PROCEEDINGS OF SCIENCE

Quality and Safety News Topic Tracking Algorithm Based on Improved K-Nearest Neighbor

Yingcheng Xu[1]
China National Institute of Standardization
No.4 Zhi Chun Road, Haidian District, Beijing, China
E-mail: yingcheng_xu@126.com

Xiuli Ning [2]
China National Institute of Standardization
No.4 Zhi Chun Road, Haidian District, Beijing, China
E-mail: springblue410@126.com

Xiaohong Gao[3]
China National Institute of Standardization
No.4 Zhi Chun Road, Haidian District, Beijing, China
E-mail: buaaseasky@163.com

Fan Yu
China National Institute of Standardization
No.4 Zhi Chun Road, Haidian District, Beijing, China
E-mail: sqshan8@163.com

According to the characteristics of deficiency for topic tracking process in initial sample, this paper chooses K-Nearest Neighbor(KNN) algorithm as the method of event and the topic tracking, and takes into account the correlation of product quality and safety WEB news to improve the KNN algorithm. The "hole shoes" event as an example is to verify the algorithm. The result shows that the proposed model is effective and feasible.

ISCC 2015
18-19, December, 2015
Guangzhou, China

[1]Speaker
[2]This research is supported and funded by the National Science Foundation of China under Grant No.71301152, the Science and Technology Support Program under Grants No. 2013BAK04B04, and and the Basic Scientific Research Business Projects 552013Y-3063.
[3]Corresponding Author

PoS(ISCC2015)018

1. Introduction

In recent years, the events of our country's product quality and safety appear frequently and severely, the "MEITAI recall" event which is caused by small detachable parts of Chinese toys leading to choking in 2007, the "baby's talcum powder containing asbestos carcinogen" event in 2009, the "bisphenol A in baby's feeding bottle" event in 2011, and the "poisonous capsule" event in 2012 etc. Product quality and safety relate to our health and safety, economic healthy development and social harmony and stability. Maintaining people's livelihood is not only the top priority of consumer concern, but also the focus of public opinion and governments concern. If the problem of product quality and safety can not be found and processed, there may be groups' injury and systemic social risk easily, and then evolve into public crisis. Therefore, monitoring product quality and safety closely, discovering significant risk which may cause systemic social issues timely, and avoiding systemic and regional product quality and safety event are the urgent priorities related to the healthy development of the national economy and the social harmony. Therefore, it is important to monitor for product safety network information, find hot and sensitive topic timely, and discriminate authenticity for network spread information and public opinion.

2. Literature Review

Topic detection and tracking (TDT) is the basis of network public opinion analysis. It is used to detect new topic automatically and track known topic dynamically. TDT is the premise and basis of researching the evolution and trend of product quality and safety event.

Foreign scholars do extensively study on topic detection. Brants (2004) applied TF-IDF method based on incremental to event detection and developed the relative system[1]. Sung (2007) studied the similarity calculation method between place names from the point of names level, names location, relations of names and time, and then, Sung applied the method to topic detection. Chen (2007) used multidimensional sentences to cluster in order to create topic by extracting the hot topic word[2]. Cataldi (2010) executed the topic detection in Twitter based on evaluation of timing sequence and social vocabulary[3]. Compared with the traditional topic retrieval, Cataldi introduced the social relationships between users calculated by PageRank into topic model; and therefore, the accuracy is improved. In public opinion analysis, the task of topic tracking is to track follow-up report of the known topic by prior topics model and history topic category set. Meanwhile, relative researchers have tried to make the TDT technology extend to social media, and Cheng (2007) proposed a kind of topic excavation model based on customer relations[4]. Xu (2014) gave three topic tracking models: a static topic model BSTM and two dynamic topic models BDTM-I, and BDTM-II by using Bayesian belief network[5]. The hidden topic detection algorithm based on related model retrieval technology was proposed by Shi (2012)[6]. Kumaran (2004), Nallapati (2004) introduced the natural language processing technology into the topic detection, and it is verified that nature language technology can improve topic detection quality effectively[7-8]. Satoshi (2004) proposed the method that uses finite mixture model to track topics trends dynamically[9]. Zhang (2011) researched information system about hot topic detection and trend tracking for community issues answer system[10]. Zhu (2008) integrated the relevance of words and user to excavate the network forum topic[11],

the model integrates topics found and new event found and topic tracking, and the model can analyze topics trends timely and dynamically.

