Heat Analysis Technology in Public Opinion Early Warning Systems: Postprint

Authors: Zheng Chuangwei, Xie Zhicheng, Chen Shaobin, Xing Gutao, Chen Yifei

Date: 2023-10-08T00:00:00+00:00

Abstract

【Objective】To enhance the accuracy and efficiency of public opinion management in newspaper groups, this study investigates the practical application effectiveness of heat analysis technology in whistleblowing systems.

【Method】 The study proposes calculations for heat and relevance, encompassing hot topic calculation, correlation analysis of keywords, and associated heat computation, ultimately achieving event heat prediction.

【Result】 The practical application of heat analysis technology satisfies the need for capturing and timely tracking of hot topics in daily operations, holding significant implications for public opinion management.

【Conclusion】This study demonstrates that methods such as heat calculation and associated heat calculation employed in whistleblowing systems substantially enhance the system' s accuracy, enabling users to efficiently and intelligently obtain target news information of concern, interest, and value from massive news data, thereby more effectively supporting business operations including public opinion monitoring, news tracking, and news production.

Full Text

Preamble

Application of Heat Analysis Technology in Public Opinion Whistleblowing Systems

Chuangwei Zheng, Zhicheng Xie, Shaobin Chen, Guta Xing, Yifei Chen
(Shenzhen Creative Smart Harbor Technology Co., Ltd., Shenzhen, Guangdong 518034)

Abstract

[**Objective**] To improve the accuracy and efficiency of public opinion management in newspaper groups, this study investigates the practical application effects of heat analysis technology in whistleblowing systems.
[**Methods**] We propose heat and relevance calculations, encompassing hot topic computation, keyword association relevance analysis, and associated heat calculation, culminating in event heat prediction.
[**Results**] Through practical application of heat analysis technology, the system satisfies the daily work requirements of capturing and tracking hot topics in real time, representing significant implementation value for public opinion management.
[**Conclusion**] This research demonstrates that the heat calculation and associated heat calculation methods employed in the whistleblowing system substantially enhance system accuracy, enabling users to efficiently and intelligently obtain targeted news information that is highly concerned, interesting, and valuable from massive news data, thereby more effectively supporting business operations such as public opinion monitoring, news tracking, and news production.

Keywords: Public opinion; Whistleblowing system; Heat; Associated heat; Relevance
Classification: P413
Document Code: A
Article ID: 1671-0134(2023)05-134-05
DOI: 10.19483/j.cnki.11-4653/n.2023.05.031
Citation Format: Zheng C, Xie Z, Chen S, Xing G, Chen Y. Application of heat analysis technology in public opinion whistleblowing systems [J]. China Media Technology, 2023(05): 134-138.

Introduction

In the current Internet era, public opinion hot events are inevitable occurrences. When such events emerge, the public rapidly focuses attention and continuously expresses viewpoints, attitudes, or emotions. These network public opinion hot events typically form a focal point representing netizens' core emotions and interest demands. In recent years, research on network public opinion analysis has gradually deepened into experimental investigations by scholars, generally focusing on social networks and applications such as Weibo and WeChat. These social scenarios contain massive active users, and once hot topics appear, their propagation speed grows exponentially. Hot network public opinion primarily spreads through the Internet, where an event' s attention, commentary, and dissemination by the public trigger broader social concern.

Regarding heat analysis, domestic researchers have employed influence propagation models to describe hot events. These models count keyword dissemination frequency, where higher values indicate greater influence and lower values indicate less influence. Such models can evaluate interaction levels between different users in social networks. Simultaneously, by analyzing topic-related messages and repost counts, researchers assess whether a topic qualifies as hot, constructing influence propagation models based on user attention to reflect an event' s impact through keyword dissemination frequency. Other scholars propose detecting hot topics through time-unit detection, which confines a topic within a specific time unit, identifies feature units based on characteristic distribution, reorganizes them, and finally generates hot topics while precisely determining their occurrence timeframe for more accurate prediction.

In this data explosion era, a critical challenge is combining massive historical news data to provide editors, journalists, and other media professionals with fast, precise, and personalized news 线索 recommendations and intelligent whistleblowing warnings, thereby enhancing public opinion situational awareness and news insight capabilities while improving office efficiency and news creation capacity. To address this, heat analysis technology enables efficient and intelligent acquisition of concerned, interesting, and valuable target news information from massive news data, powerfully supporting public opinion monitoring, news tracking, and news production.

1. Related Research on Heat Analysis Technology

Through literature review on heat analysis, we find network public opinion heat analysis can proceed from two perspectives. First, from the user perspective, analyzing topics posted by users on forums, Weibo, and other platforms. Hot topics differ from ordinary topics primarily in the amount of information users employ to describe them, network resources consumed, and discussion duration. Second, from the media perspective, analyzing forwarding and ranking of hot events by news websites like Sina and Sohu. A topic' s emergence and dissemination result from extensive public discussion and media reporting/reposting, where whether it becomes a hot topic is often measured by report quantity and frequency.

