

化工行业多源异构碳数据集构建方法解析

刘小楠^{1*}, 周 强¹, 黄 勇¹, 陆可飞², 莫 凡^{2#}

¹四川轻化工大学化学工程学院, 四川 自贡

²广东埃文低碳科技股份有限公司, 广东 广州

收稿日期: 2025年7月12日; 录用日期: 2025年8月12日; 发布日期: 2025年8月21日

摘 要

在碳达峰碳中和战略驱动下, 化工企业需要通过精准的碳排放数据分析制定减排路径, 但当前企业面临数据孤岛、质量缺陷及时效性不足等问题。现有研究虽在多源数据融合算法、工业大数据平台等方面取得进展, 但针对化工行业全流程碳数据的融合能力不足, 且缺乏对供应链环节的覆盖。本文创造性地构建了一套适配化工行业的多源异构碳数据集的分层架构, 通过整合生产、能源消耗及供应链全流程数据, 同时考虑数据可得性、准确性、时效性和可操作性。

关键词

双碳战略, 化工行业, 碳排放, 多源异构数据

Analysis of the Construction Method of Multi-Source Heterogeneous Carbon Datasets in the Chemical Industry

Xiaonan Liu^{1*}, Qiang Zhou¹, Yong Huang¹, Kefei Lu², Fan Mo^{2#}

¹College of Chemical Engineering, Sichuan University of Science & Engineering, Zigong Sichuan

²Guangdong Avi Low Carbon Technology Co., Ltd., Guangzhou Guangdong

Received: Jul. 12th, 2025; accepted: Aug. 12th, 2025; published: Aug. 21st, 2025

Abstract

Driven by the dual carbon Strategy, chemical enterprises need to formulate emission-reduction pathways through precise carbon-emission data analysis. However, current enterprises are facing

*第一作者。

#通讯作者。

文章引用: 刘小楠, 周强, 黄勇, 陆可飞, 莫凡. 化工行业多源异构碳数据集构建方法解析[J]. 计算机科学与应用, 2025, 15(8): 161-167. DOI: 10.12677/csa.2025.158206

with problems such as data silos, quality defects and inadequate timeliness. Although existing studies have made progress in multi-source data fusion algorithms, industrial big data platforms and other aspects, they are insufficient in the fusion capability of full-process carbon data in the chemical industry and lack coverage of the supply chain links. This paper creatively constructs a hierarchical architecture of multi-source heterogeneous carbon data sets adapted to the chemical industry. It integrates full-process data of production, energy consumption and supply chain, while considering data availability, accuracy, timeliness, and operability.

Keywords

Dual Carbon Strategy, Chemical Industry, Carbon Emission, Multi-Source Heterogeneous Data

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

随着全球气候变化问题日益严峻,中国提出并积极践行碳中和战略,计划于 2060 年前实现碳排放达峰并逐步实现碳中和。作为国民经济的重要支柱,化工行业不仅为各领域提供基础原材料和产品,也是温室气体排放的重点领域之一。化工行业作为碳排放重点领域,其碳排放约占全国总排放量的 5.3%,且其碳排放强度较高,生产过程中涉及的化学反应、能源消耗和原材料使用均产生大量 CO₂ 和其他温室气体。因此,化工行业迫切需要对多源异构的碳数据进行采集、分析和管理的,制定科学有效的减排路径,以响应国家碳中和目标。

目前,学界在碳排放数据多源异构融合工作进行了有益的探索。在数据整合算法层面,Dong 等提出了一种基于多维属性分析的多源数据融合算法,并将该算法应用于电力企业能源消耗和温室气体排放分析[1],但并未覆盖化工行业复杂的化学反应与物料转化过程。Bruckner 等提出了一种基于蚁群算法和模糊理论相结合的企业碳排放信息集成管理模型,虽能评估减排绩效,但未纳入供应链环节的间接排放数据[2]。在工业数据平台方面,熊肖磊等人通过使用 Redis 提升实时数据缓存效率,聚集于设备监控,未涉及碳核算的标准化处理[3]。赵德基等(2017)的 Dubbo + NoSQL 工业大数据平台虽支持多源数据接收与存储,但缺乏针对碳数据的专业化清洗规则[4]。在分析方法层面上,王宏志等(2018)的工业大数据分析框架侧重通用生产参数优化,未结合化工行业特定标准(如《中国石油化工企业温室气体排放核算方法与报告指南》)[5]。Hussain 和 Mir 等(2022)及 Hardiyansah & Zhang 等学者探讨了碳排放信息对企业价值的影响,但未涉及数据集构建的技术细节[6]-[8]。在行业适配性层面,Raihan 和 Wiedmann 从工业生态学的角度提出经济增长、可再生能源使用、技术创新与碳排放之间关系的实现有助于全球生态环境的治理和维护[9] [10],Chen 和 Gui 等对碳排放交易建立对碳排放效率的影响进行了研究[11],但均未针对化工行业高能耗、多反应路径的特性设计数据采集与处理方案。

