大数据背景下图书馆的数据存储策略优化研究

刘 瑜
（四川文理学院图书馆，四川达州 635000）

摘要：在大数据时代，数据的爆炸式增长对图书馆数据存储能力提出了严峻的挑战。论文立足于图书馆数据存储的实际操作层面对大数据存储的可能性路径以及现阶段图书馆进行大数据存储的适用方案进行探讨，认为图书馆有责任尽可能多地存储数据。对于非重要数据可利用云系统进行外挂存储，对于一般数据可通过构建图书馆联盟来实现馆际互借，对于特色数据可采用馆内存储，并用大数据压缩技术来提升存储空间，从而缓解大数据对单个图书馆存储能力的压力。

关键词：大数据；图书馆；数据存储；非结构数据
中图分类号：G250 **文献标识码：**A **DOI：**10.3772/j.issn.1674-1544.2015.06.014

Research On the Optimization Strategy of Library Storing Data at the Age of Big Data

Liu Yu
(Library , the Sichuan University of Arts and Science, Dazhou 635000)

Abstract: It is a serious challenge to library's data storage that data volumes are exploding at the age of Big Data. Based on the actual operation of library's work, this paper discusses on the possibility paths and the application schemes of the data storage at present. The author thinks that Library has the responsibility to store data as much as possible. In storage mode, these data what is not important can been external stored through the cloud system, these general data can been achieved by inter-library loan services within library consortium, these characteristic data can been stored through Library memory and can been compressed by the Big Data technology, so as to relieve the pressure which Big Data bring to a single library.

Keywords: big data, library, data storage, unstructured data

在大数据时代，数据的爆炸式增长超出了人类的想象，以知识存储为基本责任的图书馆应该如何应对呢？受技术瓶颈和经费短缺的限制，大多数图书馆要想全面升级换代现有数据库和提高现有存储容量是不可能的。在这种情况下，图书馆学情报学界往往把数据存储解决方案寄托于大数据存储技术的突破。事实上，这种研究是严重脱离和滞后于图书馆现实的，最终后果将会导致在大数据带来的巨大契机面前消极无为。鉴于此，本文立足图书馆数据存储的实际操作层面来

作者简介：刘瑜（1975—），女，法学硕士，四川文理学院图书馆员，研究方向：大数据与图书馆转型研究和图书馆大学生思想政治教育研究。

基金项目：四川文理学院高层次人才科研资助项目“大数据与图书馆转型研究”（2014XG001）。
收稿时间：2015年3月19日。

探讨这一问题。

1　大数据对图书馆数据存储能力的挑战

1.1　大数据特征分析

尽管不同业界对“大数据”（Big Data）的定义充满分歧，但都强调首先必须从数量维度去理解“大数据”的“大”。也就是说，大数据的首要特征就是数量大。国外学者一般把大数据的量级限定在10TB到1PB之间[1]。我国学者一般不主张对所谓的“大”做具体限定。徐子沛的观点最具代表性，“一般认为，大数据的数量级应该是‘太字节’的，我们也并不需要给‘什么是大’定出一个具体的‘尺寸’，因为随着技术的进步，这个尺寸本身还在不断地增大。此外，对于各个不同的领域，‘大’的定义也是不同的，无需统一”[2]。张兴旺曾经通过梳理大数据的发展历程展示了“大”的语义。认为：GB级别的数据是“超大规模数据”（Very Large Data），TB级别的数据是“海量数据”（Massive data），只有PB级别的数据才称得上是“大数据”[3]。鉴于此，从纵向维度来把握大数据的基本特征是非常可取的，因为它可以充分彰显大数据增长量大的特征。根据国际数据公司IDC监测，仅在2010年，人类生产的数据量是1.4ZB（引注：1ZB=1024PB），2011年增长到1.8ZB，到2012年达到2.9ZB[4]。如果数据按照目前的60%速度增长，这就意味着全球数据量大约每两年翻一番；预计到2020年，全球将拥有35 ZB的数据，数据增长近30倍。数据呈几何级数增长，完全超出人类世界目前的IT架构和存储能力所能承载的范围。为此，全球权威的IT 研究与顾问咨询公司 Gartner就将大数据定义为“在一个或多个维度上超出传统信息技术的处理能力的极端信息管理和处理问题”[5]；维基百科也强调大数据是“无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理的数据集合”[6]。不言而喻，数据的不断涌现与人类掌控数据能力有限性之间的矛盾关系始终是大数据时代的基本矛盾关系。因此，图书馆馆际之间的竞争是馆藏资源和空间建筑方面竞争的传统观念已经不适应大数据时代发展的需求了。

