

[数据处理]

面向发电设备故障知识图谱构建的可视化人机交互工具

文/鲍日明¹ 李强¹ 赵旭波¹ 张震伟² 彭永乐³ 杨晓蓉⁴

1. 浙江浙能绍兴滨海热电有限责任公司；2. 浙江浙能数字科技有限公司；3. 杭州电子科技大学；4. 浙江财经大学

提出一个面向发电设备故障知识图谱的可视化人机交互工具。首先通过数据预处理将语料数据进行特征提取并可视化，采样获取代表性的语料数据进行手工标注，使用打标数据训练实体抽取模型，最后识别待标注的语料数据，通过迭代优化打标的方式来获取命名实体标记，从而提高实体抽取准确率，确保知识图谱的构建质量。

我国电厂在长期运行过程中产生了大规模的电厂运行日志数据，包括故障文本数据、规程数据和定值清册等多类型的数据。在电厂故障诊断领域，日志数据、规程数据和定值清册等数据中，存在着大量的故障现象、故障原因、故障处理方案等记录，并且由于故障事件难以识别，故障的类型错综复杂，电厂的设备类型多样并相互依赖，从故障数据中构建知识图谱具有很高的研究价值。此外，随着电厂的持续运行，会产生一些新的故障事件，导致原始语料数据数量庞大。在面对海量的多类型故障数据时，如何有效整合多类型数据，快速并准确地进行知识抽取、构建知识图谱、应用知识图谱并交互式探索仍是一个待解决问题。

图谱构建，需要通过知识抽取将原始的非结构化文本数据转换为结构化数据，经过知识融合与知识加工后，构建知识图谱网络。知识图谱对专业的电力设备故障诊断领域而言，能够建立较完善的知识库，从而更好地提升故障数据的搜索能力；满足对知识管理、语义检索和数据分析等方面的需求。在发电设备领域，通过发电设备数据的语言解析相关的技术，可以把发电设备知识如维修规程和工作手册等非结构化的数据与知识图谱有机融合，推进面向电力设备维护场景的故障智能搜索方面的应用。

知识抽取主要分为实体抽取、关系抽取两个步骤，知识抽取占据图谱构建工作量的70%，是图谱构建中最重要的一环，其准确率决定图谱构建的质量。在专业领域，知识抽取的准确率不高，而手工标注效率低下，费时费力，标注的准确率受限于标注工作人员的技术水平，受人的主观影响较大。针对专业领域的自动化、半自动化标注工具研究尚是空白，大都是完全依靠手工标注方法对专业领域的数据进行标注。因此，基于人机交互的方式构建设备故障知识图谱的研究与应用有着深远的现实意义。

本文提出一个面向发电设备故障知识图谱的可视化人机交互工具，通过对大规模故障日志数据进行语义分析、聚类后，采样得出代表性的语料数据，实现一个融合专家手工标注与模型自动标注的循环工作流。通过对发电设备故障信息进行半自动标注，以提高故障数据知识抽取的准确率和效率，提高故障知识图谱构建的质量，从而实现发电设备故障诊断。

相关工作

1. 发电设备故障知识图谱可视化

知识图谱可视化广泛应用于发电设备故障领

域，由于发电设备存在故障种类多样，故障事件难以识别，故障事件动态增长等问题，而构建知识图谱并可视化能有效作出故障诊断，提供故障解决方案。Yang等人^[1]利用CiteSpace工具对工业生产知识图谱进行可视化，节点大小映射实体重要性，边的粗细映射实体之间的相关性。谢庆等人^[2]基于ALBERT-BiLSTM-Attention模型构建电力变压器运维知识图谱，使用Neo4j图数据库进行可视化。叶欣智等人^[3]利用BERT-BiLSTM-CRF模型构建电网故障知识图谱，通过计算节点相似度并可视化判断知识图谱中的冗余或异常节点。但上述工作局限于知识图谱可视化，并未对文本的语义信息做更深的可视分析。

