核电站建安调试文本语料抽取与语义图谱构建技术研究

雷玮剑[1] 单玉忠[1] 王 理[2] 司恒远[1] 汪 鹏[1]
LEI Weijian SHAN Yuzhong WANG Li SI Hengyuan WANG Peng

摘 要 为提升核电厂建安调试阶段 DEN、NCR、FCR、CR、UES、经验反馈等文本类知识语料的沉淀、检索与利用效率，设计了一种基于 Word2vec、fast-TEXT、GLOVE、Word2vec+Bi-GRU、BERT-BiGRU-Attention-CRF 等算法的上下文语义解析、标签抽取生成、关联关系计算与聚分类的方法，收集企业内真实数据，在实验室验证环境下基于开源工具，设计并搭建前述经验变更数据的标注、图谱展示、遍历推理等交互模块界面。实验结果表明，提出的技术方法及功能组件相比传统方法提高了解析识别的精准度，在核电站建安、调试业务领域达到可接受的可靠性水平，克服了传统方法在核电领域的局限性。

关键词 核电站；建安调试；非结构化；文本语料；抽取解析；语义图谱

doi：10.3969/j.issn.1672-9528.2024.01.001

0 引言

作为规模最大发展中国家、能源生产与利用主要经济体，我国能源规划发展节奏始终与国民经济社会发展紧密相关。面对能源结构优化、能源供给侧结构深入调整、我国乃至全球电力需求日趋放缓、“碳中和”与“碳达峰”政策逐步提上日程等多重复杂环境因素。核电作为一种安全、可靠、经济、高效、低碳的能源利用方式，作为国家战略产业，因其在社会绿色转型中的作用与价值，再度回归至政府、行业及公众关注视角。在技术进步的前提下，核电作为基荷电力的地位逐渐得到巩固与强化。据统计，截至 2023 年二季度，我国核电机组在运机组 55 台、装机容量位列全球第二，核电在建机组 23 台、装机容量当前位列全球第一，规划至 2030 年我国核电机组装机容量达 1.4 亿 kWh。综上所述，我国核电产业链迎来新一轮快速发展机遇。

核电站内构筑物厂房结构复杂、系统设备种类与数量多。一个典型成熟的两台百万千瓦级机组的压水堆核电站，平均包括约 400 余个工艺系统、48 000 余台套设备与备品备件，各类动力电缆、控制电缆共计 40 000 余根，长度近 5000 km。所示，核电站属于典型的大型复杂产品，相较于其他工业产品，核电站本质安全要求苛刻。核电站工程项目，具有安全质量要求高、建设标准高、劳力密集、资金密集、技术密集、周期长、资源投入多、上下游接口关系众多、多专业多单位交叉融合等典型特征，全周期全过程涉及数十个专业、上万个上下游工作接口、10 万余作业。

核电站工程全周期中，建安调试阶段自 FCD（核岛第一罐混凝土浇筑）至“168 小时试运行”结束，贯穿核电工程现场主要依据设计图纸、技术规范、技术标准、施工方案、测试大纲与测试程序等，完成土建物料与设备物资的实物构造集成、单体调试、单系统调试、联合专项试验，向运营单位交付功能齐全完整、可发电创造价值的核电站实体，典型工艺流程见图 1 所示。

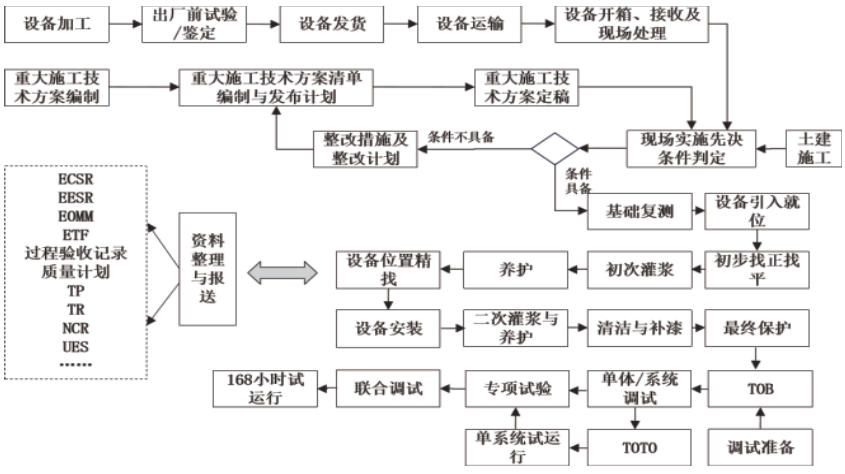

图 1 核电站建安调试阶段典型工艺流程

持续工期最长、平均约 54~60 个月，大量并行或交叉作业、作业种类繁多、从业人员流动率高、大量人员在固定物理空间内交织，进度安全质量目标刚性，高峰期现场平均有 1 万

