计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(5):87−96 [doi: 10.15888/j.cnki.csa.009102]

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

基于知识图谱的潜油电泵井故障诊断①

宫法明[1], 董文吉[1], 袁向兵[2]

[1](中国石油大学 (华东) 青岛软件学院、计算机科学与技术学院, 青岛 266580)
[2](中石化胜利油田海洋采油厂, 东营 257237)
通信作者: 董文吉, E-mail: dongwenji1997@163.com

摘　要: 潜油电泵井系统是油田开采重要工具, 具有排量大、扬程高与作业环境灵活多变等优点. 为了降低潜油电泵井系统故障危害, 需要对其发生故障部件进行快速精确定位并维修. 本文提出一种基于知识图谱的潜油电泵井故障诊断方法. 采用改进 BiLSTM-CRF 实体识别算法与 BERT 关系抽取算法提取故障数据中的专家知识, 构建潜油电泵井故障诊断领域知识图谱; 利用构建知识图谱搭建以故障征兆为初始节点的贝叶斯推理网络, 利用历史故障数据与条件概率解耦的计算方式推理出故障原因. 本文通过故障诊断真实案例进行方法验证.

关键词: 潜油电泵井; 知识图谱; 故障诊断; BiLSTM-CRF; BERT; 贝叶斯网络

引用格式: 宫法明,董文吉,袁向兵.基于知识图谱的潜油电泵井故障诊断.计算机系统应用,2023,32(5):87–96. http://www.c-s-a.org.cn/1003-3254/9102.html

Fault Diagnosis for Electric Submersible Pump Well Based on Knowledge Graph

GONG Fa-Ming[1], DONG Wen-Ji[1], YUAN Xiang-Bing[2]

[1](Qingdao Institute of Software & College of Computer Science and Technology, China University of Petroleum, Qingdao 266580 China)
[2](Sinopec Shengli Oilfield Offshore Oil Production Plant, Dongying 257237, China)

Abstract: The electric submersible pump well system is an important tool for oilfield exploitation owing to its advantages of large displacement, high head, and flexible operating environment. Reducing the hazards of the faults in the electric submersible pump well system requires the fault components to be quickly and precisely located and repaired. This study proposes a knowledge graph-based fault diagnosis method for electric submersible pump wells. The improved bi-directional long short-term memory-conditional random field (BiLSTM-CRF) entity identification algorithm and the bidirectional encoder representations from transformers (BERT) relation extraction algorithm are used to extract expert knowledge from fault data and then construct a knowledge graph in the field of fault diagnosis of electric submersible pump wells; a Bayesian inference network with fault signs as initial nodes is built with the constructed knowledge graph, and the cause of the fault is inferred by utilizing historical fault data and the calculation method of decoupling conditional probabilities. The proposed method is validated by real fault diagnosis cases.

Key words: electric submersible pump well; knowledge graph (KG); fault diagnosis; BiLSTM-CRF; bi-directional encoder representations from transformers (BERT); Bayesian networks

1　介绍

近年来油田持续开展“四化”建设, 电泵井生产数据已实现智能化传输. 通过不断改进潜油电泵数据采集技术, 海上电泵井已经实现了工艺自动化数据、井下机组参数与井筒参数采集. 以此数据资源为基础, 采用潜油电泵机理模型和数理模型两种技术方法进行工

① 收稿时间: 2022-10-20; 修改时间: 2022-12-10; 采用时间: 2022-12-23; csa 在线出版时间: 2023-03-24
CNKI 网络首发时间: 2023-03-27

况诊断应用, 但两种方法均存在一定局限性.

传统机理模型有电流卡片法与憋压诊断法两种代表性方法. 故障发生后, 两者需要专家根据记录数据特征进行分析. 采用后处理方式导致系统实时性较差, 需要丰富电泵井工况经验才能对检测结果做出准确判断; 数理模型采用大数据与深度学习技术进行实时分析, 通过分析历史工况样本数据提取参数特征建立故障诊断模型, 但故障数据难以大量积累, 影响模型诊断准确率.

近年来, 国内外已有相关文献将知识图谱应用在故障诊断领域. 许驹雄等[1] 提出了面向发动机的知识图谱构建与应用, 通过发动机生产故障与售后维修报告构建领域知识图谱, 使用知识图谱进行可视化检索与辅助决策. Bian 等[2] 提出了工程机械故障知识图谱的构建与推理方法, 通过预先设定的规则, 将工程机械的维修文档进行自动化提取三元组, 通过交替迭代训练获得辅助决策模型, 为辅助工程机械故障排除提供了新的思路.

通过与油田技术人员交流发现, 在油田领域应用知识图谱还存在诸多不确定性, 缺乏系统研究. 例如, 油田对知识图谱应用前景有所怀疑, 不确定如何将其应用到油田的故障诊断、采油、集输等流程. 此外, 目前缺乏基于知识图谱潜油电泵井故障诊断流程. 因此, 本文的主要工作如下.

(1) 建立基于知识图谱潜油电泵井故障诊断流程方法.

(2) 在知识图谱构建阶段, 改进现有知识抽取算法. 实体识别阶段, 在 BiLSTM-CRF 模型中融合 BERT 参数, 提高了模型泛化性, 有效解决了 OOV (out of vocabulary) 问题; 在关系抽取阶段, 在 BERT 模型中加入字词注意力机制层, 提高关系分类精度.

(3) 在故障推理阶段, 以知识图谱的故障征兆为导向, 构建贝叶斯推理网络[3,4]. 提出专家打分模型与 noisy-OR[5] 模型来解决拓扑结构复杂导致计算量过高与故障样本数据不足问题. 该方法解决了以往知识图谱仅用来检索、定性的性质, 真正意义上实现了知识图谱定量推理.

