推荐系统实践 (图灵原创 5)

作者：项亮

版权信息

书名：推荐系统实践

作者：项亮， 陈义， 王益

ISBN：978-7-115-28158-6

目录

序 一

推荐在今天互联网的产品和应用中被广泛采用，包括今天大家经常使用的相关搜索、话题推荐、电子商务的各种产品推荐、社交网络上的交友推荐等。但是，至今还没有一本书从理论上对它进行系统地分析和论述。《推荐系统实践》这本书恰恰弥补了这个空白。

该书总结了当今互联网主要领域、主要公司、各种和推荐有关的产品和服务，包括：

- 亚马逊的个性化产品推荐；
- Netflix的视频和DVD推荐；
- Pandora的音乐推荐；
- Facebook的好友推荐；
- Google Reader的个性化阅读；
- 各种个性化广告。

书的名称虽然是《推荐系统实践》，但作者也阐述了和推荐系统有关的理论基础和评价推荐系统优劣的各种标准与方法，比如覆盖率、满意度、AB测试等。由于这些评估很大程度上取决于对用户行为的分析，因此本书也介绍了用户行为分析方法，并且给出了计算机实现的算法。

本书对有兴趣自己开发推荐系统的读者给出了设计和实现推荐系统的方法与技巧，非常具有指导意义。

本书文笔流畅，可读性较高，是一部值得推荐给IT从业人员的优秀参考书。

吴军

腾讯副总裁，谷歌资深研究员，《数学之美》和《浪潮之巅》作者

序 二

项亮的书写完了。开始写作这本书时，我的身份是作者，但交稿时，我变成了审稿人。这让我想起了多年前流传的一个“四大傻”的段子：炒房炒成房东，炒股炒成股东，……写书写成审稿人，我看也可以并肩成为一景。

去年五六月份，图灵公司的杨海玲老师通过朋友问我有没有兴趣参与写一本推荐系统方面的书，我欣然答应。近几年推荐技术在互联网领域的应用越来越广泛，但对相关技术做系统介绍的书却非常少，相关的外文书倒是见过两三本。但一方面，对国内读者来说语言障碍或多或少会是个问题，另一方面，这些书大多以研究人员为目标读者，并不完全适合推荐技术的普及。能参与填补这项空白，何乐而不为？书开写后的最初一两个月，我的确贡献过不到万把字的内容，但随着各种不足为外人道的事务纷至沓来，能花在写作上的时间越来越少，每次答应项亮要去填补内容，最后都不了了之，一直到项亮自己把这本书写完。我最初贡献的内容，也因为写作目标和本书整体风格的逐步调整没法添加进来了。这种情况下，我实在不好意思呆在作者列表里了，所以有机会写了这篇序。

提到项亮，就不能不提Netflix推荐算法竞赛，虽然项亮自己不见得喜欢把自己定格在过去时。这项赛事，非常罕见地召集了数以万计的技术人员共同解决同一个技术问题，并且把解决方案公布出来。这为这个领域的工程人员和研究人员不同创意的碰撞提供了条件，因而产生了很多有价值的新方法，使很多以前只被少数专家掌握的技术细节能够被更广泛地传播开来，使专家们解读数据的方法、解构算法模型的思路能够被巨细无遗地发表出来。项亮在Netflix竞赛中有非常出色的表现，书中总结了很多他在Netflix竞赛以及相关研究和工程工作中学到或悟到的分析数据与设计算法的思路。虽然我一直在追踪推荐技术的发展，在书中仍然能看到很多本不了解的方法，相信其他读者读过本书也不会失望。

