高职高专旅游大类专业新形态教材

旅游与酒店业大数据应用

黄　昕　张　峰　主　编
黄婉敏　张芷晴　副主编

清华大学出版社
北　京

内 容 简 介

本书系统阐述了培养具备大数据思维和大数据应用能力、满足旅游及酒店业数字化转型升级的高素质技术技能人才所需要的大数据相关知识。全书分为导论篇和实践篇两部分，介绍了大数据基本概念、大数据思维、大数据安全与伦理、大数据处理与分析、大数据在行业中的主要应用情况等内容。本书融入了通俗易懂的案例，设计了丰富的实训任务，能够让学习者直观感受并清晰掌握相关理论知识和实践技能。

本书适用于高等职业院校旅游大类专业大数据应用与分析相关课程教学，也适合旅游行业人士阅读参考。

图书在版编目(CIP)数据

旅游与酒店业大数据应用/黄昕，张峰主编. —北京：清华大学出版社，2022.10
高职高专旅游大类专业新形态教材
ISBN 978-7-302-61933-8

Ⅰ.①旅… Ⅱ.①黄… ②张… Ⅲ.①数据处理－应用－旅游业－高等职业教育－教材 ②数据处理－应用－饭店业－高等职业教育－教材 Ⅳ.①F59-39 ②F719.3-39

中国版本图书馆 CIP 数据核字(2022)第 176321 号

责任编辑：刘士平
封面设计：傅瑞学
责任校对：袁 芳
责任印制：朱雨萌

出版发行：清华大学出版社
网 址：http://www.tup.com.cn，http://www.wqbook.com
地 址：北京清华大学学研大厦 A 座　　邮 编：100084
社 总 机：010-83470000　　邮 购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-83470410
印 装 者：三河市铭诚印务有限公司
经 销：全国新华书店
开 本：185mm×260mm　　**印 张**：19.25　　**字 数**：461 千字
版 次：2022 年 10 月第 1 版　　**印 次**：2022 年 10 月第1次印刷
定 价：58.00 元

产品编号：096949-01

前　言

随着大数据技术的日益成熟，旅游与酒店业的生产和消费方式也在不断地发生改变。大数据作为一种新的生产要素，正推动着旅游与酒店业以前所未有的创新方式进化，大数据的各种应用已经融入数字经济新格局下的旅游与酒店业的方方面面中。然而，当前具备大数据思维和大数据应用能力的旅游与酒店行业人才却十分缺乏，无法满足行业数字化转型升级对高素质技术技能人才的需求。

2021 年 8 月，全国旅游职业教育教学指导委员会启动了高等职业教育本科、高等职业教育专科、中等职业教育的专业教学标准修制订工作。新的旅游类职业教育国家专业教学标准要求适应产业优化升级需要，对接产业数字化、网络化、智能化发展新趋势，对接新产业、新业态、新模式下相关职业和岗位群的新要求，不断满足产业高质量发展对高素质技术技能人才的需求，推动职业教育专业内涵升级，提高人才培养质量。

升级后的专业目录和新研制的国家教学标准对教材提出了新的要求，需要对接新专业目录、新专业内涵，适应职业教育教学改革需求，吸收行业发展的新知识、新技术、新工艺、新方法，校企合作开发专业课教材。本书的编写团队是一支具备高等教育背景、行业数字化运营背景、大数据技术研发背景的专业团队。本书主编黄昕来自广东海洋大学商学院，主编张峰和副主编黄婉敏来自广州市问途信息技术有限公司（以下简称问途）。

问途是酒店与旅游业数字化营销技术研发与实践领域的开拓者。在教育领域，问途深度参与旅游教育数字化升级的工作，致力于将行业先进数字化技术与教育数字化技术结合，以创新教育理念和前沿技术为高等院校培养数字旅游时代的新人才创造新环境，构建产教融合新生态，以人才赋能支持中国酒店与旅游业的数字化升级。2021 年，问途创始人受邀成为旅游类专业教学标准修制订内审会专家组成员，参与旅游类高等职业教育本科和中等职业教育专业教学标准的评审工作。同时，问途还作为研制单位之一，全程参与了高职专科酒店管理与数字化运营专业教学标准的研制工作。

在国家专业教学标准的研制和评审过程中，旅游及酒店业大数据应用课程是业界对院校课程设置中呼声极高的课程之一。我们发现，现有的大数据教材基本上分为两类：一类是大数据技术类教材，需要大量的数理知识和信息技术知识为基础；另一类则是大数据通识类教材。对于旅游及酒店管理专业而言，难以找到合适的教材。大数据技术类教材受制于原旅游大类专业偏文科的特质，而大数据通识类教材则不能满足专业性和职业性教育

的需求。

教学设计目标是培养学生既具备大数据基础知识和大数据思维，又具备专业性和职业性的大数据应用能力。教材适用于旅游院校高职专科、高职本科和应用型本科院校的大数据应用课程教学，并可用于旅游与酒店业从业人员的培训。

全书分为导论篇和实践篇两大部分。

导论篇内容包括数据的本质、大数据的概念、大数据在旅游业的应用和人才需求、大数据基础技术以及与新一代信息技术关系、大数据思维及职业规划、大数据安全问题与对策、大数据个人隐私保护问题与对策、大数据伦理问题与对策。导论篇内容紧扣相关国家方针政策，每一节先进行浅显易懂的案例导入，然后翔实说明大数据在旅游业的融合应用，旨在帮助没有技术背景的旅游类专业的学生对大数据和大数据在旅游业的应用场景和挑战有清晰的认知，并能够运用大数据思维去进行职业规划。

实践篇首先介绍了大数据核心技术的实现方式，内容包括数据的采集、预处理与存储，数据的处理与分析，数据的可视化，数据的预测，数据的安全。为了让旅游类专业的学生掌握这些技术应用，还设计了 5 个实训任务，包括酒店业数据的采集与预处理、聚类算法在用户画像中的使用、关联分析算法在景区营销设计中的使用、图像识别技术在景点图片分析中的使用、Python 数据实现可视化基本操作。

实践篇还详细阐述了大数据在旅游与酒店行业的主要应用场景，包括基于大数据的旅游目的地指挥调度、基于大数据的旅游舆情应急管理、基于大数据的客户价值分析、基于大数据的市场细分、旅游统计与大数据、基于大数据的客流统计、基于大数据的客流预测。为了让旅游管理类专业的学生进一步理解大数据技术是如何在旅游业具体场景中应用的，实践篇也提供了网络爬虫在舆情管理中的使用、文本情感分析在舆情管理中的使用、酒店客户价值分析、酒店精准营销设计、旅游大数据大屏展示界面设计、酒店客流预测实操 6 个实训任务。

本书提供的多个实训任务既可以使用免费的数据挖掘软件 Python 开展实训，也可以使用问途大衍 megAnalysis ®大数据平台和相应的数字资源教学。教学方法建议均以项目为纽带、以任务为载体、以工作过程为导向，围绕旅游及酒店业大数据应用实际工作中的相关任务开展实践教学。问途大衍 megAnalysis ®大数据平台可以为有意向使用本书的院校在安排教师参与相关培训后免费开放一些实训任务。

感谢参与本书编写的两位副主编黄婉敏、张芷晴，感谢参与编写本书所有实训任务的问途公司资深研发工程师李海泉，他（她）们将来自工作的宝贵经验进行提炼和总结，并形成体系化的知识分享给读者。感谢参与本书插图和练习题设计的问途公司 UI 设计师梁志嘉、内容营销经理吴耿瑜、项目运营经理邬凯婷，他（她）们的工作使得本书能够更好地呈现在读者面前。

大数据技术在旅游与酒店业的应用和实践仍然处于不断的发展当中，本书的编写过程也是一个对旅游院校开展大数据课程教学的探索过程。对于本书存在的问题和不足之处，

恳请各位专家、老师和同学不吝批评、指正。

广东海洋大学商学院　黄昕
广州市问途信息技术有限公司　张峰
2022 年 4 月

问途大衍大数据平台资源说明

“问途大衍大数据平台”是一个适用于本科院校和高等职业院校旅游及酒店管理专业学生学习大数据应用知识的平台系统。该平台针对旅游及酒店管理专业的教学需求，设置了涵盖大数据核心技术和行业应用的多个实训任务，培养学生既具备大数据基础知识和大数据思维，又具备专业性和职业性的大数据应用能力。问途大衍大数据平台包括支持本书实训任务实践的大衍大数据实训平台和大衍旅游大数据看板两个模块。平台支持本书对应课程在旅游类专业的创新教学，从而推动旅游类专业数字化升级，支持智慧旅游技术应用专业、智慧景区开发与管理专业、定制旅行管理与服务专业、酒店管理与数字化运营专业的数字化改造。

