特约专栏

材料数据在材料创新发展中的作用与存在问题的思考

尹海清[1]，姜 雪[1]，张瑞杰[1]，刘国权[1]，郑清军[2]，曲选辉[1,3]

（1. 北京科技大学 钢铁共性技术协同创新中心，北京 100083）

（2. 美国肯纳金属有限公司，宾夕法尼亚州 15650）

（3. 北京科技大学新材料技术研究院，北京 100083）

摘 要：材料数据是材料基因组计划的三大核心工具之一，近年来在国际上引起强烈关注，美国、日本等国相继资助了大型数据库建设和数据分析的项目。材料数据的准确性与完整性，是数据分析与挖掘的根本保障，并直接影响材料数据库的建设以及材料数据价值的深度开发和应用。材料大数据的特征主要表现在材料属性的高维以及属性间的复杂关联关系，在材料数据分析挖掘中应重视与材料领域知识的充分结合，以及离群点分析上的学科特点及需求特殊性。而材料数据相关的基础教育，尤其是在本科阶段数学与计算机相关基础课程的设置，则成为今后材料数据成为材料创新发展手段的保障。本文就材料基因工程框架下材料数据长久发展进程中目前亟待重视和解决的问题加以讨论。

关键词：材料数据；材料基因工程；数据质量；数据挖掘；基础教育

中图分类号：N37 **文献标识码**：A **文章编号**：1674-3962(2017)06-0401-05

Role of Materials Data in Materials Innovation Development and Thoughts on the Existing Problems

YIN Haiqing[1], JIANG Xue[1], ZHANG Ruijie[1], LIU Guoquan[1], ZHENG Qingjun[2], QU Xuanhui[1,3]

(1. Collaborative Innovation Center of Steel Technology, University of Science and Technology Beijing, Beijing 100083, China)

(2. Kennametal Inc., Pennsylvania 15650, USA)

(3. Institute for Advanced Materials and Technology, University of Science and Technology Beijing, Beijing 100083, China)

Abstract: The materials data is one of the three key tools in materials genome initiative (MGI), which has been attracting great attention worldwide. Projects on large scale databases construction and data mining have been implemented in US, Japan and other countries. The accuracy and integrity of the materials data are the foundation of data analysis and mining and they will directly influence the quality of database construction and deep extraction of the data value. The main features of materials data are high dimensions of materials attributes and complex interactive relationships. It's worth noting that the data mining should be associated with domain knowledge of materials and the typical requirement of materials on the outlier analysis. Education on materials data and related disciplines, especially the college education on math and IT technology, will be the basic guarantee for the data being as the paradigm of innovation. The problems to be settled concerning the long term development of materials data were discussed in this paper.

Key words: materials data; materials genome initiative; data quality; data mining; college education

1 前 言

数据在当今时代发展中的作用是不容置疑的，由于计算机和互联网技术的飞速发展，数字化的信息以及数据传输已经成为社会发展的基础，而数据分析已成为国家安全、经济发展和风险分析等的重要手段，曾有人预言，支撑数据传输的电力系统如果出现全球性的停电，

收稿日期：2016-12-08

基金项目：科技部“863”计划项目(2015AA034201)；国家重点研发计划项目(2016YFB0700503)；北京市科技计划项目(D16110300240000)

第一作者：尹海清，女，1971 年生，教授，博士生导师，Email：hqyin@ ustb. edu. cn

DOI：10.7502/j.issn.1674-3962.2017.06.01

将对人类造成毁灭性打击。

科学技术的发展决定了一个国家发展的加速度。科学数据包括人文与社会科学数据和自然科学数据两大不同类别，后者以其专业性强、理论知识抽象复杂等特点，成为很小众的学科，受众群体相对少数且集中。国家科技部科技基础条件平台中心自 21 世纪初，支持了一批科学数据平台建设项目，建设了地理、天文、生物、遥感、医学、材料等领域的数据库，其中材料数据库包括两大数据库，即中国腐蚀与防护网和材料科学数据共享网。这些数据库成为各领域发展与应用的重大基础资源。2009 年发表在科学杂志上的《科学发现的第四范式》[1]中提出，数据科学是继理论分析、计算模拟和实验以外的第四种科学发现的范式，将数据正式定义为科学发现的新模式。

