中图分类号：TN915.08

论文编号：10006ZY1506221

北京航空航天大學

硕士学位论文

恶意域名检测技术研究

作者姓名　　王文博

学科专业　　计算机应用技术

指导教师　　兰雨晴　周渊

培养院系　　计算机学院

Research on malicious domain name detection technology

A Dissertation Submitted for the Degree of Master

Candidate：Wenbo Wang

Supervisor：Yuqing Lan

School of Computer Science and Engineering

Beihang University, Beijing, China

中图分类号：TN915.08
论文编号：10006ZY1506221

硕 士 学 位 论 文

恶意域名检测技术研究

作者姓名	王文博	申请学位级别	工程硕士
指导教师姓名	兰雨晴	职　　称	教授
学科专业	计算机应用技术	研究方向	网络空间安全
学习时间自	2015 年 09 月 01 日	起至	年　月　日止
论文提交日期	年　月　日	论文答辩日期	年　月　日
学位授予单位	北京航空航天大学	学位授予日期	年　月　日

关于学位论文的独创性声明

本人郑重声明：所呈交的论文是本人在指导教师指导下独立进行研究工作所取得的成果，论文中有关资料和数据是实事求是的。尽我所知，除文中已经加以标注和致谢外，本论文不包含其他人已经发表或撰写的研究成果，也不包含本人或他人为获得北京航空航天大学或其它教育机构的学位或学历证书而使用过的材料。与我一同工作的同志对研究所做的任何贡献均已在论文中作出了明确的说明。

若有不实之处，本人愿意承担相关法律责任。

学位论文作者签名：_____ 日期：　　年　　月　　日

学位论文使用授权书

本人完全同意北京航空航天大学有权使用本学位论文（包括但不限于其印刷版和电子版），使用方式包括但不限于：保留学位论文，按规定向国家有关部门（机构）送交学位论文，以学术交流为目的赠送和交换学位论文，允许学位论文被查阅、借阅和复印，将学位论文的全部或部分内容编入有关数据库进行检索，采用影印、缩印或其他复制手段保存学位论文。

保密学位论文在解密后的使用授权同上。

学位论文作者签名：_____ 日期：　年　　月　　日

指导教师签名：_____ 日期：　年　　月　　日

摘　　要

近年来，互联网用户量不断增加，在这大繁荣的网络环境下，各种网络中恶意行为也层出不穷。而域名对于互联网来说是不可或缺的关键元素，同样的很多恶意行为都会涉及到域名，这些域名我们称之为恶意域名。随着中国互联网的蓬勃发展，恶意域名的数量也在逐年增长，其危害也不断增大。针对恶意域名的检测算法主要有信誉系统、逆向工程和机器学习，这些方法或在时效性上表现不足，或技术实现难度较大，或在准确性上表现较差。被动 DNS 记录中包含了用户所有的对域名的查询，而基于被动 DNS 的网络安全相关研究尚不完善，国内外的网络环境也不尽相同，这些都造成了对现有国内被动 DNS 数据的分析不足，这些数据存在需要被进一步挖掘的需求。

本文旨在探讨将被动 DNS 数据更好的利用于恶意域名的相关研究当中的可行性，对被动 DNS 数据进行了细致的分析，利用这些数据的特点结合现有的技术方案，确定了基于被动 DNS 对恶意域名研究的需求。全面分析了现有技术方案的优势和劣势，借鉴信誉系统，提出了快速提取恶意域名相关流量的算法方案。综合国内外对恶意域名检测的相关研究，使用机器学习技术，提出了分别针对 DGA 域名和色情域名检测的算法方案。

最后探讨了国内网络环境下恶意域名检测技术应用模型，设计并实现了恶意域名相关流量提取方案和恶意域名检测算法的实验原型程序。对算法的有效性和正确性进行了验证。本研究提出的流量提取方案可以有效的在保证召回率的前提下，高效的提取出恶意域名相关流量；本研究提出的恶意域名检测算法具有很高的准确性，同时在时间维度上也保证了一定的时效性，在国内网络环境下具有很强的现实意义。

关键字：恶意域名；被动 DNS；DGA；色情域名

Abstract

In recent years, the number of Internet users has been constantly increasing. And in this prosperous network environment, malicious behaviors in various networks have been also emerging. The domain name is an essential elements for the Internet. Similarly, a lot of malicious behaviors involve domain names. These domains are called malicious domains. With the vigorous development of China's Internet, the number of malicious domain names is increasing year by year, and its harm is also increasing. The detection schemes for malicious domains mainly include reputation system, reverse engineering and machine learning. These methods are either insufficient in timeliness, or difficult to implement technology, or poor in accuracy. Passive DNS records contain all the users' queries about domain names.However, the research on network security based on passive DNS is still not perfect, and the network environment in China and abroad is also different. All of these make the analysis of the existing domestic passive DNS data insufficient, and there is a need for these data to be further mined.

The purpose of this paper is to explore the feasibility of using passive DNS data in the related research of malicious domain names. In this paper, we analyze the passive DNS data carefully and use the characteristics of these data combined with the existing technical solutions to determine the requirements of passive DNS based on the research of malicious domain names. We have comprehensively analyzed the advantages and disadvantages of the existing technical solutions and made reference to the reputation system to propose a scheme for quickly extracting the traffic related to the malicious domain names. Based on the related researches on malicious domain name detection in China and abroad, the paper presents the technical solutions for DGA domain name and pornographic domain name detection respectively using machine learning technology.

Finally, the paper discusses the application model of malicious domain name detection technology under the domestic network environment. the prototype program of the scheme is designed and implemented, and the correctness and efficiency of the scheme are verified. The traffic extraction scheme proposed in this study can efficiently extract the relevant traffic of malicious domain name under the premise of guaranteeing the recall rate. Malicious domains detection scheme proposed in this study has high accuracy, but also to ensure a certain

timeliness in the time dimension.This study has strong practical significance in the domestic network environment.

Key words: Malicious domain name; passive DNS; DGA; pornography domain name

目　　录

图　　目

表　目

第一章 绪论

1.1 研究背景

1.1.1 恶意域名数量巨大

IP 地址是由 IP 协议提供的数字型统一地址标识，作为一种逻辑地址来定义一台设备在网络之中的位置，网络设备逐渐增多 IP 地址的记忆困难显现出来，保罗·莫卡派乔斯 (Paul Mockapetris) 在 1983 年的第 882 和在南加州大学里资讯科学研究院所提出的 883 号因特网标准草案中提出 DNS 的架构，提议将其改进为分布式和动态的数据库域名系统，也就是我们今天所用的域名系统的雏形。从 1985 年 Symbolics 公司注册的第一个 com 域名到如今仅中国域名总数增长至 3698 万个[1]，域名产业飞速发展，随之而来的安全问题也越来越多的暴露出来。高级持续性威胁常态化，移动互联网黑色产业链已经成熟，巨大的利益促使大量人进行相关活动。以僵尸网络控制端（通常使用域名来进行联系）为例，2012 年木马和僵尸网络控制端数量高达 36 万余个，随着检测技术和安全意识的提高，至 2014 年数量有明显降低，仍有 10 万个左右的僵尸网络控制端，并且该数量至今相对稳定在这一水平上[3]。

1.1.2 恶意域名危害增大

我们将当前存在恶意行为或者被恶意使用的域名都视为恶意域名，这其中就包含了 DGA、DNS 放大攻击、钓鱼域名等等。正如 1.1.1 节中所述，尽管随着互联网安全监管的加强，恶意域名数量仍然庞大，并且恶意域名相关的技术在与检测技术的对抗之中不断进步，造成的安全威胁更加巨大[42]。2017 年活跃的 WannaCry[43]、Mirai 等恶意程序就是最好的佐证。

WannaCry 是一种勒索木马，感染主机后利用微软系统漏洞 EternalBlue（永恒之蓝），获取系统权限，将硬盘中的文件加密进行勒索。2017 年 5 月第一波爆发，感染了全球一百多个国家和地区超过三十万台主机，造成经济损失达百亿美元。根据对样本的观察以及之后研究人员对 WannaCry 源码的分析，我们了解到 WannaCry 是否在感染主机后进行文件的加密处理与域名“*www.iuqerfsodp9ifjaposdfjhgosurijfaewrwergwea.com*”（该域名为第一次攻击时使用的域名，之后使用其他域名）休戚相关，WannaCry 会首先尝试访问该域名，如果无法被正常访问则会继续产生恶意行为。这和以往的恶意软件很不一样，大多数恶意程序会通过域名或其他手段与攻击者产生连接，根据攻击者指令产生攻击行

为，而 WannaCry 恰恰相反，域名的成功连接作为攻击停止的信号来使用。这里所提到的域名就属于恶意域名的范畴当中。

```
static void resolve_cnc_addr(void)
{
    struct resolv_entries *entries;                    获取C&C域名
    entries = resolv_lookip(table_retrieve_val(TABLE_CNC_DOMAIN, NULL));
    table_lock_val(TABLE_CNC_DOMAIN);
    if (entries == NULL)
    {
        #if DEBUG
        printf("[main] Failed to resolv CNC adress\n")
        #endif
        return;
    }
    srv_addr.sin_addr.s_addr = entries->addrs[rand_next() % entries->addr_len]
    resolv_entries_free(entries)

    table_unlock_val(TABLE_CNC_PORT);                  获取C&C端口
    srv_ADDR.sin_port = *((port_t *)table_retrive_val(TABLE_CNC_PORT, NULL));
    table_lock_val(TABLE_CNC_PORT);

    printf("[main] Resolved domain\n")
}
```

图 1 Mirai 连接 C&C 服务器代码片断

Mirai 是一类针对物联网设备的僵尸网络木马，可以由一个感染设备对其他可感染设备进行 SYN 扫描与探测，扩散能力极强。Mirai 所感染的肉鸡通过域名与攻击者所控制的 C&C 服务器产生连接。如图 1 所示，最初的 Mirai 使用的域名还不是自动生成的，而是简单存在一个列表“*TABLE_CNC_DOMAIN*”中。被感染设备会遍历这个域名列表，攻击者只需要选择其中一个域名注册使用，即可控制所有的被感染设备。在 2016 年 10 月美国爆发的大规模网络瘫痪事件中，黑客正是利用 Mirai 控制大量的物联网设备针对 Dyn 域名服务器发起的 DDOS 攻击。如今物联网设备或者家庭智能设备数量不断的增长，黑客也将注意力更多的放在了这些设备上。在之后出现的诸多 Mirai 变种中，使用了 DGA（域名生成算法）代替了原本的固定域名列表，那么使用 DGA 有什么好处呢。

传统僵尸网络使用固定 IP 或者域名与 C&C 服务器建立连接，隐蔽性很差，极易被发现。后来出现的例如 Nugache[36]，Storm[37]，Waledac[38]，Zeus[39]等基于 P2P 的僵尸网络虽然具有较好的鲁棒性与稳定性，但实现难度和维护成本较高。如今大部分活跃的僵尸网络都采用了 DGA，依赖于集中的 C&C 服务器，相比于前两者具有简单易行，兼顾稳定性与隐蔽性的优点。诸如 Locky[7]、GameOver、Rovnix 等新型木马均使用了 DGA 来获取与 C&C 服务器连接的域名。DGA 作为一种随机算法，输出为域名，我们将算法输入称为种子（例如数字常量、当前时间、Tiwwter 动态等），按照种子产生类型以及算

法类型，可以对 DGA 进行分类。如果这个种子与时间有关，称之为 TD（time-dependent，时间相关），反之，称为 TI（time-independent，时间无关），如果种子可以估计（例如日期），称为 D（Determinism，可估计），反之（例如欧洲央行每天外汇参考利率），称为 N（Non-determinism，不可估计）。域名产生模式分别有 A（Arithmetic，算法类）、W（Wordlist，单词表类）、H（Hashing，哈希类）、P（Permutation，置换类）四种。进行排列组合理论上有 16 类 DGA，实际上只出现了 TDD-A, TID-A,TDD-W , TDD-H, TDN-A , TID-P 这六种类型[5]。可见 DGA 算法千变万化，是黑客的一大利器。

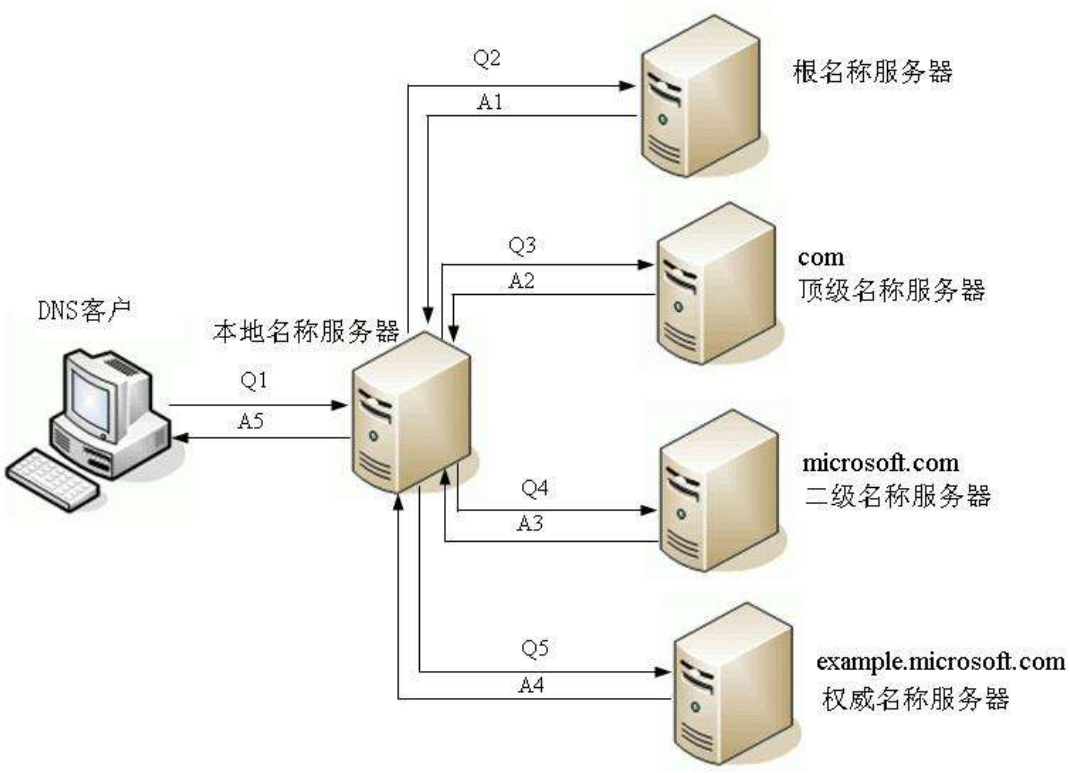

图 2 DNS 递归查询过程

上文介绍了僵尸网络在建立和连接上如何使用了域名，攻击者利用这些僵尸网络可以很轻松的发起 DDos 攻击（Distributed denial-of-service，分布式拒绝服务攻击）[32]。在 2015 年，DDoS 攻击峰值流量不断上升，甚至出现了 1T 超大流量攻击事件，全年的攻击总流量接近 28 万 Tbytes[1] ，常用手法包含有 UDP 攻击[44]、SYN 攻击[45]、ICMP 包攻击[46]等。在各种类型的 DDos 攻击中，有两种涉及到了恶意域名[40]，一是利用 DNS 解析的反射放大攻击，在利用流量来实施的 DDoS 攻击中，只要有可以利用来进行放大数据量的协议，都有可能被恶意使用。正常的 DNS 查询是从原 IP 地址向 DNS 服务器（递归或者权威），大小相对固定（70 字节左右），也就是攻击成本相对固定，DNS 返回数据由于请求域名和类型的不同，数据大小从几百到几千字节（查询“*www.baidu.com.*”

的返回数据为 302 字节），放大了四倍以上。由于利用 A 记录或者 Cname 记录放大效果并不十分明显，Akamai 研究人员在 2014 年发现了利用 TXT 记录来进行的 DNS 放大攻击，攻击者使用名为 DNS Flooder 的工具，从 guessinfosys.com 获得 TXT 记录，攻击峰值高达 4.3Gbps[6]。二是随机子域名攻击，这是一种专门针对域名服务器的 DDos 攻击。例如攻击者的利用的域名为“*example.cn*”，那么攻击者会利用僵尸网络中的主机产生一系列无意义的域名头部，与目标域名组合之后形成诸如“*iygyaid.example.cn*”，“*aduh3uh.example.cn*”之类的域名，每一台主机都会对这些域名进行 DNS 请求，如图 2 所示，由于二级域名都是合法的，导致在域名的递归查询过程中，这些 DNS 都会进行相同的二级域名服务器直到权威域名服务器，在僵尸网络足够庞大的情况下，这些递归 DNS 服务器就会因为处理这些 NXDomain（Non-exist Domain，不存在域名）而陷入瘫痪。

恶意域名中钓鱼域名也是不可忽视的一部分，钓鱼网站（phishing page）一直以来都是一个难以规避的网络问题，同时他也是一种社会工程学的攻击方式[41]。通常攻击者对恶意页面进行伪装，使得页面看起来像是银行网站或是其他可以获取关键隐私信息的网站，受害者往往由于受到邮件或者短信的欺骗，通过其中的链接导向到钓鱼网站，或者由于误植域名导向到钓鱼网站。而受害者对于网站的真伪缺少足够的判断和辨识的能力，导致银行账户密码等隐私信息被攻击者窃取。还有一个需要提及的是色情域名，由于文化以及法律法规的差异，各国对色情域名的定义大相径庭，在这里我们将他视为一类恶意域名来研究。由于网络的普及以及缺乏对网络内容的合理分级，导致网络中大量非法的负能量的内容可以轻易的被不适宜的人群所接触，尤其是青少年人群对一些内容的辨识度不够，很容易误入歧途，造成不良的社会影响。除了以上提及的恶意域名种类之外，还包含有垃圾邮件域名、域名阴影的域名等等，每一类域名都不断延伸交错形成一个个完整的灰色或者黑色产业链，对互联网环境造成了巨大威胁。

1.2 问题的提出

各类恶意域名带给互联网的安全威胁巨大，那么基于被动 DNS 的行之有效的方法来检测和识别出这些恶意域名就显得尤为重要。现阶段针对恶意域名的检测技术主要面向三个方向，一是域名信誉系统（DRS，Domain Reputation System）[47]，简要来说 DRS 就是一个给域名打分的系统，域名的良性与恶性都由分数来决定，我们熟知的黑名单也可以看作是一种最简单的信誉系统，黑名单上的域名的分数统一为 0 分。建立黑名单是

一种切实有效的防御手段，国内外很多安全厂商和团队都长期维护着黑名单，可以说这是误报率最低同时也是检测速度最快的一种防御措施。但黑名单也有着不可避免的缺点，将数据信息进行人工或者程序的审核，确认为恶意的数据信息加入黑名单，因此维护和更新的成本相对较高，而且黑名单的建立必然在恶意行为发生之后，相对滞后。

二是逆向工程[48]，逆向工程对于恶意软件防御方面而言是一种很常规的手段，Khaled Yakdan 等人的论文中正是对现在活跃的大部分 DGA 使用了逆向工程[5]，在实际使用中可以做到 FP 值为 0，他们历时数年的细致工作完美体现了逆向工程精准的特质，但是逆向工程的缺点也暴露无遗，那就是太耗费人力和时间，一旦出现可能是新的 DGA 算法，就需要专业的逆向工程师来处理、验证，这对于如今层出不穷的新型木马而言是远远不够的。

三是机器学习，机器学习现如今是一个很流行的手段，深度学习、人工智能的发展，为机器学习创造了更多的可能性。如今信息安全领域仍然是一个富数据，穷分析的领域。2016 乌云白帽大会上，phunter 做了报告《What can you get from 100 billion DNS queries, each day, in real time?》，可以想见当我们有 1000 亿条实时数据，如果仍然使用蜜罐、逆向工程，周期长度以及工程量都是无法承受的。如何做到又快又精确地对给定域名进行分析，那么机器学习就是不二之选。特征的选择体现了人类的经验，而把这些量化特征交给机器，相当于机器利用人类的经验来完成这些重复性的工作。

有幸在研究生学习期间在国家互联网应急中心实习，获得了中心提供的 PDNS（Passive DNS，被动 DNS）数据，该数据为广东省电信部门的 DNS 解析数据，数据具体详情见 3.1 节。由上述检测技术可见无论哪一种方法都存在着一定缺陷，同时由于国内外的网络环境不尽相同，针对国内被动 DNS 的挖掘还远远不足，因此本论文希望能够探讨如何利用 PDNS 数据的高效快速并且尽可能准确的检测出相关的恶意域名。综上所述，本文的主要内容包含以下几点：

1） 提出了快速提取恶意域名相关流量的方案；

2） 设计了完善的特征分析方案；

3） 提出了基于词向量空间的色情域名检测算法；

4） 提出了多维度的 DGA 域名检测算法；

5） 从样本和时间的维度对算法进行了验证，验证了算法的有效性和准确性。

1.3 论文的主要内容

本文对于基于被动 DNS 如何解决恶意域名相关的安全问题，并重点针对 1.2 节中提出的具体问题，首先探讨了现行的恶意域名检测技术；然后引入被动 DNS，介绍了我们可以用被动 DNS 做些什么；随后提出了一种快速提取恶意流量的方法，该方法可以为进一步的检测提供便利；之后对域名的访问特征和字符特征进行了细致的分析；最后对于 DGA 域名和色情域名给出了完整的检测算法和实现。

1.4 论文的组织结构

第一章：绪论，简单介绍了本文的研究背景和相关概念，针对恶意域名巨大的数量以及危害提出了提出了本文研究的问题。

第二章：介绍了恶意域名检测方法的国内外研究现状，对现有方案进行分类，分别简述了其特色以及优缺点，并对相关技术应用情况进行了介绍。

第三章：介绍了利用被动 DNS 所做的部分研究，包含被动 DNS 的相关介绍、特征分析方案、流量提取方案以及恶意域名分类算法。

第四章：完整的将检测模型原型系统设计和实现进行了展示。

第五章：对实验中涉及到的样本、特征、参数以及最终检测的分析结果进行了详尽的展示与解释。

总结与展望：对本文进行总结，并展望了下一步可以作为补充的研究工作。

第二章 国内外研究现状分析

本章主要对国内外学者在恶意域名检测技术的研究现状进行介绍和分析。对于恶意域名检测技术的研究主要分为逆向工程、信誉系统、机器学习三类。逆向工程具有误报率低、召回率高的优点，但是耗时耗力、对攻击反应不及时；信誉系统具有检测速度快、误报率低的优点，但是召回率较高，更新滞后；机器学习方法具有高效、便捷、实时性高的优点，相对前两者精度略有差距。在本章中我们主要关注信誉系统与机器学习。

