学 号 2014061216

密 级

哈尔滨工程大学学士学位论文

经济责任审计知识图谱构建方法及应用研究

院（系）名 称：计算机科学与技术学院

专 业 名 称：计算机科学与技术

学 生 姓 名： 刘绪光

指 导 教 师： 黄少滨 教授

哈尔滨工程大学

2018年6月

**经济责任审计知识图谱构建方法及应用研究 刘绪光 哈尔滨工程大学**

学 号 2014061216

密 级

经济责任审计知识图谱构建方法及应用研究

Research on the construction method and application of knowledge graph of economic responsibility audit

**学生姓名：**刘绪光

**所在学院：**计算机科学与技术学院

**所在专业：**计算机科学与技术

**指导教师：**黄少滨

**职称：**教授

**所在单位：**哈尔滨工程大学

**论文提交日期：**2018年6月

**论文答辩日期：**2018年6月

**学位授予单位：**哈尔滨工程大学

摘 要

知识图谱是Google为了提高其搜索引擎的体验而提出来的概念，它是多个学科结合的产品，在教育、医疗、金融等领域已经有相关的落地应用了。经济责任审计是我国独有的审计模式，用以对党政主要领导干部和国有控股企业领导人员在任职期间财政收支、财务收支和相关经济活动进行的审计。但是经济责任审计知识比较分散，相关数据库完备性不高、信息化程度不够，这些问题给审计和被审计人员带来了极大的不便，针对这一情况，本文构建了经济责任审计知识图谱，方便相关人员获取经济责任审计知识。

本文完成了经济责任审计这一特定领域的知识图谱构建工作。在构建过程中利用网络爬虫技术、html解析技术从相关机构网站、百科网站获取知识；通过实现基于TensorFlow和卷积神经网络中文文本分类模型来挑选领域词汇，保证获取知识的合理性与可用性；采用基于依存句法分析的开放式中文实体关系抽取方法完成实体关系抽取工作；最后使用开源图形数据库neo4j存储领域知识；本文同时实现了对已构建知识图谱的简单应用，使用Django实现了一个web项目，提供了实体查询和关系查询的功能，可以快速定位经济责任审计知识及它们之间的关系。

关键词：知识图谱；经济责任审计；实体关系抽取

**ABSTRACT**

Knowledge graph is a concept proposed by Google in order to improve the experience of its search engine. It is a combination of multiple disciplines and has been applied in fields such as education, medical care, and finance. Economic responsibility audit is an unique audit model in China that is used to audit financial revenues and expenditures, financial revenues and expenditures, and related economic activities of leading party and government leading cadres and leaders of state-controlled enterprises. However, the knowledge of economic responsibility auditing is relatively fragmented, the completeness of related databases is not high enough, and the level of information is not enough. These problems have caused great inconvenience to the audited and audited personnel. In view of this situation, this paper has constructed an economic- responsibility audit knowledge graph. To facilitate the relevant personnel to obtain economic responsibility audit knowledge.

This paper completed the work of constructing a knowledge graph of the specific area of ​​economic responsibility audit. During the construction process, web crawler technology and html parsing technology are used to obtain knowledge from related institution websites and encyclopedia websites, and domain vocabularies are selected by implementing TensorFlow and convolutional neural network Chinese text classification models to ensure rationality and availability of acquired knowledge. An open Chinese entity relationship extraction method based on dependency-syntax analysis completes the entity relationship extraction; finally uses the open source graphics database neo4j to store domain knowledge. This article also implements a simple application of the constructed knowledge map, using Django to implement a web project, providing functions of entity query and relation inquiry can quickly locate the economic responsibility audit knowledge and the relationship between them.

**Key words:** Knowledge graph; Economic responsibility audit; Entity relationship extraction

目 录

[第1章 绪论 1](#_Toc515108762)

[1.1 研究背景及意义 1](#_Toc515108763)

[1.2 国内外研究现状 2](#_Toc515108764)

[1.2.1 经济责任审计研究现状 2](#_Toc515108765)

[1.2.2 知识图谱研究现状 2](#_Toc515108766)

[1.3 课题主要研究内容 3](#_Toc515108767)

[1.4 章节安排 4](#_Toc515108768)

[第2章 相关理论分析 5](#_Toc515108769)

[2.1 知识图谱分析 5](#_Toc515108770)

[2.1.1 知识图谱定义及架构 5](#_Toc515108771)

[2.1.2 知识图谱类型 5](#_Toc515108772)

[2.1.3 知识图谱构建方法 6](#_Toc515108773)

[2.1.4 知识图谱构建流程 7](#_Toc515108774)

[2.2 网络爬虫技术分析 7](#_Toc515108775)

[2.3 领域词汇判定 8](#_Toc515108776)

[2.4 卷积神经网络介绍 9](#_Toc515108777)

[2.5 实体关系抽取分析 11](#_Toc515108778)

[2.6 图数据库理论与方法 11](#_Toc515108779)

[2.6.1 图数据库概念 11](#_Toc515108780)

[2.6.1 Neo4j图形数据库 12](#_Toc515108781)

[2.7 本章小结 13](#_Toc515108782)

[第3章 图谱构建方案设计 14](#_Toc515108783)

[3.1 数据获取方案 14](#_Toc515108784)

[3.1.1 百科类网站爬虫 14](#_Toc515108785)

[3.1.2 新闻类网站爬虫 16](#_Toc515108786)

[3.2 领域词汇判断模型 17](#_Toc515108787)

[3.3 实体关系抽取 19](#_Toc515108788)

[3.4 图谱应用设计 20](#_Toc515108789)

[3.5 本章小结 21](#_Toc515108790)

[第4章 图谱构建及其应用 22](#_Toc515108791)

[4.1 知识获取 22](#_Toc515108792)

[4.2 领域词汇判定实现 23](#_Toc515108793)

[4.3 实体关系抽取实现 25](#_Toc515108794)

[4.4 知识入库 28](#_Toc515108795)

[4.5 知识图谱应用 29](#_Toc515108796)

[4.6 本章小结 33](#_Toc515108797)

[结　　论 34](#_Toc515108798)

[参考文献 37](#_Toc515108799)

[攻读学士学位期间发表的论文和取得的科研成果 40](#_Toc515108800)

[致　　谢 41](#_Toc515108801)

# 

# 第1章 绪论

## 1.1 研究背景及意义

经济责任审计是指企事单位的法定代表人或经营承包人在任期内或承包期内应负的经济责任的履行情况所进行的审计[1]。经济责任审计的主要目的是分清经济责任人任职期间在本部门、本单位经济活动中应当负有的责任，为组织人事部门和纪检监察机关和其他有关部门考核使用干部或者兑现承包合同等提供参考依据[2]。

现在，很多单位会在领导干部离任时对其进行经济责任审计（离任审计[3]），审计结果会作为领导干部升迁的重要依据[4]，被审计人员会受到经济责任审计的极大影响。而随着国家反腐力度的加大，经济责任审计也越来越受到社会的关注，实现审计监督的全覆盖、建立任期内轮审制度[5]势在必行。

因为经济责任审计范围的扩大和被审计次数的增多，使得越来越多的审计人员和被审计人员对相关知识的需求增强，而经济责任审计知识又是极为分散的，虽然可以通过查阅相关政策法规文件、使用搜索引擎获取知识，但是这些手段并不能精准定位用户所需要的知识，需要用户用自己的专业知识去甄别有效信息，极大的浪费了用户精力与时间。

当前知识图谱主要分两种，一是面向开放领域的知识图谱，二是特定领域的知识图谱。面向开放领域的知识图谱都有大规模的知识做支撑，比如谷歌知识图谱、百度知心、搜狗知立方等都是面向开放领域的；随着大数据时代的到来，知识图谱在医疗、金融、教育等领域已经有所应用，这些领域的知识获取难度相对较小，而且数据大多以结构化或半结构化形式组织，提取难度相对小，所以构建这一类知识图谱也相对容易。

在现有的知识图谱产品中，还没有经济责任审计这一特定领域图谱，虽然审计概念早就有了，但是经济责任审计出现的比较晚，现有的金融、企业财务等领域知识图谱和经济责任审计存在少量的交叉。目前对于经济责任审计知识图谱构建的研究也比较少，领域知识的极度分散导致构建经济责任审计知识图谱的难度加大，本文将利用现有知识构建经济责任审计知识图谱，更好的组织领域知识，为后期的经济责任问答系统[6]奠定基础。

经济责任审计在遏制腐败的发生、评价任用干部等方面发挥着不可估量的作用，经济责任审计是加强领导干部的管理监督、促进领导干部依法行政的重要手段[5]。通过本文经济责任审计知识图谱技术将会极大的减轻审计和被审计人员的记忆负担，降低经济责任审计知识获取难度，帮助审计和被审计人员学习和掌握经济责任审计知识，不仅方便被审计人员规范自己的行为，也方便审计人员选择合理的审计方法对领导干部进行审计。因此，本文所构建的经济责任审计知识图谱是十分有意义的，它为日后的经济责任审计智能问答系统提供数据支持，为今后的科研任务打了良好的根底。

## 1.2 国内外研究现状

由于经济责任审计知识覆盖范围大，而且知识收集困难，目前没有对经济责任审计知识图谱的相关研究，而经济责任审计又作为我国独有的一种审计模式，更不可能在国外出现与经济责任审计的相关研究了，下面将分别就经济责任审计研究现状与知识图谱研究现状作简要介绍。

### 1.2.1 经济责任审计研究现状

经济责任审计是我国独有的审计形式，作为我国腐败监督体系的重要组成部分，对于腐败治理有着不可估量的作用。因为经济责任审计出现的时间比较晚，所以相关的法律法规都需要进一步完善。经济责任审计已经从政府部门扩展到了其他各个机构，因此经济责任审计的重视程度也变得越来越高。目前的研究主要集中在经济责任审计评价指标、量化标准[7]等方面，而针对经济责任审计领域基础知识的研究反而被边缘化了。现在也没有针对经济责任审计知识图谱的研究，虽然已经有人提出了经济责任审计要和大数据结合的观点[8]，但是仅仅是从理论层面进行分析，并没有相关的技术实现。

