

基于数字技术的中国画论基因观念识别研究^{*}

钮亮¹ 徐佳俊¹ 项玮²

(1. 中国计量大学经济与管理学院, 杭州 311222;

2. 浙江外国语学院艺术学院, 杭州 311222)

摘要:[目的/意义] 基因观念是决定文化走向的重要依据, 识别中国画论基因观念对理解中国绘画的审美取向及其在整个东方艺术中的定位有重要帮助。[方法/过程] 首先采用 BLAST 方法构建中国绘画论文本引用网络, 接着依据改进的库恩基因测度指标提取画论基因观念, 最后通过共词网络考察基因观念之间的语义联系。[结果/结论] 从研究结果看, 谢赫提倡的六法、王维提倡的山水技法、苏轼倡导的文人画理念是中国绘画发展的重要基因观念。该研究为定量识别文化基因提供了一种探索。

关键词: BLAST 方法 库恩基因测度 中国画论 基因观念

分类号: G203

DOI: 10.31193/SSAP.J.ISSN.2096-6695.2023.01.04

Dawkins 认为, 人类文化的发展受到基因控制, 因为它携带着“遗传密码”, 决定着文明的发展^[1]。基因观念应用于知识领域, 称为“知识基因”, 通常指存在于文献中不可再分割的具有完备知识观念表达的知识单位, 是一切知识管理活动的前提^[2]。梁镇涛等将知识基因引入医学信息领域, 从微观层面对以往基于文献级别同类研究进行补充扩展, 对医学知识基因的传播有了更直观的观察^[3]。孙晓玲等利用知识基因发现算法, 以石墨烯碳类纳米材料为例, 探索了石墨烯作为一种信息的碳类纳米级材料在各科学技术领域的应用和传播, 以此探究科学和技术之间的关系^[4]。但基因识别方法目前尚未见应用于画论观念识别中, 本文尝试对其进行探索。

中国画论是对我国绘画美学思想、画法技巧等进行探索的成果, 它蕴含的观念对理解中国绘画的审美体系和艺术定位有重要帮助。然而中国画论观念受外部环境和绘画实践的影响处于流变之中, 出现诸如画派上的南北分宗、绘画从“形似”到“传神”的转变等。在这些流变中, 那些

^{*} 本文系 2023 年度浙江省哲学社会科学规划课题“数字人文视域下浙江古代画学文献整理及观念演化研究”(23LMJX06YB)、2022 年度教育部人文社会科学研究规划基金项目“基于数字技术的中国图像志方法论研究”(22YJA760081) 阶段性成果。

[作者简介] 钮亮 (ORCID: 0000-0001-6934-4416), 男, 副教授, 研究方向为数字人文、中国画论, Email: niutyut@126.com; 徐佳俊 (ORCID: 0000-0003-4800-3561), 男, 本科生, 研究方向为知识基因、数字人文, Email: 824127941@qq.com; 项玮 (ORCID: 0000-0001-5546-2091), 女, 讲师, 研究方向为数字人文、近现代美术史, Email: 173216819@qq.com。

保留和传承至今, 且处于支配地位的绘画观念构成了基因观念。中国绘画的基因观念不仅对中国绘画体系构建起到推进作用, 而且对中国绘画在整个东方艺术中的定位做出了重要贡献。因此甄别中国画论中的基因观念, 考察其在中国绘画史上的作用是非常重要的。

1 理论与方法

由于观念是通过书写被确立和加固的, 如果不同朝代的画论文本之间存在引用关系, 那么观念就会在引用关系中得到传播, 传播的链条越长越有可能是基因观念。鉴于此, 甄别画论基因观念就转变为从画论文本的引用关系中识别它们。假设观念沉淀为关键词^[5], 识别基因观念就转变为识别关键词。由于古代写作鲜有现代学术的引用规范, 这使得被引用的内容和作者本人的写作混杂在一起, 很难确定画论文本的引用内容。依照 Vierthaler 等人的建议, 如果文本之间存在重用内容, 则视它们之间存在引用关系^[6]。借助文本重用确立画论引用, 然后在引用的画论文本中查找基因观念是本文的思路。

1.1 BLAST 方法及引文网络构建

文本重用通过计算文本之间的相似度进行, 粗略归为序列比照法与词袋法。序列比照法是逐字比较两个文本之间最长的相同连续字符序列的长度, 常用算法为“史密斯-沃特曼算法”(Smith-Waterman algorithm)^[7]和“BLAST 方法”(Basic Local Alignment Search Tool, 局部序列比对基本搜索工具)^[8], 两相比较, 史密斯-沃特曼算法的复杂度较高, 时间和空间开销较大^[9]。词袋法不关心比对话句中词的顺序, 而是根据词出现的概率来进行比较, 通常需要借助主题模型来计算文本的主题分布, 并以此分布来判断文本之间的相似^[10]。然而, 主题模型只能比较整个文本内容, 无法捕捉局部重叠的语句; 同时, 运用主题模型计算相似度依赖于分词, 对于分词不理想古籍文本来说, 选择主题模型进行相似度比较并不是很好的选择。综上, 对探测画论文本之间的相似度来说, BLAST 方法在保持较高精度的情况下大大减少了运行时间, 而且进行局部序列比对又无需考虑分词, 是古籍比对较好的解决办法^[6]。因此, 本文采用 BLAST 方法建立中国画论的文本引用关系网络。

