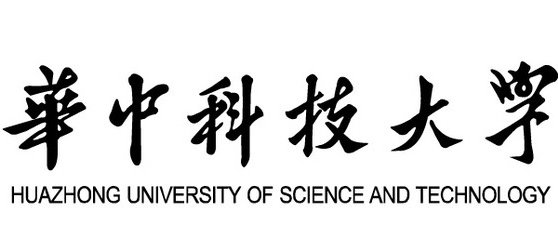
|  |  |
| --- | --- |
|  |  |
| **学 号：** | U201513350 |

****

**论 文 翻 译**

|  |  |
| --- | --- |
| **题 目** | Feature Selections for the Machine Learning based Detection of Phishing Websites |
| **出 处** | 978-1-5386-1880-6/17/$31.00 ©2017 IEEE |
| **学 号** | U201513350 |
| **班 级** | 电信1506 |
| **姓 名** | 刘晓 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 2018 | 年 | 6 | 月 | 10 | 日 |

**基于机器学习的钓鱼网站检测特征选择**

Ebubekir Buber, Önder Demir Marmara University Institute of Pure and Applied Sciences Computer Engineering Department Istanbul, Turkey ebubekirbbr@gmail.com, [odemir@marmara.edu.tr](mailto:odemir@marmara.edu.tr)

Ozgur Koray Sahingoz

Turkish Air Force Academy

Computer Engineering Department

Istanbul, Turkey

sahingoz@hho.edu.tr

摘要-钓鱼网站是恶意网站，它们冒充合法网页，目的是揭示用户重要信息，如用户ID、密码和信用卡信息。这些钓鱼网站的检测是一个非常具有挑战性的问题，因为网络钓鱼主要是基于语义的攻击，特别是滥用人类的脆弱性，而不是网络或系统漏洞。作为软件检测方案，两种主要的方法被广泛使用：黑名单/白名单和机器学习方法。机器学习解决方案能够检测零小时网络钓鱼攻击，并且他们对新类型的钓鱼攻击有很强的适应能力，因此他们是首选。要使用这种类型的解决方案，必须仔细选择输入特性。解决方案的整体性能取决于这些特征。因此，本文的目的是列出和识别基于机器学习的钓鱼网站检测的重要特征。

索引词-钓鱼，机器学习，特征，URL，域名。

**一、引言**

在过去的几十年里，互联网技术从网络社交网络发展到网络电子商务和网络银行技术，使人们的生活更加舒适。由于这种不可控的增长，网络系统出现了许多安全威胁：其中最常见的是“网络钓鱼”。网络钓鱼是一种基于网络的攻击，攻击者试图通过发送来自一个有信誉的实体或个人的电子邮件或与其他渠道进行通信来揭示一些敏感信息，如用户ID /密码或账户信息。许多用户不知不觉地点击钓鱼域名每天和每小时。袭击者不仅瞄准了用户，也瞄准了公司。根据2014年2月发布的第三微软安全指数报告（1），网络钓鱼攻击的年度影响可能高达50亿美元。造成这种巨大成本的主要原因是用户缺乏意识。但安全防范者必须采取预防措施，以防止用户从这些有害场所遭遇。为了防止网络钓鱼的破坏，这些防御者主要是试图提高公司的意识，建立强大的安全机制，可以在防止网络钓鱼攻击之前对其进行检测和预防。进行检测和预防。

在网络钓鱼攻击中，受害者通常会收到一个看起来是由一个已知的联系人或组织发送的电子邮件。此消息包含一些Web链接，其目标是针对用户计算机的某些恶意软件，或者与直接受害者对恶意网站的链接，以欺骗他们泄露个人和财务信息，例如密码、信息系统（IS）或信用账户。卡片细节。

