【史学理论】

大数据技术与传统文献学的现代转型

刘 石 李飞跃

【摘 要】大数据技术引发了传统文献的生产方式创革、结构形态新变和获取方式拓展，文献的碎片化、标准化、结构化与可视化形成各种文本集、数据库等“宏文本”“超文本”，促进了文献的关联与知识的再发现。网络分析、文献计量、主题模型等文本信息技术的应用，可以革新传统文献学的实践路径，增强传统文献研究的整体性和实证性，催生新的研究范式，促进传统文献学的现代转型。当代大数据技术改变了我们对传统文献学的认识方式和把握尺度，反映了人们对知识挖掘、组织、管理与再造能力的追求。

【关 键 词】大数据；传统文献学；知识形态

【作者简介】刘石，清华大学人文学院教授；李飞跃，清华大学人文学院副教授(北京 100084)。

【原文出处】《中国社会科学》(京)，2021.2.63～81

【基金项目】本文为国家社会科学基金重大项目“基于大数据技术的古代文学经典文本分析与研究”(18ZDA238)阶段性成果。

文献素指载有历史信息的文字资料，今已成为“记录有知识的一切载体”的代称。[①]在甲骨、金石、简帛、纸张之后，文献进入了数字化时代。数字文献是以数字代码形态存在，依赖计算机系统存取和传输的文本、图像、音频、视频等文献。大数据时代的新文献形态如电子文本、文本集、数据库、知识库、系统平台等，在体量、结构、组织、管理等方面呈现出与传统文献不同的特征。大数据的目的是将海量数据转化为知识(Big Data to Knowledge)，[②]美国塔夫茨大学古典学教授克雷恩曾提出过一个发人深省的问题：“你怎么处理100万册的图书？”[③]大规模文献整理、文本挖掘与知识转化不同于小样本研究，工具和模型的使用是大数据研究与传统文献整理及研究方式的最大区别。

传统文献学，前人又称“治书之学”，亦即围绕古代典籍进行搜集、整理与研究。它在长期发展过程中，形成了深厚的知识积累、相对确定的研究范围、自洽的理论体系和成熟的研究方式。大数据技术长于数据挖掘，而传统的文献研究者实际上也是“数据挖掘者”，只不过挖掘的对象和使用的方法不同而已。大数据技术引发了文献生产的创革、文本形态的新变和知识获取的拓展，最终将促进传统文献学的现代转型。审视大数据技术与传统文献学的通变，不仅可以发明大数据技术下新型文献学的实践功能，也可借此认识大数据技术与传统学术的深层关联。

一、文献生产的创革

传统的文献生产一般包括写抄、刊刻等文本制作，校勘、辑佚等文本整理，注疏、考辨等文本研究。与传统文献研究相类似，大数据研究同样关注文献整体特征和内在结构特征。不同的是，传统文献学的主要处理对象是较为固化的文献形态和具体知识，比较重视经验与思辨；大数据研究主要处理多种类型的文献形态和海量知识，更依赖工具与技术，例如文本分词、词性标注、命名实体识别、句法分析、特征提取、情感识别、自动纠错、可视化呈现等，同时往往会对文本库及其分层子库的数据来源、数据量、数据格式、输入机制、参数

指标、算法工具等进行说明，其对文献的处理方式和功能建构也因而极大突破了传统文献的生产方式，实现了文献知识的再发现与再生产。

(一)通过分词、标引、词向量等技术实现原始文献的碎片化与颗粒化。古代汉语分词是将汉字序列切分成单独的词并按照一定规范重新组合成词序列。古籍通过分词、标引、抽取等方式，生成各种新的知识单元，产生新的知识形态如语义网络与知识图谱。知识图谱对每一项有意义的知识单元都赋予独立标识并以标准方式(RDF)进行描述，所有数字对象均按照领域规范编码，形成数据化知识集。古代汉语字词在不同历史时期有着不同的含义与侧重，通过数字编码，每个字词的读音、义项甚至义原都有一个身份代码(ID)。数字资源唯一标识符系统(CDOI)以及语义空间向量表达，在单个字词的基础上按照读音、语义、用法等进一步颗粒化，便于按字数、字频、字量、词汇、句法等统计，从多维度呈现文本特征。

当典籍根据语义标注的粒度被划分为篇章、段落、句子、词组、词等不同的簇，文本即可作为词汇的集合进入计算分析和知识生产的过程。在空间向量模型中，文本最终会被表示为向量。以往的典籍通常以点线面的二维空间模式呈现，而向量空间模型的应用可将文本以多维和高维模式呈现。计算机通过计算可获得语言特征的实例，如特征词的词性、词间关系、词位置分布等，继而分析特征词的出现频率、分布规律和语境特征，由此归纳作品整体的特性和风格。谷歌与哈佛大学共同开发数据库，对1800年至2000年出版的近520万册书籍的单词和短语的使用频率进行统计，推出书籍词频统计器，可查询词或词组在过去数百年的典籍中出现的频率与变化趋势，用以探索名物的兴衰沿革、话题的热度变化、人物或群体的影响力等。④

词汇标注、词性标注和音韵标注，目的是表征传统文献用词的隐藏状态。通过对人物、事件、地名、职官、称谓等实体标引，可以对文本内容的微观结构重加组织。如对人名及字号、别名、谥号等标引，可使所有人物出现的文本位置都排列在主条目之下，实现相关知识的本体化。就是说，标引能够使原始文献基于实体或关系而形成相应的聚类，生成新的独立文本。将文献中的章、节、图、表、数值信息等拆分成知识单元，进行主题标引，便能够形成所需的结构化知识库。深度标注能使计算机快速准确地找到目标文献，从而有效建立文献之间的关联，进行统计分析和比较研究。

(二)通过建词表、定格式、序编码等实现传统文献知识的标准化。原生文献在经过数据化处理后以集合、向量、概率等替代物形态呈现，人、地、时、事、概念、术语、图表、图像等对象的异质性也在此过程中被抹平。所有知识都被数字化编码，碎片化与数字同一性增加了知识单元之间的关联，形成一个超大而密集的文献网络，研究者可以便捷地获取位于任何一个序列中的文本与知识集合。词表的形成建立在分词统计的基础上，同时自身也构成更多新文献与文本的基础。题名、人物、称谓、职官、典故、地名等实体名词词表的创建提高了分词的准确率，有助于生成诸如传记、年谱、资料汇编等独立文本或知识集。词表、格式、编码等标准的确立与推进，将有力促进文献整理的规范化、规模化。从原生文献中抽取出特定的事实信息，如从编年史书籍中抽取详细的年表、年谱，从方志中抽取地名，从人物传记中抽取人物行踪及相互关系，从作品集中抽取不同题材作品和相关评论，生成专题史料集，将会变得易如反掌。

电子文献的标准化及行业编码、国家标准的制订(包括电子格式、文本字符、图像格式、音声标准、索引工具等对象的标准化)将极大促进文献的传播、使用与研究。目前针对文献名目、主题、类型的规范工作，尤其对同一文献的不同责任者和同一责任者的不同称谓等进行消歧与合并已取得初步成果。针对文献作者、题名、摘要中的人名，

