第一章

大数据基础知识

大数据(Big Data)是人们获得新认知、创造新价值的源泉,是改变市场、组织机构以及政府与公民关系的方法。1980 年,著名未来学家阿尔文·托夫勒在《第三次浪潮》一书中将大数据赞颂为"第三次浪潮的华彩乐章"。如今,大数据成为人们不可忽视的重要资源,谁拥有了海量的数据资源,谁用活了数据资源,谁就拥有未来。

第一节 大数据的相关概念

信息技术与经济社会的交汇融合引发了数据迅猛增长,数据已成为国家基础性战略资源,大数据正日益对全球生产、流通、分配、消费活动以及经济运行机制、社会生活方式和国家治理能力产生重要影响。

一、大数据的定义

"大数据"并不是一个确切的概念。最初,"大数据"的概念是由最先经历信息爆炸的学科(如天文学和基因学)创造出来的。[①] 如今,这个概念几乎应用到了人类所有的发展领域,尤其是在信息技术领域大数据已

① "大数据"专刊,《自然》,2008 年 9 月 4 日。

经成为最为流行的词汇。总体来讲,“大数据”或称巨量数据、海量数据,是由数量巨大、结构复杂、类型众多的数据构成的数据集合,是基于云计算的数据处理与应用模式,通过数据的集成共享、交叉复用形成的智力资源和知识服务能力。

(一)Gartner 对大数据的定义

大数据的研究机构 Gartner 给出了这样的定义:大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的多样化、高增长率和海量的信息资产。

(二)维基百科对大数据的定义

大数据是指一些使用目前现有数据库管理工具或传统数据处理应用很难处理的大型而复杂的数据集。其挑战包括采集、管理、存储、搜索、共享、分析和可视化。

(三)百度百科对大数据的定义

大数据,指的是所涉及的资料量规模巨大到无法透过目前主流软件工具在合理时间内达到撷取、管理、处理,并整理成为帮助企业经营决策更积极目的的资讯。

(四)IBM 对大数据的定义

IBM 通过调查研究认为,大数据本身的关键属性可概括为“3V”,即:数量(volume)、多样性(variety)和速度(velocity)。但还需要考虑另一个重要的第四维度:精确性(veracity)。将精确性作为大数据的第四个属性凸显了应对与管理某些类型数据中固有的不确定性的重要性。

(五)微软对大数据的定义

微软对大数据定义也采用了“3V”模型,并进一步指出:(1)由于硬件成本在快速下降,而数据的复杂性在不断提高,并且包含大量的非结构化数据,如文件、图像、视频、Web 日志、点击流和地理空间数据,数据量仍在持续增加。(2)社交媒体的兴起极大地增加了数据的种类,数据更具有鲜明的多样化特征。(3)随着各类终端设备成为大数据的数据源,数

据产生速度飞速增加。①

(六)SAS 对大数据的定义

SAS 作为专业的商业分析软件与供应商,在对大数据的传统"3V"模型定义基础上加入了"可变性"和"复杂性"两个重要特征。可变性主要反映数据流可能具有高度的不一致性,并存在周期性的峰值。复杂性只要体现在数据来源的多样性上,处理来自多个系统的数据是一件非常复杂的事情。

二、大数据的特点

虽然大数据不具有唯一确切的定义,但从各个角度的定义分析可以看出大数据具有四大典型特点:第一,数据体量巨大。数据体量已经从 TB 级别跃升到 PB 级别(数据最小的基本单位是 bit,按顺序给出所有单位:bit、Byte、KB、MB、GB、TB、PB、EB、ZB、YB、BB、NB、DB,其中 8 bit = 1 Byte,从 Byte 开始,相邻单位之间按照进率 1024——2 的十次方来计算)。第二,数据类型多样。电子邮件、网络日志、视频、图片、购买历史、各类仪器上传数据、地理位置信息等等。第三,处理速度快。对于有些应用来说,数据的处理时间必须以秒来计算,快速的信息处理可从各种类型的数据中快速获得高价值的信息,这与传统的数据挖掘技术有着本质的不同。第四,价值密度低。价值密度的高低与数据总量的大小成反比,随着数据量增大,错误率也会相应增加,有价值的数据比例相应也会降低。但是只要合理利用纷繁复杂的相关数据,能对其进行正确、准确的分析,将会带来很高的价值回报。如何通过有效的数据处理算法更迅速地完成数据的价值"提纯"成为目前大数据背景下亟待解决的难题。

