Research on Public Opinion and Early Warning Analysis Model of Network Emergencies Based on Decision Tree

Peipei Wang
School of Accountancy, Shandong Management University, Jinan, China

Corresponding author’s e-mail: Wpp870213@163.com

Abstract. Taking the existing network public opinion cases as the research object, the paper analyzes key factors affecting the network public opinions by the machine learning method and the big data technology statistical technology, and builds the early warning system of public opinions based on this. The decision tree model is also built to realize analysis and early warning, and finally the "Qingdao Sky-price Prawns" incident is cited to illustrate the feasibility of the early warning by classification proposed in this paper.

Keywords. Public Opinion, Early Warning Analysis, Network Emergencies, Decision Tree

1. Introduction
Nowadays, the texts of public opinions represented by news reports, WeChat Moments, Q-Zone, blog information, and BBS news are springing up like mushrooms. The open and interactive texts of public opinions have become an important way for people to obtain information. With the maturing of the media, especially in recent years, a variety of network emergencies involving people's livelihood have caused great influence in society. For example, the “Weifang Gauze Gate” incident (2016) and the “China murder” happened in Liaocheng city of Shandong province (2017) and other incidents in the past two years are closely related to the people's livelihood. These incidents have been continuously spread among the majority of netizens, caused the widespread panic, affected the social stability and unity and had a bad influence on the public opinion on the people's livelihood in China and the implementation of public policies involving people's livelihood (Li Chuanjun, 2017). This calls for the special analysis and effective supervision of the network public opinion on the major network incidents that concern the people's livelihood and are closely watched by the masses.

2. Related foreign and domestic researches
Since the incident “British protest against high oil price” in 2000, the network public opinions that were widely spread through the Internet after being triggered by emergencies have attracted the attention of governments and researchers around the world. Nkpa [1], Bertolotti [2], et al. defined the information that induced the network public opinions at the earliest from the perspectives of communication science and epistemology. The literature [3,4] introduced virus propagation theory and network topology information in their study to control the speed and scope of the dissemination of the network public opinions. The above research was conducted on the condition that known network public opinions...
exist. Actually, not only automatically predicting network public opinion from network data flow is itself a challenge, but also the network public opinions are featured with dynamics, variability, emergency and other uncertain factors. These problems have caused great difficulties in the early warning of public opinions. Therefore, the network public opinion warning strategy is an indispensable step in dealing with network public opinions.

To adapt to the situation that network public opinions have multiple subjects, complex relationships, and are difficult to be measured, and reasonably evaluate the development trend of network public opinions, Zhang Peng [5] et al. considered the early warning of network public opinions as a classification issue. Their study integrated some key factors such as the attention of media and Internet users and the explosiveness of topics under the network environment as the indexes of early warning of public opinions, and these classification indexes were also introduced into the Bayesian network as the basis for the classification of the early warning of network public opinion. The research of Dong Jianfeng [6] showed that the network public opinion and emergencies influence each other and have a mutual catalytic effect, which leads to the dynamic evolution of the network public opinion and brings difficulties to the early warning of public opinions. Liu Yi [7] screened the early warning indexes of network public opinions in advance, based on which the triangular fuzzy numbers and the fuzzy analytic hierarchy process were used to establish the weight of each early warning index, thereby solving the uncertainty for the quantification of the seller’s semantic evaluation. Lin Chen [8] studied the current early-warning index system of network public opinions and found that the design of the current index system is incomplete and not easy to be implemented. Therefore, they started from the life cycle of network public opinions and monitored all factors affecting the outbreak and derivation of the network public opinion, based on which four performance indexes, including information energy, communication energy, subject energy and opinion energy, were built. Unfortunately, they did not quantify the above indexes in their study.

At present, researches on the network public opinion system mainly focus on the collection and mining of network data, the sentiment orientation analysis of network information, and the formation of early warning indexes of network public opinions. Among them, the early warning program of network public opinions based on data mining is to mine the network public opinion data by capturing network data, completing the clustering and classification of network data stream, further extracting network information features, and extracting features of public opinion related data. Finally, predicting the possibility of network public opinion on the basis of the data mining of network public opinion, and discovering public opinions in an unsupervised manner are the most effective method in the context of current massive network data.

