

东北财经大学图书馆用户借阅记录数据集*

李 荣 廖映红 李 双 (东北财经大学图书馆)

摘 要 利用汇文文献信息系统抓取了 2011–2020 年间东北财经大学图书馆用户借阅数据, 该数据包含了用户信息、借阅信息及图书信息。通过数据匹配和加密处理, 共获得 1 464 729 条记录。其中对用户信息 ID 进行了 UUID 重新编码处理, 保证了用户隐私。本次公开的数据集有助于高校图书馆准确掌握近年来文献资源利用情况、探寻用户群体借阅行为的变化趋势、挖掘用户阅读兴趣; 同时有助于为图书馆馆藏布局调整、资源与服务的有效推广及文献资源精准化服务提供决策支持。

关键词 借阅行为 借阅记录 数据集 东北财经大学

DOI: 10.13663/j.cnki.lj.2022.10.012

Dataset of Borrow and Return Records at Dongbei University of Finance and Economics Library

Li Rong, Liao Yinghong, Li Shuang (Dongbei University of Finance and Economics Library)

Abstract This paper uses Huiwen document information system to collect borrow and return records at Dongbei University of Finance and Economics Library from 2011 to 2020, which includes user information, borrowing information and book information. Through data matching and encryption processing, a total of 1,464,729 records were obtained. The UUID of user information is re-encoded to ensure user privacy. The dataset is helpful for university libraries to accurately grasp the utilization of literature resources in recent years, to explore the changing trend of user groups' borrowing behavior, and to tap users' interest in reading. At the same time, it is helpful for decision-making regarding collection layout adjustment, effective promotion of resources and services, and accurate service of literature resources.

Keywords Borrowing behavior, Borrow and return records, Dataset, Dongbei University of Finance and Economics

数据集基本信息

数据集中文名称	东北财经大学图书馆用户借阅记录数据集
数据集英文名称	Dataset of Borrow and Return Records at Dongbei University of Finance and Economics Library
数据作者	李荣; 廖映红; 李双
通信作者	李荣
作者单位	东北财经大学图书馆
版本号	V1.0

* 本文系辽宁省教育厅科学研究青年科技人才“育苗”项目“基于图书文本挖掘的用户借阅行为研究”(项目编号: LN2019Q24)和辽宁省教育厅科学基础研究项目“基于用户需求的纸质资源建设优化策略研究”(项目编号: LN2020J17)的研究成果之一。

通信作者: 李荣, E-mail: westsky2000@139.com

(续表)

数据集中文名称	东北财经大学图书馆用户借阅记录数据集
版本时间	20210126
基金项目类型 1	2019 年辽宁省教育厅项目
基金项目名称 1	基于图书文本挖掘的用户借阅行为研究
基金项目批准号 1	LN2019Q24
基金项目类型 2	2020 年辽宁省教育厅项目
基金项目名称 2	基于用户需求的纸质资源建设优化策略研究
基金项目批准号 2	LN2020J17
国家	中国
语种	中文
数据覆盖时间范围	20110101-20201231
地理区域	东北财经大学
经纬度	北纬 N38°53'4.36" 东经 E121°33'28.54"
数据格式	.csv
数据体量	数据记录 1 464 729 条, 共 590MB
关键词	借阅行为 借阅记录 数据集 东北财经大学
主题分类	研究数据; 科学数据; 图书情报学; 计量学
全球唯一标识符	hdl:20.500.12304/10566
网址	http://hdl.handle.net/20.500.12304/10566
数据集组成	共 5 个文件, 2011 年至 2020 年的数据(一个文件包含 2 年的数据)
使用条款	本数据集可以通过在网站实名注册登录获取, 可以用于科学研究和教学目的, 使用时需标注引用信息, 禁止二次分发和商业演绎