Current research on public opinion heat calculation and prediction has achieved certain academic results, but most algorithms focus on data analysis without considering network public opinion' s inherent characteristics, particularly neglecting interconnectivity between network information. Therefore, this study combines association analysis thinking with heat calculation, comprehensively considering correlations in time, location, persons, and behavior to mine relationships between different attributes, constructing a public opinion heat prediction model with associations. By analyzing relationships between related events or information, we establish corresponding regression models for heat, making heat values more realistic.

2.1 Heat Calculation

In our whistleblowing system, discovering and predicting hot topics are both crucial for the media industry. Existing research mostly uses heat calculation combined with historical data validation, which often suffers from lag and cannot effectively predict a topic' s development trend when it first emerges, nor help government departments timely and precisely regulate opinion direction or continuously track topics according to monitoring rules. Therefore, this study employs the Z-algorithm to analyze and categorize article heat and sensitivity, saving results for timely hot topic discovery.

The specific process is as follows:

First, perform binary distribution statistics on semantically decomposed public opinion data (i.e., tokenized data) to count word occurrences, obtaining binary distribution statistics.

Next, calculate each word' s heat value using the standard score Z-Score algorithm:

$$Z = \frac{X - \bar{X}}{S}$$

where X is the word occurrence count, $\bar{X}$ is the average occurrence count, S is the standard deviation, and Z represents the deviation from the mean in standard deviation units, indicating the word' s heat value.

Store words with heat values exceeding the preset upper threshold in the hot active word library, and words below the preset lower threshold in the hot inert word library. The hot word library associates with domain libraries covering news, blogs, forums, and social networking sites, enabling queries about which domains each hot word originates from.

Calculate the co-occurrence threshold for hot active words in tokenized data based on word heat values and preset hot word libraries:

$$P1 = \frac{|Wx \cap Wh|}{|Wx|}$$

where Wx is the news word set and Wh is the hot active word set.

Calculate the co-occurrence threshold for hot inert words:

$$P2 = \frac{|Wx \cap Wc|}{|Wx|}$$

where Wx is the news word set and Wc is the hot inert word set.

Then perform linear weighted calculation using thresholds $P1$ and $P2$ to obtain the heat value:

$$H = Z_i \cdot P1 + (1 - Z_i) \cdot P2$$

where Z_i is the heat value of the i-th word, $P1$ is the hot active word co-occurrence threshold, and $P2$ is the hot inert word co-occurrence threshold.

Based on heat values, classify public opinion data as hot or non-hot according to preset heat level criteria, archiving qualifying data as hot documents and non-qualifying data as non-hot documents.

For sensitivity analysis, compare the hot active word library against a preset sensitive word library to obtain the count of sensitive words contained. Calculate the sensitivity value S as news sensitivity using:

$$S = \frac{Ws}{Wn}$$

where Ws is the number of sensitive words and Wn is the number of hot active words in the domain library.

2.2 Association Relevance Analysis

Public opinion prediction requires judging a topic' s future trend. Generally, higher relevance topic heat values indicate greater probability of becoming a hot topic, meaning the predicted topic' s probability of becoming hot correlates with its related topics' heat or quantity. Association analysis primarily includes calculating and weighting relevance between different word features such as time, location, persons, and behavior.

2.2.1 Time Relevance Calculation

Time relevance between topics mainly refers to whether the time difference between two topics falls within a specified range. Calculate the time interval to determine relevance—if within range, the topics are temporally associated, with shorter intervals indicating stronger association. The formula is:

$$R(Ti, Tj) = \frac{1}{|time(Ti) - time(Tj)|}$$

where $time(T1)$ represents a topic' s time, and Ti and Tj represent the two topics requiring relevance prediction. To analyze topic occurrence sequence, simply arrange $time(Ti)$ chronologically.

2.2.2 Location Relevance Calculation

Location relevance primarily uses location names in topics as the main basis, calculating relevance values through distances between primary locations. Therefore, construct a location-related noun set specific to district-level in cities or township-level in rural areas, and establish a hierarchical tree corresponding to higher administrative regions. If predicted topics' locations are within a certain distance range, they are considered associated, with association strength calculated based on distance—closer distance indicates higher relevance. The formula is:

$$R(T1, T2) = \frac{1}{|locate(T1) - locate(T2)|}$$

where $locate(T1)$ represents a topic' s primary location, and the difference with $locate(T2)$ indicates the path length between two topic locations in the hierarchical tree.

