



项目编号: 2019048

CALIS 全国农学文献信息中心研究项目 结题报告

项目名称: 基于用户画像的高校图书馆资源推荐服务研究

项目关键词: 信息服务、资源推荐、用户画像

项目单位(盖章): 西北农林科技大学图书馆

通信地址: 陕西省咸阳市杨陵区邠城路3号

(详细地址含邮编) 西北农林科技大学图书馆 712100

项目主持人: 张莹

联系电话: 029-87080138

电子邮件: 119051135@qq.com

提交日期: 2020年5月

基于用户画像的高校图书馆资源推荐服务研究

关键词：用户画像，图书馆，资源推荐，服务

1 研究背景、目的及意义

随着高校图书馆的深入发展，其馆藏资源日益丰富，仅仅将数据资源进行简单检索和结果堆放已无法满足用户需求，海量的资源并没有为用户需求提供实质性的帮助，反倒使得用户产生了信息迷航。于是，个性化服务的概念逐渐被人们所熟知来应对用户存在的多元化服务需求或者信息需求。推荐研究作为个性化服务研究领域的重要分支之一，已经在电子商务等领域广泛应用，它通过历史记录的数据分析得到有价值的且关联性高的资源并生成个性化信息推送给用户。因此，个性化推荐服务的发展，引起用户获取信息方式的变革。图书馆领域也逐渐意识到个性化服务的重要性，个性化推荐服务在图书馆信息服务中的应用也越来越普遍。

高校图书馆资源推荐服务的核心是用户，为实现推荐资源的有效利用，研究用户信息偏好、状态、环境等所产生的影响具有重要意义。通过对用户数据的收集、分析、处理等，实现每个用户的详尽、立体而又直观的描述，从而为用户推荐提供更加有效的个性化信息，减少在检索决策中的不确定性。如何实现实时、准确性、高效性的服务，为不同用户提供不同的信息需求是图书馆领域的重要研究方向。

目前，高校图书馆个性化推荐服务已获得了令人瞩目的成绩，在提升资源推荐服务质量及资源利用率方面已取得良好的效果。但在用户需求感知方面，依然以传统的用户需求调研为主，以分析小体量结构化数据分析用户需求，难以精准、全面刻画用户需求，因而可能导致个性化信息服务质量差、效率低等问题。用户画像作为大数据环境下的用户描述工具，在用户描述与建模上具有巨大优势，因而在分析用户信息需求时，通过构建用户画像完成对用户个人习惯以及兴趣等内在需求信息的刻画，从而构建起通往用户兴趣的桥梁，为个性化推荐服务提供数据基础。所以，建立基于用户画像的高校图书馆用户服务模式，深入研究相关的理论，应用相关技术，对实现高校图书馆服务水平的提高有着较为重要的意义。

2 研究内容及方法

2.1 用户画像及研究现状

2.1.1 用户画像的内涵

交互设计鼻祖 Alan Cooper 在研究中最早提出用户画像并对这一概念给出了定义，他指出用户画像是通过不同的呈现方式对用户真实数据的刻画，是建立在一系列真实数据之上的目标用户模型^[1]。简言之，利用各种数据挖掘技术采集并清洗用户的社会属性、生活习惯和消费行为等信息，并通过标签映射的方式抽象出用户画像模型。文献[2]等将用户画像描述为一种刻画用户需求和兴趣偏好的形象集合。余孟杰^[3]认为用户画像就是从海量用户数据中挖取出特定的、潜在的用户信息，并以标签化的形式呈现出用户全貌，用户画像是用户信息面貌的虚拟刻画，其目标是实现精准服务。随着互联网和大数据理念的深入，如何利用大数据实现精准服务已成为企业日益关注的焦点，用户画像作为一种了解和预测客户、精准定位客户群体的有效工

具，被广泛应用于社交网络、金融、电子商务等领域。

David Travis 认为一个令人信服、全面、精准化的用户画像应具有基本性（Primary research）、真实性（Realistic）、目标性（Objectives）、独特性（Singular）、移情性（Empathy）、数量（Number）和应用性（Applicable）7 个基本条件^[4]。文献[5]认为图书馆领域在以数据驱动的主路线下，用户画像应该满足 6 个条件：时效性、可迭代性、交互性、区隔性、知识性以及聚类性。

