TURING 图灵程序设计丛书

MANNING

Mahout in Action

Mahout 实战

[美] Sean Owen Robin Anil
Ted Dunning Ellen Friedman 著

王斌 韩冀中 万吉 译

Apache基金会官方推荐

Mahout核心团队权威力作

大数据时代机器学习的实战经典

人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

Mahout in Action

Mahout 实战

[美] Sean Owen Robin Anil
Ted Dunning Ellen Friedman 著

王斌 韩冀中 万吉 译

人民邮电出版社
北京

图书在版编目（CIP）数据

Mahout实战 /（美）欧文（Owen, S.）等著 ；王斌，韩冀中，万吉译. -- 北京 ：人民邮电出版社，2014.3
（图灵程序设计丛书）
书名原文：Mahout in action
ISBN 978-7-115-34722-0

Ⅰ. ①M… Ⅱ. ①欧… ②王… ③韩… ④万… Ⅲ. ①机器学习②电子计算机－算法理论 Ⅳ. ①TP181 ②TP301.6

中国版本图书馆CIP数据核字(2014)第031399号

内 容 提 要

本书是 Mahout 领域的权威著作，出自该项目核心成员之手，立足实践，全面介绍了基于 Apache Mahout 的机器学习技术。本书开篇从 Mahout 的故事讲起，接着分三部分探讨了推荐系统、聚类和分类，最后的附录涵盖 JVM 调优、Mahout 数学知识和相关资源。

本书适合所有数据分析和数据挖掘人员阅读，需要有 Java 语言基础。

◆ 著 [美] Sean Owen Robin Anil Ted Dunning Ellen Friedman
译 王 斌 韩冀中 万 吉
责任编辑 毛倩倩
执行编辑 刘 帅
责任印制 焦志炜

◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路11号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
三河市海波印务有限公司印刷

◆ 开本：800×1000 1/16
印张：21.25
字数：502千字 2014年3月第1版
印数：1－4 000册 2014年3月河北第1次印刷
著作权合同登记号 图字：01-2011-7805号

定价：79.00元

读者服务热线：(010)51095186转600 印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京崇工商广字第 0021 号

前　言

追溯这本书的由来，就我个人而言，要从2005年说起。我的一个朋友当时正在创办一家公司，急需协同过滤技术。虽然当时可以找到成熟、开源的软件包，但是它们要么太过繁复，要么太学术化。所以，我决定从零开始，为这个朋友的创业公司开发了一个推荐系统的简单原型。遗憾的是，这家创业公司夭折了。然而，我却无法说服自己删除这个原型。它实在有趣，于是我对它进行整理并写了文档，用Taste这个名字将它发布为一个开源项目。

一年无声地过去了。我在业余时间为其增加了一些代码，并修复了一些问题。接着，有一两个用户出现，并提交了一些软件bug和补丁，然后又有了几个用户，再后来又增加了好一些。到了2008年，虽然小但却稳定的用户群形成了。后来，Apache Lucene的人把机器学习相关的部分剥离出来形成了Apache Mahout，他们建议把我们的两个项目进行合并。此后，本书在2009年晚些时候开始立项。而今，当看到这个项目滚雪球般地发展到2011年，并开始被大公司在生产系统中使用，我自己既惊讶又欣喜。

的确，我只是无心插柳。即便我已经是一个高级工程师，曾在谷歌工作过，也没有人会误认为我是这个领域的专家。我更像是一个博物馆的管理者，而不是一个画家，我将一个领域的伟大思想进行搜集、组织和打包，使之广为所用。这同样不失为一项有用的工作。

一些人在读过本书的初稿之后，说它是一本“通俗易懂”的机器学习书。这是一种盛誉，而我完全赞同。机器学习有其魔力所在，不过这个领域中有许多研究性的著作对于非专业人员而言就像天书，它们也与该技术的应用实践相去甚远。而本书旨在让读者易于理解，为爱好者揭示领悟的快乐，并为实践者节省工作时间。我希望你阅读本书时的惊喜比疑问多。