2.2.3 Person Relevance Calculation

Person relevance mainly examines whether persons or institutions involved in predicted topics follow each other or have other relationships. If friend or other relationships exist, the topics are considered associated in terms of persons. However, in practice, Weibo or WeChat friend relationships are often unavailable, so topic person names can be used for calculation, such as through duplicate name counts. The formula is:

$$R(T1, T2) = \frac{|people(T1) \cap people(T2)|}{|people(T1) \cup people(T2)|}$$

where $people(T1)$ is the set of person names involved in a topic, and Ti and Tj represent the two topics requiring prediction.

2.2.4 Behavior Relevance Calculation

Behavior relevance primarily collects topic behavior feature words for calculation. If behaviors are identical or similar, topics are considered relevant. The formula is:

$$R(T1, T2) = \frac{\sum_{w \in A1} maxsim(w, A2) \cdot IDF(w)}{\sum_{w \in A1} IDF(w)}$$

where $A1$ and $A2$ represent behavior feature word sets in two topics, $maxsim(w, Ai)$ is word semantic similarity, and $IDF(w)$ is obtained from corpus word information statistics.

2.3 Associated Heat Calculation

Associated heat calculation primarily segments topic heat by time, then identifies entities to calculate time relevance from temporal information, location relevance from location information, person relevance from person information, and behavior relevance from behavioral data, finally establishing a relevance connection graph.

In our whistleblowing system, we construct a news topic relationship graph, calculate heat values as initial weights for associated heat calculation within a time period. After heat calculation, relevance algorithms predict and analyze topic heat trends to achieve whistleblowing system warnings.

2.3.1 Establishing Inter-topic Relationships

Let $A = \langle V, E \rangle$ be the relationship graph, as shown in [Figure 1: see original paper], where $V = \{T_1, T_2, \ldots, T_{n-1}, T_n\}$ is the set of given topics, and $E = \{T_{1T}2, T_{1T}3, \ldots, T_{n-1}T_n\}$ is the edge set representing inter-topic relevance degrees. An edge exists only when the relevance between two vertices is not less than the threshold.

After establishing the relationship graph, the next step converts the graph into matrix form, where rows and columns represent graph vertices, and matrix values represent vertex degrees. As shown in [Figure 2: see original paper], R_{ij} represents the relevance degree between node i and node j, with values of 0 for edges where relevance is below threshold.

2.3.2 Calculating Relevance Importance of Related Topics

Define the transformation matrix M as:

$$M = d \cdot R + (1 - d) \cdot \frac{1}{n} E$$

where d is the damping coefficient ranging between 0 and 1. This matrix primarily measures each point' s influence on the point to be predicted. Matrix M has a unique stable distribution:

$$h = M^T h$$

The obtained h value can represent a topic' s importance degree in the relationship graph:

$$h = [d \cdot R + 1 - d]^T h$$

2.3.3 Heat Prediction

In whistleblowing systems, short-term heat trends with limited current information must be predicted to determine whether a topic will become hot. This study employs gray prediction methods, typically using the GM(1,1) model for topic heat prediction. The calculation process is as follows [10]:

a. Input initial sequence $X^{(0)} = (x^{(0)}(1), x^{(0)}(2), \ldots, x^{(0)}(n))$

b. Perform first-order accumulated generating operation on the initial sequence to obtain $X^{(1)} = (x^{(1)}(1), x^{(1)}(2), \ldots, x^{(1)}(n))$

c. Generate the adjacent mean sequence of $X^{(1)}$: $Z^{(1)} = (z^{(1)}(2), z^{(1)}(3), \ldots, z^{(1)}(n))$ where $z^{(1)}(k) = 0.5x^{(1)}(k) + 0.5x^{(1)}(k-1)$

d. The GM(1,1) gray differential equation model is: $x^{(0)}(k) + az^{(1)}(k) = b$, where a is the development coefficient and b is the gray action quantity. Let $\hat{\alpha} = (a, b)^T$ be the parameter vector to be estimated.

e. The least squares estimation parameter sequence satisfies: $\hat{\alpha} = (B^{TB})^{-1}B^{TY}$

f. Solve the differential equation to obtain: $\hat{x}^{(1)}(k+1) = (x^{(0)}(1) - \frac{b}{a})e^{-ak} + \frac{b}{a}$

g. Restore to original data to get: $\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k)$

This yields the heat trend prediction interval.

2.4 Heat Prediction (Event Correlation Method)

In practical work, the primary method used is event correlation-based public opinion trend prediction to determine whether a topic becomes hot. This model is based on the assumption that "events are interrelated and mutually influential," where connections exist between events that may affect or constrain each other. The algorithm framework is shown in [Figure 3: see original paper] [11].

The specific process mainly includes [12]:

(1) Retrieve recent events related to the topic to be predicted, paying attention to feature word selection when setting search terms.

