

# 融合多维语义模型的中西书画术语词表 对齐算法\*

钮亮

(中国计量大学管理科学与工程学院, 杭州 310018)

**摘要:** [目的/意义] 为纾解中国书画术语跨语言检索的“孤岛”困境, 针对中西方文化差异与词表结构不兼容的问题, 提出术语精准对齐的算法模型。[方法/过程] 构建融合文本相似度、语义相似度、领域权重与文化适配系数的多维模型, 通过四维语义拆解、动态优先级规则设计与映射关系判定, 实现中西书画术语词表的精准匹配与概念结构对齐。[结果/结论] 融合文本、语义、领域、文化差异的多维模型, 整体性能较强, 可以显著提升术语匹配的准确率, 为书画术语跨语言检索及多语言知识组织系统构建提供借鉴。

**关键词:** 概念结构对齐 语义拆解 多维度融合 数字人文

**分类号:** G254.24

**DOI:** 10.31193/SSAP.J.ISSN.2096-6695.2026.01.06

## 0 引言

文化遗产的数字化与全球化进程, 对中国书画这类富含东方哲学意蕴的艺术资源提出了跨语言精准检索的迫切需求。然而, 中西方在文化认知与分类逻辑上的根本性差异, 导致其核心术语在跨语言知识组织中形成了典型的“孤岛”现象, 也对实现有效的概念对齐带来了巨大挑战。研究表明, 中国书画核心术语在中西方词表间存在语义断层, 致使跨系统数据传递的准确率不足50%<sup>[1-2]</sup>。这一困境源自中西方文化深层次的认知差异。中文术语(如“意境”“气韵”)深刻内化了道家“虚实观”、儒家“中和之美”等哲学思想, 而西方术语则侧重于技术与形式层面的客观描述<sup>[3]</sup>。例如, 西方《艺术与建筑主题词表》(*Art & Architecture Thesaurus*, 简称 AAT) 对中国绘画中“白描”的定义止步于描述其“单色线条”的技术特征, 而忽略了其“以线传神”的美学内核, “意境”等核心概念甚至未被收录<sup>[4]</sup>。在此背景下, 如何在承认文化差异的前提下, 实现中西书画术语词表概念结构的最大对齐, 从而支撑跨语言检索, 成为一个关键且亟待解决的

\* 本文系浙江省社科联研究项目“中国画图像志受控词中英语义对应研究”(项目编号: 2025N119)的研究成果之一。

[作者简介] 钮亮, 男, 副教授, 研究方向为数字人文、知识组织, Email: niutyut@126.com。

问题。

为破解此难题,本研究致力于构建一套算法驱动的概念结构对齐框架,以实现中西书画术语词表的概念对接。为便于研究,本文选取《台北故宫博物院中国书画受控词汇》(*Taipei Palace Museum Calligraphy-and-Painting Controlled Vocabulary*,简称TPM-CPCV)作为中文术语来源,以AAT作为西方术语来源。TPM-CPCV是台北故宫博物院基于其中国书画藏品开发的专门术语集,约含2000个词条,深度涵盖中国书画特有的技法、主题与材质,如“皴法”“草虫”“端石”等,具有高度的文化专指性和学术权威性。AAT是由美国盖蒂研究中心开发的国际通用艺术术语索引典,收录约35000个概念,结构完整、层级清晰,是西方艺术领域标准的知识组织工具。因此,这两个词表能够较好地代表中国书画术语体系与西方艺术术语体系。通过实现TPM-CPCV与AAT的概念对齐,验证本研究框架在跨语言精准检索中的有效性。

本研究的核心路径为:首先,构建一个融合文本、语义、领域权重与文化适配系数的多维度量化模型,将抽象的“文化关联”转化为可计算的“文化适配系数”。在此基础上,设计一种基于“动因—目的—构成—形式”的四维语义拆解框架,对TPM-CPCV中融合的多维语义进行结构化解析,以应对其语义模糊性与文化密集性。进而,提出一套基于动态优先级的匹配类型判定规则,系统化处理术语间复杂的映射关系,提升对齐结果的可解释性与可用性。本研究的创新之处与价值主要体现在:针对中国书画术语文化内涵丰富的特点,尝试构建融合文化适配系数的多维度对齐模型,推动跨语言概念对齐从“经验判断”向“量化计算”的方法论转型;研究成果可应用于博物馆、图书馆、数字档案馆等机构的跨语言检索系统,提升中国书画资源的国际可见性与检索精度;通过系统对比TPM-CPCV与AAT的概念结构,初步揭示两者在文化表达与分类逻辑上的差异,为后续相关词表的修订、扩展或跨系统互联提供数据支持和分析依据。

