材料信息学及其在材料研究中的应用

王 卓[1,2]，王 礞[2]，雍歧龙[1]，郭艳华[3]，崔予文[3,4]

（1. 钢铁研究总院，北京 100081）

（2. 成都材智科技有限公司，四川 成都 610041）

（3. 南京工业大学材料科学与工程学院，江苏 南京 211899）

（4. IMDEA Materials Institute，C/Eric Kandel 2，Getafe，Madrid，Spain）

崔予文

摘 要： 2011 年美国奥巴马总统提出的材料基因组计划（MGI），旨在以比原先至少快两倍的速度开发和制造先进材料，且成本仅为原先的几分之一，这促使了材料信息学的快速发展。材料信息学是信息学技术在材料学中的应用，通过建设材料信息数据库、集成材料研究设计平台和材料数据挖掘方法对材料大数据进行分析和预测，快速发现决定材料性能的“基因”，也就是材料成分-工艺-组织-性能之间的定量关系，可以有效地加快材料研发设计。介绍了材料信息学的基本概念和主要研究领域，描述了材料信息学中的 3 个主要组成部分：材料信息数据库、集成材料设计平台和材料数据挖掘技术的主要内容和应用实例。材料信息数据库储存和管理各类材料数据，包括材料基础性能、晶体结构数据、模拟计算数据、试验与工艺数据、专利数据和各类出版物等；集成材料设计平台提供各种模拟计算方法，如第一性原理、分子动力学、CALPHAD 方法、相场模拟和有限元分析；数据挖掘是统计学、机器学习、信息学、可视化技术等学科的交叉领域，是从大数据中发现知识的实用方法。并介绍了成都材智科技搭建的“材智云”集成材料设计平台的框架和功能。思考了材料信息学在材料领域中应用时所面临的难题。

关键词： 材料信息学；材料数据库；材料集成设计平台；数据挖掘；大数据

中图分类号： TB30 **文献标识码：** A **文章编号：** 1674-3962（2017）02-0132-09

Materials Informatics and Its Application in Materials Research

WANG Zhuo[1,2], WANG Meng[2], YONG Qilong[1], GUO Yanhua[3], CUI Yuwen[3,4]

（1. China Iron & Steel Research Institue Group , Beijing 100081, China）

（2. Matclouds CO., Ltd, Chengdu 610041, China）

（3. College of Materials Science and Engineering, Nanjing Tech University, Nanjing 211899, China）

（4. IMDEA Materials Institute, C/Eric Kandel 2 Getafe Madrid, Spain）

Abstract: The Materials Genome Initiative (MGI) project proposed by President Obama in 2011 is aimed at two times faster of developing and manufacturing of advanced materials while at a fraction of cost than before. This project promoted the rapid development of materials informatics that is the application of information technology in materials science. Material information database, integrated materials design platform and data mining methods are used to analyze and predict materials big data, and further reveal the quantitative relationships of the constituents-process-microstructure-properties in materials science, which is the gene to determine the materials properties. The design and development of advanced materials can be effectively speeded up by materials informatics. This article describes the concepts and main research areas of materials informatics. The materials information databases provide the storage and management service of materials data, such as crystal structure data, simulated and predicted data, experimental and processing data, and even patent data and various kinds of publications. Integrated materials design platform provides multiscale simulation techniques of materials research, such as the first-principles calculation, molecular dynamics, CALPHAD method, phase-field simulation and finite element analysis, etc. The calculated data can be added to the materials databases. Data Mining is an interdisciplinary field merging methods from statistics, machine learning, information science, visualization and other disciplines. It is a very useful approach to discover knowledge from materials big data produced by combination of experiments and high throughput calculation. The application of materials databases, integrated materials design platform

收稿日期： 2016-09-08

基金项目： 国家自然科学基金资助项目（51571113）；江苏省前瞻性联合研究项目（SBY2016020451）

第一作者： 王 卓，男，1980 年生，博士研究生

通讯作者： 崔予文，男，1970 年生，教授，博士生导师，Email：ycui@ njtech. edu. cn

DOI：10. 7502/j. issn. 1674-3962. 2017. 02. 08

and data mining in materials research are introduced. Especially, the framework and function of "MATGENE" integrated materials design platform built by Matclouds technology in Chengdu are also described. Finally the challenges of materials informatics in materials research are discussed.

Key words: materials informatics; materials databases; integrated materials design platform; data mining; big data

1　前　言

目前开发新材料、替换材料和材料制造工艺的研究主要通过实验和模拟方法进行，工程量巨大并且十分耗时，获得的材料数量稀少并且依赖一定的经验和运气。如何提高材料研发设计的效率并缩短周期，成为材料科学工作者的首要目标。2011 年美国奥巴马总统提出的材料基因组计划(The Materials Genome Initiative，MGI)，旨在以比原先至少快两倍的速度开发和制造先进材料，且成本仅为原先的几分之一。实现这个雄伟的目标需要以下 3 个条件[1]：①从理论层面理解物理机制和决定材料性能的结构与性能关系；②多尺度、高通量模拟计算软件与高效的计算能力；③计算软件所需的数据库及之后有效的筛选方法。

实现材料基因组计划的 3 个条件，离不开信息技术的支撑。信息学的定义是使用计算机软件对信息进行收集、存储、管理、分类和检索，目前已经成功应用于生命科学和化学研究等领域。随着高通量实验和表征方法在现代材料研发中的不断应用，材料数据进入了爆发式增长阶段，从大量的数据中发现知识是未来材料研发的主要方法。因此结合了材料学研究和信息技术的材料信息学近些年来得到了快速的发展。

