Uncovering Topics of Public Cultural Activities: Evidence from China (Postprint)

Authors: Zixin Zeng, Bolin, Hua, Bolin, Hua

Date: 2022-11-28T00:00:00+00:00

Abstract

In this study, we uncover the topics of Chinese public cultural activities in 2020 with a two-step short text clustering (self-taught neural networks and graph-based clustering) and topic modeling approach. The dataset we use for this research is collected from 108 websites of libraries and cultural centers, containing over 17,000 articles. With the novel framework we propose, we derive 3 clusters and 8 topics from 21 provincial#2; level regions in China. By plotting the topic distribution of each cluster, we are able to shows unique tendencies of local cultural institutes, that is, free lessons and lectures on art and culture, entertainment and service for socially vulnerable groups, and the preservation of intangible cultural heritage respectively. The findings of our study provide decision-making support for cultural institutes, thus promoting public cultural service from a data-driven perspective.

Full Text

Preamble

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.: Uncovering topics of public cultural activities: Evidence from China. Data Intelligence 4(3), 509-528 (2022). doi: 10.1162/dint_a_{00121}

Data Citation: Zeng, Z., & Hua, B.: Public Cultural Activities in 2020. Data Intelligence 4(3), (2022). doi: https://doi.org/10.57760/sciencedb.j00104.00107

ChinaRxiv 中X

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

ABSTRACT

In this study, we uncover the topics of Chinese public cultural activities in 2020 with a two-step short text clustering (self-taught neural networks and graph-based clustering) and topic modeling approach. The dataset we use for this research is collected from 108 websites of libraries and cultural centers, containing over 17,000 articles. With the novel framework we propose, we derive 3 clusters and 8 topics from 21 provincial-level regions in China. By plotting the topic distribution of each cluster, we are able to show unique tendencies of local cultural institutes, that is, free lessons and lectures on art and culture, entertainment and service for socially vulnerable groups, and the preservation of intangible cultural heritage respectively. The findings of our study provide decision-making support for cultural institutes, thus promoting public cultural service from a data-driven perspective.

1. INTRODUCTION

Public cultural activities refer to cultural activities organized by eight types of public institutes under the supervision of the Ministry of Culture and Tourism in China (i.e., libraries, cultural centers, museums, art museums, community art centers, science museums, memorials, and Children' s Palaces) that aim at facilitating public welfare [1]. With the rapid development of big data theory and technology in recent years, a great number of governments and public cultural institutions have attached great importance to big data practice in public culture [2]. In China, the establishment of the national public culture cloud platform in 2017 served as a catalyst for the development of local public culture digital platforms [3], which disseminate a variety of information, including available digital resources, upcoming cultural events, and cultural services [4].

The thriving big data practice of public cultural institutes has paved the way for big data research on public culture services. By integrating and mining public cultural big data, it would be possible to gain profound insights into different areas and users, in turn supporting the decision-making process of public cultural institutes [5]. In this paper, we focus on the topics of public cultural activities, as versatile and attractive activities are crucial for promoting the participation of local citizens and building an inclusive cultural atmosphere. Furthermore, we analyze public cultural activities at the provincial level, which indicates the tendencies of each region, thus aiding cultural institutes in striking a balance between adhering to macroscopic cultural policies, following the newest trends, and establishing a unique cultural identity.

In this research, we propose a novel framework for modeling topics of public cultural activities with short text clustering. Following the work of Xu et al. [6], we train a self-taught CNN (convolutional neural network) to obtain deep text representations, then employ the K-means algorithm to assign the cultural activity texts to various clusters. The obtained cluster labels are used to compute a graph with nodes as provinces. Subsequently, SCAN (Structural Clustering Algorithm for Networks) [7] is run on the graph to derive clusters of provinces. Finally, we use LDA (Latent Dirichlet Allocation) to extract topic words for each cluster. With this two-step clustering approach, we are able to spot common patterns of public cultural activities at the province level, thus allowing for a fine-grained analysis of the features of public cultural activities across various provinces.

