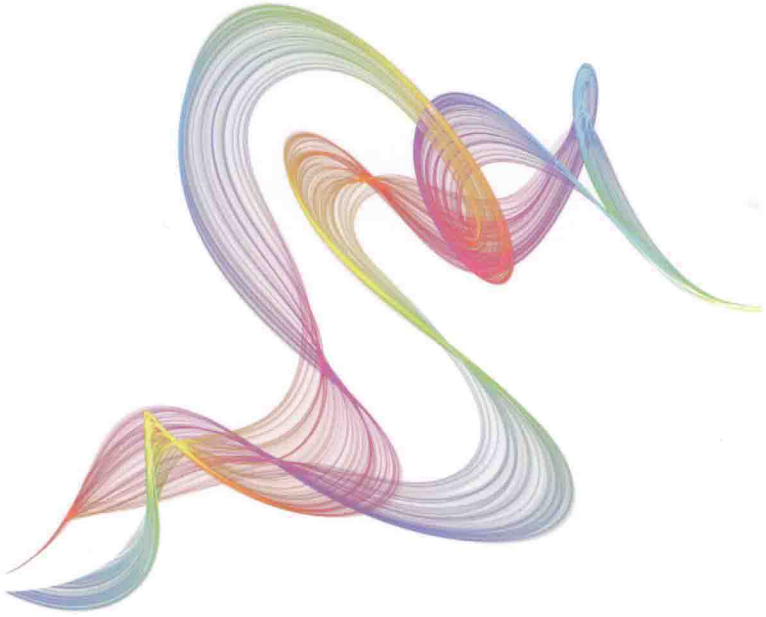
HZ BOOKS
华章科技

[PACKT]
PUBLISHING

内容全面而深入，既展示Mahout的强大功能，又全方位讲解利用Mahout进行大数据分类、聚类和预测分析的各种技术细节、方法和最佳实践

实战性强，包含丰富案例，涉及Mahout开发环境、序列文件使用方式、整合Mahout和外部资源、实现朴素贝叶斯分类器、股市预测、顶棚聚类、频谱预测、K-均值聚类等

大数据技术丛书



Apache Mahout Cookbook

Mahout实践指南

（美） Piero Giacomelli 著

靳小波◎译

机械工业出版社
China Machine Press

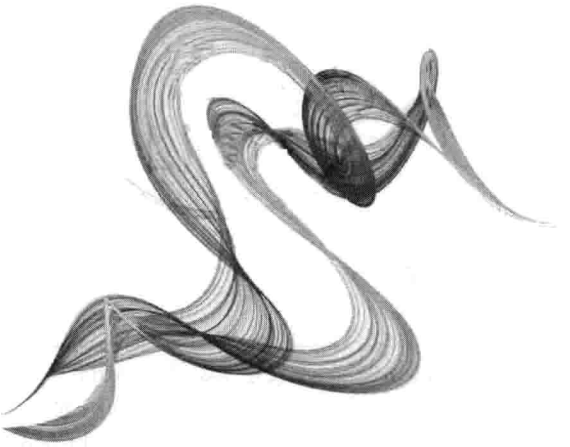
技术丛书

Apache Mahout Cookbook

Mahout实践指南

（美） Piero Giacomelli 著

靳小波◎译



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Mahout 实践指南 /（美）贾科梅利（Giacomelli，P.）著；靳小波译．—北京：机械工业出版社，2014.6
（大数据技术丛书）
书名原文：Apache Mahout Cookbook

ISBN 978-7-111-46714-4

I. M… II. ① 贾… ② 靳… III. ① 机器学习 ② 电子计算机－算法理论 IV. ① TP181 ② TP301.6

中国版本图书馆 CIP 数据核字（2014）第 100085 号

本书版权登记号：图字：01-2014-1083

Mahout 实践指南

[美] Piero Giacomelli 著

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：秦 健 责任校对：殷 虹
印 刷：北京市荣盛彩色印刷有限公司 版 次：2014 年 6 月第 1 版第 1 次印刷
开 本：186mm×240mm 1/16 印 张：12
书 号：ISBN 978-7-111-46714-4 定 价：49.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzjsj@hzbook.com

华章计算机
HZBOOKS | Computer Science and Technology

译 者 序

机器学习是人工智能领域里的一个重要分支，是进行复杂数据分析和构建智能系统的一个十分重要的研究方向。互联网数据不断地爆炸性增长标志着大数据时代的来临，机器学习领域在处理大规模数据时将面临新的挑战。研究者除了从软件方面发明新的时间复杂度更低的可扩展的算法之外，还应积极地从硬件架构方面进行改进，其中值得关注的方向有两个：将并行计算和分布式计算引入机器学习。