For the topic detection and tracking in terms of product quality and safety events, there are some blanks as follows:

① the exiting TDT studies are mainly for characteristics expression models and related metrics model of topic and event information. However, it is necessary to integrate the semantic method and timing change analysis method for product quality and safety events, but there are fewer research reports in this field;

② the existing TDT topic model belongs to general models. There is lack of proprietary model that can realize online testing and tracking for product quality and safety events.

3. Improvement of the KNN Topic Tracking Algorithm

Generally, there are some correlations between the news reports and the same event. Therefore, considering the reports sequence and the contents correlation, we can construct classifier by introducing NFL to KNN in order to track news topics. NFL is a novel pattern recognition classification method which is put forward by Stan.Z.Li etc. NFL is also used for voice classification and face recognition. It can reach better classification results by using sequential relationship and correlation among sample points.

The news report about each focus of a certain event is a point of feature space. With the development of the events evolution, the corresponding point will remain a continuous locus at feature space. The corresponding point locus of the feature space is seen as straight line at the neighboring news report. Calculate all feature lines of positive feature samples and negative feature samples respectively according to NFL based on the facts demonstrated, and find the k nearest neighbor characteristic lines apart from test samples. Compare the average distance difference between the positive feature samples and negative feature samples among the k nearest neighbor characteristic lines, if the difference is more than threshold θ, and then the test samples belong to event category, but not belong to otherwise.

The detail classification steps are as follows:

Step1: pre-treat at first and then make the training semi-structured HTML documents resolved as containing useful information document only; and then, assort with the text document and remove the stopwords;

Step2: apply the TF-IDF formula normalized word frequency to obtain vector space representation of the training documents;

Step3: take some samples of the events tracked as positive training samples category C_p, and the remaining training samples are negative training samples category C_N. It can get S subclasses representative points $(f_1, f_2, \ldots f_s)$ by executing to all of training samples in C_N on subclass representative points initialization select by applying for density function method, and

then take these points as the negative training samples representative points. Therefore, C_p and C_N represent $F_P = \{f_p \quad 0 < i \quad N_p\}$ and $F_n = \{f_n \quad 0 < i \quad N_s\}$ feature point set respectively;

Sep4: resolve all of the feature lines in feature space C_p and C_N, that is,

$S_P = \left\{\overline{f_i^p f_j^p} \quad 0 < i, j \ 9N_p, i \quad j\right\}$ and $S_N = \left\{\overline{f_i^N f_j^N} \quad 0 < i, j \ 9N_s, i \quad j\right\}$;

Step5: when the texts x to be classified come, firstly, pre-treat the texts x and determine the text vector representation f_x, and then calculate k feature lines nearest by f_x. Given that k_p belongs to positive category and k_N belongs to negative category in k nearest lines, it is necessary to note that $k_p + k_N = k$;

Step6: calculate the average distance difference value between C_p and C_N in k nearest feature lines. The calculation formula is as follow:

$$r(f_x, k, T) = \frac{1}{k_p} \sum_{\overline{f_i^p f_j^p} \in P_{K_P}} D(f_x, \overline{f_i^p f_j^p}) - \frac{1}{k_N} \sum_{\overline{f_i^N f_j^N} \in N_{K_N}} D(f_x, \overline{f_i^N f_j^N}) \tag{3.1}$$

Where, P_{K_P} and N_{K_N} are samples sets that belong to C_p and C_N in k nearest feature lines;

Step7: if the average distance difference value is more than threshold θ, thus the data in the test samples belong to tracking event, not vice versa. Generally, the initial value of θ is 0.

