基于改进马尔可夫链的高效域名生成算法

钱志业，李雪，李锁钢
（赛尔网络有限公司网络运行部，北京 100084）

摘　要：首先，结合多种策略进行子域名的初步筛选，包括字典枚举、搜索引擎挖掘和网站信息抓取。接着，使用改进的马尔可夫模型算法对筛选数据进行分析，生成并添加新的子域名到结果集中。然后，检查并验证数据的真实性，若数据未达标准，则重复分析过程。最终，形成一个经过严格验证的数据集。该算法显著提升了子域名发现的效率及覆盖范围，有效弥补了传统方法的不足。

关键词：域名生成算法；子域名挖掘；DNS 解析；网络安全
中图分类号：TP311
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024253

Domain name generation algorithm based on improved Markov chain

QIAN Zhiye, LI Xue, LI Suogang
Network Operation Department, CERNET Corporation, Beijing 100084, China

Abstract: First, a combination of multiple strategies were used to conduct preliminary screening of subdomain names, including dictionary enumeration, search engine mining, and website information crawling. Then, the improved Markov model algorithm was used to analyze the filtered data, and new subdomain names were generated and added to the result set. Thereafter, the data was checked and verified for authenticity. If the data did not meet the criteria, the analysis process was repeated. Finally, a rigorously validated data set was formed. The proposed algorithm significantly improves the efficiency and coverage of subdomain discovery, effectively making up for the shortcomings of traditional methods.

Keywords: domain name generation algorithm, subdomain mining, DNS resolution, network security

0　引言

子域名作为主域名下的一个分支，通过在主域名前添加独特的前缀来构建，以此区分不同的网站部分或服务类型。例如，在主域名“example.com”之上能够创建用于特定目的的独立域名，类似“jd.example.com”“baidu.example.com”。在这些案例中，“jd”和“baidu”作为前缀，与主域名通过点号连接，构成子域名。每个子域名都能拥有独立的网络地址、服务器配置和内容。这种方法便于网站管理员更有效地对网站进行管理和分类，以适应不同的业务需求和服务供应。文献[1]揭示了全球信息安全面临的严峻挑战。随着网络技术的快速发展和网络安全威胁的不断演变，域名作为互联网基础设施的核心组成部分，其安全管理和监控变得尤为重要。子域名作为域名结构中的关键元素，不仅承载着重要的网络服务，也是网络安全防护的前沿阵地。传统的子域名发现技术虽然能满足基本的安全检测需求，但在处理大规模数据时，常常面临效率低下和覆盖不全的问题。

本文致力于通过改进马尔可夫链模型，开发一种新的域名生成算法，以提高子域名发现的效率和准确性。该算法在传统方法的基础上引入了复杂网络理论中的概率转移矩阵，使得子域名生成不再单

收稿日期：2024–10–24

纯依赖于预定义的字典或简单的枚举，而是根据实际的域名使用模式动态预测和生成。这种方法的核心在于，它可以根据已知的域名结构学习潜在的、未被发现的子域名，从而为网络安全防护提供更全面的数据支持。

进一步地，本文将探讨改进马尔可夫链模型在实际网络环境中的应用效果，通过与传统技术的对比分析，验证其在提高子域名探测速度和准确性方面的优势。本文的研究不仅有助于推动网络安全技术的发展，也对保护网络基础设施的完整性和稳定性具有重要意义。

1　域名信息挖掘常用方法

子域名挖掘技术的宗旨在于探寻和搜集隶属于某一主域名下的全部子域名。通过这种挖掘过程，可以拓宽所掌握的资产范围，收集到更多关于该域名的数据资料，这些信息随后可被应用于安全漏洞检测、情报搜集等不同场景。子域名的信息收集有被动测量和主动探测 2 种方法。被动测量主要指流量测量与分析；主动探测则包括域名系统（DNS，domain name system）信息探测、网站信息探测、爬虫探测等[2]。下面列举一些国内外常见的主动探测方法。

1.1　DNS 字典枚举

DNS 字典枚举是一种网络安全技术，旨在通过预定义的子域名列表对特定的主域进行系统性探测，以识别潜在的活跃子域名。这种方法利用了组织在命名子域时可能遵循的常规和标准化命名模式。实施者使用专门的工具（如 dnsenum、dnsmap 等），根据一组精心挑选的候选子域名向 DNS 服务器[3]发送查询请求。根据 DNS 响应，这些工具可以有效地揭示出哪些子域名是解析成功的，从而为进一步的安全评估或网络渗透提供基础数据。这种技术在执行网络空间映射和安全漏洞评估时尤其重要，有助于揭示企业或组织的网络边界和潜在的攻击面。

