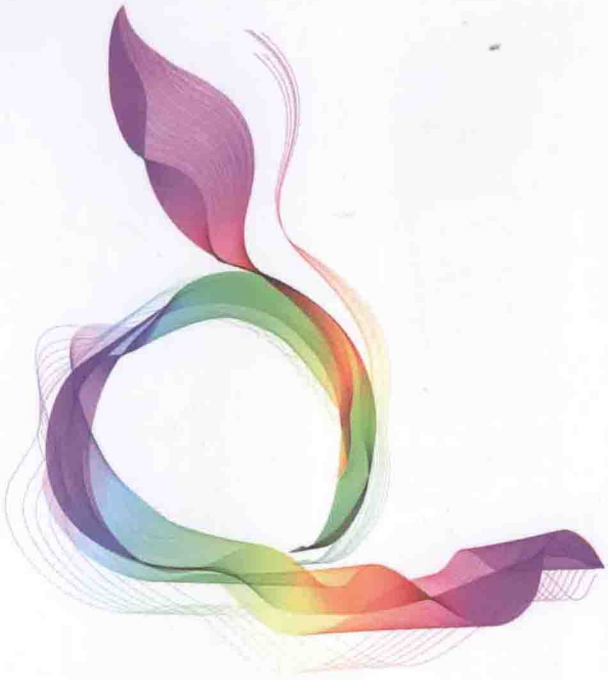
HZ BOOKS
华章科技

全面分析Mahout算法库中不同模块中各个算法的原理及其Mahout实现流程
每个算法都辅之以实战案例，同时还包括4个系统级案例，实战性强

大数据
技术丛书



Mahout in Action: Algorithm and Cases

Mahout算法解析与案例实战

樊 哲◎著

机械工业出版社
China Machine Press

大数据技术丛书

Mahout in Action: Algorithm and Cases

Mahout算法解析与案例实战

樊 哲◎著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Mahout 算法解析与案例实战 / 樊哲著 . —北京：机械工业出版社，2014.6
（大数据技术丛书）

ISBN 978-7-111-46797-7

I. M… II. 樊… 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 106826 号

本书是一本经典的 Mahout 著作，原理与实战并重。不仅全面分析了 Mahout 算法库不同模块中的各个算法的原理及其实现流程，而且每个算法都辅之以实战案例。此外，还包括 4 个系统级案例，实战性非常强。

全书 11 章共分为三个部分：第一部分为基础篇（第 1 ～ 2 章），首先介绍了 Mahout 的应用背景、Mahout 算法库收录的算法、Mahout 的应用实例，以及开发环境的搭建；第二部分为算法篇（第 3 ～ 7 章），分析了 Mahout 算法库中不同模块的各个算法的原理以及 Mahout 实现流程，同时在各章节含有每个算法的实战，让读者可以自己运行程序，感受程序运行的各个流程；第三部分为实战篇（第 8 ～ 11 章），通过对 4 个不同系统案例的分析讲解，让读者了解开发完整的云平台系统的各个流程，即需求分析、系统框架选择及构建、系统功能设计和功能开发。

Mahout 算法解析与案例实战

樊 哲 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：佘 洁 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：17.5
书 号：ISBN 978-7-111-46797-7 定 价：59.00 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

Preface 前　言

为什么要写这本书

2010 年以后，世界快速进入了大数据时代，Hadoop 成为大数据分析的首选平台和开发标准，无数数据分析软件纷纷向 Hadoop 靠拢。在 Hadoop 原有技术基础之上，涌现了 Hadoop 家族产品，它们正在配合“大数据”概念不断创新，推动科技进步。因此，新一代 IT 精英也必须顺应潮流，抓住机遇，随着 Hadoop 一起发展和成长！

简而言之，Hadoop 是由 Apache 基金会开发的一个优秀的云计算技术框架，用户在其基础上即使不了解分布式底层细节，也可以开发分布式程序。Hadoop 家族成员中的诸多成员进一步利用了这一优势，拓展了云计算的应用领域，降低了相应的软件开发门槛，而 Mahout 就是其中最难掌握，也是最有竞争力且最值得学习的项目之一。

