Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Invited paper

Outlier knowledge management for extreme public health events: Understanding public opinions about COVID-19 based on microblog data

Huosong Xia a,c, Wuyue An a, Jiaze Li b, Zuopeng (Justin) Zhang d,*

a School of Management, Wuhan Textile University, Wuhan, 430073, China
b School of Software, Zhengzhou University, Zhengzhou, 450000, China
c Research Center of Enterprise Decision Support, Key Research Institute of Humanities and Social Sciences in Universities of Hubei Province, Wuhan, 430073, China
d Coggin College of Business, University of North Florida, Jacksonville, FL, 32224, USA

A R T I C L E   I N F O

Keywords:
COVID-19
Analysis of public opinion
Natural language processing
Outlier knowledge management
Governance suggestion

A B S T R A C T

Based on complex adaptive system theory and information theory for investigating heterogeneous situations, this paper develops an outlier knowledge management framework based on three aspects—dimension, object, and situation—for dealing with extreme public health events. In the context of the COVID-19 pandemic, we apply advanced natural language processing (NLP) technology to conduct data mining and feature extraction on the microblog data from the Wuhan area and the imported case province (Henan Province) during the high and median operating periods of the epidemic. Our experiment indicates that the semantic and sentiment vocabulary of words, the sentiment curve, and the portrait of patients seeking help were all heterogeneous in the context of COVID-19. We extract and acquire the outlier knowledge of COVID-19 and incorporate it into the outlier knowledge base of extreme public health events for knowledge sharing and transformation. The knowledge base serves as a think tank for public opinion guidance and platform suggestions for dealing with extreme public health events. This paper provides novel ideas and methods for outlier knowledge management in healthcare contexts.

1. Introduction

The COVID-19 pandemic has affected many countries with increasing morbidity and mortality. On March 11th, 2020, the World Health Organization (WHO) declared COVID-19 as a global pandemic. Until June 3rd, 2020, this outbreak virus caused over 6,500,000 detected infection cases in 210 countries and territories and around 383,000 confirmed death cases. In the absence of effective treatment and vaccine, the control of COVID-19 in China was achieved by unprecedented massive non-medical public health interventions [1].

People’s online activities could significantly affect their public concerns and health behaviors. Due to the difficulty in accessing credible information from reliable sources during a pandemic, people increasingly choose to seek relevant information on the web [2]. Therefore, measuring and analyzing hot topics and public sentiment of the COVID-19 pandemic is essential for establishing effective and efficient disease control policies.

As a novel virus, COVID-19 has exhibited unique natures with heterogeneous knowledge, which is manifested in the following aspects. First, it takes closures at different levels to prevent the large-scale spread of the virus. Second, the lack of living materials and medical protection materials seriously affects normal life and disease prevention. Third, the uncertainty of the time to return to school caused by COVID-19 seriously affects economic and social development and learning efficiency. Fourth, social media and other channels on the Internet make it easy to spread the COVID-19 fake news. The outliers of the COVID-19 pandemic have impacted everyone’s life, work, study, and behavior patterns for a prolonged period, causing people’s different levels of stress reactions [3].

When extreme public health incidents occur, epidemic rumors often spread rapidly on the Internet, which affects residents’ judgments about the epidemic and further aggravates public panic. Online social media represented by microblogs and WeChat provides a platform for netizens to obtain epidemic information and express opinions [4,5]. Mining the topics of microblogs and the sentiment corresponding to the topics can help governments, enterprises, and other organizations to predict and control emergencies, judge the public’s information needs, concerns, and emotional changes, and then make quick and timely responses.
including targeted announcements, communication, emotional comfort, and educational activities, which helps promote the scientific management of public opinion emergency [6]. The outlier knowledge of COVID-19 can provide decision-makers with scientific decision-making basis to develop knowledge through data mining and feature extraction. Analyzing public opinions on COVID-19 microblog data, acquiring outlier knowledge of COVID-19, and sharing such knowledge can increase the effectiveness of decision-making.

Research on COVID-19 can be generally divided into two categories: prevention and prediction. Prevention and control are mainly carried out from the aspects of epidemiology, clinical features, and mental health. Prediction mainly explores the relationship between the outbreak and the end of epidemics and population movements. The content posted by residents on social media during the epidemic is another important data source for the study of COVID-19, which is essential for establishing effective disease control policies. However, there exist very few studies on the public opinion analysis of COVID-19 from the perspective of outlier knowledge management (OKM). The general knowledge management is based on the “data-information-knowledge” paradigm, but the universal paradigm is lacking in solving the problem of extreme public health events considering the importance of the acquired outlier knowledge. When establishing a model for acquiring outlier knowledge of COVID-19, data dimensions and object dimensions should be taken into account to obtain relatively complete outlier knowledge. For this reason, this paper is motivated to (i) use theory to help and improve the public opinion analysis and decision-making mode in outlier scenarios, and (ii) improve and repair theories in outlier scenarios to better serve practice. Accordingly, we attempt to study the following three research questions.

1. How to identify key information in outlier scenarios? For individuals, how does social media improve the acquisition of outlier knowledge related to COVID-19, that is, how to acquire outlier knowledge of COVID-19 for individuals under extreme public health events?
2. How does the information of social media in outlier situation guide the management of public opinion and epidemic in terms of emotion and behavior, that is, what are the typical characteristics of public opinion and epidemic in extreme public health events?
3. How to reveal the epidemic situation and public opinion of extreme public health events with the typical clustering representation of social media? Do the findings of these data mining promote the construction of a public health outlier knowledge release system that can make system policy recommendations for organizations to interact with the public?

In order to answer these questions, this paper applies natural language processing (NLP) technology to analyze specific microblogs content, mainly for the abruptness and heterogeneity of the COVID-19, and uses BERT to establish the sentiment domain lexicon of COVID-19. NLP technology analyzes different online public opinions during the epidemic to help us obtain outlier knowledge of COVID-19. Its purpose is to explore the relationship between organizations, public, and publishing systems in outlier situations, thereby providing a think tank for organization building, public opinion guidance, and platform recommendations under extreme public health events.

The remaining structure of this paper is as follows. Section 2 reviews relevant research on COVID-19, public opinion analysis, and OKM. Section 3 outlines our research framework and algorithmic principle. Section 4 presents the process of refining the outlier knowledge of COVID-19 and the analysis of the experimental results. Section 5 concludes the paper by highlighting the contributions, limitations, and future research directions.

