基于用户画像的数字化阅读推广平台设计

郑素萍
（宁夏回族自治区图书馆，宁夏银川 750002）

摘要：为进一步优化数字化阅读的推广效果，设计基于用户画像的数字化阅读推广平台。从信息层、数据处理层和推广层3个方面设计数字化阅读推广平台总体架构，利用爬虫原理设计数字化阅读文本采集器，结合数字化阅读推广器的设计完成平台的硬件设计；通过构建用户画像，利用自适应学习算法设计数字化阅读推广算法，完成平台的软件设计。平台性能测试结果表明，基于用户画像的数字化阅读推广平台推广路径完成度高于0.8，召回率可以达到85%以上。由此验证了设计平台在功能上可以满足设计要求，还可以通过提高数字化阅读推广的成功率和精度满足性能设计要求。

关键词：用户画像；推广平台；数字化阅读；文本采集；阅读流量
DOI：10.3772/j.issn.1674−1544.2023.01.006 **CSTR：**15994.14.issn.1674.1544.2023.01.006
中图分类号：TP393 **文献标识码：**A

Design of Digital Reading Promotion Platform Based on User Portrait

ZHENG Suping
(Library of Ningxia Hui Autonomous Region, Yinchuan 750002)

Abstract: In order to further optimize the promotion effect of digital reading, a digital reading promotion platform based on user portrait is designed. The overall architecture of the digital reading promotion platform is designed from three aspects: information layer, data processing layer and promotion layer. The digital reading text collector is designed by using the crawler principle. Combined with the design of the digital reading promoter, the hardware design of the platform is completed; through the construction of user portrait, the digital reading promotion algorithm is designed by using adaptive learning algorithm, and the software design of the platform is completed. The platform performance test results show that the completion degree of the promotion path of the digital reading promotion platform based on user portrait is higher than 0.8, and the recall rate can reach more than 85%. This proves that the designed platform can meet the design requirements in function, and can also meet the performance design requirements by improving the success rate and accuracy of digital reading promotion.
Keywords: user portrait, promotion platform, digital reading, text collection, reading traffic

0 引言

数字化阅读是一种全新而独特的文学形式，是一种以娱乐方式进行阅读的新体验，给人们带来了不同于传统书籍全新的心灵体验[1−2]。数字化阅读推广是指利用图书馆网站、微博、微信等数字平台进行阅读推广的工作，其功能主要定位于宣传与展示图书馆的阅读推广活动和资源推荐，引导读者阅读，为读者提供阅读交流与互动平台等。国内高校图书馆的数字化阅读推广主要利用了图书馆OPAC、开发专题网站、文本发布等模式。从整体上说，相较于传统阅读模式，数

作者简介：郑素萍（1969—），女，宁夏回族自治区图书馆讲师，研究方向为地方文献整理与研究、图书馆学及阅读推广研究。
收稿时间：2022 年 5 月 23 日。

字化阅读可以使用户在阅读内容、方式和资源上有更多的选择，可以营造自主阅读、探究性学习的氛围，调动用户阅读积极性，提高阅读的成效和质量。随着数字化阅读的普及，大量的新手作者开始加入，也给大众带来了更多的阅读作品。目前，较前端和知名的网站签约门槛都很高，所以很多新手作者的作品都会优先在小型网站中呈现[3-4]。而在小型网站中的推广却大多集中在榜单推荐和定时换书等传统的推荐方式上。这种方式不仅推广范围较小，而且很难有流量上的突破[5]。

崔建双等[6]提出了一种基于优选框架的数字化阅读平台推广设计，利用优化算法提取数字化阅读平台的问题特征，再基于算法性能之间的关联关系，将提取的问题特征进行分类，并对其进行资源约束，构建测试样本集，最后通过人工蚁群模拟测试样本集的支持向量机，利用优选框架对结果进行分类，实现算法的分类推荐。郭斯檀等[7]为了解决数字图书馆对于学习者的偏好推荐不准确等问题，提出了一种基于遗传学算法的数字图书馆推荐系统框架，首先提取数字图书馆的图书特征，利用遗传学对数字图书馆的特征进行优化，然后利用聚类分析法设置数字图书分类，缩小搜索空间，提高整个搜索过程的精度。但是上述方法在功能上还无法满足设计的需求。