并行计算当前的热门方向是GPU计算，将传统上同时运行在多台机器上的任务交给单台机器上的图形处理器处理，这使得并行计算的费用大大降低。Shane Cook撰写的《CUDA并行程序设计：GPU编程指南》[⊖]是这方面的经典参考书。GPU的技巧已经大规模地应用到机器学习领域以改进传统的算法，两个有代表性的GPU机器学习库是Theano[⊜]和GPUMLib[⊜]。

分布式方面最有代表性的工作是Apache Hadoop。它支持在大型集群中运行应用程序。最为重要的是，该架构是Java语言编写的开源软件框架，它实现了Google的Map/Reduce框架，可供商业或科研免费使用。Mahout库就是在这样的背景下产生的。它建立在Hadoop的基础上，主要用于处理大规模的机器学习问题，其中核心算法有聚类、分类、协同过滤。同样，该库是开源免费的，且支持商业级别的机器学习方面的应用。

针对从事机器学习应用方面的开发人员以及机器学习理论研究方面的科研人员使用Mahout，本书提供了非常有价值的参考。作者在将Mahout用于商业领域方面经验丰富，本书旨在降低Mahout初学者的入门门槛。本书特点如下：

- 通过分析大量的实例，展示了如何更好地使用Mahout算法，主要有分类算法、聚类算法以及遗传算法。
- 由浅入深讲解实例，帮助读者逐步掌握Mahout的应用方法。
- 图文并茂，让读者及时了解每一步操作之后的效果，帮助读者更好地检验学习进度。
- 写作方式独特，通过编码的方式帮助读者了解代码的目标及含义，避开代码背后复杂的机理。

⊖ 本书已由机械工业出版社引进出版，ISBN：978-7-111-44861-7。——编辑注

⊜ 参见http://deeplearning.net/software/theano/。——译者注

⊜ 参见http://gpumlib.sourceforge.net/。——译者注

❑ 避开烦琐的数学表述，通过具体而形象的描述，让读者直观了解机器学习技术。

值得一提的是，Mahout 主要用在 Linux 平台上，但是对于使用 Windows 系统的大部分读者来说，这并不是一个障碍。本书通过详尽的描述，让不熟悉 Linux 的读者也可以学到 Linux 的基本使用技巧。实际上，本书中所有的代码都是在 Windows 系统下编写的，作者通过在 Windows 上安装 Virtual Box 软件来使用 Linux 平台，这种方式为那些 Windows 系统下的开发者使用 Mahout 库提供了一个良好的建议。

为了方便读者正确、迅速地理解本书，译者对本书的一些错误进行了修正，并在某些表述不太清楚的地方添加了注释，希望对读者理解本书内容有所帮助。然而，不得不承认，尽管译者从事的研究方向是机器学习，但由于水平有限，本书难免存在错误。欢迎读者及时向出版社指出，便于再版时予以更正。

特别感谢机械工业出版社编辑为本书出版所付出的辛勤劳动。

最后，感谢夫人靖莹以及耿光刚博士在文字校稿方面给予的支持和帮助。

靳小波

前　言

在最近的 10 年，社会化网络的出现和移动设备的发明极大地改变了我们处理数据的方式。

为了帮助你了解究竟发生了什么，我们不得不提到在 2012 通过 Qmee 做的一项研究：在 60 秒里展示互联网上经常发生的事情。结果参考 http://blog.qmee.com/qmee-online-in-60-seconds/，它告诉我们在过去的每一秒里 Twitter 收到 278 000 条推文（tweet），Facebook 收到 41 000 条 post，YouTube 已上传了时长达 72 小时的视频。这些算是很大的网站了，但是即使是具有国家或国际背景的网站，因收集网站的日志而拥有上百万的记录也是很平常的。

为管理如此海量的信息，需要编写新框架来实现不同机器之间的计算任务共享。Hadoop 是 Apache 编写的算法解决方案，它可以将计算任务分配到不同的硬件架构上运行。