1.2　搜索引擎挖掘

搜索引擎挖掘是利用搜索引擎的数据和功能来发现和提取特定信息的技术。该方法尤其用于识别和收集关于目标域名和其网络资源的详细信息。通过在各大搜索引擎中输入特定查询[4]，分析人员可以发现与目标域相关的各种资料，包括子域名、关联网站、公开的服务和配置信息等。这不仅包括从搜索结果中提取链接和文本信息，还涉及解析这些信息中包含的网络特征和业务关系。搜索引擎挖掘的优势在于其能够利用庞大的搜索引擎数据库，快速获取大量分散的数据点，有助于构建目标网络的全面视图。

1.3　证书透明度日志

在进行子域名探测的研究中，证书透明度日志（CT, certificate transparency）提供了一种有效的技术手段。证书透明度是一种公开的监控系统，旨在改善 SSL/TLS 证书的安全性和信任度。通过该系统，每个颁发的 SSL/TLS 证书都会被记录在公开、不可篡改的日志中。这些日志为研究人员提供了获取域名下所有已颁发证书的子域名的机会，包括那些可能没有直接出现在 DNS 记录中的子域名。

2　域名信息挖掘框架

本文域名信息挖掘的构建流程主要分为 4 个步骤，流程构架[5]如图 1 所示。

1) 初始域名数据收集。依据字典枚举、搜索引擎挖掘、证书透明度日志等基本原理，进行初始域名数据收集。

2) 数据库的构建。根据探测结果采集的数据，利用各种类型数据的特性，将多种不同类型数据转化成统一数据集，归纳整理入数据库。

3) 基于马尔可夫模型和初始域名数据进行建模，结合既定参数生成待测域名集合。

4) 域名验证。对待测域名信息高效并发探测，收集并统计结果，构建最终域名知识图谱。

3　基于马尔可夫模型的测试域名生成方法

3.1　马尔可夫链介绍

马尔可夫链是一种数学模型，它描述了一个系统随时间演化而在一系列可能状态之间转移的过程。这些状态的特点是，未来状态的概率分布仅依赖于当前状态，而与之前的历史状态无关，这种性质称为马尔可夫性质或无记忆性[6]。

1) 基本原理

设 $X_1,X_2,X_3,\cdots$ 是一个随机变量序列，这个序列中的每一个变量 X_n 都取自一个有限或可数的状态空间 S，且满足马尔可夫性质。对于任何自然数 n 和任何状态 $i,j\in S$，马尔可夫性质可以在数学上表述为

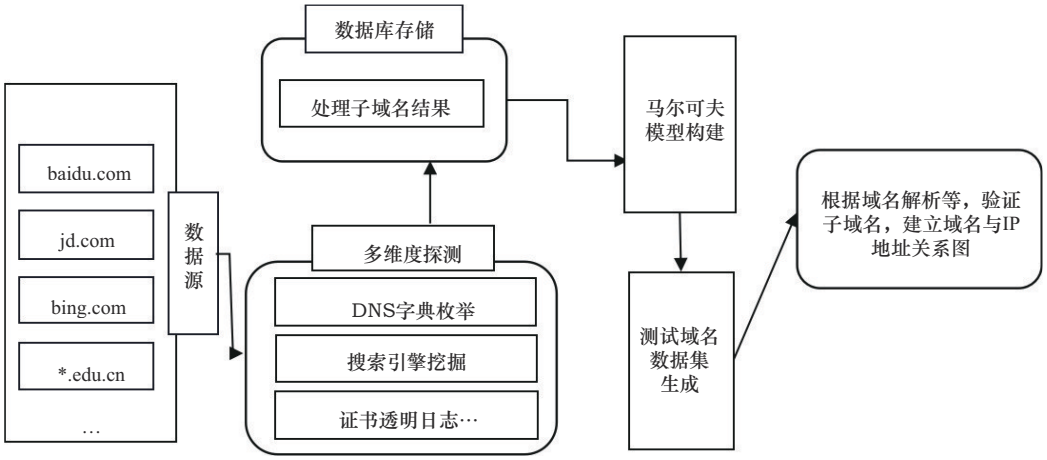

图1　域名信息挖掘流程构架

$$P(X_n+1=j|X_1=i_1,X_2=i_2,\cdots,X_n=i)=P(X_n+1=j|X_n=i) \tag{1}$$

式(1)表明，状态j在时间n+1的概率仅依赖于时间n的状态i，而与之前的状态无关。

2) 转移概率矩阵

在马尔可夫链中，从状态i转移到状态j的概率称为转移概率，通常用矩阵$\boldsymbol{P}$来表示，其中，矩阵的元素p_{ij}表示从状态i转移到状态j的概率，即

