

量化方法运用于古代文学研究的进展和问题

——以近年数字人文脉络中的个案探索为中心

赵 薇

内容提要 近年来随着数字人文的发展,量化方法在古代文学研究诸多问题上进展迅速,计算技术与文学阐释结合,形成了可操作的批评路径。《红楼梦》作者归属判断透露出计量方法的施用和解释问题;声律统计和诗歌研究向大样本、长时段的宏观层面迈进,创生了独特的研究方法;文学史研究对系地维度的重视,带动了文学地理信息平台的建设;借助于文本聚类 and 分类算法,可以从反思的角度切入文类和目录学研究;社会网络和文本网络分析成为极有潜力的方向。数字人文取向的文学研究应避免简单的工具化倾向,文学研究者须知晓工具的算法实质,将数据驱动的探索过程和论证驱动的假设—验证范式更好地结合起来,将建模整合进严密的论证中,获得有效的评价。

关键词 数字人文 古代文学研究 量化方法 计算批评 远读

近年来,数字人文的兴起已经是一个毋庸置疑的事实。诚如郑永晓先生所言,近二十多年来,我国在古籍数字化、数字文献学和古籍智能处理领域的成就是巨大的^①。尤其在汉文古籍的 OCR 识别、自动句读和分词、命名实体、版本比对和校勘等方面,因能充分发挥数字人文的“工具属性”,所达到的水平是领先于国际学界的。与此同时,在事实发现和阐释论证并重的古代文学史、文学理论和文学批评研究中,数字人文研究也在文献数据化的基础上,逐步走向了可操作的计算批评实践,形成了一套其来有自,亦可接续传承的量化方法和独特体系。这一过程在现阶段更多地表现出一种以探索为主、成绩与问题同在、机遇与挑战共存的特点。鉴于此,本文不揣谫陋,以量化方法与文学问题的结合为重点,将海内外相关探索放到数字人文自身的发展脉络中,尝试以重要问题和技术自身的进展为线索,归纳出四个方面,从众多个案中择其要者而论,以期就教于学界同仁^②。

一 计量文体学的重要应用

1. 《红楼梦》作者归属判定:计算方法的施用和解释问题

数字人文脉络中,将计量文体学(Stylometry)用于疑难文本的作者归属判定已成专门之学,历史悠久,成果足观。一般认为,1887年门登霍尔在《科学》杂志发表《创作的特征曲线》一文,用狄更

^① 郑永晓、段海蓉《古籍数字化、数字人文与古代文学研究——访中国社会科学院郑永晓教授》,《吉首大学学报》2020年第2期。

^② 受篇幅和个人能力所限,古籍数字化、数字文献和纯粹的数据库平台建设等工作取得的重要成就,不在本文讨论范围内。

斯、萨克雷等人词长频率的分布图来代表作者的“创作风格”，为一现代开端^①。这一做法为人们从文本内部找证据，以协助著作者判定奠定了方法基础。文体测量学在古代文学研究中较早见于《红楼梦》前八十回抄本和后四十回刻本作者统一性的考证，特别是后四十回作者归属这一公案上。该问题自胡适揭举“新红学”后便无定论，到20世纪50年代始有量化手段介入。从1952年高本汉的研究至今^②，已有四十余种量化分析的专论、专书出版。足见其影响之大，引发的兴趣之久，故而本文也将此问题专门列出讨论。

在早期，人们通常会比较前八十回与后四十回的功能词频率、构词法、句长、特殊词汇等等“作者指纹”在频率或分布上是否存有显著差异，继而推断它们是否为同一作者所撰。在现实中，由于不同研究纷纷选取了不同的版本、不同的语言特征、不同的抽样方法乃至分布假设——在缺乏新的有力“旁证”的情况下，便不免会得出相互抵牾的结论。如赵冈和陈钟毅于1970年检验了五个常见语助词在脂本前八十回和程甲本后四十回一百页文本中平均频次的差异，反驳了高本汉的测量方法和结论^③。又如，陈炳藻曾用计算机算出前、中、后三个四十回各八万字中几类词汇出现频率的相关情况，认定全书作者唯一^④。陈大康先生则认为这种抽样忽略了后四十回可能含有刻意仿写和部分残稿，且样本量较小，遂花一年半的时间“捺字点数”，分别在字、词、句的水平上，检验了八十八个特征在前后两部分出现频率的分布形态^⑤，最终不仅得出后四十回作者另有其人，还给出了可能含有部分残稿的章回区间。

近三十年随着文本挖掘、机器学习等多元统计技术的跟进，人们开始采用文本分类和聚类技术对接此类问题。复旦大学数学系李贤平教授是最早启用词汇的向量化表示和多维尺度分析对《红楼梦》文体风格做分类，从而对著作权进行划分的学者，也是在这一问题上引起较大争议的量化研究者^⑥。该研究以陈大康提供的四十七个虚字的频率作为特征集，计算单回文本向量间的卡方距离，降维后投射到散点图上，通过观察距离来确定回与回之间的相似性，相当于主成分分析的一个典型应用。这一做法在如今并不少见，作者确有开创之功，其做法本身并无太大问题，之所以会引发争议，更在于技术施用和结果解释的有效性上。也就是说，作者仅仅依据图的视觉效果，便认为不同的风格章回被算法分开，得出了《红楼梦》是由不同作者在不同时期撰写而成的“惊人之论”。实际上，如果按照此种文本处理方式，分离出来的却未必是作者笔法，很可能仅仅是某些字频组合的固定模式。如果改换切割文本的基本单位，也可能得到不同的离散效果，就像被其文本聚类所证实的一样。所以恰如陈大康的质疑，到底多大程度的离散，才能判定任意两回并非同一作者？“这判定本身的可靠性又究竟有多大呢？”可以说，李贤平研究最早暴露出人文计算的重点并不在于计算技术有多高超，而更在于方法要如何施用、怎么解释才有效，这其实也是一个不容忽视的阐释学问题^⑦。

有意思的是，陈大康先生还对造成各回聚散形态的可能原因做了推测，认为这仅仅是由于近代白话带来的文白夹杂所致^⑧。不论此说是否成立，都等于是从反复阅读的经验出发，对主成分究竟为何

① Mendenhall T C. “The Characteristic Curves of Composition”. *Science*, 1887, 9 (214): 237-249.

② B. Karlgren (高本汉). “New Excursions in Chinese Grammar”, *The Bulletin of the Museum of Far Eastern Antiquities*, vol. 24, 51-80, 1952.

