基于半监督学习的个性化推荐算法研究

RESEARCH ON PERSONALIZED RECOMMENDATION ALGORITHM BASED ON SEMI-SUPERVISED LEARNING

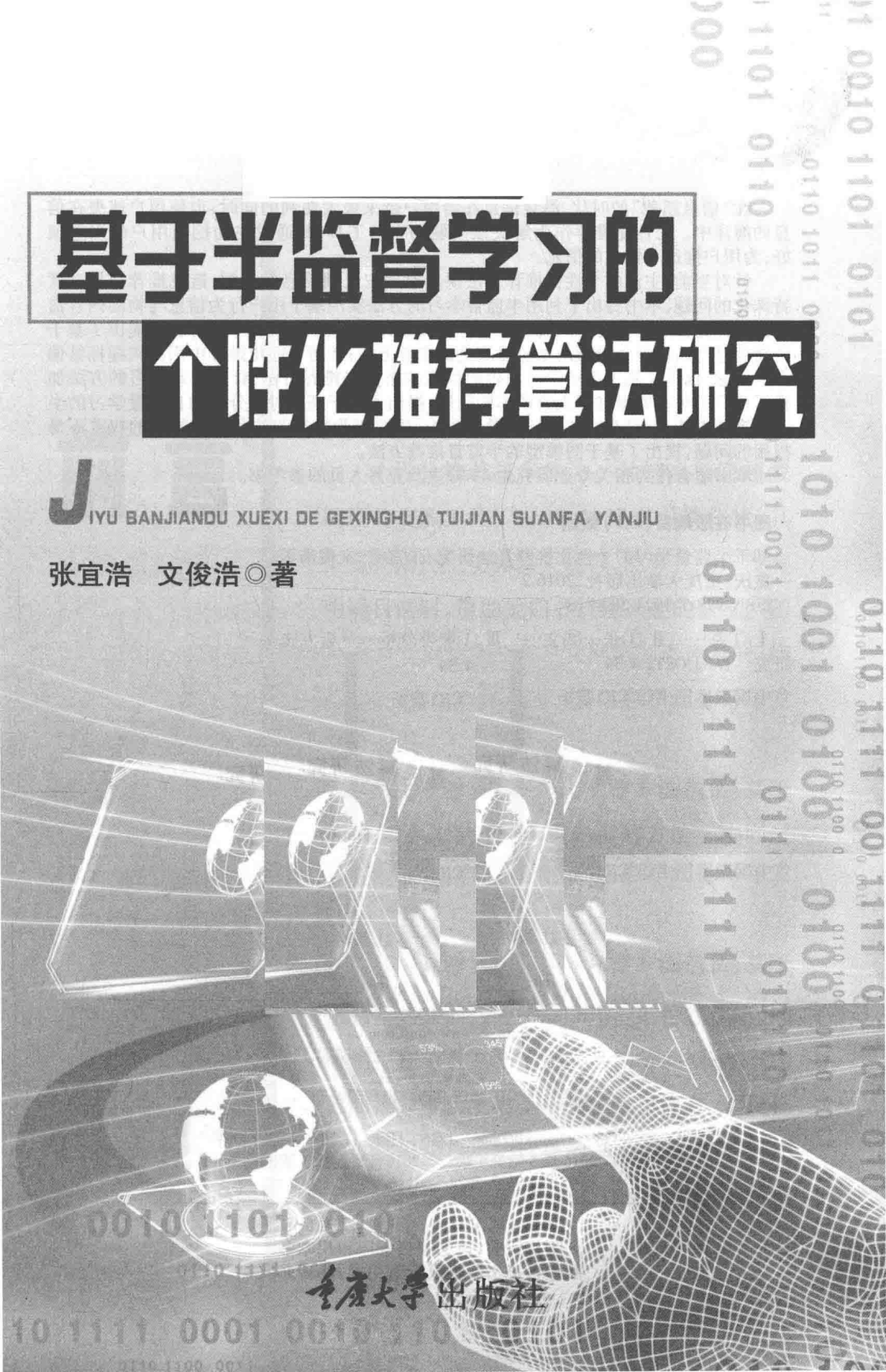
张宜浩　文俊浩◎著

重庆大学出版社
http://www.cqup.com.cn

基于半监督学习的个性化推荐算法研究

JIYU BANJIANDU XUEXI DE GEXINGHUA TUIJIAN SUANFA YANJIU

张宜浩　文俊浩◎著



重庆大学出版社

内容提要

在“信息超载”的时代，海量信息在给用户带来极大便利的同时，也使用户迷失在信息的海洋中。个性化推荐作为解决该问题的有效工具，其通过主动挖掘用户的兴趣偏好，为用户推送个性化的信息。

