DSI DATA SCIENCE AND INFORMETRICS
VOL.1, NO.1, FEB. 2021

Visual Analytics of Large-scale E-government Text Data via Simplified Word Cloud

Yanan Liu[a], Fang He[a], Jin Wen[a], Zhiguang Zhou[a, c], Jinchang Li [b*]

a. School of Information, Zhejiang University of Finance and Economics, Hangzhou, China
b. School of Data Sciences, Zhejiang University of Finance and Economics, Hangzhou, China
c. State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou, China

ABSTRACT

With the rapid development of Internet technology, a rich set of e-government data are collected by the government departments. For example, a variety of feedback text data can be obtained quickly and efficiently through various channels such as the mayor's mailbox. It is an effective way to improve the working efficiency of the government to extract hot topics from large-scale e-government text data, establish the correlation between topics and geographic space, and interactively explore the sources of public feedback problems. However, it is a difficult task to explore the large-scale e-government text data with traditional visualization methods such as word cloud, because too many words are hardly distributed in a limited space which will largely disturb the visual perception. In this paper, we propose a visual analytics system for large-scale e-government data exploration by means of simplified word cloud. Firstly, a representation learning model is used to embed the text data into high-dimensional space to quantitatively represent the semantic structure features of e-government text data. Then, the high-dimensional vectors are projected into a two-dimensional space where the coordinate distribution of points effectively expresses the semantic similarity of original words, which also presents geographic features that can be quantized by means of a similarity computing model. In order to simplify the understanding of large-scale e-government data and improve the cognitive efficiency of word could, we adopt the adaptive blue noise method to sample the topic words, which can simplify the visual expression of word cloud and improve the understanding efficiency of e-government data without losing the semantic structure features. Furthermore, an abstraction and visual analysis system for large-scale e-government text data is designed and implemented by integrating the above representation learning model, sampling-based abstraction model of word cloud, and topic and geographic correlation analysis model. This system provides convenient human-computer interaction modes and supports users to explore the analysis and extraction of the characteristics hidden in large-scale e-government data. It also helps government departments quickly locate the hot topics of public concern and their related regional distribution, and provides decision support to further improve the work efficiency of the government. Case studies based on real-world datasets further verify the effectiveness and practicability of our system.

KEYWORDS

E-government; Text mining; Text visualization; Visual analytics

* Corresponding Author

1 Introduction

E-government refers to the use of internet technology as a platform for exchanging information, providing services and transacting with citizens, businesses, and other arms of government (Scholta et al., 2019) . With the development and maturity of e-government mode, government departments pay more attention to public participation in e-government and provide efficient communication channels for the public to express their wishes(Shareef et al., 2012). Thus, a rich set of e-government data are produced, which represent the will of the public and are often collected in the form of text. Traditional e-government data analysis methods often extract key information from the text data in a manual way, and then calculate and predict the public's satisfaction with the government departments (Metaxas et al., 2017; Song & Meier, 2018) . However, the process of traditional exploration methods is always cumbersome and complex, which usually requires repeated loading of summary and statistical analysis, resulting in strong uncertainty of the results. Moreover, with the accumulation of data volume and the increasingly complex data structure, the limitations of traditional processing and analysis methods are more prominent.

In the field of visualization, we often utilize bag-of-word models to mine the text topics, and use word cloud to display the topic information. With the increase of data scale, e-government data mining and visualization face the following challenges: C1. Text topics are difficult to mine. E-government data is mostly in the term of short text, and the topic-mining model based on bag-of-word is difficult to accurately extract the semantic information. C2. Text data scale is large, and word cloud visualization method is not effective. Large-scale text data are presented in a limited word cloud space, with serious crowding and overlap, resulting in visual redundancy and difficulty in accurate topic exploration. C3. The geographical distribution is hard to be discovered intuitively in the word cloud. It is difficult to find the geospatial distribution of the topic in the word cloud because of the separation between them.

To tackle the above challenges, we design a visual analytics framework based on the simplified word cloud to explore the large-scale e-government data. Firstly, a representation-learning model Word2vec is employed to extract topic features from e-government short text data. The extracted features have realistic topic meaning and are helpful to understand people's resource preferences. Then, an adaptive blue noise sampling model is conducted in the word embedding space to extract keywords that can effectively express the semantics of original data, which are further utilized to generate simplified word clouds with semantic features preserved. Furthermore, the semantic similarity is calculated to establish the relation between extracted topics and geographical space, and help users to explore the spatial distribution of topics of interest. Finally, a rich set of convenient interaction modes are integrated into the visualization system, enabling users to explore e-government related topics and their spatiotemporal relationships. Case studies based on real-world datasets further verify the effectiveness and practicability of our system, which can provide decision-making basis for the work evaluation and follow-up reform and innovation of the relevant government departments. The main contributions of this paper are as follows:

(1) We design a semantic region-partitioning algorithm to recognize semantic topics in the vectorized space obtained through representation learning, by means of which more correct semantics will be extracted based on the essential characteristics of natural language.

(2) We propose a simplified word cloud generation method based on blue noise sampling

to present the semantic topics of the original e-government data, by means of which the overdrawn problem of words is tackled while the semantic topics are all preserved.

(3) The association between topics and geographic space is constructed in virtue of semantic similarity, which is calculated with semantic distance of words and visualized on the map. It really supports the visual analysis of spatiotemporal changes of semantic topics.

(4) A web-based visual analysis framework is implemented to integrate above models and visual designs, by means of which users can explore the topics and geographic correlation features of e-government data, and select the topics or regions of interest for specific analysis.

The organization of this paper is structured as follows. Section 2 discusses the related work of e-government data analysis. Section 3 introduces analysis tasks and workflow of the system. Section 4 describes the innovation and realization of the algorithms in detail. The visual analysis system and the intention of visual design are described in Section 5. Section 6 evaluates the effectiveness of our system with case studies and expert interviews, and discusses the shortcomings of the system. The last section summarizes the paper and looks forward to the future work.

2 Related Work

In this section, we review the related work, including e-government data analysis, text mining and visualization, and spatiotemporal data visualization

2.1 E–government data analysis

With the development of information technology, more people participate in the evaluation of government work through e-government platforms. They express their views and suggestions on the work of government departments, or consult their concerns to government departments, etc., forming e-government data (Linders, 2012) . These data provide good conditions for government departments to understand the hot issues of public concern. In recent years, it is through the e-government platform that government departments make public opinions play an increasingly important role in government performance evaluation (Bai, 2013), public decision-making support (Nabatchi et al., 2015) , etc. The work of government departments is more inclined to reflect the public value, thus reducing the phenomenon of government failure (Huang, 2004) and improving the governance ability of the government. However, the basis for the public opinion to play its real value in government governance is that government departments can correctly perceive and accept the opinions. Therefore, how to accurately mine the key characteristics of e-government data, and correctly perceive the main content of public opinion expression is extremely important.

Many scholars have conducted research on e-government data. For example, Stylios et al. (2010) use sentiment analysis method to extract public opinions automatically and emotions in online posts, to facilitate the government departments in the future work reference. Mayasari et al. carry out sentiment analysis on tweets based on machine learning method, and study the variation rule of public sentiment on government performance evaluation in Surabaya, Malaysia (Mayasari et al., 2020) . Baojun et al. (2013), aiming at the content analysis of public opinions in the context of smart cities, propose a methodological framework based on LDA topic model to extract potential topics that the government or policy makers may pay attention to and analyze the time series of discussion heat from large-scale opinion information text. Yi et al. (2019) adopt LDA model to mine e-government data for gover-

nance of bike-sharing policies, hoping to provide theoretical basis and decision-making suggestions for the government to make policies more scientifically. Yimin (2018) believes that whether urban planning and construction are well done or not is ultimately measured by the satisfaction of the masses. The public opinions of the draft of urban master plan can reflect the citizens' satisfaction with various areas of urban development in a specific period. They utilize text-mining technology to analyze the e-government data of Beijing urban planning. Zhengrong (2019) takes advantage of big data analysis technology to excavate the topic features of public concern, so as to facilitate government departments to understand the key information of public concern and better respond to public demand.

It can be seen from the above literature that some studies have focused on the mining and analysis of e-government data. However, there are still three deficiencies in this field. First, the number of literatures in this research field of e-government data analysis is still small, so the study of this paper has a certain contribution nonetheless. Second, related studies have not used professional visualization technology to make visual analysis of domain data, which makes these studies fail to directly reveal the hidden features of public opinion data. Third, the relevant research did not simplify large-scale data, did not relate the topic with the geographic space, and failed to adopt the depth mining and visual analysis of data features. In view of the above three shortcomings, this paper conducts interactive visual analysis of e-government data based on text mining and abstraction, which is of great significance to explore the semantic characteristics of topics and the correlation characteristics with geographic space.

