doi:10. 3969/j.issn. 1672-4623.2023.01.003

# 一种面向生态环境领域的知识图谱构建方法

王天一[1]，孟小亮[1*]，张　华[1]
（1. 武汉大学 遥感信息工程学院，湖北 武汉 430079）

**摘　要：**面向生态环境领域，提出了一种利用文献文本的领域知识图谱自动构建方法。首先利用机器学习技术抽取生态环境领域的关键词和关键关系；再结合统一建模语言系统构建生态环境领域知识图谱模式层，并以该模式层为指导建立生态环境领域语料库；然后采用深度学习方法从文本文献中自动抽取关键信息，通过来源指数和校验次数保证抽取知识的置信度；最终实现知识图谱的自动化构建与可视化展示。结果表明，利用2 000万字规模的文献文本可得到含29 490条知识三元组的知识图谱，利用Transformer模型抽取的总体精度为96.27%。

**关键词：**领域知识图谱；语义标注；深度学习；机器学习；生态环境
**中图分类号：**P208　　**文献标志码：**B　　**文章编号：**1672-4623（2023）01-0014-06

## Knowledge Graph Construction Method in the Ecology and Environment Field

WANG Tianyi[1], MENG Xiaoliang[1], ZHANG Hua[1]
(1. School of Remote Sensing Information Engineering, Wuhan University, Wuhan 430079, China)

**Abstract:** In this paper, we proposed a knowledge graph construction method in the ecology and environment field based on literature texts. Firstly, we applied machine learning and modeling language system to structure the model layer of domain knowledge graph. Then, we utilized a deep learning technique to extract key information from literature texts automatically. In order to guarantee the accuracy of extracted knowledge, we used the source index and verification times to ensure the confidence of extracted knowledge. Through experiments, we found that the knowledge graph containing 29 490 knowledge triplets was obtained by using 20 million words of literature texts, and the total extraction accuracy by using Transformer model was 96.27%.

**Key words:** domain knowledge graph, semantic annotation, deep learning, machine learning, ecology and environment

知识图谱是通过节点和有向边的方式来表达实体、概念及其相互之间关系的大型语义网络，是大数据时代背景下从数据互联迈向知识互联服务的基础[1-2]。目前，知识图谱在很多领域发挥作用，而应用知识图谱的前提是构建知识图谱。地学领域已有学者探究面向灾害应急响应的地质灾害链知识图谱的构建方法，但所用方法未充分利用丰富的文献文本资源[3]；部分信息管理领域学者对从文献等文本资源中提取的知识进行了研究，但通用方法对专业领域知识图谱构建具有局限性[4-5]。生态环境是人类赖以生存和发展的重要基础，人类社会的建设对生态环境安全造成了难以估量的影响，为了人类的可持续发展，生态环境的治理保护显得尤为重要[6]。生态环境领域期刊文献中的文本具有形式简洁、知识集中、现势性好等特点。本文选择生态环境领域文献作为数据源研究领域知识图谱的构建方法具有典型性。

本文设计的文献驱动的领域知识图谱构建方法如图1所示，包括数据获取与存储、领域知识图谱模式层构建、语料库语义标注、知识自动抽取与知识验证等部分。输入数据源可以是文献数据库、行业标准等，步骤①通过自动抓取程序、OCR文本识别以及PDF文本抽取等方式获取领域文献文本数据；步骤②将数据存储到原始语料数据库中；步骤③~④通过统计和机器学习方法，如TextRank、TF-IDF利用无监督关系抽取模型OPENIE抽取领域中关键的实体概念以及实体间的相关关系；步骤⑤利用领域知识和建模系统人工总结归纳、提炼得到领域知识图谱的模式层，并将其作为语义标注平台的标注依据；步骤⑥将模式层导入语义标注工具中；步骤⑦通过语义标注平台从原始语料库获取未标注的文本；步骤⑧人工参与标注领域实体和关系组建语义丰富的领域实体关系语料库；完成标注的语料库作为训练样

