科学研究范式的演化

——大数据时代的科学研究第四范式

邓仲华 李志芳 （武汉大学信息管理学院 湖北 430072）

摘要 文章分析了科学研究范式的演化背景，并阐释了其演化过程，重点解释了数据密集型科学及科学研究第四范式，最后用知识地图的方式呈现出科学研究范式的整个演化过程及体系结构。

关键词 科学研究范式 第四范式 数据密集型科学

The Evolution of Scientific Research Paradigm:
The Fourth Paradigm of Scientific Research in the Era of Big Data

Deng Zhonghua Li Zhifang (School of Information Management, Wuhan University, Hubei, 430072)

Abstract The paper analyzes the background of the evolution of scientific research paradigm and explains its evolution process. The authors importantly explain the keypoints in data-intensive science and scientific research paradigm. Finally the authors present its whole evolution process and architecture using knowledge map.

Keywords scientific research paradigm, the fourth paradigm, data-intensive science

哲学上科学研究范式是指关于研究的一系列基本观念，主要包含存有论问题、认识论问题和方法论问题。存有论问题解释实在的本质到底如何；认识论问题解释知识的本质到底如何；方法论问题解释如何获得知识。这说明在整个科学研究的过程中，科学研究一直以来是以问题为驱动的。随后范式的理论被应用到各个学科，作为研究本学科范式的基础。2009 年，微软在 The Fourth Paradigm: Data-Intensive Scientific Discovery[1] 中从科学研究方法的角度解释科学研究范式并指出新的科学研究范式，针对数据密集型科学的第四范式的产生，引起了大家对于数据密集型科学的重视，也引起了大家对于科学研究第四范式的研究。科学研究范式演化的背景怎样以及具体的演化过程是什么，是进行科学研究第四范式研究的前提。

1 科学研究范式的演化背景

1.1 科学研究范式

“范式”这一概念最初由美国著名科学哲学家托马斯·库恩 1962 年在《科学革命的结构》中提出来，指的是常规科学所赖以运作的理论基础和实践规范。“范式”是从事某一科学的科学家群体所共同遵从的世界观和行为方式，它包括三个方面的内容[3]：(1) 共同的基本理论、观点和方法；(2) 共有的信念；(3) 某种自然观（包括形而上学假定）。“范式”的基本原则可以在本体论、认识论和方法论三个层面表现出来，分别回答事物存在的真实性问题、知者与被知者之间的关系问题以及研究方法的理论体系问题。这些理论和原则对特定的科学家共同体起规范的作用，协调他们对世界的看法及其行为方式。由于产生于特定的历史时期和特定的科学家群体，“范式”的基本理论和方法不是固定不变的，而是随着科学的发展发生变化。

在范式和科学共同体基础上，库恩又提出科学知识增长模式：前学科（没有范式）—常规科学（建立范式）—科学革命（范式动摇）—新常规科学（建立新范式）。在前学科时期，科学家之间存在意见分歧，因而没有一个被共同接受的范式。不同范式之间竞争和选择的结果是一种范式得到大多数科学家的支持，形成科学共同体公认的范式，从而进入常规科学时期。在常

本文系教育部人文社会科学重点研究基地重大项目“信息资源云体系及服务模型研究”（编号：11JJD630001）和武汉大学自主科研项目（人文社会科学）的研究成果，得到“中央高校基本科研业务费专项资金”资助。

规科学时期,科学共同体的主要任务是在范式的指导下从事释疑活动,通过释疑活动推动科学的发展,“常规科学即解难题(Puzzle)”[3]。在释疑活动过程中,一些新问题和新事物逐渐产生,并动摇了原有的范式,建立新范式的科学革命随之产生。革命的结果是拥有新范式的新的科学共同体取代拥有旧范式的旧的科学共同体。新范式的产生并不表示新范式更趋近真理,只是解题能力的增强。在后库恩时期,为了进一步阐明范式,库恩提出了专业母体,又可译为学科基质,是指一个科学共同体成员共同掌握的、有待进一步发展的基础,它主要包括概括(公式)、模型(一种形而上学的假设)和范例(最具体的题解),其中范例是最基本的要素,它使原先范式概念的模糊性得到改善。