In the analysis of sentiment orientation of network information, it is found that the attitude of netizens can promote the development of network public opinions [9,10]. The early warning scheme of network public opinions based on the sentiment analysis can solve the problem of early warning of network public opinions to a certain extent. However, the early warning of public opinions based on the sentiment analysis needs to obtain emotional information in an all-round way. In effect, information from the we-media environment is disseminated in a fragmented manner, resulting in difficulties in accessing to all information, unbalanced sentiment analysis, and ultimately reducing the accuracy of public opinion warnings.

In the study on the building of the early-warning index system of network public opinions, the researchers discovered the characteristics such as the publisher, heat and growth status of public opinions during the occurrence, development and disappearance of network public opinions through the life cycle of network public opinions, and quantified these indexes as the basis of the early warning of the network public opinions to build the early-warning index system of network public opinions [11,12]. Regardless of the whole life cycle of online opinions, only the various factors that stimulate the occurrence and development and the difficulty in capturing and quantifying these factors are themselves the subjective indexes defined by the researchers, and the quantification also depends on the cognitive level of experts. The development of public opinions is easy to be misjudged in the early stage, and
there is a serious lag in the mid-term of development, which will not achieve the purpose of early warning.

This paper proposes that the scheme for the building of an early warning index system of online opinion based on the decision tree classification belongs to the category of machine learning and data mining. In this scheme, the learning characteristics of the stimuli events are automatically learned from the machine learning algorithm, and the impact on the events is captured quantity their strengths, build a network early-warning index system and achieve the purpose of early warning.

3. The building of early-warning index system of network public opinions based on the decision tree classification

All kinds of information elements of related events in the Internet may become the breaking points of public opinions. The main tasks of the early warning mechanism of network public opinions are to predict the network public opinions, identify its breaking points and give warnings to relevant functional departments. Therefore, the main tasks of building the model are to integrate the fragmented information and heterogeneous information scattered in every corner of the network, extract the features that affect the development of public opinions, and evaluate its strength of influence. The specific assessment process is shown in Figure 1.

![Figure 1. Early-warning Index System of Network public opinions Based on the Decision Tree Classification](image)

3.1. Decision tree model

The decision tree model is essentially a classification tree, consisting of nodes and a group of directed edges. The nodes can be divided into the internal nodes that identify the attributes on a branch in the tree and the leaf nodes that identify the classification results. In the grading mechanism of network public opinions, leaf nodes identify different warning levels, as shown in Figure 2.
Obviously, using the decision tree to classify early warnings is to divide the characteristic space into the non-intersected areas based on the conditional probability distribution of the feature system in the context of feature systems. Each path in the decision tree shown in Figure 2 corresponds to one conditional probability distribution that is constituted by the probability distribution of early warning at each level. The conditional probability distribution described by the entire decision tree is composed of probability distributions of early warning conditions at all levels.

Assuming that the random variables of the early warning system of network public opinions are represented by X, and the random variables of the early warning category at each level are represented by Y, the conditional probability distribution corresponding to each level of early warning in the decision tree can then be expressed as $P(Y | X)$.

### 3.2. Feature selection

Selecting features with higher classification ability is the key to constructing a decision tree for classification. Therefore, the classification ability of each feature is quantified by calculating the information gain ratio of each feature. The information gain ratio refers to the extent to which the characteristic belonging to Class Y is reduced when the feature X information is known. Training set is defined as $D$, and sample size as $|D|$. Assuming there are K known classifications (the classification warning system in this paper is divided into 4 classifications), it is defined as $C_k$, in which

$$k = 1, 2, ..., K; \ |C_k|$$

is the number of samples of classification $C_k$, and obviously $\sum_{k=1}^{K} |C_k| = |D|$. If the feature $A$ can have $n$ values $\{a_1, a_2, ..., a_n\}$, then the training set $D$ can be classified into $n$ different sub-sets according to the different values of feature $A$, and the information gain ratio can be written as:

$$G(D, A) = H(D) - H(D | A)$$  \hspace{1cm} (1)

Where, $H(D)$ describes the empirical entropy of the training set $D$; $H(D | A)$ represents the empirically conditional entropy of $D$ under the given feature $A$. The description shows that the higher the information gain ratio is, the more determined the information is.

