Ensemble Making Few-Shot Learning Stronger

Authors: Liu Yongbin

Date: 2022-11-15T00:00:00+00:00

Abstract

Few-shot learning has been proposed and rapidly emerging as a viable means for completing various tasks. Many few-shot models have been widely used for relation learning tasks. However, each of these models has a shortage of capturing a certain aspect of semantic features, for example, CNN on long-range dependencies part, Transformer on local features. It is difficult for a single model to adapt to various relation learning, which results in a high variance problem. Ensemble strategy could be competitive in improving the accuracy of few-shot relation extraction and mitigating high variance risks. This paper explores an ensemble approach to reduce the variance and introduces fine-tuning and feature attention strategies to calibrate relation-level features. Results on several few-shot relation learning tasks show that our model significantly outperforms the previous state-of-the-art models.

Full Text

Preamble

RESEARCH PAPER

Uncovering Topics of Public Cultural Activities: Evidence from China

Zixin Zeng & Bolin Hua†

Department of Information Management, Peking University, Beijing 100871, China

Keywords: Public culture; Short text clustering; Topic modeling; LDA; Big data analysis

Citation: Zeng, Z., & Hua, B. (2022). Uncovering topics of public cultural activities: Evidence from China. *Data Intelligence*, 4(3). doi: 10.1162/dint_a_{00121}

Received: July 28, 2021; **Revised:** November 27, 2021; **Accepted:** February 9, 2022

ABSTRACT

This study uncovers the topics of Chinese public cultural activities in 2020 using a novel two-step approach combining short text clustering (self-taught neural networks and graph-based clustering) with topic modeling. Our dataset comprises over 17,000 articles collected from 108 websites of libraries and cultural centers. Through this framework, we derive 3 clusters and 8 topics across 21 provincial-level regions in China. By plotting the topic distribution of each cluster, we reveal distinct tendencies among local cultural institutes: free lessons and lectures on art and culture, entertainment and service for socially vulnerable groups, and the preservation of intangible cultural heritage. These findings provide decision-making support for cultural institutes, thereby promoting public cultural services from a data-driven perspective.

1. INTRODUCTION

Public cultural activities refer to cultural events organized by eight types of public institutions under the supervision of China' s Ministry of Culture and Tourism (i.e., libraries, cultural centers, museums, art museums, community art centers, science museums, memorials, and Children' s Palaces) that aim to facilitate public welfare [**?**]. With the rapid development of big data theory and technology in recent years, numerous governments and public cultural institutions have attached great importance to big data practices in public culture [**?**]. In China, the establishment of the national public culture cloud platform in 2017 served as a catalyst for developing local public culture digital platforms [**?**], which disseminate diverse information including available digital resources, upcoming cultural events, and cultural services [**?**].

The thriving big data practices of public cultural institutes have paved the way for big data research on public culture services. By integrating and mining public cultural big data, it becomes possible to gain profound insights into different areas and users, thereby supporting the decision-making processes of public cultural institutes [**?**]. This paper focuses on the topics of public cultural activities, as versatile and attractive activities are crucial for promoting local citizen participation and building an inclusive cultural atmosphere. Furthermore, we analyze public cultural activities at the provincial level, which reveals regional tendencies and aids cultural institutes in striking a balance between adhering to macroscopic cultural policies, following emerging trends, and establishing unique cultural identities.

We propose a novel framework for modeling topics of public cultural activities through short text clustering[1]. Following Xu et al. [**?**], we train a self-taught CNN (convolutional neural network) to obtain deep text representations, then employ the K-means algorithm to assign cultural activity texts to various clus-

ters. The obtained cluster labels are used to compute a graph with provinces as nodes. Subsequently, SCAN (Structural Clustering Algorithm for Networks) [**?**] is applied to the graph to derive clusters of provinces. Finally, we use LDA (Latent Dirichlet Allocation) to extract topic words for each cluster. This two-step clustering approach enables us to identify common patterns of public cultural activities at the provincial level, allowing for fine-grained analysis of activity features across provinces. Our motivation for extracting regional features of public cultural activities is twofold. First, compared to qualitative research methods that require tedious manual labor and may be vulnerable to subjective biases, our text clustering and topic modeling approach leverages open-access Internet data to provide an efficient method for analyzing current trends in cultural activities. Second, we strive to collect data from all provinces in China to form a comprehensive view of public cultural service development, helping government officials formulate actionable insights for future policies.

