2007-2017 厦门大学图书馆纸质文献借阅记录数据集

肖铮 吴至艺 林俊伟（厦门大学图书馆 361005）

摘要 通过汇文图书馆自动化集成管理系统采集2007年至2017年间厦门大学图书馆纸质文献借阅记录，数据由读者信息、借阅信息、图书信息三部分组成，其中读者隐私信息经过加密不可恢复性脱敏处理，结果共有11137751条记录。本数据集可用于分析近年高校图书馆纸质文献的使用情况、用户阅读行为演变、经典图书榜单等；也可作为图书馆馆藏建设、借阅规则调整、馆舍布局优化等决策支撑数据；还可用于指导图书馆开展阅读推广、个性化推荐等服务，具有较高的科研价值和实用价值。

关键词 纸质文献 借阅记录 厦门大学 汇文

Dataset of print resources borrow and return records at Xiamen University Library（2007-2017）

Xiao Zheng, Wu Zhiyi, Lin Junwei (Xiamen University Library)

Abstract: This paper uses Huiwen library integration system to collect borrow and return records at Xiamen University from 2007 to 2017. The data was composed of readers, books, borrow and return information, the privacy information is desensitized, and a total of 11137751 records. The dataset can be used to analyze the use of print resources in academic library, the changes of readers reading behaviors, and the list of classic books. It can also be used as decision support data for library resources construction, lending rules adjustment and library space optimization, and can also be used to guide library reading promotion, personalized recommendation and other services. It has scientific research value and practice value.

Keywords: print resource; borrow and return records; Xiamen University; Huiwen

数据库基本信息

数据集中文名称	2007-2017 厦门大学图书馆纸质文献借阅记录数据集
数据集英文名称	Dataset of print resources borrow and return records at Xiamen University Library（2007-2017）
数据作者	肖铮；吴至艺；林俊伟
通讯作者	肖铮
作者单位	厦门大学图书馆
版本号	V1.0
版本时间	20190301
基金项目类型	无

基金项目名称	无
国家	中国
语种	中文
数据覆盖时间范围	20070101-20171231
地理区域	厦门大学
经纬度	北纬 N24° 44’26.27″ 东经 E18° 11’19.07″
数据格式	csv
数据体量	数据记录 11137751 条，共 2.52G
关键词	纸质文献；借阅记录；厦门大学；汇文
主题分类	研究数据；科学数据；图书馆情报学；计量学
全球唯一标识符	hdl:20.500.12304/10159
网址	http://hdl.handle.net/20.500.12304/10159
数据集组成	共 11 个文件，分别为 2007 年至 2017 年的数据
使用条款	本数据集可以通过在网站实名注册登陆获取，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 背景

近年来，在网络环境的影响下，图书馆纸质文献资源使用量呈现出逐年下降的趋势[1]。纸质文献作为图书馆传统的资源形式，在新环境下如何更好地建设纸质文献馆藏资源，更好地发挥其作用与价值，值得图书馆深入研究。本数据集通过厦门大学图书馆自动化集成管理系统采集 2007 年至 2017 年间厦门大学师生使用纸质文献的借阅记录，数据规范完整具有一定的代表性。该数据集有助于研究高校师生纸质文献使用，开展阅读推广活动，改进纸质馆藏建设，修改借阅规则，优化馆舍布局，提供个性化服务等科研与实务工作，具有较强的实用性。

1 数据采集和处理方法

厦门大学图书馆自 2006 年 10 月起开始使用汇文图书馆自动化集成系统，汇文系统采用标准的关系型数据库 Oracle 作为数据仓库，用户信息、书目信息、借阅历史信息、馆藏地信息分别存储在 READER 表、MARC 表、LEND_HIST 表、LOCATION_LST 表中，使用 SQL 查询语句，关联以上四个数据表，提取读者历史借阅记录。各表间关联规则和限制条件为：

（1） READER 表学工号字段与 LEND_HIST 表中的学工号字段关联。

（2） LEND_HIST 表图书 MARC 记录号字段与 MARC 表 MARC 记录号号字段关联。

（3） LEND_HIST 表中馆藏地 LOCATION 字段与 LOCATION_LST 表中馆藏地 LOCATION 字段关联。

（4） LEND_HIST 表借阅时间字段以自然年为单位，分年度提取借阅记录。如 2007 年借阅记录的限定条件为大于等于‘2007-01-01’并小于等于‘2007-12-31’。