业界将上述特点归纳为 4 个"V"——Volume(数据体量大)、Variety(数据类型繁多)、Velocity(处理速度快)、Value(价值密度低)。

---

① 鲍亮、李倩:《实战大数据》,清华大学出版社,2013 年版。

第二节 大数据的研究内容

一、大数据面临的挑战

大数据可以粗略地分成来自物理世界和来自人类社会两大类。前者多半是科学实验数据或传感数据，后者与人的活动有关系，特别是与互联网有关。未来社会的主要任务不是获取越来越多的数据，而是数据的去冗分类、去粗取精，从数据中挖掘知识。几百年来，科学研究一直在做“从薄到厚”的事情，把“小数据”变成“大数据”，现在要做的事情是“从厚到薄”，要把大数据变成小数据。要在不明显增加采集成本的条件下尽可能提高数据的质量，要研究如何科学合理地抽样采集数据，减少不必要的数据采集①。在这样的背景下，大数据研究将面临以下挑战②：

（一）数据量的成倍增长对于数据存储能力的挑战

传统的数据库追求高度的数据一致性和容错性，对数据设备的存储能力和更新扩展性要求相对较低，大部分数据存储设备不能有效存储视频、音频等非结构化和半结构化的数据。大数据时代，大数据及其潜在的商业价值则要求有专门的数据库技术和专用的数据存储设备进行支撑，当前数据存储能力的增长远远赶不上数据的增长，因此，设计最合理的分层存储架构成为信息系统的关键③。

（二）数据类型的多样性对于数据挖掘能力的挑战

从数据库的角度看，挖掘算法的有效性和可伸缩性是实现数据挖掘

① 李国杰、程学旗：《大数据研究：未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考》，《中国科学院院刊》，2012 年第 6 期。

② 陶雪娇等：《大数据研究综述》，《系统仿真学报》，2013 年 8 月增刊。

③ 李国杰：《大数据研究的科学价值》，《中国计算机学会通讯》，2012 年第 9 期。

的关键，现有的算法通常适合于常驻内存的小数据集，大型数据库中的多样化数据可能无法同时导入内存，随着数据规模的不断增大，算法的效率逐渐成为数据分析流程的瓶颈。要彻底改变被动局面，必须对现有架构、组织体系、资源配置和权力结构进行有效重组。

（三）大数据的处理速度对于数据处理时效性的挑战

传统的数据挖掘技术在数据维度和规模增大时，需要的资源呈指数增长，随着数据规模的不断增大，分析处理数据的时间相应地越来越长。大数据时代对信息处理的时效性要求越来越高，面对 PB 级以上的海量数据，NlogN 甚至线性复杂度的算法都难以接受，处理大数据需要简单有效的人工智能算法和新的问题求解方法。

（四）数据跨越组织边界传播对于信息安全的挑战

大数据时代，随着技术的发展，数据价值越来越大。大数据的海量数据，通常采用云端存储，数据管理比较分散，对用户进行数据处理的场所无法控制，很难区分合法与非法用户，容易导致非法用户入侵，窃取或篡改重要数据信息，犯罪分子更加容易获取各类保密数据、隐私数据，国家安全、知识产权、个人信息等都面临着前所未有的安全挑战。信息安全问题对数据存储的物理安全性以及数据的多副本与容灾机制提出了更高的要求。如何利用算法和特征建立相应的强大安全防御体系来发现和识别安全漏洞是保证信息安全的重要环节。

（五）大数据时代的到来对于人才资源的挑战

从大数据中获取价值必须具备三类关键人才队伍：第一，进行大数据分析的资深分析型人才；第二，精通如何申请、使用大数据分析的管理者和分析家；第三，实现大数据的技术支持人才。此外，由于大数据涵盖内容广泛，所需的高端专业人才十分广泛，既包括程序员和数据库工程师，同时也需要如物理学家、生态学家、数学和统计学家、社会网络学家和社会行为心理学家等。可以说，在未来几年，资深数据分析人才短缺问题将越来越凸显。更需要具有前瞻性思维的实干型领导者，能够基于从大数

