Broadview®
www.broadview.com.cn

清华大学数据科学研究院
Tsinghua Institute for Data Science
清华大学数据科学研究院
清华大数据产业联合会 联合力荐

智能并非从人脑或机器中凭空产生
大数据是习得智能的必由之路

互联网时代的机器学习和自然语言处理技术
MACHINE LEARNING & NATURAL LANGUAGE PROCESSING IN THE INTERNET AGE

大数据智能

BIG DATA INTELLIGENCE

刘知远 崔安颀 等著

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

互联网时代的机器学习和自然语言处理技术

MACHINE LEARNING & NATURAL LANGUAGE PROCESSING IN THE INTERNET AGE

大数据智能

BIG DATA INTELLIGENCE

刘知远 崔安颀 等著

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

内 容 简 介

本书是一本介绍大数据智能分析的科普书籍，旨在让更多的人了解和学习互联网时代的机器学习和自然语言处理技术，以期让大数据技术更好地为我们的生产和生活服务。

全书包括大数据智能基础和大数据智能应用两个部分，共 8 章。大数据智能基础部分有三章：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。大数据智能应用部分有 5 章：第 4 章介绍智能问答，第 5 章介绍主题模型，第 6 章介绍个性化推荐，第 7 章介绍情感分析与意见挖掘，第 8 章介绍面向社会媒体内容的分析与应用。最后在本书的后记部分为读者追踪大数据智能的最新学术材料提供了建议。

本书适合作为高等院校计算机相关专业的研究生学习参考资料，也适合电脑爱好者阅读。作者特别希望本书能够帮助所有愿意对大数据技术有所了解，以及想要将大数据技术应用于本职工作的读者。

图书在版编目（CIP）数据

大数据智能：互联网时代的机器学习和自然语言处理技术 / 刘知远等著. —北京：电子工业出版社，2016.1
ISBN 978-7-121-27648-4

Ⅰ. ①大… Ⅱ. ①刘… ②崔… Ⅲ. ①机器学习②自然语言处理 Ⅳ. ①TP181②TP391

中国版本图书馆 CIP 数据核字（2015）第 281768 号

统筹策划：顾慧芳
责任编辑：徐津平
特约编辑：顾慧芳
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
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：787×980　1/16　印张：14.75　字数：322 千字
版　　次：2016 年 1 月第 1 版
印　　次：2016 年 1 月第 1 次印刷
定　　价：49.00 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。

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。

服务热线：（010）88258888。

前　言

天才并不是自生自长在深林荒野里的怪物，是由可以使天才生长的民众产生、长育出来的，所以没有这种民众，就没有天才。

——鲁迅

千淘万漉虽辛苦，吹尽狂沙始到金。

——[唐]刘禹锡

大数据时代与人工智能

在进入21世纪前后，很多人预测这将会是怎样的世纪。有人说这将是生命科学的时代，也有人说这将是知识经济的时代，不一而足。现在15年过去了，随着互联网的高速发展，大量的事实强有力地告诉我们，这必将是大数据的时代，是智能信息处理的黄金时代。

自2012年美国奥巴马政府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的概念最早于2001年由麦塔集团（Meta Group）分析员道格·莱尼（Doug Laney）提出，后来被高德纳咨询公司（Gartner）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，例如真实性（Veracity），等等。既然有如此众多珠玉在前，我们推出这本书，当然希望讲一些不同的东西，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或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，英国著名学者图灵就发表了一篇重要论文（题名《计算机器与智能》*Computing Machinery and Intelligence*），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，那么就可以认为这台机器具有了智能。自 1956 年达特茅斯研讨会正式提出了“人工智能”的研究提案以来，人们开始了至今长达半个多世纪的曲折探索。

我们且不去纠结“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*），而是先来谈谈人工智能与大数据有什么关系？要回答这个问题，我们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，最后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信息：图像、声音、味觉、触觉，等等。你有没有发现，无论从数据的规模、速度还是形态来看，这些信息无疑是典型的大数据。因此，人类习得语言、思维等智能的过程，就是从大数据学习的过程。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这样认为。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，10 年内计算机将能成为国际象棋冠军，而通过简单的句法规则变换和词典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冠军卡斯帕罗夫。而在机器翻译这样更加复杂的任务（人们甚至连优质翻译的标准都无法达成共识，并清晰地告诉机器）上，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，人们远远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），我们当然可以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，简单方法就束手无策了。