$$p_{ij}=P(X_n+1=j|X_n=i) \tag{2}$$

对于所有$i,j\in S$，转移概率分别满足非负性和规范性。

3) 状态和转移概率的计算

① 确定状态空间：根据问题的具体情境定义状态空间。

② 估计转移概率：通常通过历史数据来估计转移概率。如果状态i出现N_i次，并且在这些情况中有N_{ij}次转移到状态j，则p_{ij}可估计为

$$p_{ij}=\frac{N_{ij}}{N_i} \tag{3}$$

③ 构造转移矩阵：将所有估计的p_{ij}值填充到转移矩阵$\boldsymbol{P}$中。

4) 长期行为

马尔可夫链的长期行为通常通过计算其稳态分布来分析，稳态分布$\boldsymbol{\pi}$是转移矩阵$\boldsymbol{P}$的一个左侧特征向量，满足

$$\boldsymbol{\pi P}=\pi \tag{4}$$

$$\sum_{i\in S}\pi_i=1 \tag{5}$$

其中，π_i表示长期存在于状态i的概率。

3.2　域名生成模型

子域名探测是网络安全和信息收集领域的一项重要活动，用于映射组织的网络结构并发现潜在的攻击面。传统的子域名探测技术，如穷举法或基于字典的搜索，虽广泛应用但效率低下且容易受限于预定义的数据集。基于马尔可夫链的模型通过从已知域名的结构学习概率分布，能更智能地预测可能存在的子域名，从而提高探测的效率和覆盖范围。

在本文模型中，每个字符（包括字母、数字及符号）都被定义为马尔可夫链的一个状态。模型的核心是构建一个转移概率矩阵，该矩阵基于从大量已知子域名中学习到的字符之间的转移概率。这种方法利用了子域名中字符的统计规律，使得生成的子域名与实际使用的域名在结构上更加接近。

3.2.1　数据准备与预处理

首先，需要收集一个大规模的、多样化的已知子域名数据集。这些数据可以从公开的DNS记录、历史的网络安全报告或其他可用的网络资源中获得。收集后的数据需要进行预处理，通常包括清洗非法字符、去除顶级域名部分以及统一字符的格式（例如，转换所有字符为小写）以减少状态空间的复杂性。

3.2.2　构建转移概率矩阵

每个独立的字符被视为一个状态。通过分析字符序列（即域名中的字符）的转移频次，可以构建一个状态转移矩阵。具体地，如果某个字符a后面跟着字符b的次数为N_{ab}，而字符a总共出现的次数为N_a，那么从字符a转移到字符b的概率p_{ab}为

$$p_{ab}=\frac{N_{ab}}{N_a} \tag{6}$$

转移概率矩阵$\boldsymbol{P}$的每一个元素p_{ij}表示从状态i转移到状态j的概率。这个矩阵是模型的核心，用

于在生成过程中决定下一个最可能的字符。

3.2.3　子域名生成算法

利用训练好的马尔可夫链，子域名的生成从一组起始字符（可以是频繁出现的字符或随机选择的字符）开始。根据当前字符的状态，使用累积概率分布来随机选择下一个字符，直到生成完整的子域名。这个过程可以通过设置长度限制、字符种类或达到预定义的停止条件（如点号或特定后缀）来终止。

3.2.4　模型评估与应用

生成的子域名列表需要通过实际的DNS查询来验证其存在性。此外，模型的有效性可以通过比较其预测结果与现实中已知的子域名数据进行评估。有效的评估指标包括准确率、召回率以及生成效率等。

3.3　模型优化

为了进一步增强基于马尔可夫链的子域名探测模型的预测能力，本文引入了子域名的长度分布、字符频率和词缀信息等新特征。这些特征的引入旨在捕捉子域名生成过程中的更多细节信息，从而提高模型的准确性和实用性。

3.3.1　子域名长度分布

子域名的长度可以提供关于域名构造复杂性的重要线索。例如，较长的子域名可能指示了更具体的功能或层级较深的网站结构。通过分析已知子域名的长度分布，可以调整模型以偏好生成实际环境中更常见的长度范围的子域名。

