中国科学: 信息科学 2014 年 第 44 卷 第 5 期: 676–692
www.scichina.com info.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

论 坛

大数据时代对建模仿真的挑战与思考
—— 中国科协第 81 期新观点新学说学术沙龙综述

胡晓峰*, 贺筱媛, 徐旭林

国防大学信息作战与指挥训练教研部, 北京 100091
* 通信作者. E-mail: xfhu@vip.sina.com

收稿日期: 2014–01–09; 接受日期: 2014–02–27
国家自然科学基金 (批准号: 61174156, 61174035, 61374179, 61273189) 资助项目

摘要 2013 年 9 月中国系统仿真学会承办了中国科协第 81 期新观点新学说学术沙龙, 沙龙主题为 "大数据时代对建模仿真的挑战与思考", 重点对 "以大数据为基础的第四范式是否成立? 大数据方法对仿真建模带来了什么挑战? 大数据方法对仿真建模带来了什么机遇?" 等三个议题进行了研讨, 本文对与会专家的主要观点和研讨取得的主要成果进行综述.

关键词 大数据 第四范式 建模 仿真 复杂系统

1 引言

近两年来, "大数据" 一词被广泛提及, 甚至有被滥用的嫌疑. 大数据究竟会对科学研究带来哪些影响, 尤其是对与之密切相关的建模仿真领域带来哪些困惑、挑战和机遇, 很值得我们认真加以研究探讨.

为此, 2013 年 9 月 14 日至 15 日, 由中国科协主办、中国系统仿真学会承办的 "大数据时代对建模仿真的挑战与思考" 新观点新学说沙龙 (简称 "双新沙龙") 在吉林延吉召开. 由中国工程院李伯虎院士和国防大学胡晓峰教授担任本次沙龙的领衔科学家, 与来自全国近 20 余家科研院所的专家学者一起, 从 "产、学、研、用" 等多个方面, 围绕 "大数据时代对建模仿真的挑战与思考" 这一主题, 就 "以大数据为基础的第四范式是否成立? 大数据方法对仿真建模带来了什么挑战? 大数据方法对仿真建模带来了什么机遇?" 等三个重点议题进行了深入地探讨. 本文将介绍与会专家的主要观点, 对研讨取得的主要成果进行综述.

2 问题的背景

2.1 大数据及产生背景

什么是大数据? 大数据的主要特性是什么? 这些问题至今还没有准确、统一的定义. 通常认为, 大数据具有 "4V" 特性: 一是规模性 (Volume), 即体量大, 数据量级可达 TB, PB 乃至 EB 以上; 二是多

引用格式: 胡晓峰, 贺筱媛, 徐旭林. 大数据时代对建模仿真的挑战与思考 —— 中国科协第 81 期新观点新学说学术沙龙综述. 中国科学: 信息科学, 2014, 44: 676–692, doi: 10.1360/N112014-00013

样性 (Variety), 信息的种类多、异构, 可以多种信息载体形式存在; 三是高速性 (Velocity), 是高速率的流数据, 要求处理速度在合理时间之内; 四是价值性 (Value), 或称为真实性 (Veracity), 即大数据往往含有噪声, 具有高价值低密度的特点, 或指数据包含的价值具有真实性 [1,2]. 此外, 维基 (Wiki) 百科从处理方法角度也给出了大数据的定义, 认为大数据是指常规软件工具去捕获、管理和处理数据所耗时间超过可容忍时间限度的数据集.

大数据的产生主要有以下几个方面的原因: 一是信息技术的发展创造了数据产生和处理条件. 计算环境演进使数据量高速增长, 云计算使数据存储和处理能力不断得到增强, 网络、存储设施、数据库等技术的发展, 以及目前已逐步得到广泛应用的物联网、RFID 技术、视频监控等技术的普及与应用, 为人类从大数据中筛选信息、洞察世界提供了新的可能. 二是互联网运用的广泛普及, 带来了大量的数据, 包括社交网络、博客、微信、基于位置服务、搜索服务等等. 有专家称, 近两年产生的数据等于 2010 年前人类产生数据的总和. 三是各类大数据应用产生了很好的效果并提出了更高的要求. 对数据的深度挖掘所获得的出人意料的效果, 已超越了早期以 “啤酒与尿布” 等经典案例为代表的数据挖掘, 出现了得到各界广泛关注的 “纸牌屋”、“点球成金” 等新的传奇.

2.2 大数据的影响及反思

大数据现已得到各界的广泛关注. 奥巴马政府 2012 年 3 月发布的 “大数据研究与发展倡议”, 将其作为美国未来发展的重要战略, 启动了 “大数据发展计划”, 奥巴马政府冀望于通过该计划的实施, 重蹈 “信息高速公路计划” 带来互联网霸权的覆辙, 再次获得信息技术领域广泛优势. 此次斯诺登事件的曝光, 说明了美国的互联网霸权能在其他国家浑然不觉中就将其置于非常危险的境地, 而现在美国又将目光瞄准大数据这一新的未来领域, 正是在为创造未来的大数据霸权奠定的基础.

大数据在经济领域也引起了格外关注. 在 2012 年 1 月举行的达沃斯经济论坛上, 专门以 “大数据、大影响, 全球开发的新可能” 为主题发表了大数据报告, 受到各国首脑和企业家的普遍关注.

大数据问题也引发了学术界的普遍关注. 2008 年, 英国《自然》杂志推出大数据专刊, 专门探讨 “PB 时代的科学” 以及科研形态的变化, 指出: “数据为准绳的理念指导, 以及强大的计算能力支撑, 正在驱动一次科学研究方法论的革命”. 美国《科学》杂志也在 2011 年推出专刊 “Dealing with Data”, 围绕 “数据洪流” 展开讨论, 将大数据深度分析作为未来研究的重要突破点. 此外, 各类学术机构也纷纷组织各种研究和探讨, 发表研究报告、召开各种会议、成立大数据组织等等, 相关学术研究正如火如荼的展开.

但是, 在大数据得到各界热棒的同时, 应冷静思考并回答以下 3 个方面的疑问: 一是大数据与以前一些数据概念有哪些不同? 例如, 大数据与早期提出的海量数据 (massive data)、超大规模数据 (very large data) 等有何不同? 二是大数据方法与过去的数据方法有什么差异? 三是大数据应用与过去基于数据分析的应用又有什么不同? 例如, 与过去的商业智能 (business intelligence, BI) 等一类基于数据分析的应用有何不同?

2.3 大数据的主要特征

大数据带来了全新的研究思维和方式, 其革命性特征主要有以下 4 个方面.

特征一: 从局部到全体, 将网络化的大数据作为分析对象.

首先, 大数据是直接面向全体的、网络化的数据分析, 其中, “数据大” 是关键, 不象过去的数据分析, 只是对少量样本数据进行分析, 而直接面向整体数据, 或者叫做所有数据, 甚至是全部的数据分析;

“网络化” 是核心, 它本质上终结了还原论的分解方法, 需从整体关系考虑. 其次, 大数据在规模、类型、模式、工具、对象等方面, 都与传统的数据库和分析方法有所不同. 一是大数据将 “局部和明确的数据” 转化为 “所有几乎全部且不明确的数据”. 有专家比喻说, 过去的数据处理象是在池塘里抓鱼, 对池塘里投放了多少鱼、能收获多少是心中有数的, 而现在大数据处理就象在 “大海里捞鱼”, 有鱼与否、能捞到什么鱼都不知道. 二是大数据变 “脱机” 处理为 “联网” 处理, 因与网络的关系极为密切, 处理的同时数据还可能变化.

特征二: 从单纯到繁杂, 接受数据的繁杂和不精确.

大数据以非结构化、种类繁多的数据为主, 抛弃了对有条理和纯净数据的偏爱, 容忍凌乱数据; 大数据不以 “匹配性查找、增删改管理” 为数据库应用目标, “海量”、“超大规模” 都只是数量概念, 不说明其他特征. 大数据的不确定性和涌现性特点比较突出. 它的数据来源不确定、处理模型不确定、模型参数学习也不确定, 它能体现演化模式的涌现、群体行为的涌现、甚至网络智慧的涌现.

特征三: 从因果到关联, 更强调相关性而非因果性.

大数据方法最重要的思想是放弃对事情原委的追究, 而代之以对相关性的接纳, 因此, 它更适合于回答 “是什么”, 而不是回答 “为什么”, 这就为 “知其然而不知其所以然” 的研究找到了依据, 即直接从大数据中获取答案.

