基于Hadoop的高校图书馆大数据关键技术研究*

叶春蕾
（北京农学院图书馆，北京 102206）

摘要：为解决大数据环境下高校图书馆服务面临的海量数据分布式存储、多样化数据源分布式管理以及简易灵活的大数据服务应用问题，本文深入分析大数据处理研究内容、Hadoop生态系统以及高校图书馆大数据服务需求，提出一种基于Hadoop的高校图书馆大数据整体技术框架，构建高校图书馆海量数据分布式存储管理、多样化数据源分布式管理和多样化服务管理。该技术框架充分考虑大数据环境下高校图书馆大数据特征、数据存储、数据管理及服务处理等方面的变化，能够在一定程度上解决高校图书馆大数据服务的关键技术问题。

关键词：大数据；大数据技术；高校图书馆；Hadoop
中图分类号：G250
DOI：10.3772/j.issn.1673-2286.2017.05.005

随着信息化发展，大量数字资源纷纷进入高校图书馆。移动终端的普及使用户可以不受时空限制获取知识，用户数据量呈现爆发增长趋势。同时，高校图书馆数据来源也呈现多样化特征，除传统结构化数据资源外，还包括海量半结构、非结构化的信息资源。并且随着社交网站的普遍应用，互联网数据的产生速度超过以往任何一种传播媒介，高校图书馆用户的使用数据增长量更大，形成高速发展的大数据基础。为充分发挥大数据技术对高校图书馆服务的促进作用，本文提出基于Hadoop的高校图书馆大数据技术框架[1]，从应用实践角度阐述其关键技术，并对其进行深入探讨。本文构建的技术模型，旨在解决大数据环境下高校图书馆发展中面临的三个主要问题，即海量数据的分布式存储管理、多样化数据源管理（包括结构化数据、半结构化数据和非结构化数据的分布式管理）以及简易灵活的大数据服务管理。

1 研究现状分析

1.1 大数据技术研究现状

大数据技术融合多种计算技术。从信息系统角度可将大数据处理分为基础设施层、系统软件层、并行化算法层以及应用层[2]。Hadoop作为新的分布式存储与计算架构，因具有可扩展性、低成本、高效性与可靠性等优点，在分布式计算领域得到广泛运用，并已逐渐成为工业与学术界海量数据并行处理标准之一。Hadoop借鉴Google分布式文件系统（Google File System）实现分布式文件系统（Hadoop Distributed File System，HDFS）[3]，借鉴MapReduce计算模型实现分布式计算框架[4]，这两个系统构成Hadoop的核心子系统。MapReduce为大数据处理提供了良好平台，但由于是为大数据线下批处理而设计的，其随着数据规模不断扩大，对于需要高响应性能的大数据查询分析计算问题，以Hadoop为代表的大数据处理平台通常难以满足计算要求，因此有研究者尝试在Hadoop平台上搭建Spark框架[5]，利用Apache Spark快速灵活的迭代计算能力来满足大数据环境下日益增长的速度需求。与此同时，Hadoop为提高计算性能，参考BigTable实现了分布式数据库HBase[6-7]，并提供其他配套工具（如Hive[8]、Pig[9]等），以期在一定程度上弥补MapReduce的不足。

1.2 高校图书馆大数据研究现状

大数据环境下，无论从高校图书馆数据类型、数量、价值还是从未来发展趋势来看，高校图书馆海量数据已经初步具备大数据基本特征。图书馆作为图书情报领域的实践阵地，一直关注新信息技术和应用。

从大数据处理内容来看，系统软件层主要考虑大数据存储管理和并行化计算能力，其中存储管理是关键。陈臣提出一种基于新型存储的高校图书馆分布式大数据存

* 本研究得到北京市社会科学基金项目“大数据环境下高校图书馆服务创新技术研究”（编号：16XCB006）资助。

储架构[10]，其主要设计思想源于Hadoop的HDFS系统架构；梁俊荣提出一种基于Hadoop的图书馆复合大数据存储系统[11]。目前，基于HDFS的分布式文件系统发展较成熟，其以可扩展方式支持大规模数据的存储管理。但从高校图书馆大数据技术角度来看，需根据大数据处理过程中所面临的问题，提出可行性存储方案；此外，还需要考虑多样化数据结构存储问题。目前，国内外学者在非结构化数据处理和应用方面进行了广泛研究[12-17]，而更需要解决的是如何针对高校图书馆非结构化数据的特点建立模型，并提出具体实施方案。

