Broadview
www.broadview.com.cn

MANNING

智能Web算法

（第2版）

Algorithms of the Intelligent Web

(Second Edition)

[英]Douglas G. McIlwraith
[美]Haralambos Marmanis 著
[美]Dmitry Babenko

Yike Guo 作序

达观数据 陈运文 等译

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

智能Web算法

（第2版）

Algorithms of the Intelligent Web

(Second Edition)

[英]Douglas G. McIlwraith
[美]Haralambos Marmanis 著
[美]Dmitry Babenko

Yike Guo 作序

达观数据 陈运文 等译

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

内容简介

机器学习一直是人工智能研究领域的重要方向，而在大数据时代，来自 Web 的数据采集、挖掘、应用技术又越来越受到瞩目，并创造着巨大的价值。本书是有关 Web 数据挖掘和机器学习技术的一本知名的著作，第 2 版进一步加入了本领域最新的研究内容和应用案例，介绍了统计学、结构建模、推荐系统、数据分类、点击预测、深度学习、效果评估、数据采集等众多方面的内容。本书内容翔实、案例生动，有很高的阅读价值。

本书适合对算法感兴趣的工程师与学生阅读，对希望从业务角度更好地理解机器学习技术的产品经理和管理层来说，亦有很好的参考价值。

版权贸易合同登记号　图字：01-2016-7946

图书在版编目（CIP）数据

智能Web算法/（英）道格拉斯· G. 麦基尔雷思（Douglas G. McIlwraith），（美）哈若拉玛·玛若曼尼斯（Haralambos Marmanis），（美）德米特里·巴邦科（Dmitry Babenko）著；陈运文等译.—2版.—北京：电子工业出版社，2017.7
书名原文：Algorithms of the Intelligent Web, 2nd Edition
ISBN 978-7-121-31723-1

Ⅰ.①智… Ⅱ.①道… ②哈… ③德… ④陈… Ⅲ.①互联网络—程序设计 Ⅳ.①TP393.4

中国版本图书馆CIP数据核字（2017）第121014号

策划编辑：张春雨
责任编辑：刘　舫
印　　刷：北京中新伟业印刷有限公司
装　　订：北京中新伟业印刷有限公司
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：15.5　　字数：278千字
版　　次：2017年7月第1版
印　　次：2017年7月第1次印刷
定　　价：69.00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。

将本书献给我挚爱的 Elly。

—*D.M.*

译者序

人工智能和机器学习技术近年来得到了飞速的发展，并成为计算机界乃至全社会炙手可热的话题。这些优秀的技术让每个人的生活越来越方便和智能，这让从业者感到非常欣喜。智能算法是人工智能的核心技术，不论是我当前创办的达观数据，还是之前在腾讯、盛大、百度等互联网企业的工作，都是围绕智能算法展开的，我对此有深厚的热情。因此当电子工业出版社计算机出版分社的张春雨编辑邀请我翻译这本《智能 Web 算法（第 2 版）》的时候，虽然深知翻译和审校要付出大量的时间和精力，但还是很愉快地接受了邀请并完成了翻译工作，希望本书中文版的面世，能帮助广大爱好者建立起对 Web 数据挖掘和机器学习技术全面且直观的了解。

在众多有关机器学习和数据挖掘的书籍里，本书是颇为经典的一本。其特点之一是内容覆盖面很广，有关网络数据挖掘的方方面面都涵盖到了，从数据采集、存储，到降维运算和结构抽取，以及涉及模式识别的聚类和分类、统计机器学习理论等，还有面向互联网应用的推荐系统、搜索引擎、广告点击预测等，配套的效果评估机制也有专门的章节进行讲解，读者阅读本书后可以形成较为全面的学习体系。特点之二是本书较好地在算法思想、数学原理、应用案例之间找到了平衡点。每个章节作者都由浅入深地讲解了算法的思想，并通过列举一些非常生动的案例来让读者更好地理解算法的原理。例如，列举的 Iris 数据集结构的抽取、在线电影推荐系统、金融欺诈检测、广告点击预测等实践案例的讲解都非常清晰易懂。书中对数学公式

