面向列车检修领域的知识图谱构建与应用

张俊杰　姜仕军　刘伟东
（中车青岛四方车辆研究所有限公司，266031，青岛）

摘　要　[目的]知识图谱以其强大的语义处理能力和开放组织能力，为各个领域的知识化组织和智能应用奠定了基础。为了提高城市轨道交通列车检修业务的数字化和智能化水平，应建立面向列车检修领域的知识图谱。[方法]阐述了知识图谱技术的简介，采用自上而下的方法构建知识图谱。从本体设计、信息提取、知识映射、知识存储、知识融合5个方面分析了面向列车检修领域知识图谱的构建过程，并展望了该知识图谱在列车检修领域的应用前景。[结果及结论]基于该图谱，不仅能够有效解决列车检修数据竖井化问题，还可实现列车检修领域相关数据的统一管理，为列车检修提供“一车一档案”、智能检索、列车故障趋势分析等智能化服务。

关键词　城市轨道交通；列车检修；知识图谱；本体设计；智能化检修
中图分类号　TP392：U279
DOI：10.16037/j.1007－869x.2024.09.035

Construction and Application of Knowledge Graph for Train Maintenance Field

ZHANG Junjie，JIANG Shijun，LIU Weidong
（CRRC Qingdao Sifang Rolling Stock Research Institute Co.，Ltd.，266031，Qingdao，China）

Abstract　[Objective] With its powerful semantic processing and open organization capabilities，the knowledge graph lays the foundation for knowledge-based organization and intelligent applications in various fields. In order to improve the digitization and intelligence level of urban rail transit train maintenance，a knowledge graph for the train maintenance field should be established. [Method] The knowledge graph technology is briefly introduced，and a top-down approach is adopted to construct the knowledge graph. The construction process of the knowledge graph for the train maintenance field is analyzed from five aspects，i.e. ontology design，information extraction，knowledge mapping，knowledge storage，and knowledge fusion. The outlook for the application of the knowledge graph in the field of train maintenance is described. [Result & Conclusion] Based on the above graph，not only can the problems of train maintenance data silos be effectively solved，but also the unified management of relevant data in the field of train maintenance can be realized，and intelligent services such as 'one file for one train'，intelligent retrieval，and train fault trend analysis be provided for train maintenance.

Key words　urban rail transit；train maintenance；knowledge graph；ontology design；intelligent maintenance

随着城市轨道交通运营规模日益增大，与列车检修业务相关的知识总量呈爆炸式增长，传统的知识组织和管理方式已无法满足设备维保的需要。当前，城市轨道交通列车检修数据普遍以工单为最小单位存储在运营企业的资产管理系统中，采购数据存储在运营企业的采购系统中，列车基本信息则存储在其他的信息化系统中。大部分维修数据存储颗粒度大，竖井化问题严重，造成信息重复存储、更新不同步、关联性缺失、系统功能单一等问题。

为此，本文提出以列车的基本信息、维修工单信息、采购信息等为主要数据，基于列车功能树和位置树结构，构建支持列车检修业务的知识图谱，以实现对列车检修数据的统一有效管理及智能化应用。

1　知识图谱简介

知识图谱本质上是一种基于图模型的关联网络知识表达方式。知识图谱将实体抽象为顶点，将实体之间的关系抽象为边，通过结构化的形式对知识进行建模和描述，使之实现知识可视化。此外，知识图谱还可对海量知识进行智能化处理，进而形成大规模的知识库，支撑业务应用。

在知识图谱中，有2个最重要的概念：实体、关系。实体指的是现实世界中的事物，如人、地名、概念、公司等；关系则用来表达不同实体之间的某种联系，比如人—“居住在”—北京，张三和李四是“朋友”。知识图谱用多关系图中的不同节点来表达不同的实体，用不同的边来表达不同的关系。