1.2 背景分析

科学研究第四范式的产生,一方面是由于科学研究范式本身的发展,另一方面是由于外部环境的推动。库恩提出的科学知识的增长模式说明了科学研究范式是一个不断解决新问题、不断发展的过程。而随着信息技术的发展,社会环境的变化,促使新的问题不断产生,使科学研究范式受到各个方面的挑战。主要表现如下:

1.2.1 大数据的挑战

大数据[4]是一类我们已经无法使用目前的数据管理工具在能够容忍的时间范围内去收集、分析、管理和处理的数据集。它具有“4V+1C”的特点[5]:数据量大(Volume)、数据类型繁多(Variety)、价值密度低(Value)、处理速度快(Velocity)、复杂性(Complexity)。大数据带来的挑战不仅包括数据量的挑战,而且包括数据结构、数据融合和数据处理效率方面的挑战。

大数据到底有多大?一组名为“互联网上的一天”的数据告诉我们,一天之中,互联网产生的全部内容可以刻满1.68亿张DVD;发出的邮件有2940亿封之多;发出的社区帖子达200万个;卖出的手机为37.8万台,高于全球每天出生的婴儿数量37.1万……[6]

目前,数据量已经从TB级别跃升到PB级、EB级乃至ZB级别。国际数据中心(IDC)的研究结果表明,2008年全球产生的数据为0.49ZB,2009年的数据量为0.8ZB,2010年增长为1.2ZB,2011年的数据量更是高达1.8ZB,相当于全球每人产生200GB以上的数据。而到目前为止,人类生产的所有印刷材料的数据量是200PB,全人类历史上说过的所有话的数据量大约是5EB。IBM的研究称,整个人类文明所获得的全部数据中,有90%是过去两年内产生的,而到了2020年,全世界所产生的数据规模将达到今天的44倍[1]。

1.2.2 信息技术发展的挑战

移动互联网、电子商务、物联网以及社交媒体的快速发展已经使我们进入了大数据时代。目前的数据处理工具已不能高效率地对大数据进行存储、处理与应用。IDC在2012年的互联网市场报告中指出:作为数据运营组织、互联网公司正在从大数据的存储、处理与应用等各个环节推进技术的创新,这种创新可以从空间和时间两个维度进行透视,如图1所示[1][8]。

一是从空间维度出发,以非关系数据库、分布式

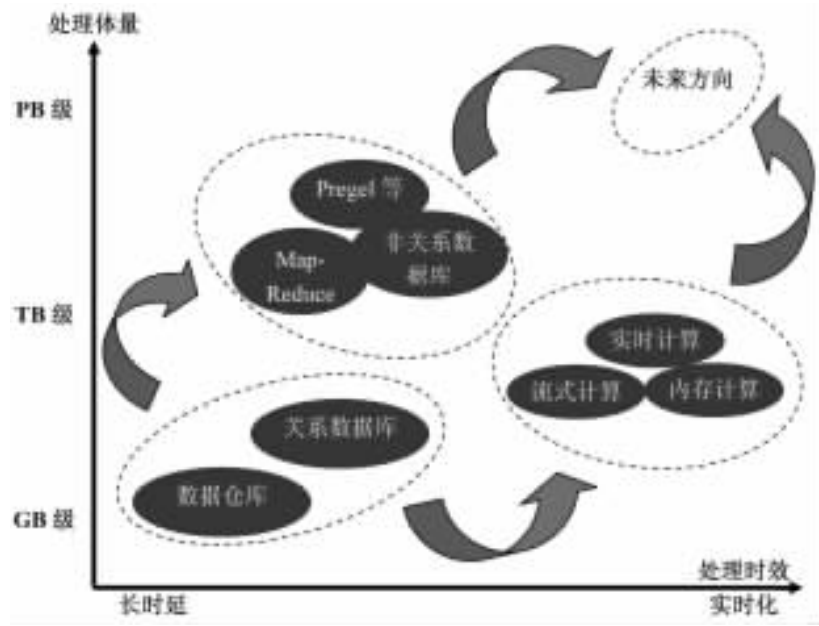

图1 中国互联网市场洞见:互联网大数据技术创新研究

计算架构等为代表,互联网公司正在不断提升数据处理的体量,尤其是强化对日益增加的非结构化数据的驾驭能力。分布式架构还让互联网公司能够利用大量相对廉价的服务器与存储设备来应对大数据集,并灵活地进行弹性部署。这意味着互联网行业正在步入数据处理的规模经济时代,在大数据潮流中走在前面的互联网公司能够赢得明显的数据成本优势。

