# 基于内容提取的短链接生成算法研究

薛　富　高一男

（中国人民公安大学网络安全保卫学院　北京　100038）

【摘　要】短网址服务通过将较长的目标网址缩短，来方便人们记忆并分享。社交网络尤其是近年来微博的盛行，使短网址服务获得广泛的应用。然而，现在短网址服务同时被不法分子盯上，他们利用短网址来伪装恶意链接，尤其是钓鱼网站链接，利用微博平台进行快速传播，最终窃取用户敏感信息甚至诈骗钱财，危害互联网安全。本文提出了一种新的短链接生成算法，通过对链接进行分析并在短网址中插入简短的目的网址内容，使用户在点击链接前能够辨识短网址的链接目标，以达到防范网络钓鱼的目的。

【关键词】反钓鱼；短网址生成；内容提取

中图分类号：TP391.41　　文献标识码：A　　文献编号：1009-6833（2014）02-114-02

## Research of URL shortening technology based on content extraction

Xue Fu，Gao Yinan

**Abstract**：URL shortening service facilitates people to remember and share URLs by compressing a URL's length into a few characters. It becomes more and more popular with the widespread of social networks and micro blogging. However，it also hides the information of real URLs as we cannot identify whether a shortened URL is legitimate or malicious. Phishers take use of shortened URLs to hide malicious especially phishing websites，spread these links by micro blogging platform，and ultimately steal sensitive information or money，endangering Internet security. This paper presents a new shortened link generation algorithm，through the analysis of the link and insert the destination URL information into its shortened URL，so that users can identify its destination before click on the shortened URL link. achieving the purpose of preventing phishing attacks.

**Keywords**：anti-phishing；shortened URLs；content extraction

## 0　引言

随着微博获得了高速发展，短链接服务也更加活跃。短链接在方便了人们在微博等平台上进行分享的同时也带来了诸多风险。由于短链接中不含有任何的目标网站信息，以至于人们将无从知晓该短网址究竟会带我们走向哪里。许多的网络钓鱼犯罪分子通过在用户界面张贴一个通向钓鱼网站的“短网址”，然后微博平台便自动将这恶意短网址分发给该用户所有的好友。由于这些社交网络平台用户间的信任关系，他们更容易点击这些恶意的短网址，最终引入这些钓鱼网站，被窃取个人敏感信息如身份证号、银行卡号、密码等，最终造成个人财产损失。

本文提出了一种新的短网址生成方法，通过在生成过程中分析目标网站的特征，并将其嵌入短网址中，建立起该短网址与目标网站 URL 之间的一种联系，使用户在点击短地址之前便能够知晓目的网址的部分信息，并提高对网络钓鱼的警惕性。这将有效的遏制网络钓鱼犯罪分子肆无忌惮的利用短网址欺骗用户的行为，对从源头减少恶意链接的生成、规范短网址生成服务提供有效借鉴。

## 1　短网址服务

短网址服务通常包含短地址生成过程和地址重定向两个过程。短网址服务提供商会提供一个包含脚本的界面，该脚本包含请求缩短的长地址，系统经过滥用预防、URL 过滤、垃圾预防、URL 验证等检查之后会生成一个随机字符串，并将该 ID 与目标地址以某种形式存储在数据库中，并返回与该 ID 相关的短地址。当用户访问该段地址时，系统就可以通过 301、302 或 META 转向等域名重定向技术将访问当前短网址的用户引导至目标网站。对于短地址 www.shorturl.com/8kiR21o，“www.shorturl.com”即是服务网站，8kiR21o 则为编码后的 ID。

## 2　网络钓鱼新形式

由于生成后的短网址与目标网站在内容上不存在任何联系，导致用户无法根据该短网址猜测目的 URL。因而，网络钓鱼分子便利用这一弱点实施网络诈骗。网络钓鱼分子会通过微博等平台发送一条包含指向钓鱼网站的短网址，并生成这是一条合法的地址，通过优惠、打折等相关词语吸引其他好友的点击。其他用户收到该信息以后认为这是将指向一个合法的购物或银行网页如淘宝、当当、工商银行等，实际上这个网址却转向了仿冒的网站。目前，传统的黑白名单方法无法在第一时间发现并提示用户安全风险，而机器学习来检测钓鱼网站的方法也因短链接的随机性而无法提取出有效的特征。

