Computer Science and Application 计算机科学与应用, 2024, 14(6), 137-148
Published Online June 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.146150

Hans汉斯

产业污染管理知识图谱构建与应用：以江门市为例

薛子如[1]，高　乐[1]，贾旭东[2]，陈　涛[1*]

[1]五邑大学电子与信息工程学院，广东 江门
[2]加州州立大学北岭分校计算机科学与工程学院，美国 北岭

收稿日期：2024年5月22日；录用日期：2024年6月21日；发布日期：2024年6月28日

摘　要

随着工业化和城市化的快速发展，产业污染问题日益严峻，对环境和人类健康构成了严重威胁。产业污染数据具有多源性和异构性，需要通过知识提取和融合技术，才能应用于产业污染治理和决策支持。本文提出利用知识图谱技术整合和建模产业污染领域的关键数据、实体和关系，以实现产业污染知识的多维度展示，包括概念、属性和实例等。以中国广东省江门市产业污染为例，本文通过知识提取、本体构建和知识存储，对公司、污染物、产品等信息进行了综合处理，构建了一个较为全面的产业污染知识图谱。实验结果表明，本文提出的知识图谱构建方法不仅能够有效且直观地揭示污染场地数据之间潜在关联，而且能为决策者提供数据支持和决策参考，同时也为相关研究和应用领域提供了共享数据。

关键词

知识图谱构建，产业污染，本体构建，环境治理

Construction and Application of Industrial Pollution Management Knowledge Graph: Taking Jiangmen City as an Example

Ziru Xue[1], Le Gao[1], Xudong Jia[2], Tao Chen[1*]

[1]School of Electronic and Information Engineering, Wuyi University, Jiangmen Guangdong
[2]School of Computer Science and Engineering, California State University Northridge, Northridge, USA

Received: May 22nd, 2024; accepted: Jun. 21st, 2024; published: Jun. 28th, 2024

Abstract

With the rapid development of industrialization and urbanization, the problem of industrial pol-

*通讯作者。

文章引用: 薛子如, 高乐, 贾旭东, 陈涛. 产业污染管理知识图谱构建与应用: 以江门市为例[J]. 计算机科学与应用, 2024, 14(6): 137-148. DOI: 10.12677/csa.2024.146150

lution has become increasingly serious, posing a significant threat to the environment and human well-being. Industrial pollution data are multi-sourced and heterogeneous, requiring the use of knowledge extraction and fusion techniques for application in industrial pollution management and decision support. This study proposes a knowledge graph approach to integrate and model key data, entities, and relationships within the field of industrial pollution industrial pollution knowledge in terms of concepts, attributes, and instances. Using Jiangmen City in Guangdong Province, China, as a case study, we constructed a comprehensive industrial pollution knowledge graph by integrating company, pollutant, product, and other relevant information through knowledge extraction, ontology construction, and knowledge storage technology. The experimental results show that our knowledge graph construction method effectively reveals potential associations among polluted site data, providing not only intuitive insights but also valuable data support and decision-making references for decision makers. Additionally, the graph contributes to the broader research community by offering accessible data for related studies and applications.

Keywords

Knowledge Graph Construction, Industrial Pollution, Ontology Construction, Environmental Governance

1. 引言

随着我国经济飞速发展，在工业化、城市化进程中，产业活动通常伴随着能源消耗、物质排放和废弃物处理等环境影响，如废水排放、固废沉降、管道泄漏、危化品运输储存等[1]。随着工业化进程的加快，工业生产规模持续扩大，导致产业污染的排放量不断上升。目前，部分地区和行业的产业污染问题依然严重，具体问题包括污染物排放超标和环境监管不力等。此外，产业污染还带来了一系列消极影响，如破坏环境系统、增加健康风险、阻碍经济发展等[2]。产业污染相关数据具有多类型、多关联、多维度和大数据量等特点，需要对其进行深度挖掘，才能更好对产业污染进行高效管理。在环境领域中，传统的数据服务知识库在语义关联方面存在不足，尤其在处理碎片化的数据，如污染场地、土壤污染和生产活动等方面。

知识图谱[3]是一种图形化表示方法，将领域知识进行知识提取，整合相关联的实体和关系，以节点和边的形式展示，用于呈现知识领域中实体之间的关系和发展过程。它通过建立实体之间的语义关系，能够有效地连接和管理碎片化数据，为环境管理提供更全面、可视的解决方案。目前，一些知名的通用知识图谱包括 DBpedia [4]、Wikidata [5]、Yago [6]等。随着知识图谱技术的发展，已经在金融、医疗、商业、企业等领域得到广泛应用[3]。诸云强等人[7]在地学知识图谱构建综述中，提出了一个适用于大规模地学知识图谱构建技术和总体框架。赵又霖等人[8]构建了面向突发事件的时空语义模型，使用本体模型构建突发事件地理本体，为应急管理领域建立完整的共享概念模型。Wu 等人[9]讨论了自动知识图谱构建的概念，该概念涉及了在没有人为干预的情况下从非结构化文本构建知识图谱。此外，知识图谱可以追踪地表水中的水文污染物运输[10]，该图谱可以更好地了解污染物在水体、河流和流域之间的运输和命运，同时利用先进的算法和工具对数据进行了分析和决策。王晓爽等人[11]通过抽取案件示例的实体和关系，构建了大气污染执法事理知识图谱，实现大气污染案件的溯源。总体而言，在污染物相关研究中，知识图的构建涉及一系列任务，有助于全面理解影响污染水平和风险的因素。

