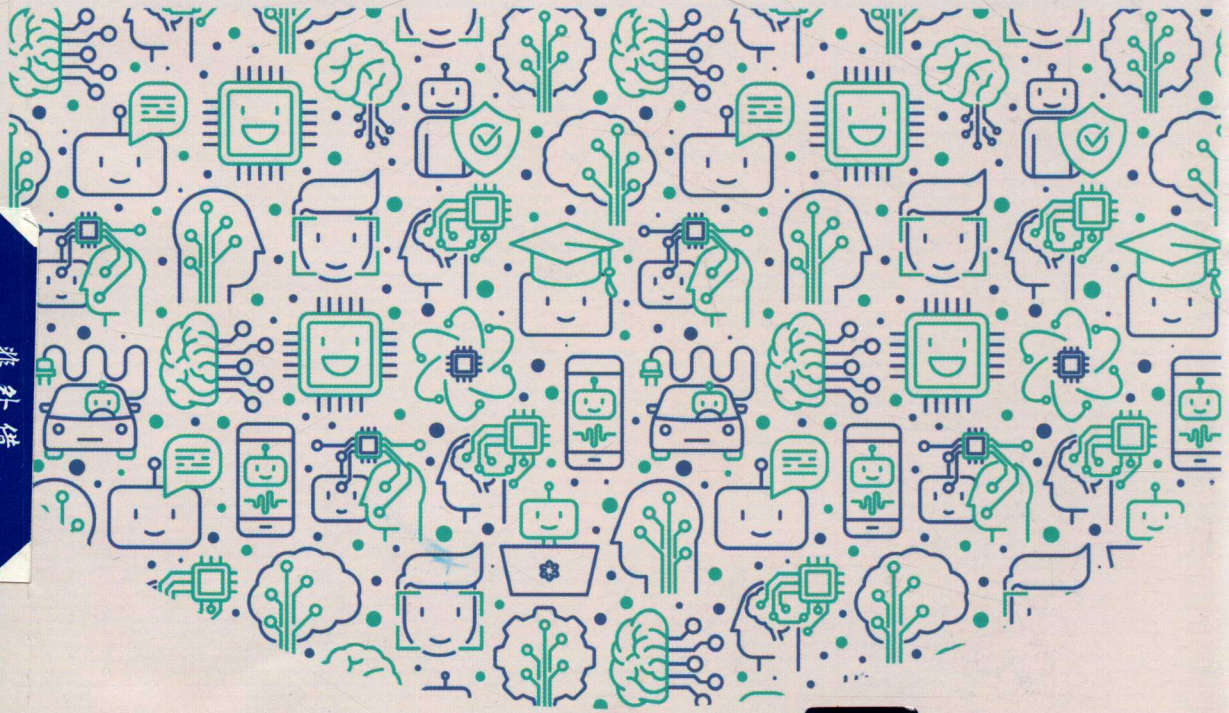
机器学习算法实践

推荐系统的协同过滤理论及其应用

王建芳 著

MACHINE LEARNING



清華大學出版社

机器学习算法实践

推荐系统的协同过滤理论及其应用

王建芳 著

清华大学出版社
北京

内 容 简 介

个性化推荐能够根据用户的历史行为显式或者隐式地挖掘用户潜在的兴趣和需求，并为其推送个性化信息，因此受到研究者的追捧及工业界的青睐，其研究具有重大的学术价值及商业应用价值，已广泛应用于大型电子商务平台、社交平台、新闻客户端以及其他各类旅游和娱乐类网站中。

本书内容丰富，较全面地介绍了基于协同过滤的推荐系统存在的问题、解决方法和评估策略，主要内容涉及协同过滤推荐算法中的时序技术、矩阵分解技术和社交网络信任技术等知识。

本书可供从事推荐系统、人工智能、机器学习、模式识别和信息检索等领域的科研人员及研究生阅读、参考。

图书在版编目(CIP)数据

机器学习算法实践：推荐系统的协同过滤理论及其应用/王建芳著. —北京：清华大学出版社，2018

ISBN 978-7-302-50783-3

Ⅰ. ①机… Ⅱ. ①王… Ⅲ. ①机器学习－算法 Ⅳ. ①TP181

中国版本图书馆 CIP 数据核字(2018)第 178631 号

责任编辑：曾　珊
封面设计：常雪影
责任校对：焦丽丽
责任印制：李红英

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954
印 装 者：三河市国英印务有限公司
经　　销：全国新华书店
开　　本：170mm×240mm　　**印　　张**：13　　**字　　数**：270 千字
版　　次：2018 年 11 月第 1 版　　**印　　次**：2018 年 11 月第 1 次印刷
定　　价：69.00 元

产品编号：080054-01

前言

PREFACE

个性化推荐与信息检索技术的目标一致，也是一种帮助用户更快速地发现有用信息的工具，但与信息检索技术不同的是，个性化推荐能够根据用户的历史行为显式或者隐式地挖掘用户潜在的兴趣和需求，为其推送感兴趣并且个性化的信息，已越来越受到研究者的追捧及工业界的青睐，其研究具有重大的学术价值及商业应用价值。如今基于个性化推荐算法的推荐系统已广泛应用于大型电子商务平台（如天猫、京东和亚马逊等）、社交平台（如新浪微博、Facebook 和 Twitter 等）、新闻客户端（今日头条、天天快报等）以及其他各类旅游和娱乐类网站（如携程网、电影音乐社区等）中，在提高用户满意度和忠诚度的同时也为自身带来了可观的经济效益。

