

电子商务 推荐系统导论

伍之昂 曹杰 著



科学出版社

电子商务推荐系统导论

伍之昂 曹杰 著

科学出版社

北京

内 容 简 介

本书以电子商务为应用背景,深入浅出、全面地介绍推荐系统基本原理和主流技术,对每种模型和算法的阐述力求易于理解,并保持数学严谨性。本书涵盖的内容有:传统推荐模型与算法、情境推荐模型与算法、社会化推荐模型与算法、推荐系统评价指标与方法、推荐系统的可信性和安全性、电子商务推荐系统与企业绩效、电子商务推荐系统架构设计以及在大数据时代下推荐系统的展望。在章节论述过程中,作者主要从学术研究角度出发,围绕推荐系统面临的挑战性问题,阐释现有研究中的主流技术,指出值得进一步研究的难题。同时,还对电子商务推荐系统设计开发涉及的工程性技术及成功的实际应用系统作了简明扼要的阐述。

本书既可作为有兴趣从事推荐系统领域研究工作的高校教师和研究生的理论基础用书,帮助他们了解本领域研究现状,把握关键问题,熟悉基本方法;亦可作为企业工程师设计和开发推荐系统时的参考书。

图书在版编目(CIP)数据

电子商务推荐系统导论/伍之昂,曹杰著. —北京:科学出版社,2014.11
ISBN 978-7-03-042265-1

I. ①电… II. ①伍… ②曹… III. ①电子商务 IV. ①F713.36

中国版本图书馆CIP数据核字(2014)第247649号

责任编辑:惠雪 曾佳佳/责任校对:朱光兰
责任印制:肖兴/封面设计:许瑞

科学出版社 出版

北京东黄城根北街16号

邮政编码:100717

<http://www.sciencep.com>

文林印务有限公司 印刷

科学出版社发行 各地新华书店经销

*

2014年11月第 一 版 开本:720×1000 1/16

2014年11月第一次印刷 印张:12 1/2

字数:250 000

定价:69.00元

(如有印装质量问题,我社负责调换)

前 言

推荐系统脱颖于数据挖掘，近年来，推荐系统在不断汲取包括计算机科学、物理学、经济学和管理学等在内的各学科最新成果的基础上快速发展，已经成为一门新兴的多学科交叉研究领域。同时，从最早的 Tapestry 邮件推荐系统开始，推荐系统逐步得到了广泛的应用，其中最典型、最具有发展潜力的领域莫过于电子商务领域，几乎所有的电子商务网站都或多或少地运用了推荐系统，人们也从日常的在线购物中感受到了个性化推荐的无处不在。因此，大量的推荐系统领域的研究工作总是从电子商务应用切入，电子商务推荐系统所面临的挑战自然而然成为了研究者们所热衷讨论和解决的议题。

有别于理论性强的其他研究领域，电子商务推荐系统研究同时受到学术界和工业界的推动，而且工业界所投入的研究努力尤其不容被忽视。毫无疑问，推荐系统成为近年来计算机科学领域最热门的研究主题之一，国际顶级会议和期刊，如 SIGKDD、SIGIR、WWW、TKDE 等，每年都有大量论文与此论题相关。美国计算机学会 ACM 也因此于 2007 年开始举办推荐系统年会 ACM RecSys，迄今为止，RecSys 已成为一个重要的学术会议。随着研究的深入，电子商务推荐系统衍生出很多子议题，比如，越来越多复杂的数学模型被引入到协同过滤、情境化和社会化推荐方面，使其成果日趋丰富，推荐系统的可信性和安全性得到更多的重视；推荐系统的评价也不再仅仅局限于准确率，而是希望与电子商务企业绩效相挂钩。从学术研究角度，总结这一系列论文所表达出的卓越理念和先进方法就演变成了目前捧在您手中的这本书。除了介绍以模型和算法为主的理论性研究成果之外，我们还分析了多个成功应用于电子商务领域的推荐系统实例，以期弥合模型算法与实际系统之间的鸿沟。

