Analysis and Mining of Internet Public Opinion Based on LDA Subject Classification

Mei Zhang*, Huihui Su and Jinghua Wen

Information Institute, Guizhou University of Financial and Economics, Guiyang, Guizhou, China
E-mail: zm_gy@sina.com
**Corresponding Author*

Received 15 July 2021; Accepted 15 August 2021;
Publication 06 November 2021

Abstract

This paper uses Python, R language, Gephi and other software to crawl and classify the comment content of Weibo hot search events. Using word cloud, co-occurrence social network graphs, LDA topic classification visualization methods, this paper regularizes and integrates public opinions of hot events. Through this research, we can get the influence of public opinion mediators, public opinion objects, and government forces on the network public opinion and put forward corresponding improvement suggestions. We hope to contribute to the government's governance and prevention of online public opinion during the spread of COVID-19 and other public hot events.

Keywords: Online public opinion, LDA, co-occurrence knowledge graphs, visualization.

Introduction

The popularization of the Internet has not only expanded people's access to information, but also expanded the impact of the Internet public opinions

Journal of Web Engineering, Vol. 20_8, 2457–2472.
doi: 10.13052/jwe1540-9589.20811

based on the Internet information dissemination. This situation has aroused great attention from the Party Central Committee. The Fourth Plenary Session of the Seventeenth Central Committee of the Communist Party of China clearly put forward the proposal-"pay attention to the analysis of the impact of online public opinion on society" [1]. Zeng Runxi [2] (2009) defined Internet public opinion as a collection of all cognitive behaviors, attitudes, emotions, and behavioral tendencies of people generated by the stimulation of various events, which spread through the Internet. With the rapid development of information technology, mobile phones and computers are used as communication medium; information technology using WeChat, Weibo, QQ, etc. as communication platforms has gradually occupied the lives of modern people. These platforms are generating massive amounts of data every moment. After these traffic data are structured, semi-structured and unstructured, they form a massive and diversified information asset, which is big data. Based on the rapid development of information technology, big data has gradually become the focus of the business community, academia and government management. For network public opinion management, big data technology is not only a way to effectively manage and supervise network public opinion, but also an important technical path for the transformation of public opinion network management methods. In recent years, sudden and infectious public health problems have frequently occurred in various regions of China. And the COVID-19 incident that broke out early this year has become a concentrated outbreak event for netizens to pay attention to. The occurrence of this incident has great research value for the government to explore public opinion management based on the development of big data.

While the combination of big data and online public opinion has just started, AnupChakraborty and Sueli Dey [3] (2012) have analyzed the big data in the public sector of the Indian government, which may have a huge impact on the model of Indian Public Sector Management. Wendy N. Whitman Cobb [4] (2015) has discussed a timely and broad issue.:Public opinion issues and the prosperity of "big data". Lately, in China, Sun Wenbo and Yan Wujie in the "*Government Supervision Research on Internet Public Opinion Governance from the Perspective of Big Data*" pointed out:The government can realize the government's macro-control of public opinion by establishing an online public opinion big data supervision system. Wang Ying, Gong Huaping [6] (2017) have used big data crawler tools and text analysis techniques to analyze the sentiment trend and sentiment dimension state during the peak period of the public opinion event of "Nanchang University's independent cleaning", which researched and judged public opinion events in the network.

As an increasingly active social media platform in China, according to Sina's financial report during the fourth quarter of 2018, Weibo has 462 million active users. The number of active users in Weibo has increased by more than 70 million for three consecutive years, the trend of the use of APP has continued to increase, for users' enthusiasm for interaction is high [7]. These above conditions make Weibo one of the important platforms for the generation and dissemination of public opinion. Therefore, this article have used Weibo's hot search data as the main data source for public opinion analysis in this article, and have used big data technologies, such as crawlers and statistical analysis methods, to collect and aggregate data on hot search data that occurred on the Weibo network. This article aims to effectively monitor public opinion by further grasping the cognition of Weibo users on hot topics during the epidemic, so as to provide an important reference for the government to make decisions.

