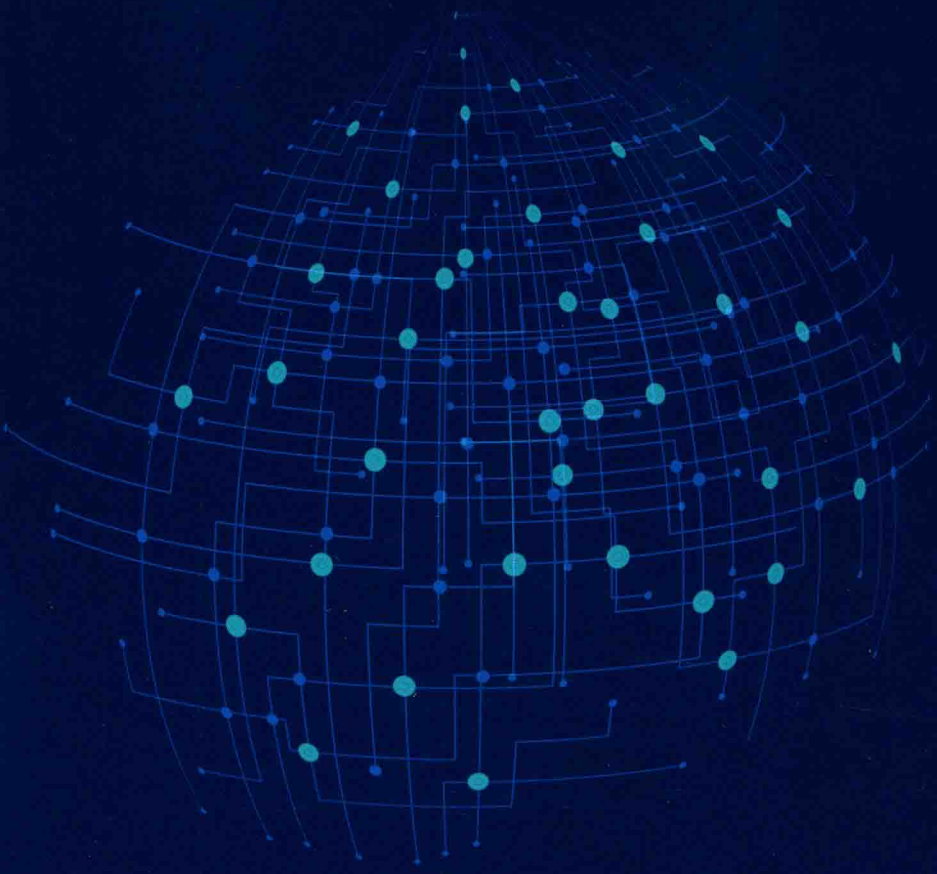
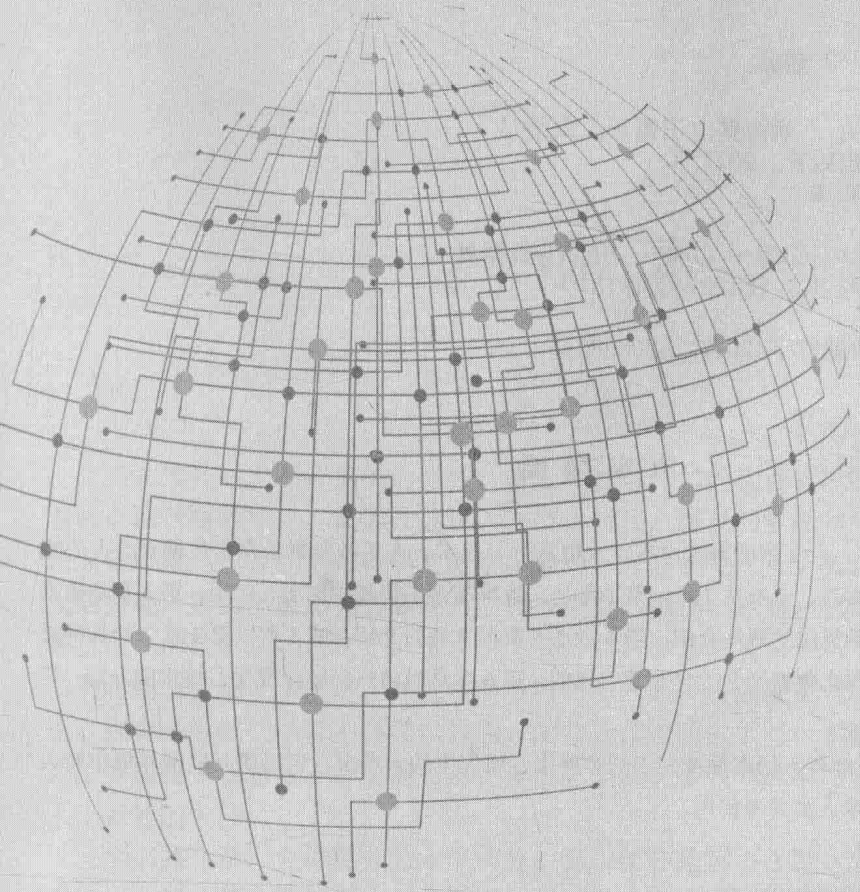