1. 深圳中广核工程设计有限公司 广东深圳 518100
2. 中广核工程有限公司 广东深圳 518124
[基金项目] 国家重点研发计划（2020YFB1711700）

名以上从业人员上岗作业，是建安调试有别于其他阶段的显著特征。因此，核电站建设阶段安全质量很大程度上决定核电站本质安全性、可靠性，随着“华龙一号”等三代核电技术的逐渐成熟与批量化建设，国内核电工程产业链相继采用模块化建造、先进建造、智能调试、开顶法施工等多重技术，以优化核电工程建设生产工艺，提升核电站的安全性、经济性、可建造性与可运维性。

1 研究背景与技术进展

1.1 研究背景

核电行业的知识语料种类较多，例如操作手册、程序制度、经验反馈、图纸文件、法规标准、设计变更、不符合项、意外事件单等，多个知识语料库间切换查询，反向加重从业人员的知识获取难度。大量的设计变更、CR、FCR、NCR、UES，成为核电工程项目现场施工、调试生产组织调度诸多困扰不便的技术源头。非结构化文本存在数据沉淀流转应用不顺畅、安全质量控制代价高、建造生产组织执行弱于策划等典型问题。核电建安调试阶段非结构化文本知识语料呈现分散、多源、异构、庞杂特点，其中蕴藏着大量核电厂建设所需的规律、知识与经验，对单一核电工程建设、后续核电工程参考借鉴具有重要的指导借鉴价值。然而，受制于“碎片化异构化存储、沉淀流转处理与应用方式传统”等客观因素，核电建安调试阶段非结构化文本语料难以有效支撑单个核电工程项目、核电工程项目集群的推进与开展。

1.1.1 非结构化文本沉淀利用上下游理解不尽一致

上游专业填写编审人、下游专业使用人、各类非结构化文本不同层级管理人员、基于 PDCA 模式的行动处理人员，因在工作经验、知识阅历、技能积累、评判依据等因素的差异，各方对同一非结构化文本数据的技术理解、要素取舍、属性字段选定方面不尽一致，例如针对同一条经验反馈数据的等级定级（作业级、业务部门级、公司级）、针对同一条 NCR 是否导致经验反馈，各方理解视角与评判尺度不一，最终等级定级、核心属性字段结果不同。

1.1.2 不同数据内容间缺乏拓扑映射关系

在“两化融合”的背景形势下，核电工程建设业务信息系统的建设与升级，各类业务数据、非结构化文本数据分散于产业链的不同参与单位，相同、相近的技术问题及文本语料会以不同界面形式重复出现于不同信息系统，且部分文本语料数据品质偏低。同时，上下游的各条数据内容间的上下游援引（例如经验反馈、偏差、事件、设计变更、不符合项、现场施工变更、澄清之间）的来源致使参考、支撑说明、影响后果类关系缺乏，主要靠专业技术人员的责任心、专业能动性进行维护构建，各条具体非结构化数据间拓扑映射关系的完整性、准确性，缺乏专业技术角度的审视与判定。

1.1.3 各方技术管理人员被稀释

受到核电项目核准连续数年中止、近年来国内核电项目核准开工节奏逐步回速、国内劳动力人口渐转老龄化、核电劳动力缺口等因素影响制约，国内核电工程产业链各方的专业技术、管理人员被稀释，导致从业队伍梯队连续性弱、专业化水平、技术素养参差不齐。同时，在以人工为主的沉淀检索利用的传统模式下，经验丰富的核电施工调试专业技术人员的稀释、新堆型、新建造工艺的引入等因素，进一步增加了核电站建造过程中知识等非结构化文本的获取、理解与应用难度。

1.1.4 非结构化文本沉淀共享利用方式传统

不同核电工程项目间、同一核电工程不同部门不同专业间、核电站建设与运维阶段间的技术问题、经验反馈等非结构化文本语料的沉淀利用交流，仍以业务人员人工方式、定期或不定期沟通交流为主，实际经验反馈传递共享、技术问题重发预防的效果，很大程度上依赖于涉及人员的经验丰富程度、责任心、厂址差异性、技术优化改进等，对负责非文本语料数据填写、审批人员素质有一定要求，核电工程产业链各方花费数量可观的成本代价确保“经验反馈的漏报、少报”。同时，文本语料长文本内容会出现未严格遵循事实、放大外因、避重就轻等问题，机组、厂房、标高、专业、房间、区域、系统、设备等结构化属性字段选取不规范不完整。

1.1.5 文本内部存储且术语专业性强

核电建安调试过程中技术密度更高的非结构文本数据（例如质量计划、问题跟踪单、经验反馈、FCR、UES、NCR 等），存储在产业链内各类数据库中，由于涉及核安全与国家安全，难以通过公开数据集方式。因此，开源社区中提供的命名实体识别算法（named entity recognition，NER），对核电建安调试领域文本数据抽取任务的适配度偏低。作为一个横跨多个专业学科的行业，核电各类技术文档、业务数据中包含着大量专业术语、缩略语、代码、代字等，互联网侧通用开源算法模型难以快速获取构建合理、全面的实体特征。

