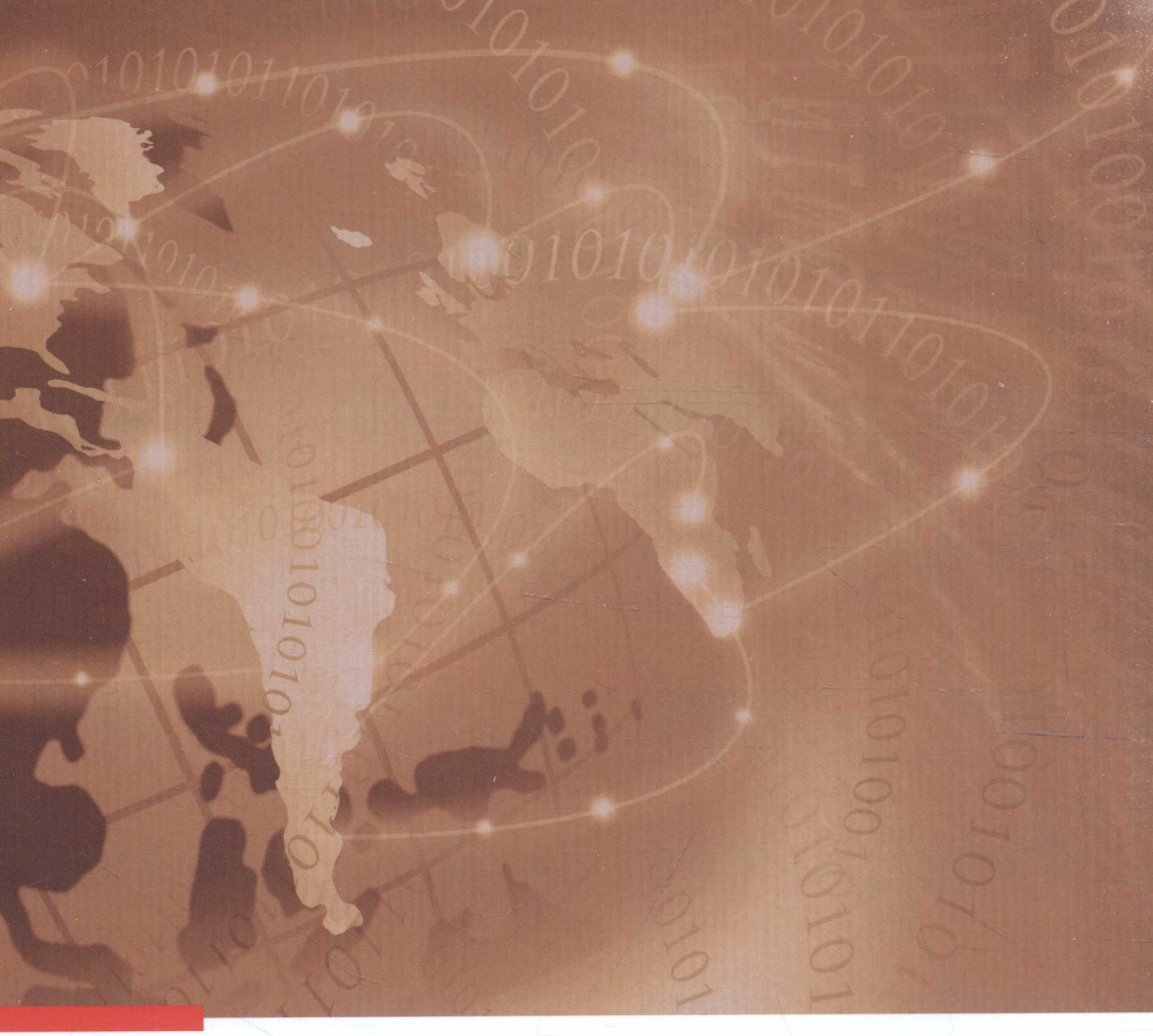

电子商务推荐系统瓶颈问题研究

李聪　马丽/著

Research on Bottleneck Problems in E-commerce Recommender Systems

科学出版社

电子商务推荐系统瓶颈问题研究

李 聪 马 丽 著

科 学 出 版 社

北 京

内 容 简 介

电子商务推荐系统是解决信息超载的重要技术。协同过滤作为推荐系统中广泛使用的、最成功的推荐算法，还存在诸如稀疏性（sparsity）、冷启动（cold-start）、可扩展性（scalability）等制约其进一步发展的瓶颈问题。本书针对稀疏性问题，提出了非目标用户类型区分理论、领域最近邻理论、基于 Rough 集理论的用户评分项并集未评分值填补方法等；针对冷启动问题，提出了一种冷启动消除方法，包括用户访问项序理论、n 序访问解析逻辑、改进的最频繁项提取算法 IMIEA、用户访问项序的 Markov 链模型等；针对可扩展性问题，提出了适应用户兴趣变化的协同过滤增量更新机制；最后设计并实现了一个电子商务协同过滤原型系统 ECRec。

本书可供管理学、计算机科学等相关领域和专业的高校师生、科研院所研究人员、IT 企业（尤其是电子商务企业）管理者及技术人员参考使用。

图书在版编目(CIP)数据

电子商务推荐系统瓶颈问题研究 / 李聪，马丽著. —北京：科学出版社，2016

ISBN 978-7-03-047158-1

Ⅰ. ①电… Ⅱ. ①李…②马… Ⅲ. ①电子商务-研究 Ⅳ. ①F713.36

中国版本图书馆 CIP 数据核字（2016）第 013595 号

责任编辑：刘 超 / 责任校对：钟 洋

责任印制：徐晓晨 / 封面设计：无极书装

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

北京中石油彩色印刷有限责任公司 印刷

科学出版社发行 各地新华书店经销

*

2016 年 1 月第 一 版 开本：720×1000 B5

