项目编号：2022012

CALIS 全国农学文献信息中心研究项目

结题报告

项目名称：大数据时代高校科研数据服务模式与设路径研究

项目关键词：大数据；高校；科研数据；服务模式

项目单位（盖章）：东北农业大学图书馆

通信地址：黑龙江省哈尔滨市香坊区木材街 59 号东北农业大学图书馆 150030

项目主持人：张晶晶

联系电话：18945041549

电子邮件：7802336@qq.com

提交日期：2023 年 5 月 24 日

项目结题验收单

专家验收表（主持人所在单位组织 3-5 名专家对项目进行验收、自评。）

项目名称	大数据时代高校科研数据服务模式与建设路径研究		
主持人	张晶晶	职务/职称	馆员
所在单位	（加盖单位公章）东北农业大学图书馆		
专家意见	当前大数据时代带来数据密集型研究范式，数据成为重要的科研生产资料，相关科研人员在科研环境和政策预期的共同作用下迫切需要专业的科研数据服务。优质的科研数据服务有利于保障数据质量和安全、促进数据共享，进一步起到节约科研经费、提高科研产出的作用。 课题研究能按申请书进度如实进行，课题组在搜集、整理、分析相关文献和资料的基础上，采用文献调查、文献统计分析、对比分析等方法，从高校提供科研数据服务的必要性、高校科研数据服务模式、高校科研数据服务建设路径三点入手，对高校科研数据服务相关问题展开了系统探讨。 课题组围绕科研项目的“科研构思——科研实施——成果整理——成果发表”四个阶段构建了基于科研项目生命周期的高校科研数据三层服务模式，其中基础层包括数据发现、数据计划、数据存档、数据发表、数据共享等服务，支撑层包括数据推广、数据素养教育、数据咨询等服务，增值层包括数据采集/生产、数据分析、数据评估/筛选等服务。课题按照按需供给、分步实施、分类开展、分层推进、由易到难、逐步深入六项原则构建了包含前期论证、基础设施建设、服务试用、服务推广、服务深入五个阶段的高校科研数据服务建设路径。 结题报告撰写详细，保证了课题研究的全面性和完整性。课题成员在研究的过程中，本着严谨、严肃的工作作风，不仅提升了业务科研能力，同时也为保障数据质量和安全、促进数据共享起到了一定的推进作用。 （如需要可增加页数）		
专家签字			
职务/职称	研究馆员	研究馆员	副研究馆员

大数据时代高校科研数据服务模式与建设路径研究

关键词：大数据；高校；科研数据、服务模式

1 研究背景、目的及意义

1.1 研究背景

科研数据的充分获取和高效管理是影响科研项目顺利进行的关键因素之一。作为人类发现、探索、解释自然与社会的忠实记录，科研数据是进行科研活动所必需的基础性生产资料，是支撑研究结论的重要事实或结果。科研数据的获取，一方面来自项目内部试验测试、遥感勘测、模型计算、调研访问、问卷统计等科研活动，另一方面来自项目外部科研数据的共享。对于多数科研项目，围绕大量科研数据的是各种繁琐的数据计划、获取、存档、分析、处理、发表、共享等工作，如何合理规划、充分获取和高效管理这些数据对于卓有成效的科研活动具有不言自明的意义。

开展科研数据服务相关工作的探索和实践已有相当长的一段历史。首先，当前科研活动呈现复杂度越来越高、耗时越来越多、分工越来越细、合作越来越密的特点；其次，与计划思考、科研实验、报告编写等工作相比，科研数据获取与处理属于科研活动中的非核心业务；最后，科研数据工作伴随绝大多数科研活动始终，占用大量的人力、物力、财力。在这些因素的共同作用下，1957 年国际科学联合会理事会成立了世界数据中心，其后众多国际组织、国家政府、资助机构、出版机构相继推出有关科研数据服务方面的政策和措施。早在 1994 年我国地学领域专家也就开展科研数据共享服务进行了呼吁。

在国外高校提供的科研数据服务中，美国、澳大利亚、英国等走在前列。早在 1970 年代美国密歇根大学就建成了“校际政治与社会研究联合数据库”（Inter-University Consortium for Political and Social Research，ICPSR），1994 年美国杜克大学就出台了科研数据保存和获取政策。澳大利亚 2007 年颁布的《澳大利亚负责任科研行为准则》明确提出高校必须建立起科研数据保存、数据所有权归属以及数据访问等政策，这直接推动了高校科研数据服务工作。英国高校科研数据服务政策制定较晚，随着英国艺术与人文研究委员会（Arts and Humanities Research Council, AHRC）、生物技术与生物科学研究委员会（Biotechnology and Biological Sciences Research Council, BBSRC）、工程与物理科学研究委员会（The Engineering and Physical Sciences Research Council, EPSRC）等科研资助机构相继制定数据管理要求，其发展迅速。

国内高校科研数据服务起步较晚，目前仍处于规划探索阶段，总体上大幅落后于国际同行。

2011 年武汉大学图书馆在 CALIS 三期的资助下，开展了科研数据管理与服务机制和平台的实践与探索；2013 年复旦大学推出了国内高校截至目前唯一的社会科研数据平台；2015 年北京大学开放研究数据平台测试版上线运行。

大数据时代，科研数据重要性空前提升，亟需专业的科研数据服务。大数据时代，科研环境转变，面对新的数据密集型科研范式，科研数据成为现代科学研究的基础性资源，“已有的数据是新研究的宝贵资产，对于已有数据的整合、挖掘和再利用为学术研究提供了一种新的资源”。面对新形势，数据的采集、存储、分析需要专业的设备、人员，这使得科研活动对于数据的获取和处理变得更为迫切。

1.2 研究目的

大数据时代的数据密集型科研是一种“大科学”模式，科研活动分工更细、协作度大幅提高，大数据的采集、存储、处理、计算都有赖于专业的设备和人员，给高校传统的以个体和小团队为主的“小科学”模式带来巨大冲击。可以说传统的个体和小团队很难完全独立胜任。科研人员没有足够的时间和精力应对科研数据获取和处理工作。即便科研人员具有相当的科研素养，也很可能无暇顾及这方面工作，当前数据获取与管理工作需要掌握更为繁琐的政策、操作更加复杂的软硬件，专业性越来越强，这意味着需要耗费的时间和精力用于数据获取与管理；另外传统科研中，科研人员为获取和处理数据花费大量的精力，但是社会发展在逐渐加速中，全社会对于科研人员的产出要求也在逐步提升，将过去的模式用在现下和未来，科研人员必将无法应对繁重的科研工作。因此提供专业科研数据服务、解放科研人员、让其回归科研核心工作不失为一个更加合理的安排。

1.3 研究意义

科研数据服务的发展大致经历了科研数据（管理和共享）服务——科研数据服务两个阶段。最早为了实现科研数据共享，主要是欧美等国科研人员将数据集中起来管理，并为此制订了完备的政策制度。然而开展科研数据共享和管理工作多年之后，欧美等国仍有不少科研数据得不到妥善管理、已有的数据得不到充分利用、甚至一些科研人员仍对科研数据共享与管理存在很多困惑。而我国由于至今未建立完善的政策规范，实践上只有国家科技部科学数据共享工程一直有效运行。

作为一种非消耗性资源，理论上科研数据可以无限制共享和重复使用，然而实际情况却不尽如人意，概因科研数据共享和管理工作专业、复杂、繁琐，对于这项耗时、耗力的非核心工作，投入过多精力和经费是一项效益低下的选择。或许正是看到问题所在，欧美各国近年来逐渐开展更为广泛的科研数据服务，以服务促进科研数据集中管理和广泛共享。

有鉴于此，国内外研究一直以来均将更多的关注直接投向科研数据共享和管理服务，直到近年来有关科研数据服务的研究才逐渐增多。即便科研数据服务与科研数据共享和管理相关研究存在诸多重叠和交叉，但科研数据服务有其自我体系，相关研究尚处于起步阶段。课题通过对高校提供科研数据服务的必要性、高校科研数据服务模式和高校科研数据服务建设路径等问题展开系统性探讨，是系统阐述高校科研数据服务理论的一个积极尝试。

2 研究内容及方法（思路、方法、具体内容）

2.1 研究思路

和多数研究一样，本课题按照“提出问题——分析问题——解决问题”的思路逐层推进（图1.1）。

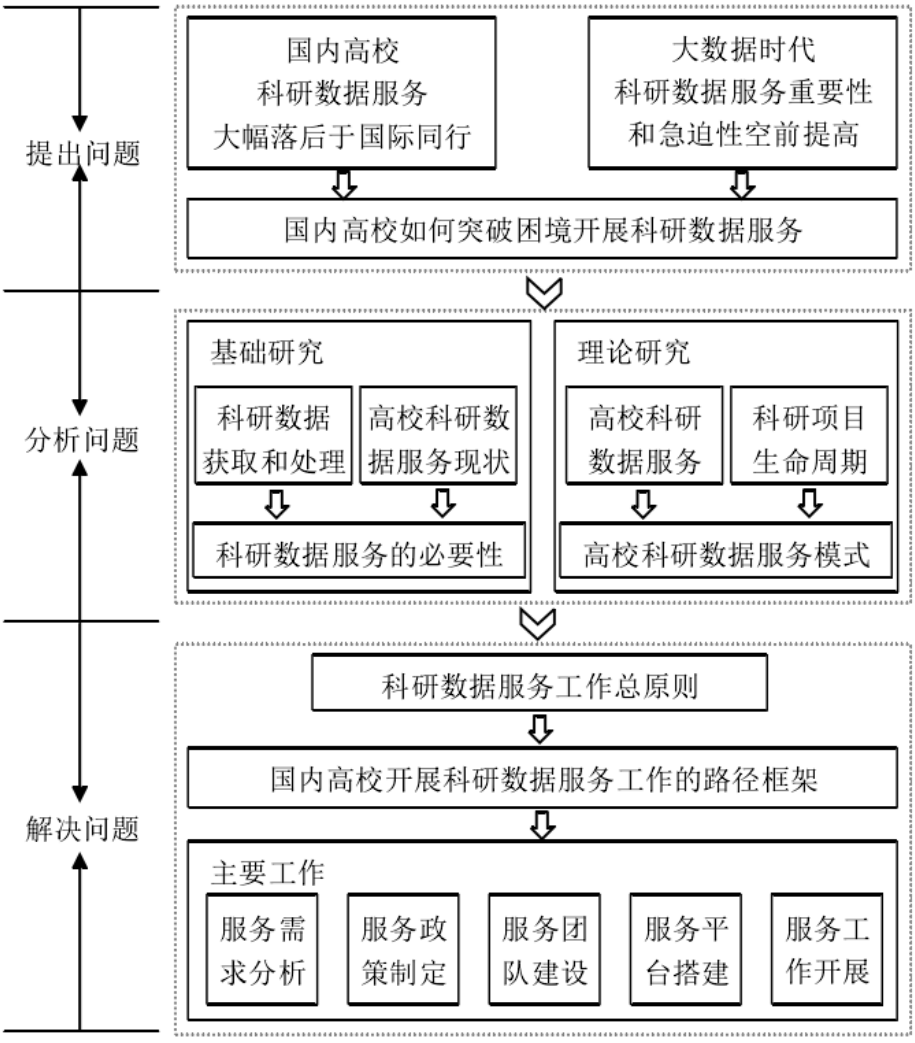

图 1.1 研究思路图

2.2 研究方法

（1）文献调查法

本课题在研究的过程中，选取 CNKI 中国知网、Emerald 数据库、百度搜索引擎，通过搜集、整理、鉴别，获得了大量有关科研数据的文献资料、数据资料，作为必要的背景材料、理论来源和事实依据。

（2）文献统计分析法

对于收集到的文献，课题通过识别、分类、统计等步骤分别从发表年代、研究领域、研究机构、资助基金等角度进行分析，从总体上把握科研数据相关研究热点和动向。

（3）比较分析法

课题中多处涉及，国内科研数据研究现状与国外科研数据研究现状的比较、科研数据概念与科技文献概念的比较，科研数据服务概念与科研数据管理概念的比较，通过比较分析抓住本质和异同点。

2.3 研究内容

课题研究主要涉及高校提供科研数据服务的必要性、高校科研数据服务的模式、高校科研数据服务的建设路径三点。三者相互依赖，必要性研究是研究得以进一步展开的前提和依据，模式研究为建设路径提供理论依据，路径研究反过来验证理论可靠性；三者共同构成一个整体系统，解决“国内高校如何突破困境开展科研数据服务”这一问题。

2.3.1 大数据时代高校亟需科研数据服务

（1）科研环境转变是高校开展科研数据服务的根本动力

我们身处大数据时代，麦肯锡全球研究院（MGI）给大数据（big data）下的定义是：“一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。”大数据一直以来都存在于我们的社会，但是没有人去采集并加以利用，关键原因在于相关处理技术还不够成熟，大数据的价值无法挖掘出来。

21 世纪初，与大数据密切相关的感知技术、存储技术、云计算和分布式处理技术日益发展与成熟，有关大数据的理论也逐步完善。2011 年 6 月 MGI 在题为《大数据：下一个创新、竞争和生产力的前沿》的研究报告中首次提出“大数据时代已经到来”的命题。大数据所描绘的美好蓝图推动了全球广泛而持久的关注。MGI 指出“当前大数据规模以及其存储容量正在迅速增长，已经渗透到各个行业和业务职能领域，成为可以与物质资产和人力资本相提并论的重要

的生产要素”。

“大数据是继传统 IT 之后下一个提高生产率的技术前沿”。牛津大学教授维克托·迈尔-舍恩伯格（Viktor Mayer-Schönberger）和《经济学人》杂志的数据编辑肯尼恩·库克耶（Kenern Cukier）合著的《大数据时代》一书指出“大数据改变了我们的思维方式，让我们从因果关系的串联思维变成了相关关系的并联思维；大数据改变了我们的生产方式，物质产品的生产退居其次，信息产品的加工将成为主要的生产活动；大数据改变了我们的生活方式，我们的精神世界和物质世界都将构建在大数据上。大数据不仅仅是一门技术，更是一种全新的商业模式，它与云计算共同构成了下一代经济的生态系统”。

毫无疑问科研活动必须借助当时所处社会的各种理论储备和技术手段，当大数据时代来临，始终处在社会发展前沿的科研活动自然不会缺席。大数据给科研活动带来三个颠覆性观念转变：a.是所有数据，而不是随机抽样（随机抽样可以看成技术能力不足的条件下人为外加的限制）；b.允许数据误差，掌握大致方向即可（样本数据较少时，数据误差容易导致结果偏差，需力求数据精确）；c.关注相关关系，而不是因果关系（大数据分析是寻找相关关系的一个重要手段，并非要否定因果关系）。这些颠覆性的改变孕育了新的科学研究第四范式，2009 年，微软在 The Fourth Paradigm: Data -Intensive Scientific Discovery 中从科学研究方法的角度解释科学研究范式并指出新的针对数据密集型（Data Intensive）科学的第四范式的产生，引起了大家对于数据密集型科学的重视。

全新的科研范式下科研数据成为科学研究的基础性资源。科学研究第四范式归根结底是对海量数据的挖掘，由于理论和技术的发展，使得我们能够突破以往抽取部分数据样本进行模拟分析的限制，进而对全样本数据进行分析。可以说只要获得数据就可以进行科研，那么“已有的数据是新研究的宝贵资产，对于已有数据的整合、挖掘和再利用为学术研究提供了一种新的资源”。Gary 认为已有数据的使用方式有 a.数据验证（Verification），通过对数据的追踪及验证有助于学者们进一步确认论文作者的学术贡献；b.数据挖掘（Mining），通过对同一数据进一步挖掘从不同角度得出更多结论；c.数据聚合（Aggregation and scaffolding），通过对不同项目和不同研究人员的数据进行聚合，形成新的想法。

