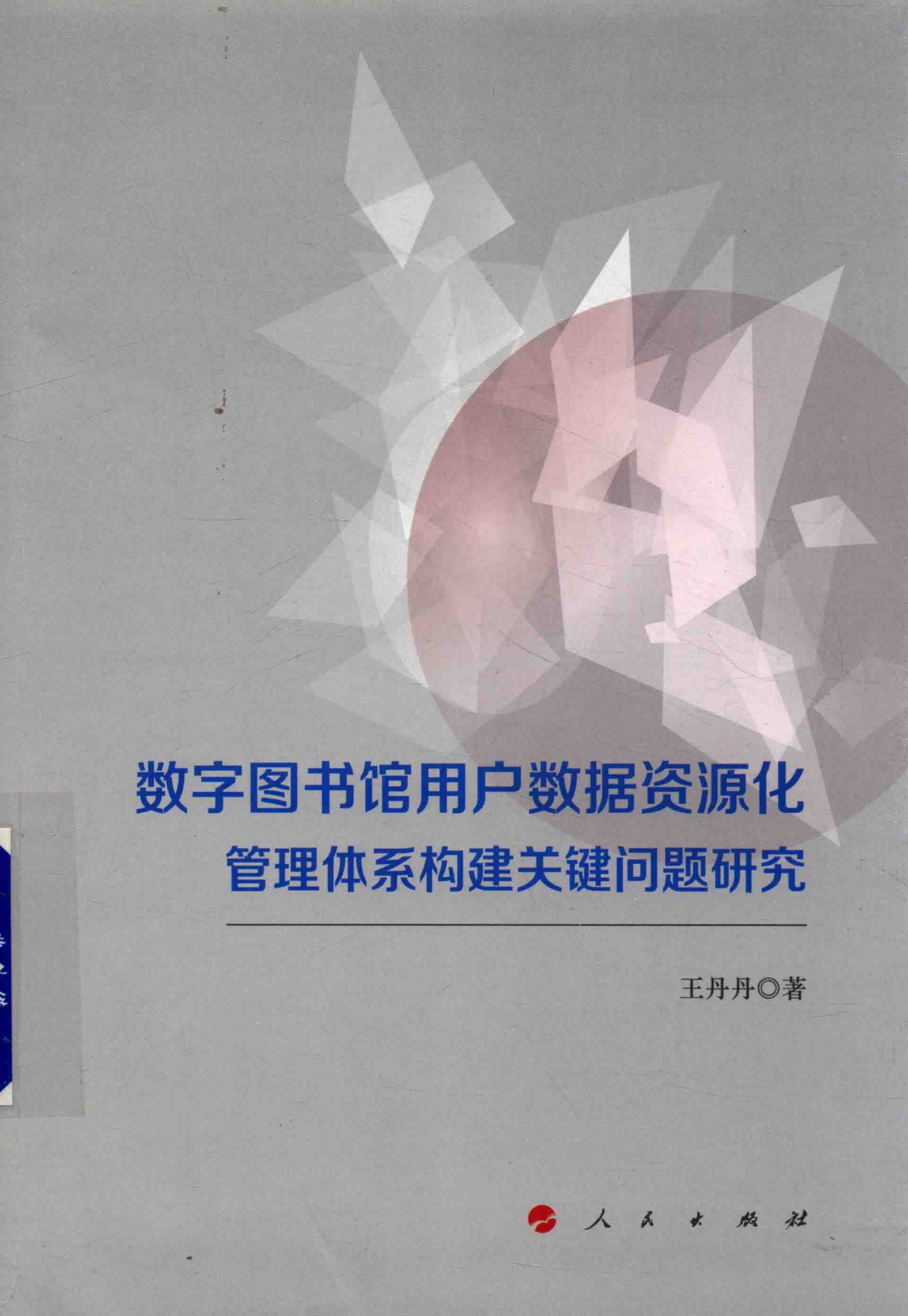




数字图书馆用户数据资源化 管理体系构建关键问题研究

王丹丹◎著



人民 出 版 社

数字图书馆用户数据资源化 管理体系构建关键问题研究

王丹丹◎著

责任编辑:李椒元

装帧设计:肖 辉

责任校对:吕 飞

图书在版编目(CIP)数据

数字图书馆用户数据资源化管理体系构建关键问题研究/王丹丹 著. —北京:
人民出版社,2018. 12

ISBN 978-7-01-019432-5

I. ①数… II. ①王… III. ①数字图书馆-信息管理-研究 IV. ①G250.76

中国版本图书馆 CIP 数据核字(2018)第 124294 号

数字图书馆用户数据资源化管理体系构建关键问题研究

SHUZI TUSHUGUAN YONGHU SHUJU ZIYUANHUA GUANLI TAXI GOUJIAN GUANJIAN WENTI YANJIU

王丹丹 著

人民出版社 出版发行

(100706 北京市东城区隆福寺街 99 号)

河北文盛印刷有限公司印刷 新华书店经销

2018 年 12 月第 1 版 2018 年 12 月北京第 1 次印刷

开本:710 毫米×1000 毫米 1/16 印张:20

字数:292 千字 印数:0,001-3,000 册

ISBN 978-7-01-019432-5 定价:42.00 元

邮购地址 100706 北京市东城区隆福寺街 99 号

人民东方图书销售中心 电话 (010)65250042 65289539

版权所有·侵权必究

凡购买本社图书,如有印制质量问题,我社负责调换。