协同过滤推荐算法是个性化推荐中运用最早和最成功的一种推荐技术，它的任务是利用用户与项目评分矩阵中的已知元素来预测未知元素的评分值并将预测评分高的项目推荐给用户。协同过滤的最大优点是对推荐对象没有特殊的要求，能处理非结构化的复杂对象（如音乐、图书、电影和资讯类新闻内容等，这类产品是难以进行机器自动内容分析的信息），避免了内容分析的不完全和不精确，而且能够根据用户的历史行为推荐个性化的信息。传统的基于邻域模型的推荐算法分为数据收集（输入）、获得最近邻集合（主要是计算相似度）和预测并推荐（输出）等步骤。目前协同过滤推荐算法还存在数据的高维稀疏性、冷启动和大数据环境下扩展性等制约其进一步发展的瓶颈问题，如何解决以上问题进而提高推荐系统的推荐质量成为个性化推荐的关键，近年来基于协同过滤的推荐算法及其相关改进模型得到了学者们的广泛关注和研究。

本书作者一直从事推荐系统理论及其应用的研究工作，提出了一系列改进推荐质量的方法，并成功应用于多种复杂的实际问题。作者的这些工作大大丰富了推荐系统理论，尤其是所关注的协同过滤推荐算法对其在其他领域的进一步研究与应用奠定了技术基础，具有重要的理论意义和实际应用价值。

本书由河南理工大学计算机科学与技术学院王建芳独立完成，是作者在本领域所发表学术论文的基础上进一步加工、深化而成的，是对已有研究成果的全面总结。全书共分 5 篇 14 章。第一篇包括第 1 章，讨论了推荐算法的分类、各类算法的基本思想和改进策略，阐述推荐算法存在的问题、实验方法和评测指标。第二篇

包括第 2 章和第 3 章，主题是围绕基于时序的协同过滤推荐算法展开研究。在推荐系统中随着时间的推移，用户的关注点在不断变化，如何捕获这一动态的时间效应是个难题。本篇针对基于时序的协同过滤推荐算法展开研究。第三篇包括第 4～11 章，主题是围绕基于矩阵分解的协同过滤推荐算法展开研究。矩阵分解模型能够基于用户的行为对用户和项目进行自动分析，也就是把用户和项目划分到不同主题，这些主题可以理解为用户的兴趣和项目属性。本篇针对 SVD、概率矩阵分解、非负矩阵分解及其与相关算法的整合分别提出相关的理论。第四篇包括第 12 章和第 13 章，主题是围绕协同过滤推荐算法与社交网络的信任展开研究，将用户的评分信息和用户的社交网络信息融入传统的矩阵分解中以提高推荐质量。第五篇包括第 14 章，从实际应用的角度用 Spark 实现一个基于矩阵分解的推荐原型系统。

在本书的撰写过程中，已毕业的硕士研究生张朋飞、李骁、武文琪以及在读研究生谷振鹏、刘冉东、苗艳玲等对书稿内容和相关实验提供了大量的帮助，在此向他们表示衷心的感谢。本书的出版得到河南省高等学校重点科研项目（项目编号：15A520074）和河南理工大学博士基金的支持，在此一并表示感谢。

推荐系统所涉及的算法，尤其是协同过滤推荐算法是一个快速发展、多学科交叉的新颖研究方法，其理论及应用均有大量的问题尚待进一步深入研究。由于作者知识水平和资料获取方面的限制，书中不妥之处在所难免，敬请同行专家和读者批评指正。

作　者

2018 年 5 月

目录

CONTENTS

第一篇 基础理论

第二篇　基于时序的协同过滤推荐算法

第三篇 基于矩阵分解的协同过滤推荐算法

第一篇　基础理论

推荐系统的传统定义可以理解为“采集用户历史行为信息，结合具体推荐模型帮助用户选择商品或提供建议的过程”。现阶段完整的个性化推荐模型主要由数据收集及预处理、推荐算法和产生推荐三部分组成。

数据收集包括收集用户属性、项目属性和用户对项目的行为信息等。收集到的数据中，有些数据无法直接使用或推荐效果很差。为了后续更好地为用户提供推荐服务，需要提前对数据进行预处理——清理和减噪。

产生推荐是通过推荐算法计算得到目标用户的最近邻集合，将最近邻评价过的项目推荐给目标用户；利用模型对未知项目进行预测，将预测评分最高的项目推送给目标用户。

推荐算法作为个性化推荐系统中的核心，将收集并处理好的数据通过推荐算法为用户产生推荐。推荐算法的优劣与个性化推荐系统的推荐质量有着直接关系。

第一篇 基础理论

第1章

理论入门

1.1 引言

信息技术的迅猛发展使人类社会由信息匮乏时代进入信息过载时代，而信息过载为用户在选择最中意的产品时带来沉重的处理负担。以电子商务网站为例，用户往往囿于潜在需求而无法用关键字表达或者搜索关键字表达不准确，从而不得不从浩如烟海的信息海洋获取真正需求的信息。

