Broadview

MANNING

智能Web算法

Algorithms of the Intelligent Web

Haralambos Marmanis
Dmitry Babenko 著

隐 陈钢 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

Algorithms of the Intelligent Web

[美] Haralambos Marmanis
Dmitry Babenko
著

阿稳 陈钢 译

電子工業出版社
Publishing House of Electronics Industry

内 容 简 介

本书涵盖了五类重要的智能算法：搜索、推荐、聚类、分类和分类器组合，并结合具体的案例讨论了它们在 Web 应用中的角色及要注意的问题。除了第 1 章的概要性介绍以及第 7 章对所有技术的整合应用外，第 2～6 章以代码示例的形式分别对这五类算法进行了介绍。

本书面向的是广大普通读者，特别是对算法感兴趣的工程师与学生，所以对于读者的知识背景并没有过多的要求。本书中的例子和思想应用广泛，所以对于希望从业务角度更好地理解有关技术的技术经理、产品经理和管理层来说，本书也有一定的价值。

版权贸易合同登记号　图字：01-2010-6423

图书在版编目（CIP）数据

智能 Web 算法 /（美）玛若曼尼斯（Marmanis,H.），（美）巴宾寇（Babenko,D.）著；阿稳，陈钢译. —北京：电子工业出版社，2015.3

书名原文：Algorithms of the intelligent Web

ISBN 978-7-121-25456-7

Ⅰ. ①智… Ⅱ. ①玛… ②巴… ③阿… 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(2015)第 020204 号

策划编辑：张春雨

责任编辑：李利健

印　　刷：北京中新伟业印刷有限公司

装　　订：河北省三河市路通装订厂

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：25　字数：440 千字

版　　次：2015 年 3 月第 1 版

印　　次：2015 年 3 月第 1 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。

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。

译者序

在这篇文章里，我会谈谈自己在翻译过程中的一些比较深刻的感受，以期让读者在正式阅读本书之前对它有一个整体的了解，比如，书中算法讲解背后所隐含的更重要的思想，又如作者使用了什么样的方式来传递书中的知识，以及本书的受众群体。

虽然书中各章论述的是不同的理论和不同的应用，但作者一直在试图传递一些这个领域里不变的朴素而实用的思想，其中一个最重要的思想是：组合不同的技术，以得到更好的结果。这无论是在学术界，还是在业界，都已经逐渐成为潮流。在 2011 年中国推荐系统峰会上，张栋博士在总结自己参加 Netflix 竞赛感受时就曾说到：一个好的推荐器无法打败无数技术组合起来的推荐器。而书中第 3 章也引述了 Netflix 竞赛获胜者的话：“我们没有发现完美的模型，我们最好的结果是来自对具有互补作用的模型预测结果的组合。”

此外，书中还反复强调了一些很具有现实参考意义的观点，限于篇幅，简要列举几点如下：

（1）对问题本质及数据性质的理解比使用算法更重要。2011 年中国推荐系统峰会也有一个类似的为人们所争议的热点，有的与会者认为以重要性而论，领域知识>数据>算法。有过一定从业经验的算法工作者，对此应能有深刻体会。

（2）实验环境的数据要能代表现实数据。从统计与抽样理论的角度来说，这个命题的重要性毋庸置疑。而对于先在离线环境中训练模型，然后在生产环境中使用模型的智能应用模式而言，这一点尤其重要，你是不可能把一个在北京计算得到的

空气质量预测模型直接拿到北海道去用的。

（3）当你想了解所做的改变产生的效果时，最好一次只改变一个因素。这其实与 AB-Test 的思想具有异曲同工之妙，要想研究某个参数或某个变量修改前后的差异，需要尽量保证除此之外的其他因素不变。

所以，这不单是一本介绍算法之术的书，更是一本介绍算法之道的书。读者在学习其中各种具体算法的同时，不妨多思考作者总结出来的这些实战性很强的经验，因为那些算法不一定是最新最有效的，但那些经验教训则是需要经历实际的成败方能总结出来的。此外，我们建议读者不要只看代码的实现，还要关注蕴含于其中的一些设计原理以及对问题建模的方式。因为在实际的应用环境中，你需要考虑的可能不仅仅是一两个算法，而是一个具体的系统问题。