根据〔15〕，人们有五个主要原因导致中招网络钓鱼；

1他们不知道URL（统一资源定位符）及其用法。

2他们不知道哪些URL可以被信任。

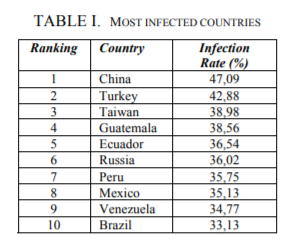
3由于重定向或隐藏URL，它们不访问/查看目标URL

4他们可以意外地点击一些URL，或者他们没有太多时间来咨询URL。

5他们无法区分合法URL和网络钓鱼

然而，所有国家都受到掠夺者的攻击，特别是发展中国家是袭击者的主要目标。反钓鱼工作组（APWG）的钓鱼活动趋势报告在最后一个季度2016强调世界上感染最多的国家是中国。其次是土耳其和台湾，如表1（16）所示。

978-1-5386-1880-6/17/$31.00 ©2017 IEEE



如果我们从攻击者的角度来看，网络钓鱼攻击背后的主要动机集中在三个主要类别〔2〕；

·财务收益：攻击者可以利用被盗凭据获得财务优势

·身份隐藏：攻击者可以出售这些被盗身份（如用户ID /密码），而不是直接使用他们，任何其他罪犯正在寻求一种方式来隐藏他们的活动和身份。

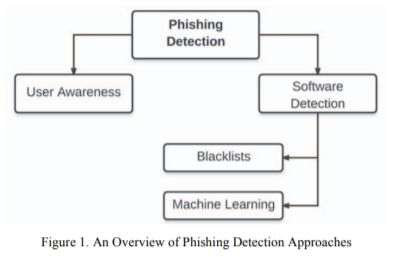
·名声和恶名：攻击者可能为了同伴身份而攻击受害者

网络钓鱼在攻击者之间很流行，因为通过激发点击恶意链接似乎比试图突破计算机的防御机制更容易欺骗某人。消息正文内的恶意链接被设计为使它们看起来使用该组织的标志或者其他合法材料进而欺骗组织及个人。

网络钓鱼攻击可以被认为是一个分层的问题，技术和人类的层面。有效的缓解将需要解决这两个层的问题。

由于网络钓鱼攻击目标利用人类（即终端用户）发现的弱点，所以很难有效地减轻它们。根据[ 3 ]，最终用户未能检测到近30%的网络钓鱼攻击，即使在受过良好的用户意识教育培训。设计强大的钓鱼检测系统以及提高用户的意识也是很重要的。

一旦检测到网络钓鱼攻击，可以对攻击应用多个动作，并且很容易降低检测到的钓鱼攻击。我们的研究重点是检测钓鱼活动的技术。网络钓鱼检测方法分为两个主要组，因为它是一个两层结构：人类层和技术层。在图1中给出了钓鱼检测方法的概述。



本研究的主要重点是覆盖钓鱼攻击检测的软件检测方法。第二部分介绍了基于黑名单的网络钓鱼检测方法，分析了仿冒领域，并从合法的领域中提取其区别特征，并在第三部分中说明了如何利用机器学习和自然语言处理技术来检测网络钓鱼。

**二 黑名单钓鱼检测**

黑名单经常更新先前检测到的钓鱼网址或IP地址列表。黑名单通常比基于机器学习的钓鱼检测系统具有更低的假阳性（FP）率[4 ]。然而，黑名单不提供保护，以防止零小时钓鱼攻击。如（5）中提到的，黑名单只能检测到20%的网络钓鱼攻击。因此，黑名单被发现对检测零小时钓鱼攻击无效。

研究（4）表明47%到83%的钓鱼URL在12小时后被列入黑名单。这个时间间隔非常重要，因为63%个钓鱼活动在2小时内结束。

有一些黑名单提供商，如谷歌安全浏览API（6），基于DNS的黑名单，PHISNET（7），自动个人白名单[ 8 ]。

谷歌安全浏览API允许客户端应用程序验证给定的URL是否存在于谷歌不断更新的黑名单中[6 ]。该服务包括2个黑名单，分别命名为Goog-PHISHIVAR和GOG恶意软件Savar分别用于网络钓鱼和恶意软件。谷歌安全浏览服务尚未完全完成。然而，它被谷歌浏览器和Mozilla Firefox使用。

