基于企业知识图谱的问答系统

肖晗 515030910113

June 21, 2020

Abstract

知识图谱(Knowledge Graph)是近些年来兴起的一项技术，被广泛应用于问答系统、风险预测、推荐系统等，如何利用现有的数据构建一套完善的知识体系是知识图谱研究的重点。本课题旨在研究以知识图谱为工具，构建一个在企业知识领域的问答系统。使用爬虫手段和API获取非结构化数据与结构化数据，然后辅以自然语言处理(NLP)的技术进行命名实体识别和基于规则的关系抽取。使用Neo4J图数据库的三元组关系存储企业知识实体和关系，搭建一个互联网企业领域的知识图谱，最后用Web的方式实现KBQA系统的展示，将用户的问题进行语义分析，然后以可视化的方式展示问题的答案。

1 第一章 绪论

本章主要概述了本课题的研究路线以及研究的意义所在。

1.1 课题研究路线

本课题旨在研究企业知识图谱的构建和应用，将重点以互联网行业为研究对象，研究这些企业的内部属性和企业之间的联系，构建一个互联网行业的企业知识图谱，并且以智能问答的应用方式进行展示。

重点针对现实中的互联网行业，将行业内的企业、人、投资企业、品牌视为实体，利用知识图谱及相关技术能够很好的挖掘与展示实体以及实体之间的关系，从而得到企业知识图谱，为企业智能问答的应用提供服务。

1. 设计并实现面向互联网行业的企业知识检索平台，结合天眼查和创头条资讯的开放数据环境，根据行业关键字自动获取与企业相关的非结构化文本信息，并从中抽取实体等要素，挖掘潜在关联关系，为企业智能问答等提供数据支持；

2. 利用爬虫技术，在创头条资讯的网站上自动获取每日最新的资讯，这些资讯是行业内的企业的非结构化文本信息，里面拥有大量的冗余信息和无效信息；

3. 对第2步抓取的每日资讯这类的非结构化文本信息，使用自然语言处理（NLP）技术，从非结构化文本信息中获取企业实体和企业实体的一些关系抽取；

4. 根据第3步中的命名实体识别（NER）获取的企业实体，根据这些实体的基本信息（名字、法人等），结合天眼查提供的外部开放数据填充实体属性，比如企业的法人、注册信息、对外投资信息、职位信息、品牌信息等等，分析实体内部以及实体之间的语义关系，使用相关匹配技术完成对多个数据源的实体匹配；

5. 将经过第4步处理的实体以及关系存储到图数据库Neo4J中，对获取的实体集合的知识库进行企业知识图谱构建，并支持语义检索功能，能在Neo4J中使用查询语句可视化地展示查询结果；

6. 结合Web应用技术，前端使用D3.js作为工具，将智能问答的查询语句的返回结果，以可视化的方式进行展示，实现能够智能问答的企业知识检索平台，并对平台扩展性、性能以及知识查询结果进行验证。