为方便教师教学选用和学生的知识学习，问途大衍大数据平台线上资源学习平台整理了本书配套的课前知识点、课前测试、课中实训实验及课后练习等模块，满足课程通过SPOC教学方法开展、全过程学生学习情况监控跟踪、针对教授对象和教学要求灵活组织教学课程以实现教学目标等需求。

如何获取“问途大衍大数据平台”试用课程资源？

平台资源对采用本书授课的本科院校及高等职业院校旅游管理、酒店管理等相关专业老师开放试用课程资源。

微信搜索公众号“数字旅游实验室”，回复关键字“大数据”，按页面提示要求完成申请即可。

目　录

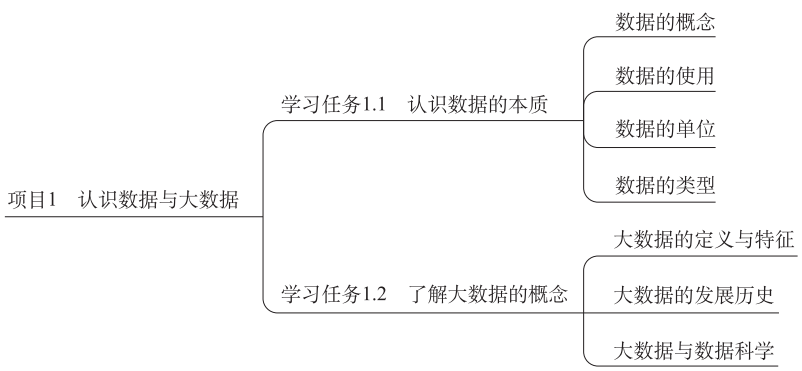
第一篇

导　论　篇

项目 1

认识数据与大数据

项目结构



学习目标

学习层次	学 习 目 标
知道	1. 认识数据的概念 2. 了解大数据的定义 3. 描述大数据的特征 4. 了解数据科学
理解	1. 掌握 DIKIW 数据转化模型 2. 解释数据的单位 3. 说明数据的类型 4. 辨别数据的来源 5. 解释大数据的发展历史 6. 说明数据科学的工作过程

学习任务 1.1 认识数据的本质

【任务概述】

数据是用于反映客观事物的性质、状态、关系和变化的原始素材，是形成信息的基本要素；信息是数据进行加工中，形成对客观事物有意义的描述，用于帮助决策或者解决问题。通过数据技术，数据的收集将更为迅速和丰富，管理者可以通过大量的数据来获得所需要的信息，准确地掌握运营情况。数据为管理决策提供了有信服力的依据，帮助管理者有效判断，并能够洞察运营的问题或者发现新的机会。多维度的数据有助于帮助企业推进精准营销，提升客户体验、提高客户黏性，增强数字服务的能力。

DIKIW 数据转化模型(Data to Information to Knowledge to Intelligence to Wisdom)说明了数据是信息、知识以及智慧的来源，揭示了数据是如何一步一步转化为信息、知识、智力和智慧的过程，描述了这些不同数据加工阶段之间的关系。企业在经营中，需要高度重视数据获取的完整性和时效性，防止因为数据不完整或者失效而导致无法产生准确的信息和有价值的知识，进而影响未来的决策。

计算机数据存储是以“字节(byte)”为单位进行存储，计算机中表示数据大小的计量是 2^n 来计量，依次增大为前一个计量单位的 1 024 倍，即 2^{10} 倍。根据数据的存储与处理方式可以将数据分为结构化数据、非结构化数据和半结构化数据；根据数据的来源，可以将数据分为第一方数据、第二方数据、第三方数据和公共数据；根据数据的拥有主体，可以将数据分为内部数据和外部数据。将数据分类，有助于企业从业务需求和使用角度，找到最有价值的数据和使用方案。

【案例导入】

走进海南“智慧大脑”——海南省大数据中心，可以看到工作人员正在一台台计算机前忙碌着。这里汇聚了海南全省基础数据库和各级政务部门的数据，可以调取实时数据并将数据显示在大厅前方的巨幅显示屏上。数年前的三亚“天价海鲜宰客门”事件让三亚旅游品牌形象严重受损。为了规范旅游市场，当地政府除了制定严格的监管和惩罚制度，更借助于数据技术对市场进行治理。

在三亚某海鲜广场，一块电子显示屏十分醒目，上面显示着当日鱼、虾、蟹等各类海鲜销售价格和政府限价。三亚相关部门定期发布各类海鲜最高限价。在将数据发布给海鲜商铺的同时，三亚主管部门也将所有的海鲜商铺的电子秤、收费系统接入海南省大数据平台。三亚市政府陆续为海鲜商铺配备了一款特殊的公平秤，通过在电子秤上安装网线，市民游客在消费时，所有交易实时反馈到数据平台上，实时采集从数量到单价的数据。由系统自动检查每笔交易，清楚显示所购买的海鲜价格以及店铺名称。通过这些数据可以判断商户是否存

在作假，以及价格是否超出政府限价。一旦信息出现异常，数据平台就会自动报警，进入联合执法系统处理。

如果您认为大数据平台只能应用于监控物价信息，那就有点小瞧这个“智慧大脑”了，通过进行政务信息整合，大数据平台对数据进行采集、挖掘及分析，快速给出应急响应方案，为精准防灾、应急指挥、科学决策提供了技术支撑。某天，三亚海上搜救分中心 12395 热线接到一位带团导游的电话。导游电话中称：“旅游团乘坐的游船在蜈支洲岛附近的海域着火遇险，情况紧急，请求救援。”海南省旅游信息和咨询服务中心借助全域旅游监管服务平台立即获取导游、旅游团、旅游大巴等相关数据，在不到 2 秒的时间内检索出了旅行团成员的数据。这是一个 29 人的旅游团，根据旅游大巴车行驶轨迹数据，他们从海口进岛后，直奔蜈支洲岛，乘船下海。管理部门在收到船舶发出的遇险警报后，利用北斗卫星导航系统的短报文通信功能与遇险船舶取得联系，定位和跟踪遇险人员。按照应急救援流程，海南省旅游信息和咨询服务中心及时向救援部门发送事发船只上旅游团人员的数据，结合事发游船发回的求救电话内容，蜈支洲岛旅游区的三亚湾海岸救援队迅速掌握了船上人员情况。事发游船发出紧急求救信号后，船长立即用灭火器实施灭火并对部分受惊吓游客进行心理安抚与疏导。三亚海上搜救分中心接警后，立即核实险情，迅速启动应急预案，安排工作人员准备好快艇与蜈支洲岛旅游区的三亚湾海岸救援队联络，带着有出海救援经验的急救医生随船出海，救援伤者。最终，在覆盖空中、水面、水下的立体救援的帮助下，不到半小时的时间，安全处理了这起海上突发事件。

上述案例只是旅游部门安排的一次演习，但这次演习不仅仅直接考验海南省全域旅游监管服务平台应对突发事件的能力，而且说明了数据的重要性。如今，通过海南省全域旅游监管服务平台，在事发地点方圆 100 千米范围内，动态掌握团队进出岛数据，包括当日团队进出岛数量、游客数等，建立游客类型画像。借助于实时掌握的旅游团队的接待情况、运动轨迹等数据，一旦发生旅游突发事件，全域旅游监管服务平台会立即向相关人员发送预警信息，为应急指挥和疏导做好准备。

一、数据的概念

数据(Data)是未经加工的原始素材，用于记录和反映客观事物的性质、状态、关系和变化。数据既可以是符号、文字、数字，也可以是图像、声音、视频等形式。人们对数据进行加工，使之成为“信息(Information)”。在上述的情景导入案例中，在海鲜市场配置公平秤采集到的交易数量、交易单价等就是未经加工的原始数据，反映了海鲜市场交易的变化。通过对这些数据进行汇总和对比分析，就能够发现是否存在异常交易的信息。在处理游船遇险救援的演习案例中，旅行团的成员背景、大巴行驶轨迹这些记录都是数据，通过对这些数据的整合加工，主管部门获得了遇险团队的全面描述，形成了救援的信息。可见，数据是用于反映客观事物的性质、状态、关系和变化的原始素材，用于统计计算、科学研究和辅助决策。通过对数据的存储、传输和加工，形成了对客户事务有意义的描述，即生成了信息与知识，数据是形成信息和知识的基本要素。

