Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 221 (2023) 330–337

Procedia Computer Science

www.elsevier.com/locate/procedia

10th International Conference on Information Technology and Quantitative Management

Risk Perception of the " Belt and Road" Countries Based on Global Media Data GDELT

Yijun Liu[a,b], Yunrui Zhang[a,b], Ning Ma[a,b], Qianqian Li[a,b]*

[a]*Institutes of Science and Development, Chinese Academy of Sciences, Beijing 100190, China*
[b]*School of Public Policy and Management, University of Chinese Academy of Sciences, Beijing 100049, China*

Abstract

Rapid and dynamic perception of risks in countries along the “Belt and Road” is of great practical significance for the construction of the “Belt and Road”. The current measurements of country risk are divided into two categories: multi-factor analysis and systematic risk modeling based on capital asset pricing theory. There are problems such as time lag in data updates and inadequate completeness. Risk perception based on big data has the characteristics of wide sources, high timeliness, multiple dimensions, and full coverage, and it can capture potential risk variations earlier and faster. In this study, based on the global media big data GDELT, it is found that the risks of the countries along the “Belt and Road” are mainly focused on politics, military risks, energy trade, terrorism, power struggles, etc. West Asia and North Africa region are at the core of the risk reports in the network of countries mentioned, with the Central and Eastern Europe and Central Asia region playing the role of "bridge" nodes. In the “country-topic risk” heterogeneous information network, when the risk topic similarity was set to 0.7, the country risk clustering effect is the best. Syria had always been at high risk. The risk of countries in West Asia and North Africa, such as Afghanistan, Iran, Israel, and Lebanon is also at high risk but slightly fluctuated from year to year. The research results show that the classification of national risks by media big data has strong consistency with existing national risk ratings, so this article proposes to use media big data to enhance the risk perception capabilities of countries along the “Belt and Road”.

Keywords: Big data, GDELT, Country risk, Risk Perception;

* Corresponding author. Tel.: 010-59358717;
E-mail address: lqqcindy@casisd.cn

10.1016/j.procs.2023.07.045

1. Introduction

In September 2013, China proposed to jointly build the "Silk Road Economic Belt" for the first time. In October of the same year, China proposed to jointly build the 21st century "maritime Silk Road" (hereinafter referred to as the "Belt and Road" initiative). The “Belt and Road” initiative aims to promote the orderly and free flow of economic factors, the efficient allocation of resources, and the deep integration of markets. However, due to the huge differences in the domestic situation, geopolitics, geographical location, economic development level, and religious culture of countries along the "Belt and Road", the construction process of the "Belt and Road" has brought severe challenges, and some projects have been shelved, canceled or postponed[1]. By analyzing and evaluating the risks of countries along the "Belt and Road", we can effectively prevent and resolve the risks and avoid losses caused by country risks[2] .

At present, the relevant quantitative researches on the risk perception of countries along the "Belt and Road" mostly use the statistical yearbook data, design the index system, use the statistical analysis method, and focus on the political risk, economic risk, financial risk, trade risk, sovereign credit risk, social risk, project risk, etc[3-7]. In addition, international well-known credit rating companies such as Standard & Poor, Moody's , Fitch as well as domestic Dagong international credit rating Co., Ltd. and China Export Credit Insurance Corporation also carry out risk ratings for various countries. The credit rating mainly carries out a comprehensive assessment of the ability and willingness of sovereign governments to repay debts in full and on time[8,9].

In recent years, with the accumulation of network big data and the maturity of data analysis technology, some studies have noticed the decision value of media report information in risk mining. By directly collecting first-hand data, media big data has the characteristics of wide sources, high timeliness, multiple dimensions and full coverage. It can objectively reflect the hot spots and changes of social concern, and it can sense the social risk situation more comprehensively and accurately as a senor of social risk[10]. However, due to the fragmentation of information in social media, it is difficult to distinguish information classification at the national level[11]. GDELT, Google's free global news dynamic database（ http://www.gdeltproject.org ）is an important data source for analyzing national risks from the perspective of media reports[12,13]. Noam Levin et al. used GDELT media event information mining, social media photo data and lighting data to analyze the conflict intensity of the "Arab Spring" movement in Arab countries[14]. It can be seen that media big data can reflect the potential risks beyond the design of the "indicator system" from bottom to top, and the completeness of data is better. The high-frequency characteristics of data also provide the possibility of rapid and dynamic perception of risk changes.