针对上述问题，为满足用户和企业的共同需求，满足不同用户偏好的推荐系统应运而生。此外，社会经济的快速发展带来种类繁多的产品类型，使得用户的购买目的更多地体现出固有的个体特性，在满足物质需求的基础上，推荐系统根据用户的历史行为，例如点击、购买和收藏等去挖掘用户的偏好信息，进而进行个性化推荐。为增加用户的黏性，越来越多的网站和社区开始采用推荐系统为用户提供个性化的优质服务。同时，随着 Web 3.0 时代的到来以及“互联网＋”理念的提出，人们越来越意识到推荐系统的重要性并纷纷投入其中。例如，亚马逊、eBay、天猫、京东等电子商务网站、Facebook、Twitter 和新浪微博等社交媒体均纷纷在原有业务的基础上增加推荐功能。事实表明，推荐系统的融入显著提高了用户的满意度和对网站的黏性，进而为其自身带来了可观的经济效益和社会影响力。

不过，单纯地以用户和项目为驱动的推荐引擎并不能满足相关用户的实际需要，用户在实际购买之中往往会结合自己的实际需要以及相关朋友（本书称为社交网络信息）的推荐来做最终选择，同时传统推荐算法往往带有很严重的“马太效应”。也就是说，推荐的商品往往都是热门的商品，因此造成热门的商品更加热门，而处在“长尾分布”上的商品仍得不到重视。为此，将社交网络与个性化推荐相结合提高推荐的精确度是近年来的研究热点。

在海量数据的虚拟环境下，电影网站提供的节目信息非常多，例如按演员来说，每天都会更新该演员出演的电影，包括蓝光、高清、标清和流畅等，这样每天网站上的数据量都有成千上万太字节（1TB＝1024GB），而仅仅通过一台微型计算机

或手机屏幕，希望观众找到一个自己真正喜欢的电影是不可能的。因此，社区或网站提供了一些智能导购的需要。例如京东的 JIMI，根据用户的兴趣推荐用户可能感兴趣的物品，用户可以很容易地找到他们所需要的或感兴趣但不容易得到的明确的项目。而且，从实际情况来看，用户的需求往往是对商品或事件的兴趣，但目前还不清楚什么商品可以满足其潜在需求。这时，如果商家基于用户的历史行为分析出其感兴趣的信息并将这些信息呈现到用户面前，就可以把用户的潜在需求转化为现实的需求，从而给用户带来惊喜。

1.2 推荐系统的形式化定义

目前推荐系统常采用的方法主要有基于内容的推荐、基于网格的推荐、基于上下文情景的推荐和协同过滤推荐。协同过滤(Collaborative Filtering，CF)推荐技术是推荐系统中最为常用且有效的方法，可分为基于内存的协同过滤和基于模型的协同过滤，前者根据用户或者项目的相似度选出与目标用户最相似的若干用户的评分来对未评分的项目进行评分预测；后者通过分析用户和项目的内部规律，预测用户对项目的偏好，其中概率矩阵分解技术是其典型代表。目前概率矩阵分解技术还存在数据的高维稀疏性和海量数据环境下的扩展性等制约其进一步发展的瓶颈问题。如何解决以上问题进而提高推荐系统的推荐质量成为个性化推荐的关键。

一个典型的电影推荐系统一般包括含有 N 个用户的用户集合 $U=\{u_1,u_2,u_3,\cdots,u_N\}$ 和含有 M 个项目的项目集合 $I=\{i_1,i_2,i_3,\cdots,i_M\}$，每个用户 $u_i\in U$ 评价了 I 中的一部分项目，评价过的项目用 $I_{u_i}\subseteq I$ 表示，用户的打分记录往往表示成 R_{NM}，如式(1-1)所示。

$$R_{NM}=\begin{bmatrix} r_{11} & \cdots & r_{1k} & \cdots & r_{1M} \\ \vdots & & \vdots & & \vdots \\ r_{21} & \cdots & r_{2k} & \cdots & r_{2M} \\ \vdots & & \vdots & & \vdots \\ r_{N1} & \cdots & r_{Nk} & \cdots & r_{NM} \end{bmatrix} \tag{1-1}$$

式中，矩阵[①]中每一行 r_i——用户 i 评价过的电影集合，所有用户集合用 U 表示；

每一列 r_j——评价电影 j 的用户集合，所有电影集合用 V 表示；

每一个元素 r_{ij}——用户 i 对电影 j 的评分，通常 r_{ij} 的取值为 1～5 的整数，数据越大表示用户对该项目越满意。

实际中 R_{NM} 非常稀疏，例如 Ciao 数据集中已有的评分数目所占比例不足 1%，因此传统推荐算法的质量才会特别差。

① 注：本书中的矩阵、向量用斜体表示，而不用黑体表示。全书统一。