NSTL 原文传递服务用户画像分析

蒋 君 王 超 张 玢

（中国医学科学院医学信息研究所/图书馆 北京 100005）

〔摘要〕 以中国医学科学院图书馆 NSTL 原文传递数据为例，采用文献计量、聚类、分类和序列分析等方法，从用户偏好、用户行为两个维度对原文传递用户构建用户画像，了解用户需求特点和规律，提出原文传递新方法和新途径。

〔关键词〕 原文传递；用户画像；图书馆；NSTL

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2019.11.017

Analysis on User Portraits of NSTL Original Text Delivery Service *JIANG Jun, WANG Chao, ZHANG Bin, Institute of Medical Information/Library, Chinese Academy of Medical Sciences, Beijing 100005, China*

〔**Abstract**〕 Take NSTL original text delivery data of Library of Chinese Academy of Medical Sciences as an example, with the methods of metrology, clustering, classification and sequence analysis of literature, the portraits of original text delivery users has been built from the two dimensions of user preference and behavior. Based on understanding the characteristics and rules of user needs, the paper proposes new methods and approaches for original text delivery.

〔**Keywords**〕 original text delivery; user portrait; library; NSTL

1 引言

随着信息时代的到来，跨领域多学科交叉研究不断深化，任何一家图书馆的馆藏资源都难以满足用户多样化信息需求，需要通过文献资源共享和文献传递的方式来解决[1]。用户画像技术能够较好地描述用户特征和信息需求，在用户和图书馆之间搭起交流桥梁，有利于驱动原文传递的创新发展。

用户画像以数据分析为工具，通过对用户属性、行为等方面的挖掘，了解并跟踪用户的需求变化，从而进行精准营销[2]。最早提出用户画像概念的是交互设计之父 A. Cooper，将其定义为基于用户真实数据的虚拟代表。Rebecca M. Quintana 将用户画像描述为一个从海量数据中获取、由用户信息构成的形象集合，通过这个集合可以描述用户偏好兴趣等个性化需求[3]。在图书情报领域，Amato G 认为信息提供者的最终目标是满足用户的信息需求，为用户定制用户画像[4]。Mao Jin 探讨基于标签的个性化推荐新方法[5]。王庆基于用户画像进行图书馆资源推荐模式设计与分析，为图书馆开展个性化服务提供新思路[6]。许鹏程在数据驱动下进行数字图书馆用户画像模型构建，以促进数字图书馆的知识服务升级[7]。陆尧针对区域图书馆联盟文献传递进行用户行为分析，提出改进意见[8]。本文在国家科技图书文献中心（National Science and Technology Library，NSTL）原文传递的基础上对用户画像进行

〔收稿日期〕 2019－09－03

〔作者简介〕 蒋君，硕士，馆员；通讯作者：张玢，副研究馆员。

分析，以便精准了解用户需求，实现资源服务内容精细化。

2 数据与方法

2.1 数据来源

中国医学科学院医学信息研究所/图书馆（以下简称医科院图书馆）为 NSTL 的医学分中心，面向全国科研单位提供医学类信息服务工作。医科院图书馆目前拥有医学及相关学科高质量数据库 91 个，电子期刊 16 300 余种，纸本期刊 4 500 余种，涵盖基础医学、临床医学、药学、公共卫生等医学各学科及化学、心理学等医学交叉学科。本文选取医科院图书馆 2018 年 1 月 1 日 – 12 月 31 日期间通过 NSTL 原文传递系统向全国医学科研机构提供的 80 866 篇原文传递文献为数据源。

2.2 研究方法

对 80 866 篇原文传递文献进行数据清洗和整理，采用文献计量学方法分析原文传递的语种、出版年等外在特征，聚类分析法计算出文献的领域特征，时序分析法分析用户申请的时间规律，从用户的行为信息和偏好兴趣两个维度对用户画像进行分析，以便优化资源建设，改进工作流程，提高工作效率和用户满意度。

