移动互联应用“赢在起点”系列图书

DASHUJU
FENXI YU YINGYONG

# 大数据分析与应用

♦赵守香 唐胡鑫 熊海涛 著

航空工业出版社

移动互联应用“赢在起点”系列图书

# 大数据分析与应用

赵守香　唐胡鑫　熊海涛　著

航空工业出版社

北　京

## 内 容 提 要

本书共分 7 章，主要内容包括：大数据与数据分析、大数据存储、大数据分析工具、大数据与信息安全、基于二部图网络的电子商务推荐算法研究、基于位置的社交网络好友推荐算法研究、基于稀有类分类的信用卡欺诈识别研究。

本书可作为大中专院校计算机、电子商务相关专业的教材，也可供渴望了解大数据知识的人士参考阅读。

**图书在版编目（CIP）数据**

大数据分析与应用 / 赵守香，唐胡鑫，熊海涛著
. -- 北京 : 航空工业出版社，2015.12
ISBN 978-7-5165-0956-2

Ⅰ. ①大… Ⅱ. ①赵… ②唐… ③熊… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2015)第 305424 号

大数据分析与应用
Dashuju Fenxi yu Yingyong

航空工业出版社出版发行
（北京市朝阳区北苑 2 号院 100012）
发行部电话：010-84936597 010-84936343
三河市祥达印刷包装有限公司印刷 全国各地新华书店经售
2015 年 12 月第 1 版 2015 年 12 月第 1 次印刷
开本：787×1092 1/16 印张：20.25 字数：468 千字
印数：1—3000 定价：48.00 元


信息化和网络化的飞速发展使得产生的数据爆炸式地增长，全球信息量预计 2015 年将达 8 ZB，数据量变引起了质变，大数据的概念被提出。

在 IT 业界，有人把大数据产业定义为：“建立在对互联网、物联网等渠道广泛大量数据资源收集基础上的数据存储、价值提炼、智能处理和分发的信息服务业”，或者如 IT 巨头概括大数据战略为：“致力于让所有用户能够从几乎任何数据中获得可转换为业务执行的洞察力，包括之前隐藏在非结构化数据中的洞察力”。

我们认为，大数据分析就是对大量、动态、能持续的数据，通过运用新系统、新工具、新模型的挖掘，从而获得具有洞察力和新价值的东西。

无论是企业、政府还是 IT 行业本身，都面临着大数据时代给它们带来的挑战和机遇。对企业和政府来说，如何充分利用现有的业务数据来获得创新灵感和市场机会，是获得持续竞争力、提升服务水平和质量的重要课题。

本书是北京市社会科学基金项目“大数据背景下的信息安全风险评估与对策研究”（项目编号：14JGA013）的阶段性研究成果，是在对大数据安全风险因素识别研究的过程中，对所收集资料的思考和分析的结果，也是研究团队近三年的研究成果。

本书共分 7 章，分别从不同侧面探讨了大数据的含义、大数据的存储、大数据分析技术、大数据安全问题，以及大数据分析在银行欺诈发现、电子商务推荐中的应用。各章内容具体如下：

- ❖ 第 1 章　大数据与数据分析
- ❖ 第 2 章　大数据存储
- ❖ 第 3 章　大数据分析工具
- ❖ 第 4 章　大数据与信息安全
- ❖ 第 5 章　基于二部图网络的电子商务推荐算法研究
- ❖ 第 6 章　基于位置的社交网络好友推荐算法研究
- ❖ 第 7 章　基于稀有类分类的信用卡欺诈识别研究

其中，第 1、2、4、6 章由赵守香编写，第 3 章由熊海涛老师编写，第 5、7 章由唐胡鑫老师编写。

本书也是 2015 北京工商大学学术专著项目（ZZCB2015-16）的成果，感谢北京工商大学科学技术处对本书出版的大力支持！

在本书的编写过程中，除参考了参考文献中列举的资料外，我们也参考了同行们无私分享在互联网上的大量资料，由于数目众多，不能一一列出，在此一并表示衷心的感谢！

由于水平有限，书中难免存在不足之处，希望广大同行批评指正。

赵守香

2015 年 9 月 30

# 目　录

# 第 1 章　大数据与数据分析

本章从大数据的出现、大数据的影响及大数据对数据处理的要求出发，分析了大数据环境下数据分析与利用的重要性。主要内容包括：

- 大数据产生的背景和特征
- 电子商务的快速发展与大数据
- 物联网的兴起与大数据
- 移动商务的快速渗透与大数据
- 数据分析给企业带来的机会与好处
- 大数据环境下数据分析的需求
- 网站分析的含义及作用
- 大数据与数据分析