2 潜油电泵井系统故障诊断知识图谱构建

潜油电泵井故障诊断知识图谱构建及应用流程[6] 包括明确潜油电泵井故障诊断领域本体类型[7]、领域实体识别与关系抽取、基于 Neo4j 图数据库构建[8]. 具体构建流程如下.

(1) 通过专家判定和应用需求明确本体类型, 确定图谱中实体和关系种类.

(2) 利用精灵标注助手软件对潜油电泵井故障诊断部分语料进行数据集标注[9], 训练故障领域实体识别模型. 利用训练模型识别故障诊断语料实体, 构建实体集合.

(3) 对上述数据集进行关系标注, 训练故障领域关系抽取模型. 使用训练好的关系抽取模型抽取故障诊断语料中的关系, 建立关系对集合.

(4) 完成实体识别和关系抽取后, 将实体和关系对集合进行三元组组合, 将三元组导入知识图谱构建工具 Neo4j, 构建出面向潜油电泵井故障诊断知识图谱.

2.1 故障领域本体构建

在构建某一领域专业图谱时, 应根据专家知识进行本体构建, 为实体识别与关系抽取提供规范. 本文图谱是针对潜油电泵井领域, 即构建潜油电泵井故障诊断领域知识图谱, 因而根据专家对潜油电泵井故障诊断的经验知识进行本体构建. 本文采用 Protege 工具进行本体设计, 本体设计可视化如图 1 所示.

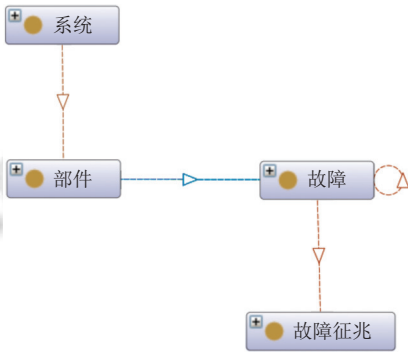

图 1 潜油电泵井故障领域本体设计示意图

潜油电泵井故障领域不同类之间的关联关系如表 1 所示, 共定义了组成、发生、导致 3 种关系.

表 1 概念关联关系表

Domain	ObjectProperty	Range
系统	组成	部件
部件	发生	故障
故障	导致	故障
故障	导致	故障征兆

2.2 基于 BERT-BiLSTM-CRF 实体识别

本文在目前使用较为广泛 BiLSTM-CRF 网络基

础上融合 BERT 参数[10], 建立 BERT-BiLSTM-CRF 模型. 改进模型可以动态调整字词语义信息, 提高同义多形词识别精度, 提高原模型识别泛化性.

2.2.1 BERT-BiLSTM-CRF 网络

BERT-BiLSTM-CRF 模型结构如图 2 所示. 句子通过 BERT 预训练模型得到其每个字的向量特征, 将构建的向量序列输入到 BiLSTM 模型中进行语义特征提取, 语义特征通过 CRF 层得到最有可能的标签标识[11].

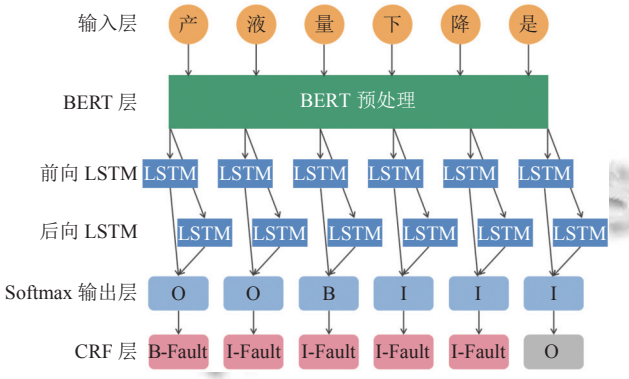

图 2 BERT-BiLSTM-CRF 模型结构图

(1) BERT 预训练语言模型

为了更好地获取句子语义特征向量, 使用 BERT 预训练模型来代替传统的 Word2Vec 方法[12]. BERT 模型可根据语境动态调整字词语义信息, 能够充分提取句子语义特征信息, 获得更好的识别结果.

(2) BiLSTM 层

BiLSTM 设置前向和后向 LSTM 网络, 输出层结果由两者共同决定. 其中, 每一个 LSTM 计算模块包含输入门、遗忘门和输出门[13]. 具体计算过程如下:

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \tag{1}$$

$$f_t = \sigma\left(w_f[h_{t-1}, x_t] + b_f\right) \tag{2}$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \tag{3}$$

$$\tilde{c}_t = \tanh(w_c[h_{t-1}, x_t] + b_c) \tag{4}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{5}$$

$$h_t = o_t * \tanh(c_t) \tag{6}$$

其中, i_t、f_t、o_t 代表在时间跨度为 t 的输入门、遗忘门、输出门输出; x_t、h_t 代表在时间跨度为 t 的输入和输出向量.

(3) CRF 层

通过 BERT-BiLSTM 网络处理, 句子能够充分提取其特征, 获得对应的序列向量. 为了保证实体识别的精度, 利用 CRF 层在序列向量上进行一层计算约束, 保证与预先设定标签的对应关系. 对于句子 X 预测序列 $Y=(y_1, y_2, \cdots, y_n)$ 概率为:

$$P(Y|X) = \frac{e^{score(x,y)}}{\sum_y e^{score(x,y)}} \tag{7}$$

$$score(X, Y) = \sum_{i=1}^{n}(P_{i,y_i} + W_{y_i,y_{i+1}}) \tag{8}$$

通过 CRF 层对 BERT-BiLSTM 层的输出序列进行处理, 能够计算得到子序列与输出标签的契合得分, 使实体识别结果更为准确.

