2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

# Research on multi-level malicious Web page Recognition based on topic Fusion and improvement of CNN

Zhou Huang
School of Computer Science and Technology,
Wuhan University of Technology
Wuhan, China
e-mail: 1239746898@qq.com

HanBing Yao
Wuhan University of Technology
Chongqing Research Institute
School of Computer Science and Technology,
Wuhan University of Technology
Wuhan, China
e-mail: 22876681@qq.com

***Abstract*—As one of the portals of the Internet, web application brings convenience to human society, but also has many problems,such as phishing attacks, malware downloads, privacy breaches, etc.In order to more accurately identify malicious web pages, reduce the risk of network attacks,this paper designs a multi-level detection model according to the different level characteristics of web pages.The model uses LDA-SECNN to learn the characteristics of web page content, uses random forest algorithm to learn the static characteristics of web page, and finally combines the two outputs to determine.Experiments show that the multi-level model can improve the accuracy of the model, and has good stability and convergence.**

***Index Terms*—Malicious web pages; LDA theme; Convolutional neural network; Random forest**

## I. Introduction

blahblah Malicious web pages refer to a collection of web pages that appear in the form of web pages, steal user privacy, install malicious programs or execute malicious code during access.In recent years, with the growth of the types and number of web pages, the frequency of malicious web pages has also increased greatly.At the same time, due to the improvement of the attacker's technical level, the content of malicious web pages is more confusing, the malicious behavior of web pages is more hidden, and it is easy to induce users to carry out unsafe operations.Malicious web page detection is a very important part of network security.

According to different malicious web page behaviors, corresponding solutions are proposed.Literature [1] uses complex recursive units to train two recursive neural networks to detect abnormal network requests. Compared with the most advanced model, it has achieved better results and can reduce the impact of feature selection.Literature [2] fully considers the relationship between word position and context in URL, uses special characters to segment URL, and then uses convolutional neural network as classifier. The experimental results show that the model has achieved good results on a large number of real data sets.Literature [3] designed a comprehensive web page association analysis model by using the technologies of topic tracking, topic anomaly discovery, web page visual similarity evaluation, web page structure analysis and URL analysis, which solved the problem of difficult unknown feature detection.Reference [4] proposed a multi-scale feature fusion detection method, which models HTTP requests from word level and character level. With the help of multi-scale feature fusion technology, it has achieved good results on real data sets.Literature [5] designed a multi-level classifier for different levels of web page features, and combined with the output of neural network and random forest for judgment. Experiments show that the multi-level model can improve the recognition rate of the model.

Through the above research, it is found that due to the complex structure of web pages, it is difficult to take into account the characteristics of different levels of web pages. The general model usually models web pages'URLs, themes, static features, and so on,Through the study of literature [5], we can see that model fusion can make up for the defects of single model and improve the overall performance of the model.Therefore, a multi-level recognition model based on topic fusion and improved convolutional neural network is proposed in this paper. Firstly, the word2vec word vector corresponding to each word of web page content is fused with LDA topic vector, and a word vector fusion algorithm is proposed according to the contribution of words to web page topic category; Then, senet is introduced into CNN to build a web page topic recognition model secnn, which can improve the performance of important features and suppress useless features, and enhance the ability of feature extraction; At the same time, the random forest classification model is used to learn the static characteristics of web pages, and finally the two outputs are synthesized for judgment.

## II. Model Design

### A. Word vector fusion algorithm

There are some problems in word2vec vector representation, such as the lack of global semantic expression and the inability

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00056

to highlight the importance of words in document categories. This vector indicates that there are problems such as the lack of global semantic expression and the inability to highlight the importance of words in the document category.The topic vector can complete the global semantic expression through the probability distribution of the words in each topic. The integration of Word2vec and LDA can better represent the text. This article uses the Word2vec fusion LDA method to improve the web topic recognition effect.First extract the URL part, content part, and JavaScript part of the web page to form an input document, then normalize the Word2vec word vector and LDA body vector corresponding to each word, and then the vector fusion vector is based on the contribution of the word to the category.The degree is weighted to highlight the importance of words in the document category.The vector fusion algorithm proposed in this paper is as follows:

(1)Use the CBOW model in Word2vec to train the corpus to obtain the word vector, d represents the dimension of the word vector, and $|P|$ is the vocabulary of the entire corpus trained with the word vector.

