

ForPKG-1.0: 林业政策知识图谱构建框架及应用分析

第一作者
工作关系/ 地址一
email@domain

第二作者
工作关系/ 地址一
email@domain

摘要

政策知识图谱能够为项目合规、政策分析、智能问答等任务提供决策支持，并可作为外部知识库，辅助相关大语言模型的推理过程。尽管有关知识图谱的相关工作已经很多，但目前尚缺乏对政策知识图谱构建方法的研究。本文面向林业领域，设计了一个完整的政策知识图谱构建框架，包括：首先提出一个细粒度的林业政策领域本体；然后提出一个无监督的政策信息抽取方法，最终构建出了一个完整的林业政策知识图谱。实验结果表明，所设计的本体具有良好的表达性和可扩展性，所提出的政策信息抽取方法比其他无监督方法能取得更好的效果。此外，通过对该知识图谱在大模型检索增强生成（Retrieval-Augmented-Generation）任务上的应用分析，证实了该知识图谱在大模型时代的实践应用价值。该知识图谱资源将发布在开源平台上，可作为林业政策相关智能系统的基础知识库，也可用于相关学术研究。此外，该研究也可为其他领域的政策知识图谱构建工作提供借鉴和参考。

关键词： 政务知识工程；知识图谱构建；大模型

ForPKG-1.0: A Framework for Constructing Forestry Policy Knowledge Graph and Application Analysis

First Author
Affiliation / Address line 1
email@domain

Second Author
Affiliation / Address line 1
email@domain

Abstract

A policy knowledge graph can provide decision support for tasks such as project compliance, policy analysis, and intelligent question answering, and can also serve as an external knowledge base to assist the reasoning process of related large language models. Although there have been many related works on knowledge graphs, there is currently a lack of research on the construction methods of policy knowledge graphs. This paper, focusing on the forestry field, designs a complete policy knowledge graph construction framework, including: firstly, proposing a fine-grained forestry policy domain ontology; then, proposing an unsupervised policy information extraction method, and finally, constructing a complete forestry policy knowledge graph. The experimental results show that the proposed ontology has good expressiveness and extensibility, and the policy information extraction method proposed in this paper achieves better results than other unsupervised methods. Furthermore, by analyzing the application of the knowledge graph in the retrieval-augmented-generation task of the large model,

the practical application value of the knowledge graph in the era of large models is confirmed. The knowledge graph resource will be released on an open-source platform and can serve as the basic knowledge base for forestry policy-related intelligent systems. It can also be used for academic research. In addition, this study can provide reference and guidance for the construction of policy knowledge graphs in other fields.

Keywords: Knowledge Engineering of Government Affairs , Knowledge Graph Construction , Large Language Model

1 引言

知识图谱是一种以图形结构表示知识的方法，其中节点表示现实世界中的实体或概念，边表示实体之间的关系 (Z. Wang et al., 2014)。知识图谱为管理、组织和利用结构化数据提供了一种有效的解决方案，其在搜索引擎、推荐系统和智能问答等方面有着广泛的应用，能够为用户提供更加智能、个性化的服务 (X. Chen et al., 2020; Fensel et al., 2020)。此外，知识图谱也在医疗健康、金融、物联网等领域发挥着重要作用，为相关领域的决策和创新提供支持 (A. Li et al., 2024; Ouyang et al., 2024)。近年来，尽管大模型 (Large Language Model, LLM) 在预训练阶段能够学习到越来越多的先验语义知识，但知识图谱仍然是辅助LLM推理的重要知识表示模式，其研究热度不减反增。

已有的知识图谱大多是开放域知识图谱，例如Dbpedia (Auer et al., 2007)、CN-Dbpedia (Xu et al., 2017)、YAGO (Suchanek et al., 2007)以及OpenConcept (叶宏彬et al., 2023)等。这些知识图谱通常刻画的是现实世界中常见的实体及其关系，例如公司、歌手、演员等实体以及父子、配偶、朋友等关系。此外也有一些针对限定域知识图谱的相关研究，例如海上事故知识图谱 (Gan et al., 2023)、军事知识图谱 (赵颜利et al., 2020)以及电商知识图谱 (X. Wang et al., 2023)等。政策知识图谱不同于其他限定域知识图谱，其重点并非是对客观世界中真实存在的对象之间的关联进行刻画，而更多的是对概念、术语、定义、解释、规范以及道义模态等抽象知识进行刻画。一个完备的政策知识图谱能够辅助合规性检测、智能政策问答、智能政策检索、政务决策支持等任务，具有显著意义。但是，据本文所知，有关政策知识图谱的相关研究并不多见。

已有的政策知识图谱相关研究仅建立了政策文件与元数据之间的关联，例如发文日期、发文机构、适用范围等 (郭佳航, 2021)，缺少对政策文本内部语义的细粒度刻画。相比而言，本文以林业领域为落脚点，为林业政策知识图谱设计了一个细粒度的本体，不仅将通用知识图谱中常见的实体和关系类型迁移过来，还根据林业政策领域特点和道义逻辑理论体系提出了领域特点鲜明的实体和关系类型。此外，知识抽取一直是知识图谱构建的关键过程。现有的基于有监督模型的知识抽取方法往往需要大量的有标签数据。在林业政策领域，有标签的数据较少，且构建成本较高。因此，本文提出一种基于开源大模型的林业政策知识抽取流程。该流程分别执行头实体抽取、关系分类和尾实体识别，从而效识别林业政策文本中的抽象关系以及长难实体类型。

本文所构建的林业政策知识图谱将在论文录用后发布到开源平台上。这一资源可作为中国林业科学研究院科信所发布的林草领域知识图谱的重要补充，为林业工程合规、林业政策分析、林业智能问答等任务提供支持和丰富。此外，本文所设计的本体和提出的抽取流程不仅局限于林业领域，还可扩展到其他政策领域，如医疗、保险、金融等。因此，本文所提供的方法论和实践经验也可对其他领域的政策知识图谱构建工作提供借鉴和参考。

总的来说，本文的主要贡献如下：

- 1) 提出了一套完整的政策知识图谱构建框架
- 2) 发布了首个林业政策知识图谱资源
- 3) 设计了首个细粒度的林业政策知识图谱本体
- 4) 提出了一个基于开源大模型的政策知识抽取流程
- 5) 探讨了政策知识图谱在大模型RAG上的应用实践

2 相关研究

2.1 知识图谱

知识图谱通过节点（实体）和边（关系）的形式化表达，刻画现实世界中的各种概念、对象及其相互之间的联系 (Peng et al., 2023; Zhong et al., 2023)。根据其应用范围，可划分为开放域知识图谱 (Zirui et al., 2021) 和限定域知识图谱 (Abu-Salih, 2021)。开放域知识图谱主要描述现实世界中常见的实体类型及其关系，例如足球运动员、演员、城市等实体类，以及父子关系、配偶关系、朋友关系等。这类知识图谱最常见，其相关研究也最多，例如Dbpedia (Auer et al., 2007)（目前最大的英文开放域知识图谱）、CN-Dbpedia (Xu et al., 2017)（使用频率最高的开放域中文知识图谱）、YAGO (Suchanek et al., 2007)、WikiData (Vrandečić and Krötzsch, 2014) 以及 OpenConcept (叶宏彬 et al., 2023) 等。然而，开放域知识图谱仅包含通用的世界知识，缺少特定领域的细粒度知识，因此在一些领域任务上无法发挥作用。

与开放域知识图谱相比，限定域知识图谱的范围更为狭窄，但它在特定领域内提供了更深入、更细致的信息和知识。因此，限定域知识图谱在不同的领域任务上都有着广泛的应用场景 (P. Chen et al., 2018; Fang et al., 2023; Tang et al., 2023)。例如，在医疗诊断任务中，知识图谱能够辅助进行疾病、病因、药物、诊疗方案之间的逻辑推理 (Abu-Salih et al., 2023; 陈静 et al., 2023)。在金融科技领域，智月平构建了一个以股票为核心的面向金融证券领域的知识图谱 (智月平, 2024)。Gan等人 (Gan et al., 2023) 构建了一个海上事故知识图谱，以显示事故重要因素之间的相关性，从而对海上事故原因进行分析并对未来潜在海上事故进行预测。此外，限定域知识图谱在电商推荐领域也发挥着重要作用，例如Wang等人 (X. Wang et al., 2023) 构建了一个图文多模态电商知识图谱，以增强电商系统的图文检索能力。在军事领域，赵等人 (赵颜利 et al., 2020) 设计了一个战例本体模型，建立了战例本体与实体的映射机制，并据此构建了一个战例知识图谱。