厦门大学图书馆在 2006 年 10 月前使用的是 ILAS 图书馆自动化集成系统，在从 ILAS 系统迁移至汇文系统时，由于两套系统的数据库结构定义不同，为了保证项目进度，对导入的数据质量要求有所降低，导致部分数据字段缺失。在十几年的使用过程中，也存在由于工作人员在操作中未完全遵照规范标准工作，导致个别字段存在数据前后不一致的问题，比如用户年级字段，早期使用“阿拉伯数字”+“级”表示，如“2005 级”，后期直接使用“阿拉伯数字”表示，如“2010”。另外，由于留级或者转专业等情况，图书馆在接收到教务处下发的留学和转专业学生名单后，会对学生年级字段进行更新，造成在借阅记录中出现看似“不可理”的数据。比如 2013 年的借阅记录中出现了年级为 2014、2015 的记录。或者由于人工操作时录入疏忽，导致个别字段缺失，比如用户性别字段。为了保证数据规范，对个别字段进行了标准化和数据清理。

（1）年级字段“READER_GRADE”：通过字符串替换，将“级”字去掉，只保留阿拉伯数字表示学生年级。将因个别工作人员在录入时，操作失误把学院字段当作年级字段的数据，通过学号中的信息，更新为正确的年级。将操作失误把备注字段当作年级字段的数据，更新为空值。将因工作人员在人工开通用户证件时，误将注册日期当作年级组录入，导致个别用户类型为“教职工”的年级字段非空，经过处理后，更新为空值。将因留级或者转专业的学生年级数据，通过学号中的信息，更新为正确的年级。

（2）ISBN 字段“M_ISBN”：图书信息数据在从 ILAS 系统导入到汇文系统过程中，因数据质量控制标准有所降低，导致导入的数据中存在错误。通过正则匹配，将 ISBN 字段中的如“CN”，“不详”等错误数据清空。因为 1987 年以前的出版物，采用统一书号，1987 年以后的出版物采用标准书号。为了提高兼容性，在从 ILAS 转入图书信息数据时，将统一书号和标准书号都存入了 ISBN 字段中。

（3）出版年字段“M_PUB_YEAR”：图书信息数据在从 ILAS 系统导入到汇文系统过程中，以及日后使用过程中工作人员操作规范不统一等原因，出版年字段存在不统一的情况。通过正则匹配，将出版年字段中的非年份字符如“c”，月份字符等删除，只保留 4 位阿拉伯数字表示出版年。

为了提供更具研究价值的完整数据，在借阅记录中包含了用户学工号，因涉及到个人敏感隐私数据，为了让使用者能够安全使用脱敏后的真实数据集，对借阅记录中学工号记录进行了数据脱敏。数据脱敏步骤为，先将学工号和密钥进行混淆，然后再采用 MD5 方式进行加密。经过处理后的学工号数据，保留了数据的一致性特性，即相同学工号经处理后的脱敏学工号数据仍然相同。同时，脱敏后的数据无法逆向解密，即无法通过脱敏后的学工号获得真

实学工号数据。

表 1 数据脱密处理样式

字段名	脱敏前	脱敏后
READER_ID	7302	0FA8064A6180441A575A6385919450ED

2 数据字典、数据样本和数据量

数据集包含了字段代码、字段中文名、字段样例值和备注，见表 2 所示。

表 2 数据字典

字段代码	字段中文名	样例值	备注
READER_ID	学工号	0FA8064A6180441A575A6385919450ED	经过脱敏处理,但保留了属性特征。
READER_SEX	性别	M	男性值为 M，女性值为 F。因录入疏漏等问题,存在个别记录的性别字段为空值。
READER_DEPT	单位	图书馆	
READER_GRADE	年级	2011	用户类型为“本科生，硕士生，博士生，嘉庚本科,成教生,交流生,大专生,预科生,博士后,海外生,马校本科”的记录,此字段为该生入学年份;其它用户类型,该字段为空值。
READER_TYPE	用户类型	本科生	用户类型共 23 种： 本科生,硕士生,博士后,博士生,成教生,大专生,短期培训,附属单位,海外生,嘉庚办证,嘉庚本科,嘉庚教工,交流生,教职工,马校本科,马校教工,聘用教师,图书馆工作人员,消费卡,校外办证,业

			务用证,预科生,阅览证。
LEND_DATE	借出时间	2017-01-0207:39:42	
RET_DATE	还回时间	2017-02-1614:11:33	
RENEW_TIMES	续借次数	1	用户对该文献的续借次数。
LOCATION_NAME	馆藏地	总馆基本书库	文献所在的馆藏地。
M_TITLE	题名	数学分析教程	
M_CALL_NO	索书号	017/714	
M_ISBN	ISBN	13209.119 7-208-1440-X 978-7-302-29303-3	ISBN 字段存在三种格式： 1. 全国统一书号； 2. 10 位 ISBN 号； 3. 13 位 ISBN 号； 部分记录中含有“-”分隔符，部分记录不含“-”分隔符。
M_AUTHOR	责任者	高孝忠编著	
M_PUBLISHER	出版社	清华大学出版社	
M_PUB_YEAR	出版年	2012	
DOC_TYPE_NAME	文献类型	中文图书	文献类型共有 15 种： 德文图书,多媒体资料,俄文图书,法文图书,规范文档,韩文图书,类型不详,日文期刊,日文图书,西文期刊,西文图书,学位论文,印刷乐谱,中文期刊,中文图书。