之所以这样做, 是因为许多事物的因果关系难以明确, 或找不到, 或根本不存在, 大数据方法认为, 海量数据的相互关系已经足以产生新的发现, 这可能是对牛顿、爱因斯坦体系下因果关系明确的还原论思想的一种完全颠覆. 正因为如此, 美国、欧盟展开了 20 余项研究计划, 如 “大脑扫描计划”、“星球皮肤计划”、“太空追踪计划” 等等, 都或多或少基于这一思想.

特征四: 从简单到深入, 更强调深度分析和间接分析.

大数据将已有的简单分析方法发展为深度分析方法. 简单数据分析方法是指对已有数据的分析, 如: 商业智能对因果关系的分析. 但大数据具有深度分析、直接分析、外推分析等特色, 可提供更多更好的数据分析功能, 并由数据量决定分析结果. 如: 苹果公司的智能语音助手 SIRI, 可基于联网数据实现数据学习功能; 外延分析可获得超出分析初衷的结果, 如: 基于搜索词的流感趋势预测; 按需分析可先有意识地产生所需数据再进行分析, 如数据客等.

3 以大数据为基础的第四范式是否成立?

3.1 科学研究的范式

科学研究的范式 (Paradigm) 概念是 Thomas Samuel Kuhn 在 1959 年《科学革命的结构》一书中首先提出的. 范式是指那些在一段时间内为科学家集团所共同接受的科学信念, 是一组假说、理论、准则和方法的总和, 用于指导现实科学研究. 一旦范式无法指导新的研究就会发生危机, 产生出新的科学成就, 这就是科学革命. 科学革命的结果就是一种新范式的诞生, 这就是 “范式转换”.

目前已存在并得到公认的科学范式包括 [3]: 第一范式是科学实验, 通过观测、记录、验证得出发现, 如: 伽利略斜塔坠球实验、天文观测等; 第二范式是理论推导, 通过逻辑推导、数学证明得出发现, 如爱因斯坦的相对论; 第三范式是科学计算 (包括建模仿真), 通过科学计算和模型仿真得出发现.

第四范式是由微软公司 James Gray 最早提出的, 他认为, 数据探索性研究方式, 即基于数据密集型的科学发现, 是未来一个非常重要的趋势. 这些科学研究从以数学模型计算为中心的方式, 转变为对海量数据处理为中心的方式. 在数据达到一定规模之后, 科学研究模式也会发生从 “量变” 到 “质

变” 的根本性转变, 这就是一种新的范式的诞生. 因此, 它可以独立于基于数学模型的科研形式, 单独成为一种新的科研范式 [4].

3.2 对第四范式是否成立的主要观点

大数据作为第四范式是否成立? 以大数据为基础的第四范式所产生的实质性变化确实需要 “范式转换”, 还是对第三范式的一种扩充, 或仅仅是一种特殊形式? 如果成立, 它应该包括哪些内容? 围绕上述问题, 与会专家展开了激烈的争论, 主要有以下 3 种观点.

3.2.1 第四范式成立

在传统建模仿真研究中, 数据只是为模型的仿真运行试验提供的基础条件, 如果说模型是 “引擎”, 数据则是 “汽油”, 数据是模型最重要的组成部分. 而现在数据可以成为发现的主体, 且数据的来源可以多种多样, 可以通过仪器采集、网络收集、仿真系统生成等方式获取数据, 之后数据就可以脱离模型成为科学发现的主体. 只要数据足够大, 只靠数据就可以完成科学发现, 因此不再需要数学模型. 这就是所谓的 “数据优先” 模式.

《连线》主编 Chris Anderson 就曾断言: “数据的洪流使传统科学方法变得过时”, “相互关系已经足够, 没有了具有一致性的模型、统一的理论和任何机械式的说明, 科学也可以进步”. 也就是说, 建模方法对于科学而言并不是必须的, 大数据方法就是一种新的科研范式.

针对上述观点, 赞同第四范式的与会专家认为, 大数据开辟了机器学习和智能科学研究的新途径, 通过大数据挖掘、基于多源数据认知分析, 将促进了认知分析学发展、扩展智能化应用, 彻底改变人们科研、学习、生活、工作模式. 大数据将对认知理念和研究方式产生革命性影响, 引发科学研究和思维方式的大变革, 颠覆前三种范式的研究模式和思维理念. 相应的, 基于网络科学的数据科学研究将迅速崛起, 数据价值挖掘和利用将会促进新兴科学发展. 因此, 以大数据为基础的第四范式将成为一种全新的科学研究体系, 并逐步形成理论、方法、技术的完整研究体系.

还有专家对大数据带来的革新性变化进行了总结. 一是研究对象具有革新性. 大数据最具代表性的研究对象应是具有 “人、机、物” 三者融合特点的. 大数据是在信息技术高速发展之后出现的, 信息技术 (information technology, IT) 的普及和发展为它提供了采集、存储、传输、处理等技术条件, 使对信息技术的关注, 从过去重点对 “T”(技术) 上升到 “I” (信息), 即通过 “I” 体现信息技术应用的最终成果, 反映人类世界和物理世界因信息技术发展而产生的逐步融合, “人、机、物” 三者中的这个 “机”, 正是实现三位一体融合的核心. 此外, 大数据本质上是网络化数据, 通过数据的多重的、网络化的关联性分析, 可以发现传统研究模式难以发现的价值和知识, 关联物不同, 所发现的知识可能也不同. 二是研究模式和思路具有革新性. 在以往的科学研究模式中, 通常是采用 “观察 – 假设 – 实验 – 应用” 的流程, 但大数据的出现完全颠覆了这一模式. 与传统研究模式不同, 大数据所采集的海量数据是否有用、价值多大, 提前是未知的, 即使是有预设研究目标的数据采集, 其研究模式和理念仍是从海量数据、稀疏价值中发现、挖掘 “未知” 的知识, 这正是拥有垄断性大数据资源的谷歌、IBM、微软、亚马逊、Facebook 等大企业热衷大数据研究的真正动力. 可以用 “光场相机” 来比喻大数据方法的创新性特点. 传统相机只能捕捉一束光, 拍摄前需要先对焦, 才能捕捉到清晰的画面. 而 “光场相机” 可以记录下整个光场里面的所有数据, 这样就可以按后续需求的不同, 清晰获得光场中对应所有焦点的清晰画面, 即从任意角度重现拍摄物. 这一主要区别为大数据后续的整体性研究、意外知识的发现提供了可能.

支持第四范式的专家认为, 以大数据为基础的第四范式与基于模型计算的第三范式相比, 主要体现出以下 6 个方面的特点: (1) 无需事先给出假设和建立系统模型; (2) 提出了具有网络化、智能化及整体、协同、关联、开放等特征的新的认识和改造系统的统计分析方法和技术; (3) 充分发挥了网络科学技术、计算机科学技术及其工具的作用; (4) 大数据方法与技术已在生命、环境、社会、科学等多个领域取得了显著成效; (5)"第四范式" 不仅是科研方式的转变, 也是人们思维方式的大变化, 即数据不再仅仅是科学研究的结果, 而且变成科学研究的灵活的基础; (6) 不仅要关心数据建模、描述、组织、保存、访问、分析、复用和建立科学数据基础设施, 更要关心如何利用泛在网络及其内在的交互性、开放性、利用大数据的可知识化、可计算化, 构造基于数据的、开放协同的研究与创新模式. 此外, 有专家强调, 第四范式不可能完全取代前三种科研范式, 将与三种范式长期并存、互为补充.

大数据方法作为一种新的科研范式, 特别是将搜索方法发展成为一种科学研究方式, 这将可能使科学研究产生 3 个重大的变化: 第一, 将一般科研活动中 "精心设计并提出问题" 的研究环节, 变为关键词的选择, 因而不再需要假设; 第二, 摆脱试验的束缚, 基于相关性理论在海量观测数据或现实镜像世界中去寻找关联, 即使有个别的模型, 也不影响其整体理论存在; 第三, 不再对研究结果进行解释, 出现了能够预测但不解释的科学. 有人说这是出现了第五种科研方法, 即在亚里士多德提出的逻辑方法、培根提出的试验方法、牛顿提出的数学方法、费米提出的模拟方法这四种原有的科研方法之外, 新增了谷歌提出的关联方法.

