

熟料等水泥基材料数据库及其应用

许成文¹, 叶家元¹, 高霖², 高国贤³, 任雪红¹, 夏凌风⁴, 王琳⁵, 张文生¹

(1. 中国建筑材料科学研究总院有限公司, 绿色建筑材料国家重点实验室, 北京 100024; 2. 合肥水泥研究设计院有限公司, 水泥制造绿色低碳技术安徽省重点实验室, 合肥 230051; 3. 中国科学院过程工程研究所, 介科学与工程全国重点实验室, 北京 100190; 4. 中存大数据科技有限公司, 北京 100024; 5. 济南大学, 山东省网络环境智能计算技术重点实验室, 济南 250022)

摘 要: 在“双碳”政策管控与人工智能浪潮的推动下, 水泥基材料亟需进行以低碳化为主要目标的数字化转型, 而建立水泥基材料数据库是数字化转型的首要任务。在科学文献中存在大量数据, 可采用人工智能方式进行自动化提取与应用。本文综述了熟料、水泥、混凝土、建筑砂浆等水泥基材料数据库的现状, 归纳了现有数据库收录的数据类别、应用方式、应用范围等信息, 分析了文献数据挖掘等人工智能技术在水泥基材料数据搜集与数据库建设中的应用前景, 指出了现有数据库存在的数据底座不统一、数据项考虑不周全、部分类别材料数据库不完善等问题, 并探讨了其未来发展方向。

关键词: 水泥基材料; 数据库; 数据挖掘; 人工智能

中图分类号: TU528; TP311.13 文献标志码: A 文章编号: 0454-5648(2025)05-1328-11

网络出版时间: 2025-01-08



水泥基材料是建筑行业最为常用的建筑材料之一。2023 年, 中国的水泥产量为 20.23 亿 t, 按照每吨水泥碳排放约 600 kg 计算^[1], 我国 2023 年水泥行业的碳排放在 12 亿 t 以上。基于我国的发展形势, 未来我国仍将长期处于基础建设的高峰期, 水泥基材料的整体需求仍将长期处于旺盛期。在我国“双碳”政策的逐步要求下, 水泥基材料行业的整体碳减排压力巨大, 低碳水泥基材料的研发逐渐成为水泥行业的重点课题之一。

传统的水泥基材料研发以实验为主, 研发效率慢, 材料创新难度较大。在水泥基材料的低碳研发目标下, 传统研发范式难以满足日益严峻的碳减排需求, 水泥行业亟需开发高效率的研发范式。随着大数据与人工智能技术在近 10 年来的快速发展, 尤其是近 3 年来由 ChatGPT 引起的大模型热潮, 以机器学习为核心的人工智能技术逐步应用于材料研发领域, 并在钢铁^[2-3]、生物^[4]、医药^[5-6]、有机^[7]等材料研发中取得了突破性进展, 实现了基于特定目标的新材料研发。水泥基材料相比于上述领域起步较晚, 目前也已开展了针对性能预测^[8-9]、组成优化^[10-11]、

低碳研发^[12]等多种目标的数字化模型开发, 主要针对模型的开发与算法的优化展开研究。

在模型开发中, 选用的数据直接影响构建模型的准确性, 主要表现为数据误差会影响模型对数据规律的提取与学习, 降低模型评估过程的可靠性, 从而影响模型的泛化能力。对于不同来源的数据, 其数据维度的差异会造成数据清洗、整合等工作难度的提升, 增加模型的复杂性与构建成本, 降低研发效率。准确、高质量的数据库/集是实现模型开发的基础, 可显著降低模型建立难度, 减少优化工作量, 提升模型预测精度与泛化能力, 促进数字化技术在水泥基材料研发中应用。因此, 高质量研发用数据库也逐渐成为水泥基材料研发的重点之一。

1 水泥基材料数据库现状

水泥基材料数据库建设的主要对象包括硅酸三钙与硅酸二钙等矿物与水泥熟料、混凝土、建筑砂浆等, 以及水泥基材料的水化产物等。水泥基材料数据库一般包括用户层、操作层、数据层、服务层等多个层级, 用户可通过用户层进行操作界面访问,

收稿日期: 2024-09-13。 修订日期: 2024-10-24。

基金项目: 中国建材集团原创技术策源地“揭榜挂帅”任务-“水泥基材料数字化研发”项目(2021YCJS01)。

第一作者: 许成文(1992—), 男, 博士。

通信作者: 张文生(1969—), 男, 博士, 教授。

Received date: 2024-09-13. Revised date: 2024-10-24.

First author: XU Chengwen (1992—), male, Ph.D.

E-mail: isaac1964620@126.com

Correspondent author: ZHANG Wensheng (1969—), male, Ph.D., Professor.

E-mail: wensheng_zhang01@163.com

数据库后台则通过服务层与数据层交互进行数据展示与功能实现, 如图 1 所示。在数据层中, 除了具

体的组成及材料数据之外, 一般还包括用户数据、模型数据、计算数据等多种类别的数据。

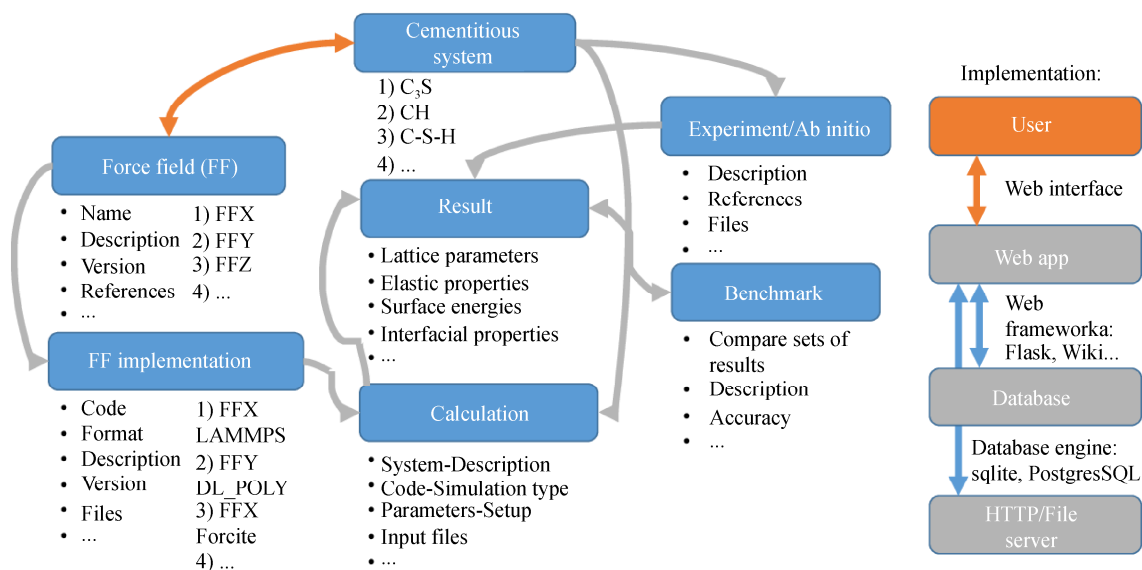


图 1 水泥基材料数据库基本框架(以 cemff 数据库为例)^[13]

Fig. 1 Basic framework of the cementitious materials database (using the cemff database as an example)

1.1 水泥熟料数据库

水泥熟料是由石灰石、黏土及铁质原料为主要原料, 按照适当比例配置成生料, 烧至部分或全部熔融, 并经冷却而获得的半成品。水泥熟料的主要矿物包括硅酸三钙、硅酸二钙、铝酸三钙、铁铝酸四钙、硫铝酸钙等, 按其矿物组成可分为硅酸盐熟料、铝酸盐熟料、硫铝酸盐熟料等。熟料的物性由其组成矿物及其比例决定, 因此, 国内外学者对熟料矿物的各种数据进行了汇总分析并建立了相应的数据库^[13-16]。熟料物相数据库主要记录不同熟料的矿物组成及其主要水化产物, 重点关注矿物的晶体学参数、热力学参数、原子力场参数、胶凝活性数据等。