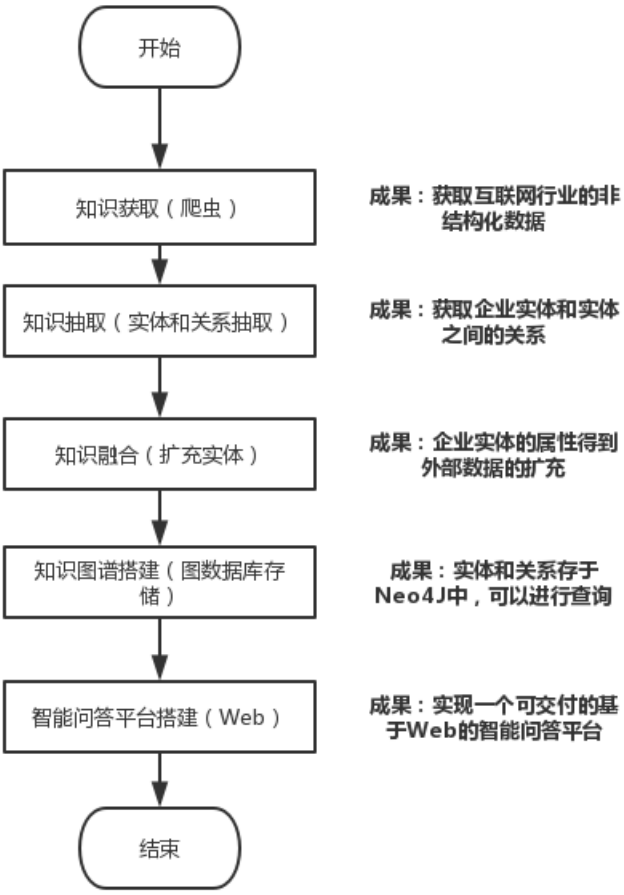

Figure 1: 知识图谱搭建流程

1.2 课题研究的意义

知识图谱在商业领域的应用，它扩展了原始科学知识地图的内涵，也使其应用场景得到了扩展。商业领域的信息比主题领域更混乱，需要对大多数信息源和非结构化文本进行语义分析，提取可用于映射知识地图和查找知识的知识单元，这种单元之间的连接非常重要。尤其在这个互联网蓬勃发展的时代，互联网行业每天的新企业都如雨后春笋般地注册产生，企业与企业之间的关系也更为紧密，企业内部的信息变更也越来越多、越来越快。

传统地，我们分析一家企业的信息，需要去互联网上搜索不同网站的不同信息，然后进行整合，得到企业实体信息，但是如果要进一步地挖掘出该企业与其他企业之间的联系，对于操作者而言难度会很大。那么，企业知识图谱的构建会高效而准确地帮助我们解决这种问题，有以下的优势：

1. 在互联网上的具有真实可靠的企业信息网站（如企业官网、天眼查这类企业信息的数据网站等）中

进行信息抓取，获取大量的企业信息的知识，然后可以更加高效地扩充、整合企业的信息，相比于人工的查询、记录，显得更加的高效和准确。

2. 要得到单独某家企业的信息靠人工完成是不难的，但是要求挖掘出企业与企业之间的深层联系，这对于人工是一件费时的事，而且，随着企业数量的增加，人工的成本开销会不可估量。此时，我们运用知识图谱及相关技术，利用之前抓取到的数据和图数据库的可视化，我们可以容易地挖掘出企业与企业之间的联系，而且这种联系是可以维护和扩充的，是一种高效而切实可行的方法。

3. 我们可以借助用企业知识构建出的图数据库，根据企业与企业之间的联系，不仅可以直观地展示联系，而且不难可以进行下一步的分析，比如可以分析出某企业的发展是否有风险，某企业如何可以进一步地发展等等，利用知识图谱不仅可以准确分析，还可以进行预测评估。

2 第二章 企业知识图谱的构建

本章将根据研究路线，依次介绍企业知识图谱的构建流程，为问答系统的构建做准备。

2.1 知识获取

非结构化数据是我们日常生活中遇到最多的数据，非结构化数据是难以使用数据库二维逻辑表或者键值对来表现的数据，比如日常生活中我们见到的文本、表格、图片、语音等等。本课题将着重从文本这种非结构化数据出发，挖掘咨询、新闻这种非结构数据中的蕴含的实体已经实体之间的关系。

爬虫即“网络爬虫”，网络爬虫的主要作用是依据一定的爬行策略自动的从网上下载网页镜像到本地，并能够抓取所有其能够访问到的网页以获取海量信息，对其中的数据进行解析，通常是HTML语言进行解析，剔除掉标签的影响，从中分析挖掘出需要的信息等。其基本工作原理如下：首先初始化一个URL作为爬行的开始位置，如果该URL没有被抓取过，解析其DNS信息，尝试与这些URL链接所在的服务器建立连接，自动提取页面上的信息保存至本地，同时提取新的URL，根据一定的遍历算法将其去重过滤后加入待爬取队列，重复以上步骤遍历所有的网页数据，直到待爬取的队列中没有可用的URL，满足停止的条件时结束爬取。

本课题采用的正是爬虫来从互联网上获取非结构化数据。“创头条”是一个和互联网行业资讯息息相关的网站，其中的“创精选”模块每日会精选新鲜出炉的热拉新闻并且发布，是获取互联网行业的非结构化数据知识的可靠网站。

