# Specialist English: Assignment 4

Rebecca J. Stones
rebecca.stones82@nbjl.nankai.edu.cn

Date due: 5 November 2018

This fourth assignment (worth 5% of the final mark) looks at improving the writing in an English-language paper.

I'll scale the marks on this assignment according to $m \mapsto \min(m, 10)$ for Master's students and $m \mapsto \lceil m/1.3 \rceil$ for Ph.D. students.

My marking will be affected by (a) your English writing, (b) your LaTeX typesetting, (c) your mathematical presentation, and (d) your understanding of the underlying computer science. Basically, I will "peer review" your assignments.

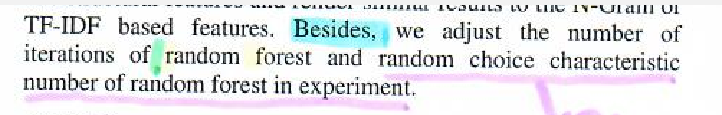
We'll look at the following paper in detail:

> C. Liu, L. Wang, B. Lang, Y. Zhou, *Finding effective classifier for malicious URL detection*, Proc. ICMSS, 2018.
>
> https://dl.acm.org/citation.cfm?id=3181352 (direct link)
>
> https://doi.org/10.1145/3180374.3181352 (DOI)

This paper uses machine learning to detect malicious URLs. I attach my annotations below, highlighting the parts I consider problematic.

**Problem 1** Improve the title by (a) fixing the grammar and (b) deleting an unnecessary word. [1 mark]

**Problem 2** I recommend not to start a sentence with "Besides, ...", which may be misinterpreted as meaning "Anyway, ...". What is a suitable alternative to "Besides, ..." in this case? [1 mark]



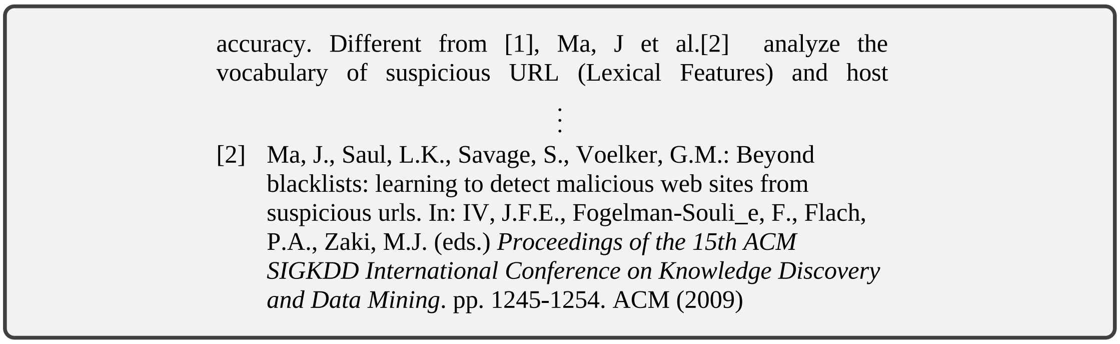
features and render similar results to the N-Gram or TF-IDF based features. Besides, we adjust the number of iterations of random forest and random choice characteristic number of random forest in experiment.

**Problem 3** What is a less childish way of saying "a lot of" in "scholars have conducted a lot of research"? [1 mark]

**Problem 4** Regarding the following snippet:

1. Writing "Different from [1], ..." looks wrong to me[1], although I don't have a good explanation as to why it looks wrong. What's an alternative to "Different from [1], ..."? [1 mark]
2. How should "Ma, J et al.[2]" be written? [1 mark]
3. The reference [2] is a mess; how should it be fixed? [1 mark]

[1] And others seem to have the same problem: https://english.stackexchange.com/questions/114734/is-a-sentence-beginning-with-different-from-not-so-good.



