

# 天津大学

## 本科生毕业论文



题目：美国科学研究系统的建模及合作模式的挖掘

学 院 管理与经济学部

专 业 信息管理与信息系统

年 级 2019 级

姓 名 蒋世华

学 号 3019209018

指导教师 王文俊 教授

# 独创性声明

本人声明：所呈交的毕业设计（论文），是本人在指导教师指导下，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，本毕业设计（论文）中不包含任何他人已经发表或撰写过的研究成果。对本毕业设计（论文）所涉及的研究工作做出贡献的其他个人和集体，均已在论文中作了明确的说明。本毕业设计（论文）原创性声明的法律责任由本人承担。

论文作者签名：

年 月 日

本人声明：本毕业设计（论文）是本人指导学生完成的  
研究成果，已经审阅过论文的全部内容。

论文指导教师签名：

年 月 日

# 摘要

科学研究是评价一个国家发展水平的重要标准,而评判一个国家的科研水平最直观的标准便是对它的相关科研成果信息进行研究和评价,因此本文着重从科研项目、科研人员、科研机构、科研论文四个本体构建关于美国科学研究系统的知识图谱,以实现美国科研发展的建模。

目前构建知识图谱的主要问题包括几个方面:数据质量问题、不同数据源问题、本体构建问题。同时在构建美国学术研究网络时也存在诸多方面的挑战,数据准确性、实体链接和消歧、知识图谱的更新和维护。

本文针对上述挑战,面向美国科研体系建模问题,收集了 52883 条美国国立卫生研究院数据与 47949 条美国国家自然科学基金会数据合计 100832 条数据通过 Py2neo 库与 Neo4j 图数据库构建了 84065 个节点、96084 个关系,形成美国学术研究网络。通过复杂网络的统计指标分析了网络中潜在的性质及隐藏信息,并在 Semantic Mediawiki 平台上分别创建了四个本体的属性、分类、模版、表单,并批量导入数据进 SMW 平台并对知识图谱可视化。

通过构建于 SMW 平台的知识图谱,用户可以通过查询和浏览知识图谱中的实体和关系,发现各种科学研究中的规律和趋势,从而提出更加深入和具有前瞻性的研究问题和方向。公众或者科研人员可以了解各种科学研究的发展历程、研究成果、研究方法等方面的信息,有助于提高公众的科学素养和科研人员的研究效率。它能够为科学研究提供更加全面和系统的视角,为公众和科研人员提供方便的参考和查阅工具,有助于推动科学研究的深入和发展。

最后通过对网络的统计指标分析。发现生成的网络属于稀疏图;度分布呈现幂律分布,且符合无标度网络的特性;科研机构或科研人员更倾向于与未合作过的进行合作,而不仅仅是大度节点与大度节点的合作,即影响力大的人员或机构并不倾向于与影响力大的合作;大部分节点都为小度节点,仅有少量节点拥有大度。

**关键词:** 科学研究系统, 知识图谱, 复杂网络, SMW 平台

# ABSTRACT

Scientific research is an important indicator of a country's development level. Evaluating a country's scientific research level can be done by studying and evaluating its relevant scientific research results. Therefore, this paper focuses on constructing a knowledge graph about the US scientific research system from four ontologies: research projects, researchers, research institutions, and research papers, in order to model the development of US scientific research.

Now constructing a knowledge graph faces some problems, including data quality, different data sources, ontology construction. And constructing an US academic research network faces various challenges, like the accuracy of data sources, entity linking and disambiguation, knowledge graph updates and maintenance.

This paper addresses the challenges mentioned above and model the US scientific research system. A total of 100,832 data were constructed using 52,883 National Institutes of Health data and 47,949 National Science Foundation data through the Py2neo library and Neo4j graph database, creating 84,065 nodes and 96,084 relationships. The potential properties and hidden information in the network were analyzed by complex network statistical indicators. Four ontologies' properties, classifications, templates, and forms were created on the Semantic Mediawiki platform, and data was imported in bulk for visualization of the knowledge graph.

Through the knowledge graph constructed on the SMW platform, users can query and browse entities and relationships in the knowledge graph, discover patterns and trends in various scientific research fields, and propose more in-depth and forward-looking research questions and directions. The public or researchers can learn about the development process, research results, research methods, and other aspects of various scientific research fields, which helps improve the public's scientific literacy and researchers' research efficiency. SMW platform provides a more comprehensive and systematic perspective for scientific research, and serve as a convenient reference and lookup tool for the public and researchers, which helps to promote the in-depth and development of scientific research.

Finally, through statistical analysis of the network, it was found that the network belongs to a sparse graph; the degree distribution follows a power-law distribution and conforms to the characteristics of a scale-free network; scientific research institutions or scientific researchers tend to cooperate with those they have not collaborated with before, rather than just collaborating with those nodes with high degrees; most nodes are small-degree nodes, and only a few nodes have high degrees.

**KEY WORDS:** Scientific research system, Knowledge graph, Complex network, Semantic Mediawiki platform

# 目 录

第一章 绪论 .....	1
1.1 研究背景 .....	1
1.2 国内外研究现状 .....	1
1.3 研究目标及方法 .....	1
1.3.1 研究目标 .....	8
1.3.2 研究方法 .....	8
1.4 可行性分析 .....	1
第二章 相关概念与技术 .....	8
2.1 自然语言处理技术 .....	8
2.1.1 自然语言处理技术的核心技术 .....	8
2.1.2 自然语言处理技术的应用场景及发展趋势 .....	8
2.2 图数据库技术 .....	9
2.2.1 图数据库技术的构建方法及核心技术 .....	9
2.2.2 图数据库技术的应用场景及发展趋势 .....	10
2.3 知识图谱 .....	9
2.3.1 知识图谱的构成和组成 .....	9
2.3.2 知识图谱的构建方法 .....	10
2.3.3 知识图谱的语义建模及数据挖掘技术 .....	9
2.3.4 知识图谱的应用领域 .....	10
2.4 复杂网络 .....	10
2.4.1 复杂网络的特点 .....	10
2.4.2 复杂网络的模型 .....	11
2.4.3 复杂网络的应用 .....	11
第三章 数据集 .....	12
3.1 Scrapy 爬虫 .....	12
3.1.1 Scrapy 爬虫整体架构 .....	12

3.1.2	Scrapy 爬虫流程 .....	13
3.2	数据集来源 .....	14
3.2.1	美国国立卫生研究院 (NIH) 数据 .....	14
3.2.2	美国国家自然科学基金会 (NSF) 数据 .....	14
3.3	数据集概况 .....	14
第四章	数据预处理 .....	18
4.1	数据清洗 .....	18
4.1.1	删除多余数据 .....	18
4.1.2	数据格式处理 .....	18
4.2	数据变换 .....	18
4.2.1	数据特征对齐 .....	18
4.2.2	特征选择 .....	19
4.3	数据融合 .....	20
第五章	知识图谱构建 .....	22
5.1	节点构建 .....	22
5.1.1	节点 .....	22
5.1.2	节点特征 .....	22
5.1.3	节点实例化 .....	22
5.2	关系构建 .....	23
5.2.1	关系 .....	23
5.2.2	关系实例化 .....	23
第六章	复杂网络分析 .....	25
6.1	复杂网络 .....	25
6.1.1	网络生成 .....	25
6.2	网络的统计指标 .....	25
6.2.1	网络密度 .....	25
6.2.2	度 .....	25
6.2.3	度相似性系数 .....	26
6.2.4	平均聚类系数 .....	27
6.2.5	k 核 .....	27
6.3	指标计算 .....	27

第七章	SMW 平台开发与可视化	29
7.1	SMW 平台开发	29
7.1.1	属性	29
7.1.2	分类	30
7.1.3	模版	31
7.1.4	表单	32
7.2	数据导入及可视化	32
第八章	总结	34
8.1	结论	34
8.2	不足与展望	34
参考文献		35
致 谢		38

## 第一章 绪论

### 1.1 研究背景

当今社会，科学研究已经成为推动社会进步和人类文明发展的重要力量。特别是在信息技术和互联网的快速发展下，科学研究的规模和复杂度也在不断增加。因此，传统的研究方法和模式已经无法满足当今科研发展的需要。在这样的背景下，科学研究的合作方式和合作伙伴选择变得越来越重要。

在这方面，美国作为当今世界上科技实力最强大的国家之一，其科学研究水平和技术创新能力一直处于世界领先地位。美国的科学研究体系包括了政府、学术机构、企业 and 非营利组织等多个方面，这些机构之间的合作和互动是美国科学研究成功的重要因素。因此，深入研究美国科学研究系统的运行机制和合作模式，具有重要的理论和实践意义<sup>[1][2]</sup>。

本文旨在通过建立美国科学研究系统的模型并利用网络分析方法<sup>[3][4]</sup>，探究美国科学研究系统的网络特征和关键节点，深入挖掘出不同领域和机构之间的科研合作模式和因素<sup>[5]</sup>，为优化研究资源配置和提高科研成果产出提供理论和实践支持<sup>[6][7]</sup>。

同时将利用网络分析方法，构建美国科学研究系统的合作网络模型。通过分析该模型的网络特征和关键节点，揭示不同领域和机构之间的科研合作模式和因素。本文还将通过数据挖掘技术，分析美国科学研究系统中的研究主题和研究热点，为科研人员提供重要的参考和指导。

本文将通过深入研究美国科学研究系统的运行机制和合作模式，探索科学研究的合作方式和合作伙伴选择，为优化研究资源配置和提高科研成果产出提供理论和实践支持<sup>[8][9]</sup>。

### 1.2 国内外研究现状

知识图谱的发展可以追溯到上世纪六十年代的语义网络研究，而自 Google 公司于 2012 年正式提出知识图谱概念后，经过近十年的发展，知识图谱已经在各行业、各领域产生了深远的影响。在这其中，对于知识图谱的技术与应用方面的研究也在不断深入。而面对冗杂的科研流程，越来越多的专家学者也将知识图谱技术应用于管理科研机构领域。在国外，Ganggao (2018) <sup>[10]</sup>等为应对知识图谱中实体在自然语言处理中存在歧义的问题，创建了基于实体的上下文词和信息