本课题使用的主要爬虫工具是Python3+Selenium。Python是目前最流行的爬虫开发的脚本软件，因为其轻量和丰富的库（比如urllib库）而被本文选择使用。本课题使用的是Selenium库中的webdriver功能，这个功能可以模拟用户登录浏览器并且访问指定的url网址，并且可以模拟点击事件的发生，比如可以模拟用户一直点击“下一页”，不断地获得非结构化的数据，非常适合于爬取那些带有简单反爬措施的网站。

因为爬虫直接访问网页，获得的仅仅是页面的html代码，而这种代码中拥有很多的标签，类似h1、p、image等等，这些标签会影响我们获取的知识的可操作性。所以在爬虫爬取网页时候，就需要使用一些技术或者库来实现爬取较为不含标签、较为干净的文本。

本课题具体将类似图像image、表格form等标签以及标签的内容删除，并且成功地使用正则表达式从HTML代码中获取了保存在指定的标签内的每日新闻的文本，即下一步将要进行知识抽取的初始文本。至此，在解析页面完毕后，本课题对于获取的文本的预处理工作就完成了。

2.2 知识抽取

2.2.1 命名实体识别

命名实体识别（Name Entity Recognition）简称NER，指在识别文本中具有特殊意义的实体，主要

包括了地名、人名、机构名、公司名等等。命名实体识别在问答系统、句法分析、信息提取等应用领域都有着重要的基础作用，是自然语言处理（Natural Language Processing，简称NLP）技术能走向实用化的重要原因之一。

国外对于英文命名实体识别早在1990年代就已经开始了，并且在一系列的国际会议上将命名实体识别作为其中的一项任务，因为英文命名实体识别只需要考虑词自身的特征而不需要考虑分词问题，所以在准确率上比起中文会高不少，实现难度也相对较低。

中文的命名实体识别区别于英文，必须要先进行词法分析，因为比如“研究”这个词，它可以是一个动词，比如“研究动物”，又可以是一个名词宾语，比如“做研究”，还可以是一定修饰定语，比如“研究所”。这就给中文命名实体识别带来了挑战，现在因为问答系统、机器翻译、语义网络等新鲜技术的兴起，中文命名实体识别的重要性和重视程度又提到了一个新的高度。

本课题使用的玻森数据Boson提供的命名实体识别工具，对之前获取的“创头条”的每日资讯内容进行知识抽取，分别将企业、人名、职位、产品名的信息保存，获取最初的数据集，为之后的企业知识图谱的构建做准备。命名实体识别效果如图2所示。

Figure 2: 玻森NER示例

2.2.2 关系抽取

本课题是要构建一个企业信息的知识图谱，其中形成实体知识库的几大实体有公司、人、品牌产品，然后基于同类实体或者不同类实体之间的联系，预先设计了几种关系：公司和公司之间的投资关系、公司和人之间的法人（拥有）关系、公司和人之间的职位关系、公司和品牌产品之间的所有关系以及品牌和品牌之间的竞争关系。

本课题使用的关系抽取的方法就是基于规则的实体关系抽取，在这里具体是基于语法规则，在上一步获取实体后，根据预设的语法规则进行关系抽取，因为其准确和高效非常适合进行初步的关系抽取，而且相对容易实现。

举例子而言，在本课题中具体研究的其中一种关系“投资关系”。本课题可以对投资关系设置几种语法规则，比如“实体（公司）投资了实体（公司）”、“实体（公司）被实体（公司）投资了xx万”、“实体（公司）对实体（公司）进行了投资”、“实体（公司）向实体（公司）提供了资金”等等，然后就是利用正则表达式到文本中去寻找定位“投资”这个关键词，并且返回“投资”这个句子中存在的实体，然后根据之前得到的实体知识库，判断句子中的实体是否满足是两个不同的公司名，如果是的话，就可以推测出这两个公司实体存在投资的关系。当然，在知识图谱中，实体之间的关系必须是有向的，所以在得到两个公司存在投资关系后，还要确定哪一个公司是投资方。

