Advances in Computer, Signals and Systems (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2026.100212
ISSN 2371-8838 Vol. 10 Num. 2

Construction of a Performance Arts Knowledge Graph Based on Large Language Models and Its Geospatial Distribution Analysis

Hualing Gao, Pingping Jiao, Jiaojiao Lu

School of Information and Intelligence Engineering, University of Sanya, Sanya, 572022, China

Keywords: Knowledge Graph, Large Language Models (LLM), Performance Arts, Entity Resolution, Visual Analysis

Abstract: Performance arts data is characterized by unstructured formats, multi-source heterogeneity, and sparse spatio-temporal distribution, rendering traditional management methods insufficient for revealing the underlying complex social networks and resource allocation patterns. This paper presents a framework for the construction and analysis of a performance arts knowledge graph powered by Large Language Models (LLMs). The approach utilizes LLMs to extract named entities and relational structures from unstructured performance records, integrated with an entity resolution algorithm to address naming redundancy and inconsistencies. A knowledge graph is constructed via the Neo4j graph database, where degree centrality and spatial radiation indices are applied to analyze core patterns in the performance market regarding subject distribution and geographical circulation. Experimental results indicate that the performance arts market exhibits a significant power-law distribution, with core leading entities demonstrating a strong spatial correlation between their total performance frequency and geographical reach.

1. Introduction

Current integration of digital media and the cultural industry has resulted in a vast accumulation of unstructured data within the performance arts market. This data typically exhibits a dispersed and fragmented distribution, lacking a unified relational logic, which hinders the analysis of resource allocation and artist mobility patterns. As a semantic organizational structure, the Knowledge Graph (KG) provides a robust approach for deciphering such complex interconnected data.

Recent studies have utilized KG technology to analyze industrial ecosystems, with significant applications in fields such as the film industry and art exhibitions (Hogan et al., 2021; Bode, 2017)[1][2]. Nevertheless, performance arts data, characterized by non-standardized naming and cross-domain heterogeneity, presents significant challenges for entity resolution in the construction of high-quality graphs. Existing research frequently relies on manual annotation or traditional rule-based matching, which are limited by scalability and efficiency bottlenecks. The semantic understanding and information extraction capabilities inherent in Large Language Models (LLMs) offer a feasible pathway for the efficient extraction of knowledge entities and relations from unstructured text.

This study establishes a framework for data governance and visual network analysis within the performance arts market. By leveraging LLMs to automate the processing of unstructured data, this work facilitates the automated extraction and standardized mapping of performance entities, while employing network analysis to quantify artists' activity levels and spatial reach. This research aims to address the gap in KG construction for the specialized performance arts sector and provide empirical evidence for the optimization of resource allocation in the cultural industry.

2. Related Work

2.1 Knowledge Graphs in the Cultural Industry

With the advancement of Digital Humanities, Knowledge Graphs (KGs) have demonstrated potential in analyzing resource distribution within the cultural industry. Existing studies primarily focus on movie recommendation systems or art auction markets, utilizing relational networks between entities to optimize resource allocation (Hogan et al., 2021)[1]. However, the performance arts market, as a sector highly dependent on geographical space, involves data characterized by significant real-time dynamics and heterogeneity. Compared with static cultural heritage data, the dynamic evolutionary characteristics of performance market data impose more stringent requirements on the construction of KGs.

2.2 Information Extraction with Large Language Models

Traditional knowledge extraction methods rely heavily on deep learning models (e.g., BERT, BiLSTM-CRF), which are typically constrained by the scale of labeled datasets and limited in their semantic understanding of unstructured text (Shen et al., 2015)[3]. Recently, the generalized representation capabilities of Large Language Models (LLMs) have provided a new paradigm for automated knowledge extraction. Existing research has validated the efficiency of LLMs in extracting information from non-standardized text, and the integration of LLMs with KGs has further expanded the boundaries of domain-specific knowledge base construction (Pan et al., 2024)[4]. Currently, research specifically targeting data cleaning and entity resolution in the performance arts sector remains in an exploratory stage, which serves as the primary research niche for this study.

3. Methodology

This study constructs an automated KG construction framework for performance arts data, consisting of three main phases: data preprocessing, LLM-driven information extraction, and rule-based entity resolution.

3.1 Data Acquisition and Preprocessing

The raw data is derived from multi-channel unstructured performance information, including artist lists, performance titles, and venue locations. To mitigate noise, we first conduct standardized preprocessing by removing irrelevant formatting characters and retaining essential entity text.

3.2 Knowledge Extraction and Representation

This research employs a few-shot instruction tuning strategy for information extraction. By constructing structured prompts, the Large Language Model is guided to transform unstructured performance text into specific entity-relation-entity triples. This process defines the relational

attributes between artists and performance events while utilizing contextual comparison techniques to distinguish entity references across semantic environments, ensuring the precision of the initial graph structure.

