

数字图书馆用户画像建模与应用实践^{*}

张洁¹ 仲跻亮¹ 岳怡然² 寇远涛^{1,3}

(1. 中国农业科学院农业信息研究所, 北京 100081; 2. 中国农业科学院研究生院, 北京 100081;

3. 农业农村部大数据重点实验室, 北京 100081)

摘要: 本文将用户画像引入数字图书馆各项服务的用户兴趣模型构建过程中, 在对比国内外用户画像概念基础上引申出数字图书馆领域用户画像概念, 通过国内外图书馆用户画像研究现状归纳出数字图书馆用户画像建设思路。从理论层面详细阐述模型设计、数据准备、数据挖掘与标签映射3项关键步骤, 之后遵循以上建设步骤以国家农业图书馆知识服务用户为研究对象开展用户画像建模及管理实践, 以期后续的研究探索提供参考。

关键词: 数字图书馆; 用户画像; 数据建模

中图分类号: G252

DOI: 10.3772/j.issn.1673-2286.2020.03.007

Edwards等^[1]通过研究发现, 1945年以后, 科研产出量每九年可翻一番, 此外计算机、通信、网络及存储技术的高速发展, 催生了科研产出数字出版的新业态。数字图书馆容纳的电子资源数量、类型和知识内容空前增长。海量资源衍生出知识冗余及知识迷航问题, 知识消费者的获得感低。新形势下, 通过对科研用户精细刻画, 实现用户需求与馆藏资源的精准匹配, 优化数字图书馆知识服务形式成为突出问题。用户画像作为数字化、虚拟化描述真实用户的技术手段, 可整合用户资源, 从动态增长的用户行为日志中挖掘用户的场景域、资源域及服务域需求。将其应用于数字图书馆领域用户建模, 一方面可充分释放馆藏资源价值, 促进图书馆各项服务增值; 另一方面, 可准确把握用户脉搏, 提升图书馆智能化、个性化服务水平。同时, 近年来用户画像在电商、智慧出行等智能信息服务领域的成功应用, 也为数字图书馆领域提供了相对成熟的技术应用经验及成功案例^[2]。

1 图书馆用户画像概述

1.1 概念界定

图书馆及信息学界对用户画像的概念界定目前尚不统一。用户画像这一概念最早源于交互设计/产品设计领域, 交互设计之父Cooper^[3]于2004年提出了用户画像概念, 并指出用户影像是真实用户的虚拟代表, 是建立在真实数据之上的目标用户模型。陈慧香等^[4]认为用户影像是建立在一系列真实数据之前的描述用户需求和偏好的目标用户模型, 该模型可全方位、立体化地反映用户特征。胡媛等^[5]认为数字图书馆将知识社区用户信息抽象化并运用聚类、关联规则及分类等数据挖掘方法汇制所得的用户可视化画像即为用户画像。陈冬玲等^[6]将用户画像称为“user profile”, 认为其是用户兴趣的描述文件, 是用户个性化需求的体现, 是个性化搜索的基础设施。总之, 由于总体设计思路及实现技术的不同, 不同学者对用户画像的理解各有侧重。

笔者引入互联网用户行为分析领域用户画像概念, 拟通过用户行为信息标签化以实现数字图书馆用户画

^{*}本研究得到中国农业科学院科技创新工程项目 (编号: CAAS-ASTIP-2016-AII)、中国工程科技知识中心建设子项目 (编号: CKCEST-2019-1-1) 和中国农业科学院农业信息研究所基本科研业务费青年探索项目 (编号: JBYW-AII-2019-21) 资助。

像的构建。笔者认为数字图书馆用户画像主要指面向真实读者用户,以用户的静态属性(人口统计特征、科研属性特征、空间和地理特征等)和动态属性(访问行为、资源检索及获取行为、学术社交行为、学术成果发表行为等)数据为基础,综合应用文本挖掘、机器学习等方法提炼出的具有显著特征的用户标签集合,该标签集合应该是关联、无歧义并且富含语义的。

1.2 国内外研究现状

以“图书馆用户画像”作为检索词搜索谷歌学术相关主题中文文献,得到800余条检索结果,发文时间在2010年之后。以“library user profile”作为检索词搜索谷歌学术外文文献,检索结果数达百万余条,最早文献发表时间可追溯至20世纪50年代。由此可见,国外相关研究起步较早,在理论及实践探索层面已相对成熟和完善,国内用户画像的研究在互联网产业的带动下开始成为热点,目前国内发文主要处于理论研究和前期探索阶段,实践层面研究成果相对较少。按照建模的数据对象来划分,用户画像包含基于用户行为及基于科研产出两类方法。