词之间语义相似度的消歧方法,提高了各个实体之间的独立性,完善了知识图谱的技术体系。Danae Pla (2016)<sup>[11]</sup>等则将目光集中于知识图谱的应用层面,该研究利用知识图谱技术通过从推特文章中检索语义信息,计算用户之间的兴趣相似度,成功降低了推特过度推荐和过度专门化等问题。而对于科研方面的知识图谱应用,Liu W (2018)<sup>[3]</sup>等基于海量科研数据构建知识图网络的方法,以从每个课题设置文档的标题信息中提取主题词为关键技术,以及课题的方向,对项目库进行分析,提取项目的基本信息,并对每个课题的文档信息和课题方向进行分析。构建属于同一字段的主题方向的知识图谱。SONG S (2023)<sup>[6]</sup>等利用技术要素和专家知识图操作生成专家知识图图像。基于科技大数据知识图谱的智能技术诊断专家匹配算法,具有一套为开发技术评估、商业计划、决策咨询、前沿分析和市场预测提供指导的增值服务。在国内,杨思洛 (2012)<sup>[12]</sup>等利用可视化方法总结了知识图谱的研究现状,重点对比分析了知识图谱研究的核心作者与所有作者的合作网络,宏观与微观机构合作网络通过高频关键词的共现分析研究知识图谱的发展趋势。李思志 (2014)<sup>[13]</sup>等则从管理科学与工程宏观的创新轨迹入手,提出利用引文共引分析、文本挖掘、特征抽取的方法,基于图谱分析平台对创新趋势进行分析。雷洁 (2020)<sup>[8]</sup>等则将知识图谱技术应用于科研档案领域的研究当中。该团队通过构建基于知识图谱的科研档案管理模型,从知识层面将科研档案资源中的科研机构、科研项目、科研成果、人员等要素与项目任务书、合同、研究报告中抽取的知识单元进行关联和融合,丰富科研档案的语义关系,推进科研档案管理系统提档升级。杜悦 (2023)<sup>[9]</sup>等提出的隐式知识图谱构建方法很好地解决了由于实体信息变动引发的数据一致性问题,适用于大规模科研知识图谱的构建,有助于科技知识的高效管理和传播利用。

综上所述,当前国内外知识图谱研究大体分为技术研究与实际应用两大方向。技术研究主要集中于数据挖掘、实体识别、关系抽取等方向;实际应用则主要集中于智能问答、辅助决策等方面。针对美国科研机构项目的知识图谱研究<sup>[4]</sup>,由于时间积累不足,主要集中在其科研档案或文献数据方面<sup>[5][7][15]</sup>,对于科研机构项目的研究较少,且并未将其作为研究主体。因此本研究将美国科研机构项目作为研究主体符合当前科研机构项目管理的迫切需要。

### 1.3 研究目标及方法

#### 1.3.1 研究目标

本文的工作目标是借助知识图谱技术,基于美国科研机构项目的开源大数据,通过对美国科研机构项目信息的语义网络分析,结合对美国科研机构项目的

实体与关系抽取，构建美国科研机构项目知识图谱，并以 Media Wiki、SMV 等开源工具为基础搭建一个美国科研机构项目流程图谱的可视化管理系统。在对该知识图谱进行管理和可视化分析的基础上，本文采用 Neo4j 图数据库与 SMW 平台实现对于知识图谱数据的构建，并通过构建的美国学术研究网络通过复杂网络挖掘之间的关联性，进而满足对科研机构项目的管理需求。

### 1.3.2 研究方法

- (1) 文献搜集法。通过搜集、整理知识图谱领域相关文献资料，归纳总结知识图谱研究现状，并总结整理知识图谱构建方法。
- (2) 系统分析与设计。借助知识图谱技术，基于美国科研机构项目的开源大数据，通过对美国科研机构项目的语义网络分析、结合针对美国科研项目的实体与关系抽取，利用 Neo4j 与 SMW 平台构建美国科研机构项目知识图谱的可视化系统。

### 1.4 可行性分析

从经济可行性上分析，本知识图谱的设计与开发基于 Media Wiki 开源环境，将先在个人 PC 端调试，之后上传网络，具有经济可行性。

从技术可行性上分析，学生具备一定的图数据库编写能力，文本语义网络分析能力，Python 等编程语言的编写能力，具有技术可行性。

从社会可行性上分析，美国科研机构项目知识图谱的可视化系统能够为相关科研机构及管理部门提供清晰、准确的参考，降低科研机构运行及管理成本，具有较高的应用推广价值，具备社会可行性。

## 第二章 相关概念与技术

本文主要使用了自然语言处理技术、图数据库技术、知识图谱以及复杂网络的相关理论知识，下文将从这三方面进行介绍。

### 2.1 自然语言处理技术

自然语言处理技术 (Natural Language Processing, 简称 NLP) 是一种人工智能技术, 它致力于研究如何让计算机能够理解、分析、处理人类自然语言的能力。自然语言是人类交流和沟通的主要方式, 而 NLP 技术的发展可以使计算机更好地理解 and 处理人类的语言信息。

#### 2.1.1 自然语言处理技术的核心技术

自然语言处理技术的核心技术包括分词、词性标注、命名实体识别、句法分析、语义分析、信息抽取等。其中, 分词是将一段连续的文本切分成一系列有意义的词语; 词性标注是为每个词语标注其在句子中的词性; 命名实体识别是识别文本中的命名实体; 句法分析是分析句子的语法结构; 语义分析是分析句子的语义含义。信息抽取是从文本中提取结构化信息的过程, 它可以用来自动化地获取文本中的实体、关系和事件等信息<sup>[16]</sup>。

#### 2.1.2 自然语言处理技术的应用场景及发展趋势

自然语言处理技术的应用场景非常广泛, 包括文本分类、情感分析、机器翻译、问答系统等。在文本分类中, NLP 技术可以对文本进行分类。在情感分析中, NLP 技术可以分析文本中的情感色彩。在机器翻译中, NLP 技术可以将一种语言翻译成另一种语言。在问答系统中, NLP 技术可以回答用户提出的自然语言问题。

自然语言处理技术的发展趋势包括深度学习在自然语言处理中的应用、多语言处理、跨模态处理等。深度学习是一种基于神经网络的机器学习方法, 它在自然语言处理中得到了广泛的应用, 在语言模型、文本分类、机器翻译等领域。多语言处理是指处理多种语言的能力, 跨语言信息检索、跨语言机器翻译等。跨模态处理是指处理多种模态的能力, 图像与文本的关联分析、语音识别与文本理解等。

深度学习在自然语言处理技术中的应用越来越广泛，随着 Chatgpt 的诞生，自然语言处理技术的不断发展将给人们带来更加便利的语言交流和更加智能化的计算机应用。

## 2.2 图数据库技术

图数据库是一种用于存储和处理图形数据的数据库系统，它使用图形结构来表示和存储数据，并提供了一组用于查询和操作图形数据的 API。与传统的关系型数据库不同，图数据库不仅可以存储实体和实体之间的关系，还可以存储实体和实体属性之间的关系，从而更好地支持复杂的数据模型和查询需求。

### 2.2.1 图数据库技术的构建方法及核心技术

图数据库通常使用基于节点和边的模型来表示数据<sup>[17]</sup>，其中节点表示实体，边表示实体之间的关系。为了支持高效的查询和操作，图数据库通常使用索引和查询优化技术。

另外，为了支持大规模数据存储和处理，图数据库通常采用分布式存储和计算架构。可以使用分布式索引和查询优化技术来提高性能和可扩展性。

图数据库的核心技术包括图形存储、查询优化、索引技术、分布式处理、事务管理等。其中，图形存储是图数据库最基本的技术，它涉及如何将图形数据存储在内存或磁盘中；查询优化是图数据库的关键技术之一，它涉及如何优化查询计划、减少查询时间和资源消耗；索引技术是图数据库的另一个重要技术，它涉及如何构建和维护节点和边的索引，以支持高效的查询和操作；分布式处理是图数据库的必要技术，它涉及如何将大规模图形数据分布式存储和计算，以提高性能和可扩展性；事务管理是图数据库的基本功能之一，它涉及如何实现多用户并发访问、数据一致性和可靠性等功能。

### 2.2.2 图数据库技术的应用场景及发展趋势

图数据库广泛应用于社交网络、知识图谱、推荐系统、网络安全等领域。Facebook 使用图数据库来存储和处理社交网络数据，Google 使用图数据库来构建知识图谱。在社交网络中，图数据库可以用于实现好友关系、消息传递、推荐系统等功能；在知识图谱中，图数据库可以用于存储实体和实体之间的关系、属性和属性之间的关系等信息。

随着大数据和人工智能技术的发展，图数据库的应用前景非常广阔。未来，图数据库将更加注重性能和可扩展性，并与其他技术如自然语言处理、机器学习

等结合,实现更加智能化的数据处理和分析。可以将图数据库与自然语言处理技术结合,实现基于自然语言的图查询和分析;将图数据库与机器学习技术结合,实现基于图的数据挖掘和预测分析。同时,图数据库还将更加注重安全性和隐私保护,以应对日益严峻的网络安全挑战。

而且,最近几年,由于数据大规模的互相连接和 Web 3.0 概念的出现,导致了对于图数据的查询和存储的需求增长迅速,一些新的图数据库技术也随之涌现,如 Neo4j、ArangoDB、JanusGraph 等,图数据库技术有广泛应用前景。

## 2.3 知识图谱

知识图谱是一种基于语义网络的知识表示方法,它的出现为人工智能领域的发展提供了新的思路和方法。知识图谱旨在将结构化、半结构化和非结构化的数据进行融合,形成一个大规模的、可访问的知识库。在这个知识库中,各种实体、属性和关系都被抽象出来,形成一个具有语义关联的知识图谱。

知识图谱的构建需要利用自然语言处理、机器学习和图论等技术。其中,自然语言处理技术用于将文本数据转换为结构化的数据,机器学习技术用于自动抽取和链接实体、属性和关系,图论技术用于构建知识图谱的结构和关系。

知识图谱的构建可以分为三个主要步骤:知识抽取<sup>[18]</sup>、实体链接<sup>[19]</sup>和关系抽取。知识抽取是将非结构化和半结构化的数据转换为结构化的数据的过程,包括实体抽取、属性抽取和关系抽取。实体链接是将不同数据源中的实体进行链接的过程,实现实体的唯一性标识和语义一致性。关系抽取是将实体之间的关系进行抽取和链接的过程,形成知识图谱的结构和语义关联。

知识图谱的构建具有许多优点。首先,它可以将各种结构化、半结构化和非结构化的数据进行融合,形成一个大规模的、可访问的知识库,为人工智能领域的研究和应用提供了重要的数据基础。其次,知识图谱可以实现实体和关系的语义一致性和唯一性标识,提高了数据的质量和可靠性。最后,知识图谱可以为各种领域的研究和应用提供重要的支持。

知识图谱是一种基于语义网络的知识表示方法,它将各种数据源中的实体、属性和关系进行抽取和链接,形成一个具有语义关联的知识图谱。知识图谱的构建需要利用自然语言处理、机器学习和图论等技术,具有重要的理论和实践意义。

### 2.3.1 知识图谱的构成和组成

知识图谱的基本组成要素一般有实体、关系、属性、属性值、本体。

在知识图谱里，通常用“实体”来表达图里的节点、用“关系”来表达图里的“边”，实体是知识图谱中的最基本元素，实体指的是现实世界中的事物，不同的实体间存在不同的关系。关系是一种表示实体之间的逻辑关系的方式，用于连接实体并表示它们的相互关系。

属性指的是图中的一条边，它指的一种实体的特征或描述性信息。它是一种关于实体的附加信息，描述了该实体的特征，性质，状态等。属性可以是定性的或定量的。知识图谱的目的之一是通过建模实体的属性，使知识能够被机器理解和使用。

