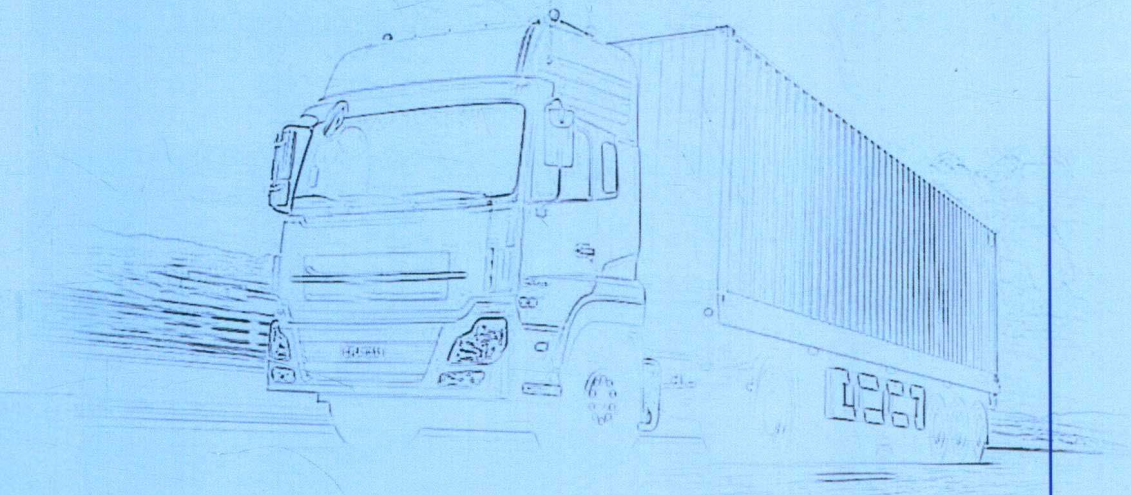
交通物流大数据

决策分析体系研究

任 鹏 丁 然 编著



人民交通出版社股份有限公司
China Communications Press Co.,Ltd.

交通物流大数据

决策分析体系研究

任 鹏 丁 然 编著

人民交通出版社股份有限公司
China Communications Press Co.,Ltd.

内 容 提 要

本书在对国内各类物流信息平台在大数据领域应用实践综合分析的基础上，结合信息系统、决策支持等理论，从背景趋势、体系框架、理论方法、决策分析、应用现状等方面，全面、详细地对交通物流大数据决策分析体系进行了系统介绍。

本书注重理论与实践相结合，既可为政府运用交通物流大数据技术开展政策制定、行业规划、标准规范、市场监管等工作以及市场企业战略规划、运营优化、协同创新等提供借鉴和参考，也可作为高等院校相关专业师生的参考资料。

图书在版编目(CIP)数据

交通物流大数据决策分析体系研究 / 任鹏，丁然编著. — 北京：人民交通出版社股份有限公司，2018.1

ISBN 978-7-114-14342-7

Ⅰ. ①交… Ⅱ. ①任… ②丁… Ⅲ. ①交通运输业—物流—数据管理—研究 Ⅳ. ①F506

中国版本图书馆 CIP 数据核字(2017)第 286016 号

Jiaotong Wuliu Dashuju Juece Fenxi Tixi Yanjiu

书　　名：交通物流大数据决策分析体系研究
著 作 者：任　鹏　丁　然
责任编辑：姚　旭
出版发行：人民交通出版社股份有限公司
地　　址：(100011)北京市朝阳区安定门外外馆斜街 3 号
网　　址：http://www.ccpress.com.cn
销售电话：(010)59757973
总 经 销：人民交通出版社股份有限公司发行部
经　　销：各地新华书店
印　　刷：北京鑫正大印刷有限公司
开　　本：720×960　1/16
印　　张：8.75
字　　数：159 千
版　　次：2018 年 1 月　第 1 版
印　　次：2018 年 1 月　第 1 次印刷
书　　号：ISBN 978-7-114-14342-7
定　　价：25.00 元

Qianyan

前　言

随着国家“互联网＋”发展战略的制定与实施，中国交通物流行业正在经历一场深刻的变革，由以信息共享、数据分析等公益性服务为基础的物流公共信息平台、以“车货匹配”和“无车承运人”为主要服务模式的市场化运作物流互联网平台和以供应链协同、运输组织优化为目的的企业物流信息平台所构成的物流信息体系，正逐步成为现代社会物流运转的基石。利用物联网、北斗导航、移动互联等新技术所采集的物流精细化数据有着极高的分析应用价值，但物流信息资源配置的无序性和自发性所带来的物流信息跨域共享、物流信息资源整合、物流信息采集、物流大数据分析等已成为困扰国家、行业和市场的主要问题。