可依靠人名库规范获取统一资源标识符(URI),再行著录。上海图书馆的开放数据平台发布了人名规范库、华人姓氏表、中国历史纪年表、地理名词表等人文词表关联数据集,涵盖人、地、时、事、物等人文信息资源要素,为古籍信息资源语义描述提供了有力支持。

古籍预处理技术的标准化将推动自动化标引的发展。荷兰莱顿大学魏希德教授主持开发的线上文本标记工具"中文古籍半自动化标记平台"(MARKUS),一方面通过与中国历代人物传记资料库(CBDB)、中国历史地理信息系统(CHGIS)等资料库关联,借助规范数据针对历代人名、地名、职官等核心词汇自动标记;另一方面亦可借助载入关键字、正则表达式等对特定词汇及特定规律的字串进行自动或手动标记。标记后的内容可由系统汇成导出多种格式的数据文件,生成各种不同的文本或知识集。基于文献自身特点或相互关系而形成的多种元数据标准,包括字符、图像、格式等标准规范的确立,将极大促进数据、库、平台之间的融通。统一化的文本编码可由通用数字语言实现文本互通,最终形成一个巨大的"宏文本"。⑤

(三)基于检索、算法、模型等实现传统文献与知识的重新关联与结构化。传统文献的线性平面形态决定了零散个体的研究价值通常要置于一个以时空划分、以文献群为单位的整体中才能被发现。文本集、数据库将知识单元按其属性类别加以集中序化和整合,这时的文献就不仅包括隐性知识的外化和显性知识的内化,也包括不同形态知识之间的转化。检索一体技术让我们从文档和数据库中获取的知识条块化,事实上是不同知识碎片的整合与单元重组。文献碎片化、知识颗粒化,基于检索、算法和模型生成超越原生文献结构的知识单元(语义单位)。深度学习模型可利用已有文献信息,自动提取、学习特征,发掘其内在的文本与知识关联。语义技术和关联数据可深度揭示知识内容,形成多层次、立体化的知识网络,也将使大规模分工协作与聚合分析成为可能。

各类传统文献基于超文本链接和各种相关性产生关联,借助宏观的大型数据库、知识库、数据平台以及微观的分词、标引、词表及词向量等数据技术方法,通过数理逻辑、语义关系、分类聚类等知识联接形成新的文献单位。根据不同需求和研究目的进行多维度的文献特征提取及相似度计算,能在任意样本空间上实现文本聚类,析出各种文本集或知识本体。知识因新的技术而重新关联,它不再是线性平面文本中的字、句、段、篇联结的方式,而是文本与文本之间建立起的交叉、立体、动态关联,由此可以观察到许多在小数据环境中很难观察到的关系与性能。知识图谱作为融概念、实体、属性和关系于一体的知识库,可实现传统文献的语义检索、全面整理与深度揭示,也可将它们联结为更大的结构化知识。

"超文本"作为线性平面文本的对立物,具有极强的关联性与查询能力。随着多场景纠错、新词发现、词义演变、语义网络等技术的发展,尤其词向量技术的进步,文献中蕴含的知识将在多维空间中投射意义。比如,常见的校笺工作有望借助数据技术以知识图谱的崭新形态呈现,异文、相同与相似语汇、典故、地名、人物、职官、事件等能够自动聚类,显现出各类实体随着时代或区域而发生的变化。一些古籍或文献专题建构了融语义词典、知识地图、跨库查询为一体的专题语义检索模型。一些古籍数据库还配备了相应工具箱,辅助文本挖掘和知识发现。

大数据技术让人类第一次有了处理大规模传统文献数据的能力。基于大数据技术的新型文献学既内在于文本,又能够出离文本,用远观、算法、模型来发现和组织知识,从海量的数据中发现隐藏在传统文献中的知识、模式、关系、趋势与规律,这在小数据时代是无法做到的。正如舍恩伯格所说:"就像望远镜让我们能够感受宇宙,显微镜让我们能够观测微生物一样,大数据正在改变我们

的生活以及理解世界的方式，成为新发明和新服务的源泉，而更多的改变正蓄势待发。”[6]在具有深厚传统的古代文献研究中，大数据技术从知识获取、标注表示、取样阐释等方面带来了根本性变革，本质上是一种方法论和研究范式的革新。由“人文计算”到“数字人文”的转变，体现的是从最初的“技术服务于人文”逐步向“领域内独特创新的方法和研究”的转变，[7]一种融合了不同媒介与资源的学术新大陆正在浮现。

二、文本形态的新变

随着数字化和数据化的发展，传统文献的计量单位从部、册、卷、篇、页、段、行、句等，向基本储存单元(位、字节、字)、扩展存储单元(KB、MB、GB、TB、PB)等转变，文本越来越多按媒介、容量、格式等分类，如结构化数据与非结构化数据。索引、算法、模型等成为知识链接的主要手段。文档、资料集、数据库、智能平台与传统的别集、总集、类书、丛书、资料汇编等共存，逐渐成为研究者使用的重要文本形态。数字化文本没有了纸张、墨色、装订等物质标志与工艺，却具有了新的特征与功能，它们通过数码、词表、图形、音像等多种方式，承传和转化传统文本的丰富信息。数据化的文本不再是线性和平面文本，而是复杂的立体网络。我们可以基于分类、关键词及属性，从时间、地点、人物、事件等各种角度进入和打开这一文本。

(一)数据库作为一种“宏文本”。数据库收录的数字化文本，文本性质并没有改变。每个数据库都可以看作一种独立文本，不同的文本基于知识、逻辑、功能等被联结成为巨大文本，是别集、总集、类书、丛书等传统文献形态的革新。据统计，古籍电子文献库在5年前已达275种，[8]特色数据库如首都图书馆“古籍插图图像数据库”，专题数据库如“中国金石总录数据库”，整合式数据库如“中华基本古籍库”等既是内容丰富、规模宏大的文本库，也是相对独立的知识主体。单一、直接的文本阅读锐减，取而代之的是数据库形态的庞大的文本集合。利用数据关联、集成分析、文本挖掘等大数据技术，文本之间不同领域、主题、角度、层次的壁障被打破，文本被表示为嵌套层次，文本的关联性在一个层次上达不到的，却可能在更多的嵌套叠层中实现。近年国家社会科学基金项目尤其重大项目，许多都以建置数据库作为一项目标，数据库作为学术成果越来越受到认可。

检索界限消失后，古籍数据库可以最大程度地“一站式”获取所需文献资料。索引数据库、书目数据库、全文数据库等，一般可按时代、地域、作者、体裁等分类浏览，提供按字词、词序、词频、词组甚至语义检索，事实上具有独立文本的特征。关系型智能化的数据库作为一种文本，其形态与功能较纸本时代有质的提升。利用关系数据库技术，既可根据关键字词将很多小数据库连接成大数据库，又可信息无损地将大数据库拆分为多个小数据库，组合成不同的数据集合。数据的采集和处理技术的进步，催生了主题性语料库。通过知识提取、实体辨识、概念语义标注、语义知识关联描述，还可以实现各种数据库和知识库资源的聚合，甚至实现众包、共享。[9]