当传统数据库在容量和数据结构上难以适应半结构化数据、非结构化数据时，NoSQL数据库技术应运而生[18]，但NoSQL数据库一般不提供SQL语言支持。大量数据库应用开发者仍然习惯于SQL编程，如果要在NoSQL上提供SQL查询机制，则需要将面向结构化数据查询的SQL与面向半结构化数据、非结构化大数据查询的NoSQL统一融合，新的数据查询技术NewSQL便是此环境下的产物（包括Apache HBase）。HBase以其分布式特点、海量存储技术以及灵活的数据定义方式在多个领域得到广泛应用[19-20]。

在大数据并行化算法层，现有研究偏向于大数据处理所需分析挖掘算法的并行化设计。大数据分析挖掘算法通常可使用MapReduce架构实现，但要求开发人员具有较高的编程能力，他们需要编写复杂的MapReduce程序以实现大数据分析和挖掘。Hive提供了一个供用户进行数据查询、分析和挖掘的仓库系统，该系统使用类似SQL的HiveQL语言描述数据处理逻辑，减少大数据处理的编程工序。吴晓英等基于Hive平台调用Mahout算法进行数据挖掘与分析[21]。一般情况下，Hive主要面向数据库的数据处理平台，但是高校图书馆的流数据也是图书馆大数据的重要来源之一，在处理数据流时可以考虑使用Pig。相比MapReduce，Pig为海量数据集的处理提供了更高层次的抽象，可以简化MapReduce任务的开发，提高Hadoop集群数据处理的便捷性。

1.3 高校图书馆大数据技术面临的问题

1.3.1 海量数据存储问题

海量数据资源存储需求对高校图书馆存储能力提出挑战。苏新宁认为大数据时代高校图书馆资源建设要注重各类再生资源的存储工作[22]。陈传夫等认为大数据环境下高校图书馆建设面临的问题之一是资金投入不足[23]。因此，在现有的资金基础上提高大数据对高校图书馆服务创新的推动作用首要解决的问题是提高海量数据存储能力。

大数据环境下，高校图书馆对数据存储的安全性、读写性能、经济性和管理效率等方面提出更高要求。数据存储的安全性是高校图书馆有效服务的关键问题之一，只有确保数据安全才能进一步提高图书馆服务可靠性，保护用户隐私。随着高校图书馆数据量指数级增长，对图书馆数据存储的读写性能也提出更高要求。海量数据要求存储系统具有高吞吐量、快速准确存取和传输能力，为高校图书馆用户服务决策支持提供保障。高校图书馆大数据存储系统在构建时需考虑存储成本问题，所以要求大数据存储架构能够对原有存储系统平台进行升级和无缝对接，在保证前期数据存储业务有效运行的同时，尽可能降低大数据存储系统建设成本。因此，在大数据环境下如何保证大数据存储系统安全、高效、经济和可靠是高校图书馆面临的一个严峻挑战。

1.3.2 多样化数据结构处理问题

陈传夫等认为：大数据环境下高校图书馆存在资源建设不合理的问题；资源同质化现象比较严重；存储在数据库中的结构化数据占比高，缺乏对非结构化数据的统一管理[23]。在高校图书馆大数据服务创新中需要解决的另一个问题是对多样化数据结构的处理。

在高校图书馆中结构化数据占比较低，非结构化数据是大数据的主要存在形式。其一方面来自图书馆自身馆藏资源，如图片、图像、论文、多媒体、数据库、自建特色数据库、RFID数据、用户行为数据、用户社交网络数据、移动设备数据等；另一方面来自图书馆外部开放资源，如即时通讯数据、网络出版与传播数据、电子商务数据、社交网络数据、馆际共享数据等。对多样化数据源，尤其是非结构化数据的有效管理将直接影响图书馆服务效果。

