Application of machine learning to identify Counterfeit Website

KuanTing, Wu	ShingHua, Chou	ShyhWei, Chen	ChingTsorng Tsai	ShyanMing, Yuan
Institute of Network	Department of Computer	Graduate Institute of	Department of Computer	Department of Computer
Engineering	Science, National Chiao	Biomedical Electronics	Science, Tunghai	Science, National Chiao
National Chiao Tung	Tung University,	and Bioinformatics,	University	Tung University,
University	Hsinchu, Taiwan	National Taiwan	Taichung, Taiwan	Hsinchu, Taiwan
Hsinchu, Taiwan	choush1119@gmail.com	University, Taipei,	ctsai@thu.edu.tw	smyuan@gmail.com
livebetterlife0820@gmail		Taiwan		
.com		d00945017@ntu.edu.tw		

ABSTRACT

Recent years the prevalence of fraudulent websites has become more severe than before. Fraudulent ecommerce websites that sell counterfeit goods not only cost financial damage to consumers but also have a great impact on Internet industry. Nowadays, there is not an effective way to confront these websites. In this paper, we look forward to achieving three goals: find the characteristics of counterfeit websites, train models for classifying ecommerce websites and provide a service to help consumers distinguish counterfeit websites from legitimate ones.

KEYWORDS

fraudulent website, counterfeit website, logistic regression, decision tree, support vector machine

1 Introduction

Fraudulent ecommerce websites selling counterfeit goods, not only cause loss to consumers but also affect digital advertising ecosystem. Usually, fraud ring rips consumers off by the means of advertisements or E-mails. In Taiwan, part of victims are drawn by the advertisements posted in social media such as Facebook and Line. In fact, these advertisements are posted by advertising agencies to help online shops gain exposure. However, they are abused by fraud ring now, which results in agencies to be reported. A digital advertising ecosystem, we can see that once fraud ring damage the reputation of agencies, the service provided by them is

affected as well. Thus, causing butterfly effect to whole digital advertising industry.

<https://doi.org/10.1145/3282373.3282407>The reason to such phenomenon is due to the difficulty of prevention in fraudulent websites. Traditionally, security companies use blacklist to block these fraudulent websites. However, fraudulent websites use URL shortening service to change their URLs and thus evade from blacklist. This severe abuse even made Google announce to close their URL shortening service on 2018/04/13, which implies that there is not an effective countermeasure to fight against fraudulent websites.

2 Backgrounds

2.1 Related Works

There have been several works on identifying malicious websites other than ecommerce websites [5-7]. The interest in identifying fraudulent ecommerce websites that sells counterfeit goods emerged until recent years. Wadleigh .et.al. [1] used three levels of features to train their models and make a discussion about prevalence of counterfeit websites in search engine. Carpineto and Romano [2] built two types of classifiers, one is to identify ecommerce websites and the other is to identify trustworthiness of these ecommerce websites subsequently. Oghaz .et.al. [3] utilized power of natural language processing (NLP) to create feature vectors rather than directly extract features from websites. Table 1 shows research mentioned above.

We not only use the text in a web page but also consider features from other aspects. Furthermore, we conduct several experiments and discuss how these features affect the performance of our models.

2.2 Classification

Classification belongs to a branch in machine learning called supervised learning. In a classification problem, the data used for training models are labeled. That is to say, we assign each data record (a website), with a given class. The main problem we are trying to solve is to decide whether a website sells counterfeit goods or not, which is a binary classification problem. Therefore, we

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
 iiWAS '18, November 19–21, 2018, Yogyakarta, Indonesia
 © 2018 Association for Computing Machinery.
 ACM ISBN 978-1-4503-6479-9/18/11 \$15.00
 DOI: <https://doi.org/10.1145/3282373.3282407>

Table 1 Summary of Previous Works

Paper	Model	Data Source (Amount)	Features	Accuracy	Precision	Recall
Wadleigh .et.al. [1]	GLM	Search Engine (602)	URL-level Page-level Website-level	85.0%	85.0%	74.8%
	SVM			86.4%	86.5%	76.9%
	AdaBoost			73.6%	71.4%	54.4%
Claudio .et.al. [2]	SVM	Search Engine (2000)	Ecommerce-or-not Legitimate-or-not	91.0%	-----	
Oghaz .et.al. [3]	LR	----- (430)	Bag-of-words	98.83%	-----	
	SGD			96.51%		
	SVM			88.37%		

define counterfeit websites are given with label 1 and legitimate websites are given with label 0 in our dataset.