From the perspective of global media coverage for the first time, this paper uses word2vec and complex network methods to analyze the risk topics and structural characteristics of associated countries along the "Belt and Road" by using global media data from 2014 to 2020. Further, the country risk topic and risk correlation are fused into a "country-risk topic" heterogeneous information network, and the heterogeneous information network node representation algorithm metapath2vec is used to calculate the country word vector representation and cluster the country risk. Finally, the policy implications of media big data for risk monitoring in countries along the "Belt and Road" are proposed.

2. Data and Modeling

2.1. Data source

GDELT is a free global news dynamic database opened by Google in 2013. It monitors broadcast, prints and online news data in more than 100 languages around the world in real time, converts multilingual news into English by using machine translation technology, and extracts news events from it. Its update frequency is every 15 minutes per time[13] . Based on the global media data information, GDELT has constructed the Global Knowledge Graph (GKG), which contains the topics of news reports and characterizes the concepts involved in risk. The risk topics include the topic table developed by the world bank and GDELT.

GDELT issues the Daily Conflict Trends Report every day, compares the real conflicts within 48 hours worldwide with the previous 48 hours and generates a global map of conflict trends. It then extracts the top 10 countries with the largest increase in real conflicts and lists the risk topics and the regions/countries mentioned in the report. A total of 65 countries along the "Belt and Road" analyzed in this paper are shown in Table 1.

Table 1. Regional distributions and names of 65 countries along the “Belt and Road”

Regin	Country
Northeast Asia (2 countries)	Mongolia, Russia
Southeast Asia (11 countries)	Singapore, Indonesia, Malaysia, Thailand, Vietnam, Philippines, Cambodia, Myanmar, Laos, Brunei, Timor Leste
South Asia (7 countries)	India, Pakistan, Sri Lanka, Bangladesh, Nepal, Maldives, Bhutan
West Asia and North Africa (21 countries)	Greece, UAE, Kuwait, Turkey, Qatar, Oman, Lebanon, Saudi Arabia, Bahrain, Israel, Yemen, Egypt, Iran, Jordan, Syria, Iraq, Afghanistan, Palestine, Azerbaijan, Georgia, Armenia
Central and Eastern Europe (19 countries)	Poland, Albania, Estonia, Lithuania, Slovenia, Bulgaria, Czech Republic, Hungary, Macedonia, Serbia, Romania, Slovakia, Croatia, Latvia, Bosnia and Herzegovina, Montenegro, Ukraine, Belarus, Moldova
Central Asia (5 countries)	Kazakhstan, Kyrgyzstan, Turkmenistan, Tajikistan, Republic of Uzbekistan

2.2. Research methods

The research process of this paper is shown in Figure 1. First, the Daily Conflict Trend Reports from 2014 to 2020 are collected from the GDELT database, with a total of 2465 days. A total of 171 countries in the world are listed on the report, and the countries along the "Belt and Road" account for nearly 40%. Based on the risk topics data, the risk topics and the country risks are clustered, and the change trends of the risk topics in the countries along the "Belt and Road" and in different regions (Northeast Asia, Southeast Asia, South Asia, West Asia, North Africa, Central and Eastern Europe, Central Asia) are observed. Then, based on the geographically linked data, this paper constructs country networks which are mentioned from 2014 to 2020, calculates the network structure measurement indicators, compares the structural differences of country networks in different years and determines the key countries in the media coverage risk through the indicators of clustering coefficient, degree and closeness. Finally, the "country-risk topic" heterogeneous information network is constructed by fusing the risk topics data and geographical correlation data. Through the heterogeneous information network node representation algorithm, the fusion node representation of country risk topics and reference information is formed, and the country risk is clustered.

