doi:10.3772/j.issn.1002-0470.2016.10－11.005

基于改进马尔可夫链的域名获取方法研究①

程亚楠②* 李正民** 迟乐军③* 许海燕* 陆柯羽*

(*哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

(**国家计算机网络应急技术处理协调中心 北京 100029)

摘 要 为解决目前域名获取方法效率低且获取域名数量较少的问题,对前期采集的大量域名进行了统计分析,以发现域名字符的组成规则以及分布特征,并根据这些特征,设计了基于马尔可夫链的域名模型,提出了一种基于改进马尔可夫链的域名生成算法。对生成的域名进行了 WHOIS 查询验证,以确认域名是否存在。通过大量实验结果证实,该算法具有较高的域名生成准确率,且与其它域名获取方法相比,该方法具有生成域名速率快、域名数量多和顶级域名覆盖广等优点。

关键词 域名,马尔可夫链,字符频率,WHOIS

0 引 言

域名是 Internet 上某一台计算机或计算机组的名称,由一串用点分隔的名字组成,用于在数据传输时标识计算机的电子方位[1]。域名要在域名注册机构注册申请通过后才可以使用,域名注册机构则将用户注册信息存入数据库,即域名 WHOIS 信息,域名注册机构一般会提供域名 WHOIS 服务器,并开放 43 端口,提供域名 WHOIS 信息查询服务,若无,则提供 Web 页面查询[2]。域名作为构成当前整个互联网的基础要素之一,用户每天都在与不同的域名进行交互,随着互联网技术的迅猛发展,互联网上的域名也越来越多。根据网络基础服务商 Verisign 的 2015 年第三季度域名趋势报告,目前全球顶级域名注册数量已经接近 3 亿,并且在持续增长中,同时根据互联网号码分配局(Internet Assigned Numbers Authority,IANA)最新统计,各类顶级域名注册数量已达到 1128 个[3]。

围绕域名的相关研究已成为热门,但目前获得大量域名的方法和途径较少,除与个别第三方域名注册机构购买外,研究人员则主要采用以下方法获取域名:(1)使用 Alexa[4] 域名网站所公布的全球排名前 100 万域名,该方法的缺点是获取的域名数量与全球域名数量相比较少,且顶级域名覆盖率较低;(2)利用集中区域数据服务(Centralized Zone Data Service,CZDS)[5] 为研究者提供的不同顶级域名的 zone file 文件 API 接口,可以获得 zone 文件并提取域名,这种方法的局限也在于域名数量较少,且未开放部分通用的顶级域名的获取权限,例如 com、net 等顶级域名,其获得的域名皆为网络中开始使用的域名,未使用的域名无法获取;(3)设计实现域名获取爬虫是最常见的方法之一,许笑等人通过设计实现广域网的分布式 Web 爬虫[6],通过该方法获得网络中网址,然后提取出域名,但是域名爬虫覆盖范围较小且与爬虫入口网址有很大关联;(4)基于域名字典的获取方法,也可以获取部分域名,但是该方法受限于字典,只能探测字典内的域名,获得域名数量

① 国家科技支撑计划(2012BAH45B01),国家自然科学基金(61100189, 61370215, 61370211)和国家信息安全计划(2014A085, 2015A072)资助项目。

② 男,1989 年生,硕士生;研究方向:域名体系安全; E-mail: mrcheng0910@gmail.com

③ 通讯作者,E-mail: qdclj@163.com

(收稿日期:2016-05-16)

少,另外其一般使用域名系统(DNS)解析来验证域名是否存在,准确率低。可以看出,以上方法都存在获取域名数量少、顶级域名覆盖率低等缺陷,这些方法所获取的域名都局限于网络开始使用的域名,而未在网络中的域名,则无法获取。针对这种情况,本文通过分析大量域名的组成分布特征,设计了基于改进的马尔可夫链的域名生成算法,该算法根据域名组成规则来生成新的域名,最后通过域名 WHOIS 信息进行验证是否存在,而不只是通过 DNS 解析验证,且该方法能够覆盖大部分顶级域名,并且充分利用已有的域名列表来生成新的域名。

1 域名组成分布特征分析

根据域名组成约定,域名具有层级性,每一层由数字、阿拉伯字母(不区分大小写)和特殊字符“-”组成,共 37 个字符,且首个字符和末尾字符不能为“-”,各层级则由小数点“.”连接,最多分为 5 级。最右边的第一层为顶级域名(top-level domain, TLD),第二级为二级域名,第三级为三级域名,依此类推,从右向左排列。例如,在域名 google. com 中,则 com 为顶级域名,google 为二级域名。在本文中所生成的域名主要是指域名中的最后一层,即主域名,例如 google,而顶级域名是指除主域名之外的部分,例如 com、cn 等。