综上,现在研究存在“三缺”:缺乏对化工全流程(生产 + 能源 + 供应链)数据的整合能力、缺乏针对碳数据的标准化处理规则、缺乏适配化工工艺的分析模型,严重制约了企业的减排决策和碳管理水平[12]。本研究聚焦化工行业碳排放检测,对涉及到碳排放的多源异构大数据处理关键技术,包括数据采集、储存、清洗、转换和分析等关键技术进行总结,为化工行业数据异构互通方法和碳排放核算提供科学指引。

2. 多源异构碳数据集构建方法

多源异构碳数据集的构建关键在于整合不同来源、格式和精度的数据,从而支撑碳排放分析、优化控制、碳足迹追踪及政策制定。本文构建的数据集架构包括数据采集、存储、处理与分析、应用服务及安全管理五层,其核心包括数据融合与互联互通,确保符合《中国石油化工企业温室气体排放核算方法与报告指南(试行)》、ISO14064 等标准,并全面覆盖企业生产、能源消耗及供应链全过程的碳排放,为企业低碳转型提供科学依据。

2.1. 数据采集

数据采集与预处理是构建高质量碳排放数据集的基础,直接决定了后续数据集成、分析、决策支持及碳排放优化的效果。

2.1.1. 数据采集的多源渠道

化工企业碳排放数据采集涵盖多个关键源,主要包括以下几类。

一是工业物联网信息采集设备。由该数据源产生的数据量是最大的,且最容易处理。通常来讲工业物联网信息采集设备会遵循行业标准,以某种固定格式将采集到的数据进行结构化输出,其输出形式包括数字、文字、表格等类型且通常是计算机软件可以直接处理的形式,该来源的数据可以由程序直接加工,根据模型直接生成可供数据服务消费的数据。

二是图片及视频。针对某些行业或项目难以直接部署工业物联网信息采集设备的情况,可以通过图片或视频的方式进行数据的采集。但是该媒介获取的数据无法被计算机直接处理,需要先通过 OCR 类技术对图片或视频中的数据进行提取,再由程序根据模型对提取后的数据进行处理,生成可消费数据。

三是行业政策法规。对于已经有成熟的行业政策法规的行业,由于企业会按照政策法规中指定的方式进行数据的收集和处理,所以通过这种方式获取到的数据是结构化的数据,可以省略数据的预处理环节,直接由程序根据模型对数据进行简单处理,即可被下一阶段消费。

四是政府或企业公开信息。政府或企业披露的公开信息经常会以 word 或 excel 等附件形式进行披露,对于这类数据源,需要有程序预先根据文本格式进行解析处理,然后提取特定数据,将数据进行一定简单的预处理后交由数据处理程序进行处理,依据模型生成最终可消费数据。

2.1.2. 数据采集的关键技术

为了解决上述挑战,化工行业数据采集层整合工业自动化系统(DCS/SCADA)、信息管理系统(ERP/MES/EMS)、IoT 传感器及政府监管数据,涵盖 API 接口、物联网协议(MQTT、OPC-UA)及区块链溯源,实现全面、实时的数据获取。具体来说,主要包括以下几类数据采集技术。

一是传感器和自动化设备。传感器是数据采集的核心工具,在化工生产过程中,各种类型的传感器用于监测温度、压力、流量、液位、pH 值、浓度等多个关键参数。随着技术的进步,传感器的精度和实时响应能力得到了大幅提升。无线传感器网络(WSN)和物联网(IoT)技术的应用,使得远程数据采集和监控成为可能。

二是实时数据采集系统,如分布式控制系统(DCS)和可编程逻辑控制器(PLC)是实时采集和控制数据的核心平台。这些系统通过与现场设备的连接,能够实时监控生产过程中的各项参数,并将采集的数据发送至中央控制系统或云平台。实时数据采集系统对于化工过程的优化、故障预警以及安全保障起到了重要作用。

三是工业互联网和物联网,随着工业 4.0 的推进,物联网(IoT)技术在化工行业的应用越来越广泛。通过互联网将各种传感器、设备和系统连接起来,能够实现数据的远程采集、实时传输和智能分析。物