水泥熟料矿物的晶体学参数是决定矿物性质的最基本信息。作为典型的无机非金属化合物, 水泥熟料矿物的晶体学参数目前已被记录于现有的多种无机化合物晶体学参数数据库中, 如表 1 所示。

如表 1 所示, 现有的无机化合物晶体学参数数据库中, 水泥熟料矿物的晶体学参数只占其很少一部分。由于原材料的杂质在水泥熟料煅烧过程中会取代熟料矿相结构中的部分原子, 从而造成晶型转变、杂质缺陷、错位等现象, 晶体学参数也因此发生变化^[17]。这些晶体结构的变化可反映离子掺杂对水泥熟料矿物活性的影响。因此, 为了更好地研究水泥熟料矿相及离子掺杂等技术手段对其性能的影响, 水泥熟料矿

相数据库在建设过程中有必要考虑离子掺杂引起的熟料矿物晶体学参数变化情况。

水泥熟料矿物的原子力场参数能够反映其水化反应机制、性能演化规律以及替代胶凝材料与载荷等对熟料及其产物的影响等多方面进行微观层面的阐述, 从而实现对水泥基材料的服役寿命进行准确预测。基于水泥熟料矿物的原子力场参数, 国内外学者建立了含黏土相材料的矿物-水界面分子模型 ClayFF^[18-19]、水泥基材料系统模型 CementFF^[20-22]和水化硅酸钙凝胶模型 CSH-FF^[23-25], 并修正了无机材料领域通用模型 IFF^[26-28]与 ReaxFF^[29-31], 用于描述水泥熟料的结构及其水化过程。Mishra 等^[13]建立了 cemff 水泥基材料原子力场参数数据库, 搜集汇总了来自 OpenKIM 与 NIST(National Institute of Standards and Technology)的原子力场参数以及国内外学者建立的水泥熟料矿物的原子力场模型, 为相关研究人员提供了原子力场数据查询与水泥熟料矿物原子力场模型验证等功能。

水泥熟料矿物的热力学参数可反映矿物形成、水化等过程的难易程度, 可反映矿物煅烧的可行性及由此矿物组成的水泥熟料水化特性(见表 1^[32-37])。CALPHAD 方法广泛应用于多元化合物体系的相图预测, 是进行材料热力学分析的主要方法之一。2000 年, 该方法首次应用于水泥熟料热力学建模、模型验证与数据计算^[38]。目前, 基于这一方法已形成成熟

表 1 现有记录水泥熟料矿物晶体学数据的数据库^[32–37]Table 1 Current databases containing crystallographic data of clinker minerals^[32–37]

No.	Database name	Introduction	Data quantity
1	Cambridge Structural Database(CSD)	A structural database of small molecule and metal organic molecule crystals based on X-ray and neutron diffraction experiments provided by Cambridge Crystal Data Center, which essentially includes all published data on organic and metal organic compound crystals with atomic numbers (including hydrogen atoms) up to 500.	Over 1.29 million crystal data, but no relevant data on cement clinker minerals has been included yet.
2	Inorganic Crystal Structure Database(ICSD)	Edited jointly by Fachinformationzentrum Karlsruhe and The Gmelin Institute from Germany, this database collects and provides information on the crystal structures of all inorganic substances, excluding metals and alloys, that have been experimentally determined to date and do not contain C—H bonds.	More than 120 000 crystal data, including over 200 data related to cement clinker minerals and hydration products.
3	Springer Materials	A comprehensive database covering multiple material categories, performance types, and applications, providing visual data display, material search, and other functions. The data is sourced from the Landolt-Börnstein collection, MSI Eureka, Linus Pauling Files, Polymer Thermodynamics Database (ATHAS), Dortmund Separation Technology Database, etc.	More than 270 000 crystal data, over 1.17 million structural performance data, and over 40 000 phase diagrams, including over 200 data related to cement clinker minerals and hydration products.
4	Crystallography Open Database(COD)	An open-source database containing organic, inorganic, metal organic compounds, and mineral crystal structures.	More than 510 000 crystal data, including over 100 data related to cement clinker minerals and hydration products.
5	American Mineralogist Crystal Structure Database(AMCSD)	A database includes all crystal structures published in <i>American Mineralogist</i> , <i>The Canadian Mineralogist</i> , <i>European Journal of Mineralogy</i> , and <i>Physics and Chemistry of Minerals</i> , as well as selected data from other journals. This database is maintained by the American Mineralogical Society and the Canadian Mineralogical Association, and is funded by the National Science Foundation of the United States.	4292 minerals, with over 8000 data, including more than 1700 pieces of data related to cement clinker minerals and hydration products.
6	International Centre for Diffraction Data(ICDD)	An International Diffraction Data Centre database based on powder diffraction data of crystal materials in published literature, covering crystal names, chemical compositions, experimental conditions, diffraction d-I/I values, etc.	More than 760 000 material data, including over 5900 data related to cement clinker minerals and hydration products.
7	China Building Materials Academy Cement Clinker Mineral Phase Database(CBMA-CCMPD)	A database established by China Academy of Building Materials Science and Technology Co., Ltd., including crystallographic data, thermodynamic data, cementitious activity data, and carbon emission data of cement clinker obtained through literature collection and laboratory testing.	More than 300 cement clinker data, including over 100 crystallographic data.

的商业软件, 所得数据已建成 3 个水泥熟料热力学商业数据库: 基于 Thermo-Calc 的 ION3 数据库、基于 FactSage 的 FToxid 数据库以及基于 MTDATA 的 NPL oxide 数据库^[39–42]。上述数据库充分研究和讨论了 CaO–SiO₂ 等二元体系的热力学模型, 并对部分三元体系的热力学模型进行了初步验证。Noirfontaine 等^[16]采用 CALPHAD 方法对现有的水泥熟料热力学商业数据库进行数据与模型的进一步验证及完善, 重点扩充了涉及 Fe₂O₃、MgO、SO₃、P₂O₅ 以及碱金属氧化物的多元体系。美国国家标准与技术研究院(NIST)建立的水泥熟料热力学数据库^[43]收录了 298.15 K 与 1bar 条件下的水泥熟料热力学数据, 并对铝酸钙的生成热与氧化钙的生成热等数据进行

了纠正^[44]。Hanein 等^[14]在上述数据库的基础上验证了熟料的矿物组成数据、煅烧过程中相变数据以及水泥熟料亚固相线温度下的矿相数据, 为研究人员提供了可用于熟料烧成的热力学数据集。

此外, 水泥熟料水化产物的热力学数据能够反映水泥熟料在不同环境或矿物掺合料条件下的水化过程, 也是水泥熟料热力学数据库建设的主要方向之一。Lothenbach 等^[15]基于 GEMS 和 PHREFQC 计算数据与近年来公布的实验数据构建了熟料水化产物热力学系列数据库 Cemdata。该系列数据库搜集了 C-S-H 凝胶、AFm 和 AFt 相、水合石榴石、水滑石、沸石和 M-S-H 凝胶等常见水泥熟料水化产物的热力学数据, 数据有效温度范围下限为 0 °C, 上限

为 100 °C。目前该系列数据库已更新至第 3 版 Cemdata18, 并囊括了 AFm、AFt、C-S-H 凝胶和 M-S-H 凝胶的固溶体模型^[45-49]。Zhou 等^[50]利用该数据库建立了低热硅酸盐水泥熟料的水化热力学模型并进行了验证, Liao^[51]与 Zhuang 等^[52]分别利用该数据库数据建立了煅烧磷石膏、熟石灰及钢渣等矿物掺合料对水泥熟料水化影响的热力学模型, 实现了对水泥熟料及其水化产物的模拟及溶解性质描述。目前, 已经建立了水泥熟料及其水化产物的热力学数据库, 国内外学者不断完善数据内容, 且基于此数据库建立并逐步验证了水泥熟料烧成过程及熟料-矿物掺合料协同作用下水化反应的热力学模型, 实现了对水泥熟料烧成过程与水化反应的准确描述。