因此，本课题提出构建一个产业污染场地的知识图谱，将知识图谱技术应用在污染场地数据管理和挖掘中，使用 BiLSTM-CRF 模型进行实体抽取，利用本体构建方法对产业所在城市、地区、类型、产物、污染物等信息进行建模，并在 Neo4j 图数据库存储和对产业污染知识图谱进行可视化。本课题构建的产业污染知识图谱方法在中国广东省江门市进行了实验，通过构建产业污染知识图谱，为决策者提供信息挖掘技术，为产业污染治理提供技术支持。相关数据和代码可以通过 tiancaiziru/BiLSTM-CRF (https://github.com/tiancaiziru/BiLSTM-CRF)网址下载。

2. 相关方法

知识图谱具有丰富的语义性，能够整合多源异构数据，将复杂的数据转化成由“实体－关系－实体”组成的三元组，从而表示并存储知识中的各种实体及其关系。这些实体和关系共同构成了一个复杂的网络，使得知识的存储相互关联并得到相互支持。考虑到污染场地数据既包括结构化数据也包括非结构化数据，如文本、图像、表格等，本研究采用了自顶向下的方法来构建污染场地的知识图谱。知识图谱构建一般方法如图 1 所示。

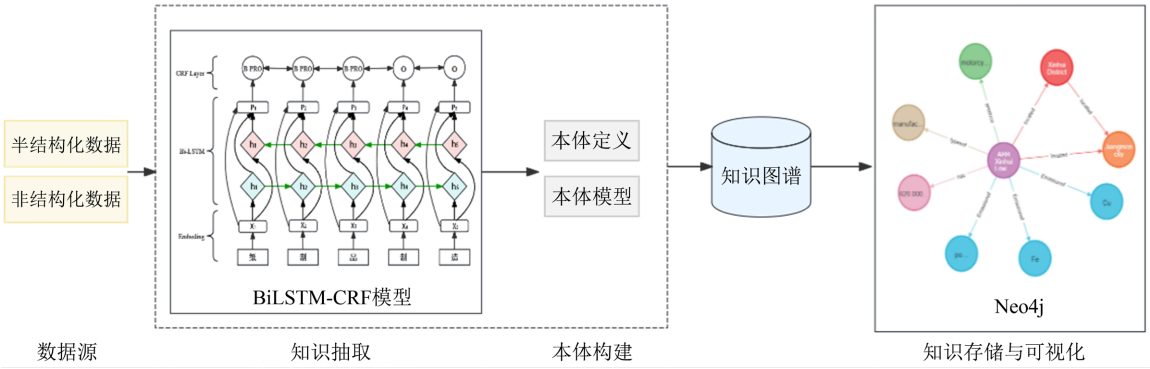

Figure 1. Framework diagram of general methods for constructing knowledge graphs
图 1. 知识图谱构建一般方法的框架图

如图 1 所示，知识图谱的构建流程主要包括以下步骤：数据收集、知识抽取、知识融合、知识存储以及知识图谱应用。首先，根据数据的不同结构，选择适宜的知识抽取方法来提取产业污染信息，以建立更加完善的语料库。其次，对抽取出的数据进行融合，并构建本体，定义其层次结构、关系、属性和规则，确立图谱构建和知识推理的统一模式。接着，使用 Neo4j 等图形数据库进行知识存储，并通过可视化与知识推理技术挖掘有效信息，进行深入推理和分析。最后，将构建完成的知识图谱整合至相关应用中。

2.1. 知识抽取

知识抽取是从不同来源和结构的数据中提取知识，并将其转化为将其结构化数据以存储于知识图谱的过程。对于结构化和半结构化数据，通常只需进行简单的预处理和映射，便能作为后续数据分析系统的输入，相关技术已相当成熟。然而，对于非结构化知识，需要应用自然语言处理技术来提取实体和关系，这需要借助信息抽取和深度学习的技术来帮助提取有效信息。目前，知识抽取技术的主要难点和研究方向包含实体抽取、关系抽取和事件抽取三个子任务。实体抽取，也称为命名实体识别(NER)，能够识别文本中具有特定意义的实体及其边界，这些命名实体包括人名、地名、时间等。NER 在自然语言处理领域有着广泛的应用，如问答系统、知识图谱构建等。在 NER 任务中，基于词典和规则的抽取方式需要

人工介入，随着信息增长，此类方法结构复杂，抽取效率较低。基于统计模型的抽取方法需要对上下文等数据进行标注，并利用统计方法训练模型，常用的模型包括支持向量机(SVM) [12]、条件随机场(CRF) [13]、隐马尔可夫模型(HMM) [14]等，但标注过程仍需人工参与。相比之下，基于深度学习的方法展现出强大的语义挖掘能力。常用模型包括循环神经网络(RNN) [15]、双向长短期记忆网络(BiLSTM) [16]、BiLSTM 与 CRF 结合模型 BiLSTM-CRF [17]等。当前，深度学习模型在命名实体识别任务中已取得显著效果。

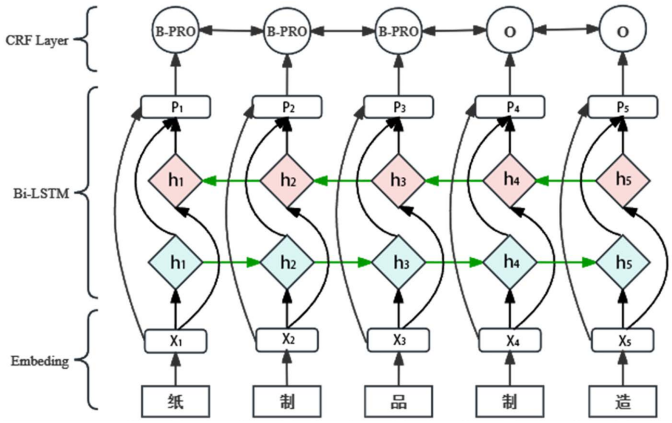

Figure 2. BiLSTM-CRF model framework diagram [17]
图 2. BiLSTM-CRF 模型框架图[17]