因此，本文以核电站施工建设阶段中产业链企业内显性知识内容源为基本数据源，进一步利用 NLP、语义算法模型，构建索引解析算法模型体系，进行语义分析、语义提取，对不符合项等非结构化文本数据中表单界面的主数据、元数据、标题等属性字段和附件中段落文字进行分词，形成核电建造领域命名实体、属性以及关系组合库。

1.2 相关理论与技术

知识图谱为知识及其语义关系结构化、向量化、矢量化

表达，通过符号描述物理世界中概念内涵及其关系。谷歌于2012年在多年积累基础上提出知识图谱，旨在完善其搜索引擎性能。2017年后，随着机器学习、深度学习、NLP等爆发式进步、涌现式应用，知识图谱相关技术与垂直应用取得长足进展。语义网络、三元组等则成为基于不同信息技术手段的知识图谱表示方式[1]。知识图谱构建，一般包括Schema设计、事件抽取、实体抽取、关系抽取、融合与消岐计算、知识存储等步骤，实现术语语义网、自然语言处理、深度学习、机器学习、图数据库、搜索引擎等技术的交叉融合，包括面向结构化数据的规则映射、面向非结构化文本数据的抽取解析，基于依存句法、情感分析、上下文语境语义关系识别计算，获取命名实体及其间语义[2]。

国内外多个专家学者、技术人员对知识图谱、自然语理解、自然语处理及相关细分技术深入研究，譬如针对短文本内容信息有限、语义模糊等问题，丁辰晖等人综合编码器-解码器模型对概念集、短文本进行编码，通过Attention机制计算获取每个概念权重值，使用Bi-GRU（双向门控循环单元）对短文本输入序列编码，短文本分类准确率均值约59.21%[3]。现有研究主要侧重于探索各种方法以构建各行业领域的知识图谱。然而，此类方法中[4-5]多数依赖于结构化或半结构化的数据源（例如数据库、表格或本体等），难以处理非结构化的文本数据（例如文档、报告或案例反馈等）。结合双层LSTM，基于语言模型上下文相关的词向量表示（embedding from language model，ELMo）由Peters提出，ELMo词向量在语义表征刻画角度更为深层详细，但在输入语句序列偏长情形下、该模型自身学习能力下降较明显[6]。常见的知识图谱应用诸如基于图谱的问答、推理、搜索与推荐，而王文广将知识检索、知识语义探索归纳为“知识图谱简单运用”，知识计算、知识推理与辅助决策总结为“知识图谱复杂运用”[7]。垂直行业实践层面，电网相较于能源领域其他细分行业走在前列，蒲天骄等人[8]提出基于NoDKG（not only domain-specific knowledge graph）的电力领域知识图谱应用架构设计，按本体、图谱相结合方式，围绕电力业务数据构建了电力领域输配电侧知识图谱构建，并在电力调度故障处理、运检工单处理、电力客服智能问答领域实践验证。另外，电网领域实现以知识图谱节点关系为语义基础，设备维修记录文本与知识图谱实体节点的关联及基于此的检索利用[9]。曲朝阳等人针对电网企业输变电知识内容，采用协同过滤推荐算法，刻画不同用户的工作行为特征，输出推荐结果排序列表，通过web、移动终端方式传递给内部用户[10]。江秀臣等人围绕电力企业核电设备领域实时状态数据，从设备部件类别、参数、生命周期维度，构建数据挖掘与分析模型，辅助开展对应设备状态监测与故障诊断[11]。基于图谱与数据解析处理的搜索方面，楼凤丹等人[12]通过运用云计算技术，构建面向电力企业集中式数据中台的全文检索索引，支撑企业内不同数据类型、不同业务场景的统一检索。

但是，很少有研究专注于核电这一特定领域。随着深度学习、机器学习、模式识别、NLP、大语言模型的技术涌现与发展，如何运用前述技术，实现核电站建设过程中非文本数据的便捷、高效的沉淀与利用，已成为业内需要解决的难题与机遇。为此，本文对比分析核电工程产业链在DEN、NCR、FCR、CR、UES、质量计划、经验反馈、安全隐患、质量隐患等小样本非结构化文本语料获取、审批、应用过程中面临的现状问题，提出了一种基于NER、SVM、图谱推理技术构建核电站建安调试领域非结构化文本的识别、知识图谱构建与推理应用的技术方法，能够有效地从核电领域的结构化、非结构化文本数据中提取和组织知识，克服现有研究的局限性，实现核电建安调试领域等非结构化文本语料的语义化、向量化表达，提供一种更有效、便捷、可靠的核电建安调试领域文本知识发现与复用的方式。通过实例验证，该技术方法在核电建安调试领域具有一定的可靠性，为该领域知识依托NLP、大模型等技术的自动化、便捷化沉淀应用，提供了技术手段与可借鉴实例。