另外,顶级域名分为 6 类,(1)通用顶级域名(generic top-level domains, gTLD),例 com, net 等;(2)国家级顶级域名(country-code top-level domains, ccTLD),例 cn, us 等;(3)基础设施类顶级域名(infrastructure top-level domains, ARPA),例 arpa 等;(4)实验性顶级域名(test top-level domains),例 test;(5)限制通用顶级域名(restricted genetic top-levels domains, rgTLD),例如 biz、pro、name 等;(6)赞助顶级域名(sponsored top-level domains, sTLD),例如 aero、tel 等[7]。

1.1 数据来源与数据整理

本文的实验域名数据来源主要为 Alexa 网站所公布的全球排名前 100 万域名以及 DMOZ[8] 网站所提供的开放式目录域名网址分类数据,这两个数据来源是目前最常用的域名获取途径,其所含有的域名具有权威性和普遍性。数据整理首先是提取出原始数据中网址的域名,去掉网址中的网络路径,并对域名做去重工作,例如,http://news. baidu. com/index. html,则提取 baidu. com。经分析,原始数据共有 2923872 条网址,数据整理后共提取出域名数量为 1883134 个,表 1 是根据顶级域名的类型,对实验数据含有的顶级域名数量和域名数量的分布进行统计。

表 1 实验数据中顶级域名数量和域名数量分布

顶级域名类型	gTLD	rgTLD	sTLD	ccTLD	总计
顶级域名数量	364	3	15	240	622
域名数量	1345690	8797	8797	522255	1883134

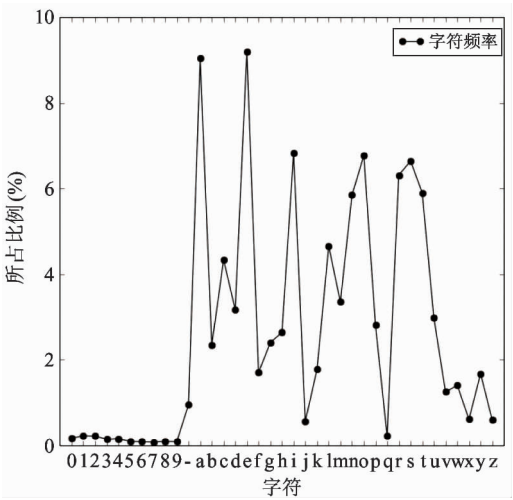
由表 1 可以看出,实验数据覆盖顶级域名的数量为 622 个,超过全球顶级域名数量的 1/2,并且覆盖了主流顶级域名,包括 com、net、org、cn 等,说明实验数据在顶级域名覆盖面具有代表性。本文中所生成的域名的顶级域名为 com,为 gTLD 类型顶级域名,目前在全球域名注册中,其域名数量最多,具有代表性,且验证比较快速方便。

1.2 域名字符分布频率特征

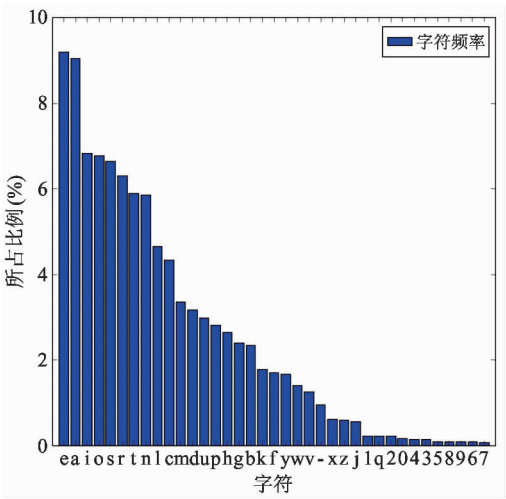
域名是由有限的字符组成,通过对大量构成域名的字符出现频率进行统计分析,可以发现域名中各个字符的分布规律。如图 1 字符分布频率所示,图 1(a)为组成域名的 37 个字符(0 ~ 9, a ~ z, –)的频率分布,图 1(b)为根据字符出现的频率的大小进行排序。

由图 1 可以发现,部分字符出现的频率相差较小,而部分字符则相差很大,特别是组成域名的字母频率特征分布不均匀,且波动大,数字频率整体较低但分布均匀。依照域名中字符出现的频率大小,将组成域名的字符频率分为 5 个等级,分别是极高频、次高频、中等频、低频和极低频,这些域名的字符分布特征,可对域名生成提供参考依据,如表 2 所示。

对域名中出现的字母频率分布进行分析,可以从英文角度来看,各个字母分布不均匀,则主要是 26个英文字母中,各个字母在单词中出现的频率也



(a) 字符分布频率(字符排序)



(b) 字符分布频率(频率排序)

图 1　字符分布频率统计

表 2　字符频率等级分布

序号	等级	字符	字符数量
1	极高频	a、e	2
2	次高频	i、n、o、r、s、t	6
3	中等频	b、c、d、g、h、k、l、m、n、p	10
4	低频	-、f、j、v、w、x、y、z	8
5	极低频	数字(0~9)、q	11