3.2.2 第四范式不成立

支持这一观点的专家认为, 范式的演变表示科学研究的一套方法及观念被另一套方法及观念所取代. 但是, 目前的大数据理论、方法与技术还处于早期的思想萌芽状态, 还缺乏系统的理论体系、技术方法和应用实践等支撑, 因此, 还达不到成为一种独立科研范式的程度, 单独将大数据分析独立出来, 并不能形成独立的科学研究模式, 更不能单独进行科学发现, 现阶段以大数据为基础的第四范式尚不能证实成立. 其一, 大数据需要获得大量的数据, 这些数据要么从现实中采集, 要么利用仿真获得, 它们都是科研过程中一个不可缺少的组成部分, 不能割裂, 只是分析方法更丰富了一些而已. 其二, 任何分析都需要模型, 没有数学模型是不可能进行分析的. 即使谷歌搜索, 也用到了各种搜索算法和匹配排序模型, 使用数据都需要初筛, 初筛就要用到模型和假设. 其三, 目前大数据研究的主要方法似乎尚未突破传统技术的范畴, 虽然大数据出现后发生了一些变化, 如过去是统计推理, 现在是推理统计, 的确是两种不同的思路, 但是否还能进一步发展形成新的技术方法体系目前尚未可知.

还有部分专家对基于大数据的第四范式持质疑态度, 理由主要包括以下几个方面: 一是大数据作为一种新的科研范式是否显得太粗糙? 新的发现依靠重复性计算, 数据量的增加虽提高了研究的质量, 但不会引起科学研究方法的本质变化. 对科学结果只预测不解释, "不解释" 只能说明认识活动还没有完成. 大数据作为第四范式是否预示着科技的 "去人类化"? 科学的神圣任务是 "揭示隐藏在混沌世界中的有序结构", 如果可以只靠网络和计算机来完成, 而不是科学家的敏锐思想, 是否意味着改变了科学家在科研活动中的角色, 也改变了科研活动的基本规律? 二是大数据能否构建具有创新性的研究体系? 在基础理论层面, 对大数据复杂性解析、计算模型等的研究尚不成熟, 尚不能总结出大家公认的理论和方法; 在核心方法与技术层面, 目前大数据研究所采用的主要方法, 在大数据概念提出之前就已经存在, 如统计方法、数据挖掘、分布式大规模批量处理、非结构化数据处理等. 但还有许多技术方法问题需要进一步研究, 包括多源异构大数据感知、融合与表示, 大数据内容建模与语义理解, 感知、存储与计算机融合的大数据架构体系等; 在应用层面, 尚需加强研究大数据的处理和软硬一体化引擎系统等. 三是大数据的研究成果能否具有普适性? 科学本身的定义是指发现、积累公认的普遍真理或普

遍真理的运用, 是已系统化和公式化的知识. 第一范式和第二范式是大家公认的科学发现的范式, 当今世界主要的普适真理和规律都是通过这两种范式发现的. 但大数据目前更多是停留在技术层面, 或是对现象解释的层面, 还只能解释部分小众世界正在发生的情况和事实, 属于认识和改造世界的范畴, 但能否用于科学发现、能否产生新的知识、是不是普适真理或普遍真理等, 尚存在很大争议. 四是大数据注重相关性研究, 能否解决因果问题尚未可知. 科学研究的目的之一就是要发现因果关系, 是因果关系的研究促进了科学体系的建立. 但从大数据的发展现状来看, 主要应用体现在企业界和工业界, 这类应用可以只要现象不要理由, 但对飞机的控制和病人的救治等, 是否可以依赖不精确、不知因果的数据并做出相应反馈? 如果不注重因果关系的研究, 是否将对整个科学的基础进行重新定义?

3.2.3 第四范式尚不成熟

除了上述两种截然相反的观点外, 还有一些专家对第四范式的观点介于两者之间. 他们认为, 数据密集型方法直接面向数据, 基于数据及其关联统计分析, 可从整体上发现复杂系统的关联规律, 能否成为一种新的科研范式有待进一步检验. 其一, 对大数据的搜索、比较、聚类、分类等分析归纳得到的相关性无法检验因果性. 因果关系的研究引发了近代科学体系的建立, 科学研究的最终目标是发现、确定研究对象的因果关系即规律. 因为基于数据的关联性, 拥抱关联性不顾因果性, 不能像实验那样明确地告知 “是什么”, 也不能像理论和仿真那样在一定程度上告知 “为什么”, 只能告知 “大概是什么”, 是一种 “知其然而不知其所以然” 的实现, 故不能说是找到了规律. 相关性在某些情况下可以包含因果性的, 但并不见得全部包含因果关系, 因果性并不是相关性的子集, 所以它无法检验因果性. 其二, 基于统计数据分析结论, 只能具有试验的效果. “科学范式” 要有普适性, 没有科学假设和模型所发现的新知识究竟有多大的普适性需要实践来检验. 任何一次统计分析结论仅具有实例效果, 基于统计学的数据挖掘只有针对某一领域并依赖于对象涉及的特殊条件和专门知识才会有效. 其三, 第四范式只是 “基于客观数据” 的模式识别, 是有限的客观性. 通常认为, 大数据的处理是客观的, 让计算机从海量的数据关联性中发现规律、知识, 较之主观的模式识别 (根据人在经验中发现的规律, 提出一个主观的假设, 再去搜集更多案例来验证这个假设) 更具有客观性. 但是, 海量数据噪音大, 数据噪声需要由人来处理, 特别是对非结构化数据, 必须首先要识别数据 (基于语料库) 才能进行挖掘. 也就是说, 第四范式只是 “让不对任何东西敏感的计算机, 人类使用有限种方法定义是否 ‘敏感’ 以发现新的东西”. 其四, 第四范式的理论基础尚需进一步明确. 大数据来自各种复杂系统的观察或记录, 是实际对象的信息映射, 作为客观事物间接存在形式的 “数据界” 一旦脱离 “物理世界” 还有其自身的共性规律吗? 大数据往往以复杂关联的数据网络这样一种独特的形式存在, 网络科学是大数据科学问题的本质吗?

支持这一观点的专家还指出, 从科学研究角度, 范式至少具备以下两个特点: 第一, 它是从事某一学科研究的群体共同遵守的世界观和行为方式, 是这个学科赖以存在的理论基础和实践规范; 第二, 它可以为科学研究提供可模仿的成功先例, 即具有推而广之的意义. 而大数据目前已初步具备了作为一种新范式的特点, 提供了一种认识和改造世界的新方法、新途径, 但将其与其它范式并列成为第四范式的理由还不充分, 能否作为独立的科研范式值得商榷. 首先, 大数据在实践规范上和研究方式上的确不同于前三种范式. 它不象第一范式那样需要精心设计试验过程, 依赖精确的数据观测和测量处理, 也不像第二范式基于一种公理体系来进行推导, 更不像第三范式那样需要预先建模. 其次, 大数据的部分方法已具有了一定的可推广性, 如 IBM 发现, 对电网的研究方法可以推广到供水、交通等相关网络上. 再次, 目前大数据更多用于解释或研究复杂系统应用问题. 从科学研究的角度来说, 大数据在研究结果的普适性、对因果关系的研究等方面, 与第一范式和第二范式还不在同一个层面上.

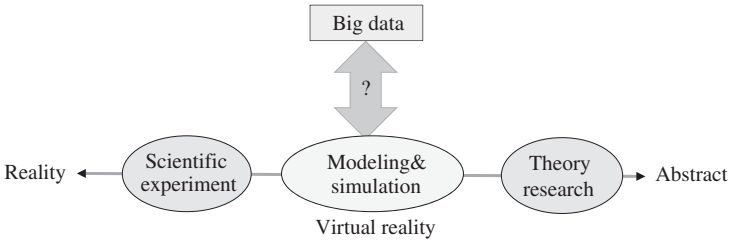

图 1　以大数据为基础的第四范式与其他范式的关系

Figure 1　The relationship of the fourth paradigm basing on big data and other paradigms

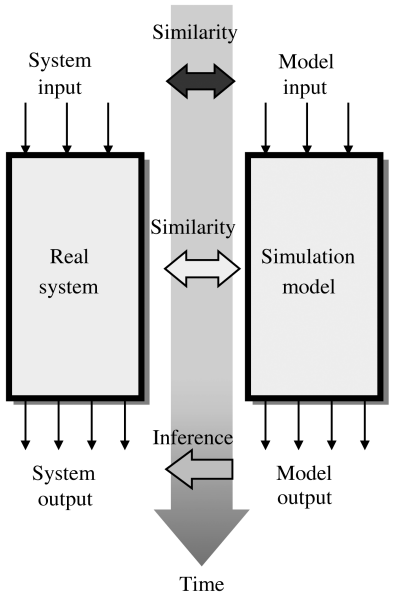

图 2　基于相似性理论的仿真

Figure 2　The simulation basing on the similarity theory

4　大数据对建模仿真带来哪些挑战?