2.1 信誉系统相关研究

佐治亚理工学院的 M.Antonakakis、P.Roberto、W.Lee 等人是最早一批深入研究 PDNS 数据的研究人员，他们先后对 DNS 缓存投毒、域名信誉系统、DGA、僵尸网络等方向都做出了很高的贡献。在 2010 年建立的 Notos[8]，是一个动态的综合性的信誉系统，首次提出了针对域名的信誉系统，利用被动 DNS 数据来输出信誉分数，而于此之前主要针对 IP 的信誉系统完全不同。Notos 使用了网络和地区的特征，通过配置信息、使用情况和域名管理情况等，能够学习良性和恶性的域名分别是如何工作的，并对每一个新域名计算一个信誉得分。对于一个域名，如果它与恶意活动(例如僵尸网络、垃圾邮件、恶意软件传播等等)有关，就给他赋予一个低的信誉值，图 3 是 Notos 系统的整体流程。

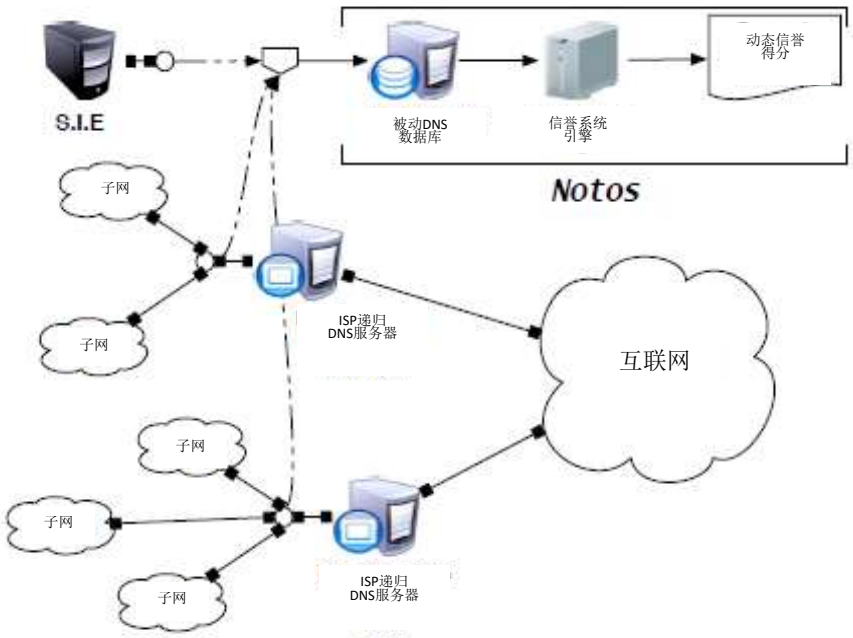

图 3 Notos 系统流程图

首先利用被动 DNS 数据构建三组特征：

1. 基于网络的特征：第一组的统计特征是从 RHIPs 的集合中提取的。计算如下特征：与 d 历史相关的 IP 地址的总数量、它地理位置的多样性、它们所在的不同的自治系统的数量等等。

2. 基于域的特征：第二组特征是从 RHDNs 集合中提取的。计算在 RHDNs 里域名的平均长度、不同 TLD 的个数、不同字符出现的频率等等。

3. 基于证据的特征：最后一组特征包括与 d 相关的不同恶意软件样本的数量、与 d 曾经指向的 IP 相关的恶意软件样本的数量等等。

Notos 的信誉引擎有两个运行模式：离线训练模式和在线分类模式。在离线模式下，Notos 使用在知识库中收集到的数据来训练信誉引擎，知识库即一组已知的恶意和合法的域名及其相关的 IP 地址。在之后的在线模式中，对于每一个新域名 d，Notos 询问信誉引擎从而为 d 计算一个信誉值。动态信誉评分系统的基础是从成功解析的域名 A 记录中获得的历史或“被动的”信息。作者使用了来自两个 ISP 的采集节点的 DNS 流量，一个位于美国东海岸(亚特兰大)，另一个位于美国西海岸(圣荷西)。另外他们也汇总了 SIE 项目中不同网络的 DNS 流量。数据库收集到了从 2009 年 7 月 19 号至 2009 年 9 月 24 号这 68 天的解析记录共 27377461 条。结果 Notos 可以以 96.8%的精度和 0.38%的误报率发现恶意域名，并且它比黑名单方法更快。

2.2 利用机器学习的恶意域名检测

2011 年 Manos Antonakakis 等人构建了一个叫做 Kopis 的恶意域名监测系统[9]，可以通过对 DNS 查询解析模式的分析得到恶意域名，与 EXPOSURE 和 Notos 这些依赖本地递归 DNS 服务器的系统相比，Kopis 使用的是上级 DNS 数据。在此基础上他们在 2012 年又提出了一个新颖的检测系统 Pleiades[2]，首次提出了使用了不存在域名（NXDOMAIN）的资源记录。NXDOMAIN 是指像域名服务器提出解析请求，但是无法解析得到对应 IP。对于使用 DGA 的被控主机大多数查询都会返回一个 NXDomain，而同一僵尸网络控制下(使用相同的 DGA 算法)的被控主机会产生有类似特征的 NXDomain 的流量。Pleiades 聚类和分类算法的结合，充分利用了机器学习的优势，将有着相似字符特征和相似访问特征的域名聚集在一起。分类算法用来将这些生成好的聚簇分配到已知的 DGA 模型中去。如果一个聚簇不能被分到已知的 DGA 中去，那么意味着可能出现了新的 DGA 变种或 DGA 家族。

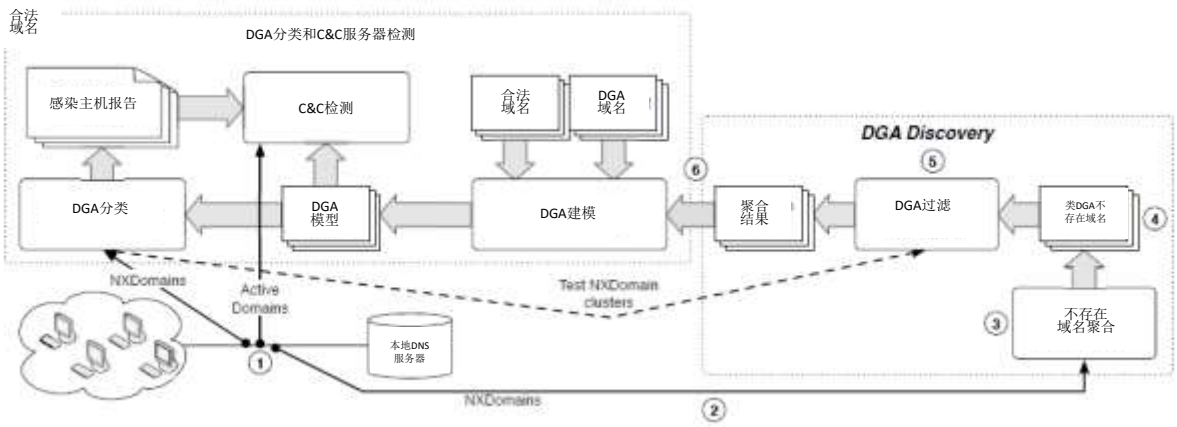

图 4　Pleiades 系统整体流程

图 4 是 Pleiades 的整体流程。DGA 发现模块分析了未成功的 DNS 解析流量，它部署在 DNS 服务器的下游。在一定时间段内网络产生的所有 NXDomains 都会被收集在内。接着，收集到的 NXDomains 会根据以下两个相似准则分别被聚类：(1)域名字符串所具有的相似的统计特征(例如相似的长度、随机性和有关字符的频率分布等)；(2)域名被一批相同或相近的 IP 地址查询。对 NXDomain 聚类的主要目的是将那些使用同一种域名生成算法的主机和域名聚类出来。由于这个聚类过程是无监督的，所以聚类出来的某些 NXDomain 簇里可能包含那些偶然错误的域名(例如由于拼写或者配置错误导致的 NXDomains)。因此对这些数据进行过滤显得极为必要。作者使用有监督的 DGA 分类器去修整这些聚类簇，修正的对象包括已经发现并建模的 DGA 产生的域名和那些与合法的域名相近的域名。DGA 发现模块最终的输出是 NXDomains 的聚簇集，其中的每一个集合都代表了已知或者未知 DGA 产生的域名。每当发现一个新的 DGA，就使用监督学习方法为这样的域名建立一个模型，该模型用来描述新 DGA 产生的域名“长什么样”。为此建立了两种不同的统计模型：(1)一个为肉鸡 产生的一组 NXDomains 赋予 DGA 标签(如 DGA-Conficker.C)的多类统计分类器；(2)基于隐马尔科夫模型判断 查询过的可能由一个 DGA 产生的活跃的单一 C&C 域名的类型。DGA 分类模块的工作过程如下。与 DGA 发现模块相似，作者监测 DNS 递归服务器中由每个主机产生的 NXDomains 流量。给定一个主机产生的 NXDomains 的子集，然后提取与其字符串有关的一系列统计特征。接着将这些特征作为特征向量传入分类器函数，分类器会输出这些 NXDomains 是否由一个已知的 DGA 产生。如果一个主机产生的 NXDomains 被贴上了相关的 DGA 标签，那么这个主机也极有可能被相关的僵尸网络所控制。一旦获得了那些感染主机的列表，就可以进行更深一步的检测。之前所有步骤都是围绕无效域名 NXDomains 展开的，下

一步则把注意力集中到这些感染主机访问的活跃域名上。其目标是确定哪些由 DGA 算法生成的域名最后解析成了 IP 地址，即识别僵尸网络的 C&C 服务器。为了实现这一目标，作者将感染主机访问过的所有可疑活跃域名都收集起来。接着对这些域名输入之前训练好的隐马尔科夫模型中，由它来决定单一活跃域名是由已知 DGA 产生的或是未知 DGA 产生的。使用隐马尔科夫模型而不是分类器做判断的原因是需要对单一域名进行检测。DGA 分类器不适合单一域名的检测，因为它根据一个感染主机产生的一组无效域名进而判断这组无效域名是哪种 DGA 产生的。

作者使用 Bobax, Sinowal, Conficker-A, Conficker-B,Conficker-C 和 Murofet 产生的 NXDomains 来引导分类器。在两种不同的模式下测试分类器的效果优劣。第一种模式使用 Conficker 的一个“超类”，它由 Conficker-A, Conficker-B 和 Conficker-C 的相同数目的样本组成。另一种模式则将 Conficker 的每个变种作为不同的类别看待。从每个类的域名中，作者随机选取了大小为 α 的 3000 个集合。其使用的 α 值有 2、5、10 和 30。这是建立不同的训练集的过程，目的就是通过实际证明哪个 α 值可以给出 DGA 模型之间的最佳的分割。

表 1 Pleiades 中不同 α 下的分类效果

	α = 5 NXDomains			α = 10 NXDomains		
Class	TP_{rate}	FP_{rate}	AUC	TP_{rate}	FP_{rate}	AUC
Bobax	95	0.4	97	99	0	99
Conficker	98	1.4	98	99	0.1	99
Sinowal	99	0.1	98	100	0	100
Murofet	98	0.7	98	99	0.2	99
Benign	96	0.7	97	99	0.1	99

Pleiades 从 2010 年 11 月的第一天就开始对 NXDomains 流量进行聚类。作者使用已知的 DGA 作为正例和一组 Alexa 域名作为反例来引导 DGA 建模过程。通过检查 NXDomains 与从恶意程序库中提取到的 NXDomains 的重叠关系来发现恶意程序家族。另外他们也在威胁情报公司的帮助下手动地检查了聚簇，检测出每个 DGA 变种有着平均 32 个感染主机，这些主机横跨全州的 ISP 网络。通过十几个月的实验，我们证明了 Pleiades 可以达到一个很高的精确度。此外，Pleiades 部署在大型 ISP 网络下的这 15 个月来，它可以发现 6 个属于已知恶意程序家族的 DGA 和 6 个之前从未报出过的 DGA。

当然 Pleiades 也存在一定的局限。例如，一旦发现一个新的 DGA，Pleiades 可以很准确地为它建立统计模型，它可以知道这种 DGA 产生的域名“长什么样”，但它不能通过学习重现它的域名生成算法。因此，Pleiades 会产生一定程度的假正和假负。但是上表的数据表明，Pleiades 可以建立一个很准确的 DGA 分类器模型，当 α =10 时假正和假负都很低。C&C 检测模块可以判断单一活跃域名的 DGA 种类，在大多数情况下表现很好。但是有一些情况下基于 HMM 的分类有部分不尽如人意的地方。作者认为这样的原因是 HMM 只考虑了域名的单一字符序列。总体来说，Pleiades 聚类和分类算法的结合，充分利用了机器学习的优势，将有着相似字符特征和相似访问特征的域名聚集在一起，分类算法用来将这些生成好的聚簇分配到已知的 DGA 模型中去，是这个领域非常重要的一篇文章。

除开佐治亚理工学院的这些人，其他地区的研究者和组织也对这个领域做出了极大的贡献。Perdisci 等提出 FluxBuster 系统, 这是一个专门针对速变域名检测的系统，将域名的 IP 变迁情况引入特征集, 共 9 组 13 个特征, 采用聚类算法来识别速变域名[11]。除了针对某一种域名的检测系统，更多的是具有广泛适用性的系统。

2011 年 L.Bilge 等人建立了 EXPOSURE 系统[12]，一个可以检测多类恶意域名的系统。该系统从 DNS 数据中分别基于时间、基于 DNS 响应、基于生存时间值（TTL）、基于域名提取了这 4 类共计 15 种特征，使用 J48 决策树训练分类器。相比于 Notos，Exposure 有着更完善的特征选取，也弥补了无法检测一个 IP 地址只被恶意使用一次的恶意域名。相比于之前文章都只能在 DNS 数据中找到特定种类的恶意域名，EXPOSURE 对恶意域名的检测更加广泛。

2013 年 Terry Nelms，Roberto Perdisci 等人构建的 ExecScent[12]是一个旨在从真实的企业网络流量中挖掘新的、从未出现过的 C&C 域名的系统，同时 ExecScent 也是第一个使用自适应 C&C 流量模型的系统。ExecScent 从已知的 C&C 通信样本中自动地学习控制协议模板(CPT)，并且这些 CPT 会匹配它们所部署的网络流量。ExecScent 构建的这种自适应模板从部署模板的网络流量中学习，这种“自适应”的方法使得 ExecScent 在保持一个很高检出率的同时极大地降低了误报率。ExecScent 自动地寻找不同恶意样本的 C&C 协议间的共同的特点，接着将这些共性编码进 CPT 集合中。每个模板都标有恶意程序家族的名称或者与该 C&C 流量相关的犯罪组织(如果有的话)。一旦将 CPT 部署在网络的边缘，任何与该模板匹配的 HTTP(S)流量就会被认为是 C&C 流量。与该流量

相关的域名就会被标记为 C&C 域名，并且会归于恶意程序家族或者与该 CPT 关联的组织。图 5 展示了 ExecScent 生成和标记 CPT 的整体过程。

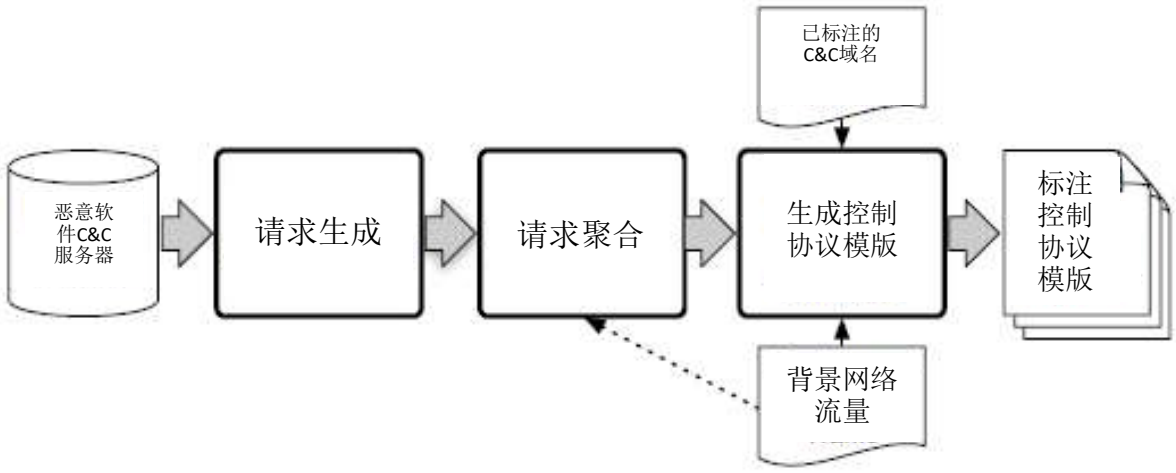

图 5 ExexScent 系统概述

给定一个大的恶意软件生成的网络痕迹，首先重构每个恶意样本的 HTTP 请求。接着，进行一个“请求泛化”的过程，这里将请求的一些参数(如 URL 参数值)替换为它们的类型和长度。当恶意程序的请求被聚类后，我们进行模板学习过程，在这里我们推导出 CPT。本质上讲，一个 CPT 高度概括了聚在一起的 HTTP 请求，并且记录了一系列重要的特征如 URL 结构、请求头的集合、每个恶意软件访问过的 IP 地址等等。此外，这些模板与恶意程序家族标签一一关联。在真正部署模板前，作者先把 CPT 放在应该部署到的网络流量中观察一段时间。特别是对于模板的每一个组件(如泛化的 URL 路径、user-agent 字符串、请求头集合等等)，计算各个组件在部署网络中出现的频率。在背景流量中“活跃”的那些组件会在该网络中得到一个低的“匹配信任”。另一方面，在流量中那些不经常(甚至是从来没)出现的组件会得到较高的信任。部署后，如果一个 HTTP 请求被 CPT 以很高的相似性和特异性所匹配，那么它就会被标记为 C&C 请求。这就是说，一个请求与 CPT 描述得很接近并且匹配的 CPT 的组件在特定的部署网络中有着很高的特异性(即不经常出现)。

2016 年 B.Rahbarinia 创新性的提出了一种基于行为的系统 Segugio[14]，Segugio 在大的 ISP 网络中通过追踪被恶意软件感染主机的 DNS 请求行为来高效的发现新增的 malware-control 域名。相比于 Notos 和 EXPOSURE，前两者建立的都是 domain-IP 映射关系模型（使用域名字符串的特征、域名承载的恶意内容等信息）而没有利用本地 DNS 服务器下游的主机请求行为，Segugio 通过监测 ISP 网络用户的 DNS 请求行为，重在精

确的追踪新增的“malware-only”域名。相比于 Kopis，他的做法和本文有相似性（Kopis 用请求者散度、请求者画像等信息），但 Kopis 利用权威或 TLD 服务器的数据，这种数据难以获得（需要与大的 DNS 区域运营商紧密合作）。Segugio 不用关心顶级域名，通过监测本地 ISP 流量（在 ISP 使用者和他们的本地 DNS 解析器之间的 DNS 流量）。因此 Segugio 可以依靠 ISP 网络权限独立开发，不需要与外部的 DNS 运营商合作。

国家互联网计算机网络应急技术处理协调中心张雪松等人提出了对算法生成恶意域名的实时检测[15]。图 6 是该算法的运行流图，有以下几个重要组成模块：数据收集及初步预处理、已知 DGA 域名的解析及访问结构构建、对新添加域名与已知域名的关联分析、未知域名的跟踪分析。该算法主要的关注点在新出现的域名上，对于从递归 DNS 服务器上收集得到的数据，与历史上出现的域名相比对，找到不曾出现过的域名。之后的检测都围绕着这些域名展开，通过关联分析以及对这部分域名的查询情况跟踪分析，将这些域名与域名生成算法的不同家族建立联系，并进行分组。尝试找到其中的可疑域名分组，进一步通过一系列的分析，包括访问特征、生存特征等，最终确定其中为域名生成算法产生的恶意域名。

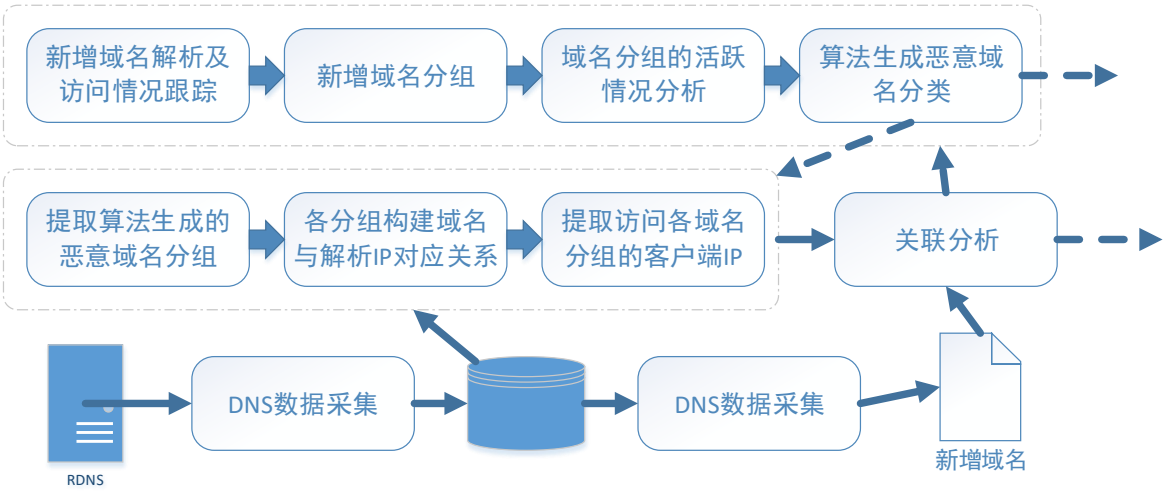

图 6　域名实时检测流程图

除了这些机构，一些高校对此也有所研究，东南大学张维维等人针对黑名单在维护和更新上存在开销大和及时性差，且攻击者常常使用算法自动生成大量的随机域名来躲避检测的不足以及实时检测开销过大的缺陷，设计了轻量级的检测算法来快速锁定监测目标，以便有针对性地使用更为复杂和更为准确的检测算法[17]。轻量级算法需要在有限的系统资源和计算时间内，尽可能多地检测出可疑域名。作者将词素这一概念引入到恶

