★ ★ ★
★ “十三五” ★
国家重点出版物出版规划项目
大数据丛书

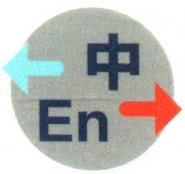
Broadview®
www.broadview.com.cn


深度计算

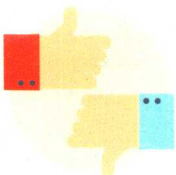

知识图谱


大数据系统

主题模型



机器翻译


情感分析与
意见挖掘


智能问答与
对话系统

大数据智能

数据驱动的自然语言处理技术

刘知远　崔安颀　等　编著

个性化
推荐系统

机器写作


社交商业
数据挖掘


智能医疗

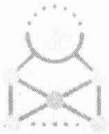

智能司法

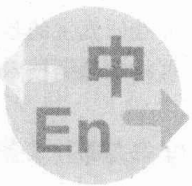
智能金融


计算社会学

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

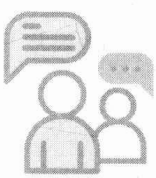
“十三五”
国家重点出版物出版规划项目
大数据丛书

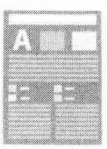



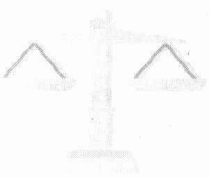
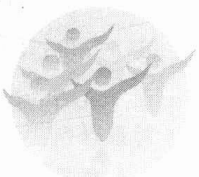
大数据智能

数据驱动的自然语言处理技术




刘知远　崔安颀　等　编著



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

内 容 简 介

本书是介绍大数据智能分析技术的科普书籍，旨在让更多人了解和学习互联网时代的自然语言处理技术，让大数据智能技术更好地为我们服务。

全书包括大数据智能基础、技术和应用三部分，共 14 章。基础部分有 3 章：第 1 章以深度学习为例介绍大数据智能的计算框架；第 2 章以知识图谱为例介绍大数据智能的知识库；第 3 章介绍大数据的计算处理系统。技术部分有 6 章，分别介绍主题模型、机器翻译、情感分析与意见挖掘、智能问答与对话系统、个性化推荐系统、机器写作。应用部分有 5 章，分别介绍社交商业数据挖掘、智慧医疗、智慧司法、智慧金融、计算社会学。本书后记部分为读者追踪大数据智能的最新学术资料提供了建议。

本书适合作为高等院校计算机相关专业研究生的学习参考资料，也适合计算机技术爱好者，特别是希望对大数据技术有所了解，想要将大数据技术应用于本职工作的所有读者阅读。

图书在版编目（CIP）数据

大数据智能：数据驱动的自然语言处理技术 / 刘知远等编著. —北京：电子工业出版社，2020.1
（大数据丛书）
ISBN 978-7-121-37538-5

Ⅰ. ①大… Ⅱ. ①刘… Ⅲ. ①数据处理②人工智能③自然语言处理 Ⅳ. ①TP274②TP18③TP391

中国版本图书馆 CIP 数据核字（2019）第 214390 号

责任编辑：郑柳洁　　特约编辑：顾慧芳
印　　刷：三河市龙林印务有限公司
装　　订：三河市龙林印务有限公司
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：23　　字数：441.6 千字
版　　次：2020 年 1 月第 1 版
印　　次：2020 年 4 月第 2 次印刷
定　　价：89.00 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：010-51260888-819，faq@phei.com.cn。

作者简介

赵　鑫

中国人民大学信息学院副教授、博士生导师。主要研究方向为数据挖掘和自然语言处理。2014年获得北京大学博士学位，在数据挖掘及其相关领域的著名国际期刊和会议上发表相关论文70余篇，曾获得CIKM 2017最佳短文提名及AIRS 2017最佳论文奖，Google Scholar统计引用2,800余次。曾获微软亚洲学者、北京大学优秀博士论文奖、中国人民大学杰出学者等荣誉称号，入选第二届CCF青年人才发展计划。长期担任国内外著名期刊和会议的评审。

