Intelligent Identification and Classification of Product Quality Incidents Based on Rules and Statistics

Yingcheng Xu¹,a, Haiyan Wang²,b, Wei Feng³,c and Xiaojun Zheng⁴,d

¹Quality Research Branch, China National Institute of Standardization, Beijing 100191, China,
²School of Management and E-Business, Zhejiang Gongshang University, Hangzhou 310018, China,
³Quality Research Branch, China National Institute of Standardization, Beijing 100191, China, * Corresponding author
⁴Sichuan Institute of Product Quality Supervision and Inspection, Sichuan, 610001, China

¹15810265898@126.com, ²sqshan8@163.com, ³springblue410@126.com, ⁴priskcn@163.com

Abstract. Product quality and safety incidents are closely related to the lives of ordinary people, and have always been the focus of all countries' governments. Using related technologies and methods, how to intelligently identify and classify product quality and safety incidents, and find out the evolution and spread of events at the first time, and deal with them quickly, has always been a difficult problem for product quality and safety supervision departments to solve. Based on rules and statistics, a multi-layer text classifier is built to identify and classify product quality incidents. The first layer implements rough segmentation of the news concerning product quality incidents based on rule classifier. The second layer matches the news in a level-wise manner from the top down by calculating the text similarity until the corresponding subclass is found. The experiments show that the proposed classifier can effectively identify the product quality incidents in Web news.

1. Introduction

In recent years, China has constantly undergone a range of product quality incidents, which caused a huge loss to both physical and property safety of consumers. Spurred by the rising Internet use, the Web has become the main way for disseminating the information about product quality incidents. To prevent the product quality incidents from springing up throughout the systems, regions and industries, it is necessary to monitor their information dissemination closely and identify and classify such incidents promptly. In general text classification methods, all classes share a classifier or each class sets a classifier[1-3]. Since the classes are all at the same level, it will be very time-consuming to extract the model if there are too many classes. Moreover, when classifying new documents, it must take a complicated process to compare them with all class models. In this context, a multi-layer text classification method based on rules and statistics is proposed to identify and classify large quantities of text in an accurate and fast way.
2. Steps for identification and classification
To classify the text of the Web news regarding product quality incidents, the documents, including the words and their weight in the documents, shall be represented by normalized vectors. The main factors like word frequency and word position are considered for calculating the words’ weight in the documents. Starting with the root node, the document vectors are then matched with various models from the top down according to each type hierarchy. In other words, their textual similarity is calculated until the appropriate subclass corresponding to the leaf node is found[4-5]. The specific process includes the following four steps:

(1) Preprocess the Web news to be classified
   It mainly includes the steps such as HTML parsing, content extraction, Chinese word segmentation, part-of-speech tagging and deletion of stop words. Let the corresponding vector be $d = (t_1, w_1, \ldots, t_m, w_m)$.

(2) Use the rule classifier to roughly classify the incidents
   The logical operations such as “AND”, “OR” and “NOT” are used to match the defined rule template with the entries combined in different dimensions. If the text belongs to a certain type of product quality incidents, go to Step 3; otherwise, go to Step 4.

(3) Use the statistical classifier to subdivide the class of incidents and topics.
   Calculate the text similarity. If the text is not a leaf node, execute the step repeatedly:
   i. Suppose the node X has n child nodes, i.e. $V_1, V_2, \ldots, V_n$, and the class model corresponding the node $V_i$ is $C_i = (t_1, w_1, \ldots, t_n, w_n)$, $i = 1, 2, \ldots, n$.
   ii. Calculate the similarity between documents and classes successively. The similarity is represented by $sim(C_i, d)$.
   iii. Find the class with the greatest similarity to the vector $d$. The text to be classified is put under the class.

(4) The end of classification.

3. Formatting the text
3.1. Rule-based identification and classification
The analysis results of product quality incidents indicate that the combination rules of product information and quality information play an effective role in identifying most of product quality incidents. The classification rules of product quality incidents are formulated by taking into account the four elements including product information, quality information, injury information and countermeasures. The specific rules are shown in Table1.

| Class                  | Rule                                                                 |
|------------------------|----------------------------------------------------------------------|
| Product information    | Laser pointer, lamp, cabinet, furniture, refrigerator, color TV, washing machine, electric blanket, water heater, hair dryer, children's diapers, children's sunglasses, toys, building blocks, magnetic beads, buck ball, magnets, plasticine, crystal clay, light clay, scooter... |
| Injury information     | Injury, death, cancer, harm, exceed standard, accident, death, die, explosion, burst, electric leakage, electric shock, on fire, fire, spontaneous combustion, mistakenly swallowed, poisoning, asphyxia, poor quality, fault, burns, scratches, cuts, crushed, scald, fire burn, radiation... |
| Countermeasures        | Investigation, recall, seal up, removal, fine, interview, criminal law... |
3.2. Statistics-based identification and classification

The second-layer classifier applies the vector distance function method for classification. When the classifier matches an object with a class, it works out a value to indicate how similar the object is to the class. If their similarity is greater than the corresponding threshold, the object is considered as belonging to the class.