在获取反映客观事物性质的数据方面，如果依赖个人经验并通过个人直觉去判断，效率将会非常低。例如，在本节情景导入案例中，如果市场管理人员每天挨家挨户查看交易数据，并通过个人经验判断是否存在缺斤少两的情况，效率可想而知。但通过借助于数据技术

对各种数据进行加工、分析、处理，就可以实现“数据—信息”这个过程的高效转化。通过数据技术，数据的收集将更为迅速和丰富，管理者可以通过大量的数据来获得所需要的信息，准确地掌握运营情况。数据为管理决策提供了有信服力的依据，帮助管理者有效判断，并能够洞察运营的问题或者发现新的机会。

综合上述，数据可以帮助管理者在复杂的现实环境中获取有效的信息，找到支持决策的依据，为管理者获得“时间差”的优势。由于数据的获取方法、来源维度不同，对企业管理和运营有不同的价值。例如，在企业和客户在线交易过程中发生的交易数据，可以用于准确描述已成交的产品的客户认可度信息，但是对于没有成交数据的产品，就缺乏交易数据去支持真实的原因。这个原因不一定是产品的问题，而有可能是产品在页面上的位置不合理、图文内容质量不高、产品的详情页加载速度太慢等原因，导致产品的转化率不高。要有效得到产品转化率低背后的信息，还需要网页上访问数据进行综合判断。

从数据来源维度看，可以将数据分为时间维度、空间维度、社交维度等多个维度的数据。企业运营人员需要从哪个维度获取数据，主要看企业当前和未来的业务对信息的要求。例如，在酒店与旅游企业开展精准营销和个性化服务的工作中，能够反映客户实时感知、产品使用和服务过程的数据是特别有价值的数据，这些数据能够准确描述“是什么样的用户(Who)”“在什么地点(Where)”“在什么时间(When)”“使用何种方式(How)”“做了什么事情(What)”，从而有助于形成判断客户需求的信息。多维度的数据有助于帮助企业推进精准营销，提升客户体验、提高客户黏性，增强数字服务的能力。

《“十四五”数字经济发展规划》中指出，要深化数字经济的发展，核心引擎是数据要素。数据对提高生产效率的乘数作用不断凸显，成为最具时代特征的生产要素。数据的爆发增长、海量集聚蕴藏了巨大的价值，为智能化发展带来了新的机遇。协同推进技术、模式、业态和制度创新，切实用好数据要素，将为经济社会数字化发展带来强劲动力。

二、数据的使用

旅游者在计划旅行的时候，通常会关注旅游目的地的天气预报，准确的天气预报内容对旅游者的出游准备和活动安排非常重要。从数据来看，天气预报的工作实际上是对原始天气数据的加工和转化。例如，某一个旅游目的地每一天的温度、湿度、降雨量这些数值都是原始数据(Data)。温度和湿度这些数值对于旅游者而言，可以转化为“热还是不热”“闷还是不闷”等有逻辑、有意义的“信息(Information)”。这些信息汇总起来，通过挖掘这些信息背后的规律，就可以总结出对这个旅游目的地气候的概括和理解，这就是“知识(Knowledge)”，它反映了人们对该领域的认知、行动能力和理解。例如，对花城广州的全年天气数据进行汇总分析后，可以总结出“广州全年平均气温为20～22℃，一年中最热的月份是7月，月平均气温达28.7℃。最冷月为1月，月平均气温为9～16℃。平均相对湿度77%，市区年降雨量约为1 720毫米。全年中，4至6月为雨季，7至9月天气炎热，10月、11月和3月气温适中，12至2月为阴凉的冬季”这个知识性判断。基于这些天气知识的运用，旅游组织者可以得出“每年10月、11月是广州最佳旅游季节”的结论，从而可以将10月和11月的旅游线路产品作为推广的重点。因此，这些基于知识所做的决策能力可以用“智力(Intelligence)”来描述。旅游组织者在服务众多旅游者过程中，不断洞察旅游者需求，对旅游者进

行细分，再对细分旅游者的需求进行洞察，最后结合天气信息和知识，就可以每年根据气候的变化，针对不同细分市场的旅游者设计不同的城市休闲旅游线路和个性化的行程安排，从而提高客户的满意度和黏性。这种面向未来需求，不断对知识和信息进行加工，演绎出最合适解决方案的能力就是“智慧(Wisdom)”。

综合上述，从“数据(Data)”加工到“信息(Information)”，从“信息(Information)”加工到“知识(Knowledge)”，从“知识(Knowledge)”加工到“智力(Intelligence)”，从“智力(Intelligence)”加工到“智慧(Wisdom)”，就形成了 DIKIW 数据转化模型，如图 1-1 所示。

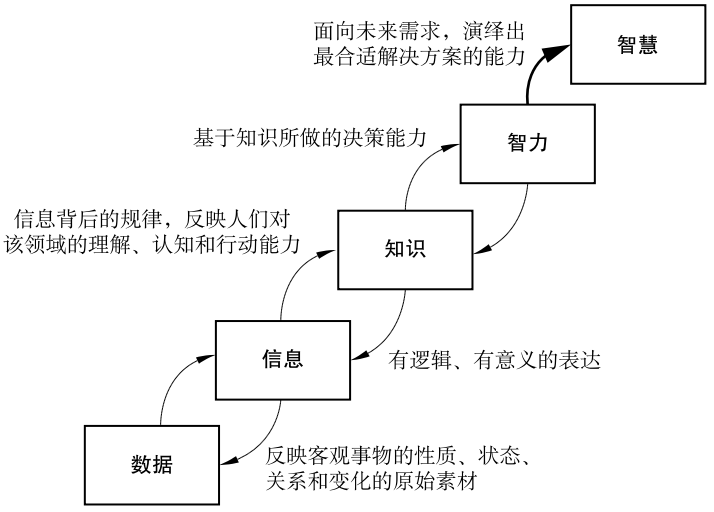


图 1-1　DIKIW 数据转化模型

DIKIW 数据转化模型说明了数据是信息、知识以及智慧的来源，揭示了数据是如何一步一步转为信息、知识、智力和智慧的过程，描述了这些不同数据加工阶段之间的关系。

因此，企业在经营中，需要高度重视数据获取的完整性和时效性，防止因为数据不完整或者失效而导致无法产生准确的信息和有价值的知识，进而影响未来的决策。

三、数据的单位

数据需要使用计算机进行存储和传输。计算机数据存储是以“字节”为单位进行存储，但是以“位”(Binary Digit，Bit，也称为“比特”)为单位进行数据传输。如图 1-2 所示，一个位就代表一个二进制的 0 或 1，每八个位组成一个字节。

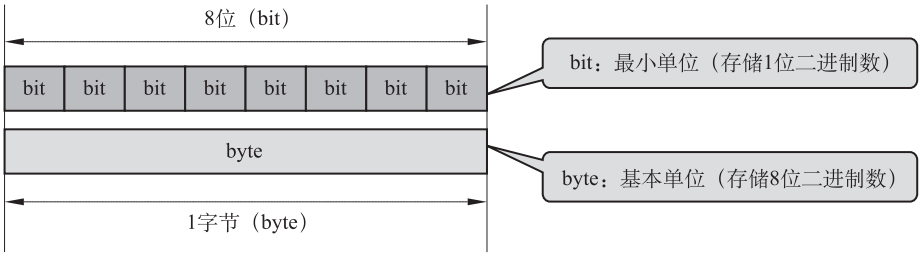


图 1-2　字节与比特单位的比对

字节是数据的最基本单位。一个英文字母(不分大小写)占一字节的空间,一个中文汉字占两个字节的空间。一个英文标点占一字节,一个中文标点占两字节。

计算机中的数据都是用二进制数表示的。现在人们对数据的认知大多数情况都是指用二进制形式编码并计量的各种数据。随着数据量的不断增加,需要用计量单位来表示数据的大小。计算机中表示数据大小的计量是 2^n 来计量,依次增大为前一个计量单位的 1 024 倍,即 2^{10} 倍,如图 1-3 所示。

