

COGNITIVE COMPUTING AND BIG DATA ANALYTICS

WILEY

认知计算与 大数据分析

【美】Judith S. Hurwitz、Marcia Kaufman、Adrian Bowles 著
张鸿涛 译



将大数据分析 with 认知计算相结合
开启人类认知计算新时代

解读人工智能、机器学习、自然语言处理和大数据分析



中国工信出版集团



人民邮电出版社
POSTS & TELECOM PRESS

WILEY

COGNITIVE
COMPUTING
AND BIG DATA ANALYTICS

认知计算与
大数据分析

【美】Judith S. Hurwitz、Marcia Kaufman、Adrian Bowles◎著
张鸿涛◎译

人民邮电出版社

北京

图书在版编目 (C I P) 数据

认知计算与大数据分析 / (美) 朱迪斯·S. 赫尔维茨
(Judith S. Hurwitz), (美) 玛西亚·A. 考夫曼
(Marcia A. Kaufman), (美) 阿德里安·鲍尔斯
(Adrian Bowles) 著; 张鸿涛译. -- 北京: 人民邮电
出版社, 2017. 1

ISBN 978-7-115-43693-1

I. ①认… II. ①朱… ②玛… ③阿… ④张… III.
①数据处理 IV. ①TP274

中国版本图书馆CIP数据核字(2016)第232858号

版 权 声 明

Judith S. Hurwitz, Marcia Kaufman, Adrian Bowles

Cognitive Computing and Big Data Analytics

Copyright (c)2015 by John Wiley & Sons Ltd..All rights reserved.

This translation published under license.

Authorized translation from the English language edition published by Wiley Publishing, Inc..

Copies of this book sold without a Wiley sticker on the cover are unauthorized and illegal.

本书中文简体字版由 John Wiley & Sons Ltd 公司授权人民邮电出版社出版, 专有版权属于人民邮电出版社。本书封底贴有 Wiley 防伪标签, 无标签者不得销售。

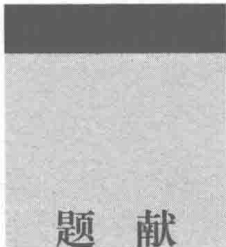
-
- ◆ 著 [美] Judith S. Hurwitz Marcia Kaufman
Adrian Bowles
译 张鸿涛
责任编辑 李 强
责任印制 彭志环
- ◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路 11 号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 <http://www.ptpress.com.cn>
大厂聚鑫印刷有限责任公司印刷
- ◆ 开本: 700×1000 1/16
印张: 17.5 2017 年 1 月第 1 版
字数: 275 千字 2017 年 1 月河北第 1 次印刷
著作权合同登记号 图字: 01-2015- 5767 号
-

定价: 69.00 元

读者服务热线: (010) 81055488 印装质量热线: (010) 81055316

反盗版热线: (010) 81055315

广告经营许可证: 京东工商广字第 8052 号



题 献

我们想要将这本书归功于协作的力量。感谢在 Hurwitz & Associates 团队里其余的人的指导和帮助：Dan Kirsch、Vikki Kolbe 和 Tricia Gilligan。

