“十二五”国家重点图书出版规划项目

大数据技术与应用

丛书策划

上海大数据产业技术创新战略联盟（上海产业技术研究院）

上海市数据科学重点实验室（复旦大学）

丛书主编

朱扬勇 吴俊伟

Big Data

Technology and Application Series

吴俊伟 朱扬勇

主编

汇计划在行动

上海科学技术出版社

大数据技术与应用

汇计划在行动

吴俊伟 朱扬勇
主编

上海科学技术出版社

图书在版编目(CIP)数据

汇计划在行动/ 吴俊伟,朱扬勇主编. —上海:
上海科学技术出版社,2015.1(2015.1 重印)
(大数据技术与应用)
ISBN 978-7-5478-2447-4

Ⅰ.①汇... Ⅱ.①吴...②朱 Ⅲ.①数据处理-研究 Ⅳ.①TP247

中国版本图书馆 CIP 数据核字(2014)第 261267 号

汇计划在行动
吴俊伟 朱扬勇 主编

上海世纪出版股份有限公司
上海科学技术出版社 出版
(上海钦州南路 71 号 邮政编码 200235)
上海世纪出版股份有限公司发行中心发行
200001 上海福建中路 193 号 www.ewen.co
苏州望电印刷有限公司印刷
开本 787×1092 1/16 印张 6
字数 130 千字
2015 年 1 月第 1 版 2015 年 1 月第 2 次印刷
ISBN 978-7-5478-2447-4/TP·34
定价:45.00 元

内容提要

本书介绍了《上海市推进大数据研究与发展三年行动计划(2013—2015年)》的编制和实施过程。系统介绍了对大数据概念、内涵、技术和应用方面的认识,介绍了在上海信息化建设的基础和现状之上,如何让大数据在上海落地,并着力解决大数据应用过程中的关键技术问题,开展数据科学前瞻研究和人才培养;对三年行动计划进行了全面解读;介绍了“上海大数据产业技术创新战略联盟”发起、组建、运行方面的情况;介绍了“上海市数据科学重点实验室”的研究方向、管理模式和开放模式。

本书的主要读者是大数据及相关专业的从业人员。

大数据技术与应用

学术顾问

大数据技术与应用

编撰委员会

前 言

2012 年，上海市科学技术委员会开始布局大数据研究项目，先期布局的项目是“城市交通大数据应用关键技术研究与示范”和“区域医疗大数据挖掘技术研究与示范应用”，并启动了大数据行动计划的编制工作。2013 年 7 月，上海市科学技术委员会正式颁布《上海推进大数据研究与发展三年行动计划》(2013—2015 年)(以下简称“汇计划”)，并成立了“推进办公室”具体负责“汇计划”的推进实施。

2013 年 7 月，上海一批企业和研究机构发起成立了“上海大数据产业技术创新战略联盟”，在全社会营造数据研究和开发的氛围，促进形成若干引领大数据产业技术创新的企业联合实体，旨在打破现有的数据流通壁垒、加强企业间数据信息合作，形成产业核心竞争力，突破核心技术，形成产业技术标准。2013 年 9 月，上海市科委批准筹建“上海市数据科学重点实验室”，重点研究大数据基础理论和关键技术，研究数据科学学科发展，编制大数据人才培养方案和开展培养工作。2015 年起，复旦大学正式招收“数据科学专业”研究生。

上海的大数据研究与发展已经取得了可喜的成绩。在大数据基础设施方面，开发了大数据一体机、实时数据库、内存计算平台等；在大数据关键技术方面，设计了一批新型大数据挖掘算法，包括：多种大数据压缩算法、多级缓存算法、大数据聚类算法、特异群组挖掘算法、图挖掘算法、流数据挖掘算法等；在大数据应用方面，健康医疗、城市交通、互联网广告、航空、航运、互联网金融等诸多领域开发了相应的大数据服务平台；另外，在大数据质量、大数据评测、大数据治理方面取得积极成果。相关成果写入了《数据密集型计算和模型》、《大数据：应用与测评技术》、《城市交通大数据》、《医疗大数据》、《城市的数据逻辑》、《智慧城市大数据》、《金融大数据》等著作。

