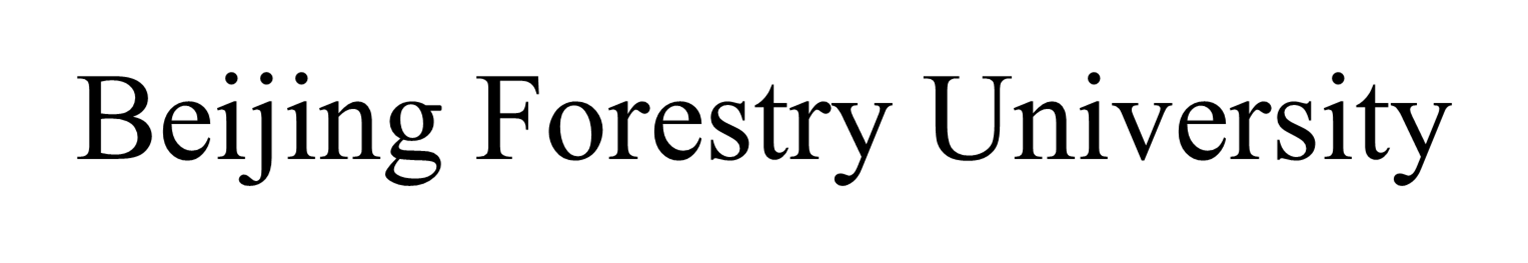
学校代码：10022





**本科毕业论文(设计)**

|  |
| --- |
| **基于知识图谱的新冠肺炎问答系统** |

|  |
| --- |
| **COVID-19 Question Answering System Based on Knowledge Graph** |

|  |
| --- |
| 尹朦朦 |

|  |  |
| --- | --- |
| **学 院** | 信息学院 |
| **专 业** | 计算机科学与技术 |
| **指导教师** | 李冬梅 教授 |
|  | 张小平 正高级工程师 |

2021 年 6 月 1 日

**独创性声明**

本人声明所呈交的论文（设计）是本人在导师指导下独立进行的设计、研究工作及取得的设计、研究成果。尽我所知，除了论文（设计）中特别加以标注和致谢的地方外，论文（设计）中不包含其他人已经发表或撰写过的研究成果，本论文（设计）中没有抄袭他人研究成果和伪造数据等行为。与我共同工作的人员对本研究所做的任何贡献均已在论文（设计）中作了明确的说明并表示了谢意。

作者签名： 日期： 年 月 日

**关于毕业论文（设计）使用授权的说明**

本人完全了解北京林业大学有关保留、使用毕业论文（设计）的规定，即：本科生在校期间毕业论文（设计）工作的知识产权单位属北京林业大学；学校有权保留并向国家有关部门或机构送交论文（设计）的纸质版和电子版，允许毕业论文（设计）被查阅、借阅和复印；学校可以将毕业论文（设计）的全部或部分内容公开或编入有关数据库进行检索，可以允许采用影印、缩印或其它复制手段保存、汇编毕业论文（设计）。

**（保密的论文在解密后应适用本授权书）**

作者签名： 指导老师签名：

日 期： 年 月 日

**摘要**

2019年底出现的新冠肺炎是冠状病毒第三次在人类中暴发，面对新冠肺炎，人们需要得到正确且清晰的信息。传统的信息检索方式仅返回相关网页集合，用户需要在冗余繁杂的信息中找到有用的信息，这样的信息获取方式效率低下，不能很好地服务用户。面对这种情况，研究针对此次新冠肺炎的问答系统满足用户信息获取的需求，显得很有必要。

本文研究并搭建了基于知识图谱的新冠肺炎问答系统，问答功能的实现采用加入朴素贝叶斯算法的模板匹配方式。对于用户输入的问题，首先进行实体识别，使用词性标注结合实体相似匹配的方式识别用户问句中的实体。接着对用户的问句意图进行预测，采用训练好的问句分类器预测类别编号，最后使用cypher查询图数据库，将返回结果输出。系统使用Flask框架搭建，除了实现问答功能，还提供关系检索和后台管理功能。其中，关系检索帮助用户清晰直观地了解实体间的关系，后台管理功能帮助管理员更新图谱数据。

本文实现的系统可以帮助用户快速获取想要查找的信息，提高用户的信息获取效率，减少用户分辨或整合信息的时间。该系统实现为包括研究人员，医务人员和公众在内的人们提供新冠肺炎医疗、健康、物资、防控、科研和人物等领域的信息，助力新冠肺炎的研究与预防。

**关键词**：知识图谱，新冠肺炎，问答系统，朴素贝叶斯

**Abstract**

The emergence of COVID-19 at the end of 2019 is the third outbreak of coronavirus in humans, and people need to get correct and clear information about COVID-19. The traditional information retrieval method only returns the collection of relevant web pages, and users need to find useful information in the redundant and complicated information. Such information retrieval method is inefficient and cannot serve users well. In the face of this situation, it is necessary to study the question answering system of COVID-19 to meet users' demand for information acquisition.

This paper studies and builds a COVID-19 question answering system based on knowledge graph. The realization of question answering function adopts the template matching method with Naive Bayes algorithm. For user input questions, entity identification is carried out first. Entity identification uses part of speech tagging combined with entity similarity matching to identify entities in user questions. Then, the user's intention of the question is predicted. Intention prediction is to use the trained question classifier to predict the category number. Finally, cypher is used to query the graph database and return the output result. The system uses Flask framework to build. The system not only realizes the function of question answering, but also provides the function of relationship retrieval and background management. Among them, the relationship retrieval helps users to clearly and intuitively understand the relationship between entities, and the background management function helps administrators to update the graph data.

The system realized in this paper helps users quickly obtain the information they want to find, it improves the efficiency of information acquisition and reduces the time for users to distinguish or integrate information. The system provides researchers, medical workers and the general public with information on COVID-19 medical, health, goods, prevention, research, and character to facilitate research and prevention of COVID-19.

**Keywords:** Knowledge graph, COVID-19, question answering system, Naive Bayes

# 目录

[目录 III](#_Toc73902165)

[1 绪论 1](#_Toc73902166)

[1.1 研究背景及意义 1](#_Toc73902167)

[1.2 国内外研究现状 1](#_Toc73902168)

[1.2.1 知识图谱研究现状 1](#_Toc73902169)

[1.2.2 问答系统研究现状 2](#_Toc73902170)

[1.3 本文主要研究内容 3](#_Toc73902171)

[1.4 本文组织结构 3](#_Toc73902172)

[2 相关理论与技术 5](#_Toc73902173)

[2.1 知识图谱 5](#_Toc73902174)

[2.1.1 知识图谱概述 5](#_Toc73902175)

[2.1.2 知识图谱关键技术 5](#_Toc73902176)

[2.2 图数据库 6](#_Toc73902177)

[2.2.1 图数据库概述 6](#_Toc73902178)

[2.2.2 Neo4j图数据库 6](#_Toc73902179)

[2.3 系统实现相关技术 7](#_Toc73902180)

[2.3.1 前端技术 7](#_Toc73902181)

[2.3.2 后端技术 8](#_Toc73902182)

[3 问答系统算法实现 10](#_Toc73902183)

[3.1 算法概述 10](#_Toc73902184)

[3.2 实体识别 10](#_Toc73902185)

[3.2.1 词性标注 10](#_Toc73902186)

[3.2.2 实体相似匹配 11](#_Toc73902187)

[3.3 意图分类 13](#_Toc73902188)

[3.3.1 TF-IDF特征提取 13](#_Toc73902189)

[3.3.2 朴素贝叶斯算法 15](#_Toc73902190)

[3.3.3 构建问句分类器 16](#_Toc73902191)

[3.4 cypher语句生成 17](#_Toc73902192)

[3.5 本章小结 18](#_Toc73902193)

[4 系统设计与实现 19](#_Toc73902194)

[4.1 系统需求 19](#_Toc73902195)

[4.1.1 功能需求 19](#_Toc73902196)

[4.1.2 性能需求 19](#_Toc73902197)

[4.2 系统设计 20](#_Toc73902198)

[4.2.1 架构设计 20](#_Toc73902199)

[4.2.2 功能设计 20](#_Toc73902200)

[4.3 系统实现 21](#_Toc73902201)

[4.3.1 知识图谱介绍 21](#_Toc73902202)

[4.3.2 关系检索 22](#_Toc73902203)

[4.3.3 问答 23](#_Toc73902204)

[4.3.4 后台管理 24](#_Toc73902205)

[4.4 本章小结 27](#_Toc73902206)

[5 工作总结与展望 28](#_Toc73902207)

[5.1 工作总结 28](#_Toc73902208)

[5.2 工作展望 28](#_Toc73902209)

[致谢 29](#_Toc73902210)

[参考文献 30](#_Toc73902211)

[附录A 问题模板 32](#_Toc73902212)

# 绪论

## 研究背景及意义

2019年12月，湖北省武汉市开始出现多起不明原因引发的肺炎病例，后证实是由一种新型冠状病毒（2019-novel coronavirus,2019-nCoV）感染所致，以下简称新冠肺炎[1,2]。2020年2月11日，世界卫生组织将新冠肺炎命名为“COVID-19”[3]。由于新冠肺炎的高传染性，现在已成为全球流行性的严重疾病。面对近百年来人类遭遇的最大公共卫生事件[4]，关于新冠肺炎，人们迫切需要得到全面、清晰、正确的信息。传统的搜索引擎根据用户输入的关键词进行检索，返回相关网页集合，用户往往需要自己辨别信息真伪，还会面对信息分散的问题，这样会造成用户的信息获取效率不高甚至获得错误信息。因此，用户更多的需要另一种信息查询服务方式——问答系统。

问答系统是自然语言处理的重要研究内容，具体是指计算机通过对问句进行深层次的语义分析，获取问句内涵的信息，然后返回最接近用户需要的答案，返回的答案越精准越能够体现计算机的智能性[5]。知识图谱本质上是表示实体之间关系的语义网络，作为大规模信息的载体，知识图谱内的三元组数据能够为问答系统提供大量的事实数据及实体间的复杂关系，因此，基于知识图谱的问答系统越来越成为人们关注与研究的重点。问答系统嵌入知识图谱后，在数据精度、数据关联性、数据结构化水平等方面得到了显著提升，可以处理更多更复杂的自然语言问题，为用户提供更加精准有效的信息，使计算机的工作变得更加人性化。

