

# 基于大数据挖掘和用户画像的高校图书馆个性化服务研究<sup>\*</sup>

陈 丹<sup>1</sup> 罗 烨<sup>1</sup> 吴智勤<sup>2</sup>

(1.江苏理工学院计算机工程学院 江苏常州 213001)

(2.江苏理工学院图书馆 江苏常州 213001)

**摘 要:** 充分利用大数据为用户提供个性化服务是高校图书馆在大数据时代的重要任务之一。文章提出基于大数据挖掘和用户画像的高校图书馆个性化服务模型框架,采用大数据挖掘发现用户的偏好、兴趣、需求、活跃度等全貌信息,构建用户画像,实现用户和资源的精准匹配,提供以用户为中心、以需求为导向的个性化服务。

**关键词:** 大数据挖掘;用户画像;高校图书馆;个性化服务

中图分类号: G258.6

文献标识码: A

## Research on Personalized Services in University Libraries Based on Big Data Mining and User Portrait

**Abstract** Making full use of big data to provide personalized service for users is one of the important tasks of university libraries in the era of big data. This paper proposes a model framework of university library personalized service based on big data mining and user portrait. Big data mining is used to discover user panoramic information of preferences, interests, demands, etc. and construct user portraits, precise matching between users and resources is realized, and user-centered and demand-oriented personalized service is provided.

**Key words** big data mining; user portrait; university library; personalized service

### 1 高校图书馆个性化服务和用户画像

个性化服务体现了图书馆以人为本的理念和价值观,是现代图书馆提高竞争力的重要途径。不同于以往的被动服务模式,图书馆个性化服务充分利用馆藏资源优势,主动地以用户为中心开展全方位服务,全面满足用户个性化需求。高等院校是国家科技创新、知识创新的重要阵地,高校图书馆是高等学校的重要职能部门之一。高校图书馆需要开展高质量的个

性化服务,更好地服务于学习、教学、科研,为文化建设和科技发展作出更大贡献。

为用户提供高质量的个性化服务,全面深刻地了解用户是前提。用户画像是近年来出现的一种全面勾画用户、联系用户与产品的良好工具。用户画像描绘用户的背景、身份、兴趣、需求、心理、性格等,全面细致地展现一个用户的信息全貌,为图书馆理解用户、制定服务策略提供依据和参考。在互联网+背景下,高校图书馆的用户数量和用户产生的行为、社交等数据迅速增加,形成了用户大数据。来源丰富、类

<sup>\*</sup> 本文系教育部人文社会科学基金青年项目“大数据时代高校图书馆多样性的个性化推荐服务研究”(项目编号:17YJC870011)和江苏省高校哲学社会科学研究项目“基于大数据挖掘的高校图书馆智慧服务研究”(项目编号:2018SJA1736)的成果之一。

型多样、规模巨大的用户大数据使精准用户画像的构建成为可能。以往的图书馆只能获取用户的少量信息,基于小样本进行个性化服务,在大数据时代,图书馆可以获得用户方方面面的信息和数据,从而更为精确地勾画用户,把数据转化为价值,使个性化服务更为精准,更好地满足用户需求,极大地提升用户的体验。

## 2 基于大数据挖掘和用户画像的高校图书馆个性化服务模型框架

本文提出基于用户画像的高校图书馆个性化服务模型框架,如图1所示。用户大数据是构建图书馆用户画像的宝贵资源,首先对图书馆用户大数据进行收集和整合,然后采用大数据挖掘算法分析和挖掘用户大数据,提取用户标签,构建用户画像,最后根据用户画像为用户提供满足其需求的图书、论文、专利等图书馆文献和资源,实现高质量的个性化服务。