BiLSTM-CRF 模型是基于深度学习的序列标注算法，将预处理好的词向量输入到 BiLSTM 模块，捕获文本上下文语义信息，最后使用 CRF 模块解码 BiLSTM 模块输出的预测标签序列，对标注任务进行全局优化，获得最佳标注结果。BiLSTM-CRF 模型能够利用大量数据进行训练，并且其自动提取文本特征的准确率高。如图 2 所示，BiLSTM-CRF 模型由一个正向 LSTM 神经网络和一个反向 LSTM 神经网络组成，形成双向 LSTM 网络。该模型将每个句子的每个单词表达成一个向量，其中 X_i 表示句子的第 i 个字在字典索引，生成每个字对应的 one-hot 向量。BiLSTM 接收每个字符的 embedding，第一层是 look-up 层，利用 embedding 矩阵将句子 X_i 中的每个字由 one-hot 向量映射为低维的字向量。第二层是双向 LSTM 层，自动提取句子特征，记作矩阵。第三层是 CRF 层，将 BiLSTM 的 Emission_score 作为输入，输出符合标注转移约束条件的、最大可能的预测标注序列。给定模型的输出 logits 和标签序列y，以及序列长度向量，CRF 损失函数计算真实标签序列的对数概率，并取其负值作为损失。

模型的损失函数公式：

$$P\left(\bar{y}\mid x\right)=\frac{\exp\left(\text{score}\left(x,\bar{y}\right)\right)}{\sum_{y}\exp\left(\text{score}\left(x,y\right)\right)} \tag{1}$$

$$\text{score}\left(x,y\right)=\sum_{i=1}^{n+1}A_{y_{i-1},y_i}+\sum_{i=1}^{n}P_{i,y_i} \tag{2}$$

其中，P 为 Bi-LSTM 的输出矩阵，$P_{i,yi}$ 表示序列 y 中第 y_i 个标签的发射得分，A 为 tag 之间的转移矩阵，$A_{y_{i-1},\,y_{i-1}}$ 表示从第 y_{i-1} 个标签到第 y_i 个标签的转移得分。y 是标记序列，x 是单词序列，Score (x, y)即单词序列 x 产生标记序列 y 的得分，得分越高，说明其产生的概率越大。

计算所有可能的标签序列的分数和：

$$\log P\left(\bar{y}\mid x\right)=\text{score}\left(x,\bar{y}\right)-\log\left(\sum_{y}\exp\left(\text{score}\left(x,y\right)\right)\right) \tag{3}$$

2.2. 本体构建

本体是知识图谱中用来描述特定领域知识和概念的一组术语和定义。它将事件转换成知识表示的概念，并对语料进行结构化表示，本体构建技术已经成为自然语言处理的一个热门技术，它能够体现信息之间的复杂关系。在知识图谱中，本体位于模式层，它定义了特定领域知识和概念的术语和定义，作为知识库的概念模板。模式层是知识图谱的概念模型，也是实例层的逻辑基础，依赖模式层中的关系和规则的定义，能实现知识的推理。在本体层构建阶段，目标是通过系统化、规范化、形式化的方式定义和表达学科领域的概念、属性、关系和规则[7]。目前，本体构建主要采用手工构建、半自动构建以及自动构建本体三种方法[18]。这些方法利用知识体系和结构化知识转换等技术，依据已有的文献、本体等数据，来构建本体层。目前，存在许多本体构建方法，方法包括 TOVE 法、METHONTOLOGY 法、骨架法和七步法等[19]。斯坦福大学医学院提出的七步法广泛应用于多个领域，有效提高本体的构建效率。对于本体的管理和维护问题，本体构建工具能够提高本体构架的效率，本体构建工具包括 OntoEdit、WebODE 和 Protégé 等[20]。

2.3. 知识存储与可视化展示

知识存储旨在为知识图谱的知识表示形式设计底层存储方式，完成各类知识的存储，以支持对大规模图数据的有效管理和计算。知识存储的对象包括基本属性知识、关联知识、事件知识、时序知识和资源类知识等。通过对数据进行处理，获取有效的信息，将不同的数据转化为结构化的三元组数据，依据数据量的大小、数据特征以及应用需求，选取合适的存储模式，将获取到的数据存储起来，形成知识图谱。知识存储是将已有的知识图谱进行存储，目前知识图谱的存储方式包括基于关系型数据库存储、基于 RDF 数据库存储和基于图数据库存储。RDF 数据库使用 XML 语法来表示数据模型，通过三元组形式描述资源的特性及其相互关系，并将这些三元组数据以文本形式存储。图数据库使用节点和边存储数据，目前主流的图数据库包括 DgraphDB、Neo4j、JanusGraph 等[21]。图数据库能够更直观的表达实体的关系，与传统数据库相比，图数据库能更直观地表达实体间的关系，并且具有更高效的复杂关系问题处理能力。

知识图谱可视化通过图形化手段展示知识图谱中的实体、关系和属性等信息，帮助用户直观理解知识图谱的结构和内容，并支持用户探索其中的实体、属性和关系。用户可以直接在知识图谱可视化工具中进行搜索、推理等操作，进而能够挖掘数据中的规律和隐藏的信息。现阶段主要的知识图谱可视化工具包括 Neo4j Bloom、D3.js、Cytoscape.js 等。