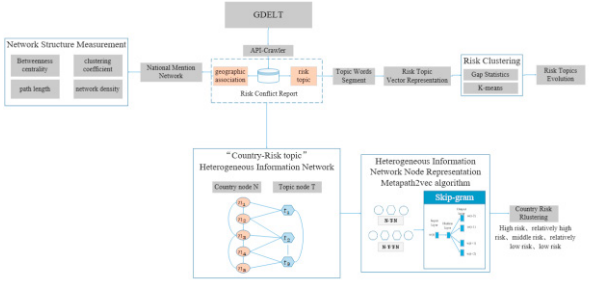


Fig. 1. Research flow chart

2.2.1 risk topic clustering based on word vector

This paper uses the word vector method to mine the risk topic semantics, and constructs the word vector of word2vec[15] after adding the risk topic information to the 100 dimensional word vector pre trained by GloVe[16]. According to the word vector representation of each topic, the similarity of any two topics can be calculated. Cosine similarity is usually used for measurement:

$$Sim_{i,j}=\frac{\sum_{k=1}^{n} theme_i \times theme_j}{\sqrt{\sum_{k=1}^{n}(theme_i)^2}\times\sqrt{\sum_{k=1}^{n}(theme_j)^2}} \quad (1)$$

Where theme_i is the vector representation of the topic i and theme_j is the vector representation of the topic j.

This paper uses k-means method to cluster risk topics. Firstly, Gap Statistic is used to determine the number of clusters. Then, the Bootstrap method is used to calculate the optimal number of clusters.

2.2.2 risk correlation network measurement indicators

According to the situation that the countries listed in Figure 1 (b) refer to other countries, this paper can build a national risk correlation network. In the network structure diagram $G = (V, E)$, the node set V is the countries along the "Belt and Road", and the edge set E is the reference relationship between countries. Thus, we can construct the national risk correlation network from 2014 to 2020 and calculate the statistical structure measurement of the network.

- Clustering Coefficient

Clustering coefficient is used to describe the degree to which vertices in a graph gather into clusters. The greater the clustering coefficient of a graph, the higher the tightness of the network connection.

- Eigenvector Centrality

Eigenvector centrality is used to measure the transmission influence and connectivity between nodes. It believes that the contribution of connecting with nodes with high scores is greater than those with low scores.

Let $A = (a_{i,j})$ be the adjacency matrix of a graph, then there is $x_i = \frac{1}{\lambda}\sum_k a_{ki}x_k$, where $\lambda \neq 0$ is a constant. The matrix representation is $\lambda x = xA$. Therefore, the center vector x is the left eigenvector of the adjacency matrix A associated with the eigenvalue λ. We select the largest eigenvalue λ among the absolute values of the matrix A.

- Closeness Centrality

Closeness centrality is the reciprocal of the total distance between a node and other nodes. The high closeness centrality of a node indicates that its path to all other nodes is the shortest.

$$C(x) = \frac{1}{\sum_y d(y,x)} \quad (2)$$

2.2.3 country risk clustering based on "country-risk topic" heterogeneous information network

In order to analyze the correlation and coupling analysis of country information and topic information, a "country-risk topic" heterogeneous information network is constructed. The nodes in this network are divided into two types: country and node. The edge relationships in the network can also be divided into three types: (1) the connection between country nodes is the mentioned relationship in the report; (2) The connection between risk topic nodes is the topic similarity relationship. (3) The connection between the country node and the topic node is the topic involved in reporting the country. On the topology of the "country-risk topic" heterogeneous information network, metapath2vec[17], a random walk method based on meta path, is used to systematically associate the country information, risk topic information and associated information between countries and topics involved in media reports, and then two meta paths are formed (Figure 2). Further, the k-means algorithm is used to cluster the country risk topics in different years, and then the clustering results of country risk under different topic similarity edge thresholds are compared (σ=0.6,0.7,0.8,0.9). In order to compare the changes in country risk topics based on media big data in different years, this paper uses Rand Index to calculate the consistency of country risk topics clustering results of different years.

