文章编号:1000-2367(2022)03-0095-09 DOI:10.16366/j.cnki.1000-2367.2022.03.012

基于用户画像的高校图书馆个性化图书推荐研究

王大阜[a],邓志文[a],贾志勇[a],安计勇[b]

(中国矿业大学 a.图书馆;b.计算机科学与技术学院,江苏 徐州 221116)

摘 要:个性化推荐服务是高校智慧图书馆的建设重点,基于此,提出了图书推荐系统整体架构.首先从读者的属性、行为、兴趣等标签维度构建用户画像模型,其次考虑读者认知能力存在差异化的特点,将读者按照不同的身份类型划分,再结合基于协同过滤、内容及属性相似度的混合推荐算法进行图书推荐.最后,通过 Hadoop 大数据平台向目标读者推荐 TOP-N 图书,实验结果表明,基于该架构模型的图书推荐系统的推荐准确度高,并且有效缓解了推荐系统的冷启动问题.

关键词:推荐系统;智慧图书馆;用户画像;冷启动

中图分类号:TP311.5 **文献标志码**:A

随着高校图书馆的馆藏纸本文献、期刊数据库、特色数据库等多样、异构资源的持续建设发展,资源的规模、体量呈现爆炸性增长态势,丰富的资源为读者学习、科研、生活提供了极大辅助作用的同时也暴露出“信息过载”的问题,读者从海量资源中找到与其兴趣匹配的、高质量的资源变得十分困难.以 OPAC 系统为例,读者通过书名、主题、关键词等条件进行检索,并从中选择感兴趣的图书.这种主动式服务的前提条件是读者有明确的检索需求,然而更多时候,读者并没有明确的需求,或是缺乏良好的检索技能,从而更倾向于被动式的个性化服务,希望系统“智慧地”向读者推荐、呈现有可能感兴趣的优质资源,如此,图书馆的阅读推广工作更见成效,同时提升了读者的服务体验.以读者为中心,为读者提供智慧服务,是智慧图书馆建设的根本宗旨,个性化推荐系统作为典型的智慧服务应用之一,已成为图书馆领域近年来的研究热点,然而真正实施技术研发并成功落地的案例并不多,多数还是采用传统的“热门图书”或 “阅读清单”这种无差异化的、宽泛的阅读推荐模式[1].事实上,推荐系统的理念、算法用于精准营销、个性推荐、广告投放等场景,在电子商务(Amazon、京东、淘宝等)、影音网站(爱奇艺、网易云音乐)、社交网站(微博、豆瓣、今日头条等)领域均有广泛的应用.

1 相关研究

1.1 用户画像

“用户画像”(Persona)最早由交互之父 Alan Cooper 于 1998 年提出,他表示用户画像是“基于用户真实数据的虚拟代表”[2].用户画像是在用户真实数据的基础上,用来勾勒用户特征,描述用户兴趣、需求的重要技术手段,能够全面细致地刻画用户的信息全貌,从而为向用户实施精准营销、个性化推荐服务奠定基础.换言之,用户画像的核心工作是为用户“打标签”.标签具有 3 个主要特征:(1)语义化,让人快速理解每个标签的含义;(2)短文本,每个标签通常只表示一种含义;(3)动态性,用户的兴趣偏好随时间推移、情境改变而变化,用户画像模型也随之需要动态修正和调整.

用户画像在图书馆学界已经受到广泛关注,汪强兵等[3]通过利用在移动端的用户手势行为数据与关键

收稿日期:2021-03-31;**修回日期**:2021-04-18.

基金项目:江苏省高校哲学社会科学研究项目(2020SJA1009)

作者简介(通信作者):王大阜(1981—),男,江苏盐城人,中国矿业大学图书馆馆员,研究方向为网络安全、云计算,E-mail:wdf@cumt.edu.cn.

词权重,挖掘用户阅读兴趣,由此构建用户兴趣画像.韩梅花等[4]根据抑郁情感词典分析微博文本,计算抑郁情感指数,获取用户画像,进而推送阅读治疗资源.王顺箐[5]以读者需求为核心,在数据采集基础上构建用户画像,构建智慧阅读推荐系统,提高阅读推广的成功率.胡媛等[6]基于读者用户画像,构建图书馆知识发现服务模型,实现图书馆的个性化、精准化知识服务,提升读者服务体验.刘海鸥等[7]构建融合情境、内容偏好、互动、会话等多维标签的用户画像模型,并以此为基础提出情境化推荐方法,为读者精准推荐个性化知识服务.以上研究表明推荐系统引入用户画像从理论和技术角度来说都是可行的.