Our motivation for extracting features of public cultural activities in each region is twofold. First, compared to qualitative research methods, which tend to require tedious manual labor and may be vulnerable to subjective views, our method based on text clustering and topic modeling leverages open-access data resources on the Internet to provide an efficient approach for analyzing current trends in cultural activities. Second, we make our best efforts to collect data from all provinces in China to form a comprehensive view of how public cultural service has developed across China and help government officials form actionable insights for future policies.

To demonstrate the effectiveness of this approach, we collected over 17,000 articles concerning public cultural activities in 2020 with web crawlers from 108 official websites of public libraries and culture centers in China and provide a comprehensive report of Chinese public cultural activities based on the data. Results indicate that the outbreak of COVID-19 hampered public cultural activities in spring, and geographical imbalance in public cultural activities can be observed from both the total number and density of cultural activities. Overall, the 21 regions we analyzed fell into three clusters (with Gansu as an outlier), and eight distinct topics were extracted from our dataset. We compare the topic distribution of each cluster and show the characteristics of each cluster.

The major contributions of this paper are:

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

The remaining paper is organized as follows. In Section 2, we discuss prior work on public culture and text clustering. In Section 3, we provide a detailed description of the proposed short text clustering framework. In Section 4, findings from Chinese public cultural activities in 2020 are discussed. Lastly, in Section 5, we present a conclusion and discussion on future directions of this research.

2. RELATED WORK

2.1 Big Data Research on Public Cultural Services

In recent years, research on big data for public cultural services has been developing rapidly. Among public cultural institutes, digital libraries have received much attention because they are viewed as a place for both social assembly and digital connectivity, as well as one of the most valuable sources of public cultural big data [2, 8]. Analytics of library big data (both catalogue data and transactional data) are found to support digital library innovations, providing immeasurable value for librarians, users, and services [9].

Cao, Liang, and Li [10] emphasized the importance of building smart libraries, which could not be achieved without smart technology, namely integrating advanced technology such as data mining and artificial intelligence. Kamupanga and Yang [11] point out that big data technologies can be used for forecasting user habits more accurately, which helps build better recommendation systems, thereby saving time and improving the efficiency of library users. With the advent of public cultural cloud platforms, other cultural institutes, especially local cultural centers, have also been analyzed from various perspectives, for example, user satisfaction, content and characteristics, and classification systems [3, 4, 12].

Partly due to the difficulty of integrating heterogeneous data from multiple sources, relatively fewer empirical and quantitative research studies have been conducted on public cultural services compared to theoretical analysis [11, 13]. Li and Hua [5] proposed the overall structure and content of big data research on public cultural services and emphasized the feasibility and necessity for data-driven research in public cultural services. Bratt and Moodley [14] analyzed economic and employment disparities by applying data mining techniques to annual survey results of public libraries in the United States and provided advice for stepping toward data accessibility and transparency. Wei [15] constructed a multi-layer regression model on survey data to explore the mechanism of cultural participation behavior. Zhang et al. [16] analyzed spatiotemporal patterns in public cultural service construction in China, which reflected the development of public cultural services in various regions.

Compared to traditional qualitative methods such as questionnaires and reviews, data-driven methods require significantly less human labor and cover a wider range of users of public cultural institutes. Data-driven research on public cultural services poses both challenges and opportunities for researchers and practitioners in the field of 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 is the act of grouping a set of texts so that texts in the same cluster are more similar to each other than those in other clusters. As most clustering algorithms are based on numerical features, transforming texts to vectors is a vital step in text clustering. Traditional approaches are based on shallow representations such as the bag-of-words model, which views each word as one dimension in the 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 problems, which has led to advances in text clustering based on deep learning techniques, namely deep clustering. In recent years, neural networks have become an increasingly popular method for computing text embeddings. Xu et al. used a self-trained convolutional neural network to obtain a denser representation of short texts [6]. Similarly, Hadifar et al. [17] proposed a multi-phase self-trained approach that finetunes an autoencoder for optimal embeddings. Other researchers tackled the issue from the neural topic modeling perspective. Wang et al. [18] applied bidirectional adversarial training based on the Dirichlet prior. Costa and Ortale [19] train the tasks of text clustering and topic modeling jointly via a Bayesian generative process. Overall, deep clustering (clustering with neural networks) has proved to be more effective than traditional methods given an ample supply of data.