美国于 2011 年 6 月由总统发布了“先进制造伙伴计划”，该框架中最为重要的一部分是“材料基因组计划”(Materials Genome Initiative，MGI)，受到全球的关注。材料基因组计划，在中国又被称作材料基因工程，是以计算模拟、实验表征与数据作为三大工具(如图 1 所示)，基于材料理论，推动材料研发从试错法向以计算为牵引的创新设计的转型，以期达到加速材料研发进程、降低研发成本的目标。材料基因组计划自发布后获得了国内外材料领域专家学者们的高度关注，美国资助了 Materials Project 等第一性原理计算相关的以电池材料为代表的功能材料数据库[2]，日本批复了材料信息学的国家级项目“Material Research on information integration” Initiative (Mi^2i)[3]，我国在过去的五年时间里展开了多次高层次的讨论，并推动形成了“十三五”期间科技部等部门的国家重点研发项目的大力度支持。

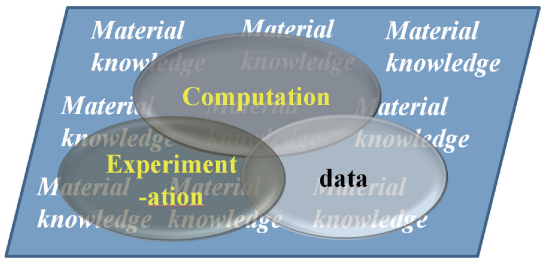

图 1 材料基因工程可以理解为构建在材料知识基础上的计算、实验和数据三大工具的创新融合

Fig. 1 Materials Genome Initiative is the innovative combination of computation, experimentation and data, based on the materials knowledge

目前，材料基因工程被业界大多数学者认识是一种新的方法论。由于材料计算与实验表征，早已经成为材料研究的两大基本手段，因此，对于材料数据的研究，被认为是材料基因工程研究的最具可能的亮点。然而，材料数据，就数量而言，尚未达到生物、地理、高能物理等领域的大数据的数量规模，但材料种类众多，影响因素错综复杂，数据关系尚待明确和梳理，本文拟对材料数据的发展及存在的潜在问题进行较为深入的讨论，以期进一步明晰材料基因组计划在材料研发创新思维的实施上的行动方案。

根据材料数据的来源，可分为计算数据、实验数据和生产数据等 3 类，数据经过收集整合并存储于数据库、数据仓库及云存储，并进一步用于数据的应用服务和深加工，其逻辑关系如图 2 所示。本文重点讨论材料数据及其在应用中的关键点或易被忽略之处。

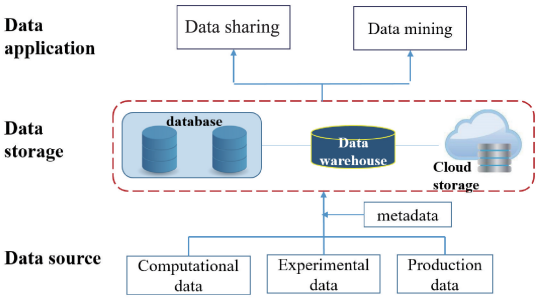

图 2 材料数据及其应用的逻辑关系

Fig. 2 Correlation between materials data and application

2 材料数据的质量

2.1 数据的准确性

数据的质量是决定数据库及其应用的根本要素。

目前，大部分材料数据是从公开发表的文章、手册等收集得来，包括已商业化的无机材料晶体结构数据库(Inorganic Crystal Structure Database，ICSD)与 Pauling file 等。尽管正在开展的高通量计算与高通量制备表征的研究在今后会产生自动化流程数据，在今后相当长的一段时间里，收集数据仍是数据来源主流。这就对数据收集者和数据库的管理提出了很高的要求。