For the word $word_j$, its corresponding Word2vec word vector is expressed as $\vec{x}_j = (p_{j,1}, p_{j,2}, \cdots, p_{j,d})$.

(2)The topic-word matrix $\varphi$ obtained by using the Gibbs sampling method to train the LDA topic model on the corpus is shown in equation (1).

$$\varphi = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,|\varphi|} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,|\varphi|} \\ \cdots & \cdots & \cdots & \cdots \\ z_{o,1} & z_{o,2} & \cdots & z_{o,|\varphi|} \end{bmatrix} (1)$$

$|\varphi|$ represents the size of the vocabulary of the topic model, and $O$ represents the dimension of the vector, which is the same as the dimension setting of the word vector. For the word $word_j$, the corresponding topic vector is denoted as $\vec{y}_j = (z_{1,j}, z_{2,j}, \cdots, z_{o,j})$ .

(3)Reference [6] mentions the explanation of word vectors, that is, word vectors can express semantics, and the addition of word vectors can also express the combination of word meanings. Therefore, this paper adopts the vector fusion method of addition to normalize the Word2vec word vector and topic vector corresponding to each word and add them to form a new vector representation.$\|\cdot\|_2$ represents the 2-norm, and the new vector representation of the word $word_j$ is shown in formula (2).

$$v'_j = [\frac{\vec{x}_j}{\|\vec{x}_j\|} + \frac{\vec{y}_j}{\|\vec{y}_j\|}](2)$$

A document is represented by the word vector corresponding to formula (3) as $d_i = [v'_1, v'_2, \cdots, v'_m]^T$, and $m$ is the sentence length of the document.

(4)Use the graph-based TextRank algorithm to extract keywords from document $d_i$, map each word in the document as a graph node, and map the co-occurrence relationship of the words in a preset window to the edges of the nodes,initialize the initial weights$w_{ji}$ of $V_j$ nodes and randomly,According to formula (3), iteratively calculate the weight of each node until convergence, $a$represents the damping coefficient, and $nt(V)$ represents the set of adjacent nodes of node $V$.

$$s(V_j) = (1-a) + a \times \sum_{V_i \in nt(V_j)} \frac{w_{ji}}{\sum\limits_{V_t \in nt(V_i)} w_{it}} s(V_i)(3)$$

After obtaining the weight of the words in the document, in order to simplify the calculation, only the top $k$ words in the weight ranking are taken out to form the keyword set $kword = \{kd_1, kd_2, \cdots, kd_k\}$.For each document, the weight factor $g(word_j)$ of $word_j$ is obtained by formula (4), and the weight factor represents the importance of the word.

$$g(word_j) = \begin{cases} s(word_j) \times \eta_j & word_j \in kd \\ 1 & word_j \notin kd \end{cases} (4)$$

Calculate the topic weight corresponding to the keywords in the text. The topic weight reflects the distribution of words in different topics. The larger the topic weight, the greater the difference in the distribution of words in each topic, and the corresponding contribution to topic recognition The greater the degree, the calculation of the topic weight is shown in equation (5).

$$\eta_j = \frac{\sqrt{\sum\limits_{i=1}^{b} \left[ df(b_i, word_j) - \frac{df(D, word_j)}{b} \right]^2}}{b} (5)$$

Among them, $\eta_j$ represents the topic weight of the word $word_j$, $b$ represents the number of topic categories, and $df(b_i, word_j)$ represents the number of texts containing the word $word_j$ in category $b_i$,$df(D, word_j)$represents the number of texts containing the word $word_j$ in the entire corpus $D$.