现实中有很多场景非常适合用知识图谱来表达。比如一个社交网络图谱里，既可以有“人”的实

体,也可以包含“公司”实体。人和人之间的关系可以是“朋友”,也可以是“同事”。人和公司之间的关系可以是“现任职于”或“曾任职于”的关系。图 1 为基于多关系图的知识图谱案例。

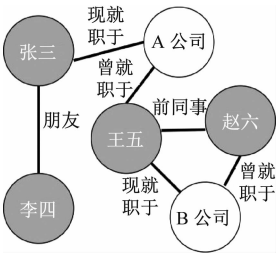

图 1　基于多关系图的知识图谱案例

Fig. 1　Knowledge graph case based on multi-relation graph

与传统的数据存储及计算方式相比,知识图谱的优势主要有:

1）关系的表达能力强。传统数据库通常需要关联表来表达关系,每种关系均需要新建关联表,这种模式不灵活,且难以支持复杂的关系网络。知识图谱基于图论和多关系图,可以处理复杂多样的关联分析,能够满足复杂关系场景下对数据的管理需求。

2）高速反馈。与传统数据存储方式相比,知识图谱采用图式的数据存储方式,数据调取速度更快,图库可计算超过百万个潜在实体的属性分布,真正实现人机互动实时响应(秒级返回结果)。

3）知识推理。知识图谱具有可解释性,可采用基于演绎的推理方式或基于归纳的推理方式,依据图谱中已有的事实或关系,推断出未知的事实或关系,进而推理出新的事实、新的关系、新的公理或新的规则,实现知识逻辑或模型的提取和沉淀。

2　知识图谱构建

针对列车检修领域数据种类繁多、关系复杂的特点,本文采用构建知识图谱的方式来实现列车检修数据的统一管理。本文采用自上而下的方法来构建知识图谱:

1）根据现有的结构化数据或专家知识库构造模式中的本体及各本体的相互关系,形成对应的概念模型和规则关系,构建出知识图谱的本体。

2）基于所构建本体中规定好的模式,从数据中抽取实体,对实体进行规则处理并完成基于图数据库[1]的存储,实现数据层的建设。

由此可见,自上而下构建知识图库的方法是从抽象到具体、先有概念而后有具体实现的构建过程,其构建过程可分为本体设计、信息提取、知识映射、知识存储、知识融合 5 个方面。

2.1　本体设计

构建知识图谱的第一步就是本体设计。本体是知识图谱的模型,是对构成知识图谱数据的一种模式约束[2]。构建一个本体,需要包括以下内容:①定义本体中的概念,将概念进行分层;②确定超类与子类关系;③定义概念的属性,并对这些属性的值予以限制;④为实例填充各属性值。

构建本体是一个创造性的过程,没有唯一正确的标准,其构建水平应通过实际应用来评价。根据对列车检修业务场景及数据的分析,本文采用功能树来描述列车各部件之间的功能所属关系,用位置树来描述具体部件之间的位置关系,并把具体部件信息与对应的位置信息进行绑定。基于功能树和位置树,得到城市轨道交通列车本体设计样例,如图 2 所示。

1）构建功能树。功能树以列车车型为单位,每种车型的功能树又细分为 3 个等级:第一级为车型,第二级是一级部件,第三级是二级部件,并根据需求继续拓展子部件。相邻级别的部件间均为双向关系,即互为对方的子部件或父部件。这样,每个部件均能追溯其父部件、子部件,进而实现全功能链的关联。在知识图谱构建时,对该本体实例化,1 种车型形成 1 个列车功能树。

2）构建位置树。位置树以具体列车为单位,定义了 1 列列车的位置结构图。1 列列车的位置树可分为 4 个等级:第一级是列车,第二级是车厢,第三级是位置,第四级是部件。组成 1 列列车的每个具体部件都要在位置树有具体位置,并与功能树的具体节点进行绑定,这样,每个部件安装在哪节车厢的哪个位置、部件在功能结构中的关系均可建立起来。如图 2 中编号为 01230000002 的转向架绑定在 0301 车的第 3 节车厢车底的位置 1,同时,该转向架属于 03A01 车型中支撑与走行系统中的转向架部件。

