

铁路内燃机车故障维修知识图谱与决策推荐

杨子赢^{1*}, 王永顺², 莫祥伦¹, 张瀚文¹

¹中国矿业大学矿业工程学院, 江苏 徐州

²徐州地铁运营有限公司, 江苏 徐州

收稿日期: 2025年5月9日; 录用日期: 2025年7月2日; 发布日期: 2025年7月10日

摘要

针对铁路内燃机车维修领域数据分散、信息孤岛等痛点, 文章提出了一种基于知识图谱构建与推理的智能辅助决策方法。首先, 采用自顶向下策略, 结合现场维修案例, 通过自定义词典和正则规则抽取设备型号、故障部位; 针对故障原因、现象和维修措施等语义复杂要素, 引入BERT-CRF模型进行抽取, 模型在构建的铁路故障语料库上实现了90.18%的Precision、90.91%的Recall与89.92%的F1值。将生成三元组数据由Neo4j图数据库存储。最后, 基于TransE、TransH、TransR和TransD四种嵌入算法开展链接预测对比, TransE在MRR (0.880)、Hits@10 (0.944)等指标上表现最佳, 选其作为推理核心。该方法能够有效整合维修数据, 提升维修决策的准确性与效率。

关键词

铁路运输, 内燃机车, 知识图谱, 决策推荐

Knowledge Graph and Decision Recommendation for Diesel Locomotive Fault Maintenance in Railway Systems

Ziying Yang^{1*}, Yongshun Wang², Xianglun Mo¹, Hanwen Zhang¹

¹School of Mining Engineering, China University of Mining and Technology, Xuzhou Jiangsu

²Xuzhou Metro Operations Co., Ltd., Xuzhou Jiangsu

Received: May 9th, 2025; accepted: Jul. 2nd, 2025; published: Jul. 10th, 2025

Abstract

In response to the challenges of fragmented data and information silos in the maintenance domain

*通讯作者。

文章引用: 杨子赢, 王永顺, 莫祥伦, 张瀚文. 铁路内燃机车故障维修知识图谱与决策推荐[J]. 交通技术, 2025, 14(4): 389-400. DOI: 10.12677/ojtt.2025.144040

of railway diesel locomotives, this paper proposes an intelligent decision-support approach based on knowledge graph construction and reasoning. First, a top-down strategy is employed, integrating on-site maintenance case studies to extract equipment models and fault locations using a custom dictionary and regular expression rules. For semantically complex elements such as fault causes, fault phenomena, and maintenance measures, a BERT-CRF model is introduced for sequence labeling, achieving a precision of 90.18%, a recall of 90.91%, and an F1 score of 89.92% on the constructed railway fault corpus. The resulting triplet data are stored in the Neo4j graph database. Finally, a comparative link prediction analysis is conducted using four embedding algorithms—TransE, TransH, TransR, and TransD—with TransE demonstrating superior performance in metrics such as MRR (0.880) and Hits@10 (0.944), and is thus chosen as the core inference model. This approach effectively integrates dispersed maintenance data, enhancing the accuracy and efficiency of decision-making and showing promising potential for practical application.

Keywords

Railway Transportation, Diesel Locomotive, Knowledge Graph, Decision Recommendation

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

铁路运输是国民经济与社会发展的命脉，内燃机车作为核心运输工具，其稳定性直接决定整个网络的效能。然而，内燃机车维修数据分散、信息孤岛严重，导致传统维修管理难以实现及时、准确的决策支持。

通用知识图谱在开放域数据融合方面具有优势，但其在工业领域的应用仍存在一定局限[1][2]；知识图谱已在医学、金融、矿业工程及铁路等领域广泛应用[3]-[7]。此外，基于规则、逻辑推理以及图谱表示学习的方法在复杂关系和推理决策中也取得了显著成果[8]-[10]。因此，本文通过融合知识图谱构建与推理技术，以提升维修决策的效率与准确性。

2. 知识图谱构建流程

知识图谱构建方法主要分为自顶向下构建和自底向上构建。自底向上方法多基于开放域大规模语料，通过弱监督或半监督学习自动生成图谱，但对专业领域术语及复杂上下文的适应性较差。与之相比，自顶向下方法依托领域专家知识，虽然需要先验本体设计，但在精度和可解释性方面具有明显优势。针对铁路内燃机车维修场景中的故障现象、维修措施等语义复杂要素，在自顶向下框架内，结合规则抽取与BERT-CRF模型，实现了90%以上的抽取精度与召回率，有效弥补了纯规则或纯深度学习方法的不足。

在矿区铁路内燃机车故障维修知识图谱构建过程中，数据采集作为关键的首要步骤，其质量和完整性决定了后续分析的可靠性。本文数据来源于某矿区铁路机务段，数据信息涵盖设备型号、故障记录、维修措施等内容。具体数据收集步骤如下：

(1) 现场数据获取：通过与矿区铁路机务段一线维修人员的现场交流，获得详细的维修记录。这些记录涉及内燃机车故障现象、故障原因、维修过程以及维修结果等关键信息。其次从设备管理部门借阅设备手册，手册中包含设备的参数、操作指南和常见故障的处理方法，为数据分析提供了技术支持。

