计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications [doi: 10.15888/j.cnki.csa.008637]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

成果地质资料知识图谱构建与可视化①

王　晴[1], 黄　进[1], 刘　鑫[1], 翟树红[2], 方　铮[3], 李剑波[4]

[1](西南交通大学 电气工程学院, 成都 611756)
[2](四川省自然资源资料馆, 成都 611756)
[3](四川省国土科学技术研究院, 成都 611756)
[4](西南交通大学 计算机与人工智能学院, 成都 611756)
通信作者: 黄　进, E-mail: jhuang@swjtu.edu.cn

摘　要: 知识图谱技术在行业领域的运用越来越广, 因此研究知识图谱技术在成果地质资料领域中的运用, 解决到馆用户的精确查询和可视化问题变得更加重要. 本文以成果地质资料为研究对象, 利用爬虫技术, 爬取成果地质资料中的矿产、地理区域、组织机构等实体信息. 结合知识图谱相关技术, 设计成果地质资料知识图谱地质实体和关系, 经过命名实体识别、关系抽取和属性抽取, 构建成果地质资料实体 266 787 个, 关系 306 686 个. 使用 Neo4j 图形化数据库存储知识图谱来提高地质资料的查询性能, 方便到馆用户的查询. 该研究可以为知识图谱在成果地质资料上面的应用提供理论支撑.

关键词: 成果地质资料; 知识图谱; NLP; Neo4j

引用格式: 王晴,黄进,刘鑫,翟树红,方铮,李剑波.成果地质资料知识图谱构建与可视化.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/8637.html

Construction and Visualization of Knowledge Graph of Geological Report

WANG Qing[1], HUANG Jin[1], LIU Xin[1], ZHAI Shu-Hong[2], FANG Zheng[3], LI Jian-Bo[4]

[1](School of Electrical Engineering, Southwest Jiaotong University, Chengdu 611756, China)
[2](Sichuan Natural Resources Museum, Chengdu 611756, China)
[3](Sichuan Academy of Land Science and Technology, Chengdu 611756, China)
[4](School of Computer and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China)

Abstract: Knowledge graph technology is used increasingly widely in industries. Therefore, it is more important to study its application in the field of geological reports to realize accurate queries and visualization for library users. Taking geological reports as the research object, this study uses crawler technology to obtain mineral, geographic area, organization and other entity information in geological reports. With the help of related technologies of knowledge graph, the study designs geological entities and relations for the knowledge graph of geological reports. After named entity recognition, relation extraction and attribute extraction, 266 787 geological report entities and 306 686 relations are constructed. Knowledge graphs are stored in the Neo4j graph database to improve the performance of geological data query and facilitate the query of library users. This research can provide theoretical support for the application of the knowledge graph to geological reports.

Key words: geological report; knowledge graph; NLP; Neo4j

新时代中国特色社会主义, 提出需要坚持“创新、协调、绿色、开放、共享”的新发展理念, 因此地质调查工作需要及时进行转型升级, 同时坚持公益性、基础性、战略性的精准定位也十分重要[1]. 地质资料主要

① 基金项目: 国家自然科学基金 (61733015); 高铁联合基金 (U1934204); 四川省自然资源科研项目 (KYL202106-0099)
收稿时间: 2021-11-17; 修改时间: 2021-12-13; 采用时间: 2021-12-21; csa 在线出版时间: 2022-05-31

包括成果资料、原始资料和实物资料 3 种类型, 同时, 地质资料也是地质工作记录和成果的表现方式. 本文主要以馆藏成果地质资料为研究对象, 利用爬虫技术、命名实体识别、关系抽取、属性抽取等相关技术和 Neo4j 图数据库来构建成果地质领域知识图谱. 知识图谱是一种结构化的语义网络知识库[2], 其主要的目的是提高搜索引擎的能力, 增强用户的搜索质量以及搜索体验[3]. 国内, 百度、搜狗等将知识图谱的研究从概念转向产品应用[4]. 陆汝钤院士提出了知见的概念[5]、Chen 等人提出了 AgriKG, 将知识图谱应用于农业领域, 构建了农业知识图谱[6]. 国外也已有较多重要的知识图谱研究成果, 如 Google Knowledge Graph、DBpedia、YAGO 和 Freebase 等[7].