基于用户行为的画像构建方面,Leung等^[7]通过搜集搜索引擎日志中的正向与反向反馈为目标用户画像并完成聚类分析。国家图书馆在其大数据项目中通过汇总读者的注册、到馆、搜索、借阅等系列行为数据,搭建HadoopMap Reduce大数据管理与计算框架,构建了包括三级标签的读者画像^[8]。

基于科研产出的画像构建方面,美国加州圣玛丽学院图书馆研究并设计了PlumX管理工具,该工具以学者兴趣领域的科研产出为对象,构建可视化学者画像以响应本校科研管理战略^[9]。Gu等^[10]以学者的研究成果为分析对象,设计MagicFG算法,以出版成果数据为对象从中抽取学者基本信息,挖掘学者研究兴趣,并构建了Aminer研究者学术搜索网站。

综上可知,基于用户行为的建模方法受限于用户行为数据的离散性;基于科研产出的建模方法则更聚焦于学术兴趣,无法兼顾行为模式研究。笔者以国家农业图书馆各项知识资源内容及应用服务用户群体为研究对象,综合使用基于用户行为及兴趣偏好的方法开展学术用户的画像建模,以期从行为模式、使用场景及学术兴趣多维度刻画目标用户。

2 数字图书馆用户画像建模

数字图书馆用户画像建模是指面向各类数字图书馆服务场景,抽象用户描述标签体系,此外综合使用多种渠道获取可信用户数据集,选取数据挖掘模型及算法实现标签抽取与映射,支撑对各类用户的精准描述与可视化呈现。整体技术路线如图1所示,主要包括模型设计、数据准备、数据挖掘与标签映射3部分工作,用户画像可为开展画像可视化、资源评价、个性化推荐及精准推送等系列个性化服务提供支持。

2.1 用户画像模型设计

信息识别是用户画像构建的重要内容,其核心工作就是给用户贴“标签”,标签通常是高度凝练的用户特征标识,将所有的标签综合起来,就可以勾勒出该用户的画像。

根据数字图书馆业务特点,笔者将画像标签分为固定属性、访问环境、忠诚度及研究兴趣4类,共计16个维度,具体标签体系如表1所示。其中,固定属性是对用户基础特征的描述,该类标签主要用于识别用户身份,标签值可直接从用户注册信息或其成果署名信息中获取;访问环境类是对用户访问场景的描述,主要记录时间、地点、硬件设备及软件环境4个要素,这类标签一般需以多值字段形式来描述;忠诚度类描述科研用户对数字图书馆服务的黏性及认可度,通过访问频率、访问深度及距离上次访问时间3个标签值来体现;研究兴趣类是数字图书馆与其他领域建模不同之处的体现,该类标签描述用户的学术属性,从关注学科主题、资源类型、作者及机构多维度表征用户对科技知识资源的偏好。

2.2 用户数据准备

围绕用户画像标签体系的设计框架,搜集图书馆自身业务系统、三方业务系统等多种渠道的可信数据,以此数据集作为下一步数据标签与标签映射的对象语料。具体来说,用户数据准备主要包括数据获取及入库存储两部分工作。