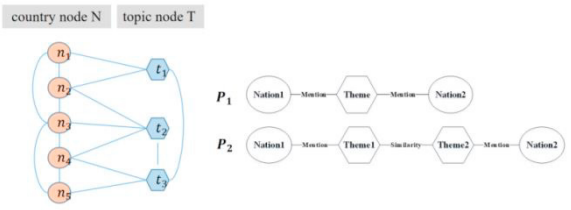


Fig.2. Schematic diagram of "country-risk topic" heterogeneous information network

3. Research findings

3.1. Risk topics and their evolution in media reports of countries along the "Belt and Road"

Clustering all risk themes from 2014 to 2020, combined with Gap Statistics calculation results and expert judgment, it can be found that the risk topics of countries along the "Belt and Road" are mainly concentrated in politics, military, politician, energy trade, terrorism, and power struggle. The minor focus is on al-Qaida, health, disaster, religion and so on (Figure 3).

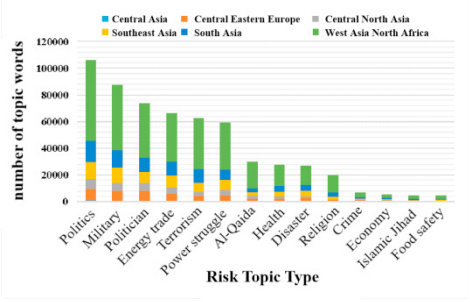
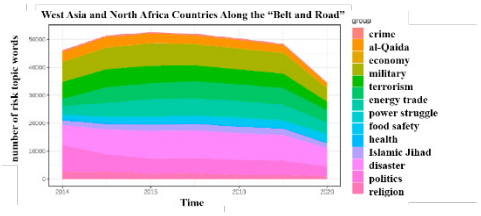


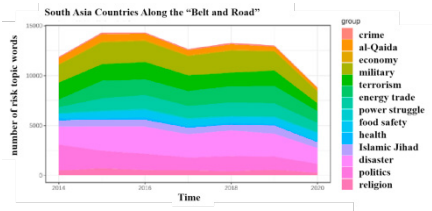
Fig.3. Regional distribution of risk topic types of countries along the "Belt and Road"

From the perspective of different regions, the total amount of risk topics in West Asia and North Africa and South Asia remains high. Risks in West Asia and North Africa are highly complex and remain high, with terrorism, security issues, and Islam being the main risk topics (Figure 4(a)). Among them, Israel, Syria, Afghanistan, and Iraq have been on the list more than 700 times. The risk topics in South Asia continue to be high, and various security risks and energy trade risks are prominent (Figure 4(b)). The countries involved are mainly Pakistan, India, and Bangladesh, which were listed 654, 666, and 334 times in conflict reports respectively.

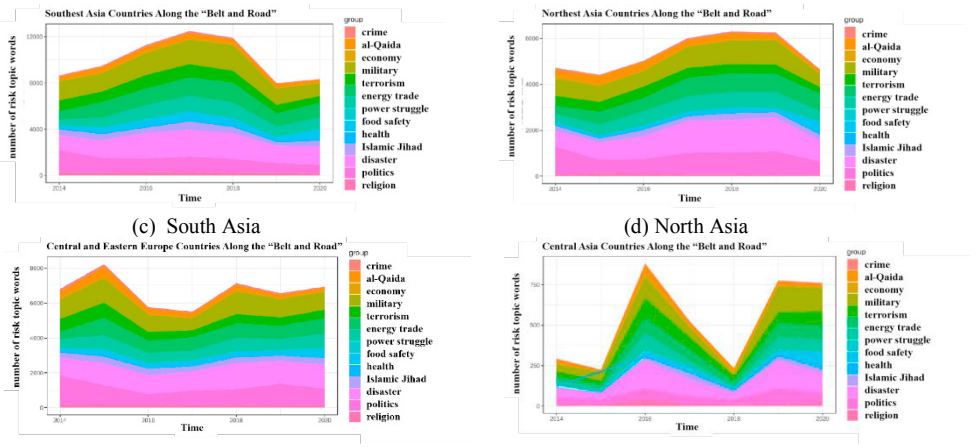
The total number of risk topics in other regions is relatively small, and fluctuations are highly correlated with the frequency of countries on the list. The country risk topics in Southeast Asia are in an inverted "U" shape, with prominent military and political risks (Figure 4(c)). The countries involved are mainly the Philippines, Indonesia, and Thailand. Indonesia and Thailand mainly face the risk of terrorism and domestic separatist forces. The risks in Northeast Asia are on the rise, and politics are the main topic (Figure 4(d)). The risks in Northeast Asia mainly come from Russia. From 2014 to 2020, Northeast Asia was on the list a total of 774 times, and Russia was on the list 772 times. Risks in Central and Eastern Europe show a downward trend and then an upward trend, with prominent political risks (Figure 4(e)). Health risks increase significantly in 2020. The scale of risk words in Central and Eastern Europe in 2020 has exceeded that of 2019. Although the risk topics in Central Asia fluctuate the most, the risks are relatively minimal (Figure 4(f)). There were peaks in 2016, 2019 and 2020, mainly triggered by politics and military affairs.



