Broadview
www.broadview.com.cn

零基础学大数据算法

王宏志　林可　编著

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

零基础学大数据算法

王宏志　林可　编著

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

内 容 简 介

本书是通俗易懂的大数据算法教程。通篇采用师生对话的形式，旨在用通俗的语言、轻松的气氛，帮助读者理解大数据计算领域中的基础算法和思想。

本书由背景篇、理论篇、应用篇和实践篇四部分组成。背景篇介绍大数据、算法、大数据算法等基本概念和背景；理论篇介绍解决大数据问题的亚线性算法、磁盘算法、并行算法、众包算法的基本思想和理论知识；应用篇介绍与大数据问题息息相关的数据挖掘和推荐系统的相关知识；实践篇从实际应用出发，引导读者动手操作，帮助读者通过实际程序和实验验证磁盘算法、并行算法和众包算法。

在讲解每一个大数据问题之前，本书都会介绍大量的经典算法和基础数据结构知识，不仅可以帮助学习过数据结构与算法、算法设计与分析等课程的同学复习，同时能够让入门的“小菜鸟”们，不会因为没有学习过经典算法而对本书望而却步，轻松地掌握大数据算法！

图书在版编目（CIP）数据

零基础学大数据算法 / 王宏志，林可编著. —北京：电子工业出版社，2016.7
ISBN 978-7-121-28937-8

Ⅰ. ①零… Ⅱ. ①王… ②林… Ⅲ. ①数据处理－算法分析 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2016）第 117341 号

策划编辑：张月萍
责任编辑：葛 娜
印 刷：北京京师印务有限公司
装 订：北京京师印务有限公司
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：787×980 1/16 印张：16.75 字数：379 千字
版 次：2016 年 7 月第 1 版
印 次：2016 年 7 月第 1 次印刷
印 数：4000 册 定价：59.00 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

这是一个互联网的时代，也是一个大数据的时代。经常有朋友问起：什么是大数据？大数据是做什么用的？我们为什么要研究大数据？应该怎么研究大数据？在寻找这些问题的答案时，许多朋友找到的内容常常是专业的概念、复杂的公式和难懂的“算法”，这让他们望而却步。很多计算机专业的新生或低年级学生在听到大数据的概念后对其非常好奇，却因没有足够扎实的专业基础知识而无法认识和理解大数据问题，更无法对大数据问题给出很好的解决办法。于是，笔者决定编写一本新生乃至非专业人士也能读懂的大数据算法教程。

本书以一个计算机专业新生小可的口吻，将他内心对大数据的好奇——询问学识渊博的Mr. 王。虽然书中的他不懂数据结构，也不懂经典算法的设计与分析，却在Mr. 王的耐心教导下一一突破了大数据背景下的亚线性算法、磁盘算法、并行算法、众包算法，了解了数据挖掘和推荐算法的基本思想，更是在Mr. 王的指导下完成了各种大数据算法的实现。在所有大数据算法的讲授中，本书无处不渗透着对各种经典算法的回顾，学过的读者可以进行充分的复习，没有学过的读者更是可以借此机会提前掌握各种经典数据结构和经典算法，使得其在今后的学习中事半功倍。这些贯穿全书的前置知识，也使得新生甚至是非专业人士能够通过本书读懂大数据，读懂大数据算法。相信非专业人士也能通过大数据算法的思想，重新认识大数据，并获得一些启迪。

本书前半部分主要以理论知识为主，文中涉及的伪代码、算法等均以简单易懂的自然语言进行了步骤描述和解释，同时给出了小规模运行的例子，使得读者可以轻松地理解。同时，笔者也深知理论与实践相结合的重要性。后半部分包含一些让读者进行实践的实验和程序。这需要具有一些基础程序设计能力，如果读者觉得不能很好地理解它们也不要心急，可以待学会这些语言后，再来尝试。即使并不完全理解这几种语言的语法，也可以按照书中详尽的步骤进行实验，体会成果出现在屏幕上的喜悦，其实尝试之后就会发现，阅读它们并不困难。

虽然本书气氛轻松、语言活泼，但讲授的知识和内容却是非常“专业”的，“算法设计与分析”是计算机学科的核心主题之一，计算机科学中所有问题的解决，都离不开算法设计与分析。本书虽讲解大数据，但无处不紧扣算法设计与分析这一要点，这也让本书带上了浓厚的计算机学科的味道，让读者能够在学习和认识大数据的过程中，学会算法设计与分析，为其他领域知识的学习打下基础。

