An Alignment Algorithm for Chinese and Western Painting and Calligraphy Terminology Vocabularies Integrating Multi-dimensional Semantic Models

Authors: Liang Niu

Date: 2026-04-01T17:39:51+00:00

Abstract

[Purpose/Significance] To alleviate the "island"dilemma of cross-lingual retrieval for Chinese calligraphy and painting terminology, and to address the issues of cultural differences between China and the West and incompatibility in vocabulary structures, this paper proposes an algorithmic model for precise terminology alignment. [Method/Process] A multidimensional model integrating text similarity, semantic similarity, domain weights, and cultural adaptation coefficients was constructed. Through four-dimensional semantic decomposition, dynamic priority rule design, and mapping relationship determination, precise matching and conceptual structure alignment of Chinese and Western calligraphy and painting terminology vocabularies were achieved. [Result/Conclusion] The multidimensional model, which integrates text, semantics, domain, and cultural differences, demonstrates strong overall performance. It can significantly improve the accuracy of terminology matching, providing a reference for cross-lingual retrieval of calligraphy and painting terminology and the construction of multilingual knowledge organization systems.

Full Text

Preamble

A Multidimensional Semantic Model for a Bilingual Chinese-Western Painting and Calligraphy Terminology Lexicon

Abstract

This study explores the construction of a bilingual terminology lexicon for Chinese and Western painting and calligraphy by integrating a multidimensional semantic model. Given the profound cultural differences and technical nuances between Eastern and Western artistic traditions, traditional translation methods often fail to capture the deep conceptual equivalence required for academic research. By leveraging machine learning and deep learning techniques, this research develops a framework that accounts for historical context, stylistic evolution, and technical specifications. The resulting lexicon aims to provide a standardized reference for scholars, facilitating cross-cultural dialogue and enhancing the precision of art historical documentation.

1. Introduction

The globalization of art history research necessitates a more rigorous approach to the translation and standardization of terminology. Chinese painting and calligraphy, with their unique aesthetic foundations and technical vocabulary, often lack direct equivalents in Western art theory. Conversely, Western artistic terms—ranging from Renaissance techniques to contemporary movements—require careful contextualization when translated into Chinese. This paper proposes a multidimensional semantic model designed to bridge these linguistic and conceptual gaps.

2. Methodology

2.1 Multidimensional Semantic Framework The core of our approach lies in the integration of multiple semantic layers. Unlike traditional dictionaries that rely on one-to-one word mapping, our model incorporates:

- **Historical Dimension:** Tracking the evolution of terms across different dynasties and artistic periods.
- **Technical Dimension:** Defining terms based on physical materials (e.g., $\mathcal{M}$ for media), tools, and execution methods.
- **Aesthetic Dimension:** Capturing the philosophical underpinnings, such as the concept of "Qi" (spirit resonance) in Chinese art versus "Mimesis" in Western traditions.

2.2 Data Collection and Processing We utilized a comprehensive corpus of art historical texts, exhibition catalogs, and theoretical treatises. To ensure

accuracy, we applied deep learning algorithms to identify semantic clusters and relationship mappings. The relationship between a term T and its semantic context C can be represented as:

$$\mathcal{S}(T) = \sum_{i=1}^{n} w_i \cdot f(C_i)$$

where w_i represents the weight of a specific semantic dimension and $f(C_i)$ denotes the feature vector of the context.

3. Lexicon Construction

The construction process involves the synthesis

摘要

Introduction

[Purpose/Significance] To alleviate the "island"dilemma of cross-lingual retrieval for Chinese calligraphy and painting terminology, this study addresses the challenges posed by cultural differences and lexical gaps between China and the West. By constructing a multilingual knowledge graph for Chinese calligraphy and painting, we aim to provide a structured semantic framework that facilitates international scholarly exchange and improves the accessibility of traditional Chinese art in a global digital context.

[Method/Process] The research utilizes a combination of machine learning and expert curation to extract, align, and standardize terminology across multiple languages. We focus on the semantic nuances of traditional artistic techniques, historical periods, and stylistic categories. By integrating heterogeneous data sources, the study develops a robust ontological model that links Chinese concepts with their closest Western equivalents while preserving the unique cultural connotations of the original terms.

[Result/Conclusion] The resulting knowledge graph demonstrates significant improvements in retrieval precision and recall compared to traditional keyword-based systems. By bridging the linguistic divide, this framework supports more nuanced cross-cultural analysis and provides a scalable model for the digital preservation of intangible cultural heritage. The study concludes that structured semantic alignment is essential for overcoming the fragmentation of art historical data in the global information landscape.

摘要

[Purpose/Significance] To alleviate the "information island"dilemma in the cross-linguistic retrieval of Chinese calligraphy and painting terminology, this study proposes a precision alignment algorithm model to address the challenges posed

by Sino-Western cultural differences and incompatible vocabulary structures. [Method/Process] A multidimensional model was constructed by integrating text similarity, semantic similarity, domain weights, and cultural adaptation coefficients. Through four-dimensional semantic decomposition, the design of dynamic priority rules, and the determination of mapping relationships, the model achieves precise matching and conceptual structure alignment between Chinese and Western calligraphy and painting vocabularies. [Results/Conclusion] The multidimensional model, which fuses textual, semantic, domain-specific, and cultural difference factors, demonstrates strong overall performance. It significantly improves the accuracy of terminology matching, providing a valuable reference for the cross-linguistic retrieval of art terminology and the construction of multilingual knowledge organization systems.

The findings provide a practical reference for future system construction.

关键词

Conceptual Structure Alignment, Semantic Decomposition, and Multi-dimensional Fusion in Digital Humanities

Abstract

In the field of Digital Humanities, the integration and analysis of heterogeneous cultural heritage data require sophisticated methods for reconciling disparate conceptual frameworks. This paper proposes a novel approach based on conceptual structure alignment, semantic decomposition, and multi-dimensional fusion. By breaking down complex historical and cultural concepts into their constituent semantic components, we facilitate a more granular alignment across different knowledge systems. Our methodology employs machine learning techniques to automate the identification of semantic overlaps and discrepancies, followed by a multi-dimensional fusion process that integrates spatial, temporal, and thematic attributes. This framework enhances the interoperability of digital archives and provides scholars with deeper insights into the evolution of cultural concepts.