0 引言

随着自动化信息系统及数据库技术在图书馆的快速普及, 图书馆存储了大量用户的借阅数据。通过借阅数据对用户特征及借阅行为影响因素进行深入挖掘, 已经成为图书馆提升馆藏文献利用率、改进馆藏建设质量、创新文献个性化推荐服务的重要手段。本数据集通过东北财经大学图书馆文献信息管理系统采集了近十年图书馆用户借还记录, 数据完善规范, 在高校图书馆拥有一定代表性^[1]。本数据集可辅助开展文献资源利用评估、用户借阅行为分析、图书文本信息的智能发现, 为用户提供智能化、精准化的文献推荐服务和知识服务, 具有较高的研究和实践价值。

1 数据采集和处理方法

本数据集涉及数据均采集自汇文文献信息管理系统后台 Oracle 数据库。用户信息数据

存储在 READER 表中, 图书借还数据存储在 LEND_HIST 表中, 图书相关信息数据分别存储在 MARC 表、INDI_ACCT 表及 LOCATION 表中, 其均为结构化数据。

1.1 数据采集和匹配

准备一台安装 SQL Server 2005 关系型数据库的服务器专门用于数据采集和数据匹配。为保证借阅数据具有利用价值, 本数据集不仅采集了发生借阅行为的用户信息, 而且对借阅文献的主题、摘要、作者简介等详细信息进行了关联。通过 SQL Server 导入功能将 Oracle 数据库的 READER 表、READER_TYPE 表、LEND_HIST 表、MARC 表、DOC_TYPE_CODE 表及 LOCATION_LST 表同步到数据采集服务器的 SQL Server 2005 数据库中。通过 SQL 语句关联 READER 表和 LEND_HIST 表, 关联限制条件为 READER 表 CERT_ID 字段与 LEND_HIST 表 CERT_ID_F 字段关联;

READER表关联 REDR_TYPE_CODE表, 关联限制条件均为 REDR_TYPE_CODE 字段; LEND_HIST表关联 MARC表, 关联限制条件为 LEND_HIST表 MARC_REC_NO_F 字段与 MARC表的 MARC_REC_NO 字段关联; LEND_HIST表关联 INDI_ACCT表, 关联限制条件为 LEND_HIST表 PROP_NO_F 字段与 INDI_ACCT表的 PROP_NO 字段关联; MARC表关联 DOC_TYPE_CODE表, 关联限制条件均为 DOC_TYPE_CODE 字段; LEND_HIST表关联 LOCATION_LST表, 关联限制条件为 LEND_HIST表 LOCATION_F 字段 LOCATION_LST表的 LOCATION 字段。

由于文献信息中关联了主题、文摘(摘要)、作者简介等信息, 上述3部分内容存储在 MARC表的同一字段中, 不便于区分。因此在实际数据采集中, 我们利用汇文文献自动化的系统管理模块中 MARC 字段导出功能, 将 606、330、314 字段的相关含有主题、文摘及作者介绍等数据导出到文本文件中, 再将文本数据导入至 SQL Server 数据库中, 最后进行用户信息表、借阅记录信息表及文献信息表的关联匹配, 一共汇总到 1 464 729 条记录。按每两年记录存储一个数据文件, 表 1 为各起止年份记录数。

表 1 2011 年至 2020 年用户借阅记录数据统计

起止日期	记录数
2011 年 1 月 1 日至 2012 年 12 月 31 日	475 037
2013 年 1 月 1 日至 2014 年 12 月 31 日	393 266
2015 年 1 月 1 日至 2016 年 12 月 31 日	292 133
2017 年 1 月 1 日至 2018 年 12 月 31 日	205 039
2019 年 1 月 1 日至 2020 年 12 月 31 日	99 254
总计	1 464 729

1.2 数据清洗

借阅用户的信息采集了学生用户的学号及教工用户的职工号, 用户的学号及职工号信息属于学校师生个人敏感数据, 为了避免隐私数据泄露, 对数据进行了脱敏处理。首先给每条用户记录增加 UUID 编码, 实际上就是用 UUID 来代替用户 ID, 保证 ID 的唯一性。UUID 是通用唯一识别码 (Universally Unique

Identifier) 的缩写^[2], 在任何一台计算机或服务服务器上生成的数字, 在所有空间和时间上被视为唯一的。对用户记录 ID 重新编码后, 可以直接抹去用户 ID, 再用 UUID 与密钥拼接进行 MD5 加密, 可完全保证使用者安全利用数据。用户信息隐私处理后的效果见表 2。