联网技术能够有效地将不同类型的设备和系统连接起来，打破数据孤岛，实现数据的统一采集和管理。

四是数据采集网关和接口技术，为了应对不同设备和系统之间的异构性，数据采集网关和接口技术成为连接不同系统的关键工具。常见的接口标准包括 Modbus、OPC、MQTT 等，它们能够实现不同设备间的数据传输和转换。数据采集网关不仅支持不同协议的数据采集，还能够进行初步的数据处理、清洗和压缩，减少数据传输的负担。

五是边缘计算与数据预处理，在化工生产中，实时数据量庞大且涉及多个环节，传统的集中式数据处理方式可能会导致延迟和带宽瓶颈。边缘计算技术可以将数据采集、处理和存储分散到靠近数据源的地方，减少数据传输的延时，提升实时性。通过在边缘设备上进行数据预处理，可以过滤噪声、进行初步的分析，只有有意义的数据才传输至中心系统[13]。

2.2. 数据储存、清洗、转换与归一化

多源异构数据集成是整合来自多个数据源的数据，屏蔽数据之间类型和结构上的差异，解决多源异构数据的来源复杂、结构异构问题，从而实现对数据的统一存储、管理和分析，充分发挥数据的价值。数据集成的关键技术包括数据存储管理、数据清洗与转换及数据降维。采集到的原始数据需经严格预处理，包括数据存储管理、数据清洗与转换及数据归一化，从而解决多源异构数据来源复杂、结构异构问题，实现用户无差别访问，充分发挥数据的价值。异构碳数据集的碳数据根据所属领域的不同，数据结构有极大的差异，以下表 1~3 为有色金属、油气、煤化工等典型化工行业一部分典型领域的碳数据字段。

Table 1. Nonferrous metals industry data

表 1. 有色金属行业数据

数据字段类型	数据字段解释
能源类型	燃气/电
热值	燃料单位热值含碳量
运输	交通运输碳排放量
进气流量	鼓风机进气流量数值
电量	用电量数值
温度	工作温度、废气温度等数值
废气排放	废气种类及排放量
基于产能换算的每吨碳量	依照产能换算表算出的碳量

Table 2. Oil, gas and chemical industry data

表 2. 油气化工行业数据

数据字段类型	数据字段解释
压力	管道或设备中的气体或液体所承受的压力值
温度	反应器或管道中的温度值
转速	设备或泵的转速值
阻塞	管道或设备中的流体流量受阻，导致流量下降值或中断值
回流量	反应器或管道中的返流流体流量
开关机状态	设备或机器的开关机状态量
仪器仪表温度	生产过程中的仪器仪表的温度
电量	机器或设备所使用的电流量

Table 3. Coal chemical industry data
表 3. 煤化工行业数据

数据字段类型	数据字段解释
煤种	烟煤、无烟煤等不同煤的种类
煤的灰分	煤中灰分的含量
气化温度	煤气化过程中的温度
气化压力	煤气化时的压力环境
合成气组成	一氧化碳、氢气等合成气成分及比例
甲醇产量	甲醇的生产数量
焦油产量	煤焦油的生成量
脱硫效率	对煤气进行脱硫处理的效率
二氧化碳排放	生产过程中二氧化碳的排放量

2.2.1. 数据储存

数据储存为整合后的数据提供一个统一且高效的存储平台。在化工领域，由于数据源的多样性和数据量的庞大，数据储存系统的设计必须具备高效的数据处理能力、可靠性和灵活性。目前常见的数据存储方式包括关系型数据库(如 MySQL、PostgreSQL 等关系型数据库适用于存储结构化数据)、NoSQL 数据库(如 MongoDB、Cassandra 等适用于半结构化或非结构化数据)，分布式存储系统(如 Hadoop HDFS、Amazon S3 对大规模数据能够提供高效的存储管理)等，确保数据高可用性和可扩展性。

2.2.2. 数据清洗

数据清洗方面，考虑化工企业的能源数据，当原始电表数据因传输干扰或设备故障，常存在缺失值与异常值。错误数据的清除可以通过人工修正或使用知识工程工具识别并修正违规数据。常用的工具包括 ETL 工具和数据迁移工具[14]。此时，对于少量连续缺失值，采用线性插值法填补，依据相邻时段数据趋势估算缺失时段用电量；对于孤立异常高值(如尖峰电流冲击导致)，结合箱线图法识别，以历史同期数据及设备额定功率为参考，判断异常值并替换为合理估计值，确保电力消耗数据连续性与合理性。数据清理直接影响数据质量，是异构互通中必不可少的部分[15]。

