

大数据技术及其背景下的 数据挖掘研究

©吴春琼 / 著



中国水利水电出版社
www.waterpub.com.cn

本书为“大数据商务智能福建省高校工程中心”建设项目阶段性成果。

大数据技术及其背景下的 数据挖掘研究

◎吴春琼/著



中国水利水电出版社

www.waterpub.com.cn

·北京·

内 容 提 要

本书围绕大数据背景下的数据挖掘及应用问题,从大数据挖掘的基本概念入手,系统地阐述了大数据的基础理论、大数据处理架构 Hadoop 以及大数据存储与管理研究;由浅入深地论述了数据挖掘的基础理论、大数据挖掘技术研究、数据挖掘的艺术,并结合实践,阐述了数据挖掘的应用。

本书对于读者掌握大数据挖掘及应用领域的基本知识和进一步研究都具有参考价值。

图书在版编目 (CIP) 数据

大数据技术及其背景下的数据挖掘研究 / 吴春琼著

· -- 北京:中国水利水电出版社,2019.1

ISBN 978-7-5170-7469-4

I. ①大… II. ①吴… III. ①数据采集—研究 IV.

①TP274

中国版本图书馆 CIP 数据核字(2019)第 031157 号

责任编辑:陈洁

封面设计:王伟

书 名	大数据技术及其背景下的数据挖掘研究 DASHUJU JISHU JI QI BEIJING XIA DE SHUJU WAJUE YANJIU
作 者	吴春琼 著
出版发行	中国水利水电出版社 (北京市海淀区玉渊潭南路 1 号 D 座 100038) 网址: www.waterpub.com.cn E-mail: mchannel@263.net (万水) sales@waterpub.com.cn 电话: (010) 68367658 (营销中心)、82562819 (万水)
经 售	全国各地新华书店和相关出版物销售网点
排 版	北京万水电子信息有限公司
印 刷	三河市元兴印务有限公司
规 格	170mm×240mm 16 开本 14.75 印张 216 千字
版 次	2019 年 4 月第 1 版 2019 年 4 月第 1 次印刷
印 数	0001-3000 册
定 价	64.00 元

凡购买我社图书,如有缺页、倒页、脱页的,本社营销中心负责调换

版权所有·侵权必究

前 言

随着大数据时代的来临，大数据分析与管理技术将具有越来越重要的战略意义，大数据已经渗透各个行业和企业的功能领域，逐渐成为一个重要的生产要素，大数据的使用将预示着生产率增长和消费者过剩的新一轮浪潮。大数据的分析和挖掘技术帮助企业用户在合理的时间内获取、管理、处理和组织海量数据，从而为企业经营决策提供积极的帮助，大数据分析和挖掘作为数据存储和挖掘分析技术，广泛应用于物联网、云计算、移动互联网等战略性新兴产业。虽然大数据在国内还处于初级阶段，但它的商业价值已经显现出来，尤其是大数据分析人才的实践经验是企业的热点资源。

数据挖掘是一个渐进的过程。在电子数据处理的早期阶段，人们试图通过一些方法来实现自动决策支持。当时，机器学习是人们注意力的焦点。机器学习的过程就是将一些已知的并已被成功解决的问题作为范例输入计算机，相应的软件通过学习这些范例总结并生成相应的规则，通常这些规则具有通用性，使用它们可以快速解决某一类的实际问题。随着神经网络技术的形成和发展，人们将注意力转向知识工程，知识工程不同于机器学习（向计算机输入范例，让它生成规则），而是直接给计算机输入已被代码化的规则，计算机通过使用这些规则来解决某些问题。

本书阐述了大数据基础理论，对大数据处理架构 Hadoop 系统、SPSS Modeler 系统进行了分析，论述了大数据存储与管理、数据挖掘基础理论、大数据挖掘技术，诠释了数据挖掘的艺术以及数据挖掘的应用。

本书在逻辑安排上循序渐进，由浅入深，内容丰富，信息量大，融入了大量本领域的新知识及新方法。在撰写本书的过程中，笔者借鉴了许多前人的研究成果，在此表示衷心的感谢！并衷心期望本书能

