

电脑与文艺鉴赏的定量化

The Computer and Quantification of Literary and Artistic Appreciation

黄鸣奋

(厦门大学中文系艺术研究所,教授 厦门 361005)

不能说对艺术的传统鉴赏方式完全没有定量分析的因素。例如,目录学家在整理和甄别文艺作品时,就很重视册数、卷数等因素。某些鉴赏家进而深入到作品内部,关注各种词汇、意象和旋律的出现频度。宋代大文豪苏轼在谈到技艺传承时还讲过一段深刻的话:“能者即数而得妙,不能者循数以得其量,其出一也。有能有不能,而精粗见焉。人见其工也,则求精于数外,而弃迹以逐妙,曰:我知酒食之所以美也,而略其方齐,舍其度数,以为不在是也,而一以意造,则其不为人所呕弃者寡矣。”^[1]文中表达了对于“数”的高度重视。尽管如此,定量分析在传统文艺鉴赏中毕竟没有成为主流。而以计算见长的电脑被引入文艺领域之后,则为文艺鉴赏别开了新面。

一、定量化鉴赏的由来

重视定量分析,是本世纪社会科学“奔向”自然科学的总趋势所决定的。在本世纪初,俄国数学家A·马尔柯夫在从事事件序列的概率研究时,将普希金的第一部史诗体小说《叶甫盖尼·奥涅金》取为对象,在分析作品中一个词汇紧跟另一个词汇的可能性问题方面取得了成就。马尔柯夫的统计方法后来被广泛应用于众多领域。计算机发明以后,日益发达的存储检索技术使作品的定量化分析变得相当便利。早期的计算机检索大多是以图书文献书目检索为基础而发展起来的,比起手工检索已有明显改进。经过30多年的研究和实践,到90年代中叶,计算机存储检索技术无论在理论上和应用上都取得了重大进展。就检索原理而言,有布尔、扩展布尔、概率、向量空间等概念模型;就文件结构而言,有倒排文件、流式文件、特征文件、图形文件、多媒体文件等;就检索操作而言,有布尔、反馈、解析、聚类等;就文献操作而言,有显示、聚类、排序、解析、分等等;从词操作上,有词干、加权、词典、截词等。^[2]上述技术在各个社会领域获得了广泛应用。值得一提的是:全文检索系统已经历了字符串检索、词检

索、智能检索等发展阶段,横跨了UNIX、DOS、Windows等多个平台,商品化软件逐渐成熟,这对于文艺鉴赏的定量化是非常有利的。

和计算机检索有着密切关系的数据库技术,是本世纪60年代开始起步的。美国、日本和欧洲各国都下大气力开发二次信息,发展数据库产业。就以美国而论,在各种久负盛名的二次信息产品从书本式向联机数据库转换的同时,开发出了一批以用户需求和市场为导向的新型产品,从而确立了世界上最大的数据库生产国和输出国的地位。近10年,我国也已开发出1000多种数据库,包括全文数据库、光盘、多媒体产品和各种类型的电子出版物。1980年,中国社会科学院就建成了《全唐诗》数据库和速检系统。使用这一系统,可以在两分钟内查找出总数达5万多首的唐诗中的任何一个字、一句诗及相应的标题、注言、异文和异作者,以及它们在《全唐诗》中的册数、页码、行数。显然,这对于进行定量分析是大有用处的。深圳大学开发的《红楼梦》数据库,有以下4方面的用途:其一,单字、双音字、成语、助词、副词等语言方面的检索;其二,修辞手法、爱情描写、心理描写等文学方面的检索;其三,戏曲音乐、绘画、医药保健、园林文艺等古代文化方面的检索;其四,原文的阅读。根据这一软件所作的统计表明:《红楼梦》前80回和后40回的语言风格相当一致,这为该书出自一人之手的假设提供了论据。相比之下,视频检索是比较困难的。但是,上海广播电影电视局节目中心组织科研人员成功研制出电视节目资料检索自动化系统,已投入运行数年。电视节目库由新闻、影视、戏剧、音舞、综艺、专题新闻、体育、少儿、社教共9个库组成,“电视节目管理系统”实施对电视节目二次文献数据库的数据登录、数据查词、数据统计和系统维护,具有检索方法优化、保留先检索词、单汉字模糊检索、正确切分和显示长字段内容、高效实现大型数据库检索结果的筛选和自由位置检索等特点。

