

数字人文资源调查与发展对策探讨

董政娥 陈惠兰 (东华大学图书馆 上海 201620)

摘要 针对数字人文特点,文章调查了数字人文文献资源分布,分析了其发展布局、来源出版物、学科方向、地区、机构和作者关键词文本挖掘,并在聚类分析基础上剖析了其研究方向,认为数字人文环境下人文社会科学可以从文本挖掘、GIS技术、文本可视化、特色馆藏和语料库等方面进行跨界合作研究,以加速人文社会科学向多元化转向。

关键词 数字人文 人文社会科学 跨界合作

Discussion of Digital Humanistic Resources Study and Development Countermeasures

Dong Zheng'e Chen Huilan (Library of Donghua University, Shanghai, 201620)

Abstract Aim at the characteristics of the digital humanities, the distribution of literature resources on the digital humanities is investigated by the analysis of the layout, source titles, academic orientation, territories, institutional and research areas. The result shows that Humanities and Social Sciences can be tapped from the text, GIS technology, text visualization and other aspects of crossborder cooperation corpus research to accelerate shift to the diversification of Humanities and Social Sciences.

Keywords digital humanities, Humanities and Social Sciences, cross-border cooperation

大数据时代不仅冲击着传统人文研究的方式和维度,而且冲击着传统人文精神的视域和语境,还将为传统人文科学的研究带来新的研究方法和研究理论,最终使信息技术与人文研究融合形成一个新的研究领域——“数字人文”(Digital humanities)^[1]。“数字人文”,也称人文计算(Humanities Computing或Computing in the Humanities)^[1],其“一端是高效的计算,另一端是人文沟通”,数字人文研究就是将计算机技术、电子技术、网络技术等应用到文学、历史学、艺术学、考古学等传统人文学科中^[3]。作为一个典型的文理交叉领域,数字人文研究项目和研究团队既包括传统人文领域(哲学、历史学、文学、语言学、艺术学、人类学等)的研究者,也包括精通计算机技术和多媒体技术的专家学者。在他们的协作下,诸如数字仓储、文本挖掘^[4]、多媒体出版^[5]、数字图书馆^[6-7]、信息可视化^[8]、虚拟现实^[9]、地理信息系统等^[10]多种信息技术开始在人文领域得到深入应用。随着数字原生(born-digital)人文数据的增

加,现代人文学科正在快速地向可计算化方向发展,在此过程中,诸如网络民族志^[11]、网络文化、虚拟社会等主题逐渐成为新的人文研究课题。由此可见,数字人文赋予人文社科领域研究更加广阔的舞台,为此,本文调查了“数字人文”研究文献在SCI-E、SSCI和CNKI数据库的分布情况,以期促进大数据时代国内人文社会科学研究的认识 and 了解,最终达到抛砖引玉、吸引更多学者关注大数据时代数字人文背景下人文社会科学研究的的目的。

1 数据来源

本文选取Web of Science核心合集数据库平台上的SCI-E(Science Citation Index Expanded, 1985年至今)和SSCI(Social Sciences Citation Index, 1995年至今)数据库引文数据和CNKI来源数据,以“数字人文or人文计算”(Digital Humanities or Humanities Computing)为关键词,分别在“主题”、“标题”和“全文”字段调查分

本文系教育部人文社会科学规划项目“纺织史研究资源保障及其服务体系构建研究”(编号:13YJA870001)的成果之一。

析数字人文相关文献的分布情况。

2 数据采集及结果分析

调查分析中将 SCI-E 和 SSCI 数据库来源数据划分为国外文献,将 CNKI 来源数据划分为国内文献。数据采集表明,数字人文在 SCI-E 和 SSCI 数据库的相关文献数据,以标题为字段检索到 102 篇相关文献;以“主题”为字段检索到 365 篇相关文献。而数字人文以“全文”为字段在 CNKI 中国知网数据库检索到 50 篇相关文献。

2.1 发展布局分析

图 1 分析了“标题”字段下数字人文研究文献的年代分布情况,由图 1 可见,国外早在 1987 年就有“数字人文”相关的文献报道,但是在之后的 20 年间(截止到 2007 年)变化均不明显,从 2008 年开始,相关研究文献明显增多,从 2009 年开始持续增加,2012 年开始增长速度加快。相比之下,中国数字人文从 1999 年开始有相关文献报道,在随后的十几年中,波动较大且较少,直到 2012 年和 2013 年开始增加,到 2014 年达到高峰,尽管如此,其文献数与国际相比还有很大差距。