服务电话:(010)65250042

国家社会科学研究基金项目

前 言

数字图书馆用户数据一方面指用户在使用数字图书馆网站和应用系统的过程中,被记录下来的反映用户访问状况、注意力或记录用户具体行为活动的的数据。另一方面,用户数据也指科研人员为研究的需要而通过各种渠道、各种方法有意识地收集的数字图书馆用户研究数据。用户数据是用户研究的基础、用户需求分析的依据,也是服务改进的源泉。用户数据资源化管理是将用户数据作为一种新型发展资源,通过对用户数据进行资源化管理,使之不断积累、整合、集成、有序和完善,成为一种服务资源进行开发利用。

对数字图书馆而言,通过有目的地采集用户数据,有目标地深度组织、多维关联用户数据,可以揭示用户科研和学习的兴趣、习惯、意图等,实现对用户信息行为和需求分析和把握,继而将其延伸到文献信息资源的组织、关联和服务过程中,实现个性化的内容服务,人性化的体验服务,智能化的知识服务,推进数字图书馆服务质量和水平不断提升。与此同时,也与 20 世纪后期以来,学术界发起了一系列旨在克服传统科学弊端,凸显“自由、开放、合作、共享”理念,与传统科学文化的封闭性形成鲜明对比的开放科学运动的趋势相呼应,促使馆员顺利实现开放学术环境下,角色和身份的转变。因此,实现用户数据资源化建设与管理,已成为数据密集型科学环境下数字图书馆提升效率和效益的一项重要研究课题。

国内外相关研究可分为两个层次,基础层次主要关注用户数据管理与建设涉及的技术问题,应用层次则主要关注用户数据功效与作用发挥方案制订问题。现有用户数据管理主要侧重于用户信息描述和获取、用户行为辅助分析,集中于用户日志和用户使用统计以及用户建模研究,尚缺乏多来源、多类

