

QR 二维码防钓鱼的研究

柴艳娜

(长安大学 信息与网络管理处, 陕西 西安 710064)

摘要: QR 码(quick response code)是一种简单易用的矩阵条码。随着移动互联网的崛起和繁荣,已经广泛应用于人们的日常活动中。它给人们带来便利的同时,还伴有钓鱼网站、病毒软件、信息泄漏等网络安全风险和恶意攻击。越来越多的钓鱼攻击由传统的诱导式电子邮件转变为一扫即开的二维码,堂而皇之地在移动互联网世界里游荡。新闻稿中也屡见各种伪造二维码、钓鱼二维码导致人们的信息和财产受到损失的报道。目前流行的具有支持扫码功能的 App 如微信、支付宝等均无法有效甄别钓鱼网站。文中分析了 QR 码的基本结构和原理,钓鱼网站以及防范钓鱼的传统技术,从 URL 结构和网页内容两个方面分析钓鱼网站的异常特征,并且相应地给出检测方法和数学模型,设计并实现 Android 平台的 QR 码钓鱼网站识别技术。

关键词: 网络安全; QR; 二维码; 钓鱼网站; 安卓

中图分类号: TP309

文献标识码: A

文章编号: 1673-629X(2020)04-0100-05

doi: 10.3969/j.issn.1673-629X.2020.04.019

Research of Anti-phishing for QR Code

CHAI Yan-na

(Dept. of Information and Network Management, Chang'an University, Xi'an 710064, China)

Abstract: QR code (quick response code), as a type of matrix barcode (or two-dimensional barcode), has become popular due to the rise and prosperity of mobile Internet. It brings convenience to people at the same time, but also accompanied by network security risks and malicious attacks like the phishing website, virus software, data leak and so on. More and more phishing website show up as faked QR code in mobile Internet instead of classical e-mail, and you can see news that the faked QR code causes money losing and data leak cases. Currently, most popular apps with support for code scanning, such as Wechat, Alipay, can't identify phishing website in QR code. Therefore, we analyze basic structure and working way of QR, research phishing website and the classical way to prevent. We can figure out exception behavior from URL structure and source code of web page from phishing website, then point out the detective way and math model, design and implement a platform to prevent phishing website for Android.

Key words: network security; QR; QR code; anti-phishing; Android

0 引言

中国互联网络信息中心(CNNIC)发布的第42次《中国互联网络发展状况统计报告》指出^[1],截至2018年6月30日,中国手机网民规模已达7.88亿,网民通过手机接入互联网的比例高达98.3%。另据美国We Are Social和Hootsuite的2018年全球数字报告,截止2018年1月,全球范围内移动设备用户已达51.35亿,贡献了52%的Web流量^[2]。人们已然处于移动互联网时代,在享受移动智能设备带来的便利的同时,也把自己暴露在日益严峻的安全风险当中。

根据360公司发布的中国手机安全状况报告,传

统的钓鱼网站、骚扰电话与垃圾短信等信息安全问题依然是影响Android用户的常见问题。2017年全年360手机卫士共拦截各类钓鱼网站攻击28.8亿次。

QR码由于其存储容量大,易于读取识别等优点,在国内被广泛应用于各个领域,尤其是移动支付方面。但是正是由于其应用的广泛,编解码模式的开放,让其成为钓鱼攻击的良好载体,极易被不法分子所利用。

2017年,在昆明、重庆等地相继发生了共享单车QR码被恶意替换的事件。不法分子将假冒的押金支付页面链接存储在二维码中,然后替换真正的二维码。当用户扫描后就会进入假冒押金支付页面,由于假冒

页面与真正的页面高度相似,致使用户不能准确做出判断,造成财产损失。所以如何有效地识别钓鱼网站并作出安全响应具有十分重大的意义。

1 QR 码与钓鱼网站分析

1.1 QR 码的结构

QR 码是由黑色方块排布在白色背景的正方形网格上所组成的,如图 1 所示。扫码设备扫描 QR 码图片并根据里德-所罗门纠错算法(Reed-Solomon error correction)处理直到图片被正确解析。码图里的方块在垂直方向和水平方向上面的不同分布代表着不同的信息内容,所以 QR 码也叫二维码^[3]。QR 码可以方便地存储数字、字母、比特或者二进制和汉字。因此,QR 码被广泛用于文本消息、URL 链接以及随移动支付兴起的支付宝或微信等支付实体信息。