2.3 分析框架

用户画像是一个长期逐步完善的过程，其目标是通过对用户行为、偏好等方面分析，给用户打上标签，以便精准快速分析用户行为习惯，为其提供个性化服务。NSTL 原文传递用户画像是在原文传递的基础上通过原文传递系统获得用户行为数据并进行预处理，形成规范化用户信息加以存储，然后对这些用户信息进行分类聚类等统计分析，勾勒出精确的用户画像，从而指导原文传递服务升级。用户画像分析框架，见图 1。

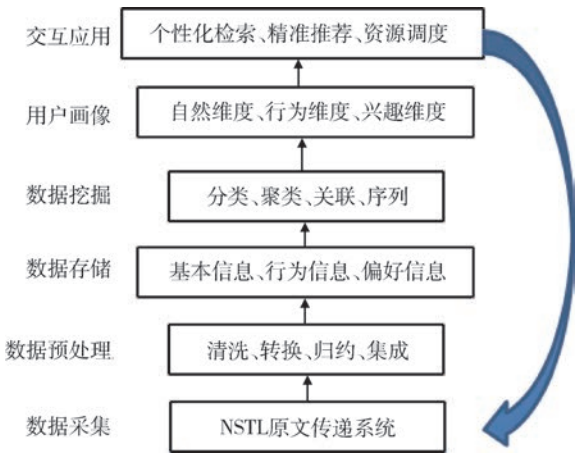

图 1 用户画像分析框架

3 用户画像构建

3.1 用户偏好

3.1.1 语种 2018 年共有 458 位用户通过 NSTL 原文传递服务平台向医科院图书馆申请原文传递服务，单个用户最高申请次数为 20 675 次。医科院图书馆发送原文传递文献 80 866 篇，去重后为 50 648 篇，其中单篇最高发送量为 81 次。将 80 866 篇原文传递文献按语种进行分类，见表 1。可以看出原文文献主要以外文文献为主，占 99.6%。同时还有 33 篇中文文献，全部是北京协和医学院的学位论文。分析其原因主要是：（1）医科院图书馆以外文文献为主，国外许多重要期刊是从创刊开始进行收录，覆盖范围广泛。（2）国际上多数医学期刊论文使用英文发表。（3）除使用英语的国家外，日本、法国、德国等国也有较好的医学专业和医学期刊，并且有些小语种（如匈牙利语、荷兰语、波兰语等）在国内是独家馆藏。（4）中文文献一般可以从中国知网或万方等网络数据库上查找全文，但北京协和医学院的硕博论文只能在图书馆获得。

表 1 原文传递文献语种分析

文献语种	文献量（篇）	百分比（%）	文献语种	文献量（篇）	百分比（%）
英语	78 993	97.68	中文	33	0.04
日语	667	0.82	匈牙利语	11	0.01
法语	468	0.58	荷兰语	9	0.01
德语	347	0.43	西班牙语	4	0.00
俄语	176	0.22	波兰语	1	0.00
意大利语	157	0.19	合计	80 866	–

3.1.2 类型 80 866 篇原文传递文献共分为 4 种类型，见表 2。在 4 种类型中期刊占绝大多数，其他 3 种类型只有少量，这与其自身特点有关：（1）期刊论文主要报道学术研究、学术创新点等成果，一般需要通过专家审稿，具有严谨性和连续性的特点，且医科院图书馆的外文医学期刊较为丰富，是医学研究人员首选。（2）会议论文是围绕某个会议主题在特定领域内的文章，是同领域内最新、最前沿的成果汇总，能及时反映学科发展趋向，有一定的参考价值[9]。（3）学位论文是作者为获得某种学位而撰写的研究报告或科学论文，具有一定独创性，参考文献多、全面，有助于对相关文献进行追踪检索[10]，并且北京协和医学院的学位论文是医科院图书馆的特色馆藏。（4）图书的内容比较系统、全面、成熟、可靠，但时效性不及其他类型文献，因此这类文献用户参考较少。对于这 4 种类型文献所包含的语种，期刊论文涉及语种较多，由除中文外的其他多语种文献组成，而学位论文仅包含中文文献，会议论文和丛书仅包含英文文献。