1.2 推荐系统

推荐系统的经典算法有两种:协同过滤(Collaborative Filtering,CF)算法和基于内容(Content Based,CB)的推荐算法,其中 CF 算法又分为基于用户的协同过滤(UserCF)算法、基于物品的协同过滤(ItemCF)算法,CF 算法原理是推荐与用户有相似兴趣的邻居用户喜欢的其他 Top-N 物品或是推荐与用户喜欢的物品相似的其他 Top-N 物品.CF 算法能够向用户推荐丰富的长尾物品,激发用户潜在的兴趣.与此同时,容易面临数据稀疏的问题,该问题对高校图书馆而言格外突出.高校图书馆的馆藏纸本书副本少,读者想借的书可能被他人借阅,造成不同读者之间借阅同一本书的共现数据稀疏,此外读者不太热衷于对图书的评分、评论,造成评分数据也同样稀疏.CB 算法原理是构造物品特征,推荐与用户喜欢的物品特征相似的其他 Top-N 物品,物品特征的表现方式可以是结构化的属性或非结构化的标签、关键词.CB 算法更适用于这种非结构化的新闻、文献等文本资源推荐,通过中文分词、TF/IDF 算法、LDA 模型等自然语言处理技术挖掘读者的兴趣关键词及权重,构造读者兴趣空间向量模型.推荐系统的冷启动问题在推荐领域中普遍存在,包括用户冷启动、物品冷启动两个层面,起因是新用户或新物品没有相关历史行为数据,造成无法为新用户推荐物品或将新物品推荐给用户.综上所述,任何一种推荐算法都有各自优缺点及适用场景,表 1 做了全面的归纳总结.

表 1 不同推荐算法的特点及适用场景

Tab. 1 Characteristics and applicable scenarios of different recommendation algorithms

算法	优点	缺点	适用场景
UserCF	长尾物品丰富	可解释性弱;存在用户和物品冷启动问题	用户数远少于物品数,如新闻网站、图书推荐
ItemCF	长尾物品丰富	可解释性弱;存在用户和物品冷启动问题	用户数远大于物品数,如购物网站、影音网站
CB	可解释性强,能解决物品冷启动问题	属性特征不易提取;用户标注标签工作量巨大;存在用户冷启动问题	文本资源推荐,如新闻、博客、电子文献数据库,社会化标签网站,如豆瓣网、网易云音乐

图书馆学界关于推荐系统的研究有:常有学等[1]基于 Spark 大数据计算技术实现高效率、高准确度的图书推荐,提高用户体验.邓志文等[8]通过社交网提取用户候选兴趣标签,结合用户-物品、物品-标签关系模型,运用朴素贝叶斯算法为用户推送信息.尹婷婷等[9]在深度学习视角下提出了以读者用户兴趣值为基础的图书馆馆藏资源推荐模型,分别从数据关联、情景分析和协同过滤技术方面进行探讨,为资源精准推荐提供参考.王仲钰等[10]采用协同过滤算法、关联算法,从用户相似性和书籍关联性两个角度探索图书推荐服务策略.王连喜[11]通过挖掘用户的兴趣特征及隐含的需求模式,研究 UserCF,CB 和基于标签多种推荐方法,实现用户与图书相互关联的个性化图书推荐服务.以上学者均将推荐算法成功应用于图书推荐服务,但都是基于有评分数据的公开数据集或用户打的标签数据进行建模,没有考虑图书馆评分、标签稀疏的现实情况.李澎林等[12]提出基于读者兴趣度与类型因子算法,建立读者兴趣度模型,解决了评分及借阅关系稀疏的问题,有很大的借鉴意义,但是选取的读者兴趣特征粒度较粗,且没有考虑读者下载行为因素.笔者结合中国矿业大学实际情况,当前在校师生人数(约 7 万)远少于图书种类数(约 67 万),如果采用 ItemCF 算法,$(67\text{ 万})^2$ 的物品高维矩阵在内存空间、计算复杂度上过高,因此更适合用 UserCF 算法.此外,为解决推荐系统中的冷启动问题,笔者综合采用了 CB 算法、属性相似度算法.

2 用户画像构建

2.1 标签分类

标签按照产生和计算方式不同可分为属性标签、统计标签、算法标签 3 种类型,属性标签是对实体基本性质的刻画,如性别、年龄、专业;统计标签是特定场景下,维度和度量的组合,如某个读者月均借阅、下载图

书的频次；算法标签是不能直接获取的，需要通过数据挖掘或计算推理得到，如读者对不同图书类别的偏好程度或感兴趣的主题.

2.2　标签维度

本文从用户、图书两个层面建立标签，并构建画像，数据来源于图书馆 OPAC 系统以及豆瓣读书的评分数据.用户标签从用户属性标签、用户行为、用户兴趣 3 个维度构建.1)用户属性标签，是对用户属性的描述.包括性别、年龄、专业、身份类型(本科生、硕士生、博士生、教师)、是否为新读者等基本属性.2)用户行为标签，是对用户活跃程度的描述.包括读者学年内月均纸质书借阅频次、电子书下载频次、续借频次及活跃度，其中活跃度是对前 3 种频次数值 Sum 求和，根据阈值判定结果，如当 Sum≥10 时为高，当 5≤Sum<10 时为中，当 Sum<5 时为低.3)用户兴趣标签，是对用户图书类别偏好的描述.豆瓣网站在用户注册账号时，会让用户选择关于电影、书籍的风格喜好，以此作为用户的兴趣标签.图书馆 OPAC 系统没有类似功能，可以从用户学年内纸本书借阅、电子书下载的历史行为数据中分析挖掘.图书标签从图书属性、图书类型两个维度构建.1)属性标签.包括图书 ID、书名、ISBN 号、作者、是否为新书(近 3 个月上架图书)等.2)图书类型标签，对应图书的中图法二级分类号.用户、图书标签体系见表 2 和表 3，图 1 是某读者的用户画像示例.

