大数据分析技术在文化资源管理中的应用

◆ 姜念云[1] 张松海[2] 谢 夏[3]
1. 北京科学学研究中心,北京 100089;
2. 清华大学,北京 100084;
3. 华中科技大学,武汉 430074

摘 要 文化资源是体现一个国家文化实力的核心要素,也是国家文化及文化产业发展的基础和源头。我国对各类物质和非物质文化资源数字化工作的开展,为我们利用大数据分析等先进技术,加强对中华文化的充分认知和深入挖掘利用提供了前所未有的契机和条件。本文利用大数据分析等技术手段,对我国如何加强文化资源管理的总体思路、技术框架和有关对策措施提出了建议。

关键词: 大数据 文化资源 数据挖掘 数据分析
中图分类号: TP315 **文献标识码:** A
文章编号: 1009-2412(2014)01-0017-04
DOI: 10.3969/j.issn.1009-2412.2014.01.003

文化资源是体现一个国家文化实力的核心要素,也是国家文化及文化产业发展的基础和源头。我国对各类物质和非物质文化资源数字化工作的开展,为我们利用大数据分析等先进技术,加强对中华文化的充分认知和深入挖掘利用提供了前所未有的契机和条件。2013 年 7 月 23 日,"西苑沙龙"举办了主题为"大数据技术在文化资源管理中的应用——需求与技术体系"的学术讨论会,本文结合此次讨论会对我国利用大数据分析等技术手段,加强文化资源管理的总体思路、技术框架和有关对策措施提出了建议。

收稿日期: 2013-9-30 **修回日期:** 2013-12-10

本文作者: 姜念云,研究员,nyjiang@htrdc.com;张松海,副教授;谢 夏,副教授,xiex@htrdc.com。

本报告是科技部发展与改革专项课题"文化科技创新战略研究"成果之一,同时也是"大数据技术在文化资源管理中的应用:需求与技术体系"西苑沙龙的研讨成果,沙龙邀请的文化与科技有关领域的专家有:关健、许鲁、孙一钢、孙茂松、李松、张加万、罗静、金海、周志华、徐波、曾成钢。

一、关于大数据与大数据分析技术

1. 关于大数据

所谓大数据(Big Data)是指数据量大。但究竟怎样的量才算大,目前并没有统一的定义。一般认为,大数据的数量级至少应该达"太字节"(Terabyte,TB)以上。因为达到了这个量级以上的数据,利用现有 IT 技术和软硬件工具将难以实现在可容忍的时间内,对其进行有效的感知、获取、管理、处理和利用,必须要开发新的数据管理和处理软硬件技术,才能满足应用需求。

除了数据量浩大外,大数据还有两个特点:一是模态繁多,包括结构化数据、半结构化数据和非结构化数据;二是生成快速,大数据往往以数据流的形式动态、快速地产生,具有很强的时效性,用户只有把握好对数据流的掌控才能有效地利用这些数据,充分挖掘其中的价值。

近年来,随着互联网、物联网、云计算和三网融合等技术的发展,大数据及其挖掘利用问题,成为了产业界、学术界与政府部门各方面关注的热门话题,并正在从不同方面促进着我们的生活、工作和思维方式的改变。

如何加强其中具有文化内涵和特征的大数据的利用,特别是从提高对各类文化资源管理和利用水平的角度,是需要我们进一步加以关注的问题。

2. 关于大数据分析技术

大数据技术可分成大数据分析技术、大数据工程、大数据科学和大数据应用等领域。目前人们谈论最多的是大数据分析技术和大数据应用。工程和科学问题尚未被重视。大数据工程指大数据的规划建设运营管理的系统工程;大数据科学关注大数据网络发展和运营过程中发现和验证大数据的规律及其与自然和社会活动之间的关系。

相比于传统的数据库应用，大数据具有数据类型庞杂、查询分析困难等特点。《计算机学报》刊登的“架构大数据：挑战、现状与展望”一文，列举了大数据分析平台需要具备的几个重要特性，对当前的主流实现平台——并行数据库、MapReduce 及基于两者的混合架构进行了分析归纳，指出了各自的优势及不足，同时也对各个方向的研究现状及作者在大数据分析方面的努力进行了介绍，对未来研究做了展望。