(a) West Asia and North Africa



(b) South Asia



(c) South Asia (d) North Asia

(e) Central and Eastern Europe (f) Central Asia

Fig.4. Risk topic evolution of different regions of countries along the "Belt and Road"

3.2. Analysis of country risk correlation networks in the global media reporting framework

Based on the risk reference information in the conflict report, the risk reference network of countries along the "Belt and Road" from 2014 to 2020 is constructed, and the basic statistical indicators of the network are calculated, including the number of nodes, the number of edges, the average closeness centrality, the average compactness, the average path length, the average path length, the average clustering coefficient and the network density (Table 2). According to the basic statistical characteristics of the country mentioned network, the number of nodes in the network remains stable, basically covering all countries along the "Belt and Road", and only Palestine is absent in 2020. The number of edges between countries reaches the maximum in 2016, indicating that the country mention relations reach the most. Overall, 60% of the countries along the "Belt and Road" have mentioned relations. In 2020, the average degree of the network is significantly higher than that of other years, mainly due to the global spread of the "COVID-19" epidemic. In 2018, the average compactness and clustering coefficient of the network are the largest, indicating that the "small world" effect of country-mentioning relationships in the network is more obvious.

Table 2 measurement indicators of country risk reference network from 2014 to 2020

Statistical index measurement	2014	2015	2016	2017	2018	2019	2020
Number of countries	65	65	65	65	65	65	64
Number of connected edges	1196	1228	**1241**	1156	1239	1198	1217
Average closeness centrality	0.411	0.414	0.408	0.411	**0.427**	0.421	0.419
Average path length	1.425	1.410	1.404	**1.444**	1.405	1.424	1.397
Average clustering coefficient	0.719	0.729	0.740	0.711	**0.744**	0.719	0.741
Network density	0.575	0.590	0.597	0.556	0.596	0.576	**0.604**

In general, in the country mention network, the measurement dimensions of eigenvector centrality, closeness centrality and clustering coefficient are not the same (Figure 5). From the perspective of eigenvector centrality, countries in West Asia and North Africa are at the core of risk reporting. Countries in West Asia and North Africa are mentioned the most times on the Internet due to their high frequency on the list. Countries related to Syria, Iran and Israel in the report have a relatively high-risk frequency. In 2020, the eigenvector centrality of Russia in Northeast Asia and Kazakhstan in Central Asia increased. The eigenvector centrality is always low mainly in East Timor, Maldives, Macedonia and so on which are in Southeast Asia, South Asia, and Central and Eastern Europe. From the perspective of closeness centrality, some countries in Central and Eastern Europe and Central Asia act as "bridges" in the country risk mention network and have more direct mention relationships with other countries. For example, countries such as the Slovak Republic, Serbia, and Albania which are in Central and Eastern Europe and Kazakhstan, Uzbekistan which are in Central Asia have high closeness centrality. From the perspective of clustering coefficients,

South Asian and Southeast Asian countries have a higher proportion of clustering coefficients above 0.9, which means the frequency of risks mentioned by countries is more intensive. On the contrary, most countries in West Asia and North Africa have lower clustering coefficients. Bhutan and Maldives in South Asia and East Timor and Laos in Southeast Asia have always maintained a high-density mutual mention relationship with other countries. In addition, in 2020, the clustering coefficients of the Slovak Republic in Central and Eastern Europe and Oman in West Asia and North Africa also increase.