2 数字图书馆用户数据资源化管理体系构建关键问题研究

型、多属性用户数据的资源化管理框架研究,缺乏用户数据的资源化应用功能探索,系统化的数字图书馆用户研究方法梳理,用户数据没有形成数字图书馆借以高效利用的资源体系,用户研究的引领指导作用没有较好体现出来。2010年11月,笔者有幸参与国家科技图书文献中心(National Science and Technology Digital Library, NSTL)的用户研究课题。该课题的主要任务是分析NSTL的用户信息需求。然而,在研究的过程中却发现,NSTL尽管积累了大量的用户数据,但是挖掘这些数据所提供的真正有价值的信息并不多,数据保存的投入和产出严重不成比例。那么,这些数据究竟应该如何保存和使用,才能发挥最大效用呢?带着这样的思考,2013年6月成功申请到国家社科基金青年项目——“数字图书馆用户数据资源化管理研究”(13CTQ012)。课题组一边进行用户数据的全方位调研、梳理与分析,一边思考和总结用户数据资源化管理所需要解决的各方面问题,力图理论结合实际,实践提升为理论。

本书是国家社科基金项目的研究成果,对国内外用户数据管理与应用实践近些年创新探索的系统梳理和客观总结。从理论和实践两个方面,系统总结和深入剖析了数字图书馆开展用户数据资源化建设需要面对和解决的关键性问题,在立足现实的基础上,前瞻性地构建了数字图书馆用户数据资源建设的基本框架,并通过大量的实例论证了不同情境中,不同类型用户数据资源化管理的不同特点和不同发展模式,突出理论与实践相结合这一研究特色,特别是提供方法上的指导和最佳实践的参考线索。全部研究内容由用户数据资源化管理框架、资源化采集方法、资源化组织与共享以及资源化管理的实践四大部分,共十二个章节构成。

用户数据资源化管理框架。从数字图书馆用户数据应用领域和数字图书馆用户研究方法两个角度入手,对数字图书馆用于服务和决策的用户数据来源、特征、用途以及优劣势进行分析。以此为基础分析数字图书馆用户数据资源建设的关键问题,并提出数字图书馆用户数据资源化建设的原理和任务。涉及1—3章的内容。

用户数据资源化采集方法。用户数据来源的复杂性决定了用户数据采集的方案必然是多样化的。为了阐释如何实现用户数据资源化采集的问题,本部分首先对数字图书馆用户数据资源的类型进行了划分。然后,以此为依据,

从外部供应商提供的用户使用统计数据的采集、机构层面以及联盟层面用户数据的资源化采集三个维度,选取代表性案例,揭示不同情境下、不同类型用户数据的资源化采集方法及应用情景。涉及4—7章的内容。

用户数据资源化组织与共享。基于用户数据资源化采集的分析结论,对数字图书馆用户数据资源化组织与共享问题进行探索。提出了实现用户数据资源化共享的FAIR原则,指出了资源化共享面临的难题,并结合现有的实践对不同难题的解决方案进行探讨。考虑到用户研究数据这一重要且特殊的资源类型,特别对可以实现用户研究数据资源化组织和共享的现有技术平台以及用户研究数据资源化建设工作的具体推进策略进行了分析。涉及8—10章的内容。

用户数据资源化管理的实践。以国家科技图书中心(NSTL)为例,分析其用户数据资源现状,提出NSTL用户本体的构建方案,结合NSTL系统用户数据现状提出改善举措。最后,综合前面的研究成果,从用户研究数据,构建数字图书馆用户研究方法库的角度,提出了NSTL引领联盟层面用户数据资源化建设的构想。涉及11章的内容。

由于数字图书馆用户数据来源多样、类型复杂,因此这样一种研究本身具有很强的探索性,难度也是很大的。由于时间和能力所限,我们虽竭尽全力,但本课题研究仍有许多不足,与期望仍存在一定差距,期盼读者和广大的学科同行斧正。但愿我们的研究成果能抛砖引玉,成为系统开展数字图书馆用户数据资源化管理研究的起点,成为推进国内数字图书馆用户服务实践的有益参考,共同推动对数字图书馆发展影响长远的用户服务事业,共同推动数字图书馆用户服务水平 and 能力的不断提升。

