



自己动手做 推荐引擎

[印] 苏雷什·库马尔·戈拉卡拉 (Suresh Kumar Gorakala) 著
左妍 译



BUILDING RECOMMENDATION ENGINES



机械工业出版社
China Machine Press

自己动手做推荐引擎

[印] 苏雷什·库马尔·戈拉卡拉 (Suresh Kumar Gorakala) 著

左妍 译



BUILDING RECOMMENDATION ENGINES



机械工业出版社
China Machine Press

图书在版编目 (CIP) 数据

自己动手做推荐引擎 / (印) 苏雷什·库马尔·戈拉卡拉著; 左妍译. —北京: 机械工业出版社, 2020.1

(自己动手系列)

书名原文: Building Recommendation Engines

ISBN 978-7-111-64108-7

I. 自… II. ①苏… ②左… III. 搜索引擎—程序设计 IV. TP391.3

中国版本图书馆 CIP 数据核字 (2019) 第 243562 号

本书版权登记号: 图字 01-2017-2013

Suresh Kumar Gorakala: Building Recommendation Engines (ISBN: 978-1-78588-485-6).

Copyright © 2016 Packt Publishing. First published in the English language under the title "Building Recommendation Engines".

All rights reserved.

Chinese simplified language edition published by China Machine Press.

Copyright © 2020 by China Machine Press.

本书中文简体字版由 Packt Publishing 授权机械工业出版社独家出版。未经出版者书面许可, 不得以任何方式复制或抄袭本书内容。

自己动手做推荐引擎

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: 刘 锋

责任校对: 殷 虹

印 刷: 中国电影出版社印刷厂

版 次: 2020 年 1 月第 1 版第 1 次印刷

开 本: 186mm × 240mm 1/16

印 张: 16.75

书 号: ISBN 978-7-111-64108-7

定 价: 79.00 元

客服电话: (010) 88361066 88379833 68326294 投稿热线: (010) 88379604

华章网站: www.hzbook.com

读者信箱: hzit@hzbook.com

版权所有·侵权必究

封底无防伪标均为盗版

本书法律顾问: 北京大成律师事务所 韩光 / 邹晓东

| 内容简介 |

本书是一本推荐引擎技术的综合入门指南，详细介绍使用R、Python、Spark、Mahout、Neo4j技术实现诸如协同过滤、基于内容的推荐引擎和情境感知推荐引擎等内容。本书也介绍了行业内广泛使用的各种推荐引擎及其实现。此外，本书还涵盖一些推荐引擎中常用的流行数据挖掘技术，并在最后简要讨论了推荐引擎的未来方向。

本书适合想要使用R、Python、Spark、Neo4j和Hadoop构建复杂预测决策系统及推荐引擎的初学者和有经验的数据科学家阅读。

About the Author 作者简介

Suresh Kumar Corakala 是一位专注于人工智能方向的数据科学家。他拥有近 10 年的专业经验，曾为多个领域的全球客户服务，并帮助他们使用先进的大数据分析技术解决业务问题。他主要从事推荐引擎、自然语言处理、高级机器学习、图数据库等方面的工作。曾参与编写《Building a Recommendation System with R》，由 Packt 出版。他热爱旅行和摄影。

我要感谢我的太太，当我在无数个深夜伏案工作时对我的包容与理解，感谢我的家人对我的支持。我也要向 Barathi Ganesh、Raj Deepthi、Harsh 和我的同事们表示深深的感谢，如果没有他们的支持，这本书很可能不会出版。我要感谢这些年我遇到的所有导师，如果没有这些导师的指点，就没有机会完成这些事情。我还要感谢本书的所有评审人员和产品经理。

技术评审员简介 *About the Reviewers*

Vikram Dhillon 是一位软件开发人员，一位生物信息学研究者，也是中佛罗里达大学 Blackstone LaunchPad 的软件教员，最近在创业，方向是医疗保健中的数据安全。他住在奥兰多，定期会参加开发者聚会和编程马拉松，业余时间探索新技术，比如区块链和开发游戏中的机器学习教程。他参与开源项目已有 5 年多，并在 opsbug.com 上发表关于技术和创业心得的文章。