2. 可视化的人机交互工具

可视化的人机交互工具旨在将专家知识融入模型构建，提高模型构建的准确性与可靠性，提高专家对复杂数据洞察的能力。张龙龙等人^[4]设计基于人机交互技术的船舶实时导航Web系统，高效查询船舶的实时导航结果，精准完成船舶作业航迹导航。李瑞连等人^[5]基于ICS框架，采用可视化人机交互系统对脱硫系统指标实时展示，为实现智能脱硫系统提供依据。高海波等人^[6]基于多元图表示的基础理论，提出基于平行坐标图的可视化人机交互技术，进行脑电图记录信号特征提取技术的研究。Zhang等人^[7]提出了LabelVizier，结合领域知识和用户特定的需求，利用可视化人机交互技术，揭示了三种数据标注的错误类型，为大规模文本多标签数据集提供更准确的标签。上述工作使用人机交互工具实现数据标注，数据展示，模型优化等功能，但并未使用人机交互工具将领域知识与模型优化结合。

系统概述

本文设计了面向发电设备故障知识图谱构建的可视化人机交互工具，流程如图1所示。

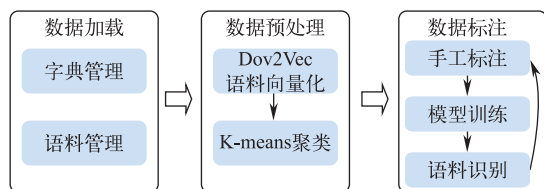


图1 系统流程图

首先，进入数据加载模块，数据加载模块中包含字典管理与语料管理。字典管理用于实体匹配，通过第三方知识库对实体标注进行修正与补充；语料管理用于存储多类型数据，提供OCR文件自动识别功能。其次，进入数据预处理模块，该模块通过文本表征与聚类算法对语料数据进行聚类，采样。在该模块中，首先通过Doc2vec将导入的语料数据投影到向量空间并可视化，其次利用K-means算法聚类语料数据，从聚类类别中均匀采样出具有代表性的语料数据。最后，进入数据标注模块，该模块是迭代式对语料数据标注，训练实体抽取模型，使模型达到较好的准确率。在手工标注环节对语料数据中的实体进行标注，然后将标注好的语料数据作为训练集输入BiLSTM-CRF命名实体识别模型训练，最后基于训练好的模型对待标注语料数据的实体进行预测，获取命名实体标记。再次进入手工标注环节，对模型预测的结果进行检验与修改，将修改后的结果再次训练实体识别模型。通过手工标注与模型训练，不断迭代优化模型识别的准确率，提高图谱构建的质量。

交互式知识图谱构建

1. 数据加载

(1) 字典管理

构建字典管理模块，用户可导入实体字典数据，用于实体抽取模型识别的字典匹配，若在实体字典数据中存在某个实体，而模型未能将该实体正确识别时将会被自动标注。用户可自定义字典的描述信息以及实体标注的快捷键，实体字典成功导入后，用户可查看和修改已经导入的数据，如图2所示。

(2) 语料管理

构建设备知识图谱的数据类型和格式多样，因此开发多类型数据导入功能是构建设备知识图谱的必要前提。系统提供了接口支持用户导入Word、Excel、TXT、PDF和图片等格式的语料数据，数据导入成功后，用户可对其进行查看、修改和导出操作。由于故障数据文件中的文字需要提取，OCR功能被集成到系统中，在导入文件时，系统进行文字自动识别。本研究中的数据格式类型主要为以往记录的故障处理文本数据、设备运转的规程数据、设备的层级

关系数据、设备的工作状态范围等，主要为非结构化和半结构化数据，其分类和格式类型见表1。

表1 电厂数据类型及格式汇总

数据类型	数据格式	数据描述
日志数据	Excel	日志数据，包括了数据的处理、操作等，由ERP系统导出
规程数据	PDF	机组集控运行规程、相关标准等
定值清册	Excel	设备的运行状态范围数据
设备树数据	Excel	设备之间的层级关系数据表
其他数据	文本、图片	操作人员记录在纸质材料上的数据，可通过扫描获得图片格式的电子数据；操作人员记录在文本工具中的数据，如Word、txt等格式类型