撰写本书希望满足两类人群的需求：其一，已经或希望以推荐系统作为其研究方向的高校教师和研究生，本书可为他们梳理研究脉络、精要介绍主流方法、列出重要参考文献，为进一步的研究提供起点。其二，从事电子商务推荐系统设计和开发的企业工程师，他们可能对以数据挖掘为基础的推荐模型和算法知之甚少，本书对每种模型和算法进行了易于理解的阐述，试图帮助他们快速了解模型算法的原理，以更好地与工业实践相结合。

本书出版得到了国家科技支撑计划资助项目(2013BAH16F01)和江苏高校优势

学科建设工程的资助，特此致谢！

最后，我们对给予书稿撰写提供实实在在帮助的朋友致以诚挚的谢意。首先，感谢我们所在的实验室团队——国家级电子商务信息处理国际联合中心，团队前期在电子商务推荐系统方面丰富的研究成果及对该领域前沿研究的清晰把握奠定了书稿的基石。其次，感谢实验室博士生王有权完成本书第 6 章的写作，硕士生陈晨对书稿图表、格式及参考文献的多次不厌其烦的修改。最后，感谢科学出版社惠雪女士多次为本书体例和内容提出的诸多专业性意见及为本书的出版所付出的艰辛努力。

尽管至善的专著是我们撰写本书时所追求的目标，但是，囿于时间和作者水平，本书不可能尽善尽美，随时欢迎读者的批评指教，联系方式：zawuster@gmail.com。

伍之昂

2014 年 7 月 11 日

目 录

前言

第 1 章 引论	1
1.1 推荐系统概述	1
1.1.1 推荐系统：动机与现状	1
1.1.2 电子商务：推荐系统最重要的应用领域	2
1.1.3 推荐系统：形式化建模	3
1.2 推荐系统研究概览	5
1.2.1 挑战性问题	5
1.2.2 本书组织结构	9
参考文献	10
第 2 章 传统推荐模型与算法	12
2.1 协同过滤模型	12
2.1.1 基于用户的协同过滤	12
2.1.2 基于项目的协同过滤	14
2.1.3 相似度计算方法	15
2.1.4 混合型协同过滤	17
2.1.5 一个例子	18
2.2 基于内容的推荐模型	19
2.2.1 项目和用户表示	19
2.2.2 用户兴趣学习模型	21
2.3 潜在语义分析模型	23
2.3.1 奇异值分解模型	25
2.3.2 概率潜在语义分析模型	27
2.3.3 潜在狄利克雷分配模型	29
2.4 基于关联规则挖掘的推荐	30
2.4.1 频繁模式与关联规则挖掘	31
2.4.2 基于关联规则的推荐	32
参考文献	33
第 3 章 情境推荐模型与算法	35
3.1 推荐系统情境信息	35

3.1.1	什么是情境信息	35
3.1.2	情境信息的层次式表征	36
3.2	融合情境信息的推荐模型	37
3.2.1	情境预过滤	38
3.2.2	情境后过滤	41
3.2.3	情境化建模	41
3.2.4	三类范式的混合使用	44
3.3	基于时空信息的推荐模型	44
3.3.1	考虑用户兴趣漂移的推荐	45
3.3.2	移动推荐	48
3.4	基于隐反馈信息的推荐模型	57
	参考文献	59
第4章	社会化推荐模型与算法	63
4.1	社会计算与社会化推荐	63
4.1.1	社会计算研究议题概览	63
4.1.2	社会化推荐	67
4.2	社会化协同过滤模型	68
4.2.1	基于信任关系的社会化协同过滤	69
4.2.2	基于矩阵分解的社会化协同过滤	71
4.3	社会化标签系统中的推荐	75
4.3.1	社会化标签系统概述	75
4.3.2	基于社会化标签的多模式推荐	77
4.4	链路预测	83
4.4.1	基于拓扑结构信息的链路预测	83
4.4.2	基于节点信息的链路预测	84
4.5	作者协作关系网络中的推荐	86
4.5.1	专家推荐	87
4.5.2	论文推荐	88
4.6	社会化推荐领域的挑战性问题	89
	参考文献	91
第5章	推荐系统评价指标与方法	97
5.1	评价任务概述	97
5.2	评价指标	98
5.2.1	准确性指标	98
5.2.2	准确度以外的指标	103
5.3	实验数据集	106