(2) 数据数字化：将文字内容录入到可编辑的文本文件中。

经文本清洗与预处理后，开展实体与关系的分析与抽取。最后将实体与关系、属性等，按已构建的
本体关系链接到图数据库 Neo4j 储存并显示为有向图的形式。图谱构建流程如图 1 所示。

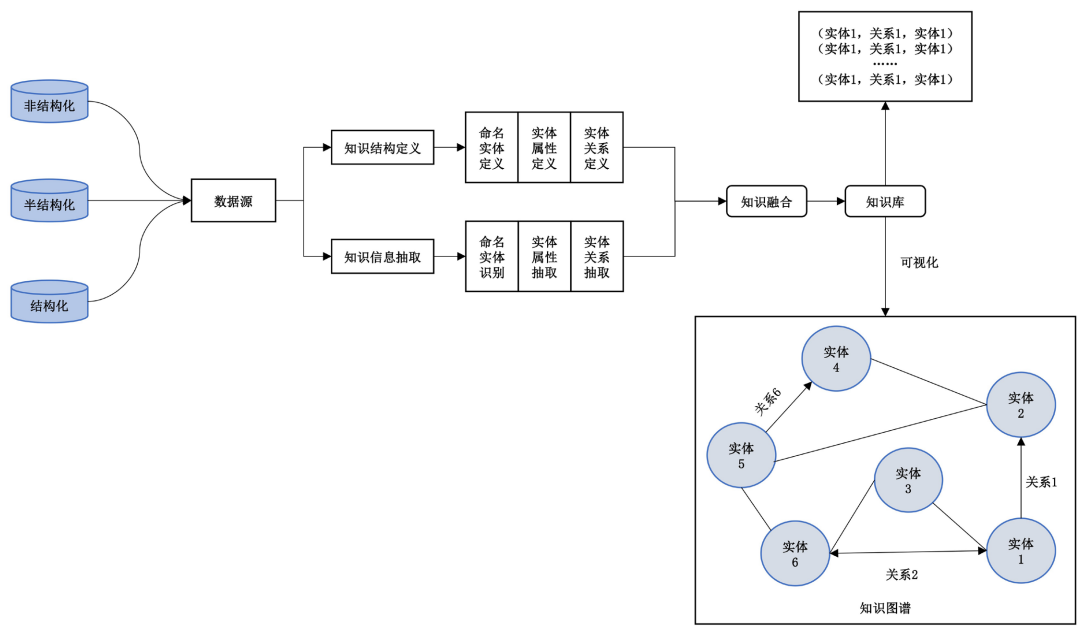


Figure 1. The construction process of the knowledge graph
图 1. 知识图谱构建流程

3. 知识建模

为构建铁路内燃机车故障维修知识图谱，需要对维修案例文本进行深入分析与全面标注，提取设备
信息、故障部位、故障原因、故障现象及维修措施等关键要素，为后续知识抽取与本体构建奠定基础。
案例文本分析如图 2 所示。



Figure 2. Maintenance case study
图 2. 维修案例分析

在对铁路内燃机车故障维修案例进行分析后，确定案例文本中的故障构成要素。实体是铁路内燃机
车故障维修的相关信息，如表 1 所示。

Table 1. Definition of entities
表 1. 实体定义

实体含义	中文含义	举例说明
EQUIPMENT_MODEL	设备型号	0080 机车
FAULT_LOCATION	故障部位	微机励磁

续表

FAULT_CAUSE	故障原因	水温高
FAULT_PHENOMENON	故障现象	卸载现象
MAINTENANCE	维修措施	更换撒砂脚踏阀后正常

关系是实体和实体之间的关系，本文定义了五种关系，如表 2 所示。

Table 2. Definition of relationship

表 2. 关系定义

关系	含义	关系说明
原因	Reason	{故障现象, 原因, 故障原因}
导致	Cause	{故障原因, 导致, 故障部位}
维修	Repair	{故障现象, 维修, 维修措施}
现象为	Phenomenon	{故障部位, 现象为, 故障现象}
属于	Belong	{故障部位, 属于, 设备型号}

4. 图谱构建

为实现原始文本到结构化知识的转化，本文采用基于规则和基于深度学习的两种混合抽取策略。

4.1. 基于规则的抽取

在规则抽取部分，设计了以下四步流程：

- (1) 文本预处理。对收集来的文本进行清洗处理，存储为 TXT 文件。
- (2) 词库构建。构建自定义词典，将领域专业词汇(如“0079 机车”、“微机励磁”、“励磁板”等)及其词性导入词典，并以 UTF-8 编码保存。
- (3) 分词及词性标注。基于上述构建的词库，对文本进行分词和词性标注。如“0079 机车”、“万向轴十字销”等词汇均被准确标注，为后续规则匹配提供语义依据。
- (4) 抽取规则构建。对于设备型号，规则采用\d{4}机车来匹配四位数字后缀“机车”的词组，如“0079 机车”。对于故障部位，则用多组上下文匹配规则，如匹配“经检查为(\w+)”、“更换(\w+)”、“检查(\w+)”等。通过对上下文关键词的匹配，提取紧随其后的内容作为故障部位。