2 核电建安调试文本知识语料图谱构建

本文面向核电建安调试中设计变更、制造类NCR、安装类NCR、经验反馈、UES、安全隐患、质量隐患等语料数据，基于NLP、NER等技术按不同样本集标注规模情况、运用不同模型算法对比将变更经验非结构化数据转为结构化数据；运用Neo4j图数据库构建核电设备变更经验知识图谱局部实例，为核电从业者基于主题知识图谱进行检索，利用应用该类知识语料提供更便捷有效的技术手段，主要构建流程如图2所示。

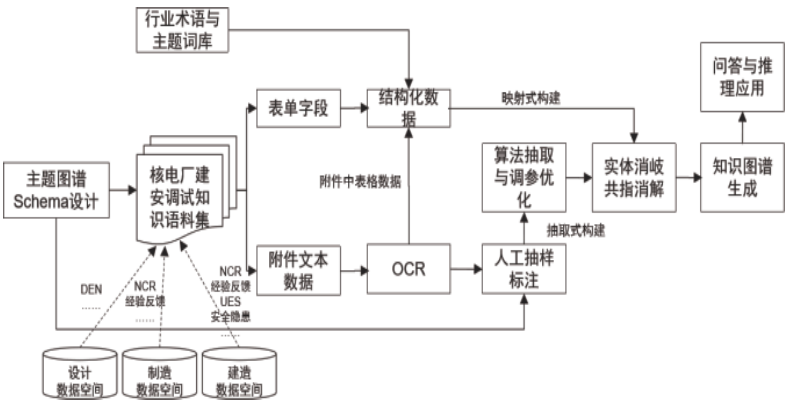

图2 核电建安调试知识语料图谱构建流程

2.1 建安调试领域知识图谱 Schema 设计

核电建安调试知识图谱 Schema 设计，主要从业务经验角度凝练形成核电建安调试领域术语组织的层次、关系的宏观图貌，设计术语主题类别、术语层次组织关系、术语间关系类别、属性字段等；设计核电建安知识语料中相关实体的词性、关系类别；词性包括动词、时段、时刻、组织、姓名、数词、量词、方位词、连接词、形容词、副词、代词、标点符号等。对领域实体分为 9 类，具体包括“专有术语”“成果物”“业务活动”“组织”“位置”“核电站实体对象”“指标参数”“问题现象”“原因”等，各命名实体 BIO 及编号设计如表 1 所示。

表 1 评价信息抽取模型

编号	实体类型	实体开始	实体内部与结尾
0	非实体	—	—
1	专有术语类	B-NOU	I-NOU
2	成果物类	B-DEL	I-DEL
3	业务活动类	B-ACT	I-ACT
4	组织类	B-ORG	I-ORG
5	位置类	B-LOC	I-LOC
6	核电站实体对象类	B-SSC	I-SSC
7	指标参数类	B-PAR	I-PAR
8	问题现象类	B-PRO	I-PRO
9	原因类	B-REN	I-REN
A	日期时间类	B-TIM	I-TIM
B	措施手段类	B-MET	I-MET

命名实体关系类别考虑上位、下位、同位、同义、组成、子类、包含、整体、位于、具象、自反、传递、对称、因果、作用、产生、从属、主体、使能。例如“ECS 终端循环回路中 ECS001PO 卡涩”采用主谓宾语义（subject-predication-object，SPO）三元组方式表示，可转化为 {[ECS001PO，属于，泵]，[ECS001PO，位于，终端循环回路]，[终端循环回路，从属，ECS]，[ECS001PO，产生，卡涩]} 等实体及其关系，进一步转化为实体节点 $\{n_1, n_2, n_3, \cdots, n_n\}$ 和实体关系 $\{r_{12}, r_{13}, \cdots, r_{n\text{-}1n}\}$ 等数学表示。

2.2 建安调试领域知识语料收集与预处理

数据收集是知识图谱构建应用的基础性导入工作，数据品质很大程度上决定了知识图谱的品质与应用效果。由于本文研究内容面向核电厂建安调试过程中具体业务场景，所以以核电设计建造领域的国家标准、行业标准、术语标准、行业词典、行业内缩略语库、核电行业主题词表、国防科技名词大典（核能）中核电相关术语作为命名实体的起始数据，整体导入图数据库。本文所提及的核电建安调试文本语料数据，主要包括设计变更、设备制造类不符合项、设备供货类不符合项、设备安装类不符合项、经验反馈等、安全隐患、质量隐患、设计要求澄清、调试意外事件单等，文本类知识语料数据收集，主要从国内某核电公司内设计变更模块、经验反馈系统、设备监造平台、智慧工地、施工管理系统下质量管理模块、调试管理系统、核电价值链协同平台中设计数据空间、制造数据空间、建造数据空间中导出、获取或下载获取原始数据，主要包括对各类变更经验的表单数据、表单附件进行获取，原始数据体量共计 423 222 条，各类数据收集到的数量规模如表 2 所示，表单数据汇合后按宽表、子表形式存储体现。