在大家一起讨论的过程中，项亮经常提到另外一本非常流行的书，即《集体智慧编程》。项亮非常希望他写的书能像《集体智慧编程》那样简明实用，帮助那些对推荐技术或数据挖掘原理完全不了解的读者快速实现自己的推荐系统。出于这个目的，本书尽可能地用代码和图表与读者交流，尽可能地用直观的讨论代替数学公式，这对于大多数工程技术人员来说应该是更为喜闻乐见的形式。另一方面，可能是因为数据资源的限制，大多数学术论文都把推荐问题看做评分预测问题，而实际应用中最常见的是TopN推荐，虽然TopN推荐问题可以归纳成评分问题，但并不是每种评分预测算法都能直接用来解决TopN推荐问题。本书大部分篇幅都在讨论TopN推荐问题，这样的安排对实际应用的实现应该帮助会更大一点。最后，本书比较系统地讨论了把推荐技术应用到真实应用场景时最常遇到的问题，希望可以帮助那些有机器学习经验的技术人员快速了解推荐技术。

最近一两年，国内大型互联网公司对个性化服务越来越重视，以个性化技术做支撑的创业公司也在不断涌现，个性化的浪潮方兴未艾，相信本书能帮助更多的技术人员投身于这一技术浪潮。能看到本书的诞生，我深感荣幸，虽然我的贡献，其实只有这篇序。

陈义

豆瓣资深算法工程师

序 三

翻翻我的邮箱，可以看到2010年6月就有项亮组织大家讨论《推荐系统实践》一书目录结构的记录。实际上最初的讨论比这还早，而且从北京初夏难得一见的暴雨砸在咖啡馆的玻璃窗上开始，一直持续到了金秋时节。讨论的焦点在于为什么要写一本关于推荐系统的书、从什么角度写以及写给谁看。

第一个问题相对好回答。推荐系统是目前互联网世界最常见的智能产品形式。从电子商务、音乐视频网站，到作为互联网经济支柱的在线广告和新颖的在线应用推荐，到处都有推荐系统的身影。这些网站和业务的开创者大都是年轻热情的工程师，或者有志于投身互联网行业的同学。虽然我们并非都有相关学术研究的背景，也并非都有在企业中积累的经验，但是大家都不乏学习的热情，而且充满着对研发成功推荐系统的期待。因此参与讨论的朋友都赞同从实践者的角度来写这本书，写给希望一起学习和实践的朋友们。讨论并不是空想。在此期间，项亮建立了一个wiki系统，样章一发布在上面，一些朋友就开始修改。经过将近一年的努力，我们看到了本书的初稿。

初识项亮是在2009年，当时项亮还是中国科学院的一名博士研究生，一方面积极参与Netflix和其他推荐系统比赛并取得了漂亮的成绩，一方面积极参与组织了recsys学术会议。作为一个有很多业界公司支持的学术交流活动，recsys在建立之初就吸引了很多同学和工程师。项亮毕业后进入Hulu公司，开始了工业级别推荐系统的开发工作，并一如既往地注意学习、总结和分享。我在recsys做了一次关于并行机器学习技术的报告后，项亮介绍我认识了本书的几位主要贡献者。随后不久，大家就开始酝酿本书的写作。项亮的经历在很大程度上决定了本书的写作目标：希望帮助在校学生了解推荐系统的业界起源和应用，把握研究方向；帮助工程师总结各类方法，迅速开发出一个推荐系统并持续优化之。

推荐系统是一个很大的话题。各种在线甚至部分离线应用中，都有各式各样目标不一的推荐系统，小到论文推荐，大到用户兴趣定向的在线广告系统。在学术圈，相关的研究成果亦可谓多矣。实际上，几周前大家还在讨论最新的机器学习方法可能给推荐系统带来的变化。可是，本书不论是写成一本学术专著，还是一部产品大全，都难免浩瀚空泛的尴尬，对大家难有帮助。因此，作者花费了大量精力在组织目录结构上，希望覆盖推荐系统的若干重要问题，同时让每个问题下既有实际产品介绍，也有技术思路介绍。为了保证可读性，本书重在常见方法和技术思路，而非全面介绍各种思想和最新研究成果。为了保证可操作性，重要的算法都配有 Python 语言的示例程序。

