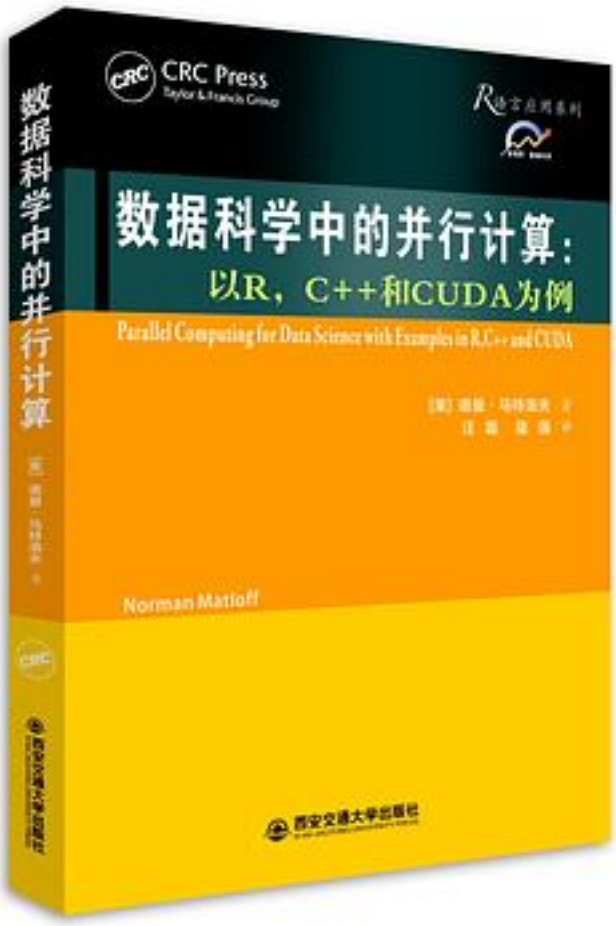
数据科学中的并行计算



数据科学中的并行计算_下载链接1_

著者:[美]诺曼·马特洛夫

出版者:西安交通大学出版社

出版时间:2017-12-12

装帧:平装

isbn:9787560599588

数据科学家的并行计算必读手册

“……一本完整、易读的并行计算入门——它适合很多学科的研究人员和学生使用。这是一本‘必备’的参考书……”

——戴维·E·吉尔斯，维多利亚大学

“这本书我会既用来当参考书，又当教材。书中的例子生动，内容也使读者直接从概念走向可用于工作的代码。”

——迈克尔·凯恩，耶鲁大学

本书是第一本并行计算领域中，注意力完全集中在并行数据结构、算法、软件工具及数据科学中具体应用的书。书中的例子不仅有经典的“n 个样本，p
个变量”的矩阵形式，还有时间序列、网络图模型，以及各种其他的在数据科学中常见的结构。本书同时也讨论了适用于多种硬件、多种编程语言的的软件包。

特点

关注数据科学中的应用，包括统计学、数据挖掘和机器学习。

讨论了数据科学中的常见数据结构，如网络图模型。

通篇强调了普遍的原理，如避免降低并行程序速度的因素。

覆盖了主流的计算平台：多核、集群以及图像处理单元（GPU）。

解释了 Thrust
包如何降低多核机器和GPU编程的难度，并使得同一份代码能够在不同的平台上工作。

在作者网站上提供了样例代码。

译者序

21
世纪的第二个十年，随着计算能力的巨大提升和移动互联网的迅猛发展，大数据时代拉开了它的帷幕。大数据时代的显著特点就是数据量大，对数据处理的速度和时效提出了苛刻的要求。

在传统的串行计算下，多核的计算机/集群等只有一个内核能够进行有效的工作，这就造成了计算性能的浪费。并行计算概念的提出，则解决了这个性能浪费的问题。它能够协调多个内核共同计算，极大地提升了计算速度，从而满足了大数据时代人们对高速处理数据的需求。

Norman Matloff
教授在加州大学戴维斯分校教授计算机科学，对计算机架构和算法了然于心。更值得一提的是，他还是该校统计系的创始人之一，不但教授本科的统计课程，还在统计系硕士和博士的考试委员会担任多个职务。对于统计理论的熟悉，使得他在使用计算机编程处理统计问题的时候，更加得心应手。该书即是他在并行计算方向上多年经验的总结。

本书不是一本并行计算的理论教材。该书别出心裁，使用实例手把手地教会读者掌握并行计算的基本概念和操作。在提纲挈领地介绍了如何在R
中使用并行方法之后，作者带领我们学习了多线程和多进程，以及并行调度等方面的知识和技能。随后，作者用详尽的篇幅讲述了如何使用R、C++
和CUDA分别来进行共享内存范式编程和消息传递范式编程。本书在讲述了当前流行的MapReduce

之后，又详细讲解了如何并行地实现串行计算下所对应的排序、扫描、矩阵乘法等经典算法。在本书的最后，作者讲述了如何使用并行计算来进行统计。此外，本书的附录中对线性代数、R 和C 也做了简明的介绍，方便不熟悉的读者迅速入门。

值得一提的是，Matloff
教授的汉语也非常熟练，在本书的翻译过程中，他也给出了相应的建议和意见。编辑李颖为本书的编辑工作提出了不少中肯的建议和意见，并为本书的顺利出版做出了巨大的努力。在这里对他们一并表示感谢。