INTRODUCTION TO BIG DATA AND ARTIFICIAL INTELLIGENCE

大数据与人工智能导论

姚海鹏　王露瑶　刘韵洁◎著

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS



INTRODUCTION TO BIG DATA AND ARTIFICIAL INTELLIGENCE

大数据与人工智能导论

姚海鹏　王露瑶　刘韵洁◎著

人民邮电出版社
北京

图书在版编目（CIP）数据

大数据与人工智能导论 / 姚海鹏，王露瑶，刘韵洁著. -- 北京 : 人民邮电出版社，2017.9
ISBN 978-7-115-46602-0

Ⅰ. ①大… Ⅱ. ①姚… ②王… ③刘… Ⅲ. ①数据处理－研究②人工智能－研究 Ⅳ. ①TP274②TP18

中国版本图书馆CIP数据核字(2017)第208980号

内 容 提 要

本书主要涉及数据工程、人工智能算法原理，大数据平台技术、人工智能算法在大数据平台上的实现、人工智能算法的应用与实践。全书共 7 章。第 1 章是大数据与人工智能的历史、应用；第 2 章是数据工程；第 3 章是人工智能基础算法的原理介绍；第 4 章是大数据平台的介绍；第 5 章以第 3 章中的几种算法为例，介绍了它们是如何在大数据平台上分布式实现的；第 6 章是当前热门的深度学习技术的介绍；第 7 章是实践。

本书可作为希望快速了解和入门本领域知识的本科生、研究生的参考书，也可供互联网领域中对人工智能算法感兴趣的工程技术人员参考使用。

◆ 著　　姚海鹏　王露瑶　刘韵洁
责任编辑　邢建春
责任印制　彭志环

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京市艺辉印刷有限公司印刷

◆ 开本：787×1092　1/16
印张：12　　2017 年 9 月第 1 版
字数：292 千字　　2017 年 9 月北京第 1 次印刷

定价：79.00 元

读者服务热线：(010)81055488　印装质量热线：(010)81055316
反盗版热线：(010)81055315

前 言

当人类开始发明计算机的时候，就已经在思考如何让计算机获取“智能”。如今，伴随着社会的日益数字化，人类社会进入大数据时代，海量的数据和云计算使人工智能进入一个快速发展期。作者在平时科研与教学中发现，许多学生对人工智能与大数据领域表现出了极大的兴趣。市面上虽然有许多优秀的相关类型的图书，然而，它们大部分都假设读者已经具备了很高的数学基础，这是许多学生尤其是非数学系的本科生所不具备的。除此之外，有的图书对细节介绍非常详尽，这使初学者容易陷入其中而忽视了整体。有的图书则完全介绍理论，也容易导致初学者不会学以致用。因此，作者萌生了写一本真正适合初学者的大数据与人工智能图书的想法，希望能指引更多有志于研究该领域的学习者少走一些弯路，顺利迈入人工智能的大门。

全书共 7 章，大体可分为 4 个部分。第一部分是第 1~2 章，介绍人工智能和大数据的基本知识；第二部分是第 3~4 章，介绍一些最经典而常用的机器学习算法和常用的大数据处理平台；第三部分是第 5~6 章，介绍一些进阶知识，包括一些机器学习算法的并行化实现和深度学习的内容；第四部分即第 7 章是实践部分，用案例来学习前面学习的算法是如何应用在实际中的。其中，第三部分的两章相对独立，读者可以根据自己的兴趣和时间情况选择使用。