的使用点到为止，力求简洁。这样既不像很多教科书那样堆砌数学公式，让很多读者望而生畏，又不像很多书籍那样只是罗列程序代码而不讲解背后的算法思想。这和作者既有工程实践经验，又有学术研究背景密不可分的。

与通常的再版书籍只是做些局部修订不同，本书第 2 版对第 1 版图书的内容进行了全面彻底的升级改写，全书有超过 80% 的篇幅与第 1 版不同，可以说是脱胎换骨的变化。这些变化具体体现在以下三个方面：首先，增加了近年来数据挖掘领域最新的一些研究成果，例如当下炙手可热的深度学习等，同时删减了一些较为陈旧的内容；其次，调整了全书的组织结构，章节的划分更为合理，每章内容更加丰富，列举的案例也更贴近实战。第三，全书的示例代码不再使用第 1 版的小众开发语言 BeanShell，而是改为机器学习界更为常用的 Python，并配合机器学习界知名的开源软件包 scikit-learn，让本书的代码阅读起来更友好，也大大增强了示例代码的实用性。

本书由于篇幅所限，虽然涉及的面很宽广，但是每个章节的内容都没有进一步深入展开。我在翻译过程中，觉得本书有些内容讲得略偏浅显，在所提及的领域都属于入门级的深度，读起来有些意犹未尽。事实上如果深究起来，本书每个章节的内容都足够扩充成一本独立的书籍。好在本书作者提供了很多参考资料，并在相应章节的脚注里细心地进行了标识，对更深入的内容感兴趣的读者，不妨按图索骥，下载相应的论文和著作来一窥究竟。

本书的翻译工作，要深深感谢电子工业出版社的张春雨、刘舫和编辑朋友们给予的大力帮助和耐心指点。同时要感谢我所在的公司——达观数据的各位亲密战友，依靠大家分工协作、共同努力，才顺利完成了全书各个章节的翻译工作，这些同事是于敬、文辉、纪达麒、纪传俊、江永青、冯仁杰、桂洪冠、高翔、王文广、张健、范雄雄、蹇智华、孟礼斌。团结才有力量，大家共同的辛勤工作和智慧结晶，让本书翻译工作顺利完成。

限于译者水平所限，在理解和翻译本书的过程中，一些知识的专递未必到位，所使用的语言也未免生涩，我们力求做到“信、达、雅”，一些不好把握的字句也反复查阅过资料，希望能较为忠实地还原作者的意图，让广大读者能享受通畅的阅读体验。如有疏漏之处，希望读者朋友阅读时多多包涵，并不吝提出各种意见和建议。

人工智能和机器学习技术正在得到越来越多的人的关注，并正在发挥着越来越大的价值。身为其中的一员，我非常荣幸自己能够生于这一历史上最火热的发展时代里，我创办的达观数据，也正在运用本书里所介绍的各种技术，来帮助中国的企

业更好地挖掘数据背后的规律，自动完成很多原本需要大量人力才能实现的功能。创业维艰，本书的很多翻译和校对工作是在出差途中和深夜完成的，感谢家人对我的理解和关怀。期望达观数据的技术服务能让很多企业提升运行效率、降低成本，从原先的粗放型增长转变为技术驱动型的精细化增长。

眼下全球技术竞争愈演愈烈，数据作为人工智能时代的原油，对其进行提炼和挖掘的技术至关重要。我希望包括本书在内的一系列国外优秀书籍被翻译引入后，能够帮助中国的技术人才、工程师、学生乃至企业管理者拓展视野、启发思维，把握业界的技术发展脉搏，成为大数据时代浪尖的弄潮儿。

陈运文

达观数据创始人兼 CEO

译者简介

陈运文，计算机博士，达观数据 CEO，ACM 和 IEEE 会员，中国计算机学会高级会员；在大数据架构设计、搜索和推荐引擎、文本数据挖掘等领域有丰富的研发经验；曾经担任盛大文学首席数据官、腾讯文学数据中心高级总监、百度核心算法工程师等工作，申请有 30 余项国家发明专利，多次参加国际 ACM 数据算法竞赛并获得冠亚军荣誉。

