大数据处理技术在信息档案管理中的运用分析

宋玉婷

（曲阜市检验检测中心　山东　济宁　273135）

[摘　要]在科学技术的支持和帮助下，加上不断深入应用社会化地理信息服务，大幅度增加信息数据，同时加大管理这些档案数据难度。在计算机技术日益成熟的今天，大数据处理技术也被广泛应用其中。基于此，本文进一步论述常见大数据处理技术，同时在此前提下阐述大数据处理技术在信息档案管理中的应用，希望给有关机构提供参考与借鉴。

[关键词]大数据处理技术；信息档案管理；运用

【DOI】10.12252/j.issn.2096-6261.2021.07.550

引言

近年来，大数据得到飞速发展，其被广泛应用于各个方面，如经济发展、社会进步等，均获得重大成果。大数据处理技术具有明显优势，这是其在各个领域得到广泛应用的重要因素，但是其相关问题也逐渐凸显出来。在管理信息档案时，主要问题包括大量档案内容、较高维护成本等。由此看来，在信息档案管理中应用大数据处理技术具有必要性和重要性，这样才能很好解决相应问题。

一、信息档案管理现状和存在问题

现阶段，国家企事业单位是开展信息档案管理工作的主要单位，开展专项管理工作是资料档案管理部门职责。在信息产业不断发展与进步背景下，越来越多的智慧城市、数字城市被建设出来，档案管理中被纳入越来越多关联地理空间位置的文档。信息档案管理存在的问题主要是：首先，较大数据量。当前，主要运用两种方式管理档案，分别是纸质文件存档、电子文档存储，现阶段急需解决的问题就是怎样以需求为依据，更好存储和扩展这些文档。其次，具有复杂结构的文档数据。信息数据不仅包括结构化数据，还包括非结构化数据，数据检索具有较大难度，耗费较长时间。最后，较低的数据安全性。考虑数据具有庞大数量，复杂格式，不能实现统一集中存储目的，导致必须分散管理数据，不能保障数据安全性，促进数据孤岛的产生。上述问题给工作人员带来重大困扰，大数据处理技术的出现，能够对这些问题进行有效解决[1]。

二、大数据处理技术

大数据处理技术主要包括以下五个方面，具体分析是：

（一）HDFS文件管理系统

当前数据具有极其复杂性，在管理档案过程中最好应用Apache基金会开发的非结构化文档分布式文件系统，其主要部分是Hadoop。主要特点表现在以下方面：第一，较低的运用成本，可以在较低廉的硬件上应用，扩容系统同样应用较低成本，具有简单方便的实施步骤。第二，对具有较高吞吐量和并发访问的运用数据给予充分支持，同时被应用在超大数据集群中，并且能够对具有较大并发量的文件实施访问操作。第三，扩张性比较强。通过对MapPeduce的应用开展分布式运算工作，对计算机集簇实施数据配置和运算操作时，可以扩展这些集簇，形成数以千计的节点。由此看来，HDFS能够将档案管理中不能集中存储文档的问题很好解决。另一方面对文档、图片、音像等相关格式进行支持，同时在包括HIVE和Impala相关的结构化数据库的HDFS数据库前提下，实施查询和更新操作[2]。

（二）知识图谱

知识图谱是一种结构，其基础是图数据，根本是由节点与边构成的语义网络。知识图谱中节点可以将现实世界中存在的实体表现出来，各个条边是实体与实体间关系。知识图谱是一种呈现关系的方式。它作为一种关系网络，能够连接不同类型信息。知识图谱可以从关系立场角度分析问题，随着知识图谱的发展和进步，大部分公司在文档管理中应用知识图谱，能够提高管理文档关系的效率，同时获取文档知识点，更好支持和帮助开展后期文档管理工作。

（三）云存储

云存储是以云计算为基础，通过不断发展而形成的概念，这是一种全新网络存储技术，借助应用集群、网络技术、HDFS等有关功能，通过软件应用，集合网络中不同种类存储设备开展协调工作，共同对外存储数据与访问业务的系统。总之，云存储作为一种新技术，通过在云上放置储存资源，能够方便人们存储和提取。

（四）非结构化数据库

现阶段，非结构化文档是档案管理的主要文件，数据类型比较多，如影像、音频、文档等，应用传统主流数据库无法实现有效存储和应用这些非结构化数据目的。通过对数据库的应用，不仅可以存储文档内容，还可以对其内容进行分析，主要方式包括Key-value非结构数据库，其中应用最广泛的是Mongo DB。

（五）Elastic Search检索

此检索的英文简称是ES，开发人员是Java，以Restful Web接口为基础，给予分布式多用户全文搜索大力支持，这是现阶段十分流行的企业级别搜索引擎。文档检索是开展信息文档管理工作中难度最大的内容。ES特点包括搜索快速、零配置等，可以在HDFS和传统Java中得到应用，同时对并发访问给予支持，它集合的工具包括主流ICTCLAS、庖丁解牛等，同样能够检索中文文档分词，实现快速搜索信息档案目标