表 2　用户标签维度表

Tab. 2　User label dimensions table

标签名称	标签主题	标签说明	标签类型	标签名称	标签主题	标签说明	标签类型
UserID	用户属性	读者 ID	属性	College	用户属性	学院/单位	属性
Name	用户属性	姓名	属性	NewReader	用户属性	是否为新读者	统计
Sex	用户属性	性别	属性	Render	用户行为	月均借阅次数	统计
Age	用户属性	年龄	属性	Download	用户行为	月均下载次数	统计
Identity	用户属性	身份类型	属性	Renewal	用户行为	月均续借次数	统计
Grade	用户属性	年级	属性	Activity	用户行为	活跃度	算法
Level	用户属性	职称	属性	Interest	用户兴趣	图书类别偏好	算法

表 3　图书标签维度表

Tab. 3　Book label dimension table

标签名	标签主题	标签说明	标签类型	标签名	标签主题	标签说明	标签类型
BookID	图书属性	图书 ID	属性	Score	图书属性	豆瓣评分	属性
Name	图书属性	名称	属性	NewBook	图书属性	是否为新书	统计
ISBN	图书属性	ISBN 号	属性	Type	图书类型	中图法分类号	属性
Author	图书属性	作者	属性				

3　混合推荐算法

3.1　划分读者类型

不同身份类型的读者由于学历背景、认知能力有所区别，他们感兴趣的书籍是有差异性的，如本科生倾向与课程相关的参考工具书，研究生倾向与某个研究主题相关的学术性书籍，教师倾向与有关学科前沿的书籍.基于此，本文首先将读者按以上 3 种身份类型进行划分，其次再结合混合推荐算法进行图书推荐.

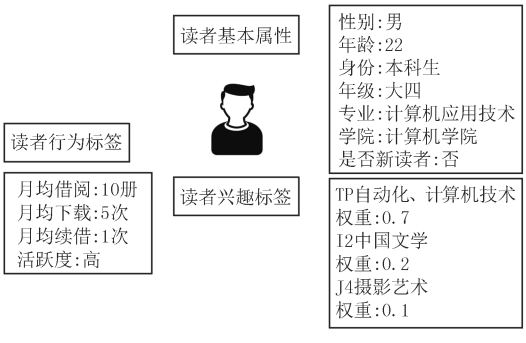

图1　某读者的用户画像

Fig. 1　User portrait of a reader

3.2 UserCF 算法

3.2.1 特征构造与 K 近邻搜寻

《中图法》目前出版第五版,中图书分类号是一种树状结构,其中大类 22 种,往下逐层扩展,分类号格式上由字母、数字、小数点等构成,如 TP391 属于第五级分类,对应分类名信息处理,TP391.1 属于 TP391 的子类,对应分类名文字信息处理.为解决读者之间借阅共现数据稀疏的问题,本文使用读者学年内对不同分类纸质图书的借阅(含续借)频次及电子书的下载频次之和作为读者的兴趣向量特征,分类层级选取的粒度太粗体现不出读者喜好图书的类别,太细导致维度过大,计算耗时长,权衡考虑,本文选取粒度为二级层级,总计 222 种小类.

假定读者的兴趣特征向量 $U=(u_1,u_2,\cdots,u_n)$,首先利用离差标准化(Min-Max)方法对特征做归一化处理,接着采用余弦相似公式计算读者之间的相似度,

$$\mathrm{sim}(u,v)=\cos(u,v)=\frac{\sum_{i=1}^{n}u_i\times v_i}{\sqrt{\sum_{i=1}^{n}(u_i)^2}\times\sqrt{\sum_{i=1}^{n}(v_i)^2}}.$$

不同读者间相似度矩阵如表 4 所示.最后从中可以搜寻与读者相似度最大的 K 个近邻读者集合,用 U_k 表示,

表 4 不同读者间相似度矩阵(3 个读者为例)

Tab. 4 Similarity matrix of different readers(three readers as an example)

	U_1	U_2	U_3		U_1	U_2	U_3		U_1	U_2	U_3
U_1	1	$\cos(U_1,U_2)$	$\cos(U_1,U_3)$	U_2	$\cos(U_1,U_2)$	1	$\cos(U_2,U_3)$	U_3	$\cos(U_1,U_3)$	$\cos(U_2,U_2)$	1

3.2.2 兴趣度计算

UserCF 算法需要结合邻居用户对某物品的评分,预测用户对该物品的评分.该评分作为用户的显式反馈(点赞、喜欢/不喜欢或打分),反映了用户对某物品的兴趣度.用户-物品评分矩阵 $R=U\times I$,行向量表示某用户的评分集合,列向量表示某物品的被评分集合.然而,高校图书馆缺乏甚至没有读者对图书的评分数据,借鉴文献[12]的方法,使用借阅持续时间、续借次数,并增加电子书下载频次、豆瓣读书栏目评分作为兴趣度提取的因子,通过多个因子综合获取读者对图书的兴趣度.

