SHUJU DAJINKUANG

数据大金矿

钱志新 著

南京大学出版社

数据大金矿

钱志新 著

南京大学出版社

图书在版编目(CIP)数据

数据大金矿 / 钱志新著. — 南京 : 南京大学出版社, 2013.8
ISBN 978-7-305-11962-0

Ⅰ. ①数… Ⅱ. ①钱… Ⅲ. ①数据处理—研究 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2013)第 187573 号

出版发行 南京大学出版社
社　　址 南京市汉口路 22 号　　邮　　编 210093
网　　址 http://www.NjupCo.com
出 版 人 左　健

书　　名 数据大金矿
著　　者 钱志新
责任编辑 陈　佳　纪玉媛　　编辑热线 025-83686308

照　　排 南京南琳图文制作有限公司
印　　刷 南京爱德印刷有限公司
开　　本 718×1000　1/16　印张 12.25　字数 172 千
版　　次 2013 年 8 月第 1 版　　2013 年 8 月第 1 次印刷
ISBN 978-7-305-11962-0
定　　价 32.00 元

发行热线 025-83594756　83686452
电子邮箱 Press@NjupCo.com
Sales@NjupCo.com(市场部)

前　言

“这是一种革命，我们确实正在进行这场革命，庞大的新数据来源所带来的量化转变将在学术界、企业界和政界中迅速蔓延开来。没有哪个领域不会受到影响，迎来大数据时代。”这是美国《纽约时报》对大数据最具代表性的精辟评论。

美国总统奥巴马发表“大数据计划”后，世界各地对大数据表现了前所未有的重视和关注，对《大数据时代》等著作起到了很好的传播和宣传作用。

中国对“大数据”的重视刚刚开始，《数据大金矿》试图从理论和实践的结合上进行系统的研究，特别注重大数据的实际应用，在数据决策、数据搜索、数据商务、数据营销、数据监测、数据仿真六个方面的应用全面展开，精心介绍分析大量典型案例，旨在为大数据的全方位推广应用作出科学指导。

Age of Big Data.

钱志新

目　录

导论　大数据革命加速到来

人类社会的历史从农业社会发展到工业社会，现在正由工业社会转向信息社会。进入21世纪以来，信息化加速发展，信息新技术不断涌现，全面应用信息技术，已经渗透到经济与社会的各个领域，信息化覆盖了整个现代社会。

一、大数据新浪潮

信息化以信息为载体。随着计算机的发明和应用，信息的承载从语言到文字再到现在的数据，大大提升了信息描述的准确性和信息传递的广泛性。无论什么样的信息都能转化为数字代码，即0和1。数据代码的单位为比特(Bite)，比特的量值是指数级的。1 000 KB为1 MB，1 000 MB为1 GB，1 000 GB为1 TB，1 000 TB为1 PB，1 000 PB为1 EB，1 000 EB为1 ZB，1 000 ZB为1 YB……数据已成为信息的结晶体。

当今，数据正在以惊人的速度增长，大数据已应运而生。根据IDC《数字宇宙膨胀：到2010年全球信息增长预测》中统计，2006年全球数据量为161 EB，大约相当于历史上全球图书馆信息总量的3 000倍。至2010年全球数据量猛增6倍，达到988 EB，年复合增长率为57%，过去3年全球数据量比以往4万年还多。专家预测，到2020年全球数据量较2010年增加43倍，达到35 ZB。数据正在宇宙级膨胀。

二、大数据广源头

物质与信息是一个整体，任何物质的存在都会有信息，通过信息人们可以更好地认识物质利用物质。信息无处不在，数据层出

不穷，当今社会的数据来源甚广，主要来自五大源头，即物理世界、政府、企业、社会、网络。

1. 物理世界

物理世界每时每刻都在产生海量信息。首先是自然界，如天体、天气、大地、江河、山川、动物、植物等万物都有大量信息活动；其次人造世界如大楼、道路、桥梁、水利工程、各种建筑物等，也产生大量信息数据，这些都是造就大数据的自然之源。

2. 政府

政府是大数据的集中之地。各国政府、各级政府、各政府部门在长期积累过程中，拥有历史的和现实的大量信息资料，成为最难得的原始数据。政府数据是战略资源，通过开放政府数据，为社会大众应用，社会价值无限。