是不同的,另外,用户在申请注册域名时,会根据个人喜好或方便记忆的需求,参考英文单词,其中可能会加入数字或特殊字符,但这种情况较少,所以数字和特殊字符频率较低。通过对牛津英文字典中大量单词进行统计分析,各个字母出现的频率如表 3 所示[9]。通过对域名中字母频率特征与英文单词字母频率特征比较,如图 2 所示,可以发现域名和单词

表 3　牛津英文字典的字母频率

字母	频率(%)	字母	频率(%)	字母	频率(%)	字母	频率(%)
A	8.4966	H	3.0034	O	7.1635	V	1.0074
B	2.0720	I	7.5448	P	3.1671	W	1.2899
C	4.5388	J	0.1965	Q	0.1962	X	0.2902
D	3.3844	K	1.1016	R	7.5809	Y	1.7779
E	11.1607	L	5.4893	S	5.7351	Z	0.2722
F	1.8121	M	3.0129	T	6.9509		
G	2.4705	N	6.6544	U	3.6308		

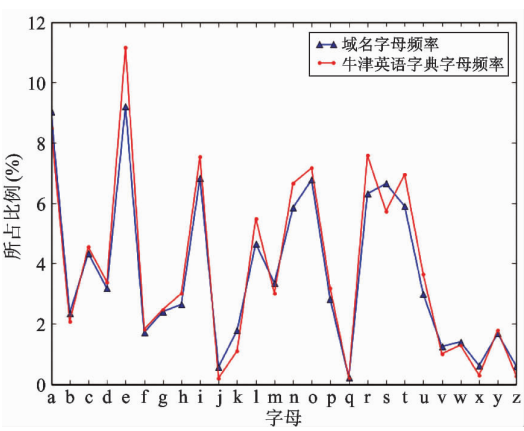


图 2　域名字母与牛津字典英语字母分布频率特征比较

中字母出现的频率特征非常接近。所以,这说明一方面域名中字母特征与单词中字母的特征特别相似,可以通过研究英文单词和字母的方法来研究域名的组成,另一方面域名中含有数字和特殊字符,所以又要加以区分进行研究。

1.3　域名长度特征

根据域名组成规定,主域名(即不包括顶级域名的部分)长度不超过 67 个字符,但是人们为了方便记忆,若无特殊原因,域名的长度特征一般不会过长。

通过对实验数据中主域名的长度进行统计，得到如图 3 所示的域名长度的分布图，经计算可知域名的平均长度为 10 个字符左右，域名较多集中在长度为5 ~ 15的区间。因此，通过遍历字符的方式来生成域名，显然是不可能，因为当域名长度达到10时，遍历所有可能字符，将会是海量数据，且真实域名的准确率非常低，同样，其他类似方法为提高域名的生成速率和准确率，需要将域名的长度限制在一定范围内。

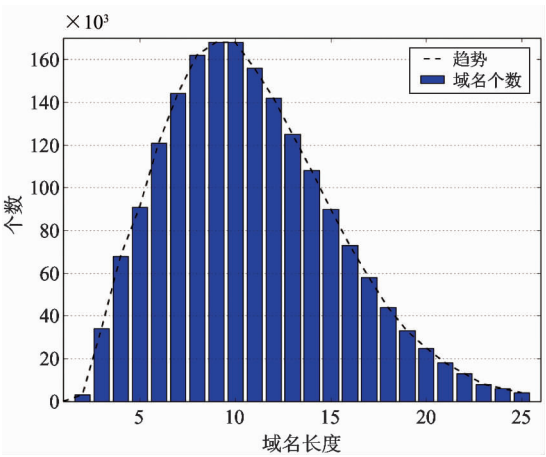


图 3　域名长度特征统计

1.4　域名不同位置字符分布频率特征

通过对域名的长度特征分析可知，域名的平均长度一般为 10 个字符左右，所以，本文对域名前 10 位置的字符出现的频率进行统计分析，统计结果如图 4 所示。可以看出，不同位置的字符分布频率不均匀，另外，随着域名长度的增加，字符出现的频率降低，这说明字符在之前出现后，之后出现的概率会减小。出现这种现象，一方面是由于英文单词的组成规律，另一方面是因为随着域名长度的增加，域名数量会减少，所以字符出现的频率降低。