为读者的学习生活以及工作实践带来帮助。由于笔者能力有限，书中难免会出现疏漏或不足之处，恳请各位专家及读者批评指正。

作者

2018年10月

目 录

前 言

第一章 大数据基础理论研究	1
第一节 大数据时代概述	1
第二节 大数据的概念及内涵研究	9
第三节 大数据与相关领域的关系	27
第四节 大数据关键技术	35
第二章 大数据处理架构 Hadoop	41
第一节 Hadoop 概述	41
第二节 Hadoop 生态系统	47
第三节 Hadoop 的安装与使用	51
第三章 大数据处理架构 SPSS Modeler	54
第一节 SPSS Modeler 软件概述	54
第二节 SPSS Modeler 文本挖掘概述	79
第三节 SPSS Modeler 文本挖掘节点	84
第四章 大数据存储与管理研究	90
第一节 分布式文件系统 HDFS 概述及其体系结构研究	90
第二节 分布式数据库 HBase 概述及其实现原理	101
第三节 NoSQL 数据库概述及其三大基石	114
第四节 云数据库概述及系统构架	129

第五章 数据挖掘基础理论研究	145
第一节 大数据与数据挖掘	145
第二节 数据挖掘的概念和原理	148
第三节 数据挖掘的内容	151
第四节 数据挖掘与相关领域的关系	158
第六章 大数据挖掘技术研究	169
第一节 数据挖掘与过程	169
第二节 大数据挖掘处理基本流程	186
第三节 数据挖掘常用算法	189
第七章 数据挖掘的艺术	198
第一节 数据挖掘目标确立的艺术	198
第二节 数据挖掘应用技术的艺术	200
第三节 数据挖掘中平衡的艺术	204
第四节 理性对待大数据时代	207
第八章 数据挖掘的应用	214
第一节 数据挖掘在银行信用卡评分中的应用——以 AHP 法为例	214
第二节 数据挖掘技术在生命科学中的作用	218
参考文献	225

第一章 大数据基础理论研究

大数据时代悄然来临，带来了信息技术发展的巨大变革，并深刻影响着社会生产和人民生活的方方面面。全球范围内，世界各国政府均高度重视大数据技术的研究和产业发展，纷纷把大数据上升为国家战略加以重点推进。企业和学术机构纷纷加大技术、资金和人员投入力度，加强对大数据关键技术的研发与应用，以期在“第三次信息化浪潮”中占得先机、引领市场。大数据已经不是“镜中花、水中月”，它的影响力和作用力正迅速触及社会的每个角落，所到之处，或是颠覆，或是提升，都让人们深切感受到了大数据实实在在的威力。

对于一个国家而言，能否紧紧抓住大数据发展机遇，快速形成核心技术和应用参与新一轮的全球化竞争，将直接决定未来若干年世界范围内各国科技力量博弈的格局。大数据专业人才的培养是新一轮科技较量的基础，高等院校承担着大数据人才培养的重任，因此，各高等院校非常重视大数据课程的开设，大数据课程已经成为计算机科学与技术专业的重要核心课程。

本章首先综述大数据的发展历程、基本概念、主要影响、应用领域、关键技术、计算模式和产业发展，其后阐述云计算、物联网的概念及其与大数据之间的紧密关系。

第一节 大数据时代概述

第三次信息化浪潮涌动，大数据时代全面开启。迅猛发展的信息科技成为大数据时代来临的技术支持，数据信息产生方式的重大变化是推进大数据时代来临的主要因素。

一、第三次信息化浪潮

到目前为止，信息化建设共产生了3次浪潮，见表1-1。

表 1-1 三次信息化浪潮

信息化浪潮	发生时间	标志	解决的问题	代表企业
第一次浪潮	1980 年前后	个人计算机	信息处理	Intel、AMD、IBM、苹果、微软、联想、戴尔、惠普等
第二次浪潮	1995 年前后	互联网	信息传输	雅虎、谷歌、阿里巴巴、百度、腾讯等
第三次浪潮	2010 年前后	物联网、云计算和大数据	信息爆炸	亚马逊、谷歌、IBM、VMWare、Palantir、Hortonworks、Cloudera、阿里云等