1 The Characteristics of Network Public Opinion Under Social Emergencies

Zeng Runxi [2] believes that compared with other forms of public opinion, online public opinion has its own characteristics such as complex contents, realistic interactions, and overall control labilities. Xu Xiaori believes that Internet public opinion has the characteristics of extensive sources and anonymity, the tendency of problem exposure and reality criticism, suddenness, group polarization, and the ability to form greater group pressure. Li Gang believes that Internet public opinion has the characteristics of suddenness, flood of information, amplification of harm, group polarization, and difficulty of control. According to the regulations of the State Council: public emergencies are "natural disasters, accidents, public health incidents, and social security incidents that occur suddenly, have caused or may cause serious social harm, and require emergency measures to deal with". Based on the existing research results, the article can summarize the characteristics of network public opinion under social emergencies:

(1) Suddenness: social emergencies spread extremely fast, which can erupt in a short period of time so that people's attention reaches its peak and can arouse strong social reactions;
(2) Dissemination: the rapid development of the Internet and social media promotes the dissemination of information to every class of society, such

as the development of the COVID-19 epidemic, the invention of the COVID-19 vaccine, and other events related to vital self-interests;

(3) Extreme tendency: because the information that the public can view is not really original information, but information that has been edited twice by editors, which makes the public easy to be misled by the edited text. After viewing the edited information, whether it is positive information or negative information, it is easy to form a subjective emotional public opinion tendency;

(4) Polymeric: online public opinion is closely related to emergencies. Many emergencies, as long as they include factors such as social stability and people's lives, can quickly arouse national enthusiasm, and may cause regional, local, and occasional issues to become public topics "onlookers" for the whole people. It requires the intervention of the central government.

2 To Establish a System for Analyzing Public Opinion on Weibo Network

The essence of Weibo network is the social network formed by the stable relationship established by Weibo users. The dissemination of information and resources in Weibo is carried out on this social network [8]. Based on big data technology, using existing public opinion analysis system flow chart (as shown in Figure 1), the article established and analyzed four analysis modules and determined the corresponding functions of each module, including the following four aspects:

(1) Data collection: There are a large number of hot searches on Weibo, so we need to use web crawlers to automatically crawl the specified data from the specified URL; using Python to crawl the keywords of the hot search events on Weibo Take, collecting its appearance time, time on the list, the top ranking volume and search volume;

(2) Preliminary processing of data: Cleaning and preliminary processing of the collected raw data, including deduplication of the raw data and word segmentation using Jieba; counting word frequency and characteristic classification and collection of word segmentation, providing convenience for later public opinion analysis;

(3) Data analysis: Analysis of topic clustering and identification, topic tracking,gathering hotspots of Weibo search, and text orientation on the preliminary processed data [9];

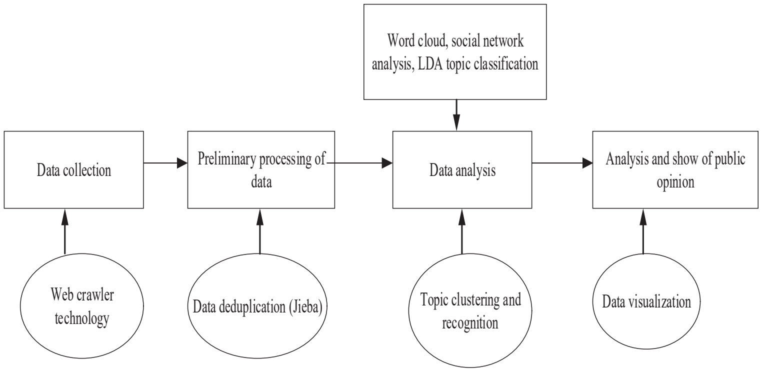
Word cloud, social network analysis, LDA topic classification
Data collection
Preliminary processing of data
Data analysis
Analysis and show of public opinion
Web crawler technology
Data deduplication (Jieba)
Topic clustering and recognition
Data visualization

Figure 1 Flow chart of Weibo's network public opinion analysis system.

(4) Analysis and show of public opinion: Using Gephi, Python's LDA topic classification technology to visualize hot topics and analyze their trends.