图 1 QR 码图样

QR 码的左上、右上及左下有 3 个“回”字型定位块,这是 QR 码的必备图形之一。其他的还有校准信息、时序信息、版本信息、格式信息以及编码过的数据信息和纠错码。图 2 是 QR 码总体结构示意图。

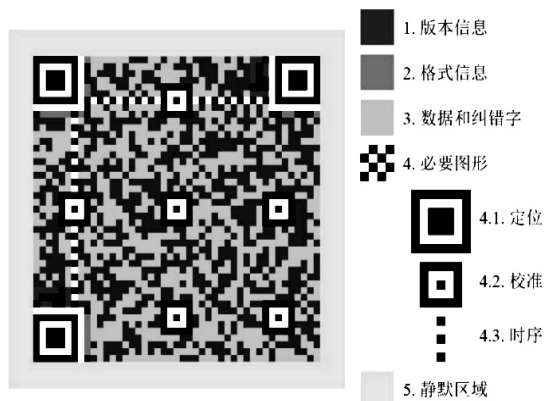


图 2 QR 码结构组成

QR 码有 40 个版本,相应的方块矩阵规格大小为 21×21 至 177×177。每个版本的 QR 码的边长都比上一版本的多 4 个模块,因此存储的数据也更多。最大的版本 40 可以存储 984 个 UTF-8 编码中文字符,4 096 个英文字符。

1.2 钓鱼网站

钓鱼是指伪装成网络中的一个可信实体,试图获取用户重要隐私和敏感信息的恶意行为,如窃取用户名、密码、信用卡及电子支付信息等。钓鱼通常是通过欺诈邮件、即时信息或短信等,将用户引诱前往以假乱真的伪造网站上输入个人信息^[4]。伪造网站和合法网站从表面看一摸一样,唯一的区别就在于不同的 URL 地址。因此,如果一个真实合法的二维码被替换成带有钓鱼链接的二维码,那么用户在使用的时候就很容易上当受骗。用户很难识别出两个二维码的不同之处,而且很多 App 在设计时,对于识别的二维码也做不到有效的甄别,导致用户信息面临被钓鱼的风险^[5]。

1.2.1 钓鱼网站 URL 特征

钓鱼网站的 URL 主要有以下几个常见特征:

(1) URL 中经常出现 IP 地址而不是域名(FQDN)。

一般正规合法的网站都会给自己申请一个简单易记识别的域名,在推广宣传方面可以带来很多便利。钓鱼网站很少会在这方面增加支出^[6]。

(2) URL 里常出现非标准端口。

标准的 HTTP 和 HTTPS 协议,其端口分别为 80 和 443。在默认情况下,URL 里省略端口,系统会按标准端口去尝试访问。如果使用了非标准端口,则须在 URL 明确指定端口,这样会加长 URL,导致普通用户记忆和使用上的不方便,正规合法网站基本上都会使用标准端口。而钓鱼网站在很多时候会租用一些便宜的共享主机或网站空间,因此很大概率使用的都是非标准端口。

(3) URL 中的域名层级超过 6 级。

www.taobao.com 是一个知名电商网站的三级域名。一般正规网站都会控制 URL 的长度以方便用户使用,所以域名的层级不会很深。有统计数据表明,当域名层级超过 6 级时^[7],有极大的可能性这是一个钓鱼网站。

(4) 使用短域名或者 URL 长度过长。

随着短链接(URL Shortener)的普及,很多钓鱼网站会用短链接隐藏自己的真实 URL,不仅缩短了 URL 长度易于传播,而且增加了用户识别的难度,只有将短链接还原后才能发现其真实面目。常见的短链接服务商有 t.cn, bit.ly, goo.gl, t.co 等。

(5) URL 中可能包含一些特殊词汇或符号。

钓鱼网站一般都会模仿登陆界面或支付界面等,以此引诱用户。所以相应的,URL 里可能会出现 login, signin, signon, signup, pay, order, buy 等单词^[8]。同时,钓鱼网站可能会重定向,可能会携带一些伪造参数,因此其 URL 极可能含有一些不常见的符号,如