2　材料信息学的概念

Agrawal 等[2]将材料研究划分为 4 个阶段(如图 1 所示)：在早期很长的一段时间内是以经验科学为主；从 17 世纪开始进入理论模型产生知识的阶段，其特征为使用数学方法得到的热力学模型；计算机的发明使材料研究进入计算模拟阶段，密度泛函理论、分子动力学等一系列模拟计算方法在这个时期得到快速应用；随着计算机运算能力的提高，采用高通量计算、组合实验等方法产生了大量数据，再结合前 3 阶段的理论知识和实验数据，材料研究进入了(大)数据推动科学发现的阶段。材料信息学将在这个阶段得到快速发展，并在材料设计领域发挥极其关键的作用。

在 1999 年美国波士顿举行的"Materials Informatics–Effective Data Management for New Materials Discovery"大会上，John R Rodgers 教授首先提出材料信息学(Materials Informatics)这一概念，认为材料信息学是对材料数据的有效管理[3]。Rodgers 和 Cebon[4]认为材料信息学是采用计算方法对材料科学和工程数据进行处理和分析。

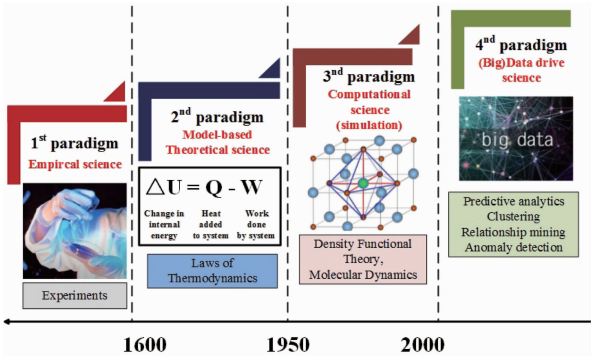

图 1　材料研究的 4 个阶段：经验、理论、计算模拟和(大)数据推动[2]

Fig. 1　The four paradigms of material research: Empirical, Theoretical, Computational, (Big) data drive[2]

Rajan[5]教授详细描述了材料信息学在材料科学与工程中的应用(如图 2)：基于数据挖掘技术的数学工具为跨尺度的集成材料科学信息提供计算引擎；信息技术提供了快速数据融合的手段，在长度和时间尺度上帮助探寻材料的结构与性能关系；材料信息学工具以联系经典材料学研究领域的信息技术为基础，将是材料研究领域内的重要工具；具备科学选择和组织数据能力的数据仓库和数据管理技术，将组成可靠的数据检索和管理系统；数据挖掘提供快速多元相关性分析；数据的科学可视化分析是评估高维信息研究的关键领域；网络基础设施可以加速信息共享、数据共享以及最重要的知识发现共享。

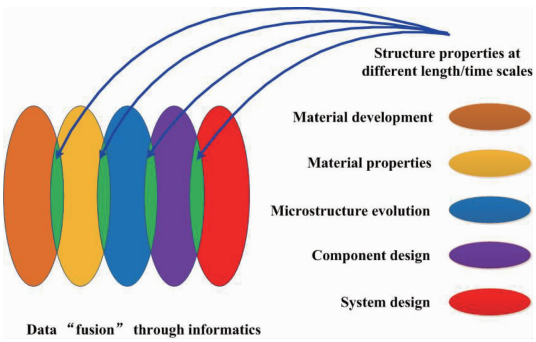

图 2　材料信息学在材料科学与工程中的应用[5]

Fig. 2　The applications of materials informatics in material science and engineering[5]

综上所述，材料信息学的含义可以归纳为材料科学与工程领域的大数据分析，通过计算机技术对海量材料数据进行数据挖掘和可视化分析，从中提取总结材料的

成分-工艺-结构-性能关系，实现知识共享，有力促进新材料新工艺的研发设计。

3 材料信息学的研究领域

材料信息学的研究领域可以划分为 3 部分：数据产生、数据管理和知识发现。现今采用组合材料科学、高通量计算等新的研究方法产生了大量结构和性能数据，需要科学的数据分析和数据挖掘方法才能揭示数据内部隐藏的知识和规律。因此材料信息学的实质是材料集成设计和材料数据库平台的搭建，以及材料领域的大数据分析。通过高通量的材料计算，获得大量的材料理论数据，结合材料的实验数据和工艺数据，构成材料的大数据集，利用数据库技术进行管理、数据挖掘方法进行分析和预测，总结新的知识，探寻决定材料结构-性能关系的“基因”，促进新材料的快速发展。目前材料信息学的主要研究领域集中在以下 4 个方面：

数据标准

目前存在大量数据形式不同的数据库，数据库之间的数据传输和信息共享十分困难。统一的数据标准是数据库之间实现数据共享的基础。因此材料信息学首要的任务是材料信息标准化的制定，以便整合这些数据库为一体。国际标准化组织(ISO)制定了一系列“产品模型数据交互规范”(Standard for the Exchange of Product Model Data，STEP，ISO10303)标准，用以描述整个产品生命周期内的产品信息，旨在实现产品数据的交换和共享。美国国家标准和技术研究院(National Institute of Standard and Technology，NIST)基于 XML 开发的 MatML，是专为材料数据信息管理和交换的可扩展标识语言。目前已经应用于 MatWeb 在线材料数据库的数据导出和下载、Granta Design 的材料数据管理软件、通用电气公司内部的数据交换等。

