基于深度学习的仿冒域名生成工具*

邹可欣[1],陈彦光[2],时金桥[1],徐 睿[3]
(1.北京邮电大学 网络空间安全学院,北京 100876;
2.OPPO 广东移动通信有限公司,广东 东莞 523860;3.华北计算机系统工程研究所,北京 100083)

摘 要:为了应对仿冒域名攻击,一种可行的思路是主动查找所有的仿冒域名,查看这些域名的使用情况,为此需要先针对受保护的域名生成所有可能的仿冒域名。在针对仿冒域名构造类型研究的基础上,设计和实现了一种仿冒域名生成工具,该工具包含基于规则的仿冒域名生成模块和基于 LSTM 神经网络的前后缀仿冒域名生成模块。经过实际的测试,证明由这种 LSTM 神经网络所生成的仿冒域名词缀和正常的词缀类似。

关键词:域名;仿冒域名;LSTM;域名生成
中图分类号:TN711 **文献标识码**:A DOI:10.16157/j.issn.0258-7998.200325

中文引用格式:邹可欣,陈彦光,时金桥,等. 基于深度学习的仿冒域名生成工具[J].电子技术应用,2020,46(7):108-112.
英文引用格式:Zou Kexin,Chen Yanguang,Shi Jinqiao,et al. Typosquatting domain name generator based on deep learning[J]. Application of Electronic Technique,2020,46(7):108-112.

Typosquatting domain name generator based on deep learning

Zou Kexin[1],Chen Yanguang[2],Shi Jinqiao[1],Xu Rui[3]
(1.School of Cyberspace Security,Beijing University of Posts and Telecommunications,Beijing 100876,China;
2.Guangdong OPPO Mobile Telecommunications Co.,Ltd.,Dongguan 523860,China;
3.National Computer System Engineering Research Institute of China,Beijing 100083,China)

Abstract: For coping with typosquatting domain name attacks, a feasible idea is to actively find all typosquatting domain names and check the usage of these domain names. To this end, all possible typosquatting domain names need to be generated for the protected domain names. Based on the research on the types of typosquatting domain names, this paper designs and implements a typosquatting domain name generation tool. The tool includes a rule-based typosquatting domain name generation module and an LSTM neural network based on domain name suffix generation module. After practical tests, it is proved that the typosquatting domain name affixes generated by this LSTM neural network are similar to normal affixes.

Key words: domain names;typosquatting;LSTM;domain name generation

0 引言

仿冒域名(Typosquatting)是指第三方恶意注册的和某些知名网站、知名企业、政府或学校等机构或部门的名称和网站相似的互联网域名。针对仿冒域名行为的防范方法主要是被动检测,很多仿冒域名往往是被攻击者实施恶意行为一段时间后才被发现,可能还有很多未被发现的仿冒域名已被注册并用于恶意行为。因此,主动去查找可能的仿冒域名,并查看这些域名是否已经被用于恶意行为,是一种很重要的仿冒域名应对和防御思路。仿冒域名生成工具可以更加便捷地生成仿冒域名,有助于检测哪些域名已被注册和使用、域名注册和使用的目的等。本文将基于主动防御的思路出发,设计和实现一种仿冒域名生成工具。

1 相关工作

1.1 仿冒域名介绍

当前研究主要针对仿冒域名的注册意图、仿冒域名的使用方式和仿冒域名的盈利策略。MOORE T[1]以 Alexa 排名中的域名为种子生成可疑仿冒域名,通过分析发现与原域名的编辑距离为 1 或 2 的域名极有可能是仿冒域名,并且绝大部分仿冒域名的主要用途是放置点击付费广告和重定向至其他网站。HALVORSON T 等人[2]对常用于成人行业的顶级域名".xxx"进行了研究,发现超过 1/3 的已注册".xxx"域名是被知名公司和个人注册,防止这些域名被用于损害形象。AGTEN P 等人[3]对由 Alexa 前 500 个域名生成的共 28 179 个域名所在网站进行了 7 个月的跟踪,发现域名抢注者会随着时间的推移改变仿冒网站的盈利策略。