目 录

前 言	1
绪 论	1
第 1 章 数字图书馆用户数据应用领域分析	14
1.1 数字图书馆用户数据应用项目介绍	14
1.2 数字图书馆用户数据主要应用领域	22
1.3 用户数据在个性化推荐中的应用	27
第 2 章 数字图书馆用户研究方法分析	39
2.1 数字图书馆用户研究整体状况回顾	39
2.2 数字图书馆用户研究数据收集方法	41
2.3 研究数据管理服务用户需求识别方法	53
第 3 章 数字图书馆用户数据资源建设要素与综合监管框架	65
3.1 应用领域分析对用户数据资源建设的启示	65
3.2 研究方法分析对用户数据资源建设的启示	70
3.3 数字图书馆用户数据资源建设要素梳理	75
3.4 数字图书馆用户数据资源化管理的框架设计	79
第 4 章 数字图书馆用户数据资源的类型分析	87
4.1 基于获取形式的用户数据分类	87
4.2 基于业务类型的用户数据分类	95

第5章	数字图书馆外部供应商提供的用户数据的采集	100
5.1	数字图书馆评价电子资源的方法	100
5.2	使用数据统计规范的出现与发展	102
5.3	自动实现使用数据采集的平台工具分析	104
5.4	采集外部供应商提供的用户数据的新要求	115
第6章	数字图书馆机构层面的用户数据资源采集	120
6.1	数字图书馆的数据认识误区	121
6.2	数字图书馆数据应用的困境	125
6.3	机构层面用户数据有效采集与应用的典型代表	127
6.4	机构层面面向资源化应用的数据采集案例分析	130
第7章	数字图书馆联盟层面的用户数据资源采集	145
7.1	构建实践社区促进用户研究数据的资源化采集	145
7.2	联盟层面用户相关数据的资源化采集与应用	157
7.3	定性数据的资源化采集与应用	167
第8章	数字图书馆用户数据的资源化组织与共享	173
8.1	数据资源化共享的指导原则	173
8.2	用户数据资源化组织与共享面临的难题	178
8.3	用户数据资源化共享与隐私保护矛盾的解决方案	179
8.4	用户数据资源化共享中访问权限与信任问题的解决方案	186
第9章	可用于用户研究数据资源化管理的现有技术平台	197
9.1	数据平台基本功能与利益相关群体	198
9.2	数据平台的用户访谈与使用测试	202
9.3	现有技术平台调研结果的总结与思考	207
第10章	机构研究数据资源化建设工作的具体推进策略	208
10.1	政策制定	209
10.2	基础设施部署	211
10.3	服务提供与能力建设	213

10.4 对用户研究数据资源化建设工作的启示	218
第 11 章 NSTL 用户数据资源化管理方案	221
11.1 NSTL 的用户数据资源概况	221
11.2 构建 NSTL 用户本体的思考	224
11.3 NSTL 系统用户数据现状与改善举措	240
11.4 构建 NSTL 用户研究数据资源体系的构想	259
第 12 章 研究结论与展望	266
12.1 主要成果	266
12.2 主要结论	269
12.3 创新与特色、不足与展望	274
参考文献	277

绪 论

数字图书馆用户数据一方面指用户在使用数字图书馆网站和应用系统的过程中,被记录下来的反映用户访问状况、注意力或记录用户具体行为活动的数据。另一方面,用户数据也指科研人员为研究的需要而通过各种渠道、各种方法有意识地收集的数字图书馆用户研究数据。用户数据是用户研究的基础、用户需求分析的依据,也是服务改进的源泉。用户数据资源化管理是将用户数据作为一种新型发展资源,通过对用户数据进行资源化管理,使之不断积累、整合、集成、有序和完善,成为一种资源进行开发利用。