材料数据库

为了满足材料工作人员的不同需求，适应材料生产和研究开发，经过良好的组织和管理汇总后的材料数据库是非常必要的。按信息内容可以将材料数据库划分为材料基础性能数据库和材料信息数据库：材料基础性能数据库的数据主要包括材料的机械性能、晶体结构、热力学动力学数据和物理性能(弹性常数、热导率、磁学性能等)，为材料设计提供基础数据；材料信息数据库则利用先进的信息技术，从文献、互联网等各个渠道中提取和管理材料数据，包括材料的生产工艺数据、性能数据和服役性能等。

材料数据可视化

可视化是指将数据和信息通过一定的方法转化为大脑易于分析和理解的视觉形式(曲线、图表、数据仪表盘等)。基于材料数据的材料结构可视化信息的构建可以助力研究人员从不同视觉维度分析和解释材料性能和材料结构之间的关系。数据一旦可视化后，原先可能在本领域工作多年的专家也很难察觉的内部特征和规律，将变得非常容易预测和识别，这将极大地促进材料知识的发现和应用。

材料数据挖掘

对于工业不断提出的大量新材料需求，通过物理模拟的方法分析成分、工艺和最终性能的影响规律是耗时耗力的。数据挖掘方法以数据输入并分析预测产生模型输出，可以利用其对材料大数据分析建模发现潜在的组织性能影响规律，其典型流程如图 3 所示[6]。

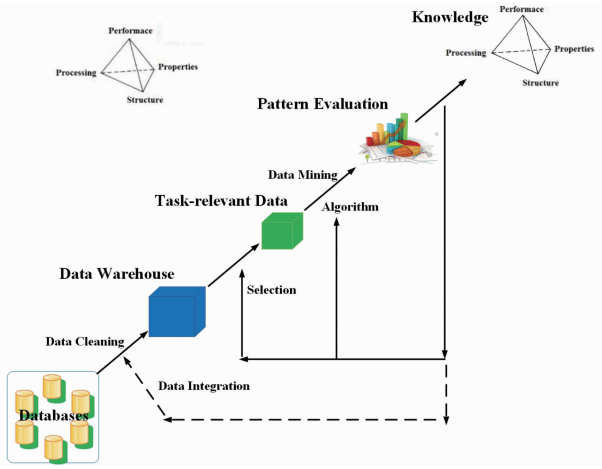

图 3 利用数据挖掘方法进行知识发现的流程[6]

Fig. 3 The knowledge discovery process by data mining[6]

4 材料信息学在材料研究中的应用

图 4 展示了材料信息学在材料科学研究中的典型应用流程[2]：通过实验、计算获得大量原始数据以不同的数据格式存储在各类材料数据库中；材料工作人员可以使用数据库里的数据进行初步的统计分析；为了建立性能预测模型，需要了解数据的格式和意义，并在建模前做必要的数据预处理以保证数据可靠性，包括删除或适当处理数据噪点、异常点、缺失值、重复数据等；完成数据预处理后采用监督式数据挖掘技术进行模型预测：以正交验证等方法使用训练数据集评估模型的精度，然后利用模型预测未知数据，发现知识和规律。除了模型预测，用户也可以根据需要使用聚类和关联挖掘。整个流程的应用对象包括实验学者、计算机和材料信息学相关专业人员。材料信息学在整个流程中的应用包括材料数据库的建立、集成数据库与模拟计算的材料研发平台和材料数据的挖掘和分析，接下来的内容将分别描述其在材料科学中的具体应用。

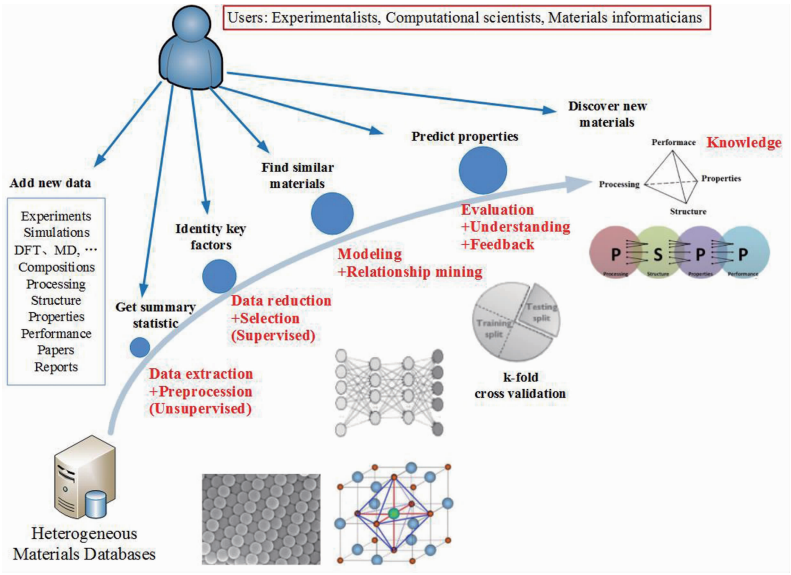

图 4　利用材料信息学进行知识发现的流程示意图[2]

Fig. 4　The knowledge discovery workflow by materials informatics[2]