意域名的检测当中，利用其构建出诸多有价值的特征，并通过实验证明了方案的准确性，同时具有很好的空间复杂度与时间复杂度。

2.3 恶意域名特征研究

2.3.1 针对 DGA 的相关研究

D.Plohmann，F.Fkie 等人针对 DGA 做了大量细致的工作[5]，他们对 43 种 DGA 恶意软件家族和变种进行一个综合性的研究，针对 DGA 提出一个分类学方法，并用它对所研究的 DGA 进行分类与比较。并重现了这些算法，预先计算所有可能的 AGD，覆盖了大部分已知的活跃 DGA，以过去八年总计一千八百万 DGA 域名的注册状态，来证实预计算得到的域名确实是是可靠的。对于 botmaster 的域名注册策略也提出了见解。

DGA 作为一种随机算法，输出为域名，作者将算法输入称为种子（例如数字常量、当前时间、Tiwwter 动态等），按照种子产生类型以及算法类型，可以对 DGA 进行分类。如果这个种子与时间有关，称之为 TD（time-dependent，时间相关），反之，称为 TI（time-independent，时间无关），如果种子可以估计（例如日期），称为 D（Determinism，可估计），反之（例如欧洲央行每天外汇参考利率），称为 N（Non-determinism，不可估计）。域名产生模式分别有 A（Arithmetic，算法类）、W（Wordlist，单词表类）、H（Hashing，哈希类）、P（Permutation，置换类）四种。进行排列组合理论上有 16 类 DGA，实际上只出现了 TDD-A, TID-A,TDD-W , TDD-H, TDN-A , TID-P 这六种类型。

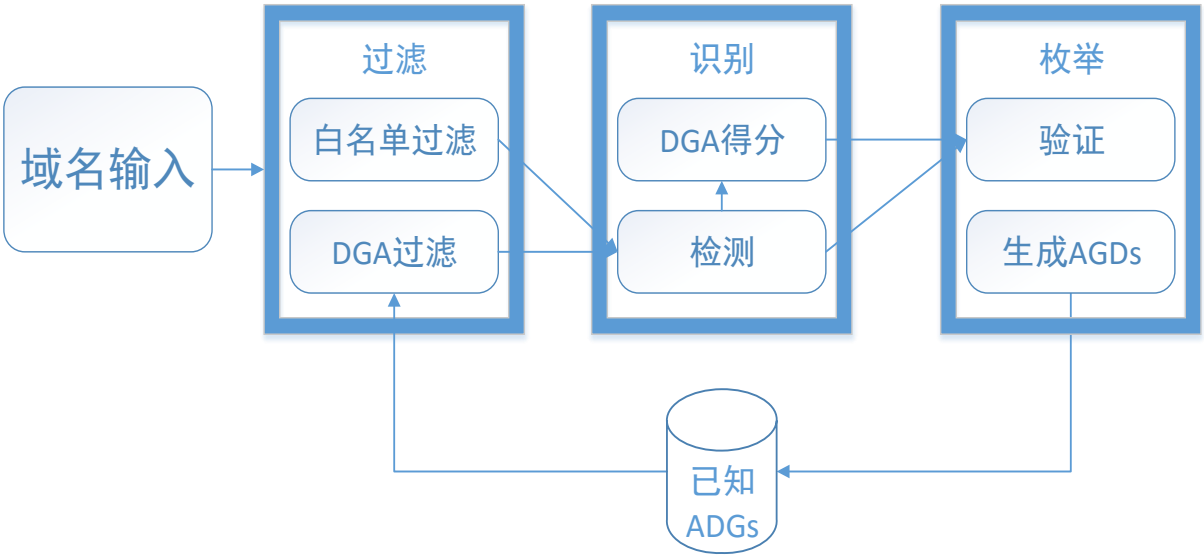

图 7 Plohmann 的 DGA 收集过程

图 7 为作者的 DGA 域名收集过程，实际工作中首先通过恶意软件分析报告和 blog，定义了最初的 22 个 DGA 家族。然后利用 Alexa 排名前一万名的域名和已知 AGD（由 DGA 算法产生的域名）进行过滤。接下来是识别的过程，构建探测器用来检测一个给定域名是否符合已知域名生成算法的输出，捕捉其产生域名的最大长度和最小长度，该域名生成算法所使用的字母表，已知顶级域名种类等字符特征。如果域名的构成模式相同，就将其标注为使用了新的种子的已知 DGA 家族；如果不一致或者域名数量不足导致不能准确判断，则通过进一步计算一系列的数值，例如 n-gram，熵值，长度等，来判定是否是新的 DGA 家族种类。当检测出来是一个可能的新 DGA 算法，人工再进行逆向工程来验证。

国内也有很多机构活跃在恶意域名的检测领域，中国科学院信息工程研究所有着大量与此相关的文章。例如张洋、柳厅文等人基于多元属性特征的恶意域名研究针对的是域名的伪装特点和跳变特点[33]，作者细致的将特征分类两类，一类是词法特征，其中包含了长度相关特征、计数相关特征、IP 相关特征等；另一类是网络特征，包括了资源记录查询的统计特征、时间上的特征以及网络划分所涉及到的一系列特征。并通过实验验证了这些特征对于恶意域名的鉴别具有较大的贡献，这些特征的选取具有很好的借鉴意义。

表 2　多元属性特征的恶意域名字符特征

编号	**词法特征**	**恶意域名可能的特征**
1	域名长度	域名长度较长
2	是否存在 IP 地址	存在 IP 地址
3	分隔符个数	分隔符较多
4	特殊字符个数	特殊字符较多
5	数字个数	数字较多
6	数字占总长度比例	数字占较大比例
7	数字字母转换频率	数字字母转换频率较大
8	大写字母个数	含有大写字母
9	域名分隔符间的最大长度	域名分隔符间的长度较大
10	连续数字的最大长度	连续数字较长
11	连续字母的最大长度	连续字母较长

表 3 多元属性特征的恶意域名访问特征

编号	网络属性特征	恶意域名可能的特征
1	TTL 平均值	TTL 平均值较小
2	A 记录个数	多个 A 记录
3	所属网段个数	所属于多个网段
4	AS 个数	AS 个数较多
5	NS 个数	NS 个数较多
6	NS 分散度	NS 分散度较大
7	NS 对应 TTl 平均值	NS 对应 TTL 平均值较小
8	注册时间	注册时间较晚
9	所属国家	国家分布不均匀

文章的实验中选择了随机森林作分类器，正样本为标注好的恶意域名，负样本为正常域名。如表 2、表 3 构建特征。随机森林在这里有许多的优势，例如对数据的维度不敏感，对数据的格式不敏感，对缺失数据不敏感等。而这些特征中，既有离散值又有连续值，网络的属性中也会出现诸多无法测量的值，刚好这些问题都可以被随机森林算法所弥补，这也是集成学习带来的优势。最终的实验结果表明，该方案的准确率、召回率都达到了 99.8%以上，显示除了良好的检测效果。

张永斌等人认为相同周期内产生的僵尸网络域名大多具有相同字符特征，被感染主机有相同行为特征，提出了他们的一个检测算法[34]。整个检测算法分为：数据预处理、不存在域名聚类分析、新域名聚类分析、恶意域名提取 4 个处理过程。他们方案的特点是针对组进行检测，无论是针对不存在域名亦或是新出现的域名，都划分成组，以组为单位提取单位并进行检测。

2.3.2 针对色情域名的研究

当前针对色情域名的研究大多数集中在针对网页内容的研究上，Luh 等人基于内容实现了对网站的分级[20]，苏贵阳等人基于页面汉字文本实现了中文色情网站识别[22]。除此之外还有一些研究人员通过网络访问日志来进行色情网站的识别，曹建勋等人基于日志中所记载的用户行为的挖掘，验证了用户访问色情网站与普通网站时的行为确实具有明显的差异[23]。他们从日志中提取了特定时间段网站的访问率、搜索引擎以及网站内部网页访问率、色情词关联度、URL 特定字符关联度。此算法会使用用户访问日志提取用

户行为特征，第一类特征是不同时间段内访问量的占比，由于色情网站浏览时间与正常网站不尽相同，例如多集中在晚上等，就会在浏览量趋势上有所不同；第二类特征是网站跳转类，进入一个网站链接的方式多种多样，例如搜索引擎、其他网站的跳转、邮件中的链接等等。色情网站通常情况下会出现较多的子网页不断跳转的情况，由于这些网页的内容为浏览者所需，因此也会有较长的页面停留时间。这里就可以形成网页链接跳转方式占比和子页面跳转占比的特征；第三类是搜索词关联度特征，这一类特征主要针对用户使用搜索引擎搜索的关联分析，首先设定好一个色情词汇的分级表，按照不同级别词汇统计搜索结果中跳转网站的比例。并且此类网站通常会经过更多的点击下一页的次数，这也是一个特征；第四类特征是域名关键字符，大量的色情域名当中会出一些明显的英文字符（例如 sex、ll、bb 等），这些字符和若干数字一起构成完整的二级域名。获得这些特征之后，使用诸如朴素贝叶斯、决策树等分类器进行训练，具体过程类似于基于字符特征的 DGA 检测算法。

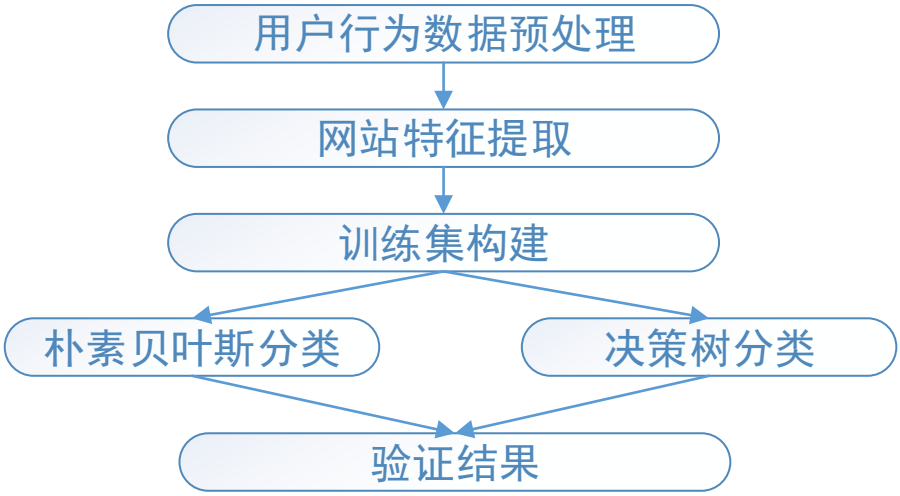

图 8　基于用户日志的色情网站检测流程

2.4 威胁情报平台相关应用

在学术领域 Notos、Pleiades、Kopis、EXPOURE 等一些列经典的系统都已在上文中仔细介绍过了，这一节的重点放在国内外工程与商业上的应用成果。

Nominum 是为网络运营商提供综合型用户、网络和安全解决方案的全球领先提供商[31]。Nominum 所建设的 N2 平台每天处理的 DNS 查询量数量级达到万亿次，实现利用 DNS 数据的应用的快速开发与无缝集成。在海量数据的支持下，始终活跃在抗击网络攻击的最前线，同时提供着最为权威的安全报告，对当前流行的、危害高的网络威胁（例如 Ddos、随机子域名攻击等）都有着最为细致的分析。

微步在线（ThreatBook）是一家从事专业威胁情报的公司，所发布的威胁分析平台VirusBook（www.virusbook.cn），是中国第一家推出综合性威胁情报分析平台的机构。随着定向攻击及 APT 的日益泛滥，业界清醒的认识到：单纯的防御方式很难消除面临的关键风险，安全检测和响应的重要性重新被发现，作为贯穿检测－定位－决策－行动全过程的威胁分析已成为当今的热点。现阶段安全分析师所使用的分析平台主要是国外的站点，如 VirusTotal、DomanTools、PassiveTotal 及一系列的开源情报站点。但是这些站点还存在着一些问题：首先，这些站点大多只提供了分析所需要的某一方面的信息，并不能完全覆盖整个分析过程，导致分析师需要到多个站点进行查询，并手动进行关联分析，这样的工作无疑是枯燥、低效的，严重影响了分析师能力的发挥；其次由于网络的问题，某些重要网站往往不能正常使用，使工作难以进行，而微步威胁分析平台实现了对于一个站点完成鉴别、定性、溯源、追踪等多重任务。

最后需要介绍一下 360 网络安全实验室所建立的一个威胁情报项目，公布了所有的检测数据并且做到每天更新，主要分为以下四个部分。一是针对全自动攻击工具 Exploit Kit 的行为检测；二是对 20 个 DGA 家族和超过 170 个种子的实时追踪；三是对任一个 IP、域名或者 MD5 值的恶意软件关联；四是当下流行的恶意代码扫描。

2.5 本章小结

本章系统的介绍了近年来国内外在恶意域名的研究过程中作出的不懈努力，其中佐治亚理工学院的 M.Antonakakis、P.Roberto、W.Lee 等人做出了尤为突出的贡献。他们从 2008 年开始，一直活跃在该领域的世界顶级平台只上，他们先后对 DNS 缓存投毒、域名信誉系统、DGA、僵尸网络等方向都做出了很高的贡献。尤其是 Pleiades 系统，诸多安全厂商皆参考此原型进行改进来对抗威胁日益增大的僵尸网络。除此之外很多国内外高校、研究机构也贡献了大量研究成果，并将大数据、人工智能等技术融入到其中，进行了许多有意义有价值的尝试，也产生了很多优秀的结果。这其中最突出的是 Noninum 公司研究人员首先尝试将自然语言处理中的 Word2vec 模型用于 DGA 域名的检测，北京大学周昌令等人也随后进行了相关研究，取得了不错的结果。

随后本章对于 DGA 域名和色情域名的相关特征的研究进行了介绍，对于 DGA 域名主要分为了访问的特征和字符的特征，色情域名的特征主要是内容上的特征，包含文本内容和其他内容，以及少量通过用户日志取得的访问特征。

最后介绍了一些威胁情报平台以及上文中相关技术在威胁情报平台上的应用，主要有 Nominum、VirusBook 等。

第三章 利用 PDNS 检测恶意域名的算法研究

本章围绕着被动 DNS 进行介绍，从相关的基础知识为起点，进一步介绍可以针对原始数据进行怎样的处理，转化为哪些特征，并针对这些特征设计了分析方案。本章更深一层的介绍了利用这些特征而设计的恶意域名相关流量的提取方案以及针对 DGA 域名和色情域名的分类算法。

3.1 PDNS 数据介绍

被动 DNS（PassiveDNS）数据是网络安全领域最为常用的资源之一，该数据收集技术由 Florian Weimer 于 2004 年提出[4]，主要目的是将 DNS 流量转换为易于访问的格式。递归域名服务器会对其接收到的来自其它域名服务器的请求进行响应，将响应记录其中的关键信息存储下来并将这部分数据复制到中央数据库当中，因此记录这些数据就掌握了 DNS 解析的历史动态。具体而言，各类可联网终端设备（包含移动终端、台式机等）在需要进行对域名进行访问时，就产生了 DNS 查询，DNS 服务器通过不断递归的方式向上查找，并返回结果，那么所有使用这个 DNS 服务器的用户都会通过某一层的递归服务器，并在上面留下相应的记录。通过对交换机、路由器端口的配置，将流向或者流出递归服务器的 DNS 流量实时的拷贝到一台专门的服务器上

Zdrnja 等人首先说明了如何利用被动 DNS 从域名中获得安全信息[18]，2008 年 Plonka 等人提出了 Treetop[19]，它可以弹性地管理逐渐增长的被动 DNS 数据，并在同时关联域和网络属性。他们的聚簇区域是基于不同的种类的网络。Treetop 依据是否符合各种 DNS RFC 标准以及解析结果来区分 DNS 流量并提供部分安全信息。

表 4 所示为被动 DNS 记录中保留的字段，在分析和实验之中主要使用到的字段有客户端 IP、请求域名、请求类型、服务器响应标识等。域名请求类型主要有 A、TXT、CNAME、ANY、MX 等。返回的域名类型和请求类型的不同之处在于没有 ANY 的返回结果，因为客户端发出 ANY 的查询请求后，递归服务器会将该域名能收集到的各种类型记录分别返回，这也是为什么在 DNS 放大攻击中攻击者通常会使用 ANY 记录查询发起攻击。另一个需要的字段是服务器响应标识，当 Rcode 为 0 时，表示没有发生异常，返回正确结果；当 Rcode 为 1 时，表示服务器收到的请求格式错误；当 Rcode 为 2 时，表示递归服务器出现错误；当 Rcode 为 3 时，表示递归服务器无法找到域名，这部分域名即 NxDomain；当 Rcode 为 4 和 5 时，分别表示服务器无法解析和拒绝解析。

表 4 被动 DNS 资源记录字段

字段	类型	备注
SIP	ipv4_addr	客户端 IP，字符串
DIP	ipv4_addr	递归 IP，字符串
DNSID	int	DNS 请求编号
DOMAIN	char(256)	请求域名
QTYPE	char(10)	请求域名类型，字符串
SCOUNT	int	请求报文重复次数
SRATE	Int	该时间间隔下报文采样率（如 10:1）
DIR	int	匹配标识（请求/应答/双向）
RD	Int	递归请求标识
AA	Int	权威请求标识
TTL	int	应答的首个 RR 的 TTL
TIME	datetime	当前获得报文的时间
PCODE	char(10)	节点编码（编码参考附录）
QLEN	Int	请求 Ip 包长
RLEN	Int	应答 IP 包长
RRTYPE	char(10)	应答的首个 RR 的类型
RCODE	Int	服务器响应标识
VALUE	char(256)	应答的首个 RR 的解析值

被动 DNS 资源记录保留的是递归服务器与客户端之间的 DNS 查询记录，具体分为 R2C（Recursive Server to Client）资源记录和 C2R（Client to Recursive Server）资源记录，两者最明显的区别在于是否包含 RRTYPE 字段，我在实验中所使用的为广东电信提供的 R2C 资源记录。

3.2 恶意域名特征介绍与分析方案

本节主要讲述域名的特征提取，这里只针对所有域名的二级域名。例如域名“*test.example1.com.cn*”和“*test.example2.com*”，这里“*.com.cn*”和“*.com*”被称为顶级域名，而“*example1.com.cn*”和“*example2.com*”被称为二级域名。大部分恶意域名都可以在二级域名这一级进行检测，例如 DGA 算法产生的域名，一般会从顶级域名选择

一个或者多个，与生成有意义或无意义的字符串拼接成一个二级域名。其他的恶意域名也类似，攻击者可以直接注册二级域名来使用。其他情况例如随机子域名攻击，攻击者会利用合法二级域名构建子域名，这些子域名多是三级的甚至是四级的，例如“*test1.example2.com*”和“*test2.test1.example2.com*”，他们仍然具有相同的二级域名“*example2.com*”，因此我们也会统计一定时间下一个二级域名的不同的子域名数量。

3.2.1 域名字符特征

一个合法域名 *d* 包含了数字、字母、“.”、“_”，例如“*www.example.com*”。最右边的部分被称为 *TLD*（top-level domain，顶级域名），例如“*.com*”。最右边的两个部分被称为 *2LD*（second-level domain，二级域名），例如“*.example.com*”。同样的，最右边三个部分被称为 *3LD*（third-level domain，三级域名），例如“*www.example.com*”。构成域名字符特征的种类很多，这里主要选取三类特征，一个域名可以得到 12 个 N-gram 特征、2 个熵的特征、3 个其他统计特征，共 17 个字符特征。

N-gram 特征：这里首先将白名单中的域名进行处理，把每个 *d* 的 *SLD* 去除 *TLD* 部分。然后在头部加上字符“*^*”，尾部加上字符“*$*”，例如合法域名“*www.example.com*”则变为“*^example$*”。将白名单中每一个域名做上述处理后用作语料库，分别计算 N-gram 的频率值，其中 n 的取值为 1、2、3、4。当 n 取值为 1 时，不考虑起始字符“*^*”和结尾字符“*$*”，则“*^example$*”处理为[‘e’, ‘x’, ‘a’, ‘m’, ‘p’, ‘l’, ‘e’]。当 n 取值为 2 时，“*^example$*”处理为[‘^e’, ‘ex’, ‘xa’, ‘am’, ‘mp’, ‘pl’, ‘le’, ‘e$’]。由此获得域名每个子字符串的 N-gram 概率，例如一个长度为 7 的字符串，我可以得到 7 个 1-gram 频率值，8 个 2-gram 频率值，7 个 3-gram 频率值和 6 个 4-gram 频率值。分别计算每组频率值的均值，中位数和标准差，最后一个域名可以得到 12 个 N-gram 特征。