## 1.1　概　述

生活在社会、经济与技术革命之中，我们已经将通信、交际、度过闲暇时光、开展业务转移到了互联网上。互联网又渗透进入我们的手机、我们家园和城市中的设备，以及工业经济的工厂中。其导致的数据爆炸和挖掘正改变着我们的世界。

互联网、移动互联网、物联网、云计算的快速兴起，以及移动智能终端的快速普及，正使得当前人类社会的数据增长比以往任何一个时期都要快。数据的爆炸式增长正在出乎人们的想象。据预计，2020 年，全球以电子形式存储的数据量将达 35 ZB，是 2009 年全球存储量的 40 倍。

与此同时，伴随着物联网、移动智能终端和移动互联网的快速发展，移动网络中数据流量的增长速度也非常迅猛。从 2011 年开始，全球移动数据流量年增长率将保持在 50% 以上，将处于一个稳定增长的态势。到 2016 年，全球移动数据流量将达到 2011 年全球移动数据流量的 18 倍，达到 129.6 EB。

数据的疯狂增长，使得适应和应对数据增长成为整个社会关注的焦点。“大数据”的概念正是在这一背景下应运而生的。图 1-1 所示为大数据概览。

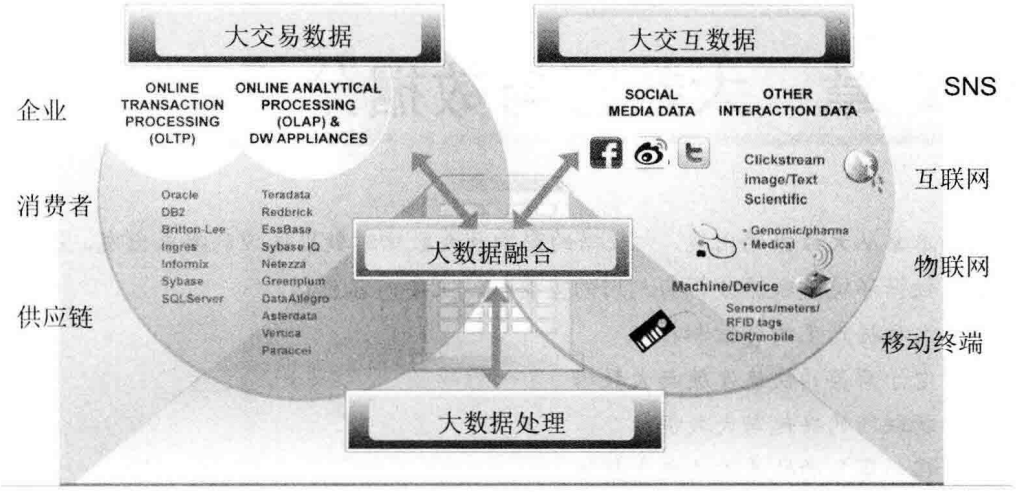


图 1-1　大数据概览


计算机存储容量单位一般用字节（B）、千字节（KB）、兆字节（MB）、吉字节（GB）、太字节（TB）、拍字节（PB）、艾字节（EB）、泽它字节（ZB）、尧它字节（YB）表示。它们之间的换算关系是：

1 KB=1 024 B　　1 MB=1 024 KB　　1 GB=1 024 MB　　1 TB=1 024 GB

1 PB=1 024 TB　　1 EB=1 024 PB　　1 ZB=1 024 EB　　1 YB=1 024 ZB

### 1.1.1　大数据的含义

自从古代有过第一次计数和农作物产量记录以来，数据收集和分析便成为社会功能改进的根本手段。17、18 世纪的微积分、概率论和统计学所提供的基础性工作，为科学家提供了一系列新工具，用来准确预测星辰运动，确定公众犯罪率、结婚率和自杀率。这些工具常常带来惊人的进步。

在 19 世纪，约翰·斯诺（John Snow）博士运用近代早期的数据科学绘制了伦敦霍乱爆发的“群聚”地图。霍乱在过去被普遍认为是由“有害”空气导致的，斯诺通过调查被污染的公共水井进而确定了“霍乱”的元凶，并同时奠定了疾病细菌理论的基础。