According to the incident-based model for vector space representation, the preprocessed candidate characteristic words are placed into three sub-vectors, namely time information, abstract of incident and full text of incident[6]. Therefore, a news report on product quality incident can be represented as:

\[ D_i = (D_{time,i}, D_{abstract,i}, D_{content,i}) \] (1)

In the equation, \( D_i \) represents the vector of the \( i \)th news report, \( D_{time,i} \) represents its component for time characteristics, \( D_{abstract,i} \) represents its component for abstract of incident, and \( D_{content,i} \) represents its component for characteristic words of full text.

The similarity between vectors is usually calculated by means of cosine similarity and sum of weights[7]. In a multi-vector representation model, the text \( d_1 \) is written as \( d_1 = (\vec{d}_{11}, \vec{d}_{12}, \vec{d}_{13}) \), while the text \( d_2 \) is written as \( d_2 = (\vec{d}_{21}, \vec{d}_{22}, \vec{d}_{23}) \). Suppose there are \( k \) methods for similarity calculation \( (M_1, M_2, \cdots, M_k) \). The \( m \)th sub-vector between the texts \( d_1 \) and \( d_2 \) of the Web news of product quality incidents uses the \( j \)th method to calculate the similarity. The calculation process can be represented by the formula below:

\[ Sim_j(\vec{d}_{1w}, \vec{d}_{2w}) = M_j(\vec{d}_{1w}, \vec{d}_{2w}) (0 < m \leq 3; 0 < j < k) \] (2)

(1) Calculate the similarity of abstract and content characteristics

Given there are two \( n \)-dimensional vectors, \( d_1 \) and \( d_2 \), their similarity in the vector model can be quantitatively denoted and calculated by using the cosine of the angle between them. For example, the similarity of the components for characteristic words of abstract can be calculated by the formula below:

\[ Sim(t_{i4}, t_{j4}) = \frac{d_i \cdot d_j}{|d_i| \times |d_j|} = \frac{\sum_{m=1}^{n} (W_{m,i} \times W_{m,j})}{\sqrt{\sum_{m=1}^{n} W_{m,i}^2} \times \sqrt{\sum_{m=1}^{n} W_{m,j}^2}} \] (3)

In the equation, \( d_i \) is the characteristic vector of test text, \( |d_i| \) and \( |d_j| \) is the norm of vector, and \( m \) is the dimension of characteristic vector.

The similarity of sub-vectors of main body is also calculated in the same way.

(2) Calculate the time similarity

The time distance between the document \( i \) and the latest document \( j \) in the training set is defined as follows:

\[ dis_{time}(i, j) = \min\{|time_i - time_j|\} \] (4)

After the time distance is obtained, the similarity between documents is calculated by the following formula[8].

\[ sim_{time} = \begin{cases} 1 & dis < 2 \\ -0.25 & dis \geq 2 \end{cases} \] (5)

(3) Calculate the global similarity
In general, linear combination and classifier are the two major methods for combining the similarity of 3D sub-vectors to form the similarity between two texts. The present study intends to use linear combination. According to the multi-vector text representation model above, the distance between documents is calculated by the following formula:

\[
Sim(d_1, d_2) = \alpha_1 \text{sim}(\overrightarrow{a}_{\text{abstract}}, \overrightarrow{a}_{\text{abstract}}) + \alpha_2 \text{sim}(\overrightarrow{a}_{\text{content}}, \overrightarrow{a}_{\text{content}}) - \beta \text{sim}(\overrightarrow{a}_{\text{time}}, \overrightarrow{a}_{\text{time}})
\]

(6)

In the formula, \(\alpha_i\) represents the weighting factor of the sub-vector for textual characteristics that describes the influence of abstract and full text on overall similarity of content, respectively; \(\beta\) represents the influence of incident distance on similarity that can be determined by training or genetic algorithms.

4. Experimental process and analysis

Due to the lack of a public data set regarding product quality incidents, an experiment was conducted to analyze the algorithms by using Web crawlers to capture relevant incidents from NetEase News, including the Poisonous School Uniform Incident, the Kumho Tire Incident, the Da Vinci Furniture Incident, and the Crocs Shoes Incident. The distribution of training set and test set of the incidents is shown in Table 2:

| Name of news topics     | Time                  | Training set | Test set |
|-------------------------|-----------------------|--------------|----------|
| Poisonous School Uniform| 2016-3-15 to 2016-4-15| 50           | 211      |
| Kumho Tire              | 2017-3-15 to 2017-4-15| 50           | 46       |
| Da Vinci Furniture      | 2015-4-16 to 2015-5-16| 50           | 71       |
| Crocs Shoes             | 2018-09-11 to 2018-10-11| 50          | 365      |
| Irrelevant incidents    | 2017-3-15 to 2017-5-16| 200          |          |

4.1. Experimental results of rough segmentation

When the incidents were identified at the first layer, the text of the Web news regarding product quality incidents was classified in accordance with the customized rule templates. A total of 200 irrelevant news reports were selected as disturbances. The results showed that 99% of the irrelevant news reports were excluded. The F values of the other four incidents after classification were 0.989, 0.986, 0.981 and 0.973, respectively. The findings suggest that the first-layer classifier can accurately identify whether the news is related to a product quality incident and classify its product field.

