Research on CO-word network topic mining and topic differences based on haze microblog data

Haiyue Lu\textsuperscript{1,a}, Xiaoping Rui\textsuperscript{2,b,*} and Runkui Li\textsuperscript{3,c,*}

\textsuperscript{1} College of Hydrology and Water Resources, Hohai University, Nanjing, Jiangsu, 211100, China

\textsuperscript{2} School of Earth Sciences and Engineering, Hohai University, Nanjing, Jiangsu, 211100, China

\textsuperscript{3} College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, 100049, China

\textsuperscript{a} email: 191309010002@hhu.edu.cn

\textsuperscript{b} email: ruixp@hhu.edu.cn, \textsuperscript{c} email: lirk@ucas.ac.cn

\textbf{Abstract:} Some studies have shown that haze not only poses a threat to people's health, but also affects the secretion of human hormones, making people depressed and endangering mental health. Microblog has the advantages of short content, rapid communication and convenient interaction. When the haze comes, a large number of topic microblogs related to the haze will be generated. Mining the topics of concern and psychological reactions contained in these microblogs is helpful for resource allocation and public opinion publicity in the case of haze. At present, the research of microblog topic mining in haze situation only involves a single research area, and few studies discuss the spatial differences of different regions. Based on this, this study collected the microblog data of seven provincial capitals in the severe haze areas in 2017, and used the community-based co-word network method to complete a series of experimental steps, such as keyword extraction, co-occurrence matrix construction, co-word network construction and topic community detection. On this basis, we detected the topic community in the microblog data set, and analyzed the horizontal differences of topics in different cities. The results show that different cities have not only the same but also different concerns about haze. The results can provide theoretical guidance for the healthy development of cities.

1. Introduction

Haze is a kind of weather phenomenon that the horizontal visibility of air is less than 10km due to the dust, moisture, smoke and water vapor in the air \cite{1-2}. It is originated from PM2.5, which is the particulate matter in the air whose aerodynamic equivalent diameter is less than or equal to 2.5 microns \cite{3}. Currently, although haze is better than in previous years, it still occurs from time to time in the northern region, which has a very serious threat to social life, economic development, especially to people's physical and mental health. Studies have shown that when the concentration of air pollutants, especially PM2.5, is too high, people living in this kind of polluted air will have acute health risks as a result, and then induce cardiovascular diseases \cite{4-5}; in addition, air pollution is not only related with a variety of human diseases, but also have a more significant impact on human psychology. GU et al.\cite{6} analyzed the effects of air pollution on the mental health by using a variety of econometric methods, and
found that the higher the PM2.5 concentration, the more prominent the four negative emotions of humans: tension, depression, weakness, and irritability.

At internet age, people are increasingly inclined to express their views on social platforms. Weibo has the characteristics of openness, rapid propagation and strong real-time, which makes it have multi-source users and multi-topic content. People of different ages and social nature can publish content on different topics according to their needs, which is a good data source for conducting various public opinion research [7]. When the haze pollution is serious, people will post, forward, and comment on related microblogs on the Weibo platform. The content of these microblogs contains a lot of opinions and information about the haze pollution. Different cities are affected by the haze at different degrees, so people’s attitudes to haze are also different. This paper takes typical cities in the north which are seriously polluted by haze as examples to collect and analyze Weibo data. The purpose of mining the topic information is to provide theoretical guidance for urban network public opinion guidance and environmental protection policymaking.

At present, there are a large number of domestic and foreign studies using Weibo data to discuss and analyze the concerns of netizens when the haze occurs. Yang et al.[8] applied the framework theory to analyze the text of Weibo, and found that China's official media for the haze concern focused on five aspects: the Government's concern, public opinion dissuasion management, public opinion factors, social haze-related news and external haze-related news; Zhang et al.[9] used Weibo data to analyze the seasonal differences in people’s perception of haze and found that people’s concerns in spring, summer, autumn and winter are focused on the causes of haze, positive emotions, prevention measures, and health effects, respectively; Wang et al.[10] conducted text analysis on the content of Weibo. Taking Harbin’s Weibo data as an example, it was found that the contents of the user's attention during the haze were divided into three categories: emotional expression and views set forth, information reminder, and individual context awareness; Lin [11] analyzed the Weibo data during the haze crisis in Singapore, and found that when the environmental crisis occurred, traditional media and new media adopted different ways in reporting relevant news, responding to major events, and releasing information to the public.