### 2. Literature review

#### 2.1. Public opinion analysis technology

The development of social public opinion research has generally experienced three stages: traditional social public opinion analysis, online public opinion analysis, and big data public opinion analysis. The key technologies of online public opinion analysis include information collection, hot spot discovery, hot spot assessment, subject tracking, and analysis and processing [7].

Current scholars mainly obtain research data through web crawler programs and website API interfaces. Commonly used hot spot discovery algorithms include the Sing-pass clustering algorithm, K-means, KNN nearest neighbor method, support vector machine (SVM) algorithm, and SOM neural network clustering algorithm. Among them, the Sing-pass clustering algorithm and K-means are widely used in current big data clustering analysis due to relatively simple rules and fast calculation speed. Hot spot evaluation, topic tracking, and analysis processing mainly include classification methods based on probability theory and information theory, such as the Naive Bayes algorithm, maximum entropy algorithm, and machine learning-based classification algorithm [8].

Microblog text topic discovery and sentiment analysis are two important aspects of public opinion analysis. The topic discovery methods for microblog text can be summarized into two categories: text clustering method and topic model method [9]. The text clustering method based on the statistical level is difficult to solve the problem of polysemy in the text, and the topic model method can overcome the shortcomings of the document similarity calculation method in text clustering, so the topic model method is often used in text topic discovery [10, 11].

Among them, the LDA (Latent Dirichlet allocation) model is widely used in the discovery of microblog text topics [12]. There are many types of research on improving the LDA model. For instance Ref. [13], proposed an LDA short text clustering algorithm based on sentiment word co-occurrence and knowledge pair feature extraction; the experimental results show that it reveals better semantic analysis ability and emotional topic clustering effect [14]. Collected questions and answers related to COVID-19 from Naver and then used the structural topic model and word network analysis to analyze the focus of people’s anxiety and worry. Their research also contributes to developing methods for measuring public opinion and sentiment in an epidemic situation based on natural language data on the Internet [1]. Collected topics related to the COVID-19 epidemic on the Sina Microblog hot search list and described the trend of public attention on COVID-19 epidemic-related topics. They found that social media (e.g., Sina Microblog) can be used to measure public attention toward public health emergencies. The importance of sentiment factors in public opinion analysis of emergencies has been widely mentioned [10]. In the field of machine learning, support vector machines, maximum entropy, and Naive Bayes classification are used in sentiment analysis. With the popularization and application of deep learning in the field of NLP, deep neural networks such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) are used for sentiment classification, for instance, using CNN to mine word, phrase, and sentence information to improve the accuracy of sentiment analysis [15].

In the era of big data, how to quickly analyze massive network data and then establish a public opinion monitoring and guidance mechanism to provide decision support for managers is the current hot research topic. Compared with traditional social public opinion analysis, social public opinion analysis in the era of big data is more focused on the collection, storage, cleaning, and text mining techniques of large amounts of network data to obtain relevant public opinion research information from a large number of low-value density data [16]. “Data” and “technology” are the key production factors of big data public opinion research. Based on the big data public opinion research path
driven by the “data-technology paradigm”, combined with the perspective of OKM, deep learning technology is used to conduct public opinion research on the new crown microblog data and then analyze and refine outlier knowledge under extreme public health event scenarios. By breaking the barriers of disciplines and creating a more open, cooperative, and inclusive research environment, our research will promote the deep development of big data public opinion research.

2.2. COVID-19 research

COVID-19 is a sudden public health event; because of its rapid and widespread spread, multiple transmission routes, and strong infectious nature, there is no specific treatment for the time being, so the virus poses a huge threat to human life and health, and it is also likely to cause tension and anxiety in the public psychological reaction [17]. Existing research on COVID-19 can be divided into five categories: etiology, epidemiology, clinical characteristics, treatment prevention and control, and mental health [18], [19]. Used the flow of international passengers in and out of Wuhan to predict how the virus will spread after a disease outbreak [20]. Used deep learning algorithms to predict the host and infectivity of COVID-19 in Wuhan and considered that bats and mink were potential hosts of COVID-19 in Wuhan from the perspective of infectivity [21]. Analyzed the 425 diagnosed patients and found that 55% of the patients who developed the disease before January 1, 2020, were related to the South China seafood market in Wuhan, and only 8.6% of the subsequent cases were related to the seafood market; they believed that there had been evidence of human-to-human transmission of COVID-19 in mid-December 2019 [22]. Conducted an online survey of the Chinese public using snowball sampling techniques to better understand their psychological impact, anxiety, depression, and stress levels in the early stages of the COVID-19 outbreak [2]. Designed a mathematical model to quantify the roles of information super-spreaders in single specific information which outbreaks rapidly and usually had a short duration period and to examine the information propagation dynamics in the Chinese Sina-microblog. Data fitting from the real data of COVID-19 obtained from Chinese Sina-microblog can identify the different contact rates and forwarding probabilities and can be used to evaluate the roles of opinion leaders in different stages of the information propagation and the outbreak unfolding. Early detection of large-scale infectious diseases and understanding of people’s ability to respond to these diseases are the concerns of governments around the world [23–25]. During an epidemic, research on public knowledge, attitudes, and behaviors can refer not only to the formulation of communication and sentiment mitigation strategies but also to the help with future prevention plans [26,27].

Social media has been recognized as an important source of new information, such as detecting earthquakes, monitoring ongoing disaster events, tracking public opinion, human behavior research, and public health issues [28]. In the field of public health, research on social media can help the government understand the public’s health information to promptly warn or intervene and make targeted recommendations [29,30]. Information published on social media has been identified as an indicator of public health issues, such as detecting influenza [31].

2.3. Research on OKM

From the perspective of knowledge management (KM), knowledge can be divided into common knowledge and outlier knowledge. Outlier knowledge comes from the valuable data part of outlier data that refers to the data that deviates significantly and is not satisfied with the general pattern and behavior of the data [32,33]. In recent years, scholars have proposed a large number of outlier data detection methods, which can be roughly summarized into six types, namely statistical methods, distance methods, density methods, depth methods, deviation methods, and clustering methods [34]. To study outlier knowledge from the perspective of outlier detection methods, the key to the analysis of outliers is to infer the properties of outliers to provide a reference for the next step of data analysis and processing, not just to explore outliers [35]. The knowledge of outliers is derived from outlier data through identification, extraction, mining, and analysis of the existing data in the database through machine learning and data mining [36].