2.2 Text mining and visualization

With the popularity of social network, the scale of social network text data is getting larger and larger. How to find valuable information quickly and accurately from these massive data has become a major challenge in the field of information science and technology. Text mining is a text processing technology that extracts meaningful information from unstructured data and discovers the potential value of large-scale text information (He et al., 2013).

The essence of text is natural language. One way is to construct semantic embedding space from the context of language to carry out text mining. Many previous works are based on representation learning to mine language features. For example, Hotho et al. (2003) use WordNet to convert word vectors into concept vectors, and measure the affinity between documents by calculating the similarity between concept vectors. Kim et al. (2015) present a hierarchical similarity measurement method based on search fragments on short texts to calculate the similarity between short texts. Other scholars focus on the exploration and analysis of linguistic models in textual data.BengioHolger et al. (2006) propose to learn the distributed representation of words and the probability function of word sequences simultaneously to counter the curse of dimensionality. Dauphin et al. (2017) develop a finite context recognition method through stacked convolution, which allows parallelization on sequential tags and can improve the processing efficiency of text data. Ghanbarpour and Naderi (2020) propose a language model based attribute specific ranking method, which sort candidate answers according to their semantic information until they reach the corresponding attribute level. Collins et al. (2009) produce the concept of multi-model semantic interaction, in which semantic interaction can be used to guide multiple models at multiple data scale levels to enable users to solve larger data problems.Angus et al. (2012) introduce Conceptual Recursive Graph to process text, which is a tool for drawing recursive graphs based on similarity of

concepts rather than terms. They build a part of speech model and apply the algorithm to measure the similarity between two sessions.

Clustering analysis of text data is another method of text feature mining.Beil et al. (2002) propose two text-clustering algorithms: FTC plane clustering based on frequent sets and HFTC hierarchical clustering.Yin and Wang (2014) present a folding Gibbs sampling algorithm based on Dirichlet Polynomial Mixed Model Short Text Clustering (GSDMM) to solve the problems caused by short text's sparseness, high dimension and large volume. The method achieves a good balance between the completeness and uniformity of clustering results. In order to understand the attention of bioinformatics community to different sub-fields, Janssens et al. (2007) deeply merge the text content with the structure of citation graph, and improve the unsupervised clustering performance of text based on Fisher reverse chi-square hybrid clustering method.

Text visualization refers to the process of transforming abstract data into visual graphics. By extracting, transforming and mapping eigenvalues of data, the data is finally displayed in the form of images, which is the basic technology of data visual analysis in this paper (Card et al., 1999). As the saying goes, a picture is worth a thousand words, and more than 80% of the information obtained by human beings from the outside world comes from the visual system (Lei et al., 2014). The presentation of text data in a visual and intuitive form is conducive to the analysis of the hidden information and knowledge behind the data. When confronted with massive texts, people need to browse the main contents of each text or the whole text set quickly, so it is necessary to display the text visually.

Word cloud is a commonly used text visualization method, which maps the size of words in two-dimensional space by taking the frequency of occurrence of words as the correlation measure. For instance, Wordle creates a presentation similar to the word cloud and uses a heuristic method to optimize the use efficiency of the visual area (Viegas et al., 2009) Seifert et al. (2008) introduce an algorithm for rendering compact visualization, which takes any convex polygon as the boundary to obtain higher space utilization. Wang et al. (2018) design a consistency preserving word cloud generation method, namely Edwordle, which allows users to move and edit words while preserving the neighborhood of their words. Paulovich et al. (2008) propose a kind of least square projection (LSP) to represent documents by arranging graphic marks in the visual space, and the distance of documents in the projection space reflect the content similarity. Andrews et al. (2002) propose a method of hierarchical organization of document sets to optimize the design of Voronoi diagrams and use boundary polygons to visualize document sets of specific levels in the hierarchy.

2.3 Spatiotemporal data visualization

The spatiotemporal attribute is an important feature of text data, which refers to the time attribute and the geographic attribute. Time attribute refers to the generation time of index data, while geographic attribute refers to the specific place where behaviors and events occur or belong. Visualization of data with spatiotemporal attributes is conducive to exploring data characteristics under different spatiotemporal conditions, to assist decision-making and management.

In the field of data visualization, many scholars have discussed how to conduct efficient visual analysis of spatiotemporal data. Wang et al. (2014) explore the characteristics of vehicle operation data at traffic checkpoints in Nanjing. They use dots on the map to describe the geographic location of traffic checkpoints, and design attributes such as color, number of ar-

rows and direction to represent the speed, direction and volume of traffic flow at different checkpoints. Users can intuitively analyze and discover important traffic hubs and traffic flow information in Nanjing. Pu et al. (2013) propose a visual analysis system T-Watcher. According to GPS data, the map is divided into raster points, which are clustered to form a regional view, and the color brightness represents the traffic flow. The dense area of taxi passengers can effectively present the distribution of hot spots in the city, thus helping the traffic department to monitor and analyze the complex traffic situation in big cities. Wu et al. (2016) design a visual analysis system, TELCOVIS, which can effectively analyze urban crowd movement behavior for recording the telecommunication data exchanged between mobile phones and base stations in Guangzhou. This system focuses on the behavioral characteristics of co-occurrence, and studies its feature extraction and association analysis. In addition, visual effects such as contour tree diagram and parallel coordinate diagram of geographic view are designed to help users quickly identify the common behavioral characteristics of the crowd, and provide assistance and support for relevant departments to study and analyze urban crowd activities and various kinds of derived social problems. Cao et al. (2012) make use of Twitter data to develop a visual analysis system Whisper, which is able to analyze social network public opinion effectively in real time by combining geographic information.

3 Task Analysis and System Overview

3.1 Data introduction

The data to verify the visual analysis system comes from the "Topic-Overview" section of a city's network political platform, which published 28,357 e-government opinion data for the city's three districts and four county-level governments between 2014 and 2020. The text data contains geographic attributes, which is suitable for the research objective of analyzing the correlation characteristics between public concerns and geographic topics in this paper. With this experimental data set, it provides a new perspective for the research on opinion data of government work, and a convenient interactive visual analysis way for government departments to understand the correlations between public concern topics and geographical space.

3.2 Requirements analysis

Through close communication and in-depth exploration with domain experts, we have a detailed understanding of the practical problems and interested directions of domain experts in e-government data analysis, and finally summarize four visual analysis tasks.

T1. Visualization of semantic structure representation

The government is very concerned about the topics that people in different geographical regions are most interested and how the level of concern varies. Data with geographical attributes integrate the geospatial features of the topics, which lead to the distribution of the topics covered by the data in different semantic regions. It is difficult for the classical topic mining methods to obtain accurate semantics when extracting topics. Therefore, how to characterize the text, construct semantic regions, and describe semantic correlations so that the extracted topics contain correct semantics is very important.

T2. Visualization of topic semantic features

Displaying all the keywords in the layout space visually will cause serious overlap and occlusion, which is not conducive to intuitive analysis of the meaning of the topics for users. How

to sample representative words from a large number of keywords to minimize the loss of semantic information and present the topic semantics clearly in a visual way is critical.

T3. Visualization of geographical distribution of topics

Topics that people are concerned not only share common characteristics, but also are affected by the particularity of geographical regions, which leads to geographical differences in topics. It is worth studying that how to establish the correlation between the topic and geographical space, and visually express the distribution of topics in geographic space, to show the hot issues that the government departments in different regions need to focus on intuitively.

T4. Visual analysis system for geographical distribution characteristics of topics

How to integrate text mining algorithm, visualization and interaction to design a system used by government departments for interactive visual analysis of topic and geographic correlation features of e-government data is of great significance. It can provide convenient data analysis tools for government departments, realize the one-stop transformation and analysis of data into visual interface, interactively choose interested regions or topics according to their own interests, and detect the changes in the topics of public concern in different regions.