二、定量化鉴赏的应用

建立在计算机存储检索技术上的定量化鉴赏,

已经应用于听媒艺术、视媒艺术、语文艺术(文学)等领域。

就听媒艺术的定量化鉴赏而言,早在 70 年代末,宾夕法尼亚大学的音乐教授弗兰克林·齐默尔曼博士就已着手开发一个使用“全存储器编曲”技术的复杂软件,称为旋律索引程序,简称 Melind。他的抱负是建立巴洛克时代的作曲家们的非常详尽的索引。这个索引工作分成三步。第一步是用第一行乐曲来编纂主旋律的索引,因为大多数作品的主旋律在最初的十个音调中就已显露出来了。第二步是跟踪主旋律的使用和借用。例如,亨得尔的作品索引就做得比较完整,他使用了 26000 个主旋律,并不断重复借用他自己的主旋律。第三步是让计算机分析旋律的风格“指纹”,以帮助确定作品的作者。目前所上市的音乐软件,有不少包含了定量化鉴赏的成分。例如,Howling Dog System 公司的音乐游艺室 Mr. Drumstix' Music Studio 为了培养儿童对于音符的鉴别能力,设计了识别乐器声音、判断演奏音符的多少、匹配音乐模式或者识别和弦类型的活动。

电脑在视媒艺术的定量化鉴赏方面可以说正在大显身手,迄今已知的用途主要有以下几个方面:

1. 鉴别艺术品的真伪。例如,美国纽约市艺术博物馆西欧装饰艺术部副主任卡尔·道特曼曾使用计算机来研究法国塞佛尔瓷器上的几组暗记。这种 18 世纪的瓷器常被人伪造,但真品都有镌刻或釉彩的暗记,表明制作者、制作年代、使用的瓷土等情况。在一件瓷器上可能出现 10 种不同的镌刻和釉彩暗记。道特曼利用计算机来比较和对照真品暗记的不同组合。有了这种资料,在发现赝品上的暗记不符正常情况时,就可以引起警惕。在一次查验中,道特曼一下就发现 13 件瓷器的暗记有可疑迹象。又如,圣何塞市立大学的吕兰·坦西,曾利用计算机数据库里的资料对前哥伦布时期的墨西哥和中美洲艺术品进行了分类和年代鉴定。我国学者宣国荣,则运用以马氏距离矢量的计算机视觉不变量对艺术品加以甄别。该方法把物体分成若干个部件,各部件点集同马氏距离形成的矢量成为不变量。艺术作品图像间是否有差别,可用原作品和待鉴定作品的图像的各自部件点集间一系列马氏距离来比较。

2. 考证艺术品产生的环境。布兰戴斯大学的乔治·考基尔是研究墨西哥前阿兹特克时期蒂奥底华坎城的起源和发展的。装有专门程序的计算机使他可以按照考古发现描画该城的地图。据此,他弄清了当时城内方圆 8~9 英里的市街情况。例如,有一种名叫“阿尔美那”的陶制屋顶饰品只能在上层居住区中找到。由于计算机能显示在遗址各处发现

的“阿尔美那”的数目,考基尔得以确定,当时的上层社会就在“大街”附近。计算机显示的不同时期的地图表明:当时的居住密度很高,但规划也颇合理。一位名叫雷伊·史密斯的学者利用计算机解读了在埃及发现的 35 000 块饰有图画石头,人们原以为这些石块都是来自同一个寺院废墟,但经过计算机分类,史密斯证明在全长 1 英里的遗址上,当时曾矗立许多寺院和建筑。