当你需要分析上亿条数据记录时，在大多数情况下，你的目的是通过信息提取来发现数据之间的新关系。传统上，数据挖掘算法就是为这个目的发展起来的。然而，当处理非常大的数据集时，无法在一个合理的时间里实现数据挖掘任务。Mahout 是一个数据挖掘框架，可以和 Hadoop 一起应用数据挖掘算法处理大规模数据集上的数据挖掘任务，它使用了封装在 Hadoop 里面的 MapReduce 例程。所以，Mahout 通过 Hadoop 架构的这个底层接口为编程人员进行数据挖掘任务提供了一个易用的框架。

本书将通过一些实例向你展示怎么使用 Mahout 进行数据挖掘，以及数据挖掘的不同方法。最关键的是，使用一种简洁通俗的方法向你展示使用 Mahout 对数据进行分类、聚类和预测的方式。本书是面向编程的，所以我们并不想在步骤中过多地引入其理论背景，但是我们会给有能力的读者一些参考文献以进行更深一步研究。当写这本书时，我们面临的主要挑战是：

- 根据我的经验，Mahout 有着非常高的学习曲线，主要原因是算法使用了 MapReduce 方法，该方法不同于序列方法。
- 数据挖掘算法本身就不太容易理解，它们在某些情况下需要一些特殊的技能，而开发人员不一定具备这样的技能。

于是我们尽力使用一种面向源码的方式让读者能抓住每一段代码的含义和目的，但是又不需要深入理解其背后的实现机制。

这种方法的效果由你来判断，而且我们希望你在阅读的时候能够发现一些乐趣，就像我

们在写作的时候获得的一样。

本书的组织

第 1 章描述如何在单台机器上创建一个完整的开发环境。通过编写一个推荐算法使得数据挖掘操作的所有代码片段均以 Hadoop 的方式呈现（包括引入的 JAR 库等），这非常清晰地展现在没有任何背景的读者面前。

第 2 章介绍序列文件。当使用 Hadoop 和 Mahout 时，序列文件是个比较关键的概念。在多数情况下，Mahout 并不直接操作要使用的数据集，所以在没有编码算法之前，我们需要描述如何对待这些特别的文件。

第 3 章详细介绍使用命令行工具和代码从 RDBMS 中读写数据。

第 4 章详细介绍如何使用朴素贝叶斯分类器分类文本文档。全面地描述如何将文档单词转化为包括单词出现次数的向量，并展示如何使用 Java 编写朴素贝叶斯分类器和互补朴素贝叶斯分类器。

第 5 章主要涉及两个算法：logistic 回归和随机森林（Random Forests）。它们展示了通过分析某些普通数据就可能预测其未来值。

第 6 章描述 Mahout 框架中最常用的算法，其中包括大数据的聚类分析和分类任务。在这一章，通过一些实例介绍使用顶棚聚类围绕聚类中心聚合数据向量。

第 7 章继续介绍 Mahout 中的聚类分析算法。该章描述了频谱聚类的使用方式，它在对图形式的链接信息进行分类时是非常有效的。

第 8 章描述了使用 K- 均值聚类（包括序列方式和 MapReduce 方式）对主题中的文本文档进行分类。我们将通过命令行方式和 Java 编码的方式解释如何使用该算法。

第 9 章介绍一个比较老的，称为频繁模式挖掘（Frequent Pattern Mining）的算法。该算法通过过去顾客的购买情况来预测哪些东西应该放在一块出售。Latent Dirichlet 算法将用于文本分类。

第 10 章描述了如何在 Mahout 中使用遗传算法解决旅行商（Travelling Salesman）问题和提取规则。我们将会看到如何使用 Mahout 的不同版本来使用这些算法。

阅读本书你需要什么

在第 1 章中，我们将介绍本书需要的所有软件。本书中所有的例子均在 Ubuntu 10.04 简易发行版和 Oracle 公司的 Virtual Box 平台上编程实现。

本书的读者

本书对希望以一种新颖、快速的方式来入门 Mahout 的开发人员是比较理想的。阅读本书不需要了解 Mahout，有经验的开发人员或系统管理人员也可以从本书中受益。

下载示例代码

你可以通过你的账号（在网站 http://www.packtpub.com）下载在 Packt 上购买的书籍的所有示例代码。如果你在别的地方购买本书，可以访问 http://www.packtpub.com/support 并注册，我们将会通过电子邮件直接把文件发送给你。