5.3.1 一般性评分数据集	106
5.3.2 融合社会网络的数据集	107
5.4 评价方法	109
5.4.1 在线评价	109
5.4.2 离线评价	110
5.5 总结	110
参考文献	111
第 6 章 推荐系统可信性与安全性	113
6.1 托攻击概述	113
6.1.1 托攻击模型	113
6.1.2 托攻击分类	115
6.2 托攻击危害性衡量指标与脆弱性分析	117
6.2.1 托攻击危害性指标	117
6.2.2 托攻击脆弱性分析	118
6.3 托攻击检测特征指标与特征选择	120
6.3.1 托攻击检测特征指标	120
6.3.2 托攻击检测指标特征选择	122
6.4 托攻击检测算法	124
6.4.1 监督学习模型	124
6.4.2 无监督学习模型	127
6.4.3 半监督学习模型	130
6.5 托攻击检测算法评价方法	132
6.5.1 仿真实验	132
6.5.2 真实案例分析	133
6.6 群组攻击	134
6.6.1 严格版本的群组攻击构建算法	135
6.6.2 松弛版本的群组攻击构建算法	136
6.7 未来的研究方向	138
参考文献	139
第 7 章 电子商务推荐系统与企业绩效	142
7.1 电子商务企业绩效对推荐系统的影响	142
7.1.1 电子商务企业绩效的长尾效应	142
7.1.2 长尾分布、齐普夫定律和幂律分布	143
7.1.3 长尾效应对推荐系统设计的思考	144
7.2 多样化推荐系统设计	146
7.2.1 多样性的定义	146

7.2.2	重排序技术	148
7.2.3	基于主题多样性的推荐	149
7.3	推荐系统冷启动问题	150
7.3.1	冷启动问题概述	150
7.3.2	种子项目选择方法	151
7.4	总结	154
	参考文献	155
第 8 章	电子商务推荐系统架构设计	156
8.1	商用推荐系统概述	156
8.1.1	通用框架	156
8.1.2	数据类型	157
8.1.3	用户画像	159
8.2	推荐系统案例分析	161
8.2.1	Amazon 推荐系统	162
8.2.2	Netflix 推荐系统	163
8.2.3	Hulu 推荐系统	165
8.3	基于 Web 日志的推荐	167
8.3.1	Web 使用挖掘技术	167
8.3.2	Google 新闻推荐系统	171
	参考文献	172
第 9 章	下一代推荐系统研究展望	174
9.1	大数据的挑战	174
9.1.1	大数据计算	174
9.1.2	大数据计算对推荐系统基础架构的变革	175
9.2	推荐系统研究议题展望	179
9.2.1	隐私保护	179
9.2.2	基于位置的推荐	180
9.2.3	群组推荐	181
9.2.4	基于演化计算的推荐模型	182
	参考文献	182
索引	186

第 1 章 引 论

1.1 推荐系统概述

1.1.1 推荐系统：动机与现状

因特网的崛起和迅猛发展使人们获得了前所未有的自由信息空间，也使得世界范围内的信息处于大爆炸状态，造成信息量大、信息质量差、信息价值低等问题，用户难以从过量信息中获取对自己有用的部分，这就是所谓的信息过载(information overload)^[1]。正如文献[2]中所言，数据是广泛可用的，所缺乏的是从中提取出知识的能力。为应对信息过载问题，大量信息过滤网络工具应运而生，人们最常用的莫过于门户网站和搜索引擎，但这些工具都是以面向人们主流需求为主，且用户获取信息的方式是“被动”的。

显然，考虑个性化需求，将有用信息以“主动”方式推送(push)给用户不仅能过滤信息，而且能增强用户体验，推送的信息也更符合用户个体的需求。简信息聚合(really simple syndication, RSS)是一种在主动信息推送服务方面的典型工具。RSS是一种消息来源格式规范，能够将新闻标题、摘要、内容按照用户的要求“送”到用户桌面。但是，在RSS中，用户兴趣通常是由用户预先给定的关键词来描述，但缺乏自动挖掘用户兴趣的功能。