在此背景下，探索基于各类物流主体需求与实施条件的交通物流大数据信息服务体系以及大数据综合决策分析应用的理论与方法有着极为重要的意义。目前，针对上述问题和研究主要集中在如下三个方面：一是交通物流大数据分析体系的构建，即通过对各类物流信息资源的分布、内容与形式的统筹规划，提升物流信息资源整合水平，为社会经济、行业发展和社会服务决策分析提供有效技术支撑。二是跨区域、跨行业、跨部门物流信息共享能力提升，即通过促进物流透明化、标准化和一体化发展，优化物流资源配置。三是完善物流市场诚信体系建设，促进物流行业规范有序发展。

本书内容建立在多个部级、省级课题研究成果的基础之上，结合相关平台及企业的运营实践，进一步总结凝练而成。全书共分为六章，主要对交通物流大数据的基本概念、背景趋势、体系框架、理论方法、决策分析、应用现状等内容进行了介绍。本书既可为政府运用交通物流大数据技术开展政策制定、行业规划、标准规范、市场监管等工作以及市场企业战略规划、运营优化、协同创新等提供借鉴和参考，也可作为高等院校相关专业师生的参考资料。

本书在编写过程中，得到了交通运输部、浙江省交通运输厅、福建省交通运输厅、贵州省交通运输厅等交通运输行业主管部门的大力支持和帮助，他们提供了很多有益的建议和丰富的案例材料，在此致以衷心感谢。同时，本书借鉴了许多国内外专家学者的研究成果，相关参考文献已在书后列出，在此一并表示感谢。

由于编者水平和时间有限，书中难免有错误与不妥之处，恳请读者批评指正。

编著者

2017 年 10 月

Zuozhe Jianjie

作者简介

任鹏，现就职于交通运输部科学研究院，高级工程师，毕业于同济大学交通运输规划与管理专业，获工学博士学位，主要研究方向为物流信息化、运输组织优化、城市配送、交通运输规划与管理等。近年来主持或参与了国家科技支撑计划、交通运输部软科学重点项目，交通运输战略规划政策研究项目等20余项课题的研究工作，发表各类学术论文30余篇。

丁然，现就职于交通运输部科学研究院，助理研究员，理学硕士，毕业于英国利物浦大学运营与供应链管理专业，主要研究方向为供应链管理、物流信息化、运输组织优化、数据挖掘。近年来承担和参与完成交通运输战略规划政策项目、交通运输部软科学重点项目、交通建设发展前期工作费研究项目、省级交通运输科技项目等省部级以上课题10余项，在国内外核心期刊等发表各类学术论文10余篇。

Mulu

目　录

第一章　大数据基础

第一节　概念与定义

随着移动互联网、物联网、北斗导航等技术和应用的兴起，全球范围内数据量迅猛增长，大数据（Big Data）时代已经来临[1]。早在1980年，阿尔文·托夫勒的《第三次浪潮》中已提出大数据对社会的重要性，但直至21世纪初，学术界和工业界才对其产生关注。2008年 *Nature* 第一次推出 *Big Data* 专刊，对科学研究中的大数据问题展开讨论，*Science* 也在2011年2月推出专刊 *Dealing with Data*。紧随其后，麦肯锡全球研究院（MGI）于2011年6月发布名为 *Big Data: The next frontier for innovation, competition and productivity* 的研究报告，报告中指出数据是新时期的基础生活资料与市场要素，其重要程度不亚于物质资产和人力资本，而大数据则将成为企业提高生产力和竞争力的主要方式与关键要素。数据资产化、产业垂直整合、泛互联网化是大数据时代的三大发展趋势。进入2012年以来，各国对于大数据的关注度与日俱增。2012年3月，美国政府发布了 *Big Data Research and Development Initiative*，正式启动“大数据发展计划”，计划在科学研究、环境、生物医学等领域利用大数据技术进行突破[2,3]。

大数据在全球范围内备受关注，对于大数据的定义也有多种：

IBM 提出“3V”概念，认为大数据具备规模性（Volume）、多样性（Variety）和高速性（Velocity）三个基本特征。其中，规模性是指数据量巨大，量级达到TB级甚至PB级；多样性指数据类型繁多，包括结构化数据和非结构化数据；高速性指数据处理和分析的速度快。在此基础上，其概念可进一步扩充为“4V”。IBM 认为大数据还具有精确性（Veracity），将精确性作为大数据的第四个属性凸显了应对与管理某类型数据中固有的不确定性的重要性；而 IDC 则认为大数据的价值往往呈现出稀疏性，因此价值性（Value）应作为大数据的另一个基本特征。