表 2 核电建安调试各知识语料收集规模

知识语料类别	核电厂 A（在建）	核电厂 B（在建）
设计变更 DEN	39 467	10 557
设计要求澄清 CR	84 092	23 143
不符合项报告 NCR	38 415	13 031
现场变更申请 FCR	136 042	51 146
意外事件单 UES	7 444	4 420
安全隐患	3189	5189
质量隐患	1851	1440
质量事件	266	72
经验反馈	3458	534

表单附件则运用 paddle OCR 工具识别转换为可编辑文字、表单原始附件通过 FTP 路径存储，在软件工具对文字内容初核基础上，由具备核电基本业务经验知识的工程师进行复核修正，附件中的表格数据自同步、补全至结构化数据。

2.3 结构化数据映射规则标记

按照设备变更经验语料类别，按照共性字段、专有字段整理识别结构化数据字段及其规则关系，结合核电内部业务逻辑，以表头实体字段为单元明确实体类别（见图 3），明确实体类别字段与属性字段间的关系、实体类别与实体类别的关系，定义其数据源头及获取路径，配置设定其数据抓取频次。

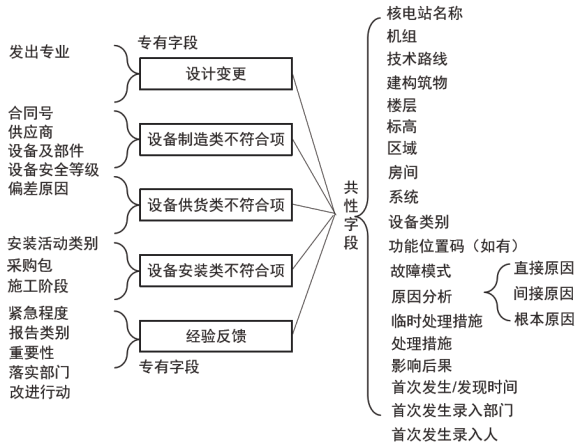

图 3 结构化数据表头字段样例

2.4 非结构化数据标注与抽取识别

对于非结构化数据，各类变更经验的原始存储于 FTP 路径，运用 paddle OCR 工具识别转换为可编辑文字，由具备核电基础知识的工程师对 OCR 转化后的文字品质进行抽查修订。分别按随机抽取总数据量的 10%、20% 作为样本数量集，先后进行两批次相互独立的分词、词性标注，运用部署在企业局域网中的 Word2vec 算法、GLOVE 算法、Word2vec+Bi-GRU、多维度语义提取算法（BERT-BiGRU-Attention-CRF），识别获取各类变更经验语料中实体、属性及边等关系，其中基于 BERT-BiGRU-Attention-CRF 算法技术路径如图 4 所示。

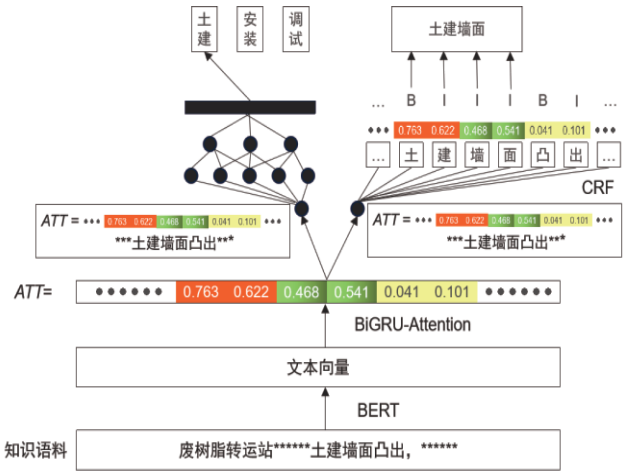

图 4 多维度语义提取模型

本文通过 BIO 标注方式（见图 5），以 10%、20% 的比例随机选取样本数据集进行标注，分别获取标记后数据 42 322 条、84 644 条。

图 5 核电厂建安调试文本语料实体标注示例

采用基于 BiLSTM 的字级处理器，提取字的信息、字与字间的语义关系。句子级处理器，以整句为单元提取信息，获取不同语境下的命名实体张量。同时，基于 self-attention 模型识别提取句子中不同实体间依赖关系，计算实体与实体间的关系权重，获取实体间位置领近关系、句子内部结构与语义信息，GCN 中 entity graph model 处理内部实体特征、adjacency graph model 处理实体间关系。按以下条件规则，通过依存句法方式识别提取实体对及其边关系，具体包括 nn- 名词组合、tmod- 时间修饰、punct- 标点修饰、nsubj- 主谓、prep- 介词修饰、pobj- 介宾关系、root- 核心关系、amod- 形容修饰、dobj- 直接宾语。最终通过三元组形式，将实体间、实体属性间关系存储于图数据库，如图 6 所示。