3. 产业污染知识图谱构建与应用

3.1. 知识图谱构建

本文使用课题组所收集的江门市产业污染数据作为主要数据源，使用 BiLSTM-CRF 模型对数据进行实体抽取，将处理好的数据转换成三元组，通过自顶向下的方法细分概念构建本体，并基于 Neo4j 构建并存储产业污染知识图谱。

3.1.1. 知识抽取

本课题数据来源于课题组收集的数据，通过网络爬虫爬取百度百科与江门市产业污染相关的企业信息，原始数据为大量文本数据和表格。数据集样式如表 1 所示。数据获取之后不能直接进行知识图谱构建，需要对获取内容进行数据预处理。对于结构化数据，数据预处理后将实体和关系存储进图数据库中。

对于非结构化数据，产业污染领域的知识抽取过程中，需要专家手动构建规则模板进行命名实体识别，

建设周期长。为了提高抽取效率和准确率，本课题使用了 BiLSTM-CRF 深度学习模型进行产业污染实体抽取。本课题的研究对象为江门市的产业污染，数据中收集到一些不属于产业污染范畴的案例，需要将其去除重复性操作，并参考实体的命名规范和结合专家意见对信息进行标注，将数据集的格式转换为 BIO 标注格式，数据集的每个字符均被标记成“B-product”、“I-product”或“O”，最终形成命名实体识别的标准数据集。使用 BiLSTM-CRF 模型进行命名实体识别，将 80%的标注语料作为训练集，10%的预料作为验证集，10%的语料作为测试集。为了评价命名实体识别算法效果，通过使用精确率(Precision)，召回率(Recall)和 F1 值作为评价识别的性能指标。精确度以预测结果为判断依据，预测为正例的样本中预测正确的比例。召回率以实际样本为判断依据，实际为正例的样本中，被预测正确的正例占总实际正例样本的比例。F1 分数被定义为精确率和召回率的调和平均数。

Table 1. Overview of part of the industrial pollution dataset in Jiangmen city
表 1. 江门市产业污染数据集部分数据概览

主体名称	住所/经营场所/驻在场所	行业门类	经营范围
台山市健冠五金塑料模具有限公司	台山市四九镇台商投资示范洞美工业区 32 号	制造业(C)	生产、销售：五金模具、塑料模具、塑胶制品、五金制品、铝制品、自动机械化设备。
台山市元超电子有限公司	台山市冲蒌镇红岭工业区红岭中路 7 号之六	制造业(C)	生产、销售：电子元器件、线路板(不含电镀)、LED 灯具及配件。
江门市冠捷电子有限公司	江门市江海区高新西路 46 号 B1 厂房三楼	制造业(C)	五金配件加工；销售：线路板材料及其配件、电子元器件。
江门市新会区金桥化工厂	江门市新会区罗坑镇天湖村委会锦龙村民小组	制造业(C)	生产、销售：水性涂料、水性漆、水性乳液、水性胶粘剂、水溶性树脂、水性聚合物、胶浆、胶粘剂。
台山市尊尚装饰工程有限公司	台山市水步镇台鹤中路 100 号新雅苑 8 幢 104 商铺	建筑业(E)	承接装饰工程；装饰设计；门窗加工、安装；销售：建筑材料。

衡量标准 F1 值公式：

$$F1 = \frac{2 \times precision \times recall}{precision+recall} \tag{4}$$

精确率公式：

$$precision = \frac{TP}{TP+FP} \tag{5}$$

召回率公式：

$$recall = \frac{TP}{TP+FN} \tag{6}$$

Table 2. Experimental results of extracting information from industrial pollution entities
表 2. 产业污染实体信息抽取实验结果

模型	precision	recall	F1	accuracy
BiLSTM-CRF	83.55%	83.01%	83.28%	96.85%

实体抽取评价结果如表 2 所示。总体在产业污染实体抽取中达到较好的效果。

针对本文提出的污染场地知识图构建方法，以广东省江门市的污染场地数据为例进行了实例分析。抽取实体如表 3 所示。共抽取场地信息、企业信息、污染信息等污染场地实体 21,221 个，对应关联关系

235,137 个。根据各地区企业的情况、生产活动、污染状况与这组基本信息相关联，便于分析某一污染物污染原因，实现后续污染场地的防治。

Table 3. Overview of entities and relationships in the industrial pollution knowledge graph statistical table
表 3. 产业污染知识图谱的实体和关系概述统计表

实体类型	实体数量	关系类型	关系名称	关系数量
公司	21,221	公司→城市	位于	21,221
城市	1	公司→区/市	位于	21,221
区/市	7	公司→行业	行业类型是	21,221
行业	519	公司→产品	生产	56,853
产品	160	公司→污染物	排放	114,408
污染物	58	行业→污染物	排放	206

3.1.2. 本体构建

本课题采用斯坦福大学提出的领域本体构建七步法，以构建产业污染知识图谱本体。首先，确定产业污染领域的本体范围，涵盖实体、属性和关系。通过文献综述、专家咨询以及政策法规分析等途径，收集与产业污染相关的知识。随后，对收集到的知识进行实体抽取，并对抽取出的实体、属性和关系进行分类与整理。接着，基于提取的信息构建本体模型，包括产业污染领域实体的结构、属性定义和关系描述。然后，对构建的本体模型进行验证和评估，并与专家进行讨论，以确保模型的准确性。最后，将本体模型应用于知识图谱，并根据监控和反馈情况进行必要的改进。

根据本体构建原则，本课题对每个本体类的概念层次进行了细分，并定义了城市产业污染环境实体的标签、关系和类型。使用基于 Java 语言开发的开源本体编辑工具 protégé[1]，对产业污染本体进行可视化处理。产业污染领域的本体构建详情参见图 3。

