A platform for automatic identification of phishing URLs in mobile text messages

Xiang Xun SUN¹, Sha DAI², Yang Xu WANG¹

¹Machine Intelligence Laboratory College of Computer Science, Sichuan University, Chengdu, Sichuan 610065, China
²China Mobile (Suzhou) Software Technology Co.,Ltd, No.78,Kunlun mountain Road, High-Tech Zone,Suzhou,Jiangsu,215000,China
sunxiangxun777@163.com

Abstract. This paper constructs an automated identification platform for phishing URLs in mobile message. This platform have the ability that including preprocessing the content of message, feature extraction and automatic identification of URLs in the message, output the result. Aiming at the problem of automatic identification of phishing web site, this paper propose a phishing URLs recognition model based on neural networks and deep learning. The experimental results show that that the model has a very good judgment effect on the phishing URLs, and the accuracy rate is 98.2%. The recall rate is 96.9%.

1. Introduction

With the rapid development of the Internet, the number of Internet users in China is over 750 million, of which the proportion of mobile Internet access is up to 96.3%[1], people enjoy the mobile online shopping, social interaction, games and other services ,but at the same time, they also face the problem about disclosure important information, the huge risk of property losses, The statistics show that the current mobile Internet for the fishing site attack has gone beyond the traditional Internet[2]. The use of pseudo-base station posing as banks, electricity and other official numbers to send text messages containing phishing sites, to induce users to click and get the user's personal information, theft of accounts and other acts to bring great economic losses.

Traditional way of fishing urls detection was based on blacklist technology, which technology can not adapt to the phishing because of it’s varied form and also that can’t deal with the problem of timeliness and flexibility[3][4][5].PAN et al.[6]extracted eight characteristics of the phishing page, and then proposed a classification model based on svm algorithm, and obtained accuracy about 82%,Xiang G[7] proposed CANTINA + detection model for classification, which based on Naive Bayesian algorithm. and also achieved great results. Although the above model has a high recognition rate, but there still remains a lot of work to do.

On the basis of summarizing research that already exists, this paper puts forward a new fishing urls identification model based on the depth learning and the characteristics of the phishing URL, also constructs an automatic identification platform. In this paper, no matter the depth of learning based on the fishing site identification model or the feature extraction, we made a corresponding improvement compare to the existing research. In the feature selection, this paper extracts feature extraction from four aspects: URL similarity, URL feature, web page text and message text. a multi-layer feedforward neural networks is used to construct the phishing URL recognition model based on URL feature. The
LSTM neural networks (Long Short-Term Memory)[8] is used to construct the phishing URL recognition model based on message text. The network carries on the integrated training to each sub-model, has achieved the good recognition effect.

2. Automatic identification platform for phishing URL in message

2.1. Platform architecture

The automatic identification system for phishing sites in the large scale mobile message consists of four modules, namely, message preprocessing, black-and-white list, feature extraction and phishing site recognition model based on deep learning. The system architecture figure as follows:

The preprocessing module uses regular expressions to separate URL and text from message and remove the irregular characters in the message. Also, by leveraging open source tool OpenCC[9] to do character conversion for the message to get the normalized URL and text.

Black-and-white list module, it uses network crawler to crawl and collect well-known urls in various industries from the public to create white list. While the black list, which is automatically added by platform identification. By leveraging open source tool NLPIR[10] and TF-IDF algorithm[11] to the text word segmentation and establish a keywords-based database.

The feature extraction module, using the network crawler and WHOIS[12] to get different characteristics of the URL, using the NLPIR to segmenting web text and message into word, using TF-IDF algorithm to extract text keywords from web pages, using Word2vec to translate message words into 400 dimensional word vectors.

Phishing site recognition model based on deep learning, which sets up the neural networks recognition model based on four aspects, URL similarity between measured URL and URL which is in the white list, URL features, the similarity between measured URL web text and whitelist keywords-based database and message.

On need basis, the platform uses the non-relational database MongoDB to store the data in a distributed way. MongoDB stores data in a document oriented manner, which has a big difference with the traditional relational database, the data structure for data storage is relatively loose. Also, it is unnecessary to know the structure definition in advance and store in same database with different structures.

2.2 Platform capability

The specific function of four modules of the platform mentioned above are shown in Table 1:
### Table 1 Comparison Between Numerical and Experimental Results

| model                        | function                                                                 |
|------------------------------|---------------------------------------------------------------------------|
| black and white list         | 1. Support regular update to crawl well-known urls including finance, e-commerce, social tools, operators, mobile phone manufacturers, banks, games, video and other areas; 2. Regularly update the black and white list |
| The preprocessing            | 1. Separate URL and text from message 2. Standardize URL and text in the message |
| The feature extraction       | 1. Extract URL web features: the number of external links, empty links, web pages, text and other information; Extract URL registration information: registrant, registration time, filing information; Extract the URL string information: URL length, the number parts of domain name, the number of abnormal characters, IP addresses which is instead of domain names and other information; 2. Extract web text keywords 3. Convert text messages into multidimensional word vectors |
| Phishing site recognition    | 1. Analyze input features and export the probability of phishing sites |
| model based on deep learning |                                                                           |

### 3. Recognition Model of Fishing Web Site Based on Deep Learning

The model of phishing URL recognition based on deep learning is constructed by four sub-models, That is:

1. Model of URL similarity calculate based on edit distance.
2. Model of phishing URL recognition based on neural networks and characters of that URL.
3. Model of phishing URL recognition based on text similarity of web page.
4. Model of phishing URL recognition based on LSTM neural networks.