本书作为《大数据技术与应用》丛书的分册之一，主要介绍大数据基本概念和内容、

大数据的发展和应用，介绍汇计划的编制形成，解读汇计划的内容，介绍“上海大数据产业技术创新战略联盟”和“上海市数据科学重点实验室”的基本情况。希望通过本书能让读者对上海大数据研究与发展的总体情况有一个全面的了解，为阅读《大数据技术与应用》丛书的其他分册提供背景知识。

本书由吴俊伟、朱扬勇策划并确定内容和组织编写，第一章由朱扬勇编写，第二章由吴俊伟、廖志成编写，第三章由吴俊伟、毛火华编写，第四章由熊赟编写。全书由吴俊伟、朱扬勇统稿。

因作者技术水平和理解能力所限，书中难免有错误不妥之处，欢迎批评指正。

作 者

2014 年 9 月

目 录

第1章

大数据

当今，几乎各个领域的人，都在使用或关注大数据。一种技术、一个概念让政界、商界、学界的各个领域都为之兴奋不已，甚至超过了当年计算机的诞生，也超过了互联网的诞生。一夜间，大数据无处不在，大数据企业遍地开花。那么，大数据究竟是什么？为什么不叫“大信息”而叫“大数据”？大数据有什么用？如何用？该如何培养数据科学家？本章将试图回答这些问题，并介绍数据科学家及其培养情况。

1.1 数据界

网络空间(Cyber Space)是指计算机网络、广播电视网络、通信网络、物联网、卫星网等所有人造网络和设备构成的空间，这个空间真实存在。信息化的本质是将现实世界中的事物转化成数据并存储到网络空间中，即信息化是一个生产数据的过程，网络空间中的所有数据构成数据界(Datanature)[1,2]。

1.1.1 数据

“数据”(Data)的含义很广，不仅指 1011、8844.43 这样一些数字，还指“dataology”、“小舟扬帆出海”、“11/11/11”等符号、字符、日期形式的数据。数据是指能够输入到网络空间中的任何东西，如数值、字符、声音、图像等，处理数据的计算机程序本身也是“数据”[2]。

“Data”一词来源于拉丁语单词“Datum”，含义为“给定的事物(thing given)”，数据的最初含义是对事物的度量，例如 39℃、96 kg、CPI 为 3.6 等。数据是人类记住事物的方式之一。试图记住知道的东西是人的天性，但是人并不能做到过目不忘，于是人类寻求辅助手段来帮助记忆，数据便成为记录事物的符号。印刷术和造纸术的发明后，大量自然界的事物(自然现象、人文和社会等)用文字和图形表示，然后印刷成书等。它们可以长期保存，大量复制，并可以广泛传播。历史的事物或经历被记载在书中被存储和传播，如四库全书、圣经、史记等。自电子计算机及其存储设备发明以来，人类开始以二进制数位(bit)的形式记录事物，这些记录在磁、光、电介质上的数据更持久，并且通过计算机网络传播更快。当前，“数据”主要指计算机系统能够处理的任何东西，或者说是网络空间中的任何东西。例如，电影、照片、微博、微信、购物记录、住宿记录、乘坐飞机记录、银行消费记录、政府文件等都是数据。

数据是指网络空间中的唯一存在，即网络空间中的任何东西，是可度量的，可处理的，

可观测的，并占有空间的。

网络空间中，有各种各样的数据，如何对数据进行分类是一个重大科学问题，是数据科学的一个重要研究方向。进行合理的数据分类需要科学家长期的努力。下面是对数据进行一些直观上的分类。

