

基于顶级域解析日志的递归 DNS 识别方法^①

胡安磊^②* ** ** ** 谢高岗*** ** ** 苑卫国** 魏金侠*** ** ** 付豪*** ** **

(* 中国科学院计算技术研究所 北京 100190)

(** 中国互联网络信息中心 北京 100190)

(*** 中国科学院计算机网络信息中心 北京 100083)

(**** 中国科学院大学 北京 100049)

摘要 递归域名系统(DNS)根据其服务的开放性、进行递归查询的目的等可分为不同的类型,递归 DNS 类型的准确识别,对于对根、顶级和各级权威 DNS 的分析与运行具有重要意义。针对递归 DNS 的准确识别问题,本文通过分析 .CN 国家顶级域名系统的解析日志,提出基于递归查询的行为特征识别递归 DNS 类型的方法。该方法从多个维度信息来筛选甄别表征全量日志信息,基于无监督特征选择方法选择重要特征,实现同类型递归 DNS 的准确聚类。实验结果表明,该方法能高效准确识别出递归 DNS 类型。

关键词 递归域名系统(DNS); 特征识别; 无监督特征选择; 聚类算法

0 引言

域名系统(domain name system, DNS)是互联网上关键基础设施之一,实现域名和网际互连协议(internet protocol, IP)地址间的转换。DNS 分为权威 DNS 和递归 DNS 两类,其中递归 DNS 是用户使用域名解析服务的入口。准确识别出真实递归 DNS,对于保障各级权威域名系统的安全、支撑对递归 DNS 的深入研究和安全管理是非常重要的。

对递归 DNS 进行准确识别需要全局和海量的递归查询行为数据支撑,局部网络(如运营商网络、校园网)内的数据存在覆盖面不够的问题。.CN 国家顶级域名系统拥有全局视角海量的递归查询行为数据,海量递归查询中也存在大量非真实和异常递归查询,会直接影响 .CN 国家顶级域名系统服务的安全稳定,目前尚未基于此类数据对递归 DNS 进行准确的识别研究。因此本文以 .CN 国家顶级域名

系统某服务节点连续一周的解析查询日志为数据集,对 .CN 国家顶级域名解析日志进行递归查询行为特征分析和递归域名服务聚类分析,在聚类的基础上结合专家经验对递归 DNS 进行准确的识别,可以支撑对递归 DNS 进行进一步研究,形成真实递归 DNS 的清单可用于 .CN 国家顶级域名系统进行服务管理和安全防护。

本文的主要贡献如下。

(1)特征处理方面。提出一种基于递归 DNS 源 IP 的递归查询行为特征选择方法,利用特征自表示性将关键具有代表性的特征选择出来,减少冗余特征对模型的影响。

(2)递归 DNS 识别方面。利用改进的聚类方法对海量递归 DNS 查询请求数据聚类,首先利用粗聚类方法结合模型的关键指标和计算时间确定 K 值,然后针对 DNS 查询行为日志进行聚类,最后通过与专家知识库的关联,识别出各种类型的递归 DNS。

(3)数据来源方面。使用真实的递归 DNS 到

^① 国家自然科学基金(62072437)和国家自然科学基金区域联合重点基金(U20A20180)资助项目。

^② 男,1979年生,博士,教授级高级工程师;研究方向:网络安全,网络测量,互联网基础资源管理;联系人,E-mail: huanlei@cnnic.cn。(收稿日期:2022-08-24)

.CN 国家顶级域名系统的查询日志作为数据源,选取 2021 年 12 月某一周的查询日志,包含约 4.2 亿条查询记录。通过实验分析与人工验证,本文方法可以将上述数据准确地划分到各个查询行为类别中。

1 相关工作

如图 1 所示,域名服务体系由 2 大类别和 4 个环节的 DNS 组成。第 1 类是权威 DNS,包括根 DNS、顶级 DNS 和其他各级权威 DNS 3 个环节,负责维护和保存各级权威域域名信息,接受递归 DNS

查询请求;第 2 类是递归 DNS,为终端用户提供域名查询服务。

顶级 DNS 处于整个域名服务体系的次顶端,目前全球顶级域名数量已超过 1500 个,主要包括国家和地区顶级域名(如中国 .CN,美国 .US),通用顶级域(如 .COM 和 .NET)和新通用顶级域(如 .TOP 和 .XYZ),特别是国家顶级域名用于标识特定国家域名空间,是国家主权在网络空间的象征。截止 2021 年 12 月,我国域名总数为 3593 万,其中 .CN 国家顶级域名数量为 2041 万,占域名总数的 56.8%^[1],.CN 国家顶级域名系统在我国整个域名服务体系中处于关键地位。

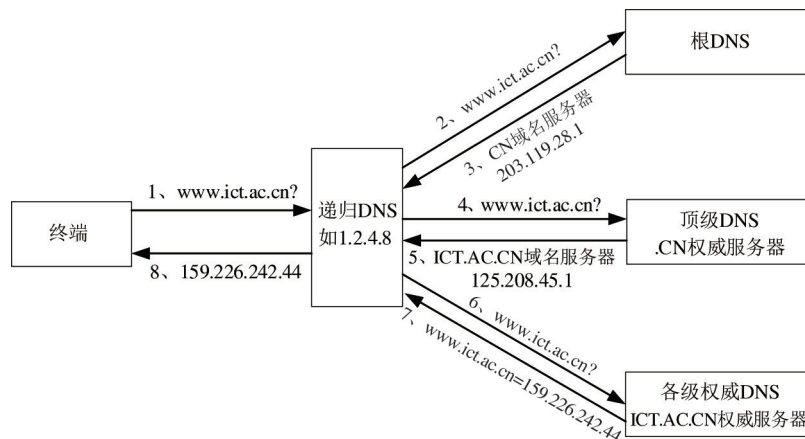


图 1 域名服务体系架构

递归 DNS 根据其服务的开放性可分为开放递归 DNS 和非开放递归 DNS;根据其进行递归查询的目的,可以分为真实的递归 DNS 和非真实的递归 DNS。上述几大类型的递归 DNS 可根据其行为特征进行进一步地类型识别。限于数据获取等原因,目前对递归 DNS 进行类型识别的研究主要通过主动扫描探测的方法对开放递归 DNS 进行识别,这种方法在递归 DNS 识别的覆盖面上存在不足。