## 1 相关研究

### 1.1 跨语言概念与词汇等同性研究的技术演进

跨语言概念与词汇等同性是术语对齐的核心理论基础,其发展围绕“概念关联方式”与“技术支撑体系”,形成三阶段演进特征。

#### 1.1.1 传统知识组织理论阶段

本阶段以人工构建结构化知识体系为核心,致力于通过厘清术语间的逻辑关系来建立跨语言语义关联,其贡献在于确立了“概念优先于语言符号”的根本原则。Madsen等<sup>[5]</sup>首次提出跨语言术语控制机制,倡导通过“概念—术语”的双向映射以消减语义歧义,其“先定义脱离具体语言的抽象概念,再匹配多语言术语”的思想,为多语言叙词表的构建奠定了理论基础。Soergel<sup>[6]</sup>进而提出“概念—术语分离模型”,明确将概念界定为“反映事物本质属性的抽象实体”,而术语仅是“概念的语言载体”,从方法论上为文化特异性术语的跨语言定义提供了支撑。Larson<sup>[7]</sup>构建的“概念内核→语言符号”跨层映射模型将研究焦点从表层的“字面匹配”转向深层的“语义本质匹配”,强调需先行提取术语的核心语义属性,再进行跨语言映射,为跨语言语义对齐提供了重要的方法论基础。

### 1.1.2 计算语言学转向阶段

进入 21 世纪, 神经网络与机器学习技术推动该领域步入“自动化模型构建”阶段, 其核心突破在于实现了语义相似度的量化计算。Mikolov 等<sup>[8]</sup>提出的跨语言词嵌入对齐范式, 通过将不同语言词汇映射至统一的高维向量空间, 并计算其欧氏距离或余弦相似度(例如, “竹”与“bamboo”的向量距离趋近于 0), 为大规模术语的批量匹配开辟了可行路径。然而, 语言向量空间固有的差异导致语义偏移。为此, Artetxe 等<sup>[9]</sup>提出了“正交约束映射算法”, 通过约束不同语言向量空间之间的正交关系, 将低频词汇及文化特异性术语的匹配误差降低了约 20%。Ruder 等<sup>[10]</sup>在对 39 种主流跨语言嵌入模型进行系统评述后指出, 基于通用语料训练的模型在“低资源语言”与“文化专属领域”的适配性上存在明显缺陷, 对领域专属术语的语义捕捉准确率不足 50%, 其根本原因在于未能有效融入领域知识与文化特征。由此可见, 该阶段虽显著提升了自动化效率, 但在文化适配性方面仍存在固有局限。

### 1.1.3 神经符号融合阶段

近年来, 为了克服纯数据驱动模型的不足, 研究进入了神经符号融合阶段, 旨在通过融合知识图谱的符号逻辑与神经网络的语义表示, 提升复杂领域术语的匹配精度。Navigli 等<sup>[11]</sup>构建的 BabelNet 是此方向的早期典范, 它构建了“概念—语义—文化”三层框架, 首先利用知识图谱梳理术语的层级关系与属性信息, 再结合词嵌入模型计算语义相似度, 从而有效提升了复杂场景下的术语对齐精度。Giunchiglia 等<sup>[12]</sup>进一步提出了“零样本跨语言概念映射”方法, 通过迁移学习利用高资源语言的概念知识, 直接预测低资源领域(如中国书画)术语的向量表示, 实现了在缺乏大量双语语料下的关联映射。这一阶段的研究, 标志着该领域从依赖海量数据的浅层匹配, 向融合先验知识与深度语义理解的精准对齐转变。

## 1.2 中国书画领域术语对齐研究现状

聚焦于中国书画领域的特殊性, 国内学者已开展了一系列针对性探索, 这些研究在取得局部进展的同时, 也暴露出该领域在规模化与精准化对齐方面所面临的共性挑战。在针对中西书画术语对齐的探索中, Chen 等<sup>[13]</sup>从批判性视角揭示了 AAT 所内嵌的“西方艺术中心主义”局限, 指出其“形式+功能”的分类逻辑与中国书画“文化内涵优先”的认知逻辑存在根本性冲突, 主张通过“本体—词表桥接”策略实现跨系统互通。这一思路颇具启发性, 但未能进一步细化出普适性的桥接规则与标准化映射机制, 致使该方案在实践层面缺乏可操作性。在实践层面, 张俊娥<sup>[14]</sup>对 AAT 中文化进行了系统研究, 其工作重点在于解决“皴法”“描法”等关键技术术语的翻译等效性问题。此项研究为跨语言检索提供了基础的词汇对照, 但其着力点停留在语言表层, 未能深入概念结构层面进行深度对齐, 因而无法从根本上解决因文化差异导致的深层语义偏差问题。张萌<sup>[15]</sup>认为中国书画术语在跨语言标注中普遍存在“语义归属混乱”现象。该研究敏锐地揭示了问题表征, 为理解现有系统的不足提供了重要依据, 但未能提出一套系统的、可计算的解决方案来根治这一顽疾。

综上所述, 现有研究虽已从本体构建、理论批判、语言翻译与问题诊断等多个维度为中西书画术语对齐奠定了基础, 但仍普遍存在以下局限: 或偏重理论模型而轻实践推广, 或局限于表层翻译而缺乏深层语义对齐, 或精于问题剖析而疏于算法实现。这些问题构成了本研究的出发点。

## 2 核心算法体系设计

为实现 TPM-CPCV 术语与 AAT 概念间的精准对齐，本研究构建了一套层次化的核心算法体系，通过确认英译词汇、多维度融合计算与匹配类型判定几个关键环节，解决因文化差异与结构不兼容导致的对齐难题。核心算法流程如图 1 所示。