2.2.2 基于 BERT-BiLSTM-CRF 实体识别实验

(1) 数据获取与处理

本文选取某海洋采油厂近 20 年历史躺井案例、知网潜油电泵井故障诊断相关文献作为语料进行分析. 通过文本预处理 (PDF 文本提取、分句等) 得到 14 856 条语句. 数据集标注规则如下.

① 将实体分为系统、部件、故障与故障征兆 4 类, 标注前系统学习潜油电泵井系统工作原理.

② 故障与故障征兆里包含的部件不进行标注, 标注标签只能存在一层, 故障诊断系统更关注故障与故障征兆.

在此基础上, 利用精灵标注助手软件进行标注, 之后采用 BIO 标注法对标注之后语料进行调整, 其中 B 表示当前字是实体的首字, I 表示是实体的非首字, O 表示不是实体的字. 表 2 给出了标注实例.

表 2 BIO 标注实例

由	于	电	缆	短	路
O	O	B-Fault	I-Fault	I-Fault	I-Fault

将标注数据语料中的 11 885 条作为训练集数据, 2 971 条作为测试集数据.

(2) 参数设置与实验过程

本文采用了 Google 发布的 BERT 中文预训练模型, 该模型包含 12 个隐藏层、768 个隐藏层单元、12 个 encode 层注意头和 110M 个参数, 最大序列长度设置为 128, epoch 设置为 100, batch_size 设置为 128, LSTM 的隐藏单元个数设置为 200, 其他设置为默认值. 实验环境依靠 TensorFlow 库[14].

实验过程中, BERT 中文预训练模型提取句子字词

语义表示向量序列; BiLSTM 网络能够根据上下文语境, 调整句子语义特征; CRF 层进行标签约束. 为了检验模型的性能, 实验过程中与其他实体识别模型进行了对比.

(3) 结果分析

为了检验模型优劣, 本文使用精确率、召回率与 $F1$ 分数进行评估, 计算公式如下:

$$precision = \frac{TP}{TP+FP} \tag{9}$$

$$recall = \frac{TP}{TP+FN} \tag{10}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{11}$$

经过实验, 各个模型测试集预测结果如表 3 所示.

从表 3 可以看出, BERT-BiLSTM-CRF 模型识别效果明显优于其他 3 个模型. 在 BiLSTM-CRF 模型中加入 BERT 中文预训练模型, 对潜油电泵井故障诊断实体识别模型效果有显著的提升.

表 3 实体识别模型实验结果

模型	精确率	召回率	$F1$分数
BiLSTM	0.61	0.57	0.60
BiLSTM-CRF	0.74	0.53	0.61
BERT-BiLSTM	0.79	0.58	0.67
BERT-BiLSTM-CRF	0.84	0.63	0.72

2.3 基于 Word-Attention-BERT 的关系抽取

获取到实体节点之后, 需根据句子与句子中实体对进行关系抽取分析. 本文使用目前较为广泛应用的 BERT 模型作为基础网络, 通过在网络模型中加入字词注意力层来提高模型关系抽取精度.

2.3.1 Word-Attention-BERT 介绍

Word-Attention-BERT 模型整体结构如图 3 所示, 整个网络结构主要包括 BERT 编码层、字词注意力层和关系分类层.

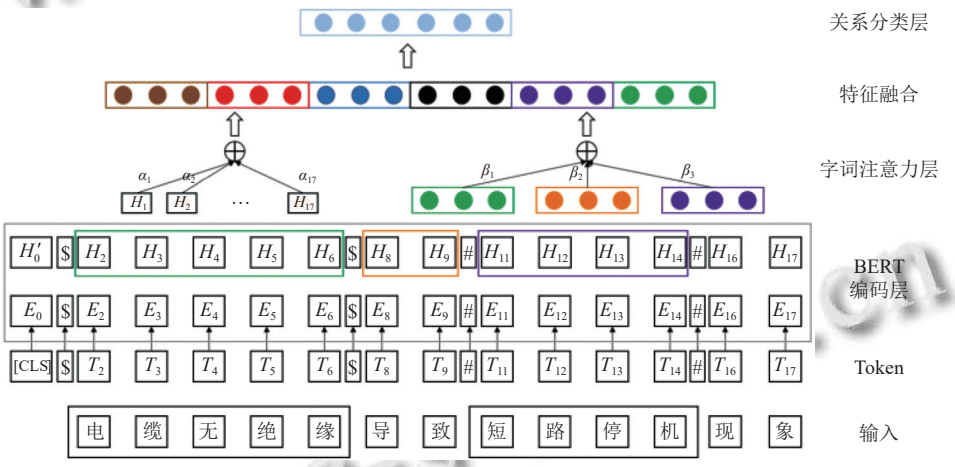

图 3 Word-Attention-BERT 模型结构图

将输入句子每个字转换为 Token, 将 Token 输入到 BERT 编码层中得到句子的向量表示. 在实际关系抽取过程中, 实体之间的关系往往是由某个词决定的, 因此在 BERT 关系抽取网络中加入字词注意力层, 能够突出关键字词的作用, 提升关系抽取精度.

(1) BERT 编码层

为了使 BERT 模型能够获取句子中两个实体位置和边界信息, 使用“$”和“#”来标记句子中实体.