收稿日期：2022-09-30。
项目来源：国家重点研发计划资助项目（2020YFD1100203）；国家自然科学基金面上基金资助项目（41971352）。
（*为通信作者）

本，通过步骤⑨对领域知识自动抽取算法进行训练；训练完成后，将未标注的大量文本数据通过步骤⑩输入领域知识抽取算法；经算法处理后，步骤⑪得到抽取的知识三元组；步骤⑫对抽取的知识三元组进行验证；步骤⑬将符合规定条件的知识三元组存储到领域知识图谱中。

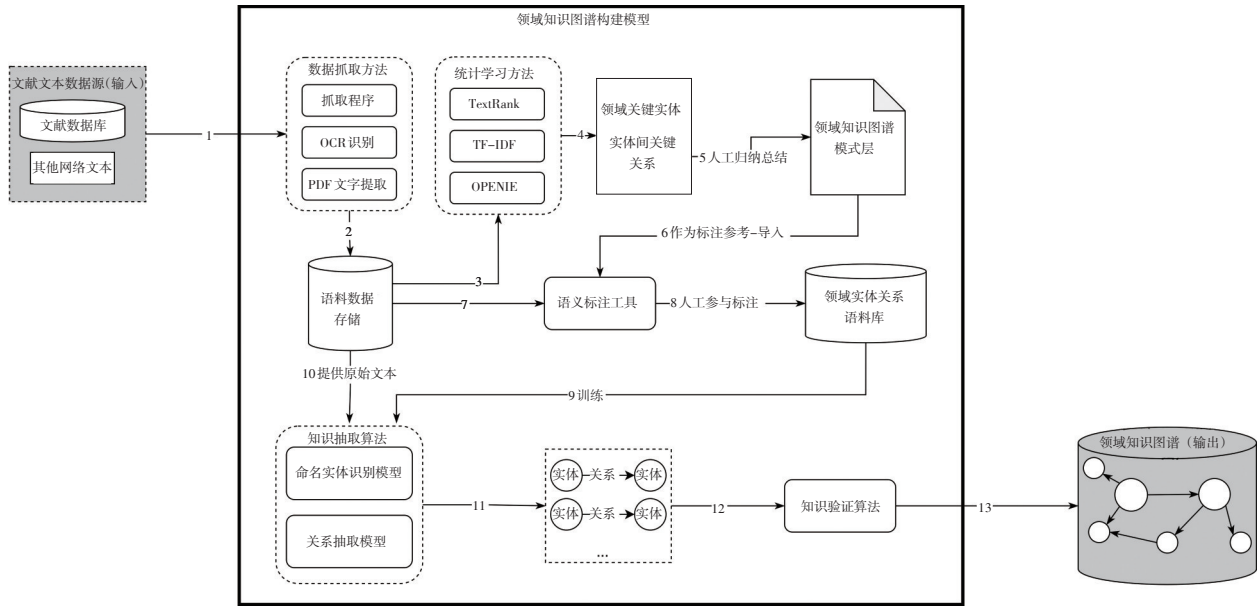



图1　文献驱动的领域知识图谱构建方法技术流程图

## 1　自动化语料内容数据获取

数据源的质量将对研究结果造成很大影响，文献文本数据规范、客观、可信。本文选取的数据来源于《生态学报》《环境科学》等专业生态环境领域中文期刊，均从收录期刊的文献数据库或期刊官网中获取。由于这些文本数据具有较规则的数据形式和特定的Dom结构，因此可通过分析页面要素的特点和规律，编写抓取程序对文章的标题、作者、关键词、摘要以及全文内容等信息进行自动获取。本文共获取2014—2020年见刊的5 000多篇文献约2 000万字的语料内容数据。

## 2　利用机器学习的图谱模式层构建

模式层构建是构建知识图谱的核心内容[7]，模式层构建即构建本体模型。本体模型的构建包括手工构建和自动构建等方法[8]，首先需明确概念、关系等内容，然后便可在某种本体语言的基础上进行本体建模[9]。传统构建方法几乎完全需要领域专家从语料文献中人工提炼，而自动构建方法可快速建立本体模型，但与人们的认知习惯有所差距[10]。因此，本文首先利用机器学习技术从大量生态环境领域文献文本中自动抽取领域关键词和领域关系，再人工从语义层面分析抽取结果，然后通过建模语言归纳出领域知识图谱的模式层，最终以构建的模式层为指导建立领域语料库。