苏劲松

厦门大学信息学院副教授、博士生导师。主要研究方向为自然语言处理和机器翻译。2011年获得中国科学院计算技术研究所博士学位，在人工智能、自然语言处理领域的著名国际期刊和会议上发表相关论文60余篇。担任CCF中文信息处理青年工作委员会常务委员，中文信息学会青年工作委员会委员，福建省人工智能学会理事，自然语言处理国际会议 NLPCC2018、EMNLP2019领域主席。

张永锋

罗格斯大学助理教授，主要研究方向为信息检索、推荐系统、机器学习及互联网经济。本科及博士毕业于清华大学计算机科学与技术系，曾获清华大学优秀博士毕业生及博士论文奖，中国人工智能协会优秀博士论文奖，研究论文发表于SIGIR、WWW、IJCAI、AAAI、CIKM、WSDM、RecSys等领域内主要会议，并担任会议审稿人。

严　睿

北京大学助理教授、研究员、博士生导师，曾任百度公司资深研究员，华中师范大学与中央财经大学客座教授与校外导师。主持研发了多个开放领域对话系统和服务类对话系统，发表高水平研究论文100余篇，担任多个学术会议（KDD、IJCAI、SIGIR、ACL、WWW、AAAI、CIKM、EMNLP等）的（资深）程序委员会委员及审稿人。受邀在EMNLP、WWW、SIGIR、IJCAI、AAAI等多个顶级国际会议上做针对人机对话系统与自然语言处理的讲习班指导报告或邀请报告。

汤步洲

哈尔滨工业大学（深圳）计算机科学与技术学院副教授、博士生导师。主要研究方向为自然语言处理、知识图谱、医学信息处理、医疗支持决策。2011年获得哈尔滨工业大学博士学位，毕业后先后赴美国范德堡大学和德州大学休斯敦医学科学中心以博士后研究员身份从事研究工作。在人工智能、医学信息学领域著名国际期刊和会议上发表相关论文80余篇，Google Scholar统计引用1,300余次。多次在相关领域国际公开评测中获得第1名。担任OHDSI China Working Group NLP方向负责人，中国中文信息学会医疗健康与生物信息处理专委会秘书长，中国中文信息学会青年工作委员会执委。

涂存超

清华大学计算机系博士后。主要研究方向为自然语言处理和法律智能。2018年获得清华大学博士学位。在人工智能及自然语言处理著名国际期刊和会议上发表相关论文10余篇。获得清华大学优秀博士毕业生、清华大学优秀博士论文奖、北京市优秀博士毕业生等荣誉，入选“博士后创新人才支持计划”。

丁　效

哈尔滨工业大学助理研究员、硕士生导师。主要研究方向为人工智能、自然语言处理、社会计算和事理图谱。2016年获得哈尔滨工业大学博士学位，已在AAAI、IJCAI、ACL、EMNLP、NAACL、COLING等人工智能领域的著名国际期刊和会议上发表相关论文20余篇。承担国家自然科学基金青年项目等省部级以上项目四项，参与国家重大科技基础设施建设项目、“新一代人工智能”重大项目、国家自然科学基金重点项目等多个科研项目。荣获全国青年人工智能创新创业大会三等奖、第五届全国青年计算语言学研讨会优秀论文奖等荣誉。担任中国中文信息学会社会媒体处理专委会秘书、常务委员，语言与知识计算专委会委员，中国中文信息学会青年工作委员会委员。

本书编委会

张开旭 | 腾讯

刘知远 | 清华大学

韩文弢 | 清华大学

赵　鑫 | 中国人民大学

苏劲松 | 厦门大学

崔安颀 | 薄言RSVP.ai

张永锋 | 罗格斯大学

严　睿 | 北京大学

汤步洲 | 哈尔滨工业大学（深圳）

涂存超 | 幂律智能

丁　效 | 哈尔滨工业大学

前言

大数据时代与人工智能

在进入 21 世纪前，很多人预测 21 世纪将会是怎样的世纪。有人说 21 世纪将是生命科学的时代，也有人说 21 世纪将是知识经济的时代，不一而足。随着互联网的高速发展，大量的事实强有力地告诉我们，21 世纪必将是大数据的时代，是智能信息处理的黄金时代。