维基百科对大数据的定义为：“巨量资料，或称大数据，指的是所涉及的资料量规模巨大到无法通过目前主流软件工具，在合理时间内达到撷取、管理、处理并整理成为帮助企业经营决策更积极目的的资讯”。

研究机构 Gartner 给出了这样的定义：“大数据”是需要新处理模式才能具有更

强的决策力、洞察发现力和流程优化能力来适应海量、高增长率和多样化的信息资产。

麦肯锡全球研究院对大数据的定义是：大数据指的是规模超出常规的数据库工具获取、存储、管理和分析能力的数据集。它具有数据规模海量、数据流转快速、数据类型多样和价值密度低四大特征。

大数据技术的战略意义不仅在于掌握庞大的数据信息，更在于对这些含有意义的数据进行专业化处理。换言之，大数据价值体现的关键，不仅在于数据量的大小，更重要的是对数据的分析处理能力。因此，大数据与云计算的关系就像一枚硬币的正反面一样密不可分。大数据必然无法用单台计算机进行处理，而必须采用分布式架构。因此，要对海量的数据进行分布式数据挖掘，必须依托云计算的分布式处理、分布式数据库和云存储、虚拟化技术。

第二节　基本特征

大数据具有数据规模大、数据类型多样、数据处理速度快、数据价值密度低四个基本特征。这些特征指出了大数据与传统数据概念里“海量数据”的区别，即大数据不仅强调数据规模，更重要的是集中体现数据复杂度、处理能力等方面的特征。

1. 数据规模大

聚合在一起的数据量非常大是大数据的首要特征。根据 IDC 的定义，至少要有超过 100TB 的可供分析的数据。导致数据规模激增的原因有很多，首先，在互联网没有广泛应用时，只有少量的机构可以通过调查、取样的方法获取数据，同时发布数据的机构也很有限，人们难以短期内获取大量的数据，而现在用户可以通过网络非常方便地获取数据，用户有意的分享和无意的点击、浏览都可以产生大量数据；其次，早期的单位化数据，对原始事物进行了一定程度的抽象，数据维度低，数据类型简单，多采用表格的形式来收集、存储、整理。数据的单位、量纲和意义基本统一，存储、处理的只是数值而已，因此数据量有限且增长速度慢。而随着应用的发展，数据维度越来越高，描述相同事物所需的数据量越来越大。以当前最为普遍的网络数据为例，早期网络上的数据以文本和一维的音频为主，维度低，单位数据量小。近年来，图像、视频等二维数据大规模涌现，而随着三维扫描设备以及 Kinect 等动作捕捉设备的普及，数据真实度大幅提升，数据的描述能力不断增强，使得数据量呈几何级数增长。此外，数据量大的另一个重要因素是人们处理数据的方法和理念发生了根本的改变。早期，人们对事物的认知受限于获取、分析数据的能力，人们一直利用采样的方法，以少量的数据来近似描述事物的全貌，样本的数

量可以根据数据获取、处理能力来设定。不管事物多么复杂，通过采样得到部分样本，使得数据规模变小，然后利用当时的技术手段来进行数据管理和分析。随着技术的发展，样本数量逐渐逼近原始的总体数据，且在某些特定的应用领域，采样数据可能远不能描述整个事物，甚至可能丢掉大量重要细节，致使得到完全相反的结论。因此，使用更多的数据可以带来更高的精确性，从更多的细节来解释事物属性，这就必然使得要处理的数据量显著增多[4]。

2. 数据类型多样

数据类型繁多、复杂多变是大数据的重要特性。以往的数据尽管数量庞大，但通常都是事先定义好的结构化数据❶。处理此类结构化数据，只需事先分析好数据的意义以及数据间的相关属性，通过构造表结构来表示数据的属性。结构化数据都以表格的形式保存在数据库中，数据格式统一，以后不论再产生多少数据，只需根据其属性，将数据存储在合适的位置，就可以方便地查询、处理，一般不需要为新增的数据显著更改数据聚集、查询、处理方法，限制数据处理能力的只是运算速度和存储空间。这种只关注结构化信息，强调大众化、标准化的属性使得处理传统数据的复杂程度呈线性增长，但新增的数据可以通过常规的技术手段处理。而随着互联网与传感器的飞速发展，非结构化数据大量涌现。非结构化数据没有统一的结构属性，难以用表结构来表示，在记录数据数值的同时还需要存储数据的结构，增加了数据存储、处理的难度。目前，网络上流动着的数据大部分是非结构化数据，人们上网不只是看看新闻，发送文字邮件，还会上传、下载照片、视频，发送微博等，产生大量非结构化数据。同时，遍及工作、生活中各个角落的传感器也不断产生各种半结构化、非结构化数据，这些结构复杂、种类多样，同时规模又很大的半结构化、非结构化数据逐渐成为主流数据[5]。截至2016年底，非结构化数据量已占到数据总量的75%以上，且非结构化数据的增长速度比结构化数据快10～50倍。此外，大数据与传统数据处理最大的不同就是重点关注非结构化信息，大数据关注包含大量细节信息的非结构化数据，强调小众化、体验化的特性使得传统的数据处理方式面临巨大的挑战。