(二)历史文本的空间化与可视化。传统文献的线性平面化特征，决定了它在空间化和可视化方面的弱势地位。可视化能够包含多重变量，具有可读性与可理解性。“目前至少存在着资料离散和时空分离两大难题，不借助数字人文技术就较难突破和解决”，[10]以可视化方式呈现主题词，可以直接观察多主题间的相关性，显示文本间的隐含内容和潜在语义联系，同时对海量信息进行抽象和概括。王兆鹏主持研发的“唐宋文学编年地图平台”，借助地理信息系统(GIS)技术，按时间、地点、人物、事件、作品等要素将分散的古代史料及研究成果集成为数据库，一定范围内解决了时空分离的难题。罗凤珠等利用“中华文明之时空基础架构(CCTS)”建立了“宋人与宋诗地理信息系统”“唐代诗人行吟地图”以及宋人、宋诗、宋诗语

言分布地图等。[11]徐永明创建了国内第一个综合性“学术地图发布平台”,将历史人物的行迹和社会关系、人群分布和物群分布配上地理信息,以一目了然的可视化方式呈现。GIS技术显著地促进了传统文献的图表化、可视化,以动态的数字化地图和知识图谱体系,改变和丰富了传统的文本形态和使用功能。

“一图胜千言”,文本内部蕴含的信息也具有可视化潜力。通过发现古代汉语文本特定的词频模式(如高频词、异常词频),可以借助文档相似性比较、主题探测、趋势发现等探索文本中特定的隐含语义关系,将难以理解的抽象数据空间转化成具体直观的视觉空间。据斯迪芬·詹尼克等总结,文本可视化方法有结构图、热力图、标签云、地图、时间线、网络图。[12]一般来说,结构图用于展现文本层级,热力图用于显现章句重复,标签云用于展示字词和意象的相对比例,地图用于呈现人物活动的空间分布,时间线用于呈现故事或情感的历史演化,网络图用于展现事件影响或人物关系。比如,可视化通过压缩信息、集中描述数据特征和差异来揭示重要模式和关系,勾勒文本及所蕴含知识的发展趋势和重大变迁。统计《左传》每个段落的时间信息,可以得到公元前722—前468年间的人物、地点出现次数的曲线,[13]辅助分期和重要时间段研究。对《全宋文》中有明确籍贯的3317名北宋作者进行精确统计,并用QGIS软件对地理分布进行可视化呈现,可以发现这些作者在时空分布上的诸多特征。[14]

(三)异质同构与传统文献的跨文本融合。大数据是多源异质的高维度知识,统一化的文本编码为信息联通提供了基础。数字化技术将各类知识载体如图书、报纸、乐曲、照片、绘画、视频、音乐和信息对象如文字、声音、图像、数表等转换为二进制,经计算机统一处理,产生新的数据文本。与传统纸质文本不同,数据文本不仅提供多对象聚合,还提供实体链接和交叉检索,最大程度地为构建叙述场景提供便利。未来的文献形态远不止于文本、图形、图像,还包括音频、视频及增强现实、虚拟现实等。动态、关联、立体,是未来文献的基本特征。

海勒斯说:“阅读总是由复杂多样的时间活动构成,但在由话语、形象、声音、动漫、图像和文字构成的21世纪阅读环境中,我们需要重新思考何为阅读,以及它是如何操作的。”[15]超文本打破了传统文献的线性平面结构,超链接和知识图谱使得阅读路径更加灵活自由。文本库与图像库、声音(方音)库、信息地理系统、社会网络、多语种翻译平台等连接,从而实现更多层次、最大范畴间的文本融通。甚至还可以纳入使用者维度,通过为用户和研究者打标签和画像,追踪知识的接受、传播与演化。

数字技术“重新媒介化”了印刷文化所塑造的社会形态,[16]语音识别、图像识别及自动翻译技术,使得数据化的古籍可以超越国别语言文字,实现同一语义层面的关联与比较,促进多语种文献和比较文化的研究,为人类知识图谱的绘制提供最大一块拼图,促进这一新时期人类文化“巴别塔”的建立。融合多媒体和超文本检索技术,实现跨时空、跨语种、跨媒介的检索,德里达所说的“万物皆文本”正在实现。

三、知识获取的拓展

传统文献通过篇章划分、页码标记等方式规定了逐段逐页、线性平面的阅读顺序。由海量文献组成的大样本和高维变量的数据集,量大质异、多源分散,超出人类在可接受时间下的收集、管理和处理能力。相较具体知识的直接获取和归纳演绎,在关联化、计量化和模型分析基础上统摄不同知识之间的复杂关系,从碎片化、多维异构的海量知识中获取并融合成系统化或创新性知识,是未来学术发展的总体趋势。

(一)知识的关联。托夫勒曾展望未来说:“‘知识就是力量’的旧观念,现在已经过时了。今天要

想取得力量，需要具备关于知识的知识。”[17]“知识的知识”就是要观照到知识之间的联系与组织。文献作为一种语言和知识系统，字、词、句之间皆可构成特定的复杂网络关系。知识超越简单的时空排序、内容关联和页码顺序，通过关键词、类别、主题、命名实体、函数、图表等实现跨文本甚至跨媒介关联，通过界面或网络联结呈现。知识网络让研究者能直观发现在词频统计之外的知识内部的更深层关系，如整体网络特征、核心人物功能与不同时期人物关系的演化模式。共被引分析通过引文之间的共现，可实现基于知识的聚合，解释知识的主题结构和新颖度。社会网络分析关注的是关系和关系模式，如弗兰克·莫莱蒂通过建立《哈姆莱特》的人物关系模型，发现了霍拉旭在情节发展中的中心地位，用网络模型将线性情节关系视觉化，展现了文体的叙事功能。[18]

古代典籍一般有着独特的结构形式和语义特征，通过数据挖掘可以发现文本中隐藏的模式。如统计先秦诸家学派中的共有词型序列，计算各学派高频词型等级之间的相关系数，可以发现儒道两派相关度最高，道家与其他各学派之间相关系数均值最大，说明道家对其他各学派的影响最大。[19]基于《左传》人物事件表，将属于同一历史事件中的人物视作关联人物，构建人物关系矩阵，可以发现《左传》中春秋时期的每个人物平均与书中其余10人通过同一事件产生联系，历史人物的关系整体上有低平均距离、高聚集性的特点，人物之间符合“四度分隔”理论，“晋国”与“孔子”在《左传》中具有特殊地位，等等。[20]此外，从强弱关系、网络中心度、结构洞等角度对社会网络进行探索，对人物关系及功能进行挖掘，可以超越“主要人物/次要人物”“圆形人物/扁平人物”“性格中心论/功能中心论”等传统理论。[21]