递归 DNS 测量对于域名系统运维与安全保障至关重要。根据被动采集递归 DNS 的查询响应日志,以及从用户端视角对全球开放数百万公共递归服务主动探测,结合递归 DNS 查询行为统计特征分析,发现递归 DNS 面临缓存投毒和拒绝服务威胁^[24],存在大量配置问题和安全隐患^[5-8],全球范围内公共递归 DNS 存在严重域名解析劫持问

题^[9-11]。上述研究主要集中在开放递归 DNS 的探测识别、测量评价、特征统计和行为分析方面,研究分析的递归 DNS 覆盖面存在不足。

递归 DNS 查询行为与网络整体运行状况和各类网络安全攻击行为紧密联系。通过对全球 300 万开放递归 DNS 的主动探测和恶意响应行为研究,可以分析递归 DNS 对互连网络安全和稳定的影响^[12];也可以通过对递归域名查询日志的深入监测和查询行为分析,开展诸多针对恶意域名^[13-15]、僵尸网络^[16-17]等危害网络安全的异常行为研究。上述研究主要集中在通过递归 DNS 查询行为分析网络中的恶意攻击行为方面,对网络攻击、恶意行为和递归 DNS 的关系研究不足。

针对 .CN 国家顶级域名解析日志查询行为的分析测量有利于了解国内用户互联网访问特征,以

及攻击异常行为及时发现。 .CN 国家顶级域名递归 DNS 和域名的查询频度遵循明显的幂律分布特征,从整体分布统计特征角度检测 DNS 查询行为是否异常^[18]。基于 .CN 国家顶级域名的递归 DNS 日志查询行为的特征提取,可基于 K-means 算法进行 DNS 查询模式分析^[19]。但上述研究在特征提取过程中,只应用到了 IP 和域名基于时间维度统计的直接特征,未考虑到不同特征间的联系和特征重要性的区别,对噪声特征比较敏感,并集中在递归 DNS 的查询的统计规律特征分析,未对递归 DNS 的识别作进一步的研究。

综上,使用 .CN 国家顶级域名系统的日志研究

递归 DNS,可覆盖我国几乎所有的递归查询行为,全面准确分析递归查询行为并进行真实递归 DNS 的识别研究,有助于支撑 .CN 国家顶级域名系统的安全保障和对递归 DNS 的进一步研究。

2 数据集与查询行为

2.1 数据集

本文采集 2021 年 12 月 21 日至 27 日连续 7 d 的 .CN 国家顶级域名系统某解析节点连续一周的查询日志,其日志信息示例见表 1,数据集全局统计信息见表 2。

表 1 DNS 查询请求日志信息示例

DNS 日志字段名称	举例 1	举例 2	举例 3
查询时间	20211224084916	20211227094901	20211227171030
客户端 IP 地址	220.187.246.34	220.187.246.34	220.187.246.34
客户端查询端口号	34585	61082	57502
客户端查询域名	www.cnnic.cn	www.sina.com.cn	mail.chinatelecom.cn
资源分类	IN	IN	IN
资源记录类型	A	A	NS
所查询的目的 IP	203.119.28.1	203.119.28.1	203.119.28.1
所查询的权威域名	cnnic.cn	sina.com.cn	chinatelecom.cn
所查询的顶级域名	cn	com.cn	cn
所查询节点名称	Alien	alien	Alien

表 2 数据集全局统计信息

参数	值
数据采集时段	21 日 23:30 ~ 27 日 23:30
大小	49.1 GB
递归查询请求总量	420 000 000
IPv4 地址量	69 386
IPv6 地址量	1
查询域名总量	93 320 616

其中例 1 “20211224084916” 为查询时间,表示 2021 年 12 月 24 日 08 时 49 分 16 秒;“220.187.246.34” 为客户端 IP 地址;“34585” 为客户端查询端口号;“www.cnnic.cn” 为客户端查询域名;“IN” 是 Resource Class 中最常见的一种,表示 Internet(另有少量 CS、CH、HS);“A” 为资源记录类型(resource re-

cord type),表示所查询的域名服务器类型,为 IPv4 类型,另外还有 IPv6 类型“AAAA”、邮件交换地址类型“MX”等;“203.119.28.1” 为查询的目的 IP,即 DNS 服务器的公网 IP 地址;“cnnic.cn” 为客户端查询的权威域名;“cn” 为查询的顶级域名;“alien” 为被查询解析节点名称。

2.2 查询行为分析

首先根据对递归 DNS 查询日志的资源记录类型统计,发现查询类型一共有 42 种,其中查询次数超过 80 万的类型有 9 个,分别包括 A、AAAA、NS、TYPE65(HTTPS)、TXT、DS、CNAME、SOA 和 MX,具体分布情况如表 3 所示。查询类型为 A、AAAA 和 NS 记录查询占比超过 93%,其他查询类型包括 TYPE65(HTTPS)、TXT、DS、CNAME、SOA 记录、MX 记录等占比不足 7%。

表 3 主要域名查询记录类型分布

序号	域名查询记录类型	查询数量	所占比例
1	A	278 926 267	68.10%
2	AAAA	74 741 679	18.25%
3	NS	37 566 719	9.17%
4	TYPE65(HTTPS)	6 423 077	1.57%
5	CNAME	4 166 518	1.02%
6	SOA	3 468 286	0.85%
7	TXT	2 331 208	0.57%
8	MX	1 058 916	0.26%
9	DS	883 609	0.22%