4.2. 基于深度学习的抽取方法

深度学习方法能够捕捉长距离上下文依赖，适用于故障原因、故障现象及维修措施等复杂要素的抽取。

4.2.1. 文本序列标注

为了确保构建出的知识图谱在规模与质量上均达到较高水平，本文在内燃机车故障维修元素抽取过程中做出了如下改进：

- (1) 文本序列标注：文本序列标注是指对数据集中每条故障文本进行逐词标记，从而生成具有序列标签的语料库。为扩大语料库的范围，本文利用现有的矿区铁路内燃机车故障数据，同时整合其他来源的文本数据(如设备手册、操作指南等)。在文本序列标注工具的选择上，采用“精灵标注助手”对案例按 BIO

标注体系进行人工标注，构建语料库。具体的处理过程如图 3 所示。

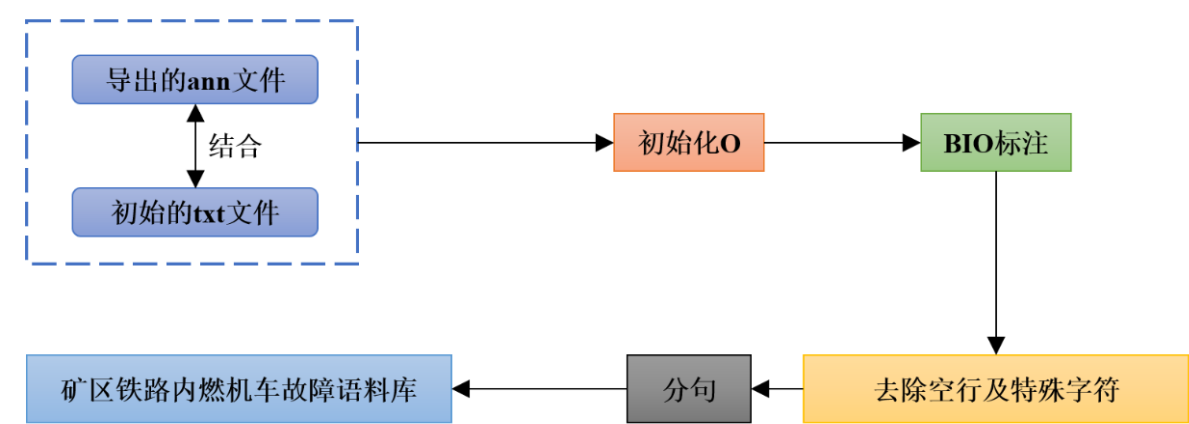


Figure 3. Corpus construction
图 3. 语料库构建

(2) 实体标注方法：为完成故障维修案例中关键元素(包括故障原因、故障现象、解决措施等)的抽取任务，需要对文本数据进行准确的实体标注。针对矿区铁路内燃机车故障文本数据，经人工初步分析后，定义了三种实体类型：故障原因、故障现象与解决措施。常用的标注体系有 BIO、BMES 和 BIOES 等。为了保证标注的清晰简洁，捕捉到领域内专业术语的细粒度信息，本文在传统 BIO 标注体系的基础上进行了适当的调整。具体的标注规则如下表 3 所示。

Table 3. Labeling rules
表 3. 标注规则

实体类别	起始标记	中间标记	结束标记
故障原因	B-fau	I-fau	I-fau
故障现象	B-phe	I-phe	I-phe
解决措施	B-mea	I-mea	I-mea

为了进一步降低纯人工标注的工作量，提高标注效率和一致性，引入半自动标注模块。利用训练好的 BERT-CRF 的模型，对原始文本数据进行自动预测，自动生成初始的实体标注结果。生成的自动标注可以通过使用标注工具进行复核和必要的校正，确保标注错误能被及时纠正。半自动标注流程在保证数据一致性和准确度方面具有显著优势，同时降低了标注工作的耗时。

(3) 语料库构建：完成文本序列标注与实体标注后，需要将生成的 ann 格式文件转化为符合 BIO 标注格式的最终语料库，部分示例如表 4 所示。

Table 4. Corpus
表 4. 语料库

词语	标注	词语	标注	词语	标注
0	O	因	O	修	O
0	O	为	O		O
7	O	整	B-fau	更	B-mea

续表

9	O	流	I-fau	换	I-mea
机	B-phe	二	I-fau	二	I-mea
车	I-phe	极	I-fau	极	I-mea
不	I-phe	管	I-fau	管	I-mea
走	I-phe	击	I-fau	即	O
车	I-phe	穿	I-fau	可	O
,	O	,	O	。	O
原	O	辅	O		

4.2.2. 深度学习模型及模型比较

文章在已构建的 BIO 标注语料上开展实验，旨在评估不同深度学习模型在故障记录实体抽取任务中的表现。为捕捉文本的基本语义信息，首先利用 Word2Vec 模型在语料上训练了 100 维词向量(窗口大小为 5，最小词频阈值为 3)。已构建的 BIO 标注语料上，采用 Word2Vec 训练了 100 维词向量(窗口大小 5，最小词频阈值 3)。基于预训练的 BERT-CRF 模型，设计了多种实体抽取模型，并在相同故障语料库上进行比较。利用 BERT-CRF 模型进行实体抽取，所有模型均统一设置学习率为 $2e-5$ 、Batch size 为 8，并在 60 个 Epoch 内完成训练。实验结果如表 5 所示。