2016 年 1 月第一次印刷 印张：10 1/2

字数：300 000

定价：88.00 元

（如有印装质量问题，我社负责调换）

前　言

2009 年 2 月，*Science* 发表社会计算学论文，阐述利用网络数据研究群体社会行为及其演化规律，标志着“社会计算”这门新兴学科正成为国际瞩目的前沿研究和应用热点。在电子商务环境下，如何实现准确、快速的个性化推荐服务是一项历久弥新的、同时也是社会计算领域的重要研究内容。纷繁芜杂的网站商品信息对消费者造成了“信息超载”(information overload)，导致其无法准确、快速地定位能满足自身需求的个性化产品。电子商务推荐系统可以帮助消费者提高购买决策的质量和效率，被认为是目前解决信息超载问题的最有效工具，其本质是面向消费者的决策支持系统。

协同过滤（collaborative filtering）作为目前电子商务推荐系统中广泛使用的、最成功的推荐理论（简称协同推荐），其采用的用户兴趣模型是通过搜集消费者对商品的评分来生成的用户-项目评分矩阵，进而模拟人类口碑相传（word of mouth）的推荐过程来实现商品推荐。经过 1992 ~ 2015 这二十多年的发展，协同推荐已在电子商务网站及社交网络、视频/音乐点播等主流 Web 2.0 服务中得到普及，并被成功引入到数字图书馆、在线学习、移动商务、多 Agent 系统、Web Service 甚至核安全评估、医疗等多个领域。但是，评分矩阵会随着网站用户及商品数量的不断增加而迅速成长为一个存在大量空缺元素的高维矩阵，从而给协同推荐带来稀疏性(sparsity，用户所给商品评分通常不及 1%)、冷启动（cold-start，稀疏性的极端情况，也称冷开始）、可扩展性（scalability，庞大的用户和商品数量使得推荐计算极为耗时）等瓶颈问题。其中，稀疏性和冷启动严重影响

协同推荐的准确性；可扩展性则使得协同推荐的实时性难以保证。

本书正是围绕上述三大瓶颈问题展开研究。我们诚挚希望书中的研究成果能够为促进协同推荐理论的完善及其在电子商务中的应用、推广和普及，以及构建电子商务智能数据挖掘与复杂机器学习系统提供有益参考。