The weighted word vector of the word $word_j$ is expressed as shown in equation (6).

$$v_j = g(word_j) \otimes v'_j (6)$$

Among them, $v_j$ represents the weighted word vector, the symbol $\otimes$ represents element-wise multiplication, and the weighted word vector of the document $d_i$ is expressed as: $d_i = [v_1, v_2, \cdots, v_m]$.

The document vector obtained by this algorithm can represent the text better than the general Word2ve word vector representation. Using the document vector obtained by this algorithm as the feature representation of the neural network for learning is helpful to improve the performance of the model.

### *B. LDA-SECNN Web Page Topic Recognition Model*

With the more and more extensive application of deep learning in NLP field[7][8], recurrent neural network (RNN) and CNN model have been widely used in the field of text classification. According to the research of Yin et al[9]., the training of CNN model has more advantages in text classification, so we decided to adopt CNN model.

After vector fusion, the semantic feature distribution of the document is unstable. In order to make the model pay more

attention to the features beneficial to the topic recognition task, the se module in senet is introduced into the convolution neural network, and a secnn model suitable for the topic recognition task is designed. The structure diagram of secnn model is shown in Figure 1.

The number of channels for each height of the convolution kernel output feature map is $C$, and the SE layer uses the feature map output by the convolution layer as output. First, perform a global average pooling operation on each feature channel $Y_C$ of the feature map to obtain the compressed The feature map $Z_C$ is shown in formula (7).

$$z_c = F_{sq}(Y_c) = \frac{1}{W' \times H'} \sum_{i=1}^{W'} \sum_{j=1}^{H'} Y_c(i,j) (7)$$

Where $W'$,$H'$ represent the size of the feature map,$c = \{1, C\}$.

Then through two fully connected operations, training parameters $W_1$ and $W_2$ and the ability to recognize the importance of each feature channel, highlight the important features for the topic recognition task,The normalized weight vector $S$ is obtained by the Sigmoid function as shown in equation (8).

$$S = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) (8)$$

Among them, $\delta$ represents the ReLU activation function, $\sigma$ represents the Sigmoid activation function,$W_1 \in R^{\frac{C}{r}C}$, $W_2 \in R^{\frac{C}{r}C}$,$r$ represents the number of hidden layer nodes in the middle layer.

Finally, the weight vector is weighted to the original feature map channel by channel through multiplication, and the final output $Y_c$ is obtained. The original feature is recalibrated in the channel dimension. The weighted feature map and the original feature map have the same dimension, as in formula (9) Shown.

$$Y_c = F_{scale}(Y_c, S_c) = S_c \cdot Y_c (9)$$

Among them,$S_c$ represents the weight vector corresponding to the response feature map, and the symbol $\cdot$ represents the vector product operation.

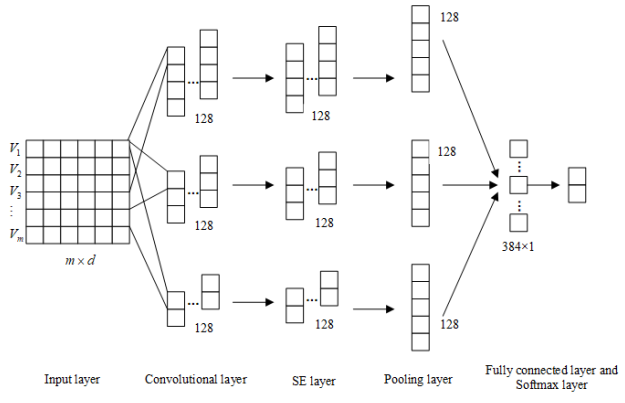


Fig. 1. LDA-SECNN structure

TABLE I

| Page features | |
|---|---|
| Total URL length | URL path part length |
| Whether to use https protocol | Are there uncommon ports |
| Does it contain '@' | The number of characters '.' |
| Number of keywords | Number of top-level domains |
| Whether to use iframe | Domain name survival time |
| URL path part length ratio | Whether to include ip address |
| URL information entropy | The number of characters '-' |
| DNS records | |

