

大数据时代数字资源整合与聚合研究^{*}

陈晓美¹, 孙中秋¹, 王秀艳², 周珊珊¹

(1. 吉林大学管理学院, 长春 130022; 2. 中国建设银行股份有限公司松原分行, 松原 138000)

摘要: 以CNKI收录的有关数字资源整合与聚合的相关文献作为研究对象, 采用文献计量、社会网络分析和可视化方法, 借助Ucinet、Sati3.2等软件, 揭示大数据环境下数字资源整合和聚合领域研究文献的外部特征和内容特征。在此基础上, 比较国内数字资源整合与聚合研究现状, 并对数字资源聚合研究的深入开展提出建议。

关键词: 大数据; 数字资源聚合; 数字资源整合; 文献计量; 社会网络分析

中图分类号: G253

DOI: 10.3772/j.issn.1673—2286.2014.06.005

数字资源整合是依据不同方式、模式、策略把分布、异构、自治的数字资源重构为新的有机体的过程^[1]。数字资源聚合是利用多种技术, 从不同维度, 挖掘数字资源之间的关联关系, 从而实现数字资源语义重组的过程, 是对数字资源整合的深化与延展。

国内外学者较为关注数字资源整合研究, 召开多次会议, 如“网络信息整合国际研讨会”、“IEEE信息再利用与整合国际大会”、“普通高等院校图书馆数字资源整合研究会议”、“数字化环境下高校图书馆网络资源的收集、整合与共享研究会”等。数字资源聚合是在数字资源整合研究基础上提出的新的理念和研究方法, 一经提出, 受到广泛关注。相关会议主要有英国JISC会议(发布《数字资源的聚合和维护报告》)、全国数字文献资源语义化、深度聚合与可视化研究专题研讨会, 研究成果颇丰。

大数据时代, 数字资源呈现出数量大、类型多、结构复杂等特征, 利用整合与聚合方法、技术、工具实现数字资源关联共享具有重要的理论和实践价值。本文采用文献计量、社会网络分析和可视化方法, 分析总结大数据环境下数字资源整合和聚合研究呈现的外部特征(作者、机构、期刊等)和内容特征(研究方法、模型、技术、工具等), 在此基础上, 对数字资源整合和聚合研究现状进行比较分析, 并对数字资源聚合研究的深入开展提出若干建议。

1 数据来源和研究方法

1.1 数据来源

本文以CNKI中国学术期刊网络出版总库为数据来源, 数字资源整合研究文献检索表达式为“KY=“数字资源整合” OR KY=(“数据库”+“电子期刊”+“电子图书”+“多媒体”) AND KY=“整合””, 数字资源聚合研究文献检索表达式为“SU=(“数字资源聚合”+“语义聚合”+“语义互联”) OR KY=“关联数据””, 检索时间、学科领域不限, 分别得到970篇和412篇文献, 由于会议和报纸文献在一定程度上也反映出研究领域的主要内容, 不删除。将检索到的文献以Endnote格式导出。

1.2 研究方法和工具

本文利用文献计量、社会网络分析和可视化方法, 对国内资源整合和聚合研究论文的外部特征和内容特征进行计量和社会网络分析, 并借助Ucinet、Sati3.2等软件, 绘制高频关键词和高产著者共现知识图谱, 从而揭示大数据环境下数字资源整合与聚合研究领域的核心著者、合著团体以及研究热点等。

文献计量法是情报学特有的研究方法, 从定性和定量两个角度揭示文献的外部和内容特征。社会网络

^{*} 本研究得到国家社会科学基金重大项目“基于语义的馆藏资源深度聚合与可视化研究”(编号: 11&ZD152)和国家自然科学基金项目“语义网络环境下数字图书馆资源多维度聚合与可视化展示研究”(编号: 71273111)资助。

分析法是对社会网络中行动者之间的关系进行量化研究的方法,借助社会网络分析软件,揭示资源网络中的关联关系并进行结构划分。

Ucinet是目前较为流行的社会化网络分析软件,数据分析能力强,可视化效果好,软件内嵌入NetDraw和Pajek软件,可以清晰地揭示文献资源网络中节点之间的语义关系。Sati是文献题录信息统计分析工具,通过字段抽取、词频统计和构建矩阵等挖掘文献之间的关联关系^[2]。