无论是多么朴素和实用的思想，传授起来也难免让人觉得枯燥，所以作者尽最大努力从工程化的角度来介绍这些算法。其中一种尝试就是尽量摒弃对数学公式的使用，而代之以论述具体问题与代码说明的方式来解释算法。是否依赖数学公式来说明问题是一个值得商榷的事情，数学公式就好像是一种共同的语言或标准，只要大家都懂这种语言，沟通基本上不会存在问题，而且简练、优雅，无二义性。但霍金也曾经说过，多一个数学公式，就会吓跑一半的读者，所以在他那本经典著作《时间简史》中只保留了最重要的一个公式。近年来，面向大众的技术书籍也有这种尽量摆脱乏味的数学公式堆砌的风潮，特别是国外的很多优秀著作，即使只用很少的公式也能把问题说清楚。

依我的观点，一本书籍对数学公式的取舍与使用量的多少，取决于作者对读者人群的界定，以及这一读者群中公认的沟通标准。无疑，在 IT 技术从业者中，按通用性而论：自然语言>代码>数学公式。虽然作为这一领域的从业者，懂得基本的数学知识是必需的，但为了使本书适用于更广泛的程序员阶层人群，作者尽力使用代码与自然语言来描述那些算法，而不是数学公式。于是，你将会很惊讶地看到，一本讲述算法的著作里居然找不到几个数学公式，而且这一点儿都不妨碍你对这些算法含义的理解。

另一方面，要填平从理论到实践、从学校到业界的鸿沟并不是件容易的事情，特别是在我国现有的教育体系下，在校期间学习的知识与现实需求的差异尤其大。本书作者既有工程的从业背景，又有非常丰富的机器学习研究经历，使得他可以用一种实用性较强的方式把业界所需要的知识点传递出来。被证明最直接有效的传递

知识的方式当然就是案例化的教学，书中每一个讲授具体算法的章节都会辅之以一个现实中的案例，比如文档搜索引擎、在线音乐推荐商店、用户的信用等级分类等。

除了案例化的知识传授方式，每章后的 To Do 事项则充分体现了西方人所崇尚的启发式教学思想，虽然有的读者也许并不乐意自己再去探索，而更愿意作者把所有的答案都告诉我们，但如果对照着现成的代码把一个个具有探索性意味的 To Do 事项完成，你获得的将是超越本书所教授的内容的，属于自己的知识。学会独立思考、独立实践，也是一名普通工程师与优秀工程师的重要分水岭。

虽然有不少的优点，但本书也仍有其不足之处。书中对智能技术的主要方面都有所论及，但在如此篇幅的书中把所有问题都论述清楚是不可能的，过于追求全面就会导致深度方面稍有缺乏，所有领域相关的问题未必能方方面面都讨论透彻，对于资深从业人员来说，未免会有点意犹未尽的感觉。此外，虽然本书是本着工程的理念来讲述的，在代码实现方面也考虑了一些工程上效率的因素，但这些代码毕竟只是供教学演示之用，为追求简明易懂而未免显得简陋，所以不能寄希望于把它们直接用于自己的生产环境中。

这不是一本适用于所有对该领域感兴趣的读者的书，它有其适用人群。

适用人群之一：软件或互联网从业工程师，如果你原来没有接触过相关的知识，而又想让自己的代码拥有更多的智能化特性，这本书也许能给你带来新的启发和新的思维。

适用人群之二：如果你已经是书中所涉及领域的从业人员，那么阅读本书也许会帮助你梳理原有的知识体系，但如果你在寻求某个领域的最新技术动向，此书则不合适，因为书中讲述的都是业界上已经获得成熟应用的概念与算法。

适用人群之三：想在数据挖掘、智能技术上发力的企业管理人员或产品经理。据我了解，一些互联网公司在发展到达一定的阶段后，纯粹的人力与资金上的增加已经无法带动效益的同比增长，他们都在寻求发展模式上的转变，转粗放型增长为技术型增长，希望这本书关于智能技术的概念与案例能给有这方面需求的人带去一定的帮助。