目前，基于知识图谱的新冠肺炎问答系统的研究是大势所趋。问答系统基于知识图谱对用户输入的自然语言问题进行分析理解，在知识图谱中搜寻对应的关系与数据，为用户提供答案，这样的问答系统能够节省用户在繁杂网页中搜索花费的时间，提高信息的获取效率，助力新冠肺炎的研究与预防。

## 国内外研究现状

### 知识图谱研究现状

知识图谱并非全新的概念，2006年，BERNERS-LEE等人提出了语义网[6]的概念，呼吁使用本体模型形式化表达数据中的隐含语义，由此产生了RDF（resource description framework）模式和万维网本体语言OWL（Web ontology language）的形式化模型。知识图谱的产生就是基于以上研究。知识图谱这一概念最早在2012年被谷歌正式提出[7]，其初衷是为了提高搜索引擎的搜索能力，增强搜索质量以及提升用户体验。目前，随着智能信息服务应用的不断发展，知识图谱已被广泛应用于智能搜索、智能问答、个性化推荐等领域。

知识图谱分为通用知识图谱和垂直领域知识图谱[8]。通用知识图谱强调的是广度，围绕通用知识图谱的项目有语言学类的 Word Net[9]、百科类的DBpedia[10]、中文通用百科知识图谱（CN-DBpedia）[11]、zhishi.me[12]等，这些知识图谱都具有较高的领域覆盖面。垂直领域知识图谱强调的是深度，描述范围是特定的行业领域，需要依靠特定的行业数据构建。垂直领域知识图谱包括阿里巴巴构建的淘宝知识图谱、地理信息领域的知识图谱 Geonames[13]、企业领域知识图谱“天眼查”等。

### 问答系统研究现状

问答系统最早可以追溯至人工智能初期，阿兰·图灵于1950年提出可以通过观察机器是否具备正确回答问题的能力来验证机器是否具有智能[14]。1966年，麻省理工学院Weizenbaum设计了名为ELIZA的聊天机器人[15]，实现了与人的简单交流。2011年，IBM公司推出了轰动一时的超级计算机Watson，Watson在美国知识竞赛类电视节目中上演了“人机大战”并最终战胜了人类选手，这被视为人工智能发展的里程碑。与国外研究相比，国内问答系统研究起步较晚。目前，国内较为活跃的研究机构有：中国科学院计算所、哈尔滨工业大学、清华大学等。其中，中国科学院计算所研发的NKI系统能够对用户的自然语言问题给出较好的回答。

从涉及的应用领域进行分类，问答系统可以分为限定域问答系统和开放域问答系统。限定域问答系统能处理的问题只限定于某个领域，比如只限定于医学、化学或者某企业的业务领域等。例如BASEBALL、LUNAR、SHRDLU等都属于限定域的问答系统。BASEBALL只能回答关于棒球比赛的问题，LUNAR只能回答关于月球岩石的化学数据的问题，SHRDLU只能回答关于积木移动的问题等。开放域问答系统不同于限定域问答系统，这类系统可回答的问题不限定于某个特定领域。

基于知识图谱的问答系统的构建方法有许多。TrueKnowledge[16]模板问答方法首先利用模板成分匹配问句中的内容，然后按照顺序匹配相应模板，同时，一个模板可以覆盖多个问题。例如问句“What is the capital of America”，首先匹配疑问词“what”，然后根据其反应的问句意图生成“what a y”，其中变量a与y映射问句中的内容，确定好待填充结构后，将“is the capital of”映射到变量a，“America”映射到变量y。Nilesh等人[17]概述了基于神经网络的问答方法，将现有方法分为三类，第一类是基于分类的方法，给定一个问题，使用神经网络模型预测它属于一组固定的类别中的哪一个；第二类是基于排序的方法，使用神经网络比较不同的候选逻辑形式与问题，选择排序最高的逻辑形式；第三类是基于翻译的方法，使用网络学习将自然语言问题转化为对应的可执行逻辑形式。知识图谱的嵌入学习方法[18]通过将知识图谱内的每个对象编码到连续空间，反映关系密切的实体及关系的相似程度，为问答的实现做了良好的铺垫。Huang等人[19]设计了一种基于知识图谱嵌入的问答系统，结合知识图谱嵌入学习在问答系统中的优点，解决谓词在问题中的不同表达和实体被识别后的消歧等问题。

## 本文主要研究内容

本文研究内容分为三部分，本文基于已有的新冠肺炎知识图谱进行研究，设计并实现基于知识图谱的新冠肺炎问答系统。问答系统使用模板匹配的方法，将用户输入的自然语言问句转化为图数据库中的关系查询，最终将查询结果呈现给用户。本文工作的第一部分是实现问句的实体识别，第二部分是对问句进行意图预测，得到对应的模板编号，第三部分是基于算法搭建问答系统。

（1）实体识别

本文使用jieba分词加载用户字典对问句进行实体识别，对于不能正确识别的实体，本文设置了相似匹配，找出与分割词组最接近的特征词作为实体识别结果。

（2）意图预测

首先使用TF-IDF对训练数据进行特征提取和向量转化，将向量化后的数据作为输入，运用朴素贝叶斯算法进行模型训练，生成问句分类器，对于用户输入的问题能够识别出对应的模板编号。根据实体识别结果和问句分类器输出的模板编号，进行模板匹配。

（3）搭建问答系统

本文搭建了基于知识图谱的新冠肺炎问答系统，方便用户操作。系统提供的功能除了新冠肺炎问答外，用户还可以对实体进行关系检索，设置后台管理功能方便管理员更新图谱数据。

## 本文组织结构

本文共分为5章。

第1章为绪论，介绍了基于知识图谱的新冠肺炎问答系统的研究背景及意义，概述了知识图谱和问答系统的研究现状，并叙述了本文的主要研究内容。

第2章介绍了相关理论与技术，描述了知识图谱、图数据库的基本内容，介绍了Neo4j图数据库，说明了系统前后端所用技术。

第3章介绍了本文采用的算法，详细叙述了从实体识别到意图预测，最后cypher查询图数据库将结果输出的实现过程。

第4章叙述了基于知识图谱的新冠肺炎问答系统的实现，详细介绍了系统需求、系统设计以及系统主要功能的实现。

第5章是对本文工作的总结与展望。

# 相关理论与技术

## 知识图谱

### 知识图谱概述

知识图谱是结构化的语义知识库，用于以符号形式描述现实世界中的概念及其相互关系[20]。知识图谱的基本组成单位是“实体-关系-实体” 三元组，知识图谱中的实体具有属性，实体间通过关系相互链接，构成网状的知识结构。从图的角度看，知识图谱中的实体对应节点，实体间的语义关系对应边，因此，知识图谱本质上是表示实体之间关系的语义网络。

知识图谱的架构包括逻辑结构和体系架构。知识图谱的逻辑架构是指，知识图谱从逻辑上可以划分为两个层次：数据层和模式层，其中模式层是知识图谱的核心。知识图谱的体系架构也叫技术架构，具体是指是知识图谱的构建过程[21]，从原始数据中对知识进行提取，将结果存入知识图谱，此过程不断迭代更新，知识图谱也不断更新完善。

### 知识图谱关键技术

根据知识获取的逻辑，知识图谱构建过程可以分为三个阶段：知识抽取、知识融合以及知识加工[20]。

（1）知识抽取

知识抽取是知识图谱构建的第一步，知识抽取是一种自动化地从半结构化或无结构化数据中抽取信息的技术，抽取的信息包括实体、关系以及实体属性等结构化信息，知识抽取涉及的关键技术包括：实体抽取、关系抽取及属性抽取。知识抽取的难点是如何从异构数据源中自动抽取得到候选知识单元。

（2）知识融合

通过知识抽取得到的信息可能存在冗余或数据错误等问题，数据间缺乏层次性和逻辑性，因此需要对数据进行清理和整合，消除概念歧义，剔除冗余和错误概念，确保知识的质量，这就是知识融合。知识融合包括实体链接和知识合并两部分，实体链接是指将从文本中抽取到的实体正确链接到知识图谱中对应的实体上的操作；知识合并主要是解决实体、属性及关系的冲突问题，去除重复的冗余数据，将处理后得到的数据要素融合到本地知识库中，实现数据的整合。

（3）知识加工

经过知识抽取与知识融合，可以得到一系列事实表达，但事实并不等于知识，要想得到结构化、网络化的知识体系，还需要进行知识加工。知识加工的技术有很多，其中最重要的是知识推理。知识推理是指从知识图谱中已有的实体关系数据出发，进行计算推理，从而建立实体间的新关联，达到扩展知识图谱的目的。通过知识推理能够从现有的知识中发现新的知识，知识推理是知识图谱构建的关键技术和重要手段。

## 图数据库

### 图数据库概述

随着互联网技术特别是移动互联网的发展，越来越多的领域如电商、社交、金融产生了海量的数据，这些数据之间存在着各种各样的关系，传统的关系数据库已经无法满足存储需求。以社交网络为例，采用关系数据库将导致数据冗余，并且不能适应社交数据的动态性，也不能很好地支持类似“好友的好友”这样的多层复杂查询。针对数据间内在关系复杂且动态变化的问题，人们将目光转向图数据库。

图数据库是NoSQL数据库的一种，是一种使用图论来存储、映射和查询关系的NoSQL数据库。与关系数据库相比，图数据库存储信息的方式更为灵活，能够存储结构化和半结构化的数据，在前期可以省去一些数据处理的工作；图数据库易于扩展，更适合存储管理动态变化的数据；图数据库能够存储图结构的数据，更适合知识图谱的存储；图数据库能够在多元且深度关联的数据上进行快速的查询分析，在2度查询和3度查询上十分高效，极大地提高了查询效率；图数据库对于关系的表示方法更为直观，用户体验好。

### Neo4j图数据库

常见的图数据库有Neo4j，FlockDB，AllegroGrap，GraphDB，InfiniteGraph，HugeGraph等，其中，Neo4j是目前图数据库技术中最为常用的一种，功能最完善、发展最快。Neo4j的主要优点有：