2 文献的外部特征

2.1 发文量和时间分布特征

通过对发文量和发文时间进行计量分析有助于掌握研究领域的发展历程和趋势,某一时间段内发文量的多少能反映此时间段内研究领域受关注程度的高低。图1是国内数字资源整合和聚合研究发文量和时间分布折线图。



图1 数字资源整合和聚合研究发文量和时间分布折线图

由图1可知,国内数字资源整合研究源于2001年,毛翠萍探讨了企业电子商务信息和企业内部信息整合,开创了国内信息资源整合研究的先河^[3]。2002年至2011年间国内数字资源整合研究领域的发文量一直处于稳步增长的态势,表明数字资源整合研究持续受到学者的关注,2012年至今,数字资源整合研究发文量有所减少,表明学者对数字资源整合研究关注度有所降低。

数字资源聚合研究源于2004年,《图书馆亘古的挑战:社会知识资源聚合控制核心动力发展论》指出知识资源社会化聚合是图书馆本质形态,实现社会知识资源聚合是图书馆发展的内驱力^[4]。2004年至今,数字资源聚合研究热度总体呈上升趋势,需要指出2007与2012年,数字资源聚合实现两次研究飞跃,发文量大幅度增加,笔者分析,第一次飞跃是由于语义网格技术的发展推进数字资源聚合研究,第二次飞跃是由于国

家立项支持(2012年国家自然科学基金信息资源管理类基金项目共27项,其中以聚合为主题索引的基金项目为3项),吸引学者关注数字资源聚合研究。

2.2 作者分布特征

2.2.1 高产作者统计

通过对高产作者及其发文量进行统计,有助于发现数字资源整合和聚合领域有影响力著者。将导出的Endnote文件导入Sati3.2,生成.xml格式文件,提取作者字段,进行频次统计,得到数字资源整合研究作者537位,数字资源聚合研究作者310位(均包含合著者),由于篇幅有限,本文仅展示发文量排在前10的作者,如表1所示。

表1 数字资源整合与聚合研究作者统计表

数字资源整合		数字资源聚合	
作者	发文量	作者	发文量
马文峰	8	眸冬梅	17
李忠强	4	毕强	11
夏明春	5	夏翠娟	9
黄晓斌	4	刘炜	9
郑燃	4	王丽伟	7
刘涛	4	张晓林	6
耿骞	3	沈志宏	6
郑燃	3	韩毅	6
杜小勇	3	李春旺	6
刘清	3	李楠	5
肖希明	3	张宁	5
刘子辉	3	范炜	5
白海燕	3	祝忠明	5

由表1可知,数字资源整合研究的高产作者为马文峰、李忠强、夏明春、黄晓斌等人,主要来自中国人民大学、贵州省图书馆、武汉大学。数字资源聚合研究高产作者为眸冬梅、毕强、夏翠娟、刘炜等人,主要来自吉林大学、上海图书馆、中国科学院国家科学图书馆和武汉大学。由此可知,高等院校和图书馆是数字资源整合与聚合研究的主要科研力量。

从高产作者及其发文量的统计来看,数字资源整合

研究领域发文数量为1篇的作者数量为1139位,发文量在2篇到3篇之间的作者数量为61位,发文量3篇以上的作者数量为69位。数字资源聚合研究领域发文量为1篇的作者463位,发文量在2篇到4篇之间的作者数量为68位,发文量4篇以上的作者23位,高产作者相对较多。

从合著论文统计来看,数字资源整合研究领域发表的970篇文献中,合著论文227篇,占论文总数的23.4%,合著率较低。数字资源聚合研究发表的412篇文献中,合作文献为120篇,占论文总数的29%,合著率也较低。

2.2.2 高产作者共现

高产作者在研究领域的科研产量和学术影响力较高,通过绘制高产作者共现图谱有助于捕捉数字资源整合研究领域有影响力的团体和学者们较为关注的研究内容。根据作者发文量,利用Sati3.2抽取作者字段,进行词频统计,生成高产作者共现矩阵,转换数据格式(##d),利用NetDraw生成高产作者共现图谱,如图2所示。