在大数据知识关联中，人们更关注的是知识信息的网络结构与流动转化。随着更多要素和变量纳入，知识会呈现不同的形态、性能与趋势。在更为宏阔的视域下，知识获取已非直接来自单个文本，亦非来自文本本身。如远读，“它使你关注比文本更细微或更宏观的知识板块：方法，主题，修辞——或者文类和体系”。[22]关联数据既提升了知识发现能力，也增强了结果验证能力。包弼德基于CBDB数据库和CHGIS平台，从人物群体和地理信息两个方面探讨社会网络对宋代道学传播的影响。[23]王军根据《宋元学案》将宋代理学衍化脉络可视化，借以观察宋代理学各门派各学说的消长流衍。[24]分析党争格局演化、诗人交游网络，考察家族兴衰因素，分析社会形态变迁，可实现传统纸质文献无法呈现的多维、动态效果，最大程度复现场景和还原史实。

(二)知识的计量。通过对知识本体、要素及关系的刻画与计量，尽量精准把握知识的特征、规律与趋势。文献学关注的作者归属、文体分类、主题异同、语义辨析等问题，正是统计分析学之所长。文体风格的基础是语言数据采集与模型建立，在词汇、语法、篇章等层面发现文本中蕴含的语言模式。通过分析作者用词习惯(尤其功能词、非主题名词或动词等边缘性词汇)或标点习惯等，可以发现作者自己也未必意识到的词语长度、句子长度、高频词汇、动名词使用比率等写作习惯。用大量已知文本训练机器识别作家的特色语言，可依据匹配程度揭示文献与未知作者的关联。

文献学的统计方法和数据模型不仅要从已知事实推及未知事实，还要借助定性和定量的描述性公式和算法进行通式建构。人们利用文本挖掘工具，从词频、意象、词汇、语义网络、字向量、情绪等维度，分析作品的常见意象、典型形象、情感倾向等。1993年，曼博教授设计用于文本相似检测的系统工具，首次提出了“指纹”概念。[25]早期针对知识产权保护而设计的文本相似度检测系统，主要是通过关键词操作和计算来判断文本相似度。后来一些技术平台依靠词句特征、实例特征、语义相似度、属性相似度算法等技术，实现了发现不同

文本间仿拟和互文等关系的功能。

文体的发展伴随着“变体”与“破体”,也就是说文体虽然代表着一些共性特征,但又始终与特征的变异并存,如宋人所谓“以文为诗”“以诗为词”等。一些新文体或分支文体的产生并非偶发的突变,而是基于经典文体之间的互相影响。可利用大数据对某一特定文本所含“文体因子”进行辨析,从而确定它受哪些文体影响,又影响了哪些文体,文体间的距离有多大,最终或可聚类不同文体的核心特征,构建出文体之间交互映射的网络。

主观文本(长文本)一般会有情感基调和情感走向,对情感词汇标引,可揭示语篇层面上的情感流动。对不同时代或群体的情感用词统计聚类,可自动生成“情感辞典”。在此基础上,可以对不同的文本单元如句子、段落和全文作出情感分析,也可以对整篇以及整部、多部作品作多维度如作者、体裁、题材、时代、地域、社团、流派等的情感分析,绘制“情感雷达”,捕捉其如何表达情感、表达什么情感,以及情感如何分布和演化等。通过情感标签与计算,能够进一步探究不同历史时期的社会现实与作品之间的对应关系。分析情感图谱上的关节点,可以探究情感症候与文献生产、文体流变之间的关系。循此思路,情感分析方法还可用于分析诗词曲格律情感表达模式,即哪些格律更多用来或者更适合用来表达哪类情感。另外,针对诗词文中的“意象”,借助深度学习技术建立意象表征和情感图谱,方便学者分析文学作品中意象与情感刻画的演化。通过数字转化及语义网络分析,人们将对文献与文本含义的丰富度产生前所未有的认识。

情感计算的本质是对语言评价义的挖掘,可利用情感词之间的相似度和语义场分析情感倾向。把每段文本按照情感强度标记分类,然后进行机器学习,再对新的文本进行测试,可以发现不同情感强度的具体表征。意大利学者史华罗研究了中国15–18世纪典籍中的情感与心境词汇,探析其情感行为包含的社会和思想含义。[26]还有一些学者利用《李娃传》等唐传奇中的情感词绘制出不同人物随情节发展而变化的情感曲线。情感倾向分析和对话情绪识别应用已在社交媒体中取得了显著成效,在文本分析中也必将有更大的开掘空间。

(三)主题模型提取。建构整体性的文史研究,靠以经典阅读、个人体验、意义阐释为代表的文本细读和个案研究等传统方式难以完全实现,它需要相应的视野、技术和方法,其中重要的一点是主题模型识别与提取。文本由词群组成,每一组词群又可以理解为一个主题,所以文本是由一个或若干个主题组成。主题是基于概率分布的词语,主题模型是用一些特定的词语分布来刻画主题。共词分析通过分层聚类揭示词与词之间的关系,进而分析它们所代表的主题与结构。

模式识别是计算机擅长的领域,计算机可根据不同文献设置参数,提取所需主题。目前智能媒体中使用的自动摘要技术,已能自动抽取关键信息,根据需求灵活控制摘要长度,并用于内容理解、智能写作等,为主题分析带来了新的契机。布雷、吴恩达和乔丹于2003年提出了“隐含狄利克雷分布”(LDA),[27]可自动完成对文献主题的归纳。艾伦等人运用LDA主题模型对汉典古籍网站上18000多篇、1亿多字的中国古典文献进行主题识别,开发了主题模型可视化工具,给出了主题分布和相关关系:当输入某一篇文本的标题时,能够给出与该文本相似度从高到低排序的其他文本及各自的主题分布;当输入某一主题时,可以得到该主题分布率从高到低的所有文本。[28]LDA主题模型能够探索一篇文本的主题构成,每个主题的词汇分布,文本之间的主题相似性,某个主题在哪些文本中分布较多及其在语料库中的意义,以及比较文本中每一段落或篇章主题的集中程度及段与段、篇章与篇章之间的主题连续性。据此,可比较不同作品在主题内容方面的复变关系、不同文体

在段落主题集中度与连续性方面的功能异同，以及不同作者在相同文体下的写作风格差异。尼克尔斯等人运用LDA主题模型对从先秦到宋代的数百万字语料库进行训练，通过比较《论语》《孟子》《荀子》主题中的词汇权重、文本中的主题权重与语料库中的主题权重后发现：《论语》与《孟子》的共有主题在《荀子》中出现概率也较高，而《论语》与《荀子》的共有主题在《孟子》中出现概率则较低；在《论语》中比重最大的主题在语料库中出现的概率也最高。[29]

新工具和方法赋予我们观察超长历史时段文化现象的新视角，也赋予我们获取新知识、发现新问题的能力。传统文献研究强调研究者的积累性、经验性、直觉性和思辨性，通过对已有研究成果的整理、分析、归纳、提炼进行知识发现和创新。大数据技术基于观察数据、实验数据、模拟数据，通过数据“发声”和获取知识，是对传统文献研究方法的拓展。在这些实证研究中，信息素养的意义可能要高于传统对知识体系的掌握，掌握获取知识的能力可能比掌握知识更重要。一些计算机专家正致力于开发主题模型工具套件，让主题提取变成简单的命令录入，从而降低应用门槛。