馆藏成果地质资料指的是地质资料汇交人将成果地质资料按照规定要求提交后, 由馆藏机构对其进行保存和提供利用的成果地质资料. 馆藏成果地质资料不仅是国家重要的基础性信息资源, 同时也是社会化的公共产品. 本文主要以馆藏成果地质资料为对象来构建地质资料领域知识图谱. 首先获取成果地质资料领域复杂多样的知识, 然后探索成果地质资料领域知识图谱的构建方法, 设计成果地质资料知识图谱的地质实体和关系, 通过知识图谱可以清晰地了解到地质矿产与地理区域、组织机构的关系. 本文的贡献主要如下:

(1) 利用序列标注工具构建了成果地质资料领域的语料库, 其中包含了矿产名称、组织机构、地理区域等相关语料实体.

(2) 利用命名实体识别、关系抽取等相关技术将成果地质资料领域的文本中的非结构化数据转化为结构化数据.

(3) 利用 Neo4j 图形化数据库构建了成果地质矿产领域的知识图谱. 这是首次将知识图谱技术应用于成果地质资料领域.

1 成果地质资料知识图谱框架设计

知识图谱主要可以分为通用知识图谱和行业知识图谱[8]. 本文主要根据四川省自然资源资料馆提供的馆藏成果地质资料为基础, 研究成果地质领域知识图谱构建与可视化. 将馆藏成果地质资料档案和网络百科的相关地质资料知识相结合, 利用爬虫技术, 爬取成果地质资料中的地质矿产、地理区域、组织机构等实体信息, 通过对得到的地质数据进行清洗、抽取和融合处理, 经过实体识别、关系抽取和属性抽取等步骤, 构建成果地质资料领域知识图谱, 属于行业领域的知识图谱, 图 1 为成果地质资料知识图谱构建流程图.

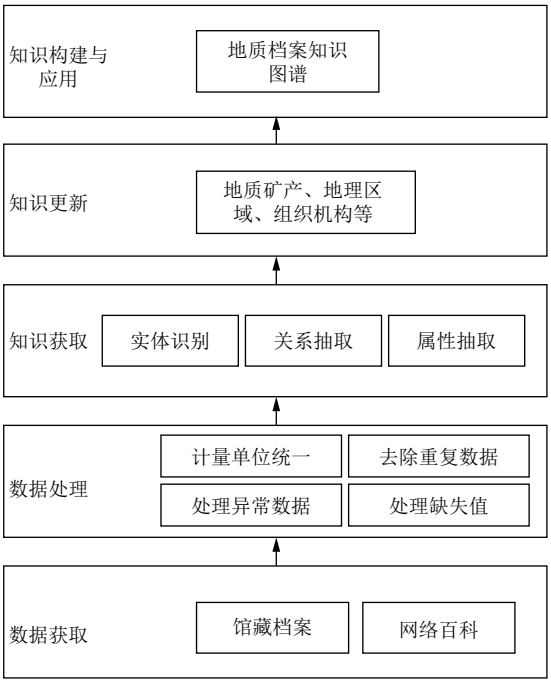

图 1 成果地质资料知识图谱构建流程图

(1) 数据获取与处理. 地质数据是地质知识模型的载体[9], 因此, 对于地质数据的研究就是对于地质知识模型的研究. 本文主要通过获取馆藏成果地质资料和网络百科来获取地质数据, 其中包含了结构化、半结构化和非结构化的数据. 对结构化的数据, 可直接利用规则的方法把地质相关实体映射到知识图谱中[10]. 比如文本数据中的“四川彭县铁矿地质简报”属于结构化的数据. 对于成果地质资料中的非结构化数据, 主要是从文本中抽取出地质实体及关系等信息. 首先对成果地质资料进行预处理, 包括分词、词性标注、句法分析等, 然后利用命名实体和关系抽取技术得到需要的地质实体和关系.

