Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00121

RESEARCH PAPER

Uncovering Topics of Public Cultural Activities: Evidence from China

Zixin Zeng[1]& Bolin Hua[1†]
[1]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
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 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.

1. Introduction

Public cultural activities refer to cultural activities organized by 8 types of public institutes under the supervision of 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 have 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 of 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 on 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 newest trends and establishing a unique cultural identity.

In this research, we propose a novel framework for modeling topics of public cultural activities with

[†]Corresponding author: Bolin Hua (Email: huabolin@pku.edu.cn; ORCID:0000-0001-9248-6455)

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00121

short text clustering[1]. 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 the 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 3 clusters (with Gansu as an outlier); and 8 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 2 clustering algorithms and 1 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 of big data on public cultural services has been developing rapidly. Among the 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 librarian, user 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

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

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00121

improving efficiency of library users. With the advent of public cultural cloud platforms, other cultural institutes, especially local cultural centers, were also 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 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 the 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 towards data accessibility and transparency. Wei [15] constructed a multi-layer regression model on survey data to explore themechanism 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 covers a wider range of users of public cultural institutes. Data-driven research on public cultural services pose both challenges and opportunities for researchers and practitioners in the field of Library and Information Science (LIS), by providing profound insights of 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 which 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 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 BERTopic2 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 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 Fig. 1. With this framework, we are able to cluster regions according to the contents of their cultural activities and explain the clustering results by topic modeling.

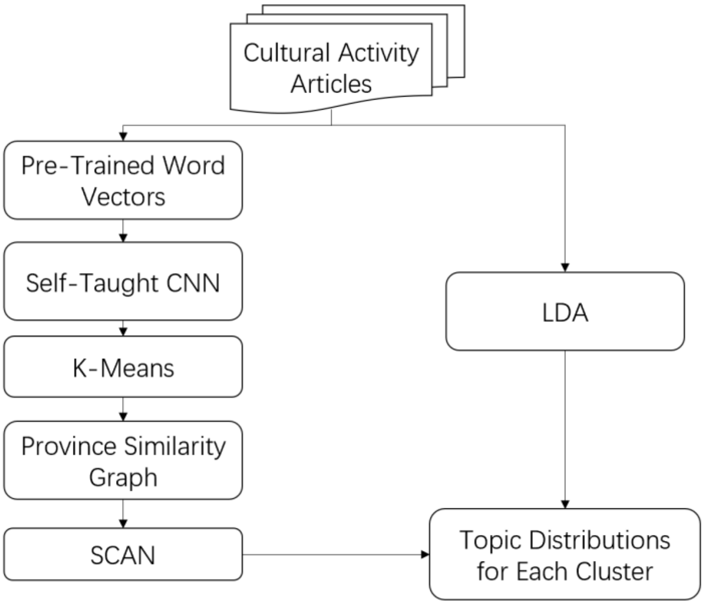
Cultural Activity Articles
Pre-Trained Word Vectors
Self-Taught CNN
K-Means
Province Similarity Graph
SCAN
LDA
Topic Distributions for Each Cluster

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, 18 were district-level centers) and 27 were managed by libraries (17 were

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

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00121

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, for 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 are 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, in the form of YYYY-MM-DD
Place	Detailed address of the cultural activity
URL	Link of cultural activity article
Remark	Description of the activity
Activity type (Optional)	Type of activity, such as exhibition, show and lecture
Organizer (Optional)	Institute or committee organizing the activity
Contact (Optional)	Phone number or e-mail address
Presenter Introduction (Optional)	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 year 2020, because many public cultural institutes just began posting articles in the past 2 to 3 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 have increased drastically over recent years, as shown in Fig. 2.

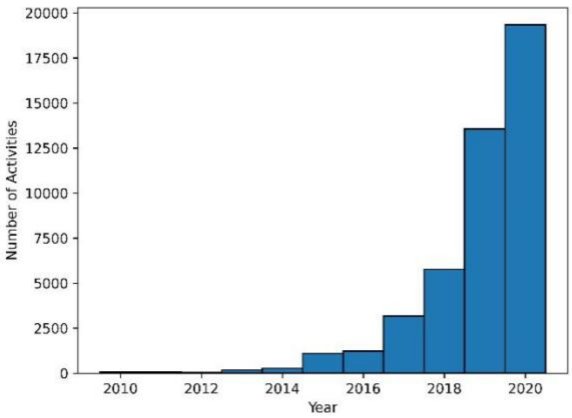

Figure 2. Articles on Cultural Activities each year

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 at 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:

- Overall, the public cultural institutes in that region were not enthusiastic about organizing cultural activities
- The public cultural institutes in that region were not accustomed to posting information online
- The region established their website recently, so few cultural activity articles have been accumulated
- 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 jieba package and remove stopwords.