文本挖掘和知识发现并非单一线索顺势而下的简单知识延伸，而是多维度、多场景的复杂知识扩展。传统文献学主要是对显性知识的处理，侧重知识体系梳理和结构特征分析，史料存在方式的分散和史料承载知识的庞大，使得人们对其中的隐性知识、动态知识和宏大知识关注较少，“就像你驾车驶过一个大陆，注意到山脉和地界，却无法看到地球的曲面。一双眼睛看不出地平线的弯曲，一个读者的记忆也不能把握人文历史的宏大模式”。[30]梁启超昔年提倡历史统计，也是基于“往往有很小的事，平常人绝不注意者，一旦把他同类的全搜集起来，分别部居一研究，便可以发见(现)出极新奇的现象，而且发明出极有价值的原则”。[31]大数据技术易于弥补传统史料存在方式的不足，帮助发现知识因规模庞大而被遮蔽的变化弧线与一般规律，从荷马著作中读出“荷马也不知的东西”(In Homer more than Homer knew)。[32]

四、传统文献学的现代转型

莫莱蒂认为，文学史就是文学的屠宰场，抽样研究致使大量书籍永无出头之日：“如果我们今天在19世纪英国的小说中选择出200多部经典(已是数量不菲)，它们也不过占全部出版小说的0.5%，那剩下的99.5%呢?”[33]中国传统文献研究集中于有较高史料价值的经典文献，但这些文本在全文数据库时代只能算是样本，不能概括或代表历史全貌。大数据技术追求的“不是随机样本，而是全体数据”，“不是精确性，而是混杂性”，“不是因果关系，而是相关关系”，[34]有望在研究的科学性、整体性与理论范式上促进传统文献学的现代转型。

(一)革新传统文献学的实践路径。目录、版本、典藏、校勘、标点、索引、辨伪、辑佚等传统文献学的主要研究内容和工作，都极大得益于计算机网络、语料库和技术工具，研究效能将得到较大提升。

大数据时代的远读即如同传统文献学中的目录，“远读也可以看作是数字文本的可视化目录。它描述了文档集合的全局特征，让研究人员对超大数据集有了整体认知”。[35]新的书目数据著录标准力求覆盖所有内容和媒介类型，具有古籍循证功能的联合目录系统正由多个单位创建，它在对传世文献全面收录的基础上，进行自动标引、信息抽取和语义规范，不仅能快速呈现作者、编者、版本、收藏、研究等信息，还能帮助用户进行资料收集、内容结构化、数据建模、多源数据融合，提供各种在线工具以支持统计分析、数据挖掘、知识推理等。

文献版本的分类除传统的标准外，还增添了信息技术层面的标准。依据文献数据碎片化、标准化和结构化的程度不同，文档集、数据库也具有

了版本意义。结构化与半结构化、标引版、切词版、清洗版等，将成为未来版本称引的新常态。未来的善本标准可能不再是传统的足、精、旧或“三性九条”，而是依据数据的精良程度来判定，如切词是否精当、标引是否准确、语义检索是否到位等。一种或一类典籍的标注越是丰富、全面，可提取和聚类的渠道越多，其版本价值越高。同时，古籍版本的计算机辅助鉴定，用地理信息系统呈现古籍的刊刻、流传情况等亦将取得引人瞩目的成果。

数字文献在典藏和流传上具有天然优势，重要文献的扫描、保存和开源共享正是大数据技术和数字人文兴起的基础。大数据技术在自动比对的广度和精度上都非人力所及，自动断句、标点、比对、文献关联性、风格相似性分析等技术手段，不仅可辅助完成一般校勘任务，也有利于发现文献的源流及相互的影响，辨伪学必将从中获益。索引是一种工具，也是一种知识的聚合与重构。在传统索引模式的基础上，把地图的可视化效果和位置解析功能同文献资源结合，可以形成兼具时空特性的直观检索集。针对大数据资源的语义空间维数高、动态增长、数据分布不规则等特点，已发展出了高维数据索引和分布式语义搜索技术，文献资源的深度挖掘与综合利用可望实现。

语言处理应用技术中的序列标注(分词、词性标注、命名实体识别)、分类任务(文本分类、情感计算)、句子关系判断(蕴含或矛盾、相似度计算)、生成式任务(机器翻译、问答系统、文本摘要)，将为文献学研究提供重要支持。词频统计、自动纠错、文档校对、语义分析、作者归属、主题模型、地理信息系统等信息技术，逐渐发展成为文献学研究的必备技能。语料库、数据库及文本工具箱已成为当今文献研究的新基础设施，大数据技术正助力传统文献学实现突破性发展。

(二)增强传统文献研究的整体性。传统文献是平面和静止的简单形态，而数字化文献是文本类型及结构复杂、数据表征及性能多样的知识系统。这一知识系统的各要素都有自己的目标和行为、自主性和主动性，存在非线性相互作用，并随时空变化而不断有新的结构、功能或状态出现。[36]原有分类和研究方法，往往是“以小见大”“小题大做”，习惯于依赖世界的某一部分来解释整体。数据化改变了我们对研究中“证据”的本质的理解，正如马修·约克斯所说：“大量数字语料库提供给我们前所未有的文献记录，也要求一种新的证据搜集方式与意义生成过程。二十一世纪的文学学者不能再满足于轶闻式的证据，不能再从那些少量的、即使可以称为‘代表性’的文本得到随机的‘事实’”。[37]大数据时代的庞大数据集合，使得之前由因果律主导的演绎法和注重实验的归纳法不免捉襟见肘。人们不再满足于简单地寻求孤立事实或线性因果，转而致力于万物相关性的发现与解释。

舍恩伯格指出，执迷于精确性是信息缺乏时代和模拟时代的产物，“只有5%的数字数据是结构化的且能适用于传统数据库。如果不接受混乱，剩下95%的非结构化数据都无法被利用……通过接受不精确性，我们打开了一个从未踏足的世界的窗户”。[38]基于大数据的复杂算法比基于小数据的精确算法更加有效，比如谷歌翻译系统拥有上万亿词汇信息的混杂语料库，包括一些不完整的句子、拼写错误、语法错误等。巨量信息的优势压倒了不够精确的缺点，反而能产生良性的结果。[39]对海量文献进行整体分析和远读，才能更有效地对某一类典籍及其蕴含的历史信息达到整体把握。我们在一定程度上应放弃对局部或细节真实的追求，转而追求对概率和趋势的认知。

随着样本量的增加和标准变化，以往建立在抽样或抽象基础上的结论可能发生改变。如CBDB数十万人的传记类聚后，“大量的数据统计分析结果告诉我们，中国过去的平均死亡年龄是61到63之间，这让我们在学校里学到的过去的平均