③ 赵冈、陈钟毅《前八十回与后四十回用字之比较》，《红楼梦新探》，香港文艺书屋1970年版。

④ B. C. Chan. “Authorship of The Dream of the Red Chamber: A Computerized Statistical Study of Its Vocabulary”. The University of Wisconsin-Madison, Ph. D. Dissertation. 1980.

⑤ 陈大康《从数理语言学看后四十回的作者——与陈炳藻先生商榷》，《红楼梦学刊》1987年第1辑。

⑥ 李贤平《红楼梦成书新说》，《复旦学报》1987年第5期。

⑦ 对这一问题的讨论，参见拙文《作为计算批评的数字人文》，《中国文学批评》2022年第2期。

⑧ 陈大康《“〈红楼梦〉成书新说”难以成立》，《华东师范大学学报》1988年第1期。

做了某种深入算法层面的推究。事实上，李文在引入技术方法时的一个问题正在于并没有将计算过程解释清楚，导致后来者都如雾里看花，频繁引用其结果的居多，却少有人去推敲其技术之实，更不用说对做法和结论做出恰当评估了。

2010年，施健军教授训练了支持向量机分类器，高度拟合了预想中的分类结果，验证了前八十回与后四十回作者完全不同这一假设^①。单纯用机器学习构造二项分类器来寻找分界点的思路，建立在只有两类归属的预判上，似乎仍未能顾及对“判定标准”本身的检验。在缺乏公认的测评数据集的情况下，采用不同的特征变量和做法也可能得到不同的分类结果，这些结果之间无从比较。这类问题在四年后开展的一项同样利用支持向量机构造分类器的研究中得到了某种程度的改进^②。该研究选取程高本前六十回和后三十回为训练集，用中间三十回做测试时也观察到了第八十回这个明确的分界线；但在前八十回上测试时，却没有显示出可分性，由此便证明了前八十回、后四十回间的确存在明显的“文体差异”。值得注意的是，研究还在多次实验中引入对照组，既纳入了曹雪芹同时代作者秦子忱的续书，也包括其他时期的章回体小说，并经多次迭代从一百九十六个特征变量中筛选出相对有效的特征项。这些都在某种程度上缓解了所谓的“判断标准”问题。

近年来，人们采取了更科学、更多元的特征指标，如 n-gram、词长、词汇丰富度、句法语义特征（如刘颖和肖天久的一系列研究^③），在特征提取和特征集的建构上也越发具有向客观探索敞开的特点。在这方面，杜协昌的工作之所以可贵，不仅因他较早设计了字频差异函数，让算法自动算出真正具有区分意义的高频词簇^④，更在于其每步实验都建立在较严密的验证和推理上，计算和挖掘的过程清晰可见，因而还可用于直接指导细读互释。此后，叶雷的研究则沿袭李贤平的思路，又弥补其不足，取了全部字的字频做特征，把两种续书以及后四十回与前八十回向量同时投射到图上，经观察便可知两部分并没有聚在一起。其不足之处，则是仍缺乏与作者自身其他文本的比对。恰如研究者也认识到的，“需要了解同一个作者文体特征在矢量空间中的分布”，才能最终确定什么样的离散程度算是真正“分开了”^⑤。

事实上，如果我们去考察作者识别问题的经典案例，会发现其中绝大多数流程，是首先找到疑似作者的大量存世语料，用于生成特征向量，使用分类算法训练出识别模型后，再用于待鉴定文本的作者判定，是谓“用已知的文本作者来推定未知文本作者的研究”^⑥。这一思路正体现在统计学家 Mosteller 和 Wallace 极负盛名的《联邦党人文集》研究中：从七十三篇作者已知的文本中统计出汉密尔顿和麦迪逊二人的功能词频分布，按贝叶斯定理计算拟合后，将剩下的十二篇文章判定给了麦迪逊^⑦。然而遗憾的是，这种做法在很可能只有绝少作品传世的“作者”身上却很难施行^⑧。也正因此，有学者

① 施健军《基于支持向量机技术的〈红楼梦〉作者研究》，《红楼梦学刊》2011年第5辑。

② Hu, X., Wang, Y., & Wu, Q. “Multiple Authors Detection: A Quantitative Analysis of Dream of the Red Chamber”, *Advances in Adaptive Data Analysis*. 6. 04 (2014): 1450012.

③ 刘颖、肖天久《〈红楼梦〉计量风格学研究》，《红楼梦学刊》2014年第4辑；肖天久、刘颖《〈红楼梦〉词和N元文法分析》，《现代图书情报技术》2015年第4期。

④ 杜协昌《利用文本采矿探讨〈红楼梦〉的后四十回作者争议》，项洁编《数位人文研究与技艺》，台湾大学出版社2014年版。

⑤ 叶雷《基于计量文体特征聚类的〈红楼梦〉作者分析》，《红楼梦研究》2016年第5辑。

⑥ 张洋、江铭虎《作者识别研究综述》，《自动化学报》2021年第11期。

⑦ Mosteller, Frederick, and David L. Wallace. “Inference in an Authorship Problem”. *Journal of the American Statistical Association* 58, no. 302 (1963): 275–309.

⑧ 此问题也即陈大康先生所谓“选错了母本”。正确的做法应是曹雪芹自己的文体惯习（如词频）作为“母本”来与抽样文本进行比对。

认为像《红楼梦》这样的题目终将无解^①。《红楼梦》的著作权问题，在未有新的曹雪芹佚文被发现的情况下，或许只能成为一种验证性研究，其关键更在于如何科学推理。

近一两年来，也有了将深度神经网络较规范地运用于此案的建模和推理的尝试。如以词嵌入和LSTM神经网络结合，通过反复测试模型的泛化性能，得出在不同训练分界点处前后差异显著性的大小，以确定文本风格的实际分界点^②。这类设计的亮点还在于，以“脂评本+程高本后四十回”的一百二十回拼接版作为对照组，发现了对照样本在八十回处前后的差别相对于程高本（一百二十回）来说更加明显，从而在一定程度上再次印证了“程高本对前80回进行了一定程度的改动”的观点。其不利之处仅在于，采用无监督的深度学习固然极大地促成了推理，却很难通过明确的特征项回溯到细读互释中去，对于进一步的文学批评和文学史研究来说也就意义不大了。