汇文系统中的用户性别是来源于学校一卡通中心同步数据, 但大部分性别不准确。为尽可能保证性别的准确性, 利用用户信息的身份证 ID 字段, 来修正用户的性别信息, 修正完成后抹去身份证字段。但早期注册用户缺失身份证 ID 信息或性别数据的, SEX 字段统一置为 NULL 值。为了区分用户的年龄段, 根据身份证 ID 信息提取用户的出生年信息存入 BIRTHYEAR 字段。但有个别用户的身份证在原始数据录入时错误或存入了非身份证信息 (如军官证号、护照号、考号等), 都会导致提取的值存在异常, 例如出现 “1016” “1086” “2886” 及 “2 (3)” 等异常值。通过与原始数据进行比对, 将明显错误的 BIRTHYEAR 字段进行了修正, 但对于无法获取用户身份证信息的 BIRTHYEAR 字段赋予 NULL 值。年级组字段 (CODE01) 在数据同步时, 不同年份学校一卡通中心提供的字段值有所区别, 因此年级组字段根据用户注册年份来修正。出版年字段中含有字母和其他非数字字符的, 需要将字母和非数字字符清洗掉, 只保留年份。单位字段 (DEPT) 和专业字段 (OCCUPATION) 中含有括弧、空格及其他特殊字符也需要清洗。另外, 对于单位或专业的文字表述, 有的年份是全称, 有的年份是简称, 需要尽可能对同一单位或专业的数据进行名称统一化处理。用户字段数据均来源一卡通同步数据, 存在部分字段为空值的记录。

借阅历史表中财产号字段 (PROP_NO_F), 可以唯一识别借阅数据中每册文献。为了隐去原有系统的财产号信息, 我们仍然采用 UUID 编码的方法生成每册文献的唯一记录号信息。数据处理中首先对文献财产表 (INDI_ACCT) 的每册图书信息进行编码存入 BOOK_ID 字段, 再将 BOOK_ID 字段关联到图书借阅信息表中。

从汇文系统管理模块导出的借阅文献主题词存储在了不同的字段, 为了方便数据使用, 我们将每个主题词之间用分号隔开, 对主题词中含有“①”“-”及空格等特殊字符进行了清洗, 并合并到一个字段 (SUB)。对于汇文系统没有主题描述的, 均赋予空值。

以上数据清洗的过程在 SQL Server 2005 数据库查询编辑器中完成, 部分清洗语句如下:

生成 UUID 编码语句 (去掉字符中的“-”):

```
select *, replace (newId ( ), '-', '') as UUID  
into 用户表 1 from 原始用户表;
```

年级组字段修正语句:

```
update 用户表 1 set CODE01=substring  
(REDR_REG_DAY, 1, 4);
```

MD5 加密语句 (去掉字符中的“0x”):

```
select *, replace (sys.fn_sqlvarbasetostr  
(HashBytes ('MD5', '拼接字符串 1'+UUID+'  
拼接字符串 2')), '0x', '') USERID into 清洗完成  
用户表 from 用户表 1;
```

性别字段修正语句:

```
Select USERID, DEPT, OCCUPATION, REDR_  
REG_DAY, CODE01, REDR_TYPE_CODE,  
case  
when len (ID_CARD)=15 and cast (substring  
(ID_CARD, 15, 1) as int) % 2 = 0 then '女'  
when len (ID_CARD)=15 and cast (substring  
(ID_CARD, 15, 1) as int) % 2 = 1 then '男'  
when len (ID_CARD)=18 and cast (substring  
(ID_CARD, 17, 1) as int) % 2 = 0 then '女'  
when len (ID_CARD)=18 and cast (substring  
(ID_CARD, 17, 1) as int) % 2 = 1 then '男'  
when len (ID_CARD)=17 and cast (substring  
(ID_CARD, 17, 1) as int) % 2 = 0 then '女'  
when len (ID_CARD)=17 and cast (substring  
(ID_CARD, 17, 1) as int) % 2 = 1 then '男'  
else NULL end as SEX
```

into 用户表 2

FROM 用户表 1;