Introduce performance metrics remember to say recall and fl score are important. Table2 shows the four situations (TP, FP, TN, FN) that predicted labels made by a classifier maps to their actual labels. We use these situations to calculate three performance metrics to evaluate a classifier, which are accuracy, precision and recall. In addition, another performance metric F1 score, which is the harmonic average of precision and recall, is also considered. The formulas of these performance metrics are listed in Table3. In this thesis, we take these four commonly used performance metrics into consideration to choose the models used in our system.

Table 2 TP, FP, TN and FN

	Actual 1	Actual 0
Predicted 1	True Positive(TP)	False Positive(FP)
Predicted 0	False Negative(FN)	True Negative(TN)

Table 3 Performance Metrics for Evaluating a Classifier

Metric
$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$
$Precision = \frac{TP}{TP + FP}$
$Recall = \frac{TP}{TP + FN}$
$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$

2.3 Clustering

Clustering is another branch in machine learning called unsupervised learning. Its purpose is to group a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. In this thesis, we use a centroid-based clustering algorithm, K-means, to help us fill the missing values in one of our experiments.

One critical problem of K-means is how to decide the optimal number of clusters (K). A solution to this problem is Average

Silhouette Method. Briefly speaking, this method will produce a silhouette score within range from -1 to 1 for each K. The higher this score is, the better clustering performance it achieves. Thus, Average Silhouette Method let us to evaluate the performance of clustering under multiple choices of K and choose the best among them.

3 Data Collection

3.1 Data Sources

The most important thing prior to training process is the preparation of the dataset. Thereby, our first job is to collect uniform resource locators (URL) of both counterfeit and legitimate websites. We collect counterfeit websites from three sources: Google Search Engine, Fraudulent Fan Page Ranking [8] and Whoscall [9]. As for legitimate websites, we collect them from Google Search Engine only.

The first data source for collecting counterfeit websites is Google Search Engine, which is also the method in previous research [1-2]. As for legitimate websites, we scrape brands from two famous online shopping platforms in Taiwan, Yahoo and Momo [10-11]. After that, we use Google Search Engine to search each of these brands' online shops.

3.2 Feature Extraction

Collecting these 800 URLs, we then extract features from these websites. In this thesis, we use 17 features in total [1] [2] and self-defined features. We divide these features into four groups by their properties. Table4 shows features group by group.

Table 4 Feature Groups

Group	Feature
URL	length_of_fqdn
	replica_in_fqdn
Content Element	num_of_currencies_seen
	num_of_duplicate_prices_seen
	percent_savings
	contain_emails
	large_iframes
	contain_phone_numbers
	has_mobile_app
	has_social_media
	has_payment_option
Content Structure	node_counts
	dom_height
	text_length
WHOIS	in_top_one_million
	china_registered
	under_a_year

Features in Group URL are transformed from URL of the websites. Features in Group Content Element are generated by checking the existence of or calculating values of certain elements. Features in Group Content Structure are our self-defined features. The idea comes from the observation that counterfeit websites have simpler structure than legitimate ones. Therefore, we quantify the concept of simplicity to produce these features by document object model(DOM), which is a tree structure used for representing web pages. Features in Group WHOIS are information about website owners when their websites were registered.

4 System Architecture

4.1 Overview

Figure 1 shows the architecture of our web service, it consists of four parts: Web Application Framework, , Message Broker and Process Pools. The general procedure is as follows and the detailed specification is in Workflows section.

Figure 1 Overview of Counterfeit Website Classification Service

- Step 1. Users provide parameters to web application framework.
Step 2~3. Web application framework saves the parameters into database and then publishes a message to the message broker.
Step 4~5. Process pools subscribes the message from the message broker, gets required information from database, process it and produce result.
Step 6~7. Process pools saves result back to database and then publishes another message to notify web application framework.
Step 8~10. Web application framework receives the message and gets the result. At the end, the result is sent back to users.