3. 分析艺术家的创作技巧。美国电报电话公司贝尔实验室的计算机制图顾问施沃尔兹(L. Schwartz),是计算机艺术领域的先驱之一。她和同事霍兹曼(G. Holzmann)在试验一种能够反转、改变比例和并置图画程序时,就达·芬奇的自画像及其名作《蒙娜丽莎》中的人物面孔作了比较。他们反转这幅自画像,并使两幅图像瞳孔中心之间距离相等而改变它们的比例;接着将它们平分,在并置这两幅拆半的肖像之前将鼻尖对准。这两位专家发现:达·芬奇本人的自画像的嘴唇、颊、颧骨和头形竟然与《蒙娜丽莎》是一致的!这些相似之处表明:达·芬奇在最初被画像的人不在场的情况下创作《蒙娜丽莎》时,把自己当作了模特儿。在被称为“变形”的计算机化转换中,去掉表示年龄的线条和胡须并使嘴角上翘后,表情严肃的达·芬奇就变成了微笑着的蒙娜丽莎了。除此之外,施沃尔兹还应用计算机技术研究了达·芬奇的其它作品以及 1623 年德罗舒特(M. Droeshout)所绘的莎士比亚肖像,对画家们的创作技巧均有独具慧眼的发现。顺便说一句,还有人利用电脑计算,获知在达·芬奇的《蒙娜丽莎》中,女主人公的脸的面积为 199.11 平方厘米,一只手占 169.86 平方厘米,据此推测她的身高为 168~173 厘米。

尽管有着“文如其人”之类众所周知的说法,如何辨认文学作品的真实创作者事实上是相当复杂的问题。电脑为解决上述问题提供了新的可能性。只要某一作者的风格具有相当特征并能以数量表示,就可用计算机进行相关的鉴定。

电脑进入文艺领域以后,就有一些研究者用它来鉴别典籍的作者。例如,旧约圣经以赛亚书的作者究竟为谁,千百年来一直困惑着有关的学者,开头的 12 章一般都认为是以赛亚所写,余下 54 章的作者则是有争议的。有些学者认为全书都出于以赛亚一人,但更多的人认为从第 12 章以后,作者是 2~4 位不知名的先知。以色列理工学院的希伯来学讲师拉迪试图用计算机来从事鉴定。他原来以为全书是以赛亚一人所写,通过计算机研究却发现作者不止一个。根据他计算的结果,第 1~35 章的作者可能也是第 40~66 章作者的概率仅为 0.00001。在鉴定新约圣经作者方面也进行过类似的研究。安德鲁·莫顿和詹姆斯·麦克利曼牧师发表的一篇分

析文章说,使徒保罗的 14 篇书简的作者包括保罗和其他人。另外,对柏拉图的第七封信和荷马史诗的作者,也有学者从事研究。问世于日本平安时代(公元 11 世纪初)的长篇小说《源氏物语》,共 54 卷,被誉为日本古典文学的最高峰。但是,由于该书的手写原本已经失传,因此,对于其作者究竟是谁的问题,至今仍然争论不休。日本文部省统计数理研究所用电子计算机对它的用语进行了数量分析,发现全书所使用的词汇达 37.6232 万个,而且,就前 44 卷和后 10 卷相比,形容词和助词的用法有所不同,名词与助动词的使用情况差别更加显著。如果从词汇搭配的角度继续进行深入研究的话,是有可能逐步揭示这一千古之谜的。

电脑既然可以用来鉴别作者,也就可以用来辨别文学作品的真伪。1976 年,英国《泰晤士报》刊登某出版商的广告,宣称新发现莎士比亚遗著。这一消息在英国读者中造成轰动,也吸引了专业人员的注意。剑桥大学文学系的两位教师运用计算机,将《莎士比亚全集》的用词习惯和所谓“遗著”加以比较,结果发现后者原来是赝品。上述鉴别技术,是由英国爱丁堡大学计算机系安德鲁·莫顿等三位博士发明的。这一技术又被英国巴新斯顿技术学院讲师托马斯·梅里用以分析传世的剧本《托马斯·莫尔爵士之书》,推翻了此剧系由安东尼·芒戴等人合著的传统看法,证实其作者是莎士比亚。其根据是它在用词习惯上和莎士比亚的剧本《裘力斯·凯撒》、《泰特斯·安德洛尼克斯》、《泰尔亲王配力克里斯》颇为相似。

电脑甚至还可以推算出作家创作特定作品的准确年代。1951 年,海明威写信给一位新闻记者说,他正在创作一个故事,即后来在 1952 年发表的《老人与海》。但达特茅斯学院的英语教授戴累尔·曼塞尔却认为海明威这个故事写成的时间比 1952 年早得多,应是在 1935 年或 1936 年,写成之后就收藏起来,直到 1951 年才把这旧稿取出拍掉积尘,做出好象正在构思的样子。为了证明这点,曼塞尔使用统计方法来对《老人与海》和另外 4 部创作时间基本无误的作品进行比较,结论是:《老人与海》和一部写于 1936 年的作品之间的相似程度要比它和写于其它年代的作品之间的相似程度大得多。