大数据时代的数据密集型科研是一种“大科学”模式，给高校传统的以个体和小团队为主的“小科学”模式带来巨大冲击。第四范式下的科研活动分工更细、协作度大幅提高，大数据的采集、存储、处理、计算都有赖于专业的设备和人员，可以说传统的个体和小团队很难完全独立胜任。大数据时代，数据成为可以重复使用的生产资料，高质量的数据获取和处理工作一定会从传统科研活动中细分出来，及早适应并紧随趋势对于做好大数据科研至关重要。

（2）科研人员需求是高校开展科研数据服务的直接动力

首先，多数科研人员在有关科研数据获取与管理方面是“无知”的。数据密集型科研范式下，科研数据获取和处理是科研工作细分的结果，是一项具有一定专业负责性的工作，当 UNC 的研究人员被问及数据管理“是否能够获得充分的基金支持”等问题时，高达 65%的回答是“我

不知道”；根据刘霞等学者的调查，高校中“超过 60%的科研人员发生过数据丢失现象”、“65%的数据由项目团队分散存储和管理”、“超过 50%的科研人员不对数据永久保存”、超过 40%的研究者对“数据管理对于科学研究的促进作用”没有明确的认识，即便如此仍有“超过 50%的科研人员对现有的数据管理手段表示满意”，可见相当数量的科研人员由于各方面原因并不了解数据获取或者管理，更不要谈论如何更专业的做好这件事。

其次，科研人员没有足够的时间和精力应对科研数据获取和处理工作。即便科研人员具有相当的科研素养，他也很可能无暇顾及这方面工作，当前数据获取与管理工作需要掌握更为繁琐的政策、操作更加复杂的软硬件，专业性越来越强，这意味着需要耗费的时间和精力用于数据获取与管理；另外传统科研中，科研人员为获取和处理数据花费大量的精力，但是社会发展在逐渐加速中，全社会对于科研人员的产出要求也在逐步提升，将过去的模式用在现下和未来，科研人员必将无法应对繁重的科研工作，UNC 超过一半的受访者认为在使数据为他人所用上需要花费时间。因此提供专业科研数据服务、解放科研人员、让其回归科研核心工作不失为一个更加合理的安排。

最后，科研人员没有足够的经费用于科研数据的获取和处理。获取和处理数据需要支付大量的软硬件和人力费用，由个人或者小团队独力支付相关费用，无疑是很不经济的做法，而且这也得不到资助机构的全力支持，当 UNC 的研究人员被问及在他们所在院系、实验室、中心或者研究群体中，在长期数据保存（5 年以上）上面是否可以获得充分基金支持，仅有 10%的人给出肯定的回答。科研经费是有限的，提供科研数据获取和处理服务，集中管理科研数据，对于节省科研经费、提升科研产出具有不证自明的作用。

大数据时代，科研环境的转变、数据政策的繁琐要求、科研数据获取和处理工作日益专业化，面对这种趋势，科研人员越来越难以很好的应对，对数据获取和处理“无知”化日趋加重，即便能够胜任相关工作，但是在繁琐的科研工作和有限的经费制约下，科研数据服务将会成为科研人员的重要需求。

2.3.2 基于科研项目生命周期的高校科研数据服务模式构建

针对高校“小科学”、数据量小、多学科、项目类型多、数据差异大、科研人员众多、数据素养差距大、政策配套滞后等特点，课题构建了基于科研项目生命周期的高校多层次科研数据服务模式（图 2.1）。

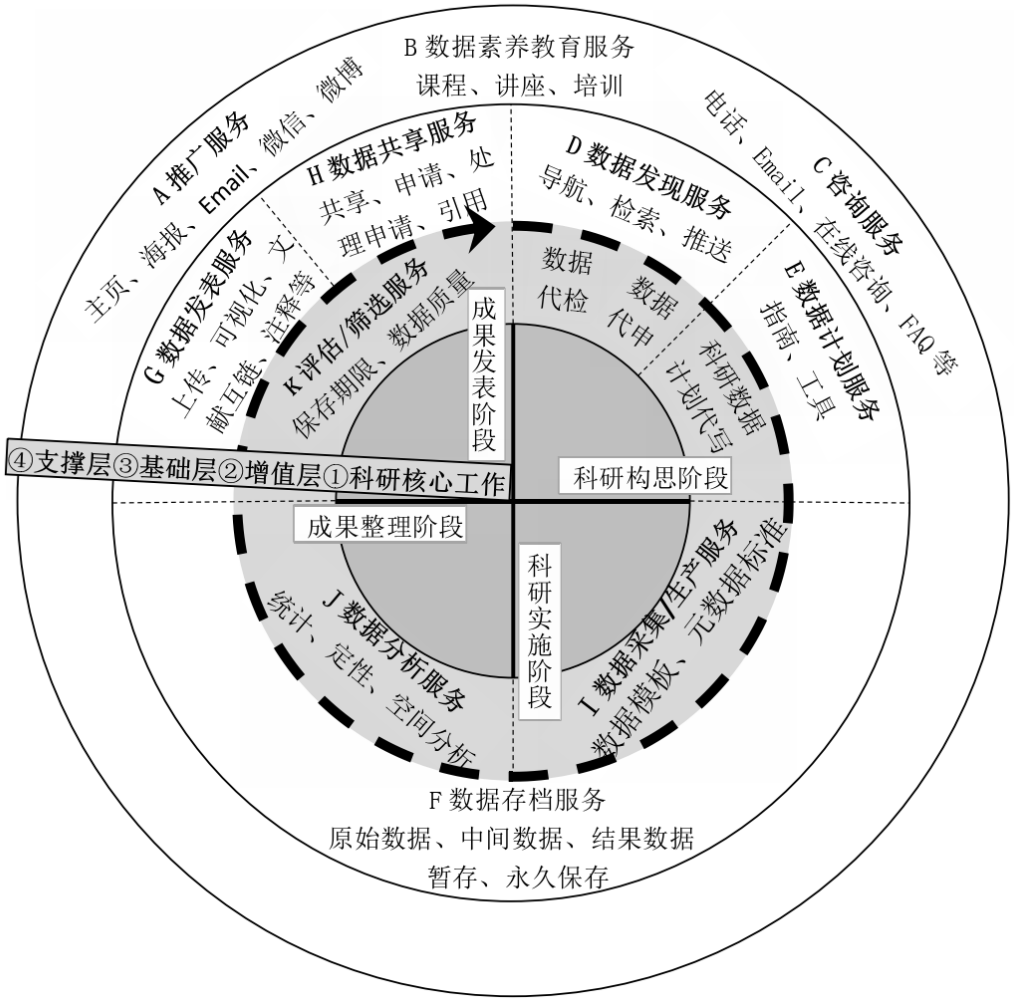

图 2.1 基于科研项目生命周期的科研数据三层服务模式

这一模式绕开高校不同学科科研数据的差别，立足于科研项目本身，按照所有项目都具备的科研构思——科研实施——成果整理——成果发表四个阶段构建。把科研数据服务视作一项服务工作，那么这项工作的服务对象应该是科研活动，他通过对科研活动中科研数据相关工作（获取科研数据、管理项目生产出来的科研数据）提供各种帮助，实现对科研活动的支持，因此以科研活动为中心构建服务框架具有天然合理性。同时这项服务的落脚点仍然在科研数据上，与科研数据服务的宗旨是相吻合的。并且由于嵌入到科研过程当中的服务，能更好的体现服务科研活动的目的，更易于被科研人员接受，激发科研人员数据服务需求，一定程度上缓解目前政策配套滞后的问题。

这一模式设计支撑服务——基础服务——增值服务三个层面。高校人员多、科研数据素养相差较大，由这些科研人员提出来的科研数据服务需求也是千差万别，课题这一模式涉及三个层次、面向所有科研人员的各种需求，让科研人员各取所需，能够很好的满足高校众多科研人员的个性化需求。

I. 基础服务

通过提供“基础服务”（见图 2.2）可以明显改善目前多数科研人员获取和处理科研数据的模式的弊端：a.可以有效的保障数据安全、最大化促进数据再利用，借助公用数据平台存档数据极大地提升了数据的安全性，同时当多数人都在平台上发表、共享数据的时候，一个良好的数据共享生态就建立起来了，这将最大化促进数据再利用；b.可以提高科研产出、有效节约大量科研经费，一方面数据重复利用（数据重新理解、数据聚合等）可以在相同科研试验投入的前提下获得更多产出，另一方面数据的开放共享可以避免大量不必要的重复试验，节约科研经费。

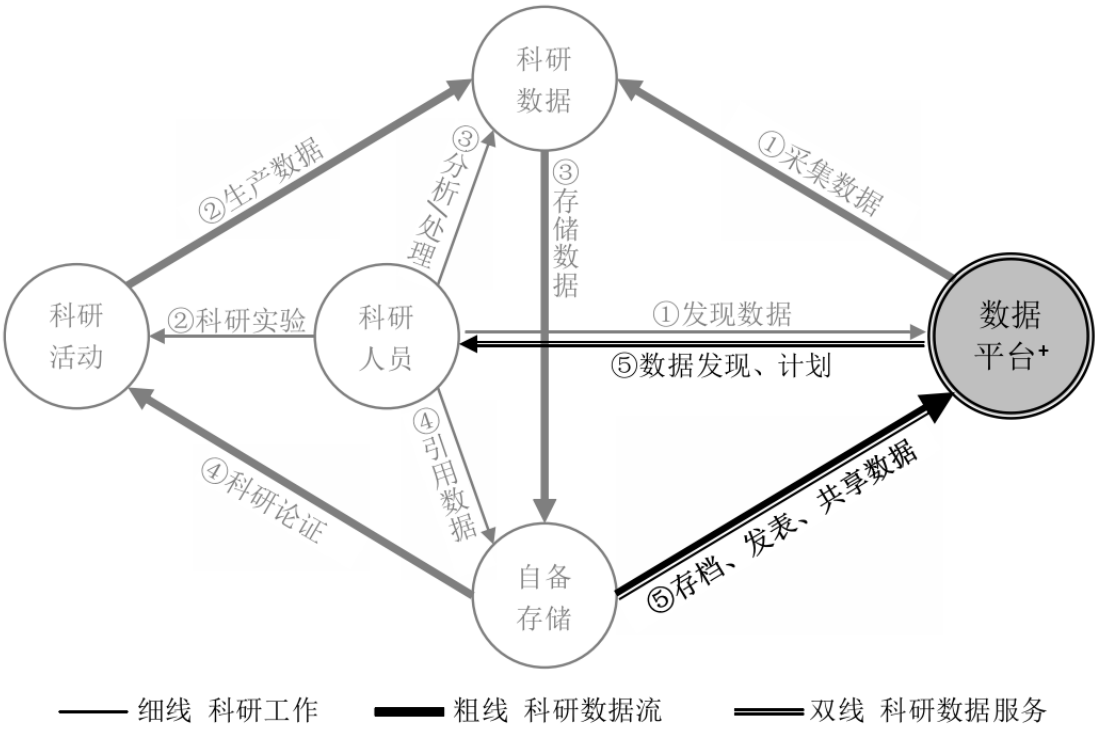

图 2.2 高校科研数据服务的“基础层次”

假使仅提供“基础服务”，那么弊端也是显而易见的。a.加重科研人员负担、耗费科研人员大量时间和精力，科研数据平台的熟悉和数据计划、存档、发表、共享都会占据科研人员大量时间，这会成为阻碍科研人员充分利用科研数据服务的重要因素；b.数据质量难以得到充分保证，由于缺乏专业人员的指导和帮助，由科研人员借助数据平台完成操作，显然缺乏必要的专业性。

II. 支撑服务

良好的“支撑服务”能够（见图 2.3），a 提高科研人员科研数据素养，通过推广、教育和咨询，让科研人员充分知悉科研数据服务、了解科研数据政策和技术、熟悉科研数据平台相关操作，适应数据密集型科研环境；b 有效促进服务利用，许多科研人员没有足够的时间和意识去了解科研数据服务，通过推广、教育和咨询，让科研人员感知科研数据服务的有用性、掌握足够的数据服务相关知识、并建立起相适应的努力预期，最终转化为科研人员对于服务的合理利用。

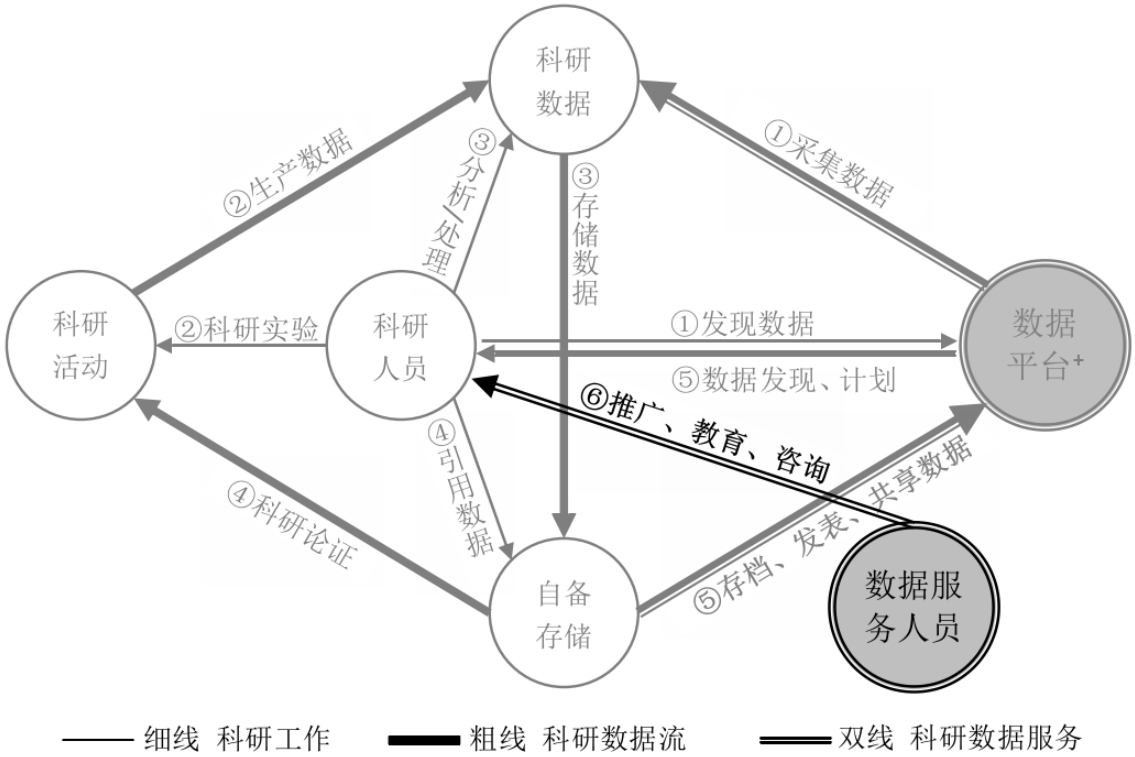

图 2.3 高校科研数据服务的“支撑服务”

但加“支撑服务”后仍然有一些问题没有解决：a 仍然耗费科研人员大量时间和精力，支撑服务让科研人员更加熟悉科研数据服务，但这是建立在额外耗费时间之上的，这仍会阻碍科研人员充分利用科研数据服务；b 数据质量仍然难以得到充分保证，支撑服务不足以让科研人员成为数据服务专家，科研人员也没有时间和精力持续更新相关知识，专业性仍显不足。

III. 增值服务

通过提供“增值服务”能够（见图 2.4）：a 为科研人员节约大量时间和精力，数据服务人员部分或全部承担科研数据获取和处理相关工作，大幅减少了科研人员相关工作，让其得以专注于科研核心工作；b 激发科研人员使用数据服务的需求；c 提高数据质量。