水泥熟料矿物的胶凝活性数据可直接反映其水化活性、强度等信息, 可判断该矿物是否适于作为熟料的矿物组成。Yan 等^[53]研究了不同磷元素取代的 C_4A_3S 矿物的胶凝活性数据, 提出了不同磷取代率下 C_4A_3S 矿物水化产物组成数据与抗压强度数据, 记录了磷元素取代下的胶凝活性数据。Kriskova 等^[54]研究了 γ - C_2S 的水化活性, 记录了机械活化与化学激发下 γ - C_2S 的水化产物组成数据与抗压强度数据。王晓东^[55]研究了磷铝酸钙矿物的水化胶凝特性, 通过分析其水化产物组成数据与抗压强度数据, 设计了高胶凝性的 $CaO-Al_2O_3-P_2O_5$ 熟料体系。目前, 虽然有大量学者针对熟料矿物的胶凝活性进行了研究^[56-59], 但由于熟料种类繁多, 且存在离子掺杂等复杂情况, 尚无有公信力的熟料矿相胶凝活性数据库。

综上所述, 目前针对水泥熟料矿物及其水化产物已建立了数据库, 部分数据库还集成了多种可用于水化分析的模型, 实现了数据库数据的高效利用。然而, 目前水泥熟料矿物的晶体学参数主要采信于无机非金属材料晶体学数据库, 尚无专门的水泥熟料矿物晶体学参数数据库。此外, 当前的晶体学参数数据库并未收录离子掺杂而导致的晶体学参数变化数据, 这在一定程度上影响了离子掺杂作用调控熟料矿物活性及低碳水泥熟料体系设计等方面的数字化研发。

1.2 水泥数据库

由于水泥的品种复杂、原材料种类多样、应用与研发场景繁多, 水泥数据库的建设主要针对特定的研发场景, 从场景需求出发进行数据搜集、整理以及数据库建设。

Blanc 等^[60]从黏土等材料的利用角度出发, 构建了水泥、沸石、黏土数据库 ThermoChimie。该数据库提供了上述 3 类材料的热力学数据, 其基于物相碱度、化学反应趋势等综合考虑, 保证了数据的一致性。Abraham 等^[61]从油井钻采对水泥基材料的性能要求出发, 建立了油井水泥材料数据库, 收录了 1000 余个水泥样本, 涵盖水泥类型、外加剂种类、不同龄期水泥性能(尤其是长龄期的性能)等数据, 整合了抗压强度、稠化时间、强度发展、密度等文献数据与实验室数据, 为油井水泥的开发与应用提供了数据基础。

目前, 分布式控制系统(DCS)等水泥产线数字化系统在国内水泥企业中广泛应用。DCS 系统的应用使得水泥企业在生产过程中积累了大量的数据。为了更好地管理与利用这些数据, 国内外学者重点关注基于产线数据的智能控制与工艺优化^[62-63]。Tkachenko 等^[64]结合地理空间计算机视觉和大语言模型技术, 为研究人员提供了全球水泥企业生产数据的精准查询, 并通过碳排放模型实现了企业碳数据的准确预测。此外, 水泥产线为了保障产品质量的稳定, 其工艺数据与实验室检测数据普遍具有数据波动小、数据范围集中等特点, 针对固定产线的产品性能预测与工艺优化具有一定效果, 但较难迁移应用于新型水泥基材料研发与制造。

1.3 混凝土数据库

混凝土是水泥的主要应用方式, 也是世界上应用最为广泛的建筑材料。混凝土的工作性能、强度、耐久性能直接影响混凝土构筑物的使用寿命。然而, 上述性能的影响因素各不相同, 国内外学者主要针对单一性能或具有密切关系相关性的数个性能进行混凝土材料性能数据库建设。

混凝土材料的强度是评价混凝土材料的主要性能指标, 包括抗压强度、抗拉强度、抗折强度等。Jayasuriya 等^[65]搜集了 115 篇文献的再生粗骨料混凝土材料强度数据, 构建了数据库; 其通过统计分析以非线性回归方程形式展示了强度变化趋势, 讨论了再生混凝土骨料替代不同比例天然骨料对混凝土性能的影响。Wendner 等^[66]建立了混凝土材料强度数据库, 并将其用于混凝土强度模型的校准和验证。该数据库包括 4 种尺寸的挠曲试验、抗拉试验、约束和无约束抗压试验数据, 以及 5 种尺寸的劈裂试验数据, 并包括加载和卸载数据。相关数据来源为 152 根有缺口和无缺口梁、32 个圆柱体试块、28 个立方体试块、46 个棱柱试块的测试结果。

混凝土材料耐久性是评价混凝土材料服役寿命的主要参考指标,包括收缩徐变、抗碳化、抗冻、抗氯离子侵蚀、抗硫酸盐侵蚀等。美国西北大学于1978年开始建设混凝土材料收缩与徐变性能数据库^[67];1992年,在美国混凝土协会(ACI)与国际材料与结构研究实验联合会(RILEM)联合扩展下形成RILEM-ACI 209 数据库^[68–71];在2008年^[72]与2015年^[73],分别进行了数据,目前拥有1400余条徐变曲线、1800余条收缩曲线,其中约800条徐变曲线和1050条收缩曲线所测试混凝土含有矿物掺合料。该数据库为混凝土材料收缩徐变研究人员提供了翔实的混凝土配合比、测试条件、试块几何尺寸数据等相关信息,为混凝土性能的可控设计提供了基础数据。方修春^[74]建立了混凝土材料碳化深度与钢筋锈蚀性能数据库,并通过BP神经网络与灰色系统模型实现了混凝土设计服役年限后碳化深度与钢筋锈蚀量的准确预测。高志斌^[75]等基于天津机场道路工程施工数据建立了混凝土材料抗冻性能数据库与评价系统,并通过抗冻性能预测模型指导了机场路面混凝土配合比优化设计。梁程瑶^[76]采用人工采集手动录入的方式收录了2500余篇文献中海洋环境下混凝土材料抗压强度数据并采用相关性分析、灰色关联分析以及主成分分析3种数理统计方法对数据关系与数据规律进行了深入挖掘,为海洋环境中复杂离子侵蚀作用下混凝土强度、耐久性等数据关系的深入挖掘奠定了基础。Ma等^[77]提出了一种智能混凝土材料耐久性数据库的建设方式,其将特定产地耐久性数据与通用耐久性数据智能混合,根据产地数据的充分与否选择是否采用产地数据进行耐久性预测。采用这种方式,既保证了特定产地数据对产地混凝土耐久性预测的正面影响,又避免了产地数据缺失时预测准确性降低的问题。其建立的耐久性数据库为珠港澳大桥耐久性研究提供了数据基础。

混凝土材料的强度、耐久性等数据经国内外学者梳理总结,已建立了一些区域性的数据库,开展了强度模型、服役寿命预测模型、抗冻性能预测模型等数据模型的研发与验证工作。但也应注意到,在混凝土材料耐久性数据库建设与数据模型开发中,现有数据库虽然涵盖了收缩徐变、抗碳化、抗冻等性能,并对海洋工程等复杂离子侵蚀环境有所研究,但仍缺乏抗氯离子侵蚀、抗硫酸盐侵蚀等性能的专用数据库。此外,只有少数学者将材料产地数据纳入数据建模的考虑范围之中,大部分学者主要针对特定产地的混凝土材料进行数据库构建与模型开发。