电脑不仅可以对具体作家、具体作品进行个案研究,而且可以对范围更广的文学对象进行综合研究。作为例证,可以举出北京大学目前正在开发的以全宋诗为首选对象的古诗研究系统,即古诗研究的计算机支持环境。该系统具有基于超文本的阅读、写作功能(研究者可将自己的心得、考据等作为新注释录入,形成电子笔记或电子眉批)、全文检索和属性检索功能,包括诗作统计、汉字统计和整体一局部对照统计在内的统计功能,以及语言、格律

及风格等方面的辅助研究功能。仅就风格研究而论,如果统计了某一诗人作品中的全部汉字及其频度,也统计了《全宋诗》的全部汉字及其频度,便可以了解这位诗人使用汉字的风格与同时代诗人是否相近。如果实现了诗句中词的切分与词典的编纂,则可以研究诗人造词、造句的风格,包括是否喜欢用典等。即使不将特定诗人的作品与全宋诗作比较,仅就他(她)自己的诗进行统计,也可以分析出若干风格上的特征。例如,计算陆游全部诗作中各种诗体所占的比率,对古体诗换韵、平声韵、仄声韵、近体诗首句押韵、平仄变化合乎常规与否等现象进行统计分析,便可以判断陆游写诗是严格遵循诗的格律还是比较自由的。利用这些统计数据,可以在概率的意义上判定某些归属存疑的诗是否陆游所作。

除上述领域之外,量化分析当然也可以适用于建筑、戏剧、电影、电视节目等综合文艺。例如,南佛罗里达大学的詹姆斯·斯特兰奇在耶路撒冷以北 90 英里的柯贝西马村修复一个约在公元 417 年被毁的犹太教会堂。在基本弄清原建设的规模后,斯特兰奇用计算机来帮助确定一些已无法查明的结构参数,并勾勒出这个建筑最可能具有的形状。^[3]

三、量化鉴赏的评价

在科学研究中,定量分析和定性分析是相辅相成的两个方面。而且,定性分析经常是定量分析的必要前提。就以用词习惯统计而言,这种技术事实上是以“作者(至少是成熟的、形成了独特风格的作者)在使用词语的类型和频度方面均有一定之规”的理论假设为基础的。当然,上述假设是否始终成立,尚待证明。1980 年,美国威斯康星大学讲师陈炳藻利用计算机对《红楼梦》前 80 回和后 40 回所使用的形容词、名词、副词、动词、虚词进行分类比较,得出了这部名作出于一个作者之手的结论。我国有些作家就不同意上述看法。他们从自己的实践经验出发,认为创作者在完成一部长篇小说时前后风格是有所变化的,陈炳藻在设计程序时并未考虑到上述因素。

就检索技术而言,目前比较成熟的只是文本方面的。不论是音乐的数字化记录、数字化摄像或语音合成芯片的开发,近年来都取得了相当可观的进展。上述进展至少说明:计算机不仅可以用来处理数字和符号,而且可以用来处理声音和视觉模式之类似模拟信息。尽管如此,比起处理数字和符号的技术来,模拟信息技术的发展是落后的。例如,对于数字和符号,计算机可以十分方便地进行检索,然而,模拟信息之间的模式匹配,则就不是一件容易的事情了。现在可将大量的音乐旋律或照片存储在

一张光盘上,但是仍然难以直接找出旋律彼此相似的音乐信息或者能使人回忆起某张照片的图画,^[4]而只能以加注关键词的方法进行间接检索。但是,音频、图像和视频信息的内容具有丰富的内涵,在许多情况下仅用几个关键词难以充分概括,而且,作为关键词的图像特征的选取也有很大的主观性。正是在这种背景下,基于内容的检索应运而生。就图像与视频检索而言,与传统的检索方式相比较,基于内容的检索融合了对于静止与动态图像的理解技术,从而可使人们从巨容的图像和视频库中根据要求进行有效检索。具体地说,对于静止图像,可以根据颜色、形状和纹理等特征,或者给出样本图像进行检索。无法给出特征和样本的图像时,再进行浏览检索。对于视频信息,必须将它们分割成镜头,并为其建立描述,以供检索之用。如果视频数据量太大,则需要添加视频图符从粗到细进行分层管理。音频信息的检索和相关的定量化鉴赏,也可用类似的方法进行。