1. Introduction

The rapid digitization of cultural heritage has resulted in a vast accumulation of data across various domains. However, the lack of standardized conceptual structures often hinders cross-disciplinary research and data integration. Traditional methods of manual alignment are labor-intensive and frequently fail to capture the nuanced semantic shifts inherent in historical data. To address these challenges, this study explores the synergy between semantic decomposition and multi-dimensional fusion as a means to achieve robust conceptual structure alignment.

2. Conceptual Structure Alignment

Conceptual structure alignment is the process of establishing correspondences between different ontologies or taxonomies. In Digital Humanities, this involves mapping terms and concepts from diverse historical periods and cultural contexts.

2.1 Semantic Decomposition Semantic decomposition involves breaking down a complex concept into its fundamental semantic units. For instance, a historical title or administrative role can be decomposed into attributes such as authority, geographical jurisdiction, and temporal duration. By analyzing these components individually, we can identify partial matches between concepts that might otherwise appear unrelated.

Let $\mathcal{C}$ represent a complex concept, which can be decomposed into a set of semantic features $F = \{f_1, f_2, \ldots, f_n\}$. The similarity between two concepts $\mathcal{C}_i$ and $\mathcal{C}_j$ is then calculated based on the overlap and weighted importance of their respective feature sets:

$$Sim(\mathcal{C}_i, \mathcal{C}_j) = \frac{\sum_{k=1}^{n} w_k \cdot \sigma(f_{ik}, f_{jk})}{\sqrt{\sum w_k^2}}$$

where w_k denotes the weight of the k-th semantic dimension and σ is a similarity

关键词

Conceptual Structure Alignment and Semantic Decomposition: Multi-dimensional Fusion in Digital Humanities

Abstract

In the burgeoning field of Digital Humanities, the integration of heterogeneous data sources remains a significant challenge. This paper proposes a novel framework centered on **conceptual structure alignment** and **semantic decomposition** to facilitate multi-dimensional data fusion. By breaking down complex humanistic concepts into granular semantic units, we enable more precise cross-disciplinary mapping and knowledge discovery. Our approach leverages machine learning techniques to bridge the gap between qualitative interpretative traditions and quantitative computational methods.

1. Introduction

The digital transformation of humanities research has led to an explosion of diverse datasets, ranging from historical archives and literary corpora to geospatial data and artistic metadata. However, the inherent ambiguity and context-dependency of humanistic data often hinder effective integration. Traditional

data fusion methods frequently fail to capture the nuanced relationships between disparate conceptual frameworks. To address this, we introduce a methodology for conceptual structure alignment that prioritizes semantic decomposition as a prerequisite for multi-dimensional fusion.

2. Conceptual Structure Alignment

Conceptual structure alignment refers to the process of identifying and mapping equivalent or related concepts across different knowledge domains or datasets. In the context of Digital Humanities, this requires moving beyond simple keyword matching.

- **Ontological Mapping**: Establishing formal relationships between entities using standardized vocabularies.
- **Contextual Normalization**: Adjusting for temporal and cultural shifts in language use to ensure that the $\mathcal{C}_{source}$ aligns accurately with $\mathcal{C}_{target}$.
- **Structural Isomorphism**: Identifying similar relational patterns within different hierarchical schemas.

3. Semantic Decomposition

Semantic decomposition involves breaking down high-level, complex concepts into their constituent semantic features. This process allows for a more granular analysis of how meanings are constructed and transformed across different media and periods.

[Figure 1: see original paper]

As illustrated in [Figure 1: see original paper], a concept such as "Urbanization" can be decomposed into dimensions including demographic shifts, architectural evolution, and socio-economic restructuring. By representing these as a vector space $\mathbf{V} = \{v_1, v_2, \dots, v_n\}$, we can apply computational models to measure semantic proximity.

4. Multi-dimensional Fusion

The ultimate goal of alignment and decomposition is the fusion of multi-dimensional data. This integration allows researchers to view a single phenomenon through multiple lenses—spatial, temporal, and thematic—simultaneously.

$$F$$

0 引言

Introduction

The digitization and globalization of cultural heritage have created an urgent demand for precise cross-linguistic retrieval of artistic resources rich in Eastern philosophical significance, such as Chinese calligraphy and painting. However, fundamental differences between Chinese and Western cultural cognition and classification logic have led to a typical "silo" phenomenon among core terms in cross-linguistic knowledge organization, posing significant challenges for effective conceptual alignment. Research indicates a semantic gap between Chinese and Western vocabularies regarding core terminology in Chinese calligraphy and painting, resulting in cross-system data transmission accuracy of less than 50% [**?**]. This dilemma stems from deep-seated cognitive differences: Chinese terms (such as *yijing* [artistic conception] and *qiyun* [spirit resonance]) deeply internalize philosophical concepts like the Taoist "void and substance"and the Confucian "beauty of harmony," whereas Western terminology tends to emphasize technical and formal objectivity. For example, the *Art & Architecture Thesaurus* (AAT) defines the Chinese painting term *baimiao* solely by its technical characteristic of "monochrome lines," overlooking its aesthetic core of "conveying spirit through line." Furthermore, core concepts such as *yijing* are entirely absent from the AAT. In this context, achieving maximum alignment between the conceptual structures of Chinese and Western calligraphy and painting vocabularies—while acknowledging cultural differences—has become a critical and urgent issue for supporting cross-linguistic retrieval.

To address this challenge, this study aims to construct an algorithm-driven conceptual structure alignment framework to facilitate the mapping between Chinese and Western calligraphy and painting vocabularies. For the purposes of this research, the *Taipei Palace Museum Calligraphy-and-Painting Controlled Vocabulary* (TPM-CPCV) was selected as the source for Chinese terminology, while the AAT serves as the source for Western terminology. The TPM-CPCV is a specialized vocabulary developed by the Taipei Palace Museum based on its collection of Chinese calligraphy and painting. Containing approximately 2,000 entries, it provides in-depth coverage of techniques, themes, and materials unique to Chinese art—such as *cunfa* (texture strokes), *caochong* (insects and plants), and *duanshi* (Duan stone)—and possesses high cultural specificity and academic authority. The AAT, developed by the Getty Research Institute, is a globally recognized art thesaurus containing approximately 35,000 concepts. With its complete structure and clear hierarchy, it serves as a standard knowledge organization tool in the Western art field.