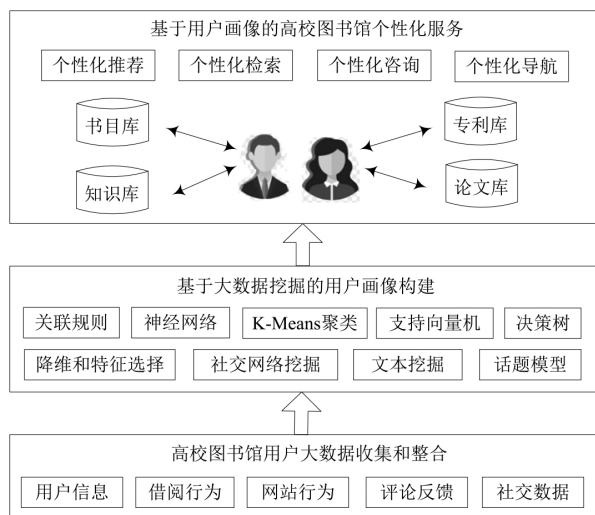


图1 基于大数据挖掘和用户画像的高校图书馆个性化服务模型框架

## 3 基于大数据挖掘的用户画像构建

用户的身份、偏好、活跃度、显性兴趣等标签可以通过传统的统计分析方法得到,而用户的隐性兴趣、潜在需求、心理、情感等则不易直接从数据中获得。通过大数据挖掘方法和技术深入分析和挖掘海量用户数据,可以洞察用户的需求、心理、情感、情绪等,提取相关标签。

### 3.1 基于大数据挖掘提取用户画像标签的途径

物联网、移动互联网、社交网络等技术在图书馆的深入应用使得图书馆数据持续不断地以超大规模增长,传统的数据统计和分析方法不能适应海量数据的处理,需要采用大数据挖掘技术分析和挖掘海量动态的图书馆大数据<sup>[1]</sup>。应用大数据挖掘技术,从图书馆用户大数据中提取用户画像标签主要有以下3种途径。

#### 3.1.1 挖掘用户行为大数据获取用户标签

应用大数据挖掘技术分析和处理用户行为大数据是提取用户画像的偏好、兴趣标签的重要途径之一。用户行为不仅包括借阅行为,也包括图书馆网站的点击、收藏、下载、阅读等行为。对于高校图书馆来说,读者行为还包括自主学习、学术文化交流以及教学和科研等内容<sup>[2]</sup>。刘春雷以浙江图书馆为例,基于数据分析对用户续借行为进行探讨,为图书馆服务工作提供参考和依据<sup>[3]</sup>。王向真以技术接受整合模型为基础,研究高校学生电子图书使用行为,进而推进电子图书资源服务的精准营销<sup>[4]</sup>。大数据挖掘算法众多,其中关联规则、神经网络、支持向量机、K-Means聚类算法等可用于分析用户行为大数据的规律和模式,从而发现用户偏好、兴趣、活跃度等特征,提取用户标签。

#### 3.1.2 挖掘用户社交大数据获取用户标签

随着社交网络技术的发展及其普遍应用,高校图书馆建立了基于微博、微信公众号、QQ群、论坛的知识服务社区,为用户和图书馆的交流和互动提供了极大的便利,也为提取用户标签提供了数据来源。用户社交网络数据由3个维度构成:用户、交流、内容。社交网络用户具有社会化、相关性强的特点,其核心是关系。社交网络使得现实社会中难以形成的关系层在互联网上可以不断涌现。社交网络中的用户关系众多,且用户之间随时进行着交流。交流包括讨论、交谈、评价、分享自己的状态更新、赞赏他人的分享和信息等。大量用户的交流形成了丰富的内容,交流内容具有类型多样性的特点。文本是交流内容中最常见的数据类型。社交网络上的文本不同于传统的文本(例如新闻),具有情感性,携带了用户或正面或负面的丰富情感。柳益君等<sup>[5]</sup>通过用户在社交网络中的兴趣相似好友来挖掘用户的多样隐性兴趣,实现多样性的阅读推荐。韩梅花等<sup>[6]</sup>根据抑郁情感词典分析用户微

博文本,计算其抑郁情感指数,得到用户画像,根据用户画像向用户推送相应的阅读治疗资源。

### 3.1.3 挖掘用户标签集得到获取用户标签

用户画像的标签体系构建是一个动态迭代过程,在图书馆用户画像建模过程中,标签不仅可以从用户数据中挖掘得到,也可以通过挖掘已有的用户标签集合来得到。郑海雁等<sup>[7]</sup>设计标签集约束近似频繁模式挖掘算法LCPP,并将该算法并行部署在MapReduce计算模型中,使之能高效处理大规模数据。周朴雄等<sup>[8]</sup>借助标签云系统的概念,对其加以改进,将其作为用户兴趣的表达方式,通过共现分析建立标签集之间的关联关系,预测用户兴趣。在已有用户标签的基础上,采用频繁模式挖掘、关联规则挖掘等大数据挖掘算法,深入分析图书馆用户画像的标签之间的关系和模式,可以预测和发现新的用户画像标签。