To demonstrate this approach' s effectiveness, we collected over 17,000 articles concerning public cultural activities in 2020 using web crawlers from 108 official websites of public libraries and cultural centers in China, providing a comprehensive report of Chinese public cultural activities. Results indicate that the COVID-19 outbreak hampered public cultural activities in spring, and geographical imbalance is evident in both the total number and density of cultural activities. Overall, the 21 regions we analyzed fell into 3 clusters (with Gansu as an outlier), and 8 distinct topics were extracted. By comparing each cluster' s topic distribution, we characterize each cluster' s unique features.

The major contributions of this paper are:

- This is the first paper to conduct a thorough data-driven analysis using text mining techniques on public cultural activities, based on a self-constructed comprehensive dataset.
- We propose a novel framework integrating two clustering algorithms and one topic modeling algorithm, which is extensible to corpora with geographic features.
- Our approach delineates characteristics of public cultural activities in each region, providing decision-making support for cultural institutes.

The remaining paper is organized as follows. Section 2 discusses prior work on public culture and text clustering. Section 3 provides a detailed description of the proposed short text clustering framework. Section 4 presents findings from Chinese public cultural activities in 2020. Finally, Section 5 concludes and discusses future research directions.

2.1 Big Data Research on Public Cultural Services

In recent years, big data research on public cultural services has developed rapidly. Among public cultural institutes, digital libraries have received con-

[1]Code is available at: https://github.com/zixinzeng-jennifer/public-culture-activity/

siderable attention as venues for both social assembly and digital connectivity, representing one of the most valuable sources of public cultural big data [?, ?]. Analytics of library big data (both catalog and transactional data) support digital library innovations, providing immeasurable value for librarians, users, and services [?].

Cao, Liang, and Li [?] emphasized the importance of building smart libraries, which cannot be achieved without smart technology—namely integrating advanced technologies such as data mining and artificial intelligence. Kamupanga and Yang [?] point out that big data technologies can forecast user habits more accurately, helping build better recommendation systems that save time and improve library user efficiency. With the advent of public cultural cloud platforms, other cultural institutes—especially local cultural centers—have been analyzed from various perspectives, including user satisfaction, content and characteristics, and classification systems [?, ?, ?].

Partly due to the difficulty of integrating heterogeneous data from multiple sources, relatively fewer empirical and quantitative studies have been conducted on public cultural services compared to theoretical analyses [?, ?]. Li and Hua [?] proposed the overall structure and content of big data research on public cultural services, emphasizing the feasibility and necessity of data-driven research. Bratt and Moodley [?] analyzed economic and employment disparities by applying data mining techniques to annual survey results of U.S. public libraries, providing recommendations for advancing data accessibility and transparency. Wei [?] constructed a multi-layer regression model on survey data to explore the mechanism of cultural participation behavior. Zhang et al. [?] analyzed spatiotemporal patterns in public cultural service construction in China, reflecting regional development of public cultural services.

Compared to traditional qualitative methods such as questionnaires and reviews, data-driven methods require significantly less human labor and cover a wider range of public cultural institute users. Data-driven research on public cultural services presents both challenges and opportunities for researchers and practitioners in Library and Information Science (LIS) by providing profound insights into the effects of policies and systems designed for user-centered public cultural services.

2.2 Short Text Clustering

Text clustering groups texts so that those in the same cluster are more similar to each other than those in other clusters. Since most clustering algorithms rely on numerical features, transforming texts into vectors is a vital step. Traditional approaches use shallow representations such as the bag-of-words model, where each word represents one dimension in vector space, often weighted by Term Frequency (TF) or Term Frequency-Inverse Document Frequency (TF-IDF).

However, this approach can be problematic for short texts due to data scarcity, leading to advances in deep learning-based text clustering—namely deep clus-

tering. In recent years, neural networks have become increasingly popular for computing text embeddings. Xu et al. used a self-trained convolutional neural network to obtain denser short text representations [**?**]. Similarly, Hadifar et al. [**?**] proposed a multi-phase self-trained approach that fine-tunes an autoencoder for optimal embeddings. Other researchers have tackled the issue from the neural topic modeling perspective. Wang et al. [**?**] applied bidirectional adversarial training based on the Dirichlet prior. Costa and Ortale [**?**] jointly train text clustering and topic modeling tasks via a Bayesian generative process. Overall, deep clustering (clustering with neural networks) has proven more effective than traditional methods given ample data.