0.1 研究背景与意义

0.1.1 研究背景

0.1.1.1 需求因素的拉动

在日趋激烈的竞争环境下,数字图书馆要想生存并具有竞争力,需要始终以用户为出发点和落脚点。然而,当前的事实却是,存在于数字图书馆不同的服务和应用系统中的各种数据,很难被用于充分地描述、刻画用户,了解其需求,数据作为资源的价值没有充分体现。主要原因在于:1)用户数据分散。用户数据分散在图书馆业务的各个模块中,通过这些零散的数据无法去全面地了解用户,致使各部门难以在统一的信息基础上为用户提供服务;2)信息不准确。很少有图书馆设立专门的数据管理部门,针对用户数据,建设采集、存储、处理和输出数据的有效且能及时更新的数据知识库,用户数据的完整性和准确性得不到保证;3)大多数图书馆员缺乏数据管理的意识和进行有效数据管

理所需的知识和技能,也不知道从何入手。鉴于此,有效转变用户数据粗放式、无意识收集与管理的状态,对这些数据进行了有意识的收集、存储和监管,使数据使用者能够相对容易的发现、获取、互操作和使用这些数据,成为数字图书馆实现数据驱动的决策的前提。

0.1.1.2 技术因素的推动

计算机网络技术的日益成熟,网络检索系统的服务器日志可以记录用户与系统交互的整个过程,包括访问时间、所输入的检索词、点击的检索结果以及点击事件、移动用户的终端设备信息等。近年来,随着数据采集设备的逐步普及,数据采集成本逐年下降,原始数据呈现出爆发式增长趋势。与此同时,科学研究活动开始围绕着研究数据进行,科学研究范式进入了不同于第一范式实验科学、第二范式理论科学、第三范式模拟科学的以数据为核心的数据密集型科学研究范式时代,被称之为第四范式^①。这一科研范式下,数据不仅是科学研究的成果,同时也是科学研究的对象和工具,科研人员基于数据来探索、设计和实施科学研究。在这种大环境下,作为社会知识服务机构的图书馆,如何借助研究数据管理的发展趋势,利用现有的技术、平台和最佳实践,实现图书馆研究数据的共享、重用,实现与同行之间的相互学习,更好地利用用户数据分析、评估图书馆产生的经济效益与图书馆的影响力,在充分了解自身优势及存在不足的基础上,变革服务模式,将服务优化调整为适合用户需求的方式具有重要意义。

0.1.1.3 图书馆管理变革的需要

个人获取和分享信息的方式在过去十年中发生了巨大变化,并将继续改变。为了使数字图书馆成为这一环境的一部分,图书馆员必须了解个人如何以及为何分享他们的信息。馆员不能指望人们改变他们的行为方式,以利用数字图书馆的虚拟和实体服务。为了吸引用户使用数字图书馆,数字图书馆需要在用户居住、学习和工作的环境中嵌入其服务。而当前的现实是,人们往往将图书馆与书籍联系起来,没有认识到图书馆与在线资源或参考服务有关。

^① Hey T, Tansley S, Tolle K. The fourth paradigm: data - intensive scientific discovery. Washington: Microsoft Research Publishing Company, 2009: 1-13.

人们可能不会想到使用图书馆来获取他们的信息,因为他们不知道服务存在,而且现有的一些服务他们不熟悉或不适合他们的工作流程。用户可以有許多不同的方式来参与信息环境,物理和数字图书馆只是信息搜索者可用的许多选项之一。数字图书馆的资源通常不是科研人员的首选,科研人员经常选择更方便,更易于使用的开放获取渠道。因此,作为传统信息中心的图书馆的地位被不断弱化,图书馆传统的价值和社会地位受到不同程度的质疑。图书馆应该如何向公众证明其服务和存在的价值,吸引资助机构的投入和支持,成为当下不得不考虑的现实问题。

