

大数据环境下高校图书馆数据管理 与分析应用实践^{*} ——以北京师范大学图书馆为例

刘斌 黄婧 李峰 李书宁
(北京师范大学图书馆, 北京 100875)

摘要: 高校图书馆利用大数据技术与服务对各类数据采集、存储、管理与分析, 能够为其提升管理与服务水平、分析和预估发展趋势提供数据支撑。本文总结高校图书馆的数据类型与特点, 在调研国内外高校图书馆相关实践工作基础上, 以北京师范大学图书馆为例, 阐述数据管理与分析系统的架构, 以及利用系统对数据进行管理与分析的过程, 指出高校图书馆利用数据管理与分析系统对不同来源数据进行采集、处理与分析, 能够直观地为图书馆决策提供数据支撑, 有助于提高自身的数据管理与分析能力, 提升科学管理与服务水平。

关键词: 大数据; 高校图书馆; 数据管理; 数据分析

中图分类号: G250

DOI: 10.3772/j.issn.1673-2286.2018.10.008

随着数字图书馆、智能图书馆的持续发展, 高校图书馆各类服务与管理业务日益增多, 逐渐积累了大量数据, 其中包括图书馆的各类馆藏资源数据、读者图书借阅数据、电子资源利用数据、网站访问数据、自助设备使用数据、咨询数据、管理数据、日志数据等。这些数据对于图书馆全面、深入地了解自身馆藏资源、读者借阅及信息利用行为、读者的信息需求变化、各类服务系统的运行效率等具有重要意义, 可以说这些数据也是高校图书馆的重要核心资产。图书馆通过分析这些数据, 能够获取大量有深度和有价值信息, 并利用这些数据的支持保证图书馆的资源配置、读者服务、系统运行的高效和满意。

通过对各类型数据有效地采集、存储和管理, 并予以科学分析和评价, 不仅能够为图书馆开展更加精准的信息服务和精细的资源管理提供科学决策依据, 也能够为图书馆进行未来规划、服务模式调整、服务创新的趋势分析与发展预估提供数据支撑。在大数据关

键技术不断发展的环境下, 技术应用与解决方案的日趋成熟为图书馆更加科学地管理分析数据提供有效的方法和工具。因此, 高校图书馆需要把握大数据环境带来的良好机遇, 充分利用相关技术工具, 构建数据管理与分析平台, 采用科学的分析与管理策略, 对积累的各类型数据进行管理与分析利用, 从而实现深入、系统地了解读者需求, 提升图书馆服务的智能化程度, 提高图书馆资源配置的合理性, 推动图书馆科学管理, 并将管理与服务决策方式由传统、主观、经验式决策, 向依靠大数据的科学、量化决策方式转变。

本文在调研国内外高校图书馆数据管理与分析实践的基础上, 介绍北京师范大学图书馆依托大数据管理技术架构和商业通用数据分析平台开展多源数据管理和应用实践, 期望能够为大数据环境下的高校图书馆实施“数据驱动”资源建设优化、服务价值提升和管理决策科学化提供有益参考。

^{*}本研究得到北京地区高校图工委2016—2018年科研基金项目“大数据环境下图书馆的数据管理与服务实践研究”(编号: BGT2016021)资助。

1 高校图书馆数据类型与特点

1.1 高校图书馆的数据类型

高校图书馆数据包括业务数据、馆藏资源数据、读者数据、管理数据。其中,业务数据包括图书馆在馆藏资源管理的业务流程中产生的采、编、流、检数据。馆藏资源是图书馆根据本机构学科专业设置规划和配置馆藏,并经过专业规范的揭示与组织后所形成的信息(文献)资源集合,包括纸质资源、订购的电子资源和自建数字资源等。从数据内容上讲,馆藏资源数据包括图书馆描述和揭示资源本身的结构化数据(即元数据)、资源的全文数据、资源的访问数据等。读者数据是读者在利用图书馆资源、服务的过程中产生的各种行为数据,包括纸质资源的借阅、电子资源及数字资源的检索与全文访问、图书馆网站的访问、自助服务的利用、信息服务的咨询等。管理数据是图书馆在运转与服务中产生的各类数据,包括馆舍数据、人事数据、财务数据、项目数据、合同数据等。这些数据“是高校图书馆最为宝贵的核心数据资产,是图书馆业务运转和开展各类信息服务的基础”^[1]。