表 2 原文传递文献类型分析

文献类型	文献语种	文献量（篇）	百分比（%）
期刊论文	多语种	80 794	99.91
会议论文	英语	40	0.05
学位论文	中文	25	0.03
图书	英语	7	0.01
合计	–	80 866	100

3.1.3 年代 原文传递文献按出版年代分布，见图 2，可以看出：（1）1995 – 2018 年每年都有申请，基本上是年代越新申请量越大（2018 年除外），说明用户非常重视文献的时效性，希望获得最前沿的科技成果。（2）2011 – 2018 年的文献占比为 51%，超过半数，2015 – 2017 年这 3 年的文献需求量最多，其中 2015 年的文献超过 6 000 篇，说明近 3 年的文献是研究人员关注的重点。（3）2000 年之前的文献约占 5.5%，主要集中在《生殖医学杂志》（86 篇，影响因子 0.452，JCR 分区 Q4，妇产科）和《神经外科学杂志》（83 篇，影响因子 4.319，JCR 分区 Q1，临床神经病外科学）等期刊，说明这些医学期刊具有长尾效应，对现在仍有影响。期刊、会议和学位论文 3 种原文传递文献数量排名前 3，将这 3 种类型分别按年代进行排序，见图 3。期刊从 1995 – 2018 年都有使用，与总体趋势一样，年代越新使用量越大（2018 年除外）；会议论文重点关注前一年（2017 年）的文献，共计 21 篇；学位论文涉及 2010 – 2017 年 10 年的文献，且每年 2 ~ 3 篇，分布比较均匀。

图 2 原文传递文献年代分析

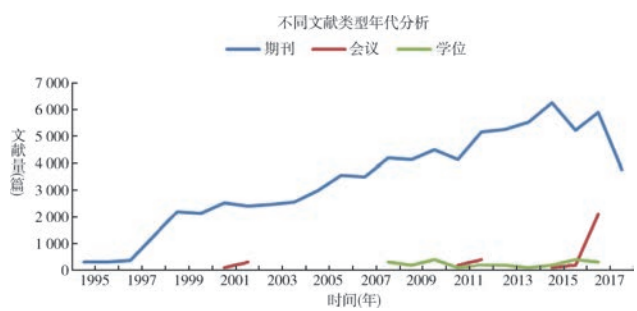

图 3 期刊类型文献年代分析

3.1.4 来源 原文传递文献共涉及 3 883 种来源文献，前 360 种期刊累计占比达 50%。按文献量倒序排列，选取排名前 20 位，累积占比 9%，见表 3。

总体来看：（1）排名前 20 位的来源文献申请次数都在 240 以上，最高达 558 次。（2）20 种来源文献的 JCR 分区，Q1、Q2、Q3、Q4 分别占 50%、10%、10% 和 25%，多数文献分布在第 1 个分区，说明申请文献的质量较高。（3）排名前 3 的是《国际病毒学杂志》、《肝脏与胃肠病学》和《白血病和淋巴瘤》，这 3 种期刊的 JCR 分区均位于 Q2 ~ Q4，说明用户相对期刊来说，更看中单篇文献的质量。（4）3 大顶级医学期刊《柳叶刀》（*Lancet*）、《新英格兰医学杂志》（*NEJM*）、《美国医学会杂志》（*JAMA*）分别位列第 4、11 和 16 位，影响因子较高，受到研究者的广泛关注。（5）还有一种期刊《印度医学会杂志》（第 8 位）未被收入 SCI 中，没有影响因子和 JCR 分区，但是文献传递量较高，说明用户关注印度相关的医学信息。