OKM has a wide range of applications in the fields of financial fraud, network intrusion, medical diagnosis, face recognition, etc. The extraction and analysis of outlier data features are the core parts of OKM [37,38]. Because COVID-19 is an extreme public health incident, under the influence of the epidemic, the public shows different psychological and behavioral features using microblogs and other important platforms to express their views and sentiments. This paper intends to analyze public opinions based on microblogs from the perspective of OKM, refine outlier knowledge of COVID-19, and promote the construction of the outlier knowledge system of public health.

3. OKM framework and algorithmic principle

3.1. Theoretical Foundation

OKM refers to the process from outlier acquisition to outlier knowledge acquisition and sharing. It can be expressed in terms of the combination and application of outlier detection and KM. In the era of data and information overload and knowledge crisis, OKM theory and methods also face challenges, especially in understanding the mechanism of outlier knowledge.

Information theory [39,40] believes that the process of transforming data into information is the process of data processing and analysis. The formation of knowledge needs to rely on certain technical acquisition capabilities. Knowledge is objectively existing and not transferred by human will.

The basic concept of complex adaptive system theory ([51] [41,42]; is the interaction and adaptability of the main body in the environment; The above two theories have shown that KM has important values in application scenarios, but the existing theories have gaps in guiding the solution of today’s problems in extreme scenarios, especially for OKM in outlier scenarios.

In the Internet environment, data is typically unstructured, heterogeneous, and from multiple sources. The location of outlier knowledge is different from anomaly detection, so the mining of outlier knowledge from outlier data cannot be satisfied in depth and breadth. The elements of outlier knowledge should be considered from both the object and the subject perspectives. The object refers to the outlier data, while the subject should include participants, i.e., characters through analyzing the behavioral features and psychological emotions during the interaction between the person and the object to extract new valuable outlier knowledge to avoid the loss of knowledge. At the same time, in different scenarios, outlier knowledge has different effects, and the acquisition of outlier knowledge in extreme public health event scenarios is more concerned at the theoretical and application levels.

Fig. 1 summarizes the above rationale based on existing theories. This paper addresses the existing theoretical gap from three aspects—dimension, object, and situation—to propose a new OKM framework.

3.2. OKM based on ‘dimension + object + situation’

The outlier of COVID-19 emerges from the situation when people express their opinions and emotions of COVID-19 on social media platforms, different from those in the general situation. In the context of extreme public health events based on COVID-19, we use advanced NLP technology to study COVID-19 content on microblogs with respect to topic distribution, keyword extraction, sentiment analysis, named entity recognition, and relationship recognition, and then acquire COVID-19 outlier knowledge to construct the extreme public health events outlier knowledge base for knowledge sharing and transformation.
Fig. 1. Theoretical foundation.

Fig. 2. COVID-19 public opinion analysis research framework.

Fig. 3. BERT input representation.
Fig. 2 shows the COVID-19 public opinion analysis research framework based on the ‘dimension + object + situation’ OKM framework and the algorithmic principles.

3.3. BERT principle

BERT can be applied to a variety of NLP tasks. It uses the Transformer [43] language model, which abandons the recursive structure and uses the attention mechanism to mine the relationship between input and output. The input representation is shown in Fig. 3. The BERT Input embedding is the sum of Token embedding, Segmentation embedding, and Position embedding [44]. In the pre-training stage, BERT uses a masked language model, that is, randomly masks some words, and then predicts them during the pre-training process, so that it can learn to merge the representation of two different directions of text.

We complete the classification task in the fine-turning stage, assign a sentiment label to each piece of data, and finally extract important features through attention weights. The higher the attention weight, the more important the word. We then calculate the attention weight with Eqs (1) and (2).

\[ f(i_a, I_o) = a_o^t \cdot \tanh(W_e \cdot I_o + U_a \cdot I_o), \]
\[ a_i = \text{softmax} \left( \frac{f(i_a, I_o)}{\sqrt{d_i}} \right), \]

in which \( a^t_o \), \( W_e \), and \( U_a \) are attention parameters, \( a_i \) represents the similarity between the word vector \( i_o \) and all word vectors \( I_o \), and \( d_i \) is the dimension of \( I_o \). The purpose of \( \sqrt{d_i} \) is to scale the soft max function to avoid \( a_i \) being too large.

3.4. BERT + LDA Topic mining model

The BERT + LDA model proposed in this paper is based on the improvement of the Lda2vec model. Lda2vec absorbs the advantages of LDA for topic distribution and Word2vec for word representation. It combines topics and document vectors and integrates the ideas of word embedding and topic models [45]. LDA is a document topic generation model, also known as a three-layer Bayesian probability model, which contains a three-layer structure of words, topics, and documents [46]. Word2vec and BERT are both classic models in language representation. In contrast, BERT has the following two advantages.

1. Word representations produced by Word2vec are static, regardless of context. BERT uses the Transformer as a feature extractor. This method naturally makes good use of context [47].
2. Word2vec is relatively simple and cannot reflect the complex characteristics of words, such as grammar and semantics. Because BERT learns a deep network, it can get different levels of features on different network layers after pre-training [48]. Given the excellent performance of the BERT pre-trained language model in NLP tasks, this article attempts to replace Word2vec in the Lda2vec model with BERT to form a new training model, as shown in Fig. 4.

4. Experiment and model verification

This paper takes COVID-19 as an example to verify the proposed model framework and conduct the public opinion analysis of the COVID-19 epidemic situation based on microblog data. Specifically, we first construct a thesaurus of outlier knowledge areas of COVID-19 and then use BERT + LDA to conduct public opinion analysis on the microblog data of high and medium epidemic conditions in Wuhan area and input case provinces with respect to topic distribution, keyword extraction, sentiment analysis, named entity recognition, and relationship recognition to obtain new outlier knowledge of COVID-19.

We searched for the keyword ‘COVID-19’ in the hot search on microblog and limited the time between January 29, 2020, and March 30, 2020. A Python web crawler was used to crawl the content of blog posts; due to the large number of microblogs during the period, we set the crawling rules when writing the crawler program by specifying the step length of page-turning crawling to 3. We then performed data cleaning by (i) removing null data, invalid data, and duplicate data; and (ii) conducting a fuzzy search for data containing words such as “news” and “daily” in the blogger’s name field, and deleting such data because official microblogs mainly publish information about newly confirmed cases and epidemic prevention and control policies and there is a lot of duplication. Finally, 7560 valid data points were retained, which were further divided into two categories: positive emotions and negative emotions. The training set, verification set, and test set were divided according to the ratio of 8:1:1.