正常的递归 DNS 查询量每日随时间变化具有显著的周期性,图 2 展示的是选取节点的 .CN 顶级

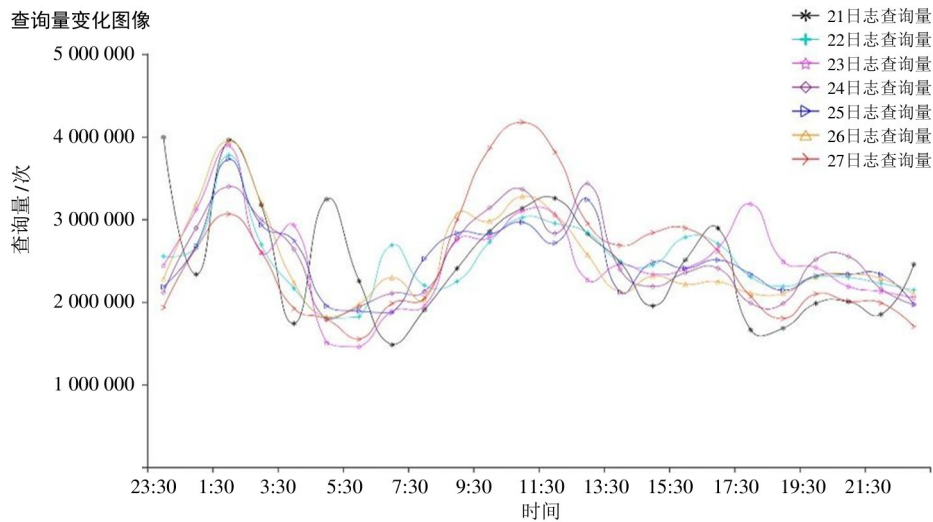


图 2 每日查询量变化趋势

本文对递归 DNS 查询数量整体频次特征分别基于源 IP 和域名 2 个方面进行实证分析。可以发现,基于源 IP 递归 DNS 查询量与其频次分布呈现一种长尾特征,约 95% 以上的递归查询请求由大约 5% 的递归 DNS 发起,如图 3 所示。另外,基于域名访问量与其频次分布也呈现一种长尾特征,约 96% 的域名查询请求次数低于 5 次,占总查询量不足 40%,也就是说占比 4% 的域名所产生查询量占比总查询量近 60%,如图 4 所示。

3 识别方法

本文提出的基于自表示特征提取的递归 DNS 行为识别流程如图 5 所示。

域名服务器在 7 d 内每日按小时统计的查询量变化情况。图中显示来自递归 DNS 的查询在凌晨和上午呈上涨趋势,分别在凌晨 1:30 左右和上午 10:30 左右达到较高值。其中出现凌晨域名查询高峰与存在注册机的域名抢注行为密切相关,在晚上呈上涨趋势,凌晨 2 点左右达到较高值,之后呈下降趋势并在早上 6 点左右开始上升,在上午 11 点左右达到较高峰值之后缓慢下降。

根据文献[18]递归 DNS 的整体查询行为在正常网络状况下遵循 Zipf's 分布,在双对数坐标图中呈现明显的线性特征(负相关),即递归 DNS 的查询请求具有整体集中分布的特点。

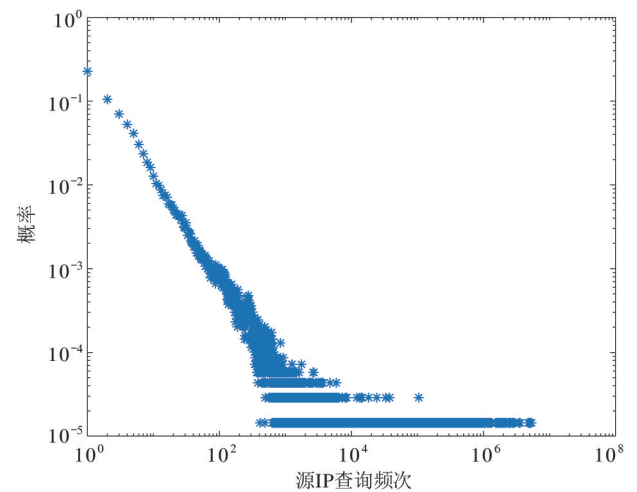


图 3 基于源 IP 的递归 DNS 查询量频次分布

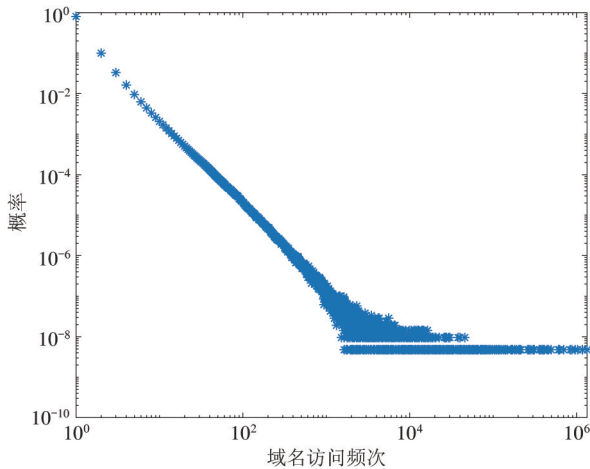


图4 基于域名的递归DNS访问量频次分布

3.1 特征扩展

对于递归DNS查询行为来说,查询日志中的IP地址来源分布、单位时间查询量变化、域名信息熵分布、查询频率等因素对递归分类结果影响比较大。同时考虑查询行为的周期性因素,为了能够更准确地对递归DNS查询行为进行分类,将周期作为关键因素考虑进来。因此,综上本文基于递归DNS的IP地址角度梳理递归查询行为共计9维的直接特征,

具体如下所述。

- (1) 查询请求总数 x_1 : 统计日志中每个IP的查询请求数量,反映了IP的活跃程度。
- (2) 每小时最大查询总次数 x_2 : 统计每个IP每个小时内的请求查询次数,并取其中最大值。
- (3) 每小时最大查询变化率 x_3 : 统计每个IP每个小时内的请求查询次数,并计算每个小时相较于上一小时的变化比率。
- (4) IP端口信息熵 x_4 : 统计日志中每个IP的所有端口,并计算信息熵, $entropy = -\sum p \cdot \log(p)$, 其中 p 是每个端口出现的概率。当DNS发生流量异常时,会引起查询源IP端口熵值的突变。
- (5) 域名种类 x_5 : 统计每个IP对应的域名的种类数,反映提交的域名请求分布情况。
- (6) 域名信息熵 x_6 : 统计日志中每个IP的所有域名字符串,并计算信息熵。
- (7) 权威域名信息熵 x_7 : 统计日志中每个IP的所有权威域名,计算信息熵。
- (8) 顶级域名信息熵 x_8 : 统计日志中每个IP的所有顶级域名,计算信息熵。
- (9) 重复查询次数 x_9 : 统计日志中每个IP所查询域名出现的平均次数,域名重复查询次数越大说明同一个域名被访问的平均时间间隔就越小。