对于政策法规领域的知识图谱构建，目前已知的相关工作不多，并且已有的工作都类似于文献网络的构建，仅建模了政策文本之间及政策文本和各类元数据之间的关系，没有深入到政策文本内部对更细粒度的语义关联进行刻画。例如，郭佳航 (郭佳航, 2021) 构建了一个科技政策知识图谱，其中的实体类包括：政策标题、主要内容、省份、发布机构、发布日期等。该研究将整个政策文本作为完整的实体，因此阻碍了对政策内部的语义进行逻辑推理。与该工作类似，支凤稳等人 (支凤稳 et al., 2024) 构建了一个科技报告政策知识图谱。尽管该工作对本体进行了更加细致的设计，但仍然只对政策文本与外部的关联进行了建模。据我们所知，目前尚不能够同时刻画政策文本外部和内部语义关联的知识图谱构建工作。

2.2 信息抽取

知识图谱的构建过程离不开信息抽取，即从自由文本中抽取结构化的实体、关系以及属性等 (L. Li et al., 2020)。传统的信息抽取方法大多将实体抽取任务建模为一个序列标注问题，并训练一个神经网络模型来预测文本中实体提及的起止位置 (Liu et al., 2022)。在识别出文本中的实体提及后，可以进一步据此来预测实体之间的关系类别 (Zeng et al., 2014)。此外，也有一些方法从文本中联合抽取实体和关系，从而避免实体抽取过程中产生的误差传播到关系识别过程 (Y. Chen et al., 2020)。尽管这些传统的信息抽取方法已经具有十分成熟的技术积累，相关研究十分深入，但是，这类方法往往需要大量的监督数据。在某些特定领域或面向特定任务的知识图谱构建任务中，由于缺乏足够的标注数据，这些方法会受到严重限制。

大语言模型 (Large Language Model, LLM) 的出现使得大多数自然语言处理任务都可以在小样本甚至零样本的条件下执行。信息抽取任务也不例外，许多研究人员提出基于LLM的实体抽取和关系识别方法 (Dagdelen et al., 2024; Huang et al., 2023)。此外，现有的许多中文LLM都提供了开源API，用户甚至可以直接键入提示词并上传待抽取的文本以获取抽取结果。然而，在本文所设计的林业政策知识图谱本体中，存在许多超长的实体提及。例如，抽象概念实体“退耕还林还草补助资金兑付”的长度达到了12个字符，而行为实体“对林业资金收支的真实、合法情况依法进行稽查监督。主要内容包括资金计划的申请、下达，资金的拨付、使用、管理及其他有关情况”的长度达到了59个字符。此外，相比于常见的实体关系，本文所设计的知识图谱中实体之间的关系也更加抽象。例如，“义务”关系，“禁止”关系等。因此，如何有效利用开源LLM的API来从林业政策法规文本中抽取结构化信息，从而构建一个完备的林业政策知识图谱，仍然是一个需要深入研究的课题。然而，据我们所知，目前尚无相关工作可以

参考。

3 本体设计

本体设计是构建知识图谱的重要步骤，尤其对于领域知识图谱来说，本体的设计直接决定了整个知识图谱的质量和范围。本文首先确定林业政策法规本体所包含的实体类型，然后确定关系类型和属性。

首先，本文将通用知识图谱中最常见的实体类型借鉴过来，初步确定了3种实体类型：组织机构、人名、地理位置。然后，本文进一步确定了2种与政策领域相关的实体类型：政策文件和类别。许多已有的工作还额外定义了城市、省份、发文机构等实体类型(支凤稳et al., 2024; 郭佳航, 2021)，但本文认为这些都可以归类到地理位置、组织机构等通用实体类别中，因此不再重复定义。此外，由于林业政策法规文本中存在大量的领域概念、领域专属物品以及定义和解释，因此本文还设置了抽象概念、具象物品以及解释/定义这3个实体类型。最后，政策文本中往往存在大量与道义逻辑(Von Wright, 1951)相关的指令句，例如“各级林业主管部门要进一步严把入口关，提高标准站建设成效”。因此，本文根据道义逻辑的理论体系定义了行为和状态这2个实体类型。值得注意的是，本文根据道义逻辑理论体系所定义的不仅有2个实体类型，还包括3个关系类型，后文将对此详述。总的来说，本文一共定义了10种实体类型，附录A1中的表2展示了所定义的全部实体类型及其描述信息。

根据所定义的实体类型，以及林业政策法规领域特点，本文定义了15种关系类型：发布、位于、隶属、任职于、义务、禁止、有权、定义/解释、相关、范畴、引用、包含以及被发布、雇佣和被引用。其中，义务关系、禁止关系以及有权关系是根据道义逻辑理论体系所定义的(Von Wright, 1951)。附录A1中的表3展示了所定义的关系类型及其相关信息。

最后，本文根据林业政策法规的特点为政策文件(DOC)实体定义了7个属性，如附录A1的表4所示。图1为林业政策法规本体的可视化展示，可以看到，该本体的设计既保留了通用知识图谱的一些特性，也融入了林业政策法规的领域特点，更融合了道义逻辑相关概念，具有良好的表达性、可扩展性以及可复用性。

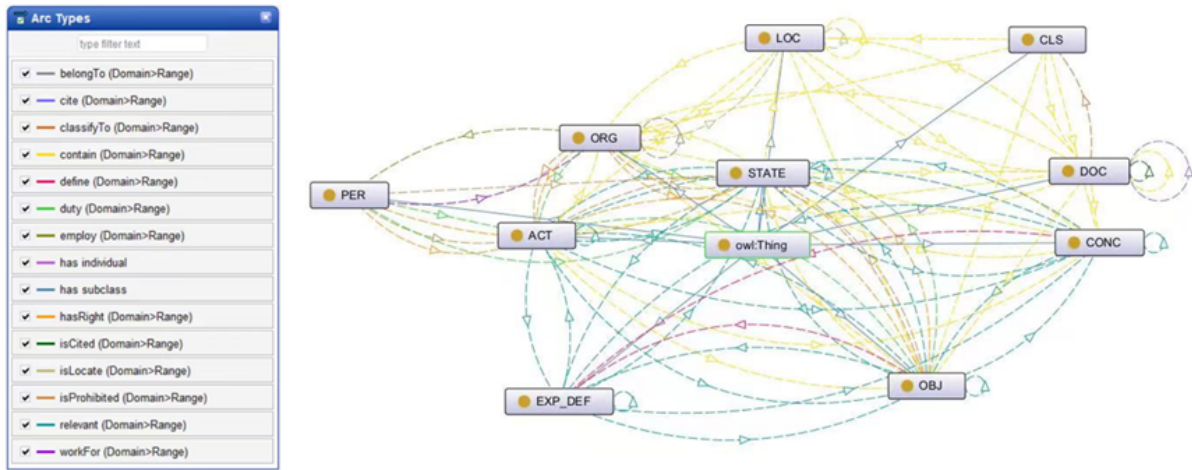


Figure 1: 本文所设计的林业政策法规本体的可视化

4 构建流程

由于有标签数据不足，现有的基于有监督信息抽取的知识图谱构建方法在特定领域上无法有效执行。大语言模型(Large Language Model, LLM)的出现有效缓解了这一现状。许多开源的中文LLM允许直接调用其API来实现零样本条件下的信息抽取，从而辅助构建领域知识图谱。然而，政策知识图谱存在以下几点特殊性：1) 一些实体提及很长，甚至可能占整个句子的一半；2) 一些关系类型较为抽象，涉及道义逻辑等理论。因此，常见的LLM调用方案并不完全适用本任务。为了解决这一问题，本文提出一种基于开源LLM的政策信息抽取流程，如图2所示。