The experiments in this paper were based on Tensor-flow deep learning framework. First, we downloaded the pre-trained BERT model from Google (Chinese_L-12_vcrH-768_A-12), and then pre-processed the data set. Since BERT is based on Transformer with self-attention being its basic component, it can extract important information according to the weights assigned to different features by the attention mechanism and form a new lexicon in the sentiment field of COVID-19 while completing the sentiment classification task. Word2vec and BERT were used as comparative experiments. Finally, the effectiveness of our model is evaluated by extracting parameters from the confusion matrix, namely, precision, recall, and accuracy. The confusion matrix contains four terms: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). For a given class \( c \), which could be either positive or negative, \( \pi \) represents the corresponding opposite class. The TP represents the number of reviews of class \( c \) that were correctly classified as class \( c \), whereas the FP indicates the number of reviews of class \( \bar{c} \) that were incorrectly classified as class \( c \). Also, TN depicts the number of reviews of class \( \bar{c} \) that were correctly classified as \( \bar{c} \), whereas FN represents the number of reviews of class \( c \) that were erroneously classified as \( c \).

These evaluation metrics are discussed briefly as follows. Precision quantifies the exactness of a model and is defined as the ratio of correctly predicted reviews (TP) to the total number of predicted reviews (TP + FP) in any class \( c \), where \( c \) may be positive or negative class. The formula of precision is

\[ \text{Precision} = \frac{TP}{TP + FP}. \]  

Recall computes the completeness of a model and is defined as the ratio of correctly predicted reviews (TP) to the total number of actual
The formula of recall is

$$\text{Recall} = \frac{TP}{TP + FN}$$

Accuracy evaluates the correctness of a model and is calculated as the ratio of correctly predicted to the total number of reviews, which is

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$  

Table 1 shows the comparison of the evaluation criteria of different models, which indicates that the performance of BERT is more superior. Through the analysis of the COVID-19 field lexicon, we find that the main negative sentiments are expressed as follows.

(1) The negative expressions of the information interaction between organizations, people, and platforms primarily include the overall situation of the epidemic, the effective allocation of medical resources, the timeliness of the epidemic report, the lack of early scientific prevention and control plans, and the informed decision-making process.

(2) The negative performance of public opinion guidance and education mainly consists of public opinion guidance and classification of public access to epidemic information, knowledge, and authoritative release platforms.

(3) The negative behaviors of the organizations on wildlife market regulation are those for market ecological regulation.

The main positive sentiments are mainly manifested in three aspects: the effective prevention and control of COVID-19, the support of medical staff and volunteers from various places in the hardest-hit areas, and the free opening of various resources by the state, which we further discuss as follows.

(1) When the overall situation of COVID-19 epidemic is easing, the epidemic is effectively prevented and controlled, and the epidemic information is announced every day, the feature words are as follows: no new addition, cure, immunity, decline, stabilize, hold on, epidemic prevention and control, openness, and transparency.

(2) When national medical staff supports the hardest-hit areas and volunteers from all over the world donated materials and mourned the medical staff who died during the epidemic, the feature words are as follows: farewell, chartered car, evacuation, donation, hardcore, tribute, warrior, retrograde, dedication, rush, support, donation, silence, mourning, Role model, and volunteer.

(3) When resources such as academic websites and video websites are open for free as well as medical treatment, living allowances, and return to work and school are on the agenda, the feature words are as follows, free access, back to school, normal, exercise, sunshine, dream, and expense.

Among them, negative sentiments mainly stem from the public’s panic about the situation of COVID-19 and the shortage of medical resources, followed by the impact of lockdown and isolation on life. In the positive sentiments, there are many vocabularies about the national medical staff supporting Wuhan, and joint efforts to fight the epidemic are the main source of the residents’ positive sentiments. During the epidemic, effective prevention and control policies such as strict isolation, wearing masks, shelter hospitals, effective rescue, supporting living measures, strong execution, and daily publication of epidemic information, have an important impact on the sentiment trend of residents. Besides, the support and unity of all parties in the face of difficulties to jointly fight the epidemic are the key measures to achieve significant results in the prevention and control of the epidemic.

The COVID-19 sentiment lexicon is shown in Table 2. The field lexicon indicates that some words express different meanings in the context of COVID-19 from those in generic situations. The construction of the sentiment lexicon is conducive to the understanding and interpretation of the topic’s distribution and helps to improve the accuracy of sentiment classification. For example, parents, college entrance examination, online courses, new words, resources, help, peak, and other similar words have a negative sentiment tendency in the context of COVID-19, whereas decline, evacuation, send-off, retrograde, no addition, mourning, hardcore, volunteer express positive sentiments.

### 4.2. COVID-19 outlier knowledge acquisition process

#### 4.2.1. Data sources and experimental process

This paper uses Python to develop web crawlers to crawl microblog data. According to the research content, the crawled content is divided into three parts: (i) all microblogs under the topic of #Life of residents after the closure of the city#, resulting in a total of 804 data points, (ii) all the contents of the homepage for posting help messages on microblogs, resulting in a total of 1128 data points, and (iii) the comments under the microblogs published by the Henan Health Commission from 2020.01.19 to 2020.02.25, resulting in a total of 7600 data points and 7500 items remaining after cleaning. A total of 18,600 pieces of data were collected from the microblog hot discussions during the high and medium operating period of each city; after data cleaning, 18,200 pieces of data were retained. The experiment was divided into three parts, and BERT + LDA was used to analyze the topic distribution and sentiment tendency of microblog topics of Wuhan citizens’ life in provinces of input cases during the closure of the city as well as the portraits of patients with the COVID-19 seeking help, such as the patient’s specific location information, gender, age, and regional distribution. Through the above analysis, we can extract outlier knowledge of COVID-19 and provide reference and guidance for the prevention and control of extreme public health events.