## 3　内容相关短链接生成算法

基于内容的短链接生成算法通过构建目的网址与生成的短网址的一种联系，使得在地址得到缩短的同时，用户能够在短链接中猜测到目的网址的部分内容，从而达到防范网络诈骗的目的。算法主要思想来源于阿拉伯语中不包含短元音，却同样可以用于交流而不会产生障碍。因此，我们尝试将去掉元音的关键信息融合进短网址中，使得用户在去掉元音的时候仍能够猜测目的网址；同时添加一标志位用于存放对目标网址预处理的结果，便于其他组织对该链接进行自动化检测。

### 3.1　生成算法

首先提取目的网址的站点名称。这里的站点名称指 URL 中排除了协议、顶级域名、路径等之后最能体现目标网站类型的部分。例如：“http：//tieba.baidu.com/index.html”这一链接，我们将首先提取域名部分“baidu.com”，然后提取站点名称“baidu”，这一过程可以使用正则表达式直接实现。然后通过去掉元音字母、数字和连接符的方法来生成一个简短的相关词。上文中站点“baidu”去掉元音“ai”、“u”后会得到相关词“bd”，并将其全部转换为小写。当站点名称没有任何辅音的时候我们将会通过其他附加规则来生成相应的词。如网易站点 163 将使用相关词“3N”，大写字母“N”表示数字类型，3 表示所含数字个数。

检查相关词的登记信息。当用户使用长地址缩短服务时，将首先检查该目的 URL 是否已被注册。如果目的地址已被注册，则直接返回相应的短链接。如果目的地址或相关词未被注册，则对该相关词进行增量计数。生成的结果将包含相关词和计数两部分，确保不同站点的相关词得以区分。例如百度公司“www.baidu.com”和美国 BD 公司 “www.bd.com”将根据请求短网址服务的先后顺序生成“bd_0 和 bd_5”。相同站点下的链接将通过对其内部 ID 进行编码区分。（下转第 116 页）



21994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

报警模块——具有告警信息显示，同时形成篡改日志，并将日志形成文件存储，为用户日后查阅提供依据，也通过电子邮件告知管理员。

软件将由使用者自主选择需要保护与监测的网页文件及文件夹，软件自动扫描并对其进行消息摘要，提供文件完整性保护，保证页面被任意改变时可及时发现被篡改痕迹。在散列算法 MD5 得出文件摘要后对重要文件进行备份，保存重要文件作为备份以供恢复时使用。软件允许使用者根据自身需要手动调整待扫描目录、文件摘要与备份存储路径，以便对文件进行即时扫描检测。

当文件遭遇删除时，管理者可以根据软件先前对文件的备份进行手动选择恢复，也可以通过软件提供的界面快速恢复；当出现恶意文件时，管理者可以根据提示进行快速删除；当正常文件被篡改内容时，可以根据备份文件的协助进行修改恢复。这一切检测都可以通过软件自行设定检测频率，当异常发生时，会有提示信息及时反馈给管理人员以便其根据实际情况进行处理。

正常更新时，只需要每次进行更新网页内容时同时更新消息摘要即可，这样既保证消息摘要值的不定期动态更新，提高系统的安全性，又使系统的更新没有冗余操作，简单易行。

此外，我们通过研究核心内嵌技术，实现了简单的过滤器，一方面在动态网站的请求发生时，截获用户的请求，对篡改进行实时处理，使被篡改的页面不被用户看到，另一方面也集中对服务端的网站目录进行监控，对发生的改变进行记录。这样做的好处是在测试时，用户永远无法看到被篡改的页面，而对于本地防护而言，用户在浏览器端还是有可能看到被篡改页面的缓存。

## 4 结束语

随着信息化建设的不断发展和提高，许多部门机构的业务数据都通过门户网站开展业务，个人或小集体通过建站来进行产品宣传推广也往往通过网站的形式来发布信息。因此保护 Web 应用系统安全变得越来越广泛。在现有网站安全防护的基础上部署网页防篡改系统，采用低成本有效的防篡改系统，不仅可以多一层保障，同时，网页防篡改系统详细的日志信息也为我们日常的安全维护工作提供帮助。一旦发生安全事件，防篡改系统还将为之后的调查取证工作提供有价值的线索和依据。

尽管网页防篡改系统能进一步提高网站的安全，但仍然存在一些缺陷，需要不断的研究实践加以改进。

参考文献：

[1]陈小兵，范渊，孙立伟. Web 渗透技术及实战案例解析[M]. 北京：电子工业出版社，2012.4.

[2]Noah，Inotify，FAM，Gamin[EB/OL].http：//www.noah.org/wiki/Inotify，_FAM，_Gamin，2010.4.

[3]孔辉. 一种网页防篡改系统的设计[D]. 北京：北京邮电大学，2011.