2.1.2 国内外研究现状

随着人工智能、大数据领域的悄然兴起，围绕用户画像的相关研究也相继展开，并逐步成为众多学者关注的热点。用户画像研究在国外图书馆领域开展较早，在 20 世纪 80 年代，英国国家书目和 Blaise-line 通过对用户的使用情况进行相关性分析，并以此为依据对其服务进行优化^[6]。用户画像技术算法是国外学者主要的研究方向，如统计分析、聚类分析算法、加权算法、分类算法、朴素贝叶斯网络、深度学习方法等受到学者们的关注，不同的研究采用不同的算法构建用户画像模型，从而刻画用户显性的特征和潜在的特征。

国内图书馆领域对用户画像研究还起步发展阶段，相关研究成果较少，并且大多研究还停留在理论阶段。陈慧香^[7]等人详细阐述了用户画像的概念、特性以及作用，并深入分析了用户画像建模的相关算法与技术、框架模型设计及构建流程，为图书馆基于用户画像的服务重新研究奠定了基础。刘速^[8]详细阐述了天津图书馆用户画像构建的具体方法和步骤，并对常用的用户画像分析方法，如用户关系图谱、多维度交叉分析、可视化统计描述等方法进行了比较。潘宇光^[9]将用户画像理论融入到构建智慧图书馆的过程中，从图书馆的智慧构建和智慧服务角度，提出用户画像的构架与实现途径。

2.1.3 研究述评

通过上述国内外研究现状的总结，可看出国内外对用户画像的理论和实践探索中，国外深入探讨用户画像的构建方法、技术和具体实践，并逐步应用在图书馆的规划设计中，而国内用户画像引入初期就与大数据背景紧密结合，主要集中在用户画像的定义、模型构建等理论浅层面研究，实践化程度不高，还未形成具体的实践案例。

2.2 高校图书馆用户画像设计与实现

2.2.1 数据收集

用户画像在海量的用户数据中产生和形成的，构建用户画像是为了还原用户信息，数据来源于所有用户相关的数据，因此数据源是构建用户画像的基础。随着大数据技术的发展和运用，图书馆各信息资源系统之间实现了互通互联，尤其是统一身份认证的应用，以及移动终端的普及，给用户带来方便的同时，也产生了异构化、立体化和多样化的数据，丰富了用数据类型，拓展了用户数据的来源。高校图书馆构建用户画像的首要工作是确定数据源，图书管理系统、书目检索系统、图书馆网站、移动图书馆等都是用户数据的主要来源，用户数据可分为相对稳定的静态信息数据和变化较快的动态信息数据两类。

2.2.2 标签建模

标签是基于用户数据分析后人为规定的高度精炼的用户特征标识，人能很方便地理解每个标签含义，使模糊的用户立体化、形象化。为用户建立标签模型是构建用户画像的核心工作和首要任务，用户标签建模就是对用户的静态信息数据和动态信息数据进行分析以及语义化抽象出，从而形成短文本化标签。高校图书馆从读者的信息需求偏好去考虑，对采集的数据进行清洗分析，可以从读者基本属性、行为痕迹、信息资源偏好、读者活跃度等多角度去构建用户画

像的标签体系。其中，读者属性标签主要是对读者基本特征的描述，由采集的读者属性数据构成。行为属性标签是对读者操作图书馆系统以及数字资源的登录、浏览、下载等行为特征的描述，由所采集的读者行为数据构成。信息资源偏好标签主要是对读者信息资源需求的潜在特征的描述，主要由读者检索的信息资源内容、类别、来源等构成。而读者活跃度标签主要是对读者信息需求的流失概率进行预测，是对读者登录图书馆系统的频次、浏览的时间、访问量等行为特征进行描述，主要也是由读者行为数据构成。将上述读者属性标签和时间、路径痕迹等进行组合，可以形成多级立体的读者标签体系。

2.2.3 模型构建

用户画像构建一般包含数据整合层、数据标签层以及画像应用层。数据整合层又可以分为数据采集和数据处理，通过对图书馆各种来源数据进行采集，再进行集成、转换、归约、清洗等预处理操作后形成原始数据库文件。用户画像构建的关键是数据标签层，通过对用户信息进行统计、分类、聚类分析等技术，将用户信息语义化和短文本化，以标签的形式呈现立体化的用户形象。画像应用层是在建模结果分析的基础上，采用不同的可视化技术，将用户各个维度的标签呈现出来，图书馆用户画像框架如图 1 所示。