年龄是35变得没有意义”。[40]以往学者是基于大家名集得出一些规律性的声韵规则，一旦扩大到全样本，这些规则或显示其错误，或变得不那么确定，或一些未曾重视的特征开始凸显。古代文学批评话语中存在大量的引用、因袭、转述、暗指、解释、模仿、互文等诸多文本间性的具体表现形式，这些正是建立文学术语网络、文人关系网络、文学流派网络的第一手资料。借助计算机实现实体名词自动抽取，建立表征术语、文本、文人间关联性的文本网络，利用复杂网络或社会网络分析的手段深入挖掘其间的关系和模式，将成为大数据时代的文本细读。

(三)促进传统文献研究的实证化。相较基于信念、经验、学理的传统研究，大数据技术对古典文献进行全面高效的分析，可以提高研究结论的精准性、稳定性及可验证性。以往的研究多采取特征描述和定性分析，主要是例举式，一般通过若干案例和评价性意见进行分析判断。大数据技术则可以用科学的方法来解决那些感性和偶然提出的问题，如关于文学研究中的文体学和风格学问题，计算语言学、信息统计学等领域有大量成熟的方法和技术用来抽取和表征文体、风格特征。机器学习、数据挖掘、复杂网络分析等计算机研究领域中也存在大量经典算法可以帮助总结文体模式、分析文体演化。通过用词、句式、声律、用典、态度、情感甚至段落过渡、篇章组织等多重要素的复合定量分析，文献学研究的客观性和精密性就变得明显，文献学学科的科学性也会显著增强。

汉儒、清儒以“实事求是”相标榜的实证精神是古典文献学的优秀传统，大数据技术将极大地推动实证方法在学术研究中的应用。比如，使用深度学习的词向量嵌入技术，根据字、词的上下文建立语义表征；借鉴计算语言学中基于词典释义的词义消歧思想，计算诗文中字、词、表达式与语料库中的典故相似度，来识别诗文用典及其意旨；识别文学流派的特征，表征其内涵，比对“差异性中的一致性”，辅助流派间相互影响的判断；借助复杂时序网络分析技术，利用术语节点的度中心性、特征向量中心性的历时变化，分析术语及其蕴含的文学观念在古代文论史中的兴衰生灭；利用术语节点的介数中心性的历时变化，识别文学批评史的关键人物、概念范畴；等等。

迈阿尔说：“早晚会有一天，实证研究将统领整个文化研究领域。人们会通过实证来研究理论观念……正如近两个世纪以来科学以实证主义方法将人们从神学和迷信的控制中解放出来一样。”[41]大数据利用信息消减不确定性，语料库和检索技术的发展使定量证据激增，极大提升了学术研究的实证性与科学性。理论因材料的扩充而不断被检验和发展，一些体悟、感受与思辨内容也随着认识的细化深入，逐渐具有被标识与表示的可能。传统文献学的实证化与科学化，不仅为其他学科的发展奠定更坚实的文献基础，而且，在大数据视域中，文献学与诸多学科完全可以更紧密地结合为难以分割的一个整体。

(四)催生新的研究范式。大数据技术进一步缩小了定性研究与定量研究之间的鸿沟，在经典理论和实践经验之间架设了一座桥梁，有可能发现和提出新的重要理论。莫莱蒂的“远读”(distant reading)、艾伦·刘的“简式人文+”(short-form humanities+)研究、列夫·曼诺维奇的“文化解析”(cultural analytics)及理查德·罗杰斯的“数字方法”(digital methods)等学说，[42]都是大数据技术应用于人文研究而产生的新理论方法。其中，“远读”不参与实际文本的阅读，而是“让我们着眼于比文本更小或更大的单位”，[43]目的是在更长时段、更广范围中探讨文本的复杂性，彰显知识的语境性，比传统文献学强调的从文本出发更具多种维度。

计算模型扩展了审视维度，如基于相关系数、正态分布、方差分析等统计方法能发现更多整体性和深层次的知识。利用大数据技术协作构建庞大的新型文献数据库和知识库，有望绘制出古代

物质、精神世界的隐性结构，从而完成传统学科不可想象也因而从未被纳入学科范畴的目标。2011年，以哈佛大学让一巴蒂斯塔·米歇尔为首的研究团队在《科学》杂志上发表《使用百万数字化书籍的文化定量分析》一文，提出“文化组学”(Culturomics)的概念，认为词汇像基因一样包含可继承的信息，通过在海量数据中提取并分析某些词汇在图书文献中的增长、演变、消亡等趋势，有可能观察到大范围内文化的发展趋势和演变规律，为理解人类文化的变迁提供直观证据和结论。[44]与之相似的还有“群体传记学”(Prosopography)，通过收集人物姓名、出生地点、生卒年月、家庭成员、教育背景、职业经历、社会关系、宗教信仰等信息，可以认识社会群体的共性特征和个性差异。对于有悠久传统的中国古典文献学而言，运用这种文化组学理论研究观念嬗变、情感特征、文体风格、互文重出、分期分派等，可望增添前所未见的研究视角，催生前所未有的研究成果。

在大数据基础上，除了涌现的“计算机+”的跨学科交叉研究，还催生了一些具有学理特征的研究门类，尤其在文体风格方面出现了诸如统计文体学(Statistical Stylistics)、计算文体学(Computational Stylistics)、计算话语学(Computational Textlinguistics)、语料库文体学(Corpus Stylistics)，通过对文本语义、词性、句法标注，在自动标点、自动注释、词典编纂、风格溯源、自动摘要、机器翻译等领域已经产生了日益广泛的影响。

在大数据时代，传统文献学正面临着前所未有的大转型。随着电子化、数字化尤其大数据技术应用于人文研究，更具方法论和本体论的信息科学的出现已是不争的事实。近年出现的“电子文献学”“数字文献学”“数字目录学”“人文计算”“数字人文”等概念，其内涵无不体现了大数据技术与传统文献学血脉相承而又功能各异的特征。大数据时代的文献学将拓展到古籍字符识别与编码、数字文本处理、图文设计与制作、元数据标准、数据挖掘、地理信息系统、文献数据库设计、古籍信息系统与智慧平台开发等领域。传统文献学的目录、版本、校勘、典藏等知识门类都将得到升级迭代，可以说是继秦汉以来最大一次文献、文本、知识的管理变革。大数据“重新定义何为我们所认为的遗产，并要求我们找到新的方法和工具来概念化和管理这些日益增长的物质”，[45]它极大拓展了人类的经验范畴和知识能力，是对以内省式研究为主的传统研究范式的超越。

传统文献学是一门建立在文献获取、分类、整理与研究基础上的学术门类，历史上的字书、韵书、类书、丛书，哈佛燕京学社编纂的各种“引得”以及中法汉学研究所编纂的各种“通检”，乃至作家的年谱长编、研究资料汇编等，都是传统文献学时代的“数字化”，可以看成小样本时代的大数据技术，而大数据技术则是大样本时代的文献学。文献学和大数据技术都有方法论与知识本体特征，大数据时代的文献、文本与知识以数据形式存在，而数据不仅最终仍然要转换为文献、文本和知识，毋宁说，数据也是文献，是现代文献学赖以生存和发展的新型文献。从知识的发现、组织、管理及应用来看，二者目的和功能一致，实质相通，都反映了人们对知识挖掘、组织、管理与再造能力的追求。