3. 数据处理速度快

要求数据的快速处理，是大数据区别于传统海量数据处理的重要特性之一。随着各种传感器和互联网等信息获取、传播技术的飞速发展、普及，数据的产生、发布越来越容易，产生数据的途径增多，数据呈爆炸式快速增长。新数据的不断涌现，必然要求数据处理的速度相应提升。这样，大量的数据才能得到有效的利用，

❶结构化数据是将事物向便于人类和计算机存储、处理、查询的方向抽象的结果。在抽象的过程中，忽略一些在特定的应用下可以不考虑的细节，抽取有用的信息。

否则,不断激增的数据不但不能为解决问题带来优势,反而成为快速解决问题的负担。同时,数据并非静止不动,而是在互联网络中不断流动,且数据的价值随着时间的推移而迅速降低,如果数据未得到及时、有效的处理,就失去了价值,大量的数据就没有意义。此外,在许多应用中要求能够实时处理新增的大量数据,例如大量在线交互的电子商务应用要求很强的时效性。大数据以数据流的形式产生,快速流动、迅速消失,且数据流量通常不是平稳的,会在某些特定的时段突然激增,数据的涌现特征明显,而用户对于数据的响应时间通常非常敏感。心理学实验证实,从用户体验的角度来说,瞬间(moment,3s)是可以容忍的最大极限,对于大数据应用而言,很多情况下都必须要在1s内形成结果,否则处理结果就是过时和无效的。因此,对不断激增的海量数据进行实时处理,是大数据与传统海量数据处理技术的关键差别之一。

4. 数据价值密度低

数据价值密度低是大数据关注的非结构化数据的重要属性。传统的结构化数据,依据特定的应用,对事物进行了相应的抽象,每一条数据都包含该应用需要考量的信息,而大数据为了获取事物的全部细节,不对事物进行抽象、归纳等处理,直接采用原始的数据,保留了数据的原貌,且通常不对数据进行采样。由于减少了采样和抽象,大数据呈现所有数据和全部细节信息,虽然可以分析更多的信息,但也引入了大量没有意义的信息,甚至是错误的信息。因此,相对于特定的应用,大数据关注的非结构化数据的价值密度偏低。以当前广泛应用的视频监控为例[6],在连续不间断的监控过程中,大量的视频数据被存储下来,对于某一特定的应用,比如获取犯罪嫌疑人的体貌特征,有效的视频数据可能仅仅有一两秒,大量不相关的视频信息增加了获取这有效的一两秒数据的难度。但是,大数据的数据密度低是指相对于特定的应用,有效的信息相对于数据整体是偏少的,信息有效与否也是相对的,对于某些应用是无效的信息对于另外一些应用则可能成为最关键的信息。此外,数据的价值也是相对的,有时一条微不足道的细节数据可能造成巨大的影响,例如一条几十个字符的微博,就可能通过转发而快速扩散,导致相关的信息大量涌现,其价值不可估量。因此,为了保证对于新产生的应用具有足够多的有效信息,通常必须保存所有数据,这样一方面使得数据的绝对数量激增,一方面使得数据包含有效信息量的比例不断减少,数据价值密度降低。

第三节　应用现状

大数据应用是利用大数据分析的结果,为用户提供辅助决策,挖掘潜在价值的过程。早在大数据概念提出之前,以数据分析为驱动的应用演变就已广泛地应用

于各个领域，在结构化数据分析、文本分析、网站分析、多媒体分析、网络分析和移动分析构成的六个关键分析领域产生了巨大影响。如今，大数据已在网络通信、医疗卫生、农业研究、金融市场、气象预报、交通管理、新闻报道等领域得到广泛应用，极大地促进了各行各业的发展，体现出其巨大的发展潜力及应用价值。

一、数据分析演化

20 世纪末，以数据分析为驱动的应用程序被广泛地应用于各个领域。例如，20 世纪 90 年代，“商业智能”就成为一个在商界流行的术语，21 世纪早期便出现了基于海量数据挖掘处理的网站搜索。数据分析的应用演变在不同领域都体现出了巨大潜力和应用价值，且产生了极大的影响。