属性值是一个节点，它主要指实体的某个属性的值，用来具体描述一个实体本身所具有的某种性质或特性。

本体是一个节点，这是一种特殊的实体，它是具有同种特性的实体构成的集合，主要用来描述集合、类别、对象类型或事物的种类。本体与实体之间具有从属关系。

而构成知识图谱的基本单位是“三元组”，分为以下两种。

第一种为“实体-关系-实体”，这种三元组一般用来描述实体之间的关系；第二种为“实体-属性-属性值”，这种三元组一般用来描述实体本身具有的特征。

### 2.3.2 知识图谱的构建方法

对于知识图谱的构建方法主要有自顶向下的构建方法和自底向上的构建方法。其主要区别在于模式层的定义上。

自顶向下的方法是指首先为知识图谱定义本体或进行 Schema 设计，包括本体的上下位关系和本体的约束等，然后逐步细化构建实体。常用于行业知识图谱。

这种方法的优点是可以保证知识图谱的准确性和一致性，同时可以更好地满足特定领域的需求。但是，这种方法的缺点是需要耗费大量的时间和人力来定义本体和概念体系，同时可能会忽略一些实际存在的实体和关系等信息。

自底向上的方法是指首先构建实体，然后通过实体相似度计算等方式，逐步往上抽象形成本体。自底向上的构建方法一般是从开放链接的数据源中提取实体、属性和关系，加入到知识图谱的数据层；然后将这些知识要素进行归纳组织，逐步往上抽象为本体或概念，最后形成模式层。常用于通用知识图谱。

这种方法的优点是可以快速构建知识图谱，同时可以利用大量的数据来丰富知识图谱的内容。但是，这种方法的缺点是可能存在一些噪声数据和不准确的信息，需要进行后期处理和清洗。

### 2.3.3 知识图谱的语义建模及数据挖掘技术

知识图谱的语义建模是指将知识图谱中的实体、关系和属性等信息进行语义化处理,使得机器可以理解这些信息的含义和关系。语义建模的核心是将实体、关系和属性等信息映射到本体和概念体系中,从而建立起实体、关系和属性等之间的语义关系。通过语义建模,可以使得知识图谱更加准确、完整和可靠,从而提高知识图谱的应用价值。

在语义建模中,本体和概念体系的设计是非常重要的。本体是指用于描述某个领域中的实体、关系和属性等信息的一组概念和定义。概念体系是指用于描述某个领域中的概念和关系等信息的一组概念和定义。本体和概念体系的设计需要考虑到领域的特点和需求,同时需要与知识图谱中的实体、关系和属性等信息相对应。

知识图谱的数据挖掘技术是指通过对知识图谱中的数据进行分析和挖掘,发现其中的模式、规律和关联等信息。数据挖掘技术可以帮助人们更好地理解知识图谱中的数据,从而为知识图谱的应用提供更多的支持。通常包括关联规则挖掘、聚类分析、分类分析、异常检测、图形分析等。

关联规则挖掘是一种常用的知识图谱数据挖掘技术,它可以发现知识图谱中实体之间的关系和属性之间的关联规律;聚类分析是一种将知识图谱中的实体按照某种相似度进行分组的方法,可以帮助人们更好地理解知识图谱中的实体之间的关系<sup>[3]</sup>;分类分析是一种将知识图谱中的实体按照某种特征进行分类的方法,可以帮助人们更好地理解知识图谱中的实体之间的差异和相似性;异常检测是一种检测知识图谱中异常实体的方法,可以帮助人们更好地发现知识图谱中的异常情况;图形分析是一种将知识图谱中的实体和关系等信息进行可视化展示的方法,可以帮助人们更好地理解知识图谱中的数据<sup>[12]</sup>。

### 2.3.4 知识图谱的应用领域

知识图谱是一种基于语义网络的知识表示和推理技术,它可以将各种不同类型的数据和知识连接起来,形成一张结构化的知识图谱。这张图谱可以帮助机器理解和推理人类的知识和语言,从而实现更加智能化的服务和应用。

知识图谱目前已经被应用在了很多领域。例如在智能问答方面,知识图谱可以帮助机器理解用户提出的问题,并且从知识图谱中找到最合适的答案<sup>[20]</sup>;在推荐系统方面,知识图谱可以帮助机器理解用户的兴趣和需求,并且从知识图谱中找到最适合用户的产品或服务<sup>[11][21]</sup>;在语义搜索方面,知识图谱可以帮助机器理解用户的搜索意图,并且从知识图谱中找到最相关的结果<sup>[22]</sup>。同时在机器翻译

[23]、智能客服[24]、智能家居[25]、企业知识管理[26]、金融风控[27]、医疗诊断[28][29]、智慧城市[30]等方面均已用相关的研究及应用。

总之，知识图谱在各个领域都有广泛的应用前景，将为社会的发展带来巨大的推动力。

## 2.4 复杂网络

复杂网络是指由大量节点和连接构成的网络结构，其中节点之间的连接方式具有一定的复杂性和随机性。复杂网络的分析和研究可以运用复杂网络理论进行模拟、预测和优化。最常用的方法是基于网络的结构、节点度数、聚集系数、介数中心性、PageRank 等指标对网络进行分析。

### 2.4.1 复杂网络的特点

复杂网络具有以下几个特点：

- 1、大规模性：复杂网络通常由大量节点和连接构成，规模较大。
- 2、非均质性：复杂网络中的节点之间连接的方式不同，节点的度数和连接强度也不同。
- 3、高聚类性：复杂网络中的节点往往会聚成一些密集的群体，形成高聚类的特点。
- 4、小世界性：复杂网络中的节点之间通常存在短路径，即任意两个节点之间的距离都很短。
- 5、无标度性：复杂网络中的节点度数分布呈现出幂律分布的特征。

### 2.4.2 复杂网络的模型

为了更好地研究和应用复杂网络，目前已经提出了许多复杂网络的结构和模型，常见的有以下几种：

- 1、ER 随机网络模型：是最早提出的复杂网络模型之一，节点之间的连接是随机的。
- 2、WS 小世界网络模型：在 ER 模型的基础上，增加了一定的重连机制，使得节点之间存在短路径。
- 3、BA 无标度网络模型：少数节点的度数极高，而大部分节点的度数较低。
- 4、静态网络模型：假设网络结构不随时间变化，常用于研究网络的拓扑结构和性质。



5、动态网络模型：考虑网络结构随时间变化的情况，常用于研究网络演化和动态性质。

### 2.4.3 复杂网络的应用

复杂网络在许多领域都有广泛的应用，例如社交网络、生物网络、交通网络、金融网络等。具体应用包括：

1、社交网络分析：通过分析社交网络中的节点和连接关系，了解社交网络的结构和特征，以及节点之间的影响和传播。

2、生物网络研究：通过构建生物网络，分析生物体系中的基因、蛋白质和代谢物之间的关系，探索生物体系的结构和功能<sup>[31]</sup>。

3、交通网络优化：通过分析交通网络中的节点和连接关系，优化交通流量和路径选择，提高交通效率和安全性。

4、金融网络研究：通过构建金融网络，分析金融体系中的投资和风险传播，预测金融市场的变化和趋势。

## 第三章 数据集

本文使用了 Scrapy 爬虫从美国国立卫生研究院([National Institutes of Health](https://www.nih.gov), 简称 NIH)<sup>1</sup>公开的数据中爬取了从 2010 ~ 2021 年期间、由美国机构承担的总共 52883 条科研项目信息, 从美国国家自然科学基金会([National Science Foundation](https://www.nsf.gov), 简称 NSF)<sup>2</sup>公开的数据中爬取了从 2010 ~ 2023 年期间、由美国机构承担的总共 47949 条科研项目信息。合计总共 100832 条科研项目信息。

### 3.1 Scrapy 爬虫

本文利用了 Scrapy 爬虫从 NIH 和 NSF 的公开网站上分别爬取了 52883 条和 47949 条科研项目信息的数据, 并将这总计 100832 条科研项目信息作为知识图谱构建的主要数据集来源。

#### 3.1.1 Scrapy 爬虫整体架构

Scrapy 是一个 Python 编写的开源网络爬虫框架, 可以用于抓取网站并从中提取结构化的数据。它具有高效、可扩展和可配置的特点, 可以自动化地从网站中提取数据, 支持多种数据格式和存储方式, 并且可以通过中间件和插件来扩展其功能。Scrapy 还提供了强大的调试工具和自动化测试功能, 可以帮助开发者快速开发和测试爬虫。同时, Scrapy 还支持异步请求和分布式爬取, 可以提高爬取效率和稳定性。Scrapy 是一个强大的网络爬虫框架, 适用于各种数据抓取和处理场景, 是 Python 爬虫开发的重要工具之一。

Scrapy 爬虫的整体架构如下:

- 1、引擎(Scrapy Engine), 用来处理整个系统的数据流处理, 触发事务。
- 2、调度器(Scheduler), 用来接受引擎发过来的请求, 压入队列中, 并在引擎再次请求的时候返回。
- 3、下载器(Downloader), 用于下载网页内容, 并将网页内容返回给蜘蛛。
- 4、蜘蛛(Spiders), 蜘蛛主要用来获取网页内容, 用它来制订特定域名或网页的解析规则。编写用于分析 response 并提取 item(即获取到的 item)或额外跟进的 URL 的类。每个 spider 负责处理一个特定(或一些)网站。

---

<sup>1</sup> <https://grants.nih.gov>

<sup>2</sup> <https://www.nsf.gov>

5、项目管道(ItemPipeline), 负责处理有蜘蛛从网页中抽取的项目, 他的主要任务是清晰、验证和存储数据。当页面被蜘蛛解析后, 将被发送到项目管道, 并经过几个特定的次序处理数据。

6、下载器中间件(DownloaderMiddlewares), 位于 Scrapy 引擎和下载器之间的中间件, 主要是处理 Scrapy 引擎与下载器之间的请求及响应。