本书两位译者的协作，跨越了大洋和时差。翻译的日子中酸甜苦乐，都化作段子互相慰藉。不禁让人想到，人的命运啊，当然要靠自我奋斗，但是也要考虑到历史的行程。时代带给我们的，永远值得珍惜。

译者于北京

2017 年8 月

前言

感谢你对本书感兴趣。我很享受写书的过程，也希望这本书对你非常有用。为达此目的，这里有几点事情我希望说清楚。

本书目标：

我很希望这本书能充分体现它标题的含义数据科学中的并行计算。和我所知道的其他并行计算的书籍不同，这本书里你不会碰到任何一个求解偏微分方程或其他物理学上的应用。这本书真的是为了数据科学所写无论你怎样定义数据科学，是统计学、数据挖掘、机器学习、模式识别、数据分析或其他的内容。

这不仅仅意味着书里的实例包括了从数据科学领域中选取的应用，这也意味着能够反映这一主题的数据结构、算法和其他内容。从经典的“n 个观测，p 个变量”
的矩阵形式，到时间序列，到网络图模型和其他数据科学中常见的结构都会囊括其中。

本书包含了大量实例，以用于强调普遍的原理。因此，在第1
章介绍了入门的代码实例后（没有配套的实例，这些普遍的原理也就没有任何意义），我决定在第2
章里解释可以影响并行计算速度的一般因素，而不是集中介绍如何写并行代码。这是一个至关重要的章节，在后续的章节中会经常提到它。事实上，你可以把整本书看成如何解决第2 章开头所描述的那个可怜的家伙的困境：

这里有一个很常见的情景：一个分析师拿到了一台崭新的多核机器，这台机器能做各种神奇的事情。带着激动的心情，他在这台新机器上写代码求解他最喜欢的大规模的问题，却发现并行版本的运行速度比串行的还慢。太令人失望了！现在让我来看看究竟什么因素导致了这种情形……

本书标题里的计算一词反映了本书的重点真的是在计算上。这和诸如以Hadoop
为代表的分布式文件存储等的并行数据处理不同，尽管我还是为相关话题专门写了一个章节。

本书主要涵盖的计算平台是多核平台、集群和GPU。另外，对Thrust
也有相当程度的介绍。Thrust 极大地简化了在多核机器和GPU
上的编程任务，并且同样的代码在两种平台上都可以运行。我相信读者会发现这部分材料非常有价值。

需要指出一点，这本书不是一本用户手册。尽管书中使用了诸如R 的parallel和Rmpi 扩展包、OpenMP、CUDA
等特定工具，但这么做仅仅是为了让问题具体化。本书会给读者带来有关这些工具的非常扎实的入门介绍，但不会提供诸如不同函数的参数、环境选项等内容。本书的目的是，希望读者阅读完本书后，为进一步学习这些工具打下良好基础，更重要的是，读者今后可以使用多种语言编写高效的并行代码，无论是Python、Julia，还是任何其他语言。

必要的背景知识：

如果你认为你已经可以相对熟练地使用R，那本书的大多数内容你应该都可以读懂。在一些章节里，我们需要使用C/C++，如果你想仔细阅读学习相关章节，需要具备相关的背景知识。然而，即使你不怎么了解C/C++，你也应该会发现这些章节很容易读懂，并且相当有价值。附录里包括了针对C 程序员的R 简介和针对R 用户的C 语言简介。

你需要熟悉基础的矩阵运算，主要是相乘和相加。有时我们也会使用一些更高级的运算，比如求逆（以及与之相关的QR 分解）和对角化。这些内容在附录A 中有涉及。

机器设备：

除了特别说明的地方，本书中所有的计时实例都运行在一台16
核允许两个超线程的Ubuntu 机器上。我一般使用2 到24
个核，这应该和多数读者可以使用的平台类似。希望读者可以使用4 到16
核的多核系统，或者一个有几十个节点的集群。但即使你只有一个双核机器，应该仍会发现本书的材料非常有用。

对于那些少数有幸可以使用拥有几千个内核的集群的读者，书中的内容仍然适用。依据本书中对这种系统的观点，那个著名问题“这可扩展吗？” 的答案一般是否定的。

CRAN 扩展包和代码：

本书使用了我在CRAN，即R
的软件贡献库（http://cran.r-project.org）上的几个扩展包：Rdsm、partools 和matpow。

本书的示例代码都可以从作者的网站下载，
http://heather.cs.ucdavis.edu/pardatasci.html。

作者介绍:

Matloff
博士出生于洛杉矶，在东洛杉矶和圣盖博谷两个地方长大。他在加州大学洛杉矶分校获得了纯粹数学的博士学位，学术研究方向为概率论和统计。他在计算机科学和统计学方向发表了大量论文，现在的研究方向是并行处理、统计计算和回归方法。他也是Journal of Statistical Software 的编委之一。

Matloff 教授曾是国际信息处理联合会11.3
工作组的成员，该组织是联合国教科文组织（UNESCO）下设的一个数据库软件安全国际委员会。他也是加州大学戴维斯分校统计系的创始人之一，并参与了该校计算机科学系的建立。他在戴维斯分校被授予了杰出教学奖和杰出公众服务奖。

目录: 第1章 R语言中的并行处理入门

数据科学中的并行计算_下载链接1_

标签

R

并行计算

评论

数据科学中的并行计算_下载链接1_

书评

数据科学中的并行计算_下载链接1_