纵观七十年来《红楼梦》作者归属权的研究历程，汉学家和中国学者们启用量化方法的历史并不算短。只不过相对于作者唯一的案例，像《红楼梦》这样成书过程极端复杂的作品，仍然呼唤研究模式和方法的创新。无监督的自动识别方法离不开精巧可靠的实证设计，需要在充分论证方法施用有效性的前提下，再来验证已有假设。这也使得一定阶段内，文体测量学在《红楼梦》这类问题上多会以推理和验证为主导的校勘批评（textual criticism）的面目出现。一个例证是：在前述诸多实验中，尽管研究者们所使用的版本、“繁/简”系不尽相同，其结果却都指向了六十七回作者另有其人，或为后四十回作者在残稿基础上“补做”这一推测，这便和传统红学在六十七回问题上疑议迭出的情况正相吻合^③，坐实了“原本”在传抄、过录过程中发生的诸多变异情形。实际上，明清文本的生成形态错综复杂，抛开版本谈作者问题并无太大意义。文体测量在异文推定，版本溯源和辨伪，乃至作者探佚等方面还大有可为，有待于融入更加深入、细致、全面的文献工作中，需要将文本之外的情况也考虑进来。如李友仁在探索兰陵笑笑生的真实身份时，就不仅只依靠文体测量，还对万历间《金瓶梅词话》问世之前的手稿在精英社会网络中的流通状况专门建模，为归属判定提供了另一条线索。

也许还应引起讨论的是，文体测量学的出发点说到底还在于，须承认某种特征组合模式在同一作者身上具有绝对稳定性。这种稳定性是跨越文类、语体、年代、心境，或者其他偶然因素而存在的，正所谓“文如其人”，也等于是否定了“巨奸为忧国语，热中人作冰雪文”的“饰伪”现象。然而是否真有这样一种挥之不去、深入骨髓的语言“格调”，可以为模型所捕捉、构建出来呢？退一步说，算法识别出来的模式，是否就是读者所辨认出的那个“调”呢？这里当然还牵涉一些难以回避的文体—风格的基本理论问题，关系着创作个性、语体风格、体裁规范和阅读惯习等不同层面上的风格稳定性和同一性，能否将其考虑进来，关乎这一类“实证”研究的信度。尽管李贤平先生曾认为这更是一个“实践问题”，然而国外同行的经验告诉我们，在中文里，当人们选取特征时的高频词临界值大于某个值时，体裁对风格的影响常会模糊作者身份信号。这似乎透露出基本理论也关系着整个实验的分层架构，仍然是不可回避的。

2. 体裁与目录学研究：基于聚类、分类的文体计量批评

聚类、分类任务是基于统计的自然语言处理的一大类重要应用，近十年来在目录学和文学文类、

① 黄水清《人文计算与数字人文：概念、范式及关键环节》，《图书馆建设》2019年第5期。

② 朱东旭、严广乐《基于LSTM的〈红楼梦〉文本风格分界点识别方法》，《智能计算机与应用》2020年第10期。

③ 李贤平所用第六十七回为刻印程甲本。施健军的前八十回用庚辰本，第六十七回由程甲本补配。杜协昌的前八十回以脂评本体系为底本，后四十回以程甲、程乙本为底本，更接近于程乙本。第六十七回采用戚序、戚宁、列藏、甲辰本，即所谓“繁系”。同样，叶雷也取蔡校本数据。上述个案几乎都判定第六十七回为异文。关于“繁/简系”之说，参见曹雪芹著，蔡义江校注《红楼梦》，浙江文艺出版社1993年版，第904页。

文体风格研究中已显示出一定潜力，与作者归属判断有异曲同工之妙，仍然可以算作文体计量分析的一部分。在实践中，人们可以使用基于各种分类、聚类算法的模型对文献和文本做分类，观察分类结果与传统观点的差异，弄清算法究竟是基于什么做出了分类，继而对历史观念的生成展开计算批评和分析。这一思路为大数据背景下的文学文献研究提供了新可能，也不再是简单的工具思维，而是带有了的反思性。

专治明清文学的青年汉学家李友仁曾经做过一个大胆的实验，他想通过无监督的聚类算法，来看看在概念上边界模糊的明清“野史”与“正史”“小说”的关系到底怎样^①。与作者归属判定的做法类似，他选取了高频字的频次来做文本的向量化表示，在不同规模的语料库中分别做聚类和分类测试，降维后的主成分分析显示，不仅小说和正史被分开，野史最终与演义一起，填补了它们之间的空间。可以说相对于小说，野史具有更自由、更不稳定的位置，符合其复杂的文体性质和文化定位。这个研究发表后，招致圣母大学英语系筭章难的批评。在她看来，切割万字片段的预处理步骤人为增多了数据点，有制造“分开”效果之嫌；不仅如此，她还认为之所以有此效果，更是由于内容不同，就像小说会和野史接近，是因为它们都来自于“道听途说”。然而，真的是这样吗？

事实上，研究者在这一系列实验中关注的并非不同标签下的文本可以被分开，而在于它们被什么分开。将文本等值均分这一扩大文体信号的做法，是为了便于察觉区分因素究竟为何。重要的是如何调动先验知识，对主成分轴究竟为何做出合理的解释。而细查载荷图后不难发现，各种语料的组合无不显示了白话/文言向度上的差异，其中最重要的二十个特征词恰好遵循了从正载荷到负载荷的分布规律。鉴于这一证据，作者认为：历史修辞的结构影响，仍然是通过语言的正式性（语域），而非所谓“历史内容”来施加的，这一批评性的结论意味深长。

从上述个案中或许不难发现，各种聚类、分类的算法虽然是一定的，但用到实际文本数据中的分类标准和依据却并非一成不变，这部分地取决于文本表示模型的选择和数据预处理的方式，也是一个人为调参的结果，归根结底，仍然由人文学者想要论证和阐释的具体问题决定。因而相较于套用模型得到一个貌似客观的计算结果，弄清楚算法究竟基于什么做出了分类，才是人们更关心的。在最近的一项研究中，几位学者则索性以此为契机构建了他们的批评框架。在这篇题为《〈汉书·艺文志〉目录分类再审视》的文章中，为了反思《汉志》的“分类标准”问题，团队运用基于神经词向量的模型和无监督算法，对《汉书·艺文志》中七十八部存世文献做了六十余次模拟分类实验，希望用这种摆脱了“《汉志》本身的描述”的自动分类技术，来考镜早期文献蕴蓄的多种分类可能^②。有趣的是，自动聚类结果不仅显示四分至六分的方式较为稳定，而且在他们看来，也展示出了依据言说方式（“四部分类法”）和文本内容（“义”）的两种不同的分类原则。研究者遂借对分类标准和有效性的讨论，进一步重审了后世目录学从《七略》的“六分法”到四部“四分法”这条人为规划的“演化路线”的内在理路。