（1）采用图存储结构，将数据存储为节点与关系两种类型，数据存储量大，支持更快速的事务处理能力与数据查询能力；

（2）具有高灵活性，在保留原有数据的情况下通过迭代更新对数据库进行延伸与扩展，节省开发时间；

（3）具有完整的ACID支持，保证事务的正确可靠；

（4）数据库本身提供完善的图查询语言cypher，支持各种图挖掘算法。

Neo4j的安装与使用十分简单，在官方网站上下载好Neo4j，开启服务后，在浏览器上输入“localhost:7474”即可进入Neo4j访问界面，如图 2.1所示，进入之前要输入用户名和密码进行验证。

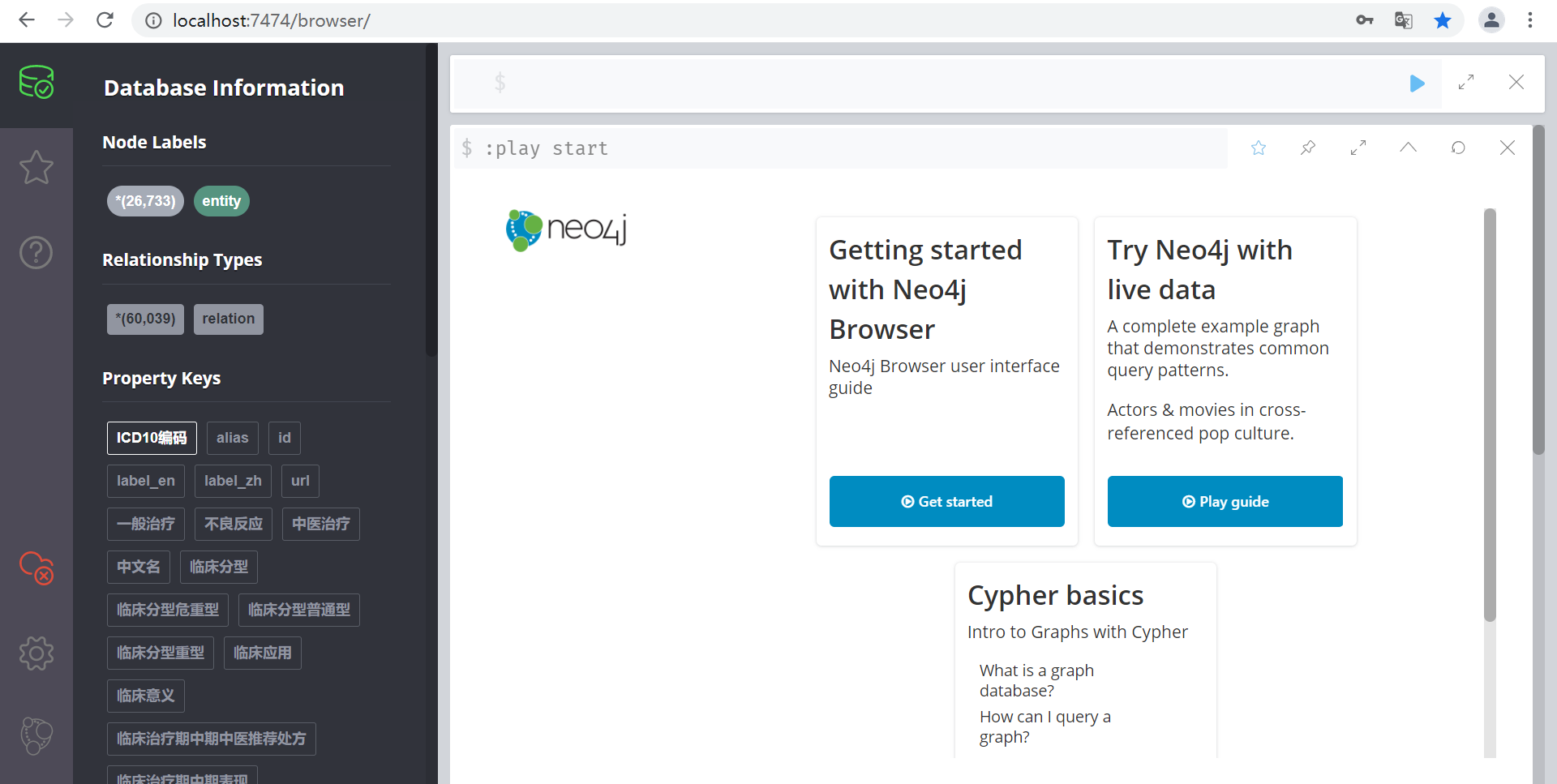


图 2.1 Neo4j使用页面

Fig. 2.1 Usage page of Neo4j

## 系统实现相关技术

本文以Python语言为基础，使用Flask框架实现基于知识图谱的新冠肺炎问答系统，系统前端使用Bootstrap框架，使用Echarts实现数据可视化，下面从分别前端和后端介绍系统使用的技术。

### 前端技术

（1）Bootstrap

Bootstrap是由美国Twitter公司的设计师Mark Otto和Jacob Thornton合作开发的一款基于HTML、CSS、JavaScript的前端开发框架。Bootstrap一经推出就备受广大开发者的青睐，一直是GitHub上的热门开源项目。Bootstrap包含了丰富Web组件，根据这些组件，可以快速地搭建一个漂亮、功能齐全的网站。

（2）Echarts

Echarts是一款基于JavaScript的数据可视化图表库，最初由百度团队开源，于2018年捐赠给Apache基金会。Echarts缩写来自Enterprise Charts（商业级数据图表），Echarts可以流畅的运行在PC和移动设备上，兼容IE、Chrome、Firefox、Safari等绝大部分浏览器。Echarts底层依赖轻量级的Canvas类库ZRender，提供直观、生动、可交互、可高度个性化定制的数据可视化图表，其创新的拖拽重计算、数据视图、值域漫游等特性大大提升了用户体验，赋予了用户对数据进行挖掘、整合的能力。图 2.2展示了部分Echarts的示例。

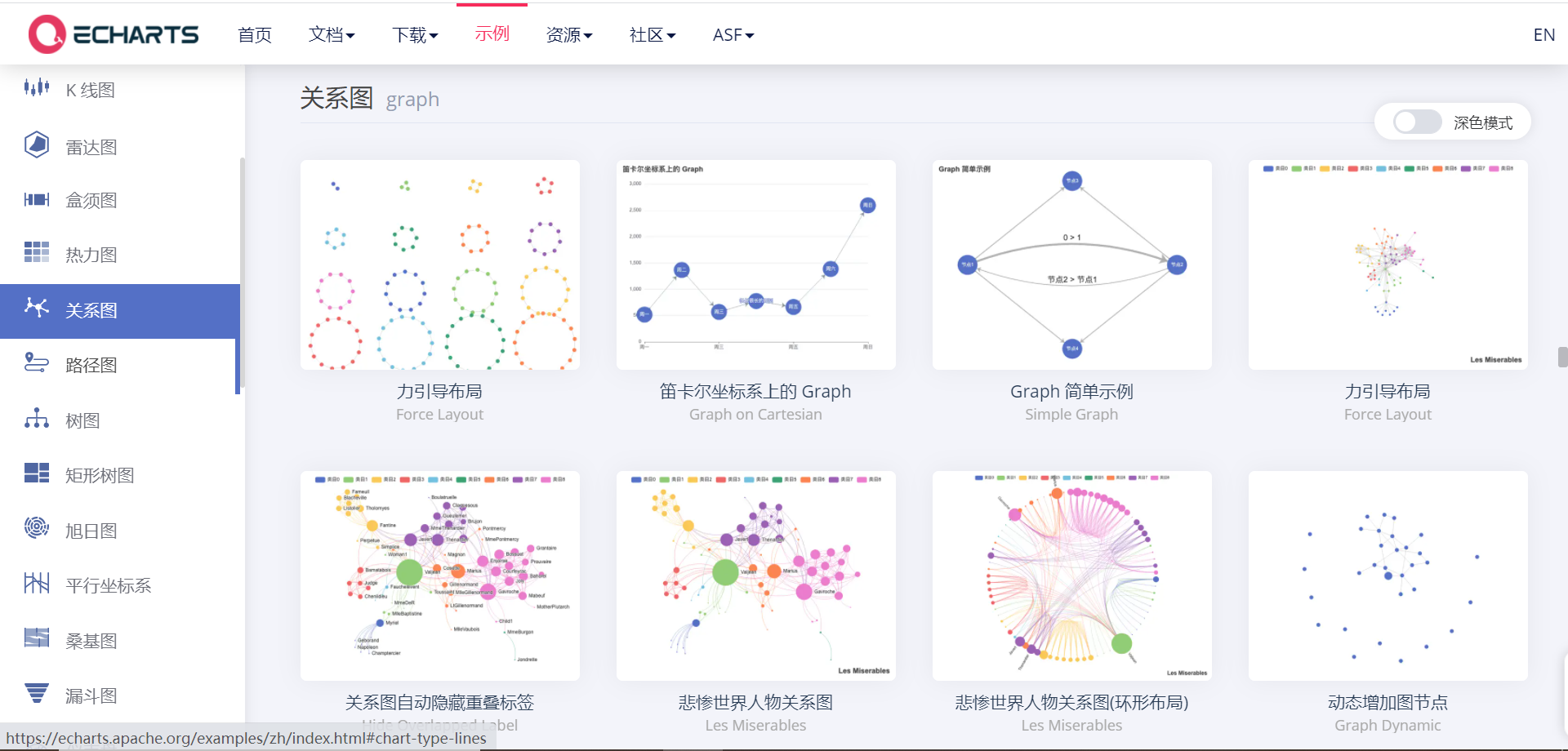


图 2.2 Echarts示例

Fig. 2.2 Echarts examples

### 后端技术

Flask是目前十分流行的Web框架，使用Python语言编写，它被称为微框架（microframework），“微”并不表示整个应用都塞在一个Python文件中，也不代表Flask功能不强，微框架中的“微”表示Flask旨在保持代码的简介和易于扩展。Flask相比其他同类框架更为灵活、安全、轻便且容易上手，核心构成比较简单，具有强的扩展性和兼容性，适合小型网站的开发。此外，Flask框架具有很强的定制性，用户可以根据需求来添加相应的功能，其强大的插件库支持用户实现个性化的网站定制。

Flask主要包括Werkzeug和Jinja2两个核心函数库，它们分别负责业务处理和安全方面的功能。Werkzeug库的功能十分强大，支持URL路由请求集成，一次可以响应多个用户的访问请求；支持Cookie和会话管理，通过身份缓存数据建立长久连接关系，提高用户访问速度；可以处理HTTP基本事务，快速响应客户端推送过来的访问请求。Jinja2库支持自动HTML转移功能，能够很好控制外部黑客的脚本攻击。