表 3 前 20 位原文传递文献来源分析

序号	来源文献	文献量（篇）	累积百分比（%）	影响因子	JCR 分区	序号	来源文献	文献量（篇）	累积百分比（%）	影响因子	JCR 分区
1	《国际病毒学杂志》	558	0.69%	2.936	Q2	11	《新英格兰医学杂志》	343	5.90%	79.26	Q1
2	《肝脏与胃肠病学》	549	1.37%	0.792	Q4	12	《抗癌研究》	341	6.33%	1.865	Q4
3	《白血病和淋巴瘤》	523	2.02%	2.644	Q2/Q3	13	《临床与实验妇产科》	336	6.74%	0.404	Q4
4	《柳叶刀》	500	2.64%	53.254	Q1	14	《国际皮肤病学杂志》	281	7.09%	1.541	Q3
5	《生殖医学杂志》	469	3.22%	0.452	Q4	15	《神经外科学杂志》	274	7.43%	4.319	Q1
6	《血液》	399	3.71%	15.132	Q1	16	《美国医学会杂志》	267	7.76%	47.661	Q1
7	《临床肿瘤学杂志》	367	4.16%	26.36	Q1	17	《神经外科》	257	8.08%	4.475	Q1
8	《印度医学会杂志》	363	4.61%	–	–	18	《临床精神病学杂志》	245	8.38%	4.247	Q1
9	《欧洲妇科肿瘤学杂志》	357	5.05%	0.617	Q4	19	《临床毒理学》	244	8.68%	4.381	Q1
10	《皮肤病药物杂志》	344	5.48%	1.527	Q3	20	《临床微生物学杂志》	241	8.98%	4.054	Q1

3.1.5 学科分类 将文献按《中国图书馆图书分类法》（以下简称中图法）进行整理，除去没有分类的 1 934 篇（暂归为其他）外，共涉及中图法 12 个大类，超过中图法大类的 50%，见图 4。其中 R 医药、卫生领域最多，约占 92%；其次是 Q 生物科学，占 4%；再次是 O 数理科学和化学、T 工业技术、N 自然科学总论、D 政治法律、S 农业科学等与医学相关学科；此外还包括 X 环境科学、G 文化科学、B 哲学宗教、C 社会科学总论和 P 天文学等边缘学科，表明这些学科与医学有交叉研究。在 2 级类目中，R73 肿瘤学、R9 药学的文献传递量最大，其次是 R6 外科学、R75 皮肤病学与性病学、R74 神经病学与精神病学等，由此得出这些领域是目前医学人员研究的重点。在非医药卫生领域，Q5 生物化学、Q2 细胞生物学、O6 化学等领域文献较多。

3.1.6 标题聚类 从文献标题入手，运用 Gephi 可视化关系网络分析软件对内容进行分析。首先将所有标题进行分词，去除没有意义的代词、介词、副词、量词等停用词，选取词频在 500 以上的词，对其进行统计和聚类，揭示词与词之间的关联关系，见图 5。通过分析可知这些文献主要聚为 4 类：以临床（clinical）为代表的粉色图标、以肌肉内（intramuscular）为代表的绿色图标、以治疗（treatment）为代表的橙色图标和以影响（effect）为代表的蓝色图标。粉色图标主要与临床、症、腺癌、肝脏、肺、肿瘤、分子、基因、血清等有关，代表腺癌、肝癌、肺癌等肿瘤在分子、基因和血清等方向的临床研究；绿色图标主要与肌肉、治疗管理、原发性、淋巴瘤、案例、外科、剂量、诊断等有关，代表原发性淋巴瘤、肌肉瘤等案例的诊断和手术；

橙色图标主要与治疗、注射、药物、慢性、急性、疾病、风险等有关，代表慢性或急性疾病的注射或药物治疗及其风险因素。蓝色图标主要说明相关影响。各类内关系密切，各类间也有相互关联，连接的粗细代表关联强 度。

