

《数据算法》

图书基本信息

书名：《数据算法》

13位ISBN编号：9787512395949

出版时间：2016-10-1

作者：Mahmoud Parsian

译者：苏金国,杨健康

版权说明：本站所提供下载的PDF图书仅提供预览和简介以及在线试读，请支持正版图书。

更多资源请访问：www.tushu000.com

《数据算法》

内容概要

《数据算法：Hadoop/Spark大数据处理技巧》介绍了很多基本设计模式、优化技术和数据挖掘及机器学习解决方案，以解决生物信息学、基因组学、统计和社交网络分析等领域的很多问题。这还概要介绍了MapReduce、Hadoop和Spark。

主要内容包括：

- 完成超大量交易的购物篮分析。

- 数据挖掘算法（K-均值、KNN和朴素贝叶斯）。

- 使用超大基因组数据完成DNA和RNA测序。

- 朴素贝叶斯定理和马尔可夫链实现数据和市场预测。

- 推荐算法和成对文档相似性。

- 线性回归、Cox回归和皮尔逊（Pearson）相关分析。

- 等位基因频率和DNA挖掘。

- 社交网络分析（推荐系统、三角形计数和情感分析）。

《数据算法》

作者简介

Mahmoud Parsian，计算机科学博士，是一位热衷于实践的软件专家，作为开发人员、设计人员、架构师和作者，他有30多年的软件开发经验。目前领导着Illumina的大数据团队，在过去15年间，他主要从事Java (服务器端)、数据库、MapReduce和分布式计算的有关工作。Mahmoud还著有《JDBC Recipes》和《JDBC Metadata, MySQL, and Oracle Recipes》等书（均由Apress出版）。

书籍目录

序 1
前言 3
第1章二次排序：简介 19
二次排序问题解决方案 21
MapReduce/Hadoop的二次排序解决方案 25
Spark的二次排序解决方案 29
第2章二次排序：详细示例 42
二次排序技术 43
二次排序的完整示例 46
运行示例——老版本Hadoop API 50
运行示例——新版本Hadoop API 52
第3章 Top 10 列表 54
Top N 设计模式的形式化描述 55
MapReduce/Hadoop实现：唯一键 56
Spark实现：唯一键 62
Spark实现：非唯一键 73
使用takeOrdered()的Spark Top 10 解决方案 84
MapReduce/Hadoop Top 10 解决方案：非唯一键 91
第4章左外连接 96
左外连接示例 96
MapReduce左外连接实现 99
Spark左外连接实现 105
使用leftOuterJoin()的Spark实现 117
第5章反转排序 127
反转排序模式示例 128
反转排序模式的MapReduce/Hadoop实现 129
运行示例 134
第6章移动平均 137
示例1：时间序列数据（股票价格） 137
示例2：时间序列数据（URL访问数） 138
形式定义 139
POJO移动平均解决方案 140
MapReduce/Hadoop移动平均解决方案 143
第7章购物篮分析 155
MBA目标 155
MBA的应用领域 157
使用MapReduce的购物篮分析 157
Spark解决方案 166
运行Spark实现的YARN 脚本 179
第8章共同好友 182
输入 183
POJO共同好友解决方案 183
MapReduce算法 184
解决方案1: 使用文本的Hadoop实现 187
解决方案2: 使用ArrayListOfLongsWritable 的Hadoop实现 189
Spark解决方案 191
第9章使用MapReduce实现推荐引擎 201

- 购买过该商品的顾客还购买了哪些商品 202
- 经常一起购买的商品 206
- 推荐连接 210
- 第10章基于内容的电影推荐 225
- 输入 226
- MapReduce阶段1 226
- MapReduce阶段2和阶段3 227
- Spark电影推荐实现 234
- 第11章使用马尔可夫模型的智能邮件营销 .253
- 马尔可夫链基本原理 254
- 使用MapReduce的马尔可夫模型 256
- Spark解决方案 269
- 第12章 K-均值聚类 282
- 什么是K-均值聚类? 285
- 聚类的应用领域 285
- K-均值聚类方法非形式化描述：分区方法 286
- K-均值距离函数 286
- K-均值聚类形式化描述 287
- K-均值聚类的MapReduce解决方案 288
- K-均值算法Spark实现 292
- 第13章 k-近邻 296
- kNN分类 297
- 距离函数 297
- kNN示例 298
- kNN算法非形式化描述 299
- kNN算法形式化描述 299
- kNN的类Java非MapReduce 解决方案 299
- Spark的kNN算法实现 301
- 第14章朴素贝叶斯 315
- 训练和学习示例 316
- 条件概率 319
- 深入分析朴素贝叶斯分类器 319
- 朴素贝叶斯分类器：符号数据的MapReduce解决方案 322
- 朴素贝叶斯分类器Spark实现 332
- 使用Spark和Mahout 347
- 第15章情感分析 349
- 情感示例 350
- 情感分数：正面或负面 350
- 一个简单的MapReduce情感分析示例 351
- 真实世界的情感分析 353
- 第16章查找、统计和列出大图中的所有三角形 354
- 基本的图概念 355
- 三角形计数的重要性 356
- MapReduce/Hadoop解决方案 357
- Spark解决方案 364
- 第17章 K-mer计数 375
- K-mer计数的输入数据 376
- K-mer计数应用 376
- K-mer计数MapReduce/Hadoop解决方案 377