勘误表

尽管我们尽了最大的努力来确保内容的准确性，但错误在所难免。如果你在书中找到错误，无论在文中或代码中，我们会非常感谢你给我们报告了这些错误。你这么做，可以避免其他读者遭受困扰，并且我们将在随后的再版中改进。如果你发现任何错误，请通过访问 http://www.packtpub.com/submit-errata 来报告它们：选择你所购买的书名，点击“errata submission form”链接，输入错误的详细细节。一旦你报告的错误得到确认，你的提交将会被接受，并且勘误表将会在我们的网站上更新或者加入已经存在的勘误列表中。通过网址 http://www.packtpub.com/support 选择书号，你可以看到现有的勘误条目。

关于评阅者

Nicolas Gapaillard 是 Java 架构方面的一位热情的自由撰稿人，他了解 Java 和开源领域中的创新项目。

他曾在开源软件公司 Linagora（http://www.linagora.com）的证券部门从事开发工作，由此开始他的职业生涯。该部门旨在开发一个围绕交易安全的开源软件，包括证书管理、密钥文档的存储和认证机制。

之后，他在 Smile 开源软件整合项目中任职 Java 技术方面的开发人员、培训人员和技术领导者。

有了上述从业经历之后，他决定创办自己的公司（名为 BIGAP,http://bigap.fr），该公司主要做自由撰稿业务，这使得他有更多的时间来学习和研究创新项目。

其中有一个业务是为名为 Onecub 的法国公司实现根据顾客的类别自动分类电子商务方面的电子邮件。当时，仅仅 Mahout 可以提供“拿来即用”的算法来解决这些问题。从那以后，Nicolas 开始深入地研究 Mahout 项目和数据挖掘领域。

某一天，Packt 出版社看到他撰写的文章（http://nigap.blogspot.fr）并邀请他为该书撰写评论，他非常愉快地接受了这项任务。

我非常感谢本书作者为保证书的质量而做出的努力，我也感谢 Packt 出版社，他们信任我，让我撰写该书的评论，他们非常仔细地管理整个流程，并且允许我评论该书的修订。我还想感谢其他的评论人为本书的修订和内容的质量而提供的帮助，感谢我的妻子让我有自由的时间来写这些评论。

Vignesh Prajapati 是 Pingax 公司大数据方面的科学家。他热爱开源技术（比如 R 语言、Hadoop、MongoDB 和 Java 语言），主要工作就是使用机器学习、R 语言、RHadoop 和 MongoDB 进行数据分析。他在多个算法方面是专家，例如数据 ETL、电子商务、Google 历史分析和其他数据集的生成推荐、分析和行为定位等。他也撰写了几篇文章来阐述使用 R 语言、Hadoop 和机器学习实现高效的智能大数据应用。他的联系方式是 vignesh2066@gmail.com 或 http://in.linkedin.com/in/vigneshprajapati/。

除了本书以外，Packt 还有两本书与他有关：他是《Big Data Analytics with R and

Hadoop》一书的作者（Packt 出版，https://www.packtpub.com/big-data-analytics-with-r-and-hadoop/book），他也为《Data Manipulation with R》一书（作者：DeMystified，Packt 出版）撰写评论。

我感谢 Packt 出版社提供的这个难得的机会，感谢我的家庭、朋友和 Packt 出版团队激励和支持我为开源技术贡献自己的力量。

Shannon Quinn 将在 Carnegie Mellon 大学（匹兹堡）攻读计算生物方向的博士学位。他的研究兴趣是和他的导师 Chakra Chennubhotla 博士一起将谱图理论、机器视觉和模式识别用于生物图像识别，为生物监控构建实时分布式的框架。他还参与了 Apache Mahout 和其他开源项目的开发工作。

致　　谢

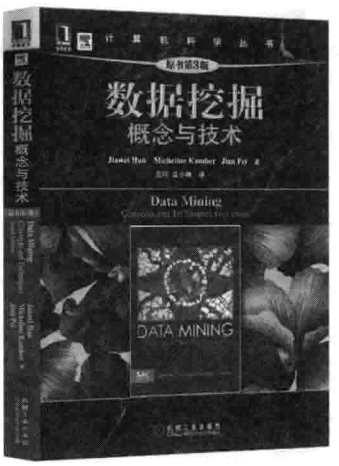
感谢我的家庭在我写书的最为紧张和激动人心的几个月里给予的支持。

感谢我的妻子 Michela，她每天激励我成为一个更好的人，而我母亲 Milena 在我结婚前就已经这样做了。另外，要感谢 Lia 和 Roberto，每次只要我们有所求，他们都会尽力帮助。