### 2.1　领域关键词和领域关系抽取

关键词可集中反映相关领域的特征，本文利用TF-IDF与TextRank算法，并融合词频统计的方式自动挖掘生态环境领域关键词。TF-IDF算法兼顾了词频特征和逆文本频率指数，TextRank算法利用图的思想在提取时顾及了语义信息。在使用Python中文分词库jieba的基础上，利用各方法抽取领域关键词。为保证领域关键词抽取结果的可靠性，根据各方法的提取能力定权，加权计算综合重要性指标。$I_{综合}$为综合重要性指标，$I_{词频}$、$I_{TF-IDF}$、$I_{TextRank}$分别为归一化后的词频、TF-IDF值和TextRank值。利用式（1）得到1 500个关键词语及其重要级。为了可视化分析和直观展示抽取结果，本文制作了领域关键词云图（图2），可以看出，“污染”、“群落”、“生物”、“环境”等词语是生态环境领域中的关键概念，但“土壤”、“碳”等一些词语是要素实例，还存在一些高频词汇与生态环境领域相关性不高，如“评价”、“可能”、“说明”等。

$$I_{综合}=0.2I_{词频}+0.4I_{TF-IDF}+0.4I_{TextRank} \quad (1)$$

关系是指实体或概念间的联系，是知识图谱模式层的核心要素。本文首先对文本数据进行分句处理，然后通过OPENIE工具（来源于Stanford CoreNLP自然语言处理工具包）进行无监督开放信息抽取，最终提炼出135 971条关系三元组。人工筛除含义不明的关系后，统计关系出现的频次，归纳得到具有生态环境




图2 生态环境领域关键词云图

领域含义的关系，如表1所示。

表1 生态环境领域关系抽取与筛选结果

| 关系 | 频次 | 关系 | 频次 | 关系 | 频次 |
|---|---|---|---|---|---|
| 具有 | 1 274 | 产生 | 203 | 释放 | 84 |
| 存在 | 666 | 影响 | 221 | 胁迫 | 82 |
| 高于 | 645 | 降低、减少 | 223 | 提高、改善 | 141 |
| 分布、位于 | 863 | 形成 | 131 | 促进 | 66 |
| 排放 | 459 | 属于 | 131 | 增加、随着 | 65 |
| 处理 | 424 | 造成 | 130 | 下降 | 61 |
| 吸附 | 308 | 导致 | 112 | 抑制 | 56 |
| 去除 | 294 | 提取 | 107 | 不利于 | 56 |
| 分析 | 281 | 研究 | 95 | 萃取 | 54 |
| 采样 | 267 | 组成 | 89 | – | – |

### 2.2 领域知识图谱模式层建立

UMLS是一种工具化、集成化、跨领域的语言系统[11]。本文将UMLS的顶层概念（图3）与提取的领域关键词和关系相结合，构建生态环境领域知识图谱的模式层。

参照UMLS体系，将领域知识体系在顶层概念中分为概念实体、物理实体、现象过程和活动4个实




图3 UMLS顶层概念

体，含义分别为人类主观认知的实体，客观存在的实体，唯物的、无意识的自然现象，计划性的、可归因的人类行为。通过对生态环境领域关键词和关系的归类，归纳得到生态环境实体列表与关系列表（表2、3）。