基于DNS的黑名单（DNSbl）提供者使用标准DNS协议。当建立SMTP连接时，系统可以通过向DNSbl服务器发送DNS RR查询来验证连接源是否在钓鱼黑名单（9）中列出。由于使用了标准DNS规范，任何配置的DNS服务器都可以充当DNSBL。

为了检测钓鱼URL，需要在分析的URL和黑名单中的任何URL之间进行精确匹配。如果对钓鱼网络URL进行任何更改，将导致没有匹配。这意味着，即使是高度类似钓鱼网络URL的URL也不能被检测为PHISH。为了解决精确匹配限制，PHISNET（7）进程黑名单URL（父母），并通过5种不同的方法产生父URL的多个变体。

自动化的个人白名单（AIWL）维护一个白名单，描述用户提交他/她的凭据的可信登录用户界面（路易斯）。每一个LUI都会发出警告，除非被信任。如果标记为可信的LUI，其特征将被本地存储在白名单中。AIWL的结构由两个主要部分组成：白名单，自动白名单维护者。白名单是可信路易斯的列表。自动白名单维护者跟踪用户登录活动并决定URL是否可疑。维护者的分类度量是终端用户通过目标LUI成功地登录足够数量的时间，然后信任该LUI。当一个LUI被检测为可信时，它被添加到白名单中。

1. **机器学习的钓鱼检测**

通过机器学习钓鱼网页检测的钓鱼检测可以被认为是文档分类或聚类问题，其中利用K-近邻（K-NN）、决策树等机器学习算法来创建模型；随机森林、支持向量机（SVM）、k-均值和基于密度的具有噪声的空间聚类（DBSCAN）等。

（如K-均值聚类算法，DBSCAN算法的实例在一个unsupervised）分区的方式。对构建的集群，它是不需要知道实例的类标签）。这些算法的实例，根据实例的聚类之间的相似。主要用途是类似的单实例是一组集群。

一般的聚类算法基于一些假设。例如，有两个单独的类的网络钓鱼检测问题。一个类的实例代表钓鱼而其他实例则代表合法的。假设说，具有低实例计数的集群表示钓鱼实例，并且具有高实例计数的集群代表合法的实例。这是钓鱼检测算法的主要假设。

检测钓鱼域名被认为是一个分类问题。因此，需要标记样本作为训练阶段中的仿冒域和合法域的样本。在训练阶段使用的数据集是建立成功的检测机制的关键点之一。检测系统应该使用准确知道类别的样本。因此，被标记为钓鱼的样本必须被绝对检测为PHISH。同样，被标记为合法的样本必须被绝对检测为合法的。否则，如果我们使用不确定类信息的样本，系统就不能正常工作。为此，为钓鱼而创建了多个公共数据集。一些著名的是PhishTank（10）和TechHelpList [ 11 ]。这些数据源是学术研究中常用的数据源。

收集合法域名是另一个问题。为此，通常使用站点信誉服务。网站信誉服务分析和排名可用的网站。这个排名可能是全球性的或可能是基于国家的。排名机制取决于各种各样的特征。具有高等级分数的网站被认定为经常使用的合法站点。著名的信誉排名服务之一是Alexa（12），研究人员正在使用Alexa的顶级列表作为学术研究中钓鱼检测的合法站点。

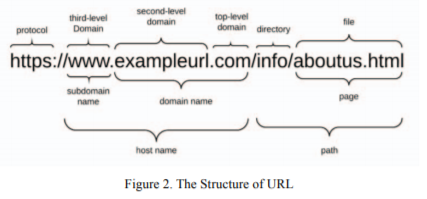
当我们有钓鱼和合法网站的原始数据时，下一步应该是处理这些数据并从中提取有意义的信息来检测欺诈性域名。用于机器学习的数据集必须包括这些特征。因此，我们必须处理从Alexa、PiStCar或其他数据资源中收集的原始数据，并创建一个新的数据集，用机器学习算法来训练我们的系统。价值应该根据我们的需要和目的来选择，并且应该为每个人计算。