适用人群之四：在自己的职业规划之路中，有算法工程师、数据挖掘工程师、数据分析师这些关键词的在校学生。

本书由我及我的搭档陈钢联合翻译完成。我们很希望，该书的出版能为正在思考自己出路的在校学生，或每天在计算机面前敲打着代码的工程师们，提供多一种

职业发展的可能性。只有更多优秀的人才投身于这个领域，才能使国内 IT 公司对智能技术应用产生广泛的重视，进而推动国内 IT 行业的创新。如果这本书能成为未来许多年算法工程师的入门书籍，则是它最大的成功之处。

这本书的翻译出版，首先要感谢我所任职的公司豆瓣网，如果没有这个宽松自主的工作环境，我很难把两三个月几乎所有的业余时间都花在这件事情上。当然，如果没有在这个地方三年来的从业经验，以及在那帮优秀的同事身边成长的经历，我从一开始就不会有信心把这件事情做好。我还要特别感谢豆瓣电台，正是它知心动听的歌曲陪伴我度过了那一个个枯燥的夜晚。还有有道词典，它在网络释义与例句上给予了我莫大的帮助。而恰巧，这两个个性鲜明的互联网产品正是本书所叙述的智能 Web 技术应用的杰出代表。同样要感谢电子工业出版社博文视点公司的编辑们，以及给我们提出过宝贵意见的中国推荐社区（http://resyschina.com）的朋友们。

因为自身水平有限，在理解与翻译本书[1]的过程中，一些知识的传递未必到位，但认识总是可以通过交流与思考来加深的。所以，接下来我们也希望借助豆瓣的读书笔记功能把我们在翻译过程中得到的一些认识写下来与大家分享交流。大家在阅读过程中也不妨通过这种形式来分享自己的理解，提出自己的问题，形成一种思想上的互动，因为与同行交流是促进思考与进步的最重要手段。读者可以在博客或豆瓣上找到我们。

阿稳：http://www.wentrue.net/blog/或http://www.douban.com/people/wentrue/

阿稳

感谢昆明理工大学的彭玮老师和中国推荐系统小组中的各位同学在百忙之中抽时间审读我们的译稿。感谢电子工业出版社博文视点的各位编辑在本书翻译过程中给予的指点和帮助。特别要感谢本书的另外一位译者阿稳在技术上的指点。没有你们的支持和帮助，本书的翻译工作不可能这么顺利。在翻译本书的几个月里，尤其是最后审稿的阶段，我不得不将本应该陪伴已怀孕的妻子的时间投入到本书的翻译工作中，特将本书献给我的妻子王倩和我们即将到来的孩子。

陈钢：http://gossipcoder.com/或 http://www.douban.com/people/oldbeggar/

陈钢

1 由于本书篇幅较大，为了节省成本和便于读者对照原书阅读，我们用“⊏⊐”标出了原书对应的页码，本书的索引所列页码为原英文版页码。

前　言

在读研究生时，我开始接触到机器学习，尤其是模式识别。我的工作主要是数学建模和数值模拟，而海量数据的模式识别其实在很多领域都有着广泛的应用。以前也未曾想到，这些年我会如此深入地探索机器学习领域。

1999 年，我完成学业，开始进入企业工作。在我担任顾问的一个项目中，我们试图根据患者的心电图判断出他们患心脏病的概率。显然，对这种问题，不存在也不可能推导出一个精确的数学公式。现实中，心脏病专家已经对大量的患者患心脏病的风险做出了诊断，而我们建模所使用的方法要能从这些病历中学习如何预测患心脏病的风险。通俗地说，我们要寻找的是能从用户输入的数据中不断地“学习”新知识的方法。

同时，在 20 世纪 90 年代，各种因素汇聚在一起导致了一个新产业的飞速发展——网络变得无处不在！根据摩尔定律，CPU 的运行速度变得更快，而且价格更便宜。RAM 模组、硬盘等各种计算机硬件的性能也日新月异，而价格则是一降再降。随之而来的是，网络连接的带宽不断增长，价格也能被更多的人接受。此外，健壮的 Web 应用开发技术已经成熟，而各种开源项目的蓬勃发展更是促进了相关技术的进步。所有的这些因素构成了现在我们称为 Web 的庞大生态系统。