3. 企业

企业无时无处不在制造信息数据。在企业生产经营过程中，大量的物流、商流、生产流、技术流、资金流等都会产生海量数据，人的行为产生的行为数据，机器产生的生产数据和交易数据都源源不断。特别在市场营销中，十分宝贵的是交易数据和行为数据，对开发客户资源具有重要作用。与此同时涌现的金融大数据更有广泛的价值。在美国，公司所有部门中至少有 100 TB 的数据存储，许多公司的数据存储已经超过 1 PB。

4. 社会

社会由人群组成，所有的人都在制造并分享数据，人们在工作、学习、生活、交往、娱乐、购买等活动中时刻都在产生大数据。社会中的各个领域，从科技、教育、文化、卫生、医疗、治安到社区等每天涌现的新数据层出不穷，社会是大数据的强大源头。

5. 网络

互联网是最大的信息源头，每时每分每秒都在创造海量数据。到 2011 年底，中国互联网行业持有的数据总量已达到 1.9 EB，预计 2015 年将增长到 8.2 EB 以上。据统计，AT&T 的网络每天流动 1.6 PB 的数据，Google 每天处理 20 PB 的数据，Facebook 每天存储 1 PB 的照片。Opera 浏览器每月处理 1 PB 数据，BBC 的 iPlayer 每月有 7 PB 的数据流，Youtube 每月存储 31 PB 的流媒体

数据。网络上的数据量都是天文数字。

物理世界、政府、企业、社会、网络五大源头都是大数据的集中来源。据麦肯锡全球研究院(MGI)估计,全球企业2010年在硬盘上存储了超过7EB的新数据,消费者在PC和笔记本电脑等设备上存储了超过6EB的新数据,两者数据总量相当于美国国会图书馆中存储数据的5.2万倍。大数据爆发式增长势不可挡。

三、大数据大智慧

信息是实体的表现形式,信息与实体是融为一体的。信息世界与实体世界互相对应,如果说实体世界为"阳",则信息世界为"阴",阴阳是互生互动的。实体世界中产生的众多问题与矛盾,往往表现在实体世界,而根源在信息世界。由于信息缺失、信息不对称和信息虚假,使信息世界发生混乱,从而造成实体世界的矛盾。信息世界大量数据,通过分析、整合、挖掘,可形成数据产品,返回到实体世界,对实体世界的发展起到优化、提升的巨大作用。

信息化的发展已经进入第三阶段。第一阶段是数字化,即E化,由于计算机的出现,将感知的信息数据化,成为可计算资源;第二阶段是互联化,即U化,由于互联网的出现,将各种数据通过电脑系统互联互通,成为一个数据的网络世界;第三阶段是智慧化,即I化,由于互联网的出现,将物体与物体互联,即M2M,可将万物互联。通过数据的智慧化应用,使实体系统变得"聪明"起来,成为智慧化的实体世界。信息化发展的EUI轨迹,展现了信息化向高级阶段智慧化发展的广阔前景。

智慧化发展是一个系统工程,包括大数据、云计算、物联网、移动通信等诸多新技术,其中大数据是基石,大数据重现数据之间的相关性,而非因果关系,是真正的智慧之源。

数据的智慧化应用,旨在指导实体世界的智慧化发展,大数据广泛应用于政府、社会和企业的各个领域,包括农业、工业、服务业、城建、国防、金融、科技、教育、文化、卫生、气象、环境、能源等方方面面。在大数据引领下,智慧城市、智慧企业、智慧学校、智慧医院、智慧交通、智慧社区等将蓬勃发展,必将形成一个智慧大世界。