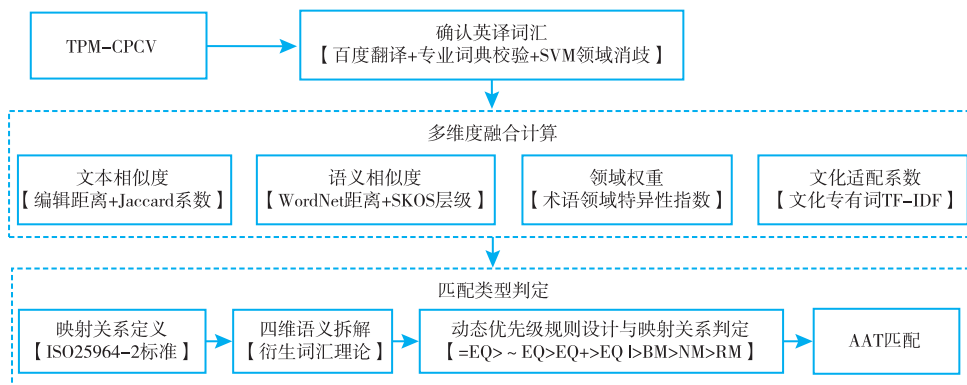


图 1 核心算法流程图

### 2.1 确认英译词汇

为建立跨语言匹配的桥梁，首先需将 TPM-CPCV 转化为标准英文表述。本研究调用百度翻译 API，生成 3~5 个候选英译，并融合《英汉汉英艺术词典》等专业词典进行校验。通过“专家投票 + 领域适配度评分”机制确定最优译法。针对一词多义现象，采用基于 WordNet3.0 同义词集<sup>[16]</sup>与支持向量机 (Support Vector Machine, SVM) 领域分类算法，过滤无关领域含义。例如，“山水”的候选含义包括“landscape painting” (书画领域)、“geography” (地理领域) 等，SVM 模型可精准判定“landscape painting”为目标含义，为后续计算奠定基础。

### 2.2 多维度融合计算

在完成 TPM-CPCV 的英译以后，构建加权综合评分模型。该模型通过集成文本相似度、语义相似度、领域权重与文化适配系数，为每个 TPM-CPCV 术语与其 AAT 候选概念计算出一个综合的量化匹配度 (Similarity)，作为后续实现自动化匹配判定的核心依据，多数术语可以通过该模型实现匹配。计算公式如下：

$$Similarity = \alpha \times Text\_Sim + \beta \times Semantic\_Sim + \gamma \times Domain\_Weight + \delta \times Cultural\_Adapt \quad (1)$$

其中， $\alpha$ 、 $\beta$ 、 $\gamma$ 、 $\delta$  为特征权重，通过层次分析法由领域专家确定； $\alpha = 0.3$  (文本相似度)， $\beta = 0.3$  (语义相似度)， $\gamma \in [0.2, 0.4]$  (领域权重)， $\delta \in [0.05, 0.15]$  (文化适配系数)<sup>[17]</sup>。各维度计算方式如下：

(1) 文本相似度 (Text\_Sim)：将编辑距离归一化为相似度值，再与 Jaccard 系数进行平均计算。

① 编辑相似度 (Edit\_Sim): 编辑距离 (EditDistance) 精确量化两个字符串的“绝对差异”(需要编辑多少次), 在这里体现为计算术语英译与 AAT 概念标签的字符修改次数, 然后将其归一化为相似度值, 通常采用公式为:

$$Edit\_Sim = 1 - \frac{EditDistance(s1,s2)}{\max(len(s1),len(s2))} \quad (2)$$

其中, len() 是一个函数, 用于计算字符串的长度 (即包含多少个字符)。编辑相似度取值在 [0,1] 区间, 数值越大表示相似度越高。例如, “竹” 英译 “bamboo” 与 AAT “bamboo(plants)[300311500]” 的编辑距离为 8, 算法自动剔除 “(plants)” 等通用领域后缀, 修正后 Edit\_Sim=1。

② Jaccard 系数: 计算关键词交集与并集的比例, 取值在 [0,1] 区间, 数值越大表示越相似。公式为:

$$Jaccard\_sim = \frac{s1 \cap s2}{s1 \cup s2} \quad (3)$$

例如, “后妃” 英译 “empresses and concubines” 与 AAT “empresses [300150492]” 的关键词交集为 {empresses}, 并集为 {empresses,concubines}, Jaccard\_sim=0.5。

(2) 语义相似度 (Semantic\_Sim): 结合 WordNet 语义距离与 AAT 的 SKOS 层级关系, 评估深层语义关联, 语义相似度取语义距离与关联强度均值。

① 语义距离 (WordNet\_Sim): 基于 WordNet 3.0, 计算 TPM-CPCV 术语与 AAT 概念的语义距离。公式为:

$$WordNet\_Sim = \frac{1}{1 + distance} \quad (4)$$

例如, “松” 与 AAT “pine trees (plants)[300174989]” 同属一个同义词集, 距离 =0, WordNet\_Sim=1.0。