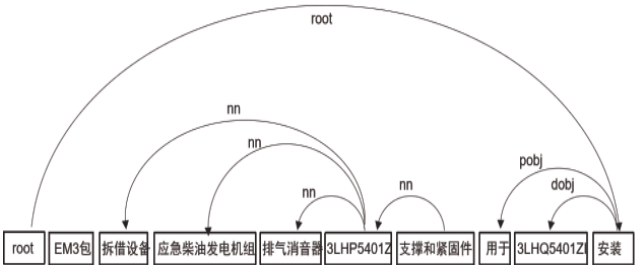

图 6 依存句法边关系提取原则

随机划分 80% 的语料作为训练集、10% 作为测试集、10% 作为开发集，最终，采用实体识别、语义提取领域常用的评价指标及公式对任务精确度、算法模型均衡性进行评价为：

$$P=\frac{T_P+T_N}{T_P+T_N+F_P+F_N} \quad (1)$$

$$R=\frac{T_P}{T_P+F_P} \quad (2)$$

$$F_1=\frac{2PR}{P+R} \quad (3)$$

式中：T_P 为识别完整正确，T_N 为本不应识别确未识别出，F_P 为模型错误判定为实体，F_N 为本应识别但未识别的实体。不同标注比例下各抽取模型结果分见表 3、表 4 所示。之后，对抽取识别后的实体进行指代消岐检查。实证结果表明，BERT-BiGRU-Attention-CRF 的组合模型抽取结果优于其他模型，且标注比例越高、抽取结果质量越好。

表 3 10% 标注比例设备变更经验语义抽取结果

算法模型	准确率 P/%	召回率 R/%	F_1 值 /%
Word2vec	81.328	79.037	80.505
fast-TEXT	81.053	76.025	78.459
GLOVE	87.052	83.026	84.991
Word2vec+Bi-GRU	89.046	86.013	87.503
BERT-BiGRU-Attention-CRF	95.075	91.072	93.030

表 4 20% 标注比例设备变更经验语义抽取结果

算法模型	准确率 P/%	召回率 R/%	F_1 值 /%
Word2vec	85.039	81.023	82.982
fast-TEXT	83.039	79.089	81.016
GLOVE	86.098	85.167	85.069
Word2vec+Bi-GRU	91.085	88.031	89.532
BERT-BiGRU-Attention-CRF	97.042	95.363	96.195

2.5 建安调试知识图谱生成与可视化

在 Neo4j 图形数据库存储抽取解析后的设备变更经验语义数据基础上，在 MyBatis 框架下进行 SQL 语句整合，执行 Neo4j 中 Graph 模块即可生成主题知识图谱，示例效果如图 7 所示。