Table 2. Number of Cultural Activity Articles in Each Region

	Region	**Number of Cultural Activities Articles in 2020**
1	Chongqing	3,589
2	Hunan	2,977
3	Guangdong	2,837
4	Jiangsu	1,453
5	Shandong	1,353
6	Shanghai	1,101
7	Beijing	840
8	Yunnan	653
9	Anhui	614
10	Tianjin	437
11	Zhejiang	351
12	Hubei	201
13	Shaanxi	184
14	Gansu	162
15	Sichuan	144
16	Fujian	115
17	Hainan	100
18	Jilin	88
19	Inner Mongolia	86
20	Shanxi	72
21	Liaoning	51
22	Jiangxi	40
23	Hebei	35
24	Xinjiang	9
25	Guangxi	8
26	Henan	1

27	Tibet	1
28	Qinghai	0
29	Ningxia	0
30	Heilongjiang	0
31	Guizhou	0
	Total	**17,402**

3.2 Neural Short Text Clustering

As can be observed from Fig. 3, 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)[3], 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 threshold, which is used to train CNN.The structure of our CNN model is shown in Fig 4. Our model used pretrained vectors developed by Li et al. [26] and dropout with 50% rate was employed for regularization. Afterwards, the classic K-means algorithm is applied to assign each article to a cluster.

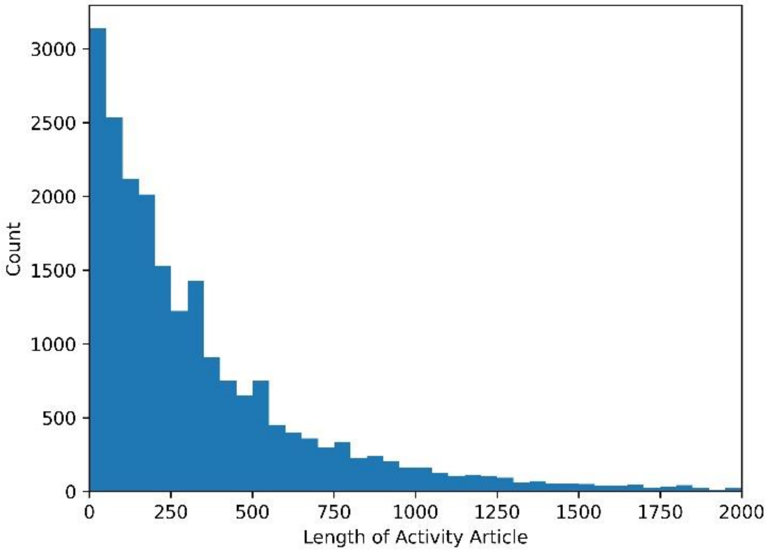

Figure 3. Length of Cultural Activity Articles

[3] This algorithm was chosen according to evaluation results in prior work.

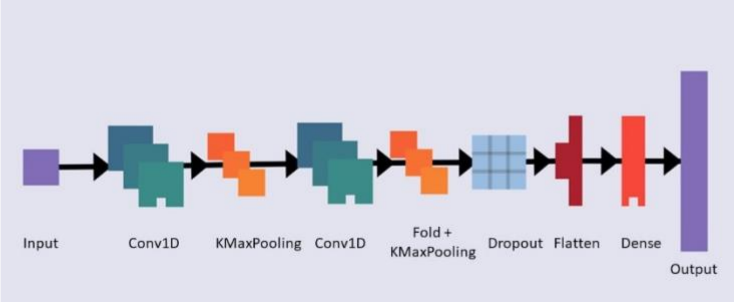

Figure 4. Self-Taught CNN model

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 (namely the Silhouette Score, Calinski-Harabasz Score, and Davies-Bouldin Score). The evaluation results are summarized in Table 3, which shows that the self- taught CNN produces significantly better clusters compared to baseline methods.