(2) 命名实体识别. 命名实体识别是自然语言处理的一项基础任务, 主要是因为命名实体任务性能的提高将有利于非结构化文本朝结构化文本的转化[11]. 成果地质资料具有丰富的领域性特征且文本具有高度非结构化的特征, 梳理地质实体的不同类型、固有的关系和属性, 完成地质实体的识别与标注工作, 建立“成果地质内容标签”语料库. 在 BERT 框架下研究中文地质命名实体识别方法, 采用预训练语料库模式从规模

化的地质非结构化文本数据中自动抽取出实体信息. BERT 预训练模型如图 2 所示, 主要包含预训练和微调两个阶段. BERT 只需一个额外的输出层就可以对预先训练的模型进行微调[12]. 比如成果地质资料数据中的“受西南地质调查所安排进行调查. 铁矿产于侏罗纪中下部地层中, 矿石为赤铁矿, 具鲕状或砾状结构”等非结构化数据, 我们需要提取出组织名称“西南地质调查所”和地质矿产名称为“赤铁矿”等实体内容.

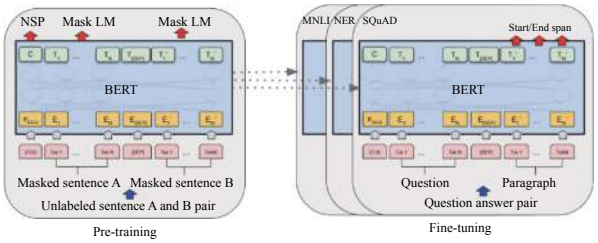

图 2 BERT 预训练模型

(3) 知识更新. 成果地质资料知识图谱的构建与应用, 将提取到的地质实体、关系和属性等结合成果地质资料领域知识的特点, 构建了成果地质资料知识图谱. 利用 Neo4j 图数据库来负责成果地质资料知识图谱节点的存储, 将提取到的地质实体、关系和属性导入到 Cypher 查询模板中, 实现成果地质资料知识的精确查询[13], 从而便于地质资料领域知识更新和到馆用户的查询.

2 关键技术设计

2.1 多源异构数据的获取与处理

多源异构数据指的是不同来源、不同结构的数据[14]. 将多源异构数据转化为符合知识图谱构造的三元组形式是非常重要和关键的技术. 成果地质资料数据主要来源于四川省自然资源资料馆、在线百科等. 馆藏成果地质资料数据具有结构复杂、类型多样的特征, 研究多源异构数据的采集、清洗、脱密、脱敏和集成关键技术, 研究对于半结构化和非结构化数据的实体抽取、关系抽取和属性抽取等知识抽取技术. 对于结构化的数据可以采用规则映射的方法, 对于半结构化和非结构化的数据需要进行命名实体识别、关系抽取从而将它转化为结构化的数据, 本文采用深度学习的方法进行处理, 从而获得地质实体和关系.

2.2 成果地质资料领域语料库构建

语料库是指大量文本数据的集合, 所以文本数据都需要经过一定的预处理后才能成为后续的研究的基础数据[15]. 本文采用 BIO 格式的序列标注方法[16], 将成果地质资料中的一部分数据拿来制作语料库, 把一部分数据的每个字标注为“B-X”“I-X”或者“O”格式. “B-X”表示该字为实体的首字属于 X 类型且在实体的开头, “I-X”表示该字属于 X 类型且在实体类型的中间位置, 其中, “X”就在本文中就包括了地质矿产名称、地理区域名称、组织机构、地质简报名称、人物名称以及时间等信息. “O”表示不属于任何类型的实体. BIO 格式构建的语料库如表 1 所示. 比如“西南地质调查所”的首个字标注为“B-ORG”表示“西”是这个实体的首字且属于“ORG“类型的实体, 其他部分标注为“I-ORG”, 表示该字是实体的非首字.

表 1 BIO 格式构建语料库

中文预料	标签
受	O
西	B-ORG
南	I-ORG
地	I-ORG
质	I-ORG
调	I-ORG
查	I-ORG
所	I-ORG
安	O
排	O
进	O
行	O
调	O
查	O
.	O

2.3 成果地质资料实体识别和关系抽取

命名实体识别指识别人名、组织名、地名等. 对标注后的语料进行训练可以得到实体抽取的结果, 如表 2 所示. 从表中可以看到抽取到的实体包括地理位置、组织机构、地质矿产、人物、时间等信息. 其中“LOC”代表识别到的是地理区域实体, “ORG”代表识别到的是组织机构实体, “ROCK”代表识别到的是地质矿产实体, “PER”代表识别的是人物名称实体, “TIME”表示识别到的是时间实体.

