DOI：10.13998/j.cnki.issn1002-1248.2015.09.022

大数据时代数字图书馆发展态势分析

赵来娟，王惠英
（甘肃农业大学图书馆，甘肃　兰州　730070）

摘　要：大数据的快速发展，使数字图书馆建设进入一个全新的时代。阐述了大数据在国内外图书馆界研究和应用现状，分析了大数据环境下数字图书馆建设发展的机遇和挑战，提出了大数据时代数字图书馆发展策略。

关键词：大数据；数字图书馆；发展策略；非结构化

中图分类号：G250　　**文献标识码：**A　　**文章编号：**1002-1248（2015）09-0091-05

The Theoretical Analysis on Development Strategy of Digital Library in the era of big data

ZHAO Lai-juan, WANG Hui-ying
(Library of Gansu Agricultural University, Gansu Lanzhou 730070, China)

Abstract: With rapid development of big data, the construction of digital library entered into a new era. Based on the elaboration of the status quo of big data in the research and application of libraries at home and abroad, the article analyzed the opportunities and challenges in the construction of digital library in big data environment and provided the development strategies and suggestions of digital library.

Keywords: Big data; Digital library; Development strategy; Unstructured

1　大数据：一个时代特征

1.1　什么是大数据，大数据在哪里

“大数据”一词最早出现在《自然》杂志 2008 年 9 月专刊发表的文章“Big Data: Science in the Petabyte Era”[1]。2011 年 6 月，美国咨询界的翘楚 McKensey 咨询公司发布了《大数据：下一个竞争、创新和生产力的前沿领域》的研究报告[2]。这份长达 150 余页的报告称：“数据，已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来”。大数据在物理学、生物学、环境生态学等领域以及军事、金融、通讯等行业已经广泛应用，引起人们的极大关注[3]。到 2012 年为止，Farecast 系统用了将近十万亿条价格记录来帮助预测美国国内航班的票价，且预测的准确度已经高达 75%[4]。淘宝网作为国内最大的 B2C 机构，运用大数据技术，在 2014 年的双“十一”，淘宝系网站的销售总额达到 570 亿元人民币[5]。在各大企业看来，大数据已经成为一种巨大的生产力。据国际数据资讯公司(IDC)监测，全球数据量大约每两年翻一番，意味着人类在最近两年产生的数据量相当于之前产生的全部数据量，表明大数据时代已经到来，将给各行各业带来数据使用方式的根本性变革。

1.2　大数据的 4 个主要特征

“大数据”是一个用来描述海量的结构化和非结构化数据的流行短语，这些数据的容量巨大，结构复杂，以至于很难用传统的数据库和软件技术进行存储、管理和处理。其特征可归纳为 4 个“V”，即 Volume（容量）、Variety（多样性）、Velocity（速度）和 Value（价值）。容量(Volume)指向数据集合的规模不断扩大；多样性(Variety)指向大数据类型繁多，包括结构化、半结构化和非结构化数据；速度(Velocity)指向数据通常以数据流的形式动态、快速地产生，具有很强的时效性；价值(Value)指向庞大的数据量蕴含着巨大财富。大数据是人们获得新

收稿日期：2015-04-28

作者简介：赵来娟（1981-），女，硕士，甘肃农业大学图书馆，馆员。王惠英（1975-），女，硕士，甘肃农业大学图书馆，馆员。

的认知、创造新的价值的源泉，大数据还是改变市场、组织机构，以及政府与公民关系的方法。

2 大数据与数字图书馆

以计算机技术、网络技术和数字化技术为支撑的数字图书馆在国内外迅速发展，成绩斐然，但仍存在简单复制传统图书馆功能、信息资源系统共享不足、囿于图书馆基本功能等局限，为其发展带来了潜在瓶颈[6]。大数据时代的到来为突破数字图书馆局限性带来了机遇，目前已成为国内外图书情报学研究的热点。

2009 年，欧洲一些领先的研究型图书馆和科技信息研究机构建立了伙伴关系，致力于改善在互联网上获取科学数据的简易性。2012 年 3 月，美国政府启动了《大数据研究和发展计划》[7]，同时，奥巴马还强调政府必须和公司、大学合作结盟，全民动员来应对“大数据”时代的挑战。Hiptype 公司应用大数据技术来分析读者使用电子书的阅读习惯和爱好，构建知识服务社区实体行为智能分析引擎，从而有针对性地开展服务，取得了较好的成效[8]。哈佛大学图书馆已将“大数据”的服务引入数字图书馆，将图书馆大数据向读者公布[9]。在欧洲，英、法、德等国也组织力量致力于大数据的收集、储存与分析研究等方面的工作。

