

企业级卓越人才培养解决方案“十三五”规划教材

ZOUJIN DASHUJU YU RENGONG ZHINENG

走进大数据与 人工智能

天津滨海迅腾科技集团有限公司◎编著

 天津大学出版社
TIANJIN UNIVERSITY PRESS

企业级卓越人才培养解决方案“十三五”规划教材

走进大数据与人工智能

天津滨海迅腾科技集团有限公司 编著



 天津大学出版社
TIANJIN UNIVERSITY PRESS

图书在版编目(CIP)数据

走进大数据与人工智能 / 天津滨海迅腾科技集团有
限公司编著. —天津: 天津大学出版社, 2018.8
企业级卓越人才培养解决方案“十三五”规划教材
ISBN 978-7-5618-6224-7

I. ①走… II. ①天… III. ①数据处理—教材 ②人工
智能—教材 IV. ①TP274 ④TP18

中国版本图书馆CIP数据核字(2018)第183592号

主 编: 王 翔 周 勇 畅玉洁 房 屹
副主编: 史永博 孙学朋 万冬娥 董善志
向 元 陈怀玉

出版发行 天津大学出版社
地 址 天津市卫津路92号天津大学内(邮编:300072)
电 话 发行部:022-27403647
网 址 publish.tju.edu.cn
印 刷 天津泰宇印务有限公司
经 销 全国各地新华书店
开 本 185mm×260mm
印 张 15.75
字 数 387千
版 次 2018年8月第1版
印 次 2018年8月第1次
定 价 49.00元

凡购本书, 如有缺页、倒页、脱页等质量问题, 烦请向我社发行部门联系调换
版权所有 侵权必究

企业级卓越人才培养解决方案“十三五”规划教材 编写委员会

- 指导专家：** 周凤华 教育部职业技术教育中心研究所
李 伟 中国科学院计算技术研究所
张齐勋 北京大学
朱耀庭 南开大学
潘海生 天津大学
董永峰 河北工业大学
邓 蓓 天津中德应用技术大学
许世杰 中国职业技术教育网
郭红旗 天津软件行业协会
周 鹏 天津市工业和信息化委员会教育中心
邵荣强 天津滨海迅腾科技集团有限公司
- 主任委员：** 王新强 天津中德应用技术大学
- 副主任委员：** 张景强 天津职业大学
宋国庆 天津电子信息职业技术学院
闫 坤 天津机电职业技术学院
刘 胜 天津城市职业学院
郭社军 河北交通职业技术学院
刘少坤 河北工业职业技术学院
麻士琦 衡水职业技术学院
尹立云 宣化科技职业学院
王 江 唐山职业技术学院
廉新宇 唐山工业职业技术学院
张 捷 唐山科技职业技术学院
杜树宇 山东铝业职业学院
张 晖 山东药品食品职业学院
梁菊红 山东轻工职业学院
赵红军 山东工业职业学院
祝瑞玲 山东传媒职业学院

王建国 烟台黄金职业学院
陈章侠 德州职业技术学院
郑开阳 枣庄职业学院
张洪忠 临沂职业学院
常中华 青岛职业技术学院
刘月红 晋中职业技术学院
赵 娟 山西旅游职业学院
陈 炯 山西职业技术学院
陈怀玉 山西经贸职业学院
范文涵 山西财贸职业技术学院
任利成 山西轻工职业技术学院
郭长庚 许昌职业技术学院
李庶泉 周口职业技术学院
许国强 湖南有色金属职业技术学院
孙 刚 南京信息职业技术学院
夏东盛 陕西工业职业技术学院
张雅珍 陕西工商职业学院
王国强 甘肃交通职业技术学院
周仲文 四川广播电视大学
杨志超 四川华新现代职业学院
董新民 安徽国际商务职业学院
谭维奇 安庆职业技术学院
油俊伟 天津大学出版社

