域名画像系统的设计与实现*

包正晶[1],苏马婧[1],康 彬[2],薛继东[1],刘 红[1]
(1.华北计算机系统工程研究所,北京 100083;2.中国人民解放军 96941 部队,北京 100080)

摘 要: 网络空间逐渐成为人类生产活动的第二空间,网络空间测绘对人们了解认识网络空间资源分布、网络关系和威胁情况等具有重要意义。当前对域名这一网络空间重要资产的测绘研究相对较少,因此针对域名资产进行探测分析,结合多源域名数据对域名的基础属性、谱系关系、规模状况和时空变化等情况进行分析,形成域名画像。该研究有助于用户掌握互联网域名整体发展情况,可对网络流量过滤和恶意域名检测、网络空间资产属性识别等提供支撑。

关键词: 域名画像;网络空间测绘;谱系构建;时空变化分析
中图分类号: TP311.1 **文献标识码:** A **DOI:** 10.19358/j.issn.2096-5133.2021.06.001
引用格式: 包正晶,苏马婧,康彬,等. 域名画像系统的设计与实现[J].信息技术与网络安全,2021,40(6):1-8.

Design and implementation of domain name portrait system

Bao Zhengjing[1], Su Majing[1], Kang Bin[2], Xue Jidong[1], Liu Hong[1]
(1.National Computer System Engineering Research Institute of China, Beijing 100083, China;
2.Unit 96941 of PLA, Beijing 100080, China)

Abstract: Cyberspace has gradually become the second space for human production activities. Cyberspace surveying and mapping is of great significance for people to understand the distribution of cyberspace resources, network relationships and threats. However, there are relatively few researches on surveying and mapping domain names, which are important assets in cyberspace. Therefore, this article conducts detection and analysis on domain name assets, and analyzes the basic attributes, genealogical relationships, scale status, and temporal changes of domain names based on multi-source domain name data to map domain name portrait. The research can help grasp the overall development of Internet domain names, and can provide support for network traffic filtering, malicious domain name detection, and network space asset attribute identification.

Key words: domain name portrait; cyberspace mapping; pedigree construction; spatiotemporal change analysis

www.pcachina.com

0 引言

随着网络技术的飞速发展,网络空间逐步成为人类生产活动的第二空间,网络空间测绘也逐渐成为学术界研究的热点。网络空间测绘旨在将网络空间、地理空间和社会空间进行相互映射,绘制一份动态实时可靠的网络空间地图[1]。当前网络空间测绘以面向实体资源测绘的IP资产属性、地址位置、网络拓扑关系的研究和以面向虚拟资源测绘的人物画像、服务画像等为主。

域名的相关研究集中在域名分类研究[2-3]、域名安全性研究[4-9]、恶意域名检测[10-12]、域名发展情况及现状的研究[13-15],对域名的全面刻画和动态刻画的研究还相对较少。然而通过对域名的属性刻画和发展趋势研究能够更好地了解和认识网络空间域名的分布情况、域名的规模、域名间的相互关系、域名的历史变化情况,有助于间接了解互联网整体的发展情况,可为恶意域名识别、恶意流量监测、流量访问控制等提供支撑。

因此,本文提出了域名画像这一概念,设计并实现了一套域名画像系统,涵盖域名基础属性识别、谱系特征识别和时空变化特征识别,对网络空间测绘具有重要作用。本文的主要研究内容和成果

* 基金项目:国防基础科研计划项目(JCKY2019211B001)

如下：

(1)本文提出了域名画像概念，从名称、域名证书、注册时间、到期时间、对应证书、状态信息、域名所有者、解析路径、历史解析情况、谱系关系、位置分布情况等 20 个维度实现对域名刻画；

(2)设计并实现多源域名数据获取模块，实现 18 亿域名(含子域名)自 2019 年 12 月至 2020 年 12 月的历史解析信息的获取，2 000 多万域名的注册信息、证书信息、主页信息等基本信息的获取；

(3)设计了域名谱系构建方案，实现了 200 个顶级域名的谱系构建，为域名规模的分析提供数据支持；

(4)设计并实现了域名时空变化分析流程，对全球 18 亿域名数据进行时空变化分析。

1 域名画像的概念

1.1 域名画像的含义

定义 1(域名画像)：将域名抽象成基础属性、谱系关系、时空轨迹等一系列相关属性的方法，是采用多维属性信息描述域名的模型，如图 1 所示。

定义 2(域名基础属性)：域名基础属性是刻画和描述域名某一时刻静态特征的集合，包括域名对应 IP 地址、域名拥有者、注册时间、到期时间、对应证书、证书加密方式、状态信息、邮箱信息、更新时间、注册链接等相关属性信息。