1. 商业应用的演变

最早的商业数据通常为结构化数据，各公司从其旧有系统中收集这些数据并把它们存储到关系型数据库管理系统中。这些系统中使用的分析技术在 20 世纪 90 年代非常流行，通常都很直观、简单，例如报表、仪表盘、条件查询、基于搜索的商业智能、联机事务处理、交互式可视化、记分卡、预测建模、数据挖掘等[7]。自 21 世纪以来，互联网和网站给各类组织机构提供了一个在线展示其业务并和客户直接互动的机会。大量的产品和客户信息，包括点击流数据日志、用户行为等，均可以从网站上获取。这样，通过各种文本和网站挖掘分析技术就可以实现产品布局优化、客户交易分析、产品建议和市场结构分析。

2. 网络应用的演变

早期的网络主要提供电子邮件和网页服务，文本分析、数据挖掘和网页分析技术也相应地用于挖掘电子邮件内容、构建搜索引擎等，网络数据占据了全球数据量的大多数。如今，Web 的应用愈发广泛，且充满各种不同类型的数据，例如文本、图像、视频、照片和交互内容等。大量用于半结构化或非结构化数据的高级技术应运而生，例如，利用图像分析技术可以从照片中提取有用的信息进行面部识别等，这种多媒体分析技术可以应用于商业、执法和军事应用中的自动化视频监控系统中。2004 年后，在线社交媒体，如论坛、网上群体、网络博客、社交网站、社交多媒体网站等，为用户创建、上传并分享内容提供了更为便捷的方式，社交数据开始呈现爆发式增长。此外，网络应用所产生数据不再仅源自互联网，移动网络和物联网也成为网络数据的重要来源。据相关文献[8]，2011 年，移动电话和平板电脑的数量第一次超过了笔记本电脑和个人计算机的数量，移动电话和基于传感器的物联网正在开启新一轮网络应用的演变。

3. 科学应用的演变

许多领域的科研都在通过高通量传感器和仪器获取大量数据，从天体物理学

和海洋学,到基因学和环境研究,无不如此。美国国家科学基金会(NSF)此前公布了 BIGDATA 方案,以利于信息共享和数据分析。目前,一些学科已经开发了海量数据平台,并取得了相应的收益。例如,在生物学中,iPlant 正应用网络基础设施、物理计算资源、协作环境、虚拟机资源、可互操作的分析软件和数据服务来协助研究人员、教育工作者和学生建设所有的植物学科。iPlant 数据集形式变化多端,其中包括规范或参考数据、实验数据、模拟和模型数据、观测数据以及其他派生数据。几种具有代表性的大数据应用领域及其特征见表 1-1。

典型的大数据应用领域及其特征　　表 1-1

应用领域	实例	用户数量	反应时间	数据规模	可靠性	准确性
科学计算	生物信息	小	慢	TB	适中	很高
金融	电子商务	大	非常快	GB	很高	很高
社交网络	Facebook	很大	快	PB	高	高
移动数据	移动电话	很大	快	TB	高	高
物联网	传感网	大	快	TB	高	高
Web 数据	新闻网站	很大	快	PB	高	高
多媒体	视频网站	很大	快	PB	高	适中

二、主要应用领域

1. 电子政务

大数据的发展,极大改变了政府现有的管理模式和服务模式。具体而言,就是依托大数据,减少政府投入,提升公共服务能力,及时有效地进行社会监管和治理。以大数据应用支撑政务活动为例,美国积极运用大数据推动政府管理方式变革和管理能力提升,越来越多的政府部门依托数据及数据分析进行决策,将之用于公共政策制定、舆情监控、犯罪预测、反恐等活动中。例如,作为大数据的强力倡导者,奥巴马及其团队创新性地将大数据应用到竞选活动中,通过对近 2 年搜集、存储的海量数据进行分析挖掘,寻找和锁定潜在的己方选民,运用数字化策略定位拉拢中间派选民及筹集选举资金,成为将大数据价值与魅力发挥得淋漓尽致的典型。此外,借助大数据,还能逐步实现立体化、多层次、全方位的电子政务公共服务体系,推进信息公开,促进网上电子政务开展,创新社会管理和服务应用,增强政府和社会、百姓的双向交流和互动[9]。

2. 网络通信业

大数据与云计算相结合所释放出的巨大能量,几乎波及所有的行业,在信息、互联网和通信产业尤为突出。特别是通信业,在传统语音业务低值化、增值业务互联网化的趋势中,大数据与云计算有望成为其加速转型的动力和途径。对于大数