Vimal Romeo 是罗马 Ernst and Young 的数据科学家，拥有罗马路易斯大学商学院大数据分析硕士学位，并拥有印度 XIME 的 MBA 学位和 CUSAT 计算机科学与工程学士学位。他也是 MilanoR 的作者，该博客主要介绍 R 语言。

我要感谢我的妈妈 Bernadit 和我的兄弟 Vibin，感谢他们一如既往的支持。还要感谢一直关心我的朋友 Matteo Amadei、Antonella Di Luca、Asish Mathew 和 Eleonora Polidoro。特别感谢 Packt 出版社的 Nidhi Joshi 对我的鼓励。

本书是一本推荐引擎技术的综合入门指南，介绍使用 R、Python、Spark、Mahout、Neo4j 技术实现诸如协同过滤、基于内容的推荐引擎和情境感知推荐引擎的内容。本书也介绍了广泛用于行业内的各种推荐引擎及其实现。当然，也涵盖了一些推荐引擎中常用的流行数据挖掘技术，并在本书最后简要讨论了推荐引擎的未来方向。

主要内容

第 1 章为数据科学家提供思路，也为入门者提供指南。该章主要介绍了人们生活中常常遇到的推荐引擎及其优缺点。

第 2 章通过一个构建电影推荐的例子简单介绍了推荐引擎，以便将读者引入到推荐引擎的世界。

第 3 章主要介绍目前流行的不同推荐引擎技术，如基于用户的协同过滤、基于项目的协同过滤、基于内容的推荐引擎、情境感知推荐引擎、混合推荐引擎、使用机器学习模型和数学模型的基于模型的推荐系统。

第 4 章阐述了推荐引擎中使用的各种机器学习技术，如相似度度量、分类、回归和降维技术，并且介绍了用以测试推荐引擎预测能力的评估指标。

第 5 章介绍如何使用 R 和 Python 构建基于用户与基于项目的协同过滤，并介绍了这两种语言中广泛用于构建推荐引擎的库。

第 6 章介绍如何使用 R 和 Python 构建个性化推荐引擎，并介绍用于构建基于内容的推荐系统和情境感知推荐引擎的各种库。

第 7 章介绍使用 Spark 和 MLlib 构建实时推荐系统的基础。

第 8 章引入 GraphDB、Neo4j 的概念和基础知识，以及如何使用 Neo4j 构建实时推荐系统。

第 9 章主要介绍 Hadoop 和 Mahout 在构建可扩展的推荐系统上的构建模块及体系结构，以及使用 Mahout 和 SVD 的详细实现。

第 10 章主要是对之前章节的总结，包括应用于构建决策系统的最佳实践以及对未来的展望。

书中代码的开发环境

为使用 R、Python、Spark、Neo4j、Mahout 进行推荐引擎的不同实现，我们需要以下软件：

章号	软件及版本	下载地址	系统环境
2,4,5	R studio Version 0.99.489	https://www.rstudio.com/products/rstudio/download/	WINDOWS 7+/Centos 6
2,4,5	R version 3.2.2	https://cran.r-project.org/bin/windows/base/	WINDOWS 7+/Centos 6
5,6,7	Anaconda 4.2 for Python 3.5	https://www.continuum.io/downloads	WINDOWS 7+/Centos 6
8	Neo4j 3.0.6	https://neo4j.com/download/	WINDOWS 7+/Centos 6
7	Spark 2.0	https://spark.apache.org/downloads.html	WINDOWS 7+/Centos 6
9	Hadoop 2.5 -Mahout 0.12	http://hadoop.apache.org/releases.html http://mahout.apache.org/general/downloads.html	WINDOWS 7+/Centos 6
7,9,8	Java 7/Java 8	http://www.oracle.com/technetwork/java/javase/downloads/jdk7-downloads-1880260.html	WINDOWS 7+/Centos 6