3）本体建模。对已有列车检修信息化系统里的数据进行本体建模,并与上面的功能树、位置树建立关联。以故障工单为例,提取工单编号、故障描述(离线)、故障状态、列车检修人员等作为实体,并将这些实体[3]进行相关性关联。这样,可以把列车从出厂到运行故障、列车检修、部件更换等全生

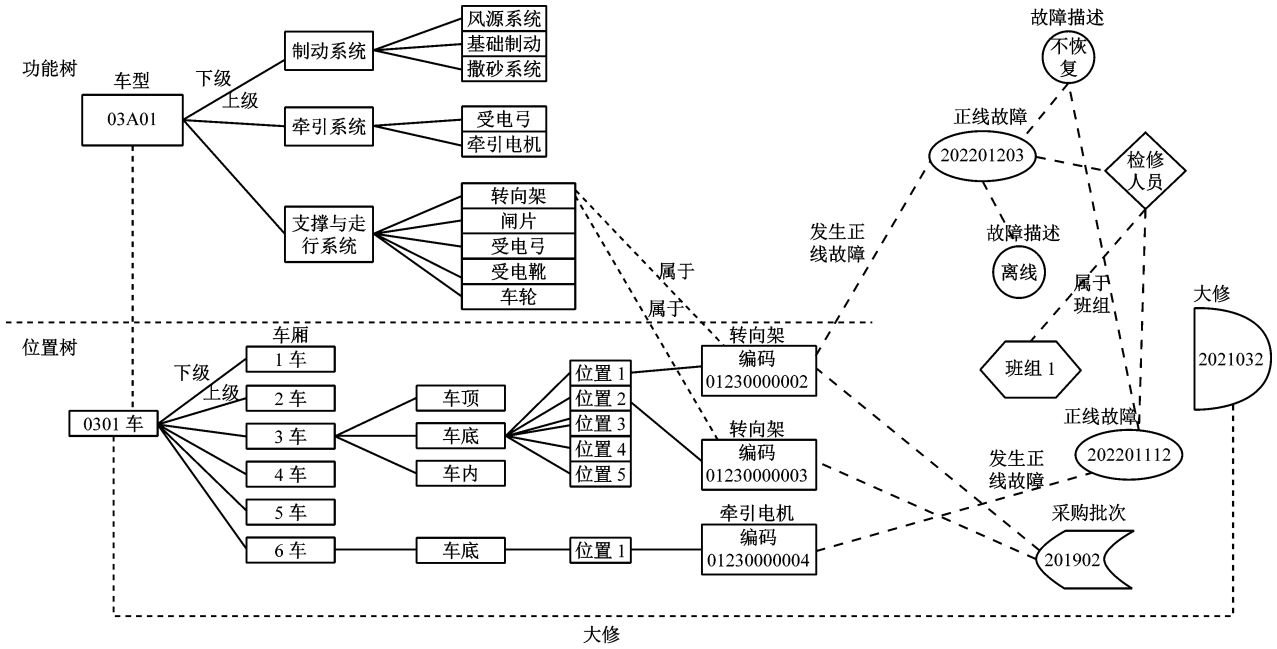

图 2　城市轨道交通列车本体设计样例示意图

Fig. 2　Schematic diagram of urban rail transit train ontology design example

命周期信息进行本体抽取并予以关联，一个面向列车检修的知识图谱本体设计就完成了[4]。

2.2　信息抽取

完成知识图谱的本体设计后，需要将所有业务数据按照设计好的本体来进行实体及关系信息的抽取[5]。以某地铁公司为例，其设备采购信息、列车检修派单信息、列车运行信息等分别存储于多个不同的信息化系统中，且以半结构化和结构化数据为主，需要提供插件化、配置化、可扩展的信息抽取模块，结合相应的配置及源数据，自动完成信息的抽取。对于结构化数据，只需根据原信息化系统格式进行定制化抽取并进行相应转化即可。