二是从时间维度出发,流式处理、实时计算、内存计算等技术的涌现,体现了数据处理高度实时化的新趋势。MapReduce等模型尽管能够以优异的性能完成数据的块式处理,但面对许多在线业务每秒上万次的动态并发查询,仍然表现得力不从心,而流式计算等架构则能够更好地应对这种业务场景,将大数据的处理进一步推向实时。

IDC认为,今后这两个方向将进一步相互融合,在数分钟甚至几十秒内,完成TB级乃至PB级数据集的计算,并从中提取富含商业价值的结论,将成为互联网行业的新常态。

IDC对于信息技术在大数据环境下的发展方向的预测,一方面表明现有的信息技术解决不了大数据的问题;另一方面,信息技术在科学研究中的应用加快了数据的产生。

1.2.3 科学研究过程的挑战

在不同的学科领域,学者们依据研究重点的不

同，对科学研究的过程有不同的理解。所以有多种模型描述科学研究的过程。福建师范大学外国语学院李荣宝[9]认为，科学研究的目的是提示研究对象内存的一般规律并提出一套能够对研究对象进行充分描写和解释的抽象理论。科学研究的这种基本目标决定了其基本过程。他认为，科学研究的首要步骤是熟悉相关研究领域的基本情况，弄清有哪些基本理论，采用哪些研究范式等；其次是就相关的研究提出新的科学命题或理论假设；再次是验证，它为理论提供肯定的事实依据；最后一个环节是理论的表述。张晓林提出科研环境中存在知识生命周期的理论，这个周期至少涉及以下过程：把握学科发展的趋势和重点；寻求研究问题的知识框架和解决路径；构造解决方案和获取相关信息；知识组织与交流[10]。英国科学与技术设施研究理事会(STFC) 提出了数字科研环境下科研活动的生命周期模型，其研究过程可以描述为：熟悉相关研究领域的发展情况，产生新的研究思路，设计解决方案，进行实验或模拟，收集数据，分析数据，发布成果[11]。

数据分析和管理是科学研究过程中必不可少的一步，在科学研究过程中，已不是跟踪别人正在做什么或者解决尚未解决的问题，而是要从数据中发现自己不知道的问题。另一方面，发布成果时应该重视支撑这些成果的数据的存储和再利用。

1.2.4 科学数据管理贡献价值的挑战

科学数据管理贡献价值的挑战主要是指目前科学数据本身内在的贡献价值与人类目前能够挖掘出的价值的矛盾，即目前的数据处理和分析技术无法应对大数据状况，获得大数据背后隐藏的价值。

科学数据的科学价值主要表现在它是科学研究的基础，同时它也是科学研究的“牵引力”。在科学研究中，对于科学数据的利用其实是一个过程或一条相互关联的链，如图 2 所示。而一直以来，尤其是在人文社科中，学者们过分重视科学数据加工后的成果，只取金字塔顶的价值来利用，而忽视了形成该成果之前链条的价值。位于塔底的支撑数据和过程演化数据的科学价值其实比结果的利用价值更高。一方面它们可以帮助学者在做相似研究时更好地理解最终成果的形成；另一方面，它们可以作为其他学者的研究材料，实现数据共享，而且可以起到学术监督的作用。

2 科学研究范式的演化过程

微软在 The Fourth Paradigm：Data- Intensive Scientific Discovery 中指出[11]，科学研究的范式包括四个(如图 3 所示) ：几千年前，是经验科学，主要用来描述自然现象；几百年前，是理论科学，使用模型或归纳法进行科学研究；几十年前，是计算科学，主要模拟复杂的现象；今天是数据探索，统一于理论、实验和模拟。它的主要特征是：数据依靠信息设备收集或模拟产生，依靠软件处理，用计算机进行存储，使用专用的数据管理和统计软件进行分析。