用户画像作为一种勾画目标用户、联系用户诉求与设计方向的有效工具，用户画像在各领域得到了广泛的应用。为此，本文拟基于以上研究背景，利用用户画像设计数字化阅读推广平台，以满足其功能和性能的设计要求。

1　数字化阅读推广平台硬件设计

1.1　总体架构

目前的阅读推广平台或读书推荐平台主要是在已有图书内容基础上的推荐，而数字化阅读涉及较多的是线上阅读资源，其需要较好的数据搜索性能和较高的用户兴趣定位，已有平台在推荐数字化阅读时无法满足功能和性能的设计要求。为了提高数字化阅读推广平台的信息推广能力，提升网络搜索引擎与数据库匹配的挖掘能力，本文首先对数字化阅读推广平台的总体架构进行优化设计。利用人工蚁群算法[8]提取数字化阅读的相似语义[9]，并融合数字化信息特征，构建数字化阅读推广关键词之间的关联关系，利用聚类分析法构建数字化阅读推广模型[10]，实现对数字化阅读平台的推广管理。数字化阅读推广平台总体架构，如图1所示。

从图1可以看出，数字化阅读推广平台总体架构模型分为3个层次，分别是数据层、业务层和应用层。

（1）数据层主要是建立数字化阅读数据库，根据数据挖掘技术，实现数字化阅读推广平台与网络之间的交互作用[11-12]。

（2）业务层主要负责完成数字化阅读信息的特征分析，并通过逻辑控制器对数字化阅读数据进行逻辑分析。根据搜索引擎的搜索数据智能地与数据库信息进行匹配，实现对数字化阅读推广平台信息的解析。

（3）应用层主要通过I/O接口[13]通信根据用

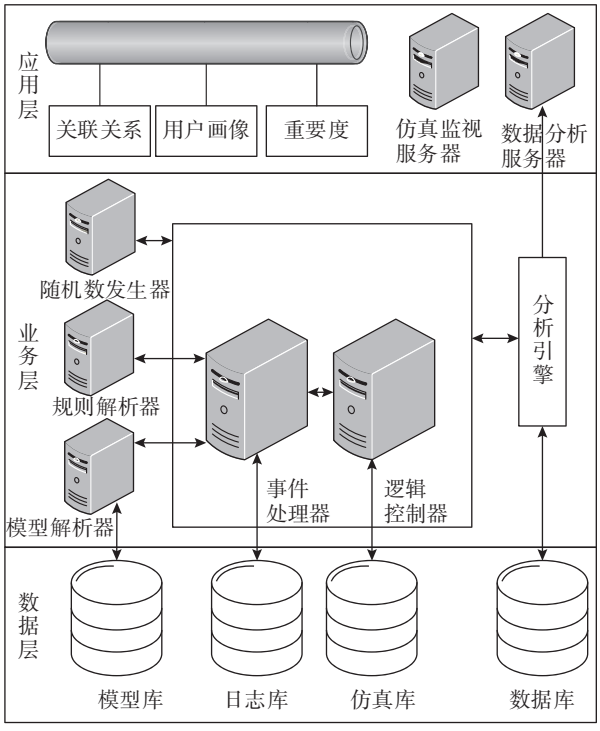

图1　数字化阅读推广平台总体架构模型

户的搜索关键词，结合架构模型对整体数据进行分析，完成高精度的推广任务。

1.2 数字化阅读文本采集器设计

数字化阅读文本采集器通常利用爬虫原理进行信息采集[14-15]。该工具位于设计平台的数据层，是数字化阅读推广平台设计中的关键资源。文本采集器决定了数字化阅读推广平台资源的可用度，对整个平台的运行具有重要的影响作用。