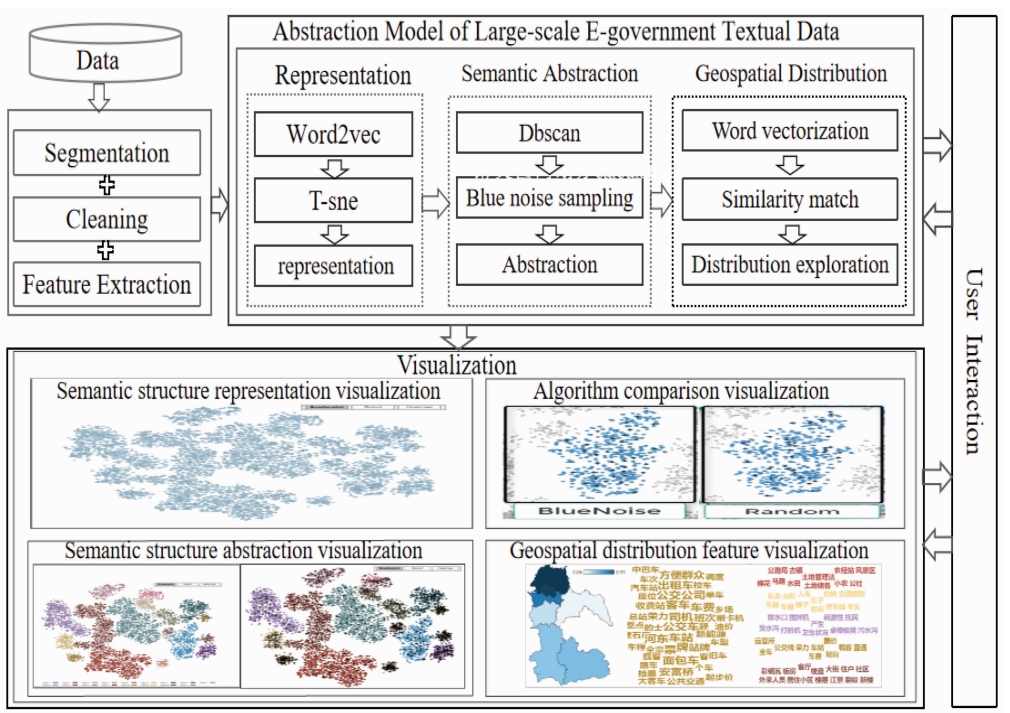


Figure 1 Flow chart of visual analysis system for large-scale e-government text data

3.3 System overview

The system flow chart of this paper is shown in Figure 1. Firstly, the data is segmented and cleaned, and the geographical features of the data are extracted. Secondly, Word2vec, a classical model of word representation learning, is used to construct the semantic space of the text data. t-SNE is employed to project the high-dimensional semantic vector into a two-dimensional plane, so as to facilitate the exploration of the semantic structure features of the

data, and prepare for mining topics by dividing semantic regions according to semantic structure. Then the blue noise sampling technique is adopted to extract the key semantic features of each topic in order to determine which topics the public discusses. After that, based on the results of representation learning, the semantic similarities between topics and geographic opinions are matched by the semantic similarity calculation, to establish their correlations. At last, by means of visualization technology, word representation learning, blue noise sampling and semantic similarity calculation methods along with convenient interaction are integrated into a visual analysis system effectively. It helps the government to explore the topics of public concern for government work and the correlations between topic and geographic space deeply.

4 E–government Data Visualization

4.1 Visualization of semantic structure representation

Large-scale e-government data targeted at different regional governments integrate the working characteristics of different functional departments and the special concerns of people in many geospatial regions, making the data characteristics complex and diverse. Compared with probability-based topic mining, it is more advantageous to extract topics from the perspective of semantic structure. In this paper, the text data is represented as word vectors, embedded into high dimensional semantic space, and the semantic similarity of words is judged from the spatial distance, to obtain the hot topic information of e-government data. The specific process is as follows:

(1) Word embedding

Word2vec is a natural language processing method that converts words into vectors through representation learning to express semantic information of text. By embedding words into a semantic space where semantically similar words are close(Shen et al., 2014), it is easy to judge the similarity between words according to geometric distance. Compared with the classical methods of learning text representation through bag-of-word model, Word2vec model takes full account of contextual semantic information and produces a higher learning quality(Tang et al., 2015). Therefore, in this paper we use Word2vec to represent text data and construct semantic space to reveal semantic structure of words. Each word is defined as a vector composed of word and word ID, and the corpus is generated by a series of words, as shown in Equation 1.

$$D=(w_1, w_2, w_3, ..., w_N) \quad (1)$$

where N is the count of words and D is the corpus generated by Word2vec training. In the process of word representation learning, the setting of text window size has an important influence on the result. Formula 2 is used in this paper to optimize the learning result of the model.

$$\frac{1}{T}\sum_{i=k}^{T-k} \log p(w_t | w_{t-k}, ..., w_{t+k}) \quad (2)$$

where T is the size of text window, and the corresponding context words of a given word is represented by $p(w_t | w_{t-k}..., w_{t+k})$.

(2) Dimensionality reduction

Words are converted to high-dimensional vectors by means of Word2vec model. However, it is difficult to visually explore the semantic structure of data and calculate semantic distance in a high-dimensional vector space with hundreds of dimensions. Therefore, to solve

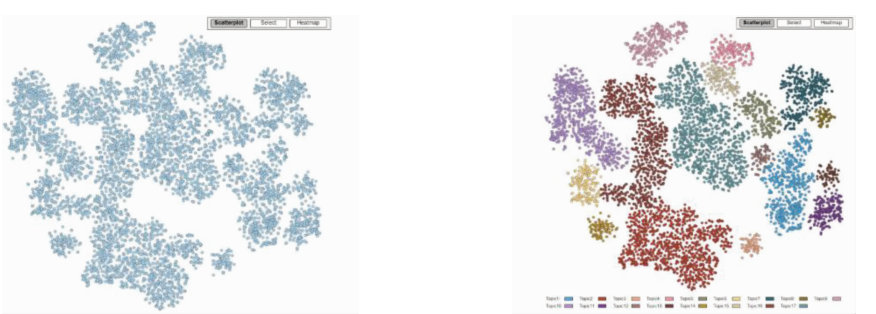
the problem of visual occlusion and computation difficulty of high-dimensional vectors, it is necessary to project high-dimensional vectors into two-dimensional space. t-SNE (t-distributed Stochastic Neighbor Embedding) (van der Maaten & Hinton, 2008) is an effective way for dimensionality reduction of high-dimensional vectors. It is able to capture both local and global features of data, and effectively retain the original features of data during dimension compression (Wattenberg et al., 2016). Due to the good performance in dimension reduction, t-SNE is used to project the representation learning results of Word2vec, to better reveal the semantic similarity of Word2vec in two-dimensional semantic space through the connection and closeness of words (Xia et al., 2018; Zhao et al., 2019).

(3) Semantic region division of topics

In order to further extract semantic features from the semantic space and obtain semantic categories, so as to extract topics of government work opinion data, in this paper we use DBSCAN (van der Maaten & Hinton, 2008) density clustering algorithm to classify words in the semantic space into meaningful categories according to their densities. The core of DBSCAN clustering is the density of clustering objects, which defines the cluster as the maximum set of density connection points. It can divide regions with high enough density into clusters and treat data points with low-density values as outliers. Since it is impossible to predict the number of topics concerned by the public, it is obviously not feasible to control the learning process by applying supervision conditions, such as setting independent variables and the number of target clusters. Compared with supervised clustering methods such as K-means (Hartigan & Wong, 1979) and GMM (Ebeida et al., 2014), DBSCAN has good performance under unsupervised conditions, so we choose DBSCAN clustering method to mine topics.

(4) Structure representation

In order to capture the representation results more conveniently and verify the effectiveness of the proposed method, we further design the representation projection view to explore the representation results quickly. Figure 2(a) shows the results of public opinion representation. Each data point in the figure represents a word, and the closer the distance between data points is, the more semantically similar the words are. It can be seen that the compactness of semantic structure is different among different semantic sections. Figure 2(b) displays the topics that people are concerned, which are represented by different colors.



(a) Word2vec semantic structure representation (b) Topic of DBSCAN for clustering

Figure 2 Semantic structure representation and clustering

4.2 Visualization of topic features

(1) Blue noise sampling

Based on large-scale corpus, if all the words in a topic are arranged in the word cloud,

they will overlap and block each other in the limited screen space, resulting in a lot of visual confusion. Thus, it is impossible to perceive the semantics of the topic of interest clearly, which affects the effective analysis of data. Therefore, sampling representative words from large-scale data is an effective method to solve the serious occlusion of large-scale data layout in limited space. In order to simplify the words in original topic and maintain the semantic features, so that the sampled words can represent the original semantics of the topic to the maximum extent, we use the blue noise sampling algorithm based on Poisson disc to perform adaptive sampling for each topic. Blue noise sampling is a commonly used sampling algorithm in the field of graphics, which simultaneously satisfies the randomness and uniformity of the distribution of sampling point sets and can maximize the retention of the original semantics of data. It has a wide application in point cloud sampling, texture rendering, geometric processing and other aspects (Yan et al., 2015). In the sampling process, an active point is selected randomly as the center by throwing dart, and the radius is set by the sampling rate to generate the sampling disk. The generated Poisson disk must meet the minimum distance characteristic, that is, only if the distance between the centers of any two Poisson disks is greater than the sampling radius, the generated sampling point is valid. If the generated disk is inconsistent with the previous one, it will be rejected (Godwin et al., 2017). To maintain the semantic structure of the original data, we use kernel density estimation to evaluate the semantic structure density of the samples, with the formula as follows:

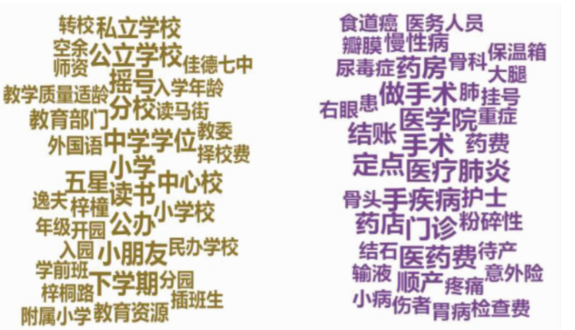
$$f(\mathrm{p}) = \sum_{i}^{n} k_h (p - p_j) \qquad (3)$$

where P = $\{p_1, p_2, \ldots, p_m\}$ is the coordinate position of a sequence of words, kh represents the Gaussian kernel function, h represents the bandwidth used to control the smoothness of the constructed density domain, and n is the total number of points in the local region. R= r_a/f (p) is used to obtain the sampling radius, where ra is the sampling rate, which is the number of subject words that the analyst needs to display.