定义 3(域名谱系关系)：域名谱系关系是指通过域名谱系、域名同源特征、域名解析路径等属性描述域名的渊源关系。域名谱系是指根据域名产生的渊源关系、所有者关系、解析路径关系对域名进行分类，构建如图 2 所示的谱系关系图。其中根域名是指域名分层结构中最高层级的域，用一个点表示，在使用过程中一般不做显示；顶级域作为根域的下一层级，也称一级域名，一般按国家划分或按组织性质划分，按国家划分一般使用国家代码作为域名，例如美国、中国、日本、俄罗斯、法国分别使用 us、cn、jp、ru、fr 等字母表示，按组织性质划分一般使用能够代表组织或机构简写的字母作为域名，

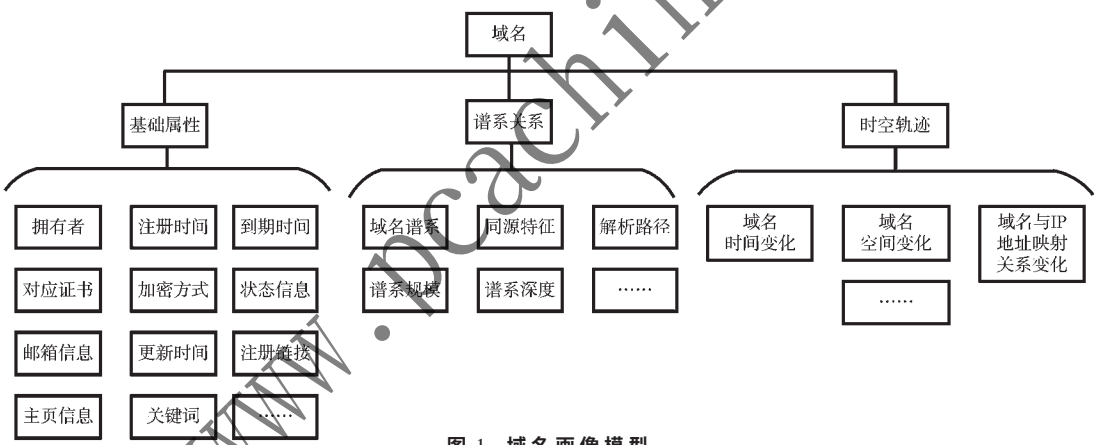

图 1 域名画像模型

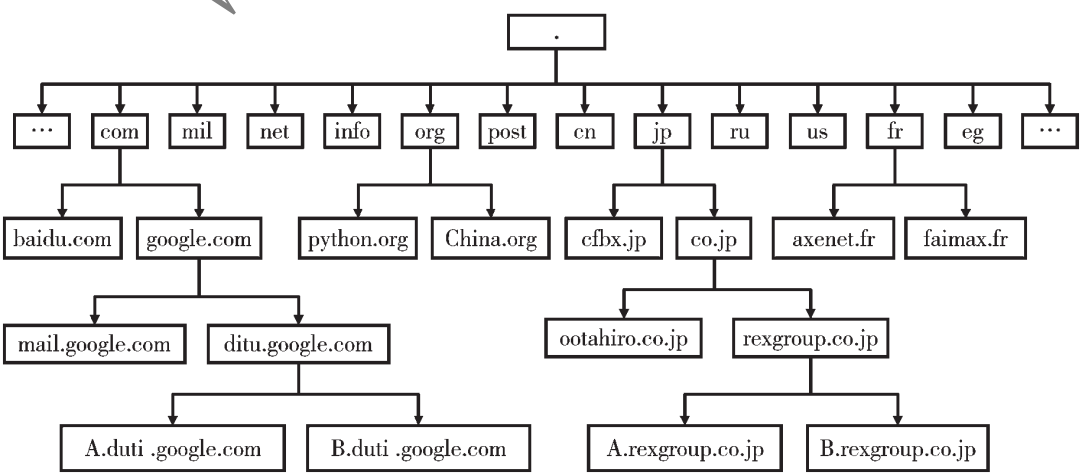

图 2 谱系关系示意图

www.pcachina.com

例如 com 表示商业机构，org 表示非营利性组织等；二级域名是顶级域的下一层级，是公司、组织、个人都可以注册的普通域，例如 baidu.com，google.com；三级域名是在二级域名的基础上添加一些字符，用于对二级域名进行扩展的域名，例如 map.baidu.com。域名的同源特征是指两个或多个域名具有共同的祖先域名节点。域名的解析路径是指 DNS 在实现域名与 IP 相互映射关系时先后请求的所有服务器及 IP 地址，将服务器、IP 被请求顺序记录下来就得到该域名的解析路径。

定义 4(域名时空轨迹)：是指通过域名时间变化、域名空间变化、域名与 IP 地址映射关系的变化描述域名随时间所产生的数量变化、空间位置变化及解析 IP 地址的变化关系等，是域名的动态刻画。