Table 3: Evaluation of Clustering Results

Clustering Method	**Silhouette Score**	**Calinski Harabasz Score**	**Davies Bouldin Score**
BoW + K-Means	0.012	103.782	7.049
AE + K-Means	0.077	305.344	5.398
Self-Taught CNN + K-means	0.557	97,756.102	0.731

[a]Number of clusters is 6.

3.3 Graph-Based Clustering

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

(1)

where denotes region *x*, and 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 .

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

(2)

where 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 reachableby any cluster).

3.4 Topic Modeling

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00121

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:

(3)

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

(4)

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

4.Results

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 Fig. 5. As can be seen from Fig. 5, over 2,500 cultural activities were organized in January 2020, which was around the time of 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 patternwas related to quarantine policies for COVID-19, for 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.

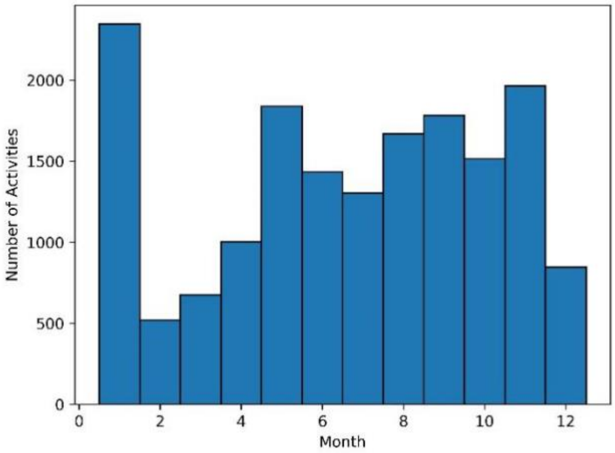

Figure 5. Number of Cultural Activities Organized Each Month

In Fig. 6, the total number of cultural activities in each region is plotted in the map. It is possible to divide regions into five categories according to the number of cultural activities, namely 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 5 cultural activities per day. Regions with frequent culturalactivities, like Beijing, Shandong, Jiangsu, Hunan, have approximately 2 cultural activities per day on average. Regions with moderate cultural activities organize 1 cultural activity per day. Finally, regions with sparse cultural activities only organize 1 cultural activity per week. Generally speaking, there are more cultural activities in Southern China and Eastern China.

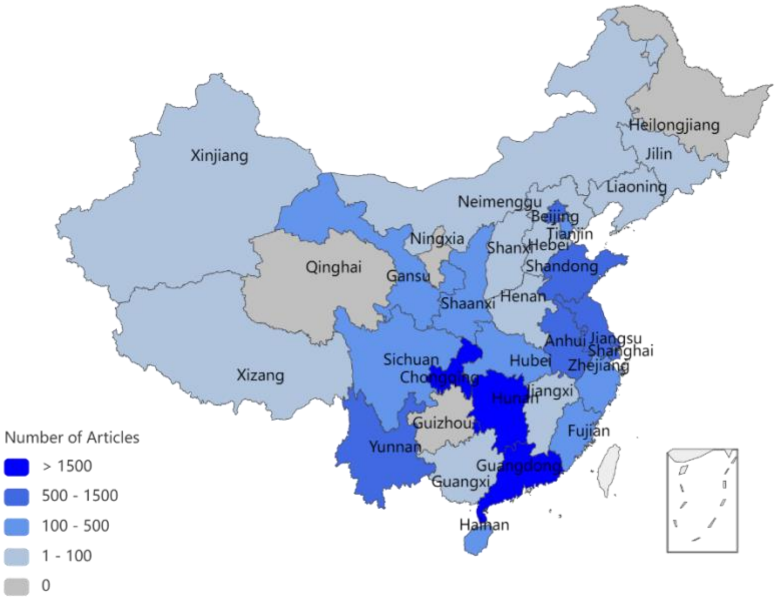

Figure 6. Number of Cultural Activities in Each Region

In Fig. 7, the density of cultural activities in each region is plotted in the map, which is defined as