显然，作为软件工程师和 Web 开发人员，我们首要的任务就是为构建健壮、可扩展、美观的 Web 应用提供足够的技术保障。正是如此，在过去的十年里，人们为此做出了巨大的努力，也获得了可观的成绩。当然，没有最好，只有更好，我们依然有进步的空间。虽然我们一直在追求更健壮、可扩展性更好、更美观的 Web 应

用，然而我们已经遇到了瓶颈。在我们看来，单调乏味的互联网应用已经成为过去，仅仅是聚合数据，简单地根据预定逻辑工作的用户请求/响应模型也已经走到了尽头。

现在，在某些应用中已经出现了一股新的浪潮，让人们对互联网应用有了新的认识。这就是本书中所说的智能应用（intelligent application）。不同于传统的应用，智能应用能根据用户的输入调整自己的行为，就像我那个能根据心电图预测患心脏病概率的建模软件。

最近五年，我渐渐地发现，对于大部分软件开发人员来说，构建智能应用的技术依然未曾掀开神秘的面纱。在我看来，这是由两方面的原因造成的。一方面，这些技术潜在的商业价值可以带来巨大的经济回报。所以从经济方面考虑，对这些应用进行保护，隐藏其中的关键细节是可以理解的。另一方面，几乎所有的相关技术都源自学术研究，需要较强的数学背景才能理解。对于第一个原因，我们无能为力，但在随时能获取海量知识的今天，第二个原因依然是不可逾越的障碍吗？我可以简短而明确地回答“不是!”。如果想要详细地回答，那就阅读本书吧！

我决定写这本书，是为了说明这些技术是可以用算法来表示的，并不需要读者有很强的数学基础。本书的目的是让读者掌握一些有助于在应用中构建智能行为的技术，同时尽可能地降低掌握这些技术的数学门槛。代码以算法的形式包含了所有必要的数学知识。

最初，我想用开源的库来演示这些技术，但大部分的此类库都是为了解决具体问题，而不是为了演示底层的技术而开发的。因此，这些库的源代码通常都是冗长且晦涩难懂的。显然，如果能有清晰、带注释的代码，一定会让本书的读者获益更多。Dmitry就是在这个时候加入了本书的写作，并最终编写完成了本书中的大部分代码。

尽管增长缓慢，但关于这个激动人心的新领域的书籍肯定将逐渐增多。本书只是一本有关这个依然在迅速增长的大领域的入门书籍。所以，本书所涉及的算法是很有限的，对算法的解释也比较简要。我的目标是选择并探讨一些有代表性的话题，而不是写一本代码手册或是有可能让读者晕头转向，内容包罗万象的书。

我希望能通过以下四个方面来实现我的目标：

- 集中精力关注清晰易懂的例子。
- 使用高级脚本语言来演示算法的使用，就像读者在自己的应用中使用这些算法一样。

- 通过大量的 To Do 事项让读者有机会尝试并思考这些代码。
- 编写高水平的、易读的代码。

那么，端着您最喜爱的热饮，坐好，来试试这些聪明的应用吧！它们就在本书中！

Haralambos Marmanis

致 谢

感谢 Manning 出版公司的人让我们有机会出版这本书。他们不仅让本书从书稿最终变成一本书，而且在我们的写作进度一再延迟的情况下非常耐心地帮助我们最终完成了本书。我们尤其要感谢 Marjan Bace、Jeff Bleiel、Karen Tegtmeyer、Megan Yockey、Mary Piergies、Maureen Spencer、Steven Hong、Ron Tomich、Benjamin Berg、Elizabeth Martin，以及 Manning 出版公司所有为本书做出贡献的人们，感谢你们的努力工作。

对审稿人和 Author Online 论坛访问者在本书上耗费的时间和精力，以及给我们的宝贵意见，我们也深表谢意，你们的建议从诸多方面提升了本书的质量。我们知道真正的“自由”时间对于专业人员是非常有限的，所以请接受我们对你们的贡献表示诚挚的谢意。

