科学通报 2015 年 第 60 卷 第 8 期: 694 ~ 703
专题: 大数据与化学 评 述 www.scichina.com csb.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

大数据与化学数据挖掘

刘言, 蔡文生, 邵学广*

南开大学化学学院分析科学研究中心, 天津 300071
* 联系人, E-mail: xshao@nankai.edu.cn

2014-11-03 收稿, 2014-12-08 接受, 2015-01-16 网络版发表

摘要 数据是重要的战略资源, 大数据挖掘技术已成为学术界、企业界甚至各国政府关注的热点. 本文介绍了大数据的基本概念及发展现状, 综述了与化学研究有关的大数据研究状况, 讨论了大数据在基础理论与关键技术2个层面上的主要问题以及大数据挖掘技术在化学各领域中的应用, 并对大数据发展的未来及其在化学学科中的应用前景进行了展望.

关键词
大数据
数据挖掘
可视化
云计算
化学

1 大数据的基本概念

随着人类对自然和社会认识的进一步加深及人类活动的进一步扩展, 科学研究、互联网应用、电子商务、移动通讯等诸多领域产生了多种多样、数量巨大的数据. 在此背景下, 一个崭新的概念——大数据(big data)应运而生, 成为世界各国关注的热点. 大数据挖掘技术及其应用创造了巨大价值, 对国家治理模式、企业决策、组织和业务流程以及个人生活方式都将产生巨大影响.

大数据尚无统一的定义. 一般认为, 大数据是一种新现象, 具有4个带"V"字的特点: (1) 数据体量(volume)巨大, 达TB级, 甚至PB级; (2) 数据种类(variety)繁多、来源复杂、格式多样, 除了结构化数据, 还有半结构化和非结构化数据; (3) 价值(value)密度低, 在大量的数据中, 有价值的信息比例不高. 例如在连续监控视频中, 有用数据可能仅为1, 2 min, 甚至1, 2 s. 但是大数据中蕴藏的信息非常丰富, 可挖掘价值很高; (4) 速度(velocity)快, 数据的产生和增长速度快, 对数据的处理的速度也要快.

当前, 各行各业都遇到大数据问题. 例如, 商界利用大数据关联分析, 通过了解消费者行为模式的变迁而发现新的商机[1]、优化库存和物流缓和供需矛盾、控制预算开支、提高服务质量. 在医疗领域, 大数据分析被用于复杂疾病的早期诊断[2]、心血管病的远程治疗[3]、器官移植[4]、HIV抗体的研究[5]等已经取得了一定的效果. 在生命科学领域, 大数据技术被用于基因组学[6]、生物医学[7]、生物信息学[8]等研究. 此外, 大数据技术还被用于温室气体排放的检测[9]以及政府信息管理[10]等公共领域.

2 大数据的发展现状

2008年, *Science*发表文章"Big data: Science in the petabyte era"[11]. 2011年, 麦肯锡公司发布了《大数据: 下一个前沿, 竞争力、创新力和生产力》的调研报告, 指出大数据研究将带来巨大价值. 2012年, 美国奥巴马政府宣布投资2亿美元启动"大数据研发计划", 旨在提高和改进从海量和复杂数据中获取知识的能力, 加速美国在科学和工程领域发明的步伐, 巩固国家安全. 大数据从此成为世界关注的热点.

各国纷纷提出了自己的大数据研究计划, 其中美国和中国的投入最大. 在美国, 联邦政府建立了统一的门户开放网站——Data.Gov, 开放部分公共数据, 鼓励民众对其进行自由开发. 美国的国家科学基

引用格式: 刘言, 蔡文生, 邵学广. 大数据与化学数据挖掘. 科学通报, 2015, 60: 694–703
Liu Y, Cai W S, Shao X G. Big data and chemical data mining (in Chinese). Chin Sci Bull, 2015, 60: 694–703, doi: 10.1360/N972014-01106

金委员会(NSF)、美国国家卫生研究院(NIH)、美国能源部(DOE)、美国国防部(DOD)、美国地质勘探局(USGS)等部门联合推出了大数据计划，旨在提升从大量复杂数据中获取知识和洞见的能力[12]. 中国工业信息化部发布了物联网“十二五”规划，把信息处理技术作为4项关键技术创新工程之一. 海量数据存储、数据挖掘、图像视频智能分析是大数据研究的重要组成部分. 另外3项，即信息感知技术、信息传输技术和信息安全技术，也与大数据密切相关. 2012年中国科学院启动了“面向感知中国的新一代信息技术研究”战略性先导科技专项，其任务之一就是研制用于大数据采集、存储、处理、分析和挖掘的未来数据系统. 同时，中国计算机学会成立了大数据专家委员会；为探讨中国大数据的发展战略，中国科学院计算机研究所举办了以“网络数据科学与工程——一门新兴的交叉学科”为主题的会议，与国内外知名专家学者一起为中国大数据发 展战略建言献计；2013年，中华人民共和国科学技术部正式启动国家高技术研究发展计划“面向大数据的先进存储结构及关键技术”，启动了多个大数据课题.

有关大数据的基础和应用研究近几年得到了迅速发展. 图1是web of science核心期刊数据库以“big data”为关键词进行检索得到的历年发表文章数的统计结果(截止日期为2014-11-28). 从图中可以清楚地看出，近几年与大数据相关的文献数量呈现出爆炸性增长态势. 2004年前后与大数据相关的文献每年仅有几篇，到2010年左右文献数量增加到每年十几篇. 而到2012年，这一数字跃增到256篇，2013年更是突增到985篇. 截止到2014年11月，发表文章数目已达到843篇. 预计大数据研究将会持续升温.

