



中青年经济学家文库
ZHONGQINGNIAN JINGJIXUEJIA WENKU

电子商务推荐系统的理论与应用研究

崔春生 / 著

ANZI SHANGWU TUIJIAN XITONG DE LILUN YU YINGYONG YANJIU



经济科学出版社
Economic Science Press

014005582

F713.36
874

电子商务推荐系统的 理论与应用研究

崔春生 著



F713.36

874

经济科学出版社



北航

C1693445

图书在版编目 (CIP) 数据

电子商务推荐系统的理论与应用研究 / 崔春生著.

—北京: 经济科学出版社, 2013. 9

ISBN 978 - 7 - 5141 - 3886 - 3

I. ①电… II. ①崔… III. ①电子商务 - 研究
IV. ①F713. 36

中国版本图书馆 CIP 数据核字 (2013) 第 243139 号

责任编辑: 刘 瑾

责任校对: 徐领柱

责任印制: 邱 天

电子商务推荐系统的理论与应用研究

崔春生 著

经济科学出版社出版、发行 新华书店经销

社址: 北京市海淀区阜成路甲 28 号 邮编: 100142

总编部电话: 010 - 88191217 发行部电话: 010 - 88191522

网址: [www. esp. com. cn](http://www.esp.com.cn)

电子邮件: [esp@ esp. com. cn](mailto:esp@esp.com.cn)

天猫网店: 经济科学出版社旗舰店

网址: [http://jjkxcbs. tmall. com](http://jjkxcbs.tmall.com)

固安华明印刷厂印装

880 × 1230 32 开 6.75 印张 200000 字

2013 年 10 月第 1 版 2013 年 10 月第 1 次印刷

ISBN 978 - 7 - 5141 - 3886 - 3 定价: 28.00 元

(图书出现印装问题, 本社负责调换。电话: 010 - 88191502)

(版权所有 翻印必究)

前 言

推荐系统是 20 世纪末随着互联网和电子商务的发展而迅速发展起来的一种技术,其直观的作用体现在避免用户迷失于海量的产品信息中,帮助用户选择需要的产品,提高产品的销售量。伴随电子商务的快速发展、“大数据”时代的到来,推荐系统逐渐成为电子商务发展中不可缺少的一项关键技术。然而,理论滞后于应用的不争事实给推荐系统的研究带来了诸多不便。

本书研究了基于 Vague 集的电子商务推荐系统的理论与应用,结合电子商务推荐系统对覆盖率和准确率的要求,借助 Vague 集的方法,另辟蹊径,从推荐系统的输入、算法、输出三个环节探讨其理论与应用,并借助 MATLAB 验证算法的可靠性。

全书共分为 8 章,具体章节内容如下:第一章,主要介绍本书的选题背景、研究意义以及研究的思路和运用的各种方法,并从本书的主体内容框架出发,描述本书研究特色。第二章,相关基础理论及推荐系统研究现状。第三章,基于 Vague 集的产品描述及排序,介绍借助 Vague 值实现非个性化产品推荐的突破口。第四章,推荐系统输入研究,探讨不同输入方式的用户聚类方法。第五章,基于 Vague 集的内容推荐算法研究,利用 Vague 集基本理论,找到基于 Vague 集的一般内容推荐算法。第六章,基于泛函网络的组合推荐算法研究,通过寻找泛函网络方法与推荐算法的结合点,构架推荐系统前融合的泛函网络拓扑结构。第七章,推荐系统输出结果分析,采用集团序的方法,研究具有个性化、差异化的 N 值取值

策略。第八章,结论与展望。从系统输入、推荐算法、系统输出三个方面对本书的研究成果做简单的总结,并从研究内容以及推荐系统的发展两个角度分析以后的研究方向,为作者的进一步研究奠定基础。

本书可作为高等院校管理科学、信息科学、系统工程、数量经济等专业本科生和研究生的教材,也可作为相关专业的教师、科研人员和企业管理人员的参考用书。

衷心感谢我所参阅过的所有文献的作者,你们的工作开阔了我的视野和思路!