推荐系统(recommender systems)是一种为用户提供建议的智能化软件工具，是解决信息过载问题非常有潜力的方法^[3]。推荐系统的研究与开发始于20世纪90年代初期。第一个推荐系统是Tapestry^[4]，是针对电子邮件的推荐系统。该推荐系统通过分析用户阅读邮件的历史行为来对新邮件重新排序，试图将用户更感兴趣的新邮件推荐给他/她。更为重要的是，Tapestry首次提出了协同过滤(collaborative filtering, CF)的概念^[4]。后续的推荐系统领域研究表明，协同过滤对推荐系统的发展产生了深远影响，成为关注最多、应用最广泛的方法之一。

推荐系统与以搜索引擎为代表的信息检索技术本质区别^[5]在于：①搜索注重结果之间的关系和排序，而推荐则需研究用户兴趣模型，即结合考虑用户与推荐对象的特征；②搜索由用户主导，包括输入查询词和选择结果，若结果不好，用户会修改查询再次搜索，而推荐则由系统主导用户的浏览顺序，主动引导用户发现需要的结果。

从 20 世纪 90 年代初期到现在，推荐系统受到学术界和工业界的长期关注，成为炙手可热的研究议题。推荐系统是一门典型的交叉学科，涵盖计算机科学、电子商务、人类社会学、物理学、经济学等诸多领域，图 1.1 给出 2013 年国际顶级会议 ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) 投稿数量的分布，可以看出，推荐系统在 20 个领域中排名第三，其热度由此可见一斑。从 2007 年开始，ACM 设立了专门讨论推荐系统的国际会议，ACM Recommender Systems Conference (RecSys) 至今已举办 8 届，其知名度持续上升。从 2006 年开始，美国密歇根大学 Paul Resnick 教授开设了“推荐系统”课程。

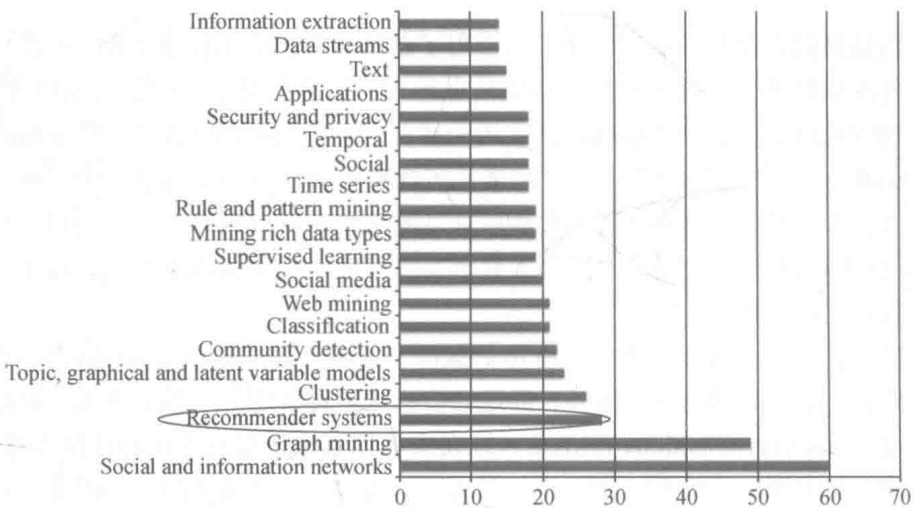


图 1.1 KDD 2013 年投稿数量领域分布

随着社会网络、物联网、大数据等新概念和新技术的出现和升温，信息过载问题将日益严重，同时，工业界的需求导致用户对信息获取方式的要求也变得越来越。推荐系统作为既能应对信息过载，又能智能化地为用户主动推送信息的强有力手段，可以预计，在今后数十年内，推荐系统仍然将获得持续关注，更加有效和高效的推荐系统将在人类信息化进程中扮演越来越重要的角色。

1.1.2 电子商务：推荐系统最重要的应用领域

Companies need to shift from the old world of mass production where “standardized products, homogeneous markets, and long product life and development cycles were the rule” to the new world where “variety and customization supplant standardized products”.