按照大数据的功能，主要应用在以下方面。

第一，数据决策。通过大数据的收集、整理和分析，为政府、企业、社会等提供科学决策依据，预测未来发展趋势，对军事活动具有特别重要意义。

第二，数据搜索。通过搜索大量社会信息特别是网络媒体信息，为法人和个人提供舆情分析，评价可能风险，挖掘客户资源。

第三，数据商务。对企业内外信息进行整合、评价和挖掘，指导企业研发设计、生产制造和物流财务等，形成智慧供应链的整体发展。

第四，数据营销。对企业市场营销中大量交易数据、行为数据进行系统分析和挖掘，管理好优质客户，开拓新的客户，增强客户黏性。

第五，数据监控。通过软件系统，远程监控实体的数据，实时动态分析运行效率、安全保障、环境监测和医疗监护等。

第六，数据仿真。应用大数据系统仿真实体运行和模拟结果，达到重大项目优化设计方案和避免盲目投资的效果。

麦肯锡将大数据的价值定量化描述，认为大数据将显著提升五大方面的价值：① 美国医疗服务业，大数据至少每年能产生约3 000亿美元的价值；② 美国零售业，大数据提高净利润达到60%以上；③ 美国制造业，利用大数据可节约成本50%，降低营运资本7%；④ 欧洲公共管理部门，大数据每年能创造价值2 500亿欧元；⑤ 全球个人位置服务，服务商每年收入至少为1 000亿美元，最终用户价值高达7 000亿美元。预计全球未来五年大数据创造价值将达到58%的复合增长率。

大数据引领大智慧，大数据创造大价值。数据资源是战略资源，与物质资源相比，数据既不会枯竭，又不排放废物，取之不尽，用之不竭，已成为亟待开发的“大金矿”。未来将是一个以“数据”为引领的智慧时代，谁掌握了大数据，谁就将引领未来。

第一章　数据理论

大数据时代已然来临，大数据正以一种新的视角并作为一种新的工具来解析世界。政府的宏观规划、企业的分析决策、个人的生活消费等都已和大数据接轨，大数据催生新的商业模式，改变传统生活方式，创造新的发展机遇。物联网、云计算、互联网的发展都离不开大数据的支持，大数据以其高速处理、海量数据、多样化结构和潜在价值等特性，在整个智慧化进程中起着基础性的作用。

一、大数据的四大特性

描述数据特征的传统观点认为可以从“3V”角度来分析数据的基本特征，即“容量”（Volume）、“速度”（Velocity）和“种类”（Variety）。大数据应该具有数据量大、存储和处理速度快、数据多样化等三大特征。近来数据价值（Value）被认为是大数据的第四大特征，从海量数据中获取有价值的信息需要多种数据挖掘技术、分析工具和模型方法的支持，这也正好印证了大数据的前三大特征。从某种意义上讲，发觉数据的内在价值是实现数据智慧化的重要途径。大数据除了量大、处理速度快、结构种类多之外，实现数据价值才是大数据的主要内涵，数据价值化赋予数据生命力，使得大数据有“肉体”，也有“灵魂”。

（一）海量数据——淹没在数据海洋中

海量数据是大数据的最基本特征，海量数据不仅指数据量大，而且指数据关联复杂，海量数据是“数据海洋”，也是“数据网络”。伴随互联网的快速发展和社交网络的兴起，数据信息在虚拟世界中犹如汪洋大海般浩瀚无际。据阿里研究中心统计，《红楼梦》含标点在内共 87 万字（不含标点为 853 509 字），每个汉字占两个字

节，容量为 1 TB 的硬盘可以存储 631 903 部《红楼梦》，据 MGI(麦肯锡全球研究院)估计，2010 年全球企业在硬盘上存储了超过 7 EB(1 EB 等于 100 万 TB)的新数据，同时，消费者在 PC 和笔记本等设备上存储了超过 6 EB 的新数据。[①] 除此之外，信息量与日俱增，每天不断地更新。一天之中，百度大约要处理 60 亿次搜索请求，达到几十 PB 的数据；淘宝网站的交易量达数千万笔，单日数据量超过 20 TB。

大数据信息产生的渠道很多，主要有四个方面：一是互联网，这是个人获取数据的最常用渠道，主要指 SNS、微博、视频网站、电子商务网站获取的数据；二是物联网，主要指移动设备、终端中的商品、传感器采集的数据；三是从通信和互联网运营商获取数据；四是天文望远镜拍摄的图像、视频数据、气象学里面的卫星云图数据等。这些数据通过各种传输方式被终端所接收，经过加工后变成信息，然后通过各种载体被人们所获取。数据与数据之间以各种层级的关联度结合在一起，成为不同信息。目前，物联网、云计算和互联网可以将人和物的所有轨迹和信息记录并分析，整个互联网的范围变得更大，网络中的核心节点不再是网页，而是一个个数据节点，数据遍布在网络的任何环节，任何数据单元的重组都会变成新的信息，可见大数据背后的信息量是多大！