——Sean Owen

我对机器学习的兴趣可以回溯到2006年上大学的那段日子。那时，我作为实习生和一组人共同设计一个个性化的推荐引擎。这个小组后来成长为Minekey公司；我也被邀请加入，成为其核心开发人员。后来的四年，我一直从事机器学习技术的实现与试验。在此期间，我偶然间发现了Mahout，并开始作为一个Google Summer of Code的参赛学生加入这个项目。我记得，接下来的事情就是不断为它的代码库贡献算法和补丁，做性能调优，以及帮助邮件列表中的其他人。

由机器学习开发者、研究者和爱好者组成的社区非常出色，正在不断成长，而我有幸成为这个团队的一员。随着越来越多的公司采用Mahout，它正在成为机器学习的主流软件库。我衷心希望你在阅读本书时能够乐在其中。

——Robin Anil

我（Ted）在机器学习上是先做研究后做项目。我早期从事的是学术工作，后来参与了一些创业团队，从而得以将机器学习在实际中应用。

我（Ellen）以前在生物化学和分子生物学实验室工作。在研究大量数据的同时，我还写了许多技术文章。此番经历，让我痴迷于数据及其蕴含的意义。我努力把这种内在的东西写进本书里。

我们两个人一致认为开源有赖于一个有大量活跃用户参与的社区。Mahout的成功主要来自于那些使用这个软件的人，他们通过在邮件列表中展开讨论、修复bug，以及提供建议，把使用经验回馈到这个项目中。

为此，本书不仅给出代码的实用注解，而且引出了代码背后的一些概念。介绍隐藏于代码背后的框架，会使你更有效地加入到Mahout的讨论中，并从中获益。我们希望本书不仅能够帮助读者，而且能够使Mahout自身得到完善和发展。

——Ted Dunning和Ellen Friedman

致谢

本书的出版离不开众人的努力。作者对他们致以衷心的感谢，但限于篇幅，致谢名单只列出了其中一部分人，排名不分先后。

- 在机器学习领域发表核心文章的研究者，详见附录C。
- 花时间测试试用版软件的Mahout用户，他们寻找与解决bug，为软件打补丁乃至提出建议。
- Mahout提交者，他们致力于Mahout的发展、完善和提升。
- Manning出版社投入了大量时间和精力将本书出版并投入市场。特别感谢Katharine Osborne、Karen Tegtmeyer、Jeff Bleiel、Andy Carroll、Melody Dolab和Dottie Marsico，你看到的最终稿与他们的工作密不可分。
- 在本书写作过程中提供了宝贵反馈的审校者：Philipp K. Janert、Andrew Oswald、John Griffin、Justin Tyler Wiley、Deepak Vohra、Grant Ingersoll、Isabel Drost、Kenneth DeLong、Eric Raymond、David Grossman、Tom Morton，以及Rick Wagner。
- Alex Ott在本书印刷前一刻对全稿进行了细致的技术审核。
- 在作者在线论坛上发帖评价本书的MEAP读者[①]。
- 每个在Mahout邮件列表中提问的人。
- 在本书长时间写作过程之中，给予我们无尽支持的家人和朋友！

① MEAP，全称Manning Early Access Program。因为图书出版周期较长，为让读者可以时刻了解热门技术，Manning出版社推出了这一图书抢鲜阅读项目。参与其中的读者可以在图书未编辑完成之际，一章一章地阅读。——编者注

关于本书

你可能还有疑问：这本书是否适合我？

如果你正在寻找一本机器学习教材，答案是否定的。本书不会对诸多算法和技术的理论以及推导过程给出全面解释。如果你了解机器学习技术，并熟悉矩阵和向量等相关的数学概念，这有助于阅读本书，但并非必要条件。

如果你正在开发先进、智能的应用，答案则是肯定的。本书从实践而非理论入手来诠释这些技术，并给出完整的例子与解决方案。它在教授用Mahout解决问题的同时，带给你实践者的经验与领悟。

如果你是人工智能、机器学习及相关领域的研究人员，答案亦是肯定的。你所面对的最大障碍，很可能就是把新的算法应用到实践中。对于新的大规模算法的测试与部署，Mahout提供了一个成熟的框架、一系列模式以及现成的组件。本书是你在复杂的分布式计算框架上学习开发机器学习系统的“快车票”。