我们尤其要感谢下面这些在本书撰写过程中反复阅读书稿，并提出宝贵意见的审稿人：Robert Hanson、Sumit Pal、Carlton Gibson、David Hanson、Eric Swanson、Frank Wang、Bob Hutchison、Craig Walls、Nicholas C. Heinle、Vlad Gorsky、Alessandro Gallo、Craig Lancaster、Jason Kolter、Martyn Fletcher 和 Scott Dawson。很重要的是，我们要诚挚地感谢本书的技术校对员 Ajay Bhandari，他在本书付梓之前仔细地阅读了每一章的内容，并检查了所有的代码。

H. Marmanis

我要感谢我的父母——Eva 和 Alexander，他们循序渐进地引导我对学习的好奇

心和热情，这正是我废寝忘食写作和研究的原动力。一生都难以报答他们的养育之恩。

我由衷地感谢我亲爱的妻子 Aurora 和我的三个儿子：Nikos、Lukas、Albert，他们是我生活中最大的骄傲和乐趣，感谢他们对我的爱、耐心和理解。孩子们永不停歇的好奇心促使我不断地学习。此外，还要感谢我的岳父岳母 Cuchi 和 Jose、我的姐妹 Maria 和 Katerina，以及我最好的朋友 Micheal 和 Antonio，感谢他们长期以来给我的鼓励和无条件的支持。

无论如何，我都不会忘记 Drs. Amilcar Avendaño 和 Maria Balerdi 的大力支持，他们教给我很多关于心脏的知识，还资助了我关于“学习”的早期研究。此外，还要感谢 Leon Cooper 教授和布朗大学所有让人啧啧称奇的人们，他们对人类大脑的研究热情感染了很多像我一样的人，并激发了我对智能应用的研究兴趣。

对于曾经和现在的同僚 Ajay Bhandari、Kavita Kanetkar、Alexander Petrov、Kishore Kirdat 等在智能研究中给予我的鼓励和支持，我的感激之情远远超出这寥寥数行文字。

D. Babenko

首先，我要感谢我亲爱的妻子 Elena。在撰写本书一年多的时间里，作为丈夫，我的时间几乎都花在了工作或是本书的撰写上，她对此却毫无怨言。她的支持和鼓励为本书的顺利完成提供了一个完美的氛围。

我还要感谢所有现在或曾经与我共事，并影响了我的专业生涯，给我带来无数灵感的同事：Konstantin Bobovich、Paul A. Dennis、Keith Lawless 和 Kevin Bedell。

最后，我要感谢我的合著者 Marmanis 博士让我参与这项工作。

关于本书

现代 Web 应用那绚丽流畅的用户界面经常为人津津乐道。这些应用的另一个方面则不太为人所知，那就是利用各种技术对信息进行智能化的处理，带来其他方法所不能给予的价值。这些技术的成功例子包括我们常见的 Google、Netflix 和 Amazon。这些应用的核心智能是由一些算法实现的，而本书要介绍的就是这些算法。

本书涵盖五类重要的算法：搜索、推荐、聚类、分类和分类器融合，其中的每一个算法都可以写成一本完整的书，所以面面俱到并不是本书的目的。本书只会对这五类算法做基本的介绍，我们的目的是展示智能应用中的基本算法，而不是覆盖计算智能中的所有算法。本书是为普通读者撰写的，所以尽可能地降低了对读者知识背景的要求。

本书的一大特点是在每章结尾都有一个很特殊的小节，我们称为 To Do，其目的不仅仅是提供额外的参考资料，每个 To Do 小节还会指导读者更深入地了解这一章的主题，激起读者思考其他可能的好奇心，以及现实应用中可能要面对的挑战。

本书大量地使用了 BeanShell 脚本库，目的有两个：其一，让读者从更高的层次上理解算法，避免过早地陷入细节；其二，清楚地描述如何将这些算法整合到读者自己的应用中。在大部分情况下，读者只要写很少的几行代码就能使用本书附带的库。不仅如此，为了维护这些源代码，确保其时效性，我们还在 Google Code（http://code.google.com/p/yooreeka/）上专门建了一个项目。

