Construction of a Popularity Measurement Model for Online Public Opinion Tweets: Post-print

Authors: Huang Wei, Liu Yi, Xu Yejing, Sun Yue

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] Data acquisition is the initial stage of online public opinion research. When confronted with massive amounts of data, constructing a popularity measurement model for online public opinion tweets can efficiently screen out data usable for online public opinion research. [Method/Process] Drawing upon the definition of average self-information in information theory, we construct an online public opinion popularity measurement model using the Analytic Hierarchy Process and the Haker News ranking algorithm. [Results/Conclusion] By crawling data from Weibo, we calculate the popularity threshold specific to this dataset and verify the accuracy of the popularity measurement model. The results demonstrate that the online public opinion tweet popularity measurement model can effectively accomplish the calculation of tweet popularity and achieve relatively high computational accuracy.

Full Text

Preamble

ChinaXiv Collaborative Journal
Vol. 63, No. 20, October 2019

Construction of a Heat Assessment Model for Network Public Opinion Tweets

Huang Wei, Liu Yi, Xu Yejing, Sun Yue
School of Management, Jilin University, Changchun 130022

Abstract

[**Purpose/Significance**] Data acquisition represents the first stage of network public opinion research. When confronted with massive datasets, construct-

ing a heat assessment model for network public opinion tweets enables rapid screening of data suitable for such research. [**Method/Process**] Drawing upon the definition of average self-information from information theory, this study employs the Analytic Hierarchy Process (AHP) and the HackerNews ranking algorithm to construct a network public opinion heat assessment model. [**Result/Conclusion**] By crawling data from Weibo and calculating the heat threshold for this dataset, the accuracy of the proposed heat assessment model is verified. The results demonstrate that the model effectively completes tweet heat calculations while achieving high computational accuracy.

Keywords: network public opinion; tweet heat; analytic hierarchy process
Classification Numbers: G250; G206
DOI: 10.13266/j.issn.0252-3116.2019.20.002

With the development of internet technology, the number of network tweets has grown exponentially, with various self-media platforms generating massive amounts of tweet information daily. In network public opinion research, crawling all public opinion tweet data from the internet simultaneously would create a data disaster. How to selectively and purposefully capture high-heat data that may cause public opinion events while filtering out low-heat data represents a significant challenge in the information acquisition stage of network public opinion research. Current research on network public opinion heat can be divided into two aspects: heat studies at single time points and heat trend studies within time periods. In single-time-point heat research, the primary focus is constructing evaluation systems for network public opinion tweets using quantitative data such as repost counts, comment numbers, likes, and follower counts. For example, Liang Changming et al. categorized tweet influence factors into blogger characteristics, content features, transmission features, and audience characteristics to construct a Weibo heat evaluation index system [1]. Du Hui et al. utilized causal models to describe topic heat from article quantity, click counts, comment numbers, and source counts, with time as a key variable [2]. In time-period network public opinion trend research, Xu Yini employed Markov chains to analyze relationships between Weibo quantitative data at different time points, plotted public opinion heat curves, and predicted trends [3]. Huang Wei et al. analyzed the temporal patterns of Weibo reposts and comments to measure the aging degree of Weibo public opinion [4]. Overall, current network public opinion heat research primarily focuses on quantitative data such as reposts, comments, likes, and follower counts, with insufficient attention paid to the tweet content itself, lacking discussion on how tweet content and author influence affect tweet heat.

This study approaches from tweet content analysis, combining heat calculations from both tweet quantitative data and author quantitative data, while comprehensively considering the influence of time on tweet heat. By incorporating the concept of average self-information from information theory and the HackerNews ranking algorithm from heat calculations, we establish a network public opinion tweet heat assessment model that filters tweets based on heat levels,

achieving preliminary filtering in the network public opinion crawling process. This model addresses the current deficiency in network public opinion heat assessment models that inadequately consider tweet content. Additionally, for tweet heat calculation, we not only consider the quantity of tweet ancillary information such as likes, reposts, and comments but also incorporate time dimensions including author network age and tweet survival duration, examining the rate of ancillary information per unit time as a factor affecting tweet heat.