本书亦算是大数据算法领域的“敲门砖”，本书可以引导读者形成设计大数据算法的思维。在阅读本书之后，读者可以带着设计大数据算法的基本思想去阅读更加深入的专著或论文，进一步的阅读必对大数据算法的学习大有裨益。

本书成书时间仓促，笔者水平亦有限，书中内容、表述、推理等方面的各种不当之处在所难免，敬请各位读者在阅读过程中不吝提出宝贵意见。

目　录

第 1 篇　背景篇

第 2 篇　理论篇

第 3 篇　应用篇

第 4 篇　实践篇

第 1 篇　背景篇

原书缺页

原书缺页

第1章 何谓大数据

1.1 身边的大数据

小可：王老师，那什么是大数据呢？

Mr. 王：你还真是一下就问了个很复杂的问题。其实大数据是一个很模糊的概念，很多学者和学术组织都对其提出过自己的定义，但是至今还没有公认的定义。我们先不谈确切的定义，先来举几个例子说明吧。你平常用社交网络吗？

小可：嗯，是的。

Mr. 王：你有很多好友吧？他们是不是每天都会发很多的状态和消息？

小可：是的，甚至有很多新闻我都是首先通过社交网络知道的。社交网络传递信息的速度真的很快，朋友们每天发布的状态我都看不完，而且不仅有原创的内容，还有很多来自他们好友的转载内容。

Mr. 王：其实社交网络上的这些信息就是一种典型的大数据。

小可惊讶地说：原来这就已经是大数据了？我一直以为大数据都在实验室里面呢。

Mr. 王：此言差矣，其实大数据就在我们身边。我们常用的社交网络上就有着非常巨大的信息量，虽然一个人发布的状态非常有限，但由于使用的人数众多，加之转载和评论，巨大的数据规模就使得社交网络信息无法在短时间内由人工或者由少量的几台计算机存储和管理。站在社交网络之外看待它，就会发现里面有很多且杂乱无章的信息和内容，同时其规模非常大。这就是大数据的一个典型例子。

小可恍然大悟地说道：哦，原来这就是大数据啊，那其实我每天都在接触大数据。

Mr. 王笑道：的确，大数据就在我们每个人的身边，随着信息时代的到来，我们每个人每天接触到的数据量都是非常大的。但你在查看这些消息的时候，有没有看到除字面内容以外的东西呢？

小可想了一下，说：好像没有什么，我关注的只是消息本身。

Mr. 王：我们研究大数据不只是能知道它的数据量很大，或者说仅仅研究如何把它们存储起来，我们还要发掘在大数据中隐藏的知识和有价值的信息。

小可：哦？大数据中隐藏着知识？

Mr. 王：是的，从表面上看，大数据可能只是一些简单的文本、杂乱的符号或者是一些数字的序列或者集合，但是从这些文本或者数字的背后，我们可以发掘其作为一个群体所具有的一些性质，从而发现一些对我们有意义、有价值的信息，所以我们才要研究大数据。

小可：大数据不是很大很大吗？那么我们研究它不就会变得很困难吗？

Mr. 王：不错，大数据的量很大很大，我们单单是把其中的信息逐个地访问一遍都很困难，所以发掘其中的知识就更加困难了，这就是研究大数据要解决的重要问题，也就需要我们这些研究大数据的人、热爱大数据的人加倍地努力了。

小可思考片刻后，说：那在超市里面，每年都会有很多人去买东西，他们的购物单上又会包含着很多内容，对超市来说，这些购物的记录就是“大数据”吧？而通过分析这些购物单，发现顾客更喜欢买哪些商品，这算不算一种通过大数据分析出的知识呢？

Mr. 王：很聪明嘛，你举了一个很好的例子。商业数据也是大数据的一个重要体现，超市购物的明细记录、公司运营的详细账目这些数据量都是很大的，处理起来非常费时费力，而其中又包含着有价值的信息，通过这些信息不仅可以分析出本年度公司的运营情况，同时可以指导下一年度公司的营销战略，这些数据对公司来说可谓是价值连城。