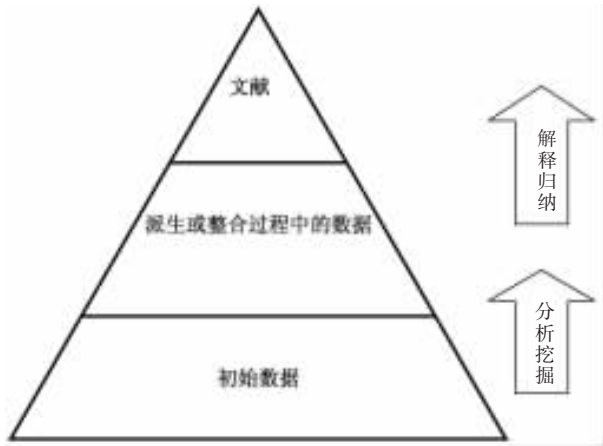

图 2 科学数据层次

2.1 经验科学

经验科学[13]是“理论科学”的对称，指偏重于经验事实的描述和明确具体的实用性的科学，一般较少抽象的理论概括性。在研究方法上，以归纳为主，带有较多盲目性的观测和实验。一般科学的早期阶段属经验科学，化学尤甚。在恩格斯《自然辩证法》中，专指 18 世纪以前搜集材料阶段的科学。在《史学理论大辞典》中，经验科学指西方学者用于概括自然科学和社会科学共同性的一个术语，它是指自然科学和社会历史科学都是从感觉经验出发的，都是以经验材料为其研究对象的，因此都具有经验科学的性质。

“经验科学”亦称“实验科学”[13]，是以实验方法为基础的科学。这种方法自从 17 世纪的科学家 Francisc Bacon 阐明之后，科学界一直沿用着。他指出科学必须是实验的、归纳的，一切真理都必须以大量确凿的事实材料为依据，并提出一套实验科学的“三表法”，即寻找因果联系的科学归纳法。其方法是先观察，进而假设，再根据假设进行实验。如果实验的结果与假设不符合，则修正假设再实验。

模型：科学实验。

范例：伽利略的物理学、动力学。伽利略是第一个把实验引进力学的科学家，他利用实验和数学相结合的方法确定了一些重要的力学定律。在 1589~ 1591 年间，伽利略通过对落体运动做细致的观察之后，在比萨斜塔上做了“两个铁球同时落地”的著名实验，从此推翻了亚里士多德“物体下落速度和重量成比例”的学说，纠正了这个持续了 1900 年之久的错误结论。牛顿的经典力学、哈维的血液循环学说以及后来的热力学、电学、化学、生物学、地质学等都是实验科学的典范。

2.2 理论科学

理论指人类对自然、社会现象按照已有的实证知识、经验、事实、法则、认知以及经过验证的假说，经由一般化与演绎推理等方法，进行合乎逻辑的推论性总结。人类借由观察实际存在的现象或逻辑推论，而得到某种学说，如果未经社会实践或科学试验证明，只能属于假说。如果假说能借由大量可重现的观察与实

验而验证，并为众多科学家认定，这项假说可被称为理论[14]。理论科学[15]是“经验科学”的对称，指偏重理论总结和理性概括，强调较高普遍的理论认识而非直接实用意义的科学。在研究方法上，以演绎法为主，不局限于描述经验事实。在恩格斯《自然辩证法》中，指19世纪以后成熟起来的，处于整理材料阶段的科学。

模型：数学模型。

范例：数学中的集合论、图论、数论和概率论；物理学中的相对论、弦理论、卡鲁扎—克莱恩理论（KK理论）、圈量子引力理论；地理学中的大陆漂移学说、板块构造学说；气象学中的全球暖化理论；经济学中的微观经济学、宏观经济学以及博弈论；计算机科学中的算法信息论、计算机理论。

2.3　计算科学

计算科学[16]，又称科学计算，是一个与数据模型构建、定量分析方法以及利用计算机来分析和解决科学问题相关的研究领域。在实际应用中，计算科学主要用于对各个科学学科中的问题进行计算机模拟和