据中获得的见解和分析，制定相应策略并贯彻执行。

二、大数据的处理过程

为了应对上述五大挑战，2012 年由微软、IBM、谷歌、HP、MIT、斯坦福、加州大学伯克利大学、UIUC 等产业界和学术界的数据库领域专家共同发布了一个关于大数据的白皮书。白皮书建议用现有的成熟技术解决大数据带来的挑战，并提出了大数据的处理过程大致可分为数据获取/记录、信息抽取/清洗/注记、数据集成/聚集/表现、数据分析/建模和数据解释等五个主要阶段①。它贯穿所有节点，具体分析步骤，如图 1－1 所示。

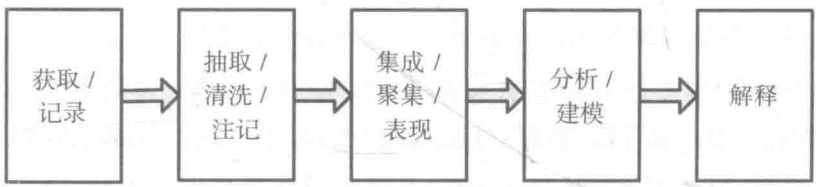

图 1－1 大数据处理框架

（一）数据获取/记录

大数据一般都来自多个不同的源头，并且大多是以动态数据流的形式产生。因此，大数据中常常包含有不同形态的干扰数据。另外，数据采样算法缺陷与设备故障也可能会导致大数据的干扰。大数据普遍存在冗余现象，这是由于大数据的多源性导致了不同源头的数据中存在有相同的数据，从而造成数据的绝对冗余；另外，就具体的应用需求而言，大数据可能会提供超量特别是超精度的数据，这又形成数据的相对冗余。消除干扰和消除冗余是提高数据质量、降低数据存储成本的基础，而数据获取正是进行这方面的工作。对原始的数据进行智能化的处理，将不需要的

① 鲍亮、李倩：《实战大数据》，清华大学出版社，2013 年版。

信息进行过滤，生成正确的数据源并进行记录是大数据处理系统的第一步。

（二）数据抽取/清洗

通常，在信息搜集的时候会得到很多不符合要求的数据，如不完整的数据、错误的数据和重复的数据等。这些数据不能直接用来分析，必须将需要的数据从众多类型的基础数据中抽取出来。数据清洗是一个反复的过程，它的任务是过滤那些不符合要求的数据，将过滤的结果交给业务主管部门，确认是否过滤掉还是由业务单位修正之后再进行抽取。

（三）数据集成/聚集/表现

大数据处理不能仅仅对数据进行记录，因为存在大量的不同格式、特点和来源的数据，如果这些数据不经处理直接进行存储，那么其他人就无法查改数据，更无从谈起对数据的使用。数据集成是把不同来源、格式、特点性质的数据在逻辑上或物理上有机地集中，从而为使用者提供全面的数据共享。当前，在企业数据集成领域，已经有了很多成熟的框架可以利用，最常用的是联邦式、基于中间件模型和数据仓库等方法来构造集成的系统，这些技术在不同的着重点和应用上解决数据共享和为用户提供决策支持。

（四）数据分析/建模

数据建模是指对现实世界各类数据的抽象组织，确定数据库需管辖的范围、组织形式等，最终转化成现实的数据库。数据建模过程中的主要活动包括：确定数据及其相关过程、定义数据、确保数据的完整性、定义操作过程、选择数据存储技术。传统的小数据分析是建立在关系数据模型之上的，主题之间的关系被创立在系统内，分析也在此基础上进行。而在信息高度发达的现实世界里，很难在所有的信息之间以一种正式的方式建立关系，非结构化数据以图片、视频、移动产生的信息、无线射频识别等的形式存在，在被考虑进大数据分析的过程中，绝大多数的分析基于纵列数据库之外。