序言

万维网（World Wide Web）是互联网信息社会里的最根本的基础设施，数以亿计的人们把它作为主要的交互联系工具。互联网上信息服务的发展也带动了工业的进步。今天，随着云计算和无线通信技术的成熟，Web 不仅成为人们发布和获取信息的平台，而且成为为数亿人随时随地提供信息服务开发、部署和应用的平台。大数据为构建多样性的服务提供了丰富的内容，也为智能化的服务创造了价值，让 Web 上服务的用户体验逐步提升。智能服务的 Web 正在改变人们的日常生活：它帮助我们寻找合适的酒店、安排完美的假期旅行，让我们购买到几乎任何商品，以及建立起丰富多彩的社群，而这些智能来自对 Web 内容和用户间交互所产生的数据的深度分析。因此建立 Web 智能是当今数据科学发展领域里的核心技术。

非常荣幸能由我来为大家介绍这本精彩的《智能 Web 算法（第 2 版）》，本书由一位年轻但经验丰富的数据科学家 Douglas McIlwraith 博士修订，目的是为大家揭示智能 Web 应用的精髓：实现智能所依赖的各种算法。这是一个宏伟的目标，但是 Doug 博士用朴实无华的语言，在不到 250 页的篇幅里成功将丰富的知识通俗易懂地呈现了出来。

本书涵盖了丰富的应用场景和常见的流行算法，并通过严谨的数学推导和简洁

的 Python 代码对这些算法进行了清晰的介绍。我非常顺畅地通读了本书，也希望能与你一起分享阅读的乐趣。更为重要的是，我希望当你阅读完本书后，发现自己可以用学会的很多知识和技能，打造出更智能的 Web！

Yike Guo
教授 & 总监
数据科学研究所
伦敦帝国理工

前言

非常荣幸我们能投身于当今时代最令人激动的一个技术领域。在短短数十年间，稚嫩的互联网就蓬勃发展成如今连接全世界的万维网，让每个身在其中的人随时随地进行通信交流，让大家拥有了瞬间就能得到几乎任何问题答案的能力。

智能算法的研发充分运用了信息的价值，在塑造我们新的生活方式上扮演了重要角色。反过来我们也越来越依赖智能算法来引领我们线上和线下的生活，这也促使我们将更宽的视野和更多的数据用于算法的训练和测试。若干年前神经网络算法还是被学术界所摈弃的方法，但是如今随着大规模高可用的数据技术的发展，神经网络技术再次大放异彩。

我们刚刚进入一个新纪元，在这里我们能与手机对话，让它预测我们的需求、预订我们的约会、建立我们的通信连接。在不久的将来，我们也许能看到无人驾驶汽车和虚拟现实技术的普及，所有这些应用都牢牢地扎根于计算机科学技术对真实世界问题的回应，智能算法是其中的重要部分，也是本书的核心。

不幸的是，进入机器学习和数据科学的世界看上去令人生畏，这里充满了数学和统计学，你的直觉有时也会误导你！通过修订本书，我们希望介绍第一版面世以来该领域的最新发展，也为新入行的朋友们提供指引。在本书中我们提供了通俗易懂的实例、真实问题的解决方案，以及相应的代码片段。我们尽可能地越过繁复的

数学公式来重点阐述技术的核心思想，希望我们对此拿捏得足够好。

在本书中你将看到，我们把内容划分为 8 个章节，每个章节涵盖智能 Web 的一个重要的算法领域。本书最后的附录部分讲解了智能 Web 应用中的数据处理流程，我们希望通过这部分内容，来为实践者展示在系统中将快速变化的数据有效地运转起来是多么重要且困难。

读者服务

轻松注册成为博文视点社区用户（www.broadview.com.cn），扫码直达本书页面。

- 下载资源：本书如提供示例代码及资源文件，均可在下载资源处下载。
- 提交勘误：您对书中内容的修改意见可在提交勘误处提交，若被采纳，将获赠博文视点社区积分（在您购买电子书时，积分可用来抵扣相应金额）。
- 交流互动：在页面下方读者评论处留下您的疑问或观点，与我们和其他读者一同学习交流。

页面入口：http://www.broadview.com.cn/31723

致谢

感谢在本书撰写过程中参与的各位伙伴：编辑 Marjan Bace 以及出版发行团队的所有成员，包括 Janet Vail, Kevin Sullivan, Tiffany Taylor, Dottie Marsico, Linda Recktenwald，以及幕后的很多工作人员。