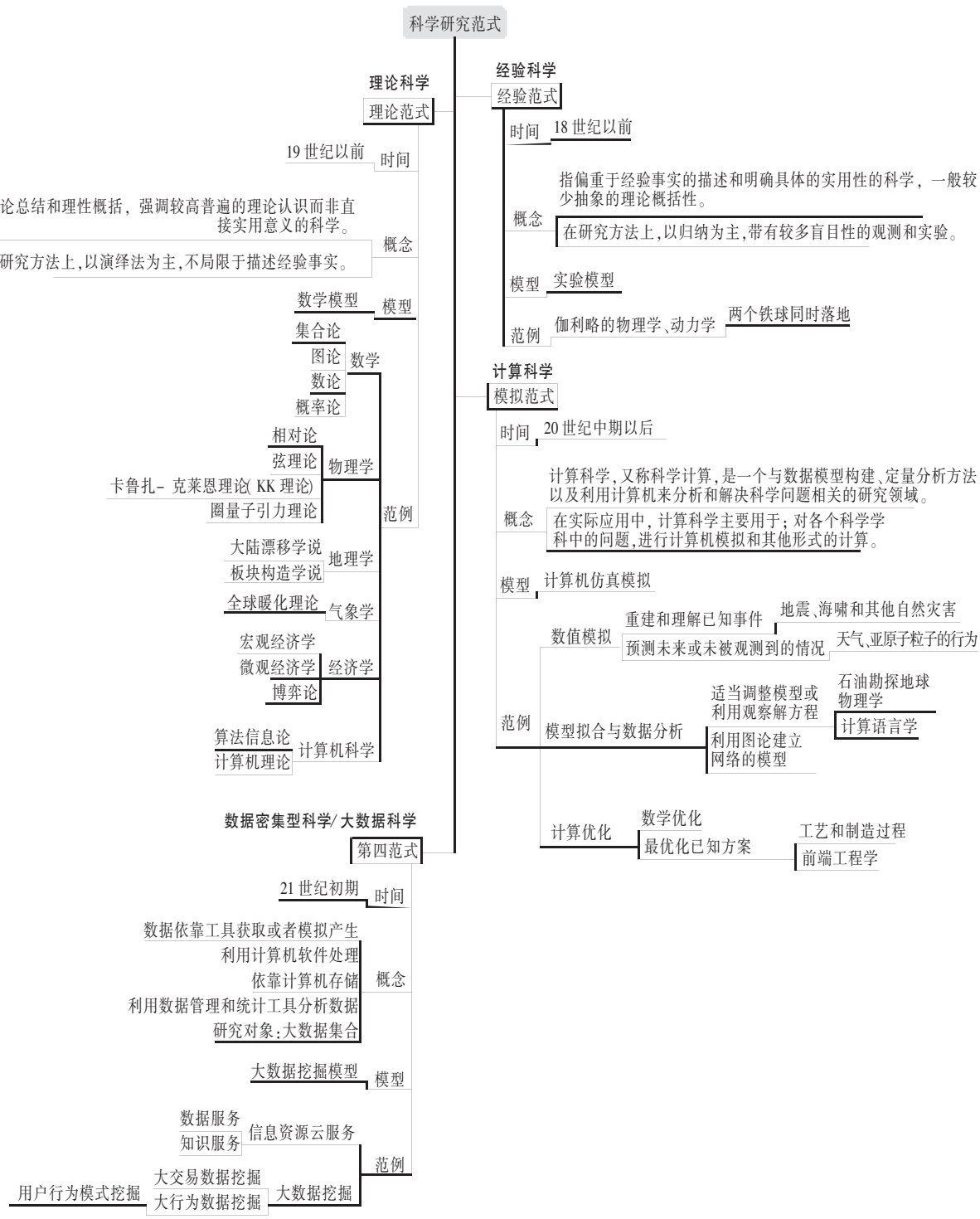

图 3　科学研究范式体系

其他形式的计算。其问题域包括：

（1）数值模拟。数值模拟有各种不同的目的，取决于被模拟的任务的特性。重建和理解已知事件（如地震、海啸和其他自然灾害）；预测未来或未被观测到的情况（如天气、亚原子粒子的行为）。

（2）模型拟合与数据分析。适当调整模型或利用观察来解方程，不过也需要服从模型的约束条件（如石油勘探地球物理学、计算语言学）；利用图论建立网络的模型，特别是那些相互联系的个人、组织和网站的模型。

（3）计算优化。数学优化；最优化已知方案（如工艺和制造过程、前端工程学）。

模型：计算机仿真/模拟。

范例：人工智能、热力学和分子问题、信号系统等。

2.4 数据密集型科学

科学研究第四范式是针对数据密集型科学，由传统的假设驱动向基于科学数据进行探索的科学方法的转变。数据依靠工具获取或者模拟产生；利用计算机软件处理；依靠计算机存储；利用数据管理和统计工具分析数据。