2.2.3. 数据转换

数据转换针对不同数据源格式与单位差异。化工企业中，生产部门物料流量数据可能采用体积流量(立方米/小时)，而碳排放核算需转换为质量流量(千克/小时)，依据物料密度特性(如不同纯度硫酸密度各异)，通过实时查询物料主数据中的密度值进行换算。时间格式统一也不可或缺，部分设备记录采用本地时间，而企业分析要求统一为格林威治标准时间(GMT)，利用时间函数库在数据集成阶段进行批量转换，保障时间序列数据一致性，便于后续按日、月、年不同周期分析碳排放趋势。

2.2.4. 数据归一化

归一化处理保障不同量级数据可比性。在考量化工企业多种能源碳排放贡献时，电力消耗数据数值较大(兆瓦时量级)，天然气消耗数据相对较小(立方米量级)，采用最小 - 最大归一化方法，将各能源消耗值映射到[0, 1]区间，使不同能源数据在同一尺度下参与碳排放强度计算，避免因数据量级差异掩盖能源结构对碳排放的真实影响，经预处理后的数据质量显著提升，为构建高质量碳数据集奠定坚实基础。

2.3. 数据分析与决策支持模块

数据分析模块是多源碳数据集架构的高级功能模块, 涵盖碳监测、足迹追踪、核算报告、能效优化及碳交易策略。该模块基于《中国石油化工企业温室气体排放核算方法与报告指南(试行)》等标准计算碳排放, 结合机器学习和大数据方法, 对集成后的碳排放数据进行分析, 识别碳排放热点、优化工艺参数, 揭示碳排放的规律、趋势, 预测未来的碳排放量, 为决策者提供有力支持。主要需包含以下功能。

一是描述性分析与可视化, 通过数据可视化工具(如 Tableau、Power BI 等)展示碳排放的基本情况、分布情况、趋势变化等, 为决策者提供直观的分析结果。描述性分析帮助决策者理解现有的碳排放水平。

二是预测性分析与建模, 利用机器学习和统计学方法(如时间序列分析、回归分析、神经网络等), 进行碳排放量的预测。这些预测可以为碳排放减排措施的制定提供依据。例如, 结合 K-means 聚类等无监督学习方法, 对实际化工生产过程中的能耗与碳耗进行归类分析。在识别低能耗和低碳排放特征工况的基础上, 借助数据可视化与归类技术, 构建生产参数影响因素优化模型, 进而提取关键工艺参数的最佳运行区间, 为节能减排提供决策依据。同时参照改进的 K-means 算法[16]。另一种利用机器学习分析预测的方法是 BP 神经网络算法, 可参考陈世超[17]、陈天生和邱嘉艳[18]、叶鑒芳等人[19]的研究。

三是基于分析结果, 系统进一步通过优化算法为碳排放优化提供具体建议。

3. 结论与展望

本研究围绕化工行业多源异构碳数据集的构建展开深入探讨, 并取得了重要实践成果。基于多源数据融合、数据标准化和碳足迹核算理论, 构建了涵盖数据采集与储存、清洗转化和应用分析的分层架构。该架构全面整合生产、能源消耗、供应链等多源数据, 运用先进技术实现精准采集、高效预处理、安全存储与深度分析, 成功构建出高质量碳数据集, 打破企业数据孤岛, 提升数据完整性与质量。总体来说, 本研究成果紧密贴合化工企业碳中和实践需求, 为企业精准碳排放管控、科学制定减排策略提供了坚实的技术支撑, 助力推动化工行业绿色低碳转型。

未来, 化工行业多源异构碳数据集构建将呈现多维度发展趋势。一是跨学科研究深度融合是关键方向, 将计算机科学、化学工程、环境科学、碳排放管理学等多学科知识深度交织, 从化工工艺流程革新、碳排放源精准识别、数据管理系统科学构建、企业减排策略优化等全方位协同发力, 为碳中和目标下的化工行业转型提供系统性解决方案。二是新技术融合应用将持续深化, 5G、物联网、区块链等新兴技术在化工碳数据管理各环节将实现更紧密融合。三是国际合作交流将日益紧密, 我国化工企业可借鉴国际先进碳数据核算标准、管理模式, 参与国际规则制定, 同时输出自身特色技术与实践成果, 通过国际合作提升我国化工行业碳数据管理国际影响力。

基金项目

四川省自然科学基金青年基金项目(2025ZNSFSC1264); 四川轻化工大学“652”科研创新团队(SUSE652A003); 广西重点研发计划项目(桂科 AB24010156)。