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00121

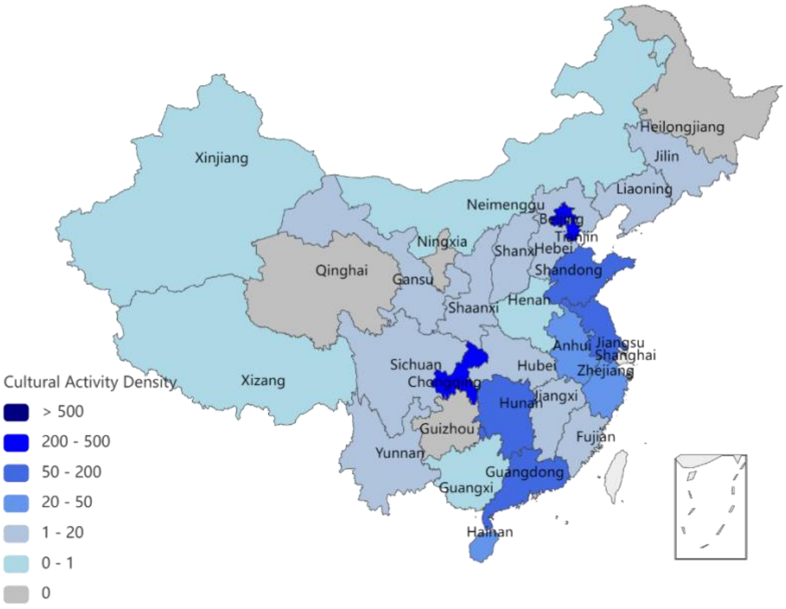

Figure 7. Density of Cultural Activities in Each Region

where A denotes a region, and the area of each region is measured in km^2. It is worth noting that regions with the highest cultural activity density are the 4 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 correlatedto overall economic development. In fact, the Spearman correlation coefficient r of disposal income per capita[4] 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 3 clusters (containing 10, 6, 4 regions respectively) and 1 outlier, summarized in Table 4. In Fig. 8, we visualized the graph weconstructed according to the Jaccard coefficient, positioning nodes with Kamada-Kawai layout. To further analyze if adjacent regions were more likely to belong to the same cluster, we also marked the clusters on a map, as shown in Fig. 9.

Table 4. Summary of Clustering Results

Clusters	Size	Members
Cluster 1	10	Liaoning, Hubei, Fujian Hainan, Zhejiang, Shanxi, Shaanxi Jilin, Inner Mongolia, Sichuan

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

Cluster 2	6	Beijing, Anhui Yunnan, Tianjin, Shandong, Shanghai
Cluster 3	4	Guangdong, Hunan, Jiangsu, Chongqing
Outlier	1	Gansu