1.3.3 多样化服务应用问题

程学旗等认为大数据价值挖掘需要对其内容进行分析与计算，主要包括深度学习、知识计算和可视化技术[24]。高校图书馆开展服务创新活动应关注数据分析[25]。在理论条件下，高校图书馆大数据技术可以满足服务创

新需求[22]，但要实现这些技术必须考虑其在服务中的灵活性和简便性。

按照时效性划分，高校图书馆大数据资源主要包括两类数据：第一类是对时效性要求不高的数据，主要包括系统日志、用户行为、阅读关系及系统配置数据等历史数据；第二类是对时效性要求较高的数据，主要包括用户个性化阅读即时需求、用户位置信息等实时数据。针对第一类数据，传统数据服务方式通常借助数据仓库，使用各类数据挖掘算法或工具提供数据服务，但大数据环境下，传统数据仓库处理方式很难有效地完成数据多样化处理；而对于时效性要求较高的第二类数据来说，传统数据服务方式通常会使用数据库或文件方式进行读、写、分析等处理，但从使用效率角度来看，很难满足海量数据实时转换、导入并加载到分布式数据库管理系统的需求。

基于以上的分析可以看出，在大数据环境下，高校图书馆服务创新面临海量数据存储问题、多样化数据结构处理问题、多样化服务应用等问题。目前，Hadoop所提供的多样化、灵活性和可扩展的系统成员能够完成大数据处理要求。陈吉荣等提出，Hadoop生态系统将成为中小企业在面对大数据问题时的首选解决方案[26]。张红介绍了国家图书馆联合软件开发商自主研发的“文津搜索”系统[27]，该系统引入Hadoop系统和各类NoSQL技术，实践了大数据技术在图书馆资源服务领域的应用。因此，本文将结合大数据技术内容、Hadoop生态系统架构以及高校图书馆大数据服务需求，提出一种基于Hadoop的高校图书馆大数据技术框架，并对框架中的关键技术进行深入讨论，以期为高校图书馆大数据服务创新实践提供技术参考。

