文章编号：2095-2163(2020)08-0183-04 中图分类号：G250.7 文献标志码：A

智能革命下数据驱动的智慧图书馆建设分析

郑智泉[1]，杨 楠[2]
（1 贵州民族大学 传媒学院，贵阳 550025；2 贵州民族大学 数据科学与信息工程学院，贵阳 550025）

摘 要：本文介绍了大数据背景下计算机技术的应用领域，对智慧图书馆体系和服务模式的特点进行了描述。从图书馆数据资源特点和业务逻辑方面入手，分析了数据类型和现有技术难点。结合馆藏数据收录标准、隐私安全、数据存取、数据挖掘处理四个方面分析了基于大数据的智慧图书馆建设在未来可能面临的挑战。

关键词：大数据；智慧图书馆；图书馆服务模式

Analysis of theconstruction of the data-driven smart library against intelligent revolution

ZHENG Zhiquan[1], YANG Nan[2]
(1 College of Journalism and communication, Guizhou Minzu University, Guiyang 550025, China;
2 College of Data Science and Information Engineering, Guizhou Minzu University, Guiyang 550025, China)

【Abstract】 Nowadays, artificial Intelligence technology has shined based on big data, which have had a profound impact on the development of all walks of life.This paper first introduces the application field of computer technology under the background of big data, and also give descriptions of the characteristics of smart library system and service mode, which briefly expatiates on overseas and domestic research status. This essay also analyzes the data types and existing technical difficulties by starting with the characteristics and business logic of library data resources. Meantime, it indicates that the construction of smart library could face a significant challenge in the future by analysis from 4 aspects of collection data collection standards, privacy security, data access and data mining processing.

【Key words】 big data; smart library; library service model

0 引 言

近年来，各行各业动态产生数据的速度呈指数增长，数据规模达到了前所未有的水平。这得益于计算机软硬件的高速发展以及成本的快速下降，使得 20 世纪几乎不可能实现的人工智能技术，在今天成为了可能。尤其是 2010 年以后，安卓系统、IOS 系统等在商业上的成功，导致社交网络、移动互联网、传感器等方面的多源数据容量迅猛增长，人们开始使用各类便携式设备为生产、生活助力。时下，以海量数据为基础的人工智能算法在交通、医疗、投顾、VR/AR、无人飞行器、智能家居、日常购物和办公、个性化定制学习等方面都有广泛的应用。

对于个体学习而言，获取信息的途径越来越多元，面对的信息体量越来越大，如何制定更加高效的学习策略和知识体系的构建路径，在大数据时代显得尤为重要。信息传播过程中的迭代效应使得人们在网络上获取的内容，其核心价值在经过多级传播之后往往变的面目全非。在信息社会如何帮助人们构建更加牢固准确的知识体系，而不是碎片化、断章取义的内容，这一点尤为重要。将图书馆融入到以大数据为基础的现代网络体系中，通过计算机算法为每个人定制个性化学习策略，使之成为能够满足每个人个性化需求的服务体系，是本文要探究的目的。

1 研究现状和服务特点

欧美国家的大学图书馆、公共图书馆和博物馆最早提出智慧图书馆的概念[1]。将图书馆构建在大数据之上，结合最新的算法和硬件设备，对新型图书馆的服务体系与服务模式进行探索。构建新型图书馆服务体系需要高度聚合人、资源、技术、服务等元素[1]。目前，传统图书馆基础硬件设施难以满足智慧图书馆服务体系建设需求，各图书馆之间的数字资源建设缺乏统一的管理标准。此外，图书馆现有的服务体系仍是数据的增删查改，是以管理者为中心的、为用户机械提供图书咨询以及借还服务的

基金项目：贵州民族大学“部校共建”专项项目(GZMDBXSZM1908)。

作者简介：郑智泉(1990-)，男，硕士研究生，实验师，主要研究方向：统计模型与统计计算、网络安全；杨 楠(1997-)，女，硕士研究生，主要研究方向：海量数据统计与分析。

收稿日期：2020-06-13

管理模式。在服务模式探索方面,刘桂峰等人基于“WEB3.0”和”情报 3.0”的基本理念,以及图书馆数据来源和分析方法的差异,提出了图书馆“服务3.0”模式[2]。传统图书馆服务体系多采用被动接收问题的方式来为读者提供咨询,而管理人员则是响应式回答问题,有关数据的记录单一且并不全面。技术的进步从外部打破了原有的服务体制,读者不仅仅是被服务的对象,更是构建整个图书馆服务体系的参与者。便携式智能终端的普及为数据的爆发式增长提供了可能,而数据是构建新型服务体系所需技术的基石。系统通过数据统计分析、可视化分析、语义分析、预测性分析, 从海量数据中发现隐性的知识关联等功能[1]。