也感谢参与本书各阶段校对的人员：Nii A-Okine, Tobias Bürger, Marius Butuc, Carlton Gibson, John Guthrie, Pieter Gyselinck, PeterJohn Hampton, Dike Kalu, Seth Liddy, Radha Ranjan Madhav, Kostas Passadis, Peter Rabinovitch, Srdjan Santic, Dennis Sellinger, Dr. Joseph Wang, Michael Williams。感谢你们反复阅读，认真进行校对，你们提供的宝贵意见在本书中得到了充分体现。

本书中引用的很多系统、函数库、程序包并非作者原创，而是来自本领域的众多社区开发者、数据科学家、机器学习专家，在此对以上所有人表示感谢。

回想起最初讨论修订《智能 Web 算法》时的情形，记得我当时心里想“嘿，这本书的第一版已经写得很好了，修订的工作量不会很大吧？”但最后结果是，很大。该领域的变化很快，有太多有趣的工作我想拿来与人分享，因此我不得不仔细地选择哪些该舍弃、哪些该删减、哪些该修订、哪些该增加。因此本书花费了比我预料更多的时间，但我很幸运获得了很多优秀的人们的支持、鼓励和忍耐。

首先也是最重要的，我想感谢我的未婚妻，Elly。你的爱心、忍耐、鼓励，是我生命中永恒的存在。如果没有你，本书是难以完成的。我爱你。

其次，我想感谢我的父母和家人，在我遇到挫折时永远呵护和支持我，希望你们能喜欢本书，你们的养育之恩我永远铭记。

第三，感谢我的众多朋友和同事，和杰出的你们在一起工作是一件非常幸运的事，你们让我每一天都过得很开心，谢谢你们！

我还想感谢我的两位编辑 Jeff Bleiel 和 Jennifer Stout，你们的指导帮助本书最终完成。Jennifer，你的乐观和热情给了我坚持的动力，谢谢你！

Douglas McIlwraith

我想感谢我的父母 Eva 和 Alexander，他们无微不至的关心，让我在夜以继日的写作和研究中，始终保持着好奇心和热情。这是我毕生难忘的恩情。

我衷心感谢我珍爱的妻子 Aurora 和我的三个孩子：Nikos, Lukas 和 Albert——你们是我人生的骄傲和乐趣。我永远感激你们给予的爱心、耐心和理解。孩子们无尽的好奇心不断地激发我学习的灵感。非常感谢我的岳父母 Cuchi 和 Jose，我的姐妹 Maria 和 Katerina，以及我最好的朋友 Michael 和 Antonio，感谢你们持续的鼓励和无条件的支持。

一定不能遗忘的是感谢 Amilcar Avendaño 博士和 Maria Balerdi 博士给予的众多帮助，让我学会了很多心脏学的知识，并打下了我早期的学习基础。感谢 Leon Cooper 教授以及布朗大学的众多杰出朋友，你们不仅揭示了很多大脑运行的规律，还鼓励我开展智能应用的工作。

鼓励和支持我进行各种智能相关的积极工作的过去和现在的同事：Ajay Bhadari, Kavita Kantkar, Alexander Petrov, Kishore Kirdat，等等，虽然这里只能写下寥寥数语，但是我对你们的感激之情溢于言表。

Haralambos Marmanis

首先也是最重要的，我想感谢我亲爱的妻子 Elena。

我还想谢谢我过去和现在的同事：Konstandin Bobovich, Paul A. Dennis, Keith Lawless 和 Kevin Bedell，你们伴随了我的职业生涯，是我的灵感源泉。

Dmitry Babenko

关于本书

本书为读者提供了设计和创造智能算法的路线指引，本书汲取了计算机科学很多领域的知识，包括机器学习和人工智能，并结合了很多作者的实践和思考。书中融入了不少实际操作技巧，介绍了本领域最新涌现的前沿技术，还提供了若干真实可行的实例，读者可以对其修改后在实践中使用。

本书适用的读者

本书主要针对已掌握了扎实的编程技能和基本数学及统计学知识的智能算法初学者。在本书写作过程中我们尽量淡化数学推导，更多地为你勾勒方法的适用性的整体印象。当然如果你更愿意探究数学内容，我们鼓励你深入推敲细节。本书的读者最好具备基本的编程经验并学习过大学数学课程。