余论

大数据技术是一场知识革命与思维革新，促进了传统文献学的转型与拓展。通过改变知识的切分、标引、聚类与呈现方式，大数据技术可以让原本庞大的文献及其间蕴含的知识变得更加浩瀚无穷，同时也为学者提供更多差异化、整体性、趋势性研究的可能。资料、检索和认知边界的同时拓展，正在使传统文献学实现“轮廓重绘”。

大数据技术促进了研究对象的交叉融合、理论方法的移植渗透，利用大数据来拟合、校准、检验因果关系与相关性，利用模式识别、算法模型等辅助处理海量和随机数据，可帮助我们超越一般

经验和思维，更为全面、客观、动态地把握和认识世界。大数据思维让我们具有前所未有的大局观，寻求处理传统文献的新路径和新方法，在开放共享中提升学术研究的造诣和境界。

需要指出的是，大数据技术在传统文献研究中也存在着局限和问题。数据资源不可能全部获得，“用数据说话”不等同于数据即是客观事实。数据量大不一定等于有用的信息多，大量的含偏差数据会引起语义整体性的忽视与破坏。大数据简化了人们对数据差异性的认知，为了提炼与呈现一部分结构化信息而致使大部分文本内容被省略或变形。“远读”有可能导致“文字被机械化切割，粗暴地重组为自动聚类的短信流和维基词条一般的快读素材”。[46]随着经典文献收集完成，资料的价值密度显著衰减。大数据技术在一定范围内穷尽史料之后，“所期待的‘史料大发现’的时代并没有到来，我们依旧要在那几部最基本史著的字里行间寻求突破；技术手段的更新，也并没有带来终极意义上的学术思维革命，前辈学者经典学说的理论框架短期内尚难以全面突破，我们所做的只是在修正、完善和细化”，[47]这也可能是一段时间内我们必须要面对的现实。

同时，也需警惕技术方法的局限性和负面影响。检索生成数据较容易，原因的分析则较难。大数据抵消了少数个体的特殊性，减损了读者对文本信息进行深度理解的意愿。量化数据库通过建立变量分析史料属性，而如何清理数据、设计和提取变量，还需借鉴传统文献学的考据方法。大数据的精准分析和标准化归纳，夸大了纯粹理性的中立原则，使数据所关联的行为主体缺乏深度在场感。研究者对分析工具的认识不足而误用、统计方法单一、缺乏有机模型和统一理论的支持、机器学习算法的黑箱问题等，可能导致结果的误差或结果可解释性的匮乏。

因此，我们应充分认识到大数据技术并未改变知识的本质或人追求知识的本质，它依然只是一种认识世界的工具和方法，是人的延伸，不能也不会替代人。其开放性虽然拓展了传统量化分析方法的空间，但如何避免以抽象运算取代解释性理解，如何注意弥补大数据在数据信度、主体呈现和因果解释等方面的缺陷，却同时变得紧迫。

但也应该看到，随着数量和维度的增多，知识的高度语境特异性反而可以让研究者有条件更多关注审美、情感、意义等层面的问题，发挥人文经典通约性和稳定性的共情能力，让我们成为“我们”。同时，人机之间的互动与互补，也将促进科学与人文之间的知识重构与认知升级，在一个更高层面带来新的确定性，实现意义重置和世界重建。这也是人类对自我局限的一次突破和面临世界巨变的一次调整。

大数据及其相应技术已经成为当代科技发展的重大标志，渗透到当今社会包括学术研究在内的各个领域，对我们的知识体系、认知观念及思维方式产生了重大影响。傅斯年曾说：“凡一种学问能扩张他所研究的材料便进步，不能的便退步”，“凡一种学问能扩充他作研究时应用的工具的，则进步，不能的，则退步。”[48]人类科研范式从实验科学、理论科学、计算科学，发展到了以算法、模型为基础的数据科学。在数据科学阶段，传统文献的自闭被打破，文献的概念延展至更为广阔、复杂、多样的领域。甚至可以说，每个人都是一个文本，每类人都是一种文献，人类这一文本与文献中寓含的物质、文化、行为、观念等知识都可以用数字标示和还原，并通过各种模型和算法来呈现和解释。虚拟现实、增强现实及混合现实正在模糊以至瓦解虚拟与现实的边界，研究对象重塑与历史场景复原将极大改变人们对文献的理解与感悟。文献、文本、知识以数据的形式被重新塑造，数据不仅被看作刻画事物关系的参数，承载了新的知识形态，还被赋予世界本体的意义。伴随着万物互联，“一切皆文本”的趋势越来越明显。

大数据将以往被分裂和隔绝的事物重新连

接，改变了我们对文献、文本、知识的认识路径和把握尺度。麦克卢汉说："任何媒介(即人的任何延伸)对个人和社会的任何影响，都是由于新的尺度产生的；我们的任何一种延伸(或曰任何一种新的技术)，都要在我们的事务中引进一种新的尺度。"⑲文本与文本之间关联，文本与声音、图像、视频等关联，结构化宏文本与超文本让文本的界面大开，同时也让世界文本化、数据化而被把握和理解。在精神世界和物质世界之间出现了一种新的"数据世界"，虚实结合的赛博格世界带给人们完全不同的体验与认知。在抄印本和电子化时代日渐被穷尽的文献，因为数据化而重新变得不可穷尽。今日之大数据只是明日的小数据，大数据技术掌握的知识只是未来知识系统中的一小部分。大数据技术的发展带来了工具方法、知识形态和思维观念的革新，而机器学习、深度学习乃至更具广泛意义的人工智能技术方兴未艾，推动着研究范式和认知方式不断升级。大数据时代的新型文献学，或者说，大数据作为一种新型文献学，必将在更大尺度、更小粒度和更多维度中绽放知识之花。

注释：

①参见《中华人民共和国国家标准GB 3792.1-83——〈文献著录总则〉》，《图书馆学通讯》1983年第4期。《中国大百科全书》增加了"信息"内涵："记录有知识和信息的一切载体。"(《中国大百科全书·图书馆学、情报学、档案学》，北京：中国大百科全书出版社，1993年，第465页)古拉丁语中Documentun和Literatura都有"文献"之意，前者包括印刷品在内的一切文字记录，如碑文、古币图文等；后者一般只包括通常意义上的图书资料。

②Philip E. Bourne et al.,"The NIH Big Data to Knowledge (BD2K) Initiative," Journal of the American Medical Informatics Association, vol.22, no.6(November 2015), p.1114.

③Gregory Crane,"What Do You Do with a Million Books?" D-Lib Magazine, vol.12, no.3(March 2006), http://www.dlib.org/dlib/march06/crane/03crane.html, 12th November 2020.

④J. B. Michel et al.,"Quantitative Analysis of Culture Using Millions of Digitized Books," Science, vol.331(January 2011), pp.176-182.