1980 年前后，个人计算机（PC）开始普及，使得计算机走入企业和千家万户，大大提高了社会生产力，也使人类迎来了第一次信息化浪潮，Intel、IBM、苹果、微软、联想等企业是这个时期的标志。随后，在 1995 年前后，人类开始全面进入互联网时代，互联网的普及把世界变成“地球村”，每个人都可以自由徜徉于信息的海洋，由此，人类迎来了第二次信息化浪潮，这个时期也缔造了雅虎、谷歌、阿里巴巴、百度等互联网巨头。时隔 15 年，在 2010 年前后，云计算、大数据、物联网的快速发展，拉开了第三次信息化浪潮的大幕，大数据时代已经到来，也必将涌现出一批新的市场标杆企业。

二、信息科技的进步是大数据时代的技术支撑

信息科技需要解决信息存储、信息传输和信息处理 3 个核心问题，人类社会在信息科技领域的不断进步，为大数据时代的到来提供了技术支撑。

（一）存储设备容量的增加促进了数据量的增加

数据被存储在磁盘、磁带、光盘、闪存等各种类型的存储介质中，随着科学技术的不断进步，存储设备的制造工艺不断升级，容量大幅

增加,速度不断提升,价格却在不断下降。

早期的存储设备容量小、价格高、体积大,例如,IBM 在 1956 年生产的一个早期商业硬盘,容量只有 5MB,不仅价格昂贵,而且体积有一个冰箱那么大,如图 1-1 所示。

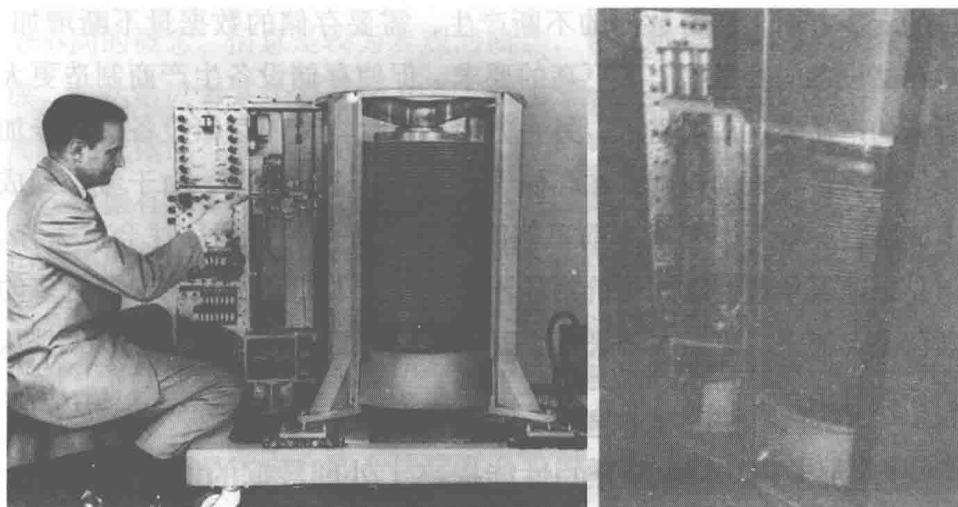


图 1-1 IBM 在 1956 年生产的一个早期商业硬盘

相反,今天容量为 1TB 的硬盘,大小只有 3.5 英寸(约 8.89cm),读写速度达到 200MB/s,价格仅为 400 元左右。廉价、高性能的硬盘存储设备,不仅提供了海量的存储空间,同时大大降低了数据存储成本。

与此同时,以闪存为代表的新型存储介质也开始得到大规模的普及和应用。闪存是一种新兴的半导体存储器,从 1989 年诞生第一款闪存产品开始,闪存技术不断获得新的突破,并逐渐在计算机存储产品市场中确立了自己的重要地位。闪存是一种非易失性存储器,即使发生断电也不会丢失数据;因此,可以作为永久性存储设备,它具有体积小、质量轻、能耗低、抗振性好等优良特性。

闪存芯片可以被封装制作成 SD 卡、U 盘和固态盘等各种存储产品,SD 卡和 U 盘主要用于个人数据存储,固态盘则越来越多地应用于企业级数据存储。一个 32GB 的 SD 卡,体积只有 24 mm × 32 mm × 2.1 mm,质量只有 0.5g。以前 7200 r/min 的硬盘,一秒钟读写次数只

有 100 IOPS (Input/Output Operations Per Second), 传输速率只有 50 MB/s, 而现在基于闪存的固态硬盘, 每秒钟读写次数有几万甚至更高的 IOPS, 访问延迟只有几十微秒, 允许人们以更快的速度读写数据。