② 关联强度 (SKOS\_Sim): 基于 AAT 的 SKOS 层级关系, 计算层级距离与关联强度。见表 1。例如, “高古游丝描” 与 AAT “line drawing [300100194]” 为 broader 关系 (上位—下位), 层级距离 =1, 关联强度 =0.9, SKOS\_Sim=0.9。

表 1 SKOS 层级距离与关联强度映射

| SKOS 关系类型 | 层级距离 | 关联强度 |
|-----------|------|------|
| 等同        | 0    | 1.0  |
| 直接上位      | 1    | 0.9  |
| 间接上位      | 2    | 0.7  |
| 关联        | 1    | 0.6  |

(3) 领域权重 (Domain\_Weight): 采用 “术语领域特异性指数” (DSI) 量化术语的领域专属程度。

$$DSI = \frac{Freq_{art}}{Freq_{art} + Freq_{general}} \quad (5)$$

基于艺术领域语料库与通用语料库统计频次。例如，“皴法”在艺术领域出现500次，通用领域出现10次， $DSI=500/(500+10) \approx 0.98$ ， $\gamma=0.4$ ；“竹”在艺术领域出现800次，通用领域出现2000次， $DSI=800/(800+2000) \approx 0.29$ ， $\gamma=0.2$ 。领域权重取值公式为：

$$Domain\_Weight = \gamma_{min} + (\gamma_{max} - \gamma_{min}) \times DSI \quad (6)$$

(4) 文化适配系数 (Cultural\_Adapt)：构建“中国书画文化专属词表”（含“意境”“气韵”“笔墨”“天人合一”等200个文化专属词，源自《中国美术大辞典》<sup>[18]</sup>与《中国美学范畴辞典》<sup>[19]</sup>，基于TF-IDF计算术语的文化词频次与总语义词频次，据此计算其文化适配系数。公式为：

$$Cultural\_Adapt = 0.05 + 0.1 \times \frac{Cultural\_Freq}{Total\_Freq} \quad (7)$$

例如，“意境”的文化词频次=3（含“情景交融”“虚实相生”“道家思想”），总语义词频次=10， $Cultural\_Adapt=0.05+0.1 \times 3/10=0.08$ 。因“意境”为强文化专属术语，调整  $Cultural\_Adapt=0.15$ 。“竹”的文化词频次=0， $Cultural\_Adapt=0.05$ 。

### 2.3 匹配类型判定

通过映射关系定义、四维语义拆解、动态优先级规则设计与映射关系判定，实现TPM-CPCV与AAT术语匹配类型的精准判定。

#### 2.3.1 映射关系定义

由于TPM-CPCV与AAT术语的匹配需覆盖“完全等同”“部分关联”“层级包含”“复合构成”等多类场景，根据ISO25964-2标准，并针对书画术语特性，定义七种映射关系，明确各类关系的适用范围与语义逻辑，为后续判定提供分类依据，如表2所示。

表2 七种映射关系

| 代码  | 对应类型   | 适用情况  |
|-----|--------|---|
| =EQ | 精准等同   | 不同词表中两个概念的范围完全一致，语义无偏差，在跨语言应用中可双向互换，且无需额外语义修正                   |
| ~EQ | 近似等同   | 概念核心语义一致，但范围存在部分重叠或文化表述差异，仅在特定场景下可互换，需标注语义偏差点                   |
| EQ+ | 交叉复合等同 | TPM-CPCV中某一复杂术语，语义等同于AAT中两个及以上概念的交集，需通过多概念组合实现完整匹配，映射方向不可逆      |
| EQ  | 并列复合等同 | TPM-CPCV中某一术语，语义涵盖AAT中两个及以上不重叠的同级概念，需通过多概念并集实现全覆盖，映射方向不可逆       |
| BM  | 上位映射   | TPM-CPCV术语为下位概念，其语义范围完全包含于AAT某一上位概念中，需通过层级关联实现匹配，支持从细分概念向上位概念检索 |
| NM  | 下位映射   | TPM-CPCV术语为上位概念，其语义范围完全包含AAT某一下位概念，需通过层级关联补充细分信息，支持从整体概念向下位概念检索 |

续表

| 代码 | 对应类型 | 适用情况   |
|----|------|--|
| RM | 关联映射 | 两类概念无等同或层级包含关系, 但在语义 (如文化背景、应用场景) 上存在关联, 仅能通过间接关联实现匹配, 需标注关联依据 |

### 2.3.2 四维语义拆解

由于文化上的差异, TPM-CPCV 的某些术语在与 AAT 术语的对齐中存在语义模糊、文化内涵密集的难题, 本研究基于 Pustejovsky 的衍生词汇理论<sup>[20]</sup>, 设计了四维语义框架, 将术语的综合性语义拆解为动因、目的、构成与形式四个维度。①动因维度: 解析艺术表现的根本意图与哲学思想。②目的维度: 阐释希望达成的审美功能与文化象征。③构成维度: 解构具体的技术要素与物质材料。④形式维度: 描述典型的视觉样式与形态特征。通过此框架, 术语被转化为一组结构化的语义描述 (示例见表 3), 为后续的多维度融合计算与复杂映射关系判定提供了结构化语义基础。对于少数因文化内涵高度独特或量化分数处于阈值边界而难以自动判定的复杂术语, 可基于该理论的属性结构分析, 为其提供深层的理论解释与结构化释义, 辅助完成最终映射。