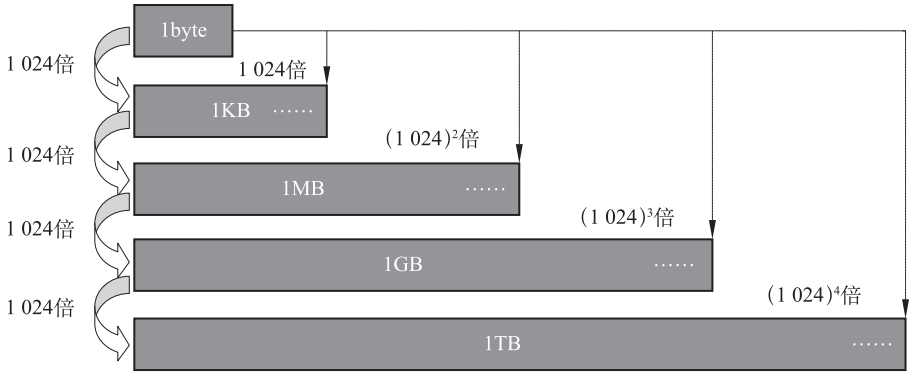


图 1-3　计算机中表示内存大小(存储容量)的单位

数字经济时代,计算力已成为核心生产力,数据是基础性资源,是新的生产要素。人类社会进入了数据体量井喷的时代,数据存储呈爆炸性增长态势,如表 1-1 所示。

表 1-1　数据单位对比表

单　位	简写	计　量	说　明
字节(byte)	B	8B	一个英文字母或一个英文标点符号
千字节(kilobyte)	KB	1 024B	512 个汉字
兆字节(megabyte)	MB	1 024KB	一首 MP3 格式的歌曲大约 4MB。如果用更低音质的 wma 格式存储,一首歌 2MB 左右
吉字节(gigabyte)	GB	1 024MB	未经压缩的 1080P 格式电影在 20GB 以上
太字节(trillionbyte)	TB	1 024GB	中国公共图书馆 2020 年年底面向全国共享的数字资源超过 145TB
拍字节(petabyte)	PB	1 024TB	人类大脑的信息存储容量至少为 1PB(美国索尔克生物学研究所)
艾字节(exabyte)	EB	1 024PB	人类所说过的所有的话约 5EB
泽字节(zettabyte)	ZB	1 024EB	2020 年人类创建、捕获、复制和消耗的数据总量约为 59ZB
尧字节(yottabyte)	YB	1 024ZB	
珀字节(brontobyte)	BB	1 024YB	
诺字节(nonaByte)	NB	1 024BB	
刀字节(doggaByte)	DB	1 024NB	

四、数据的类型

根据数据的存储与处理方式、数据的来源、数据的拥有主体等不同维度，可以将数据分为不同的类型。了解数据的类型，有助于从业务需求和使用角度，找到最有价值的数据和使用方案。

1. 根据数据存储和处理方式分类

从数据的存储和处理方式来看，数据类型可以分为三种：结构化数据（Structured Data）、非结构化数据（Unstructured Data）和半结构化数据（Semi-structured Data）。

结构化数据是指能够用二维表结构来表达实现，通过关系型数据库进行存储和管理的数据，通常以行为单位，一行数据代表一个实体的信息，每行数据的属性是相同的，严格遵循数据格式和长度规范，数据的存储和排列是有规律的。例如，以行和列这种表格形式存在和组织的数据都是结构化数据，例如酒店管理系统（Property Management System，PMS），旅行社管理系统中的数据都是结构化数据，包括预订数据、客户数据、销售数据、支付数据等，如图 1-4 所示。结构化数据为决策者提供了丰富的信息，但这些信息通常都是解释已经发生的事实。对于正在发生的事件，通过结构化数据能够产生的信息非常有限。

订单号	状态	预订人	入住人	提交时间	预订产品	数量	入住日期	离店日期	间夜	订单金额	价格计划	平均售价	在线支付金额	支付状态
00010897	已确认	酒店会员 / 微信粉丝 / 666******91 - 展**	展**	2021-12-03 12:39	DR - 豪华单床房	2	2021-12-03	2021-12-04	2	820.00	粉丝专属价格	410.00	820	已支付
00010898	未支付	酒店会员 / 微信粉丝 / 666******69	666******69	2021-12-28 21:32	DR - 豪华单床房	1	2021-12-28	2021-12-29	1	410.00	粉丝专属价格	410.00	410	未支付
00010899	已确认	酒店会员 / 微信粉丝 / 666******72	666******72	2021-12-29 23:48	BR - 商务大床房	1	2021-12-29	2021-12-30	1	480.00	粉丝专属价格	480.00	480	已支付
00010900	已确认	酒店会员 / 微信粉丝 / 666******15 - 王**	王**	2022-01-12 13:13	DR - 豪华单床房	2	2022-01-12	2022-01-13	2	720.00	粉丝专属价格	360.00	720	已支付
00010901	已确认	酒店会员 / 微信粉丝 / 666******20 - 张**	林**	2022-01-12 20:57	DR - 豪华单床房	3	2022-01-12	2022-01-13	3	1080.00	粉丝专属价格	360.00	1080	已支付
00010902	已确认	酒店会员 / 微信粉丝 / 666******89	666******89	2022-01-13 10:46	DR - 豪华单床房	2	2022-01-13	2022-01-14	2	720.00	粉丝专属价格	360.00	720	已支付
00010903	未支付	酒店会员 / 微信粉丝 / 666******39	666******39	2022-01-13 10:47	DT - 豪华双人房	3	2022-01-13	2022-01-14	3	1200.00	粉丝专属价格	400.00	1200	未支付
00010904	已确认	酒店会员 / 微信粉丝 / 666******11	666******11	2022-01-13 10:48	DT - 豪华双人房	3	2022-01-13	2022-01-14	3	1200.00	粉丝专属价格	400.00	1200	已支付
00010905	已确认	酒店会员 / 微信粉丝 / 666******90	罗**	2022-01-13 20:53	BR - 商务大床房	1	2022-01-13	2022-01-14	1	430.00	粉丝专属价格	430.00	430	已支付

图 1-4　结构化数据示例

非结构化数据就是没有固定的结构，无法用传统的二维表结构来表现以及用传统数据库来存储的数据。例如，旅游目的地的天气数据和地形数据、旅行社的网站文本、社交媒体上的图片和短视频，酒店与客人的交流或录音记录、酒店经营场所的监控和视频等数据，都是非结构化数据，如图 1-5 所示。这类数据在全球可以使用的数据总量中，占了绝大多数。受限于储存和处理能力，非结构化数据一直没有得到很好的分析和利用，因为非结构化数据需要更大、更好的存储空间，而且数据处理和分析技术更为复杂。直到大数据技术的飞速发展，非结构化数据才得到了越来越深度的应用。

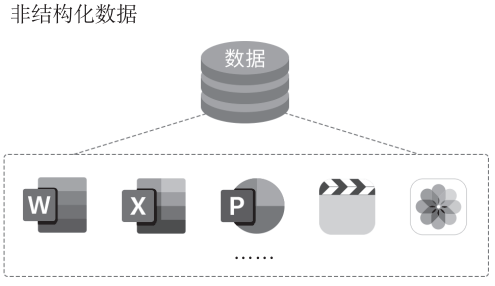


图 1-5　非结构化数据示例

半结构化数据是介于结构化数据和非结构化数据之间的数据，它有一定的结构，但又不能完全结构化，结构变化很大。例如，企业到大学去进行校园招聘，同学们给的简历有一定的结构性，一般包括姓名、性别、专业、政治面貌、特长等。但有的同学的简历可能很丰富，包括兼职经历、学生干部经历，甚至有一些个性化的数据，如个人生活照片等。这些有一定结构但又不完全结构化的简历数据就是半结构化数据。此外，用户在互联网上通过浏览器读取的网页文件都是由 HTML 格式组成，也属于半结构化数据，如图 1-6 所示。

```
<!DOCTYPE html>
<html class="no-js">

<head>
    <meta charset="utf-8">
    <title>Dossm v3 Workbench</title>
    <meta name="description" content="">
    <meta name="HandheldFriendly" content="True">
    <meta name="MobileOptimized" content="320">
    <meta name="viewport" content="width=device-width, initial-scale=1, minimal-ui">
    <meta http-equiv="cleartype" content="on">
    <link rel="stylesheet" href="/dist/icon/style.css">
    <link rel="stylesheet" href="/dist/css/dossm3.0.css">
    <link rel="icon" href="/dist/images/favicon.ico">
    <script src="/dist/jquery/jquery-1.12.0.min.js"></script>
    <style type="text/css">
        .quick {
            position: absolute;
            top: 10px;
            right: 15px;
            display: block;
            width: 40px;
            height: 40px;
            cursor: pointer;
        }
```

