TURING 图灵原创

推荐系统实践

项亮 编著　　陈义 王益 审校

人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵原创

推荐系统实践

项亮 编著　　陈义 王益 审校

人民邮电出版社
北　京

图书在版编目（CIP）数据

推荐系统实践 / 项亮编著. -- 北京 : 人民邮电出版社, 2012.6
（图灵原创）
ISBN 978-7-115-28158-6

Ⅰ. ①推… Ⅱ. ①项… Ⅲ. ①计算机网络 Ⅳ. ①TP393

中国版本图书馆CIP数据核字(2012)第095118号

内 容 提 要

本书通过大量代码和图表全面系统地阐述了和推荐系统有关的理论基础，介绍了评价推荐系统优劣的各种标准（比如覆盖率、满意度）和方法（比如AB测试），总结了当今互联网领域中各种和推荐有关的产品和服务。另外，本书为有兴趣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，并解答了在真实场景中应用推荐技术时最常遇到的一些问题。

本书适合对推荐技术感兴趣的读者学习参考。

图灵原创

推荐系统实践

◆ 编著 项 亮
审校 陈 义 王 益
责任编辑 毛倩倩

◆ 人民邮电出版社出版发行 北京市崇文区夕照寺街14号
邮编 100061 电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
三河市海波印务有限公司印刷

◆ 开本：800×1000 1/16
印张：13.5
字数：319千字 2012年6月第1版
印数：1－5 000册 2012年6月河北第1次印刷

ISBN 978-7-115-28158-6

定价：49.00元

读者服务热线：(010)51095186转604 印装质量热线：(010)67129223
反盗版热线：(010)67171154

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，《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。

目　录

第 1 章　好的推荐系统 …… 1
1.1　什么是推荐系统 …… 1
1.2　个性化推荐系统的应用 …… 4
1.2.1　电子商务 …… 4
1.2.2　电影和视频网站 …… 8
1.2.3　个性化音乐网络电台 …… 10
1.2.4　社交网络 …… 12
1.2.5　个性化阅读 …… 15
1.2.6　基于位置的服务 …… 16
1.2.7　个性化邮件 …… 17
1.2.8　个性化广告 …… 18
1.3　推荐系统评测 …… 19
1.3.1　推荐系统实验方法 …… 20
1.3.2　评测指标 …… 23
1.3.3　评测维度 …… 34

第 2 章　利用用户行为数据 …… 35
2.1　用户行为数据简介 …… 36
2.2　用户行为分析 …… 39
2.2.1　用户活跃度和物品流行度的分布 …… 39
2.2.2　用户活跃度和物品流行度的关系 …… 41
2.3　实验设计和算法评测 …… 41
2.3.1　数据集 …… 42
2.3.2　实验设计 …… 42
2.3.3　评测指标 …… 42
2.4　基于邻域的算法 …… 44
2.4.1　基于用户的协同过滤算法 …… 44
2.4.2　基于物品的协同过滤算法 …… 51
2.4.3　UserCF 和 ItemCF 的综合比较 …… 59
2.5　隐语义模型 …… 64
2.5.1　基础算法 …… 64
2.5.2　基于 LFM 的实际系统的例子 …… 70
2.5.3　LFM 和基于邻域的方法的比较 …… 72
2.6　基于图的模型 …… 73
2.6.1　用户行为数据的二分图表示 …… 73
2.6.2　基于图的推荐算法 …… 73

第 3 章　推荐系统冷启动问题 …… 78
3.1　冷启动问题简介 …… 78
3.2　利用用户注册信息 …… 79
3.3　选择合适的物品启动用户的兴趣 …… 85
3.4　利用物品的内容信息 …… 89
3.5　发挥专家的作用 …… 94