2 Construction of Tweet Heat Assessment Model

2.1 Concepts of Tweet Heat and Self-Information

2.1.1 Concept of Tweet Heat Tweet heat represents the degree of attention, discussion, and dissemination that articles, images, and videos published by authors on self-media platforms such as Weibo and WeChat receive. In the network public opinion information acquisition process, heat calculation filters tweets that receive more attention, have broader dissemination ranges, and are discussed more frequently within unit time. These tweets are more likely to serve as latent network public opinion and trigger sensational events. Simultaneously, tweets with low attention, narrow dissemination ranges, and infrequent discussions are filtered out to prevent data disasters in network public opinion research. This study quantifies tweet heat as a number between 0 and 1 for standardized comparison. Tweet heat is influenced by three intermediate-level indicators: author influence, content appeal, and network dissemination power. Author influence includes three factor-level indicators: author fan growth rate, author publication growth rate, and author following growth rate. Content appeal includes content richness and average self-information. Network dissemination power includes weighted repost rate, weighted comment rate, and weighted like rate of tweets.

2.1.2 Concept of Self-Information According to Shannon's information theory, the self-information of an event x = a in event set X is defined as:

$$I_X(a_i) = -\log P_X(a_i)$$

where $0 \leq P_X(a_i) \leq 1$ represents the probability of event x = a occurring, and $\sum_{i=1} P_X(a_i) = 1$.

Formula (1) indicates that self-information represents the uncertainty of an event before it occurs and the amount of information contained after it occurs. Higher self-information corresponds to higher event uncertainty, while lower self-information indicates greater event certainty.

Considering that network public opinion tweets contain text, video, and image information, and that extracting multimedia semantics without semantic recognition is difficult, this study uses tweet text content, video titles, and image

titles as research objects. By statistically analyzing keyword frequencies in network public opinion tweets, we treat the keyword collection as event set X, a specific keyword as event x = a, and the frequency of a keyword's occurrence as $P_X(a_i)$, thereby calculating the self-information of specific words.

This study calculates the information content of individual tweets using average self-information:

$$AvgI_k = \frac{\sum_{i=0}^{n} -\log P_X(a_i)}{n}$$

where k is the tweet number, n is the number of useful words in the k-th tweet, a_i is the i-th useful word appearing in the k-th tweet, and $P_X(a_i)$ is the frequency of that useful word across all tweets.

2.2 AHP Model and Judgment Matrix Construction

2.2.1 AHP Model Construction The AHP model aims to describe network public opinion tweet heat using various quantitative data. By constructing judgment matrices, AHP reduces the difficulty of comparing different indicators and improves accuracy, thereby effectively calculating weight relationships among indicators. This study establishes the hierarchical analysis model shown in Figure 1 [Figure 1: see original paper]. The model's objective is to calculate network public opinion tweet heat (the target layer). Based on Malcolm Gladwell's three elements of popularity theory—requiring key personnel rules, environmental power rules, and content stickiness rules for objects to become popular—and drawing on references [1, 5-6] regarding network public opinion heat influence factors, we identify three criterion-level indicators:

(1) **Author Influence** refers to an author's influence scope and activity level on self-media platforms. Tweets published by authors disseminate within their interpersonal networks; authors with broader and more active networks enable faster tweet dissemination, resulting in higher tweet heat. This includes three bottom-level indicators: Author fan growth rate reflects the radiation degree of the author's influence scope. Faster fan growth rates indicate faster coverage expansion, meaning more people may potentially follow the tweet in the future, thereby increasing tweet heat. Author publication growth rate serves as a measure of author activity on the platform, positively influencing tweet heat. Author following growth rate also measures author activity and influence, positively affecting tweet heat values. All three metrics are calculated from crawled data on fan counts, publication counts, following counts, and the time span from account creation to tweet publication.

(2) **Content Appeal** refers to the quality and attractiveness of published tweets. Higher-quality tweets receive more attention; for example, tweets containing images and videos are more attractive than pure text. Content

appeal has two bottom-level indicators: Content richness, calculated from crawled tweet content by counting words and video presence; Average self-information, representing the frequency of a tweet or similar tweets on the platform. Higher average self-information indicates lower discussion frequency of the tweet's content, and vice versa. If a tweet type frequently appears on self-media platforms, it can be considered a current hot topic with relatively high attention.