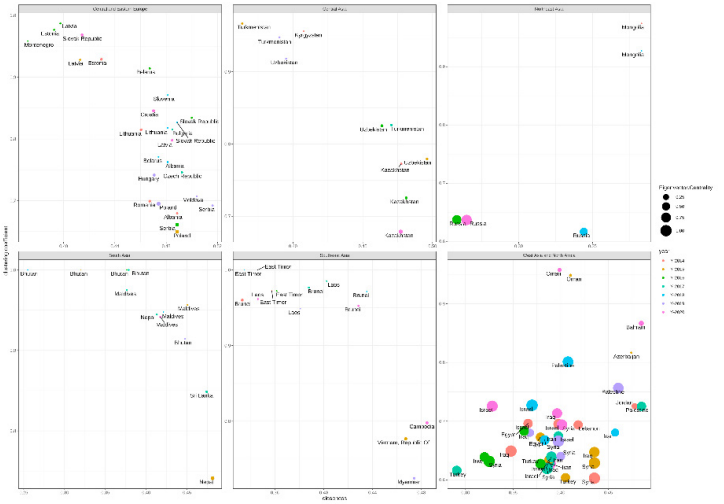


Fig.5. Distribution map of eigenvector centrality, closeness centrality and clustering coefficient of country mention network

3.3 Country risk classification based on "country-risk topic" heterogeneous information network

By constructing a "country-risk topic" heterogeneous information network, the relationship between country-topic reports and country-risk mention relations is correlated and coupled. Further, through the clustering algorithm, the country risks from 2014 to 2020 are clustered from the perspective of media reports. In this paper, the country risk is divided into five categories: high risk, relatively high risk, medium risk, relatively low risk and low risk. This paper analyzes the differences in country risk clustering results in different years under the scenarios of different similarity thresholds $\sigma = 0.6$、0.7、0.8、0.9. In general, under different parameter scenarios, the average Rand Index of country risk clustering results is about 0.7. Compared with other cases, when the risk topic similarity threshold is 0.7, the average value of Rand Index is higher and the fluctuation range is the smallest, so the classification result is the best when the topic similarity (σ) is 0.7. Based on the specific clustering Rand Index for topic similarity, the country risk clustering in 2014 has the highest consistency with the country risk clustering in other years, so we taking the country risk clustering result in 2014 as an example, the results of high risk and relatively high risk are as follows:

Table 3 Country risk classification based on media reports (2014)

Risk level	Country name
High risk	Saudi Arabia, Jordan, Ukraine, Qatar, Russia, Bahrain, Afghanistan, Egypt, Iran, Iraq, Israel, Lebanon, Pakistan, Syria, Turkey, Yemen, India, Philippines
Relatively high risk	Vietnam, Nepal, Brunei, Indonesia, Myanmar, Armenia, Thailand, United Arab Emirates, Greece, Croatia, Serbia, Romania

4. Conclusions

The concealment of risk increases the difficulty of risk perception[18]. By directly collecting "first-hand data", media big data has the characteristics of wide sources, high timeliness, multiple dimensions, and complete coverage. It can objectively reflect social concerns and changes. Using it as a social risk sensor can assess risks more comprehensively and accurately. This paper analyzes GDELT media big data through natural language processing and complex network analysis technology. It finds that countries along the “Belt and Road” have a large number of risks on the list, and

most risk topics focus on politics risks, military risks and energy trade risks. Countries in West Asia and North Africa are at the core of risk reporting, with a consistently high-risk frequency. Countries in Central and Eastern Europe and Central Asia also have more connections to other countries in their risk-reporting networks. Finally, through risk clustering, these countries are divided into five risk levels.