生物界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。而巴甫洛夫关于的狗的条件反射实验，则向我们证明了相对更高级的智能水平，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，也就是能根据铃声推断食物即将出现。人类智能则是智能的最高级形式，拥有了语言理解、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理解人类的语言，并像人一样思维。受到生物智能的启示，我们可以总结出如图 0.1 所示的基本结论，不同大小数据的处理，需要不同的计算框架，带来不同级别的智能。

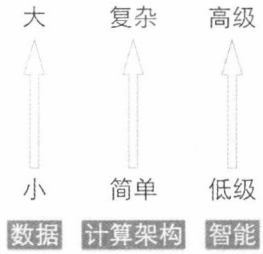

图 0.1 不同规模的数据需要不同的计算架构，产生不同级别的智能

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）却与飞鸟有本质不同；同样，生物界都在用双脚或四腿奔跑行走，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 曾对大数据作过类似结论：“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进展。我们可以确信地说，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。近年来随着计算机硬件、大数据处理技术和深度学习等领域取得了突破性进展，涌现出一批在技术上和商业上影响巨大的智能应用，这让人工智能发展道路日益清晰起来。

大数据的价值并非水落石出这样显而易见。我们认为，近年来人工智能的突破性进展，主要是在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的支持下，方能披沙拣金实现大数据智能。

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，让人类社会行为数据越来越多地汇聚到网上，触手可及。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。

（1）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，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变小，摩尔定律会受到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、多机并行等新框架的提出，计算机已经能够较好地支持大规模数据处理所需的计算能力。

（2）计算架构返璞归真。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，人工神经网络常被人诟病存在可解释性差、学习稳定性差、难以找到最优解等问题。然而，正是由于大规模数据和高性能计算能力的支持，才让以人工神经网络为代表的机器学习技术在大数据时代焕发出蓬勃生机。

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巨大成绩。甚至有媒体和公众已经开始恐慌人工智能取代人类的可能性。然而，从理性来看，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，仅依靠大数据智能，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无法逾越的天花板，进入瓶颈期。

人脑拥有现有计算框架不可比拟的优势。例如，虽然人脑中信号传输速度要远低于计算机中的信号速度，但是人脑在很多智能任务上的处理效率则远高于计算机，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。可以想象，在计算速度受限的条件下，人脑一定拥有某种独特的计算框架，才能完成这些叹为观止的智能任务，可谓大自然的鬼斧神工。

那么人工智能的下一个里程碑是什么呢？我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，并作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控制和监测大脑活动的能力，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。

当下，社会大数据、计算能力和计算框架三方面的发展融合产生出了大数据智能。我们相信更大规模数据、更强计算能力和更合理计算框架的推出，会不断推动人工智能向前发展。然而，正如前几年社会各界对物联网、云计算的追捧，最近社会对大数据和人工智能概念的炒作愈演愈烈，产生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》中有详细介绍）我们看到，在过度的期望破灭之后，随之而来的往往就是严冬。现在大数据智能万众瞩目，我们不妨心中默念凛冬将至。

事物总是在不断自我否定中螺旋前进的，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，重要的在于一颗无论冷门热门都执着的心、坚持不懈的信念。就像现在深度学习领域的巨人 Geoffrey Hinton、Yann LeCun 等学者，在这之前坐了十几年的冷板凳，研究成果屡屡被拒。对于真正的学者，研究领域冷门热门也许都不重要，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

这本书并不想在已经火得发紫的大数据火堆上再添一把柴。这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，它的局限性以及未来可能的发展前景。

本书从大数据智能基础和应用两个方面展开介绍。

基础部分有三章：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。应用部分有五章：第 4 章介绍智能问答，第 5 章介绍主题模型，第 6 章介绍个性化与推荐，第 7 章介绍情感分析与意见挖掘，第 8 章介绍面向社会媒体内容的分析应用。这基本涵盖了文本大数据智能处理的主要应用场景。以后如有机会再版，还计划纳入文档摘要、计算广告学等主题。

大数据智能仍然是个高速发展的领域。可以想象这本书出版的时候，很多内容已显陈旧。为了让读者能够跟踪这个领域的最前沿进展，本书专门设置后记，为初学者追踪大数据智能的最新学术材料提供建议。