——作者

致我的丈夫，以及我的两个孩子——Sara 和 David。我也将这本书献给我的父母 Elaine 和 David Shapiro。

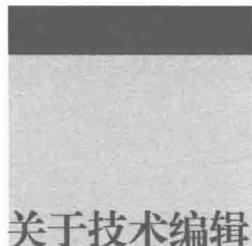
——Judith Hurwitz

致我的丈夫 Matt，以及我的孩子们——Sara 和 Emily，感谢他们对我整个写作过程的支持。

——Marcia Kaufman

致 Jeanne、Andrew、Chris 和 James，他们无尽的爱和支持使得我能够长时间埋头写作。

——Adrian Bowles

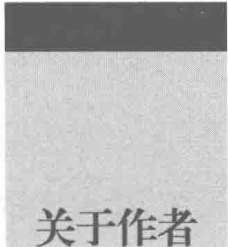


关于技术编辑

Al Nugent 是就职于 Palladian Partners 有限责任公司的合伙人。他是一位经验丰富的技术领导者，同时也是一位从业 30 多年的资深人士。在 Palladian Partners，他领导了组织的技术评估和战略实施。他曾担任 CA 技术公司的执行副总裁、首席技术官、高级副总裁，以及企业系统与管理事业部的总经理。之前，他是 Novell 公司的高级副总裁和 CTO，并且曾在 BellSouth 和 Xerox 担任 CTO。他是 Telogis 和 Adaptive Computing 董事会的独立成员，并且是几个早中期技术和医疗创业公司的顾问。他也是《写给大家看的大数据》(Big Data For Dummies) 这本书的合著者 (John Wiley & Sons, 2013)。

James Kobiulus 是 IBM 的一位大数据专员，并且是产品市场和大数据分析解决方案的高级项目主管。他是一位行业资深人士，一位受欢迎的演讲家、社交媒体达人，在大数据、Hadoop、企业数据仓库、高级分析、商业智能、数据管理、下一最优行动技术领域，他是思想领导者。

Michael D. Kowolenko 博士现在是北卡罗来纳州普尔管理学院创新管理研究中心 (CIMS) 从事工业研究的研究员。他的研究关注于技术与商业决策制定的接口。在加入 CIMS 之前，他是惠氏生物科技技术执行与产品供应 (TO&PS) 的高级副总裁，为发展中的整合和跨功能全球商业决策提供了战略和执行上的领导意见。



关于作者

Judith S. Hurwitz 是 Hurwitz & Associates 有限责任公司的董事长兼 CEO。这是一家研究和咨询公司，关注大数据、认知计算、云计算、服务管理、软件开发以及安全与管理等新兴技术。她是一位技术战略家、思想领导者和作家。作为预测技术创新和采用的先行者，多年来，她曾担任许多产业领导者值得信赖的顾问。她帮助这些公司向新的商业模式转型，专注于新兴平台的商业价值。她是 CycleBridge 的创始人，这是一家生命科学软件咨询公司。同时她也是 Hurwitz Group 的创始人，这是一家研究和咨询公司。她在很多家公司工作过，包括 Apollo Computer 和 John Hancock。她在企业及分布式软件领域有大量著作。2011 年，她创作出版了《聪明还是幸运？技术领导者是怎样将机遇转化为成功的》（Jossey Bass, 2011）一书。

Judith 是六册《达人迷》（*For Dummies*）书籍的合著者。这六册书包括 *Big Data for Dummies*、*Hybrid Cloud For Dummies*、*Cloud Computing For Dummies*、*Service Mangement For Dummies*，以及 *Service Oriented Architecture For Dummies*，第 1 版和第 2 版（John Wiley & Sons）。

Judith 拥有波士顿大学的本科和硕士学位。她在几家新兴公司的顾问委员会任职。她是波士顿大学校友委员会的成员。她在 2005 年被授予“波士顿大学艺术与科学学院杰出校友”的称号。她同时也是 2005 年曼彻斯特技术领导

委员会奖金的获得者。

Marcia A. Kaufman 是 Hurwitz & Associates 有限责任公司的 COO 和原理分析师，这是一家关注包括大数据、认知计算、云计算、服务管理、软件开发及安全和管理等新兴技术的研究和咨询公司。她发起了许多关于高级分析的研究，并且写了大量关于云基础设施、大数据和安全的著作。Marcia 在商业战略、产业研究、分布式软件、软件质量、信息管理及分析方面有着超过 20 年的经验。Marcia 在金融服务、制造业及服务产业都工作过。她在 Data Resources 公司 (DRI) 任职期间，开发了计量经济学产业模型并进行预测。她拥有康涅狄格学院数学和经济学的文学学士学位，并且拥有波士顿大学的 MBA 学位。Marcia 也是上述六册《达人迷》(*For Dummies*) 书籍的合著者。