4.1　挑战何来?

大数据为基础的第四范式是从第三范式 (科学计算、建模仿真) 中独立出去的 (见图 1). 如果从一个坐标轴来看, 以现实观测为主要形式的科学实验范式占据着坐标轴的一端, 完全抽象的理论研究占据着坐标轴的另一端, 而建模仿真则以 “虚拟现实” 的方式, 正好居于这个轴的中间.

谷歌研究部主任 Peter Norvig 曾经预言: “所有的模型都是错误的, 进一步说, 没有模型也可以成功”. 《复杂》杂志说过: “量子力学和混沌摧垮了精确预测的希望, 哥德尔和图灵的结果摧垮了数学和计算无所不在的希望”. 那么, 大数据出现后将会对传统建模仿真学科带来哪些影响和挑战? 是否会动摇或变革原有仿真理论的基础? 是否会摧毁我们原来的一些观点呢?

4.1.1　对仿真基本理论的挑战

以往的系统仿真是建立在相似性理论基础之上 (见图 2), 通过对实际系统进行建模, 使两者之间具有相似性. 这样, 如果输入相似, 则认为输出也应该相似, 从而达成仿真的目的, 用相似性中的类比

方法来获取结果, 是仿真科学最基础的观点.

仿真的目的就是发现问题和预测未来, 通过时间轴前推, 仿真能实现 “发现问题或预测未来” 的目的. 但是, 大数据的 “数据优先” 模式在某些情况下却可以做得更好, 例如在预测方面. 有报道称, 依据相关性, 通过外延效应、间接预测等方法, 谷歌对流感的预测与官方结果相关性高达 97%. 那么, 仿真可否基于相关理论进行? 基于相关理论可否部分取代基于相似理论的建模仿真? 这些问题将对重构建模仿真科学带来一系列的挑战.

众所周知, “仿真是基于模型的试验”, 而试验的目的是为了发现事物的规律, 达到认识世界的目的. 既然大数据无需模型也可以得到结果, 那么是否意味着仿真也可以利用搜索或统计来完成? 例如, “棱镜计划” 中使用的搜索和统计分析, 解决了用很多模型无法解决的问题; 采用生物地理信息学的方法, 从大量元数据中寻找可能的联系人, 来预测 “恐怖分子” 位置等. 这些建立在数据分析基础上的模型, 是通过对现实数据搜索来完成的, 并且试验结果越来越趋近于真实. 这种以搜索统计分析为基础的方式是否也是一种 “仿真”?

此外, 利用大数据为建模仿真服务, 还需要解决平台问题. 现有仿真平台能否满足大数据的要求, 如: 具有搜集到所有大数据的能力吗? 能否突破获取、隐私、安全等限制? 是否具备处理和存储 PB 乃至更大数据能力? 能够运行镜像模型吗? 能源是否是问题? 能够具备深度分析和挖掘的软件和方法吗? 等等.

4.1.2 对建模方法的挑战

有没有不要数学模型的仿真? 大数据提供了利用 “数据模型” 的新途径. 某些复杂的事物未必有可行的模型, 如复杂度非常高、计算量非常大、在可行性时间内做不到等等. 但在大数据时代, 针对这类 “可以描述但不能用模型方程解释” 的现象, 可以通过建立起认识问题的 “数据模型”, 如谷歌的关联研究等, 因而 “绕开理论 (不再建模), 直接获取答案” 就成为了一种新方法. 那么, 基于数据模型的仿真是否存在?

是否会产生出新类型的模型? 传统的数学模型仅是对问题某一侧面的描述, 这存在两方面的问题: 一是模型简化必然带来模型使用的风险. 没有一个模型能够模拟事物的全部特征, 正如建模理论先驱物理学家费利浦 • 安德森在诺贝尔获奖仪式上所说: “建模的艺术就是去除与问题无关的部分, 建模者和使用者都面临一定的风险. 建模者可能遗漏至关重要的因素, 使用者则有可能无视模型只是概略性的, 意在揭示某种可能, 而太过生硬的理解和使用实验及计算的具体结果样本”. 也就是说, 模型使用者和和建模者的理解可能是不一样的. 二是对于一些复杂的事物, 由于很多规律无法了解, 即使建立其某种模型, 也很难真正起到作用. 比如说火灾模型、经济模型、人群模型等, 能不能建立起这种复杂的模型, 建模对象是否根本就不存在可以用数学模型描述的规律? 或者至少现在我们的知识水平还无法描述这些模型?

但是, 大数据的出现对解决以上这两方面问题提供了新的可能, 可能会产生出一些新的类型的模型. 最具代表性的范例之一是镜像模型. 大数据可以利用真实世界镜像, 充当比较完美的现实缩微模型, 如以小社会代替大社会, 来研究情感变化、社会舆论、信息传播等巨量的人群行为. 那么, 这种不是 “以假代真”, 而是以 “小真” 代 “大真”, 是否算是一种新的模型? 同时, 利用镜像模型可以完成 “真实版” 的仿真, 如 Twitter, Facebook 的情绪分析可以提前 25 分钟预测股市涨跌; 利用不同事件的相关性, 可预测相关其他事件的发生, 可称其为 “纠缠相关”, 如: “奥巴马遇刺新闻乌龙与股市的涨跌”. 那么, 这种镜像模型能否称其为仿真? 第二个具有代表性的范例是 “嵌入式” 平行仿真. 基于大数据的 “嵌入式” 模型可以兼顾 “过程” 与 “结果”, 通过嵌入式平行仿真, 可一边仿真一边分析预测, 这种方

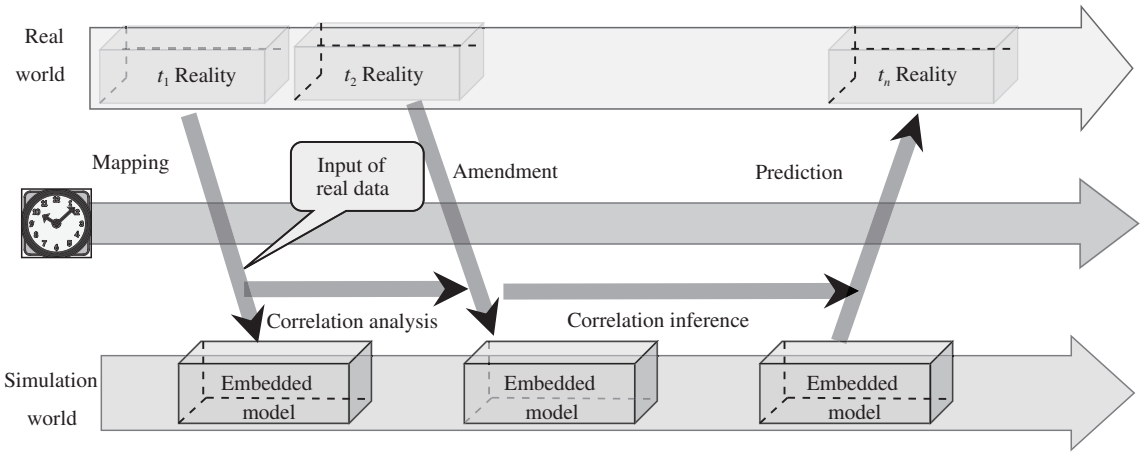

图 3 “半现实”的嵌入式平行仿真

Figure 3 The semi-real embedded parallel simulation

式可称之为 “半现实仿真”; 可以根据 “过去和未来” 的全面情况进行深入分析, 实现超前预测和处置, 即过去是现实的真实数据, 而未来是仿真出来的结果 (如图 3 所示), 通过模型建立两者之间的映射关系, 并利用数据进行修正、校验, 最后推演出未来的某一时段的过程结果. 那么, 这种基于 “半现实” 数据的嵌入式平行仿真是否也是一种新的仿真类型?