4.1　材料数据库

为了有效地管理和分析由组合实验产生的大量数据，建立相应的材料数据库是十分有必要的。在材料基因组计划中，材料数据库和集成计算材料工程(ICME)以及材料实验测试是材料研究的 3 大基本工具，其重要性不言而喻。

4.1.1　材料数据库的建设情况

早期的数据库主要为离线数值型数据库，如 Granta 开发的 CMS 和 ASM 开发的“Mat. DB”。随着 Web 技术的发展，数据库类型逐渐转变为在线数据库。目前著名的在线材料数据库为美国的 MatWeb 和日本的 MatNavi。MatWeb 目前拥有超过 115000 种材料的性能数据，涵盖金属、塑料、陶瓷和化合物，数据主要源自制造商产品检验，其余来源于数据手册或专业协会。MatWeb 还具备 ANSYS、SolidWorks 等 CAD/CAM 软件的数据输出的功能。MatNavi 由日本国立材料科学研究院(NIMS)组建，拥有 9 个基础性能数据库(计算相图、计算电子结构、中子嬗变、扩散数据库等)、5 个结构材料数据库(蠕变、疲劳、腐蚀等)、4 个工程应用数据库(金属材料、CCT 曲线、材料风险信息平台)和 5 个数据应用系统，目前已经有超过 149 个国家的 11 万用户注册使用。目前中国较为系统的在线数据库为国家材料科学数据共享网，该数据库以北京科技大学为中心，汇集了全国 30 余家科研单位的数据，整合了超过 60 万条各类材料科学数据。

在材料研究工作中，晶体结构数据库起到了良好的助力作用。结合数据库的晶体结构数据，利用 Pettifor Maps 对实验数据进行分类预测，是预测晶体结构最佳经验方法之一。服务器位于德国的 FIZ Karlsruhe 的无机晶体结构数据库(Inorganic Crystal Structure Database，ICSD)拥有超过 185000 条矿物、金属和其他无机固体化合物的晶体结构数据(2032 条元素单质、34587 条二元化合物、68064 条三元化合物、66817 条四元及多元化合物)。剑桥晶体学数据中心创建的剑桥结构数据库(Cambridge Structural Database，CSD)，具有超过 80000 条数据，主要为小分子有机物和金属有机化合物晶体；皮尔森晶体结构数据库具有 274000 条数据，涵盖 157500 种相的原子坐标和占位参数，接近 17900 幅衍射花样，约 255000 幅计算相图；Pauling File 无机材料数据库中收集了从 1900 年至今超过 21000 出版物中的数据，涵盖了晶体结构、衍射、相图和物理性能，旨在创建集成数据挖掘以及其他软件的材料设计平台。

近期出现了很多以 ICSD 数据库为基础的计算材料结构和性能数据库：如 Materials Project 计划通过超级计算集群计算所有材料的性能；以 DFT 为基础的材料计算数据库 Automatic Flow(AFLOW)管理了超过 80 万中化合物的超过 7200 万条性能数据，其重心为高通量计算；由高通量密度泛函理论(HT DFT)计算所得的材料热力学和结构数据组成的 Open Quantum Materials Database(OQMD)数据库，目前已经存储了超过 28 万种各类化合物的计算数据。这些以密度泛函理论为基础的计算材料性能数据库的不断增长，体现了材料科学研究工作者对由数据驱动的材料研发的兴趣和努力。

4.1.2　材料数据库的应用

通过对材料数据库进行数据挖掘发现知识，是现代

材料研究的重要手段之一。Spark 等[7]采用数据挖掘和机器学习算法在热电数据库中分析了成千上万的化合物的热电性能，再结合了 DFT 计算，预测未知的三元相图中的低热导率相。Agrawal 等[8]使用日本国立材料科学研究院(NIMS)创建的 MatNavi 在线材料数据库建立了钢铁疲劳强度的预测模型，结果显示神经网络、决策树和多元多项式回归等先进的数据分析方法可以显著地提高预测模型的精度：其 R^2 值≥0.97。Meredig 等[9]利用现有的 DFT 数据库中的计算结果，建立了经 DFT 数据(包括结构信息)训练的正向模型预测材料性能(如形成能)。模型建立后无需输入晶体结构信息即可预测新材料的形成能，R^2 值超过 0.9。Takahashi 等[10]利用密度泛函理论中的 GPAW(Grid-Based Projector-Augmented Wave)方法建立材料数据库，预测金属间化合物的性能数据，目标材料的数据不纳入机器学习的训练数据集，预测的点阵常数和实验数据基本一致(如表 1 所示)，说明采用第一性原理计算结合机器学习预测材料的合成和设计是完全可行的。

随着信息技术的发展，新的材料信息数据库将涵盖材料基础性能数据库，并整合工艺数据、文献专利、各国标准、专业图书和行业信息统一管理，利用数据挖掘技术对材料数据库中的大量数据进行分析和预测，快速发现新的知识和规律，是未来数据驱动材料研发的主要研究领域。

表 1 利用 GPAW 计算的数据库和机器学习对其它材料性能的预测(括号内为实验数据)[10]

Table 1 Materials properties predicted by GPAW calculation and machine learning("( )" are experimental data)[10]