针对当前，主流的个性化推荐方法缺乏对用户反馈信息的挖掘，造成推荐结果过度特殊化的问题，本书提出了利用半监督学习的方法实现基于用户行为信息与物品内容信息的个性化推荐。针对协同过滤推荐方法存在计算相似度方式单一等问题，提出了基于距离度量与高斯混合模型的半监督聚类的推荐方法；针对个性化推荐中用户兴趣标签偏少的问题，提出了基于主动学习和协同训练的半监督推荐方法，针对主动学习的方法加重了用户的负担或增加了人力成本的问题，提出了基于高斯对称分布的自增量学习的半监督推荐方法；针对在构建特征向量过程中，用户行为特征与物品内容特征的权重不易权衡的问题，提出了基于图模型的半监督推荐方法。

本书适合作为相关专业研究生、本科生及业界人员的参考书。

图书在版编目(CIP)数据

基于半监督学习的个性化推荐算法研究/张宜浩，文俊浩著.
—重庆：重庆大学出版社，2016.2
ISBN 978-7-5624-9681-6

Ⅰ.①基… Ⅱ.①张…②文… Ⅲ.①聚类分析—分析方法—研究 Ⅳ.①O212.4-34

中国版本图书馆 CIP 数据核字(2016)第 033485 号

基于半监督学习的个性化推荐算法研究

张宜浩 文俊浩 著
策划编辑：彭 宁 何 梅
责任编辑：陈 力 版式设计：彭 宁 何 梅
责任校对：邬小梅 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
POD：重庆书源排校有限公司

*

开本：787mm×960mm 1/16 印张：11.25 字数：140 千
2016 年 5 月第 1 版 2016 年 5 月第 1 次印刷
ISBN 978-7-5624-9681-6 定价：39.00 元

前言

个性化推荐技术作为一种解决信息超载问题的有效工具，与传统的搜索引擎相比，其不需要用户主动提供关键词，能够在用户没有明确目的时，帮助他们寻找感兴趣的信息。随着近年来电子商务和社交网络的迅速发展，作为其核心组成部分的个性化推荐就显得越发重要了。

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，从而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。

半监督学习作为一种通用的机器学习方法在数据挖掘、自然语言处理等诸多领域都有着广泛的应用，并显示了其独特的优越性。针对传统的推荐方法在挖掘物品内容信息与用户标签信息上的不足，本书阐述了利用半监督学习的方法实现个性化推荐。

首先，针对协同过滤推荐方法存在计算相似度方式单一等问题，提出了基于距离度量与高斯混合模型的半监督聚类的推荐方法。传统的协同过滤方法时间复杂度和用户数的增长近似于平方关系，当用户数很大时，计算非常耗时。本书提出利用聚类分析的方法替代用户兴趣的相似度计算，且综合考虑了用户行为偏好和物品内容信息。具体在聚类分析中，算法不仅考虑了数据的几何特征，也兼顾了数据的正态分布信息。

其次，针对个性化推荐中用户兴趣标签偏少的问题，提出了基于主动学习和协同训练的半监督推荐方法。传统的基于分类模型的推荐方法，当有标签数据偏少时，对挖掘用户潜在的兴趣偏好非常不利，本书利用主动学习的策略抽取数据集中具有最大信息量的样本，通过咨询(Query)方式或领域专家标注的方式获得相应的标签，增加了训练模型的样本空间，以改进个性化推荐的质量。

然后，针对主动学习的方法加重了用户的负担或增加了人力成本的问题，提出了基于高斯对称分布的自增量学习的半监督推荐方法。该方法充分利用了大量无标签的数据，并结合一定的有标签数据进行建模。具体在算法中，通过挑选具有高置信度且高斯对称分布的数据进行自增量学习，以改进个性化推荐的质量。

最后，针对在构建特征向量过程中，用户行为特征与物品内容特征的权重不易权衡的问题，提出了基于图模型的半监督推荐方法。算法通过 SELF 等方法计算权衡因子，且根据用户的行为信息构造基于最近邻图的权重矩阵。算法利用 Sigmoid 映射函数来度量两个用户的兴趣相似度，并在算法的损失函数中包括用户行为相似性约束和物品内容相似性约束，且两部分约束的权重由一个平衡因子权衡。

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“大数据环境下基于用户行为分析的推荐结果个性化过滤研究”（NO.kj1500920），重庆市自然科学基金“基于半监督聚类的协同过滤电影推荐研究”（NO. cstc2014jcyjA1772）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 2 月