2.3 Topic Modeling

Topic modeling extracts topics (similar semantic patterns) from document sets, often approached computationally as a dimensionality reduction problem. Latent Semantic Analysis (LSA) constructs a term occurrence matrix from the corpus and uses singular value decomposition to extract low-dimensional text representations [**?**]. Due to its matrix factorization procedure, LSA is not scalable for large text collections. Probabilistic Latent Semantic Analysis (PLSA) is a generative model using latent class variables to generate each word in a document [**?**], but it is prone to overfitting with large document collections [**?**]. These disadvantages led to LDA, a generative probabilistic topic modeling algorithm based on Bayesian statistics. More recent advances include Topic2Vec [**?**], which learns distributed topic representations similarly to Word2Vec [**?**] but measures similarity via distance metrics such as cosine similarity, potentially making topics highly correlated and difficult to interpret. Some Transformer-based approaches such as BERTopic[2] have been proposed, which uses BERT embeddings as input for class-based TF-IDF (c-TF-IDF) topic extraction. For a comprehensive summary of topic modeling algorithms, we refer readers to literature reviews [**?**, **?**]. This paper uses LDA because it is easy to implement and has demonstrated robust performance across diverse real-world applications.

3. METHODOLOGY

The framework of this study is illustrated in Figure 1 [Figure 1: see original paper]. This framework enables clustering regions according to cultural activity content and explaining clustering results through topic modeling.

Figure 1. Text Clustering and Topic Modeling Framework.

3.1 Data Preparation

Our study collects cultural activity articles from 108 official websites of public cultural activities and culture centers across multiple Chinese regions using

[2]https://github.com/MaartenGr/BERTopic/

Scrapy, a popular web crawling framework. These official websites were selected based on the following criteria:

- Out of China' s 34 provincial-level administrative regions, we eliminated the three special administrative regions (Hong Kong, Macau, and Taiwan).
- For the remaining regions, our primary data sources were provincial-level public libraries and cultural centers/public cultural cloud platforms.
- When provincial-level institutes suffered from data scarcity, we supplemented our corpus with public cultural activity articles from city-level or district-level public cultural institutes in that region.

Among the 108 official websites, 81 were managed by cultural centers (29 province-level, 34 city-level, and 18 district-level) and 27 by libraries (17 province-level and 10 city-level). It is not unusual for lower-level public cultural institutes to have more abundant data, as some serve as demonstrative zones. Complementing the corpus with corresponding city-level or district-level data is also plausible because these institutes are often tightly bound administratively. For more information on these public cultural institutes, please refer to our supplemental materials.

The public cultural articles used in this study are notices of upcoming activities and should be differentiated from news articles reporting past events. Table 1 presents metadata for the collected cultural activity articles. Note that the availability of some variables, such as activity type, varies depending on each website' s design and is therefore marked as optional.

Table 1. Metadata of Cultural Activity Articles.

Variable Name	Explanation
Pav Name	Public culture institute managing the website
Activity Name	Name of the cultural activity
Activity Time	Starting date of the cultural activity (YYYY-MM-DD format)
Place	Detailed address of the cultural activity
Remark	Link to cultural activity article
Activity type (Optional)	Description of the activity
Organizer (Optional)	Type of activity (e.g., exhibition, show, lecture)
Contact (Optional)	Institute or committee organizing the activity
Presenter Introduction (Optional)	Phone number or email address

For this study, we limit our data scope to events organized in 2020 because many public cultural institutes only began posting articles in the past 2-3 years,

making a cross-sectional study more suitable. In fact, only 52 public institutes had public cultural articles before 2019, and the total number of articles has increased drastically in recent years, as shown in Fig. 2 [Figure 2: see original paper].

Each activity was assigned to a provincial-level administrative region based on the geographical location of the public culture institute. Table 2 describes our data distribution. Imbalance in data collection across regions is evident, primarily because some regions posted few cultural activities on their websites. We eliminated regions with fewer than 50 articles from text clustering and topic modeling analysis due to data scarcity, resulting in 21 retained regions. Possible reasons for data scarcity include:

- Public cultural institutes in that region were not enthusiastic about organizing cultural activities.
- Public cultural institutes in that region were not accustomed to posting information online.
- The region established its website recently, so few cultural activity articles have accumulated.
- Some regions limited public cultural activity organization in accordance with COVID-19 quarantine measures.

We use activity name and description to constitute our dataset. In the data preprocessing step, we remove duplicate articles by computing Levenshtein Distance and perform word segmentation with the jieba package while removing stopwords.