小可：那么大数据在别的方面又有哪些体现呢？

Mr. 王：你应该对生物遗传有所了解吧。

小可点点头道：是的，人体通过 DNA 携带遗传信息。

Mr. 王：在医疗和生物计算领域中，每次对 DNA 序列的分析都会产生大量的数据，这个数据量已经不是用 GB 可以衡量的了，甚至要达到 PB 级别或者更大。而这么大的数据，不仅计算机的内存装不下，而且一般计算机的硬盘都已经存不下了。即使是扫描一遍，在上面发现一个小序列都需要一些时间，在这些数据上面做分析将是一件更困难的事情。这也是一种大数据。

不仅在生物学中如此，而且在很多科学仪器的使用过程中也都会产生大量的数据，比如天文观测、显微观测，现在逐渐应用的传感器和传感器网络在使用过程中都会记录下大量的数据。这些仪器不停地记录下的数据，都涉及如何存储、如何分析研究的问题，这些都是大数据。

生活中的大数据

小可：嗯。

Mr. 王：那我们就给大数据下个定义吧。

定义 1：所涉及的数据量规模巨大到无法通过人工，在合理时间内达到截取、管理、处理并整理成为人类所能解读的信息。（Dan Kusnetzky, What is "Big Data"？）

定义 2：不用随机分析法（抽样调查）这样的捷径，而采用所有数据的方法。（维克托·迈尔 - 舍恩伯格、肯尼斯·库克耶，"大数据时代"）

定义 3："大数据"是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。（"大数据"（Big Data）研究机构 Gartner）

有了前面的那些例子，这些定义是不是相对好理解一些呢？

小可：嗯，我懂了。

1.2 大数据的特点和应用

Mr. 王：大数据具有较大的数据量，和一般的数据相比，其具有如下一些特点。

- 在数据量上，大数据是通过各种设备产生的海量数据，其数据规模极为庞大，远大于目前互联网上的信息流量，PB 级别将是大数据的常态。
- 在多样性上，大数据种类繁多，在编码方式、数据格式、应用特征等多个方面存在差异性，多信息源并发形成大量的异构数据。
- 在速度上，涉及感知、传输、决策、控制开放式循环的大数据，对数据实时处理有着极高的要求，通过传统数据库查询方式得到的“当前结果”很可能已经没有价值。
- 在价值上，数据持续到达，并且只有在特定时间和空间中才有意义。

Mr. 王：我们分析大数据、研究大数据，是希望能够利用它们获得我们需要的知识。我们可以利用大数据进行：

- 预测
- 推荐
- 商业情报分析
- 科学研究

等发现大数据中的价值，使用大数据、利用大数据的过程。由此可知，对大数据的研究还是非常重要而有意义的。

小可：有种大数据中有黄金的感觉啊。

Mr. 王：正是如此，从大数据中挖掘出来的价值，真是难以估量啊。今天时间不早了，你先回去吧，下节课咱们讨论一下关于算法的问题，要讨论大数据算法，必须先了解算法的相关知识。

小可：谢谢老师，那我下次再来。

第2章 何谓算法

2.1 算法的定义

小可：王老师您好，了解了什么是大数据，今天我来听听关于算法的内容。

Mr. 王：你好，想要学懂大数据算法，算法的基础知识是一定要了解的，你必须要知道如何设计和分析算法，我们才能谈及如何在大数据集合上研究算法。

小可：我经常会听到“算法”这个词，计算机专业的同学总会提到，那么到底什么是算法呢？

Mr. 王：至今为止，算法都没有一个准确的定义。每个计算机科学家和工程师都在设计算法、使用算法、分析算法、实现算法，但对于算法的定义依然是众说纷纭，很多书籍都曾经给出自己的定义，与其说那是一个定义，不如说都是对算法的一种诠释。的确，算法是一个很模糊、抽象的概念。

小可：那么我该如何搞懂什么是算法呢？

Mr. 王：在解释“算法”这个概念之前，我们首先来谈谈，一个计算机科学家是如何解决问题的。我先问问你，计算机是用来做什么的呢？

小可：计算、办公、游戏、影音娱乐，还有上网。

Mr. 王：不错，宽泛地谈起计算机的用途真是数不胜数。不过总结起来，其实计算机做的事情就一个——解决问题。

小可：解决问题？

Mr. 王：对，生活中有很多问题，其中有些问题人工解决起来很费时费力，于是我们发明了许多工具，在这个层面上，其实计算机也是一种工具，本质上它就是解决问题的一种工具。

比如：

- 升空卫星的轨道是怎样的?
- 从哈尔滨到深圳，走怎样的一条路线最短、最省路费?
- 模拟一次比赛的结果，分析到底谁的胜算更大?
- 我们希望知道，某个游戏或者博弈中是不是有必胜的策略?