据而言，信息已经成为企业战略资产，市场竞争要求越来越多的数据被长期保存。并且，从管道、业务平台、支撑系统中能够不断产生海量有价值的数据，基于这些大数据的商业智能应用会为通信运营商带来了巨大机遇和丰厚利润。例如，电信业者可通过数以千万计的客户资料，分析出多种使用者行为和趋势，将其卖给需要的企业，这便是全新的资源经济；又如中国移动通过大数据分析，可以对企业运营的全业务进行针对性监控、预警、跟踪，使系统在第一时间自动捕捉市场变化，再以最快捷的方式推送给指定负责人，使其在最短时间内获知市场行情。

3. 医疗卫生行业

伴随医疗卫生行业信息化进程的发展，医疗业务活动、健康体检、公共卫生、传染病监测、人类基因分析等医疗卫生服务过程也会产生海量高价值的数据。这些数据内容主要包括：医院的 PACS 影像、B 超、病理分析、大量电子病历、区域卫生信息平台采集的居民健康档案、疾病监控系统实时采集的数据等[10]。面对大数据，医疗行业遇到前所未有的挑战和机遇。例如，Seton Healthcare 是采用 IBM 最新沃森技术医疗保健内容分析预测的首个客户。该技术允许企业找到大量病人相关的临床医疗信息，通过大数据处理，更好地分析病人的信息。在加拿大多伦多的一家医院，针对早产婴儿，每秒钟有超过 3000 次的数据读取[11]。通过对这些数据分析，医院能够提前知道哪些早产儿可能出现问题并且有针对性地采取措施，避免早产儿夭折。大数据可以让更多的创业者更方便地开发产品，比如通过社交网络来收集数据的健康类 APP。又如，社交网络为许多慢性病患者提供临床症状交流和诊治经验分享平台，医生借此可获得在医院通常得不到的临床效果统计数据；基于对人体基因的大数据分析，可以实现对症下药的个性化治疗；对于公共卫生部门，可以通过全国联网的患者电子病历库，快速检测传染病，对疫情进行全面监测，并通过集成的疾病监测和响应程序，快速进行响应。

4. 能源行业

能源勘探开发涉及的数据类型众多，不同类型数据包含的信息各具特点，只有综合各种数据所包含的信息才能得出真实的地质状况。能源行业企业对大数据产品和解决方案的需求集中体现在：可扩展性、高带宽、可处理不同格式数据的分析方案。例如，智能电网现在欧洲已经做到了终端，也就是所谓的智能电表。在德国，为了鼓励利用太阳能，会在家庭安装太阳能，除了卖电给用户，还可在用户太阳能有多余的时候买回来。通过电网每隔 5 min 或 10 min 收集一次数据，收集来的这些数据可以用来预测客户的用电习惯等，从而推断出在未来 2 ~ 3 个月时间里，整个电网大概需要多少电。预测后，就可以向发电或供电企业购买一定数量的电，进而降低采购成本[11]。又如维斯塔斯风力系统，依靠的就是 BigInsights 软件和 IBM 超级计算机，然后对气象数据进行分析，找出安装风力涡轮机和整个风电场的

最佳地点。以往需要数周的分析工作,利用大数据,现在仅需要不到 1 h 便可完成。

5. 零售行业

根据 IDC 和 MGI 对大数据研究结果的总结,大数据主要能在以下 4 个方面挖掘出巨大的商业价值:对顾客群体细分,然后对每个群体量体裁衣般地采取独特的行动;运用大数据模拟实境,发掘新的需求和提高投入的回报率;提高大数据成果在各相关部门的分享程度,提高整个管理链条和产业链条的投入回报率;进行商业模式、产品和服务的创新。

在商业领域,沃尔玛公司每天通过 6000 余个商店,向全球客户销售超过 2.67 亿件商品,为了对这些数据进行分析,HP 公司为沃尔玛公司建造了大型数据仓库系统,数据规模达到 4 PB(1024 GB = 1 TB,1024 TB = 1 PB),并且仍在不断扩大[11]。沃尔玛公司通过分析销售数据,了解顾客购物习惯,得出适合搭配在一起出售的商品,而且还可从中细分顾客群体,为不同群体的顾客提供个性化服务。在金融领域,华尔街德温特资本市场公司通过分析 3.4 亿条微博账户留言,判断民众情绪,依据人们高兴时买股票、焦虑时抛售股票的规律,决定公司股票的买入或卖出。阿里巴巴公司根据在淘宝网上中小企业的交易状况筛选出财务健康和讲究诚信的企业,对他们发放无需担保的贷款。当我们去购物时,系统会结合历史购买记录和社交媒体数据来为我们提供优惠券、折扣和个性化优惠。零售企业也可通过监控客户的店内走动情况以及与商品的互动,将这些数据与交易记录相结合来展开分析,从而在销售哪些商品、如何摆放货品以及何时调整售价上作出决策。此类方法已经帮助某些领先运用大数据的零售企业减少了 10% 以上的存货,同时在保持市场份额的前提下,增加了高利润率自有品牌商品的比例。