最后，本书能够顺利出版，得益于国家自然科学基金青年科学基金项目（编号：71202165）的资助和科学出版社的支持。对此，我们表示衷心的感谢！

李　聪　马　丽

2015 年初冬

目　　录

第1章 绪 论

1.1 问题的提出

随着互联网(Internet)和电子商务的迅猛发展，人类已经进入信息社会时代。中国的电子商务市场发展潜力巨大，同时保持了持续高速增长势头。截至2015年6月，中国网民规模已达6.68亿人，互联网普及率为48.8%，较2014年年底提升了0.9个百分点[1]。《2014年中国网络购物市场研究报告》[2]的数据显示，2014年中国网络零售交易额为2.79万亿元(继续保持全球第一)，同比增长49.7%，相当于同期社会消费品零售总额(26.2万亿元)的10.6%；网络购物用户规模达3.61亿人，较2013年增长19.7%；网民使用网络购物的比例从48.9%提升至55.7%；中国网络购物市场依然保持着较高的活跃度，全年交易总次数为173亿次，年度人均交易次数为48次；网络购物金额占日常消费采购支出比例的平均值则为14.2%。

电子商务网站不仅使企业节省了传统经营模式下必要的实体投资成本，而且还拥有一个巨大的优势，即消除了传统店面的商品陈列空间限制，为购物者提供了一个庞大的(也可以说是无限大)商品陈列柜台供其选择。人们通过访问电子商务网站，可以享受足不出户选购商品的快乐和方便。

但是，面对电子商务网站提供的大量商品，顾客无法通过小小的计算机屏幕在短时间内浏览所有商品，并且也缺少现实商店中促销人员的精心导购，从而面临“信息超载”(information overload)[3]。信息超载指网站为用户提供的商品信息量过多，导致其难以迅速找到所需商品，并且在这之

前难免会浏览大量不相关信息，从而很容易使用户产生疲劳直至失去购物兴趣并离开。基于上述情况，电子商务网站面临着一个严峻的问题：如何在用户浏览网站时将适合该用户的商品推荐到他／她面前，克服信息超载带来的不利影响，从而促成更多的交易以增加企业销售额？

电子商务推荐系统(E-commerce recommender systems) 就是解决信息超载问题的一种方案[4]、一种实现电子商务网站“一对一营销”(one - to - one marketing) 战略的技术[5]，可作为网站客户关系管理(customer relationship management，CRM) 的有益组成部分[6]。早在 1982 年，美国计算机学会主席 Denning[7] 就指出需要将注意力从“制造信息”(generating information) 更多地集中到接收信息(receiving information，即控制和过滤信息并使其传到必须使用它们的人) 上来。美国学者 Pine[8] 则指出，现代企业应该从大规模生产(标准化产品) 向大规模定制(为多类客户的多样需求提供多种商品) 转变，并列出了五种达到大规模定制的方法，其中有四种都能通过电子商务推荐系统来实现，包括“围绕标准化的产品和服务来定制服务”、“创建可定制的产品和服务”、“提供交货点定制”和“提供整个价值链的快速响应”，因此电子商务推荐系统也是电子商务网站进行自动化大规模定制的一条关键途径[4]，它使得网站能适应每一个消费者并为其提供具有个性化的商品展现平台和购物体验。由于用户对自身需求不甚明确时，其注意力并不专注于某特定目标，因此推荐系统所给建议被用户采纳的概率将相对较大。正如 Jeff Bezos(Amazon 公司 CEO) 所言：“如果我在网络上有 300 万个用户，我就应有 300 万个网上商店”[5, 9]。具体而言，电子商务推荐系统的作用表现在三个方面[5]。

(1) 将电子商务网站浏览者转变为购买者(converting browsers into buyers)

电子商务网站的访问者经常只是浏览一下，并没有购买商品的意愿。电子商务推荐系统能够帮助这些浏览用户找到他们愿意购买的商品，

从而将浏览者转变为购买者。

(2) 提高电子商务网站交叉销售能力(increasing cross - sell)