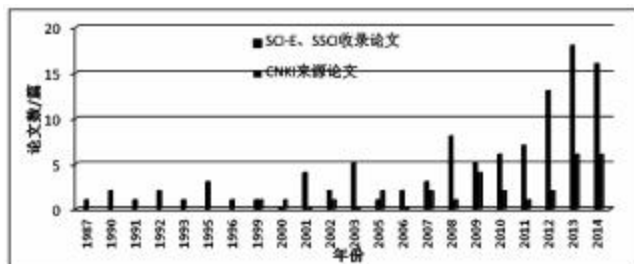


图 1 “数字人文”相关文献的年代分布

2.2 来源出版物分析

为了进一步分析“数字人文”相关文献分布情况,图 2 分析了数字人文相关外文文献来源出版物分布情况,由图 2 可见,自 1987 年以来,有关数字人文的 102 篇外文论文共分布在 48 种刊物上,其中 34 种刊物分别只收录 1 篇相关论文(如图 3(a)); 10 种刊物分别收录 2 篇论文(如图 3(b)); 收录 3 篇论文的只有 Library Quarterly, Historical Social Research Historische Sozialforschung 和 Differences a Journal of Feminist Cultural Studies 均收录了 8 篇相关论文; Computers and the Humanities 收录了 13 篇论文; Literary and Linguistic Computing 收录了 18 篇论文。

为了进一步了解数字人文中文文献分布情况,图 4 分析了数字人文相关中文文献来源出版物分布情况。由图 4 可见,数字人文有关中文文献共 35 种,

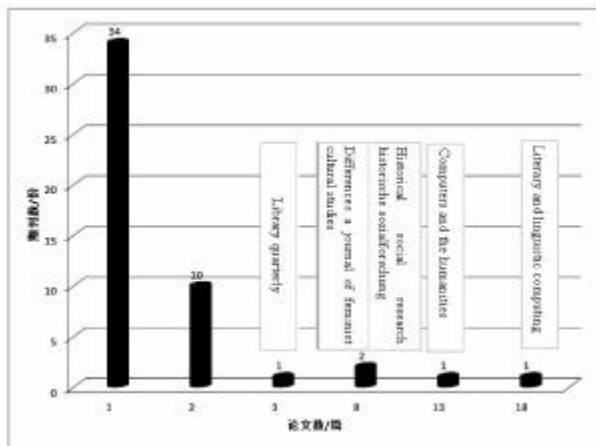


图 2 数字人文相关文献来源出版物分布

其分布可分三个区域:

第一区域:该区域只有《图书馆建设》发表了 3 篇相关论文。

第二区域:该区域共涉及 5 种期刊,均发表论文 2 篇,包括《现代图书馆情报技术》、《大学图书馆学报》、《国家图书馆学刊》、《图书情报工作》和《中国社会科学院报》。

第三区域:该区域共涉及 29 种期刊,均只发表论文 1 篇,包括《东方丛刊》、《南京社会科学》、《南通大学学报 社会科学版》和《地球信息科学学报》等。

2.3 学科方向分析

为了进一步分析目前数字人文研究文献涉及的研究方向,图 5 绘制分析了相关外文文献研究方向分布结构图,由图 5 可见,目前数字人文研究外文文献共涉及 27 个相关研究方向,根据其覆盖论文数,将其划分为三个梯队,具体如下:

第一梯队(收录论文在 10 篇以上):共涉及信息科学图书馆学、文学、语言学、计算机科学跨学科应用、计算机科学信息系统和历史 6 个研究方向,其中以语言学和信息科学图书馆学居首,分别收录 36 篇和 35 篇;其次是文学、计算机科学跨学科应用、计算

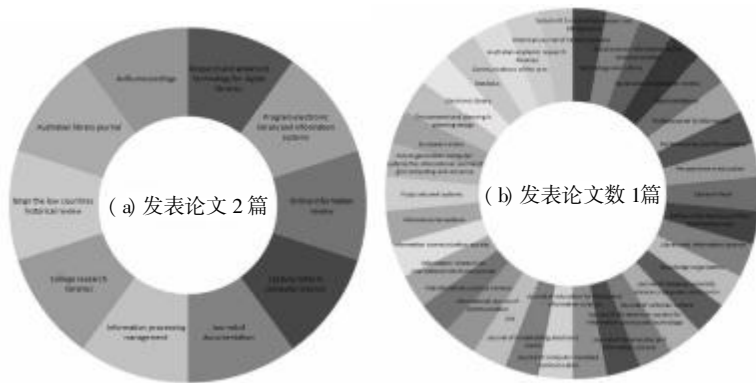


图 3 数字人文相关外文文献来源出版物分布(发表论文 1 篇和 2 篇)