至此，本课题已经获得了企业知识图谱的最初实体库和实体之间的关系，概念层和实例层如图3所示。

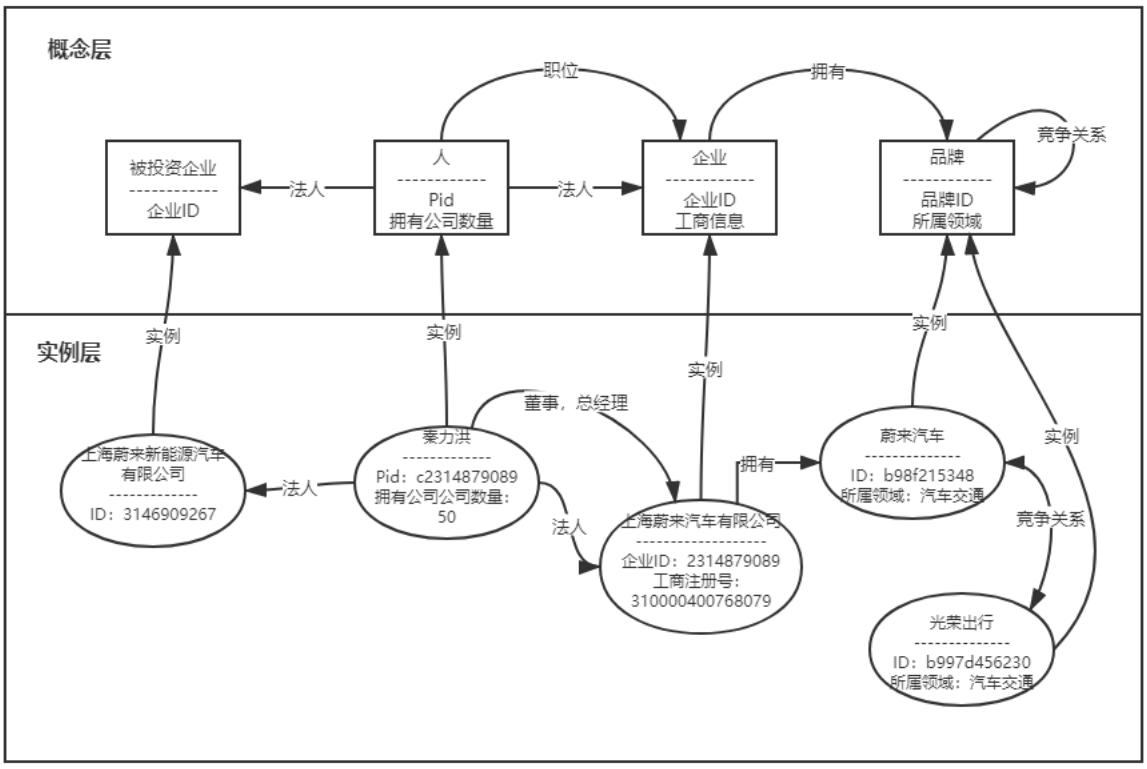

Figure 3: 企业知识图谱的模式结构

2.3 知识融合

本课题在完成命名实体识别的步骤后，获得了一个初步的实体知识库，里面包含了公司、人、品牌产品三个类型的实体，虽然已经可以将这些知识存储到数据库中，保证数据的持久化，但是由于NER技术的不够成熟，实体知识库中难免会存在海量的冗余信息以及无效甚至错误的信息。那么如何删除冗余、无效、错误的实体信息就是要提高知识库质量和知识图谱的质量的重要课题，而知识融合就可以在一定程度上帮助解决这个问题。

2.3.1 利用外部数据源对实体扩充

因为已经获取的实体知识库中的实体是从非结构化数据的文本中获取的，而这种非结构数据中蕴含的实体的属性本来就少之又少，而且关系抽取是抽取的实体与实体之间的关系，那么对于实体自身的属性就不够丰满，不能够为企业知识图谱提供丰富的查询功能。

天眼查是一个以公开数据为切入点、以关系为核心的产品，这种特性刚好与知识图谱的实体、关系的概念相契合，所以从天眼查中获取企业实体的属性，然后用来填充到实体知识库中无疑是一个明智的选择。天眼查的开放平台提供数据的api接口，用户可以利用这些api获取企业的结构化的信息，这种结构化的信息可以不经过其他的格式处理、筛选就可以直接添加为企业实体的属性，这大大地丰富了企业知识图谱的企业实体的信息量。