交叉销售在现代商业中应用非常普遍。通过交叉销售，能够引导用户发现和购买自己确有潜在需求但在购买过程中未曾想到的商品。电子商务推荐系统可以在用户浏览某商品时根据用户购物车(shopping cart)中的商品向其推荐该商品的相关产品，从而提高销售量。

(3) 建立电子商务网站客户忠诚度(building loyalty)

赢得客户忠诚度是一项基本的商业策略。在互联网上，用户只需要点击一两下鼠标，便能从当前的电子商务网站转到其竞争对手网站。电子商务推荐系统通过建立网站与客户之间的增值关系(value - added relationship)来提高客户忠诚度。一方面，电子商务推荐系统通过了解和学习客户的兴趣偏好来推荐满足其需求的合适商品；另一方面，客户越多地使用电子商务推荐系统，系统就越能了解其兴趣偏好，从而给出的推荐结果质量越高。这种良性循环一旦形成，将大大增加客户忠诚度，提高客户和网站之间的“黏性”(stickiness)。

许多大型网站早已开始使用电子商务推荐系统[10]，如全球最大的网上书店 Amazon (http: //www. amazon. com)[11](其1/3的销售额来自推荐系统[12, 13])、最大的网上拍卖站点 eBay(http: //www. ebay. com)、最大的网上音乐商店 CDNow (现已被 Amazon 收购)、最大的搜索引擎 Google[14](在 Google Q3 2006 earnings call[15] 中，Google CEO Eric Schmidt 明确指出向用户提供个性化信息是 Google 的最终使命)、主流门户网站 Yahoo(http: //www. yahoo. com)[16, 17]、最大的中文网上书店当当网(http: //www. dangdang. com) 以及搜狐商城(http: //store. sohu. com)、网易商城(http: //mall. 163. com) 等。而作为大型在线电影租赁公司，Netflix“3/4的新订单都来自推荐系统”[12]。

目前，电子商务推荐系统已成为电子商务领域的一大研究重点和热点。从本质上讲，推荐系统属于决策支持系统[18]，帮助在线用户进行购物决策[19, 20]。美国学者 Pennock[21, 22]、Yager[23] 和日本学者 Iijima[24] 等分别从社会选择理论(social choice theory)、模糊集(fuzzy set)、多准则决策(multi - criteria decision making) 的角度对其进行了剖析。协同过滤(collaborative filtering, CF) 是目前电子商务推荐系统中广泛使用的、最成功的推荐算法[9, 25, 26]，但还存在诸如稀疏性(sparsity)[9]、冷启动(cold - start)[9]、可扩展性(scalability)[9] 等制约其进一步发展的瓶颈问题。对于这些瓶颈问题，需要研究和解决的关键问题主要体现在以下方面。

1) 在稀疏性问题研究方面，如何使得目标用户的最近邻搜寻更为准确?

2) 在冷启动问题研究方面，如何对一个没有提供任何评价信息的新用户进行推荐服务，从而在其访问网站的初期便能留住他 / 她?

3) 在可扩展性问题研究方面，如何使得协同过滤算法能在用户和项目数量不断增长的情况下尽量降低在线推荐所需的响应时间?

由于我国电子商务的推荐功能相对国外存在较大差距[27]，因此如何有效解决上述问题，对于推动和促进我国电子商务推荐系统的研究、应用以及缩小与国外的差距，具有十分重要的理论研究价值和应用价值。本书的研究工作正是基于这样的背景所展开。

1.2 研究目的与意义

本书的研究目的在于通过对协同过滤进行深入研究，提出能够有效缓解或解决协同过滤稀疏性、冷启动、可扩展性等瓶颈问题的相应理论、方法、模型和机制，从而为推动中国电子商务推荐系统的更快、更好发展起到积极的作用。