At present, most of the existing studies are related to the same area. Few studies involve the study of spatial differences of haze and public opinion concerns in different regions. Based on this, this study considers spatial differences and incorporates different areas into the scope of the study. The main innovation and contribution are to explore the public opinion response of Weibo users in the case of haze for the first time by using the co-word network method, and the difference in the degree of attention to haze in different cities. Firstly, the keywords of each Weibo post are extracted according to the TF-IDF algorithm by crawling the data dominated by "haze", and the co-occurrence triples are constructed by using the co-occurrence relationship of keywords, and then the co-word network is constructed. Then, the topic community is mined through the community detection algorithm. On this basis, the differences, and the degree of differences of public opinion topics in different urban areas are compared. From the perspective of public opinion development, this paper provides differentiated theoretical guidance for cities to deal with haze. For example, the government should provide timely and appropriate emotional guidance to avoid the occurrence of bad behavior; for the negative effects caused by haze, such as health threat and traffic impact, corresponding measures should be taken to reduce the damage and improve the sense of happiness; for other different concerns, corresponding strategies should be adopted to promote sustainable urban development.

2. Materials and Methods

2.1. Data

This paper mainly adopted two kinds of data: air quality data and Weibo data. The urban air quality data were collected from the World Air Quality Index (AQI) website (https://aqicn.org/map/china/cn/), and the scope of collection includes typical cities in North China, including Beijing, Tianjin, Shijiazhuang, Taiyuan, and Hohhot, as well as Shenyang, a city in Northeast China, and Jinan, a city in East China, where the haze was relatively serious. The time for data collection was winter, including December,
January and February, when the haze was the most serious compared with other seasons, and the related Weibo data were richer and more representative and valuable for the study. The data contained information in relation to city, date, PM2.5, PM10, carbon monoxide (CO), nitrogen dioxide (NO2) and sulfur dioxide (SO2). Since PM2.5 was the main cause of haze pollution, this paper applied PM2.5 content to illustrate the degree of haze pollution. The Weibo data were captured by the Houyi Caiji [23], taking the “haze” and “air” as the search keywords to fetch the relevant Weibo data of seven cities in winter 2017. The fetched data of each Weibo post included user ID, post, time range, number of “Likes” and “comments”, etc. A total of 34,373 relevant Weibo data were collected in the end.

The above-mentioned raw data of Weibo obtained through the collector fell into the category of unstructured data with many noises, such as duplicate data, commercial advertisements, special symbols, etc., which could not be processed directly by computers. In order to improve the accuracy and efficiency when mining topics, it was of necessity to carry out pre-processing of data cleaning, word segmentation and stop-word filtering. This paper discarded emojis, hashtags, web links, etc. by using Python language programming to realize data cleaning. The jieba library in Python was called to segment the cleaned data into units, and the stop-words such as “Wo, De, Le” were filtered by loading the Stop-word List of HIT. The raw Weibo data was finally converted into structured data consisting of several words for subsequent analysis.

2.2. Methods

Based on the community-based co-word network method, considering the network characteristics of social media, the keyword node is used to construct a co-word network. The more the Weibo community with the same keywords, the closer the connection, so the co-word network can be expressed in the form of "network community, topic node, edge" [12]. Topic extraction depends on the division of communities containing different vocabularies, and the degree of modularity between communities [13] determines the accuracy of community division, that is, the same community should contain as many as possible of the same keywords, while different communities should contain as few as possible of the same keywords. For example, Ding Sheng-chun [14] considered the WeChat characteristics and user behaviors in the process of propagation, and found the potential theme of Wei Zexi event; Wang Yandong, et al. [15] used co-words network method to mine topic community in network public opinion text data, and detected the development stage and trend of disaster. This type of method can automatically identify the number of topics, use the online community as the basic unit of topics, and map the real social network to the virtual cyberspace, which conforms to the feature of the gathering of small communities in Weibo content, and has great advantages in the topic mining field. Therefore, the co-word analysis method is selected to study the difference of people's attention to haze and the degree of difference in different cities in the case of haze. The research flow chart is shown in figure 1:
2.2.1 Keywords extraction