Materials	Crystal structure	Lattice constant	Bulk modulus
FeAl	BCC	2.5 ~ 3.0(2.91)	100 ~ 200(197)
FeNi	FCC	3.0 ~ 3.5(3.57)	100 ~ 200(151)
FeTi	BCC	3.0 ~ 3.5(2.98)	100 ~ 200
CuZn	BCC	3.0 ~ 3.5(2.95)	100 ~ 200(110)
CoTi	BCC	3.0 ~ 3.5(3.00)	100 ~ 200
MgAg	BCC	3.0 ~ 3.5(3.31)	0 ~ 100(86)
ScAl	BCC	3.0 ~ 3.5(3.39)	0 ~ 100(69)

4.2 材料集成设计平台

材料集成设计平台是以 MGI 为指导，集成材料数据库、高通量材料计算、材料测试与表征，材料数据管理和分析系统为一体的现代材料研发设计平台。Liu 等[11]提出了材料计算和模拟的集成多尺度方法的框架图（如图 5)，集成了第一性原理、CALPHAD 相图计算、相场模拟和有限元分析 4 种主要的材料结构和性能的模拟计算方法：通过原子尺度的第一性原理计算预测热力学性质、晶格常数、以及单元、二元和三元化合物和固溶体的动力学数据；CALPHAD 方法建立热力学性质、晶格常数、

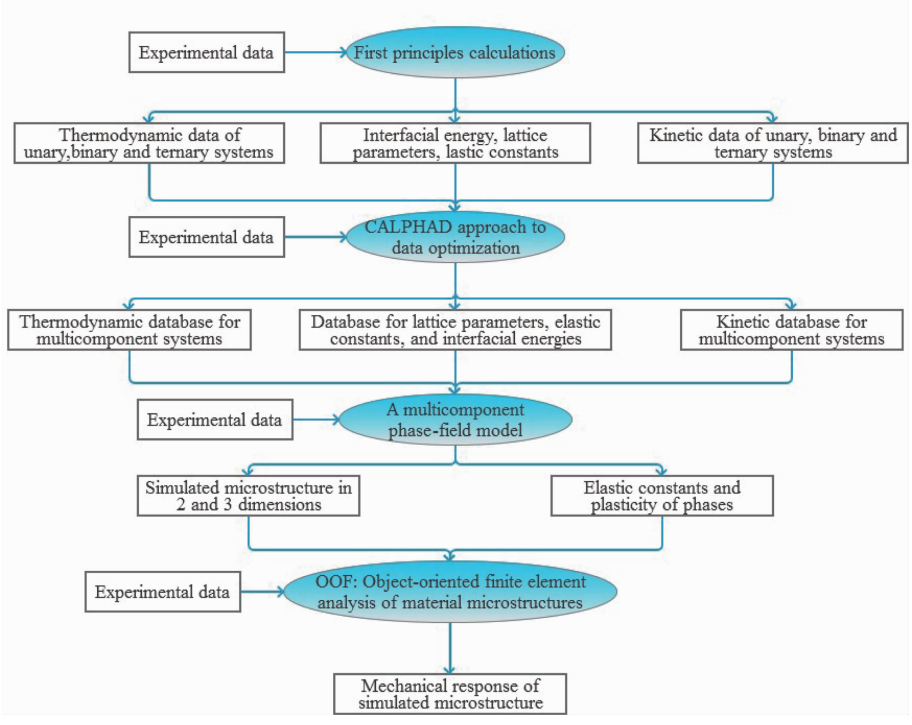

图 5 材料计算和模拟的集成多尺度方法[11]

Fig. 5 An integrated multiscale approach for materials modeling and simulation[11]

多元体系动力学数据模型；利用多元相场方法在二维和三维尺度上预测微观组织的演变规律；采用有限元分析方法从模拟组织中计算材料的机械性能。通过进行从量子力学到材料服役的跨尺度高通量的材料计算，获得大规模、多源异构的材料数据，利用信息学方法进行材料大数据分析，发现材料成分-工艺-组织-性能-服役之间的定量关系（决定材料性能的“基因”），将大大加快新材料的研发进度，摒弃传统“试错法”（或炒菜法）的材料设计方法，有效地缩短了材料的设计研发周期。

目前在建的材料集成设计平台有美国的 Automatic Flow（AFLOW）和中国科学院计算机网络信息中心组建的 Matcloud。AFLOW[12]是美国基于 VASP 建立的高通量结构能量计算平台，并集成了超过 15 万行 C++代码的一系列软件工具，其主要特征是完全并行式和多线程。AFLOW 实现了以特定数据集或大的结构数据库为对象自动计算一系列可观测量，同时只需很少的人力进行数据输入、计算运行和输出数据整理。对于不需要高通量计算和建立数据库的用户，AFLOW 还提供了结构分析和处理工具。Matcloud[13]是基于材料基因组计划中的材料集成设计理念开发的设计平台，目前支持 CASTEP 软件，已经初步实现了与中国科学院超级计算环境的集成、晶体结构计算模型的在线建立、高通量计算作业的在线提交和监控、计算与数据自动传输等。此外，源于剑桥大学的 Granta Design 公司开发的 Granta MI 实现了企业材料数据实时并可溯源的存储、检索、应用、可视化分析；MI：Gateway 确保了本地数据库和 CAD/CAM 设计软件之间的信息高速无误的传输。

