 所示，首先通过分词、词性标注对用户问题进行有效提取，

并在问题分类中输入得到的结果。识别出问题的类型后，用户的部分提问意图很大程度上也为系统所感知。当问题对应的知识图谱命名实体抽取出来后，命名实体便被链接到知识图谱中已有的实体上，继而利用问题求解中的查询模板检索图数据库，得到查询的答案，最终答案生成模块将其以自然语言的形式反馈给用户。

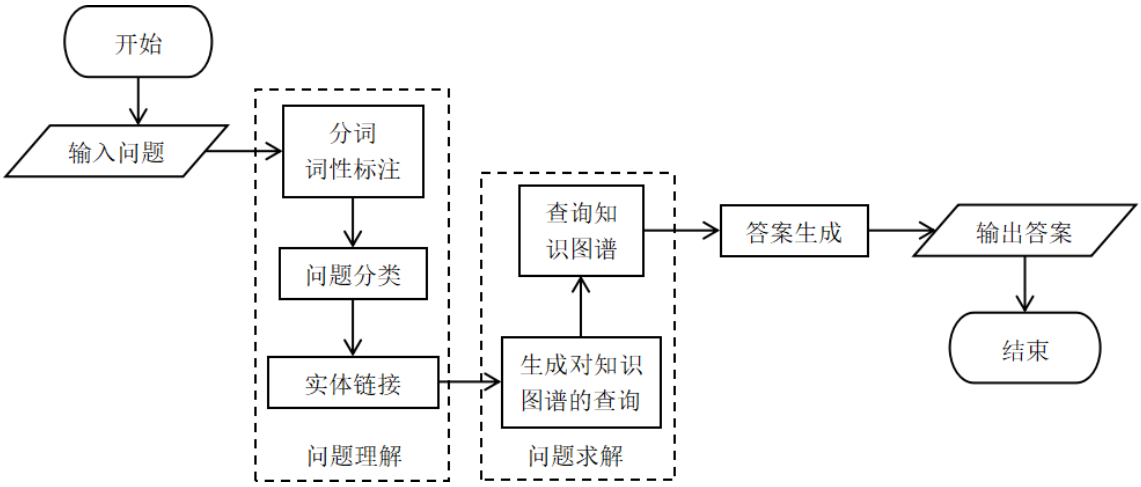

图 3-1 系统流程图

Figure 3-1 System flow chart

3.2.2 分词和词性标注

由于用户问题针对的是知识图谱中的实体，所以用户的提问总是可以根据知识图谱中的元素拆分开。通过分析，对不同的问题进行分类，如：问某一类事物在某一方面的属性、判断一个实例与另一个实例之间是否存在某个关系等。具体样例如下：

Adam 是什么时候发布的？

Swumanoid 和游泳机器人有什么关系？

用户问题首先通过句法分析，生成对句子中的分词、词性标注结果，这些结果被输入到问题模板中。该系统采用 jieba 对用户所输入的自然语言进行分词处理和词性标注。

jieba 是基于的 Python 中文分词组件，它由一系列模型与算法构成，主要是借用一系列的默认词典与用户自定义词典来分析处理用户所输入的自然语言。利用 jieba 中文分词组件对输入的语句进行解析，可以得到句子的分词结果和词性。例如，输入 “我爱北京天安门”，jieba 组件能够在很快的时间内接受并剖析每个词

语的词性和关系分解，如图 3-2 所示。其中，r 表示代词，v 表示动词，ns 表示地名。

```
>>> import jieba
>>> import jieba.posseg as pseg
>>> words = pseg.cut("我爱北京天安门") #jieba默认模式
>>> jieba.enable_paddle() #启动paddle模式。 0.40版之后开始支持，早期版本不支持
>>> words = pseg.cut("我爱北京天安门",use_paddle=True) #paddle模式
>>> for word, flag in words:
...     print('%s %s' % (word, flag))
...
我 r
爱 v
北京 ns
天安门 ns
```

图 3-2 jieba 词性标注示例

Figure 3-2 Example of speech annotations for jieba

当用户所输入的自然语言被 jieba 组件分解后，系统便会自动分析替换语句中的复杂成分以便下一阶段程序正常运行。例如将“我”替换为 r（代词），“爱”替换为 v（动词），“北京”替换为 ns（地名）等。经过这一步骤，系统利用朴素贝叶斯分类法分析预测用户所要表达的含义，进而套用提前设置的句法类型和结构，只有程序能够正确解读用户输入的自然语言的含义，智能问答系统才能给出正确的答案。

3.2.3 基于朴素贝叶斯问句类型识别算法

对于每一类问题，本文实现了相对应的问题模板来描述该类问题的特点，用于识别该类问题，并同时抽取出问题中对知识图谱中实体的指称词，即命名实体。对于问题分类，采用了机器学习的方法，使用朴素贝叶斯分类器进行分类。

该系统会将一系列的训练文本以不同的特征属性为基础进行划分，之后利用朴素贝叶斯分类算法进行分析处理，得出针对每个特征属性的分类器模型。系统在分类模型的指导下分析并判断用户输入的语句与其模型的对应度，将最符合模型计算结果的答案作为对用户问题理解的含义。句子的特征属性代表了句子的含义，如“nr 发布时间”就是一系列询问机器人的发布时间的特征属性。另外，由于输入文本与系统预设的模板间可能存在的差异性，智能问答系统在处理用户问题的类别方面配备了权重衡量标准。而朴素贝叶斯分类器就可以充当智能问答系统在

读取用户问题时“大脑”的角色。朴素贝叶斯分类器的特点是即使每一个特征属性中的训练文本数量较少，也可以估算出必要的参数[50]，如变量的方差或均值等。不过如果某个特征属性内的训练文本数量不足，则会影响到系统对语言匹配的准确率。

通过 jieba 组件分析并替换用户输入的语句这一步骤完成后，系统利用朴素贝叶斯分类器衡量权重并继续处理得到的结果，猜测出用户的问题意图，最终使用程序内预设的语法结构进行回答。图 3-3 呈现出朴素贝叶斯分类算法的整个流程。

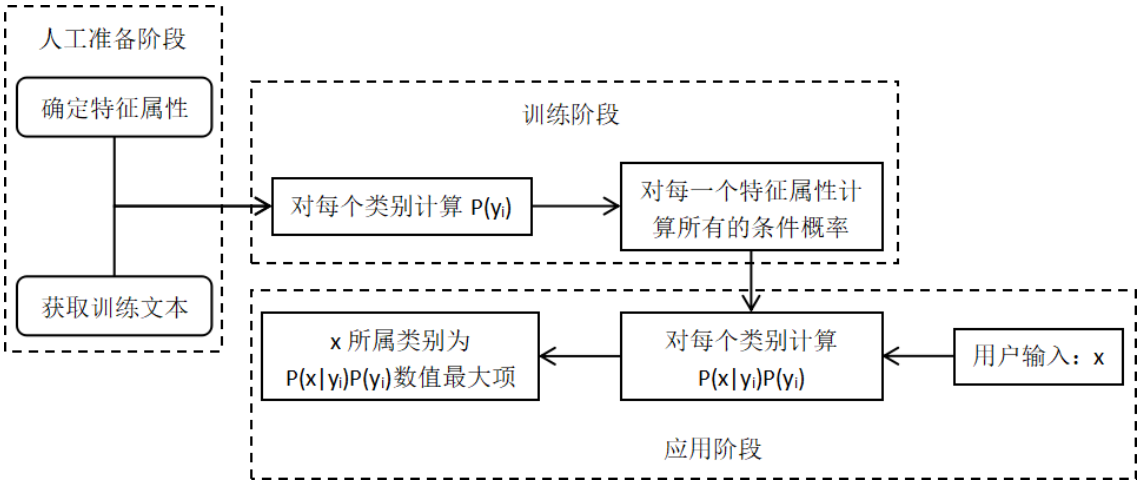

图 3-3 朴素贝叶斯算法流程图

Figure 3-3 Flow chart of naïve bayes algorithm

朴素贝叶斯分类算法主要由人工准备、训练和应用三个阶段组成。

(1) 人工准备阶段。这个阶段主要是确定特征属性，并对每个特征属性进行划分，然后由人工对一部分的待分类项进行分类，形成训练样本的集合。这一阶段的输入是待分类数据，输出的是特征属性和训练样本。