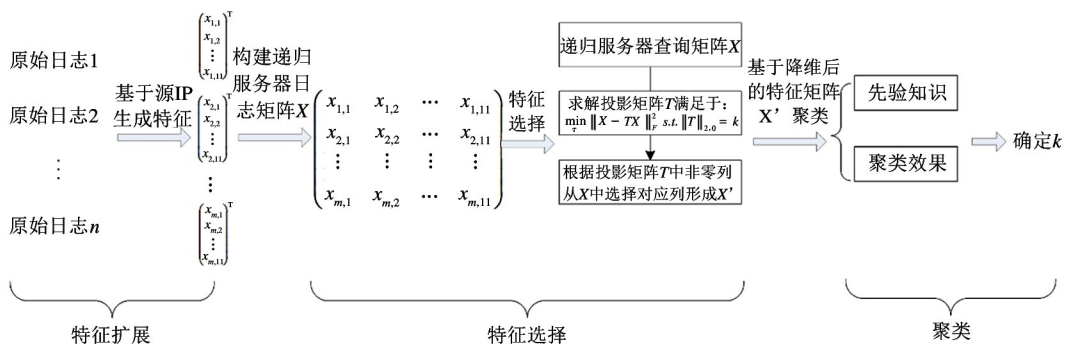


图5 基于自表示特征提取的递归DNS行为识别流程

3.2 基于系数表示的无监督特征选择

本文采用一种基于稀疏表示的无监督的特征选择方法,基于特征之前的自表示性将噪声特征去掉,留下关键特征。基于特征自表示性进行关键特征选择的原理是利用稀疏约束项^[20]对初始特征进行重新表示,形成特征重表示矩阵,然后通过特征重表示矩阵与初始特征矩阵差异最小化来求解稀疏矩阵的非零列,非零列即为要选择的关键特征。

稀疏表示目前得到广泛应用,现有分析结论中

已发现了冗余特征具有自表示性^[21]。对于DNS查询日志数据特征矩阵 X , x_j 为矩阵 X 的第 j 行,每一行表示一条日志,行数表示查询日志样本个数, $x_j = (x_{j1}, x_{j2}, \dots, x_{j9})$, $j \in [1, M]$; 每一列表示一维特征,初始维度共9维,每一维特征用 x_i , $i \in [1, 9]$, 来表示。

令 T 为投影矩阵, t_i 表示矩阵 T 的第 i 行, $t_i = (t_{i1}, t_{i2}, \dots, t_{iM})$, $i \in [1, 9]$ 。投影矩阵 T 的列向量可以反映不同特征的重要性。当矩阵 T 只有 k 列不为

0 时,与之对应的 DNS 查询日志特征矩阵 X 只有 \bar{k} 维特征被选择,其余特征没有被选择。

DNS 查询日志数据重构损失项可以表示为 $\|X - \sum_{i \in [1,9]} t_i \cdot X\|_F^2$ 。该重构损失项的含义是 DNS 查询日志数据的每个特征由其他维度的特征进行表示, t_i 为数据特征矩阵 X 的重构系数,描述了第 i 维特征 x_i 对数据整体特征重构的贡献度。如果 t_i 为 0 向量,则对应的第 i 维特征 x_i 的贡献度为 0。

为了同时满足 DNS 查询日志重构矩阵 $T \cdot X$ 与 DNS 查询日志特征矩阵 X 之间的误差最小,且投影矩阵 T 只有 k 列不为 0,则 DNS 查询日志重构损失项的约束优化问题可以表示为以下形式:

$$\min_T \|X - T \cdot X\|_F^2, \text{ s. t. } \|T\|_{2,0} = \bar{k} \quad (1)$$

利用交错方向乘子法将上述优化问题变换为拉格朗日函数,并利用迭代优化求解变量的方法进行求解,得到 \bar{k} 值和投影矩阵 T 。根据投影矩阵 T 的 \bar{k} 个非零列,可以对 DNS 查询日志矩阵 X 中的特征进行选择。

3.3 聚类分析

K-means 算法是一种基于距离聚类的方法,将 M 个 DNS 查询日志数据样本划分到 k 个类别中,要求满足同一个类别中的样本相似度较高而不同类别中的样本相似度较低。与其他算法相比,其适用于大规模数据的场景,并且收敛速度比较快,其每个类别均用该类中所有数据的平均值来表示,这个平均值即被称作为聚类中心。并且对于数值属性的数据,能很好地体现出聚类在集合和统计学上的意义,目前在聚类算法中是被应用最广泛的算法。选取的数据是连续 7 d 的 .CN 国家顶级域名系统某解析节点连续一周的查询日志,考虑到其数据量大、没有标签、大多特征为统计特征等原因,本文选择用 K-means 算法来对全量日志进行聚类,实现对大规模查询日志的快速分组。

设待分类的 DNS 查询日志数据集为 $\{x_1, x_2, \dots, x_M\}$, 计划将这些样本分为 k 类(粗聚类确定),步骤如下。

(1) 首先针对全量的查询日志数据直接进行粗

聚类,即在没有确定簇值的情况下先对数据进行多簇值聚类,然后根据 Inertias 值变化趋势、计算时间随簇数量变化趋势来综合分析确定簇值 k 。

(2) 任意选定 k 个样本作为初始聚类中心 $\{x_1^{(0)}, x_2^{(0)}, \dots, x_k^{(0)}\}$, 令 $\bar{t} = 0$ 。

(3) 将待分类的 DNS 查询日志数据样本按照最小距离原则,将每一个查询日志样本分别划分到 k 类中的每一类,即:若 $d_j^{(i)} = \min_v \{d_v^{(i)}\}$, ($j = 1, 2, \dots, M$), 式中 $d_j^{(i)}$ 表示 x_j 和中心 $x_j^{(i)}$ 的距离,则可以产生新的聚类簇 $\omega_j^{(i+1)}$, ($j = 1, 2, \dots, k$), 其中上角标表示迭代的次数。