3.2 Neural Short Text Clustering

As shown in Fig. 3 [Figure 3: see original paper], most cultural activity articles are relatively short, containing fewer than 500 characters. For short texts, vectors obtained from the Bag of Words model are extremely sparse, which can be problematic for clustering algorithms.

Inspired by Xu et al. [**?**], we train a self-taught CNN model to obtain embeddings for each article. We use Laplacian Eigenmaps (LE)[3], an unsupervised dimensionality reduction method, to produce denser representations Y of each text; subsequently, real-valued vectors Y are transformed into binary codes B using the median as threshold, which trains the CNN. Our CNN model structure is shown in Fig. 4 [Figure 4: see original paper]. The model uses pretrained vectors developed by Li et al. [**?**], and dropout with a 50% rate was employed for regularization. Afterwards, the classic K-means algorithm assigns each article to a cluster.

Figure 3. Length of Cultural Activity Articles.

Figure 4. Self-Taught CNN Model.

[3]This algorithm was chosen based on evaluation results in prior work.

We compare self-taught CNN to two baselines: bag-of-words (BoW) representation with TF-IDF weights and average embedding (AE) with TF-IDF weights, using three commonly-used clustering evaluation metrics (Silhouette Score, Calinski-Harabasz Score, and Davies-Bouldin Score). Table 3 summarizes the evaluation results, showing that self-taught CNN produces significantly better clusters than baseline methods.

Table 3. Evaluation of Clustering Results.

Clustering Method	Silhouette Score	Calinski-Harabasz Score	Davies-Bouldin Score
BoW + K-Means	0.097	97,756.102	2.314
AE + K-Means	0.123	124,832.541	2.156
Self-Taught CNN + K-means	**0.156**	**156,234.877**	**1.987**

Note: Number of clusters is 6.

3.3 Graph-Based Clustering

The similarity between cultural activities in two regions is computed using the Jaccard similarity coefficient:

$$Jaccard(P_x, P_y) = \frac{|S_x \cap S_y|}{|S_x \cup S_y|}$$

where P_x denotes region x, and S_x denotes all cluster labels assigned to cultural activity articles in region x. A simple undirected graph $G = \langle V, E \rangle$ is defined, with regions as vertices, and an edge drawn between two vertices if their Jaccard similarity coefficient exceeds threshold m.

We employ a graph-based clustering algorithm named SCAN [**?**] to cluster regions. The similarity between two vertices is defined as:

$$\sigma(x, y) = \frac{|C(x) \cap C(y)|}{|C(x)| \cdot |C(y)|}$$

where $C(x)$ denotes the set including vertex x and all its adjacent vertices, so vertices sharing more similar neighbors have higher similarity. The algorithm starts from core vertices and searches for clusters based on connectivity, marking two special vertex types: hubs (vertices reachable by more than one cluster) and outliers (vertices unreachable by any cluster).

3.4 Topic Modeling

To explain topics underlying each cluster, we employ LDA, a classic topic modeling algorithm. LDA assumes each article is generated through a sampling process where each document has a topic distribution and each topic has a word distribution:

$$p(w|d) = \sum_t p(w|t) \cdot p(t|d)$$

where w denotes word, d denotes document, and t denotes topic. From the algorithm' s output, we assign a topic to each document by defining the article' s topic as:

$$Topic(d) = \arg\max_t p(t|d)$$

Each topic t is characterized by the top k words with highest conditional probability $p(w|t)$.

4.1 Exploratory Data Analysis

This subsection visualizes spatiotemporal features of cultural activities in our dataset.

Figure 5 [Figure 5: see original paper] shows the number of cultural activities organized each month. Over 2,500 activities were organized in January 2020, around the time of Chinese Spring Festival—one of China' s most important holidays. However, the total number dropped sharply in February and gradually increased in subsequent months. Activity numbers rose steadily from March to May, fluctuated mildly from June to November, and moderately decreased in December. This pattern likely relates to COVID-19 quarantine policies, which were most rigid in February, with citizens gradually returning to school and work from March to May. The December decrease may reflect annual reviews when many institutes wrap up and reflect on the year' s work.

Figure 5. Number of Cultural Activities Organized Each Month.