(2) 训练阶段。将生成分类器作为当前阶段的主要任务，负责数据类型的输入和分类器模型的输出并估算数据样本中不同类别属性相似的概率，记录每一个种类出现的频率大小。这个阶段的输入数据是特征属性和训练样本，输出的是分类器模型。

(3) 应用阶段。采用上阶段中分类模型作为工具，对用户输入的待分类项进行分类，输出的是待分类项与类别的归属关系。

详细的公式及算法如下：

(1) 贝叶斯公式如(3-1)所示：

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)} \tag{3-1}$$

(2)假设用户输入的自然语言被分类之后为 x = {a_1, a_2, …, a_n}，其中，每一个 a 为用户输入的自然语言的一个特征属性，x 为用户输入的自然语言的特征属性的汇总。

(3)总类别集合 c = {y_1, y_1, …, y_m}，其中，每一个 y 为已分类项，c 为所有已分类项的集合，即为系统所能回答出的所有问题的集合。

(4)计算出 x 被分类到每一个 y 的几率，即 p(y_1|x), p(y_2|x)…p(y_m|x)。

(5)对于第(4)步中的条件概率可以进行如下统计，如 p(a_1, |y_1), p(a_2, |y_1)…, p(a_n, |y_1)，即为 p(y_1|x)的概率。重复第(5)步，直到 p(y_1|x)为止。

(6)统计 x 的条件下，每一个 y 类别出现的概率，并将此项分类到概率最大的 y 类别之中。如果 p(y_i|x) = max{p(y_1|x), p(y_2|x)…, p(y_m|x)}, 则可以得到 x∈y_i，即 x 可以归到 y 类之中。

(7)如果各个特征属性是条件独立的，则根据贝叶斯定理可以进行以下推导，见公式(3-2)：

$$P(Y_i|x)=\frac{P(x|Y_i)P(Y_i)}{P(x)} \tag{3-2}$$

在第(6)步之后，智能问答系统就会将 y 类归类反馈给数据库，使用数据库语言检索抽取出所需答案。

3.2.4 基于 TF-IDF 和余弦相似度问句匹配算法

当一个新的用户问题出现，将用户问题转化成对应模板形式。接下来计算用户问题模板和所有模板的相似度，将相似度大于某个阈值且最高的模板的意图作为用户问题的意图，并执行对应的 Cypher 查询语句，最终获取用户问题的结果。

计算文本的相似度有很多种手段，首先是对文本进行向量化，也就是将文本转化成计算机理解的语言。单位为词或字的文本向量化手段有 Word2Vec、TF-IDF、N-gram、词袋模型和词集模型这几类。对文本进行向量化处理后，要计算向量距离，也就是计算相似度。计算向量距离的方法有很多，比如 Jaccard 相似性系数、曼哈顿距、欧氏距离、余弦相似度等方法，要按照具体的数据种类与实际情况作出选择。

TF-IDF 是词频-逆文档词频的缩写（Term Frequency-Inverse Document Frequency），是一种用于信息检索与数据挖掘的常用加权技术。TF，也就是词频，代表文本中任一词的出现频率频数的统计，具体统计方法参照公式(3-3)。在全部文本中出现频率高的词就是停用词，对于文本来说，这些停用词并不是十分重要，停用词所获得的权重不宜太高。此处权重的含义是逆文档频率，某个词的出现频率与逆文档频率大小成负相关，具体关系如公式(3-4)。了解了逆文档词频和词频的含义后，综合起来就得到 TF-IDF 值，计算方法如公式(3-5)。某个词在文章中越不可或缺，词本身的 TF-IDF 值就越高。

$$\text{词频（TF）}=\frac{\text{某个词在文本中出现次数}}{\text{文本总词数}} \tag{3-3}$$

$$\text{逆文档词频（IDF）}=\log\left(\frac{\text{语料库中的文档总数}}{\text{包含该词的文档数+1}}\right) \tag{3-4}$$

$$TF-IDF=\text{词频（TF）}\times\text{逆文档词频（IDF）} \tag{3-5}$$

假设有两个对象 X 和 Y，都包括 N 维特征，X = $(x_1, x_2, x_3, \cdots, x_n)$，Y = $(y_1, y_2, y_3, \cdots, y_n)$，计算 X 和 Y 的相似度，余弦相似度计算见公式(3-6)：

$$\cos\theta=\frac{\sum_{i=1}^{n} x_i * y_i}{\sqrt{\sum_{i=1}^{n} x_i^{\,2}\sum_{i=1}^{n} y_i^{\,2}}} \tag{3-6}$$

利用词频-逆文档词频的技术计算余弦相似度，如下图 3-4 所示。先对问题和全部模板分词，统计全部词条出现的次数，计算出任一词条的逆文档词频的数值。当总共存在 k 条词语时，利用 K 维向量表示出任一模板与问句，每个维度上数值表示该词在问句或模板中出现的次数乘上该词语对应的 IDF 值。计算任一模板与问句的余弦值，余弦值越大说明问句和模板相似度越高。设定一个阈值 m，当余弦值的数值超出阈值 m 时，根据相似度最大的模板对应的用户意图作为用户问题的意图，完成 Cypher 语句的执行。

由图 3-4 可知，基于模板匹配的知识图谱，系统的问答体系的解答范围被限制在问题模板列表的大小中。现阶段问题模板列表为人工编写，可能对用户多样化的自然语言输入不能完全覆盖。但如果智能问答系统在线上运行一段时间，就会有一定数量的用户查询日志。这些日志反应了用户的常用问题，基于此日志制作的模板和 Cypher 语句能解决大部分的用户需求。

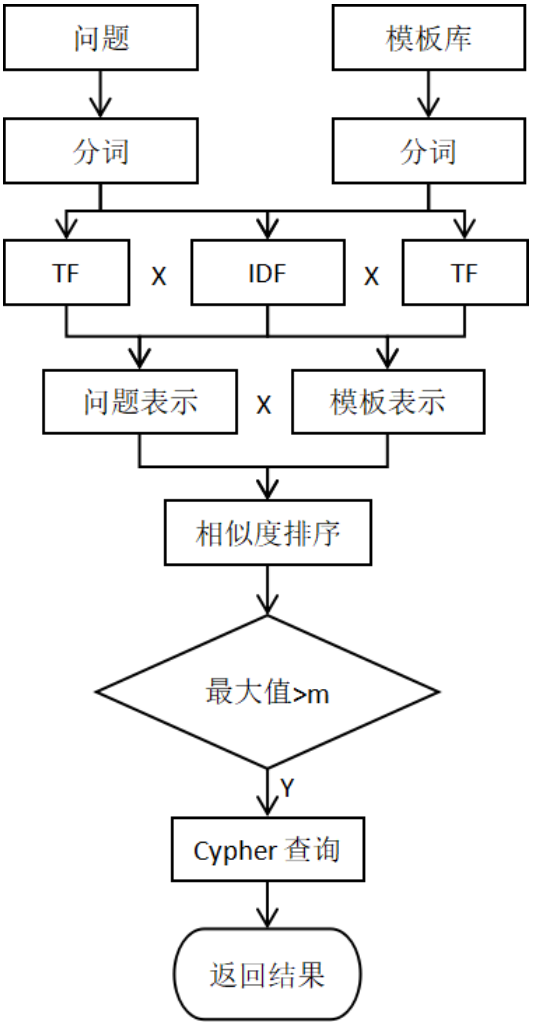

图 3-4 问题和模板匹配流程图

Figure 3-4 Problem and template matching flow chart

3. 2. 5 Cypher 查询语句模板生成

通过 3.2.3 节基于朴素贝叶斯问句类型识别算法和 3.2.4 节基于 TF-IDF 关键词提取算法对输入的自然语言形式问句处理之后，下一步就是要根据这些信息构造知识图谱能够支持的查询语句进行信息的检索。