图 1-6　半结构化数据示例

2. 根据数据来源分类

根据数据的来源，数据类型可以分为四种：第一方数据（First-party Data）、第二方数据（Second-party Data）、第三方数据（Third-party Data）和公共数据（Public Data）。

第一方数据是指企业通过各种客户接触点收集的数据，以及企业自身不断积累的数据。注册会员时填写的个人资料、浏览企业的网站或 App 时产生的行为数据等属于第一方数据。

第二方数据是指从其他合作机构购得或者交换来的第一方数据。例如，广告服务商、运营服务商、合作媒体、上下游合作企业的数据。

第三方数据是指从没有直接合作关系的企业或者组织获得或者购买的数据。例如，数据交易平台、广告营销代理商、互联网服务提供商等提供的数据都属于第三方数据，主要是对于第一方数据的增强。

公共数据是指从公共渠道获得的低成本、对外公开的数据。

一栋新公寓楼的业主正在寻找租户。若使用第一方数据，只能访问之前访问过他们网站的人。借助第三方数据，业主可以覆盖更广泛的人群，找到近期在线搜索过附近公寓的人。

3. 根据数据的拥有主体分类

根据数据的拥有主体，数据类型可以分为两种：内部数据和外部数据。内部数据是指企业能够自行采集、拥有、管理和应用的数据。这些数据可以是结构化、非结构化或者半结构化。例如，企业的会员数据、销售数据、财务数据、员工数据等。对企业来说，获取内部数据的成本很低，但企业需要确保内部数据的安全。

外部数据是相对于内部数据而言，通过公开渠道或者第三方渠道获取的数据。例如，政

府和研究机构发布的人口普查数据、行业调研数据、行业业务统计数据；通过第三方数据平台获取的人群画像数据等。外部数据相对于内部数据而言，更加丰富、全面和复杂。它可以和内部数据结合，为企业提供更全面的分析视角，更全面地洞察事物本质。

【主要术语】

1. 数据：数据(Data)是未经加工的原始素材，用于记录和反映客观事物的性质、状态、关系和变化。数据既可以是符号、文字、数字，也可以是图像、声音、视频等形式。

2. DIKIW 模型：用于揭示从“数据(Data)”加工到“信息(Information)”，从“信息(Information)”加工到“知识(Knowledge)”，从“知识(Knowledge)”加工到“智力(Intelligence)”，从“智力(Intelligence)”加工到“智慧(Wisdom)”的全过程的模型，说明了数据是信息、知识以及智慧的来源，揭示了数据是如何一步一步转化的过程和阶段，描述了这些不同数据加工阶段之间的关系。

3. 字节：字节是数据的最基本单位。计算机数据存储是以“字节”为单位进行存储，但是以“位”(Binary Digit，Bit，也称为“比特”)为单位进行数据传输。一个位就代表一个二进制的 0 或 1，每八位组成一字节。计算机中的数据都是用二进制数表示的，表示数据大小的计量是 2^n 来计量，依次增大为前一个计量单位的 1 024 倍，即 2^{10} 倍。

4. 结构化数据：能够用二维表结构来表达实现，通过关系型数据库进行存储和管理的数据，通常以行为单位，一行数据代表一个实体的信息，每行数据的属性是相同的，严格遵循数据格式和长度规范，数据的存储和排列是有规律的。

5. 非结构化数据：没有固定的结构，无法用传统的二维表结构来表现以及用传统数据库来存储的数据。

6. 半结构化数据：介于结构化数据和非结构化数据之间的数据，它有一定的结构，但又不能完全结构化，结构变化很大。

7. 第一方数据：企业通过各种客户接触点收集的数据，以及企业自身不断积累的数据。

8. 第二方数据：从其他合作机构购得或者交换来的第一方数据。

9. 第三方数据：从没有直接合作关系的企业或者组织获得或者购买的数据。

10. 公共数据：从公共渠道获得的低成本、对外公开的数据。

11. 内部数据：企业能够自行采集、拥有、管理和应用的数据。

12. 外部数据：通过公开渠道或者第三方渠道获取的数据。

【练习题】

一、自测题

1. 请举例说明什么是“数据”以及这些“数据”如何加工成为有意义的“信息”。

2. 有人说，只有在线的数据才是最有价值的数据。你认同吗？为什么？

3. DIKW 数据转化模型和 DIKIW 数据转化模型在应用上有什么区别？

4. 除了本节对大数据时代存储单位(如太字节、拍字节等)的说明，请对这些大数据存储单位另外举例描述。

5. 请根据旅游或酒店企业的数据特点，举例说明结构化数据、非结构化数据和半结构

化数据。

二、讨论题

1. 请在互联网上搜索一个旅游活动的实施案例或者一个旅游事件的处理案例，运用“数据”和“信息”的概念，讨论该活动或事件是如何通过数据进行决策的。

2. 某酒店集团拥有几十万名会员，为了快速发展会员，集团与合作银行建立了合作关系，通过会员系统对接将银行的用户引导为酒店集团会员。另外，该集团还和某大数据公司合作，将该集团的不同级别、不同地域的会员抽取样本数据后，由合作的大数据公司提供该组会员数据的标签，如学历、职业、收入、喜好等数据。请讨论该酒店集团有哪些第一方数据、第二方数据和第三方数据。

三、实践题

1. 请仔细分析本节中的 DIKIW 数据转化模型案例，然后联系一家酒店的订房部或者市场营销部进行访谈，运用 DIKIW 模型描述一下从数据一步一步转化到智慧决策的全过程，请用 PPT 进行小组演示。

2. 请搜索一下数据在世界和中国的起源，并结合数据的定义，做成一个图文并茂的短视频向同学们进行介绍。

学习任务 1.2　了解大数据的概念

【任务概述】

大数据是一种具有体量巨大、来源多样、生成极快且多变等特征，并且难以用传统数据体系结构有效处理的包含大量数据集的数据。广义上来说，大数据是思维、技术、数据和应用的结合，包括数据特征、分析方法、处理技术、商业模式、思维方式等。大数据的特征可以用“6V”来定义：巨大的数据量（Volume）、数据种类和来源多样（Variety）、数据处理速度快（Velocity）、数据类型多变（Variability），较高的商业价值（Value），以及数据准确性和可信度高（Veracity）。大数据的发展历史可以分为 1.0、2.0 和 3.0 阶段，大数据 1.0 阶段，是以 20 世纪 70 年代的关系型数据库技术、80 年代的数据网络化传输和 90 年代的互联网上不同主机之间的信息共享为主要应用场景；大数据 2.0 阶段，包括结构化和半结构化的海量数据规模促使专门为存储和分析大数据集而创建的开源框架和非关系型数据库技术的发展；大数据 3.0 阶段，大数据技术得到快速发展，移动互联网、物联网的广泛应用促使大数据成为一个重要的生产因素。数据科学成为一个新兴的研究领域。数据科学的工作过程包括：制定目标、数据理解与准备、数据建模、模型评估、结果呈现和模型部署。

【案例导入】

根据杭州网的相关报道，杭州在 2017 年组建了杭州旅游经济实验室并成立了杭州旅游

大数据中心。2018 年，旅游大数据平台与杭州城市大脑系统全面对接。杭州城市大脑是杭州为城市生活打造的数字化界面，市民通过它享受城市服务，城市管理者通过它配置公共资源，进行决策和治理。旅游大数据平台与城市大脑对接后，就形成了杭州城市文旅系统，用于建立文旅市场的智能监管机制和响应机制，实现文化旅游运营和公共管理的双赢。该大数据系统能够通过采集与分析游客的行为轨迹、消费轨迹、时空轨迹等个体数据，以及交通、气象、公安、舆情等公共数据进行动态监测和精准分析。

杭州已经实现以分钟级的颗粒度完成对游客的整体画像和旅游市场的立体描述，同时为游客提供精准的掌上旅游服务。例如，在春节黄金周，借助旅游大数据平台，杭州市文化广电旅游局发布了相关数据报告，对来杭州的游客数量、住宿偏好、消费习惯、饮食习惯，游客在杭关键词搜索、行动轨迹、消费评价等数据进行了精准分析，刻画用户画像。这些数据用于游客流量预测、引导分流、精准营销和企业决策辅助等数据赋能的用途。

