大数据之美

挖掘、Hadoop、架构，更精准地发现业务与营销

• 黄宏程　舒　毅　欧阳春　舒　娜　编著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

大数据之美

挖掘、Hadoop、架构，更精准地发现业务与营销

黄宏程　舒　毅　欧阳春　舒　娜　编著

電子工業出版社
Publishing House of Electronics Industry
北京 • BEIJING

内 容 简 介

本书从大数据的基本概念出发，深入解析了大数据应用的关键技术与应用。以大数据的数据挖掘技术、大数据的存储与处理、大数据应用的总体架构三个方面为线索，详细阐述了大数据挖掘的诸多常用算法，介绍了 Hadoop、HDFS 及 MapReduce 等大数据存储与处理的关键技术与应用、大数据应用的框架与构架。本书以通信运营商及互联网电子商务等应用为背景，从典型实例的角度系统地介绍了大数据挖掘应用从目标构建、算法建模到程序实现，再到大数据分析及结果描述应用的整个过程，以期为读者提供从理论到实务的有效借鉴。

本书适合信息产业从事海量信息处理分析的相关工程技术人员、研究人员以及高校师生阅读，也可作为高等院校大数据分析与处理相关课程的教学用书。

图书在版编目（CIP）数据

大数据之美：挖掘、Hadoop、架构，更精准地发现业务与营销 / 黄宏程等编著. —北京：电子工业出版社，2016.8

ISBN 978-7-121-29344-3

Ⅰ. ①大… Ⅱ. ①黄… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2016）第 156113 号

策划编辑：宋　梅
责任编辑：底　波
印　　刷：北京季蜂印刷有限公司
装　　订：北京季蜂印刷有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本：720×1000　1/16　印张：16.25　字数：415 千字
版　　次：2016 年 8 月第 1 版
印　　次：2016 年 8 月第 1 次印刷
印　　数：3000 册　定价：49.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：mariams@phei.com.cn。

前　言

大数据已经成为当今知识信息时代的一个强烈的音符，几乎所有的领域都在寻找着来自大数据的灵感，几乎每个与海量数据相关的应用都有大数据奏响的乐章。

大数据是指那些数据量特别大、数据类别特别复杂的数据集，这些数据无法用传统的数据库进行存储、管理和处理。大数据的主要特点为数据量大（Volume）、数据类别复杂（Variety）、数据处理速度快（Velocity）和数据真实性高（Veracity），合起来称为 4V。

大数据中的数据量巨大，甚至达到了拍字节级别。这些庞大的数据中不仅包括如数字、符号等结构化数据，还包括如文本、图像、声音、视频等非结构化数据。这使得传统的关系型数据库很难满足大数据的存储、管理和处理的需要。在大数据之中，有价值的信息往往深藏其中，这就需要对大数据的处理速度非常快，才能在短时间之内从大量的复杂数据之中获取有价值的信息。在大数据的大量、复杂的数据之中，通常不仅包含真实的数据，虚假的数据也可能混杂其中，这就需要对大数据进行清洗处理，将虚假的数据剔除，利用真实的数据来分析，得出可靠的结果。

大数据表面上看就是大量、复杂的数据，这些数据本身的价值并不高或难以直观获取，但对这些大量、复杂的数据进行分析、处理后，却能从中提炼出很有价值的信息。对大数据的分析主要有数据挖掘算法（Date Mining Algorithms）、分布式计算（Distributed Computing）、预测性分析能力（Predictive Analytic Capabilities）、可视化分析（Analytic Visualization）等。

数据挖掘算法是大数据分析的理论核心，其本质是根据数据处理模型建立起的一组算法，将收集到的数据作为输入，从而能够从大量、复杂的数据中提取有价值的信息。著名的“啤酒和尿布”的故事就是数据挖掘算法的经典案例。沃尔玛通过对啤酒和尿布购买数据的分析，挖掘出以前未知的两者之间的联系，并利用这种联系，提升了商品的销量。淘宝、当当等电子商务系统的推荐引擎和百度