### C. Webpage Static Feature Recognition Model

The random forest algorithm is an integrated machine learning method that can be used to control overfitting. In addition, with its forest structure, the instability of individual decision trees may disappear. Therefore, this experiment uses the random forest algorithm to detect web page features. The classifier first extracts web page features and generates a training set. Then randomly select a number of training subsets, and construct a decision tree for the training subsets. Finally, the output of the decision tree is obtained, and the average value of the output is obtained to obtain the detection result of the webpage.

As shown in Table I, the article extracts 15 features of web pages, including URL string features, DNS information, and web HTML features.

### D. Threshold setting

As shown in Figure 1, the threshold $\alpha$ determines whether the web page performs the second level of detection.As shown in the formula, if the ratio of the maximum and minimum values between the normal webpage probability p1 and the malicious webpage probability p2 output by LDA-SECNN is less than $\alpha$ , then web URL features, web HTML features, and DNS information need to be extracted for the second level of detection;otherwise, the output result of LDA-SECNN will be directly judged. $\alpha$ is initialized to 1, and then the recognition accuracy rate of the multi-level detection model is output, $\alpha$ is added by 1, and then the optimal recognition accuracy rate is output and converged.

$$\begin{cases} \frac{\max(p_1,p_2)}{\min(p_1,p_2)} > \alpha, Direct judgment \\ \frac{\max(p_1,p_2)}{\min(p_1,p_2)} \leq \alpha, \text{Sec}ondary\text{detection} \end{cases}$$

## III. Experiments and Analysis

### A. Evaluation index

This article uses accuracy (A), precision (P), recall (R) and F1 value to evaluate the performance of the model. Among them, accuracy is the most important evaluation index. The precision rate indicates the proportion of the samples that are predicted to be positive, and the recall rate indicates the proportion of the positive samples that are predicted to be correct. The F1 value is the harmonic average of the recall rate and the precision rate.

TABLE II

| algorithm | A | P | R | F1 |
|---|---|---|---|---|
| CNN | 95.68 | 95.53 | 95.78 | 95.66 |
| LSTM | 95.21 | 95.90 | 94.43 | 95.16 |
| LDA-SECNN | 97.22 | 96.94 | 97.45 | 97.20 |

$$A = \frac{TP+TN}{TP+TN+FN} \times 100\%$$
$$P = \frac{TP}{TP+FR} \times 100\%$$
$$R = \frac{TP}{TP+FN} \times 100\%$$
$$F1 = \frac{2\times P\times R}{P+R} \times 100\%$$

In the above formula, TP represents the total number of positive samples predicted as positive samples, FP represents the total number of negative samples predicted as positive samples, and TN represents the total number of negative samples predicted as negative samples. , FN indicates that the negative samples are predicted, which is actually the total number of positive samples.

### *B. Experimental environment and data set source*

The programming language used in this experiment is Python3.5, the CPU is AMD Ryzen 7 4800U with Radeon Graphics 1.80 GHz, and the RAM is 16G. Part of the data of benign webpages comes from websites published by Alexa. In order to ensure the purity of benign webpages, the top 1000 websites are selected. At the same time, 20,634 malicious web pages were crawled from Phishtank, as well as a web page data set from the data science competition platform kaggle[10]. After deduplication and cleaning, the final web page data set contained 36,696 benign web pages and 35,478 malicious web pages. For LDA-SECNN, set the batch bit to 64, epoch to 20, and use 5-fold cross-validation during the experiment to ensure the stability of the model.