用户画像基础数据集由用户静态基本属性、动态行为数据和科研成果数据3部分组成,以上3类数据均以结构化数据为主。其中,用户静态基本属性主要包括

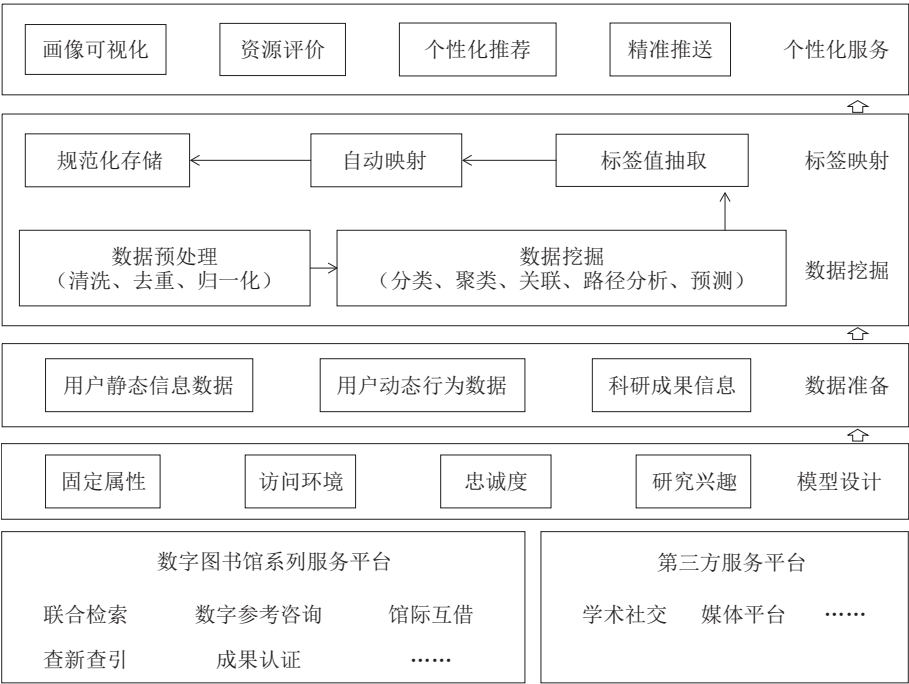


图1 用户画像建模技术路线

表1 数字图书馆用户画像标签体系

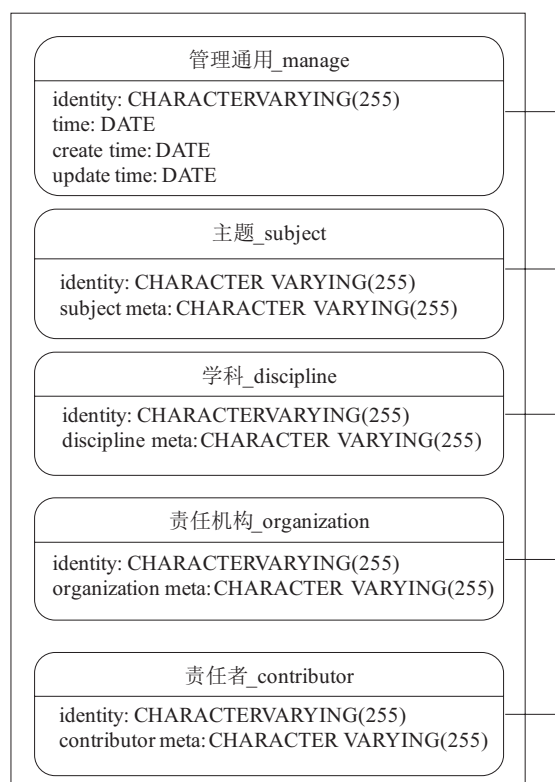
标签类别	标签名称
固定属性	用户ID
	电子邮箱
	姓名
	性别
	所属机构
访问环境	访问时段
	访问地点
	访问浏览器
	访问设备
忠诚度	访问频率
	访问深度
	距离上次访问时间
研究兴趣	研究领域
	兴趣资源
	兴趣作者
	兴趣机构

用户标识、姓名、电子邮箱、性别和工作机构等信息，这些信息相对较好采集，通常采用系统直接导入的方式。动态行为数据主要包括用户纸质与电子资源的查找、检索及借阅行为，项目立项的查新查引需求、学术社交

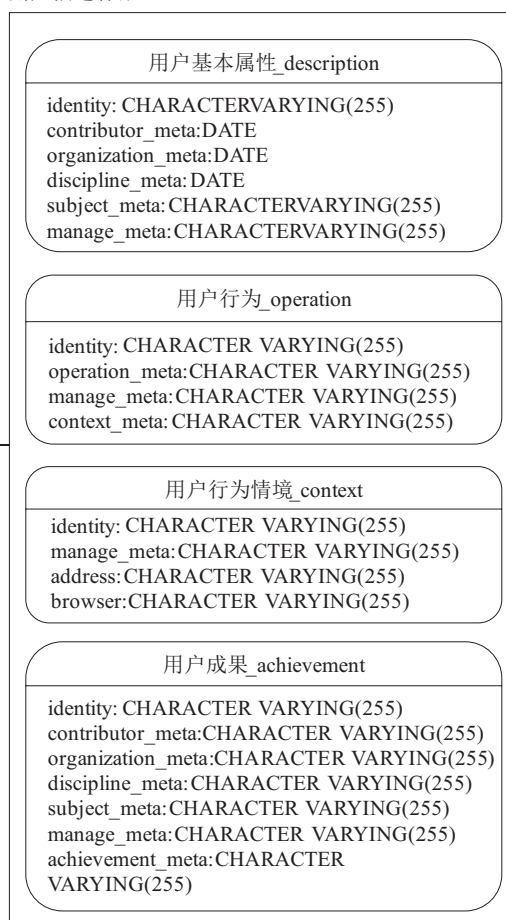
网站的互动行为等数据，这类数据较为分散，主要通过锁定信息来源后应用网络爬虫和日志记录技术进行提取。其中用户日志记录的采集主要包括WEB日志、JavaScript标记（代码埋点方式）和包嗅探器3种方式。相比而言，JavaScript标记方式收集数据灵活，可定制性强；可以记录缓存、代理服务器访问；对访问者行为追踪更为准确^[11]。科研成果数据主要包括用户作为科技创新主体的科研项目、论文、专利及获奖成果等各类成果描述信息，该类数据可从机构知识库及成果数据库中对应抽取。