quoted from “Mass Customization” by Joe Pine^[6]

以上内容引自 Joe Pine 的著作“*Mass Customization*”，说明了企业由传统模式过渡到现代模式的法则，需要从传统标准化商品过渡到多样化的定制商品，其隐含意思就是充分考虑用户个性化需求。这段话恰当地说明了推荐系统为何在电子商务领域得到良好发展和应用的原因。用户的需求通常是模糊的，而商家期望能够把满足用户模糊需求的商品推荐给用户，把用户潜在需求转化为现实需求，从而达到提升商品销售量的目标。

电子商务将企业的销售行为从线下实体店销售搬到了在线销售，用户行为和商品信息的数字化更加促进推荐系统的应用。目前，几乎所有的大型电子商务网站都不同程度地采用了推荐系统^[5]，电子商务成为推荐系统最重要的应用领域，同时，电子商务企业的应用和研究也促进了推荐系统的发展，比如，Amazon 在 10 年前提出了基于项目的协同过滤方法^[7]，并沿用至今。推荐系统为电子商务企业带来了新的营销观念，即“一对一营销”，正如 Amazon CEO Jeff Bezos 所说，“如果我们有 200 万个网络客户，我们就需要在线储备 200 万种商品”。总的来说，推荐系统能为电子商务企业带来以下三方面的收益^[8]。

(1) 提升浏览到购买的转化率。网站访问者经常漫无目的地浏览网站，却不购买任何商品，推荐系统能帮助访问者发现其感兴趣的商品，从而促进访问者到购买者的转化。

(2) 提升交叉销售量。推荐系统通过向用户推荐与其已购买商品相关联的商品，用户在购买的最后步骤，网站会向用户推荐与购物车内商品极为相关的附属商品。

(3) 提升用户忠诚度。如何提升用户忠诚度或黏粘度是电子商务网站的核心商业策略之一，推荐系统建立起网站与用户之间的联系，为提升用户忠诚度提供附加值，用户往往会重复访问与其兴趣偏好相匹配的网站，甚至会将这些网站推荐给自己的朋友圈。

在竞争日趋激烈的环境下，电子商务推荐系统能够为电子商务网站有效保留用户，提高电子商务系统绩效，产生巨大的经济效益。构建一个推荐效率高、精度高而且 UI 界面合理的商品推荐系统对于电子商务网站的发展十分重要，推荐系统也因此成为工业界的关注热点。

1.1.3 推荐系统：形式化建模

一般地，推荐系统包含两类实体：用户(user)和项目(item)。用户是指被推荐对象，无需过多解释。而项目则是指推荐对象，推荐系统领域的所有英文文献都使用“item”一词来表达，本书统一将“item”译为“项目”，泛指不同领域的推荐对象。比如，电子商务推荐系统中，项目表示商品或物品；音乐和电影推荐系统中，项目则表示 CD 和 DVD；新闻推荐系统中，项目又表示新闻网页，等等。尽管“项目”有可能与英语“project”一词相混淆，但是，将推荐实体“item”译为“项目”在

推荐系统领域的中文文献中广泛出现^[5, 9-12]。

推荐系统输入数据的最简单形式是用户对项目的评分数据,称为用户-项目评分矩阵(user-item rating matrix),大部分数据集采用5分制(或5星制)评分,即5分表示最满意,依次递减。尽管实际电子商务系统提供了用户对购买中不同方面的评分,如图1.2所示的淘宝网评分体系,就支持“宝贝与描述相符”、“卖家服务态度”、“卖家发货速度”3种不同子项。但是,在推荐系统形式化建模时却通常只考虑综合评分。同时,用户-项目评分矩阵不一定是显式给出,也可以基于隐含的用户和项目关系推测得到。如果用户对某项目评过,其偏好已经从用户-项目评分矩阵中反映出来,已评分项目就不作为推荐给用户的候选对象。因此,推荐系统工作时试图预测用户对未评分项目的偏好程度,由预测分值反映。基于上述分析,推荐系统本质是基于已有的用户-项目评分矩阵(称为训练集),寻找一个准确的预测函数 P :

$$P: Users \times Items \rightarrow \hat{r}_{ui}, \quad \forall u \in U, i \in I, r_{ui} = \phi \quad (1.1)$$