路线图

本书由 8 个章节和 1 个附录构成。

- 第 1 章介绍了智能算法的概况和一些关键特性，也提供了本书其余部分的整体指引。

- 第 2 章讨论了数据内部结构的概念，尤其是介绍了特征空间的概念，在本章中我们详细讲解了期望最大和特征向量。
- 第 3 章介绍了推荐系统。我们介绍了协同过滤技术，并讨论了 Netflix 竞赛。
- 第 4 章概述了分类技术，介绍了逻辑回归，我们使用逻辑回归来解决缺陷检测问题
- 第 5 章通过一个案例讲解了在线广告中的点击预测问题。我们概述了在线广告系统的后台运作机制，并基于一个公开的网页点击数据集，提供了一套可运转的点击预测程序实例。
- 第 6 章是关于深入学习和神经网络的内容。我们对神经网络进行了短小精要的介绍。从神经网络最终的原型到近年来深度网络学习的最新进展都有涉猎。
- 第 7 章概述了如何做出最优的选择。我们讨论了 A/B 测试中统计的重要性，以及将多臂赌博机技术用于在线学习的若干种方法。
- 第 8 章介绍了智能 Web 的前瞻性总结。
- 附录讨论了我们应该如何处理快速变化的事件流和用之来构建智能算法。我们讨论了几种 Web 日志处理的设计模式，提炼了几种需要避免的关键误区。

大部分情况下，这些章节内容独立可以分别阅读，但是第 5 章的案例分析依赖于在第 4 章中介绍的逻辑回归的知识。

资料下载

运行本书中实例时用到的所有代码和数据都可以从出版社的网站上下载到（www.manning.com/books/algorithms-of-the-intelligent-web-second-edition），或者从 GitHub 下载（https://github.com/dougmcilwraith/aiw-second-edition）。唯一例外的是 Criteo Display Challenge 数据集，因为尺寸很大，你需要从 Criteo 的官网直接下载。第 5 章提供了下载方式的说明。

所有代码都已经在 Ubuntu 14.04.2 环境下用 Python 2.7.10 测试过。各种运行依赖在下载包中的需求文件中可以找到。在文件中也可以找到确保本书中代码样例兼容性的运行环境的说明书。

编码规范

本书提供了很多示例，以清单形式展示的源代码和文本中的代码都用等宽字体来和普通文本区分开来。我们在部分区域增加了换行符或调整缩进符，以让书页容纳下较长的代码。在代码过长的情况下我们甚至使用了行连接符（➥）。另外，当源代码在文章中已有介绍的时候，我们会把代码中的注释去掉。在代码清单中有时伴随一些注释说明来突出显示重要的概念。

数学规范

文章中会使用很多数学公式来辅助说明代码和概念。全书中我们都遵循了标准的公式标记规范[译注1]，矩阵采用正体加粗大写字母表示，例如 **M**；正体加粗小写字母表示向量，如 **v**；标量则用斜体小写字母表示，例如 λ。

关于作者

Douglas McIlwraith 博士在剑桥大学计算机科学系获得了学士学位，而后在帝国理工大学获得了博士学位。他是一位机器学习专家，目前他在位于伦敦的一家广告网络公司担任数据科学家职位。他在分布式系统、普适计算、通用感知、机器人以及安全监控方面都贡献了研究成果，他为让技术更好地服务人们的生活而无比激动。

Haralambos Marmanis 博士是将机器学习技术引入工业解决方案的先驱，在专业软件研发方面拥有 25 年经验。

Dmitry Babenko 为银行、保险、供应链管理、商业智能企业等设计和开发了丰富的应用和系统架构。他拥有白俄罗斯国立信息和无线电大学计算机硕士学位。

关于封面图片

本书的封面图片来自一本介绍特色服装的法文书籍 *Encyclopedie des Voyages*，作者是 J. G. St. Saveur，1706 年出版。在当时，旅游还是一件很新鲜的事情，类似该书这样的手册很受欢迎。无论是旅行者还是足不出户的读者，都能从书中了解到

译注1 本书中使用的公式标准与我国国标中的公式标准有所不同，但因本书中涉及的公式较多，故沿用原书中的公式标准。