对应数据类型特点及标签描述需要，预先为上述3类信息设计元数据描述与存储规范。图2展示了包括以上3类数据的数据关联描述模型^[12]，该模型设计了通用容器和用户描述容器两类数据描述集合，通用容器类主要包括管理通用、主题、学科、责任机构、责任者5类公共描述元素，用户描述容器类主要包括用户基本属性、用户行为、用户行为情景及用户成果4类用户描述元素。后者将在描述目标对象时直接引用通用容器中各类描述元素。遵循上述各类元数据描述规范，综合考虑数据管理工具的安全性及稳定性，选取合适的数据库管理工具并设计定时冷备份机制来完成原始数据从关系型数据库到大数据存储工具的备份。

通用容器



用户描述容器



is cited by

图2 用户数据关联描述模型

2.3 数据挖掘与标签映射

数据挖掘与标签映射阶段主要以用户描述模型为依据,设计标签挖掘计算模型及规则,从各类用户数据集中对应挖掘并抽取用户标签值,设计标签管理流程,实现标签值提取、规范化、标引及存储等系列操作,并支持个性化服务对各类画像标签的灵活调用。该管理流程主要包括标签值提取、自动映射及标准化存储3个关键步骤,见图3。具体来说,标签抽取是指按照标签值是否直接可见将用户描述标签分为两类,遵循对应数据模型并基于ETL工具实现标签值抽取。自动映射是指完成标签值的去重、合并、消歧归一等系列规范化处理并生成最终标签值,以实现自动化批量标引的过程。需要去重及合并处理的主要为访问浏览器、设备、访问时段、访问地点等多值类标签;需要消歧归一的主要为研究领域、兴趣作者及兴趣机构等可能存在同义词、中外文对照词及别名等多值类标签。标准化存

储规范是为了兼顾单值标签与多值标签的存储要求同时满足前端多项个性化服务模式对画像数据的灵活调用,设计了索引的存储规范,并选择以Solr、ES为代表的索引管理工具实现用户画像标签库的索引构建及调用响应。

4类标签中,固定属性类、访问环境类及忠诚度类标签大都属于显性标签,隐性标签则主要包括跨渠道用户标识、研究兴趣及兴趣实体的标签值确定,下面详细介绍以上3项隐性标签挖掘的实现思路。

(1) 跨渠道用户标识打通。数字图书馆用户在科技创新的全生命周期中会用到包括联合检索、参考咨询、馆际互借、查新查引及成果认证等多个图书馆服务平台,此外,这些用户也会使用包括Research Gate、LinkedIn等在内的第三方学术社交平台来跟踪国内外同行的最新研究和成果,因此用户画像的数据来源包括来自数字图书馆本地及第三方的多个平台,为实现对目标用户的数据化建模,需要集合多渠道用户行为