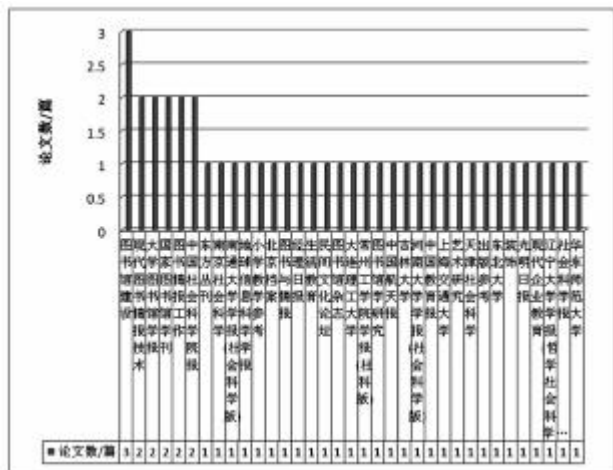


图4 数字人文相关中文文献来源出版物分布

机科学信息系统和历史。

第二梯队(收录论文在2~10篇):共涉及10个研究方向,社会科学跨学科和文化学居首,其次是女性年代研究和劳资关系,社会科学历史和计算机科学理论方法居第三位,紧随其后的是通讯,历史科学、计算

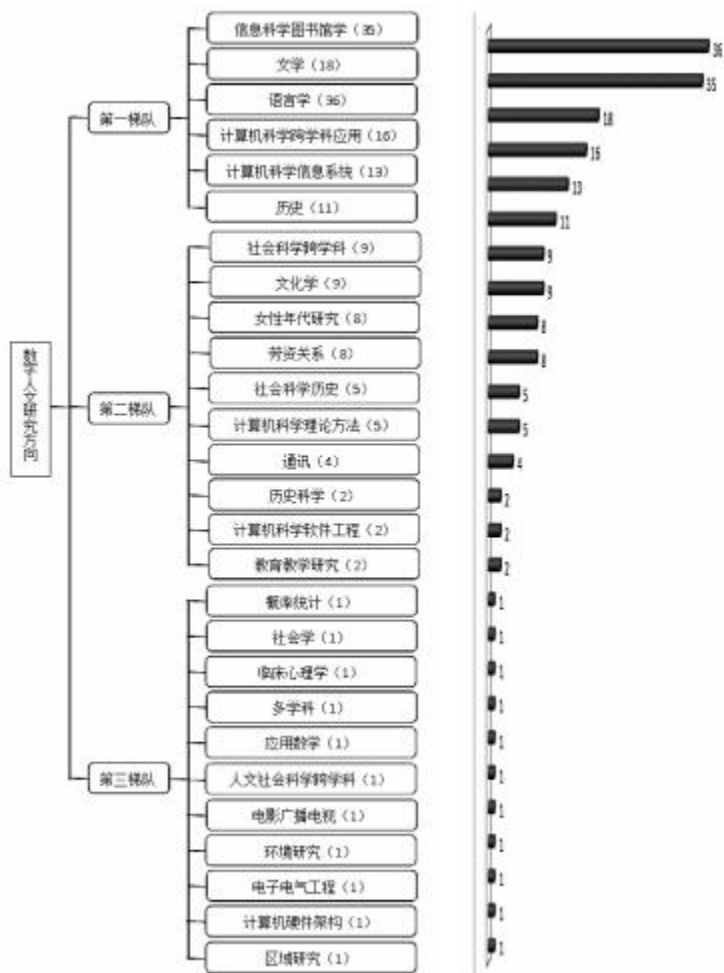


图5 数字人文相关文献的学科方向分布结构

机软件工程和教育教学研究则分别收录2篇论文。

第三梯队(收录论文1篇):共涉及11个研究方向,分别为概率统计、社会学、临床心理学、多学科、应用数学、人文社会科学跨学科、电影广播电视、环境研究、电子电气工程、计算机硬件架构和区域研究。