6. 气象行业

与世界大数据时代的进程相同,气象数据量成几何倍数增长。目前,每年的气象数据已接近 PB 量级。以气象卫星数据为例,虽然气象卫星是用来获取与气象要素相关的各类信息的,然而在森林草场火灾、船舶航道浮冰分布监测等方面,气象卫星却同样也能发挥出跨行业的实时监测服务价值。气象卫星、天气雷达等非常规遥感遥测数据中包含的信息十分丰富,有可能挖掘出新的应用价值,从而拓展气象行业新的业务领域和服务范围。比如,可以利用气象大数据为农业生产服务。美国硅谷有家专门从事气候数据分析处理的公司,从美国气象局等数据库中获得数十年来的天气数据,然后将各地降雨、气温、土壤状况与历年农作物产量的相关度做成精密图表,可预测各地农场来年产量和适宜种植的品种,同时向农户出售个性化保险服务。气象大数据应用还可在林业、海洋、气象灾害等方面拓展新的业务领域。

除了上述行业应用外，大数据在教育科研、生产制造、金融保险、交通运输等行业也有密切应用。大数据在金融行业可用于客户洞察、运营洞察和市场洞察；在智能交通、智慧城市建设方面也有出色表现。随着社会、经济的发展，各行业各类用户对于智能化的要求将越来越高，今后大数据技术会在越来越多领域得到广泛应用。通过对大数据的采集、存储、挖掘与分析，大数据在营销、行业管理、数据标准化与情报分析和决策等领域将大有作为，并将极大提升企事业单位的信息化服务水平。随着云计算、物联网、移动互联网等技术的快速发展，大数据未来发展空间将更加广阔。

第四节　发展趋势

在过去的几年中，大数据发展非常迅速，其发展和智能手机的普及有着紧密的关系。同时，物联网的浪潮正在酝酿之中，线上与线下的接合将带来更深度的数据关联，涉及消费者全渠道的行为收集。

大数据的长期发展过程，主要体现出以下六大趋势：应用无线化、信息数据化、交易无纸化、人工智能化、决策实时化、线下线上化。以上六大趋势和我们的生活密切相关，并已经成为我们现实生活的一部分。应用无线化，提供了更大的便利性与移动性，让终端设备与资料采集的作业可以更为弹性而有效率；信息数据化，让信息的流通、交换、加工、运用更趋标准化及结构化，数据处理时代数据的应用变得更即时、直接；交易无纸化，彻底改变了交易行为与资金流的流向，并赋予未来微经济商业模式更多创新思考的可能性；人工智能化，描绘了大数据所产生的创新价值如何与人类交互并深入生活之中，使人的思维与新科技产生前所未有的碰撞；决策实时化，透过大数据实时采集及加工改变了决策与信息的关系；线下线上化，也就是大家热议的全渠道议题，未来仍将呈现线下更多运用线上数据的趋势，线上与线下连接在一起，不可分割。

上述六大趋势会在各自的体系内不断深化发展与创新，而未来，大数据的发展将集中在两个方向：第一，其价值会体现在各行业当中，数据技术会成为各行各业的优化工具或催生其颠覆性创新；第二，大数据本身的发展也会被自我颠覆，数据的采集、更新、识别、关联将会变得越来越自动化。未来，我们每个人都将离不开大数据，必将生活在大数据时代里，大数据资源也将真正成为企业和国家的核心竞争力。

大数据技术的缘起，可以追溯到2004年谷歌公司提出的MapReduce模型[12]。在十几年的时间里，大数据技术从概念走向应用，形成了以Hadoop为代表的一整套技术。时至今日，大数据技术仍在快速发展之中，无论是基础框架、分析技术和

应用系统都在不断演变和完善。据统计,2015 年美国大数据初创企业获得的融资额达到 66.4 亿美元,占整个技术领域总融资额的 11%。这代表着大数据领域具有蓬勃的活力并受到市场的肯定。大数据技术的发展方向是技术发展与应用需求相互推进的结果,对大数据技术发展趋势进行分析,有助于从更本质的层面理解这个领域的现状。