读者对象

本书适合想要使用 R、Python、Spark、Neo4j 和 Hadoop 构建复杂预测决策系统及推荐引擎的初学者与有经验的数据科学家阅读。

格式约定

新术语和重要词汇用黑体显示。



表示警告或重点提示。



表示技巧。

下载示例代码及彩色图像

本书的示例代码及所有截图和样图，可以从 <http://www.packtpub.com> 通过个人账户下载，也可以访问华章图书官网 <http://www.hzbook.com>，通过注册并登录个人账户下载。

目 录 Contents

作者简介

技术评审员简介

前言

第1章 推荐引擎介绍..... 1

1.1 推荐引擎定义..... 1

1.2 推荐系统的必要性..... 3

1.3 大数据对推荐系统的推动
作用..... 4

1.4 推荐系统类型..... 4

1.4.1 协同过滤推荐系统..... 4

1.4.2 基于内容的推荐系统..... 5

1.4.3 混合推荐系统..... 6

1.4.4 情境感知推荐系统..... 7

1.5 推荐系统技术的发展..... 8

1.5.1 Mahout 在可扩展推荐系统中的
应用..... 8

1.5.2 Apache Spark 在可扩展实时推荐
系统中的应用..... 9

1.6 本章小结..... 12

第2章 构建第一个推荐引擎..... 13

2.1 构建基础推荐引擎..... 14

2.1.1 载入并格式化数据..... 15

2.1.2 计算用户相似度..... 17

2.1.3 为用户预测未知评级..... 18

2.2 本章小结..... 24

第3章 推荐引擎详解..... 25

3.1 推荐引擎的发展..... 26

3.2 基于近邻算法的推荐引擎..... 27

3.2.1 基于用户的协同过滤..... 29

3.2.2 基于项目的协同过滤..... 30

3.2.3 优点..... 32

3.2.4 缺点..... 32

3.3 基于内容的推荐系统..... 32

3.3.1 用户画像生成..... 35

3.3.2 优点..... 36

3.3.3 缺点..... 36

3.4 情境感知推荐系统..... 37

3.4.1 情境定义..... 38

3.4.2 前置过滤法..... 40

3.4.3 后置过滤法..... 40

3.4.4 优点..... 41

3.4.5 缺点..... 41

3.5 混合推荐系统..... 41

3.5.1 加权法..... 42

3.5.2 混合法..... 42

3.5.3 层叠法..... 42

3.5.4 特征组合法..... 42

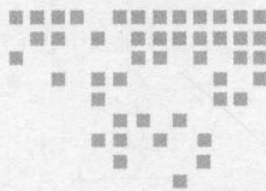
3.5.5 优点..... 43

3.6 基于模型的推荐系统..... 43

3.6.1 概率法	44	5.3 探讨数据集	88
3.6.2 机器学习法	44	5.4 使用 recommenderlab 构建基于	
3.6.3 数学法	44	用户的协同过滤	89
3.6.4 优点	45	5.4.1 准备训练数据和测试数据	90
3.7 本章小结	45	5.4.2 创建一个基于用户的协同	
第 4 章 数据挖掘技术在推荐引擎		模型	90
中的应用	46	5.4.3 在测试集上进行预测	92
4.1 基于近邻算法的技术	47	5.4.4 分析数据集	93
4.1.1 欧氏距离	47	5.4.5 使用 k 折交叉验证评估推荐	
4.1.2 余弦相似度	48	模型	95
4.1.3 Jaccard 相似度	51	5.4.6 评估基于用户的协同过滤	96
4.1.4 皮尔逊相关系数	51	5.5 构建基于项目的推荐模型	99
4.2 数学建模技术	53	5.5.1 构建 IBCF 推荐模型	100
4.2.1 矩阵分解	53	5.5.2 模型评估	103
4.2.2 交替最小二乘法	55	5.5.3 模型准确率度量	104