熵的特征：给定一个域名，将每个域名 *d* 的 *SLD* 去除 *TLD* 部分之后，按照公式(3.1)来计算熵值 *H(d)*，其中 *p(c)*为二级域名中每一个字符的出现概率，熵值可以表现出一个域名构成的随机程度。然后计算 *H(d)*除以字符串长度 *l*，得到字符平均熵，这里我们每一个域名可以获得两个熵相关的特征。图 8 展示了黑名单与白名单中的二级域名长度和熵值之间的关系，在长度一致的时候，dga 的域名往往与有更大的熵值。

$$H(d) = -\sum p(c) * Log_2 p(c) \tag{3.1}$$

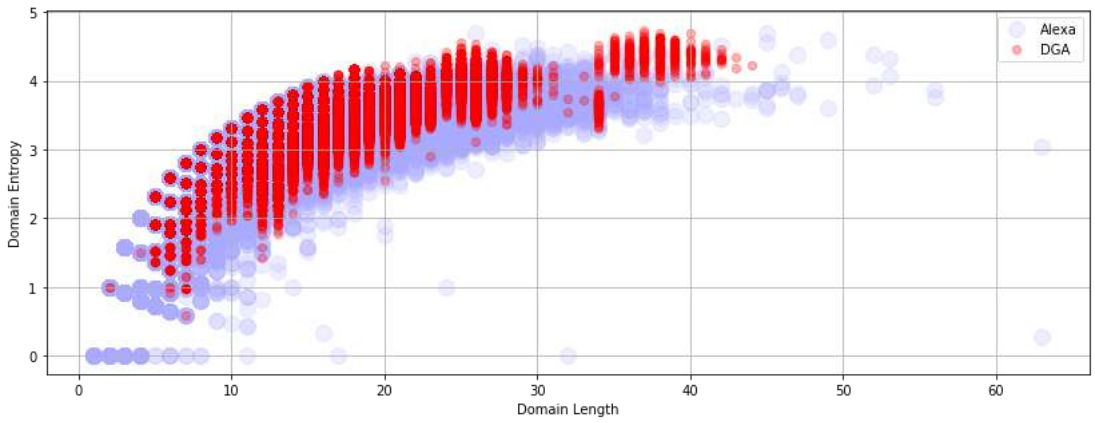

图 9　域名熵值随着域名长度的变化

其他统计特征：给定一个域名，将每个域名 *d* 的 *SLD* 去除 *TLD* 部分之后，统计其长度 *L*，及其中字母数量 *La* 和数字数量 *Ln*。由于“*.com*”顶级域名下恶意域名数量和其它顶级域名下恶意域名数量的比例差别很大[16]（监管原因和价格原因），因此还需要记录该域名的顶级域名部分是否是“*.com*”。这一部分统计以上四个特征。

3.2.2 域名访问特征

在被动 DNS 中记录了每一个 IP 对域名的请求，接下来讨论一下如何记录 IP 和域名的访问关系特征，在 Antonakakis 等人的经典论文“从丢弃的流量中检测僵尸网络”中，他们使用一个二部图来记录相关特征[10]。感染了同一种 DGA 的恶意软件的主机有着很大的可能生成互相重叠的 NXDomains 集合。另一方面，其他“非 DGA”的 NXDomains 则不会被多个主机查询。例如，在一段时间内多个用户同时犯同样的拼写错误是不太可能的。他们由此构建了一个稀疏矩阵 M，其中行代表一段时间内查询过两个以上 NXDomains 的主机，列代表 NXDomains，并对 M 进行谱聚类将相近主机共同查询过的 NXDomains 聚簇在一起。再统计聚簇的特征，做训练和分类使用。这样做的缺点是无法获取单个域名的特征，而且过程十分复杂，无法有效的将域名访问特征体现出来，因此我参考周昌令等人的文章将域名映射到向量空间之中[24]，在保留域名关联的同时，将出现达到一定次数的域名转化为向量，同时也避免了使用诸如 one-hot 编码导致向量过长，便于进一步的计算。这里将产生一个 200 维的词向量。除此之外，流量计数相关统计特征 165 个，其中原始流量直接统计得到的特征有 139 个，衍生出来的特征有 26 个。

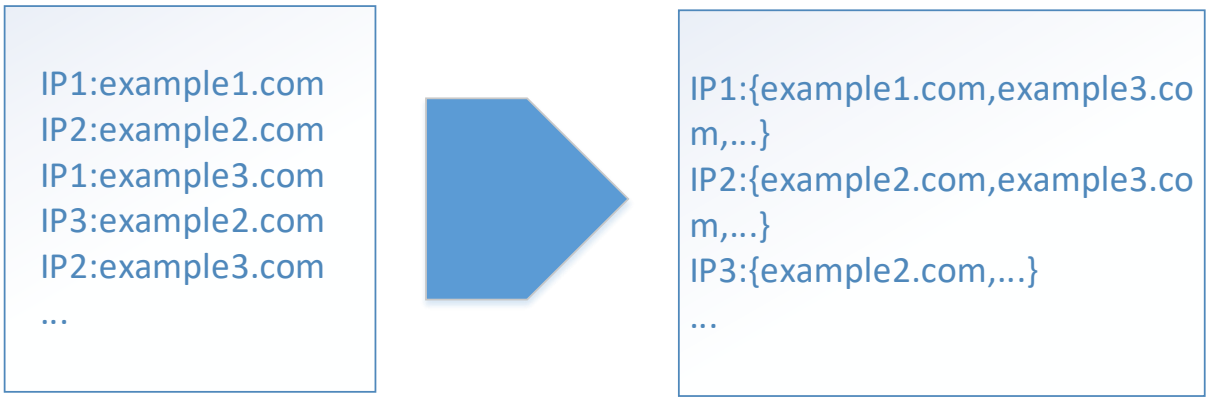

图 10 被域名查询列表构建

Word2vec 特征：word2vec 是 2013 年由谷歌开发出一个词向量工具[25]，该算法既可以通过上下文来推测目标词语构建向量（CBOW 模式），也可以通过词来推测上下文以构建向量（Skip-Gram 模式），本文利用的是 CBOW 模式。在经过 3.3.1 节的基本数据清洗之后，将每一个 IP 一定时间下的 DNS 查询按照时间戳的先后顺序并提取其 SLD 串联起来，相同的 SLD 可以多次重复的出现，如图 9 构建被域名查询列表[24]。这样构成的每一个 IP 的被域名查询序列看成是一句话，在本文中仅考虑所有的 A 记录查询。然后利用用基于 gensim 包的 Word2Vec 模块框架训练得到词向量。

访问统计相关特征：以二级域名为关键字，统计访问特征，包含了各种记录类型的查询总数以及各小时内的查询记录，也根据 rcode 的不同进行了相关统计，具体特征见下表。

表 5 域名查询统计特征

特征	数据类型	备注
QAC	Integer	该域名被查询 A 记录的总次数
QCNAMEC	Integer	该域名被查询 CNAME 记录的总次数
QTXTC	Integer	该域名被查询 TXT 记录的总次数
QMXC	Integer	该域名被查询 MX 记录的总次数
QNSC	Integer	该域名被查询 NS 记录的总次数
QDNSKEYC	Integer	该域名被查询 DNSKEY 记录的总次数
QANYC	Integer	该域名被 ANY 查询的总次数
QSOAC	Integer	该域名被查询 SOA 记录的总次数
QSPFC	Integer	该域名被查询 SPF 记录的总次数
QSRVC	Integer	该域名被查询 SRV 记录的总次数

QDSC	Integer	该域名被查询 DS 记录的总次数
QNAPTRC	Integer	该域名被查询 NAPTR 记录的总次数
SUCCESS	Integer	该域名 A 记录或 AAAA 记录查询成功的总次数
WRONGQUERYC	Integer	该域名 A 记录或 AAAA 记录查询不存在的次数
WRONGSERVERC	Integer	该域名 A 记录或 AAAA 记录查询格式错误总次数
NONEXISTC	Integer	该域名 A 记录或 AAAA 记录查询不存在总次数
SIPC	Integer	不相同的请求 IP 总个数
SECDOMAINC	Integer	该域名下被查询的不相同子域名总个数
QAC_N	Integer	第 n 小时该域名 A 记录被查询的次数
QCNAMEC_N	Integer	第 n 小时该域名 CNAME 记录被查询的次数
QTXTC_N	Integer	第 n 小时该域名 TXT 记录被查询的次数
QMXC_N	Integer	第 n 小时该域名 MX 记录被查询的次数
QNSC_N	Integer	第 n 小时该域名 NS 记录被查询的次数
QANYC_N	Integer	第 n 小时该域名 ANY 记录被查询的次数
QSRVC_N	Integer	第 n 小时该域名 SRV 记录被查询的次数
SUCCESS_N	Integer	第 n 小时该域名 A 记录记录查询成功的次数
NONEXISTC_N	Integer	第 n 小时该域名 A 记录记录查询不存在的次数
SIPC_N	Integer	第 n 小时不相同的请求 IP 个数
SECDOMAINC_N	Integer	第 n 小时该域名下被查询的不相同子域名个数

注：由于数据的原因，其中 n 为 13 至 23 共 11 个小时

除此之外，还有一些由上述统计特征衍生出来的特征，见下表。

表 6　域名查询统计特征的衍生特征

特征	数据类型	备注
DNSAMPLIFICATION_SCORE	Double	DNS 放大攻击得分（公式(3.5)）
PRSD_SCORE	Double	随机子域名攻击得分（公式(3.6)）
QAR	Double	该域名 A 记录查询数量占比
QCNAMER	Double	该域名 CNAME 记录查询数量占比
QTXTR	Double	该域名 TXT 记录查询数量占比
QSUCCESSR	Double	该域名成功查询数量占比

QNONEXISTR	Double	该域名不存在查询数量占比
QACPERIP	Double	平均每个 ip 的 A 记录查询量
VP_N	Double	第 n 小时该域名查询量占总查询比例
QAC_AVG	Double	该域名 A 记录每小时查询数平均值
QAC_SD	Double	该域名 A 记录每小时查询数方差
SUCCESSC_AVG	Double	该域名每小时被成功查询数平均值
SUCCESSC_SD	Double	该域名每小时被成功查询数方差
NONEXISTC_AVG	Double	该域名每小时被查询不存在平均值
NONEXISTC_SD	Double	该域名每小时被查询不存在方差
VP_SD	Double	该域名每小时查询数占比方差

3.2.3 特征分析方案

现有特征较为庞大，数据类型和表现形式上均有所不同，拟针对不同类型的特征给出不同的特征分析方案。对于 3.2.2 节中的域名访问特征，由原始的统计特征衍生出了一系列特征，针对这些衍生变量我需要探索是否和原本的统计变量存在简单的线性相关，因此需要计算这些衍生变量和原本统计变量之间的简单相关系数，如公式(3.2)所示，其中 *Cov(X,Y)*表示 X 和 Y 的协方差，*Var(X)*表示 X 的方差。

$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var[X]Var[Y]}} \tag{3.2}$$

进一步的对于这些域名访问特征拟通过信息值来判断特征的优劣，IV（Information Value，信息价值）是一种对自变量预测能力的量化衡量指标。直观上来说，在分类的问题中，一个变量 X_i 对 Y 的判别贡献越大，那么这个变量的 IV 值就越高。再进一步介绍信息价值之前需要首先了解 WOE（Weight of Evidence，证明权重）的概念。在计算 WOE 时第一步要做的就是对变量 X_i 进行分箱操作，即对连续数据进行离散化的操作。通常有等距离散、等频离散这两种基本的分箱方法，当然也有卡方检验方法、信息增益方法等优化的离散方法，本文中使用的是 riv 包中针对最大化信息价值的优化分箱方法。进一步的根据公式(3.3)计算出每一个分箱中的数据对分类结果有怎样的一种影响和其影响能力的大小。以二分类为例，$\#P_i$ 表示该变量的第 i 个分箱中正例的数量，$\#P_T$ 表示样本中正例的总量，$\#N_i$ 表示该变量的第 i 个分箱中反例的数量，$\#N_T$ 表示样本中反例的总量。

通过该公式(3.3)可以看出 WOE 反映了一个变量 X_i 每个分箱下正例和反例的变化差异，即可以观察分辨是正影响或者是负影响，也可以通过数值反映其影响得大小。

$$WOE_i = \ln \frac{\#P_i/\#P_T}{\#N_i/\#N_T} \tag{3.3}$$

IV 值是对与每一个变量的分箱中 WOE 值分派一个权重之后的累加和。其中对于第 j 个特征的信息价值 IV_j，每一个分箱下的权值由第 i 个分箱中正例的数量与正例总数量之比和第 i 个分箱中反例的数量与反例总数量之比的差值，在表现出正反例差距的同时在数值上保持了大于零的结果。表 7 为不同的 IV 值下某一个特征的预测能力对应强弱关系。

$$IV = \sum \left(\left(\frac{\#P_i}{\#P_T} - \frac{\#N_i}{\#N_T} \right) * \ln \frac{\#P_i/\#P_T}{\#N_i/\#N_T} \right) \tag{3.4}$$

表 7　IV 值对应预测能力关系

IV 值	预测能力
[0.3, 1.0]	Very Strong
[0.2, 0.3)	Strong
[0.1, 0.2)	Average
[0.01, 0.10)	Weak
[0.00, 0.01)	Very Weak

3.3 快速提取恶意域名相关流量的算法

图 11　快速提取恶意流量算法流程

网络中充斥着大量的恶意请求，其中绝大部分是是放大攻击、随机子域名攻击和僵尸网络产生的流量，在利用 DNS 流量进行一些恶意行为的分析的时候，往往需要对每一个域名单独进行检测，而数据量是很大的，简单的过滤无法迅速降低需要检测的域名数量，这样一来在计算资源有限的前提下，需要花费大量的时间。本节介绍了一种针对

被动 DNS 流量的提取方案，我们针对流量中占比最大的三种攻击流量，分别设计了简单快捷的提取手段。

3.3.1 数据预处理

快速提取恶意域名相关流量的方案对采集来的原始数据进行三层预处理。分别是数据清洗、数据过滤和特征提取。首先对所有数据进行清洗，一个合法的域名只包含 26 个英文字母（包括大写和小写）、数字、中划线和用来分割成每一段的点。通过构建正则表达式很容易可以将这些数据清理干净。还有一部分是无顶级域名的域名，产生这种情况的原因比较多，有配置或者人为的各种原因。

其次，进行数据的过滤操作。第一部分先过滤反向解析域名，域名反向解析是指从 IP 地址到域名的映射，其主要应用于邮件服务器阻止垃圾邮件。为了实现逆向域名解析，因此有一个专门 DNS 服务器负责反向解析，返回数据包较小，不适合用作放大攻击，因此恶意流量中没有出现利用反向解析记录进行攻击的行为。第二部分要过滤配置错误产生的域名，这类域名极为常见，这其中以“*.local*”,“*.localhost*”为后缀的域名出现最多。第三部分过滤国际化域名，国际化域名是指非英语国家推广本国语言域名系统的一个总称，使用 punycode 编码。

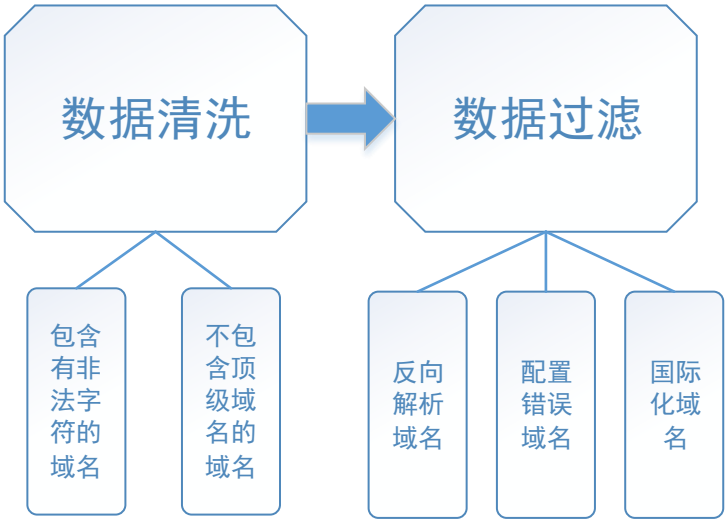

图 12 数据预处理过程图

3.3.2 针对 DNS 放大攻击相关域名的提取

这一部分主要目的是要将放大攻击中作为跳板的那部分域名找出来。攻击者想要利用这些域名，必然是利用其 TXT 记录或者 ANY 查询返回该域名所有资源记录。我们获取 ANY 类型请求比例 qar 和 TXT 类型请求比例 qtr，带入公式(3.5)中，β 为我们设定的

一个参数，当 qar+qtr<=β 时，结果为 0，当 qar+qtr>β 时 qar+qtr 与 s1 成正相关，同时设定阈值 α，其中 s1>α，我们认定为疑似放大攻击的流量。这里 α 取值为 0.1，β 取值为 0.05，阈值的取值将在下一章仔细讨论。

$$S_1 = \max\left\{0, 1 - e^{-\frac{qar+qtr-\beta}{\beta}}\right\} \tag{3.5}$$

3.3.3 针对随机子域名相关域名的提取

这一部分目的是将随机子域名攻击所使用的那部分域名找出来，攻击者会在二级域名下伪随机的生成大量子域名，这些域名都是不存在的。因此我们用 sdc 与 nxdr 相乘来表示被恶意使用的可能性大小，而这个值范围较大，我们使用公式(3.6)将结果变到 0 到 1 之间。这里我们的 θ 取值为 0.3。

$$S_2 = \frac{e^{\theta(sdc*nxdr)} - 1}{e^{\theta(sdc*nxdr)} + 1} \tag{3.6}$$

3.3.4 针对 DGA 域名的提取

域名生成算法产生的域名也是二级域名的部分，所以这部分的检测关注针对流量中的二级域名，我们模型使用黑白名单训练，数据来源在 4.1 节中介绍。将名单中每个域名的二级域名提取出来，分别计算其长度、熵值、n-gram，其中 n=2,3,4。分类器这里选择随机森林，训练每棵树时，从全部训练样本（样本数为 N）中选取一个可能有重复的大小同样为 N 的数据集进行训练（即 bootstrap 取样）。随机森林的优点为训练速度快，并且可以平衡误差。

3.4 恶意域名分类算法

本节旨在介绍 DGA 域名和色情域名的检测算法，不同算法之间的主要区别主要集中在域名特征的获取渠道和构建方式。

3.4.1 基于不存在域名检测算法

该算法主要使用被动 DNS，通过提取出其中不存在域名的流量，首先进行统计特征的聚类。给定监控网络下主机访问过得所有 NXDomains 的集合 NX，将 NX 分为长度为 α 的子集若干，假设 m 是 NX 中不重复的 NXDomains 的个数，将集合 NX 分割成[m/α]个子集，其中 α=10。对于得到的每一个子集NX_k，计算若干个字符统计特性。将每个 NX_k 转换成对应的特征向量后，这里使用 X-means 聚类算法。X-means 会将 NX_k 聚成 X

个簇，X 是由 X-means 本身的自动优化过程计算出来的。此时，给定一个大小为l个子集的聚簇 C = $\{NX_k\}_{k=1...l}$，可以简单地将 C 里的NX_k求并作为一个 NXDomain 聚类。这里之所以使用 X-means，是因为该算法对于 K-means 进行了改进和优化，针对每一轮迭代耗时长的问题，X-means 使用 kd-tree 对每一轮的迭代进行加速；针对 K-means 需要指定 k 的问题，X-means 使用 BIC 分数选取最优 k 值；针对容易收敛到局部最优解的问题，X-means 每一轮迭代只使用 2-means[27]。

输入：稀疏矩阵$M \in \Re^{l \times k}$,行代表 l 个主机，列代表 k 个 NXDomains

[1]：归一化 M：$\forall j = 1, \ldots, k \sum_{i=1}^{l} M_{i,j} = 1$

[2]：依据 M 计算相似矩阵 $S=M^T \cdot M$

[3]：通过特征值分解依据 S 计算ρ个特征向量

令$U \in \Re^{\rho \times k}$为从 S 特征值分解得来的包含 k 个维度为ρ的向量$u_1, \ldots, u_k$的矩阵(向量u_i代表第 i 个 NXDomain 压缩成ρ维后的向量)

[4]：使用 X-means 算法聚类向量(也就是 NXDomains)$\{u_i\}_{i=1,\ldots,k}$

输出：NXDomains 簇

图 13 基于不存在域名检测算法伪代码

感染了同一种 DGA 的恶意软件的主机有着很大的可能生成互相重叠的 NXDomains 集合。另一方面，其他“非 DGA”的 NXDomains 则不会被多个主机查询。例如，在一段时间内多个用户同时犯同样的拼写错误是不太可能的。构建一个稀疏矩阵 M，其中行代表一段时间内查询过两个以上 NXDomains 的主机，列代表 NXDomains。为了降低矩阵的维度，抛弃只查询一个 NXDomain 的主机，因为鉴于它们产生的 NXDomains 太少，这些主机上不太可能运行着 DGA。如果主机 h_i 没有查询过 $NXDomain_j$，令矩阵元素 $M_{ij}=0$. 相反，如果 h_i 查询过 n_j，令 $M_{ij}=w_i$, 其中 w_i 是权重。所有与h_i相关的非 0 项都被赋予权重 $w_i \sim 1/k_i$, 其中 k_i 是 h_i 查询过的 NXDomains 的个数。很明显，M 可以看成是一个二部图，其中主机顶点 V_{hi} 与 NXDomains 顶点 V_{nj} 被一个权重为 w_i 的边相连。当计算出 M 后，我们使用基于谱聚类的图划分策略(归纳见图 13)。首先，计算 M 的 ρ 个特征向量(实验中 ρ =15)，接着将每个 NXDomain(M 的每一列)映射到 ρ 维向量上。这样做能

很大程度上将 NXDomains 的向量维度从所有主机(M 的每一行)减小至 ρ 。接着我们对得到的 ρ 维 NXDomains 向量使用 X-means 算法，从而基于它们的“主机关系”进行聚类。即被相近主机共同查询过的 NXDomains 会被聚簇在一起.