The output of model is the probability of that URL is a phishing URL.

#### 3.1 model construction

The construction of model that recognize phishing URL based on deep learning have two parts like construction of sub-model and integration of sub-model.

#### 3.1.1 construction of sub-model

1. Separate domain information from the URL named A, calculate the edit distance \( L \) of domain A and domain B in white list. Edit distance is the least steps that A convert to B operations need like delete, insert and replace. Calculate of Similarity \( S \) like:

\[
S = \frac{1}{L+1}
\]  
\[(1)\]
2). Construction a Multilayer feedforward neural networks. Use features extracted from URL as input of the network. take gradient descent strategy update weight of network. calculate mean square deviation of network. set a neural networks recognition model based on URL feature.

3). Use the open source tool NLPIR segment web page text. Use TF-IDF algorithm(term frequency-inverse document frequency) extract keyword setA{A1,A2,A3,A4,A5}, calculate jaccard similarity of set A and keyword set B{B1,B2,B3,B4,B5} in white list:

\[
J = \frac{|A \cap B|}{|A \cup B|}
\]  

(2)

4). Construct a multilayer LSTM neural networks to extract deep expression of text. add a layer of logistic to classify. Use word vector as input of LSTM, and train it. build a phishing URL recognition in MSG based on LSTM neural networks.

Some expressions of LSTM like:

\[
i_t = \sigma(w_i x_t + w_h h_{t-1} + b_i)
\]  

(3)

\[
f_t = \sigma(w_f x_t + w_h h_{t-1} + b_f)
\]  

(4)

\[
g_t = \tanh(w_g x_t + w_h h_{t-1} + b_g)
\]  

(5)

\[
c_t = f_t \odot c_{t-1} + i_t \odot g_t
\]  

(6)

\[
o_t = \sigma(w_o x_t + w_h h_{t-1} + b_o)
\]  

(7)

\[
h_t = o_t \odot \tanh(c_t)
\]  

(8)

Where \(i\) is Input Gate. \(f\) is Forget Gate, \(o\) is Output Gate, \(\sigma\) is sigmoid function. \(w\) is weight matrix, \(b\) is bias of network, \(c\) is activation vector of Memory cell, \(h\) is hidden neuron. For a text sequence \(X = (x_1, x_2, \ldots, x_T)\) with length of \(T\). The \(t\)th word input of LSTM is \(x_t\), \(c_{t-1}, h_{t-1}\), and corresponding output is \(o_t, h_t\).

3.1.2. integrated of sub-model

1. Construct a multi layer BP neural networks, add logistic layer in the end to output classification probability. Use result of step 1)2)3)4) as input of network, activation function is sigmoid function. Train the network with gradient decent algorithm.

3.2 Evaluation index

Evaluation of model preference is as import as model’s algorithm. Commonly used precision (Precision), recall rate (Recall) and F value (F-measure) as a measure. Accuracy is an indicator of the accuracy of the identification of phishing sites, defined as follows:

\[
P = \frac{TP}{TP + FP}
\]  

(9)

where FP identifies the number of errors for the phishing URL, and TP identifies the correct number for the phishing URL.

Recall rate is used to measure the accuracy of the identification of normal web site indicators, defined as follows:

\[
R = \frac{TN}{FN + TN}
\]  

(10)
FN is the number of normal URL recognized error, TN is the correct number of normal URL recognition.

F value for the correct rate P and recall rate R harmonic mean, more commonly used is the F1 value, as follows:

\[ F1 = 2 \frac{P \times R}{P + R} \]  

(11)

In this paper, the accuracy of the model, the recall rate and F1 value to assess the pros and cons of model perform

4. Result

4.1 white&black list, preprocess and feature extraction

(1) white & black list

White list includes a total of 28029 mainstream urls, including financial, electricity, social tools, operators, mobile phone manufacturers, banks, games, video, news and other fields, blacklist collection of 3058 fishing sites.

(2) preprocess

For the preprocessing of message, this paper uses the regular expression to separate all the text and URL of the text message and remove the alien characters in the text and URL, and use OpenCC to convert the traditional characters into simplified characters to normalize the majority of the text.