适用于大数据的技术，包括大规模并行处理(MPP)数据库、数据挖掘电网、分布式文件系统、分布式数据库、云计算平台、互联网和可扩展的存储系统。为了开展大数据分析，中文信息处理、模式识别、知识挖掘等计算机信息处理技术以及相关的软硬件系统，都是其重要的核心基础技术。

二、关于文化资源管理

按照维基百科的解释①，所谓文化资源管理(Cultural Resource Management，CRM)是针对任何的文化相关资产的管理，主要包括历史的、技术的、社会的、建筑的或科学价值的文化遗产等，也包括当代的、创新的科技与文化资产。

因此，对于一个国家和民族来讲，文化资源是其文明发展历史过程中沉积形成的独有资产，具有唯一性和不可扩展等特点。因此，具有不可估量的文化、经济和社会价值和意义，是代表一个国家文化软实力的核心内容和象征要素，也是各类文化艺术产品创作的基础资料和源泉。所以，我们应该从战略的高度来重视文化资源的管理、保护和利用问题。

文化资源包含了多种形式和种类。从总体上可分为有形和无形资产，即物质与非物质两大类。也可从可再生和不可再生文化资源、历史和当代文化资源等多个角度进行分类。按照我国现行关于文化行业的分类，文化资源管理环节涉及文化遗产保护服务、文化研究、社团服务、图书馆与档案馆等子类。

三、加强文化资源管理对大数据分析技术的需求

随着各类数字化文化资源信息的不断产生，各类数字化文化资源库的不断建立与完善，在客观上为我们建立了一个庞大的、具有大数据特征的数据库和资源库。这就为我们进一步利用大数据分析等先进的信息技术手段，实现对这些文化资源信息的整合、梳理、分析和凝练提供了前所未有的基础和条件。

文化资源大数据在总体上可分为两类，即新生文化资源大数据和基于数字化的历史文化资源大数据。从文化资源管理角度看，这两类大数据都存在并具有很大的利用价值。

1. 新生文化资源大数据

根据数据来源的不同，新生文化资源大数据可分为随机信息和行业信息两大类。

其中随机信息是指基于互联网、物联网用户的大量网络搜索、下载、点击、上传等随机产生的大量多形态数据，可称为非结构化数据或随机大数据。对这些数据进行挖掘分析的一个重要目的就是对文化消费行为的分析。通过对不同互联网用户群体的文化消费特点和偏好的分析，将有利于更全面地了解各类文化产品、文化活动的市场需求，更有针对性地开发和创作相关内容、形式的文化产品，以满足各类消费者的需要，这对于提高文化产业的生产效率是具有重要意义的。

另一类则是按照一定的计划和规则，有意识地将各类文化产品分类汇集而产生的结构化数据或有序的大数据，如媒体资源库、数字出版库等。这类数据有约定俗成的格式规范，对于各类文化信息服务、历史文献管理及研究等都具有很大利用价值。

2. 基于数字化的历史文化资源大数据

这一类文化资源大数据是有计划地对各类历史文化资源数字化所形成的大数据信息。对这类数据的有效管理、充分挖掘和利用，或许是大数据及其分析技术更为重要的应用角度和需求。

随着数字化技术在文化资源管理中的应用，各类博物馆、图书馆以及其他社会组织，都在对各类物质与非物质文化遗产开展数字化保护工作，以便更好地实现对历史文化资源的保护、保存和利用。

由此，不但可以大大提高我们对于中华文化内涵、特点和历史的研究效率，更有可能实现与得到很

① http://zh.wikipedia.org/zh/文化资源管理

多仅依靠传统的研究方法所无法得到的，甚至难以想象的效果和结果。现在国际上已有利用大数据分析技术进行画作鉴别、古文献修复、历史文物分析等工作的先例，并取得了惊人的成果，如梵高、勃鲁盖尔等大师画作鉴别精准度达到了95%以上，而“死海古卷”机器自动修复的效率已经与数百人类专家过去一个世纪的成果相当。