之后再通过关联聚类取得训练集，DGA 分类器基于多类的决策树算法(ADT)。ADT 利用了 Boosting 的高分类精度，它产生紧凑的分类规则，可以更容易被解释。为了检测出感染了 DGA 恶意程序的主机，我们监视受控网络下每个主机产生的 NXDomains 并定期将这些信息传入 DGA 分类器。给定主机 h_i 产生的 NXDomains 集合 NX_{hi}，我们将 NX_{hi} 分割成长度为 α 的子集，每个子集我们提取统计特征。如果其中的一个子集被 DGA 分类器标记为是由给定的 DGA 产生的，我们就将 h_i 标记感染主机并且将它的 IP 地址和 DGA 标签记入恶意程序检测结果中。

3.4.2 基于词向量空间的检测算法

经典的统计语言模型通常是基于概率来进行的，例如 HAL（Hyperspace Analogue to Language method）[28]、LAS（Latent Semantic Analysis）[29]、COALS[30]等。他们的基础都是根据贝叶斯公式之类的方法计算后验的条件概率，最简单的情况就是类似马尔科夫链的形式，如公式(3.7)所示，当前的这一个单词仅仅与他的若干个前序单词相关，这就构成了 n 元的 gram 语言模型。

$$P(w_i|w_1,\dots,w_{i-1}) = P(w_i|w_{i-n+1},\dots,w_{i-1}) \tag{3.7}$$

但这些统计语言模型都无法有效的将域名转化到向量空间当中，要做到这一点最初的做法也是最简单的做法就是使用 one-hot 编码，即将离散值用一个只有 0 和 1 的向量来表示，有多少个不同的离散值这个向量就有多少的维度，其中对应离散值数值表示为 1，其余位置都为 0。但是这么做的缺点也很明显，很容易使得维度过于庞大，而且构成的矩阵也是十分的稀疏。Mikolov 使用无监督的方法提出一种精简的语言模型学习固定长度的特征表示[30]，这就是 Word2Vec 的框架。接下来进一步以 CBOW 模式为例说明该模型。假设目标词为 *w*，那么 *w* 的上下文为 *Ctx(w)*，给定一个正样本（*w*，*Ctx(w)*），给出目标函数公式(3.8)，其中 NEG(w)为随机负采样得到的样本子集。

$$g(w) = \prod_{u\in\{w\cup NEG(w)\}} p\big(u\big|Ctx(w)\big) \tag{3.8}$$

其中 *p(u | Context(w))*为

$$p\big(u\big|Context(w)\big) = [\sigma(x_w^T\theta^u)]^{L^w(u)}\cdot[1-\sigma(x_w^T\theta^u)]^{1-L^w(u)} \tag{3.9}$$

我们想要最大化 *g(w)*，等价于最大化$\sigma(x_w^T\theta^u)$，即在上文问 *Ctx(w)*的环境下，*w* 出现概率最大。这是针对一句话中一个词 *w* 的形式，接下来给出语料库 *C*，则有：

$$\mathrm{G} = \prod_{w\in C} g(w) \tag{3.10}$$

对 *G* 取对数，则有：

$$\mathrm{L} = \log G \tag{3.11}$$

$$= \sum_{w\in C}\sum_{u\in\{w\cup NEG(w)\}} \{L^w(u)\cdot\log[\sigma(x_w^T\theta^u)] + [1-L^w(u)]\cdot\log[1-\sigma(x_w^T\theta^u)]\}$$

使用梯度上升法进行优化，分别对θ^u和x_w进行求导。

$$\frac{\partial \mathrm{L}(\mathrm{w},\mathrm{u})}{\partial \theta^u} = [L^w(u) - \sigma(x_w^T\theta^u)]x_w \tag{3.12}$$

$$\frac{\partial \mathrm{L}(\mathrm{w},\mathrm{u})}{\partial x_w} = [L^w(u) - \sigma(x_w^T\theta^u)]\theta_w \tag{3.13}$$

则参数更新公式为

$$\theta^u := \theta^u + \eta[L^w(u) - \sigma(x_w^T\theta^u)]x_w \tag{3.14}$$

$$v(\tilde{w}) := v(\tilde{w}) + \eta \sum_{u\in\{w\cup NEG(w)\}} \frac{\partial \mathrm{L}(\mathrm{w},\mathrm{u})}{\partial x_w}, \qquad \tilde{w}\in Context(w) \tag{3.15}$$

CBOW 模型的网络结构是通过上下文预测中间词，输入为多个词，输出层输出多维向量。而将这个模型框架利用在域名相关的研究当中的关键点就在于如何构建语料库，构建语料库的方法已在 3.2.2 节中进行了介绍，即将一个 IP 在一定时间内的域名 A 记录请求按顺序排列视为一句话，其中的每一个域名即为 *w*，其前后若干域名为 *Ctx(w)*，据此构建语料库进行训练。

3.4.2 基于域名字符特征的检测算法

很多时候我们无法获得大量有效的用户访问数据，而黑、白名单则相对较为容易获得，在利用黑名单和白名单为训练集来训练分类器的情况下，可以选择的特征非常有限，这时就需要充分的利用域名本身的特征来训练分类器。对于具体的操作方法就如 2.3 节中所描述的那样，从多个角度进行特征提取，诸如长度相关的、熵相关的、字符顺序相关的等等，本论文中对与字符特征有一个很好的总结，创新性的提取了多达 17 个域名本身的字符特征。如图 14 所示，该算法使用黑白名单训练分类器，这样做的最大优点在于数据的易收集性，相比与用户日志和被动 DNS 数据，黑名单可以在各类威胁情报

平台或安全实验室的公开数据中找到，白名单通常使用 Alexa 排名前一万或者一百万的数据，也可以很方便在公开的网络环境中找到。利用这些数据构建的字符特征，在训练集上实验，找到最优的分类器，分类器的选择通常有线性回归、决策树、梯度提升树等。最终在真实的流量数据中使用训练好的模型分类，观察效果，并不断重复这个过程进行优化。

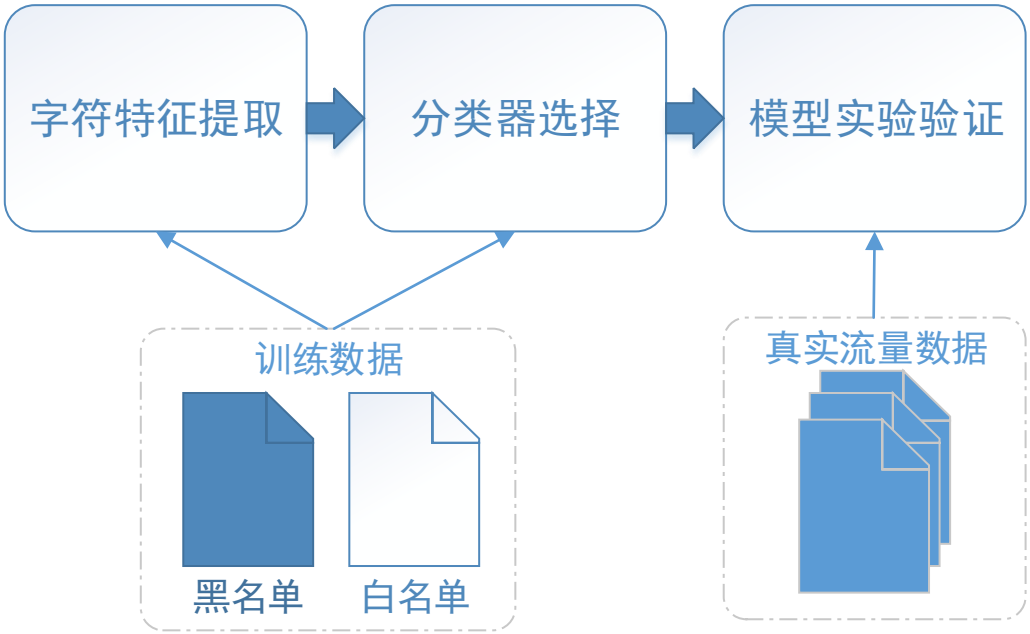

图 14　基于域名字符特征的检测算法

3.5 本章小结

本章由被动 DNS 的知识介绍切入，简明扼要的说明了被动 DNS 的来源和构成，以及解析之后的字段含义。 随后在此基础上给出了如何从被动 DNS 记录之中提取特征以及如何对这些特征进行分析。提取的特征主要分为了两类，第一类是域名的字符特征，其中包含了 N-gram 特征、熵的特征、长度特征等等；第二类是域名的访问特征，这里包含了时间维度上的访问统计特征，以及使用 Word2Vec 框架得到的向量特征。而针对这些特征的不同特点，设计了不同的特征分析方案，例如一些特征由基础的统计特征生成，如果新特征与就特征存在线性相关的关系，那么新生成的特征就没有什么意义，这时我就会选择去分析不同特征之间的协方差进行判断。其他的分析方案还有计算信息值来判断该特征的预测能力，或者通过直接训练对训练结果的 AUC、KS 分析来判断特征的优劣等。

接下来介绍了从被动 DNS 中提取出恶意域名相关流量的方案，现在网络中主要存在着三类恶意流量，分别是 DNS 放大攻击流量、DGA 域名流量和随即子域名攻击涉及的流量。首先经过一个初步的预处理，将不合法的域名以及一些不会涉及到这些恶意域名的流量筛除。而后针对这三类恶意流量分别设计了不同的提取方案，或通过评分或通过简单的分类器。恶意域名的分类算法相比较而言就要复杂一些，介绍的检测算法关键点大都在特征的构建上面，当然也有在数据的选取上的。通过对数据的不同划分可以产生更加丰富的特征，具体到分类算法上主要有随机森林、GBDT 等。具体到色情域名的算法上，现有的研究基本都直接针对网页的内容，或者是利用用户访问日志分析的算法，鲜有利用被动 DNS 数据来检测色情域名的算法，本文中创造性的提出了基于词向量空间的色情域名检测算法，并给出思路。

第四章 恶意域名检测应用原型系统设计与实现

本章是主要根据上文提出的快速提取恶意域名相关流量和恶意域名分类算法，设计并实现了原型程序。

4.1 需求分析

4.1.1 PDNS 预处理评分模型

如图 15 所示，网络中充斥着大量的恶意请求，其中绝大部分是是放大攻击、随机子域名攻击和僵尸网络产生的流量，在利用 DNS 流量进行一些恶意行为的分析的时候，往往需要对每一个域名单独进行检测，而数据量是很大的，简单的过滤无法迅速降低需要检测的域名数量，这样一来在计算资源有限的前提下，需要花费大量的时间。因此需要找到一个合理的方法快速并且便捷的将这部分流量提取出来。

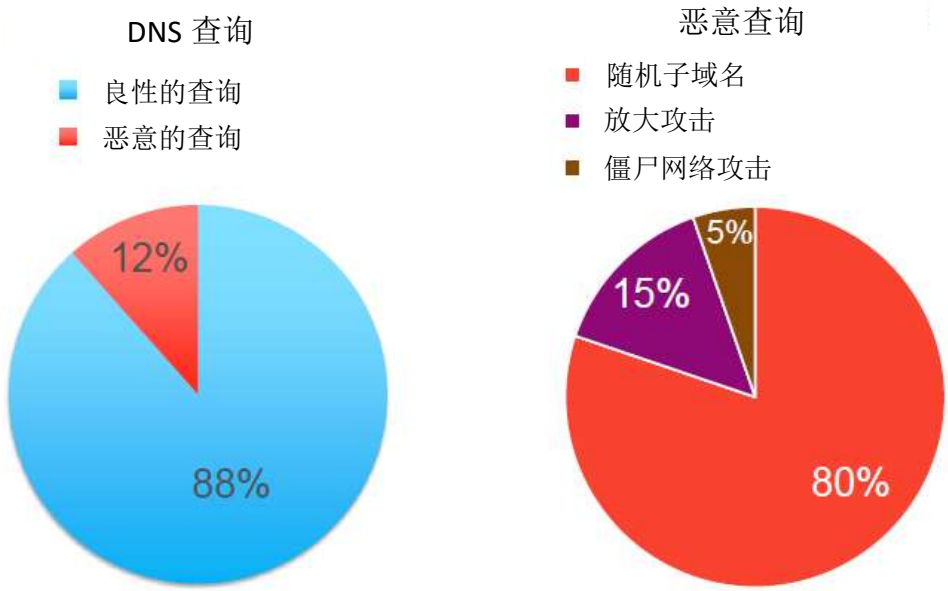

图 15 互联网恶意流量比例图示

4.1.2 域名特征提取模块

无论是域名的预处理或者恶意域名的检测，都有必要对域名特征做提取的操作。由于所需特征的种类不同，在提取时所使用的方式也有所不同。针对每一条被动 DNS 记录，都需要想先获取其二级域名，进一步对于每一个二级域名需要获取其字符特征，然后将同一 IP 下的被动 DNS 记录按照 3.2.2 节中的方式进行组织，使用 Word2Vec 框架进行运算，取得每一个二级域名在向量空间下的对应向量。再将同一二级域名下的被动

DNS 记录聚合，取得其他统计特征。以分布式的角度来看，每一个被动 DNS 记录通过 map 操作，将 key 值设定为其对应二级域名和 IP，第一个 reduce 操作将 key 值为相同二级域名的数据进行操作，并按时间统计得到特征，第二个 reduce 操作将 key 值为相同 IP 的数据聚合，将结果通过 Word2Vec 嵌入到向量空间。

4.1.2 恶意域名分类模型模块

面对流量中的恶意域名，我们最希望的就是能够将其检测出来，现阶段主要的三类检测手段有以下三种，一是信誉系统，信誉系统通常检测数据较快、较为便捷，但是较为滞后，并且准确性不高；二是逆向工程，逆向工程检测准确率极高，但是也较为滞后，并且对专业人员素质要求较高；三是机器学习，检测速度较快，较为及时，但是准确性比较一般。这三类检测方法都有所缺陷，但是机器学习的方法却有着较为突出的优点，因此，该模块需要利用好其优点，在此基础上通过特征工程以及算法的优化，尽可能的提前高其检测的精度。

4.2 总体设计

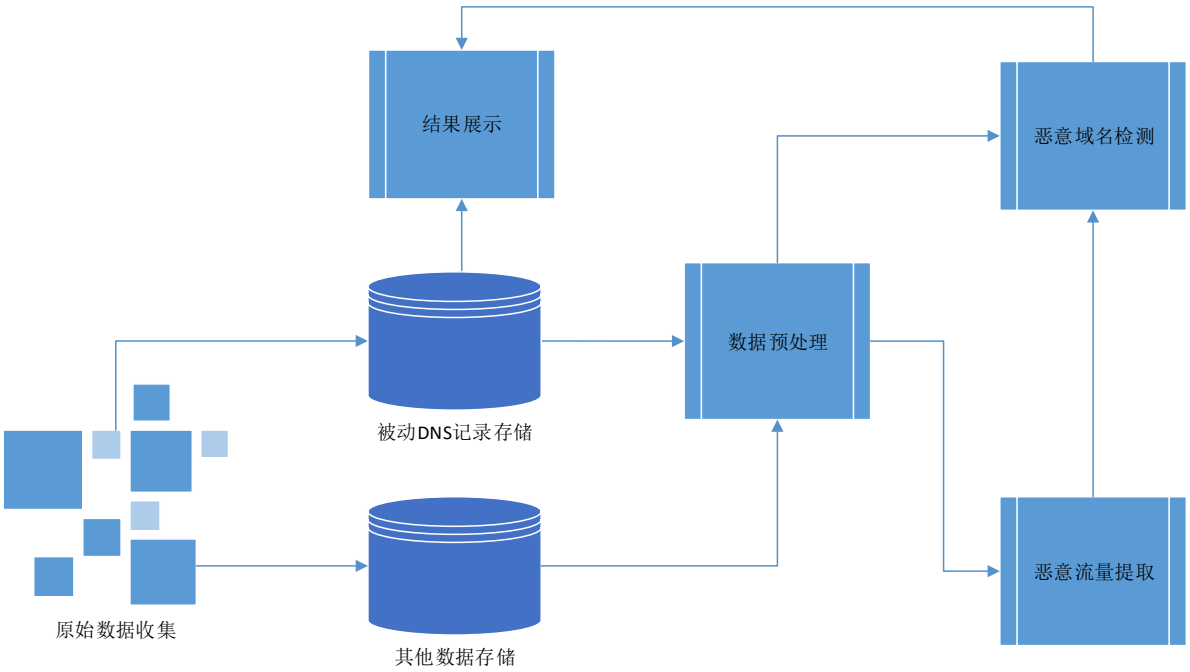

图 16 论文总体流程图

如图 16 所示为系统的整体流程图，原始数据的收集包含了被动 DNS 的收集和其他数据的收集。被动 DNS 存储在分布式的存储系统之上，其他数据存储包含了收集的有标

注的数据，例如 DGA 域名的黑白名单、Alexa 排名的域名等等。以上收集到的数据经过如 3.3.1 节中所述的数据预处理之后进行恶意流量的提取，其中包含随机子域名相关流量提取、DNS 放大攻击流量提取、DGA 域名涉及流量提取。处理之后的流量对其中的 DGA 域名和色情域名按照不同的检测方案进行检测得到结果。最终结果和流量监控一同在展示模块中展示出来。

4.3 功能实现

4.3.1 特征提取

第一步是准备语料，原始语料共有两部分，一部分是原始被动 DNS 数据，另一部分是黑白名单。

第二步是预处理，预处理的部分包含了数据清洗和数据过滤。数据清洗的目的是合法的域名保留下来，去除非法域名一个合法的域名只包含 26 个英文字母（包括大写和小写）、数字、中划线和用来分割成每一段的点。通过构建正则表达式很容易可以将这些数据清理干净。还有一部分是无顶级域名的域名，产生这种情况的原因比较多，有配置或者人为的各种原因。数据过滤是过滤掉不需要的或者是干扰性的数据。第一部分先过滤反向解析域名，域名反向解析是指从 IP 地址到域名的映射，其主要应用于邮件服务器阻止垃圾邮件。为了实现逆向域名解析，因此有一个专门 DNS 服务器负责反向解析，返回数据包较小，不适合用作放大攻击，因此恶意流量中没有出现利用反向解析记录进行攻击的行为。第二部分要过滤配置错误产生的域名，这类域名极为常见，这其中以“*.local*”,“*.localhost*”为 后缀的域名出现最多。第三部分过滤国际化域名，国际化域名是指非英语国家推广本国语言域名系统的一个总称，使用 punycode 编码。

第三步是数据初加工，对于初加工，我要做的操作有以下几点，一是将将原始数据中的域名使用该域名的二级域名替换，以便之后对于字符特征的提取；二是将原始数据按照 3.2.2 节中的方式将每个 IP 的查询按时间顺序组织起来，并使用 Word2Vec 框架进行训练，以便之后获取向量特征；三是使用白名单训练 N-gram 模型供域名字符特征计算时使用。

第四步为特征的提取，利用训练好的 Word2Vec 模型，将二级域名作为输入，得到 200 维的向量特征；利用初加工的数据做统计，得到访问统计变量，并在此基础上计算得到衍生变量；利用训练好的 N-gram 模型得到相关概率特征，其中 n 的取值为 1、2、

3、4，并计算每组频率值的均值，中位数和标准差；利用二级域名本身得到熵的特征、长度、域名中字母数量、顶级域名等特征。

第五步为特征的筛选，按照 3.2.3 节中的特征分析方案，针对不同类型的特征，使用 AUC 和 KS 计算分析、相关性分析、信息值计算分析等方法进行特征的筛选，确认最终使用的特征，记录并保留下来。

4.3.2 流量选择

如图 17 所示，分别针对 DNS 放大攻击、随即子域名攻击、DGA 域名流量并行的操作。从特征获取 qar、qtr 带入到公式(3.5)中，其中 s1>α，我们认定为疑似放大攻击的流量，这时我们记录下这些域名。类似的从特征中获取 sdc、nxdr，带入公式(3.6)中，我们得到意思的随机子域名攻击相关域名。域名生成算法生成的域名我们使用预处理之后的流量，按照 3.3.4 中的方案，利用黑白名单分别计算每个二级域名的长度、熵值、n-gram，其中 n=2,3,4。分类器我们选择随机森林，对流量中的域名进行分类。之后根据得到的疑似域名清单，回查原始流量，将流量中包含这些二级域名的流量取出。

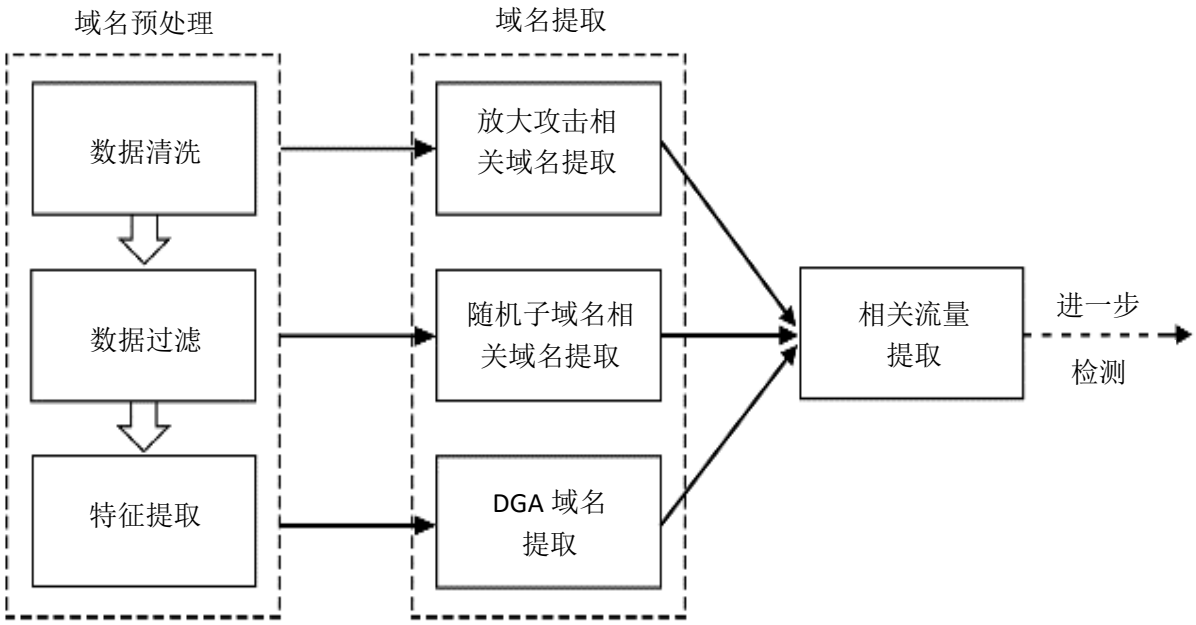

图 17 恶意流量提取流程图

4.3.3 流量监控

在获取了原始的被动 DNS 之后，需要一个切实有效的可视化方案。这里我选择使用 Influxdb 加 Grafana 构建一个流量监控系统。Influxdb 是一个开源的分布式时序、时间和指标数据库，使用 go 语言编写，无需外部依赖。它有三大特性：一是时间序列，因此与时间相关的函数使用非常的灵活（诸如最大、最小、求和等）；二是度量，可以对实时大量数据进行计算；三是事件，它支持任意的事件数据[35]。同时 Influxdb 还具有无结构（无模式）、扩展性强、支持大量统计函数、原生的 HTTP 支持、强大的类 SQL 语法等一系列优点。我选择 Influxdb 作为数据库也正是基于此。