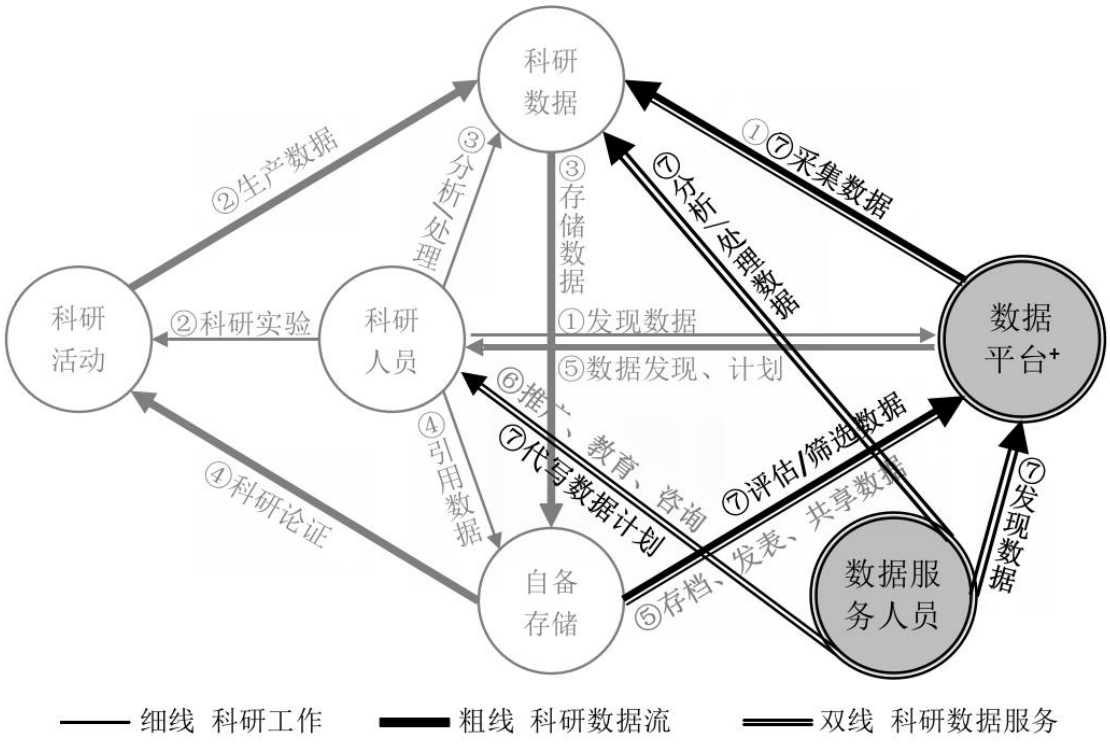

图 2.4 高校科研数据服务的“增值服务”

通过三个层次的服务，能够更好的解决现有科研人员获取和处理科研数据所面临的问题，充分保障数据质量和安全、促进数据再利用、节约科研人员时间和国家科研经费。

2.3.3 高校科研数据服务的建设路径

前面两节内容分别回答了“是否需要提供科研数据服务”、以及“提供什么样的科研数据服务”的问题，本节将回答“如何提供科研数据服务”的问题。基本思路是在总体目标和原则指导下，提出“国内高校开展科研数据服务工作的路径框架”，并对科研数据服务建设的主要工作进行逐项阐述。

（1）国内高校开展科研数据服务工作的基本思路

①推进科研数据服务工作的总体目标和原则

一旦决定启动科研数据服务建设，首当其冲的是要确定一个目标，然后进行合理规划，进而进入实施阶段。科研数据服务建设的目标包括服务的范围、服务的内容和方式等。如康奈尔大学图书馆启动其科研数据服务项目时明确告知用户，这是一个试验性项目，定位为一个面向本校学者的数据阶段型存储库（Data Staging Repository，DataStaR），建设目标是一个过渡性数据监护平台和一套完整服务方案；CALIS 三期项目中，武汉大学图书馆开展的科研数据服务明确定位为试点项目，从试点院系选择、软件选型、服务项目设计各方面均确定了有限目标，并明确告知用户。

课题为便于讨论，本研究将目标定位于通过一系列建设工作，最终要能向校内所有师生甚

至校外用户提供覆盖科研项目生命周期的全方位服务。在这个总目标之下，我们分解为若干原则，便于逐条对照制定服务建设路径。

I. 按需供给原则

按需供给是服务建设的总原则，体现的是一种服务意识，“需要什么给什么”把科研人员是否需要作为判断服务能否开展并继续下去的唯一标准，a 总体上科研人员需要什么服务就开展什么服务；b 所有科研人员最迫切需要的服务最先开展；c 科研人员暂时不太接受的服务缓一缓；d 从个人来讲按需供给就是个性化服务，通过总体上提供菜单式的若干服务选项，针对每位科研人员、每一项科研课题，由科研人员自由选择相应服务。

II. 分步实施原则

提供科研数据服务是一项系统的工程，涉及方面甚广，从总体上划分若干阶段，逐步展开相关工作，有利于项目有条不紊的实施。

III．分类开展原则

针对高校若干类别的项目，纵向课题、横向课题、校内课题、自筹经费课题、本硕博毕业论文等，每一类项目其科研数据服务需求都是不一样的，很难有一个统一的模式，需要针对不同类别项目区别开展服务；高校科研数据服务客体包括教师、学生、校外人员，每类人需求也有所差异，服务建设也要有针对性。

IV．分层推进原则

根据科研数据服务的支撑层、基础层、增值层，逐层推进相关服务有助于服务稳步开展。

V．由易到难原则

一项服务的开展是否顺利一方面取决于是否存在需求，另一方面与服务提供者也有很大关系，开展科研数据服务不得不考虑科研数据服务团队的业务娴熟程度，科研数据服务从无到有对相关提供者是一个巨大的考验，因此有必要适当兼顾服务团队，优先开展难度较低的服务项目，随着服务逐步成熟，再行提供较难的服务。

VI．逐步深入原则

从科研数据服务介入科研项目的程度来讲，介入越深的服务，越是个性化的服务，需要的人力、专业要求都越高，因此在服务开展的初期以提供共性化的基础服务为主，后期逐步深入提供增值服务。

②国内高校开展科研数据服务工作的路径框架

在前述原则指导下确立的科研数据服务建设路径分为如下 5 个阶段：

I. 前期调研论证阶段

这部分工作进行的好坏直接关系到后续工作的开展。这部分的主要工作依次是需求分析、政策调研、案例调研、经费预算，在此基础上对科研数据服务建设项目可行性做出合理评估，如果有必要建设，那么一份总体建设方案可以看作是这一阶段的成果。

II．基础设施建设阶段

基础设施建设是后续服务正常开展的保障。基础设施建设阶段的主要工作依次包括团队建设、政策制定、平台建设，完整的团队、齐备的政策、运行稳定的数据服务平台是这一阶段工作结束的标志。

III．服务试用调整阶段

对于一项新的科研支持服务，试用调整阶段是一个缓冲润滑剂。在广泛推广之前，优选重要性较低、推行难度不大、总量适度的项目（如教师的校内课题）提供试用服务，这一阶段开展的科研数据服务依次包括数据咨询、数据发现、数据计划、数据存档、数据发表、数据共享这六项服务，除数据咨询服务，其它五项服务都是基础层服务，与数据平台密切相关。通过试用阶段充分暴露服务问题、并加以调整解决，为后续服务全面开展做好储备。

IV．服务全面推广阶段

在人员、平台、资源、政策、技术调整到位的基础上，面向全校师生、所有课题（包括学生毕业论文）全面开展科研数据服务，这一阶段一方面做好服务试用调整阶段已经在开展的服务工作，另外还需开展数据推广和数据素养教育服务。

V．服务深入科研阶段

随着服务的全面推广，数据服务人员业务技术已经较为纯熟、数据平台已经逐步稳定、政策制度也已经相当完备，并且逐步形成了较为合理的科研数据服务模式。在这样的前提下，可以依次逐步开展科研数据增值服务，具体包括科研数据代检/代申、科研数据计划代写、科研数据采集生产、科研数据评估/筛选、数据分析等服务。并且服务可以逐步向校外科研人员开放。

（2）国内高校科研数据服务建设的主要工作

①科研数据服务需求调研

I．明确科研数据服务需求调研的目标和基本原则

了解科研人员真实需求是开展科研数据服务的前提和基础，是开展科研数据服务必不可少的环节。陈大庆在调研英美 30 所高校科研数据服务的基础上构建了高校科研数据服务路线，其第一步是进行需求评估，他认为“评估科研人员的数据管理需求及行为有助于确定制定数据管理政策的必要性、使用的数据管理工具的类型、数据基础设施建设的投入力度、数据管理培训的侧重点、数据管理服务的具体内容等。”

需求调研的主要目标无外乎描述事实、解释现象、预测未来，科研数据服务需求调研的目标通常定位为“分析研究人员的数据管理需求，找出当前的现状和你期望的研究数据管理服务”。

科研数据服务需求调研需要符合的基本原则大致有三项：a 客观性原则，是指科研数据服务需求调研过程中收集资料、分析资料以及得出结论都不掺杂研究者的主观因素；b 科学性原则，是指科研数据服务需求调研必须借助各门科学研究的有关成果，并建立起具有自我规律的体系；c 系统性原则，是指科研数据服务需求调研要从系统的角度出发，进行全面和充分的分析。

II. 确立科研数据服务需求调研的内容和分析框架

i. 明确科研数据服务需求调研的内容

科研数据服务需求调研的内容大致可分为四类；a 科研活动现状情况，通过对科研背景、科研项目情况调研有助于从总体上把握高校总体科研活动的规模、层次、特征，更深入理解高校科研数据和数据服务特征；b 科研数据工作现状，作为科研活动的一部分，直接针对数据获取、数据处理、数据工作执行情况进行调研，整体把握科研数据相关工作现状；c 科研数据素养现状，认知和技能决定行动，对数据认知和数据技能现状进行调研，有助于理解科研数据工作现状和科研数据服务需求；d 服务项目需求，从培训需求、服务需求、其它需求三个方面直接调查科研人员的期望。

ii．建立科研数据服务需求分析框架

科学的方式和方法有助于提高研究成果的可信度，有关科研数据服务需求，国外已经形成了数套成熟的整体解决方案，课题将其称之为“分析框架”，每个分析框架都包括了科研数据服务需求分析所要用到的各种方法、具体操作流程等内容。反观国内相关研究，多数仅是对调查问卷结果的统计分析，并没有形成系统的需求识别方法和流程。

实际操作中，首先要分析自身具体情况，作出预判，其次再根据预判优选成熟的科研数据服务需求分析框架，最后对所选择的需求分析框架分析框架做出适当微调，最后形成自己的一套分析框架。

②基础资源建设阶段的工作

I. 科研数据服务团队建设

通过基础调研阶段的可行性论证，科研数据服务进入基础资源建设阶段，这个阶段最先开展的工作就是团队建设，确定了团队才能逐步开展后续工作。

i. 确立图书馆作为统一向科研人员提供科研数据服务的窗口部门

纵观国外成熟经验，美国、英国、澳大利亚等国高校科研数据服务多由图书馆提供（分两种形式，一种图书馆独力开展，另一种其他部门协助图书馆开办），概因图书馆一直以来通过提供文献资源服务于教学和科研，数据资源与文献资源有很多相似之处，图书馆拥有经验丰富的人力资源，并且和科研人员建立了密切的联系，具有开展相关服务的天然优势；另外从国外实践来看，提供科研数据服务离不开一支专业的、训练有素的团队，多数欧美国家都在图书馆设置了专业数据馆员、数据专家，由于科研数据服务涉及面甚广、对服务人员的知识储备和操作技能有很高的要求，缺乏相关知识技能将在很大程度上制约服务的开展。

ii．确立科研数据服务团队架构

科研数据服务是一个系统而庞杂的服务，涉及到学校多个部门，离开他们的支持，服务举步维艰。结合国内高校的实际运作情况，本课题提出了高校开展科研数据服务的团队架构，整个架构含决策层、管理层、支撑层、窗口层四层结构。这种复杂的架构在当前国内高校治理现状下，有助于新服务工作的开展。

iii. 明晰团队成员的分工

作为一个复杂的团队，清晰的分工是避免内耗、顺利开展工作的重要保证。因此有必要对科研数据服务领导委员会、科研数据服务协调办公室、政策标准组、数据平台组、数据服务组这五个机构的角色、人员构成、主要工作内容做出明确规定（表 2.1），其中科研数据服务领导委员会是服务决策部门、科研数据服务协调办公室是服务管理部门、政策标准组和数据平台组是服务保障部门、数据服务组是直接面向科研人员的一线服务执行部门。

表 2.1 高校科研数据服务团队相关机构、角色、人员构成、主要工作内容

机构	角色	人员构成	主要工作内容
科研数据服务领导委员会	服务决策部门	主管副校长、图书馆负责人、科技处负责人、信息(IT)部门负责人、相关职能部门负责人、院系负责人、科研人员代表	不定期召开会议，对政策、平台、人员、经费、服务等重要事项进行表决
科研数据服务协调办公室	服务管理部门	图书馆人员（设置在图书馆）	全面领导和协调科研数据服务相关工作。
政策标准组	服务保障部门	图书馆人员、科技处人员、相关职能部门人员	调研并持续关注国内外相关政策和标准、制定和更新政策和标准
数据平台组	服务保障部门	图书馆技术部人员	负责数据平台的建设和维护
数据服务组	服务执行部门	图书馆服务部人员	直接面向最终用户负责具体服务开展工作

II. 科研数据服务相关政策规范建设

在服务团队确定以后，首要任务就是制定相关政策规范，在规范确定以后，其它工作的开展才有据可循。

本课题从总体上列出了高校科研数据服务主要政策规范框架（图 2.5），包括总体政策、数据资源政策、数据服务规范，其中总体政策是一些宏观的规范，数据资源政策包括数据标准、数据存储、数据传播、数据安全等规范，数据服务规范是对具体服务工作的约束。