本文是基于知识图谱建立了机器人领域的知识库，借助优秀的开源图形数据库 Neo4j 进行知识的存储，Neo4j 为用户提供了功能强大的高性能查询语言 Cypher。因此，本文基于 Cypher 构造通用的 Cypher 查询模板。

Cypher 查询模板划分为两类：属性查询类和关系查询类，正好契合了知识图谱的知识表示方法。Cypher 构造模板如表 3-1 所示：

表 3-1 Cypher 查询模板

Table 3-1 Cypher query template

Cypher 模板类型	Cypher 查询模板
属性查询类	match (p) where p.name = NAME and p.attribute = ATTRIBUTE return p.value as ANSWER
关系查询类	match (a{name: ENTITY_A})-[r]-(b{name:ENTITY_B}) return r.name as R_SHIP, r.value as R_SHIP_EX

其中，属性查询类模板中 NAME、ATTRIBUTE、ANSWER 均为变量。NAME 代表输入的机器人名称，ATTRIBUTE 代表输入的属性，ANSWER 代表查询返回的内容。例如问句“ApriPetit 的产地是哪里”，关键词集合为[ApriPetit, 产地]，模板中的 NAME 被替换为“ApriPetit”，ATTRIBUUTE 被替换为“产地”，ANSWER 就是返回的结果。

关系查询类模板中 ENTITY_A、ENTITY_B、R_SHIP、R_SHIP_EX 也是变量。ENTITY_A、ENTITY_B 代表两个存在关系的实体，例如问句“Swumanoid 和游泳机器人有什么关系”的关键词集合为[Swumanoid, 游泳机器人]，模板中 ENTITY_A 被替换成“Swumanoid”，ENTITY_B 被替换成“游泳机器人”。R_SHIP 代表返回的关系结果，R_SHIP_EX 是对返回的关系的一个说明。

当问句通过 3.2.3 节基于朴素贝叶斯问句类型识别算法和 3.2.4 节基于 TF-IDF 关键词提取算法处理之后，系统就可以根据问句的类型获取相应的 Cypher 模板，然后将模板中的变量替换成问句对应的关键词，这样就生成可执行的 Cypher 语句，然后传递给 Neo4j 执行，就可以获取到问句的答案了。

3.3 领域知识来源和存储

领域知识库在问答系统中有很重要的地位，知识库能合理存储并整合领域相关知识。本文选择知识图谱作为机器人领域知识库，将领域知识以图的方式组织起来。

在搭建机器人领域知识图谱体系过程中，使用者要先了解机器人领域的基础

知识。这次整理的机器人领域基础知识是为了从相关机器人网站上爬取的几十种不同类型的机器人，包括机器人名称、简介、产地(研发机构)、发布时间、类型等不同属性。以此为机器人领域的语料，作为此次课题的实验数据。

机器人领域相关语料收集完后，经过人工预处理，将其变为规范化的实验数据。然后，挑选合适的数据库进行机器人领域知识图谱的存储。

Neo4j 数据库是建立在 JVM 基础上的 NOSQL 数据库。是现阶段使用广泛的图形信息库，它具有非常直观和形式化的模型，利用图结构存储领域方面的知识。针对关联度比较高的数据，Neo4j 数据库比关系型数据库的存储和检索速度快上很多，被广泛应用在各个领域，例如生物、基因、医学、金融、社交等领域。

本篇论文选取 Neo4j 数据库作为机器人领域知识图谱的存储工具。Neo4j 能够为上层应用和使用者执行可视化、数据分析、知识检索等操作。可视化主要是根据知识图谱的形式化表示功能来展示，数据分析是基于知识图谱的分析功能，知识检索提供基于知识图谱的知识抽取功能。

3.4 本章小结

本章介绍了智能问答系统的设计及关键技术。首先对系统需求进行了分析。接着对系统整体设计流程进行了介绍。其中，详细讲解了 jieba 分词组件、基于朴素贝叶斯问句类型识别算法、基于 TF-IDF 和余弦相似度问句匹配算法和 Cypher 查询语句模板的生成。最后介绍了领域知识来源和存储。

第 4 章　基于知识图谱的问答系统的实现

根据上一章所构建的智能问答体系的算法框架和流程，本章实现一个基于机器人知识图谱的智能问答系统。该智能问答系统后台使用 Python 编程，开发框架使用 web.py 轻量级框架，前端使用 Html、Css、JavaScript 技术开发，并使用 Bootstrap 框架进行页面美化。本章首先交代系统的开发环境和系统架构。接着从数据层、逻辑层、展示层三个层面对系统的实现过程进行详细介绍。最后对系统进行了展示，对本文设计的算法和系统的问答效果进行了实验，并对实验结果进行了分析。

4.1　开发环境

本系统后台使用 Python 编程，开发框架使用 web.py 轻量级框架，前端使用 Html、Css、JavaScript 技术开发。该系统实现智能问答系统的开发环境如下：

操作系统：Windows10 64 位

CPU：Intel(R) Core(TM) i7-8550U CPU @ 1.80Hz 1.99GHz

内存：8G

开发语言：Python、Css、Html、JavaScript

数据库：Neo4j

Neo4j 数据库是一个高性能的、开源的、功能繁多的图形信息库，这种数据库属于非关系型数据库，利用图结构储存数据。本文利用 Neo4j 数据库完成知识图谱的构建，并在数据库中存储机器人领域知识图谱的相关数据，利用 Neo4j 数据库提供的查询语言 Cypher 来对知识进行检索。

在编程实现系统的过程中还使用了其他工具，包括：

(1) sklearn：scikit-learn 是 Python 的重要机器学习库，简称 sklearn，支持包括分类，回归，降维和聚类四大机器学习算法。还包括了特征提取、数据处理和模型评估三大模块。这里本文用到了 sklearn 中的 tfidf 工具和朴素贝叶斯分类器。

(2) jieba：jieba 是一款开源的中文自然语言处理软件包，该软件包采用 Python 语言实现，实现了中文自然语言处理中包括分词、词性标注、关键词提取等常用功能，并且具有很高的准确率。本文借助 jieba 处理论文中和自然语言处理相关的问

题。

4.2 系统架构

MVC（Model View Controller）设计思想经常被用于 web 开发,本文正是在此思想的基础上，利用展示层、数据层和逻辑层相分离的方法设计智能问答系统。系统架构如图 4-1 所示。