大数据不仅仅用于政府治理和决策用途，还为消费者提供了诸多便利条件。在杭州的桐庐高铁站，游客只要搭乘一人一座、票价 3 元的酒店穿梭巴士，便能直接抵达酒店门口。通过“数据＋算法”，依据高铁时刻表，桐庐已经开通了十余条数字旅游专线，目的地涵盖了酒店、景区与主题乡村。桐庐站日均有 60 多趟高铁停靠，依据高铁时刻表，游客走出桐庐站就可以直接搭乘准点发车的穿梭巴士、旅游专线，直接前往酒店、瑶琳仙境景区、白云源景区、江南古村等，而根据季节特色，桐庐还会设计不同主题的数字旅游专线。桐庐的数字旅游专线，是杭州文旅针对旅游过程中最常见的排队、等候、盲从、资讯获取慢以及“游占比”低下等“痛点”问题，所进行的一次“场景革新”。依托游客轨迹、铁路预订、公交运行等数据，进行后台关联分析与运算，科学规划旅游专线，动态调度班次时间。数字旅游专线不仅仅在桐庐县有，杭州下辖的淳安县的所有接驳巴士信息也都已经与百度、云公交、铁路等平台互通，日均发车 214 班次，日均输送旅客 3 479 人次。方便查询之外，游客的“游占比”明显提高。这也意味着游客在淳安的旅游候车时间至少可以节省 16 分钟。

2019 年，杭州全面布局的 10 秒找空房、20 秒景点入园、30 秒酒店入住、数字旅游专线、长三角 PASS 卡、杭州文化旅游年卡六大便民应用场景，将数字文旅服务由景区扩展至杭州全域范围，累计服务游客已超过 550 万人次。以 10 秒找空房小程序为例，前端空房展示平台对接了 8 561 家酒店，覆盖杭州各区（县市），通过“找空房”小程序和“96123”旅游咨询服务热线两个渠道，游客可以基于当前定位和价格偏好，找到附近的“性价比”高的酒店空房，而 20 秒景点入园场景，正在加速覆盖杭州各大景区（景点）。

杭州城市大脑文旅系统将按照“数据线上跑、用户线下游”的理念，通过部门间合作开放数据，打破数据壁垒。在游客出行、停车、景区入住、酒店入住、消费引导等旅游体验方面，多方位优化数据应用，打造互联互通、信息对称的国际体验旅游目的地。

一、大数据的定义与特征

1980 年，美国未来学家阿尔文-托夫勒（Alvin Toffler）在他所撰写的《第三次浪潮》（*The Third Wave*）中预言，大数据将成为“第三次浪潮的华彩乐章”。所谓第三次浪潮，就是在农业文明阶段、工业文明阶段之后的新文明，数字技术与生物技术是第三次浪潮的核心驱动力。

根据已经发布的中华人民共和国国家标准《信息技术 大数据 术语》(GB/T 35295—2017),大数据被定义为"具有体量巨大、来源多样、生成极快且多变等特征,并且难以用传统数据体系结构有效处理的包含大量数据集的数据"。大数据以拍字节(Petabyte,PB)为单位进行生成和存储,这和传统的最多以太字节(Trillionbyte,TB)为单位进行生成、存储和分析的技术差异很大。大数据所指的数据规模已经达到无法用传统的数据体系结构和技术来有效处理的程度,数据科学家需要用特定的技术、工具和方法,用于快速采集、存储和分析海量数据,进行建模和实现预测的目的,从而实现传统技术无法创造的竞争优势和创新能力。这些大数据技术包括 Apache Hadoop、Apache Spark、Apache Kafka 等,通过运用大数据技术,帮助企业通过正确的方式,在正确的时间从可用的数据中获取正确的决策支持,并寻找新的商业机会来确保企业的核心竞争力和创新发展能力。

大数据作为一个用来描述海量数据的流行语,它不仅仅是一个技术概念,也是一个商业概念。大数据为决策者们提供了获取关键信息、知识和智慧的途径,是企业建立核心竞争力的关键技术和策略之一。中华人民共和国国家标准《信息技术 大数据 术语》(GB/T 35295—2017)文件中指出,国际上对于大数据的特征通常直接用 Volume(体量)、Variety(多样性)、Velocity(速度)和 Variability(多变性)予以表述,即"4V"。大数据是新技术、新知识、新方法和新的商业模式的混合体,除了上述"4V"特征外,还有不少专家认为 Value(价值)和 Veracity(真实性)也是大数据的重要特征,即"6V",如图 1-7 所示。

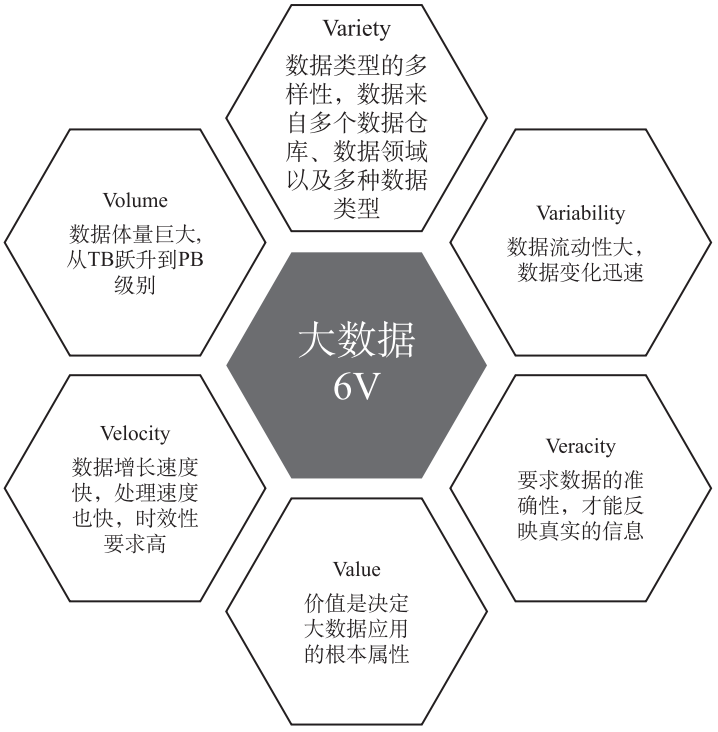


图 1-7　大数据"6V"特征

1. Volume:数据体量巨大

大数据的"大"首先体现在数据量上。随着移动互联技术的普及,相关应用产生的数据

量也在急剧增加。

根据国际数据公司(IDC)2019 年发布的《数字化世界——从边缘到核心》白皮书,全球数据圈将从 2018 年的 33ZB(十万亿亿字节)增至 2025 年的 175ZB。1ZB 相当于全世界海滩上的沙子数量的总和,若一般家用计算机硬盘以 100GB 为单位计算,那 1ZB 数据要存满百亿台数量级的电脑硬盘。淘宝网每天有数千万的交易量,其单日数据产生量超过 50TB(万亿字节);百度每天处理约 200 亿次搜索请求,处理数据量达数百 PB(千万亿字节)。

根据国际数据公司(IDC)《IDC:2025 年中国将拥有全球最大的数据圈》白皮书,中国在 2025 年将成为全球最大的数据圈。中国的数据圈将以 30%的年平均增长速度领先全球,2025 年中国数据圈增至 48.6ZB,占全球 27.8%。中国正处于数字经济的发展关键阶段,数据量增速之快,应用程度之深前所未有,将成为中国未来发展的关键资源。

2. Variety:数据类型多样

多样性体现数据来自多个数据仓库、数据领域以及多种数据类型。数据根据存储和处理方式可分为三种类型:结构化数据、非结构化数据和半结构化数据。在大数据中,80%～90%是非结构化数据,如社交媒体对话、视频、图片、地理位置、传感器数据、语音记录数据等。

多种类型的数据对数据处理能力提出了更高的要求。随着互联网时代的发展和大数据时代的到来,人们逐渐从信息匮乏的时代走向信息过载的时代,任何形式的数据都可以产生作用。不同类型的数据结合起来,使得人类可以从数据中汲取更多关键信息。这是传统时代无法想象的,例如,抖音的首页推荐是根据用户浏览视频的历史行为和兴趣进行推荐。这就是建立在非结构化的视频浏览数据挖掘基础上的应用,为用户提供个性化的信息服务和决策支持。

3. Velocity:处理速度快

处理速度快是大数据区别于传统数据挖掘的最显著特征。在海量数据场景下,传统的数据库技术已经无法满足其海量存储、高效处理和实时挖掘数据潜在价值的要求。对于很多业务场景来说,数据的时效性非常重要。例如,在携程等大型的 OTA 平台上浏览酒店后,平台会立即向用户推荐相关同类酒店或者与此酒店产品相搭配的其他产品,这也是根据用户当前浏览的商品数据进行实时推荐。在数据处理速度方面,有一个著名的“1 秒定律”,即要在秒级时间范围内给出分析结果,处理模式已经开始从批处理转向流处理。如果数据不能做到时效性,就失去了作用和价值。