美国奥巴马政府于 2012 年发布大数据研发倡议以来，关于大数据的研究与思考在全球蔚然成风，已经有很多专著面世，既有侧重趋势分析的，如舍恩伯格和库克耶的《大数据时代》（盛杨燕和周涛教授译）、涂子沛的《大数据》和《数据之巅》，也有偏重技术讲解的，如莱斯科夫等人的《大数据》（王斌教授译）、张俊林的《大数据日知录》、杨巨龙的《大数据技术全解》，等等。相信随着大数据革命的不断深入推进，会有更多的专著出版。

前人已对大数据的内涵进行过很多探讨与总结，其中比较著名的是所谓的“3V”定义：大容量（volume）、高速度（velocity）和多形态（variety）。3V 的概念于 2001 年由麦塔集团（Meta Group）分析师道格·莱尼（Doug Laney）提出，后来被高德纳咨询公司（Gartner Group）正式用来描述大数据。此外，还有很多研究者提出更多的“V”来描述大数据，如真实性（veracity），等等。既然有如此众多的“珠玉”在前，我们推出本书，当然希望讲一点不同的东西，这点不同的东西就是智能。

人工智能一直是研究者们非常感兴趣的话题，并且由于众多科幻电影和小说作品的影响而广为人知。1946 年，第一台电子计算机问世之后不久，英国数学家艾伦·麦席森·图灵就发表了一篇名为《计算机器与智能》（*Computing Machinery and Intelligence*）的重要论文，探讨了创造具有智能的机器的可能性，并提出了著名的“图灵测试”，即如果一台机器与人类进行对话，能够不被分辨出其机器的身份，就可以认为这台机器具有了智能。自 1956 年在美国达特茅斯举行的研讨会上正式提出“人工智能”的研究提案以来，人们开始了长达半个多世纪的曲折探索。

且不去纠结“什么是智能”这样哲学层面的命题［有兴趣的读者可以参阅罗素和诺维格的《人工智能——一种现代方法》（*Artificial Intelligence*: *A Modern Approach*），以及杰夫·霍金斯的《智能时代》（*On Intelligence*）］，我们先来谈谈人工智能与大数据的关系。要回答这个问题，我们先来看一个人是如何获得智能的。一个呱呱坠地、只会哭泣的婴儿，长成思维健全的成人，至少要经历十几年与周围世界交互和学习的过程。从降临到这个世界的那一刻起，婴儿无时无刻不在通过眼睛、耳朵、鼻子、皮肤接收着这个世界的数据信息：图像、声音、味道、触感，等等。你有没有发现，这些数据无论从规模、速度还是形态来看，无疑是典型的大数据。可以说，人类习得语言、思维等智能的过程，就是利用大数据学习的过程。智能不是无源之水，它并不是凭空从人脑中生长出来的。同样，人工智能希望让机器拥有智能，也需要以大数据作为学习的素材。可以说，大数据将是实现人工智能的重要支撑，而人工智能是大数据研究的重要目标之一。

但是，在人工智能研究早期，人们并不是这样认为的。早在 1957 年，由于人工智能系统在简单实例上的优越性能，研究者们曾信心满怀地认为，计算机将在 10 年内成为国际象棋冠军，而通过简单的句法规则变换和单词替换就可以实现机器翻译。事实证明：人们远远低估了人类智能的复杂性。即使在国际象棋这样规则和目标极为简单清晰的任务上，直到 40 年后的 1997 年，由 IBM 推出的深蓝超级计算机才宣告打败人类世界顶级国际象棋大师卡斯帕罗夫。而在机器翻译这样更加复杂的任务上（人们甚至在优质翻译的标准上都无法达成共识，更无法清晰地告诉机器），计算机至今还无法与人类翻译的水平相提并论。当时的问题在于，人们低估了智能的深度和复杂度。智能是分不同层次的。对于简单的智能任务（如对有限句式的翻译等），我们简单制定几条规则就能完成。但是对于语言理解、逻辑推理等高级智能，简单方法就显得力不从心。