表 3 四维语义拆解示例

| TPM-CPCV 术语 | 动因维度             | 目的维度                 | 构成维度                        | 形式维度                   |
|-------------|------------------|----------------------|-----------------------------|------------------------|
| 工笔          | 精细描绘物象形态, 还原客观真实 | 展现物象的细节美感与质感层次       | 使用细笔 (狼毫小楷)、重彩 (矿物颜料), 层层渲染 | 线条细腻均匀, 色彩浓艳饱和, 构图严谨对称 |
| 意境          | 传达画家的精神感悟与自然体验   | 实现 “情景交融、虚实相生” 的审美境界 | 依赖笔墨技法、构图布局, 融入文化思想         | 无形的精神氛围, 虚实结合, 情景交融    |

### 2.3.3 动态优先级规则设计与映射关系判定

由于 7 种映射关系的语义关联强度与判定优先级存在差异, 若直接按统一阈值判定, 易出现 “类型混淆” (如将 “近似等同” 误判为 “精准等同”)。结合 TPM-CPCV 术语匹配实际需求, 依据 “语义关联强度从高到低” 的原则, 设计动态优先级规则 =EQ > ~EQ > EQ+ > EQ | > BM > NM > RM。在完成术语的语义解析后, 对齐流程进入核心的计算与自动判定阶段。首先, 对每个 TPM-CPCV 术语执行 “多维度融合计算”, 通过集成文本、语义、领域及文化四维特征, 为每个术语生成综合量化匹配度 (Similarity) 及文化适配系数 (Cultural\_Adapt) 等关键指标。随后, 系统遵循上述动态优先级规则与表 4 所列的映射关系量化判定条件, 自动将每个术语归类至特定映射类型 (如 “精准等同” =EQ、“上位映射” BM 等), 完成对应。该规则的核心逻辑为: 优先判定语义关联最紧密、无歧义的类型, 再逐步判定关联强度较弱或结构更复杂的类型。例如, 若某一 TPM-CPCV 术语同时满足 “=EQ” 与 “~EQ” 的部分条件, 因 “=EQ” 代表语义完全一致, 优先级高于 “~EQ”, 故优先判定为 “=EQ”; 仅当不满

足高优先级类型的判定条件时，才进入低优先级类型的判定流程，避免因判定顺序混乱导致结果偏差。

表4 映射关系量化判定条件

| 代码  | 核心判定条件   | 语义依据  |
|-----|--|---|
| =EQ | Similarity ≥ 0.95 且 Cultural_Adapt ≤ 0.05          | 四维特征均高度一致，无文化偏差。例：竹→bamboo                                |
| -EQ | 0.85 ≤ Similarity < 0.95 且 Cultural_Adapt ≥ 0.1    | 核心维度重叠，但文化动因导致范围偏差。例：后妃→empresses                         |
| EQ+ | 与 ≥ 2 个 AAT 概念组合后，整体 Similarity ≥ 0.80             | 语义被拆解为多个独立维度。例：春景山水→spring+landscapes                     |
| EQ↓ | 与 ≥ 2 个同级 AAT 概念并集后，整体 Similarity ≥ 0.80           | 构成维度包含多个并列子类。例：梅(白·红·腊梅)→多品种并集                            |
| BM  | 0.70 ≤ Similarity < 0.85, 且 AAT 为 TPM-CPCV 的明确上位概念 | TPM-CPCV 是 AAT 的具体表现形式，但文化内涵更丰富。例：高古游丝描→linedrawing       |
| NM  | 0.70 ≤ Similarity < 0.85, 且 TPM-CPCV 为 AAT 的明确上位概念 | AAT 仅是 TPM-CPCV 宽泛构成下的一个具体子类。例：花鸟主题→peony and bird themes |
| RM  | 0.5 ≤ Similarity < 0.70 且 Cultural_Adapt ≥ 0.15    | 核心文化语义(动因/目的)在 AAT 中无直接对应。例：意境→aesthetics                 |

### 3 TPM-CPCV 与 AAT 概念匹配及验证

#### 3.1 数据来源

(1) TPM-CPCV: 源自台北故宫博物院《元数据需求规范(书画藏品)》，术语筛选范围参考陈淑君<sup>[21]</sup>对中国书画词汇的界定，经“关键词匹配+SVM分类+专家验证(Kappa=0.87)”，人工筛选出581个书画术语(主题543个、技法38个)，部分内容见表5。

(2) AAT数据: 提取美国盖蒂研究所AAT中聚焦与中国书画关联度最高的三个面“活动”“风格与时期”“材料”概念，转换为JSON格式构建候选库。

(3) 辅助数据: 为支撑领域权重、文化适配等维度的计算，构建了专门的艺术领域语料库，涵盖《中国美术大辞典》《历代名画记》《画继》等中文典籍，以及AAT术语定义、权威艺术史英文文献等，用于相关统计与分析。