以北京师范大学图书馆数据为例,按照数据产生源头划分,可大致将数据分为:①资源数据,包括纸质、电子、数字资源等资源的元数据、全文数据、访问数据等;②业务数据,包括图书馆员业务流程中产生的采访、编目、调拨等工作数据;③读者数据,包括读者个人信息,读者使用图书馆纸质资源产生的图书浏览、借阅数据,使用图书馆电子或数字资源产生的在线访问、检索、浏览、下载等信息行为数据,使用图书馆网站产生的访问、浏览数据,使用图书馆空间产生的入馆、选座、研究间利用数据等;④管理数据,包括图书馆内部的人事数据、财务数据、固定资产数据等。

1.2 高校图书馆的数据特点

从上述数据看,高校图书馆数据具有3个特点。

(1)数据类型复杂多样。首先,高校图书馆资源类型丰富,包括多种文献类型、载体形式和呈现格式;其次,高校图书馆业务系统烦杂,数据来源于各类不同系统,数据结构各不相同,既有来源于关系型数据库的结构化数据,也有来源于日志文件、XML文件等半结构化数据,还有大量没有固定结构的文本、多媒体等资源

对象的非结构化数据。

(2)数据量大。高校图书馆经过多年的建设、服务积累,已经拥有大量的文献信息资源数据、读者信息行为数据、系统运行日志数据、业务运行数据、管理数据等。

(3)数据增长速度快。通过利用物联网技术支持的实体资源管理,利用互联网、移动技术实现的各类资源访问,以及引入社交网络技术的读者服务,促进了图书馆业务及读者数据流动的加速,为图书馆贡献了大量在线、实时数据,加速了图书馆各类数据的生成。

高校图书馆的数据具有类型复杂、数据量大、增长速度快的特点,呈现出大数据的部分特征,给数据管理带来较大挑战。同时,为高校图书馆积极利用大数据相关技术与服务,对各类型数据进行管理与分析带来良好机遇。

2 国内外图书馆数据管理与分析应用实践

2.1 国外图书馆的实践

2001年Laney首次提出“Big data”之后^[2],大数据逐渐成为全球诸多国家的重要战略资源。在这样的时代背景下,国外图书馆界学者开始关注图书馆大数据应用和服务需求并讨论技术问题,其中美国图书馆界在大数据应用于图书馆服务方面做了较多实践尝试。

早在2005年,耶鲁医学图书馆从所有可能的数据源分析纸本刊和电子刊的利用率,揭示读者对于纸本刊和电子刊的喜好程度,图书馆以此为依据调整期刊订购策略^[3]。华盛顿大学图书馆自2006年至今,先后采用Tableau、平衡记分卡等多种工具和方法分析并展示图书馆统计数据,支持馆内的战略决策,目前该馆将图书馆数据分为馆藏、服务和空间3种类型进行数据共享^[4]。2014年,Niu等^[5]通过分析普渡大学图书馆的2种资源发现系统VuFind和Primo的读者日志数据,理解读者检索行为和使用偏好,据此提出一套检索行为评估方法,为发现系统中检索点、分面的设置及检索结果的展示提供了依据。美国哈佛大学图书馆将“大数据”的服务引入图书馆中并付诸实践,将图书馆大数据向读者公布。“这些数据包含1 200多万种资料,有书目数据、地图、手稿、音视频等,并在美国数字公共图书馆中提供下载服务”^[6]。