在半结构化数据中（特别是在工单数据的列车故障描述中），提取相应关键字来形成实体，是信息抽取的一个难题。本文以工单数据中的列车故障描述关键字为例，阐述半结构化数据文本中关键字抽取的方式[6]。工单数据的文本信息具有数据量大、专业词汇多等特点，本文采用“人工校对 + 机器学习”的方式实现信息抽取。

图 3 为信息抽取流程图，其步骤主要包括：

步骤 1　从 10 000 个工单中随机抽取 2 000 个工单作为数据的语言材料（以下简称“语料”），再从这 2 000 个工单中随机抽取 200 个样本，通过人工方式对故障描述中的关键字进行标注。

步骤 2　把完成人工标注的 200 个工单数据作为正样本，对所有样本数据进行半监督学习，并实现对关键字的自动标注。

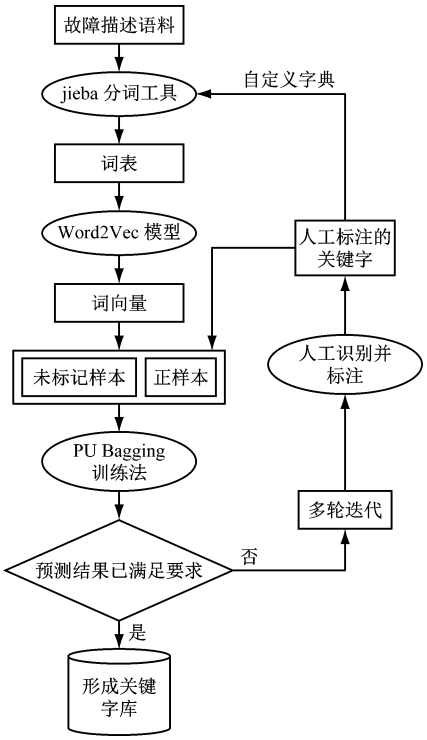

图 3　信息抽取流程图

Fig. 3　Flow chart of information extraction

步骤 3　对机器自动标注的关键字进行人工校对，并记录机器自动标注的准确率。人工校对时，系统会提供一个网页版的标注工具，以提高人工核

对的效率。

步骤 4　以上过程反复进行几轮，根据自动标注的准确率来确定最终的关键字库。

进一步阐述每个步骤的处理过程：

1）人工标注。本文开发 1 个网页版的简易标注工具，通过该工具，用户可在线阅览每个工单的故障描述文本，并可实现对关键字的人工标注。将人工标注结果录入正样本，并自动保存到自定义字典中。

2）文本处理。对于原始故障描述语料，首先读取其中的文本内容，并进行去除空格、换行符等操作；针对城市轨道交通列车维修工单语料专业词汇多的特点，基于自定义字典，使用分词工具jieba 对文本内容进行分词操作，将每篇语料表示成一组词表；采用所有语料的词表训练模型 Word2Vec，再使用训练得到的模型将所有语料词表中的每个词都转换成词向量，这样，所有故障描述语料均存储在同一个词向量表中。

3）基于 PU Bagging 法进行训练及预测。PU Bagging 是一种使用正例样本和未标记样本进行训练的半监督学习方法。将上一步生成的词向量分为正样本和未标记样本 2 个部分，随机抽取未标记的子集和所有正样本，创建 1 个均衡的训练集（采用“bootstrap”数据库格式）；用 “bootstrap”数据库构建 1 个分类器，将正样本视为 1，将 unknown（未知）样本视为 0；将被采样的 unknown 样本称为 OOB（袋外样本），预测 OOB 在训练集中的概率，并重复多次此流程，计算 OOB 的平均分。