通过 CNKI 学术资源总库检索国内学者研究的相关文献，2012 年，有关“大数据”与“图书馆”的论文只有 8 篇，2014 年增至 326 篇，从论文数量上的猛增态势可以预见在未来几年图情界对于“大数据”的研究将成为新的热点。张兴旺、郭自宽等人[10-11]从宏观上描述了国内外关于大数据的研究现状和学术环境，以及大数据生态系统在图书馆中的应用过程。吴志荣[12]从文献资源建设角度提出了大数据时代的文献发现理论，认为“文献发现”是当代图书馆新的社会职能和核心竞争力。韩翠峰[13]认为大数据将使图书馆在数据存储、数据挖掘、数据分析等方面面临巨大机遇与挑战。

总体来看，国内图书情报学界的相关研究更多地是关注大数据对图书馆事业发展的影响与挑战等方面，而很少有从技术层面研究大数据如何在图书馆付诸应用的。大数据应用于数字图书馆是大势所趋，有着非常广阔的应用与研究空间。国内信息资源产业的排头兵 CNKI 、超星、方正等已在尝试将大数据的搜集、挖掘、分析、运用等技术引进，改善基础服务体系，拓展增值性的附加服务，助推数字图书馆创新发展。

3 大数据环境下数字图书馆面临的差距与挑战

3.1 数字图书馆与实体馆的“同质”问题

数字图书馆与实体馆的“同质”问题表现为两个方面：一是与实体馆资源的同质，二是与实体馆用户的同质。首先，数字图书馆处理的问题是数据资源数字化、音频视频信息的转换、存贮和检索以及多媒体信息技术的扩展深化，但缺乏对海量数据的加工处理与管理服务。其次，从长远发展的角度来看，数字图书馆应该进行数字资源的深层次开发，拓展对原始数据的挖掘、采集、组织、保存与利用，开拓一条数据资源主导型的发展新模式。另外，技术上的差距并不难弥补，最大的瓶颈或差距是收集数据的意识。收集数据的意识不强，对于数据在决策当中的重要性认识不够，这才是数字图书馆最大的制约因素。用户同质问题方面，因数字图书馆资源内容多数为馆藏资源的数字化，实体馆用户同质的现象较为普遍。一些数字图书馆系统自成体系，难以走出实体馆的象牙塔，实现不同用户群体的信息共享与利用。

3.2 数字图书馆的非结构化数据空白

大数据可以分成两种类型：一是结构化数据，即行数据，是存储在数据库里，可以用二维表结构来实现的数据；二是半结构化或者非结构化数据（如电子邮件、办公处理文档，以及许多存储在 Web 上的信息及图像、音频和视频等）。当前数字图书馆多数为文献资料类数据库建设，非结构化的数据内容所占比重非常低，并且缺乏大数据的分析，数字图书馆很难融入企业等用户群体的细节服务。随着云计算、RFID、物联网、语义网、传感网、社交网、移动互联网等新的渠道和技术在数字图书馆中的应用，产生了大量电子邮件、数据日志、文字处理文件以及大量发布在网络上的新闻等无序、未加工整理的海量信息资源，加之数据库使用统计数据、书籍借阅数据、网站点击数据等都未被数字图书馆收集、加工、处理，造成了这些半结构化、非结构化数据的缺失。至 2012 年，非结构化数据占有比例将达到互联网整个数据量的 75%以上[14]。大数据时代，大数据的缺失使得数字图书馆成为实体馆的象牙塔；大数据分析技术的缺乏，使得数字图书馆很难融入各用户群体的细节服务。

非结构化数据作为一个新的尚未开发的信息源，可以使数字图书馆资源的结构更加全面，更加顺应时代的发展，更加适应读者的需求。通过对半结构化、非结构化数据的分析，可揭示以前很难或无法