2.2 国内图书馆的实践

近年来国内学者也开始关注高校图书馆大数据服务和技术问题,并在实践领域有所进展。在理论研究方面,马晓亭^[7]提出一种采用多层次的系统结构的图书馆大数据资源整合平台的框架。在应用技术方面,Chen等^[8]探讨了图书馆大数据的存储、数据挖掘及个性化服务等具体技术,提出以“Hadoop+MapReduce”并行架构的大数据应用方案。在实践领域,上海交通大学图书馆在2012年自主开发完成覆盖图书馆主要业务的一站式统计平台,数据涉及馆藏资源、流通活动、学科服务、应用系统、科研数据、基础信息等,实现所有数据的集成管理,为图书馆各项工作提供指导^[9]。2013年,清华大学图书馆尝试从海量权威的元数据仓储中提取关键词等信息,一方面分析关键词走向,以时间轴展示某学科的发展趋势;另一方面分析作者与合作者的关系,建立以学者为中心的知识关联网络^[10]。2016年,国家图书馆初步建成基于读者与资源核心业务系统的图书馆大数据平台,应用该平台通过一系列分析方法,对国家图书馆服务情况、主要服务对象和整体资源利用情况进行深入分析^[11]。

国内外实践表明,对于高校图书馆来说,优化资源建设、提升服务价值的“数据驱动”特征已经非常明显,通过有效管理和分析来实现数据潜在价值的挖掘成为提高高校图书馆服务水平的发展要求。

3 北京师范大学图书馆数据管理与分析系统应用实践

3.1 早期实践与问题

2015年,北京师范大学图书馆尝试利用图书馆大数据开展数据管理,并为读者提供数据服务。为此,该馆自主开发了面向院系的高校毕业生图书馆记忆系统,整合来自图书馆集成管理系统、门禁系统、座位管理系统、研究间预约系统的数据,通过数据清洗与关联分析,形成可为毕业生提供的图书馆资源与服务利用数据,并提供个人数据的查询与展示。毕业生可通过系统浏览个人的到馆记录、借阅历史清单、图书馆座位使用信息、研究间预约情况等^[12]。该系统通过将多来源的数据整合到同一数据库,分析和挖掘图书馆基础数据的价值,并通过为读者提供的数据服务,吸引众多毕业生的

参与和互动,较好地提升图书馆的影响力。

但是该系统开发的主要目的是为特定的读者服务需求提供系统支持,重点在于数据服务的实现,因此在系统架构与实现方案上存在一定局限。系统不具备前期数据的采集、清洗等功能,同时无法对后期数据进行系统性分析。为了能够满足对图书馆整体的数据管理和常规的数据分析需求,需要应用更合理的系统架构和丰富的功能,实现图书馆数据管理与分析平台的开发与应用。

3.2 当前数据状况

北京师范大学图书馆于2017年9月成功实施并上线了下一代图书馆服务平台ALMA,与早先已实施上线的PRIMO检索和发现系统一起,为该馆统一资源管理与服务提供平台。目前该馆通过ALMA及PRIMO系统,实现了对纸质资源、电子资源、部分数字资源的统一管理和服务,同时为图书馆资源、业务及读者数据的统一管理与整合提供良好的系统条件。通过ALMA系统,图书馆可获取整合的资源数据、采编流业务数据、读者个人信息数据及资源利用数据,基本覆盖图书馆的核心数据。同时,ALMA为第三方系统提供良好的API接口支持,通过ALMA分析模块的API接口,可为第三方数据管理与分析系统提供数据调用。

此外,图书馆的主页系统、门禁系统、座位管理系统及研究间管理系统等为图书馆提供了网站访问、入馆及空间利用的统计数据,这些数据均来源于各系统的关系数据库。ALMA、PRIMO、主页系统、门禁系统、座位管理系统及研究间管理系统等提供的数据,主要为结构化数据,日积月累,数据量逐渐庞大,数据可分析和挖掘的价值逐渐增多,为图书馆管理决策与服务提供了很好的数据基础(见表1)。

3.3 数据管理与分析系统架构

为解决图书馆整体的数据管理与分析需求,北京师范大学图书馆通过与商业公司开展技术合作,尝试利用通用的数据管理与分析系统,解决图书馆数据的统一管理与分析问题。为此,图书馆以读者对图书馆资源、空间的利用数据为实验数据,对系统进行初步应用。利用该系统,采集读者纸质资源借阅数据、门禁数据、座位及研究间利用数据,并通过数据处理,生成的