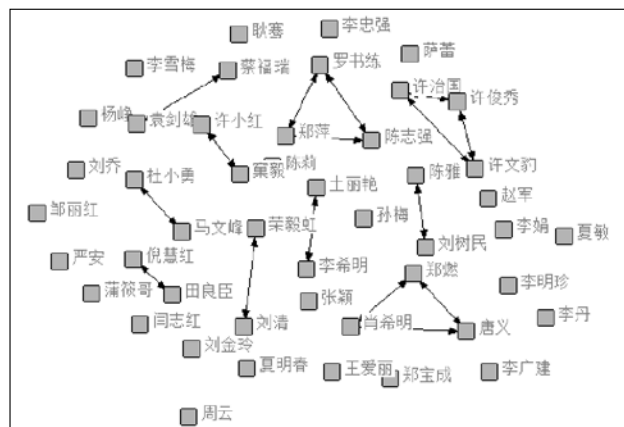


图2 数字资源整合研究高产作者共现图谱

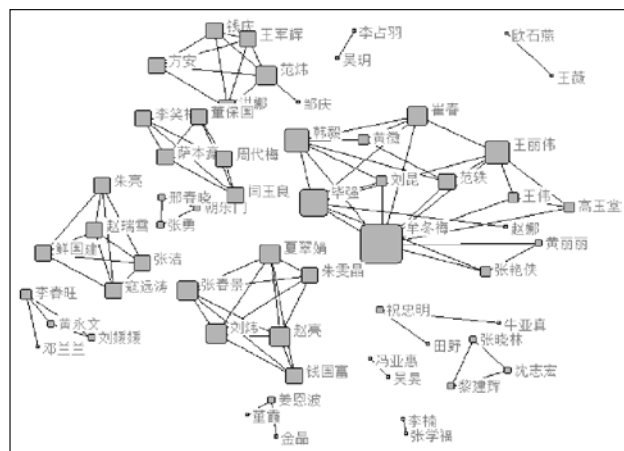


图3 数字资源聚合研究高产作者共现图谱

由图2和图3可知,图中不同节点代表不同作者,联线的粗细代表作者间合作次数多少,合作次数越多,则连线越粗。

数字资源整合研究领域形成以罗书练、郑然、许治国为核心的三大合作团体,较为关注医学、图书馆、档案馆和博物馆等的数字资源整合^[5]。但高产作者间的合作现象较少,且没有形成较大的合作群体,尤其是马文峰、王爱丽、夏明春、黄晓斌、郑燃等高产作者很少与其他高产作者合作,说明数字资源整合研究作者间的合作强度有待加强。

数字资源聚合领域形成五大合著团体,较有影响力的代表人物主要有牟冬梅、毕强、夏翠娟、刘炜等人,高产作者间合作紧密,合著群体较大,主要关注语义聚合的模式、策略、因素、方法,大数据与关联数据技术、模型等研究。高产作者间交流较少,合作水平也有待提高。

2.3 期刊分布特征

通过对期刊及其载文量进行统计,有助于寻找数字资源整合和聚合领域的核心期刊,可为后续研究工作提供高质量的文献资源。利用Sati3.2软件,抽取期刊来源字段,经过词频统计,得到表2数字资源整合和聚合研究期刊分布表(部分)。

表2 数字资源整合和聚合研究期刊分布表

数字资源整合	数字资源聚合
现代情报	26 现代图书情报技术
科技情报开发与经济	17 图书情报工作
图书馆学研究	16 图书馆学研究
中国教育科技装备	16 中国图书馆学报
教育教学论坛	14 情报理论与实践
图书馆学刊	14 图书与情报
科技信息	13 图书馆杂志
情报探究	13 情报科学
中国科技创新导刊	13 图书馆理论与实践
甘肃教育	11 情报杂志
情报科学	11 新世纪图书馆
中国信息技术教育	10 图书馆论坛
图书馆论坛	9 图书馆学刊

3.1 数字资源整合研究热点内容

由图4可知,数字资源整合研究主要集中在以下五个方面:

(1) 数字资源整合模式

数字资源整合模式是指利用方法、手段或借助统一平台实现资源重组和共享,是数字资源整合研究的重要内容。学者指出数字资源整合模式主要有基于导航系统、链接系统、OPAC、跨库检索系统^[6]的数字资源整合,基于开源软件的数字资源整合^[7]、基于数据、系统和服务的数字资源整合^[8]等。此外,从系统、数据等方面进行数字资源整合模式的比较研究^[9]也是学者较为关注的研究热点。