表2 生态环境领域实体列表

| 实体类型 | 相关词语 |
|---|---|
| 环境、生态环境要素 | 土壤、环境、水体、大气、沉积物、城市、流域、污泥、典型、水稻、河流、有机、物质、表面、地下水、水库、农业、气溶胶、农田、湖泊、底泥、径流、地表水、长江、稻田、地表、城区 |
| 量化指标、评价标准 | 浓度、含量、评价、水平、质量、活性、水质、范围、温度、指标、指数、表层、排放量、能力、粒径、功能、强度、通量、均值、梳理、生物量、性质、特性、产量、空气质量 |
| 生态环境风险—污染物 | 重金属、污染、污染物、风险、颗粒、污水、臭氧、废水、抗生素、氮磷、化合物、金属 |
| 分析、研究、处理手段 | 研究、模型、方法、分析、吸附、处理、采样、反应器、工艺、方式、测定 |
| 生物、生物群落 | 生物、植物、细菌、物种、秸秆、生物质、藻类、根系、人类 |
| 数据、参数 | 样品、因子、数据、组分、参数、水样 |
| 属性—空间信息 | 区域、地区、中国、空间、北京、分布、上海、美国 |
| 生态环境现象、过程 | 过程、硝化、氧化、降雨 |
| 生态环境风险—污染源 | 来源、工业、污染源、机动车、企业 |
| 属性—时间 | 时间、季节 |
| 人类活动 | 标准、经济 |

表3 生态环境领域关系列表

| 关系 | 相关词语 |
|---|---|
| 处理分析研究 | 处理、吸附、去除、分析、采样、提取、研究、萃取、采集 |
| 排放产生造成导致 | 排放、产生、造成、导致、释放、来源于、是因为 |
| 含有属性—空间 | 分布、位于 |
| 影响相关促进抑制 | 高于、影响、降低减少、胁迫、提高改善、促进、增加、随着、下降、抑制、不利于、 主要受 |
| 含有其他属性 | 具有 |
| 包含 | 属于、组成、含 |
| 存在 栖息于 | 存在 |
| 关系 | 相关关系 |
| 处理分析研究 | 处理、吸附、去除、分析、采样、提取、研究、萃取、采集 |

生态环境的核心是环境要素和生物体。人类活动造成的损害与生态治理过程中的分析都不断影响着生态环境[12]。沿此思路，本文将知识图谱模式层归纳为生态环境风险、生态环境主体、生态功能现象过程、人类活动、分析方法五大要素。UMLS顶层概念将实体间关系分为上下位关系和相关关系，其中相关关系涵盖环境实体间、属性实体间、环境与属性实体间的关系，环境实体间的关系又包含相关关系、概念关系、等价关系等。整合上述生态环境实体与关系，得到生态环境领域知识图谱模式层，如图4所示。