1.2　图书馆存储能力面临的挑战

（1）图书馆不可能存储所有数据

大数据时代，基本矛盾关系决定了图书馆不可能存储所有数据。该结论不仅对单个图书馆存储容量适用，而且对整个图书馆事业的存储能力也适用。在这种境况下，要求任一图书馆只能根据自己的读者需求、办馆特点和发展定位，有选择性地存储“有意义”的数据。但这也并非意味着对其他数据可以置之不理，因为任何数据都是有潜在价值的，只是针对不同对象而言。

（2）图书馆不易存储非结构数据

大数据之所以“大”关键在于半结构数据和非结构数据（以下统称为“非结构数据”）飞速增长。

第一，非结构数据增长量大。按照大数据发生学的解释，信息技术开发与应用的高度融合直接催生出云计算、移动网络和社交平台，而它们的出现又为人们能够随时随地地利用智能手机、平板电脑或导航系统等现代通信工具去生成、发送和获取数据提供了便利和平台，于是就形成了大数据赖以存在的生态环境[7]。据IDC 2012年《数据世界》报告显示，全球结构化数据增长率为32%，非结构数据增长率是65%，至2012年，结构数据的数量略高于互联网数据总量的10%[8]。

第二，非结构数据没有相应的数据库可供存储。“非结构数据”特指那种非线性的数据类型，它主要相对于以“事务”为中心而建构起来的关系数据，亦即“结构数据”而言的。在“小数据”时代，结构数据占统治地位，对应的IT架构是“关系型数据库”；在大数据时代，传统的关系型数据库已经无法直接存储这些带有异质构造性质的“非结构数据”。

第三，非结构数据占用的存储空间较大。非结构数据主要是与传感器、图像、视频、音频、微博、微信、帖子、点击等数据紧密联系在一起的，完全以“碎片”的形式存在于物理空间。在一般情况下，它们占用的物理空间都非常大且不易整理。

（3）图书馆不得不存储非结构数据

目前，图书馆还习惯于存储结构数据，但非结构数据的所占有比例远远高于结构数据的比例。而这些来自人类日常生活世界且占主要份额的非结构数据同样是人类生存体验、社交对话和情感互动的缩影，同样是人类智慧的“呈现者”，因而在很大程度上更富有“隐性知识”的价值和意义。从人类生存论上看，大数据时代已经悄然来临，不管你是否意识到都已经身临之中，都必须借助数据与世界“打交道”。可见，非结构数据具有不可或缺和不可忽视的价值，所以大数据背景下的图书馆不仅要关注结构数据，而且更应该重视非结构数据。

当前，图书馆要与时俱进地进入大数据视域，就会感受到大数据对图书馆存储容量的压力。以大数据视野审视图书馆的建设，就会发现当今图书馆数据存储的难题：一方面是不能完全把控大数据但又不得不试着去存储它；另一方面是如果要存储它，又不得不面临半结构数据不易存储的问题。

2　图书馆大数据存储的可能性路径

在大数据时代，图书馆的数据存储问题主要是“怎样对非结构数据进行存储”。解决该问题，学理上有两种可能路径：一是寄托信息工程技术领域的突破，能够构建出与非结构数据性质相适应或相兼容的数据库，亦即IT界所说的“非关系型数据库”；二是借用可资利用的“大数据技术”，通过专业化的数据处理，把半结构数据和非结构数据勾连、转换或改造为结构型数据，使之与现存的关系型数据库同质化。

2.1　坐等数据存储技术成熟是一项消极被动的路径选择

图书馆界学人大多主张走第一条路径，但他们却又无时无刻不在感叹图书馆的基础设施建设的滞后。按照这样的思维逻辑推演可以预见，在大数据带来的巨大契机面前必将碌碌无为。原因有3点：一是非关系型数据库建构观念能否转变为现实还是一个未知数。很多IT界权威机构以及资深人士预计，还需要再经历10年以上时间，大数据存储技术的应用前景才能基本清晰[9]。如果把解决问题的方案寄托于一个似是而非的设想，那是非常不可取的。二是建构出的非关系型数据库也不一定能解决所有数据存储问题。如前所述，大数据之“大”是因为数据总量超出人类存储、管理和处理能力。该矛盾关系始终构成大数据时代存在的现实基础，因为假如当人类有能力把控所有数据之时，数据也就无所谓“大”了。三是非结构数据还不是完全意义上的知识，不能直接运用。即使人类拥有了非关系型数据库，如果不加整理地把全部非结构数据都装了进去，仍然不是知识形态的数据。图书馆的基本职责不仅要存储知识，而且还要提供知识服务。“传统意义上的数据、信息和知识具有完全不同的概念。数据是信息的载体、信息是有背景的数据，而知识是经过人类的归纳和整理，呈现规律的信息。[10]”这也就是说，数据要成为知识还需要一个复杂的转换过程。按照当代图书情报学的观念来讲，图书馆存储的文献资源如果是零利用率则属于资源的浪费[11]。这不仅对纸质书籍适用，而且对数据也同样适用。这即是说，大数据给图书馆的数据存储管理提出了更高的要求。首先，要有针对性地采撷、提取、挖掘能够满足读者需求的数据。这与大数据本质完全契合，即我们不可能存储全部数据，但可以有选择地存储有用数据。其次，要把特定的非结构数据“知识形态化”。最后，要把知识形态化了的数据方便快捷地推送给读者利用。