其中, U 和 I 分别表示用户和项目的集合,任意待预测的用户项目对 (u, i) 在训练集中的评分为空值。一旦利用预测函数 P 对用户-项目评分矩阵缺失值计算出预测分值后,就很容易产生推荐,即将预测分值最高的项目(或 N 个分值最高的项目集合)推荐给用户,这一过程形式化为

$$\forall u \in U, \quad i_u = \operatorname{argmax}_{i \in I} P(u, i) \quad (1.2)$$



图 1.2 淘宝网评分体系

集合 U 和 I 在实际中都非常大,用户规模很容易达到百万级,项目规模则更大,这导致超大规模的用户-项目评分矩阵,因此,依靠遍历评分矩阵整个空间设计预测函数 P 在计算上不可行。围绕计算预测值的效率问题,推荐系统提出了大量各种各样的模型和算法,如基于内容的、协同过滤、混合型、基于矩阵分解的方法等。

推荐算法(recommendation algorithm)决定了预测函数 P 的设计和计算,是整个推荐系统最核心和关键的部分,在很大程度上决定了推荐系统的类型和性能优劣。推荐算法也因此成为探讨推荐系统的核心部分,本书后续章节正是围绕主流推荐算法来介绍电子商务推荐系统。尽管推荐系统是一门交叉学科,与其联系最紧密的是数据挖掘(data mining),数据挖掘是自动地从大规模数据集中发现有用信息的过程,几乎所有的推荐算法都可以被认为是一种数据挖掘技术^[13]。数据挖掘过程大致可分为三个阶段:数据预处理、数据分析以及结果解析,图 1.3 描述了这三个阶段中与推荐算法密切相关的数据挖掘方法。在数据预处理阶段,距离计算、抽样、维度约简等技术经常被用于协同过滤算法。在数据分析阶段,分类技术经常被用于基于内容的推荐算法,关联规则挖掘形成了一种单独的推荐算法,聚类可用于具有相似兴趣的用户群组发现及具有相似属性的商品群组发现。

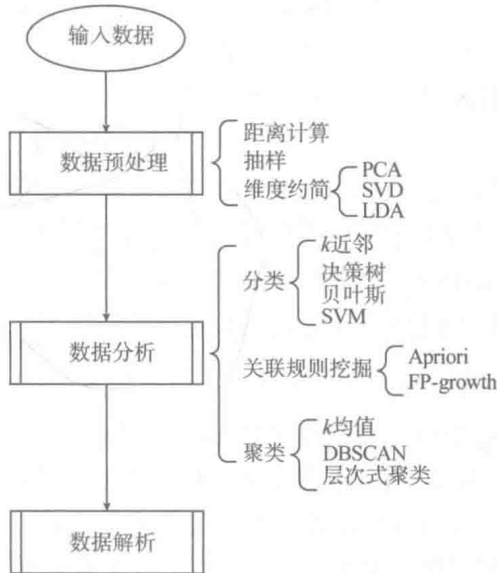


图 1.3 推荐系统中经常使用的数据挖掘技术(图片来源:文献[13])

1.2 推荐系统研究概览

1.2.1 挑战性问题

推荐系统的研究是围绕其面临的挑战性问题而展开的,因此,首先应在清楚了

解挑战性问题的基础上,来进一步阐释已有的推荐系统研究工作。本书后续章节将围绕针对推荐系统所面临的挑战性问题的研究工作而展开论述,试图为推荐系统这一活跃的研究领域勾勒出较为全面、清晰的概貌。