* 基金项目:中央高校基本科研业务费专项资金资助项目(24820192019RC56)

1.2 仿冒域名生成方法

根据仿冒方法的不同，可以总结出以下的仿冒域名构造类型：

(1)Typosquatting 类型：这类域名指因为可能的键盘输入错误等原因而产生的仿冒域名，也是提到仿冒域名最先想到的类型。Wang Yimin 等人[4]对最常见的 Typosquatting 仿冒方式进行了研究，给出了 5 种构造 Typosquatting 域名的方式。

(2)Bitsquatting 类型：由 DINABURG A 首次提出[5]，与原域名的区别是某一位二进制位上的不同，如 mic2osoft.com 相比 microsoft.com。NIKIFORAKIS N[6]对仿冒 Alexa 前 500 个域名进行了追踪，270 天内发现了有 5 366 个 Bitsquatting 域名被注册，其中 60%的域名被用于停放、重定向、出售、广告和放置恶意软件。

(3)Soundsquatting 域名：这类域名指与目标域名读音近似的域名，利用读音的近似性迷惑用户，让用户进入仿冒网站。NIKIFORAKIS N 等人[7]对 Alexa 前一万个域名进行了 Soundsquatting 的仿冒研究，总结了仿冒域名常用的同音异义词替换。

(4)Homoglyph 域名：利用与目标域名在视觉上的近似吸引用户进入错误网站，如 arnazon.com（利用 rn 仿冒 m）、0ffice.com 等域名[8]。国际化域名(IDN)的广泛应用大大增加了 Homoglyph 域名的数量[8-9]。

(5)Abbrevsquatting 域名：由 LV P 等人提出[10]，主要是针对机构域名，攻击者可以通过从现有的缩写和全名中挖掘缩写模式，并使用非官方但可以表示官方网站意义的缩写生成一个仿冒域名，并伪造成该机构的网站以实施非法行为。

(6)Combosquatting 域名：组合仿冒域名。该类仿冒域名是指在原有域名的基础上添加一些前后缀形成的新域名，例如 yahoo-mail.com。

1.3 深度学习技术与文本生成

文本生成和预测是深度学习运用在自然语言处理中的一个重要类别。利用神经网络实现文本生成的思路是通过上下文实现对某个字或词的预测，而在传统神经网络中所有的输入(和输出)都是相互独立的，为了解决这个问题，循环神经网络 RNN 及其变体长短时记忆网络(LSTM)应运而生。RNN 可以将上一步感知到的内容和输入样例一并输入下一步的网络之中，而 LSTM 在 RNN 的基础上通过名为“门”的结构对单元状态添加或删除信息。RNN 和 LSTM 的特性使得它们被大范围地运用在文本生成和预测工作中[11-12]。

2 仿冒域名生成框架

2.1 仿冒域名生成总体架构

2.1.1 软件架构

仿冒域名生成工具的软件架构可以分为 4 个模块：输入模块、仿冒模块、数据模块、输出模块。模块结构如图 1 所示。

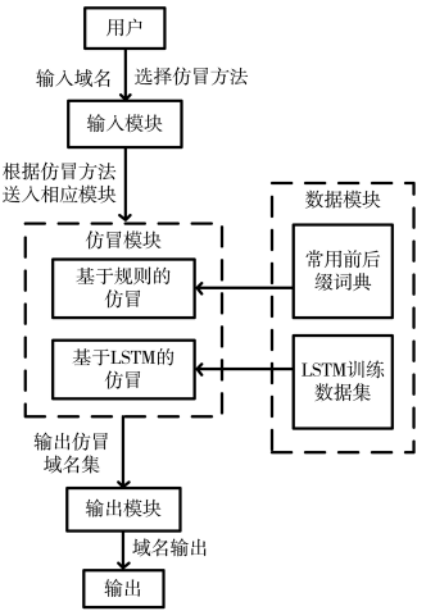

图 1 生成工具的模块示意图

(1)输入模块：该模块包括域名的输入和仿冒方法选择的输入。输入内容为种子域名和仿冒方法选择序号，支持命令行终端或文件输入两种方式。输出内容为待处理的域名，输出内容将输出至仿冒模块。