3 Data Collection and Preliminary Processing

Data Collection

From the occurrence and follow-up of the Weibo hot search event "The owner of a private car unclaimed for several months in Hubei has passed away", until the "rumor that a car mover was detained for 3 days due to the death of the new crow", we collected information and analyzed changes and characteristics of the Internet public opinion.

Using Python to collect data on Weibo hot search comments of a series of events, we got 8,350 comments on Weibo. After manually deleting invalid and repeated information in these data, we got 6,153 comments.

The cause of the incident was that Wang saw a car in the community that had not been removed by the owner for several days. He fabricated a rumor that the owner of a car was infected by the COVID-19 virus and died as a volunteer during the COVID-19 epidemic. This incident occurred during the COVID-19 outbreak, so it caused many media to spread and quickly boarded the hot search list of Weibo.

The information requirements of public opinion were to track the relevant public opinion information of this incident, analyze the public's sentiment on the incident, and deal with the subsequent development of the incident.

Results of public opinion events were that after a quick investigation and verification by the public security authorities of Xiaogan City, it was found

Table 1 The time table of the event

	Hot Search Keywords	Time of First Appearance	Total Time on the List (h)	Highest Ranking List	The Highest Search Volume on the Hot Search List
1	The owner of the unclaimed private car in Hubei Province has died	June 21, 18:08	14	2	2873176
2	It is a rumor that car owner in Hubei Province cannot move their car due to COVID-19 death	June 22 at 13:06	8.08	2	3325746
3	The rumors of the death of the car owner in Hubei Province meet with the car owner	June 23 at 08:27	8.4	15	479157
4	The car owner in Hubei Province who was unable to move the car due to COVID-19 was detained for 3 days	June 23 at 19:39	13.52	8	1016571

that the owner was in good health, this incident is a rumor. The rumors were detained for 3 days.

Sorting out public opinion hotspots: As shown in the timeline combing diagram of Table 1, this hot search appeared at 18:08 on June 21, 2020. On June 22, 2020, the public security department of Xiaogan City, Hubei Province immediately checked that the owner had received the news and the owner was not dead and dispelled the rumors. In accordance with the law, the rumors will be detained for 3 days. Figure 2 shows the trend chart of the search volume of hot searches on Weibo for this event, it can be seen from Figure 2 that the public opinion attention to this incident soared rapidly after the incident. After the police on the 22nd dispelled the rumors about the incident, the attention of Weibo users reached a peak. Since then, the popularity of the incident has declined slightly, and it has rebounded slightly when the police issued the punishment result.

3.1 Preliminary Processing of Data

The captured data is processed for deduplication, stop words and word segmentation. After the repeated data is deleted, stop words are processed for words that are frequently used in this article but are of no practical use

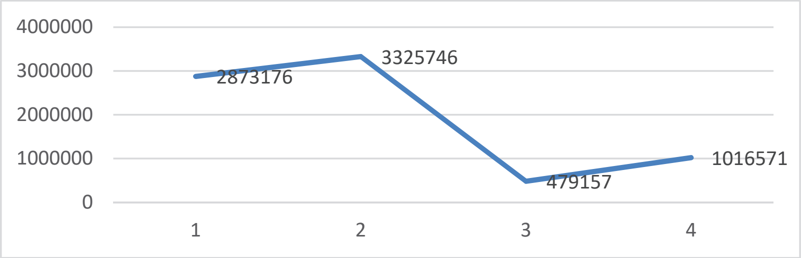

Figure 2 The Weibo search volume trends of the event.

Table 2 Word segmentation of Weibo comments (the part of whole table)

Serial Number	The Comments
1	Anti-epidemic warrior hero way farewell
2	Media Broke the news come out Rubbish The news Report Prior to Check Deny Fact No way Car owner Hear Pissed off
3	Suggestion Weibo Hot search Open Refute rumors Column
4	People Unharmed Severe Punishment Rumors
5	Now news Report Both Seeking truth from facts
6	Enthusiastic Citizen Pull out Really Anger

Table 3 High-frequency keywords

Key Words	Word Frequency	Key Words	Word Frequency
Spread rumors	1216	Fabricate	336
Media	672	Fake	258
News	588	Forward	256
Owner of the car	361	Verify	242
Death	337	Rumor	241

to actual data processing. Such words as “where” and “what” have no actual meaning. Finally, word segmentation is performed on the processed text. Get Weibo comment segmentation Table 2 and high-frequency topic thesaurus 3.