从上文可以看出，目前没有针对经济责任审计的知识图谱研究，因此本文所构建的经济责任审计知识图谱是具有开创性意义的。

### 1.2.2 知识图谱研究现状

“Knowledge graph”[9]是谷歌于2012年5月提出的概念，并宣布将以知识图谱为基础构建下一代智能化搜索引擎，表1.1中列举了部分知识图谱相关产品，多半属于国外产品，因此相比国内，国外对于知识图谱的研究是走在前列的。近几年，国内的学者也陆续开始了知识图谱的研究，并将之与医疗、商业、教育等领域结合，取得了一定的进展，复旦大学、清华大学、上海交通大学纷纷发布了自己的研究成果。伴随谷歌知识图谱系列产品的不断上线，以及相关产品表现出来的良好体验效果，知识图谱技术引起了学术界的广泛关注，但是目前还没有特别成熟的产品，在使用谷歌搜索引擎搜索知识时，也只有少部分关键词会在右侧显示其详细信息，大部分依旧是返回已经排好序的网页链接供用户筛选，表1.1列出了现有部分知识图谱产品[10]。

表1.1 现有知识图谱部分产品

|  |  |  |
| --- | --- | --- |
| Knowledge Base | Products | Data Source |
| Konwledge Graph | Google Seach Engine  Google Now | Wikipedia,Freebase,Web Open Data |
| Wolfram Alpha | Apple Siri | Mathematica |
| Sarori/Probase | Bing Seach Engine | Wikipedia,Web Open Data |
| Watson KB | IBM Watson System | Web Dictionaries,The World Book,Encyclopedia |
| DBpedia KB | DBpedia | Wikipedia |
| YAGO KB | YAGO | Wikipedia |
| NELL KB | NELL | Web Open Data |
| Facebook KB | Shopycat | Social Network Data |
| Zhilifang KB | Sougou Seach Engine | Web Open Data |
| Zhixin KB | Baidu Zhixin Platform | User Generated Content |
| Cross-Lingual KB | XLORE | Chinese/English Encyclopedia,Wikipedia |
| Zhishi.me KB | Zhishi.me | Chinese Encyclopedia |

上面这些知识图谱相关产品都属于领域无关知识图谱，并不能够直观的展现经济责任审计相关知识之间的联系，因此很有必要构建经济责任审计这一限定领域的知识图谱，为审计和被审计人员提供快速的知识获取渠道。

目前商业领域对知识图谱的应用有三种，分别为搜索引擎、智能问答系统和聊天机器人，对知识图谱的应用大多集中在医疗、金融、教育行业，因此本文所构建的经济责任审计知识图谱，不论从学术角度还是商业角度，都有极强的现实意义。

## 1.3 课题主要研究内容

从上文的分析可以看出，目前没有经济责任审计这一限定领域的知识图谱，本文的目标是构建一个相对完善的经济责任审计知识图谱，然后以已构建知识图谱为基础，实现对知识图谱的简单应用。目前没有便利的获取经济责任审计知识的手段，因此本文的主要研究内容是如何利用现有网络开放知识库构建经济责任审计知识图谱。从技术上来看，难点集中在网络爬虫技术、领域文本分类、实体关系抽取、知识应用。主要研究内容如下：

1. 使用网络爬虫技术获取领域相关知识；
2. 通过深度学习技术实现领域知识的筛选工作；
3. 实现基于依存句法分析的开放式中文实体关系抽取方法；
4. 将知识表示成三元组形式并采用图数据库存储；
5. 对已构建知识图谱进行适当的简单应用，实现一个简易搜索引擎。

## 1.4 章节安排

第1章为绪论，主要介绍课题研究背景及意义、国内外研究现状以及课题主要研究内容；

第2章为相关理论分析，主要介绍本文在知识图谱构建过程中所使用到的技术理论知识，包括知识图谱理论、卷积神经网络、实体关系抽取等，并结合本文实际情况进行理论分析，选择切合实际的理论基础；

第3章为图谱构建设计，基于第2章所分析的理论知识对本文所实现的各部分任务进行了详细的设计；

第4章为图谱构建及其应用，主要介绍了构建经济责任审计知识图谱的具体工作及对知识图谱的应用；

论文最后的部分是对本篇论文的总结与结论。

# 

# 第2章 相关理论分析

本章对经济责任审计知识图谱构建过程中所用到的技术理论进行简要的介绍和分析，介绍内容包括知识图谱相关技术、网络爬虫技术、领域词汇判定、卷积神经网络、实体关系抽取和图形数据库的理论与方法，通过这些理论的介绍与分析，可以大致掌握本文的理基础，以及为什么选择这样的理论来支撑知识图谱构建及应用任务，对理解第3章的设计和第4章的实现有巨大的帮助。

## 2.1 知识图谱分析

### 2.1.1 知识图谱定义及架构

Google在2012年提出的知识图谱是从现有概念演变过来的，大致经历了语义网络、本体论、万维网、语义网、链接数据、知识图谱，共6个概念的演变。知识图谱是结构化的网状知识库，它的基本单位是“实体—关系—实体”三元组，通过关系将具备属性的实体连接起来[10]，形成一种概念网络，网络中的结点就是现实世界的概念，实现了对物理世界符号表达，概念之间的依靠关系进行连接。以数据结构的角度来看，知识图谱中的实体就是结点，图的边则由不同的语义关系组成，因此知识图谱为新的信息检索方式提供了支撑，实现从网页链接到概念链接的巨大转变。

知识图谱的架构分为自身逻辑结构和构建过程所采用的技术架构。一般将知识图谱分为数据层和模式层两个层级[10]，数据层中存储的是以知识为事实三元组为单位的知识，一般使用图数据库进行存储；为了提炼知识及方便管理，在数据层之上建立了模式层，常见的是使用本体库管理数据库中的知识[11]，也可以使用本体库的约束条件来规范实体、关系以及实体的属性，属于同一个本体库的实体应该是相同的类型、具备相同的属性，百度也提供了免费的知识图谱校验工具，本文所有的实体只使用了一个类进行组织，因此不需要使用本体进行约束。

### 2.1.2 知识图谱类型

现在用于商用的知识图谱主要分为以下几种类型：

（1）领域无关知识图谱，这种类型的知识图谱一般不限于特定领域，它尽可能的包含更多的知识，内容比较广泛，谷歌知识图谱就是一个典型的领域无关知识图谱，百度所构建的也是一个领域无关的知识图谱[11]，这种知识图谱常使用本体对知识进行分类校验。

（2）特定领域知识图谱，这类知识图谱都是针对某一特定领域，其内容没有领域无关知识图谱广泛，但是在特定领域的知识会更加全面而深入[11]，更加具有针对性，现在热门的医疗、金融、教育知识图谱都是特定领域知识图谱。

（3）跨语言知识图谱，这类知识图谱包含了多种语言的描述，比如Wikidata[12]、Freebase[13]都属于跨语言知识图谱。

本文所构建的知识图谱属于第二种类型，即经济责任审计这一特定领域的知识图谱，所有知识都是以经济责任审计为中心进行扩展的。

### 2.1.3 知识图谱构建方法

自顶向下方法和自底向上方法是构建知识图谱的两种方式。自顶向下常从百科类网站等结构化的数据中提取知识并加入到知识库中；自底向上则通过一定的技术手段，从公开的非结构化数据中提取知识，选取置信度较高的知识加入知识库中。本文同时借助了百科类网站的半结构化数据，和新闻类网站的非结构化数据进行构建。下面介绍了按数据来源进行区分的四种不同构建方法。

（1）基于网络百科资源的知识图谱构建方法，这种方法通常会从互动百科、维基百科、百度百科等资源中抽取自己所需的知识[11]。以互动百科为例，它提供了百科知识分类树，可以很方便的根据这棵分类树查找自己所需的知识，每一个词条都是一个单独与之对应的页面内，该页面首先是词条名称、然后会列出该词条的详细解释，紧接着会列出词条的其它属性信息，通过html解析技术就可以提取出其中的值-属性对信息，而且不需要添加任何规则就可以很方便的提取知识，达到了快速高效的理念。

（2）基于结构化数据的知识图谱构建方法，知识图谱主要的工作就是将现实世界复杂的信息变成结构化的数据，方便计算机的存储，知识的一个重要来源就是结构化数据，因为结构化数据的质量高[14]，易于处理，所以首先都会考虑将结构化知识加入知识库中。现在广泛采用的结构化数据有Yago[15]、DrugBank[16]等知识库。

（3）基于半结构化数据的知识图谱构建方法，现在有很多中文百科类网站提供了大量的半结构化知识，由于我国互联网起步较晚，中文百科网站的结构化程度还远远赶不上维基百科，在表格中隐藏了很多的实体信息，为了提取这些信息就需要使用出具抽取技术处理半结构化数据。一般我们可以通过观察具体网站来针对性的编写知识提取程序，因为同一个百科类网站的数据组织方式基本都是一样的，我们只需要找到它组织数据的方式，然后再通过定义一些规则就可以实现了知识的抽取。

（4）基于非结构化数据的知识图谱构建方法，这是目前学术前沿的研究方向，因为对于特定领域的知识图谱，仅仅依靠百科类数据资源无法满足需求，所以还需要从非结构化数据中提取领域知识。与前文三种数据来源相比，现在从非结构化数据中获取知识主要还是依靠人工，使用自动化或半自动化程序抽取知识的难度非常大，因此应用并不不广泛，对非结构化数据的处理对技术的要求非常高，而且现在对非结构化数据进行处理的理论知识也极度欠缺，而对中文领域开放文本的处理也并不成熟，可想而知使用非结构化数据构建知识图谱难度之大。

本文同时使用了（1）、（3）和（4）的结合，因为经济责任审计知识的比较分散，网络中可得到的相关知识也很少，所以如果仅仅使用一种方式，会造成所构建的知识图谱不够完善，本文的实体信息是基于互动百科构建的；实体关系一部分来自于百科网站知识库，一部分是从半结构化和非结构化数据中提取的。

### 2.1.4 知识图谱构建流程

不同的科学家对知识图谱的绘制流程划分不一样，可归结为以下几个步骤：

（1）知识获取，高质量的原始语料是构建优质知识图谱的基础，一般采用网络爬虫技术获取网页内容作为原始语料。

（2）信息抽取，信息抽取即从已获取的语料中提取可用的知识，包括实体抽取、属性抽取、关系抽取，信息抽取是知识图谱构建的技术瓶颈，该阶段最常用到的技术是自然语言处理。

（3）知识融合，知识融合主要是对已抽取信息中存在的冗余、错误进行处理，比如共指消解、实体消歧、实体连接等。

（4）知识表示及存储，从原始语料中提取出了可用的高质量知识后，下一步面临的就是知识的表示及存储，可以使用本体、三元组等形式表示知识；由于知识图谱的特殊性，使用关系型数据库进行存储存在一定的缺陷，大多数采用图形数据库存储。

