An Improved KNN Algorithm for Product Quality and Safety Incident Information Tracking

Yingcheng Xu[1, a], Zhongbao Sun[2,b] and Wei Jiang[2,c *]

[1] China National Institute of Standardization, Beijing 100191, China

[2] Jiuquan Satellite Launch Center, China

[a]yingcheng_xu@126.com, [b]springblue410@126.com, [c]buaaseasky@163.com

Keywords: Web news; product quality and safety incident; KNN

Abstract: In recent years, product quality and safety event shave happened frequently, which poses a threat to the health and property safety of consumers. With the development of the Internet, Web news has become one of the important channels for information release regarding product quality and safety issues. Therefore, identifying and tracking these incidents information plays an important role in providing early warning for product quality and safety problems. This paper presents K-Nearest Neighbor(KNN) algorithm as the methods of event and the topic tracking, and takes into account the correlation of product quality and safety WEB news to improve KNN algorithm. The "hole shoes" event as an example is to verify the feasibility of the proposed algorithm.

1 Introduction

Product quality and safety is directly related to the health and safety of consumers. Therefore, close monitoring of product quality and safety in a timely manner could avoid those issues which may cause significant problems. Product quality and safety relates 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 caused groups injury and systemic social risk easily, and then evolves public crisis. Therefore, monitoring product quality and safety closely, discovering significant risk which may caused systemic social issues timely, avoiding systemic and regional product quality and safety event are the urgent priority related to the national economy healthy development and the social harmonious. Therefore, it is important to monitor for product safety network information, find hot and sensitive topic timely, discriminate authenticity for network spread information and public opinion.

2 Related researches

Topic detection and tracking (TDT) is a research area closely related to Web information flow analysis. TDT study began in 1996 as an information processing method to identify unknown topic and track media information flow. Topic tracking is designed to monitor the information flow of news reports, to find out the news reports related to the known topic, which is equivalent of a special binary classification task. 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. Lo (2002) embed feedback self learning module in topic tracking system that weaken topic drift effectively by intercepting the second threshold value by using topics updated model by follow-up reports. Yang (2002) proposed news topic tracking algorithm based on text by decision tree. Satoshi (2004) proposed the method that uses finite mixture model to track topics trends dynamically. The model integrates topics found and new event found and topic tracking. And the model can analyze topics trend timely and dynamically. Zhang (2011) researched information system about hot topic detection and trend tracking for community issues answer system. The hidden topic detection algorithm based on related model retrieval technology is proposed by Shi (2012). Yipeng Zhou (2012) proposed temporal situation

* Corresponding author. Tel.: +86-10-5881-1138

topic model that used in topic tracking according to the feature that internet information usually contains context data as release time, place, etc.

3 Deficiencies of the traditional KNN algorithm

KNN algorithm is a simple method which doesn't need specific training set. KNN algorithm determines the test sample category by judging k reference sample categories around the test samples. At the time of verification, just the reference point set of the sample is given. We determine the sample which category belongs by judging the major k nearest neighbors among the considering test sample belongs to. The details decision procedures are as follows:

For a given test sample x, we determine the k nearest samples by the measuring the distance between the two samples in all the $C_1, C_2, \cdots C_{\mathrm{m}}$ samples sets.

Where, there are $n_i (i = 1, 2, \ldots m, \sum_{i=1}^{\mathrm{m}} n_i = k)$ samples belong to C_1. The discrimination function of the sample is defined as:

$$f_i(x) = n_i, i = 1, 2, \ldots, m \tag{1}$$

Therefore, the classification decision rule is:

$$\text{IF} \quad f_j(x) = \max_i (n_i) \quad i = 1,2,3 \ldots m, \quad \text{THEN} \quad x \in C_j \tag{2}$$

In order to overcome the bias situation caused by lack of positive samples in the process of classification, the researchers of the Carnegie Mellon University (CMU) use the improved KNN algorithm, the formulas are as follows:

$$r(x, k_p, k_n, D) = \frac{1}{|\mathrm{U}_{k_p}|} \sum_{y \in \mathrm{U}_{k_p}} \cos(x,z) - \frac{1}{|\mathrm{V}_{k_n}|} \sum_{z \in \mathrm{V}_{k_n}} \cos(x,z) \tag{3}$$

$$r(x, k, D) = \frac{1}{|\mathrm{P}_k|} \sum_{y \in P_k} \cos(x,z) - \frac{1}{|\mathrm{Q}_k|} \sum_{z \in Q_k} \cos(x,z) \tag{4}$$

Where, U_{k_p} presents k_p related documents vector sets that nearest with x. V_{k_n} presents k_n unrelated documents vector sets that nearest with x. And P_k and Q_k are the positive samples sets and negative samples sets between training samples and test samples in the k closest samples.