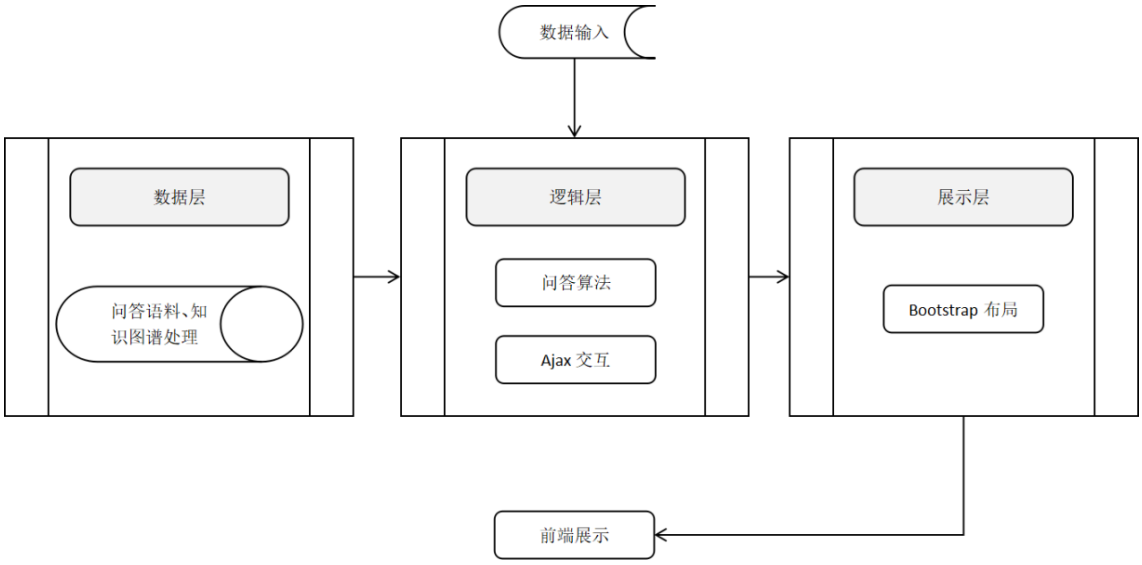

图 4-1 智能问答系统框架图

Figure 4-1 Framework of Intelligent QA

(1) 数据层

问答系统要想正常运转，数据层是必不可少的一个部分，数据层的主要功能是处理数据信息，首先是处理知识图谱数据和问答语料，具体的处理方法是归一化、过滤、建构实体字典等。紧接着需要处理系统运行数据，具体的处理方法是数据分析、查询日志等。所有的处理操作都是依托后台 Python 代码进行。

(2) 逻辑层

本文设计智能问答系统，其后台逻辑层主要分为问题理解、问题求解和答案生成三个子模块。问题理解主要是对自然语言形式的问句进行解析，通过语义信息的提取，包括分词、词性标注、问句类型识别、关键词提取等操作将自然语言进行解析；问题求解主要是将问题与 Cypher 模板相匹配，得到相应的 Cypher 查询语句，通过数据库的调动而查询到所需答案；答案生成主要是检验问句的相似度，和库中

的问答模板匹配，使用正确的模板显示答案。

(3)展示层

系统展示指用户进入人机交互的页面，本系统选择 Python 轻量级 web 框架 web.py。它的优势在于轻量级，支持丰富的扩展功能。前端页面采用 Html、Css、JavaScript 开发。前端主要是用于展示信息和获得用户输入的自然语言提问，将收集的信息利用 Ajax 技术提交到后端，经过处理解析后再返回答案至前端界面。

4.3 系统实现

4.3.1 数据层实现

在系统实现时，本文采用上一章中领域知识来源和存储这一节中介绍和构造的机器人数据作为本系统数据来源。下面将对数据采集、数据存储两个方面进行介绍。

(1)数据采集

本文的结构化数据为从互联网上的机器人网站上爬取。爬虫采用 Scrapy 框架进行搭建，通过代码逻辑自动操控浏览器进行访问和页面的渲染，最后将爬取到的知识内容使用 python 编程写入到 CSV 格式的文件中。

本文所构建的机器人领域知识图谱是以机器人名称实体作为主体，因此首先考虑对机器人相关属性的提取。由于垂直领域网站中的机器人信息均是以机器人作为分类来组织信息结构的，所以先要在垂直领域网站上爬取机器人列表及其对应的 URL，之后通过访问对应的机器人页面并解析来获得该机器人下的属性列表。然后依次访问属性 URL，通过解析页面信息得到该机器人的属性信息，如图 4-2 所示。结构化数据主要是各种表格数据，其中包含了机器人的名称、简介、发布时间、产地和类型等属性信息。

(2)数据存储

本文选择 Neo4j 作为机器人领域知识图谱的存储工具。Neo4j 可以将 CSV 格式的文件批量上传到图形数据库中，并且可以建立实体与实体之间关系的关联和实体内部属性的关联。

CSV 格式的文件设置好表头，将其导入到 Neo4j 文件夹下的 import 目录下，便可进行数据的批量上传。本文数据上传部分实现代码如图 4-3 所示。

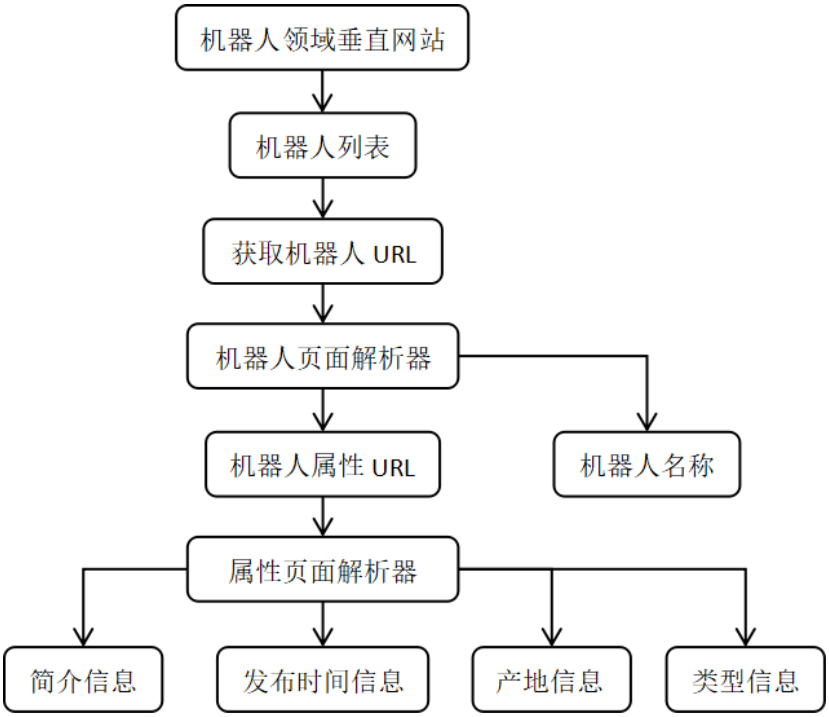

图 4-2 爬虫流程图

Figure 4-2 Crawler flow chart

```
//导入节点 机器人类型
LOAD CSV WITH HEADERS FROM "file:///genre.csv" AS line
MERGE (p:Genre{gid:toInteger(line.gid),name:line.gname})

//导入节点 机器人属性信息
LOAD CSV WITH HEADERS FROM 'file:///robot.csv' AS line
MERGE (p:Robot{ rid:toInteger(line.rid),releasedate:line.releasedate,name:line.name,
introduction:line.introduction,place_origin:line.place_origin,url:line.url})

//导入关系  机器人是什么类型 == 1对多
LOAD CSV WITH HEADERS FROM "file:///robot_to_genre.csv" AS line
match (from:Robot{rid:toInteger(line.rid)}),(to:Genre{gid:toInteger(line.gid)})
merge (from)-[r:is{rid:toInteger(line.rid),gid:toInteger(line.gid)}]->(to)
```

图 4-3 数据上传部分代码实现

Figure 4-3 Data upload part of code implementation

机器人领域数据库字段说明如表 4-1 所示：

表 4-1 数据库字段

Table 4-1 Database field

节点标签	属性	说明
Robot	rid	机器人唯一标识
	name	机器人名字
	releasedate	机器人发布时间
	place_origin	机器人产地
	introduction	机器人简介
	url	机器人链接
Genre	gid	类型唯一标识
	gname	类型名称

所有机器人相关数据信息均来自机器人领域垂直网站。本文爬取了 63 种机器人共 378 类不同的机器人关系和属性集合(机器人名称、简介、发布时间、产地、类型和 url 等)，最终大约抽取了 600 多个实体，3000 多个三元组集合。知识图谱在 Neo4j 上的局部结构如图 4-4 所示。