如果你正领导一个产品小组或创业团队，想利用机器学习创造一种竞争优势，那么本书同样适合你。通过实际示例，它让你了解这些技术的多种应用方式。它还会让小型技术团队变得高效，能够处理大量数据，而这在以前只有具有大量技术资源的组织机构才做得到。

路线图

本书分为三部分，分别介绍了Apache Mahout中的协同过滤、聚类和分类。

首先，第1章整体介绍Apache Mahout。这一章为你阅读后续各章奠定基础。

第一部分（第2章~第6章）由Sean Owen编写，主要介绍协同过滤与推荐。第2章基于Mahout构造推荐引擎并评估其性能。第3章探讨如何高效呈现推荐引擎使用的数据。第4章介绍可在Mahout中利用的所有推荐算法，并比较它们的优缺点。在此背景之下，第5章给出一个案例，将第4章介绍的推荐系统实现应用在该案例的真实问题中，配以某些特定属性的数据，从而建立一个可为生产环境所用的推荐引擎。第6章介绍Apache Hadoop，通过研究基于Hadoop的推荐引擎，首次为你展现分布式环境中的机器学习算法。

第二部分（第7章~第12章）探索Apache Mahout上的聚类算法。通过Robin Anil所做的技术说明，你可以把看起来类似的数据片段组织为一个个集合或者说簇（cluster）。聚类有助于揭示大规模数据中有趣的信息组合。这部分从聚类中的简单问题开始介绍，并给出了Java示例。接下来，

作者引入更多实际示例，并展示如何让Apache Mahout以Hadoop作业的方式运行，从而轻而易举地聚类大量数据。

第三部分（第13章~第17章）由Ted Dunning和Ellen Friedman编写，探索如何用Mahout进行分类。首先，作者带你了解如何通过一组示例“教会”一个算法，从而建立和训练分类器模型。接下来，你会了解如何评估并微调分类器模型以得到更好的结果。这部分最后以一个分类实战案例结束。

代码约定及下载

本书源代码均采用等宽字体印刷，列为代码清单，并对重点进行注释。代码清单旨在简单明了，重点突出。它们通常不给出Java导入包、类声明、Java注释，以及其他对代码的讨论无关紧要的东西。

本书中的类名亦采用等宽字体，放于文本之间，以显示它们是可以在Apache Mahout源代码中找到并研究的类名。例如，`LogLikelihoodSimilarity`是Mahout中的一个Java类。

一些代码清单中列出了可执行命令。它们是为Mac OS X和Linux发行版等类Unix环境而写的。如果使用了类Unix的Cygwin环境，它们也可以在微软Windows系统下运行。

本书关键代码清单中的源代码均可编译，且均可从www.manning.com/MahoutinAction下载[①]。这些都是独立的Java源文件，并不包括编译脚本。为了方便起见，你可以把它们解压到Mahout源代码发布包的examples/src/java/main目录下。这样，Mahout的编译环境将会自动编译这些代码。

多媒体资料

四位作者均录制了音频和视频片段，与书中多数章中的特定节互相补充，为相应话题提供了附加信息。你可以从本书英文版电子书中看到或听到这些音视频片段，该电子书对英文纸质书的拥有者免费；你还可以从www.manning.com/MahoutinAction/extras免费获取。通过书中的音频和视频图标，你可以获知其所涉及的话题，以及发言者是谁。这些多媒体资料的清单详见“关于多媒体资料”。

作者在线

本书英文版读者可以免费访问Manning出版社专门维护的一个论坛，并可以发表评论、提出技术问题，并获得作者和其他论坛用户的帮助。你可以通过网页www.manning.com/MahoutinAction进入和订阅该论坛。完成注册后，你可以了解如何使用该论坛、该论坛所能提供的帮助，以及论坛的行为规范。