数字化阅读文本采集器主要通过爬虫原理的第三方jar包Jsoup实现信息采集的。通过爬虫原理构建数字化阅读文本库，不仅功能成熟，而且操作方便快捷[16-17]。通过用户搜索内容与数字化阅读文本库内容进行匹配处理。数字化阅读文本采集器爬虫原理如图2所示。

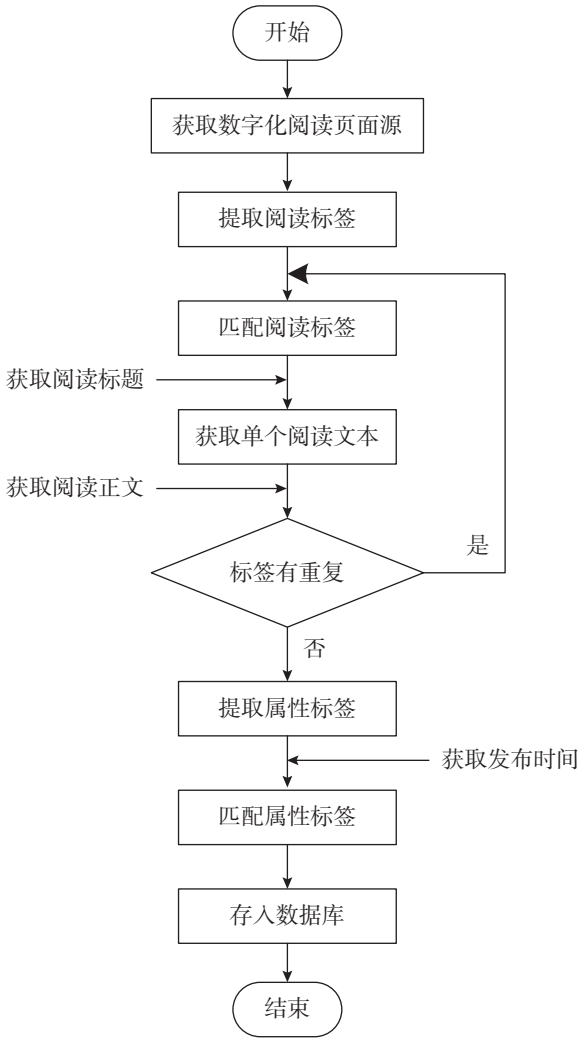

图2 阅读文本采集器的爬虫原理

爬虫步骤1：构建实体类的数据库，包含数字化阅读的标题和主要内容关键字，用于与搜索内容进行匹配[18]。

爬虫步骤2：在浏览器中，查看数字化阅读文本采集器的网页源码，从而获得数字化阅读的主要内容和文章简介。

爬虫步骤3：判断标签是否有重复。

爬虫步骤4：通过阅读文本采集器的爬虫完成数字化阅读文本采集器的设计。

1.3 数字化阅读推广器设计

数字化阅读推广器位于设计平台的应用层，其主要由信息模型、学习者模型和检索模型3个模块构成，如图3所示。

对于数字化阅读推广平台的管理者，数字化阅读推广器应具备信息资源上传和管理的功能。该工具需要根据数字化阅读的特性构建初始推广架构，利用数字化阅读采集器构建的数据库里的资源[19]，根据检索模型对数据库内的数据进行匹配搜索。在信息检索资源得到一定的累积后，通