本书的内容汇集了笔者近几年的研究成果,虽则尽力,但水平所限,差错疏漏之处在所难免,敬请各位专家和读者指正。

目 录

第一章 绪论	1
一、研究背景和意义	1
二、研究思路及方法	9
三、本书的主要内容	11
第二章 推荐系统研究现状及相关基础理论	17
一、推荐系统研究现状	17
二、Vague 集基本理论	33
三、推荐系统国内外研究趋势	39
四、本章小结	44
第三章 基于 Vague 集的产品描述及排序	45
一、推荐系统描述	45
二、电子商务推荐系统与 Vague 集的对接	50
三、基于 Vague 值的非个性化产品推荐	57
四、非个性化产品相似性的度量	65
五、关于 Vague 相似性的进一步思考	69
六、本章小结	73

第四章 推荐系统输入研究	75
一、推荐系统输入存在的问题	75
二、基于显式评分输入的用户聚类分析	81
三、基于隐式浏览输入的用户聚类分析	94
四、本章小节	104
第五章 基于 Vague 集的内容推荐算法研究	106
一、基于内容的推荐算法	106
二、基于内容推荐算法的 Vague 求解	110
三、案例应用	123
四、本章小节	128
第六章 基于泛函网络的组合推荐算法研究	130
一、泛函网络	130
二、推荐系统组合推荐算法	136
三、基于泛函网络的推荐算法结构模型	138
四、基于泛函网络的推荐系统学习过程研究	141
五、本章小节	148
第七章 推荐系统输出结果分析	149
一、Top- N 技术发展	149
二、基于集团序方法的推荐输出	151
三、案例计算	158
四、本章小结	166
第八章 结论与展望	168
一、结论	168

二、展望.....	170
附录.....	174
参考文献.....	190

表索引

表 2-1 隐式反馈行为分类表	19
表 2-2 主要推荐技术	25
表 3-1 用户/产品评分矩阵表	46
表 3-2 推荐与喜欢关系表	47
表 3-3 推荐与接受关系表	48
表 3-4 浏览过程及其推断表	52
表 3-5 产品基本数据	55
表 3-6 产品统计数据	56
表 3-7 记分函数计算结果	63
表 3-8 相似度计算结果	68
表 4-1 用户/产品聚类组合表	90
表 4-2 用户社区密度表	92
表 4-3 各种社区结构的紧致与分离性效果函数	92
表 5-1 主要推荐技术优缺点	109
表 5-2 用户评分表	122
表 5-3 用户情况表	124
表 5-4 用户 2 及用户 5 部分预测值	125
表 5-5 用户 35 及用户 40 部分预测值	126
表 5-6 MAE 值	127
表 6-1 训练集用户评分预测值	146
表 7-1 用户 804 和用户 901 的预测值	158
表 7-2 用户 804 和用户 901 的推荐度	160

表 7-3	$N=10$ 时推荐结果	162
表 7-4	用户 804 推荐结果质量评价	165
表 7-5	用户 901 推荐结果质量评价	165

图索引

图 1-1	中国网民规模	2
图 1-2	本书研究思路	10
图 1-3	内容结构图	15
图 2-1	Vague 集示意图	35
图 4-1	产品分类图	98
图 5-1	基于内容推荐系统描述	112
图 5-2	基于 Vague 集内容推荐算法步骤描述	114
图 5-3	不同 α 值条件下 MAE 对比图	128
图 6-1	人工神经网络结构图	132
图 6-2	泛函网络结构图	133
图 6-3	泛函网络学习过程图	135
图 6-4	基于内容推荐算法结构图	139
图 6-5	基于内容推荐算法泛函网络图	139
图 6-6	协同过滤推荐算法结构图	140
图 6-7	协同过滤推荐算法泛函网络图	141
图 6-8	前融合推荐算法结构图	142
图 6-9	前融合推荐算法泛函网络图	142
图 6-10	程序主界面	145
图 6-11	用户 901 推荐结果	147
图 7-1	集团排序流程	154
图 8-1	“灰羊”问题	172