本课题使用的是Python脚本语言，使用的是python上的requests的库，通过requests.post的方法将密钥和文本进行传输，然后会收到一个以JSON格式返回的response恢复，接下来可以使用requests.json()方法来解析JSON格式里面的键值对关系，以结构化数据的方式存储为企业实体的属性，对实体进行了扩充。

但是，在这里值得一提的是，由于天眼查提供的api非常昂贵，单词查询一家企业的信息就需要5角

上下不等，而本课题的实体大约在一万上下，全部使用api开销太大，所以后期仍然使用爬虫获取天眼查的企业信息的页面，扩充知识库的实体内容。

2.3.2 知识合并

上面提到的由于NER技术的不够成熟，所以实体知识库中难免会有海量的冗余信息以及无效甚至错误的信息。比如“联通”、“中国联通”、“中国联合网络通信集团有限公司”这三个名次会在实体知识库中以三个不同的实体存在，但是本质上这三个实体其实是同一个实体三种叫法问题，将这三个实体合并为一个实体，消除了知识库中存在的这类冗余，这就是知识合并。

使用天眼查上的企业数据扩充实体，本课题在天眼查上获取的企业数据之一就是工商信息，工商信息包括了法定代表人、工商注册号、组织机构代码、注册资本、成立日期、统一社会信用代码等信息，在获取到工商信息后，都先以企业实体的属性存进实体知识库，为后续的知识合并做准备。如图4，根据观察，可以发现“联通”、“中国联通”、“中国联合网络通信集团有限公司”这三个实体在法定代表人、工商注册号上是完全一致的，说明这三个实体所对应的的工商信息是同一份，因为在中国一家企业/公司只能对应一个工商信息，相当于是企业的身份证，所以如果工商信息相同，那么说明上面的三个“联通”的企业其实是同一个企业实体，至此，就可以在企业知识库中删去“联通”、“中国联通”这两个实体，保留了一个名字最为完全、最为标准的实体，这就是知识合并要做的事。

至此，完备的实体知识库和关系知识库基本已经就位了，为了使这些数据和知识能够持久化，下一步对知识的存储就是至关重要了。

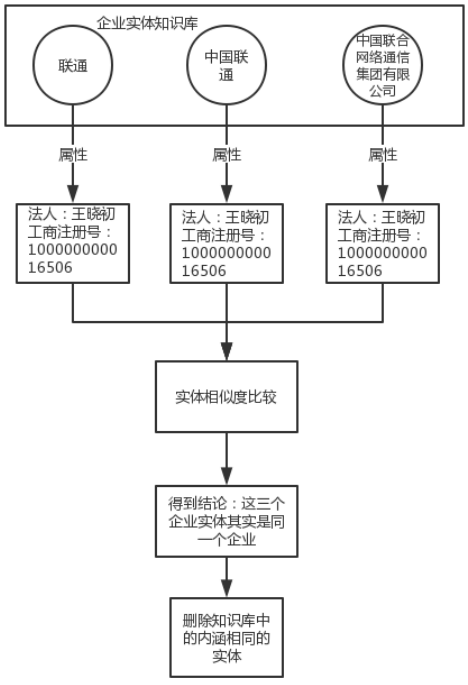

Figure 4: 知识合并的逻辑过程

2.4 知识存储

在知识融合这步完成后，已经得到了比较完善、内容比较具体的实体知识库和关系知识库，而这样还不能称之为知识图谱，因为数据没有持久化。数据持久化是知识存储的目的，在数据库中就可以通过查询语句对实体和关系进行查询，而且也为之后基于企业知识图谱的智能问答平台做准备。本课题选择使用Neo4J图数据库作为知识存储的工具，cypher语言作为数据库查询的语言。

Neo4J是一种非关系型数据库，属于NoSQL的一种，作为图数据库，Neo4J将结构化的数据存储到了图/网络中，以节点和边的形式进行刻画。之所以使用Neo4J这种图数据库，是因为它的几大优势：第一，Neo4J是一个高性能的图数据库，也可以理解为高性能的图引擎，用户可以使用他快速地进行数据存储、数据查询等等操作；第二，多平台的特征，可以使用Neo4J Browser直接在浏览器上对存在本地的图数据库进行操作，也可以通过Neo4J Desktop在本地的环境下进行操作；第三，作为新鲜的图数据库，Neo4J在中国网络上有大量的使用经验和新手入门教程，这就对用户很友好了。