Keywords not only reflect the topic relevance in the text, but also reflect the importance of the words [16]. Therefore, a certain keywords extraction technology needs to be used to screen out the key feature words that have a large contribution to the construction of the co-word network. In the keyword extraction technique, a more classic keyword extracting method is TF-IDF algorithm [17 - 18]. TF-IDF is a weighting method commonly used in the field of information retrieval and text mining. The main idea is: if a word appears frequently in a document and appears less frequently in other documents, the word has a good ability to distinguish categories. In the TF-IDF algorithm, TF refers to the term frequency (Term Frequency), which reflects the importance of a word to a text. IDF refers to the inverse document frequency (Inverse Document Frequency), which reflects the importance of a word to a collection of texts. TF-IDF actually refers to TF*IDF, which means that the importance of a word is directly proportional to the number of times that the word appears in the text, and is inversely proportional to the frequency of the word appearing in the whole text collection. If a word has a high TF value in the text (more times in the text) and a high IDF value (less times in other texts), it means that the word can represent the central content of the text. The specific calculation formula of TF-IDF value is:

\[
TF - IDF = TF \times IDF
\]

\[
TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}
\]

(1)

\[
IDF_i = \log \frac{|D|}{|\{j: t_i \in d_j\}|}
\]

Where: \(n_{i,j}\) represents the number of words appearing in the text \(d_j\), \(\sum_k n_{k,j}\) represents the total number of words appearing in the text \(d_j\), \(|D|\) represents the total number of texts in the text collection, \(|\{j: t_i \in d_j\}|\) represents the number of texts containing words \(t_i\). The importance weight is calculated...
by the formula. Combined with the core keyword screening index, the key feature words that appear more in some texts but less in the whole text collection can be selected to build a co-word network.

2.2.2 Co-word network

The construction of the co-word network depends on the formation of the keyword co-occurrence matrix. The co-occurrence matrix is established according to the frequency of keyword co-occurrence, which is the basis of subsequent statistical analysis[19]; the co-word network is used to describe keywords and co-occurrence relationships. The mathematical graph model is $G = (V, E)$, where $V$ is a non-empty set, representing a node (Node) formed by keywords; and $E$ is also a non-empty set, representing an edge formed by the co-occurrence relationship between keywords (edge); $e_{ij}(G)$ represents the co-occurrence edge between node $V_i$ and node $V_j$ in G; $\omega_{ij}$ represents the weight, which is the number of co-occurrence between node $V_i$ and node $V_j$ [20-21]. The co-word network structure is shown in figure 2:

![Fig. 2 Schematic diagram of co-word network structure](image)

2.2.3 Louvain community detection algorithm

The structure of the co-word network is usually not disordered but will show certain rules. Generally speaking, some nodes in the network will gather into small groups. The closer the group connection, the more the same keywords and the more topics it contains, and the more similar it will be[12]. This small group structure is also known as the community. The discovery and description of topic in Co-word network can be transformed into the discovery of topic community, and community contains words representing the content of the topic. At present, compared with other algorithms, Louvain algorithm[13] is more applicable to large-scale networks. Therefore, this research uses this algorithm for community discovery of co-word networks, which is based on modularity[22] for optimization and heuristic calculations. It has the characteristics of strong explanatory calculation results and support for large-scale networks. The modularity is defined as follows:

$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{k_i \times k_j}{2m}] \times \delta(C_i, C_j)$$

(2)

Where, $m$ is the total number of edges in the graph, $k_i$ represents the sum of the weights of all the edges that point to node $i$, $k_j$ represents the sum of the weights of all the edges that point to node $j$, $A_{ij}$ represents the weight of the edges between nodes $i$ and $j$, $C_i$ and $C_j$ respectively
represents the community to which node i and j belong. When i and j belong to the same community, δ is 1, otherwise, they do not belong to the same community, δ is 0. By continuously optimizing the degree of modularity, communities with different topic content can be divided, and each community has as many same keywords as possible, and there are as few same keywords as possible between communities[16], so different communities can be clearly distinguished.