企业级卓越人才培养解决方案简介

企业级卓越人才培养解决方案(以下简称“解决方案”)是面向我国职业教育量身定制的应用型、技术技能人才培养解决方案,是以教育部—滨海迅腾科技集团产学研协同育人项目为依托,依靠集团研发实力,联合国内职业教育领域相关政策研究机构、行业、企业、职业院校共同研究与实践的科研成果。本解决方案坚持“创新校企融合协同育人,推进校企合作模式改革”的宗旨,消化吸收德国“双元制”应用型人才培养模式,深入践行基于工作过程“项目化”及“系统化”的教学方法,设立工程实践创新培养的企业化培养解决方案。在服务国家战略:京津冀教育协同发展、“中国制造 2025”(工业信息化)等领域培养不同层次的技术技能人才,为推进我国实现教育现代化发挥积极作用。

该解决方案由“初、中、高”三个培养阶段构成,包含技术技能培养体系(人才培养方案、专业教程、课程标准、标准课程包、企业项目包、考评体系、认证体系、社会服务及师资培训)、教学管理体系、就业管理体系、创新创业体系等,采用校企融合、产学融合、师资融合“三融合”的模式在高校内共建大数据(AI)学院、互联网(商务)学院、软件学院、电子商务学院、设计学院、智慧物流学院、智能制造学院等,并以“卓越工程师培养计划”项目的形式推行,将企业人才需求标准、工作流程、研发规范、考评体系、企业管理体系引入课堂,充分发挥校企双方优势,推动校企、校际合作,促进区域优质资源共建共享,实现卓越人才培养目标,达到企业人才招录的标准。本解决方案已在全国几十所高校开始实施,目前已形成企业、高校、学生三方共赢的格局。

天津滨海迅腾科技集团有限公司创建于 2004 年,是以 IT 产业为主导的高科技企业集团。集团业务范围已覆盖信息化集成、软件研发、职业教育、电子商务、互联网服务、生物科技、健康产业、日化产业等。集团以科技产业为背景,与高校共同开展“三融合”的校企合作混合所有制项目。多年来,集团打造了以博士、硕士、企业一线工程师为主导的科研及教学团队,培养了大批互联网行业应用型技术人才。集团先后荣获:天津市“五一”劳动奖状先进集体、天津市政府授予“AAA”级劳动关系和谐企业、天津市“文明单位”“工人先锋号”“青年文明号”“功勋企业”“科技小巨人企业”“高科技型领军企业”等近百项荣誉。集团将以“中国梦,腾之梦”为指导思想,在 2020 年实现与 100 所以上高校合作,形成教育科技生态圈格局,成为产学研协同育人的领军企业。到 2025 年形成教育、科技、现代服务业等多领域 100% 生态链,实现成为教育科技行业“中国龙”的目标。

前 言

大数据与人工智能是当今前沿的计算机科学技术。近年来,关于大数据与人工智能的讨论与研究一直在持续。越来越多使用大数据与人工智能技术的软件与网站出现在我们的生活之中,为无数人的生活与工作带来了便利。与此同时,大数据技术与人工智能技术渗透到了各个行业之中,颠覆了很多常规行业的运作模式。

本书主要介绍了大数据与人工智能的相关概念,通过案例事件的叙述,讲述了大数据与人工智能技术的来龙去脉及发展历程。结合大量的案例讲解了大数据与人工智能技术在现实生活中的应用,通过对大数据与人工智能现状的深入剖析,展望了大数据与人工智能的发展方向与未来。

本书共九个学习情境,以“大数据与人工智能概论”→“大数据技术与数据处理”→“大数据在各行业应用案例”→“语音、语义识别”→“计算机视觉识别”→“人工智能芯片”→“机器人”→“人工智能在各行业应用”→“大数据与人工智能未来”为线索,全方位地阐述了在学习大数据与人工智能技术之前,需要了解和掌握的指导性思想与相关概念。