本课题对之前命名实体识别阶段的实体继续细分了一下，从天眼查上除了获取工商信息外，还进一步通过api获取了企业的对外投资情况，因此将企业实体分为了有详细信息的企业实体Company和被投资的没有详细信息的被投资企业实体InvestCompany，另外还有人Person实体和品牌Brand实体，节点之间的关系也分为了投资关系invest、法人关系islegalperson，职位关系staff，拥有品牌关系havebrand，品牌竞争关系havecompetition。

Neo4J使用的查询语言是cypher语言，简称cql，句法是match…where…return与sql其实大致上相同，但是在match的时候是以三元组的关系进行描述的，比如(n:Company)-[r:invest]-(m:InvestCompany)。在完成Neo4J的知识存储后，可以通过可视化的方式看到企业知识图谱的样子，如图5所示。这种可视化的展示方式也是下一步搭建智能问答平台想要实现的结果可视化的示例。

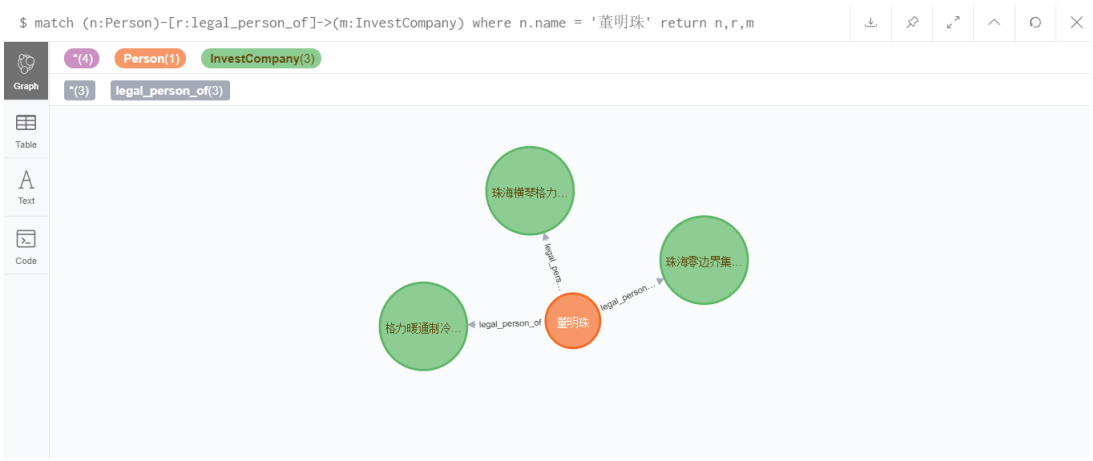

Figure 5: 使用Neo4J来查询语句后返回的可视化结果

3 第三章 企业信息KBQA的搭建与应用

使用Neo4J完成知识存储后，一个持久化的知识图谱已经构建完成了，但是要想对其直接应用，同时保留以点、线的形式的可视化的特征，在此基础上，本课题决定在Web的基础上，搭建一个面向互联网行业的企业知识的智能问答系统。

3.1 可视化智能问答系统的优势

基于企业知识图谱的智能问答平台相比起以往传统的企业信息查询软件拥有以下的几点优势：

1. 智能问答系统支持自然语言的问题的输入，比如“请问格力投资了哪些公司？”这种自然语言，不再是传统的查询模式，需要固定几个参数，然后返回结果。智能问答系统会对自然语言进行语义分析，然后将其转变为cypher查询语句。

2. 以可视化的方式返回结果。如图5,这是在Neo4J里使用查询语句返回的可视化结果，而智能问答系统使用了D3.js这个前台可视化的JS工具，将cypher返回的数据重新以点、线的方式进行了展示，更加还原了知识图谱的场景。而传统的企业信息查询软件，几乎总是以文本的方式返回结果，如

图5的返回结果会变为“格力暖通制冷设备（成都）有限公司、珠海横琴格力商业保理有限公司、珠海零边界集成电路有限公司”，这种文字的展示方式与可视化的结果相比，在用户体验上差了不少。使用D3.js可以轻松实现查询结果的可视化，如图6所示。