### *C. Result analysis*

As shown in Table II, compared with CNN and LSTM detection models, LDA-SECNN is better than the former in accuracy, recall, precision, and F1 value. Through the comparison of Figure 2, the multi-level detection model designed in this paper performs better on indicators such as accuracy. As shown in Figure 3, when the value of $\alpha$ is 219, the recognition accuracy of the multi-level detection model converges, and it is significantly better than the LDA-SECNN model and the RF model.

## IV. Conclusion

This paper integrates the content of the webpage into the word vector expression, designs a multi-level detection model combining neural network and random forest algorithm, and compares it with the single-level detection model. Experiments have proved that the multi-level detection model has better results in terms of accuracy, F1 value and other indicators. There are many types of malicious webpages. In the future, we can study the multi-classification of malicious webpages, which is helpful for targeted defense.

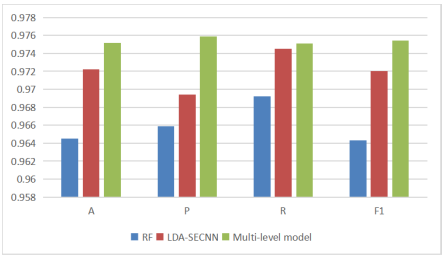


Fig. 2. Comparison chart of RF, LDA-SECNN, multi-level detection model

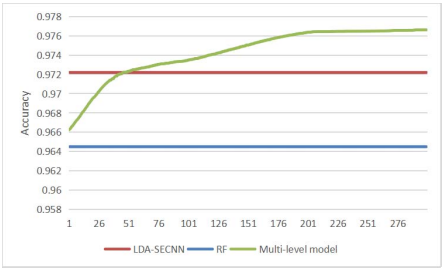


Fig. 3. Threshold $\alpha$ change graph


## Acknowledgments

The work described in this paper was supported by the Research Project of Wuhan University of Technology Chongqing Research Institute (YF2021-10).


## References


[1] Jingxi Liang, Wen Zhao, and Wei Ye. 2017. Anomaly-Based Web Attack Detection: A Deep Learning Approach. In Proceedings of the 2017 VI International Conference on Network, Communication and Computing (ICNCC 2017). Association for Computing Machinery, New York, NY, USA, 80–85. DOI:https://doi.org/10.1145/3171592.3171594

[2] Wu Haibin,Zhang Dongmei. Malicious URL detection technology based on context information[J]. Computer engineering Software,2019,40(1):63-68. DOI:10.3969/j.issn.1003-6970.2019.01.013.

[3] Senhao Wen, Zhiyuan Zhao, and Hanbing Yan. 2018. Detecting Malicious Websites in Depth through Analyzing Topics and Web-pages. In Proceedings of the 2nd International Conference on Cryptography, Security and Privacy (ICCSP 2018). Association for Computing Machinery, New York, NY, USA, 128–133. DOI:https://doi.org/10.1145/3199478.3199500

[4] WuJiahong,Zhen Guo Yang,Liu Wenyin.Malicious HTTP request detection method based on multi-scale feature fusion[J].Application Research Of Computers,2021,38:871-874+880.

[5] Zhang Shikun.Research on malicious web page detection technology based on multi-layer classifier[J].Modern computer,2020(18):64-68.

[6] ZHOU Wanting. Research on vector representation of text sentiment analysis based on word vectors[D]. Changchun: Northeast Normal University, 2019.

[7] Johnson R, Zhang T. Effective Use of Word Order for Text Categorization with Convolutional Neural Networks[J]. Eprint Arxiv, 2014.

[8] Wen Y, Zhang W, Luo R, et al. Learning text representation using recurrent convolutional neural network with highway layers[J]. 2016.

[9] Yin W, Kann K, Yu M, et al. Comparative Study of CNN and RNN for Natural Language Processing[J]. 2017.

[10] Singh A. K. (2020). Malicious and Benign Webpages Dataset. Data in brief, 32, 106304.