出版年字段修正语句:

```
UPDATE 文献信息表 SET PUB_YEAR=  
replace (replace (PUB_YEAR, 'c', ''), '.', '') where  
PUB_YEAR like 'c%';
```

主题词特殊字符处理和合并语句:

```
update 借阅历史表 set SUB = ISNULL  
(SUB1+'; ', '') + ISNULL (SUB2+'; ', '')  
+ ISNULL (SUB3+'; ', '');
```

```
update 借阅历史表 SET SUB = LEFT (SUB,  
len (SUB) - 1) where LEN (SUB) > 0;
```

表 2 数据脱敏处理样式

字段名	处理前 (UUID)	处理后 (MD5 加密)
USERID	FD7874B25D014728	72ee861be6baa73ad
	BBC84EB69A2905F9	d3420677aa019a6

2 数据字典、数据样本和数据量

数据集包含用户信息、借阅信息及文献信息三大部分, 其字段名、字段样例值及字段的备注信息见表 3。

3 数据质量控制

本数据集涉及所有数据均由汇文系统后台 Oracle 数据库和系统管理模块采集。为了保证数据质量, 通过汇文流通模块的流通日志和统计模块对近十年的用户借阅记录进行了核实, 流通借阅日志和统计模块的借阅统计数据均与从后台提取的借阅记录条数一致, 确保了数据采集准确无遗漏。另外选取多个用户 (不同用户类型), 从流通管理模块用户个人借阅历史记录查询, 并与本数据集比对, 数据信息内容均一致^[1]。最后通过汇文统计模块的书刊信息查询功能, 对本数据集的借阅文献的摘要、主题词、作者介绍等信息进行了抽查核对, 所采集信息均与 MARC 借阅文献信息一致, 确保了数据真实可靠。

4 数据价值

本数据集包含 2011–2020 年间东北财经大学图书馆用户在利用图书馆文献过程中产生的借阅行为数据, 数据字段较为丰富 (22 个), 对高校图书馆用户的文献利用评估、用户借阅行为分析和用户需求预测等具有重要意义。近十年用户借阅数据是对图书馆馆藏进行评价的数据基础, 通过对这些数据的分析, 不仅可全面评估现有馆藏结构及文献保障水平^[3], 而且

表 3 数据字典

字段名称	中文名称	样 例 值	备 注
USERID	用户 ID	72ee861be6baa73add3420677aa019a6	利用通用唯一识别码 (32 位) 编码再进行脱敏处理保证唯一性 空值
BIRTHYEAR	出生年份	1988	空值代表一卡通未提供
SEX	性别	女	空值代表一卡通未提供
DEPT	单位或院系	会计学院	空值代表一卡通未提供
OCCUPATION	专业	会计学	空值代表一卡通未提供
CODE01	年级组	2008	空值代表一卡通未提供
REDR_RYPE_NAME	用户类型名称	本科生	学生用户代表入校年份, 教工用户代表注册年份
LEND_DATE	借阅时间	2012-03-02 09:28:55	本数据集中一共涉及 13 种: 本科生、研究生、教工、教师、图书馆工作人员、离退休、非计划内学生、馆际互借、国际留学生、国内进修生、外籍教师、外聘国内专家、临时读者
RET_DATE	归还时间	2012-03-09 10:11:44	空值代表该文献未归还
RENEW_TIMES	续借次数	0	用户续借的次数
TITLE	题名	股利政策理论与实证研究	
BOOK_ID	图书 ID	2D026D3966F7484996FCBB4ACF3B73E6	图书唯一记录号
ABSTRACT	摘要 / 简介	本书对我国上市公司股利分配动因、支付形式、股利政策的特点及其形成原因等进行了较为深入的剖析, 并就我国上市公司如何制定股利政策提出了决策思路	空值代表原始 MARC 数据没有摘要信息
SUB	主题词	股份有限公司; 利润	空值代表原始 MARC 数据没有主题信息
CALL_NO	索书号	F279.246/18	
AUTHOR	作者信息	李常青著	
AU	作者简介	李常青 (1968—), 安徽省太湖县人, 厦门大学 MBA 教育中心副教授, 著有《股利政策理论与实证研究》	空值代表原始 MARC 数据没有作者简介信息
PUBLISHER	出版社	中国人民大学出版社	
ISBN	ISBN 号	7-300-03853-0	
PUB_YEAR	出版年	2001	
LOCATION_NAME	馆藏地	辅助书库三	
DOC_TYPE_NAME	文献类型	中文图书	