大数据分析技术在历史文化分析研究的成果，对于我们进一步加深对中华民族文明发展历史的认知，辨识中华文化“基因”、延续文脉，明确我国文化建设应加强保护、传承和对外传播的重点内容，制定国家文化发展战略具有重要的意义。

四、建立面向历史文化资源管理应用的大数据分析系统

1. 现状与问题

中华文明几千年发展的历史为我们留下了丰富的文化遗产和资源。目前，我国很多博物馆、图书馆以及非物质文化遗产保护组织与部门，正在以不同方式、为不同的应用目的，开展着对各类历史文化资源的数字化工作，客观上形成了一个前所未有的、难得的中华文化资源大数据汇集。但由于这些数字化资源分散在不同的单位和部门，基本没有统一的格式标准，形成了一系列新的“信息孤岛”，难以充分发挥其应有的作用。

如何在现行体制下，通过一定的技术手段以及适宜的共享共建机制，构建一个实际或虚拟的数据交汇中心或平台，整合各类数据资源。并在此基础上，进一步发挥计算机中文信息处理、模式识别、知识挖掘等大数据分析技术的优势，面向各类文化研究、文化艺术创作、文化管理等用户提供更为优质和高效的信息服务，便成为了一个需要文化与科技相关领域共同探讨和推进的任务。

2. 面向历史文化资源管理的大数据分析系统框架

根据历史文化资源数据信息来源及其结构多元、主要应用领域、方式与用户多样等特点，面向历史文化资源管理的大数据分析系统，应是数据来源和应用端开放的、能够实现对数据提供相关主体和各类用户共建共享的数据管理平台，其框架结构如图1所示。

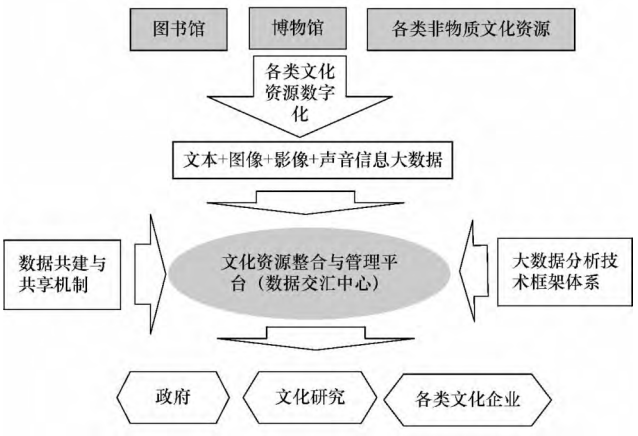

图1 基于大数据分析的数字文化资源汇集应用框架示意图

3. 数据交汇中心技术构架及需解决的关键问题

为建立开放共享的数据交汇中心，需要建立的技术系统构架如图2所示。其中需要解决的主要技术问题为：

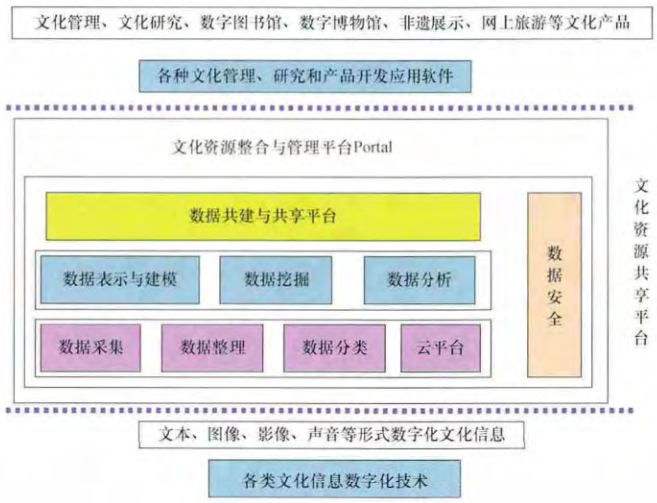

图2 数字文化资源大数据分析技术框架示意图

(1) 建立数字化标准

数字化标准是数字化设计遵守的一系列可重复使用的规则、导则或特性文件。文化大数据的采集方式多种多样，有视频、图像、纸质版本等。因此，我们需要建立统一的数字化标准，统一和规范数据格式，以便于存储和二次利用，主要包括：基础标准①、