4.2. Experimental results of incident identification

(1) Select characteristic number

Given that a different number of characteristic items have a great impact on the classification results and high-dimensional characteristic items impose significant costs on the experiment, the number of characteristic items was carefully selected by test. The relationship between the number of characteristic items and the classification results after training is shown in Figure 1.

The figure above shows that when the number of characteristic items ranges from 30% to 40%, the performance of the classifier can reach a stable value. Given the amount of computation, the number of characteristic items is finally set at 30%.

(2) Weight of text similarity

Different weight ratios of abstract to full text (ranging from 1 to 8) were set for test. Experimental results show that when the weight remains constant, the F1 value of classification effect of the four
incidents is 84.2%, 83.2%, 78.5% and 85.4%, respectively. The F1 value increases with the rise of the ratio of abstract to full text and reaches its peaks at 4:1, 5:1 and 6:1, respectively. A comprehensive comparison indicates that the classification accuracy of all four incidents is close to 90% when the ratio is 5:1. Therefore, the weighting factor for the ratio of abstract to full text ($\alpha$) is set at 5:1, as shown in Figure 2.

![Figure 1](image1.png)

**Figure 1.** The influence of different characteristic vectors on incident identification

![Figure 2](image2.png)

**Figure 2.** The influence of different weight ratios on classification effect

### 4.3. Experimental results of topic classification

| Incident class          | Description | Countermeasures, Impact | Cause analysis | Others | Total |
|-------------------------|-------------|-------------------------|----------------|--------|-------|
| Poisonous School Uniform| Training set 15 | 10 | 10 | 10 | 5 | 50 |
|                         | Testing set 78 | 54 | 58 | 12 | 9 | 211 |
| Kumho Tire              | Training set 10 | 15 | 10 | 10 | 5 | 50 |
|                         | Testing set 8 | 23 | 8 | 5 | 2 | 46 |
| Da Vinci Furniture      | Training set 15 | 10 | 10 | 10 | 5 | 50 |
|                         | Testing set 28 | 15 | 14 | 12 | 2 | 71 |
In the stage of text processing, the four incidents were divided into six topics by manual sorting. The specific classification is shown in Table 3.

Based on the two experiments above, the candidate words with the top 30% of characteristic weight were selected as characteristic words. During similarity calculation, the weight ratio of three dimension-vectors including time information, abstract and full text was $\beta: \alpha_1: \alpha_2 = 1:5:1$. A multi-layer classifier was used to calculate the similarity for identifying and classifying incidents. The experimental results are shown in Fig. 4-19. As revealed by the experimental results, the recall ratio, precision ratio and F1 value under each topic all exceed 85%, suggesting that the classifier plays an effective role in classification.

5. Conclusion
A similarity calculation method for matching multi-dimensional vector space is set up based on incident-oriented text representation of vector space. A multi-layer classification method based on rules and statistics is proposed to classify the product quality incidents. Finally, the four typical incidents including the Poisonous School Uniform Incident, the Kumho Tire Incident, the Da Vinci Furniture Incident, and the Crocs Shoes Incident were selected to perform an experimental analysis of the identification and classification of product quality incidents. The results indicate that the proposed classifier is effective and feasible.

Acknowledgments
We would like to acknowledge that this Study is supported and funded by the Market Supervision Technology Assurance Project under Grant No. 2020YJ043, the National Science Foundation of China under Grant No. 91646122, 91746202, the Basic Scientific Research Business Projects 552018Y-5927-2018.

References
[1] R E Schapire. BoosTexter: A Boosting based System for Text Categorization . Machine Learning, 1999, 39 (223) :135-168
[2] Du A N, Fang B X. Novel Approaches for Web Filtering Based on User Interest Focusing Degree . International Journal of Innovative Computing Information and Control, 2008. 4(4): 1325-1334
[3] Jyoti Prakash Singh, Yogesh K. Dwivedi, Nripendra P. Rana. Event classification and location prediction from tweets during disasters. Annals of Operations Research, 2019, 283 (3):737-757.
[4] Yao-Tsung Chen, Meng Chang Chen. Using chi-square statistics to measure similarities for text categorization. Expert Systems with Applications.2011, 38 :3085–3090
[5] Hora M., Baopuji H., Rth A.V. Safety hazard and time to recall: The role of recall strategy, product defect type, and supply chain player in the U.S. toy industry. Journal of Operations Management, 2011,29(7-8):766-777
[6] Zhang Z Z, Zhou S G, Wang Z Y, Shen Z. A geometric growth model interpolating between regular and small-world networks. J Phys A: Math Theo, 2007, 40: 11863-11876
[7] S. Rajeshbabu,B.V. Manikandan. Detection and classification of power quality events by expert system using analytic hierarchy method. Cognitive Systems Research,2018,52:729-740
[8] Shengyi Jiang, Guansong Pang, Meiling Wu, Limin Kuang. An improved K-nearest-neighbor algorithm for text categorization. Expert Systems with Applications.2012(13):1503–1509