Adrian Bowles 博士是 STORM Insights 公司的创始人，这是一家研究顾问公司，为新兴技术市场上的买方、卖方以及投资者提供服务。之前，Adrian 为 Object Management 集团、101 Communications 公司的 IT 合规研究所及 Atelier Research 公司组织了治理、风险管理与合规圆桌会议。他曾在 Ovum (Datamonitor)、Giga Information Group、New Science Associates 和 Yourdon 公司担任行政职务。Adrian 对于认知计算和分析的关注自然地延续了他在读硕士期间的研究。(他的第一个自然语言仿真应用发表于控制论和软件国际研讨会期间。) 他也在德雷赛尔大学和纽约州立大学宾汉姆顿分校担任计算机科学方面的学术职务，并且是纽约大学商学院和波士顿学院的特约教授。他获得了纽约州立大学宾汉姆顿分校的心理学本科学位和计算机科学硕士学位，并且获得了西北大学的计算机科学博士学位。

致 谢

写一本话题复杂性类似认知计算的书籍是需要大量研究的。我们的团队阅读了几百篇关于这个领域技术基础多个方面的技术文章和书籍。此外，我们很幸运地找到了许多愿意慷慨地与我们交流的专家。我们希望能够兼收并蓄。因此，我们要感谢很多人。我们可能会遗漏一些我们在会议中遇到的人，他们提供了能够影响这本书主题的有见解的讨论。我们同样要感谢为完成这本书而在我们三人间建立的合作伙伴关系。我们也要感谢我们在 Wiley 的编辑们，包括卡罗尔·朗和汤姆·丁泽。我们感谢来自于三位技术编辑的见解和协助，他们是 Al Nugent、James Kobielus 和 Mike Kowolenko。

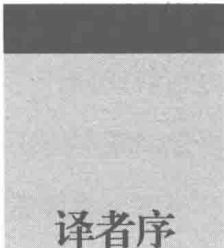
以下的各位慷慨地付出了他们的精力并提供了见解：Manny Aparicio 博士；Avron Barr, Aldo Ventures; Jeff Cohen, Welltok; Umesh Dayal 博士，Hitachi Data Systems; Stephen DeAngelis, Enterra; Rich Y. Edwards, IBM; Jeff Eisen, IBM; Tim Estes, Digital Reasoning; Sara Gardner, Hitachi Data Systems; Murtaza Ghadyali, Reflexis; Stephen Gold, IBM; Manish Goyal, IBM; John Gunn, Memorial Sloan Kettering Cancer Center; Sue Feldman, Synthesis; Fern Halper 博士，TDWI; Kris Hammond 博士，Narrative Science; Ed Harbor, IBM; Martenden Haring, Digital Reasoning; C. Martin Harris 博士，Cleveland Clinic; Larry Harris 博士；Erica Hauver 博士，Hitachi Data Systems; Jeff Hawkins, Numenta and The Redwood Center for Theoretical

Neuroscience; Rob High, IBM; Holly T. Hilbrands, IBM; Paul Hofmann 博士, Space-Time Insight; Amir Husain, Sparkcognition, Inc.; Terry Jones, WayBlazer; Vikki Kolbe, Hurwitz & Associates; Michael Karasick, IBM; Niraj Katwala, Healthline Networks, Inc.; John Kelly 博士, IBM; Natsuko Kikutake, Hitachi Consulting Co., LTD; Daniel Kirsch, Hurwitz & Associates; Jeff Margolis, Welltok; D.J. McCloskey, IBM; Alex Niznik, Pfizer; Vincent Padua, IBM; Tapan Patel, SAS Institute; Santiago Quesada, Repsol; Kimberly Reheiser, IBM; Michael Rhoden, IBM; Shay Sabhikhi, Cognitive Scale; Matt Sanchez, Cognitive Scale; Chandran Saravana, SAP; Manoj Saxena, Saxena Foundation; Candy Sidner 博士, Worcester Polytechnic Institute; Dean Stephens, Healthline Networks, Inc.; Sridhar Sudarsan, IBM; David E. Sweenor, Dell; Wayne Thompson, SAS Institute; Joe Turk, Cleveland Clinic; Dave Wilson, Hitachi Data Systems.