命名实体识别任务常采用的评价指标有精确率:

$$Precision = \frac{TP+TN}{TP+FN+FP+TN}$$

其中, *TP* 指将正预测为真, *FN* 将正预测为假, *FP* 指将反预测为真, *TN* 指将反预测为假.

在整个成果地质资料档案知识图谱构建过程中, 关系抽取[17] 至关重要, 基于地质档案的关系抽取包括

了空间关系抽取、语义关系抽取[18]、时间关系抽取几个部分, 其技术流程如图 3 所示. 首先, 馆藏档案资料通过规则建立来进行空间关系抽取, 然后通过关系融合进行实体链接. 通过对档案资料数据结构分析, 其中包含了结构化数据、半结构化数据和非结构化数据, 然后进行知识抽取, 包括空间、语义、时间的关系抽取, 最后进行实体链接.

表 2 实体抽取示例

实体类型	实体标签
纂江县	LOC
麻柳滩	LOC
白石塘	LOC
西南地质调查所	ORG
四川地质调查所	ORG
四川省国土资源厅	ORG
赤铁矿	ROCK
磁铁矿	ROCK
黄铜矿	ROCK
菱铁矿	ROCK
张其泽	PER
张宇	PER
李大宽	PER
1899	TIME
2014/11/4 0:00	TIME

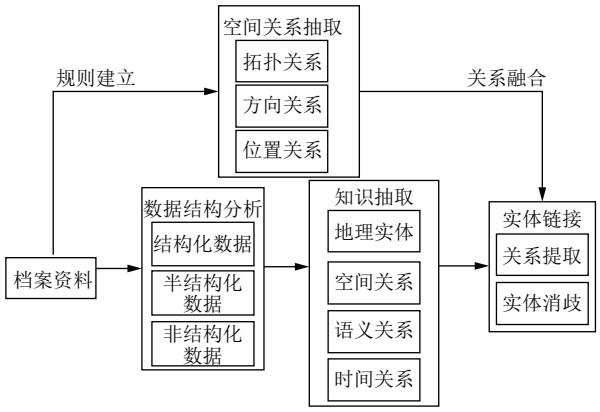

图 3 地质档案知识图谱关系抽取流程图

2.4 知识图谱设计

知识图谱是一种对事实的结构化表征. 当获取的数据比较大并且结构复杂时, 用知识图谱来表示会更加的清晰准确[19]. 研究知识图谱动态演化的事件图谱可视化技术, 满足不同业务场景的智能服务需求, 进一步提升馆藏服务水平. 经过命名实体识别、知识抽取后, 整理成果地质资料包含的地质矿产类型、行政区名称、矿产名称等实体. 实体类型设计如表 3 所示. 比如矿产类型包含了闪锌矿、磁铁矿. 行政区类型包含了攀枝花市、会理县等. 根据地质资料的实体类型和关系模型, 从而构建“地质实体-关系-地质实体”三元组[20], 地质资料领域三元组设计如表 4 所示, 其中包含了含矿种类的关系, 比如攀西地区含矿类型为钒钛磁铁矿. 包含了矿区隶属地的关系, 比如矿区隶属于攀枝花市东区银江镇马坎村等.

表 3 知识图谱实体类型设计

实体类型	中文含义	举例
矿产	矿产	闪锌矿、磁铁矿
行政区	行政区	攀枝花市、会理县
地层	地层	寒武系、震旦系
矿区	矿区	白马矿区、宝鼎矿区
金属元素	金属元素	TFe、mFe
方位	方位	西南部、西侧
组织机构	组织机构	四川地勘局
数字	数字	0.088平方公里
时间	时间	2012年底
项目	项目	全国国土遥感综合信息系统建设
人物	人物	孙永军、姜琦刚