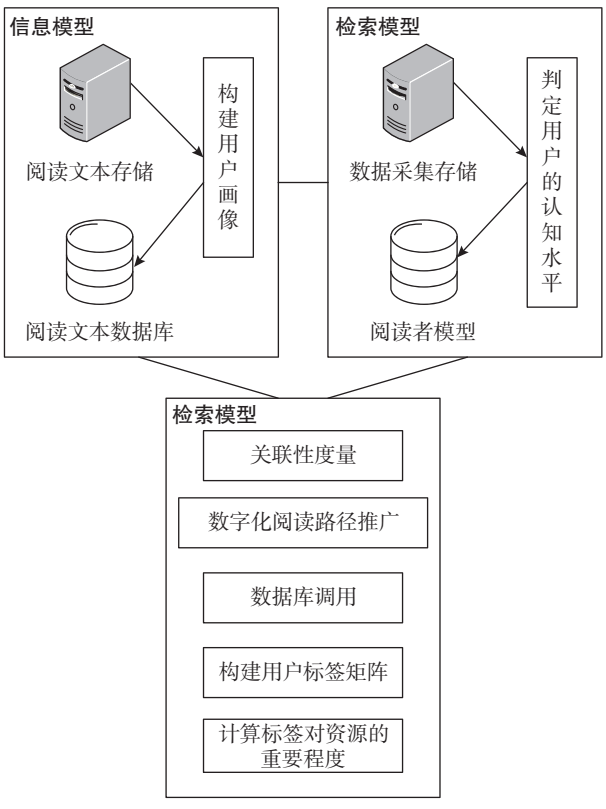

图3 数字化阅读推广器结构示意

过关联法则计算出阅读文本的关键词，从而更新数字化阅读的推广数据库。

对于数字化阅读推广平台的使用者而言，初始的使用流程较复杂。使用者首次使用数字化阅读推广器，需要填写个人信息，需要详细填写检索关键词，方便后期的推广。使用者在每次登录前，要在登录页面进行推广信息检索，通过个人兴趣偏好选择检索到的结果[20-21]。数字化阅读推广器通过对检索信息的采集与分析处理，得到使用者的个人喜好，并结合数据库为学习者提出推广建议，匹配相关的数字化阅读需求，在满足推广需求的同时，为使用者提供方便快捷的服务。

2　数字化阅读推广平台软件设计

2.1　构建用户画像

用户的画像通常包括用户的数字化阅读相关信息，主要有用户的个人偏好、背景技能、目标和任务等。基于用户画像构建的数字化阅读推广平台，可以从用户的兴趣出发为用户推广相关的数字化阅读内容，以及相关主题的内容。可以根据客户的实际阅读数据，对文章的内容进行检索分类，为用户提供更加个性化的推广服务[23]；可以针对用户画像内的背景和知识技能，为用户提出相关的阅读建议；可以根据客户的目标任务，为用户提供数字化阅读推广建议，但该项结果会受到平台和应用程序领域的影响。

利用自适应学习算法[24]计算出用户画像特征的属性值，即：

$$C_m^* = \sqrt{\frac{d_{in}(y_1)}{d_{out}(x_1)+d_{in}(y_1)}} \tag{1}$$

式中，$d_{out}(x_1)$表示以x_1为背景的用户画像属性值；x_1表示用户画像信息集合；$d_{in}(y_1)$表示以y_1为背景的用户画像属性值；y_1表示用户画像节点信息集合。

如果将C_m^*融入以x_1为背景的用户画像特征向量中，就会对用户画像特征产生干扰，表示为：

$$\overline{R}_{pq} = \sum_{x_1 \in N_x}^{N_x} \sqrt{\frac{d_{in}(y_1)}{d_{out}(x_1)+d_{in}(y_1)}} R_{jq} \tag{2}$$

式中，$\overline{R}_{pq}$表示自适应学习算法输入背景x_1对画像特征输入的预测评分；R_{jq}表示自适应学习算法输出背景x_1对画像特征输出的模拟评分。

基于式（2）的干扰，提取出用户画像特征，即：

$$T = \prod_{x=1}^{r} \left[N\left(R_{pq}\right) g(x) \right] \tag{3}$$

式中，$g(x)$表示用户画像的背景特征；$N\left(R_{pq}\right)$表示用户画像的前景特征。

根据提取出的用户画像特征，构建用户画像，利用用户画像构建阅读用户的推广模块，扩展用户画像的推广内容。

2.2　设计数字化阅读推广算法

利用采集器和推广器，根据构建的用户画像，设计数字化阅读推广算法。

步骤1：构建数字化阅读的标签矩阵，在数字化阅读的标签矩阵中，根据用户搜索的关键词，计算出各个标签使用的次数[25]，b_{1n}为用户1使用标签n的次数，以此类推。

$$user\_tags = \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \cdots & \cdots & \cdots \\ b_{n1} & \cdots & b_{nn} \end{bmatrix} \tag{4}$$