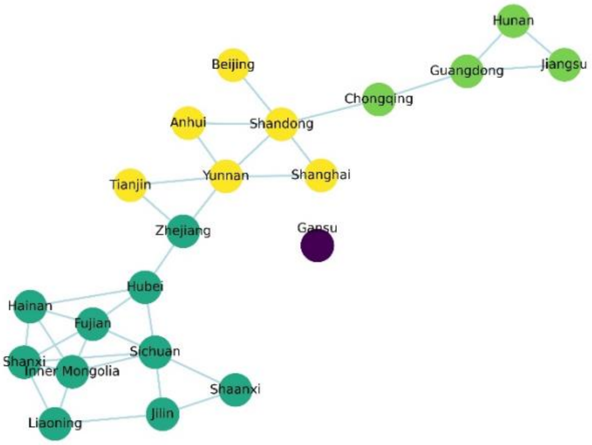

Figure 8. Clustering Results

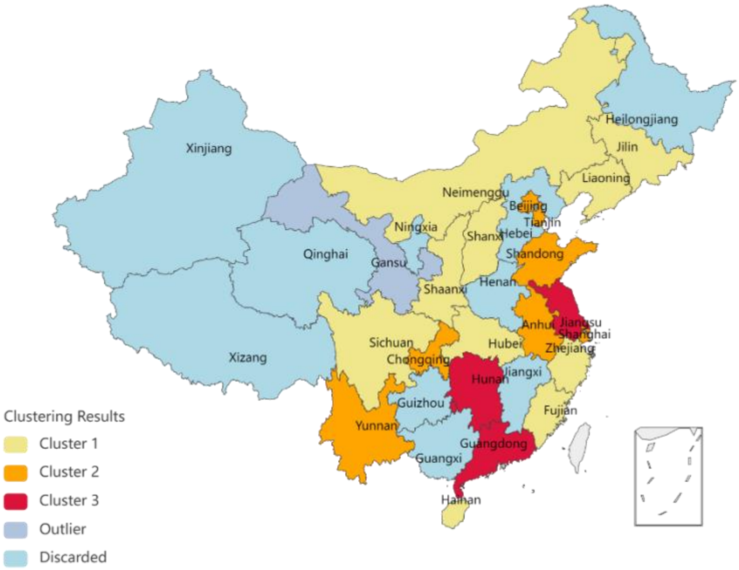

Figure 9. Clustering Results on Map

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00121

From Table 4, Fig. 8 and Fig. 9, 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 3 out of the 4 municipalities were assigned to Cluster 2, which may be due to fact that Chongqing has a much larger area than the other 3 municipalities. What's more, 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. 10. In this way, 8 topics were extracted from the cultural activity articles and the top 10 keywords for eachtopic, which is summarized in Table 5. We can see that the 8 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.

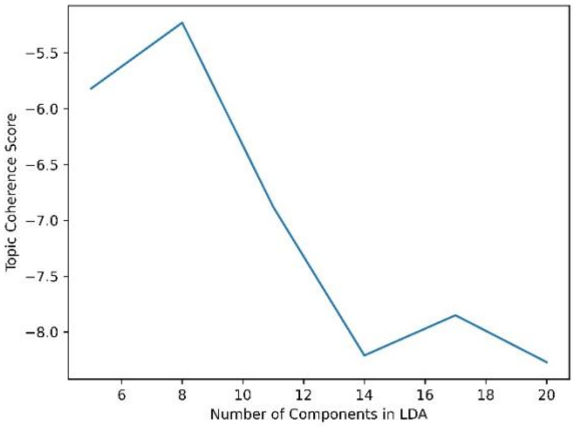

Figure 10. Topic Coherence Score

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
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

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. 11. It can be observed from Fig. 11 that for cluster 1, 2 and 3, the most frequent topic is topic 1, namely lectures on various forms of art and zeitgeist atlocal libraries. This shows that local libraries play an extremely important role in the organization of public cultural activities.

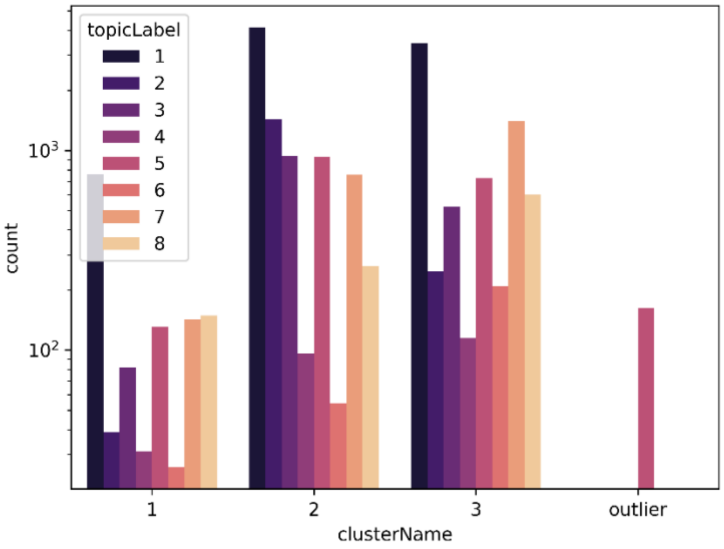

Figure 11. Topic Distribution of Each Cluster

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

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00121

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, its 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. For instance, as a provincial area with over 25 ethnic minorities, Yunnan has strived to preserve the arts of its residents. The Cultural Center of Yunnan has a program that introduces intangible cultural heritage to the public, including the knife dance of Yi, knitting techniques of Wa, and wall paintings of Dai. The Library of Yunnan also has a multimedia database that contains records of the 15 ethnic minority groups that are unique to Yunnan.