Table 5. Model comparisons

表 5. 模型比较

模型	Precision	Recall	F1
BERT	88.30%	87.25%	87.77%
BERT-BILSTM	89.21%	88.38%	88.79%
Bert-CRF	90.18%	90.91%	89.92%
Bert-BILSTM-CRF	89.65%	89.80%	88.23%

实验结果表明，BERT-CRF 模型在复杂语境下对实体的抽取效果最佳，较其他模型显著提升了准确率和召回率。例如，对“0079 机车第二空压机不打风，经检查为 ZYC 接触器损坏，更换 ZYC 接触器后使用正常”这一记录，模型准确识别出“第二空压机不打风”(故障现象)、“ZYC 接触器损坏”(故障原因)及“更换 ZYC 接触器”(维修措施)，验证了其实用性。

4.3. 数据存储

完成三元组结构化转换后，为实现对故障维修知识的高效存储与检索，需将生成的三元组导入图数据库 Neo4j。Neo4j 数据库的数据导入方式主要包括两种：基于 Cypher 语句的逐条插入和通过 Import 工具实现批量加载。针对非结构化数据转化后的大量三元组存储需求，本文采用 Import 工具完成数据的高效入库。最终构建的部分知识图谱如图 4 所示。

在图谱中，深蓝色代表设备型号，褐色表示故障部位，紫色代表解决措施，红色代表故障原因，而绿色表示故障现象；关系线的颜色分别为：黄色表示“现象为”关系，蓝色代表“导致”或“原因”关系(可通过深浅区分细微差别)，深绿色表示“属于”关系，而粉色则对应“维修”关系。这种颜色区分不仅可以在海量数据中迅速定位和理解关键信息，还能直观地呈现系统中各要素之间的因果和归属关系。