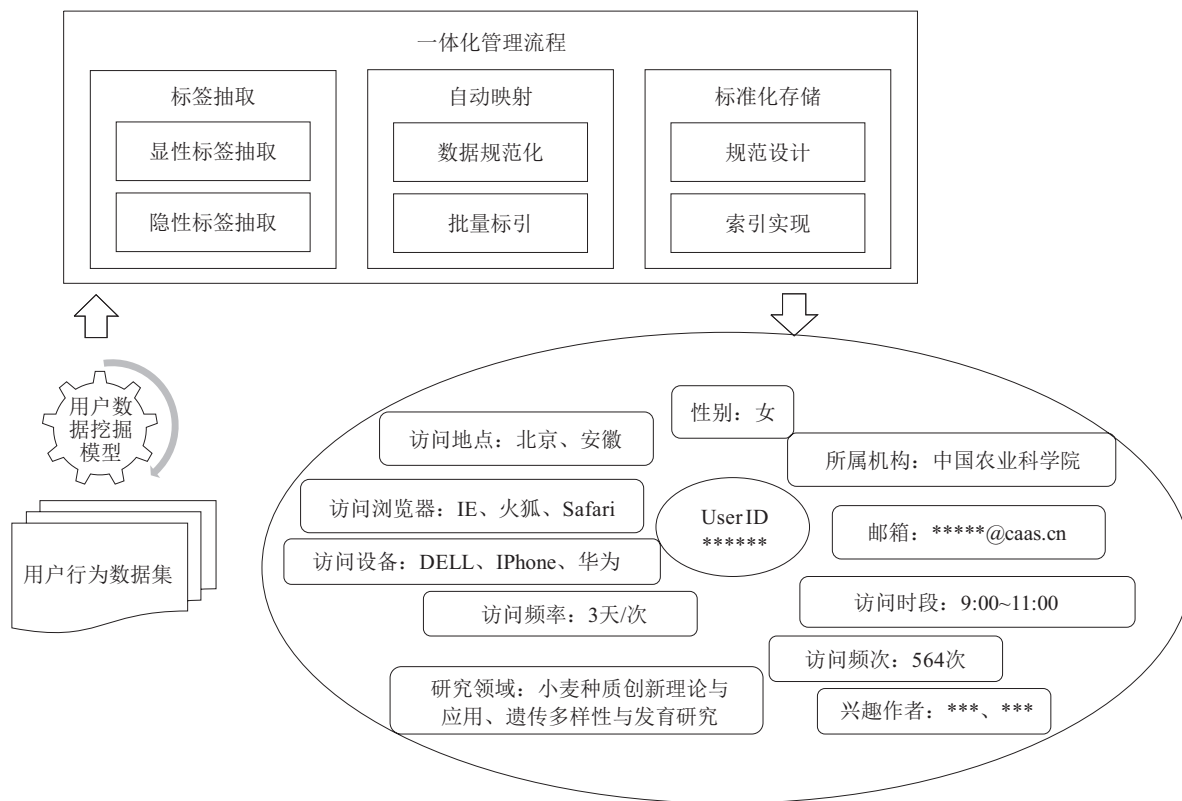


图3 标签映射管理流程

数据，完成标识间的打通串联，实现单用户跨系统用户行为的关联。目前跨渠道用户标识打通主要基于id-mapping算法，以包括MAC（Media Access Control）、Android ID、IDFA、手机号码及电子邮箱等终端访问及信息标识为关联依据，为不同访问途径下记录下了不同ID。基于ID间的共现关系，该算法将不同ID进行路径链接，这些相连路径则可被认定为同一位用户。

（2）研究领域识别。研究领域识别是指综合行为模式及科研成果，识别图书馆用户所关注的研究主题。因此，该过程可转化为对用户历史互动数据的文本集合进行主题挖掘，其中历史互动数据包括检索词、借阅书目及文献等。目前文本主题挖掘的实现方法按照是否需要先验知识可以分为文献计量及概率主题模型两类方法，前者以基于关键词的词频分析方法和共词分析方法为代表，后者以LDA、DMM、BTM、CTM等潜在主题信息挖掘方法为代表，此外随着词向量模型的应用优势，结合深度学习思想的概率主题模型也在近几年崭露头角^[13]。

（3）兴趣实体识别。同领域专家学者和专业机构也是用户在使用数字图书馆各项信息与知识服务过程中重点关注的命名实体类型。对于数字图书馆各项服

务来说，用户具有多角色属性，一方面是各类科技信息资源的消费者，另一方面作为专家学者也是各类科技信息资源的供应者。故此，可以从用户资源使用行为及成果发表行为两类数据中识别用户兴趣专家及机构标识。用户资源使用行为中，根据用户资源检索、查阅各类资源的描述文本，抽取责任作者、责任机构等信息，根据不同操作行为的质量权重，进行加权求和。根据求和结果降序排列，抽取规定阈值数目的作者及机构名单作为目标用户兴趣专家及机构标签值。用户成果发表行为中，抽取目标用户的合作发文作者及机构网络，将阈值范围内的合作专家及机构补充作为该用户的兴趣作者和兴趣机构标签值。