作者认为，对于初学者而言，应该适当增加学习的广度而降低学习的深度。本书对数据工程、机器学习、大数据以及机器学习的并行化实现、深度学习均予以介绍，已经涵盖了大部分人工智能的基础性内容。需要指出的是，理论上机器学习只是人工智能的一种解决方案。然而，近年来，机器学习已经在人工智能领域中占据了绝对主导地位。因此本书作为一本旨在服务初学者的图书，并不会区分它们。作者在本书中试图尽可能地少使用数学知识，对于一些不可避免的部分，力求展现其中的精华，同时亦在本书的最后介绍了一些确实不可避免的线性代数以及概率论知识。作者在保证广度和淡化深度的同时，避免了罗列知识，而是有机地将各方面知识串接起来。对于应用实践部分，一些基础性编程语言的掌握是必不可少的，作者在附录中介绍了 Java 和 Python，供有需要的读者阅读。

本书的作者都是在大数据和人工智能领域具有丰富教学实践经验或研究经验的专家，可以说凝结了多人的智慧和心血，其中北京邮电大学未来网络理论与应用实验室的姚海鹏副教授带领研究生团队编写了第 1、3、4、5、6、7 章，北京工业大学未来网络高精尖创新中心的王露瑶博士参与编写了第 2 章。中国工程院院士刘韵洁对本书做了重要的指导。全书由姚海鹏副教授通稿。

特别感谢北京邮电大学未来网络理论与应用实验室的研究生付丹阳、刘惠文、王淇艺、张博、陈旭、董理、刘冲以及北京工业大学未来网络高精尖创新中心的研究生李飞翔、买天乐、

张贵娟、郭倩影、范春明、贾耀宗、张楠。他们为本书的调研、材料收集、书稿撰写做了大量的研究工作，同时也结合自身研究实际，为本书提出了大量建设性建议。

大数据和人工智能发展迅速，目前已发展成为多个学科。作为一本基础性教材，本书希望能够带领读者入门，为读者进一步在这个领域深造打下坚实的基础。作者自认为才疏学浅，同时编写书稿时间也比较仓促，书中个别谬误之处在所难免，还望相关专家学者批评指正。

姚海鹏

2017 年 7 月于北京

目 录

第1章 绪 论

1.1 日益增长的数据

随着移动通信技术和智能终端设备的飞速发展，全球数据通信总量也逐年激增。一方面，由于数据产生方式发生了从手工生产到自动化生产的改变，人类为了实现对信息的全量化收集，大量使用传感器（目前全球有 3 B~5 B 个传感器），这些传感器 24 h 都在产生数据，加快了信息的爆发式增长；另一方面，由于人类活动越来越离不开数据，人类的日常生活已经与数据成为密不可分的整体。伴随着移动智能设备的普及（图 1-1 中的数据显示了近几年全球网络用户数量的变化），移动端的数据已经逐步增长并成为最主要的数据来源：社交通信中产生的文字、语音、图像、视频，生活应用中的位置信息、查询请求信息，娱乐购物产生的产品介绍信息、订单请求信息等无时无刻不在人们周围产生并传递。举例来说，Youtube 上每天会有来自全球 28.8 k 小时的视频上传量，Twitter 上每天大概会新增 50 M 条信息，亚马逊每天产生 6.3 M 笔订单……欧洲粒子物理研究所的大型强子对撞机，每秒产生的原始数据量高达 40 TB。2000 年斯隆数字巡天项目（SDSS，Sloan Digital Sky Survey）启动的时候，位于墨西哥州的望远镜在短短几周内收集到的数据比之前天文学历史上收集的数据总和还要多。从科研领域到医疗卫生领域，从银行业到互联网行业，各行各业都面临着需要解决爆发式增长的数据量的难题。