2. 数据预处理

完成字典和语料数据加载后，需要对语料数据进行预处理，分为两个步骤：数据聚类与数据采样。数据聚类步骤依据文本的语义信息进行聚类及可视化；数据采样步骤从聚类中均匀采样出具有代表性的语料数据。

用户在页面上侧两个输入框内分别输入向量维度 n 与聚类类数 k ，随后点击聚类按钮。系统使用Doc2Vec模型学习语料数据的语义信息，将语料数

据投影到向量空间，将每一条文本数据转换为长度为 n 的向量。然后，基于用户输入的聚类类数 k ，使用K-means算法将语料数据聚类为 k 个类别，随后利用T-SNE降维算法将 n 维向量的文本数据投影到2维向量，将2维文本数据使用散点图可视化，用颜色区分不同类别的文本数据，同时将文本数据分词的结果使用词云图可视化，如图3所示。

用户拖动“采样率”滑动条以选择采样率 x ，点击“采样”按钮，点击“执行打标”按钮。系统从聚类后的文本数据中依据采样率 x 从每个类别中均匀采样，目的在于获取具有代表性的文本数据，降低重复标注相似语料数据的可能性，减少手工标注数据耗费的人力物力，提高实体抽取模型的泛化能力。

3. 数据标注

手工标注环节中，用户需要对语料数据标注实体，构建实体抽取模型所需的训练数据。页面上侧的提示框提示用户手工标注的当前轮次，以及标注所有数据的进度，如图4a所示。5个实体类别标签出现在提示框下方，实体类别标签中含有实体类别以及对应的标注快捷键，用颜色区分实体类别标签，如图4b所示。在实体打标工作区中，用户通

☒	实体名称	实体描述	快捷键	标签颜色	包含词量 (个)	实体容量 (字)	操作
☒	设备	设备描述	E	<div></div>	5	40	编辑 导出 删除
☒	故障名称	故障	F	<div></div>	105	773	编辑 导出 删除
☒	故障现象	现象	P	<div></div>	81	4338	编辑 导出 删除
☒	故障原因	原因	R	<div></div>	49	4289	编辑 导出 删除
☒	处理方法	处理方法	T	<div></div>	164	33586	编辑 导出 删除

图2 实体字典

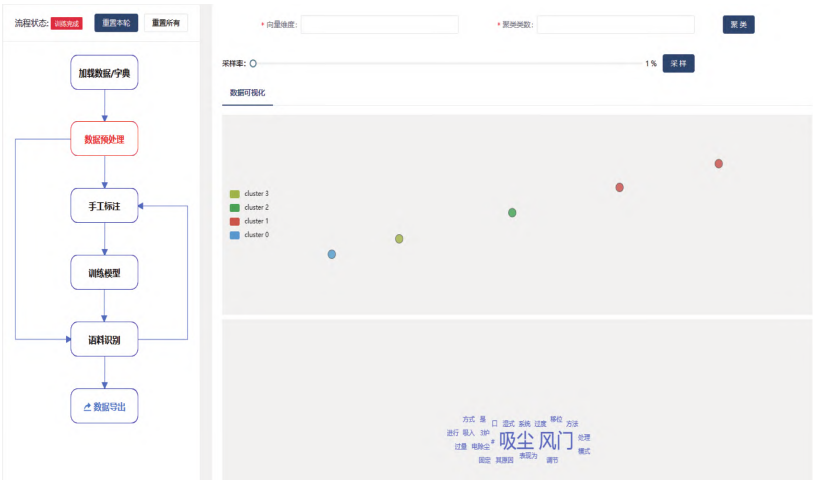


图3 数据预处理界面

过鼠标选中连续的文本，使用快捷键对该文本进行实体标注，标注文本的颜色与实体标签中的颜色一致，如图4c所示。页面下方实体打标结果中，将标注结果按照实体类别进行分类，若实体标注、实体匹配的结果有误，用户可在当前区域中删除该结果，重新标注，如图4d所示。待当前轮次的语料标注全部完成，进入模型训练环节。