### 2.3 领域语料库建立

建立语料库是自动抽取知识的先决条件。目前生

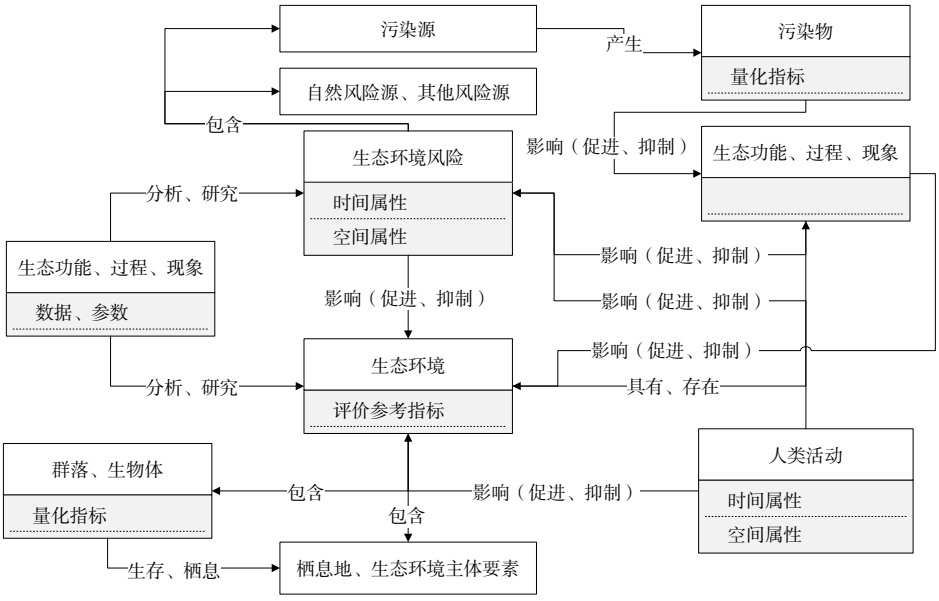



图4　生态环境领域知识图谱模式层

态环境领域还没有公开可查询的语义标注集，且诸多语义标注软件在用户体验上不能满足需求。语义标注应从知识图谱的模式层出发，包含实体和关系的标注。关系表达为三元组 $(E_1, R, E_2)$，$R$ 表征实体 $E_1$ 和 $E_2$ 的关系。本文设计和开发了语义标注软件工具，主界面如图5所示。语义标注流程为：①标注实体，标注人员在选定文字后，选择实体类型与空间信息，点击标注按钮即可完成实体标注；②标注关系，先在实体列表上设置头实体和尾实体，再设置关系类型点击标注关系按钮即可标注一条关系。本文共标注了40万字的语料，其中实体类别标签2.2万条，关系类别标签1.1万条。




图5　设计实现的语义标注软件工具

## 3　领域知识图谱自动化构建

在生态环境领域，已有学者针对生态环境领域命名实体识别进行了研究，如蒋翔[13]等基于BiLSTM-IDCNN-CRF模型进行了生态治理技术领域命名实体识别。本文采用BERT-BiGRU-CRF模型进行生态环境领域实体提取，并讨论了最佳的实验参数；采用Transformer[14]特征提取网络对生态环境领域实体之间的关系进行抽取；利用校验次数和来源指数对抽取的知识进行质量验证，从而自动构建生态环境领域知识图谱。

### 3.1　实体信息自动抽取

在构建的语料库的基础上，通过程序生成训练样

本，实体利用程序转换并输出参考BIO方案的标注结果，标记数据以60%、20%、20%的比例随机用于训练、验证和最终测试。BERT[15]是应用广泛的词向量编码模型，几乎替代了以Word2vec[16]为主流的词向量编码模型，在许多自然语言处理任务中精度更高；对于长序列文本处理，GRU[17]比LSTM[18]模型所需的训练样本更少，CRF[19]（条件随机场）是序列化标注算法。本文利用样本训练BERT-BiGRU-CRF模型并测试，相关文献指出模型学习率与BatchSize比值越大泛化性越好[20]，以此为指导权衡训练时间，最终调整BatchSize为4、学习率为1e-5时，平均准确率为88.64%、平均召回率为88.90%，各类的识别准确率在85%~99%之间。

### 3.2 实体概念间关系自动抽取

本文利用Transformer网络进行关系抽取。基于该网络的编码器组件，将数据输入到Transformer网络的Encoder中，多头注意力机制将从语句中提取重要信息，然后将注意力层的输出与输入进行合并和正则化。为充分抽取不同特征等级的句子信息，此处借鉴ResNet网络的方法，将Transformer的结果拼接并传送给全连接层输出分类结果。经试验得到，学习率为1e-4、BatchSize为16时，取得最优的平均准确率（97.25%）。

### 3.3 知识验证与知识图谱存储

生态环境领域实体和关系自动抽取完毕后，得到了数目庞大的领域实体和关系构成的三元组，即知识图谱的知识单元。这些通过深度学习算法抽取的知识单元并非完全可靠，不能直接构建知识图谱，需先将其暂存在Elasticsearch数据库中，再利用知识质量验证算法过滤筛选得到可靠的知识，最终将验证后的知识数据同步到Neo4j数据库。