3.3.2　字符频率分析

每个字符在子域名中出现的频率是预测其出现在新子域名中位置的关键因素。通过统计大量子域名中各个字符的出现频率，可以对马尔可夫链的转移概率矩阵进行调整，使其更加符合真实世界数据的分布。

3.3.3　词缀和语义分析

子域名常常包含具有特定意义的词缀，如-test、-dev等，这些词缀有助于反映子域名的用途或属性。通过识别和分析这些常见词缀，模型可以更精确地预测符合逻辑和实用性的子域名结构。此外，引入语义分析能够帮助识别包含特定意义的完整词汇，如mail、shop等，这对于生成更为相关和实用的子域名尤其重要。

这些新引入的特征被整合到马尔可夫模型中，通过修改转移概率矩阵和状态定义来实现。具体地，模型训练过程中将考虑这些特征对于状态转移概率的影响，以及它们如何影响最终子域名的生成。这一整合不仅提升了模型的预测精度，也增强了其对于实际应用场景的适应性。

4　实验及结果分析

4.1　实验数据来源及处理

本文实验的数据取自QS世界大学排名中TOP 500名的非中国大陆高校，数据由官方平台提供。在数据准备阶段，本文专注于收集这些高校的官方网站主域名，并将域名进行字典枚举、爬虫网站、网站信息探测等多维度处理合并，形成了较为全面的数据集。以下为数据处理的部分成果，占比为某大学子域名个数占总子域名个数的比例，具体如表1所示。

表1　　数据处理结果

大学名称	子域名个数/个	占比
MIT	135	0.12%
University of Cambridge	609	0.57%
University of Oxford	372	0.34%
...	...	
Umea University	141	0.13%
...	...	
总计	106 787	100%

由表1可以看出，TOP 500高校子域名总数为106 787个，其中子域名数量超500个的有30余所高校。

依据域名的命名要求，只允许以数字、字母开头，因此字符包括两类：数字、字母。由于域名中不区分大小写，可认为域名字符集共36个字符[7]，具体如表2所示。

表2　　域名字符集合

分类	字符集
数字	0~9
字母	a~z、A~Z

域名由特定的字符集构成，通过对众多域名中字符出现的频次进行统计分析，可以探索出域名内各个字符的分布模式。除此之外，针对域名的级别与长度也需要统计分析。为方便理解研究内容，

表3~表5及图2以University of Cambridge的域名分析为例展开描述。

表3　　域名的级别

域名级数	数量/个
4层	16
5层	456
6层	131
7层以上	6
总计	609

由表3可以看出，域名的级数主要以5层、6层域名为主，共占96%以上。

表4　　数据标签起始字符分布

字符	概率	字符	概率	字符	概率	字符	概率
0	0.0000%	9	0.0000%	i	3.7829%	r	4.4408%
1	0.0000%	a	6.2500%	j	1.8092%	s	10.3618%
2	0.1645%	b	3.6184%	k	0.8224%	t	3.2895%
3	0.0000%	c	12.6645%	l	3.2895%	u	0.3289%
4	0.0000%	d	4.1118%	m	7.0724%	v	1.6447%
5	0.0000%	e	4.6053%	n	3.1250%	w	9.8684%
6	0.0000%	f	2.1382%	o	2.4671%	x	0.3289%
7	0.0000%	g	2.1382%	p	7.8947%	y	0.0000%
8	0.0000%	h	3.1250%	q	0.3289%	z	0.3289%

由表4可以看出，c、w、s、m、a、p作为首字符出现的概率最大，均超过5%。字符出现的概率可大致体现出命名该域名人的喜好，通过概率大小实现域名的生成。

由表5可以看出，域名长度基本处于17到27之间。因此，可以将域名长度限制到较小的范围内，且以不同的概率取不同的长度，以此缩小探测空间。

4.2　实验结果

使用基于马尔可夫链的域名生成算法，产生大量不同长度及数量级别的域名，最终通过DNS解析验证子域名真实性[8]，统计生成域名数与验证成功的域名数占比，计算准确率。通过该算法，University of Cambridge又重新挖掘出430个子域名，实验结果部分展示如图2所示。