当前，针对协同过滤瓶颈问题展开研究具有重要的意义，其具体体现在现实应用和理论研究两个方面。

1）在现实应用方面，随着现代电子商务的迅猛发展，用户和企业都迫切需要有效的电子商务推荐系统。没有客户就没有利润，拥有足够的客户群是企业生存和发展的充要条件。在电子商务环境下，企业面临的主要问题之一就是如何为用户提供个性化程度更高、更符合用户需求的商品和服务。这是电子商务企业价值链的源头和市场营销的起点。但是，由于电子商务网站的商品种类和数量太多，导致用户需要花费较多的时间来寻找自己所需要的商品，因此类似于促销助手的电子商务推荐系统就显得格外有用。通过电子商务推荐系统，可以为用户创造出一个方便快捷的个性化购物环境。用户可能感兴趣的商品将被实时地推送到用户眼前，使得用户能够在尽量少的时间内迅速找到心仪的商品。而电子商务企业也可以基于这种主动的、自动化的、实时的、准确的“one - to - one”销售方式，在激烈的市场竞争中尽可能满足不同顾客对不同商品的个性化需求，以赢得更多的客户，获得更高的销售业绩。因此，电子商务推荐系统研究符合现代电子商务发展的要求，是用户和电子商务企业即买卖双方的迫切需求，具有较高的现实意义。协同过滤作为目前电子商务推荐系统中广泛应用的、最成功的推荐算法，已成为电子商务推荐系统的核心组成部分和主要研究内容。

2）在理论研究方面，正是现实应用需求的大力推动，使得电子商务推荐系统研究日渐成为国内外学术界关注的热点。电子商务推荐系统中的协同过滤算法作为当前的主流推荐技术，由于其在支持新异推荐（serendipitous recommendations）、处理非结构化复杂对象（视频、图像等）以及推荐质量等方面优于其他推荐技术，更是受到广大研究人员的重视。与此同时，传统的协同过滤算法所固有的稀疏性、冷启动、可扩展性等问题已经成为阻碍电子商务推荐系统发展的瓶颈而需亟待解决。同时，协同过滤不仅仅可以应用于电子商务领域，在群决策（如谈判支持系统[28]）、Web 站点导航[29]、数字图书馆（digital library）[30]、在线学习（E - learning）[31] 等领域都有广泛的应用前景。因此，针对协同过滤瓶颈问题展开进一步研究，具有较高的理论研究意义。

1.3 电子商务推荐系统概述

1.3.1 定义及任务

电子商务推荐系统被电子商务网站用作虚拟店员(virtual salespeople)向客户提供商品信息和建议，帮助用户决定应该购买何种商品[5]。电子商务推荐系统作为一种强大的新兴技术，能够帮助用户找到他们喜爱的商品，反过来也提高了电子商务网站的销售额，因此很快成为电子商务网站一种至关重要的工具[32]。虽然目前还没有公认的电子商务推荐系统的标准定义，但众多学者和研究人员给出的说法大同小异。本书综合众多文献的描述，给出电子商务推荐系统的定义如下。

定义 1.1(电子商务推荐系统) 电子商务推荐系统是一套用于电子商务商品推荐的软件系统，通过对能够反映用户兴趣偏好的数据进行统计分析和机器学习，向用户推荐适合其兴趣的商品项或商品项集合，从而将网站浏览者转变为购买者、提高网站交叉销售能力及建立客户忠诚度。

对于电子商务推荐系统的任务(即其需要解决的问题)，本书给出如下形式化描述：

Given：

对于电子商务站点 ω，令其商品项集合 $I=\{I_i \mid I_i \in I, i=1,2,\cdots,|I|\}$。对于站点用户 $u_a \in U$，U 为 ω 的用户集合，令 $v(u_a, I_i)$ 表示 u_a 对 I_i 的兴趣度，则对于 u_a，存在一个兴趣度函数 F_{u_a} 和有限商品项集合 $I_{u_a} \subset I$，使得

$$F_{u_a}=\underset{I_i \in I_{u_a}}{\operatorname{argmax}} v(u_a, I_i)$$

式中，$I_i \in I_{u_a}$，$i=1, 2, \cdots, |I_{u_a}|$。

Goal：

如何求得 I_{u_a}、$v(u_a, I_i)$？

如同其他搜索引擎一样，电子商务推荐系统在实际应用中存在两种可能的错误[33]：一是拒真(false negatives，即错误否定)，即有些用户喜欢的商品未能推荐出来；二是纳伪(false positives，即错误肯定)，即有些不被用户喜欢的商品却被推荐出来。在这两种错误中，最需要避免发生的是纳伪，因为错误的商品推荐将导致用户生气乃至离开站点。电子商务网站通常有用户愿意购买的众多商品，因此没有理由冒险推出用户不喜欢甚至厌恶的商品，宁可推荐得“保守”一点。