表1 北京师范大学图书馆数据的主要来源

数据类型	数据内容	系统来源
资源数据	纸质、电子、数字资源	ALMA
业务数据	采访、编目、调拨等	
读者数据	读者个人信息	
	纸质图书浏览、借阅数据	
	电子资源利用数据	PRIMO
	数字资源利用数据	
	检索、全文获取行为数据	数据库商
	其他电子资源利用数据	
	图书馆网站利用数据	图书馆主页
	入馆数据	门禁系统
	选座数据	座位管理系统
管理数据	研究间利用数据	研究间管理系统
	人事、财务、固定资产等	手工处理

结果应用于分析读者群体的资源与空间利用状况。该数据管理与分析平台的架构主要包括数据层、分析层和展示层。其中数据层主要包括数据采集、数据整合、数据ETL（Extract-Transform-Load）处理、数据集市；分析层主要实现对数据查询结果的分析，同时系统支持通过R语言实现的深度分析操作；展示层主要实现对

数据分析结果的图表化展示，并支持移动端的展示功能。具体架构见图1。

3.4 数据管理与分析过程

3.4.1 梳理数据源, 进行数据采集

首先，针对读者对图书馆资源、空间的利用问题，详细梳理涉及的数据来源及数据内容（见表2）。其次，针对涉及的数据来源及各系统数据库情况和开放程度，分别对接各来源系统的数据源。根据数据来源系统的技术开放程度，主要采取3种数据源连接方式。①通过建立ODBC接口直接利用SQL查询语句从对应系统的数据库中实时获取感兴趣的数据（包括My SQL、SQL Server、Oracle等关系数据库的数据），通过此种方式，实现对门禁系统、座位管理系统及研究间管理系统的数据采集。②通过ALMA系统的统计分析模块API接口，实现对ALMA中读者个人信息数据及纸质资源借阅数据的采集。③系统支持数据导入方式，即将工作人员人工录入并保存在Excel、CSV等表单中的数据导入系统。

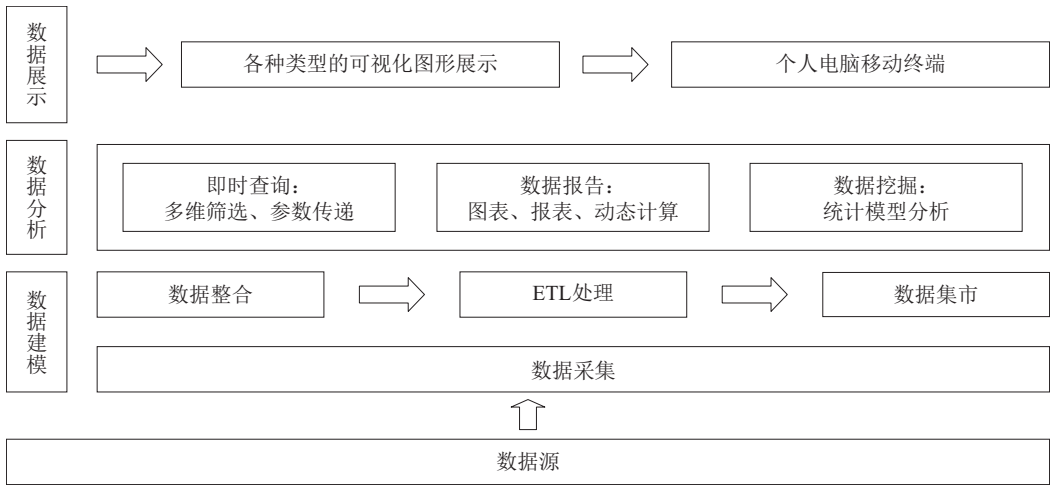


图1 系统整体架构

表2 各数据源的采集内容

数据来源	数据内容	数据字段
ALMA平台	读者信息	读者ID、姓名、学历、院系
	读者借阅统计	读者ID、借阅图书数量、借阅图书信息、借阅时间
门禁系统	读者入馆统计数据	读者ID、入馆时间
座位管理系统	读者选座数据	读者ID、选座时间、座位位置
研究间管理系统	读者预约研究间数据	读者ID、预约时间、研究间位置