总体而言, 数据量和存储设备容量二者之间是相辅相成、互相促进的。一方面, 随着数据的不断产生, 需要存储的数据量不断增加, 对存储设备的容量提出了更高的要求, 促使存储设备生产商制造更大容量的产品满足市场需求; 另一方面, 更大容量的存储设备进一步加快了数据量增长的速度, 在存储设备价格高企的年代, 由于考虑到成本问题, 一些不必要或当前不能明显体现价值的数据往往会被丢弃。但是, 随着单位存储空间价格的不断降低, 人们开始倾向于把更多的数据保存起来, 以期在未来某个时刻可以用更先进的数据分析工具从中挖掘价值。

(二) CPU 不断提升的性能提高了处理数据的能力

CPU 处理速度的不断提升也是促使数据量不断增加的重要因素。性能不断提升的 CPU, 大大提高了处理数据的能力, 使得人们可以更快地处理不断累积的海量数据。从 20 世纪 80 年代至今, CPU 的制造工艺不断提升, 晶体管数量不断增加, 运行频率不断提高, 核心 (core) 数量逐渐增多, 而同等价格所能获得的 CPU 处理能力也呈几何级数上升。在 30 多年里, CPU 的处理速度已经从 10 MHz 提高到 3.6 GHz, 在 2013 年之前的很长一段时期, CPU 处理速度的增加一直遵循“摩尔定律”, 性能每隔 18 个月提高一倍, 价格下降一半。

(三) 网络带宽的增加促进了大数据时代的信息传输

1977 年, 世界上第一条光纤通信系统在美国芝加哥市投入商用, 数据传输速率为 45 Mbit/s, 从此, 人类社会的信息传输速度不断被刷新。进入 21 世纪, 世界各国更是纷纷加大宽带网络建设力度, 不断扩大网络覆盖范围和传输速度。移动通信宽带网络迅速发展, 3G 网络基本普及, 4G 网络覆盖范围不断加大, 各种终端设备可以随时随地传输数据。大数据时代, 信息传输不再遭遇网络发展初期的瓶颈和制约。

三、数据产生方式的变革对大数据时代来临的促进作用

数据是人们通过观察、试验或计算得出的结果。数据和信息是两个不同的概念。信息是较为宏观的概念，它由数据的有序排列组合而成，传达给读者某个概念方法等；而数据则是构成信息的基本单位，离散的数据没有任何实用价值。

数据有很多种，比如数字、文字、图像、声音等。随着人类社会信息化进程的加快，人们在日常生活和生产中每天都会产生大量的数据，比如商业网站、政务系统、零售系统、办公系统、自动化生产系统等，数据随着分秒源源不断地出现。如今数据在各行各业和任何工作范围内都会产生，是决定生产过程和企业获取核心竞争力的关键要素。数据资源是国家和社会持续安全发展的重要影响因子，和物质资源、人力资源一样是国家拥有的重要战略资源，因此，数据也被称为“未来的石油”。

数据产生方式的变革，是推进大数据时代来临的主要因素。数据大体经过了运营式系统、用户原创内容和感知式系统三个阶梯过程的产生方式。

（一）数据产生的运营式系统阶段

人类社会最早大规模管理和使用数据，是从数据库的诞生开始的。大型零售超市销售系统、银行交易系统、股市交易系统、医院医疗系统、企业客户管理系统等大量运营式系统，都是建立在数据库基础之上的，数据库中保存了大量结构化的企业关键信息，用来满足企业各种业务需求。

在这个阶段，数据的产生方式是被动的，只有当实际的企业业务发生时，才会产生新的记录并存入数据库。比如，对于股市交易系统而言，只有当发生一笔股票交易时，才会有相关记录生成。

（二）数据产生的用户原创内容阶段

互联网的出现，使得数据传播更加快捷，不需要借助于磁盘、磁带等物理存储介质传播数据，网页的出现进一步加速了大量网络内容的产生，从而使得人类社会数据量开始呈现“井喷式”增长。但是，互联网真正的数据爆发产生于以“用户原创内容”为特征的 Web 2.0 时代。Web 1.0 时代主要以门户网站为代表，强调内容的组织与提供，大量上网用户本身并不参与内容的产生。而 Web 2.0 技术以 Wiki、博客、微博、微信等自服务模式为主，强调自服务，大量上网用户本身就是内容的生成者，尤其是随着移动互联网和智能手机终端的普及，人们更是可以随时随地使用手机发微博、传照片，数据量开始急剧增加。