借助于新的技术,知识与知识之间将不再孤立。在学习的过程中,每个人接纳吸收内容的最佳方式有所不同,知识体系的理解顺序也稍有差异。当今社会,由于人们接收信息绝大多数来自网络,而网络中的内容质量参差不齐,在没有该领域比较完备的先验知识的前提下,多数人对于网络上的内容是没有判断能力的。图书馆如何借助计算机算法结合海量数据对每个个体进行有效分析,进而推荐学习路径和内容,成为构建智慧图书馆的难点。此外,关系型数据库已经无法满足建设智慧图书馆的需求,半结构化、非结构化数据占据了图书馆数据 85% 以上[4]。多源异构、分布广泛、动态增长、价值密度低、先有数据后有模式已经成为时下数据的主要特点[3],这些特点并不能采用传统的方法去分析它,也无法在合理的时间范围内得到期望的结果。

2 数据类型与技术发展

夏立新等人表示,图书馆服务的本质是围绕用户需求的知识服务[5]。不同的时代、不同的社会背景下,服务形式虽有不同,但知识服务的内涵却从来没有改变。对馆藏数据和用户网络行为数据进行深度挖掘和利用,通过自然语言处理、图像识别等技术让图书馆服务更加高效且智能化是研究者努力的方向。

狭义上的图书馆数据几乎是围绕图书馆这个机构本身而产生的,主要划分为 4 类,即馆藏数据、业务管理数据、服务数据、用户行为数据[1]。值得说明的是,这里的用户行为数据并不包含用户自身的网络行为数据,这些数据依旧来源于图书馆。从广义上来说,图书馆大数据还应该包含用户的网络记录。其中包括:社交平台记录、购物记录、娱乐倾向记录、作息时间记录、网络学习内容等,这些数据都将成为个性化定制判断的依据。图书馆由于自身的特殊性,其本身拥有的馆藏数据非常庞大,包含各种近现代书籍、古籍、杂志、字画和契约等。数字技术的发展又为图书馆增添了另一个庞大的数据分支,那就是各类实体资源的数据化。数据化后的馆藏资源主要分为 3 类,即图文类、影音类、虚拟现实(VR)/增强现实(AR)类。丰富的馆藏资源是影响图书馆读者数量的重要因素之一,伴随读者数量而来的是大量的业务管理数据和服务记录。如,资源建设使用情况、RFID 数据、关卡出入记录、检索与下载记录、咨询评价服务体系数据等。硬件成本的不断降低,使得获取这些数据变得异常廉价且容易。

(1)数据管理方面。以 MapReduce 为代表的非关系数据管理和分析技术异军突起,在扩展性、容错性和大规模并行处理方面均有不俗表现[6]。实时记录的海量数据、高性能计算机以及各类硬件终端设备相结合,使智能算法在语音识别、图文分析、数据实时关联分析等方面取得了突破性进展。硬件的支撑、算法的日趋完善、高速移动互联网的形成使得构建新型的智慧图书馆体系成为了可能。

(2)语音识别方面。一套完备的语音识别系统应包含前端的信号处理模块、中间的语义识别模块和自然语言处理模块。识别自然状态下的语音内容难点很多,在内容表示、语义理解、情感分析等方面需要做出适当判断。由此引发的研究策略包括知识图谱、对话管理、机器翻译等。自然语言处理 NLP 是利用计算机技术对人类语言进行自动分析和表征的方法及理论的总称[7]。对话系统通常也被叫作聊天机器人,或者基于自然语言的人机交互。通常分为两种:一种是面向特定任务的,目的是帮助用户完成特定的任务;一种是开放领域的,以聊天交流为主要目的。任务导向的对话系统可以完成类似预定酒店、提供餐厅信息和获取公交时间表等任务。这类系统通常依赖结构化的本体或者数据库,提供系统交谈所需要的领域知识;而开放领域对话不是以提供信息为目的,一般是以与用户交流的情感体验为目标[8]。

(3)图像识别方面。如何快速准确的构建数字资源库是扩充、完善馆藏资源的必经之路,而快速识别读者身份是提升服务体验与个性化推送的必要前提。由于图像并不能被计算机直接指认,机器的理解逻辑与人有所区别,如何建立图片与文字之间的映射关系成了交互过程中的核心。视频和图片在某种意义下属于同一类数据类别,而图像预处理和分割、特征提取和判断匹配等问题是处理过程中的常用手段。

(4)网络行为数据方面。分析网络行为数据可

以更好的进行个性化学习定制。而海量的信息环境、毫无规律和关联性可言的数据资源，无形中给用户带来巨大的学习负荷[5]。如何从中找到适合自身需要的内容，进而制定学习策略无疑是难上加难。知识内容环环相扣，知识体系的构筑节点也有先后之分。体系结构中的每个节点相互交错，彼此联系，节点的构筑顺序虽不唯一，却有最优解。从一个知识点所引发的相关知识有时会让人不知所措，如果去粗取精有针对性地解决当下最为紧迫的知识诉求尤为重要。