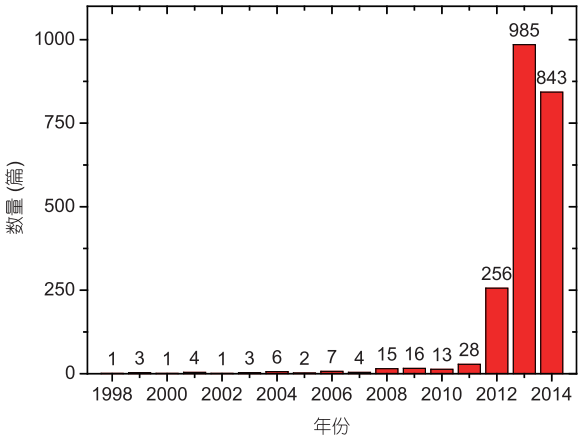

图 1 (网络版彩色)Web of science 上以“big data”为关键词检索得到的历年文献数

Figure 1 (Color online) The number of literatures in each year by searching the key words “big data” on web of science

正是由于中美两国的巨大投入，在大数据方面的研究成果也最为突出. 图2是web of science核心期刊数据库以“big data”为关键词进行检索得到的相关文献按国籍进行统计的结果(截止日期为2014-11-28). 从图中可以清晰地看出，美国发表的与大数据相关的文献占了总数的39.56%，在所有国家中列第1位. 这一数量超过了排名第2~4位国家文献数量的总和，也超过了排名在第5位之后的所有国家文献数量的总和. 中国以15.62%排名第2位，虽然文献数量比排名第3的英国(6.26%)和第4的德国(5.39%)高出不少，但是与美国相比仍然存在不小的差距.

从web of science核心期刊数据库的检索结果还可以看出大数据研究的学科分布. 统计结果表明，计算科学、工程和电信类的文献数量排在前3位，相关文献数多达1116，608和157，分别占文献总数的50.98%，27.78%和7.17%. 这一结果表明针对大数据的基础理论研究以及大数据应用上某些关键技术的研究仍是目前科学界关注的重点. 而排名4~8位的则是大数据应用比较广泛的商业、健康保障服务和医疗信息学等领域. 这一结果说明大数据在这些领域应用广泛，相关的研究工作也在进行展开. 但是与化学学科相关研究方向的文献数量则相对较少，生物化学和分析生物学领域的文献数量排在第11位，而化学类文献数量则更少，只排在第20位，文献数量仅有31篇，占总数的1.42%. 因此，与计算机、商业等领域相比，化学领域与大数据相关的文献数量仍然比较少，大数据技术在化学及其相关学科之中的应用与发展，仍然处于起步阶段，有着很大的上升空间. 在当前化学数据飞速增加的时代，化学大数据的挖掘

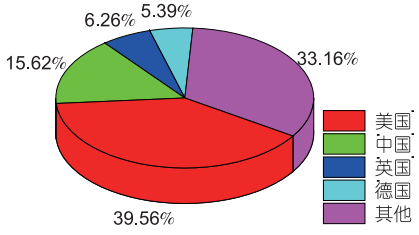

图 2 (网络版彩色)Web of science 上以“big data”为关键词检索得到各国文献数所占比例

Figure 2 (Color online) The percentage of literatures in each country by searching the key words “big data” on web of science

仍需要更大的投入.

3 大数据的研究内容

一般认为, 大数据的处理过程包括采集、处理与集成、分析和解释4个步骤[13]. 大数据研究的主要内容涉及这4个步骤在实际实施过程中的相关问题. 数据采集是大数据处理流程中最为基础的一步, 即使用传感器收取、射频识别(RFID)、搜索引擎、条形码识别等数据采集技术, 从外界获取数据. 大数据的"大", 原本就意味着数量多、种类复杂, 因此, 通过各种不同的方法获取数据信息便显得格外重要. 数据的处理与集成主要是对已经采集到的数据进行适当的处理并进一步集成后进行存储. 大数据另一个特点便是其多样性, 这就决定了经过各种渠道获取的数据种类和结构都非常复杂, 这给之后的数据分析处理带了极大的困难. 通过数据处理与集成, 将结构复杂的数据转换为单一或便于处理结构的数据, 为以后的数据分析打下良好的基础. 同时, 由于采集到的数据中往往会掺杂很多噪音和干扰, 还需要对这些数据进行"去噪"和"清洗", 以保证数据的质量以及可靠性. 常用的方法是在数据处理的过程中设计一些数据过滤器, 通过聚类或关联分析的规则方法将无用或错误的离群数据挑出来过滤掉, 防止其对最终数据结果产生不利影响. 然后将这些整理好的数据进行集成和存储. 目前主要的方法是针对特定种类的数据建立专门的数据库, 将这些不同种类的数据信息分门别类的放置, 这样可以有效地减少数据查询和访问的时间, 提高数据提取速度.

数据分析是整个大数据处理流程里最为核心的部分, 在数据分析的过程中, 会发现数据的价值所在. 由于大数据其本质上来说仍然是数据, 因此传统的数据处理分析方法, 包括聚类分析、因子分析、相关分析、回归分析[14]等仍然可以用于对大数据进行分析. 但这些方法在处理大数据时也存在这许多问题. 首先, 传统数据分析方法大多数都是通过对原始数据集进行抽样或者过滤, 然后对数据样本进行分析, 寻找特征和规律, 其最大的特点是通过复杂的算法从有限的样本空间中获取尽可能多的信息由于大数据极大的数据量, 而大数据本身巨大的数据量对于机器硬件以及算法本身都是严峻的考验. 其次, 大数据的应用常常具有实时性的特点, 算法的准确率不再是大数据应用的最主要指标, 很多实际应用过程中算法需要在处理的实时性和准确率之间取得一个平衡, 这便要求传统的分析方法能够根据应用的需求进行调整. 最后, 当数据量增长到一定规模以后, 可以从小量数据中挖掘出有效信息的算法并不一定适用于大数据. 正是由于这些局限性, 传统的分析方法在对大数据进行分析时必须进行调整和改进. 此外, 为了更好地对大数据进行分析, 出现了许多专门针对大数据的分析方法. 大数据分析方法与传统分析方法的最大区别在于分析的对象是全体数据, 而不是数据样本, 其最大的特点在于不追求算法的复杂性和精确性, 而追求可以高效地对整个数据集的分析. 目前一些大数据具体处理方法主要有散列法[15]、布隆过滤器(Bloom Filter)[16]、Trie树[17,18]等. 同时, 针对不同类型的数据, 也存在不同的分析方法. 如对文本进行分析的自然语言处理(NLP)技术[19]、对Web进行分析的Page Rank法[20]和CLEVER法[21]、对多媒体进行分析的摘要系统以及对社交网络进行分析的概率法[22]和线性代数法[23]等.