2.2　利用现有技术推进数据知识形态化是一项积极主动的路径选择

数据知识形态化就是要把非结构数据勾连、转换或改造为结构型数据。这就进入到第二条路径的语境，也是一条非常符合图书馆实际情况的大数据存储的路径选择。同时，图书馆经过多年的信息化和智能化建设，已经具有了一定的大数据存储管理的特征[12]。这主要表现在以下4个方面：一是图书馆馆藏文献资源种类繁多，不仅有纸质印刷品资源、数据库资源、光盘资源等结构

化数据，也有大量的读者信息、服务日志等大量的非结构数据；二是图书馆存储信息资源的容量也在迅速增长。单个图书馆的资源总量也许不能达到PB量级，但全国所有图书馆的资源加起来却是ZB量级的，如全国文化共享工程的资源总量就达到108TB。由此可以看到：我们完全可以统筹安排整个图书馆行业的存储空间，再通过分布式共享，消除在数据存储过程中产生的重复数据，从而最大限度地扩展存储空间。三是根据读者需求，图书馆的采访、编辑工作也出现个性化、学科化和团队化的趋势。四是图书馆自动化水平进入一个新的水平，不仅大量读者行为信息被记载和统计，而且还能实现读者服务信息被即时传送。总之，经过多年实践积累，图书馆已经具备大数据存储管理的一些经验和优势，我们在大数据面前并非束手无策。另外，信息技术发展已经为大数据存储提供了一系列相对实用的工具系统。在数据存储方式方面，已为广大图书情报学研究者熟知的有：网络附加存储（Network Attached Storage，简称NAS）、存储域网络存储（Storage Area Network，简称SAN）、直接外挂存储（Direct Attached Storage，简称DAS）。在非结构数据处理软件方面，主要有EMC、Hadoop和Datameer。数据转换工具主要有语义关联分析、网络分析、聚类分析、可视化分析、数据融合和数据集成等。充分利用这些现有的大数据技术，再结合已经积累的大数据存储经验，我们完全可以在大数据存储方面大展身手。

3　图书馆大数据存储方案

3.1　利用云系统存储非重要数据

大数据首先是对图书馆存储的硬件设施，尤其数据存储容量提出了严峻的挑战。为应对这一难题，有些图书馆学情报学研究者把希望全部寄托于“云系统”技术的推广应用。云系统（Cloud Computing）的核心思想是“分布式共享”，具有动态性、开放性、自治性、可靠性、可用性等特征[13]。在具体操作上，图书馆员们只需将相关数据输入到“云端”，就可以自由地在上面进行存储、访问、修改、反馈或提取。云系统无限地扩展了图书馆的存储容量，而且馆员还无须为技术问题而烦恼（一切技术操作都有云服务商来解决）。但是，云系统在给予我们便利的同时也带来高技术转让费以及知识产权、技术标准、信息安全、管理体制等方面的难题[14]，所以，我们在使用时应该谨慎待之。目前，最稳妥的办法就是把图书馆的数据、信息、知识资源进行分门别类处理。可以根据知识产权、技术标准和信息安全的不同程度把所有数据细分成不同的、特点明确的类型，然后按照其类型采取相应的存储方式。具体说来，可以把重要数据如特色数据、有知识产权要求的数据和保密程度高的数据进行馆内存储，把“非重要数据”进行外挂存储。这样，既能够在一定程度上规避可能的风险，又能够最大限度地提升图书馆自身的存储容量。

