基于优化支持向量机的个性化推荐研究

RESEARCH ON PERSONALIZED RECOMMENDATION BASED ON OPTIMIZED SUPPORT VECTOR MACHINE

王喜宾　文俊浩◎著

重庆大学出版社

基于优化支持向量机的个性化推荐研究

JIYU YOUHUA ZHICHI XIANGLIANGJI DE GEXINGHUA TUIJIAN YANJIU

王喜宾　文俊浩◎著

重庆大学出版社

内容提要

在实际应用中，个性化推荐存在小样本、高维度和非线性等问题。针对这些问题，本书提出了基于支持向量机的个性化推荐方法，实现对项目内容与用户行为信息的综合分析。针对不同的推荐问题先后提出了基于支持向量分类机的推荐方法、基于支持向量机先分类再回归的推荐方法、基于平滑技术和核减少技术的对称支持向量机推荐方法以及基于主动学习的半监督直推式支持向量机推荐方法。

图书在版编目(CIP)数据

基于优化支持向量机的个性化推荐研究/王喜宾，文俊浩著.
—重庆：重庆大学出版社，2017.4
ISBN 978-7-5689-0484-1

Ⅰ.①基… Ⅱ.①王…②文… Ⅲ.①向量计算机—研究
Ⅳ.①TP38

中国版本图书馆 CIP 数据核字(2017)第 062179 号

基于优化支持向量机的个性化推荐研究

王喜宾 文俊浩 著
策划编辑：鲁 黎
责任编辑：陈 力 版式设计：鲁 黎
责任校对：秦巴达 责任印制：赵 晟

*

重庆大学出版社出版发行
出版人：易树平
社址：重庆市沙坪坝区大学城西路 21 号
邮编：401331
电话：(023) 88617190 88617185(中小学)
传真：(023) 88617186 88617166
网址：http://www.cqup.com.cn
邮箱：fxk@cqup.com.cn (营销中心)
全国新华书店经销
万州日报印刷厂印刷

*

开本：720mm×960mm 1/16 印张：11.5 字数：155 千
2017 年 4 月第 1 版 2017 年 4 月第 1 次印刷
ISBN 978-7-5689-0484-1 定价：45.00 元

前言

当前，主流的个性化推荐方法包括：基于协同过滤的方法和基于内容的方法。协同过滤的方法通过计算用户兴趣偏好的相似性，为目标用户过滤和筛选感兴趣的物品。它主要是基于用户的行为信息进行推荐，而没有真正利用物品的内容信息和用户的标签信息，同时也存在着数据稀疏和冷启动等问题。基于内容的推荐，本质上则是一种信息过滤技术，仅仅通过学习用户历史选择的物品信息，缺乏对用户反馈信息的挖掘，这也往往会造成推荐结果过度特殊化。

个性化推荐在实际应用中存在小样本、高维度和非线性等问题，鉴于支持向量机在小样本学习，解决非线性问题时可以较好地克服“维度灾难”，以及处理高维稀疏数据方面的优势，本书阐述了基于支持向量机的个性化推荐方法，实现对项目的内容信息以及用户行为信息的综合分析与挖掘。

首先，针对传统的协同过滤推荐方法存在相似度计算方式单一，不易利用项目的内容信息和冷启动等问题，提出了利用支持向量分类机方法来代替传统的相似度计算，不仅考虑了用户的行为信息，而且也利用了项目的内容信息和用户的人口统计学信息。同时，利用带收缩因子的动态惯性权重自适应粒子群优化算法对支持向量分类机的参数进行优化，以期提高推荐模型的准确率。

其次，针对实际应用中，不仅需要推荐列表，而且还需要详细的评分信息（在某种程度上反映了用户的兴趣度），提出了基于支持向量机先分类再回归的推荐方法。该方法根据“用户-项目”关联关系信息，构造特征向量并训练一个分类模型，预测项目的类别，形成一个初始推荐列表；然后，在该推荐列表上建立一个回归模型，预测项目的具体评分；并且在建立分类模型和回归模型时，采用提出的带进化速度和聚集度的自适应粒子群优化算法来优化预测模型。