表5 TPM-CPCV 书画术语分类统计

| 部分分类 | 术语数量(个) | 部分代表性词汇                         |
|------|---------|---------------------------------|
| 自然景观 | 20      | 山水、春景、冬景(雪景)、雨景、云、江河湖海、瀑布、石壁岩洞  |
| 人物   | 31      | 帝王、后妃、高士(士人隐士)、仕女、孩童、农夫、渔夫船夫、行旅 |
| 佛道人物 | 22      | 佛、菩萨、观音、罗汉(应真尊者)、神仙、僧(和尚尼姑)     |
| 树木   | 25      | 松、竹、杨柳、柏、芭蕉、梧桐、银杏、菩提            |
| 花草   | 86      | 梅(白红腊梅)、兰薰、菊、牡丹、芍药、莲荷、海棠        |
| 果蔬   | 46      | 梨、柿、樱桃、枣、苹果、石榴、笋、荔枝、莲蓬          |
| 五谷   | 7       | 稻、麦、玉米、稷、小米、高粱                  |

续表

| 部分分类 | 术语数量(个) | 部分代表性词汇                         |
|------|---------|---------------------------------|
| 走兽   | 28      | 马、牛、羊、虎、鹿、龙、蛇、猴、骆驼              |
| 翎毛   | 55      | 鹰、鹤、喜鹊、雁、麻雀、鸳鸯、孔雀、雉(锦鸡雉类)       |
| 草虫   | 19      | 蝶蛾、蟋蟀、蜻蜓、蝉、蜂、蜘蛛、蚂蚁、蝗            |
| 建筑   | 48      | 宫殿、台阁、亭、桥、房舍、寺庙、塔、庭院、茅草屋        |
| 器用   | 55      | 文房用具、乐器、家具(屏风)、饮食器、花器、扇、灯笼、香炉火盆 |
| 节令   | 16      | 年节、清明、中秋、麒麟、凤凰、蝙蝠、辟邪            |
| 基础技法 | 13      | 白描、工笔、写意、界画、没骨、双钩、指画、皴法         |
| 细分技法 | 25      | 披麻皴、斧劈皴、高古游丝描、铁线描、钉头鼠尾描、柳叶描     |
| 总数   | 496     |                                 |

### 3.2 匹配结果

基于公式(1)所示的多维度融合计算模型与映射关系量化判定规则(表4),对581个TPM-CPCV术语进行自动化对齐计算与映射判定。最终,各术语对均被判定为7种映射关系(=EQ、~EQ、EQ+、EQ|、BM、NM、RM)之一,并输出对应的量化匹配度(Similarity)与文化适配系数(Cultural\_Adapt)等关键指标。表5所示的术语分类中,共计产生175个“=EQ”、107个“~EQ”、13个“EQ+”、15个“EQ|”、71个“BM”、173个“NM”及27个“RM”匹配对。

### 3.3 模型性能与匹配效果验证

为系统验证模型性能与匹配效果,本研究从整体性能、细分类型适配性、核心模块有效性三个维度展开评估。在不同维度衡量匹配效果时,采用准确率(正确匹配术语数/总术语数)与Kappa系数(衡量模型结果与专家人工标注结果的一致性,Kappa ≥ 0.8视为“极强一致”)两类量化指标。

#### 3.3.1 整体性能对比

鉴于本研究模型对中国书画术语的特别考量,除考察模型整体准确率与Kappa系数外,还计算文化专属术语准确率与多维度术语准确率,并与传统方法进行对比,以全面分析模型性能,见表6。

结果显示,本研究模型准确率达89.2%,较传统字面匹配法(准确率为59.5%)提升29.7%;Kappa系数=0.89,与专家标注一致性极强。传统方法因仅依赖文本相似度,对文化专属术语(如“意境”“气韵”)与多维度术语(如“寒江独钓”)匹配准确率不足40%,而本模型通过引入文化适配系数与结构化语义分析,有效提升了此类术语的对齐精度。

表6 模型整体性能对比

| 方法/模型                  | 准确率(%) | Kappa系数 | 文化专属术语准确率(%) | 多维度术语准确率(%) |
|------------------------|--------|---------|--------------|-------------|
| 传统字面匹配法 <sup>[2]</sup> | 59.5   | 0.58    | 38.2         | 35.7        |
| 本研究模型                  | 89.2   | 0.89    | 82.1         | 88.5        |

### 3.3.2 细分类别适配性

针对7种映射关系分别计算各匹配类型准确率，如表7所示。可以看出，各类型准确率存在差异： $=EQ$ （精准等同）、 $EQ+$ （交叉复合等同）与 $EQ|$ （并列复合等同）三种类型的准确率最高，均超过92%； $BM$ （上位映射）与 $NM$ （下位映射）类型的准确率也较高，分别为88.7%和90.2%；而 $RM$ （关联映射）类型的准确率最低，为70.0%。