4.2.3 奇异值分解	55	5.5.4 模型准确率绘图	105
4.3 机器学习技术	57	5.5.5 IBCF 参数调优	107
4.3.1 线性回归	57	5.6 使用 Python 构建协同过滤	110
4.3.2 分类模型	59	5.6.1 安装必要包	110
4.4 聚类技术	69	5.6.2 数据源	110
4.5 降维	71	5.7 数据探讨	111
4.6 向量空间模型	75	5.7.1 表示评级矩阵	113
4.6.1 词频	75	5.7.2 创建训练集和测试集	114
4.6.2 词频 - 逆文档频率	76	5.7.3 构建 UBCF 的步骤	115
4.7 评估技术	78	5.7.4 基于用户的相似度计算	115
4.7.1 交叉验证	79	5.7.5 预测活跃用户的未知评级	116
4.7.2 正则化	80	5.8 使用 KNN 进行基于用户的	
4.8 本章小结	82	协同过滤	117
第 5 章 构建协同过滤推荐引擎	83	5.9 基于项目的推荐	118
5.1 在 RStudio 上安装		5.9.1 评估模型	119
recommenderlab	83	5.9.2 KNN 训练模型	120
5.2 recommenderlab 包中可用的		5.9.3 评估模型	120
数据集	85	5.10 本章小结	120

第6章 构建个性化推荐引擎	121	7.5.4 模型选择和超参数调优	174
6.1 个性化推荐系统	122	7.6 本章小结	179
6.2 基于内容的推荐系统	122	第8章 通过 Neo4j 构建实时推荐	180
6.2.1 构建一个基于内容的推荐系统	123	8.1 图数据库种类	181
6.2.2 使用 R 语言构建基于内容的推荐	123	8.2 Neo4j	183
6.2.3 使用 Python 语言构建基于内容的推荐	133	8.2.1 Cypher 查询语言	184
6.3 情境感知推荐系统	144	8.2.2 节点语法	184
6.3.1 构建情境感知推荐系统	144	8.2.3 关系语法	185
6.3.2 使用 R 语言构建情境感知推荐	145	8.2.4 构建第一个图	185
6.4 本章小结	150	8.3 Neo4j Windows 安装	192
第7章 使用 Spark 构建实时推荐引擎	151	8.4 Neo4j Linux 安装	194
7.1 Spark 2.0 介绍	152	8.4.1 下载 Neo4j	194
7.1.1 Spark 架构	152	8.4.2 设置 Neo4j	195
7.1.2 Spark 组件	154	8.4.3 命令行启动 Neo4j	195
7.1.3 Spark Core	154	8.5 构建推荐引擎	197
7.1.4 Spark 的优点	156	8.5.1 将数据加载到 Neo4j	197
7.1.5 Spark 设置	156	8.5.2 使用 Neo4j 生成推荐	200
7.1.6 SparkSession	157	8.5.3 使用欧氏距离进行协同过滤	201
7.1.7 弹性分布式数据集	158	8.5.4 使用余弦相似度进行协同过滤	206
7.1.8 关于 ML 流水线	158	8.6 本章小结	209
7.2 使用交替最小二乘法进行协同过滤	160	第9章 使用 Mahout 构建可扩展的推荐引擎	210
7.3 使用 PySpark 构建基于模型的推荐系统	162	9.1 Mahout 简介	211
7.4 MLlib 推荐引擎模块	163	9.2 配置 Mahout	211
7.5 推荐引擎方法	164	9.2.1 Mahout 单机模式	211
7.5.1 实现	164	9.2.2 Mahout 分布式模式	218
7.5.2 基于用户的协同过滤	172	9.3 Mahout 的核心构建模块	220
7.5.3 模型评估	173	9.3.1 基于用户的协同过滤推荐引擎组件	220
		9.3.2 使用 Mahout 构建推荐引擎	223
		9.3.3 数据描述	223