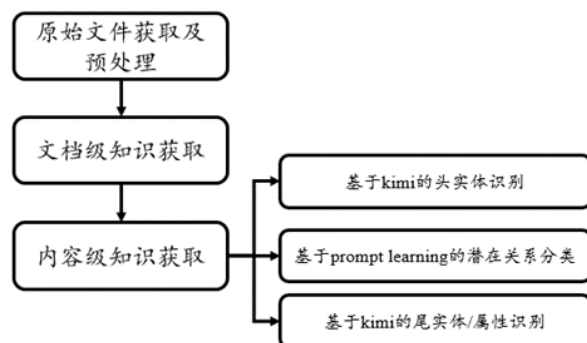


Figure 2: 基于开源LLM的林业政策知识图谱构建流程

其中，文档级知识获取过程旨在获取林业政策文件的发布机构、发布时间等元数据以及政策文件的所属类别、政策文件之间的相似性等，不涉及政策文件内部的知识。内容级知识获取过程则深入到文本内部，通过调用LLM的API来从文本中抽取实体及其关系。

4.1 原始文件获取及预处理

本文的林业政策法规原始文件全部来源于中国林业信息网¹，该网站由中国林业科学研究院林业科技信息所建设和管理，是国内林业行业中信息量最大、涵盖面最广的权威性行业网站。本文初步选取了国家林业局于1998年-2018年颁布的577项林业政策法规与国家林业和草原局于2018-2024年颁布的180项林业政策法规进行研究，总共757项文件，以TXT文件格式下载后进行整理，保留了其中的元数据。其中的国家林业和草原局于2018年3月成立，国家林业局在国务院机构改革后不再保留。本文使用关键词匹配的方法，将获取的757项文件按照18个类别划分，各类别的关键词以及每种类别的数量统计如附录A2中的表5所示。

4.2 文档级知识获取

文档级知识指的是不涉及到文本内容的知识，包括文档的发文机构、发布时间、实施时间、关键词、时效性等。这类知识的获取较为容易，在从网站获取政策文件时，可以直接从HTML结构中获取元数据。图3展示了官网所发布的政策文件及其元数据。

元数据			
名 称	中华人民共和国森林法	发布时间	2019年12月28日
文 号	中华人民共和国主席令第三十九号	来 源	新华网
发布机构	全国人民代表大会常务委员会	时效状态	现行有效
业务类型	森林资源管理		

中华人民共和国森林法	
(1984年9月20日第六届全国人民代表大会常务委员会第七次会议通过 根据1998年4月29日第九届全国人民代表大会常务委员会第二次会议《关于修改〈中华人民共和国森林法〉的决定》第一次修正 根据2009年8月27日第十一届全国人民代表大会常务委员会第十次会议《关于修改部分法律的决定》第二次修正 2019年12月28日第十三届全国人民代表大会常务委员会第十五次会议修订)	

政策法规正文

Figure 3: 政府官网所发布的政策文件正文及其元数据

除此之外，在本阶段还计算任意两个政策文件之间的相关性，从而在政策文件(DOC)实体之间建立相关(relevant)关系。给定政策文本集合 $D = d_1, d_2, \dots, d_N$ ，本文使用预训练语言模型(Zaheer et al., 2020)的中文检查点将每篇政策文本编码为语义向量，得到 $\mathbf{V} = \vec{v}_1, \vec{v}_2, \dots, \vec{v}_N$ 。然后，计算每两个向量之间的余弦相似度，当政策文件 d_i 和 d_j 之间的相似度值大于设

¹<http://www.lknet.ac.cn/>

定的阈值 λ 时，则在政策文件实体 d_i 和 d_j 之间建立一条类别为相关(relevant)的可自反关系边 $\langle d_i, r_{i,j}^{rel}, d_j \rangle$ 。之所以使用BigBird来表征整个政策文本，是因为其支持长达8000的文本输入长度。

4.3 内容级知识获取

从林业政策文本中抽取内容级知识需要模型能够深刻理解文本的语义。中文LLM是在海量中文语料上预训练的大型语言模型，因此具有良好的语义理解能力。常见的中文LLM包括ChatGLM(Du et al., 2022)、LLaMa-Chinese(Touvron et al., 2023)、BaiChuan(Yang et al., 2023)以及kimi²等。本文从模型表现、使用便捷性、使用成本等方面综合考量了上述模型，最终确定使用kimi。本文发现仅在一轮对话中一次性地询问kimi，让其从整个政策文本中直接抽取出全部<实体-关系-实体>三元组，效果极差。因此，本文设计了分三步的抽取流程：第一步是基于kimi的头实体识别；第二步是基于Prompt learning的潜在关系分类；第三步是基于kimi的尾实体识别。下面分别详细介绍每一步。

4.3.1 基于kimi的头实体识别

根据本文所设计的林业政策本体，政策文本中存在许多超长的实体提及，例如“由地方政府负责人担任林长，负责森林资源的保护和管理（对林长制的解释）”。Kimi对这类实体的抽取较为困难。通过观察，本文发现这类超长的实体提及几乎都是尾实体，因此决定转而先让kimi识别短小而简单的头实体。此外，为了提供给kimi更多的上下文信息，从而提高其识别能力，本文将整篇文本作为输入单位而非单个句子。最终，本文设计如下提示模板，来驱动kimi执行政策文本上的头实体识别任务：

```
Messages=[{"role": "system",
  "content": "你是一个实体抽取工具，从文本中抽取以下实体：组织机构、人名、地理位置、抽象概念、具象物品"。
    抽象术语是指林业领域的术语、理论和方法，例如科学绿化和木材安全。具象物品是指与林业相关的具体工具、
    物品等，例如森林防火设施和生态修复材料。},
  {"role": "user",
    "content": d}]
```

Figure 4: kimi执行头实体识别任务的提示模板

其中， d 为待抽取的林业政策文本。实践中发现，即使不解释人名和地理位置实体的含义，模型也能取得良好的抽取效果，而抽象术语和具象物品这两个实体需要向模型解释其含义并演示一些示例，模型才能取得令人满意的抽取效果。

4.3.2 基于Prompt learning的潜在关系分类

该步骤旨在根据给定的文本及其中所包含的头实体，识别出文本潜在的关系类型。本文反复测试了多种提示模板，均无法有效驱动kimi模型从给定的林业政策文本中识别出预定义的关系类型。考虑到该过程本质上可以视为一个文本分类任务，因此，本文转而使用Prompt learning来训练一个本地的文本分类模型。之所以采用Prompt learning技术，是因为其能够在小样本场景下取得良好的模型微调效果(Schick and Schütze, 2021)。众所周知，获取特定领域上的有标签数据需要花费昂贵的人力和时间成本。本文为每个关系类别精心构造了10条训练样本，总共得到150条样本。随后，使用这150条训练数据对DeBERTa模型进行Prompt learning微调，详细过程如算法1所示。

其中， s_i 表示待识别的文本， h_i 是其中所包含的头实体， y_i 表示金标关系类型。标签词(Verbalizer)设置为关系类别自身的名称文本。预定义的模板为： s_i 中的头实体为 h_i ，其中潜在的关系类别为[MASK]。候选关系类别包括：发布，位于，隶属，任职于，义务，禁止，有权，定义/解释，相关，范畴，引用，包含，被发布，雇佣，被引用。

训练好的模型记作 \mathcal{M} ，给定待识别潜在关系的文本 s 以及其中所包含的头实体 h ，训练好的模型 \mathcal{M} 能够预测 s 中潜在的关系类别，如公式(1)所示。

$$\hat{y} = \mathcal{M}(s, h) \quad (1)$$

²<https://kimi.moonshot.cn/>

算法 1: 基于Prompt learning的关系分类模型微调

输入: 训练集 $Train = \{(s_1, h_1, y_1), (s_2, h_2, y_2), \dots, (s_{150}, h_{150}, y_{150})\}$, 初始模型参数 θ_{init}
输出: 调优的模型参数 θ_{opt}