## 3.2 基于关联规则的隐性兴趣标签预测

关联规则算法是数据挖掘中的经典算法之一。关联规则挖掘算法最初用来解决购物篮分析问题,通过关联规则挖掘,发现顾客的购物篮里不同物品之间的关联,从而帮助商家制定营销策略。关联规则挖掘技术在零售、电商、金融、搜索引擎、智能推荐等领域有广泛应用。

应用关联规则算法分析用户画像的习惯、偏好和兴趣等标签之间的关联性,挖掘用户在学习、科研等方面的隐性兴趣和需求,将之作为新的用户画像标签,进一步完善用户画像。对所有用户画像进行关联规则挖掘计算量大且意义不明显,可针对目标用户,在该目标用户画像与其近邻用户画像的范围内进行关联规则挖掘,这样计算量小且结果更有意义。本文提出基于关联规则的图书馆用户画像隐性兴趣标签预测流程,见图2。例如,目标用户画像的兴趣标签中有“机器学习”而无“数据挖掘”,对目标用户及其近邻的用户画像进行显性兴趣标签关联规则挖掘,得到“机器学习 $\Rightarrow$ 数据挖掘”的兴趣规则,将“数据挖掘”作为目标用户的隐性兴趣预测标签加入用户画像。

## 3.3 用户画像构建关键问题

用户画像构建需要关注时效性、颗粒度、隐私保护等问题。

### 3.3.1 用户画像的时效性

构建图书馆用户画像需要考虑画像的时效性。用户画像对于时效性很敏感,某一时刻的用户画像对该

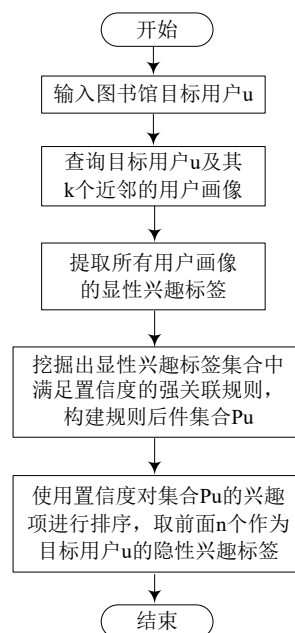


图2 基于关联规则的图书馆用户画像隐性兴趣标签预测流程

时刻的个性化服务最为有效。用户画像的来源数据中,用户属性数据涉及人口统计特征,是静态的,相对稳定。但是,海量行为数据、社交数据等具有较强的动态性,用户的行为随时间持续增加,新行为会使现有用户画像失效。此外,用户会受环境、其他用户等各种不确定因素的影响而改变偏好、兴趣和需求,这就需要用户画像随之改变以适应用户新状况。用户画像并非一成不变,而是实时动态变化的,具有时效性。图书馆要提供精准的个性化服务,需要设计合理有效的用户画像动态更新机制,以准确刻画用户的当前状态。

### 3.3.2 用户画像的颗粒度

构建图书馆用户画像需要考虑颗粒度,即画像细化程度。颗粒度小的用户画像对用户描述很精细,有利于提高图书馆个性化服务的精准性。但是,颗粒度越小意味着用户数据越细化,这会导致用户画像建模成本增加,也会降低用户画像适用性。例如,“机械工业出版社的机器学习书籍”和“机器学习书籍”两个阅读兴趣标签,前者颗粒度更小,但是只能代表某一特定出版社出版的机器学习类书籍,使得服务目标过于单一,后者颗粒度更大,但是适用性更好。需要根据图书馆具体业务需求选择合适的颗粒度,构建立体清晰且适用性强的用户画像。