9.3.4 基于用户的协同过滤	225	10.2 推荐系统的发展阶段	243
9.4 基于项目的协同过滤	228	10.2.1 一般的推荐系统	243
9.5 协同过滤评估	231	10.2.2 个性化推荐系统	244
9.6 基于用户的推荐评估	231	10.2.3 未来的推荐系统	245
9.7 基于项目的推荐评估	232	10.2.4 下一个最佳举措	249
9.8 SVD 推荐系统	235	10.2.5 使用案例	249
9.9 使用 Mahout 进行分布式 推荐	236	10.3 流行方法	251
9.10 可扩展系统的架构	240	10.4 推荐引擎的时效性	252
9.11 本章小结	241	10.4.1 A/B 测试	253
第 10 章 推荐引擎的未来	242	10.4.2 反馈机制	254
10.1 推荐引擎的未来	242	10.5 本章小结	255



推荐引擎介绍

我们是如何决定购买某些商品的呢？在日常生活中，我们在决定购买之前会询问朋友或者家人；网购时，我们会阅读匿名用户对产品的评论，对产品进行比较，之后才会决定是否购买。当今互联网中的信息正以指数级速度增长，寻找有效的信息将是一项挑战，而增加用户对检索结果的信心则是更大的挑战。

推荐系统的出现可以为用户提供优质的相关信息或重要信息，一些大型网站的参与则更好地推动了推荐引擎的普及。例如，亚马逊的商品推荐，Facebook 的人脉圈推荐，Twitter、LinkedIn、YouTube 上的视频推荐，谷歌新闻推荐等。这些案例的成功为诸如旅游、医疗健康、金融等领域的其他行业打开了一扇窗。

1.1 推荐引擎定义

推荐引擎是从信息检索和人工智能派生出来的技术，是分析巨量数据（特别是产品和用户信息）强有力的工具，能够基于数据挖掘提供相关推荐。

在技术上，推荐引擎要求开发一个数学模型或者目标函数，用来预测用户对商品的喜欢程度。

假设 $U = \{\text{用户}\}$ ， $I = \{\text{商品}\}$ ， F 表示目标函数，通过 F 计算 I 对 U 的有用性，表示成：

$$F: U \times I \rightarrow R, \text{ 其中 } R = \{\text{推荐商品}\}$$

对于每位用户 u ，我们要选择使目标函数最大的商品 i ，表示成：

$$u \in U, I'_u = \operatorname{argmax}_u(u, i)$$

推荐系统的主要目的是在用户做决定的时候提供相关推荐，以便在线用户从网络上的大量可用选项中做出更好的决定。一个优秀的推荐系统更趋向于个性化推荐，它可以通过收集用户有价值的数字足迹（如人口统计、事务细节、交互日志）和关于产品的信息（例如规格、用户反馈、与其他产品比较等），来完成推荐之前的数据分析。



用户标签

推荐引擎最大的挑战是如何给参与者提供好的推荐，一个好的推荐系统会充分考虑消费者信息（如上图所示用户标签）和销售信息。从消费者角度来说，一个有价值的推荐可以影响消费者的最终决定，所以推荐的动机是考虑如何增强消费者的购买信心。从商家的角度来说，为不同层次的消费者提供个性化推荐，做到精准营销是十分必要的。伴随着网购的兴起壮大，很多公司通过收集大量用户交互信息日志，对用户行为进行越来越深入的分析。此外，对推荐引擎的实时性也有更高的要求。随着技术和研究的进步，将大数据分析和人工智能技术加入推荐引擎也是一个挑战。下图显示了推荐引擎的一些应用。



1.2 推荐系统的必要性

我们知道构建推荐系统是非常复杂的，它需要大量的工作量和人员、技术、资金的投入。如此巨大的投入值得吗？来看一些事实：