在钓鱼检测问题中，我们有两个类：“钓鱼”和“合法”。

为了建立检测机制，系统根据标签的需要和目的计算出我们选择的特征，并将其用于训练。然后，通过系统对没有标签的实例进行分类。

**A.钓鱼领域的特点**

要了解攻击者在创建仿冒域时的想法，需要知道URL结构。创建统一资源定位器（URL）来处理网页。图2显示了典型URL的结构中的相关部分。

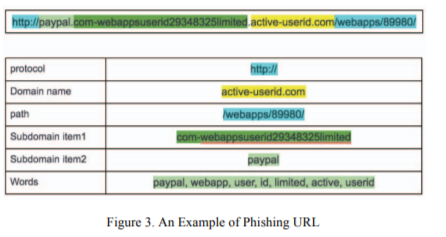


URL从用于访问页面的协议开始。完全限定的域名标识宿主Web页面的服务器。它包括一个注册域名（第二级域名）和后缀，我们称之为顶级域名（TLD）。域名部分受到限制，因为它必须与域名注册机构注册。主机名由子域名称和域名组成。PHisher对子域部分有完全控制，并可以对其设置任何值。URL也可以有路径和文件组件，也可以由PHEDER随意更改。子域名和路径完全由PHIDER控制。在这项研究中，我们使用术语“自由URL”来引用URL中的那些部分来继续文章。

攻击者可以注册以前没有注册的任何“域名”。URL的这一部分只能设置一次。PHisher可以随时更改免费URL以创建新的URL。网站的独特部分是“域名”，这就是为什么安全捍卫者挣扎着检测网络钓鱼“域名”。当域被检测为欺诈时，在用户访问该域之前很容易阻止该域。

一些威胁情报公司（13, 14）检测并发布欺诈网页或IPS作为黑名单，从而防止这些有害资产被其他人变得越来越容易。

攻击者必须智能地选择域名，因为目标应该是说服用户，然后设置免费的URL，使检测困难。图3给出了钓鱼URL的一个例子。

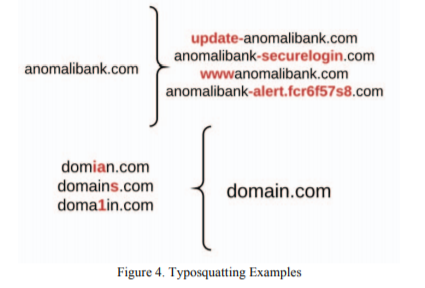


在例子中给出的图，攻击者试图使域名看起来像“贝宝。com”加入免费网址，虽然真正的域名是“paypal.com”。当用户看到“paypal.com”开头的网址，他们可以信任的网站连接，然后向该欺诈网站分享他们的敏感信息。这是攻击者常用的方法。

攻击者经常使用的其他方法是网络抢占和抢注。

域名抢注（也称为域名抢注），是注册、贩卖或使用具有恶意的域名，目的是从属于他人的商标的善意中获利。域名抢注者可以将域名出售给一个人或某公司，该人或公司以一个夸大的价格拥有包含在商标内的商标，或者可以将其用于诸如钓鱼的欺诈目的。例如，你的公司的名字是“ABCORE”，你把域名注册为“ABCONT.COM”。然后，钓鱼者可以注册“abc.net”、“abcor.org”、“abcb. Biz”，他们可以用它来欺诈。

TypQuice，也称URL劫持，是一种网络抢占的形式，它依赖于错误，例如当用户将网址输入到网络浏览器或基于快速阅读时很难注意的排版错误时，互联网用户所犯的排印错误。用Typosquatting创建的URL看起来像是一个可信域。用户可能意外地输入不正确的网站地址或点击看起来像可信域的链接，这样，他们可以访问由PHisher所拥有的另一个网站。一个著名的例子是Goggel.com，一个极其危险的网站。另一个是Yuube网站，它就像GoggLeGo网站，但它的目标是YouTube用户。类似地，www. ArrFr.com被称为“www. ARIFRANC.com”，将用户转移到一个叫卖折扣旅行的网站。其他一些例子；PayWPAL.com，MyRoFoT.com，Apple Lead，AppIE.com。另一个例子如图4所示。



