Hotspot Exploration of Museums in China Based on the K-means Model

Zhongsheng Wang[1,a *], Minlu Wang[1,b] and Qiang Shi[1,c]

[1] Xi'an Technological University

[a]email:wzhsh1681@163.com, [b]milo124@163.com, [c]75883289@qq.com

Abstract. [Objective/Significance] Over the past two decades, museums have witnessed flourishing growth and have evolved into open public spaces that encompass functions such as exhibition, education, collection, research, and services. Interdisciplinary and cross-domain research has emerged as a prevailing trend in academic investigations. Analyzing the literature hotspots within the context of museums serves to assist scholars in swiftly grasping prominent issues and forthcoming trends in this field, as well as furnishing a theoretical foundation for interdisciplinary convergence. [Method/Process] This study takes the years from 2016 to 2023, focusing on museums as its principal theme. It employs complex network techniques to construct a co-occurrence network of literature keywords. Gephi network analysis and visualization tools are harnessed for the exploration of the thematic landscape embedded within museum-related literature. [Conclusion] Following the identification of hot research topics in museum-related literature, this study offers a comprehensive analysis of the salient features characterizing existing literature hotspots. It endeavors to proffer hot topics and research directions in the domain of museum-related studies, thereby providing valuable guidance and reference points for subsequent research endeavors in this field.

Keywords: Museums; Social networks; Literature hotspots

1. Problem Statement

Research on "museums" plays a crucial role in the academic domain, facilitating an in-depth exploration of multiple dimensions within art, culture, history, and society. It offers a rich array of research themes and interdisciplinary opportunities, contributing significantly to cross-disciplinary research and theoretical development in academia. Through scholarly investigations into museums, we gain a better understanding of the diverse and intricate aspects of human culture. Social network analysis holds a pivotal position in contemporary research, aiding not only in comprehending relationships between pieces of information but also in revealing the hotspots and future trends within research fields [1].

This paper focuses on museums as the subject of study and employs keyword co-occurrence networks and social network techniques [2]. Utilizing data mining and visualization tools, we conduct an in-depth examination of relevant literature from the period spanning 2016 to 2023. The aim of this paper is to propose hot topics and research directions within the domain of museum-related studies, offering valuable guidance and references for future research endeavors. We aspire to contribute to the advancement of research and development in the field of museums.

2. Research Methodology and Framework

This study employs a systematic research approach to investigate and analyze research themes and hotspots within the chosen subject. The research framework is depicted in Figure 1 and consists of the following key steps:

Extraction of Semantic Representations using TF-IDF: To commence, we utilize the TF-IDF (Term Frequency-Inverse Document Frequency) method to extract semantic representations from the literature. This step involves quantifying the significance of specific terms within the corpus and capturing the underlying semantics of the documents.

Discovery of Research Clusters using the K-means Clustering Algorithm: Subsequently, we apply the K-means clustering algorithm to identify clusters within the literature where research

topics are concentrated. This clustering technique helps uncover coherent themes within the body of literature, facilitating subsequent analysis and interpretation.

Construction of a Keyword Co-occurrence Network using Complex Network Techniques: To further elucidate keyword relationships and reveal the interconnections between research topics, we employ complex network techniques to construct a keyword co-occurrence network. This network highlights the associations between different keywords, allowing us to identify central terms and their interactions.

Gephi Network Analysis and Visualization: Utilizing the Gephi network analysis and visualization tool, we delve into the constructed keyword co-occurrence network. This step enables us to visually explore the hot research topics within the field of our study. By analyzing the network's structure, we can pinpoint key themes and their relationships, shedding light on overarching trends and research focal points.

Through these systematic procedures, our aim is to provide valuable insights into the current landscape of research related to our subject, identifying prominent areas of interest. The research methodology and framework, as depicted in Figure 1, enable us to contribute to a comprehensive understanding of the dynamic field of study.

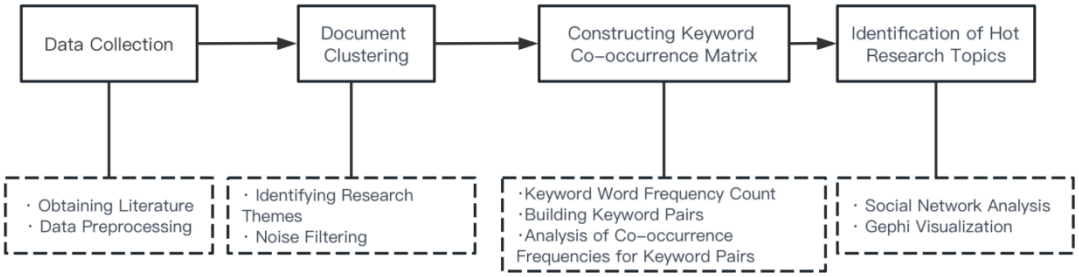

Figure 1. Research Approach and Framework

2.1 Data Collection. To acquire literature related to the keyword "museum," data were retrieved from the CNKI database for the time range from 2016 to 2023. After multiple searches, a web crawler code was developed using Python and the Selenium framework to extract data such as titles, authors, keywords, abstracts, and publication years from the search results. In total, 3782 pieces of literature were collected.