Web application framework has two components, which are frontend and backend. Frontend receives user parameters and sends these parameters to backend. Backend is responsible for interacting with the components in process pools through database and message broker.

Process pools contains four components: URL Crawler, Feature Crawler, Model Caller and Model Updater. Each component is a process pool that can designate multiple processes to execute its specific job. URL crawler is responsible for crawling new URLs of both counterfeit and legitimate websites periodically. Feature crawler crawls features from a given website and fill the missing values if there is any. Model caller is in charge of making prediction of a website while model updater's duty is to do feature selection and train new models afterward.

4.2 Predict Workflow

In this paper, user needs to provide the URL of a website and the system will return the prediction result. The prediction result consists of three things: predicted label of the website, possibility of that website belongs to counterfeit websites and the reason why system classify it as a counterfeit or legitimate website. We define two situations in this workflow. The first situation is URL_NOT_EXIST, which means that a website is not in the database since the system has not predicted this website before. Therefore, backend will publish a message to feature crawler to and follows the procedure shown in the Figure3 As for the other situation, URL_EXIST, means that a website has been predicted. Hence, backend will access the prediction result directly from the database.

5. Experiments

5.1 Data Preprocessing

Data preprocessing is to process the raw data before training the models. The importance of data preprocessing is that it affects the quality of data and thus determines the limit of the model performance. In this paper, we adopt two steps in preprocessing, one is filling missing values and the other is data normalization, before experiments in the next section.

The purpose of filling missing values is to keep the dataset as complete as possible. To deal with missing values, we can either discard them or fill them. The former is suitable only when the dataset is large enough, but we only got 800 websites in our case. As a result, we choose to the latter one and fill the missing values of each feature with 0, no matter its type is numerical or Boolean.

The range of features may be various from one to one, which results in the prediction made by models dominated by some large-range features. Hence, data normalization is to scale the features into the same range such that the effect of each feature to be the same. In this paper, we adopt Min-Max Normalization for this purpose and normalize features to the range from 0 to 1. Equation 1 is the formula of Min-Max Normalization.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

5.2 Result

We collect 400 Chinese counterfeit websites and 400 Chinese legitimate websites as our training set for feature extraction .We run three classification algorithms with 5-fold cross-validation; the 5-fold mean result is in Table5. We use three model: Logistic Regression (LR), Decision tree (DT) and Support vector machine (SVM). The rest of the rows represents that we eliminate one feature to A/B test whether the feature influence models significantly. It shows that all four performances metrics increase to near and even exceed 90% for three algorithms. Additionally, the feature has social_media seems to be the main reason for such enhancement since four performance metrics drop the most after we eliminate this feature.

Table5 A/B Testing Result

	test_accuracy			test_precision			test_recall			test_f1		
	LR	DT	SVM	LR	DT	SVM	LR	DT	SVM	LR	DT	SVM
all	90.88	88.00	90.38	89.49	88.60	89.97	92.75	87.25	91.00	91.05	87.90	90.43
w/o contain_phone_numbers	90.88	88.50	90.50	89.49	89.09	89.81	92.75	87.75	91.50	91.05	88.40	90.60
w/o has_mobile_app	91.00	89.00	90.50	89.65	89.84	89.81	92.75	88.00	91.50	91.15	88.89	90.60
w/o has_payment_option	91.00	87.75	90.25	89.71	88.17	89.74	92.75	87.25	91.00	91.16	87.70	90.32
w/o has_social_media	84.63	83.50	83.88	85.74	85.59	84.40	83.25	81.00	83.25	84.38	83.01	83.76

Because has_social_media is the main cause, we calculate the proportion of social media links in Table 6 and find two facts. One fact is that up to 93% of counterfeit websites do not have links to social media while up to 80% legitimate ones do; and the other is that among three types of social media, i.e, Facebook, Instagram and Line, there is up to 74% of legitimate websites have links to fan page of Facebook.

Table 6 Proportion of Social Media Links

Type of Website Proportion of Links(%)	Counterfeit	Legitimate
fb	2.18	74.42
ig	0.73	26.04
line	6.30	40.55
at_least_one_social_media	7.75	80.18

6 Conclusion

This thesis finds out that counterfeit websites have following characteristics from coefficients of Logistic Regression:(1)URL: shorter URL length; (2)Content Element: higher percent savings, lower number of duplicate prices, no links to social media and mobile applications;(3)Content Structure: shorter and smaller DOM;(4)WHOIS: usually registered in a year.