3. Results & Discussion

Fig.3 Co-word network of Beijing

Fig.4 Co-word network of Jinan City

Fig.5 Co-word network of Shenyang

Fig.6 Co-word network of Shijiazhuang
Figures 3-9 show the co-word network of seven cities. In the co-word network figures, the node color determined the contents of the topic communities and the node size determined the size of the topic communities. Integrating the content of nodes with different colors in the co-word network figures and the attributes of nodes in Gephi allowed us to understand the situation of topic communities. Further, the differences of people’s opinion concern in seven cities under the haze event were quantitatively analyzed, as shown in Table 1.

| Topic Community: Haze Control | Words | Control | Problem | Environment | Need | China |
|-------------------------------|-------|---------|---------|-------------|------|-------|
| Pagerank                      | 0.003949 | 0.003759 | 0.003698 | 0.002557 | 0.004875 |
| Words                         | Health | Cough   | Cold    | Throat      | Comfortable |
Due to space limitations, the details of the remaining topic communities were not shown in this paper, and the following summary table 2 of topic communities was available using the same analysis method.

Table 2 List of co-word topic communities of cities

| City | Topic | Beijing | Jinan | Shenyang | Shijiazhuang | Taiyuan | Tianjin | Hohhot |
|------|-------|---------|-------|----------|-------------|---------|---------|--------|
| 1    | Explore the reasons | Intuitive feeling | Intuitive feeling | Intuitive feeling | Intuitive feeling | Intuitive feeling | Official voice |
|      | 0.262736 | 0.367953 | 0.354467 | 0.374295 | 0.30826 | 0.118189 | 0.316407 |
| 2    | Haze control | Travel & entertainment | Appeals & advocacy | Negative emotions | Haze control | Expert explanation | Haze control |
|      | 0.112274 | 0.046091 | 0.080266 | 0.029020 | 0.052288 | 0.071614 | 0.039412 |
| 3    | Optimistic attitude | Traffic impact | Flight cancellation | Haze control | Optimistic attitude | Health threats | Government regulation |
|      | 0.085384 | 0.030276 | 0.071757 | 0.015699 | 0.044031 | 0.061363 | 0.026896 |
| 4    | Intuitive feeling | Scenic spots | Optimistic attitude | Health threat | Traffic control | Appeals & advocacy | Intuitive feeling |
|      | 0.084136 | 0.027046 | 0.018363 | 0.012693 | 0.033482 | 0.053920 | 0.020484 |
| 5    | Travel and entertainment | Countermeasures | Holiday & rest | Traffic impact | Flight cancellation | Daily life | |
|      | 0.045813 | 0.007663 | 0.006785 | 0.008338 | 0.025348 | 0.047906 | 0.002097 |

Table 2 showed that when the haze occurred, in respect of the degree of people’s attention to the topics, the topics of concern in Beijing were explore the reasons, haze control, optimistic attitude, intuitive feeling, travel and entertainment; the topics of concern in Jinan were intuitive feeling, travel and entertainment, traffic impact, scenic spots, and countermeasures; the topics of concern in Shenyang were intuitive feeling, appeals and advocacy, flight cancellation, optimistic attitude, holidays and rest; the topics of concern in Shijiazhuang were intuitive feelings, negative emotions, haze control, health threats, and traffic impacts; the topics of concern in Taiyuan were intuitive feeling, haze control, optimistic attitude, traffic control, and flight cancellation; the topics of concern in Tianjin were intuitive feeling, expert explanation, health threats, appeals and advocacy, and daily life; and the topics of concern in Hohhot were official voice, haze control, government regulation and intuitive feeling.