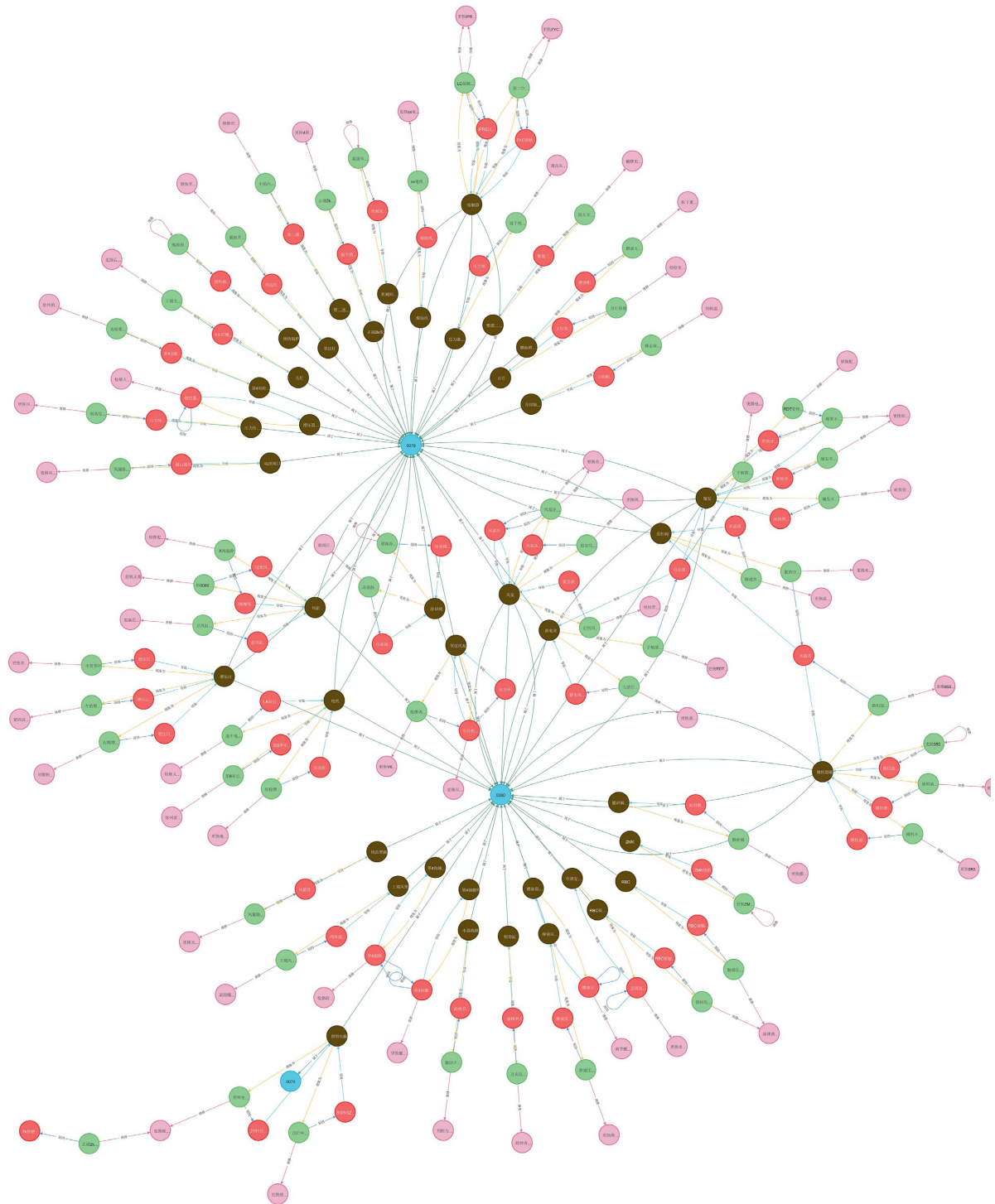


Figure 4. Knowledge graph visualization
图 4. 图谱可视化

5. 辅助决策推荐

知识图谱可以从历史故障维修记录中获取故障维修方法和故障原因等相关信息。然而，知识图谱的用途不仅限于查询数据库，本文将进一步研究链接预测任务，通过比较 TransE、TransH、TransR 和 TransD

四种嵌入算法在铁路内燃机车故障维修知识图谱上的表现,选取在 MRR、MR、Hits@1、Hits@3、Hits@10 等指标上表现最优的模型,作为辅助故障维修方案推荐的推理核心。

5.1. 实验过程及可视化分析

5.1.1. 实验数据预处理

基于构建好的知识图谱,利用 APOC 工具将实体间的关系以三元组形式展现,从 Neo4j 数据库中提取了所有实体与关系,导出结果如表 6 所示。

Table 6. Mapping table

表 6. 映射表

Entity		Relation		Entity_relation_entity		
名称	Id	名称	Id	头实体 id	关系	尾实体 id
0079	1	原因	0	127	1	126
0080	2	导致	1	224	2	1
ZMK 烧损	19	属于	2	216	3	190
冷却阀冷却单节漏水	29	现象为	3	209	0	200
...	...	...	...	...	...	...
照明电路的 8/24 与 15/22 阀断路	151	维修	4	174	4	65

5.1.2. 实验过程与可视化分析

首先,从知识图谱中提取所有正例三元组,为三元组中的每个头实体、尾实体和对应关系分别生成初始向量表示;接着,采用随机替换法生成负样本,对每个正例随机替换头实体或尾实体,构造出与正例数量等量的错误三元组。在超参数调优方面,采用网格搜索与随机搜索相结合的策略,对关键超参数(如学习率、批尺寸、嵌入向量维度以及训练的 Epoch 数)进行调参,以确保模型在不同配置下均能达到稳定的性能。例如,在学习率范围[0.001, 0.01]内进行初步搜索,最终确定学习率为 0.01 时,模型整体表现最佳。