本书采用基于工作过程系统化的设计思路,每个学习情境由多个学习任务组成,每个学习任务都分为问题导入、学习目标、学习概要、学习内容、知识回顾五个模块来讲解相应的知识点。

本书由王翔、周勇、畅玉洁、房屹任主编,由史永博、孙学朋、万冬娥、董善志、向元、陈怀玉等共同任副主编,畅玉洁、房屹负责统稿,王翔、周勇负责全面内容的规划,史永博、孙学朋负责整体内容编排。具体分工如下:情境一至情境三由史永博、孙学朋编写,王翔负责全面规划;情境四至情境七由董善志、向元共同编写,周勇负责全面规划;情境八至情境九由陈怀玉、万冬娥共同编写,王翔负责全面规划。

本书的主旨是大数据与人工智能技术的概念普及,定位于大数据与人工智能技术的初学者。与市面上现有的理论指导性书籍不同,本书通过采用大量的案例与历史事件,生动、形象且不失深度地讲解了大数据与人工智能的概念。通过对本书的学习,读者可以提高对大数据与人工智能技术的理论性认识。

天津滨海迅腾科技集团有限公司
技术研发部

目 录

学习情境一 大数据与人工智能概论	1
任务一 大数据概论	1
问题导入	1
学习目标	1
学习概要	2
学习内容	2
知识回顾	11
任务二 人工智能概论	11
问题导入	11
学习目标	12
学习概要	12
学习内容	12
知识回顾	26
学习情境二 大数据技术与数据处理	27
任务一 大数据技术	27
问题导入	27
学习目标	27
学习概要	28
学习内容	28
知识回顾	42
任务二 数据处理工具与数据安全	43
问题导入	43
学习目标	43
学习概要	44
学习内容	44
知识回顾	48
学习情境三 大数据在各行业应用案例	49
任务一 企业大数据	49
问题导入	49
学习目标	49
学习概要	50

学习内容	50
知识回顾	62
任务二 企业级大数据应用	63
问题导入	63
学习目标	63
学习概要	64
学习内容	64
学习情境四 语音、语义识别	77
任务一 语音识别技术概述	77
问题导入	77
学习目标	77
学习概要	78
学习内容	78
知识回顾	84
任务二 语音识别应用	85
问题导入	85
学习目标	85
学习概要	86
学习内容	86
任务三 语义识别技术概述	97
问题导入	97
学习目标	97
学习概要	98
学习内容	98
知识回顾	103
任务四 语义识别应用	104
问题导入	104
学习目标	104
学习概要	105
学习内容	105
学习情境五 计算机视觉识别	111
任务一 计算机视觉概述	111
问题导入	111
学习目标	111
学习概要	112
学习内容	112
知识回顾	124

任务二 计算机视觉应用	125
问题导入	125
学习目标	125
学习概要	126
学习内容	126
学习情境六 人工智能芯片	138
任务一 人工智能芯片概述	138
问题导入	138
学习目标	138
学习概要	139
学习内容	139
知识回顾	146
任务二 人工智能芯片性能分析	147
问题导入	147
学习目标	147
学习概要	148
学习内容	148
知识回顾	158
任务三 人工智能芯片产业生态分析	159
问题导入	159
学习目标	159
学习概要	160
学习内容	160
知识回顾	168
学习情境七 机器人	169
任务一 机器人概述	169
问题导入	169
学习目标	169
学习概要	170
学习内容	170
知识回顾	185
任务二 机器人案例分析及发展趋势	186
问题导入	186
学习目标	186
学习概要	187
学习内容	187
知识回顾	194