- 1 初始化提示模板: $T(\cdot)$;
- 2 初始化标签到标签词的映射: $\phi: \mathcal{Y} \rightarrow \mathcal{V}$;
- 3 初始化损失: $\mathcal{L} = 0$;
- 4 **for** $epoch = 1 \rightarrow epochs$ **do**
- 5 **for each** (s_i, h_i, y_i) **in** $Train$ **do**
- 6 $\ell = \log P([MASK] = \phi(y_i) | T(s_i, h_i))$ //计算掩码损失;
- 7 $\mathcal{L} += \ell$;
- 8 **end**
- 9 更新参数: $\theta_{init} = \theta_{init} - \alpha \cdot \frac{\partial \mathcal{L}}{\partial \theta_{init}}$;
- 10 **end**
- 11 $\theta_{opt} \leftarrow \theta_{init}$;
- 12 **return** θ_{opt} ;

4.3.3 尾实体识别

通过前述过程, 已经可以获得给定文本所包含的头实体及其所蕴含的关系类型, 接下来对尾实体的获取则较为容易。通过观察, 本文发现大部分尾实体都紧邻关系词的后面。例如“灌木一般系指高3米以下, 没有明显主干、呈丛生状态的木本植物”中, 关系词是“一般系指”, 而尾实体是“高3米以下, 没有明显主干、呈丛生状态的木本植物”。因此, 我们首先让kimi识别关系词 w_r 。给定头实体和关系类型, 我们通过以下提示词来驱动kimi的API执行关系词识别:

```
Messages=[{"role": "system",
            "content": "我将给你一段文本以及其中所包含的头实体和关系类型, 请你帮我识别出其中的关系词。"},
           {"role": "user",
            "content": "文本: d; 头实体: h; 关系类型: y"}]
```

Figure 5: kimi执行关系词识别的提示词

其中, d 表示待识别的文本, h 表示其中的头实体, y 表示其潜在关系类型。有了关系词 w_r , 接下来, 直接匹配 w_r 后面的文本作为尾实体 t 。

5 实验与分析

本节对所提出的林业政策知识图谱构建框架的可行性和先进性进行分析, 具体包括: 对本体和信息抽取方法进行分析。此外, 我们还通过基于大模型的RAG来分析该知识图谱的实践应用效果。

5.1 本体分析

5.1.1 实验1: 知识图谱查询准确性实验

本研究旨在验证知识图谱在响应用户查询方面的准确性。为此, 我们邀请了林业领域的专家志愿者, 在ChatGPT的辅助下精心设计了1000条高质量的Cypher查询语句及其相应的预期查询结果。查询语句的详细构建流程参见附录A3。图6展示了这些查询的样本示例。我们运用Neo4j的推理引擎, 在林业政策知识图谱上执行查询。实验结果显示, 知识图谱成功返回预期答案的准确度高达82.5%。这一结果表明, 我们构建的林业政策知识图谱能够有效地覆盖并响应用户提出的大多数查询需求。此外, 附录A4还通过图12展示在Neo4j数据库上的具体查询案例。

5.1.2 实验2: 知识图谱扩展性实验

我们将林业政策知识图谱分别与五个通用知识图谱进行融合, 检验其融合后的查询响应能力, 从而反应所设计本体的灵活性和可扩展性。图7展示了将林业政策知识图谱分别

```

MATCH (d:DOC {name: "中华人民共和国主要草种目录"})
MATCH (d)-[:isPublished]->(o:ORG)
RETURN o.name AS org_name

MATCH (o:ORG {name: "林业检测站"})-[:duty]->(a:ACT)
RETURN a.name AS act_name

MATCH (c:CONC {name: "国家储备林"})-[:define]->(e:EXP_DEF)
RETURN e.name AS exp_def_name

```

Figure 6: 林业领域专家志愿者设计的Cypher查询语句示例

与OwnThink、OpenKG、CN-Dbpedia、DBpedia以及OpenConcepts融合后，响应用户查询的准确率提升情况。从实验结果中可以看出，与五个通用知识图谱融合后，其响应用户查询的准确率均明显提升。这说明，我们设计的本体可以充分兼容通用领域或其他特定领域知识图谱的本体，从而进一步提升在不同真实场景中的应用效果。

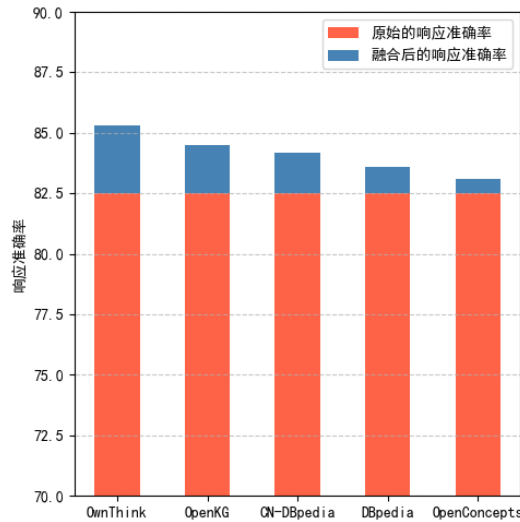


Figure 7: 将林业政策知识图谱分别与五个通用知识图谱进行融合后的响应准确率提升情况

5.2 信息抽取分析

5.2.1 整体分析

本节整体地分析所提出的基于开源大模型的政策信息抽取方法的有效性和先进性。为此，我们与以下基线进行对比：1) DeepKE (Zhang et al., 2022)是由浙江大学开发和维护的开源知识图谱抽取框架，支持低资源和篇章级的信息抽取；2) LTP (Che et al., 2021)是由哈尔滨工业大学开发的中文处理基础平台，提供命名实体识别和关系分类功能；3) Jiagu³是由思知机器人公司开发的针对中文的实体及关系抽取开源工具；4) DeepDive (Niu and Ré, 2012)是由斯坦福大学InfoLab实验室开发的开源知识抽取系统，OPENKG平台修改了DeepDive以使其支持中文。这些基线都是开箱即用的无监督信息抽取框架，代表了无监督信息抽取技术的通用水准。此外，鉴于我们的方法是建立在LLM强大的语言推理能力基础上的，因此我们也将直接使用大模型执行政策信息抽取作为对比基线，包括：1) 直接调用Kimi的API执行篇章级或段落级的政策信息抽取；2) 直接使用LLaMa2-Chinese执行篇章级或段落级的政策信息抽取。

我们评估上述基线以及我们提出的方法在50篇林业政策文本上三元组抽取效果。具体来说，首先人工标注出50篇林业政策文本中所包含的三元组，共标注出了1126个三元组。然后，计算方法在这50篇文本上的精确率和召回率，最终的实验结果如表1所示。

³<https://github.com/ownthink/Jiagu>

从表1中可以看出，现有的开箱即用的信息抽取框架在政策信息抽取任务上效果较差。主要原因有两点：第一，政策信息中包含大量具有超长名称的实体，例如状态实体、行为实体、解释实体等，这类实体名称有时能占整个句子的一半，因此给常规的信息抽取框架带来了挑战；第二，政策信息中的许多实体和关系与通用领域的实体及其关系在语义分布上也有着较大的差距，因此非定制化的通用的信息抽取框架很难对齐抽取目标。

从表1中我们还可以看出，直接调用LLM的API无法取得与我们方法相当的抽取结果。这说明我们的方法在利用LLM强大的语义推理能力的同时，更适配政策信息抽取的特点，对政策概念、术语以及超长的实体及其之间的特殊关系有着良好的。此外，尽管我们方法的效果远超其他基线，但仍尚未达到80%，这意味着政策信息抽取任务的难度较大，未来仍然需要深入探索。

方法	精确率	召回率
开箱即用的无监督信息抽取框架		
DeepKE	22.3%	5.2%
LTP	16.5%	2.2%
Jiagu	32.9%	3.6%
DeepDive	19.1%	1.9%
直接调用LLM的API		
Kimi-段落级抽取	43.2%	17.9%
Kimi-文档级抽取	50.7%	23.3%
LLaMa2-Chinese-段落级抽取	38.6%	12.8%
LLaMa2-Chinese-文档级抽取	30.7%	9.8%
我们的方法	76.2%	62.6%