For regions in cluster 3, its public cultural institutes emphasize the importance of intangiblecultural heritage, followed by free art and music lessons at culture centers for public welfare. Public cultural institutes in these regions are enthusiastic about preserving cultural traditions. For example, the culture center of Guangdong hosts lectures on traditional Chinese medicine each month, which helps citizens gain a better understanding of traditional medical practices that have flourished in China for thousands of years.

Last but not least, Gansu's public cultural activities all shared the topic of free art and music lessons at culture centers for public welfare, and it may be the case that Gansu has more monotonous topics compared to other regions. As a matter of fact, Gansu is one of the economically underdeveloped regions in China, and it is not surprising that Gansu has lagged behind other regions in terms of cultural activities. At present, cultural institutes in Gansu are encouraging more citizens develop reading habits by hosting reading activities regularly.

5.Conclusion

In this study, we collect over 17,000 articles from 108 official websites of public libraries and culture centers across China. Analysis of the spatiotemporal features of our dataset reveal that there were fewer public cultural activities in spring, possibly due to the outbreak of COVID-19, and as the quarantine measures relaxed, more cultural activities were organized across all regions. The total number and density of public cultural activities are imbalanced across regions, with more cultural activities in Eastern and Southern China (especially in Yangtze River Delta and Pearl River Delta), as well as the highest cultural activity density in the four municipalities (Shanghai, Tianjin, Chongqing and Beijing), which shows that the number and cultural activities may be related to the level of economic development, informatization and the population density in that region. Also, regions with many ethnic minorities, like Yunnan, also have rich cultural activities.

We further uncover the topics of cultural activities with a two-step text clustering and topic modeling approach. A self-taught CNN is trained for embeddings of each article, on which the classic K-Means algorithm is applied to obtain cluster labels. Afterwards, we compute an undirected graph based on the Jaccard similarity coefficient of cluster labels of articles from each pair of regions. A graph-based clustering algorithm called SCAN is employed to derive clusters of regions. To explain the clustering result, we use LDA, a topic modeling algorithm, to derive various topics, which are characterized by the most important keywords in each topic. By plotting the topic distribution of each cluster, we are able to uncover unique tendencies of local cultural institutes when organizing cultural activities.

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

Our clustering and topic modeling results show that different regions vary as to their focus on public cultural activities. Some regions have strived to provide rich educational resources for its residents, while other regions focus on promoting public cultural services for special social groups (e.g. rural residents and ethnic minorities), and still other regions place great emphasis on preserving cultural traditions. Our study also reveals that cultural activities in Gansu lack diversity, which may have negative influence on participation.

6. Discussion

The crucial aim of providing public cultural service is to promote public welfare by satisfying the cultural needs of citizens, such as the need to receive education, the need to preserve traditional culture, and enjoying cultural works are a form of entertainment, etc. Essentially, providing better public cultural service is about understanding the needs of citizens and allocating resources of public cultural service institutes efficiently to provide these services. Many government officials and scholars have proposed different theories and roadmaps for improving the quality of public cultural service, including providing equal and accessible public cultural service for all society members [28, 29], extending providers of public cultural service [30, 31], emphasizing regional characteristics [32, 33]. Despite the relative abundance of theoretical frameworks on public cultural service, there have been few papers focusing analyzing public cultural service using empirical data, and prior work often rely heavily on qualitative methods such as field surveys and are limited in scope (i.e., they are often written in the form of case studies). Data-driven methods such as text mining techniques proves to be a promising way for understanding both the cultural needs of citizens and public cultural services currently provided by public cultural citizens.

In this paper, we focus on understanding the trends of public cultural activities (an important component of public cultural service). To this end, we propose a text clustering and topic modeling framework for providing fine-grained analysis on the characteristics of public cultural activities in China and assess trends of public cultural activities in 2020 based on a self-constructed dataset. Compared to traditional methods such as surveys or fieldwork, our approach provides satisfactory analysis results despite requiring significantly less manual labor. Our paper is also the first study to provide a comprehensive overview of public cultural service from a national perspective, which we hope can provide insights for the formulation of future policies. While public cultural service has been organized relatively independently by regional cultural institutes in the past, with the recent advent of the National Public Cloud Platform (https://www.culturedc.cn/) we observe the feasibility and necessity of assessing public cultural service from a more holistic perspective. We hope that our findings will help government officials gain actionable insights from current trends in public cultural service, and serve as the basis for the formulation of future cultural policies.