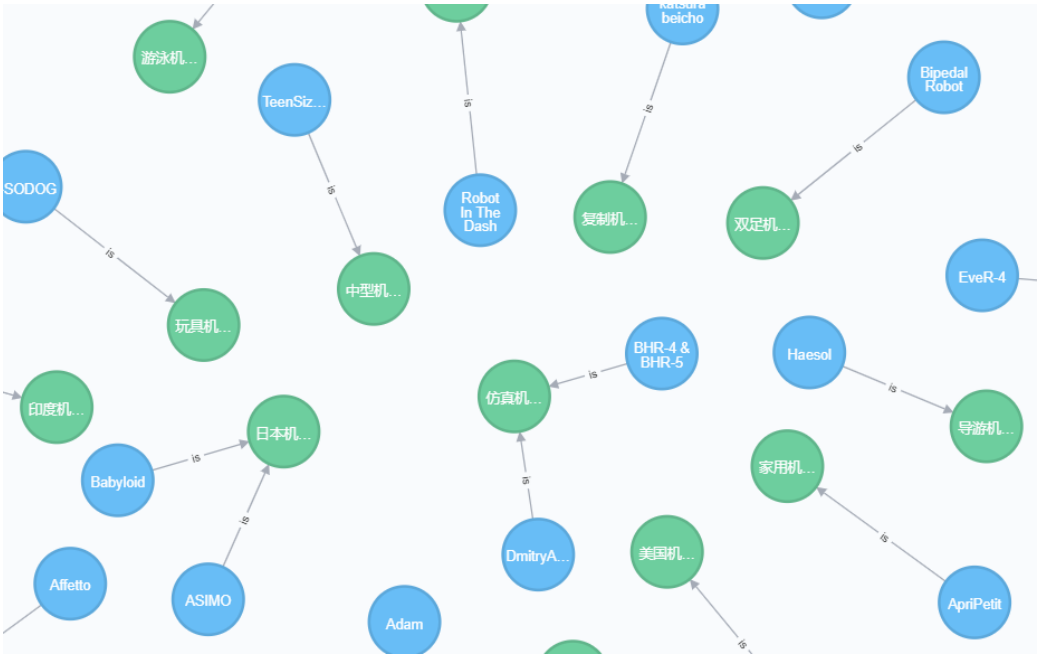

图 4-4 机器人知识图谱局部结构

Figure 4-4 Local structure of robot knowledge graph

4.3.2 逻辑层实现

逻辑层主要实现的是智能问答相关算法。分析用户输入的自然语言，将其中的知识提取出来，然后与知识图谱相匹配，得到所需的答案，并显示给用户。

当用户输入的自然语句被识别后，需要对该语句问题的目的进行解析。这就用到了问题理解子模块。只有问题的意图理解对了，才能得到正确的答案。因此，在该智能问答系统中，问题理解子模块的重要性显而易见，这个模块的主要工作是对用户输入的语句进行分析，抓取语句中的核心信息。信息从前端界面输入到问题理解子模块后，需要对其进行分词、词性标注、问句类型识别、关键词提取等处理，然后将处理过的信息传输到问题求解模块。

(1)分词、词性标注

这个部分的实现工具是 jieba 提供的函数。但是 jieba 的核心字典中并未含有机器人领域的概念词，因此在分词时可能会将专有名词进行拆分，所以需要将机器人领域的概念词纳入到用户的自定义词典中，从而有效的降低拆分概念词现象的发生。自定义词典中的词通过从构建的机器人领域知识图谱中获取，然后对于机器人名称这样的名词定义其词性为“nr”，代表其为机器人名。对于像机器人类型这样的词定义其词性为“ng”，代表类型名词。图 4-5 包含了用户自定义词典中的一些机器人领域的专有词汇。

```
AcYut nr
Affetto nr
ASIMO nr
Autom nr
Babyloid nr
半身仿人机器人 ng
双足机器人 ng
家用机器人 ng
仿真机器人 ng
车载机器人 ng
```

图 4-5 机器人领域自定义词典

Figure 4-5 Custom dictionaries for robotics

(2) 问句类型识别

问句经过 jieba 处理之后就需要分析问句的类型，此时需要用到类型识别程序。本文所设计的类型模板如表 4-2 所示。本系统类型识别的算法采用朴素贝叶斯分类算法进行问题的分类，朴素贝叶斯分类算法在 3.2.3 节中已经进行了详细的介绍，这里就不再说明了。

表 4-2 问句类型模板

Table 4-2 Question type template

序号	问句类型	例子
1	nr 简介	Adam 是什么
2	nr 发布时间	Autom 是什么时间发布的
3	nr 产地	Autom 是哪里生产的
4	nr 类型	Adam 是什么类型的机器人
5	nr 链接	SAMI 的链接地址是什么

(3) 问句关键词提取

问句被正确的识别归类之后，就可以经过关键词提取程序获得问句中的语义关键词。本系统问句关键词提取程序的算法采用 3.2.4 节中介绍的 TF-IDF 算法。下表 4-3 展示部分问句经过关键词提取算法处理之后的关键词集合。

表 4-3 部分问句关键词提取结果表

Table 4-3 Partial question keyword extraction result table

问句	关键词集合
Adam 是什么	[Adam, 简介]
Adam 是什么类型的机器人	[Adam, 类型]
Autom 的发布时间	[Autom, 发布时间]
Autom 来自哪里	[Autom, 产地]
SAMI 的链接地址	[SAMI, 链接]

问题求解子模块接收来自问题理解模块处理后的信息，主要以问句关键词和问句类型的形式展示，然后经过处理后将问题的答案输出。

Cypher 查询语言是本文设计的智能问答系统信息检索时的主要工具。所以，

在 Cypher 模板中找到与用户输入的问题类型相匹配的模板，然后将问题关键词代入其中，实现变量替换，进而得到一个可执行的 Cypher 语句，然后通过 neo4j-driver 工具包提供的驱动将 Cypher 语句传递给 Neo4j 执行，得到与问题类型和关键词所匹配的答案。图 4-6 对 Cypher 执行语句的生成过程进行了展示。

问题求解子模块对于大部分问题可以得到一个初步的答案，但是也存在失效的情况，此时就需要答案生成子模块进行辅助。此时，可以利用答案生成模块对用户的提问进行分析，然后在答案库中寻找相似度最高答案，主要是运用相似度计算的方法，紧接着对答案和问句的相似度进行评判，看是否大于设定的阈值，倘若是，则可以将此答案返回给用户，倘若不是，则在答案栏显示“无法找到相关知识”。问句匹配算法在 3.2.4 节中已经介绍过了，使用的是基于 TF-IDF 的余弦相似度算法。

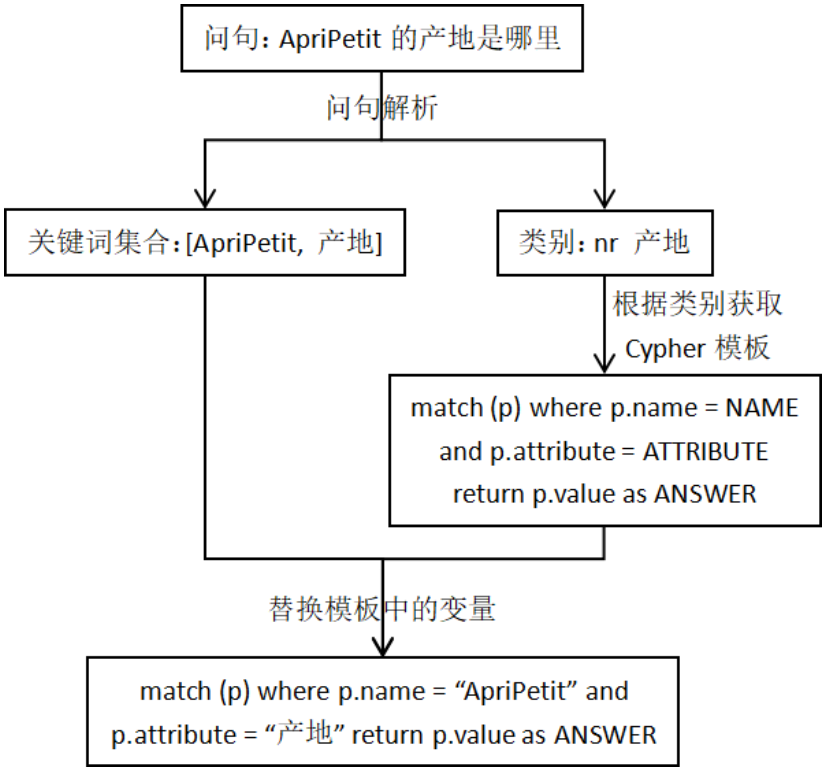

图 4-6 问句到 Cypher 执行语句的生成过程

Figure 4-6 Question to Cypher to execute the statement generation process

4.3.3 展示层实现

若用户想查询机器人相关知识，可以通过前端交互模块进行查询。本文基于

web.py 框架设计开发了一个 B/S 架构的智能问答系统，前端利用 JavaScript、Css、Html 进行系统界面的设计，并使用 Ajax 异步通信技术与后端的处理系统通信。系统界面如图 4-7 所示。由于前端技术开发不是本课题的重点，这里就不做过多介绍了。