确定的重要相互关系，可以获得更加丰富、深入和更加准确的用户，可以深入理解读者并给予智慧型的解决方案，最终提高数字图书馆的核心竞争力。

3.3 数字图书馆处在象牙塔　远离创新前沿

党的十八大提出要构建以企业为主体、市场为导向、产学研相结合的技术创新体系，而目前的数字图书馆处在象牙塔，远离创新前沿。任何人在任何时间、任何地点，可以获得所需要的任何知识，这是数字图书馆建设的美好愿景。而当前，多数数字图书馆服务系统都是基于门户网站开展的服务。少数数字图书馆的服务范围从互联网向移动通信网、广播电视网等网络平台逐步拓展，开展了移动图书馆等新媒体服务建设，但服务功能有限。国内数字图书馆对用户信息需求与信息获取习惯的变化还不够敏感，缺乏创新理念与服务机制，缺乏与业务流程的融合，数字图书馆处在象牙塔，远离创新前沿。

3.4 数字图书馆面临新的研究需求

科学研究的变化，要求数字图书馆适应新的研究需求。在 E-Science、海量数据、科教结合、协同创新、产学研结合、第四范式等科学研究发展的新理念、新模式下，数字图书馆面临新的研究需求。大数据时代的研究需求更多地是数据驱动的研究，研究侧重于面向问题的研究、面向数字与模拟的研究、面向决策支持的研究、面向协同创新的研究。越来越依赖数据科学研究的不断变化转型，对数字图书馆的大数据利用提出了要求，然而数字图书馆缺乏大数据的利用，这无法迎合科学研究的变化要求。

4 数字图书馆发展的“大数据”策略

4.1 以大数据创新数字图书馆资源建设模式

大数据时代数字图书馆应该进行数字资源的深层次开发，在注重诸如电子邮件、仪器仪表数据、日志数据、科学研究数据、丰富的媒体数据等半结构化、非结构化数据建设的同时，拓展对原生大数据和特藏大数据的挖掘、采集、组织、保存与利用，开拓一条数据资源主导型的发展新模式。

首先，应尽快开发新一代数字图书馆应用支撑平台，以支持多种异质文档及其元数据的管理，支持多媒体文档的存储、保管、检索和管理[15]，支持结构化数据与非结构化数据的统一管理。其次，在各类教育、科研和文化对象正逐渐走向信息化、数字化、网络化的全面、泛在的数字信息环境中，基于社交网络的“非正式”数字内容、原生数字内容、开放数字内容等在许多领域已经成为主要的信息资源，搜索引擎就是网络信息资源最有力的组织者，通过百度、Google 等进行信息检索与利用已经成为许多人查找信息的首选。这时，数字图书馆需要借鉴搜索引擎搜集网络信息的优势，开放集成网络环境下的各类数字内容，根据用户的需求来有机链接所需要的内容及其分析利用工具，使自己真正成为信息社会的知识服务枢纽。最后，建设基于大数据的特色数据库。能够体现一个图书馆与另一个图书馆之间差别的地方在于：对图书馆所在单位或机构特色资源的开发、整理和建设。目前国内外在这一方面都有行动，比如北京大学的北京历史地理数据库[16]、美国国会图书馆开发的 24 个不同专题特色库[17]，中国拓片（Chinese Rubbings Collection）专题[18]等。大数据时代的数字图书馆应加强特色资源的建设。

因此，在数据爆炸性增长、新类型数据不断涌现、数据结构更趋复杂的大数据时代，数字图书馆的资源建设也将发生深刻的变化，在这全面和泛在的数字信息环境，数字图书馆需要开放集成网络环境下的各类数字内容。