4. Variability:数据多变性

大数据的体量、多样性和速度一直都是处于变化之中。在 2020 天猫双 11 全球狂欢季纪录之夜,根据双 11 实时交易数据,11 月 11 日 00:00:26,天猫双 11 迎来了流量洪峰,订单创建峰值达 58.3 万笔/秒。这个新纪录是 2009 年第一个天猫双 11 活动的 1 457 倍。

相对于传统的数据,大数据的质量、规模、结构化程度等都在随时变化,难以把握趋势,阻碍了有效处理和管理数据的进程。例如,百度、神马等互联网搜索引擎每秒钟都会产生和存储许多不同类型的数据,这些数据不断快速变化。此外,同样的数据在不同的语境中可能有不同的含义。因此,数据分析必须在上下文中进行,这给算法带来了挑战。

5. Value:价值

“Value(价值)”是指大数据有较高的商业价值。无论大数据有多大的容量和多少种类

型，如果没有应用价值，就没有意义。如何将业务逻辑与强大的机器算法相结合，从而能够最大限度地挖掘数据的价值？这是大数据应用的关键所在。例如，今日头条、抖音、西瓜视频等字节跳动旗下的 App，有一个共同的特点，就是基于大数据技术的精准推送机制。利用大数据分析，App 可以对用户特征、环境特征和内容特征进行匹配，进行这种操作，实现对不同用户的信息精准推送，让用户在信息过载、碎片化的互联网时代迅速获取自己关心的内容，相应地提升了用户使用时长及满意度。

6. Veracity：真实性

“Veracity（真实性）”是指数据的可信度，反映数据的质量。高质量的决策必须依赖于高质量的数据，而从现实世界中采集到的大数据很多情况下难以确保质量和准确性。一方面，要确保大数据在采集时候的有效性；另一方面，要通过大数据技术对“问题数据”进行预处理，达到“去伪存真”的效果。例如，现在大数据技术已经能够处理错别字、缩写等不准确的数据。微软在 Windows 版的 Microsoft Word 中引入了文本预测，使用机器学习功能，根据上下文和机器学习来提示用户接下来可能想要输入的文本来节省用户时间。这项功能还可以减少拼写和语法错误，并在用户使用一段时间后，根据他们的写作风格给予最佳建议。

二、大数据的发展历史

从数据本质来说，大数据并不是新的事物，数据分析以及相关技术自古就有。例如，在古代的战争中，统帅需要分析敌我军队的相关数据，确定最佳的策略和布局。

随着人类社会的发展和数据技术的进步，数据的数量、速度、类型也在不断提升。到了互联网时代，数据已经超出了人工以及传统数据技术可以处理的程度。随着非结构化的数据在数据总量中占比越来越大，大数据技术开始兴起，并随着移动互联网、物联网的普及而不断进步，因为移动互联网、物联网的应用使得数据以前所未有的速度增加。大数据的发展历史大致可以分为如下几个阶段。

1. 大数据 1.0 阶段

20 世纪 70 年代，为了有效管理和访问大量的数据资源，减少数据冗余和存储成本，关系型数据库技术应运而生。

20 世纪 80 年代，个人便携式计算机时代开始，释放了巨大的经济增长动力。而 20 世纪 90 年代兴起的互联网实现了数据的网络化传输，从而真正彻底改变了人类社会的信息交流和互动方式，造就了一批成功的企业并导致了一些企业的消亡，这些成功的企业通过破坏性创新推动了社会、企业发展和个人的成长，并进一步导致了数据的爆炸性增长和发展。

随着数据挖掘理论和数据库技术的逐步成熟，一些商业智能工具和知识管理技术开始得到应用，如数据仓库、专家系统、知识管理系统等。企业和组织对内部数据进行统计、分析和挖掘利用。

20 世纪 90 年代，万维网技术的出现促进了互联网上不同主机之间的信息共享。首先转型的是媒体行业，以门户网站为代表的网络媒体，改变了人们获取信息的方式，逐渐削弱了报纸、广播、电视等传统媒体的影响力。

2. 大数据 2.0 阶段

搜索引擎的诞生，改变了人类利用互联网获取信息的方式。数据挖掘技术用于分析和

处理网站用户访问 Web 服务器时产生的日志数据，从而发现 Web 用户的访问模式和兴趣爱好等，用于辅助站点管理和决策支持等，深入了解用户的需求和行为。

21 世纪初，随着移动互联网、社交媒体和新一代信息技术的发展，数据来源的丰富性使得大数据不仅包含传统的结构化数据，还包含半结构化数据和非结构化数据。用户通过社交媒体和其他在线服务平台产生了海量数据。这些数据的规模超过了典型数据库软件工具能够支持的存储、管理和分析能力。Hadoop（一个专门为存储和分析大数据集而创建的开源框架）在同一年被开发出来。NoSQL（非关系型数据库）也在这一时期开始流行起来。

3. 大数据 3.0 阶段

移动互联网、物联网的普及与应用，产生了大量的感知数据，如城市视频监控的流媒体数据和手机用户的使用数据。

互联网接入终端开始向移动设备发展。移动设备不仅可以分析用户行为数据（如点击和搜索查询），还可以存储和分析基于位置的数据（GPS 数据）。例如，当人们在运动时携带智能手机、智能手表、智能手环等移动设备，接入这些智能设备的大数据相关应用系统可以采集用户的运动数据，实时监测和反馈分析运动数据，甚至为健康提供基于数据分析的建议。

物联网（Internet of Things，IoT）是在互联网的基础上利用射频标签和无线传感器网络技术建立的覆盖所有人与物的网络化信息系统。物联网的兴起，带动了智慧城市的发展，可以利用数据、通信和技术来改善城市问题。通过物联网，居民与城市中的设备，如交通、自来水道、电力设备等形成有效的互动，最终提高政府的效率，改善人们的生活质量。

大数据已经渗透到今天的每一个行业和商业功能，并成为一个关键的生产要素。对海量数据的挖掘和使用预示着新一轮的生产力增长。随着越来越多的人、设备和传感器通过网络连接起来，人类产生、传输、分享和访问数据的能力正在发生革命性的变化。在旅游业，大数据技术得到快速发展，大数据应用变得越来越普遍，并改变了人们的旅行方式。例如，旅游者能够在网上搜索旅游信息、预订酒店、旅行途中打车或者租车、获取目的地旅游信息等，整个旅程都离不开大数据在背后的支持。对于能够利用大数据开展业务的旅游企业要比没有应用大数据开展业务企业要有更强的竞争优势和创新能力。

三、大数据与数据科学

大数据技术的迅速发展使得数据的处理越来越自动化，越来越多的数据采集、处理和分析都是由计算机进行自动化处理，而不是再通过人工去处理。因此，一门新兴的交叉学科——数据科学（Data Science）应运而生。

美国机器学习专家德鲁·康威（Drew Conway）总结了一个维恩交叉图来描述数据科学的概念。如图1-8所示，数据科学是由计算机科学、数学与统计、专业知识三个不同领域的结合体，不同领域的结合需要机器学习、软件工程和数据分析等技术支持。数据科学的目标是从数据中提取有意义的知识、洞察力和决策力。

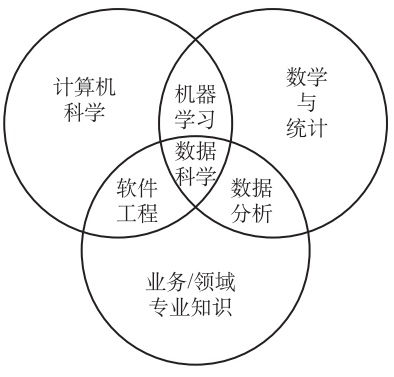


图 1-8　数据科学的概念

数据科学通常包括数据挖掘、预测、机器学习、预测分析、统计和文本分析等领域。这些领域过于专业，以至于寻找一个合适的数据科学专家并不容易。随着人工智能（Artificial Interlligence，AI）技术和机器学习算法的不断进步，如今，企业已经可以智能化使用数据科学家通过 AI 技术建立的各种分析模型来对数据进行自动化分析。这不仅提高了工作效率，而且使得非数据科学背景的业务人员也可以比以往任何时候都更加容易地使用各种分析模型，从业务数据中提取信息、知识和洞见，为业务增长建立竞争优势。