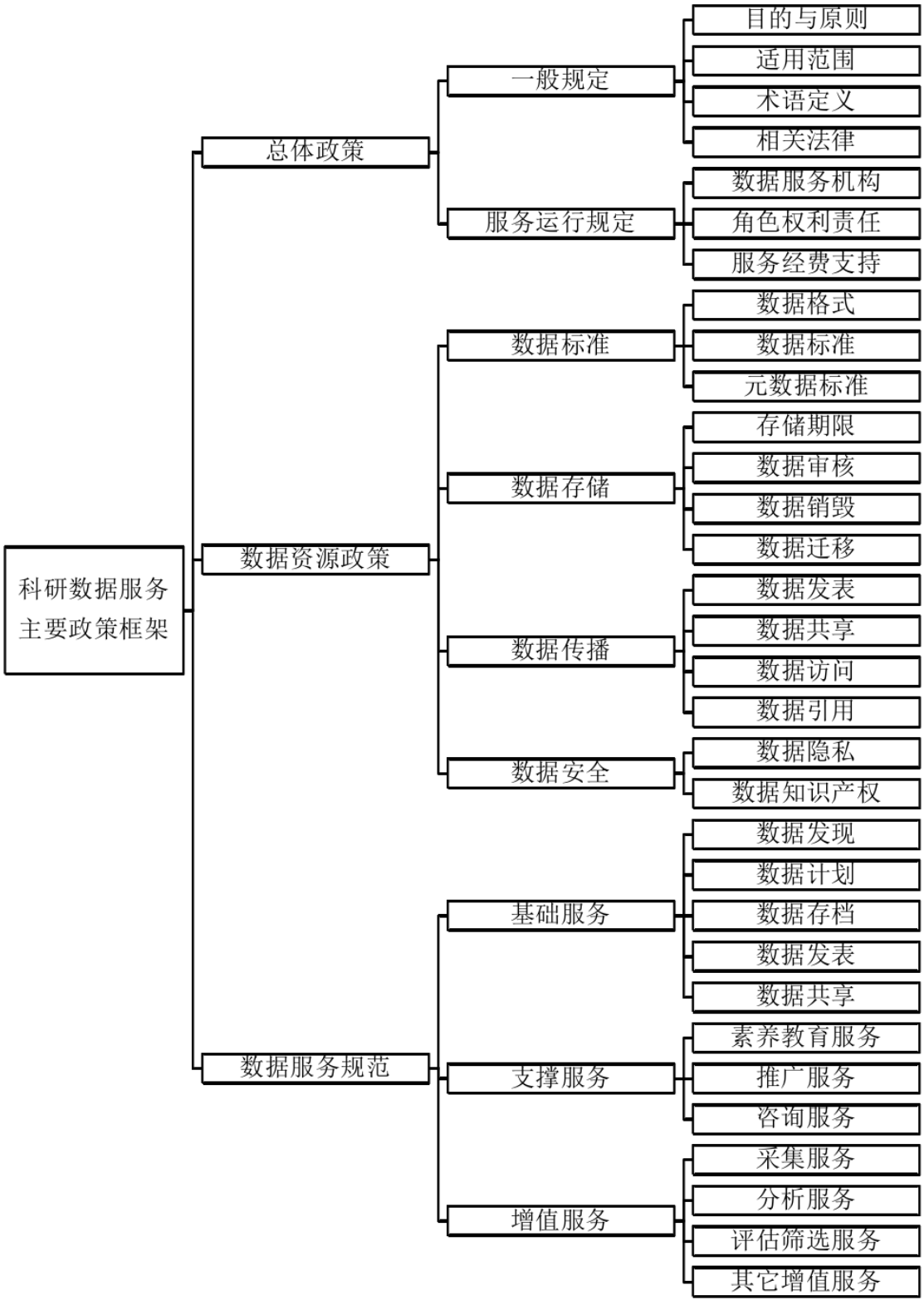

图 2.5 高校科研数据服务主要政策规范框架

III. 科研数据服务平台建设

科研数据服务平台建设是基础资源建设阶段很重要的一环，平台建设直接关系到相关信息的传递、数据资源的展示、数据服务的推广，关系到用户数据服务的操作和申请，数据平台还

关系到服务人员服务工作的开展是否方便。平台建设的方式包括：

i. 自建平台

对于服务量达到一定规模的高校，可以优选成熟的开源软件，在其基础上按照需求进行二次开发。目前已经形成了很多开源平台，如 Dataverse、Data conservancy、CKAN、Dryad、PURR、ICPSR、Genbank、Figshare、Nesstar；多数高校在实践中都是基于开源软件进行二次开发，以此为基础建设数据服务平台。

ii．联合开放

为节约经费，也为提供数据平台的使用效率，合作共建数据存储服务也是一个不错的选择。如哈佛—麻省理工数据中心（Harvard-MIT Data Center，HMDC）就是由两所学校联合建立的数据仓储中心；

iii．其他方式

对于服务量较小的高校可以灵活处理，采用租用他校平台、推荐公用平台等方式解决平台问题。在已经开展科研数据服务的高校中，就存在一些学校没有建设数据服务平台，转而向研究人员推荐外部仓储；

无论是那种方式，在平台的规划、选型和建设中一定要注意数据平台的功能是否完善（图 2.6），科研数据服务平台主要功能模块应该包括业务层和应用层。业务层面向服务人员，作为服务人员开展工作的平台；应用层则需要包括面向大众和面向个人两个模块。

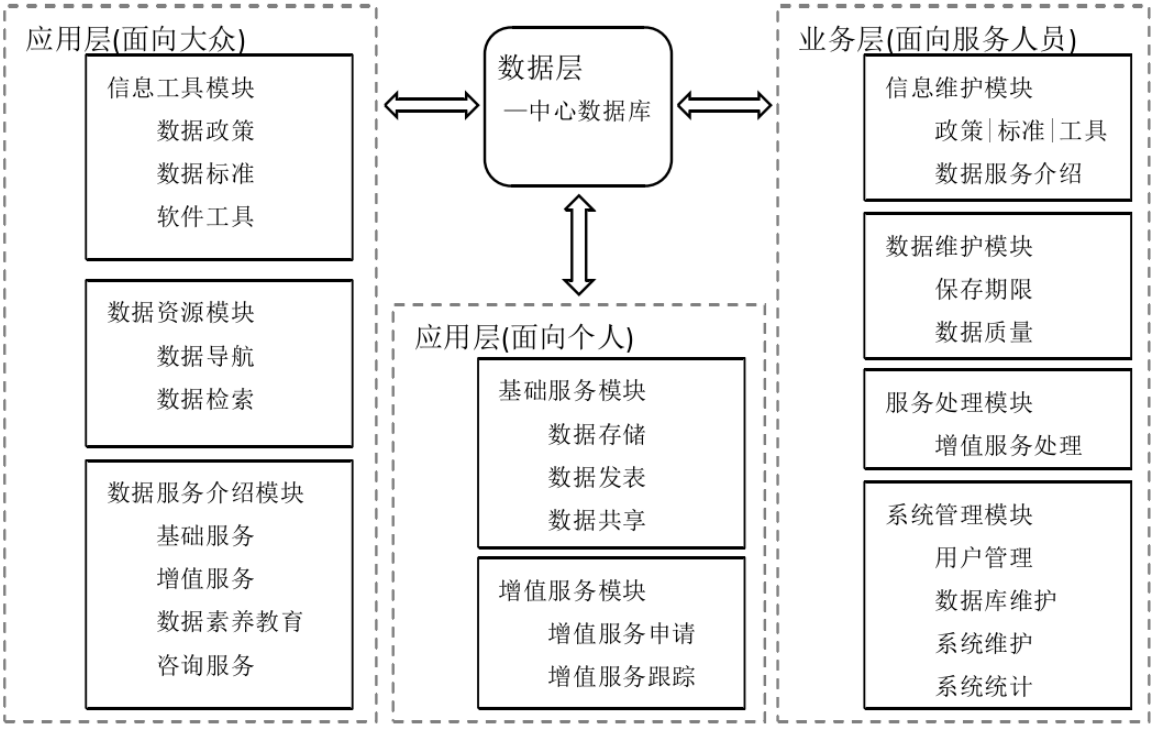

图 2.6 科研数据服务平台主要功能模块

③科研数据服务开展阶段的工作

由于在前文的论述中，已经或多或少提及科研数据服务的供给内容和方式，此处仅在前文

的基础上做一些补充和深化。

I．数据推广服务

数据推广服务，直白的讲就是宣传工作，对于国内高校而言，科研数据服务仍然属于一个新生事物，借助主页、海报、Email、微信、微博等一系列宣传手段让科研人员知道科研数据服务的存在，进而逐步培养科研数据服务意识，激发科研数据服务需求，这是相关服务工作得以顺利开展的重要保证。

II．数据素养教育

数据是一类特殊的信息资源，数据素养教育的方式可以是开设课程、讲座、培训等。对于教师和专兼职科研人员开设讲座和短期培训是比较合适的方式。对于博硕士研究生，数据素养教育可以作为信息素养教育的一部分，直接嵌入到信息素养教育课程中去。

III．数据咨询服务

数据咨询服务的模式通常是科研人员提出问题，由科研数据服务人员借助先进的手段和科学的方法，提供专业的建议或解决方案。数据咨询服务的途径可以是电话、Email、在线咨询、FAQ 等各种方式。

IV．数据发现服务

数据发现是促进科研数据重复利用的重要环节。数据发现的主要方式包括数据导航、数据推送、数据检索、数据代检、数据代申等。数据导航是指利用网页形式组织和揭示科研数据；数据推送是指主动通过 Email 等方式有针对性的向科研人员发送相关科研数据信息；数据检索主要是指在数据服务平台上设置检索功能和接口；数据代检/代申是指主动介入科研人员项目，帮助科研人员代为检索和申请数据。

V．数据计划服务

数据计划服务主要是指向科研人员提供数据计划撰写指南、数据计划撰写工具，进而帮助科研人员确定数据库类型、数据组织方式、元数据标准、数据保存和共享计划等；另外，数据服务人员也可直接帮助科研人员代写数据计划。

VI．数据存档服务

存档服务包括数据的短期暂存和数据长期保存。数据存档服务应当按照科研的情况和数据服务相关政策制度要求，针对原始数据、中间数据、结果数据分别提供不同的存储策略。

VII．数据发表服务

数据发表是数据共享的前提，数据发表包括数据上传、数据可视化、文献互链、数据注释等服务工作。数据上传是指将科研数据上载到数据平台；数据可视化是指借助现代计算机技术将数据资源以图形图像等媒体技术展示给科研人员；文献互链是指将科研数据与科研文献建立关联；数据注释则是对数据解释和说明工作。

VIII．数据共享服务

有关数据共享服务的工作包括协助科研人员制定共享策略，确立科研数据共享时间、共享

范围、共享方式；数据共享服务还包括提供在线申请、处理数据申请、提供数据引用服务等工作。

IX．数据采集/生产服务

从数据生产的源头上介入有助于从根本上提升数据质量，数据采集/生产服务正是出于这个目的。如协助科研人员创建数据收集模板，对数据格式、精度、注释等做出详细得约束，确定元数据标准也是数据采集/生产服务的一个主要方向。

X．数据分析服务

数据分析是未来科研数据服务的一个主要增长点，软件和工具不断推陈出新，科研人员的时间精力是有限的，借助科研数据服务人员对数据进行专业分析是一个不错的选择。数据分析可以是统计分析、定性分析、空间分析等。

XI．数据评估/筛选服务

科研人员在获取和处理数据的过程中，时常需要做出判断，例如数据价值大小，数据需要保存多久等。科研数据服务人员可以帮助科研人员制定数据保存期限策略，还可以按照数据质量标准对科研数据的质量和价值做出评估和筛选。

3 结论与建议

国外长期以来更多的关注科研数据共享和科研数据管理领域，有关研究已经较为成熟，近年来也出现了不少有关科研数据服务的研究，但其中不少关注的是科研数据管理服务（并非完整意义上的科研数据服务），国内相关研究多数是对国外理论研究和实践探索的跟踪，有关科研数据服务的研究也少于科研数据管理和科研数据共享，因此科研数据服务研究总体上较为落后，并不成体系。

正因如此，课题在搜集、整理、分析相关文献和资料的基础上。采用文献调查法、文献统计分析法、对比分析法等方法，从高校提供科研数据服务的必要性、高校科研数据服务模式、高校科研数据服务建设路径三点入手，对高校科研数据服务相关问题展开了系统探讨。科研数据服务归根结底可以理解为高校的一项科研支持服务，课题力图站在更高的一个视角，立足科研数据服务、吸收科研数据服务理论和实践研究成果，但又跳出科研数据服务的事务性工作，从总体上把握科研数据服务。

课题首先对科研数据概念、科研数据特征和分类、大数据时代科研人员对于科研数据的获取和处理等分别进行探讨。其次本课题对科研数据服务的概念、科研数据服务的客体和特征、科研数据服务的服务项目内容和方式、科研数据服务的内容分层、高校科研数据服务模式分别展开研究。最后本课题对推进科研数据服务工作的总体目标和原则、高校开展科研数据服务工作的路径框架、高校科研数据服务建设部分关键措施进行探讨。

研究过程中的不足，①科研数据服务选题涉及管理学、图书情报学、信息技术/计算机、行

为学、营销学等学科，这一选题于我还是很有挑战性的，限于时间和精力，在需求调研、团队建设、政策规范建设、数据平台建设等多个问题上都是简单阐述，研究不够深入；②研究的过程是一个螺旋式上升的过程，随着研究加深和论文推进，有些观点不断调整，最后限于时间和精力，只能做适当调整，带来部分内容说理不够透彻。

在许可的情况下，未来将继续沿着这个思路继续深入一些问题，①国外的研究热衷于总结出各种模式、框架等，沿着这一思路，未来可以深化文章中所有的模式、框架，使之更为合理；②相比与基础服务、支撑服务，增值层的服务直接嵌入到科研活动过程当中，这部分研究也是较少的，可以继续进一步深化；③有关培养和激发科研数据服务需求的研究，科研人员由于多种原因不了解科研数据服务，或者不愿意使用，最后导致科研数据服务需求不足，如何通过一系列措施将科研人员潜在的科研数据需求激发出来是一个具有实践意义的选题。

4 项目成果（发表的文章、开发的软件、取得的实践效果等）

在项目研究及建设过程中，课题组成员以课题研究为基础，已发表学术论文 1 篇，并被 SCI、EI 收录检索，具体发表文章信息如下：

Zhang, JJ & Chi Y. Data Management and Service Mode of Library Based on Data Mining Algorithm[J]. *Scientific Programming*, vol. 2022, Article ID 2414830, 2022.

SCI、EI 检索报告如下：

Web of Science™

1 record(s) printed from Clarivate Web of Science

第 1 条，共 1 条

标题: Data Management and Service Mode of Library Based on Data Mining Algorithm

作者: Zhang, JJ (Zhang, Jingjing); Chi, Y (Chi, Yang)

来源出版物: SCIENTIFIC PROGRAMMING **卷:** 2022 **文献号:** 2414830 **DOI:** 10.1155/2022/2414830 **出版年:** SEP 21 2022

Web of Science 核心合集中的 "被引频次": 0

被引频次合计: 0

入藏号: WOS:000892084800004

文献类型: Article

地址: [Zhang, Jingjing; Chi, Yang] Northeast Agr Univ, Lib, Harbin 150030, Heilongjiang, Peoples R China.

通讯作者地址: Chi, Y (通讯作者)， Northeast Agr Univ, Lib, Harbin 150030, Heilongjiang, Peoples R China.

电子邮件地址: zhjj623@163.com; cy@neau.edu.cn

Web of Science Index: Science Citation Index Expanded (SCI-EXPANDED)

ISSN: 1058-9244

eISSN: 1875-919X

输出日期: 2023-01-04

End of File

1. Data Management and Service Mode of Library Based on Data Mining Algorithm

Accession number: 20224112858827
Authors: Zhang, Jingjing (1); Chi, Yang (1)
Author affiliation: (1) Library, Northeast Agricultural University, Harbin, Heilongjiang; 150030, China
Corresponding author: Chi, Yang(cy@neau.edu.cn)
Source title: Scientific Programming
Abbreviated source title: Sci. Program
Volume: 2022
Issue date: 2022
Publication year: 2022
Article number: 2414830
Language: English
ISSN: 10589244
CODEN: SCIPEV
Document type: Journal article (JA)
Publisher: Hindawi Limited
Abstract: Data management for large-scale data library services with mining procedures improves the availability and readiness of heterogeneous sources. The heterogeneous data sources are assimilated as a single entity through mining procedures to meet the data demands. This article introduces connectivity-persistent data mining method (CDMM) to improve the data handling precision with boosting availability. The proposed method relies on federated learning for identifying the service demands, thereby providing data mining. The learning paradigm accumulates information on shared data library existence over various services. Based on the availability, further mining demands are forwarded to the data management system. If the existence verified by the federated learning is adaptable, then sharing-enabled mining is endorsed for the connected users. The data management then augments several heterogeneous shared libraries to meet the mining requirements. This process is reversible based on the service mode and existence. Therefore, the proposed method improves data availability with less mining and access time and fewer failures. © 2022 Jingjing Zhang and Yang Chi.
Number of references: 29
Main heading: Information management
Controlled terms: Data handling - Data mining - Libraries
Uncontrolled terms: Data library - Data mining algorithm - Data mining methods - Heterogeneous data sources - Heterogeneous sources - Large scale data - Library services - Management modes - Service demand - Service mode
Classification code: 723.2 Data Processing and Image Processing - 903.4.1 Libraries
DOI: 10.1155/2022/2414830
Compendex references: YES
Database: Compendex
Data Provider: Engineering Village

Hindawi
Scientific Programming
Volume 2022, Article ID 2414830, 12 pages
https://doi.org/10.1155/2022/2414830

Research Article

Data Management and Service Mode of Library Based on Data Mining Algorithm

Jingjing Zhang and Yang Chi

Library, Northeast Agricultural University, Harbin 150030, Heilongjiang, China

Correspondence should be addressed to Yang Chi; cy@neau.edu.cn

Received 9 July 2022; Revised 8 August 2022; Accepted 20 August 2022; Published 21 September 2022

Academic Editor: Juan Vicente Capella Hernandez

Data management for large-scale data library services with mining procedures improves the availability and readiness of heterogeneous sources. The heterogeneous data sources are assimilated as a single entity through mining procedures to meet the data demands. This article introduces connectivity-persistent data mining method (CDMM) to improve the data handling precision with boosting availability. The proposed method relies on federated learning for identifying the service demands, thereby providing data mining. The learning paradigm accumulates information on shared data library existence over various services. Based on the availability, further mining demands are forwarded to the data management system. If the existence verified by the federated learning is adaptable, then sharing-enabled mining is endorsed for the connected users. The data management then augments several heterogeneous shared libraries to meet the mining requirements. This process is reversible based on the service mode and existence. Therefore, the proposed method improves data availability with less mining and access time and fewer failures.