## 2.2 网络爬虫技术分析

网络爬虫（Web Crawl），又称为网络蜘蛛（Web Spider）或Web信息采集器[17]。它是搜索引擎的一个重要组成部分，可以自动的从网页中下载想要的内容。一般网络爬虫从种子集开始爬行，在爬行过程中会遇到新的URL并存入待爬行URL集合中，直到待爬行URL集合为空或达到某个终止条件才结束爬虫程序；爬虫程序中根据实际需要，用于存储待爬行URL的数据结构可选择栈或队列，使用栈存储待爬行URL是web数据进行深度优先遍历，用队列则是进行广度优先遍历。

因爬虫技术可以快速获取相关知识，在数据挖掘领域，网络爬虫是不可或缺的数据获取手段。但是网络爬虫同样也会给被爬虫网站的服务器造成极大的压力，大量的快速访问势必导致服务器存在宕机的危险，因此很多网站都采取了不同程度的反爬虫手段，保证网站后台的健壮性。另一方面，网络中大部分的页面其实是重复的，客户端和服务端脚本语言的使用导致指向相同页面的URL呈指数级增长。这些特征都给网络爬虫造成了一定的困难，主要体现在单位时间内获取的可用知识少、容易被服务器识别出爬虫程序导致自己的IP被禁止访问。

网络爬虫分为四种类型，分别为通用网络爬虫、聚焦网络爬虫、增量式网络爬虫、深层网络爬虫[17]。而在实际的运用过程同，通常是根据需要将不同的类型进行结合。（1）通用网络爬虫常用于商业领域，因此它的技术细节公开资料较少。一般是从一些种子URL扩充到整个web，这类网络爬虫对爬行速度和存储空间要求较高，而对于页面的顺序相对要求较低，通常采用并行工作方式[18]，搜索引擎和大型Web服务提供商的数据采集工作可以使用通用网络爬虫；（2）聚焦网络爬虫也称为主题网络爬虫，它会筛选页面，只爬行与主题有关的页面，相比其他类型网络爬虫，聚焦网络爬虫多了一个评价模块，用于给链接和内容打分，而且也节省了网络与硬件资源，适用于特定领域的信息获取；（3）增量式网络爬虫是增量式的对已经下载的网页进行更新，即只爬行新产生或发生过变化的网页[17]。它的优点在于，可以保证爬取的页面都是新的，可以减少数据下载量。当然，其算法的复杂度和实现难度也会相应增加；（4）深层网络爬虫是爬行那些不可以通过静态链接访问的页面，比如一些要求用户注册后才可以发现的网页内容。深层网络爬虫可访问的数据是表层网页的几百倍，是互联网最大的信息资源。

本文所有知识都是从网络开放数据中获取的，获取渠道包括审计署门户网站、纪检委门户网站、新浪新闻、百度百科、互动百科以及wikidata。针对不同的网站需要编写与之对应的爬虫程序，就本文的实际情况，可以划分为两大类：新闻类网站爬虫和百科类网站爬虫。因本文需要对抓取内容进行筛选，所以通用网络爬虫不适于本文的任务，聚焦网络爬虫以及增量式网络爬虫是本文采用的方式。

## 2.3 领域词汇判定

领域词也称为某个领域的术语，领域词具有较强的特征表示性。领域词识别是本文的关键任务之一，因为本文是面向经济责任审计这一特定领域，需要去除那些与本领域无关的词汇。目前领域词判定主要有以下三种办法：

（1）基于规则的方法，这种方式依赖人工构建领域词库，通过人工对领域词的规则进行发现和总结，是一个极为低效的方法，另一方面，不同个体的观点不同，不能保证统一性，目前运用较少。

（2）基于统计的方法，这种方法利用左右信息熵、词频分布等统计量来对候选词进行评分排序，得到评分在指定范围内的词汇作为领域词。这种方法目前使用较多，可以结合多个特征进行统计，适用于大规模语料库的统计，但是这种方式对语料库的质量要求比较高。

（3）规则和统计结合的方法，这种方法同时结合了规则和统计的优点，候选领域词汇使用统计方法得到，然后再定义具体规则进行筛选，这是目前的研究热点，效果明显。

本文是基于卷积神经网络技术进行领域词汇判定，结合世界情况，本文已经有了百科类网站的数据，这些数据中对每个词条都有详细的解释说明，另一方面单个词汇的可提取特征也比较少，所以本文的工作偏向于使用卷积神经网络实现的中文句子分类。卷积神经网络将在下一节进行介绍。

## 2.4 卷积神经网络介绍

卷积神经网络是一种前馈神经网络[19]，它的设计目的是为了处理图像和分类任务，常由一个或多个卷积层以及全连接层组成，最开始在图像处理领域取得非常不错的成绩，近年来在自然语言处理领域也表现出了其强大的功能，本文使用卷积神经网络完成中文文本的分类任务。

卷积神经网络的基本单元是神经元，每个神经元都是有若干个输入的运算单元，其对应的公式如式（2-1）所示：

（2-1）

其中是神经元激活函数，激活函数可以加入非线性因素，进而解决线性模型表征能力弱的缺点，常用的激活函数如表2.1所示，当输入达到一定条件时，神经元会被激活，常用的有双曲正切函数、Sigmoid函数、Relu函数等[20]。

卷积神经网络中的卷积又是指分析数学中的一种运算操作，是两个函数乘积的积分，卷积运算定义式如（2-2）所示：

（2-2）

其中x(.)是观测函数，w(.)是权重函数，t代表观测时间。卷积在自然语言处理中是将一个文本中相邻词语的向量拼接为一个新的向量，然后再对这个词向量做线性映射得到新的向量，即原来相邻词语的一个特征表示。

表2.1 常用激活函数

|  |  |  |  |
| --- | --- | --- | --- |
| 名称 | 公式 | 优点 | 缺点 |
| Tanh |  | 可以不断增强特征 | 存在梯度饱和问题 |
| Sigmoid |  | 反向传播计算简单 | 存在梯度饱和或者消失的问题 |
| Relu |  | 可以减轻梯度弥散，不会产生梯度消失 | 容易产生死神经元 |

参数共享和局部感知是卷积神经网络的核心，其结构分为特征提取层和特征映射层。卷积神经网络使用多个有固定窗口的卷积核提取文本中的局部信息，对这些信息进行整合得到多个特征映射，就是文本的特征。本文就是利用文本的特征来筛选领域词汇。

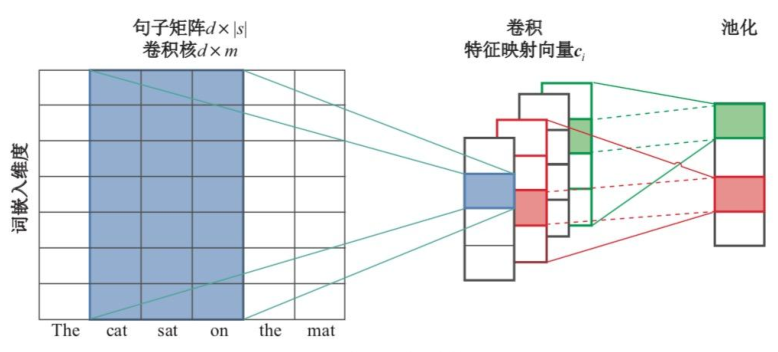
卷积神经网络结构如图2.1[21]所示：

图2.1 卷积神经网络结构

从图2.1可以看出，卷积神经网络主要由卷积层和池化层组成。

卷积核是卷积层的核心，每个卷积核实际上是一个权值矩阵，矩阵中每一个值实际上就是这个提取特征时该点所对应的权重，通常会有多个卷积核用于同一个卷积层。每个神经元只跟前一层输入的一个局部区域相连接[19]，这个区域的大小就是卷积核的大小，采用卷积计算公式进行卷积的计算。

池化层也叫做下采样层，即使用卷积层的计算结果进行分类。之所以采用池化层，是因为如果把所有解析出来的特征都关联到一个分列器，那样的计算量将无比巨大，因此是不明智的。池化层最大的作用就是把特征区域求最大值或者均值后才表示这个区域，本文所采用的是求最大值，即max pooling。

## 2.5 实体关系抽取分析

实体关系（事实三元组）抽取是自然语言处理领域一直以来的研究热点和难点，因为其面向的是开放的文本，毫无规则，所以抽取难度极大。目前常用的实体关系抽取方法主要有三种：

（1）人工抽取，人工抽取是目前最可靠的方式，自然语言本身就是人类发明的，人类可以很轻松的理解文本数据，但是机器无法理解，所以人工抽取的事实三元组是目前认为最可靠的。但是其缺点也很明显，需要大量的人力资源支撑，同时也耗时耗力，效率底下是最大的缺点。

（2）基于深度学习的实体关系抽取，这种方法目前是既高效又相对可靠的抽取方式，它通过大量的数据进行训练，最后用训练所得的模型抽取实体关系，在效率上得到了极大的提高，同时所提取的实体关系质量也能和人工媲美。这种方式最大的问题是训练集的标注，对训练集的要求很高，而且需要大量的训练集才能得到好的效果。训练集的标注是症结所在，所以本文没有采用这种方法。

（3）基于依存句法分析的实体关系抽取[22]，这种方法是一种基于规则的抽取方法，首先把句子分析成一棵依存句法树，然后再通过规则遍历这棵依存句法树得到实体关系。而且对于依存句法分析的研究现在已经比较成熟，本文直接采用开源汉语言处理包HanLP提供的依存句法分析接口得到依存句法树。

Zhang等人[23]采用了基于实例的无监督学习方法，所得的实验效果也比较好，本文参考了文献[22]所提供的方法，使用java语言实现了实体关系抽取程序，并参考文献[24]所提出的方法，实现了非动词性关系的抽取。相比较第（1）、（2）种方法，基于依存句法分析的方法不需要大量的训练集标注，也不需要大量的人力资源，是本文唯一的可行方案。

## 2.6 图数据库理论与方法

### 2.6.1 图数据库概念

数据存储一直是人类长期研究的重要领域，关系型数据库[25]具备扎实的理论基础，它具备较高的安全性，而且经过多年的研究发展，关系型数据库的技术已经非常成熟了，它可以依赖简单的二维表结构来表达丰富的语义信息。但是随着数据量的加大，关系型数据库也表现出了其一直未能解决的瓶颈，那就是速度开始变慢，尤其是使用关联查询时，将耗费大量的资源，性能随之也会变差。另一方面，关系型数据库的事务机制也占用了大量的资源，因此越来越多的学者开始了对非关系型数据库（NoSQL[26]）的研究。

图数据库[27是非关系型数据库的一种，它可以很方便的将点、线、面等元素按照一定数据结构存储下来，如果需要表示不同事物之间的联系，关系型数据库需要进行关联查询才能实现，而图形数据库的优点在于它能直观的表示事物之间的关系。图数据库包括两种基本元素：结点（Node）和关系（Relationship）。使用结点表示现实世界的事物，即实体，实体可以有自己的属性（Property）；使用关系表示结点之间的联系，即实体与实体之间的联系，关系也可以具备自己的属性。