(2)仿冒模块：该模块从输入模块获取种子域名和仿冒方法选择的标志符号，利用仿冒模块中基于规则和 LSTM 神经网络的构造方法生成疑似仿冒域名，将生成的仿冒域名添加到仿冒域名集中，送入输出模块。

(3)数据模块：该部分主要存储仿冒模块需要使用的数据文件，主要包括 Combosquatting 规则的常用前后缀字典和 LSTM 神经网络的训练数据集。

(4)输出模块：该部分从仿冒模块获取生成的仿冒域名集，并根据用户的选择通过终端或者文件输出。

2.1.2 仿冒模块的架构

本文的仿冒模块设计图如图 2 所示，其中包含 8 个可用的仿冒方法。

由于二级域名是表明域名注册人类别的部分，因此各种仿冒域名构造方法主要是针对二级域名进行仿冒。仿冒模块的主要步骤如下：

(1)从输入模块读取域名，将域名拆分为 SLD(二级域名)和 TLD(顶级域名)。

(2)根据输入模块中的方法选择，将 SLD 送入各类仿冒方法中。SimpleTLD 方法，TLD 也会送入仿冒方法；AbbrevSequatting 方法，需要额外输入机构名称；Combosquatting 方法，需要选择前后缀字典。

(3)获取仿冒方法中生成的各类仿冒域名，添加到仿冒域名集中。

(4)将仿冒域名集送入输出模块。

2.2 基于规则的仿冒模块的方法

基于规则的仿冒域名方法可以分为以下 7 大类：

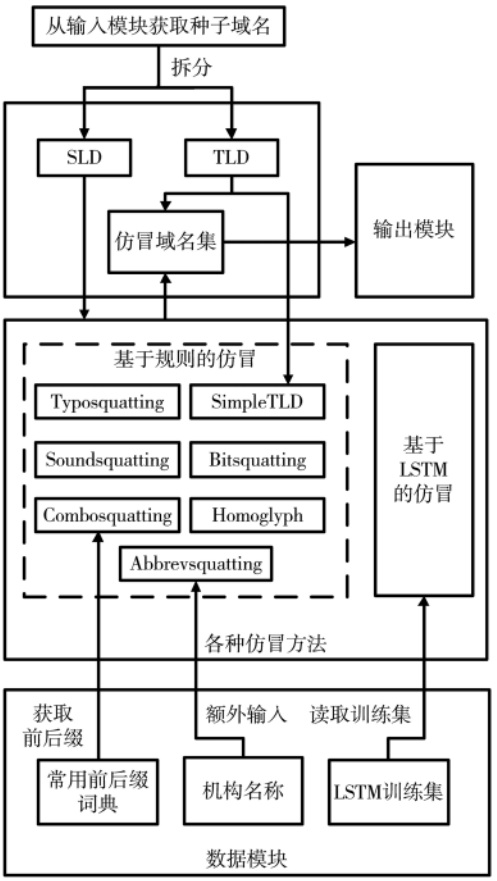

图 2 仿冒模块设计图

(1)Typosquatting：添加单字、删除单字、添加重复字、添加临近字、临近字替换、添加“-”、添加“.”和字符换位方法；

(2)Bitsquatting：对域名的某一个二进制位进行比特反转；

(3)Soundsquatting：元音替换方法；

(4)Homoglyph：使用英语字符和数字字符之中就原本就存在的相似字符进行替换，和使用非英语的字符来与相似的字符进行替换来生成仿冒域名；

(5)Abbrevsquatting：通过获取网站所属机构的名称，生成该名称的其他中英文缩写来生成仿冒域名；

(6)Combosquatting：借助词缀字典对种子域名添加前后缀来生成仿冒域名；

(7)SimpleTLD：针对顶级域名生成仿冒域名。

2.3 基于 LSTM 神经网络的仿冒模块

为了生成多样化的仿冒域名，本文设计和实现了一种基于 LSTM 神经网络的仿冒域名构造方法。

本文所用 LSTM 的训练网络如图 3 所示。

LSTM 神经网络的搭建步骤如下：