(3) **Network Dissemination Power** refers to the dissemination speed and interaction capability of user-published tweets. User reposting behavior best reflects participation and dissemination levels. Reference [5] suggests that commenting, liking, and reposting all demonstrate user interaction participation, which influences overall tweet heat. This study designs network dissemination power as three bottom-level indicators: weighted repost rate, weighted comment rate, and weighted like rate, derived from crawled repost counts, comment numbers, like counts, and the time span from publication to crawling. Regarding weighting, this study designs an ideal value to exclude author fans' participation in tweet dissemination, weakening the correlation between network dissemination power and author fan counts. This empirically derived ideal value is 0.75; further definition, improvement, and refinement of this value will be addressed in future research.

2.2.2 Judgment Matrix Construction After establishing the network public opinion tweet heat hierarchical model, we compare the relative importance of indicators to construct judgment matrices. Let network public opinion tweet heat be A, author influence be B1, content appeal be B2, network dissemination power be B3, author fan growth rate be C1, author publication growth rate be C2, author following growth rate be C3, content richness be C4, tweet average self-information be C5, weighted repost rate be C6, weighted comment rate be C7, and weighted like rate be C8.

Through literature review and expert surveys: author influence B1 is slightly more important than content appeal B2; network dissemination power B3 is slightly more important than author influence B1; network dissemination power B3 is significantly more important than content appeal B2. Author fan growth rate C1 is slightly more important than author publication growth rate C2; C1 is strongly more important than author following growth rate C3; C2 is significantly more important than C3. Average self-information C5 is strongly more important than content length C4. Weighted repost rate C6 is slightly more important than weighted comment rate C7; C6 is between slightly and significantly more important than weighted like rate C8; C7 is more than equally important compared to C8.

Based on Table 1 and the above analysis, the judgment matrix is constructed as shown in Figure 2 [Figure 2: see original paper].

Table 1 Judgment Matrix Scale and Meaning

Scale	Meaning
1	Equal importance
3	Slight importance
5	Strong importance
7	Very strong importance
9	Absolute importance
2,4,6,8	Intermediate values between adjacent judgments

After constructing the judgment matrix, consistency must be tested using the criteria shown in formulas (3) and (4):

$$CI = \frac{\lambda_{max} - n}{n - 1}$$

$$CR = \frac{CI}{RI}$$

where λ_{max} is the matrix's maximum eigenvalue, n is the matrix order, and RI values are shown in Table 2 .

Table 2 Consistency Index RI Values

Matrix Order	1	2	3	4	5	6	7	8	9
RI Value	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45

Calculations yield $CR_A = 0.0214$, $CR_{B1} = 0.0102$, and $CR_{B2} = 0.036$, all satisfying consistency tests. The resulting weights for the public opinion tweet heat assessment model are:

- Author influence weight: 0.2583
- Content appeal weight: 0.1047
- Network dissemination power weight: 0.6370
- C1 (fan growth rate) weight: 0.6491
- C2 (publication growth rate) weight: 0.2790
- C3 (following growth rate) weight: 0.0719
- C4 (content richness) weight: 0.1250
- C5 (average self-information) weight: 0.8750
- C6 (weighted repost rate) weight: 0.6250
- C7 (weighted comment rate) weight: 0.2385
- C8 (weighted like rate) weight: 0.1365

2.3 Tweet Heat Calculation and Threshold Models

2.3.1 Tweet Heat Calculation Model Common tweet ancillary information heat calculation models include the Reddit ranking algorithm [7], PageRank [8], and HackerNews ranking algorithm. Reddit primarily applies to tweets with upvote/downvote statistics, while PageRank targets webpage links; neither supports the Weibo tweet heat calculation in this study. The HackerNews model parameters involve publication time and like counts, showing higher compatibility with this study's heat calculation model. Therefore, we modify the HackerNews algorithm to construct a new tweet heat calculation model.

The HackerNews ranking algorithm expression is:

$$\frac{(P-1)}{(t+2)^G}$$

where P is tweet vote count, t is time in days, and G is the gravity factor. Larger G values cause tweet rankings to drop faster over time, typically set to G = 1.8. The numerator subtracts 1 to exclude the author's own vote.

Based on the hierarchical model, author influence corresponds to indicators: author fan count (Fans#), author publication count (Publications#), author following count (Follow#), and account creation time span (t_{person}). Network dissemination power corresponds to: tweet repost count (Forward#), tweet comment count (Comment#), tweet like count (ThumbUp#), and tweet publication time span ($t_{article}$). Content appeal corresponds to: video presence (Video), tweet effective word count (Word#), tweet average self-information (SelfInformation#), and ideal value ($\rho = 0.75$).