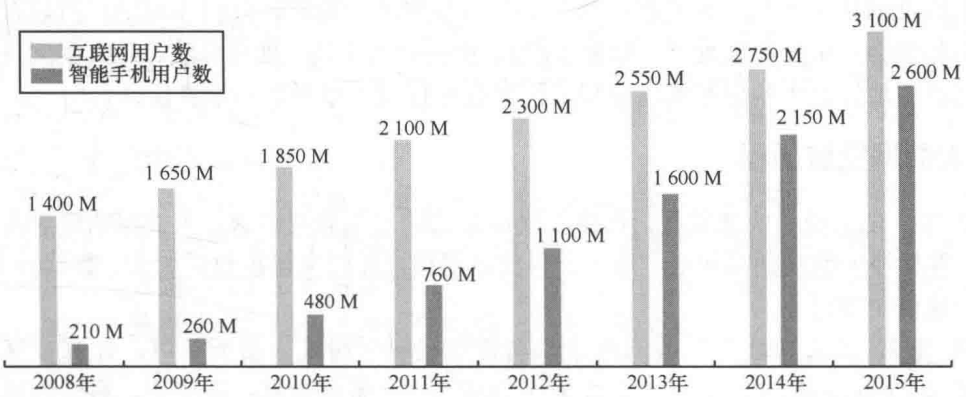


图 1-1 2008~2015 年全球网络用户数目增长情况

根据南加州大学嫩伯格通信学院马丁的研究，人类在2007年存储了超过300 EB的数据，也就是3×10^{11} GB的数据量（数据单位转化如表1-1所示），已经进入了数据海量激增的时代。人类存储信息量的增长速度比世界经济的增长速度快4倍，而计算即数据处理能力的增长速度比世界经济的增长速度快9倍。大数据的时代已经到来。

表1-1 数据单位转化

单位	定义	字节数
Byte	1 Byte	2^0
Kilobyte	1 024 Byte	2^{10}
Megabyte	1 024 Kilobyte	2^{20}
Gigabyte	1 024 Megabyte	2^{30}
Terabyte	1 024 Gigabyte	2^{40}
Petabyte	1 024 Terabyte	2^{50}
Exabyte	1 024 Petabyte	2^{60}
Zettabyte	1 024 Exabyte	2^{70}
Yottebyte	1 024 Zettabyte	2^{80}

1.1.1 大数据基本概念

“大数据”一词最早出现在20世纪90年代的美国，但直到2012年之后，大数据才逐渐获得了业界更多的关注和重视。其覆盖面之广涉及物理学、生物学、环境生态学、金融学以及军事领域、通信领域。当下对“大数据”的相关研究也很火热。

那么，何谓大数据？其实在最开始的时候，大数据并没有一个确切的概念，而是指需要处理的信息量很大，已经超过一半电脑在处理数据时所能使用的内存量，所以迫使工程师们必须改进处理数据的工具。在这种驱动下，谷歌的MapReduce和开源的Hadoop平台的出现使人们可以处理的数据量大大增加，从而提升了对大量数据的处理能力。在本书中，大数据的含义主要是指海量乃至巨量数据，并且数据规模大到无法通过目前普及的计算机系统在用户可容忍时间内获取、存储、处理的数据。

对大数据的认识和利用需要通过相关工具对数据进行提取、分析和利用。所以在后面的内容中相继会对常见数据处理（数据工程）、相关处理工具、处理算法、经典案例进行描述，本书将主要针对大数据基本理论和工程实战进行叙述以帮助初学者快速入门。

1.1.2 大数据发展历程

2008年，电子商务快速发展，传统手段已无法满足其业务需求，大数据的理念和技术被雅虎、谷歌等大型互联网和电子商务公司尝试采用，并用来解决数据量大、数据种类多、数据流动速度快等问题。

2008年末，“大数据”被美国部分知名计算机科学研究人员所认可，计算社区联盟（Computing Community Consortium）发表了白皮书《大数据计算：在商务、科学和社会领域创建革命性突破》。它打破人们的思维局限，使人们的思维不再局限于用以处理数据的机

器，而且指出大数据重要的是其新用途和新见解，而不再是数据本身。

2010 年，随着 Web2.0 时代的到来，社交网络的飞速发展使人类进入自媒体时代，互联网数据进一步激增。智能移动终端的普及也使网络中的数据激增，大数据已如影随形，彻底融入人类社会的生产与生活之中。