图 18 流量监控展示

Grafana 是一个开箱即用的可视化工具，具有功能齐全的度量仪表盘和图形编辑器，有灵活丰富的图形化选项，可以混合多种风格，支持多个数据源特点。

4.3.4 恶意域名检测

图 19 所示为恶意域名检测图示，底层使用 HDFS 存储数据，通过 hive 来管理数据，利用 spark 在此基础上进行统计和计算。其中包含了域名字符特征和访问特征的统计以及词向量特征的计算。具体而言，恶意域名检测使用的分类算法为 xgboost，针对 DGA 域名，使用了筛选之后的全部的 295 维特征，包含了 word2vec 特征的 200 维特征、78 个访问统计相关的特征以及 17 个的域名字符特征。分别进行样本外验证和时间外验证，其中样本外验证完全使用 27 日数据，进行五折交叉验证，并进行三组对比实验，一组不包含访问特征，一组不包含字符特征，一组包含所有特征。而时间外验证是为了验证模型在时间维度上的衰减，训练集为 27 日数据，分别使用 28 日和 29 日数据进行验证，这里评价标准使用 AUC。对于色情域名特征只选择了 Word2Vec 特征，进行时间外验证，评价标准使用了 AUC 与 KS。

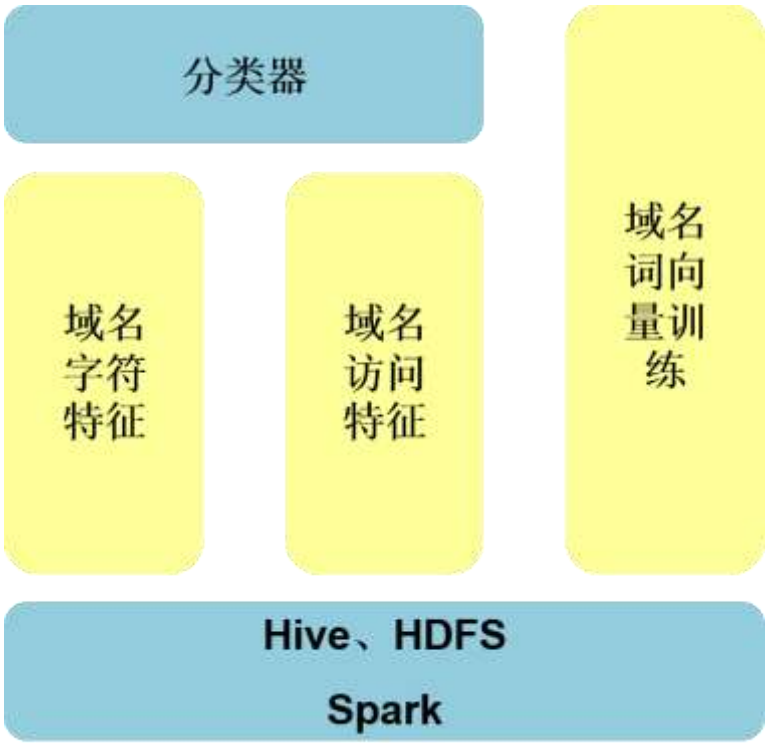

图 19 恶意域名检测图示

4.4 本章小结

本章是第三章所述方案的应用和验证，主要是针对被动 DNS 当中出现的恶意流量以及恶意域名，设计了提取和检测的模型，并设计实现了一个恶意域名检测应用原型系统程序，以验证方案的正确性和可行性。

第五章 实验结果与分析

本章主要对按照前文所述设计的原型系统所产生的实验数据和结果进行分析。

5.1 总体情况

根据第四章所设计的系统进行实验，使用被动 DNS 完成的实验结果将在本章完整的呈现出来。样本分析的部分，对共计三部分数据进行了整体的概括，初步做了统计上的分析。进一步的对恶意域名检测所需特征进行了分析，主要包含相关性测试、AUC 分析、KS 分析和信息值分析，并根据结果给出特征筛选结果。随后对第三章中所涉及到的参数，给出了实验结果，进行了参数的选择。最后一节为恶意域名检测的结果，分别在样本维度上和时间维度上对 DGA 域名和色情域名做了分类的实验，以 AUC 来验证结果，取得了很好的结果，证明所设计的算法具有优秀的效果。

5.2 样本分析

分别对恶意域名快速提取的样本和恶意域名检测的样本进行了统计上的分析，对于有标注的数据，给出了正负样本比例。恶意域名检测的样本为抽样之后的样本，训练集和验证集保持了相近的正负样本比例。

5.2.1 恶意域名快速提取样本分析

我们使用了来自 CNCERT/CC 提供的数据，包括了山西省的中国电信 PDNS 数据和广东电信的 PDNS 数据。其中山西省数据为 2015 年 10 月 15 日 23 个小时的数据，如图 20 所示，DNS 记录的总量接近 20 亿条，每小时不重复的二级域名数量在 10 万到 20 万之间，共标注 DDOS 相关恶意域名 101 个，DGA 相关域名 322 个。广东省的数据为 2017.4.14 9 点到 16 点共计 9 个小时的数据，如图 21 所示，DNS 记录总量达 11 亿余条，每小时不重复的二级域名数量达到 60 万左右，共标注 DDOS 相关恶意域名 163 个，DGA 相关域名 265 个。

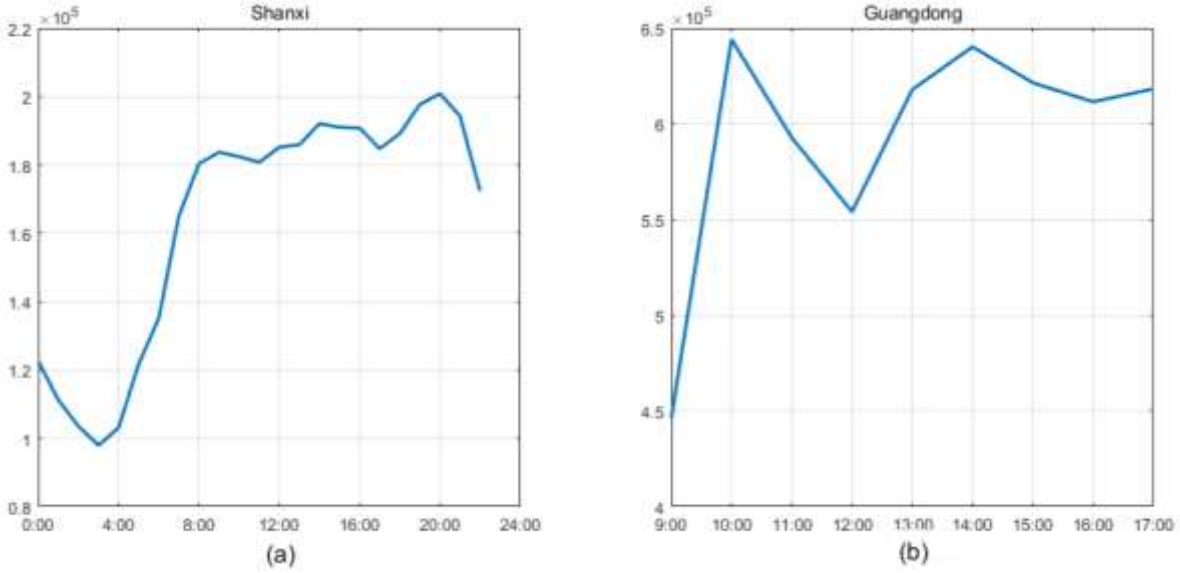

图 20　山西省和广东省电信不重复二级域名数量

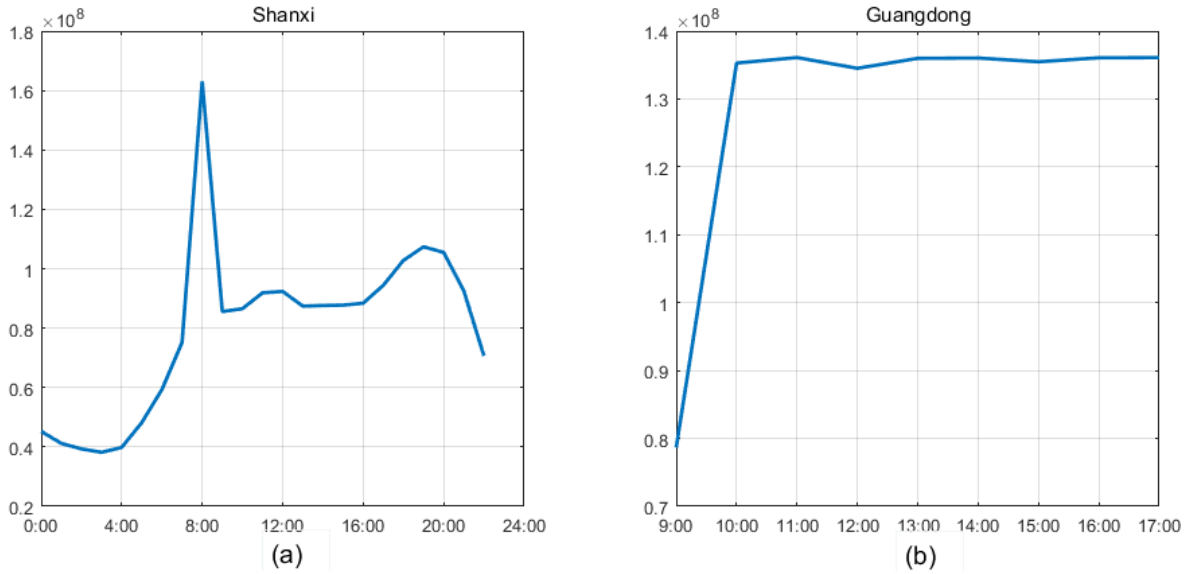

图 21　山西省和广东省电信 DNS 记录数量

在进行 DGA 流量的提取中，我们使用 alexatop100 万的域名列表作为白名单，黑名单我们将从 360 安全实验室下载的 DGA 黑名单作为黑名单，包含 1037304 条二级域名。

5.2.2 DGA 域名检测样本分析

由于原始数据时间跨度有限，因此无法在原始被动 DNS 数据中进行实验和验证，因此只保留了二级域名出现次数大于五次的域名。我们将 DGA 黑名单中出现的域名定为正样本，黑名单中未出现的域名视为负样本，正负样本比例大概为 1：20。正负样本过于不均衡，因此使用了下采样的方法，对负样本进行了抽样，保留所有正样本，负样本每天抽取 25000 个，使正负样本比例达到 1：5 左右。

表 8 DGA 分类样本概况

	20171127	20171128	20171129	合计
正样本	5211	6072	5451	16734
负样本	25000	25000	25000	75000
总样本	30211	31072	30451	91734
正样本比例	17.249%	19.542%	17.901%	18.242%

5.2.3 色情域名样本分析

表 9 色情域名分类样本概况

	20171127	20171128	20171129	合计
正样本	159	303	193	655
负样本	746	1493	997	3236
总样本	905	1796	1190	3891
正样本比例	17.569%	16.871%	16.218%	16.834 %

色情域名分类样本与 DGA 域名分类样本情况类似，使用相同的三天数据，采用下采样的方案，保留所有正样本。

5.3 特征分析

这部分内容是对各类特征进行初步的可用性分析，针对每一个单类特征进行分析，使用 AUC、信息值等指标来对特征的好坏进行一个评估，评估使用 2017 年 11 月 27 日数据。

Word2Vec 特征分析：使用 gensim 训练好的 Word2Vec 模型，随机挑选域名“*078mvrxcg4j3b49b.net*”，该域名属于 Chinad 家族木马产生的域名，表 所示为余弦相似度最高的十个域名。明显可见，虽然没有使用域名的字符特征训练余名，而只是利用访问序列训练出来的向量，具有很高的实用性。

表 10 “*078mvrxcg4j3b49b.net*”相似域名

域名	DGA 家族	余弦相似度
wbopgyg26s4jtdjb.net	Chinad	0.9648110270500183
o1cw6qug9ixwhwlo.info	Chinad	0.9643963575363159
60hwibc80hgdd853.net	Chinad	0.964003324508667

18o67n7qrnvoca9k.info	Chinad	0.9635913968086243
x9se3frkdzvhob2q.biz	Chinad	0.9630709290504456
rcsw95z947xdx09q.cn	Chinad	0.9628362059593201
ait1kf8sae7cdjhm.info	Chinad	0.9628314971923828
7l8zc789bucj50ky.org	Chinad	0.9625819325447083
1uw34d1ywa6pf3gg.org	Chinad	0.962489664554596
nexd7g076ppa6n6w.ru	Chinad	0.9624853730201721

进一步的将向量作为特征放入 xgboost 模型中，粗略调参，查看效果。这里 xgboost 具体参数设置为 nrounds = 500,max_depth = 6,eta = 0.02。使用五折交叉验证，将 27 日的向量数据分随机分为五份，其中四份用于训练，一份用于验证。图 22 所示为验证集的 ROC（Receiver Operating Characteristic）曲线，其中横坐标为 FPR（false positive rate），纵坐标为 TPR（true positive rate），这样不必设定阈值就可以看出这个分类器的性能，曲线下方的面积为 AUC（Area Under Curve）。AUC 是二值分类器非常重要的评价指标之一，他表示从样本验证集的正负样本中各取一个样本，其中正样本的模型预测结果数值大于负样本的模型结果预测数值的概率，如果是完全随机分布的话，AUC 的值应该约为 0.5，这里的计算结果为 0.915，可以看出效果极好。图 23 所示为验证集的 K-S 曲线，横坐标为阈值，纵坐标为在不同阈值下的 TPR 和 FPR，ks 数值为 TPR 与 FPR 差值最大的地方，即图中的蓝线，K-S 曲线可以很好的表示模型对正负样本的区分能力。这里 K-S 的值为 0.649，说明该模型在使用 Word2Vec 特征的情况下具有良好的正负样本分类能力。

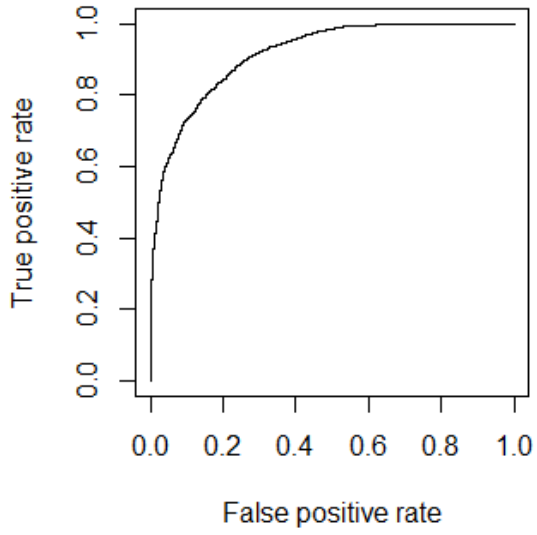

图 22 使用词向量预测 DGA 域名的 ROC 曲线

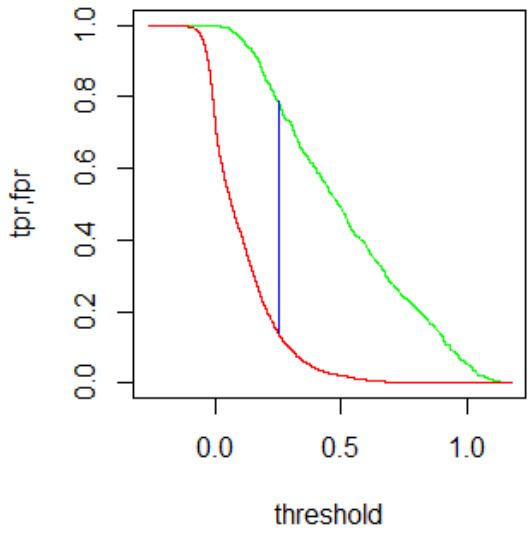

图 23 使用词向量预测 DGA 域名的 KS 曲线

访问记录特征：首先计算衍生变量和原本变量的简单相关系数，避免衍生变量与原本变量具有过强的相关性，造成特征的冗余。图 24 所示为相关系数的展示图，总体来说相关性较为稀疏，有部分变量具有明显的相关性，例如 qnonexistr（域名查询不存在比例）和 qsuccesser（域名查询成功占比）具有明显的负相关，qac（域名 A 记录查询数量）和 qsuccesser 具有较为明显的正相关。

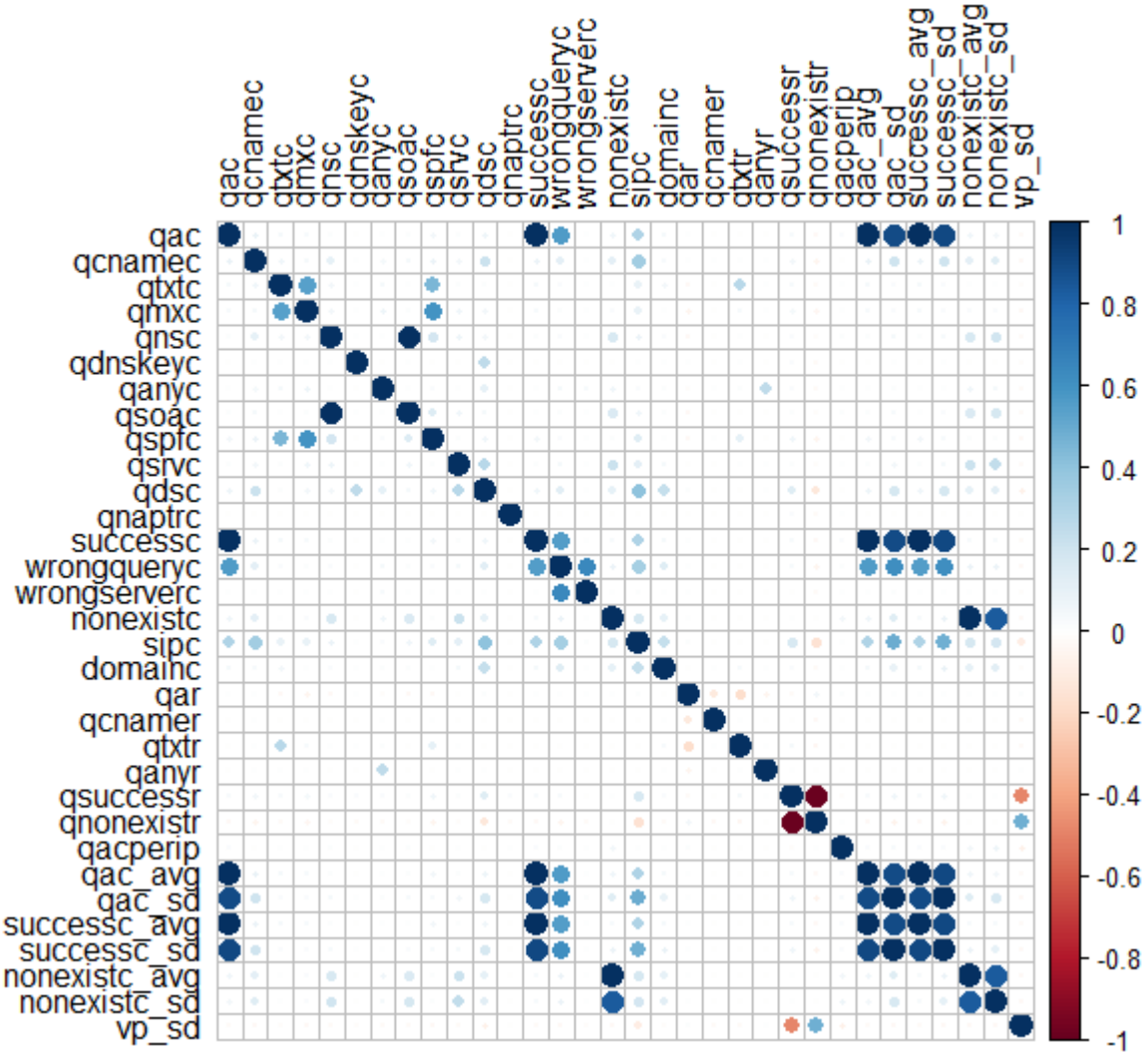

图 24　域名访问统计特征相关系数

图 25 为域名访问统计特征的 IV 值结果，共有 51 个预测能力好以及非常好的特征，具体来看图中的结果具有很强的可解释性。不存在域名数量的方差预测能力好的原因是是 DGA 域名的爆发时间具有一定的集中性，在一段时间内大量出现，而这些域名大多是不存在，造成方差大于正常域名；不存在域名访问数和不存在域名的均值预测能力强原因是 DGA 域名大多是 NxDomain，两者的 IV 值大致相等，同时由图 可见，这两个特征具有较强的相关性，因此是冗余的，入模时只需保留其一即可；各小时内的成功查

询数预测能力好是因为在这些时间段内 DGA 域名出现了活跃，合法非恶意域名可以正常访问与 DGA 域名产生了区别；每小时查询占比的方差预测能力强的原因是合法域名每个小时的访问量波动较小。同时该图也说明了 A 记录相关的特征具有较好的预测能力，而诸如 CNAME、TXT 等记录的特征效果很小，无论是每小时的或是整体的，同样的相关特征的衍生变量也有类似的表现。