2.3 Topic Modeling

Topic modeling is the task of extracting topics (similar semantic patterns) from a set of documents and has often been approached computationally as a dimension reduction problem. Latent Semantic Analysis (LSA) constructs a term occurrence matrix from the corpus and uses a matrix factorization method called singular value decomposition to extract a low-dimensional representation of each piece of text [20]. Because of the matrix factorization procedure, LSA is not scalable on large amounts of text. Probabilistic Latent Semantic Analysis (PLSA) is a generative model that uses latent class variables to generate each word in a document [21]. Unfortunately, PSLA is prone to overfitting when the number of documents is relatively large [22]. These disadvantages led to the proposal of LDA, a generative probabilistic topic modeling algorithm based on Bayesian statistics.

More recent advances in topic modeling include Topic2Vec [23], which learns distributed topic representations with a mechanism that is similar to Word2Vec [24] but measured based on distance metrics such as cosine similarity, so topics can be highly correlated, and it may be hard to extract meaningful topic words for each topic. Some Transformer-based topic modeling approaches such as BERTopic have been proposed, which uses BERT embeddings as the input of class-based TF-IDF (c-TF-IDF) for the extraction of topics. For a more comprehensive summary of topic modeling algorithms, we refer the reader to literature reviews [22, 25]. In this paper, we use LDA for topic modeling because this algorithm is easy to implement and has been observed to perform robustly in a wide range of real-world applications.

3. METHODOLOGY

The framework of this study is illustrated in Figure 1 [Figure 1: see original paper]. With this framework, we are able to cluster regions according to the contents of their cultural activities and explain the 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 from multiple regions in China with Scrapy, a popular web crawling and scraping framework. The aforementioned official websites were selected according to the following criteria.

Out of the 34 provincial-level administrative regions in China, we eliminated the three special administrative regions (Hong Kong, Macau, and Taiwan). For the remaining regions, our primary data sources were the provincial-level public libraries and culture center/public cultural cloud platform in that region. In case the provincial-level institutes suffered from data scarcity, we would complement our corpus with public cultural activity articles from the city-level or district-level public cultural institutes in that region.

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

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

Table 1. Metadata of cultural activity articles.

Variable Name	Explanation
Pav	Public culture institute managing the website
Name	Name of the cultural activity
Activity Time	Starting Date of the cultural activity, in the form of YYYY-MM-DD
Place	Detailed address of the cultural activity
Remark	Link of cultural activity article
Activity type (Optional)	Description of the activity
Organizer (Optional)	Type of activity, such as exhibition, show, and lecture
Contact (Optional)	Institute or committee organizing the activity
Presenter Introduction (Optional)	Phone number or e-mail address Introduction of individual(s) presenting the activity, e.g., lecturer

For the purpose of this study, we limit the scope of our data to events organized in the year 2020 because many public cultural institutes just began posting articles in the past two to three years, thus it would be more suitable to conduct a cross-sectional study. In fact, only 52 public institutes have public cultural articles before 2019, and the total number of articles on public cultural activities has increased drastically over recent years, as shown in Fig. 2 [Figure 2: see original paper].

Each activity was assigned to a provincial-level administrative region according to the geographical position of the public culture institute. The distribution of our data is described in Table 2 . Imbalance in the amount of data collected from different regions can be clearly observed, which was primarily because these regions posted few cultural activities on their websites. We eliminated the regions with fewer than 50 articles in our analysis with text clustering and topic modeling due to scarcity of data, resulting in a total of 21 regions. Possible reasons for scarcity of data include: (1) overall, the public cultural institutes in that region were not enthusiastic about organizing cultural activities; (2) the public cultural institutes in that region were not accustomed to posting information online; (3) the region established their website recently, so few cultural activity articles have been accumulated; and (4) in concordance with quarantine measures of COVID-19, some regions limited the organization of public cultural activities.