学习情境八 人工智能在各行业应用	195
任务一 人工智能 + 金融	195
问题导入	195
学习目标	195
学习概要	196
学习内容	196
知识回顾	199
任务二 人工智能 + 家居	200
问题导入	200
学习目标	200
学习概要	201
学习内容	201
知识回顾	208
任务三 人工智能 + 医疗	209
问题导入	209
学习目标	209
学习概要	210
学习内容	210
知识回顾	214
任务四 人工智能 + 安防	215
问题导入	215
学习目标	215
学习概要	216
学习内容	216
知识回顾	221
学习情境九 大数据与人工智能未来	222
任务一 大数据技术的未来	222
问题导入	222
学习目标	222
学习概要	223
学习内容	223
知识回顾	232
任务二 人工智能的未来	233
问题导入	233
学习目标	233
学习概要	234
学习内容	234
知识回顾	240

学习情境一 大数据与人工智能概论

任务一 大数据概论

问题导入



学习目标

通过对大数据概论的学习，了解什么是大数据，熟悉大数据从概念提出至今经历了哪些阶段，掌握大数据迅速发展的原因，根据大数据现状分析大数据的机遇。在任务实现过程中：

- 了解什么是大数据。
- 熟悉大数据经历的阶段。

- 掌握大数据发展的原因。
- 根据大数据现状分析大数据的机遇。

学习概要



学习内容

早在 1980 年,美国作家阿尔文·托夫勒在其未来学著作《第三次浪潮》中对“大数据”就有所提及。在书中,阿尔文·托夫勒将大数据预言为:将是第三次浪潮的华彩乐章。而对大数据的定义,直到近几年且在各方拥有保留意见的情况下才被确立下来。

大数据是一个体量大、数据类别多的数据集合,并且无法在一定时间范围内使用传统数据库工具对其内容进行抓取、解析、管理和处理。

一、什么是大数据

自大数据产生以来,就有很多机构或组织想要给大数据下一个权威的定义来规范大数据的特性。对大数据定义最具代表性的是认为大数据必须满足三个特点才能被称为“大数据”。大数据的三个特点又被称为 3V 特性,即规模性(Volume)、多样性(Variety)和高速性(Velocity)。如图 1.1 所示。

1. 规模性

所谓规模性就是指数据量庞大、数据存储体量大和计算量大。目前,整个社会各个行业每天要产生 EB 级别的数据量,因此大数据中的数据计算单位已经不能再用传统的 GB 或者 TB,而要用 PB、EB 甚至 ZB 为计量单位。

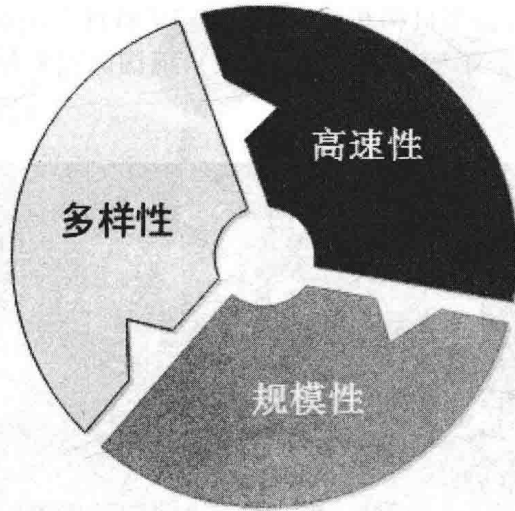


图 1.1 大数据的三个特性

2. 多样性

多样性是指数据的种类繁多。造成数据种类繁多的原因是互联网技术和科学技术在不断发展。由于传感器的规格、数据来源的网站类型不同,数据的格式也不同。数据可分为结构化数据、半结构化数据和非结构化数据三种类型。大数据处理中三种类型数据比例约为 1:3:6。如图 1.2 所示。