Flask的基本模式是在程序里将一个视图函数分配给一个URL，每当用户访问这个URL时，系统就会执行给该URL分配好的视图函数，获取函数的返回值并将其显示到浏览器上，其工作过程见图 2.3。



图 2.3 Flask框架工作过程图

Fig. 2.3 Working process diagram of Flask framework

# 问答系统算法实现

## 算法概述

为实现基于知识图谱的新冠肺炎问答系统，本文采用加入朴素贝叶斯的模板匹配方法，算法流程为实体识别、意图预测和cypher语句生成三部分。对于用户输入的问题，首先进行实体识别，本文采用词性标注结合实体相似匹配的方式进行实体识别。采用朴素贝叶斯算法进行意图预测，首先构建训练数据，使用TF-IDF对训练数据进行特征提取和向量转化，对向量化后的数据使用多项朴素贝叶斯进行模型训练，得到问句分类器，对输入的问题进行意图预测。最后为每类模板编写对应的cypher语句查询Neo4j图数据库，将查询结果输出。算法流程如图 3.1所示。

**图 3.1 算法流程图**

Fig. 3.1 Algorithm flow chart

## 实体识别

实现问答系统必不可少的一步是进行实体识别，本文采用词性标注结合实体相似匹配的方式进行实体识别，下面具体介绍实体识别的实现。

### 词性标注

词性标注也叫做语法标注，是指将语料库中单词的词性按其含义和上下文内容进行标记的文本处理技术。本文采用人工标注结合jieba的词性标注功能对单词进行标注，人工标注了disease，drug，symptom，people，virus和way等六类词性，词性的含义解释见表 3.1。

表 3.1 词性含义解释

Table 3.1 Explanation of part of speech meaning

|  |  |  |
| --- | --- | --- |
| 词性 | 含义 | 举例 |
| disease | 疾病 | 新型冠状病毒肺炎 |
| drug | 药物 | 利托那韦 |
| symptom | 症状 | 咳嗽 |
| people | 人物 | 钟南山 |
| virus | 病毒 | 禽流感病毒 |
| way | 预防措施对象 | 学生 |

人工标注具体工作：首先将可能用到的单词从Neo4j图数据库中导出，以disease为例，输入cypher语句查询实体名称，查询结果如图 3.2所示，选择csv格式导出，对导出的文件进行人工标注。将所有标记结果放入一个文件构成用户字典。

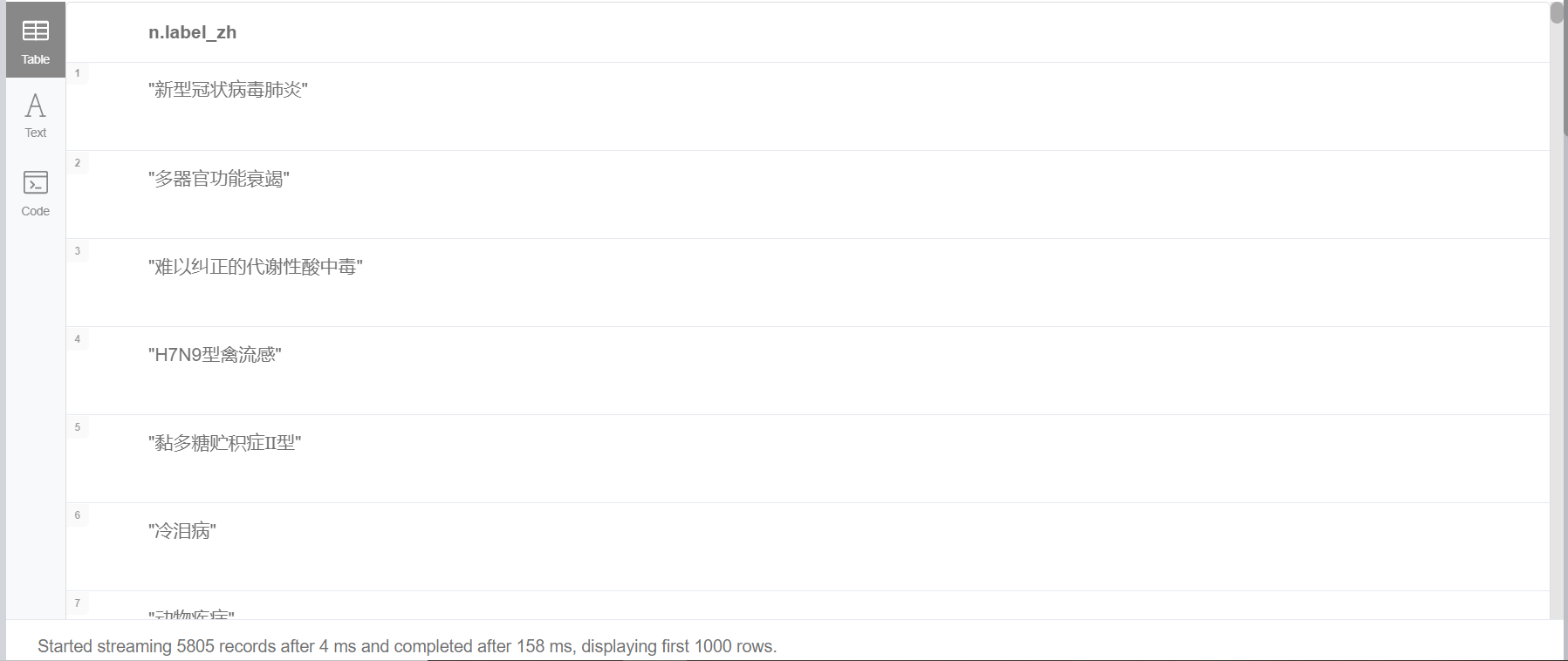


图 3.2 cypher查询结果

Fig. 3.2 Cypher query results

jieba加载用户字典后能提高分词的准确性并对其进行正确的词性标注，例如输入语句“新型冠状病毒肺炎的临床表现”，加载用户字典前的分词及词性标注结果为['新型/b', '冠状病毒/l', '肺炎/n', '的/uj', '临床表现/i']，加载用户字典后的结果为['新型冠状病毒肺炎/disease', '的/uj', '临床表现/i']，显然，后者更符合用户问句的真实意图。

根据disease词性获取“新型冠状病毒肺炎”实体。

### 实体相似匹配

在词性标注未能获取到实体的情况下对实体进行相似匹配，通过求解输入语句词组与实体特征词之间的字重叠率和编辑距离两个指标的平均值得到与实体特征词的相似度[22]，选取相似度最大的实体特征词作为输入语句的实体识别结果，具体流程如图 3.3所示。



图 3.3 相似匹配流程图

Fig. 3.3 Flow chart of similarity matching

实体特征词库由Neo4j导出：首先使用cypher语句查询相关实体名称，下载csv文件，将所有查询结果放入一个文件，得到实体特征词库。对于输入语句分割词和实体特征词库特征词，首先计算分割词与特征词的字重叠率（式3.1），得到第一个结果。

（3.1）

式（3.1）中*num*指两个单词的重复字数，*c*指两个单词的不同字符的总个数。

第二个结果通过计算编辑距离得到。编辑距离（Edit Distance），又称Levenshtein距离，是指两个字串之间，由一个转换成另一个所需的最少编辑操作次数。编辑操作包括将一个字符替换成另一个字符，插入一个字符，删除一个字符。例如计算两个字符串“kitten”和“sitting”的编辑距离时，首先将“kitten”中的“k”替换为“s”，再将替换后的“sitten”中的“e”替换为“i”，最后添加“g”得到“sitting”。这个过程中一共发生了两次替换操作和一次添加操作，所以“kitten”和“sitting”的编辑距离是3。一般来说，编辑距离越小，两个串的相似度越大。

用表示字符串*a*，*b*的编辑距离，其中和分别表示字符串*a*，*b*的长度，的计算方法见式（3.2）。

（3.2） 表示*a*的前*i*个字符和*b*的前*j*个字符之间的距离，表示当前遍历到字符串*a*的位置*i*，字符串*b*的位置*j*。为分段函数，时说明当前至少有一个字符串的长度为0，此时为*i*与*j*中的最大值。时，为以下三种情况的最小值：

（1）表示删除；

（2）表示删除；

（3）表示替换。

本文采用的编辑距离相似度计算方法见式（3.3），为两个字符串的编辑距离，为两个字符串的最大长度。

（3.3）

## 意图分类

本文首先利用TF-IDF特征提取方法对文本进行特征提取和向量转换，然后采用朴素贝叶斯算法构建问句分类器，对用户问题进行正确分类。

### TF-IDF特征提取

TF-IDF（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（term frequency）指词频，即统计文本中某个词出现的频率，计算方法见式（3.4）。

（3.4）

其中表示特定单词w在文章中出现的次数，*n*表示文章的单词总数。该算法认为单词在文章中出现的频率能够衡量该单词对文章主旨表达的重要程度，但需要注意的是，一些通用的词语对于文章主题没有太大的作用，所以单纯的使用TF是不合适的。权重的设计必须满足：一个单词预测主旨的能力越强，其权重越大，反之，权重越小。在所有统计的文章中，一些词只是在其中很少几篇文章中出现，那么这样的词对文章主旨的作用很大，这些词的权重应该设计的较大。IDF就是在完成这样的工作。IDF（inverse document frequency）指逆文本频率指数，某一特定词语w的IDF可以由式（3.5）计算得出。

（3.5）

其中，表示语料库中文章的总数，表示出现单词w的文章的数目，1的作用是防止分母为0。TF-IDF的主要思想是：如果某个词在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词具有很好的类别区分能力，适合用来分类。TF-IDF的计算公式见公式（3.6）。

（3.6）

使用TF与IDF相乘，词的重要性随着它在文章中出现的次数成正比增加，同时会随着它在语料库中出现的频率成反比下降。

本文构建了21类问题模板作为训练数据，每种类别样本数据放在一个文件中，文件名含有类别编号，例如“【0】drug的临床应用.txt”，文件内容如图 3.4所示（全部模板及其文件内容见附录A）。