第 4 章　利用用户标签数据 …… 96
4.1　UGC 标签系统的代表应用 …… 97
4.1.1　Delicious …… 97
4.1.2　CiteULike …… 98
4.1.3　Last.fm …… 98
4.1.4　豆瓣 …… 99
4.1.5　Hulu …… 99
4.2　标签系统中的推荐问题 …… 100
4.2.1　用户为什么进行标注 …… 100
4.2.2　用户如何打标签 …… 101
4.2.3　用户打什么样的标签 …… 102
4.3　基于标签的推荐系统 …… 103
4.3.1　实验设置 …… 104
4.3.2　一个最简单的算法 …… 105
4.3.3　算法的改进 …… 107
4.3.4　基于图的推荐算法 …… 110
4.3.5　基于标签的推荐解释 …… 112
4.4　给用户推荐标签 …… 115
4.4.1　为什么要给用户推荐标签 …… 115

4.4.2 如何给用户推荐标签 …… 115
4.4.3 实验设置 …… 116
4.4.4 基于图的标签推荐算法 …… 119
4.5 扩展阅读 …… 119

第 5 章 利用上下文信息 …… 121
5.1 时间上下文信息 …… 122
5.1.1 时间效应简介 …… 122
5.1.2 时间效应举例 …… 123
5.1.3 系统时间特性的分析 …… 125
5.1.4 推荐系统的实时性 …… 127
5.1.5 推荐算法的时间多样性 …… 128
5.1.6 时间上下文推荐算法 …… 130
5.1.7 时间段图模型 …… 134
5.1.8 离线实验 …… 136
5.2 地点上下文信息 …… 139
5.3 扩展阅读 …… 143

第 6 章 利用社交网络数据 …… 144
6.1 获取社交网络数据的途径 …… 144
6.1.1 电子邮件 …… 145
6.1.2 用户注册信息 …… 146
6.1.3 用户的位置数据 …… 146
6.1.4 论坛和讨论组 …… 146
6.1.5 即时聊天工具 …… 147
6.1.6 社交网站 …… 147
6.2 社交网络数据简介 …… 148
社交网络数据中的长尾分布 …… 149
6.3 基于社交网络的推荐 …… 150
6.3.1 基于邻域的社会化推荐算法 …… 151
6.3.2 基于图的社会化推荐算法 …… 152
6.3.3 实际系统中的社会化推荐算法 …… 153
6.3.4 社会化推荐系统和协同过滤推荐系统 …… 155
6.3.5 信息流推荐 …… 156
6.4 给用户推荐好友 …… 159
6.4.1 基于内容的匹配 …… 161
6.4.2 基于共同兴趣的好友推荐 …… 161
6.4.3 基于社交网络图的好友推荐 …… 161
6.4.4 基于用户调查的好友推荐算法对比 …… 164
6.5 扩展阅读 …… 165

第 7 章 推荐系统实例 …… 166
7.1 外围架构 …… 166
7.2 推荐系统架构 …… 167
7.3 推荐引擎的架构 …… 171
7.3.1 生成用户特征向量 …… 172
7.3.2 特征-物品相关推荐 …… 173
7.3.3 过滤模块 …… 174
7.3.4 排名模块 …… 174
7.4 扩展阅读 …… 178

第 8 章 评分预测问题 …… 179
8.1 离线实验方法 …… 180
8.2 评分预测算法 …… 180
8.2.1 平均值 …… 180
8.2.2 基于邻域的方法 …… 184
8.2.3 隐语义模型与矩阵分解模型 …… 186
8.2.4 加入时间信息 …… 192
8.2.5 模型融合 …… 193
8.2.6 Netflix Prize 的相关实验结果 …… 195