第一章

绪论

研究背景和意义

(一) 研究背景

1. 电子商务迅猛发展。自 20 世纪末互联网开始至今, Internet 以无法想象的速度迅猛增长, 接入 Internet 的服务器数量和 World-Wide-Web 上的网页数目都呈现指数增长的态势, 互联网改变了传统的信息传播模式和商品交换模式。

中国互联网络信息中心 (CNNIC) 发布的《第 31 次中国互联网络发展状况统计报告》显示, 截至 2012 年 12 月, 我国网民规模达 5.64 亿人 (如图 1-1 所示), 普及率达到 42.1%, 保持低速增长, 网民规模较 2011 年年底增长 5 090 万人, 较 2011 年年底提升 3.8%。2012 年我国手机网民数量为 4.2 亿, 年增长率达 18.1%, 远超网民整体增幅。此外, 网民中使用手机上网的比例也继续提升, 由 69.3% 上升至 74.5%, 其第一大上网终端的地位更加稳固,

但是手机网民规模与整体 PC 网民（包括台式电脑和笔记本电脑）相比还有一定差距。截至 2012 年 12 月，我国网络购物用户规模达到 2.42 亿，网络购物使用率提升至 42.9%。与 2011 年相比，网购用户增长 4 807 万人，增长率为 24.8%。在网民增速逐步放缓的背景下，网络购物应用依然呈现快速的增长势头。中国网上零售市场规模达 13 205 亿元，较 2011 年增长 64.7%，预计 2013 年全国网络零售市场交易额则有望达到 18 155 亿元。数据显示，2012 年中国网络零售市场交易规模已占到当年社会消费品零售总额的 6.3%。而这个比例在 2011 年仅为 4.4%。这也或许意味着电商改变零售业格局的开始。中国电子商务研究中心还预计，这一比例将保持扩大态势，到 2013 年达到 7.4%。

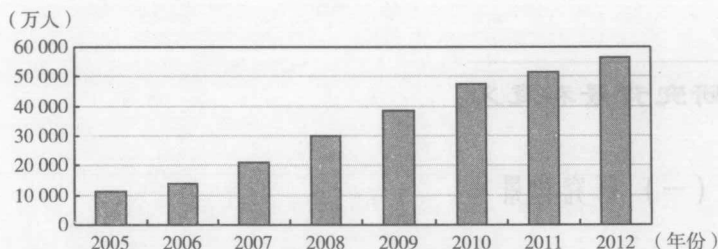


图 1-1 中国网民规模

在 B2C 模式中，电子商务网站除了使企业节省了传统经营模式下的实体投资成本外，还拥有巨大的优势，即消除了传统店面的产品陈列空间限制，为购物者提供了一个庞大的（也可以说是无限大的）产品陈列柜台供其选择。用户通过访问电子商务网站，可以享受足不出户选购产品的快乐和方便。

2. 信息超载推动推荐系统发展。互联网规模和覆盖面的迅速增长为人们带来了巨大的惊喜与变革，使人们方便地得到各式各样的海量信息，例如，Netflix 上有数万部电影，Amazon 上有数百万本书，淘宝有四十多万家店铺。大数据时代的到来，人们迷失于海

量的数据中,数据量的增长已经不是以我们所熟知的多少 G 和多少 T 来描述了,而是以 P (1 千 T), E (1 百万 T) 或 Z (10 亿 T) 为计量单位。万维网自 1991 年出现以来已经发展成为一个巨大的全球化的信息空间。“新摩尔定律”指出:Internet 上的信息正以每 6 个月翻一番的速度爆炸性地增长着。很显然,信息爆炸时代已经到来,信息爆炸带给我们的不仅仅是便利与惊喜,同时也是困惑与无奈。信息的爆炸使得信息超载 (Information-overload) 现象日益凸显。