Consequently, these two vocabularies effectively represent the terminological systems of Chinese calligraphy and painting and Western art, respectively. By achieving conceptual alignment between TPM-CPCV and AAT, this study validates the effectiveness of the proposed framework for precise cross-linguistic retrieval.

The core methodology of this research is as follows: First, a multi-dimensional quantitative model is constructed by integrating text, semantics, domain weights, and cultural adaptation coefficients, thereby transforming abstract "cultural associations" into computable "cultural adaptation coefficients." Building on this, a four-dimensional semantic decomposition framework based on "Motivation-Purpose-Composition-Form" is designed to perform a structured analysis of the multi-dimensional semantics embedded in TPM-CPCV, addressing its semantic ambiguity and cultural density. Furthermore, a set of matching type determination rules based on dynamic priorities is proposed to systematically handle complex mapping relationships between terms, enhancing the interpretability and usability of the alignment results. The innovation and value of this study are primarily reflected in several areas: addressing the rich cultural connotations of Chinese calligraphy and painting terminology, it attempts to build a multi-dimensional alignment model incorporating cultural adaptation coefficients, promoting a methodological shift in cross-linguistic conceptual alignment from "empirical judgment" to "quantitative calculation." The research findings can be applied to cross-linguistic retrieval systems in museums, libraries, and digital archives to improve the international visibility and retrieval precision of Chinese calligraphy and painting resources. Finally, by systematically comparing the conceptual structures of TPM-CPCV and AAT, this study reveals preliminary differences in cultural expression and classification logic, providing data support and analytical foundations for the future revision, expansion, or cross-system interconnection of related vocabularies.

1.1 跨语言概念与词汇等同性研究的技术演进

Cross-linguistic conceptual and lexical equivalence serves as the core theoretical foundation for terminology alignment. Its development has evolved around "conceptual association methods" and "technical support systems," characterized by a three-stage evolutionary process.

1.1.1 传统知识组织理论阶段

This stage centers on the manual construction of structured knowledge systems, aiming to establish cross-linguistic semantic associations by clarifying the logical relationships between terms. Its primary contribution lies in establishing the fundamental principle that "concepts take precedence over linguistic symbols." Madsen et al. first proposed a cross-linguistic terminology control mechanism, advocating for the reduction of semantic ambiguity through bidirectional mapping between "concepts" and "terms." Their approach—defining abstract concepts independently of specific languages before matching them with multilingual terminology—laid the theoretical foundation for the construction of multilingual thesauri.

Building on this, Soergel proposed the "Concept-Term Separation Model," which

explicitly defines a concept as an "abstract entity reflecting the essential attributes of an object," while a term serves merely as the "linguistic carrier of the concept." Methodologically, this model provides critical support for the cross-linguistic definition of culture-specific terminology. Furthermore, the "Conceptual Core → Linguistic Symbol" cross-layer mapping model developed by Larson shifted the research focus from surface-level "literal matching" to deep-level "semantic essence matching." This model emphasizes the necessity of extracting the core semantic attributes of a term before performing cross-linguistic mapping, providing an essential methodological basis for cross-linguistic semantic alignment.

1.1.2 计算语言学转向阶段

Entering the 21st century, the development of neural networks and machine learning technologies propelled the field into a stage of "automated model construction." The core breakthrough of this era lies in the realization of quantitative calculations for semantic similarity. The cross-lingual word embedding alignment paradigm proposed by Mikolov et al. [**?**] maps vocabularies from different languages into a unified high-dimensional vector space. By calculating Euclidean distance or cosine similarity—for instance, ensuring the vector distance between "竹" (zhu) and "bamboo" approaches zero—this approach opened a feasible path for the large-scale batch matching of terminology.

However, inherent discrepancies between different linguistic vector spaces often lead to semantic shifts. To address this, Artetxe et al. [**?**] proposed an "orthogonal Procrustes alignment algorithm." By enforcing orthogonal constraints on the mapping between different language vector spaces, they reduced the matching error for low-frequency words and culture-specific terms by approximately 20%.

Despite these advancements, limitations remain regarding specialized domains. In a systematic review of 39 mainstream cross-lingual embedding models, Ruder [**?**] pointed out that models trained on general-purpose corpora exhibit significant deficiencies in adapting to "low-resource languages" and "culture-specific domains." Specifically, the accuracy of semantic capture for domain-specific terminology was found to be less than 50%. The fundamental reason for this failure is the inability to effectively integrate domain knowledge and cultural characteristics. Consequently, while this stage significantly improved automation efficiency, it still faces inherent limitations regarding cultural adaptability.

1.1.3 神经符号融合阶段

In recent years, to overcome the limitations of purely data-driven models, research has entered a stage of neuro-symbolic integration. This approach aims to improve the matching precision of complex domain terminology by fusing the symbolic logic of knowledge graphs with the semantic representations of neural networks.

Navigli et al.' s BabelNet [?] represents an early paradigm in this direction. It establishes a "concept-semantic-cultural" three-layer framework that first utilizes knowledge graphs to organize the hierarchical relationships and attribute information of terms, and then combines this with word embedding models to calculate semantic similarity. This method effectively enhances term alignment precision in complex scenarios. Furthermore, Giunchiglia et al. [?] proposed a "zero-shot cross-lingual concept mapping" method. By employing transfer learning to leverage conceptual knowledge from high-resource languages, this approach directly predicts the vector representations of terms in low-resource domains (such as Chinese painting and calligraphy), enabling associative mapping even in the absence of extensive bilingual corpora.

Research in this stage marks a significant transition in the field: moving away from shallow matching dependent on massive datasets toward precise alignment that integrates prior knowledge with deep semantic understanding.