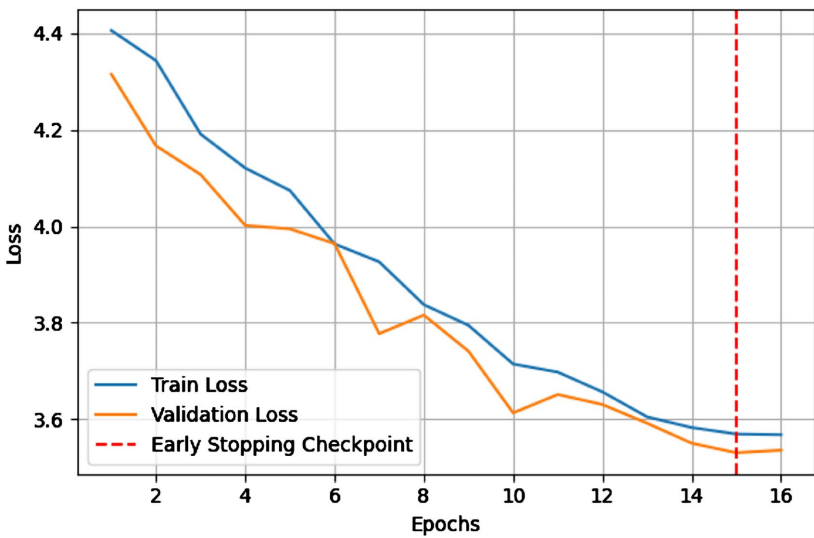


Figure 5. Loss function

图 5. 损失函数

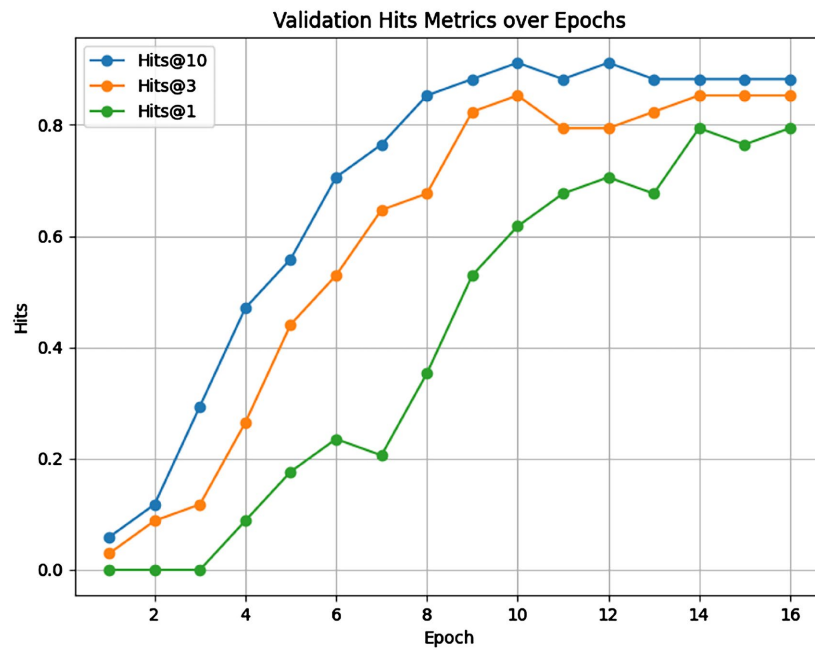


Figure 6. Hits@K metric

图 6. Hits@K 指标

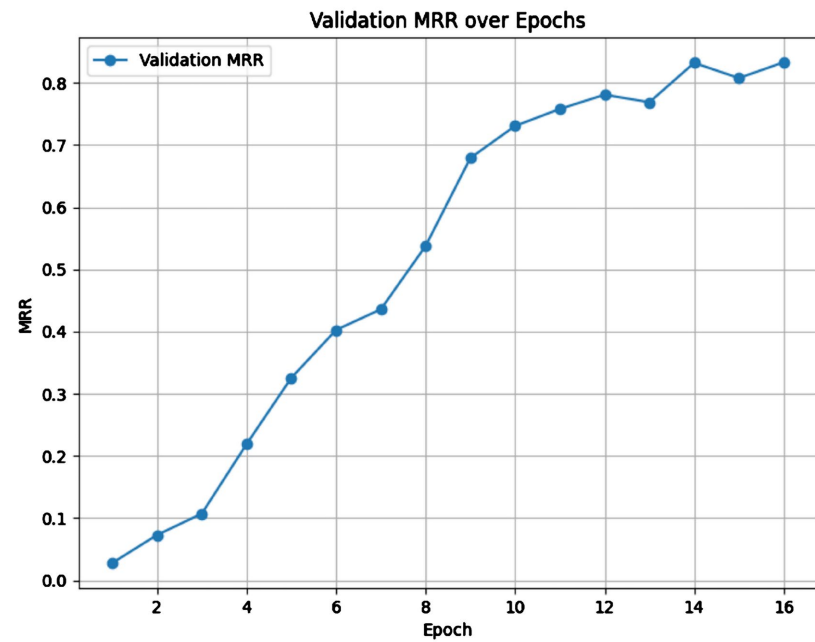


Figure 7. MRR indicator

图 7. MRR 指标

在训练过程中，对于每种嵌入算法(TransE、TransH、TransR、TransD)均采用相同的超参数设置，包括学习率 0.01、margin 值 4.5、Batch size16 以及训练 16 个 Epoch，并以 margin-based ranking loss 为目标函数，通过梯度下降等优化方法多次迭代，使正样本与负样本之间的距离差异达到最小化。

(1) 训练损失函数如图 5 所示。

损失曲线如图 5 所示, 显示了模型的收敛趋势: 训练集(蓝色)和验证集(橙色)损失均从 3.6 下降至 4.4 左右。第 15 个 Epoch 时, 早停策略生效, 验证损失保持稳定, 有效防止过拟合。这图直观展示了模型由快速下降至稳定收敛的过程。

(2) Hits@K 指标分析。

实验结果如图 6 所示, 随着训练轮次增加, Hits@K 指标稳步上升, 其中 Hits@10 在早期就达到较高水平, Hits@3 和 Hits@1 指标也持续改善, 证明模型能有效将正确答案排在靠前位置。

(3) MRR 指标分析

结果如图 7 所示, 随着训练轮次的增加, MRR 指标显著提升并趋于稳定, 这表明模型的排序能力得到了持续优化。总体来看, 这些量化结果验证了我们采用的训练策略、超参数设置以及负样本生成方法在提升模型语义匹配和排序性能方面的有效性, 为后续模型优化及实际应用提供了坚实的理论依据。

最后, 训练完成后, 通过 MRR、Mean Rank、Hits@10 和 Hits@3 等评价指标, 对各个模型在链接预测任务上的表现进行比较评估, 实验结果汇总如表 7 所示。

Table 7. Model comparison
表 7. 模型对比

模型	MRR	MR	Hits@10	Hits@3
TransE	0.880	6.111	0.944	0.917
TransR	0.143	49.556	0.389	0.194
TransH	0.813	12.833	0.917	0.889
TransD	0.694	11.083	0.889	0.861