(二) 数据多样性——非结构化数据来袭

多样性是大数据的结构特征，传统意义上的海量数据是指结构化数据和半结构化数据，大数据具有多样性，不仅包含结构化数据，还包含非结构化数据。据统计，企业中 20%的数据是结构化的，80%的数据则是非结构化或半结构化的。当今世界结构化数据增长率约为 32%，而非结构化数据增长率是 63%。至 2012 年，非结构化数据占有比例将达到互联网整个数据量的 75%以上，而非结构化数据中 50%～75%的数据都来源于人与人的互动。

目前企业可分析的运行数据和交易数据大多属于结构化数据，典型的就是事务数据、定量数据。企业收集这些数据，利用数

① 阿里研究中心：《大数据时代》，2012。

据挖掘技术，分析企业内在问题和发展瓶颈，制定科学合理的战略方向、优化运营，提高企业综合竞争力。结构化数据能直观提供业务发展的多维信息，在制定报表、预判指标等方面十分有效。如今在电子商务、移动应用、社交网络等领域积存着大量的诸如图片、视频、音频、地理位置等非结构化或半结构化数据，其规模或复杂程度超出了常用技术按照合理的成本和时限捕捉、管理及处理这些数据集的能力。非结构化数据来袭，企业内部大量的影像资料、办公文档、扫描文件、Web 页面、电子邮件、微博、即时通信以及音视频等非结构化数据爆发式增长。相对于存储在关系型数据库里结构化数据而言，非结构化数据大多不便用二维逻辑结构来表示，但非结构化数据中蕴含了大量特定的价值，这些价值在通过量化分析等方法解析后得以产生和放大，将成为企业新的核心竞争力。

（三）数据的高速处理——行驶在数据高速公路

现实生活常常会遇到交通堵塞的情况，设想一下如果互联网也遇到网络堵塞是多么令人头痛的问题。尽管计算机存储技术和计算技术发展迅速，但面对大数据的侵袭似乎还有所不足。大数据的速度特征一方面表现在数据的存储速度，在高速网络中，通过高处理技术快速处理服务器存储网中传送过来的数据包，在这个过程中常常需要建立数据中心和计算中心专门用于数据的存储和传输；另一方面，大数据要求数据能够快速移动、处理和反馈，要求能根据用户的需求对海量非结构化数据快速地检索、计算和交互。目前，通过基于高速处理器创建实时数据流已成为解决大数据高速处理的主要技术。

畅游在快速的数据网络中，就如同驾车行驶在高速公路上，道路堵塞的概率会非常的小。高速数据的处理速度需要两大技术的支持，一是提高硬件存储处理性能，尤其是提高处理器的性能，高性能硬件设施是数据高速读写的基础；二是提高软件优化性能，软件性能的提升将能够帮助用户得到精确的数据，保证数据检索的科学性和时效性。只有硬件技术、软件技术的最佳融合才能构筑起大数据的“高速公路”。

（四）数据价值的稀疏性——挖掘数据潜在价值

数据背后隐藏着巨大的数据价值，需要通过数据挖掘来生成。商业智能无疑是将数据价值通过软硬件技术得以提炼的有效途径，然而在大数据中发现数据价值也绝非易事。大数据的价值具有“稀疏性”“不确定性”和“多元化”等几个主要特点。首先大数据的价值是存在的，但却是稀疏的。“网络女皇”Mary Meeker 曾这样比喻：把大数据比喻成整整齐齐的稻草堆，而数据价值就如同散落在稻草堆中的一个个缝衣针，利用大数据技术可以在稻草堆中找到你所需要的东西，哪怕是一枚小小的缝衣针。[①] 其次，大数据的价值具有不确定性。数据在大多数情况下需要通过统计分析、主题性挖掘才能直观地得到数据的潜在信息，才能产生价值。但难处在于无法辨认哪些数据是有数据价值，哪些数据是没有用的，数据是否有价值都与挖掘需求有关，这就如同视频监控录像，如果没有遇到偷盗事件或是其他事件需要回放，就根本无法判别哪天的监控录像是有用的。再次，数据价值呈现多元化。数据是原始生产资料，利用原始生产资料可以生产多种产品或是服务，数据的内在价值通过各种各样的产品和服务得以体现。即使是同一产品，使用目的不同也会使数据产生不同的价值。例如淘宝的部分交易数据，客户通过排名信息去寻找信誉好的产品或商家；商家利用评论信息去改善客户服务质量和产品质量；淘宝官方则利用交易数据改善网站购物体验和运营方式。