Table 1: 三元组抽取的精确率和召回率

5.2.2 细节分析

本节逐一评估我们所提出方法中关键步骤的有效性，具体包括：基于Kimi的实体识别；基于Prompt learning的潜在关系分类。为了更直观地展示结果，我们利用雷达图来描绘这些步骤在处理不同实体类型和关系类型时的抽取效果。

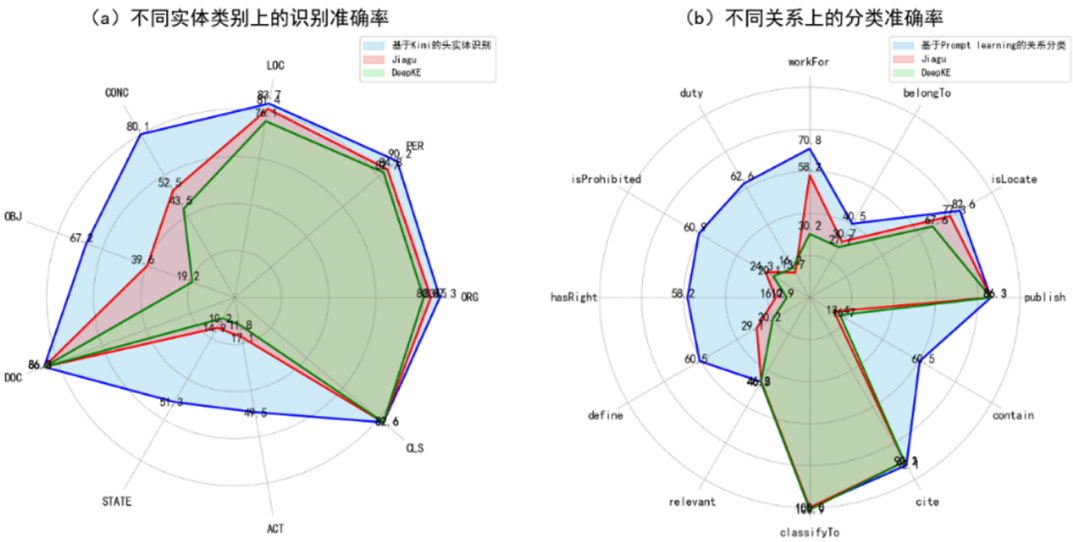


Figure 8: 在不同实体类型和不同关系类型上的准确率

图8的左侧呈现了基于Kimi的实体识别在各类实体类型上的识别准确率（蓝色区域）。观察结果可以发现，“行为（ACT）”和“状态（STATE）”实体的识别准确度相对较低，而“解释/定义（EXP_DEF）”实体的抽取效果也不尽人意。这三类实体都属于超长实体类型，这表明

在政策信息抽取任务中，对超长实体的识别依然是一个主要挑战。与此同时，我们也注意到，与通用信息抽取一致的组织机构、地理位置以及人名仍然是识别效果最好的实体类型。

图8的右侧展示了基于Prompt learning的潜在关系分类在不同关系类型上的分类准确率（蓝色区域）。从图中可以明显看出，“相关（relevant）”这一关系类型的识别准确度最低。我们推测，这可能是由于该关系类型的含义较为模糊，且与其他关系类型的界限不甚清晰，从而导致了对其效果评估的歧义性。此外，“隶属于（belongTo）”关系类型的识别效果也不理想，我们认为这可能与该关系类型在林业政策文本中的出现频率较低有关，这导致了准确率难以反映真实情况。

雷达图中也显示了额外两个基线效果（红色和绿色区域）。综合分析来看，我们方法中的关键步骤虽然仍有不足之处，但整体上已经在政策信息抽取任务上表现出了令人满意的效果，优于其他基线。因此，我们认为，我们所提出的基于LLM的政策信息抽取方法是可行的且先进的。

5.2.3 训练数据依赖性分析

我们致力于实现政策信息抽取流程的自动化，以最大限度地减少人工干预。然而，在基于Prompt learning进行关系分类的环节中，我们仍需对部分数据进行手动标注。如4.3.2节所述，我们为每个关系类别精心挑选并设计了10个训练实例，总计150个训练实例。本节旨在探究在基于Prompt learning的关系分类过程中，是否能够减少训练实例的数量而不影响分类效果。实验结果如图9所示，其揭示了一个明显的趋势：随着训练实例数量的减少，Prompt learning在关系分类任务上的表现显著降低。这说明对于每个关系类别而言，少于10个训练实例是不足以保证分类质量的。此外，通过与当前主流的关系分类技术进行对比，我们发现基于Prompt learning的关系分类方法在数据稀疏场景下拥有明显的优势。

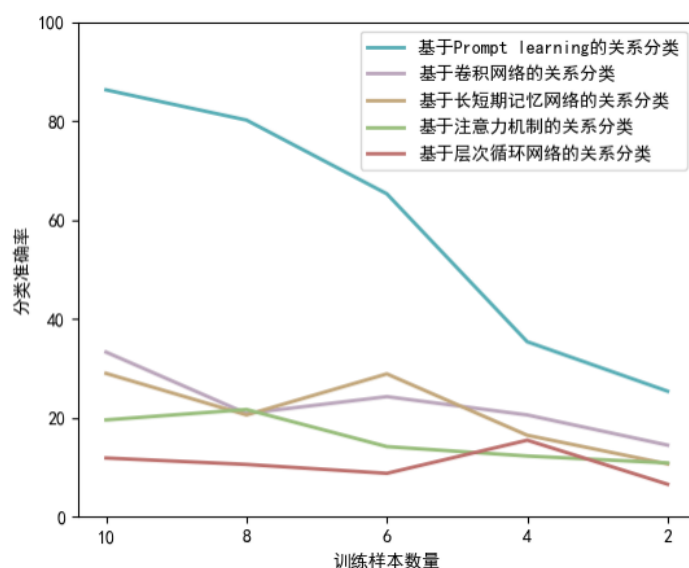


Figure 9: 训练样本依赖性分析

5.3 应用实践分析

为了进一步分析所构建的林业政策知识图谱的实际应用效果，我们将该知识图谱作为大模型的RAG知识源。我们使用LangChain框架构建了一个基于LLaMa-Chinese的RAG系统，并使用GPT-4评估接入林业政策知识图谱后大模型的文本生成效果是否有所提升。评估标准涵盖文本的正确性、有效性与流畅性三个维度，评分范围为1至10分。我们精心设计了一套引导GPT-4执行评估任务的模板，旨在确保评估过程的准确性和一致性，模板如图10所示。

图11清晰地展示了在引入林业政策知识图谱后，LLaMa-Chinese模型在所有三个评估维度上的得分均有显著提升。为了更直观地呈现这种提升，附录A5中的表6提供了一个具体的对比案例分析，进一步印证了知识图谱在提升文本生成质量方面的显著优势。

请你作为一个公正的判别器，评估模型对以下用户问题的回答质量。你的评价应该考虑正确性、有效性、流畅性三个维度。尝试发现并纠正模型回答中的任何错误，且评价要尽可能客观。提供解释性说明后，你必须严格遵循以下格式，在1到10的范围内对响应进行评分，例如[分数=5]。

Figure 10: GPT-4执行评估任务的模板

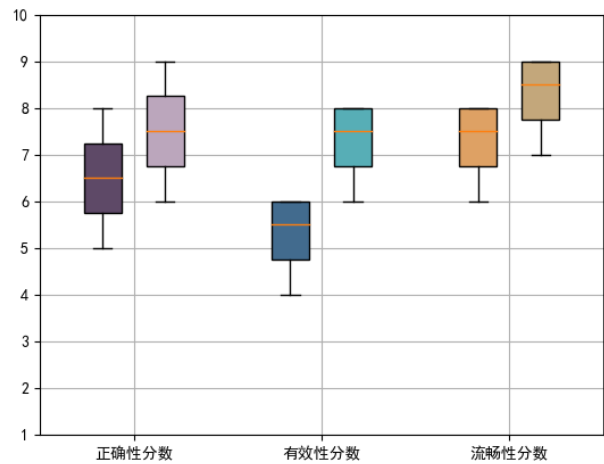


Figure 11: LLaMa-Chinese模型在集成林业政策知识图谱前后的答案生成质量评估结果

6 构建结果展示

我们使用Neo4j数据库来存储所构建的林业政策知识图谱实体及关系，其可视化展示参见附录A6中的图13。未来，我们将会进一步扩充该知识图谱的规模，并细化其本体、剔除噪声三元组等，然后对其进行封装，提供在线API，以供相关领域的研究人员使用。