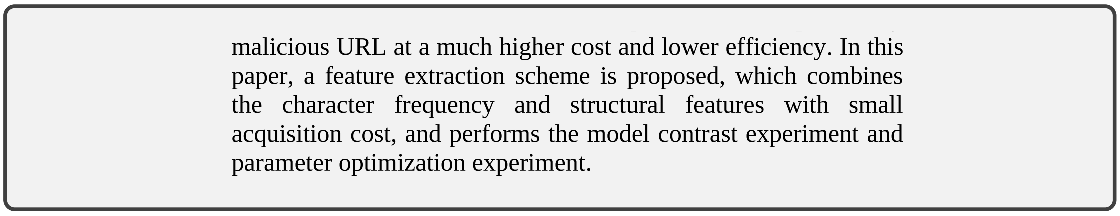
accuracy. Different from [1], Ma, J et al.[2] analyze the vocabulary of suspicious URL (Lexical Features) and host

⋮

[2] Ma, J., Saul, L.K., Savage, S., Voelker, G.M.: Beyond blacklists: learning to detect malicious web sites from suspicious urls. In: IV, J.F.E., Fogelman-Souli_e, F., Flach, P.A., Zaki, M.J. (eds.) *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 1245-1254. ACM (2009)

(I also recommend using simple past tense for past papers; in this case "analyzed" instead of "analyze". Also "suspicious URL" should be plural, i.e., "suspicious URLs".)

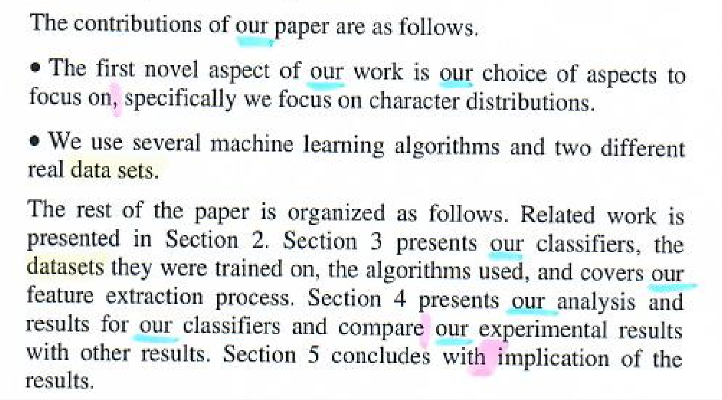
**Problem 5** What is ambiguous about "the character frequency and structural features with small acquisition cost"? [1 mark]



malicious URL at a much higher cost and lower efficiency. In this paper, a feature extraction scheme is proposed, which combines the character frequency and structural features with small acquisition cost, and performs the model contrast experiment and parameter optimization experiment.

**Problem 6** Regarding the following snippet:

1. What's a more suitable punctuation mark in "... choice of aspects to focus on, specifically ..." than a comma? (This is called a comma splice[2].) [1 mark]
2. I recommend against using "our" as it's exclusive and unfriendly (except when it's technically correct, e.g., "our data set" is suitable when the reader does not have access to the data set). From the snippet below, give two examples of how we can avoid using "our". [2 marks]



The contributions of our paper are as follows.

- The first novel aspect of our work is our choice of aspects to focus on, specifically we focus on character distributions.
- We use several machine learning algorithms and two different real data sets.

The rest of the paper is organized as follows. Related work is presented in Section 2. Section 3 presents our classifiers, the datasets they were trained on, the algorithms used, and covers our feature extraction process. Section 4 presents our analysis and results for our classifiers and compare our experimental results with other results. Section 5 concludes with implication of the results.

**Problem 7** From the highlighted text, identify any three distinct problems (other than the problems mentioned above). [3 marks]

[2] https://en.wikipedia.org/wiki/Comma_splice

# Finding effective classifier for malicious URL detection

Chunlin Liu
State Key Lab of Software Development Environment
School of Computer Science and Engineering,
Beihang University
Beijing, China
jackliu@buaa.edu.cn

Lidong Wang
National Computer Network Emergency Response
Technical Team/Coordination Center of China
Senior Engineer
Beijing, China
wld@cert.org.cn

Bo Lang
State Key Lab of Software Development Environment
School of Computer Science and Engineering,
Beihang University
Beijing, China
langbo@buaa.edu.cn

Yuan Zhou
National Computer Network Emergency Response
Technical Team/Coordination Center of China
Senior Engineer
Beijing, China
13910020110@163.com

## ABSTRACT

Malicious URL is an important security issue to the Internet, which has a significant economic impact. By now, it is still a challenging problem. In this paper, we propose that combining statistical analysis of website URLs with machine learning techniques will give a more accurate classification of malicious URLs. We focus on the Character features of malicious URLs by statistical methods to obtain char distribution features and structural features. Then, In order to find effective classifier for malicious URL detection, we use six different classifiers to perform cross training. The experimental results on our data set demonstrate that the combination of the URL features extracted in this paper and the Random Forest classification algorithm can achieve 99.7% precision with a false positive rate of less than 0.4%. We also show that these features render better performance than the previously used features which combine lexical features and structural features and render similar results to the N-Gram or TF-IDF based features. Besides, we adjust the number of iterations of random forest and random choice characteristic number of random forest in experiment.