We also analyze the total number of cultural activities per region. Regions can be divided into five categories: dense (more than 5 activities daily), frequent (approximately 2 activities daily), moderate (1 activity daily), scarce (1 activity weekly), and unavailable data. Dense activity regions such as Guangdong, Hunan, and Chongqing typically host over 5 activities per day. Frequent regions like Beijing, Shandong, Jiangsu, and Hunan average about 2 activities daily. Moderate regions organize 1 activity daily, while sparse regions only organize 1 weekly. Generally, Southern and Eastern China have more cultural activities.

We define cultural activity density as the number of activities divided by regional area (in km^2). The regions with highest density are the four municipalities:

Beijing, Tianjin, Shanghai, and Chongqing. As municipalities represent some of China' s most economically prosperous regions, cultural activity density may be positively correlated with overall economic development. Indeed, the Spearman correlation coefficient r between disposable income per capita[4] and cultural activity density was as high as 0.766 ($p < 0.001$).

4.2 Text Clustering and Topic Modeling

This subsection discusses short text clustering and topic modeling results. Using the two-step clustering approach described above, we derived 3 clusters (containing 10, 6, and 4 regions respectively) and 1 outlier, summarized in Table 4 . Figure 6 [Figure 6: see original paper] visualizes the graph constructed according to the Jaccard coefficient, positioning nodes using the Kamada-Kawai layout.

Table 4. Summary of Clustering Results.

Clusters	Members
Cluster 1	Liaoning, Hubei, Fujian, Hainan, Zhejiang, Shanxi, Shaanxi
Cluster 2	Jilin, Inner Mongolia, Sichuan, Beijing, Anhui, Yunnan, Tianjin, Shandong, Shanghai
Cluster 3	Guangdong, Hunan, Jiangsu, Chongqing
Outlier	Gansu

Figure 6. Clustering Results.

Table 4 and Figure 6 show that provincial regions in Cluster 1 (except Zhejiang) are located in central China with moderate economic development. Three of the four municipalities were assigned to Cluster 2, possibly because Chongqing has a much larger area than the other three municipalities. Moreover, provinces in Cluster 3 all scored relatively high on cultural activity density.

Using the LDA algorithm, we estimate the appropriate number of components with a topic coherence measure proposed by Mimno et al. [**?**]. The document frequency term in this metric was estimated by sampling 140,000 articles from THUCTC, a large-scale Chinese news dataset. The topic coherence score for each LDA model was computed by averaging across all topics, and the model with the highest score was selected, as shown in Fig. 7 [Figure 7: see original paper]. This process extracted 8 topics from the cultural activity articles, with the top 10 keywords for each topic summarized in Table 5 . The 8 topics partially overlap but each remains unique. For example, both Topic 1 and Topic 8 concern

[4]Statistics available at http://www.stats.gov.cn/

lectures organized by public cultural institutes, but Topic 1 emphasizes lecture content and location while Topic 8 emphasizes lecture audiences.

Figure 7. Topic Coherence Score.

Each article' s topic is defined as the topic with the highest conditional probability. Figure 8 [Figure 8: see original paper] shows each cluster's topic distribution, calculated by counting topic labels for all articles in each cluster. For clusters 1, 2, and 3, the most frequent topic is Topic 1—lectures on various art forms and zeitgeist at local libraries—demonstrating that local libraries play an extremely important role in organizing public cultural activities.

Table 5. Topics of Cultural Activities.

Topic	Top Keywords (Translated)	Explanation
1	works, library, art, read, lecture, service, calligraphy, learn, people, zeitgeist	Lectures on zeitgeist and various art forms at local libraries
2	show, benefit the people, drama, rural, grassroots, opera, campus, culture center, group, education	Cultural shows for rural residents and students
3	community, volunteer, civilized, traditional holidays (Spring Festival, Mid-Autumn Festival, Dragon Boat Festival), health, service, elder, harmonious	Voluntary service at community centers, especially for the elderly during traditional holidays
4	museum, relic, exhibition, paper-cutting, book review, history, archaeology, skill, hulusi (flute), the Forbidden City	Exhibitions on art, history, and archaeology at museums
5	culture center, training, citizen, class, art, music, concert, public welfare, free, face-mask	Free art and music lessons at culture centers for public welfare
6	lotus, recreational, rural, competition, photography, fishing, recommend, scenic spot, popularize knowledge, relic	Recreational activities at scenic spots in rural areas
7	intangible cultural heritage, dance, travel, program, tradition, competition, ethnicity, people, drama, music	Intangible cultural heritage, especially of ethnic minorities
8	children, story, movie, history, literature, lecture, university, language, parent, wisdom	Lectures on various topics (e.g., literature) for children

Figure 8. Topic Distribution of Each Cluster.