的广告系统都大量使用了数据挖掘算法。

对于如何处理大数据，通常采用分布式计算的方式进行分布式存储和分布式处理。Hadoop 作为大数据处理的杰出代表，成为分布式计算事实上的国际标准，其采用 MapReduce 分布式计算框架，以 HDFS 分布式文件系统作为存储系统，并开发了 HBase 数据存储系统。

预测性分析能力是大数据分析最重要的应用领域。从大量、复杂的数据中挖掘出规律，建立起科学的模型，通过将新的数据输入模型，就可以预测未来的事件走向。预测性分析能力常常被应用在业务分析、辅助决策、科学研究等领域。

可视化分析是普通消费者常常可以见到的一种大数据分析结果的表现形式，可视化分析将大量复杂的数据转化成直观形象的诸如文字、图表等形式，使其能够更加容易地被用户所接受和理解。

本书力图系统地呈现包括数据挖掘算法、Hadoop 大数据存储处理系统等大数据关键技术，并通过通信运营商及互联网电子商务等应用为背景的案例，详尽介绍大数据应用从目标构建、算法建模、程序实现到数据分析与结果呈现的整个过程。

本书由黄宏程、舒毅、欧阳春、舒娜编著，参加编写工作的还有陆卫金、王言通、孙欣然、杨立娜、黄春妮、魏青、冯榆斌。在本书的编写过程中，得到了重庆邮电大学胡敏老师及通信软件工程研究中心的老师和研究生们的诸多帮助，同时也得到了电子工业出版社的大力支持，特表示衷心感谢。本书的部分内容在编著过程中参考了业界的出版物，未能在书中穷尽，在此一并向原作者表示诚挚的感谢！

大数据所涉及的技术内容较多，其发展也非常迅速，由于作者水平有限，书中疏漏之处在所难免，恳请广大读者批评指正。

编著者

2016 年 3 月

目　录

第 1 章　大数据概述 …… 1

1.1　大数据的概念 …… 1

1.1.1　什么是大数据 …… 1

1.1.2　大数据的产生和来源 …… 2

1.1.3　大数据的技术 …… 3

1.1.4　大数据的特征 …… 8

1.1.5　数据、信息与知识 …… 10

1.2　大数据的价值与挑战 …… 10

1.2.1　大数据的潜在价值 …… 11

1.2.2　大数据对业务的挑战 …… 12

1.2.3　大数据对技术架构的挑战 …… 13

1.2.4　大数据对管理策略的挑战 …… 14

1.3　大数据与相关领域的关系 …… 16

1.3.1　大数据与统计分析 …… 16

1.3.2　大数据与数据挖掘 …… 16

1.3.3　大数据与云计算 …… 17

1.4　大数据发展状况 …… 20

参考文献 …… 23

第 2 章　大数据挖掘技术 …… 24

2.1　数据挖掘与过程 …… 24

2.1.1　数据挖掘的七大功能 …… 24

2.1.2　数据挖掘的实质 …… 25

2.2　数据挖掘过程 …… 26

2.2.1　定义挖掘目标 …… 27

2.2.2 数据取样 …… 28
2.2.3 数据探索 …… 30
2.2.4 数据预处理 …… 32
2.2.5 数据模式发现 …… 37
2.2.6 模型评价 …… 40
2.3 常用算法 …… 47
2.3.1 决策树 …… 48
2.3.2 回归 …… 50
2.3.3 关联规则 …… 54
2.3.4 聚类 …… 59
2.3.5 贝叶斯分类方法 …… 66
2.3.6 神经网络 …… 69
2.3.7 支持向量机（SVM） …… 73
2.3.8 假设检验 …… 77
2.3.9 遗传算法 …… 81
参考文献 …… 84

第 3 章 大规模存储与处理技术 …… 86

3.1 Hadoop 概述 …… 86
3.1.1 什么是 Hadoop …… 86
3.1.2 Hadoop 发展简史 …… 88
3.1.3 Hadoop 的优势 …… 90
3.1.4 Hadoop 的子项目 …… 90
3.2 HDFS …… 92
3.2.1 HDFS 的设计目标 …… 93
3.2.2 HDFS 文件系统的原型 GFS …… 93
3.2.3 HDFS 文件的基本结构 …… 95
3.2.4 HDFS 的文件读/写操作 …… 97
3.2.5 HDFS 的存储过程 …… 101
3.3 MapReduce 编程框架 …… 105
3.3.1 MapReduce 的发展历史 …… 105