（五）解释

大数据的效果价值在于让用户理解数据的分析结果，而仅仅向用户提供分析结果是不够的，还需要向用户解释这种结果是如何产生的。因此，对数据的解释不能凭空出现，通常包括检查所有提出的假设并对分析过程进行追踪，不断向用户提供附加资料，由于大数据本身的复杂性，这一过程特别具有挑战性，是一项非常重要的内容。

上述数据处理过程还可以归结为数据的采集、存储、分析及呈现，涉及的具体技术将在第四章阐述。

第三节 大数据的研究现状

国内外的学术界、产业界和政府机构都越来越意识到大数据的巨大发展和应用潜力，纷纷加入了对大数据的研究序列。各国的顶尖学术机构和具有影响力的公司都积极进行跨界联合，对大数据进行深入系统的研究。很多发达国家相继制定实施大数据战略性文件，大力推动大数据发展和应用。目前，我国互联网、移动互联网用户规模居全球第一，拥有丰富的数据资源和应用市场优势，大数据部分关键技术研发取得突破，涌现出一批互联网创新企业和创新应用，一些地方政府已启动大数据相关工作。

一、国外大数据研究状况①

1989 年在美国底特律召开的第 11 届国际人工智能联合会议专题讨论会上，首次提出了“数据库中的知识发现”的概念。1995 年召开了第一届知识发现与数据挖掘国际学术会议，随着与会人员的增多，数据库中的

① 鲍亮、李倩：《实战大数据》，清华大学出版社，2013 年版。

知识发现国际会议发展为年会。1998 年在美国纽约举行了第四届知识发现与数据挖掘国际学术会议，不仅进行了学术讨论，而且 30 多家软件公司展示了自己的产品，标志着数据研究已成为产业界的重要目标。

在学术方面，美国著名的加州大学伯克利分校利用政府注资的 1000 万美元，开展了旨在采用机器学习技术和云计算技术解决大数据问题、挖掘大数据中的重要信息的“大数据研究与开发”项目研究。加州大学伯克利分校 Lawrence 国家实验室的研究人员联合了 7 所大学和 5 所其他国家实验室，主要从事大数据管理、分析和可视化方面的研究工作。同年，加州大学伯克利分校还开设了一门关于大数据的公开课，着重从软件工程的角度介绍大数据的分析技术，探索解决大数据问题的方法。麻省理工学院的计算机科学和人工智能实验室则是与英特尔联合进行大数据研究项目。其关注点在于计算平台、可伸缩的算法、机器学习和理解及隐私和安全等四个方面的问题研究与解决。华盛顿大学设立了一个跨学科的大数据方面的博士学位，还开设了一个关于数据科学方面的认证项目，提供相关的教育与培训服务。计算机科学与工程系利用自身在数据管理、机器学习和开放信息抽取方面的传统优势，开展了大数据管理、数据可视化、大数据系统 Web 上的大数据、大数据和发现等多项科研项目。在斯坦福大学，学校提供了大规模数据挖掘认证课程，校内的学生可以选修相关课程获得认证。医学系还专门成立了生物医学专业大数据组，定期组织生物学、医学、计算机等方面的专家针对大数据问题进行研讨，其目的在于跨学科的研究和讨论大数据问题。

在产业界，谷歌、IBM、微软、SAS、EMC、Teradata、亚马逊等都进行了大数据的研究和应用，并取得了出色的成绩。谷歌先后推出了 MapReduce 模型和 BigQuery 服务。MapReduce 是面向大数据及处理的编程模型，由于其简单而又强大的数据处理接口核对大规模并行执行、容错和负载均衡等实现细节的隐藏，该技术一经推出便迅速在机器学习、数据挖掘和数据分析等领域得到广泛应用。BigQuery 服务对于开发者来说提供了