① 术语与符号标准、分类与编码标准、数据表达标准等。该部分以采用国际、国家、行业相关现行标准为主，并针对需求进行适当剪裁。

数字化设计标准①、数字化采集与分析标准②、数字化测试与试验标准③、数字化管理标准④、数字化支撑标准⑤和信息技术安全标准⑥等。

由于数字化标准具有内容广泛、技术前瞻、更新周期短等特点，为保证数字化标准具有一定的先进性、前瞻性、时效性和可操作性，要采用“快速制定、动态修正、及时反馈”等措施，并采取工程技术、标准化和信息技术专业结合，以及标准制定与验证应用结合等策略，以加快标准制定，提高标准质量。

为了实现文化大数据信息全面数字化，除了制订设计全面数字化的标准与规范以外，还必须把标准与规范固化到有关软件、数据库和图形库中，才能保证系统正常运行。

(2) 设计统一的跨平台中间件

由于文化大数据采集的种类多样，采集的时间跨度大(例如，敦煌壁画的三维立体视像采集了约10年时间)，我们需要一个统一的跨平台中间件。这个跨平台中间件采用多层次服务模式：架构即服务(IaaS)、平台即服务(PaaS)和软件即服务(SaaS)，利用云计算环境提供服务。

(3) 实现高效的数据共享

虚拟化技术为资源的动态配置和高效共享提供了有效手段，不同的文化资源数据既可以在同一主机上进行处理，也可以存放于云端同一存储设备。云平台常采用流行的虚拟机管理器进行构建，具有高度的开放性，为上层应用提供了有力的支持。

(4) 加强对应用于文化资源大数据分析的基础技术的集成利用

中文信息处理、模式识别、知识挖掘等计算机信息处理技术，以及相关的软硬件系统，都是开展大数据分析的重要核心基础技术。加强对这些领域相关研究成果的集成应用，是实现大数据分析需求的重要技术基础。

文化资源的数据种类丰富、类型繁多，是典型的多源异构性数据。通过主动机器学习和群体计算多种数据处理方法的有效交互、融合与归纳是提高数据挖掘效率的重要方法。分布式计算引擎针对多源异构动态大数据挖掘的计算需求，支持交互处理、高效可扩展的数据分析。基于多源异构数据表示和能效优化的多粒度存储机制，以优化计算效率为目标，利用灵活的异构数据分布式存储与索引结构，面向通信密集型计算，能同时支持高效数据批量传输和消息频繁传递的分布式通信机制。

对于多媒体等数据流，分布式图计算引擎面向大图数据分析与挖掘，支持基于大规模消息传递，中间计算结果自适应保存，按需优化容错开销，实现高效性与可扩展性的统一，从而保证了文化大数据的有效利用。

五、相关对策与措施建议

加强各类数字文化资源的共享利用，不但需要有适宜的技术支撑，同时也需要面对应用需求和客观现实条件加强顶层设计，在组织管理模式和保障措施上有所创新。

1. 加强对我国文化数字资源管理的系统规划

由文化、科技、宣传主管部门共同组织、统筹协调，开展对我国文化数字资源管理和利用工作的系统研究，明确总体目标、任务和发展战略，提出有利于促进国家文化资源信息大数据管理利用的，由国家和社会相关机构共享共建的组织机制、商业模式和技术框架与标准。

2. 组织实施国家文化资源管理与共享专项

可依托“国家文化科技创新工程”，实施“国家文化资源管理与共享专项”，组织图书、文物、非遗保护以及各类文化企业、研究机构等有关单位和组织机构，共同围绕国家文化资源管理的发展战略和目标，开展基于大数据技术的各类历史文化数字资源库的建设、价值挖掘研究及其综合利用产品的开发。