We use activity name and description to constitute our dataset. In the data preprocessing step, we remove duplicate articles by computing the Levenshtein

Distance and complete word segmentation with the jieba package and remove stopwords.

3.2 Neural Short Text Clustering

As can be observed from Fig. 3 [Figure 3: see original paper], the majority of cultural activity articles are relatively short, with less than 500 characters. For short texts, vectors obtained from the Bag of Words model are extremely sparse, which is potentially problematic when clustering algorithms are applied.

Inspired by the work of Xu et al. [6], we train a self-taught CNN model to obtain embeddings for each article. We use Laplacian Eigenmaps (LE), an unsupervised dimensionality reduction method, to produce a denser representation Y of each text; subsequently, the real-valued vectors Y are transformed to binary codes B using the median as a threshold, which is used to train CNN. The structure of our CNN model is shown in Fig. 4 [Figure 4: see original paper]. Our model used pretrained vectors developed by Li et al. [26], and dropout with a 50% rate was employed for regularization. Afterwards, the classic K-means algorithm is applied to assign each article to a cluster.

Figure 3. Length of Cultural Activity Articles.

Figure 4. Self-Taught CNN model.

Table 3. Evaluation of Clustering Results.

Clustering Method	Silhouette Score	Calinski Harabasz Score	Davies Bouldin Score
BoW + K-Means	0.097	56,756.102	1.234
AE + K-Means	0.123	67,891.445	1.156
Self-Taught CNN + K-means	0.156	97,756.102	0.987

3.3 Graph-Based Clustering

The similarity of cultural activities in two regions is computed with the Jaccard similarity coefficient:

Jaccard(Px, Py) = |Px   Py| / |Px   Py|

where Px denotes region x, and Sx denotes all the cluster labels assigned to cultural activity articles in region x. A simple undirected graph G = <V, E> is defined, with the regions as vertices, and an edge is drawn between two vertices if their Jaccard similarity coefficient exceeds threshold m.

A graph-based clustering algorithm named SCAN [7] is employed to cluster the regions. The similarity between two vertices is defined as:

$\sigma(x, y) = |C(x) \quad C(y)| / \sqrt{}(|C(x)| \times |C(y)|)$

where C(x) denotes the set including vertex x and all adjacent vertices of x, so the more similar neighbors two vertices have, the higher their similarity. The algorithm starts from core vertices and searches for clusters based on connectivity and marks two special types of vertices: hubs (vertices that are reachable by more than one cluster) and outliers (vertices that are not reachable by any cluster).

3.4 Topic Modeling

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

$p(w|d) = \Sigma t\ p(w|t) \times p(t|d)$

where w denotes word, d denotes document, and t denotes topic. From the output of the algorithm, we can assign a topic to each document by defining the topic of an article to be:

$\mathrm{Topic}(d) = \mathrm{argmaxt}\ p(t|d)$

Each topic t is characterized by the top k words with the highest conditional probability.

4. FINDINGS

4.1 Exploratory Data Analysis

In this subsection, we visualize spatiotemporal features of the cultural activities in our dataset. The number of cultural activities organized each month is visualized in Figure 5 [Figure 5: see original paper]. As can be seen from Fig. 5, over 2,500 cultural activities were organized in January 2020, which was around the time of the Chinese Spring Festival, one of the most important holidays in China. However, in February, the total number of activities dropped sharply and gradually increased in the following months. The number of cultural activities rose steadily from March to May and fluctuated mildly from June to November, with a moderate decrease in December. It is very likely that this pattern was related to quarantine policies for COVID-19, as such measures were most rigid in February, and citizens gradually returned to school and work from March to May. The number of cultural activities may have dropped in December due to annual reviews, when many institutes wrap up and reflect on the entire year' s work.