1.4 其他相关数据库

粉煤灰、矿渣等辅助胶凝组分对水泥混凝土的性能有着重要影响,因此其对应的数据库建设显得尤为必要。

粉煤灰是火力发电行业的主要固体废弃物,由于其具有胶凝活性而受到了国内外学者广泛研究,与之相关的数据库建设工作也开展得最早。1990年,McCarthy等^[78]建立了最早的粉煤灰数据库,收录了北美地区粉煤灰的化学成分、矿物组成与物理性能数据。该数据库将178组粉煤灰样品按照氧化钙含量分为低钙(<10%)、中钙(10%~20%)和高钙(>20%)3类,并依据数据库记录信息分析了粉煤灰矿物组成与原煤种类的关系、粉煤灰中晶体与玻璃相分布规律、高钙粉煤灰的性质稳定性等。Diaz等^[79]搜集了16组不同燃煤发电厂的粉煤灰样本,汇集了化学元素分析数据、X射线衍射分析数据、粒度分布数据以及其制备的混凝土性能(凝结时间、抗压和抗折强度、弹性模量和泊松比等)数据,并分析了粉煤灰性能对混凝土性能的影响,对丰富粉煤灰数据库起到了积极作用。

随着矿渣、硅灰、精炼渣等固体废弃物在混凝土中的规模化应用,这些原材料数据库建设也受到了重视^[42, 80–81]。目前,这些原材料的相关数据库建设主要集中于数据汇集与数据模型开发阶段,其中矿渣数据库主要考虑了生产条件对其组成的影响。

除了混凝土之外,建筑砂浆也是水泥的主要产品之一。但由于其应用场景较少,数据积累难度较大,国内外学者对其进行的数据库建设研究较少,部分学者主要关注砂浆砌块强度^[82]、裂缝^[83]等信息,砂浆数据库的研发以及基于数据库的材料模型研发仍处于起步阶段。

1.5 水泥基材料数据的应用

水泥基材料数据目前的主要应用方向包括:

1) 数据“增删改查”。数据的添加、删除、修改与查询是数据库的基础功能,能为相关学者提供准确的数据信息。在数据的“增删改查”中,既完善了数据结构与内容,又提升了数据准确性。

2) 建立水化反应模型。水泥基材料的原子力场参数与热力学参数能够从不同角度建立水泥基材料水化反应模型,采用水化产物物相检测等数据对模型进行验证分析,可实现水泥基材料水化过程的准确描述。

3) 实现性能预测。数据库中的大量数据能够为相关学者提供数据建模基础,为生产工艺优化、水

泥性能预测、混凝土耐久性设计等提供模型与工具。

2 人工智能应用于数据库建设

传统的水泥基材料数据库的数据来源主要有以下几个方面: 1) 实验室数据积累; 2) 产线、工程数据获取; 3) 文献数据提取。对于这 3 个方面的数据, 一般通过人工采集的方式将数据录入现有数据库, 数据存储效率较低, 文献数据利用效率也不高, 存在数据范围集中、波动小等问题。随着 ChatGPT 引起新一轮人工智能热潮的到来, 基于数据挖掘的材料研发技术逐渐成熟, 并已在金属、生物、医药等领域取得了突破性进展^[2, 4, 6, 84]。在这一过程中, 越来越多的学者认识到高质量数据库对于人工智能赋能材料研发的重要性。因此, 以人工智能技术进行文献数据挖掘并建立基于文献的高质量数据库成为学者们的关注重点。

文献挖掘算法的本质是对采用自然语言处理算法对科学文献中的文本进行分析与识别, 从而挖掘出文献中所含的有价值信息。命名实体识别与文本关系提取是实现文献挖掘的两大重要任务。命名实体识别一般采用字典查询、规则制定、机器学习等方法, 需要相关领域专家进行字典构建、数据标注等工作。ChemDataExtractor^[85]采用内嵌机器学习算法与语法规则的字典, 实现了化学命名实体的识别与化合物关系提取。Court 等^[86]将 ChemDataExtractor 算法与半监督学习方法 Snowball

结合, 从数十万篇文献中提取了居里温度和奈尔温度。Wang 等^[2]开发了一种自动化自然语言处理算法, 结合启发式文本多重关系提取距离算法实现了无标记样本的合金数据自动采集, 准确性远高于 ChemDataExtractor 算法。此外, Wang 等^[3]还开发了一种提取合金合成制备工序的半监督文本挖掘方法, 能够在较少的专家知识与数据标注下生成完整的工艺流程字典, 实现较为准确的文本数据挖掘。

文献中的关键信息多为具有一定规律的数据, 以 word2vec^[87]和 GloVe^[88]为代表的单词嵌入算法可将重点单词表中的单词视为向量, 评价其在文献语句组成的高维空间中的规律, 从而迅速成为了实现文献挖掘的基础方法, 并在材料基因组计划中实现了材料化学组成的数据提取。作为以 ChatGPT 为代表的大语言模型的基础, Transformer 算法对单词嵌入算法进行了深入开发, 在捕捉单词依赖关系、解析上下文中专业术语的出现位置方面表现出了优异的性能^[89]。基于 Transformer 算法, 相关学者开发了 BERT 算法(Bidirectional Encoder Representations from Transformers, 基于 Transformer 的双向编码表示算法)^[90]。该算法具有数亿个可调节参数, 目前已经在通用文献、材料科学以及生物医药等方面实现了应用。Gilligan 等^[84]开发了一种基于 BERT 算法的无规则工作流程 BERT-PSIE, 以文本标注取代基于规则的语法设计, 实现了文献文本数据挖掘与结构化数据库的自动化构建。

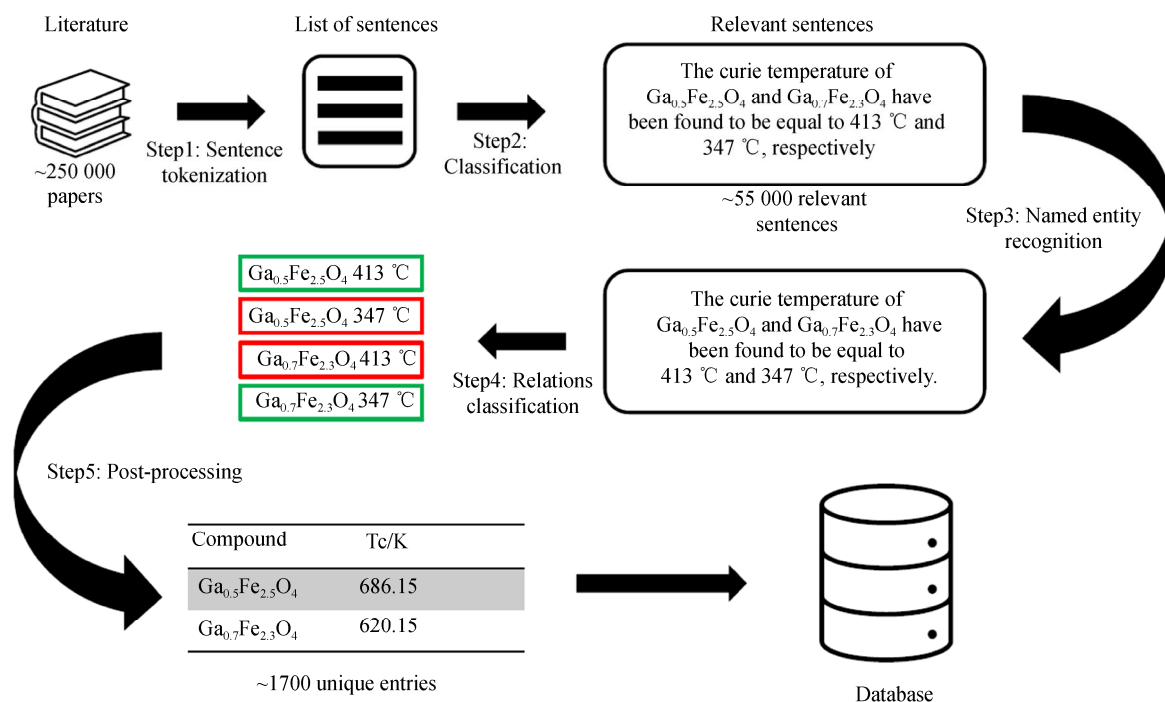


图 2 典型文献挖掘算法处理流程(以 BERT-PSIE 为例)^[84]