In addition, the comparative experiment between random sampling and blue noise sampling algorithm is added to further prove the superiority of blue noise sampling in the sampling of topic words. The basic principle of random sampling is to select one sampling data point randomly at a time, and the algorithm will automatically traverse all the data until the total number of sampling data points meets the preset sampling rate requirements. Random sampling is also one of the commonly used sampling methods in data abstraction.

(2) Word cloud

As mentioned above, topics with different semantics are obtained based on representation learning, and then expressed by sampled words through blue noise sampling. In order to display the semantic features of each topic visually, we use word cloud to show the topic keywords that people are concerned. Word cloud is a very convenient and effective text visualization technology. Words displayed in the word cloud are the topic words obtained by blue noise sampling. The size of the word represents the occurrence probability of the word in the document. The greater the occurrence probability, the larger the size of the word in the word cloud. Figure 3 shows the word cloud view of topic 8 and topic 11, in which we can directly see that topic 8 focuses on education, while topic 11 focuses on medical care. Both of them have clear semantics.



(a)Topic 8 (b)Topic 11

Figure 3 Word cloud of topic

Moreover, we also design a multi-topic semantically preserved word cloud to display the top five hot topic words of each region simultaneously. Firstly, the layout space of the word cloud is divided, and the same topic words are placed in the same area. Secondly, according to the attribution of words to a topic, the keywords of different topics are rendered with the corresponding topic colors in the representation space.

4.3 Visualization of geographical distribution features of topics

(1) Exploration of geographical distribution features of topics

In order to further explore the distribution characteristics of topics in geographical space, we introduce the concept of semantic similarity calculation to analyze the correlation between topics and geographical space. Semantic similarity calculation (Palangi et al., 2014) refers to a method to calculate semantic association of text by calculating semantic distance between two texts to carry out similarity matching. Semantic similarity computing has been widely used in intelligent search and matching, machine translation, etc. Drawing on the concept of semantic similarity computing, we match the topic with the semantics of public opinion in different geographical regions to obtain their associations. The vector coordinates of the words in the topic in the two-dimensional semantic space are defined as:

$$L_i = ((x_1, y_1)(x_2, y_2)\dots(x_n, y_n)) \tag{4}$$

where n is the number of words in the topic corpus, (x_i, y_i) is the vector coordinate of word i in semantic space, and L_i represents the coordinate set of word vector. The vector of words in the semantic space of the geographic regional corpus is defined as:

$$L_j = ((x_1, y_1)(x_2, y_2)\dots(x_m, y_m)) \tag{5}$$

where m is the number of words in the geographic corpus, (x_j, y_j) is the vector coordinate of word j in semantic space, and L_j represents the coordinate set of word vector. Considering the difference of word vectors in semantic direction, we use cosine distance to match the text, and the formula is shown in 6:

$$D = \frac{\vec{L_i} \cdot \vec{L_j}}{||L_i|| \cdot ||L_j||} = \frac{\sum_i^n (x_i, y_i) \cdot \sum_j^m (x_i, y_i)}{\sqrt{\sum_i^n (x_i^2 + y_i^2)} \cdot \sqrt{\sum_j^m (x_j^2 + y_j^2)}} \tag{6}$$

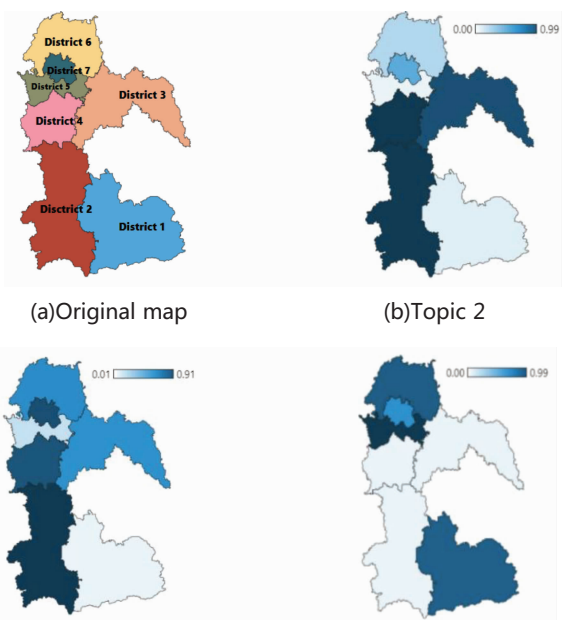
In our system, the semantic similarity between topic corpus and geographic corpus is used to represent the association between topic and geographic information. The more similar the semantics is, the greater the degree of correlation is, indicating that the attention of that

topic in the geographic area is higher.

(2) Visualization of geographic distribution features of topics

This section designs the visualization scheme of the geographic distribution features of topics, including the visualization of the geographic distribution features of a certain topic and the visualization of the hot topics of a certain region. Firstly, a visual display of the distribution of attention heat in different regions is designed for a certain topic. We use map view to facilitate government departments to analyze the distribution characteristics of the topic in different districts and counties. Figure 5(a) is a map of the administrative divisions of a city, including 7 districts. Different colors are used to fill different districts according to how much attention each county pays to the topic. The darker the color is, the higher the attention of the topic in a certain district. In Figure 5, (b) shows topic 2, which deals with the management and planning of land and resources in rural construction; (c) displays topic 3, which is about the renovation and demolition of dilapidated houses; (d) is topic 15, which is related to construction, contract signing and tax payment. From the distribution differences of different topics on the map, it can be seen that topic 2 has a higher attention heat in districts 2, 3 and 4; topic 3 has the highest attention heat in districts 2, 4 and7; and topic 15 has a higher attention heat in districts 1, 5, and 6.

Secondly, in order to support the government departments to further select specific regions of interest and analyze the hot topics that people are concerned about in a certain region, we further design the visualization of hot topics in a region. By clicking on a region in Figure 4(e), the corresponding top 5 hot topics in that region will be highlighted in Figure 4 (b), and the topic word cloud in Figure 4(c) will be replaced with the corresponding topic word cloud for a region of concern.



(a)Original map (b)Topic 2

(c)Topic 3 (d)Topic 15

Figure 4 Geographic distribution features of topics

4.4 Visual analysis system

In order to facilitate government departments to understand the differences in topics of

public concern in different regions and adjust the direction of work more accurately, a visual analysis system of topic and geographic correlation characteristics is designed to help government departments conduct convenient interactive visual analysis of public opinions. The interface of the system is shown in Figure 5. Its main views include: (a) Control panel, which helps users adjust and control relevant parameters. (b) Representation visualization of semantic structure of opinions, which is used to show the semantic structure of public opinions and the topic of concern. (c) Topic semantic feature cloud visualization, describing the semantic features of the topic obtained based on Word2vec, DBSCAN and blue noise sampling, and showing the semantic features of the top five hot topics in each district. (d) Data overview window, showing topics obtained after multiple processing and the number of words on topics. (e) Visualization of the geographic distribution characteristics of topics, which describes the distribution characteristics of each topic in different districts and counties. (f) Opinions display window for displaying the corresponding original opinions of different districts and counties, and the topic words contained in the original opinions.

In order to facilitate the government departments to choose the corresponding topics or regions according to their own interests for analysis, the system provides a large number of convenient man-machine interaction window linkage operation. Users can quickly analyze and explore the hot topics of public concern and the geographical spatial distribution characteristics of the topics. The text mining algorithm encapsulates the required features in the back-end database after mining. By loading data for visual display in the front-end system, government departments can conduct visual analysis of data according to their own interests.