3.2.2.1 借阅时长与续借

读者对某本图书借阅时间越长表明对该该图书越感兴趣,某本图书的借阅时长百分比为:

$$p=\frac{T_a(u,i)-T_b(u,i)}{T_c},$$

式中,$T_a(u,i)$ 表示读者 u 归还图书 i 的时间,$T_b(u,i)$ 表示读者借阅图书 i 的时间,T_c 为图书馆规定的超期有效期.此外,当读者对图书非常喜欢时,会对图书进行续借,限制续借 1 次,此时 $p=5$.最后将 p 映射成 1 ~ 5 兴趣度值,公式如下:

$$pref_1(u,i)=\begin{cases}1, 0\le p<0.25,\\ 2, 0.25\le p<0.5,\\ 3, 0.5\le p<0.75,\\ 4, 0.75\le p<1,\\ 5, \text{续借}.\end{cases}$$

3.2.2.2 电子书下载

当前,高校图书馆的资源建设部门对纸本图书的采购副本量逐渐减少,转化为以电子书为主导资源.以中国矿业大学图书馆为例,目前全馆馆藏书籍 60 余万种,230 万余册,基本实现全覆盖数字化.新采购的图书编目后不久,对纸本图书扫描加工成电子书,同时将电子书嵌入到 OPAC 系统中,方便读者下载阅读.由于读者试读电子书后,兴趣度很高才会产生下载行为,所以该因素分值设置相对较高.兴趣度公式如下:

$$pref_2(u,i)=\begin{cases}3, t=0,\\ 4, t=1,\\ 5, t\ge 2.\end{cases}$$

3.2.2.3　豆瓣评分

以上借阅时长、电子书下载存在不确定因素,譬如读者借阅了某本书,可能忙于学业,没时间还书,造成有效的借阅时间有偏差,电子书下载可能出现试读时满意,待整本书阅读后却不满意的情况.鉴于此,本文引入豆瓣评分,豆瓣网中读书栏目中关于图书的星级评分、评论是来自兴趣相投的网友们真实反馈,较为公正、准确,最终根据算法计算得出综合评分(1～10 分),本文将评分除以 2,作为图书的评分 $pref_3(u,i)$.豆瓣网每本图书的书目、评分信息通过编写 python 爬虫脚本采集.

最终读者 u 对图书 i 的综合平均兴趣度值 $pref(u,i)=[pref_1(u,i)+pref_2(u,i)+pref_3(u,i)]/3$.

接着使用 UserCF 算法计算用户 u 对图书 i 的兴趣度 $pref(u,i)$,公式如下:

$$pref(u,i)=\frac{\sum_{v\in u_k} pref(v,i)\times \mathrm{sim}(u,v)}{\sum_{v\in u_k}\mathrm{sim}(u,v)}.$$

3.2.2.4　时间衰减因子

读者的学习、研究兴趣具有时间效应,会随着时间上下文的推移而有所变化,如计算机专业的本科生借阅《机器学习》入门书,该学生进入研究生阶段时,可能会借阅《机器学习》进阶书.本文在计算两个用户相似度时,增加“时间衰减函数”.∞为时间衰减因子,t_{ui},t_{vi}分别为用户 u 和 v 借阅图书 i 的时间.

$$f(|t_{ui}-t_{vi}|)=\frac{1}{1+\infty|t_{ui}-t_{vi}|},$$

最终用户 u 对图书 i 的兴趣度为

$$pref(u,i)=\frac{\sum_{v\in u_k} pref(v,i)\times \mathrm{sim}(u,v)}{\sum_{v\in u_k}\mathrm{sim}(u,v)}\times\frac{1}{1+\infty|t_{ui}-t_{vi}|}.$$

3.3　冷启动问题

用户冷启动解决如何为新读者推荐合适图书的问题,新读者类型包括:新入学的学生、新入职的教师以及尚未借阅过图书的读者;物品冷启动解决如何为新书找到受众读者的问题,图书馆每年花费大量资金购置新书,致力于第一时间向读者推荐新书,从而提高图书资源利用率.