Fig. 2 Typical literature mining algorithm processing flow (taking BERT-PSIE as an example)^[84]

基于人工智能的文献挖掘技术在钢铁、生物、医药等领域的应用,为水泥基材料的数据库建设提供了可行新思路。然而,在具体的算法开发中,水泥基材料的文献数据挖掘技术仍存在着技术瓶颈:

1) 数据字典建立难度大。不同于钢铁等材料,水泥基材料种类繁多、成分复杂,在建立数据字典时,即便只考虑水泥熟料,其复杂的矿相组成与离子掺杂等作用会增加数据字典建立的难度。此外,不同水泥基材料存在着相同属性的性能数据,但必须要考虑不同因素在数据字典中的通用表达。例如,28 d 抗压强度是熟料、水泥、混凝土等水泥基材料的重要物理性能指标,但骨料类别、砂率、用水量等因素各不相同,必须建立准确的通用表达以实现对熟料、水泥、混凝土等不同实体的区分。

2) 实体关系复杂,文本关系提取困难。水泥基材料为多矿相组成的各向不均匀材料,且其还存在水化过程这一影响最终物相组成与微观结构的化学反应过程。此外,其强度等性能受成分与养护方式等多种因素的影响。这些性能、成分、养护方式数据以及这些数据之间的对应关系均需要通过文本关系提取来获得,这就要求研发人员必须充分理解水泥基材料性能检测方式、养护制度等数据,并具备数字化归纳能力。

3) 存在一定量图片数据准确识别需求。为了更好地展示数据规律,水泥基材料相关的文献中有相当比例的数据以图片形式呈现。为了更好地获取数据,研发人员需涉猎图片识别与处理技术。不同图片数据的识别难度存在差异,这就要求研发人员基于研发需求进一步开发相关算法。

3 水泥基材料数据库存在的问题及未来发展

3.1 存在的问题

目前大部分数据库只能满足水泥基材料研发的部分需求,主要问题如下:

1) 未考虑离子掺杂等作用对水泥熟料的影响。目前数据的描述主要集中于硅酸三钙等常见矿物,且并未考虑离子掺杂等作用对水泥熟料矿物晶体结构、水化活性、力学性能、制备方式等方面的影响,尚没有建立更为全面的熟料数据库。

2) 缺乏统一的水泥基材料数据底座与数据库建设规范,不同数据库之间的数据无法有效对接。同一熟料矿相的不同类别数据,以及熟料与水泥数据、水泥与混凝土数据等之间不能有效衔接,无法形成完整的数据链条,在一定程度上影响了数据建

模分析与材料新体系研发。

3) 水泥、混凝土等数据库建设主要集中于强度、徐变等信息,部分研究涉及抗冻、抗碳化等数据,缺少抗氯离子侵蚀、抗硫酸盐侵蚀等耐久性数据库建设。水泥基材料的性能复杂,除了工作性能、力学性能等基本性能之外,耐久性更关系到构筑物的使用和服役效能,但目前缺乏相应的耐久性数据库对已有文献数据或实验数据进行有效整理与充分应用。

4) 辅助胶凝组分等其他相关材料数据库建设尚不完善。虽然已经针对粉煤灰等原材料建立了完善的数据库,但矿渣、硅灰、钢渣等辅助胶凝组分的数据多以数据集的形式应用于数据建模中,缺少完善的数据库系统对其特性进行归纳和整理。

上述问题使得水泥基材料现有数据库表现为数据不完善、数据利用价值受限,只能为相关研发人员提供数据查询与单一方向的数据建模功能。

3.2 未来展望

随着人工智能技术的逐步发展,越来越多的学者认识到统一的数据库对于材料研发的重要性。鉴于现有水泥基材料数据库存在的数据缺陷问题,应从以下几个方向完善水泥基材料数据库:

1) 统一研发数据底座,有效衔接上下游数据。梳理水泥基材料上下游数据特点,研究矿物-熟料-水泥-砂浆/混凝土的全链条数据关系,重点梳理熟料矿物组成与熟料性能、水泥性能之间的关系数据,建立水泥基材料的一站式研发数据底座,形成前后紧密衔接的水泥基材料研发数据链。

2) 完善水泥基材料数据库建设,补充辅助胶凝材料等相关数据。以矿渣为代表辅助胶凝材料及以减水剂为代表的化学外加剂是水泥混凝土的重要组成部分,建设辅助胶凝材料、外加剂数据库,可数据化描述这些组分对水泥基材料工作性能、耐久性能等性能的影响,形成完善的水泥基材料数据库体系。

3) 发展适用于水泥基材料的文献挖掘技术,开发文献挖掘数据有效匹配现有数据库中数据、数据库自动更新等算法,加快数据搜集效率。水泥基材料的种类繁多、成分复杂,单个团队进行试验或人工收集数据的效率很低,难以完成水泥基材料全链条数据关系打通、数据库体系完善、数据填充等工作。大语言模型等人工智能技术能够极大地加快数据搜集效率,实现数据自动清洗、验证、填充等工作。然而,考虑到水泥基材料组成与性能的复杂多样性,进行文献挖掘算法开发时首先应从单一性能

出发, 结合数据搜集与工艺过程提取两方面需求, 对文献中有效数据进行全面挖掘, 再逐步扩展到多性能提取。此外, 在从文献中获取数据时, 应对已获取数据进行甄别, 部分数据可能会存在重复情况, 要做好数据匹配与数据库自动更新相关算法的配套开发。

通过上述方式完善的水泥基材料数据库, 可为水泥基材料数据模型构建提供全面准确的数据支撑, 提高数据模型在性能预测、结构设计、组分优化等方面的准确性, 提升数据模型的鲁棒性, 扩大其在水泥基材料研发、制造、应用等流程中的适用范围, 促进新型低碳水泥基材料的发现与应用。

4 结论

水泥基材料行业面临着数智化与低碳化的双重转型, 建立高质量数据库成为行业共识。本文综述了水泥基材料数据库的现状, 包括熟料数据库、水泥数据库、混凝土数据库等, 并讨论了人工智能技术在数据挖掘中的作用, 得出如下结论:

1) 当前已初步建成熟料、水泥、混凝土等材料数据库, 涵盖晶体学参数、热力学参数、工作性能、力学性能等多个类别的数据, 为水泥基材料数据建模奠定了基础。

2) 水泥基材料数据库体系仍不完善, 存在数据底座不统一、数据之间无法衔接等问题。

3) 数据种类存在“偏科”, 重点集中于水泥熟料数据库, 混凝土耐久性数据库等还不完善。

4) 大语言模型等人工智能技术可显著提高文献数据挖掘效率, 且已在钢铁等行业实现了应用, 可以为水泥基材料文献数据挖掘提供一定借鉴。

参考文献:

