基于机器学习的钓鱼网站的检测

贾怀宇 2020090916007

论文原名：Phishing URL Detection:A Network-based Approach Robust to Evasion

摘要：

许多网络攻击是从传播网络钓鱼网址开始的。当点击这些钓鱼网址时，我们的私人信息会被泄露给攻击者。目前已经提出了几种机器学习方法来检测钓鱼网站。我们简要介绍了几种常见的机器学习方法，其中有一种能够有效检测有规避性的网站。原文提出了一种基于网络的推理方法，以准确地检测出伪装成合法模式的网络钓鱼网址，对逃避行为具有鲁棒性。除此之外，本报告还列举了其他的应用在钓鱼网站的的机器学习方法。我们希望通过对比这些检测手段，来获得更高效的钓鱼网站侦察方法。

论文原址：<https://arxiv.org/pdf/2209.01454.pdf> [1]

关键词：phising detection; classification; network-based inference

**1.引言**  
**1.1 背景**

网络攻击在信息时代已经层出不穷，许多攻击者会使用主动攻击，欺骗等手段，这对我们的社会造成了巨大的损害，许多网络攻击是从网络钓鱼开始的。网络钓鱼是为了欺骗人们将他们的敏感信息透露给攻击者。特别是，网络钓鱼的URL被伪装成人们看起来很熟悉的URL。粗心的人就会点击它们，导致他们的私人信息被泄露。比如说QQ邮件，欺骗性的购物网站，人们在不知不觉中就会泄露自己的信息。

因此，许多检测钓鱼url的方法已经被开发出来，作为回应，攻击者开始考虑用合法模式进行伪装的规避技术。因此，防止使用规避的网络钓鱼攻击是最重要的。

已有人提出机器学习方法来检测网络钓鱼。它们可以被分为两种类型：基于内容和基于URL字符串。基于内容的方法下载并分析网页内容，要收集这样的训练数据并不容易。基于内容的方法并不总是首选。基于字符串的方法主要依赖于URL字符串模式分析，因为众所周知，钓鱼网站的字符串模式非常容易区分。因此，人们提出了许多检测钓鱼网站的词汇特征（见第2节）。由于基于字符串的方法在计算上是轻量级的，并能提供较高的准确性，许多研究者倾向于使用这种方法。然而，众所周知，它的准确度很一般。

**1.2 方法**

几乎所有现有的基于字符串的方法都没有考虑规避问题[45]。逃避是指攻击者通过操纵钓鱼网站的模式来欺骗防御者的检测方法，从而创造出看似合法的钓鱼网站的技术。由此，本文设计了一个新颖的自然语言处理统一框架和基于网络的方法来检测钓鱼网站。本文将每个URL视为一个句子，并将其分割成子串(字），考虑到URLs的语法和标点符号--URLs在英语中具有明确的语法。之后，我们建立一个由异质实体组成的大网络，如URL、域名、IP地址、权威名称服务器和子串，并执行我们定制的信念传播来检测钓鱼网站。

**2.知识补充**

**2.1 url**

Internet上的每一个网页都具有一个唯一的名称标识，通常称之为URL（Uniform Resource Locator, 统一资源定位器）。[4]它是www的统一资源定位标志，简单地说URL就是web地址，俗称“网址”。 URL是对互联网上得到的资源的位置和访问方法的一种简洁表示，是互联网上标准资源的地址。URL它具有全球唯一性，正确的URL应该是可以通过浏览器打开此网页的。



图1：url: https://www.uestc.edu.cn/

URL由三部分组成：资源类型、存放资源的主机域名、资源文件名。

URL的一般语法格式为：

protocol :// hostname[:port] / path / [;parameters][?query]#fragment

protocol（协议）

指定使用的传输协议，最常用的HTTP协议，它是目前WWW中应用最广的协议。

常见的属性：

|  |  |  |
| --- | --- | --- |
| http | 通过 HTTP 访问该资源 | 格式 HTTP:// |
| https | 通过安全的 HTTPS 访问该资源 | 格式 HTTPS:// |
| ftp | 通过 FTP访问资源 | 格式 FTP:// |