句子和实体经过 BERT 编码层编码过后得到 $H=\{H_0', H_1, \cdots, H_{n+4}\}$, H_0'是标记符号“[CLS]”对应的编码输出. H_0'作为整个句子语义表示, 是由 H_0 添加 tanh 激活函数并经过线性变换后得到. 实体信息通过计算实体所包含 H_t 的平均值, 经过 tanh 激活函数并经过线性变换得到实体最终表示. 词向量直接计算词所包含的 H_t 的平均值. 具体计算过程如下:

$$H_0' = W_0(\tanh(H_0)) + b_0 \tag{12}$$

$$e_1 = W_1\left[\tanh\left(\frac{1}{j-i+1}\sum_{t=i}^{j} H_t\right)\right] + b_1 \tag{13}$$

$$e_2 = W_2\left[\tanh\left(\frac{1}{m-k+1}\sum_{t=k}^{m} H_t\right)\right] + b_2 \tag{14}$$

$$Word=\frac{1}{s-r+1}\sum_{t=s}^{s}H_t \tag{15}$$

其中, W_0、W_1、W_2 表示维度为 $d\times d$ 的可训练权重矩阵; b_0、b_1、b_2 表示维度为 $d\times 1$ 的可训练权重向量; d 表示字向量维度; e_1、e_2 表示实体; i、j 分别表示实体 1 的起止位置; k、m 表示实体 2 的起止位置; r、s 表示词的起止位置. $Word$ 表示句子中词的泛指.

(2) 字词注意力层

在故障诊断领域“由于”“带来”“因为”等字词往往决定了实体之间的关系, 所以在融合句子特征时加入字词注意力层, 增强关键字词特征以提高关系抽取精度.

经过 BERT 编码层编码之后的句子 $H=\{H_0', H_1, \cdots, H_{n+4}\}$, 使用字词注意力机制增强关键字词特征. 将 $\{H_1, \cdots, H_{n+4}\}$作为输入, 采用一个两层的神经网络, 得到一个维度为 $1\times(n+4)$、取值范围为 (0, 1) 的重要度向量 α. 同理在词尺度上得到重要度向量 β. 最后对字词的向量进行加权, 并将加权后的向量级联起来, 最后通过 tanh 激活函数激活并经过线性变换得到最终的句子融合特征表示. 具体计算过程如式 (16)–式 (21) 所示:

$$\alpha=Softmax(q_c\tanh(Q_cH_c)) \tag{16}$$

$$\beta=Softmax(q_w\tanh(Q_wH_w)) \tag{17}$$

$$h'=\sum_{t=1}^{n+4}\alpha_tH_t \tag{18}$$

$$w'=\sum_{s=1}^{Num_{Word}}\beta_sH_s \tag{19}$$

$$h''=W_h\tanh(h')+b_h \tag{20}$$

$$w''=W_w\tanh(w')+b_w \tag{21}$$

其中, α、β 分别表示字和词层面的重要度向量; q_c、q_w 表示维度为 $1\times d$ 的可训练权重向量; Q_c、Q_w 表示维度为 $d\times d$ 的可训练权重矩阵; H_c、H_w 分别表示字和词层面的输入向量; h'、w'分别表示字和词层面的重要度加权句意特征向量; h''、w''表示 h'、w'经过 tanh 激活函数并经过线性变换得到的句意特征向量的最终表示; W_h、W_w 表示维度为 $d\times d$ 的可训练权重矩阵; b_h、b_w 表示维度为的可训练权重向量.

(3) 关系分类层

将句子特征向量表示、实体特征向量表示、字层面的句意特征向量和词层面的句意特征向量级联融合, 然后输入 *Softmax* 函数进行关系分类, 得到最终结果. 具体计算过程如式 (22) 和式 (23) 所示:

$$r=W_r[concat(H_0', e_1, e_2, h'', w'')]+b_r \tag{22}$$

$$outcome=Softmax(r) \tag{23}$$

其中, r 表示句意的融合特征向量; W_r 表示维度为 $L\times 5d$ 的可训练权重矩阵; b_r 表示维度为 $L\times 1$ 的可训练权重向量; L 表示关系种类数目.

2.3.2 基于 Word-Attention-BERT 的关系抽取实验流程及结果分析

(1) 数据标注与处理

本文将潜油电泵井故障诊断领域实体间的关系定义为 4 类: 组成、发生、导致、未知, 其中未知表示提取关系不是我们关注内容. 选取某海洋采油厂近 20 年的历史躺井案例、知网潜油电泵井故障诊断相关文献作为语料进行标注. 利用精灵标注助手软件进行标注之后, 将其进一步调整为 JSON 格式. 之后将处理后的数据按 8:2 的比例划分为训练集和测试集, 分别为 11 885 条和 2 971 条.

(2) 参数设置与实验过程

在关系抽取实验过程中同样使用了 Google 发布的 BERT 中文预训练模型, 并在此基础上针对关系抽取任务进行微调. 学习率设置为 0.000 5、epoch 设置为 200、batchsize 设置为 10, 最大序列长度为 128.

实验过程中, 首先将语料中的句子进行预处理, 获得两个实体的位置信息, 之后通过 BERT 编码层获得具有语义表示的字向量与句向量, 再将字向量送入字词注意力层获得分别在字词层的句意向量表示, 最后通过 *Softmax* 层输出两个实体的关系类别.

(3) 实验结果分析

目前 BERT 的关系抽取模型显著优于其他模型. 在潜油电泵井故障诊断领域关系抽取数据集上, 本文仅选用 BERT 与 Word-Attention-BERT 进行对比, 评价指标同样选用精确率、召回率和 $F1$ 分数, 具体结果如表 4 所示.

表 4 关系抽取实验结果

模型	精确率	召回率	$F1$分数
BERT	0.84	0.89	0.86
Word-Attention-BERT	0.92	0.90	0.91

从表 4 可以看出, Word-Attention-BERT 在准确率、召回率和 $F1$ 分数的表现均优于普通的 BERT

关系抽取模型. 这说明 Word-Attention-BERT 能更好地利用句子中的关键信息提高识别精度.