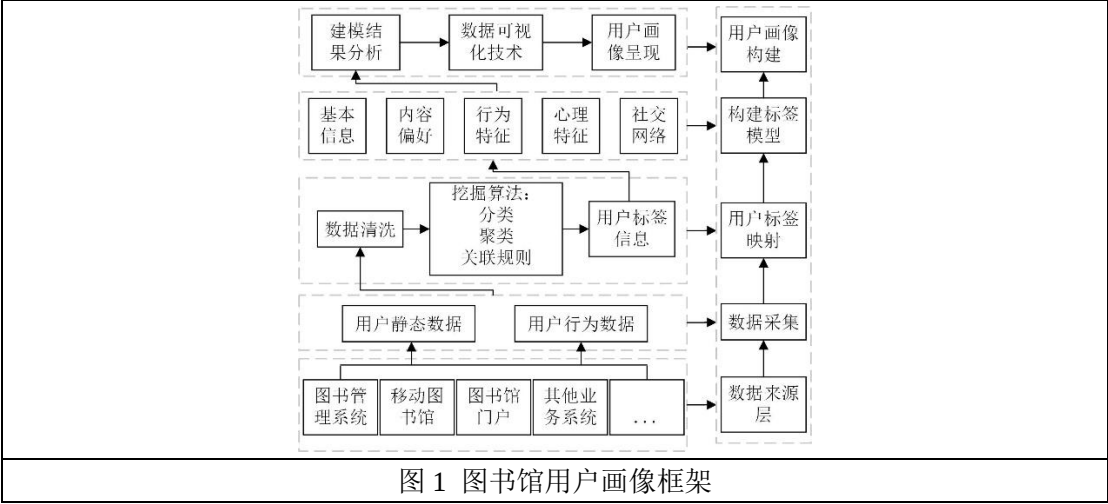


图 1 图书馆用户画像框架

2.2.4 图书馆用户画像结果可视化

本研究构建的画像包括学生和教工两类画像，学生画像选择西北农林科技大学图书馆 2018 年 1-6 月的用户借还书作为主要数据来源；教工画像数据来源主要为西北农林科技大学教工发表的论文和获得的专利，利用爬虫技术从 CNKI 中抓取，并进行规制清洗和筛选形成教师数据库文件。学生用户的属性数据反映了借还书用户的明确数据类型，包括用户的姓名、学号、级别、院系，图书数据包括图书的题名、作者、条码、借阅日期、归还日期。教工用户的属性数据反映了该教工研究的兴趣和研究方向。基本数据通过收集整理、标签建模、提取关键词。汇总后，采用用户画像工具词云。