(2) Search the group's local database, compare with Internet searches, analyze inter-topic relationships, and obtain text information data related to public opinion events. After data collection, denoising and other processing are required to ensure accuracy.

(3) Use clustering algorithms on organized text information to extract potential topic quantities.

(4) Sort text data chronologically, set time periods according to actual needs, and calculate inter-topic relevance based on event time, persons, location, and behavior in each period to obtain the relationship connection graph.

(5) Analyze different topics' importance degrees, predict associated heat, and finally calculate the possibility of a topic or information becoming hot.

3.1 Experimental Design

After predicting public opinion heat, our whistleblowing system uses posterior difference testing to verify experimental effectiveness. Specific steps include:

(1) Calculate the mean of the original sequence;

(2) Calculate the mean square deviation S_1 of the original sequence;

(3) Calculate residual mean values;

(4) Calculate residual mean square deviation S_2;

(5) Calculate the ratio $C = S_2/S_1$;

(6) Calculate small residual probability P.

This study uses P and C values to measure sudden public opinion prediction effectiveness and designs a corresponding posterior difference test discrimination reference table (see).

3.2 Experimental Results

Using the "Sun Xiaoguo Case" data in the database for heat prediction, the experimental results are as follows (see):

The results demonstrate that the associated heat calculation method achieves excellent prediction effectiveness for sudden public opinion, verifying the feasibility and effectiveness of the heat analysis technology used in this whistleblowing system.

This study provides in-depth analysis of heat calculation, association relevance analysis, associated heat calculation, and heat prediction used in the newspaper group whistleblowing system, listing relevant formulas and model factors such as time, location, persons, and behavior. Through these methods and practical application, the newspaper group whistleblowing system proves to have good accuracy, enabling users to efficiently and intelligently obtain concerned, interesting, and valuable target news information from massive news data, thereby strongly supporting public opinion monitoring, news tracking, and news production. Governments can also use this system to guide public opinion direction, enabling rapid response to major opinion events. This can help suppress negative public emotions toward opinion events to some extent, facilitating correct guidance of public opinion development trends and maintaining social harmony and stability.

References

[1] Liang Xiuming. Public crisis communication governance paths in new media environments [J]. China Media Technology, 2019(5): 48-50.

[2] Feng Xiaodong, Li Zhuoya, Shi Zhihui. Natural disaster impact assessment analysis based on network public opinion heat [J]. Information Exploration, 2020(1): 16-22.

[3] Yuan Ran. Data technology application practice in all-media communication [J]. China Media Technology, 2021(7): 21-23.

[4] Gao Ping, Zhou En. Research on public opinion guidance and countermeasures for government crisis PR under big data background—taking government Weibo as an example [J]. Yinshan Academic Journal, 2019(6).

[5] Mao Tong, Xie Chaode. Research on credit public opinion index construction and empirical study based on Baidu big data [J]. Credit Investigation, 2020(1): 11-20.

[6] Wang Wenyu, Yin Xueying. Government real-time network public opinion monitoring based on fuzzy evaluation method and heat model construction [J]. China Management Informationization, 2019(23): 170-.

[7] Zou Jiacheng, Ma Yuanyuan, Liu Ting, Tang Bochao, Liu Zhenguo, Gao Hui. Liquor industry public opinion information monitoring platform based on big data [J]. Brewing Science & Technology, 2020(3).

[8] Zhang Picui, Yang Jianwu, Shi Shuicai. Public opinion situation awareness in cyberspace [J]. Information Security Research, 2019(11): 1013-1020.

[9] Li Jingyun. Public opinion governance strategies for hot events in new media environments [J]. News Tide, 2019(10): 44-45+48.

[10] Zeng Runxi. The key to network public opinion governance is "preventive treatment" [J]. China Media Technology, 2018(12): 12-14.

[11] Zhang Yuanqi. Discussion on identification factors affecting network public opinion heat evaluation [J]. News Research Guide, 2022(1): 127-129.

[12] Wang Qianyi, Du Mingkun, Zhang Shan. Research on network public opinion heat based on deep learning [J]. Wireless Internet Technology, 2020(22): 16-17.

Author Biographies:
Zheng Chuangwei (1978-), male, from Shantou, Guangdong, Senior Engineer, research direction: big data, artificial intelligence;
Xie Zhicheng (1980-), male, from Shantou, Guangdong, Intermediate Title, research direction: big data, cloud computing;
Chen Shaobin (1973-), male, from Jieyang, Guangdong, Intermediate Title, research direction: big data;
Xing Guta (1984-), male, from Wenchang, Hainan, research direction: cloud

computing;
Chen Yifei (1981-), male, from Zhanjiang, Guangdong, Intermediate Title, research direction: big data.

(Executive Editor: Zhang Xiaojing)