For Cluster 1 regions, other popular topics include free art and music lessons at culture centers, intangible cultural heritage (especially of ethnic minorities), and lectures for children. This shows that public institutes in Cluster 1 focus on public libraries with special emphasis on reading and learning. For example, Zhejiang Library has invested heavily in electronic resources, offering diverse academic and popular databases to the public, including KUKE (digital music) and Scopus. These rich resources provide excellent opportunities for local citizens to learn new knowledge at libraries.

For Cluster 2 regions, public cultural institutes often organize activities on cultural shows for rural residents and students, followed by voluntary service at community centers (especially for the elderly). Cultural institutes in Cluster 2 emphasize promoting cultural services for special social groups, potentially improving social equality from a public cultural perspective. For instance, as a province with over 25 ethnic minorities, Yunnan strives to preserve residents' arts. Yunnan Cultural Center runs programs introducing intangible cultural heritage to the public, including the Yi knife dance, Wa knitting techniques, and Dai wall paintings. Yunnan Library also maintains a multimedia database containing records of 15 ethnic minority groups unique to Yunnan.

For Cluster 3 regions, public cultural institutes emphasize intangible cultural heritage preservation, followed by free art and music lessons at culture centers for public welfare. Public cultural institutes in these regions enthusiastically preserve cultural traditions. For example, Guangdong Cultural Center hosts monthly lectures on traditional Chinese medicine, helping citizens better understand traditional medical practices that have flourished in China for millennia.

Finally, Gansu' s public cultural activities exclusively share the topic of free art and music lessons at culture centers for public welfare, suggesting Gansu may have more monotonous topics than other regions. As one of China's economically underdeveloped regions, it is unsurprising that Gansu lags behind in cultural activity diversity. Currently, Gansu' s cultural institutes encourage citizens to develop reading habits by regularly hosting reading activities.

5. CONCLUSION

This study collected over 17,000 articles from 108 official websites of public libraries and cultural centers across China. Analysis of spatiotemporal features reveals fewer public cultural activities in spring, likely due to COVID-19, with activity numbers increasing as quarantine measures relaxed. The total number and density of public cultural activities are imbalanced across regions, with more activities in Eastern and Southern China (especially the Yangtze River Delta and Pearl River Delta), and the highest cultural activity density in the four municipalities (Shanghai, Tianjin, Chongqing, and Beijing). This suggests that activity numbers may relate to regional economic development, informatization,

and population density. Regions with many ethnic minorities, like Yunnan, also exhibit rich cultural activities.

We further uncovered cultural activity topics using a two-step text clustering and topic modeling approach. A self-taught CNN was trained for article embeddings, on which classic K-means was applied to obtain cluster labels. We then computed an undirected graph based on the Jaccard similarity coefficient of cluster labels from each region pair. The graph-based clustering algorithm SCAN derived region clusters. To explain clustering results, we used LDA to extract various topics characterized by the most important keywords in each topic. By plotting each cluster' s topic distribution, we uncovered unique tendencies of local cultural institutes when organizing activities.

Overall, most regions organized lectures on art and zeitgeist at local libraries. Public cultural institutes play crucial roles in knowledge dissemination and art popularization. Local libraries are enthusiastic promoters of knowledge, often hosting reading activities and educational lectures. Cultural centers focus more on art popularization by organizing performances in underdeveloped areas and building databases and archives for intangible cultural heritage.

Our clustering and topic modeling results show that different regions vary in their public cultural activity focus. Some regions strive to provide rich educational resources, others focus on promoting services for special social groups (e.g., rural residents and ethnic minorities), and still others emphasize preserving cultural traditions. Our study also reveals that Gansu' s cultural activities lack diversity, potentially negatively influencing participation.

6. DISCUSSION

The crucial aim of providing public cultural service is promoting public welfare by satisfying citizens' cultural needs—such as receiving education, preserving traditional culture, and enjoying cultural works as entertainment. Essentially, better public cultural service requires understanding citizens' needs and efficiently allocating public cultural institute resources. Many government officials and scholars have proposed theories and roadmaps for improving public cultural service quality, including providing equal and accessible services for all society members [**?**, **?**], extending public cultural service providers [**?**, **?**], and emphasizing regional characteristics [**?**, **?**]. Despite abundant theoretical frameworks, few papers analyze public cultural service using empirical data, with prior work often relying heavily on qualitative methods like field surveys and limited in scope (often case studies). Data-driven methods such as text mining prove promising for understanding both citizens' cultural needs and current public cultural services.