——Judith Hurwitz

——Marcia Kaufman

——Adrian Bowles



译者序

随着人类的进步和发展，人类计算经历了从简单到复杂，从低级到高级的转变。IBM 将人类计算从历史上划分为三个时代。第一个时代是制表时代 (Tabulating Computing)，始于 19 世纪，主要标志是能够执行详细的人口普查和支持社会保障体系。第二个时代是可编程计算时代 (Programming Computing)，始于 20 世纪 40 年代，内容涵盖了从太空探索到互联网。第三个时代是认知计算时代 (Cognitive Computing)。在过去几十年里，科学技术取得了巨大的进步。然而，人类的计算水平并没有获得明显的提高，仍然停留在可编程计算时代，即根据相应的逻辑编写代码以实现预期的目的和功能。如今，新一代的信息系统——认知计算正在兴起。认知计算源自模拟人脑的计算机系统的人工智能，认知系统能够从自身与数据以及与人的交互中学习，从而不断地实现自我提高。认知计算目前虽然还处于发展的初步阶段，但认知系统已经在各领域扩展了人类的认知，在实际生活的各方面为人类提供便利。因此，认知计算具有广阔的应用前景。

本书是专门介绍认知计算原理、发展及未来应用前景的著作。书中深入地探讨了认知计算的基本构成以及怎样使用它来解决问题，探究了在拥有充足文本信息的条件下，分析具有复杂数据和流程的系统时所涉及的人类行为，给出了各种用来使决策更加系统化的因素，包括数据、先前的知识和相互的

关联，以及人类的假设和偏见对决策过程的影响。本书的主要功能是引导读者理解认知计算的各部分组成，以及通过详细探索每个主题，获得对认知计算更加深入的认识。

第 1~2 章主要涉及认知计算系统的基础以及设计原则。

第 3~7 章主要讨论的是现有人工智能方法在未来认知计算系统中扮演的角色和作用，包括认知系统如何使用自然语言处理技术以及这些技术是如何起作用的，大数据技术在认知系统中的支柱作用，如何利用分类学和本体论构建认知系统，认知计算系统对高级分析方法的要求，以及如何将云计算和分布式计算应用到认知计算系统中。

第 8~9 章分析了认知计算的商业意义，通过分析 IBM 沃森（Watson）认知系统，分析认知系统如何构建，以及如何应用到商业中。

第 10~12 章分析了构建认知应用过程中需要考虑的问题，通过认知医疗系统和智慧城市详细分析了这一过程。构建认知应用首先需要明确应用的对象和领域，之后再进行问题的探索和语料库的建立与更新。认知医疗系统主要通过认知应用来改善电子病历和临床教学，以区别于传统医疗系统的方案改善患者的健康状况。而智慧城市主要是指更智能的城市运作和交通基础设施。

第 13~14 章分析了认知计算出现的领域，并展望了认知计算的未来应用。目前认知计算已经逐渐渗透到零售业、旅游业、运输与物流以及通信业，未来的认知计算应用的突破点主要在于认知计算系统构建过程的优化和现有人工智能技术水平的提升。

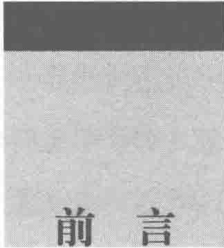
本书是认知计算产业界工程师的理想读物，读者对象可以涵盖研究、开发、系统设计等认知计算领域的相关从业人员。认知计算及相关专业的高年级本科生、研究生和教师也可将本书作为参考书籍。

本书译者长期从事认知计算、未来网络研究工作，具有丰富的理论基础和实践经验。本书主要由北京邮电大学张鸿涛翻译，李榜旭、杨梓华、孟娜、陈莹、牛沐楚等参与了部分章节的翻译，全书由张鸿涛统稿和审校。本书得到了国家自然科学基金资助（61302090）。最后，还要感谢人民邮电出版社的大力支持和高效工作，使本书能尽早与读者见面。