所谓信息超载是指:信息同时呈现使得用户无法从中获取对自己有用的内容,信息使用效率反而降低。现有的很多网络应用,比如门户网站、搜索引擎和专业数据索引等,本质上都是帮助用户过滤信息,然而这些工具只满足主流需求,没有个性化的考虑,仍然无法很好地解决信息超载的问题。目前,解决信息超载的常用方法有信息检索、信息过滤和推荐系统。

信息检索 (Information Retrieval) 通常是由用户以关键字的形式提出一个查询,计算机通过关键字匹配,返回可能相关的结果。用户查看返回的结果,并且通过自己的分析,获取所需的信息。这种技术的优点在于查询领域广泛,并可以快速返回相关信息;缺点是准确率较低,用户需要花费大量的时间查找、选择。

信息过滤 (Information Filtering),也称为信息的选择性传播,该技术关注的是用户一段时间内的固定信息需求。信息过滤不能处理非结构化信息,一般只处理半结构化和结构化的信息,并且面对的是相对固定,且具有相对稳定的规律性和周期性的任务,这种技术记录并利用用户的偏好信息。其优点在于对操作者的要求很低,并具有一定的个性化推荐的功能;缺点是缺乏对新信息的推荐能力,需要信息的内容描述。

推荐系统 (Recommender System) 依赖于用户对产品的偏好数据集来实现,并且不要求这些产品必须有文本描述。因此推荐系统,包括个性化搜索 (Personalization Retrieval),被认为是当前解

决信息超载问题最有效的工具之一。目前,国际上很多知名网站,如 Microsoft、AOL、IBM、Google、CNN、Lycos 等,都推出了个性化功能;很多电子商务网站,比如 Amazon、Ebay、Expedia、Bestbuy 等,也发现了个性化服务的巨大商机,开始提供个性化服务;国内一些知名的门户网站,如新浪、搜狐、网易网也推出了个性化服务。

目前,我国电子商务网站中可以看到推荐形式很多,从个性化程度上区分包括非个性化推荐、个性化推荐。例如进入搜狐、新浪等门户网站就可以看到很多网站链接,这些形式在某种意义上就属于一种在线非个性化推荐形式,再如艾瑞的网站排名 (<http://www.iwebchoice.com/>)、卓越亚马逊的热荐商品 (<http://www.amazon.cn/>)、百度 MP3 中的新歌 TOP100 (<http://list.mp3.baidu.com/top/top100.html>) 等,都是典型的非个性化在线推荐。酷讯网 (<http://www.kuxun.cn/>) 中的机票查询产生的机票价格对比、淘宝网 (<http://www.taobao.com/>) 中的其他顾客浏览产品、互动网 (<http://www.china-pub.com/>) 中的相近产品等则属于个性化在线推荐。

从实时性角度来看,包括在线推荐、线下推荐等。例如,去哪儿网 (<http://www.quna.com/>) 当中的机票预约方式,系统将会根据预约信息,将推荐结果发送到用户的邮箱。国内各大航空公司、国内知名招聘网站以及诸如凡客诚品 (<http://www.vancl.com/>)、艺龙 (<http://www.elong.com/>)、橡果国际 (<http://www.chinadrvtv.com/>) 等专业购物网站也会经常向注册用户的邮箱中发送产品信息,这些产品信息有些是非个性化的推荐信息,有些则是基于用户历史行为的个性化信息。这些都属于线下 E-mail 推荐形式。

因此随着网络以及电子商务的发展,推荐系统逐渐在第三产业中的娱乐、旅游、金融、计算机服务、消费品、教育、交通运输、体育等各个领域遍地开花,并且出现了迅速发展的趋势。