⑤胡易容、张克：《从"数字化生存"到"符号的栖居"——论数字人文学的符号学界面》，《华南师范大学学报》2016年第2期。

⑥维克托·迈尔——舍恩伯格等：《大数据时代——生活、工作与思维的大变革》，盛杨燕等译，杭州：浙江人民出版社，2013年，第1页。

⑦N. Katherine Hayles,"How We Think: Transforming Power and Digital Technologies," in David M. Berry, ed., Understanding Digital Humanities, London: Palgrave MacMillan, 2012, pp.42-66.

⑧其中中国大陆166种、中国台湾70种、中国香港9种、中国澳门2种，国外22种，合作开发6种。参见张三夕、毛建军主编：《汉语古籍电子文献知见录》，广州：世界图书出版广东有限公司，2015年。

⑨如耶鲁大学研发的"广厦千万间项目"(10000 Rooms)允许用户上传文件，支持多人处理同一文本。参见https://tenthousandrooms.yale.edu/，12th November 2020.

⑩王兆鹏、邵大为：《数字人文在古代文学研究中的初步实践及学术意义》，《中国社会科学》2020年第8期。

⑪参见罗凤珠等：《唐代诗人行吟地图建构：李白、杜甫、韩愈》，《第四届中国古籍数字化国际学术研讨会论文集》，北京：北京国学时代文化传播股份有限公司，2013年，第169-171页。

⑫S. Jänicke et al.,"On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges," Eurographics Conference on Visualization, Cagliari, Italy, May 2015, pp.83-103.

⑬ 参见李斌等：《数字人文视域下的古文献文本标注与可视化研究——以〈左传〉知识库为例》，《大学图书馆学报》2020年第5期。

⑭参见王小红：《〈全宋文〉北宋作者的时空分布特征——基于QGIS等技术的研究》，《大数据时代的史料与史学》，北京：人民出版社，2017年，第251-265页。

⑮N. Katherine Hayles, How We Think: Digital Media and Contemporary Technogenesis, Chicago: The University of Chicago Press, 2012, p.79.

⑯关于"重新媒介化"(remediation)，参见Jay David Bolter and Richard Grusin, Remediation: Understanding New Media, Cambridge, Massachusetts: The MIT Press, 2000.

⑰阿尔温·托夫勒：《预测与前提——托夫勒未来对话录》，粟旺、胜德、徐复译，北京：国际文化出版公司，1984年，第113页。

⑱Franco Moretti,"Network Theory, Plot Analysis," New Left Review, vol.68(March/April 2011), pp.80-102.

⑲参见马创新、梁社会、陈小荷:《先秦诸家学派的相关系数与特征词研究》,《中文信息学报》2019年第12期。

⑳参见许超、陈小荷:《〈左传〉中的春秋社会网络分析》,《南京师范大学文学院学报》2014年第1期。

㉑参见赵薇:《网络分析与人物理论》,《文艺理论与批评》2020年第2期。

㉒Franco Moretti,"Conjectures on World Literature," New Left Review, vol.1(January 2000), p.57.

㉓参见包弼德:《群体、地理与中国历史:基于CBDB和CHGIS》,《量化历史研究》第3、4合辑,北京:科学出版社,2018年,第213-246页。

㉔参见王军:《从人文计算到可视化——数字人文的发展脉络梳理》,《文艺理论与批评》2020年第2期。

㉕Udi Manber,"Finding Similar Files in a Large File System," Proceedings of the USENIX Winter 1994 Technical Conference, San Francisco(January 1994), pp.1-10.

㉖参见P.史华罗:《明清文学作品中的情感、心境词语研究》,庄国土、丁隽译,北京:中国大百科全书出版社,2000年,第1-2页。

㉗David M. Blei, Andrew Y. Ng and Michael I. Jordan,"Latent Dirichlet Allocation," Journal of Machine Learning Research, vol.3(January 2003), pp.993-1022.

㉘Colin Allen et al.,"Topic Modeling the Hàn diǎn Ancient Classics," Journal of Cultural Analytics, October 2017, pp.1-34.

㉙Ryan Nichols et al.,"Modeling the Contested Relationship between Analects, Mencius, and Xunzi: Preliminary Evidence from a Machine-Learning Approach,"The Journal of Asian Studies, vol.77, no.1(February 2018), pp.19-57.

㉚Ted Underwood, Distant Horizons: Digital Evidence and Literary Change, Chicago: The University of Chicago Press, 2019, pp.ix-x.

㉛梁启超:《历史统计学》,《史地学报》1923年第2期。

㉜Jonathan Swift, Robert A. Greenberg and William Bowman Piper, The Writings of Jonathan Swift: Authoritative Texts, Backgrounds, Criticism, New York: Norton, 1973, p.570.

㉝Franco Moretti,"The Slaughterhouse of Literature," Modern Language Quarterly, vol.61, no.1(March 2000), p.207.

㉞维克托·迈尔-舍恩伯格等:《大数据时代——生活、工作与思维的大变革》,第27、45、67页。

㉟王军:《从人文计算到可视化——数字人文的发展脉络梳理》,《文艺理论与批评》2020年第2期。

㊱参见黄欣荣:《从复杂性科学到大数据技术》,《长沙理工大学学报》2014年第2期。

㊲Matthew L. Jockers, Macroanalysis: Digital Methods and Literary History, UrbanaChampaign: University of Illinois Press, 2013, p.8.

㊳维克托·迈尔-舍恩伯格等:《大数据时代——生活、工作与思维的大变革》,第64页。

㊴参见江大白、徐飞:《大数据:科学方法的新变革》,《自然辩证法研究》2016年第1期。

㊵包弼德:《数字人文与中国研究的网络基础设施建设》,《图书馆杂志》2018年第11期。

㊶David S. Miall,"On the Necessity of Empirical Studies of Literary Reading," Frame: Utrecht Journal of Literary Theory, vol.14, 2000, pp.43-59.

㊷参见大卫·M.贝里、安德斯·费格约德:《数字人文——数字时代的知识与批判》,王晓光译,大连:东北财经大学出版社,2019年,第132页。

㊸Franco Moretti,"Conjectures on World Literature," New Left Review, vol.1(January 2000), p.57.

㊹文化组学是"文化"与"基因组学"两个词的合并,旨在通过文本定量分析来揭示人类行为和文化发展趋势。J. B. Michel et al.,"Quantitative Analysis of Culture Using Millions of Digitized Books," Science, vol.331(January 2011), pp.176-182.

㊺James Murphy,"How to Do Things with Networks: A Response to Franco Moretti," 18th August 2011, https://magmods.wordpress.com/, 12th November 2020.

㊻Jaron Lanier, You Are Not a Gadget: A Manifesto, New York: Vintage Books(A Division of Random House, Inc.), 2010, "Preface".

㊼陈爽:《回归传统:浅谈数字化时代的史料处理与运用》,《史学月刊》2015年第1期。

㊽傅斯年:《历史语言研究所工作之旨趣》,《傅斯年全集》第4卷,台北:联经出版事业公司,1980年,第256页。

㊾马歇尔·麦克卢汉:《理解媒介——论人的延伸》,何道宽译,北京:商务印书馆,2000年,第33页。