BLAST 方法采用的是一种启发式算法, 它的工作过程是: 首先, 通过常用的词汇识别语言模型“N-gram 方法”计算每篇画论的词项作为基因候选词(由于 N-gram 成词的覆盖性, 不存在关键词丢失的情况, 其召回率是百分百, 保证了观念基因的全面), 将输入的文本切割成子片段(N 为子片段所含的字数), 每个片段赋予片段编号、片段在文本中的位置、文本编号, 从而构建可索引的 BLAST 数据库, 同时, 构建 k 个可能的种子词。然后, 将种子词与预先索引好的 BLAST 数据库中的序列片段进行比对。接着, 选择种子词中连续打分较高的位置采用动态规划算法进行延伸, 延伸过程利用 Levenshtein 算法计算相似度得分, 当得分低于阈值, 延伸过程终止, 输出得分高的序列。由于文本之间可能存在多条相似序列, 其总得分为加总全部匹配序列得分, 公式为 $score = \sum_i^n length_i \times similarity_i$, 其中每条序列得分等于匹配序列的字符数量乘以其相似度。最后, 选取得分大于一定阈值的画论进行两两连边, 构建画论引文网络。BLAST 方法计算相似度和引文网络构建分别通过 python 的 Levenshtein 包和 networkx 包进行, 得到的相似句子见表 1。

表1 Blast方法搜索相似语句(部分数据)

源画论句子	相似画论句子	源画论作者信息	相似画论作者信息	相似度
古人云有笔有墨	古人云有笔有墨	清-王槩-学画浅说	明-莫是龙-画说	1
笔墨二字人多不晓画岂无笔墨哉	笔墨二字人多不晓画岂无笔墨哉	清-王槩-学画浅说	明-莫是龙-画说	1
作诗必此诗定知非诗人	赋诗必此诗定知非诗人	明-董其昌-思白论画	宋-苏轼-东坡论画	0.8754
正如摩诘所云行到水穷处坐看云起时是也	行到水穷处坐看云起时	清-王昱-东庄论画	宋-郭熙郭思父子-林泉高致	0.7669
丈山尺树寸马豆人此其法也	丈山尺树寸马分人	五代-荆浩-山水赋	唐-王维-山水论	0.7204

1.2 库恩基因观念测度及改进

基因观念识别多基于文献的引证关系进行,其中库恩(T.kuhn)等提出基于引文网络的粘滞系数(Sticking)和分散系数(Sparking)的基因观念测度指标^[11],该指标被认为能够有效识别基因观念^[12],公式如下:

$$P_m = \frac{d_{m \rightarrow m}}{d_{\rightarrow m} + noise} \bigg/ \frac{d_{m \rightarrow \bar{m}} + noise}{d_{\rightarrow \bar{m}} + noise} \quad (1)$$

$$f_m = d_m / d_{all} \quad (2)$$

$$M_m = P_m * f_m \quad (3)$$

其中, P_m 为基因观念的传播分数, $d_{m \rightarrow m}$ 表示包含基因观念 m 且引用了至少一个包含 m 的文献数量, $d_{\rightarrow m}$ 表示所有引用了至少一篇包含 m 的文献数量, $d_{m \rightarrow \bar{m}}$ 表示包含基因观念 m 且没有引用包含 m 的文献数量, $d_{\rightarrow \bar{m}}$ 是所有没有引用包含基因观念 m 的文献的数量。noise 是噪声元素,其作用一是防止出现分母为零导致最终得分为无穷的情况,二是防止最后筛选出没有意义的传播单一的基因观念,起到惩罚传播单一基因观念的作用。 $\frac{d_{m \rightarrow m}}{d_{\rightarrow m} + noise}$ 构成基因观念 m 的粘滞系数,表示该基因观念在引用包含该基因观念的文本中出现的比例,代表基因观念被继承和传递的程度。 $\frac{d_{m \rightarrow \bar{m}} + noise}{d_{\rightarrow \bar{m}} + noise}$ 构成基因观念 m 的分散系数,表示独立包含该基因观念的文献数量与没有引用该基因观念的文献数量的比率,代表基因观念在扩散过程中被割裂的程度。

f_m 为基因观念 m 在所有文献中出现的频率。 d_m 为包含 m 的文献数量, d_{all} 为所有文献的数量。引入 f_m 能够有效地控制低频基因观念对结果的影响。

M_m 为库恩值,表示所选基因观念的重要程度。

然而, 库恩算法只考虑观念在引用网络中传播的层级关系和引用频次, 对于含有时间意涵的观念探测没有考虑。事实上, 越早期的观念对后来的文化越有形塑作用, 将时间靠前的观念发掘出来非常重要。因此, 本研究对库恩等人的测算方法进行了改进, 保持 P_m 和 f_m 不变, 考虑基因观念 m 出现的时间顺序, 设计了奖励因子。以函数 $\text{contain}(m \rightarrow t)$ 判断观念基因 m 是否存在于时间 t 中, 某个观念在不同时间中出现的机会越多, 它越重要, 表现了传承性, 记为 cour1 ; 某个观念出现的时间越早, 它越重要, 表现了优先性, 记为 cour2 。公式为:

$$\text{cour1} = \sum_{t=1}^T \begin{cases} 0 & \text{contain}(m \rightarrow t) = \text{False} \\ \frac{1}{t} & \text{contain}(m \rightarrow t) = \text{True} \end{cases} \quad (4)$$

$$\text{cour2} = (T - \text{begin_t}) / T \quad (5)$$

$$M'_m = P_m * f_m * \text{cour1} * \text{cour2} \quad (6)$$

T 为全部时段, t 为每一个具体时段, begin_t 为基因 m 第一次出现的时段, M'_m 为运用改进的库恩测度算法计算的库恩值。

为了更好地理解改进的库恩测度算法, 本文通过测试用例进行说明。测试用例使用包含 10 个节点【 $n1$ 、 $n2$ 、 $n3$ 、 $n4$ 、 $n5$ 、 $n6$ 、 $n7$ 、 $n8$ 、 $n9$ 、 $n10$ 】的引文网络, 时期划分为 $T1$ 、 $T2$ 、 $T3$ 、 $T4$, 时间依次由远及近。各节点分布在不同时期, 且具有先后顺序, 其引用网见图 1 所示。