目录

第1章 绪 论

1.1 研究背景

随着互联网与信息技术的蓬勃发展，网络上的资源呈爆炸式增长。一方面，人们能从网络上获取越来越丰富的信息，给生活带来了极大的便利；另一方面，在海量的信息空间带给用户更多元化选择的同时，反而使用户迷失在信息的海洋中，极大地增加了用户搜索自己感兴趣信息的难度和成本。人们逐渐地从信息匮乏的时代步入了信息超载（information overload）的时代[1]。特别是在即将步入的“大数据”时代，无论是对信息生产者还是对信息消费者都提出了极大的挑战[2]：对于信息生产者，让自己生产的信息脱颖而出，受到大量用户的关注，而不至于安静地躺在网络的旮角不为人所知，是一件十分困难的事情；对于信息消费者，从大量

的信息中获取自己感兴趣的信息也不是一件容易的事情[3]。推荐系统正是解决这一突出问题的有力工具:其一方面帮助用户搜索对自己有价值的信息,另一方面让信息呈现于对它感兴趣的用户面前。从根本上说,推荐问题就是代替用户评估它从未看过的物品或信息[4]。

为解决信息过载问题,已有无数科学家与工程师提出了众多的解决方案。从信息检索的方式来看,这些有代表性的解决方案大致分为3个阶段:门户网站、搜索引擎、推荐系统[5]。

①门户网站。著名的互联网公司 Yahoo 凭借分类目录起家,其将著名的网站分门别类,从而方便用户查找。但是随着网页数量的增长与互联网规模的不断膨胀,分类目录网站也只能覆盖少量的热门网站,越来越不能满足用户的需求。

②搜索引擎。随着信息量的不断增长,分类目录帮助人们搜索信息的局限性越来越明显,因此搜索引擎诞生了。并随着网络上信息的大量涌现,搜索引擎行业也不断地发展壮大,其中最具代表性的莫过于 Google。搜索引擎可依据用户输入的关键词,快速地返回给用户与关键词相关的信息。但搜索引擎需要用户主动提供准确的关键词来搜寻信息,因此它不能解决用户的很多其他需求。当用户不能提供准确描述自己的需求时,搜索引擎也就无能为力了;对于搜索用户而言,搜索引擎也不能考虑他们之间的需求差异,只要用户输入的是相同的关键词,最终获得的网页信息及排序也将是相同的。

③推荐系统。与搜索引擎一样,推荐系统也是一种帮助用户快速搜寻有用信息的工具。不同的是,推荐系统不需要用户提供明确的需求,它是通过分析用户的历史行为记录对用户的兴趣建模,从而主动给用户推荐其可能感兴趣的信息和需求。从某种意义上说,搜索引擎满足了用户有明确目的的主动查找需求,而推荐系统能够在用户没有明确目的时帮

助他们寻找感兴趣的信息，其根据搜索用户的特点为不同的用户返回不同的搜索结果。这一观念在现在的搜索引擎中也有所体现，如 Google 允许用户定制自己网页的重要性，百度董事长兼首席执行官李彦宏在“百度技术创新大会”上也提出了智能框计算技术，这些都可称之为个性化网页搜索[6,7]。

个性化推荐技术作为一种解决信息超载问题的最有效工具，与传统的搜索引擎相比，它不需要用户主动提供关键词，能够在用户没有明确目的时，帮助他们寻找感兴趣的信息。随着近年来电子商务的迅猛发展，作为其核心组成部分的个性化物品推荐越发显得重要了。2006 年 Netflix Prize 电影推荐竞赛、2012 年 KDD Cup 腾讯微博用户推荐大赛、2013 年百度电影推荐系统算法创新大赛等，一系列推荐大赛将推荐系统的研究推向了一个前所未有的高度。

从信息服务的角度出发，个性化推荐通过分析用户的习惯、偏好等行为，能够及时跟踪用户的需求变化，进而主动调整信息服务的内容与方式，并定制地向用户推荐其感兴趣的信息和服务。与传统搜索引擎提供的“一对多”式的信息服务方式不同，个性化推荐系统反馈的结果更符合用户需求，同时用户的参与度也更低，从而极大地降低了用户搜寻信息的成本与难度。个性化推荐作为一种崭新的智能信息服务方式，能有效地解决“信息超载”带来的一系列问题，其已成为当前各主流网站不可或缺的新一代信息服务形式。