(1)读取训练文件，将文件中的域名分成单字，将所有的域名连在一起生成一个长队列，并使用“.”作为不同域名之间的分隔符；

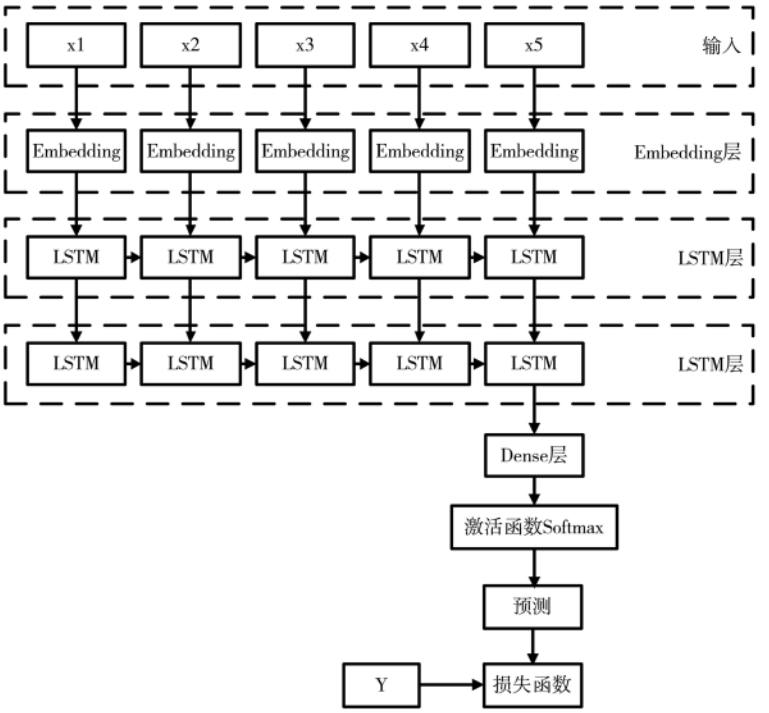

图 3 LSTM 模型训练网络结构图

(2)将分字后的结果与数字下标对应，构建 word-to-index 和 index-to-word 两个索引，其中包含了 26 个字母、10 个数字和符号“-”、“.”和“_”及空格符(结束符)、PAD 符和 Unknown 符；

(3)对分词后长字符串构造训练数据集和验证数据集，逐次读取长队列的 6 个连续字符，将前 5 个字符作为输入 x_data，将第 6 个字符作为输出 y_data，并将 x_data 处理成 tensor 模式，y_data 处理成 one-hot 模式；

(4)构建语言模型，根据前 5 个单字预测下一个单字，将其输入到双层 LSTM 网络中进行学习和验证；

(5)Dense 层，经过 Softmax 函数的处理输出概率。

3 实验设计和结果

3.1 数据集的获取和预处理

本文从以下渠道收集和获取域名，并构造了用于 LSTM 的域名数据集：

(1)Alexa 前一百万个域名，选择了其中排名靠前的 2 万个域名；

(2)恶意木马 gozi、matsnu 和 suppobox 所使用的 DGA 生成的域名集中可被认为是仿冒域名的恶意域名，共 11 983 个域名；

(3)Github 上的钓鱼域名数据集 Phishing Database 中可被认为是仿冒域名的恶意域名，共 2 935 个域名；

(4)Github 上的 APT 报告合集中总结出的仿冒域名，共 469 个域名(截止到 2019 年 4 月 10 日)。

本文中基于 LSTM 的仿冒模块所使用的数据集一共包含了 35 387 个域名，其中包括 20 000 个正常域名和 15 387 个恶意域名。为防止数据混淆，实验中全部删除

了它们的顶级域名。

3.2 训练情况及预测方法说明

本文选取80%的数据用于训练，20%的数据用于验证。神经网络的参数选择如下：2层LSTM，Embedding层输出为800维，第1层LSTM输出为800维，第2层LSTM输出为1 600维，使用Softmax函数进行归一化，优化算法为adam，损失函数为categorical crossentropy，一次载入的样本数量batch size设为256，训练次数epoch设为50。