0.1.2 研究意义

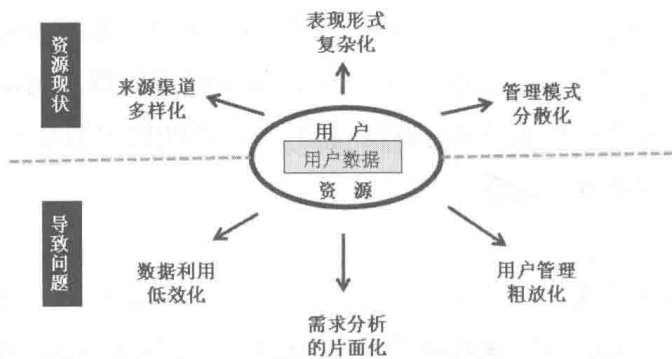
数字图书馆大量用户数据处于零散、无序,甚至流失状态,需要对其进行资源化建设和管理,将用户数据提到一种战略资源的高度,放在图书馆转型发展层面,构建形成数字图书馆用户数据资源体系。本研究针对数字图书馆用户数据来源渠道多样化,表现形式复杂化和管理模式分散化这一现状,以及由此所导致的用户数据利用低效化、需求分析片面化和用户管理粗放化等现实问题,以用户数据应用为出发点,系统总结数字图书馆用户数据资源化管理的相关理论、方法,提炼数字图书馆用户数据资源建设的主要思想、目标、内容、模式、机制和方法。

首先厘清数字图书馆多种来源渠道且表现形式复杂的用户数据类型和特点,应用领域及存在的问题;其次,以此为基础,在用户数据的资源化管理原理探寻基础上,分析用户数据资源化管理的目标任务,构建用户数据资源化管理框架体系,分析不同类型数据的获取方法和集成管理方式;然后,探索用户数据资源化采集方法、资源化组织方式和资源化共享机制等关键问题,最后,提出用户数据资源化应用模式,并进行实证研究。其意义体现在两个方面:

理论意义:用户数据资源化管理是用户研究走向科学化的基础,对数字图书馆用户数据的本质特征、关联关系、保存策略和应用场景进行研究,构建数字图书馆用户数据资源化管理框架,有利于丰富图书馆用户研究的理论体系,创新用户数据驱动决策的服务模式。

应用价值:为数字图书馆思考如何采集、整合、关联和利用用户数据提供

方法论指导和最佳实践的启发,指导数字图书馆尝试有意识的整合孤立、零散的多来源用户数据信息,形成用户数据知识库,构建借以高效利用的用户数据资源的体系,为数字图书馆各种决策提供证据数据库。



0.2 国内外研究现状评述

国内外相关研究可分为两个层次,基础层次主要关注用户数据管理与建设涉及的技术问题,应用层次则主要关注用户数据功效与作用发挥的方案制订问题。

0.2.1.1 用户数据标准化处理与获取

数字图书馆用户数据被不同的图书馆、内容提供商和系统集成商所掌控,既有原始数据又有统计数据,既有普查数据又有抽样数据。国内外科研人员针对这些不同来源的用户数据,研究了其标准化处理与自动获取方法。如推行实施使用数据统计 COUNTER 标准,实现使用统计数据自动收割的 SUSHI 协议,以及帮助图书馆收集用户使用统计数据的自动化工具, ScholarlyStas、Usta、360 Counter 等^①。现在虽然已经出现了大量的用户自适应系统,然而由于用户自适应系统之间缺乏互操作性和同步,造成在应用程序和领域之间重

^① Robin A.Paynter.Commercial library decision support systems:an analysis based on collection managers'needs.Collection Management,2008,34(1):31-47.

复构建用户模型的工作。因此, Carmagnola F 等认为下一代用户模型应该是可互操作的, 即能够交换用户模型部分并且使用已被交换的信息来丰富用户体验。为此, 概述了有关用户模型互操作性的一些解决方案。总结了一组维度, 用于描述工作分类的用户模型互操作性过程^①。Ma FC 等提出了提出一种在大数据环境中改善用户需求信息组织的概念模型——需求网络模型, 用于组织和互连用户需求。该模型包含数据层、语义层和应用层, 与现有方法相比, 该模型满足了大数据环境中用户需求信息组织的要求, 提供了四个方面的要点, 即为用户需求信息的智能处理提供依据和条件, 使用 RDF 作为描述规范, 实现用户需求信息的互联和设定各种协议, 保护用户隐私^②。赵洁针对图书馆本地原始数据, 提出一种多粒度用户行为数据收集方法, 以可配置的插件形式嵌入服务器端收集数据^③。朱庆华等针对用户产生的数据, 设计了采集规则和元数据方案^④。胡昌平基于用户数据管理的需求研究信息资源的异构整合机理^⑤。于长锐等人提出了本体化的 Web 信息、基础日志数据与用户识别和追踪相结合的方法^⑥。