1）依据数据表示的含义来划分

从数据表示的含义来分，数据可以分为两类：一类是表示现实事物的数据，称为现实数据，另一类则不表示现实事物，只在网络空间中存在，称为非现实数据。

现实数据主要包括：

（1）感知数据　是指通过感知设备感知现实世界获得的数据，包括感知生命的数据。这类数据是客观世界的直接反映。

（2）行为数据　是指人类科学研究、劳动生产、生活行为等所产生的数据。这类数据是人类行为的直接反映。

非现实数据种类繁多，目前还不能很好地进行分类，举例如下：

（1）计算机病毒　是指能够自我复制的一组计算机指令或程序代码，只在数据界中存在，而在自然界没有映射。

（2）网络游戏　包括与自然界对应的场景映射到网络空间中，和只在网络空间中的游戏场景设置。

（3）垃圾数据　是指没有任何含义的数据。

2）依据数据的权属来划分

数据权属还没有法律的界定，从目前数据的属性和数据被占有的情况来看，数据可以分成以下四类：

（1）私有数据　主要指个人隐私数据和个人工作数据。而个人工作数据涉及内容繁多，包括是工作单位的数据、个人工作需要收集到的数据和其他需要获得的数据等，还有散落在互联网络上的个人数据。

（2）企业数据　主要是指企业生产经营数据、企业的客户数据、企业的竞争对手数据、行业数据等，这些数据主要存储在企业的计算机系统中。

（3）政府数据库　主要指存储在政府计算机系统中的数据。

（4）公共数据库　主要是指发布在公共网站上的数据，这些数据能够通过搜索引擎访问到。

3）依据数据的组织形式来划分

从数据的组织形式来看，数据主要有下列一些组织形式：

（1）专用格式数据　有相当多的数据是由专用数字化设备产生的数据，如医学影像数据（X射线片、MR、CT等）、遥感数据、GPS数据等。这些数据的处理需要专门的设备或专门的软件。

（2）通用格式数据　在信息化早期阶段，大多数数据库都是存储在文件或通用数据库

中的，由文件系统或通用的数据库管理系统来管理。这些数据结构清楚，处理方便。

(3) 互联网数据　互联网上的数据，种类和格式繁多，还包括很多垃圾数据、病毒数据，关键是如何找到有用的数据。由于互联网数据的存在，使整个网络空间中的数据更加显现出自然界的一些特征。

1.1.2 数据资源

经过国民经济与社会信息化发展战略的实施，信息技术(Information Technology，IT)已被大众所熟悉，今天的工作、学习、生活无不依赖于信息技术。现在，很难想象如果没有银行卡如何出差旅游、买房买地；很难想象如果没有收银机超市如何运行；很难想象如果办公室没有计算机该如何工作等。信息化给我们的工作、学习、生活带来便利，我们已经不能退回到信息化之前的时代了，这是信息化的成就。

那么，信息化到底做了什么呢？信息化是将我们过去手工做的事情转换成计算机来做，并且会更加准确、方便、高效；信息化还将现实的事物通过摄像头、麦克风、传感器等采集到计算机中。透过信息化给人类带来好处的现象，所有信息化的结果是在计算机系统中形成了很多数据，所以我们不断地购买存储系统、买硬盘、买光盘、买 U 盘，不断地做备份，不断地保安全，为的是保存好信息化的成果、保存好工作成果、保存好值得纪念的东西等。因此，从网络空间的视角来看，信息化的本质是生产数据的过程。

随着信息化产生的数据逐渐积累，自然而然地形成了一个新的概念——数据资源。有含义的数据集结到一定规模后形成数据资源[2]。

“一定规模”是数据资源的要求，没有“一定规模”不能称为数据资源。当少数人、少数实体、少数工作实施信息化阶段，数据并不形成资源。但到了信息化时代，信息化的广度和深度都达到了相当水平，数据就成为资源。以个人数据为例，一个人的身份数据不能称为数据资源，但是一个城市所有居民的身份数据是很重要的数据资源。另外，计算机使用者也已经产生了很多数据资源，很多计算机用户个人都会有 TB 级别的硬盘、GB 级别的 U 盘或是 TB 级别的移动硬盘，他们在其中存储了大量文档资料、数码照片、家庭视频，以及他们收集到的其他数据，这些都是个人数据资源。更大的数据资源来自科学研究、广播电视和整个互联网等。在国民经济与社会信息化建设过程中，国家正在致力于建设自然人数据库、法人数据库、空间地理数据库和宏观经济数据库，这些都是很重要的数据资源。