### 2.6.1 Neo4j图形数据库

Neo4j[28]数据库是一个高性能的NoSQL图形数据库，它将结构化数据存储在网络而不是二维表中，是一个嵌入式的、基于磁盘的、具备完全的事务特性的Java持久化引擎[29]。Neo4j具备以下四个特征：

（1）高可用性：它可以很方便的集成到任何应用中，不会受到具体业务的约束。

（2）易扩展性：如果单个结点无法满足数据需求时，Neo4j的分布式集群部署支持可以解决这一问题。

（3）完整的数据库事务支持：数据库事务正确执行的四个基本要素是原子性、一致性、隔离性、持久性。Neo4j可以完整的支持这四个特性，保证数据的准确性。

（4）快速检索：可以通过Neo4j的遍历工具进行数据的快速检索。

Neo4j数据库是在java虚拟机的基础上进行开发的，因此要使用neo4j数据库的前提条件是安装了jdk，本文所使用的jdk版本是1.8。它的安装方法很简单，只需要根据提示同时neo4j也提供了一个web版操作页面供用户使用，默认访问地址为“127.0.0.1:7474”。

Cypher是Neo4j的官方查询语言，是一个类SQL语言，可以方便的对图形数据库进行查询和更新[30]。与SQL相比，Cypher语法简单，而且功能强大，很多SQL种无法解决的任务都可以用它来实现，它常用的子句有：

MATCH子句：通常用于匹配数据库中的数据。

WHERE子句：WHERE子句一般作为MATCH子句一部分，用于指定查询的条件。

RETURN子句：指定查询需要返回的内容。

CREATE子句：用于创建结点、关系、属性。

除了Cypher查询语言外，为了方便数据的导入Neo4j还支持数据的批量导入，同时也为java、python、javascript等语言提供了对应的API，方便编程人员直接调用和开发，本文所使用的是Neo4j为python提供的查询接口。

## 2.7 本章小结

本章主要介绍了知识图谱架构、类型极其大致构建方法与流程，并将构建知识图谱过程中的四个重要步骤：知识获取、信息抽取、知识融合、知识表示及存储，映射为对网络爬虫技术、实体关系抽取和领域词汇判定、图数据库的介绍，而领域词汇判定又使用了卷积神经网络进行实现，所以也对卷积神经网络做了简要的介绍。下一章将根据本章介绍的理论知识进行详细的应用设计。

# 第3章 图谱构建方案设计

本章将对经济责任审计知识图谱构建过程中所使用到的算法、方案提供理论可行的设计方案，结合第2章所介绍的理论知识，对本文实现的任务进行设计，为相关任务的实现奠定基础。

## 3.1 数据获取方案

所有数据均来自于网络，使用网络爬虫程序进行数据的抓取，主要涉及到新闻类网站和百科类网站的爬虫，下面将分别进行介绍。

### 3.1.1 百科类网站爬虫

目前没有经济责任审计领域词库，因此本文考虑借助网络百科全书平台获取经济责任审计领域词汇，而且网络百科全书拥有上亿用户的智慧，大部分都是准确且严谨的，很适用于领域词库的数据来源。

具体的数据来源包括：互动百科、百度百科、wikidata四个平台，他们的组织结构相似，此处利用百度百科举例说明。百度百科词条信息如图3.1所示。



图3.1 百度百科词条信息

百度百科的词条页面包括了词条的解释、词条其他属性。词条的解释十分详细，同时还带有外部链接词汇，在图中表现为蓝色超链接字体。这些蓝色字体可以作为领域候选词汇，与大多数爬虫不同的是，本文采用爬行过程中就进行了领域词汇的判定，而不是在爬虫之后进行领域词汇判定。

对百科类网站进行爬虫，首先需要人工设定种子词汇，利用图3.1中的特性，在爬行过程中把新的词汇加入到待抓取集合中，直到递归结束。初始种子词汇来源于经济责任审计相关法律法规。对这些法规使用HanLP进行分词、去除停用词、去重，最后再人工筛选，得到种子词汇库。利用网络爬虫技术抓取这些种子词汇对应的页面，如果有蓝色超链接字体，则将对应的词汇加入到待抓取集合中，有一些词汇在百科类网站无法检索到，这类词汇直接放入到领域词库中，为了提高抓取的精准度，本文采用的是广度优先遍历算法。

百科类网站爬虫算法描述如下：

**算法3.1：**获取领域词汇

输入：种子词汇队列

输出：领域词及其属性

1. 定义待抓取链接队列urls，将种子放入改队列
2. 定义已经获取链接列表has\_urls
3. while(urls不为空):
4. 获取urls中第一个链接，记为url
5. 队列urls第一个链接出队
6. if 待抓取url不在has\_urls列表中:
7. 将url加入has\_urls列表中
8. 抓取url对应的页面
9. 对抓取的页面进行解析
10. 利用领域文本判定模型判断是否为领域词汇
11. if 是领域词汇:
12. 存入领域词库
13. 提取页面蓝色字体超链接
14. if 蓝色字体超链接不在has\_urls中:
15. 将蓝色字体超链接加入urls中
16. end if
17. end if
18. end if
19. end while

从上述算法第（9）行可以看出，因为抓取的页面是带有html标签且有大量无用信息的，所以需要对页面进行解析。Python提供了不同的网页解析器，用户可以根据自己的喜好进行选择，本文使用的网页解析器是BeautifulSoup，它可以传入css选择器来定位自己想要的信息，简洁快速是它最大的优点。最后把抓取的信息整理成json格式，方便后期处理。

从百度百科得到领域词汇后，然后直接针对每一个词汇爬取其在互动百科上面的解释信息，本文采用互动百科提供的数据而不采用百度百科的数据，是因为互动百科对每个词条解释更加详细，其属性也更加准确。

### 3.1.2 新闻类网站爬虫

新闻类网站获取的数据主要是为领域词汇判断模型提供数据支持（训练集、验证集、测试集）。本文新闻类网站主要爬行了审计署、纪检委、新浪新闻三个网站，下面以审计署门户网站爬虫进行举例说明。

图3.2为审计署门户网站首页，可以看到在右上角为用户提供了一个搜索框，用于在网站进行全局搜索；以“经济责任审计”为关键词进行搜索，网站会打开一个新的页面用于显示结果，搜索结果如图3.3所示。可以看到页面以超链接列表的形式组织，每一个超链接对应一篇新闻或法律法规的内容，爬虫时直接抓取这些链接对应的页面内容即可。



图3.2审计署首页 图3.3 审计署搜索结果页面

在实际测试过程中，发现政府类网站非常容易被封禁IP，而且使用网站内提供的搜索功能在不同时间段所生成的随机验证串是不一样的，本文使用了chromedriver，它可以模拟用户的操作过程来访问具体网站，可以解决被封IP和站内验证的问题，也能够较直观的展现爬虫过程。

新闻类网站爬虫算法描述如下：

**算法3.2：**获取经济责任审计类文本

输入：待爬虫网站地址

输出：经济责任审计相关文本

1. 加载chromedirver作为浏览器
2. 使用chromedriver打开门户网站首页
3. 人工在右上角输入框输入搜索关键字
4. 点击搜索按钮
5. 获取新打开窗口句柄handle
6. 关闭门户网站首页窗口
7. 定义下一页按钮next-page
8. while next-page不为空值：
9. 抓取搜索页面链接列表为urls
10. for url in urls:
11. 抓取url对应的页面
12. 对页面进行解析
13. 将解析结果存入文件
14. end for
15. 程序模拟点击下一页next-page
16. 更新下一页按钮next-page
17. end while

上述算法中可以看到，第三部需要人工输入搜索关键字，点击搜索按钮之后，算法将会自动开始执行，指导搜索结果全部爬行完毕，即最后一页所有超链接抓取完毕后，算法结束。

## 3.2 领域词汇判断模型

领域词汇判断采用了第二章所介绍的卷积神经网络进行设计。根据所抓取数据的特点，本文直接将使用词条的解释而不是使用词条进行预测，词条的解释是一段文本，相比单个的词条信息，可提取的特征更加丰富，可以提高预测模型的准确率。

模型训练过程中，直接把新闻类网站抓取的文本信息作为训练集，本文实际上是使用卷积神经网络设计了一个文本分类模型[31,32]。最后通过该分类模型进行领域词汇的判定，筛选出属于经济责任审计领域的词条信息。

文字是人类最重要的发明，有了文字之后，可以很方便的传达信息，最重要的是前人的经验都因为文字的发明得以保存，人类依赖文字来传播和记录信息。但是计算机是数值计算工具，无法理解人类世界的复杂语言，因此需要将原始的文字形式转换为计算机可理解的数值形式。

本文使用了常用的词向量来表示我们的语言。比如现在有两个句子“经济责任审计又叫离任审计”和“经济责任审计是一种审计形式”。对这两个句子进行分词处理可得到“【经济责任审计、又叫、离任审计】”和“【经济责任审计、是、一种、审计形式】”。我们取出两个分词集合的并集，可以得到“【经济责任审计、又叫、离任审计、是、一种、审计形式】”。那么原始的两个句子可以采用如下形式进行表述。

【经济责任审计（1）、又叫（2）、离任审计（3）、是（4）、一种（5）、审计形式（6）】

经济责任设计又叫离任审计：【1，2，3，0，0，……】

经济责任审计是一种审计形式：【1，4，5，6，0，……】

可以看出，本文所使用的词向量表达方式和常见的方式不太一样，每个句子被转换成一个长度为600的序列，如果本省序列不够600，那么添0补齐，如果句子长于600，则舍弃600以后的词汇。通过这种方法，可以很方便的将对计算机来说晦涩难懂的自然语言转换为方便理解的数值向量。而且不存在因为数据量变得巨大而导致的词向量变得稀疏的问题。模型的大致结构如图3.4所示。

图3.4 本文分类模型结构

完成中文文本到词向量的转换之后，将词向量作为卷积神经网络模型的训练数据输入，经卷积神经网络提取特征后输出的领域文本预测模型就可以用于本文的领域词汇判断任务。从上图可以清晰的看出本文所设计模型进行分类时所进行的每一步操作，只需要给该模型输入一段文本，模型提取该段文本的特征并判断其所属类别。

## 3.3 实体关系抽取

本文使用基于依存句法分析的开放式中文实体关系抽取方法进行事实三元组的抽取，此方法所抽取的关系表述都是基于谓语动词的，主要分三步：分词、依存句法分词、实体关系抽取。