知识来源是衡量知识置信度的重要指标。本文设置$Q$为来源指数（指数越高表明知识越可信），图谱知识来源包括人工标注文献文本、模型识别文献文本、模型识别其他文本等，模型的训练样本是经人工标注的文献文本，且基于最大似然的思想，认为文献文本数据、人工标注数据相对于其他来源的数据可信度更高，综上得出式（2）。

$$Q_{\text{人工标注}} > Q_{\text{模型抽取}},\ Q_{\text{文献}} > Q_{\text{其他}} \tag{2}$$

$T$为校验次数，是指同一个知识三元组被尝试插入知识图谱多次，其中来源指数不小于现有图谱中该三元组来源指数的次数。当某知识校验次数大于阈值，可将该知识视为正确。知识三元组需经过$Q$、$T$指数的考核才能录入图谱数据库，知识更新的流程为：当知识图谱中插入新的知识三元组时，首先从知识图谱中查询是否具有头尾实体相似的三元组，若无，则该三元组是新知识，设置知识来源后插入知识图谱；若有，如果关系相同，判断来源是否更加权威，更新现有相似知识三元组的来源指数和校验次数，如果关系不同，则比较知识三元组置信水平，倘若图谱中知识三元组校验次数超过阈值，则认为知识正确，倘若校验次数不及阈值，则比较知识来源的权威性，新增知识权威性较高则插入，反之则丢弃，否则比较模型识别的置信度，新知识置信度较高则插入知识，随后更新$Q$、$T$指数，如图6所示。

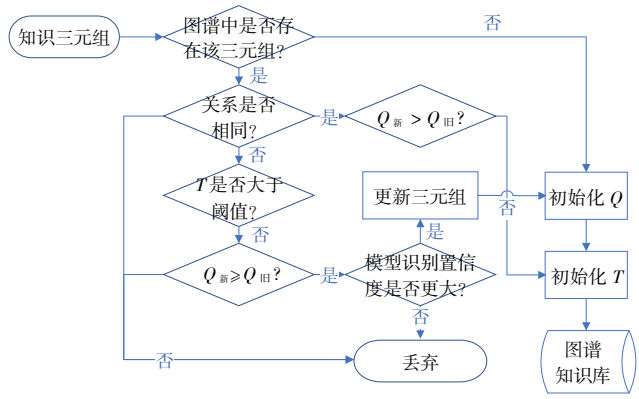



图6 知识验证流程图

本文通过上述步骤自动构建知识图谱，并将部分结果做可视化处理，结果如图7所示，通过认证次数确定知识节点的符号大小，不同节点颜色表示不同类别的知识节点，知识节点的名称通过文本显示。

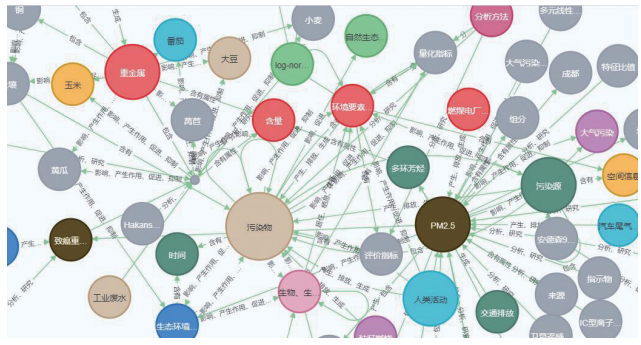


图7 部分知识图谱可视化

## 4 结 语

当前海量生态环境文献文本中蕴含了丰富的非结构化领域专业知识，这些数据可被用于构建知识图谱。本文提出了一种生态环境文献文本驱动的自动化领域知识图谱构建方法。该方法以生态环境类期刊文献文本为数据源，通过构建生态环境领域知识模型，标注生态环境领域语料库，并基于深度学习方法抽取

生态环境领域信息，实现生态环境领域知识图谱的自动化构建。

1）本文共获取5 000余篇生态环境领域文献，共计2 000万余字，采用机器学习方法抽取保留了重要性指数前1 000的词组和关系，再从抽取结果中总结归纳构建生态环境领域知识图谱模式层。