全书结构

本书共分 7 章，第 1 章是简介，第 2～6 章分别介绍搜索、推荐、聚类、分类和分类器组合，第 7 章介绍如何把前几章中的算法整合到一个具体的应用中。

尽管章节之间有一些联系，但这并不会妨碍你单独阅读第 1～5 章中的任何一章。第 6 章是以第 5 章为基础的，如果单独阅读第 6 章，则可能有些难度。第 7 章涉及本书所有的内容，单独阅读该章也会有些困难。

第 1 章介绍了智能应用的概况，并举例说明了智能应用的意义。这一章从实践的角度定义了智能 Web 应用和一些设计原则，接着介绍了六大类 Web 应用，这些应用都可以利用本书中介绍的智能算法加以改进。这一章还讲述了本书所涉及的算法的历史起源，及其与人工智能、机器学习、数据挖掘、软计算等领域的关系。最后还总结了八条具有重要实践意义的设计原则。

第 2 章首先描述了依赖于传统信息获取技术的搜索方法。对传统方法稍作总结后，逐步转向不仅仅是索引的搜索，其中包括最负盛名的链接分析算法——PageRank。还有一节介绍了如何对用户的点击进行分析来提高搜索结果的质量。这项技术能获知用户对某个网站或话题的喜好，而且有很大的改进潜力，可以扩展出很多新的特性。

第 2 章还介绍了一个用于非网页文档搜索的新算法——DocRank。这个算法有一定的前景，但更重要的是这个算法说明了稍做改动，链接分析中的基本数学原理就能快速地扩展到其他应用中。另外还介绍了一些处理超大网络时可能会遇到的挑战。最后，介绍了有关搜索结果的可信度和验证的问题。

第 3 章介绍了两个重要的概念，即距离和相似度。然后介绍两大类构建推荐系统的技术——协同过滤和基于内容的方法。该章以一个虚拟的在线音乐商店为例，介绍了如何为其开发推荐系统，还介绍了两个更通用的例子。第一个例子是一个假想的网站，利用 Digg 的 API 获取用户感兴趣的内容，然后据此向用户推荐其没看过的文章。第二个例子是关于电影推荐的，引入了数据规范化（data normalization）的概念。本章还介绍了基于均方根误差的推荐系统精确性评价的方法。

第 4 章介绍了聚类算法。聚类有着广泛的应用领域，从理论上说，任何由多个对象组成的数据集都可以根据给定的属性进行聚类。在该章中，我们会介绍论坛帖子的分组，以及如何识别相似的网站用户。同时还介绍了不同类型的聚类算法，以及六种算法的完整实现：单链接（single link）、平均链接（average link）、最小生成

树单链接（minimum spanning tree single link）、k 均值（k-means）、ROCK 和 DBSCAN。

第 5 章介绍分类算法，这是智能应用所不可或缺的组件。该章首先描述了本体（ontology），它包含三个组成部分——概念（concept）、实例（instance）和属性（attribute）。所谓分类，就是将实例赋予最合适的概念。不同的分类器之间的差别就在于它们表示和衡量最优赋值方案的方法。该章简要介绍了分类问题，包括二分类和多分类、统计算法和结构算法。本章还介绍了使用分类器的三个步骤：训练、验证和生产阶段。

第 6 章介绍了分类器的组合——一种可以提升单个分类器准确性的高级技术。该章主要的例子是评价抵押申请的可靠性，同时会详细探讨 bagging 和 boosting 两种技术。另外还介绍了 Breiman 的 arc-x4 boosting 算法的一个实现。

第 7 章举例介绍了这些智能算法在一个新闻门户网站中的应用。我们讨论了其中的技术问题，以及这些智能算法给应用所带来的新的业务价值。例如，聚类算法可以用于新闻的分组，还可以利用新闻之间的相互引用增加新闻的曝光度。在该章中，我们介绍了智能算法的实际应用，勾勒出了将各种智能算法组合在一起实现特定目标的大致框架。

有特色的 TO DO 小节

从第 2 章开始，每一章的最后一节会提供一些引导读者深入学习的内容。作为一个软件工程师，我们发现 To Do 这种形式非常有吸引力：带有祈使的语气，但又不像练习（exercise）那样正式。