In practical applications of Web news topic tracking events, several news reports are difficult to cover all the details about an event. However, KNN algorithm ignores sequential relationship and relevance of the events reported caused by severe dependent on the number and distribution of the sample points of the classification set. Meanwhile, traditional KNN algorithm makes a decision by comparing the test sample points' similarity with all the reference sample points' similarity. And it is necessary to note that the cost of the traditional algorithm realization is heave because the distance between every test sample point and every reference sample point should be given in the procedure of traditional algorithm realization.

4 Improvement of the KNN topic tracking algorithm

Generally, there are some correlations among the news reports to the same event. Therefore, considering the reports sequential 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 detail classification steps are as follows:

Step1: Pre-treatment at first and then make the training semi-structured HTML documents resolve as containing useful information document only. And then, assort to the text document and remove the stopwords.

Step2: Applying for 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 as $F_P = \{ f_p \quad 0 < i \le N_p \}$ and $F_n = \{ f_n \quad 0 < i \le N_s \}$ feature point set respectively.

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

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

Step5: When the texts x to be classified come. Firstly, pre-treat to 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 belong to positive category and k_N belong 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 follows:

$$r(f_x, k, T) = \frac{1}{k_p} \sum_{\overline{f_i^p f_j^p} \in P_{K_P}} D(f_x, \overline{f_i^p f_j^p}) - \frac{1}{k_N} \sum_{\overline{f_i^N f_j^N} \in N_{K_N}} D(f_x, \overline{f_i^N f_j^N}) \tag{5}$$

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.

5 Experiments and results

This paper excavates web pages related to product quality and safety by web crawlers. The problem events contains 415 items about “hole shoes” between 2012-09-11 to 2012-10-11, 500 news pages and other events page. And then it needs to pre-process and manual tag for 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, excluding the news which is unconcerned to product quality and safety. Lastly, tracking events and topics based on improved KNN algorithm. Organization of the Text.

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

Table 1 Compared of the improved KNN tracking performance

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%

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 "events effect" and "cause analysis". The F1 closes to 90%. It can be seen that improved KNN algorithm has a good effect on the text classification with less training set. It is shown in figure 1.

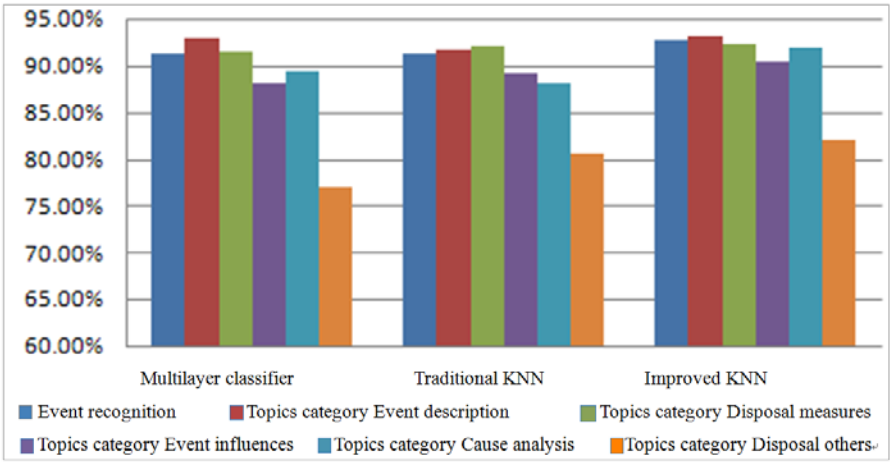

Figure 1 Compared of the improved KNN tracking performance

6 Summary

This study is focusing on Web news product quality and safety incidents tracking. Based on K-Nearest Neighbor classification algorithm, a topic-tracking model is set up. The model is verified via the design of "hole shoes"experiments, the results show the the proposed model is effective and feasible.

Acknowledgements

We would like to acknowledge that 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 Quality Inspection Project under Grant No. 201510203.

References

[1] Lo Y., Gauvain J. L.. The LIMSI Topic Tracking System for TDT 2002[C]. In: Topic Detection and Tracking Workshop. Gaithersburg, USA, 2002.

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

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

[4] Yang Y. M., Zhang J.. Topic-conditioned Novelty Detection[C]. In proceedings of the International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, Washington: ACM Press, 2002, 688-693.

[5] Yipeng Zhou, Junping Du. Theme tracking based on time and space situational model[J]. South China University of Technology (Natural Science). 2012, 40(8):82-87.

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