值得一提的是，该研究的文本表示建立在前期近七亿字《四库全书》训练出的一个三百维的古代汉语神经词向量的模型上，可以为任意古汉语文本中的单字在高维语义空间中定位，继而以取平均的办法为小数据集中的每个文本定位，最后以各种距离算法来计算文本相似度。可以看到，在今天，以语言的向量化表示和深度学习为基础的大规模预训练模型已经可以直接用于聚类、分类这样的下游任务，这或许预示了未来数字人文取向的巨量文献研究的一种趋势。但同时也应明白，数字人文在这类实践中引入模型的价值恰恰产生于研究者“主观介入”的环节。这是因为，大模型固然让结果更趋精准，其内部的每一个特征维度究竟是什么却变得不可解释。以上下文中的词共现为基础的表达模型，

① 李友仁著，汪蘅译《虚构与历史：明清文学中的对立与风格梯度》，《数字人文》2020年第2期。

② 诸雨辰、李绅《〈汉书·艺文志〉目录分类再审视》，《数字人文》2021年第3期。

使聚类结果本来就既有可能是基于语义也有可能是基于形式的，只能通过人为调节参数，观察结果变化，反推其机制。经过不断调整 k 值并细察结果，研究者对分几类和分类依据的关系，做出了并非客观的推测，这种推测成为其进一步展开批评研究的基础。总之，该研究可视为初具计算批评雏形的一种思路，未来人们或许可设计出更精巧的实验，在努力破解算法黑箱的同时，对有价值的历史问题给予新解释，在这个方向上，还有诸多令人期待的结合点。

二 诗歌研究

1. 声律与节奏的统计分析

声律体格问题古已有之，大概是众多形式问题中技术性最强的一宗。从传统音韵学、诗体学角度研究声律问题的成果，至 20 世纪已蔚为大观。20 世纪 90 年代后，经邝健行、何伟棠、杜晓勤等时贤努力，统计数字成为重要证据被纳入史论中来。早在 1993—1994 年间，杜晓勤先生即采取量化方法，对五言诗的声律实践做分类归纳，统计出齐永明至初唐各种“联间组合”的出现率，尤其以“粘式律”的比率变化为重点，来衡量各个时期律化成就的高低，再对读时人和后人观念，不仅对沈约、赵执信等人的永明声律说有所纠偏，也勾画出了总体演进趋势^①。1997 年以来，杜晓勤、尹小林两位先生还合作开发“中国古代诗歌声律分析系统”^②，使用计算机软件对古诗的用韵方式、五言诗的“四声八病”情况进行大批量的自动标识和数据统计，为研究先秦两汉用韵特点和韵部变化，考察永明诗歌声病情况，近体诗律形成与定型，提供了最基础的数据资源，可谓功莫大焉。

近年来，充分利用大样本和自动标注算法，以事实数据来验证前人声律学说成为可行的方向。为考察拗救现象，诸雨辰、胡韧奋曾以卷帙浩繁的《全唐诗》为底本，专门对律诗中的 13683 个拗句的平仄规律及其分布情况做详细的统计分析^③。选取出频次大于百分之一的单、双拗句，再统计“拗而不救”的发生频次，将后者与前两者相减后再与王士禛、赵执信的拗救说相参，便可验证真伪，或对其存废的标准做讨论。可以说是经统计发现了真正有典型意义的拗救句式，重新提炼出唐诗平仄的基本原则，是以现象叩诸事理，从后验概率来实事求是地阐明问题的做法，立论扎实，思路清晰。

循此思路，清华大学中国古典文献研究中心近年来和统计学中心合作，采用自动标引与人工校补结合的方式，对《全唐诗》进行精确注音，对现存五七言四八句体诗（共 35090 首唐诗）展开细粒度的统计。“对标准句型、二四异声、粘对、三字尾、换头、护腰、相承、挟声、拗救等声律特征进行校验，证实或证伪一些流传已久的命题，并发现一些曾被认为非律的句联篇模式出现几率较高。”^④这种做法其实是取更开放的态度来考察“格律”观念的建构，即在抽样时尽力摒除明清和今人格律标准的先验性，而将对象放宽到“全部五七言四八句”，追求所谓“最全”的数据样本，来获得真实的创作情形。其研究观念已经从狭隘的经典观转向了更大的文学生产场。

通观量化声律研究的进展，统计手段为其注入科学性，“大数据”的视角带来了观念的更新，让人们认识到“唐诗格律更像是一个动态平衡系统”（《唐诗格律的统计分析及其问题》），是一个历史生成过程，而并非靠一时之规就可以框定的，这是脚踏实地得出的结论。同时也应看到，在今天，人们也

① 杜晓勤《齐梁诗歌向盛唐诗歌的嬗变》，北京大学出版社 2009 年版。

② 杜晓勤《“中国古典诗歌声律系统”的研发过程和学术价值》，《石河子大学学报》2016 年第 4 期。

③ 诸雨辰、胡韧奋《清人“拗救说”再审视——以〈全唐诗〉15290 首律诗为样本》，《中国诗歌研究》第 18 辑，社会科学文献出版社 2019 年版。

④ 李飞跃《唐诗格律的统计分析及其问题》，《文学遗产》2022 年第 5 期。

许可以根据统计结果,轻易推翻王力先生和启功先生总结的规则,但如何根据大样本形成新的认识,而非满足于所谓“数据驱动”带来的冗杂结果,却仍非易事。在这方面,还须具备一定的抽象建模能力、足够的远距离视野,把诸多不确定因素本身也纳入模型中来,以便发现整体上的趋势。

在声律建模方面,一个较富启发性的例子来自波尔多蒙田大学廖学盈教授的《诗经》研究^①。为了挖掘兴体诗潜藏的节奏模型,作者以一种频率—位置组合序列分析算法辨识,抽绎出兴体诗的起句和应句之间的联结模式,继而将音韵、修辞和主题融为一身,对其“修辞节律”做量化探源。从实际意义说,作者相信通过结果和传统注疏的对勘,可以进一步考察文体被建构的过程,也能重新审视过去的诠释。而颇有意味的是,这种序列分析模型最后还沉淀为一种释读方法,即由句组间对位所形成的节奏复沓。这种释读似乎离析出了更恒久动人的抒情结构,正是经过这一系列建模操作上的现象学还原过程,近乎完美地诠释了“兴”是如何在“涵泳讽诵”的释读传统中成为可能的。

2. 诗歌的词汇语象和影响研究

近十年来,计算语言学和文本挖掘的一些基本思想方法,也被用于古诗的语言、意象、风格和影响研究,而尤其是对《全唐诗》《全宋诗》等语言体式的测量。这类研究的结论虽然多从微处着墨,但整体视野和气象已不同从前,而且还在向更大的样本拓展。