表7 各匹配类型准确率

| 映射关系 / 匹配类型 | 术语数量 (个) | 正确匹配数 (个) | 准确率 (%) |
|-------------|----------|-----------|---------|
| $=EQ$       | 175      | 172       | 98.3    |
| $\sim EQ$   | 107      | 91        | 85.0    |
| $EQ+$       | 13       | 12        | 92.3    |
| $EQ $       | 15       | 14        | 93.3    |
| $BM$        | 71       | 63        | 88.7    |
| $NM$        | 173      | 156       | 90.2    |
| $RM$        | 27       | 19        | 70.0    |

这一结果客观反映了西方词表的结构性局限——AAT未收录“意境”“气韵”等中国书画核心美学范畴，仅能通过“关联映射”将其与“aesthetics”等宽泛概念关联，导致匹配精度较低，也印证了本研究提出“文化适配系数”的必要性。

### 3.3.3 核心模块有效性

“文化适配系数”“四维语义拆解”“动态优先级规则”是本研究模型三个核心模块，为验证这三个模块的有效性，本研究设计了三组消融实验，从完整模型中移除某一模块，测算剩余模型的准确率，对比结果见表8。

表8 消融实验结果

| 实验设置          | 准确率 (%) | 准确率下降幅度 (%) |
|---------------|---------|-------------|
| 完整模型          | 89.2    | 0           |
| 移除“文化适配系数”模块  | 78.5    | 10.7        |
| 移除“四维语义拆解”模块  | 75.3    | 13.9        |
| 移除“动态优先级规则”模块 | 62.1    | 27.1        |

依次移除这三个核心模块后，准确率分别下降10.7%、13.9%、27.1%，明确了其对模型性能的贡献度，验证了核心算法模块的必要性与协同价值。

“文化适配系数”模块是捕捉“意境”等文化专属术语语义的核心，以及量化文化维度、保障匹配精度的基础；“四维语义拆解”模块通过结构化语义提取，避免某些复杂术语的语义丢失，是多维度计算的基础；“动态优先级规则”能有效避免匹配类型的误判，是保障匹配类型判定准

确的关键。三者协同作用，形成了一个覆盖文化维度量化、语义解析与关系判定的完整技术方案，为同类复杂人文领域的术语对齐提供了可复用的方法论。

## 4 结论

### 4.1 算法创新与核心发现

本研究针对中国书画术语跨语言检索中的“孤岛”困境，构建了一套融合文化适配的算法驱动对齐框架，并实现了基于 TPM-CPCV 与 AAT 的实证验证，算法的创新之处与发现如下。

(1) 提出了一种融合文化语义的多维度对齐模型，实现从“经验判断”到“量化计算”的方法演进。本研究将“文化关联”量化为可计算的“文化适配系数”，并结合文本、语义与领域权重，构建加权综合评分模型。实验表明，该模型整体匹配准确率达 89.2%，较传统字面匹配方法提升 29.7%；文化专属术语（如“意境”“气韵”）匹配的准确率提升至 82.1%，有效解决了文化维度缺失导致的语义捕捉不足问题。

(2) 设计基于动态优先级的映射判定机制，显著提升对齐结果的可解释性与系统性。通过定义 7 类映射关系并设计“=EQ > ~EQ > EQ+ > EQ | > BM > NM > RM”的动态判定顺序，系统化处理了术语间复杂的语义竞争与边界情况，避免了类型混淆。消融实验表明，该模块对整体准确率的贡献率达 27.1%，是保障匹配类型判定一致性的关键。

(3) 量化揭示了中西方艺术术语体系的结构性差异，为跨语言知识融合提供路径依据。结果表明：约 30% 的术语可通过“精准等同”实现对接，为系统低成本落地提供切入点；而“意境”等核心美学范畴因在 AAT 中缺失，仅能通过“关联映射”实现间接对齐，准确率相对较低（70.0%），从数据层面印证了中西分类逻辑的根本差异。

### 4.2 研究局限与展望

尽管本研究取得了预期效果，但仍存在以下局限，有待在未来工作中进一步完善。

(1) 语料覆盖范围有待扩展。当前实验主要基于书画的主题与技法术语，未来应纳入材料、装裱等更多维度的术语，以检验模型的全面性。

(2) 文化适配深度可进一步挖掘。当前的“文化适配系数”主要基于词频计算，未来可探索融入历史语境与跨文化阐释差异，以提升对强文化关联术语的量化精度。

(3) 模型的泛化能力需持续验证。本框架在其他多语言词表（如 UDC、Iconclass）以及低资源语言书画术语上的有效性，是下一步的重要验证方向。

(4) 知识贡献机制的缺失与建构。目前的研究主要集中在 TPM-CPCV 与 AAT 的对照和差异分析上，尚未形成从“发现差异”到“补充完善”的完整机制。未来希望将那些 AAT 中尚未收录、但具有中国书画特色的核心概念（如皴法、题跋、钤印等），通过规范的提案流程，正式纳入 AAT 体系，实现中国书画术语在国际词表中的系统性贡献。

### 【参考文献】

- [1] Chen S J, Zeng M L, Chen H H. Alignment of conceptual structures in controlled vocabularies in the domain of

Chinese art: a discussion of issues and patterns [ J ] . International Journal on Digital Libraries, 2016, 17(1): 23–38.