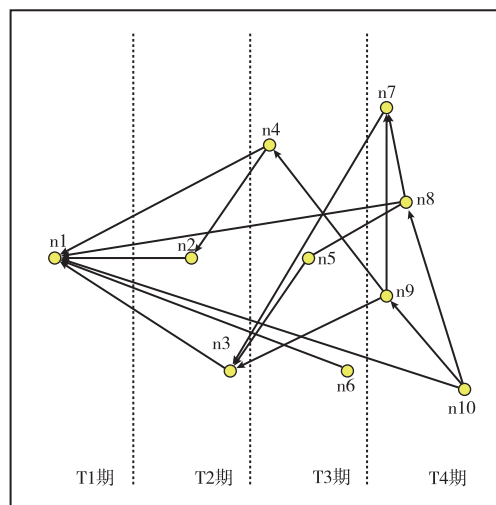


图 1 引用网示例

测试所用的基因观念为【 $M1$ 、 $M2$ 、 $M3$ 、 $M4$ 】, 它们的引用情况见表 2, 传播网络见图 2。选择噪音系数 noise 等于 1, 原库恩测度算法和改进的库恩测度算法的计算结果见表 3。 rank 为库恩值得分排名, rank' 为改进算法库恩值得分排名。

表 2 基因观念引用情况

节点	基因观念	引用情况
n1	M1	\
n2	M1; M2	n1
n3	M1; M3	n1
n4	M2	n1; n2
n5	M3	n3
n6	M1; M2	n1
n7	M1; M4	n3
n8	M1; M3; M4	n1; n5; n7
n9	M2; M4	n3; n4; n7
n10	M1; M4	n1; n8; n9

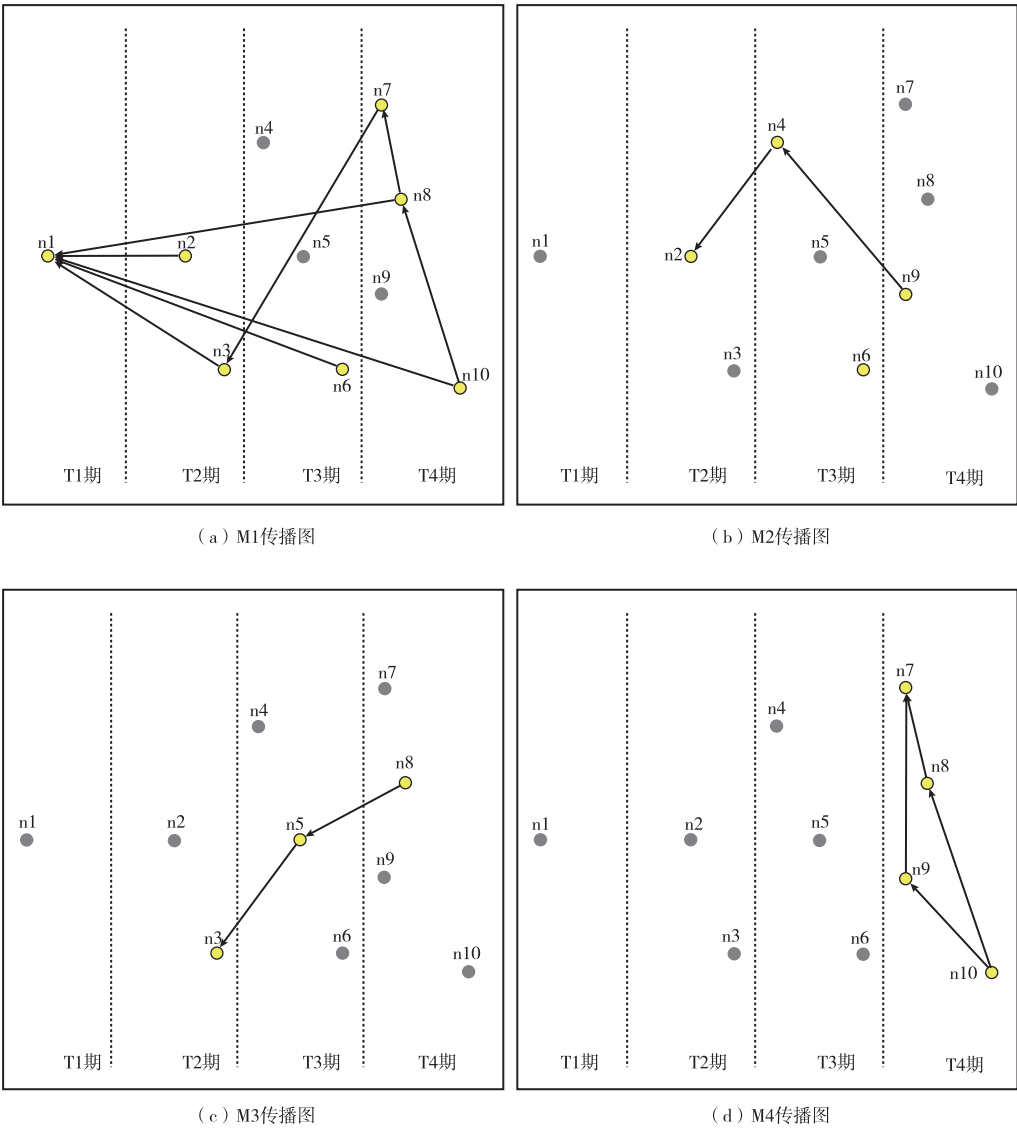


图 2 基因观念传播图

表 3 测度结果

基因观念	begin_t	P_m	f_m	cour1	cour2	M_m	rank	M'_m	rank'
M1	T1	0.6	0.7	1	1	0.42	3	0.42	1
M2	T2	1.33	0.4	0.75	0.75	0.53	2	0.3	2
M3	T2	1	0.3	0.75	0.75	0.3	4	0.17	3
M4	T4	3	0.4	0.25	0.25	1.2	1	0.08	4