7、蜘蛛中间件(SpiderMiddlewares), 介于 Scrapy 引擎和蜘蛛之间的中间件, 主要工作是处理蜘蛛的响应输入和请求输出。

### 3.1.2 Scrapy 爬虫流程

Scrapy 流程图如图 3-1 所示, 各个组件之间获取网页数据的具体流程步骤如下。

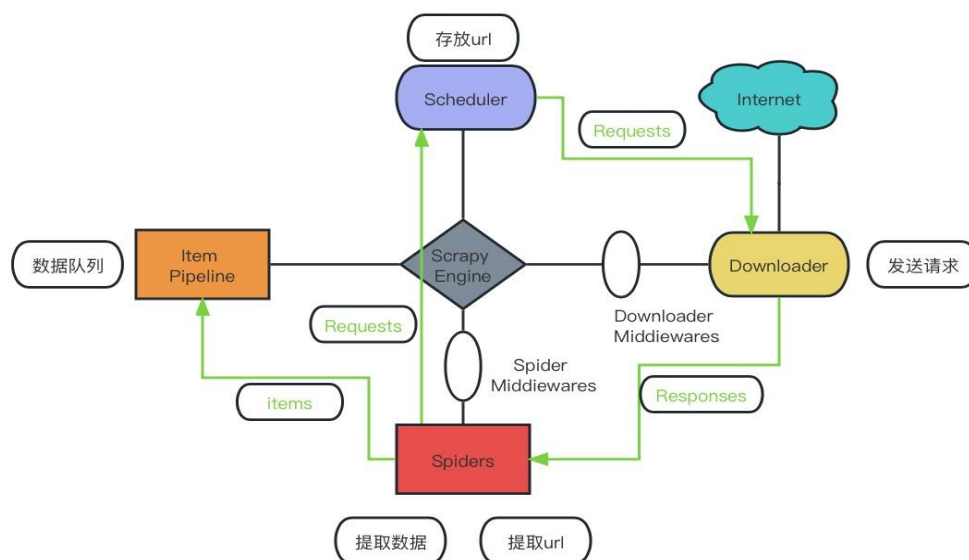


图 3-1 Scrapy 爬虫流程图

1. Spider 向 Scheduler 请求下一个要爬取的 URL。
2. Scheduler 返回下一个要爬取的 URL，并将其添加到队列中。
3. Downloader 从队列中获取下一个要爬取的 URL。
4. Downloader 向该 URL 发送一个 HTTP 请求。
5. 目标网站接收到请求并返回一个 HTTP 响应。
6. Downloader 将响应传递给 Spider。
7. Spider 解析响应，提取目标数据，并生成 Item。
8. Spider 还可以生成新的请求，并将其添加到 Scheduler 队列中。

9. 步骤 3-8 循环执行，直到队列为空或 Spider 被关闭。
10. 当 Spider 关闭时，它将生成一个包含所有提取的数据的输出文件。
11. Scrapy 引擎将输出文件传递给 Pipeline 处理。
12. Pipeline 处理 Item，可以进行数据清洗、去重、存储等操作。
13. 最终，Scrapy 引擎将处理后的数据输出到指定的存储介质中，例如数据库或文件。

## 3.2 数据集来源

### 3.2.1 美国国立卫生研究院(NIH)数据

NIH 是美国政府最大的医学研究机构，成立于 1930 年。其主要任务是推动医学研究，促进健康和预防疾病。NIH 负责资助和支持全球范围内的医学研究项目，包括基础研究、临床研究和转化研究等多个领域。NIH 在医学研究领域拥有世界领先的专家和科学家，致力于推动医学研究的进步，为全球的健康事业做出贡献。

### 3.2.2 美国国家自然科学基金会(NSF)数据

NSF 是美国政府主要的科学研究机构之一，成立于 1950 年。其主要任务是资助和支持基础科学研究，包括数学、物理、化学、生物学、计算机科学、工程学等多个领域。NSF 资助的研究项目涵盖了从基础研究到应用研究的全过程，旨在推动科学技术的发展和创新，为人类社会的发展做出贡献。NSF 在科学技术领域拥有世界领先的科学家和专家，致力于推动科学技术的进步和创新。

## 3.3 数据集概况

数据集主要包括了科研项目、科研机构、科研人员、科研论文四个本体，每一条科研项目信息都由一个科研机构承担、一个或很多个科研人员参与，而每个科研论文又由一个或很多个科研人员发表，对于 NIH 和 NSF 数据集中存在的部分特征及特征释义如表 3-1、表 3-2 所示。

表 3-1 NIH 部分特征

特征	特征类型	特征释义
ORG_CITY	Str	组织所属城市
ORG_COUNTRY	Str	组织所属国家
PROGRAM_OFFICER_NAME	Str	项目办公室名称
PROJECT_TERMS	Str	项目所属领域
PROJECT_TITLE	Str	项目标题
STUDY_SECTION	Str	研究所属部分
STUDY_SECTION_NAME	Str	研究所属部分名称
SUPPORT_YEAR	Int	资助时长
TOTAL_COST	Int	总花销
TOTAL_COST_SUB_PROJECT	Int	子项目总花销

表 3-2 NSF 部分特征

特征	特征类型	特征释义
Name	Str	组织名称
StateName	Str	组织所属州
ZipCode	Int	组织邮政编码
PI_FULL_NAME	Str	主要研究人员名称
AUTHOR_LIST	Str	论文作者
AbstractNarration	Str	项目介绍
AwardID	Int	项目ID
AwardEffectiveDate	Date	项目开始时间
AwardExpirationDate	Date	项目结束时间
AwardTitle	Str	项目标题
AwardAmount	Int	总花销
CFDA_NUM	Int	国家食品药品监督管理总局编号

## 第四章 数据预处理

数据预处理是指在数据挖掘之前对原始数据进行处理和转换,以便更好地适应挖掘算法和模型的需要。数据预处理包括数据清洗、数据变换、数据融合等过程。数据清洗主要是删除冗余、过滤错误、填补缺失等数据清理工作;数据变换主要是将数据进行规范化、离散化、标准化或特征选择等处理,提取有用的特征信息;数据融合主要是将来自不同数据源或不同数据格式的数据统一起来。通过数据预处理,使得挖掘算法可以更好地处理数据,提高数据挖掘的效果和可信度。

### 4.1 数据清洗

#### 4.1.1 删除多余数据

在读取数据至 Python 中以后,首先删除多余的特征,如 NIH 中的 ACTIVITY、ARRA\_FUNDED、ED\_INST\_TYPE、SERIAL\_NUMBER、SUFFIX 和 NSF 中的 Award、Value、PI\_FILL\_NAME、POR、DRECONTENT;并删除了 NSF 中为空值的数据,以避免知识图谱构建空节点。

#### 4.1.2 数据格式处理

NIH 中的 PI\_NAMES 和 PI\_IDS 的特征值中存在(contact)字符串,这可能会影响到科研人员的节点构建,因此删除该字符串部分。并且在 NIH 中的 PI\_NAMES 和 PI\_IDS 还存在多个科研人员之间用“;”分隔,为了实现每个科研人员成为一个独立的节点,把每条科研项目信息分为了多行,分别对应该科研项目的一名参与的科研人员;在论文的数据中,也存在着 AUTHOR\_LIST 中每个科研人员用“;”分隔的情况,本文采用了同样的处理方式。

而在 NIH 和 NSF 都存在部分特征值为 NAN,本文把 NAN 的部分都改为了“ ”的空字符串,以便于之后的知识图谱的节点特征的构建。NSF 中存在部分特征值含有“\n”和“\r”的字符串,本文把这两个字符串都改为了“ ”形式。

### 4.2 数据变换

#### 4.2.1 数据特征对齐

在 NIH 和 NSF 中存在某些含义相同但是特征名并不相同的情况,本文将 NSF 数据集中的特征对齐 NIH 的特征进行修改,PI\_FULL\_NAME、Name、AwardTitle、CityName、ZipCode、CountryName、StateName、AwardEffectiveDate、

AwardExpirationDate、AwardAmount、AbstractNarration、CFDA\_NUM、AwardID、LongName、Code、Abbreviation、Division 分别改为了 PI\_NAMES、ORG\_NAME、PROJECT\_TITLE、ORG\_CITY、ORG\_ZIPCODE、ORG\_CITY、ORG\_STATE、PROJECT\_START、PROJECT\_END、TOTAL\_COST、ABSTRACT\_TEXT、CFDA\_CODE、APPLICATION\_ID、NSF\_OGR、NSF\_ORG\_Code、NSF\_ORG\_Abbreviation、NSF\_ORG\_Division, 如表 4-1 所示。

表 4-1 特征修改前后

特征修改前	特征修改后
PI_FULL_NAME	PI_NAMES
Name	ORG_NAME
AwardTitle	PROJECT_TITLE
CityName	ORG_CITY
ZipCode	ORG_ZIPCODE
CountryName	ORG_CITY
StateName	ORG_STATE
AwardEffectiveDate	PROJECT_START
AwardExpirationDate	PROJECT_END
AwardAmount	TOTAL_COST
AbstractNarration	ABSTRACT_TEXT
CFDA_NUM	CFDA_CODE
AwardID	APPLICATION_ID
LongName	NSF_OGR
Code	NSF_ORG_Code
Abbreviation	NSF_ORG_Abbrevia tion
Division	NSF_ORG_Division

#### 4.2.2 特征选择

本文根据 Schema 设计, 分别根据科研人员、科研机构、科研项目以及科研论文的特征进行筛选, 四个本体的特征如图 4-1 中的 Schme 设计所示。

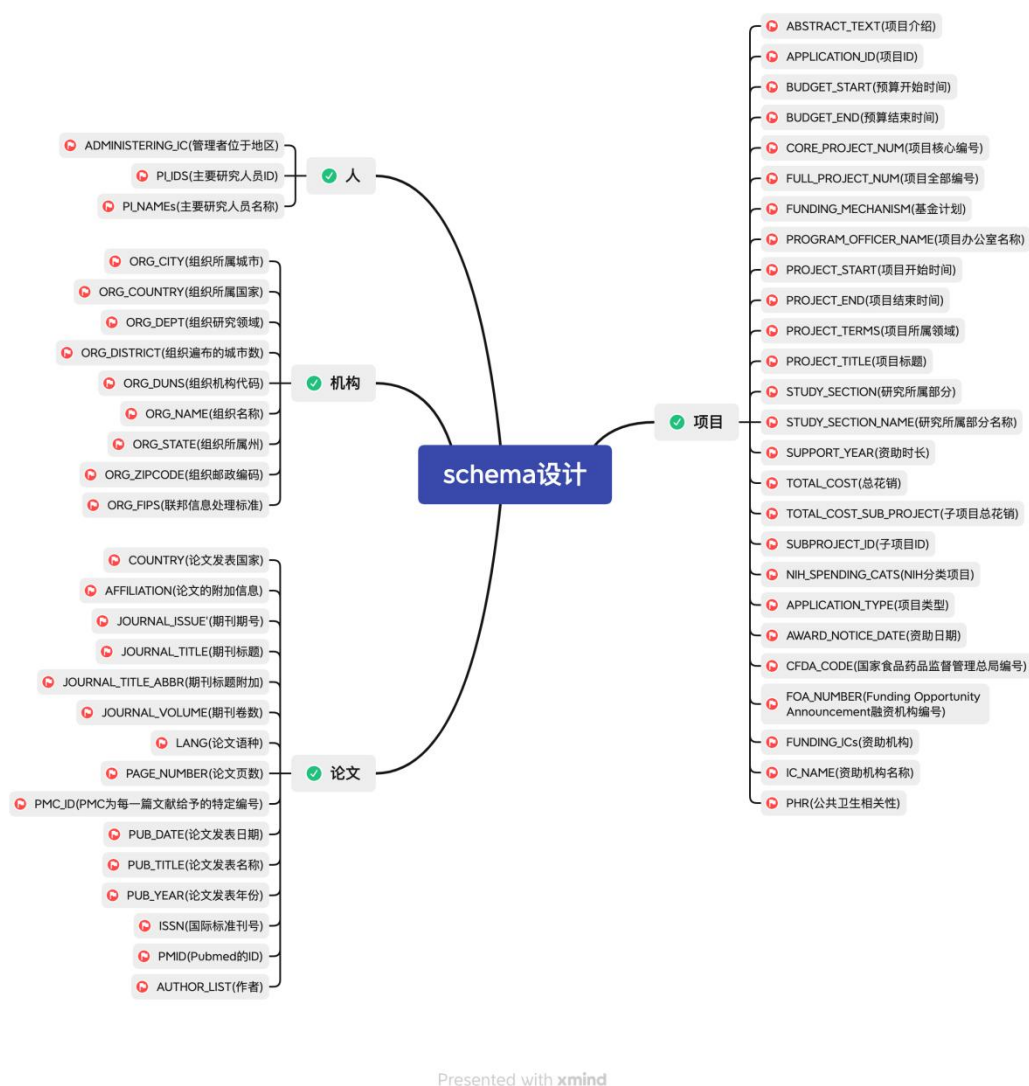


图 4-1 Schema 设计

### 4.3 数据融合

首先将 NIH 的数据合并为一个数据集，把每条科研项目的信息和科研项目的介绍数据根据项目的 ID——APPLICATION\_ID 进行合并；并根据每条科研项目的 CORE\_PROJECT\_NUM 与该科研项目所对应发表的论文 ID——PMID 合并；接着再把不同年份的科研项目合并为一个数据集。