如前所述, 大数据本身巨大的数据量对于机器硬件以及算法本身都是严峻的考验. 随着数据量的膨胀, 单台机器在性能上已经无法满足分析和处理的需要. 为了实现对大数据的分析, 并行计算和分布式的存储与管理, 也就是云技术势在必行[24]. 云技术最早由Google公司提出, 主要由分布式文件系统(GFS)[25]、分布式数据库(BigTable)[26]、批处理技术(MapReduce)[27,28]以及开源实现平台(Hadoop)[29]4大部分组成. 其中, GFS是基于分布式集群的大型分布式处理系统, 通过数据分块、追加更新等方式实现海量数据的高效存储, 为MapReduce计算框架提供低层数据存储和数据可靠性的保障; BigTable是分布式数据库, 通过一个多维稀疏排序表以及多个服务器实现对大数据的分布管理. MapReduce是云技术的核心, 即通过批处理的方法实现对大数据的分析; MapReduce技术主要由Map和Reduce 2部分组成, 首先将用户的原始数据源进行分块, 然后分别交给不同的Map任务区处理. Map任务从输入中解析出链/值(Key/Value)对集合, 然后对这些集合执行用户自行定义的Map函数得到中间结果, 并将该结果写入本地硬盘. Reduce任务从硬盘上读取数据之后会根据Key值进行排序, 将具有相同Key值的组织在一起; 最后用户自定义的Reduce函数会作用于这些排好序

的结果并输出最终结果. MapReduce的设计思想在于将问题分而治之, 同时把计算推到数据而不是把数据推到计算, 有效地避免数据传输过程中产生的大量通信开销. Hadoop是一个由Java编写的云计算开源平台, 通过Hadoop可以将前面提到的传统数据分析技术以及专门针对大数据的分析技术编写成基于MapReduce计算框架的程序, 实现对大数据的分析. 云技术使得前面叙述的各类分析方法能够在实际应用中得到实现, 意义十分重大. 因此, 出现了大量针对云技术的研究与应用, 如针对GFS的改进, 出现了Colosass, Hay-stack和TFS等新的管理系统; 针对MapReduce的改进, 出现了Pregel, Dremel和Dryad等新的并行计算方法; 同时也出现了与BigTable功能类似的Dynamo和PNUTS等新的数据库; 而各种对Hadoop改进并将其应用于各种场景的大数据处理, 更是成为新的研究热点[30~32].

对于广大的数据信息使用者来讲, 最关心的并非是数据的分析处理过程, 而是对大数据分析结果的解释与展示. 因此, 在一个完善的大数据分析流程中, 数据结果的解释步骤至关重要. 若数据分析的结果不能得到恰当的显示, 则会对大数据使用者产生困扰, 甚至会误导使用者. 传统的数据展示方式是用文本形式下载输出或用户个人电脑显示处理结果, 但随着数据量的加大, 数据分析结果往往也越复杂, 用传统的数据显示方法已经不足以满足大数据分析结果输出的需求. 因此, 为了提升对大数据的解释和展示能力, 数据可视化技术作为一种解释大数据最有力的方式, 得到了广泛的应用和蓬勃的发展. 通过可视化结果分析, 抽象的数据表现成为可见的图形或图像在屏幕上显示出来, 以图形化的方式更形象地向使用者展示数据分析结果, 方便使用者对结果的理解和接受[33~35]. 目前, 学术科研界不停地致力于大数据可视化的研究, 发展出了基于集合的可视化技术、基于图标的技术、基于图像的技术、面向像素的技术和分布式技术等. 同时, 商业上已经有了很多经典成功的可视化应用案例. 如网络上用于标示不同标签对象的标签云(Tag Cloud)技术[36], 用于可视化文档编辑的历史流图(History Flow)[37]等. 最近, 俄罗斯工程师Ruslan Enikeev将196个国家的35万个网站数据整合起来, 并根据这些网站相互之间的链接关系设计开发了互联网宇宙(the Internet Map, http://internet-map.net/).

4 化学及其相关学科中的大数据研究

目前, 由于实验方法的丰富和学科之间交流的加快, 化学学科的发展同样进入了一个数据量爆炸性增长的时期. 在化学学科中的某些领域中也出现了大数据的身影, 给大数据技术在化学领域的应用带来了极大的空间. 与其他学科和领域不同, 化学是一门比较保守的学科, 在研究时不擅于分享数据, 化学家们对于从数据中得到结论的重视程度远大于数据本身. 而这一点正随着大数据的产生而发生改变, 越来越多的化学家们认识到了数据收集和交流的重要性. 以化学信息搜索和分析为研究领域的化学信息学家, 敏锐地发现这一点, 许多工作也因此而展开.

为了方便化学家更好地进行交流, 对化学物质名字进行统一和标准化成为了一项重要的工作. 为此, 国际纯粹与应用化学联合会(IUPAC)推出了International Chemical Identifier(InChI)以及与之配套的InChIKey. 该系统取代了旧有的Simplified Molecular-Input Line-Entry System(SMILES)方法, 成为一种标准化的、可以被索引和机器识别的化学结构表达方式, 这极大地方便了数字时代下的化学家之间的交流和研究工作. 在一些与计算化学和分子模拟等与计算机相关的领域, 大数据的研究和应用工作正在进行. 一些学者尝试将各种各样的分子描述符进行统一和集成, 以便统一进行管理, 方便机器查找和索引. 同时, 旧有的信息分析平台如Cambridge Structural Database(CSD)和Protein Data Bank(PDB)被改造和升级以适应大数据时代的需要, 更有许多新的数据检索平台, 如Collaboratory for Multi-scale Chemical Sciences (CMCS)和Chemical Informatics and Cyberinfrastructure Collaboratory(CICC)等出现以方便化学家进行数据的收集和交流[38].