中国的材智科技率先开展集成多尺度材料计算和材料信息数据库的材料设计平台研究与开发工作。旗下的“材智云”产品作为材料基因组计划的技术支撑，拟整合材料信息数据库、多尺度材料模拟云计算平台、材料测试平台和第三方数据交易综合服务平台，旨在搭建材料行业公共知识库和专业技术服务平台（如图 6）：其中材料信息数据库整合了材料性能数据库（全球 63 个标准体系、25 万个金属牌号、超过 1000 万条性能数据）、材料基础数据库（30000 条晶体结构数据、3000 幅相图、5000 幅微观组织）、各国专利（约 2000 万项），还有科技报告、行业资讯等海量数据；各类模拟计算软件（第一性原理、分子动力学、CALPHAD 方法、相场模拟等）支持用户输入数据或导入材料基础数据，实现跨尺度材料模拟计算，快速获得各类计算数据。“材智云”拟为用户提供快速准确的材料数据检索、模拟计算和材料大数据分析等数据推动研发的一站式服务（如图 7），可准确指导用户选材和产品市场分析，显著加快用户材料研发进程并有效降低成本。

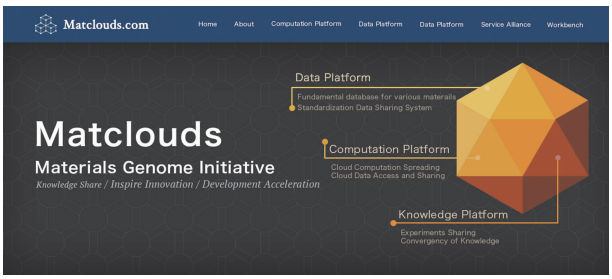

图 6　集成材料设计平台—材智云

Fig. 6　Matclouds, the integrated materials design platform

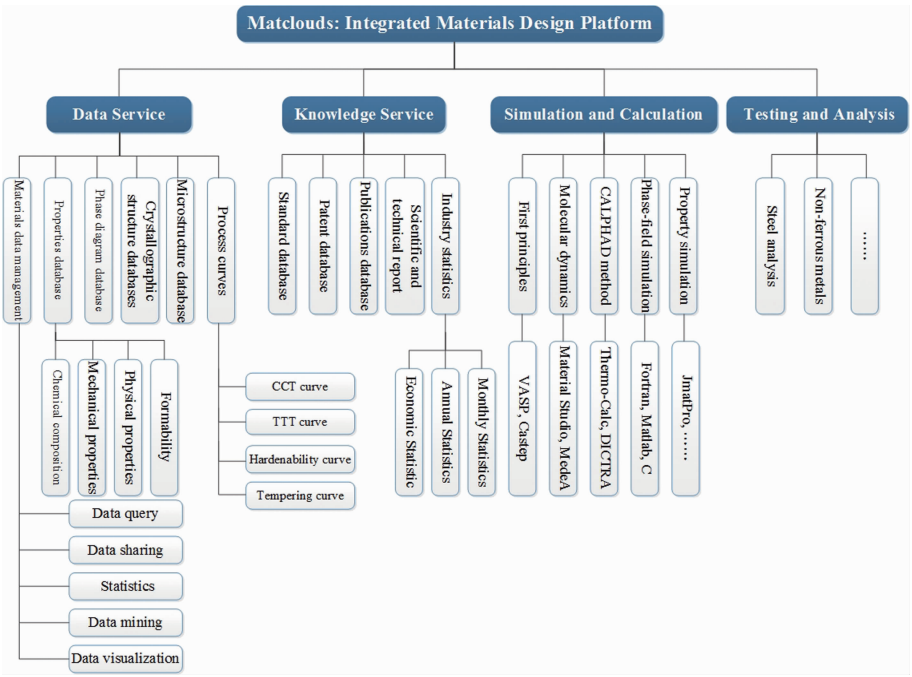

图 7　材智云结构功能示意图

Fig. 7　The framework and function of Matclouds

4.3 材料的数据挖掘

数据挖掘（Knowledge-Discovery in Databases，KDD）是使用特定的算法对大数据集进行搜索，提取数据库中的知识的过程。该过程主要包括数据输入、数据预处理（数据汇合、数据清洗、特征选择等）、数据挖掘和后处理（模式过滤、可视化等），最终得到有用的信息（知识）。服务于材料科学研究的数据挖掘主要是建立在对材料性能和服役的理解基础之上的模式识别和模式预测。模式识别是从分散的数据中发现相关性、趋势、簇类、轨迹和异常现象的基础，模式预测的本质则是对材料物理与化学的理解。在很多情况下数据挖掘和在工程材料研究中的以结构-性能关系为中心类似。

4.3.1 材料研究中的数据挖掘技术

传统的数据挖掘技术主要有线性和非线性分析、回归分析、因素分析和聚类分析，随着数据挖掘技术的飞速发展，决策树理论（Decision Trees）、人工神经网络（Artificial Neural Network，ANN）等新的技术不断应用于材料研究中。决策树是通过概率论的直观运用建立的树形结构，其中每个内部节点代表一个属性上的测试，每个分支代表一个测试输出，每个叶节点代表一种类别。决策树是分类模型的非参数方法，不需要昂贵的计算，非常容易理解。常用的决策树算法有ID3、C4.5、CART等。Georgilakis等[14]利用决策树方法为能源电力变压器的缠绕材料选材，准确度达到94%并且十分迅速。