1. 基础架构

历经多年发展,大数据基础设施正在向着快速、便捷与整合的方向发展。Hadoop 框架是大数据分析的重要基础框架,但它存在着计算速度慢、运维复杂等问题。基于 Hadoop 衍生出了如 Spark、Pig 等框架,正在不断提升计算性能和优化处理流程。与 Hadoop 相比,Spark 的抽象层次更高,计算速度更快,编程更加简便。更重要的是,Spark 提供了统一的数据平台,通过不同的模块支持了不同类型的数据应用。如通过 Spark Core 支持批处理;通过 Spark SQL 支持数据交互;通过 Spark Streaming 支持流式存储;通过 MLLib 支持机器学习;通过 GrphaX 支持图计算等。

在大数据基础设施中,各种新技术不断产生,数据湖(Data Lake)和雾计算(Fog Computing)分别从数据的集中与分布的不同角度给出了解决方案。数据湖是大型的基于对象的存储库,数据以其原始格式存储,不需要对数据进行转换,就可以进行全面的监控和分析,并建立数据模型。与一般意义的数据汇聚不同,数据湖不需要改变原始数据的结构,而是支持分析原始数据。这个方式消除了数据抽取、转换和加载 ETL 的成本。为了达到不改变数据结构而直接存储和技术的目标,数据湖对元数据有很高的要求。目前,数据湖技术仍在起步阶段,还存在原始数据差别大、类型复杂、分析应用困难等问题,但它有助于企业完成更长远的数据规划,建立数据治理结构,并预先解决安全问题[13]。数据湖与一般大数据汇集方式的对比见表 1-2。

数据湖与一般大数据汇集方式对比　　表 1-2

属性 / 类别	汇聚方式	元数据要求	数据结构	计算	安全性
数据湖	原始格式汇聚	高,需要通过元数据实现聚合	结构化、非结构化整合	不移动数据	低
一般大数据	需花大量时间转换格式	低	结构化、非结构化分立	移动数据	高

与数据湖侧重数据的聚集不同,雾计算则提出了一种分布式解决方案。雾计算这一名词最早来自网络安全领域,后来由思科(Cisco)公司借用,并赋予了分布式计算的含义。思科将雾解释为“更贴近地面的云”,即雾计算是云计算的延伸。与云计算不同,雾计算并非由性能强大的服务器组成,而是由性能较弱、更为分散

的各类计算模块和智能网络设置组成，这些低延迟且有能力进行位置感知的模块可以融入各类基础设施，乃至生活用品之中[14]。

可以预见，随着物联网的不断发展，来自各类终端的数据量会激增。面对这一情况，云计算的瓶颈可能会凸显。在雾计算中，数据、分析和应用都集中在网络的终端节点，只在需要的时候汇集到云中。云计算与雾计算的对比见表1-3。

云计算与雾计算特性对比　　表1-3

属性 / 类别	计算单元	计算能力	节点数量	分布情况	数据存储	延迟	位置感知
云计算	服务器	强	较少	物理集中与分散并存	全部存储于云端	较高延迟	无
雾计算	各类处于网络边缘设备	弱	较多	物理分散	存储于网络边缘设备	低延迟	有

雾计算将计算能力延伸到了网络边缘的各类智能设备。在这种模式下，智能设备的管理与交互就变得非常重要。例如，比特币的底层技术“区块链”(Block Chain)形成了行动登记、权属确认和智能管理模式，这为通过网络实现各种智能终端和设备实现自我管理和智能交互提供了新的技术支持[15]。数据湖和雾计算着眼于大数据的源头和终端，从分布和集中两个角度提供了解决方案。诚然，这些方案需要通过实践进行检验。但总体而言，数据湖和雾计算代表着大数据分析基础设施的发展趋势，即采用更灵活的方式获取和处理终端数据，合理分布计算负载，对核心数据进行广泛汇集，通过制定标准实现数据治理。

2. 分析技术

分析技术是基于大数据进行模型构建，并进行评价、推荐和预测等具体应用的基础。大数据分析技术在近年得到快速发展，智能化、实时化和易用性是其主要特征。

(1)智能化。

在分析技术方面，大数据与机器学习相结合形成的新型人工智能，已经成为近年最为引人瞩目的项目。大数据与机器学习正使得数据分析在统计分析的基础上，更快速地实现智能关系发现和预测，如图1-1所示。AlphaGo就是这一趋势的典型应用范例。在海量数据的基础上，以深度学习为代表的创新算法，通过大规模并行计算，不断迭代演化，最终形成了能够战胜人类的数据智能。

大数据与机器学习整合所实现的人工智能，其意义不限于特定的领域应用，而是实现了一般性人工智能技术的突破。这一突破将在以医疗、交通、金融和教育等为代表的各个应用领域产生重大影响。从更为广阔的角度，以智慧城市为代表的智能化系统解决方案，预示着智能化大数据技术综合应用的未来前景。由各类设