(2) 数字资源整合方案与策略

数字资源整合方案指对数字资源整合过程中涉及的目标、原则等进行预先的设定和编排。研究者指出,数字资源整合的目标主要是实现访问与检索功能、资源集成、提供个性化和智能化服务^[10]等,资源整合的原则主要包括树立大资源观、强调数字图书资源的应用^[11]等。

数字资源整合策略的制定随着整合目标、方式等变化而变化,有学者指出从资源组织方式和知识管理的角度出发,数字资源整合策略可划分为基于主题、面向问题和基于知识地图的资源整合三种^[12]。此外,学者们认为主题词表、学科知识体系以及行业需求^[13]也是数字资源整合策略的分析维度。

(3) 数字资源整合影响因素

数字资源整合影响因素分析是开展数字资源整合研究的必要基础和前提。数字资源整合影响因素随着聚合目标、环境、方法等变化而变化,因此,对数字资源整合因素的分析要视情况而定。研究者主要从政治、经济和技术等视角^[14]分析数字资源整合的影响因素,也有学者从数字资源本身、环境、聚合团队等方面分析数字资源整合的影响因素。

(4) 数字资源整合标准与协议

数字资源整合标准与协议贯穿数字资源整合研究的全过程,标准和协议的制定是数字资源整合研究的重要任务。通过统一、完善、规范化的标准和协议使得数字资源整合成为可能。目前,数字资源整合研究过程中涉及的标准与协议主要有用于开放档案元数据获取的互操作协议OAI-PMH、用于元数据级的检索和数据传送的OpenURL、定义计算机之间进行信息查询和提取的服务协议Z39.50、用于在不同的安全域中传输身份验证和授

权凭证的SAML和用于不同系统间互操作的OAIS等。

(5) 数字资源整合服务

数字资源整合的最终目的是为用户或科研人员提供集成化的数字资源服务,促进数字资源共享。数字资源整合应用服务主要包括:①资源导航。数字资源导航主要是帮助用户全面了解数字资源现状,并提供相应的浏览和检索服务,数字资源整合模式、方法、策略的差异性导致资源导航的多样性;②个性化推荐。数字资源整合有利于发现用户或研究人员的信息需求,从而有针对性地提供个性化的数字信息资源,节约检索的时间和经济成本;③异构检索。创建统一的数字资源检索平台既是数字资源整合的方式,也是数字资源整合成果的重要应用形式,以“元数据”为核心的商业平台^[12]、数据库、检索平台^[15]等实现图书馆联盟成员馆间资源共享是异构检索的典型应用。

3.2 数字资源聚合研究热点内容

由图5可知,数字资源聚合研究主要集中在以下三个方面:

(1) 数字资源聚合模型

构建科学系统的聚合模型可为数字资源聚合研究提供理论指导,国内数字资源聚合模型研究主要集中在语义关联检索模型、互联应用模型、跨媒体资源检索模型以及馆藏资源聚合模型等方面。

(2) 数字资源聚合方法与技术

数字资源聚合是借助聚合技术和软件,挖掘数字资源之间语义关联关系的过程,聚合技术研究是数字资源聚合研究的重中之重。数字资源聚合方法主要有主题词表、本体、分众分类法、文献计量学和社会网络分析法等。数字资源聚合不是单一维度下的聚合,而是多视角下的资源聚合探索,因此,聚合方法的互补融合也是聚合研究的热点内容。数字资源聚合技术主要包括数据挖掘、云计算、本体构建、SFX和可视化技术等。随着数字资源语义聚合研究的深入,学者们还提出了概念格、关联数据分析、语义标注和语义聚合等技术。

(3) 数字资源聚合应用

数字资源聚合主要用于挖掘图书馆馆藏文献资源^[16]、数据库学术资源^[17]以及网络数字资源^[18]的语义关联关系,通过数据挖掘、语义标注和语义检索技术为用户提供语义导航服务。此外,信息聚合服务(RSS)在高校图书馆应用现状、影响因素^[19]的研究也是数字资

源聚合应用研究的重要分支。

4. 数字资源整合与聚合研究现状分析

4.1 数字资源整合与聚合研究比较分析

4.1.1 外部特征比较

(1) 发文时间与发文量。国内学者对数字资源整合研究开展较早,且发文量较多,对数字资源聚合研究开展较晚,发文量较少。从长远来看(尤其2012年以后),数字资源深度聚合必然成为国内学者探索的新目标。在借助数字资源整合研究已有成果基础上,从研究主题、研究方法、研究技术等方面开展数字资源聚合研究是大势所趋。