3.3.1　用户冷启动

用户冷启动可以根据用户性别、年龄、年级、职称、学院/单位自然属性计算用户与用户的相似度,将相似度高的用户借阅图书推荐给目标用户.具体方法是:首先提取所有读者相关属性,作为读者的向量特征,其中性别离散型属性使用 0、1 表示,年龄连续型属性采用 Min-Max 方法归一成[0,1]数值,年级、职称、学院等离散型属性可用 One-Hot 编码表示.接着找出与新读者相似度高的 K 个近邻老读者集合.当为新读者推荐旧书时,采用 UserCF 算法中计算兴趣度的公式,预测新读者对旧书的兴趣度,评选 TOP-N 旧书推荐给读者.当为新读者推荐新书时,获取为老读者推荐的新书集合,进行去重处理后向新读者进行 TOP-N 推荐.

3.3.2　物品冷启动

物品冷启动最简单的处理办法是将新书随机性展示,这显然不够个性化,展示的新书很大概率是读者不喜欢的.采用 CB 算法可以解决物品冷启动问题,具体方法是:首先为新书构造特征向量,并提取用户兴趣特征向量,然后计算两者之间的相似度,将相似度高的新书推荐给目标用户.假定读者的兴趣特征向量 $U=(u_1,u_2,\cdots,u_n)$,对应的兴趣权重特征向量 $U'=(u'_1,u'_2,\cdots,u'_n)$,特征 u'_i 指读者对某种图书类型数目占所有图书类型总数目的比重,值区间为[0,1],图书的特征向量 $I=(i_1,i_2,\cdots,i_n)$,特征为判断图书是否属于某种图书类型,0 表示否,1 表示是,并赋予读者兴趣权重,赋权后的图书特征向量 $I'=(i'_1,i'_2,\cdots,i'_n)=(i_1\times u'_1,i_2\times u'_2,\cdots,i_n\times u'_n)$. 接着对读者兴趣与图书的相似度采用余弦相似度公式计算,公式如下.最后对相似度进行排序,形成相似度较高的 TOP-N 新书推荐给读者.

$$\cos(U,I)=\frac{\sum U'_i\times I'_i}{\sqrt{\sum U'^2_i\times\sum I'_i}}$$

3.4 算法流程图

1)旧书推荐算法流程:

步骤 1 根据读者登录 OPAC 系统的 ID 号,即学工号,识别读者的身份类别;

步骤 2 判断读者是否是新读者,即是否曾经借阅、下载过图书,如果是老读者,按照步骤 3 至步骤 5 顺序执行,如果是新用户,按照步骤 6、7 顺序执行;

步骤 3 提取与读者身份类型相同群体的读者-图书类型兴趣向量,计算读者之间相似度,获取相似度较高的近邻老读者集合;

步骤 4 提取读者-图书兴趣度矩阵,根据近邻读者对所借阅书籍的隐式兴趣度,预测目标读者对这些图书的兴趣度,并进行兴趣度排序;

步骤 5 过滤预测兴趣度低及目标读者已借阅过的图书,并向目标读者推荐 TOP-N 图书列表;

步骤 6 提取新读者性别、年龄、年级(学生)、职称(教师)、学院特征属性,并与老读者进行相似度计算;

步骤 7 获取与新读者特征向量距离相近的近邻老读者集合,按照步骤 4、5 顺序执行.

2)新书推荐算法流程:

步骤 1 判断读者是否是新用户,如果是老读者,按照步骤 2 至步骤 3 顺序执行,如果是新读者,按照步骤 4 至步骤 6 顺序执行;

步骤 2 提取目标读者的兴趣特征向量,提取新书的特征向量,计算两者之间相似度,并进行相似度排序;

步骤 3 过滤相似度低且目标读者已借阅过的图书,并向目标读者推荐 TOP-N 图书列表;

步骤 4 提取新读者相关特征属性,并与老读者进行相似度计算;

步骤 5 获取与新读者特征向量距离相近的近邻老读者集合,获取老读者新书推荐集合;

步骤 6 对新书推荐集合进行去重处理,并向目标读者推荐 TOP-N 图书列表.

具体如图 2 所示.