2.2 Data Preprocessing. Data preprocessing involved several steps to refine the dataset for analysis. Duplicate entries were removed, and non-Chinese literatures as well as empty content were excluded. This resulted in a dataset of 3001 pieces of literature that met the research requirements. Keyword extraction was also performed as part of data preprocessing, with the goal of preparing text data for subsequent analysis and modeling. The first step involved tokenization, which was carried out using the Jieba Chinese text segmentation tool to split Chinese text into individual words or tokens. The second step was stopword removal, where common words with little analytical value, known as stopwords, were eliminated from the text data to reduce noise and retain meaningful vocabulary, thereby improving the quality of subsequent analysis.

2.3 Text Vectorization - Feature Matrix Construction. Text vectorization was performed using the Bag-of-Words (BOW) model, which disregards factors such as word order and grammar, treating text as a simple collection of words. Each word's occurrence in a document is considered independently, irrespective of other words. Sentence vectors were first constructed, with the dimensionality of each sentence matching that of the vocabulary. The value in the i-th dimension represented the frequency of the word with ID i in that sentence. The CountVectorizer class from the sklearn library was employed to implement text frequency counting and vectorization, converting vectorized text data into a word frequency matrix. CountVectorizer transforms a collection of text documents into a token count matrix, where each word in the text is assigned a unique number, and the matrix represents the count of each word in the original text.

2.4 Clustering Analysis of Literature on "Museums". The K-means algorithm, an unsupervised clustering algorithm based on prototypes, was applied to cluster 3001 pieces of literature related to "museums." This algorithm partitions samples into K clusters based on the shortest Euclidean distance, aiming to make points within clusters as close as possible and maximize the distance between clusters. K-means is known for its efficiency and is one of the most commonly used clustering algorithms. [3][4]

To visualize high-dimensional datasets, t-SNE (t-distributed Stochastic Neighbor Embedding) was used for dimensionality reduction, as shown in Figure 2. Compared to other dimensionality reduction techniques like PCA, t-SNE creates a smaller feature space where similar samples are modeled by nearby points, and dissimilar samples are modeled by high-probability distant points. t-SNE converts the similarity between data points in the original space into conditional probabilities. Similarity in the original space is represented by a Gaussian joint distribution, while similarity in the embedded space is represented by a Student's t-distribution. The KL divergence, a measure of dissimilarity between probability distributions, is used to evaluate the quality of the embedding. The KL divergence function is employed as the loss function to minimize through gradient descent, resulting in a converged solution.

Through experimentation, a total of 10 clusters were obtained:

Cluster 0: Contains 484 pieces of literature with the theme of theoretical research and curation.

Cluster 1: Includes 60 pieces of literature focused on exhibition-related topics.

Cluster 2: Encompasses 347 pieces of literature centered on art and theoretical research.

Cluster 3: Comprises 117 pieces of literature related to the modernization of China.

Cluster 4: Contains 64 pieces of literature providing exhibition information.

Cluster 5: Involves 120 pieces of literature introducing museums.

Cluster 6: Consists of 213 pieces of literature focused on academic research, including architectural aspects.

Cluster 7: Comprises 73 pieces of literature concentrated on art research.

Cluster 8: Encompasses 109 pieces of literature related to public research.

Cluster9: Contains 1412 pieces of literature pertaining to art education, appreciation, and opening consultations.

Additionally, there are 1792 pieces of literature that were not clustered. After in-depth analysis, these unclustered literatures can be categorized into three situations:

2.4.1 The first situation involves literature with broad research themes and a wide scope, lacking a specific focal point, such as "Frontiers of Art Management: Building Museums in the New Era – An Interview with Li Lei, Executive Curator of the China Art Palace" ("Art Management," 2019).

2.4.2 The second situation includes literature with clearly defined research topics but a niche focus, such as "Wei Encircling Zhao: On the 'Fine Pillar' Strategy" ("Architectural Technique," 2021).

2.4.3 The third situation comprises noisy data that, despite preprocessing, still contains some noise, such as "Quotations" ("Art Market," 2022). This category includes interviews that appeared as consultation data or data that remained difficult to filter through clustering due to their noisy nature.

It's important to note that multiple experiments were conducted to achieve a balanced and effective clustering result. Based on summarization and categorization, the above-mentioned themes can be grouped into four categories: disciplinary development and theoretical research, functional research related to public services, research on the operational models of museums, and trends in data intelligence development.