7 总结与展望

本文定义了一个细粒度的林业政策知识图谱本体。相比于已有的工作，本文所定义的本体深入到了政策内容上，具有更细粒度的表达能力。此外，本文提出了一个基于开源大模型的分三阶段的政策知识抽取流程，并通过该流程构建了一个较为完备的林业政策知识图谱。最终的构建结果展示及分析表明，所提出的政策知识抽取流程是有效的，所构建的林业政策知识图谱具有良好的林业政策知识表达及推理能力。

本文仅在林业政策知识图谱构建任务上做了初步的探索。未来，可以进一步对本体进行细化，以提升对林业政策知识的表达能力。此外，政策知识抽取过程仍然存在一些问题，例如无法抽取跨句子甚至跨段落的实体关系、无法抽取重叠实体关系。未来，将提出更先进的政策知识抽取方法。最后，未来将扩大该知识图谱的规模，并提出一系列基于该知识图谱的林业工程合规、林业智能问答、林业政策分析等方法论或设计实际可用的系统。

参考文献

- Abu-Salih B. 2021. Domain-specific knowledge graphs: A survey. *Journal of Network and Computer Applications*, 185:103076.
- Abu-Salih B., Al-Qurishi M., Alweshah M., Al-Smadi M., Alfayez R., and Saadeh H. 2023. Healthcare knowledge graph construction: A systematic review of the state-of-the-art, open issues, and opportunities. *Journal of Big Data*, 10(1):81.

- Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R., and Ives Z. 2007. Dbpedia: A nucleus for a web of open data. *International Semantic Web Conference*, 722–735.
- Che W., Feng Y., Qin L., and Liu T. 2021. N-LTP: An Open-source Neural Language Technology Platform for Chinese. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 42–49.
- Chen P., Lu Y., Zheng V. W., Chen X., and Li X. 2018. An automatic knowledge graph construction system for K-12 education. *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, 1–4.
- Chen X., Jia S., and Xiang Y. 2020. A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, 141:112948.
- Chen Y., Sun Y., Yang Z., and Lin H. 2020. Joint entity and relation extraction for legal documents with legal feature enhancement. *Proceedings of the 28th International Conference on Computational Linguistics*, 1561–1571.
- Dagdelen J., Dunn A., Lee S., Walker N., Rosen A. S., Ceder G., Persson K. A., and Jain A. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.
- Du Z., Qian Y., Liu X., Ding M., Qiu J., Yang Z., and Tang J. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 320–335.
- Fang Y., Zhang Q., Zhang N., Chen Z., Zhuang X., Shao X., Fan X., and Chen H. 2023. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, 5(5):542–553.
- Fensel D., Şimşek U., Angele K., Huaman E., Kärle E., Panasiuk O., Toma I., Umbrich J., Wahler A., Fensel D., et al. 2020. Introduction: What is a knowledge graph? *Knowledge Graphs: Methodology, Tools and Selected Use Cases*, 1–10.
- Gan L., Ye B., Huang Z., Xu Y., Chen Q., and Shu Y. 2023. Knowledge graph construction based on ship collision accident reports to improve maritime traffic safety. *Ocean & Coastal Management*, 240:106660.
- Huang A. H., Wang H., and Yang Y. 2023. FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.
- Li A., Han C., Xing X., Wei Q., Chi Y., and Pu F. 2024. KGSCS-a smart care system for elderly with geriatric chronic diseases: A knowledge graph approach. *BMC Medical Informatics and Decision Making*, 24(1):73.
- Li L., Wang P., Yan J., Wang Y., Li S., Jiang J., Sun Z., Tang B., Chang T.-H., Wang S., et al. 2020. Real-world data medical knowledge graph: Construction and applications. *Artificial Intelligence in Medicine*, 103:101817.
- Liu P., Guo Y., Wang F., and Li G. 2022. Chinese named entity recognition: The state of the art. *Neurocomputing*, 473:37–53.
- Niu F., and Ré C. 2012. *DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference*.
- Ouyang K., Liu Y., Li S., Bao R., Harimoto K., and Sun X. 2024. Modal-adaptive Knowledge-enhanced Graph-based Financial Prediction from Monetary Policy Conference Calls with LLM. *arXiv Preprint arXiv:2403.16055*.
- Peng C., Xia F., Naseriparsa M., and Osborne F. 2023. Knowledge graphs: Opportunities and challenges. *Artificial Intelligence Review*, 56(11):13071–13102.
- Schick T., and Schütze H. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 255–269.
- Suchanek F. M., Kasneci G., and Weikum G. 2007. Yago: A core of semantic knowledge. *Proceedings of the 16th International Conference on World Wide Web*, 697–706.

- Tang X., Feng Z., Xiao Y., Wang M., Ye T., Zhou Y., Meng J., Zhang B., and Zhang D. 2023. Construction and application of an ontology-based domain-specific knowledge graph for petroleum exploration and development. *Geoscience Frontiers*, 14(5):101426.
- Touvron H., Martin L., Stone K., Albert P., Almahairi A., Babaei Y., Bashlykov N., Batra S., Bhargava P., Bhosale S., et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv Preprint arXiv:2307.09288*.
- Von Wright G H. 1951. Deontic logic. *Mind*, 60(237):1–15.
- Vrandečić D., and Krötzsch M. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.
- Wang X., Wang C., Li L., Li Z., Chen B., Jin L., Huang J., Xiao Y., and Gao M. 2023. Fashionklip: Enhancing e-commerce image-text retrieval with fashion multi-modal conceptual knowledge graph. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, 149–158.
- Wang Z., Zhang J., Feng J., and Chen Z. 2014. Knowledge graph embedding by translating on hyperplanes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1).
- Xu B., Xu Y., Liang J., Xie C., Liang B., Cui W., and Xiao Y. 2017. CN-DBpedia: A never-ending Chinese knowledge extraction system. *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 428–438.
- Yang A., Xiao B., Wang B., Zhang B., Bian C., Yin C., Lv C., Pan D., Wang D., Yan D., et al. 2023. Baichuan 2: Open large-scale language models. *arXiv Preprint arXiv:2309.10305*.
- Zaheer M., Guruganesh G., Dubey K. A., Ainslie J., Alberti C., Ontanon S., Pham P., Ravula A., Wang Q., Yang L., et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.
- Zeng D., Liu K., Lai S., Zhou G., and Zhao J. 2014. Relation classification via convolutional deep neural network. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2335–2344.
- Zhang N., Xu X., Tao L., Yu H., Ye H., Qiao S., Xie X., Chen X., Li Z., and Li L. 2022. DeepKE: A Deep Learning Based Knowledge Extraction Toolkit for Knowledge Base Population. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 98–108.
- Zhong L., Wu J., Li Q., Peng H., and Wu X. 2023. A comprehensive survey on automatic knowledge graph construction. *ACM Computing Surveys*, 56(4):1–62.
- Zirui C., Xin W., Lin W., Dawei X., and Yongzhe J. 2021. Survey of Open-Domain Knowledge Graph Question Answering. *Journal of Frontiers of Computer Science & Technology*, 15(10).
- 叶宏彬, 张宁豫, 陈华钧, 邓淑敏, 毕祯, 陈想. 2023. OpenConcepts: 一个开放的细粒度中文概念知识图谱. *中文信息学报*, 37:46–53.
- 支凤稳, 孟佳琪, 郑彦宁. 2024. 科技报告政策细粒度知识图谱构建及应用研究. *情报理论与实践*, 1–15.
- 智月平. 2024. 基于深度学习的金融证券知识图谱的构建与应用. 硕士学位论文. 浙江科技大学.
- 赵颜利, 李连军, 余红梅, 史玉彬. 2020. 面向战例知识图谱的本体构建. *海军航空工程学院学报*, 35:407–413.
- 郭佳航. 2021. 基于深度学习的科技政策知识图谱构建及应用研究. 硕士学位论文. 北京邮电大学.
- 陈静, 张文泰, 陈清财, 户保田, 冯铭. 2023. 面向智能诊疗的疾病文本知识表示体系. *中文信息学报*, 37:67–76.