小可：嗯，其中有些问题人工解决起来确实很费劲。那么计算机科学家又是如何解决这些问题的呢?

Mr. 王：首先，如果希望计算机能真正地解决一个实际问题，我们先要将现实世界中的事物转化为模型，这个模型可以被计算机理解和处理，它可以表示成数据和指令等。这个过程我们称之为**建立模型**。在此过程中，我们需要把一个实际问题抽象成计算机可以理解的语言，或者说计算机可以理解的问题，才可以用计算机求解。

小可：哦，这就是所谓的“建模”吧。

Mr. 王：其次，我们要知道这个问题是不是可计算的。计算机可以解决很多问题，也有很多问题解决不了。那么，由此诞生的，研究一个问题是不是计算机可计算的、可解的计算机科学分支叫作**可计算理论**。

小可：还有计算机解决不了的问题吗?

Mr. 王：当然，比如著名的“停机问题”。在这方面做出卓越贡献的科学家是非常著名的阿兰·图灵。图灵曾经提出过很多对计算机科学产生深远影响的理论，直到现在，我们使用的电

子计算机在模型上依然可以称之为“图灵机”。停机问题在很多资料中也称作“图灵停机问题”。图灵已经进行了证明，停机问题是不可计算的。

小可：那是不是说，我的计算机内存太小、CPU 太慢，有些特别大型的问题在我这里就“计算不出来”，这就是一个不可计算的问题呢？

Mr. 王：不，这是不对的，不可计算的问题并不是出于 CPU 速度和内存大小等资源的限制而无法在一定的时间内完成，而是不论给计算机多大的内存、给它多快的 CPU 都是无法求解的。如果你的计算机 CPU 太慢、内存太小，在一台内存更大、CPU 更快的机器上，还是能够求解的，那么这样的问题不是一个不可计算的问题。

不过这里有一个问题：某个问题虽然是可计算的，但是对于这个问题我们急着要的结果要算几年时间，那么这个问题恐怕也“相当于”解决不了，或者说交给计算机解决已经没有什么意义了。所以我们必须要知道的一件事情是，这个问题能不能用计算机在我们可以接受的时间或者空间界限内解决。研究某个问题被计算机解决的时空下界限的计算机科学分支称为**计算复杂性理论**。计算复杂性理论主要研究的是某个问题可以被计算机求解的时间和空间下界限。它研究给出的问题的时空下界限，是任何算法都无法突破的，研究比下界限更快的算法是没有意义的，当发现某一个算法已经足够快到可以和计算复杂性理论中得出的下界限是同一个级别时，我们就不必再去提升它的效率了。同时，研究这种时空下界限有另一方面的目的，就是可以用来评价我们现在设计出来的算法与这个问题可以被解决的极限时间空间界限还有多远的距离，很多时候我们设计出来的算法还不足以达到这个极限界限，但有了这个界限，可以给我们接下来的研究指明一个方向。

总结起来，可计算理论和计算复杂性理论都是一个研究“问题”的范畴。

研究过问题之后，我们要考虑的就是如何去解决问题。这也是计算机作为一种工具的重要属性。想要解决问题，就需要我们设计算法。

小可：那么到底什么是算法呢？

Mr. 王：这里我们还是举个生活化的例子吧。比如，我们现在想要煮一锅汤，这就是一个问题。根据生活经验，我们认为它是可解的（可计算分析），也是理论上可以在我们接受的时间范围内解决的（计算复杂性分析）。这时，需要我们设计一个解决它的方法或者说一系列步骤。

小可：我想想，煮汤我们可以想到的步骤就是洗菜、切菜、烧水、煮汤、出锅。哈哈。

煮汤的“算法”

Mr. 王：很好，你已经在设计一个算法了。

小可吃惊地说：啊？！这就是一个算法了？