2.4 潜油电泵井故障领域知识筛选与表达规约

非结构化潜油电泵井故障领域知识文本经过 BERT-BiLSTM-CRF 实体识别模型与 Word-Attention-BERT 关系抽取模型处理, 抽取出来的知识可能存在歧义与重复的数据; 另外, 抽取出来的故障机理知识存在错误与个例的情况, 可能对故障诊断模型存在负影响. 所以本节对抽取的知识进行筛选与表达规约, 具体处理方法如下.

(1) 歧义消解: 抽取出来的知识存在有歧义的实体, 例如, “绝缘为零”在不同文本中含义不同, 可能代表“电缆绝缘为零”或者“电机绝缘为零”, 因此应当结合上下文明确实体的含义.

(2) 重复合并: 抽取出来的知识存在一种实体在不同文本表达不一的情况, 例如, “油管有圆洞”与“油管有砂眼”表示同一个含义, 因此需要对指向同一含义的实体进行合并处理.

(3) 知识纠正: 由于抽取语料质量不同, 必然存在抽取知识有误的情况, 例如, “保护器故障”导致“短路停井”是偷换概念, 经专家分析, “保护器故障”是导致“电机进井液”, 进而引发“短路停井”, 因此需要对抽取的知识进行知识纠正, 改正抽取有误的知识.

(4) 非典型知识删除: 由于抽取语料过多, 存在抽取知识是个例的情况, 例如, “气锁”导致“井口温度上升”是从特殊案例中抽取的知识, 不具有典型性, 甚至对故障诊断模型产生负影响, 因此需要将非典型知识进行删除, 保证构建知识图谱中知识的典型性.

2.5 基于 Neo4j 的知识图谱存储

抽取出的知识经筛选与表达规约得到三元组集合, 通过三元组集合构建知识图谱. Neo4j 具有快速、灵活与开发敏捷性等特点, 因此本文选用 Neo4j 作为知识图谱构建工具[15].

将潜油电泵井故障诊断领域三元组集合导入 Neo4j 进行存储, 部分知识存储结果如图 4 所示. 其中绿色代表潜油电泵井系统、粉色代表组成部件、蓝色代表故障、橙色代表故障征兆.

3 知识图谱在潜油电泵井系统故障诊断中的应用

本文以某采油厂近 20 年 1 476 份潜油电泵井故障诊断分析报告以及潜油电泵井故障诊断知网文献 76 篇作为知识抽取语料, 利用第 2 节所提的知识图谱构建方法构建潜油电泵井故障诊断知识图谱并将其存储在 Neo4j 图数据库中, 一共包含 1 486 个实体节点和 2 372 条关系. 使用 Cypher 语言在 Neo4j 图数据库中进行检索, 通过输入检测到的故障征兆, 便可以快速得到与之相关的故障图谱[16]. 以往根据故障征兆进行检索无法准确定位故障原因, 需要故障检修人员参与检查.

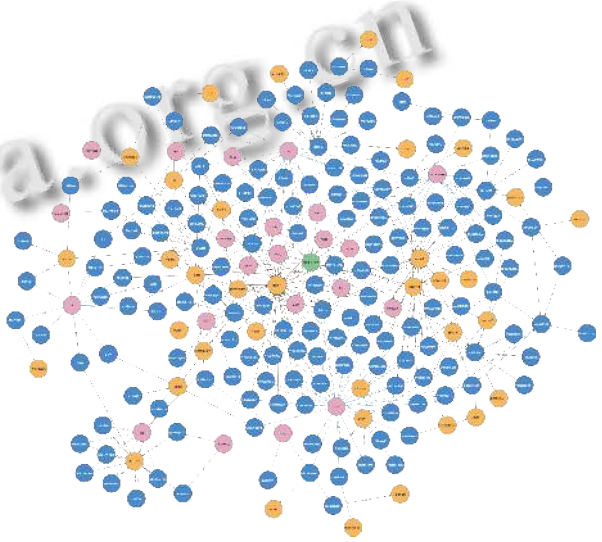

图 4 部分潜油电泵井故障诊断知识图谱

贝叶斯网络作为一种概率图模型, 可以通过变量节点间的关联关系进行推理, 即通过一些变量的状态来推测获得其他变量各个状态概率, 得到节点变量最有可能结果[17]. 上文构建的潜油电泵井故障诊断知识图谱为贝叶斯网络提供了合适的网络结构. 通过观测的故障征兆输入 Neo4j 进行检索, 将检索图谱查询子图作为贝叶斯推理基础网络. 根据专家知识与以往故障记录, 确定贝叶斯网络关联关系与先验概率, 由此来定位故障.

3.1 建立贝叶斯网络故障诊断模型

3.1.1 贝叶斯网络

贝叶斯网络是一个概率推理网络. 具体来说, 贝叶斯网络 B 由网络架构 G 和参数 θ 构成, 即 $B=(G, \theta)$. 其中网络架构 G 是一个有向无环图. 参数 θ 定量地描述了节点间关联关系的强度[18]. 具体计算公式如下:

$$\theta_{x_i|\pi_i} = P_B(x_i|\pi_i) \tag{24}$$

$$P(A|B) = \frac{P(AB)}{P(B)} \tag{25}$$

$$P(A)=\sum_i P(A|B_i)P(B_i) \tag{26}$$

$$P(B_i|A)=\frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n}P(B_j)P(A|B_i)} \tag{27}$$

$$P(X_1,X_2,\cdots,X_n)=\prod_{i=1}^{n}P(X_i|\pi_i) \tag{28}$$

其中, 式 (24) 表示每个变量条件概率表示, π_i 为变量 x_i 在 G 中父节点集; 式 (25) 表示条件概率公式, $P(A|B)$ 表示事件 B 发生的条件下事件 A 发生的概率, $P(AB)$ 表示事件 A 和事件 B 同时发生的概率, $P(B)$ 表示事件 B 发生的概率; 式 (26) 表示全概率公式; 式 (27) 表示贝叶斯公式, 由条件概率公式和全概率公式推导得出; 式 (28) 表示联合概率分布, X_i 表示贝叶斯网络中的节点, π_i 表示 X_i 父节点集合.