3.2　通过馆际联盟存储次重要数据

由于单个图书馆存储设备容量始终是有限的，而读者需求数据又呈几何级数增长，即使仅存储核心数据也会很快“爆棚”。比如当前，综合图书馆每年必需数据增量大约为20TB，如此浩瀚的数据量，对于一座拥有100 TB存储容量的大型图书馆也仅能满足5年左右的存储需要。因此，图书馆之间寻求合建数据存储库来实现科学数据的收集、共享和服务也成为必然选择。近10年，图书馆为解决文献资源不全而构建的“馆际互借系统”和为形成文献资源互补优势而建立的图书馆联盟都为这种分布式共享奠定了物质基础，提供了技术支撑，比如北京地区高等教育文献保障系统（BALIS）就是在北京地区高校图工委的统一领导下建构的。它采用集中式门户平台建设和分布式存储相结合的方式，不仅提高了文献资源的利用率，而且减少了单个高校图书馆的存储容积。这种模式完全可以借鉴到大数据图书馆联盟建设上。对图书馆来说，如果要想提高数据存储容量而又能规避云系统带来的安全威胁，组建或加入图书馆联盟是可行的策略之一。 这是因为图书馆之间具有很多天然的同一性，图书馆在应对潜在风险方面的立场基本一致。图书馆馆

际之间合作的深化还有利于克服重复建设以及资源浪费方面的弊端。这一解决措施反映在图书馆的软硬件建设上，就必须由过去追求高端服务设施向中低端软硬件基础设施构建的大规模分布式计算机群集转变，将分块、分类的大数据复制到集群服务器节点上进行处理[15]。

3.3　利用馆内存储来建设自己的特色数据库

图书馆行业的内部竞争决定了任何一个图书馆都不愿意把自己具有核心竞争力的特色数据通过链接方式予以共享。与之相反，各图书馆都在竞相自建独具特色的数据库。而特色数据库具有高度的可靠性和安全性要求，但存储周期长、数据类型多、数据量大，这同样会对内部存储容量构成巨大的挑战。在图书馆的实际管理经验中，经常是采用整理碎片的技术来提高存储空间利用率和数据查询效率，这对小容量操作切实可行，但对于大型数据系统却是远远不够的。鉴于目前图书馆存储设备容量利用率不到50%的现实情况[16]，提高数据库存储最有效的办法是运用大数据压缩技术[17]。相比较于自动精简配置技术和重复数据删除两项传统压缩技术，大数据压缩技术不仅兼顾了它们两者功能，而且功效更强大，它可以针对整个图书馆系统内两个或多个文件之间数据的相同性和相似性，通过分析比较，删除多余数据，达到数据压缩的目的。图书馆在自建特色数据库时，总会遇到大量的非结构数据。在进行“数据知识形态化”过程中，针对复杂、多样的非结构数据管理需求，可以结合OLTP、Datameer和Hadoop等IT技术给予解决。具体操作程序是：首先利用Datameer提供采集和读取不同类型数据库的平台，然后将“二次生成数据”植入Hadoop开源框架之中，凭借其提供的分析工具对数据进行可视化分析、预测性分析、智能语义分析，从而建立“名副其实”的语义引擎，最后把“三次生成数据”进行OLTP技术处理。这样，原初的非结构数据经过一套蕴含多重深度分析工具程序的改造之后，就能直接存储到关系数据库之中。当然，由于“数据知识形态化”程序的技术含量高，这必然给图书馆员提出了更高层次的业务能力要求。

4　结语

在大数据存储技术并不成熟的情况下，图书馆员们已经在具体的实践活动中摸索出一些大数据存储经验。作为这种感性经验的理论总结，本文力图指出，图书馆并非在海量数据面前无所作为；在现阶段，最切实可行的数据存储路径是利用现有的一些数据处理工具进行“数据知识化”处理工作；经济适用的数据存储方案是根据不同数据对本馆馆藏的重要程度，选择、分类并有针对性地进行外挂存储、馆际存储或馆内存储。当然，在研讨过程中不难发现，图书馆作为政府主导下的公益型事业，图书馆员在对大数据的认识上以及对大数据存储难题的探索上总面临内驱力不足的问题。这是图书馆学情报学界亟待解决的另一个问题。

参考文献

[1] Terence K. Big data, Big Future[J].Computer in Librar−ies,2012(6):21−22.

[2] 徐子沛.大数据:正在到来的数据革命,以及它如何改变政府、商业与我们的生活[M].桂林:广西师范大学出版社,2012:40.

[3] 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):13.

[4] IDC. The Digital Universe[EB/OL]. [2015−10−29]. http://www.emc.com/collateral/analyst−reports/idc−extracting−value−from−chose−ar.pdf.

[5] Gartner.What Is Big Data? [EB/OL].[2015−12−13]. http://www.gartner.com/it−glossary/big data/.