表 4 知识图谱关系设计

实体关系类型	举例
含矿种类	“攀西地区”含矿“钒钛磁铁矿”
面积	“测区”面积“7332平方公里”
矿区隶属地	“矿区”隶属于“攀枝花市东区银江镇马坎村”
实施单位	“全国国土遥感综合调查与信息系统建设”实施单位“中国国土资源航空物探遥感中心”
委托机构	“煤攀公司”委托“四川省煤田地质勘查设计研究院”
矿物	“I-1号铁矿”, 厚度, “2.98m”
负责人	“全国国土遥感综合调查与信息系统”负责人“孙永军”
项目时间	“四川省冕宁县大桥铁多金属矿预查”时间“2013年”
资源量	“I-2号矿”资源“5.1万吨”
相距距离	“灰浦煤矿”16公里“太平乡”
组织机构地点	“志森工贸有限责任公司”位于“攀枝花市”

3 知识图谱系统构建与应用

3.1 地质资料知识图谱的构建与实现

知识图谱的核心思想是将数据表示为图形, 节点表示具体的对象、信息或概念, 边表示语义关系[21]. 根据馆藏成果地质资料来获取关于地质矿产、组织机构、地理位置、地质简报名称等数据. 将数据导入到 Neo4j 图数据库之后, 我们可以得到馆藏成果地质资料领域的知识图谱. 如图 4 所示为馆藏成果地质资料领域的知识图谱, 同一种颜色的“圆”属于同一种地质实体类型, 不同颜色的“圆”代表不同的地质实体类型, 不同颜色的“圆”之间的连线代表地质实体与实体之间的关系. “圆-线-圆”对应“地质资料实体-关系-地质资料实体“三

元组. 三元组是知识图谱的通用表示形式. 其中, 红色代表地质矿产, 绿色代表地理位置, 黄色代表地质简报的名称, 蓝色代表组织机构名称, 不同实体之间的关系通过线来连接, 从而构建了馆藏成果地质领域的知识图谱.