人工神经网络（ANN）是模拟生物神经系统，由一组相互连接的节点和有向链组成网络。每个节点代表一种特定的输出函数，也就是激励函数（Activation Function），每两个节点间的连接都代表一个对于通过该连接信号的加权值，称之为权重。ANN的特点为[15]：可以用来近似任何目标函数，但需要选择合适的拓扑防止模型的过拟合；可以处理冗余特征，冗余权值非常小；对训练数集的噪声非常敏感；当隐藏节点数量巨大时，ANN的训练相当耗时，但测试分类非常快。Liu等[16,17]通过人工神经网络方法成功预测了热轧C-Mn钢的机械性能，以及常规热轧和TMCP工艺下C-Mn钢和HSLA钢的组织演变。Wu等[18]对C-Mn钢的工业生产大数据进行数据清洗后，采用贝叶斯正则神经网络建立了性能预测模型，屈服强度和抗拉强度的预测准确度分别达到96.64%和99.16%，预测值和测试值的绝对误差在±30 MPa范围内，85.71%的样本延伸率预测值和测试值之间的绝对误差不超过±4%。

遗传算法（Genetic Algorithms）借鉴自然选择和生物进化规律，是一种通过模拟“适者生存”和遗传学生物进化过程以搜索最优解的方法。它是计算机科学在人工智能领域中用于寻找最优化的一种搜索启发式算法，属于一种进化算法。遗传算法和传统搜索算法的不同点在于：①遗传算法搜寻全局最优多峰函数的群体解，而非单个解；②遗传算法可以处理无导数信息的非连续目标函数；③遗传算法处理参数集的编码而非参数本身；④遗传算法使用诸如选择、交叉和变异概率型算子，而不是那些确定型算子。遗传算法常用于确定满足所需性能的化合物和内部结构，以及确定化合物结构设计中的堆垛顺序[19]。

当系统中存在多种描述符描述的各种变量时，采用统计方法对每一个描述符进行计算是非常昂贵费时且无效率的，可以采用主成分分析法（Principal Component Analysis，PCA）解决这个问题[20]。PCA采用因素分析和主坐标分析等技术，将具有高维属性的复杂数据集投影至易于可视化的低维空间，使数据集中的描述符大幅减少，从而使数据易于可视化、分类和预测。PCA的运用须建立在相关数据库的基础上，例如已知化合物的计算能量或理论化合物的晶体结构。在常规多元法受限的情况，例如观测值少于预测变量时，可以使用PLS（偏最小二乘法）回归。PLS回归可以用于选择合适的预测变量和在经典线性回归前识别异常点。

4.3.2 数据挖掘方法在材料中的应用

数据挖掘方法很适合应用于晶体结构研究，因为晶体结构数据是离散非连续的，因此非常适合采用数据挖掘方法进行分析和预测。传统的Pettifor Maps方法广泛地应用于预测晶体结构，但也存在一定的局限：一次只能应用于一种化合物，对数据很少的晶体结构预测十分困难。为了克服传统方法的不足，研究人员使用数据挖掘或机器学习技术分析计算和实验获得的数据并预测未知的晶体结构。Morgan等[21]提出了一种结合数据挖掘的Pettifor Maps方法：采用数据算法将晶体学数据库中的数据变换为Pettifor Maps，然后使用Pettifor Maps对未知的晶体结构进行预测，通过交叉验证方法发现，Pettifor Maps预测AB和A_3B型化合物时生成的5种备选结构的准确率为86%，无未知结构的情况下准确率达到95%。Ceder[22]的团队采用数据挖掘建立了一个具有114维的结构形成能空间，然后使用PCA方法分析具有不同结构的不同材料的ab型从头能量之间的关系，设计了一个根据已有信息预测未知晶体结构的贝叶斯算法。通过以上的方法，他们能够使用数据挖掘技术从第一性原理计算中获得的大量化合物的可能结果中筛选最可能的晶体结构。

数据挖掘技术同样能够快速可靠的预测材料的组织、性能和服役行为。Liu等[23]以机器学习方法优化Fe-Ga

合金的组织、提高其机械性能和磁致伸缩效应为例对这个问题作出了回答。他们开发了由随机数据生成、特征选择和分类算法组成的系统的框架，同时满足线性和非线性属性约束的 5 个设计问题的实验表明，比起传统优化方法，计算框架的平均耗时下降了 80%，并且所得结果优于其他任何方法。另一方面，ANN 分类和回归树（CART）在处理分类数据和数据缺失方面更有优势。Sinha 等[24]采用多目标遗传算法设计 Ti-Ni 合金的工艺来优化机械性能和形状恢复行为，成功的在形状回复率和硬度及 *H/E* 比率之间建立了均衡关系；设计了以 ANN 技术为基础的数据模型解释加工条件和性能之间的经验关系，揭示了可恢复应变最大化情况下的不同工艺参数的作用。

5　材料信息学应用于材料研究中的问题

数据挖掘和信息技术方法给材料研究设计带来了新的机遇，随着可用的材料数据的规模不断增加，将会孕育不经传统实验分析而从数据中归纳科学原理和设计规则的技术。目前阻碍材料信息学进一步应用的因素主要如下。

（1）使用大数据资源时的问题的经验积累。正如大量数据库和可用的数据不断产生，但能够从处理大数据资源并提取出有用信息的用户仍然较少。而且，当无法从某个数据库中获得所需数据时，向其他数据库请求数据和从不同数据库中整合信息也很困难。计算数据和实验数据吻合性也是个难点，因为实验进行时所引用的晶体结构数据或其他数据已经无迹可寻。而使用计算数据也相当棘手：研究人员必须充分理解分析方法的误差，在某些情况下误差可能相当大并且即使相当有经验的专家可能也无法准确估计。