各数据源连接完成后,基于具体的数据需求,通过创建同步查询,对核心数据进行同步提取,将数据库的数据保存到本地,它的主要功能是使数据展示和数据操作与其他查询模块产生的结果一致。同步查询可以通过定时任务进行定时同步。

3.4.2 数据处理与分析

各数据源数据采集完成后,基于具体的数据分析指标,进行数据分析。本文围绕读者对图书馆资源及空间的利用这一核心需求,主要提取和分析读者ID关联的个人信息、纸质资源数据、入馆数据、座位数据和研究间预约数据,进而对读者的图书借阅、空间利用行为进行具体分析,核心指标与统计纬度见表3。基于核心指标与统计纬度的设定,将相关字段设置在系统的数据查询模块中。数据查询模块通过系统的查询编辑器定义。此查询编辑器可以将多种数据源的数据混合,构建返回业务逻辑的查询。

表3 核心指标与统计纬度

核心指标	统计纬度
读者信息纬度	学历、院系
图书借阅指标	借阅热门图书、借阅数量
入馆行为指标	入馆次数、入馆高峰
选座行为指标	选座次数、选座位置热点
研究间预约分析	预约次数、预约位置热点

3.4.3 数据分析与报告展示

在系统中将数据分析指标设置完成,就可以通过仪表盘(dashboard)方式对编辑好的查询表数据进行分析图表的创建和展示。系统通过仪表盘编辑器对数据图表进行定制。通过系统提供的各种组件来展现数据。系统支持多种图表类型,选择合适的图表类型后,将上一步中提取的数据查询结果与图表进行绑定,然后选择计量指标与统计纬度,就可生成数据分析的图表展示结果。同时系统支持通过过滤器组件实现对展示图表的数据过滤,图书馆可通过数据过滤器选择特定过滤条件对展示的图表数据进行联动查询。

4 结语

高校图书馆利用大数据技术与服务对各类数据采

集、存储、管理与分析,能够为其提升管理与服务水平、分析和预估发展趋势提供数据支撑。国内外实践表明,挖掘数据潜在价值已成为提高高校图书馆服务水平的发展要求。通过北京师范大学图书馆依托大数据管理技术架构和商业通用数据分析平台开展的多源数据管理和分析应用实践,可以发现高校图书馆利用数据管理与分析系统能够对来自不同数据源的各类数据进行采集、处理与分析,并通过系统生成的可视化分析结果,直观地为图书馆资源、服务及管理决策提供数据支撑。图书馆需要全面地梳理自身数据情况及分析需求,明确图书馆各方面业务及服务的数据统计指标,充分利用数据管理与分析系统完成数据的管理与分析。同时图书馆也可以利用专业的统计分析模型,对采集的数据进行建模分析,完成数据的深度解析,实现对图书馆各方面发展的科学预测,从而进一步提高图书馆数据管理与统计分析能力,提升图书馆的科学管理与服务水平。

参考文献

- [1] 包冬梅,范颖捷,李鸣.高校图书馆数据治理及其框架[J].图书情报工作,2015,59(18):134-141.
- [2] LANEY D. 3D data management: Controlling data volume, velocity and variety [R]. Technical Report, Meta Group, 2001.
- [3] GALLAGHER J, BAUER K, DOLLAR D M. Evidence-based librarianship: Utilizing data from all available sources to make judicious print cancellation decisions [J]. Library Collections Acquisitions and Technical Services, 2005, 29(2): 169-179.
- [4] CHEN H L, DOTY P, YU J C, et al. Library assessment and data analytics in the big data era: Practice and policies [J]. Proceedings of the Association for Information Science and Technology, 2015, 52(1): 1-4.
- [5] NIU X, ZHANG T, CHEN H L. Study of user search activities with two discovery tools at an academic library [J]. International Journal of Human-Computer Interaction, 2014, 30(5): 422-433.
- [6] 程莲娟. 美国推进大数据的应用实践及其有益借鉴——基于图书馆视角的分析 [J]. 情报资料工作, 2013, 34(5): 110-112.
- [7] 马晓亭. 大数据时代图书馆数据整合系统构建研究 [J]. 图书馆建设, 2014(6): 83-87.
- [8] CHEN M, MAO S, LIU Y. Big data: a survey [J]. Mobile Networks and Applications, 2014, 19(2): 171-209.