经过50轮训练后，训练集误差loss为0.708 5，准确率acc为0.747 7；验证集最低误差val_loss为1.46332，验证集准确率val_loss为0.596 1。

通过迭代方式生成长度为200字符的字符串，使用wordninja库对字符串进行分词，排除和原域名相同的字符串和长度不小于2的分词结果，最终生成的词缀列表便是基于LSTM神经网络生成的词缀列表。

3.3 结果与评价

对于生成的各个域名，统计生成的词缀中各字符使用的频率，只统计26个英文字母和10个数字字符。

对于基于规则的仿冒域名构造方法，使用20个常见的域名SLD（google.com、baidu.com、youtube.com等）作为输入，输出的仿冒域名的字符使用频率，与这20个域名原本的字符使用频率进行比较。对于测试有以下几点说明：

(1)字符统计的是生成的整个域名(SLD+TLD)，因此包括“-”和“.”字符；

(2)Homoglyph方法只统计其中使用英文和数字字符替换原字符的情况；

(3)AbbrevSquatting难以通过热门域名生成仿冒域名，不计入统计。

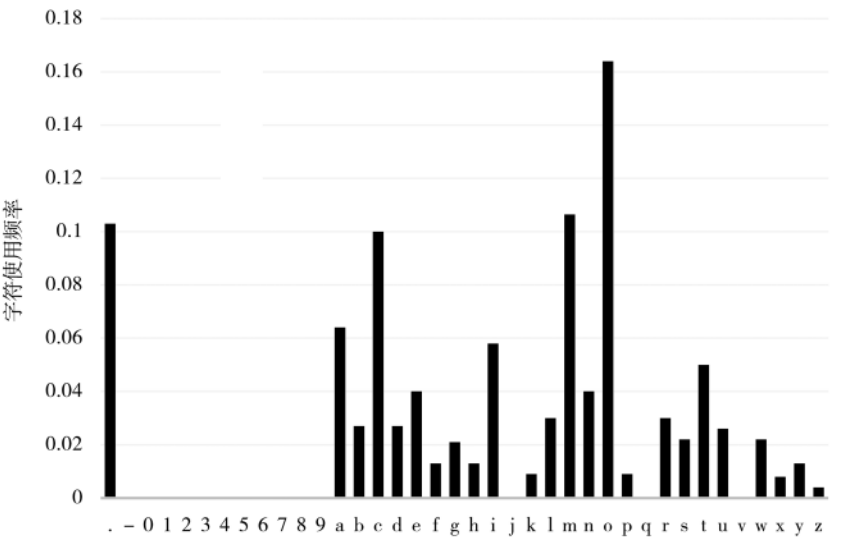

图4 基于规则的仿冒模块使用的20个测试域名的字符频率统计

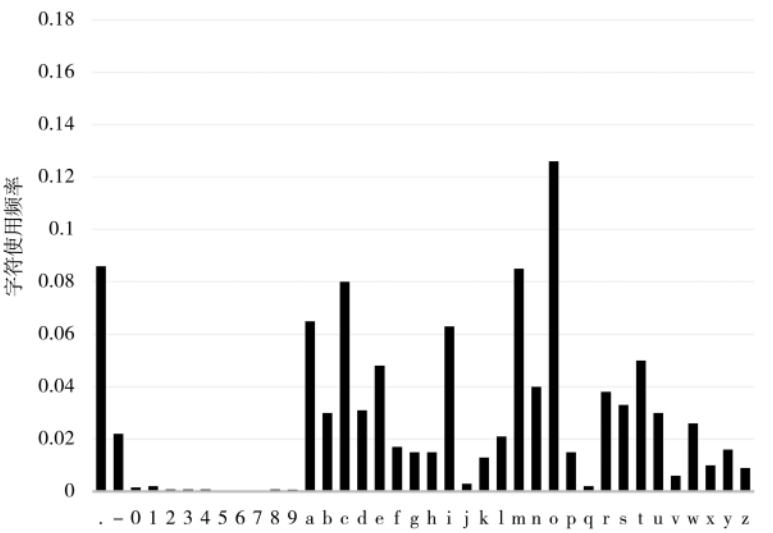

图5 基于规则的仿冒模块生成的仿冒域名的字符频率统计

图4和图5分别为原始域名和其生成的23 624个可测试的仿冒域名的字符频率统计。可以看出，基于规则的仿冒模块生成的域名和原域名在字符使用频率上是相似的，也因此可以说明基于规则的仿冒模块生成的仿冒域名与原域名具有很强的相似性。