基于 Web 服务的编程接口，使得开发者可以利用谷歌后台的架构对超大规模数虚拟据库进行操作；对于用户来说，该服务允许用户上传超大规模数据，并对数据进行分析，客户端不需要做任何事情就可以得到海量数据的实时分析结果。IBM 推出了包括 BigInsights 和 Stream 两个产品系列的 InfoSphere 大数据分析平台。这两个产品中 BigInsights 适用于大规模的静态数据分析，能够在常用、低成本的硬件上运行，可用于支持半结构化或非结构化的信息，不需要烦琐的预处理，允许跨信息类型动态添加结构和关联。它还可以支持主动风险管理与预测、实体识别与情绪趋势分析等新型工作负载，同时配备了高级文本识别功能。Stream 则是一个实时大数据分析平台，它是一个擅长处理流动数据的高性能计算平台，能够对数以万计的信息源进行快速采集、分析和关联操作，及时捕捉并处理关键业务数据，能够满足用户对反应时间和扩展性的要求，并且支持高容量、结构化和非结构化流数据源。微软公司则在数据检索、数据处理和数据存储等方面对大数据进行了研究，开发出了一系列的产品。在数据检索方面，为了呈现高质量的搜索结果，微软分析了超过 100PB 的数据。在数据处理方面，微软为高性能计算提供了分布式的运行和编程模型，并利用相关技术提高了系统的处理能力和扩展性。在数据存储方面，微软还提出了并行数据仓库的概念，提供了企业级的计算能力，能够处理超过 600TB 的大数据量。SAS 是世界上较早提供大数据分析服务的公司，他们研究的大数据产品主要包括高性能分析服务器、SAS 可视化分析和 SAS DataFlux 数据流处理引擎，能为科学计算、时间序列趋势预测、作业成本管理、金融大数据整体解决方案、客户智能、财务智能、政府行业解决方案等提供有效支持。EMC 针对大数据推出了 Greenplum 数据引擎软件，可以为新一代数据仓库所需的大规模数据和复杂查询功能提供支持。特别是针对数据处理工程中的协作问题，EMC 还推出了专门的大数据处理社交工具集，使得数据科学家可以通过类似 Facebook 的方式进行数据处理协作。针对大数据，Teradata 则推出了 Aster Data 产品，它将数据分

析推向了数据库内分析，提高了数据分析的性能。Facebook 新的大数据处理分析平台 PUMA，通过对数据多处理环节区分优化，相比之前单纯采用 Hadoop 和 Hive 进行处理的技术，数据分析周期从 2 天降到 10 秒以内，效率提高数万倍。

不仅是学术界和产业界，各政府机构也在不断加强对大数据的关注。联合国曾发起过一个名为“全球脉搏”的项目，该项目旨在利用消费互联网的数据推动全球发展。其利用自然语言解码软件，对社交网络和手机短信中的信息进行情绪分析，从而对失业率增加、区域性开支降低或疾病暴发等进行预测。基于该项目的研究成果，2012 年联合国在纽约总部发布了一份名为“大数据促发展：挑战与机遇”的政务白皮书，书中指出大数据对联合国和各国政府来说都是一个历史性的机遇，要探讨如何利用包括社交网络在内的大数据资源造福人类。建议联合国成员国建立“脉搏实验室”，开发网络大数据的潜在价值。目前印度尼西亚和乌干达已经率先建成了脉搏实验室。2012 年 3 月，奥巴马政府公布了“大数据研发计划”，还发布了《大数据研究开发倡议》，旨在提高和改进人们从海量、复杂的数据中获取知识的能力，发展收集、储存、保留、管理、分析和共享海量数据所需要的核心技术。大数据成为继集成电路和互联网之后美国政府信息科技关注的重点，在这一背景下，美国政府各个部门纷纷展开了大数据的相关研究。2012 年 7 月，日本总务省也推出了 ICT 战略研究计划，该计划重点关注大数据应用所需的云计算、传感器、社会化媒体等智能技术开发。大数据将会为医疗技术开发、缓解交通拥堵等公共领域提供便利与贡献。根据日本相关研究机构的分析，日本大数据应用带来的经济效益将超过 10 万亿日元。

二、国内大数据研究状况

与国外相比，国内对于大数据的研究起步稍晚，在学术界，中国科学院、清华大学、北京航空航天大学、中国人民大学等学术机构已经逐步开

展了对大数据的研究，但还未形成整体力量。在企业界，企业使用数据挖掘技术尚不普遍，但近几年以百度、阿里数据、新浪、腾讯等为代表的企业对大数据的研究应用出现了蓬勃发展的态势。