图 3.4 训练数据示例

Fig. 3.4 Training data example

在构建训练数据时先遍历所有文件，获取文件名中的数字，如果文件名含有数字，则读取该文件。设置当前文件下的数据标签，即该文件内数据对应的类别标签，对应训练样本的y数据。接着读取文件内容，每一行作为一条训练样本，每条训练样本的x数据由空格相连的词组组成。训练样本的x数据及y数据如图 3.5所示。

采用TfidfVectorizer将x数据向量化，为后续训练模型做准备，向量化后的数据如图 3.6所示。

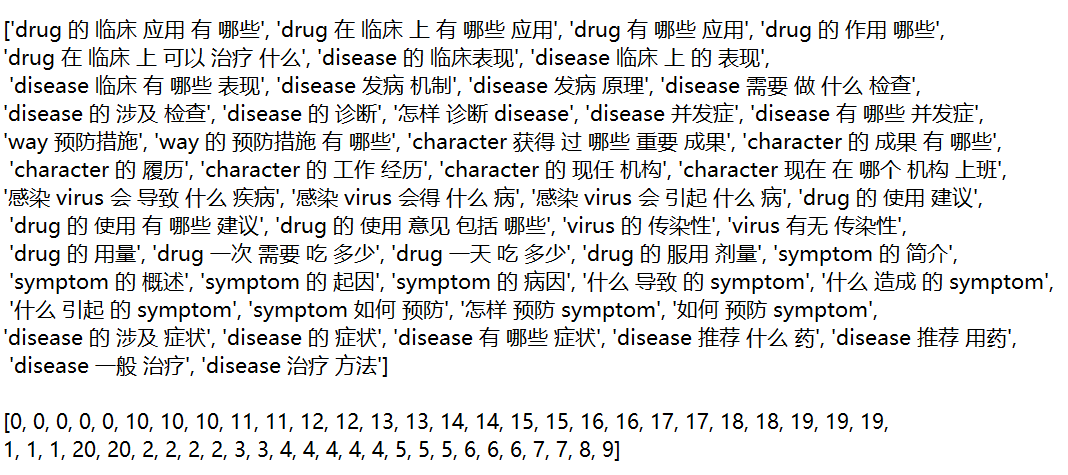


图 3.5 训练样本数据

Fig. 3.5 Training sample data

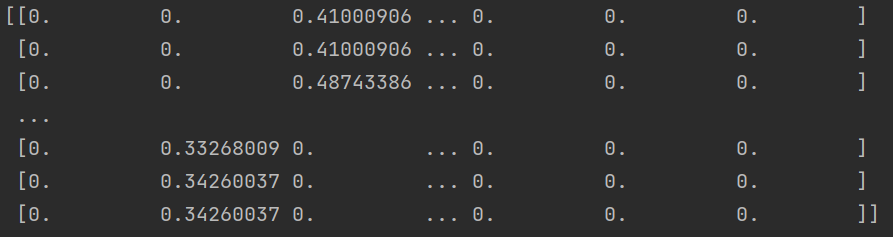


图 3.6 向量化后的数据

Fig. 3.6 Vectorized data

### 朴素贝叶斯算法

朴素贝叶斯算法（Naive Bayesian algorithm）[23,24]是目前应用最为广泛的分类算法之一。朴素贝叶斯算法是以贝叶斯定理（公式3.7）为基础并假设特征条件之间相互独立的方法，先通过给定的训练集，以特征词之间独立作为前提假设，学习从输入到输出的联合概率分布，再基于学习到的模型，输入X求出使得后验概率最大的输出Y。

（3.7）

给定训练集（X,Y），每个样本*x*都包含n维特征，即，类标记集合含有k种类别，即。判断一个新样本x的类别的问题转化为求x属于哪个类别的概率最大，即求解。根据贝叶斯定理，的计算方法见式（3.8）。

（3.8）

根据全概率公式，式（3.8）可以进一步表示为式（3.9）。

（3.9）

条件概率，假设各个维度的特征相互独立，则条件概率可以转化为式（3.10）。

（3.10）

式（3.10）代入式（3.9）得到式（3.11）。

（3.11）

于是朴素贝叶斯算法的公式计算为式（3.12）。

（3.12）

对于所有的，式（3.12）的分母值相同，所以忽略分母部分，朴素贝叶斯算法的计算公式最终表示为式（3.13）。

（3.13）

朴素贝叶斯算法假设了数据集属性之间是相互独立的，因此算法的逻辑性十分简单，并且算法较为稳定，当数据呈现不同特点时，朴素贝叶斯算法的分类性能不会有太大的差异。本文使用朴素贝叶斯算法对问句分类的工作流程如下：

（1）构建训练数据

这是为使用朴素贝叶斯算法做的准备工作。本文的相关工作是使用TF-IDF对数据特征提取和向量转化，为后续生成分类器准备数据，具体工作见3.3.1。

（2）生成分类器

本文在这个阶段的任务是构建问句分类器，具体工作见3.3.3。

（3）应用阶段

本阶段将使用问句分类器对问句意图进行预测，即判断输入问句属于哪一类问题。本阶段的输入是问句分类器和问题，输出是对于该问题预测的类别编号。

### 构建问句分类器

对训练数据进行TF-IDF特征提取和向量转化后，采用scikit-learn模块的朴素贝叶斯模型建模。scikit-learn模块有三种不同类型的朴素贝叶斯，分别是：高斯分布型，多项式型和伯努利型，本文使用多项朴素贝叶斯（multinomial Naive Bayes）。多项朴素贝叶斯实现服从多项分布数据的朴素贝叶斯算法，适用于文本分类。

多项朴素贝叶斯在计算先验概率和条件概率时会做一些平滑处理。计算先验概率的公式见式（3.14）。

（3.14）

其中*N*是样本个数，*k*是总的类别个数，是类别为的样本个数，是平滑值。计算条件概率的公式见式（3.15），其中*n*是特征的维数，是类别为的样本中，第i维特征的值是的样本个数。

（3.15）

平滑值应用于训练数据中没有出现的特征并防止在计算中后验概率为0，本文中将设置为0.01建立多项朴素贝叶斯模型，并利用TF-IDF向量化后的x数据和y数据训练模型，得到问句分类器。

## cypher语句生成

问句分类器对输入问题进行意图预测后，对每一类问题设置特定的cypher语句，用于Neo4j数据查询。例如输入问句“新型冠状病毒肺炎的症状”，经过实体识别获取“新型冠状病毒肺炎”，经过意图分类获取问题编号，对应症状的模板，其生成的cypher语句为：

MATCH (p)-[r:relation{label\_zh:'症状'}]->(n) WHERE p.label\_zh='新型冠状病毒肺炎' RETURN n.label\_zh

Neo4j图数据库查询结果如图 3.7所示。

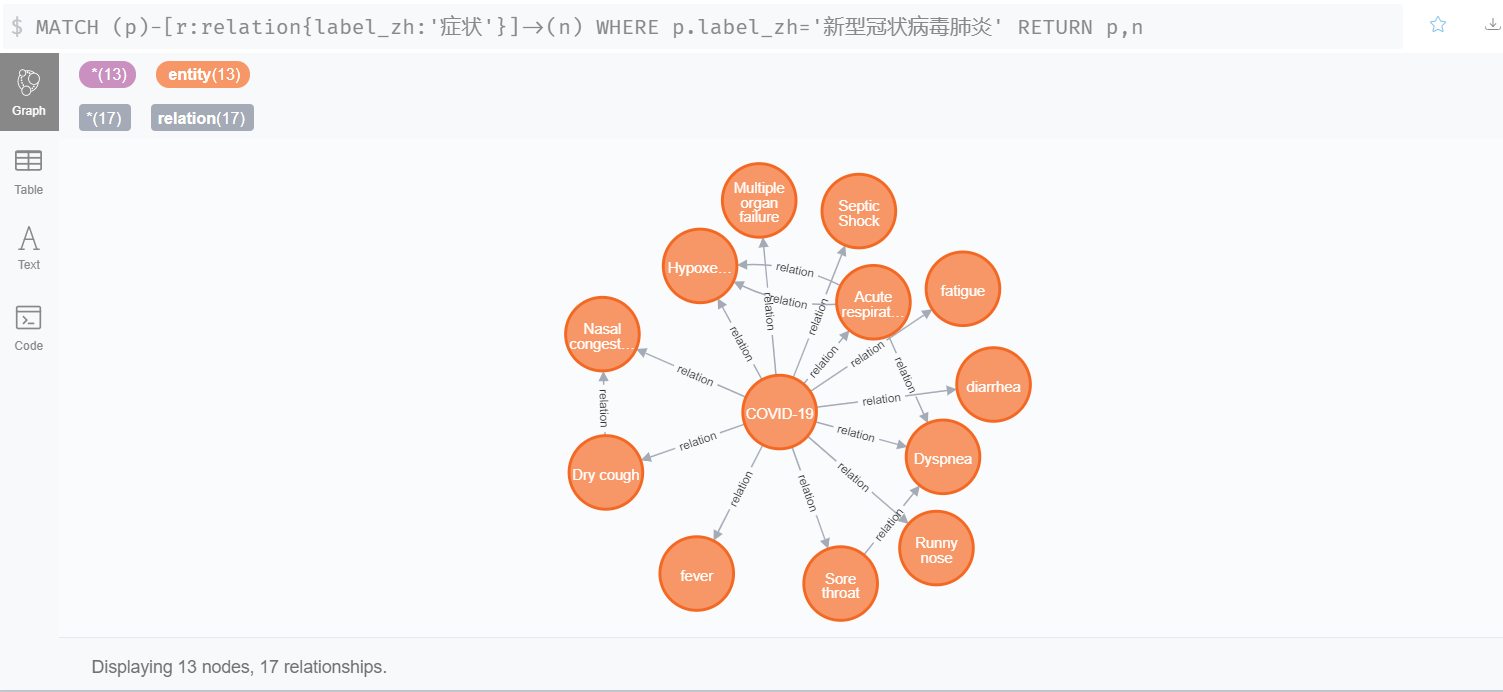


图 3.7 Neo4j查询结果

Fig. 3.7 Neo4j query results

## 本章小结

本章具体讲述了基于知识图谱的新冠肺炎问答系统的问答算法原理及本文对于相关算法的使用。本章首先对算法进行概述，总体介绍算法，然后分别从实体识别、意图分类和cypher语句生成三部分进行具体介绍，每部分给出了详细的算法原理和实现过程。