Manning出版社承诺为读者和作者提供一个进行深入对话的场所，但不对作者的参与程度做要求，他们对于该论坛的贡献是出于自愿且无报酬的。我们建议读者尽量向作者提一些具有挑战性的问题，让他们保持兴趣！

本书在印期间，读者均可访问作者在线论坛，并查看之前的讨论。

---

① 亦可在图灵社区（iTuring.cn）本书页面免费注册下载。——编者注

关于多媒体资料

本书附带的多媒体资料可以在www.manning.com/MahoutinAction/extras/上免费收听或收看。本书中空白处的音频或视频文件图标（如下所示），指出了书中哪些地方可参考附加资料。

Audio icon

Video icon

No. 1 音频 p2
Sean介绍了Mahout项目以及他参与的事项。

No. 2 音频 p19
Sean讨论了推荐系统的工作。

No. 3 音频 p29
Sean阐述为什么他认为人们有可能过度“聆听”数据。

No. 4 音频 p42
Sean谈论皮尔逊相关系数的实现。

No. 5 音频 p63
Sean讨论了诠释性能指标的价值。

No. 6 音频 p84
Sean解释了Mahout和Hadoop之间的关系。

No. 7 音频 p108
Robin解释了如何为一个数据集选择正确的距离测度方法。

No. 8 音频 p114
Robin扩展了苹果的类比示例。

No. 9 音频 p127
Robin解释了k-means聚类迭代过程。

No. 10 音频 p165
Robin讨论改善聚类质量的策略。

No. 11 音频 p179
Robin解释了如何改进大规模聚类的性能。

No. 12　视频　p208
Ellen展示了如何训练一个模型使之逐步优化。

No. 13　视频　p234
Ted和Ellen展示了Logistic回归的内部机制。

No. 14　视频　p238
Ted比较了使用串行算法与并行算法的优势。

No. 15　音频　p249
Ted和Ellen讨论了AUC评估方法。

No. 16　音频　p252
Ted和Ellen讨论了为什么对数似然法意味着“永不说不”。

关于封面

封面上是“一个来自Rakov-Potok的男人”。Rakov-Potok是克罗地亚北方的一个村庄。该图取自克罗地亚19世纪中叶传统服饰影集的一个副本，作者为Nikola Arsenovic，由Ethnographic博物馆在2003年出版于克罗地亚的斯普利特。该图得自于乐于助人的Ethnographic博物馆馆员，这个博物馆位于该城镇在中世纪罗马时的核心位置，是公元304年左右罗马皇帝戴克里先的宫殿遗址。这本书包含来自克罗地亚不同地域的颜色精美的插图，附有服饰和日常生活的说明。

Rakov-Potok是一个风景如画的乡村，位于Samobor山脚下、萨瓦河土地肥沃的河谷中，距Zagreb城不远。它有着悠久的历史，在那里，你会与许多城堡、教堂和中世纪甚至罗马时期的遗迹不期而遇。封面上的人物身着白色羊毛长裤和白色羊毛外套，上面有着大量的红色和蓝色绣花——这是该地区山区居民的典型装束。

过去200年间，人们的着装和生活方式已经发生变化，曾经如此丰富的地域多样性已渐渐消失了。现在，各大洲的居民已经很难分辨，更遑论不同小村或距离只有几英里的人。也许我们用文化多样性换来的是更多样化的个人生活——必然是更为丰富和快节奏的技术生活。

Manning出版社在此类古老书籍的插图中取材，基于两个世纪前丰富多样的地域生活来制作图书封面，借此颂扬计算机行业的创造力和首创精神。

目　录

第1章 初识Mahout

本章内容

- Apache Mahout是什么？从哪里来？
- 现实中的推荐引擎、聚类和分类一览
- Mahout安装

大概你已经从书名中猜到了，本书主要讲解一个特殊的工具——Apache Mahout——在现实生活中的高效应用。它具备三个明显的特征。

首先，Mahout是一个来自Apache的、开源的机器学习（machine learning）软件库。它所实现的算法归属于机器学习或集体智慧（collective intelligence，也常常译为群体智慧或群体智能）这个广阔的领域。这意味着有许多事情可做，但对于此时此刻的Mahout，它主要关注于推荐引擎（协同过滤）、聚类和分类。