(?)

## CCS Concepts

• **Security and privacy → Network security**

## Keywords

URL classification; char distributions; Random Forest

## 1. INTRODUCTION

In recent years, the rapid development of the Internet has created enormous convenience for people's daily life. At the same time, convenient network service also attracted the network attackers through phishing, spam and malicious software and other ways to profit illegally. Although the purposes and means of these illegal activities are different, but they require unsuspecting users to access malicious URLs provided by attackers. Even if the user's computer to install anti-virus software or firewall, which prevent most of the malicious attacks, various forms of malicious URLs, such as "viruses, spam, phishing sites, especially web hanging horse", the users inadvertently will be caught and cause a huge loss of information and property. Therefore, how to detect malicious URLs, to reduce the loss of ordinary users and maintain public Internet security, is of great significance.

虽然…
但是…
Chinglish

Beyond blacklisting heuristics, domestic and foreign scholars have conducted a lot of research work on malicious URL detection. There are two main research ideas in the literature. One is that attempt to find the best classification features which can be extracted only from the URL, and the other is that visit the potentially malicious site and extract features from the crawled content. For example, Choi et al. [1] uses a lot of lexical features, host-based features, link breadth features (the number of websites linked to this site) and web content features to achieve a 98% accuracy. Different from [1], Ma, J et al.[2] analyze the vocabulary of suspicious URL (Lexical Features) and host attributes (Host - Based Features), and use the bag of words model (Bag-of-Words) to represent features, and obtain thousands of features. In lexical features, on the one hand consider the host name length, URL length, URL point number; on the other hand, for each word symbol host and path in URL, using the bag of words model to build a two valued feature. In the host features, the IP address attributes, WHOIS attributes, domain name attributes and geographical location attributes are considered. Finally, the accuracy rate of 95-99% is obtained. Zhang Y et al. [3] uses a web content based feature to achieve a 94% accuracy. Fu et al. [4] based on Web image similarity detection, this method calculates the similarity between the images by EMD distance, but it can only be applied to the case where malicious URL is very similar to a benign URL and have certain limitations.

bag-of-words

Then, let's look at some other examples. Garera et al. [5] analyzed the URL structure of phishing sites, and introduced the selection process of feature set in detail. Logistic Regression Filter was used to classify URL. URL structure of phishing sites is analyzed, and 4 types of URL structure are obtained. Feature set consists of 18 features: page feature, domain name feature, type feature and word feature. The page features with Google search engine, select Page Rank of URL , Page Rank of Host, PageRank Present in Crawl Database, Page Present in Index ,the two page quality evaluation (Page Quality Score) a total of 6 characteristics; In the domain name feature selections ,the domain name in the white list

*ICMSS 2018*, January 13–15, 2018, Wuhan, China
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5431-8/18/01…$15.00
DOI: https://doi.org/10.1145/3180374.3181352

(?)

represents 1 feature ; Type feature select the type of I, II, and III ,3 features; Word features select secure, account, webscr, login, ebayiapi, signin, banking and confirm 8 words . Finally, the accuracy rate of 97.31% was obtained. Kan et al. [6] uses word level n-grams features to classify and obtain 76% accuracy when there is no combination of web page text features. When combined with web page text features, 81% accuracy is achieved. Aldwairi et al. [7] extracts n-grams, TF-IDF and content features from URL, and achieves an average accuracy of 87%.In 2010, Liu Gang et al. [8] took link relations, keywords sorting relations, text similarity relations, hierarchical similarity relations as statistical features, using unsupervised learning algorithm DBSCAN (Density-Based Spatial Clustering of Applications with Noise) for fishing URL attack target recognition. Experimental results show that the proposed method can identify 91.44% phishing targets, and the false positive rate is controlled at 3.4%.