表一：常见协议与属性

一般来说，https开头的URL要比http开头的更安全，因为这样的URL传输信息是采用了加密技术。

hostname（主机名）

是指存放资源的服务器的域名系统(DNS) 主机名或 IP 地址。有时，在主机名前也可以包含连接到服务器所需的用户名和密码（格式：username:password@hostname）。

port（端口号）

HTTP缺省工作在TCP协议80端口，用户访问网站http:// 打头的都是标准HTTP服务。HTTPS缺省工作在TCP协议443端口。

path（路径）

由零或多个“/”符号隔开的字符串，一般用来表示主机上的一个目录或文件地址。

parameters（参数）

这是用于指定特殊参数的可选项。

query(查询)

可选，用于给动态网页（如使用CGI、ISAPI、PHP/JSP/ASP/ASP。NET等技术制作的网页）传递参数，可有多个参数，用“&”符号隔开，每个参数的名和值用“=”符号隔开。

fragment（信息片断）

信息片断，字符串，用于指定网络资源中的片断。例如一个网页中有多个名词解释，可使用fragment直接定位到某一名词解释。

**2.2 网络钓鱼**

网络钓鱼不仅是一种网络攻击技术同时也是一项最常见的社会工程技术，网络犯罪分子或网络攻击者通过尝试伪装为可信任个人或公司组织来进行发送信息，来获取企业或私人的敏感信息（包括用户名称、密码、手机号码、银行卡账号密码、个人邮箱信息等等）。通过欺骗伪装形式来操纵收件人泄露敏感信息、下载恶意软件或资金或资产错误的转移到攻击者指定的账户。



图二：钓鱼网站

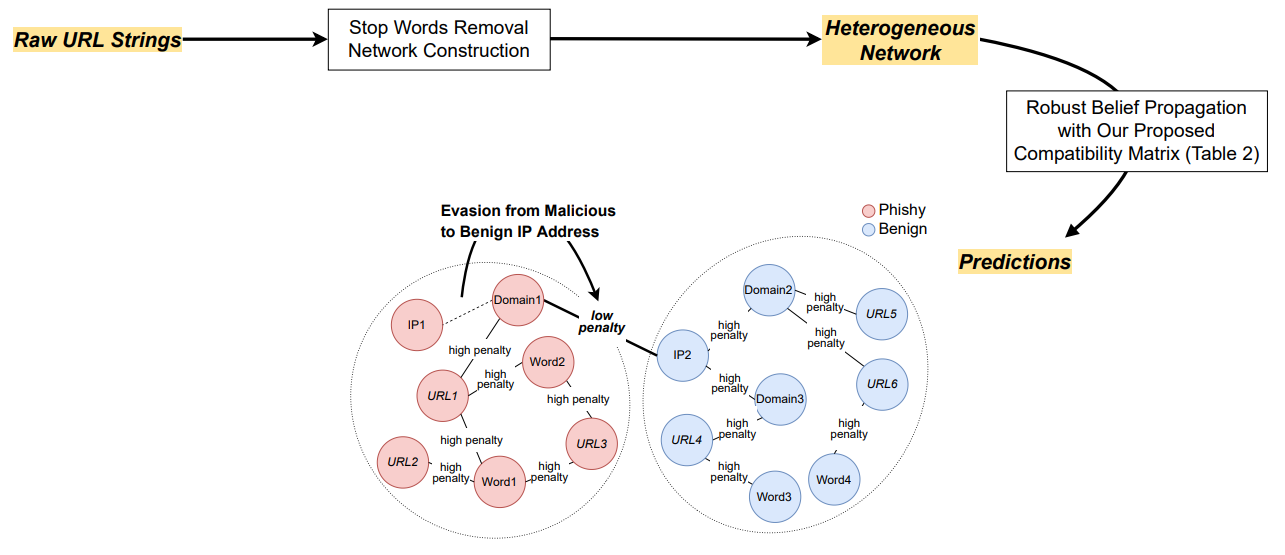
**3. 模型**

**3.1 总述**

(1) 我们从phishtank.com抓取许多URL，并下载其他作品的开放数据集。

(2) 如前所述，我们创建了一个由URLs、域名、IP地址、名称服务器和子串（词）组成的异质网络。我们使用标准的自然语言处理技术将URLs分割成子串（词），并在URL和子串之间画出边框。