今天，数据比以往任何时候都更加深入地与我们的生活交织在一起。我们期待着用数据解决各种问题、改善福利，以及推动经济繁荣。数据的搜集、存储与分析技术不断提升，这种提升看上去正处于一种无限的向上轨迹之中。它们的加速是因为处理器能力的增强、计算与存储成本的降低，以及在各类设备中嵌入传感器的技术的增长。

这些趋势还将持续下去。我们只是处在所谓的“物联网”（The Internet of Things）的相当初级的阶段。在物联网中，我们的各种应用设备、运输工具以及持续增长的“可穿戴”技术产品将可以彼此交换信息。

在 2014 年 12 月 12 日电商的促销期，淘宝网推出“时光机”——一个根据淘宝买家几年来的购买商品记录、浏览点击次数、收货地址等数据编辑制作的“个人网购志”，从而记录和勾勒出让人感怀的生活记忆。背后，是基于对 4.7 亿淘宝注册用户网购数据的分析处理，这正是大数据的典型应用。

随着传统互联网向移动互联发展，全球范围内，除了个人电脑、平板电脑、智能手机、游戏主机等常见的计算终端之外，更广阔的、泛在互连的智能设备，比如智能汽车、智能电视、工业设备和手持设备等都连接到网络之中。基于社会化网络的平台和应用，让数以百亿计的机器、企业、个人随时随地都会获取和产生新的数据。

互联网搜索引擎是大数据最为典型的应用之一。百度日处理数据量达到数十 PB，并呈现高速增长的态势。如果一张光盘容量为 1 GB，这相当于垒在一起的几千万张光盘。微软 Bing（中文名：必应）搜索引擎，一周需要响应 100 亿次量级的搜索请求。通过和 Facebook 的合作，每天有超过 10 亿次的社交网络搜索请求通过 Bing 来处理。

简单地讲，大数据就是那些超过传统数据库系统处理能力的数据。但是，大数据的问题并不仅仅是规模，数据产生的速度以及数据的多样性同样是大数据不可忽略的两个基本特性。根据摩尔定律，计算能力每一年半到两年的时间将增加一倍。可是，现有的网络带宽并没有以同样的速度在增加。因此，如此之迅猛的数据洪流的产生，正在给电信运营商的网络运营带来极大的挑战。

在 IT 业界，有人把大数据产业定义为：建立在对互联网、物联网等渠道广泛大量数据资源收集基础上的数据存储、价值提炼、智能处理和分发的信息服务业。或者如 IT 巨头概括大数据战略为：致力于让所有用户能够从几乎任何数据中获得可转换为业务执行的洞察力，包括之前隐藏在非结构化数据中的洞察力。

总之，大数据是对大量、动态、能持续的数据，通过运用新系统、新工具、新模型的挖掘，从而获得具有洞察力和新价值的东西。

### 1.1.2　大数据的定义

根据维基百科的定义，大数据是指难以用常用的软件工具在可容忍时间内抓取、管理以及处理的数据集。大数据具有数据体量巨大、数据类型繁多、要求的处理速度快等显著特征。

大数据技术涵盖了从数据的海量存储、处理到应用多方面的技术，包括海量分布式文件系统、并行计算框架、NoSQL 数据库、实时流数据处理以及智能分析技术如模式识别、自然语言理解、应用知识库等。

关于“大数据”有许多种定义，这种差别取决于你是一位计算机科学家，还是一位金融分析师，抑或是一位为风险投资人推销一个概念的企业家。多数定义都反映了那种不断增长的捕捉、聚合与处理数据的技术能力，而这个数据集在数量、速率与种类上持续扩大。换言之，数据可以更快获取，有着更大的广度和深度，并且包含了以前做不到的新的观测和度量类型。

更确切地说，大数据集是庞大的、多样化的、复杂的、纵深的和/或分布式的，它由各类仪器设备、传感器、网上交易、电子邮件、视频、点击流，以及现在与未来所有可以利用的其他数字化信号源产生。

### 1.1.3 大数据的特征

虽然有多种解读，但业界一般认为，大数据有 4 个“V”字开头的特征：Volume（容量）、Variety（种类）、Velocity（速度）和最重要的 Value（价值）。

1．Volume（容量）

Volume 是指大数据巨大的数据量与数据完整性。IT 业界所指的数据，诞生不过 60 多年。而一直到个人电脑普及前，由于存储、计算和分析工具的技术与成本限制，许多自然界和人类社会值得记录的信号，并未形成数据。几十年前，气象、地质、石油物探、出版业、媒体业和影视业是大量、持续产出信号的行业，但那时 90%以上采用的是存储模拟信号，难以通过计算设备和软件进行直接分析。拥有大量资金和人才的政府和企业，也只能把少量最关键的信号，进行抽取、转换、装载到数据库中。