2011 年，麦肯锡咨询公司的研究报告《大数据：创新、竞争和生产力的下一个新领域》分析了数字、数据和文档爆发式增长的状态，阐述了处理这些数据的潜在价值，分析了大数据相关的经济活动和业务价值链。

2012 年大数据发展空前繁荣。2012 年 3 月，白宫网站发布了《大数据研究和发展倡议》，2012 年 4 月，第一家大数据公司 Splunk 在纳斯达克上市，同年 7 月，联合国发布了一份大数据政务白皮书，阿里巴巴设立“首席数据官”管理职位。大数据已成为全球热门领域之一。

2014 年 4 月，世界经济论坛围绕“大数据的回报与风险”主题发布了名为《全球信息技术》的报告。报告认为，在未来几年中，针对各种信息通信技术的政策会显得更加重要。同年 5 月，美国白宫发布了研究报告《大数据：抓住机遇、守护价值》，鼓励用数据推动社会进步，同时也呼吁相应的框架、结构与研究支撑相关进展。

2015 年，国务院印发《促进大数据发展行动纲要》，明确提出要推动大数据的发展和应用，建立大数据下的经济新体制，大数据正式进入中国国家发展战略。

1.1.3　大数据的特征

大容量（Volume）：从前面的例子中可以体会到，人类社会活动产生的数据量已经超越 300 EB 级，并且这个数据还在逐年递增。

多种类（Variety）：多样化的数据往往都归类为结构化数据、半结构化数据和非结构化数据。与以往的结构化数据为主要数据的局面不同，现如今的数据多为非结构化数据，而这些包括网络日志、社交网络信息、地理位置信息等类型的数据都对数据处理提出了挑战。

速度快（Velocity）：面对如此大的数据体量，须用高效快速的处理方式对数据进行处理，提取有用信息，提高价值密度。

真实性（Veracity）：可靠的数据来源能够保障数据的真实性，而只有根据真实可靠的数据才能制定确实可靠的决策。

非结构性（Nonstructural）：在获得数据之前无法提前预知其结构，目前绝大多数数据都是非结构化数据，而不是纯粹的关系数据，传统的系统对这些数据无法完成处理。大量出现的各种数据本身是非结构化的或者说弱结构化的，如图片、视频数据等都是非结构化的，而网页等是半结构化数据。

时效性（Timeliness）：大数据的处理速度非常重要。数据规模越大，分析处理时间越长。如果设计一个处理固定大小数据量的数据系统，其处理速度可能会非常快，但这种方法并不适用于大数据的要求。在许多情况下，用户要求即时得到数据的分析结果。因此，还需要在实践、处理速度与规模之间折中考虑，寻求新的方法。

安全性（Security）：由于大数据高度依赖数据存储和共享，必须寻求更好的方法来消除各种隐患与漏洞，才能有效地管控风险。数据的隐私保护是大数据分析和处理的一个重

要问题，而隐私保护也是一个社会问题，一旦对个人数据的使用不当，尤其是涉及大批量的关联数据泄露，将会导致严重的后果。

1.1.4 大数据的基本认识

量变导致质变，大数据的价值正是体现在这里。一方面对大数据的分析和利用可以带来经济利益；另一方面，大数据已经开始融入现代生活的方方面面，从商业到医疗、政府、教育等各领域的决策都离不开对大数据的依赖。

大数据的核心功能是预测，通过将数学算法运用到海量的数据上来预测事情发生的可能性。而这些预测系统的关键在于它们是建立于海量数据的基础之上。系统的数据越多，算法就能更好地改善自己的性能。

大数据的精髓在于分析信息时的三个转变，这些转变是初学者需要注意的概念。第一个转变就是，在大数据时代可以分析更多的数据，有时候甚至可以处理和某个特别现象相关的所有数据，而不再依赖于随机采样；第二个改变就是，研究数据量之多使大家不再热衷于追求精确度；第三个就是由于上述两个转变，大数据侧重对相关关系的发掘而不再注重于因果关系。

1.2 人工智能

1.2.1 认识人工智能

人工智能(AI，Artificial Intelligence)最初是在 1956 年的 Dartmouth 学会上提出的。自此以后，众多的研究者们在发展理论的过程中，人工智能的概念也就逐渐扩散开来。从计算机应用系统的角度出发，人工智能是研究如何制造智能机器或智能系统来模拟人类智能活动的能力，以延伸人类智能的科学。