#### 4.2.2. Analysis of experimental results

##### 4.2.2.1. Topic distribution

First, we used BERT + LDA to extract the topic distribution of Wuhan citizens’ microblogs, conducted a comparative experiment with Lda2vec, and selected a suitable topic number K by calculating KL divergence (Kullback–Leibler divergence) [49]. The KL distance was used to calculate the non-symmetric difference or dissimilarity between two probability distributions. Given the word probability distribution $U$ for Topic 1 and $Y$ for Topic 2, the KL divergence between Topics 1 and 2 is defined as:

$$KL(U | Y) = \sum_{i=1}^{V} U(i) \log \frac{U(i)}{Y(i)},$$

where $|V|$ is the total size of the vocabulary. Because of the non-symmetric difference, the KL distance between the two topics is calculated as:

$$KL(U, Y) = \frac{(KL(U | Y) + KL(Y | U))}{2}.$$

In this study, we selected the appropriate topic number $K$ with the maximum average KL distance between topics. For example, when the topic number $K$ is set to three, the average KL distance between any two of the three topics is defined as:

### Table 1
Comparison of evaluation criteria of different models.

| Method   | Accuracy | Recall | Precision |
|----------|----------|--------|-----------|
| Word2vec | 0.847    | 0.86   | 0.903     |
| BERT     | 0.899    | 0.9    | 0.945     |
KL values of different models (Wuhan citizens)

Table 3

| Topics number | High operating period | Median operating period |
|---------------|-----------------------|------------------------|
|               | BERT-LDA              | Lda2vec                |
| 3             | 9.25                  | 8.99                   |
| 4             | 8.71                  | 8.08                   |
| 5             | 8.16                  | 7.15                   |
| 6             | 7.03                  | 7.08                   |
| 4             | 8.63                  | 7.15                   |
| 5             | 8.99                  | 7.15                   |
| 6             | 8.08                  | 7.15                   |

KL(k = 3) = (KL(U, Y) + KL(U, Z) + KL(Y, Z))/3

As Table 3 shows, in the high operating period, BERT-LDA was used to extract the topic distribution; when the topic K was set to 3, the maximum value of KL was 9.25. The topic distribution was extracted with Lda2vec; when the topic K was set to 3, the maximum value of KL was 8.96. The KL value of the BERT-LDA model was greater than that of Lda2vec; therefore, the BERT-LDA model proposed in this paper can achieve a better classification effect. During the median operating period, BERT-LDA was used for topic distribution extraction. When topic K was set to 4, KL obtained the maximum value of 8.52. The KL value of the BERT-LDA model was greater than Lda2vec, and the model proposed in the paper was better than Lda2vec; therefore, the proposed model was superior to Lda2vec.

Table 4 shows the topic distribution of Wuhan citizens’ microblogs and keywords under specific topics during the high and medium operating periods of the epidemic. In the high operating period, the contents of the citizens’ microblogs include three topics: epidemic prevention, material needs, and daily life. Residents’ attention mainly focuses on the prevention and control of COVID-19. In the medium operating period, the topic of expressing emotions appears; because of the continued decline of the diagnosed cases, the citizens’ emotions have changed, appearing happy, cheer, etc. Besides, due to the long-term closure, citizens may lack supplies and feel psychological anxiety, so words such as bored, too difficult, and unable to buy start to appear. By observing the keywords of the high and medium operating period of the epidemic under the topic of daily life, we can find that words such as group purchase, group, organization, property, and distribution appear in the keywords during the medium operating period. At the present stage, all communities focus on group purchase and distribution. Observing the keywords of the high and medium operating period of the epidemic under the topic of material demand, we find that the food types in the keywords of the medium operating period are relatively richer than those in the high operating period. Residents have different needs and sentiments at different stages of epidemic control. In the early stage of the epidemic, correct prevention and control measures are crucial, such as quarantine and isolation, allocation of medical resources, open and transparent information, and avoiding the spread of rumors and false information. With the continuous advancement of epidemic prevention and control, in order to enrich the lives of citizens during the lockdown period, the government distributes caring vegetables to community residents and organizes them to purchase materials by way of group purchase to meet the needs of life. Major academic resource websites are also available for free to facilitate scholars and students to conduct scientific research; many websites publish online fitness courses and free movies and video resources to Wuhan citizens.

Second, we analyzed the microblog hot discussions of the provinces of the input cases. As Table 5 shows, in the high operating period, we first used BERT-LDA for topic distribution extraction; when the topic K was set to 4, the maximum value of KL was 10.6. Lda2vec was then used for topic distribution extraction; when the topic K was set to 3, the maximum value of KL was 9.77 and the KL value of the BERT-LDA model was greater than that of Lda2vec. Therefore, the BERT-LDA model proposed in this paper can achieve a better classification effect. During the median operating period, we used BERT-LDA for topic distribution extraction; when topic K was set to 4, KL obtained the maximum value of 13.69. Lda2vec was used for topic distribution extraction; when topic K was set to 4, KL obtained the maximum value of 12.03. The KL value of the BERT-LDA model was greater than Lda2vec, and the model proposed in the paper was better than Lda2vec. Therefore, the proposed model was superior to Lda2vec.

Table 6 shows the four topics hotly discussed on microblogs in the high and medium operating period, and the keywords under each topic. During the high and medium operating periods of the epidemic, the number of topics included in the hot discussion on microblogs did not change, but the content of the topics changed. For example, the keywords under the topic of returning to work and school varied in different periods. In the high operating period, due to the impact of the epidemic, time that should have been spent at work and schools was completely spent at home; there was an urgency of returning to work and school. During the medium operating period, newly diagnosed cases in many areas had achieved zero growth, so some areas had begun to open to traffic and the order of life had slowly returned to normal; returned people began to go through procedures such as passes and health certificates to prepare for the return to work or school. For the topic of epidemic protection, in the high operating period, the keywords were epidemic, mask, diagnosis, etc. In the median operating period, the keywords were zero growth, isolation, cure, hardcore, etc. By observing the keywords under this topic in the high-median operating periods, we found that the epidemic prevention and control had achieved remarkable results, and the changes in the keywords were manifested from the epidemic situation to the epidemic prevention and control effect. Regarding the topic of emotional attitude, both positive and negative sentiments appeared during the high-median operating period. During the high operating period, negative sentiment words expressed fear and panic about COVID-19 as well as the worries about the confirmed cases around them; positive sentiment words mainly expressed the anticipation of the arrival of effective control of the epidemic. During the median operating period, positive sentiments were affected by the effective control of the epidemic situation and the declining number of newly confirmed cases, whereas negative sentiment words were due to the emergence of new confirmed cases in certain areas, which had increased...
4.2.2.2. Analysis of sentiment tendency. Fig. 5(a) shows the sentiment trend of comments released on microblogs by Henan Provincial Health Commission, the information released by the official microblogs of the Health Commission is more authoritative and credible, the content is mainly about the epidemic situation, prevention and control policies, and the travel trajectory of confirmed cases. Observing the sentiment curve, we can find that the residents’ sentiment fluctuations are large, and the sentiment scores are sometimes high and sometimes low. During the period from February 6 to 11, the spread of the epidemic continued to expand, and the number of people diagnosed with COVID-19 in Henan surged. On February 11, the spread reached the maximum, and the number of confirmed cases with COVID-19 reached a peak. During this period, the trend of the sentiment curve was relatively low, indicating that residents’ sentiment fluctuations were mainly affected by the growing trend of the epidemic. Fig. 5(b) shows the sentiment curve of Wuhan citizens during the lockdown period. Compared with imported case provinces, it can be seen that the overall sentiment scores of Wuhan citizens are low. The epidemic information and prevention and control measures have a greater impact on the psychology and sentiments of Wuhan citizens. The main reason is that Wuhan is in the hardest-hit area of the epidemic and has attracted national and even global attention. The epidemic prevention and control policy is closely related to emotional fluctuations, for example, on January 23, Wuhan government announced the closure of the city, and the public sentiment curve showed a sharp downward trend; 11 new square cabin hospitals were built on February 4, which can provide more than 10,000 beds for 7 days, and the sentiment curve rises. On February 10, 19 provincial counterparts supported 16 cities, prefectures, and county-level cities in Hubei except for Wuhan, and the sentimental curve trend reached the highest.