1.3.2 用户偏好数据

电子商务推荐系统需要将能够反映用户兴趣偏好的信息作为输入数据以生成推荐结果。真实的用户兴趣偏好信息对推荐结果的准确程度起关键作用。一般而言，用户的注册数据、交易数据、评分数据、购物篮数据、浏览数据等都可以作为电子商务推荐系统的输入，具体可以分为显式评分(explicit ratings)和隐式评分(implicit ratings)两类。

(1) 显式评分

显式评分要求用户向电子商务推荐系统提供自己的兴趣偏好信息，主要是用户对系统给出的推荐项进行反馈和评价，也包括用户注册时提供的人口统计学数据(demographic data)和感兴趣领域。曾春等[34]将其称为显式跟踪。基于用户的显式评分数据，系统能够向其提供有针对性的推荐服务。在实际生活中，用户购买某商品后未必对该商品满意，因此有些推荐系统让用户重新对已购商品给出评价，使得系统能产生更精确的推荐。例如，CDNow允许用户对其已购商品作出“拥有并且喜欢它”(own it and like it)或“拥有但不喜欢它”(own it but dislike it)的区分[5]。

但是显式评分也存在若干不足[35]：① 用户需要停止浏览和阅读以进行显式评分输入；② 如果用户感到不能从提供评分中得到好处，将不会提

供评分；③用户浏览的项目大大多于他们所评分的项目；④协同过滤要求每个项目都有相当数量的评分才能提供精确的推荐结果。

(2) 隐式评分

相对于显式评分，隐式评分具有更高的自动化程度[3]，因为显式评分对于系统用户而言是一个额外负担。隐式评分需要电子商务推荐系统通过自动学习用户行为信息来了解用户兴趣偏好，从而缓解用户评分数据的稀疏性，用户甚至感觉不到推荐系统的存在。曾春等[34]将其称为隐式跟踪。隐式评分包括用户行为分析(查询/访问页面、在页面中搜索文本、保存/删除书签、点击/移动鼠标、拖动滚动条、剪切/粘贴/保存/打印页面、用E-mail发送页面等)、Web日志挖掘(获取页面点击次数、停留时间、访问顺序等)、购物篮数据、购买历史等，系统将这些用户信息转化为反映用户兴趣偏好的数据并应用于推荐生成。例如，GroupLens系统[3, 9, 36, 37]的研究人员以及Adomavicius和Tuzhilin[38]均指出用户对文档的阅读时间有助于预测该用户对文档的评分；Nichols[39]总结了购买、保存/打印、回复、查询等13种隐式评分信息类型，Oard和Kim[40]则进一步将可观察的隐式反馈行为归纳为审查(examination)、保持(retention)、参考(reference)三大类；Claypool等[35]将能够反映用户兴趣的隐式评分信息称为隐式兴趣指示器(implicit interest indicators)并开发了一个Web浏览器“Curious Browser”，可以对用户花在网页上的阅读时间以及对网页滚动条、鼠标、键盘等相关操作进行统计分析以推测用户对该网页的兴趣大小，他们通过实验证明用户的阅读时间和在网页上拖动滚动条的次数确实能较好地反映用户兴趣；Yoda系统[41]则采用遗传算法(genetic algorithm)来学习用户的访问行为，从而自动调整用户描述(user profile)；此外，客户对商品的退货行为也可认为是对该商品作出了“否定评分”(negative ratings)[4]。

隐式评分的优势在于[35]：①免去了用户对项目进行评分的开销