学生画像，根据学生借阅记录（类别、次数、时长），从基本属性、兴趣爱好、访问偏好等维度构建画像，如图 2 所示。

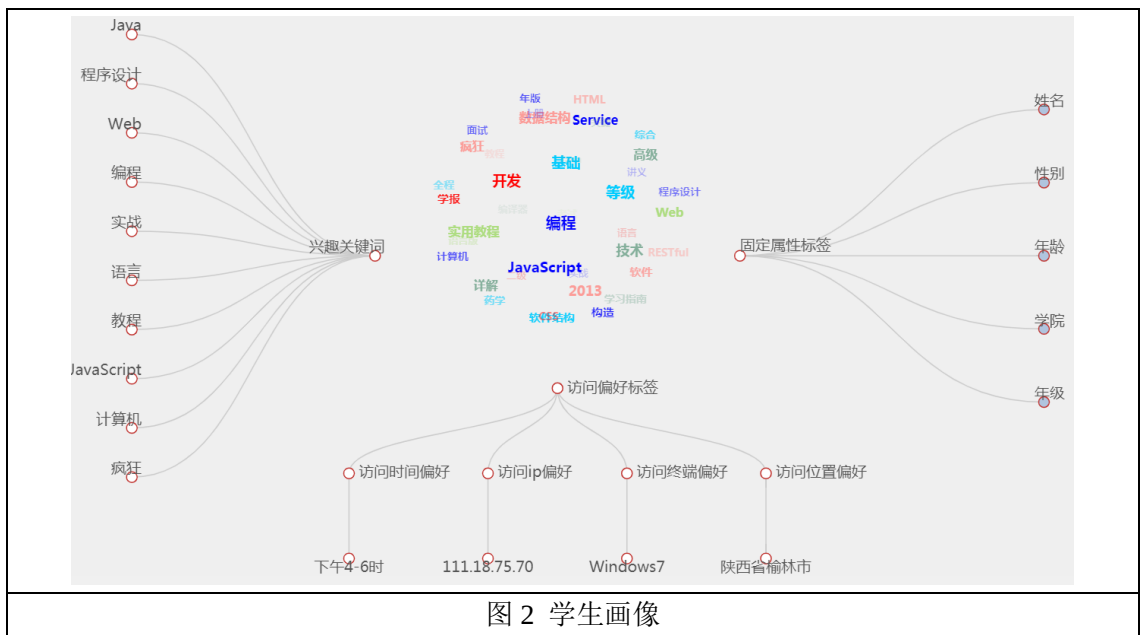


图 2 学生画像

教工画像，根据所发表的论文和专利等，进行学术研究领域画像，从基本属性、学术领域和社交属性（合作伙伴、相似用户）三个维度构建画像，如图 3 和图 4 所。

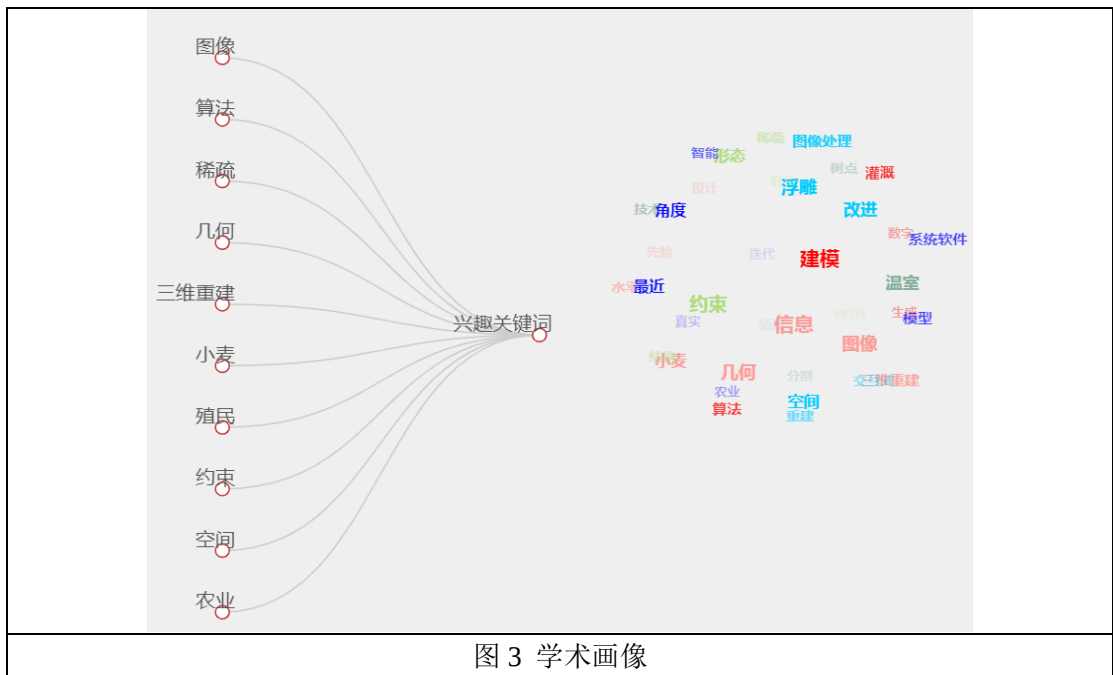


图 3 学术画像

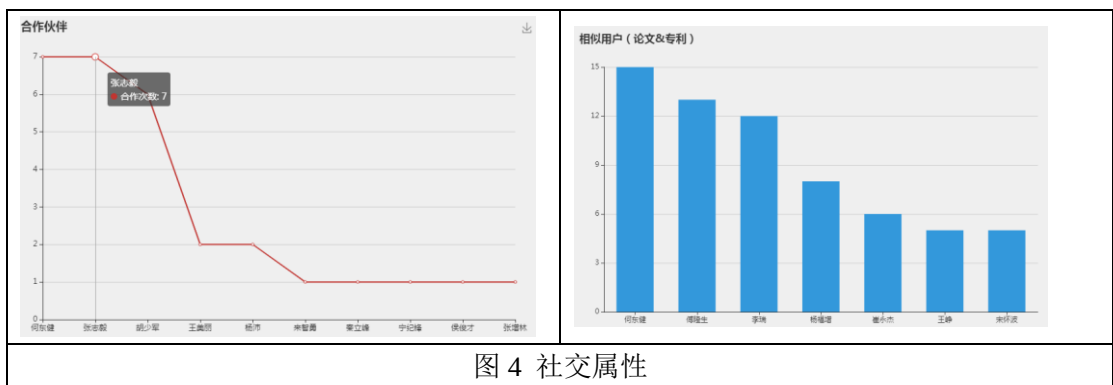


图 4 社交属性

2.3 基于用户画像的高校图书馆资源推荐

在用户画的基础上采用 K-means 聚类算法对用户属性信息和用户行为数据进行聚类，将用户划分成不同的类别；然后计算目标用户与各类别的距离，并选择最近距离的类别作为目标用户的检索空间；最后在目标用户的检索空间中通过用户相似度计算确定目标用户的最近邻居，在此基础上产生推荐列表，提高推荐精度，降低推荐所耗时长。

2.3.1 聚类算法

数据挖掘的主要任务之一是聚类分析，它将物理或抽象对象的集合划分成多个类簇，同一个簇中的对象彼此相似，不同簇中的对象彼此相异。在众多的聚类算中最常用，经典的是 K-means 聚类算法，K-means 聚类原理是：依据两个对象的距离越大，相似度越小的准则，通过最小化误差函数优化将所有对象划分为 K 个簇，K 为预先确定的常数。K-means 聚类算法具有速度快，质量高，伸缩性好等优点，经常被应用到各种聚类任务中，其主要流程为：

(1) 假设一个数据对象包含了 n 个对象，在这些对象中随机选择 K 个对象作为聚类中心的初始值。

(2) 通过距离公式对每个对象到聚类中心的距离进行计算，从中根据最近原则将对象依次分配到距离它最近的聚类中心存在簇中，计算误差平方和准则函数 E 的值。

(3) 设立新的聚类中心，通过计算各个簇中的所有对象的平均值作为各个簇的聚类中心以此来得到误差平方和准则函数 E 的值。

(4) 把步骤 (3) 的 E 值同前一次计算 E 值来对比，假如两个 E 值的差的绝对值小于等于之前设定好的阈值，则聚类的准则函数收敛，转到步骤 (5)，否则转到步骤 (2)。