生物界中，从简单的单细胞生物进化到人类的过程，也是智能不断进化的过程。最简单的单细胞生物草履虫，虽然没有神经系统，却已经能够根据外界信号和刺激进行反

应，实现趋利避害——我们可以将其视作最简单的智能。而俄国高级神经活动生理学奠基人伊万·彼得罗维奇·巴甫洛夫的关于狗的条件反射实验，则向我们证明了相对更高级的智能水平：能根据铃声推断食物即将出现，也就是可以根据两种外界信号（铃声与食物）的关联关系实现简单的因果推理。人类智能则是智能的最高级形式，拥有语言理解、逻辑推理与想象等独特的能力。我们可以发现，低级智能只需小规模的简单数据或规则的支持，而高级智能则需要大规模的复杂数据的支持。

同样重要的，高级智能还需要独特计算架构的支持。很显然，人脑结构就与狗等动物有着本质的不同，因此，即使将一只狗像婴儿一样抚育，也不能指望它能完全学会和理解人类的语言，并像人一样思维。受到生物智能的启发，我们可以总结出如下图所示的基本结论：不同规模数据的处理，需要不同的计算框架，产生不同级别的智能。

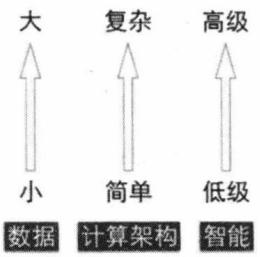


图　不同规模数据的处理，需要不同的计算架构，产生不同级别的智能

关于人工智能是否要完全照搬人类智能的工作原理，目前仍然争论不休。有人举例：虽然人们受到飞鸟的启发发明了飞机，但其飞行原理（空气动力学）与飞鸟有本质不同；同样，生物界都在用双脚或四腿行走、奔跑，人们却发明了轮子和汽车实现快速移动。然而不可否认，大自然无疑是我们最好的老师。人工智能固然不必完全复制人类智能，但是知己知彼，方能百战不殆。生物智能带来的启示已经在信息处理技术发展中得到了印证。谷歌研究员、美国工程院院士 Jeff Dean 曾对大数据做出过类似结论：“对处理数据规模 X 的合理设计可能在 $10X$ 或 $100X$ 规模下就会变得不合理。（Right design at X may be very wrong at 10X or 100X.）”也就是说，大数据处理也需要专门设计新颖的计算架构。而与人工智能密切相关的机器学习、自然语言处理、图像处理、语音处理等领域，近年来都在大规模数据的支持下取得了惊人的进展。我们可以确信，大数据是人工智能发展的必由之路。

大数据智能如何成真

虽然大数据是实现人工智能的重要支持，但如何实现大数据智能，却并非显而易见。近年来，计算机硬件、大数据处理技术和深度学习等领域取得了突破性进展，涌现出了一批在技术上和商业上影响巨大的智能应用，让人工智能发展道路日益清晰。触手可及的人类社会大数据、高性能的计算能力，以及合理的智能计算框架，为大数据智能的实现提供了有力的支持。

人类社会大数据触手可及。如前所述，这是大数据的时代，互联网的兴起、手机等便携设备的普及，让人类社会的行为数据越来越多地汇聚到网上。这让机器从这些大数据中自动学习成为可能。但是，大数据（如大气数据、地震数据等）并非现在才出现，只是在过去，我们限于计算能力和计算框架，难以从中萃取精华。因此，大数据智能的实现还依赖以下两个方面的发展。