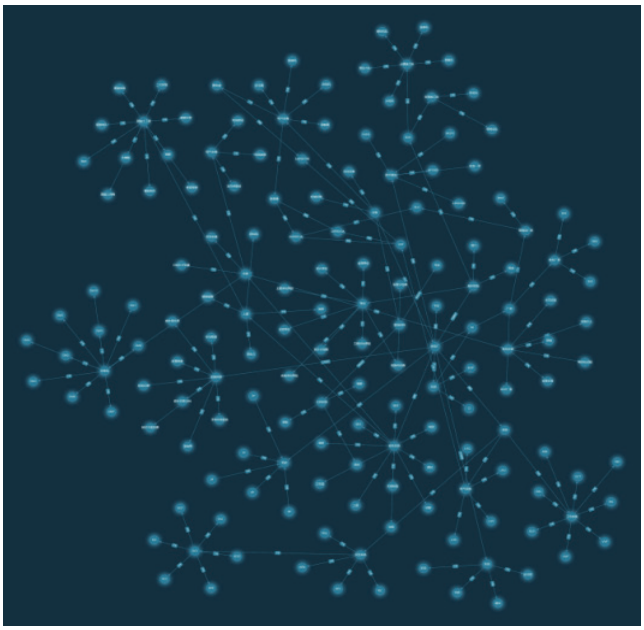

图 7 建安调试知识语料图谱实例

基于领域知识语义图谱，运用 FCM 聚分类算法，对建安、调试领域知识语料中获取的原因类、措施手段命名实体进行统计分析，部分典型结果如表 5 所示。

表 5 事件原因及措施手段类聚分类分析结果

阶段	征兆现象	原因分析及频次	措施手段及频次
建安	设备（或附件）无法安装 / 安装错误	土建基础公差超允许公差 [11]；不符合图纸要求 [3]；预埋件缺失 [6]；预埋件、预留孔洞位置偏差或漏项 [14]；设备供货滞后 [22]；设计设备基础荷载、实际采购设备基础荷载要求不符 [3]。	连接结构调整 [20]；土建基础返工修正 [8]；部分部件焊接形式改为现场焊 [6]；设备供货计划与安装工艺相适配 [22]；配筋设计修改 [3]。
建安	墙体 / 底板裂纹、起沙、空鼓、麻面、夹层、露筋或蜂窝。	混凝土配比不当 [3]；混凝土搅拌不均匀、和易性弱、搅拌时间不够 [5]；插筋过密 [1]；墙板间结构复杂（弧形等）[1]；温度应力 [2]；降雨或降温 [3]；混凝土入模温度高 [1]；模板缝隙未堵严实 [4]。	调整坍落度等搅拌参数 [8]；从设计源头优化插筋数量与布置 [2]；埋设温度应变传感器以监控温度和应力 [6]；增设混凝土养护措施，例如增设养护棚 [2]；搅拌机配置液冷、风冷装置 [2]。
建安	漆面破损	吊运安装中，磕碰坚硬物品 [3]。	补漆，定期保养 [3]。
建安	阀门自带线芯接反	设计图纸未明确线芯、端子对应关系 [3]；施工方工作惯性，未参照厂家图纸 [4]。	安排经验丰富的设计人员 [3]；关于阀门自带线端接，施工方应参考厂家图纸 [4]。

表 5（续）

阶段	征兆现象	原因分析及频次	措施手段及频次
建安	管道类，清洁度、坡度不符设计要求、预制不合格，支吊架固定间距超差。	未按施工方案或图纸要求作业 [2]；施工人员技术要求理解不充分 [3]。	强化从业人员质量意识 [3]；加强上岗前及作业中的技能培训与辅助提醒 [2]。
建安	预埋件位置偏差（中心轴线移位、标高偏差等）	不符合预埋件布置图要求 [2]；土建误差 [3]；测量放线偏差 [2]。	按规定开展作业返工 [4]；设备固定，据实、按需拆分为埋件、厂房模块 [3]。
调试	拉杆机构卡涩	设备运抵现场时间太早 [2]；润滑油脂失效 [5]。	优化匹配设备供货计划 [2]；加强相关部件润滑保养 [5]。
调试	阀门内漏	阀瓣变形 [3]；选型出错 [1]；阀门安装方向设计错误 [1]。	更换阀门或阀瓣 [4]；升版设计文件，加强设计图纸审核 [1].
调试	系统功能异常	设备部件安装不完整 [1]；设备拒开或拒动作 [3]；部件安装位置高且隐蔽 [1]；工艺流程图纸未体现该部件 [1]。	升版工艺流程图纸 [1]；加强设计图纸的校核 [3]；优化安装或调试程序文件 [3]。

3 结果应用

专业技术人员通过运用本文所构建的核电建安调试知识图谱，可直观查询电厂对象、业务活动、组织专业、空间位置、原因、措施等各类命名实体的语义关系，进而对核电建安调试阶段各个专业技术领域、全周期业务、变更经验数据演变关系具有整体认知，通过实体节点与界面浮窗方式，查阅与该实体节点关联度、依据杰卡德系数（Jaccard Index）自高往低排序的抽取后知识内容。在压水堆核电工程建安、调试业务过程中，基于知识图谱、搜索引擎排序机制算法（Boolean Model、Topic Sensitive PageRank、Intelligent Surfer Model、Backward Forward Step 等）的结合，实现核电建安、调试中设计变更、制造 NCR、施工 NCR、经验反馈的主动推送与查阅获取。

借助知识图谱的呈现载体、专业技术人员可便捷遍历实体节点之间的关系路径、览阅最短关系路径，可按需查阅每台设备、每个部件、介质的上下游工艺关系，加载呈现每条知识语料内容间的上下游引用参考关系。当核电工程项目出现以前未曾发生过的异常情况或后续批量化核电工程中面临新的技术改进项目，前期已积累的知识经验、变更数据无法直接复用或相似参考时，知识图谱已呈现的工艺逻辑、语义

推理关系，可为专业技术人员开展原因定位排查、技术要素分析提供“地图导航式服务”。另外，基于领域知识图谱对核电站三维模型的 tag 解析，计算形成各个模型或部件相关程度自高往低的知识语料排序，如图 8 所示，为人工查看读取形成便捷化、一站式界面。同时，通过分类模型、预测模型与知识图谱的结合，可实现设备经验变更知识中表单数据的建议生成，可减轻核电设备设计、制造、安装业务过程中，相关知识填写阶段的录入工作量，提高相关字段数据的规整性与准确性。