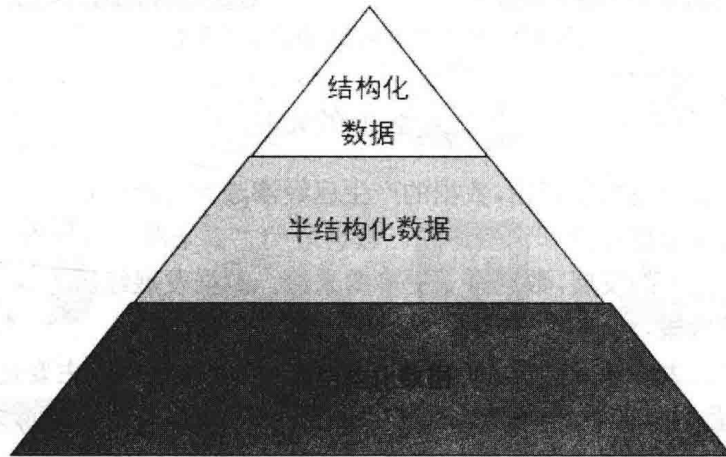


图 1.2 三种类型数据比例

3. 高速性

数据的高速性主要体现在两个方面,具体如下。

一方面指数据的增长速度十分迅猛。在短短的 60 秒之内,淘宝和天猫就有超过 14 万人访问,有 774 人产生交易(除促销外);百度要处理 340 万次以上的搜索请求并极快地返回结果;新浪微博会发送超过 9 万条新微博。中国互联网 60 秒发生的事情如图 1.3 所示。

另一方面指数据存储、传输等处理速度十分迅捷。如对灾难预测,在灾难发生后,需要极快地对灾难的发生程度、影响范围进行预测。例如:在 2011 年 3 月 11 日日本大地震发生后仅 9 分钟,美国国家海洋和大气管理局(NOAA)就发布了详细的海啸预警。

除上述三个特性之外,很多机构和公司尝试在 3V 特性上增加大数据的新特性。国际数据公司 (IDC) 认为大数据还具有价值性 (Value), 而国际商业机器公司 (IBM) 认为大数据拥有真实性 (Veracity)。

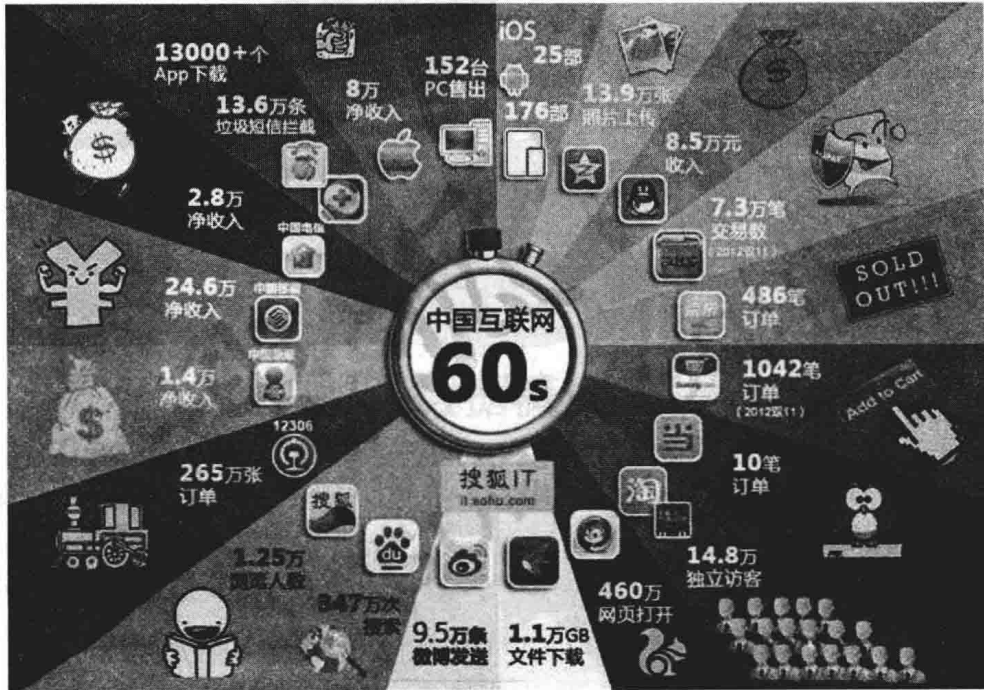


图 1.3 中国互联网 60 秒发生的事情

二、大数据的发展

大数据并不仅仅是猜想和理论,数据的产生已经渗透到传统行业之中,对大数据的挖掘和运用也为人们带来新的生产和消费方式。数据最初主要来源于数据库(结构化数据),但随着互联网和科学技术的发展,数据有了更多的来源。数据发展经历了以下三个阶段。