由于之前已经完成了 NSF 中的数据特征与 NIH 中的数据特征对齐，于是这里只需把 NSF 中不同年份的科研项目合并。

接着合并 NIH 和 NSF 数据集，同时为了减小数据集的大小，本文保留了论文数据集中包含的科研人员姓名而删除了多余的论文数据。数据集的 UML 类图



和 protege 本体构建如图 4-2、图 4-3 所示<sup>[24]</sup>。

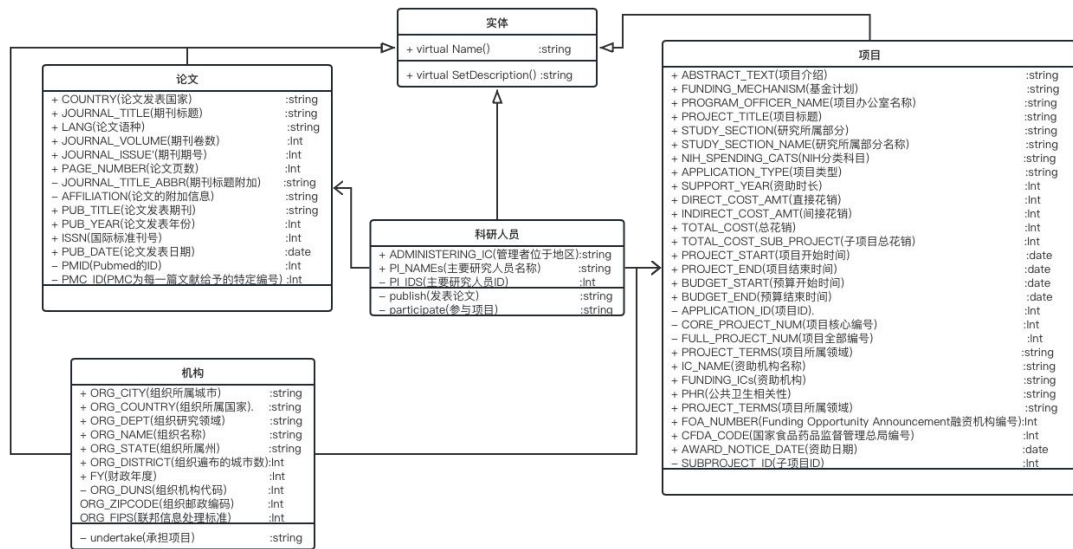


图 4-2 UML 类图

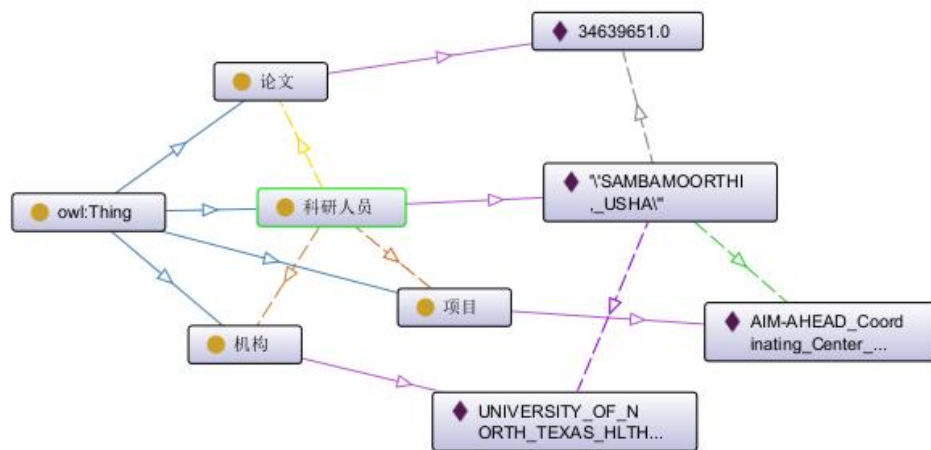


图 4-3 Protege 本体图

## 第五章 知识图谱构建

本文主要采取了图数据库中的 Neo4j 构建了美国科学研究系统的知识图谱, 使用 Python 构建 Neo4j 的知识图谱需要用到 Py2neo 库。

### 5.1 节点构建

#### 5.1.1 节点

数据集的节点和关系数量庞大, 因此本文使用了 `concurrent.futures` 库来进行多线程的节点构建, 根据 4 种不同的节点类型创建了一个字典, 再把字典转为全局变量实现高效的大量节点的构建。

#### 5.1.2 节点特征

根据生成的节点字典, 通过节点的 `name` 索引到该节点后再分别把该节点对应的节点特征添加至该变量中。

#### 5.1.3 节点实例化

构建完成节点与节点特征之后, 批量实例化节点即可, 这时 Neo4j 已经完成了所有节点的创建并可在 Neo4j 中可视化节点, 分别生成科研人员 34725 个节点、科研机构 4751 个节点、科研项目 44485 个节点、科研论文 104 个节点, 总计 84065 个节点, 如表 5-1 所示。

表 5-1 节点

节点类型	值
科研人员	34725
科研机构	4751
科研项目	44485
科研论文	104
总计	84065

## 5.2 关系构建

### 5.2.1 关系

数据集中每条信息可对应到每个科研人员参与到科研项目的情况、科研机构承担科研项目的情况与科研人员发表科研论文的情况，因此在数据集中构建不同实体之间的关系作为新列，并把每列实体与它们之间的关系生成一个列表。

根据列表中的节点特征 Name 构建不同节点之间的关系。

### 5.2.2 关系实例化

构建完成关系之后，批量实例化节点即可，这时 Neo4j 已经完成了所有关系的创建并可在 Neo4j 中可视化关系，分别生成科研人员参与科研项目 47725 个关系、科研机构承担科研项目 48083 个关系、科研人员发表科研论文 276 个关系，总计 96084 个节点，如表 5-2 所示<sup>[24]</sup>。

表 5-2 关系

关系类型	值
科研人员-科研项目	47725
科研机构-科研项目	48083
科研人员-科研论文	276
总计	96084

本文可视化了知识图谱关系最多的一个节点及 4 种不同节点间存在的 3 种关系，如图 5-1、图 5-2、图 5-3、图 5-4 所示。该节点为科研机构 Regents of the University of Michigan - Ann Arbor，其包含了总共 765 条关系，机构所在城市为 ANN ARBOR，机构所处国家为 United States，机构所属洲 Michigan，机构的邮政编码为 481091340。