4 Public Opinion Analysis and Display Based on Topic Capture and Classification

4.1 Construction of Word Cloud and Co-Occurrence Network Knowledge Graph

In the traditional public opinion research work, it takes very complicated calculations to get visualization results of the data. However, the network public opinion analysis system in the era of big data has the advantages of storage space for massive data and processing of unstructured data. And combined with the technical architecture of the data analysis system [10], this technology greatly shortens the time to process data, reduces workload and improves work efficiency. But the public opinion mining and opinion analysis are still in the exploratory stage. Most of them draw on the technology and methods of collecting product reviews in the field of e-commerce to capture and gather topics and tendencies in the text.

According to the collected comment data to summarize the public opinion information of the incident, and based on the public appeals based on Weibo users as a template, suggestions for governance and prevention are put forward, in terms of public opinion on public emergencies.

In order to analyze and observe this incident more intuitively, using Python to perform word segmentation and word frequency statistics on Weibo comments to construct a common word matrix; Using R to perform word cloud processing on word frequency. The purpose is to be able to further discover the relationship between the status of high-frequency keywords in the entire event and the high-frequency keywords. The article takes the co-occurrence frequency as the weight, weights the nodes and edges, and uses Gephi software to realize the visualization of social network analysis spectrum [11]. Selecting the top 64 high-frequency keywords to get the word cloud diagram (Figure 3) and social network analysis diagram (Figure 4) of this event.

From the word cloud (Figure 3) and the social network analysis graph (Figure 4), high-frequency keywords appearing in Weibo comments, such

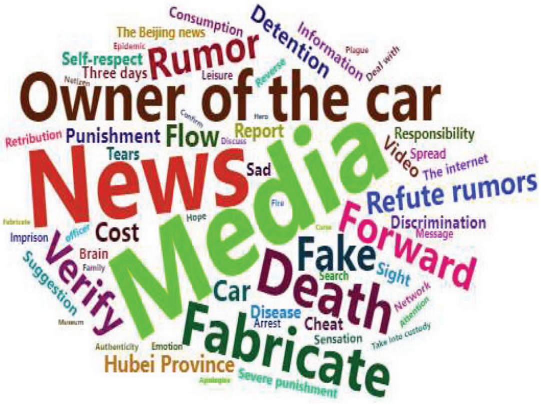

Figure 3 Word cloud.

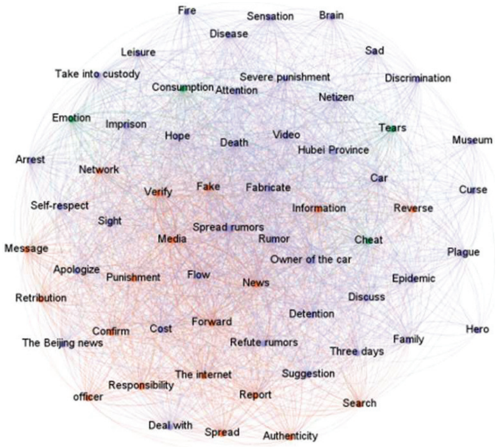

Figure 4 Social network analysis.

as "Spread rumors", "Rumor", "News", "Media", "Media", "Flow", "Fabricate", "Owner of the car", "Fake", "Verify", "Forward" and so on, they also occupy a central position in the social network analysis graph. This phenomenon shows that Weibo users are more concerned about the authenticity of news in this incident, and Weibo users have doubts about the authenticity of media reports.

4.2 LDA Topic Model Classification

LDA (Latent Dirichelt Allocation) topic model is a Bayesian topic model composed of document layer, subject layer and topic word layer. It can be used for the generation of document topics [12]. Using Python to classify the words extracted from the file through the LDA topic model. After the file is tested step by step, the optimal number of topics is 3, and finally the keywords of each topic are obtained. As shown in Table 4, 10 high-frequency words are selected for each topic for display.