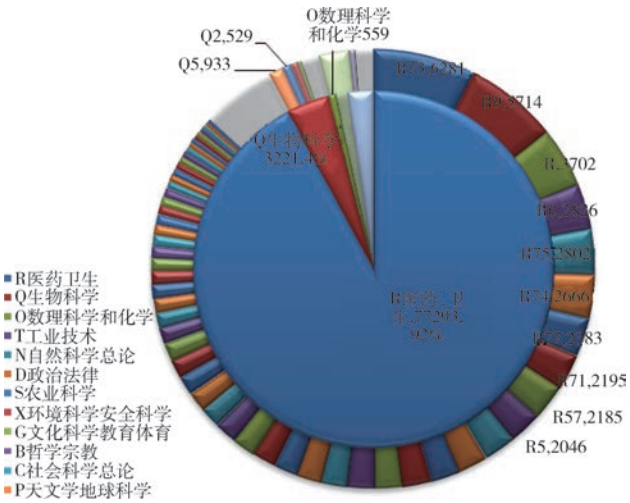

图 4 原文传递文献学科分类

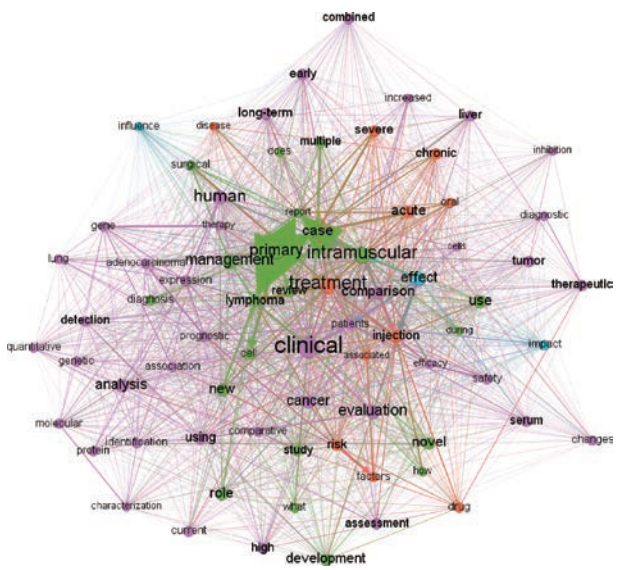

图 5 标题聚类分析

3.2 用户行为

提交申请时间不仅可以反映用户工作习惯，还便于医科院图书馆根据需求量大小更好地安排工作。以下从工作日、月份、周期和时段 4 个时间维度对用户提交时间进行分析。从工作日来看，周一到周五的原文传递提交量相对较多，周末较少。周三是用户提交需求的高峰期，周二和周四其次，两头最少，见图 6。从提交月份来看，10 月、11 月是最高峰，其次是 7 月和 9 月，然后是 1 月、3 月和 5 月，见图 7。这与申报奖项、课题和职称评定有一定关系。从提交周期来看，第 41 周（10 月 7 – 13 日）提交需求最多，其次是第 17 周（4 月 22 – 28 日）、37 周（9 月 9 – 15 日）和 48 周（11 月 25 日 – 12 月 1 日），基本上都在提交月份的高峰期内，见图 8。其中第 41 周与十一放假后需求积压反弹有关联。从提交时段来看，在 24 个时点中有 3 个高峰期：9 – 11 点是第 1 个高峰期，也是提交量最大的时间段，12 – 16 点是第 2 个高峰期，20 – 22 点又迎来一个小高峰，见图 9。从用户原文传递申请的全年时间分布分析，总体来说用户在 10 月节后第 1 个星期和 4 月、11 月最后 1 个星期的周二到周四上午 9 – 11 点提交的申请最多，可以提前做好准备，根据需求量大小对工作人员进行相应调整，以便更好地为用户服务。