我国在化学信息搜索和分析方面也做出了大量的工作. 李晓霞课题组[39~42]开发了化学深层网检索引擎ChemDB Portal, 具备通过不同检索方式, 包括名称、分子式、CAS号检索、结构检索等方式, 实时在线检索多来源网络数据库的功能, 实现了化合物数据信息的多途径集成检索和利用. 利用ChemDB Portal, 用户仅需输入一次查询请求(可以是1个化合物的CAS号/名称/分子式或者在线画出的化学结构图或提交分子结构的mol文件), 该系统就可自动检索网络上的多个专业数据库(包括物化性质、化合物安全数据表MSDS、试剂供应商等), 把从各库检索得到

结果统一返回给用户. 目前, ChemDB Portal索引了约50万个化合物、超过100万种产品的信息, 可在线同时检索十几个化学数据库的物性数据、MSDS等数据源. 以此为基础, ChemDB Portal可以逐步衍生出更多的数据服务如建立化学品与化学文献的动态链接、建立原始实验数据的respository、构建基于化学品的在线计算服务如毒性预测等各种功能. 姚建华课题组[43]开发了化学信息管理系统CISOC-ChIMS, 具有化学结构检索以及文字检索2大检索功能, 可以进行数据库的维护、中文处理、图形存储, 尤其是中文处理功能弥补了其他国外开发的化学信息管理系统在中文处理上的不足. Hou课题组[44]开发了作为计算生物学和计算机辅助药物设计(CADD)相关软件的开发基础的函数库(molecular objects and relevant templates, MORT). 与其他的一些函数库相比, MORT使用C++编写, 充分利用了C++的面向对象的思想, 使其易于理解并具有良好的可拓展性; 同时, 在表征分子时, MORT采用了关系模型, 与那些使用层次模型的函数库相比有着更大的灵活性; 此外, MORT中包含了大量的功能函数, 能对一个分子进行各种处理, 这对于计算生物学和CADD的程序开发者来说是极大的便利. Li课题组[45~48]设计开发了基于结构特异性得分矩阵(SPSSM)的蛋白质二级结构的数据库. 该数据库记录了900万种蛋白质序列的结构特异性得分矩阵, 通过该数据库可以很容易地对未知蛋白质的二级结构进行预测, 是一种比较成功的蛋白质二级结构预测工具.

在药物化学领域, 大数据的出现已经深远地影响了药物化学家的开发和研究新型药物的方式. 传统的药物开发由设计、合成、测试、评价4个流程的交替循环组成[49], 但这一流程随着药物化学领域数据量的直线上升而受到极大地冲击. 根据Chemical Abstract Services Registry 2014年提供的数据, 已知的药物基准物质已经达到了74000000种, 而这一数量还在逐年增加. 同时, 随着实验技术的提高, 各种检测手段层出不穷, 这也使得实验数据与以往相比呈现了级数式的增长. 分析这些海量的数据并作出决策, 使用传统的分析手段往往需要耗费大量的时间, 而在分析的过程中, 往往又会产生了大量的新实验数据. 由于数据的更新速度大于决策速度, 而更新产生的数据又有可能改变设计决策的方向, 这使得制定设计决策变得越来越困难. 因此, 必须加强和大数据相关的计算机领域的合作, 借鉴和学习其管理与分析大数据的经验. 为了方便药物化学家进行大数据的管理与决策, 许多专业的数据存储库以及决策支持工具, 如Integrated Project View(IPV)[50], ArQule公司的ArQiologist[51], Amgen公司的Amgen's Data Access Analysis Prediction Tools (ADAAPT)[52], Actelion公司的OSIRIS[53]和Johnson&Johnson公司的Advanced Biological and Chemical Discovery System (ABCD)[54]等被开发出来. 在这些管理软件的帮助下, 实验者们可以在自己电脑屏幕上分析和管理自己的实验数据, 分析和决策也变得相对容易. 同时, 大数据的出现对药物化学本身也提出了新的要求. 为了对大数据进行分析, 常用的数据分析方法主成分分析、线性回归、k均值聚类、贝叶斯方法、交叉验证等各种监督学习、模型预测、聚类分析、数据挖掘理论成为了药物化学家必须掌握的基础理论. 药物化学家也要由传统的根据研究做出决策的研究模式改为根据数据做出决策的研究模式. 数据的来源变得多样化, 可以是自己实验获得的, 也可以是公共数据和他人的数据. 许多的研究成果甚至可以不进行实验, 仅对数据库中的数据进行分析就可以得到重要的结论, 如Lipinski通过对2245个药物分子进行分析, 得到口服药物的通用性质[55]、通过对数据库进行分析得到G蛋白偶联受体的标靶药物的通用性质[56~58]等.

微流控芯片技术, 作为化学领域一个比较热门的领域, 从诞生之初就倍受关注. 近年来, 随着微流控芯片技术的发展, 芯片实验室产生的数据量和数据种类大量增加, 大数据的出现, 为管理和研究这些数据, 提供了一个可行的方案. 例如, Ozcan课题组[59]提出了的一种微流控芯片大数据管理平台BioGames, 对于下一代微流控芯片数据的管理有很大的启示作用. BioGames的核心是一种基于智能群体分包(crowed-scoured)的二元判定(binary decision)系统. Ozcan及其团队开发了一款可以在手机、电脑及平板上运行的游戏, 游戏的内容十分简单, 只需要玩家根据给定的图像在另一组图像中找出与之类似的图像. 其中, 给定的图像为微流控技术得到的患有某类疾病的人体细胞图像, 另一组图像则为微流控技术得到的疑似病人的细胞图像. 通过收集游戏玩家的选择结果, 开发者们对疑似病人进行二次判断, 从而得到最终的诊断结果. 作者以疟疾为例对该平台的诊断效果进行了检测, 超过60个国家接近1000名

玩家参与了该游戏, 结果显示, 大量未经训练的普通人参与游戏后统计得到的诊断结果与专家的判断结果类似, 系统的有效性得到了很好的验证. 随着便携低成本的成像、传感技术与高通量的微流控芯片技术相结合, 将会有大量多尺度的生物医学、环境等方面的数据出现. BioGames通过智能群体分包和数字游戏的策略来实现诊断的概念可以帮助我们更好地处理下一代成像、传感、微流控技术产生的大数据.