数据的数值超过可能的取值范围或不合理的数值等明显错误，通过数据库建设时对数据的规范性约束，在存储等环节可以被计算机自动识别发现。然而由于数据录入人员的疏忽等原因造成的数值的非明显错误，对于存有大量数据的数据集，管理者和使用者是难以发现的，至今国际上尚无明确的方法或技术能够对材料数据库中大量数据的准确性进行逐一把关或验证，而此类数据的存在，对今后的数据分析与挖掘的准确性的影响不容忽视。

因此，数据规范的建设，对于不同材料的数据的整合是关键而有效的，同时，数据生产者和收集人的知识水平和工作态度是数据库质量的保证。今后，数据采集

的自动化操作，可能成为解决问题的手段之一，但由于目前实验数据的完整采集，生产环节数据记录的人工介入，以及计算的跨尺度需要等现状或问题，可知数据采集的自动化过程尚需时日。

2.2　数据的完整性

数据的完整性指的是一条数据包含的信息的完整性。

材料基因工程的新材料设计、现有材料性能提升以及新工艺的优化，对材料信息完整性的需求是不同的。如以选材和材料替代为目的的数据需求，材料的成分与性能数据是核心，数据来源的可靠性可以成为评价数据质量的有效标准。而以发现新材料为目的的数据需求，则对材料数据的内容的完整性提出了更高要求，仅仅有成分与性能数据是远远不够的，需包括计算的边界条件和初始条件、模型、算法等，实验工艺及其详细参数，表征方法及设备的基本参数指标等。我们基于国家材料科学数据共享网的建设经验与教训，制定了《材料数据提交规范》(草案)[4]，对计算数据、实验数据和生产数据所应包括的内容提出了通用格式规范。

数据的完整性与数据的准确性相辅相成，信息缺失的不完整数据在数据清洗中将被过滤掉。一条数据，如果出现信息缺失，那么该条数据的质量是不够好的，如果关键内容缺失，那么该数据的质量将被视为不合格的。只有信息完整，对数据准确性的评价，以及重复性验证才有可能，正如在众多领域的实验研究中形成的共识，即实验结果如果不能被重复出来，往往结果被质疑，甚至被认为是错误或无效的。

2.3　数据的数量与科学覆盖面

大数据的概念在当今时代已是耳熟能详的术语了。对于材料数据，其产生途径难以形成测绘卫星或正负离子对撞机产生数据的规模，在数据量上难以同高能物理等领域的数据量相提并论，但材料数据间关联关系的复杂性是材料数据能够被称之为大数据的核心，同时 MGI 强调的高通量计算与高通量实验的发展与应用，将成为材料数据量快速增长的途径之一。梅宏院士曾指出，真正的大数据应该体现在多源数据的融合，绝不仅仅是数据的“海量”[5]。数据融合与数据仓库(Data Warehouse)、数据一体化(Data Integration)不同。它的目的不是将一个企业(Enterprise)或组织的所有数据集中在一起并标准化而产生唯一的真相(Single Truth)。它是以产生决策智能为目标将多种数据源中的相关数据提取、融合、梳理整合成一个分析数据集[6]。

除了数据量，材料数据的覆盖面及其科学性和系统性是影响材料数据分析处理质量但常常为人们忽略的因素。因为如果数据大量集中在某些方面，则会造成盲人摸象的现象，导致分析结果的偏差和应用上的误导。而数据的科学性和系统性，往往是领域专家才能给出的正确定义和范畴，单纯的数据专家是难以胜任的，因此，在材料领域，同其他自然科学领域一样，领域专家与数据专家的紧密合作，是促进数据成为科学发现第四范式的基础要素。