实验结果显示, 在相同参数配置下, TransE 模型在链接预测任务中表现最佳。

5.2. 推理流程

辅助故障维修方法利用 TransE 嵌入模型进行链接预测。整个推理流程如下。

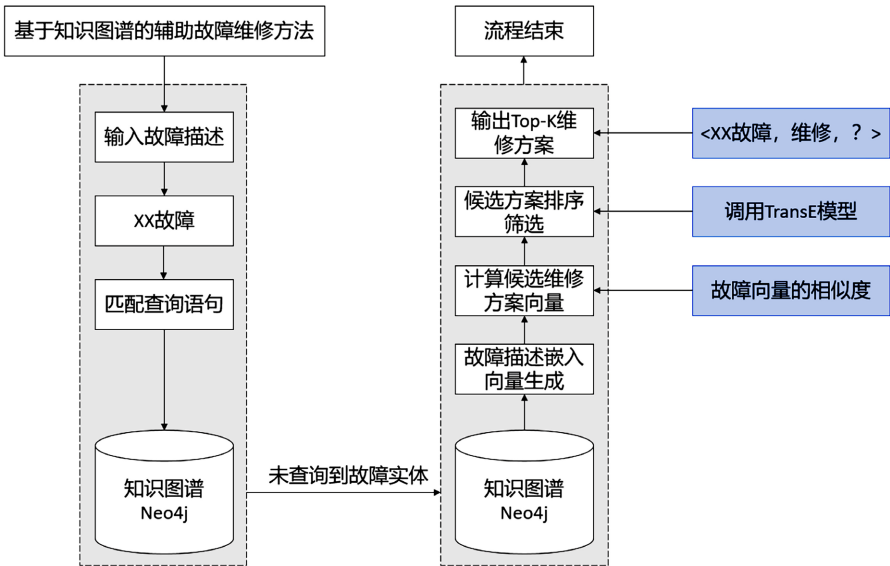


Figure 8. Inference process
图 8. 推理流程

- (1) 故障描述输入：接收文本形式的故障描述，进行分词及预处理，生成与知识图谱实体同构的向量表示。
- (2) 知识图谱匹配：首先在 Neo4j 中检索是否存在与输入描述完全匹配的历史维修记录；若无精确匹配，则计算输入描述向量与所有故障实体向量的余弦相似度，选取相似度最高的实体作为推理起点。
- (3) 链接预测与候选生成：将查询格式化为“<故障描述，维修，？>”，并在 TransE 嵌入空间中，依据头实体向量与关系向量之和与各尾实体向量的距离差异，生成维修措施候选列表。
- (4) 方案排序与输出：对候选维修方法按向量距离(或余弦相似度)进行排序；对距离分值归一化后，输出前三名推荐方案以辅助决策。推理流程如图 8 所示。

5.3. 案例分析

对于输入故障现象<辅发不发电，维修，？>，系统首先在 Neo4j 中尝试精确匹配历史记录；若未命中，则进入链接预测阶段，具体过程如下：

- (1) 历史记录匹配。系统利用 Neo4j 的图查询功能，查找与“辅发不发电”对应的故障节点及其历史维修方案。如果找到了完整匹配的历史记录，可直接返回相应的维修方法。
- (2) TransE 链接预测。当精确匹配失败时，系统对故障描述进行分词和预处理，生成与知识图谱中实体同构的嵌入向量。将输入故障节点作为起点，在整个知识图谱中筛选出所有候选维修方案对应的节点。使用 TransE 模型计算目标故障向量与各候选维修措施向量的距离(或余弦相似度)。对计算得到的距离分值进行归一化，使得不同尺度下的距离可比，再从小到大排序，最终输出 Top-3 推荐方案。如图 9 所示，图中直观展示了连接预测的路径。