表 3 各年份数据记录数统计

年	记录数	年	记录数
2007	1548175	2013	882944
2008	1509457	2014	791968
2009	1376070	2015	637856
2010	1255691	2016	555382

2011	1145562	2017	487037
2012	947609	总计	11137751

3 数据质量控制

为了保证数据的完整性和准确性，我们使用了汇文系统的统计模块提供的读者借阅统计功能，逐年统计了年度总借阅数。经过比较，汇文系统统计的年度总借阅数与我们提取的当年度借阅记录总条数一致，保证了数据的完整性。另外，使用汇文系统的流通模块的读者借阅历史统计功能，抽取若干用户在某年度的借阅历史，与我们提取的数据中当年度该读者的借阅记录相比较，借阅记录条数和借阅信息均一致，从而保证了数据的准确性。

4 数据价值

近年来，随着移动化、数字化、碎片化阅读方式的兴起，纸质文献的使用量有所下降，在新环境下如何更好地发挥纸质文献价值，改进纸质文献资源建设方法，促进纸质文献的使用，是图书馆面临的新课题。本数据集基于 2007 年至 2017 年间厦门大学图书馆汇文系统中的真实借阅历史记录提取生成，数据采集时间跨度大，数据量大，完整准确，具有较好地代表性。数据集包括脱敏后的用户信息、图书信息、借阅信息，可从不同维度对数据集进行分析研究。对年度借阅量进行统计，可帮助研究者掌握近年高校图书馆纸质文献的使用情况[2]，了解纸质文献近年的借阅变化趋势，分析影响纸质文献借阅量的因素。对各类型读者的借阅数据分类统计，可帮助研究者对比分析不同类型读者近年阅读行为变化，根据不同类型读者推出适合的推荐阅读书目，制定不同的阅读推广方案。基于借阅信息统计热门图书、热门作者、热门出版社，帮助图书馆采访人员跟踪采购，优化纸质文献馆藏建设方案[3]。借阅数据还可以作为借阅规则更改、书库空间调整、馆舍布局等决策的支撑数据。此外，还可作为推荐系统和深度学习的训练数据集，用于验证算法，优化建模等研究。本数据集具有较高的科研价值和实用价值[5]。

5 数据使用方法和建议

本数据集可以使用通用的关系型数据库软件、编程语言 Pyhton 和 R、数据统计分析软件 SPSS、数据表格软件 Excel 等工具进行统计分析。

（1） 通过对年度借阅总量统计，使用数据可视化工具，分析近十年纸质文献使用变化趋势。

（2） 根据用户类型、学院、年级用户数据的聚类分析，分析各类型用户的借阅量变化趋势，对比不同类型读者借阅行为变化。统计各类型读者阅读习惯，平均借阅周期，重复借阅

行为。

（3） 统计各类型、各学院、各年级用户的热门图书，分析读者借阅倾向，形成各类型读者的经典必读书单。

（4） 通过数据挖掘，分析读者与读者，读者与图书，图书与图书之间的关联关系，对读者和图书进行画像分析。

（5） 作为图书推荐系统算法的训练数据，构建高质量的图书推荐模型。

参考文献：

[1]庄小峰,马凌云.近十年高校图书馆读者图书借阅偏好及变化研究——以上海师范大学为例[J].河北科技图苑,2018,31(04):57-62.

[2]李艳琼,崔湛,郑艳.高等院校图书馆资源利用调查分析——以云南农业大学为例[J].图书情报工作,2018,62(S1):56-59+63.

[3]鞠兰萍.基于流通借阅统计分析的馆藏优化策略研究[J].图书馆,2012(04):69-72.

[4]冯向春.对高校图书馆纸质文献借阅现状的思考[J].图书馆研究,2015,45(02):49-53.

[5]李伟.数字信息时代高校图书馆纸质馆藏的价值与服务对策[J].图书情报工作,2017,61(03):79-85.

作者简介：

肖铮 男 厦门大学图书馆 高级工程师。研究方向：数字图书馆。作者贡献：主题策划、确定数据遴选标准、数据提取论文写作。Email：zhengx@xmu.edu.cn 厦门 361005

吴至艺 男 厦门大学图书馆 馆员。研究方向：数字图书馆。作者贡献：数据提取和数据检查。

林俊伟 男 厦门大学图书馆 工程师。研究方向：数字图书馆。作者贡献：数据库测试环境搭建。

联系方式：

通讯地址：福建省厦门市思明南路 422 号厦门大学图书馆

手机：13515964765