[6] Wikipedia.Big Data [EB/OL].[2015−12−13].https://it.wikipedia.org/wiki/Big−data/

[7] 金茵,储娟. 图书馆服务与发展[J].当代图书馆, 2013(3):44.

[8] 霍娜.非结构数据来袭[N].中国计算机报,2013−07−11(A24).

[9] Gartner. Gartner’s 2014 Hype Cycle for Emerging Technologies Identifies “Tip−ping Point” Technologies That Will Unlock Long−awaited− technology Scenarios [EB/OL]. [2015−12−13].http://www.gartner.com/it/page.jsp.id=2124315.

（下转第106页）

文件各自的特点，有基于兴趣的读者描述文件和基于行为的读者描述文件。读者描述文件可以用文件来组织，也可以用关系数据库或其他数据库来组织。目前，有一些个性化系统采用基于XML的RDF（Resource Definition Framework）来表达读者描述文件，并利用支持XML的数据库系统来存储读者描述文件，这样不仅利用了XML的优点，也保持了系统的性能。

3.4　读者兴趣模型构建

读者兴趣模型构建是从有关读者兴趣和行为的信息中归纳出可计算的读者兴趣模型的过程。可计算性是构建读者兴趣模型的基本要求。在个性化服务系统中的读者兴趣模型不是针对读者个体的一般性描述，而是一种面向算法的、具有特定数据结构的形式化的读者描述。读者建兴趣模型是个性化服务的基础和核心，无论何种形式的个性化服务都需要建立对读者的描述，然后才能据此对不同读者提供个性化服务。读者兴趣模型有读者手工定制模型和自动读者建立模型。手工定制模型由读者自己手工输入或选择，如读者自己输入感兴趣的关键词列表或感兴趣的栏目。自动读者建立模型是根据读者的浏览内容和浏览行为自动构建读者模型，建模过程无须读者主动提供信息，不需要读者参与，由系统自动完成。

4　结语

数字图书馆个性化服务在国内开展得比较晚，也比较少。对数字图书馆个性化服务缺乏系统的研究。数字图书馆个性化服务节省了读者的时间、提高了读者科研效率，必将受到读者的欢迎。数字图书馆引进个性化信息服务模式，是适应图书馆读者需求多样化的需求，是让图书馆的信息服务向更深、更广的方向发展，是提升图书馆读者服务质量的重要手段。信息社会中，个性化服务将成为图书馆读者服务的重要方向。数字图书馆应开展不同类型的个性化服务、为读者提供个性化信息，吸引更多的读者进入到图书馆，从而提升图书馆的社会地位和竞争力。

参考文献

[1] 文露霜.基于Web2.0的数字图书馆个性化信息服务模式研究[D].武汉:华中师范大学,2009.

[2] 陈玉梅.高校图书馆个性化服务问题与策略[J]. 企业导报,2011(6): 41−43.

[3] 刘阳.基于个性化学习的远程职业培训研究[D]. 沈阳:沈阳师范大学,2009.

[4] 周灵威.试析图书馆个性化信息服务研究的现状与问题[J].图书馆论坛, 2005,25(2): 160−162,132.

[5] 严根荣,朱毅洁.高校图书馆个性化服务探析[J].安徽文学, 2008, (10): 392−393.

[6] 陈亮.基于多Agent的网络信息服务的技术研究[D].武汉: 湖北大学, 2009.

[7] 张伟娜.馆藏中医古籍数字化共享平台的组建设想[J]. 医学信息, 2014, 27(1): 26−27.

[8] 张泯泯.基于自适应随机元胞自动机的数据挖掘技术[D]. 杭州:浙江大学,2004.

（上接第86页）

[10] 徐子沛.大数据及其成因[J].北京:科学与社会,2014(1):14.

[11] 彭凤,黄力军.高校图书馆文献资源建设组织体系构建[J].四川图书馆学报,2014(3):25.

[12] 杨海燕.大数据时代的图书馆服务浅析[J].图书与情报,2012(4):121−122.

[13] 张建勋,古志良,郑超.云计算研究进展综述[J].计算机应用研究,2012,27(2):429−433.

[14] 王晴.云计算大数据时代图书馆的挑战和机遇[J].公共图书馆,2013(1):48−49.

[15] 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):24.

[16] Peter Spitzforn, Pongracz Sennyey.A Vision for the Future of Academic Library Collections[J]. International Journal of the Book,2007(4):4.

[17] 罗彬,阳静,袁赟.数字图书馆中大数据存储的应用研究[J].科技与企业,2013(18):122.