In Xing Pengfei's research on online public opinion, the participants of online public opinion are divided into four elements: The subject of public opinion (participants of public opinion), Intermediary of public opinion (media of communication), The object of public opinion (events that occur), The power of the government. As can be seen from Table 4, the theme of public opinion intermediary mainly focuses on the media's influence on this incident. Key words such as "Flow", "Forward", "Fake news" and "Report" reflect the attention of Weibo users to the unconfirmed media reporting

Table 4 Display of high-frequency words in 3 topics

Theme	Key words
(1) Public opinion intermediary	Media, Forward, Verify, Flow, Fake news, Vedio, Report, Network, Spread, Sensation
(2) The subject of public opinion	News, Owner of the car, Self, Death, Hubei province, Suggestion, Discrimination, Deem, Reverse, Brain
(3) The power of the government	Spread rumors, Good intention, Refute rumors, Detention, Cost, Tears, Punishment, Severe punishment, Deal with, Too low

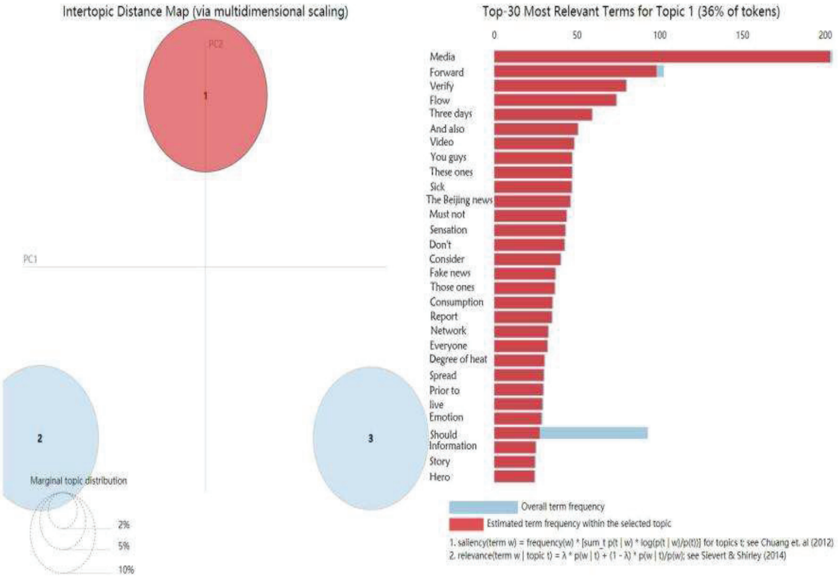

Figure 5 Visualization of LDA subject classification.

and dissemination of this incident. The theme of subject of public opinion mainly focused on Weibo users'views on other users' opinions on blindly following the trend and making comments regardless of whether the matter is true or not. The theme of power of the government mainly focuses on the level of handling rumors and spreading rumors. Key words such as "Punishment", "Detention", "Spread rumors" and "Deal with " reflect the attitudes of Weibo users towards those who want to spread rumors and the hope that the government will do something to handle the problems.

The effect of topic classification in this article is interactively visualized by the pyLDAvis software package in Python, and the results of the topic classification are shown in Figure 5. It can be seen from Figure 5 that the comments are three thematic circles, and there is no overlap between themes, indicating that the themes have a good cohesion effect. When the mouse moves to the range of topic 1, the keywords of the topic will be displayed in red bars, and the length indicates the frequency of the keywords of topic 1 [12]. It can be seen from Figure 5 that the word frequency of the three keywords “Media”, “Forward” and “Verification” are in the top 3 and their frequencyis relatively high. In theme 1, Weibo users paid more attention to the role of the media in the entire incident.