综上，材料数据的质量是数据应用的基础与根本保证，直接影响数据共享、知识应用及价值提取，其关系如图 3 所示。

3　数据分析挖掘的质量

数据分析与挖掘的质量受材料数据质量的影响，并直接影响数据的应用(图 3)。例如基于计算数据的分析，众所周知，计算往往是对真实环境进行了简化或特定理想条件下获得的，如果分析方法或模型选取不当，数据分析时势必造成误差的累积，导致数据分析的结果难以令人信服。因此，同计算结果需要实验验证的作用

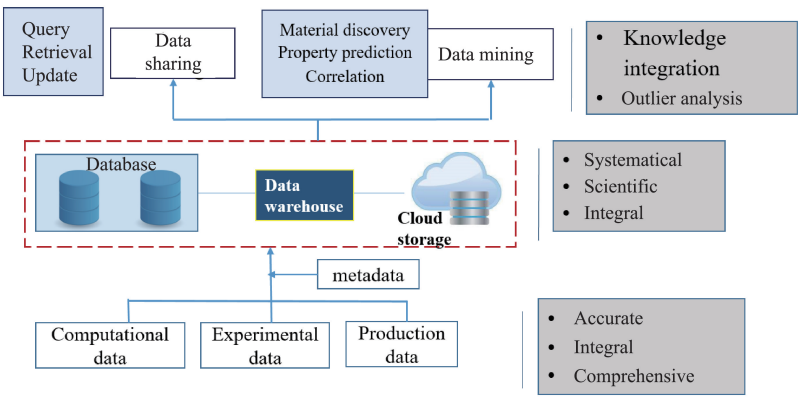

图 3　数据质量与数据存储及数据挖掘质量的关系

Fig. 3　Correlation among the data quality, data storage and quality of material data mining

相同，材料数据分析挖掘结果的实验验证，是今后数据分析人员的工作重点之一。

3.1 数据分析挖掘与材料知识的融合

当数据分析与挖掘方法应用于材料科学领域时，只有基于材料科学的基础知识、自身特点以及发展规律的数据分析与挖掘，才有可能得出有价值的结果。材料数据的分析早在 20 世纪就已经应用在不同材料研发上了，但数据量较小，多数来源于实验室自产数据。

在大数据时代，材料数据量的绝对数值相对较小，但数据间的复杂关系的融入使得在分析和处理过程中需要更多的专业人员的介入，并将相应的关系在分析模型和算法中体现出来[6]。在 MGI 高通量计算上，Gerbrand Cedar 教授[7,8]带领团队构建了 Material project 数据库，用于电池材料设计，并取得了显著成果，已发现几种性能优异的成分。Stefano Curtarolo 等[9,10]构建了 Aflow 数据库，并以特征值(descriptor)作为筛选的依据，目前，研究中所用的特征值基本是单一参数，而基于多参数组合的特征值的数据分析将是今后研究的方向之一。跨尺度计算是 MGI 架构下材料计算的研究方向之一，在材料制备过程中，单一工序的参数优化难以获得系统最终性能的最优，而数据分析与挖掘，是目前研究者正在尝试的实现跨尺度计算和实验过程的系统优化的手段与技术。Agrawal A 等[11]基于日本国家材料研究所(NIMS)的数据研究金属材料的高温疲劳性能，用成分和工艺的数据，采用多项技术来预测钢的疲劳性能，发现采用神经网络、决策树以及多元多项式回归等技术可以得到较为理想的预测精度。Singh S 等[12]利用人工神经网络与贝叶斯算法等方法，实现了对钢的工艺过程的参数优化和成分对最终性能的作用规律的揭示。Jae Hoon Jeong 等[13]采用降维和线性回归技术确定了材料成分、中间阶段性能和最终性能间的相关关系。

然而对数据含义的理解不足，或数据集选取不当，可能会导致不符合材料科学规律的结果出现。例如 Agrawal A 等[11]对不同影响因素的重要性分析时得到的一个结论是回火温度的重要性高于固溶处理温度，显然这与材料知识相悖，分析其原因在于，回火温度较固溶温度的波动大，而作者选择了几种不同材料的数据，回火处理可能是低温、中温或者高温回火，回火温度相差可达几百度。