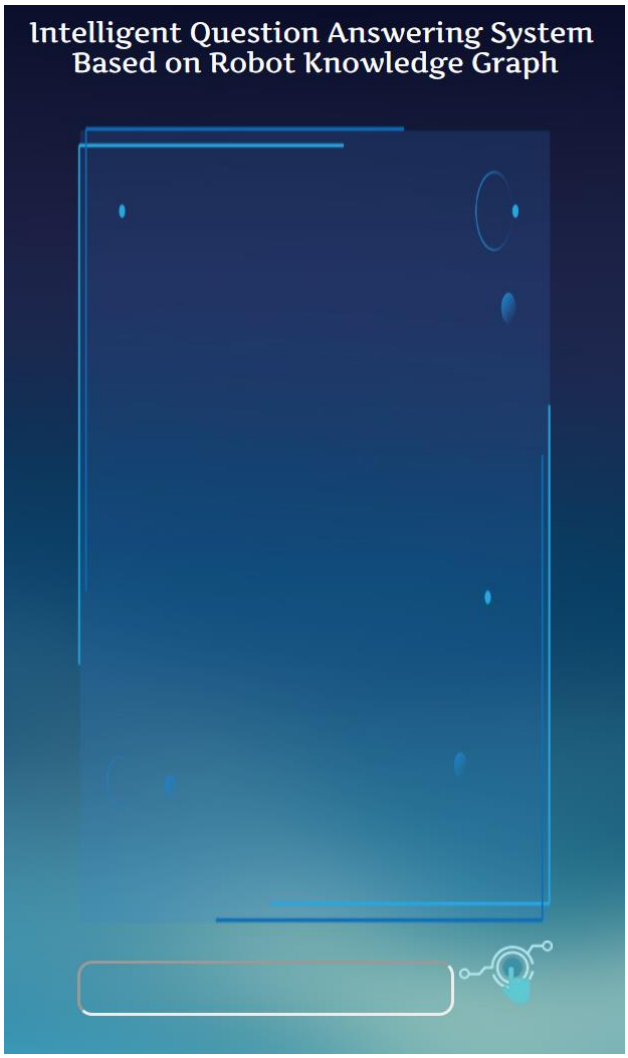

图 4-7 智能问答系统前端界面

Figure 4-7 Intelligent QA front-end interface

4.4 系统展示

最终，本文实现的基于机器人知识图谱的智能问答系统的问答效果见图 4-8 所示。

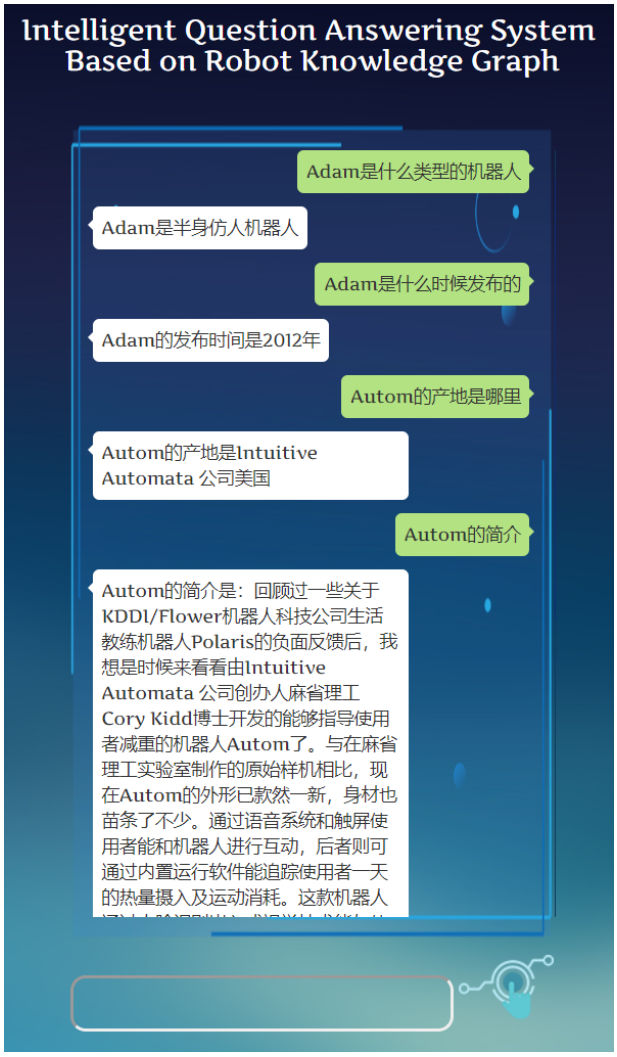

图 4-8 问答效果图

Figure 4-8 Question answering effect diagram

用户提问一个问题“Adam 是什么类型的机器人”，后台在获取到用户所提问题后会经过问题理解模块对问题进行分词处理等操作，然后利用朴素贝叶斯分类器对问题进行分类，得到问题的归属类别。接着提取关键词，利用问题求解模块从知识图谱中检索到答案。最后答案生成模块生成答案返回给前台页面展示。

4.5 实验与结果分析

本节对本文设计并实现的基于机器人知识图谱的智能问答系统的相关算法和

问答效果进行实验。通过相关评价指标，得出实验结果，并对实验结果进行分析。

4.5.1 测试数据

知识库的规模和质量是系统问答效果的关键。因此，从问答的实际情况也可以窥探出基于机器人知识图谱的问答系统构造的程度。

基于规则匹配的知识图谱的问答系统的问答能力取决于 Cypher 查询函数的能力。本文初步设计了 4 类 Cypher 查询函数，如下表 4-4 所示。函数针对一些通用而且简单的问题，具有一定的泛化能力。get_robot_releasedate 可以查询机器人的发布时间；get_robot_placeoforigin 可以查询机器人的产地信息；get_robot_introduction 可以查询机器人的简介；get_robot_genre 可以查询机器人的所属类别。

表 4-4 Cypher 查询函数

Table 4-4 Cypher query functions

Cypher 查询函数	**功能**
get_robot_releasedate	查询机器人的发布时间
get_robot_placeoforigin	查询机器人的产地
get_robot_introduction	查询机器人的简介
get_robot_genre	查询机器人的类别

针对这四个查询函数，结合实际用户案例，手工编写了 200 条用户问题(各 50 条)和对应答案以检测算法的泛化能力。

4.5.2 评价指标

基于知识图谱的问答系统属于问答系统，问答系统的回答的答案可能出现三种结果，第一种是找到知识库中的实体（属性），第二种是找到知识库中的实体关系，第三种是没有在知识库中找到答案。利用精确率（Precision）、召回率（Recall）、F1-Score 对系统的问答效果进行评价。如表 4-5 所示。