1. 运营式系统阶段

数据库的出现大大降低了数据管理的复杂程度。现实中数据库主要是作为运营系统存储数据,或作为运营系统的数据管理子系统存在的,比如超市的销售记录系统、医院病人的医疗记录系统和银行的交易记录系统等。由于运营式系统广泛使用数据库来存储数据,人类社会数据存储量第一次有了大的飞跃。这个阶段最主要的是数据伴随着一定的运营活动而产生并记录在数据库中,比如消费者进行一次交易,就会在银行的交易系统数据库中产生对应数据。此阶段数据的产生方式是被动的。常见数据库种类如图 1.4 所示。

2. 用户原创内容阶段

社交媒体和智能设备的出现,导致数据存储量发生了第二次大的飞跃。智能设备可以使用户全天候不间断地连接到互联网及社交媒体(如微博、微信、QQ、抖音、快手等)。互联网及社交媒体也可以同时在不同的智能设备上运行。二者的结合导致了数据量的暴增,用户产生数据的意愿也更加强烈。此阶段数据的产生方式是主动的。常见社交媒体如图 1.5 所示。

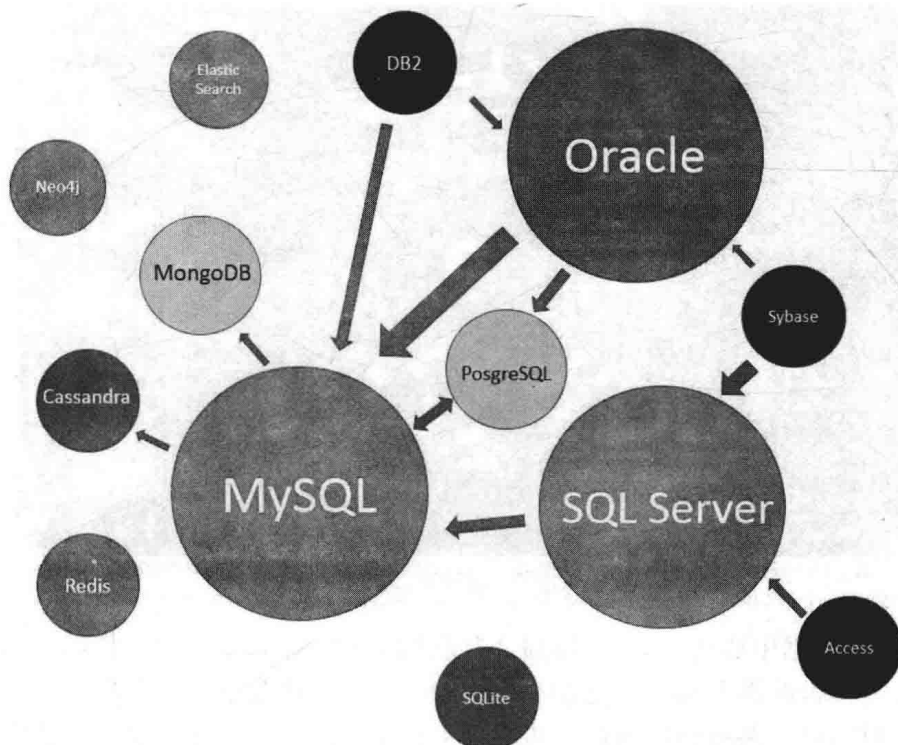


图 1.4 常见数据库种类



图 1.5 常见社交媒体

3. 感知式系统阶段

感知式系统的广泛应用,导致了数据存储量发生第三次大的飞跃,最终导致了大数据的产生。由于技术的发展,人类已经有能力制造极其微小的、带有处理功能的传感器用来收集数据,并开始将这些设备广泛地布置于社会的各个角落,通过这些设备来对整个社会的运转进行监控和信息存储。这些设备会源源不断地产生新数据。此阶段数据的产生方式是自动的,如大型的粒子对撞机平均每秒就可以产生 1GB 的数据。粒子对撞机如图 1.6 所示。