(2) 高产作者及合著水平。数字资源整合和聚合研究高产作者均较少,且发文量较少,表明数字资源整合与聚合的研究力度有待加强。数字资源聚合研究领域的高产作者的合著率略高于数字资源整合研究领域,表明数字资源聚合研究领域学者间交流互动机会较多,有助于数字资源聚合研究的深入开展。

(3) 期刊与机构。图书情报工作、情报科学、图书馆学研究等期刊,武汉大学、南京大学和华中师范大学等机构不仅在数字资源整合研究中载文量(发文量)较多,在数字资源聚合研究中贡献也较大,可见数字资源整合研究为聚合研究提供科研力量和科研阵地。此外,图书情报技术、情报理论与实践等期刊是数字资源聚合研究的核心期刊,表明数字资源聚合研究偏重技术和实践应用。吉林大学、上海大学等机构为数字资源聚合研究注入新力量。机构与期刊的合作,可为数字资源整合与聚合研究深入开展提供新契机。

4.1.2 内容特征比较

(1) 研究体系。数字资源整合研究内容体系较为完整,理论研究主要包括整合模式、方案、策略、服务和影响因素等,理论挖掘透彻,且较为全面,实践研究主要为整合标准、协议、方法、技术等,研究成果已被应用到众多领域,技术方法较为成熟;数字资源聚合研究内容体系不够丰满,理论研究方面较为关注模型、服务和应用研究,理论研究点较少,实践研究主要集中在聚合技术和方法研究,还未有学者从事聚合标准与协议的相关研究。

(2) 研究侧重点。数字资源整合研究较为注重理论探索,而数字资源聚合研究则更偏重方法与技术探究。整合研究中虽然有众多的技术和标准研究,但依靠单一方法和技术已经不能满足大数据时代信息处理的需要,而数字资源聚合研究则试图整合利用多种聚合方法和技术,实现资源的多维度聚合,如本体技术与分众分类法融合实现馆藏资源聚合等。显而易见,数字资源聚合研究中的方法和技术更能满足大数据环境下用户的信息需求。

4.2 数字资源聚合研究建设性意见

通过上述的比较分析可知,大数据时代,国内数字资源整合与聚合研究均取得较多研究成果,但数字资源聚合研究因起步较晚、技术突破难等原因,研究深度需加强,广度需延展。鉴于此,笔者对国内数字资源聚合研究提出以下建议:

(1) 借助数字资源整合研究成果。数字资源整合研究与聚合研究密切相关,整合研究的众多理论与方法也适合于聚合研究,如本体方法同时适用于两者。数字资源聚合研究应充分利用整合研究的已有成果、科研力量、应用领域等,并在其基础上创新、扩展和完善。

(2) 提高合著水平。大数据环境,海量信息处理考验的是团队合作能力,“合则共赢”已成为研究者的共识。数字资源聚合研究的合著率虽然略高于整合研究,但与其他较为成熟的学科比较,合著群体尤其是有影响力的合著团体较少,且规模较小,这可能制约研究人员的信息交流与科研创新。合著研究是短时间内提高数字资源聚合研究数量和质量的有效手段,既有利于资源共享,也有助于增强作者间的学术凝聚力,加强著者尤其是高产著者间的合作是数字资源聚合研究的重要任务。

(3) 坚持“标准先行”的原则。数字资源形式多样、结构复杂是大数据背景下信息资源的特点,缺乏科学统一的聚合标准严重制约数字资源聚合的效果。与数字资源整合研究比较发现,国内数字资源聚合标准研究较少,标准问题是数字资源聚合研究需要解决的首要问题,数字资源聚合研究可以参照数字资源整合研究的相关标准,并在其基础上进行细化,如聚合技术标准、聚合应用标准等。聚合标准的制定需要坚持促进聚合技术的原则,结合聚合需求的具体情况而定。

(4) 融合聚合方法和技术。大数据时代,信息资源的新特点、用户需求的个性化均要求聚合方法和技术

的推陈出新,而聚合方法和技术的互补融合研究不仅能弥补单一聚合方法或技术的局限性,而且能扩展聚合维度、优化聚合效果。数字资源聚合研究领域学者已开始从事聚合方法互补融合研究,但随着信息环境和用户需求的变化,聚合方法和技术融合研究仍是数字资源聚合研究中的难点和突破点。

