

一种港口危险货物信息知识图谱构建技术的研究

谢佳杰¹ 陈昱丞² 刘玥骅² 马佳涛¹ 陈超¹

1. 浙江海港嘉兴港务有限公司, 中国·浙江 嘉兴 314200

2. 中交第三航务工程勘察设计院有限公司, 中国·上海 200032

摘要: 港口危险货物种类繁多、理化性质各异、储存和运输规则琐碎复杂, 且其在储运过程中具有高风险性, 一旦发生事故, 会造成严重的人员伤亡、经济损失与环境污染。本文研究了一种以危险货物信息为主的知识图谱构建技术。首先根据危险货物的知识体系进行了要素分析, 构建了港口危险货物信息知识图谱的模式层; 在此基础上, 选择基于规则和模板的方法对实体进行抽取, 并制定了关系抽取的规则模板进行匹配。该方法可以将危险货物的知识形成结构化的语义网络, 简化查询程序, 完善查询结果。除此之外还可以作为后续危险货物信息智能问答系统的数据基础。

关键词: 知识图谱; 命名实体识别; 危险货物; 关系抽取

Study on a construction technique of Knowledge Graph for Port Dangerous Goods Information

Xie Jiajie¹, Chen Yucheng², Liu Yuehua², Ma Jiatao¹, Chen Chao¹

1. Zhejiang Seaport Jiaxing Port Affairs Co., Ltd, Zhejiang Jiaxing 314200

2. CCCC Third Harbor Consultants Co., Ltd, China Shanghai 200032

Abstract: The port handles a wide range of dangerous goods, each with distinct physicochemical properties and complex storage and transportation regulations. this study first establish the schema layer of this knowledge graph. Furthermore, this study employ rule-based and template-based methods for entity extraction and formulate rule templates for relationship extraction. This methodology transforms dangerous goods knowledge into a structured semantic network, simplifying the querying process and improving query results. Additionally, it also lays the foundation for an intelligent question-answering system specific to dangerous goods information.

Keywords: Knowledge graph; Named entity recognition; Dangerous goods; Relation extraction

0 引言

危险货物种类繁多, 其中纳入《危险化学品目录》的危化品就达到了 2828 个条目。危险货物的储存和运输条件要求严格, 并且其理化性质与储运规则相关知识零碎, 从而导致在港口生产过程中经常会发生不安全的违规行为。一旦发生事故, 会造成严重的人员伤亡、经济损失与环境污染。这不仅形成了巨大的安全风险, 而且增加了监管机构的压力。因此, 通过更好地表达和关联这些知识、简化查询程序、实施智能化监管危险货物堆存情况以确保危险货物储运过程中的安全, 对于港口安全生产具有重要意义^[1]。在数据库的基础上, 已有大量研究人员将知识图谱技术应用于港口危货品管理。Zhang 等人^[2]基于 IMDG 规则构建了海运危货品知识图谱, 有助于简化危货品专业知识的检索过程, 实现货物储存和隔离的自动判断。陈鹏冲等人^[3]

基于“人机料法环”方法, 对海关危化品实验室风险因素进行识别和分类, 构建了危化品实验室安全检查风险知识图谱。

本文针对危险货物种类众多, 理化性质、储存隔离知识琐碎等问题, 使用权威数据来源, 分析风险因素, 制定知识抽取策略, 研究了一种港口危险货物信息知识图谱的构建技术。

1 知识图谱的概念

知识图谱是一种基于图数据结构的语义知识库。基本组成为“实体-关系-实体”三元组, 以及实体及其相关属性-属性值对, 实体间通过关系相互联结, 构成网状的知识结构。从图的角度来看, 三元组中的实体是图中的节点, 而关系则为图中的边。知识图谱能够改变现有的基于简单的字符串匹配的信息检索方式, 通过其结构化的语义