3.1.2 故障先验概率设定

故障先验概率表示潜油电泵井运行历史中各类故障概率统计. 但是完备故障数据在实际生产中很难获得, 根据现有的数据无法保证先验概率有效性. 文献 [19] 提出了一种概率推理方法, 只需要统计潜油电泵井历史上发生的概率既可设定故障先验概率.

根据贝叶斯定理, 故障征兆 s 由故障 f 导致概率为:

$$P(f|s)=\frac{P(s|f)P(f)}{P(s)} \tag{29}$$

因此, 先验概率 $P(f)$ 可表示为以下形式:

$$P(f)=\frac{P(f|s)P(s)}{P(s|f)} \tag{30}$$

其中, $P(s|f)$ 表示故障 f 导致故障征兆 s 的概率. 为简化计算, 不妨设故障 f 发生必然引发故障征兆 s, 即 $P(s|f)=1$, 则式 (30) 变为:

$$P(f)=P(f|s)P(s) \tag{31}$$

式 (31) 即为故障 f 先验概率计算公式. 由于故障征兆 s 已经出现, 故有 $P(s)=1$; $P(f|s)$ 表示故障 f 在导致故障征兆 s 出现所有记录中所占的概率. 例如, 某口井“油压波动”的 118 例故障记录中, 有 16 例是由“游离气影响”引起的, 则 $P(f|s)=16/118=0.136$.

除此之外, 对于缺乏统计资料的故障, 需要根据专家经验设定故障的先验概率. 本文提出由多位专家进行故障评估方法, 能够一定程度上避免人为主观影响. 基于这种策略, 本文在经典贝叶斯网络中加入附加节点, 融入专家知识[20].

改进后的贝叶斯网络诊断模型见图 5, 节点 e 即为引入的附加节点, 该节点基于不同专家知识获得故障先验概率, 然后进行加权处理. 具体计算过程如下:

$$P_j=\frac{\sum_{i=1}^{k}\alpha_i P_{ij}}{k},\ j=1,2,\cdots,n \tag{32}$$

其中, P_j 表示第 j 个故障的先验概率; k 表示专家个数; α_i 表示第 i 个专家置信度权值; $P_{ij}=P(f_j=T|e_i)$ 即表示第 i 个专家对第 j 个故障先验发生概率的评估.

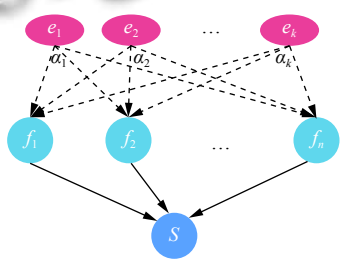

图 5 修改后的贝叶斯网络诊断模型

3.1.3 基于因果机制独立条件下的故障与征兆关联强度分析

由于潜油电泵井系统复杂性, 潜油电泵井故障与故障征兆间关联关系受很多因素影响, 经典贝叶斯网络建模方法需要根据条件概率来表示关联强度[21]. 然而, 传统贝叶斯网络建模方法参数量的设置随着节点数的增加呈指数式增长. 另外, 条件概率需要保证其精确程度, 必须建立潜油电泵井故障领域完整的概率统计信息. 因而限制了贝叶斯网络的实际应用.

本文采用 noisy-OR 模型量化潜油电泵井与故障征兆间的因果关系. 通过因果机制独立假设[22,23] 分析潜油电泵井故障事件之间的关联性. 如图 6 所示, 设 $f_1,\cdots,f_n$ 是引发故障征兆 s 出现的 n 种故障, 其与故障征兆 s 间的因果机制相互独立. 令 $P_i=P(s=T|f_i=T, f_j=F_{[\forall j,j\neq i]})$ 表示故障 f_i 单独导致故障征兆 s 出现的概率. 若所有故障均为二态取值, 多种故障 $f_1,\cdots,f_n$ 与故障征兆 s 的耦合效果如下所示:

$$\begin{cases}P(s=F|pa(s))=\prod\limits_{i:f_i\in pa(s)^+}(1-P_i)\\ P(s=T|pa(s))=1-\prod\limits_{i:f_i\in pa(s)^+}(1-P_i)\end{cases} \tag{33}$$

其中, $pa(s)^{+}$表示节点 s 的取值为真的父节点子集.

采用 noisy-OR 模型后, 节点间关联强度通过单个父节点作用效果耦合得到. 基于此, 其所需先验知识大为减少, 从而降低了贝叶斯网络故障诊断建模的复杂性.

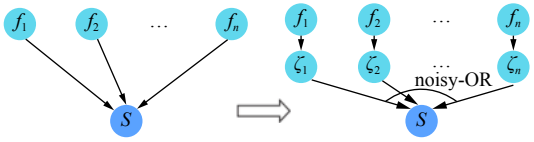

图 6 noisy-OR 模型图

3.2 案例验证

接下来通过案例来介绍如何基于知识图谱与贝叶斯网络进行潜油电泵井故障诊断. 假设已知故障现象“欠载停机”. 首先, 在 Neo4j 中使用 Cypher 查询语句检索与这一故障征兆相关的故障, 可以得到图 7 所示的知识图谱查询子图.