“大数据”概念正在引领中国互联网行业的新一次浪潮。2011 年 12 月，微软研究院宣布与中国科学院计算机网络信息中心签署合作备忘录，为中国的科研人员提供先进的云计算资源。未来两年，微软将每年向中国科学院计算机网络信息中心的研究项目捐赠 200 万小时的 Windows Azure 云资源，以及 15TB 的 Windows Azure 存储空间。此外，微软还将向科研人员提供额外的 300 万小时的 Windows Azure 资源和 25TB 的存储空间，用于双方共同选择的 6 至 12 个研究项目。英特尔公司与中国科学院自动化研究所联合成立了“中国英特尔物联技术研究院”，以攻克大数据处理技术、传输技术和智能感知等物联网核心技术为目标。同时该研究院还将与国际和国内一流科研院所、院校和企业合作，建立一个开放式的研究中心。清华大学的计算机科学技术系和地球系统科学研究中心等机构研究了清华云存储系统、大数据存储系统、大数据处理平台、社交网络云计算和海量数据处理系统等。在以当前互联网和大数据时代新型信息技术为牵引、创造新的学术领域和应用增长点为目的前提下，北京航空航天大学计算机学院、爱丁堡大学信息学院、香港科技大学计算机系、宾夕法尼亚大学和百度公司于 2012 年联合创立了“大数据科学与工程”国际研究中心。此外，北京航空航天大学还创办了国内第一个“大数据科学与应用”软件工程硕士专业，目的是以实际需求为牵引，结合企业内训和项目实践，让大学生掌握大数据管理、系统开发、数据分析与数据挖掘等方面的核心技能。中国人民大学的“云计算与大数据实验室”，则主要关注云计算、非结构化数据、海量数据、数据库等方向的研究。

在企业应用开发方面，百度作为最大的中文搜索引擎公司，拥有海量的数据，并且数据种类繁多，更新极快，为此百度自行开发了能够实现数据实时性、一致性和可扩展性的数据存储系统。淘宝是当前中国最大的

购物网站，为了研究用户数据的长期走势、购买商品的人群特征、商品成交排行等信息，阿里数据进行了一系列的研究，成立了商业智能部门，不断开发数据分析产品，先后推出了淘数据、数据门户、云梯、数据魔方、观星台、地动仪、黄金策、淘宝指数、淘宝时光机等一系列的大数据产品。众所周知，新浪微博是新浪公司在大数据应用领域推出的主要产品。在新浪微博页面聚合了用户的兴趣爱好、社交关系的综合展示，各类话题、图书、音乐、餐饮美食等内容都能在微博上生成专属的页面，网友也可以很方便地查找到有价值的内容。腾讯公司的大数据产品线非常广泛，包括门户网站、微博、视频、电子商务、无线、开放平台等多个跨平台领域。腾讯的大数据战略主要是为个人用户提升使用体验，从商家用户处获得广告收益。腾讯将7亿活跃账户的数据作为门户服务的支持，打造基于用户社交关系链的“下一代腾讯网络”，它将利用大数据和关系链为用户筛选、推荐适合本人的内容。同时，腾讯的广告也更多转向基于用户社交关系链的口碑营销。

当前我国正处在全面深化改革阶段，正在致力于加强顶层设计和统筹协调，大力推动政府信息系统和公共数据互联开放共享，加快政府信息平台整合，消除信息孤岛，推进数据资源向社会开放，增强政府公信力，引导社会发展，服务公众企业，努力营造以企业为主体的宽松公平环境，加大数据关键技术研发、产业发展和人才培养力度，着力推进数据汇集和发掘，深化大数据在各行业创新应用，促进大数据产业健康发展，通过促进大数据发展，加快建设数据强国，释放技术红利、制度红利和创新红利，提升政府治理能力，推动经济转型升级。值得欣喜的是，国务院2015年8月31日通过的《促进大数据发展行动纲要》提出，到2017年底前，明确各部门数据共享的范围边界和使用方式，跨部门数据资源共享共用格局基本形成。2018年底前，建成国家政府数据统一开放平台。2020年底前，逐步实现信用、交通、医疗、卫生、就业、社保、地理、文化、教育、科技、资源、农业、环境、安监、金融、质量、统计、气象、海洋、企业登记监管等民生