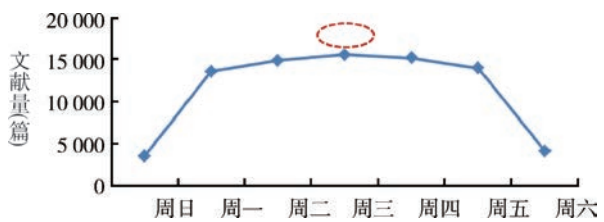

图 6 提交工作日分析

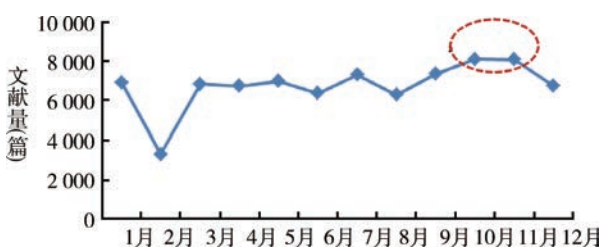

图 7 提交月份分析

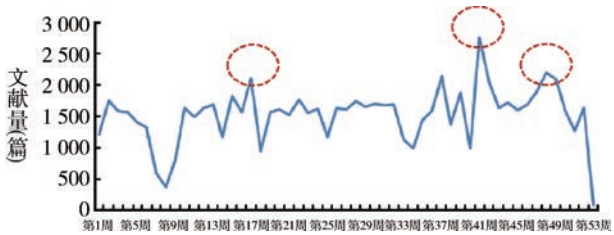

图 8 提交周期分析

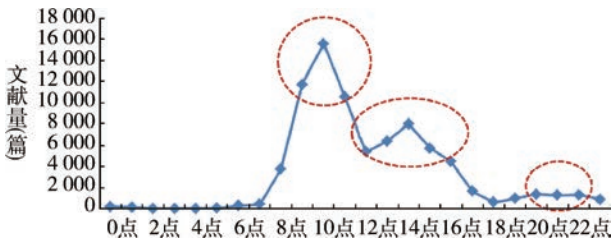

图 9 提交时段分析

4 启示与建议

4.1 关注用户需求，适时调整馆藏资源

从原文传递的分析结果可以看出大部分用户关注肿瘤学、药学、外科等领域资源及一些重点期刊，适当加强相应学科的资源建设将更好地满足用户需求。原文传递需求的学科分布将是加强针对性资源建设的参考，应定期向资源建设部反映馆内发送申请的情况，以便图书馆在购买新增资源时参考。

4.2 注重文献种类多样性，文献语种多样化

从用户需求特征看，期刊文献的需求量最大，但会议论文、学位论文和丛书也有需求，可能以后还包括标准、专利、科技报告等其他类型文献。从用户对文献语种的需求看，除英文外小语种文献也占有一定比例。为满足用户需求的多样化和个性化，建议文献采集时尽量扩充文献类型和语种。

4.3 拓展特种文献保障，完善馆藏资源揭示

原文传递的关键是找到用户需要的文献资源，这就意味着对文献要进行全面的揭示和完善的查询。目前图书馆中有少数馆藏资源只保存纸质版，尚没有进行数字化加工，难以实现统一揭示，尤其是一些珍贵的特藏文献，目前只有纸版保存。因此建议尽量实现图书馆的数字化处理，有利于文献的长久保存和有效利用。

4.4 预估工作强度，合理安排时间

根据用户提交申请时间可以推断出原文传递在每年的 10 月、11 月达到高峰期，在每天的 9 – 11 点是一个高峰时段，可以根据分析结果预估工作量，做好工作安排，快捷高效地为用户服务。

4.5 加强宣传推广，提高用户满意度

相对于馆藏 16 300 余种电子期刊和 4 500 余种纸本期刊，原文传递文献使用量相对较少。为使用户能够有效使用图书馆资源，应不断进行原文传递服务的宣传和推广工作。可以采用发放宣传手册、举办讲座、走进课堂或者以公众号的方式进行宣传，重点介绍图书馆馆藏资源和原文传递的使用方法，为用户提供参考。此外可以向注册和潜在用户发放调查问卷，收集相关需求，以便及时改进，更好地为用户服务。

5 结语

原文传递是数字时代传统图书馆开展主动服务的一种表现形式。本文通过分析原文传递数据，描述用户画像特征，建立以用户需求和满意度为出发点的原文传递服务形式，根据用户画像中的需求调整馆藏资源，注重文献种类的多样性，完善馆藏资源揭示，根据用户请求时间分布，更加合理地安排工作，加强宣传推广，最终提高用户满意度。

参考文献

1 百度百科. 原文传递 [EB/OL]. [2019 – 03 – 01]. https://baike.baidu.com/item/%E5%8E%9F%E6%96%87%E4%BC%A0%E9%80%92/4091664?fr=aladdin.