要想提高文艺鉴赏的科学性,有必要将定性分析和定量分析有机结合起来,建立相应的专家系统。所谓专家系统,指的是能够像专家一样解决某些领域中的问题的智能计算机系统。它能够处理符号和概念,并根据推理找出结论和答案,因此比仅能根据规定的算法处理数据执行计算任务的普通数据处理系统高出一筹。专家系统主要由知识库、推理机和数据库三部分组成。推理机是专家系统的核心,能接受所要解决的问题,通过正向、反向、混合三种形式的推理获得答案。上述推理是在知识库和数据库的支持下进行的。知识库存贮了从专家那儿所得的专门知识,这些知识以状态空间搜索、语义网络、产生式规则等方法加以表示。数据库的用途是储存初始数据以及推理过程中所得的各种中间信息。专家系统与自然语言理解、机器人学并列为人工智能三大研究方向,于60年代中期成为科技界的研究课题。1971年,世界上第一个专家系统DENDRAL由美国斯坦福大学的爱德华·菲根鲍姆研制成功。这一系统能够像化学家那样分析化合物质谱,推断未知的化学结构。从那时以来,全世界已研制出2000多种专家系统,主要分布在医疗、

地质、化学、物理、生物、军事等领域。尽管业已研制成功的专家系统相当多,但还未见到关于文艺鉴赏专家系统的报道。之所以出现这种情况,和下述两条原因不无关系:其一是文艺鉴赏专家系统目前的实用价值相对来说还不是非常大。在许多情况下,人们倾向于相信自己的审美判断力、保留自己独特的审美趣味、满足于欣赏随手得到的文艺作品,而不想向专家或专家系统请教。其二是文艺鉴赏专家系统的开发难度相当高。国内一位学者在分析文艺数量化研究的现状时曾列举出三条障碍:从主体的角度看,文艺研究很难做到纯客观;从客体的角度看,文艺所反映的是社会生活的整体,又表现出作家的个体思想感情,因此,难以用数量化关系来表示;三是文艺作品贵在创新,因此,依据重复出现的文艺现象去进行研究以求得数量关系的测定,似乎就不可思议。^[5]这些障碍是确实存在的。由于这些缘故,文艺数量化研究的成果目前主要集中在作品的形式方面。只有在基于内容的检索技术充分完善之后,对于文艺作品的定量化鉴赏才可能是深入的,这种深入是建立文艺鉴赏专家系统的必要条件。至于文艺鉴赏专家系统的实用价值,将会随着当代通信技术的发展、文艺作品数量空前迅速的增长而日益增加。在即将到来的新世纪中,这种专家系统将帮助人们从海量的作品中寻找上乘之作,使我们的宝贵时光不致于被“艺术垃圾”所占用,艺术赝品也不再那么容易蒙住鉴赏者的眼睛。

参考文献

- [1]苏轼文集第二册:盐官大悲阁记,中华书局,1986年3月版,P387
- [2]曾民族,中文全文文本检索系统的技术特点,中国计算机报,1996年3月28日,第27版
- [3](美)唐纳德·桑德斯,计算机·信息·社会,周赞明等编译,知识出版社,1985年12月版,P271~272
- [4](日)神沼二真、松本元,生物计算机——日本的下一代计算机,顾仲梅等译,国防工业出版社,1996年6月版,P19
- [5]姚鹤鸣,文艺数量化研究之问题与探讨,收入钱中文、李衍柱主编的《文学理论:面向新世纪》,山东人民出版社,1997年7月版,P318~326

(责任编辑 王宏章)

(上接第52页)

3. 为制止各类高新技术开发区之间的不规范竞争,政府有关部门应对进入我国的国外高新技术资源进行筛选和控制,设立国家技术引进中心、国家技术创新中心,专门致力于吸收和消化国家引进的高新技术资源,对经国家技术引进中心论证的,并有能力进行模仿创新的高新技术原则上由国家

创新中心负责“物化”的实现,或指定有条件的高新技术开发区负责“物化”的实现,然后向其它开发区转移。

4. 加快培育高新技术市场,搞活技术要素市场的流通渠道,消解技术扩散、技术转移过程中的种种不利的“壁垒”。

(责任编辑 孙立明)