3.3 Entity Resolution

To address pervasive naming heterogeneity in performance data, this study proposes a canonical key-based entity resolution algorithm. This approach first extracts the core name of an entity using regular expression matching and subsequently utilizes atomic merge operations in the graph database to redirect all semantically identical heterogeneous nodes to a unique canonical entity. By executing node merging and relationship migration, the system ensures the logical integrity of historical performance associations during graph evolution, effectively reducing graph redundancy.

4. Experiments and Results Analysis

This section presents a quantitative analysis of the constructed performance arts knowledge graph to validate the resource allocation patterns in the performance market.

4.1 Data Scale and Governance Effectiveness

Following the preprocessing and entity resolution algorithm, this study successfully constructed a performance arts knowledge graph in the Neo4j graph database, comprising 767 artist nodes and 4,735 performance event nodes, with 1,562 valid "PERFORMED_IN" relations established. During the data governance phase, the canonical-key-based disambiguation algorithm identified and merged 16 redundant nodes resulting from naming heterogeneity, ensuring the uniqueness of entity identities within the collaboration network. Comparing data density before and after governance demonstrates that this algorithm corrected centrality calculation biases caused by naming heterogeneity, thereby improving the statistical accuracy of degree centrality for core artists.

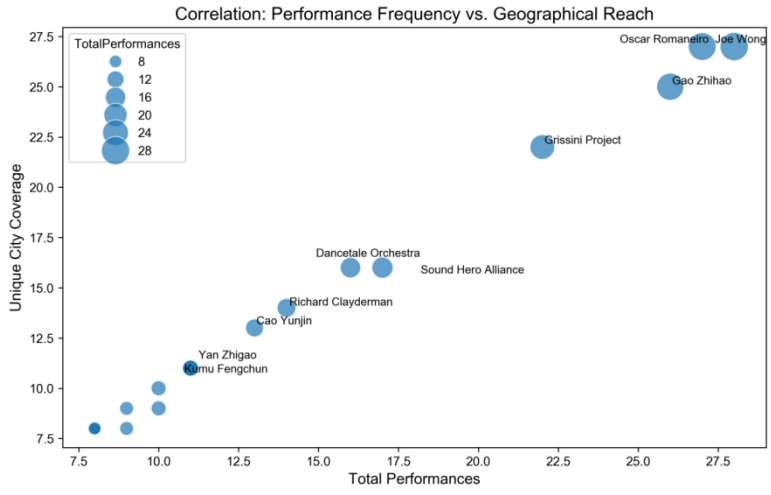

Figure 1: Correlation between performance frequency and geographical reach of top artists

4.2 Spatial Circulation Analysis

Figure 1 illustrates the distribution of performance subjects concerning their Total Performances and Unique City Coverage. The experimental data reveals a high positive correlation between these variables. As depicted, leading performance subjects achieve extensive geographical reach through

high-frequency touring strategies, constructing a nationwide influence matrix. This distribution validates the head effect within the performance market, confirming that performance frequency is a primary driver for the expansion of geographical radiation capacity.

4.3 Centrality Characteristics of Geospatial Resource Circulation

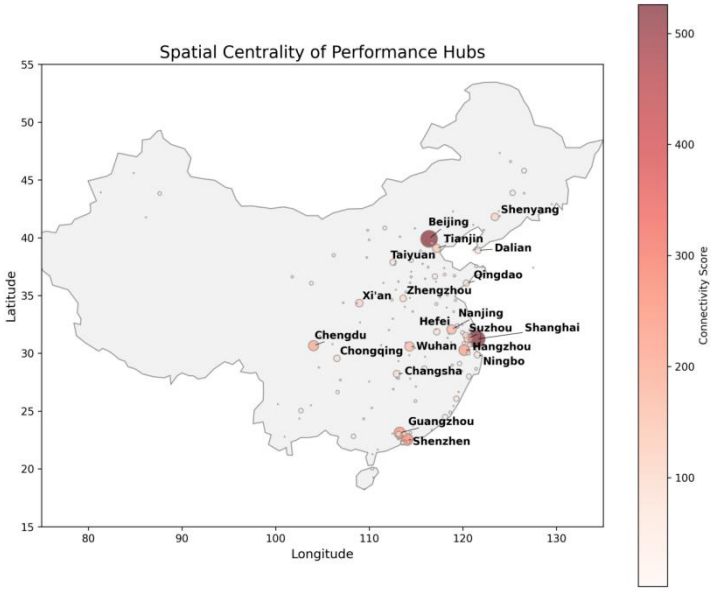

Figure 2. Spatial polarization and diffusion pattern of performance arts resources

The evolution of the performance market reflects the logic of resource allocation across geographical space. Figure 2 illustrates the geospatial distribution of the performance market in China, which reveals a distinct spatial polarization effect. Quantitative analysis of degree centrality indicates that Shanghai and Beijing, with connectivity scores of 526 and 515 respectively, occupy the core positions in national performance resource allocation, forming the primary centers of the national market. They are closely followed by regional central cities such as Shenzhen, Guangzhou, and Hangzhou, which collectively support the secondary radiation network of the performance market. From a macroscopic perspective, the geographical distribution of performance resources is non-uniform, following a "Core-Periphery" hierarchical diffusion model.