参考文献

- [1] Dong, F., Zhu, J., Li, Y., Chen, Y., Gao, Y., Hu, M., et al. (2022) How Green Technology Innovation Affects Carbon Emission Efficiency: Evidence from Developed Countries Proposing Carbon Neutrality Targets. *Environmental Science and Pollution Research*, 29, 35780-35799. <https://doi.org/10.1007/s11356-022-18581-9>
- [2] Bruckner, B., Hubacek, K., Shan, Y., Zhong, H. and Feng, K. (2022) Impacts of Poverty Alleviation on National and Global Carbon Emissions. *Nature Sustainability*, 5, 311-320. <https://doi.org/10.1038/s41893-021-00842-z>
- [3] 熊肖磊, 王春伟, 赵炯, 等. 基于 Redis 与 SSM 的大型设备数据运用系统设计[J]. 现代机械, 2018(6): 29-34.

-
- [4] 赵德基, 王力, 狄军峰. 基于 Dubbo+NoSQL 的工业领域大数据平台研究[J]. 数字技术与应用, 2017(7): 64-67.
- [5] 王宏志, 梁志宇, 李建中, 等. 工业大数据分析综述: 模型与算法[J]. 大数据, 2018, 4(5): 62-79.
- [6] Hussain, M., Mir, G.M., Usman, M., Ye, C. and Mansoor, S. (2020) Analysing the Role of Environment-Related Technologies and Carbon Emissions in Emerging Economies: A Step towards Sustainable Development. *Environmental Technology*, **43**, 367-375. <https://doi.org/10.1080/09593330.2020.1788171>
- [7] Hardiyansah, M., Agustini, A.T. and Purnamawati, I. (2021) The Effect of Carbon Emission Disclosure on Firm Value: Environmental Performance and Industrial Type. *The Journal of Asian Finance, Economics and Business*, **8**, 123-133.
- [8] Zhang, W., Zhu, Z., Liu, X. and Cheng, J. (2022) Can Green Finance Improve Carbon Emission Efficiency? *Environmental Science and Pollution Research*, **29**, 68976-68989. <https://doi.org/10.1007/s11356-022-20670-8>
- [9] Raihan, A., Begum, R.A., Said, M.N.M. and Pereira, J.J. (2022) Relationship between Economic Growth, Renewable Energy Use, Technological Innovation, and Carbon Emission toward Achieving Malaysia's Paris Agreement. *Environment Systems and Decisions*, **42**, 586-607. <https://doi.org/10.1007/s10669-022-09848-0>
- [10] Wiedmann, T., Chen, G., Owen, A., Lenzen, M., Doust, M., Barrett, J., et al. (2020) Three-Scope Carbon Emission Inventories of Global Cities. *Journal of Industrial Ecology*, **25**, 735-750. <https://doi.org/10.1111/jiec.13063>
- [11] Chen, J., Gui, W.L. and Huang, Y.Y. (2022) The Impact of the Establishment of Carbon Emission Trade Exchange on Carbon Emission Efficiency. *Environmental Science and Pollution Research*, **30**, 19845-19859. <https://doi.org/10.1007/s11356-022-23538-z>
- [12] 贺雅琪. 多源异构数据融合关键技术研究及其应用[D]: [硕士学位论文]. 成都: 电子科技大学, 2018.
- [13] 马吉军, 贾雪琴, 寿颜波, 等. 基于边缘计算的工业数据采集[J]. 信息技术与网络安全, 2018, 37(4): 91-93.
- [14] Chu, X. and Ilyas, I.F. (2016) Qualitative Data Cleaning. *Proceedings of the VLDB Endowment*, **9**, 1605-1608. <https://doi.org/10.14778/3007263.3007320>
- [15] Yang, D.H., Li, N. and Wang, H.Z. (2016) Optimization of Parallel Big Data Cleaning Process Base on Task Merging. *Chinese Journal of Computers*, **1**, 97-108.
- [16] 徐健锐, 詹永照. 基于 Spark 的改进 K-Means 快速聚类算法[J]. 江苏大学学报(自然科学版), 2018, 39(3): 316-323.
- [17] 陈世超, 崔春雨, 张华, 等. 制造业生产过程中多源异构数据处理方法综述[J]. 大数据, 2020, 6(5): 55-81.
- [18] 陈天生, 邱嘉艳. 人工神经网络预测煤的发热量[J]. 煤质技术, 2006(4): 56-58.
- [19] 叶鑫芳, 钟志鹏, 郑仁广. 基于碳电强度的碳排放监测方法[J]. 能源与环境, 2023(1): 40-44.