(5) 输出 K 个聚类。

2.3.2 协同过滤推荐算法

作为当前应用最广泛的个性化推荐技术之一，协同过滤推荐算法在用户行为中寻找特定模式，创建用户特有的推荐内容。协同过滤推荐算法分为三类：基于记忆方法、基于模型方法和基于用户。协同过滤本算法优点是无需了解领域知识，大多数情况推荐结果令人满意。基于用户的协同过滤算法包含以下步骤：

(1) 相似度计算构建目标用户兴趣相似的用户集合，Pearson 相关系数常用来计算两个用户之间的相似度。

$$S_{uv} = \frac{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{i \in I_{uv}} (r_{vi} - \bar{r}_v)^2}} \quad (1)$$

式中 S_{uv} 表示目标用户 u 与用户 v 的相似度， I_{uv} 是用户 u 和 v 都评分过的共同文， r_{ui} 是目标用户 u 对物品 i 的评分， r_{vi} 是目标用户 v 对物品 i 的评分， $\bar{r}_u$ 和 $\bar{r}_v$ 表用户 u 和用户 v 所有评分的平均分

(2) 评分预测计算预测目标用户对其没有评价过文献的评分，基于用户相似集合通过评分预测预测评分。

$$P_{ui} = \frac{\sum_{v \in K} (r_{vi} - \bar{r}_v) \times S_{uv}}{\sum_{v \in K} S_{uv}} \quad (2)$$

公式中, P_{ui} 是目标用户 u 对物品 i 的评分预测, 得出的 P 评分预测越高, 目标用户 U 感兴趣的程度或者几率就越高。

(3) 最终按照评分的高低可以给予目标用户 TOPN 的推荐列表。

2.3.3 实验及分析

(1) 数据收集及数据预处理

以西北农林科技大学图书馆图书借阅数据为基础, 通过选择借阅图书数量超过 20 次的读者产生的借阅记录, 并经清洗、集成和归约等预处理形成实验数据集, 总共生成 150000 条评分的记录, 包含 2350 个用户及 3700 本图书和 150000 条评分的记录作为实验数据集, 将数据集 80% 为训练集, 20% 作为测试集。