网络可以实现推理从而达到复杂的语义搜索功能^[4]。

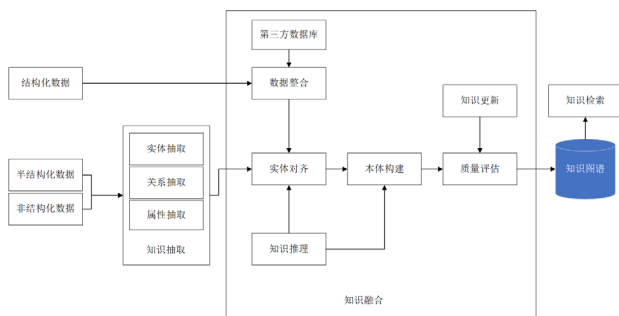


图1 知识图谱的体系架构

知识图谱的体系架构如图 1 所示。知识图谱的构建有三个步骤：(1) 知识抽取；(2) 知识融合；(3) 知识加工。其中知识抽取是指从多源异构数据源中提取出实体、属性以及实体间的关系的过程，包括实体抽取和关系抽取。知识融合指的是消除矛盾和歧义，将来自不同数据源的知识在同一规范下进行整合、对齐、加工，例如实体对齐就是指消除异构数据中实体冲突、指向不明等不一致性问题，创建一个统一的实体库。

2 港口危险货物信息知识图谱构建

港口危货品管理规则与危险化学品信息内容完整丰富，知识体系结构清晰，所以本文将采用自顶向下的方式。首先构建模式层，对实体概念、实体属性、实体间关系进行描述。再通过知识抽取技术构建数据层。

2.1 港口危险货物信息要素

本文以港口危险货物本身的信息为核心，采用“危货品、危险货物分类、危险货物属性、物理危险、健康与环境危害、消防措施、隔离规则”七类关键因素作为基本维度。根据这七类关键因素，可以分析得到所有的实体类型与实体属性类型，如图 2 所示。

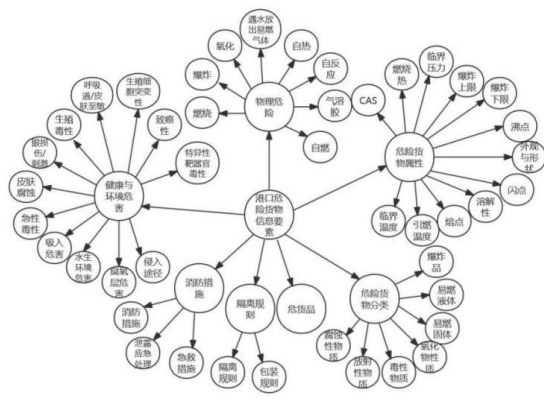


图2 实体类型示意图

2.2 知识抽取

知识抽取包括实体抽取和关系抽取，其中实体抽取又

可以叫作命名实体识别，关系抽取包括实体与实体之间的关系和实体与实体属性值之间的关系。

2.2.1 实体抽取

当前危险货物管控领域缺乏权威数据集，采用机器学习与深度学习的方法难以获得标注之后的训练集语料。袁金斗等人^[5]研究了基于规则和词典的用电安全领域命名实体识别方法。本文采用实体词典 + 特征字符规则匹配 + 词性组合规则匹配结合的方法完成实体抽取。技术架构如图 3 所示。

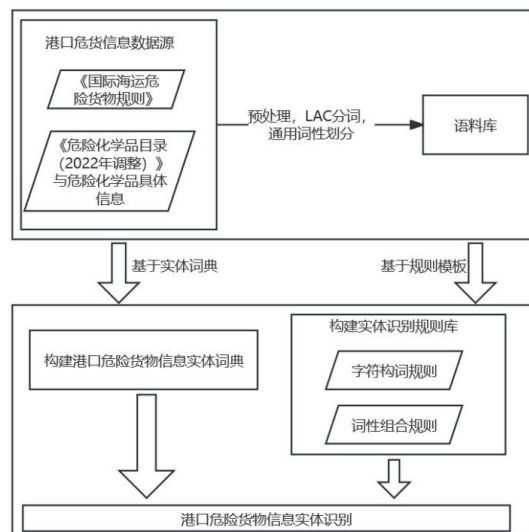


图3 实体抽取架构图

首先按照实体类别人工建立港口危险货物信息实体词典，词典中存在的实体不再被拆分。《危险化学品目录》中罗列的危险化学品及其 CAS 号可以直接作为实体加入词典中，其余实体通过人工标注的方式加入词典中。词典的构建完成之后，使用基于词典的正向最大匹配算法进行第一轮实体识别，得到实体集合 1。