- K-mer计数Spark解决方案 378
- 第18章 DNA测序 390
 - DNA测序的输入数据 392
 - 输入数据验证 393
 - DNA序列比对 393
 - DNA测试的MapReduce算法 394
- 第19章 Cox回归 413
 - Cox模型剖析 414
 - 使用R的Cox回归 415
 - Cox回归应用 416
 - Cox回归 POJO解决方案 417
 - MapReduce输入 418
 - 使用MapReduce的Cox回归 419
- 第20章 Cochran-Armitage趋势检验 426
 - Cochran-Armitage算法 427
 - Cochran-Armitage应用 432
 - MapReduce解决方案 435
- 第21章 等位基因频率 443
 - 基本定义 444
 - 形式化问题描述 448
 - 等位基因频率分析的MapReduce解决方案 449
 - MapReduce解决方案, 阶段1 449
 - MapReduce解决方案, 阶段2 459
 - MapReduce解决方案, 阶段3 463
 - 染色体X和Y的特殊处理 466
- 第22章 T检验 468
 - 对bioSet完成T检验 469
 - MapReduce问题描述 472
 - 输入 472
 - 期望输出 473
 - MapReduce解决方案 473
 - Spark实现 476
- 第23章 皮尔逊相关系数 488
 - 皮尔逊相关系数公式 489
 - 皮尔逊相关系数示例 491
 - 皮尔逊相关系数数据集 492
 - 皮尔逊相关系数POJO 解决方案 492
 - 皮尔逊相关系数MapReduce解决方案 493
 - 皮尔逊相关系数的Spark 解决方案 496
 - 运行Spark程序的YARN 脚本 516
 - 使用Spark计算斯皮尔曼相关系数 517
- 第24章 DNA碱基计数 520
 - FASTA 格式 521
 - FASTQ 格式 522
 - MapReduce解决方案: FASTA 格式 522
 - 运行示例 524
 - MapReduce解决方案: FASTQ 格式 528
 - Spark 解决方案: FASTA 格式 533
 - Spark解决方案: FASTQ 格式 537

- 第25章 RNA测序 543
 - 数据大小和格式 543
 - MapReduce工作流 544
 - RNA测序分析概述 544
 - RNA测序MapReduce算法 548
- 第26章 基因聚合 553
 - 输入 554
 - 输出 554
 - MapReduce解决方案（按单个值过滤和按平均值过滤） 555
 - 基因聚合的Spark解决方案 567
 - Spark解决方案：按单个值过滤 567
 - Spark解决方案：按平均值过滤 576
- 第27章 线性回归 586
 - 基本定义 587
 - 简单示例 587
 - 问题描述 588
 - 输入数据 589
 - 期望输出 590
 - 使用SimpleRegression的MapReduce解决方案 590
 - Hadoop实现类 593
 - 使用R线性模型的MapReduce解决方案 593
- 第28章 MapReduce和幺半群 600
 - 概述 600
 - 幺半群的定义 602
 - 幺半群和非幺半群示例 603
 - MapReduce示例：非幺半群 606
 - MapReduce示例：幺半群 608
 - 使用幺半群的Spark示例 612
 - 使用幺半群的结论 618
 - 函子和幺半群 619
- 第29章 小文件问题 622
 - 解决方案1：在客户端合并小文件 623
 - 解决方案2：用CombineFileInputFormat解决小文件问题 629
 - 其他解决方案 634
- 第30章 MapReduce的大容量缓存 635
 - 实现方案 636
 - 缓存问题形式化描述 637
 - 一个精巧、可伸缩的解决方案 637
 - 实现LRUMap缓存 640
 - 使用LRUMap的MapReduce解决方案 646
- 第31章 Bloom过滤器 651
 - Bloom过滤器性质 651
 - 一个简单的Bloom过滤器示例 653

《数据算法》

版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:www.tushu000.com