用户数据包括用户固有特征数据以及用户的借阅的图书数据, 数据有字符型、数值型, 需要标准化预处理, 同时为满足聚类模型, 需要用户数据对进行预处理及向量特征表示过程如下。

(1) 用户身份属性特征表示: 通过数值符号将包括年龄、性别、职业、专业及年龄等用户身份数据数字化, 如将性别数据转化为 0 和 1, 将专业, 职业等数据转化为数字, 通过数字符号转换形成如 $U1=(46,1,13,5)$ 的用户属性向量, 其表示为用户是一个计算机学院年龄为 46 岁, 从事计算机教学的女教师。

(2) 用户行为特征表示: 首先依据图书的类别将每一本图书映射成为一个对应类别位置为 1, 其它为 0 的向量, 由于图书有 22 个类别, 因而图书特征有 22 个维度; 其次, 将用户所借阅图书的向量进行累加, 用图书向量的累加向量表征用户行为特征。

(2) 评价指标

本文采用推荐算法中常用的平均绝对偏差 MAE、准确率和运行时间对预测精准度进行分析。它通过将目标用户的预测评分与实际评分进行度量对比, MAE 的值越小, 表示算法的推荐质量越高; 值越大表示推荐质量越低。

(3) 结果分析

通过多次实验当聚类数目 $K=10$ 时准确率最高, 因此在本文的实验中将用户聚类数 K 设置为 10, 为了验证文本方法在推荐中的优势, 分别采用推荐中常用的基于内容和基于用户协同过滤推荐进行试验, 3 种不同方法的 MAE 结果如表 1 所示。

	平均绝对偏差	准确率	运行时间 (秒)
本文方法	20.6575	0.7652	5504
基于内容推荐	27.6832	0.5987	5076
基于用户协同过滤	24.5368	0.6982	9154

从表 1 中可以看出, 通过与基于内容推荐和基于用户协同过滤这两个算法的实验结果比较可知, 本文提出的推荐方法在平均绝对偏差和准确率方面都要优于基于内容推荐和基于用户协同过滤推荐算法; 并且运行时间相比基于用户协同过滤有很大提升, 这是因为通过用户聚类后, 目标用户的相似用户搜索空间大大降低, 使得推荐效率更高。

3 结论与建议

目前高校图书馆领域对用户画像研究还在起步发展阶段, 相关研究成果较少, 并且大多研究还停留在理论阶段。各种类型图书馆应该对读者进行潜在的行为与需求分析, 勾画出读者画像, 准确把握读者的需求特征, 从读者需求角度入手设计图书馆的资源推荐服务, 为读者提供

智慧化、个性化精准资源服务，从而提高图书馆服务质量与资源利用率。根据本次研究给出以下建议。

(1) 整合用户数据、预测读者需求，实现精准资源推荐

用户画像在用户描述与建模上具有天然巨大的优势，各类高校图书馆需要首先将分散在各个系统读者数据进行整合和分析，对读者进行比较完整的用户画像，从画像层面动态地监测读者的信息需求和行为变化趋势，明确读者的信息需求偏好，从而预测读者的未来资源需求，最终为图书馆个性化精准资源推荐提供可靠的支持。

(2) 重视读者需求，优化馆藏资源建设

高校图书馆可利用用户画像来了解用户的整体兴趣和需求走向，根据用户使用图书馆资源的行为习惯，设计或调整图书馆资源的配置、空间布局、科学管理。高校图书馆管理层可借助用户画像技术及时发现用户的特征信息和兴趣偏好，掌握用户需求的变化，从而为决策制定的科学合理性、高效性提供可靠的数据支持。比如，通过用户画像的指标发现读者用户对电子资源的使用频率远远高于纸质资源的使用率，这说明高校图书馆在未来的资源配置率需要调整。如果出现读者用户对图书馆空间的利用需求大大增加，那么图书馆可以根据实际需要调整图书馆的物理空间和功能进行优化。