- [1] 魏军晓, 耿元波, 沈镭, 等. 基于国内水泥生产现状的碳排放因子测算[J]. 中国环境科学, 2014, 34(11): 2970–2975.
WEI Junxiao, GENG Yuanbo, SHEN Lei, et al. China Environ Sci, 2014, 34(11): 2970–2975.
- [2] WANG W R, JIANG X, TIAN S H, et al. Automated pipeline for superalloy data by text mining[J]. NPJ Comput Mater, 2022, 8: 9.
- [3] WANG W R, JIANG X, TIAN S H, et al. Alloy synthesis and processing by semi-supervised text mining[J]. NPJ Comput Mater, 2023, 9: 183.
- [4] LEE J, YOON W, KIM S, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234–1240.
- [5] YANG L, XU S, SELLERGREN A, et al. Advancing multimodal medical capabilities of gemini[EB/OL]. 2024: arXiv: 2405.03162. <http://arxiv.org/abs/2405.03162>
- [6] SAAB K, TU T, WENG W H, et al. Capabilities of gemini models in medicine[EB/OL]. 2024: arXiv: 2404.18416. <http://arxiv.org/abs/2404.18416>
- [7] BAUM Z J, YU X, AYALA P Y, et al. Artificial intelligence in chemistry: Current trends and future directions[J]. J Chem Inf Model, 2021, 61(7): 3197–3212.
- [8] 张文生, 曹传荔, 郑晓, 等. 机器学习方法用于水泥基材料的研究进展[J]. 硅酸盐学报, 2024, 52(11): 3617–3630.
ZHANG Wensheng, CAO Fuli, ZHI Xiao, et al. J Chin Ceram Soc, 2024, 52(11): 3617–3630.
- [9] 徐潇航, 胡张莉, 刘加平, 等. 基于机器学习回归模型的三峡大坝混凝土强度预测[J]. 材料导报, 2023, 37(2): 41–49.
XU Xiaohang, HU Zhangli, LIU Jiaping, et al. Mater Rep, 2023, 37(2): 41–49.
- [10] 陈云, 付前旺. 机器学习和元启发式算法优化混凝土配合比研究进展[J]. 复合材料学报, 2024: 1–39.
CHEN Yun, FU Qianwang. Acta Materiae Compositae Sinica, 2024: 1–39.
- [11] 胡定宇. 基于机器学习和优化算法的混凝土性能预测及配合比设计[D]. 重庆: 重庆交通大学, 2022.
HU Dingyu. Performance prediction and mix design of concrete based on machine learning and optimization algorithm[D]. Chongqing: Chongqing Jiaotong University, 2022.
- [12] 任强生. 碱渣-矿渣基胶凝材料的改性及充填材料的制备研究[D]. 秦皇岛: 燕山大学, 2023.
REN Qiangsheng. Study on modification of soda residue-ground granulated blast slag based cementitious materials and preparation of cemented paste backfills[D]. Qinhuangdao: Yanshan University, 2023.
- [13] MISHRA R K, MOHAMED A K, GEISSBÜHLER D, et al.: A force field database for cementitious materials including validations, applications and opportunities[J]. Cem Concr Res, 2017, 102: 68–89.
- [14] HANEIN T, GLASSER F P, BANNERMAN M N. Thermodynamic data for cement clinkering[J]. Cem Concr Res, 2020, 132: 106043.
- [15] LOTHENBACH B, KULIK D A, MATSCHEI T, et al. Cemdata18: A chemical thermodynamic database for hydrated Portland cements and alkali-activated materials[J]. Cem Concr Res, 2019, 115: 472–506.
- [16] DE NOIRFONTAINE M N, TUSSEAU-NENEZ S, GIROD-LABIANCA C, et al. CALPHAD formalism for Portland clinker: Thermodynamic models and databases[J]. J Mater Sci, 2012, 47(3): 1471–1479.
- [17] 魏丽颖, 刘姚君, 汪澜, 等. 重金属在水泥熟料矿物中的固溶机理研究进展[J]. 硅酸盐通报, 2014, 33(10): 2541–2546.
WEI Liying, LIU Yaojun, WANG Lan, et al. Bull Chin Ceram Soc, 2014, 33(10): 2541–2546.
- [18] SHAHSAVARI R, PELLENG R J M, ULM F J. Empirical force fields for complex hydrated calcio-silicate layered materials[J]. Phys Chem Chem Phys, 2011, 13(3): 1002–1011.
- [19] CYGAN R T, LIANG J J, KALINICHEV A G. Molecular models of hydroxide, oxyhydroxide, and clay phases and the development of a general force field[J]. J Phys Chem B, 2004, 108(4): 1255–1266.
- [20] SCHRÖDER K P, SAUER J, LESLIE M, et al. Bridging hydroxyl groups in zeolitic catalysts: A computer simulation of their structure, vibrational properties and acidity in protonated faujasites (H–Y zeolites)[J]. Chem Phys Lett, 1992, 188(3/4): 320–325.
- [21] FREEMAN C L, HARDING J H, COOKE D J, et al. New forcefields for modeling biomineralization processes[J]. J Phys Chem C, 2007, 111(32): 11943–11951.
- [22] LEWIS G V, CATLOW C A. Potential models for ionic oxides[J]. J Phys C Solid State Phys, 1985, 18(6): 1149–1161.