(4) 对于每一个簇,重新计算聚类中心, $x_v^{(i+1)} = \frac{1}{n_v^{(i+1)}} \sum_{x_j \in \omega_v^{(i+1)}} x_j$, ($v = 1, 2, \dots, k$), 其中 $n_v^{(i+1)}$ 是 $\omega_v^{(i+1)}$ 类所含样本的个数。

(5) 如果 $x_j^{(i+1)} = x_j^{(i)}$, ($j = 1, 2, \dots$), 聚类中心不在该表则结束;否则 $\bar{t} = \bar{t} + 1$, 转到第(2)步。

4 实验结果分析

4.1 聚类参数的选择

本文通过对 .CN 国家顶级域名系统收到的来自递归 DNS 的查询数据进行分析,利用先验知识与样本特征对聚类模型性能的影响选取恰当的聚类数。在先验知识方面,考虑当前常见的递归查询行为有真实递归 DNS 发起的规范查询,还有各类非规范的递归查询,如探测查询、随机查询、错误查询、攻击查询等,因此,可以选择聚类参数 $k > 5$ 。

在样本特征对聚类模型性能的影响方面,本文选择 Inertias 与不同 k 值情况下的聚类时间综合选取 k , 其中 Inertias 值是 K-means 模型对象的属性,作为没有真实分类结果标签下的非监督式评估指标,表示样本到最近的聚类中心的距离总和。该值越小越好,越小表示样本在类间的分布越集中。

针对实验样本进行聚类,聚成多个类别,观察类别数量与 Inertias 值的对应关系变化,结果如图 6 所示,横坐标表示聚成的类的数量,纵坐标表示 Inertias 值。实验结果显示,聚类类别 $k \geq 20$ 时,曲线变化率接近平缓,Inertias 值下降不明显。

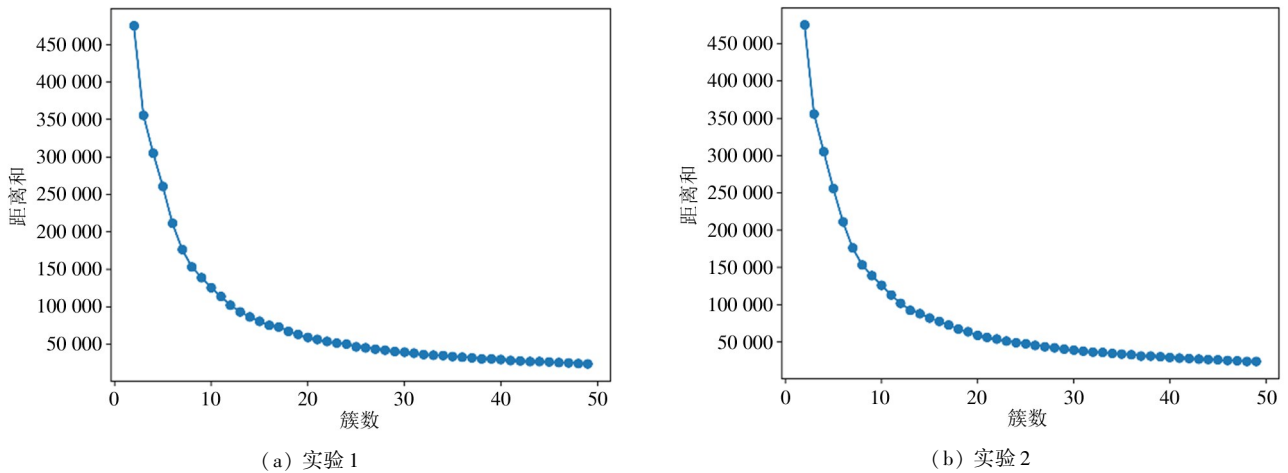


图 6 聚类算法 Inertias 值变化趋势图

本文为了验证聚成类别数量对模型性能的影响,做了 5 次聚类实验,结果如图 7 所示。从图中的结果可以看出,聚成的类别越多,模型运算时间越长,整体呈现出阶段性的线性增长。第 1 个阶段聚成的类别数量小于 12 类,运算时间在 0.05 s 以内;第 2 个阶段聚成的类别数量小于 21 类,运算时间在 0.062 s 以内,同比第 1 个阶段时间增长率为 24%;第 3 个阶段聚成的类别数量小于 32 类,运算时间在

0.08 s 以内,同比第 2 个阶段时间增长率为 29%;第 4 个阶段聚成的类别数量小于 50 类,运算时间在 0.1 s 以内,同比第 3 个阶段时间增长率为 25%。通过分析可知,随着聚类数量的增加,时间增长率最低的是从第 1 个阶段到第 2 个阶段,即 $k \leq 21$ 时,聚类时间增长率比较小。因此,综合考虑先验知识、随着聚类数量变化引起的聚类模型 Inertias 值以及聚类时间变化等因素,选取聚类数量 $k = 20$ 。