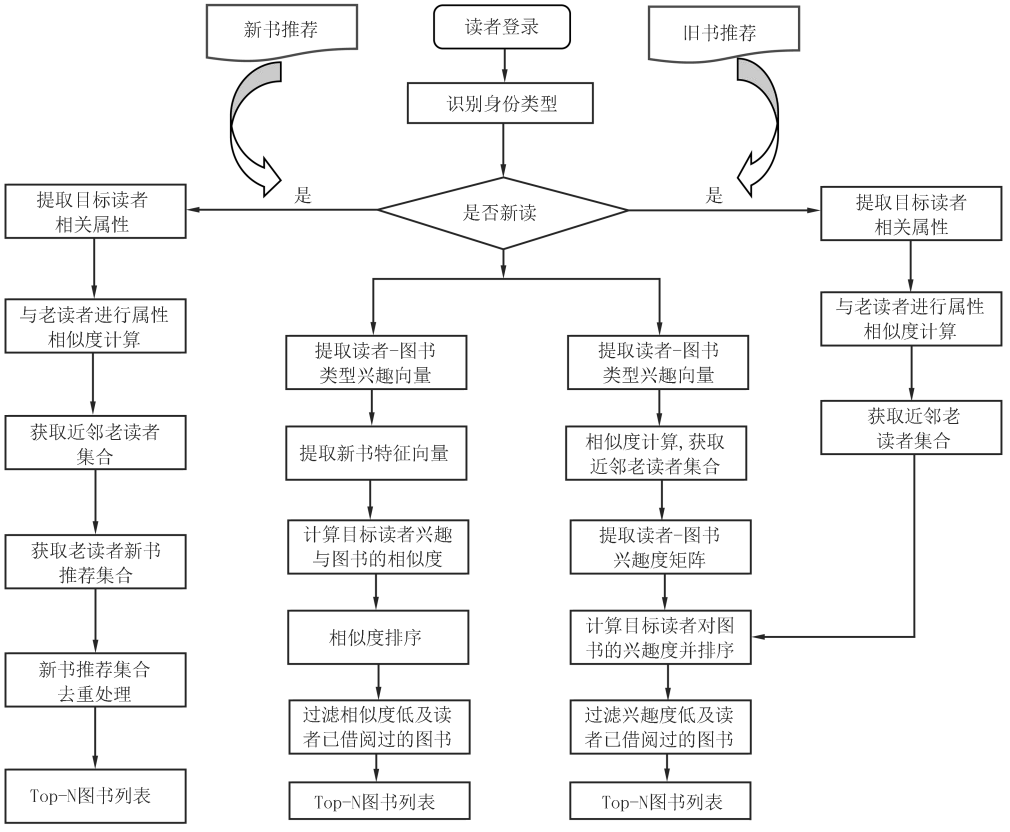

图2 混合推荐算法流程图

Fig.2 Flow chart of hybrid recommendation algorithm

4　推荐系统架构

推荐系统体系架构如图 3 所示,分为数据层、处理层、逻辑层、表现层.为避免推荐系统数据量庞大造成性能瓶颈,本文设计的推荐系统体系架构在 Hadoop 分布式环境下进行部署实现,Hadoop 版本选择 Cloudera 公司的集成化的发行版本 CDH5.Hadoop 是目前流行的针对大规模数据分析的开源分布式系统基础架构,由提供分布式文件存储(HDFS)和并行计算框架 MapReduce 组成,能够以高可靠、高性能、高扩展性的优势处理海量数据.