1. Introduction

Data mining is a process that extracts certain patterns and useful details from a large set of data. Data mining provides the necessary set of data for the analysis process. Various methods and techniques are used to perform the data mining process. Data mining is a complicated task in every application [1]. Data mining also identifies the problems identified by the data analysis process. A data management system for mining services is a crucial task that manages a huge amount of data. Data management is a process that protects, store, collect, organize, and manage data that provide an appropriate set of data for various processes [2]. Data mining services are a process that converts the raw data into a useful set of data that is used for further processes. Various management services are used for the data mining process using the machine learning (ML) approach [3]. A data management system improves the performance and efficiency rate of the system, improving the accuracy rate in the decision-making process. Data management systems manage the data collected by an application and organization. Storing and managing a data management system is mostly used for data mining. Data mining services and details are handled by a management system [4, 5].

Various data mining types are available to identify the dataset's important patterns. An organization widely uses the service demand-based data mining method. The data mining process plays a major role in every organization that helps enhance an organization's performance and feasibility [6]. The organization gives requirements and preferences that provide a set of demands over the data mining process. The service demand-based data mining process provides an accurate dataset for the decision-making process that reduces the failure rate [7]. Organizations demand a certain set of services for the data mining process. The real-time data mining process is a complicated task to perform in every management system [8]. The classification method is used in the service demand-based data mining process. The classification method classifies the dataset by combining it with given service demand. Various demands and requests are

demanded by an organization for the data mining process. Companies and industries demand a certain set of services that improve the accuracy rate in the data mining process [9, 10].

Machine learning (ML) techniques are widely used for various applications to perform prediction and analysis. ML techniques improve the accuracy rate in both the analysis and prediction process. ML techniques are also used in data mining to enhance the service accuracy rate. ML technique-based data mining process identifies the important features and patterns from a huge set of data [11, 12]. The convolutional neural network (CNN) algorithm is commonly used for data mining. The feature extraction process is used in CNN to extract the features presented in a given raw dataset [13]. The classification process classifies the features extracted from the feature extraction process. CNN predicts the actual data necessary for an application [14]. The support vector machine (SVM) algorithm is also used for data mining. SVM first trains the dataset with an important set of features collected by the analysis. SVM reduces the latency and error rates in the computation process, which improves the efficiency rate of the system. The data analysis process analyzes the raw data stored in the database [15]. The main contribution of CDMM is as follows.

(i) The suggested method focuses on federated learning for recognizing the service requests and consequently enabling data mining. The learning paradigm accumulates information about shared data library presence over numerous services.

(ii) The data management then augments numerous heterogeneous shared libraries to match the mining needs. This process is adjustable based on the service mode and existence.

(iii) Therefore, the suggested strategy improves data availability with less mining and access time and fewer failures.

2. Related Works

Huang et al. [16] introduced a new algorithm for fast mining frequent patterns using a distributed computing system. Frequent pattern mining identifies the important patterns that are presented in a given dataset and reduce the latency rate in the analysis process. The big data analysis process is used here to analyze the huge amount of data and produce an optimal dataset for further data mining. The proposed method improves the accuracy rate in the execution process, enhancing the system's performance. The proposed method reduces time and energy consumption in the execution process.

Xie et al. [17] proposed an information filtering and mining method for big data analysis. A support vector machine (SVM) algorithm is used here to analyze the data necessary for the mining process. The proposed method is mainly used for the retrieval process that retrieves educational images. Certain features and patterns are identified by filters that produce an optimal dataset for further analysis. The proposed method improves the performance rate and efficiency of the system.

Obregon et al. [18] introduced the data mining information as a flow method for social networking services (SNSs). The proposed discussion flow model identifies the data and provides appropriate details for the data mining process. Data mining captures interaction among communities, producing effective information about discussions. The proposed method enhances the feasibility and reliability of the system. The proposed method reduces the complexity rate and improves the mobility rate of SNS.

Bhattacharya et al. [19] proposed a mobile blockchain (MB) based data mining method as-a-service (MB-MaaS) for the Industrial Internet of Things (IIoT). MB is used here to enhance the effectiveness of rata in the analysis process. The proposed method identifies the group discussion and interaction of users in IIoT. The experimental results show that the proposed method achieves a high accuracy rate in the mining process, which improves the system's performance.

Zhang et al. [20] introduced a massive data mining-based method for mobile libraries. The filtering technique is used here to filter the candidate's available datasets in mobile libraries and produce a feasible set of data for further process. The Apriori algorithm is used here to provide optimal rules for the candidates, reducing unnecessary problems in the management system. The proposed method reduces energy and time consumption in the computation process. The proposed mining method also improves the execution time of the system.

Dhelim et al. [21] proposed a personality-aware hybrid filtering-based mining method for a social network. The personality filter first filters the traits and personalities of users and produces necessary information for the mining process. The data analysis process collects the data available in a social network that provides appropriate data for the mining process. The proposed method maximizes the accuracy rate in the data mining process that provides appropriate services to the users.

Wang et al. [22] introduced a new framework for library services and immigrant needs. The proposed framework identifies the cause of problems that are occurred in libraries. Social networks provide necessary information about the candidates, reducing the time consumption rate in the searching process. Finally, the proposed framework provides various guidelines and rules for libraries that improve the appropriate services to the users.

Xiao et al. [23] proposed a fine-grained sentiment analysis-based preference mining method. The sentiment analysis approach finds out the important emotions and characteristics of users. User features are identified by a pretraining language model that produces a feasible set of data for preference mining. Both numerical and text-relation information is analyzed by preference mining, reducing the execution process's latency rate. The proposed method achieves a high-performance rate in providing services for the users.

Peng et al. [24] introduced a fuzzy convolutional neural network (FCNN) based on big data mining and analysis (BDMA). The feature extraction approach is used here to extract the important features available in the dataset.

The feature extraction method collects an appropriate dataset for the big data analysis. The FCNN algorithm is mostly used for the recognition process that enhances the system's feasibility. The proposed method maximizes the system's effectiveness and efficiency rate, improving the accuracy rate in the big data analysis process.

Alkathiri et al. [25] proposed a multidimensional data mining method using the MapReduce technique for a distributed environment. The MapReduce technique is used here for the ecosystem data analysis process that finds the features presented in a given dataset. Machine learning (ML) techniques are also used here to enhance the system's feasibility. The proposed method reduces the error rate in the data mining process, improving the system's performance.

Deng et al. [26] introduced a jointed neural network-based multimedia data stream is an information mining model. The soft clustering technique is used here to cluster the huge data available in the database. A joint neural network is implemented here to train the dataset necessary for the data mining process. The proposed data mining approach addresses the problems presented in an application. The proposed model achieves high efficiency and effectiveness rate in the mining process.

Ju et al. [27] proposed a data mining-based commodity recommendation method for online shopping. The proposed method is mostly used in e-commerce and online shopping applications. The commodity recommendation method identifies users' preferences, requests, and browsing history that provide relevant details for an application. The data mining approach analyzes the given set of data and produces a feasible set of data for the recommendation process. The proposed method improves the performance and feasibility rate of the online shopping system.

Zhou et al. [28] introduced a new data mining approach using particle swarm optimization (PSO)-based back-propagation (BP) neural network. Internet of Things (IoT) is used here to enhance the communication process among users and organizations. PSO is used here to train the dataset necessary for the data mining process. IoT collects real-time data that users produce. The proposed method increases the accuracy rate in the prediction and analysis process.

2.1. Proposed Connectivity-Persistent Data Mining Method. The data source repository is the maintenance of databases by collecting data from multiple sources meeting the objective function. It is a database infrastructure that aggregates, manages, and stores datasets mined for data analysis. It makes sharing data easier by managing it and maintaining metadata for the study of data. The aggregated data are reviewed for the type of data based on which the data are stored. The data in the repository are loaded with an increasing volume of data. In Figure 1, the proposed method is illustrated.

Service mining is influenced by federated learning to validate its existence for further sharing. This learning further operates on different service demands. If any deficiency is found, a data management system ensures data existence and availability for varying users (refer to Figure 1). The request-based services from users are generated in a particular time slot where the total number of requests r from the users is denoted as$\omega_r(t)$. The request from the users allocated to the data source repository s is denoted as$\gamma_{rs}(t)$. ${\gamma_{rs}}^{\max}$ be the number of maximum requests from users to the data source repository as shown in the following equations:

$$\omega_r(t) \leq {R_r}^{\max}, t \in [1,], \tag{1}$$

$$\omega_r(t) = \sum_S \gamma_{rs}(t), t \in [1, \tau]. \tag{2}$$

To handle the requests, the capacity$\mu_s(t)$ of the data source repository with its pricing $\rho_s(t)$ of data to be provisioned is calculated. Thus, the cost$\partial(t)$ of the data source repository for the request is obtained from

$$\partial(t) = \sum_s \mu_s(t).\rho_s(t). \tag{3}$$

The delay in addressing the request to the data source repository based on the quality of experience is calculated considering the network and queuing delay. The following equation denotes the delay of the network:

$$s = s_{nw} + s_{qe}. \tag{4}$$

The network delays_{nw} and the queuing delays_{qe} to fulfil the request depending on the factors such as transmission delay and propagation delay. The queuing delay is obtained from the workload network delay on the distance between the user and the data source repository. The delay in making a decision incurs further delay, which is represented ass_{dm}:

$$\chi_s(t) = \sum_{\beta_s(t)} s_{nw} + \sum_{\beta_s(t)} s_{qe} + s_{dm} \tag{5}$$

$$\chi_s(t) = \gamma_{rs}(t) s_{nw}(r,s) + \sum_{\beta_s(t)} s_{qe} + s_{dm}. \tag{6}$$

From 5the above equations, $\beta_s(t)$ is the request allotted to the data source repository. The network delay for the request is as$s_{nw}(r,s) = p.(s_{rs})^{v}$. s_{rs} is the distance between the user request to the data source repository. p, v are the parameters considered to scale the distance and maintain the function's convex property. The decision-making based on the delay factors for data existence verification is presented in Figure 2.

The user requests are influenced by $\mu_s(t)$and s such that $\omega_r(t)$ is sustained for the entire allocation intervals. The data availability and existence are verified $\forall Interval \in (1,n) such\ that\ {r_{rs}}^{\max}$ is satisfied. The learning process relies on $\chi_s(t) such\ that\ s_{nw}\ and\ s_{que}$ are distinguished for their existence (Figure 2). The queuing delay for the request allocated to the data source repository is obtained using the following equation:

$$\sum_{\beta_s(t)} s_{qe} = \max\left[K_s(t) - \rho_s(t)\mu_s(t)\sigma, 0\right]. \tag{7}$$

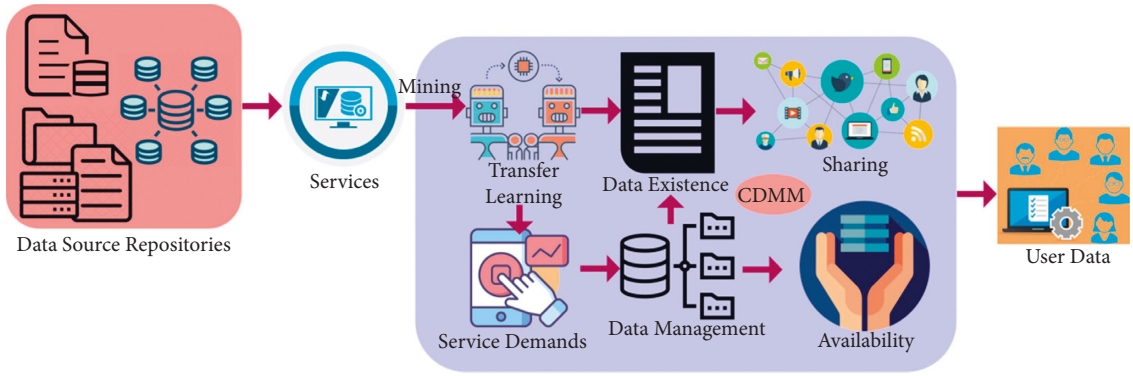

Figure 1: Proposed method illustration.

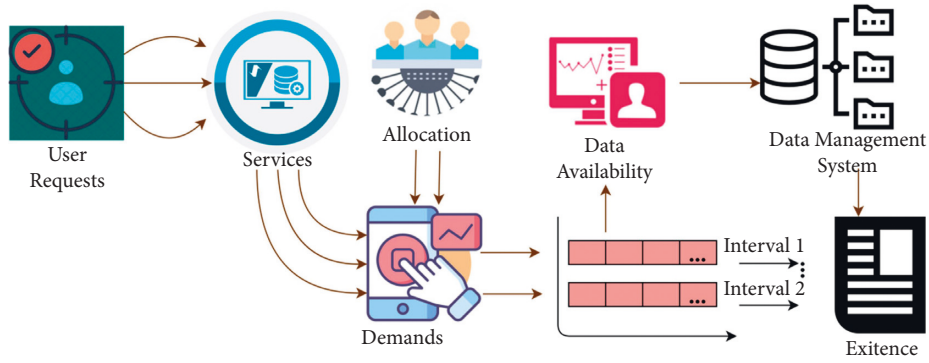

Figure 2: Decision-making for data existence verification.

The workload to be processed is represented as $K_s(t)$ for the data source repository at a particular time slot. The service time provisioned by the data source repository allotted for the request is given by$\rho_s(t)\mu_s(t)\sigma$. From (7), the deficiency of the service time at the data source repository is obtained bymax $[K_s(t) - \rho_s(t)\mu_s(t)\sigma, 0]$. Thus, the upcoming request from the users must have to wait for their request to be processed. To maintain fair processing of request based on the heterogeneity of the data, the quality of experience by the users is calculated by considering the tolerable delay and the actual delay. Thus, it can be defined as in the following equations:

$$\psi(\zeta, \phi) = \frac{\psi_{\max}}{\phi(a-1)}[a.\phi - \zeta], \zeta > \phi, \tag{8}$$

$$\psi(\zeta, \phi) = \psi_{\max}, \zeta \leq \phi, \tag{9}$$

$$\psi(\zeta, \phi) = 0, \zeta > a.\phi. \tag{10}$$

From the above equations, the tolerable delayϕ and the actual delayζ of the request in a particular time slot are represented with$\psi_{\max}$. The above equations denote the quality of experience by parameter$\psi_{\max}$. The user is processed before a tolerable delay, and then, the requests from the users are mentioned with$\psi_{\max}$. If the request is not processed within the tolerable delay, then it is considered that the users are not fulfilled and the waiting time of the users is expired, and a is the parameter that mentions the rate of declination representing the quality of experience. Based on the conditions above, the quality of experience by the user request in the data source repository within the time slot is defined by

$$d_s(t) = \psi(\zeta_s(t), \phi). \tag{11}$$

Based on the above estimation, the validations (8), (9), and (10) are performed using the federated learning model. This is depicted in Figure 3.