后记 …… 196

图表目录

图 1-1 推荐系统的基本任务是联系用户和物品，解决信息过载的问题 2
图 1-2 推荐系统常用的 3 种联系用户和物品的方式 3
图 1-3 亚马逊的个性化推荐列表 4
图 1-4 单击 Fix this recommendation 按钮后打开的页面 5
图 1-5 基于 Facebook 好友的个性化推荐列表 6
图 1-6 相关推荐列表，购买过这个商品的用户经常购买的其他商品 6
图 1-7 相关推荐列表，浏览过这个商品的用户经常购买的其他商品 7
图 1-8 亚马逊的打包销售界面 7
图 1-9 Netflix 的电影推荐系统用户界面 8
图 1-10 视频网站 Hulu 的个性化推荐界面 9
图 1-11 Pandora 个性化网络电台的用户界面 10
图 1-12 Last.fm 个性化网络电台的用户界面 11
图 1-13 豆瓣个性化网络电台的用户界面 11
图 1-14 Clicker 利用好友的行为给用户推荐电视剧 13
图 1-15 用户在 Facebook 的信息流 14
图 1-16 不同社交网站中好友推荐系统的界面 14
图 1-17 Google Reader 社会化阅读 15
图 1-18 Zite 个性化阅读界面 16
图 1-19 FourSquare 的探索功能界面 17
图 1-20 Gmail 的优先级邮箱 18
图 1-21 Facebook 让广告商选择定向投放的目标用户 19
图 1-22 推荐系统的参与者 19
图 1-23 AB 测试系统 22
图 1-24 Hulu 让用户直接对推荐结果进行反馈，以便度量用户满意度 24
图 1-25 豆瓣网络电台通过红心和垃圾箱的反馈来度量用户满意度 24
图 1-26 不同网站收集用户评分的界面 25
图 1-27 Epinion 的信任系统界面 31
图 2-1 当当网在用户浏览《数据挖掘导论》时给用户推荐“购买本商品的顾客还买过”的书 36
图 2-2 各种显性反馈界面 37
图 2-3 物品流行度的长尾分布 40

图 2-4 用户活跃度的长尾分布 40
图 2-5 MovieLens 数据集中用户活跃度和物品流行度的关系 41
图 2-6 用户行为记录举例 45
图 2-7 物品-用户倒排表 47
图 2-8 Digg 的 My News 界面 51
图 2-9 亚马逊提供的用户购买 iPhone 后还会购买的其他商品 52
图 2-10 Hulu 的个性化视频推荐 52
图 2-11 一个计算物品相似度的简单例子 54
图 2-12 一个简单的基于物品推荐的例子 56
图 2-13 UserCF 和 ItemCF 算法在不同 K 值下的召回率曲线 61
图 2-14 UserCF 和 ItemCF 算法在不同 K 值下的覆盖率曲线 62
图 2-15 UserCF 和 ItemCF 算法在不同 K 值下的流行度曲线 62
图 2-16 两个用户在豆瓣的读书列表 65
图 2-17 雅虎首页的界面 71
图 2-18 用户物品二分图模型 73
图 2-19 基于图的推荐算法示例 74
图 2-20 PersonalRank 的简单例子 75
图 2-21 不同次迭代中不同节点的访问概率 76
图 3-1 Pandora 的用户注册界面 79
图 3-2 IMDB 中不同美剧的评分用户的性别分布 80
图 3-3 一个基于用户人口统计学特征推荐的简单例子 81
图 3-4 Lastfm 数据集中男女用户的分布 84
图 3-5 Lastfm 数据集中用户年龄的分布 84
图 3-6 Lastfm 数据集中用户国家的分布 84
图 3-7 Jinni 在新用户登录推荐系统时提示用户需要给多部电影评分 86
图 3-8 Jinni 让用户选择自己喜欢的电影类别 86
图 3-9 Jinni 让用户对电影进行评分的界面 87
图 3-10 给用户选择物品以解决冷启动问题的例子 88
图 3-11 关键词向量的生成过程 90
图 3-12 通过 LDA 对词进行聚类的结果 93
图 3-13 Jinni 中专家给《功夫熊猫》标注的基因 94
图 4-1 推荐系统联系用户和物品的几种途径 96
图 4-2 Delicious 中被打上 recommender 和 system 标签的网页 97
图 4-3 Delicious 中“豆瓣电台”网页被用户打的最多的标签 97
图 4-4 CiteULike 中一篇论文的标签 98
图 4-5 Last.fm 中披头士乐队的标签云 98
图 4-6 豆瓣读书中《数据挖掘导论》一书的常用标签 99
图 4-7 Hulu 中《豪斯医生》的常用标签 99
图 4-8 标签流行度的长尾分布 101
图 4-9 著名美剧《豪斯医生》在视频网站 Hulu 上的标签分类 102