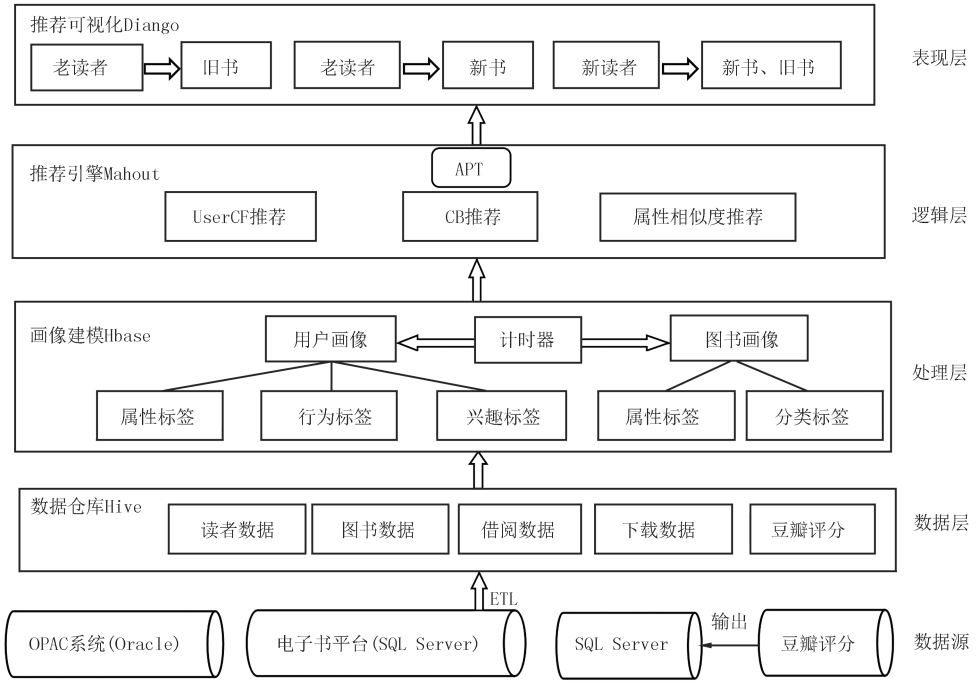

图3　推荐系统体系架构图

Fig.3　Recommendation system architecture diagram

1)数据层的元数据来源于 OPAC 系统(读者、图书、借阅数据)、电子书平台(下载数据)、豆瓣读书评分数据 3 个方面,其中豆瓣评分是通过爬虫技术采集并输出至 SQL Server 数据库.ETL(抽取、转换、装载)是构建数据仓库关键步骤,实现数据源到目标数据仓库的迁移,并在迁移过程数据完成了必要的数据清洗.具体做法是:通过 Sqoop 数据导入/导出工具将各种数据库数据导入到 Hive 数据仓库,再通过 HQL 语句实现数据清洗.数据清理主要从以下几个方面着手:①格式不规范的数据,如单词存在空格、数值数据中有字母或者输成全角数字字符、日期格式不正确,可以通过 HSQL 语句修正.②缺失值填充,如读者性别、专业、学院等信息缺失等,通过编写特定的语句从抽取的数据中过滤出这些数据,然后人工补全再写入数据库.③噪声数据,指源业务系统没有严格的数据校验造成的数据录入错误,比如日期越界、年龄巨高等.这些噪声数据可通过正态分布检测、基于模型检测方式检测出异常,并在数据源中进行修正再抽取[9].

2)处理层负责建立用户、图书画像模型,画像数据存储于面向列的、适合大数据实时查询的 HBase 分布式数据库.读者借阅、下载数据以及图书数据会动态更新,用户画像、图书画像也会相应产生变化,因此设定计时器,每隔 24 小时更新画像.

3)逻辑层是整个体系架构的核心,综合了 UserCF,CB 和属性相似度算法,形成多元化推荐引擎,实现为新读者与老读者分别推荐新图书与旧图书.推荐引擎基于 Mahout 实现,Mahout 是 Hadoop 生态圈的一个开源项目,提供分类、聚类、推荐引擎等机器学习算法,与 MapReduce 开发相比,非常简单便捷[13].

4)表现层调用逻辑层的 API,为读者提供可视化界面,向其展示推荐的新图书、旧图书.可视化推荐基于 Diango 框架实现,Diango 是基于 Python 语言开发的、采用 MVC 模式的 Web 应用框架,通过 Python 程序

调用 Mahout 的接口返回图书推荐列表,接着根据列表中的图书 ID,读取 Hbase 数据库中的图书信息,并向读者展示推荐结果.

5 实验论证

在本文的实验中,提取 2018－2019 学年师生读者的纸本书借阅数据以及电子书下载数据,其中读者 73 978人,图书 671 095 种,借阅数据量 268 145 条,下载数据量 38 212 条.在建立用户画像模型和推荐系统模型后,利用读者数据对模型进行训练,选取 30 名读者做 TOP-10 图书推荐,并对推荐结果做问卷评估,最后针对不同的 K 值,使用精确率(Precision,P)和平均绝对误差(MAE)评估推荐结果的准确度.精确率和平均绝对误差定义、计算公式如下:

精确率,表示正确预测(T_P)用户喜欢的图书在所有预测(T_P+F_P)用户喜欢的图书中所占比例,

$$P=\frac{T_P}{T_P+F_P}.$$

平均绝对误差,表示预测评分值和真实评分值之间的差值取绝对值再求和之后的平均值,

$$MAE=\frac{\sum_{i=1}^{m}|y_i-y'_i|}{m}.$$

实验结果如表 5 所示,结果表明,当 $K\leq 20$ 时,精确率逐渐提升,平均绝对误差逐渐降低;当 $K>20$ 时,精确率逐渐降低,平均绝对误差逐渐平稳.因此 $K=20$ 时,精确率和平均绝对误差是最优化的.

表 5 图书推荐准确度评价指标值

Tab. 5 Evaluation index values of book recommendation accuracy

K	5	10	15	20	25	30
精确率	67.2%	71%	73%	76.3%	74.7%	73.8%
平均绝对误差	1.39	1.1	0.92	0.8	0.78	0.77

6 结 语

用户画像作为大数据时代的产物,在电子商务领域已经成功应用于精准营销、广告投放,本文将用户画像应用于图书推荐服务,通过对读者的基本特征、行为、兴趣进行精准刻画,洞悉掌握读者的用户特征及需求.在此基础上,采用混合推荐算法,实现向读者进行个性化、精准化的图书推荐.未来将从以下 3 个方面进行探索、优化:1)高校图书馆拥有庞大丰富的资源,如论文数据库、学术视频库、特色资源库等,研究如何将各种资源融合,形成图书馆本地化的知识发现系统,并为读者推荐多样化的资源;2)高校学者作为推动学校学科发展的中坚力量,更希望获取与研究主题相匹配的、高品质的图书,而不追求推荐效果的多样化,可以将研究主题作为学者的兴趣特征,通过聚类算法,构建不同学科学者的群体用户画像[2],从而为其推荐适配的优质图书资源;3)针对高活跃度学者用户,通过学科馆员介入,对图书推荐资源进行人工干预提取,通过邮箱主动推送,巩固并维持用户活跃度;针对活跃度为中或低的学者用户,通过关联推荐算法拓展图书推荐资源,激发并提高用户活跃度.