在传统门户网站时代，与流量伴生的数据价值长期被低估。进入大数据时代后，如何从这些海量数据中挖掘、分辨出用户的行为模式、兴趣偏好等变得特别重要。比如，用户对资讯的偏好不仅和兴趣相关，也和所处的阅读场景、资讯的关联性等其他方面相关。通过对这些方面日常数据的累积和挖掘，就可以很准确地向用户推荐最适合的内容，打造专属于

每个人的智慧门户。

从物品的角度出发，推荐系统可以更好地发掘物品的长尾（long tail）。根据长尾理论，传统的80/20（80%的销售额来自于20%的热门物品）原则在互联网电子商务中不断受到挑战。由于网络货架成本低廉，与传统零售业相比，网络中出售的不热门物品数量极其庞大，也许会超过热门物品带来的销售额。热门物品往往代表绝大多数用户的需求，而长尾物品则代表了小部分用户的个性化需求。发掘物品的长尾，需要通过分析用户的行为来分析用户的个性化需求，从而将长尾物品准确地推荐给对其感兴趣的用户，这正是个性化推荐系统主要解决的问题。

同搜索引擎相比，个性化推荐系统需要依赖用户的行为数据，因此在目前其一般都是作为一个应用存在于各大网站的后台。个性化推荐系统在这些网站中的主要作用就是通过分析大量用户的行为偏好，然后向不同用户展示不同的个性化页面。就现阶段而言，推荐系统主要利用领域包括：电子商务、电影和视频、个性化音乐网络电台、社交网络、个性化阅读、基于位置的服务、个性化邮件、个性化广告等。

1.2 国内外研究现状

个性化推荐系统的研究最初源于其他领域的工作，其雏形可追溯于1979年在认识科学领域中Elaine Rich提出的Grundy系统[8]，其利用stereotypes机制建立用户模型，并通过模型向用户推荐相关书籍。然而直到20世纪90年代，个性化推荐系统的研究才作为一个独立的概念被提出来[9,10]，此后推荐系统的研究和应用得到了飞速发展。特别是在20世纪90年代中期，出现了一大批基于协同过滤算法的推荐系统研

究[9-11]，推荐系统也逐渐成为一个独立的研究领域，得到了国际学术界的广泛关注。

从1999年开始，ACM每年都举行一次电子商务的研讨会（ACM-EC），在研讨会上，一个重要议题就是电子商务的个性化推荐，而且随着互联网技术的发展，关于个性化推荐的研究成果逐年增加。与此同时，ACM领导下的数据挖掘特别兴趣组（SIGKDD组）设立了WEBKDD讨论组，专门研究电子商务中Web挖掘和推荐的相关技术。另外1999年人机界面会议（SIGCHI-99）也专门设立了推荐系统特别兴趣组。在2001年召开的研究和发展会议上，ACM下信息检索特别兴趣组（SIGIR组）专门将推荐系统作为一个研讨主题。进入21世纪，特别是最近几年，在人工智能、数据挖掘、机器学习等领域的顶级国际会议中（如AAAI，KDD，SIGIR，ICML等），都将推荐系统以及相关推荐算法研究作为会议的一个重要议题。

特别是2006年9月，ACM和SIGIR在西班牙组织召开了“推荐系统的现在与未来”暑期班。该研讨班吸引了来自世界各地研究推荐系统的机构与科研人员，针对推荐系统的技术方法、应用领域、发展前景等方面，进行了深入详细的探讨与交流。鉴于此次学术研讨班取得的良好效果，2007年10月，ACM在美国的明尼苏达召开第一届推荐系统国际会议（ACM International Conference on Recommender Systems 2007，RecSys 2007），为广大的研究机构与科研人员提供了一个专业的学术交流平台。至此，每年都会在世界各个地方举办一次推荐系统国际会议。特别值得一提的是，2013年10月16日，第七届推荐系统国际会议（RecSys 2013）在中国香港举行，这次会议由华为在香港的诺亚方舟实验室协办，会议主席为诺亚方舟实验室主任杨强教授，是首次在亚洲举行的世界顶级推荐系统国际会议，具有特别重大的意义。

伴随在推荐系统的研究过程中，推荐算法是研究的核心，它是推荐系统的最重要组成部分，决定了整个推荐系统的工作方式与推荐策略[12]。依据推荐算法的策略不同，推荐系统一般可分为：基于内容的推荐（Content-Based Filtering，CBF）、协同过滤推荐（Collaborative Filtering，CF）、混合推荐（Hybrid Recommendation，HR）以及其他推荐方法[4]。推荐系统的分类如图 1.1 所示。