Combined with the algorithm function and according to the research objectives of this paper, the interaction design of the system is as follows. After the system loads the data, first, when clicking the Word2Vec button in Figure 5(a), the system displays the Word2Vec semantic structure representation projection diagram in Figure 5(b). If the DBSCAN button is clicked, Figure 5(b) shows 17 topics, each represented by a cluster of different colors. Secondly, click on any one in the topic bar chart in Figure 5(d) to highlight the position of the topic in Figure 5(b) and the sampled words of the topic synchronically. All other topics are diluted. At the same time, the word cloud of this topic is shown in Figure 5(c), and the heat distribution of this topic is interactively shown in Figure 5(e) as well. Finally, when clicking on a district in Figure 5(e), the top 5 hot topics and sampled words of this region are highlighted in Figure 5(b), and word clouds of the top 5 hot topics of this region are shown in Figure 5(c). The original opinion data of the district and county are displayed in the opinion display window in Figure 5(f), and the topic words of each district and county are highlighted in the opinion display, too. Interaction design will be demonstrated in case studies more realistically.

5 Evaluation

In this paper, we adopt React, Python, D3.js and other technologies to implement a Web-based visual analysis system. It supports users to explore topic and geographic associations for large-scale e-government text data. The effectiveness and usefulness of the algorithm and visual analysis system are verified through a series of cases on real data sets.

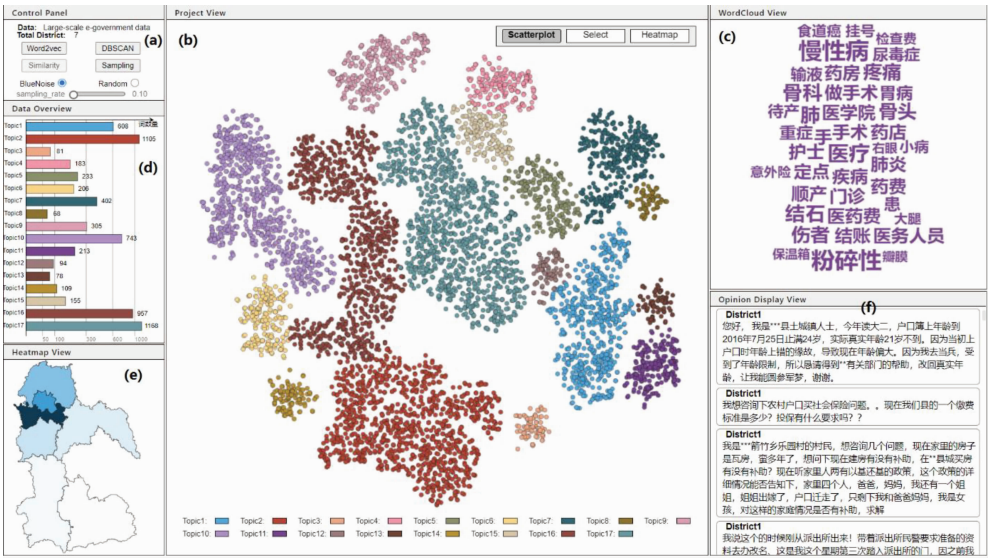


Figure 5 Visual interface of visual analysis system

5.1 Case Study

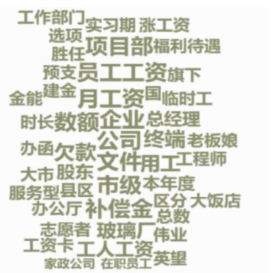
In the process of specific case study, we invite users with experience and needs of large-scale e-government text data analysis to use the designed system. The data analysis process and feedback of users are summarized, recorded and analyzed from the visual analysis perspectives of topic representation structure and semantic features, algorithm comparison and geographic spatial distribution features of topics.

Case1 Visual analysis of topic representation structure and semantic features

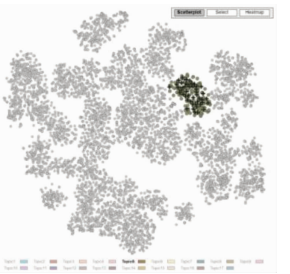
This section makes a visual analysis of the representation structure and semantic features of the topics of public concern through specific case data. Users select topics 5, 6, 9 and 14, whose word cloud diagrams and the corresponding semantic structure representation views are shown in each sub-graph in Figure 6. The highlighted cluster in Fig. 6(b) is the semantic structure representation view of topic 5 in semantic space, in which the highlights with the cross symbol are the sampled data points that correspond to the words in Fig. 6(a). These sampled points are evenly distributed in the topic cluster. It can be seen that the representative words of topic 5 in the word cloud keep the original semantics of topic 5 well, without serious loss of original semantics due to sampling. Further analysis of the topic semantic representation structure diagrams of topics 6, 9 and 14 selected by users, namely 6 (d), 6(f) and 6(h), shows that the distribution of sampling highlights is also relatively uniform. Therefore, it can be inferred that blue noise sampling can keep the semantic information of the original topic to the maximum extent, and optimize the spatial distribution of the topic words in the word cloud to avoid the problem of incomplete perception of the semantic information of the topic caused by visual clutter and occlusion.

Using the four topics selected by users, the semantic features expressed by the topics are further analyzed visually, to judge which topics the public concerns. As shown in Figure 6(a), the representative topic words are salary increase, employee salary, in-service employee welfare, treatment, internship, enterprise, company, etc. It can be seen that topic 5 is about wages, labor relations, labor security and other aspects that people pay close attention. The government should formulate more comprehensive labor law and related regulations, and

strictly enforce the law to ensure the benefits of workers. In Figure 6(c), the topic words are urban roads, vehicles, tickets, street vendors, no parking, traffic flow, roadblocks, streetlights, speed bumps, etc. It can be seen that topic 6 is about urban management, which reminds the city to improve in the management of traffic and street vendors.

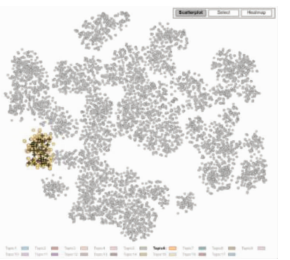


(a)Topic 5

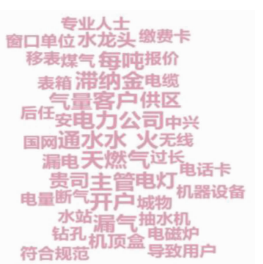


(b)Semantic structure representation for topic 5

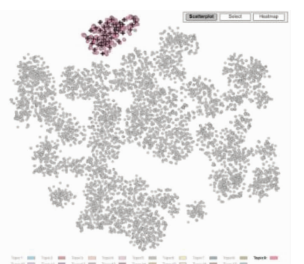
(c)Topic 6



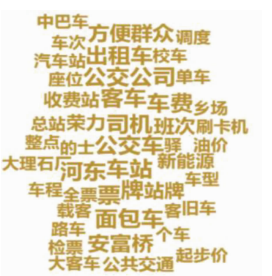
(d)Semantic structure representation for topic 6



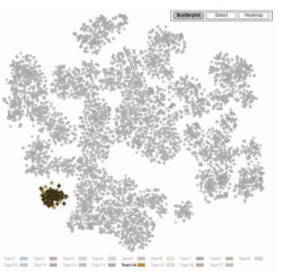
(e)Topic 9



(f)Semantic structure representation for topic 9



(g)Topic 14



(h)Semantic structure representation for topic 14

Figure 6 Visual analysis of topic representation structure and semantic features

In Figure 6 (e), the subject words are electric quantity, natural gas, power company, leakage, meter box, faucet, etc. It can be seen that topic 9 is about water supply, power supply and gas supply. People have a strong demand for basic conveniences of life, and

relevant departments should listen to public opinions, do a good job in providing services and adjust prices reasonably. In Figure 6(g), the topic words are bus, bus company, driver, public transportation, train number, station, ticket, toll booth, etc., and users can easily perceive that topic 14 is about transportation. This shows that the public's demand for convenient transportation is very strong, and the government departments should actively provide more public transportation services.

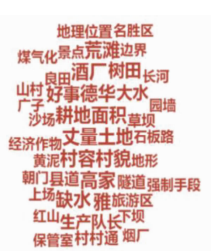
Case2 Visual analysis of algorithm comparison

This section further verifies the validity of blue noise sampling applied to semantic feature preservation of the topic by comparing it with random sampling algorithm. As shown in Figure 7, we selected three topics to compare the differences between blue noise sampling algorithm and random sampling algorithm in semantic feature preservation of topics.