4. 模型训练

模型训练环节将标注过的实体数据作为训练数据输入到实体识别模型中训练。用户通过下拉框选择实体抽取的模型BiLSTM-CRF；设置模型的训练数量，即模型训练的轮次；设置模型输入的维度；点击“训练”按钮。实体抽取模型将文本数据作为输入的特征，文本数据中的实体标记作为输入的类别标签。

BiLSTM网络层学习和提取语料文本中双向的上下文特征信息，获取语料文本中每个实体的各个标签概率值，再利用CRF层预测标签与标签之间的关系，对文本序列进行标注，最终获取最大概率的标注结果。

训练完成后，在模型训练页面左侧绘制训练时模型准确率随迭代次数变化的折线图，右侧是模型训练准确率随手工标注轮次变化的折线图，如图5所示。训练后的模型即可用于语料识别并进行新一轮次的语料数据实体标记。

5. 语料识别

语料识别环节将模型运用在未标注的文本数据上，进行实体识别。用户拖动“采样率”滑动条设置采样率 y ；通过下拉框选择BiLSTM-CRF模型；点击“识别”，如图6所示。根据采样率 y 从剩余未



图4 手工标注界面

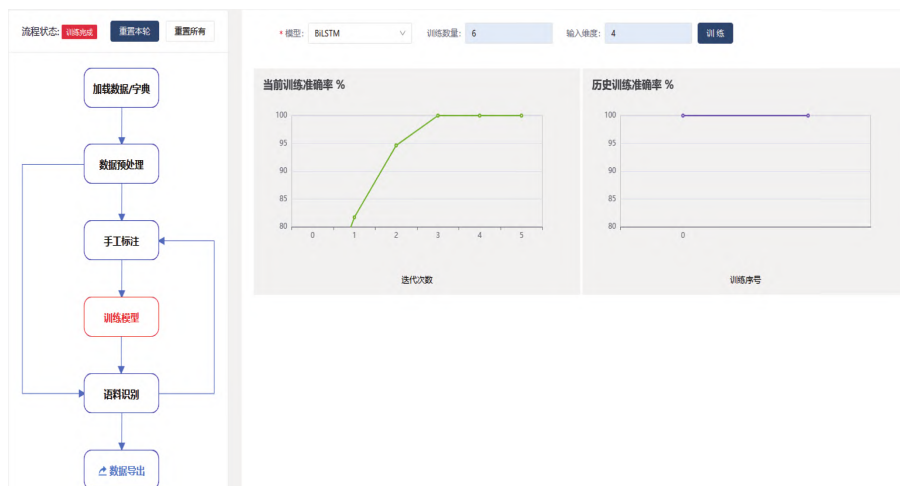


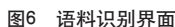
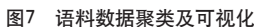
图5 训练模型界面

语料。

手工标注环节中,在实体标注工作区对文本数据进行标注并将其编注后的数据作为训练案例,实体标签包括设备、故障名称、故障现象、故障原因及处理方法,如图8所示。

本文设计文本标注工具主要面向电力领域的电力设备故障信息，并支撑电力设备知识图谱的构建。因此使用电厂部分实际设备故障日志数据对设计进行测试，即测试标注工具最终运行结果与预期效果是否相一致。

在文本标注测试模块的准备工作中，导入故障日志数据，在加载数据/字典界面查询已选中字典、



训练模型环节中，在训练模型子页面中设置模型为BiLSTM-CRF，“训练数量”为6和“输入维度”为4后，对BiLSTM-CRF实体识别模型进行训练。

训练完成后，页面绘制模型训练过程中的正确率折线图，如图9所示。

语料识别环节中，设置“采样率”为100%，以选取所有剩余的语料，点击“采样”按钮，点击“识别”按钮，识别未标注的语料数据，如图10所示，并使用字典匹配功能，将未识别正确的语句中的实体进行自动标注。

语料识别成功后系统将自动跳转到手工标注界面，以此能够实现迭代进行手工标注并训练模型的操作直至实体识别模型自动标注准确率达标。在所有数据标注完成后，系统可导出打标好的实体数据。