步骤2：计算标签关键词对用户的权重。利用TF-IDF 算法[26-27]，计算每个标签关键词对用户的权重信息，根据用户使用标签的次数，计算出各个标签的权重值，并将其应用到文档中，构建关键词矩阵，计算出用户的个人偏好和每个关键词出现的次数。用户m与标签关键词f之间的关联度为$R(m,f)$，可根据式（5）进行计算，其中N表示通过标签f进行检索的用户总数量。

$$R(m,f) = \frac{a_{m,f}}{\lg(1+N)} \tag{5}$$

步骤3：构建数字化阅读的关键词资源矩阵，见式（6）。式中，d_{11}代表资源1在资源库中被搜索1次，d_{1n}代表资源1在资源库中被搜索n次，以此类推。

$$item\_tags = \begin{bmatrix} d_{11} & \cdots & d_{1n} \\ \cdots & \cdots & \cdots \\ d_{n1} & \cdots & d_{nn} \end{bmatrix} \tag{6}$$

步骤4：计算标签对数字化阅读推广资源库的

权重。计算方法与步骤 2 的计算公式类似，则标签f对数字化阅读推广资源库的权重值计算公式如式（7），W为定义标签所有资源库的关键词个数。

$$R(i,f)=\frac{d_{i,f}}{\lg(1+W)} \tag{7}$$

步骤 5：根据式（7）完成用户画像的扩展，预测用户m对标签f关键词的喜爱程度，并用$P(m,f)$表示。根据计算结果，为用户推广排名靠前的数字化阅读资源[28]。

$$P(m,f)=R(m,f)\times R(i,f)=\frac{a_{m,f}}{\lg(1+N)}\times\frac{b_{i,f}}{\lg(1+W)} \tag{8}$$

以上 5 个步骤完成了数字化阅读推广算法的设计，根据计算结果可以为用户提供优质精准的推广服务。

3 测试分析

3.1 测试环境

为了验证基于用户画像的数字化阅读推广平台是否满足设计要求，可以对平台的功能和性能进行测试。图 4 显示了平台测试环境的结构。

3.2 测试工具

本文主要采用UI Automator工具[29]和Monkey工具[30]完成对数字化阅读推广平台进行自动化测试。UI Automator工具编写的脚本可以在不同操作平台上运行，实现用户的不同操作，还可以对平台的各项功能生成详细的报告，方便开发人员随时发现平台问题；Monkey工具是一种命令执行工具，可以将其安装在真实设备中，向推广平台发送用户的相关信息，实现对推广平台性能的测试。

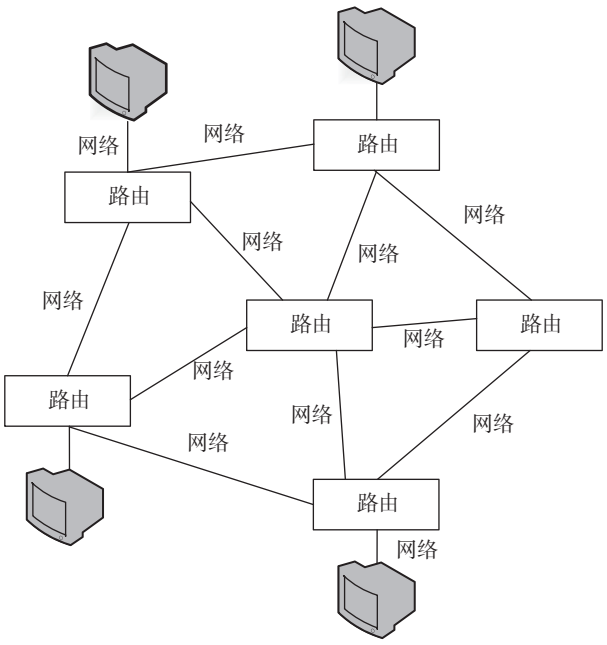

图 4　实验环境结构

3.3 测试参数

在平台测试过程中涉及的参数如表 1 所示。

表 1　测试参数

参数编号	参数名称	参数大小
1	CPU	E7−7640/4.00 GHz
2	操作系统	Ubuntu 15.08
3	内存	32 GB
4	硬盘	RAID1 SAS 500 GB × 2