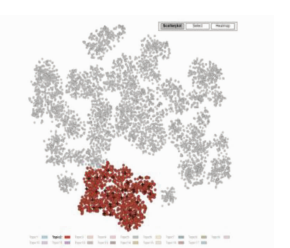
Figure 7(a) is the word cloud of topic 2, and the red highlighted part in figure 7(b) is the position of topic 2 in the semantic space, where the data points marked by crosses are the words sampled by the blue noise sampling algorithm, corresponding to the topic words in the word cloud. The red highlighted part in Figure 7(c) is also the position of topic 2 in the semantic space, where the data points marked by the cross are the words sampled by the random sampling algorithm. According to the distribution of sampled data points in Fig. 7(b) and Fig. 7(c), it can be perceived that the blue noise sampled data points have a relatively uniform distribution in the semantic space of the topic. While the results of random sampling show an irregular distribution, with either too many local sampling points or too few local sampling points. In Figure 7(c), the three unsampled local regions are further framed. That is to say, the semantics of these three local regions will be missing when the random sampling algorithm is used to sample representative topic words, resulting in incomplete semantics when the semantic features of the topic are perceived in the word cloud.

Next, for topic 7, whose word cloud is shown in 7(d), from which we can know that topic 7 is about teaching qualifications and campus life. From Figure 7(e) and 7(f), it can be seen that the distribution of sampling points of highlighted topics is still relatively uniform in the case of blue noise sampling, and relatively uneven in the case of random sampling whose data points in two local semantic regions are not sampled.

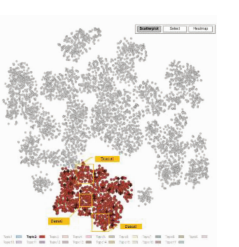
Finally, through the analysis of topic 10, it can be seen that the blue noise sampling points in Fig. 7(h) present a relatively uniform distribution. However, the random sampling points in 7(i) are not evenly distributed. In the upper left and lower right parts of the purple highlighted cluster, there is a large semantic area that does not involve any sampling points, respectively. It can be seen that the result of random sampling leads to serious loss of semantic information.



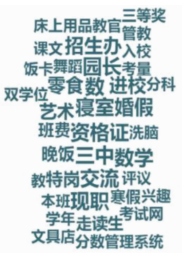
(a)Topic 2



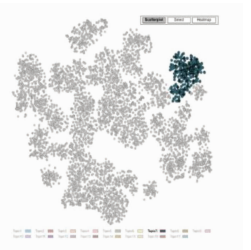
(b)Distribution of blue noise sampling for topic2



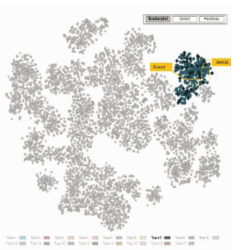
(c)Distribution of random sampling for topic2



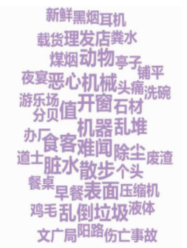
(d)Topic 7



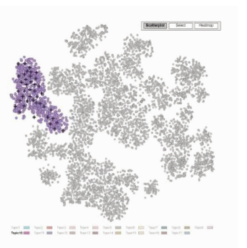
(e)Distribution of blue noise sampling for topic7



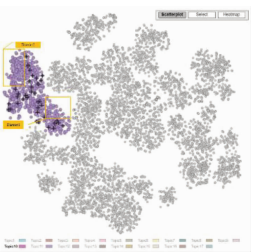
(f)Distribution of random sampling for topic7



(g)Topic 10



(h)Distribution of blue noise sampling for topic10



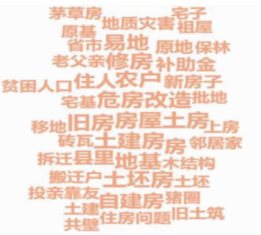
(i)Distribution of random sampling for topic10

Figure 7 Comparison between blue noise sampling and other sampling algorithms

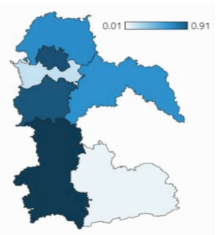
Case3 Visual analysis of geographic distribution characteristics of topics

This section further provides the visual design of the geographical spatial distribution features of topics, to facilitate the government departments to perceive the geographical spatial distribution of the topics of public concern visually. The specific visual design includes selecting a topic of interest, exploring the distribution characteristics of the topic in the geographic space and selecting a region of interest to explore the top 5 topics most relevant to the region.

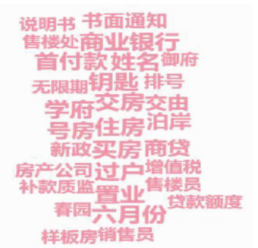
As shown in Figure 8, the geographic distribution characteristics of the topics are discussed through the user's selection of 4 topics. First, from the word cloud in Figure 8(a), it can be seen that topic 3 is about related to the renovation of dilapidated houses and the demolition of houses, which attracts the highest attention in districts 2,4 and 7. Therefore, people in these three districts pay more attention to the topic related to the reconstruction and demolition of houses than those in others. Then, the user clicks topic 4, and it can be seen from Fig. 8(c) that topic 4 focuses on issues related to house purchase, mortgage and deed tax. Fig. 8(d) shows that the topic has the highest attention in districts 5 and 16, and relatively low attention in districts 2, 3 and 4. When the user chooses topic 13, it can be found from Fig. 8(e) that the semantics of topic 13 is about social security, and as shown in Fig. 8(f), this topic has the highest attention in districts 6 and 7. Finally, the user chooses topic 15, and it can be found from the word cloud in Figure 8(g) that topic 15 is related to project construction, project tax payment and project contract signing. Topic 15 attracts more attention in districts 2 and 1 than other counties, as shown in Figure 8(h).



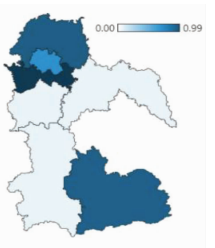
(a)Topic 3



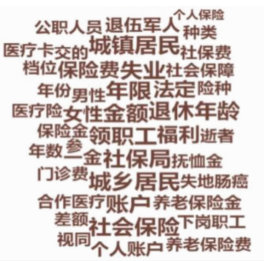
(b)Geographic distribution features for topic 3



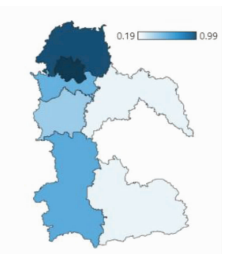
(c)Topic 4



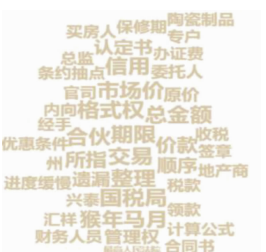
(d)Geographic distribution features for topic 4



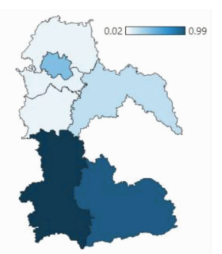
(e)Topic 13



(f)Geographic distribution features for topic 13



(g)Topic 15

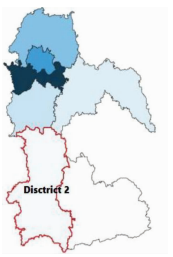


(h)Geographic distribution features for topic 15

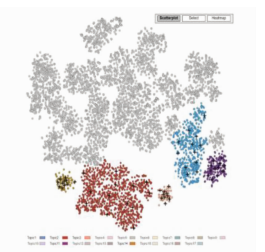
Figure 8 Visual analysis of geographic distribution features for topics

Furthermore, visual analysis of hot topics in a certain region is provided. By selecting a district or county, users can visually see the top 5 topics of concern in the region. As can be seen from Figure 9(a)-(c), the top five topics that district 2 pay attention to are Topic 1: household registration, family planning and marriage registration, etc., Topic 2: rural construction and land management, Topic 3: housing demolition and reconstruction, etc., Topic 11: medical treatment, and Topic 14: public transportation. When the user selects district 3 in Figure 9(d), Figure 9(e) shows the top five topics of concern for the district, and Figure 9(f) synchronization shows the corresponding word cloud. Users can perceive that the top five topics of concern in district 3 include topic 2: rural construction and land resource management, topic 6: city traffic management, topic 10: the urban environment and pollution con-

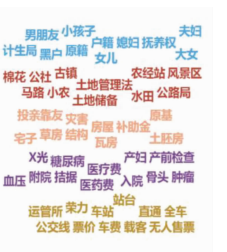
trol, topic 14: public transportation, and topic 16: community management. The user further clicks district 4 in Figure 9(g). Figure 9(h) shows the top 5 topics most relevant to district 4, and Figure 9(i) shows the word cloud of the 5 topics. The users can easily perceive the five most relevant topics of district 4 as topic 2: rural construction and land resource management, topic 3: housing demolition and reconstruction, etc., topic 6: urban traffic management, topic 14: public transportation, and topic 16: community management.