对于基于LSTM神经网络的仿冒域名生成方法，由于LSTM神经网络最重要的是生成仿冒域名词缀，而词缀使用的字符是难以预测的，因此本文通过统计词缀字符使用频率的方式以对生成效果进行评价。

本文选择了20个常见的域名作为统计输入，生成了783个词缀，将这些词缀的字符统计结果与正常域名集Alexa Top 1M的字符统计结果(如图6所示)进行比较，输出域名的统计结果如图7所示。可以看出，除了数字使用较少以外，使用LSTM生成的仿冒域名词缀在英文字母和数字上的使用频率与Alexa中的正常域名相似，说明生成的词缀可以被认为和正常的词语和词缀相似。因此可以认为基于LSTM的仿冒域名构造模块生成了和正常域名相似的Combosquatting类仿冒域名。

4 结论

本文设计和实现了一种基于双层LSTM神经网络的一种仿冒域名生成工具，使用35 387个域名作为数据集对神经网络进行了训练，训练后的模型可以比较好地生成仿冒域名词缀。通过对生成的词缀的字符频率统计，得出结论是生成的仿冒词缀与正常的域名集的字符频率相似，可以认为生成的词缀是有意义的词缀。相比于传统的只能使用特定字典进行域名生成的算法而言，基于LSTM神经网络的仿冒域名生成技术可以生成更加多样的仿冒域名，对于仿冒域名的研究意义重大。