We calculate author influence (r_{person}) and network dissemination power ($r_{article}$) using HackerNews combined with the above weights:

$$r_{person} = \frac{(0.6491 \times Fans\# + 0.2790 \times Publications\# + 0.0719 \times Follow\#)}{(t_{person}/(24 \times 60) + 2)^{1.8}}$$

$$r_{article} = \rho \times \frac{(0.6250 \times Forward\# + 0.2385 \times Comment\# + 0.1365 \times ThumbUp\#)}{(t_{article}/(24 \times 60) + 2)^{1.8}}$$

Treating a video as 100 useful words, we standardize tweet average self-information and word count using formula (8), then calculate content richness heat using formula (9):

$$g(x) = \frac{x - x_{min}}{x_{max} - x_{min}}$$

$$r_{content} = \begin{cases} 0.8750 \times (1 - g(SelfInformation\#)) + 0.1250 \times g(Word\#), & Video = False \\ 0.8750 \times (1 - g(SelfInformation\#)) + 0.1250 \times g(Word\# + 100), & Video = True \end{cases}$$

After calculating author influence, network dissemination power, and content appeal, we standardize the data using formula (10) to restrict results between 0 and 1. Considering that tweet influence expresses heat levels more strongly than network dissemination power and content appeal, we weight tweet influence to calculate overall heat using formula (11):

$$f(x) = \frac{x - x_{min}}{x_{max} - x_{min}}$$

$$r = 0.2583 \times f(r_{person}) + 0.1047 \times f(r_{content}) + 0.6370 \times f(r_{article})$$

2.3.2 Tweet Heat Threshold Calculation Model To calculate tweet heat thresholds more precisely, we establish the following mathematical model:

$$f(r_k) = \begin{cases} 1, & r_k \geq x \\ 0, & \text{else} \end{cases}$$

$$Err = \sum_{k=1}^{N} (Target_k - f(r_k))^2$$

where r_k represents the heat of the k-th tweet, x is the heat threshold, N is total tweet count, and $Target_k$ is the actual label. We optimize this model by selecting appropriate x values to minimize Err.

2.3.3 Keyword and Sensitive Word Libraries This study employs two word libraries for algorithmic refinement: a keyword library and a sensitive word library [9]. In practice, some heat-qualified tweets may have low relevance to network public opinion research (e.g., weather forecasts, inspirational quotes, lottery activities). Filtering these out using high-recognition keywords improves accuracy. Conversely, some heat-unqualified tweets may be highly relevant (e.g., local incidents not yet sensational, tweets containing sensitive information). Extracting these through sensitive word screening improves accuracy.

Sensitive words come from the open-source "2017 Sensitive Word Library" compiled by CSDN users, containing categories such as pornography, violence, reactionary content, corruption, and livelihood issues, used to screen tweets from the unqualified set. The keyword library includes weather, lottery, and advertisement categories compiled from Weibo tweets, used to filter irrelevant tweets from the qualified set.

2.4 Tweet Heat Assessment Process and Evaluation Model

2.4.1 Tweet Heat Assessment Process Based on the hierarchical model and heat calculation model, we apply the tweet heat assessment process shown in Figure 3 [Figure 3: see original paper] to measure network public opinion information heat. The heat calculation model is embedded in the hierarchical model for comprehensive heat calculation. Using the threshold calculation model, we estimate threshold standards: tweets above the threshold are considered heat-qualified, while those below are unqualified. Through keyword filtering and sensitive word screening, we finalize the network public opinion tweet sets.

2.4.2 Tweet Heat Assessment Evaluation Model We evaluate the model using accuracy and recall rates. Let T be the actual heat-qualified tweet set, ~T the unqualified set, CT the calculated qualified set, and ~CT the calculated unqualified set. Accuracy (ACC) and recall (REC) are calculated as:

$$ACC = \frac{|T \cap CT| + | \sim T \cap \sim CT|}{N}$$

$$REC = \frac{|T \cap CT|}{|T|}$$

Accuracy reflects overall calculation correctness, while recall represents the proportion of actual qualified tweets correctly identified, measuring false negative rates.

3 Data Acquisition and Cleaning

3.1 Data Acquisition

Data were collected from Sina Weibo using the Octoparse web crawler tool, gathering 70,833 tweets from 500 bloggers as initial data. The dataset includes 12 columns: author Weibo name, tweet acquisition time, tweet publication time, tweet content, repost count, comment count, like count, video presence, author account creation time, author following count, author fan count, and author publication count. After filtering duplicates and cleaning illegal characters, 63,816 valid Weibo entries were obtained.