3.3.2 MapReduce 的基本工作过程 …… 107
3.3.3 MapReduce 的特点 …… 110
3.4 建立 Hadoop 开发环境 …… 111
3.4.1 相关准备工作 …… 111
3.4.2 JDK 的安装配置 …… 113
3.4.3 SSH 无钥登录 …… 113
3.4.4 安装、配置 Hadoop 环境变量 …… 115
3.5 大数据处理系统分类 …… 118
3.5.1 批量数据处理系统 …… 118
3.5.2 流式数据处理系统 …… 119
3.5.3 交互式数据处理 …… 122
3.5.4 图数据处理系统 …… 124
3.6 大数据查询和分析技术：SQL on Hadoop …… 126
3.6.1 数据库简介 …… 126
3.6.2 图数据库 …… 128
3.6.3 Hive：基本的 Hadoop 分析 …… 130
3.6.4 实时互动的 SQL：Impala 和 Drill …… 134
3.7 以通信业务分析为例的大数据的技术环境部署 …… 136
3.7.1 应用架构规划与设计 …… 136
3.7.2 技术环境部署与配置 …… 137

第 4 章 大数据应用的总体架构和关键技术 …… 148

4.1 大数据的业务分析 …… 148
4.2 大数据的总架体构模型 …… 152
4.3 大数据高级分析 …… 161
4.3.1 数据仓库与联机分析处理技术 …… 162
4.3.2 大数据分析与传统分析 …… 167
4.3.3 非结构化复杂数据分析 …… 168
4.3.4 实时预测分析 …… 177
4.4 可视化分析 …… 181
4.4.1 可视化技术 …… 181

4.4.2 可视化工具 …… 192
参考文献 …… 195

第 5 章 运营商数据分析 …… 196

5.1 案例背景 …… 196
5.1.1 大数据运营已为大势所趋 …… 196
5.1.2 采取大数据运营的原因 …… 196
5.1.3 大数据分析如何提升电信行业绩效 …… 197
5.1.4 大数据的社会价值 …… 199
5.2 挖掘目标的提出 …… 200
5.3 案例分析 …… 201
5.3.1 体系架构 …… 201
5.3.2 Hadoop 集群抽取模块 …… 202
5.3.3 数据处理模块 …… 208
5.3.4 数据分发 …… 211
5.4 MapReduce 操作 …… 218
5.5 结果分析 …… 221

第 6 章 互联网电影推荐系统 …… 223

6.1 背景描述 …… 223
6.2 业务目标 …… 224
6.3 业务需求 …… 225
6.4 协同过滤推荐系统建模 …… 225
6.4.1 推荐系统概述 …… 225
6.4.2 基于对立用户的协同过滤模型 …… 227
6.5 项目处理过程 …… 229
6.5.1 项目数据 …… 229
6.5.2 数据预处理 …… 230
6.5.3 Hadoop 并行算法 …… 242
6.6 总结 …… 250

第1章

大数据概述

随着互联网的普及与发展，我们来到了数据爆炸的时代。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，各种销售记录、公司业绩、股票交易等；在通信行业中，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？

1.1 大数据的概念

不同领域的组织和专家对于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我们所用，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。

1.1.1 什么是大数据

根据维基百科，大数据是指所涉及的数据量规模巨大到无法通过人工在合理时间内达到截取、管理、处理，并整理成为人类所能解读的信息。从这个定义中，我们能够看出这是从大数据本身的出发得出的。而网上一些人认为，大数据是在多样的或者大量数据中，迅速获取信息的能力。这显然是从大数据的作用来定义

的。对于大数据定义，要达成共识非常困难。一种逻辑上的选择是接受所有的大数据定义，其中每种定义反映了大数据的特定方面[1]。

从宏观世界角度来讲，大数据是融合物理世界、信息空间和人类社会三元世界的纽带，因为物理世界通过互联网、物联网等技术有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、移动互连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无限的发展空间以及无穷的潜在价值。

1.1.2 大数据的产生和来源

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我们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与产生[2-3]。