在英文中，不需要分词，词与词之间有空格隔开，但是中文需要借助分词技术，而且本文是针对经济责任审计这一特定领域的任务，所以需要进行个性化分词。本文所采用的分词工具是HanLP汉语言处理包所提供的分词接口，它可以很方便的导入用户词典就实现了个性化分词，且分词效率高。

目前中文领域的依存句法分析技术相对比较成熟，本文直接借助了哈工大提供的依存句法分析工具。在进行依存句法分析之前，需要先对句子进行分词和词性标注，使用的是HanLP，然后将分词和词性标注结果传入依存句法分析模块，得到依存句法分析结果。然后通过遍历依存句法分析树，根据不同的情况来抽取实体关系。

本文对中文句子中的关系表述分为 7种情况，分别为：主谓宾关系、动补结构、状动结构、状动补结构、定于后置关系、介宾关系和宾语前置结构。关系表述参考文献[14]进行了改进，使用（3-1）文法产生：

*中文关系表述：状语\*动词+补语？*[22] （3-1）

其中\*表示出现0次或任意多次，+表示至少出现一次，？表示出现1次或0次。本文所设计的关系抽取算法较复杂，对每个句子的依存句法分析树进行了重新组织，针对每一个词语都为其存储一个依存字典，构建句法依存列表的算法描述如下：

**算法3.3：**构建依存句法列表

输入：依存句法分析结果parser

输出：词语依存字典

1. 定义依存句法列表list
2. for curword in parser：
3. 定义依存句法字典dict
4. for word in parser：
5. if curword是word的依存依存词语：
6. if 字典已经包含word与curword的依存关系rel：
7. 将word加入键rel所对应的列表中
8. else：
9. 将rel：word对应的键值对加入到依存字典dict中
10. end if
11. end for
12. 将依存字典dict加入到依存句法列表list中
13. end for
14. return 依存句法列表list

可以看出，经过上述算法对依存句法分析结果重新进行组织之后，使得后面的实体关系抽取算法更加容易设计。因为实体关系抽取的情况较多，且每种情况的抽取形式又大致相似，所以此处只用主谓宾做简单举例说明，主谓宾关系抽取算法描述如下：

**算法3.4：**主谓宾关系抽取算法

输入：句法依存字典dict

输出：事实三元组

1. if dict中同时包含键值“主谓关系”和“动宾关系”：
2. 定义关系rel为dict所属词语
3. 定义entity1为dict中键值为“主谓关系”的词语
4. 定义entity2为dict中键值为“动宾关系”的词语
5. return entity1, rel, entity2
6. end if

上述算法只做理解说明使用，是实际使用算法的简化版。可以看出，算法实际是在依存句法分析的结果之上定义了自己的一套规则来抽取事实三元组。

## 3.4 图谱应用设计

当前业界对知识图谱的应用主要有三个个方向，第一种是像谷歌、搜狗一类的搜索引擎，它们以知识图谱为数据支撑，为用户提供了更好的搜索体验；第二种种是像阿里小蜜一类的问答系统，能够快速给用户提供相应问题得精准答案；第三种是微软小冰一类的聊天机器人，聊天机器人和问答系统在某些方面是相似的，不同的是，聊天机器人更关注的是给用户提供一个陪伴服务。

本文实现了简单的知识图谱可视化模块，相当于一个简易的搜索引擎，可以快速的定位经济责任审计知识及它们之间的关系。可视化模块主要是对知识进行展示，主要是展示实体详细信息和实体之间的关系。

实体详细信息展示直接将实体词条的解释展示出来，如果有其他属性也一并添加到页面展示，实体图片信息从网络中获取。

实体关系展示使用javascript插件charts.js实现，使用一个中心结点表示查询的实体，其他结点表示与该实体有关系的所有实体，结点与结点之间的连线表示它们之间的关系，用于直观展示实体之间的关系。除了按实体查询关系，还需要设置可以按关系查询所需知识，即应查询数据库中符合某一关系的所有三元组[33]。

除此之外，还需要在关系图的下方放置一个表格形式组织的实体关系表，用于详细展示实体关系。

## 3.5 本章小结

本章对数据获取、领域词汇判定、实体关系抽取和数据可视化进行了详细具体的设计，对主要使用算法进行了详细的描述，为经济责任审计知识图谱构建做了充足的准备，并做了知识图谱应用的相关分析，结合本文情况，设计了知识图谱应用方案，下一章将实现本章所设计的内容，完成经济责任审计知识图谱的构建工作及知识图谱应用实现

# 第4章 图谱构建及其应用

本章所使用的语言包括java和python3，分别以Intellij IDEA 2017.1.4和JetBrains PyCharm Community Edition 2017.2.3作为编译器，按照第三章的设计进行了实现。经济责任审计知识图谱总体实现路线如图4.1所示。

图4.1 经济责任审计知识图谱技术路线

上图清晰直观的展现了经济责任审计知识图谱构建流程，下面将分别进行详细的阐述说明。

## 4.1 知识获取

本文所有数据均来源于网络绝大部分数据均使用网络爬虫技术抓取得到，训练集中有部分数据来源于清华大学自然语言处理实验室整理的新浪网新闻。网络数据获取所使用的技术主要包括：python3.6、BeautifulSoup网页解析器、scrapy分布式爬虫框架和chromedriver。

在抓取审计署门户网站、新浪新闻这一类新闻网站时，都是使用chrmedriver模拟用户访问。程序首先需要获取一个ChromeDriver实例，为方便理解，可以把ChromeDirver当作一个受程序控制的浏览器，程序中设置了打开ChromeDriver后应该进入的第一个页面。为了尽可能保证所获取的新闻文本都是与经济责任相关的，所以在打开第一个页面后，程序不会停止，而是采用显示等待，直到用户在页面右上方输入搜索关键字，并点击搜索按钮后，网站会自动打开新的页面并显示搜索结果。此时程序将开始自动运行，首先获得第一个窗口的句柄并使用该句柄第一个窗口关闭。程序的结束条件为，搜索结果页面不再有“下一页”按钮，否则将会一直获取当前页面的超链接列表，对列表中的所有链接进行抓取并存储为UTF-8编码的纯文本格式，然后程序会自动模拟用户点击“下一页”，如此循环直到程序结束。

百科类网站由于其数据量大，爬虫时间较久，为了防止网络爬虫程序在爬行过程中由于网络、内存异常、其他未知错误等造成的非正常结束，所以对百科类网站的爬虫使用了scrapy爬虫框架。Scrapy的优点在于它可以根据实际需求方便的进行修改，不仅有web2.0支持，还提供了很多爬虫类型的基类。

对百度百科使用的是增量式爬虫方案，初始的种子词汇由人工手动建立，在爬虫过程中，如果遇到新的相关词汇，将会把该词汇加入到待爬行队列中，直到待爬行队列为空时，爬虫程序运行结束。与普通的爬虫程序不同的是，本文所实现的爬虫程序会在爬行过程中判断当前页面信息是否属于领域文本，如果属于才将对应的词条加入到领域词汇库中并存入文件。

最后使用scrapy爬虫框架针对上文所构建的领域词汇库中每一个词条抓取详细信息，此处是基于互动百科获取实体属性的，为了对页面中不同属性进行抓取，本文使用了BeautifulSoup网页解析器对内容进行筛选，它的用法与css选择器类似，只需要传入选择器字符串，BeautifulSoup就可以返回该选择器所对应的信息，达到了方便快速的效果。

因为本文的实体详细信息全部来自于互动百科的词条解释，所有的词条解释信息都已得到行业专家和大众的认可，所以可以简单的认为本文所构建知识图谱实体信息的准确率为100%。

## 4.2 领域词汇判定实现

领域词汇判定，本文通过卷积神经网络实现，依赖环境为python3.6、TensorFlow 1.3、numpy、scikit-learn、scipy。所使用的数据集分两部分，一部分为从网络抓取的经济责任审计相关新闻文本，另一部分由清华大学自然语言处理实验室根据新浪新闻RSS订阅频道所筛选的新闻文档，所有数据均为UTF-8纯文本格式。本文构建的是经济责任审计领域的知识图谱，所以只需要区分哪些文本不属于经济责任审计领域，因此本次训练只使用了2个分类，即经济责任审计和其他类别。

数据集的划分为：训练集6215篇文本；验证集495篇文本；测试集796篇文本。它们内部审计类文本和其他类文本都各占约50%。

数据预处理阶段会构建使用字符级表示的词汇表，并将这一词汇表存储下来，避免每一次的重复处理，然后再将词汇表转换为字典格式形式存储，其内部存数形式为：{词:id}的格式。除了对文本进行处理，还需要对分类进行处理，本文只使用了两个分类，也将其表示为：{分类:id}的格式，为方便神经网络的训练，此处把所有的文字转换为固定长度的id序列表示，序列长度固定为600。经处理后的训练集是3018\*600的矩阵、测试集是198\*600的矩阵、验证集是398\*600的矩阵；标签数据则分别是3018\*2、198\*2、398\*2的矩阵。

卷积神经网络配置项如下所示：

embedding\_dim = 64 # 词向量维度

seq\_length = 600 # 序列长度

num\_classes = 2 # 类别数

num\_filters = 128 # 卷积核数目

kernel\_size = 5 # 卷积核尺寸

vocab\_size = 5000 # 词汇表达小

hidden\_dim = 128 # 全连接层神经元

dropout\_keep\_prob = 0.5 # dropout保留比例

learning\_rate = 1e-3 # 学习率

batch\_size = 64 # 每批训练大小

num\_epochs = 10 # 总迭代轮次

print\_per\_batch = 100 # 每多少轮输出一次结果

save\_per\_batch = 10 # 每多少轮存入tensorboard

在实际训练过程中，总共经过了7轮迭代才停止，在验证集上的最佳效果表现为93.43%，在测试集上的准确率达到了96.04%。

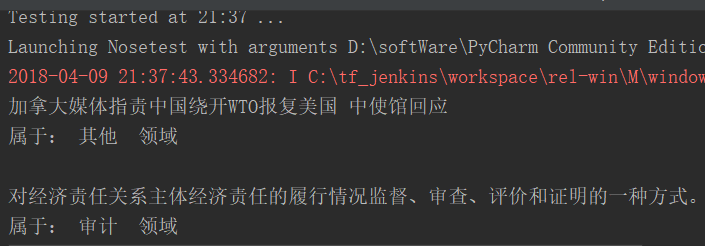
然后将一段文本作为模型的输入，模型就能预测这段文本是否属于经济责任审计领域，如果文本属于经济责任审计领域，那么就认为该文本所对应的词条属于经济责任审计领域，将该词条放入领域词汇库。如图4.2，是领域文本预测模型运行效果。