Given the dataset's large size, complete manual annotation was impractical. We employed TensorFlow's open-source natural language processing tools and GitHub's convolutional neural network classification tools for sentiment analysis, labeling negative information as 1 and positive/neutral as 0. Tweets with negative sentiment but minimal interaction were marked as 0. Manual screening of the small amount of data labeled 1 removed tweets irrelevant to network public opinion research, including weather, inspirational quotes, welfare, government announcements, gossip, pets, showmanship, and lottery types, marking

them as non-crawl tweets. The final dataset contained 51,531 non-crawl tweets (labeled 0) and 12,285 crawl-required tweets (labeled 1).

3.2 Chinese Word Segmentation

Using Java programming and the ansj Chinese word segmentation tool from Maven, we implemented Chinese word segmentation for tweet content. Considering pronouns, prepositions, modal particles, and other semantically irrelevant words, we applied ansj's stopword tool to filter 20 part-of-speech categories: general nouns, personal names, transliterated names, place names, transliterated place names, organization names, other proper nouns, nominal idioms, nominal morphemes, new words, location words, general verbs, auxiliary verbs, gerunds, verbal morphemes, adjectives, adverbial adjectives, nominal adjectives, adverbs, and distinctive words. Additional stopwords like "有" (have), "没有" (no), "还" (still), "是" (is), and "也" (also) were manually added.

Based on segmentation and stopword settings, we conducted word frequency statistics, calculating individual word frequencies. Due to small frequency values, we used Java's BigDecimal package for precise calculation. The segmentation and self-information calculation code is as follows:

```
// HashMap for word counts
HashMap<String, Integer> map = new HashMap<String, Integer>();
// Segmentation result list
ArrayList<Result> ls = new ArrayList<Result>();
for (int i = 0; i < sheet.getRows(); i++) {
    // Core segmentation code
    String cellInfo = sheet.getCell(3, i).getContents();
    Result strs = ToAnalysis.parse(cellInfo);
    ls.add(strs);
    for (java.util.Iterator<Term> itr = strs.iterator(); itr.hasNext();) {
        Term temp = itr.next();
        // Stopword filtering
        if (expectedNature.contains(temp.getNatureStr()) && !stopWords.contains(temp.getNatu
            String tempString = temp.getName();
            if (map.containsKey(tempString)) {
                map.put(tempString, map.get(tempString) + 1);
                total++;
            } else {
                map.put(tempString, 1);
                total++;
            }
        }
    }
}
for (Entry<String, Integer> entry : list) {
    String key = entry.getKey();
```

```
    Integer value = entry.getValue();
    // Precise frequency calculation using BigDecimal
    BigDecimal bValue = new BigDecimal(value);
    BigDecimal bTotal = new BigDecimal(total);
    BigDecimal percentage = bValue.divide(bTotal, 10, RoundingMode.HALF_{UP});
    // Calculate self-information
    double selfInformation = -Math.log(percentage.doubleValue()) / Math.log(2);
}
```

4 Experiments and Results

4.1 Experimental Process

4.1.1 Word Segmentation and Average Self-Information Calculation In the Java environment using the ansj Chinese word segmentation tool, we achieved segmentation results shown in Figure 5 [Figure 5: see original paper]. Using HashMap to record each word and its frequency, we calculated self-information per word using formula (1), as shown in Figure 6 [Figure 6: see original paper]. Finally, we calculated average self-information per tweet using formula (2).

4.1.2 Heat Calculation After calculating average self-information, we computed single tweet publication duration (from publication to crawling time) and account creation duration (from account creation to crawling time) in minutes. The resulting basic tweet data is shown in Figure 7 [Figure 7: see original paper]. Using formulas (6)-(11), we obtained the calculation results shown in Figure 8 [Figure 8: see original paper].

4.2 Experimental Results

Applying the tweet heat model to quantitative data yielded heat values for 63,817 tweets. Normalized heat values ranged from 0 to 1, with the lowest at 0.00002 and highest at 0.935. Partial results are shown in Figure 9 [Figure 9: see original paper] (high-heat tweets) and Figure 10 [Figure 10: see original paper] (low-heat tweets). High-heat tweets show high fan counts, following counts, publication counts, and interaction numbers, with low average self-information, representing hot network topics. Most were marked as crawl-required, except for some advertisements and inspirational quotes. Low-heat tweets show minimal interaction (mostly zero), low author influence metrics, and high average self-information, representing niche content.