同样,图6分析了数字人文相关中文文献的学科方向分布情况,由图6可见,中文文献共涉及36个学

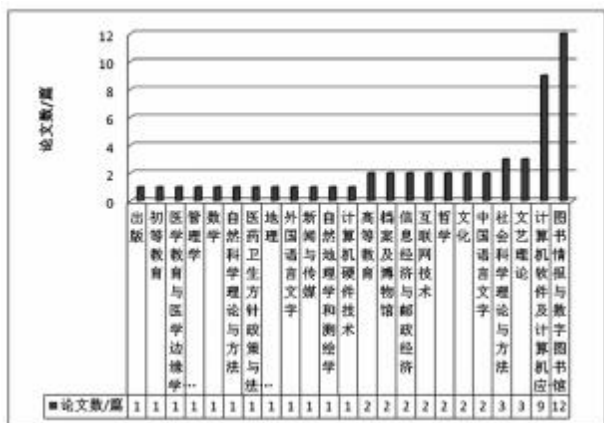


图6 数字人文相关中文文献的学科方向分布结构

科方向,可划分为4个区域:

第一区域:该区域图书情报与数字图书馆发表论文最多,为12篇,其次为计算机软件及计算机应用(9篇)。

第二区域:该区域中文艺理论和社会科学理论与方法分别发表论文3篇。

第三区域:该区域包括7个学科领域,分别涉及中国语言文学、文化、哲学、互联网技术、信息经济与邮政经济、档案及博物馆和高等教育。

第四区域:该区域均发表论文1篇,共涉及30个学科,具体涉及计算机硬件技术、自然地理学、测绘学和外国语言文学等。

2.4 地区分布分析

图7所示数字人文相关外文文献的国家和地区分布表明,目前数字人文相关外文文献共涉及22个国家和地区,其可划分为三个大区间,具体如下:

第Ⅰ区间(4~43篇论文):如图7a所示,该区域共涉及6个国家和地区,其中以美国代表,占总数的41%,其次是英格兰(占总数的18%),其他4个国家和地区所占份额较少,分别为加拿大、澳大利亚、荷兰和德国。

第Ⅱ区间(分别2篇论文):如图7b所示,该区域的威尔士、南非、苏格兰、墨西哥、日本、意大利和法国共7个国家和地区分别被收录2篇相关论文。

第Ⅲ区间(分别1篇论文):如图7c)所



图7 数字人文相关外文文献的国家和地区分布

示,该区域瑞典、西班牙、苏格兰、挪威、爱尔兰、芬兰、丹麦、博茨瓦纳、比利时共计9个国家和地区分别被收录1篇相关论文。

2.5 机构分布分析

数字人文相关外文文献共涉及114个机构,其发表相关论文数及分布如图8所示。由图8可见,UNIVERSITY OF LONDON(伦敦大学)发表论文最多,为17篇;其次是UNIVERSITY COLLEGE LONDON(英国伦敦大学学院)和KINGS COLLEGE LONDON(伦敦国王学院),分别为9篇和8篇;除此以外,UNIVERSITY SYSTEM OF MARYLAND(马里兰大学系统)、UNIVERSITY SYSTEM OF GEORGIA(佐治亚大学系统)、UNIVERSITY OF MARYLAND COLLEGE PARK(马里兰大学学院公园)、UNIVERSITY OF CALIFORNIA SYSTEM(加州大学系统)、CITY UNIVERSITY OF NEW YORK CUNY SYSTEM(纽约城市大学的城市大学系统)5个机构分别发表3篇论文;VASSAR COLI(瓦萨学院)和UNIVERSITY OF VICTORIA(维多利亚大学)等16个机构均发表2篇相关论文;而只发表1篇论文的机构共计90个。

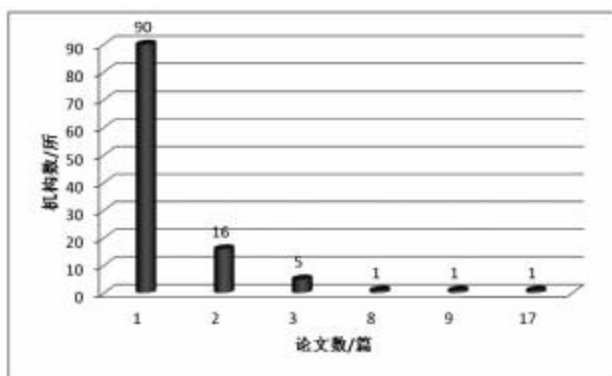


图8 数字人文相关外文文献涉及的机构分布

为了了解中文相关文献涉及机构分布情况,图9分析了数字人文相关中文文献涉及机构分布情况。由图9可见,数字人文相关中文文献涉及的机构可以划分为三个区域:

第一区域:以南京大学和武汉大学为代表的大学

分别在该领域发表论文4篇。

第二区域:以华东师范大学、华东理工大学、常州工学院和北京大学为代表的大学分别在该领域发表论文2篇。

第三区域:以盐城市毓龙路实验学校、大连理工大学、福建师范大学、上海交通大学、中国国家图书馆、中国科学院-香港中文大学地球信息科学联合实验室、中国科学院自动化研究所和中国科学院武汉文献情报中心为代表的相关机构均在该领域发表论文1篇。

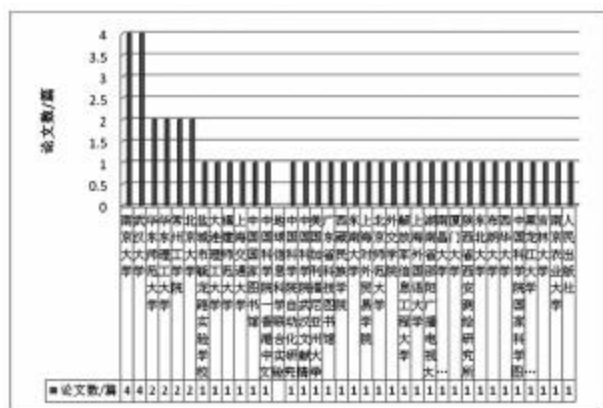


图9 数字人文相关中文文献涉及的机构分布

2.6 研究方向分析

为了进一步了解数字人文国内外相关文献涉略的研究方向,图10、图11分别对国外文献作者关键词进行了聚类分析。由图10可见,国外文献共涉及798个作者关键词,其中出现一次的关键词最多,大概占总数的88%;出现2次的占总数的7%;出现3次的占总数的2%。为了充分分析数字人文研究方向,根据作者关键词出现次数可将国外数字人文研究方向划分为以下四个区域:

第一区域(作者关键出现4次及以上):该区域共占总作者关键词的3%,图11(a)对出现频次最多作者关键词分布分析表明,除数字人文(Digital humanities)、人文(Humanities)和人文计算(Humanities computing)出现最多外,数字图书馆(Digital libraries)、科技

文艺评论(Literary criticism)、方法论(Methodology)、开放存取(Open access)、研究基础设施(Research infrastructures)、可持续发展(Sustainability)、文本挖掘(Text mining)、用户研究(User studies)、网络信息计量学(Webometrics) 和万维网(World Wide Web) 受到关注。

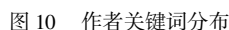
(1) 数字化: 数字计算 (Computation)、数字档案 (Digital archives)、数字收藏 (Digital collections)、数字通讯系统 (Digital communication systems)、数字哲学 (Digital philosophy)、数字保存 (Digital preservation)、数字技术 (Digital technologies)、关联数据 (Linked data)、元数据 (Metadata)、云计算 (Cloud computing)、收藏品管理 (Collections management)。

(3) 文本及其分析: 文本编码(Text encoding)、文本分析(Textual analysis)、社交网络(Twitter)、虚拟现实(Virtual reality)、引文分析(Citation analysis)、通讯(Communications)。

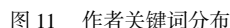
(4) 其他: 基础设施 (Infrastructure)、跨学科 (Interdisciplinarity)、词典编纂 (Lexicography)、图书馆信息学 (Library and information science)、文献研究 (Literary studies)、文献理论 (Literary theory)、地图 (Map)、移动设备 (Mobile devices)、建模 (Modeling)、博物馆 (Museums)、本体论 (Ontology)、绩效 (Performance)、政治经济学 (Political economy)、后人类 (Posthuman)、公众史学 (Public history)、研究方法 (Research methods)、研究工作 (Research work)、学术交流 (Scholarly communication)、标准置标语言 (SGML)、社会科学 (Social sciences)、教学与学习策略 (Teaching/ learning strategies)、文化研究 (Cultural studies)。

第四区域(作者关键词出现1次): 该区域作者关键词聚类分析表明,在数字人文背景下,以下主题受到关注:

(1) 补充计量学(Altmetrics)、基于图像的人文计算(Image-based humanities computing)、计算机科学哲学(Philosophy of



第二区域(作者关键词出现 3 次):该区域作者关键词分布(图 11(b))分析表明,基于数字人文、大数据(Big data)、通讯技术(Communication technologies)、数字媒介(Digital media)、历史(History)、互联网(Internet)、



computer science)、计算哲学(Philosophy of computing)、信息哲学(Philosophy of information)、科学哲学(Philosophy of science)、可持续发展哲学(Philosophy of sustainable development)。

(2) 特色馆藏(Special collections): 古书(Ancient book)、考古信息学(Archaeology informatics)、建筑史(Architectural history)、艺术史(Art history)、文物(Artifacts)、文物遗产(Artistic heritage)、小说(Baltasar Llopis Minguez)、陶瓷科技(Ceramics)、以收藏为主的研究如收藏数据库(Collection access)、收藏建设(Collection building)和收藏(Collections),并延伸出计算生物学(Computational biology)、计算新闻学(Computational journalism)、计算社会科学(Computational social science)、网络考古学(Cyber-archaeology)、文化传播(Cultural dissemination)、文化遗产(Cultural heritage)、文化遗产文本(Cultural heritage texts)、文化理论(Cultural theory)、伦理学(Ethics)、民族志(Ethnography)、图查询语言(Graph query language)、图表搜索(Graph search)、网格(Grid)及其计算、遗产(Heritage)、文物收藏(Heritage collections)、历史地理学(Historical geography)、历史学(Historiography)、数字人文历史(History of Digital Humanities)、教育史(History of education)、地貌历史(History of geomorphology)、泌尿外科史(History of urology)、城市历史(Urban Pond Procession)、非物质文化遗产(Intangible heritage)、新闻业(Journalism)、新闻历史(Journalism history)、山水画(Landscape)、文献地图学(Literary cartography)、音频打印(Musical printing)、音频分享(Musical analysis)、音频数据(Musical data)、音频数据库(Musical databases)、音频信息(Musical information)、音乐学(Musicology)和开放档案信息系统(Open Archival Information System,简称OAIS)。

(3) 语料库(Linguistic data bases): 语料库语言学(Corpus linguistics)、语言救助(Linguistic salvage)。

(4) 数据: 数据和知识传播(Data and knowledge transmission)、数据收集(Data collection)、数据保管(Data curation)、数据处理(Data handling)、数据管理(Data management)、数据挖掘(Data mining)、数据质量(Data quality)、数据登记(Data registries)、数据可视化(Data visualization)、数据仓储(Data warehousing)、数据密集研究(Data-intensive research)和数据库管理(Database management)。

(5) 数字化: 数字考古(Digital archaeology)、数字共享(Digital commons)、数字内容(Digital contents)、数字保存(Digital curation)、数字设备(Digital devices)、数字档案(Digital documents)、数字生态系统(Digital ecosystems)、数字环境(Digital environment)、数字民族志(Digital ethnography)、数字人文遗产(Digital humanities and heritage)、数字人文中心(Digital humanities cen-

ters)、数字人文实践(Digital humanities practices)、数字信息环境(Digital Informational Environment)、数字文化(Digital literacy)、数字地图(Digital maps)、数字方法(Digital methods)、数字对象保存(Digital object preservation)、数字对象(Digital objects)、数字语言学(Digital philology)、数字平台(Digital platforms)、数字出版(Digital publishing)、数字研究(Digital research)、数字学术(Digital scholarship)、数字仓储(Digital storage)、数字转化(Digital transformations)、数字视频转化(Digital video files)、数字化学术资源(Digitised scholarly resources)、纪录片(Documentary filmmaking)、记录语言学(Documentary linguistics)、数字学习(E-learning)、数字研究(E-resources)、数字社会科学(E-social science)、濒危语言(Endangered languages)、数字科学(E-science)。

(6) 知识共享(Knowledge as commons)、知识社区(Knowledge communities)、文学评论(Literary analysis)、文化批判(Cultural criticism)之外,情景感知(Context aware)、情景化(Contextualization)、动觉体验(Kinaesthetic experience)、认知(Knowing)、位置感知(Location-aware future)、位置感知学习(Location-based learning)、位置感知服务(Location-based services)受到关注。