5 Conclusions and Recommendations

5.1 Conclusion

Network public opinion incidents as a social issue in the network society, it has the relevant attributes of sociology [11]. What happened this time is a microcosm of the development of online public opinion in modern society. This article uses Python, R language, Gephi and other tools to crawl data, display word cloud, analyzes co-occurrence social network, and make LDA topic classifications to study the Internet public opinion that occurred during the COVID-19 epidemic, and visualizes the results. Based on the research of the “The unclaimed owner of a private car in Hubei has passed away due to COVID-19 for several months”, the article researched the hot search list on the Weibo platform and the user’s evaluation. We extracted the people’s attitude towards rumors and their sensitivity to the series of consequences of the COVID-19 virus during the sudden public hot spot event. Combining the results of data analysis and the characteristics of online public opinion, we can draw the following conclusions:

(1) Intermediary of public opinion: Weibo, Tencent, Tik tok and other platforms have become the main channels for people to obtain information on a daily life, these platforms constitute the main channels for the spread of online public opinion and have major impact on the development of the network of public opinion. However, due to the incomplete information identification capabilities of the new media represented by the above-mentioned social media platforms, its users are suspicious of the published events.

(2) The subject of public opinion: Since the public may be extremely inclined towards the information they browse, especially when it is unfavorable or rumored, they often render and disseminate information again under their impulse, thereby intensifying the spread of false statements and videos.

(3) The power of the government: The government is the most important part of online public opinion governance. The efficiency and result of the government's handling of online public opinion play an important role in promoting the transformation of government functions, achieving democratic decision-making effect and establishing a good image of the government.

5.2 Recommendation

The Fourth Plenary Session of the 19th Central Committee of the Communist Party of China put forward the requirement of "improving and adhering to the correct guidance of public opinion guidance". It is required to "improve and innovate positive publicity, improve the public opinion supervision system, and improve the public opinion guidance mechanism for major public opinions and emergencies". Based on the above research conclusions, this article proposes the following three suggestions:

(1) Intermediary of public opinion: Social media platforms increase their ability to identify false information, and use big data technology to monitor online public opinion to reduce the formation and spread of online rumors.

(2) The subject of public opinion: Internet users should reduce or control their extreme tendency when browsing information in social media. When browsing information, Internet users should reduce or control the extreme tendency when browsing information in social media, increase scientific literacy, and do not believe in rumors or spread rumors without confirmation, internet users should avoid impulsive actions that promote the spread of rumors.

(3) The power of the government: The 19th National Congress of the Chinese Middle Class pointed out:"The government needs to increase credibility and execution, and build a service-oriented government that satisfies the people." The government needs to face up to its position in the online public opinion and at the same time strengthen cooperation with public opinion agencies such as social media. The government should keep abreast of public opinion information on public hot spot events in a timely manner, strengthen the notification of incidents, the publicity of event information, transparency, and response to hot spots on the Internet [13].

Acknowledgments

The authors would like to thank the project supported by Science and Technology Program of Guizhou Provincial Science and Technology Agency (Contract number:Qian KeHe Basic[2020]1Y156) and the major project supported by Guizhou University of Finance and Economics (Project approval number: 2019XZD02).

References

[1] Zeng Runxi, Xu Xiaolin. The Spread Law of Internet Public Opinion and Netizen Behavior: An Empirical Study [J]. Chinese Administration, 2010(11): 16–20.

[2] Zeng Runxi. Research on the Working Mechanism of Network Public Opinion Control [J]. Library and Information Service, 2009, 53(18): 79–82.

[3] Anup Chakraborty, Sueli Dey. Big Impact through Big Data: Potential of Big Data in the Indian Public Sector [C]. International Conference on Public Administration,2012, 25th, October. Hyderabad, India

[4] Wendy N. Whitman Cobb. Trending now: Using big data to examine public opinion of space policy [J]. Space Policy, Volume 32, May 2015, Pages 11–16.

[5] Sun Wenbo, Yan Wujie. Research on Government Supervision of Internet Public Opinion Governance from the Perspective of Big Data [J]. Journal of Changsha University, 2017(1): 41–45

[6] Wang Ying, Gong Huaping. Big Data Based on Emotional Dimensions Analysis and Research on the Emotional Tendency of Public Opinions in Data Networks-Taking "Nanchang University's Independent Cleaning" Weibo Public Opinion Events as an Example [J]. Information Science, 2017, 35(04): 37–42.