“@”。

(6) 钓鱼网站的域名与正规合法网站的域名很相似。

一些做的更加逼真的钓鱼网站,它们的网址也会尽可能模仿正规网站的,真假李逵,难以辨别。一般它们会将正规网站域名中的个别字符进行近似的替换,比如将字母“o”替换成数字“0”,将字母“l”替换成数字“1”,以达到混淆视听的目的^[9]。比如用 ta0ba0.com 来模仿淘宝网的域名制作一个钓鱼网站。

1.2.2 钓鱼页面特征

钓鱼攻击者会精心设计伪造目标网站的页面和功能,因而在页面代码中会留下一些关键的特征,常见的有:

(1) a、link 及 img 等资源链接标签的 href、src 属性通常指向域外攻击者一般会直接将目标网站的前端页面代码保存下来做少量修改,以此进行快速伪造,因此会大量复用目标网站的静态资源,如 CSS 以及各种

图片。所以,有很大的可能是钓鱼网站的静态资源,会跨域引用到正常网站的资源^[10]。

(2) ICP 经营许可证。

所有的网络内容服务商(ICP),包括网站,在中国大陆的法律法规下,其必须取得 ICP 经营许可证,在网站上必须注明 ICP 备案号。通过向主管机关查询备案号,可以获得网站的注册信息,其中便包含有域名信息(见图 3),便可据此进行分辨。



图 3 ICP 备案号

2 钓鱼网站检测

2.1 检测模型

QR 二维码防钓鱼系统的主要框架如图 4 所示。

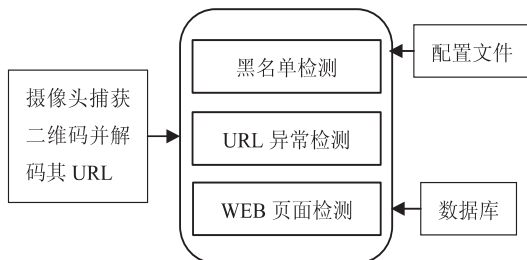


图 4 防钓鱼主要框架

(1) QR 扫描解码。

安卓 App 可以利用智能手机的摄像头对 QR 码进行拍照,扫描并读取其内含信息。由于 QR 码协议本身的开放,现在开源社区存在很多活跃高效的 QR 码图片处理库可以直接使用,比如 ZXing,就是一个支持多种格式的 1D/2D 条码图像处理库。所以,系统可以直接利用 ZXing 将摄像头拍摄的 QR 码进行解析,从而提取信息以作后续分析之用。

(2) 黑名单匹配检测。

传统的钓鱼网站识别方法之一便是黑名单技术,各大浏览器内置的一些安全机制都会根据黑名单来检查用户访问的链接是否安全从而给出相应的提醒^[11]。

PhishTank、中国反钓鱼联盟及安全联盟等都是较为权威的反钓鱼组织。个人用户都可以将自己遇到的钓鱼网址在这些组织上进行分享,使其他用户可以提高警惕^[12]。因此,防钓鱼系统可以直接利用这些已经被鉴定的钓鱼网址,分析并创建一份黑名单列表,当用户所扫描的 QR 码信息中含有黑名单里的 URL 时,则直接对用户进行风险提示。

黑名单匹配检测的实现技术简单且可重复利用,

可以避免对已鉴定的钓鱼网址进行重复判别,节约了大量的时间成本。黑名单必须保持更新,更新频率越高,鉴定样本越多,则黑名单的命中率越高。黑名单的更新需要依赖对钓鱼网址的发现,这天然存在的滞后性,在日益复杂的网络世界里,愈发制约其有效性。

2.2 异常特征检测

如果用户扫描 QR 码后所得的 URL 在黑名单中匹配失败,则它就是一个安全性未确定的 URL。上文已经对钓鱼网站的 URL 特征和页面特征进行过分析,对这些异常特征集合进行研究,可以发现其行为模型^[12]。

一、URL 异常特征向量。

从 URL 地址的结构和词汇两方面出发,可以提取出如表 1 所示的 7 个钓鱼网站 URL 异常特征向量。

(1) F_1 : URL 不使用域名,而是用 IP 地址代替,这是一个钓鱼链接比较明显的特征。判断 URL 中是否含有 IP 地址相对简单,可以用正则表达式直接匹配。对最终的匹配结果可以用函数 f_1 表示:

$f_1 = 1$, URL 中含有 IP 地址;