表5　　域名的长度

域名长度	数量/个	占比
12	1	0.2%
13	4	0.7%
14	2	0.3%
15	3	0.5%
16	9	1.5%
17	37	6.1%
18	47	7.7%
19	28	4.6%
20	48	7.9%
21	65	10.7%
22	64	10.5%
23	45	7.4%
24	47	7.7%
25	38	6.3%
26	28	4.6%
27	33	5.4%
28	25	4.1%
29	18	3.0%
30	13	2.1%
31	8	1.3%
32	10	1.6%
33	7	1.2%
34	3	0.5%
35	6	1.0%
36	6	1.0%
37	2	0.3%
38	5	0.8%
39	1	0.2%
40	1	0.2%
41	1	0.2%
43	1	0.2%
45	1	0.2%
49	1	0.2%

表6的首字符概率为算法挖掘出的430个子域名的首字符概率，实验结果与样本概率相差不大，由此证明本文所提算法的正确率。

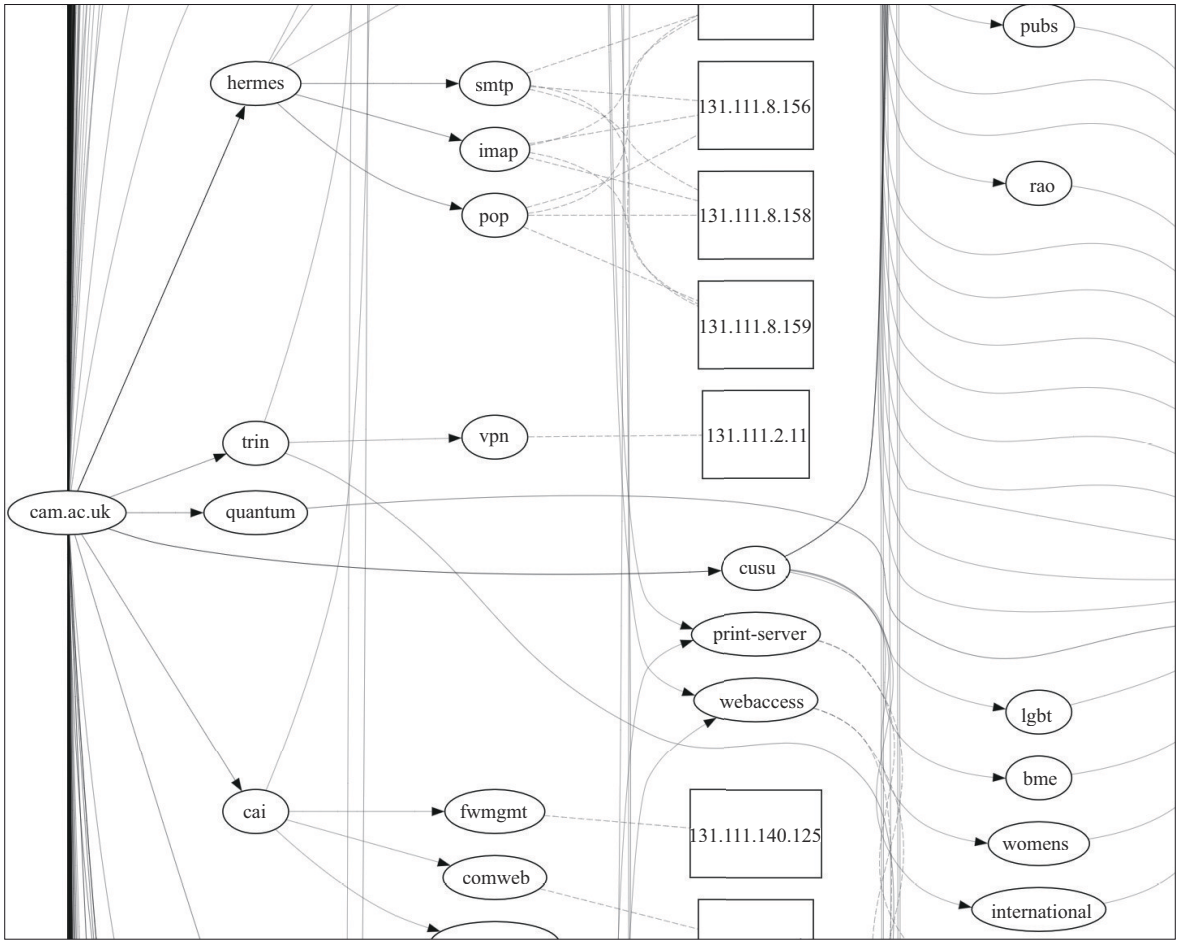

图2　University of Cambridge子域名信息结果部分展示

表6　　数据标签起始字符分布

字符	概率	字符	概率	字符	概率	字符	概率
0	0.000 0%	9	0.000 0%	i	4.418 6%	r	2.325 6%
1	0.232 6%	a	6.511 6%	j	1.162 8%	s	7.674 4%
2	0.232 6%	b	3.023 3%	k	1.395 3%	t	5.814 0%
3	0.000 0%	c	12.558 1%	l	5.348 8%	u	2.325 6%
4	0.000 0%	d	6.744 2%	m	6.511 6%	v	2.325 6%
5	0.000 0%	e	3.488 4%	n	3.255 8%	w	6.744 2%
6	0.000 0%	f	3.023 3%	o	1.627 9%	x	0.000 0%
7	0.000 0%	g	4.651 2%	p	2.790 7%	y	0.000 0%
8	0.000 0%	h	3.720 9%	q	1.627 9%	z	0.465 1%

5　结束语

本文提出了一种基于改进马尔可夫链的域名生成算法，该算法显著提升了子域名的发现效率和精确性。通过实证研究，本文证明了算法在动态生成与现实世界数据相符的子域名方面的有效性。此外，算法的实现不依赖传统的枚举或字典方法，而是基于从现有域名数据中学习得到的概率模型，这一点体现了其创新性和先进性。本文的成功展示了利用统计学习理论改进信息安全技术的潜力。通过引入马尔可夫链中的状态转移概率矩阵，不仅优化了域名生成的过程，还为网络安全提供了一个更为精准的工具，能够识别和预测潜在的网络安全威胁[9]。实验结果支持了模型的实用性，展示了其在提高探测速度和覆盖范围上的优势。

参考文献：

[1] 刘阳, 刘晶 . 2023 年全球网络安全态势综述[J]. 保密科学技术, 2024 (2): 37-43.
LIU Y, LIU J. Overview of global cybersecurity situation in 2023[J]. Secrecy Science and Technology, 2024(2): 37-43.

[2] 胡荣贵, 许成喜, 汪永益, 等. 马尔科夫链在域名信息探测中的应用[J]. 计算机应用与软件, 2015, 32(6): 152-155.