4.2 “大数据”成为重要研究对象

与会专家认为, 可以按数据来源将大数据分为两大类: 一类来自自然世界, 多半是科学实验数据或工程传感数据. 主要是大体量的、异构的结构化数据. 科学实验是科技人员设计的, 数据采集代价较高, 数据如何采集和处理事先已确定, 不管是检索还是模式识别, 都有一定的科学规律可循; 另一类来自互联网的人类社会活动数据, 有许多不同于自然科学数据的特点, 包括多源异构性、交互性、时效性、社会性、突发性和高噪声等, 不但非结构化数据多, 而且数据的实时性强, 大量数据都是随机动态产生的. 数据的采集成本较低, 价值密度低, 许多数据是重复的或者没有价值, 即具有 “4V” 特性.

有专家认为, 对建模仿真研究而言, “大数据” 问题早已存在, 只是其内涵具有相对性, 是相对于计算能力、数据来源、数据处理目的而言的, 并将随数据采集能力提高、计算机处理需求和能力的发展而不断变化. 变化的 “大数据” 内涵使建模仿真需要面对和处理的问题越来越多, 并对建模仿真带来了诸多挑战. 以对模型概念的挑战为例, 从最早的仿真控制开始, 仿真对象和目标一直在不断扩大, 但大数据将使仿真对象发生前所未有的扩展, 如: 大数据带来对图形图像抽象的仿真需求, 故需重新看待模型的内涵.

有专家从数据发展历史的角度分析了大数据演化带来的仿真数据管理模式的变革及技术挑战. 他们认为, 大数据的特征必然使数据管理发生本质性的变化. 大数据注重整体性, 数据量本身的增大仅仅是大数据的特征之一. 在整个大数据的处理的过程中, 要注重时效性和精确性以及完备性三者之间的关系, 在数据量增长和数据种类增加的过程中, 数据处理的时效性变化成了整个仿真系统生存关键性的问题, 甚至可能需要牺牲精确和完备性来追求时效性. 此外, 大数据具有密集性特点, 这既意味着可能得到整个世界的原貌, 也意味着在整个数据里可用数据的稀疏, 特别是在网络数据里, 稀疏性更明显. 以对论坛意见的挖掘为例, 期望收集到所有数据与处理其噪声数据成为影响整个分析结果的一对矛盾因素. 也就是说, 数据的密集和可用数据的稀疏之间的矛盾, 或将成为制约仿真大数据处理的

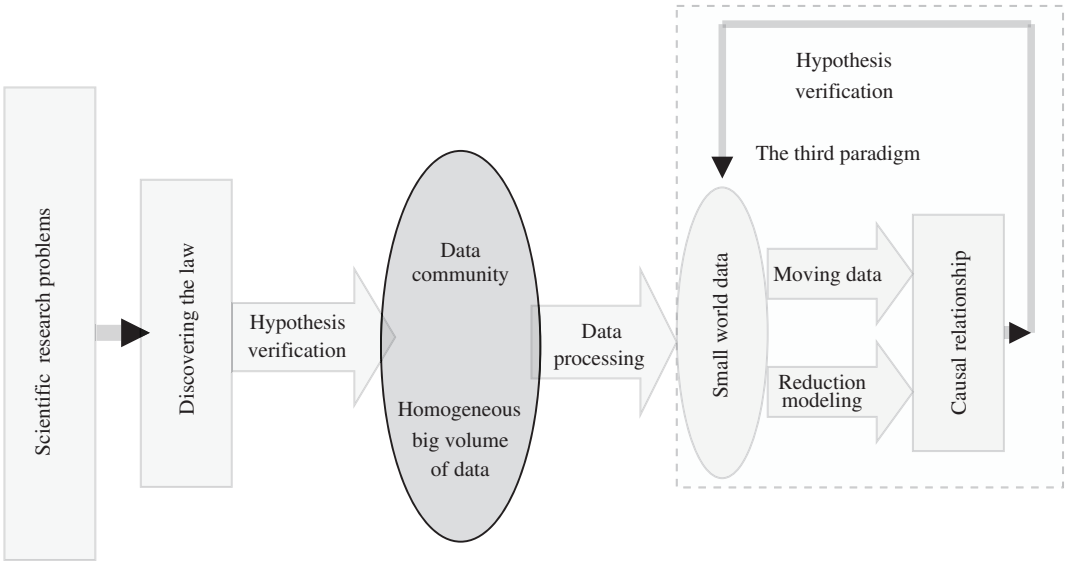

图 4 基于仿真范式的同构大体量数据处理模式

Figure 4 The homogeneous and big volume of data processing mode basing on simulation paradigm

重要问题.

4.3 从第三、第四范式的比较看挑战

仿真范式对数据的处理过程具有以下特点: 一是基于模型的, 以还原论为基础. 其核心理念在于 "世界由个体 (部分) 构成", 复杂对象由较低层次的简单构件组装而成, 对研究对象要不断进行分解, 化复杂为简单; 二是面向小世界, 是基于因果关系的. 小世界是指有边界、有确定性定义的系统, 基于小数据的因果分析是可行的, 如开普勒发现行星三大定律、牛顿发现力学三大定律等; 三是对目标、边界、实体、属性、状态、约束等进行了预定义, 是可重复、可复现的.

仿真范式中数据的处理过程见图 4 所示. 可见, 仿真范式是面对小世界数据的, 是一种 "移动数据" 模式: 通过基于还原论的建模来建立因果关系, 并依据从数据界中发现的规律, 作为假设或验证, 然后输送到第三范式, 对数据进行处理以后再进入小世界数据, 这一过程面向的主要是同构化的数据.

与之相比, 数据密集型科学范式具有以下特点: 一是面向大世界. 无需定义边界, 无需规定规模, 只受限于数据; 二是无需模型, 不受还原论约束, 是一种整体论的解决方法, 可不受时间、空间尺度影响, 由数据发现涌现性、演化机制, 可适应开放复杂大系统的研究要求; 三是将计算用于数据, 而非数据用于计算. 基于数据及其关联网络形成的数据界, 通过机器学习、数据挖掘, 发现这些节点和链接的关联, 从而获得整体的知识.

数据密集型科学范式中数据的处理过程见图 5 所示. 可见, 第四范式以决策支持为目标, 而不是为了发现因果规律, 是一种 "移动问题" 模式: 即把问题提交给大数据系统, 由其调动众多资源进行计算和统计分析, 才使众包模式和云计算这样大规模的研究成为可能, 如谷歌的 "流感趋势" 项目、地球引擎项目, 都是众包的模式, 大家共同做一个课题, 才有可能将计算用于数据, 而不是数据用于计算.

通过上述比较可知, 第三范式当前面对的主要挑战包含以下两个方面: 一是难以满足处理来自互联网的人类社会活动大数据的需求. 仿真本质上属于程序引导的密集计算, 模型的求解、分析都基于