[3]。

三、大数据处理技术在信息档案管理中的应用

大数据处理技术具有重要意义，为了提高其综合成效，本文进一步探讨其在信息档案管理中的应用策略，具体分析是：

（一）信息档案资源大数据处理平台的构建

第一，通过利用Hadoop HDFS，将档案管理集群构建出来，考虑Hadoop HDFS对单一节点硬件资源提出较低要求，建设初期可以应用档案馆内已有硬件资源，将50-100节点的档案管理集群构建出来，通过集群中所有节点的协同工作，对数据实施存储与计算操作，后期以档案管理需求为依据随时扩展节点，从而使日益增长的档案数据对存储空间和计算性能的弹性需求进行满足。第二，通过对Hbase数据工厂的运用，处理大量半结构化、非结构化数据，促进结构化数据的生成，进而储存在Hadoop的数据库中。建筑HIVE数据仓库的布置，使管理结构化数据入库目的得以实现。联合使用HIVE和Hbase之后，能够集群式存储和管理结构化与非结构化两类地理信息档案数据。第三，借助MapReduce开展分布式计算工作，通过对Hadoop集群中并行计算的应用，使同步处理大量数据成为现实，最终形成具有多种功能的信息档案大数据处理平台[4]。

（二）信息档案资源的数据挖掘和知识发现

近几年，持续发展的信息数据获取方式，大幅度增加信息档案资源数据，加上这些年积累起来的信息数据档案，促进海量信息档案资源的形成，想要将其中蕴含的宝贵知识财富挖掘出来，仅仅依靠人工方式无法实现目标，必须运用技术，当前最有效的方法是大数据处理技术。

整个大数据处理的核心与关键是数据分析，怎样充分利用数据挖掘和知识发现研究分析大数据是大数据时代运用知识服务发展的重要问题，与此同时，是档案管理部门创新服务的有效途径。通过清洗、抽取、集成大量多元异构数据，并将其转换为容易分析形式，载入文件系统数据库分布式处理模型中，进而将完整的信息档案大数据处理平台搭建出来，从而具备挖掘数据与知识发现的条件。通过与历史图谱的结合，借助ES全文搜索引擎，实施语义理解数据操作，促进搜索质量的提高，将更加准确的信息找出来，提供给用户，最终做出全面总结并且将有深度的相关信息提供出来[5]。

有大量隐藏信息存在于信息档案资源中。借助挖掘信息档案资源的数据，获取相关有价值资料，通过查询相同或相似地理信息档案，可以实施关联性提取与推荐操作。例如，在以知识图谱为基础的信息档案管理系统中，将某区域农业用地变化情况查询出来，系统会获取国情普查数据，同时，与最新基础信息数据相结合，叠置分析所有关联农业用地的基础信息图层，通过与最新遥感数据的结合，修正分析结果，最终促进制图模板的自动化选择和应用，促进农业用地变化状况的动态生成，并将其返回给用户，从而可以将一些隐藏知识发现[6]。

类似于非空间数据挖掘方法，信息档案数据挖掘技术主要有：预处理档案数据技术，如筛选、过滤信息数据，分割信息档案区域等；档案空间特点与空间模式提取技术，比如分类信息档案、提取信息档案规则、预测地理信息、聚类信息档案等，一方面包含地理信息档案监督学习，另一方面包含信息档案无监督学习。

检索信息

完成统一处理海量档案信息任务之后，提取并挖掘整合后的档案信息时，借助文本挖掘技术，能够将文本处理中的高质量信息提炼出来。通过对这些信息实施摘要、压缩处理、融合、分类操作，促进文档关键字的形成。档案管理人员检索文档中任意关键字时，能够使数据搜索范围大大缩小；由于关键字是深入分析文本提炼出来的高质量信息，因此，借助关键字搜索功能，可以将关联关键字的数据搜索出来，从而大大提高用户搜索数据的精准性[7]。

当不能提供搜索关键字时，可以综合利用多种方式，如日期、文件名等实施检索信息操作，进而从中将用户可能的隐性诉求提取出来。多样性的检索方式提高信息查询效率和质量。

结束语

总而言之，大数据处理技术在信息档案管理中的应用能够使其最大程度发挥作用和价值。在HDFS作为基础的前提下，通过对集中存放的数据开展处理工作，再借助应用Hadoop信息档案资源大数据处理平台，能够促进信息档案管理工作效率和质量的提高，为最终获取精准信息档案提供支持和帮助，使信息档案管理得到不断优化与完善。

参考文献

[1]胡瑛.大数据处理技术在地理信息档案管理中的应用[J].现代测绘，2016，39（5）：56-58.

[2]姚丽丽，邢艳娟.基于大数据时代高职院校档案信息化建设探究[J].辽宁高职学报，2019，21（11）：105-108.

[3]冯淋潇.基于大数据事业单位人事档案信息化建设[J].卷宗，2021，11（21）：146-147.

[4]庄瑞弋.大数据思维与档案信息化建设路径[J].卷宗，2019，9（5）：20.

[5]任力.大数据技术时代档案信息管理思维与方式的变革[J].科技资讯，2016，14（32）：23-24.

[6]马润霞.信息化发展中大地测量电子档案与纸质档案并存的思考[J].测绘技术装备，2017，19（1）：67-68.

[7]闫爱玲.大数据时代工程档案信息管理与创新研究[J].工程技术研究，2019，4（6）：163-164.