Although the above studies made some progress in the detection of malicious URL, but there are still some shortcomings, such as no general features; link breadth, host-based features need to get through third party tools , so these features are expensive to acquire and are less efficient while comparing with lexical features, n-grams, TF-IDF, structure characteristics, character distribution. Web content features require access to potentially malicious URL at a much higher cost and lower efficiency. In this paper, a feature extraction scheme is proposed, which combines the character frequency and structural features with small acquisition cost, and performs the model contrast experiment and parameter optimization experiment.

passive

The contributions of our paper are as follows.

- The first novel aspect of our work is our choice of aspects to focus on, specifically we focus on character distributions.

- We use several machine learning algorithms and two different real data sets.

The rest of the paper is organized as follows. Related work is presented in Section 2. Section 3 presents our classifiers, the datasets they were trained on, the algorithms used, and covers our feature extraction process. Section 4 presents our analysis and results for our classifiers and compare our experimental results with other results. Section 5 concludes with implication of the results.

## 2. RELATED WORK

Many researchers have presented their work on malicious URLs detection, so there are some similarities between our work and theirs, but also some key differences. The smart reader will see from the presentation below, there are also several similarities among the features used by previous researchers, since the URL is basically a string of symbols.

We use some same features and classifiers as the below authors do. The authors of [9] do use some structural features such as IP address and lengths of domain and URL and a strict subset of the classifiers (SVM, logistic regression, and a naive Bayes classifier). Unlike their work we have not used a bag-of-words feature implementation because our goal is to demonstrate the effectiveness of a small feature vector and improve efficiency. The work by [10] has been influential in the detection of phishing websites. There is some overlap between their work and ours, specifically in their URL analysis section of their methods. We share a set of four features: whether or not the URL is an IP address (also used by Ma et al. and others), the number of dots in the domain name, the length of domain in the malicious URL, the depth of a URL path. However, bigger differences between our work and theirs are that they also use the PageRank and calculate domain similarity and test their approach on smaller datasets.

Some recent studies are as follows. Recently in [11] the authors are able to accurately classify 99.3% of their dataset, with a false positive rate of less than 0.4% by using a two-sample Kolmogorov-Smirnov test along with other features. Our method can get slightly better results by using different feature combination. In [12], the authors have created a system based only on URLs to detect phishing websites. The authors achieve good accuracy, but lower than ours, with a confidence weighted classifier, using a bag-of-words implementation that generates a huge set of 369,000 features. Additionally, the authors have some hand selected features that correspond to ours, such as IP detection, and Length analysis. The authors of [13] extract n-grams and tf-idf scores from malicious URLs. They train and test using Logistic Regression, Linear SVC, open-source XGBoost, and Multilayer Perceptron and their classifiers achieves good accuracy with low false true rate. We get similar results on their dataset with focusing on the character of malicious URLs.

## 3. CLASSIFIERS

### 3.1 Feature

This section discusses our feature extraction algorithms. Prior to feature extraction "http://" and "https://" were removed from the URL, so as not to the features such as number of punctuation, and to prevent URLs classified without such parts from being misclassified.

In malicious attacks, the evildoers always try to absorb victims into clicking a URL pointing to the malicious site. They usually obfuscate malicious URLs through various methods. Every method attaches some features to the malicious URLs and these features can differentiate it from a legal one. Therefore, the URL features are essential to detect the malicious activities. By analyzing the malicious URLs we collect, the prominent features of a malicious URL are listed as follows:

1) Mixing IP address in the malicious URL. According to 5000 malicious URLs and 5000 legitimate ones, we can find that legitimate URLs containing an IP address almost do not happen.

2) Obfuscating the domain with a mass of dots. Malicious URLs usually use lots of dots to confuse users, for example,http://https.secure.paymentaccount.customer.service.com .setup-onlinecustomerhelp.inc-faq.cgi-ticket-help.desk.updated-setup-connect.uroproct.com.br/myaccount/.This kind of URL rarely exists in a legitimate URL.

3) Confusing users with abnormal depth of a URL path. In other words, there are many "/" in malicious URLs.

4) Confusing users with some special characters, such as "@", "~", "-". These special characters are often found in malicious URLs.

5) Abnormal length of domain in the malicious URL. Under normal circumstances, the string appearing between the http:// and the first "/" is considered as domain and the length is relatively longer in parts of malicious URLs.

6) The character distribution of benign URL and malicious URL is quite different. We have statistics of each character in URL (A-Z (ignore case), 0-9) the number of occurrences, then the total length divided by, then calculate the entropy, to construct a