图 4-10 Jinni 让用户对编辑给的标签进行反馈 …… 110
图 4-11 简单的用户-物品-标签图的例子 …… 111
图 4-12 SimpleTagGraph 的例子 …… 111
图 4-13 豆瓣读书的个性化推荐应用“豆瓣猜”的界面 …… 112
图 4-14 Last.fm（左）和豆瓣（右）的标签推荐系统界面 …… 115
图 4-15 豆瓣给我推荐的《MongoDB 权威指南》一书的标签 …… 118
图 5-1 sourcetone.com 个性化音乐推荐系统，该图右侧的圆盘可以让用户选择现在的心情 …… 122
图 5-2 facebook、twitter 和 myspace 3 个词的搜索变化曲线 …… 123
图 5-3 手机品牌的搜索量变化曲线 …… 124
图 5-4 一些食品相关搜索词的搜索量变化曲线 …… 124
图 5-5 不同数据集中物品流行度和物品平均在线时间的关系曲线 …… 126
图 5-6 相隔 T 天系统物品流行度向量的平均相似度 …… 127
图 5-7 推荐系统实时性举例 …… 128
图 5-8 时间段图模型示例 …… 134
图 5-9 BlogSpot 数据集的召回率和准确率曲线 …… 137
图 5-10 NYTimes 数据集的召回率和准确率曲线 …… 137
图 5-11 SourceForge 数据集的召回率和准确率曲线 …… 138
图 5-12 Wikipedia 数据集的召回率和准确率曲线 …… 138
图 5-13 YouTube 数据集的召回率和准确率曲线 …… 139
图 5-14 左图是大众点评提供的附近商户推荐，右图是街旁网提供的探索功能界面 …… 140
图 5-15 Hotpot 地点推荐界面 …… 140
图 5-16 一个简单的利用用户位置信息进行推荐的例子 …… 142
图 6-1 Facebook 提供的导入电子邮件好友的方式 …… 145
图 6-2 Facebook 在用户注册时让用户提供的一部分信息 …… 146
图 6-3 社交网络（Slashdot）中用户入度的分布 …… 149
图 6-4 社交网络（Slashdot）中用户出度的分布 …… 149
图 6-5 视频推荐网站 Clicker 利用 Facebook 好友信息给用户推荐视频 …… 150
图 6-6 亚马逊利用 Facebook 好友信息给用户推荐商品 …… 150
图 6-7 社交网络图和用户物品二分图的结合 …… 152
图 6-8 融合两种社交网络信息的图模型 …… 153
图 6-9 Twitter 的用户信息流 …… 156
图 6-10 Facebook 的用户信息流 …… 157
图 6-11 Jilin Chen 的用户调查实验结果 …… 159
图 6-12 Twitter 的好友推荐界面 …… 159
图 6-13 LinkedIn 的好友推荐界面 …… 160
图 6-14 Facebook 的好友推荐界面 …… 160
图 6-15 新浪微博利用用户的学校、公司、位置、标签给用户推荐好友 …… 161
图 7-1 推荐系统和其他系统之间的关系 …… 166
图 7-2 3 种联系用户和物品的推荐系统 …… 168
图 7-3 基于特征的推荐系统架构 …… 168

图 7-4 亚马逊同时给用户推荐电子产品和图书 …… 169
图 7-5 亚马逊的社会化推荐结果中包含了各种物品 …… 170
图 7-6 亚马逊给用户推荐最新加入的物品 …… 170
图 7-7 豆瓣电台考虑用户来源的上下文（该页面地址链接中加入了 context 参数） …… 170
图 7-8 推荐系统的架构图 …… 171
图 7-9 推荐引擎的架构图 …… 172
图 7-10 相关物品之间流行度之间的关系 …… 176