The learning induces multiple ψ as defined in (8), (9), and (10) for different$\omega_r(t)$. Based on the sharing output, user service mining and allocations are performed. This requirement is fulfilled based on the availability factor. The delay and existence impacts are mitigated using the maximum sharing ratio and learning implication (refer to Figure 3).

2.2. Learning Implications for Data Management. Federated learning is a technique where devices are decentralized with collaboration processing service demands considering user requests. The networks with several users have been partitioned based on their interests. This number of users share the data among themselves. Data resembling common interests among the users have been identified to verify the available data. If similar data are available, then the data are shared in a decentralized manner. The model with users $\{U_1, ...U_n\}$ and their data is $\{I_1, ..., I_n\}$. These users

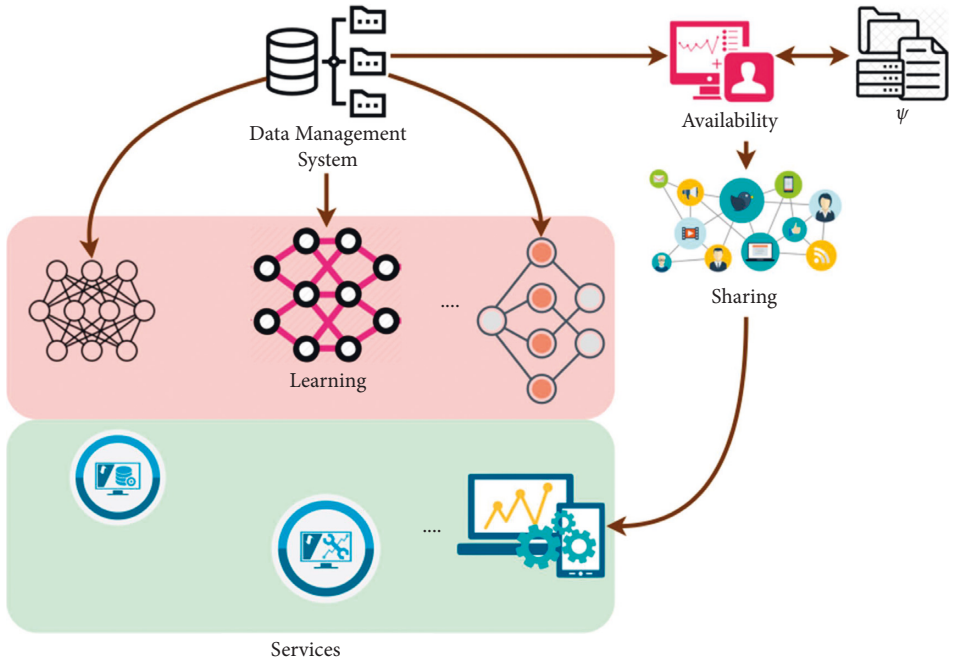

Figure 3: Federated learning model for service experience.

with the data information collaborate to identify the existence of data. The users in the network are combined with the data $I = \{I_1 \cup .... \cup I_n\}$, which is used to train the model M.

The users in the network share this model used for training; for each new data in the network among the users, a common interest procedure is to be followed. In the proposed, horizontal federated learning is considered where the users in the network communicate with each other to update the model M.

Based on the data from the users, a common interest group is created, enhancing network efficiency. The users in this group have their data. By aggregating the data from all the members, a dataset is generated. Thus, each common interest group maintains its own set of data within it. If the user in the network wishes to leave the common interest group, the user may leave with the data. Contrarily, if a user wishes to join the common interest group, the data are verified by some users in the common interest group for the relevant data. Each common interest group has its reputation for maintaining relevant or accurate data. The rewards are shared among the users in the common interest group based on the size and the relevant data they offer.

The common interest group in the network with model M., *the number of users in the common interest group,* update their model M. The proposed model improves when the users join the common interest group, so the data availability for the users also improves. Each user in the common interest group is provided to access the shared model M. The users use the model to calculate the existence of data by finding the similarities between the requested data by the users and the availability of data in the common interest group. A cosine similarity index is used to find the potentiality of the data by identifying the similarities between the requested data and the available data, as shown in the following:

$$\cos\theta = \frac{\overrightarrow{\tau}.\overrightarrow{\kappa}}{\overrightarrow{\tau}\,\overrightarrow{\kappa}}. \tag{12}$$

The availability of data is accepted only when most users in the common interest group find the resemblance of data between the requested data and the available data. The users in the common interest group are provided with rewards based on the amount of data that is being made available. The users in the common interest group must be made available with some sort of data by generating the data and updating the model M. Else, the user in the group might be expelled from the group. The users are requested to maintain some sort of space for the data allocation. The data from other users in the common interest group are stored in the maintained space. The subset of the data sent to the other users in the joint interest group is checked for relevancy. The data are verified whether it remains fixed to maintain the data within the common interest group. It asks for recommendations from common interest groups to ensure the availability of the data. Each user is provided with some functions to maintain the reliability of the users in the common interest group. Each user interacts with a common interest group; the data are shared with its functions key. Suppose these function key does not match with the available function key list. In that case, a warning update is provided, which is shared with model M. On receiving this model update, all the users in the common interest group verify its function key. If the function key fails, the corresponding user is removed from the common interest group. The learning process for$M$ in maximizing data sharing for different mining requests is presented in Figure 4.

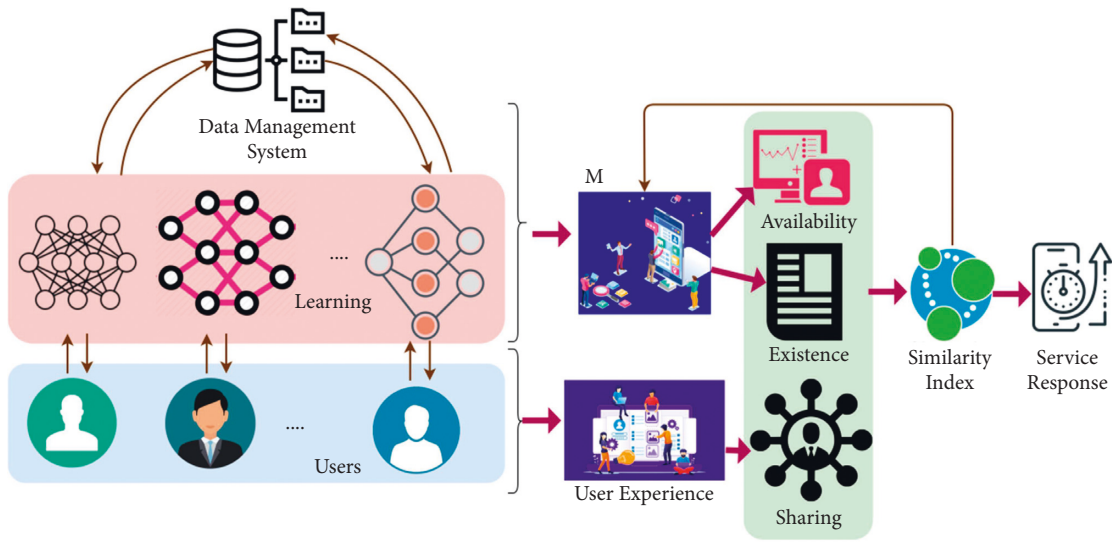

Figure 4: Learning Process for**M**.

The management system eyes on M for different ψ such as sharing for availability and existence. This M is modified based on $\cos\theta$ such that service responses are granted with better mining outcomes. The allocated requests are granted from the mining demands (Figure 4). Once the existence of data is verified with the cosine similarity, it is allocated for the request placed by the users in a particular time slot. Suppose the requested service is not available within the common interest group. In that case, a service demand is placed upon the request based on which, using the federated learning, the service demand is addressed. The data management system using federated learning enables the storage facility to enhance the network performance minimizing the access time. It is designed to operate asynchronously, making the proposed technique more flexible. It maintains a model at both the data source repository and the user side. On registration of users with the data source repository, a global model is designed. Using this global model, the data management makes arrangements for a suitable space to store the data based on the request. If the available storage space is enough, then the version information with the model data is stored. If the higher data version is available, then the model is updated. It gets updated to the advanced version from the existing version of data. The data source repository has the privilege of designing its global model and sharing it with the users. On sharing the global model, the local model gets updated. The version of the global model is initially checked before updating the local model.

The user requests the data source repository by importing the global model data based on which the local model data are generated. In the absence of a request from the user regarding the version update, the data management system monitors the version and sends the update to the users. The data source repository aggregates the updated local model from users concerning the corresponding version of the global data. It maintains the aggregation until adequate quality is obtained. This proposed method, namely, connectivity-persistent data mining method (CDMM) managed by storing the characteristics of data from the users where the requests are stored. The data source repository uses this information to process the service demands and allocate the required data. The requests from the clients include parameters such as request ID, version of the models, and device ID. The request ID represents the identity of the request where the users and the data source repository perform a task.

The data source repository addresses the request of the users by allocating the data. The data management system searches for the data using federated learning and provides the data to the users. The version of the global model is used to update the local model; the global model gets updated by connectivity between the users and the clients. Based on the model updates, the users and data source repository manage the service demands based on the requests. The device ID represents the specific ID given to the users. A particular user can be provided with the requested data using this device ID. The corresponding data communication based on the global and local model is performed with these parameters. The proposed method achieves better data available to the users by reducing the dependency on the traditional request/service demand with minimal access time and fewer failures by providing asynchronous peer-to-peer communication between the users and the data source repository. The self-analysis for varying capacity, similarity index, and demand factors is presented in Figures 5, 6, and 7, respectively.

Figure 5 presents the analysis of cost factors and requests allocated for the varying $\mu_s(t)$. This method allocates $\omega_r(t) \forall s_{nw} and\ s_{que}\ such\ that\ X_s(t)$ is performed for the increasing mining requests. The $\beta_s(t)$ is validated based on $\psi(\zeta, \phi)$ such that ϕ is accounted for maximizing $d_s(t)$. Therefore, the allocations are maximized in intervals $\in (1, n)$. The learning segregates existence and availability for the requests such that $\sum_{\beta_s(t)} s_{que}$ is reduced. This single factor

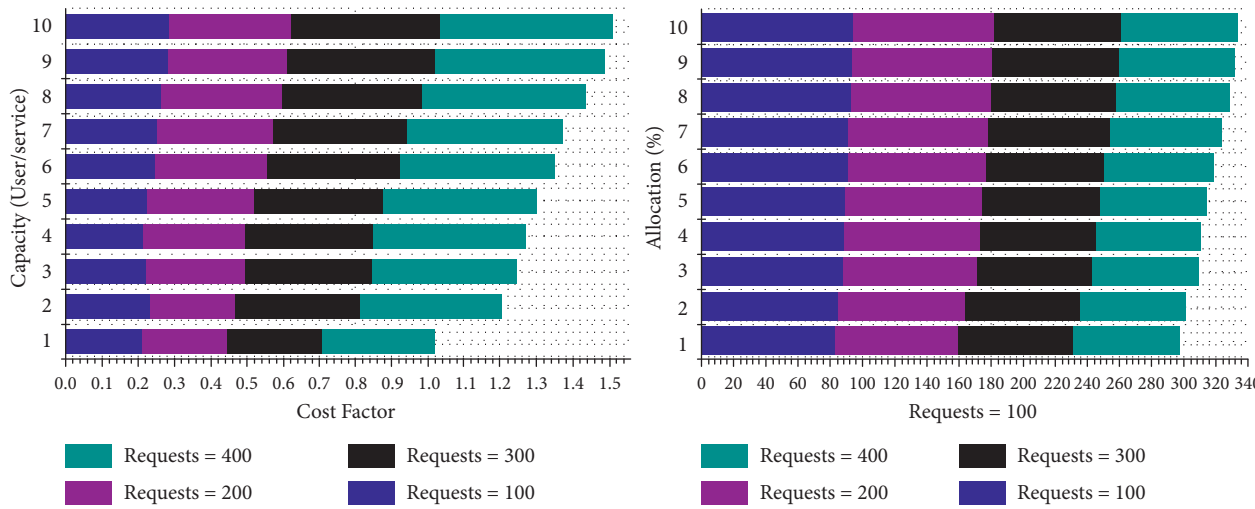

Figure 5: Analysis for cost factor and allocated (%) by varying capacity.

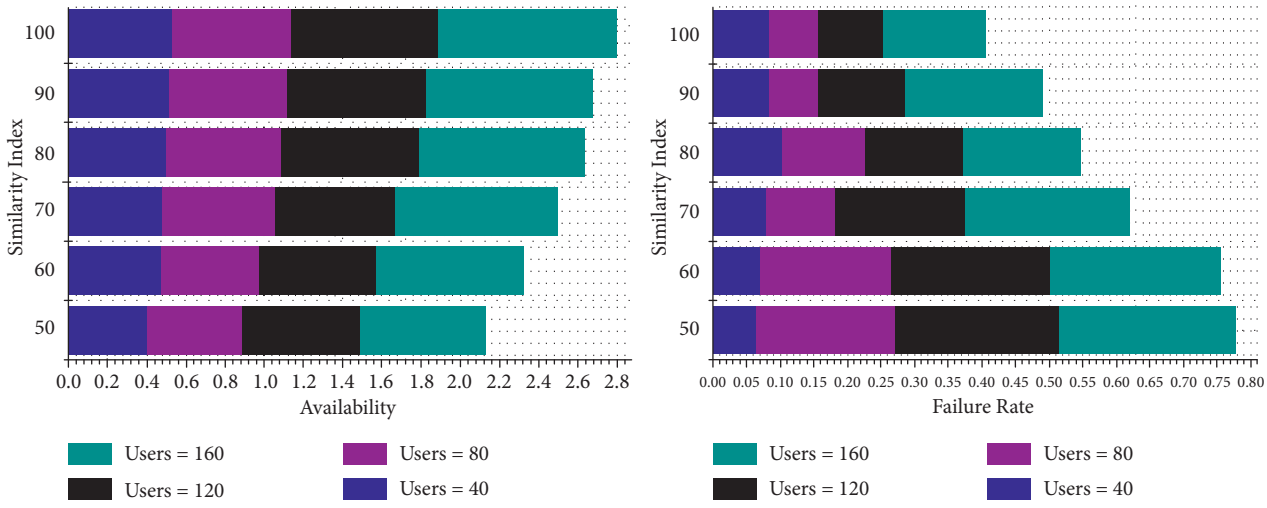

Figure 6: Analysis for availability and failure rate by varying the similarity index.

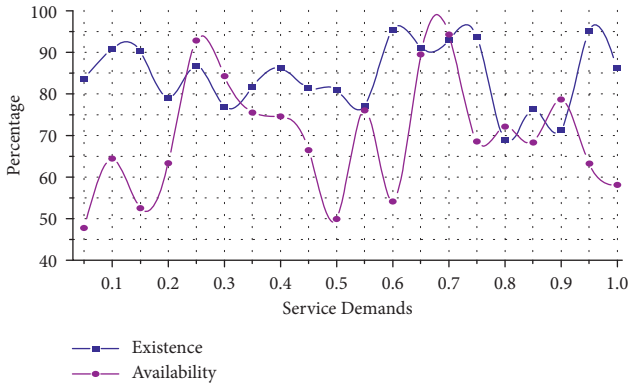

Figure 7: Analysis for existence and availability by varying service demand.

maximizes the responses by reducing the wait time for which $\partial(t)$ is reduced. Based on the $\chi_s(t)$ and learning output, the further $d_s(t)$ is performed. In the process, $\psi(\zeta, \phi)$ is used for maximizing the allocation.