2 基于Hadoop的高校图书馆大数据技术框架

本文在充分研究大数据技术内容、Hadoop生态系统架构、高校图书馆大数据现状和问题的基础上，提出一种基于Hadoop的高校图书馆大数据技术框架，如图1所示。

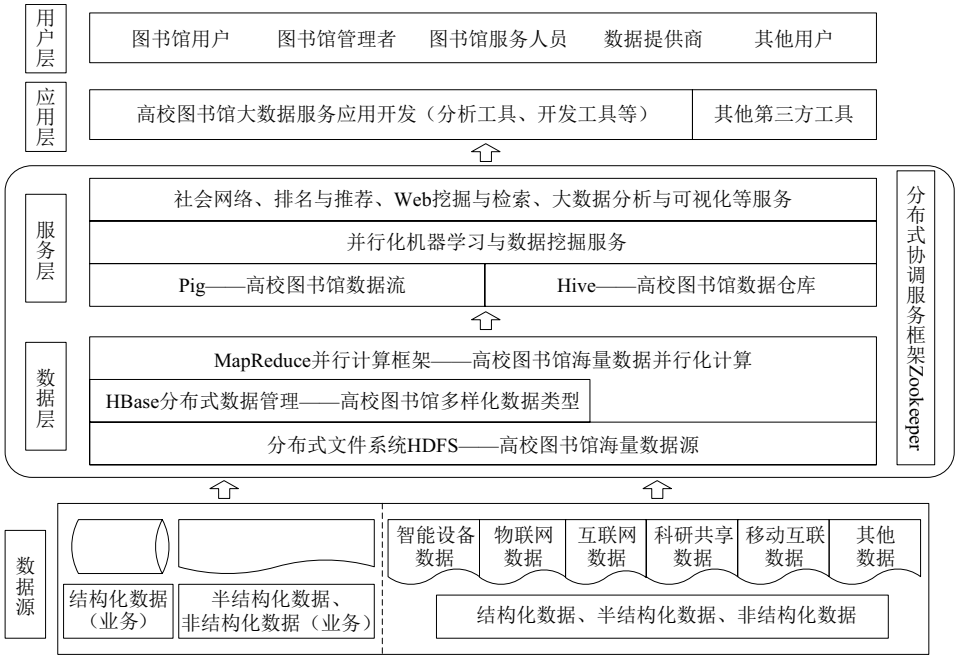

图 1 基于Hadoop的高校图书馆大数据技术框架

基于Hadoop的高校图书馆大数据技术框架主要分为四个层次，分别是数据层、服务层、应用层和用户层。应用层主要利用传统分析工具、开发工具等进行大数据服务应用开发；用户层主要针对各级用户提供基于工具的服务应用；其中关键技术问题主要集中在以Hadoop生态系统为支撑的数据层和服务层，这两层主要解决海量数据分布式存储管理、多样化数据源分布式管理以及大数据多样化服务管理三方面问题。

2.1 基于HDFS的海量数据分布式存储管理

为更好地解决海量数据分布式存储面临的安全性、

读写性能、经济性和管理效率等问题，本文建立基于HDFS的高校图书馆大数据存储结构，如图2所示。

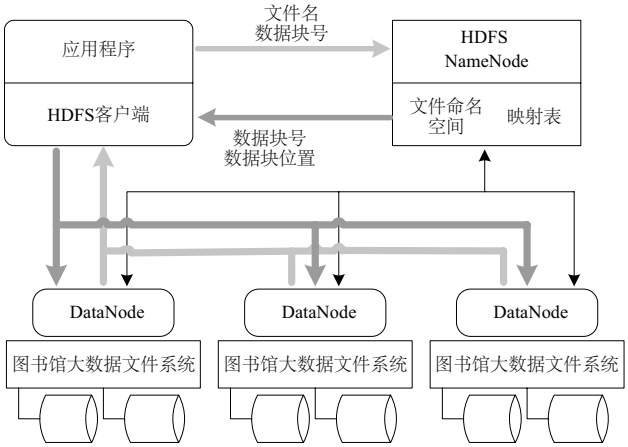

图 2 基于HDFS的高校图书馆大数据存储结构

在基于HDFS的高校图书馆大数据存储结构的实现过程中，可以从一组普通商用服务器中选择一台性能较好的服务器作为主控节点NameNode，其他服务器作为从节点DataNode。高校图书馆大数据文件（包括结构化数据、半结构化数据以及非结构化数据）作为Linux本地文件被存储在DataNode节点服务器中。主控节点和从节点服务器的目录结构不同，主要由其身份决定。NameNode负责管理文件系统的命名空间和元数据，同时处理外部文件访问请求。NameNode保存高校图书馆大数据文件系统的3种元数据：命名空间（整个分布式文件系统的目录结构）、数据块与文件名映射表以及每个数据块副本（默认有3个副本）的位置信息。HDFS对外提供命名空间，保证用户数据可存储在文件中；但HDFS对内，文件可能被分成若干个数据块。DataNode用于存储和管理文件的数据块，为防止数据丢失，每个数据块默认有3个副本，且3个副本分别复制在不同节点上，以避免因一个节点失败而造成一个数据块的彻底丢失。因此，基于HDFS的高校图书馆大数据分布式管理能在一定程度上解决高校图书馆海量数据的存储问题。