5 化学计量学中的大数据问题

作为化学领域中专门处理数据的学科, 化学计量学有着特殊的地位. 通过统计学或数学方法将对化学体系的测量值与体系的状态之间建立联系, 化学计量学实现了对化学数据的分析与挖掘. 目前, 化学计量学的方法已经广泛应用于化学的各个领域, 分析与挖掘各种类型的化学数据. 分子模拟、计算机辅助药物设计、虚拟筛选(VHTS)和定量构效关系(QSAR)等化学计量学技术推动了生命科学和生物医药领域的发展, 促进了新药的研发和创制[60~63]. 理论化学在理解物质结构和性质、解释化学反应机理等方面取得了飞速发展, 在结构化学、材料科学和生命科学领域中发挥着不可替代的作用[64]. 由于多元校正及模式识别技术的发展, 近红外光谱(NIR)技术得到了广泛应用, 已成为复杂体系分析、产品质量评价与控制、环境检测与控制、生命与健康等领域的关键技术之一[65~67]. 同时, 复杂信号和高维分析化学信号的解析技术推动了分析化学的发展, 大大增强了分析化学解决实际问题的能力[68].

随着化学计量学在化学各个领域的深入发展, 分析数据的数量级逐渐变大, 许多数据分析的过程中均出现了“大数据化”的特征, 而相应的方法也随着数据量的增大而随之发展. 如在分子模拟领域, 随着图形处理单元(graphics processing unit, GPU)快速发展, GPU在计算能力和存储器带宽上的优势使之为提高分子动力学模拟的计算能力提供了新的可能. GPU作为一种具有极强运算能力的多核处理器, 成为高性能计算领域的主要发展方向, 大量的研究工作也随之展开[69~71]. 在药物设计领域, 研究者发现生物体内存在大量被称为化学基元(chemoyl)的基本结构单元, 这些结构单元在生物的活动过程中起着重要作用. 在此基础上, 出现了以超级计算与大数据挖掘技术为基础, 研究各种化学基元的结构、组装与演化的基本规律的药物分子设计的新理论——化学基元学[72,73]. 化学基元学通过揭示生物系统制备化学多样性的规律, 发展仿生合成方法制备类天然化合物库(quasi natural product libraries)以供药物筛选, 成功解决了药物设计领域药物筛选资源日益枯竭这样一个瓶颈问题. 目前, 该理论已发展出了在超级计算支持下基于分子动力学的虚拟筛选方法(MDVS)[74], 基于GPU的分子三维叠合并行算法gWEGA[75], 面向系统性疾病治疗药物设计的药理网络[76]以及分子活性构象预测的新技术[77]等. 在近红外光谱的应用领域, 由于大量在线数据的出现, 传统的定性定量分析开始逐渐向在线分析与过程质量控制进行转变[78,79]. 在许多领域, 基于近红外光谱的物联网系统和数据库系统也在逐渐形成并成为发展的主要趋势.

大数据的可视化问题一直是大数据研究的热点问题. 在化学计量学领域, 学者们提出探索性资料分析(exploratory data analysis, EDA)的概念[80], 用于对不同类型的化学数据进行挖掘, 以研究其中的规律. 其中, 主成分分析(PCA)和偏最小二乘(PLS)是2种最为常用且有效的分析方法. 两者均是基于数据本身潜在结构的投影模型, 原始数据通过投影计算被表示成几个不同主成分(principle component)或者潜变量(latent variable)下的得分, 并通过得分图(score plot)显示出来. 由于得分图具有直观的表现形式, 可以让研究人员很容易地发现数据内部潜在的规律, 成为了一种非常行之有效的可视化工具. 然而, 随着数据量的增大, 大量样品的得分在传统的得分图上往往由于重叠无法很好地进行观察, 这在一定程度上影响到了研究人员从得分图中获得有效信息. 同时, 数据量的增大也降低了PCA与PLS的计算速度, 对于某些数据而言, 其分析计算的速度甚至赶不上数据更新的速度, 从而严重影响到了数据分析的有效性. 为此, Camacho[81]提出了压缩得分图(compressed score plots)的概念, 对传统的得分图进行改进, 使之能够直观地表现大容量和快速更新的化学数据. 对于大容量的数据, 使用聚类的方法来减少得分图上的数据点数量, 以绘制聚类的中心点来代替原始数据点的得分, 有效减少了得分图上的数据点数. 同时, 为了最大限度地保留原始得分图上的信息, 对于聚类得到的中心点, 以中心点的大小来表示该点中包含原始数据点的多少. 为了减少每次计算的耗时, 使用并行计算的理论(基于分布式文件系统的Hadoop)来

进行计算和编程. 对于更新速度较快的数据, 采用指数加权移动平均(exponentially weighted moving average)的方法来对其进行更新操作, 避免了对全部数据的重复计算, 有效减少了计算耗时. 化学计量学领域的此类方法, 对于解决大数据可视化问题, 有着很重要的借鉴意义.

6 大数据的未来及其对化学学科发展的影响

随着近年来大数据热潮的不断升温, 人们认识到"大数据"并非是指"大规模的数据", 而是一种规模更大、种类更多、数据更广泛、价值更高同时处理难度更大的全新数据模式. 大数据的出现, 对产业界、学术界和教育界正在产生巨大影响. 随着科学家对大数据研究的不断深入, 人们意识到对数据的利用可以为其生产生活带来巨大便利的同时, 也带来了不小的挑战. 其中, 大数据的安全与隐私问题、大数据的集成与管理问题、大数据的IT技术架构问题以及大数据的生态环境问题成为大数据发展过程中出现的亟待解决的几个重要问题. 如何面对这几个问题的挑战, 对大数据未来的发展至关重要.

对化学学科而言, 大数据在其中的应用仍然处于起步阶段. 目前化学领域大数据的应用都是数据标准化、数据挖掘、数据可视化等比较简单的应用. 而大型数据的管理与分析、云计算以及基于网络的数据传输和运算, 大型分析软件的开发等大数据的核心技术以及真正的优势部分, 在化学领域的应用体现的不够多. 这从另一个角度说明大数据在化学学科内的应用存在着广阔的应用空间. 在未来, 随着大数据技术的发展和完善, 以采集、处理、分析为基础的传统分析仪器将会逐步被小型化、便携式的新型分析仪器所取代. 分析仪器最终将简化为一个带有数据传输功能的检测器, 在采集数据之后将数据直接传输到大数据的分析平台上, 所有的数据处理与分析功能均在这个平台上完成. 同样, 基于PC机、小容量、统一数据类型的传统化学数据管理方式也会逐渐被以云技术为代表的大数据管理和存储模式所取代. 新的数据管理和存储模式以大型服务器为基础, 可以轻松管理海量不同领域、不同类型的化学数据. 而在化学信息领域, 设计和建设以大数据算法为基础的化学搜索引擎和化学信息数据库可以帮助研究者进一步研究和挖掘各种类型的化学信息, 加深对研究内容的理解. 同样, 在化学计量学领域, 开发以大数据分析技术为基础的新型数据分析方法将帮助研究者们更容易地处理大容量、复杂来源的化学数据. 而大数据的可视化技术也可以直观地帮助研究者们表达和解释研究的结果. 可以看到, 大数据对化学, 尤其是化学仪器、化学数据的管理与分析、化学信息学和化学计量学将产生深远而巨大的影响.