0.2.1.2 用户数据语义组织与集成整合

不同特征的用户数据间具有丰富的语义联系。Aquin Md 等使用 RDF 整合日志数据, 使用 SPARQL 和轻量级本体对整合的结果进行推理, 实现以用户为中心的分析方法^⑦。参考咨询数字仓储项目证明了对参考咨询用户服务

① Carmagnola F, Cena F, Gena C. User model interoperability: a survey. *User Modeling and User-Adapted Interaction*, 2011, 21(3): 285-331.

② Feicheng Ma, Ye Chen, Yiming Zhao. Research on the organization of user needs information in the big data environment. *The Electronic Library*, 2017, 35(1): 36-49.

③ 赵洁、董振宁、张沙清等:《一种多粒度 Web 使用数据收集方法》,《现代图书情报技术》2011 年第 2 期。

④ 赵宇翔、吴克文、朱庆华:《基于 IPP 视角的用户生成内容特征与机理的实证研究》,《情报学报》2011 年第 3 期。

⑤ 胡昌平、汪会玲:《信息资源的异构整合及其对平台建设的要求》,《中国图书馆学报》2006 年第 3 期。

⑥ 石季辉、于长锐、刘兰娟:《网络社区用户兴趣挖掘的数据预处理研究》,《情报理论与实践》2010 年第 9 期。

⑦ Aquin Md, Elahi S, et al. Semantic technologies to support the user-centric analysis of activity data[EB/OL]. [2017-05-03]. <http://oro.open.ac.uk/29914/>.

数据进行集成化管理的可行性^①。宾夕法尼亚大学图书馆的 DATA FARM 项目,提出数据整合/处理的 MetriDoc 框架,实现了数据提取、转化、读取、查询的一体化^②。用户档案通常被用于提高可用性以及支持个性化、适应性和其他以用户为中心的功能。到目前为止,应用设计师主要以特殊方式模拟用户档案文件,从而阻碍用户档案层面的应用互操作性,增加了待完成的工作量以及档案文件模型中出现错误或遗漏的可能性。为此,Golemati M 等创建一个用户档案文件本体,其中包含用于建模用户档案文件的概念和属性。已经考虑了与用户情景和分析领域相关的研究成果、应用和本体,以便创建一个通用,全面和可扩展的用户模型。本体可作为参考模型,以减轻上述问题。该模型可供下载,应用于个人信息管理和自适应可视化两个不同领域^③。Dolog P 等人针对 E-Learning 环境,设计了一种学习者模型,通过建立学习者本体和领域本体,实现不同系统对学习者的共同理解。然后将聚合后的用户模型存储在用户模型服务器上,引入 Web 服务技术以支持不同客户端对同一个用户模型的访问,并通过 Java API 支持对用户模型的检索、插入和更新。与此同时,不同服务间也可以通过 API 进行用户模型的交换^④。Liu CH 指出用户本体应该包含静态和动态两个部分,引入了类“statements”和“validtime”来表达用户本体中动态的部分^⑤。

刘兹恒指出用户信息是开展信息集成服务的重要依据,用户信息应该共建共享^⑥。张太平探讨了用户信息的组织以及建立用户信息资源体系的问题

① Nicholson, Scott, Lankes, et al. The digital reference electronic warehouse (DREW) project: creating the infrastructure for digital reference research through a multi-disciplinary knowledge base [EB/OL]. [2016-02-17]. <http://bibliomining.com/nicholson/drewfinal.htm>.