1.2 中国书画领域术语对齐研究现状

Focusing on the specific characteristics of Chinese painting and calligraphy, domestic scholars have conducted a series of targeted explorations. While these studies have achieved localized progress, they also reveal common challenges regarding large-scale and high-precision alignment in this field. In exploring the alignment of Chinese and Western art terminology, Chen et al. [?] critically revealed the inherent "Western art-centrism" within the Art & Architecture Thesaurus (AAT). They pointed out a fundamental conflict between the AAT' s "form + function" classification logic and the "cultural connotation priority" cognitive logic of Chinese painting and calligraphy, advocating for a "bridge strategy between ontology and vocabulary" to achieve cross-system interoperability. Although this approach is highly insightful, it fails to further refine universal bridging rules or standardized mapping mechanisms, rendering the proposal difficult to implement in practice. On a practical level, Zhang Jun [?] conducted a systematic study on the localization of the AAT into Chinese, focusing on the translational equivalence of key technical terms such as *cunfa* (texture strokes) and *miaofa* (line drawing methods). While this work provides a foundational vocabulary reference for cross-lingual retrieval, its focus remains on the linguistic surface. It fails to achieve deep alignment at the conceptual structure level, and thus cannot fundamentally resolve the deep-seated semantic deviations caused by cultural differences. Furthermore, some scholars [?] argue that Chinese art terminology suffers from a widespread phenomenon of "semantic attribution confusion" in cross-lingual annotation. This research acutely identifies the symptoms of the problem and provides an important basis for understanding the shortcomings of existing systems, yet it stops short of proposing a systematic, computable solution to address this persistent issue.

In summary, while existing research has laid a foundation for the alignment of Chinese and Western art terminology through various dimensions—including ontological construction, theoretical critique, linguistic translation, and problem

diagnosis—several limitations remain prevalent. These studies tend to emphasize theoretical models at the expense of practical application, focus on surface-level translation while lacking deep semantic alignment, or excel at problem analysis while neglecting algorithmic implementation. These unresolved issues constitute the starting point for the present study.

2 核心算法体系设计

To achieve precise alignment between TPM-CPCV terminology and AAT concepts, this study constructs a hierarchical core algorithmic framework. This system addresses alignment challenges arising from cultural differences and structural incompatibilities through several key stages: the verification of English translations, multi-dimensional fusion computation, and the determination of matching types. The workflow of the core algorithm is illustrated in Figure 1 [Figure 1: see original paper].

2.1 确认英译词汇

To establish a bridge for cross-linguistic matching, the TPM-CPCV must first be converted into standard English terminology. In this study, we utilized the Baidu Translation API to generate three to five candidate English translations, which were then verified against professional references such as the *English-Chinese Chinese-English Dictionary of Art*. The optimal translation was determined through an "expert voting + domain adaptation scoring" mechanism. To address polysemy, we employed a domain classification algorithm based on WordNet 3.0 synonyms and Support Vector Machines (SVM) to filter out irrelevant meanings. For example, while candidate meanings for "shanshui" (山水) include "landscape painting" (fine arts domain) and "geography" (geographical domain), the SVM model accurately identifies "landscape painting" as the target meaning, thereby establishing a rigorous foundation for subsequent calculations.

2.2 多维度融合计算

After completing the English translation of the TPM-CPCV, a weighted comprehensive scoring model is constructed. By integrating text similarity, semantic similarity, domain weights, and cultural adaptation coefficients, this model calculates a comprehensive quantitative matching degree (Similarity) for each TPM-CPCV term and its corresponding AAT candidate concepts. This score serves as the core basis for subsequent automated matching decisions, enabling the matching of the majority of terms. The calculation formula is as follows:

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

In this model, α, β, γ, and δ represent feature weights, which are determined by domain experts using the Analytic Hierarchy Process (AHP). For the purposes of

this study, the weights are assigned as $\alpha = 0.3$ (representing the text similarity weight), with the remaining coefficients calibrated accordingly to reflect the relative importance of each feature in the evaluation process.

···similarity), $\beta = 0.3$ (semantic similarity), $\gamma \in [0.2, 0.4]$ (domain weight), and $\delta \in [0.05, 0.15]$ (cultural adaptation coefficient). The calculation methods for each dimension are as follows:

(1) Text Similarity ($Text\_Sim$): The edit distance is normalized into a similarity value, which is then averaged with the Jaccard coefficient.

The Edit Similarity (Edit_{Sim}) metric utilizes Edit Distance to precisely quantify the "absolute difference" between two strings—specifically, the number of character-level operations required to transform one into the other. In this context, it is used to calculate the number of modifications needed to align the English translation of a term with the corresponding AAT (Art & Architecture Thesaurus) concept label. This value is then normalized into a similarity score, typically calculated using the following formula:

$$Edit\_Sim = 1 - \frac{EditDistance(s_1, s_2)}{\max(len(s_1), len(s_2))} \tag{2}$$

In this context, `len()` is a function used to calculate the length of a string (i.e., the total number of characters it contains). The value of the edit similarity metric is normalized.

The similarity value ranges within the interval $[0, 1]$, where a higher value indicates a higher degree of similarity. For example, the edit distance between the English translation of "竹" (bamboo) and the AAT term "bamboo(plants) [300311500]" is 8. The algorithm automatically removes general domain suffixes such as "(plants)" ; after this correction, the $Edit\_Sim = 1$. Jaccard Coefficient: This measures the ratio of the intersection to the union of the keywords. The value also ranges within the $[0, 1]$ interval, with higher values representing greater similarity. The formula is as follows:

$$Jaccard\_sim = \frac{|s_1 \cap s_2|}{|s_1 \cup s_2|} \tag{3}$$

For example, the English translation of "后妃" (empresses and concubines) and the AAT keyword "empresses [300150492]" have an intersection of {empresses} and a union of {empresses, concubines}, resulting in a $Jaccard\_sim = 0.5$.

(2) Semantic Similarity ($Semantic\_Sim$): This metric evaluates deep semantic associations by combining WordNet semantic distance with the SKOS hierarchical relationships of the AAT. The semantic similarity is calculated as the mean of the semantic distance and the association strength.