[7] https://data.weibo.com/report/reportDetail?id=433&display=0&retcode =6102

[8] Peng Hao, Zhou Jie, Zhou Hao, Zhao Dandan. Public opinion analysis based on topic discovery in Weibo network [J]. Telecommunications Technology, 2015, 55(06): 611–617.

[9] Ma Mei, Liu Dongsu, Li Hui. Research on network public opinion analysis system model based on big data [J]. Information Science, 2016, 34(03): 25–28+33.

[10] Xia Sheng. Network public opinion analysis system in the era of big data [J]. Electronic Technology and Software Engineering, 2016(17): 187.

[11] Mou Dongmei, Shao Qi, Han Nannan, Wang Ping, Jin Shan, Jin Chunyan. Multi-dimensional social attribute analysis and visualization of public opinion on Weibo-Taking a vaccine incident as an example [J]. Library and Information Service, 2020, 64(03): 111–118.

[12] Xing Pengfei, Li Xinxin. Research on the formation mechanism and guiding strategy of online public opinion in the prevention and control of major epidemics-Based on the qualitative analysis of the text of online public opinion during the new crown pneumonia epidemic [J]. Journal of Information, 2020, 39(07): 67–74+15.

[13] Li Wanlian, Gao Guanghan. Research on the Generation Mechanism of Internet Public Opinion Heat in Public Emergencies-A Qualitative Comparative Analysis of Fuzzy Sets Based on 48 Cases (fsQCA) [J]. Journal of Information, 2020, 39(07): 94–100.

Biographies

Mei Zhang was born in Guanzhou Village, Tong Ren; China in 1974. She received her master's degree from the Department of Computer Science at Guizhou University, in 2002 and the Ph.D. degree in Photogrammetry and Remote Sensing from Wuhan University, CHINA, in 2008.

From 2006 to 2010, she is an associate professor of computer science at the information institute of guizhou institute of finance and economics. She was promoted to professor of computer science in 2010. In recent years, she published three monographs in science publishing house, published more than 50 papers, among which sci, ei, istp included more than 30 as the first author.She has published extensively in various research areas of Computer Science, such as computer vision, data mining, information safety, green data center, e-commerce, and three-dimensional reconstruction.

Prof. Mei Zhang won the third prize in natural science of Guizhou province in 2019, ranking first. In May 2017, she was awarded the Third Prize for Advanced Individuals in Scientific Research at Guizhou University of Finance and Economics in 2016. Her paper "the realization of pb7.0 general arbitrary field query technology", published as an independent author, won the third prize of the first excellent academic paper of natural science in guizhou province in 2006 "mobile communication cup".

Huihui Su was born on Marchr 28, 1994 in Zhongzhao Township, Neihuang County, Anyang City, Henan Province. She received a bachelor's degree in accounting from Cheng Yi College, Jimei University in July 2018.

She studied for a master's degree in Library and Information from 2018 to 2021 at the School of Information, Guizhou University of Finance and Economics. She is a master candidate in Guizhou University of Finance and Economics. Her major is Library and Information. His research interests include Financial data analysis, Hotspot analysis of rural inclusive finance, e-commerce, big data, library and information science.

Jinghua Wen was born in RenHe Village, Tong Ren; China in 1975. He received his master's degree from the Institute of Computer Software and Theory at Guizhou University, in 2002 and the Ph.D. degree in Computer Software and Theory from Guizhou University, CHINA, in 2006.

From 2003 to 2007, he was a lecturer in computer science at the information institute of Guizhou institute of finance and economics. He was promoted to professor of computer science in 2007. He published three monographs in science publishing house, published more than 50 papers, among which sci, ei, istp included more than 30 as the first author or a second author or correspondence author. His research interests include computer software and theory, Computer vision, information safety, big data, e-commerce and Cloud computing.

Prof. Jinghua Wen won the third prize in natural science of Guizhou province in 2019, ranking second. He won the second prize for scientific and technological progress in Guizhou province in 2013, ranking second. He won the first Youth Innovation Talent Award in Guizhou Province in 209 years, ranking first.