因此，作为材料数据的分析挖掘的第一步，依据材料基本知识对数据集进行初步认识和预处理，是保证分析质量的主要步骤。

3.2 离群点的分析

离群点是指在数值上远离数值的一般水平的极端大值和极端小值，也称为歧异值，由于离群值跳跃度比较大，会直接影响分析模型的拟合精度，因此常被认为是坏的数据而在数据清洗中丢弃。然而，材料科学与工程的研究与应用发展到今天的水平，从数据中寻找主流已经难于满足向国际一流水平前进的需求了，而在一些关键点上发现问题并形成突破往往是当前的思路，如对最低值的分析，可以发现问题存在和影响规律、作用机理等。Paul Raccuglia 等[14]从失败的实验数据中发现了规律，就充分证实了这一点。在材料学科中，关键点往往存在于一些离群点上，在微观组织的图像分析上尤为明显，即在相界、晶界和界面等处，在数据分析中为简化起见，如果直接用一条简单曲线代替，使原有界面上的信息都丢失了[15]。可见，离群点分析，在材料大数据的分析中显得尤为重要，可能成为服务于材料优化设计的有效手段。

4 MGI 的基础教育

MGI 作为 IT 和互联网技术发展下的新的材料研究方法论，具有典型交叉学科的发展特点。MGI 对材料计算与实验表征的融合提出了更高的要求，而两者在新材料设计开发的时效性的需求下形成的高通量技术，以及钢铁等材料生产中设备的数据自动采集功能，则将数据科学引入材料研究和生产中，引导材料研究人员以一种全新的方法来开展研究。可见，MGI 要求计算、实验与数据三者融合，对材料工作者的能力要求显著提高，沿用原有的教育教学方案已经不能满足 MGI 实施的需求。因此应从本科生教学入手，开展相关专业课程的建设，尤其是数据及其分析处理的知识。因为相比材料计算和实验，数据分析与挖掘是一个全新的课程，尤其对于材料制备加工，将材料成分、组织与复杂工艺相结合，研究变得很复杂，仅凭对单一参数的优化，无法获得最终性能与工艺的最优方案，而基于数据的分析可以考虑多个参数的作用，形成一个全局性的解决思路。

在课程的设置上，加入数据分析与挖掘的内容，不仅要考虑数据处理技术，而且需要这些技术在材料或相关科学中应用的示例，使之真正成为一门数据挖掘技术在材料科学中的应用课程。同时因为数据分析与挖掘技术大量涉及计算机和数学等方面的知识，相应地在这些基础课程的学习过程中需要加大难度。

在授课教师的选拔和教师队伍的建设上，需要一支具有交叉学科知识、勇于创新精神、肯于和善于不断学习新知识新技术的人才队伍。不同于金属结构材料计算中多尺度计算的跨层次的要求，也不同于功能材料的第一性原理计算结果的单一参数作为判据的筛选，材料数

据的分析与挖掘作为一门相对独立的方向时，由于数据的生产途径和代表的含义不同，其研究内容覆盖了几乎材料研究的所有内容，即成分、工艺、组织、性能及服役等材料五要素间的复杂关系和交互影响。

因此，作为一个全新的材料研究方法论，材料基因工程的基础教育开展的难度是不容忽视的，但其意义重大，关乎材料创新发展的步伐。抓住信息时代的机遇，培养能够满足时代需求的材料人才，是材料基因工程得以长期发展的关键基础设施。

5　存在的问题与展望

材料基因组计划与大数据计划在美国的相继提出，催生了中国的材料基因工程热，继而出现材料数据热，使材料数据的发展获得了新生。然而，对于材料数据的理解和材料大数据的理解，仍处于初级阶段，甚至对于数据的整合和数据库的建设，仍理解为一件很简单的事，使得在材料数据库的建设初期，就出现因材料专家对数据库专家的过高且不现实的要求而导致工作进程推动缓慢的情况。因此，个人观点认为，对材料数据存在虚热，需要等待降温后留下的一批真正热爱材料数据的人，将材料数据扎实而稳步地开展下去。