参考文献

1 Spiess J, T'Joens Y, Dragnea R, et al. Using big data to improve customer experience and business performance. Bell Labs Tech J, 2014, 18: 3–17

2 Liu R, Wang X D, Aihara K, et al. Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network. Med Res Rev, 2014, 34: 455–478

3 Hsieh J C, Li A H, Yang C C. Mobile, cloud, and big data computing: Contributions, challenges, and new directions in telecardiology. Int J Environ Res Public Health, 2013, 10: 6131–6153

4 Massie A B, Kuricka L M, Segev D L. Big data in organ transplantation: Registries and administrative claims. Am J Transplant, 2014, 14: 1723–1730

5 Gonzalez-Diaz H, Herrera-Ibata D M, Duardo-Sanchez A, et al. ANN multiscale model of anti-HIV drugs activity vs AIDS prevalence in the US at county level based on information indices of molecular graphs and social networks. J Chem Inf Model, 2014, 54: 744–755

6 O'Driscoll A, Daugelaite J, Sleator R D. "Big data", Hadoop and cloud computing in genomics. J Biomed Inform, 2013, 46: 774–781

7 Costa F F. Big data in biomedicine. Drug Discov Today, 2014, 19: 433–440

8 Dai L, Gao X, Guo Y, et al. Bioinformatics clouds for big data manipulation. Biol Direct, 2012, 7: 43–49

9 Tang H, Yang X, Zhang Y J. Effort at constructing big data sensor networks for monitoring greenhouse gas emission. Int J Distrib Sens Netw, 2014, 14: 1–7

10 Kim G H, Trimi S, Chung J H. Big-data applications in the government sector. Commun ACM, 2014, 57: 78–85

11 Graham-Rowe D, Goldston D, Doctorow C, et al. Big data: Science in the petabyte era. Nature, 2008, 455: 1–50

12 Huang Z X, Cao F Y, Li J J, et al. Developing sea cloud data system key technologies for large data analysis and mining (in Chinese). J Netw New Media, 2012, (1): 20–26 [黄哲学, 曹付元, 李俊杰, 等. 面向大数据的海云数据系统关键技术研究. 网络新媒体技术, 2012, (1): 20–26]
13 Meng X F, Ci X. Big data management: Concepts, techniques and challenges (in Chinese). J Comput Res Dev, 2013, 50: 146–169 [孟小峰, 慈祥. 大数据管理: 概念、技术与挑战. 计算机研究与发展, 2013, 50: 146–169]
14 Wu X D, Zhu X Q, Wu G Q, et al. Data mining with big data. IEEE Trans Knowl Data Eng, 2014, 26: 97–107
15 Andoni A, Indyk P. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. Commun ACM, 2008, 51: 117–122
16 Bloom B. Space/time tradeoffs in hash coding with allowable errors. Commun ACM, 1970, 13: 422–426
17 Buddhikot M M, Suri S, Waldvogel M. Space decomposition techniques for fast layer 4 switching. Protoc High Speed Netw, 1999, 25–41
18 Geraci F, Pellegrini M, Pisati P, et al. Packet classification via improved space decomposition technique. In: Makki K, Knightly E, eds. Proceedings IEEE Infocom 2005. Piscataway, NJ : IEEE, 2005. 1: 304–312
19 Melton C B, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc, 2005, 12: 448–457
20 Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Comput Netw ISDN Syst, 1998, 30: 101–117
21 Konopnicki D, Shmueli O. W3QS: A query system for the worldwide Web. In: Dayal U, Gray P M D, Nishio S, eds. VLDB'95. Proceedings of the 21st International Conference on Very Large Data Bases. San Francisco, CA: Morgan Kaufmann, 1995. 54–65
22 Ninagawa A, Eguchi K. Link prediction using probabilistic group models of network structure. In: Shin S Y, Ossowsi S, Schumacher M, eds. Proceedings of the 2010 ACM Symposium on Applied Computing. New York: ACM, 2010. 1115–1116
23 Dunlavy D M, Kolda T G, Acar E. Temporal link prediction using matrix and tensor factorization. ACM Trans Knowl Discov Data, 2011, 5: 10
24 Li Q, Zheng X. Research survey of cloud computing (in Chinese). Comput Sci, 2011, 38: 32–37 [李乔, 郑啸. 云计算研究现状综述. 计算机科学, 2011, 38: 32–37]
25 Ghemawat S, Gobioff H, Leung S T. The google file system. Operat Syst Rev, 2003, 37: 29–43
26 Chang F, Dean J, Chemawat S, et al. Big table: A distributed storage system for structured data. ACM Trans Comput Syst, 2008, 26: 4–15
27 Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. Commun ACM, 2008, 51: 107–113
28 Li C H, Zhang X F, Jin H, et al. MapReudce: A new programming model for distributed parallel computing (in Chinese). Comput Eng Sci, 2011, 31: 129–135 [李成华, 张新访, 金海, 等. MapReduce: 新型的分布式并行计算编程模型. 计算机工程与科学, 2011, 31: 129–135]
29 Shafer J, Rixner S, Cox A L. The Hadoop distributed filesystem: Balancing portability and performance. In: IEEE, ed. 2010 IEEE International Symposium on Performance Analysis of Systems & Software (ISPASS 2010). Piscataway, NJ: IEEE, 2010. 122–132
30 Yang C, Zhang X Y, Zhong C M, et al. A spatiotemporal compression based approach for efficient big data processing on Cloud. J Comput Syst Sci, 2014, 80: 1563–1583
31 Zhang X Y, Liu C, Nepal S, et al. A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud. J Comput Syst Sci, 2014, 80: 1008–1020
32 Zhao J Q, Wang L Z, Tao J, et al. A security framework in G-Hadoop for big data computing across distributed Cloud data centres. J Comput Syst Sci, 2014, 80: 994–1007
33 He Q B. The development and application of visualization technique (in Chinese). Sci Technol West China, 2008, 7: 4–7 [贺全兵. 可视化技术的发展及应用. 中国西部科技, 2008, 7: 4–7]
34 Liu K, Zhou X Z, Zhou R T. Data visualization research and development (in Chinese). Comput Eng, 2002, 28: 1–2 [刘勘, 周晓峥, 周洞汝. 数据可视化的研究与发展. 计算机工程, 2002, 28: 1–2]
35 Rysavy S J, Bromley D, Daggett V. DIVE: A graph-based visual-analytics framework for big data. IEEE Comput Graph Appl, 2014, 32: 26–37
36 Kaser O, Lemire D. Tag-cloud drawing: Algorithm for cloud visualization. Comput Res Reposit, 2007, 70: 109–118
37 Vigas F B, Wattenberg M, Dave K. Studying cooperation and conflict between authors with history flow visualizations. In: Dykstra-Erickson E, Tscheligi M, eds. CHI'04 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM, 2004. 575–582
38 Bird C L, Frey J G. Chemical information matters: An e-Research perspective on information and data sharing in the chemical sciences. Chem Soc Rev, 2013, 42: 6754–6776
39 Chu C M, Li X X, Guo L. Directed query engine application in the integrated retrieval of chemical Web databases. Comput Appl Chem, 2005, 22: 659–666
40 Zhuo L Y, Li X X, Guo L. Chemical deep Web data extraction with XML-based technology (in Chinese). Comput Appl Chem, 2006, 23: 1137–1141 [卓流艺, 李晓霞, 郭力. XML 技术在化学深层网数据提取中的应用. 计算机与应用化学, 2006, 23: 1137–1141]
41 Yuan X L, Li X X, Guo L, et al. Using open-source software in the structure searching of chemical database (in Chinese). Comput Appl Chem, 2008, 25: 1143–1146 [袁小龙, 李晓霞, 郭力, 等. 开源软件在化学数据库分子结构检索中的应用. 计算机与应用化学, 2008, 25: 1143–1146]