针对二分类的问题，正类指的是关注的类，负类指的是其他类，预测测试数据集时分类器会出现两种结果，正确或者不正确，将四种情况发生的总数计为以下四种。

False Negative（FN）：指正例样本被错误地标记为负例的数目

False Positive（FP）：指负例样本被错误地标记为正例的数目

True Negative（TN）：指负例样本被分类器正确分类的数目

True Positive（TP）：指正例样本被分类器正确分类的数目

明显可以得出，测试样本总数=FN+TN+FP +TP。

表 4-5 评价指标

Table 4-5 Evaluation index

	正例（预测结果）	负例（预测结果）
正例（真实情况）	TP	FN
负例（真实情况）	FP	TN

精确率：指的是实际正类在分类器预测的正例样本中的占比，即：

$$precison = \frac{TP}{TP + FP} \tag{4-1}$$

召回率：指的是在总的正样本数量中，预测正确的正例样本的占比，即：

$$recall = \frac{TP}{TP + FN} \tag{4-2}$$

通常情况下，召回率和精确率是反比的关系，前者越低，后者越高，反之亦然，将召回率和精确率求调和平均数，计为 F1-Score，当二者都很高时，F1-Score 的值也会越大，即：

$$F1 = \frac{2}{\frac{1}{precison} + \frac{1}{recall}} = \frac{2 \times precison \times recall}{precison + recall} \tag{4-3}$$

4. 5. 3 算法实验结果

针对 3.2.3 节提出的基于朴素贝叶斯问句类型识别算法和 3.2.4 节基于 TF-IDF 和余弦相似度问句匹配算法进行实验。本次共对 200 个问句进行实验，实验结果见表 4-6 所示。

表 4-6 算法实验结果

Table 4-6 Experimental results of algorithm

算法	精确率	召回率	F1-score
问句类型识别算法	86%	89%	87.5%
问句匹配算法	85%	82%	83.5%

从上表的实验可以得出，本文提出的基于朴素贝叶斯问句类型识别算法、基于 TF-IDF 和余弦相似度问句匹配算法能够有效识别用户输入的问题。精确率没有达到更高的原因可能是测试集中有一部分问句的句式结构过于复杂。本文提出的算法对常见句式的提取效果明显，对复杂问句提取效果不是特别好。

4. 5. 4 问答效果和分析

在对设计的问答进行测试后，得到表 4-7 所示结果。

(1)无论哪一个函数，精确率、召回率和 F1-Score 的值都取得了不错的实验结果，所以可以判断算法的鲁棒性和泛化能力可以得到保证。

(2)和前三个函数相比，第四个函数的问答效果表现不佳。可能是由于阈值设的较高，加上问题表达的多样化，使得模板的匹配难度加大。在数据量大的情况下，可以加入类似 Word2Vec 的语义信息，以增强对句子含义的理解。

表 4-7 问答效果实验结果

Table 4-7 Experimental results of question answering effect

Cypher 查询函数	精确率	召回率	F1-score
get_robot_releasedate	89%	91%	90%
get_robot_placeoforigin	85%	83%	84%
get_robot_introduction	88%	84%	86%
get_robot_genre	79%	75%	77%
average	85%	83%	84%

从实验结果可以看出，利用本文提出的算法设计的智能问答系统能够有效地回答用户提出的机器人领域的相关问题。该实验证明了本文提出的构建机器人领域智能问答系统的方案的有效性。

4.6 本章小结

本章对于在机器人知识图谱基础上实现的智能问答系统进行了详细的介绍，主要从开发环境、系统架构和实现过程三个部分进行讲解。其中分别从数据层、逻辑层、展示层三个方面详细阐述了智能问答系统的具体实现，并对系统进行了展示。最后对本文设计的算法和问答效果进行了实验，对实验结果进行了分析。

总结与展望

本文首先对智能问答系统的发展背景和研究状况进行了简单概述，紧接着对基于知识图谱的智能问答系统所涉及的理论基础和关键技术进行深入分析和探讨。本文对于智能问答系统的实现，设计了基于朴素贝叶斯问句类型识别算法、基于 TF-IDF 和余弦相似度问句匹配算法和 Cypher 查询语句模板的生成。同时设计实验对本文设计的算法和系统的问答效果进行了测试。以下几点是对本文工作的归纳和总结。

⑴深入研究包括知识图谱、自然语言处理、机器学习等在内的理论和技术，为基于机器人知识图谱的智能问答系统的设计和实现奠定知识基础。

⑵学习并分析基于自然语言处理和机器学习技术的相关算法。本文首先通过相关自然语言处理技术，使用 jieba 工具包对用户输入的问题进行分词、词性标注等处理。接着使用基于朴素贝叶斯的分类算法，对输入问题进行分类。等问题正确归类后，使用 TF-IDF 算法提取出关键词，然后与问题模板进行相似度匹配，这里使用余弦相似度算法。最后生成答案，返回给用户。

⑶使用本文设计的相关算法框架，设计并实现了一个基于机器人知识图谱的智能问答系统，并通过实验对相关算法和问答效果进行测试，试验结果表明，该系统能较好地对用户提出的问题进行回答，证明了本文提出的方法的有效性。

作为综合性的研究领域，智能问答技术涉及多个学科领域，如人工智能、自然语言处理、信息检索等。但现今学术界对于基于知识图谱的智能问答技术并未建立起完善的研究体系，而是仍处于探索阶段。基于本文在该领域的研究和初步探索，仍需在以下方面深入分析并探索。

⑴目前，本文只是对结构化的数据完成抽取工作，为了能够给问答系统和知识推理提供足够的数据，知识库应当不断进行扩大。对知识库进行完善的同时应当将非结构化的文本数据纳入考虑范围，而不是仅考虑结构化数据。

⑵所构建的智能问答模型还不能支持一些复杂的问句，本文的研究关注在单关系的知识图谱问答，今后的研究中可以对多关系的知识图谱问答进行研究。

⑶对基于知识图谱的问答系统进行研究时可结合强化学习的方式。

参考文献

1 毛先领，李晓明. 问答系统研究综述[J]. 计算机科学与探索，2012，6(03)：193-207.

2 D R Radev, H Qi, Z Zheng, et al. Mining the Web for Answers to Natural Language Questions[J]. Acm Cikm, 2001:143-150.

3 D Ravichandran, E Hovy. Learning Surface Text Patterns for a Question Answering System: Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 2002[C].

4 Aliyu F M, Uyar A. Evaluating search features of Google Knowledge Graph and Bing Satori[J]. Online Information Review, 2015, 39(2):197-213.

5 郑立新，王振强. 义务教育阶段机器人模块内容标准解读[J]. 中国电化教育，2012，11：28-30.

6 Angelino E，Larus-Stone N，Alabi D，et al. Learning certifiably optimal rule lists for categorical data[J]. The Journal of Machine Learning Research，2017，18(1)：8753-8830.

7 Xu Kun，Feng Yansong，Zhao Dongyan，et al. Semantic Comprehension of Chinese Natural Language Questions in Knowledge Base[J]. Journal of Peking University (Natural Science), 2014，50(1)：85-92.

8 L S Zettlemoyer，M Collins. Learning to Map Sentences to Logical Form：Structured Classification with Probabilistic Categorial Grammars：Conference on Uncertainty in Artificial Intelligence，2005[C].

9 刘康，张元哲，纪国良等. 基于表示学习的知识库问答研究进展与展望[J]. 自动化学报，2016 42(06)：807-818.

10 Voorhees，E. M. (2001). The trec question answering track. Natural Language Engineering，7(4)，361-378.

11 Voorhees EM，Buckland L. Overview of the TREC 2003 Question Answering Track. InTREC 2003 Nov 19 (Vol. 2003, pp. 54-68).

12 Hovy EH, Gerber L, Hermjakob U, Junk M, Lin CY Question Answering in Webclopedia. InTREC 2000 Nov 13 (Vol.52,pp.53-56).

13 Kwok C, Etzioni O, Weld DS. Scaling question answering to the web. ACM Transactions on Information Systems (TOIS). 2001 Jul 1;19(3):42-62.