在后续图谱构建模块中，将该结构化的数据构建到故障知识图谱中。在测试中通过智能检索、故障事件搜索与方案处置，专家知识累计入库等形式能够对设备运行状态和异常情况作出判断，并对故障类型、故障现象及原因进行诊断。



图8 手工标注界面标注语料



图9 训练BiLSTM-CRF模型

进行图谱构建之前，在原有故障知识图谱上，如图11a所示，搜索“温度失准”的故障事件，未出现与之故障相关的事件，显示了与“温度偏差”相关的故障事件，点击其中一个故障，显示故障为“一期湿电碱液箱区域安全淋浴器管漏水”，解决方案为“淋浴器管道破损，更换淋浴器，现已正常”，如图11b所示。在故障知识图谱构建完成后，在知识图谱中查询“温度失准”的故障事件，标注工具最终运行结果与预期效果相一致，出现了“温度失准”事件的原因，现象和解决方案，故障

为“6#炉主蒸汽出口温度3失准（温度1:535、温度2:538及温度3:546）”，原因为“6#炉主蒸汽出口温度3拆回校验不合格”，解决方案为“已更换一支合格的温度元件。处理结果目前温度偏差为实际偏差”，如图11c所示，证明了该人机交互工具成功对文本数据结构化，提高了实体抽取的效果。实现了对发电领域数据资源的实体标注，将原有的半结构化数据或非结构化数据进行知识提取后构成知识图谱，对于电厂智能数据处理与应用具有极大的意义。