(3) 我们运行我们定制的信念传播算法来推断未知URL的虚假/良性标签，这是我们的主要贡献。特别是，这种类型的推断被称为过渡性推断。在我们的案例中，训练样本和测试样本同时存在于一个网络中，测试样本的标签是根据网络结构从其他已知的训练样本的标签中推断出来的。



图三：模型架构

**3.2 网络构建**

钓鱼网站有许多共同的字符串模式，而且各种实体相互交叉关联，因此我们创建了一个网络来表示多个实体（顶点）之间的复杂关系，如URL、其域名、IP地址、权威名称服务器和子字符串：

1.一个URL和它的域之间画一条边。

2.一个域名和它所解决的IP地址之间画一条边。

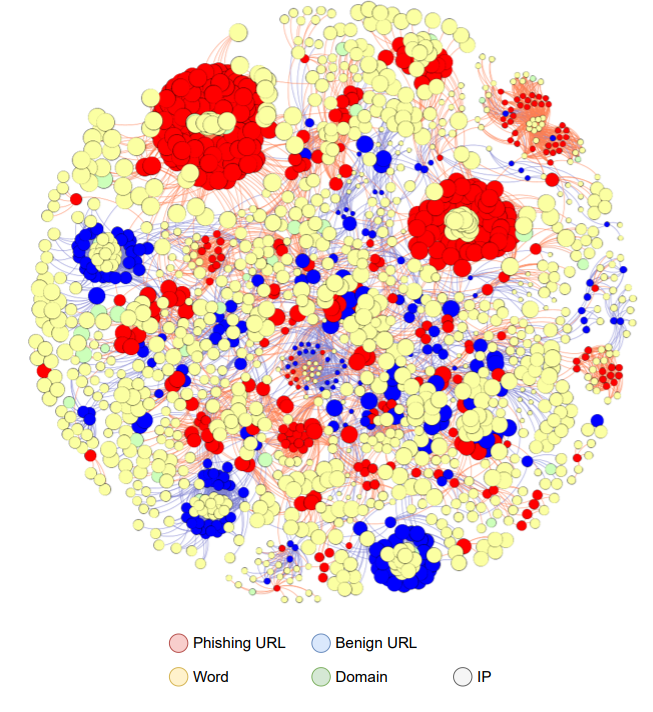
3.一个域和它的权威名称服务器之间画一条边。

**3.3 url划分**

如何将一个URL分割成单词。一个URL是用来定位互联网上的资源的。它由几个部分组成：方案、用户名、密码、主机、端口号、路径和查询字符串--其中一些可能丢失。我们在每个部分使用定制的单词分割策略，如下所示：

1. 方案指的是协议，例如，http和https。只可能有两个词。
2. 用户名和密码可以在主机之前指定。我们使用标点符号来分割它们，即'//'、':'和'@'。
3. 主机名可以简单地用'.'分割成单词。
4. 有时路径可以很长，用'/'隔开。我们使用所有可能的标点符号，如'/'、'.'、'！'、'&'、','、'#'、'$'、'%'和';'，将路径部分分割成单词。
5. 查询字符串能够包含多个由'&'分隔的查询，每个查询由一个查询名称和一个值组成。

下图是拟创建的网络，红色代表钓鱼网站，蓝色代表良性网站。其他颜色表示非URL实体--由于名称服务器的重要性低于其他实体，所以没有显示。请注意，存在着许多群集。顶点的大小代表一个顶点所属集群的强度。