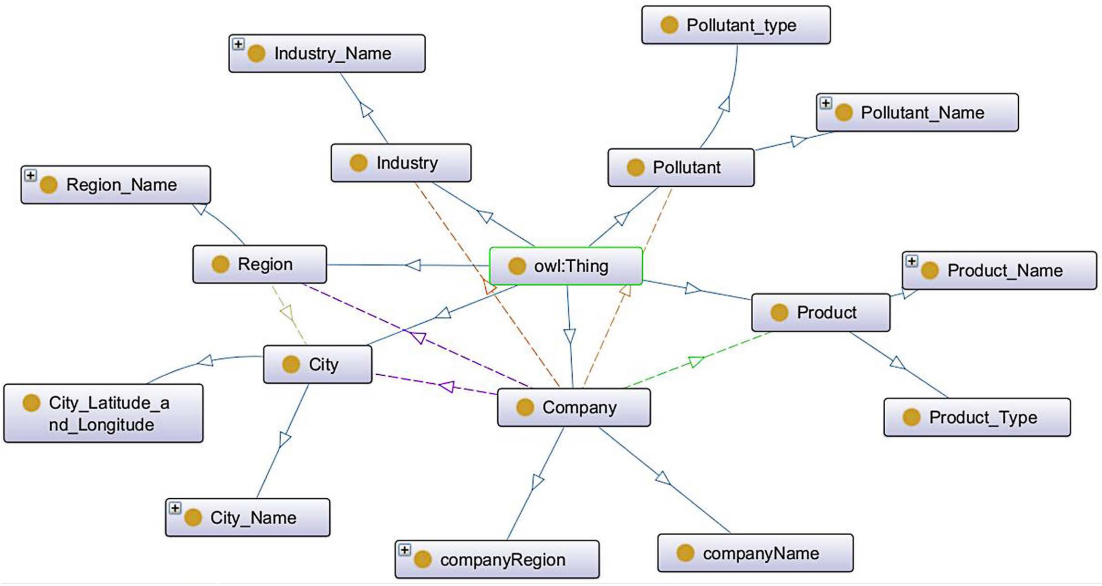

Figure 3. Schematic diagram of the ontology layer division of industrial pollution knowledge graph
图 3. 产业污染知识图谱本体层划分示意图

[1]protégé 软件下载地址：https://protege.stanford.edu/。

从图 3 可以看出，本课题构建的本体包括城市、地区、公司、污染物、产品和行业等部分。例如，在污染物类型，根据对环境影响不同，将其分为土壤污染、水污染和大气污染三类；根据排污行业特点，分为制造业、采矿业、建筑业和供应业等类别。

3.1.3. 知识图谱存储与可视化

知识图谱中，每个实体都是一个节点，每个关系都是一条边，采用节点和边的属性存储知识。构建的知识图谱是将实体－关系－实体、实体－属性－属性值中的首尾部分作为节点，属性和关系以边的形式储存，实现结构化知识三元组到图中节点和边的映射。基于图数据库的存储优势，在图数据库中实体、关系和属性等被映射为节点和边，形成结构化的知识三元组，与知识图谱中的图结构相契合。并且图数据库提供查询语言，能够进行复杂查询，支持多种路径查询算法，查询速度优于关系型数据库。基于所构建的产业污染知识图谱，通过图查询语言和图挖掘算法，进行关系延伸计算，能够实现知识图谱的知识推理和知识补全等应用，提高查询效率。

在存储产业污染知识图谱方面，Neo4j 图数据库展现出更大的灵活性。Neo4j 能够有效地表示实体间的关系、属性以及复杂的拓扑结构，支持快速查询和遍历大规模图形数据集，并提供可视化工具以便于理解数据间的关系，同时允许批量导入节点进行存储。综合以上考虑，本课题使用 Neo4j 图数据库存储知识图谱，实现从概念、属性、实体展示产业污染知识图谱，并使用图数据库 Neo4j 的 Cypher 查询语言遍历查询关系图。

通过分析污染场地在各行业的分布特征，可以更轻松地理解和分析数据，进而帮助决策者制定适当的治理措施，并确定处理的优先级。并为不同利益相关方之间的交流和协商提供有效的依据，从而提高治理效果，更有效地减少污染带来的风险。目前，文本数据的可视化主要是图表展示，但许多方法仅停留在数值比较和基础数据列举上，未能直观地揭示数据间的关联性。考虑到污染场地通常具有多种生产活动和多功能分区的特点，如果不能直观准确地展示数据之间的关系，就无法为后续的污染修复工作提供实质性的方案支持。因此，决定采用关系图可视化的方法来反映文本数据之间的关联性。将实体与关系导入图数据库中，能够根据需求进行数据可视化，展示江门市产业污染的知识图谱。图 4 展示了江门市产业污染中江门地区与公司、公司排放的污染物、生产的产品、所在行业实体之间的关系。不同颜色的节点表示不同的实体，实现产业污染知识图谱三元组的描述，将产业污染的地区、公司、污染物、行业等关系进行表达。