2.2 基于HBase的多样化数据源分布式管理

为有效解决多样化数据源，尤其是非结构化数据的分布式管理问题，本文对非结构化数据源建立统一的数据模型。在高校图书馆中，多数非结构化数据的产生与特定用户行为有关，虽然该类数据格式各异，但可通过用户行为抽取出具有共同特征的属性集。同样，海量非结构化数据也存在一定关联，其关联性通过主题体现，可以从数据主题特征中抽取出相关主题属性集。高校图书馆非结构化数据模型主要包括基本属性集、内容属性集、特征属性集、行为属性集和主题属性集。其中，基本属性集描述非结构化数据对象的一般属性，包括与非结构数据文件相关的信息，如文件名称、文件路径、文件类型、可操作权限类型、数据负责人、数据所属单位等；内容属性集描述与非结构化数据内容相关的信息，包括非结构化数据标题、数据主题信息、所属科学领域、数据内容语言等；特征属性集描述与非结构化数据类型特征相关的特有属性，如媒体属性、文档属性、音频数据、视频属性以及图像属性等；行为属性集描述与非结构化数据用户行为相关的属性，如最近访问时间、访问累计时长、所属服务名称、服务类别等；主题属性集描述与非结构数据主题相关的属性，如非结构数据所属主题在检索结果中的数量相同主题数以及非结构数据所属主题被检索次数等。通过属性集描述非结构化数据，并将其纳入多样化数据源分布式管理模型中，能促使该分布式数据管理模型更好地完成高校图书馆用户行为检测和模式提取、高性能数据检索、数据分析以及可视化服务应用等。

HBase是一个基于HDFS的分布式可扩展NoSQL数据库，提供对结构化、半结构化以及非结构化大数据的实时读写和随机访问。HBase提供了一个基于行、列和时间戳的三维数据管理模型，在HDFS实际的存储中，直接存储每个字段数据所对应的完整键值对“{row key, column family, column name, timestamp}->value.”。如图书馆用户访问资源的键值对可以表示为“{key3, userInfo, dataSource, t2}->’http://www.cnki.net’”。HBase的每行每列族中保存一个Map映射表，列不需要静态定义，每列都可以动态增加或减少。利用HBase实现多样化数据源管理主要包括三个步骤：对非结构化数据的各类属性进行描述，提取并设置属性值；以属性集合及其属性值建立对应的HBase数据表（见表1）；基于HBase数据表进行各类灵活的查询、分析等操作。

HBase通过灵活的键值对为高校图书馆非结构化数据属性集提供精确的保存方式。同时，HBase中每个数据表的记录数（行数）可以多达几十亿条，每条记录可以拥有上百万字段。其存储能力不需要特殊硬件，普通服务器集群即可胜任。因此，对于高校图书馆海量、多样化数据来说，基于HBase的技术框架是可行的。

表 1 HBase数据表的基本结构

row key	BasicInfo			ContentInfo			CharacInfo			BehaviorInfo		TopicInfo	
	FName	FPath	……	Title	Topic	……	Type	Size	……	Time	……	TName	……
k1	f1	p1	……	ti1	to1	……	ty1	s1	……	time1	……	n1	……

2.3 基于Hive、Pig的大数据多样化服务管理

为更好地提高高校图书馆大数据服务的灵活性和简便性，本文针对时效性要求提出不同管理方案。对于高校图书馆对时效性要求不高的数据来说，可以使用Hive数据仓库。首先，Hive是建立在Hadoop上的数据仓库基础架构，早期被Facebook用于处理和分析大量用户日志数据；其次，作为Hadoop的数据仓库工具，Hive可将结构化数据文件映射到数据库表，并提供简单的数据分析功能；再次，Hive还提供一系列工具，可以进行数据提取、转换和加载；最后，Hive定义了简单的类SQL查询语言（HiveQL），方便熟悉SQL的用户执行简单的数据查询操作。此外，该语言也允许熟悉MapReduce的开发者开发自定义的Mapper和Reducer，完成复杂的数据分析工作。

对于高校图书馆时效性要求较高的数据来说，可以使用Pig平台从数据流层面解决这一问题。Pig Latin是一种面向数据流的语言。其提供数据排序、过滤、求和、分组和关联功能，同时允许用户自定义函数，以满足特殊数据处理需求。当处理海量数据时，首先需要使用Pig Latin语言编写脚本程序，然后在Pig中执行该脚本程序。Pig将用户编写的Pig Latin程序编译为MapReduce作业程序，并上传到集群中运行。对用户来说，底层的MapReduce工作是完全透明的，用户只要了解Pig Latin语言就可以自行处理海量数据。基于Hive和Pig的大数据多样化服务管理的流程图如图3所示。