1. 来自数据库的大数据

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（EPR）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Web的大数据

在Web 1.0时代，网站服务商提供了大部分Web内容。Web 2.0的发展与盛行带来了数据的爆发性增长，用户通过网页交互大量参与、贡献Web内容，我们从使用数据摇身一变成为数据的生产者。国内的新浪微博、淘宝网、百度，国外的Facebook、Twitter等，每时每刻都有大量的新数据产生。

3. 来自移动互联网的大数据

智能手机的普及使移动互联网占据了我们生活极其重要的一部分，人们通过手机等移动终端获取社会资讯、与其他用户进行交互，随时随地生产数据。据工信部数据显示，截止到2014年1月，移动互联网用户总数达到8.38亿人；移动

互联网接入流量 1.33 亿 GB，同比增长 46.9%；户均移动互联网接入流量达到 165.1MB，其中手机上网流量占比提升至 80.8%，月户均手机上网流量达到 139.3M。这些事实说明移动互联网在互联网中有着举足轻重的地位，而其迅猛的发展趋势也让更多人参与到数据的生产中来。

4. 来自物联网的大数据

根据维基百科，物联网（The Internet of things）是一个基于互联网、传统电信网等信息承载体，让所有能够被独立寻址的普通物理对象实现互连互通的网络。随着传感器、视频以及各种智能设备的发展，其产生的数据是极其庞大的，并且数据的生成方式也有了根本性的不同。

1.1.3 大数据的技术

目前大数据技术主要有大数据科学、大数据工程和大数据应用[4]。大数据科学是通过寻找在大数据网络的快速发展和运营过程中的规律，并用其来验证大数据与社会活动之间的复杂关系；大数据工程是通过规划建设大数据并进行运营管理整个系统。大数据应用主要体现在业务需求方面，而在此之前，大数据需要对大量的数据进行有效处理，其中包括大规模并行处理（MPP）数据库、分布式文件系统、数据挖掘电网、云计算平台、分布式数据库、互联网和可扩展的存储系统。开源与商用两个生态圈是目前用于分析大数据的工具。开源大数据生态圈主要有 Hadoop HDFS、HadoopMapReduce、HBase 等，商用大数据生态圈包括一体机数据库、数据仓库及数据集市。由于大型数据集分析需要大量计算机持续高效分配工作，而大量非结构化数据需要大量时间和金钱来处理分析关系型数据库，因此大数据分析常和云计算联系到一起。相比于传统的大数据分析，目前的大数据分析存在数据仓库数据量大、查询分析复杂问题。

目前大数据把时间作为处理要求，把处理方式分为流处理和批处理[5]。两种处理方式的不同将给相关的平台带来体系结构上的不同。流式处理是指假设数据的潜在价值是数据的新鲜度，因此该处理方式应尽可能快地处理数据并得到相应的结果。在数据连续到达的过程中，由于流携带了大量数据，只有小部分的流数据被保存在有限的内存中。流处理理论和技术的研究已相对成熟，其代表性的开源系统有 Storm、S4 和 Kafka。流处理方式用于在线应用，通常工作在秒或毫秒级别。批处理是指在批处理方式中，数据首先被存储，随后被分析。MapReduce 是非常重要的批处理模型。MapReduce 的核心思想是，数据首先被分为若干小数据块 chunks，随后这些数据块被并行处理并以分布的方式产生中间结果，最后这

些中间结果被合并产生最终结果。由于简单高效，MapReduce 被广泛应用于生物信息、Web 挖掘和机器学习中。

大数据应用使人们的思维不局限于数据处理机器，重要的是新用途和新见解，对大规模信息的处理需求从根本上推动了大数据相关技术的发展，超级计算机的发明、大数据的存储和处理技术以及大数据分析算法的研发最终导致了教育、金融、医疗等多方面大数据广泛应用。

从数据生命周期的角度，从数据源、数据特性等方面总结比较了主要的数据分析方法，包括结构化数据分析、文本分析、Web 数据分析、多媒体数据分析、社交网络数据分析和移动数据分析。企业可以针对自身的需求来应用某种数据分析方法来分析自身拥有的数据，从数据中发现问题，如产品设计问题、运营策略问题、战略规划问题。