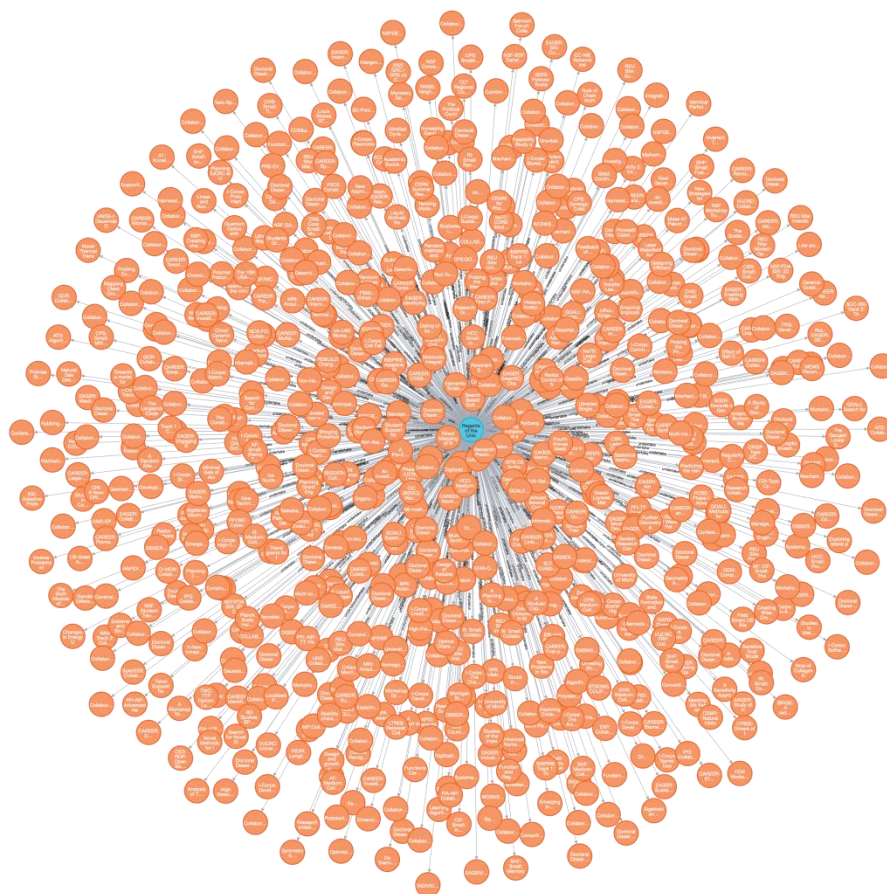


图 5-1 知识图谱可视化

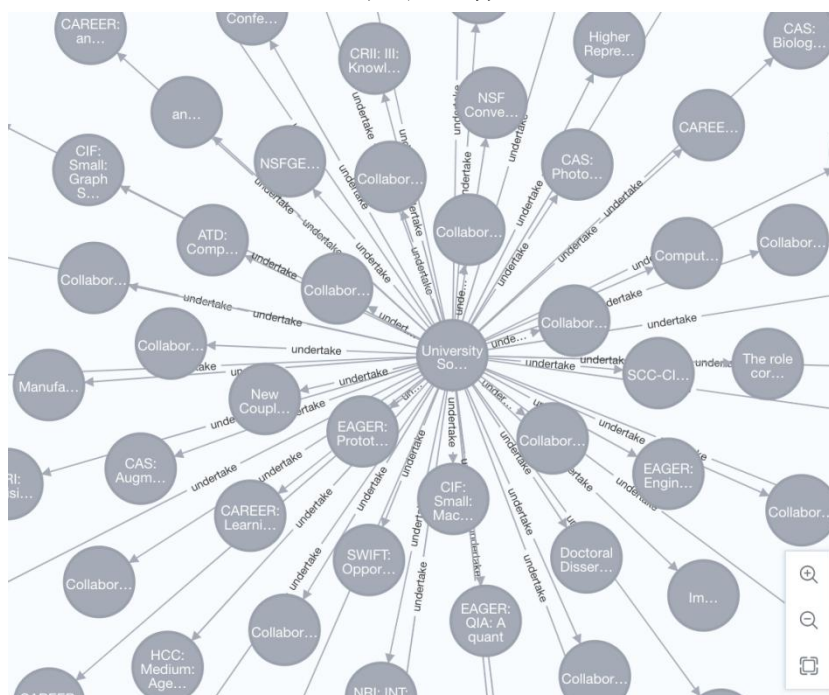


图 5-2 科研机构-科研项目

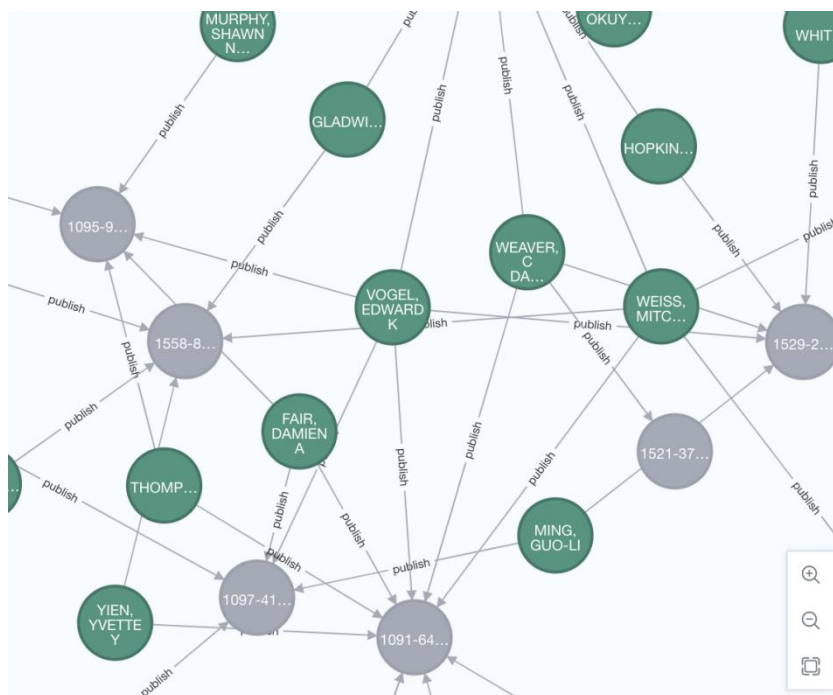


图 5-3 科研人员-科研论文

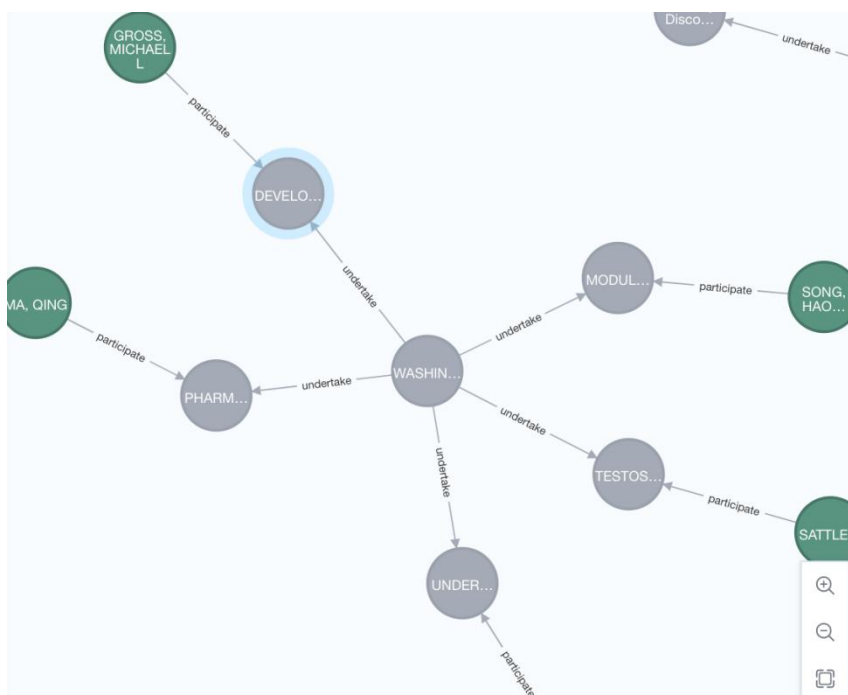


图 5-4 科研人员-科研项目

## 第六章 复杂网络分析

### 6.1 复杂网络

#### 6.1.1 网络生成

为了计算复杂网络的统计指标以便进一步获取知识图谱网络中潜藏的更多信息，使用 Cypher 语句将网络的信息转为了 networkx 网络进行更深层次的网络信息挖掘。

同时，该知识图谱的网络为有向无权重的网络，于是本文获取了节点、节点特征以及它们之间的关系后将网络转为了 networkx 中有向无权重的网络，并获取了网络的极大连通子图进行分析，其中极大连通子图包含节点数 71489 个、关系数 86177 条，如表 6-1 所示。

表 6-1 极大连通子图

指标	数量
节点	71489
关系	86177

### 6.2 网络的统计指标

#### 6.2.1 网络密度

网络密度是指一个网络中实际存在的连接数与可能存在的连接数之比。在复杂网络中，网络密度通常用来描述网络中节点之间的紧密程度，即网络中节点之间的联系程度。如果一个网络中的节点之间联系非常紧密，那么该网络的密度就会很高，反之则会很低。网络密度可以帮助我们了解网络中节点之间的关系，进而研究网络的结构、功能和演化等问题。在实际应用中，网络密度也常常被用来作为网络性能的评价指标，在社交网络中，高密度的网络通常意味着社交圈子比较紧密，信息传递速度比较快。网络密度计算公式如 (6-1) 所示。

$$D = \frac{m}{\frac{1}{2}n(n-1)} \quad (6-1)$$

其中  $m$  表示网络中实际存在的边数， $n$  表示节点的数量，若为无向图，则分母为  $\frac{1}{2}n(n-1)$ ，若为有向图，则分母为  $n(n-1)$ 。

## 6.2.2 度

在复杂网络中，度是指一个节点连接的边的数量。最大度是指网络中所有节点的度中的最大值，最小度则是指所有节点的度中的最小值。平均度是指网络中所有节点的度的平均值。

度可以帮助我们了解网络中节点的重要性和连接情况。一个节点的度越高，通常意味着它在网络中的地位越重要，它所连接的其他节点也越多。最大度和最小度则可以帮助我们了解网络中节点度数的分布情况，进而研究网络的结构特征。如果一个网络的最大度比较高，那么该网络中可能存在一些重要节点，这些节点对整个网络的影响比较大。如果一个网络的最小度比较低，那么该网络中可能存在一些孤立节点，这些节点对整个网络的影响比较小。

平均度可以帮助我们了解网络中节点的平均连接情况。一个网络的平均度越高，通常意味着网络中节点之间的联系比较紧密，信息传递速度也比较快。平均度还可以帮助我们了解网络的稠密程度，如果一个网络的平均度比较高，那么该网络可能比较密集，节点之间的联系比较紧密。

当我们研究复杂网络的拓扑结构时，度分布是一个非常重要的指标。度分布指的是网络中各个节点的度数出现的频率分布情况。具体来说，度分布可以用一个概率分布函数来表示，该函数描述了在网络中一个节点的度数为  $k$  的概率。

在研究度分布时，常常会将网络中的节点按照度数从小到大排序，然后统计每种度数出现的次数，最终得到一个度分布序列。度分布序列可以帮助我们了解网络中节点度数的分布情况，进而研究网络的结构特征。

在实际应用中，度分布通常呈现出幂律分布的特征。幂律分布指的是一个概率分布函数，如 (6-2) 所示。幂律分布的特点是在一定范围内，只有少数节点的度数非常大，而大部分节点的度数比较小。这种分布形式在很多复杂网络中都得到了验证，例如社交网络、互联网、蛋白质相互作用网络等。

$$P(k) \propto k^{-\gamma} \quad (6-2)$$

其中  $x$  表示度数， $\gamma$  是一个常数。

## 6.2.3 度相似性系数

复杂网络的度相似性系数是用来描述网络中两个节点之间连接的相似程度的一个指标。它是指在一个网络中，两个节点的邻居节点中相同的节点数占它们邻居节点总数的比例。度相似性系数可以用来衡量网络中节点之间的相似性。

在同配性和异质性的研究中，度相似性系数也有着重要的应用。同配性是指网络中度相近的节点之间更容易相连，而异质性是指网络中度不同的节点之间更

容易相连。度相似性系数可以用来衡量同配性和异质性。当系数为正时，表示度较大的节点倾向于连接度较大的节点，网络呈现同质性，反之则相反，网络呈现异配性。度相似性系数计算公式如 (6-3) 所示。

$$r = \frac{\sum_{xy} xy(e_{xy} - a_x b_y)}{\sigma_a \sigma_b} \quad (6-3)$$

其中  $x$ 、 $y$  分别表示两个不同节点的度数， $e_{xy}$  表示度为  $x$  的节点和度为  $y$  节点连接的边数再除以总边数， $a_x$ 、 $b_y$  分别表示出度为  $x$ 、 $y$  的节点集，其出度总和占总边数的比例  $\sigma_a$ 、 $\sigma_b$  分别表示  $a_x$  和  $b_y$  的标准差。

#### 6.2.4 平均聚类系数

复杂网络的平均聚类系数是用来描述网络中节点聚集程度的一个指标。它是指一个节点的邻居节点之间实际连接数与可能连接数之比的平均值。其中，可能连接数是指一个节点的邻居节点之间可能存在的连接数，即邻居节点数目的二项式系数。平均聚类系数可以用来衡量网络中节点之间的紧密程度。如果一个节点的邻居节点之间连接较多，那么该节点的平均聚类系数就较高；反之，如果一个节点的邻居节点之间连接较少，那么该节点的平均聚类系数就较低。节点  $i$  的聚类系数及平均聚类系数计算公式如 (6-4)、(6-5) 所示。

$$C_i = \frac{2x}{k_i(k_i-1)} \quad (6-4)$$

其中  $C_i$  表示与第三个节点连接的一对节点被连接的概率， $x$  表示节点  $i$  的度数， $k_i$  表示节点  $i$  的邻居节点数。

$$C = \frac{1}{N} \sum_{i=1}^N C(i) \quad (6-5)$$

其中  $N$  表示节点数， $C(i)$  表示节点  $i$  的聚类系数。

#### 6.2.5 k 核

复杂网络的  $k$  核是指网络中最大的一个子图，其中每个节点至少与  $k$  个节点相连。 $k$  核可以用来衡量网络中节点之间的稳定性和连接程度。如果一个节点在  $k$  核中，那么它与其他节点之间的连接比较稳定，且至少与  $k$  个节点相连。 $k$  核可以用来识别网络中的重要节点和社区结构。在社交网络分析、生物信息学、交通网络和电力系统等领域， $k$  核有着广泛的应用。可以用来识别互联网中的重要网站和社区结构，从而进行网络优化和安全监测。