附录

A1 所定义的林业政策领域本体

表2展示了所定义的林业政策法规本体中的实体类型。共包含10种实体类型：组织机构、人名、地理位置、政策文件、类别、抽象概念、具象物品、解释/定义、行为、状态。表2中第一列展示实体类型的名称，第二列为标识符，第三列为对该实体类型的描述信息，第四列展示了该实体类型的一些示例。

表3展示了所定义的林业政策法规本体中的关系类型，包括：发布、位于、隶属、任职于、义务、禁止、有权、定义/解释、相关、范畴、引用、包含。表3中第一列为关系类型，第二列为标识符，第三列为该关系的定义域，第四列为该关系的值域，第五列为一些示例。此外，表中最后一列展示关系之间的互逆性。

表4展示了为政策文件实体所定义的属性，包括7个属性类型：ID、关键词、发布时间、逝世日期、时效性、效力位阶、来源网址。

实体类型	标识符	描述	示例
组织机构	ORG	与林业相关的公司、科研机构以及政府机构、事业单位等	林业观测站、黑龙江省林业和草原局
人名	PER	林业领域的个人，包括政府公务员，科学家、政策制定者、环保活动家、林业企业家等	侯绪珉（黑龙江省林业和草原局局长）
地理位置	LOC	通用地理位置或与林业相关的地理实体	黑龙江省、内蒙古敕勒川自然保护区
政策文件	DOC	用文件标题来代表政策文件	《国家林业和草原局公告（2023年第15号）（发布41项行业标准）》
类别	CLS	所涉及的政策类别	湿地保护、野生动物保护
抽象概念	CONC	林业领域的术语、理论和方法	科学绿化、木材安全
具象物品	OBJ	与林业相关的具体的工具、物品等	森林防火设施，生态修复材料
解释/定义	EXP_DEF	对林业概念的解释或定义	由地方政府负责人担任林长，负责森林资源的保护和管理（对林长制的解释）
行为	ACT	有行动能力的主体可以执行的具体行为	监督林业安全生产过程，巡查林业防火预警系统
状态	STATE	无行动能力的主体所呈现的状态	原有样貌保持不变（指古建筑硬保持原有样貌），有涂改痕迹（指林业相关材料不得有涂改痕迹）

Table 2: 本文定义的林业政策法规本体中的实体类型

关系类型	标识符	定义域	值域	示例	互逆关系
发布	publish	ORG	DOC	<国家草原和林业局, 发布, 《关于贯彻实施新修订森林法的通知》>	被发布 (isPublished)
位于	isLocate	ORG/LOC	LOC	<国家草原和林业局, 位于, 北京市>、<鸡西市前进林场, 位于, 鸡西市>	包含 (contain)
隶属	belongTo	ORG	ORG	<国家野生动物保护监测中心, 隶属于, 国家林业和草原局>	包含 (contain)
任职	workFor	PER	ORG	<侯绪珉, 任职于, 黑龙江省林业和草原局>	雇佣 (employ)
义务	duty	PER/ORG/ OBJ	ACT/STATE	<侯绪珉, 义务(应当), 在4月1日之前到任工作>、<各检测站, 义务(负责), 检测森林火情>、<古建筑, 义务(要), 保持原有样貌>	/
禁止	isProhibited	PER/ORG/ OBJ	ACT/ STATE	<社会团体, 禁止(不得), 未经请示进入林场>、<相关材料, 禁止(不得), 有涂改痕迹>	/
有权	hasRight	PER/ORG/ OBJ	ACT/ STATE	<下属林场, 有权(可以), 在周边街道设置警戒线>、<森林防火设置, 有权(可以), 放置在公共区域>	/
定义/解释	define	CONC/OBJ	EXP_DEF	<退耕还林还草, 解释, 旨在将部分耕地恢复为林地或草地以改善生态环境>	/
相关	relevant	CONC/OBJ/ EXP_DEF/ ACT/ STATE	CONC/OBJ/ EXP_DEF/ ACT/ STATE	<监督林业安全生产过程, 相关, 林业安全生产>	该关系具有自反性
范畴	classifyTo	DOC	CLS	<《国家林业和草原局公告(2023年第15号)》(发布41项行业标准)》, 范畴, 行业标准>	包含 (contain)
引用	cite	DOC	DOC	<《国家林业和草原局公告(2023年第15号)》(发布41项行业标准)》, 引用, 《土地退化类型与分级规范》>	被引用 (isCited)
包含	contain	DOC/LOC/ ORG/STATE/ ACT/CLS	DOC/LOC/ ORG/CONC/ OBJ	<行业标准, 包含, 《国家林业和草原局公告(2023年第15号)》(发布41项行业标准)》>	/

Table 3: 本文定义的林业政策法规本体中的关系类型

属性	描述
ID	根据政策文件的获取顺序为其自动分配的10位数ID
关键词	官网上发布政策文件大多都附带一个关键词
发布时间	政策文件在官网上发布的准确时间
实施日期	政策文件开始生效的日期
时效性	包含“现行有效”和“失效”两个候选值
效力位阶	包含“行政法规”、“部门规章”、“公告”等候选值
来源网址	该政策文件的获取网址

Table 4: 本文为政策文件(DOC)实体所定义的属性

A2 原始文件的类别划分信息

表5展示了原始文件的类别划分。此外，还展示了划分类别所采用的关键词以及每个类别上的文件数量。

类别	类别名称	所用关键词	文件数量
1	绿化条例	“绿化”，“植被”，“城市美化”，“公园”，“花坛”，“草坪”，“行道树”，“景观设计”，“绿色空间”，“生态规划”	247
2	环境保护	“污染控制”，“生态平衡”，“环境治理”，“减排”，“污水处理”，“废气治理”，“垃圾分类”，“环境监测”，“可持续发展”，“清洁能源”	108
3	森林防火	“森林火灾”，“火源管理”，“防火隔离”，“火险预警”，“森林警察”，“应急响应”，“火情监测”，“森林植被”，“火灾预防”	92
4	自然保护区条例	“自然保护区”，“生物多样性”，“生态保护区”，“野生动植物”，“保护区管理”，“生态旅游”，“物种保护”，“生态系统”，“环境教育”	244
5	风景区保护条例	“风景名胜”，“文化遗产”，“旅游景区”，“自然景观”，“历史遗迹”，“旅游规划”，“景区管理”，“文化保护”，“环境保护”	78
6	水土保持	“土壤侵蚀”，“水土流失”，“河流治理”，“水土保持法”，“土地整治”，“水资源管理”，“生态修复”，“水土保持技术”，“泥石流防治”	56
7	森林资源保护	“森林覆盖”，“林木资源”，“森林生态”，“森林经营”，“森林保护区”，“森林生态系统”，“森林旅游”，“森林资源利用”，“森林保护法”	168
8	湿地保护	“湿地生态系统”，“湿地恢复”，“水生植物”，“湿地保护区”，“湿地管理”，“湿地生态系统服务”，“湿地生物多样性”，“湿地资源利用”，“湿地保护法”	22
9	城市公园保护条例	“城市绿地”，“公共空间”，“休闲设施”，“城市公园”，“城市绿化”，“城市生态”，“市民休闲”，“城市景观”，“城市环境美化”	14
10	野生动物保护	“野生动物”，“物种保护”，“生态平衡”，“野生动物保护区”，“野生动物管理”，“野生动物救助”，“生物多样性”，“野生动物研究”，“野生动物法律”	162
11	野生植物保护	“植物多样性”，“濒危植物”，“植物保护区”，“野生植物资源”，“植物保护法”，“植物生态”，“药用植物”，“野生植物研究”，“植物资源利用”	16