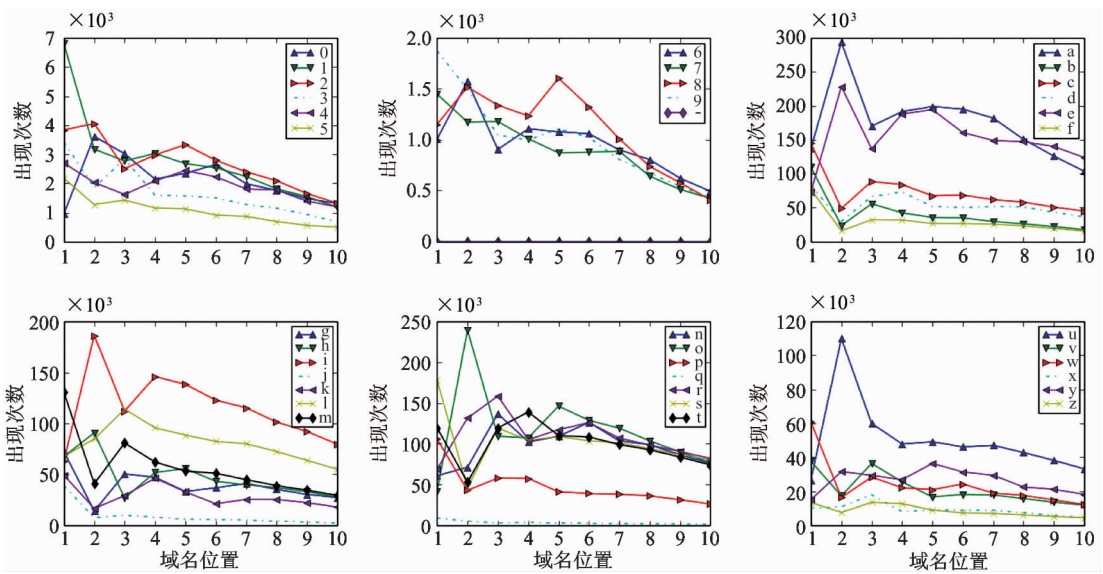


图 4　字符在域名不同位置出现的频率

同时可以发现，极高频和次高频的字符在第二个字符出现的频率会高于第一个字符出现的频率，第三个字符开始降低，且其整体频率在各个字符都高于其他等级的字符，相反地，中等频在首字符出现频率高，在第二个字符会出现下降，在第三个又开始增高，出现了跟单词字母很相似的规律。

图 5 所示为域名前 4 位置出现的字符频率分布统计。可以看出，第 1、2 位置中字符频率下降较快，而第 3、4 位置则出现基本符合线性降低，另外，部分字母在不同位置出现频率变化较大，例如字母 o，而数字则变化较小。

域名的首字符在一定程度上反映出人们命名域名的习惯，或者是从英语单词命名延伸出来。在基于改进马尔可夫链中的域名生成模型中，通过概率大小反映出这种命名偏好，表 4 为域名首字符频率分布。

1.5　顶级域名分布特征

顶级域名作为域名的重要组成部分，通过分析

其分布特征,可以更准确地为域名生成提供参考依据。图 6 为实验数据中各个顶级域名含有的域名数量分布统计。

通过图 6 可知,实验数据中各个顶级域名(前 25

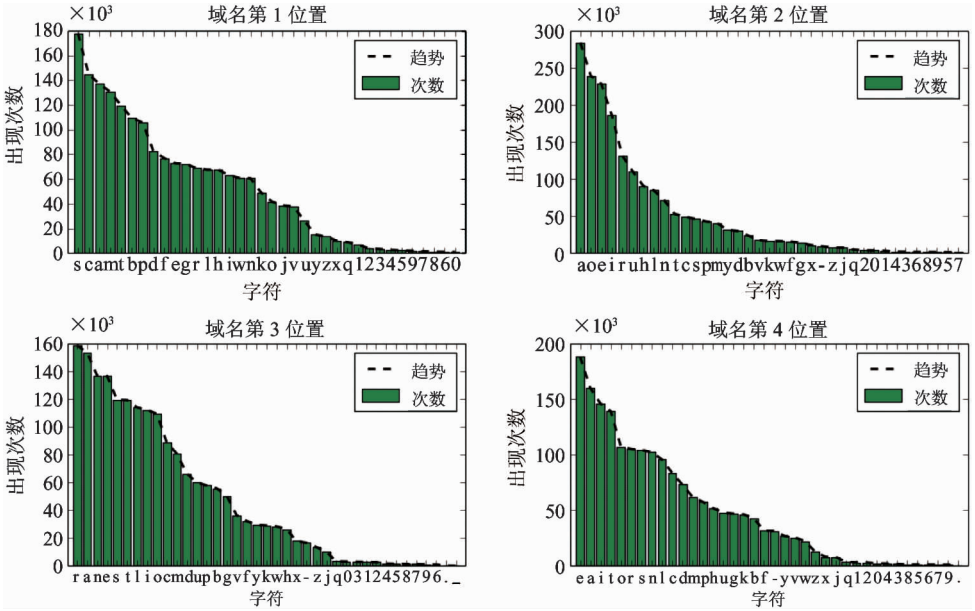


图 5　域名前 4 位的字符频率特征

表 4　域名首字符频率

字符	频率(%)	字符	频率(%)	字符	频率(%)	字符	频率(%)
s	9.45	e	3.87	o	2.21	2	0.20
c	7.68	g	3.83	j	2.05	3	0.18
a	7.27	r	3.65	v	2.00	4	0.14
m	6.93	l	3.59	u	1.40	5	0.12
t	6.34	h	3.58	y	0.80	9	0.10
b	5.80	i	3.35	z	0.71	7	0.08
p	5.62	w	3.23	x	0.54	8	0.06
d	4.38	n	3.22	q	0.47	6	0.05
f	4.08	k	2.60	1	0.36	0	0.05