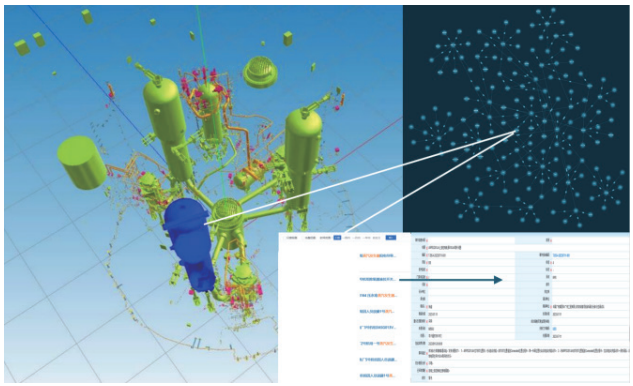

图 8 基于建安调试知识图谱的路径计算与内容推荐

4 总结与讨论

本文通过公开数据与核电工程产业链内部数据相结合，以核电工程产业链内部数据为主的方式，收集整理相关数据，数据种类覆盖面广，是真实、较全面的研究数据，分析核电厂施工调试阶段 FCR、NCR、CR、UES、经验反馈等文本类知识语料离散、孤岛、内容长短不一等问题。

完成领域图谱 Schema 设计，按不同标注样本集比例的条件下，通过 Word2vec、fast-TEXT、GLOVE、Word2vec +Bi-GRU、BERT-BiGRU-Attention-CRF 等算法模型对比及 Neo4j 图数据库结合，设计了建安调试文本语料自动化索引解析、主题聚集展示、去重重构的算法模型体系，构建生成核电建安调试知识语料语义图谱。最终，利用组织内语料文本，围绕关系路径推理查询、原始变更经验语料与实体节点关联匹配、知识表单字段自动预测填充、三维模型自动标签及语料数据计算等场景进行实例验证，验证其技术可信性与合理性。结果表明，基于领域主题知识图谱的构建与延展应用，可为核电行业知识发现与应用自动化形成新技术手段，解决传统模式下以专业技术人员为主进行知识沉淀与应用时的低效、完整性欠缺等问题。

然而，由于时间有限及技术成熟度因素，本文尚未联同 CV（计算机视觉）技术、声音识别技术实现知识图谱与多模态语料的关联解析与交互应用，尚未探索图文互搜，尚未结合大语言模型对知识图谱构建工作模式优化、更为自然的知识发现与内容组织进行验证，尚未通过智慧工地、智能监造等移动终端进行文本知识语料的规模化推送应用，尚未与工作流系统界面进行细粒度精准化刻画解析。未来下一步研究过程中，基于知识图谱的多模态知识语料融合分析、核电行业领域大语言模型与核电设备知识图谱构建双向支撑交互，核电知识自动化、知识挖掘与图谱推理计算必将成为主流方向与热点焦点，将通过方法手段研究支撑于核电设备变更、经验反馈便捷化沉淀应用。

参考文献：

[1] 赵琛，王一帆，李思颖，等 . 中国未来核电发展趋势与关键技术 [J]. 能源与节能 ,2020(11):46-49+67.

[2] 陈华钧 . 知识图谱导论 [M]. 北京：电子工业出版社 ,2021.

[3] 丁辰晖，夏鸿斌，刘渊 . 融合知识图谱与注意力机制的短文本分类模型 [J]. 计算机工程 ,2021,1(47):99.

[4] 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 [J]. 计算机研究与发展 ,2018,55(12):13.

[5] 郑泳智，朱定局，吴惠彝，等 . 知识图谱问答领域综述 [J]. 计算机系统应用 ,2022,31(4):1-13.

[6] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[J].Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies, 2018(1): 2227-2230.

[7] 王文广，知识图谱：认知智能理论与实战 [M]. 北京：电子工业出版社，2022.

[8] 蒲天骄，谈元鹏，彭国政，等 . 电力领域知识图谱的构建与应用 [J]. 电网技术 ,2021,45(6): 2080-2091.

[9] 王琼，魏军，闫润珍 . 知识图谱在智能电网的应用 [J]. 电子元器件与信息技术，2020,4(1):135-137,147.

[10] 曲朝阳，周宁，曲楠，等 . 基于知识关联度的电力大数据协同过滤推荐算法 [J]. 东北师范大学报 (自然科学版), 2018, 50(1):74-78.

[11] 江秀臣，盛戈皞 . 电力设备状态大数据分析的研究和应用 [J]. 高电压技术 ,2018,44(4):1041-1047.

[12] 楼凤丹，裴旭斌，王志强，等 . 基于云计算及大数据技术的电力搜索引擎技术研究 [J]. 电网与清洁能源，2016, 32(12): 86-92.

【作者简介】

雷玮剑（1988—），男，陕西渭南人，硕士研究生，高级工程师，研究方向：自然语处理、知识图谱及核电知识中心等。

（收稿日期：2023-09-19）