**B.用于钓鱼URL检测的特征**

在文献和商业产品中，有很多算法和各种各样的钓鱼类型检测的数据类型。网络钓鱼URL和相应的页面有几个区别于恶意URL的特征。例如，攻击者可以注册一个冗长而混乱的域名来隐藏实际域名（域名抢注，Typosquatting）。在某些情况下，攻击者可以使用直接IP地址而不是使用域名。这种类型的事件超出了我们的范围，但它可以用于相同的目的。攻击者还可以使用与任何合法品牌无关的短域名，并且不具有任何免费URL添加。这些类型的网站也超出了我们的范围，因为它们与欺诈域名更相关，而不是钓鱼域名。

在学术研究的检测过程中使用了机器学习算法中的各种特征。从学术研究收集的特征与机器学习技术的钓鱼领域检测分组如下。

1.基于URL的特征

2.基于领域的特征

3.基于页面的特征

4。基于内容的特征

在这部分的延续中解释了这些特征组。

1. 基于URL的特征
2. URL是分析网站以确定它是否是钓鱼网站的第一件事。网络钓鱼域名的URL有一些独特的点。当处理URL时，获得与这些点相关的特征。下面给出了一些基于URL的特性。

·URL中的数字计数

·URL的总长度

·检查URL是否为Typosquatted

·检查是否包括合法品牌

·URL中的子域数

·TLD是常用的一种吗？

2）基于领域的特征

钓鱼域名检测的目的是检测钓鱼域名。因此，与域名相关的被动查询，我们希望分类为钓鱼或不提供有用的信息给我们。下面给出一些有用的基于领域的特征。

·域名还是其IP地址在知名信誉服务黑名单中？

·从域名注册到现在有多少天了？

·注册者姓名是否被隐藏？

3）基于页面的特征

基于页面的特征使用关于计算信誉排名服务的页面的信息。这些特性中的一些给我们提供了一个网站是多么可靠的信息。下面给出了一些基于页面的特性。

·全球网页排名

·国家网页排名

在Alexa 前1百万的网站的排名位置

一些基于页面的特性告诉我们关于目标站点上的用户活动的信息。下面给出这些特性中的一些。获取这些类型的特征并不容易。有一些付费服务来获得这些类型的特征。

·估计每天的、每周的或每月的域的访问次数

·每次访问的平均浏览量

·平均访问时间

•每个国家的网络流量共享

·计算从社交网络到给定域的引用

·领域类别

·类似的网站等。

4）基于内容的特征

获得这些类型的特征需要主动扫描到目标域。页面内容为我们处理，以检测目标域是否用于钓鱼。下面给出了一些关于页面的处理信息。

·页面标题

·元标签

·隐藏文本

·正文中的正文

·图像等

通过分析这些信息的网页截图，我们可以收集信息，如：

·是否需要登录网站？

·网站类别

·有关观众简介等信息。

所有这些解释的特征对于钓鱼领域检测是有用的。在某些情况下，由于某些限制，使用这些特征可能不是合乎逻辑的。例如，使用基于内容的特征来开发快速检测机制可能是不合乎逻辑的，该机制能够在一天内分析100到200之间的域的数量。另一个例子，如果我们想分析新注册的域名，基于页面的功能不是很有用。因此，检测机构所使用的特征取决于检测机构的目的。在检测机构中需要仔细选择哪些特征。

**四、结论**

网络钓鱼是一个主要的问题，它利用社会工程和技术欺骗来获取用户的重要信息，如财务数据、电子邮件和其他私人信息。网络钓鱼利用人的脆弱性，因此，大多数保护协议不能阻止整个网络钓鱼攻击。他们中的许多人使用黑名单/白名单的方法，但是，这不能检测到零小时钓鱼攻击，并且他们无法检测到新类型的网络钓鱼攻击。因此，机器学习用于钓鱼检测是一个在这一领域有效果和有效率的工具。使用ML方法，需要大量的（标记）数据，此外，这些数据的特征是非常重要的。因此，本文的目的是列出和识别基于机器学习的钓鱼网站检测的重要特征。