This paper focuses on understanding public cultural activity trends (an important public cultural service component). We propose a text clustering and topic modeling framework for fine-grained analysis of Chinese public cultural activity characteristics and assess 2020 trends using a self-constructed dataset. Com-

pared to traditional surveys or fieldwork, our approach provides satisfactory results with significantly less manual labor. This study is also the first comprehensive public cultural service overview from a national perspective, offering insights for future policy formulation. While public cultural services have been organized relatively independently by regional institutes, the recent National Public Cloud Platform (https://www.culturedc.cn/) demonstrates the feasibility and necessity of assessing services more holistically. We hope our findings help officials gain actionable insights from current trends and inform future cultural policies.

Our public cultural activity dataset provides detailed, authentic information on service content and characteristics across regions. Despite extensive collection efforts, we observed data availability imbalance, eliminating articles from several regions due to scarcity. With public cultural institutes' informatization, we anticipate richer public culture data will become available. We also note that textual data from public cultural institutes are highly unstructured, so further investigation of information extraction algorithms may help understand various activity aspects, such as organizers, presenters, and overall scale.

To uncover cultural activity topics across regions, we jointly used short text clustering (self-trained CNN), graph clustering (SCAN), and topic modeling (LDA). While confident this framework suits our purposes, recent natural language processing advances offer alternative approaches. For example, texts could be encoded via Transformer models like BERT, fuse structural and semantic information via graph convolutional networks, and use newer topic modeling algorithms like BERTopic. These NLP algorithms warrant exploration in future public cultural service text mining research. Our LDA algorithm relies on the bag-of-words model and conditional independence assumption, losing semantic information. Ideally, topics could be mined more integrally, perhaps using multi-text summarization techniques. Additionally, evaluating clustering and topic modeling quality remains challenging, though we used unsupervised metrics like Silhouette Score and Topic Coherence Score to guide hyperparameter selection. Further examination by public cultural experts could validate our findings.

We highlight several future public cultural service research directions. First, this paper focuses only on text mining to understand public cultural activities. Future work could collect empirical data on citizens' cultural needs and other service aspects, helping officials and institutes better understand trends and challenges from a data-driven perspective. Second, summarizing and presenting data mining findings is an important research direction. Our text clustering and topic modeling results could be integrated into a public cultural service visualization system incorporating various aspects—such as geographic institute locations, citizen participation numbers, and activity topic distributions—to help citizens and officials better understand public cultural services. Finally, studying temporal evolution of public cultural service trends over longer periods will become feasible as more data accumulates on official websites. Understanding

how services change over time will reveal consistency and future directions for public cultural service provision.

ACKNOWLEDGEMENTS

This article is an outcome of the key laboratory project "Research on the Wisdom Mode Clustering and Dynamic Display System of Public Culture" (No. 2020008) supported by the Ministry of Culture and Tourism of the People' s Republic of China.

AUTHOR CONTRIBUTIONS

Z. Zeng (zixinzeng_{jennifer}@pku.edu.cn) was responsible for data collection, code implementation, and writing this paper. B. Hua (huabolin@pku.edu.cn) proposed the research topic and revised this paper.

REFERENCES

[1] Wan, L.: Public culture and its development in contemporary China. *Journal of Renmin University of China* 1, 98–103 (2006)

[2] Cao, L., Ma, C.: The study of domestic and international big data practice in public culture. *Library Journal* 34(12), 9–15 (2015)

[3] Wei, J., Wang, Y.: Empirical research on user satisfaction of National Public Culture Cloud Platform. *Information and Documentation Services* 41(4), 30–38 (2020)

[4] Chen, Z., Liu, Y., Nie, Q.: Analysis of service content and characteristics of public cultural cloud in China. *Library* 8, 27–31, 46 (2018)

[5] Li, G., Hua, B.: A tentative model for big data research on public cultural services. *Library Tribune* 38(7), 62–71 (2018)

[6] Xu, J., et al.: Self-taught convolutional neural networks for short text clustering. *Neural Networks* 88, 22–31 (2017)

[7] Xu, N., et al.: SCAN: A structural clustering algorithm for networks. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 824–833 (2007)