Semantic Distance ($WordNet\_Sim$): Based on WordNet 3.0, this measures the semantic distance between TPM-CPCV terms and AAT concepts. The formula is as follows:

$$WordNet\_Sim = \frac{1}{1 + distance} \tag{4}$$

For example, "pine" (松) and the AAT term "pine trees (plants) [3 00174989]" belong to the same synset, resulting in a distance of 0. (2) Association Strength (SKOS_{Sim}): This metric calculates the hierarchical distance and association strength based on the SKOS hierarchical relationships within the AAT. These relationships are detailed in Table 1 . For instance, "Gao Gu You Si Miao" (high ancient gossamer line drawing) and the AAT term "line drawing [3 00100194]" share a "broader" relationship (superordinate-subordinate), with a hierarchical distance of 1 and an association strength of 0.9, resulting in a SKOS_{Sim} value of 0.9.

SKOS Relationship Types (3) Domain Weight (Domain_{Weight}): The "Domain Specificity Index" (DSI) is employed to quantify the domain exclusivity of a given term.

$$DSI = \frac{Freq_{art}}{Freq_{art} + Freq_{general}}$$

Statistical frequencies are calculated based on a specialized arts domain corpus and a general-purpose corpus. For example, if the term "Cunfa" (texture strokes) appears 500 times in the arts domain and 10 times in the general domain, its Domain Specificity Index (DSI) is calculated as $DSI = 500/(500 + 10) \approx 0.98$, with a corresponding weight $\gamma = 0.4$. Conversely, if the term "Bamboo" appears 800 times in the arts domain and 2,000 times in the general domain, its $DSI = 800/(800 + 2000) \approx 0.29$, with a corresponding weight $\gamma = 0.2$. The formula for determining the domain weight is defined as follows:

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

4. Cultural Adaptation Coefficient ($Cultural\_Adapt$)

To quantify the cultural alignment of the generated content, we constructed a "Chinese Painting and Calligraphy Cultural Lexicon." This lexicon comprises 200 culture-specific terms—including concepts such as "Artistic Conception" (*Yijing*), "Rhythmic Vitality" (*Qiyun*), "Brush and Ink" (*Bimo*), and "Unity of Heaven and Humanity" (*Tianren Heyi*)—sourced from authoritative references such as the *Dictionary of Chinese Art* and *Categories of Chinese Aesthetics*.

The Cultural Adaptation Coefficient is calculated based on the Term Frequency-Inverse Document Frequency (TF-IDF) method. Specifically, we measure the

frequency of these specialized cultural terms relative to the total semantic word count within the generated text. The formula for this coefficient is defined as follows:

$$Cultural\_Adapt = 0.05 + 0.1 \times \frac{Cultural\_Freq}{Total\_Freq} \tag{7}$$

For example, the cultural word frequency for the term "Artistic Conception" (意境, *yìjìng*) is 3 (including related concepts such as "blending of scene and emotion," "interplay of void and substance," and "Taoist thought"), while the total semantic word frequency is 10. Consequently, the cultural adaptation score is calculated as $Cultural\_Adapt = 0.05 + 0.1 \times 3/10 = 0.08$. Since "Artistic Conception" is a highly specialized cultural term, the $Cultural\_Adapt$ value is adjusted accordingly.

2.3 匹配类型判定

By defining mapping relationships, performing four-dimensional semantic decomposition, designing dynamic priority rules, and implementing mapping relationship determination, we achieve the precise identification of terminology matching types between TPM-CPCV and AAT.

2.3.1 映射关系定义

Matching TPM-CPCV terms with AAT terms requires covering various scenarios, such as "exact equivalence," "partial association," "hierarchical inclusion," and "composite construction." In accordance with the ISO 25964-2 standard and the specific characteristics of Chinese painting and calligraphy terminology, we define seven types of mapping relationships. These definitions clarify the scope of application and semantic logic for each relationship type, providing a classification framework for subsequent determinations, as shown in Table 2 .

The scope of two concepts from different vocabularies is identical, with no semantic deviation; they are bi-directionally interchangeable in cross-linguistic applications without requiring additional semantic correction. While the core semantic meaning of the concepts is consistent, there is partial overlap in scope or differences in cultural expression; they are interchangeable only in specific contexts and require the annotation of semantic deviation points. In "Intersecting Composite Equivalence," a complex term in TPM-CPCV is semantically equivalent to the intersection of two or more concepts in AAT, requiring a combination of multiple concepts to achieve a complete match; this mapping direction is irreversible. In "Coordinate Composite Equivalence," a term in TPM-CPCV covers the semantic scope of two or more non-overlapping sibling concepts in AAT, requiring a union of multiple concepts to achieve full coverage; this mapping direction is also irreversible. A TPM-CPCV term functions as a subordinate concept whose semantic scope is entirely contained within a

broader superordinate concept in AAT; matching is achieved through hierarchical association, supporting retrieval from narrow concepts to broader ones. A TPM-CPCV term functions as a superordinate concept whose semantic scope entirely encompasses a specific subordinate concept in AAT; matching is achieved through hierarchical association to supplement granular information, supporting retrieval from general concepts to specific ones.

The two types of concepts share no equivalence or hierarchical inclusion relationship but are semantically related (e.g., through cultural background or application scenarios). Matching can only be achieved through indirect association, and the basis for the association must be explicitly annotated.

2.3.2 四维语义拆解

Due to cultural differences, certain terms in TPM-CPCV present challenges such as semantic ambiguity and dense cultural connotations when aligned with AAT terminology. To address this, the present study designs a four-dimensional semantic framework based on Pustejovsky' s Generative Lexicon theory, decomposing the comprehensive semantics of terms into four dimensions: motivation, purpose, composition, and form. The motivation dimension analyzes the fundamental intent and philosophical thought behind artistic expression. The purpose dimension elucidates the intended aesthetic functions and cultural symbolism. The composition dimension deconstructs specific technical elements and physical materials. The form dimension describes typical visual styles and morphological characteristics. Through this framework, terms are transformed into a set of structured semantic descriptions (see for examples), providing a structured semantic foundation for subsequent multi-dimensional fusion calculations and the determination of complex mapping relationships. For a small number of complex terms that are difficult to determine automatically—due to highly unique cultural connotations or quantitative scores residing at threshold boundaries—this framework provides deep theoretical explanations and structured interpretations based on attribute structure analysis to assist in finalizing the mapping.