[ 2 ] 陈淑君, 陈雪华. 中国艺术领域的中英控制词汇语义对应 [ J ] . 图书资讯学刊, 2015, 13 ( 2 ) : 161–208.

[ 3 ] Whorf B L. Gestalt technique of stem composition in Shawnee [ C ] //Carroll J B. Language thought and reality: Selected writings of Benjamin Lee Whorf. Cambridge: The MIT Press, 1956: 160–172.

[ 4 ] 庄颖. 面向人工智能的博物馆藏品知识组织: 以故宫博物院“中国古代可移动文物概念参考模型”为例 [ J ] . 故宫博物院院刊, 2023 ( 11 ) : 126–136.

[ 5 ] Madsen B N, Thomsen H E, Vikner C. Principles of a system for terminological concept modelling [ C ] //Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon: European Language Resources Association, 2004: 15–19.

[ 6 ] Soergel D. Organizing information: principles of data base and retrieval systems [ M ] . Orlando: Academic Press, 1985.

[ 7 ] Larson M L. Meaning–based translation: a guide to cross–language equivalence [ M ] . 2nd ed. Lanham: University Press of America, 1998.

[ 8 ] Mikolov T, Le Q V, Sutskever I. Exploiting similarities among languages for machine translation [ J/OL ] . arXiv, 2013: 1309.4168. <https://arxiv.org/abs/1309.4168>.

[ 9 ] Artetxe M, Labaka G, Agirre E. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance [ C ] //Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2016: 2289–2294.

[ 10 ] Ruder S, Vulić I, Søgaard A. A survey of cross–lingual word embedding models [ J ] . Journal of Artificial Intelligence Research, 2019, 65(1): 569–631.

[ 11 ] Navigli R, Ponzetto S P. BabelNet: the automatic construction, evaluation and application of a wide–coverage multilingual semantic network [ J ] . Artificial Intelligence, 2012, 193: 217–250.

[ 12 ] Giunchiglia F, Gábor Bella, Nair N C, et al. Representing interlingual meaning in lexical databases [ J ] . Artificial Intelligence Review, 2023, 56(10): 11053–11069.

[ 13 ] Chen S J, Chen H H. Mapping multilingual lexical semantics for knowledge organization systems [ J ] . The Electronic Library, 2012, 30(2): 278–294.

[ 14 ] 张俊娥. 艺术和建筑词表AAT的中文化研究 [ J ] . 数字图书馆论坛, 2015 ( 7 ) : 36–43.

[ 15 ] 张萌. 图像志与图像数据库构建 [ D ] . 北京: 中国艺术研究院, 2024: 142.

[ 16 ] Miller G A. WordNet: a lexical database for English [ J ] . Communications of the ACM, 1995, 38(11): 39–41.

[ 17 ] Chen S J, Wu D, Peng P W, et al. AAT–Taiwan: toward a multilingual access to cultural objects [ C ] //Lalmas M. Research and Advanced Technology for Digital Libraries: Proceedings of the 14th European Conference, ECDL 2010. Berlin: Springer, 2010: 389–392.

[ 18 ] 邵洛羊. 中国美术大辞典 [ M ] . 上海: 上海辞书出版社, 2002.

[ 19 ] 成复旺. 中国美学范畴辞典 [ M ] . 北京: 中国人民大学出版社, 1995.

[ 20 ] Pustejovsky J. The generative lexicon [ M ] . Cambridge: The MIT Press, 1995.

[ 21 ] 陈淑君. 索引典的中英词汇语义对应之研究: 以中国艺术领域为例 [ D ] . 台北: 台湾大学, 2012.

# A Multidimensional Semantic Model-Based Alignment Algorithm for Chinese and Western Painting and Calligraphy Terminology Thesauri

Niu Liang

(School of Management Science and Engineering, China Jiliang University, Hangzhou 310018, China)

**Abstract:** [ **Purpose/Significance** ] To address the “island” dilemma in cross-language retrieval of Chinese painting and calligraphy terminology, and to tackle the challenges posed by cultural differences between China and the West along with incompatible vocabulary structures, this study proposes an algorithmic model for precise term alignment. [ **Method/Process** ] A multidimensional model was constructed, integrating text similarity, semantic similarity, domain-specific weighting, and cultural adaptation coefficients. Through four-dimensional semantic decomposition, dynamic priority rule design, and mapping relationship determination, accurate matching of Chinese and Western painting and calligraphy terminology vocabularies and alignment of conceptual structures were achieved. [ **Result/Conclusion** ] The multidimensional model, which incorporates textual, semantic, domain, and cultural differences, demonstrates strong overall performance. It significantly improves the accuracy of term matching, offering valuable insights for cross-language retrieval of painting and calligraphy terminology and the construction of multilingual knowledge organization systems.

**Keywords:** Conceptual structure alignment; Semantic decomposition; Multi-dimensional fusion; Digital humanities

( 本文责编: 王秀玲 )

ChinaXiv:202604.00085v1