[4]王茂昌，黄甜，王普彪，赖培辉. 网站安全性研究[J].安阳师范学院学报，2011.

[5]赵 莉，邓 峰. 基于核心内嵌技术中安全散列函数的探讨[J].科技信息，2012-12.

[6]胡宏银，姚 峰，何成万.一种基于文件过滤驱动的 Windows 文件安全保护方案[J].计算机应用，2009-1.

作者简介：

周晗（1993—），男，汉，江西省抚州市，本科，学生，研究方向：计算机科学与技术。

（上接第 114 页）生成链接检查标识。在提供短网址服务时同时将对目的链接进行简单的安全性检查，检测其是否含有钓鱼网站特征，并在短网址中添加一个标志位，既能便于用户了解更多的安全性信息，又能方便第三方组织根据该特征位实现自动化检测。首先将检测目的地址是已经是短链接，若是则进一步判定其是否为本站点提供的短链接，为本站点提供的短链接则提取标识位，否则还原其目的地址。下一步对目的 URL 检测钓鱼网站特征。如 O 代表普通网址，I 表示链接为 IP 地址，P 指示使用非标准端口，H 表示含十六进制编码等。最终“http：//tieba.baidu.com/index.html” 将被缩短为为“www.shorturl.com/bd_0OiR21o”，而“www.bd.com” 将被缩短为为“www.shorturl.com/bd_5OeR4to”。

### 3.2 结果分析

我国的网址命名一般按照拼音、谐音、英文含义等方式将单位或组织名称嵌入域名中，如拼音形式的“baidu.com”“renren.com”，谐音形式的“sina.com”“vancle.com”，以及简写“ruc.edu.cn”等形式。这样做符合人们的阅读习惯并方便人们记忆。而我们所研究的去掉元音保留辅音的方法和人们常用的使用拼音的首字母代替该汉字有相似之处，人们可以很自然地根据缩短的相关词去推测其全文含义，而不需要过多的加以引导。例如“baidu”缩写为“bd”、“renren”缩写为“rnrn”，“vancle”缩写为“vnl”。

常见的钓鱼网址类型，通过该生成算法得到的短网址和被仿冒网站生成短网址有较好的区分度。而高明的钓鱼网址仿冒类型，如通过替换相似字母将“i”替换成“l”，使工商银行网址“www.icbc.com.cn”变成“www.lcbc.com.cn”，但是短网址中“i”为元音将被去掉，而“l”却会被保留，由此产生的短网址“www.shorturl.com/b_0OqrSC” 和 “www.shorturl.com/lb_Lqs5i” 能够被很好地区分。其他网站名称和 IP 地址类型等则更容易区分。另外，通过添加一位标志位，将更好地显示出原网址的特征，提醒用户对于相关词不易区分但暗藏风险的站点多加提防。

## 4 总结

短网址的广泛应用，给网络诈骗带来了可乘之机。该算法从短链接生成时便提供了网络钓鱼的防范机制，使得网络钓鱼分子不能随心所欲地利用短网址服务进行变形和伪装的，从源头切断短网址传播网络钓鱼链接这一方式，同时统一的短网址生成格式也有助于其他网络钓鱼探测系统对该短网址进行进一步的分析，解决了机器学习难以提取有效特征的问题。净化网络环境，打击网络犯罪离不开各方的共同努力，必须多措并举，共同营造积极健康的网络环境。

参考文献：

[1]黄华军，王耀钧，姜丽清 .网络钓鱼防御技术研究 [J].信息网络安全，2012，（04）：30-35.

[2]蔡岳峰.网易短网址服务系统的设计与实现 [D]. 北京：北京交通大学，2012.

[3]成亦陈，黄淑华.恶意短链接欺骗的防护对策研究[J].信息网络安全，2013，（074）：32-33.

[4]S. Chhabra，A. Aggarwal，F. Benevenuto，andP. Kumaraguru. Phi.sh/$oCiaL：the phishing landscape through short URLs. In CEAS ’11. ACM Request Permissions，Sept. 2011.

[5]C. Grier，K. Thomas，V. Paxson，and M. Zhang.@spam：the underground on 140 characters or less. In CCS ’10，pages 27–37，New York，NY，USA，2010.ACM.

作者简介：

薛富（1989—），男，山东，中国人民公安大学网络安全保卫学院 2011 级研究生，主要进行信息安全及反钓鱼方面研究与学习。

高一男（1990—），男，吉林，中国人民公安大学网络安全保卫学院 2013 级研究生，主要进行信息安全及反钓鱼方面研究与学习。



21994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net