由于认知计算技术演进日新月异，翻译时间比较仓促，疏漏错误之处在所难免，敬请读者原谅和指正。

译 者

于北京邮电大学



前言

在过去的 30 年里，科技取得了巨大的进步，但从数据中获得见解并采取行动的能力却并没有太大的改变。总的来说，应用程序仍然被设计用于执行预定的功能，或者使业务流程自动化，因此，它们的设计者必须为每一个应用场景做准备，并且按照相应的逻辑编写代码。它们不能使自己适应数据中的变化或者从经验中学习。如今的计算机速度更快、更便宜，但是并没有更智能。当然，人类也没有比他们 30 年前聪明多少。但人类和机器的这种情况将发生变化。新一代的信息系统正在兴起，这种系统脱离了旧的进程自动化的计算模型，将为我们提供一个探索的合作平台。这些系统的第一波已经在各种各样的领域扩展了人类认知，作为人类用户的伙伴或合作者，这些系统可能从大量的自然语言文本中获取意义，基于比人类一生能够吸收的还多的数据，在几秒内生成和评估假设。这就是认知计算的未来。

人工智能+机器智能

传统应用擅长自动执行定义明确的进程。从库存管理到天气预报，当速度是成功的关键因素而且进程是事先知道的时候，定义需求、编写逻辑、运行应用的传统方法是足够的。然而，当我们需要动态地寻找和利用数据元素之间模糊的关系时，尤其是在一些数据复杂度急速增加的领域，这种方法将

会失效。改变、不确定性以及复杂度是传统系统的敌人。

认知计算——基于不需要改编程序的软硬件和自动认知任务，展现出一个具有吸引力的应用发展模式或范例。我们开始思考如何提高人脑在新的应用能力下能够达到的最好的表现，而不是使我们已经开展的业务形式自动化。我们开始了从企业的内部和外部摄取数据的进程，增加用于在大量或者非结构化数据集合中辨别、评估模式和复杂关系的函数，例如学术期刊、书籍、社交媒体、图片或音频中的自然语言文本。这样做的结果就是系统可以通过评估上下文的数据，支持人类的推理法，并且呈现与证明结果正确的证据相关的发现。像传统的应用一样，这个方法使用户更加有效率，而它也更有影响力——因为部分推理和学习的进程被自动分配给一个不知疲劳的、快速的协作者。

与传统计算的基础类似，智能机器背后的概念并不新鲜。甚至在数字计算机出现之前，工程师和科学家就推断学习型机器会发展到模仿人类解决问题和沟通的技巧。虽然一些基础技术之下的概念，包括机器智能、计算语言学、人工智能、神经网络以及专家系统，在 10 年前或者更早之前就已经被应用在传统的解决方案中，但我们看见的仅仅是开端。智能计算新纪元受到许多因素共同的驱动：

- 系统、智能设备、传感器、视频等产生的数据量的增长；
- 计算机存储器价格和计算能力的下降；
- 以数据产生的速度分析复杂数据的技术复杂度的增加；
- 全球新兴公司对调查并挑战人机合作成果的长期理念的深度研究。

碎片整合

在你把大数据技术、不断变化的计算经济同工商业智能化的需求结合起

来时，你就开始了根本性的转变。这个模式的转变有很多名称：机器学习、认知计算、人工智能、知识管理和学习型机器。但随着新兴计算系统以前所未有的速度解释多种类型的海量数据，无论你怎么叫它，这种变化实际上是人类对世界认知的最好整合。但它依旧不足以解释或分析数据。认知计算的最新方法就是必须收集有关特定主题的大量数据，与主题专家进行互动，并了解这一主题的背景和语言。这种新的认知时代正处于起步阶段，但由于这个系统具有重要而直接的市场潜力，我们就开始了这本书的写作。认知计算并不玄妙。它是一种实在的方法，通过将改变市场和行业的学习型机器为人们解决问题提供支持。