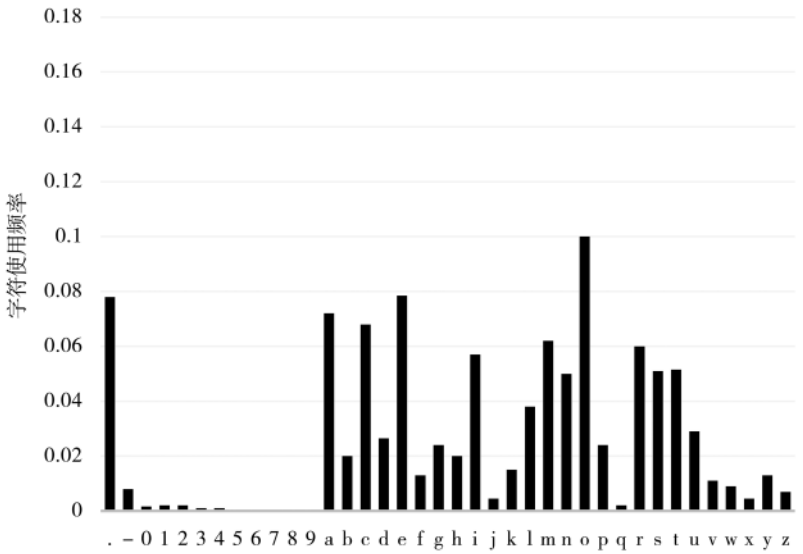

图 6 正常域名集(Alexa Top 1M)的字符频率统计结果

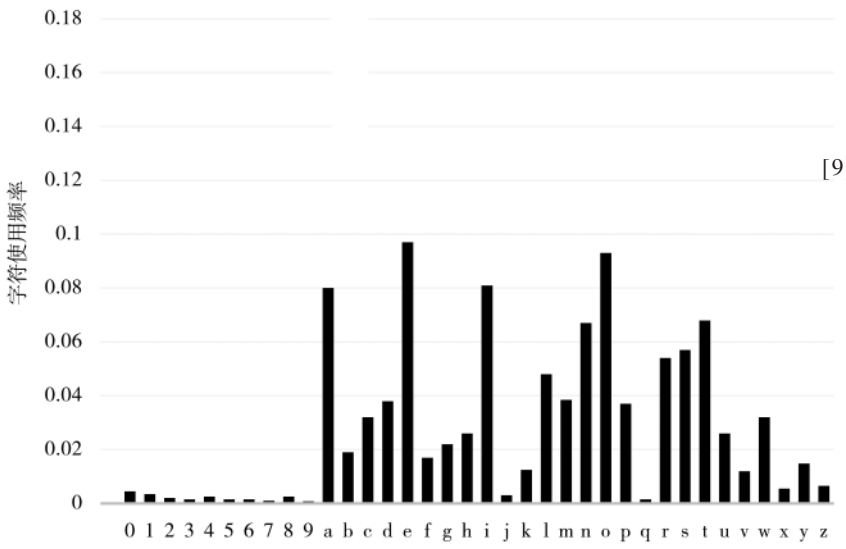

图 7 基于 LSTM 神经网络生成的长字符串的字符频率统计结果

参考文献

[1] MOORE T, EDELMAN B. Measuring the perpetrators and funders of typosquatting[C]. In. Radu Sion. Financial Cryptography and Data Security, 14th International Conference, FC 2010, Tenerife, Canary Islands, January 25-28, 2010, Revised Selected Papers. Tenerife, Canary Islands, 2010: 175-191.

[2] HALVORSON T, LEVCHENKO K, SAVAGE S, et al. XXXtortion?: inferring registration intent in the.XXX TLD[C]. ACM Proceedings of the 23rd International Conference on World Wide Web. Seoul, Korea, 2014: 901-912.

[3] AGTEN P, JOOSEN W, PIESSENS F, et al. Seven months' worth of mistakes: A longitudinal study of typosquatting abuse[C]. In. Internet Society. Network and Distributed System Security Symposium. San Diego, USA, 2015: 1-13.

[4] Wang Yimin, BECK D, WANG J, et al. Strider typo-patrol: discovery and analysis of systematic typo-squatting[C]. USENIX Association. SRUTI'06: 2nd Workshop on Steps to Reducing Unwanted Traffic on the Internet. Berkeley, USA, 2016: 31-36.

[5] DINABURG A. BitsQuatting: DNS hijacking without exploitation[EB/OL]. (2011-07-xx)[2019-05-17]. http://dinaburg.org/Bitsquatting.html/.

[6] NIKIFORAKIS N, ACKER S V, MEERT W, et al. Bitsquatting: exploiting bit-flips for fun, or profit?[C]. Proceedings of the 22nd International Conference on World Wide Web, 2013, 5: 989-998.

[7] NIKIFORAKIS N, BALDUZZI M, DESMET L, et al. Soundsquatting: uncovering the use of homophones in domain squatting[C]. Springer. International Conference on Information Security. Hong Kong, China, 2014: 291-308.

[8] HOLGERS T, WATSON D E, GRIBBLE S D. Cutting through the confusion: a measurement study of homograph attacks[C]. USENIX Association. Usenix Technical Conference. DBLP. Boston, USA, 2008: 261-266.

[9] HELOU J A, TILLEY S. Multilingual web sites: internationalized domain name homograph attacks[C]. IEEE International Symposium on Web Systems Evolution. IEEE, 2010: 89-92.

[10] LV P, YA J, LIU T, et al. You have more abbreviations than you know: a study of abbrevsquatting abuse[J]. Computational Science-ICCS 2018, 2018, 6: 221-233.

[11] 牛伯浩. 循环神经网络实现文本智能预测[J]. 智能城市, 2018, 4(10): 21-23.

[12] NIKOLAEV R. Generating drake rap lyrics using language models and LSTMs[EB/OL]. (2018-03-09)[2019-05-17]. https://towardsdatascience.com/generating-drake-rap-lyrics-using-language-models-and-lstms-8725d71b1b12/.

(收稿日期: 2020-04-22)

作者简介:

邹可欣(1997-),女,本科,主要研究方向:网络与信息安全。

陈彦光(1997-),男,本科,主要研究方向:网络安全、移动安全。

时金桥(1978-),通信作者,男,博士,教授,主要研究方向:匿名与隐私保护、高级威胁检测溯源、大数据智能分析,E-mail: shijinqiao@bupt.edu.cn。

版权声明