信息化形成的数据资源非常巨大。当前，世界各国都在利用卫星、电子望远镜等设备，开展太空探测、深海探测、地球勘探等活动，收集宇宙、大气、地球、海洋等自然数据，形成自然数据资源；也利用 DNA 测序获得关于生命的数据，形成生命数据资源；国民经济与社会信息化则产生了社会发展和人类行为的数据，形成了经济社会资源。例如，在国民经济领域，有国家统计数据、证券交易数据、海关数据等；在社会领域，有民政数据、交通数据、医疗保险数据，以及大量的互联网行为(电子商务行为、网络游戏行为、电子邮件行为、网络社

区)等;在科学研究领域,有国家建设的地球系统科学数据共享平台、国土资源科学数据共享网、中国气象科学数据共享网等。

当前,整个社会已经离不开网络空间。事实上,社会是运转在网络空间中的。社会运转依据数据进行,并在运转中生产新的数据,人类行为以数据的形式记录在网络空间中。人类的社会、政治和经济活动都将依赖于数据资源,而石油、煤炭、矿产等自然资源的勘探、开采、运输、加工、产品销售等无一不是依赖数据资源,离开了数据资源这些工作都将无法开展。因此,数据资源是一种重要的现代战略资源,其重要程度将越来越明显,在21世纪有可能超过石油、煤炭、矿产等自然资源,成为人类最重要的资源之一。

对网络空间数据资源的占有、开发和利用在一定程度上也将是未来国家政治的战略竞争之所在。"茉莉花事件"、"震网事件"和"斯诺登事件"表明,网络空间中的政治、军事手段的威力将远超过核武器的威力,所谓的"货币战争"也是在网络空间中发生的战争。美国于2009年成立了网络部队司令部,开始了对网络空间的战略占领和控制。

数据资源的战略性表现在以下几个方面:

(1) 掌握数据资源将在国际上掌握主动权。不论是反倾销诉讼、铁矿石谈判、汇率问题、节能减排、碳关税谈判等重大国际政治、经济事务,无一不依靠数据进行决策。相关数据在网络空间中是存在的,需要将它们开发出来,为政治、经济服务。

(2) 掌握数据科学技术就是掌握未来经济。掌握数据科学技术才能开发利用数据资源,数据资源开发利用是未来产业的制高点。数据产业是战略型新兴产业,发展数据产业可以产生巨大的经济和社会效益,使国家从"国民经济与社会信息化战略"转向"基于网络空间的现代国家管理发展战略"。

特别需要注意的是,作为一种资源,数据应该有相应的权益。数据权益是指数据的所有权和获益权,需要建立相应的法律来保护数据的所有者权益。

亚马逊前首席科学家表示"数据是原油,但石油需要加以提炼后才能使用,从事海量数据处理的公司就是炼油厂"。国民经济与社会信息化形成的自然数据资源、经济社会数据资源、网络行为数据资源等非常巨大,正是这些数据资源的开发利用构成了当前的大数据热潮。

1.1.3 数据界的特征

人类社会的进步发展是人类不断探索自然(宇宙和生命)的过程,当人们将探索自然界的成果存储于网络空间中的时候,却在不知不觉中创造了一个数据界。虽然是人生产了数据,并且人还在不断生产数据,但当前的数据已经表现出不为人控制、未知性、多样性和复杂性等自然界特征。

1) 数据不为人类控制

数据爆炸式增长,人无法控制它,人们还无法控制计算机病毒大量出现和传播,垃圾邮件泛滥,网络攻击数据阻塞信息高速公路等。现在的日常生活中,人们都在不断生产数据,