(3) Feature extraction

Feature extraction mainly uses reptiles and WHOIS tools to collect and URL-related features, using open source tools NLPIR and Word2Vec to collect textual features. In the feature extraction, the URL that returns the error code 404 and can not be pinged multiple times is considered invalid. Training part of the collection characteristics of the situation shown in Table 2, of which 2558 samples of phishing sites, the normal number of 2,500 samples.

| Sample Type | registration information | URL string information | Web features | Text information |
|-------------|--------------------------|------------------------|--------------|-----------------|
| Phishing URL | 1223                     | 2558                   | 1791         | 2558            |
| Normal URL  | 1974                     | 2500                   | 2160         | 2500            |
| Total       | 3197                     | 5058                   | 3951         | 5058            |

4.2 Phishing URLs recognition model based on deep learning

In this paper, a phishing URL recognition model based on deep learning is proposed. In the experiment of evaluate model performance, the message sample containing the phishing URL is defined as the positive sample and the negative sample containing the normal urls is negative. In order to evaluate the performance of the model comprehensively, the positive and negative samples of the model training were randomly disrupted and divided into five different samples. The subset of the samples was divided as shown in Table 3.
Table 3 Sample subsets table

| Set Name | Train Set | Test Set |
|----------|-----------|----------|
| S1       | 1000      | 450      |
| S2       | 2000      | 450      |
| S5       | 3000      | 450      |
| S4       | 4000      | 450      |
| S5       | 5000      | 450      |

Experiments were performed on each subset of the samples in Table 3, and the result of different sample subsets on the accuracy and recall rate of the model is shown in figure 2:

![Figure 2: Accuracy and recall rate for different sample subsets](image)

It can be seen from Figure 3 that the accuracy of the model, the recall rate and the F1 value increase with the number of training set samples, but when the sample reaches 4000, the growth rate slows down. In order to validate the accuracy and effectiveness of the proposed algorithm, we choose a model based on NB and SVM proposed by GU X Q[13], which we can call it NB_SVM and SONG MQ[14] proposed PhishDetector, a kind of phishing URLs recognition model, remember as PharyDetector, and ZHANG WF[15], proposed another one, which is based on the Hungarian Matching Algorithm model, we can remember as Hungarian Matching. That can be regard as comparative verification model. The accuracy and recall rates of each model are shown in Table 4.

Table 4 Each model performance

| Model        | Precision | Recall | F1   |
|--------------|-----------|--------|------|
| PhishDetector| 0.980     | 0.900  | 0.938|
| This parper  | 0.982     | 0.969  | 0.975|
| NB_SVM       | 0.942     | 0.937  | 0.939|
| Hungarian    | 0.978     | 0.890  | 0.931|

5. Conclusion

Based on the analysis of a large number of phishing sites, this paper constructs an automated identification platform for phishing urls, which can automate the preprocessing of incoming messages, collate to black and white list, extract feature and phishing web site analysis. On the basis of summarizing the existing machine learning methods, this paper proposes a phishing urls recognition model based on deep learning. The experiment results show that the model has good accuracy and recall rate for the identification of phishing sites. The next step is to improve the extraction speed and effect of the feature extraction model.

Acknowledgements

This work was supported in part by the Ministry of Education-China Mobile Research Foundation(NO. MCM20160307)
References

[1] China Internet Network Information Center. The 40th China Statistical Report on Internet Development [R]. 2017.7.

[2] China Internet Network Information Center. Global Chinese Phishing Sites Report (2016) [R]. 2017.6.

[3] Steve Sheng, Brad Wardman, Gary Warner, et al. An Empirical Analysis of Phishing Blacklists [C]. In Proc. of the sixth Conference on Email and Anti-Spam, California USA, July 16-17 2009.

[4] Lin Yadong. Detection of Phishing URL Based on Abnormal Feature [J]. Electronic Test, 2014, (3): 70-72.

[5] Christian Ludl, Sean McAllister, Engin Kirda, et al. On the Effectiveness of Techniques to Detect Phishing Sites [C]. In Proc. of the 4th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, Lucerne Switzerland, July 12-13 2007: 20-39.

[6] Pan Y, Ding X. Anomaly based web phishing page detection [C]. 22nd Annual Computer Security Applications Conference (ACSAC 2006), Sep. 2006.

[7] Xiang G, Hong J, Rose CP, Lorrie C. CANTINA+: A feature-rich machine learning framework for detecting phishing web sites [J]. ACM Trans. on Information & System Security, 2011, 14(2): 613–613.

[8] Mike Schuster, Kuldip K Paliwal. Bidirectional recurrent neural networks [J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681.

[9] OpenCC: https://github.com/argolab/OpenCC/.

[10] NLPIR: http://ictclas.nlpir.org/.

[11] Huang D, Shanshan W, Ren Fu-ji. Creating chinese-english comparable corpora [J]. IEEE Trans. on Information and Systems, 2013, 96(8): 1853-1861.

[12] WHOIS: https://www.whois.com/.

[13] Phishing detection approach based on Naïve Bayes and support vector machine [J]. Computer Engineering and Applications, 2015, 51(4): 87-90

[14] SONG M Q, CAO X Y. Phishing detection method based on sensitive characteristics of phishing webpage [J]. Journal of Dalian University of Technology, 2013, 53(6): 903-907.

[15] ZHANG W F, ZHOU Y M, XU L, XU B W. A Method of Detecting Phishing Web Pages Based on Hungarian Matching Algorithm [J]. CHINESE JOURNAL OF COMPUTERS, 2010, 33(10).