因此，通过对技术框架中各关键技术的分析可以看出，在高校图书馆海量数据资源、资金有限、多样化服务需求条件下，本文构建的高校图书馆大数据技术框架将为高校图书馆大数据服务提供充分支持，能够有助于充分挖掘图书馆海量数据资源的潜在价值，进而提升图书馆服务创新水平。

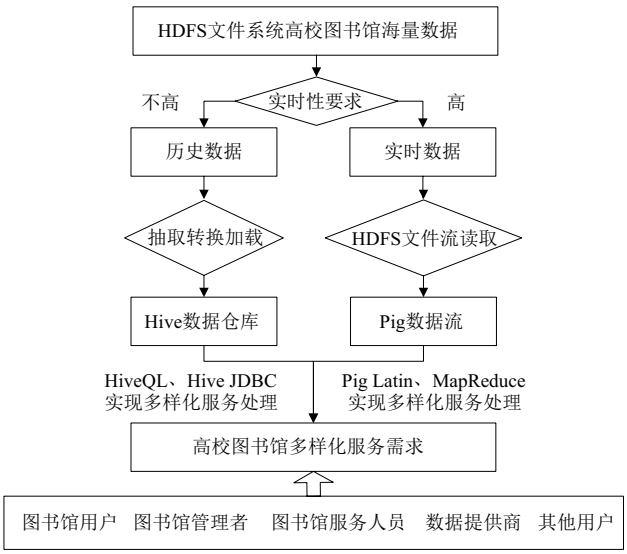

图 3 高校图书馆大数据多样化服务流程

3 结论

高校图书馆数字资源无论从数据类型、数量、价值还是从未来发展趋势来看，都初步具备大数据基本特征。因此，为有效解决大数据环境下高校图书馆服务创新面临的问题，本文深入分析高校图书馆大数据特征、大数据技术研究内容、Hadoop生态系统架构以及高校图书馆大数据技术面临的问题，提出一种基于Hadoop的高校图书馆大数据技术框架，并进一步提出分布式存储技术（旨在解决图书馆大数据海量存储问题）、多样化数据源分布式管理技术（旨在解决图书馆多样化数据类型管理问题）、多样化服务处理技术（旨在提供灵活简便的图书馆大数据服务）。该技术框架充分考虑大数据环境下高校图书馆大数据特征、数据存储与计算、数据管理及服务处理等方面的变化，能够在一定程度上解决高校图书馆大数据技术实施问题。

尽管Hadoop存在不足，特别是基于内存计算模式对速度的响应程度远低于Spark，但因其具有可扩展性、低成本、高效性与可靠性等优点，在分布式计算领域得到广泛的应用。Hadoop正努力扩展现有计算模式框架，以解决现有版本在计算性能、计算模式等方面的不足。针对Hadoop MapReduce难以支持迭代计算的缺陷，后续研究会考虑在Hadoop平台上搭建Spark框架以解决这类问题。本文为高校图书馆大数据技术框架的

模型探讨阶段，具体实证将在后续研究中进一步展开，以更好地验证该技术框架的有效性和可行性。

参考文献

[1] Hadoop[EB/OL].[2017-01-18].http://hadoop.apache.org/.

[2] 黄宜华.深入理解大数据:大数据处理与编程实践[M].北京:机械工业出版社,2014:24.

[3] GHEMAWAT S,GOBIOFF H,LEUNG S-T.The Google file system[C]//Proceedings of the 19th ACM Symposium on Operating System Principles(SOSP 2003),October 19-22,2003,New York:[s.n],2003:29-43.

[4] DEAN J,GHEMAWAT S.MapReduce:simplified data processing on large clusters[J].Communications of the ACM,2008,51(1):107-109.

[5] 陈虹君,吴雪琴.基于Hadoop平台的Spark快数据推荐算法分析与应用[J].现代电子技术,2016(10):18-20.