通过基于构词特征字符规则的实体识别方法完成第二轮的实体识别。BIOES 标注规则是 Begin, Inside, Outside, End, Single 的缩写。

上文中定义的七大实体类别中，危货品与危货品分类实体类中的实体已确定，其余五大实体类及其标签为，危险货物属性 (PROP)、物理危险 (DGR)、健康与环境危害 (HAZ)、消防措施 (MEA)、隔离规则 (IR)。根据 BIOES 标注规则构建字符标签集合 LabelSet={B-PROP, I-PROP, E-PROP, S-PROP, B-HAZ, I-HAZ, E-HAZ, S-HAZ, B-DGR, I-DGR, E-DGR, S-DGR, B-MEA, I-MEA, E-MEA, S-MEA, B-IR, I-IR, E-IR, S-IR, O}。从数据源中划分训练集，对常见实体按照 BIOES 模式序列化

标签进行标注。例如对于措施类实体的标注：“脱去污染的衣着”“用肥皂水和清水彻底冲洗皮肤”，其中脱去标注为 B-MEA，污染标注为 I-MEA，衣着标注为 E-MEA，肥皂水标注为 B-MEA，清洗标注为 I-MEA，皮肤标注为 E-MEA。标注完成之后，将其按照数组的形式存储。例如，物理危险实体的前缀（字符标签为 B-DGR）的字符存储在 begin_dgr[] 中，后缀存储在 end_dgr[] 中。然后采用基于字符的序列标注匹配规则算法进行实体识别，可以得到实体集合 2。

接下来通过基于词性组合特征规则的实体识别得到实体集合 3。该方法首先构建词性规则组合模版，分别研究四类实体常见的词性组合特征，然后得到正则表达式。例如对于消防措施，其前缀通常为动词 (v)、动名词 (vn) 或名词 (n)，后缀为 E-MEA，所以规则模版为：(v*)(vn*)[E-MEA]、(n*)[E-MEA]。使用 LAC 工具对训练语料进行分词以及词性标注，之后进行正则匹配的方式进行实体识别，可以得到实体集合 3。融合三个实体集合，得到最终的实体集合。

2.2.2 关系抽取

关系抽取是指利用语言学、统计学、信息科学等学科的方法技术，从文本中发现实体间的语义关系。关系抽取包括实体之间的关系抽取和属性抽取，实体之间的关系描述两个实体的关联状态，实体属性描述实体内部的信息。

任安琪等人将关系抽取分为句子级和文档级两类进行综述。根据本文知识图谱构建所采用的数据源，其危险货物的理化性质、危险情况、救援措施都是呈现在一句话或者一段话中的，所以本文采用句子级的关系抽取方法。通过上文对数据源的处理，可以得到实体集合与描述这些实体的相关语句（包含了实体的属性与关系）。由于知识图谱是一个具有语义信息的网络结构，其储存需要采用图数据库，目前 Neo4j 图数据库为主流的知识图谱存储数据库。

车金立等人研究的军事装备知识图谱通过挖掘每个实体的所属类别与每个实体的介绍得到实体间的关系。根据危险品实体的分类，从上到下的创建索引并建立索引间的隶属关系，上文将危险品分为 7 个大类，定义大类和危险品的关系为属于。按照分类依次导入上文中得到的实体，可以得到关于危险品之间的分属关系。