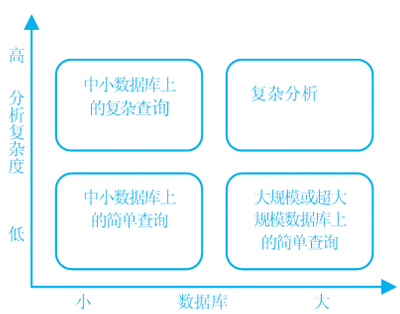
4.2 以大数据助推数字图书馆技术创新

大数据的意义并不在于大容量、多样性等特征，而在于如何对数据进行存取、组织管理和分析，以及因此而发掘出的价值。

大数据开发必须能够高速获得海量的复杂结构化、半结构化和非结构化数据。一方面，大规模数据的激增，对图书馆的基础设施提出了挑战。出于成本的考虑，大数据应用机构将软硬件资源的建设由以前的高端服务器向中低端的大规模计算机集群转变[19]。另一方面，复杂结构的数据对数据仓库提出了更高的要求。目前，处理结构化大数据的关系数据库管理技术已经非常成熟，如商业型 Oracle、Sql Server、开源型 MySql 等，均具备了强大的结构化数据管理功能，并且均拥有较为强大的数据仓库功能，已经形成了基本固定的模式和方法，有力地推动了数字图书馆的资源揭示。针对复杂的半结构化非结构化数据，关系数据管理系统的扩展性遇到了前所未有的障碍，大数据时代需要一些能够处理大型非结构化数据的工具和平台。以 Hadoop 为代表的 HDFS 文件系统和 MapReduce 数据处理框架将结构化、半结构化和非结构化数据的有效管理变为现实。

大数据分析技术可以帮助数字图书馆的业务范

围由对资源组织和服务的集成检索调整到对资源的深度聚合、满足用户对情报的分析统计和对知识的发现评价上来。巨量数据中蕴含着大量有价值的情报信息。为了从数据中挖掘知识、发现价值并加以利用，指导人们的决策，必须对数据进行超越常规报表的深入分析。如附图所示,人们不仅需要通过数据了解现在发生了什么，更需要利用数据对将要发生什么进行预测,以便在行动上做出一些主动的准备[20]，例如预测客户流失情况。通过强大的大数据分析平台还可实现可视化分析、数据挖掘、图形分析、文本分析、智能语义分析、空间信息分析、语义引擎、数据质量和数据管理等功能，从而发现新知识。



附图　数据分析的趋势

4.3 以大数据促进数字图书馆服务升级

在数字化、网络化、规模化和集约化等共性技术特征的基础上，大数据支撑的数字图书馆信息服务模式可以概括为以下几点。

4.3.1 个性化服务

大数据支撑的数字图书馆信息服务是个性化主动推送服务。越来越多的用户期望能从海量数据中得到具有针对性的、个性化的信息服务和用户支持，从“个人计算机”向“个人计算”过渡。个人计算千差万别，大数据支撑的数字图书馆以用户为导向，对数字信息资源和信息服务整合集成，通过深度、广度以及常规分析预测，能根据用户个人爱好和特点，在用户希望的时间和希望的地点，主动推送各取所需、各得其所的个性化服务，将会大大提升数字图书馆的知识服务质量。

4.3.2 按需服务

大数据时代，数字图书馆的信息服务是一种由用户需求驱动的知识服务新模式。从用户的需求看，不是因为图书馆拥有什么就利用什么，而是根据自己的需要，组织数据库资源、网络资源、服务资源等，包括半结构化、非结构化和结构化数据，构建包括物理图书馆在内的信息环境。大数据数字图书馆融合了物联网、传感网、云计算、可信计算和信息物理融合系统等新兴信息技术。面对海量数据，及时检索所需数据、快速进行知识挖掘，不仅为用户提供所需要的知识，而且为用户提供潜在的但和用户需求相关的有价值的信息。

4.3.3 不确定性服务

大数据支撑的数字图书馆信息服务是不确定性服务。对于用户信息处理需求不具备惟一解，而是用大数据技术和方法，依据用户大数据处理需求，形成知识服务解集合。这是由大数据的不确定的特性所决定的。在大数据时代，允许不精确的出现已经成为一个新的亮点，而非缺点。因为放松了容错的标准，人们掌握的数据也多了起来，还可以利用这些数据做更多新的事情，用大量数据创造了更好的结果。用户通过大数据数字图书馆信息服务平台提出大数据处理需求，并按用户自主需求构建的大数据知识服务组合模型，部署服务实施方案。大数据数字图书馆信息服务平台通过支持语义 Web2.0 技术、智能优化算法，对用户所提出的大数据信息服务需求进行语义相似度计算、智能推荐和检索提示等，从搜索到的符合用户需求的大数据信息服务解集合中，选择合适的服务参与组合,并通过协同优化算法从所有可能的大数据信息服务解集合中优选与组合出最佳的一组组合来协同完成用户请求。