(3) 建立读者信用机制，提升资源利用

盗窃图书、恶意下载资源、图书超期、破坏公共环境及书刊污染等不良用户行为时有发生，针对这些严重的用户失信行为，图书馆往往采用警告或轻微罚款方式，其效果不是很好，起不到威慑失信用户的作用。用户信用管理是抑制用户失信现象的有效途径之一，通过收集用户数据，分析用户信用数据，将用户信用纳入个人征信记录，对高校图书馆不同层次用户进行信用评分分级，建立用户信用分级管理。对图书馆不同信用评分的用户，赋予不同权限，并给予一定的奖励和处罚，可以优化用户管理，灵活服务，促进馆藏资源利用。

(4) 建立读者评价机制，提升资源服务质量

高校图书馆需要建立科学的反馈评价机制来有效指导资源推荐服务。全方位收集读者用户信息，对不同阶段的数据进行比较分析，形成用户对不同资源或服务的反馈意见大数据，并对这些数据进行评估，不断完善资源推荐服务。此外，根据读者用户的反馈，能够真实地反映出图书馆阅读推广方面存在的问题，实现读者与图书馆的互动交流，拉近与读者的距离，使得读者真正参与到图书馆未来的建设和服务中去，将有助于图书馆精细资源服务。

4 项目成果

发表论文两篇：

- [1]. 基于用户聚类的图书协同推荐算法研究[J].科技资讯,2020,18(09):198-199.
- [2]. 图书馆用户画像设计与应用研究[J].数字通信世界,2020(02):34-57.

5 参考文献

- [1]. 杨燕.图书馆用户画像及其应用研究[J].四川图书馆学报,2018(1):63-65.
- [2]. GAUCH S,SPERETTA M.User profiles for personalized informationaccess [C].The Adaptive

Web,Methods and Strategiesof Web Personalization DBLP,2007: 54-89.

- [3]. 余孟杰.产品研发中用户画像的数据建模 [J]. 设计艺术研究, 2014, 4(6):62-64.
- [4]. Travis D.E-Commerce Usability:tools and techniques to perfect the On-Line experience[M].CRC Press,2002.
- [5]. 许鹏程,毕强.数据驱动下数字图书馆用户画像模型构建[J].图书情报工作, 2019 (2) :30-37.
- [6]. BISHOPJ,LEWISPR.BLAISE-LINEAnd The British National Bibliography:profiles of users and uses[J].Journal of librarianship & information science,1985(2):119-136.
- [7]. 陈慧香,邵波.国外图书馆领域用户画像的研究现状及启示[J].图书馆学研究,2017(20):16-20.
- [8]. 刘速.浅议数字图书馆知识发现系统中的用户画像—以天津图书馆为例[J].图书馆理论与实践,2017(6):103-106.
- [9]. 潘宇光.高校智慧图书馆读者信息需求的用户画像[J].合肥工业大学学报(社会科学版)[J],2018(02):114-119.

项目结题验收单

项目名称	基于用户画像的高校图书馆资源推荐服务研究				
主持人	张莹	职务/职称	馆员		
所在单位	西北农林科技大学图书馆(加盖公章)				
专家意见	图书馆 1、该项目阐述了用户画像的概念及特性，系统梳理了用户画像的国内外研究现状，对当前用户画像的研究进行了述评，并给出了高校图书馆用户画像的构建流程和基本框架。 2、项目以西北农林科技大学图书馆借阅记录和该校教工所发论文和专利为基础数据，从基本属性、兴趣爱好、访问偏好等维度分别构建学生和教工画像，并利用多种可视化技术进行展示。 3、该项目在用户画像的基础上，利用聚类算法和协同过滤算法构建基于用户画像的高校图书馆资源推荐方法，并在西北农林科技大学图书馆借阅记录数据上进行了实验，结果表明，基于内容推荐和基于用户协同过滤推荐算法相比，该方法能有效提高推荐精度，降低推荐所耗时长。 4、该研究项目发表了 2 篇相关的论文，“图书馆用户画像设计与应用研究” 2020 年 2 月发表于《数字通信世界》期刊；“基于用户聚类的图书协同推荐算法研究” 2020 年 3 月发表于《科技资讯》。 5、张莹同志已完成了预期的研究目标，其研究成果达到了结题的要求和标准。专家一致同意该项目结题。				
专家签字	王香兰	赵慧清	王芹		
职务/职称	研究馆员	研究馆员	副研究馆员		