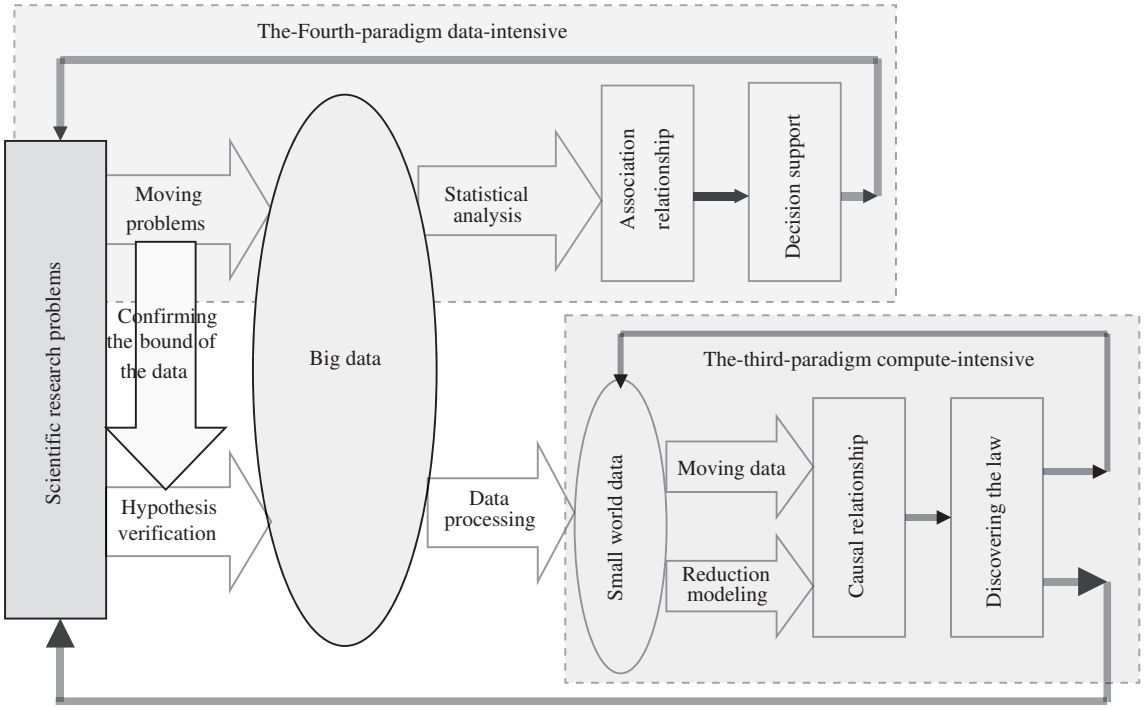

图 5　基于数据密集型科学范式的 “4V” 数据处理模式

Figure 5　The “4V” data processing mode basing on data-intensive paradigm

移动数据, 只能适应同构的大体量数据, 难以适应 4V 特征的大数据, 计算的时效性也只有在实时 (在线) 仿真条件下才能处理时效性问题, 即小世界的时效性, 对大世界来说时效性比较困难. 二是难以解决无组织的大世界问题. 无组织的大世界问题是指难以定义无边界、开放、变结构的系统, 难以处理不确定性、涌现性、非平稳随机过程等, 尤其是对于开放复杂的巨系统, 各部分之间可能互为因果, 现在的 “因” 可能是过去的 “果”, 此处的 “果” 可能是别处的 “因”, 相互纠缠, 且隐藏在整个系统之中.

因此, 在面对来自科学实验与工程的大数据时, 第三范式应采取 “以大化小”, 在还原论的指导下, 基于已知小世界的规律, 建立与运行模型, 以发现有组织的更大更复杂的系统的因果规律; 在面对来自互联网的人类社会活动数据时, 应运用数据密集方法, 基于统计分析等技术方法, 从整体上研究大世界的相关性.

面对这一挑战, 应采用集成和融合的方式来发展仿真范式. 在大数据时代需要将计算密集与数据密集两种模式融合起来, 实现密集计算与密集数据的集成, 实现 Bottom-up 与 Top-down 的集成, 以克服第三范式难以解决无组织大世界的困难, 实现无组织复杂系统因果规律的发现, 也使得数据密集型方法能满足科学研究的因果性、普适性及客观性的要求. 两种范式相融合的研究模式如图 6 所示.

4.4　仿真工程与科学面临的主要挑战

美国世界科技评估中心 (world technology evaluation center, WTEC) 曾发表研究报告认为, 许多国家已经实现了科学数据密集应用, 包括 (实时) 实验与观测数据和建模与仿真的集成, 以加速发现和工程问题的解决等, 典型应用涉及生命与医疗、粒子物理、天气预报、基因学、地震预报等多个领域. 与会专家认为, 面对来自科学与工程领域的大数据, 基于仿真的工程科学 (simulation-based engineering

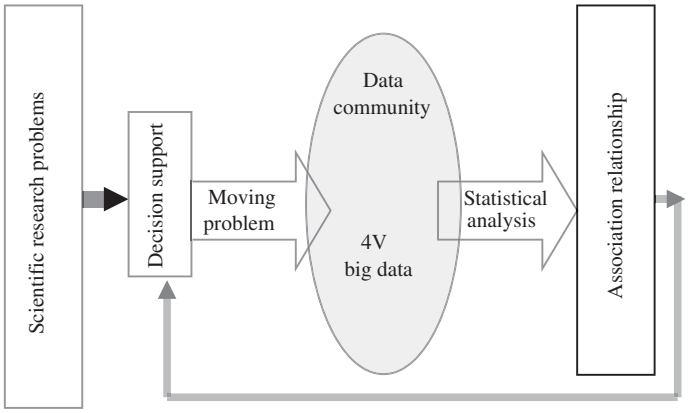

图 6 数据密集和计算密集相融合的科研过程

Figure 6 The scientific research process merging data-intensive paradigm and compute paradigm

and science, SBE&S) 面临以下 4 个方面的挑战 [5]: 一是高性能计算架构和算法. 超级计算机的计算速度已经超过了千万亿次, 而在未来十年甚至可能达到亿亿次 (exascale) 的量级, 但高性能计算的算法开发仍然非常滞后, 阻碍仿真技术发展的拦路虎将不会是硬件的性能, 而将主要是理论算法和软件. 二是多尺度建模和仿真. 仿真技术本身的发展正在从单尺度走向多尺度, 包括多学科、多尺度模型的融合、演化与集成技术, 总体上还处在发展阶段, 远未成熟. 三是科研范式转换的挑战. 科学研究不仅仅是以获取知识为目的的原始性发现和基本数据产出, 应由假设驱动转向基于探索的科学. 而 "科研智能" 如何能跟上 "感知能力", 需要考察大数据时代科学研究的未来, 探索支持科学研究的新范式. 四是面对来自互联网的社会活动大数据的分析处理. 新型的应用将致力于为实际的决策提供信息, 最终目的是帮助科学家、研究人员、决策者及社会大众做出有充分信息依据的决定.

大数据时代 SBE&S 的进一步发展还需应对以下 3 个方面的问题: 一是大数据技术本身尚不成熟, 需要结合 SBE&S 的要求开展探索性研究. 在大数据表示方法方面, 需要通过统一的数据格式构建融合人、机、物三元世界的统一信息系统; 在大数据去冗降噪方面, 大数据的多源性导致绝对的冗余, 同时数据采样算法的缺陷与设备故障会导致大数据的噪声; 在大数据管理方面, 传统的关系数据库技术无法胜任半结构化和非结构化数据的管理要求等. 二是需要发展仿真范式, 实现密集计算与密集数据的集成, 以实现无组织复杂系统因果规律的发现. 数据密集方法可由数据从整体上分阶段发现涌现性、演化机制下的结果, 而计算密集方法在部分时段或部分区域 (空间) 上满足了相似性 (行为特性、作用规则、演化流程等) 研究的需要, 为实现整体上的可预测性, 即通过模型运行来揭示相应复杂性系统的运行规律, 必须将数据密集与计算密集集成起来. 三是大数据环境下构建 SBE&S 环境的挑战. 包括大数据驱动下的建模与仿真、基于数据的建模/模型修改/模型校核与确认等技术.

此外, 也有专家特别强调了大数据对建模仿真方法和手段带来的挑战. 这些挑战包括: 各类复杂系统已经产生了大量具有 4V 特点的大数据, 但现有的建模方法还不能建立相应的系统模型, 关联和处理这些大数据; 现有的仿真支撑方法手段还不能适应对分布、异构复杂系统的大数据进行感知、采集、挖掘、处理、应用的需求; 现有的仿真应用工程技术对复杂系统所产生大数据的处理还不能全面、充分、及时地用于推动各行各业的发展等等. 因此, 需要变革建模仿真的方法和手段, 包括: 模型内涵及其建模 (一次、二次) 方法和技术, 仿真支撑技术及仿真硬、软件系统 (数学仿真、半实物仿真、人在回路仿真), 仿真应用工程技术 (结果处理、工作流程及 VV&A 技术等) 等, 以应对大数据发展带来

的一系列问题.

5 大数据对建模仿真带来哪些机遇?

有人说, 大数据的出现, 代表着一个以数据优先、数据为王的 “大数据时代” 的到来 [6]. 大数据时代是信息社会从 “量变” 走向 “质变” 的表征, 或者说, 信息化社会在大数据时代才算真正到来, 建模仿真也许会在这个门槛上发生根本性的变化, 甚至可能会重构仿真科学的体系、增强仿真科学的活力. 那么大数据时代会给建模仿真带来哪些机遇呢?

5.1 机遇何来?

有专家指出, 要把握这一划时代变革对仿真建模带来的机遇, 必须认真思考以下 3 个方面的问题.

第一, 能否为仿真结果分析提供更好手段? 传统的仿真结果分析大多是比较直接和简单, 而大数据可以提供更深入的分析和预先的处理. 例如: 在科学实验领域, 用于对粒子碰撞所产生的物理数据生成与分析, 在寻找希格斯玻色粒子 “万亿分之一” 的概率中取得重要进展; 在大规模仿真数据处理方面, 原有的一些仿真科学方法, 如: 数据分析、数据挖掘、数据耕耘方法等等, 与大数据的思路是一致的, 它既需要随时产生新的数据, 也是对仿真数据的一种筛选, 是两者不断迭代的过程, 而大数据的出现, 可以为解决这种大规模的仿真数据处理提供新的思路.