The public cultural activity dataset we constructed provides detailed and authentic information on the content and characteristics of public cultural services in various regions. Despite the extensive efforts we made in our data collection process, we observed imbalance in the amount of available data for different regions, and articles from several regions were eliminated due to data scarcity. With the informatization of public cultural institutes, we are optimistic that richer data on public culture will be available in the near future. We would also like to mention that textual data collected from public cultural institutes are highly unstructured, so further investigation on information extraction algorithms may be helpful for understanding various aspects of public cultural activities, such as organizers, presenters and the overall scale of activities.

In order to uncover the topics of public cultural activities in different regions, we used short text clustering (self-trained CNN), graph clustering (SCAN) and topic modeling (LDA) algorithms jointly. While we are confident that this framework is suitable for the purpose of this paper, recent advances in natural language processing technologies have also provided us with several alternative approaches. For example, it is possible to encode texts via Transformer models such as BERT, fuse both structural and semantic information via graph convolutional networks, and use more recent topic modeling algorithms such as BERTTopic2. These NLP algorithms can be further explored in future text mining research on public cultural service. The topic modeling algorithm (i.e., LDA) we used in this study is based on the bag of words model and conditional independence assumption, so semantic information was lost in the analyzing process. It may be more ideal if we could mine the topics of cultural activities from a more

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00121

integral level, perhaps with multi-text summarization techniques. What's more, it is quite difficult to evaluate the quality of clustering and topic modeling, even though we have used unsupervised metrics such as Silhouette Score and Topic Coherence Score to guide the choice of hyperparameters in our paper. Further examination by public cultural experts may help validate our findings.

At the end of this paper, we would like to highlight some directions for future research on public cultural service. First of all, our paper only focuses on leveraging text mining techniques to understand public cultural activities. In the future, it is possible to collect empirical data to study the cultural needs of citizens, as well as other aspects of public cultural service, which will help government officials and public cultural institutes gain a better understanding of trends and challenges of public cultural services in China from a data-driven perspective. Besides, summarizing and presenting the findings of data mining research is also an important research direction. For example, our text clustering and topic modeling results can be integrated in a visualization system for public cultural service. Such a visualization system may incorporate various aspects of public cultural service, such as the geographic locations of public cultural institutes, the number of citizens actively participating in cultural activities, topic distributions of cultural activities and so on to help citizens and officials gain a better understanding of public cultural service. Last but not least, we believe it is meaningful to study the temporal evolution of trends in public cultural service over a longer period of time, which will become feasible as more data accumulates on the official websites of public cultural institutes. By understanding how public cultural service changes over time, we will be able to gain a better understanding of the consistency and future directions for providing public cultural service.

Author Contributions

Z. Zeng (zixinzeng_jennifer@pku.edu.cn) was responsible for data collection, code implementation, as well as writing this paper. B. Hua (huabolin@pku.edu.cn) put forward the research topic and revised this paper.

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.

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., Wang, P., Zheng, S., Tian, G., Zhao, J., Xu, B.: Self-Taught Convolutional Neural Networks for short text clustering, Neural Networks 88, 22-31 (2017)

[7] Xu, N., Yuruk, N., Feng, Z., Schweiger, T. A. G.: 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. Melbourne, Australia: University of Melbourne (2015)

[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] Kamupunga, 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, Cape Town, South Africa, Aug. 15–21 (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., Deng, S., Kong, J., Yan, X.: 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., Sterckx, L., Demeester, T., Develder, C.: 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., Dai, X., Zhang, J., Chen, J.: Topic2Vec: Learning Distributed Representations of Topics. In: 2015 International Conference on Asian Language Processing (IALP), pp. 193-196 (2015)

[24] Mikolov, T., Chen, K., Corrado, G., Dean, J.: 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(IJACSA) 6(1), 147-153 (2015)

[26] Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., Du, X.: 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., Wallach, H., Talley, E., Leenders, M., McCallum, A.: 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+2 (2021)

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00121

Author Biography

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

Bolin Hua received his PhD degree of Information Resource Management in Nanjing University. He is currently an associate professor at the Department of Information Management in Peking University, and has published over 60 papers on the topics of text mining, intelligence analysis based on big data, and big data of public culture service.

ORCID:0000-0001-9248-6455