4.3.4 智能型服务

大数据时代数字图书馆的服务内容，是以预测分析为基础的知识型、智能型服务。大数据将图书馆从知识的聚集地变革成为知识的加工地、新知识的孵化地。大数据的核心就是预测。它通常被视为人工智能的一部分，把数学算法运用到海量的数据上来预测事情发生的可能性。用户往往需要通过海量非结构化、半结构化数据了解现在发生了什么，甚至需要利用数据预测未来将要发生什么，以便在行动上作出利于发展的主动准备。如通过预测用户的流失，预先采取行动,或预测竞争对手下一步行动以便采取主动等。大数据支撑的数字图书馆根据用户科学研究主题，及时跟踪该研究领域的核心著作、期刊、科研课题等资源，了解该研究领域权威专家的学术科研动态;运用大数据分析技术，预测学术趋势和发展方向，以便于用户对新研究方向进行宏观把握，为后续科学研究工作的开展奠定基础。另外还可以通过行为数据分析读者喜好和偏爱。

5 结语

作为国家信息化战略重要组成部分的数字图书

馆建设，应积极迎合大数据技术的涌现，建立融数据、文献、新型处理技术、创新服务于一体的新型数字图书馆。形成数据与多类信息资源融合的互操作架构。随着大数据在数字图书馆建设中的研究不断深入，必将促使数字图书馆的内涵不断发展与完善，促进数字图书馆的服务不断深化与增值。

参考文献：

[1] Nature. Big data: Science in the Petabyte Era [EB/OL].[2014-11-04]. http://www.nature.com/nature/journal/v455/n7209/edsumm/e080904-01.html.

[2] James Manyika, Michael Chui, et al .Big data: The next frontier for innovation, competition, and productivity [OL].[2014-11-04].http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation.

[3] Zhou AY. Data intensive computing-challenges of data management techniques [J].Communications of CCF,2009,5 (7):50-53 (in Chinese with English abstract).

[4] [英] 维克托·迈尔-舍恩伯格,等.大数据时代:生活、工作与思维的大变革[M].杭州:浙江人民出版社,2013.

[5] 2014 淘宝双 11 直播间：双 11 淘宝销售额或达 600 亿[EB/OL].[2014-11-12].http://www.sxdaily.com.cn/n/2014/1112/c73-5553611-24.html.

[6] 张晓林. 从数字图书馆到 E-Knowledge 机制 [J]. 中国图书馆学报,2005,31(4):5-10.

[7] The White House. Obama administration unveils “big data” initiative: Announces $200 million in new R&D investments[EB/OL].[2014-11-04].http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf.

[8] Digital Book world. New Start-Up Aims to Be Google Analytics for E-Books [EB/OL].[2014-03-10].http://www.digitalbookworld.com/2012/new-start-up-aims-to-be-google-analytics-for-e-books/.

[9] The New York Times.Harvard Releases Big Data for Books[EB/OL].[2014-8-11].http://bits.Blogs. ny-times.com/2012/04/24/Harvard-releases-big-data-for-books/.

[10] 张兴旺.图书馆大数据体系构建的学术环境和战略思考[J].情报资料工作,2013,(2):12-17.

[11] 郭自宽,张兴旺,麦范金.大数据生态系统在图书馆中的应用[J].情报资料工作,2013,(2):23-28.

[12] 吴志荣.文献发现:当代图书馆的重要命题[J].图书馆杂志,2013(9):4-7.

[13] 韩翠峰.大数据时代图书馆的服务创新与发展[J].图书馆,2013,(1):121-122.

[14] 李国杰.大数据研究:未来科技及经济社会发展的重大战略领域[J].中国科学院院刊,2012,(6):647-657.

[15] Kannan R,Andres F. Digital Library for Mulsemedia Content Management[C].In:Proceedings of the International Conference on Management of Emergent Digital EcoSystems.2010:275-276.

[16] 北京历史地理[EB/OL].[2012-01-29].http://bjhg.lib.pku.edu.cn/.

[17] Digital Collections & Services [EB/OL].[2014-01-29]. http://www.loc.gov/library/libarch-digital.html.

[18] Chinese Rubbings Collection [EB/OL].[2014-01-29].http://nrs.harvard.edu/urn-3:hul.eresource:chinrubc.

[19] Audrey Watters. Strata Week: Harvard library releases big data for its books: Harvard offers big data for books, cloudera´s new hadoop distribution, splunk goes public [EB/OL].[2014-08-11].http://radar.oreilly.com/2014/04/harvard-book-data-cloudera-hadoop -splunk-ipo.html.

[20] Cohen J, Dolan B, Dunlap M, Hellerstein JM, Welton C. MAD skills: New analysis practices for big data[J].PVLDB,2009,2(2):1481-1492.