本文采用基于规则的方式抽取关系，首先根据实体定义关系抽取的规则，然后对文本执行模式匹配。危险品的分属关系规则可以不用定义，直接按照分类导入即可。

关系抽取的流程如下：

表1 算法流程

算法1 关系抽取算法	
输入：	分词处理之后的待抽取文本Text，关系抽取规则集合rules，七大类实体集合entities
输出：	N行3列的二维数组形式存储的三元组
1	Relations=[];初始化关系集合
2	for rule in rules:#对于每一条规则，使用正则表达式在文本中找到匹配
3	pattern = re.compile(rule['pattern'])
4	matches = pattern.finditer(text)
5	for match in matches:
6	entity1 = match.group('entity1')
7	relation = rule['relation']
8	entity2 = match.group('entity2')
9	relations.append((entity1, relation, entity2))
10	return relations

上述算法流程中的 rules 为正则表达式集合。部分规则的正则表达式如表 2 所示。

表2 部分正则表达式

关系规则	正则表达式
危险品-燃烧热-燃烧热值	(?P<entity1>[\u4e00-\u9fa5]+\s*-\s*燃烧热值\s*-\s*(?P<value>\d+(\.\d+)?)\s*kj/mol
危险品-爆炸-爆炸情况	(?P<entity1>[\u4e00-\u9fa5]+\s*-\s*爆炸情况\s*-\s*(?P<description>.+)
危险品-吸入危害-危害情况	(?P<entity1>[\u4e00-\u9fa5]+\s*-\s*吸入危害\s*-\s*(?P<description>.+)

通过 Python 中的库函数可以实现基于规则模版匹配的关系抽取。关系抽取的结果由数组的形式存储，之后将其转换为 3 列的表格形式，添加至 Neo4j 数据库中，完成实体之间的链接，最终完成港口危险货物信息知识图谱的构建。

3 结语

本文研究了使用基于规则和模版的知识抽取方法的港口危险货物信息知识图谱构建技术，通过对《国际海运危险货物规则》以及《危险化学品目录》中的危险化学品信息进行分析，得到了港口危险货物信息的实体模型，其中包括 7 个实体大类和 44 个实体小类，采用实体词典 + 特征字符规则匹配 + 词性组合规则匹配结合的方法进行实体抽取，可以得到实体集合。根据句子级的关系抽取要求，本文定义了 34 条关系抽取的规则，采用基于抽取规则匹配算法得到实体间的关系三元组，研究了一套完整的港口危险货物信息知识图谱的构建技术。本文研究的港口危险货物信息知识图谱构建技术可以实现危险货物知识的语义网络化表示，清晰地展现知识结构，可以作为后续危险货物检索系统、问答系统的数据基础。本文也存在一些问题，面对日新月异的危险货物知识，存在着词典更新不及时，规则模板冗余不全面，抽取结果的召回率与准确率下降的问题，未来可以结合基于深度神经网络的知识抽取方法进

行深入研究。

参考文献:

[1] 王少雄, 何静, 昌雪玲等. 知识图谱技术在港口多元危险货物储存中的应用[J]. 职业卫生与应急救援, 2022,40(3):341-346+380.

[2] Zhang Q ,Wen Q Y , Han D , et al. Construction of knowledge graph of maritime dangerous goods based on IMDG code[J].The Journal of Engineering,2020,2020(13):361-365.

[3] 陈鹏冲, 陈举师, 赵雨薇等. 海关危化品实

验室安全检查知识图谱构建方法[J]. 工业安全与环保, 2023,49(9):29-33.

[4] 刘峤, 李扬, 段宏等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016,53(3):582-600.

[5] 袁金斗, 潘明明, 张腾等. 基于规则和词典的用电安全领域命名实体识别[J]. 电子技术应用, 2022,48(12):22-27.

作者简介: 谢佳杰 (1990-), 男, 汉族, 浙江嘉兴人, 助理工程师, 大学本科, 研究方向: 港口系统集成、智能化、数字化改造。