Based on heat calculations, we optimized formulas (12) and (13) to find the minimum error of 0.059 at threshold $x = 0.4$, as shown in Figure 11 [Figure 11: see original paper]. At threshold 0.4, formulas (14) and (15) yield 94% accuracy and 91% recall. This indicates many tweets in the low-heat group warrant crawling. Using sensitive word screening on the low-heat set extracted

131 tweets, while keyword filtering on the high-heat set removed 1,128 tweets. After filtering, accuracy rose to 95% and recall to 92%.

While the algorithm's accuracy is slightly inferior to convolutional neural networks, it offers significant advantages in time complexity. With O(N) linear complexity compared to machine learning algorithms' $O(N^2)$ starting point, this algorithm provides clear speed advantages for preliminary data filtering in network public opinion research. In the experiment with 63,816 tweets, 50,865 irrelevant entries were filtered, collecting 12,951 valuable network public opinion data points—filtering 79% of useless data at 95% accuracy, demonstrating good performance.

5 Summary and Outlook

This study constructs a complete heat assessment model through hierarchical analysis and heat calculation models. By analyzing tweet authors, content, and ancillary information, we calculate tweet heat and improve assessment accuracy through keyword screening and sensitive word filtering. The heat calculation provides technical and data support for information filtering in subsequent network public opinion acquisition. This is a static heat assessment model that can be reapplied at different time nodes to study dynamically changing tweet heat.

This study only crawled and analyzed Weibo tweets. Future research should extend to multimedia network public opinion information including images, videos, and audio. The weight settings in tweet dissemination weighted rates also require further refinement.

References

[1] Liang Changming, Li Dongqiang. Empirical research on Weibo heat evaluation index system based on Sina Hot Platform[J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(12): 1278-1283.

[2] Du Hui, Guo Yan, Fan Yixing, et al. Topic heat calculation and prediction method based on causal models[J]. Journal of Chinese Information Processing, 2016, 30(2): 50-55.

[3] Xu Yini. Analysis of network public opinion heat trends for media spectacles based on Weibo[J]. Information Science, 2017, 35(2): 92-97, 125.

[4] Huang Wei, Wang Jiejing, Zhao Jiangyuan. Research on aging measurement of Weibo public opinion information[J]. Information and Documentation Services, 2017(6): 6-11.

[5] He Yue, Cai Bochi. Weibo heat evaluation model based on factor analysis[J]. Statistics & Decision, 2016(18): 52-54.

[6] Rao Hao, Wen Haining. Real-time linear model for Weibo topic early warning analysis[J]. Library and Information Service, 2017, 61(15): 130-137.

[7] GLENSKI M, PENNYCUFF C, WENINGER T. Consumers and curators: browsing and voting patterns on reddit[J]. IEEE transactions on computational social systems, 2017, 4(4): 196-206.

[8] BERKHIN P. A survey on pagerank computing[J]. Internet mathematics, 2005, 2(1): 73-120.

[9] Deng Yigui, Wu Yuying. Sensitive word decision tree information filtering algorithm based on text content[J]. Computer Engineering, 2014, 40(9): 300-304.

Author Contributions

Huang Wei: Framework design and overall concept supervision
Liu Yi: Model construction and paper writing
Xu Yejing: Data collection and experimentation
Sun Yue: Data collection and paper proofreading

The Construction of Heat Assessment Model for Tweets of Network Public Opinion

Huang Wei, Liu Yi, Xu Yejing, Sun Yue
School of Management, Jilin University, Changchun 130022

Abstract: [Purpose/significance] Data Collection is the first step of the study of Network Public Opinion. The construction of Heat Assessment Model for Tweets of Network Public Opinion will rapidly screen useful data over dramatic number of data. [Method/process] This paper cites the definition of Average Self-Information, applies Analytic Hierarchy Process (AHP) and HackerNews Ranking Algorithm to construct a Heat Assessment Model for Tweets of Network Public Opinion. [Result/conclusion] Through the calculation of data collected from Weibo, this paper obtains the threshold of this dataset. Then this paper tests the accuracy of the model, which proves this model could achieve the heat calculation precisely.

Keywords: network public opinion; heat of Tweets; AHP

Note: Figure translations are in progress. See original paper for figures.