4.2.2.3. Portrait of pneumonia patient for help. The microblog platforms provide a helpful venue for patients with new-type coronary pneumonia who cannot receive timely assistance. The portrait of the patients is shown in Fig. 6(a): 52.85% of patients seeking help were between 60 and 80 years old, and the proportion of infancy and young patients seeking help was relatively low. As shown in Fig. 6(b), a larger proportion of women were seeking help. As shown in Fig. 6(c), 92.2% of the patients were located in the Wuhan area. Wuchang, Hanyang, and Jiang’ an districts in Wuhan City had more people seeking help, and the number of people in Jiangxia district asking for help was small. Fig. 6(e) is a word cloud of help content. The words dad, mom, grandma, old man, uncle, and husband can be seen as the relatives of COVID-19 patients. There were too many middle-aged and elderly people with the COVID-19, and there may be obstacles for them to use the microblog help channel. Cough, fever, and severe illness are the symptoms of patients seeking help. From the words including urgent, urgent need, rescue, bed, treatment, and admission, we can observe the urgency of the need for patients seeking help to be admitted to hospitals.

4.3. Discussion

The analysis of microblog data based on text mining and our proposed OKM framework lead to the following implications.

First, positive emotions and negative emotions tend to have typical features in different situations. Personal expression features when extreme public health events occur differ from those in common situations. For instance, words such as parents, college entrance examinations, online courses, new additions, resources, help, and peak are related negative emotions in the COVID-19 pandemic, whereas words such as descent, evacuation, farewell, retrograde, no new additions, mourning, hardcore, and volunteers express positive emotions in the COVID-19 pandemic.

Second, the sentiment curve of extreme public health events has outlier features. The first category is that in the provinces with imported cases, the sentiment fluctuations of the residents during the high operating period of the epidemic are relatively large and the trend of the...
sentiment curve is low. During the medium operating period of the epidemic, the sentiment curve gradually increases. By extracting the hot topic discussion on microblogs, we find that the topics discussed during the high and medium operating period include the four topics—return to work and school, epidemic prevention, sentiment attitude, and closure of the city, whereas during the medium operating period, the keywords under the topic of return to work and school include health certificate, certificate handling, office, work, rework, and pass, indicating that most of the personnel are preparing for returning to work or school. During the high operating period, the keywords under this topic are school, college entrance examination, returning to work, working, earning money, online courses, express delivery, kneel to beg, go out to play, and want to see. The other category is that in the Wuhan area, the citizens’ sentiment fluctuations during the high and medium operating periods are not large, and the sentiment scores are lower than the overall sentiment scores of the imported provinces. Through topic distribution detection and keyword extraction, we notice that after the closure of the city, the living materials of Wuhan citizens were purchased through community groups. The basic living requirements can be met but the quality and richness were not high. In general, the life and mood changes of the citizens of the province with imported cases were more obvious than those of Wuhan citizens during the high and medium operating period.

Finally, the portraits of patients seeking help in extreme public health events also have outlier features. The features of COVID-19 patients who seek help show that on the one hand, the regional features were in a blowout state. COVID-19 patients who seek help were located in Wuhan city, and some patients belonged to the urban area surrounding Wuhan. In Wuhan, Wuchang district and Jiang'an district had more patients asking for help, while the Jiangxia district had the least number of patients asking for help. On the other hand, there were large differences in age and gender. COVID-19 patients accounted for a large proportion of the 60–80 age group, and the proportion of female patients was relatively large.

5. conclusion

From the OKM perspective, this paper conducts public opinion mining on COVID-19, constructs a vocabulary (positive and negative) in the field of extreme public health events, and adds new outlier knowledge in the knowledge base of extreme public health events. Our research helps to provide a reference for the prevention and control of sudden public health events in the future. It also offers a new research perspective for public opinion analysis.

5.1. Theoretical contribution

The data features and analysis of data mining in this research make theoretical contributions to the existing literature, which we summarize as follows.

First, we use complex adaptive system theory and information theory to help and improve public opinion analysis and decision-making models under extreme major public health events. The theory of complex adaptive systems emphasizes the initiative and adaptability of the subject, and constantly “learns” or “accumulates experience” to adapt to the environment. Based on this theory, the paper acquires outlier knowledge in the context of COVID-19 and proposes a think tank for the prevention, control, and management of extreme public health events. Information theory emphasizes the importance of data mining methods for obtaining valuable information. In the context of extreme major public health events, microblog public opinion is an important data source for obtaining outlier knowledge. Based on the above two theories, the paper uses advanced NLP technology to conduct data mining on the content of COVID-19 on microblogs, and obtain outlier knowledge of COVID-19, which enriches and enhances the prevention and control of extreme public health events.

Second, we improve the OKM framework for dealing with extreme public health events. The paper constructs a new “dimension-object-situation” outlier knowledge acquisition mode. To avoid the loss of knowledge, especially outlier knowledge, the idea of classification should be integrated when acquiring knowledge, including data dimension classification, situation classification, and the classification of different subjects. Under the context of extreme public health events, we obtain outlier knowledge through the analysis of public opinions on microblogs of different subjects of COVID-19 and propose a think tank for the prevention, control, and management of extreme public health events based on the outlier knowledge of COVID-19.