2 王凌霄，沈卓，李艳. 社会化问答社区用户画像构建 [J]. 情报理论与实践，2018，41 (1)：129 – 134.

(下转第 85 页)

特色化文献信息库的框架与模式。针对不同类型文献信息库的具体要求确定数据库结构，对参与文献信息库建设的人员进行技术培训，开展文献信息库内容的收集整理以及数字化加工工作。首先确立文献搜集的范围和检索策略，分别交由各临床科室或研究室的人员进行文献资料的搜集整理工作；其次通过信息采集系统对各种数据库及网页内容的解析和抓取，结合词表、自动识别技术，对采集数据内出现的内容实体进行自动识别和抽取并进行存储；最后通过数字化加工技术对已有内容资源的结构化拆分解析并将拆分的结构进行结构化存储，为不同文献类型数据库的建设奠定数据基础。将检索到的文献逐条分析，按类别进行标引、著录，导入到相应的文献信息库中，形成文献信息库的整体模型。

4.3 运行

将试点科室的文献信息库建设方案和成果逐步推广到其他临床科室及研究室，逐步建立系统、完善、全面反映基地临床及科研成果的文献信息系统应用平台。

5 结语

临床科研平台文献信息库基于广安门医院数字图书馆的平台，建立集综合检索、开放获取、学术分析、个性化服务于一体的中医药特色文献信息库及服务系统，为中医临床研究基地建设提供强有力的信息保障。

参考文献

1 马红敏，邓文萍，孙静. 国家中医临床研究基地标准信息管理系统研究与设计［J］. 中国数字医学，2014，9（10）：34－36.

2 卢传坚，陈淑慧，蔡桦杨. 国家中医临床研究基地科研创新平台设计初探——基于广东基地的实证研究［J］. 中国卫生事业管理，2016，33（5）：360－362.

3 赵爽. 医院信息系统安全分析与管理［J］. 医学信息学杂志，2018，39（11）：32－34.

4 刘红丽. “互联网＋”时代医学高校数字图书馆知识发现系统研究［J］. 医学信息学杂志，2017，38（5）：11－15.

5 艾金勇. 藏学文献特色数据库建设实践［J］. 智能计算机与应用，2016，6（4）：45－47.

6 于琦，崔蒙，李园白. 中医药文献数据库建设规范研究［J］. 世界科学技术—中医药现代化，2014，16（11）：2304－2307.

7 王静. 基于元数据异构共享的艺术院校图书馆特色数据库建设研究［J］. 图书馆学刊，2018，40（6）：53－57.

（上接第 80 页）

3 Quintana RM, Haley SR, Levick A, etal. The Persona Party: using personas to design for learning at scale［C］. New York: 2017 ACM SIGCHI Conference on Human Factors in Computing Systems , 2017.

4 Amato G, Straccia U. User Profile Modeling and Applications to Digital Libraries［C］. Berlin: 3rd European Conference on Research and Advanced Technology for Digital Libraries, 1999.

5 Mao J, Lu K, Li G, et al. Profiling Users with Tag Networks in Diffusion－based Personalized Recommendation［J］. Journal of Information Science, 2016, 42（5）: 711－722.

6 王庆，赵发珍. 基于“用户画像”的图书馆资源推荐模式设计与分析［J］. 现代情报，2018，38（3）：105－109，137.

7 许鹏程，毕强，张晗，等. 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，2019，63（3）：30－37.

8 陆尧，杨代庆. 区域图书馆联盟文献传递用户行为分析［J］. 图书馆论坛，2019，39（5）：88－94，126.

9 百度百科. 会议文献［EB/OL］.［2019－03－01］. https://baike.baidu.com/item/%E4%BC%9A%E8%AE%AE%E6%96%87%E7%8C%AE/10605294?fr=aladdin.

10 百度百科. 学位论文［EB/OL］.［2019－03－01］. https://baike.baidu.com/item/%E5%AD%A6%E4%BD%8D%E8%AE%BA%E6%96%87/4678889?fr=aladdin.