[8] Wyatt, D., McQuire, S., Butt, D.: Library as producer of public culture. In: *Public Libraries in a Digital Culture*, pp. 20–27. Available at: http://arts.unimelb.edu.au/___{data}/assets/pdf_{file}/0005/1867865/PublicLibrariesinaDigitalCulture.pdf. Accessed 27 November 2021

[9] Liu, S., Shen, X.: Library management and innovation in the big data era. *Library Hi Tech* 36(3), 374–377 (2018)

[10] Cao, G., Liang, M., Li, X.: How to make the library smart? The conceptualization of the smart library. *The Electronic Library* 36(5), 812–825 (2018)

[11] Kamupanga, W., Yang, C.: Application of big data in libraries. *International Journal of Computer Applications* 178(16), 34–38 (2019)

[12] Sun, J., Zheng, J.: Research on the framework of classification system for the big data of public cultural services. *Library Tribune* 40(9), 28–35 (2020)

[13] Liao, X.: Review of the research on big data of public culture and estimation of the research trends. *Library* 7, 42–49 (2019)

[14] Bratt, S., Moodley, K.: Promoting public library sustainability through data mining: R and Excel. In: *IFLA World Library and Information Congress 2015*, pp. 1–14 (2015)

[15] Wei, Y.: Individual motivation and community moderation of residents' cultural participation: Based on multi-layer linear model. *Library Tribune* 41(6), 56–66 (2021)

[16] Zhang, Y., et al.: Study on spatio-temporal differentiation and influencing factors of public cultural service construction in China. *Library Development* 6, 165–174, 183 (2021)

[17] Hadifar, A., et al.: A self training approach for short text clustering. In: *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pp. 194–199 (2019)

[18] Wang, R., et al.: Neural topic modeling with bidirectional adversarial training. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 340–350 (2020)

[19] Costa, G., Ortale, R.: Jointly modeling and simultaneously discovering topics and clusters in text corpora using word vectors. *Information Sciences* 563, 226–240 (2021)

[20] Landauer, T.K., Foltz, P.W., Laham, D.: Introduction to latent semantic analysis. *Discourse Processes* 25(2–3), 259–284 (1998)

[21] Hofmann, T.: Probabilistic latent semantic indexing. *ACM SIGIR Forum* 51(2), 211–218 (2017)

[22] Likhitha, S., Harish, B.S., Keerthi Kumar, H.M.: A detailed survey on topic modeling for document and short text data. *International Journal of Computer Applications* 178(39), 1–9 (2019)

[23] Niu, L., et al.: Topic2Vec: Learning distributed representations of topics. In: *2015 International Conference on Asian Language Processing (IALP)*, pp. 193–196 (2015)

[24] Mikolov, T., et al.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781v3 (2013)

[25] Alghamdi, R., Alfalqi, K.: A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications* 6(1), 147–153 (2015)

[26] Li, S., et al.: Analogical reasoning on Chinese morphological and semantic relations. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 138–143 (2018)

[27] Mimno, D., et al.: Optimizing semantic coherence in topic models. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 262–272 (2011)

[28] Wanyan, D., Wang, Z.: An empirical analysis of the regional equalization of public digital culture service in China. *Research on Library Science* 5, 50–58, 66 (2020)

[29] Xiao, X., Wanyan, D.: Research on the practice of promoting the equalization of basic public cultural services by digitization. *Library Work and Study* 8, 5–10 (2016)

[30] Pan, Y., Sun, H., Zheng, J.: Research on the development path of rural public culture under the background of culture and tourism integration. *Library Tribune* 41(3), 68–77 (2021)

[31] Li, S., Wang, T.: Participation logic and behavior strategy of multi-dimensional subject in public cultural services—an observation of policy implementation on creation demonstration area of national public cultural service system. *The Journal of Shanghai Administration Institute* 19(5), 61–69 (2018)

[32] Zhong, Y.: Research on the public cultural service system in promoting the construction of local characteristic information resources. *The Library Journal of Shandong* 2, 5–9 (2018)

[33] Lin, T.: Shaping communities: Building the regional embeddedness of public cultural services. *Administrative Tribune* 28(5), 105–114, 112 (2021)

AUTHOR BIOGRAPHY

Zixin Zeng is an undergraduate student at Peking University. Her research interests include text summarization, machine translation, and knowledge graph.

Bolin Hua received his Ph.D. in Information Resource Management from Nanjing University. He is currently an Associate Professor in the Department of Information Management at Peking University and has published over 60 papers on text mining, intelligence analysis based on big data, and big data of public cultural services.

ORCID: 0000-0001-9248-6455

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