表 1-1 使用了 Facebook Instant Personalization 工具的网站 …… 13
表 1-2 离线实验的优缺点 …… 21
表 1-3 获取各种评测指标的途径 …… 33
表 2-1 显性反馈数据和隐性反馈数据的比较 …… 37
表 2-2 各代表网站中显性反馈数据和隐性反馈数据的例子 …… 38
表 2-3 用户行为的统一表示 …… 38
表 2-4 MovieLens 数据集中 UserCF 算法在不同 K 参数下的性能 …… 48
表 2-5 两种基础算法在 MovieLens 数据集下的性能 …… 48
表 2-6 MovieLens 数据集中 UserCF 算法和 User-IIF 算法的对比 …… 50
表 2-7 利用 ItemCF 在 MovieLens 数据集上计算出的电影相似度 …… 54
表 2-8 MovieLens 数据集中 ItemCF 算法离线实验的结果 …… 57
表 2-9 MovieLens 数据集中 ItemCF 算法和 ItemCF-IUF 算法的对比 …… 58
表 2-10 MovieLens 数据集中 ItemCF 算法和 ItemCF-Norm 算法的对比 …… 59
表 2-11 UserCF 和 ItemCF 优缺点的对比 …… 61
表 2-12 惩罚流行度后 ItemCF 的推荐结果性能 …… 63
表 2-13 MovieLens 数据集中根据 LFM 计算出的不同隐类中权重最高的物品 …… 69
表 2-14 Netflix 数据集中 LFM 算法在不同 F 参数下的性能 …… 70
表 2-15 MovieLens 数据集中 PersonalRank 算法的离线实验结果 …… 76
表 3-1 年轻用户和老年用户经常看的图书的列表 …… 83
表 3-2 年轻用户比例最高的 5 本书和老年人比例最高的 5 本书 …… 83
表 3-3 4 种不同粒度算法的召回率、准确率和覆盖率 …… 85
表 3-4 常见物品的内容信息 …… 89
表 3-5 MovieLens/GitHub 数据集中几种推荐算法性能的对比 …… 91
表 4-1 Delicious 和 CiteULike 数据集的基本信息 …… 103
表 4-2 Delicious 和 CiteULike 数据集中最热门的 20 个标签 …… 103
表 4-3 基于标签的简单推荐算法在 Delicious 数据集上的评测结果 …… 107
表 4-4 Delicous 和 CiteULike 数据集上 TagBasedTFIDF 的性能 …… 107
表 4-5 Delicous 和 CiteULike 数据集上 TagBasedTFIDF++的性能 …… 108
表 4-6 CiteULike 数据集中 recommender_system 的相关标签 …… 108
表 4-7 Delicious 数据集中 google 的相关标签 …… 109
表 4-8 考虑标签扩展后的推荐性能 …… 109
表 4-9 10 个用户最满意的主观类标签 …… 114

表 4-10 10 个用户最满意的客观类标签 …… 114
表 4-11 3 种标签推荐算法在 N=10 时的准确率和召回率 …… 117
表 4-12 HybridPopularTags 算法在不同线性融合系数α下的准确率和召回率 …… 117
表 5-1 离线实验数据集的基本统计信息 …… 125
表 5-2 美国、英国、德国用户兴趣度最高的歌手 …… 141
表 6-1 3 种不同好友推荐算法的召回率和准确率 …… 163
表 6-2 不同好友推荐算法的问卷调查结果 …… 164
表 7-1 电子商务网站中的典型行为 …… 167
表 7-2 离线相关表在 MySQL 中的存储格式 …… 173
表 8-1 评分预测问题举例 …… 179
表 8-2 MovieLens 数据集上不同平均值方法的 RMSE …… 184
表 8-3 MovieLens 数据集中对平均值方法采用级联融合后的效果 …… 194
表 8-4 Netflix Prize 上著名算法的 RMSE …… 195