TPM-CPCV utilizes fine brushes (weasel-hair small script brushes) and heavy colors (mineral pigments). The lines are delicate and uniform, while the colors are rich and saturated. The composition is rigorous and symmetrical, meticulously depicting the forms of objects to restore objective reality. It showcases the detailed beauty of the subjects and conveys the painter' s spiritual insights, achieving an aesthetic realm where "scenery and emotion blend, and the virtual and real coexist." This process relies on brush and ink techniques and compositional layout, integrating cultural thought to create an intangible spiritual atmosphere through the combination of void and substance.

2.3.3 动态优先级规则设计与映射关系判定

Due to differences in the semantic correlation strength and determination priority among the seven mapping relationships, applying a uniform threshold for determination can easily lead to "type confusion" (e.g., misidentifying an "Approximate Equivalent" as an "Exact Equivalent"). To meet the practical requirements of TPM-CPCV terminology matching, a dynamic priority rule was designed based on the principle of "semantic correlation strength from high to low" : $= EQ >\sim EQ > EQ+ > EQ| > BM > NM > RM$. After completing the semantic analysis of the terminology, the alignment process enters the core phase of calculation and automatic determination.

First, a "multi-dimensional fusion calculation"is performed for each TPM-CPCV term. By integrating four dimensions of features—textual, semantic, domain-specific, and cultural—the system generates key indicators for each term, such as the comprehensive quantitative matching degree ($Similarity$) and the cultural adaptation coefficient ($Cultural\_Adapt$). Subsequently, following the aforementioned dynamic priority rules and the quantitative determination conditions for mapping relationships listed in , the system automatically classifies each term into a specific mapping type (e.g., "Exact Equivalent"$= EQ$, "Broader Mapping" BM, etc.) to complete the correspondence.

The core logic of this rule is to prioritize the determination of types with the closest semantic correlation and no ambiguity, and then gradually determine types with weaker correlation strengths or more complex structures. For example, if a TPM-CPCV term simultaneously satisfies some conditions for both "$= EQ$" and "$\sim EQ$" , the system will prioritize the "$= EQ$" determination because "$= EQ$" represents complete semantic identity and holds a higher priority than "$\sim EQ$" . A lower-priority determination process is only initiated when the conditions for higher-priority types are not met.

This approach avoids result bias caused by a disordered determination sequence. The core determination conditions are as follows: For $Similarity \geq 0.95$ and $Cultural\_Adapt \leq 0.05$, the four-dimensional features are highly consistent with no cultural deviation (e.g., 竹 → bamboo). For $0.85 \leq Similarity < 0.95$ and $Cultural\_Adapt \geq 0.1$, the core dimensions overlap, but cultural motivations lead to a deviation in scope (e.g., 后妃 → empresses). When the overall $Similarity \geq 0.80$ after combining with ≥ 2 AAT concepts, the semantics are decomposed into multiple independent dimensions (e.g., 春景山水 → spring + landscapes). When the overall $Similarity \geq 0.80$ after the union of ≥ 2 AAT concepts at the same level, the constituent dimensions include multiple parallel subcategories (e.g., 梅 (白・红・腊梅) → multiple varieties). For $0.70 \leq Similarity < 0.85$ where AAT is a clear broader concept of TPM-CPCV, the TPM-CPCV term is a specific manifestation of the AAT concept but possesses richer cultural connotations (e.g., 高古游丝描 → line drawing).

For $0.70 \leq Similarity < 0.85$ where TPM-CPCV is a clear broader concept of AAT, the AAT term is merely a specific subcategory under the broad compo-

sition of the TPM-CPCV term (e.g., 花鸟主题 → peony and bird themes). For $0.5 \leq Similarity < 0.70$ and $Cultural\_Adapt \geq 0.15$, the core cultural semantics (motivation/purpose) have no direct correspondence in AAT (e.g., 意境 → aesthetics).

3 TPM-CPCV and AAT Concept Matching and Verification

3.1 数据来源

(1) TPM-CPCV: Derived from the *Metadata Requirement Specification (Calligraphy and Painting Collections)* of the National Palace Museum in Taipei. The scope of terminology selection refers to the definitions of Chinese calligraphy and painting vocabulary. Following a process of "keyword matching + SVM classification + expert verification (Kappa=0.87)," 581 calligraphy and painting terms (543 subjects and 38 techniques) were manually screened. Representative content is shown in . (2) AAT Data: Concepts from the three facets most closely associated with Chinese calligraphy and painting— "Activities," "Styles and Periods," and "Materials" —were extracted from the Getty Research Institute' s Art & Architecture Thesaurus (AAT) and converted into JSON format to construct a candidate library. (3) Auxiliary Data: To support the calculation of dimensions such as domain weight and cultural adaptation, a specialized art domain corpus was constructed. This corpus covers Chinese classical texts such as the *Dictionary of Chinese Art*, *Record of Famous Paintings of Successive Dynasties (Lidai Minghua Ji)*, and *The Sequel to the Record of Paintings (Huaji)*, as well as AAT term definitions and authoritative English-language art history literature, used for relevant statistical analysis.

Number of Terms / Representative Vocabulary

Number of Terms / Representative Vocabulary: Horse, ox, sheep, tiger, deer, dragon, snake, monkey, camel; Eagle, crane, magpie, wild goose, sparrow, mandarin duck, peacock, pheasant; Butterflies and moths, cricket, dragonfly, cicada, bee, spider, ant, locust; Palace, pavilion, terrace, bridge, house, temple, pagoda, courtyard, thatched hut; Scholar' s objects, musical instruments, furniture (screens), eating and drinking vessels, flower vessels, fans, lanterns, censers and braziers; Annual festivals, Qingming, Mid-Autumn, Qilin, Phoenix, Bat, Bixie (evil-warding symbols); Baimiao (line drawing), Gongbi (meticulous brushwork), Xieyi (freehand style), Jiehua (architectural painting), Mogu (boneless technique), Shuanggou (double-outline), Zhihua (finger painting), Cunfa (texturing methods); Pima Cun (hemp-fiber stroke), Fupi Cun (ax-cut stroke), Gaogu Yousi Miao (ancient silk-thread line), Tiexian Miao (iron-wire line), Dingtou Shuwei Miao (nail-head rat-tail line), Liuye Miao (willow-leaf line).