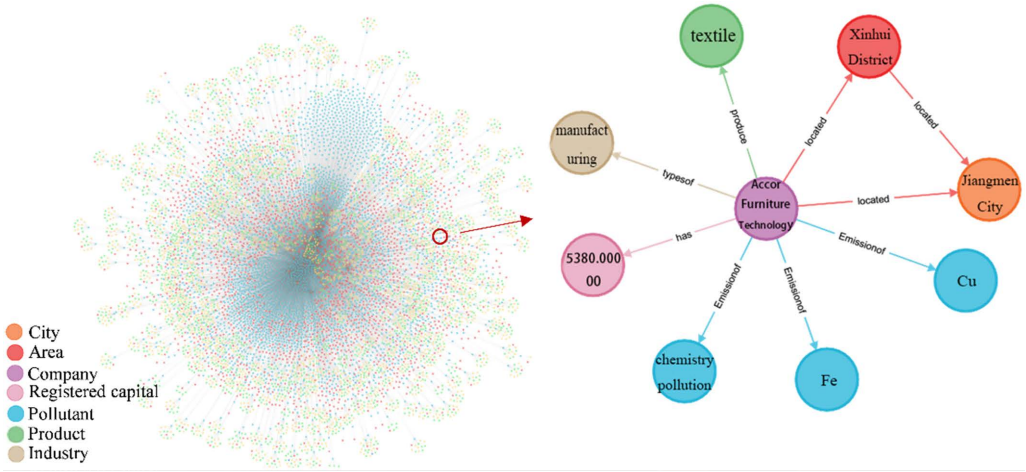

Figure 4. Overview of industrial pollution knowledge graph
图 4. 产业污染知识图谱概览

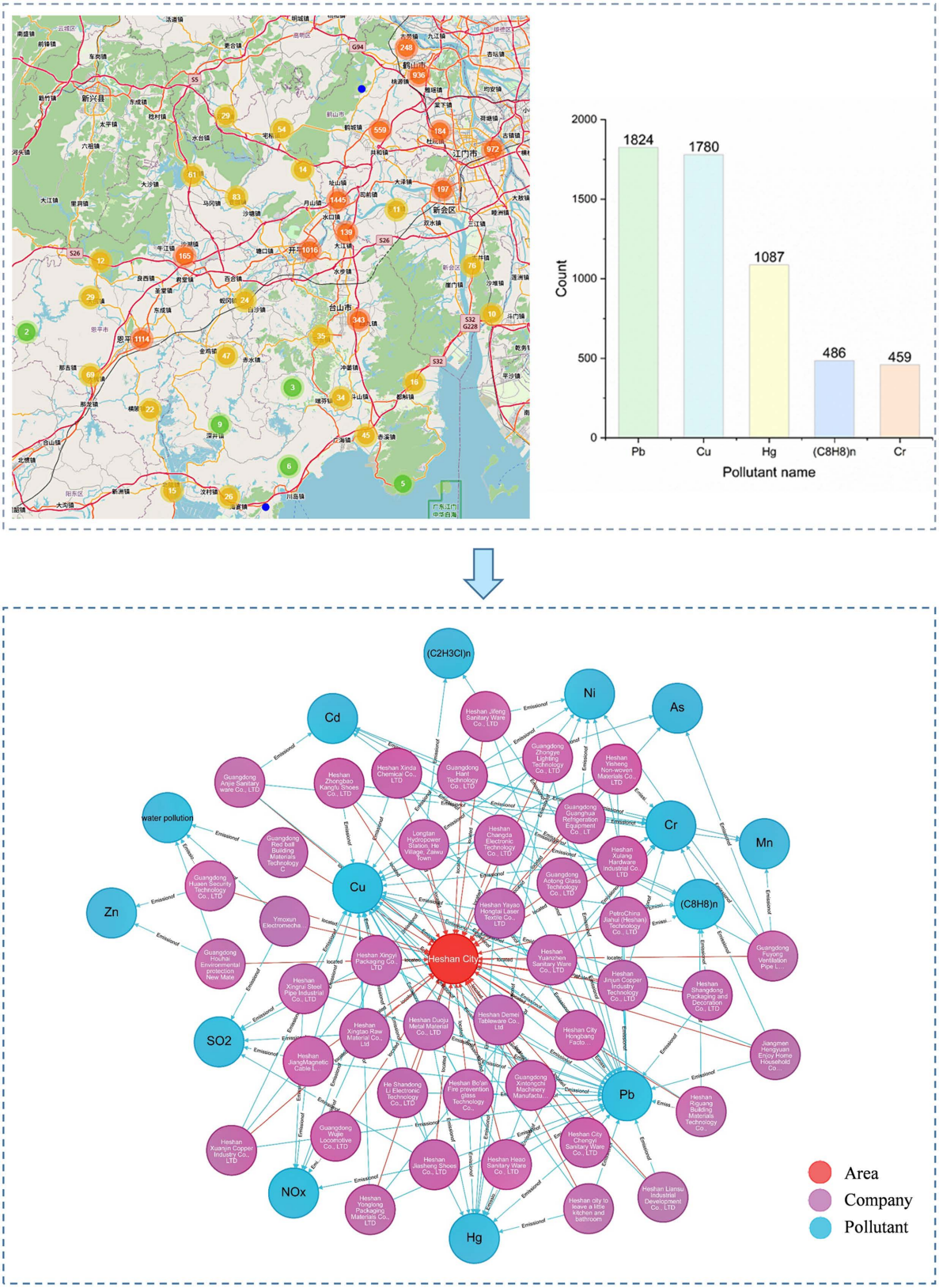

Figure 5. Regional emission of pollutants based on path mining: a case study of Heshan in Jiangmen city
图 5. 基于路径挖掘的地区排放污染物——以江门市鹤山为例

3.2. 知识图谱应用

3.2.1. 路径挖掘

路径挖掘作为知识图谱中的一种方法，能够揭示实体间的潜在关系和隐藏信息。利用知识图谱的路径挖掘技术可以深入了解数据挖掘过程，进而发现数据中的潜在关系和模式。通过结合领域知识，可以提高数据挖掘的准确性和效率。本课题运用路径挖掘技术，针对产业污染知识图谱，分析了特定地区的行业分布以及特征污染物的区域分布情况。针对江门市，本课题利用知识图谱关联公司所在行业、地区及排放的污染物，以此展示与公司节点相连接的地区。以拥有最多公司数量的地区为例，我们统计了与各污染物的连接数量，并据此对污染物进行排序。最终，我们识别出了该地区主要排放的重金属污染物。