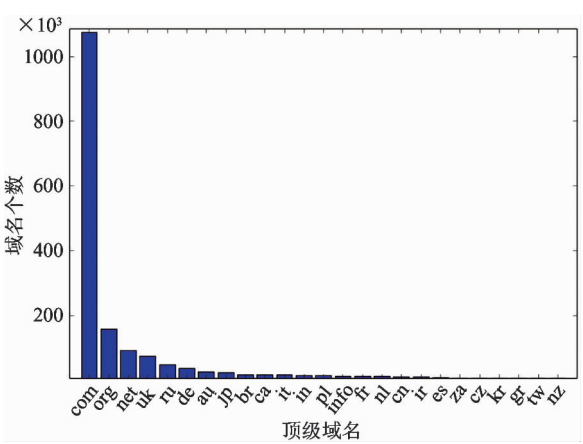


图 6　顶级域名含有的域名数量统计

个顶级域名)所含有的域名数量,com 含有域名数量远远高于其他顶级域名,并且前3名都为通用顶级域名,该顶级域名特征可为域名组成中顶级域名的选择做出参考。

2　基于改进马尔可夫链的域名生成算法

马尔可夫链通常用来建模排队理论和统计学中

的建模,还可作为信号模型用于熵编码技术,马尔可夫链也有众多的生物学应用,特别是增殖过程,可以帮助模拟生物增殖过程的建模[10]。

2.1　马尔可夫链定义

马尔可夫链是符合马尔可夫性质的随机变量 X_1, X_2, …, X_n 的一个序列,这些变量的范围为它们所有可能的取值的集合,称为“状态空间”,若当前状态、将来状态和过去状态是相互独立的[11],有如下公式:

$$P(X_{n+1} = x \mid X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = P(X_{n+1} = x \mid X_n = x_n) \quad (1)$$

其中 x 为过程中的某个状态,该性质称为马尔可夫性质。

设 $X_n(n \geqslant 0)$ 是 (S, P, Q) 上的随机过程,满足公式

$$P(X_{n+1} = x) = P(X_{n+1} = x \mid X_n = x_n) \quad (2)$$

则称 X_n 为马尔可夫链。说明如下:

(1) S: 系统所有可能的状态所组成的非空的状态集,用 $S = \{s_1, \cdots, s_N\}$ 表示,其中 N 为状态集元素数目;

(2) P:$P = [P_{ij}]_{n\times n}$ 是系统的状态转移矩阵,P_{ij} 表示系统当前处于 s_i 状态,转移到下一状态 s_j 的概率大小,N 是系统所有可能的状态值,其中 $\sum_{j=1}^{N} P_{ij} = 1$;

(3) Q: $Q = [q_1, q_2, \cdots, q_n]$ 是系统的初始概率分布,q_i 是系统在初始时刻处于状态 i 的概率,满足 $\sum_{i=1}^{N} q_i = 1$。

2.2　基于马尔可夫链的域名模型

通过对域名组成特征的分析,假设组成域名的字符序列满足马尔可夫性质,即根据域名不同位置的字符,只与该位置的前一字符有关,而与之前的字符无关。所有字符的组合是系统状态的集合[12]。

基于马尔可夫链的域名模型,需要提取域名的特征包括:

(1) 组成域名的字符集合, $C = \{c_1, c_2, \cdots, c_m\}$, 即为字母(a – z、A – Z)、数字(0 – 9),特殊字符(-);

(2) $first\_char_i$, 该参数为域名的首位出现各个字符的次数;

(3) $next\_char_{i,j}$, 该参数为字符 i 后跟字符 j 的次数;

(4) $first\_char\_freq_i$, 该参数对应马尔可夫链的初始分布概率 Q, 即域名的首位中各个字符出现的概率,满足公式

$$first\_char\_freq_i = \frac{first\_char_i}{\sum_k first\_char_i} \quad (3)$$

(5) $domain\_length$, 该参数为域名的长度,该参数根据域名长度分析所得;

(6) $next\_char\_freq_i$, 该参数对应马尔可夫链的状态转移矩阵,表示字符 i 之后为字符 j 所占所有字符中的比例,见公式

$$next\_char\_freq_{i,j} = \frac{next\_char_{i,j}}{\sum_K next\_char_{i,k}} \quad (4)$$