再次，针对大规模数据中的推荐效率和实时性等问题，提出了基于平滑技术和核减少技术的对称支持向量机推荐方法。该方法采用平滑技术对对称支持向量机进行变换，避免了大规模矩阵的求逆运算，降低了算法的时间复杂

度。为进一步提高大规模数据的处理能力，采用了核减少技术进一步降低算法的时间复杂度和空间复杂度。同时，鉴于用户的兴趣和偏好会随着时间、地点等不断演化，对推荐系统的实时性要求较高。为此，引入反馈机制，将用户的评分数据及时加入历史数据中，并设计训练规则，启动模型的重新训练，使模型具有一定的自适应能力，从而提高模型的推荐质量。

最后，针对个性化推荐中有标签数据价值高但稀少，同时对无标签数据标注存在耗时、耗力、代价高等问题，提出了基于主动学习的半监督直推式支持向量机推荐方法。首先，挖掘用户评价信息中有价值的评论信息，并将其加入“用户-项目”关联关系数据集中；然后，采用批采样的主动学习策略对大量无标签的“用户-项目”数据中具有最高信息量的样本进行查询并标注，获得对分类器提升最有价值且尽可能小的样本集，从而降低样本标记的代价，提高了分类器的性能。同时，为了更好地利用无标签数据的分布特征，在目标函数中引入基于图的流形正则项，进一步提升了模型的推荐效果。

本书受国家自然科学基金面上项目“基于异构服务网络分析的 Web 服务推荐研究”（NO. 61379158），重庆市教委科学技术研究项目“大数环境下基于用户行为分析和上下文感

知的个性化推荐研究”(NO. KJ1600437)等项目的资助。

限于本书作者的学识水平,书中疏漏之处在所难免,恳请读者批评指正。

著　者

2016 年 9 月

目录

第1章 绪 论

1.1 研究背景与意义

随着互联网技术与信息技术的发展，人们的生活、学习和生产方式发生了巨大改变，互联网为人们提供了丰富的信息资源，使人们可以随时随地通过互联网获取信息。但是面对如此繁多、质量参差不齐的信息资源，人们淹没在了信息的海洋中，难以找到自己感兴趣的信息资源，甚至使他们忘却或不能明确自己的真实需求[1]，导致了信息丰富但选择困难的两难境地，产生了“信息过载”问题[2,3]。如何帮助用户从庞大的信息资源中，快速、准确地找到自己所需的信息资源，成了当今信息技术研究中的一个重大挑战，也成了学术界研究的热点和难点。为此，先后提出了信息检索技术和信息过滤技术来解决该问题[4,5]。

信息检索技术在一定程度上可以帮助用户找到自己所需要的信息，并得到了成功和广泛的应用。在日常生活中，常用的搜索工具有百度、雅虎、好搜、必应和谷歌等，都属于信息检索系统的范畴[6]。但是随着信息

量的急剧增长，信息过载问题越发严重，尤其是到了大数据时代，信息检索系统更加不能满足人们的需求，于是信息过滤技术应运而生[7]。它根据用户提供的需求或兴趣偏好，对动态信息资源进行筛选，自动检索出符合用户需求或感兴趣的信息，并将其呈现给用户[8]。

个性化推荐系统（Personality Recommender System，PRS）作为信息过滤的一种重要方式，为解决“信息过载”问题提供了很好的解决方法，也是当前广泛应用的方法之一，被广大电子商务网站和个性化网站所采用[9]。个性化推荐系统与信息检索系统（以搜索引擎为代表）的主要区别如下：

①PRS 根据收集到的用户特征数据（如行为特征等），建立个性化的推荐模型，然后将符合用户偏好或需求的信息资源推荐给用户；而信息检索系统关注的重点是检索结果之间的客观关系和排序。