### 3.3.3 用户画像的隐私保护



用户隐私保护是图书馆用户画像构建和应用中一个令人关注的问题。构建用户画像的过程中不可避免地要收集用户个人信息,在大数据环境下尤其如此。图书馆需要在“告知与同意”的隐私保护框架下,实施更加有效的措施加强用户隐私保护。用户画像中的用户隐私管理不仅需要技术方法和手段,也需要完善相关条例和法规。在技术上,保护用户敏感信息,保证用户隐私数据的安全,防范各种风险,如数据不适当公开、数据非法获取和使用、数据损坏或修改、数据丢失和泄露等。在用户画像建模算法中融入隐私保护技术,或者对用户画像信息划分等级,在不同级别的应用中使用相应等级的用户信息。在图书馆管理条例和法规中,规范图书馆在授权范围内对用户隐私数据的使用、超时销毁等行为,保障用户对敏感信息和隐私数据的控制权,最终在保障用户隐私的前提下构建出清晰有效的用户画像。

#### 4 基于用户画像的个性化服务探讨

随着经济的发展和科技的进步,人类进入了知识社会的新时代。知识的激增在促进社会进步的同时,也让人们迷失在信息和知识的海洋中,难以找到所需要的信息和知识,人们普遍面临着知识迷航和信息过载的困境。高校图书馆用户以学生、教师、科研人员为主,他们需要图书馆的个性化服务来帮助自己摆脱信息过载的困境。但是,没有对用户全面充分的了解,高校图书馆为用户提供个性化服务便有盲目性。有了用户画像,高校图书馆为用户提供个性化服务不再盲目,而是有据可依。通过用户画像展示的用户背景、兴趣、需求、活跃度等全貌信息,高校图书馆可以充分洞察用户,进而有针对性地开展以用户为中心、以满足用户需求为目标的高质量个性化服务,帮助用户摆脱信息过载的困境。

通过用户画像关联图书、论文、专利等各类纸质和电子馆藏资源,图书馆可以为用户提供符合其兴趣、需求的资源和服务。这里以个性化图书推荐为例,探讨用户画像在个性化服务中的应用。图3给出了

基于用户画像的用户-图书关联示意。一本图书与个性化服务相关的属性有:①作者。用户可能会喜爱几位特定作者的书籍。②类别。按中国图书馆分类法得到的图书类别,用户可能会喜爱某些类别或方向的图书。③出版社。用户可能会偏爱某些出版社的图书。④出版年份。用户可能会偏爱某些时期的图书。⑤媒介类型。用户可能会偏爱某些媒介类型的图书,如纸质媒介,或pdf、图像、音视频等格式的电子媒介。⑥语言。用户可能会偏爱某些语言的图书。除了以上6种属性,也可以从图书的内容描述或用户的评论等非结构化信息中提取与之相关的特征。通过用户画像的图书偏好和兴趣标签关联用户和馆藏书目库,为用户提供精准性的个性化图书推荐服务。

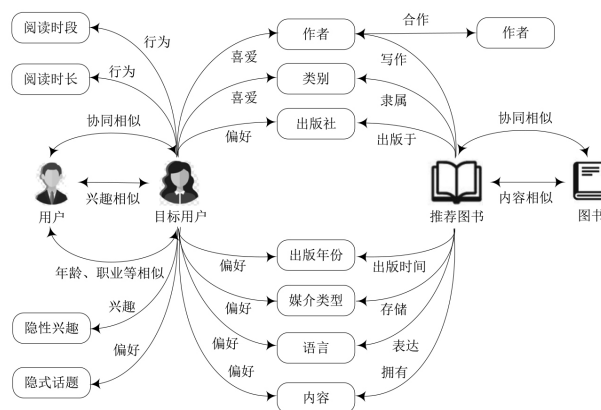


图3 基于用户画像的用户-图书关联示意图

#### 5 结语

在大数据时代背景下,高校图书馆需要充分利用大数据,开展基于大数据分析的个性化服务,更好地服务于知识传播、科技创新。本文首先提出基于用户画像的高校图书馆个性化服务模型框架,利用大数据挖掘技术分析和挖掘图书馆大数据,获取用户的全貌信息,构建用户画像,以用户画像为依据提供高质量的个性化服务,并探讨了基于大数据挖掘的用户画像构建,以及用户画像支持下的个性化服务。本文的研究对于应用用户画像提升高校图书馆服务水平有一定借鉴意义。

#### 参考文献:

[1] 柳益君,蔡秋茹,何胜,等.高校移动图书馆的场景化资源推荐服务:要素、模型和技术[J].图书馆学研究,2018(1):67-71.