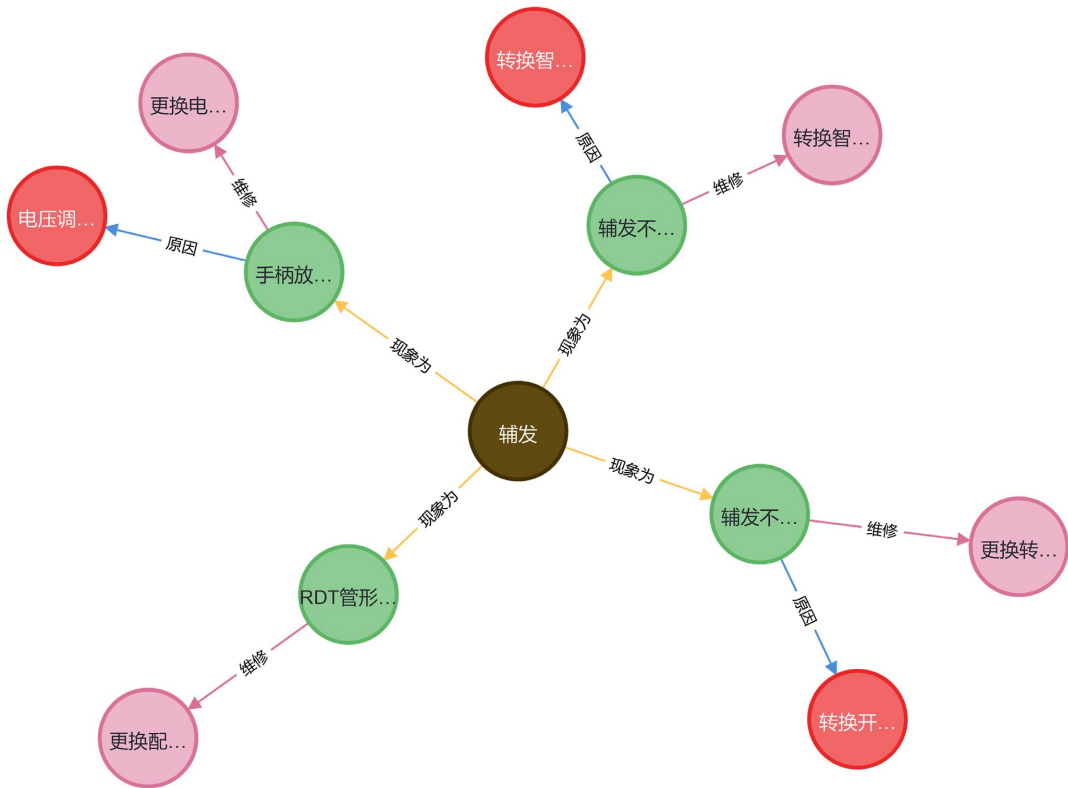


Figure 9. Link prediction path
图 9. 链接预测路径

根据排序结果，系统输出如表 8 所示。

Table 8. Inference result
表 8. 推理结果

<FAULT_PHENOMENON, Repair, ? >	TopK	Recommended Repair Methods
<辅发不发电, 维修, ? >	1	更换转换开关后正常
	2	更换配件后使用正常
	3	更换电压调整电阻 RDT 后正常

从结果可见，“更换转换开关后正常”与输入故障向量最为接近，同时也是历史维修数据中出现频次最高的方案，因此被排在第一位；“更换配件后使用正常”与第一方案具有很高的相似度，符合实际维修逻辑；“更换电压调整电阻 RDT 后正常”：虽然语义匹配略逊，但依然作为备选方案具备参考价值。

这种基于知识图谱嵌入和链接预测的混合方法，不仅发掘了故障与维修方案之间的潜在语义关联，同时依托历史记录的统计信息增强了推荐的准确性，为现场维修决策提供了高效且有力的辅助支持。

6. 结论

针对铁路内燃机车维修中存在的信息孤岛和故障诊断不精准的问题，本文提出了一种基于知识图谱构建与推理的辅助决策方法。首先采用自顶向下的策略设计领域本体，并结合规则抽取和 BERT-CRF 模型，实现对历史维修案例中设备型号、故障部位、故障原因、故障现象及维修措施等关键信息的自动化提取，将抽取结果以三元组形式存储于 Neo4j 图数据库中。其次，通过对 TransE、TransH、TransR 与 TransD 四种知识图谱嵌入算法的系统对比，验证了 TransE 在 MRR、Hits@10 等评估指标上的优越性，最终以其为核心实现链接预测和方案推荐，为维修人员提供高效准确的故障维修候选方案。

本研究在知识抽取准确度等方面还有待提高，未来将进一步优化抽取算法、丰富模型训练数据，实现更精准的维修方案推荐，从而为铁路内燃机车的智能维修提供更全面的数据支撑和决策依据。

参考文献

- [1] Rosyiq, A., Hayah, A.R., Hidayanto, A.N., Naisuty, M., Suhanto, A. and Avuning Budi, N.F. (2019) Information Extraction from Twitter Using Dbpedia Ontology: Indonesia Tourism Places. 2019 *International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, Jakarta, 24-25 October 2019, 91-96. <https://doi.org/10.1109/icimcis48181.2019.8985194>
- [2] Pellissier Tanon, T., Weikum, G. and Suchanek, F. (2020) YAGO 4: A Reason-Able Knowledge Base. In: Harth, A., et al., Eds., *The Semantic Web*, Springer, 583-596. https://doi.org/10.1007/978-3-030-49461-2_34
- [3] 吴涛. 基于知识图谱的肝脏疾病问答系统的研究与实现[D]: [硕士学位论文]. 太原: 中北大学, 2023.
- [4] 李伟光. 基于知识图谱的理财基金问答系统的研究与应用[D]: [硕士学位论文]. 长春: 吉林大学, 2024.
- [5] 耿念. 煤矿核心灾害知识图谱构建及问答研究[D]: [硕士学位论文]. 北京: 中国矿业大学, 2024.
- [6] 姚剑. 铁路客站设备健康管理 with 智能管控关键技术研究[D]: [博士学位论文]. 北京: 中国铁道科学研究院, 2023.
- [7] 吕志凡. 面向高铁车载设备故障诊断的知识图谱构建及应用研究[D]: [硕士学位论文]. 兰州: 兰州交通大学, 2024.
- [8] 孙建强, 许少华. 基于可微神经计算机和贝叶斯网络的知识推理方法[J]. 计算机应用, 2021, 41(2): 337-342.
- [9] Bordes, A., Nicolas, U., Alberto, G.D., et al. (2013) Translating Embeddings for Modeling Multi-Relational Data. *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Lake Tahoe, 5-10 December 2013, 2787-2895.
- [10] 孟小艳, 蒋同海, 周喜, 等. 一种改进的自适应知识图谱嵌入式表示方法[J]. 计算机应用研究, 2021, 38(1): 39-43.