②在信息检索系统中，用户是主导者，必须明确自己的需求，输入查询条件，系统返回匹配的信息资源，然后用户筛选返回的结果。如果返回的结果不符合用户的需求，则可以修改查询条件继续查询。而在个性化推荐系统中，系统根据收集的用户偏好特征，为用户推荐他们感兴趣的信息资源，是一种引导性的信息消费，同时用户可能并不知道推荐信息的存在，甚至不知道怎样才可以找到这样的信息资源。此外，个性化推荐系统具有一定的个性化和实时性，可以根据用户偏好特征的变化，调整其推荐策略，及时为用户提供最新的推荐列表。

由于个性化推荐技术的特定优势，在各大网站得到了广泛应用，例如，在 YouTube、土豆、优酷等视频网站中为用户推荐视频；在淘宝、京东、亚马逊等电商网站中为用户推荐喜欢的商品；网易、新浪、雅虎等新闻网站根据用户的浏览行为为用户推荐感兴趣的新闻等。这些成功应用的反作用又进一步促进了个性化推荐技术的发展，成了当今各大主流网站不可缺少的一种信息服务形式。

相对于传统的信息门户时代，用户根据兴趣和爱好寻找所需信息资源所付出的代价非常高，同时信息的价值也一直被忽视。尤其到了大数

据时代，各大网站收集的信息量和种类也越来越多，如何从这些海量的信息中发掘用户的兴趣爱好和行为模型成了各大网站推销自己和维持用户的重要手段。而个性化推荐技术能够根据收集的“用户-项目”关联数据，包括用户的浏览历史记录、评价信息、用户的人口统计学特征、项目的内容信息等建立个性化的推荐模型，从而为用户提供个性化的服务，向用户推荐最适合的项目，使用户在浏览网站的同时，能很快发现自己感兴趣的内容，提升了网站的服务质量，增进了用户的黏着性。

正是由于这些巨大的需求，个性化推荐技术自20世纪90年代初期被提出以来，就得了广泛的研究和应用。其中，具有代表性的方法主要有基于协同过滤的推荐和基于内容的推荐。然而前者存在冷启动、数据稀疏性等问题；而后者在内容不易分析时，将无法很好地分析和推荐。同时，也有很多采用机器学习方法来实现个性化的推荐，例如聚类、关联规则分析、回归分析、决策树和神经网络等。基于机器学习的个性化推荐方法可以很好地解决相似度计算方式单一、相似度计算复杂度高、不易利用用户的标签信息以及用户的人口统计学信息等问题，而用户的标签信息和人口统计学信息在解决冷启动方面相当有效，是发现用户潜在兴趣的有价值的信息。

1.2 国内外研究现状

关于个性化推荐系统的产生和应用可以追溯到20世纪70年代，当时协同过滤算法已具雏形。到了20世纪90年代，个性化推荐系统的理论框架已经基本成熟。

早期使用协同过滤算法的推荐系统Tapestry[10]是由美国施乐公司的Palo Alto研究中心开发的邮件过滤系统，但该系统需要人工对邮件进行标注，自动化程度低，导致该系统只能用于小型的邮件系统。

卡内基梅隆大学联合莲花公司开发了主动协同过滤系统[11]，并将其

融合到了办公系统中,实现了部分数字文档的个性化推荐。

明尼苏达大学的 GroupLens 研究团队开发的 GroupLens[12] 推荐系统,采用基于自动协同过滤的推荐方法,在 Usenet 新闻、文章推荐中得到了成功应用。后来,又创建了 Movielens 数据集用于学术研究,该数据集分为 3 个不同的版本用于不同的科研目的。除了卡内基梅隆大学和明尼苏达大学外,还有密歇根大学、纽约大学、微软研究院、谷歌等研究机构对推荐系统的发展也作出了巨大贡献。特别是从 2006 年起,密歇根大学开设了推荐系统相关课程。