There are two main issues to be overcome on the KNN algorithm improvement based on NFL:

(1) considering the relevance and sequence of the news event, introduce NFL to KNN and calculate the k feature lines nearest by test samples;

(2) generally, negative samples are more than positive samples in tracking news event. In order to reduce calculation load and the negative samples, we take the representative samples of the negative reference samples as the new negative reference sample by calculating the negative sample density function.

4. Experiment and Analysis of Topic Tracking

4.1 Experiment Design

There is no common data set in product quality and safety field currently. This paper excavates web pages related to product quality and safety by web crawlers. The problem events contain 415 items about “hole shoes” between 2012-09-11 to 2012-10-11, 500 news pages and other events pages; and then it is needed to pre-process and manually tag the news pages in the identified "hole shoes" event. Firstly, the experiment selects 50 texts as the training set to train the improved KNN classifier in the second layer classification; secondly, for the remaining test text, exclude the news which are unconcerned to product quality and safety; lastly, track events and topics based on improved KNN algorithm. The specific experimental process design is shown in Fig. 1:

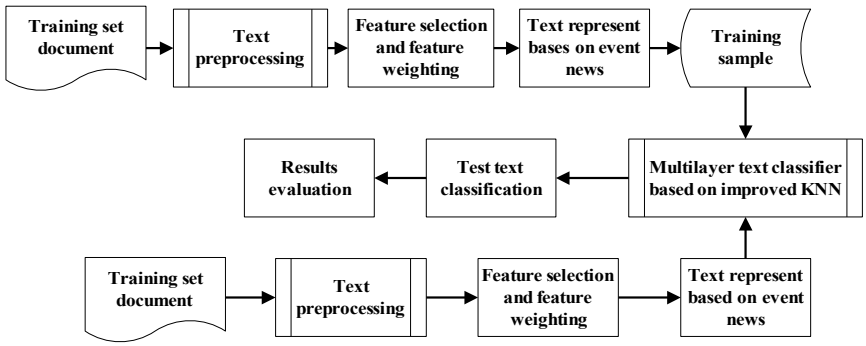

Figure 1: Experimental Process Design

4.2 Selecting the K

For the improved KNN classification algorithm, the effect of different K values on system performance is compared by experiments. This paper selects eight different values of parameter K: 1, 3, 5, 7, 9, 11, 13, 15, and tracking system performance of each event at different K-values is shown in Fig. 2:

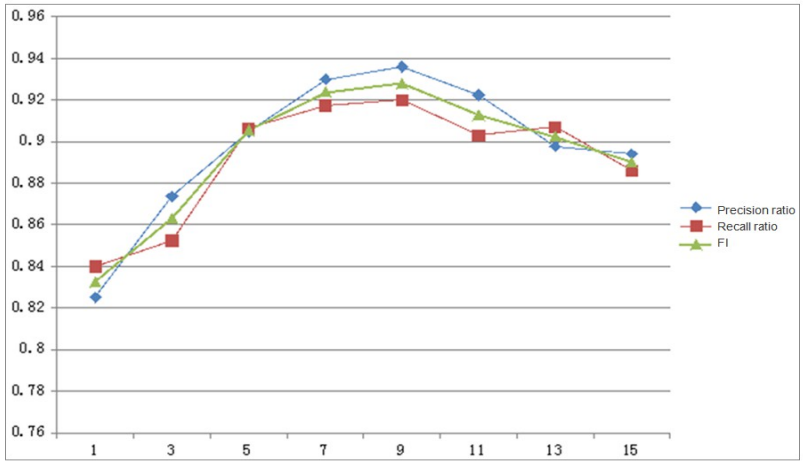

Figure 2: System Tracking Algorithm Performance for Different K

It can be seen from the experimental results that classifiers effect improves with the K rising from 1-9. The tracking system achieves the best classification results when the K takes 9. With the subsequent increase of K, classification quality doesn't improve, but declines, and therefore, K is taken as 9 in the tracking process of "melamine" event.