1. 结构化数据分析

在科学研究和商业领域产生了大量的结构化数据，这些结构化数据可以利用成熟的 RDBMS、数据仓库、OLAP 和 BPM 等技术管理，而采用的数据分析技术则是前面介绍的数据挖掘和统计分析技术。近年来深度学习（Deep Learning）逐渐成为一个主流的研究热点。许多当前的机器学习算法依赖于用户设计的数据表达和输入特征，这对不同的应用来说是一个复杂的任务。而深度学习则集成了表达学习（Representation Learning），学习多个级别的复杂性/抽象表达。

2. 文本分析

文本数据是信息储存的最常见形式，包括电子邮件、文档、网页和社交媒体内容，因此文本分析比结构化数据具有更高的商业潜力。文本分析又称为文本挖掘，是指从无结构的文本中提取有用信息或知识的过程。文本挖掘是一个跨学科的领域，涉及信息检索、机器学习、统计、计算语言和数据挖掘。大部分的文本挖掘系统建立在文本表达和自然语言处理（NLP）的基础上。文档表示和查询处理是开发矢量空间模型、布尔检索模型和概率检索模型的基础，这些模型又是搜索引擎的基础。NLP 技术能够增加文本的可用信息，允许计算机分析、理解甚至产生文本。词汇识别、语义释疑、词性标注和概率上下文无关文法等是常用的方法。基于这些方法提出了一些文本分析技术，如信息提取、主题建模、摘要（Summarization）、分类、聚类、问答系统和观点挖掘。

3. Web 数据分析

对于互联网企业来说，精通数据分析技术、精通如何监测和测量数据指标，

目前成为企业运营的核心技术，而 Web 数据分析的目标是从 Web 文档和服务中自动检索、提取和评估信息以发现知识，涉及数据库、信息检索、NLP 和文本挖掘，可分为 Web 内容挖掘、Web 结构挖掘和 Web 用法挖掘（Web Usage Mining）。

4. 多媒体数据分析

多媒体数据分析是指从多媒体数据中提取有趣的知识，理解多媒体数据中包含的语义信息。多媒体数据的来源非常丰富，其不再是我们以往认为的图像，而是来源于各种可以产生丰富的图像、视频、语音数据的智能设备。除此之外，还有在现实生活中的各种监控摄像设备、医疗图像设备、物联网传感设备、卫星等都能产生大量的图像、视频数据。因此，多媒体数据在很多领域比文本数据或简单的结构化数据包含更丰富的信息，提取信息需要解决多媒体数据中的语义分歧。以新浪微博为例，用户的微博含有大量的图片、视频等链接，即体现在被大量关注和转发的微博上。而用户对于纯文本的微博信息关注程度比较低。再者，目前微信的使用量居高不下，其主要凭借以语音作为信息载体，改变了以往以纯文本的形式进行社交的方式，使得微信的应用具有一定的竞争优势，现在大家经常能在街上看见用微信来与好友对话的人。为此，多媒体数据分析研究覆盖范围较广，包括多媒体摘要、多媒体标注、多媒体索引与检索、多媒体推荐和多媒体事件检测。

5. 社交网络数据分析

随着在线社交网络的兴起，网络分析从早期的文献计量学分析和社会学网络分析，到 21 世纪的社交网络分析。社交网络包含大量的联系和内容数据，其中联系数据通常用一个图拓扑表示实体间的联系；内容数据则包含文本、图像和其他多媒体数据。显然，社交网络数据的丰富性给数据分析带来了前所未有的挑战和机会。通过对社交网络数据分析，可以发现潜在内部的商机，如发现某个用户的活动商圈是否在企业的商圈覆盖范围内、某个用户的消费能力、用户的兴趣爱好及近期的购买哪些商品的习惯、用户购买自己产品的概率、竞争对手的策略。对结果分析还能达到一个效果预测。这不仅能促进企业的发展，还能给企业带来危机预警，以防止企业遇到危机时无从下手解决。对于危机预警来说，是对一些网络中突然发布的一条可能对企业产生危机的信息即时监控起来，并实时追踪其传播路径，最终找到其中的关键节点。利用“乱石”打散其传播轨迹，从而让危机尽量消失。这就类似舆情爆发和控制，经过对社交网络数据分析来控制这类舆情的爆发。从以数据为中心的角度看，社交网络的研究方向目前主要有基于联系的结构分析和基于内容的分析。今后，社交网络将会成为我们预测未来趋势的有利工具，而企业的发展也将借助对社交网络数据分析来制定更精准、广泛、有效的