第二, 能否为复杂系统的建模仿真开辟新的出路? 复杂系统仿真一直就是难题, 采用传统的还原论方法很难做到. 一是非线性性质与不确定性结果带来的挑战, 复杂性既然导致因果关系不能确定, 那又如何建模? 二是系统动态结构对系统适应性建模带来挑战, 复杂系统结构总在不断演化中, 模型又如何适应这种变化? 三是涌现性仿真, 很多人认为涌现性很难得到, 目前常用的基于 Agent 的方法是合适的方法吗? 上述这些问题是造成社会仿真、经济仿真、战争仿真非常困难的主要原因.

但是, 大数据的出现为整体性分析提供了条件. 大数据的出现抛弃了对因果关系的追求, 这就避开了一个最难以解决的问题, 从而把重心放到了寻找相关关系上. 对基于传统科学的分解方法仍然解决不了的社会、经济、战争等复杂系统问题, 放弃还原论的分解建模研究, 代之以对 “整体数据” 的分析, 或承认对复杂事物无法建模, 转而直接从 “现实” 中去寻找问题答案, 这可能是复杂系统研究的新出路. 同时, 大数据可以将分解出来的各种碎片又重新组装成网络, 使得我们再次回到整体而不仅仅只关心局部.

与上述新思路密切相关的问题是: 谁来收集或产生这些 “大数据”? 是直接利用真实的镜像数据, 还是由大型仿真系统来产生这些数据? 针对这一问题, 有两个值得关注的案例. 第一个案例是欧盟拟投资 10 亿欧元的 “活地球” 模拟器项目, 试图对欧洲数据进行全面仿真和收集, 来分析和解决其面临的经济、交通、人口等复杂现实问题, 这一项目得到了几十位诺贝尔奖获得者的支持, 虽然最终胎死腹中未能得以实施, 但其创意却极具启发性. 第二个例子是建立在真人演习与仿真系统运行基础上, 收集获取全面的仿真数据, 不做过多筛选, 用于后续的深入研究. 这种方式和过去仿真数据收集方式的不同之处在于, 数据获取前不预设问题的 “某一侧面”, 而是全维产生并多角度分析数据, 这样更能反映出它的整体性, 而不是局部性, 局部性可能会切掉一些你认为无关紧要、但实际却可能是致命的东西.

第三, 智能仿真能否真的可以实现? 智能仿真是复杂系统仿真的一种, 但更具挑战性, 大数据方法未来或将为智能仿真带来曙光. 目前有两种智能仿真方式, 一种是以 IBM 的 “深蓝” 和 “更深的蓝” 为代表的方式, 它对人类逻辑和数学推理能力的仿真, 主要靠的是更精准的数学算法和棋谱数据, 并

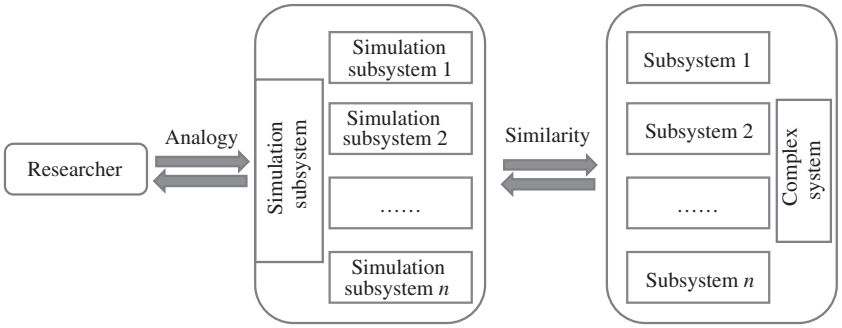

图 7 传统的复杂系统建模仿真研究思路

Figure 7 The conventional research method on modeling and simulating complex system

将每一步棋的可能性通过数学模型进行深度分析和计算, 最终找出最佳的方法战胜了人类; 第二种是以 IBM 的 "沃森" 问答系统为代表的方式, 是对人类认知的仿真, 它是从大量实际问题数据中学习形成知识体系, 通过对问题进行深入地相关性分析, 找到最有可能正确的答案. 显然, "沃森" 的模拟方式更接近人类常规的智能, 因为大量数据使得找到知识关联更为可行, 而非建立固定的因果模型.

总之, 大数据为我们提供了一个解释不明现象的新颖视觉, 带来许多值得思考和研究的问题. 大数据提供了一种绕开理论直接走向应用的新途径, 它是否真的挑战了 "观察、假设、实验、应用" 的科研流程, 是否真的找到了可以避开建模而直接获得答案的方法? 大数据带来了许多值得研究的科学新问题, 如: 预测问题时模型是必须的吗? 仿真是可以替代的吗? "模型优先" 与 "数据优先" 两者异同或矛盾在什么地方? 等等.

5.2 复杂系统建模仿真的新方法

与会专家一致认为, 复杂系统建模与仿真已经成为研究各类复杂系统的最佳手段之一 [7], 大数据将为复杂系统建模仿真提供新的出路. 在以往的复杂系统建模仿真中, 研究重点主要集中在复杂适应系统、非适应系统 (如元胞自动机)、标度、自相似、复杂性的度量等方面, 研究领域主要集中于生命系统、大脑神经系统和社会经济系统等 [8]. 其中, 复杂适应系统 (complex adaptive system, CAS) 建模方法的核心是通过在局部细节模型与全局模型间的循环反馈和校正, 来研究局部细节变化如何涌现出整体的全局行为, 其模型组成一般是基于大量参数的适应性主体, 其主要手段和思路是正反馈和适应, 认为环境是演化的, 主体应主动从环境中学习. 正是基于 CAS 理论所具备的这一独具特色的新功能, 为模拟生态、社会、经济、管理、军事等复杂系统提供了巨大的潜力. 元胞自动机 (cellular automata) 是一类模型的总称, 或者说是一个方法框架, 其特点是时间、空间、状态都离散, 每个变量只取有限多个状态, 且其状态改变的规则在时间和空间上都是局部的. 元胞自动机的构建没有固定的数学公式, 构成方式繁杂, 变种很多, 行为复杂. 复杂系统建模仿真目前常用的方法 [9] 有: 基于智能技术的复杂系统建模与仿真, 如遗传算法, 神经网络等方法; 基于数学手段的复杂系统仿真方法, 如参数优化方法、模糊仿真方法等; 基于离散事件动态系统的复杂系统建模与仿真, 如 Petri 网、任务/资源图建模等等. 目前复杂系统建模仿真的研究思路如图 7 所示, 通常采用类比方法, 即基于相似性理论, 将复杂系统化解成多个简单的系统, 先进行子系统的构建, 再形成一个大系统. 但是, 这种通过局部子系统仿真还原全局系统、逐渐逼近原系统的传统建模方式, 目前已遇到了严重瓶颈, 已难以满足当前复杂系统仿真研究的需求.

而大数据揭示了基于关联数据网络的共性问题, 大数据往往以复杂关联的数据网络这样一种独特形式存在, 它观察各种复杂系统得到的大数据, 虽然直接反映的往往是一个个孤立的数据和分散的连接, 但这些反映相互关系的连接整合起来就是一个网络, 这一关系网络就是基于关联数据的网络, 它的参数和性质 (如: 平均路径长度、度分布、聚集系数、核数和介数等等) 就能刻画大数据背后的网络共性, 要理解这些大数据就要对其后的关系网络进行深入分析 [10]. 故大数据面临的科学问题本质上可能就是网络问题, 复杂网络数据分析将成为数据科学的重要基石.