2）从领域知识图谱模式层出发，分析了标注过程的需求，自主开发了语义标注工具，并利用该工具识别语料40万字，完成了2.2万条实体与1.1万条关系的语义标注工作。

3）本文进行了生态环境领域信息抽取算法的训练与调优。实验结果表明，BERT-BiLSTM-CRF模型在BatchSize(4)、Learningrate（3e-5）时效果较好、效率较高，抽取总体精度为89.09%，召回率为88.60%，F1值为88.85；Transformer模型在BatchSize（16）、Learningrate（1e-4）时效果最好，抽取总体精度为96.27%，召回率为97.33%，F1值为96.80%。抽取数据经过本文提出的知识验证算法筛选后存储到知识图谱数据库中，最终得到29 490条知识三元组。

本文还可在以下两个方面进一步开展工作：①语义标注时，在选取文本契合度方面可进一步利用专家知识，并研究提高标注效率的方法；②本文使用的实体提取和关系提取模型仍为通用模型，可进一步对生态环境领域实体提取和关系提取的网络结构进行优化。

## 参考文献


[1] 陆锋,余丽,仇培元. 论地理知识图谱[J]. 地球信息科学学报,2017,19(6):723-734

[2] 徐增林,盛泳潘,贺丽荣,等. 知识图谱技术综述[J]. 电子科技大学学报,2016,45(4):589-606

[3] 邱芹军,吴亮,马凯,等. 面向灾害应急响应的地质灾害链知识图谱构建方法[J/OL]. (2022-08-30)[2022-09-15]. https://kns.cnki.net/kns8/Detail?sfield=fn&QueryID=11&CurRec=1&recid=&FileName=DQKX20220827003&DbName=CAPJLAST&DbCode=CAPJ&yx=Y&pr=&URLID=42.1874.p.20220829.1828.014

[4] 姚萍,李坤伟,张一帆. 知识图谱构建技术综述[J]. 信息系统工程,2020(5):121,123

[5] 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

[6] Steffen W, Broadgate W, Deutsch L, et al. The Trajectory of the Anthropocene: The Great Acceleration[J]. Anthropocene Review,2015,2(1):81-98

[7] 董晓晓,周东岱,黄雪娇,等. 学科核心素养发展导向下教育领域知识图谱模式构建方法研究[J]. 电化教育研究,2022,43(5):76-83

[8] 虞为,曹加恒,陈俊鹏. 基于本体的地理信息查询和排序[J]. 计算机工程,2007,33(21):157-159

[9] 张斌,李宏伟,崔铁军. 全息地理空间本体模型构建[J]. 地理空间信息,2021,19(12):1-6

[10] 张恒才,张晶,陈文倩. 基于空间认知的校园本体的构建[J]. 地理空间信息,2008,6(5):88-91

[11] 白海燕,王莉,梁冰,等. UMLS及其在智能检索中的应用[J]. 现代图书情报技术,2012(4):1-9

[12] 欧阳志云,王效科,苗鸿. 中国生态环境敏感性及其区域差异规律研究[J]. 生态学报,2000,20(1):9-12

[13] 蒋翔,马建霞,袁慧. 基于BiLSTM-IDCNN-CRF模型的生态治理技术领域命名实体识别[J]. 计算机应用与软件,2021,38(3):134-141

[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is All You Need[J]. Advances in Neural Information Processing Systems,2017(30):126-132

[15] Devlin J, CHANG M W, Lee K, et al. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv Preprint arXiv:1810.04805,2018

[16] Mikolov T, CHEN K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. arXiv Preprint arXiv:1301.3781,2013

[17] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning Phrase Representations Using RNN Encoder-decoder for Statistical Machine Translation[J]. arXiv Preprint arXiv:1406.1078,2014

[18] SHI X, CHEN Z, WANG H, et al. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting[J]. Advances in Neural Information Processing Systems,2015(28):136-147

[19] SONG S, ZHANG N, HUANG H. Named Entity Recognition Based on Conditional Random Fields[J]. Cluster Computing,2017(1):1-12

[20] HE F, LIU T, TAO D. Control Batch Size and Learning Rate to Generalize Well: Theoretical and Empirical Evidence[J]. Advances in Neural Information Processing Systems,2019(32):265-274


**第一作者简介：**王天一（2001—），研究方向为空间信息与数字技术。