3. 推荐系统及其优点。推荐系统的迅猛发展,一方面源于信息超载的推动,另一方面源于 Web 2.0 技术的成熟。有了这个技术,用户不再是被动的网页浏览者,而成为电子商务的主动参与者。推荐系统与以搜索引擎为代表的信息检索系统最大的区别在于:(1) 搜索注重结果之间的关系和排序,但推荐系统除此之外还研究用户模型 (User Profile) 和用户的喜好,基于社会网络 (Social Network) 进行个性化的计算 (Personalization); (2) 搜索的进行由用户主导,包括输入查询词和选择结果,如果结果不好,用户会修改查询条件再次搜索;而推荐是由系统主导用户的浏览顺序,引导用户发现需要的结果。高质量的个性化推荐系统能使用户对系统产生强烈的依赖感。因此,推荐系统可以为用户提供个性化的推荐服务,也可以与用户建立长期稳定的内部互动关系,提高用户对系统的忠诚度,防止用户流失。

推荐系统的研究直到 20 世纪 90 年代才被作为一个独立的概念提出来,并逐渐形成为一门独立的学科。目前被广泛引用的推荐系统的非形式化概念是雷斯尼克 (Resnick) 和瓦里安 (Varian) 在 1997 年给出的:“推荐系统是利用电子商务网站向客户提供产品信息和建议,帮助用户决定应该购买什么产品,模拟销售人员帮助客户完成购买过程。”可以看到,推荐系统有 3 个组成要素:推荐候选对象 (产品)、系统用户和推荐算法。用户可以向推荐系统主动提供个人偏好信息或推荐请求 (显示输入),或者用户不提供,而是系统主动采集 (隐式浏览输入)。推荐系统可以使用不同的推荐方法进行推荐,如将采集到的用户信息和产品数据进行计算得到推荐结果,或者直接基于已建模的知识数据库进行推荐,将推荐结果返回给用户使用。

作为提高电子商务销售能力的强大引擎,推荐系统的优点主要体现在以下三个方面:

(1) 将电子商务系统的浏览者转变为购买者 (Converting Browsers into Buyers): 具有明确购物目标的用户也许可以借助搜索

引擎找到自己需要的东西,但对于大多数漫无目的的浏览者,或是需求不确定的购买者来说,很少有耐心在海量的产品目录中寻找是否有自己喜欢的产品。推荐系统可以通过合适的推荐方法,将一个浏览者变为购买者或潜在的客户。

(2) 提高电子商务网站的交叉销售能力 (Cross-Sell): 推荐系统在用户购买产品的过程中,根据用户的喜好向他提供其他相似或有价值的产品推荐,用户能够从系统提供的推荐表中购买自己确实需要,但在推荐之前没有购买计划的产品,从而有效提高电子商务系统的交叉销售能力。

(3) 提高客户的忠诚度 (Building Loyalty): 客户忠诚度是实体商务和电子商务共同追求的目标。与传统的商务模式相比,电子商务系统使得用户拥有更多的选择,用户更换商家极其方便,只需简单地点击鼠标就可以在不同电子商务网站之间转换。电子商务推荐系统分析用户的购买习惯以及产品特征,根据用户的需求及产品属性向用户推荐有价值的产品。如果电子商务推荐系统的准确率很高,那么用户会对推荐系统产生强烈的依赖感,进而有效提高客户的忠诚度,防止客户流失。

B2C 电子商务是推荐系统最典型的应用领域,并且具有良好的发展和应用前景。商家根据用户的兴趣、爱好推荐用户可能感兴趣或满意的产品(如书籍、音乐、电影等)。然而用户的需求通常是不明确的、模糊的,如果商家能够把满足用户模糊需求的产品推荐给用户,将不确定需求转化为确定需求,就可以把用户的潜在需求转化为现实需求,从而达到提高产品销售量的目的。目前,几乎所有的大型电子商务系统,如 Amazon、eBay、互动、淘宝等,都不同程度地使用了各种形式的推荐系统。

(二) 研究意义

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,美国学者彭诺克(Pennock, 1999, 2000)、杨戈(Yager, 2003)和日本学者饭岛淳一(2007)等人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。