In case of the training set $D$ and the feature $A$, the information gain ratio of the feature $A$ against the training set $D$ is written as $G(D, A)$, and the information gain ratio can be calculated by the following steps:

1. Calculating the empirical entropy $H(D)$ of the training set $D$:

$$H(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|}$$  \hspace{1cm} (2)
(2) Calculating the empirically conditional entropy \( H(D \mid A) \) of the feature \( A \) against the training set \( D \):

\[
H(D) = \sum_{i=1}^{s} \frac{|D_i|}{|D|} H(D_i) = -\sum_{i=1}^{s} \frac{|D_i|}{|D|} \sum_{k=1}^{c} \frac{|C_{ik}|}{|D_i|} \log \frac{|C_{ik}|}{|D_i|}
\]

(3) Calculating the information gain ratio of the feature \( A \) against the training set \( D \):

\[
G(D, A) = H(D) - H(D \mid A)
\]

3.3. Construction of decision tree

The CART (Classification and Regression Tree, CART) algorithm is a decision algorithm that can be used for classification, and can be used to make early warnings and classifications of public opinions. CART realizes the regression algorithm through the decision tree. It uses the minimum residual variance to determine the optimal partition of the decision tree. The following is the flow of CART:

(1) Building the main function of decision tree:
   a. Training sets and classification tags
   b. The minimum residual variance is used to calculate the optimal division of the regression tree, and the division nodes of each feature are created based on the different features.
   c. The division nodes are used to divide the training set into two parts.
   d. Based on the results of the binary data set, the left and right nodes of the decision tree are built as branches of the decision tree.
   e. Determine whether to meet the recursive termination conditions.
   f. Input the new nodes as new sub-sets and classification labels, and conduct the recursion of the above steps.

(2) The minimum residual variance sub-function is used to calculate the optimal division of the data set.

(3) Dichotomous training set: the training set is divided into two data sets according to the given optimal division.

(4) Pruning strategy: pruning the decision tree by pre-pruning and post-pruning.

4. Cases and results analysis of model application

4.1. Case data set

This paper captured the hottest 60 cases from 2013 to 2014, such as “A Malaysia airline lost contact” and “Hong Kong Occupies the Central” in advance by the web crawler, from which 30 events were selected randomly as the experimental training set, and the rest 30 topics as a test set. To normalize the incident warning into a unified framework system, the experiment abstracts the topical features into four categories: topic heat, topic sensitivity, topic tendency, and topic burstiness. Among them, the topic heat calculation is commonly quantified by topics such as the number of views, the times to be replied, and the times to be forwarded. The topic sensitivity is subject to the sensitivity classification based on a Bayesian classifier. The topic tendency analyzes the sentimental tendencies contained in the content of the report by the tendency analysis method in Chapter 4 of this report. For the burstiness, the detection scheme for the emergencies of the preliminary research program of the research group is used to quantify the burstiness of each event.

4.2. Pruning estimation of decision tree model

If the decision tree is set too deep during the fitting process of real data, the model may fit the training set well, but it may be difficult to fit the test set well, resulting in a decrease in the generalization ability of the decision tree model and followed with the overfitting. In order to avoid over-fitting during the training process, this paper tests the generalization ability of the constructed decision tree in advance.
Figure 3 to Figure 8 describe the degree of fit of different depths on the same data set, where the horizontal coordinate describes the time period of the event, and the vertical ordinate describes the number of decision tree nodes after the data set is decomposed. These nodes correspond to different warning labels in the paper.