一方面，计算能力突飞猛进。受到摩尔定律的支配，近半个世纪以来，计算机的计算和存储能力一直在以令人目眩的速度提高。摩尔定律最早由英特尔（Intel）创始人之一戈登·摩尔提出，其基本思想是：保持价格不变的情况下，集成电路上可容纳的元器件的数目大约每隔 18 到 24 个月就会增加一倍，性能也将随之提升一倍。也就是说，每一块钱能买到的计算机性能将每隔 18 到 24 个月提升一倍以上。虽然人们一直担心，随着微处理器器件尺寸逐渐变小，摩尔定律会受量子效应影响而失效，但至少从已有发展历程来看，随着多核、多机并行等框架的提出，计算机已经能够较好地提供大规模数据处理所需的计算能力了。

另一方面，计算框架返璞归真。近年来，深度学习在图像、语音和自然语言处理领域掀起了一场革命，在图像分类、语音识别等重要任务上取得了惊人的性能突破，在国际上催生了苹果 Siri 等语音助手的出现，在国内则涌现了科大讯飞、Face++等高科技公司。然而我们可能很难想象，深度学习的基础——人工神经网络技术，此前曾长期处于无人问津的境地。在深度学习兴起以前，人工神经网络常因存在可解释性差、学习稳定性差、难以找到最优解等问题而被诟病。然而，正是由于大规模数据和高性能计算能力的支持，以人工神经网络为代表的机器学习技术才得以在大数据时代焕发出勃勃生机。

人工智能的下一个里程碑

当下，以深度学习为代表的计算框架在很多具体任务上取得了重大的成果，甚至有媒体和公众已经开始因人工智能取代人类的可能性而恐慌。然而，理性地看，深度学习的处理能力和效率与人类大脑相比仍有巨大差距。因此，大数据智能并非孕育人工智能的终极之道。随着技术的进步和研究的深入，现有解决方案必然触及天花板，进入瓶颈期。

人脑拥有现有计算框架不可比拟的优势。例如，虽然人脑中的信号传输速度要远低于计算机中的信息传递速度，但是人脑在很多智能任务上的处理效率远高于计算机，例如在众多声音中快速识别出叫自己名字的声音，通过线条漫画认出名人，复杂数学问题的推导求解，快速阅读理解一篇文章，等等。可见，在计算速度受限的情况下，人脑一定拥有某种独特的计算框架，才能完成这些令人叹为观止的智能任务。

那么人工智能的下一个里程碑是什么呢？我们猜想，可能是神经科学及其相关学科。一直以来，神经科学都在探索各种观测大脑活动的工具和方法，并做出了大量的实证和建模工作。随着光控基因技术（optogenetics）和药理基因技术（pharmacogenetics）等新技术的发展，人们拥有了在时间和空间上更加精确地监测和控制大脑活动的能力，从而有望彻底发现人脑的神经机制。一旦人脑的神经机制被发现，有理由相信，人们可以迅速通过仿真等方式，在计算机中实现类似甚至更高效的计算框架，从而推动实现人工智能的最终目标。此外，量子计算、生物计算、新型芯片材料等领域的发展，都为我们展现出无限可能的未来。

当社会大数据、计算能力和计算框架三方面发展到一定阶段，融合产生了大数据智能。相信随着更大规模数据、更强计算能力和更合理计算框架的推出，人工智能也会不断向前发展。然而，正如前几年社会各界对物联网、云计算的追捧，最近社会上对大数据和人工智能概念的炒作愈演愈烈，产生了很多不切实际的幻想和泡沫。对于这个领域重新得到青睐，我们当然感到欣慰，但是，也不妨多一些谨慎和冷静。鉴古知今，回顾人工智能的曲折发展史（《人工智能——一种现代方法》一书中有详细介绍），我们看到，在过度的期望破灭之后，随之而来的就是严冬。在大数据智能万众瞩目的今天，我们不妨心中常存对于凛冬将至的警惕。

事物总是在不断自我否定中螺旋式前进的，人工智能的探求之路也是如此。我们相信大数据是获得智能的必由之路，但现在的做法不见得就一定正确。多年之后，我们也

许会用截然不同的办法处理大数据。然而这些都不重要，重要的是一颗执着的心和坚持不懈的信念。就像深度学习领域的巨人 Geoffrey Hinton、Yann LeCun 等，曾坐了十几年的冷板凳，研究成果屡屡被拒，到了 2019 年才荣膺计算机领域最高奖“图灵奖”。对真正的学者而言，研究领域是冷门还是热门也许不重要，反而会成为对从业者的试金石——只有在寒冬中坚持下来的种子，才能等到春天绽放。