电子商务推荐系统的理论滞后于应用的发展,带来了理论研究的断层,所以本书的研究在理论上有一定的价值和意义。同时,本书的研究是在大数据时代的到来、买方市场全球化市场环境业已形成、电子商务销售比重不断攀升、物流产业快速发展以及物联网逐渐形成的背景下进行的,因此,本书的研究在应用方面具有很强的价值和意义。

1. 理论意义。

(1) 相似性研究。目前推荐系统研究中还存在一些理论难点,如推荐质量问题。推荐系统评价的标准是查全和查准,查全问题已经受到广泛的重视,并在理论界取得了相当的研究成果,但查准问题一直困惑着理论研究者。应用广泛的基于内容的推荐系统以及协同过滤推荐依赖的是产品(或用户)的相似性,因此探讨合适的问题描述工具以及选择科学的相似性理论是解决这一问题的重要理论突破点。林德康(Dekang Lin, 1998)曾从信息论的角度对相似性给出了一个统一的、与应用领域无关的非形式化的定义。他认为, X 与 Y 之间的相似性一方面与它们的共性相关,共性越多,相似性越高;另一方面与它们的区别相关,区别越大,相似度越低;当 X 与 Y 完全相同时,相似性达到最大值。推荐系统研究领域常用的相似性计算方法有向量空间相似度、Pearson 相关、Spearman 相关、熵、余弦方法、相关性方法等。同时,Vague 集研究领域大量研究的都是 Vague 集(值)之间的相似性。Vague 集(值)之间的相似性度量是 Vague 集理论在模糊决策、模式识别、模糊推理、机器学习、社会选择等方面应用的理论基础,也是近年来该领域学者们普遍关注的热点。可见,借助于 Vague 集的理论研究成果

及研究方法,对于推荐系统的查准问题(相似性)会有一定的推动作用。

(2) 稀疏矩阵的约简。稀疏矩阵是影响推荐系统发展的一个瓶颈问题,推荐系统中存在大量的产品和用户,但用户一般只能给很少的产品打分,而每一个产品也只能获得很少用户的打分,因此评分矩阵非常稀疏。矩阵稀疏性容易导致相似用户或相似产品难以获得,相似关系不明确以及部分信息容易丢失等问题。在显性评分输入的研究中,采用用户聚类的方法,借助社区思想以及矩阵调整手段,寻求稀疏评分矩阵约简的方法。该方法的推广将对稀疏矩阵约简的研究有一定的推动作用。

(3) 基于内容推荐及产品特征表示。相比协同过滤推荐算法,基于内容推荐算法的研究略显滞后。一方面是因为协同过滤算法产生较早,并取得了良好的推荐效果,进而推动了该算法的发展;另一方面是因为基于内容推荐系统中特征提取问题一直是困扰学者的一个难点。在信息检索领域中,文本等对象特征的提取技术经由关键词得到,已经很成熟。但推荐系统面对的对象不一定具有文本特征,尤其是网络中广泛存在的多媒体数据,如音乐、视频、图像、电影等,自动化的特征提取方法需要结合多媒体内容分析领域的相关技术并借助于不确定性问题解决方法。Vague 集方法是一种很好的解决不确定性问题的工具,借助于 Vague 集可以从新的角度为基于内容推荐系统的研究开辟新的研究道路。

(4) 前融合推荐算法。推荐算法是推荐系统研究中的关键问题,日趋成熟的各类推荐算法还不能完全满足推荐系统的质量要求,很多学者提出采用融合的方法研究推荐算法。无论是前融合方式、中融合方式还是后融合方式,选择合适的融合工具具有一定的理论指导意义。本书探讨符合顾客心理变化及心理需求的前融合方式,可以很好地描述推荐系统问题,以期获得良好的推荐结果。

(5) Vague 集应用。Vague 集理论在研究的过程中已经取得了重大的理论突破,但是在应用方面缺乏一定的解释机制。将推荐系