其次，Mahout是可扩展的。它旨在当所处理的数据规模远大于单机处理能力时成为一种可选的机器学习工具。在当前的Mahout系统中，这些可扩展的机器学习实现都是用Java来写的，而且有些部分是建立在Apache的Hadoop分布式计算项目之上的。

最后，它是一个Java软件库，并不提供用户接口、预装服务器（prepackaged server）或安装程序（installer）。它打算为开发者提供一个可用可改的工具框架。

出于阶段安排的需要，本章将通过一些常见的真实案例简要介绍一下推荐引擎、聚类和分类这几种机器学习手法，而Mahout通过它们帮助你处理数据。

若要在阅读本书时做到对Mahout随学随用，还要做一些必要的系统搭建与安装工作。

1.1 Mahout的故事

首先来了解Mahout的背景知识。你可能还搞不清Mahout该如何发音：就是其通常的英语发音（[mə'haʊt]），它和trout押韵。它来自北印度语，意为驱象人，若要解释它，还有个小故事。

Mahout是2008年作为Apache Lucene的子项目出现的。Lucene项目推出了一个同名的著名开源搜索引擎，并给出了搜索、文本挖掘（text mining）和信息检索技术的先进实现方法。在计算机科学领域，这些术语和机器学习技术中的概念很相近，比如聚类（clustering），并在某种程度

上与分类（classification）相近。这样一来，某些Lucene贡献者的工作更多落入机器学习领域，从而逐渐脱离出来形成了独立的子项目。之后不久，Mahout吸纳了开源的协同过滤项目Taste。

No. 1

图1-1给出了Mahout在ASF（Apache Software Foundation，Apache软件基金会）中的部分传承关系。到2010年4月，Mahout已经成为了一个独立的顶级Apache项目，并发布了一个全新的驱象人徽标。

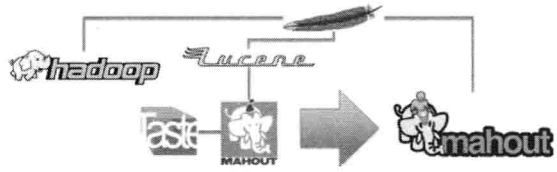

图1-1　Apache Mahout及其在ASF中的相关项目

Mahout所做的大量工作不仅体现在以高效和可扩展的方式实现这些经典算法，而且将部分算法进行转换使其可以在Hadoop上处理大规模的问题。Hadoop的吉祥物是一头象，Mahout项目的名字便由此而来！

从Mahout孵化出了许多技术和算法，其中有许多仍在开发或实验阶段（https://cwiki.apache.org/confluence/display/MAHOUT/Algorithms）。在该项目的早期，有3个明确的核心主题：推荐引擎（协同过滤）、聚类和分类。虽然它们绝非Mahout的全部，但在本书写作时，它们是最突出和最成熟的主题，也因此成为了本书的焦点。

也许你在阅读本书时已经了解了这三种技术的魅力，但为了不漏掉什么，请继续读下去。

1.2　Mahout的机器学习主题

虽然Mahout项目在理论上可以实现所有类型的机器学习技术，但实际上当前它仅关注机器学习的三个主要领域，即推荐引擎（协同过滤）、聚类和分类。

1.2.1　推荐引擎

在目前采用的机器学习技术中，推荐引擎是最容易一眼就被认出来的。服务商或网站会根据你过去的行为向你推荐书籍、电影或文章。它们会推测你的品味与爱好，并找到某些你可能感兴趣的物品。

- 在部署了推荐系统的电子商务网站中，亚马逊大概是最有名的。亚马逊基于交易行为和网站记录为你推荐你可能感兴趣的书籍和其他物品（见图1-2）。
- 与之类似，Netflix为用户推荐其可能感兴趣的DVD，为了鼓励研究者改善其推荐质量，它给出了一份1 000 000美元的奖金，这使它颇具盛名。
- 像Líbímseti这样的约会网站（稍后讨论）还能把一个人推荐给另一个人。