社会化营销体系，也能不断提高自己的服务质量。

6. 移动数据分析

随着移动计算的迅速发展，更多的移动终端（手机、传感器和 RFID）和应用逐渐在全世界普及。移动应用是移动互联网的重要载体之一，而移动应用的数据分析是指在获得移动应用的用户使用等基本数据情况下，进行数据分析，深入挖掘用户的使用特点和潜在的价值，从而找到企业产品设计的不足，发现机遇，优化产品及营运策略，提升移动应用的质量。如图 1.1 所示，体现了移动数据分析的意义。移动数据分析的思路是研究者由最初对移动数据分析的研究一头雾水的情况，到觉得研究移动数据分析是件有趣的事的变化过程，也由基础数据分析到深度数据分析的演化。基础数据分析包括用户的新增和启动、活跃分析、时段分析、地域分析、设备机型等；深度数据分析包括用户留存、用户的流失、用户的生命周期、用户的回访次数、日启动次数等。移动数据分析的流程就是一个发现问题、分析问题和解决问题的过程，这与其他大数据分析方法的流程一样。在做移动数据分析之前，必须想好三个问题，如图 1.2 所示。移动数据分析要达到移动应用和产品、运营、市场三者平衡，如图 1.3 所示。2012 年年末，移动数据流量每月达到 885 PB。巨量的数据对移动分析提出了需求，但移动数据分析面临着移动数据特性带来的挑战，如移动感知、活动敏感性、噪声和冗余。