使用 Neo4j 软件中的 neomap 插件，结合公司的经纬度信息，采用图聚类算法在知识图谱中识别出密集连接的子图。图 5 展示了江门市公司的聚类情况，以江门市鹤山为例，该地区在知识图谱可视化结果中显示公司数量最多。定义了一组规则来描述预期挖掘的路径模式，并根据这些规则在知识图谱中搜索匹配的路径。图 5 展示了名为“鹤山市”的公司实体与各类污染物之间的最短路径，并根据江门市鹤山的排放情况对污染物进行排序，从而确定该地区主要排放的重金属污染物为铅(Pb)、铜(Cu)和汞(Hg)。同时，根据相同方法，挖掘出鹤山主要行业为制造业。在有限的环境管控条件下，这些企业在处理废水、重金属污染物等方面也面临着挑战。这些挑战增加了环境污染事故发生的可能性，进而增加了污染的风险。通过知识图谱的可视化分析，我们可以识别和评估污染因素，预测潜在的环境风险，并提前采取相应措施，以预防或降低环境污染事故的风险。在识别出主要排放污染物的地区和行业后，可以更有针对性地分配环保资源，比如监测设备、清洁技术投资和政策执行力度。

3.2.2. 信息检索

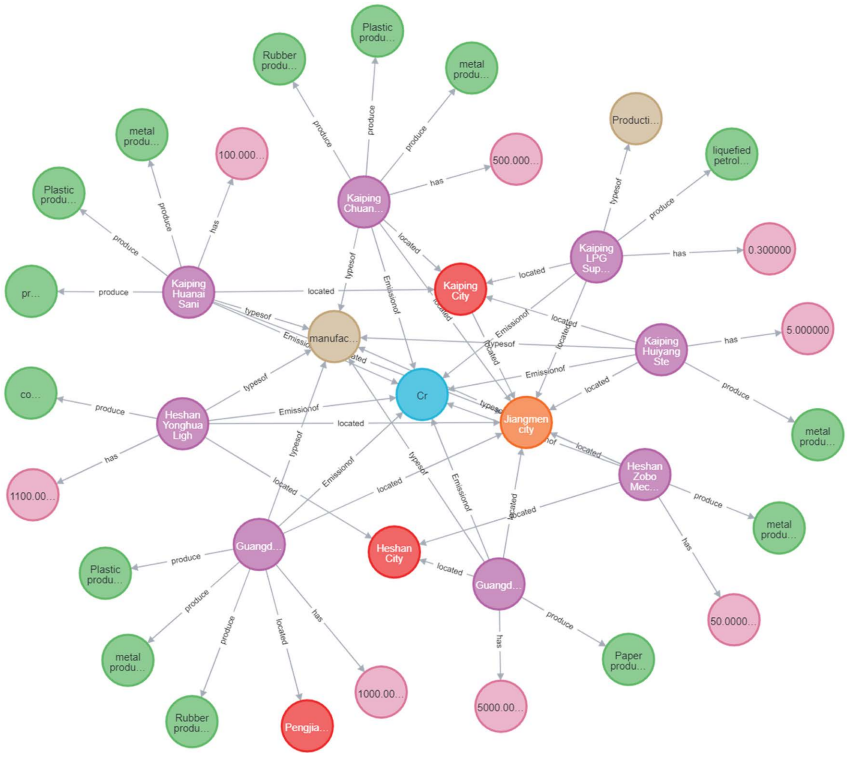

Figure 6. The relevant information of Cr pollutants in the knowledge graph constructed in this project
图 6. 本课题构建的知识图谱中 Cr 污染物的相关信息

在产业污染领域，污染物的致污因素至关重要，它们是污染场地修复和管理的关键依据。这有助于工作人员分析污染物的来源和传输途径，进而评估并制定相应的治理措施。在产业污染知识图谱中，公司相关的产品、污染物、行业、城市、地区和注册资本等信息均通过三元组来表示。在 Neo4j 数据库中，使用 Cypher 语言来检索相关信息。通过设定复杂的规则和语法，我们能够搜索特定的 Cr(铬)污染物。查询结果显示了排放 Cr 的公司、所属行业类型、公司所在地区、生产的产品及注册资本等信息(见图 6)。这不仅揭示了数据间的关联，还实现了对 Cr 污染物信息的知识整合与分析，同时能够为管理者提供了智能信息检索功能。通过构建的知识图谱，可以识别出公司生产的产品对环境的潜在影响，并判断其是否为主要的污染排放源。这些公司所在的地区面临着环境挑战和风险，有助于地方政府分析污染物的来源和传输途径，对潜在的环境风险进行评估，从而更好地规划环境修复和管理措施，可以制定更为针对性的治理措施，如对特定行业或区域实施更严格的排放标准。