# 系统设计与实现

## 系统需求

### 功能需求

基于知识图谱的新冠肺炎问答系统包括以下三项主要功能：

（1）关系检索。搜索框输入实体名称，系统将以图的形式返回与该实体相关的全部实体，用户可以清晰直观的看到该实体与其他实体的关系。

（2）问答。用户在搜索框里输入问题，系统将返回最符合用户问句意图的答案。

（3）后台管理。系统提供给管理员后台管理的功能，管理员登录后台可以对知识图谱的实体及关系进行增加、删除和修改操作。

系统共有两类用户角色，分别是普通用户和管理员，普通用户在系统中无需登录，只能进行关系检索和问答；管理员除了进行关系检索和问答外，在进行身份验证后可以进行后台管理。

### 性能需求

对于一个完备的系统，系统性能需求如下：

（1）稳定性

对于一个系统而言，在用户输入错误时要能够进行处理，保证系统依然能够正常运行，不因用户操作错误而崩溃。例如，当用户输入的实体名称在数据库中没有对应的实体时，系统应该继续正常运行，并将实体不存在的信息反馈给用户。

（2）安全性

系统应保证只有通过验证的管理员可以对数据进行操作，用户不可以跳过验证进入后台管理。系统应保证数据库中图谱数据，管理员的用户名、密码等信息的安全不被泄露。

（3）可扩展性

当产品需求发生变化时，系统能够做到向外扩展。系统的架构应保证新模块增加时不会对系统原有功能和架构产生影响。

（4）响应时间要求

系统应具有快速响应的特性，用户打开界面和提交事务的平均响应时间应少于1.5秒；用户在线实时查询业务操作的数据处理时间应少于5秒。

## 系统设计

### 架构设计

系统采用B/S架构，使用Neo4j图数据库和MySQL关系数据库进行数据存储，其中，Neo4j图数据库存储知识图谱，MySQL关系数据库存储管理员信息。系统使用Python语言编写，服务器端框架使用Flask框架处理用户请求。系统前端使用Bootstrap框架进行开发，知识图谱的可视化使用Echarts技术。图 4.1展示了系统的架构设计。



图 4.1 系统架构图

Fig. 4.1 Architecture diagram of system

### 功能设计

本系统的核心功能为关系检索、问答和后台管理。系统包含普通用户和管理员两类用户角色，关系检索和问答面向所有用户，管理员进行身份验证后能够进行后台管理。系统功能结构如图 4.2所示，下面详细介绍系统功能。

（1）关系检索

关系检索实现知识图谱的可视化，当用户输入实体名称，回车或点击搜索后，系统将以图的形式返回与该实体有关的全部实体并进行动态展示，用户看到的结果接近知识图谱内部的图数据结构。

（2）问答

用户在问答界面的搜索框内输入问题，回车或点击搜索后，系统将返回最接近用户问句意图的答案。对于用户输入不准确的实体名称，系统返回的答案中显示知识图谱内准确的实体名称。对于系统不能处理的问题，返回“我也还不知道。”

（3）后台管理

管理员进行身份验证后，进入后台管理界面。管理员在后台能够进行修改实体、删除实体、添加实体、添加关系和删除关系等五类操作。

图 4.2 系统功能结构图

Fig. 4.2 Functional structure diagram of system

## 系统实现

### 知识图谱介绍

本文基于清华大学AMiner团队构建的COVID-19开放知识图谱进行研究，图谱下载地址：<https://covid-19.aminer.cn/kg/>。该图谱包含505个概念，393个属性，26733个实体，32352个知识三元组，涵盖了有关新冠肺炎的医疗、健康、物资、防控、科研和人物等几个重要领域。此外，本文根据中国科学院软件研究所刘焕勇的开源项目（https://github.com/liuhuanyong/QASystemOnMedicalKG）中整理的json数据进行处理，对原有图谱的数据在疾病、药物、检查等方面进行了补充。最后，如图 4.3所示，存储在Neo4j图数据库中的COVID-19知识图谱共包含29508个实体与75278条关系。

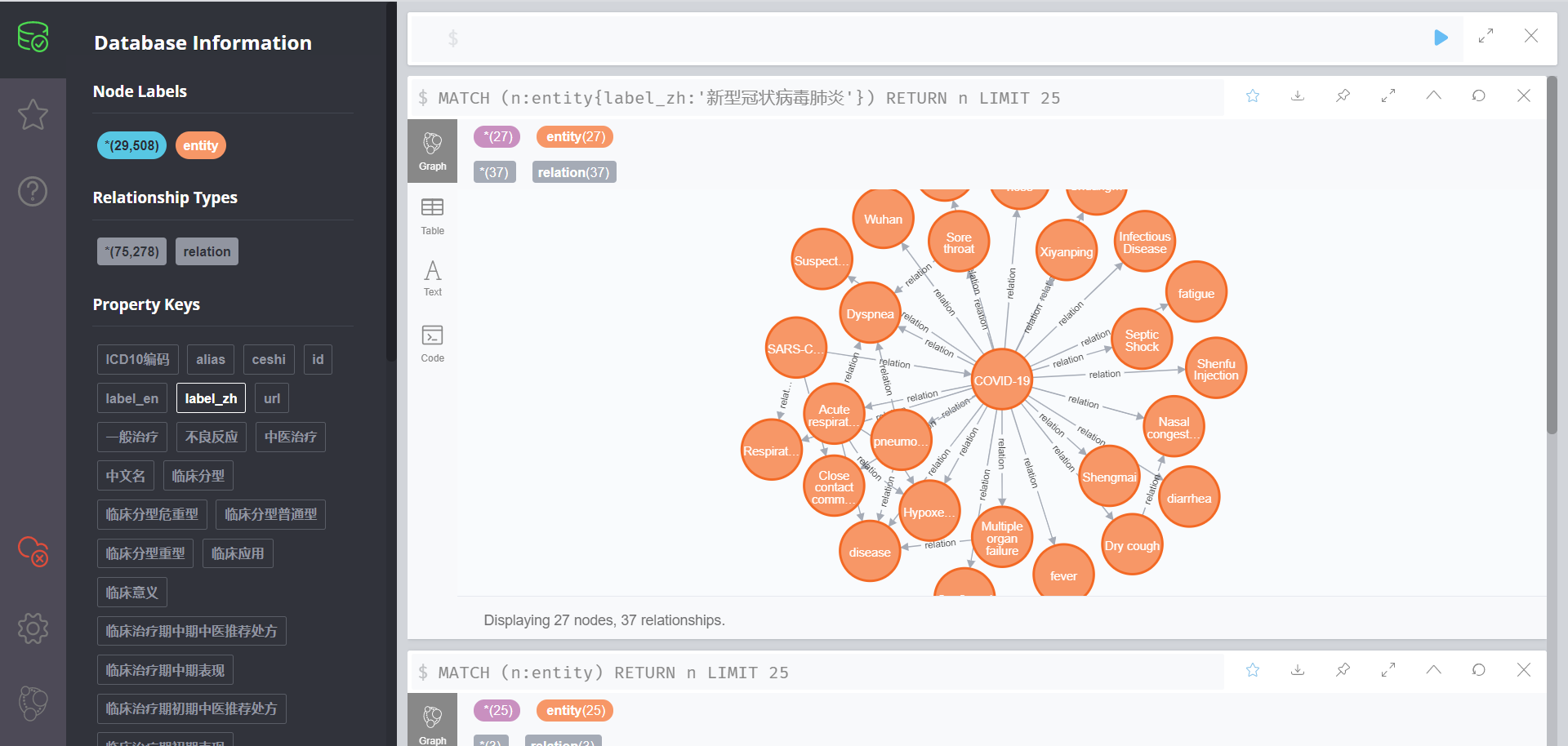


图 4.3 COVID-19知识图谱

Fig. 4.3 COVID-19 knowledge graph

### 关系检索

由于知识图谱数据存储在Neo4j中，当用户输入实体名称后，点击搜索或回车，将触发绑定的search函数，search函数将用户输入的数据传递给服务器端进行处理。服务器端得到传入的数据后先将其字符串化，然后连接数据库，在Neo4j图数据库中查询相关信息，将Neo4j返回的信息处理为json数据，包含节点数据和关系数据。服务器端将查询到的数据以json格式传递给前端，运用Echarts将数据可视化。例如查询“利托那韦”，点击搜索或回车后，系统返回结果如图 4.4所示。



图 4.4 关系检索示例

Fig. 4.4 Example of relation search

### 问答

用户进入问答界面，在搜索框中输入问题后，点击搜索或回车，前端以表单的形式将数据传递给服务器，服务器获取用户输入的问题后经过实体识别、意图分类和cypher语句生成，查询存储在Neo4j图数据库中的COVID-19知识图谱，最终将查询到的结果返回给前端，显示在界面上。问答系统框架如图 4.5所示。当用户输入问题“新型冠状病毒肺炎的症状”时，点击搜索或回车后，经过实体识别获取“新型冠状病毒肺炎”实体名称，经过意图分类，由问句分类器得出该问句对应的模板编号为6，生成cypher查询语句查询Neo4j图数据库，最终将结果输出，系统返回结果如图 4.6所示。



图 4.5 问答系统框架图

Fig. 4.5 Question answering system frame diagram



图 4.6 问答示例

Fig. 4.6 Q&A example

### 后台管理

管理员身份验证后进入后台管理界面，如下图 4.7所示，管理员可以对图谱数据进行增加、删除和修改操作。删除实体操作与修改实体操作同属于一个界面，因此，下面分别介绍添加实体、修改实体、删除关系、添加关系四个操作。