台湾政治大学刘昭麟教授团队于2015年发表《〈全唐诗〉的分析、探勘与应用——风格、对仗、社会网络与对联》^②,较早利用词频统计和信息检索等手段对《全唐诗》做全景式扫描。此后,考察对象被扩展为周至明两千多年间的九部代表性诗歌总集,共290241首诗^③,对诗歌发展做了词汇水平上的长时段俯瞰。这类研究将逐词索引(Concordance)、词语搭配(Collocation)等传统的信息检索功能发挥到极致,用于比较某些特定词汇语象(如“白日”)在不同诗人、文集或诗体间的使用频率、出现位置、搭配模式(如“白日”与“青山”对仗)、分布规律,及其在不同年代间的传承和演变关系。如果和人物传记数据结合起来,还可将词汇史的远读和具体诗人挂钩,例如可研究盛唐或南宋时期的诗人们如何使用“红颜”一词,或者发现“惺忪”一词很可能是晏殊在宋代的首创。同一时期,郑文惠教授等则主要借助词语搭配功能,深入探究了中唐诗人“白”色语象的构词特点。从14566首诗中撷取以“白”为前后缀的词汇,找到与这些种子词构成对仗或搭配关系的词,链结出有主题意义的关系网并探究其背后的文化因由,据此归纳出了中唐诗人的“白色抒情系谱”^④。与此相类,邱伟云、严程还曾研究过中唐诗人所偏好的空间方位词汇,以方位词前词缀与字频计算,凸显了“人间”主题在中唐的涌现,从另一个侧面探讨了唐人气象之结穴所在^⑤。

可以看到,仅仅求助于一两个检索功能,便可探珠渊海、手到擒来,将有价值的信息网罗殆尽,为进一步精研细读、考镜源流服务。搜索算法的另一大功用,还体现在文本比对和对勘上。二十年前,即有北京大学李铎先生主持的“全唐诗分析系统”和“全宋诗分析系统”问世,可按格律和重出诗等功能检索,还可以按字及词组的频率分布做统计。而真正将检索落到文学研究实处的,是刘

① 廖学盈《〈诗经〉的量化研究:挖掘兴体诗的隐藏节奏》,《数位典藏与数位人文》第4期,2019年。

② 刘昭麟等《〈全唐诗〉的分析、探勘与应用——风格、对仗、社会网络与对联》,The 2015 Conference on Computational Linguistics and Speech Processing, ROCLING 2015, pp. 43–57。

③ 刘昭麟、余泰明、康森杰著,潘亦迎译《用数字工具探索中国古典诗歌:语言学、文学及历史视角之例证》,《数字人文》2021年第2期。

④ 郑文惠等《情感现象学与色彩政治学:中唐诗歌白色抒情系谱的数位人文研究》,项洁编《数位人文:在过去、现在和未来之间》,台湾大学出版中心2016年版。

⑤ 邱伟云、严程《心寄乐园,凝望人间:中唐诗空间方位的数字人文研究》,《西南民族大学学报》2020年第8期。

京臣十多年来就唐诗对宋词的影响展开的全面研究^①。该研究结合诗词的集校注本、词话和历代汇评,计算每位作者的“影响因子”和“接受因子”,特拈出十大影响源,用文本比对软件查找十大诗人诗作和宋词之间的全部重出、相似文本,对其间有意味的“证据”做穷尽式释读,着重分析的是词人对诗人用语、意象、题材和意境的广泛取法、化用和变异。作者对痕迹的捕捉一丝不苟,条分缕析,又并不泥于此间细小征象,而是务必贯通其流脉,考问来由,在此基础上讨论影响发生的“规律”和“范式”,或牵连、回应一些更有意思的话题,展示出一种绵延不绝的互文性,也显出极强的说服力。

不过,此类研究仍然基于字符串匹配和检索,考察的是字句层面的“偷语”现象。诚如研究者认识到的,如果想要将更难捉摸的“偷意”和“偷势”也一并勘探出来,则需要借助更高阶的深层语义检索和分析功能(《晚唐诗对宋词影响研究》,第340页)。目前,这类功能也被大平台开放,如清华大学自然语言处理与社会计算实验室的“九歌类义句搜奇”,基于预训练模型 BERT-CCPoe,改进了句向量的相似度算法,即便没有任何一个字词相同,也可以将语义和意境相似的文句查询、提取出来。只不过,此类工具如何用于具体研究还有待进一步开拓,或许至少如郑永晓先生所言,可用于特定意象词群的抽取和异同分析,统计它们被历代作家使用频次多寡,再用于更别致的研究主题^②。

三 文学史研究的“系地”化:文学地理学与平台建设,空间计算、现地研究

古代文学研究历来重视文学地理学的分支维度。近年的数字人文大发展中,文学地理学也是率先提出实证化和技术化要求的一个领域。早在2006年,郑永晓先生就曾一口气提出GIS融入古典文学研究的十个构想^③。今天看来,这些应用已成为人们最关注的几个面向,在文献整理和文学史研究中越来越集中地体现为一种“系地”的努力^④,渐渐带上了较明显的数字人文色彩。在历史地理信息化的驱动下,一些文学地图平台逐步建成并向学界开放,旨在突破既往作家年谱和文学史研究注重时间编年而不注重空间系地的局限,有助于对历代文人的行旅、写作和相关文化现象做远读式研究^⑤。

在此,如果以“文学版图”“文人群体的地理分布”“文化中心的变迁”这一类典型题目的进展为切片,或可管窥一些变化趋势。早先的文人地域分布研究多以籍贯可考的占籍数据为本。昔日有唐圭璋先生《两宋词人占籍考》发轫于前;20世纪90年代初陈尚君先生为复旦史地所《中国文化地图集》作《唐诗人占籍考》;同期亦有曾大兴先生《中国历代文学家之地理分布》,对西周至清的六千二百九十三位作家分布做初步统计和成因详解;新世纪后,刘跃进先生以《汉书·地理志》中一百零三郡国为统计单位,对六百二十余位秦汉文人做明确的“系地”工作。早期的钩沉整理繁难艰巨,大都以史传或方志、工具书中的籍贯信息为据,以文化中心或南北分布格局的变化及因由为贯穿始终的关注点,尚未融入GIS信息,也没有实现基于地图的可视化操作论证。

若单就唐宋元的情况而言,2016年王小红采集《全宋文》中的传记资料,较早使用QGIS软件绘

① 刘京臣《盛唐中唐诗对宋词影响研究》,中国社会科学出版社2014年版 《晚唐诗对宋词影响研究》,中国社会科学出版社2021年版。

② 《古籍数字化、数字人文与古代文学研究——访中国社会科学院郑永晓教授》。

③ 郑永晓《以GIS为例看信息技术在古典文学研究中的应用》,《重庆教育学院学报》2006年第5期。

④ 系地一说,自20世纪90年代起渐为学界所重视,后为王兆鹏教授主倡。参见曾大兴《中国历代文学家之地理分布》,湖北教育出版社1995年版,第1页;王兆鹏、蒋晓晓《时空一体化——唐宋文学编年地图平台的学术理念与学术价值》,《三峡论坛》2020年第5期。