5.2. Practical contribution

Our research findings indicate that the scientific release system can promote good interaction between government and people under the scenario of extreme public health events. We highlight the specific practical values of our research as follows.

First, the credible and timely construction of an outlier knowledge release system is crucial for strengthening the interactions between the government and the public. Therefore, it is necessary to establish clinical data document standards to accumulate individual cases, store personal health data in a unified form, form logical associations between cases, build outlier knowledge bases to tap the value of outlier knowledge, and provide the necessary foundation for other similar emergencies events in macro medical management. The publishing system should be “flat” and
“flexible”, so the prevention and control data link and emergency intelligence architecture in the system can facilitate real-time data sharing. The publishing system should also reflect the authority and the timeliness of the mining and dissemination of outlier knowledge. Through establishing intelligent retrieval technology, the scientific discovery and progress of the world can obtain the required outlier knowledge from the publishing system and provide scientific guidance and governance for scientific defense and epidemic situations.

Second, the scientific and efficient construction of organizational systems helps develop an organization’s think tank. It is necessary to use data mining on time to form an outlier knowledge graph for public opinion analysis. Therefore, emotion words that are different from public governance in generic situations can be identified to analyze the positive emotion words and effective measures of governance, to correlate negative emotion words and invalid measures, to continuously optimize policy measures, and to release effective information in a timely manner. In terms of organizational governance, it is necessary to make good use of news media and scientific analysis and mining, to form outlier portraits with different characteristics, to scientifically allocate and optimize defense and treatment resources according to the portraits, and to optimize the scheduling of living resources. In organizational decision-making, an automated intelligent epidemic analysis visualization system can be established to provide a scientific basis for government decision-making.

Third, a good public opinion ecosystem can guide and educate people’s behavior, improve their knowledge of extreme public events, and help them constantly learn new knowledge. For individuals, the occurrence of extreme public health events takes place when there is no scientific mechanism to explain it, so the existing effective defense knowledge should be used to strengthen personal epidemic prevention capabilities. The outlier knowledge enables people to face the challenges scientifically and optimistically and guide and tolerate their misconduct caused by ignorance and panic. People can enhance their ability to scientifically identify the true and false of different data sources and improve how science uses new media to express their claims. The hierarchical and classified governance measures will help eliminate rumors and clarify the facts promptly without large-scale screening and deletion of posts, saving social costs. For the rumors that need time to verify, the progress information should be released to prevent the spread of rumors and cause panic. For rumors with uncertain consequences, we must trace their sources, listen to the opinions of all parties, screen and delete posts in time, and publish highly persuasive and positive information.

5.3. Limitations and future research

From the perspective of OKM, this paper analyzes public opinions on microblogs using advanced NLP technology and proposes a new outlier knowledge acquisition framework in the context of extreme public health events. However, there are still some limitations to this research. For instance, the granularity of the analysis is not detailed enough, such as microblog topics evolution in the time dimensions, the weight of each topic, and the weight of keywords under that topic. At the same time, the reasons for the fluctuation of the sentiment curve were explained through the key node events of news reports and we did not conduct a more in-depth analysis through experimental methods. In the follow-up research, we can combine the microblog sign-in data and use GIS technology and spatial econometrics to study the relationship between urban spatial factors and virus transmission.

![Fig. 5. Sentiments tendency.](image1)

![Fig. 6. Visualization of the portrait of patients.](image2)
CRediT authorship contribution statement

Hua Xia: Conceptualization, Methodology, Writing - original draft. Wuyue An: Data curation, Software, Writing - original draft. Jiaze Li: Visualization, Software, Investigation. Zuopeng (Justin) Zhang: Supervision, Writing - review & editing.

Acknowledgement

This research has been supported by the National Natural Science Foundation of China: (NSFC, 71871139), Model of Risk knowledge acquisition and Platform governance in FinTech based on deep learning; (NSFC, 71571139), Outlier Analytics and Model of Outlier Knowledge Management in the context of Big Data. We deeply appreciate the suggestions from fellow members of Xia’s project team and Research Center of Enterprise Decision Support, Key Research Institute of Humanities and Social Sciences in Universities of Hubei Province (DSS202007A1).