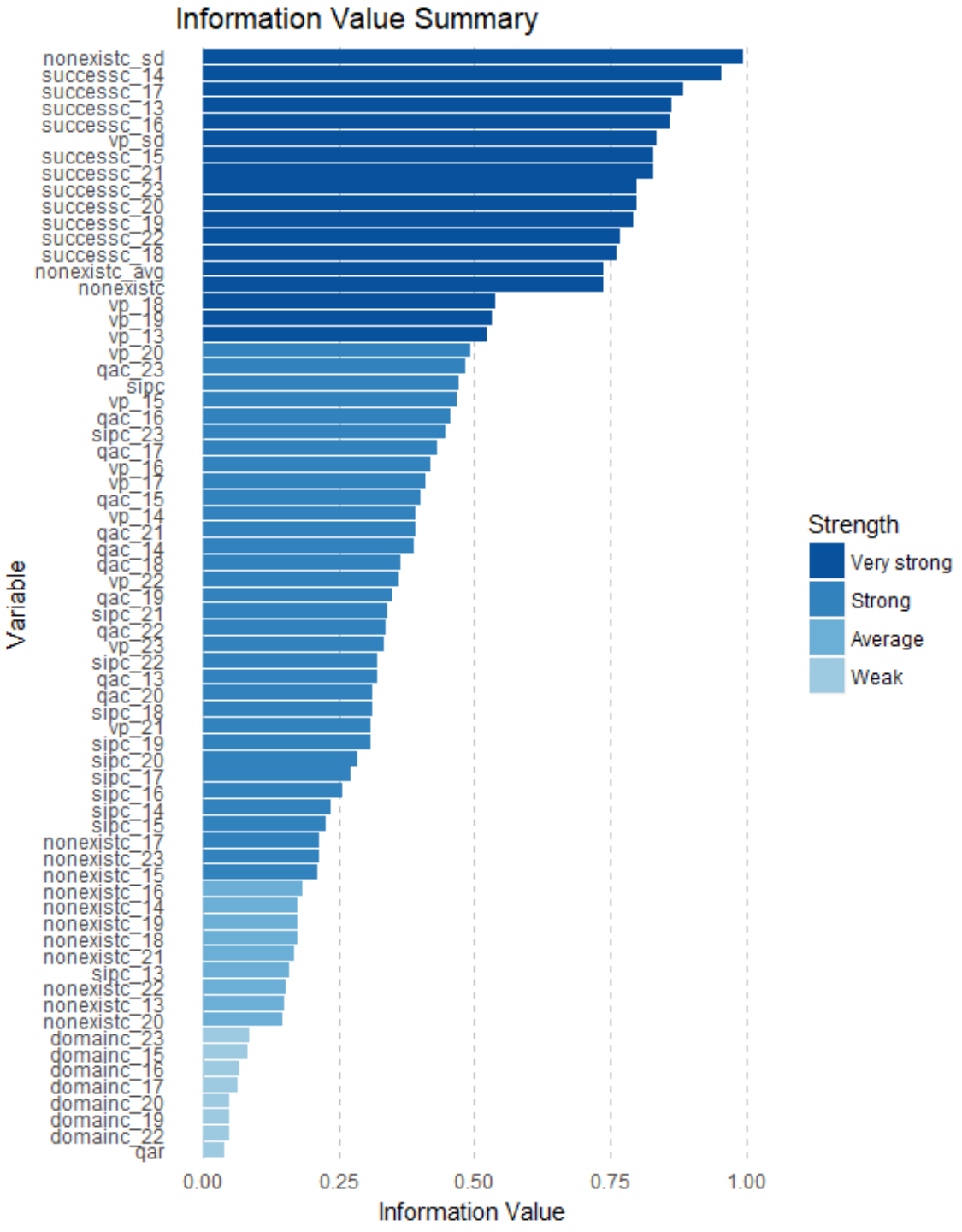

图 25　域名访问统计特征 IV 值结果

域名字符特征分析:同样的对域名的字符特征计算其信息值，查看这部分的特征对 DGA 域名的预测能力强弱。与 5.2.1 节中对域名熵值的分析类似，从图 26 中可以看出长度、熵值的确可以较好的区分正常域名和 DGA 域名。其次是 N-gram 系列的域名，信息值都达到了 0.2 以上，也具有好的预测能力。这部分特征与访问特征不同的一点是，访问特征的覆盖率达到了 100%，但是由于样本中少数域名长度较短，无法计算其 4-gram 系列特征，这相关的 3 个特征在样本中覆盖率为 99.89%。

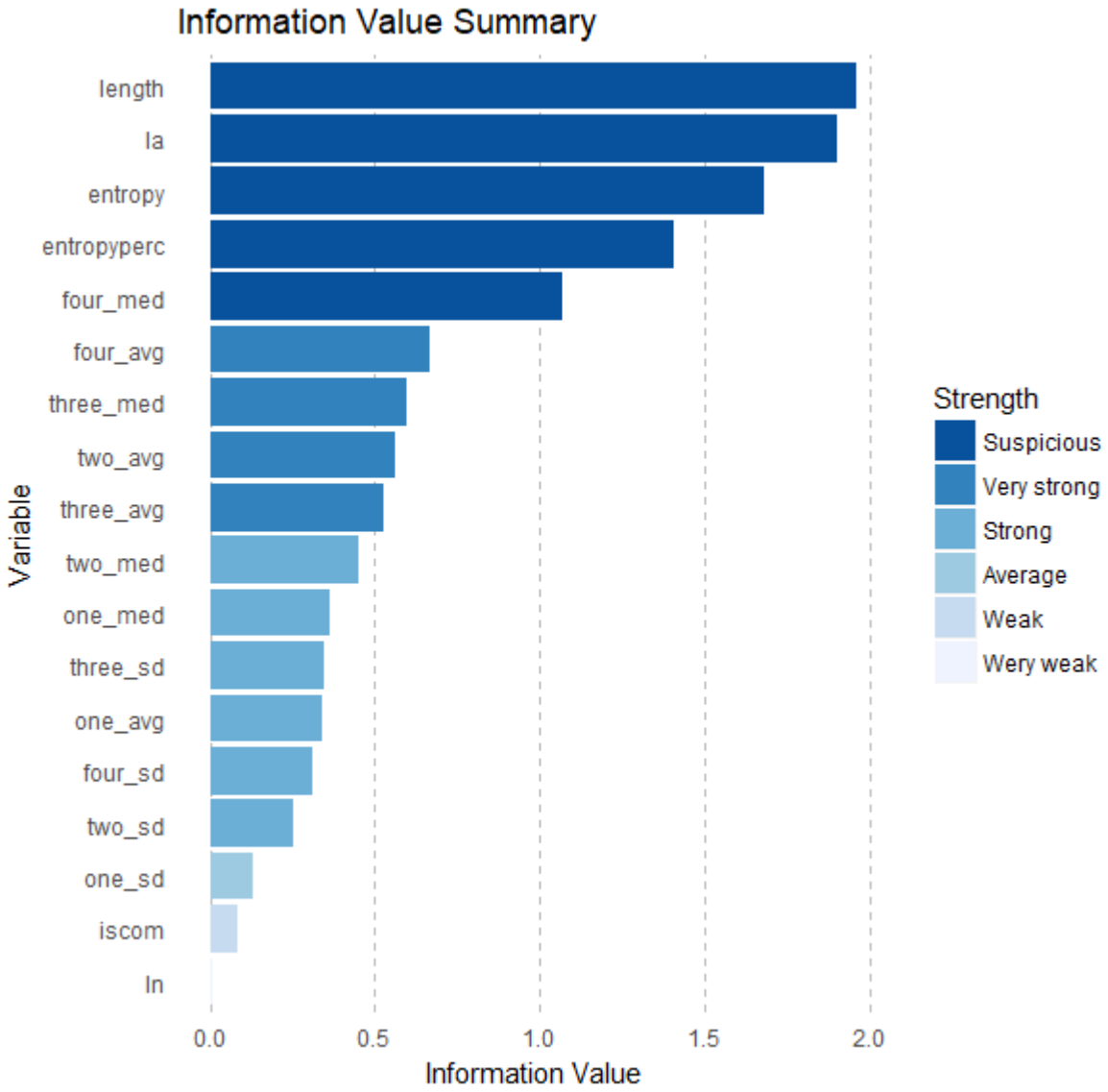

图 26　域名字符特征 IV 值结果

综上所述，无论是域名的访问特征或者是域名的字符特征在对于 DGA 域名的分析中都有非常显著的作用。这些特征之中，出现了一些冗余的特征以及预测能力极低的特征，因此在实验中会将这部分特征初步剔除，在 5.5.1 节中会对进一步讨论这些剔除的特征对实验结果的影响。当前 word2vec 特征的 200 维特征全部保留使用；164 个访问统

计相关的特征剔除预测能力差的以及冗余的特征后保留使用其中 78 维特征；18 维的域名字符特征剔除其中预测能力较差的数字数量特真正保留剩余 17 个特征，共计 295 维特征。

5.4 参数的讨论

本节当中将对第三章中所提到的参数进行讨论和试验，并给出分析结果。

5.4.1 针对 DNS 放大攻击提取相关参数

我们先 3.3.2 节中的参数进行讨论，这里的参数涉及到了三个，分别是时间间隔、判断域名是否可疑的阈值 α 、公式(3.5)中的参数 β 。其中时间间隔取值为 10、30、60、120，α 取值为 0、0.05、0.1、0.2、0.3、0.5、0.6， β 的取值为 0.07、0.1、0.15、0.2，如图 27 所示，z 轴代表召回率，当时间间隔为 10 时，召回率始终在 0.98 以上，当时间间隔为 120 时，召回率最高为 0.94，都无法满意的对 α 与 β 取值。表 1 展示了在不同参数的取值与可疑域名数量之间的关系。可以看到时间间隔选为 30 和时间间隔为 60 所能达到的效果差别不大，而前者的执行次数为后者的两倍，因此我们将时间间隔定为 60 分钟，为了获得尽可能小的域名数量，我们将 α 设定为 0.1， β 设定为 0.05。

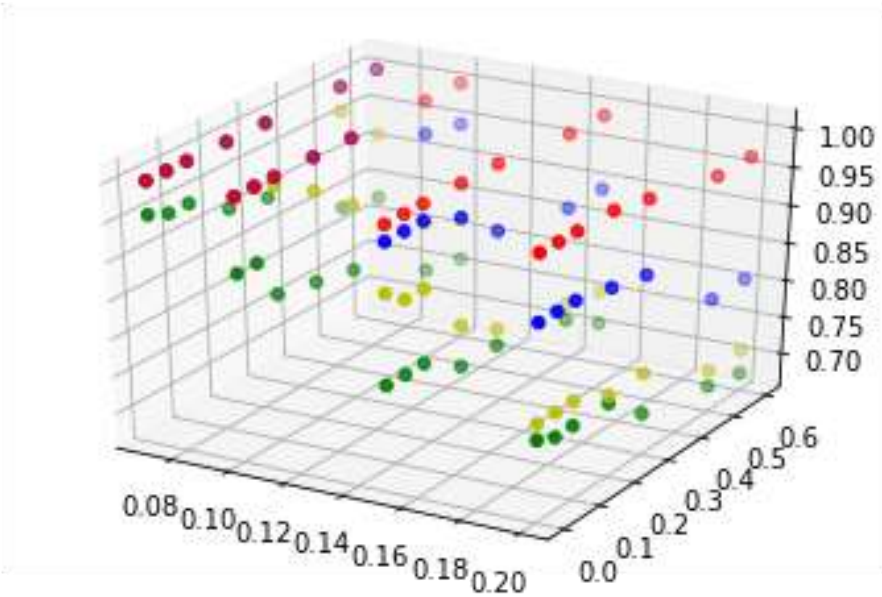

图 27　不同参数下 DNS 放大攻击域名提取效果

5.4.2 针对随机子域名提取相关参数

另一个需要讨论的参数是 3.3.3 节中公式(3.6)中的 θ，θ 的值越小，这个函数的曲线越平滑。如图 28 所示，描述了 θ 的取值和提取出来的域名的数量之间的关系。其中当 θ 为 0.1 时，召回率为 50%。当 θ 取值大于 0.2 时，召回率达到 100%。从图中可以看出 θ 的取值域名数量保持正相关，为避免过拟合，我们将 θ 取值为 0.3。

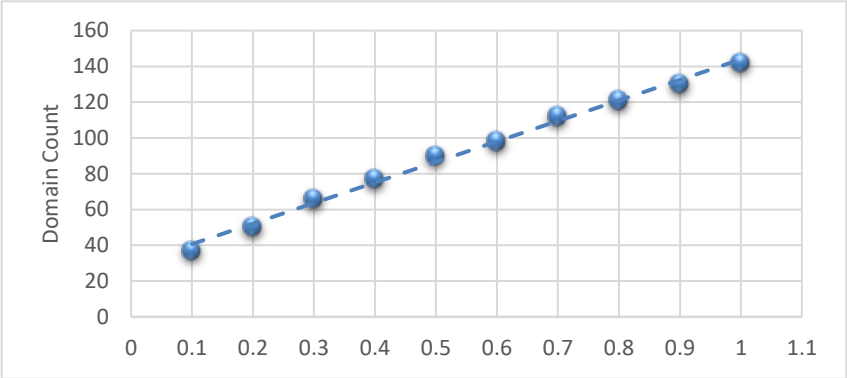

图 28　不同参数下随机子域名涉及域名提取效果

5.5 恶意流量提取结果

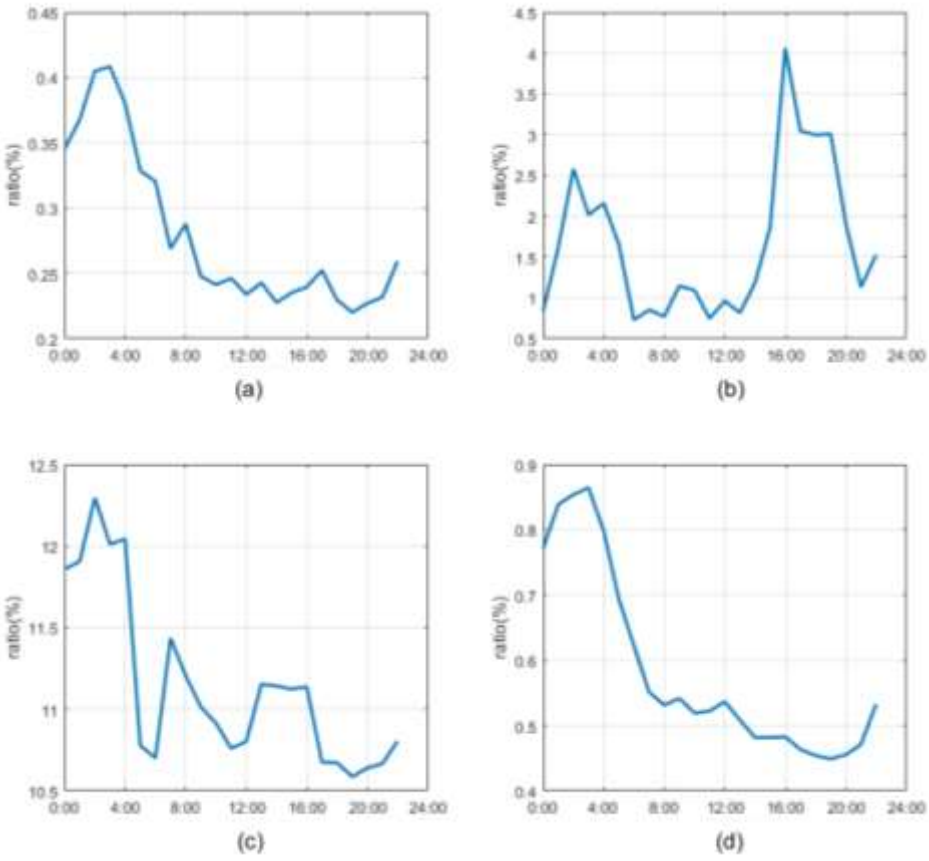

图 29　山西省电信恶意流量提取结果

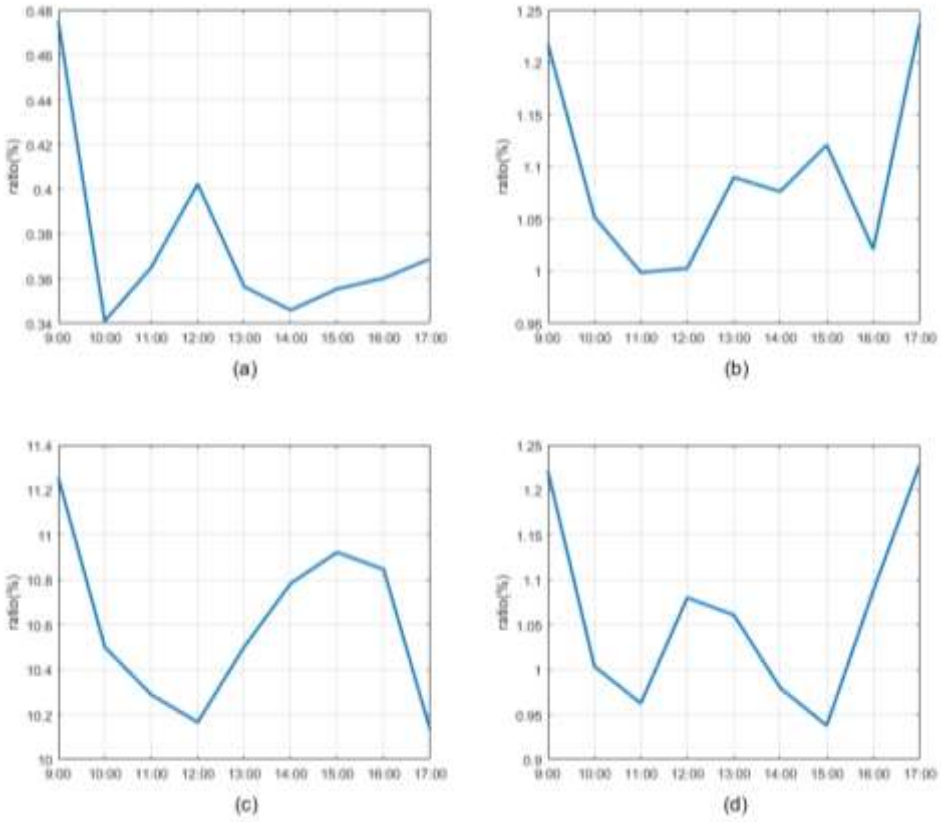

图 30　广东省电信恶意域名流量提取结果

图 29，图 30 为分别使用山西省和广东省电信的进行恶意流量提取得到的实验结果，参数使用 5.4 节中最后选择的参数。(a)图表示提取的随机子域名攻击涉及域名和 DNS 放大攻击涉及域名数量所占比例，(b)图表示提取的随机子域名攻击涉及域名和 DNS 放大攻击涉及流量所占比例，(c)图表示提取的 DGA 域名数量所占比例，(d)图表示提取的 DGA 域名流量所占比例。其中随机子域名和 DNS 放大攻击涉及域名提取的召回率达到 100%，DGA 域名的召回率达到 92%。

5.5 恶意域名检测结果

本节对 DGA 域名和色情域名的分类结果进行展示和分析，包含了样本上的验证结果以及时间上的验证结果。

5.5.1 DGA 域名检测结果

根据 5.3.1 中对特征的分析结果，现每个域名对应一个 295 维的特征向量，使用这些特征在训练集中来训练 DGA 检测模型并在验证集中检验模型效果，实验分为样本外的验证和时间外的验证，使用 AUC 来判断模型效果的优劣。旨在查看方法结果的有效性之外，观察模型效果在时间上的衰减情况。

样本外验证：分别使用以下三种情况的特征进行验证，一是不使用域名访问特征；二是不使用域名字符特征；三是使用所有特征。如图 31，DGA 域名检测样本外验证结果的 ROC 曲线所示，三种情况都得到了较好的结果，使用所有特征的结果显然要优于另外两种。

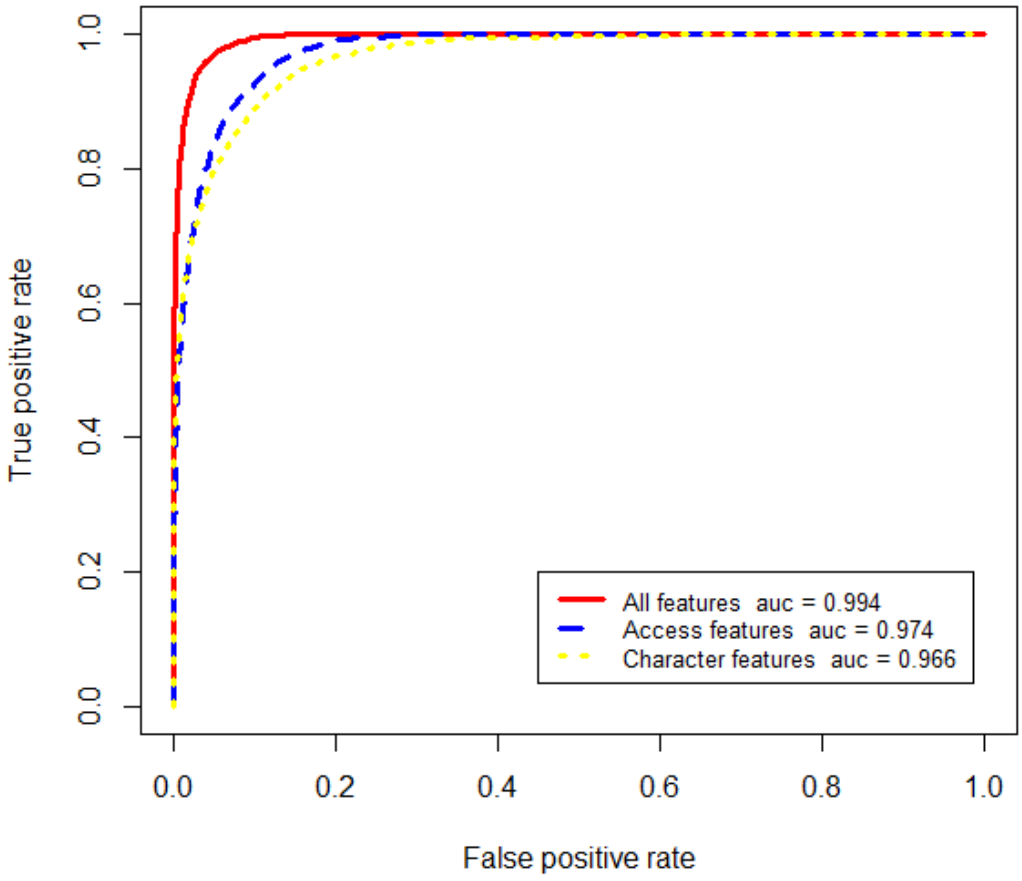

图 31 DGA 域名检测样本外验证结果

时间外验证：使用 27 日的数据作为训练数据，对 28 日和 29 日的数据进行验证，表 11 展示了 DGA 域名检测时间外的验证结果，分类效果随时间的衰减较为弱，连续两天皆取得了很好的结果。可以看见字符特征的效果衰减要小于访问特征。