14 T Kwiatkowski, L Zettlemoyer, S Goldwater, et al. Inducing Probabilistic Ccg Grammars From Logical Form with Higher-Order Unification: Conference on Empirical Methods in Natural Language Processing, 2010[C].

15 X Yao, B V Durme. Information Extraction Over Structured Data: Question Answering with Freebase: Meeting of the Association for Computational Linguistics, 2014[C].

16 W T Yih, X He, C Meek. Semantic Parsing for Single-Relation Question Answering: Meeting of the Association for Computational Linguistics, 2014[C].

17 Y Zhang, K Liu, S He, et al. Question Answering Over Knowledge Base with Neural Attention Combining Global Knowledge Information[J]. 2016.

18 Z Xie, Z Zeng, G Zhou, et al. Knowledge Base Question Answering Based On Deep Learning Models[J]. 2016:300-311.

19 F Yang, L Gan, A Li, et al. Combining Deep Learning with Information Retrieval for Question Answering: International Conference on Computer Processing of Oriental Languages, 2016[C].

20 徐增林，盛泳潘，贺丽荣，等. 知识图谱技术综述[J]. 电子科技大学学报，2016, 45(4):589-606.

21 胡芳槐. 基于多种数据源的中文知识图谱构建方法研究[D]. 华东理工大学，2015.

22 刘娇，李杨，段宏等. 知识图谱构建技术综述[J]. 计算机研究与发展，2016, 53(03):582-600.

23 Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z. Dbpedia: A nucleus for a web of open data. The semantic web. 2007:722-35.

24 Suchanek FM, Kasneci G, Weikum G. Yago: a core of semantic knowledge. InProceedings of the 16th international conference on World Wide Web 2007 May 8 (pp. 697-706). ACM.

25 Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J. Freebase: a collaboratively created graph database for structuring human knowledge. InProceedings of the 2008 ACM SIGMGD international conference on Management of data 2008 Jun 9 (pp. 1247-1250). ACM.

26 G Lample, M Ballesteros, S Subramanian, et al. Neural Architectures for Named Entity Recognition[J]. 2016:260-270.

27 Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. In Advances in neural information processing systems 2013 (pp.

2787-2795).

28 李一夫. 基于生成对抗学习的知识图谱问答系统研究[D]. 浙江大学，2018.

29 单良，刘欣. 基于中国历史人物知识的智能问答系统构建[J]. 情报探索，2019(6)：101-105.

30 张淼. 基于中文知识图谱的智能问答系统设计与实现[D]. 华中师范大学, 2018.

31 Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality In Advances in neural information processing systems 2013 (pp. 3111-3119).

32 Bordes, A., Chopra, S., & Weston, J. (2014). Question Answering with Subgraph Embeddings. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 615-620).

33 Dong, L., Wei, F., Zhou, M., & Xu, K. (2015). Question Answering over Freebase with Multi-Column Convolutional Neural Networks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1：Long Papers) (pp. 260-269).

34 Lai, W S., Huang, J. B., & Yang, M. H. (2017). Semi-Supervised Learning for Optical Flow with Generative Adversarial Networks. In Advances in Neural Information Processing Systems (pp. 353-363).

35 王文辉，吴敏华，骆力明等. 基于相似度算法的英语智能问答系统设计与实现[J]. 计算机应用与软件，2017，34(06)：62-68.

36 刘济源.旅游领域知识图谱的构建及应用研究[D]. 浙江大学，2019.

37 钱强，庞林斌，高尚.一种基于词共现图的受限领域自动问答系统[J].计算机应用研究，2013, 30(3):841-843.

38 李思珍.基于本体的行业知识图谱构建技术的研究与实现[D].北京邮电大学，2019.

39 陶杰. 住房公积金领域自动问答系统关键技术研究[D]. 哈尔滨工程大学，2017.

40 杜泽宇，杨燕，贺裸，等. 基于中文知识图谱的电商领域问答系统[J]. 计算机应用与软件，2017(5)：153-159.

41 钱宏泽. 基于中草药语义网的自动问答系统的研究与实现[D]. 浙江大学，2016.

42 Y Lai, Y Lin, J Chen, et al. Open Domain Question Answering System Based On Knowledge Base[M]. Springer International Publishing, 2016. 722-733.

43 张克亮，李伟刚，王慧兰. 基于本体的航空领域问答系统[J]. 中文信息学报，2015, 29(4)：192-198.

44 刘建炜，燕路峰. 知识表示方法比较[J]. 计算机应用系统, 2011, 20(03):242-246.

45 Nadeau D, Sekine S. A survey of named entity recognition and classification[J]. Lingvisticae Investigationes, 2007, 30(1):3-26.

46 Lafferty J, McCallum A, Pereira F C N. Conditional random fields：Probabilistic models for segmenting and labeling sequence data[J]. 2001.

47 Zhang M L, Zhou Z H. ML-KNN: A lazy learning approach to multi-label learning[J]. Pattern recognition, 2007, 40(7):2038-2048.

48 王俊华，左万利，闫昭. 基于朴素贝叶斯模型的单词语义相似度度量[J]. 计算机研究与发展，2015(7)：1499-1509.

49 赵胜辉，李吉月，徐碧路，等.基于 TFIDF 的社区问答系统问句相似度改进算法[J]. 北京理工大学学报，2017, 37(9):982-985.

50 ZHANG H. The Optimality of Naïve Bayes[C]. Florida: Proceedings of the 17th International FLAIRS Conference, 2004.

攻读硕士学位期间发表的学术论文

1　黄春梅, 王松磊. 基于词袋模型和 TF-IDF 的短文本分类研究[J]. 软件工程. 2020(3).

原创性声明、使用授权书

哈尔滨师范大学学位论文独创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文题目：基于机器人知识图谱的智能问答系统的设计与实现

学位论文作者签名：王松磊　　日期：2020 年 5 月 30 日

哈尔滨师范大学学位论文授权使用声明

本人完全了解并遵守哈尔滨师范大学有关保留、使用学位论文的规定，即：学校有权保留学位论文并向国家主管部门或其指定机构送交论文的电子版和纸质版。有权将学位论文的标题和摘要汇编出版。有权将学位论文提交《中国学术期刊（光盘版）》电子杂志社在《中国优秀硕士学位论文全文数据库》和《中国博士学位论文全文数据库》中发表，可以采用影印、缩印或扫描等复制手段保存学位论文。保密的学位论文在解密后适用本规定。

作者签名：王松磊　　日期：2020 年 5 月 30 日

导师签名：黄春梅　　日期：2020 年 5 月 30 日

致谢

光阴如梭，两年的研究生生涯转眼即将结束，回想起这两年来的点点滴滴，往事涌上心头。在此，对曾经帮助我、关心我的所有人表示最衷心的谢意。

在研究生期间，深深受益于许多老师的关心、爱护和淳淳教导。首先要感谢指导教师姜春茂教授，姜老师学识渊博、治学严谨，实践经验丰富，在我攻读硕士学位期间，姜老师在学习和科研方面给予了我耐心的培养与教育，并带着我参加了很多实验室项目的开发与实践，为学生提供了一个很优秀的平台。同时还要感谢黄春梅副教授，黄老师在我这两年的研究生生涯中对我的学习和工作都提供了非常中肯的建议。在遇到困难时，她的建议总会给我提供新的思路。在我迷茫时，她的指导也会给我非常大的鼓励。在此谨向姜老师和黄老师表示我最诚挚的敬意和感谢。

感谢一直关心与支持我的同学和朋友们，感谢你们的鼓励和帮助。两年来，我们朝夕相处，共同进步，感谢你们给予我的所有关心和帮助。同窗之宜，永生难忘。还要感谢父母和亲人，正是因为他们的一如既往的支持才让我可以一直孜孜不倦的学习和提高。

最后要感谢哈尔滨师范大学，哈尔滨师范大学为我提供了非常好的学习平台，不仅可以得到来自老师的指导，也有同学带给我各种帮助。正是由于老师、同学的指导和帮助，我才能克服困难走到最后。