(下转第59页)

- [5] 闫小芬.基于职业化视角的学科馆员职业专业性、公共性调查分析[J].图书情报工作,2017,61(17):76-83.
- [6] 李超,徐建华,霍丽敏,等.当代图书馆员“快乐指数”调查主报告[J].图书情报工作,2007(6):6-11.
- [7] 俞碧颀,徐建华.群体比较视角下图书馆员内隐刻板印象实验研究[J].中国图书馆学报,2014,40(5):52-67.
- [8] 徐建华,付娇.图书馆员的职业生涯开发与管理[J].中国图书馆学报,2003(1):18-21.
- [9] 王凤娥.高校图书馆员职业倦怠与干预模式分析[J].大学图书馆学报,2008(4):89-93.
- [10] 肖希明,吴钢,刘畅,等.图书馆学专业教育与图书馆员职业竞争力:来自图书馆馆长的调查与分析[J].图书情报知识,2008(1):11-16.
- [11] 肖希明,宋琳琳,水亮,等.图书馆学专业教育与图书馆员职业竞争力:来自图书馆员的调查与分析[J].图书情报知识,2008(1):17-21.
- [12] 鄂丽君,王启云.高校图书馆专业馆员职业能力认识与需求调查研究[J].图书与情报,2016(5):97-104.
- [13] 陶倩,刘泳洁.美国专业图书馆员职业能力标准的内容分析[J].图书情报工作,2016,60(24):20-25.
- [14] 王姗姗.高校图书馆员工招聘偏好分析:基于扎根理论的一个探究性研究[J].上海高校图书馆情报工作研究,2016,26(3):55-58,25.
- [15] Glaser B G, Strauss A L. The Discovery of Grounded Theory: Strategies for Qualitative Research[M]. New Jersey: Transaction Publisher, 2009: 45.
- [16] 孙晓娥.扎根理论在深度访谈研究中的实例探析[J].西安交通大学学报(社会科学版),2011,31(6):87-92.
- [17] 柯平,张文亮,李西宁,唐承秀.基于扎根理论的馆员对公共图书馆组织文化感知研究[J].中国图书馆学报,2014,40(3):37-49.
- [18] 王平,茹嘉祯.国内未成年人图书馆服务满意度影响因素:基于扎根理论的探索性研究[J].图书情报工作,2015,59(19):41-46.
- [19] 刘建花,杨蕙馨.消费者响应企业社会责任的内在机理与干预路径:基于扎根理论的探索性研究[J].现代财经(天津财经大学学报),2013,33(4):12-20.

**作者简介:** 杨俊丽(1976—),女,郑州大学图书馆研究馆员;陶雪然(1994—),女,郑州大学信息管理学院硕士研究生。

**收稿日期:** 2018-07-25

(上接第53页)

- [2] 陈军相.移动互联环境下高校图书馆读者行为分析[J].图书馆研究与工作,2017(6):92-95.
- [3] 刘春雷.基于数据分析的读者续借行为探讨:以浙江图书馆为例[J].图书馆研究与工作,2015(2):46-48.
- [4] 王向真.基于UTAUT的高校学生电子图书使用行为研究:以广西中医药大学为例[J].图书馆研究与工作,2018(11):82-88.
- [5] 柳益君,何胜,吴智勤,等.基于用户社交网络分析的高校图书馆主题多样性阅读推荐[J].图书情报工作,2018,62(8):67-73.
- [6] 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(6):105-110.
- [7] 郑海雁,王远方,熊政,等.标签集约束近似频繁模式的并行挖掘[J].计算机工程与应用,2015(9):135-141.
- [8] 周朴雄,陈蓓蓉.基于标签云系统共现分析的用户兴趣预测模型[J].情报杂志,2017,36(5):187-191.

**作者简介:** 陈丹(1980—),女,硕士,江苏理工学院计算机工程学院讲师,研究方向为智能信息处理、图书馆大数据应用;罗烨(1975—),女,江苏理工学院计算机工程学院讲师,研究方向为智能信息处理、图书馆大数据应用;吴智勤(1977—),女,江苏理工学院助理馆员,研究方向为图书馆信息服务。

**收稿日期:** 2018-12-02