3. 建立国家数字文化资源管理虚拟平台

建立面向数据提供方和应用方双向开放的“国家数字文化资源管理虚拟平台”。平台的决策管理可采用联盟机制，由各数据提供单位和重点应用部门等共同组成，负责制定文化资源数字化相关各类

(下转第27页)

① 文化大数据数字化定义标准、设计分析与优化标准、存储标准等。该部分重点反映综合优化技术、建模与仿真、设计知识管理、数字化协同设计等方面的标准需求。

② 数字化采集设计标准、数字化加工设计标准等。

③ 数字化测量与诊断标准、3D仿真试验标准、数字博物馆系统仿真试验标准、数字化应用试验评价标准等。

④ 面向用户的服务资源服务标准、文化大数据电子商务类标准、电子政务类标准等。

⑤ 应用平台构建标准、基础数据库标准、系统集成与接口标准、软件工程标准、计算机与网络标准等。

⑥ 文化大数据基础信息安全技术标准、系统与平台信息安全标准、信息安全测试与评估标准、信息安全法律法规等。

应用场景、技术指标及运营模式的需求判断，给出未来电力系统储能的技术需求和发展趋势。

（2）开展含规模化储能的电力系统基础理论和关键技术研究。包括规模化、多样化储能技术在电力系统应用中的建模仿真、系统分析、运行控制及规划等研究，不同应用场景下对储能载体、中间转换及电网接入的关键技术研究，以及储能技术在电力系统的适用性和评价技术研究。

（3）开展储能基础理论、关键材料和元器件、功率变换、系统集成和控制、多类型组合应用的研究。电池以及超级电容器重点解决规模化应用时的适用性、安全性、系统管理和评测表征等问题；抽水蓄能、压缩空气储能、储热、蓄冷和制氢重点解决提高效率和改善动态性能的问题；飞轮、超导储能重点解决基础材料等科学问题。

（香山科学会议　杨炳忻供稿）

Brief Introduction to Xiangshan Science Conferences of Nos. 466—470

（上接第20页）

标准及有关共建共享规则，共同投入，利益共享。平台的运营管理则可由第三方技术支持部门承担，主要负责在技术上为各类、各源文化资源大数据的整合、交汇和综合利用，以及安全维护、利益分配和知识产权保护等提供支撑。

4. 开展形式多样的应用示范

以“国家文化资源管理与共享专项”为引导、以“国家数字文化资源管理虚拟平台”为依托，选择若干重点文化主题，组织开展系统深入的文化研究，以进一步从系统上加深对中国文化发展历史和中华文明精髓的认知，为制定国家文化发展战略提供依据；充分发挥市场和各类文化企业的作用，开展基于中国历史文化资源、形式多样的文化和艺术产品的开发，为提升我国文化产品的文化内涵和中华文化的国际传播力、影响力提供支撑。

参考文献

[1] 侯经川，方静怡. 大数据时代的数据引证研究：进展与展望. 中国图书馆学报，2013，01：112—118

[2] 李宗桂. 重视优秀传统文化现代价值发掘和阐释，寻找核心价值观的历史支撑. 人民论坛，2013，10：72—73

[3] 孙红军，李　红. 基于大数据时代的战略管理研究——以文化产业为例. 绿色科技，No. 1，2014，01：207—210

[4] 吴祐昕，吴　波，麻　蕾. 互联网大数据挖掘与非遗活化研究. 新闻大学，2013，03：66—71，53

The Application of Big Data Mining in Culture Resource Management

Jiang Nianyun，*Zhang Songhai*，*Xie Xia*

1. Beijing Research Center for Science of Science，Beijing 100089；

2. Tsinghua University，Beijing 100084；

3. Huazhong University of Science & Technology，Wuhan 430074

Culture resource is key factor to support China's culture power，which also is the basement and fountainhead of China's culture and culture industry development. Nowadays，we focus on developing digital work on material culture and intangible cultural resource by big data analysis technology. It is an unprecedented opportunity and challenge for deep mining and understanding well about China's culture. This paper propose how to strength culture resource management and give some effective suggestions.

Keywords: big data；culture resource；data mining；data analysis