1. 推荐算法的准确度和可扩展性

推荐算法一直是推荐系统研究中最核心的问题,关注焦点包括准确度和可扩展性两个方面。准确度是衡量算法的有效性(effectiveness),可扩展性是衡量算法的效率(efficiency)。目前,如何设计效率高且准确度高的算法仍然是一个开放性话题,百家争鸣,各种各样的推荐算法被提出。导致这一挑战的根本原因在于数据呈现高度稀疏性,即用户和项目规模通常达到百万级,甚至千万级,而已知的评分数据又比较稀少,基于如此稀疏的数据,设计兼顾有效性和效率的算法却是一大挑战。目前已有的研究大多沿着两种思路来设计推荐算法:一种是以稀疏的用户-项目评分矩阵为输入数据,依托数据挖掘技术,提出新的模型和算法,包括协同过滤及其各种变例、基于内容的推荐、基于维度约简的各类模型以及基于关联规则挖掘的推荐等,将在第2章介绍这些算法;另一种是在用户-项目评分矩阵的基础上,融入更多的额外数据以缓解稀疏性。将在第3章讨论的情境推荐模型与算法中考虑推荐的上下文信息,可能包括用户购买意图、季节、节日、位置、同伴、天气等因素。将在第4章讨论的社会化推荐模型与算法中则考虑用户的社会关系信息,将推荐系统与社会网络结合到一起。

从算法采用的模型来看,无论是单一的评分矩阵,还是融入额外数据的复杂输入源,模型本身仍然是数据挖掘中一些经典模型针对具体数据或具体应用的扩展,比如,奇异值分解(SVD)技术广泛用于基于单一评分矩阵的传统推荐和社会化推荐,协同过滤模型也被扩展到能融合社会关系数据。

2. 推荐系统的评价

评价推荐系统的效率与传统算法评价一样,可以通过时空复杂度(大 O 表示法)或运行时间进行比较。但是,推荐系统有效性的评价一直是困扰研究者的一大难题,到目前为止,推荐系统有效性的评价仍然未能形成完整的、公认的体系。几乎所有的推荐算法设计者仅仅只能从某个角度验证其算法是优秀的,而不可能声称某个推荐算法是十全十美的。而我们在对数据挖掘算法研究时很少遇到算法评价的困扰,比如,评价分类算法约定俗成地采用准确率(precision)、召回率(recall)及F-measure;评价聚类算法则往往使用互信息熵(NMI)、纯度(purity)、Rand Index(R_n)等^[14]。

除了推荐系统评价指标体系的缺失,数据的高度稀疏性为离线评价只能评价极少量的已评分记录,在线评价则是一种理想的评价途径,即将推荐结果发布到网站,让用户提交满意度反馈,或通过点击量来推测,但是,在线用户测试需依托具有一

定影响力(访问量大的)实际系统,而这些与工业界缺乏较强合作的研究者则很难实施。将在第5章介绍推荐系统评价方面的工作。

3. 推荐系统的健壮性

电子商务推荐系统与经济利益相挂钩,网店店主如果能够经常让自己的商品进入推荐列表,同时打压对竞争对手商品的推荐,显然能够攫取可观的收益。同时,电子商务推荐系统应用最为广泛的协同过滤算法面对托攻击(shilling attack)时非常脆弱,这使得恶意用户有了可乘之机^[11]。大量研究都是围绕如何提升推荐系统健壮性而展开,大体上沿着两种思路进行。其一,提升推荐算法本身的健壮性,即抗托攻击的能力,通常的思路是放弃协同过滤方法,改用诸如关联规则挖掘、维度约简的方法,但是,这种思路难以改变电子商务推荐系统已经广泛采用协同过滤的现状。其二,研究托攻击检测方法,作为协同过滤的补充手段。这样的研究思路与网络安全极为类似,利用防火墙希望“御敌于国门之外”,对于“漏网之鱼”,则希望利用入侵检测技术在尽量短时间内发现它们,并赶出系统。将在第6章详细阐述推荐系统的健壮性领域的研究工作。

4. 多样性和准确性

管理学方面的大量研究指出,电子商务企业的绩效并不遵循帕雷托法则(二八定律),而是符合重尾分布。这一现象为推荐系统提出了新要求,以推荐流行商品为主的系统固然能提升准确性,但是,提升流行商品销售量不能使电子商务企业获得巨大效益,即对于企业而言,希望提升那些销售量低且处于“长尾”上的商品的销售量。这就为推荐系统设计带来矛盾,既要保持高准确性,又要兼顾推荐结果的多样性。