2.3　基于改进马尔可夫链的域名生成算法

由于组成域名的字符在域名不同位置出现概率不同,本文在基于马尔可夫链的域名模型基础上,提出了改进马尔可夫链的域名模型,通过在域名的各个位置上增加阈值 $threshold\_char_{i,j}$ 来限定选择的字符,该值表示字符 j 在域名的第 i 位置概率最小值,若字符出现的频率小于阈值,则不生成该字符,继续下一步操作。通过添加阈值,不仅可以加快域名的生成效率,同时也可以提高域名准确率。图 7 为改进后的基于马尔可夫链的域名获取模型图,其中 s_i 表示域名第 i 位的状态,即字符, $T_{i,j}$ 表示域名第 i 位出现字符 O_j 的阈值, O_j 表示组成域名的字符。

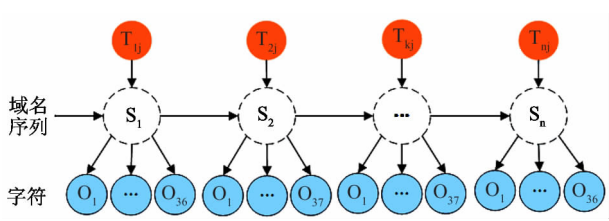


图 7　改进后的基于马尔可夫链的域名模型

根据以上改进马尔可夫链模型的描述,域名的生成算法的描述如下:

```
INPUT:
  域名列表(domain_list)、顶级域名(domain_prefix)、域名字符(domain_char)、
  current_domain(当前域名)
OUTPUT:
  新域名(new_domain)
BEGIN
  first_char_freq←根据domain_list计算得到首字符频率
  next_char_freq←根据domain_list计算得到转移矩阵
  threshold_char←根据domain_list计算得到各个字符在某位置的阈值
  domain_char←组成域名字符(0-9,a-z,-)
  last_char←get_domain_last_char(current_domain) //得到域名最后字符
  for i in domain_char
    next_char =Max (last_char, next_char_freq)
    if next_char >=threshold_char
      return merge (current_domain, next_char, domain_refix) //合并并输出
    else
      next_char_freq = next_char_freq – current_next_char_freq // 去掉不合适的char
    end if
    if next_char_freq is Null
      return False //无合适域名
    end if
  end for
END
```

2.4 基于 WHOIS 信息的域名验证方法

域名申请注册成功后,会在域名注册机构生成 WHOIS 记录,因此,通过查询域名的顶级域名所在的 WHOIS 服务器,可获得其 WHOIS 记录,通过该记录来验证域名是否注册。本文主要使用 WHOIS 信息验证域名存在,而未使用较为普遍的域名 DNS 查询,主要是因为,WHOIS 验证更加准确,域名 DNS 查询虽然速度较快,但是其所查询的域名是在网络中流通才可以解析正确,否则无法解析,错误率较高,而域名只要真实存在,则一定有域名 WHOIS 信息[13]。

对于已在域名注册机构注册的域名,例如 abcd.com,该域名 WHOIS 信息如图 8(a)所示,而未注册的域名 WHOIS 信息,例如 zottdd.com,如图 8(b)所示。

```
Domain Name: ABCD.COM
Registrar: CSC CORPORATE DOMAINS, INC.
Sponsoring Registrar IANA ID: 299
Whois Server: whois.corporatedomains.com
Referral URL: http://www.cscglobal.com/global/web/csc/dig
ml
Name Server: ORNS01.DIG.COM
Name Server: ORNS02.DIG.COM
Name Server: SENS01.DIG.COM
Name Server: SENS02.DIG.COM
```

(a)域名存在的 WHOIS 信息

```
Whois Server Version 2.0

Domain names in the .com and .net domains can now be regis
tered
with many different competing registrars. Go to http://www
.internic.net
for detailed information.

No match for "ZOTTDD.COM".
>>> Last update of whois database: Tue, 26 Apr 2016 13:36:
20 GMT <<<
```

(b)域名不存在的 WHOIS 信息

图 8 域名 WHOIS 信息格式

3 实验结果及分析

使用改进马尔可夫链的域名生成算法,生成不同数量级别的域名,并通过域名 WHOIS 信息验证是否真实存在,根据生成域名的不同长度,统计分析算法生成域名的效率和准确率,为使生成域名的数

量多,且验证速率快,所有生成域名的顶级域名为 com。同时,把该方法与原理相似的字符遍历生成域名的方法进行了比较,分别在生成相同数量域名其正确率和生成相同数量真实域名其探测次数两个方面进行。

3.1 域名生成结果分析

使用算法生成数量级分别为 1000、10×10^3、100×10^3、200×10^3、500×10^3、1000×10^3、2000×10^3、5000×10^3 的域名,根据生成域名的长度分布整理数据,可得表 5 所示结果。