保障服务相关领域的政府数据集向社会开放。

在具体目标方面，一是打造精准治理、多方协作的社会治理新模式。将大数据作为提升政府治理能力的重要手段，通过高效采集、有效整合、深化应用政府数据和社会数据，提升政府决策和风险防范水平，提高社会治理的精准性和有效性，增强乡村社会治理能力；助力简政放权，支持从事前审批向事中事后监管转变，推动商事制度改革；促进政府监管和社会监督有机结合，有效调动社会力量参与社会治理的积极性。2017 年底前形成跨部门数据资源共享共用格局。二是建立运行平稳、安全高效的经济运行新机制。充分运用大数据，不断提升信用、财政、金融、税收、农业、统计、进出口、资源环境、产品质量、企业登记监管等领域数据资源的获取和利用能力，丰富经济统计数据来源，实现对经济运行更为准确的监测、分析、预测、预警，提高决策的针对性、科学性和时效性，提升宏观调控以及产业发展、信用体系、市场监管等方面管理效能，保障供需平衡，促进经济平稳运行。三是构建以人为本、惠及全民的民生服务新体系。围绕服务型政府建设，在公用事业、市政管理、城乡环境、农村生活、健康医疗、减灾救灾、社会救助、养老服务、劳动就业、社会保障、文化教育、交通旅游、质量安全、消费维权、社区服务等领域全面推广大数据应用，利用大数据洞察民生需求，优化资源配置，丰富服务内容，拓展服务渠道，扩大服务范围，提高服务质量，提升城市辐射能力，推动公共服务向基层延伸，缩小城乡、区域差距，促进形成公平普惠、便捷高效的民生服务体系，不断满足人民群众日益增长的个性化、多样化需求。四是开启大众创业、万众创新的创新驱动新格局。形成公共数据资源合理适度开放共享的法规制度和政策体系，2018 年底前建成国家政府数据统一开放平台，率先在信用、交通、医疗、卫生、就业、社保、地理、文化、教育、科技、资源、农业、环境、安监、金融、质量、统计、气象、海洋、企业登记监管等重要领域实现公共数据资源合理适度向社会开放，带动社会公众开展大数据增值性、公益性开发和创新应用，充分释放数据红利，激发大众创业、万众创新活力。五是培

育高端智能、新兴繁荣的产业发展新生态。推动大数据与云计算、物联网、移动互联网等新一代信息技术融合发展，探索大数据与传统产业协同发展的新业态、新模式，促进传统产业转型升级和新兴产业发展，培育新的经济增长点。形成一批满足大数据重大应用需求的产品、系统和解决方案，建立安全可信的大数据技术体系，大数据产品和服务达到国际先进水平，国内市场占有率显著提高。培育一批面向全球的骨干企业和特色鲜明的创新型中小企业。构建形成政产学研用多方联动、协调发展的大数据产业生态体系①。

在具体措施方面，国家层面，国务院《"十二五"国家战略性新兴产业发展规划》中提出，海量数据存储、处理技术的研发与产业化将作为我国未来战略新兴产业的一个方面，工业和信息化部也将大数据产业的相关内容纳入了关键技术创新工程。地方政府层面，广东省率先启动大数据战略推动政府转型，广东大数据战略坚持以"开放共享"推动大数据应用，以"开放应用"带动大数据在国内发展，再试图通过大数据的发展来促进社会创新，为"智慧广东"建设助力。广东省政府准备在财政、环保、招投标等领域率先开展数据公开试点，通过互联网等形式开放数据。上海市发布了大数据研发三年行动计划，计划在三年内重点选取医疗卫生、食品安全、终身教育、智慧交通、公共安全、科技服务等具有大数据基础的社会领域，探索交互共享、一体化的服务模式，建设大数据公共服务平台。除了推进六大公共平台，该计划还将在金融证券、互联网、制造业等六个重点领域开展行业应用研发，并力争在大数据一体机、新型架构计算机、大数据分析软件等技术方面取得突破。北京市近年来各相关单位积极筹建各类信息化系统，在基础设施建成的基础上，逐步实现信息、数据共享，建成了北京市现代城市交通网络系统、数字政务管理系统及北京旅游信息系统等，在政务信息公开、舆情监控、预警预案等多方面都进行了大数