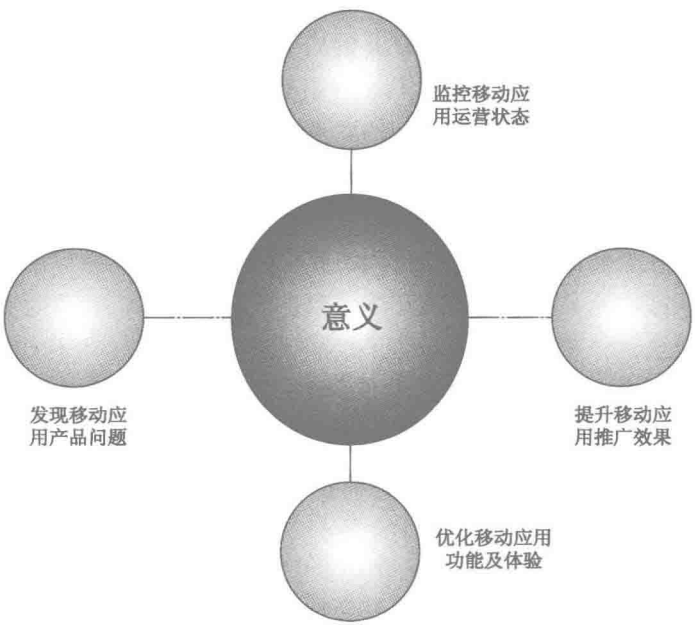

图 1.1　移动数据分析的意义

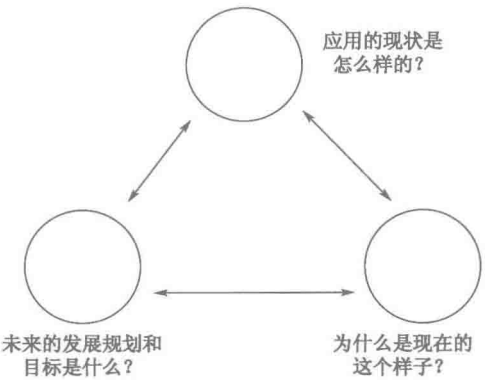

图 1.2　移动数据分析的三个问题

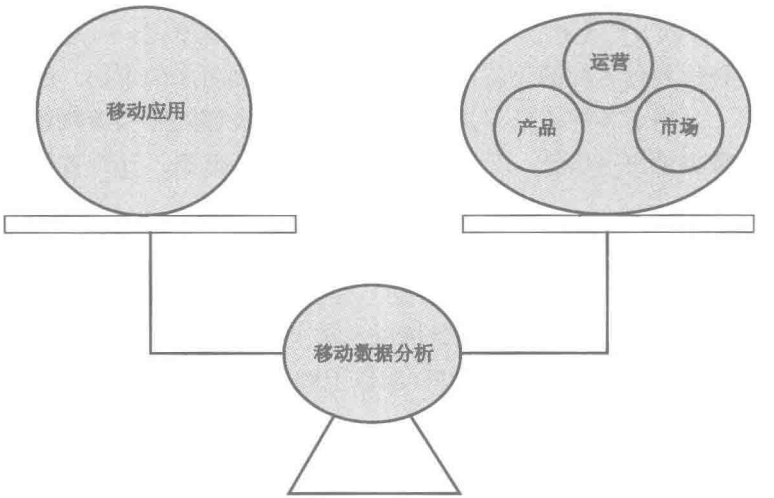

图 1.3　移动数据分析的平衡

移动数据分析的核心是预测。“无尺度网络模型”的作者艾伯特-拉斯洛•巴拉巴西就认为：人类的 93%的行为是可以预测出来的。而数据作为人类活动的痕迹，就像金矿等待发掘。但前提是明确自己的业务需求，这样才能发掘出数据潜在的价值，进而为你所用。

综上几种对大数据的分析方法，数据分析的最终目标都是找出数据背后隐藏的规律，通过利用该规律来进行企业的运营策略以及对目前自己的产品做出一定的定量，也可以预测出未来市场的变化趋势。

1.1.4 大数据的特征

大数据具有4个典型的特征，即通常说的4个V——Volume、Variety、Velocity、Value[1]。从技术研究和开发的角度来看，Volume、Variety、Velocity这3个特征是大数据的根本特点；从商业应用的角度来看，Value才是大数据的核心和关键。大数据的“4V”特征表明了数据量巨大，同时也指出了对于大数据的分析会更加复杂、更加追求速度和更加注重实效。

1. 数据量巨大（Volume）

数据量巨大是大数据和传统数据最显著的区别，它不仅指数据需要的存储空间大，也指数据的计算量巨大，通常可以达到PB级以上的计量，而一般数据的数据量在TB级。产生这么巨大的数据量有多方面的原因：一是由于技术的发展，人们会使用各种各样的设备，使人们能够了解到更多的事物，而这些数据都可以保存；二是由于各种通信工具的使用，使人们能够随时保持联系，这就使得人们交流的数据量快速增长；三是由于集成电路价格低廉，让许多设备都有智能的成分。

数据量的大小间接体现了大数据技术处理数据的能力。数据的基本单位是字节（Byte）。对于传统企业来说，数据量一般在TB级，而对于一些大型企业，比如大型搜索引擎百度、谷歌、新浪微博以及淘宝网等数据量则达到PB级。目前的大数据技术处理的数量级一般指PB级以上的数据。

2. 数据类型多样化（Variety）

大数据拥有多种多样的数据类型，既可以是单一的文本形式或结构化的表单，也可以是半结构化的数据或非结构化的数据，比如视频、图像、语音、网络日志、地理位置信息、订单等。

结构化的数据便于人和计算机对事物进行存储、处理和查询，在结构化的过程中，直接抽取了有价值的信息，而对于新增数据可以用固定的技术进行处理。非结构化的数据由于没有统一的结构属性，导致其在保存数据时还需要保存数据的结构，这就加大了对数据进行存储和处理的困难。目前非结构化的数据已经占了总数据的四分之三以上，而且随着数据的迅猛增长，新的数据类型越来越多，传统的数据处理已经越来越不能满足需求。

大数据不仅量大，并且种类繁多。在这庞大数据量中，五分之四的数据属于非结构化数据，它们来自于物联网、社交网络等各个领域，只有小部分属于结构化的数据。