由表 5 可以看出,设置生成域名的数量与实际生成的域名数量相比,大约少 9% 个域名,主要原因是生成域名的最后阶段,当某概率下将要生成的域名数量为小数时,会舍去小数部分,所以导致生成的域名数量稍少,但不影响整体结果。

表 5 生成域名的长度分布

生成域名数量($\times10^3$)	不同长度的域名分布								实际域名
	1	2	3	4	5	6	7	8	
1	36	718	158	--	--	--	--	--	912
10	36	1070	6385	1489	--	--	--	--	8980
100	36	1296	18894	58542	12306	2	--	--	91076
200	36	1296	23080	103083	53688	455	--	--	181638
500	36	1296	28887	180850	226485	16420	--	--	453974
1000	36	1296	33595	252200	515163	104164	239	--	906693
2000	36	1296	38177	333791	1001466	432194	7157	--	1814127
5000	36	1296	42821	464702	2017262	1857306	155181	78	4538682

图 9 描述在不同的域名数量级别下,不同长度域名的分布情况,可以看出,当域名数量达到 100×10^3 时,长度为 4 的域名数量开始增长迅速,到 500×10^3 时,则因为长度为 5 的域名数量增加,而所占百分比相应减少,这也符合域名数量的分布规律。

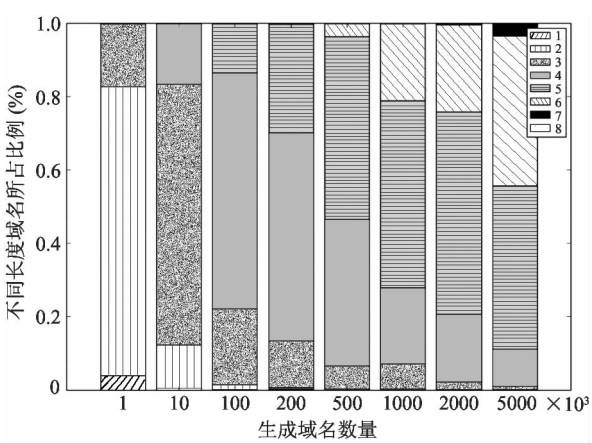


图 9 不同长度域名组成情况

3.2 域名验证结果分析

使用域名 WHOIS 查询的方法对生成的域名进行验证,确认其是否真实存在,并且根据其不同长度的域名分布,判断该方法的准确性,该方法的优势在于无论域名是否在网络中,只要域名注册后,就会有域名 WHOIS 信息。

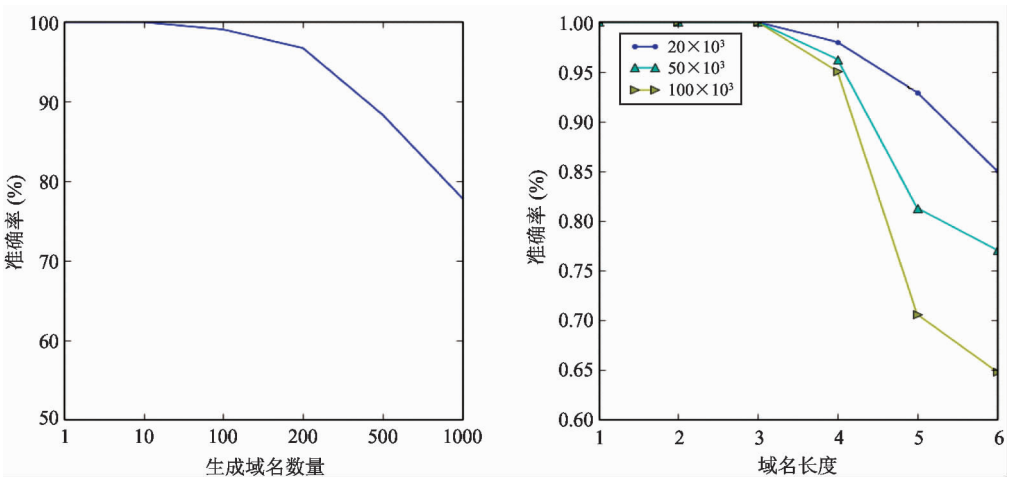
对所有生成的近 100 万个域名进行 WHOIS 验证其是否存在,验证结果如表 6 所示,同时对不同长度的域名正确率进行纵向和横向比较,结果如图 10 所示。

由表 6 域名验证结果可知,当生成的域名数量较少时,域名长度在 4 个字符之内,则真正注册的域名分布多,生成域名的准确率为 100%,随着生成域名数量的增加,域名长度增加,其准确率开始降低。

由图 10(a)可知,生成域名数量较少时,域名真实存在的准确率达到 90% 以上,随着生成域名数量的增加,其准确率开始降低。通过对生成的不同域名长度进行比较,由图 10(b)可知,准确率的降低主要原因在于长度较长的域名数量增加以及其准确率较低,算法对于生成的较长域名的准确性有待提高。