3.4 性能测试

为了验证平台的性能，选取 1 000 名读者作为实验样本。被试者包括学生、教师、教育从业者、其他工作者等多种职业类型，年龄为 15 ～ 35 岁，每种职业和人数均为随机选取。将基于用户画像的数字化阅读推广平台与基于多分类支持向量机的推广平台和基于模糊本体和遗传算法的推广平台进行对比，利用推广路径完成度衡量数字化阅读推广的成功率，完成度越高，说明推广的成功率越高，反之则越低；利用召回率衡量数字化阅读推广的精度，召回率越高，说明推广的精度越高，反之则越低。

3 个平台的推广路径完成度测试结果如图 5 所示。从图 5 可以看出，随着读者人数的增加，3 个平台的数字化阅读推广路径完成度逐渐下降，但是基于用户画像的数字化阅读推广平台的推广路径完成度仍然可以保持在 0.8 以上，说明文中平台在推广数字化阅读时的成功率更高，具有更好的稳定性。

3 个平台的数字化阅读推广召回率测试结果如图 6 所示。从图 6 可以看出，基于用户画像的数字化阅读推广平台的召回率可以达到 85% 以上，说明在推广数字化阅读时，该平台的推广精度更高，可以保证平台的稳定性。

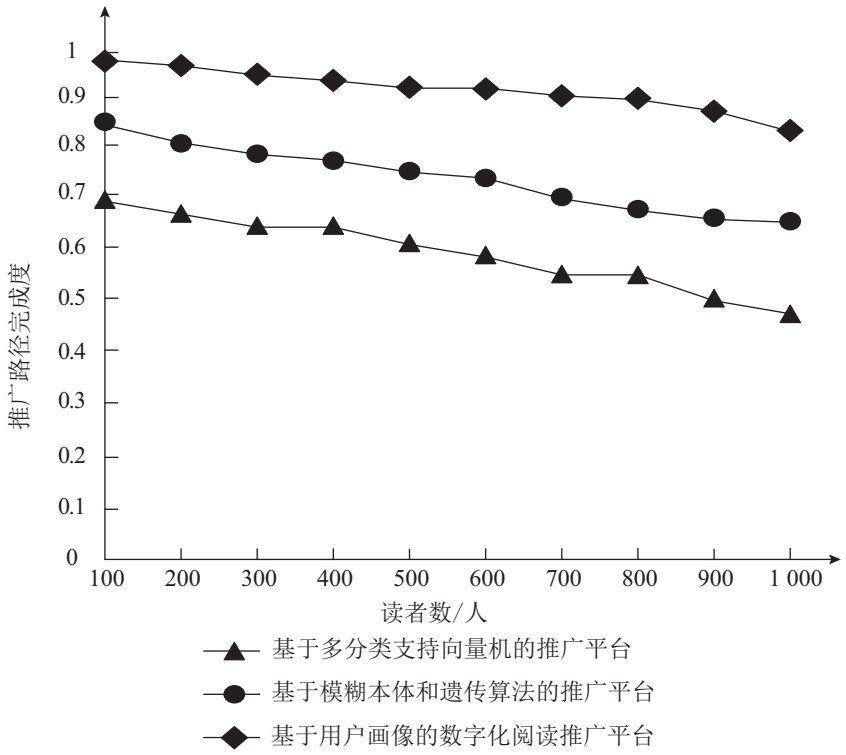

图 5　推广路径完成度测试结果

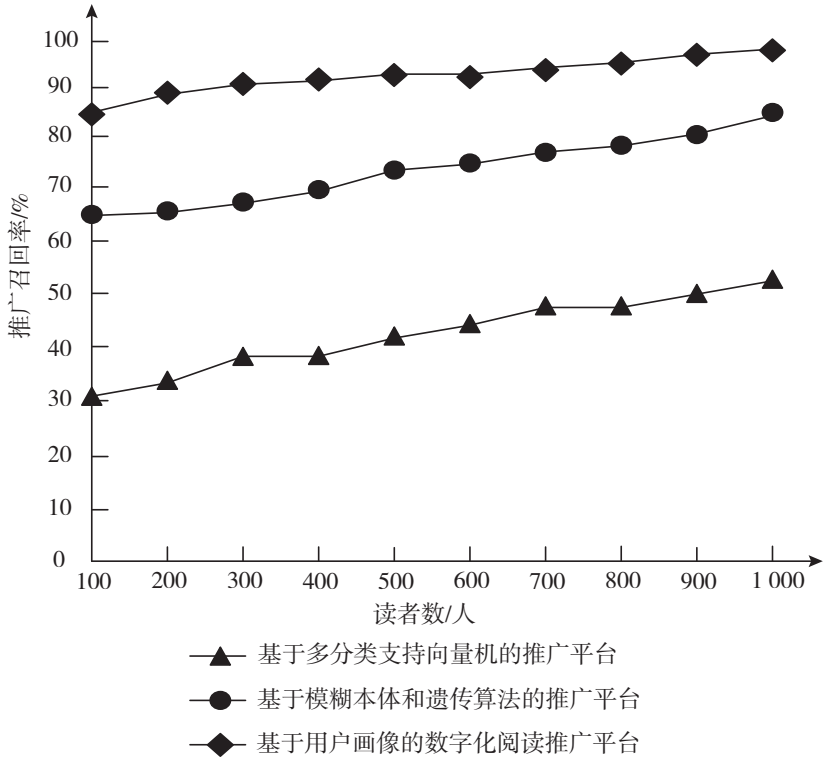

图 6　数字化阅读推广召回率测试结果

4　结语

为了提高数字化阅读推广平台的信息推广能力，提升网络搜索引擎与数据库匹配的挖掘能力，优化推广效果，本文设计了基于用户画像的数字化阅读推广平台。经测试发现：与基于多分