个人学识有限，深怕在自己不擅长的领域说出外行话甚至错误连篇。因此，我邀请熟识的同学朋友撰写他们所擅长的章节。除了前言、第 2 章、第 8 章和后记由我操刀外，我请同门师弟张开旭博士撰写第 1 章，清华大学计算机系统方向博士韩文弢撰写第 3 章，清华大学信息检索方向博士崔安颀撰写第 4、7 章，北京大学自然语言处理方向博士、现中国人民大学信息学院教师赵鑫撰写第 5 章，清华大学个性化推荐方向博士生张永锋同学撰写第 6 章。他们都在相关领域开展了多年研究工作，发表过高水平论文。最后，我对全书做了统稿和校对，北京邮电大学毕业的林颖同学在我们实验室实习期间帮助我做了大量的书稿整理工作。

致谢

本书能够出版，无疑得到了很多人的支持和帮助。

首先，感谢这本书的几位合作者张永锋、崔安颀、张开旭、赵鑫和韩文弢，他们的热情、无私与认真，让我相信这本书能够真的为读者提供及时有用的知识。

其次，感谢我的导师和领导清华计算机系的孙茂松教授，是他将我带入了这个精彩纷呈的研究领域，也是他为我提供了宽松的写作环境，能够让这本书顺利问世。

我还要感谢刘洋（清华大学）、付杰（新加坡国立大学）、来思惟（中科院自动化所）

等同事、同学和好友，在本书撰写过程中提供了很多最新进展和热情帮助。特别感谢林颖同学所做的书稿整理和封面设计工作。

最后，我要特别感谢电子工业出版社副总编辑兼计算机分社社长郭立老师的热情邀请和大力支持，以及本书编辑、清华计算机系学长顾慧芳老师的不断激励和鼎力相助，让我鼓起勇气敢于接下这个选题，也能在我拖延症反复发作时耐心地等待，经过了两年多时间的酝酿、收集资料、研究分析以及整理撰写，终于变成了你手中的这本书。

欢迎交流

当今世界，大数据智能是一个涉及非常广泛、而且发展非常迅猛的领域，这个领域的研究成果将帮助人类加速认识世界、探索宇宙，也将极大地影响到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，作为抛砖引玉，旨在为需要了解与学习大数据智能技术的朋友提供帮助，甚至加入到大数据智能分析这一充满惊奇和魅力的领域中来。

当然，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被介绍到，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的任何反馈，无论是指出错误还是改进建议，请直接发邮件给我：liuzy@tsinghua.edu.cn。我们会专门开辟网站维护勘误清单，如果本书有机会再出下一版的话，也会尽量改正所有发现的错误。

刘知远博士
清华大学计算机科学与技术系　助理研究员
2015 年 8 月于北京清华园

目 录

第 1 章 深度学习——机器大脑的结构 1

1.1 概述 3
1.1.1 可以做酸奶的面包机——通用机器的概念 3
1.1.2 连接主义 5
1.1.3 用机器设计机器 6
1.1.4 深度网络 6
1.1.5 深度学习的用武之地 7
1.2 从人脑神经元到人工神经元 8
1.2.1 生物神经元中的计算灵感 8
1.2.2 激活函数 9
1.3 参数学习 10
1.3.1 模型的评价 11
1.3.2 有监督学习 11
1.3.3 梯度下降法 12
1.4 多层前馈网络 13
1.4.1 多层前馈网络 14
1.4.2 后向传播算法计算梯度 16
1.5 逐层预训练 17
1.6 深度学习是终极神器吗 19
1.6.1 深度学习带来了什么 19
1.6.2 深度学习尚未做到什么 20
1.7 内容回顾与推荐阅读 21

1.8 参考文献 21

第 2 章 知识图谱——机器大脑中的知识库 23

2.1 什么是知识图谱 25
2.2 知识图谱的构建 27
2.2.1 大规模知识库 27
2.2.2 互联网链接数据 28
2.2.3 互联网网页文本数据 29
2.2.4 多数据源的知识融合 29
2.3 知识图谱的典型应用 30
2.3.1 查询理解（Query Understanding） 30
2.3.2 自动问答（Question Answering） 32
2.3.3 文档表示（Document Representation） 33
2.4 知识图谱的主要技术 34
2.4.1 实体链指（Entity Linking） 34
2.4.2 关系抽取（Relation Extraction） 35
2.4.3 知识推理（Knowledge Reasoning） 37
2.4.4 知识表示（Knowledge Representation） 38
2.5 前景与挑战 39
2.6 内容回顾与推荐阅读 40
2.7 参考文献 41