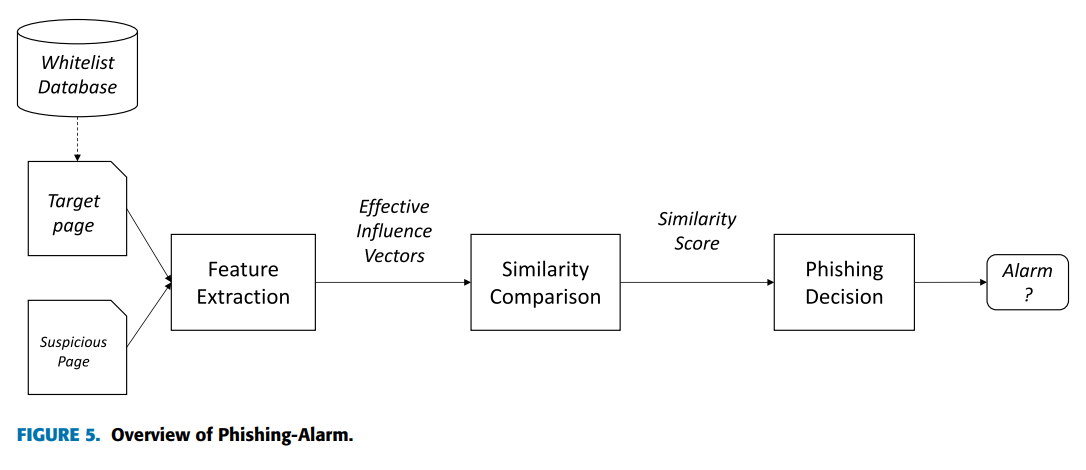


图三：由上述方法创建的网络

**4. 相关工作**

**4.1 基于网页内容特征的钓鱼网址检测方法 [3]**

在本篇论文中，我们提出了一种算法，根据网页之间视觉外观的相似性来量化网页的可疑性评级。由于层叠样式表（CSS）是跨浏览器实现的指定页面布局的技术，我们的方法使用CSS作为基础来精确量化每个页面元素的视觉相似性。



图四：基于CSS方法的检测

在第一阶段，即特征提取阶段，给定一个可疑页面Ps，我们提取其CSS结构CSS(Sus)，并将其转换为感染向量，以代表页面Ps的视觉布局的静态特征。我们还维护了一个白名单数据库，其中包含了被网络钓鱼攻击的热门网页。我们用同样的方法从数据库中提取页面特征。

在第二阶段，基于影响向量，我们对可疑网页和白名单数据库中的网页的相似性进行匹配。

最后，我们通过比较页面的相似度得分和预设的阈值来做出决定。如果相似度得分超过，并且存在其他线索表明两个测试页面是不同的，那么可疑页面将被视为钓鱼页面。

**4.2****使用递归神经网络对钓鱼网站进行分类 [2]**

本文将URL的词法和统计分析与随机森林分类器结合起来，根据URL的特征对钓鱼网站进行分类的方法。首先，用特征工程的方法提取一系列重要的变量。然后，使用分类算法来建立模型。

**4.2.1 特征提取：**

我们确定了一组14个特征，可用于创建URL的词汇和统计分析。

- 域名存在于Alexa排名中：如果该域名存在于前100万个Alexa域名中。Alexa排名是一个按互联网受欢迎程度排列的域名列表。大多数网络钓鱼网站都托管在被黑的合法网站或新域名中。如果钓鱼网站被托管在一个被劫持的网站中，那么该域名不太可能是顶级Alexa域名的一部分，因为排名靠前的域名往往有更好的安全措施。

- 子域长度。这需要子域的URL长度。钓鱼网站试图通过使用合法网站的域作为其子域来模仿合法网站的URL。真正的网站往往有一个短的子域。

- URL长度。这需要URL的长度。一个长的URL会增加混淆用户的几率。

- 路径长度。这需要URL的路径长度。网络钓鱼的URL往往比合法的URL有更长的路径。

- URL熵。计算URL熵。一个URL的熵越高，它就越复杂。由于钓鱼网站往往有随机文本，我们可以尝试通过其熵来找到它们。

- 长度比。计算URL长度和路径长度之间的比率。在[20]中，他们得出结论，钓鱼网站的比率往往比合法网站高。

- '@'和'-'计数。对URL中的@和-字符进行计数。根据[20]，该功能被添加。在URLs中，@左边的所有内容都会被忽略。有鉴于此，钓鱼网站的URL使用它来欺骗用户。例如goodURL.com@phishURL.com。