尽管业界对达到怎样的数量级才算是大数据并无定论，但在很多行业的应用场景里，数据集本身的大小并不是最重要的，是否完整才最重要。

2．Variety（种类）

Variety 则意味着要在海量、种类繁多的数据间发现其内在关联。互联网时代，各种设备通过网络连成了一个整体。进入以互动为特征的 Web 2.0 时代，个人计算机用户不仅可以通过网络获取信息，还成为了信息的制造者和传播者。这个阶段，不仅是数据量开始了爆炸式增长，数据种类也开始变得繁多。

这必然促使我们对海量数据进行分析、处理和集成，找出原本看来毫无关系的那些数据的“关联性”，把似乎没有用的数据变成有用的信息，以支持我们做出的判断。

不仅是数据的数量正在快速增长，它的格式也越发多样，来源也越发广泛。有些数据是“天生数字化的”（born digital），意思是说它就是特别创造出来用于计算机和数据处理系统的。这些例子存在于电子邮件、网页浏览，或 GPS 定位之中。其他数据是“天生模拟的”（born analog），这是说它从物理世界中发散出来，但可以不断被转化成数字格式。模拟数据的例子包括手机、相机或摄像设备录制的语音或可视信息，或者还有通过可穿戴设备监测到的身体活动数据，如心率或排汗量。“数据融合”（data fusion）能够将分散的

数据源整合在一起，随着这种能力的提升，大数据可以带来一些远见卓识。

3．Velocity（速度）

Velocity 可以理解为更快地满足实时性需求。数据的实时化需求正越来越清晰。对普通人而言，开车去吃饭，会先用移动终端中的地图查询餐厅的位置，预计行车路线的拥堵情况，了解停车场信息甚至是其他用户对餐厅的评论。吃饭时，会用手机拍摄食物的照片，编辑简短评论发布到微博或者微信上，还可以用 LBS（基于位置的服务）应用查找在同一间餐厅吃饭的人，看有没有好友在附近……

如今，通过各种有线和无线网络，人和人、人和各种机器、机器和机器之间产生无处不在的连接，这些连接不可避免地带来数据交换。而数据交换的关键是降低延迟，以近乎实时（这意味着小于 250 毫秒）的方式呈献给用户。

数据采集与分析的执行速度越来越接近即时时间，这意味着对于一个人就其周边环境或生活所做的决定产生即时的影响而言，大数据分析有着越来越大的潜力。高速数据的例子包括记录使用者在线与网页互动活动的点击流数据，即时追踪定位的移动设备获得的 GPS 数据，以及得到广泛分享的社交媒体数据。客户与公司希望通过分析这种数据使其即刻获益的要求越来越高。事实上，如果手机定位应用不能即时准确地确认手机位置，它根本就不会有什么用处，并且，在确保我们的汽车安全运行的计算机系统中，实时操作就至为关键了。

4．Value（价值）

但比前面 3 个“V”更重要的，就是 Value，它是大数据的最终意义——获得洞察力和价值。大数据的崛起，正是在人工智能、机器学习和数据挖掘等技术的迅速发展驱动下，呈现这么一个过程：将信号转化为数据，将数据分析为信息，将信息提炼为知识，以知识促成决策和行动。

就大数据的价值而言，就像沙子淘金，大数据规模越大，真正有价值的数据相对越少。所以真正好的大数据系统，重要的不是越多越好，其实越少越好。开始数据要多，最好还是要少，把 ZB、PB 最终变成一个比特，也就是最后的决策。这才是最关键的。

正如我们常说的：书刚开始越读越厚，到最后就越读越薄了。

数据挖掘和应用可以多方位创造价值：数据可用性可以提高 10%，各行业员工销售额提高百分比统计如图 1-2 所示。图 1-3 展示了海量数据对商业模式的影响。

“卖数据”称为直接赢利模式，例如淘宝推出的“数据魔方”收费标准为 300 元/月，直接创造经济价值。

数据采集、存储与处理成本的下降，连同像传感器、相机、地理位置及其他观测技术提供的新的数据来源，意味着我们生活在一个数据采集几乎无处不在的世界中。采集与处理的数据量是空前的。从基于网络的应用、可穿戴技术与先进传感器到监测生命体征、能源使用状况与慢跑者跑步速度的监测仪，由此带来的数据爆炸将推进人们对于高性能计算技术的需求，并推动针对最复杂数据的管理能力的提升。