目前,越来越多的研究者开始关注推荐结果多样性的问题,普遍认为,推荐系统不应当只推荐或经常推荐流行商品,因为流行商品很容易被找到,无需推荐系统的帮助,而一个好的推荐结果列表应该包含一些用户意想不到的结果,或许这也正是用户所期待的。将在第7章介绍这一有趣问题以及一些已有的对策。

5. 冷启动现象

冷启动(cold start)问题一直是推荐系统研究面临的一个难题,本质上,冷启动问题仍然是由用户-项目评分矩阵的稀疏性引起的,即大量用户只给出少量评分,同时大量项目也仅得到了少量评分。对于实际电子商务网站,新用户注册、新商品上架就自然而然引起冷启动现象。解决或减轻冷启动现象的应用最广泛的手段是融入额外的信息。将在第3、4章介绍的情境推荐和社会化推荐中,大量模型和算法都声称引入情境信息和社会关系之后有助于解决冷启动问题。这样的声称显然是合理的,融入额外数据后,新用户和新商品有了来自于其他数据源的描述,推荐系统

也就有了可供使用的输入数据。

除了引入更多的额外信息,挑选种子项目成为一种简洁实用的解决方案。种子项目带来了两种针对冷启动问题的应用方案:其一,直接将种子项目推荐给冷启动用户,类似于将热门排行榜上的商品推荐给一个陌生用户,这似乎是一种“不求有功但求无过”的解决方案;其二,为新用户设置简单的口味测试,将种子项目作为口味测试对象,根据用户提交的对种子项目评价来挖掘用户偏好。如何挑选种子项目成为以上两个解决方案中的核心问题。将在第7章介绍几类种子项目选择方法。

6. 推荐系统的用户接口

用户接口影响推荐系统的实际应用效果,推荐系统的用户接口主要是指以何种形式将推荐结果推送给用户,甚至收集用户对推荐结果的反馈。这就要求用户接口重点突出、时机合理、界面简洁、易于反馈。目前,大部分电子商务网站都以图片配合文字的形式呈现推荐结果,如 Amazon、Netflix、淘宝网、当当网等,少量网站使用视频方式呈现推荐结果,如 Hulu。除了结果展示形式,文献[15]指出推荐结果的可解释性至关重要,这类似于以往专家系统给出决策支持建议时,需要给出建议的依据。用户不希望推荐系统看起来像个“黑匣子”(black box),更希望能看到推荐系统对结果解释其合理性,Amazon 就注意到这个问题,在推荐结果下方会出现“为什么推荐这些商品”的链接。在第8章介绍电子商务推荐系统案例时,特别比较了不同设计风格的用户接口,这值得引起推荐系统工程化建设和开发人员的重视。

7. 使用评分数据之外的数据源

尽管主流推荐系统都依赖于用户-项目评分矩阵,但在实践中,推荐系统设计经常依赖于企业或网站所拥有的数据类型,一个经营时间不短的企业或网站都能积累大量的原始数据,多种类型的数据都能反映用户偏好和兴趣,比如,用户点击网站留下的日志数据、商品评论文本、售后投诉记录等。相对于显式评分数据,这些数据能隐式地反映用户偏好和兴趣,而合理使用隐式数据有助于提升推荐系统准确度,并缓解冷启动现象。

目前,由于文本和 Web 日志数据的普遍性,基于文本数据和日志数据挖掘的推荐技术逐渐得到重视。利用观点挖掘(opinion mining)发现文本中隐含的用户观点,及细粒度的对商品具体属性的态度,将有助于构建用户兴趣模型,甚至检测恶意用户。有趣的是,有些推荐系统专门用于推荐有用的评论^[16]。另外,用户点击网站 HTML 页面留下的 Web 日志数据成为企业必不可少的数据类型,且往往占据总数据量很高的比例,利用数据挖掘技术从 Web 数据中发现有用模式称为 Web 使用挖掘(Web usage mining),目前,依托 Web 使用挖掘技术设计推荐系统受到大量关