Mahout 是一个基于 Hadoop 的机器学习和数据挖掘的分布式计算框架，在 MapReduce 模式下封装实现了大量数据挖掘经典算法，为 Hadoop 开发人员提供了数据建模的标准，从而大大降低了大数据应用中并行挖掘产品的开发难度。在掌握了 Mahout 之后，Hadoop 开发人员可以直接调用相关算法模型的接口，方便、快捷地创建智能应用程序，从而大幅提升商业智能软件的大数据分析处理能力。

但是，目前关于 Mahout 的参考资料比较少，比较有名的是 Sean Owen 编写的《Mahout in Action》，更多时候开发者只能通过 Mahout 的官网或者网络上一些技术爱好者发布的博客内容来进行学习。《Mahout in Action》是一本全英文的书籍，而且出版年份比较早，对国内的一些 Mahout 爱好者来说，阅读此书有一定的难度，因此，笔者就有了结合自己的经验写一本与 Mahout 有关的书籍的想法。本书针对 Mahout 算法库目前收录的大多数算法进行了分析，同时收录了笔者开发的 4 个简单系统，作为读者学习和实践的实例。

读者对象

- ❑ Hadoop 用户和爱好者
- ❑ 云平台系统架构师
- ❑ Mahout 代码二次开发者
- ❑ 云平台系统开发者
- ❑ 使用 Mahout、Hadoop 的相关用户
- ❑ 开设相关课程的大专院校学生

如何阅读本书

本书分为三大部分：

第一部分为基础篇（第 1 ～ 2 章），首先对 Mahout 的应用背景以及 Mahout 算法库收录的算法进行了简单介绍，同时分析了 Mahout 的应用实例。接着介绍其开发环境并详细分析了它的配置，使读者可以搭建一个自己的开发环境，为后面实战做好准备。

第二部分为算法篇（第 3 ～ 7 章），分析了 Mahout 算法库中不同模块的各个算法的原理以及 Mahout 实现流程，同时在每章末尾都有算法实战，让读者自己运行程序，感受程序运行的各个流程。

第三部分为实战篇（第 8 ～ 11 章），通过对 4 个不同系统案例的分析讲解，让读者了解开发完整的云平台系统的各个流程，即需求分析→系统框架选择及构建→系统功能设计→功能开发和界面开发。

其中第三部分以接近实战的案例来讲解云平台算法和当前流行框架的结合，此部分内容需要读者有一定的 Spring、Struts 2、Hibernate 等框架的基础。第一、第二部分则是 Mahout 基础知识，如果读者对 Mahout 不熟悉，建议从第 1 章内容开始阅读。

勘误和支持

除封面署名外，还有很多人对本书的写作提供了帮助，分别为：张汉锐、张良均、刘名军、庄思待、曾祥柱、曾健荣等。由于作者的水平有限，加之编写时间仓促，书中难免会出现一些错误或者不准确的地方，恳请读者批评指正。为此，读者可以通过笔者微博（http://weibo.com/fansy1990）或 CSDN 地址（http://blog.csdn.net/fansy1990）反馈有关问题。如果你有更多的宝贵意见，也欢迎发送邮件至邮箱 fansy1990@foxmail.com，期待能够得到你们的真挚反馈。