42 Liu Z C, Li X X, Yuan X L, et al. Management of chemical data knowledge framework based on SSH and ExtJS (in Chinese). Comput Appl Chem, 2008, 25: 1147–1151 [刘增才, 李晓霞, 袁小龙, 等. 基于 SSH+ExtJS 架构的化学数据知识框架管理. 计算机与应用化学, 2008, 25: 1147–1151]

43 Shen T X, Li F, Yao J H. CISOC-ChIMS: Chemical information management system (in Chinese). Comput Appl Chem, 2007, 24: 130–132 [沈天翔, 李丰, 姚建华. CISOC-ChIMS: 化学信息管理系统. 计算机与应用化学, 2007, 24: 130–132]

44 Zhang Q, Zhang W, Li Y Y, et al. MORT: A powerful foundational library for computational biology and CADD. J Cheminformatics, 2014, 6: 36–45

45 Sun J M, Li T H, Cong P S, et al. Retrieving backbone string neighbors provides insights into structural modeling of membrane proteins. Mol Cell Proteomics, 2012, 11: 1–7

46 Song Q, Li T H, Cong P S, et al. Predicting turns in proteins with a unified model. PLoS One, 2012, 7: 1–8

47 Li D P, Li T H, Cong P S, et al. A novel structural position-specific scoring matrix for the prediction of protein secondary structures. Bioinformatics, 2012, 28: 32–39

48 Cong P S, Li D P, Wang Z H, et al. SPSSM8: An accurate approach for predicting eight-state secondary structures of proteins. Biochimie, 2013, 95: 2460–2464

49 Lusher S J, McGuire R, van Schaik R C, et al. Data-driven medicinal chemistry in the era of big data. Drug Discov Today, 2014, 19: 859–868

50 Baede E J. Integrated project views: Decision support platform for drug discovery project teams. J Chem Inf Model, 2012, 52: 1438–1449

51 Rojnuckarin A. ArQiologist: An integrated decision support tool for lead optimization. J Chem Inf Model, 2005, 45: 2–9

52 Cho S J. ADAAPT: Amgen's data access, analysis, and prediction tools. J Comput Aided Mol Des, 2006, 20: 249–261

53 Osiris S T. An entirely in-house developed drug discovery informatics system. J Chem Inf Model, 2009, 49: 232–246

54 Agrafiotis D K. Advanced biological and chemical discovery (ABCD): Centralizing discovery knowledge in an inherently decentralized world. J Chem Inf Model, 2007, 47: 1999–2014

55 Lipinski C A. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev, 1997, 46: 3–26

56 Lowrie J F. The different strategies for designing GPCR and kinase targeted libraries. Comb Chem High Throughput Screen, 2004, 7: 495–510

57 Balakin K V, Tkachenko S E, Lang S A, et al. Property-based design of GPCR-targeted library. J Chem Inf Comput Sci, 2002, 42: 1332–1342

58 Sprous D G. QSAR in the pharmaceutical research setting: QSAR models for broad, large problems. Curr Top Med Chem, 2010, 10: 619–637

59 Mavandadi S, Dimitrov S, Feng S, et al. Crowd-sourced BioGames: Managing the big data problem for next generation lab-on-a-chip platforms. Lab Chip, 2012, 12: 4102–4106

60 He L J, Cui C X, Li S Y, et al. Study on interaction of naringin with human serum albumin (in Chinese). Chem Res Appl, 2014, 26: 1195–1199 [何丽君, 霍彩霞, 李生英, 等. 柚皮苷与人血清白蛋白相互作用的研究. 化学研究与应用, 2014, 26: 1195–1199]

61 Xue W, Jiao P, Liu H, et al. Molecular modeling and residue interaction network studies on the mechanism of binding and resistance of the HCV NS5B polymerase mutants to VX-222 and ANA598. Antiviral Res, 2014, 104: 40–51