Figure 5. Number of Cultural Activities Organized Each Month.

We also analyze the total number of cultural activities in each region. It is possible to divide regions into five categories according to the number of cultural activities: regions with dense, frequent, moderate, scarce activities, and regions where such data could not be obtained. Regions with dense cultural activities, such as Guangdong, Hunan, and Chongqing, typically have more than five cultural activities per day. Regions with frequent cultural activities, like Beijing, Shandong, Jiangsu, and Hunan, have approximately two cultural activities per day on average. Regions with moderate cultural activities organize one cultural activity per day. Finally, regions with sparse cultural activities only organize one cultural activity per week. Generally speaking, there are more cultural activities in Southern China and Eastern China.

We define the density of cultural activities in each region as the number of activities divided by the area of each region (in km^2). We discover that regions with the highest cultural activity density are the four municipalities, namely Beijing, Tianjin, Shanghai, and Chongqing. As the municipalities represent some of the most economically prosperous regions, it may be the case that cultural activity density is positively correlated with overall economic development. In fact, the Spearman correlation coefficient r of disposable income per capita and cultural activity density was as high as 0.766 ($p<0.001$).

4.2 Text Clustering and Topic Modeling

In this subsection, we discuss our results of short text clustering and topic modeling. With the two-step clustering approach previously described, we derived three clusters (containing 10, 6, and 4 regions respectively) and one outlier, summarized in Table 4 . In Figure 6 [Figure 6: see original paper], we visualized the graph we constructed according to the Jaccard coefficient, positioning nodes with Kamada-Kawai layout.

Table 4. Summary of Clustering Results.

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

Figure 6. Clustering Results.

From Table 4 and Figure 6, it can be observed that the provincial regions in Cluster 1 (except for Zhejiang) are located in the central region of China and have a moderate level of economic development. It can also be observed that three out of the four municipalities were assigned to Cluster 2, which may be due to the fact that Chongqing has a much larger area than the other three municipalities. Moreover, the provinces in Cluster 3 all scored relatively high on cultural activity density.

With the LDA algorithm, we estimate the appropriate number of components with a topic coherence measure proposed by Mimno et al. [27]. 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 topic coherence score was chosen, as shown in Fig. 7 [Figure 7: see original paper]. In this way, eight topics were extracted from the cultural activity articles and the top 10 keywords for each topic, which are summarized in Table 5 . We can see that the eight topics do partially overlap with each other, but each of them is unique. For example, topic 1 and topic 8 are both about lectures organized by public cultural institutes, but topic 1 emphasizes the content and place of lectures, while topic 8 emphasizes the audience of lectures.

Figure 7. Topic Coherence Score.

The topic of each cultural activity article is defined as the topic with the highest conditional probability. The topic distribution of the clusters, calculated by counting topic labels for all articles in each cluster, is shown in Fig. 8 [Figure 8: see original paper]. It can be observed from Fig. 8 that for clusters 1, 2, and 3, the most frequent topic is topic 1, namely lectures on various forms of art and zeitgeist at local libraries. This shows that local libraries play an extremely important role in the organization of 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 forms of art 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 at 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

Topic	Top Keywords (Translated)	Explanation
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 those 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 regions in cluster 1, other popular topics are free art and music lessons at culture centers, intangible cultural heritage (especially those of ethnic minorities), and lectures for children. This shows that public institutes from regions in Cluster 1 focus on public libraries, with a special highlight on reading and learning. For example, the Library of Zhejiang has invested much in its electronic resources and offers a variety of both academic and popular databases to the public, including KUKE (a database featuring digital music) and Scopus. These rich resources provide great opportunities for local citizens to learn new knowledge at local libraries.

For regions in cluster 2, their public cultural institutes often organize activities with the topic of cultural shows for rural residents and students, followed by voluntary service at community centers (especially for the elderly). Cultural institutes belonging to cluster 2 place great emphasis on promoting cultural services for special social groups, which may serve to improve social equality from a public cultural perspective.

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