致谢

首先要感谢 Apache 基金会，感谢其开源的伟大精神。如果没有 Apache 基金会，没有其 Mahout 开源项目，那么就不会有这本书。

感谢华南师范大学的薛云老师，在写本书的过程中，他给了我很多指导，为我指明了方向，同时提供了很多支持。

感谢 CSDN 社区，它提供了一个技术交流平台，笔者在这个平台学到了很多知识和技能，为笔者编写本书提供了很多帮助和支持。

感谢机械工业出版社华章公司的编辑杨福川、姜影和佘洁，在这一年多的时间中始终支持我的写作，你们的鼓励和帮助引导我顺利完成全部书稿。

最后感谢我的爸爸、妈妈，感谢你们将我培养成人，感谢你们在我人生的每个阶段都能给我提供不同的建议，并时时刻刻支持我。无论我做什么决定，你们的支持是我坚持的动力。

谨以此书献给我最亲爱的家人，以及众多热爱 Mahout 的朋友！

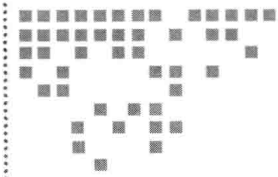
目 录 *Contents*

第三部分 实战篇

第一部分 Part 1

基 础 篇

- 第1章 Mahout简介
- 第2章 Mahout安装配置



Chapter 1 第 1 章

Mahout 简介

当今社会什么技术最牛？什么技术最火？也许很多人会说是云计算，它可以说是近几年来一直被热议的“高深莫测”的词汇。大家都在说云计算，但是很少人能把云计算说得彻底且明白，大多数人还是有“云里雾里”的感觉。虽然如此，但是随着最近几年云计算概念的普及，云计算神秘的面纱正在慢慢地被揭开。云计算的核心重点是云平台下算法的开发，有了算法的支撑才能发挥云计算的最大优势。Mahout 开源项目就是一个 Hadoop 云平台的算法库，已经实现了多种经典算法，并一直在扩充中，其目标就是致力于创建一个可扩容的云平台算法库。

下面就让我们开始 Mahout 探索之旅吧。

1.1 Mahout 应用背景

随着互联网的发展，企业拥有的数据也越来越多，比如 Facebook 公司，从公司成立之初的 100 万用户数到 2010 年的 1.34 亿用户数，再到 2014 年的 13.1 亿用户数，其用户增长速度达到了令人惊叹的地步，单单用户数目的增长已经达到了如此地步，更不用说每个用户所产生的数据量了。很明显，面对如此庞大的数据量，企业再用以前的数据处理方式显然已经不能满足要求了。

正所谓，变则通，通则久。企业若想长久发展，面对日益增长的数据，在以前传统的数据处理方式显得力不从心的时候，就需要“变”。所谓“变”，其实就是对现有方式的创新。在此情况下，“云计算”便应运而生。所谓“云计算”是一种基于互联网的计算方式，通过这种方式，共享的软硬件资源和信息可以按需提供给计算机和其他设备，这样可以最

大限度、最大效率地利用计算机资源，达到快捷、高速地处理数据的目的。

但是，单单有云计算平台还不够，还需要有适合云平台的算法。云计算的核心就是计算，要研究可以在云平台上实现的算法，这样才能发挥云计算的最大威力。以前的数据挖掘算法是在单机上实现的，单机实现的算法其编程思路和模式与云平台下的编程思路和模式很不一样，如果还是按照以前的思路，那么肯定是行不通的。

目前开源的云平台有多种，本书所述的云平台是 Hadoop 云平台。Hadoop 云平台是一个用于处理大数据的分布式应用的开源框架，提供分布式存储和高效计算能力。Hadoop 具有以下优势：

- 同时提供分布式存储和计算能力。
- 具有极高的可扩展性。
- 其主要的组件之一 HDFS 具有很高的数据吞吐量。
- 具有软件和硬件容错性。
- 允许大数据的并行工作。

在 Hadoop 云平台下编程不仅要求用户对 Hadoop 云平台框架比较熟悉，还要对 Hadoop 云平台下底层数据流、Map 和 Reduce 原理非常熟悉，这是基本的编程要求。此外，用户要编写某一个算法还需要对该算法的原理比较熟悉，即需要对算法原理理解透彻。总体来看，编写云平台下的算法程序是属于高难度的开发工作了。但是，如果使用 Mahout，情况就会有很大的不同，用户再也不用自己编写复杂的算法，不需要掌握太高深的云平台的框架和数据流程的理论知识。用户所需要了解的只是算法的大概原理、算法实际应用环境和如何调用 Mahout 相关算法的程序接口。当然，在具体的项目中，用户还应该根据实际需求在 Mahout 源代码基础上进行二次开发以满足具体的实际应用情况。

Mahout 是 Apache 基金会的开源项目之一。Apache Mahout 起源于 2008 年，当时它是 Apache Lucene 的子项目。在使用 Hadoop 云平台的基础上，可以将其功能有效地扩展到 Hadoop 云平台中，提高其运算效率。2010 年 4 月，Apache Mahout 最终成为了 Apache 的顶级项目。创建此项目的用意是建立一个可扩容的云平台算法库。目前，Mahout 已经实现了多种经典数据挖掘算法，算是比较完备的算法库了。Mahout 目前还在扩充中，由世界上对这个项目感兴趣的云平台算法编程高手们一起进行开发、测试，然后进行算法扩充，任何对这个项目感兴趣的个人或者组织都可以加入到该项目的社区中，为该项目做出贡献。