5 结语

继承与发展是学科得以延续和兴旺的关键。数字资源聚合作为新兴研究领域,起步较晚,需借助数字资源整合研究的成果,避免数字资源整合研究中出现的问题。同时,也要满足大数据环境下数字资源和用户需求,不断完善和细化聚合理论体系,深化研究内容,扩展研究维度和应用领域,当然,这些研究的开展需要相关聚合技术和标准作为支撑。

参考文献

- [1] 马文峰.数字资源整合研究[J].中国图书馆学报,2002,28(4):64-67.
- [2] 文献题录信息统计分析工具SATI [EB/OL] [2014-04-13]. <http://sati.liuqiyan.com/>.
- [3] 毛玉萃.试论企业电子商务与企业内信息系统的整合[J].管理信息系统,2001(5):43-45.
- [4] 郭正武.图书馆亘古的挑战:社会知识资源聚合控制核心动力发展论[C]//福建省图书馆学会2004年学术年会论文集,2004:1-6.
- [5] 罗书练,郑萍,陈志强.加强数字资源的整合 提供良好信息服务[J].医学信息,2005(8):017.
- [6] 谢宝义.高校图书馆数字资源整合模式研究与实践:以石家庄铁道大学图书馆为例[J].图书馆建设,2011(2):13.
- [7] 陈和,王爽.基于开源软件实现馆藏数字资源整合与统一检索[J].现代图书情报技术,2009(6):70-75.
- [8] 李和娟,惠青.混合型图书馆的数字资源整合[J].情报探索,2009(8):56-57.
- [9] 张巧娜.海峡两岸高校图书馆数字资源整合的比较研究[J].图书情报工作,2012,56(19):43-47.
- [10] 李红霞.图书馆馆藏数字资源整合方案研究[J].现代情报,2006(8):144-145.
- [11] 赵荣.浅谈图书馆服务工作如何应对数字化阅读的挑战[J].图书馆工作与研究,2011(7):99-101.
- [12] 邓三鸿,金莹,秦嘉杭.基于知识管理的数字资源整合策略[J].情报科学,2006,24(10):1489-1493.
- [13] 邱琳,郑怀国,孙素芬,等.基于知识管理的专业图书馆数字资源整合策略[J].现代情报,2007,27(8):90-93.
- [14] 袁村平.图书馆与出版企业数字资源共享影响因素分析[D].湘潭:湘潭大学,2013.
- [16] 王倩,刘金玲.基于商业元数据整合的区域图书馆联盟数字资源揭示系统研究[J].图书馆学研究,2013(11):23.
- [15] 郭瑞华.图书馆信息资源整合及元数据应用[J].图书情报工作,2006,50(10):100-103.
- [16] 邱均平,王菲菲.基于共现与耦合的馆藏文献资源深度聚合研究探析[J].中国图书馆学报,2013(3):25-33.
- [17] 潘薇.“长尾理论”与图书馆学术信息资源聚合[J].图书情报工作,2010(S2):83-85.
- [18] 李爱明.数字图书馆中基于语义关联挖掘的跨媒体检索研究:模型设计与实验分析[J].情报科学,2014,32(1):85-88.
- [19] 刘军,金淑娜.基于RSS订阅中心的高校图书馆网络信息资源聚合研究[J].图书馆学研究,2013(19):60-63.

作者简介

陈晓美,女,1973年生,吉林大学管理学院副教授。

孙中秋,女,1990年生,吉林大学管理学院硕士研究生,通讯作者,E-mail: 466386531@qq.com。

The Research of Digital Resources Integration and Aggregation in The Area of Big Data

SUN ZhongQiu¹, CHEN XiaoMei¹, WANG XiuYan², ZHOU ShanShan¹

(1. School of Management, Jilin University, Changchun 130022, China; 2. Songyuan Branch, China Construction Bank, Songyuan 138000, China)

Abstract: With the relevant literatures of digital resources integration and aggregation in CNKI as the research objects, using the bibliometrics, social network analysis and visualization methods, revealing the external characteristics and contents characteristics of digital resources integration and aggregation in the area of big data. On this basis, we compare the research status of digital resources integration and aggregation, and give some suggestions to further research of digital resources aggregation.

Keywords: Big data; Digital resources integration; Digital resources aggregation; Bibliometrics; Social network analysis

(收稿日期: 2014-04-30)