图4.2 领域文本预测模型运行效果图

从图4.2可以看出，领域文本分类模型分别对“加拿大媒体指责中国绕开WTO报复美国 中使馆回应”和“对经济责任关系主体经济责任的履行情况监督、审查、评价和证明的一种方式。”进行了预测，且预测结果正确，效果理想。本文使用预测模型共得到领域词汇101397条。

因为本文的实体详细信息全部来自于互动百科的词条解释，所有的词条解释信息都已得到行业专家和大众的认可，所以可以简单的认为本文所构建知识图谱实体信息的准确率为100%。

## 4.3 实体关系抽取实现

经分析，基于依存句法分析的开放式中文实体关系抽取是本文唯一的可行方案。在进行依存句法分析之前，需要先对文本进行分词，本文所示用的分词工具为HanLP1.6，它支持用户直接导入用户词典实现个性化分词，使用个性化分词是为了防止分词工具将长词切分为短词，比如“经济责任审计”在导入用户词典之前会被切分为“经济”、“责任”、“审计”三个词汇，然而对于经济责任审计这一特定领域，很明显应该划分“经济责任审计”为单个词汇，所以用户词典中的词汇，都是由领域长词组成。

本文所是使用的用户词典主要来源于权威法律法规，采取人工方式从中挑选领域长词。最后共计挑选了3232个，用户词典部分词汇如表4.1所示。然后分别对101397条词条的详细解释信息和3524条经济责任审计相关的新闻文本进行了分词，然后对分词结果进行词性标注。

表4.1 用户词典部分词汇

|  |  |  |  |
| --- | --- | --- | --- |
| 经济责任审计 | 离任审计 | 财务管理理论 | 损益审计 |
| 审计证据 | 经济责任关系 | 审计机关审计抽样准则 | 审计工作报告 |
| 长期负债审计 | 资金信托 | 注册会计师 | 审计事务所 |
| 任期离任审计 | 中华人民共和国审计署 | 独立审计师 | 资本公积金 |

依存句法分析使用了哈工大LTP语言云提供的依存句法分析接口，通过定义相应的规则对依存句法分析结果进行实体关系抽取，可得到句子中的事实三元组，即“实体-关系-实体”。下面以简单句子“刘小绪喜欢跑步。”做具体说明。图4.3是“刘小绪喜欢跑步”的依存句法分析结果。

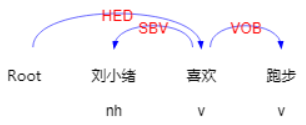


图4.3 依存句法分析结果

从分析结果可以看出，“刘小绪”和“喜欢”构成主谓关系，“喜欢”和“跑步”构成动宾关系。通过第3章所设计依存句法分析列表构建算法，将上述分析结果可以组成如图4.4的形式。

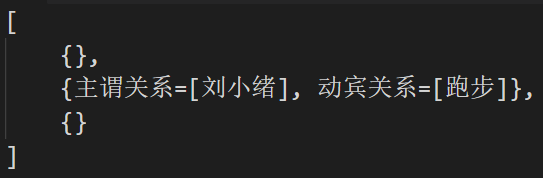


图4.4 依存句法分析列表

然后遍历依存句法分析列表，如果当前词语是动词，则遍历该词语的依存字典，本例中会遍历“喜欢”的依存字典，在遍历过程中会发现主谓关系列表中有“刘小绪”，动宾关系列表中有“跑步”，就可以组成（刘小绪，喜欢，跑步）这样一个事实三元组。事实三元组抽取部分核心代码如下：

if (dic.containsKey("主谓关系") && dic.containsKey("动宾关系")){

String relation = word.LEMMA;

for (CoNLLWord entity1: dic.get("主谓关系")) {

for (CoNLLWord entity2: dic.get("动宾关系")) {

//System.out.println("主谓宾关系==========");

String preEntity = completeEntity(parser, dict, entity1.ID-1);

entityMap.clear();

String rearEntity = completeEntity(parser, dict, entity2.ID-1);

entityMap.clear();

result.add(entity1.LEMMA + "," + relation + "," + entity2.LEMMA);

result.add(preEntity + "," + relation + "," + rearEntity);

}

}

}

上述代码段实现的功能是主谓宾关系三元组抽取，其他结构的三元组抽取与其类似，知识判断条件不同，为节省文章篇幅，故不展示。图4.5是基于依存句法分析抽取实体关系程序的运行效果图。

从图中可以看出，本文所实现程序在大部分结构的中文语句中表现都比较好，准确率和置信度也相对较高。

对于所抽取关系的测试，本文采用了随机验证的方案进行测试。随机抽取了100段文本，然后人工从这100段文本中抽取出实体关系，再使用实体关系抽取程序进行抽取，最后对比两种方式所抽取的实体关系，通过对比计算实体关系抽取的准确率。实体关系准确率计算公式如（4-1）。

（4-1）

通过上述公式，计算所得本文所构建经济责任审计知识图谱的实体关系准确率为79.46%，说明本文所实现的基于依存句法分析的开放中文实体关系抽取方法是可行的方案，验证了第3章设计的正确性，不足之处在于，准确率不高，本文的后续工作将是以经济责任审计知识图谱为基础构建经济责任审计智能问答系统，以及对实体关系抽取程序的优化。

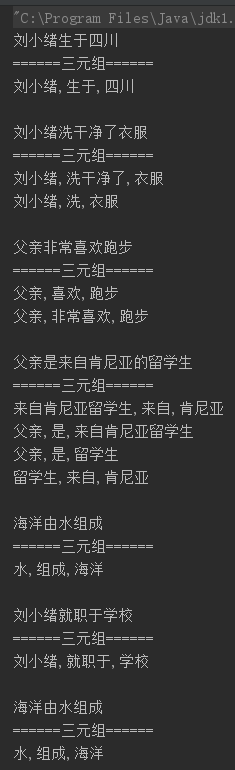


图4.5 基于依存句法分析抽取实体关系程序的运行效果图

## 4.4 知识入库

Neo4j支持多种数据导入方式，既可以手动导入，也可以采用csv文件直接导入，本文的关系数据采用csv批量方式导入，首先需要将关系数据整理成csv格式，每一行表示一条关系。导入csv文件数据之前，必须将被导入文件放入neo4j安装目录下的import文件夹下才可以，关系数据导入命令如下：

LOAD CSV WITH HEADERS FROM "file:/// relation.csv" AS line

MATCH (entity1:Hudong{title:line.HudongItem}) ,

(entity2:Hudong{title:line.NewNode})

CREATE (entity1)-[:RELATION { type: line.relation }]->(entity2)

因本文所抓取的实体详细信息数据中存在大量的回车符和制表符来保证文本格式的美观性，所以实体信息的导入没有采用csv的方式，而是采用了neo4j提供的java接口，由java程序读取记录，然后再由java为每一条数据创建一个结点，即将数据存入了数据库，主要程序代码如下所示：

session.run( "CREATE (a:Hudong {title: '" + json.get("title") + "',url:'"

                                + json.get("url") + "',image:'"

                                + json.get("image") +"',openTypeList:'"

                                + json.get("openTypeList") + "',detail:'"

                                + json.get("detail") + "',baseInfoKeyList:'"

                                + json.get("baseInfoKeyList") + "',baseInfoValueList:'"

                                + json.get("baseInfoValueList") + "' })");

在所有实体信息入库之后，为保证知识图谱的完备性，本文所采用的完备性测试方案为：随机抽选一篇与经济责任审计相关的新闻，然后将该篇新闻进行分词，去除其中的形容词、动词等肯定不是实体的词汇，然后去除停用词，最后再把所得到的词汇分别与数据库进行匹配，统计未匹配到的数量。本文所使用的新闻为“教育部就《教育部经济责任审计整改工作办法》答问”（来源：央视网；2017年08月16日），从中共提取出213个实体名称，然后逐个与数据库进行匹配，共有7个未匹配成功，使用（4-2）计算所得领域知识覆盖率为：96.71%。说明本文所构建的经济责任审计知识图谱对领域知识的覆盖率很高，但仍然存在少量的知识未能入库，需要进一步进行完善。

（4-2）

## 4.5 知识图谱应用

为了方便对构建的经济责任审计知识图谱进行展示，本文也采用web技术实现了知识图谱的展示功能，可供用户进行实体和实体关系的查询。知识存储使用了neo4j图数据库进行存储，它可以很方便的以图的形式展示知识，是知识存储的优先选择。

知识图谱展示所使用技术主要包括：python3、django框架和用于连接数据库的py2neo。在实体查询部分，由用户输入想要查询的实体，点击查询按钮，系统会将待查询的实体传入后台程序，由python负责到neo4j数据库中查询是否存在该实体，如果存在则返回true，不存在则返回false。用户界面会根据后台所返回的结果进行相应的处理，如果接收到的结果为false，则向用户弹出提示信息：“数据库中暂未添加该实体”；如果为true，则由javaScript脚本语言从当前页面跳转到实体详情页面，在跳转过程中会再词提交该实体名称，后台程序接收到该实体名称后会从数据库中去除该实体所有详细信息，并由js脚本语言渲染到实体详情页面，供用户查看。图4.6是实体查询页面，图中将以“温家宝”为查询字段进行查询；图4.7是查询字段得到的实体详细信息结果。

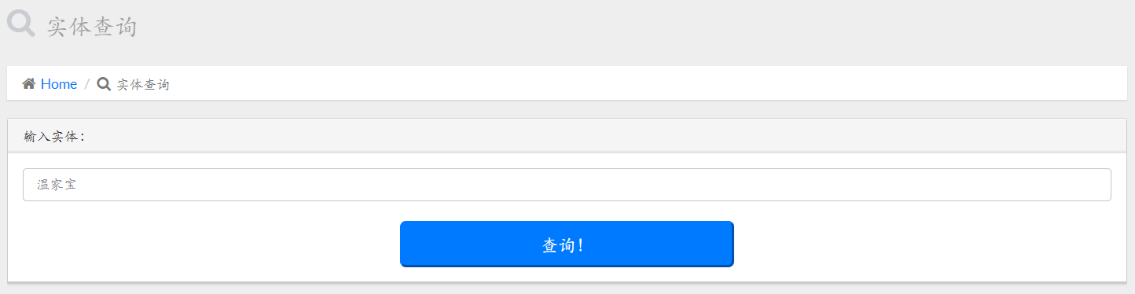


图4.6 实体查询页面



图4.7 实体详情页面

实体关系查询页面设置了三个输入框，分别对应实体1、关系、实体2的查询条件，用户可以选择其中任意一个或多个进行查询条件的录入，后台会根据具体的填写情况进行相应的数据库匹配工作，最后把查到的信息返回给前台页面。前端页面使用了charts.js插件来渲染实体关系图，用户可以很直观的通过关系图查看实体与实体之间的关系。在关系查询中，本文的亮点之处在于用户可以只输入关系进行查询，后台程序会匹配所有以此关系进行连接的实体。