材料数据的整合是数据挖掘的基础，而挖掘是数据储备后的延伸工作。目前数据分析与挖掘技术在应用于材料科学领域时，需要与材料理论知识以及现有发展成果相结合，在方法选择和建模等步骤中都要将材料已有成果抽象化进行考虑。这就需要材料学科专业人员的大量介入和对数据分析处理的知识储备。然而，由于目前基础教育尚未跟进数据时代的发展，导致一个较为普遍的情况是材料与数据分析人员的需求无法对接，材料专业人员在数据分析上的知识匮乏导致双方的协同难以尽快推进。

材料数据的发展不是孤立的。数据来源于材料计算模拟和实验表征，本身被赋予了材料的含义及其与其他知识与数据的复杂关系，材料数据最终将成为材料知识的载体，成为新材料发现和发展的基础和手段。以需求出发的材料知识的抽象化和数字化可能成为今后的发展趋势之一。

6　结　语

依托于大量材料数据库资源和不断激增的数据，对材料数据的研究和分析挖掘正在成为新的材料研发模式。包括数据准确性和完整性的材料数据质量，以及基于对材料知识充分理解的材料数据的分析挖掘的质量，是决定材料数据作为研发新模式的发展进程的关键。材料数据分析挖掘要求材料知识与计算机和数学知识的高度融合，从本科生的基础教育抓起，才能保证新的研发模式的充分应用和可持续发展。

致　谢　本研究得到了国家科技部科技基础条件平台建设项目“材料科学数据共享网”(2005DKA32800)、国家高技术研究发展计划(“863”计划)“基于材料基因工程的高性能材料设计、制备与表征技术”(2015AA034201)、国家重点研发计划项目“材料基因工程专用数据库和材料大数据技术”(2016YFB0700503)，北京市科技计划项目(D16110300240000)以及美国肯纳金属有限公司的支持。

参考文献　References

[1] Hey T, Tansley S, Tolle K. *The Fourth Paradigm: Data-Intensive Scientific Discovery*[M]. Washington: 2009: 109-130.

[2] The White House. [EB/OL]. [2016-08-01]. https://www.whitehouse.gov/blog/2016/08/01/ materials-genome-initiative-first-five-years.

[3] Austin T. *Materials Discovery*[J], 2016(3): 1-12.

[4] 材料科学数据共享网[EB/OL]. (2016-07-01)[2017-01-10] http://matsec.ustb.edu.cn/uploadFiles/shujutijiao.pdf.

[5] Nosengo N.*Nature*[J], 2016, 533: 22-26.

[6] Jain A, Persson K, Ceder G. *APL Materials*[J], 2016, 4(053102): 1-14.

[7] Jain A, Ong S P, Hautier G, *et al*. *APL Materials*[J], 2013, 1(1): 011002.

[8] Richards W D, Tsujimura T, Miara L J, *et al*. *Nature Communications*[J], 2016, 7: 11009.

[9] Curtarolo S, Hart G L, Nardelli M B, *et al*. *Nat Mater*[J], 2013(12): 191-201.

[10] Perim E, Lee D, Liu Y, *et al*. *Nat Commun*[J], 2016, 7: 12315.

[11] Agrawal A, Deshpande P D, Cecen A, *et al*. *Integrating Materials and Manufacturing Innovation*[J], 2014, 3: 8-26.

[12] Singh S, Bhadeshia H, MacKay D, *et al*. *Ironmak Steelmak*[J], 1998(25): 355-365.

[13] Jeong J H, Ryu S K, Park S J, *et al*. *Computational Materials Science*[J], 2015(100): 21-30.

[14] Raccuglia P, Elbert K C, Adler P D F, *et al*. *Nature*[J], 2016(533): 73-77.

[15] Rajan K. *Informatics for Materials Science and Engineering*[M]. Elsevier Inc., 2013: 21-23.

(编辑　惠　琼)