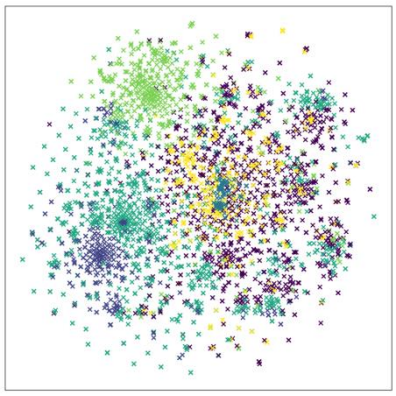

Figure 2. Dimensionality-Reduced Visualization of Clustering Results

In the previous step, we obtained a total of 1,209 pieces of literature across four main themes, containing a combined total of 8,970 unique keywords. After removing duplicate words, we extracted 5,007 distinct keywords. When two different keywords co-occur in the same literature, they form a binary relationship. Due to the extensive nature of the keyword co-occurrence network, we selected the top 200 keywords for analysis, resulting in 40,000 binary relationships, as depicted in Figure 3.

From a social network perspective and utilizing Gephi visualization techniques, this study calculated the co-occurrence frequency of different keywords in each piece of literature within the dataset, thereby constructing a keyword co-occurrence network. It's important to note that in this network, the multiple keywords within each piece of literature are considered as parallel, with no directional relationships, making the keyword co-occurrence network an undirected network.

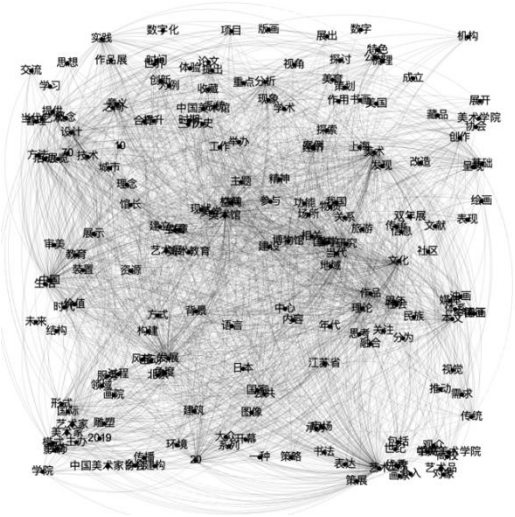

Figure 3. Gephi Visualization Analysis

3 Research Findings

3.1 Research Directions of Hot Topics

3.1.1 Rooted in Disciplinary Development and Theoretical Research

The integration of disciplines and theories with the current Chinese context is an emerging research direction of interest to scholars. Some scholars have proposed the concept of "New Museum Studies," suggesting that it should transcend traditional museum studies and approach the relationship between museums and artistic disciplines from a more comprehensive perspective. This approach focuses on cultural awareness, encompassing research on both the historical evolution and institutional development of traditional museums and the contemporary issues raised during the development of modern museums, including exhibitions, collections, and public education.

Scholars emphasize three core aspects of the "New Museum": the role of "people" as subjects, the multidimensional issue of "public space," and the critical examination of "institutional" frameworks. [5]

In addition, some scholars have explored theoretical aspects by drawing inspiration from Susan Sontag's collection of essays, "Against Interpretation." They emphasize the significant value of this theory in shaping public educational methods in museums. They propose that, in the realm of museum public education, greater emphasis should be placed on the most basic and direct communication between individuals and artworks. This approach involves respecting the audience's most immediate experiences, providing opportunities for interaction, touch, and sensory perception, and aligning with human nature and adaptability. [6]

3.1.2 Research on the Functions of Public Services

Public cultural services have long been a hotspot in the field of "museum" research. Museums, as public platforms for displaying and studying artworks, play a crucial role in disseminating art knowledge and transmitting human cultural values. They also serve as the primary battleground for fostering positive values, enhancing the aesthetic and innovative abilities of the public. [6] Numerous studies focus on the public nature of museums and call for the establishment of a public cultural service system. Scholars like Du Qun have delved into the development of public cultural service systems in state-owned museums in the new era. They emphasize three important research directions: 1) optimizing the forms of exhibition services and innovating public educational service carriers, enhancing the efficiency of academic research and collection utilization to improve service levels; 2) improving the utilization of digital applications and expanding public service channels through cross-boundary spatial expansion, enhancing the hardware and soft power of museum facilities to improve the public service environment; and 3) cultivating talent and strengthening safety awareness to ensure the management of public services.