62 Jiao P, Xue W, Shen Y, et al. Understanding the drug resistance mechanism of hepatitis C virus NS5B to PF-00868554 due to mutations of the 423 site: A computational study. Mol Biosyst, 2014, 10: 767–777

63 Xue W, Ban Y, Liu H, et al. Computational study on the drug resistance mechanism against HCV NS3/4A protease inhibitors vaniprevir and MK-5172 by the combination use of molecular dynamics simulation, residue interaction network, and substrate envelope analysis. J Chem Inf Model, 2014, 54: 621–633

64 Li J, Xu L W, Hu J, et al. Hydrolysis reaction mechanism of 2,4-dichlorophenoxy acetic acid metabolism (in Chinese). Acta Phys Chim Sin, 2013, 29: 1923–1930 [李佳, 徐雯丽, 胡静, 等. 2,4-二氯苯氧乙酸代谢中的水解反应机理. 物理化学学报, 2013, 29: 1923–1930]

65 Cai W S, Li Y K, Shao X G. A variable selection method based on uninformative variable elimination for multivariate calibration of near-Infrared spectra. Chemom Intell Lab Syst, 2008, 90: 188–194

66 Xu L, Zhou Y P, Tang L J, et al. Ensemble preprocessing of near-Infrared (NIR) spectra for multivariate calibration. Anal Chim Acta, 2008, 616: 138–143

67 Li Y K, Cai W S, Shao X G. A consensus least squares support vector regression (LS-SVR) for analysis of near-infrared spectra of plant samples. Talanta, 2007, 72: 217–222

68 Xia A L, Wu H L, Li S F, et al. Alternating penalty quadrilinear decomposition algorithm for an analysis of four-way data arrays. J Chemom, 2007, 21: 133–144

69 Shi J, Li X X, Liu Z L, et al. GPU-enabled implementations of Particle-Mesh-Ewald method (in Chinese). Comput Appl Chem, 2012, 29: 517–522 [石静, 李晓霞, 刘忠亮, 等. Particle-Mesh-Ewald(PME)算法在 GPU 上的实现. 计算机与应用化学, 2012, 29: 517–522]

70 Liu Z L, Li X X, Shi J, et al. GPU-based implementation of LINCS constraint algorithm for molecular dynamics simulation (in Chinese). Comput Appl Chem, 2012, 29: 907–912 [刘忠亮, 李晓霞, 石静, 等. 分子动力学模拟 LINCS 约束算法的 GPU 并行化. 计算机与应用化学, 2012, 29: 907–912]

71 Li X, Zheng M, Liu J, et al. Revealing chemical reactions of coal pyrolysis with GPU-enabled ReaxFF molecular dynamics and cheminformatics analysis. Mol Simul, 2015, 41: 13–27
72 Xu J, Gu Q, Liu H, et al. Chemomics and drug innovation. Sci China Chem, 2013, 56: 71–85
73 Gu Q, Yan X, Xu J. Drug discovery inspired by mother nature: Seeking natural biochemotypes and the natural assembly rule of the biochemome. J Pharm Pharm Sci, 2013, 16: 331–341
74 Ge H, Wang Y, Li C, et al. Molecular dynamics-based virtual screening: Accelerating drug discovery process by high performance computing. J Chem Inf Model, 2013, 53: 2757–2764
75 Yan X, Li J, Liu Z, et al. Enhancing molecular shape comparison by weighted gaussian functions. J Chem Inf Model, 2013, 53: 1967–1978
76 Wang L, Gu Q, Zheng X, et al. Discovery of new selective human aldose reductase inhibitors through virtual screening multiple binding pocket conformations. J Chem Inf Model, 2013, 53: 2409–2422
77 Chen N, Zhou J, Li J, et al. The concerted cyclization of lanosterol c-ring and d-ring under human oxidosqualene cyclase catalysis: An *ab initio* QM/MM MD study. J Chem Theory Comput, 2014, 10: 1109–1120
78 Liu J J, Xu H, Cai W S, et al. Discrimination of industrial products industrial products by on-line near Infrared spectroscopy with an improved dendrogram. Chin Chem Lett, 2011, 22: 1241–1244
Liu J J, Ma X, Wen Y D, et al. On-line near Infrared spectroscopy combined with alternating trilinear decomposition for process analysis of industrial production and quality assurance. Ind Eng Chem Res, 2011, 50: 7677–7681
79 Yu T. An exploratory data analysis method to reveal modular latent structures in high-throughput data. BMC Bioinformatics, 2010, 11: 440–452
80 Camacho J. Visualizing big data with compressed score plots: Approach and research challenges. Chemometrics Intell Lab Syst, 2014, 135: 110–125

Big data and chemical data mining

LIU Yan, CAI WenSheng & SHAO XueGuang

Research Center for Analytical Sciences, College of Chemistry, Nankai University, Tianjin 300071, China

Big data is fast becoming an important resource and a hot topic in academic research, business and government. In this paper, we introduce the concept of big data, and review advances in big data research, including technology for big data collection, cloud computing technology like Google's file system, BigTable, MapReduce and Hadoop, and data mining and visualization methods for big data. Big data are commonly defined by the so-called 4 V's, i.e., volume, variety, velocity, and value. High volume data with large variety make the analysis of big data much more difficult. Since velocity is important, fast high performance analysis methods are needed for big data. Moreover, the high value of big data is precisely the reason for the importance of and research activity in this area. In this paper, we also summarize various applications of big data in chemistry. Professional information platforms like the Collaboratory for Multi-scale Chemical Sciences (CMCS) and Chemical Informatics and Cyberinfrastructure Collaboratory (CICC) have been developed to manage and research chemical big data, while search engines like the ChemDB Portal have been established to extract chemical information from the internet. Software like the Integrated Project View and ArQiologist can be used to assist in the design of new medicines in medicinal chemistry. A data management system called BioGames has been proposed to analyze microfluidics big data. Moreover, graphics processing units are widely used to improve the computational capabilities of molecular dynamics simulations, while compressed score plots have been proposed to solve visualization issues in the field of chemometrics. In the era of big data, the analytical instruments, chemical data systems, and even the research methods may need to be changed and therefore, new strategies and techniques are still needed for the generation and processing of big data.

big data, data mining, visualization, cloud computing, chemistry

doi: 10.1360/N972014-01106