- 标点符号计数。. !"的计数。# $ % & , . ; ' 在URL中。在[20]中，他们发现钓鱼网站通常显示出更多的标点符号计数。

- 其他顶级域名计数。出现在URL路径中的顶级域名的数量。钓鱼网址试图通过在路径中使用其域名和TLD来冒充合法的网址。

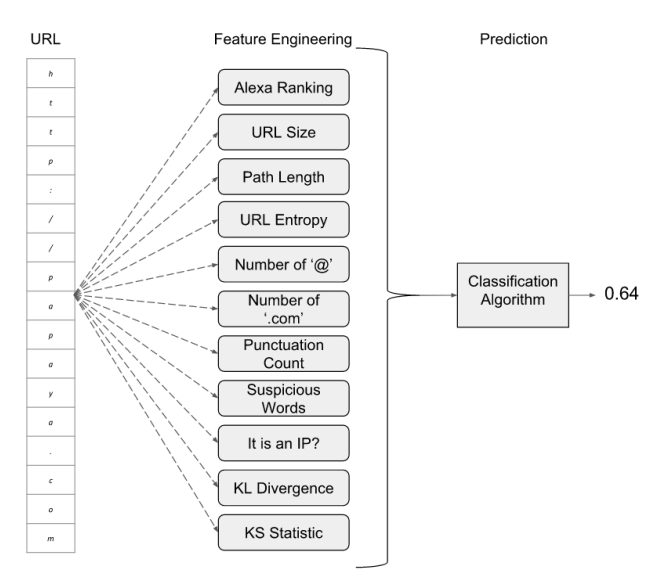
- 是IP：如果URL是一个IP而不是一个域名。这是在文献中已经使用的一个特征。

- 可疑字数。URL中的可疑词的数量。可疑词语包括'确认'、'账户'、'安全'、'webscr'、'登录'、'签到'、'提交'、'更新'、'登录'、'安全'、'wp'、'cmd'和'admin'。它们是通过观察钓鱼网站手动选择的。

- Euclidean距离。URL中英文字符之间的欧几里得距离。这个特征和下面两个特征试图衡量URL与普通英语的差异程度。

- Kolmogorov-Smirnov统计学。计算URL和英语之间字符频率的双样本Kolmogorov-Smirnov统计。

- Kullback-Leibler分歧。计算URL和英语之间的字符频率的Kullback-Leibler分歧。



图五：基于url特征的检测分类

**4.2.2 分类**

一旦特征被提取出来，就用所提出的URL特征来训练一个二元分类器。本文使用随机森林（RF）方法来实现这一目标。随机森林是一种分类算法，它依靠较弱的模型来建立一个较强的模型，并以较弱的模型反应的平均值为依据。所用的较弱的模型是分类树，每一个模型都根据特征值递归地分割数据集，当所有输入实例都属于同一类别时，就停止当前的分割。我们选择了RF，因为它被广泛使用，并且可以被训练为并行运行。

**5.结论：**

尽管已经提出了许多（机器学习）方法来检测钓鱼网站，但人们忽略了攻击者可以使用规避技术来化解它们。第一种方法中解决了规避后检测钓鱼网站的重大问题。在将URLs分割成单词并创建一个由交叉相关实体组成的异质网络后，并配备了定制的边缘潜力机制。此外，我们还回顾了其他的钓鱼网站监测方法，一种是基于网页内容特征的钓鱼网址检测方法，它更关注网页的CSS样式，另一种是使用递归神经网络对钓鱼网站进行分类，它着重url的词汇分类，并且基于url特征进行神经网络分类。可以看到，机器学习检测钓鱼网站有一定作用，并且具有优化的可能性。

**文献：**

[1] Kim T, Park N, Hong J, et al. Phishing URL Detection: A Network-based Approach Robust to Evasion[C]//Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. 2022: 1769-1782.

[2] Bahnsen A C, Bohorquez E C, Villegas S, et al. Classifying phishing URLs using recurrent neural networks[C]//2017 APWG symposium on electronic crime research (eCrime). IEEE, 2017: 1-8.

[3] Mao J, Tian W, Li P, et al. Phishing-alarm: Robust and efficient phishing detection via page component similarity[J]. IEEE Access, 2017, 5: 17020-17030.

[4] Berners-Lee T, Masinter L, McCahill M. Uniform resource locators (URL)[R]. 1994.

[5] Manadhata P, Yadav S, Rao P, et al. Detecting malicious domains via graph inference[C]//Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop. 2014: 59-60.

[6] Justin Ma, Lawrence K Saul, Stefan Savage, and Geoffrey M Voelker. 2009. Be yond blacklists: learning to detect malicious web sites from suspicious URLs. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery

and Data mining (KDD). 1245–1254.