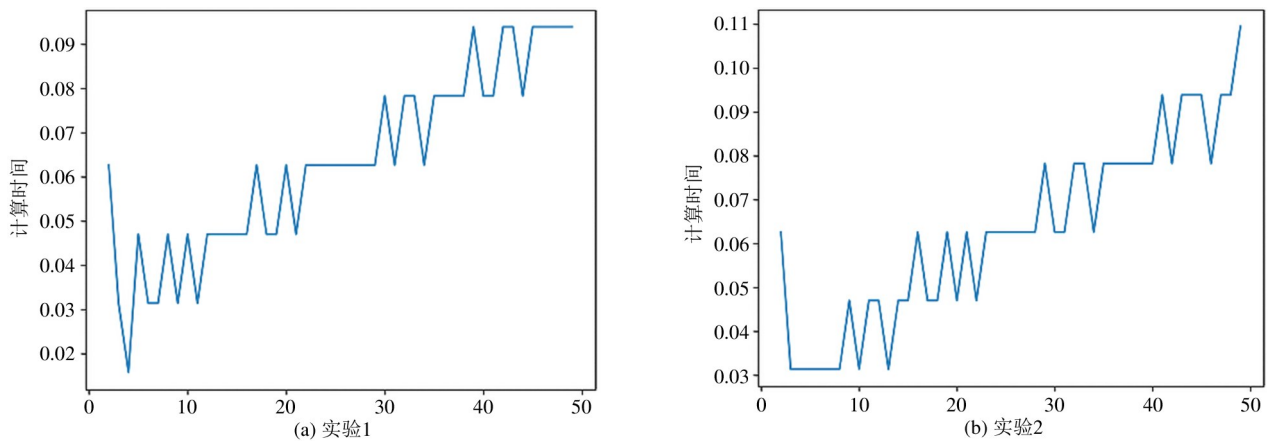


图 7 聚类类簇由 2 增加到 50 计算时间变化趋势图

4.2 实验结果分析

本文实验选择了表 2 中 69 386 个源 IP 的查询日志来验证本文方法的有效性。实验主要分为 2 组,分别对具有 9 维特征的样本进行聚类 and 选择出重要特征的样本进行聚类。

(1) 第 1 组是对样本直接进行聚类形成 20 个簇,经过预处理之后生成初始特征是 3.1 节所梳理

的 9 维特征。采用 K-means 算法对 69 386 个 IP 特征向量进行聚类,选择 $k = 20$, 结果详见表 4。

分析表 4, 针对聚类形成的 20 个簇进行归类合并, 形成查询行为特征相似的递归 DNS 集合; 结合每个簇内递归 DNS 具体查询日志内容, 发现基于查询行为, 可以将递归 DNS 识别为公共递归 DNS、企业级递归 DNS、自服务递归 DNS、探测递归 DNS、攻

击递归 DNS、域名抢注递归 DNS 等 6 类集合,结果如表 5 所示。针对表 5 结果进行集合内验证,结论如下。

1)公共递归 DNS。查询行为聚到类 4、6、14 中

的递归 DNS 识别为公共递归 DNS,为大范围网络(如互联网、ISP 网络)内的用户提供递归解析服务,属于真实递归 DNS。其查询特征是查询数量大(如类 4 中平均查询总量超 96 万次),查询具有明显时

表 4 初始选定的九维特征聚类结果

簇序号	IP 数量	平均查询总量	每小时最大总查询次数	每小时最大查询变化率	IP 端口信息熵	域名种类	域名信息熵	权威域名信息熵	顶级域名信息熵	重复查询次数
1	7294	120	10	4.3	5.5	34.2	4.0	3.4	0.7	2.8
2	20 288	17	1.4	0.0	0.3	1.1	0.1	0.0	0.0	8.0
3	8	3 125 155	133 269	5909.4	15.9	2 835 348	21.3	21.1	1.0	1.1
4	85	965 884	23 649	79.6	15.4	460 949.3	16.7	15.3	1.5	219.3
5	2	616 197	6015	0.9	15.8	4.5	0.2	0.0	0.0	138 669.9
6	649	202 655	5137	150.8	14.9	109 722.5	15.3	14.4	1.4	2.4
7	9927	7	2.1	0.7	2.1	4.4	1.8	1.6	0.9	1.4
8	1	738 195	56 218	54 093	15.9	716 284	19.4	19.3	1.1	1.0
9	3099	14 407	347	44	10.8	3186.9	9.3	8.7	1.4	37.2
10	4	5 041 359	473 290	590.8	16	5 034 163.2	22.3	15.6	0.7	1.0
11	5804	25	4	2	3.5	13.7	3.2	2.9	1.5	1.6
12	34	831 549	88 715	3152.9	15.9	727 252.8	19	18.2	1.0	1.2
13	2	352 453	4682	0.9	15.7	5.5	0.5	0.0	0.0	64 654
14	21	2 768 950	54 541	231.2	15.7	1 266 840.7	18.7	17.1	1.4	2.8
15	3716	2550	98	15	8.5	53.4	1.5	0.7	0.3	165.6
16	1	865 496	77 064	22 606	15.9	849 875	19.7	19.6	0.8	1.0
17	56	1 563 412	10 555	3191.5	14.3	86 038.7	14.6	14.2	1.5	2.5
18	6536	622	27	8.6	7.3	114.5	5.8	5.1	1.4	3.4
19	26	106 558	1271	3.8	14.6	4.7	0.7	0.2	0.0	17 302.7
20	11 833	9	2.3	0.7	2.3	3.8	1.6	1.1	0.0	2.5

表 5 递归 DNS 识别和查询行为主要特征分析

序号	集合	簇类	IP 地址举例
1	公共递归 DNS	类 4、6、14	220.187.246.*/ * 浙江省绍兴市电信;202.101.173.*/ * 浙江省杭州市电信;61.188.7.*/ * 四川省成都市电信;36.25.250.*/ * 浙江省湖州市电信
2	企业级递归 DNS	类 15	59.61.241.*/ * 福建省泉州市 电信;183.11.73.*/ * 广东省深圳龙岗区电信
3	自服务递归 DNS	类 1、2、7、11、18、20	122.225.70.*/ * 浙江省嘉兴平湖市 电信;113.116.71.*/ * 广东省深圳龙华区电信;183.246.66.*/ * 浙江省温州瑞安市 移动;219.145.103.*/ * 陕西咸阳秦都区 电信;61.164.63.*/ * 浙江省杭州市 电信
4	探测递归 DNS	类 5、9、13、19	49.7.30.*/ * 北京电信;116.228.111.*/ * 上海电信;218.205.81.*/ * 浙江省杭州移动;171.15.198.*/ * 河南省郑州电信
5	攻击递归 DNS	类 10	101.226.162.*/ * 上海电信
6	域名抢注递归 DNS	类 3、8、12、16、17	106.11.86.*/ * 上海市阿里云;203.119.171.*/ * 北京市阿里云;140.205.129.*/ * 6 上海市阿里云;47.106.97.*/ * 广东省深圳市阿里云

间周期性,查询量变化相对平缓且少尖峰,查询域名种类多(如类4中查询域名种类平均46万次)且多数为有意义域名对象,域名重复查询次数很少(如类6平均小于3次),IP源端口随机变化,说明查询域名类别相对丰富,顶级域名信息熵较大存在查询变化。