类支持向量机的推广平台和基于模糊本体和遗传算法的推广平台相比，该平台的推广路径完成度可以保持在 0.8 以上，说明在推广数字化阅读时，该平台成功率更高，具有更好的稳定性；该平台的召回率可以达到 85% 以上，说明在推广数字化阅读时，该平台的推广精度更高，可以保证平台的稳定性。因此，基于用户画像的数字化阅读推广平台的功能和性能都可以满足数字化阅读推广的设计要求。在今后的研究中，希望可以设计一套流量控制系统，提高读者的留存率。

参考文献

[1] 谢修娟, 莫凌飞, 李香菊, 等. 情境感知的移动阅读个性化推荐算法研究 [J]. 高技术通讯, 2019, 29(7): 640−647.

[2] 田秀峰. 新媒体时代移动终端数字化阅读特点分析: 以“扇贝阅读”APP 为例 [J]. 传媒, 2021(4): 61−63.

[3] 刘旭青, 柯平. 目录学知识在阅读推广活动中的应用价值 [J]. 图书馆论坛, 2019, 39(6): 133−138.

[4] 张立, 熊秀鑫, 周琨, 等. 对近年来数字出版评优产品的追踪测评及分析(Ⅱ): 传统书刊报单位 PC 端网站测评报告 [J]. 科技与出版, 2021(8): 47−57.

[5] 赵泽昱, 陈健, 张月琴. 基于情感空间的用户阅读兴趣模型研究 [J]. 计算机工程, 2019, 45(1): 308−314.

[6] 崔建双, 车梦然. 基于多分类支持向量机的优化算法智能推荐系统与实证分析 [J]. 计算机工程与科学, 2019, 41(1): 153−160.

[7] 郭斯檀, 潘广贞, 赵利辉, 等. 基于模糊本体和遗传算法的推荐系统 [J]. 计算机工程与设计, 2019, 40(3): 241−245.