Figure 3 and Figure 4 respectively describe the decision results when the depths are “1” and “2” in the decision tree model. It is easy to see that a decision curve with a depth of “1” can hardly fit the data set and is not a good prediction. When the depth is “2”, relatively speaking, it can better fit the data set compared with the depth of “1”, and the fitting curve can capture the change of data in the second time window, but the data are difficult to be captured with the time changes and still fail to fit the dataset very well.

To better fit the training data and avoid overfitting of the model, the depths of “3”, “4”, “5” and “6” are selected in the paper. The specific fitting results are shown in Figure 5 to Figure 8.
Figure 6. Result with a model depth of “2”

Figure 7. Result with a model depth of “1”

Figure 8. Figure 8 Result with a model depth of “2”

Figure 5 shows the forecast results with a decision tree depth of “3”. It can be seen that the fitting curve fits the training data well. More than that, over time the forecast results show a balanced change. When the depth is greater than 3, the fitting curve has large fluctuations. Although it can fit the training set well, the model tends to be more complex due to its large fluctuations, which will eventually lead to overfitting of the model. For this reason, this paper finally determines the depth as “3”.

4.3. Analysis of decision cases

In order to illustrate the feasibility of this method, the “Qingdao Prawn Event” is used as an example here. The data distribution of Qingdao prawn is shown in Table 1.

| Table 1. Data Distribution of Qingdao Prawn with Sky-high Price |
|---------------------------------------------------------------|
|                  | Oct. 5 | Oct. 6 | Oct. 7 | Oct. 8 |
| Baidu search     | 795    | 10883  | 73114  | 105015 |
| BBS              | 7092   | 5025   | 18519  | 13599  |
| MicroBlog        | 35724  | 19799  | 87130  | 46419  |
According to the data in Table 1, it is possible to calculate the heat and business of the incident of “Qingdao Prawns with Sky-high Price” happened on October 28. Among them, the heat value is calculated according to the third chapter of this report. The information gains of hotspot words are calculated on the basis of keywords such as “Qingdao Prawns”, “Prawns with Sky-high Price” and “Rip Off”. At the same time, the growth rate of the report is calculated as a measure index of the business of the events described in this paper.

The reporting and comments of the incident involve the life topics such as “Rip Off”, “active response of relevant departments” and “properly handling”. The global tendency distribution is shown in Table 2.

| Negative tendency | Neutral | Positive tendency |
|-------------------|---------|------------------|
| 0.85              | 0.1     | 0.05             |

From the data in the table, it can be seen that netizens generally hold negative emotional tendency against the incident of “Qingdao Prawns with Sky-high Price”.

According to the data provided in the “Blue Paper of Public Opinion Analysis,” the “Qingdao Prawns with Sky-high Price” incident is affiliated to the “price” field. Although prices are widely concerned by the Internet users, it is not a common phenomenon and generally does not cause group events. Therefore, the sensitivity of all reported news is defined as the neutral tendency.

After the above preprocessing, the report on this incident was taken from October 6, 2015 to October 8, 2015, to study the early warning of the incident. The following uses October 6 as an example to illustrate the building process of the decision model. To illustrate the method of this paper, the data of Table 1 is cleaned in advance, to remove the duplicate data, graphic marks and other records. Finally the information is sorted by relevance, and the most relevant 1024 records are taken as experimental data to make prediction.

Firstly, Table 3 is generated based on the tendency segmentation table.

| Counts | Tendency | Heat | Burstiness | Sensitivity | Warning Grade |
|--------|----------|------|------------|-------------|---------------|
| 276    | Negative | High | High       | Middle      | Red           |
| 364    | Negative | High | Middle     | Middle      | Red           |
| 292    | Negative | Middle| High       | Middle      | Orange        |
| 92     | Neutral  | High | High       | High        | Orange        |

Similarly, the data is segmented based on heat, sensitivity and burstiness, and then the information entropy needs to be calculated in the prediction process.