1.2.2 人工智能的现状与应用

随着 AI 技术的发展，现如今几乎各种技术的发展都涉及人工智能技术，人工智能技术已经渗透到许多领域，应用范围主要包括以下 9 个方面。

（1）符号计算

计算机最主要的用途之一就是科学计算，科学计算可分为两类：一类是纯数值的计算，通常是对函数、公式的求值；另一类是符号计算，也称代数运算，这种运算是对符号进行运算，并且符号可以代表整数、有理数、实数和复数，也可以代表多项式、函数、集合等。

（2）模式识别

模式识别就是通过计算机对数据样本进行特征提取，并用数学方法来研究模式的自动处理和判读。这里常说的模式是指文字、语音、生物特征、数字水印等环境与客体的结合体。

（3）机器翻译

机器翻译是通过计算机把一种自然语言转换成另一种自然语言的过程，用以完成这一

过程的软件系统叫作机器翻译系统。它是计算语言学（Computational Linguistics ）的一个分支，涉及计算机、认知科学、语言学、信息论等学科，是人工智能的终极目标之一，具有重要的科学研究价值。

（4）机器学习

机器学习是机器具有智能的重要标志，同时也是机器获取知识的根本途径。机器学习是一个难度较大的研究领域，它与认知科学、神经心理学、逻辑学等学科都有着密切的联系，并对人工智能的其他分支，如专家系统、自然语言理解、自动推理、智能机器人、计算机视觉、计算机听觉等方面，也会起到重要的推动作用。

（5）逻辑推理与定理证明

逻辑推理是人工智能研究中最持久的领域之一，其中特别重要的是要找到一些方法，只把注意力集中在一个大型数据库中的有关事实上，留意可信的证明，并在出现新信息时适时修正这些证明。

（6）自然语言处理

自然语言的处理是人工智能技术应用于实际领域的典型范例，经过多年艰苦努力，这一领域已获得了大量令人瞩目的成果。目前该领域的主要课题是：计算机系统如何以主题和对话情境为基础，注重大量的常识——世界知识和期望作用，生成和理解自然语言。

（7）分布式人工智能

分布式人工智能在 20 世纪 70 年代后期出现，是人工智能研究的一个重要分支。分布式人工智能系统一般由多个智能体（Agent）组成，每一个 Agent 又是一个半自治系统，Agent 之间以及 Agent 与环境之间进行并发活动，并通过交互来完成问题求解。

（8）计算机视觉

计算机视觉主要研究的是使计算机具有通过二维图像认知三维环境信息的能力，这种能力不仅包括对三维环境中物体形状、位置、姿态、运动等几何信息的感知，而且还包括对这些信息的描述、存储、识别与理解。

（9）专家系统

专家系统是目前人工智能中最活跃、最有成效的一个研究领域，它是一种具有特定领域内大量知识与经验的程序系统。人类专家因其丰富的知识，能够高效、快速地解决相应领域的众多问题，基于这一事实，给计算机程序学习并使其灵活运用这些知识，也就能解决人类专家所解决的问题，而且能帮助人类专家发现推理过程中出现的差错。

1.2.3　当人工智能遇上大数据

大数据的发展离不开人工智能，而任何智能的发展，都是一个长期学习的过程，且这一学习的过程离不开数据的支持。而近年来人工智能之所以能取得突飞猛进的进展，正是因为这些年来大数据的持续发展。而各类感应器和数据采集技术的发展，人类开始获取以往难以想象的海量数据，同时，也开始在相关领域拥有更深入、详尽的数据。而这些数据，都是训练相关领域“智能”的基础。

与以前的众多数据分析技术相比，人工智能技术立足于神经网络，并在此基础上发展

出多层神经网络，从而可以进行深度机器学习。与以往的传统算法相比，这一算法不像线性建模，需要假设数据之间的线性关系之类多余的假设前提，而是完全利用输入的数据自行模拟和构建相应的模型结构。这一算法特点决定了它是更为灵活的依据不同的输入来训练数据而拥有的自优化特性。