3 国家农业图书馆用户画像实践探索

国家农业图书馆研建了农业科技信息资源共建共享平台，该平台以整合知识检索及获取为核心，为农业及相关学科的科研主体提供知识资源发现及多渠道全文供给。笔者以该系统及其用户群体为对象，遵循第2章所述用户画像模型，完成用户行为数据准备工作，研发用户画像管理工具，该工具支持对用户画像的可视化

展示及标签化维护。

3.1 用户数据准备

通过对系统用户使用逻辑的分析梳理,笔者确定了该系统用户画像所需的基础数据体系,主要包括用户

基本属性、科研属性、访问行为、知识资源检索行为、知识资源获取行为及知识资源浏览行为6类信息,具体记录字段如图4所示。其中,右侧2类属于静态信息,可直接从用户注册信息表中获得;左侧4类属于动态信息,使用JavaScript标记方式实现对4类动态信息的记录及实时入库。

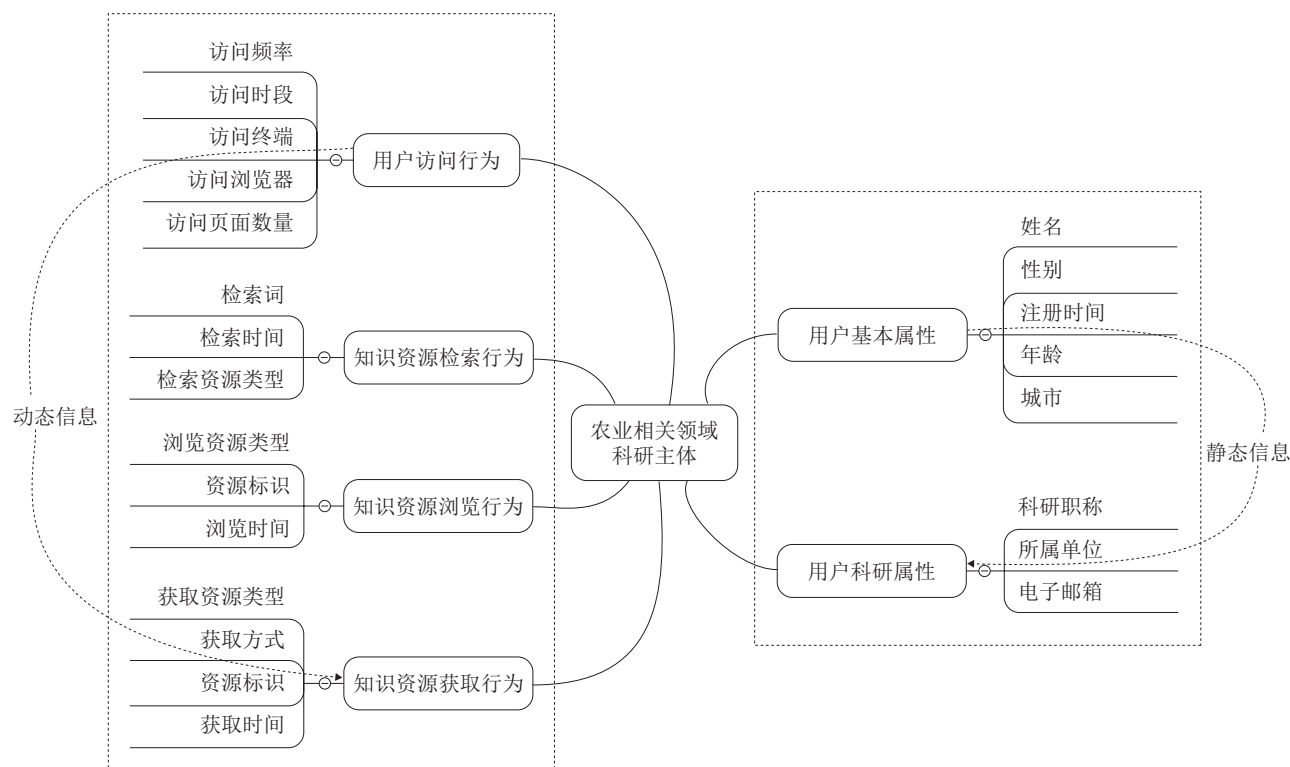


图4 农业领域知识服务用户行为数据体系

适应上述各类数据的来源及数据规范,设计数据实时传输、解析及入库规则,以结构化形式存储在数据表中,构建完成的用户行为数据集主要包括用户属性表、访问场景表、关键行为表,其中关键行为表又包含资源检索、资源浏览及资源获取3类子表。以资源检索为例,图5展示了资源检索行为中检索时间、检索词及资源类型等关键字段的记录代码及已记录数据示例。