$f_1 = -1$, URL 中不包含 IP 地址。

表 1 钓鱼网址特征向量

特征向量	特征
F_1	URL 包含 IP 地址
F_2	URL 含有“@”字符
F_3	URL 使用非标准端口,如 80 和 443
F_4	URL 的域名层级过深,其“.”个数超过 6 个
F_5	URL 链接长度超过 23 个字符
F_6	URL 中存在诸多敏感词汇 { account, login, alibaba, taobao, paypal, alipay, ebay, amazon, bank, jd, icbc, apk, mobile, webapp }
F_7	URL 是短链接 sina. lt, t. cn, dwz. cn, goo. gl, bit. ly, qq. cn. hn, tb. cn. hn, jd. cn. hn, j. mp

(2) F_2 :合法的网址一般不会含有“@”,钓鱼链接却经常会使用“@”进行伪装。对于这一特征,可以直接使用正则表达式进行匹配,结果可以用函数 f_2 表示:

$f_2=1$,URL 中含有@;
 $f_2=-1$,URL 中不包含@。

(3) F_3 :钓鱼链接一般会租用便宜甚至免费的网络服务,这些服务一般会使用到非标准的网络端口,如 Web 服务但不是标准的 80 和 443 端口,此时就会在 URL 中显示出来。同样可以使用正则匹配,结果用函数 f_3 表示:

$f_3=1$,URL 中含有端口号;
 $f_3=-1$,URL 中不包含端口号。

(4) F_4 :通常用户访问的网站域名不会超过三级,如果域名中含有 3 个以上的“.”,则很可能为钓鱼网站,检测结果可以用函数 f_4 表示:

$f_4=1$,URL 中“.”的个数大于 3;
 $f_4=-1$,URL 中“.”的个数不大于 3。

(5) F_5 :如果 URL 长度大于 23 的话,根据一些研究成果,有很大概率是一个钓鱼链接,此项检测结果用函数 f_5 表示:

$f_5=1$,URL 长度大于 23;
 $f_5=-1$,URL 长度不大于 23。

(6) F_6 :钓鱼链接 URL 中大多会出现敏感词汇以达到冒充效果,因此可以用正则表达式检测如表 1 所示的敏感词,检测结果可以用函数 f_6 表示:

$f_6=1$,URL 含有敏感词;
 $f_6=-1$,URL 不含敏感词。

(7) F_7 :URL 是短链接,而不是一般网站常用的自己域名的链接,此种特征可以用网址中是否含有常见短链接服务商所提供的域名后缀进行匹配,如表 1 所示,结果为函数 f_7 :

$f_7=1$,URL 是短链接;
 $f_7=-1$,URL 不是短链接。

二、页面异常特征向量。

如 1.2.2 分析,钓鱼网址的页面内容可以提炼出下列 2 个需要关注的异常特征,见表 2。

表 2 钓鱼网页内容特征向量

特征向量	特征
F_8	跨域引用资源
F_9	ICP 注册信息异常

(1) F_8 :一般钓鱼网页会做的与官方网页非常相似,但是其自身的域名肯定与官方域名是不相同的,所以为了与官方保持一模一样的页面和体验,它们一般会照搬官方网页上的资源,诸如图片, CSS 样式, JavaScript 脚本等。因此,可以分析抓取到的钓鱼网页内容,分析其 HTML 结构和标签,检测是否存在引用跨域资源,将结果表示为函数 f_8 :

$f_8=1$,页面内容存在跨域资源;
 $f_8=-1$,页面内容不存在跨域资源。

(2) F_9 :国内的互联网管理条例规定,正规网站的运营肯定会取得 ICP 许可。因此,抓取钓鱼网站页面,检查其是否有 ICP 许可备案号,以及检查该备案号是否真实有效且与网页所在域名一致,将结果表示为函数 f_9 :

$f_9=1$,没有备案或者备案无效或不符;
 $f_9=-1$,真实有效的 ICP 备案。

2.3 Logit 模型

在根据 URL 的异常特征来检测其是否安全时,不同的特征对判断结果的区分度是不同的。这意味着不同的特征向量具有不同的权重。权重大的特征,表现的更明显,对结果影响更大,更易识别和使用。权重越小则越难判定,识别效果就越差。因此可以利用 Logit 模型,将统计分析所得的各个特征向量的权重代入其中,以权重为因子修正模型,获得更加准确的分类结果^[13]。