The analysis of availability and failure rate for the varying similarity index is presented in Figure 6. The proposed method maximizes availability by reducing cost and s. In the federated learning for $M, \beta_S(T)$ is maximized for which existence is verified. If the verification fails, then $\varnothing$ is analyzed, and hence, availability is maximized. Therefore, the allocations are performed to improve the allocations post $\cos\theta$ and $d_s(t)$. The failures based on ${\gamma_{rs}}^{\max}$ is rectified by assigning $\partial(t) less\, \chi_s(t)$ such that new allocations are performed. The demands are supported in achieving fair sharing depending on the available sources. As the sharing increases, the availability is maximized by reducing failures.

Figure 7 presents an analysis of the existence and availability of the varying service demands factors. This analysis relies on $\mu_s(t)$ and $\partial(t)$ such that $\beta_s(t)$ is performed. However, the existence is high compared to the availability such that ψ determines its allocation. This is required by the M for further sharing and $\cos\theta$ analysis. Based on this, further, allocation is performed to improve availability.

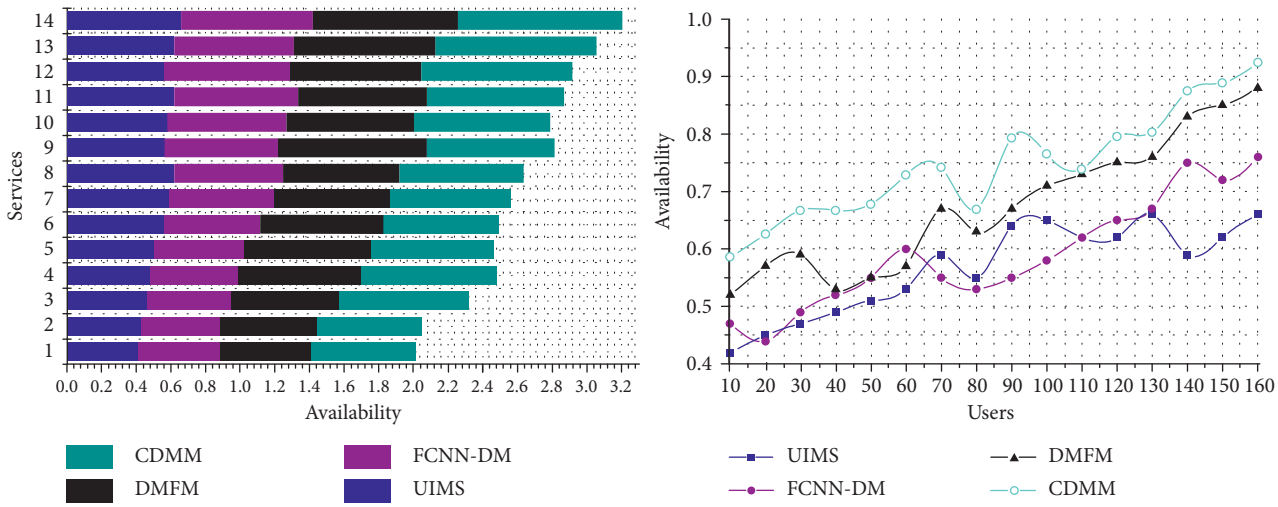

Figure 8: Availability comparison.

2.3. Performance Assessment. This section discusses the comparative analysis results of assessing the proposed CDMM using the dataset [29]. This dataset contains Flipkart product data classified under 16 fields for 30 K products. The mining process is performed by searching a product by its "ID," "Category," "Title," and "Price Range." Such queries are reverted with appropriate "Purchase," "Description," "Offers," and "Availability" information for 160 users. Similarly, the services are held for 12–20 mins for a user. From this detailing, the metrics of data availability, mining time, access time, failure rate, and sharing ratio are compared with the existing UIMS [21], FCNN-DM [24], and DMFM [16] methods.

2.4. Availability Comparison. The comparative analysis for availability is presented in Figure 8 for the varying services and users. The proposed method identifies $\mu_s(t)$ for $\beta_s(t)$ improvements. This improvement is analyzed$\forall s$ in the preallocation andϕ for the tolerable level verification. Based on these assessments, the federated learning validatesψ and$M$ individually. The common outputs are merged across different $\sum_{\beta_s}(t)$ such that availability is maximized. In particular, the availability is maximized using $\chi_s(t)$ between two successive intervals. In the repeated assessment, ${\gamma_{rs}}^{\max}$ is satisfied using $\partial(t)$ minimization. Therefore, the conventional request allocation and service assignments are maximized. In the mining process, the available resources are shared acrossψ satisfied intervals. Therefore, the user accessing intervals are maximized with $d_s(t)$ based on $\cos\theta$. This is carried forward for all$\chi_s(t)$ based on learning outputs. Hence, this proposed method maximizes data/resource availability.

2.5. Mining Time Comparison. The proposed CDMM achieves less mining time for the varying services and users, as presented in Figure 9. First, the influencing factors for s for s_{nw} ands_{que} is estimated. Based on the delay estimation, $w_r(t)$is assigned using$\mu_s(t)$ maximization. In the consecutive allocations, $\chi_s(t)$- and$M$-based federated learning influences the delay causing factors such that$\sum_{\beta_s}(t)$ is reduced. In the available allocation intervals, $\chi_o(t)$ is the deciding factor for preventing increasing mining time for multiple resources. If the service and user concentration increase, then the ψ factor as in(8), (9), and (10) is assessed for different ϕ conditions. These conditions are based on the time factor for preventing additional delay, and therefore, the allocation consecutively aids existence. This is unanimously pursued for $d_s(t)$ and $\beta_s(t)$ such thatψ is improved by reducing delay. Contrarily, for the varying users, $\mu_s(t)$ is varied such that all $w_r(t)$ is allocated from the available resources. Therefore, the wait time, that is, s_{que}, is reduced, preventing additional mining time.

2.6. Access Time Comparison. The access time for the proposed method's varying users and services is less than the other methods (refer to Figure 10). The queuing and mining time in the proposed method is reduced by assigning $\beta_s(t)$ based on $\mu_s(t)$. This is required to improve the $w_r(t)$ allocation and processing rate. Based on the allocation capacity and accessing intervals, the availability is maximized. First, the s_{nw} is reduced by mining concurrent resources across varying $\sum_{\beta_s}(t)$ such that ${\gamma_{rs}}^{\max}$ is achieved. Depending on$\chi_s(t)\forall s_{d_m}$ and(r,s) the further access grant is provided. In particular, $\cos\theta$ using the federated learning is improved for$d_s(t)$ such that $(\beta)s(t)$ is increased. This is pursued to improve the existence, wherein $\partial(t)$ is reduced. However, in the varying user concentration, $d_s(t)$ varies across multiple $\chi_s(t)$ preventing the balance in (r,s). Therefore, s_{nw} is also reduced balancing $(\zeta,\phi)\forall$interval$\in(1,n)$. The successful $d_s(t)$ is increased for achieving less access time for any service $\forall$ users in the same interval.

2.7. Failure Rate Comparison. The resource allocation failure in the proposed method is less than in other methods. Following the varying services, $w_r(t)$is maximized by

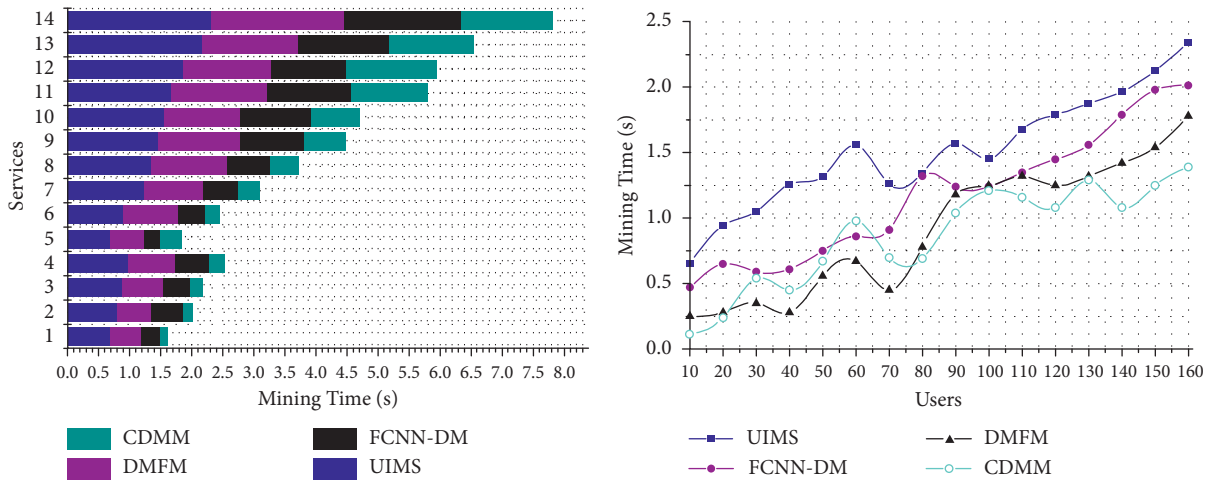

Figure 9: Mining time comparison.

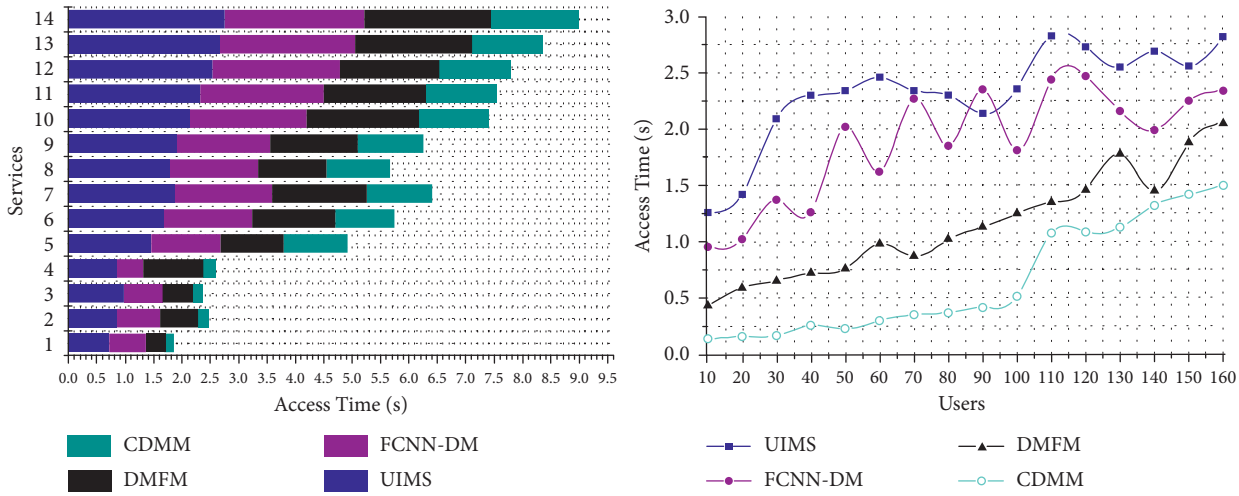

Figure 10: Access time comparison.

reducings in the pre-$\chi_{s(t)}$ assessment. In the learning-based validation, ψ assessment achieves fairϕ across the varying users. This is confined between 1 to n intervals such that $\cos\theta$ is the same. If the similarity index is high, then $\partial(t)$-based allocations are performed. Therefore, the maximum requests are assigned with a resource in the interval$n$. For the varying services, $\mu_s(t)$is varied for admitting $\omega_r(t)$ in the continuous intervals. The proposed method identifies ϕ in all the assigned$n$ such that $d_s(t)$ is maximized. This is required for maximizing $\cos\theta$, wherein the learning operates independently. From the M and ψ designed by the learning model, further allocations are performed, preventing $d_s(t)$ reduction. This is required for s mitigation and ϕ balancing between r and s. In the learning output assessment, $\partial(t)$-based allocations are prevented from interfering the $\chi_s(t)$ a decision such that existence is updated. Therefore, the sharing (shared) resource augments the demand suppression and reduces failures (refer to Figure 11).

2.8. Sharing Ratio Comparison. The proposed method achieves a fair sharing ratio compared to the other methods, as presented in Figure 12. The sharing is enabled by reducing the delay in queuing, access, and mining, as discussed earlier. The $\chi_s(t)$-varying users and services are streamlined using the deviating delay and mining time to prevent additional failures. The $\max[K_s(t) - \rho_s(t)\mu_s(t)\sigma, 0]$ process is responsible for performing the allocation across different tolerance factors. In the consecutive resource allocation,$\psi_{\max}$ is the estimating factor for maximizing the sharing ratio. The data management is performed for the above factor and M independently to maximize the mining process. In this process, the learning for similar features is streamlined to achieve a high repository allocation level. The process is prevented from avoiding requesting fewer allocations in the consecutive repository mining process. The collaborative allocations are performed for varying services such that$\mu_s(t)$ is maximized. In this process, the cost suppression is maintained such that the delay is also confined. The learning process further augments the data management system for improving the availability and retaining its existence until the interval $\in (0,1)$. Therefore, the repository is available for varying users and requests to improve the sharing ratio. This is not

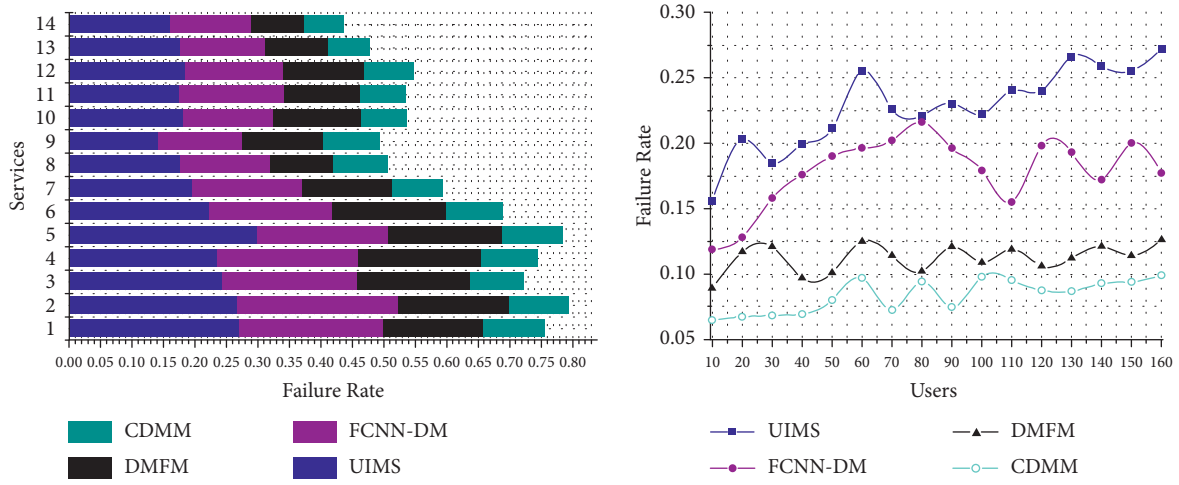

Figure 11: Failure rate comparison.