- [23] GALE J D. Empirical potential derivation for ionic materials[J]. *Philos Mag B*, 1996, 73(1): 3–19.
- [24] MOGHADDAM S E, HEJAZI V, HWANG S H, et al. Morphogenesis of cement hydrate[J]. *J Mater Chem A*, 2017, 5(8): 3798–3811.
- [25] BONACCORSI E, MERLINO S, KAMPF A R. The crystal structure of tobermorite 14 Å (plombierite), a C–S–H phase[J]. *J Am Ceram Soc*, 2005, 88(3): 505–512.
- [26] MISHRA R K, FERNÁNDEZ-CARRASCO L, FLATT R J, et al. A force field for tricalcium aluminate to characterize surface properties, initial hydration, and organically modified interfaces in atomic resolution[J]. *Dalton Trans*, 2014, 43(27): 10602–10616.
- [27] MISHRA R K, FLATT R J, HEINZ H. Force field for tricalcium silicate and insight into nanoscale properties: Cleavage, initial hydration, and adsorption of organic molecules[J]. *J Phys Chem C*, 2013, 117(20): 10417–10432.
- [28] HEINZ H, LIN T J, MISHRA R K, et al. Thermodynamically consistent force fields for the assembly of inorganic, organic, and biological nanostructures: The INTERFACE force field[J]. *Langmuir*, 2013, 29(6): 1754–1765.
- [29] DEHNE G C. Review of McNaught & Wilkinson (1997): Compendium of chemical terminology, IUPAC recommendations[J]. 1997, 4(2): 347–351.
- [30] VAN DUIN A C T, DASGUPTA S, LORANT F, et al. ReaxFF: A reactive force field for hydrocarbons[J]. *J Phys Chem A*, 2001, 105(41): 9396–9409.
- [31] VAN DUIN A C T, STRACHAN A, STEWMAN S, et al. ReaxFFSiO reactive force field for silicon and silicon oxide systems[J]. *J Phys Chem A*, 2003, 107(19): 3803–3811.
- [32] ALLEN F H, TAYLOR R. Research applications of the Cambridge Structural Database (CSD)[J]. *Chem Soc Rev*, 2004, 33(8): 463–475.
- [33] HELLENBRANDT M. The inorganic crystal structure database (ICSD)—Present and future[J]. *Crystallogr Rev*, 2004, 10(1): 17–22.
- [34] DOWNS R T, HALL-WALLACE M. The American Mineralogist crystal structure database[J]. *Am Mineral*, 2003, 88(1): 247–250.
- [35] GRAŽULIS S, DAŠKEVIČ A, MERKYS A, et al. Crystallography Open Database (COD): An open-access collection of crystal structures and platform for world-wide collaboration[J]. *Nucleic Acids Res*, 2012, 40(Database issue): D420–D427.
- [36] BLANTON T N, HUANG T C, TORAYA H, et al. JCPDS—International Centre for Diffraction Data round robin study of silver behenate. A possible low-angle X-ray diffraction calibration standard[J]. *Powder Diffr*, 1995, 10(2): 91–95.
- [37] 袁欣. SpringerMaterials: 物理、化学研究中必不可少的数值型数据库[J]. *清华大学学报(自然科学版)*, 2013, 53(9): 1360–1364.
- YUAN Xin. *J Tsinghua Univ Sci Technol*, 2013, 53(9): 1360–1364.
- [38] BARRY T I, GLASSER F P. Calculations of Portland cement clinkering reactions[J]. *Adv Cem Res*, 2000, 12(1): 19–28.
- [39] ANTONOV S, DETROIS M, ISHEIM D, et al. Comparison of thermodynamic database models and APT data for strength modeling in high Nb content γ - γ' Ni-base superalloys[J]. *Mater Des*, 2015, 86: 649–655.
- [40] BALE C W, BÉLISLE E, CHARTRAND P, et al. FactSage thermochemical software and databases—Recent developments[J]. *Calphad*, 2009, 33(2): 295–311.
- [41] GISBY J, TASKINEN P, PIHLASALO J, et al. MTDATA and the prediction of phase equilibria in oxide systems: 30Years of industrial collaboration[J]. *Metall Mater Trans B*, 2017, 48(1): 91–98.
- [42] KANG Y B, PELTON A D. Thermodynamic model and database for sulfides dissolved in molten oxide slags[J]. *Metall Mater Trans B*, 2009, 40(6): 979–994.
- [43] MANION J A, HUIE R E, LEVIN R D, et al. NIST Chemical Kinetics Database, NIST Standard Reference Database 17, Version 7.0 (Web Version), Release 1.6.8, Data version 2015.09[Z]. Gaithersburg, Maryland: National Institute of Standards and Technology, 2015.
- [44] COUGHLIN J P. Heats of formation of crystalline $\text{CaO} \cdot \text{Al}_2\text{O}_3$, $^{12}\text{CaO} \cdot 7\text{Al}_2\text{O}_3$, and $3\text{CaO} \cdot \text{Al}_2\text{O}_3$ [J]. *J Am Chem Soc*, 1956, 78(21): 5479–5482.
- [45] MYERS R J, BERNAL S A, PROVIS J L. A thermodynamic model for C-(N-) A-S-H gel: CNASH_{ss}. Derivation and validation[J]. *Cem Concr Res*, 2014, 66: 27–47.
- [46] KULIK D A, KERSTEN M. Aqueous solubility diagrams for cementitious waste stabilization systems: II, end-member stoichiometries of ideal calcium silicate hydrate solid solutions[J]. *J Am Ceram Soc*, 2001, 84(12): 3017–3026.
- [47] KULIK D, TITS J, WIELAND E. Aqueous-solid solution model of strontium uptake in CSH phases[J]. *Geochim Cosmochim Acta*, 2007, 71(S 1): A530.
- [48] LOTHENBACH B, LE SAOUT G, BEN HAHM M, et al. Hydration of a low-alkali CEM III/B– SiO_2 cement (LAC)[J]. *Cem Concr Res*, 2012, 42(2): 410–423.
- [49] KULIK D A. Improving the structural consistency of C-S-H solid solution thermodynamic models[J]. *Cem Concr Res*, 2011, 41(5): 477–495.
- [50] ZHOU Y F, LI W W, PENG Y X, et al. Hydration and fractal analysis on low-heat Portland cement pastes using thermodynamics-based methods[J]. *Fractal Fract*, 2023, 7(8): 606.
- [51] LIAO Y S, YAO J X, DENG F, et al. Hydration behavior and strength development of supersulfated cement prepared by calcined phosphogypsum and slaked lime[J]. *J Build Eng*, 2023, 80: 108075.
- [52] ZHUANG S Y, WANG Q. Inhibition mechanisms of steel slag on the early-age hydration of cement[J]. *Cem Concr Res*, 2021, 140: 106283.
- [53] YAN J H, WU F N, LI S S, et al. Mechanical and chloride ions solidification performance of $\text{C}_4\text{A}_3(\text{S}, \text{P})$ mineral as promising marine engineering material[J]. *Constr Build Mater*, 2022, 323: 126553.
- [54] KRISKOVA L, PONTIKES Y, ZHANG F, et al. Influence of mechanical and chemical activation on the hydraulic properties of gamma dicalcium silicate[J]. *Cem Concr Res*, 2014, 55: 59–68.
- [55] 王晓东. $\text{CaO}-\text{Al}_2\text{O}_3-\text{P}_2\text{O}_5$ 系胶凝材料的设计、制备、表征与性能研究[D]. 深圳: 深圳大学, 2013.
- WANG Xiaodong. Design, preparation, characterization and properties of $\text{CaO}-\text{Al}_2\text{O}_3-\text{P}_2\text{O}_5$ cementitious materials[D]. Shenzhen: Shenzhen University, 2013.
- [56] 刘垒. 基于硫硅酸钙结构与活性调控的新型低碳熟料体系制备研究[D]. 北京: 中国建筑材料科学研究总院, 2023.
- LIU Lei. Study on preparation of new low-carbon clinker system based on structure and activity regulation of calcium sulfosilicate[D]. Beijing: China Building Materials Research Institute, 2023.
- [57] 陶勇. 水泥熟料矿物晶体结构与水化活性分子模拟研究[D]. 武汉: 武汉理工大学, 2021.
- TAO Yong. Study on crystal structure and hydration activity of cement clinker by molecular simulation[D]. Wuhan: Wuhan University of Technology, 2021.
- [58] 张子良. Fe、Mg、Na、K对硫铝铁系胶凝材料熟料矿相形成及水化进程的影响研究[D]. 济南: 山东大学, 2023.
- ZHANG Ziliang. Study on the influence of Fe, Mg, Na and K on the formation of mineral phase and hydration process of clinker of