② Collins P. Fear and loathing in cooperative collection development [J]. Interlending & Document Supply, 2012, 40(2): 100-104.

③ Golemati M, Katifori A, Vassilakis C, et al. Creating an ontology for the user profile: method and applications [EB/OL]. [2017-02-04].

④ Dolog P, Henze N, Nejd W, et al. Personalization in distributed e-learning environments [EB/OL]. [2016-02-17]. <http://www.www2004.org/proceedings/docs/2p170.pdf>.

⑤ Liu CH, Chen JJY. Mobile user agent with user ontology for personalized web service access [C]. Web Information Systems Engineering-WISE 2007 Workshops. Springer Berlin/Heidelberg, 2007: 325-336.

⑥ 刘兹恒、楼丽萍:《用户信息在图书馆工作中的应用》,《图书馆杂志》2002 年第 2 期。

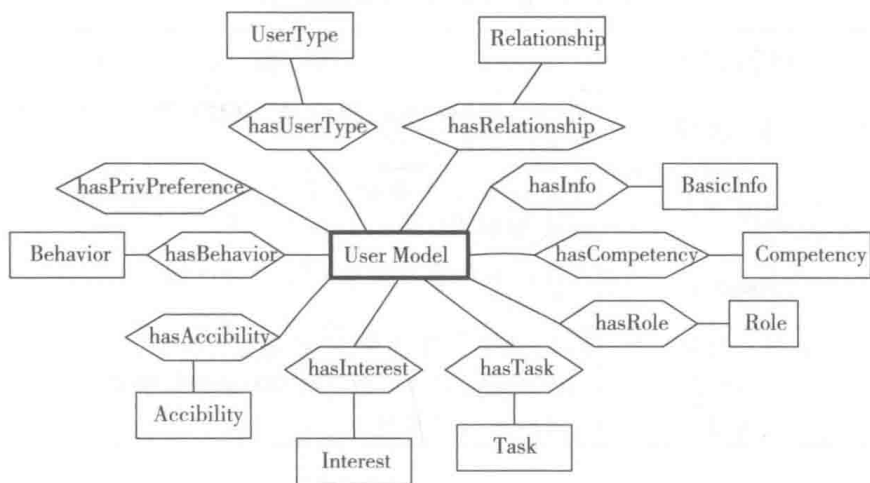


图 0-2 用户信息维度

题,包括实例、体系构架、数据库管理系统的建立等^①。朱静以汽车业供应链中客户信息共享为例,提出应用本体建模来解决汽车业供应链上异构信息系统中客户信息共享问题的方法,并设计了汽车供应链上客户信息本体模型^②。宋志正认为应该根据用户角色的分类构建多维角色模型,每个模型中都包含特定领域中的个性化服务系统所需的用户特征信息^③,如图 0-2。周倩提出构建图书馆通用用户本体的概念列表,如表 0-1,以及利用用户本体进行数据挖掘的设想^④。丁雪等人构建了用户本体,包括用户基本信息本体、用户个性本体和用户需求本体,并分析了本体间的相互作用机制^⑤。李书宁则系统抽象数字信息服务中用户服务的相关情景要素,讨论情景要素之间的关系,构建了用户服务情景本体^⑥。

① 张太平:《用户信息组织问题》,《图书情报工作》2002 年第 3 期。

② 朱静、吴家春、蒋馥:《汽车业供应链客户信息共享问题研究》,《情报科学》2005 年第 7 期。

③ 宋志正:《支持跨系统个性化服务的用户模型研究》,燕山大学 2007 年硕士学位论文。

④ 周倩:《基于 User-Ontology 的图书馆用户数据挖掘研究》,《图书馆杂志》2006 年第 10 期。

⑤ 丁雪、张玉峰:《基于本体的智能数字图书馆个性化推荐用户本体研究》,《现代情报》2009 年第 12 期。

⑥ 李书宁:《用户情景敏感数字信息服务若干关键技术研究》,中国科学院文献情报中心 2009 年博士学位论文。