（2）为晶体学等建立材料描述符。在过去几年这方面取得了一定的成果，但目前仍没有关于描述晶体的描述符的算法。这类描述符包括材料性质、限定条件、量化的结构评价等。目前研究人员通过构图法向机器学习算法描述晶体结构是非常困难的。

（3）对机器模型适当性和转移性的评估。这些评估以性能导向为指标，如交叉验证。但是，产生具有误导性的性能指标的原因是多方面的。交叉验证错误会受到交叉验证类型、设计模型的选择和数据如何分解并拟合的影响。掌握机器学习模型的精确度也是非常重要的，因为具有最小交叉验证误差的模型同时也最复杂（如神经网络和随机森林），并且无法做出科学的预测。当与传统的、可解释的模型和方法中提取的知识冲突时，材料学家是否应该相信由费解的机器学习模型做出的预测，还需进一步实践。

6　结　语

材料信息学，其核心内容为材料的大数据分析，是采用先进的数据挖掘方法从各种材料信息数据库中提取知识和预测规律的研究方法。建立涵盖材料基础性能数据库、产品生产工艺数据、文献专利、各国标准、科技报告和行业信息统一管理的材料信息数据库，以及集成从原子到宏观的跨尺度高通量材料模拟计算软件和材料数据挖掘工具搭建的集成材料设计平台，将是未来材料研发极其重要的工具。采用先进的数据挖掘方法对材料信息数据库中的大数据进行分析和预测，帮助快速发现材料成分-工艺-组织-性能-服役之间的定量关系，也就是决定材料性能的“基因”，摒弃传统的“试错法”（或炒菜法）的材料设计方法，将极大地加快新材料的研发进度，达到缩短材料开发周期、降低材料研发成本的最终目的。

参考文献　References

[1] Seshadri R, Sparks T D. *Apl Materials* [J], 2016, 4 (5): 25.

[2] Agrawal A, Choudhary A. *Apl Materials* [J], 2016, 4 (5): 1-17.

[3] John R Rodgers. *Materials Informatics-Effective Data Management for New Materials Discovery* [M]. Boston: Knowledge Press, 1999.

[4] Rodgers J R, Cebon D. *Mrs Bulletin* [J], 2006, 31: 975-980

[5] Rajan K. *Informatics for Materials Science & Engineering* [J], 2013, 15 (4): 1-16.

[6] Doreswamy, Hemanth K S. *International Journal of Database Management Systems* [J], 2012, 3 (1): 512-522.

[7] Sparks T D, Gaultois M W, Oliynyk A, *et al*. *Scripta Materialia* [J], 2016, 111: 10-15.

[8] Agrawal A, Deshpande P D, Cecen A, *et al*. *Integrating Materials & Manufacturing Innovation* [J], 2014, 3 (1): 1-19.

[9] Meredig B, Agrawal A, Kirklin S, *et al*. *Physical Review B* [J], 2014, 89 (9): 82-84.

[10] Takahashi K, Tanaka Y. *Computational Materials Science* [J], 2016, 112: 364-367.

[11] Liu Z K, Chen L Q, Raghavan P, *et al*. *Journal of Computer-Aided Materials Design* [J], 2004, 11 (2-3): 183-199.

[12] Curtarolo S, Setyawan W, Hart G L W, *et al*. *Computational Materials Science* [J], 2013, 58: 218-226.

[13] Wang Zhuo (王 卓), Yang Xiaoyu (杨小渝), Zheng Yufei (郑宇飞), *et al*. *Chinese Science Bulletin* (科学通报) [J], 2013 (35): 3733-3742.

[14] Georgilakis P S, Gioulekas A T, Souflaris A T. *Journal of Materials Processing Technology* [J], 2007, 181 (1): 281-285.

[15] Tan Pang-Ning. *Introduction to Data Mining* (数据挖掘导论: 完整版) [M]. Translated by Fan Ming and Fan Hongjian (范 明, 范

宏建译). Beijing: People Post Press, 2011: 150-156.

[16] Liu Z Y, Wang W D, Gao W. *Journal of Materials Processing Technology* [J], 1996, 57 (3-4): 332-336.

[17] Tan W, Liu Z Y, Di W U, *et al*. *Journal of Iron & Steel Research International* [J], 2009, 16 (2): 80-83.

[18] Wu S W, Zhou X G, Cao G M, *et al*. *Iron and Steel* [J], 2016, 51 (5): 88-94.

[19] Kulkarni A J, Krishnamurthy K, Deshmukh S P, *et al*. *Materials Science & Engineering A* [J], 2004, 372 (1-2): 213-220.

[20] Krishna Rajan. *Materials Today* [J], 2005, 8 (10): 38-45.

[21] Morgan D, Rodgers J, Ceder G. *Journal of Physics Condensed Matter* [J], 2003, 15 (25): 4361-4369.

[22] Fischer C C, Tibbetts K J, Morgan D, *et al*. *Nature Materials* [J], 2006, 5 (8): 641-6.

[23] Liu R, Kumar A, Chen Z, *et al*. *Scientific Reports* [J], 2014, 5.

[24] Sinha A, Chattopadhyay P P, Datta S. *Materials & Design* [J], 2012, 46: 227-234.

(本文为本刊约稿，编辑 盖少飞)