依据原库恩测度算法得到的库恩值得分排名为【M4, M2, M1, M3】,改进后的库恩测度算法得到的库恩值得分排名为【M1, M2, M3, M4】。其中 M4 第一次出现的时期为 T4, 粘滞系数为 0.75, 分散系数为 0.25, 传播分数 P_m 为 3, 出现频率 f_m 为 0.4; M1 第一次出现的时期为 T1, 粘滞系数为 0.60, 分散系数为 1, 传播分数 P_m 为 0.6, 频率 f_m 为 0.7。原库恩测度算法得到的 M4 库恩值为 M1 的近三倍。但是, 引入奖励因子 cour1 和 cour2 后, 对出现频次高、出现时期较早的基因观念 M1 具有增益效果, 最终 M1 从原有排名的第三位升为第一位, 符合基因观念的历史地位。

1.3 基因观念共词网络构建

运用库恩基因测度方法可以提取文献中的重要基因观念, 但这些基因观念是独立存在的。知识的演进通常是多组基因观念共同作用的结果, 彼此关联的基因观念既相互说明, 反映领域知识的重点, 又反映了知识演进的走向。因而, 发掘基因观念之间的关系成为了解知识演进过程的重要突破口, 在这方面, 常用的方法是共词网络分析。共词网络分析是研究词与词之间深层关联的一种方法, 它以文本的段落或语句作为窗口单元, 通过统计出现在同一窗口单元中的词汇并对其两两相连, 来确立词汇间的联系。就基因观念共词网络的构建来说, 过程如下^[13]:

- (1) 以库恩算法找到的基因观念作为词典;
- (2) 依照窗口单元(段落或语句)切分全部文献;
- (3) 让指针指向窗口单元的起始位置;
- (4) 通过正向最大匹配算法, 提取窗口单元中包含的词典中的基因观念;
- (5) 将出现在窗口单元中的基因观念两两相连, 统计累计关联次数;
- (6) 重复步骤(3–5)至下一窗口单元, 直至所有单元循环完毕。

构建基因观念共词网络后, 对其中的基因节点按照它们的库恩值排序前 20 进行子网提取, 节点大小受库恩值影响。提取出的基因观念子网中的节点一方面具有库恩基因测度的重要性, 另一方面网络拓扑结构又揭示了基因观念在关系中的重要性。由于基因观念网络是以窗口单元(段落、语句)为构建基础, 保证了基因观念的语义完整, 选出的基因观念更准确。

2 中国画论基因观念实证分析

2.1 研究框架

根据上述理论与方法, 针对中国画论基因观念的实证分析, 本文确定整体研究框架如图 3 所示:

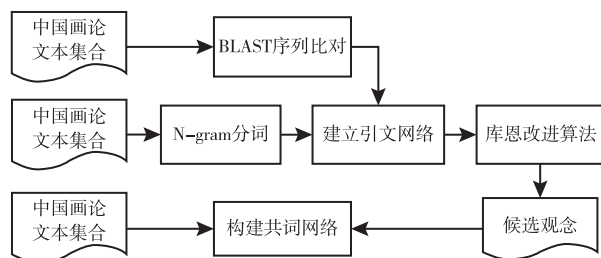


图3 本文整体研究框架

(1) 通过 BLAST 方法构建中国画论引文网络。

(2) 运用改进的库恩基因识别算法考察带有时间意涵的中国画论基因观念，选取候选基因观念。

(3) 使用正向匹配法遍历中国画论中的候选基因观念，对出现在同一段落中的基因观念建立联系，进而构建共词网络，探测基因观念之间的语义关系，确立中国画论的观念体系。

2.2 数据来源及预处理

本研究的数据来源主要出自俞剑华先生整理的自先秦至明清的《中国历代画论大观》，全书共十编，收录 453 篇不同朝代的代表画论。选择俞剑华《中国历代画论大观》作为语料库有几个原因^[14]：(1) 收录较为全面。集成了画理、画法、画诀、画诗、画品、画评、画谱、画说、画鉴等内容，除了常规画论外，涉及画法讨论颇为精深的题跋也一并收入。(2) 时代为经，性质为纬。依照朝代顺序收录画论文本，为基因观念发掘提供了基础材料。(3) 真实性和纯粹性。对画论的真伪进行了辨别，摒除了画史文字，保证了画论论理的纯粹性。

首先，对《中国历代画论大观》的纸质文本进行矢量化处理，人工校对后以“作者－朝代－画论名称”为题保存为文本文件。然后，通过 BLAST 方法对整理好的中国画论文本进行文本重用计算，搭建中国画论引用网，如图 4 所示。图中节点大小代表数据中所获总分多少，有向边线宽度为画论之间引用的重复次数。

其中，度中心性得分最高的节点是《清－恽格－南田题跋》，总得分为 11 032.0；其次是《清－沈宗骥－芥舟学画编》，总得分为 9 238；第三名是《清－汤贻汾－画筌析览》，总得分是 5 739.067；第四名是《清－方薰－山静居画论》，总得分为 5 322.0；第五名是《清－笪重光－画筌》，总得分为 5 237.067。

2.3 中国画论基因观念测度

中国画论基因观念沉淀于关键词中，为了较为全面地捕捉中国画论基因观念，对不同朝代画论文本进行分词是首要工作。依照基因观念的意义，选择 2-gram（二字词）和 4-gram（四字词）对画论文本进行分词。然后利用库恩基因测度算法计算中国画论的基因观念。

不同朝代的画论文献数量差距悬殊，明清两代的画论数量约是之前朝代的三倍。根据文献的增长理论，任何研究领域的文献会随着时间的推移不断增长且增长速度不断增大直至趋于饱和^[15]。由于中国画论在明清时期出现爆炸式增长，彼此之间的引用相比其他朝代也多。由于库恩算法只考虑观念在引用网络中传播的层级和引用频次，所以明清时期出现的关键词，比如