3.2 匹配结果

Based on the multi-dimensional fusion calculation model shown in Equation (1) and the quantitative mapping determination rules presented in , an automated alignment calculation and mapping determination were performed on 581 TPM-CPCV terms. Ultimately, each term pair was categorized into one of seven mapping relationships ($= EQ$, $\sim EQ$, $EQ+$, $EQ|$, BM, NM, or RM). The system also outputted key indicators for each pair, including quantitative matching scores (Similarity) and cultural adaptation coefficients (Cultural_{Adapt}).

As shown in the term classification in , the process yielded a total of 175 "$= EQ$" pairs, 107 "$\sim EQ$" pairs, 13 "$EQ+$" pairs, 15 "$EQ|$" pairs, 71 "BM" pairs, 173 "NM" pairs, and 27 "RM" matching pairs.

3.3 模型性能与匹配效果验证

To systematically validate model performance and matching effectiveness, this study conducts evaluations across three dimensions: overall performance, suitability for specific sub-types, and the effectiveness of core modules. When measuring matching performance across these different dimensions, two types of quantitative indicators are employed: accuracy (the ratio of correctly matched terms to the total number of terms) and the Kappa coefficient (which measures the consistency between model results and expert manual annotations, where a Kappa ≥ 0.8 is considered to indicate "extremely strong consistency").

3.3.1 整体性能对比

Given the specific considerations of Chinese painting and calligraphy terminology in this study, we evaluated the model' s overall accuracy and Kappa coefficient, alongside the accuracy of culture-specific and multi-dimensional terms. These metrics were compared against traditional methods to provide a comprehensive analysis of model performance, as shown in . The results indicate that the proposed model achieved an accuracy of 89.2%, representing a 29.7% improvement over the traditional literal matching method (59.5%). The Kappa coefficient of 0.89 demonstrates strong consistency with expert annotations. While traditional methods rely solely on text similarity—resulting in accuracy rates below 40% for culture-specific terms (e.g., "Yijing/Artistic Conception," "Qiyun/Spirit Resonance") and multi-dimensional terms (e.g., "Solitary Angler on a Wintry River")—the proposed model significantly improves alignment precision for these categories by incorporating cultural adaptation coefficients and structured semantic analysis.

Method/Model | Accuracy (%) | Kappa Coefficient | Culture-Specific Term Accuracy (%) | Multi-Dimensional Term Accuracy (%)

3.3.2 细分类型适配性

The accuracy rates for each of the seven mapping types were calculated, as shown in . The results indicate significant variation across the different categories. The highest accuracy rates were achieved by the =EQ (Exact Equivalence), EQ+ (Intersecting Compound Equivalence), and EQ | (Parallel Compound Equivalence) types, all of which exceeded 92%. The accuracy rates for BM (Broader Mapping) and NM (Narrower Mapping) were also relatively high, at 88.7% and 90.2%, respectively. In contrast, the RM (Related Mapping) type exhibited the lowest accuracy at 70.0%.

These results objectively reflect the structural limitations of Western vocabularies. For instance, the Art & Architecture Thesaurus (AAT) does not include core aesthetic categories of Chinese painting and calligraphy, such as "Yijing" (artistic conception) or "Qiyun" (spirit resonance). Consequently, these terms can only be linked to broad concepts like "aesthetics" through "Related Mapping," leading to lower matching precision. This further validates the necessity of the "Cultural Adaptation Coefficient" proposed in this study.

3.3.3 核心模块有效性

The "Cultural Adaptation Coefficient," "Four-Dimensional Semantic Decomposition," and "Dynamic Priority Rules" constitute the three core modules of the model proposed in this study. To verify the effectiveness of these modules, we designed three sets of ablation experiments. By removing a specific module from the complete model and measuring the accuracy of the remaining architecture, we evaluated their individual contributions. The comparative results are presented in .

Upon sequentially removing these three core modules, the model' s accuracy decreased by 10.7%, 13.9%, and 27.1%, respectively. These results clearly quantify the contribution of each component to the overall performance, thereby validating the necessity and synergistic value of the core algorithmic modules. The "Cultural Adaptation Coefficient" module serves as the core for capturing the semantics of culture-specific terminology, such as "artistic conception" (*yijing*); it functions as the foundation for quantifying cultural dimensions and ensuring matching precision. The "Four-Dimensional Semantic Decomposition" module utilizes structured semantic extraction to prevent the loss of meaning in complex terms, providing the basis for multi-dimensional computation. Finally, the "Dynamic Priority Rules" effectively prevent the misclassification of matching types, serving as the key mechanism for ensuring the accuracy of matching type determination.

The synergy between these three components forms a comprehensive technical solution covering cultural dimension quantification, semantic parsing, and relationship determination. This framework provides a reusable methodology for terminology alignment within similarly complex domains of the humanities.

4.1 算法创新与核心发现

To address the "information silo"dilemma in the cross-lingual retrieval of Chinese painting and calligraphy terminology, this study constructs an algorithm-driven alignment framework integrated with cultural adaptation. This framework was empirically validated using TPM-CPCV and the Art & Architecture Thesaurus (AAT). The algorithmic innovations and findings are summarized as follows:

(1) We propose a multi-dimensional alignment model that integrates cultural semantics, facilitating a methodological evolution from "empirical judgment" to "quantitative calculation." This study quantifies "cultural correlation"into a computable "cultural adaptation coefficient."By combining text, semantics, and domain weights, we constructed a weighted comprehensive scoring model. Experimental results demonstrate that the model achieves an overall matching accuracy of 89.2%, representing a 29.7% improvement over traditional literal matching methods. Notably, the accuracy for matching culture-specific terms (such as *yijing* [artistic conception] and *qiyun* [spirit resonance]) increased to 82.1%, effectively addressing the issue of insufficient semantic capture caused by the absence of cultural dimensions.