There is abundant space for further progress in counterfeit website classification. We can discuss about this thesis's future works from three aspects: dataset, model and system.

So far, the dataset we used for training models consists of only Chinese websites. However, the features extracted from these websites can also apply to websites in other language families. Hence, we can verify if the characteristics of counterfeit websites in other countries are the same as ones in Taiwan.

The models used in the current system is for classifying counterfeit websites specifically. Nevertheless, there are other kinds of fraudulent websites on the Internet. We can collect more features by referencing other researches and attempt to train new models for classifying more general fraudulent websites.

REFERENCES

- [1] John Wadleigh, Jake Drew and Tyler Moore, "The E-Commerce Market for "Lemons": Identification and Analysis of Websites Selling Counterfeit Goods", Proceedings of the 24th International Conference on World Wide Web (WWW15), pp. 1188-1197, 2015.
- [2] Claudio Carpineto and Giovanni Romano, "Learning to detect and measure fake ecommerce websites in search-engine results", Proceedings of the International Conference on Web Intelligence (WI17), pp. 403-410, 2017.
- [3] Mahdi Maktabdar Oghaz .et.al, "Content-based Fraudulent Website Detection Using Supervised Machine Learning Techniques", HIS 2017: Hybrid Intelligent Systems, pp. 294-304, 2017
- [4] Digital Advertising Ecosystem. [Online]. Available: <https://hsienblog.com/2017/06/25/%E6%95%B8%E4%BD%8D%E5%BB%A3%E5%91%8A%E6%9E%B6%E6%A7%8B-rtb-dsp-ssp-dmp-ad-exchange-ad-network/>
- [5] Birhanu Eshete, Adolfo Villa!orita, and Komminist Weldemariam, "Judging a site by its content: learning the textual, structural, and visual features of malicious web pages", Proceedings of SecureComm 2012: Security and Privacy in Communication Networks, pp. 149-166, 2012.
- [6] S. N. Bannur, L. K. Saul, and S. Savage, "BINSPECT: Holistic Analysis and Detection of Malicious Web Pages", Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, pp. 1-10, 2011.
- [7] D. Canali, M. Cova, G. Vigna, and C. Kruegel, "Prophiler: a fast filter for the large-scale detection of malicious web pages", Proceedings of the 20th International Conference on World Wide Web (WWW '11), pp. 197-206, 2011.
- [8] Fraudulent Fan Page Ranking. [Online]. Available: <http://page.board.tw/rank.php?tagid=74>
- [9] Whoscall. [Online]. Available: <http://cybercrime.whoscall.com/>
- [10] Yahoo Shopping Platform. [Online]. Available: https://tw.mall.yahoo.com/merchant_homepage?searchby=sname
- [11] Momo Shopping Platform. [Online]. Available: https://www.momoshop.com.tw/brand/Main.jsp?FTOOTH=21&Area=tooth&cid=T21&oid=L21999&mdiv=1000100000-bt_0_157_01-&ctype=B
- [12] Nhamo Mtetwa, Mehdi Yousefi, Viseshini Reddy, "Feature selection for an SVM based webpage classifier", 2017 IEEE 4th International Conference on Soft Computing & Machine Intelligence (ISCMCI), pp. 85-88, 2017.
- [13] Gayathri. S, "Phishing Websites Classifier using Polynomial Neural Networks in Genetic Algorithm", 2017 4 th International Conference on Signal Processing, Communications and Networking (ICSCN -2017), pp. 1-4, 2017.
- [14] Prasad Shende, Shivani Mehendarge, Sachin Chougule, Pallavi Kulkarni, Umesh Hatwar, "Innovative ideas to improve shopping mall experience over E-commerce websites using Beacon technology and Data mining algorithms", 2017 International Conference on circuits Power and Computing Technologies [ICCPCT], pp.1-5, 2017.
- [15] Ebubekir Buber, Önder Demir, Ozgur Koray Sahingoz, " Feature selections for the machine learning based detection of phishing websites", 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), pp.1-5, 2017.