大数据必将为复杂系统仿真建模带来新的机遇. 从哲学层面来看, 认识世界的方法论有还原论和整体论. 还原论是把复杂的事物简单化, 用简单的运动规律来代替高效运动的规律; 将世界万物不断分解到最小单位, 是通过解构系统还原给人们单个节点和链接的理论. 但随着系统复杂度的增强, 这一传统的仿真建模方法已不能从本质上正确认识复杂系统. 而基于的整体论网络理论则反其道而行之, 通过组装这些节点和链接, 来帮助我们重新看到整体. 基于大数据对复杂系统进行整体性的研究, 也许将为研究复杂系统提供新的途径. 从这种意义上看, "网络数据科学" 是从整体上研究复杂系统的一门科学, 两者结合将使仿真建模方法更能胜任于复杂系统研究. 从逻辑推理层面来看. 原有的认识客观世界的方法有基于实验的归纳推理、基于理论的演绎推理、基于仿真的类比推理, 传统仿真建模中采用的是归纳和演绎推理, 而大数据扩展了用于认识客观世界的、具有更广泛意义及新意的逻辑推理方法 —— 合情推理. 合情推理是根据已有的事实和正确的结论、实验和实践的结果以及经验和直觉等, 推测某些结果的推理过程, 它离不开人类的知识、试验和实践的结果, 试验是我们有目的的设计, 实践是我们不一定有目的的设计产生的大量数据, 以及人类提炼出来的经验. 将大数据融入仿真建模, 所带来的第四种逻辑推理方法 —— 合情推理, 将会更胜任于复杂系统的仿真.

5.3 建模仿真科学迎来历史性发展机遇

将大数据方法与仿真建模方法相融合, 将为仿真科学技术的发展与应用带来崭新的历史机遇. 主要表现在以下 4 个方面: 一是将革新现有仿真科学的思维方式和科研模式. 包括要建立从大数据中获取知识的理念、进一步实现还原论和整体论的融合、引入合情推理方式和数据智能方式等. 二是将革新现有的建模方法学. 例如从现有的机理与非机理建模方法, 拓展到基于大数据的建模方法. 三是将革新现有的仿真支撑体系. 包括建立基于泛在网络的、面向服务的、高效处理大数据的一体化、智慧化云仿真系统架构; 在现有的仿真算法、软件、硬件、系统中融入大数据的高速并行处理软件框架, 如 Map Reduce/Hadoop 技术, 网络数据采集、多维数据预处理、数据流处理等大数据预处理方法; 大数据文件存储、No SQL 数据库等大数据管理技术, Hive 和 Mahout 海量数据挖掘技术等; 引入大数据技术, 重构甚至替代现有半实物仿真系统和人在回路仿真系统, 构成新型的人/机/物融合的仿真系统等等. 四是将革新现有的仿真应用工程技术. 如研究基于大数据技术的 VV&A 技术; 研究融合大数据技术的智能化仿真结果处理系统; 研究引入大数据技术的智能可视化系统; 研究基于大数据技术的嵌入式仿真系统; 研究有效处理大数据的仿真应用组织与管理模式等等.

此外, 也有专家预测, 大数据技术将对不断扩展新的仿真研究领域带来机遇, 特别是对人体仿真、社会仿真和人脑仿真等带来的机遇, 并为仿真用于社会治理、预测、城镇化等提供前所未有的机遇. 还有专家指出, 大数据技术的发展, 对采用建模仿真方法研究 Cyberspace 等新型虚拟信息空间提供了重要机遇. 众所周知, 建模仿真技术的优势在于提供了一个从现实世界通往虚拟空间的桥梁 (从实到虚), 为研究人类社会、物理世界的未知领域问题提供了一个可替代物. 然而, Cyberspace 是一个完全不同于物理世界、人类社会的全新虚拟空间, 这个虚拟空间不仅是 "人、机、物" 三者相融合的空间, 同时

它还具有多层网状、跨域关联等特点, 对这一全新虚拟空间的研究是当前科学研究的热点和难点. 那么, 在充分利用大数据研究成果的基础上, 能否采用平行系统、嵌入式仿真等方法, 建立起通往这类新型空间的桥梁 (从虚到虚), 为研究它们之间的交互影响或行为特点提供一个虚拟的替代物? 能否利用大数据的成果不断修正、检验所构建的模型, 这些都是建模仿真科学进一步发展需要关注的问题.

6 结束语

此次沙龙重点讨论了大数据时代对建模仿真的挑战、机遇与变革等部分问题. 可以预见, 仿真模式与手段必将随着大数据科学的不断发展而发展, 将大数据科研模式与现有的理论科研模式、实验科研模式、仿真计算科研模式等柔性、有机的融合, 将为社会、生命、工程、军事、科学等领域的研究, 特别是复杂系统研究提供更为高效的研究模式和手段. 此次沙龙的举办, 必将积极促进我国仿真科学技术在大数据时代的发展, 为推动我国建模与仿真理论与技术的研究、应用、产业和人才的培养做出新的贡献.

致谢 参加此次研讨的专家有 (按拼音字母排序): 毕长剑、丁刚毅、曹建文、范文慧、费敏锐、淦文燕、郝文宁、胡晓惠、金炜东、李伯虎、李伟、刘洋、吕金虎、马帅、邱晓刚、汤大权、王积鹏、吴琳、肖田元、易东云、张宏军、赵沁平. 在此向各位专家表示衷心的感谢!

参考文献

1 Li G J. The scientific value of big data research. Communication of the CCF, 2012, 09: 8–12 [李国杰. 大数据研究的科学价值. 中国计算机学会通讯, 2012, 09: 8–12]
2 Meng X F, Ci X. Big data management: concepts, techniques and challenges. 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]
3 Li B H, Xiao T Y. Simulation: The Third Method to Recognize and Change the World. Beijing: China Science and Technology Press, 2007 [李伯虎, 肖田元. 仿真 —— 认识和改造世界的第三种方法吗. 北京: 中国科学技术出版社, 2007]
4 Hey T, Tansley S, Tolle K. The Fourth Paradigm: Data-Intensive Scientific Discovery. Pan J F, Zhang X L, trans. Beijing: Science Press, 2012 [潘教峰, 张晓林, 译. 第四范式: 数据密集型科学发现. 北京: 科学出版社, 2012]
5 Glotzer S C, Kim S. International Assessment of Research and Development in Simulation-Based Engineering and Science. London: Imperial College Press, 2011
6 Lohr S. The Age of Big Data. New York Times, 2012
7 Li B H, Hu X F. The Puzzle and Thinking on the Simulation of Complex System. Beijing: China Science and Technology Press, 2012 [李伯虎, 胡晓峰. 复杂系统建模仿真中的困惑与思考. 北京: 中国科学技术出版社, 2012]
8 Fei M R, He G S, Cao J L. Life system modeling and simulation research and application progress, In: Proceedings of System Simulation Society Annual Conference, 2003. 710–715 [费敏锐, 何国森, 曹家麟. 生命系统建模仿真的研究和应用进展. 系统仿真学会学术年会论文集, 2003. 710–715]
9 Xu G B, Zeng L Z. Simulation-based complex system study. Comput Simulat, 2013: 1–4 [徐庚保, 曾莲芝. 基于仿真的复杂系统研究. 计算机仿真, 2013: 1–4]
10 Bakshi K. Considerations for big data: architecture and approach. In: Proceedings of Aerospace Conference. Piscataway: IEEE, 2012. 1–7

Simulation in the big data era — review of new ideas and new theories in the 81st Academic Salon of China Association for Science and Technology

HU XiaoFeng*, HE XiaoYuan & XU XuLin

Department of Information Operation and Command Training, National Defense University of PLA, Beijing 100091*, China*
*E-mail: xfhu@vip.sina.com

Abstract In September 2013, the Chinese Association for System Simulation undertook the 81st New Ideas and New Theories Academic Salon of China Association for Science and Technology. The theme is "challenges and thinking of modeling and simulation in the era of big data", which focused on 3 topics, including "Is big data really the fourth paradigm", "How the big data challenge simulation" and "What opportunities the big data bring to simulation". This paper introduces the main opinions of the experts and shares the achievements of this salon.

Keywords big data, the fourth paradigm, modeling, simulation, complex system

HU XiaoFeng was born in 1957. He is now a professor and doctoral advisor. His research interests include military simulation, virtual reality and multimedia systems, and military information systems. He is the vice president of Chinese Simulation Society, the vice president of Chinese Military Operations Research Society, the committee chairman of War Complex Systems, the vice committee chairman of Military Systems Engineering Society, the director of Chinese System Engineering Society, the special committee of Chinese Institute of Computer Multimedia.

XU XuLin was born in 1978. He received the Ph.D. degree from Nankai University, Tianjin, in 2010. Currently, he is working at the National Defense University. His research interests mainly focus on complex system, nonlinear dynamics, and complex networks.

HE XiaoYuan was born in 1968. She received the Ph.D. degree from National Defense University, Beijing, in 2009. Currently, she is an associate professor at the National Defense University. Her research interest includes war simulation system, military operation research. She directed many research projects, including 863 national plan projects, national nature science foundation projects, and military projects so on. She has published more than 30 papers in various. She is a member of CCF.