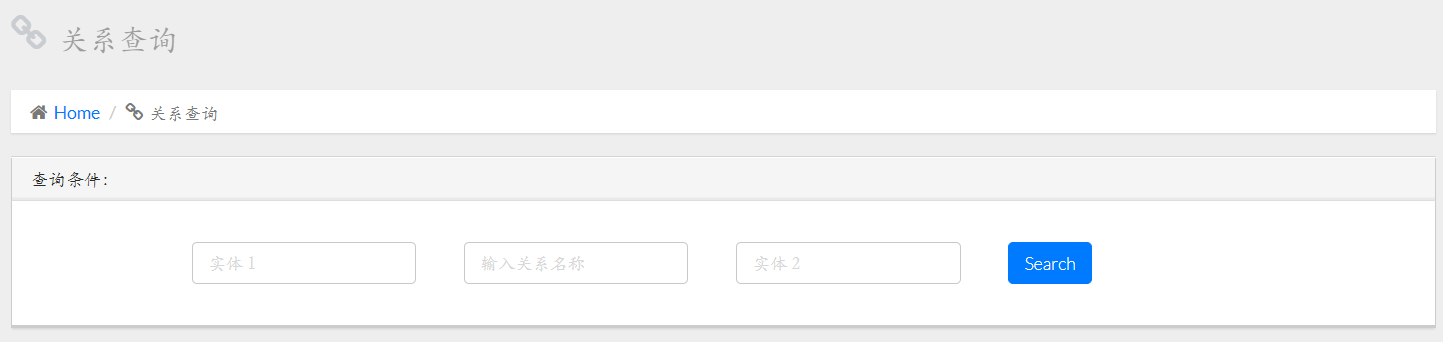
虽然关系图可以很直观的展现实体与实体之间的关系，但是却没有对关系信息进行有序的组织，所以在关系图的下方，还实现了一个关系列表，方便用户查看所有的关系数据。如图4.8是关系查询页面。

图4.8 关系查询页面

在图4.8实体1输入框中，填入“温家宝”，点击搜索按钮，所得到的关系图如图4.9所示，如果选择实体2输入框进行搜索，那么所得到的结果将是所有指向“温家宝”的关系和实体，而不是从“温家宝”指出的关系。如果只查询关系，可以只在关系名称输入想要查询的关系，如图4.10是以“职业”作为查询条件的所得到的结果。

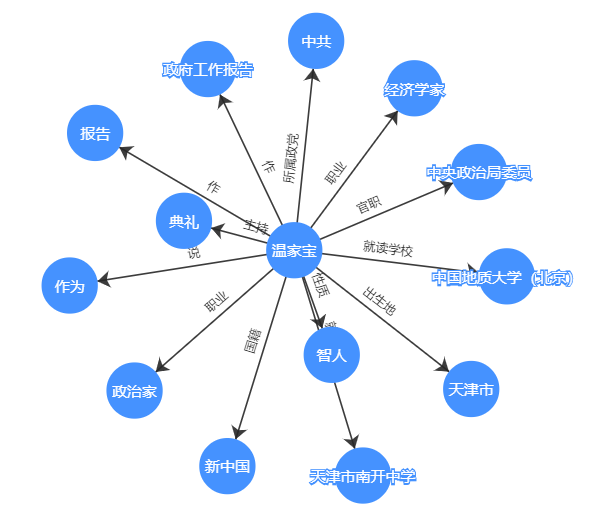


图4.9 按实体查询关系

从图4.9中可以看出，在本文所构建的经济责任审计知识图谱中，实体“温家宝”共与14个实体有关系，其中有少部分关系很明显不符合生活逻辑，这说明基于依存句法分析进行实体关系抽取所获得的结果并不全都是符合逻辑的，实体关系抽取仍然属于自然语言处理领域的难题。本文所得到的关系数量巨大，没有相关方法对所获得的实体关系进行筛选，如果使用人工进行筛选，那么效率将极其低下，费时费力，故本文不再进行处理。

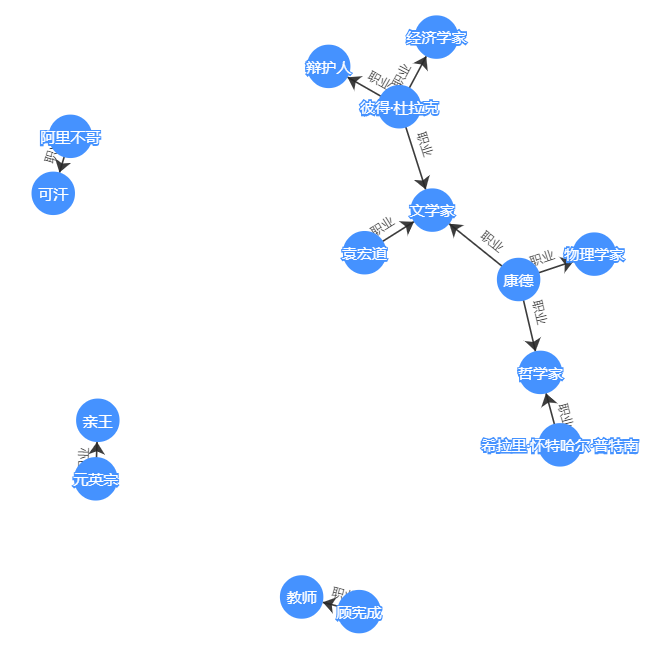


图4.10 按关系查询实体关系

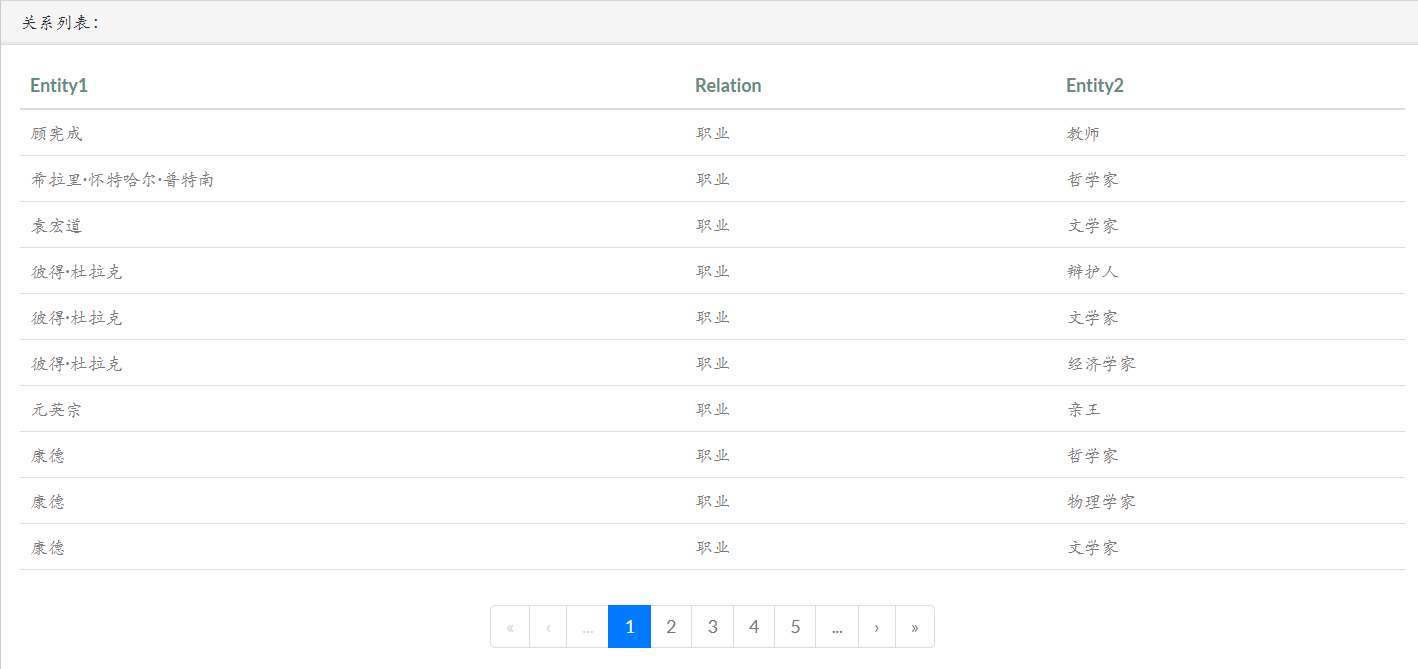
所有的查询结果不仅有关系图的直观展示，也有以列表形式组织的详细信息，图4.11是图4.10的列表形式展示。

图4.11 关系列表

从图4.11和图4.10对比可以刻看出，实际上查询到的结果远不止关系图中所展现的那么多，而没有在关系图中全部显示的原因是图形数据库的展示对硬件的要求较高，如果全部使用关系图的方式展示出来，那么将给硬件造成极大的负担，甚至可能导致程序出现未知错误而异常退出，所以在关系图的展示中，本文做了相应的限制。

## 4.6 本章小结

本章分别就知识获取、领域词汇判定、实体关系抽取、知识入库、知识图谱应用的实现做了详细的介绍，严格遵从第三章的设计进行实践，得到了较好实验效果，同时也说明第三章的设计是完全可靠的。并对已构建的知识图谱进行准确性与完备性测试，分析其优点与缺点，为以后的研究指明方向，并对后期工作进行展望。

# 结　　论

本文的工作是构建经济责任审计知识图谱。前文详细介绍了知识图谱的构建流程及实现工作，基本覆盖了业界公认的知识图谱构建流程：（1）知识获取；（2）信息抽取；（3）知识融合；（4）知识表示及存储；同时还对以构建知识图谱做了简单的实际应用。

在知识获取阶段，主要使用了网络爬虫技术，分别获取了新闻类文本和百科类词条知识。新闻类文本数据直接作为领域分词模型的训练数据和实体关系抽取的部分数据，百科类词条知识经整理后直接入库，作为领域实体，因此本文可以看作是基于百科知识进行经济责任审计知识图谱的构建。