The research in this paper shows that national risk management and decision-making driven by media big data can realize rapid perception from high-frequency data, deep customization from data mining, and cross-form data decision support from the fusion of heterogeneous information. In order to enhance the risk perception ability of countries along the "Belt and Road", this paper believes that the top-down traditional national risk measurement of political stability, economy and finance, social development, and business environment should be combined with the bottom-up risks of media big data. Form a joint force to strengthen the top-level design of the risk media big data early warning system of countries along the "Belt and Road"; at the same time, related departments should establish a 24-hour all-media information monitoring platform for countries along the "Belt and Road to strengthen risk calculation capabilities; a regular daily newspaper for "Belt and Road" risk research and judgment mechanism should also be established to strengthen risk prediction and early warning capabilities.

Acknowledgment

The authors acknowledge financial support from the National Natural Science Foundation of Beijing (No. 9222030), National Natural Science Foundation of China (No. 72074205，71774154，T2293772)

References

[1] Wang L G (2020) " Risks, Causes and Implications of The 'Belt and Road' Initiative(BRI) Process." (in Chinese) *Journal of China Emergency Management Science* **7**:58-68.
[2] Hu J C, Wang D D (2016) " The Research of China's Banking Regulation in the Concept of Finacial Inclusion." (in Chinese) *On Economic Problems* **5**:1-6+43
[3] Li B, Yan X C (2018) " China's New Comparative Advantage of Trading with Belt and Road Countries: Public Security Perspectives." (in Chinese) *Economic Research Journal* **53(1)**:183-197.
[4] Zhao M Y, Dong S C, Wang Z, et al. (2016) " Assessment of Countries' Security Situation along the Belt and Road and Countermeasures." (in Chinese) *Bulletin of Chinese Academy of Sciences* **31(6)**:689-696.
[5] Yang G, Huang X, Huang J, et al (2020)" Assessment of the effects of infrastructure investment under the belt and road initiative." *China Economic Review* **60**: 101418.
[6] Yuan Y, Liu Y, Wei G (2017)" Exploring inter-country connection in mass media: A case study of China." *Computers Environment and Urban Systems* **62**: 86–96.
[7] Shang B (2020)" Impact of Terrorism on Bilateral Trade Between China and Five Central Asian Counteries: Based on an Expended Gravity Model . "*Transformations in Business & Economics* **19(3C)**: 431–447.
[8] Liu H M, Hu S L, Fang K, et al (2019)." A comprehensive assessment of political, economic and social risks and their prevention for the countries along the Belt and Road." (in Chinese) *Geographical Research* **38(12)**:2966-2984.
[9] Wang Z, Yang J. "Study on Risk Rating of 'The Belt & Road Initiative' Countries." (in Chinese) *Journal of Beijing Technology and Business University(Social Sciences)* **33(4)**:117-126.
[10] Lin K-P, Hung K-C, Lin Y-T, et al (2018) " Green Suppliers Performance Evaluation in Belt and Road Using Fuzzy Weighted Average with Social Media Information. " *Sustainability* **10(1)**: 5.
[11] Sun Z, Zhao H, Wang Z S (2021) "Analysis on the Association and Evolution Path of Internet Public Opinion." (in Chinese) *Library and Information Service* **65(7)**:143-154.
[12] Ward M D, Beger A, Cutler J, et al (2013) "Comparing GDELT and ICEWS Event Data." *Analysis* **21(1)**: 267-97.
[13] Leetaru K, Schrodt P A (2013) " GDELT: Global Data on Events, Location and Tone,1979-2012." *in ISA annual convention*: Citeseer
[14] Levin N, Ali S, Crandall D (2018) "Utilizing remote sensing and big data to quantify conflict intensity: The Arab Spring as a case study" *Applied Geography* **94**: 1–17.
[15] Mikolov T, Sutskever I, Chen K, et al (2013) " Distributed Representations of Words and Phrases and their Compositionality." *Advances in neural information processing systems* **26**: 3111-3119.
[16] Pennington J, Socher R, Manning C (2014) " Glove: Global Vectors for Word Representation." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*: 1532–1543.
[17] Dong Y, Chawla N V, Swami A (2017) "metapath2vec: Scalable Representation Learning for Heterogeneous Networks. " *Kdd'17: Proceedings of the 23rd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*: 135–144.
[18] Huang Y S, Wang Y B (2020) " Governance Perspective and China Program of Global Risk Society." (in Chinese) *China Soft Science* **8**:10-19.