⑤ 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。

制出北宋作者的地理分布密度图^①，与刘尊明、王兆鹏多年前的计量结果相呼应^②，认为南方文人的数量在宋代是远超北方文人的。同年王兆鹏先生以《全唐诗》《全宋诗》等资料中籍贯地为据，贯通唐宋，发现诗坛的重心亦发生了明显的南移，这种转移始于晚唐五代，完成于北宋，而并非前人所认为的“始于安史之乱，成于靖康之役”^③。此后，复从更全面的资料中提取、考订信息，揭示出“孤篇存世”的作者占一半左右，浙江、福建、江西作者占“天下之半”这样一种“分层分布”的结构状况^④。同一时期，徐永明教授团队同样重视籍贯数据和地理信息相结合。早先曾依《全元文》《全元诗》中籍贯信息，对元代文学的作者分布做可视化分析，从文人教育背景与心态、理学思想的渗透，以及诗歌传统的生成等角度对诗、文作者在地域分布上的精微差异做出解释^⑤。近来又加入了宋代数据，宋元合观，力图描述经历蒙宋战争后的文学版图又发生了怎样的变化。

与上述仅以籍贯分布这一“静态数据”为指标不同，王兆鹏团队曾借助唐宋文学编年系地信息平台，用一条条人工标注的、描述作家活动的编年系地数据，将作家一生途经或寓居某地的频次也统计进来。据此可以从“北方占籍诗人虽然多于南方，但活动人次却少于南方”的现象中，再次得出唐代诗人大多是向南流动的观点，且大都是向江西、两湖、两广和重庆移动^⑥。

近年来空间技术与文化地理学渐相汇通，为人们从量化建模的角度深入此类问题提供了新思路。2014年发表于《科学》杂志的研究用过去两千年间欧洲和北美十五万文化名人的生卒地作节点，以动态网络的形式再现了西方文化中心的盛衰演变^⑦。2020年，武汉大学地理信息团队与中南民族大学文传学院合作，在十一万条文人足迹数据上构建了唐宋文人的大规模迁徙网络^⑧，在不同时期的网络图中让全部城市按PageRank值显示大小，直观地印证了先前关于文化中心的大量结论。他们还专门对南北城市的吸引力差异变化给出时间函数的精确建模。观察曲线可以清晰看到，安史之乱对于吸引力中心南移的影响其实是“暂时性”的，战乱时北方吸引力急剧下降，战乱之后又恢复到北方强于南方的状态，至靖康之乱时，文化的吸引力南移“已经完成”。经这一模型评估，研究者再次强化了文化中心南移完成于北宋的认识，修正了此前的“三次历史事件驱动说”。可以说推进至此，该研究已经从多学科交叉的角度为这一课题画上阶段性的句号，颇有一锤定音的效果。

回顾十多年来的探索历程，几个团队围绕所谓文化南移说的相关假设，从简单的描述统计，到引入移动数据和地理信息、空间人文的技术和视角，背后的关切不仅仅是一个说法，还有对地理中心与文化中心、文人迁徙和文学中心、文人分布与作品分布等多种变量之间内在关系的持续兴趣^⑨。随着

① 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》，《史学月刊》编辑部编《大数据时代的史料与史学》，人民出版社2017年版。

② 刘尊明、王兆鹏《唐宋词的定量分析》，北京大学出版社2012年版。

③ 王兆鹏《唐宋诗歌版图的空间分布与位移》，《中国人民大学学报》2016年第6期。

④ 王兆鹏、齐晓玉《宋代诗文人作者的层级与时空分布》，《中南民族大学学报》2022年第1期。

⑤ 徐永明、唐云芝《〈全元诗〉作者地理分布的可视化分析》，《浙江大学学报》2019年第1期。

⑥ 王兆鹏、王艳《唐代诗歌版图的静态分布与动态变化——基于〈唐代文学编年系地信息平台〉的数据分析》，《中南大学学报》2020年第1期。

⑦ Schich, Maximilian, Chaoming, Song, Yong-Yeol, & Ahn, et al. (2014). "A Network Framework of Cultural History". *Science*, 345 (6196), 558-562.

⑧ 应申、侯景洋、周钰笛、窦小影、王兆鹏、邵大为《基于唐宋文人足迹集聚性分析的中心文化城市变迁》，《地球信息科学学报》2020年第5期。

⑨ 明代的情况参见李菁、徐永明《明代作家的空间分布与可视化分析——以〈中国文学家大辞典·明代卷〉为例》，《浙江社会科学》2022年第2期；Lixiang Qian. Distribution Maps of Chinese Poets in the Ming Dynasty (1368-1644): A Geographical Visualization Experiment. *Library Trends*, Vol. 69, No. 1, 2020.

元数据类型和数据量不断扩充,观测距离拉得越来越远,同样的题目在不同的数据基础、不同的统计建模和论证角度下有了一做再做的价值。这也使得人文数据和经验的平台化愈发成为一种现实需要。

从数字人文实践看,平台既是研究成果的载体,也是数据的管理、加工、再生产、对外展示和提供服务的工具。地理信息平台有助于整合史料文献并将其转化为空间数据,“以厘清历史时空变迁过程中的各种因果关连”^①。通过结构化数据表的多图层嵌套,可以实现 GIS 或 WebGIS 的初步架构,是建设历史和文学地理信息平台的关键。中南民族大学和搜韵网合作开发的“唐宋文学编年地图平台”^②,浙江大学和哈佛大学合作的“学术地图发布平台”^③都在此架构上建成。以后者为例,除国家认可的“天地图”外,其底图主要来源于各种历史地图集,并借助更大的地理信息系统(如 CHGIS),嵌入地名字典、各级行政区划沿革的历史地理数据。与此相辅相成的则是从年谱、传记、方志等文献和研究中抽取信息,打上 GPS 坐标后,制作成点线面结合的矢量图层,便于编辑要素,获得地理属性,制作“行迹”并上传分享。如参照《汤显祖年谱》,可以绘制包含生卒地、任职地、创作地等四十三个地点、一百四十八项生平事迹的轨迹图层。若调用该图层则不仅可查看汤显祖行迹,还可通过地名索引查询其在某地活动的时间及事迹。各种人物的分布、行迹图层也可嵌套在不同时期、不同内容的底图上,通过时空对比来剖析某一事象的区域演变,或发现所关心的变量(如进士分布和文人聚集)之间的关系^④。