1.2 Mahout 算法库

Mahout 自从 2008 年兴起以来，发展迅速，从最开始的只有推荐系统到现在的多个算法模块，涵盖了很多行业。这些模块有聚类算法、分类算法、协同过滤算法和频繁项集挖掘算法，每个模块都含有一个或者几个不同的实现算法，下面分别进行介绍。

1.2.1 聚类算法

中国有句古谚语“物以类聚，人以群分”。一个聚类即是一类物体的集合，集合中的个体是相似的，不同聚类中的个体是不相似的。聚类的二维图如图 1-1 所示。

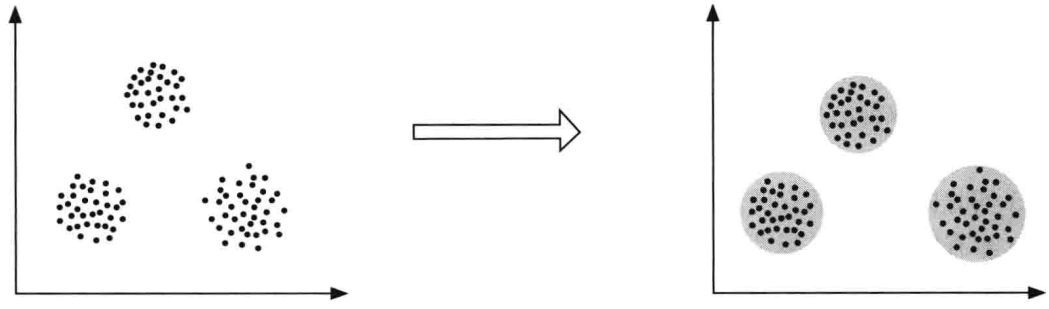


图 1-1 聚类二维图

针对上面的数据，我们可以很容易地把它们分为右边阴影中的 3 类，这里的分类依据是不同点之间的距离：对于两个或者多个数据点，当它们之间的距离达到一定程度的时候，我们就把它们分为一个类，采用这种方式的聚类称做基于几何距离的聚类。

可以看到，聚类的目的就是把一组无标签的数据加上标签。那么，如何去评价一个模型的好坏？如何去评判一个模型把一组无标签的数据“完美地”贴上了标签呢？事实上，没有一个绝对的标准来衡量这些模型算法，所以，一般都是用户根据自己的需要评测一个模型的好坏，而且还要求模型的参数要根据用户的不同数据加以调整以适应具体的情况。

Mahout 算法库中聚类模块包含的算法有：Canopy、K-Means、Fuzzy K-Means、Mean Shift、Hierarchical、Spectral、Minhash、Top Down，其中在小括号中标注“开发中”的算法其编写还不是很完善。下面对这些算法分别进行简要分析。

（1）Canopy 算法

Canopy 算法是一种非常简单、快速的聚类方法。Canopy 算法经常用于其他聚类算法的初始步骤，比如 K-Means 算法等。

（2）K-Means 算法

K-Means 算法是一种相对简单但是广为人知的聚类算法，一般聚类问题都可以使用聚类算法。在 Mahout 中，该算法在每次循环时都会新建一个任务，对于算法来说，增加了很多外部消耗。

（3）Fuzzy K-Means

Fuzzy K-Means 是 K-means 的扩展，是一种比较简单且流行的聚类方法。相比于 K-Means 聚类方法用于发现严格的聚类中心（即一个数据点只属于一个聚类中心），Fuzzy K-Means 聚类方法用于发现松散的聚类中心（即一个数据点可能属于几个聚类中心）。

（4）Mean Shift 算法

Mean Shift 算法最开始应用于图像平滑、图像分割和跟踪方面，在 1995 年一篇重要的

文献发表后，Mean Shift 才被大家所了解。Mean Shift 算法比较吸引人的地方是该算法不需要提前知道要聚类的类别数（K-Means 算法就需要），并且该算法形成的聚类形状是任意的且与要聚类的数据是相关的。