4.3 Comparison of the Classification Performance

Compared multilayer text classifier with traditional KNN and improved KNN text classifier by experiment, experimental results are shown in Table 1.

It can be seen from the experimental results that the classification results have slightly improved by improved KNN multilayer classifier, especially in the topic of "event influences" and "cause analysis". It can be seen that improved KNN algorithm has a good effect on the text classification with less training set.

Classification method	Event recognition	Topics category				
		Event description	Disposal measures	Event influences	Cause analysis	Others
Multilayer text classifier	91.2%	92.91%	91.43%	88.06%	89.32%	77.05%
Multilayer text classifier based on traditional KNN	91.34%	91.75%	92.1%	89.1%	88.2%	80.63%
Multilayer text classifier based on improved KNN	92.8%	93.14%	92.4%	90.45%	91.96%	82%

Table 1: Comparison of the Improved KNN Tracking Performance

5. Conclusion

In order to solve the problem of bad classification performance with less training set in the early stage, the second layer classification method is improved based on multi-classifier in the product quality and safety event in this paper. Two improvements are considered in K nearest neighbor classification method. Firstly, take into account relevant news and time relevance and introduce the nearest feature line to the KNN method, and then seek the nearest K feature lines from text samples; secondly, take into account that the negative samples will be more than positive examples generally in tracking some news and then reduce negative samples number by calculating the counter-example density function and select some typical samples among counter-examples sample as new counter-examples reference sample, and therefore, the amount of calculation is reduced; finally, the improved tracking algorithm is verified by "hole shoes" event that the new method is effective for tracking WEB news topic of product quality and safety events.

References

[1] T. Brants , F. Chen, A. Farahat. *A System for New Event Detection*. In Proceedings of the 26th Annual International ACM SIGIR Conference. New York, NY, USA: ACM Press, 2003, 330-337

[2] C. C. Chen, Y. T. Chen, and M. C. Chen, *An aging theory for event life-cycle modeling*. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 2007, 37(2): 237-248.

[3] M. Cataldi, D. L. Caro, C. Schifanella. *Emerging topic detection on Twitter based on temporal and social terms evaluation*. In Proceedings of the Tenth International Workshop on Multimedia Data Mining. New York, NY, USA: ACM Press, 2010, 1-10.

[4] V. Cheng, C. Li. *Topic Detection via Participation Using Markov Logic Network*. In Third International IEEE Conference on Signal-Image Technologies and Internet-Based System. New York, NY, USA: ACM Press, 2007, 85-91.

[5] J.M, Xu, S.F. Wu, Y. Hong. Topic tracking with Bayesian belief network. *Optik - International Journal for Light and Electron Optics*, 2014,125(9): 2164-2169

[6] K. S. Shi , L. M. Li. *A Close-to-linear Topic Detection Algorithm using Relative Entropy based Relevance Model and Inverted Indices Retrieval*. International Journal of Computational Intelligence, 2012, 5(4):735-744.

[7] G. Kumaranand J. Allan. *Text classification and named entities for new event detection*. In: Proceedings of the SIGIR Conference on Research and Development in Information Retrieval. Sheffield, South Yorkshire: ACM, 2004, 297-304.

[8] R. Nallapati, A. Feng, F. C. Peng, J. Allan. *Event Threading within News Topics*. In Proceedings of the thirteenth ACM international conference on Information and knowledge management, New York, NY, USA, Washington: ACM Press, 2004, 446-453.

[9] M. Satoshi, Y. Kenji. *Tracking Dynamics of Topic Trends Using a Finite Mixture Model*. In Proceedings of tenth ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, Washington: ACM Press, 2004, 811-816.

[10] Z. F. Zhang, Q. D.Li. *QuestionHolic: Hot topic discovery and trend analysis in community question answering systems*. Expert Systems with Applications, 2011, 38(6):6848-6855.

[11] M. Zhu, W. Hu, O. Wu. *Topic detection and tracking for threaded discussion communities*. In 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008, 77-83.