表5 库恩基因测度算法 4-gram 得分

基因观念	朝代	$d_{m \rightarrow m}$	$d_{\rightarrow m}$	$d_{m \rightarrow \bar{m}}$	$d_{\rightarrow \bar{m}}$	P_m	f_m	M_m
荆关董巨	明	13	48	6	405	11.56	0.04	0.48
古人作画	宋	11	36	7	417	11.85	0.04	0.47
青绿山水	明	12	55	3	398	13.83	0.03	0.46
用笔用墨	明	12	44	8	409	9.56	0.04	0.42
自然生动	清	6	23	1	430	24.98	0.02	0.39
黄鹤山樵	元	16	65	12	388	6.13	0.06	0.38
活泼泼地	清	5	15	3	438	20.42	0.02	0.36
骨法用笔	唐前	8	35	3	418	14.77	0.02	0.36
米家父子	明	11	57	4	396	10.45	0.03	0.35
气韵生动	唐前	15	62	14	391	5.35	0.06	0.34

中国画论的基因观念需要考虑先前性和传承性两个维度，出现地越早且跨越更多朝代的观念更容易是重要的基因观念。利用库恩改进算法实现中国画论基因的测度，在改进算法中， t 为 m 第一次出现的朝代。在 T 的划分中，之所以以唐为分界，是由于绘画在唐进入了自觉期。梁启超以为“中国绘画，大体上，中唐以前是一个时代，开元、天宝以后是另一个新时代，分野在开元初年。”^[16] 滕固认为“唐朝是中国绘画史的黄金时期。这个时期的重大意义在于：一方面，山水画开始独立发展，形成中国的绘画主流；另一方面，佛教绘画脱离了外来影响，并形成中国风格。”^[17] 苏轼在《书吴道子画后》说“知者创物，能者述焉。非一人而成也。君子之于学，百工之于技，自三代，历汉至唐而备矣。故诗至于杜子美，文至于韩退之，书至于颜鲁公，画至于吴道子。而古今之变，天下之能事毕矣。道子画人物，如以灯取影，逆来顺往，旁见侧出，横斜平直，各相乘除，得自然之数，不差毫末，出新意于法度之中，寄妙理于豪放之外，所谓游刃余地，运斤成风，盖古今一人而已。”^[17] 梁启超以开元初年作为分野，滕固则从绘画的自觉性上将唐代作为中国绘画的黄金时期，苏轼将唐朝作为中国文化最完备的时期。之后朝代画家、画作、画论都渐趋发达，因此将画论时间分为唐前、唐、五代、宋、元、明、清大抵是比较合适的。利用改进的库恩算法得到基因观念得分，结果见表 6、表 7 所示。

可以看出两字词中唐前有“用笔”“山水”“自然”“气韵”等观念，唐代有“笔墨”“画家”观念，都占据非常重要的位置。这些观念在之后朝代的画论中持续存在，它们的传承系数 $cour1$ 和时间系数 $cour2$ 都非常高，均达到或接近 1。说明了这些两字词观念的基因属性。

四字词中唐前有“气韵生动”“骨法用笔”“经营位置”，唐代有“元气淋漓”“凡画山水”“顾陆张吴”“远人无目”，宋代有“平淡天真”“山水人物”等基因观念贯穿了整个中国古代画论。说明中国绘画对“气韵生动”“骨法用笔”“经营位置”“平淡天真”等观念的讨论普遍存

在于众多画论之中, 这些基因观念在某种程度上决定着中国绘画的走向。以四字词为例, 对中国画论基因观念进行库恩算法和改进算法的探测比较, 总体而言, 使用改进后的库恩基因算法能够捕捉到更多的中国画论基因观念, 只有清代部分相反, 这是因为, 清代是本研究选取的最后一个朝代, 之后的传承已经无法在本研究体现了。比较结果见图 5 所示。

表 6 改进后的基因测度算法 2-gram 得分

基因观念	朝代	$d_{m \rightarrow m}$	$d_{\rightarrow m}$	$d_{m \rightarrow \bar{m}}$	$d_{\rightarrow \bar{m}}$	P_m	f_m	<i>cour</i> 1	<i>cour</i> 2	M'_m
用笔	唐前	73	98	52	355	4.70	0.28	1.00	1.00	1.30
山水	唐前	88	113	131	340	1.94	0.48	1.00	1.00	0.94
古人	唐前	85	103	110	350	2.51	0.43	0.86	1.00	0.92
书画	唐前	75	110	90	343	2.47	0.36	1.00	1.00	0.90
自然	唐前	60	101	55	352	3.53	0.25	1.00	1.00	0.90
气韵	唐前	55	92	54	361	3.70	0.24	1.00	1.00	0.89
画法	唐前	56	89	48	364	4.38	0.23	0.86	1.00	0.86
树石	唐前	40	77	37	376	4.74	0.17	1.00	1.00	0.81
笔墨	唐	81	102	93	351	2.84	0.38	0.86	0.86	0.80
画家	唐	66	95	64	358	3.63	0.29	0.86	0.86	0.77

表 7 改进后的基因测度算法 4-gram 得分

基因观念	朝代	$d_{m \rightarrow m}$	$d_{\rightarrow m}$	$d_{m \rightarrow \bar{m}}$	$d_{\rightarrow \bar{m}}$	P_m	f_m	<i>cour</i> 1	<i>cour</i> 2	M'_m
气韵生动	唐前	15	62	14	391	5.35	0.06	0.86	1.00	0.29
骨法用笔	唐前	8	35	3	418	14.77	0.02	0.71	1.00	0.26
经营位置	唐前	10	49	8	404	7.12	0.04	0.86	1.00	0.24
元气淋漓	唐	11	55	8	398	6.91	0.04	0.71	0.86	0.18
古人作画	宋	11	36	7	417	11.85	0.04	0.57	0.57	0.15
凡画山水	唐	5	37	5	416	6.55	0.02	0.86	0.86	0.11
山水人物	宋	13	62	12	391	5.25	0.06	0.57	0.57	0.09
顾陆张吴	唐	7	50	5	403	6.70	0.03	0.57	0.86	0.09
平淡天真	宋	8	38	3	415	13.59	0.02	0.43	0.57	0.08
远人无目	唐	2	13	6	440	6.15	0.02	0.86	0.86	0.08