关于本书

本书前身《大数据智能——互联网时代的机器学习和自然语言处理技术》出版于 2016 年，作为一本技术科普书，在社会上得到了一些正面的反响。于是，我们邀请更多作者加入，在原有的 8 章内容基础上新增了 6 章内容。此外，对原有章节内容进行了适当更新，使内容更加全面。

本书并不想在已经熊熊燃烧的大数据火堆上再添一把柴。本书希望从人工智能这个新的角度，总结大数据智能取得的成果、局限性及未来可能的发展前景。本书共分 14 章，从大数据智能基础、技术和应用三个方面展开介绍。

本书基础部分有 3 章。第 1 章以深度学习为例介绍大数据智能的计算框架；第 2 章以知识图谱为例介绍大数据智能的知识库；第 3 章介绍大数据的计算处理系统。在大数据智能的技术和应用部分，我们选择文本大数据作为主要场景进行介绍，主要原因在于，语言是人类智能的集中体现，语言理解也是人工智能的终极目标，图灵测试的设置是以语言作为媒介的。技术部分有 6 章，分别介绍主题模型、机器翻译、情感分析与意见挖掘、智能问答与对话系统、个性化推荐系统、机器写作等数据智能关键技术。应用部分有 5 章，分别介绍社交商业数据挖掘、智慧医疗、智慧司法、智能金融、计算社会学等典型应用场景。

大数据智能仍然是一个高速发展的领域。为了让读者能够了解这个领域的前沿进展，本书专门设置后记，为初学者追踪大数据智能的最新学术资料提供了建议。

大数据智能方向众多，每位学者术业有专攻，很难独力完成所有章节内容。因此，我们邀请了多位作者撰写他们所擅长方向的章节。他们都在相关领域开展了多年研究工作，发表过高水平的论文。

致谢

本书能够出版，无疑得到了很多人的支持和帮助。首先，感谢本书的几位合作者：丁效、韩文弢、苏劲松、汤步洲、涂存超、严睿、张开旭、张永锋、赵鑫。他们的热情、无私与认真，让我们相信本书能够真的为读者提供及时、有用的知识。还要感谢各位同事、同学和好友，在本书撰写过程中提供了很多最新研究资料和热情的帮助。

我们特别感谢电子工业出版社副总编辑兼博文视点公司总经理郭立老师的热情邀请和大力支持，以及本书策划兼特约编辑、清华大学计算机系 1964 届学长顾慧芳老师的不断激励和鼎力相助，让我们鼓起勇气接下这个选题，也能在我们拖延症反复发作时耐心地等待。在书稿的准备过程中，特别感谢本书责任编辑郑柳洁老师对书稿的悉心修改，对封面设计和每章内容都提供了大量中肯的建议，让本书焕然一新。

欢迎交流

当今世界，大数据智能是一个涉及面非常广泛、发展非常迅猛的领域，而且这个领域的研究成果将加速人类认识世界、探索宇宙，也将极大地影响人们日常生活的方方面面。因此，笔者想在从事学习和自然语言处理等基础技术和最新进展研究工作的同时撰写一本介绍这一领域的科普书籍，抛砖引玉，旨在为需要了解与学习大数据智能技术的朋友提供帮助，使更多有志之士加入大数据智能分析这一充满惊奇和魅力的领域中。

笔者尽量以开放的态度梳理每个方向的相关成果和进展，然而大数据智能日新月异，而我们所知有限，难免有挂一漏万之憾。如有重要进展或成果没有涉及，绝非作者故意为之，敬请大家批评指正。我们欢迎读者对本书做出任何反馈，无论是指出错误还是改进建议，请直接发邮件至 liuzy@tsinghua.edu.cn。我们会在书中改正所有发现的错误。

刘知远　崔安颀

2019 年 11 月于北京

反侵权盗版声明

目录