数据密集型科学的研究对象是科学数据。笔者综合考虑，将其研究对象确定在四类：即时收集到的观察数据、源自实验室仪器设备的实验数据、源自测试模型的模拟仿真数据、互联网数据。其中互联网数据受信息技术革新的影响在互联网环境下而产生的大行为数据和大交易数据。大行为数据主要产生于社会网络中，例如 Twitter、新浪微博、虚拟社区等；大交易数据的产生主要基于电子商务的社会化。

模型：大数据挖掘模型。

范例：信息资源云服务、大数据挖掘服务。信息资源云服务是在信息资源云平台（云存储平台和云服务平台）上进行的数据和知识的存储以及数据和知识的服务体系。大数据挖掘服务主要是基于大行为数据的用户行为特点的挖掘和基于大交易数据的市场预测。

由于数据密集型科学中的科学研究第四范式的发展刚刚起步，其模型和范例并没有形成统一的标准，本文所提出的关于科学研究第四范式的模型和范例是综合目前的发展状况以及笔者正在做的相关研究而提出来的。

3 科学研究范式之间的关系

经验科学是理论科学的实践基础，重复实验直至完全准确，则形成了理论，如果理论从未被推翻，则形成定律。理论科学是经验科学的指导，经验科学是在已有的理论基础上进行实验的。两者是互相联系、互相补充、互相推进的。计算机科学是对经验科学和理论科学中的科学方法的补充和优化，而数据密集型科学是处理经验科学和计算机科学中出现的大数据处理问题，是对前三种科学的补充。

4 结语

科学研究范式是对科学研究的规范，在进行科学研究时必须遵循本学科已经形成的大家公认的科学理论体系。信息密集型科学的出现使科学研究以数据为中心、以数据为驱动的特征越来越突出。从大数据中探索“不知道自己不知道”的现象和规律，成为科学研究中必不可少的部分。科学从经验科学到理论科学再到计算机科学，现在发展到数据密集型科学，科学范式也相应地从经验范式发展到理论范式再到计算机模拟范式到第四范式。每一个范式都有各自相应的特征和范例，清楚认识各个范式的特点和所包含的范例，对于科学研究第四范式的发展有着重要的意义。其最重要的是加强能够在合理的时间范围内处理数据密集型问题的软件工具的开发。

参考文献

[1] Tony Hey, Stewart Tansley, Kristin Tolle. The fourth paradigm[M]. Microsoft Press, 2009.

[2] 范式和范式转移[OL]. [2012-11-10]. http://www.360doc.com/content/06/1016/18/12749_232037.shtml.

[3] T·S·库恩. 科学革命的结构[M]. 李宝恒，纪树立，译. 上海：上海科学技术出版社，1980：29.

[4] http://zh.wikipedia.org/wiki/大数据.

[5] 大数据时代的特点[EB/OL]. [2012-11-18]. http://www.5lian.cn/html/2012/xueshu_0417/32237.html.

[6] 大数据到底有多大？[EB/OL]. [2012-11-18]. http://www.ciotimes.com/ea/data/72789.html.

[7] 大数据时代降临[EB/OL]. [2012-11-18]. http://today.banyuetan.org/jrt/120922/70953.shtml.

[8] IDC：互联网拥抱大数据，数据即服务（DaaS）时代到来[EB/OL]. [2012-11-18]. http://chinasourcing.mofcom.gov.cn/c/2012-05-24/121764.shtml.

[9] http://wenku.baidu.com/view/fd0871a1284ac850ad0242ca.html.

[10] 张晓林. 从数字图书馆到 E-Knowledge 机制[J]. 中国图书馆学报，2005(4)：5-10.

[11] How JISC is helping researchers: Research lifecycle diagram [OL]. [2012-11-18]. http://epubs.stfc.ac.uk/bitstream/3857/.

[12] http://baike.baidu.com/view/443426.htm.

[13] 实验科学[OL]. [2012-11-18]. http://gongjushu.cnki.net/refbook/BasicSearch.aspx?kw=%E5%AE%9E%E9%AA%8C%E7%A7%91%E5%AD%A6.

[14] http://zh.wikipedia.org/wiki/%E7%90%86%E8%AE%BA.

[15] http://baike.baidu.com/view/443344.htm.

[16] http://zh.wikipedia.org/wiki/计算科学.

[作者简介] 邓仲华，男，1957 年生，武汉大学信息管理学教授，博士生导师。

李志芳，女，1987 年生，武汉大学信息管理学院博士研究生。

收稿日期：2013-03-16