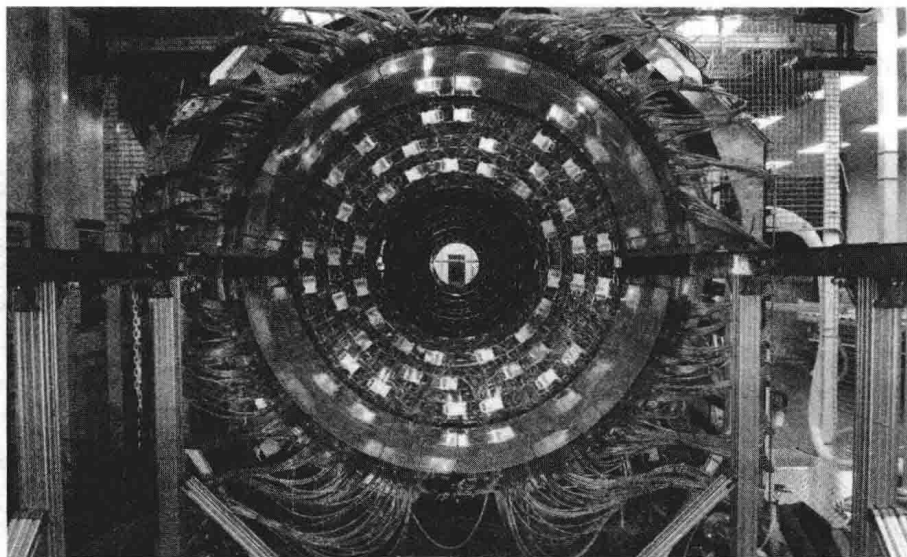


图 1.6 粒子对撞机(每秒可以产生 1GB 数据)

美国互联网数据中心研究显示,现如今互联网上的数据每年将增长 50% 以上,每两年便将翻一倍。目前,世界上 90% 以上的数据是最近几年才产生的。全世界的工业生产产品有着无数的数码传感器,通过即时测量和传递有关位置、运动、震动、温度、湿度等信息产生了海量的数据信息,这些都是大数据的数据来源。

三、大数据迅速发展的原因

在科学技术迅速发展的今天,大数据和人工智能有着前所未有的机遇。政策扶持、资本注入是大数据与人工智能技术最大的两个机遇。

1. 政策扶持

2015 年 8 月 31 日,中华人民共和国国务院正式印发《促进大数据发展行动纲要》(以下简称《纲要》)。《纲要》中提出主要任务是加快政府数据开放共享,推动资源整合,提升治理能力;推动产业创新发展,培育新兴业态,助力经济转型;强化安全保障,提高管理水平,促进健康发展。图 1.7 为《纲要》印发通知(图片来源:中国政府网)。

国务院关于印发促进大数据发展 行动纲要的通知

国发〔2015〕50号

各省、自治区、直辖市人民政府,国务院各部委、各直属机构:

现将《促进大数据发展行动纲要》印发给你们,请认真贯彻落实。

国务院

2015年8月31日

图 1.7 《纲要》印发通知

2016 年 3 月 16 日第十二届全国人大第四次会议审查通过了《中华人民共和国国民经济和社会发展第十三个五年规划纲要》(以下简称《“十三五”规划》)。《“十三五”规