本书重点

本书深入地探讨了认知计算的基本构成，以及怎样使用它来解决问题。本书同时探究了在改进一个拥有充足文本信息来分析复杂数据和流程的系统时，比如医疗、制造、交通、零售以及金融服务，所涉及的人类行为。这些系统是人类和机器合作的设计。本书调查了各种用来帮助决策更加系统化的项目。训练有素并且有丰富经验的专业人员如何利用数据、先前的知识以及相互的关联来做出明智的决策。由于知识的深度，有时这些决策是正确的。然而，在其他时候，由于学者也会将他们的假设和偏见带入决策过程，因此这些决策就有可能是不正确的。许多在实施他们第一个认知系统的组织在寻求能够将深度的经验和复杂的大数据分析机制结合利用的技术。虽然这个产业很年轻，但这里有许多东西可以从具有开拓性的认知计算工作中学习。

本书和技术概述

本书的作者 Judith Hurwitz、Marcia Kaufman、Adrian Bowles 都是计算机

产业的资深人士。我们都有自己的主张，并且是独立于产业的分析师和咨询师，以综合的观点看待不同技术之间的关系，以及它们怎么改变商业和产业。我们是作为真正的合作者来完成这本书的写作的。从软件开发、新兴技术评估到重要技术创新的深入研究，我们每个人都贡献了不同的经验。

就像许多新兴的技术一样，认知计算也不简单。首先，认知计算代表一种创造应用的新方式，以支持商业和研究目标。其次，它是不同技术的结合，而这些技术对于商业化应用已经足够成熟。因此，你可能会注意到，本书里详细介绍的大多数技术早已在研究中，介绍的产品已经出现了几年甚至数十年。被视为人工智能应用的一些技术或方法，比如机器学习算法和自然语言处理（Natural Language Processing, NLP）已经出现了数十年。其他的技术，例如高级分析，随着时间在不断演进并且变得更加精准。部署模型的巨大变化，比如云计算和分布式计算技术，已经提供了足够的能力和经济规模，将计算能力带入一个仅 10 年前都不可能达到的水平。

本书并非要取代众多出色的技术书籍，那些书籍具有独立主题，比如机器学习、NLP、高级分析、神经网络、物联网、分布式计算和云计算。事实上，我们认为明智的做法是，利用本书来理解这些部分是怎么组合的，之后通过详细探索每个主题获得更加深入的知识。

这本书是如何组织的

本书涵盖了对于构建认知系统十分重要的基本原理和基础技术，同时也涵盖了认知计算的商业驱动力和较早采用认知计算的一些产业。本书的最后一章提出了对于未来的一些思考。

第 1 章：认知计算的基础。这一章提出了从人工智能、机器学习向认知计算演进的观点。

第2章：认知系统的设计原则。这一章使你理解认知计算的架构以及各部分是如何组织在一起的。

第3章：自然语言处理支持下的认知系统。这一章解释了认知系统如何使用自然语言处理技术，以及这些技术如何产生理解力。

第4章：大数据和认知计算的关系。大数据是认知系统的支柱之一。这一章说明了大数据技术以及认知系统中基础的方法。

第5章：在分类学和本体论中表示知识。为了构建认知系统，对于内容，需要存在一个组织性的结构。这一章解释了本体论怎样为没有结构的内容提供含义。

第6章：应用于认知计算的高级分析方法。为了评估有结构 and 无结构内容的含义，这就要求使用大范围的分析技术和工具。这一章探讨需要什么技术工具。

第7章：认知计算中云和分布式计算的作用。没有分布式计算容量和资源的能力，衡量认知系统将会很困难。这一章解释了大数据、云服务和分布式分析服务之间的联系。

第8章：认知计算的商业意义。为什么商业需要构建认知计算的环境？这一章解释了一个组织或企业从认知计算获益的情况。

第9章：IBM 沃森（Watson）——一个认知系统。IBM 通过开启一个“巨大的挑战”开始构建认知系统。设计“巨大的挑战”是为了用来测试它能否对抗这个世界上最优秀的《危险边缘》（Jeopardy）游戏者。这个实验的成功使得 IBM 创造了一个叫做沃森的认知平台。

第10章：建立认知应用的过程。对于一个组织而言，构建自己的认知系统需要做些什么呢？这一章概述了这个过程以及组织需要考虑的问题。

第11章：建立认知医疗系统。每个认知应用根据所属的领域而有所不同。