- sulfur-aluminum-iron cementitious materials[D]. Jinan: Shandong University, 2023.
- [59] 郑国法. 碱对铁铝酸盐水泥熟料矿相形成与水化的影响研究[D]. 济南: 济南大学, 2022.
- ZHENG Guofa. Effect of alkali on mineral phase formation and hydration of aluminoferrite cement clinker[D]. Jinan: University of Jinan, 2022.
- [60] BLANC P, VIEILLARD P, GAILHANOU H, et al. ThermoChimie database developments in the framework of cement/clay interactions[J]. Appl Geochem, 2015, 55: 95–107.
- [61] ABRAHAM J J, DEVERS C, TEODORIU C, et al. The need for a comprehensive cement database - A novel approach to best practices by cataloging cement properties[C]//Day 3 Wed, November 17, 2021. Abu Dhabi, UAE. SPE.
- [62] 梅细燕. 水泥窑仿真系统的建模研究及其基于web的数据库管理[D]. 武汉: 武汉理工大学, 2002.
- MEI Xiyen. Modeling research of cement kiln simulation system and its database management based on web[D]. Wuhan: Wuhan University of Technology, 2002.
- [63] 王亚峰. 基于WINCC与SQL数据库的水泥散装集中控制系统的设计与应用[J]. 中国水泥, 2019(7): 103–105.
- WANG Yafeng. China Cem, 2019(7): 103–105.
- [64] TKACHENKO N, TANG K, MCCARTEN M, et al. Global database of cement production assets and upstream suppliers[J]. Sci Data, 2023, 10(1): 696.
- [65] JAYASURIYA A, SHIBATA E S, CHEN T, et al. Development and statistical database analysis of hardened concrete properties made with recycled concrete aggregates[J]. Resour Conserv Recycl, 2021, 164: 105121.
- [66] WENDNER R, VOREL J, SMITH J, et al. Characterization of concrete failure behavior: A comprehensive experimental database for the calibration and validation of concrete models[J]. Mater Struct, 2015, 48(11): 3603–3626.
- [67] BAŽANT Z P, PANULA L. Practical prediction of time-dependent deformations of concrete[J]. Matériaux Constr, 1978, 11(5): 307–316.
- [68] BAŽANT Z P, KIM J K. Improved prediction model for time-dependent deformations of concrete: Part 3—Creep at drying[J]. Mater Struct, 1992, 25(1): 21–28.
- [69] BAŽANT Z P, KIM J K. Improved prediction model for time-dependent deformations of concrete: Part 2—Basic creep[J]. Mater Struct, 1991, 24(6): 409–421.
- [70] BAŽANT Z P, KIM J K. Improved prediction model for time-dependent deformations of concrete: Part 4—Temperature effects[J]. Mater Struct, 1992, 25(2): 84–94.
- [71] BAŽANT Z P, KIM J K, PANULA L. Improved prediction model for time-dependent deformations of concrete: Part 1—Shrinkage[J]. Mater Struct, 1991, 24(5): 327–345.
- [72] BAZANT Z P, LI G. Comprehensive database on concrete creep and shrinkage[J]. ACI Mater J, 2008, 105(6): 635–637.
- [73] HUBLER M H, WENDNER R, BAŽANT Z P. Comprehensive database for concrete creep and shrinkage: Analysis and recommendations for testing and recording[J]. ACI Mater J, 2015, 112(4): 547–558.
- [74] 方修春. 混凝土耐久性试验数据库管理系统的构建与应用研究[D]. 重庆: 重庆大学, 2004.
- FANG Xiuchun. Construction and application of database management system for concrete durability test[D]. Chongqing: Chongqing University, 2004.
- [75] 高志斌, 刘岩. 机场道面混凝土抗冻耐久性评价预测软件系统的研发[J]. 建筑科学, 2011, 27(S1):230–234.
- GAO Zhibin, LIU Yan. Buind Sci, 2011, 27(S1):230–234.
- [76] 梁程瑶. 海洋环境下混凝土强度演变的数据挖掘和数据库建立[D]. 南京: 东南大学, 2018.
- LIANG Chengyao. Data mining and database establishment of concrete strength evolution in marine environment[D]. Nanjing: Southeast University, 2018.
- [77] MA P F, ZHANG Y, LI K F, et al. Smart database design for concrete durability analysis - An application in the Hongkong-Zhuhai-Macau bridge[J]. Cem Concr Res, 2023, 163: 107033.
- [78] MCCARTHY G J, SOLEM J K, MANZ O E, et al. Use of a database of chemical, mineralogical and physical properties of North American fly ash to study the nature of fly ash and its utilization as a mineral admixture in concrete[J]. MRS Online Proc Libr, 1989, 178(1): 3–33.
- [79] DIAZ E I, ALLOUCHE E N. Recycling of fly ash into geopolymer concrete: Creation of a database[C]//2010 IEEE Green Technologies Conference. Grapevine, TX, USA. IEEE, 2010: 1–7.
- [80] NAFEEES A, AMIN M N, KHAN K, et al. Modeling of mechanical properties of silica fume-based green concrete using machine learning techniques[J]. Polymers, 2021, 14(1): 30.
- [81] PELLEGRINO C, FALESCHINI F. Experimental database of EAF slag use in concrete[M] //PELLEGRINO C, FALESCHINI F. Sustainability Improvements in the Concrete Industry: Use of Recycled Materials for Structural Concrete Production. Cham: Springer International Publishing, 2016:141-175.
- [82] CORRADI M, BORRI A. A database of the structural behavior of masonry in shear[J]. Bull Earthq Eng, 2018, 16(9): 3905–3930.
- [83] DE ASSIS GARCIA SOBRINHO R, PIAUHY NETO F, FERNANDES H. Public database of cracks images in mortar coating with different types of surface finishes[J]. Buildings, 2023, 13(7): 1872.
- [84] GILLIGAN L P J, COBELLI M, TAUFOR V, et al. A rule-free workflow for the automated generation of databases from scientific literature[J]. NPJ Comput Mater, 2023, 9: 222.
- [85] SWAIN M C, COLE J M. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature[J]. J Chem Inf Model, 2016, 56(10): 1894–1904.
- [86] COURT C J, COLE J M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction[J]. Sci Data, 2018, 5: 180111.
- [87] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. 2013: arXiv: 1301.3781. <http://arxiv.org/abs/1301.3781>
- [88] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 1532–1543.
- [89] VASWANI A. Attention is all you need[J]. arXiv preprint arXiv: 1706.03762, 2017.
- [90] DEVLIN J, CHANG M, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//BURSTEIN J, DORAN C, SOLORIO T eds. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, Association for Computational Linguistics, 2019: 4171–4186.

Databases of Cementitious Materials Including Clinker and Their Applications

XU Chengwen¹, YE Jiayuan¹, GAO Lin², GAO Guoxian³, REN Xuehong¹, XIA Lingfeng⁴, WANG Lin⁵, ZHANG Wensheng¹

(1. State Key Laboratory of Green Building Materials, China Building Materials Academy, Beijing 100024, China; 2. Anhui Key Laboratory of Green and Low-carbon Technology in Cement Manufacturing, Hefei Cement Research & Design Institute Corporation Ltd., Hefei 230051, China; 3. State Key Laboratory of Mesoscience and Engineering, Institute of Process Engineering, Chinese Academy of Sciences, Beijing 100190, China; 4. ZhongCun Big Data Technology Co., Ltd, Beijing 100024, China; 5. Shandong Provincial Key Laboratory of Ubiquitous Intelligent Computing, Jinan University, Jinan 250022, China)

Extended Abstract

In light of the policy of ‘carbon peaking and carbon neutrality’ and the advent of artificial intelligence, there is an urgent need for research and development in the field of cementitious materials to advance low-carbon alternatives. Given the lengthy research periods associated with traditional experimental methods, digital research and development has recently emerged as a dominant trend in the field. Currently, digital models are being developed for purposes such as performance prediction, composition optimization, and low-carbon innovation, with a primary focus on algorithm optimization. However, the success of these models relies heavily on the availability of accurate, high-quality databases, which serve as the foundation for model implementation. Utilizing such databases can significantly simplify model construction and reduce the need for extensive optimization procedures. Consequently, building comprehensive databases for cementitious materials research and development has become a key objective in this field.

This paper reviews the current status of databases for cementitious materials, including those related to clinker, cement, concrete, mineral admixtures, and construction mortar. The types of data, application methods, and scope of application encompassed by these databases are summarized. Databases for clinker were among the earliest to be developed, primarily cataloging diverse mineral properties, including hydration products. They include crystallographic data, thermodynamic data, and force field data, with some databases integrating models for analyzing cement hydration. Cement databases, while recording thermodynamic properties, cement types, and characteristics, have largely been focused on production control in cement plants. Moreover, the construction of databases for concrete and related materials has often been driven by specific research objectives, such as studying chemical composition, strength, and durability. Some researchers have also conducted property predictions based on their collected datasets.

Despite the existence of numerous databases on cementitious materials, a significant volume of relevant data within scientific literature remains underutilized. The advancement of artificial intelligence in natural language processing has enabled the adaptation of data extraction algorithms across various domains, including metals, medicine, and biology. Named entity recognition and textual relationship extraction—two critical components of literature data mining—can be implemented through AI algorithms such as ChemDataExtractor and BERT. ChemDataExtractor and related algorithms have demonstrated accurate chemical data extraction from compounds and semi-supervised relationship extraction. Similarly, BERT, like ChatGPT, is a state-of-the-art language model developed using the Transformer architecture and has been successfully applied in automated text data extraction. However, the complex mineralogical and chemical composition, multi-scale particle characteristics, and hydration processes of cementitious materials pose challenges to the direct application of these algorithms.

Summary and prospects The construction of comprehensive databases represents the cornerstone of the digital transformation of low-carbon cementitious material research and development. Although existing databases contain extensive datasets on cementitious materials, several challenges persist, such as non-standardized database structures, insufficiently considered data categories, and incomplete material coverage. To address these issues, future database development should prioritize unifying data formats and linking upstream and downstream processes to create a cohesive and interconnected database. Such a unified database would enable the establishment of a fully connected data chain for cementitious materials, enhancing the accuracy of predictions, supporting reverse design processes, and saving time on data cleaning.

Particular emphasis should also be placed on refining databases for related materials, such as mineral admixtures. These refined databases would provide critical data for improving the durability of cementitious materials and reducing their carbon emissions. Furthermore, combining artificial intelligence algorithms, such as ChemDataExtractor and BERT, with domain expert knowledge holds significant potential for advancing literature mining techniques tailored to cementitious materials. This process could begin by focusing on individual performance attributes and gradually expand to encompass the comprehensive extraction of data for all cementitious materials.

The future of cementitious materials database development is promising, with the potential to drive innovations in low-carbon materials research and development, ultimately contributing to achieving global carbon neutrality goals.

Keywords cement-based material; database; data mining; artificial intelligence