此外,台湾“中山大学”简锦松团队最近也推出了“明清水陆路程与文学”全开放平台^⑤,为其多年主张的“现地研究”之副产品。即今人可以利用现代科技工具,随文献而进入古人的时空现场寻找遗迹;而古人的方志、文集、山河、驿道等,也可以作为时人亲身实见的证据被“邀请到研究中来”,达成一种“山川为证”的效果^⑥。如以稼轩词与多种史料相参互证,结合实地考察、GPS 测量和制图,对词人罢官回乡时宁选迂回多艰的长江余干线而不走江南运河的原因一探究竟^⑦,花大力气考证信州上饶、铅山两县山水的古今之变,以推测其“避开临安,不入行在”的心理,再来对读两首晚年词作中的怨恨心境,便别有一番体会在心头了。此外,还有一些个案也吸收了现地研究之精神,唐宸利用天文软件对李商隐用典、陈子昂作品系年的考证^⑧,刘京臣对几部行录笔记的再研读,都有精湛的体现^⑨。

在空间人文这个交叉方向上,一些团队还致力于深挖小说文本数据,用空间技术来诠释文本中的空间文化。哈尔滨工业大学建筑学院何捷团队近年便从空间大数据分析入手,对唐传奇人物的出行信息编码,用虚构人物的“移动数据”构建社会网络,进行文学制图,试图从网络形态的变化中看出有唐一代长安人“感知结构”的变迁^⑩。这些研究也从全新的角度激起人们对“小说证史”这一类问题的再思考,做法本身更耐人寻味。

① 陈刚《“数字人文”与历史地理信息化研究》,《南京社会科学》2014年第3期。

② <https://sou-yun.cn/PoetLifeMap.aspx>

③ <http://amap.zju.edu.cn>

④ 这两个团队也正在建设更高级的知识图谱平台“知识图谱”(<https://cnkgraph.com>) 和“智慧古籍平台”(<https://csab.zju.edu.cn>)。

⑤ <https://www.dhlib.cn/mqrl>

⑥ 简锦松《山川为证:东亚古典文学现地研究举隅》,台湾大学出版中心2021年版。

⑦ 简锦松《现地研究与辛弃疾词的新读法》,《数字人文》2020年第1期。

⑧ 唐宸《天象模拟在古代文学研究中的运用——以Stellarium软件为例》,《数字人文》2020年第1期。

⑨ 刘京臣《大数据视阈中的文学地理学研究——以〈入蜀记〉、〈北行日记〉等行路笔记为中心》,《文学评论》2017年第1期。

⑩ 马昭仪、何捷、刘帅帅《中国古典叙事文学的时空叙事数字模型研究——以〈李娃传〉为例》,《地球信息科学学报》2020年第5期《从唐小说中的空间交互看都城长安的社会感知变迁》,《数字人文》2022年第1期,即出。

四 社会网络与文本网络研究

网络分析以网络理论为依托,是专门为关系型数据发展出的一套计算分析方法,将人、事、组织等实体之间的关系看成“节点”和“边”组成的网络,通过各种统计指标的计算,发现网络中的关键节点和网络的拓扑结构特征,形成关于社会历史结构或文本结构的新认识。网络分析在文学研究中可大致分为社会网络分析和文本网络分析。

社会网络分析最常见于文学世家、文学社群和文学活动研究。这一做法较早的样板来自 CBDB 这类大型人物传记资料库所支持的群体传记学研究,可以把历代精英的亲属、任职、师承、通信、著述等关系信息直接导入网络分析软件,呈现不同时期士人群体的核心人物、派系划分等情况。徐永明就曾结合年谱和 CHGIS 中的系地信息,将汤显祖的社会关系准确投射到明代地图上^①。刘京臣为了发掘明清进士家族中的文学家族,爬梳中华寻根网中的家谱目录来弥补 CBDB 关系表示上的缺陷,将家族成员的亲属关系重新呈现在世系图中。调用登科录数据后,便可从多角度验证“一经传家”的复杂问题。而如果引入家集诗词文赋中的唱和活动,还可进一步印证家族群体之间的交游关系^②。这一思路深得社会网络思想之精髓,其本质在于把人放在多重社会身份的网络中考察。当前数据库属性变量不完备,编码粒度不够细,反而促使研究者发挥能动性,把论证不断推向了精细严谨的境地。同样是唱和关系,严程从清代女词人顾太清与其友人的著作编年中提取关系人,建构酬唱往来网络,发现了女性诗人交游群体在“秋红吟社”存续期间的两次结构变化,“破解”了它的“中断之谜”^③。这些研究虽然还未诉诸更具体的量化计算,但网络样态自身已作为关键的可视化证据而得到使用,在一定程度上增强了论说力。相类似的运用还可见于邱诗文对桐城派文人群体的可视化分析^④。

相对于真实的社会关系网络,在文本网络分析中,各项结构指标对应了何种意涵则由任意两个节点间连线的编码方式来决定。诸如人物的互动频率、对话强度、词语共现次数、地点之间的出发和到达关系等等,皆可形成关系网。Jack Chen 等人较早运用网络分析研究了《世说新语》中人物的互动关系,观察到这部轶事集中的人物按时间顺序聚集的趋势,而不同的聚类又分明代表了不同的主题^⑤。此后,同样的方法被秦颖更系统地运用于《唐语林》研究中^⑥。结合“点击度分析”,她进一步提出了影响对话网络全局的“玄宗因素”及其在后半部分发生的“反转”。作者认为下一步可考察这一特征是如何反映中晚唐轶事材料对玄宗时期的关注的。也就是说,经网络分析得出的模型,为后续更大尺度的问题提供了一个通用的、可检验的度量指标,体现的也是一种模型检验的思维。不难看出,此类虚构性文本的网络建模,关心的更是深层的叙述意图和写作观念,与目前看到的大量定义简单,或仅仅基于词共现的关系提取方式不同,往往会寻求更有意义的关系界定方法。在这方面,廖俊凡在 2010 年便曾开发出一种描述会话角色关系的算法,用于《儒林外史》的对话网络建构^⑦。通过计算来发现

① 徐永明《中国古典文学研究的几种可视化途径——以汤显祖研究为例》,《浙江大学学报》2018 年第 2 期。

② 刘京臣《大数据视域中的明清进士家族研究——以 CBDB、中华寻根网为例》,《北京大学学报》2019 年第 4 期。

③ 严程《顾太清交游网络分析视野下“秋红吟社”变迁考》,《山东社会科学》2018 年第 7 期。

④ 邱诗文《可视性社会关系网络辅助文学流派界定方法探析——以桐城派文人群体为例》,项洁编《数位典藏与数位人文》第 7 期,2021 年。

⑤ Jack W. Chen, Zoe Borovsky, Yoh Kawano & Ryan Chen. “The Shishuo xinyu as Data Visualization”, *Early Medieval China*, 2014: 20, 23–59.