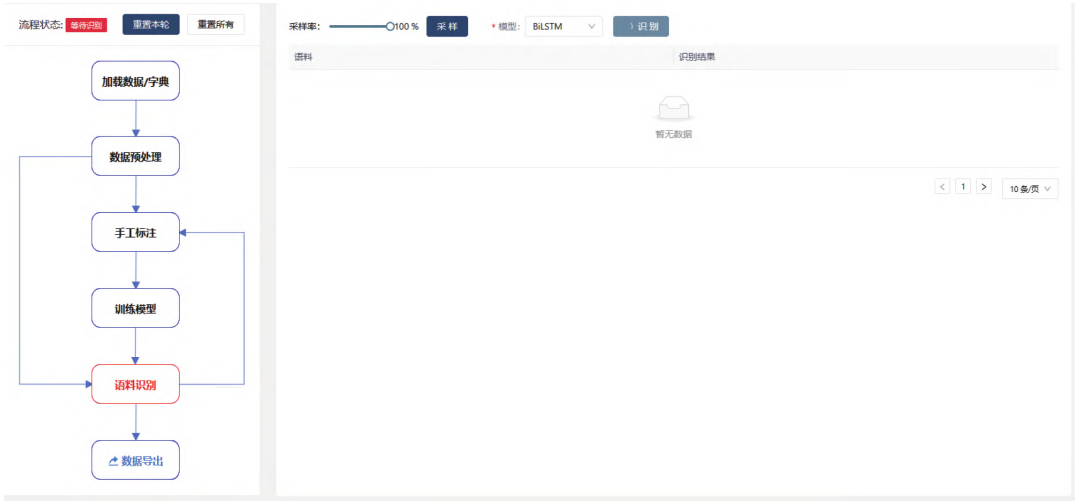


图10 模型识别剩余语料

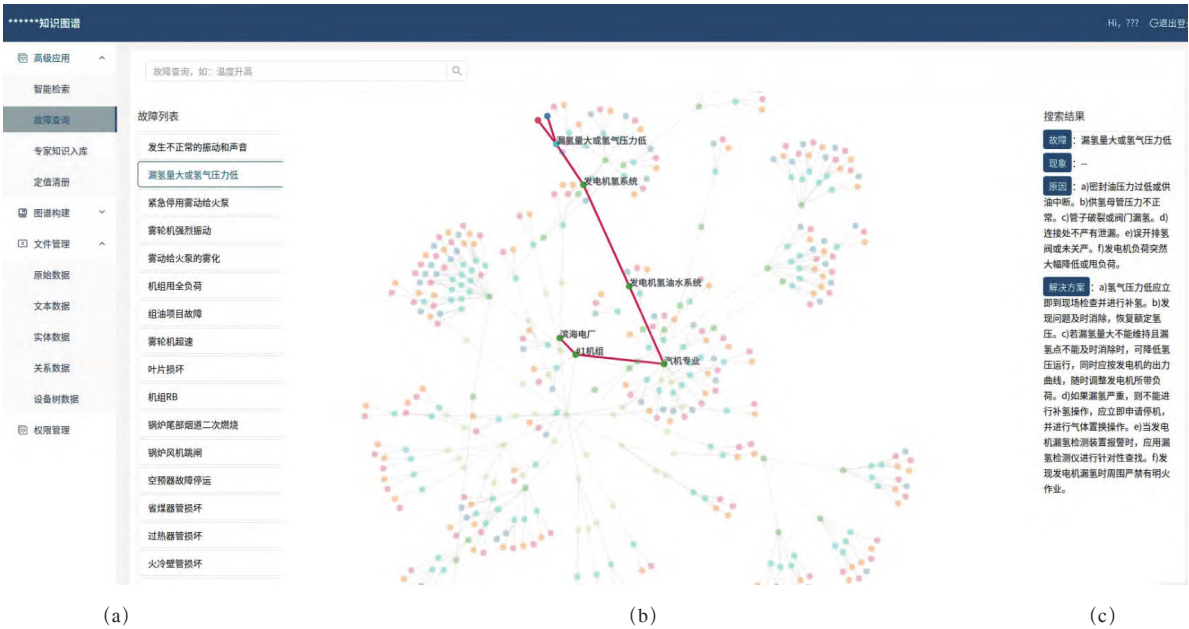


图11 故障知识图谱及故障检索

结束语

针对电厂发电设备故障数据多源异构,故障事件难以识别,故障类型复杂多样,故障事件动态增长等问题,文本提出了一个面向发电设备故障知识图谱的可视化人机交互工具,减少专业领域数据标注的人力物力成本,提高实体抽取的准确率,提高知识图谱构建的质量。该方法使电力系统获得挖掘和分析大规模文本信息中有关知识的能力,从而有针对性地解决当今电厂面临着电力设备类别多样、设备数量繁多、故障类型众多、耦合关系复杂和相关数据多源等问题。测试结果表明,本文提出的面向发电设备故障知识图谱的可视化人机交互工具,较好地解决了实体抽取效率低、关系构建不明确、知识推理准确率低、自动标注准确率低、手工标注费时费力等问题。**EA**

(收稿日期:2024-01-03)

参考文献

- [1] Weihua, Y., Dong, X. Visual analysis of industrial knowledge graph research based on Citespace. In 2021 7th International Conference on Condition Monitoring of Machinery in Non-Stationary Operations (CMMNO). IEEE. 2021, 6: 297-300.
- [2] 谢庆,蔡扬,谢军,等.基于ALBERT的电力变压器运维知识图谱构建方法与应用研究[J].电工技术学报,2023,38(1):95-106.
- [3] 叶欣智,尚磊,董旭柱,等.面向配电网故障处置的知识图谱研究与应用[J].电网技术,2022,46(10):3739-3749.
- [4] 张龙龙,袁静雯,金喆洙.基于人机交互技术的船舶实时导航系统[J].舰船科学技术,2023,45(11):131-134.
- [5] 李瑞连,曾德良,刘吉臻,等.湿式脱硫系统智能化改造架构及关键技术[J].热力发电,2023,52(7):74-86.
- [6] 高海波,崔建新,洪文学,等.用基于平行坐标图的可视化人机交互技术提取EEG信号特征的方法[J].高技术通讯,2010,20(05):518-523.
- [7] Zhang, X., Xuan, X., Dima, A., LabelVizier: Interactive Validation and Relabeling for Technical Text Annotations. In 2023 IEEE 16th Pacific Visualization Symposium (PacificVis) IEEE. 2023-4: 167-176.