- [9] 孙翌, 郑巧英, 徐璟. 多数据来源的高校图书馆统计系统研究与实践[J]. 国家图书馆学报, 2014, 23(4): 64-69.
- [10] 邓景康. 大数据环境下清华大学图书馆的实践[EB/OL]. (2013-08-29) [2017-08-29]. <http://www.ids.tsinghua.edu.cn/index.php?p=3100>.
- [11] 杨帆, 张红, 薛尧予. 基于核心业务系统的图书馆大数据平台构建策略研究[J]. 图书馆学研究, 2017(6): 38-42.
- [12] 李峰, 李书宁, 于静. 面向院系的高校毕业生图书馆记忆系统[J]. 现代图书情报技术, 2016, 32(5): 99-104.

作者简介

刘斌, 男, 1979年生, 硕士, 副馆长, 研究方向: 信息技术管理。

黄婧, 女, 1981年生, 硕士, 副研究馆员, 研究方向: 数字资源建设与管理。

李峰, 女, 1982年生, 硕士, 副研究馆员, 通信作者, 研究方向: 系统开发与应用, E-mail: lif@lib.bnu.edu.cn。

李书宁, 男, 1978年生, 博士, 副研究馆员, 研究方向: 业务数据管理与统计分析。

The Practice on Data Management and Analysis in Academic Libraries in the Environment of Big Data: Taking Beijing Normal University Library as an Example

LIU Bin HUANG Jing LI Feng LI ShuNing
(Beijing Normal University Library, Beijing 100875, China)

Abstract: To collect, store, manage and analyze different kinds of data in academic library by using big data technology and service, can provide data support for the library to improve management and service level, to analyze and predict the trend of development. This paper summarizes the data types and characteristics of university libraries. Then taking the library of Beijing Normal University as an example, it illustrates the framework of data management and analysis system and the process of data management and analysis by using the system, on the basis of investment at home and abroad. Through the system application, collecting, processing and analyzing data from different sources can directly provide data support for library decision-making, help to improve their own data management and analysis capabilities, and enhance scientific management and service level.

Keywords: Big Data; Academic Library; Data Management; Data Analysis

(收稿日期: 2018-10-08)

■ 书 讯 ■

《网络环境下叙词表编制与发展》

叙词表作为一种有效的知识组织工具, 在网络环境下继续发挥着重要作用。中国科学技术信息研究所常春研究馆员及其项目研究团队, 依托国家社科基金项目“网络环境下叙词表的编制模式与应用方式研究”(10BTQ048), 对这一课题开展研究, 并于完成《网络环境下叙词表编制与发展》。

该著作主要论述网络环境下叙词表的编制、维护与应用的理论和方法。编制方法包括网络环境下总体策略、总体形态、选词方法、词间关系建立方法、编制管理机制、维护方式方法等; 应用研究包括网络环境下相关技术的突破给叙词表带来的各类新的应用方式, 从术语服务、多语种翻译、概念组配、知识单元、概念映射、国外应用等多个方面, 阐述了网络环境下叙词表的发展方向。最后按年代顺序介绍了国内历年编制的、可查阅的重要中文叙词表, 理、工、农、医四大领域20多个可从网络上在线获取的英文叙词表。该书可供图书馆学、情报学专业相关专业人员参考使用。

《网络环境下叙词表编制与发展》于2015年4月由科学技术文献出版社出版, 定价38.00元。