Logit 模型(Logit Model)也就是逻辑回归,是最

早的、应用最广泛的离散选择模型,是统计实证分析方面的常用方法。Logit 模型求解速度快,是一款高性能的分类评定模型。特征检测实质上也是一个对 URL 根据特征向量进行安全筛查分类的过程。同时,通过曹玖新等人的研究发现,在样本集和向量空间相同的情况下,Logit 模型比其他分类模型的性能更强大^[14]。

Logit 模型其实是个逻辑分布公式,最终获得是一个似然概率,这个概率的计算公式为^[15]:

$$P(Y = 1 | X = x) = e^{\text{logit}} / (1 + e^{\text{logit}})$$

其中, $\text{logit} = \omega_0 + \omega_0 x_0 + \omega_1 x_1 + \cdots + \omega_n x_n$, x_0, x_1, \cdots, x_n 是样本数据的特征变量, $\omega_i, i \in [1, n]$ 则是各个特征变量对应的权重数值。

App 可以分析 URL, 获得的各项数据, 再通过 Logit 模型得到最后的分类概率, 利用这个概率与实验研究验证过的阈值相比较, 如果大于阈值, 则将 URL 判定为钓鱼链接。

3 结束语

QR 码技术的简便快捷, 使得其广泛应用于日常生活中。各种支付场景和信息共享都出现二维码的身影。同时, 其广泛应用所带来的安全风险也与日俱增。文中探讨了 QR 码技术, 并分析了钓鱼链接的特征, 提出根据特征向量, 设计模型对钓鱼链接进行辨识, 以防范 QR 码的钓鱼风险。

参考文献:

- [1] 倪天华, 朱程荣. 网络钓鱼防御方法研究[J]. 计算机技术与发展, 2008, 18(9): 115-118.
- [2] LIU N. China's digital economy: a leading global force[J]. China's Foreign Trade, 2018(3): 20-21.
- [3] 庞 爽. 面向 QR 码的网络钓鱼防御研究[D]. 长沙: 中南林业科技大学, 2016.
- [4] 蔡洪民. 校园网钓鱼邮件监控系统的研究与实现[J]. 计算

机技术与发展, 2013, 23(10): 103-106.

- [5] 周诚诚, 张代远. 利用图像识别技术过滤海量可疑钓鱼网站[J]. 计算机技术与发展, 2012, 22(11): 246-249.
- [6] 林佳华, 杨 永, 任 伟. QR 二维码的攻击方法与防御措施[J]. 信息安全, 2013(5): 29-32.
- [7] 梁雪松. 基于浏览器的钓鱼网站检测技术研究[J]. 信息安全与通信保密, 2007(11): 53-55.
- [8] 黄华军, 钱 亮, 王耀钧. 基于异常特征的钓鱼网站 URL 检测技术[J]. 信息安全, 2012(1): 23-25.
- [9] 庄蔚蔚, 叶艳芳, 李 涛, 等. 基于分类集成的钓鱼网站智能检测系统[J]. 系统工程理论与实践, 2011, 31(10): 2008-2020.
- [10] 周治平, 杜彦辉, 戴明星. 网络钓鱼网站探测系统分析与设计[J]. 计算机安全, 2008(3): 86-88.
- [11] ALSHARNOUBY M, ALACA F, CHIASSON S. Why phishing still works; user strategies for combating phishing attacks [J]. International Journal of Human - Computer Studies, 2015, 82: 69-82.
- [12] BASNET R B, SUNG A H. Mining web to detect phishing URLs[C]//2012 11th international conference on machine learning and applications. Boca Raton, FL: IEEE, 2012: 568-573.
- [13] LI M, ZHU W, JIANG X, et al. A rapid payment confirm scheme base on QRCode[C]//Proceedings of 2014 international conference on industrial engineering and information technology. Beijing: IEEE, 2014: 133-141.
- [14] KROMBHOLZ K, FRUHWIRT P, RIEDER T, et al. QR code security - how secure and usable apps can protect users against malicious QR codes [C]//2015 10th international conference on availability, reliability and security (ARES). Toulouse: IEEE, 2015: 230-237.
- [15] VIDAS T, OWUSU E, WANG S, et al. QRishing: the susceptibility of smartphone users to QR code phishing attacks [M]//Financial cryptography and data security. [s. l.]: Springer, 2013: 52-69.