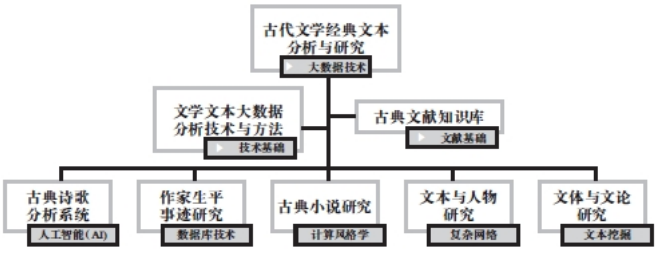
大数据时代的古典文学研究

2018年10月15日 12:32 来源： 光明日报



【编者按】对人文学者来说，作为工具的计算机，已从文献检索时代进入到数据分析时代。计算机不仅能帮助我们从海量文献中快速检索到所需的资料，还能以数据为基础帮助我们发现问题和分析问题。随着数字人文技术的发展，数据分析的技术和方法越来越有针对性和强效性，能清晰地揭示隐藏在文学史背后的作家与社会之间、作家与作家之间、文本与文本之间的直接与间接、显性与隐性的多种关联，能以全知型的视角系统整体地还原和呈现文学史的立体景观，改变传统的思维方式和文学研究范式。

目前的中国古代文学研究，在数据分析方面虽然已经起步，但还没有完全跟上数字人文的发展步伐，可用于统计分析的关系型文学数据库建设还比较薄弱，适用于古代文学研究的分析工具、分析方法、分析模型还相当有限。近些年学界和业界推出了相当丰富的数字化的文献资源库，如《中国基本古籍库》《中华经典古籍库》等，但主要用于检索，还不是结构化的能进行统计分析和再生知识的数据库。运用数字人文的分析工具和技术方法来研究古代文学，也取得了一定的实绩，但还处在尝试性阶段，未成规模，影响不大。

为推进数字人文技术在古代文学研究中的应用与突破，本期约请清华大学中国古典文献研究中心数字人文研究团队的刘石、孙茂松、张力伟和刘京臣四位先生从不同的角度笔谈他们的构想和规划。刘石、孙茂松先生构建了古典文学研究的分析模型，刘京臣先生阐述了基于社会网络分析的文本与人物研究的理路，既有理论的前瞻性，也有方法的可操作性；张力伟先生提出了建设“中国古典知识库”（CCKB）的宏大构想，令人期待！（王兆鹏）

一

20世纪60年代，电脑就被西方国家运用于人文学科研究，称为“人文计算”。美、英、法、德等国利用大数据技术研究文学开展得早、影响大，相继成立了国家级项目组或研究中心，致力于莎士比亚戏剧、法国中世纪诗歌等多语种文学经典的内容分析，产生了一批引人注目的理论著述与应用成果。

进入新世纪，一些研究机构及企业开始对书籍进行大规模数据化。谷歌与哈佛大学共同研发的数据库可对1600年至2000年间出版的500多万册书籍的单词和短语的使用频率进行统计，通过关键词使用频率的变化，可以崭新的视角揭示500年来人类文化发展史的总体趋势。伴随人工智能技术的进步，机器的深度学习在文本分析方面展现了惊人效率。《布谷鸟的呼唤》原是《哈利·波特》的作者J. K. 罗琳于2013年匿名发表的小说。牛津大学的Peter Millican和杜肯大学的Patrick Juola运用法律语言学的分析方法对比分析，推测它很可能是罗琳的新作，最后，罗琳承认这部小说确出己手。

国内在20世纪80年代也出现了“人文计算应用”的概念，一些学者开始致力于运用电脑技术研究人文课题。早期对古典文学尤其诗词的研究多为计算机或统计专业的学者。厦门大学周昌乐教授课题组针对宋词风格“豪放与婉约”的分类问题，研创了基于字和词为特征的风格分类模型、基于频繁关键字共现的诗歌风格判定方法以及基于词和语义为特征的风格分类模型。首都师范大学尹小林教授最早研发了“《全唐诗》检索系统”，北京大学李铎教授也研发了“《全宋诗》分析系统”“《全唐诗》分析系统”“《资治通鉴》分析系统”等。北京大学杜晓勤教授研发的“中国古典诗文声律分析系统”首次实现对中国古典诗歌及有关韵文进行批量四声自动标注和八病标识、数据统计功能，不仅有助于研究永明体诗歌的声病情况，还可考察永明诗律向近体诗律演变的环节和过程。中南民族大学王兆鹏教授是较早采用量化分析研究古代文学经典的专家，他先后主持了“中国古代诗歌史的计量分析”“20世纪唐五代文学研究论著目录检索系统与定量分析”等多个项目，尤其是唐宋诗词名篇的定量分析（排行榜）及国家社科重大项目“唐宋文学编年系地信息平台”引发了社会的普遍关注。

郑永晓先生数年前已经呼吁古典文学研究从数字化向数据化的转变。基于大数据技术对古代文学经典文本进行高效和深度分析，可将文学研究纳入到一个更宏观的视野，提高研究结论的精准性、稳定性及可验证性，促生新的研究理念、方法与范式。但总体来看，古典文学研究领域目前还基本处在古籍数字化、数字化检索和少数专题数据平台建设阶段。

二

现阶段数字人文研究的主要技术方法，包括机器学习与人工智能、数据库建设、计算语言学、社会网络与地理信息系统、数据与文本挖掘等方面。这些技术方法可分别用于古典诗歌分析系统的尝试、作家生平事迹研究、古典小说研究、文本与人物研究、文体与文论研究，涵盖了古典文学研究的主要方面。