还要特别感谢整个 Packt 出版团队，感谢 Gaurav Thingalaya、Amit Singh、Venu Manthena、Shiksha Chaturvedi、Llewellyn F. Rozario、Amey Varangaonkar、Angel Jathanna 和 Abhijit Suvarna，他们对我给予了足够的耐心，尽管他们没有特别的原因要这么做。

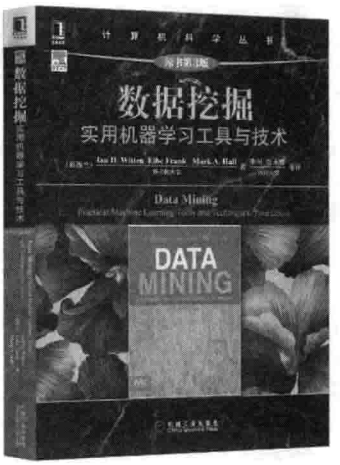
最后，当我写这本书的时候，我找到一个新工作，这要感谢 Giuliano Bedeschi。他和他的两个儿子 Giovanni 和 Edoardo 一手创办了 SPAC 公司，这是我值得为之骄傲的为数不多的公司之一。在这个过渡期间，SPAC 公司的同事们真正地给予了我很大的帮助。

推荐阅读



数据挖掘：概念与技术（原书第3版）

作者：Jiawei Han 等 ISBN：978-7-111-39140-1 定价：79.00元



数据挖掘：实用机器学习工具与技术（原书第3版）

作者：Ian H. Witten 等 ISBN：978-7-111-45381-9 定价：79.00元

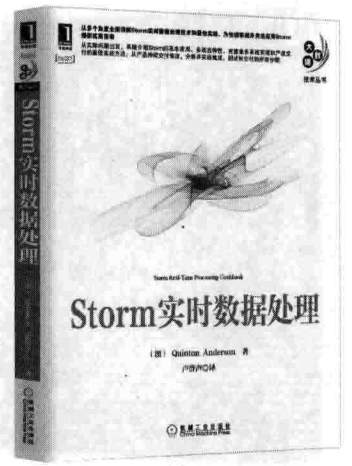
大数据管理：数据集成的技术、方法与最佳实践

作者：April Reeve ISBN：978-7-111-45905-7 定价：59.00元

大规模分布式系统架构与设计实战

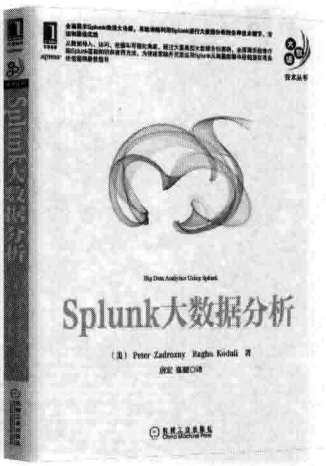
作者：彭渊 ISBN：978-7-111-45503-5 定价：59.00元

推荐阅读



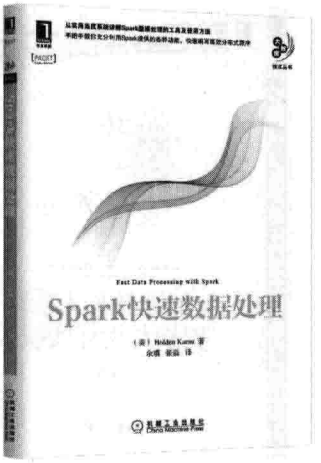
Storm实时数据处理

作者：Quinton Anderson ISBN：978-7-111-46663-5 定价：49.00元



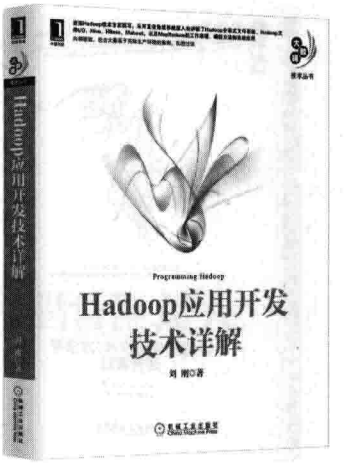
Splunk大数据分析

作者：Peter Zadrozny 等 ISBN：978-7-111-46429-7 定价：69.00元



Spark快速数据处理

作者：Holden Karau ISBN：978-7-111-46311-5 定价：29.00元



Hadoop应用开发技术详解

作者：刘刚 ISBN：978-7-111-45244-7 定价：79.00元

目　录