4. 结论与展望

产业污染知识图谱能够将大量多源异构数据转换为图数据，利用节点和边的形式表示信息，并挖掘其中的隐藏信息，从而实现高效的数据管理。知识图谱在数据存储管理、数据关联和信息挖掘方面发挥着重要作用，基于知识图谱的特点，本课题选择中国广东省江门市中 21,222 家公司作为研究对象，构建产业污染知识图谱，并探究其在城市产业污染管理中的应用潜力。针对不同类型的数据，设计了产业污染知识地图的构建流程，提出了产业污染数据的实体识别技术。利用本体工具对产业污染本体的概念属性进行可视化，实现了产业污染知识概念和关系的系统化、规范化、形式化表达。根据图数据库有助于揭示特定场地内部和周边的完整信息的特点，减少潜在风险的隐匿性。通过对知识图谱进行路径挖掘和信息查询，获取地区、行业受污染情况，提供受污染相关信息，辅助污染成因。分析研究结果表明，知识图谱技术在污染场地管理和污染预测方面具有巨大潜力。本课题为政府机构提供了宝贵的启示，增强了对污染问题的理解与应对能力，有助于更精确地制定环境紧急事件的管理策略。此外，它还能提供快速、可靠的决策支持，推动环境保护和社会经济的可持续发展。尽管本课题在知识抽取过程中取得了进展，但仍存在诸多挑战。未来的工作需要改进技术，收集更多的产业污染数据，并进一步将产业污染知识图谱应用于应急管理、分析与评估中，以提升数据的应用价值。

参考文献

[1] 周沛婕, 李娟, 童晓静. 浅析工业源污染减排的现状及对策[J]. 皮革制作与环保科技, 2022, 3(11): 186-188.

[2] 刘沛, 黄慧敏, 余涛, 等. 我国新污染物污染现状、问题及治理对策[J]. 环境监控与预警, 2022, 14(5): 27-30+70.

[3] Peng, C., Xia, F., Naseriparsa, M. and Osborne, F. (2023) Knowledge Graphs: Opportunities and Challenges. *Artificial Intelligence Review*, **56**, 13071-13102. https://doi.org/10.1007/s10462-023-10465-9

[4] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., *et al*. (2009) Dbpedia—A Crystallization Point for the Web of Data. *Journal of Web Semantics*, **7**, 154-165. https://doi.org/10.1016/j.websem.2009.07.002

[5] Vrandečić, D. and Krötzsch, M. (2014) Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*, **57**, 78-85. https://doi.org/10.1145/2629489

[6] Suchanek, F.M., Kasneci, G. and Weikum, G. (2008) YAGO: A Large Ontology from Wikipedia and WordNet. *Journal of Web Semantics*, **6**, 203-217. https://doi.org/10.1016/j.websem.2008.06.001

[7] 诸云强, 孙凯, 胡修棉, 等. 大规模地球科学知识图谱构建与共享应用框架研究与实践[J]. 地球信息科学学报, 2023, 25(6): 1215-1227.

[8] 赵又霖, 庞烁, 吴宗大. 社会感知数据驱动下突发事件应急管理的时空语义模型构建研究[J]. 情报科学, 2021, 39(2): 44-53.

[9] Wu, X., Wu, J., Fu, X., Li, J., Zhou, P. and Jiang, X. (2019). Automatic Knowledge Graph Construction: A Report on the 2019 ICDM/ICBK Contest. 2019 *IEEE International Conference on Data Mining* (*ICDM*), Beijing, 8-11 November 2019, 1540-1545. https://doi.org/10.1109/icdm.2019.00204

[10] Cole, D.L., Ruiz-Mercado, G.J. and Zavala, V.M. (2023) A Graph-Based Modeling Framework for Tracing Hydrological Pollutant Transport in Surface Waters. *Computers & Chemical Engineering*, **179**, Article ID: 108457. https://doi.org/10.1016/j.compchemeng.2023.108457

[11] 王晓爽, 李吉东, 徐海红, 等. 顾及时空特征的大气污染执法事理图谱构建方法研究[J]. 地理与地理信息科学, 2022, 38(3): 1-8.

[12] Lee, K., Hwang, Y., Kim, S. and Rim, H. (2004) Biomedical Named Entity Recognition Using Two-Phase Model Based on SVMs. *Journal of Biomedical Informatics*, **37**, 436-447. https://doi.org/10.1016/j.jbi.2004.08.012

[13] Yu, B. and Fan, Z. (2019) A Comprehensive Review of Conditional Random Fields: Variants, Hybrids and Applications. *Artificial Intelligence Review*, **53**, 4289-4333. https://doi.org/10.1007/s10462-019-09793-6

[14] Rabiner, L.R. (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, **77**, 257-286. https://doi.org/10.1109/5.18626

[15] Ma, Z., Zhang, H. and Liu, J. (2023) MM-RNN: A Multimodal RNN for Precipitation Nowcasting. *IEEE Transactions on Geoscience and Remote Sensing*, **61**, 1-14. https://doi.org/10.1109/tgrs.2023.3264545

[16] Woźniak, M., Wieczorek, M. and Siłka, J. (2023) BiLSTM Deep Neural Network Model for Imbalanced Medical Data of IoT Systems. *Future Generation Computer Systems*, **141**, 489-499. https://doi.org/10.1016/j.future.2022.12.004

[17] Chen, T., Xu, R., He, Y. and Wang, X. (2017) Improving Sentiment Analysis via Sentence Type Classification Using BiLSTM-CRF and CNN. *Expert Systems with Applications*, **72**, 221-230. https://doi.org/10.1016/j.eswa.2016.10.065

[18] 何琳, 杜慧平, 侯汉清. 领域本体的半自动构建方法研究[J]. 图书馆理论与实践, 2007(5): 26-27+38.

[19] 王向前, 张宝隆, 李慧宗. 本体研究综述[J]. 情报杂志, 2016, 35(6): 163-170.

[20] 杨郑子衿, 徐倩, 王安莉, 等. Protégé 在构建中医药本体中的应用[J]. 医学信息学杂志, 2021, 42(6): 37-42+47.

[21] 王鑫, 邹磊, 王朝坤, 等. 知识图谱数据管理研究综述[J]. 软件学报, 2019, 30(7): 2139-2174.