12	林木种子 条例	“种子质量”，“林木繁殖”，“种子库”，“种子生产”，“种子检验”，“种子加工”，“种子市场”，“种子保护”，“种子技术”，“种子法规”	61
13	林木(公益林、防护林等)	“公益林”，“防护林”，“森林功能”，“森林生态系统”，“森林资源”，“森林经营”，“森林保护”，“森林旅游”，“森林管理”，“森林法规”	277
14	生态建设	“生态系统”，“绿色发展”，“生态修复”，“生态工程”，“生态保护”，“生态文明”，“生态经济”，“生态环境”，“生态规划”，“生态监测”	257
15	草原保护	“草原生态”，“草场恢复”，“草原管理”，“草原保护区”，“草原生态旅游”，“草原生态系统”，“草原生物多样性”，“草原资源利用”，“草原保护法”，“草原生态监测”	11
16	名木古树 保护	“古树名木”，“历史树木”，“树木保护”，“名木普查”，“古树保护法”，“树木养护”，“树木研究”，“树木文化”，“树木生态”，“树木旅游”	30
17	森林公园 保护	“森林公园”，“生态旅游”，“森林资源”，“森林保护区”，“森林生态系统”，“森林游憩”，“森林景观”，“森林保护法”，“森林管理”，“森林环境”	259
18	其他	无关键词	224
共757			

Table 5: 本文原始数据的类别划分

A3 查询语句的构建过程

为了构建查询语句及其对应的预期答案，我们首先招募三名初级志愿者（林业相关专业的本科生）分别撰写80条自然问句。然后，我们使用ChatGPT将总共240条自然问句转换为Cypher查询语句。接下来，我们邀请两名计算机领专家从中选出质量最好的200条Cypher查询语句。

A4 在Neo4j数据库上的具体查询案例

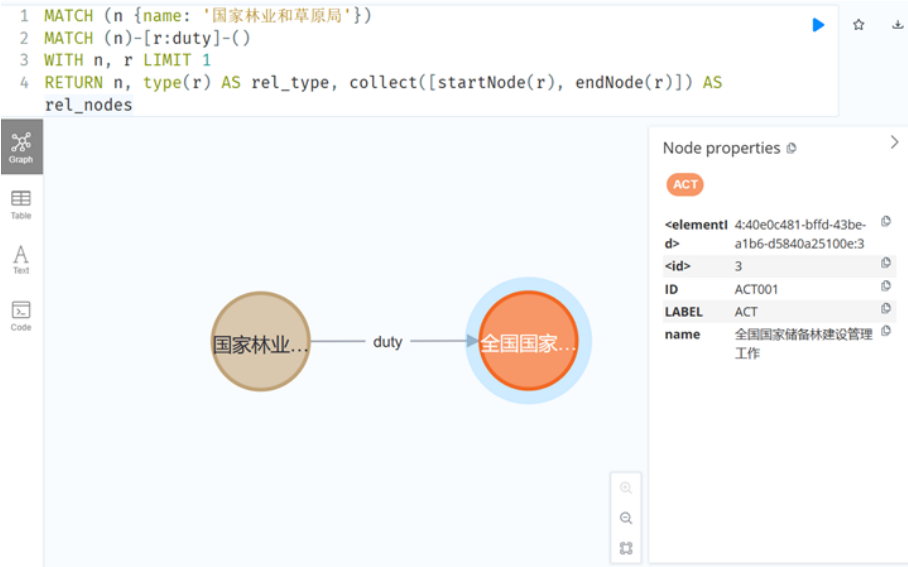


Figure 12: 在所构建的知识图谱上查询“国家林业和草原局的职责”

在Neo4j数据库上的具体查询案例。可以看出，当查询“国家林业和草原局”实体的责任时，显示其负责“全国国家储备林建设工作”。

A5 知识图谱对大模型回答的增强案例分析

问句	什么是林业生态环境建设
LLaMa-Chinese生成的回答	<p>林业生态建设是指通过科学规划和管理，利用林业技术和措施，保护、恢复和改善森林生态系统，增强其生态功能，实现生态环境的可持续发展的过程。其主要内容包括：</p> <ol style="list-style-type: none">1) 森林保护：保护现有森林资源，防止乱砍滥伐和森林火灾，维护森林生态系统的稳定性和多样性。2) 植树造林：通过人工造林和天然更新，扩大森林面积，提升森林覆盖率，改善生态环境。3) 生态修复：恢复被破坏的森林生态系统，如退耕还林、封山育林等，恢复生态功能和生物多样性。4) 水土保持：通过植被恢复、坡地治理等措施，防止水土流失，保护水资源，改善水环境。5) 生物多样性保护：保护珍稀濒危植物和动物，维护森林生态系统的生物多样性。6) 森林经营管理：科学管理和合理利用森林资源，发展林业经济，实现生态效益、经济效益和社会效益的协调发展。
林业政策知识图谱增强的LLaMa-Chinese生成的回答	<p>林业生态环境建设是指通过各种林业活动和措施，保护和改善森林生态系统，以促进生物多样性、维持生态平衡、提供生态服务和产品，以及实现可持续发展的过程。具体来说，林业生态环境建设包括以下几个方面：</p> <ol style="list-style-type: none">1) 森林保护和培育：保护现有森林资源，防止非法伐木和森林退化，同时通过植树造林等活动增加森林面积。2) 湿地保护和恢复：湿地作为生态系统的重要组成部分，对维持生物多样性和生态平衡具有重要作用。林业生态环境建设中湿地的保护和恢复是关键一环。3) 防沙治沙和水土流失治理：通过植被恢复和合理管理，减少沙漠化和水土流失，保护和改善土地资源。4) 野生动植物保护：加强自然保护区的建设和管理，保护濒危物种和生态系统的完整性。5) 林业产业发展：发展以森林资源为基础的产业，如木材生产、非木质林产品加工、森林旅游等，同时确保这些活动的可持续性。6) 生态文化建设：提升公众的生态文明意识，通过教育和文化活动，增强人们对森林和生态环境重要性的认识。7) 循环经济：林业是规模最大的循环经济体之一，通过科学经营和合理利用森林资源，实现资源的高效利用和循环再生。8) 生物质能源开发：利用林业资源开发可再生、可降解的绿色能源，如生物柴油和木质燃料，以减少对化石燃料的依赖。9) 生态文明建设：林业在生态文明建设中承担着重要使命，不仅要承担生态建设的重任，还要成为生态文化的建设者。10) 碳汇功能：森林作为陆地生态系统的主体，具有强大的碳汇功能，对实现“双碳”目标（即碳达峰和碳中和）具有重要作用。

Table 6: 知识图谱对大模型回答的增强案例分析

A6 最终构建的知识图谱可视化展示

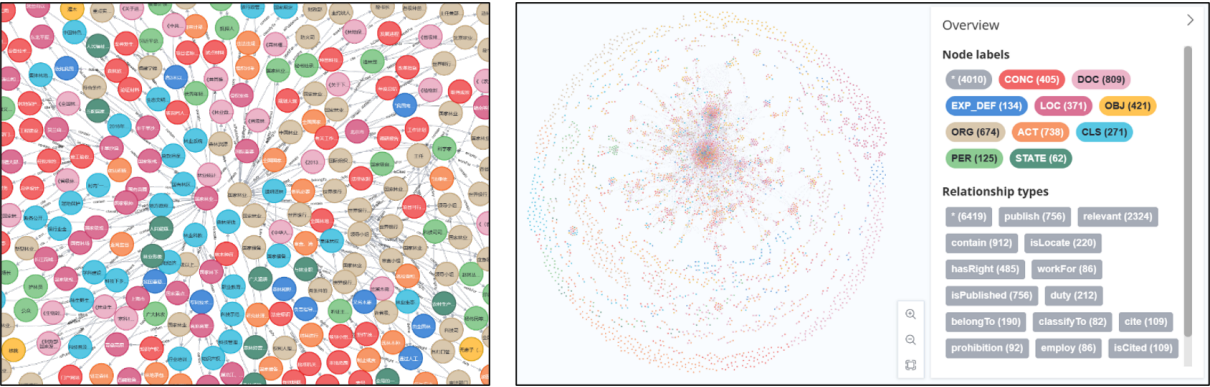


Figure 13: 所构建的林业政策知识图谱在Neo4j数据库中的可视化展示（左侧为局部细节图，右侧为全局图）