### 6.3 指标计算

对于复杂网络的统计指标，主要对网络的网络密度、最大度、最小度、平均度、度相似性系数、平均聚类系数、最大  $k$  核进行计算，如表 6-2 所示，并绘制



了度的相关分布情况以及 k 核的相关分布情况，如图 6-1、图 6-2 所示。

表 6-2 网络的统计指标

统计指标	值
网络密度	0.00001686239
最大度	765
最小度	1
平均度	2.41091636476
度相似性系数	-0.0543612236
平均聚类系数	0.000000000000
最大 k 核	3

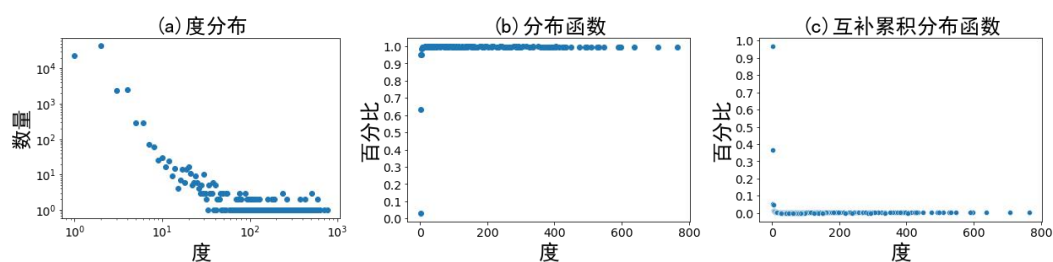


图 6-1 度分布及分布函数情况

(a) 度分布; (b) 度分布函数; (c) 度互补累积分布函数

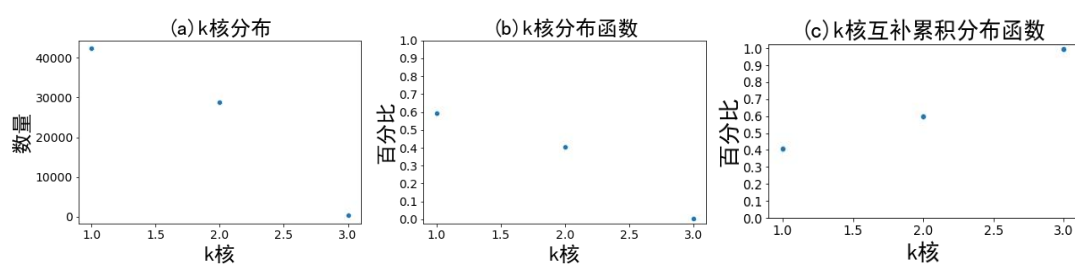


图 6-2 k 核分布及分布函数情况

(a) k 核分布; (b) k 核分布函数; (c) k 核互补累积分布函数

## 第七章 SMW 平台与可视化

本文实现了对 Semantic Mediawiki (简称 SMW) 平台的开发, 以平台的页面作为知识图谱中的节点, 页面之间的联系作为关系, 在 SMW 平台上实现了美国科学研究系统的知识图谱的构建。

### 7.1 SMW 平台开发

在创建 SMW 的页面之前, 首先需要创建四个本体的属性、分类、模版、表单, 接着再把本体对应的数据导入页面中。

#### 7.1.1 属性

“属性”为 MediaWiki 的页面添加了更细致的数据管理; 属性及类型是 SMW 中定义结构化数据 (或语义数据) 的基本方式。可以将属性视为对页面 (或实体) 的一系列规范化的描述词。

定义 SMW 平台中页面的属性相当于知识图谱中每个节点的节点特征, 因此在创建页面之前先对四个本体分别创建它们的属性, 平台中期刊论文的属性如图 7-1 所示。

自定义主表字段:

显示 10 记录

搜索:

导出 ▼

数据层级	字段名	字段含义	Wiki 属性类型	Wiki 值列表	Freetext(非结构化)字段	InfoBox(结构化)字段	SQL 类型	属性特征	转置属性	修订详细描述
主表	URL	论文页面链接	URL	否	否	是	VARCHAR	无		赵莹修订
主表	Affiliation	论文的附加信息	Text	否	否	是	VARCHAR	无		蒋世华修订
主表	JournalVolume	期刊卷数	Text	否	否	是	VARCHAR	无		蒋世华修订
主表	PMID	Pubmed 的 ID	Number	否	否	是	VARCHAR	无		蒋世华修订
主表	PageNumber	论文页数	Number	否	否	是	VARCHAR	无		蒋世华修订
主表	JournalIssue	期刊期号	Number	否	否	是	VARCHAR	无		蒋世华修订
主表	ISSN	国际标准刊号	Number	否	否	是	VARCHAR	无		蒋世华修订

图 7-1 SMW 平台属性

### 7.1.2 分类

分类是 MediaWiki 的一项功能，能够自动索引页面，为读者提供某一主题下的页面实体列表。只需给页面的维基文本中加上一个或多个 Category 标记即可将页面归类。这些标记将在页面底部创建指向分类页面的链接，从而可以很方便地查看同一分类下的其他相关文章。在语义化 Wiki 中，常常使用分类来筛选页面实体。

定义 SMW 平台中页面的分类相当于知识图谱中不同标签的个体，对四个个体分别创建它们的分类，如图 7-2 所示。



图 7-2 SMW 平台分类

### 7.1.3 模版

模板是一种标准 wiki 页面，但它主要会被嵌入到其它页面中。模板的页面名称最前面都有“Template:”或者“模板:”——将它分配到该命名空间。

定义 SMW 平台中页面的模版为 SMW 平台的页面提供了统一的标准化，模版为后续批量导入数据并创建页面提供了统一的页面格式，如图 7-3 所示。

这是“科研论文基本信息”模板。调用它时应该采用如下格式：

```
{{科研论文基本信息
|Country=
|Affiliation=
|JournalIssue=
|JournalTitle=
|JournalTitleAbbr=
|JournalVolume=
|Language=
|PageNumber=
|PMCID=
|PubDate=
|PubTitle=
|PubYear=
|ISSN=
|PMID=
|Author=
}}
```

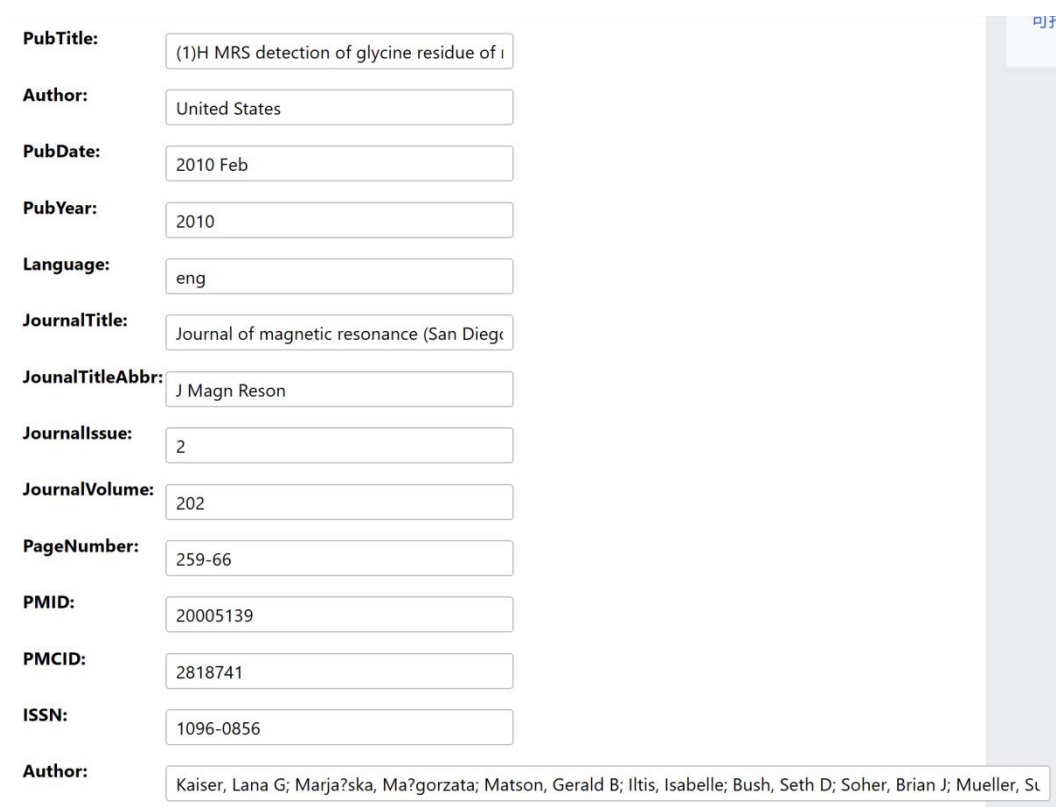
编辑页面以阅读模板文本。

图 7-3 SMW 平台模版

### 7.1.4 表单

在完成了创建属性、创建分类、创建模板之后，对于 MediaWiki 的页面实体控制已经可以实现批量化、自动化的创建和修改了。而为了便于非专业用户编辑和使用，Page Forms 还提供了“表单”功能。

定义 SMW 平台中页面的表单便于在 SMW 平台中展示每个实体的属性，对四个本体分别创建它们的表单，如图 7-4 所示。



PubTitle:	(1)H MRS detection of glycine residue of i
Author:	United States
PubDate:	2010 Feb
PubYear:	2010
Language:	eng
JournalTitle:	Journal of magnetic resonance (San Dieg
JournalTitleAbbr:	J Magn Reson
JournalIssue:	2
JournalVolume:	202
PageNumber:	259-66
PMID:	20005139
PMCID:	2818741
ISSN:	1096-0856
Author:	Kaiser, Lana G; Marja?ska, Ma?gorzata; Matson, Gerald B; Iltis, Isabelle; Bush, Seth D; Soher, Brian J; Mueller, Su

图 7-4 SMW 平台表单

## 7.2 数据导入及可视化

知识图谱中存在大量的节点和关系，因此本文使用了 Python 批量创建 SMW 平台的页面，并使用 Network 插件可视化页面之间的联系，从而构建知识图谱。Network 插件允许通过一个交互式网络图可视化维基页面之间的连接。实体页面是网络图中的节点，实体页面之间的相互链接关系是网络图中的有向边。平台界面及知识图谱展示如图 7-5、图 7-6 所示。