基于这样的理解，我们拟以先秦至明清品类纷繁的古代文学经典文本为中心，利用计算机、统计学、信息科学等学科的新兴技术手段，形成如右上图所示的研究结构。

研究的流程是文学专家提出问题——技术专家设计算法模型——借助知识库或数据库等平台进行文本分析——文学专家对分析结果进行解析和研究。数据库建设、技术创新运用与文本研究三位一体。数据库是基础，文本分析技术是关键，最终要落实到发掘依靠阅读经验难以发现

的文本组织特征及相互关系，通过定量统计、定性分析，解决古典文学研究领域长期存在的疑而难决的作品归属、作品辨伪、异文辨析、修辞特色、风格生成、题材变迁、因革影响等方面的问题，期望在以下诸方向有所推进：

1.重新验证已有成说的经典史论问题。比如，提出“文必秦汉，诗必盛唐”的明代前后七子为代表的文人群体，其诗文创作是否落实和如何落实其文学创作的主张？利用共词分析、语义分析、人物事件交杂等技术思路，尝试全新分析和解决诸如文体形式、社团流派、人物关系、情节演进、阶段特征、历史影响等问题。

2.解决人力难以彻底解决的疑难问题，为作品归属、重出异文、改编续写、风格流派、文类划分等提供新的证据、思路与方法。如唐宋诗“体格性分之殊”的判断，诗词曲三种相近文类格律、用韵、题材、语词、典故、句法、意象、风格的穷尽性统计，为定性分析提供数据支撑，可以提高研究结论的精确性、稳定性及可验证性。

3.超越主观感受与印象分析层面，科学梳理文学史长时段中存在的特征、规律、关联性问题。比如陆游诗近万首，词自中唐产生而历经各代，他或它们的题材、修辞、风格变化轨迹究竟如何，数者之间的关系怎样？通过对一个作家或一类作品的“深度学习”（计算语言学专业术语），发挥其文本比对、关联分析等技术优势，追踪挖掘以往不曾注意到的迹象或线索，以期提高文学经典研究的可靠性与科学性。

三

利用大数据技术研究中国古代文学，对学术发展和学科建设的意义是明显的，特别体现在研究范式与思维方式的革新。

傅斯年认为，“凡一种学问能扩张他所研究的材料便进步，不能的便退步”。大数据技术可以实现相关研究史料的全覆盖，是对以往研究资料的极大扩充。目前研究中普遍存在的检索依赖会造成史料的类型遮蔽，特别是反证材料的遮蔽。检索依赖也会导致对史料的解读脱离历史语境，无数孤零零的没有历史气息的材料断片的组合，无法反映真实的历史场域中的问题。文学研究者接受的信息如果是非全息的，文史研究的科学性和有效性必然值得怀疑。全数据分析模式抛弃了随机性的样本研究模式，让研究者具有“上帝视角”，重视对事情整体系统的感知，又强调基于全数据的细节化，提高认知的精确度，是一种理想的学术研究模式。

传统的文献材料彼此间基本上呈现出相对明显的线性关系，可以找到前因后果，进而形成相对完整和自洽的因果链。大数据时代面对的只是具有相关性的海量数据，几乎不可能找到每个数据的微观因果链，如果坚持因果路径，将陷入无穷无尽的因果关系之中而茫然无措。因此，大数据时代不必非得知道现象背后的原因，而是让数据自己发声。对思想、情感和艺术为主体的古典文学学科而言，强调差异性、变异性和独特性的相关性分析方法比因果性分析方法可能具有更强的裁断力。

大数据技术的兴起，使数据采集、存储和处理极大地智能化、自动化。“全数据模式”将与问题相关的数据一网打尽，最大限度地摆脱客观条件局限造成的以局部论全部，问题可以得到更系统、更全面、更整体的刻画，从而得到更精确、更彻底的解决。这是数据化带来的一种严格意义上的整体论，将使思维方式从还原性思维走向整体性思维。

历史与逻辑、事实与价值的统一是人文社科研究的基本方法，大数据时代的研究尊重全体材料、重视量化分析和兼顾所有关系，这将有助于促进人文学科的研究由“解释性”向“求是性”转向。随着人的思想、情感、心理的数据化，人文学科的研究对象也能够实现数据化，可以通过数据挖掘、数据分析和数据建模来进行研究，这样人文学科也就由以往被认作非科学的学科跻身于科学成员的大家庭中，进而发展出人文科学。

总之，大数据思维为人文社科研究的变革与创新带来了千载难逢的历史机遇，正如美国康奈尔大学教授杰弗里·汉考克（Jeffrey T. Hancock）所说：“这是社科研究的一个全新时代，就好比显微镜的诞生对化学科学发展所起到的促进作用。”

需要指出的是，古典文学研究中新技术手段的应用需要充分依靠计算机科学和统计学的专业技术，在尚缺乏此类技术力量的今天，必然会促进学术研究人力资源的整合，倒逼跨学科合作研究的开展。但文学性问题的提出和分析处理不可能完全交给机器，也就不可能完全交给技术专家。相反，从问题的设置到语料的选取再到分析结果的解读、意义的阐释、体系的建构等，都将由古代文学和文献学相关领域高水平的专家学者完成。（作者：刘石，系清华大学人文学院教授；孙茂松，系清华大学计算机科学与技术系教授）