References

[1] Zhao YX, Cheng SX, Xu XY. Chinese public’s attention to the COVID-19 epidemic on social media: observational descriptive study. J Med Internet Res 2020;22(5).
[2] Yin F, Xia X, Song N, Zhu L, Wu J. Quantify the role of super spreaders – opinion leaders – on coronavirus information propagation in the Chinese sina-microblog. PLoS One 2020;15(6).
[3] Ying L, Gayle AA, Amies WS, Joachim R. The reproductive number of covid-19 is higher compared to sars coronavirus. J Trav Med 2020;27(2):1–4.
[4] Alexander DE. Social media in disaster risk reduction and crisis management. Sci Eng Ethics 2014;20(3):717–32.
[5] Li J, Li X, Yen DC, Zhang P. Impact of online review grouping on consumers' system usage behavior: a system restrictiveness perspective. J Global Inf Manag 2019;26(4):45–66. https://doi.org/10.4018/JGIM.2016100103.
[6] Kim J, Hastak M. Social network analysis: characteristics of online social networks after a disaster. Int J Inf Manag 2018;38(1):86–96, 3.
[7] Rasmussen A, Mader LK, Reher S. With a little help from the people? the role of public opinion in advocacy success. Comp Polit Stud 2018;51(2):139–64.
[8] Zhang W, Zhu YC, Wang WP. An intelligent textual corpus big data computing approach for lexicons construction and sentiment classification of public emergency events. Multimed Tool Appl 2019;78(21):30159–33.
[9] Ravi K, Ravi V. A survey on opinion mining and sentiment analysis tasks, approaches and applications. Knowl Base Syst 2015;89:14–46.
[10] Karimi A, Shah V, Vaersli R, Banal A. Twitter speaks: a case of national disaster situational awareness. J Inf Sci 2019;46(3):313–24.
[11] Han J, Huang Y, Kumar K, Bhattacharya S. Time-varying dynamic topic model: a better tool for mining microblogs at a global level. J Global Inf Manag 2018;26(1):104–19. https://doi.org/10.4018/JGIM.2018100106.
[12] Nagamunruja R, Pethalakshmi A. A novel framework based on bi-objective optimization and LAN2FIS for Twitter sentiment analysis. Social Netw. Analys. 2020;35(2):1314–33.
[13] Wu D, Yang R, Shen C. Sentiment word co-occurrence and knowledge pair feature extraction based LDA short text clustering algorithm. J Intell Inf Syst 2020;1–23.
[14] Jo W, Lee J, Park J. Online information exchange and anxiety spread in the early stage of the novel coronavirus (COVID-19) outbreak in South Korea: structural topic model and Network Analysis. J Med Internet Res 2020;22(6).
[15] Ragini JR, Anand PMR, Bhamak V. Big data analytics for disaster response and recovery through sentiment analysis. Int J Inf Manag 2018;42(5):13–24.
[16] Zhang W, Wang M, Zhu YC. Does government information release really matter in regulating contagion-evolution of negative emotion during public emergencies? From the perspective of cognitive big data analytics. Int J Inf Manag 2020;50:498–514.
[17] Spinnelli A, Pellino G. Covid-19 pandemic: perspectives on an unfolding crisis. Br J Surg 2020;107(7):785–7.
[18] Sohrabi C, Alshar Z, O'Neill N, Khan M, Agba R. World health organization declares global emergency: a review of the 2019 novel coronavirus (covid-19). Int J Surg 2020;76:1–6.
[19] Bogoch II, Alexander W, Andreu TB, Carmen H, Kremsner MUG, Kamran M. Pneumonia of unknown etiology in wuhan, China: potential for international transmission via commercial air travel. J Trav Med 2020;26(2).
[20] Guo Q, Li M, Wang C, Wang P, Fang Z, Tan J, et al. Host and infectivity prediction of Wuhan 2019 novel coronavirus using deep learning algorithm. 2020. bioRxiv. https://doi.org/10.1101/2020.03.20.008789.
[21] Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med 2020;382:1199–207.
[22] Wang C, Pan R, Wan X, Tan Y, Xu L, Ho HS, Ho RC. Immediate psychological responses and associated factors during the initial stage of the 2019 coronavirus disease (COVID-19) epidemic among the general population in China. Int J Environ Res Public Health 2020;17(5).
[23] Johnson EJ, Haririhan S. Public health awareness: knowledge, attitude and behaviour of the general public on health risks during the h1n1 influenza pandemic. J Public Health 2017;25(3):1–5.
[24] Caligiuri P, Cieri HD, Minhaeva D, Verbeke A, Zimmermann A. International HRM insights for navigating the covid-19 pandemic: implications for future research and practice. J Int Bus Stud 2020;51(2).
[25] Syed IB. Diet, physical activity, and emotional health: what works, what doesn’t, and why we need integrated solutions for total worker health. BMJ Public Health 2020;20(152):1–9.
[26] Walter S, Böhmer MM, Reiter S, Krause G, Wichmann O. Risk perception and information-seeking behaviour during the 2009/10 influenza A (H1N1) pdm09 pandemic in Germany. Euro Surveill 2012;17(15):20131.
[27] Tomazelli G, Larr G, Gupta V, Xuereb PA, Buttigieg SC, Vassallo P. Healthcare systems and corporate social responsibility communication: a comparative analysis between Malta and India. J Global Inf Manag 2018;26(4):52–66. https://doi.org/10.4018/JGIM.2018100104.
[28] Bachner J, Hill KW. Advances in public opinion and policy attitudes research. Pol Stud J 2014;42(1):51–70.
[29] Arami A, Mykkänen S, Morita M. Twitter catches the flu: detecting influenza epidemics using twitter. Empirical methods in natural language processing. 2011, 403–412.
[30] Robinson B, Sparks R, Power R. Social media monitoring for health indicator(s): 21st international congress on modelling and simulation. 2015. p. 1862–8.
[31] Sharma GD, Talan G, Srivastava M, Yadav A, Chopra R. A qualitative enquiry into strategic and operational responses to Covid-19 challenges in South Asia. J Publ Health 2020;2195.
[32] Genoreux M, Lafontaine M, Eykelboosh A. From science to policy and practice: a critical assessment of knowledge management before, during, and after environmental public health disasters. Int J Environ Res Public Health 2019;16(4).
[33] Leroux AC, Jones H, Clenet A, Tisseyre B. Knowledge discovery and unsupervised editing. J Glim 2018;51(3):66. https://doi.org/10.4018/JGIM.2016100103.
[34] Gang R, Taeho H. Investigating online destination images using a topic-based classification. Knowl Base Syst 2020;187:104831.
[35] Bachner J, Hill KW. Advances in public opinion and policy attitudes research. Pol Stud J 2014;42(1):51–70.
[36] Zou L, Lam NS, Cai H, Qiang Y. Mining twitter data for improved understanding of an influenza epidemic among the general population in China. Huosong Xia is a professor in the school of management at Wuhan Textile University. In 2018, she got her bachelor’s degree from the school of software at Huosong Xia graduated from Huazhong University of Science and Technology in China. Huosong Xia is a professor in the school of management at Wuhan Textile University. He was a visiting scholar at Eller College of Management of the University of Arizona, USA from 2006 to 2007. His main research interests are knowledge management, data science, e-Commerce, and logistics information system. His publications have appeared in over 100 referred papers in journals, book chapters, and conferences, such as Journal of Knowledge Management, International Journal of Knowledge Management, Journal of Knowledge Management Practice, International Journal of Management, Journal of Systems Science and Information, Journal of Computing Information Technology, Journal of Grey System, Financial Innovation (Springer), and World Journal of Social Science Research. He has obtained research funding from 4 projects with National Social Science Foundation of China and National Science Foundation of China.
Zhengzhou University. Her main research interests are knowledge management, data mining and, e-Commerce, and Logistics Information System.

Ms. Jiaze Li is an undergraduate student in the school of software at Zhengzhou University. Her research interest is information security.

Dr. Zuopeng (Justin) Zhang is a faculty member in the Coggin College of Business at University of North Florida. He was previously an Associate Professor of Management, Information Systems, and Analytics at State University of New York at Plattsburgh. He received his Ph.D. in Business Administration with a concentration on Management Science and Information Systems from Pennsylvania State University, University Park. His research interests include economics of information systems, knowledge management, electronic business, business process management, information security, and social networking. He is the editor-in-chief of the Journal of Global Information Management, an ABET program evaluator, and an IEEE senior member.