Besides, there was a cross section of concerns in each city. In terms of intuitive feelings about the haze, people in each city described what they saw and thought when the haze occurred, with Shijiazhuang, Jinan, Shenyang, Taiyuan, Tianjin, Beijing, and Hohhot, in descending order; in terms of discussion on haze control, four cities were concerned about it, with Beijing, Taiyuan, Hohhot and Shijiazhuang in order of concern; in terms of appeals and advocacy, Shenyang was better than Tianjin by degree of discussion; in terms of travel and entertainment, Beijing and Jinan were equally concerned; in terms of optimistic attitude, Beijing, Taiyuan and Shenyang were in order of degree; in terms of discussion of the traffic impact, Jinan and Shijiazhuang were in order of degree; in terms of concern about flight cancellation, Jinan and Taiyuan were in order of degree; and in terms of concern about health threats caused by haze, Tianjin and Shijiazhuang were in order of degree.

In addition to common topics of concern, each city had its own unique focus: the people of Beijing focused on the causes of the haze; the people of Jinan discussed specific measures to deal with the haze; the people of Shenyang were more concerned about holidays and rest in hazy weather; the people of Shijiazhuang may not be as tolerant of the haze as several other cities; the people of Taiyuan paid more
attention to traffic control; the people of Tianjin paid more attention to experts’ explanations of the haze phenomenon; and the people of Hohhot paid more attention to the official media’s response and the government’s remediation of haze pollution. Among the seven cities, Hohhot showed the best air quality, to some extent due to the timely attention and active response to the haze.

4. Conclusions

By using the community-based co-word network method, this paper explored the differences of people’s concerns in Weibo in different cities when haze pollution occurred. It took seven typical cities in North China, East China, and Northeast China as the study areas, and it was found that:

(1) Despite the subtle differences in each city's concerns about haze, people in each city would have timely perceptions of haze pollution when it occurred and would discuss various aspects of the effects brought about by haze.

(2) Out of the need for a healthy life, people in each city paid different degrees of attention to the haze management, such as haze control, appeals and advocacy, countermeasures and other concerns.

(3) Besides, the attitude of most cities towards the haze event was dominated by optimism, and a few cities would have abnormal emotions, such as the key words of hard, helpless and depression in the co-word network of Shijiazhuang.

The research results may serve as some theoretical guidance for urban health development. For example, in response to the abnormal emotion of the public, measures can be taken to achieve early psychological intervention to reduce the chance of bad behavior. In response to the negative impact of haze pollution on health, relevant medical departments can be added and medical resources can be deployed to meet health needs. In response to the road safety problems caused by haze, relevant departments can strengthen safety tips and increase road patrol control to reduce traffic accidents and create a good traffic safety environment for residents to travel. Managers can develop differentiated response strategies according to different haze public opinion concerns to achieve symptomatic management and improve management efficiency.

This paper constructed a co-word network using the co-occurrence relationship between keywords and discovered topic communities with Louvain community detection algorithm as a base unit to study the difference of concerns. In particular, keyword extraction, based on TF-IDF algorithm, would filter out certain important words and affect the accuracy of the detection of topic community. Besides, the topic mining in this paper was based on a static time period, and people’s concerns would change with the development of time. Therefore, improving the keyword extraction algorithm and the dynamic topic evolution of haze public opinion will be the focus of the later research.

Acknowledgements

This research was funded by the National Key Reasearch and Development Projects(2017YFB0503605), the National Natural Science Foundation of China(41771478), and the Fundamental Research Funds for the Central Universities(2019B02514).

References

[1] Sun, Y.L., Zhuang, G.S., Tang, A.H. (2006) Chemical characteristics of PM\textsubscript{2.5} and PM\textsubscript{10} in haze-fog episodes in Beijing. Environmental Science & Technology, 40: 3148–3155.

[2] Tan, J.H., Duan, J.C., Chen, D.H., Wang, X.H., Guo, S.J., Sheng, G.Y., He, K.B., Fu, J.M. (2009) Chemical characteristics of haze during summer and winter in Guangzhou. Atmospheric Research, 94: 238–245.

[3] Huang, J.P., Mcqueen, J., Wilczak, J., Djalalova. I. (2017) Improving NOAA NAQFC PM\textsubscript{2.5} predictions with a bias correction approach. Weather and Forecasting, 32: 407–421.