2)企业级递归DNS。查询行为聚到类15中的递归DNS识别为企业级递归DNS,为一定网络范围(如企业局域网)内的用户或特定应用提供递归解析服务,属于真实递归DNS。其查询特征是查询域名总量不大(如类15中总共包括3718个IP对象,平均查询量两千多),每小时查询变化率不大,所查询域名种类数量均不大,域名信息熵都很小,所查询对象的二级域名基本相同且重复查询次数很大(如类15平均超过160次),主要集中在特定域名对象如DNS委托服务、证书服务、软件升级、游戏网站和组织机构等。

3)自服务递归DNS。查询行为聚类到簇1、2、7、11、18、20中的递归DNS识别为自服务递归DNS,仅为自身的网络应用提供递归解析服务,属于真实递归DNS。其查询特征总访问量不大且比较平均(如类18中总共包括6536个IP对象,平均日查询量622次),域查询种类不大(如类18平均低于120次),且绝大多数为正常访问的娱乐新闻类等网站域名,IP端口信息熵较小,域名重复查询次数较小(如类18平均3次左右)。

4)探测递归DNS。查询行为聚类到簇5、9、13、19中的递归DNS识别为探测递归DNS,为特定用途(如搜索引擎爬虫、网络监控等)提供递归解析服务,不属于真实递归DNS。其特征是查询量很大(如类5查询总量超150万),基本为定时查询且无查询变化率或很小,IP端口信息熵正常,所查询域名种类固定(如类9主要固定探测gov.cn、bj.cn、hk.cn、sc.cn、tw.cn等各省几十种二级域名状态),探测重复查询次数最大(类5超13万),其权威域名信息熵和顶级域名信息熵不大。类5(IP:49.7.*.*)也为定时探测,所探测目的域名主要是固定域名解析状态等互联网基础服务(如*.tv.ctdns.cn),类19(IP:10.10.*.*)为定时探测某些重点

域名解析状态是否正确(如cnnic.cn和95538.cn等)。

5)攻击递归DNS。查询行为聚类到簇10中的递归DNS识别为攻击递归DNS,其为某些网络攻击提供递归查询服务或直接发起针对权威DNS的攻击,不属于真实递归DNS。其查询特征是查询量短时间集中且查询量非常大(如类10中总共包括4个IP对象,域名查询总量超500万,每小时超50万),域名查询种类非常大(如类10中域名种类总量也超500万),且绝大多数为具有典型DGA特征域名。

6)域名抢注递归DNS。查询行为聚类到簇3、8、12、16、17中的递归DNS识别为域名抢注递归DNS,为域名行业中掉线域名抢注这一特定行为提供递归解析服务,不属于真实递归DNS。其查询特征域名查询访问时间周期性明显,域名查询总访问量很大且随时间变化很大(如类3总量超300万,每小时最大变化率近13万),域名查询种类非常大(如类3超280万次),绝大多数为个有意义网站域名,且查询NS记录比例很高,域名信息熵和权威域名信息熵很高,域名重复查询次数很小。

(2)对于递归DNS查询日志来说,国家顶级域名系统每天产生的日志量达到百亿级别,直接对全维度的日志进行分析,处理过程比较复杂,分类模型的运算时间将直接影响该方法在国家顶级域名系统查询日志的实际应用。为提高模型的聚类性能,第2组实验主要是给出从3.1节梳理的9维特征中选取关键特征的相关结果。

表6给出了利用无监督特征选择方法从初始9维特征中依次选择2、...、8特征的结果,其中表格中行表示特征初始维度,列表示被选择的特征对应的维度。

从表6和7的综合结果可以看出,选择出5维特征时模型的整体效果比较好,与全维度特征聚类结果相近,且运算时间也具有显著优势。具体选择出来的5维特征包括查询请求总数、端口信息熵、域名信息熵、权威域名与顶级域名信息熵。其中查询总数反映了IP的活跃程度,正常源IP的网络行为通常不会有过高的DNS请求;端口信息熵当DNS发生流量异常时,必定会引起查询源IP端口熵值的

表 6 选择不同维度特征结果

选择的特征维度	1	2	3	4	5	6	7	8	9
2				✓				✓	
3				✓		✓		✓	
4				✓		✓	✓	✓	
5	✓			✓		✓	✓	✓	
6	✓			✓	✓	✓	✓	✓	
7	✓		✓	✓	✓	✓	✓	✓	
8	✓		✓	✓	✓	✓	✓	✓	✓

表 7 不同维度特征聚类结果

选择特征的数量	类内样本数大于 10 000 的簇	类内样本数大于 5000 小于 10 000 的簇	类内样本数大于 1000 小于 5000 的簇	类内样本数大于 500 小于 1000 的簇	类内样本数大于 100 小于 500 的簇	类内样本数大于 50 小于 100 的簇	类内样本数大于 10 小于 50 的簇	类内样本数小于 10 的簇	计算时间/s
二维	0	0	0	0	2	15	2	1	1.00
三维	0	0	0	0	4	13	2	1	1.02
四维	0	0	0	0	3	13	3	1	0.98
五维	5	4	0	0	1	2	4	2	0.92
六维	3	1	1	3	1	6	3	2	0.93
七维	4	2	2	1	1	4	4	2	0.98
八维	6	4	1	0	1	2	5	1	0.97
九维	6	4	1	0	1	2	4	2	0.97

突变;域名信息熵反映查询域名的分散或集中程度,当 DNS 发生流量异常时,必定会引起查询域名熵值的突变;权威域名信息熵与顶级域名信息熵与域名信息熵意义类似。

5 结论

本文从 .CN 国家顶级域名系统面对的实际问题和实际需求出发,基于真实数据设计了一种递归 DNS 行为特征提取和识别方法。首先通过最小化特征的重构误差选择出递归 DNS 查询日志中具有代表性的重要特征;然后利用粗聚类方法确定聚类簇的数量;最后基于递归 DNS 的查询日志通过聚类的方式全面准确识别出公共递归 DNS、企业级递归 DNS、自服务递归 DNS 等真实递归 DNS,以及探测递归 DNS、攻击递归 DNS、域名抢注递归 DNS 等非真实递归 DNS。该研究结果可形成完整准确的真实递归 DNS 清单,可支撑 .CN 国家顶级域名系统的