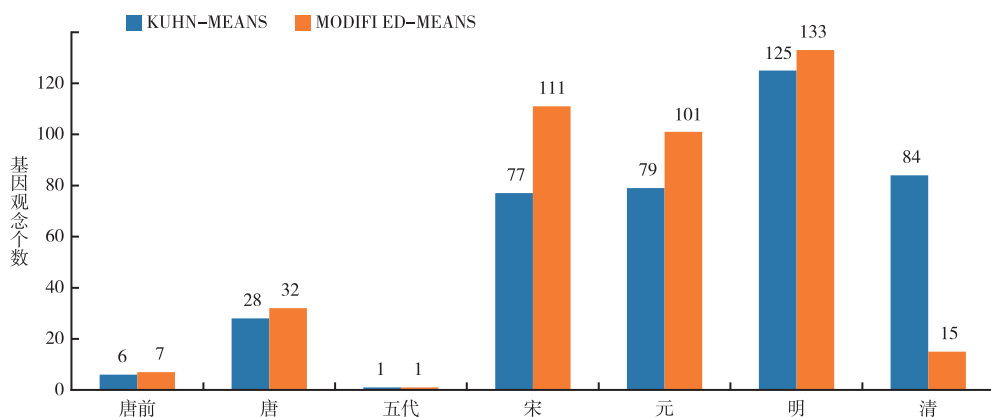


图5 库恩算法和改进后的4-gram库恩算法各朝代基因观念个数对比（以前400为例）

2.4 基因观念共词网络构建

运用改进的库恩算法提取中国画论核心基因观念后，再通过共词网络揭示基因观念之间的关系。本文以画论文本的段落作为窗口单元，选取两字词和四字词基因观念作为字典，依照最大匹配算法遍历全部中国画论段落，对出现在同一段落中的基因观念进行两两相连，构建共词网络，提取基因观念排序前20的词汇进行子网显示，节点大小依照库恩值排序。见图6、图7。

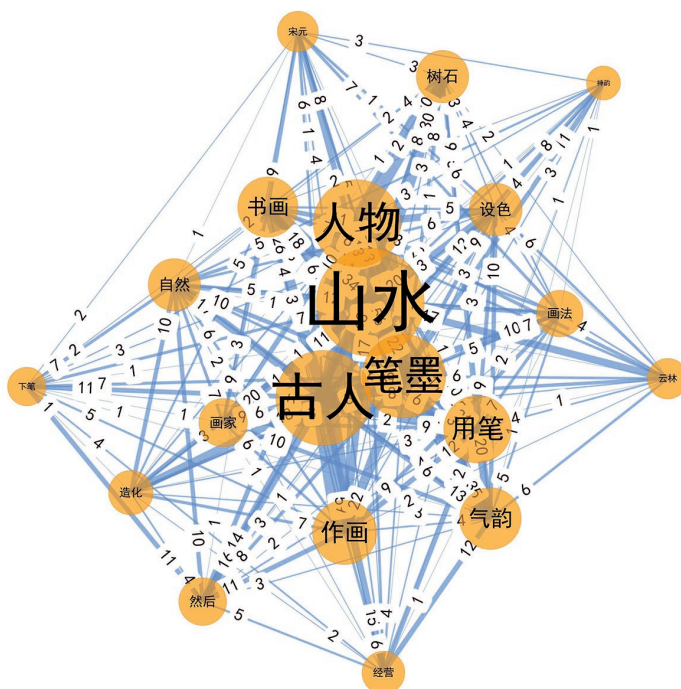


图6 两字词基因观念排名前20的共词网络

可以看出，在两字词基因观念中，“山水”“笔墨”“人物”“古人”等处于共词网络的中心位置，这说明“山水”“笔墨”“人物”“古人”等基因观念不仅在库恩算法中排序靠前，而且在网络拓扑结构中与其他关键词连接频繁，体现了语义的重要性。例如库恩算法排名第一的“用笔”

观念在共词网络中排名靠后, 而排名第二的“山水”观念在共词网络中变得非常醒目, 对画论的思想发展具有举足轻重的作用。中国文化大致可以概括为虚实、阴阳、天地、气理、心意、神这六个概念。前三个(虚实、阴阳、天地)属于宇宙观的范畴, 后三者(气理、心意、神)则意味着人的精神同宇宙之间的微妙联系^[18]。能够将宇宙和精神之间的连接变成可见对象且投注思考的就是山水画了。可见, 它之所以能被成功地提出, 满足了由外向的再现造化之磅礴生命转至内敛的感应自然的寓兴的功能^[19]。由于山水观念扎根于中国文化的土壤, 即便其发生一些变化, 也是风格意义上的, 其作为绘画母题是不变的。在山水变迁中, 张彦远认为“山水之变, 始于吴(道子), 成于二李(思训、昭道)。”^[20]郭若虚说:“画山水惟营丘李成、长安关仝、华原范宽, 智妙入神, 才高出类, 三家鼎峙百代, 标程前古。虽有传世可见者如王维、李思训、荆浩之伦, 岂能方驾。”^[21]王世贞说:“山水至大小李一变也, 荆关董巨又一变也, 李成范宽又一变也, 刘李马夏又一变也, 大痴黄鹤又一变也。”^[22]上述三条文献对于“山水”的关注分别出自 9、11、16 世纪的画史家之手, 而对其他绘画科目诸如人物、花鸟等鲜有这种持续的兴趣。石守谦认为, 这些史家“观察角度、关怀重点或有些差异, 不过, 值得注意的则是他们对于山水之史的叙述, 都表达了一种高度的兴趣。”^[23]这反映了山水观念持续的活力和人们对其探索的热情, 这是山水观念成为中国绘画基因观念的原因。