（5）Spectral 算法

Spectral 算法相对于 K-Means 算法来说更加有效和专业化，它是处理图像谱分类的一种有效的算法，主要针对的数据也是图像数据。

（6）Minhash 算法

Minhash 算法只负责将原始内容尽量均匀随机地映射为一个签名值，原理上相当于伪随机数产生算法。对于传统 hash 算法产生的两个签名，如果相等，说明原始内容在一定概率下是相等的；如果不相等，除了说明原始内容不相等外，不再提供任何信息，因为即使原始内容只相差一个字节，所产生的签名也很可能差别极大。从这个意义上来说，要设计一个 hash 算法，使相似的内容产生的签名也相近，是更为艰难的任务，因为它的签名值除了提供原始内容是否相等的信息外，还能额外提供不相等的原始内容的差异程度的信息。

（7）Top Down 算法

Top Down 算法是分层聚类的一种，它首先寻找比较大的聚类中心，然后对这些中心进行细粒度分类。

1.2.2　分类算法

分类是一种基于训练样本数据（这些数据都已经被贴标签）区分另外的样本数据标签的过程，即另外的样本数据应该如何贴标签的问题。举一个简单的例子，现在有一批人的血型已经被确定，并且每个人都有 M 个指标来描述这个人，那么这批人的 M 个指标数据就是训练样本数据，根据这些训练样本数据，建立分类器（即运用分类算法得到一些规则），然后使用分类器对测试样本集中的未被贴标签的数据进行血型判断。分类算法和聚类算法的不同之处在于，分类是有指导的学习，而聚类是一种无指导的学习。有指导和无指导其实是指在训练的时候训练样本数据是否提前被贴上了标签。图 1-2 为分类算法的一般过程。

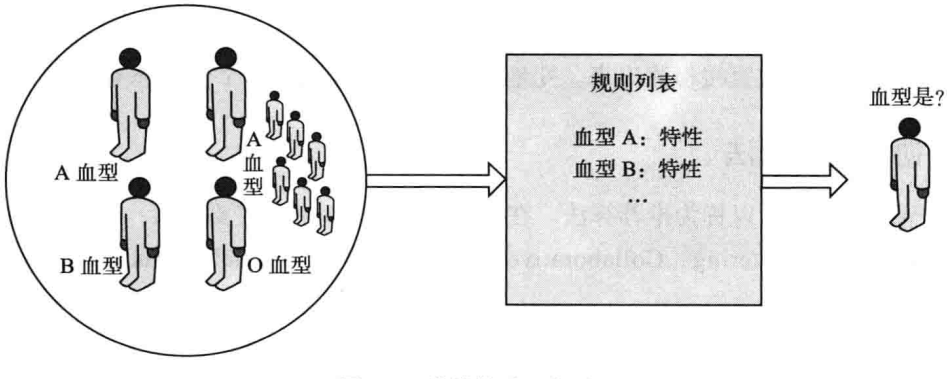


图 1-2　分类算法一般过程

Mahout算法库中分类模块包含的算法有：Logistic Regression、Bayesian、Support Vector Machine、Random Forests、Hidden Markov Models。

（1）Logistic Regression

Logistic Regression是一种利用预测变量（预测变量可以是数值型，也可以是离散型）来预测事件出现概率的模型。其主要应用于生产欺诈检测、广告质量估计，以及定位产品预测等。在Mahout中主要使用随机梯度下降（Stochastic Gradient Decent，SGD）思想来实现该算法。

（2）Bayesian

通常，事件A在事件B发生的条件下的概率，与事件B在事件A发生的条件下的概率是不一样的；然而，这两者是有确定的关系，贝叶斯（Bayesian）定理就是这种关系的陈述。通过联系事件A与事件B，计算从一个事件产生另一事件的概率，即从结果上溯源。

在Mahout中，目前已经有两种实现的贝叶斯分类器了，其中一种是朴素贝叶斯算法，另外一种是互补型的朴素贝叶斯算法。

（3）Support Vector Machine

Support Vector Machine（支持向量机）属于一般化线性分类器，也可以认为是提克洛夫规范化（Tikhonov Regularization）方法的一个特例。这种分类器的特点是它能够同时最小化经验误差与最大化几何边缘区，因此支持向量机也称为最大边缘区分类器。