1) Calculating the information entropy required for classification in a given sample

As shown in Figure 3, the entire data set is divided into two categories: “red” and “orange”. Where, $S_1$(red) = 640 , $S_2$(orange) = 384 ; the probability of $S_1$ is 0.625, and $S_2$ as 0.375. The following can be obtained according to the formula of information entropy:

$$
I(S_1, S_2) = I(640,384) \\
= -p_1 \log p_1 - p_2 \log p_2 \\
= -(p_1 \log p_1 + p_2 \log p_2) \\
= 0.9544
$$

(5)

2) Information entropy of computational features

Here only the tendency index is cited as an example to illustrate the method.
The tendency is divided into two groups: “negative” and “neutral”. The “negative” probability was 0.91; the ratio of having the warning grade as “red” in the negative group is \( \frac{640}{932} = 0.6867 \), and the “orange” as \( \frac{292}{932} = 0.3133 \).

\[
I(S_1, S_2) = I(640, 292) = 0.8969 
\]

The ratio of having the warning grade as “red” in the “neutral” group is 0, and the “orange” as 1.

\[
I(S_1, S_2) = I(0, 92) = 0 
\]

By the same method, the maximum information gain is selected according to the information entropy of the other three features as the division data set of root nodes. For the first time, this report divides the data set by the tendency, and finally October 6 is forecasted as the day with a red warning.

5. Conclusion
Taking the existing network public opinion cases as the research object, the paper analyzes key factors affecting the network public opinions by the machine learning method and the big data technology statistical technology, and builds the early warning system of public opinions based on this. The decision tree model is also built to realize analysis and early warning, and finally the "Qingdao Sky-price Prawns" incident is cited to illustrate the feasibility of the early warning by classification proposed in this paper.

References
[1] Nkpa N K 1996 Rumor mongering in war time. The journal of Social Psychology, no. 1, pp. 27-35.
[2] Tommaso B, Lorenzo M 2014 An epistemological analysis of gossip and gossip-based knowledge. Synthese, pp. 1-31.
[3] Nam P. Nguyen, Guanhua Yan, My T. Thai, Stephan Eidenbenz 2012 Containment of Misinformation Spread in Online Social Networks// In: Proceedings of the 3rd Annual ACM Web Science Conference, pp. 213-222.
[4] Zhao L J, Wang Q, Cheng J J, Chen Y C, Wang J J, Huang W 2011 Rumor spreading model with consideration of forgetting mechanism: A case of online blogging Live, Journal. Physica A, vol. 390, pp. 2619-2625.
[5] Zhang P, Lan Y X, Li H Q, et al 2016 Research on Comprehensive Classification of Online Rumors Based on Cognitive Process. Library and Information, no. 4, pp. 8-15.
[6] Dong J F 2014 Research on the Early Warning of Public Opinion based on Web Mining. Modern Information, vol. 34, no. 2, pp. 43-51.
[7] Liu Y 2012 Building of Warning Index System of Network Public Opinions Based on Triangular Fuzzy Numbers. Statistics and Decision, vol. 350, no. 2, pp. 12-15.
[8] Lin C 2015 Research on the Building of Public Opinion Index System Based on the Formation Process of Network Public Opinion. Information Science, vol. 33, no. 1, pp. 146-161.
[9] Cao J X, Chen G J, Wu J L, et al 2016 Mining of Social Opinion Leaders based on Multi-Dimensional Feature Analysis. Acta Electronica Sinica, vol. 44, no. 4, pp. 898-905.
[10] Liu C. J, Liu Z, Chai Y J, et al 2016 Study on Visual Calculation Method of Social Group Emotion based on Weibo Text Data Analysis. Acta Scientiarum Naturalium Universitatis Pekinensis (Science and Technology), vol. 52, no. 1, pp. 178-186.
[11] Zeng R X 2010 The Building of Early Warning Index System of Network Public Opinions Emergency. Information Studies: Theory & Application, no. 1, pp. 77-80.
[12] Qu Z K, Lan Y X, Liu Y, et al 2016 Research on Early Warning of Group Events Based on Synthetic Analysis of Network Public Opinion Information. Electronic Government, no. 11, pp. 20-30.