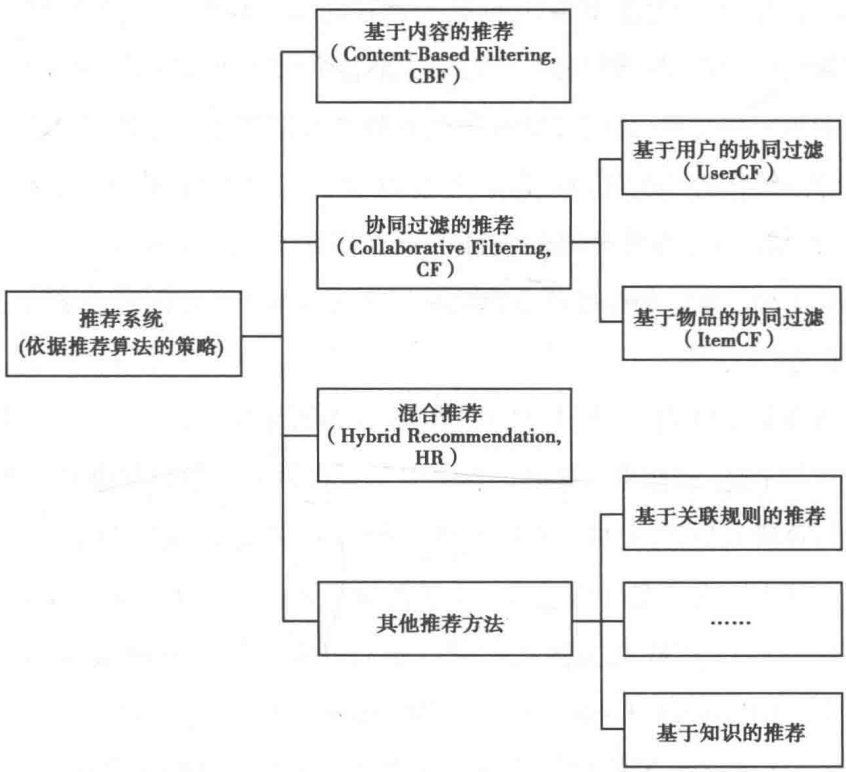


图 1.1　推荐系统的分类

1.2.1　基于内容的推荐

基于内容的推荐系统源于信息检索领域，利用了很多信息检索中的相关理论、方法以及技术。其一般流程是：首先分析推荐物品的内容信

息,抽取出推荐物品的特征描述;然后根据用户感兴趣物品的内容信息进行用户建模,从而形成基于内容的用户兴趣特征描述;最后通过计算用户未访问物品的特征与用户兴趣描述间的相似性,选择相似度最大的物品进行推荐。

基于内容的推荐方法在互联网推荐中得到了大量的应用。麻省理工学院的 Malone 等人[13]开发了电子邮件过滤的系统(Information Lens),采用了基于内容的半结构化模块,实现了对邮件信息的过滤。斯坦福大学的 Balabanovic 等人[14]构建了针对网页推荐的智能代理,该系统利用内容的搜索规则对互联网进行搜索,并将搜索结果页面推荐给用户;当用户对推荐的网页进行评价后,系统也会根据用户的评价反馈,对搜索规则进行更新,以完善后续的推荐结果,实现了较传统搜索引擎更为个性化的搜索内容。加州大学的 Pazzani 等人[15]利用用户对已浏览网页的评分信息,建立了 Syskill & Webert 推荐系统,它利用贝叶斯分类器构建用户的兴趣模型,实现多样化的推荐。卡内基梅隆大学的 Joachims 等人[16]开发了网页浏览路径推荐代理系统(Web Watcher),该系统通过对用户浏览网页的超链接进行分析,并结合 Agent 的历史推荐浏览路径,对用户的浏览行为进行学习建立模型。卡内基梅隆大学的 Zhang 等人[17]提出了利用自适应过滤技术更新用户配置文件,其主要思想是利用用户的喜好信息构建配置文件并将用户兴趣归纳为几个主题,然后计算未知 Web 文件内容与主题文件的相似度,进而选择相似度较高的 Web 文件实现推荐。Degemmis 等人[18]利用 WordNet 构建基于语义学用户的配置文件,而配置文件是通过机器学习算法得到的,结果表明这种方法可以提高推荐的准确性。田超等[19]利用自然语言处理技术对用户评论进行情感分析,构建推荐系统的 SuperRank 框架。Chang 等[20]通过赋予短期感兴趣的关键词更高的权重,建立新的关键词更新树,大大减少了更新配置文件的代