HU R G, XU C X, WANG Y Y, et al. Application of Markov chain in domain Name system information detection[J]. Computer Applications and Software, 2015, 32(6): 152-155.

[3] 刘宇, 岳明, 汤锦淮, 等. 互联网域名解析系统安全分析[J]. 信息网络安全, 2010, 10(12): 14-16.
LIU Y, YUE M, TANG J H, et al. Security analysis of Internet domain names system[J]. Netinfo Security, 2010, 10(12): 14-16.

[4] 刘文峰, 方滨兴, 张文佳. 针对新顶级域名的Web浏览器行为测试与分析[J]. 智能计算机与应用, 2020, 10(2): 333-338.
LIU W F, FANG B X, ZHANG W J. The behavior test and analysis of new top-level domain names in Web browsers[J]. Intelligent Computer and Applications, 2020, 10(2): 333-338.

[5] 胡昌秀, 张仰森, 刘洋, 等. 面向域名解析系统的知识图谱构建与应用方法[J]. 科学技术与工程, 2023, 23(23): 9979-9990.
HU C X, ZHANG Y S, LIU Y, et al. Knowledge graph construction and application method for domain Name system[J]. Science Technology and Engineering, 2023, 23(23): 9979-9990.

[6] 程亚楠, 李正民, 迟乐军, 等. 基于改进马尔可夫链的域名获取方法研究[J]. 高技术通讯, 2016, 26(10): 857-866.

[7] 张宗国. 马尔可夫链预测方法及其应用研究[D]. 南京: 河海大学, 2005.
ZHANG Z G. Research on Markov chain prediction method and its application[D]. Nanjing: Hohai University, 2005.

[8] ZENGX M, CHENX S, SHAOG L, et al. DTA-HOC: online HTTPS traffic service identification using DNS in large-scale networks[J]. Tsinghua Science and Technology, 2020, 25(2): 239-254.

[9] CASINO F, LYKOUSAS N, HOMOLIAK I, et al. Intercepting hail hydra: real-time detection of algorithmically generated domains[J]. Journal of Network and Computer Applications, 2021, 190: 103135.

[作者简介]

钱志业（1997– ），男，河南周口人，赛尔网络有限公司网络安全工程师，主要研究方向为主干网运行管理和网络安全。

李雪（2000– ），女，山西临汾人，赛尔网络有限公司网络安全工程师，主要研究方向为主干网网络安全。

李锁钢（1978– ），男，内蒙古包头人，博士，赛尔网络有限公司高级工程师，主要研究方向为主干网运行管理和网络安全。