4）形成关键字库。经过以上流程的多轮迭代后，若预测准确率已达到 95%，则基于该模型对所有工单数据进行分析，识别出所有关键字，形成关键字库。后续对每个工单进行信息抽取时，都与该关键字库进行匹配，匹配成功则识别为“实体”，匹配不成功则不是“实体”[3]。

2.3　知识映射

经过信息抽取后的数据，需要结合已定义的本体进行知识映射，完成内部知识的统一表示。统一的内部表示有利于多源知识融合、知识图谱应用推理等。实现知识映射的关键是：已完成信息抽取的三元组的每个节点或关系，均需要与本体中相应的节点或关系建立关联。因本文研究的实体均属于城市轨道交通列车领域，节点和关系均已知，因此，可采用人工定制 Map（键值对）文件的方式实现映射，最后通过代码实现实体和本体的映射关系，最后将所有的映射关系统一存储在图数据库 Neo4j 中。

2.4　知识存储

图数据库 Neo4j 可实现知识存储[7]。传统关系型数据库在处理大数据复杂关系问题时，会产生一系列的问题，如数据表结构复杂且灵活度低、关联查询效率低下、储存空间浪费、数据库应用扩展困难等。与关系型数据结构不同的是，图数据库 Neo4j 以“节点”和“边”为基本存储单元存储数据，利用节点之间物理上相互“指向”的特点，对相邻节点提供“无索引”的关联操作。由此，图数据库 Neo4j 具备存储数据类型多样、处理关联数据表现好、操作简单等优点，是专门用于处理海量关系的数据库。

本体的三元组数据与实体的三元组数据进行映射后，全部数据均存储在图数据库 Neo4j 中，实现海量关系数据的统一存储。

2.5　知识融合

在列车检修领域，对同一个实体或概念的描述信息可能存在多个数据源，此时需要采用知识融合措施，对来自不同数据源的知识进行异构数据整合并消歧。本文选用属性计算融合方法进行知识融合。每个知识都有自己的属性列表，相同知识在融合时均会比较属性之间的差距。当该差距小于设定的阈值时，知识会进行融合，否则知识将不融合。知识融合的主要计算方法有：

1）余弦距离。利用大规模 Word2Vec 分布式和 TFIDF（词频-逆文档频率）权重对属性字符串进行文本表示，然后对比 2 个属性字符串间的余弦距离。

2）局部敏感哈希。局部敏感哈希可实现海量高维数据的快速近似查找，通过比较数据点之间的距离或相似度来确定是否进行信息融合。

3）编辑距离。对 2 个属性字符串的差异程度进行量化量测，其量测方式是看至少需要多少次的处理，才能将 1 个字符串变成另 1 个字符串。

4）杰卡德距离。采用杰卡德距离来衡量 2 个集合的差异性时，先将文本分词，将数据变成文本集合后再进行计算。

5）恒等式。该方法的理念为只有 2 个属性完全相同的知识才能进行融合。

综合上述各知识融合方法，对列车检修领域几

个典型实体属性设置了相应的知识融合方法，其阈值设置如表 1 所示。

表 1　几个典型列车实体属性的知识融合方法及阈值设置

Tab. 1　Several typical knowledge fusion methods for train ontology attributes and threshold setting

实体属性	计算方法	操作符	阈值
列车编号	恒等式	=	0
列车故障名称	编辑距离	<	1
列车故障描述	余弦距离	>	0.92
列车故障处置方案	编辑距离	<	1.20

3　知识图谱在列车检修中的应用

3.1　建立“一车一档案”制度

上文所述的知识图谱覆盖了城市轨道交通列车的各业务阶段，记录了列车采购信息、列车运行信息、列车维修信息、列车配置信息等核心业务数据，进而建立了“一车一档案”制度。“一车一档案”包含的信息主要包括列车的静态履历信息、故障信息、采购信息、部件拆装记录、工艺信息、质量信息、生产工单信息及耗用物料信息等，可实现列车全生命周期数据的统一管理。