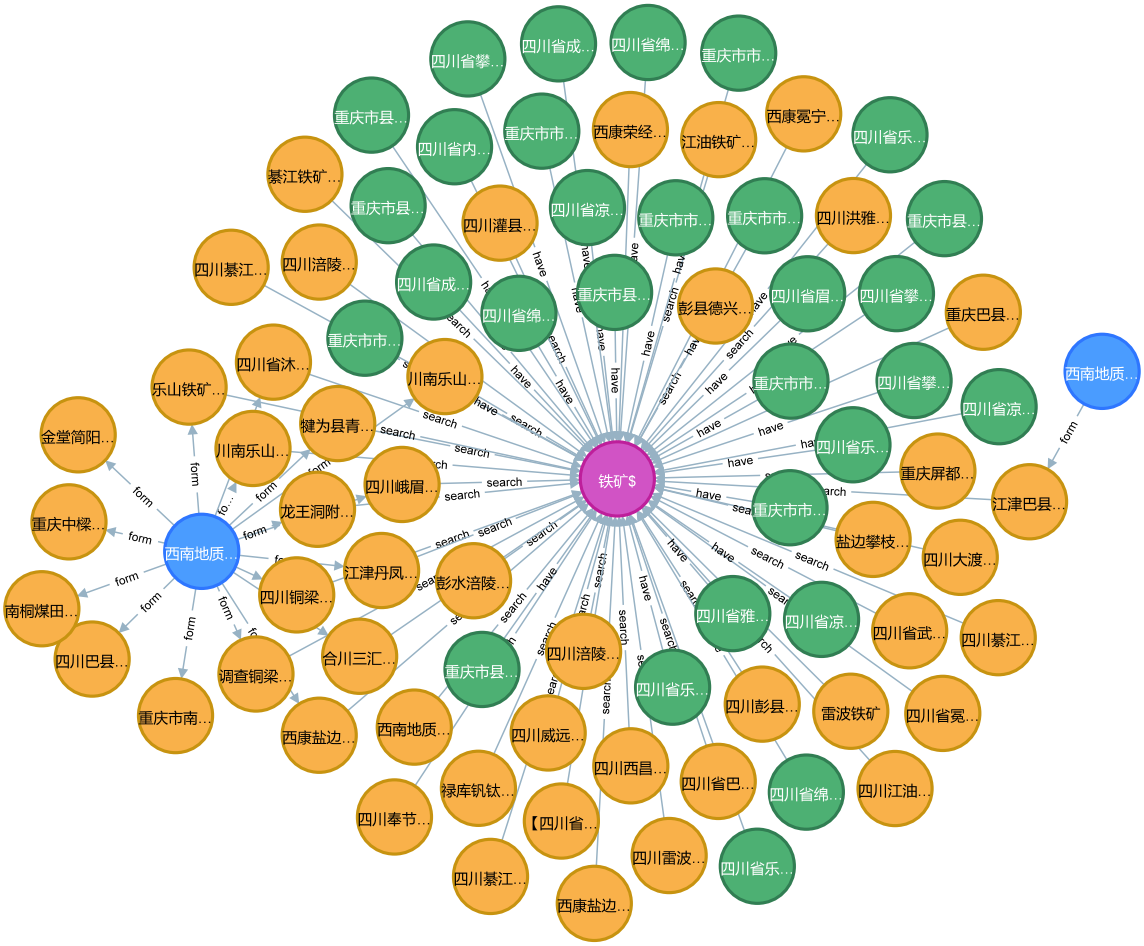

图 4 Neo4j 构建的地质资料知识图谱

3.2 矿产知识查询与可视化

知识图谱的可视化可以让人直观地了解到图谱中的关系信息. Neo4j 数据库里面的 Cypher 语言可以对数据库进行 CRUD (create, read, updata, delete) 的一系列相关操作, 从而方便实现地质资料领域数据的查询和更新功能. Neo4j 图形化数据库的高查询性能以及查询语言可定制化, 不仅可以查询地质实体与实体之间的关系, 还可以实现地质矿产的查询, 以返回快速、精准、结构化的知识. 如图 5 所示, 展示了西南地质调查所节点的相关信息, 从图中可以看到西南地质调查所与它所形成的简报名称相连, 从而可以实现成果地质图谱的可视化.

4 结论与展望

目前, 知识图谱技术是人工智能的热门研究方向, 并且还会在未来很长一段时间有长足的发展. 随着地质领域信息化的发展, 成果地质数据有了一定的积累. 构建成果地质领域的知识图谱, 可以从海量数据中提炼出地质资料相关知识, 并合理高效的对其进行管理、共享及应用, 对现今的地质行业有着重要意义, 也是许多研究机构的研究热点. 本文构建了成果地质资料的语料库, 其中包含了矿产名称、组织机构、地理区域等相关语料实体. 利用命名实体识别、关系抽取等相关技术将成果地质矿产领域的文本中的非结构化数据转化为结构数据. 利用 Neo4j 图形化数据库构建了成果地质矿产领域的知识图谱. 在未来的研究工作中, 我们将结合知识图谱的问答系统来实现地质资料知识的智能问答, 这也是接下来我们工作的研究重点. 虽然目前对于成果地质资料知识图谱的研究有了许多

很有意义的尝试, 但总的来说还不够完善和深入, 需要更进一步的研究.

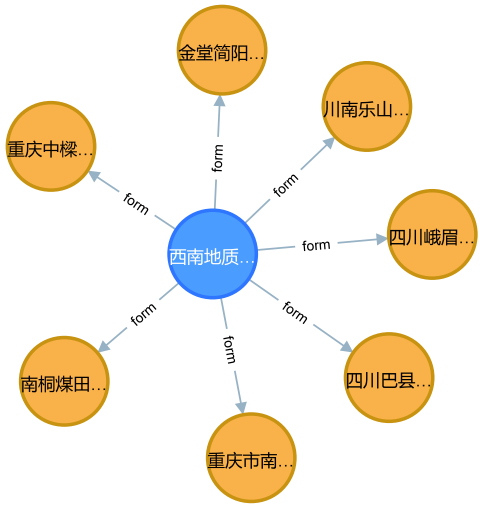

图 5 地质资料知识图谱的可视化

参考文献

1 陈婷玉. 推进新时代地质成果资料信息有效服务. 兰台世界, 2019, (6): 105–108. [doi: 10.16565/j.cnki.1006-7744.2019.06.30]

2 杨佳琦. 基于中文自然语言处理的糖尿病知识图谱构建 [硕士学位论文]. 包头: 内蒙古科技大学, 2020. [doi: 10.27724/d.cnki.gnmgk.2020.000567]