有些 To Do 的内容是更加深入地探讨本章的内容，但有些是向读者展示与本章主题有关的其他内容。完成这些任务可以让读者更加深入和广泛地理解智能算法。

本书中所有标注了“TO DO”的代码都可以在各种 IDE 中查看，例如，在 Eclipse IDE 中可以单击 Tasks 面板。单击任何一个任务，都会显示与之相关的部分代码。

谁适合阅读本书

对于想学习在商业上取得巨大成功的算法的软件工程师和 Web 开发人员来说，本书正是为你们而写的。因为本书的源代码是基于 Java 编程语言的，所以本书对 Java 用户可能更具吸引力。尽管如此，使用其他编程语言的读者也能从本书中获益，或许还能将书中的代码转换成其他编程语言的版本。

本书中的例子和思想应用广泛，对于希望从业务角度更好地理解有关技术的技

术经理、产品经理和管理层来说，也有一定的价值。

最后，尽管在本书的标题中有 Web 一词，但本书中的内容也同样适用于其他类型的软件应用，包括移动应用，以及诸如文本编辑器和电子表格一类的传统桌面应用等。

代码约定

本书中所有的源代码都是等宽字体，并且是与上下文分开的。本书中的大部分代码清单都是用于说明代码中的关键概念的，而有些清单有时则是与代码有关的附加信息，某些很长的代码行会有行继续符号。

本书中所有的源代码都可以从 http://code.google.com/p/yooreeka/downloads/list 或出版社的网站 www.manning.com/AlgorithmsoftheIntelligentWeb 中获得。

我们假设读者使用的是微软 Windows 操作系统，否则，读者需要自行修改我们提供的脚本以适用于其他操作系统。将下载的文件解压到 C 盘。压缩文件的顶层目录的名字是 iWeb2，本书中所有的目录都是相对于这个顶层目录的。例如，如果说 data/ch02 目录，指的就是 C:\iWeb\Data\ch02 目录。

解压之后，就可以运行 Ant 构建脚本。很简单，切换到构建目录，然后运行 Ant。无论将文件解压在什么位置，Ant 脚本都能正常运行。现在就可以根据附录 A 的内容来运行 BeanShell 脚本了。

Author Online 论坛

购买本书的同时，读者也获得了免费访问 Manning Publications 论坛的权限，在这个论坛中，读者可以对本书进行评价、咨询技术问题，并从作者或其他读者那里获得帮助。在浏览器中输入 www.manning.com/AlgorithmsoftheIntelligentWeb，就能访问和订阅该论坛的内容，这个页面中说明了读者在注册后如何访问该论坛，可以获得哪些帮助以及论坛的规则，同时还有链接指向本书中例子的源代码、勘误表和其他下载。

Manning 出版社致力于提供一个用户之间以及用户和作者之间的交流平台。对作者参与该论坛的交流并没有强制要求，所有 Author Online 上的贡献都是自愿的（当然也没有报酬）。建议读者尝试问作者一些有挑战性的问题，作者对这样的问题会更有兴趣。

只要本书还在出售，读者就可以在出版商的网站上访问 Author Online 论坛和所有讨论的文档。

关于封面设计

本书的封面设计来自法国的一本服装设计书，即 J. G. St. Saveur 在 1796 年出版的 Encyclopedie des Voyages。旅游在当时还是一个比较新鲜的事物，诸如这样的旅行手册很受欢迎，无论是旅行者还是足不出户的读者，都能从书中了解到世界上其他地方的风土人情，以及法国和欧洲其他地区的特色服饰。

Encyclopedie des Voyages 一书中用丰富的图片生动地展示了 200 年前世界各地的特色。在那个时代，两个人即使是来自两个相隔不过十来英里的地方，也可以轻易通过着装区分出来。不仅如此，在那个时代，通过一个人的服饰还能轻易地判断出这个人的社会地位、行业和种族。

在那以后，不同地区之间服饰的差异逐步缩小。现在，仅仅根据服饰已经很难区分出来自不同大洲的人。或许，乐观地看，我们告别了一个文化和服饰极具特色的世界，换来了多姿多彩的个人生活，或者说得到了更丰富有趣的智能化高科技生活。

本书封面上两个世纪前极具地方特色的服饰就来自这本旅游手册，Manning 出版社以此来庆祝计算机产业的创造力、进取心和其中的乐趣。