Figure 6: 智能问答系统的可视化结果

3.2 基于D3.js的Web前端可视化

因为要实现最后的结果以可视化的方式来展示，尽可能地还原Neo4J中的节点与关系，所以在前端就需要有一个可视化的工具，D3.js刚好提供了一个轻量、方便的前端可视化JS工具。要具体实现前端可视化，就必须前后台结合，相互传递数据。本课题选择的是前端用D3.js来进行可视化的开发，后端使用Java Web来接收前端的request的请求，具体的语义分析和访问Neo4J数据库使用的是Python脚本，因为其访问的方便性。实现的具体步骤如下：

1. 用户从前端输入查询的自然语言，比如“董明珠是哪些公司的法人？”
2. 前端通过request.post的方法向后端发送这条查询语言，以字符串的方式发送
3. 后端通过runtime库来第三方调用本地的Python脚本文件，将自然查询语句以参数的方式传递
4. Python脚本被调用运行后，接收到了自然查询语言，通过语义分析和正则表达式的手段，确定了自然语言问的是哪一种实体关系，以cypher查询语言的语法生成一条cql指令，连接Neo4J，请求查询，Neo4J的查询结果，以数组的方式返回
5. Python脚本对Neo4J发回的数组的数据，进行重新整理格式，整理成为JSON字符串的格式返回给Java程序，Python脚本运行完毕自动退出
6. Java Web的后端收到了来自Python返回的JSON字符串，直接将其设置为参数，也用request的方式发回给前端
7. 前端的JavaScript收到了后端发来的JSON字符串，使用JS自带的JSON.parse()函数将JSON字符串转换为JSON格式对象，并将节点和边的信息分别赋给了D3.js画图所需要的Nodes和Edges两个参数。D3.js以力Force的方式最终画出了节点和线的图，还原了Neo4J的效果。

4 第四章 总结与未来工作

4.1 成果总结

本课题最终成果是设计开发一个基于Web的面向互联网行业的企业知识图谱的构建与应用系统，应用系统的展示是以智能问答的方式来实现的。其包含的主要成果如下：

1. 实现开发了一个可交付的面向互联网行业的企业知识智能问答平台，平台在Java Web上开发，具有很好的可靠性，表现为同一查询语句的返回结果总是相同的，具有良好的稳定性，表现为在使用过程中，不会因为一些设计问题而发生闪退等报错，导致不能继续使用，本系统设计了一个异常处理的方法，即使用户输入出错，也不会导致系统崩溃，而是返回错误类型。

2. 实现了一个面向互联网行业的企业知识图谱的构建，严格按照知识图谱的构建的四个主要步骤进行：知识获取、知识抽取、知识融合、知识存储。知识图谱的设计具有良好的可扩展性，可以扩展新的现存类型的实体，也可以扩展新的类型的实体和关系。为之后的其他应用领域做准备，不局限于智能问答系统的开发，也可以用于风险预测等应用领域。

4.2 未来的工作

下一步的工作将着重解决该系统目前存在的不足之处，然后扩展知识图谱的应用领域，使得搭建的企业知识图谱内容更加丰富和可靠，面向互联网行业知识的智能问答平台更加的高效，拥有更多的功能，将从以下几点出发：

1. 完善智能问答平台现有的功能，使得平台的外观更加的好看，操作更加的易于理解。使用更好的可视化工具，图文并茂地实现问答系统；将自然查询语言的处理和对Neo4J的访问从Python平台迁移到Java平台，并且提升中文语义分析的效率和准确率，使得智能问答平台的结果展示更加的快速、准确。

2. 将企业知识图谱的应用领域扩展，未来将根据企业之间的关系，尝试在风险预测领域应用知识图谱帮助预测。

References

[1] Berant J, Chou A, Frostig R, et al. Semantic Parsing on Freebase from Question-Answer Pairs//EMNLP. 2013, 2(5): 6.

[2] 谭刚,陈聿,彭云竹.融合领域特征知识图谱的电网客服问答系统[J/OL].计算机工程与应用:1-11[2019-09-17].

[3] 邢超. 智能问答系统的设计与实现[D].北京交通大学,2015.