# (1)H MRS detection of glycine residue of reduced glutathione in vivo.

20005139

Varian Incorporated, Palo Alto, CA, USA. lana.kaiser@varianinc.com

目录

隐藏

- 1 知识图谱
- 2 期刊论文编码
- 3 期刊论文发表
- 4 期刊
- 5 科研论文期刊附加

三 语义属性 表单编辑

期刊论文基本信息

PubTitle	(1)H MRS detection of glycine residue of reduced glutathione in vivo.
Country	United States
PubDate	2010 Feb
PubYear	2010
Language	eng
JournalTitle	Journal of magnetic resonance (San Diego, Calif. :

图 7-5 SMW 平台界面

知识图谱

编辑

JournalTitleAbbr	1997) J Magn Reson
JournalIssue	2
JournalVolume	202
PageNumber	259-66
PMID	20005139
PMCID	2818741
ISSN	1096-0856
Author	Kaiser, Lana G, Marja?ska, Ma? gorzata, Matson, Gerald B, Iltis,

图 7-6 SMW 平台知识图谱界面

31

## 第八章 总结

### 8.1 结论

由网络中节点数量、关系数量、网络密度以及平均聚类系数可看出该网络属于稀疏图；且最大度为 765，最小度为 1，平均度仅为 2.41091636476，而度分布又呈现幂律分布，在网络中存在较少的节点拥有较大的度，而大部分节点的度都较小，符合无标度网络的特性；由度相似性系数为-0.0543612236 可看出度较大的节点可能更倾向于与度较小的节点相连，在美国科学研究系统中，科研机构或科研人员更倾向于与未合作过的进行合作，而不仅仅是大度节点与大度节点的合作，网络呈现了一定的异配性；而在图 5-3(a)所示的 k 核分布中，随着 k 核越大数量越少也符合度分布的情况，大部分节点都为小度节点，仅有少量节点拥有大度。

构建于 SMW 平台的知识图谱为用户提供了一些简单的分析工具，可以帮助用户对知识图谱进行分析和探索。用户可以通过查询和浏览知识图谱中的实体和关系，发现各种科学研究中的规律和趋势，从而提出更加深入和具有前瞻性的研究问题和方向。

同时 SMW 平台还能够为公众或者科研人员提供方便的参考和查阅工具。通过浏览知识图谱中的实体和关系，公众或者科研人员可以了解各种科学研究的发展历程、研究成果、研究方法等方面的信息，有助于提高公众的科学素养和科研人员的研究效率。

在 SMW 平台上构建知识图谱能够为科学研究提供更加全面和系统的视角，为公众和科研人员提供方便的参考和查阅工具，有助于推动科研的深入和发展。

### 8.2 不足与展望

美国学术研究网络的知识图谱已经初步完成图谱可视化和相关网络性质的探索，但是很多方面还不够完善，存在数据源不完整、数据处理过程相对粗糙、网络的性质未探索完全等等。

从数据的角度看待，首先数据源不够全面，导致数据处理过程过于粗糙。其次数据 Schema 仍不够细致，存在部分重要的信息没有包含在内，导致在爬取及数据处理过程中缺少部分重要信息。其次科研人员的姓名可能存在重名等问题。

构建一个更加强大和智能的知识图谱对数据集以及数据集中各类本体、关系的明确提出了一个更高的要求，而目前在大数据及数据的全面性方面仍有一定的局限性，在构建知识图谱时这方面的挑战仍然是一个有待进一步研究的课题。

## 参考文献

- [1] 李志民.美国科研机构概览[J].世界教育信息,2018,31(05):6-10.
- [2] 育东.美国科研机构的运作与管理[J].全球科技经济瞭望,1995(04):5-8.
- [3] Liu, W., et al. Method of constructing knowledge graph network based on massive scientific research data, involves extracting subject word from title information of setting document of each topic as key technology, and clustering direction of topics, State Computer Network & Information Saf (St cn-C).[P].2018-04-27.
- [4] Z. Ye, Y. J. Kumar, G. O. Sing, F. Song and J. Wang, "A Comprehensive Survey of Graph Neural Networks for Knowledge Graphs," in IEEE Access, vol. 10, pp. 75729-75741.
- [5] Huang, S., Wan, X. (2013). AKMiner: Domain-Specific Knowledge Graph Mining from Academic Literatures. In: Lin, X., Manolopoulos, Y., Srivastava, D., Huang, G. (eds) Web Information Systems Engineering – WISE 2013. WISE 2013. Lecture Notes in Computer Science, vol 8181. Springer, Berlin, Heidelberg.
- [6] Song, S., et al. Science and technology big data knowledge graph based in telligent technical diagnosis expert matching algorithm, has set of instructions for developing technology evaluation, business plan, decision consultation, front analysis and market prediction value-added service, BEIJING WAN FANG SOFTWARE CO LTD (BEIJ- Non-standard).[P].2022-12-13.
- [7] H. Cai, Z. Liu and C. Wang, "Intelligent recommendation system based on knowledge graph for scientific research teams," 2021 13th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, China, 2021, pp. 204-207.
- [8] 雷洁,赵瑞雪,李思经等.知识图谱驱动的科研档案大数据管理系统构建研究[J].数字图书馆论坛,2020,No.189(02):19-27.
- [9] 杜悦,常志军,董美,钱力,王颖.一种面向海量科技文献数据的大规模知识图谱构建方法[J].数据分析与知识发现,2023,7(02):141-150.
- [10] Zhu G , Iglesias C A . Exploiting Semantic Similarity for Named Entity Disambiguation in Knowledge Graphs[J]. Expert Systems with Applications, 2018, 101(JUL.):8-24.



- [11] Danae Pla Karidi,Yannis Stavrakas,Yannis Vassiliou. Tweet and followee personalized recommendations based on knowledge graphs[J]. Journal of Ambient Intelligence and Humanized Computing,2018,9(6):2035-2049.
- [12] 杨思洛,韩瑞珍.知识图谱研究现状及趋势的可视化分析[J].情报资料工作,2012(04):22-28.
- [13] 李思志,李佳骏,李艳红.管理科学与工程领域的创新轨迹研究——基于 TOP 期刊的文献计量和文本挖掘视角[J].中国管理科学,2014,22(S1):56-62
- [14] 王洪旭,温晓会,刘万明.基于知识图谱的高校科研经费管理研究[J].行政事业资产与财务,2022(05):108-110.
- [15] 雷洁,赵瑞雪,李思经,鲜国建,寇远涛.知识图谱驱动的科研档案大数据管理系统构建研究[J].数字图书馆论坛,2020,No.189(02):19-27.
- [16] Raza Shaina,Schwartz Brian. Entity and relation extraction from clinical case reports of COVID-19: a natural language processing approach[J]. BMC Medical Informatics and Decision Making,2023,23(1):1-1.
- [17] Saad Mohamed,Zhang Yingzhong,Tian Jinghai,Jia Jia. A graph database for life cycle inventory using Neo4j[J]. Journal of Cleaner Production,2023,393(1):1-1.
- [18] Zhang Y, Li X, Yang Y, Wang T. Disease- and Drug-Related Knowledge Extraction for Health Management from Online Health Communities Based on BERT-BiGRU-ATT. International Journal of Environmental Research and Public Health. 2022,19(24):16590.
- [19] D'Auria Daniela,Moscato Vincenzo,Postiglione Marco,Romito Giuseppe,Sperli Giancarlo. Improving graph embeddings via entity linking: A case study on Italian clinical notes[J]. Intelligent Systems with Applications,2023,17(1):1-1.
- [20] Bratsas C, Chondrokostas E, Koupidis K, Antoniou I. The Use of National Strategic Reference Framework Data in Knowledge Graphs and Data Mining to Identify Red Flags. Data. 2021,6(1):2.
- [21] Shen, Y., et al. Knowledge graph constructing, searching and visualization method, involves finishing query in knowledge graph presenting user search content in visual pattern form in responding to personalized retrieval requirement of user, and improving user search experience, Univ Shanghai (Ush n-C).[P].2022-11-08.

- [22] Liang Lu, Li Yong, Wen Ming, Liu Ying. KG4Py: A toolkit for generating Python knowledge graph and code semantic search[J]. Connection Science, 2022, 34(1):1384-1400.
- [23] Ahmadnia Benyamin, Dorr Bonnie J., Kordjamshidi Parisa. KNOWLEDGE GRAPHS EFFECTIVENESS IN NEURAL MACHINE TRANSLATION IMPROVEMENT[J]. COMPUTER SCIENCE-AGH, 2020, 21(3):1-1.
- [24] 王冬梅. 高校财务智能客服知识图谱研究[J]. 中国管理信息化, 2022, 25(16):77-79.
- [25] 鲁效平, 江民圣, 孙明等. 基于 CiteSpace 的全球智能家居行业研究知识图谱分析[J]. 家电科技, 2022, 415(02):56-60.
- [26] 周晟宇. 基于 CiteSpace 的企业知识管理研究分析[J]. 科技资讯, 2023, 21(02):219-222.
- [27] 袁俊, 刘国柱, 梁宏涛等. 知识图谱在商业银行风控领域的研究与应用综述[J]. 计算机工程与应用, 2022, 58(19):37-52.
- [28] Liu, Y., et al. Medical use-based knowledge graph system, has medical application layer for providing service to user based on knowledge graph, where medical application layer comprises decision assistance module that provides decision assistance scheme for user, Univ Northeastern (Unen-C).[P]. 2022-11-04.
- [29] 孙欣. 基于大数据的中医药高层次人才学术研究合作网络的知识图谱——以南阳地区为例[J]. 中国科技信息, 2022, No.687(22):117-119.
- [30] 蓝乾栋. 我国智慧城市研究进展及趋势——基于 CiteSpace 的知识图谱分析[J]. 韩山师范学院学报, 2022, 43(06):25-32+37.
- [31] Mubeen Sarah, Domingo Fernández Daniel, Díaz del Ser Sara, Solanki Dhvani M., Kodamullil Alpha T., Hofmann Apitius Martin, Hopp Marie T., Imhof Diana. Exploring the Complex Network of Heme-Triggered Effects on the Blood Coagulation System[J]. Journal of Clinical Medicine, 2022, 11(19):5975.

## 致 谢

长达半年的毕业设计终于要告一段落，在完成毕业设计论文的同时，我要衷心感谢我的导师王文俊教授，在整个研究过程中给予我的指导、鼓励和支持。他的深刻见解、建设性的批评和耐心的指导对于塑造我的研究方向和帮助我成长为一名研究人员起到了至关重要的作用。

我还要感谢李天鹏博士和刘基业硕士，在我的项目中提供了宝贵的帮助和建议。他们的专业知识和深刻见解在帮助我克服各种挑战和障碍方面发挥了不可替代的作用。

最后，我要感谢我的家人和朋友们在我学术之旅中给予我的不懈支持和鼓励。他们的爱、鼓励和对我的信任一直是我不断前行的动力和源泉，我深深感激他们的坚定支持，希望本人能在研究的道路上继续不断钻研、不忘初心。