(5)数据存储及处理方面。存储和计算海量数据所导致的内存溢出和所需时间是人们不得不考虑的问题，对数据本身的度量引发了处理方式的探索。应用场景的不同导致多样的数据类型产生，但数据处理方式相对固定，有批处理和流式处理两种。2004 年谷歌公司提出的 MapReduce 编程模型是最具代表性的批处理模式[9]。针对于每天动态生成的海量数据，两年后又提出云计算的概念。这不仅解决了大数据处理过程中单台设计计算能力不足的难题，同时也催生了以 Hadoop 为代表的一系列云计算平台。是一个包含文件系统、数据库、数据处理等功能模块在内的完整生态系统，多功能模块的集合让其成为处理大数据的利器。在解决了编程模型和计算能力不足的问题后，数据的存储方式也需要做相应变革。近年来，以谷歌、亚马逊、雅虎等公司为代表的企业或组织均研发了适用于大数据的数据库模型。其中，Bigtable、Dynamo、PNUTS 均是典型代表，这些系统在商业上的成功引发了新时代人对于非关系型数据库的深入思考以及关系型数据库在大数据时代的应有之义。目前对于非关系型数据库的认识，政府、企业、学界和各类组织并没有制定统一的标准，只要能够支持海量的数据存储需求，能够保证数据的一致性，且能够方便高效的利用即可。

3　面临的挑战

知识共享是图书馆的社会职能之一，而馆藏数据收录标准没有统一，各图书馆之间管理模式、管理流程存在差异化，这些因素降低了图书馆之间数字资源共享的可能性。将智慧图书馆体系中的数字资源库与应用层实现有效分离，是解决这一问题的良好途径。由于图书馆之间采用不同的硬件设备，使用不同的数据库软件，在现有的基础上，实现馆藏资源与应用的分离，难度非常大。图书馆馆藏资源数据冗余问题严重，这很大程度上不是来源于技术上的难度，而是由系统差异、地域差异等因素导致的。现阶段，我国也在深入推进图书馆数字资源共享以减少数据冗余问题，但其中面临的挑战和不可预测因素也非常多。

信息安全和隐私问题是人们长久关注的话题之一，实时记录的数据内容如果不能进行合理的监管，后果将是灾难性的。移动终端的普及让数据获取的方式更加便捷，内容也更加详实。有些人可能会刻意隐瞒或规避自身的敏感数据，这在以前或许可行，但在大数据时代，如果规避自身敏感数据，比如关闭手机定位功能，导致的直接后果就是该功能无法正常使用。数据的公开和隐私保护在大数据时代是一个矛盾的共同体，监管这些数据在技术上、伦理上都会是一个难题。在网络的传输过程中，如何保证个体的偏好数据不被网络黑客获取，如何防止个人偏好数据成为信息贩卖者攫取利益的工具，这些都是需要考虑的问题。实际生活中，用户的数据会在不经意间主动或被动上传，妥善保管和利用这些数据会避免对用户本身造成损害。有文献表示，隐私保护可细分为位置信息保护、标识符匿名保护、连接关系匿名保护等[10]。大数据时代数据的动态性特征以及数据本身的预测性质，让原有的静态数据信息防护技术显得无能为力，如何在复杂的环境中实现数据的有效利用和信息安全是异常严峻的挑战。

一般情况下，非关系型数据库系统中都存储着海量的数据资源，这些数据是无法全部持续停留在内存当中的，所以数据的读取方式一般都面向磁盘来进行读取[11]。此外，分布式数据存储机制效仿计算机内部“磁盘-内存-缓存”模式，然后供中央处理器处理数据的三级存储策略，这类缓存方式的核心需要确定热点数据。目前，多数云端存储框架采用分布式文件存储系统，数据存储模型采取键值对的方式进行，数据间的关联度相对弱化，如何快速的在这些海量数据中查找所需内容是问题的关键所在。在本地图书馆，如何构建硬件系统缓存本地阅览者热点数据，以快捷处理本地事务是重要的研究方向。

自然语言处理、图像识别正确率依然无法达到期望水平。如何让机器像人一样思考并解决问题，是人工智能算法面临的巨大挑战。基于海量数据的深度学习等算法可以让机器变得越来越智能，但就目前而言仍然没有实现真正意义上的拟人化。人们的自然交流方式场景很复杂，不同环境下语言内容的实质含义会有所差异，如何从众多差异中挑选最合适的映射关系并不容易。以中文为例，一词多义、

(下转第 191 页)