美国贝尔通信研究所开发的视频系统(Video Recommender)[13] 实现了电影的个性化推荐,该系统采用电子邮件的方式来收集用户对电影的评分数据。在电影推荐中,备受关注的是 2006 年启动的一场历时三年的 Netflix 100 万美元大赛,这个机器学习和数据挖掘的大赛主要目的是解决电影评分预测问题,并奖励使 Cinematch 推荐系统[14] 的准确率提高 10%的团队或个人,该大赛吸引了来自全球 186 个国家的专家、学者组成的上万支队伍参加。Youtube 作为目前世界上最大的视频网站,允许用户自由地观看、上传、下载和分享各类视频,并且可以对视频进行评价来不断改善视频的推荐质量,Youtube 的个性化推荐算法是在 2008 年上线运行,并在 2013 年获得了美国国家电视艺术与科学学会授予的“技术与工程艾美奖”,因为它可以从海量的视频中发现用户的爱好,提供深度的个性化体验,延长了用户的注意力。当然,在国内也有类似的视频网站,例如优酷、土豆和酷 6 等也都引入了推荐技术为用户提供视频推荐服务。

亚马逊(Amazon)作为最早采用推荐系统的电商网站,同时也是推荐系统运用到实际应用中非常成功的典范,对推动推荐系统的研究与发展起到了积极作用。据统计,35%的商品销售额由亚马逊的推荐系统提供[15]。这引起了学术界和工业界的极大重视,也推动了研究人员对推荐系统的研究激情。与亚马逊类似的京东、当当、淘宝、eBay 等也都引入了推荐技术,为用户推荐喜欢的商品。

同时,学术会议的召开也促进了推荐系统的研究与发展。近年来,关

于研究个性化推荐系统以及相关技术的国际会议特别多，例如 KDD、ICML、AAAI、IJCAI 和 PKDD 等。在国内，特别是中国计算机学会（China Computer Federation，CCF），每年都会组织一次关于推荐系统的学术会议和若干期关于推荐系统与数据挖掘相结合的学科前沿讲习班。

推荐系统的发展，其本质上是推荐算法不断发展和演化的结果，因为它是推荐系统的核心部分，它的性能决定了推荐系统的性能，也决定了推荐系统的推荐策略和工作方式，因此对推荐系统的研究实际上是对推荐算法的研究。根据推荐算法的工作机制，可以将推荐系统分为[16]：基于内容的推荐（Content-Based Filtering，CBF）、基于协同过滤的推荐（Collaborative Filtering，CF）（包括：基于项目和基于用户的方法）、混合推荐（Hybrid Recommendation，HR）以及其他推荐方法等。

（1）基于内容推荐的研究现状分析

基于内容的推荐算法是对信息过滤技术的发展和延伸，其核心是提取和分析项目内容的特征信息，为用户推荐那些与用户历史感兴趣且相似度高的项目。

在不同的应用场景中，基于内容的推荐算法得到了广泛运用，它最早被用在网页推荐和邮件过滤等方面。在网页过滤方面，比较著名的是 Fab 系统[17]，该系统通过从网页中提取权重最大的 128 个词组作为网页的特征词来描述网页，并对网页进行分析，从而实现网页推荐；针对网页推荐，斯坦福大学的 Balabanovic 等[18]开发了智能代理 LIRA，它采用基于内容的搜索规则来搜索互联中的网页，并向用户推荐符合规则的网页，然后用户评价推荐的网页，并将评价结论作为有价值的信息反馈给系统，根据反馈结果更新搜索规则，这样可以为用户推荐更符合要求的个性化搜索内容；针对网页浏览，麻省理工学院的 Lieberman[19]开发了辅助推荐智能代理系统 Letizia，该系统通过对用户的浏览行为进行隐性跟踪，主动学习用户的兴趣模型，并根据学习得到的兴趣模型在后台搜索网页，最后将那些符合兴趣模型的网页推荐给用户；加州大学的 Pazzani 等[20]为实现多元化的推荐形式，利用用户对已浏览网页的评分信息实现了基于内容