[8] 巫红霞, 谢强. 基于加权社区检测与增强人工蚁群算法的高维数据特征选择 [J]. 计算机应用与软件, 2019, 36(9): 285−292, 301.

[9] 阮光册, 谢凡, 涂世文. 基于 Word2vec 的图书馆推荐系统多样性问题应用研究 [J]. 图书馆杂志, 2020, 39(3): 124−132.

[10] 刘海鸥, 黄文娜, 苏妍嫄, 等. 大数据深度融合的移动图书馆情境化推荐 [J]. 情报科学, 2019, 37(1): 70−75.

[11] 程秀峰, 张孜铭, 孟亚琪, 等. 知识找回场景下推荐系统模拟实现及评价研究 [J]. 图书情报工作, 2019, 63(16): 72−83.

[12] 郑鹏怡, 陈进朝. 基于发布订阅的实时交互平台 NetDDS 的设计与实现 [J]. 高技术通讯, 2021, 31(4): 435−440.

[13] 田鸿运, 武林平, 董勇, 等. 面向大规模集群的并行 I/O 用户层配置优化策略 [J]. 国防科技大学学报, 2020, 42(2): 23−30.

[14] 王茹芳, 宁璐. 基于用户画像的图书馆推荐系统研究 [J]. 图书馆建设, 2020(S1): 100−102.

[15] 杨戈, 杨麓涛. 基于爬虫和 TFIDF−NB 算法的微博情感分析 [J]. 电子技术应用, 2021, 47(4): 59−62, 66.

[16] 谢蓉蓉, 徐慧, 郑帅位, 等. 基于网络爬虫的网页大数据抓取方法仿真 [J]. 计算机仿真, 2021, 38(6): 439−443.

[17] 王征, 梁建华. 高校内部在线教学资源快捷推荐模型研究 [J]. 情报理论与实践, 2021, 44(5): 180−186.

[18] 张艳. 一种数字图书馆资源聚合质量推荐模型 [J]. 信息技术, 2021, 45(8): 5.

[19] 田野. 关联数据驱动的学术资源语义检索推荐系统框架 [J]. 图书馆理论与实践, 2019(2): 49−54.

[20] 吴谈, 周栋, 包恒泽. 基于用户类别兴趣偏好的个性化排序方法 [J]. 湖南科技大学学报(自然科学版), 2020, 35(1): 104−112.

[21] 黄小根. 基于 Web 知识发现的图书数字资源个性化检索系统 [J]. 计算机系统应用, 2021, 30(8): 111−117.

[23] 赵海燕, 汪静, 陈庆奎, 等. 主动学习在推荐系统中的应用 [J]. 计算机科学, 2019, 46(S2): 153−158, 184.

[24] 杜晨杰, 杨宇翔, 伍瀚, 等. 旋转自适应的多特征融合多模板学习视觉跟踪算法 [J]. 模式识别与人工智能, 2021, 34(9): 787−797.

[25] 王帅, 孙喜民, 高亚斌, 等. 基于神经协同过滤的个性化商品推荐方法 [J]. 信息技术, 2021(6): 143−147.

[26] 郑鑫. 传统出版业数字化转型路径探究 [J]. 中国传媒科技, 2020(8): 29−31.

[27] 许丽, 焦博, 赵章瑞. 基于 TF−IDF 的加权朴素贝叶斯新闻文本分类算法 [J]. 网络安全技术与应用, 2021(11): 31−33.

[28] 艾宪仓, 岳铁军. 基于深度学习的小目标检测区域数据推荐算法 [J]. 信息技术, 2020, 44(5): 54−57, 63.

[29] 秦利园. 移动端测试辅助工具安卓平台远程桌面监控系统的研发 [D]. 长春: 吉林大学, 2016.

[30] 叶佳, 葛红军, 曹春, 等. 规则驱动的 Android 应用 DFS 测试技术 [J]. 计算机科学, 2018, 45(9): 99−103, 118.