在计算机运算能力取得突破以前，这样的算法几乎没有实际应用的价值（因为运算量实在是太大了）。在十几年前，用神经网络算法计算一组并不海量的数据，辛苦等待几天都不一定会有结果。但如今，高速并行运算、海量数据、更优化的算法，打破了这一局面，并共同促成了人工智能发展的突破。

1.3 大数据与人工智能的机遇与挑战

1.3.1 大数据与人工智能面临的难题

人工智能已经发展了 60 多年，虽然在研究解释和模拟人类智能、智能行为及其规律这一总目标来说，已经取得了很大的进展。但从整体发展情况来看，人工智能发展过程曲折，而且还面临着不少难题，主要集中在以下几个方面。

（1）机器翻译

机器翻译遇到的最主要的问题是歧义性问题。构成句子的单词和歧义性问题一直是自然语言理解（NLU，Natural Language Understanding）中的一大难关。不同的使用场景，句子的含义也可能天差地别。所以要想消除歧义，正确解释句子语意必须结合具体语境。但现有的翻译方式通常都是将句子甚至词组作为理解单元，翻译结果往往忽视具体语境。另外，即使对原文语意理解到位，如何将其正确地表示成另一种语言，也是一个难题。现有的 NLU 系统无法随着时间增长而提高解读能力，学习深度不够。

（2）自动定理证明

自动定理证明需要机器拥有一套智能系统，不仅能够对现有条件进行合理演绎，并且能够做出正确判定。这一领域的代表性工作是 1965 年鲁宾孙提出的归结原理。归结原理虽然简单易行，但它所采用的方法是演绎，而这种形式上的演绎与人类自然演绎推理方法是截然不同的。基于归结原理的演绎推理要求把逻辑公式转化为子句集合，从而丧失了其固有的逻辑蕴含语义。

（3）模式识别

虽然使用计算机进行模式识别的研究与开发已取得大量成果，有的已成为产品投入实际应用，但是它的理论和方法与人的感官识别机制是全然不同的。一方面，人的识别手段、形象思维能力是任何最先进的计算机识别系统望尘莫及的；另一方面，在现实世界中，生活并不是一项结构严密的任务，一般的动物都能轻而易举地对付，但机器不会，这并不是说它们永远不会，而是说目前不会。技术的发展总是超乎人们的想象，要准确地预测人工智能的未来是不可能的。但是，从目前的一些前瞻性研究可以看出，未来人工智能可能会向以下几个方面发展：模糊处理、并行化、神经网络和机器情感。

1.3.2　大数据与人工智能的前景

人工智能作为一个整体的研究才刚刚开始，离其预定的目标还很遥远，但人工智能在某些方面将会有大的突破。

（1）自动推理是人工智能最经典的研究分支，其基本理论是人工智能其他分支的共同基础。一直以来，自动推理都是人工智能研究的最热门内容之一，其中知识系统的动态演化特征及可行性推理的研究是最新的热点，很有可能取得大的突破。

（2）机器学习的研究取得长足的发展。许多新的学习方法相继问世并获得了成功的应用，如增强学习（Reinforcement Learning）算法等。也应看到，现有的方法在处理在线学习方面尚不够有效，寻求一种新的方法以解决移动机器人、自主 agent、智能信息存取等研究中的在线学习问题是研究人员共同关心的问题，相信不久会在这些方面取得突破。

（3）自然语言处理是 AI 技术应用于实际领域的典型范例，经过 AI 研究人员的艰苦努力，这一领域已获得了大量令人瞩目的理论与应用成果。许多产品已经进入了众多领域。智能信息检索技术在 Internet 技术的影响下，近年来迅猛发展，已经成为 AI 的一个独立研究分支。由于信息获取与精化技术已成为当代计算机科学与技术研究中迫切需要研究的课题，将 AI 技术应用于这一领域的研究是人工智能走向应用的契机与突破口。从近年的人工智能发展来看，这方面的研究已取得了可喜的进展。

人工智能一直处于计算机技术的前沿，其研究的理论和发现在很大程度上将决定计算机技术的发展方向。如今，已经有很多人工智能的研究成果进入人们的日常生活。未来，人工智能技术的发展将会给人们的生活、工作和教育等带来更大的影响。