3.2 画像管理实践

以农业科技信息资源共建共享平台用户行为数据集为基础语料,对应固定属性、访问环境、忠诚度和研究兴趣4类标签体系,完成对应属性值抽取及标注。为实现对数字图书馆用户画像的可视化及标签体系管理,笔者构建了用户画像管理工具,该工具为数字图书馆的用户运营管理提供综合看板、标签管理及用户画

像呈现等系列功能。

综合看板以雷达图形式集中展示所有用户的农业知识服务访问情况,并支持从PV、UV、搜索量、停留时间、下载量、注册时间等多个维度自定义排序筛选用户访问情况,页面示例如图6。

标签管理是通过标签组定义、标签增删改等功能提供对用户画像标签体系的维护及集中式管理。使用该管理功能,按照标签组添加、标签名添加、标签值管理的流程,实现农业科技信息资源共建共享平台用户的画像标签体系自定义维护与管理。

用户画像呈现是基于数据建模及可视化技术,实现对包括用户基本情况、综合访问表现、用户标签及历史搜索关键词的整合显示,以真实用户为例,使用画像管理工具对其画像数据进行可视化展示,页面效果见图7。

通过JavaScript完成资源检索行为记录

Input:

```
GlobSearchlogC c=new GlobSearchlogC();
    c.setUserid (userid);
    c.setUserip (req.getRemoteAddr());
    c.setSessionid (session.getId());
    c.setSearchtext (inputWord);
    c.setSource (language); c.setMemeberid(memberId);
    c.setCount (String.valueOf(count));
    .....
}
```

Output:

id	user_id	username	searchtext	sessionid	userip	source	tag	createime	memberid
21		湖南常德	湖南常德	F88F1389A4B	220.168.96.106	ch	default	2016-12-27 22:29:37.567	29
22		大豆育种	大豆育种	CE25606DAE4	124.17.34.58	en	default	2016-12-27 22:31:50.463	1
23		制梨白粉病分子机制的研究	制梨白粉病分子机制的研究	F88F1389A4B	220.168.96.106	ch	default	2016-12-27 22:38:50.487	29
24		制梨白粉病分子机制的研究	制梨白粉病分子机制的研究	F88F1389A4B	220.168.96.106	ch	default	2016-12-27 22:38:52.223	29
25		制梨白粉病分子机制的研究	制梨白粉病分子机制的研究	F88F1389A4B	220.168.96.106	ch	default	2016-12-27 22:38:55.973	29
26		制梨白粉病分子机制的研究	制梨白粉病分子机制的研究	F88F1389A4B	220.168.96.106	ch	default	2016-12-27 22:38:55.500	29
27		元数据	元数据	817D04374971	124.17.34.59	en	default	2016-12-27 22:45:23.153	[Null]
28		棉花目标价格改革成效及建议	棉花目标价格改革成效及建议	DE45C318120	211.101.136.177	en	default	2016-12-27 22:49:36.550	6
29		棉花目标价格改革成效及建议	棉花目标价格改革成效及建议	DE45C318120	211.101.136.177	ch	default	2016-12-27 22:49:45.583	6
30		棉花目标价格改革成效及建议	棉花目标价格改革成效及建议	DE45C318120	211.101.136.177	ch	default	2016-12-27 22:49:45.613	6
31		棉花目标价格改革成效及建议	棉花目标价格改革成效及建议	DE45C318120	211.101.136.177	ch	default	2016-12-27 22:49:47.317	6
32		棉花目标价格改革成效及建议	棉花目标价格改革成效及建议	DE45C318120	211.101.136.177	ch	default	2016-12-27 22:49:48.787	6
33		棉花目标价格改革成效及建议	棉花目标价格改革成效及建议	DE45C318120	211.101.136.177	ch	default	2016-12-27 22:49:50.130	6

图5 农业科技信息资源共建共享平台用户行为数据记录及存储(示例)