（三）数据产生的感知式系统阶段

物联网的发展最终导致了人类社会数据量的第三次跃升。物联网中包含大量传感器，如温度传感器、湿度传感器、压力传感器、位移传感器、光电传感器等，此外，视频监控摄像头也是物联网的重要组成部分。物联网中的这些设备，每时每刻都在自动产生大量数据，与 Web 2.0 时代的人工数据产生方式相比，物联网中的自动数据产生方式，将在短时间内生成更密集、更大量的数据，使得人类社会迅速步入“大数据时代”。

四、大数据的发展历程

大数据的发展历程总体上可以划分为 3 个重要阶段：萌芽期、成熟期和大规模应用期（表 1-2）。

表 1-2 大数据发展的 3 个时段

进程	时间	详情
第一阶段：萌芽期	1990—2000 年	产生商业智能工具和知识管理的信息技术、知识管理工具和知识管理软件。例如数据仓库、专家系统、知识管理系统开始被应用于市场，说明了数据挖掘理论和数据库技术开始出现
第二阶段：成熟期	2000—2010 年	用传统的方法应对 Web 2.0 应用的迅猛发展非常困难。随着非结构化数据巨量出现，大数据技术开始快速突破，其解决方案也越趋成熟，计算与分布式系统两大技术开始形成，谷歌的 GPS 和 MapReduce 反响强烈，Hadoop 平台也开始广泛应用
第三阶段：大规模应用期	2010 年以后	大数据推进精准决策，社会中各个领域的发展已经离不开大数据。大数据的大规模应用进一步提高了信息社会的智能化

下面是对大数据发展历程的简要回顾。

(1) 全世界极具权威的未来学家有学者在 20 世纪 80 年代就曾提出大数据是“第三次浪潮的华彩乐章”，其著作《第三次浪潮》中对于大数据的赞扬引起世界各国学者的广泛关注和热烈讨论。

(2) 发表在美国计算机学会的数字图书馆的文章提及“大数据”一词。这是迈克尔·考克斯和大卫·埃尔斯沃思在第八届美国电气和电子工程师协会（IEEE）上提出的，并于 1997 年 10 月发表了《为外存模型可视化而应用控制程序请求页面调度》，对大数据技术展开了进一步佐证和阐述。

(3) 1999 年 10 月，在美国电气和电子工程师协会关于可视化的年会上，设置了名为“自动化或者交互：什么更适合大数据？”的专题讨论小组，探讨大数据问题。

(4) 2001 年 2 月, 梅塔集团分析师道格·莱尼发布题为《3D 数据管理: 控制数据容量、处理速度及数据种类》的研究报告。10 年后, “3V” (Volume、Variety 和 Velocity) 作为定义大数据的 3 个维度而被广泛接受。

(5) 2005 年 9 月, 蒂姆·奥莱利发表了《什么是 Web2.0》一文, 并在文中指出“数据将是下一项技术核心”。

(6) 2008 年, 《自然》杂志推出大数据专刊。同时计算社区联盟 (Computing Community Consortium) 发表文章《大数据计算: 在商业、科学和社会领域的革命性突破》, 文章中阐述了大数据技术及其面临的一些挑战。

(7) 2010 年 2 月, 肯尼斯·库克尔在《经济学人》上发表了一份关于管理信息的特别报告《数据, 无所不在的数据》。

(8) 2011 年 2 月, 《科学》杂志推出专刊《处理数据》, 讨论了科学研究中的大数据问题。

(9) 2011 年, 维克托·迈尔·舍恩伯格出版著作《大数据时代: 生活、工作与思维的大变革》, 引起轰动。

(10) 2011 年 5 月, “大数据时代到来”这一概念被提出。麦肯锡全球研究院在文章《大数据: 下一个具有创新力、竞争力与生产力的前沿领域》中阐述了这一概念。

(11) 2012 年 3 月, 美国政府在推出信息高速公路计划之后为进一步适应社会的高速发展, 继而发布“大数据发展计划”这一项国家发展战略。在《大数据研究和发展倡议》中, 美国政府详尽说明了这一举措的必要性和重要意义。