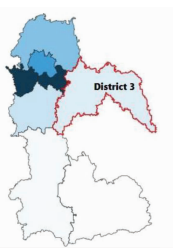
(a)District 2



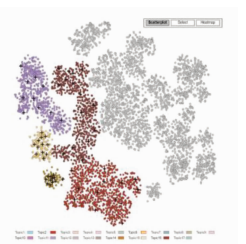
(b)Top 5 hot topics for district 2



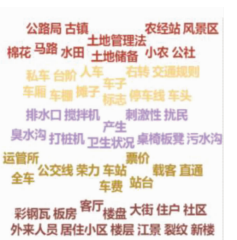
(c) Word cloud of top 5 hot topics for district 2



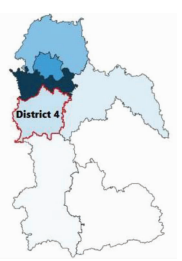
(d)District 3



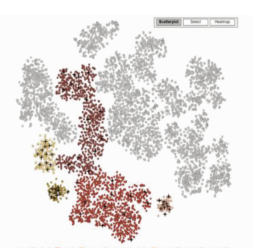
(e)Top 5 hot topics for district 2



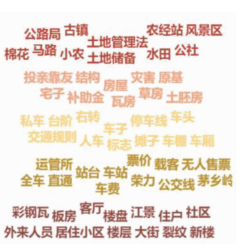
(f)Word cloud of top 5 hot topics for district 2



(g)District 4



(h)Top 5 hot topics for district 2



(i)Word cloud of top 5 hot topics for district 2

Figure 9 Visual analysis of hot topics

From the analysis of the above figure, it can be seen that although there are some differences in the hot topics concerned by different districts, there are also some common points. For example, from the point of view of traffic development, the three counties are more concerned about traffic problems. In combination with the reality, these three districts and counties are relatively deficient in economic and transportation development. People rely heavily on public transportation for travel, so they may pay more attention to the development of transportation and road construction. In addition, the three hot topics of regional discussion are all related to rural construction, land resource management and crop cultiva-

tion. Combined with the actual situation, we found that these three areas belong to the suburban counties, which have natural advantages in the development of crop planting and animal husbandry. Moreover, compared with the southern mountains, these three areas belong to the plain terrain and the terrain is relatively gentle. Therefore, when addressing livelihood issues, government departments should not only pay attention to the topics of common concern, but also take measures according to local conditions and plan public service and resource allocation policies according to the actual problems in each region.

5.2 Discussion

In this paper, the method of large-scale e-government text data exploration is to use the word representation learning method to obtain the semantic structure features of large-scale text data, and then use the density clustering method to explore the semantic region in the representation space to obtain the topic of government work opinion data. Then the topic words are obtained based on blue noise sampling. This method can divide topics based on semantic structure density, and effectively retain semantic information of original data while reducing visual clutter of large-scale network, so that users can better perceive topics. In addition, the association between topic and geography is constructed based on semantic similarity calculation, which draws on the method of calculating semantic similarity between two text datasets in the field of natural language processing. This method can well measure the degree of association between topic and geographic information. However, there are still some problems in this paper that have not been well solved and need to be further studied.

Firstly, by embedding words into vectorized space, geometric distance can be used to effectively represent semantic similarity of words. Nevertheless, due to the randomness of word representation learning and the approximation of t-SNE dimension reduction, some errors will inevitably occur. In the future work, this study will try to design a better topic mining model, so as to mine the topics of public concern more accurately and reduce the errors caused by the randomness of the model. Secondly, in addition to the deviation caused by representation learning and dimensionality reduction, blue noise sampling will also lead to the loss of original information. In future research, we will try our best to design a better algorithm to optimize blue noise sampling, or avoid using this method, but still optimize the layout of the subject word in the visual space. Finally, when visual analysis of the topic and geographic association features is conducted, due to the limitation of data acquisition, only the data of different districts and counties in a certain city is used. Therefore, in the research on the association between people's concern topics and geographical space, the difference analysis of people's concern topics in different regions is only limited to the comparative analysis of small geographical space areas. Even though different districts and counties have differences in economic conditions, terrain conditions, political status, social culture and other aspects, which makes the focus of people in different regions will be different to some extent. But we hope that future work will be able to take data from more places over a larger geographic area, so that we can study whether people's concerns are more markedly different in places that are geographically distant. This will also have a greater reference value for the government's more macro policy control.

6 Conclusion

In this paper, we explore the abstraction of large-scale e-government text data, and design

a visual analysis system to analyze large-scale text data in a convenient visual way. Firstly, the word representation learning method is used to embed the text data into the high dimensional vector semantic space. The two-dimensional semantic space is constructed by dimensionality reduction. According to the semantic structure of words, the density clustering algorithm based on the semantic structure is selected to divide the semantic region of the constructed two-dimensional plane, so as to mine the topic of public concern. Then the blue noise sampling algorithm is used to mine the representative topic words. After that, the association between topic and geographical space is established based on semantic similarity calculation. Secondly, we integrate the algorithms t0 design the flexible interactive visual analysis system of large-scale e-government text data. Finally, the validity of the algorithm and the system is evaluated through the real data set to further verify the practical value of the system.

Acknowledgements

This work was supported by the National Natural Science Foundation of China(No.61872314, No.61802339), the Natural Science Foundation of Zhejiang Province (No.LY18F020024), the Humanities and Social Sciences Foundation of Ministry of Education in China (No. 18YJC910017), the Major Humanities and Social Sciences Research Project in Zhejiang Province (2018QN021), and the Open Project Program of the State Key Lab of CAD&CG of Zhejiang University(No.A2001).

References

Andrews, K., Kienreich, W., Sabol, V., Becker, J., Droschl, G., Kappe, F., Granitzer, M., Auer, P., & Tochtermann, K. (2002). The InfoSky Visual Explorer: Exploiting Hierarchical Structure and Document Similarities. *Information Visualization, 1*, 166–181. https://doi.org/10.1057/palgrave.ivs.9500023

Angus, D., Smith, A., & Wiles, J.(2012). Conceptual Recurrence Plots: Revealing Patterns in Human Discourse. *IEEE Transactions on Visualization and Computer Graphics, 18* (6), 988–997. https://doi.org/10.1109/tvcg.2011.100

Bai, W.(2013). A Public Value Based Framework for Evaluating the Performance of e–Government in China. *iBusiness, 5*(3), 26–29.

Baojun, M., Nan, Z., & Tao, S.(2013). Big data analysis of public feedback in the context of smart cities: the perspective of probabilistic topic modeling. *E–Government*(12), 9–15.

Beil, F., Ester, M., & Xu, X.(2002). Frequent Term–Based Text Clustering. *KDD ´02 :Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 436 –442. https://doi.org/10.1145/775047.775110

BengioHolger, Y., SchwenkJean–Sébastien, SenécalFréderic, & Gauvain, M.–L.(2006). A Neural Probabilistic Language Models. In D. E. Holmes & L. C. Jain (Eds.), *Innovations in Machine Learning. Studies in Fuzziness and Soft Computing*(Vol. 194, pp. 137–186). Springer.

Cao, N., Lin, Y.–R., Sun, X., Lazer, D., Liu, S., & Qu, H.(2012). Whisper: Tracing the Spatiotemporal Process of Information Diffusion in Real Time. *IEEE Transactions on Visualization and Computer Graphics,18* (12), 2649–2658. https://doi.org/10.1109/tvcg.2012.291

Card, S., Mackinlay, J., & Shneiderman, B.(1999). *Readings in Information Visualization: Using Vision To Think.* Morgan Kaufmann Publishers.

Collins, C., Carpendale, S., & Penn, G.(2009). DocuBurst: Visualizing Document Content using Language Structure. *Computer Graphics Forum, 28*(3), 1039–1046. https://doi.org/10.1111/j.1467–8659.2009.01439.x

Dauphin, Y. N., Fan, A., Auli, M., & Grangier, D.(2017). *Language Modeling with Gated Convolutional Networks* Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning

Research. http://proceedings.mlr.press

Ebeida, M., Mitchell, S., Awad, M., Park, C., Swiler, L., Manocha, D., & Wei, L.–Y.(2014). Spoke Darts for Efficient High Dimensional Blue Noise Sampling. *ACM Transactions on Graphics, 37*. https://doi.org/10.1145/3194657

Ghanbarpour, A., & Naderi, H.(2020). An Attribute–Specific Ranking Method Based on Language Models for Keyword Search over Graphs. *Ieee Transactions on Knowledge and Data Engineering, 32*(1), 12–25. https://doi.org/10.1109/tkde.2018.2879863

Godwin, A., Wang, Y., & Stasko, J. T.(2017). TypoTweet Maps: Characterizing Urban Areas through Typographic Social Media Visualization.