第 3 章 大数据系统——大数据背后的支撑技术 43

3.1 概述 45
3.2 高性能计算技术 46
3.2.1 超级计算机的组成 47
3.2.2 并行计算的系统支持 48
3.3 虚拟化和云计算技术 52
3.3.1 虚拟化技术 52

3.3.2 云计算服务 54
3.4 基于分布式计算的大数据系统 55
3.4.1 Hadoop 生态系统 55
3.4.2 Spark 61
3.4.3 典型的大数据基础架构 63
3.5 大规模图计算 63
3.5.1 分布式图计算框架 64
3.5.2 高效的单机图计算框架 65
3.6 NoSQL 66
3.6.1 MongoDB 简介 67
3.7 内容回顾与推荐阅读 69
3.8 参考文献 70

第 4 章 智能问答——智能助手是如何炼成的 71

4.1 概述 73
4.2 问答系统的主要组成 77
4.3 文本问答系统 78
4.3.1 问题理解 78
4.3.2 知识检索 81
4.3.3 答案生成 83
4.4 社区问答系统 84
4.4.1 社区问答系统的结构 85
4.4.2 相似问题检索 86
4.4.3 答案过滤 86
4.5 多媒体问答系统 87
4.6 大型问答系统案例：IBM 沃森问答系统 89
4.6.1 沃森的总体结构 89
4.6.2 问题解析 90
4.6.3 知识储备 90

4.6.4 检索和候选答案生成 91
4.6.5 可信答案确定 92
4.7 内容回顾与推荐阅读 93
4.8 参考文献 94

第 5 章 主题模型——机器的智能摘要利器 97

5.1 概述 99
5.2 主题模型出现的背景 100
5.3 第一个主题模型潜在语义分析 102
5.4 第一个正式的概率主题模型 104
5.5 第一个正式的贝叶斯主题模型 105
5.6 LDA 的概要介绍 106
5.6.1 LDA 的延伸理解——主题模型广义理解 109
5.6.2 模型求解 111
5.6.3 模型评估 112
5.6.4 模型选择：主题数目的确定 113
5.7 主题模型的变形与应用 114
5.7.1 基于 LDA 的模型变种 114
5.7.2 基于 LDA 的典型应用 115
5.7.3 一个基于主题模型的新浪名人话题排行榜应用 118
5.8 内容回顾与推荐阅读 122
5.9 参考文献 123

第 6 章 个性化推荐系统——如何了解电脑背后的 TA 129

6.1 概述 131
6.1.1 推荐系统的发展历史 132
6.1.2 推荐无处不在 133
6.1.3 从千人一面到千人千面 133
6.2 个性化推荐的基本问题 134
6.2.1 推荐系统的输入 135

6.2.2 推荐系统的输出 137
6.2.3 个性化推荐的形式化 137
6.2.4 推荐系统的三大核心问题 138
6.3 典型推荐算法浅析 139
6.3.1 推荐算法的分类 139
6.3.2 典型推荐算法介绍 140
6.3.3 基于矩阵分解的打分预测 146
6.3.4 推荐的可解释性 151
6.3.5 推荐算法的评价 153
6.3.6 我们走了多远 156
6.4 参考文献 160

第 7 章 情感分析与意见挖掘——计算机如何了解人类情感 165

7.1 概述 167
7.2 情感分析的主要研究问题 172
7.3 情感分析的主要方法 175
7.3.1 构成情感和观点的基本元素 175
7.3.2 情感极性与情感词典 177
7.3.3 属性－观点对 182
7.3.4 情感分析 184
7.4 主要的情感词典资源 188
7.5 内容回顾与推荐阅读 189
7.6 参考文献 190

第 8 章 面向社会媒体大数据的语言使用分析及应用 195

8.1 概述 197
8.2 面向社会媒体的自然语言使用分析 197
8.2.1 词汇的时空传播与演化 198
8.2.2 语言使用与个体差异 200

8.2.3 语言使用与社会地位 202
8.2.4 语言使用与群体分析 203
8.3 面向社会媒体的自然语言分析应用 206
8.3.1 社会预测 206
8.3.2 霸凌现象定量分析 207
8.4 未来研究的挑战与展望 208
8.5 参考文献 209

后 记 214

国际学术组织、学术会议与学术论文 214
国内学术组织、学术会议与学术论文 216
如何快速了解某个领域的研究进展 217

第 1 章

深度学习——机器大脑的结构

为了实现高层抽象表征的复杂能力，我们需要深层结构。

——[美]尤舒·本吉奥（Yoshua Bengio）