参 考 文 献

[1] 常有学,刘建胜,刘旭波.基于 Spark 的高校图书馆书目推荐系统[J].现代电子技术,2019,42(14):64-67.
CHANG Y X,LIU J S,LIU X B.Spark-based bibliographic recommendation system for university libraries[J].Modern Electronics Technique,2019,42(14):64-67.

[2] 王庆,赵发珍.基于"用户画像"的图书馆资源推荐模式设计与分析[J].现代情报,2018,38(3):105-109.
WANG Q,ZHAO F Z.Design and analysis of library resource recommendation model based on user profile[J].Journal of Modern Information,2018,38(3):105-109.

[3] 汪强兵,章成志.融合内容与用户手势行为的用户画像构建系统设计与实现[J].数据分析与知识发现,2017,1(2):80-86.
WANG Q B,ZHANG C Z.Constructing users profiles with content and gesture behaviors[J].Data Analysis and Knowledge Discovery,

2017，1(2)：80-86.

[4] 韩梅花，赵景秀.基于“用户画像”的阅读疗法模式研究：以抑郁症为例[J].大学图书馆学报，2017，35(6)：105-110.
HAN M H，ZHAO J X.Research on bibliotherapy model based on user profile-take depression as an example[J].Journal of Academic Libraries，2017，35(6)：105-110.

[5] 王顺箐.以用户画像构建智慧阅读推荐系统[J].图书馆学研究，2018(4)：92-96.
WANG S Q.Using the smart recommendation system to promote the personalized reading[J].Research on Library Science，2018(4)：92-96.

[6] 胡媛，毛宁.基于用户画像的数字图书馆知识社区用户模型构建[J].图书馆理论与实践，2017(4)：82-85.
HU Y，MAO N.User modeling of digital library knowledge community based on user portrait[J].Library Theory and Practice，2017(4)：82-85.

[7] 刘海鸥，姚苏梅，黄文娜，等.基于用户画像的图书馆大数据知识服务情境化推荐[J].图书馆学研究，2018(24)：57-63.
LIU H O，YAO S M，HUANG W N，et al.Library big data knowledge service situational recommendation based on user profile[J].Research on Library Science，2018(24)：57-63.

[8] 邓志文，都平平，秦丽，等.面向社交网的图书馆信息主动推送方法研究：以“人人网”为例[J].图书馆杂志，2015，34(3)：84-89.
DENG Z W，DU P P，QIN L，et al.Study on social-network-oriented methods of active push service of library：taking renren web as an example[J].Library Journal，2015，34(3)：84-89.

[9] 尹婷婷，曾宪玉.深度学习视角下图书馆馆藏资源推荐模型设计与分析[J].现代情报，2019，39(4)：103-107.
YIN T T，ZENG X Y.Design and analysis of the library resources recommendation system based on the deep learning[J].Journal of Modern Information，2019，39(4)：103-107.

[10] 王仲钰，刘凯俐.基于协同过滤和关联分析的图书推荐系统[J].现代商贸工业，2019，40(35)：209-211.
WANG Z Y，LIU K L.Book recommendation system based on collaborative filtering and association analysis[J].Modern Business Trade Industry，2019，40(35)：209-211.

[11] 王连喜.一种面向高校图书馆的个性化图书推荐系统[J].现代情报，2015，35(12)：41-46.
WANG L X.Personalized books recommender system for university library[J].Journal of Modern Information，2015，35(12)：41-46.

[12] 李澎林，洪之渊，李伟.基于兴趣度与类型因子的高校图书推荐算法[J].浙江工业大学学报，2019，47(4)：425-429.
LI P L，HONG Z Y，LI W.Book recommendation algorithm base on the interest and type factor for university[J].Journal of Zhejiang University of Technology，2019，47(4)：425-429.

[13] 奉国和，黄家兴.基于 Hadoop 与 Mahout 的协同过滤图书推荐研究[J].图书情报工作，2013，57(18)：116-121.
FENG G H，HUANG J X.Research on collaborative filtering book recommendation based on hadoop and mahout[J].Library and Information Service，2013，57(18)：116-121.

Research on personalized book recommendation of university library based on user profile

Wang Dafu[a]，Deng Zhiwen[a]，Jia Zhiyong[a]，An Jiyong[b]

(a. Library；b. School of Computer Science and Technology，China University of Mining and Technology，Xuzhou 221116，China)

Abstract：Personalized recommendation service is the focus of the construction of University Smart Library. This paper puts forward the overall architecture of book recommendation system. Firstly，the user portrait model is constructed from the tag dimensions of readers'attributes，behavior and interests. Secondly，considering the characteristics of differences in readers' cognitive ability，the readers are divided into different identity types，and then combined with collaborative filtering a hybrid recommendation algorithm based on content and attribute similarity is used for book recommendation. Finally，Top-N books are recommended to target readers through Hadoop big data platform. The experimental results show that the book recommendation system based on this architecture model has high recommendation accuracy and effectively alleviates the cold start problem of the recommendation system.

Keywords：recommendation system；smart library；user profile；cold start

[责任编校　陈留院　赵晓华]