的推荐系统 Syskill & Webert，该系统利用贝叶斯分类器训练用户的兴趣模型。在邮件过滤方面，麻省理工学院的 Malone 等[21]实现了 Information Lens 系统，该系统采用基于内容的半结构化模块，实现了电子邮件的简单过滤；黄志刚[22]建立了基于贝叶斯分类模型的中文垃圾邮件过滤系统，该系统采用了改进的断句算法和数据挖掘算法，可以很好地发现邮件中的不良信息；刘伍颖等[23]提出了历史域分类器效力线性组合权和当前域文档分类能力线性组合权，对网页进行过滤。此外，还有基于本体的推荐，该方法可以利用本体提供的知识网络代替采用关键字进行资源过滤的方式，梁俊杰等[24]根据网页中的标注信息和对应的本体概念实现网页的分类，并通过用户兴趣模型与网页类别的匹配为用户推荐网页。其他方面的应用也有很多，例如：辛菊琴等[25]利用本体语义推理机制实现资源聚类，在推荐过程中通过实时分析用户浏览行为捕获用户个性化偏好的变化，动态实时推荐内容，实验结果表明动态更新推荐列表，更加贴近用户的真实需求；田超等[26]借助多属性决策手段，提出了智能网上商城推荐系统 SuperRank 框架，能很好地符合用户的评论偏好，是一种有效的方法。

(2)基于协同过滤推荐的研究现状分析

基于内容的推荐其关键是分析项目的内容信息，并提炼能够描述项目内容的特征词，从而挖掘文本内容的实质。但在互联网中，除了易分析的文本资源外，还有大量的多媒体资源(如视频和音频等)都难以分析，导致了基于内容的推荐无法完成。而基于协同过滤的推荐则是通过用户关于项目的评分来计算相似度，从而为用户提供推荐。该方法不受对象内容的限制，应用领域更为广泛，且具有简单、通用和可以很好地发现用户新兴趣的特点，因此该方法自产生以来就得到了快速的发展。但随着电子商务的发展，协同过滤系统的缺点也逐渐凸显，例如稀疏性、冷启动、相似度计算等。

降维是解决数据稀疏性问题的一种重要方法，该方法通过将高维的用户评分空间映射到隐式的低维语义空间，降低推荐算法对数据稀疏性

的敏感度。在该类方法中矩阵的奇异值分解与主成分分析是两种具有代表性的方法。Billsus 等[27]提出的奇异值分解方法将那些对相似度计算影响不大的用户或项目评分直接移除来提高评分矩阵的密度;著名的隐语义索引模型(Latent Semantic Indexing,LSI)[28]其本质也是利用奇异值分解方法来对用户向量进行降维,然后计算用户间的相似度。Glodberg 等[29]将推荐过程分为两个步骤,在离线阶段用主成分分析对评分矩阵进行降维,在线阶段为用户提供推荐;Kim 等[30]提出了迭代的主成分分析来实现矩阵的降维处理。虽然该方法在某种程度上能够缓解数据稀疏性问题,但是它舍弃了部分的用户评分或用户,不可避免地要损失一些有价值的信息。

缺失值填充也是缓解数据稀疏性问题的一种重要方法,该方法采用有效的预测方法对缺失值进行预测并填充,来提高数据的密度。最简单也是最方便的方法是采用平均评分值、评分中值、众数等来填充[4];张锋等[31]确定候选邻居集的方法是依据用户评分矩阵交集的大小,并且采用 BP 神经网络对未评分的项目进行评分预测,减小了候选邻居评分矩阵的稀疏程度;Ma 等[32]对基于项目和基于用户的方法进行结合,来预测评分矩阵中的缺失值,提高了矩阵的密度和推荐精度;Sun 等[33]采用多种方法,例如贝叶斯分类预测、均值填补、预测均值匹配、线性回归预测等对评分矩阵填充,并对比分析了各种方法的准确性。

为解决传统相似度计算方法对矩阵稀疏性比较敏感的问题,学者们提出了一些新的相似度计算方法。周军锋等[34]提出了一种优化的协同过滤算法,该方法采用修正的条件概率方法计算项目间的相似度,得到了更准确的结果;张光卫等[35]提出了一种新的基于云模型的相似度计算方法,并且在数据极其稀疏的情况下可以得到理想的推荐效果;Luo 等[36]提出了分别计算用户的局部相似度和全局相似度,并根据相似度的大小选择各自的近邻,然后对两种最近邻的预测评分进行计算,并采用权值来平衡两种预测的重要性;Choi 等[37]考虑了目标项目与所有项目的相似度,凡是与目标项目相似度越高的项目,在最近邻搜寻中起到的作用