表 6　域名验证结果

实际域名数量	存在域名数量	不同长度的域名分布							准确率
		1	2	3	4	5	6	7	
912	912	36	718	158	--	--	--	--	100%
8980	8980	36	1070	6385	1489	--	--	--	100%
91076	90235	36	1296	18894	57986	12021	2	--	99.07%
181638	175700	36	1296	23080	101008	49893	387	--	96.73%
453974	401104	36	1296	28887	174083	184144	12658	--	88.35%
906693	705974	36	1296	33595	239691	363735	67541	80	77.86%



(a)不同数量域名准确率变化情况　　(b)不同长度域名准确率变化情况

图 10　生成域名准确率统计

3.3　与域名字符遍历方法比较

基于改进马尔可夫链模型的域名生成算法,主要是解决了不通过字符遍历的方式获得域名,使生成域名的准确率更高,效率更快。因为域名的每个位置有 36(首末字符不能为"-")或 37 种可能字符,当域名长度较长时,遍历所有字符并且验证域名是否存在,这个数量级会是天文级,显然是不可能,数据量过于巨大,且生成的域名准确率非常低。

经测试,当使用两种算法同时生成 90 万个域名时,基于改进马尔可夫链的域名生成算法生成的真实域名数量占到 77.86%,而使用遍历方法生成 90 万个域名时,准确率只有达到 61.51%,且随着生成域名数量和域名长度的增加,其准确率急剧降低。

当使用两种算法同时生成 175700 个真实存在域名时,基于改进马尔可夫链的域名生成算法,只需要生成 181638 个域名即可,而使用遍历的方法,则需要生成 226266 个域名,同样,随着生成域名的数量增加,遍历方法效率越来越低。可以看出,基于改进马尔可夫链的域名生成算法有很好的准确率和效率。

4　结　论

使用本文提出的改进马尔可夫链算法生成了大量域名,并且验证了域名是否存在。实验结果证明,该方法具有较高的域名生成准确率,而且通过使用 WHOIS 信息进行验证表明,该方法比其他域名获取方法更加准确与快速,可以使用它获取大量域名,为域名研究提供基础数据。此外,随着域名数量的不断增加,能够丰富初始域名列表,提高算法的准确率。本文接下来进一步的研究,主要围绕算法准确性的提高和加快域名验证速率展开。

参考文献

[1] Mockapetris P. RFC 1035-Domain names-implementation and specification, Nov. 1987

[2] Daigle L. WHOIS Protocol Specification. IETF RFC 3912, Sept. 2004

[3] Internet Assigned Numbers Authority. http://www.iana.org: IANa, 2016

[4] Alexa.com. Alexa - Actionable Analytics for the Web. http://www.alexa.com: ALEXA, 2016

[5] CZDS. Centralized Zone Data Service. https://czds.icann.org: ICANN, 2016

[6] 许笑, 张伟哲, 张宏莉等. 广域网分布式 Web 爬虫. 软件学报, 2010, 21(5):1067-1082

[7] Wikipedia. Top-level domain. https://en.wikipedia.org/wiki/Top-level_domain: WIKIPEDIA, 2016

[8] dmoz.org. DMOZ - the Open Directory Project. http://www.dmoz.org: DMOZ, 2016

[9] Oxforddictionaries. Which letters in the alphabet are used most often. http://www.oxforddictionaries.com/words/which-letters-are-used-most: OXFORDDICTIONARIES, 2016

[10] Teugels J L. Markov chains: Models, algorithms and applications. *Journal of the American Statistical Association*, 2008, 103(483):1325-1325

[11] Wikipedia. Markov Chain. https://en.wikipedia.org/wiki/Markov_chain: WIKIPEDIA, 2016

[12] 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. 计算机应用与软件, 2015, 32(6): 152-155

[13] 张尼, 郭莉, 方滨兴. 一个分布式网址定位平台的设计和实现. 计算机工程, 2005, 131(24): 138-140

A study of domain name acquisition method based on improved Markov chain

Cheng Yanan*, Li Zhengmin**, Chi Lejun*, Xu Haiyan*, Lu Keyu*

(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

(**National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029)

Abstract

To solve the problem that current domain name acquisition methods have the low efficiency and can only acquire a small number of domain names, the study conducted the statistical analysis of the quantifies of domain names collected in the early stage to find the composition rules and distribution characteristics of domain name characters, and then designed a domain name model based on Markov chain according to these characteristics, and proposed a domain name generation algorithm based on the improved Markov chain. The generated domain names were verified with WHOIS records to confirm whether the domain names exist. The experimental results show that the proposed algorithm has a high accuracy in domain name generating. And compared with other domain name acquisition methods, this method has the faster generating speed, and can generate more domain names with a wider coverage of Top-Level Domains.

Key words: domain name, Markov chain, character frequency, WHOIS