参考文献

1. Microsoft, “Microsoft Security Index Report, ”https://news.microsoft.com/en-sg/2014/02/11/microsoftconsumer-safety-index-reveals-impact-of-poor-online-safetybehaviours-in-singapore/#sm.0000c8bivc14h3dyfvxak6545kbcz #4FXDf2H3VbYmD1b1.97 , accessed May 2017.
2. W. D. Yu, S. Nargundkar, and N. Tiruthani, "A phishing vulnerability analysis of web based systems." in Proceedings of the 13th IEEE Symposium on Computers and Communications (ISCC 2008). Marrakech, Morocco: IEEE, July 2008, pp. 326- 331.
3. S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs, “Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions,” in Proceedings of the 28th international conference on Human factors in computing systems, ser. CHI ’10. New York, NY, USA: ACM, 2010, pp. 373–382.
4. S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang, “An empirical analysis of phishing blacklists,” in Proceedings of the 6th Conference in Email and Anti-Spam, ser. CEAS’09, Mountain view, CA, July 2009.
5. Khonji, M., Iraqi, Y. and Jones, A., 2013. Phishing detection: a literature survey. IEEE Communications Surveys & Tutorials, 15(4), pp.2091-2121.
6. Google, “Google safe browsing API,” http://code.google.com/apis/safebrowsing/ , accessed May 2017.
7. P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, “Phishnet: predictive blacklisting to detect phishing attacks,” in INFOCOM’10: Proceedings of the 29th conference on Information communications. Piscataway, NJ, USA: IEEE Press, 2010, pp. 346–350.
8. Y. Cao, W. Han, and Y. Le, “Anti-phishing based on automated individual white-list,” in DIM ’08: Proceedings of the 4th ACM workshop on Digital identity management. New York, NY, USA: ACM, 2008, pp. 51–60.
9. Rbldnsd, http://www.corpit.ru/mjt/rbldnsd.html, accessed on May 2017.
10. PhishTank, https://www.phishtank.com/, accessed on May 2017.
11. TechHelpList, https://techhelplist.com/pastes/, accessed on May 2017.
12. Alexa, http://www.alexa.com/about, accessed on May 2017.
13. Cymon, https://cymon.io/, accessed on May 2017.
14. Firehol, http://iplists.firehol.org/, accessed on May 2017.
15. M. Volkamer, K. Renaud, B. Reinheimer, A. Kunz, “User experiences of TORPEDO: TOoltip-poweRed Phishing Email DetectiOn”, Computers & Security (2017), doi: 10.1016/j.cose.2017.02.004
16. Anti-Phishing Working Group (APWG), ‘‘Phishing activity trends report — last quarter 2016. http://docs.apwg.org/reports/ apwg\_trends\_report\_q4\_2016.pdf, accessed on May 2017

个人观点与评论：恶意网页识别是信息安全领域的热点问题，随着网络攻击技术和防御技术的不断发展，该问题一直受到研究人员的广泛关注。在网络攻击日益复杂、恶意代码层出不穷的今天，机器学习算法在恶意代码检测中的应用逐渐受到学术界很大重视。本文对基于机器学习算法识别电鱼网页检测的特征选择和方案进行了梳理，这一工作将为新型基于机器学习的钓鱼网站检测的设计和实现提供参考。

但基于机器学习的钓鱼网站检测领域仍属于发展阶段，还存在着许多未来工作和挑战，我查阅了相关资料和学习了相关知识后发现如下对于基于机器学习的钓鱼网站检测的有待提升指出：机器学习算法基于好的特征选择虽然可以提供高准确率的恶意网页分类，但分类器一般作为黑盒机制被加以使用，安全人员缺乏对结果的理解。结果往往在不质疑分类器性质的情况下直接被使用，因此分类结果受经验阈值和数据特征的影响，出现一定倾向性。所以研究传统量化分析（如准确率、误报率）之外的统计学方法，如可信度，科学地评价和比较底层的机器学习算法，还有待提高到新的水平。