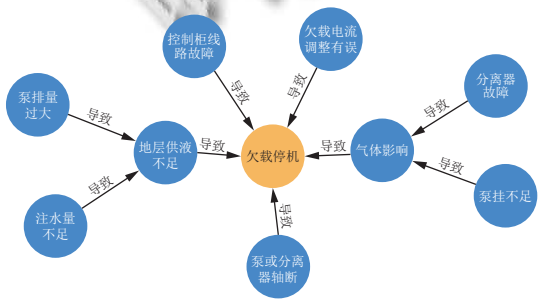

图 7 欠载停机故障现象检索结果

然后根据知识图谱故障现象查询子图构建贝叶斯网络. GeNIe 是一款专业的贝叶斯网络可视化软件, 建模工作者可以利用 GeNIe 构建贝叶斯网络模型进行推理[24]. 本文在知识图谱故障现象检索结果的基础上, 利用 GeNIe 根据知识图谱故障原因查询子图进行贝叶斯网络结构建模, 如图 8 所示.

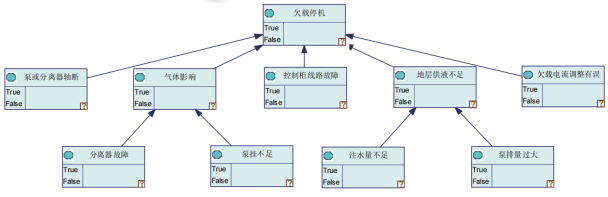

图 8 贝叶斯网络故障诊断模型

在利用贝叶斯网络推理之前, 要确定贝叶斯网络中根节点的先验概率和其他节点的条件概率取值. 本文根据某采油厂历史故障记录与第 3.1.2 节专家打分法确定根节点的先验取值, 根据第 3.1.3 节方法确定节点之间的关联强度, 如表 5、表 6 所示.

表 5 根节点先验概率表

贝叶斯网络根节点	先验概率
泵或分离器轴断	0.06
分离器故障	0.08
泵挂不足	0.13
注水量不足	0.07
泵排量过大	0.05
控制柜线路故障	0.06
欠载电流调整不正确	0.08

表 6 关联强度表

关联关系	关联强度
泵或分离器轴断-欠载停机	0.75
气体影响-欠载停机	0.85
地层供液不足-欠载停机	0.8
控制柜线路故障-欠载停机	0.55
欠载电流调整不正确-欠载停机	0.45
分离器故障-气体影响	0.75
泵挂不足-气体影响	0.7
注水量不足-地层供液不足	0.8
泵排量过大-地层供液不足	0.85

将上述先验概率、由关联强度和式 (33) 确定的条件概率输入到贝叶斯推理模型中. 设置初始节点状态, 检查各故障的逆反事件, 发现排量正常, 排除泵或分离器轴断故障; 欠载电流阈值正常, 排除欠载电流调整不正确故障; 电机温度正常, 排除泵挂不足故障. 故将欠载停机节点状态设置为 True, 泵或分离器轴断故障、泵挂不足与欠载电流调整不正确节点状态设置为 False, 如图 9 所示.

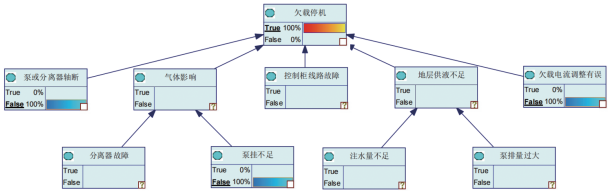

图 9 贝叶斯网络故障诊断初始节点设置图

利用 GeNIe 进行贝叶斯网络推理, 得到引发欠载停机最有可能的原因. 从图 10 可以看出, 分离器故障最有可能引发欠载停机. 现场工作人员在检修过程中发现分离器多处穿孔, 导致大量游离气的井液进入潜油电泵机组, 井液中游离气的含量大于离心泵的设计允许值, 使离心泵工作性能不稳定, 效率下降, 导致欠

载停机. 可以看出人工检修结果与贝叶斯网络推断出的故障原因一致, 这证明本文方法是有效的.

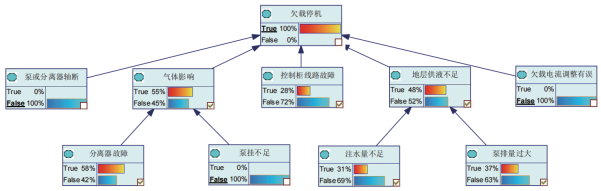

图 10 贝叶斯网络故障诊断结果

4 结论

为了解决潜油电泵井故障诊断快速准确定位问题, 本文提出了一种基于知识图谱的潜油电泵井故障诊断方法. 提出一种自动抽取历史故障案例与专业知识三元组算法, 实现自动构建潜油电泵井故障诊断知识图谱; 利用 Neo4j 图数据库进行存储管理; 提出专家打分系统与 noisy-OR 方法相结合来改善故障样本少且计算复杂问题; 通过案例验证此方法的有效性, 为工作人员提供了便捷.

潜油电泵井故障领域复杂, 而本文语料数据来源较少, 存在知识图谱自动构建模型训练不够充分等问题; 贝叶斯网络先验概率也缺少大数据支撑. 未来可以聚焦在整个潜油电泵井故障领域大规模语料数据构建与故障数据的扩充, 提高模型推理效果.