3 刘燕, 贾志杰, 闫利华, 等. 知识图谱研究综述. 赤峰学院学报 (自然科学版), 2021, 37(4): 33–36. [doi: 10.13398/j.cnki.issn1673-260x.2021.04.008]

4 徐增林, 盛泳潘, 贺丽荣, 等. 知识图谱技术综述. 电子科技大学学报, 2016, 45(4): 589–606. [doi: 10.3969/j.issn.1001-0548.2016.04.012]

5 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述. 计算机研究与发展, 2016, 53(3): 582–600. [doi: 10.7544/issn1000-1239.2016.20148228]

6 Chen YZ, Kuang J, Cheng DW, *et al*. AgriKG: An agricultural knowledge graph and its applications. In: Li GL, Yang J, Gama J, *et al*., eds. Database Systems for Advanced Applications. Cham: Springer, 2019. 533–537. [doi: 10.1007/978-3-030-18590-9_81]

7 聂莉莉, 李传富, 许晓倩, 等. 人工智能在医学诊断知识图谱构建中的应用研究. 医学信息学杂志, 2018, 39(6): 7–12. [doi: 10.3969/j.issn.1673-6036.2018.06.002]

8 黄恒琪, 于娟, 廖晓, 等. 知识图谱研究综述. 计算机系统应用, 2019, 28(6): 1–12. [doi: 10.15888/j.cnki.csa.006915]

9 李敏, 傅洁, 陈安蜀, 等. 面向知识服务的地质资料管理转型研究. 地质与资源, 2021, 30(1): 92–98. [doi: 10.13686/j.cnki.dzyzy.2021.01.012]

10 李家瑞, 李华昱, 闫阳. 面向多源异质数据源的学科知识图谱构建方法. 计算机系统应用, 2021, 30(10): 59–67. [doi: 10.15888/j.cnki.csa.008218]

11 焦凯楠, 李欣, 朱容辰. 中文领域命名实体识别综述. 计算机工程与应用, 2021, 57(16): 1–15. [doi: 10.3778/j.issn.1002-8331.2103-0127]

12 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 4171–4186.

13 许鑫, 岳金钊, 赵锦鹏, 等. 小麦品种知识图谱构建与可视化研究. 计算机系统应用, 2021, 30(6): 286–292. [doi: 10.15888/j.cnki.csa.007986]

14 宋伟, 张游杰. 基于环境信息融合的知识图谱构建方法. 计算机系统应用, 2020, 29(6): 121–125. [doi: 10.15888/j.cnki.csa.007424]

15 赵京胜, 肖娜, 高翔. 基于自然语言处理的能源领域知识图谱. 信息技术与信息化, 2018, (5): 55–58. [doi: 10.3969/j.issn.1672-9528.2018.05.014]

16 邓小政. 面向法律文书知识图谱构建研究 [硕士学位论文]. 南昌: 江西财经大学, 2021.

17 覃晓, 廖兆琪, 施宇, 等. 知识图谱技术进展及展望. 广西科学院学报, 2020, 36(3): 242–251. [doi: 10.13657/j.cnki.gxkxyxb.20201027.009]

18 修晓蕾, 吴思竹, 崔佳伟, 等. 医学知识图谱构建研究进展. 中华医学图书情报杂志, 2018, 27(10): 33–39. [doi: 10.3969/j.issn.1671-3982.2018.10.006]

19 朱木易洁, 鲍秉坤, 徐常胜. 知识图谱发展与构建的研究进展. 南京信息工程大学学报, 2017, 9(6): 575–582. [doi: 10.13878/j.cnki.jnuist.2017.06.002]

20 侯梦薇, 卫荣, 陆亮, 等. 知识图谱研究综述及其在医疗领域的应用. 计算机研究与发展, 2018, 55(12): 2587–2599. [doi: 10.7544/issn1000-1239.2018.20180623]

21 Opdahl AL. Knowledge graphs and natural-language processing. Big Data in Emergency Management: Exploitation Techniques for Social and Mobile Data. Cham: Springer, 2020. 75–91. [doi: 10.1007/978-3-030-48099-8_4]

(校对责编: 牛欣悦)