3.1.3 Exploration of Museum Operation Models

The exploration of museum operation models is crucial for the development of private museums. Some scholars emphasize the value of museums and argue that curating, as a profession, have evolved in the field of museum studies. They propose a shift in the research paradigm toward the cultural function and strategic significance of "museum curation." They analyze classic cases of global museum curation within the theoretical framework of the "cultural happening field." They highlight the urgency of upgrading technological concepts, emphasize the cultural production function of "museum curation," and reveal that the "cultural domain" of museums has evolved into the new concept of a "cultural happening field." [7]

Others focus on the narrative aspects of exhibitions, aiming to establish communicative temporary communities that engage with the audience. [8] They believe that exhibitions should serve as platforms for meaningful communication between the museum and its visitors, fostering an environment where visitors can actively engage in dialogue and exchange ideas.

These various research directions highlight the multidimensional nature of museum studies and underscore the evolving roles of museums in society. Researchers are increasingly exploring the intersections of art, culture, and public engagement, as well as the dynamic relationship between museums and their communities.

3.1.4 Trends in the Smart Development of Museums

In the context of ongoing digital transformation, some scholars believe that the application and development of museums are undergoing a transformation towards a digital strategy characterized by high technology, high intelligence, and extensive services. The museums of the future are expected to leverage emerging technologies such as digital technology, the Internet, cloud computing, and big data to create a new architectural system that is audience-centric, real-time, responsive, and service-oriented. This system aims to digitize core operations like exhibitions, collections, academic research, and public education while breaking down data silos, establishing connections and sharing of business data, facilitating intelligent decision-making, enhancing the capacity for public cultural services, and ultimately creating museums of the future that are entirely dedicated to their audiences. [9]

After studying the impact of digitization on museums, some scholars have shifted their focus to examine how digitization affects researchers. They have researched the integration of social media and digital image collection within American cultural institutions, with a particular emphasis on the challenges and opportunities this technology brings to image collection platforms. [10]

4. Summary

This paper adopts a social network analysis approach to examine research literature related to "museums" from 2016 to 2023. It constructs a keyword co-occurrence network and applies the k-means algorithm to categorize pre-processed text data. The text data, after processing, undergoes dimensionality reduction using t-SNE for visualization. Gephi network analysis and visualization tools are utilized to explore the processed text data. The analysis identifies four popular research directions: "Theoretical Studies," "Public Services," "Operational Models," and "Smart Development." Finally, through cross-validation with a substantial body of literature, this paper interprets the directions of these popular themes, providing guidance and references for future researchers.

References

[1] Wu Xiaoqiu and Lv Na, Research on Hotspot Analysis Method Based on Keyword Co-occurrence Frequency, Information Studies: Theory & Application, Vol. 35, no. 8, pp. 115 – 119, 2012, doi: 10.16353/j.cnki.1000-7490.2012.08.026.

[2] Zhang Jie and Wang Hong, Comparative Analysis of Domestic and Foreign Research Hotspots in Mobile Learning Based on Word Frequency Analysis and Visualization of Co-occurrence Networks, Modern Distance Education, no. 2, pp.76 – 83, 2014, doi: 10.13927/j.cnki.yuan. 2014.02.006.

[3] Number of Clusters and Initial Center Point Self-Determination of K-means Algorithm - China National Knowledge Infrastructure. Accessed: Sep. 28, 2023. [Online]. Available: https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C44YLTlOAiTRKibYlV5Vjs7i0-kJR0HYBJ80QN9L51zrP2Z0XrEa7w0kSjiiSgVMaLdv-AJ_Uc_SXyX_4QZXH7Am&uniplatform=NZKPT

[4] Quan Yufei et al., Multi-objective Stochastic Programming and Clustering Analysis of Reservoir Group Water Storage Scheduling, Journal of Hydroelectric Engineering, pp. 1 – 10.

[6] Han Jing, Reflection on the Public Education Mode of Art Museums—Review of 'Public Education in Art Museums', Educational Development Research, Vol. 41, no. 11, p. 2, 2021, doi: 10.14121/j.cnki.1008-3855.2021.11.001.

[7] Wang Meng, From the Production of 'Museum Curator' Behavior to the 'Cultural Generative Field', Art, no. 12, pp. 108 – 111, 2020, doi: 10.13864/j.cnki.cn11-1311/j.006189.

[8] He Xiaote, Image History and Its Art History Issues: On the Curatorial Ideas of 'Towards the Ocean: Works on the Theme of Ocean Construction in the 1950s to 1970s in Guangzhou Academy of Fine Arts', Art Observation, no. 11, pp. 35 – 36, 2019.

[9] Du Qun, Governance and Exploration of 'Digital Intelligence' in Future Art Museums, New Art, Vol. 42, no. 4, pp. 243 – 248, 2021.

[10] Zhang Shaoqian, From Reception to Interaction: Authors and Readers in Digital Image Sharing, Art Observation, no. 4, pp. 21 – 23, 2021.