3.2　智能检索文档及数据

以知识图谱及 AI（人工智能）算法为核心，对各类检索语句或关键词进行意图识别，对知识图谱的数据及推理出的隐含知识进行信息匹配，并形成符合业务需要的内容展示形式，方便业务人员快速检索列车全寿命周期中的相关文档和数据，掌握列车运维过程中的相关情况。

3.3　分析列车故障趋势

基于知识图谱强大的关联数据查询能力，可轻松找出列车的高发故障类型、高磨耗零部件种类等信息，对列车检修质量、部件故障情况进行趋势分析，形成趋势图。知识图谱的趋势分析结果可以为工作人员全面了解列车的检修整体状况及故障趋势提供数据支撑。

4　结语

本文构建了面向列车检修领域的知识图谱，并基于该图谱建立了“一车一档案”，实现了智能检索、趋势分析等功能在列车检修业务中的应用。面向列车检修领域的知识图谱不仅能够有效解决数据竖井化问题，实现列车检修领域相关数据的统一管理，促进列车检修业务的数字化，还可基于知识图谱的知识推理能力，提供智能检索、故障趋势分析等更多形式的智能化服务。

参考文献

[1] 陆晓华，张宇，钱进. 基于图数据库的电影知识图谱应用研究[J]. 现代计算机，2016(7)：76.
LU Xiaohua，ZHANG Yu，QIAN Jin. Implementation of movie knowledge graph based on graph database[J]. Modern Computer，2016(7)：76.

[2] 项威. 事件知识图谱构建技术与应用综述[J]. 计算机与现代化，2020(1)：10.
XIANG Wei. Reviews on event knowledge graph construction techniques and application[J]. Computer and Modernization，2020(1)：10.

[3] 祝锡永，吴炀，刘崇. 基于 CTD-BLSTM 的医疗领域中文命名实体识别模型[J]. 计算机系统应用，2020，29(8)：173.
ZHU Xiyong，WU Yang，LIU Chong. Chinese named entity recognition in medical field using CTD-BLSTM model[J]. Computer Systems & Applications，2020，29(8)：173.

[4] 杜亚军，吴越. 微博知识图谱构建方法研究[J]. 西华大学学报(自然科学版)，2015，34(1)：27.
DU Yajun，WU Yue. Research on constructing the knowledge graph based on microblog[J]. Journal of Xihua University (Natural Science Edition)，2015，34(1)：27.

[5] 杨玉基，许斌，胡家威，等. 一种准确而高效的领域知识图谱构建方法[J]. 软件学报，2018，29(10)：2931.
YANG Yuji，XU Bin，HU Jiawei，et al. Accurate and efficient method for constructing domain knowledge graph[J]. Journal of Software，2018，29(10)：2931.

[6] 张华丽，康晓东，李博，等. 结合注意力机制的 Bi-LSTM-CRF 中文电子病历命名实体识别[J]. 计算机应用，2020，40(增刊1)：98.
ZHANG Huali，KANG Xiaodong，LI Bo，et al. Medical Name entity recognition based on Bi-LSTM-CRF and attention mechanism[J]. Journal of Computer Applications，2020，40(S1)：98.

[7] 姜惠娟，郭文龙. 基于 Neo4j 的药膳方图数据库设计与优化[J]. 中央民族大学学报(自然科学版)，2019，28(3)：48.
JIANG Huijuan，GUO Wenlong. Design and optimization of medicated diet's diagram database based on Neo4j[J]. Journal of Minzu University of China (Natural Sciences Edition)，2019，28(3)：48.

· 收稿日期：2022-05-12　修回日期：2022-06-22　出版日期：2024-09-10
Received：2022-05-12　Revised：2022-06-22　Published：2024-09-10

· 通信作者：张俊杰，高级工程师，zhangjunjie612@163.com