可从借阅时间变化趋势、不同用户类型借阅特征、热门借阅文献、借阅文献主题词热度等多维度出发,为高校图书馆文献资源建设创新服务、用户多元化服务、个性化推荐服务等提供科学的决策依据。

5 数据使用方法和建议

本数据集可采用 Tableau、Excel、SPSS、Python 语言及关系型数据库等软件工具进行统计分析^[1]。例如,针对年度热门借阅文献的特点进行可视化分析;根据近十年年度借阅量统计数据进行分析用户文献借阅量的变化趋势;根据用户的年龄、性别、单位、专业等属性来分析不同用户的借阅特征和借阅频率差异;通过时间序列分析借阅文献主题词热度在一年中的变化趋势;根据近十年的用户借阅记录进行统计性描述分析,选择目标用户群体和某大类图书作为研究对象,对图书的摘要简介和主题词进行分析,借助相关模型进行主题挖掘,将主

题化后的特征与用户的借阅行为特征进行融合分析,探索用户的借阅关系、兴趣变化趋势,设计特征融合的推荐模型,为用户提供更友好、更准确的图书推荐和导引,提高文献服务的效率^[4];可引入深度学习相关算法,将近十年历史数据形成用户借阅行为的共现矩阵,研究满足用户多样化需求的推荐模型^[5];还可结合用户的社交媒体行为、科研行为、校内学习行为等图书馆外部数据进行融合分析,尽可能全面和精确地构建用户画像,满足深入研究图书馆用户借阅行为的需求,提高数据的使用价值^[6]。

5.1 数据引用格式

李荣,廖映红,李双.东北财经大学图书馆用户借阅记录数据集[DB/OL].[日期].hdl:20.500.12304/10566.

5.2 论文引用格式

李荣,廖映红,李双.东北财经大学图书馆用户借阅记录数据集[J].图书馆杂志,2022(10):104-109.

参考文献

- [1] 肖铮,吴至艺,林俊伟.2007-2017年厦门大学图书馆纸质文献借阅记录数据集[J].图书馆杂志,2020,39(6):113-117.
- [2] 百度百科.UUID[EB/OL].[2020-11-30].<https://baike.baidu.com/item/UUID/5921266?fr=aladdin>.
- [3] 李荣,王庆石,李双,等.基于读者行为大数据的图书馆精准服务与管理创新研究——以东北财经大学图书馆为例[J].图书馆研究,2018,48(3):19-23.
- [4] 施国良,张潇潇,杨小莉.高校读者群体差异对其借阅行为和阅读偏好的影响研究[J].图书馆,2020(4):59-64;78.
- [5] 阮光册,谢凡,涂世文.基于 Word2vec 的图书馆推荐系统多样性问题应用研究[J].图书馆杂志,2020,39(3):124-132.

- [6] 阮光册,任金玥,涂世文.数据驱动下的图书馆用户行为研究[J].图书馆杂志,2019,38(12):28-37.

李 荣 东北财经大学图书馆,副研究馆员。研究方向:图书馆自动化及数据管理。作者贡献:数据论文撰写、数据采集、数据脱敏。E-mail:westsky2000@139.com 辽宁大连 116025

廖映红 女,东北财经大学图书馆,副研究馆员。研究方向:文献资源建设。作者贡献:研究策划、数据论文修订。辽宁大连 116025

李 双 女,东北财经大学图书馆,高级实验师。研究方向:数字图书馆及计算机应用。作者贡献:数据处理方案指导、数据论文修订。辽宁大连 116025

(收稿日期:2021-01-27 修回日期:2021-03-19)