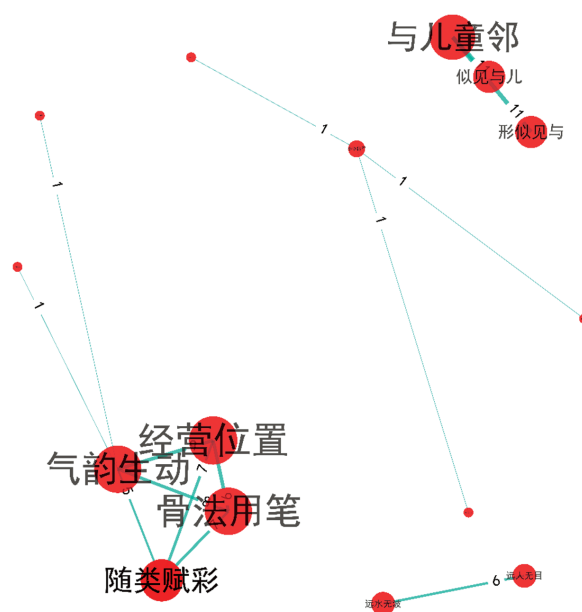


图 7 四字词基因观念排名前 20 的共词网络

词语由单字向多字的不断演变中, 其意义指向越来越明确, 对于绘画来说, 演变而成的四字词形成了绘画的指导性和体系建构。从图 7 共词网络中排名靠前的 20 个观念基因中可以看到, “骨法用笔”“气韵生动”“经营位置”“随类赋彩”这几个词经常出现, 占据主要位置。它们是谢赫提出的绘画六法中的重要组成。谢赫说“夫画品者, 盖众画之优劣也。图绘者, 莫不明劝戒,

著升沉，千载寂寥，披图可鉴。虽画有六法，罕能尽该。而自古及今，各善一节。六法者何？一气韵生动是也，二骨法用笔是也，三应物象形是也，四随类赋彩是也，五经营位置是也，六传移模写是也。”^[24] 共词网络中“应物象形”和“传移模写”没有处于显著位置，是由于中国绘画从“形似”的路子转向了“传神”，绘画的主体或者说为绘画张目的主体从画工转向了文人，文人画渐趋发达。董其昌的南北分宗是文人画发展的结果。“对于文人画家来说，就像苏东坡的老师欧阳修所称的，想处理空间、高度、距离的策略都属于‘画工之艺’。文人画家关心的不是山水的实物实景。”^[25] 因此表达真实山水的“应物象形”和“传移模写”也就不再那么重要。正如倪瓚说的“余之竹聊以写胸中逸气耳，岂复较其似与非，叶之繁与疏，枝之斜与直哉！或涂抹久之，他人视以为麻为芦，仆亦不能强辨为竹。真没奈览者何，但不知以中视为何物耳？”^[26] 至于“骨法用笔”“气韵生动”“经营位置”和“随类赋彩”，因构成了无论是文人画或院体画的普遍要求，所以被保留下来，这可以视为中国绘画的法则。“远水无波”“远人无目”来自于王维《山水论》“凡画山水，意在笔先。丈山尺树，寸马分人。远人无目，远树无枝。远山无石，隐隐如眉。远水无波，高与云齐，此是诀也。”^[27] 后来荆浩在《山水赋》中也提到了“远人无目，远树无枝，远山无皴，高与云齐，远水无波，隐隐似有，此其式也。”^[28] 王维和荆浩具体化了山水画的绘制技巧，可以被视为中国绘画的技法传授。“与儿童邻”“似见与儿”“形似见与”尽管四字词缺乏独立的意义，但它们的语境被保留下来，是苏轼在《鄢陵王主簿所画折枝二诗》中的句子，“论画以形似，见与儿童邻。赋诗必此诗，定知非诗人。诗画本一律，天工与清新。边鸾雀写生，赵昌花传神。何如此两幅，疏淡含精匀。谁言一点红，解寄无边春。瘦竹如幽人，幽花如处女。低昂枝上雀，摇荡花间雨。双翎决将起，众叶纷自举。可怜采花蜂，清蜜寄两股。若人富天巧，春色入豪楮。悬知君能诗，寄声求妙语。”^[29] 苏轼诗歌讨论了绘画的形似与神似，他更重视传神的作品，这种讨论实际上是回应了文人画和院体画两个方向，之后的绘画主流被文人画占领。从四字词的共词网络中可以看出，这些词反映了中国绘画法则、中国绘画方向和中国绘画技法，它们作为观念基因是合理的。

3 结 语

本文将 BLAST 方法、库恩基因观念测度以及共词网络分析应用于中国画论基因观念的研究，是数字人文在中国画论方面的实践和尝试。研究发现，“用笔”“山水”“自然”“气韵生动”“骨法用笔”“经营位置”等词项贯穿整个中国画论，是中国画论的核心观念。在研究过程中也得到一些结论：

(1) 改进的库恩算法能够有效捕捉具有时间效应的基因观念，这对获取具有历史意涵的文化基因观念有很大帮助。

(2) 采用共词网络分析获得了基因观念的语义关联性，使得库恩算法中一些离散排序的基因获得了有效关联，更能帮助理解绘画的整体关照。

(3) 二字词中的“山水”“自然”“气韵”等基因观念有效地将中国文化的虚实、阴阳、天地、气理、心意、神这些概念连接起来，表达了宇宙客体和人的主体的微妙联系，实现了天地人三才的视觉化表达。

(4) 四字词基因观念“骨法用笔”“气韵生动”“经营位置”“随类赋彩”(谢赫),“与儿童邻”“似见与儿”“形似见与”(苏东坡),“远水无波”“远人无目”(王维),反映了中国绘画的法则、文人画方向和山水画技法。

(5) 改进的库恩算法具有方法论推广的意义,适用于寻找长段历史中的基因观念。如果某种文化的语料库完整、时间跨度长、语料库中的文献存在着引用关系,那么即可对该语料库进行文化基因识别,诸如哲学史中的观念演变、中国书法中的观念演变以及儒释道等具有长段历史文献的观念演变等。

本研究还存在诸多的不足,主要包括以下两个方面:

(1) 由于古代文献写作不像现代文献有明确的引文规范,本研究的引文网是依照文本序列的相似度对比得出引用关系的,缺乏画论之间的直接引用。

(2) 不同时期的画论数量差异悬殊,这使得数据稀疏成为研究的一大难题,后续可以根据不同历史时期的画论数量给予一定的权重,以增加研究的可靠性。

【参考文献】

- [1] Dawkins R. The selfish gene [M]. New York: Oxford University Press, 1976: 360.
- [2] 文庭孝, 罗贤春, 刘晓英, 等. 知识单元研究述评 [J]. 中国图书馆学报, 2011, 37 (5) : 75-86.
- [3] 梁镇涛, 毛进, 操玉杰, 等. 基于知识基因级联网络的领域知识扩散模式分析 [J]. 情报理论与实践, 2020, 43 (4) : 40-46.
- [4] 孙晓玲, 丁堃. 基于知识基因发现的科学与技术关系研究 [J]. 情报理论与实践, 2017, 40 (6) : 23-26, 17.
- [5] 金观涛, 刘青峰. 隐藏在关键词中的历史世界 [J]. 东亚观念史集刊, 2011 (1) : 55-83.
- [6] Vierthaler P, Gelein M. A BLAST-based, language-agnostic text reuse algorithm with a MARKUS implementation and sequence alignment optimized for large Chinese corpora [J]. Journal of Cultural Analytics, March 18, 2019: 1-19.
- [7] Funk, Kellen, Mullen, et al. The spine of American law: Digital text analysis and U.S. legal practice [J]. The American Historical Review, 2018, 123(1): 132-164.
- [8] Stephen F Altschul, et al. Basic local alignment search tool [J]. Journal of Molecular Biology, 1990, 215(3): 403-10.
- [9] Linder F, Desmarais B, Burgess M, et al. Text as policy: Measuring policy similarity through Bill Text Reuse [J]. Policy Studies Journal, 2018(2): 1-29.
- [10] 王振振, 何明, 杜永萍. 基于LDA主题模型的文本相似度计算 [J]. 计算机科学, 2013, 40 (12) : 229-232.
- [11] Tobias Kuhn, Matjaž Perc, Dirk Helbing. Inheritance patterns in citation networks reveal scientific memes [J]. Physical Review X, 2014, 4(4): 041036.
- [12] 逯万辉, 谭宗颖. 基于知识基因游离与重组的领域主题演化研究 [J]. 情报理论与实践, 2019, 42 (2) : 101-107.
- [13] 钮亮. 《论语》的观念结构: 一个数字人文的视角 [J]. 图书馆论坛, 2021, 41 (2) : 67-76.
- [14] 俞剑华. 中国历代画论大观: 第一编 先秦至五代画论 [M]. 南京: 江苏凤凰美术出版社, 2015: 1-5.
- [15] 李久平, 姚乐野. 知识管理文献增长模型研究 [J]. 图书馆理论与实践, 2012 (5) : 36-39.
- [16] 梁启超. 中国历史研究法(外二种) [M]. 石家庄: 河北教育出版社, 2000: 350-351.

- [17] 滕固. 中国美术小史 唐宋绘画史 [M]. 上海: 上海书画出版社, 2016: 187.
- [18] 扎瓦茨卡娅. 外国学者论中国画 [M]. 长沙: 湖南美术出版社, 1986: 64-80.
- [19] 洪再辛. 海外中国画研究文选 [M]. 上海: 上海人民美术出版社, 1992: 250.
- [20] 俞剑华. 中国历代画论大观: 第一编 先秦至五代画论 [M]. 南京: 江苏凤凰美术出版社, 2015: 117-118.
- [21] 俞剑华. 中国历代画论大观: 第二编 宋代画论 [M]. 南京: 江苏凤凰美术出版社, 2015: 18-19.
- [22] 俞剑华. 中国历代画论大观: 第四编 明代画论 [M]. 南京: 江苏凤凰美术出版社, 2015: 89.
- [23] 石守谦. 中国山水画和观众的历史 [M]. 上海: 上海书画出版社, 2019: 1.
- [24] 俞剑华. 中国历代画论大观: 第一编 先秦至五代画论 [M]. 南京: 江苏凤凰美术出版社, 2015: 62.
- [25] 高居翰. 李渝 译. 图说中国绘画史 [M]. 北京: 生活·读书·新知三联书店, 2014: 130.
- [26] 俞剑华. 中国历代画论大观: 第三编 元代画论 [M]. 南京: 江苏凤凰美术出版社, 2015: 144-153.
- [27] 俞剑华. 中国历代画论大观: 第一编 先秦至五代画论 [M]. 南京: 江苏凤凰美术出版社, 2015: 155-156.
- [28] 俞剑华. 中国历代画论大观: 第一编 先秦至五代画论 [M]. 南京: 江苏凤凰美术出版社, 2015: 160-161.
- [29] 俞剑华. 中国历代画论大观: 第二编 宋代画论 [M]. 南京: 江苏凤凰美术出版社, 2015: 212-214.

Research on the Identification of Genetic Concepts in Chinese Painting Based on Digital Technology

Niu Liang¹ Xu Jiajun¹ Xiang Wei²

(1. College of Economics & Management, China Jiliang University, Hangzhou 311222, China;
2. School of Art, Zhejiang International Studies University, Hangzhou 311222, China)

Abstract: [**Purpose/significance**] Genetic concepts are an important basis for determining the direction of culture, and identifying the genetic concepts of Chinese painting theory is important for understanding the aesthetic orientation of Chinese painting and its positioning in the overall Eastern art. [**Method/process**] In this study, the BLAST method was first used to construct a Chinese painting thesis text citation network, then the painting theory gene ideas were extracted based on the improved Kuhn gene measure, and finally the semantic links between the gene ideas were examined through the co-word network. [**Result/conclusion**] From the results of the study, the six methods advocated by Xie He, the landscape techniques advocated by Wang Wei, and the literati painting advocated by Su Shi are important genetic concepts in the development of Chinese painting. This study provides an exploration for quantitative identification of cultural genes.

Keywords: BLAST; Kuhn gene measure; Chinese painting theory; Genetic concepts

(本文责编: 王秀玲)