（4）Random Forests

Random Forests（随机森林）是一个包含多个决策树的分类器，并且其输出的类别由个别树输出的类别的众数而定。这里的众数是指个别树输出类别重复最多的一个类别数值。随机森林算法在决策树的基础上发展而来，继承了决策树的优点，同时弱化了决策树的缺点。

（5）Hidden Markov Models

Hidden Markov Models（隐马尔科夫模型）主要用在机器学习上，比如语音识别、手写识别及自然语音处理等。隐马尔科夫模型是一个包含两个随机变量O和Y（O和Y可以按照顺序改变它们自身的状态）的分析模型。其中，变量Y是隐含变量，包含{y_1，…，y_n}个状态，其状态不能被直接检测出来。变量Y的状态按照一定的顺序改变，其状态改变的概率只与当前状态有关而不随时间改变。变量O称为可观察变量，包含{o_1，…，o_m}个状态，其状态可以被直接检测出来。变量O的状态与当前变量Y的状态有关。

1.2.3 协同过滤算法

协同过滤算法也可以称为推荐算法。在Mahout算法库中，主要包括：Distributed Item-Based Collaborative Filtering、Collaborative Filtering using a parallel matrix factorization，下面进行简要分析。

（1）Distributed Item-Based Collaborative Filtering

Distributed Item-Based Collaborative Filtering是基于项目的协同过滤算法，其简单思想

就是利用项目之间的相似度来为用户进行项目推荐。项目之间的相似度通过不同用户对该项目的评分来求出，每个项目都有一个用户向量，两个项目之间的相似度便是根据这个用户向量求得的。求得项目之间的相似度，便可以针对用户对项目的评分清单来推荐与清单中极为相似的项目。

（2）Collaborative Filtering using a parallel matrix factorization

Collaborative Filtering using a parallel matrix factorization 在 Mahout 的介绍中是以 Collaborative Filtering with ALS-WR 的名称出现的。该算法最核心的思想就是把所有的用户以及项目想象成一个二维表格，该表格中有数据的单元格（i, j），便是第 i 个用户对第 j 个项目的评分，然后利用该算法使用表格中有数据的单元格来预测为空的单元格。预测得到的数据即为用户对项目的评分，然后按照预测的项目评分从高到低排序，便可以进行推荐了。

1.2.4　频繁项集挖掘算法

在 Mahout 算法库中，频繁项集挖掘算法主要是指 FP 树关联规则算法。传统关联规则算法是根据数据集建立 FP 树，然后对 FP 树进行挖掘，得到数据库的频繁项集。在 Mahout 中实现并行 FP 树关联规则算法的主要思路是按照一定的规则把数据集分开，然后在每个分开的部分数据集建立 FP 树，然后再对 FP 树进行挖掘，得到频繁项集。这里使用的是把数据集分开的规则，可以保证最后通过所有 FP 树挖掘出来的频繁项集全部加起来没有遗漏，但是会有少量重叠。

1.3　Mahout 应用

作为 Apache 基金会的顶级项目之一，Mahout 的应用也极其广泛，一般分为商业应用和学术应用。

在商业应用中，Adobe AMP 公司使用 Mahout 的聚类算法把用户区分为不同的圈子，通过精确定位营销来增加客户。Amazon 的个人推荐平台也是使用 Mahout 的算法库来进行推荐的。AOL 使用 Mahout 来进行购物推荐。DataMine Lab 使用 Mahout 的推荐算法以及聚类算法来提高客户广告投放的精确度。iOffer 使用 Mahout 频繁项集挖掘算法和协同过滤算法为用户推荐项目。Twitter 使用 Mahout 的 LDA 模型为用户推荐其感兴趣的东西。Yahoo 公司的邮件使用 Mahout 的关联规则算法。

在学术应用中，Mahout 也被广泛应用。在 TU Berlin 大学的“Large Scale Data Analysis and Data Mining”课程中，使用 Hadoop 和 MapReduce 来进行数据并行分析的教学。在 Nagoya Institute of Technology，Mahout 被用来在一个研究项目中进行数据分析。

在本书中，笔者结合自身经验，设计并开发了 4 个简易系统，分别是 Friend Find 系