[6] CHANG F,DEAN J,GHEMAWAT S,et al.Bigtable:a distributed storage system for structured data[C]//Proceeding of the 7th Symposium on Operating Systems Design and Implementation(OSDI).Seattle:[s.n],2006:205-218.

[7] HBase[EB/OL].[2017-02-15].https://hbase.apache.org/.

[8] Hive.Getting started with Apache hive software[EB/OL].[2017-02-15].https://hive.apache.org/.

[9] Hadoop.Welcome to apache Pig[EB/OL].[2017-02-15].https://pig.apache.org/.

[10] 陈臣.一种基于新型存储的数字图书馆分布式大数据存储架构[J].现代情报,2015(1):100-103.

[11] 梁俊荣.基于Hadoop的图书馆复合大数据存储系统研究[J].现代情报,2017(2):63-67.

[12] FERRUCCI D,LALLY A.UIMA:an architectural approach to unstructured information processing in the corporate research environment[J].Natural Language Engineering,2004(10):327-348.

[13] DOAN A,NAUGHTON J F,BAID A,et al.The case for a structured approach to managing unstructured data[C]//Proceedings of the 4th Biennial Conference on Innovative Data Systems Research.Asilomar:CIDR,2009:1-10.

[14] 韩晶,鄂海红,宋美娜,等.基于主体行为的非结构化数据模型[J].计算机工程与设计,2013(3):904-908.

[15] 白如江,冷伏海.“大数据”时代科学数据整合研究[J].情报理论与实践,2014(1):94-99.

[16] 郭春霞.大数据环境下微信公众平台非结构化数据融合研究[J].现代情报,2015(8):141-143,150.

[17] 陈臣.基于Hadoop的图书馆非结构化大数据分析与决策系统研究[J].情报科学,2017(1):24-28.

[18] 申德荣,于戈,王习特,等.支持大数据管理的NoSQL系统研究综述[J].软件学报,2013(8):1786-1803.

[19] 王远,陶烨,袁军,等.一种基于HBase的智能电网时序大数据处理方法[J].系统仿真学报,2016(3):559-568.

[20] 徐爱萍,王波,徐武平.HBase中基于时空特征的监测视频大数据关联查询研究[J].计算机应用研究,2017(5):1-7.

[21] 吴晓英,明均仁.基于数据挖掘的大数据管理模型研究[J].情报科学,2015(11):131-134.

[22] 苏新宁.大数据时代数字图书馆面临的机遇和挑战[J].中国图书馆学报,2015(6):4-12.

[23] 陈传夫,钱鸥,代钰珠.大数据时代的数字图书馆建设研究[J].图书情报工作,2014(7):40-45.

[24] 程学旗,靳小龙,王元卓,等.大数据系统和分析技术综述[J].软件学报,2014(9):1889-1908.

[25] 程结晶.大数据时代图书馆服务创新的内容及其策略研究[J].情报理论与实践,2016(3):57-62.

[26] 陈吉荣,乐嘉锦.基于Hadoop生态系统的大数据解决方案综述[J].计算机工程与科学,2013(10):25-35.

[27] 张红.基于大数据技术的资源发现平台构建——以国家图书馆“文津搜索”系统为例[J].数字图书馆论坛,2016(1):61-67.

作者简介

叶春蕾，女，1975年生，博士，副教授，研究方向：情报分析、大数据技术研究，E-mail：yechunlei2014@126.com。

Study on the Key Technology of University Library’s Big Data Based on Hadoop

YE ChunLei
(Library of Beijing University of Agriculture, Beijing 102206, China)

Abstract: In order to solve the problems that the distributed storage of the massive data, the distributed management of the diverse data sources, the simple and flexible application of the big data services in university libraries in China, this paper proposes a framework of the big data technology in university libraries based on Hadoop. The framework builds the distributed storage of the mass data, the distributed management of the diverse data sources and the diversified service processing. The technical framework can solve the key technical problems of the big data service of university libraries to a certain extent.

Keywords: Big Data; Big Data Technology; University Library; Hadoop

（收稿日期：2017-04-11）