我想，这本实践者写给实践者的书，留下的是作者对“思考”和“学习”的辩证足迹。我希望本书的出版能带动更多的朋友一起把足迹走成大路，而大路的前方，是更多成功的互联网应用和完美的技术方法。

王益

腾讯公司情境广告中心总监

前言

说起本书，还要追溯到2010年3月份的ResysChina推荐系统大会。在那次会议上，我遇到了刘江老师。刘老师看过我之前写的一些推荐系统方面的博客，希望我能总结总结，写本简单的书。当时国内还没有推荐系统方面的书，而国外已经有这方面的专业书了，因此图灵公司很想出版一本介绍推荐系统的书。所以，去年7月博士毕业时，我感觉有时间可以总结一下这方面的工作了，于是准备开始写这本书。

写这本书的目的有下面几个。首先，从个人角度讲，虽然写博士论文时已经总结了读博期间在推荐系统方面的工作，但并没有全部涉及整个推荐系统的各个方面，因此我很希望通过写作这本书全面地阅读一下相关的文献，并在此基础上总结一下推荐系统各个方面的发展现状，供大家参考。其次，最近几年从事推荐系统研究的人越来越多，这些人中有些原来是工程师，对机器学习和数据挖掘不太了解，有些是在校学生，虽然对数据挖掘和机器学习有所了解，却对业界如何实现推荐系统不太清楚。因此，我希望能够通过本书让工程师了解推荐系统的相关算法，让学生了解如何将自己了解的算法实现到一个真实的工业系统中去。

一般认为，推荐系统这个研究领域源于协同过滤算法的提出。这么说来，推荐系统诞生快20年了。这期间，很多学者和公司对推荐系统的发展起到了重要的推动作用，各种各样的推荐算法也层出不穷。本书希望将这20年间诞生的典型方法进行总结。但由于方法太多，这些方法的归类有很多不同的方式。比如，可以按照数据分成协同过滤、内容过滤、社会化过滤，也可以按照算法分成基于邻域的算法、基于图的算法、基于矩阵分解或者概率模型的算法。为了方便读者入门，本书基本采用数据分类的方法，每一章都介绍了一种可以用于推荐系统设计的、新类型的用户数据，然后介绍如何通过各种方法利用该数据，最后在公开数据集上评测这些方法。当然，不是所有数据都有公开的数据集，并且不是所有算法都可以进行离线评测。因此，在遇到没有数据集或无法进行离线评测的问题时，本书引用了一些著名学者的实验结果来说明各种方法的效果。

为了使本书同时适合工程师和在校学生阅读，本书在写作中同时使用了两种介绍方法。一种是利用公式，这样方便有一些理论基础的同学很快明白算法的含义。另一种是利用代码，这样可以方便工程师迅速了解算法的含义。不过因为本人是学生出身，工程经验还不是特别足，所以有些代码写得不是那么完美，还请工程师们海涵。

本书一开始写的时候有3位作者，除了我之外还有豆瓣的陈义和腾讯的王益。他们两位都是这方面的前辈，在写作过程中提出了很多宝贵的意见。但因为二位工作实在太繁忙，所以本书主要由我操刀。但书中的很多论述融合了大家的思想和经验，是我们很多次讨论的结果。因此在这里感谢王益和陈义二位合作者，虽然二位没有动笔，但对这本书做出了很大的贡献。

其次，还要感谢吴军老师和谷文栋为本书作序。感谢谷文栋、稳国柱、张夏天各自审阅了书中部分内容，提出了很多宝贵的意见。感谢我在Hulu的同事郑华和李航，郑华给了我充分的时间完成这本书，对这本书能够按时出版功不可没，而李航审阅了书中的部分内容，提出了很多有价值的修改意见。

最后感谢我的父母和妻子，他们在我写作过程中给予了很大照顾，感谢他们的辛勤付出。