⑥ 秦颖《〈唐语林〉中对话网络的可视化和统计分析初探》,《数字人文》2022 年第 1 期,即出。

⑦ 廖俊凡《中国古典白话小说中的社会网络分析:以〈儒林外史〉为例》,台湾大学 2010 年硕士论文。

叙述中的关键人物及其分布形态,探讨不同指标所表征的人物功能,还可以用于近代小说人物体系和人物观的演化研究,这一做法其实有更深远的理论来源,也确乎为一条可行之路^①。

近年在深度学习和知识图谱的热潮中,关系提取已经发展为自然语言处理中的常规任务,很多人工智能实验室都推出了基于大模型的批量提取和可视化工具。但对文学研究来说,能否接受这一类现成的建模方式,还取决于它是否提供了研究者所需要的解释,可否施以意义明确的计算。从根本上说,这需要人们相信一套最基本的语义分布假说,包括它的认识论基础和知识表征的方式。也许,在很长一个时段内,前述基于特征的建模方式将会是文学研究所青睐的方向。

余 论

把数字人文脉络中量化方法的发展置于近一二十年来的古代文学研究中观察,会发现无论是作者归属判断、诗歌研究、文学地理学这些投入力量较多,眉目也渐清晰的领域;还是网络分析、空间人文、计量批评等崭露头角的新方向,都取得了极大进展,其成就是有目共睹的。这是一个几代学人艰难求索的历程,无疑令人钦佩。与此同时,放眼当下大量自由探索的个案,在量化结果的解释,可视化证据的信度、效度,实验设计和论证的科学性,以及问题意识等方面还存有诸多可商榷之处。这不仅与研究者的统计水平和社会科学素养、古典文学的造诣有关,数字人文自身一直面临的来自知识基础、方法论和评价标准等方面的挑战也不容忽视。这便要求研究者不断摸索有效的结合路径,突破重围。

首先是所有量化方法都须恰到好处地融入具体的论证框架中。近年来,越来越多的工具被摆在前台,颇有乱花渐欲迷人眼之势,但是有效的施用仍然稀有,为之增色的更难得一见。问题的关键在于,施用者是否真正了解了工具?文学建模作为观念迹化和操作化的形式,是推理论证的一种新手段,本来是相当灵活自由的。如果没有能力自己建模,人文学者能否理解各种现成工具的内置算法,能否贴合具体问题来有针对性地设计实验、选取文本表示模型和计算模型,然后才谈得上对结果予以合理的解释。正如本文中列举的大多数个案,这个过程每一步都不可割裂,都有其现实依据,这是一个运用之妙、存乎一心的过程。

原则上讲,人文研究并不允许过程的黑箱存在。在不得已的情况下,我们需要反复地调节输入、观察输出来推测模型的分类机制,需要在不同样本的水平上完成对模型自身的检验。这种用文本(数据)来对话算法,在人机交互中破解算法黑箱的操作,是为了在较细的粒度上完成文学理解和阐释。否则的话,单纯工具性地引入高科技,不仅会产生大量问题,在方法论的层面也几无新意可言。

与此相关的一个重要前提是,应认清数字人文研究的阐释本质,实事求是地看待“大数据”“人工智能”“大模型”的效力。技术正在飞速的迭代升级是事实,然而并非越高级、越“智能”的技术就是不二之选。正如网络分析可被视为结构主义和文学社会学的理论工具,技术也是人们认识世界的理论透镜。在文体研究中,文本表示的方式已然决定了所要研究的问题。数字人文的文学研究,是通过一套别样的表征体系来形成关于文学现象的认识,必然先在地代入了对表征自身的前理解和认同,因而这种研究说到底是一种阐释取向的批评研究,或者说计算批评。要知道,一切“新发现”归根结底是为了更好的给批评性观点提供证据,即便作者归属判定这类考证问题,其本质仍是由一环接一环的“假设—表示(建模)—推理—验证”构成的量化批评实践。因而越是在大数据的背景下,越应注意将“数据驱动”的开放探索和“论证驱动”的假设—验证范式结合起来,这种结合,也为主体审美判断的介入留下了必要的余地与契机。如何将审美感兴和细读体验用于假设的提出和模型检验,让其

^① 参见拙文《网络分析与人物理论》,《文艺理论与批评》2020年第2期。

成为批评的基石，亦是文学研究成为文学研究的重心所在。反过来说，技术本身并不分高低贵贱，能说明问题的模型就是好模型。辨识作者风格的 LSTM 神经网络是模型，评估唐诗经典化程度的影响力公式亦是模型^①，哪怕只是简单的点数统计，只要能够将结果令人信服地融入论证且经得起检验，就是一种合理、有效的运用。

数字人文取向的文学研究对反思性提出了更高要求。人文研究若想保留超越性的维度，就要对方法本身做出充分的自反性思考。在目前研究作家分布的多项课题中，还少见对取样和数据库构成的思考和讨论，这便难免会有将表征等同于“现实”的风险。但若能够以此为切口，反观总集或选本的内在编撰机制，如在聚类/分类实验中看到的思路，就有可能提出以前不易察觉的问题，又岂不妙哉^②。

最后，从长远而言，应处理好文献基础设施建设和具体研究之间的关系。大规模基础设施建设的指归是资源的存续、联通和共享。现阶段，我们对面向用户的古籍和基础文献处理平台还需加大开放力度，而非仅满足于提供查询检索和封装软件，否则，好平台也难以用于实际的研究产出，还可能引来误解，让人文研究落入全面工具化的窠臼，这种批评似乎也已不绝于耳了^③。

[作者简介] 赵薇，女，中国社会科学院文学研究所助理研究员。发表过论文《社会网络分析与“〈大波〉三部曲”的人物功能》等。

(责任编辑 刘京臣)

① 王兆鹏、邵大为、张静、唐元《唐诗排行榜》，中华书局2011年版，第8页。

② 又如郑永晓先生对《佩文韵府》的研究思路（参见郑永晓《〈佩文韵府〉的编纂与康熙朝后期的诗坛取向》，《文学遗产》2017年第3期）。

③ 参见陈平原先生近年来就数据库检索对文学研究带来的影响问题的多次发声。