表 11 DGA 域名检测时间外验证结果

	28	29
所有特征	0.983	0.977
只包含访问特征	0.939	0.92
只包含字符特征	0.963	0.96

5.5.2 色情域名检测结果

图 32 为色情域名检测在时间上的验证结果，首先看 AUC 与 KS 的结果，AUC 数值远高于 0.5，KS 的数值也都达到了 0.55 以上，可见模型的分辨能力和预测能力都很优秀。具体来看，在曲线的起始阶段，即 FP 为 0.0 时，TP 迅速的增大，证明在某一阈值下，该模型对色情域名准确率可以达到接近 100%。

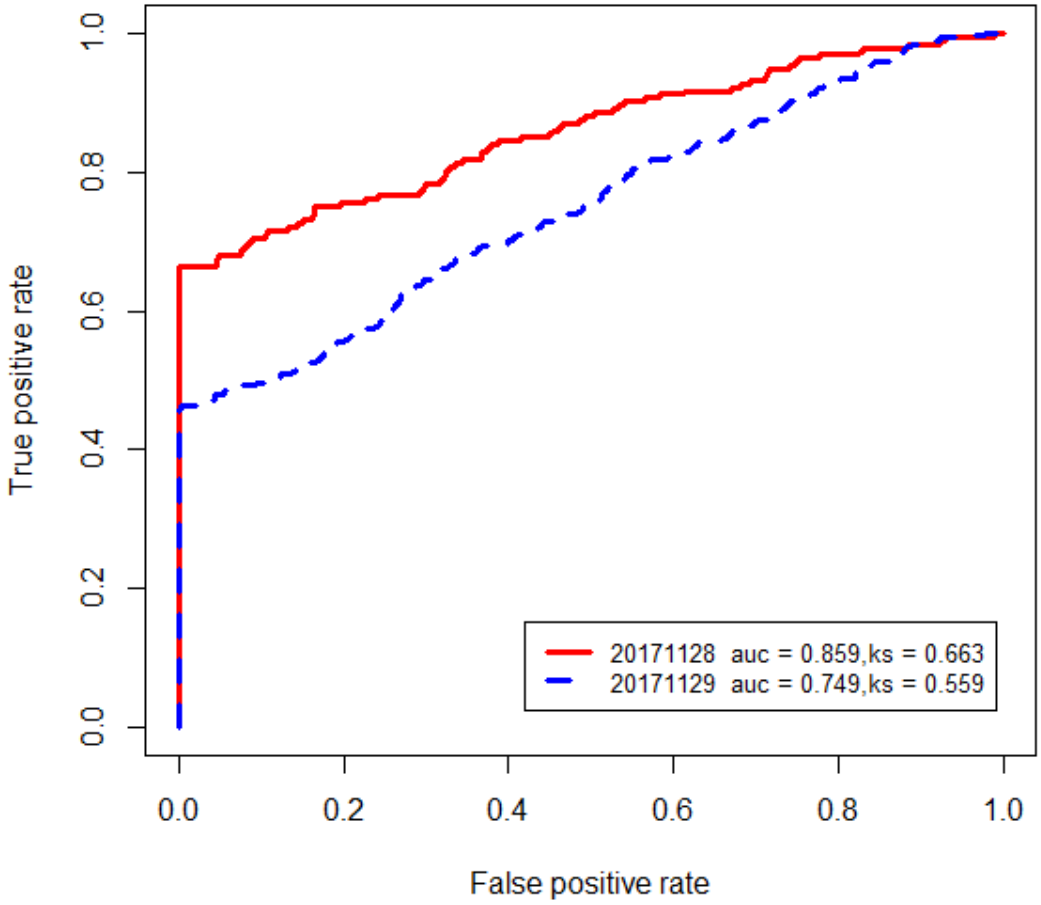

图 32 色情域名检测样本外验证结果

5.6 本章小结

根据之前两章所描述的解决方案以及系统原型设计，在本章中对实验结果进行了展示，并针对细节进行了分析和讨论，这其中包括了样本分析的结果、特征分析的结果、参数分析的结果以及最终恶意域名检测的结果。实验证明，该算法切实有效，取得了良好的效果，满足了 4.1 节中所提出的需求。

总结与展望

本文针对如何利用被动 DNS 从域名的角度保护网络安全的问题，首先分析了现有恶意域名检测技术的优缺点和技术应用发展状态，提出了快速提取恶意域名相关流量的需求和具体的恶意域名检测的需求。前者根据 DNS 放大攻击、随机子域名攻击、DGA 域名的不同特点，参考信誉系统提出了利用评分的提取方案；后者主要针对 DGA 域名和色情域名，综合考量访问特征和字符特征，提出了分类检测算法。最终原型系统实验的结果表明，本文提出的恶意域名流量提取方案切实有效的缩减待检测流量数量，分别针对 DGA 域名和色情域名的分类检测也取得很好的效果，很好的满足了利用被动 DNS 检测恶意域名的需求。具体而言，论文的主要研究成果包括三个方面，恶意域名相关流量的快速提取，利用被动 DNS 的 DGA 域名和色情域名的检测。

研究工作总结

本论文中涉及到的主要研究工作包括以下几点：

1、进行了充分的网络安全相关的调研，阅读了大量的文献，在深入了解网络、域名、DNS 系统的基础上，结合网络安全知识和当前流行的黑客攻击技术，确定了了本文的研究方向。在总结和对比了现有的恶意域名检测技术的基础上，提出了本文的研究目标和研究内容；

2、对恶意域名相关流量的快速提取进行了研究，针对恶意域名中在流量中占比最大的三类恶意域名，DNS 放大攻击涉及域名、随机子域名攻击涉及和 DGA 域名，借鉴了域名信誉系统的评分方法以及机器学习的方法，分别设计了有效的提取方案，并通过大量的实验对参数进行选择；

3、对于 DGA 域名的检测进行了深入的研究，分别利用域名的字符特征和访问特征进行实验，使用了协方差、IV 计算、AUC 分析、KS 计算等方式对特征进行了细致的分析。对于所选特征选择使用 GBDT 作为分类算法，进行了时间外的验证和样本外的验证，证明了特征选取以及算法方案的有效性以及表现了在时间维度上效果的轻微衰减；

4、针对色情域名的检测设计了基于 Word2Vec 的域名查询行为向量空间映射算法，该算法与现有的基于用户访问日志和基于网页内容的检测算法完全不同，通过对被动 DNS 中的域名记录按时间和用户 IP 进行整理之后，用深度学习的机制将域名嵌入到向量空间中，以这些向量为特征进行了时间维度上的验证，证明了该算法的有效性。

未来工作展望

在本文的工作基础上，未来可以针对以下几个问题展开进一步的深入研究：

1、本文实验中虽然进行了时间上的验证，但是受限于更多的数据，无法进行更多的时间上的衰减验证，无法确认训练好的模型随时间效果衰减的过程。而且本文中恶意域名检测实验中所使用的数据为周一、周二、周三三天的时间的被动 DNS 数据，都是周中的时间，并没有使用周末的时间。在实验中使用了大量的访问特征，主观来看周中的用户访问规律会和周末的用户访问有较大的区别，例如工作日的工作时间各类娱乐网站的查询比例相比于周末就会小很多，这个问题也是值得探索的；

2、研究过程中缺少长期连续性的数据的很重要的原因之一就是机器硬件的限制，想要克服这种限制的一种很好的方式就是找到一个更加合理的被动 DNS 记录存储方式，如果每一条记录都完整的存储下来，不但在统计时要花费较长的时间，而且占据的硬盘空间极大，例如广东省电信一天的被动 DNS 记录的大小就在 T 的数量级上。针对这一问题，简单的设想就是损失一部分的信息，例如一个 IP 对一个域名的查询一段时间内只保留一条记录，同时记录下查询次数，第一次和最后一次的查询时间。但是很明显这么做也存在一定的缺陷，可见找到一个合理的被动 DNS 记录存储方式还有诸多问题需要考虑；

3、在恶意域名流量的提取当中，也存在一些限制，由于没有长期的连续数据支持，很多统计特征没有办法提取出来，例如大多数 DGA 域名的生存周期都会小于 30 天，这其中的大部分域名生存周期不到 7 天，如果可以统计得到一个域名在之前的 7 天、30 天内是否出现过或者是否被成功的查询过，都可以很好的对模型进行补充。另外，这部分实验里对放大攻击涉及流量的提取中，没有一个好的办法将一些访问量高的网站剔除出来，可以从这一点入手，在保证召回率的同时进一步的减少提取出来的流量；

4、在针对恶意域名的检测当中，大量的工作都是针对域名的特征提取，而对于模型本身没有太过于关注，模型的选择是简单使用了当前流行的 Xgboost 分类模型，虽然取得了非常好的效果，但在时间维度上的效果衰减，我认为还存在着很多提升的空间，这时就可以考虑从模型选择入手，尝试模型组合或者现在深度学习中流行的 lstm 等；

5、在进行恶意域名检测的实验中，由于正负样本差距过大，我采用了负采样的做法，使得正负样本比例在一定的范围内，同时保证样本集和验证集的正负样本比例大致相当。但是这部分实验没有在真实的环境中测试，而真实环境中正负样本比例和实验的样本差

别很大，本实验只验证了方法的有效性，无法代表可以在生产中使用，如何将此方法投入生产环境，这一点也可以深入研究；

6、文中所做的恶意域名相关流量的提取，借鉴了信誉系统的建设模式，通过构建简要的指标进行筛选，同样也说明该方案具有完善成为一个完整的信誉系统的潜力。进一步的话，可以通过完善指标，使用更多的数据测试，构建橙一个完整的信誉系统。

参考文献

[1] 中国互联网络信息中心 、新华网等综合汇编.CNNIC 发布第 38 次《中国互联网络发展状况统计报告》[J].中国教育网络,2016(09):16.

[2] Bilge L, Kirda E, Kruegel C, et al. EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis.[C]// Network and Distributed System Security Symposium, NDSS 2011, San Diego, California, Usa, February -, February. DBLP, 2011.

[3] 中国互联网络信息中心.《2016 中国互联网网络安全报告》[J]. 中国教育网络,2017(03):16.

[4] Vissers T, Joosen W, Nikiforakis N. Parking Sensors: Analyzing and Detecting Parked Domains[C]// Network and Distributed System Security Symposium. 2015:53-53.

[5] Plohmann Daniel,Fkie Fraunhofer,Yakdan Khaled,Klatt Michael. A Comprehensive Measurement Study of Domain Generating Malware[C]// USENIX Security Symposium.2016:263–278.

[6] 杜云.《2015 全年 DDoS 威胁报告》报告[J].计算机与网络,2016,42(09):48.

[7] Attackers are increasingly leveraging large Domain Name System (DNS) TXT records in an effort to amplify the impact of their distributed denial-of-service (DDoS) attacks, Akamai's Prolexic Security Engineering and Research Team (PLXsert) warned on Tuesday[EB/OL],http://www.securityweek.com/large-dns-text-records-used-amplify-ddos-attacks-akamai, 2016-08.

[8] Antonakakis M, Perdisci R, Dagon D, et al. Building a Dynamic Reputation System for DNS.[C]// Usenix Security Symposium, Washington, Dc, Usa, August 11-13, 2010, Proceedings. DBLP, 2010:273-290.

[9] Antonakakis M, Perdisci R, Lee W, et al. Detecting malware domains at the upper DNS hierarchy[C]// Usenix Conference on Security. 2011:27-27.

[10] Antonakakis M, Perdisci R, Nadji Y, et al. From throw-away traffic to bots: detecting the rise of DGA-based malware[C]// Usenix Conference on Security Symposium. 2012:24-24.

[11] Perdisci R, Corona I, Giacinto G. Early Detection of Malicious Flux Networks via Large-Scale Passive DNS Traffic Analysis[J]. IEEE Transactions on Dependable & Secure Computing, 2012, 9(5):714-726.

[12] Bilge L. EXPOSURE : Finding Malicious Domains Using Passive DNS Analysis[J]. http://www.cs.ucsb.edu/~chris/research/doc/ndss11_exposure.pdf, 2011.

[13] Nelms T, Perdisci R, Ahamad M. ExecScent: mining for new C&C domains in live networks with adaptive control protocol templates[C]// Usenix Conference on Security. 2013:589-604.

[14] Rahbarinia B, Perdisci R, Antonakakis M. Segugio: Efficient Behavior-Based Tracking of Malware-Control Domains in Large ISP Networks[C]// Ieee/ifip International Conference on Dependable Systems and Networks. IEEE, 2015:403-414.

[15] Rahbarinia B, Perdisci R, Antonakakis M. Efficient and Accurate Behavior-Based Tracking of Malware-Control Domains in Large ISP Networks[M]. ACM, 2016.

[16] 张雪松,徐小琳,李青山. 算法生成恶意域名的实时检测[J]. 现代电信科技,2013,07:3-8.

[17] 张维维,龚俭,刘茜,刘尚东,胡晓艳. 基于词素特征的轻量级域名检测算法[J]. 软件学报,2016,09:2348-2364.

[18] Bsufka K, Kroll-Peters O, Albayrak S. Intelligent Network-Based Early Warning Systems[M]// Critical Information Infrastructures Security. Springer Berlin Heidelberg, 2006:103-111.

[19] Zdrnja B, Brownlee N, Wessels D. Passive Monitoring of DNS Anomalies[M]// Detection of Intrusions and Malware, and Vulnerability Assessment. Springer Berlin Heidelberg, 2007:129-139.

[20] Plonka D, Barford P. Context-aware clustering of DNS query traffic[C]// ACM SIGCOMM Conference on Internet Measurement 2008, Vouliagmeni, Greece, October. DBLP, 2008:217-230.

[21] Lee L, Luh C. Generation of pornographic blacklist and its Incremental update using an inverse chi-square based method[J]. Information Processing and Management, 2008, 44 (5):1698-1706

[22] Su Guiyang, Li Jianhua, Ma Yinghong, et al. A KNN algorithm on Chinese erotic text filtering [J]. Journal of Shanghai Jiaotong University, 2004 , 38: 86-79

[23] 曹建勋,刘奕群,岑荣伟,马少平,茹立云.基于用户行为的色情网站识别[J].计算机研究与发展,2013,50(02):430-436.

[24]周昌令,栾兴龙,肖建国.基于深度学习的域名查询行为向量空间嵌入[J].通信学报, 2016, 37(3):165-174.

[25]Tomas Mikolov. Word2vec project[EB/OL]. [2014-09-18]. https://code.google.com/p/word2vec/.

[26]MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in Neural Information Processing Systems. c2013:3111-3119.

[27]Pelleg, Dan, Moore, et al. X-means: Extending K-means with Efficient Estimation of the Number of Clusters[M]// Intelligent Data Engineering and Automated Learning — IDEAL 2000. Data Mining, Financial Engineering, and Intelligent Agents. Springer Berlin Heidelberg, 2000:17-22.

[28]Lund K, Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence[J]. Behavior Research Methods, Instruments, & Computers, 1996, 28(2): 203-208.

[29]Deerwester S C, Dumais S T, Landauer T K, et al. Indexing by latent semantic analysis[J]. JAsIs, 1990, 41(6): 391-407.

[30]Rohde D L T, Gonnerman L M, Plaut D C. An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence[J]. Communications of the Acm, 2005, 8:627-633.

[31]Le Q V, Mikolov T. Distributed Representations of Sentences and Documents[J]. 2014, 4:II-1188.

[32]张永铮,肖军,云晓春,王风宇.DDoS 攻击检测和控制方法[J].软件学报,2012,23(08):2058-2072.

[33]张洋,柳厅文,沙泓州,时金桥. 基于多元属性特征的恶意域名检测[J]. 计算机应用,2016,04:941-944+984.

[34]张永斌,陆寅,张艳宁.基于组行为特征的恶意域名检测[J].计算机科学,2013,40(08):146-148+185.

[35]ZERO,influxdb 快速入门[EB/OL], http://blog.csdn.net/zero__007/article/details-/52097696, 2016-08-02.

[36]Stover S, Dittrich D, Hernandez J, et al. Analisys of the storm and nugache trojans: P2P is here[J].The magazine of USENIX & SAGE, 2007, 32.18-27.

[37] Wikipedia, The storm botnet[OL], http://en. wikipedia.org/wiki/Storm_botnet, 2010.

[38] J.Williams, What we know (and learned) from the waledac takedown[OL]. http://tinyurl.com/ 7apnn9b, 2010.

[39] abuse.ch, ZeuS Gets More Sophisticated Using P2P Techniques[OL]. http://www.abuse.ch/ ?p=3499, 2011.

[40] Afek Y, Bremlerbarr A, Cohen E, et al. Efficient Distinct Heavy Hitters for DNS DDoS Attack Detection[J]. 2016.

[41] Weaver R, Collins M P. Fishing for phishes:applying capture-recapture methods to estimate phishing populations[C]// DBLP, 2007:14-25.

[42] Holz T, Gorecki C, Rieck K, et al. Measuring and Detecting Fast-Flux Service Networks[J]. Ndss, 2008, 1(5):487 - 492.

[43] Mattei T A. Privacy, Confidentiality and Security of Healthcare Information: Lessons from the Recent WannaCry Cyberattack.[J]. World Neurosurgery, 2017.

[44] Treseangrat K, Kolahi S S, Sarrafpour B. Analysis of UDP DDoS cyber flood attack and defense mechanisms on Windows Server 2012 and Linux Ubuntu 13[C]// International Conference on Computer, Information and Telecommunication Systems. IEEE, 2015:1-5.

[45] 曹玥, 李晖, 吕东亚. 基于 DDoS 的 TCP SYN 攻击与防范[J]. 电子科技, 2004(2):19-23.

[46] Izaddoost A, Othman M, Rasid M F A. Accurate ICMP TraceBack Model under DoS/DDoS Attack[C]// International Conference on Advanced Computing and Communications. IEEE, 2008:441-446.

[47] Buckingham J T, Mehr J D, Rehfuss P S, et al. Network domain reputation-based spam filtering: US, US 7487217 B2[P]. 2009.

[48] Csete M E, Doyle J C. Reverse Engineering of Biological Complexity[J]. Science, 2002, 295(5560):1664-9.

攻读硕士学位期间取得的学术成果

攻读博士学位期间取得的研究成果：

[1] Wenbo Wang,Tianning Zang, Yuqing Lan. The Rapid Extraction of Suspicious Traffic from Passive DNS[A].ICISSP 2018,2018(08).已录用

致谢

在北航的两年余的研究学习过程中，从开始的基础知识的学习，到接触安全相关的任务，再到利用国家互联网中心大量的数据进行研究。这是一个不断完善自我的学习水平、能力的过程，同时也是一个提升修养、培养品格的过程。这一路走来，靠的是我不断的坚持与努力，也是身边的老师、朋友、同学、长辈给予我的鼓励和支持，在这里我向你们表示由衷的感谢。

首先，我要感谢我的校内导师兰雨晴教授，兰老师的科研功底深厚，有着非常丰富的理论知识和实践经验，对很多研究领域都有很多新颖独到的见解。兰雨晴老师在操作系统领域有着深厚的功底，但是当我与他交流网络安全相关问题时，兰老师的很多建议都令我受益匪浅，眼界大开。兰老师不仅有很高的科研成就，而且其本人平易近人、温和儒雅，在北航的两年多期间，对我的校园生活也关怀备至。尤其是毕业论文撰写期间，老师不仅在论文选题和开题上给予了悉心的指导，在论文写作及实验中也经常督促我并提出改进意见，经常在百忙中过问研究进度，并在文章结构，语言表述等方面提出了很多有益的建议。没有老师的耐心指导和宝贵意见，这次的毕业论文将不可能如此顺利完成。在此我怀着感恩的心情，向兰老师致以最衷心的感谢！

同时我也要感谢国家互联网应急中心的两位老师，周渊博士和臧天宁博士。周渊老师可以说是我在网络安全道路上的领路人，正是周老师的指导，我这才确定了研究生期间的研究方向。虽然和周渊老师交流时间有限，每次和周老师交流，我都收获良多，他总会用很通俗易懂的例子帮助我理解一些晦涩难懂的理论和技术，启发我的思维，感受科研的乐趣。周老师开阔的眼界也对我产生了很大的影响，使我的眼界和思维不再局限于一个小的领域，帮助我在科研的路上成长的更快。臧老师则是在我论文和具体实验过程中对给予了我巨大的帮助，臧老师非常的务实，从他的身上我看到了属于科研工作者的那种努力专研的光芒，与他交流论文、交流实验、交流工程上遇到的点滴细节，可能不能每次都直接找到我想要的答案，但总能找到通向答案的那扇门。

其次，我要感谢实验室的师兄师弟们。感谢夏庆新、韩涛、白立旺、谢刚、运明纯诸位师兄师姐，是你们给我树立了榜样，也在我的论文完成过程中提供了许多建议，与你们交流使我避免了很多弯路。感谢 G308 一起努力的同学们，于思民、王铖成、房大梁、任冠华，是你们的不断鼓励和相互支持帮助，才会让我更好地完成我的毕业设计。感谢蒋波、谭伟良等师弟，一起科研，共同进步，建立了深厚的友谊。

再次，我要感谢我的女朋友左一萌。在我的忙碌的科研生活之外，你带给我了太多的色彩。长久以来，我也感觉到了对你的照顾不周，你却对我报以最大的包容和耐心，让我得以更大的精力投入科研工作当中。

我也要衷心的感谢我的家人、长辈，尤其是我的父母和我的姥姥。每一次通话我都能感受到那超越一切的亲情，你们每次都像汇报任务一样告诉我身体都好，家里也安好，就是怕我分心于此，你们是我最坚强的后盾。

最后感谢评审老师在百忙之中抽出时间评阅我的论文。