图 4.7 后台管理界面

Fig. 4.7 Background management interface

（1）添加实体。添加实体界面如图 4.8所示。

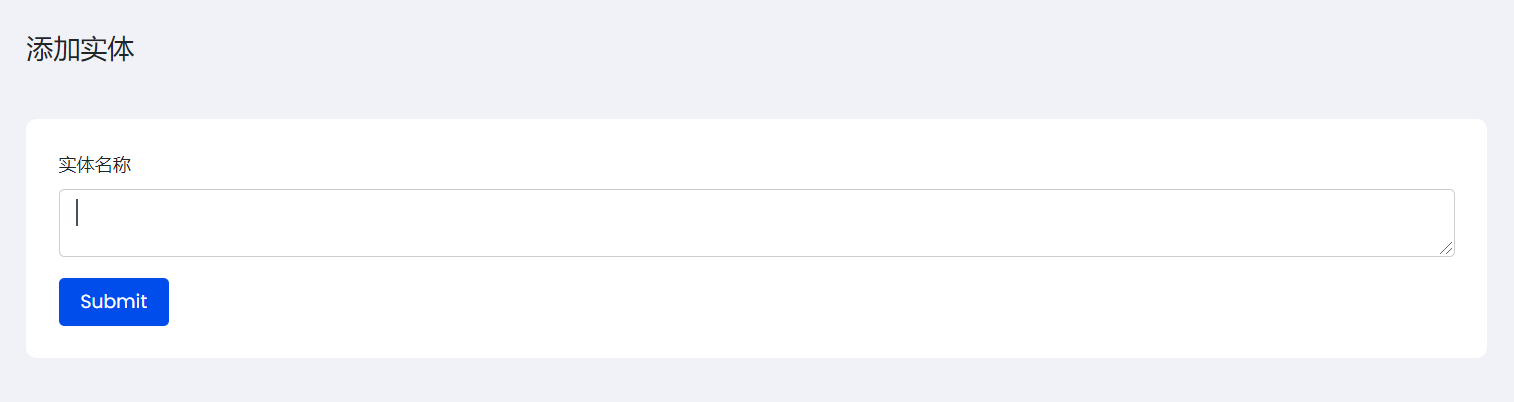


图 4.8 添加实体

Fig. . Add entity

输入实体名称，点击“Submit”，系统判断数据库中是否存在该实体，若存在，系统不会重复添加，否则，向Neo4j图数据库中添加一个实体。例如添加实体“结巴”，管理员输入实体名称后，点击“Submit”，数据库中没有名称为“结巴”的实体，系统将向Neo4j图数据库中添加“结巴”实体，添加成功后系统自动跳转至实体查询界面，管理员在搜索框内输入添加实体的名称，返回界面如图 4.9所示。

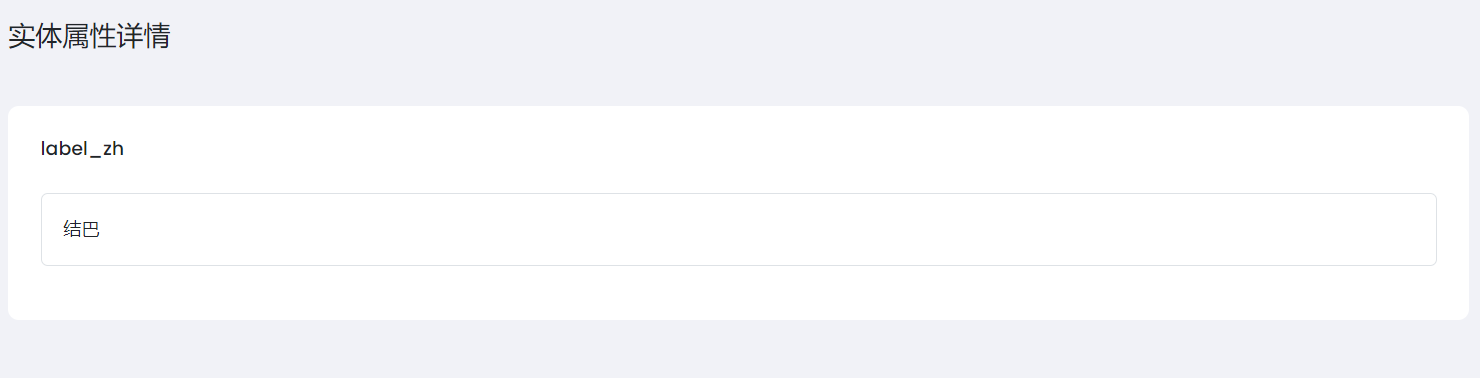


图 4.9 实体添加成功界面

Fig. 4.9 Interface for adding entity successfully

（2）修改实体。管理员点击“修改实体”后，搜索想要操作的实体，系统根据实体名称查找，若数据库中不存在该实体，系统将实体不存在的信息反馈给管理员；若存在该实体，系统返回实体属性详情界面，如图 4.10所示。管理员在此界面可以增加实体属性、删除实体属性、对已有属性修改以及删除该实体。点击“删除实体”，将删除该实体及与其相关的所有关系，该操作不可逆转。

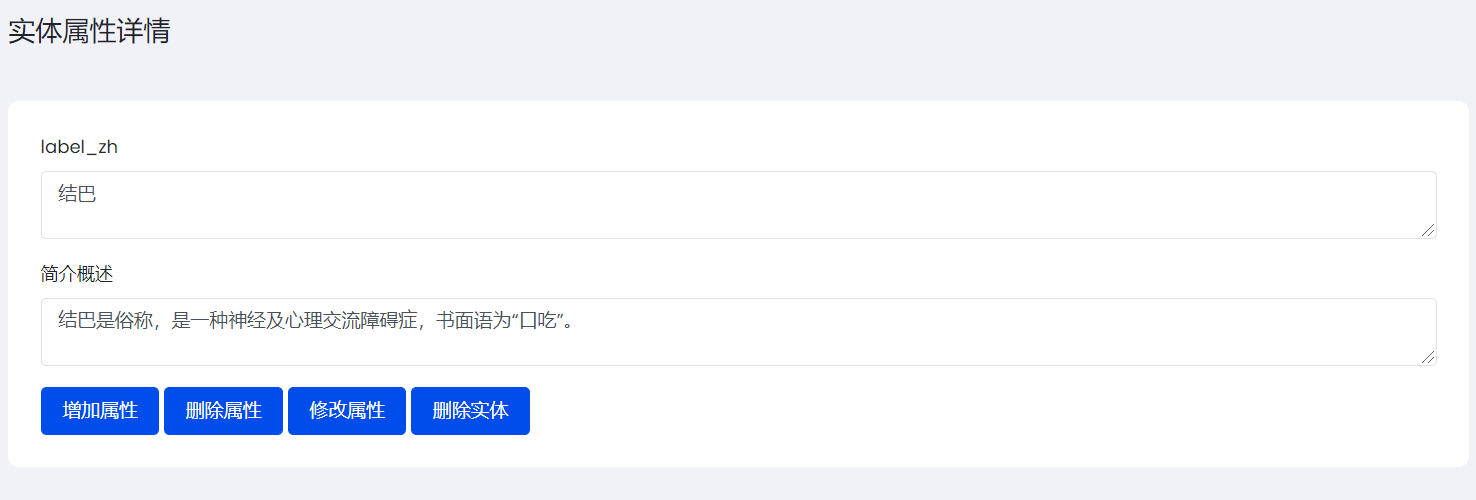


图 4.10 修改实体界面

Fig. 4.10 Modify entity interface

（3）删除关系。管理员可以对Neo4j图数据库中已有实体之间的关系进行操作。管理员点击“删除关系”后，首先在搜索框内输入想要删除关系的实体进行查找，系统在数据库中查找实体，若该实体不存在，系统返回提示信息；若系统中包含对应实体，找到该实体后进入操作界面，界面显示该实体的关系图，管理员可以在此界面进行操作。例如，管理员搜索“利托那韦”，系统返回界面如下图 4.11所示。

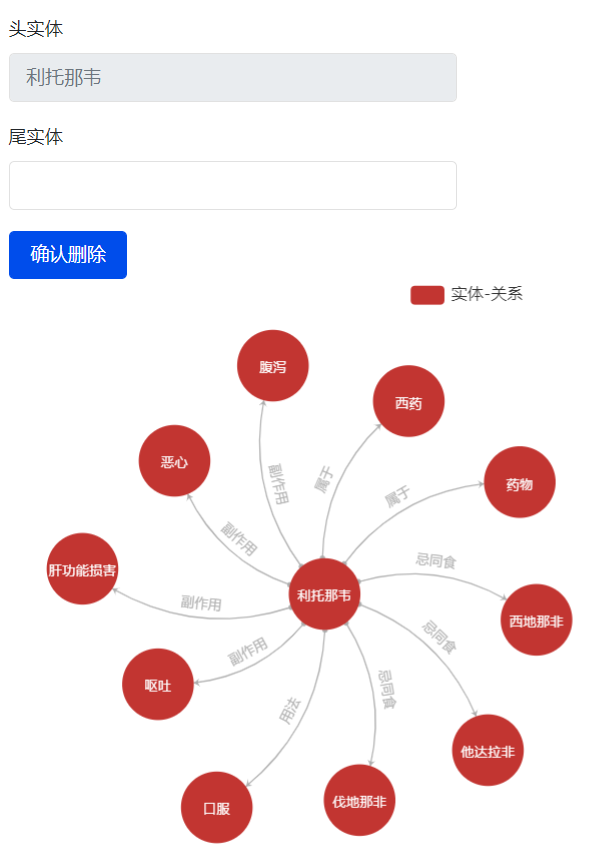


图 4.11 删除关系前利托那韦关系图

Fig. 4.11 Ritonavir diagram before deleting relationship

头实体表示管理员正在查找的实体，在删除关系界面不可修改。管理员在尾实体处输入实体名称，点击“确认删除”，即可删除由头实体指向尾实体的关系。若输入的尾实体不存在或关系不存在，则关系图不会有任何修改，否则，返回删除关系后的关系图。例如输入尾实体为“腹泻”，存在由“利托那韦”指向“腹泻”的关系，点击“确认删除”后，系统将删除这条关系，关系图更新。

（4）添加关系。添加关系与删除关系操作类似，管理员点击“添加关系”，搜索想要操作关系的实体，系统判断是存在该实体，若不存在，系统返回提示信息；否则，系统返回添加关系界面，该界面显示实体关系图。不同于删除关系界面，添加关系时，管理员需指定添加关系的名称，如下图 4.12所示。若管理员输入的尾实体不存在，则不会创建关系，关系图不更新。



图 4.12 指定关系名称

Fig. 4.12 Specify the relationship name

## 本章小结

本章首先介绍了基于知识图谱的新冠肺炎问答系统的功能需求和性能需求，接着对系统进行了架构设计和功能设计，并对COVID-19知识图谱的数据进行了介绍，最后分别叙述了系统的三个主要功能即关系检索、问答和后台管理的实现。

# 工作总结与展望

## 工作总结

在后疫情时代，包括世界卫生组织在内的科学界一致认为，新型冠状病毒将与人类长期共存。人们要在这场与病毒的战役中保护好自己，这离不开数据的支持。为了提高信息检索效率以及帮助人们更好地了解新冠肺炎相关信息，本文实现了基于知识图谱的新冠肺炎问答系统，本文的主要研究工作总结如下：