(7) 文本及其分析备受关注: 诸如文本分析(Text analysis)、文本编码和呈现(Text encoding and rendering)、文本检索(Text retrieval)、文本/图像结合(Text/image coupling)、文本目录学(Textual bibliography)、文献学(Textual criticism)、文本研究(Textual studies)和文本变异(Textual variation)、量化文本分析(Quantitative text analysis)。

(8) 可视化是研究的另一个重点: 如视频游戏(Videogames)、虚拟重建(Virtual reconstruction)、虚拟研究环境(Virtual research environments)、虚拟工作台(Virtual workbench)、虚拟世界(Virtual worlds)、虚拟未来(Visual culture)、可视化查询语言(Visual query language)、可视化社会学(Visual sociology)可视化方法(Visualisation methods)。

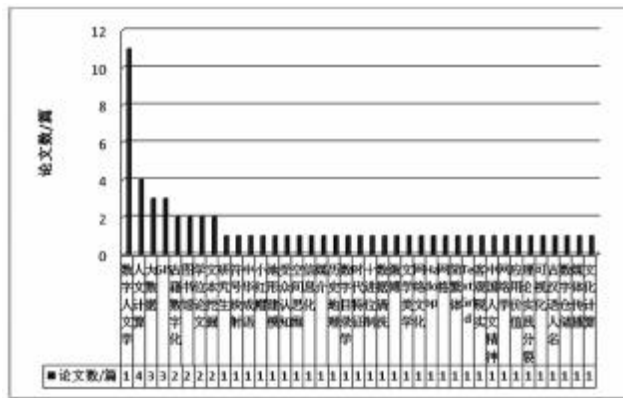
为了进一步了解国内人文数字文献涉及的研究方向,图12分析了国内文献作者关键词的分布情况。由图12可见,和国外研究相比,中文文献共涉及39个关键词,其中可划分为如下五个区域:

第一区域: 以“数字人文学”为代表的关键词,共出现11次。

第二区域: 以“人文计算”为代表的关键词共出现4次。

第三区域: 以“大数据”和“GIS”为代表的关键词分别出现3次。

第四区域: 以“古籍数字化”、“图书馆”、“学位论文”和“文本挖掘”为代表的关键词均出现2次。



第五区域:以“符号映射”、“受众认知”、“空间思维”、“信息化”、“数据清洗”、“数字目录学”、“数字仓储”、“文化计算”、“Hadoop”、“历史地理”、“地形建模”和“可视化”等31个关键词只出现1次。

综上所述分析可得,数字人文凸显了大数据时代跨界合作的重要性。数字人文近年来的文献分布表明,在人文社会科学研究领域,数字人文才刚刚起步,针对特色馆藏、档案、考古、文物收藏、历史、艺术、音频、小说、新闻、遗产、民族志等载体,以“人文计算”、数字化、信息化、网络可视化分析、文本挖掘、数据分析为特征的研究方法越来越多地被采纳进行虚拟再现、共享平台建构和数字人文中心成立等。因此,正如南京大学孙建军教授所讲,大数据时代人文社会科学研究取向和热点问题将着眼于跨学科协作、跨平台协作、海量资料加工以及人文社会科学计算化^[1]。人文社会科学某种程度上具有批判性与人文关怀的特性,尽管大数据技术不是万能的,不能解决一切问题,而只是决策的一种量化手段,但是,我们必须正视大数据给了人文社会科学向多元化转向的机遇和挑战^[2],只有善用数据说话^[3],从数据到决策,才能加速人文社会科学的进一步发展,与时代对接。由此出发,数字人文环境下人文社会科学可以从文本挖掘、GIS技术、文本可视化、特色馆藏和语料库等方面入手开展跨界合作研究。与此同时,凭借资源优势、人员优势和技术优势,图书馆在数字人文广泛发展过程中将扮演重要角色,其中之一即可通过文献资源调整,在发展布局、来源出版物、学科方向、地区、机构和研究方向分析基础上提供相应发展导向和支持服务。

[1] 范佳.“数字人文”内涵与古籍数字化的深度开发[J].图书馆学研究,2013(3):29-32.

[2] 王晓光.“数字人文”学科的产生、发展与前沿[EB/OL]. [2011-11-24]. <http://www.xzbu.com/4/view-3204229710.htm>.

- [作者简介] 董政娥,女,1978年生,东华大学图书馆副研究馆员。
陈惠兰,女,1964年生,东华大学图书馆副馆长,研究馆员。
收稿日期:2015-06-04

109