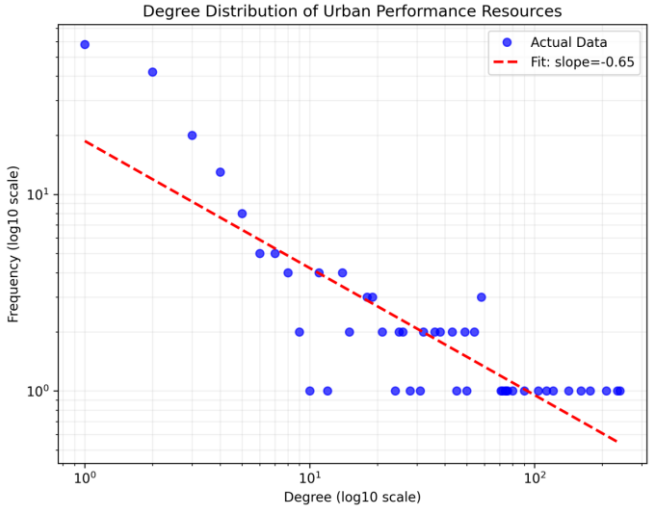

Figure 3: Degree distribution of performance resources across city nodes (log-log scale).

The performance market exhibits a pronounced polarization, following a typical power-law distribution. This structure aligns with the characteristics of scale-free networks in complex social systems (Barabási, 2009)[5], revealing that the diffusion of performance resources is not a linear process, but rather highly dependent on the spillover effects of core hubs. As shown in Figure 3, quantitative analysis of the degree distribution of city nodes confirms this power-law characteristic. Although the power-law exponent $\gamma \approx 0.65$ suggests a persistent head effect, the distribution demonstrates greater resource diffusion compared to standard scale-free networks. This indicates that while the Chinese performance market maintains the dominance of core hubs, it is increasingly undergoing a structural transition characterized by resource downward-filtering toward mid-sized cities.

5. Discussion

This study constructs a knowledge graph for the performance industry and reveals the characteristics of resource allocation in the market through quantitative analysis. Experimental results indicate a significant "head effect" in the distribution of performance resources, reflecting the market's high dependency on influential artists. Furthermore, the strong positive correlation between the geographical reach and performance frequency of artists confirms that "touring," as a method of spatial capital operation, serves as a core strategy for artists to establish nationwide influence and optimize resource allocation.

Although this study has achieved effective outcomes in data governance and knowledge graph construction, certain limitations remain: first, information extraction based on Large Language Models (LLMs) may involve semantic deviations; second, the current graph update relies on offline data collection, which limits its timeliness. Future research can explore the integration of real-time streaming data to further enhance the capabilities of the knowledge graph in dynamic analysis and trend prediction for the performance market.

6. Conclusion

This study proposes a framework for the construction and analysis of a performance arts knowledge graph based on Large Language Models. By employing automated information extraction and entity resolution methods, this research converts unstructured performance data into a structured knowledge network, addressing the challenges of data sparsity and integration in the performance industry. Quantitative analysis based on the graph reveals the logic of resource allocation across geographical space, validating the distinct "Core-Periphery" distribution pattern and providing a quantitative basis for understanding the macro-circulation mechanism of performance resources. Furthermore, the methodology proposed in this paper demonstrates strong generalizability and offers a reference for constructing domain-specific knowledge graphs in other cultural and artistic sectors.

Acknowledgement

This research was supported by the Higher Education Scientific Research Project of the Hainan Provincial Department of Education (Grant No. Hnky2025-29) and the National Training Program of Innovation and Entrepreneurship for Undergraduates (Grant No. 202513892029).

References

[1] Hogan A , Blomqvist E , Cochez M ,et al. Knowledge Graphs[J]. ACM Computing Surveys (CSUR), 2021.
[2] Bode, Katherine. The Equivalence of "Close" And "Distant" Reading; Or, toward a New Object for Data-Rich

Literary History[J]. Modern Language Quarterly, 2017, 78(1):77-106.
[3] Shen W, Wang J, Han J. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions[J]. Knowledge & Data Engineering IEEE Transactions on, 2015, 27(2):443-460.
[4] Pan S, Luo L, Wang Y, et al. Unifying Large Language Models and Knowledge Graphs: A Roadmap[J]. IEEE Transactions on Automatic Control, 2024, 36(7):20.
[5] Barabási, Albert-László. Scale-Free Networks: A Decade and Beyond [J]. Science, 2009.