数据科学是一个旨在将数据转化为实际价值的跨学科领域，它包括数据提取、数据准备、数据探索、数据转换、存储和检索、计算基础设施、各种类型的挖掘和学习、解释和预测的展示，并同时考虑到道德、社会、法律和商业等方利用（Van Der Aalst，2016）。

2001 年美国统计学教授威廉·克利夫兰（William S. Cleveland）发表了《数据科学：拓展统计学的技术领域的行动计划》（*Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics*），奠定了数据科学的理论基础，数据课程成为一门独立的学科。

数据科学的工作过程如下。

1. 制定目标

制定目标的前提是理解业务，明确要解决的商业或业务问题是什么？比如，“如何提高客户满意度”，这是企业自上到下都很关心的问题，但从数据科学的角度，这个问题太空泛了，需要将这个业务问题进行分解为若干个小问题，形成可研究和可测试的问题。例如，“客户满意度的数据是如何衡量的？”“如何获得客户满意度的数据？”“现有的满意度是多少？”“创造什么条件会增加客户满意度？”

2. 数据理解与准备

当目标制定后，就要基于要解决的现实问题来理解和准备数据。数据理解和准备，一般需要解决下面的问题。

（1）需要哪些数据指标。

（2）数据指标的含义是什么。

（3）数据的质量如何。

（4）数据能否满足需求。

（5）数据是否还需要加工。

（6）探索数据中的规律和模式，进而形成假设。

在实际应用当中，数据理解和准备工作往往可能需要尝试很多次。因为在复杂的大数据场景下，发现数据中存在的模式是一个不断试错、不断优化的过程。

3. 数据建模

在准备好的数据基础上，建立数据模型。这种模型可能是机器学习模型，也可能不需要机器学习等高深的算法。选择什么样的模型，是根据要解决的问题和制定的目标来确定的。

当然也可以选择两个或多个模型来进行对比，同时不断调整优化模型的参数，使得模型的效果不断优化。

4. 模型评估

模型效果的评估有两个方面：一是模型是否解决了需要解决的问题（是否还有没有注意

和考虑到的潜在问题需要解决)；二是模型的精确性是否符合要求(误差率或者残差是否符合正态分布等)。

5. 结果呈现

结果呈现主要关注以下方面。

(1) 模型解决了哪些问题。

(2) 模型的解决效果如何。

(3) 解决问题的具体操作步骤是什么。

6. 模型部署

数据模型确定后，一般情况需要通过技术环境部署落实，以便业务人员进行有效使用。业务人员不需要了解数据模型背后的复杂技术原理和算法，但是可以选择出最优的模型帮助决策。

同时，还要注意部署后的数据模型并非一成不变，而是需要持续优化，应该是一个周期性循环的过程，如图 1-9 所示。

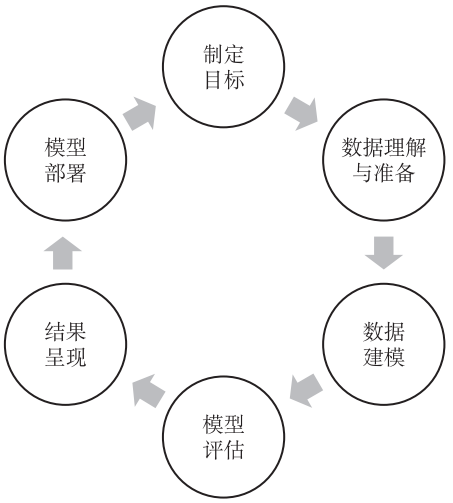


图 1-9　数据科学工作过程

【主要术语】

1. 大数据(Big Data)：是具有体量巨大、来源多样、生成极快且多变等特征并且难以用传统数据体系结构有效处理的包含大量数据集的数据。

2. 数据挖掘(Data Mining)：是通过算法搜索隐藏在大量数据中的信息的过程。通常与计算机科学有关，通过统计学、在线分析处理、情报检索、机器学习、专家系统和模式识别等多种方法来实现。

3. 数据库(Data Base)：是一个长期存储在计算机中，有组织、可共享、统一管理的大型数据集合。

4. 峰值(Peak)：是起伏变化的数值中的最大值。

5. 体量(Volume):是指大数据的数据量。大数据的“大”首先体现在数据量上。采集、存储和计算数据量的起始计量单位往往是 TB(1 024GB)、PB(1 024TB)。

6. 多样性(Variety):是指类型和来源的多样化,种类上包括结构化数据、半结构化数据和非结构化数据,具体包括网络日志、音频、视频、图片、地理位置信息等。

7. 速度(Velocity):这是大数据区分于传统数据挖掘的最显著特征。例如,搜索引擎要求几分钟前的新闻能够被用户查询到,个性化推荐算法尽可能要求实时完成推荐。

8. 多变性(Variability):是指数据的含义总是在快速变化。例如,在同一条推文中,一个词可以有完全不同的含义。

9. 价值(Value):是指所有可用的数据可以为组织、社会和消费者创造巨大的价值。如何结合商业逻辑,通过强大的机器算法来挖掘数据的价值,是大数据时代需要解决的最重要的问题。

10. 真实性(Veracity):组织需要确保数据的真实性,以保证数据分析的正确性。如果数据本身是虚假的,那么就失去了存在的意义,通过虚假数据得出的任何结论都可能是错误的,甚至是相反的。

11. 搜索引擎(Search Engine):是一种信息检索系统,旨在协助搜索存储在计算机系统中的信息。

12. 算法(Algorithm):是对一个问题的解决方案的准确和完整描述,包含多个步骤,可以对数据进行操作,以解决特定问题。

13. 移动互联网(Mobile Internet):是将移动通信和互联网两者合二为一。是将互联网的技术、平台、商业模式和应用与移动通信技术相结合并进行实践的活动的总称。

14. 关系型数据库(Relational Database):最典型的数据结构是表,是一种由二维表和表之间的联系组成的数据组织。

15. 数据仓库(Data Warehouse):是一种面向商务智能(Business Intelligence,BI)活动(尤其是分析)的数据管理系统,使用数据仓库的人主要是管理和运营人员,通过对历史数据的分析和洞察来做出相应的商业决策。

16. 专家系统(Expert System):是一个智能计算机程序系统,它包含了大量的知识和经验,达到了某一领域专家的水平,能够使用人类专家的知识和解决问题的方法来处理该领域的问题。

17. 知识管理系统(Knowledge Management System):是收集、处理和分享一个组织的全部知识的信息系统,通常由计算机系统支持。

18. 万维网(World Wide Web):是文件、图片、多媒体和其他资源的集合,使用统一资源标识符(URL)标识,由许多超链接互相链接形成。

19. 日志数据(Log Data):用户名、用户执行的程序名称、日期、时间等被写进日志文件(通常以 *.log 结尾)。这是为了将来系统出现故障时可以进行追踪。

20. 非关系型数据库(Non-Relational Database):与关系型数据库相比,通过减少用不到或很少用的功能,来大幅度提高产品性能。

21. 数据科学(Data Science):数据科学是由计算机科学、数学与统计、专业知识三个不同领域的结合体,不同领域的结合需要机器学习、软件工程和数据分析等技术支持。数据科学的目标是从数据中提取有意义的知识、洞察力和决策力。

22. 数据模型(Data Model):是数据库设计的起始阶段,通常由属性、实体类型、完整性规则、对象关系和定义组成。

【练习题】

一、自测题

1. 数据和大数据有什么区别?

2. 大数据“6V”特征具体表现在哪些方面?请举例说明。

3. 大数据的发展历史分为几个阶段?每个阶段的发展重点是什么?

4. 数据科学具体包括哪些内容?

5. 数据科学的工作过程可分为哪几个步骤?

二、讨论题

1. 如今,大数据已经初具规模,并在很多方面得到了应用。以小组为单位,找出不少于3个大数据在日常生活中的应用实例,并说明对人类生活有哪些便利(或不便)的影响。

2. 请根据大数据的“6V”特征,讨论一下旅游或者酒店行业对应的应用场景。

三、实践题

大数据被广泛应用于各个行业和领域,带来商业变革、管理变革和思维变革,在此背景下,包括旅游业与酒店业在内的以终端消费者为服务目标市场的行业迎来了新的发展机遇。以小组为单位,选取一家酒店(集团)或者旅游企业,通过包括但不限于电话访谈、问卷调查、实地调研等方法,了解该企业大数据的应用现状和未来大数据应用的设想,并撰写调研报告进行汇报。