(12) 2013 年 12 月, 中国计算机学会发布《中国大数据技术与产业发展白皮书》, 系统总结了大数据的核心科学与技术问题, 推动了我国大数据学科的建设与发展, 并为政府部门提供了战略性的意见与建议。

(13) 2014 年 5 月, 美国政府发布 2014 年全球“大数据”白皮书《大数据: 抓住机遇、守护价值》, 报告鼓励使用数据来推动社会进步。

(14) 2015 年 8 月, 国务院印发《促进大数据发展行动纲要》, 全面推进我国大数据发展和应用, 加快建设数据强国。

(15) 2016 年 5 月, 工信部在“2016 大数据产业峰会”上透露, 我国将制定出台大数据产业“十三五”发展规划, 有力地推进了我国大数据技术创新和产业发展。

(16) 2017 年 11 月, 我国首个由行业主管协会起草的大数据人才培养发展方向的通识性标准——《中国大数据人才培养体系标准》正式发布, 预示着中国评定大数据人才的核心标准转向是否能够综合使用工具为企业或客户创造商业价值。

(17) 2018 年 1 月, 今日头条大数据公布“中国”成为 2017 年度标题常用字。

第二节 大数据的概念及内涵研究

随着互联网的普及与发展, 世界迎来了数据爆炸的时代。来自企业、互联网以及日常生活的数据日积月累, 形成了巨大的数据海洋。比如, 在常见的商业活动中, 能产生各种销售记录、公司业绩、股票交易等; 在通信行业中, 能产生全球通信网数据、搜索引擎数据、社交网络数据等。这些数据的爆发式增长以及其巨大的潜在价值引起了各界的广泛关注, 也让大数据的分析与挖掘前景呈现出生机勃勃的景象。那么, 具体来说, 什么是大数据? 它来自哪里? 它有什么特点? 它将对企业的业务和营销带来什么影响?

一、大数据的概念

不同领域的组织和专家对大数据的理解略有不同, 但其内在的价值却得到了一致的肯定。“大数据”这一概念的提出并不早, 从 2009 年提出至今, 人们对它的认知都还不够全面, 处于探索阶段。要想充分挖掘出数据中的价值和知识并为人们所用, 了解大数据的基本概念、把握大数据的特征及类型、理解大数据与信息知识的内在联系是最基本的工作。

（一）认识大数据

大数据到底该怎样定义呢？维基百科给出了这样的定义，大数据是一种人工在合理时限内无法处理整合的规模巨大的抽象信息。数据规模如此之大，以至于人类没有办法完成对其进行精准的科学活动。

从宏观角度上来看，连接物理世界、信息空间和人类社会的纽带就是大数据。因为物理世界通过互联网、物联网等技术实现了在信息空间的大数据反映，而人类社会则借助人机界面、脑机界面、移动互联网等手段在信息空间中产生自己的大数据映像。从信息产业的角度来讲，大数据还是新一代信息技术产业的强劲推动力。

“大数据”一词从 2009 年提出以来，在互联网 IT 行业逐渐流行，但仍然没有严谨的定义，这也说明这一概念在数据分析行业具有无限的发展空间，以及无穷的潜在价值。

（二）大数据的来源

随着新一代信息技术的飞速发展和广泛应用，其中互联网、移动互联网、社交网络等技术的迅猛发展带领世界进入了一个大数据的时代。下面从几个方面阐述大数据的来源^①。

1. 来自数据库

从企业的角度来看，企业内部的管理系统，比如企业资源计划（ERP）系统、办公自动化（OA）系统、客户关系管理（CRM）系统等，其产生的数据通过多年的积累和沉淀形成企业内部数据，这些数据在企业决策方面具有重要作用。从传统数据库角度来看，业务系统的运行伴随着数据的产生，数据库便是这些数据的“容器”。

2. 来自物联网

根据维基百科，物联网（The Internet of Things）是一个基于互联网、传统电信网等信息的承载体，让所有能够被独立寻址的普通物理对象实现互联互通的网络。随着传感器、视频以及各种智能设备的发

^① 赵刚. 大数据技术与应用实践 [M]. 北京：电子工业出版社，2013：14-16.