图6 用户画像综合看板页面效果



图7 用户画像详情页面效果

4 结语

用户画像为数字图书馆科研用户的数字化建模提供了有效解决途径,可有力支持数字图书馆各项知识

资源的采购、编目及组织工作从粗放型运营逐渐过渡到精细化运营,并在此基础上为包括信息过滤、知识服务设计等系列个性化知识服务模式提供基础数据及决策支撑。

国家农业图书馆引入用户画像思想,结合业务需求构建了用户画像理论模型,基于已有用户基础初步实践了用户标签化建模,构建了用户画像管理工具,支持对标签体系的维护及画像的可视化呈现,为数字图书馆用户精细化管理提供抓手。然而标签批量标引及隐性标签挖掘实现等方面的研究相对薄弱,下一步应针对性地开展相应工作,一方面提升用户描述标签的准确性,另一方面提升用户画像管理工具的易用性。

参考文献

- [1] EDWARDS M A, ROY S. Academic research in the 21st century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition [J]. Environmental engineering science, 2017, 34 (1): 51-61.
- [2] 张锐. 基于动态精准画像的图书馆个性化推荐服务研究 [J]. 情报探索, 2019, 256 (2): 102-105.
- [3] COOPER A. The Inmates are Running the Asylum: Why High-Tech Products Drive Us Crazy and How to Restore the Sanity [M]. Indianapolis: Sams Publishing, 2004.
- [4] 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.
- [5] 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.
- [6] 陈冬玲, 王大玲, 于戈. 支持个性化检索的User Profile研究综述[J]. 小型微型计算机系统, 2008, 29 (10): 1903-1907.
- [7] LEUNG K W, LEE D L. Deriving concept-based user profiles from search engine logs [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22 (7): 969-982.
- [8] 杨帆. 画像分析为基础的图书馆大数据实践——以国家图书馆大数据项目为例[J]. 图书馆论坛, 2019, 39 (2): 58-64.
- [9] WONG E Y, VITAL S M. PLUMX: a tool to showcase academic profile and distinction [J]. Ocle Systems & Services, 2017, 33 (4): 305-313.
- [10] GU X T, YANG H, TANG J, et al. Profiling web users using big data [J]. Social Network Analysis and Mining, 2018, 8 (1): 24.
- [11] zolalad. 网站分析数据(即用户行为数据)的三种收集方式详解 [EB/OL]. [2019-07-15]. <https://blog.csdn.net/zolalad/article/details/37809165>.
- [12] 赵瑞雪, 鲜国建, 罗婷婷, 等. 中国工程科技知识中心元数据规范(1) [M]. 北京: 中国农业科学技术出版社, 2017: 16.
- [13] 黄佳佳, 李鹏伟, 彭敏, 等. 基于深度学习的主题模型研究 [J/OL]. 计算机学报: 1-30 [2019-11-29]. <http://kns.cnki.net/kcms/detail/11.1826.TP.20191030.1633.004.html>.

作者简介

张洁, 女, 1991年生, 硕士, 馆员, 并列第一作者, 研究方向: 数字图书馆构建关键技术研究。
仲跻亮, 男, 1980年生, 硕士, 助理研究员, 并列第一作者, 研究方向: 信息系统和数字图书馆关键技术研究。
岳怡然, 女, 1996年生, 硕士研究生, 研究方向: 用户画像体系及其应用场景构建研究。
寇远涛, 男, 1982年生, 博士, 副研究馆员, 通信作者, 研究方向: 信息系统和数字图书馆关键技术研究, E-mail: kouyuantao@caas.cn。

Digital Library User Profile Modeling and Application

ZHANG Jie¹ ZHONG JiLiang¹ YUE YiRan² KOU YuanTao^{1,3}

(1. Agricultural Information Institute of Chinese Academy of Agricultural Sciences, Beijing 100081, China; 2. Graduate School of Chinese Academy of Agricultural Sciences, Beijing 100081, China; 3. Key Laboratory of Big Agri-data of Ministry of Agriculture and Rural Areas, Beijing 100081, China)

Abstract: This paper introduces user profile into the digital library user interests modeling. Based on the comparison of domestic and foreign user profile concepts, it gives out the concept of user profile in the field of digital library. Through the analysis of application status at home and abroad, it summarizes construction route of digital library user profile. The three key steps of model design, data preparation, data mining and label mapping are elaborated theoretically. Then taking the National Agricultural Library knowledge service user as an example, this paper carries out profile modeling and management practices in order to provide reference ideas for subsequent research and exploration.

Keywords: Digital Library; User Profile; Data Modeling

(收稿日期: 2020-03-13)