(2) We designed a mapping determination mechanism based on dynamic priorities, significantly enhancing the interpretability and systematic nature of alignment results. By defining seven categories of mapping relationships and designing a dynamic determination sequence of "$= EQ >\sim EQ > EQ+ > EQ \mid> BM > NM > RM$," the system systematically handles complex semantic competition and boundary cases between terms, thereby avoiding type confusion. Ablation experiments indicate that this module contributes 27.1% to the overall accuracy, serving as a critical component for ensuring consistency in matching type determination.

(3) This research quantitatively reveals the structural differences between Chinese and Western art terminology systems, providing a path-based foundation for cross-lingual knowledge fusion. The results show that approximately 30% of terms can be linked through "exact equivalence," providing an entry point for low-cost system implementation. Conversely, core aesthetic categories such as *yijing* are absent in the AAT and can only be aligned indirectly through "associative mapping." The relatively lower accuracy (70.0%) for these terms empirically confirms the fundamental differences between Chinese and Western classification logics at the data level.

4.2 研究局限与展望

Although this study achieved its expected results, several limitations remain that warrant further refinement in future work. (1) The scope of the corpus needs to be expanded. Current experiments primarily focus on thematic and technical terminology in painting and calligraphy; future research should incor-

porate additional dimensions, such as materials and mounting techniques, to test the model' s comprehensiveness. (2) The depth of cultural adaptation can be further explored. The current "Cultural Adaptation Coefficient" is primarily based on word frequency calculations. Future studies could explore integrating historical context and cross-cultural interpretive differences to improve the quantitative precision of terms with strong cultural associations. (3) The generalization capability of the model requires continuous verification. A key direction for the next stage is to validate the effectiveness of this framework on other multilingual vocabularies (such as UDC and Iconclass) and on painting and calligraphy terminology in low-resource languages. (4) The construction of a knowledge contribution mechanism is currently missing. Present research focuses on the comparison and discrepancy analysis between TPM-CPCV and the AAT, but it has not yet established a complete mechanism for moving from "discovering differences" to "supplementing and improving." In the future, we hope to formally incorporate core concepts unique to Chinese painting and calligraphy that are not yet included in the AAT—such as *cunfa* (texture strokes), *tiba* (inscriptions), and *qianyin* (sealing)—into the AAT system through a standardized proposal process, thereby achieving a systematic contribution of Chinese art terminology to international vocabularies.

【References】 [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] Chen S J, Chen H H. Semantic mapping of Chinese-English controlled vocabularies in the field of Chinese art [J]. Journal of Library and Information Science, 2015, 13(2): 161- [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] Zhuang Y. Knowledge organization of museum collections for artificial intelligence: A case study of the "Conceptual Reference Model for Ancient Chinese Movable Cultural Relics" of the Palace Museum [J]. Palace Museum Journal, 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.

University Press of America, 1998. [8] Mikolov T, Le Q V, Sutskever I. Exploiting similarities among languages for machine translation [J/OL] . arXiv, [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, Vuli gaard A. A s urvey 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, G bor 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] Zhang June. Research on the localization of the Art & Architecture Thesaurus (AAT) [J]. Digital Library Forum, 2015(7): 36-43.

[15] Zhang Meng. Iconography and the construction of image databases [D]. Beijing: Chinese National Academy of Arts, 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] Shao, Luoyang. *Dictionary of Chinese Art* [M]. Shanghai: Shanghai Lexicographical Publishing House, 2002.

[19] Cheng, Fuwang. *Dictionary of Chinese Aesthetic Categories* [M]. Beijing: Renmin University of China Press, 1995.

[20] Pustejovsky, J. *The Generative Lexicon* [M]. Cambridge: The MIT Press, 1995.

[21] Chen, Shujun. *A Study of Chinese-English Lexical Semantic Correspondence in Thesauri: A Case Study of the Chinese Art Domain* [D]. Taipei: National Taiwan University, 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

Abstract

This study explores the methodological framework of conceptual structure alignment, semantic decomposition, and multi-dimensional fusion within the context of digital humanities. By integrating machine learning and deep learning techniques, we propose a systematic approach to bridge the gap between heterogeneous data sources and complex humanistic inquiries. The research emphasizes the importance of maintaining semantic integrity while enabling cross-disciplinary data synthesis.

1. Introduction

The rapid evolution of digital humanities has necessitated more sophisticated methods for processing and interpreting vast arrays of cultural and historical data. Traditional computational approaches often struggle with the nuanced and polysemic nature of humanistic texts. To address these challenges, this paper introduces a framework centered on conceptual structure alignment and semantic decomposition. These processes allow for a more granular analysis of information, facilitating a multi-dimensional fusion that respects the original context of the data while enabling large-scale comparative studies.

2. Conceptual Structure Alignment

Conceptual structure alignment serves as the foundational layer for integrating disparate datasets. In the realm of digital humanities, data often originates from different eras, languages, and archival standards. Alignment involves identifying equivalent or related concepts across these diverse schemas.

2.1 Mapping Heterogeneous Schemas

The alignment process begins with the identification of core entities and their relationships. Given two conceptual spaces $\mathcal{S}_1$ and $\mathcal{S}_2$, the goal is to find a mapping function $f : \mathcal{S}_1 \to \mathcal{S}_2$ such that the semantic distance between aligned nodes is minimized. This is often represented as:

$$\min \sum_{i,j} w_{ij} \cdot d(c_{1,i}, c_{2,j})$$

where d is a distance metric in the embedding space and w_{ij} represents the strength of the conceptual link.

3. Semantic Decomposition

Semantic decomposition involves breaking down complex concepts into their constituent parts to better understand their underlying meanings and historical shifts. This is particularly useful for analyzing evolving terminology in historical documents.

3.1 Decomposition Models

By applying deep learning architectures, we can decompose a high-level concept C into a set of latent features $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$. This allows researchers to track how specific dimensions of a concept—such as its political, social, or religious connotations—change over time. Using a decomposition matrix, we can represent the concept as

Note: Figure translations are in progress. See original paper for figures.