二、智慧化三大架构

当今，物联网、云计算和互联网的快速发展，加快了智能互联和智能计算的步伐。如果物联网被比作人类的四肢、眼、鼻、耳等“感官”，是人类实现对物理世界的“感、知、控”的工具，那么云计算就好比人类的“大脑”，通过对大量信息提取、分析和反馈，指导人类的行为决策。不管是物联网还是云计算都离不开数据中心的支

① 网络女皇 MaryMeeker：《2012 互联网趋势分析报告》，2012。

持，数据起着基础性的作用，其地位好比人类的“心脏”，维系着信息的采集、传输、分析、决策等整个过程的延续。物联网、云计算、大数据构建了智慧化的三大架构。

（一）物联网

物联网就是物物相连的互联网。这有两层意思：第一，物联网的核心和基础仍然是互联网，物联网是在互联网基础上延伸和扩展的网络；第二，其用户端延伸和扩展到了任何物品与物品之间，进行信息交换和通信。[①]

1. 物联网基本内涵

广义上的物联网被定义为利用条码、射频识别（RFID）、传感器、全球定位系统、激光扫描器等信息传感设备，按约定的协议，实现人与人、人与物、物与物在任何时间、任何地点的连接，从而进行信息交换和通讯，以实现智能化识别、定位、跟踪、监控和管理的庞大网络系统。联网的本质就是将 IT 基础设施融入到物理基础设施中，也就是把感应器嵌入到电网、铁路、公路、桥梁、隧道、供水系统、油气管道等各种物体中，实现信息的自动提取。[②]

物联网的发展离不开互联网的快速发展，物联网是信息化建设历程中的重要里程碑，其实现了物理世界和数字世界的无缝对接。物联网不同于感知信息搜集的传感器网络，也不同于信息传输的互联网。物联网的核心元素是建立一个由相互连接的物体构成的网络。通过通信网络，这些物体不仅能从它们周围环境获取信息，与物质世界进行互动，而且能够使用现有的互联网标准来提供服务。物联网充分体现了物理世界和信息空间的深度融合，使人类可以融入到一体化的智能生态环境中，实现人、机、物的协同统一。

作为崭新的综合性信息系统，物联网包括信息的感知、传输、处理决策、服务等多个方面。物联网自身显著的特点：第一，是对客观物理世界的全面感知，它不仅表现在对单一的现象或目标进

① 引自百度百科。

② 杨运平：《物联网平台在安防领域的应用概述》，中安网，2013。

物联网依靠互联网扩展

（来源：互动百科）

行多方面的观察获得综合的感知数据，也表现在对现实世界各种物体现象的普遍感知；第二，是物联网实体间的泛在互联，表现在各种物体经由多种接入模式实现异构互联，也突出表现在物联网不仅包括互联网、电信网等公共网络，还包括电网和交通网等专用网络；第三，是智慧的信息处理和决策，它体现在物联网中从感知到传输到决策应用的信息流，并最终为控制提供支持，也广泛体现在物联网中大量的物体和物体之间的关联和互动。物体互动经过从物理空间到信息空间，再到物理空间的过程，形成感知、传输、决策、控制的开放式循环。①

2. 物联网体系

一般认为物联网的体系架构可以分为感知层、网络层和应用层。其中感知层负责完成感知、识别物体，采集、捕获信息；网络层具备网络运营和信息运营的能力；应用层是将物联网技术与行业应用相结合，实现广泛智能化应用的解决方案集。

① 孙利民，沈杰，朱红松：《以云计算到海计算：论物联网的体系结构》，《中兴通讯技术》，2011 年第 1 期。