① 国务院：《促进大数据发展行动纲要》，2015 年 8 月 31 日。

据应用的探索。此外,北京大数据交易服务平台正式上线运行,平台致力于为政府机构、科研单位、企业乃至个人提供大数据“交易服务”,盘活数据资产,实现数据资源的有效利用,逐步打破“数据割据”“数据孤岛”的不良发展局面,建立可信的数据交易机制,为数据所有者提供大数据变现的渠道,为数据开发者提供统一的数据检索、开发平台,为数据使用者提供丰富的数据来源和数据应用。山东省成立了农业大数据战略联盟以提升农业竞争力。联盟成员单位充分利用各自优势,开展大数据研究与开发服务,采用大数据研究手段,在搜集和存储气象、土地、水利、农资、农业科研成果、动物和植物生产发展情况、农业机械、病虫害防治、生态环境、市场营销、食品安全、公共卫生、农产品加工等诸多环节大数据的基础上,通过专业化处理,对海量数据快速“提纯”并获得有价值的信息,为政府、企业乃至各种类型单位的决策和发展提供支持,为公众提供便捷的服务。这将填补国内在农业领域应用大数据研究手段的空白。

第四节 大数据的应用价值

虽然大数据的概念至今没有统一的定论,但这对于大数据的研究而言并不是最重要的,如何充分地使用大数据才是关键。一切对大数据的研究其最终目的都是更好地应用大数据。国际咨询公司麦肯锡 2012 年大数据报告中的一组数据显示,大数据产业为美国医疗系统带来每年 3000 亿美元的收益;为欧洲公共管理部门带来 2500 亿欧元的收益;为零售业增加 60% 的净利润;为制造业减少 50% 的产品研发等成本。而 Canner 认为,2015 年超过 85% 的财富 500 强企业将在大数据竞争中失去优

势①。据市场调研机构 IDC 预测,大数据技术与服务市场将从 2010 年的 32 亿美元攀升到 2015 年的 169 亿美元,实现 40% 的年增长率,相当于 IT 与通信产业增长率的 7 倍②。当前,大数据的研究与应用已经在商业智能、互联网、咨询与服务以及医疗服务、零售业、金融业、通信等行业显现,并产生了巨大的社会价值和产业空间。

一、大数据在零售业领域的应用

美国最著名的商业零售巨头沃尔玛公司(Wal - Mart Stores)拥有世界上数一数二的数据仓库,是最早应用数据挖掘技术的企业之一,也是数据挖掘技术的集大成者。20 世纪 90 年代,沃尔玛尝试将一种通过分析购物篮中的商品集合,从而找出商品之间的关联关系的算法——Aprior 算法引入到了 POS 机数据分析中,并根据商品之间的关系,找出了客户的购买行为,从而获得了成功,产生了最经典的大数据应用案例——"啤酒与尿布"的故事③。沃尔玛的超市管理人员在一次分析销售数据时发现,在某些特定的情况下"啤酒"与"尿布"两件看上去毫无关系的商品会经常出现在同一个购物篮中,经过后续调查发现,这种现象出现在年轻的父亲身上。这是因为在美国有婴儿的家庭中,一般是母亲在家中照看婴儿,年轻的父亲前去超市购买尿布。父亲在购买尿布的同时,往往会顺便为自己购买啤酒,这样就会出现啤酒与尿布这两件看上去不相干的商品经常会出现在同一个购物篮的现象。如果这个年轻的父亲在卖场只能买到两件商品之一,则他很有可能会放弃购物而到另一家商店,直到可以一

① 顾芳、刘旭峰、左超:《大数据背景下运营商移动互联网发展策略研究》,《邮电设计技术》,2012 年第 8 期。

② GANTZ J,R EINSEL D. 2011 digital universe study: extracting value from chaos[EB/OL]. (2011 - 07). http://www.b - eye - network.com/blogs/devlin/archives/2011/071.

③ 杨旭、汤海京:《数据科学导论》,北京理工大学出版社,2014 年,第 17—18 页。