Hartigan, J. A., & Wong, M. A.(1979). A K–Means Clustering Algorithm. *Journal of the Royal Statistical Society: Series C(Applied Statistics), 28*(1), 100–108. https://doi.org/https://doi.org/10.2307/2346830

He, W., Zha, S., & Li, L.(2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management, 33* (3), 464–472. https://doi.org/10.1016/j.ijinfomgt.2013.01.001

Hotho, A., Staab, S., & Stumme, G.(2003, NOV 19–22, 2003). *Ontologies improve text document clustering* Third Ieee International Conference on Data Mining, Proceedings, MELBOURNE, FL. <Go to ISI>://WOS:000188999400077

Huang, D.(2004). *History of Western Administrative Doctrine(Revised Edition)*. Wuhan University Press.

Janssens, F., Glanzel, W., & De Moor, B.(2007, AUG 12–15, 2007). *Dynamic Hybrid Clustering of Bioinformatics by Incorporating Text Mining and Citation Analysis* Kdd–2007 Proceedings of the Thirteenth Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Jose, CA <Go to ISI>://WOS:000266628300037

Kim, Y., Han, J., Yuan, C., & Assoc Comp, M.(2015, AUG 10–13, 2015). *TOPTRAC: Topical Trajectory Pattern Mining* 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Univ Technol Sydney, Adv Analyt Inst, Sydney, AUSTRALIA. <Go to ISI>://WOS:000485312900063

Lei, R., Yi, D., Shuai, M., Xiao–Long, Z., & Guo–Zhong, D.(2014). Visual Analytics Towards Big Data. *Journal of Software, 25*(9), 1909–1936.

Linders, D.(2012). From e–government to we–government: Defining a typology for citizen coproduction in the age of social media. *Government Information Quarterly, 29* (4), 446–454. https://doi.org/10.1016/j.giq.2012.06.003

Mayasari, R., Fithriasari, K., Iriawan, N., & Winahju, W.(2020). Surabaya Government Performance Evaluation Using Tweet Analysis. *MATEMATIKA*, 36, 31–42. https://doi.org/10.11113/matematika.v36.n1.1176

Metaxas, T., Makaratzi, E., & Terzidis, K.(2017). Improving service quality to local communities via a citizen satisfaction measurement in Greece: The ´MUSA´ approach. *The Journal of Developing Areas*, 51 (3), 77–101.

Nabatchi, T., Becker, J. A., & Leighninger, M.(2015). Using Public Participation to Enhance Citizen Voice and Promote Accountability. In J. L. Perry & R. Christensen(Eds.), *Handbook of Public Administration*(3rd ed., pp. 137–151). Wiley.

Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X., & Ward, R.(2014). Semantic Modelling with Long–Short–Term Memory for Information Retrieval.

Paulovich, F. V., Nonato, L. G., Minghim, R., & Levkowitz, H.(2008). Least square projection: A fast high–precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics, 14*(3), 564–575. https://doi.org/10.1109/tvcg.2007.70443

Pu, J., Liu, S., Ding, Y., Qu, H., Ni, L., & Ieee.(2013). *T–Watcher: A New Visual Analytic System for Effective Traffic Surveillance.* https://doi.org/10.1109/mdm.2013.23

Scholta, H., Mertens, W., Kowalkiewicz, M., & Becker, J.(2019). From one–stop shop to no–stop shop: An e–government stage model. *Government Information Quarterly, 36* (1), 11–26. https://doi.org/10.1016/j.giq.2018.11.010

Seifert, C., Kump, B., Kienreich, W., Granitzer, G., & Granitzer, M.(2008). On the beauty and usability of tag clouds. In E. Banissi, L. Stuart, M. Jern, G. Andrienko, F. T. Marchese, N. Memon, R. Alhajj, T. G. Wyeld, R.

A. Burkhard, G. Grinstein, D. Groth, A. Ursyn, C. Maple, A. Faiola, & B. Craft(Eds.), *Proceedings of the 12th International Information Visualisation*(pp. 17–+). https://doi.org/10.1109/iv.2008.89

Shareef, S. M., Jahankhani, H., & Dastbaz, M.(2012). E–Government Stage Model: Based On Citizen–Centric Approach in Regional Government in Developing Countries. *International Journal of Electronic Commerce Studies, 3*(1), 145–164.

Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G.(2014). A Latent Semantic Model with Convolutional–Pooling Structure for Information Retrieval. *CIKM 2014 – Proceedings of the 2014 ACM International Conference on Information and Knowledge Management*, 101–110. https://doi.org/10.1145/2661829.2661935

Song, M., & Meier, K. J.(2018). Citizen Satisfaction and the Kaleidoscope of Government Performance: How Multiple Stakeholders See Government Performance. *Journal of Public Administration Research and Theory, 28*(4), 489–505. https://doi.org/10.1093/jopart/muy006

Stylios, G., Christodoulakis, D., Besharat, J., Vonitsanou, M.–A., Kotrotsos, I., Koumpouri, A., & Stamou, S.(2010). Public Opinion Mining for Governmental Decisions. *Electronic Journal of e–Government, 8* (2), 202–213.

Tang, D., Qin, B., & Liu, T.(2015). Deep learning for sentiment analysis: successful approaches and future challenges. *Wiley Interdisciplinary Reviews–Data Mining and Knowledge Discovery, 5* (6), 292–303. https://doi.org/10.1002/widm.1171

van der Maaten, L., & Hinton, G.(2008). Visualizing Data using t–SNE. *Journal of Machine Learning Research, 9*, 2579–2605. <Go to ISI>://WOS:000262637600007

Viegas, F. B., Wattenberg, M., & Feinberg, J.(2009). Participatory Visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics, 15*(6), 1137–1144. https://doi.org/10.1109/tvcg.2009.171

Wang, Y., Chu, X., Bao, C., Zhu, L., Deussen, O., Chen, B., & Sedlmair, M.(2018). EdWordle: Consistency–preserving Word Cloud Editing. *IEEE Transactions on Visualization and Computer Graphics, 24*(1), 647–656. https://doi.org/10.1109/tvcg.2017.2745859

Wang, Z., Ye, T., Lu, M., Yuan, X., Qu, H., Yuan, J., & Wu, Q.(2014). Visual Exploration of Sparse Traffic Trajectory Data. *IEEE Transactions on Visualization and Computer Graphics, 20*(12), 1813–1822. https://doi.org/10.1109/tvcg.2014.2346746

Wattenberg, M., Viégas, F., & Johnson, I.(2016). How to Use t–SNE Effectively. *Distill, 1*. https://doi.org/10.23915/distill.00002

Wu, W., Xu, J., Zeng, H., Zheng, Y., Qu, H., Ni, B., Yuan, M., & Ni, L. M.(2016). TelCoVis: Visual Exploration of Co–occurrence in Urban Human Mobility Based on Telco Data. *IEEE Transactions on Visualization and Computer Graphics, 22*(1), 935–944. https://doi.org/10.1109/tvcg.2015.2467194

Xia, J., Ye, F., Chen, W., Wang, Y., Chen, W., Ma, Y., & Tung, A. K. H.(2018). LDSScanner: Exploratory Analysis of Low–Dimensional Structures in High–Dimensional Datasets. *IEEE Transactions on Visualization and Computer Graphics, 24*(1), 236–245. https://doi.org/10.1109/tvcg.2017.2744098

Yan, D.–M., Guo, J.–W., Wang, B., Zhang, X.–P., & Wonka, P.(2015). A Survey of Blue–Noise Sampling and Its Applications. *Journal of Computer Science and Technology, 30* (3), 439–452. https://doi.org/10.1007/s11390–015–1535–0

Yi, Y., Yi, Z., Mei, L., & Wen, D.(2019). Research on the adoption of public feedback opinions based on the LDA model: a comparative analysis of the revision of shared bicycle policy and data mining. *Information Science, 37*(1), 86–93.

Yimin, W.(2018). Comparative analysis of the public opinions of the two editions of Beijing´s master plan based on text mining. *Beijing Planning and Construction, 2*(1), 87–94.

Yin, J., & Wang, J.(2014). A Dirichlet multinomial mixture model–based approach for short text clustering. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2623330.2623715

Zhao, Y., Luo, F., Chen, M., Wang, Y., Xia, J., Zhou, F., Wang, Y., Chen, Y., & Chen, W.(2019). Evaluating Multi–Dimensional Visualizations for Understanding Fuzzy Clusters. *IEEE Transactions on Visualization and Computer Graphics, 25*(1), 12–21. https://doi.org/10.1109/tvcg.2018.2865020

Zhengrong, W.(2019). *Research on the effectiveness of online political–civilian interaction based on public perception* Lanzhou University]. Lanzhou.