在编写百科网站爬虫程序时，为了防止因网络、操作系统以及其他未知原因造成程序异常结束，所以使用了scrapy分布式爬虫框架；而针对审计署门户网站等新闻类网站则采取chromedriver模拟用户访问，其目的是为了防止被网站禁封ip以及较直观的观察爬虫程序进度。使用chromedriver模拟用户访问网站的方式进行数据抓取是本文在知识获取阶段的亮点，而存在的不足之处在于，依赖于网站本身提供的搜索框进行新闻的过滤，造成了很多潜在数据的丢失；在爬虫程序中加入领域词汇判断模型也是本文的一大亮点，直接在数据获取时就进行了数据的筛选工作，节省了硬盘资源和网络资源，不足之处在于训练集过少，由于无法获取到足够的经济责任审计领域文本作为训练集，所以本文所使用的模型依旧存在很大的缺陷。

本文所使用的实体信息来源于互动百科网站，因此实体属性也直接采用了互动百科所提供的属性，虽然在技术上没有难度，但是相比从开放中文文本中提取实体属性，这种方法所获得的知识更加可靠、准确。实体关系也分两部分来源，一部分来源于wikidata所提供的现有关系列表，这部分直接使用网络爬虫技术得到；另一部分来源于对开放中文文本进行实体关系抽取，这是本文的一大难点，基于依存句法分析的实体关系抽取方法也是本文唯一可行的方案，通过对抓取的新闻类文本和实体详细信息进行依存句法分析，然后再针对依存句法分析定义实体关系表述的规则，通过编写基于依存句法分析的实体关系抽取程序进行实体关系的抽取，最后对抽取的实体关系进行筛选，效果差强人意，存在的最大问题就是准确率不高，这也是业界一直在努力解决的问题，限于水平有限，没有继续做深入研究。

知识融合主要是将已获取知识进行整理，本文没有使用本体进行实体的归类，所有实体都归属为一个类，只进行了同义词替换、繁体转简体、英文转中文的工作，相对其他工作较为简单，所以没有做过多介绍。

本文使用三元组形式进行知识的表示，即“实体-关系-实体”的形式，每一条三元组就表示一条知识，本文共计构建了24万余条知识。而如此大量的数据如果采用关系型数据库进行存储，在查询时就需要进行关联查询，而由于本文的特殊性，所有的查询都会使用关联查询，关系型数据库的缺点就是在关联查询时会变得比较慢，所以综合考虑各方面因素，最终采用neo4j图形数据库进行知识的存储。

本文所实现的经济责任审计知识图谱demo是为了方便对知识图谱进行简单的应用，通过现在常用的web框架实现了对实体详细信息，实体关系的展示，用户可以根据实体名称或者关系名称对知识进行浏览。由于时间有限，本文只搭建了一个简单的搜索引擎，不支持模糊查询，也不支持知识推理，这些将作为下一步经济责任审计问答系统的研究工作。

本文在对知识图谱的演变历史与研究现状进行了建明扼要的阐述，同时也对我国独有的经济责任审计进行了现状分析，结合国内外对知识图谱和经济责任审计的研究现状，将知识图谱技术应用于经济责任审计领域。通过具体的分析与实现，构建了一个相对完善的经济责任审计知识图谱，具体来说，得到的结论如下：

1、研究探讨了知识图谱相关技术的发展情况，讨论了构建知识图谱的一般流程和技术，以及对知识的表示存储，并对比了使用关系型数据库和图形数据库的优缺点，对图形数据库做了详细的介绍；

2、明确了本文的目的，并通过文献、网络等资源了解知识图谱技术和经济责任审计的研究现状，为经济责任审计知识图谱的构建设计了具体的流程与技术方向；

3、明确了具体的技术方向后，使用java、python等基本编程语言，结合本文的实际情况，实现了知识获取、信息抽取、知识表示等任务，并通过技术路线图直观的展示了经济责任审计知识图谱的构建流程。

本文在构建经济责任审计知识图谱过程中所存在的不足之处有如下几点：

1、领域词汇判定模型使用的训练集较少，导致模型在个别情况下表现不佳；

2、实体信息是基于互动百科所得到的，没有从相关法律法规文献中提取实体和属性；

3、使用依存句法分析进行实体关系的抽取，因为依存句法分析本身的局限性，和本文所定义规则的不完善性，致使实体关系抽取准确率不高，依存句法分析阶段的错误在实体关系抽取时得到进一步的放大；

4、欠缺对已构建知识图谱的良好应用，本文只实现了简易版的搜索引擎功能，没有实现更高层次的应用，比如智能问答系统。

# 参考文献

1. 中共中央办公厅、国务院办公厅党政主要领导干部和国有企业领导人员经济责任审计规定（中办发〔2010〕32 号）[Z].2010.
2. 许建超. 谈谈企业经济责任审计的重点[J]. 中国盐业, 2017(15).
3. 郑石桥. 离任审计[M]. 中国审计出版社, 2003.
4. 李小红. 深化经济责任审计工作的几点体会[J]. 现代审计与经济, 2014(1):20-21.
5. 靳素平, 王逵轩. 深化经济责任审计工作的思考[J]. 中国内部审计, 2008(9):86-87.
6. Zheng Z. AnswerBus question answering system[C]// International Conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc. 2002:399-404.
7. 常思雯. 经济责任审计评价指标体系构建问题研究[J]. 中国集体经济, 2018(1):51-52.
8. 付磊. 试析大数据背景下推进经济责任审计全覆盖[J]. 宿州教育学院学报, 2017, 20(5):36-37.
9. Singhal A. Introducing the knowledge graph: things, not strings [EB/OL].Official google blog, 2012. https://googleblog. blogspot. co.za/2012/05/introducing-knowledge-graph-things-not. html.
10. 刘峤, 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(3):582-600.
11. 袁旭萍. 基于深度学习的商业领域知识图谱构建[D]. 华东师范大学, 2015.
12. Kinzler D, Pintscher L. Wikidata[C]// The International Symposium. 2014:1-1.
13. Bollacker K, Evans C, Paritosh P, et al. Freebase:a collaboratively created graph database for structuring human knowledge[C]// SIGMOD Conference. 2008:1247-1250.
14. 张观林, 欧阳纯萍, 邹银凤,等. 知识图谱及其在医疗领域的应用[J]. 湖南科技学院学报, 2016, 37(10):73-75.
15. Amarilli A, Galárraga L, Preda N, et al. Recent Topics of Research around the YAGO Knowledge Base[M]// Web Technologies and Applications. Springer International Publishing, 2014:1-12.
16. Dan B. DrugBank[M]// Dictionary of Bioinformatics and Computational Biology. John Wiley & Sons, Ltd, 2014.
17. 孙立伟, 何国辉, 吴礼发. 网络爬虫技术的研究[J]. 电脑知识与技术, 2010, 06(15):4112-4115.
18. 王芳, 陈海建. 深入解析Web主题爬虫的关键性原理[J]. 微型电脑应用, 2011, 27(7):32-34.
19. Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. Biological Cybernetics, 1980, 36(4):193-202.
20. Xu B, Wang N, Chen T, et al. Empirical Evaluation of Rectified Activations in Convolutional Network[J]. Computer Science, 2015.
21. Severyn A, Moschitti A. Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks[C]// The, International ACM SIGIR Conference. ACM, 2015:373-382.
22. 李明耀, 杨静. 基于依存分析的开放式中文实体关系抽取方法[J]. 计算机工程, 2016, 42(6):201-207.
23. Zhang Y, Zhou J F. A trainable method for extracting Chinese entity names and their relations[C]// The Workshop on Chinese Language Processing: Held in Conjunction with the, Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2000:66-72.
24. Mausam, Schmitz M, Bart R, et al. OPEN LANGUAGE LEARNING FOR INFORMATION EXTRACTION[J]. In Proceedings of EMNLP-CoNLL, 2012, 4590(5):523-534.
25. Ramani R, Kanna R. Relational Database Management System[J]. Betascript Publishing, 2005, 20(13):4611-4627.
26. Atzeni P, Bugiotti F, Rossi L. Uniform Access to Non-relational Database Systems: The SOS Platform[M]// Advanced Information Systems Engineering. Springer Berlin Heidelberg, 2012:160-174.
27. Graph Database[M]. Part Press, 2012.
28. Vukotic A, Watt N, Abedrabbo T, et al. Neo4j in Action[M]. Manning Publications Co. 2014.
29. 吕明育, 李小勇. NoSQL数据库与关系数据库的比较分析[J]. 微型电脑应用, 2011, 27(10):55-58.
30. 王仁武, 袁毅, 袁旭萍. 基于深度学习与图数据库构建中文商业知识图谱的探索研究[J]. 图书与情报, 2016(1):110-117.
31. Kim Y. Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.
32. Zhang X, Zhao J, Lecun Y. Character-level Convolutional Networks for Text Classification[J]. 2015:649-657.
33. 王秋月, 覃雄派, 曹巍,等. 扩展知识图谱上的实体关系检索[J]. 计算机应用, 2016, 36(4):985-991.

# 攻读学士学位期间发表的论文和取得的科研成果

# 致　　谢

为期14周的毕业设计已经接近尾声，回首这一学期的生活，过的充实而忙碌，收获颇多。作为计算机科学与技术学院计算机科学与技术专业的学生，我一直都在努力的提升自己的专业技能，大四这一年更是加倍努力的复习基础知识。虽然在做毕业设计的过程中遇到了很多困难，也曾一度想要放弃，但是最终都通过查阅相关资料，解决了各种难题。

本文是是在黄少滨老师的指导下完成的。在此由衷的感谢黄少滨老师对我的悉心指导，不仅让我参加实验室的研讨会，还让我跟着研究生学姐一起学习知识图谱构建的技术。黄老师对科研求真务实、对工作认真敬业的态度对我产生了很大的影响，尤其是黄老师教给我解决实际问题的方法，让我在面对难题时不至于惊慌失措，在撰写论文的过程中，黄老师也给予了我极大的帮助，这里由衷的感谢黄少滨老师。

其次，我要感谢赵国冬老师、刘越学姐和牛家彬、王辉两位学长，以及我的朋友和同学。他们在我大学的成长历程中起到了不可估量的作用，在我遇到困难的时候，无私的为我提供帮助，也指出了我毕业设计的不足，并帮助我加以改正。

作为一名哈尔滨工程大学的学生，我无比的自豪。感谢哈尔滨工程大学为我提供了一个良好的学习环境，尤其是免费的文献查询条件让我受益良多，感谢哈尔滨工程大学对我的悉心栽培，感谢同级哈尔滨工程大学各位同学的陪伴。

最后，感谢我的导员和2班全体同学，在大家的陪伴下，我度过了充实的本科生活，感谢家人一直以来对我默默的支持。感谢评阅论文的各位老师，谢谢你们为我的毕业设计和论文提出宝贵的意见，在这里献上最诚挚的祝福。愿你们阖家欢乐，万事如意。