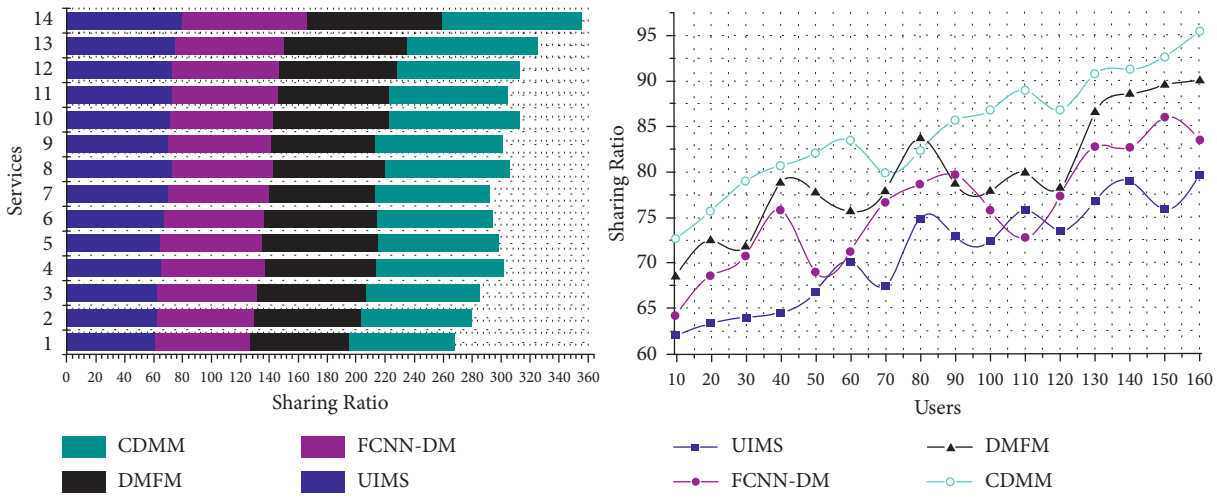

Figure 12: Sharing ratio comparison.

Table 1: Analysis summary for services.

Metrics	UIMS	FCNN-DM	DMFM	CDMM
Availability	0.66	0.76	0.84	0.942
Mining time (s)	2.328	2.15	1.86	1.48
Access time (s)	2.765	2.456	2.23	1.523
Failure rate	0.161	0.1296	0.0823	0.0634
Sharing ratio	79.45	86.66	92.44	96.34

Inference: The proposed method maximizes the availability and sharing ratio by 9.43% and 10.16%, respectively. It reduces the mining time, access time, and failure rate by 10.47%, 12.91%, and 6.09%, respectively.

Table 2: Analysis summary for users.

Metrics	UIMS	FCNN-DM	DMFM	CDMM
Availability	0.66	0.76	0.88	0.924
Mining time (s)	2.34	2.015	1.78	1.39
Access time (s)	2.82	2.34	2.05	1.498
Failure rate	0.272	0.177	0.126	0.099
Sharing ratio	79.67	83.45	90.03	95.46

Inference: The proposed method maximizes the availability and sharing ratio by 10.49% and 11.54%, respectively. It reduces the mining time, access time, and failure rate by 1.068%, 12.57%, and 9.27%, respectively.

repeated until the next allocation prevents additional access time. The above analysis is summarized for varying services and users in Tables 1 and 2 respectively.

3. Conclusion

This article introduced a connectivity-persistent data mining method for improving the sharing and allocation of a library of resource-based services. The proposed method relies on federated learning for validating data existence and availability for diverse user services. The mining process is performed for heterogeneous resources based on capacity-based allocation and delay mitigation. The user service demands are satisfied using experience and tolerance-based mining assimilations for improving resource availability. Besides, the available data are shared between the users and requests based on their existence. This existence is provided by maximizing request allocation and mining between connected users. The distinct service modes through existence and allocations are performed using the federated learning process through precise decisions from the data management system. Therefore, the proposed method maximizes existence and sharing regardless of the demands across the various intervals. The proposed method maximizes the availability and sharing ratio for the varying services by 9.43% and 10.16%, respectively. It reduces the mining time, access time, and failure rate by 10.47%, 12.91%, and 6.09%, respectively.

Data Availability

Data cannot be made available due to restrictions.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was supported by the research project of China Academic Library & Information System (CALIS) National Agricultural Literature Information Center in 2022 "*The Research about the Service Mode and Construction Path of the Scientific Research Date in Universities by the Era of Big Data.*" Project Number: 2022012.

References

[1] A. Zainab, A. Ghrayeb, D. Syed, H. Abu-Rub, S. S. Refaat, and O. Bouhali, "Big data management in smart grids: technologies and challenges," *IEEE Access*, vol. 9, Article ID 73046, 2021.

[2] Z. Xu, X. Zhou, A. Kogut, and J. Watts, "A scoping review: synthesizing evidence on data management instruction in academic libraries," *The Journal of Academic Librarianship*, vol. 48, no. 3, Article ID 102508, 2022.

[3] O. Zide and O. Jokonya, "Factors affecting the adoption of data management as a service (DMaaS) in small and medium enterprises (SMEs)," *Procedia Computer Science*, vol. 196, pp. 340–347, 2022.

[4] R. Tang and Z. Hu, "Providing research data management (RDM) services in libraries: preparedness, roles, challenges, and training for RDM practice," *Data and Information Management*, vol. 3, no. 2, pp. 84–101, 2019.

[5] J. Masinde, J. Chen, D. Wambiri, and A. Mumo, "Research librarians' experiences of research data management activities at an academic library in a developing country," *Data and Information Management*, vol. 5, no. 4, pp. 412–424, 2021.

[6] W. Didimo, L. Grilli, G. Liotta, L. Menconi, F. Montecchiani, and D. Pagliuca, "Combining network visualization and data mining for tax risk assessment," *IEEE Access*, vol. 8, Article ID 16073, 2020.

[7] M. Er Kara, S. Ü. Oktay Fırat, and A. Ghadge, "A data mining-based framework for supply chain risk management," *Computers & Industrial Engineering*, vol. 139, Article ID 105570, 2020.

[8] C. H. Lai and C. Y. Hsu, "Rating prediction based on combination of review mining and user preference analysis," *Information Systems*, vol. 99, Article ID 101742, 2021.

[9] S. A. Mohd Selamat, S. Prakoonwit, and W. Khan, "A review of data mining in knowledge management: applications/findings for transportation of small and medium enterprises," *SN Applied Sciences*, vol. 2, no. 5, pp. 818–915, 2020.

[10] J. C. Kim and K. Chung, "Mining based time-series sleeping pattern analysis for life big-data," *Wireless Personal Communications*, vol. 105, no. 2, pp. 475–489, 2019.

[11] Y. Zhong, L. Chen, C. Dan, and A. Rezaeipanah, "A systematic survey of data mining and big data analysis in internet of things," *The Journal of Supercomputing*, pp. 1–49, 2022.

[12] C. H. Lai, D. R. Liu, and K. S. Lien, "A hybrid of XGBoost and aspect-based review mining with attention neural network for user preference prediction," *International Journal of Machine Learning and Cybernetics*, vol. 12, no. 5, pp. 1203–1217, 2021.

[13] S. Hosseini and S. R. Sardo, "Data mining tools-a case study for network intrusion detection," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 4999–5019, 2021.

[14] M. Chen, W. Z. Li, L. Qian, S. L. Lu, and D. X. Chen, "Next POI recommendation based on location interest mining with recurrent neural networks," *Journal of Computer Science and Technology*, vol. 35, no. 3, pp. 603–616, 2020.

[15] M. Mahmud, M. S. Kaiser, T. M. McGinnity, and A. Hussain, "Deep learning in mining biological data," *Cognitive computation*, vol. 13, no. 1, pp. 1–33, 2021.

[16] P. Y. Huang, W. S. Cheng, J. C. Chen, W. Y. Chung, Y. L. Chen, and K. W. Lin, "A distributed method for fast mining frequent patterns from big data," *IEEE Access*, vol. 9, Article ID 135144, 2021.

[17] Y. Xie, P. Wen, W. Hou, and Y. Liu, "A knowledge image construction method for effective information filtering and mining from education big data," *IEEE Access*, vol. 9, Article ID 77341, 2021.

[18] J. Obregon, M. Song, and J. Y. Jung, "InfoFlow: mining information flow based on user community in social networking services," *IEEE Access*, vol. 7, Article ID 48024, 2019.

[19] P. Bhattacharya, F. Patel, S. Tanwar, N. Kumar, and R. Sharma, "MB-MaaS: mobile blockchain-based mining-as-a-service for IIoT environments," *Journal of Parallel and Distributed Computing*, vol. 168, pp. 1–16, 2022.

[20] Z. Zhang, "Mining method of massive data of mobile library under information asymmetry facing large-scare database," *Microprocessors and Microsystems*, vol. 81, Article ID 103730, 2021.

[21] S. Dhelim, N. Aung, and H. Ning, "Mining user interest based on personality-aware hybrid filtering in social networks," *Knowledge-Based Systems*, vol. 206, Article ID 106227, 2020.
[22] C. Wang, R. Huang, J. Li, and J. Chen, "Towards better information services: a framework for immigrant information needs and library services," *Library & Information Science Research*, vol. 42, no. 1, Article ID 101000, 2020.
[23] Y. Xiao, C. Li, M. Thürer, Y. Liu, and T. Qu, "User preference mining based on fine-grained sentiment analysis," *Journal of Retailing and Consumer Services*, vol. 68, Article ID 103013, 2022.
[24] W. Peng, "Big data mining and analysis based on convolutional fuzzy neural network," *Arabian Journal for Science and Engineering*, pp. 1–11, 2021.
[25] M. Alkathiri, A. Jhummarwala, and M. B. Potdar, "Multidimensional geospatial data mining in a distributed environment using MapReduce," *Journal of Big Data*, vol. 6, no. 1, pp. 82–34, 2019.
[26] L. Deng and D. Li, "Multimedia data stream information mining algorithm based on jointed neural network and soft clustering," *Multimedia Tools and Applications*, vol. 78, no. 4, pp. 4021–4044, 2019.
[27] C. Ju, J. Wang, and G. Zhou, "The commodity recommendation method for online shopping based on data mining," *Multimedia Tools and Applications*, vol. 78, no. 21, pp. 30097–30110, 2019.
[28] H. Zhou, G. Sun, S. Fu, J. Liu, X. Zhou, and J. Zhou, "A big data mining approach of PSO-based BP neural network for financial risk management with IoT," *IEEE Access*, vol. 7, Article ID 154035, 2019.
[29] Data World, "Promptcloud," https://data.world/promptcloud/flipkart-products-dataset.

5 参考文献

[1] 李伟绵. 基于生命周期理论的研究数据管理服务评估研究[D]. 北京理工大学, 2016.

[2] Management of Research Data and Records Policy (MPF1242) [EB/OL].
[2016-8-30]. http:// policy.unimelb.edu.au/MPF1242.

[3] NIH Grants Policy Statement[EB/OL]. [2016-8-31]. http://grants.nih.gov/grants/policy/nihgps_2011/nihgps_ch2.htm.

[4] 程娟, 张慧, 裴雷. 信息检索[M]. 天津: 天津大学出版社, 2014 年.

[5] 张瑶, 吕俊生. 国外科研数据管理与共享政策研究综述[J]. 图书馆理论与实践, 2015(11): 47-52.

[6] 关于政协十二届全国委员会第三次会议第 1071 号(科学技术类 054 号)提案答复的函[EB/OL]. [2016-8-31].
http://www.most.gov.cn/mostinfo/xinxifenlei/qgzxtafwgk/201603/t20160304_124467.htm.

[7] 丁培. 国外大学科研数据管理政策研究[J]. 图书馆论坛, 2014(5) :99-106.

[8] 孙继周. E-Science 环境下高校图书馆开展科学数据管理与共享的路径研究[J]. 图书馆, 2016(5): 66-71.

[9] Marchionini Gary, 杨冠灿, 芦昆. 科研数据管理:保障数据质量,促进 iS chools 新科学研究[J]. 图书情报知识. 2013(4): 4-9.

[10] 邓仲华, 李志芳. 科学研究范式的演化——大数据时代的科学研究第四范式[J]. 情报资料工作, 2013, 34(4):19-23.

[11] What is research data[EB/OL]. [2016-8-30].
http://www.ands.org.au/guides/what-is-research-data.

[12] 孙九林, 黄鼎成, 李晓波. 我国科技数据管理和共享服务的新进展[J]. 世界科技研究与发展, 2002, 24(5):15-19.

[13] 李晓辉. 图书馆科研数据管理与服务模式探讨[J]. 中国图书馆学报, 2011(5):46-52.

[14] 张凯勇. 数据密集型科学环境下的高校图书馆科学数据服务[J]. 图书馆学研究, 2014(3): 69-72+96.

[15] 陈传夫. 中国科学数据公共获取机制: 特点、障碍与优化的建议[J]. 中国软科学, 2004(2):8-13.

[16] 王学勤, Amy Stout, Howard Silver. 建立数据驱动的 e-Science 图书馆服务: 机遇和挑战[J]. 图书情报工作, 2011, 55(13):80-83.

[17] 张莉. 中国农业科学数据共享发展研究[M]. 北京: 中国农业科学技术出版社, 2007.

[18] Explanation of Terms[EB/OL]. [2016-8-30]. http://www.lib.cam.ac.uk/preservation/incremental/

glossary.html.
[19] 凌晓良, Belbin LEE, 张洁, 等. 澳大利亚南极科学数据管理综述[J]. 地球科学进展, 2007, 22(5):532-539.
[20] 韩涛. 科学数据与科学文献相关性研究——以生物信息学为例[J]. 图书情报知识, 2008(3):42-46.
[21] 徐菲, 王军, 曹均, 等. 康奈尔大学嵌入式科研数据管理服务探析[J]. 图书馆建设, 2015(12): 54-59.
[22] 梁彧文. 高校图书馆科研数据管理服务探析[J]. 农业图书情报学刊, 2015, 27(3):189-191.
[23] 刘霞, 饶艳. 高校图书馆科学数据管理与服务初探——武汉大学图书馆案例分析[J]. 图书情报工作. 2013, 57(6):33-38.
[24] 张秋彦. 高校科学数据监护研究[J]. 情报科学, 2013(5): 42-45.
[25] 高芹, 钟晓莉. 高校图书馆科学数据监护平台构建研究[J]. 图书馆学刊, 2015(8): 113-116.
[26] 张梦霞, 顾立平. 数据监管的政策研究综述[J]. 现代图书情报技术, 2016, 32(1):3-10.
[27] 许鑫, 刘甜, 于霜. Data One 项目及其对我国数据监管工作的启示[J]. 图书与情报, 2014(6): 109-116.
[28] 王超. 高校图书馆科研数据管护模型研究[D]. 辽宁师范大学, 2015.
[29] 介凤, 王娟. 科研数据策管对我国高校图书馆的挑战[J]. 新世纪图书馆, 2015(8): 22-25.
[30] 时婉璐, 任树怀. 数据策管: 图书馆服务的新创举[J]. 图书馆杂志, 2012(10): 24-27.
[31] 沈婷婷. 北美科学数据服务的实践与启示[J]. 现代情报, 2014, 34(12):145-147.
[32] 朱彩萍. 高校图书馆提供科学数据服务的途径与内容[J]. 图书与情报, 2014(3):97-99.
[33] 刘桂锋, 卢章平, 阮炼. 美国高校图书馆研究数据管理服务内容研究[J]. 图书馆论坛, 2015(8): 137-144.
[34] 王婉. 澳大利亚高校图书馆参与科研数据管理服务研究[J]. 图书馆论坛, 2014(3):130-136 +149.
[35] 胡昕. 高校图书馆科研数据管理研究[J]. 管理观察, 2014(30): 138-139.
[36] 穆向阳, 洪跃. 学科馆员在科研数据管理中的角色分析[J]. 新世纪图书馆, 2015(8):17-21.
[37] 杨鹤林. 康奈尔大学:让研究者共享科学数据[J]. 中国教育网络. 2014(4): 32-34.
[38] 王应密, 张乐平. 我国高校开展院校研究数据库建设的困境与对策[J]. 高等工程教育研究. 2012(6): 139-144.
[39] 钱鹏. 高校科学数据管理研究[D]. 南京大学, 2012.
[40] 谢江宁. 高校科研数据可视化关键技术研究[D]. 山东大学, 2014.
[41] 徐坤. 基于本体的科学数据监护平台研究[D]. 吉林大学, 2014.
[42] 谢艳秋. 高校科学数据共建共享实现机制研究[D]. 东南大学, 2015.
[43] 吕欣. 高校图书馆社会科学数据管理与服务研究[D]. 东北师范大学, 2015.