参考文献

1 许驹雄, 李敏波, 刘孟珂, 等. 发动机故障领域知识图谱构建与应用. 计算机系统应用, 2022, 31(7): 66–76. [doi: 10.15888/j.cnki.csa.008592]

2 Bian JN, Mao ZH, Liu Y, *et al*. Construction and reasoning method of fault knowledge graph with application of engineering machinery. Proceedings of the 2021 China Automation Congress (CAC). Beijing: IEEE, 2021. 2577–2581. [doi: 10.1109/CAC53003.2021.9727906.]

3 Luo WH, Cai FT, Wu CN, *et al*. Bayesian network-based knowledge graph inference for highway transportation safety risks. Advances in Civil Engineering, 2021, 2021: 6624579. [doi: 10.1155/2021/6624579]

4 Pan HL, Yang XH. Intelligent recommendation method integrating knowledge graph and Bayesian network. Soft Computing, 2023, 27(1): 483–492. [doi: 10.1007/s00500-021-05735-z]

5 Ji CY, Su X, Qin ZF, *et al*. Probability analysis of construction risk based on noisy-OR gate Bayesian networks. Reliability Engineering & System Safety, 2022, 217: 107974. [doi: 10.1016/j.ress.2021.107974]

6 聂同攀, 曾继炎, 程玉杰, 等. 面向飞机电源系统故障诊断的知识图谱构建技术及应用. 航空学报, 2022, 43(8): 625499. [doi: 10.7527/S1000-6893.2021.25499]

7 Tiwari S, Al-Aswadi FN, Gaurav D. Recent trends in knowledge graphs: Theory and practice. Soft Computing, 2021, 25(13): 8337–8355. [doi: 10.1007/s00500-021-05756-8]

8 Liu HB, Jiang GY, Su LH, *et al*. Construction of power projects knowledge graph based on graph database Neo4j. Proceedings of the 2020 International Conference on Computer, Information and Telecommunication Systems (CITS). Hangzhou: IEEE, 2020. 1–4. [doi: 10.1109/cits49457.2020.9232609]

9 陈曦. 基于领域知识图谱的柴油发动机故障诊断研究 [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2021. [doi: 10.27061/d.cnki.ghgdu.2021.001490]

10 Wu GH, Tang GG, Wang ZR, *et al*. An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition. IEEE Access, 2019, 7: 113942–113949. [doi: 10.1109/ACCESS.2019.2935223.]

11 Meng FQ, Yang SS, Wang JD, *et al*. Creating knowledge graph of electric power equipment faults based on BERT-BiLSTM-CRF model. Journal of Electrical Engineering & Technology, 2022, 17(4): 2507–2516. [doi: 10.1007/s42835-022-01032-3]

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 Yu Y, Si XS, Hu CH, *et al*. A review of recurrent neural networks: LSTM cells and network architectures. Neural Computation, 2019, 31(7): 1235–1270. [doi: 10.1162/neco_a_01199]

14 Sanchez SA, Romero HJ, Morales AD. A review: Comparison of performance metrics of pretrained models for object detection using the TensorFlow framework. IOP Conference Series: Materials Science and Engineering, 2020, 844: 012024. [doi: 10.1088/1757-899x/844/1/012024]

15 Wiese L. Data analytics with graph algorithms—A hands-on tutorial with Neo4j. Proceedings of the BTW 2019 Workshopband. Bonn: Gesellschaft für Informatik, 2019.

259–261. [doi: 10.18420/btw2019-ws-26]

16 Holden WJ. Analyzing open shortest path first (OSPF) networks with Neo4j and Cypher. https://www.wjholden.com/neo4j-ospf.pdf. (2021-07-26).

17 Amin MT, Khan F, Imtiaz S. Fault detection and pathway analysis using a dynamic Bayesian network. Chemical Engineering Science, 2019, 195: 777–790. [doi: 10.1016/j.ces.2018.10.024]

18 Ibne Hossain NU, Nagahi M, Jaradat R, *et al*. Modeling and assessing Cyber resilience of smart grid using Bayesian network-based approach: A system of systems problem. Journal of Computational Design and Engineering, 2020, 7(3): 352–366. [doi: 10.1093/jcde/qwaa029]

19 王金鑫. 船用柴油机多故障解耦与智能诊断方法研究 [博士学位论文]. 哈尔滨: 哈尔滨工程大学, 2020. [doi: 10.27060/d.cnki.ghbcu.2020.001749]

20 Nababan M, Laia Y, Sitanggang D, *et al*. The diagnose of oil palm disease using naive Bayes method based on expert system technology. Journal of Physics: Conference Series, 2018, 1007: 012015. [doi: 10.1088/1742-6596/1007/1/012015]

21 王红, 郭笑丹, 祝寒. 基于贝叶斯网络的民航突发事件因果关系分析方法研究. 计算机应用研究, 2019, 36(3): 711–715. [doi: 10.19734/j.issn.1001-3695.2017.09.0926]

22 Chiappa S, Isaac WS. A causal Bayesian networks viewpoint on fairness. Proceedings of the 13th IFIP International Summer School on Privacy and Identity Management. Vienna: Springer, 2019. 3–20. [doi: 10.1007/978-3-030-16744-8_1]

23 Luthfi A, Janssen M, Crompvoets J. A causal explanatory model of Bayesian-belief networks for analysing the risks of opening data. Proceedings of the 8th International Symposium on Business Modeling and Software Design. Vienna: Springer, 2018. 289–297. [doi: 10.1007/978-3-319-94214-8_20]

24 陈志煌, 刘国恒, 王莹莹, 等. 基于动态贝叶斯网络的水下连接器故障诊断. 中国安全科学学报, 2020, 30(5): 81–87. [doi: 10.16265/j.cnki.issn1003-3033.2020.05.013]

(校对责编: 牛欣悦)