（1）本文基于清华大学AMiner团队构建的COVID-19开放知识图谱进行研究，并对其中的数据进行了补充。对用户输入的问句进行实体识别，从Neo4j中查询并导出相关实体，进行人工标注并将结果存入用户字典，jieba加载用户字典对用户问题进行词性标注。对于词性标注未能识别到的实体，采用相似匹配。

（2）本文根据知识图谱中的数据构建了21类问题模板，将构建好的问题模板处理为训练数据，首先运用TF-IDF对其进行特征提取和向量转化，然后运用朴素贝叶斯算法训练问题分类器，对于用户输入的问题能够正确识别对应的模板编号。设计cypher语句查询Neo4j图数据库，将查询结果以文本形式显示在Web端。

（3）搭建系统。本文以Python语言为基础，选用Flask框架进行系统搭建，前端使用Bootstrap进行开发，运用Echarts实现知识图谱可视化，帮助用户直观清晰地看到实体间的相互关系。经过需求分析后本文确定系统分为两类用户角色，主要实现三大功能。普通用户使用此系统除了问答功能外，还能够对图谱中的数据进行关系检索。管理员除了上述两个功能外，系统还提供了后台管理功能，帮助管理员更新图谱数据。

## 工作展望

本文设计并实现了基于知识图谱的新冠肺炎问答系统，但是仍有一些可以改进的地方：

（1）本文使用的知识图谱数据不够完整，后续可以继续收集数据补充到现有的知识图谱中，丰富实体属性和关系，提升用户体验。

（2）本文设计的问题模板不够全面，没有覆盖全部问题类型，后续可以对问题模板进行补充设计。

# 致谢

随着毕业论文的完成，我的四年大学生活也即将结束，在这四年里，我学习到了许多计算机的专业知识，认识了许多帮助我的老师和同学，非常感谢能够遇到他们。

首先，我要感谢我的本次毕业设计指导老师李冬梅老师。李老师面对我的问题总会耐心解答，给我提供帮助。十分感谢李老师对我的真诚指导。感谢李老师认真负责地督促，帮助我按时完成毕业设计论文。感谢李老师带我第一次接触并认识了知识图谱和自然语言处理，整个毕业设计过程让我学会了很多。非常荣幸李老师能成为我的毕业设计指导老师，为我大学四年的求学生涯画上完满的句号。

其次，我要感谢我的企业指导老师张小平老师，感谢他为我提供医药领域的专业知识指导，帮助我的项目更加的专业化。

我还要感谢帮助过我的学长学姐，感谢他们耐心回复我技术上遇到的问题。感谢同学和朋友们对我的鼓励和支持，他们帮助我在遇到瓶颈时保持良好的心态，鼓励我努力地去探究和解决问题。

最后，感谢父母和家人对我生活上的支持和精神上的鼓励。

# 参考文献

1. Nanshan Chen, Min Zhou, Xuan Dong, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study[J]. The Lancet, 2020, 0(prepublish):507-513.
2. 张路,马冉,韩雯雯,徐娟年,程灏.哈尔滨市新型冠状病毒肺炎疫情分析与复工建议[J].重庆师范大学学报(自然科学版),2020,37(02):107-111.
3. 人民网.世卫组织将新冠肺炎命名为“COVID-19”[EB/OL].(2020-02-14)[2020-03-11]. http://world. people.com.cn/n1/2020/0212/c1002-31582421.html.
4. 张志强,张邓锁,胡正银．突发重大公共卫生事件应急集成知识咨询服务体系建设与实践——以新冠肺炎(COVID-19)疫情事件为例[J]. 图书与情报,2020,000(002):1-12.
5. 王智悦,于清,王楠,王耀国.基于知识图谱的智能问答研究综述[J].计算机工程与应用,2020,56(23):1-11.
6. BERNERS-LEE T, HENDLER J, LASSILA O. The semantic Web[J]. Scientific American Magazine, 2008, 23(1):1-4.
7. AMIT S. Introducing the knowledge graph[R]. America: Official Blog of Google, 2012.
8. Lv Q, Xu L, Jie Y, et al. Research on domain knowledge graph based on the large scale online knowledge fragment[C]. 2014 IEEE Workshop on Advanced Research and Technology in Industry Applications (WARTIA), 2014.
9. Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11):39-41.
10. Auer S, Bizer C, Kobilarov G, et al. Dbpedia: A nucleus for a web of open data[M]. The semantic web. Springer, Berlin, Heidelberg, 2007:722-735.
11. Xu B, Xu Y, Liang J, et al. Cn-dbpedia: A never-ending Chinese knowledge extraction system[C]. Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, 2017:428-438.
12. Niu X, Sun XR, Wang HF, et al. Zhishi.me-weaving Chinese linking open data[C]. Proceedings of the 10th International Semantic Web Conference. Bonn, Germany. 2011. 205–220.
13. Francesca Frontini; Riccardo Del Gratta; Monica Monachini. GeoDomainWordNet: Linking the Geonames Ontology to WordNet[J]. Human Language Technology. Challenges for Computer Science and Linguistics, 2016, Vol. 9561:229-242.
14. Turing A M.Computing machinery and intelligence[J].Mind, 1950, 59(236):433-460.
15. Weizenbaum J.ELIZA—a computer program for the study of natural language communication between man and machine[J].Communications of The ACM, 1966, 9(1):36-45.
16. Tunstall-Pedoe W.True knowledge: open-domain question answering using structured knowledge and inference[J]. AI Magazine, 2010, 31(3):80-92.
17. Chakraborty N, Lukovnikov D, Maheshwari G, et al. Introduction to neural network based approaches for question answering over knowledge graphs[J]. arXiv preprint arXiv:1907.09361, 2019.
18. Ji S, Pan S, Cambria E, et al. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021.
19. Huang Xiao, Zhang Jingyuan, Li Dingcheng, et al. Knowledge graph embedding based question answering[C]//Proceedings of the 12th ACM International Conference on Web Search and Data Mining, 2019:105-113.
20. 刘峤,李杨,段宏,刘瑶,秦志光.知识图谱构建技术综述[J].计算机研究与发展,2016,53(03):582-600.
21. 徐增林,盛泳潘,贺丽荣,王雅芳.知识图谱技术综述[J].电子科技大学学报,2016,45(04):589-606.
22. 李贺,刘嘉宇,李世钰,吴迪,金帅岐.基于疾病知识图谱的自动问答系统优化研究[J].数据分析与知识发现,2021,5(05):115-126.
23. HOFF P D.A first course in Bayesian statistical methods[J]. Journal of the Royal Statistical Society, 2010, 173(3):694-695.
24. 王爱平,张功营,刘方.EM算法研究与应用[J]. 计算机技术与发展,2009,19(09):108-110.

# 附录A 问题模板

问题模板文件名及文件内容见表A1。

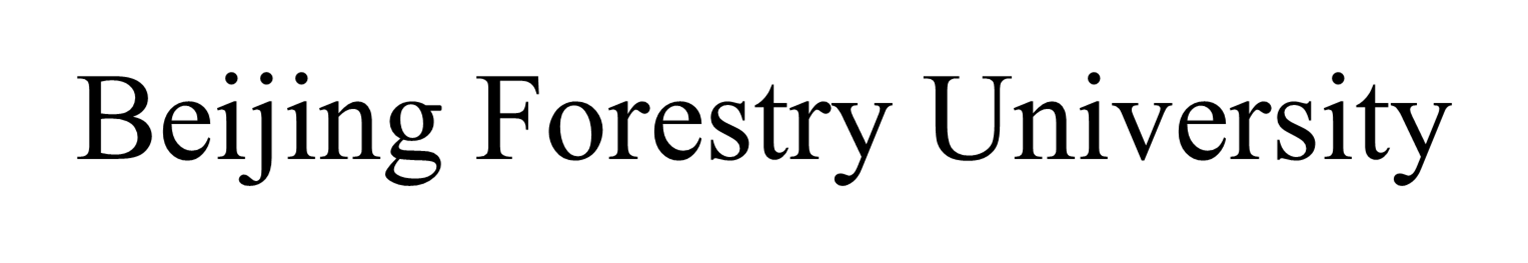
**表A1问题模板**

**TableA1 Question template**

|  |  |
| --- | --- |
| 文件名 | 文件内容 |
| 【0】drug的临床应用 | drug的临床应用有哪些  drug在临床上有哪些应用  drug有哪些应用  drug的作用哪些  drug在临床上可以治疗什么 |
| 【1】drug的使用建议 | drug的使用建议  drug的使用有哪些建议  drug的使用意见包括哪些 |
| 【2】drug的用量 | drug的用量  drug一次需要吃多少  drug一天吃多少  drug的服用剂量 |
| 【3】symptom的简介 | symptom的简介  symptom的概述 |
| 【4】symptom的起因 | symptom的起因  symptom的病因  什么导致的symptom  什么造成的symptom  什么引起的symptom |
| 【5】symptom的预防 | symptom如何预防  怎样预防symptom  如何预防symptom |
| 【6】disease的涉及症状 | disease的涉及症状  disease的症状  disease有哪些症状 |
| 【7】disease推荐药物 | disease推荐什么药  disease推荐用药 |
| 【8】disease一般治疗 | disease一般治疗 |
| 【9】disease治疗方法 | disease治疗方法 |
| 【10】disease临床表现 | disease的临床表现  disease临床上的表现  disease临床有哪些表现 |
| 【11】disease发病机制 | disease发病机制  disease发病原理 |
| 【12】disease涉及检查 | disease需要做什么检查  disease的涉及检查 |
| 【13】disease诊断 | disease的诊断  怎样诊断disease |
| 【14】disease并发症 | disease并发症  disease有哪些并发症 |
| 【15】way预防措施 | way预防措施  way的预防措施有哪些 |
| 【16】character获得过哪些重要成果 | character获得过哪些重要成果  character的成果有哪些 |
| 【17】character的履历 | character的履历  character的工作经历 |
| 【18】character的现任机构 | character的现任机构  character现在在哪个机构上班 |
| 【19】virus会导致什么 | 感染virus会导致什么疾病  感染virus会得什么病  感染virus会引起什么病 |
| 【20】virus传染性 | virus的传染性  virus有无传染性 |









树木树人

知山知水