服务管理和安全防护,也可支撑对递归 DNS 的进一步深入研究。

参考文献

- [1] 中国互联网络信息中心. 第 49 次中国互联网络发展状况统计报告[R]. 北京:中国互联网络信息中心, 2022.
- [2] MAN K Y, ZHOU X A, QIAN Z Y. 2021. DNS cache poisoning attack: resurrections with side channels[C]// Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. New York: Association for Computing Machinery, 2021:3400-3414.
- [3] YEHUDA A, ANAT B, LIOR S. NXNSattack: recursive DNS inefficiencies and vulnerabilities[C]// Proceedings of the 29th USENIX Conference on Security Symposium. Berkeley: USENIX Association, 2020:631-648.
- [4] ZHENG X F, QI A X, LU C Y, et al. Poison over troubled forwarders: a cache poisoning attack targeting DNS forwarding devices[C]// Proceedings of the 29th USENIX Conference on Security Symposium. Berkeley: USENIX Association, 2020:577-593.
- [5] GAO H, YEGNESWARAN V, CHEN Y, et al. An empirical reexamination of global DNS behavior[J]. ACM Sigcomm Computer Communication Review, 2013, 43

- (4):267-278.
- [6] SCHOMP K, CALLAHAN T, RABINOVICH M, et al. On measuring the client-side DNS infrastructure [C] // Proceedings of the 2013 Conference on Internet Measurement Conference. New York: Association for Computing Machinery, 2013:77-90.
- [7] CALLAHAN T, ALLMAN M, RABINOVICH M. On modern DNS behavior and properties [J]. ACM SIGCOMM Computer Communication Review, 2013,43(3):7-15.
- [8] PATRICIA C, CUEVAS R, VALLINA-RODRIGUEZ N, et al. Measuring the global recursive DNS infrastructure: a view from the edge [J]. IEEE Access, 2019, 7: 168020-168028.
- [9] KUHRER M, HUPPERICH T, BUSHART J, et al. Going wild; large-scale classification of open DNS resolvers [C] // Proceedings of the 2015 ACM Conference on Internet Measurement Conference. Tokyo: Association for Computing Machinery, 2015:355-368.
- [10] LIU X M, SUN Y, HUANG C Y, et al. Fast and accurate identification of active recursive domain name servers in high-speed network [C] // Proceedings of the 2016 ACM International on Workshop on Traffic Measurements for Cybersecurity. Xi'an: Association for Computing Machinery, 2016:40-49.
- [11] LIU B J, LU C Y, DUAN H X, et al. Who is answering my queries; understanding and characterizing interception of the DNS resolution path [C] // Proceedings of the 27th USENIX Conference on Security Symposium. Baltimore: USENIX Association, 2018:1113-1128.
- [12] PARK J, JANG R, MOHAISEN M, et al. A large-scale behavioral analysis of the open DNS resolvers on the Internet [J]. IEEE/ACM Transactions on Networking, 2022,30:76-89.
- [13] 张斌, 廖仁杰. 基于关联信息提取的恶意域名检测方法 [J]. 通信学报, 2021,42(10):162-172.
- [14] ZHAUNIAROVICH Y, KHALIL I, YU T, et al. A survey on malicious domains detection through DNS data analysis [J]. ACM Computing Surveys, 2018,51(4):1-36.
- [15] LIU Z, ZENG Y, ZHANG P, et al. An imbalanced malicious domains detection method based on passive DNS traffic analysis [J]. Security and Communication Networks, 2018, 6510381:1-7.
- [16] CHOI H, LEE H. Identifying botnets by capturing group activities in DNS traffic [J]. Computer Networks, 2012, 56(1):20-33.
- [17] SOLTANAGHAEI E, KHARRAZI M. Detection of fast-flux botnets through DNS traffic analysis [J]. Scientia Iranica, 2015,22(6):2389-2400.
- [18] YUCHI X B, WANG X, LEE X D, et al. A new statistical approach to DNS traffic anomaly detection [C] // The 6th International Conference on Advanced Data Mining and Applications. Berlin:Springer-Verlag, 2010:302-313.
- [19] 季成, 李晓东, 袁坚, 等. 基于 k—means 算法的 DNS 查询模式分析 [J]. 清华大学学报 (自然科学版), 2010,50(4):601-604, 608.
- [20] FAN M Y, CHANG X J, ZHANG X Q, et al. Top-k supervise feature selection via ADMM for integer programming [C] // Proceedings of the 26th International Joint on Artificial Intelligence. Melbourne: AAAI Press, 2017: 1646-1653.
- [21] ZHU P F, ZUO W M, ZHANG L, et al. Unsupervised feature selection by regularized self-representation [J]. Pattern Recogn, 2014,48(2):438-446.

A recursive DNS identification method based on top-level domain resolution log

HU Anlei^{* ** ****}, XIE Gaogang^{**** *****}, YUAN Weiguo^{**}, WEI Jinxia^{**** *****}, FU Hao^{**** *****}

(* Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

(** China Internet Network Information Center, Beijing 100190)

(*** Computer Network Information Center, Chinese Academy of Sciences, Beijing 100083)

(**** University of Chinese Academy of Sciences, Beijing 100049)

Abstract

Recursive domain name system (DNS) can be categorized into different types according to the characteristics in terms of the resolution service openness and the purpose of recursive queries. The accurate identification of recursive DNS types has an important impact on the analysis and operation of root, top-level and all levels of authoritative DNS. The accuracy of traditional method based on the character features needs to be further improved. Aiming at the accurate identification of the types of each recursive DNS, this paper first analyzes the query log data from .CN national top-level DNS, and then proposes a recursive DNS type identification method based on the observed behavioral characteristics of recursive query. Specifically, this method distills the full amount of log information from multiple dimensions and selects important features based on unsupervised feature selection, in order to realize accurate clustering of recursive DNS. Experimental results show that this method can identify recursive DNS types efficiently and accurately.

Key words: recursive domain name system (DNS), feature recognition, unsupervised feature selection, clustering algorithm