[4] Boldo, E., Medina, S., Tertre, A.L., Hurley, F., Mucke, H.G., Ballester, F., Aguiler, I. (2006 ) Aphesis: health impact assessment of long-term exposure to PM\textsubscript{2.5} in 23 European cities. European Journal of Epidemiology, 21: 449-458.
[5] Janssen, N.A., Hoek, G., Simic-Lawson, M., Fischer, P., Cassee, F.R. (2011) Black carbon as an additional indicator of the adverse health effects of airborne particles compared with PM$_{10}$ and PM$_{2.5}$. Environ Health Perspect, 119: 1691-1699.

[6] Gu, H.J., Yan, W.R., Elahi, E., Cao, Y. (2020) Air pollution risks human mental health: an implication of two-stages least squares estimation of interaction effects. Environmental Science and Pollution Research, 27: 2036-2043.

[7] Tian, X.Y., Yu, G., He F. (2016) An analysis of sleep complaints on Sina Weibo. Computers in Human Behavior, 62: 230-235.

[8] Yang, F., Muhamad, J.W., Yang, Q.H. (2019) Exploring environmental health on Weibo: a wextual analysis of framing haze-related stories on Chinese social media. International Journal of Environmental Research and Public Health, 16: 2374.

[9] Zhang, Q., Chen, J.S., Liu, X.Y. (2019) Public perception of haze weather based on Weibo comments. International Journal of Environmental Research and Public Health, 16: 4767.

[10] Wang, S.Y., Bai, H. (2014) A case study on micro-blog communication characteristics related to haze during the Harbin haze disaster.In: 2014 International Conference on Management Science and Engineering (ICMSE). Helsinki. pp. 120-125.

[11] Lin, T. (2019) Communicating haze crisis online: comparing traditional media news and new media perspectives in Singapore. Environmental Communication, 13: 864-878.

[12] Wang, Y.D., Fu, X.K., Li, M.M. (2018) A topic mining method for social media data based on co-word network. Journal of Wuhan University: Information Science Edition, 2008: 2287-2293.

[13] Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E. (2008) Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008: 1-11.

[14] Ding, S.C., Wang, N., Wu, C.L.Y. (2018) Research on microblog hot topic recognition based on keyword co-occurrence and community discovery. Modern intelligence, 38: 10-18.

[15] Wang, Y.D., Li, M.M., Fu, X.K., Shao, S.W., Liu, H. (2020) Disaster situation detection method based on Social Media Co word network. Journal of Wuhan University (Information Science Edition), 45: 691-698, 735.

[16] Liu, Z.Y. (2011) Research on keyword extraction method based on document topic structure. Tsinghua University, Beijing.

[17] Zhang, X.C., Song, C.B. (2007) Keyword extraction of Chinese documents based on text category information entropy. In: Research on Chinese Computing Technology and Language Problems: Proceedings of the 7th International Conference on Chinese Information Processing. Beijing. pp. 369-373.

[18] Luo, Y., Zhao, S.L., Li, X.C., Han, Y.H., Ding, Y.F. (2016) Text keyword extraction method based on word frequency statistics. Computer Applications, 36: 718-725.

[19] Wei, L., Zhang, Y.Q. (2015) Hot spot monitoring of Neurology in China based on CO-word analysis and visualization. Chinese Journal of Medical Library and Information, 224: 61-67.

[20] Shang, X.L., Wang, X.D. (2016) Text feature extraction method based on dynamic co word network in microblog topic recognition. Library and Information Knowledge, 3: 80-88.

[21] Liu, Z.Q., Yue, L.X., Xu, H.Y., Fang, S. (2020) Research on construction and dynamic visualization of temporal co word network. Acta Sinica Sinica, 39 : 186-198.

[22] Newman, M.E.J. (2006) Modularity and Community Structure in Networks. Proceedings of the National Academy of Sciences, 103: 8577-8582.

[23] Kuaiyi Technology Co., Ltd. [2020-08-04] Houyi collector. http://www.houyicaiji.com/.