1.2 域名画像的目标

域名画像的目标是快速、大规模地获取各类域名相关数据，使用多维度特征对域名全方位刻画，为域名领域概况研究提供支撑。包括以下几方面：

(1)全面掌握互联网域名的发展规模、行业分布，不同顶级域名所包含子域名数量等特征。

(2)全面获取域名的基础属性信息。

(3)针对不同来源、不同时刻、不同方式获取的多源、动态数据进行属性抽取，并对比不同来源信息的真实性、时效性，确保属性抽取结果具有真实、可靠、时效性等特征，对不同属性进行融合分析，形成对域名的认识和知识表达。

(4)域名历史分布变化和数量增减数据统计，对域名动态进行跟踪，识别时空变化、预测发展趋势。

1.3 域名画像难点与挑战

(1)大规模域名的发现

域名画像需要全面获取当前互联网中可用域名信息，然而由于域名的动态性、广泛性等特性以及缺乏有效的索引机制，快速全面获取域名全集是域名画像首要解决的问题，例如 google.com 所包含的二级域名和三级域名数量随着 Google 业务数量或者业务场景的变化呈现动态增长或消亡的特征，对 Google 域名全集的获取带来很大的困难。

(2)域名属性填充问题

为满足域名基础属性、谱系属性、时空轨迹等属性的填充，需要对海量域名的 Whois 信息、证书信息、解析路径等信息进行获取。由于域名数据量大、属性特征复杂，因此需要对互联网中不同网站的信息进行主动请求，并且在域名数据获取之后需要具备一定的自然语言处理或者行业专业知识的人员，对信息正确性、完整性、时效性以及重复率进行分析推断，以满足域名属性的时效性、完整性和准确率。

(3)大规模数据存储问题

由于域名画像数据具有数据量大、数据描述维度多、时空变化动态性强等特点，需要针对性地设计数据存储方式和存储结构，以提高查询检索和数据分析的效率。

(4)面向谱系、时空变化等海量数据的分析

为实现域名谱系和时空变化分析，需要对所有域名解析路径进行获取和持续监测，由于域名(含子域名)数量巨大，导致所有域名进行解析路径获取和历史解析数据分析开销较大，另外对海量域名历史解析数据分析需要基于域名多维特征对域名进行排序，保证分析优先级，给系统性能带来很大挑战。

2 系统设计与实现

本文设计并实现了一套大规模域名画像系统，该系统由数据层、业务逻辑层、表示层三层组成，系统流程如图 3 所示。为实现域名画像目标，本文以主动探测和获取开源数据相结合的方式，获取域名的基础属性信息、关联 IP 信息、网站信息、域名解析信息等各类基础数据，利用域名谱系识别技术和域名时空变化分析技术，形成大规模域名的多维画像库；在此基础上，对域名资产数据进行了初步分析，通过域名整体概况了解互联网域名发展状况、分布情况和变化情况。

本文系统中，数据层主要获取域名的历史解析数据、Whois 信息、域名的证书信息、IP 定位信息、域名解析路径等信息，为后续分析提供数据支持；业务逻辑层实现域名的谱系识别和域名的时空变化分析，域名谱系识别主要通过对域名所有者信息的统计、域名证书一致性判断、解析路径相似性匹配、域名字符串层次结构划分等多种方式实现，域名的时空变化分析主要以域名对应 IP 地址的变化和域名空间变化分析为主；表示层是基于业务逻辑层的分析结果进行呈现；直观清晰地表示域名属性信息、谱系规模和域名的时空变化情况。

2.1 多源域名数据获取系统

为解决前文提到的大规模域名发现和域名属

www.pcachina.com

性填充所带来的困难，本文设计了如图 4 所示的多源域名数据获取系统，使用第三方数据和主动探测相结合的方式实现了大规模域名的发现，并对通过主动请求获取的域名 Whois 信息、证书信息、解析路径、解析 IP、主页信息等域名相关信息，完成域名基础属性的填充。系统主要包含以下三个模块：

(1)域名发现模块：该模块以 ICANN 组织官网的域名相关文件、顶级域名的区域文件、反向 DNS 记录等第三方数据作为大规模域名获取的基础；通过对网络 IP 地址存活情况扫描，并对存活 IP 地址和端口进行访问，以获取存活 IP 地址对应域名的方法和网站相关链接嵌套爬取的方法作为大规模域名获取的补充。通过第三方数据和主动探测数据相结合有助于获取到全面的域名数据集，应对大规模域名发现的挑战。

(2)域名静态信息获取模块：该模块主要实现 3 个功能，一是实现以 Whois 信息为基础的属性填充，请求互联网中各大网站提供的 Whois 数据库，获取域名对应的所有者信息、注册时间、到期时间、状态信息；二是实现以域名解析为基础的属性填充，为每一个域名进行迭代查询，并记录查询结果和所请求的域名服务器的地址及先后顺序，完成域名解析

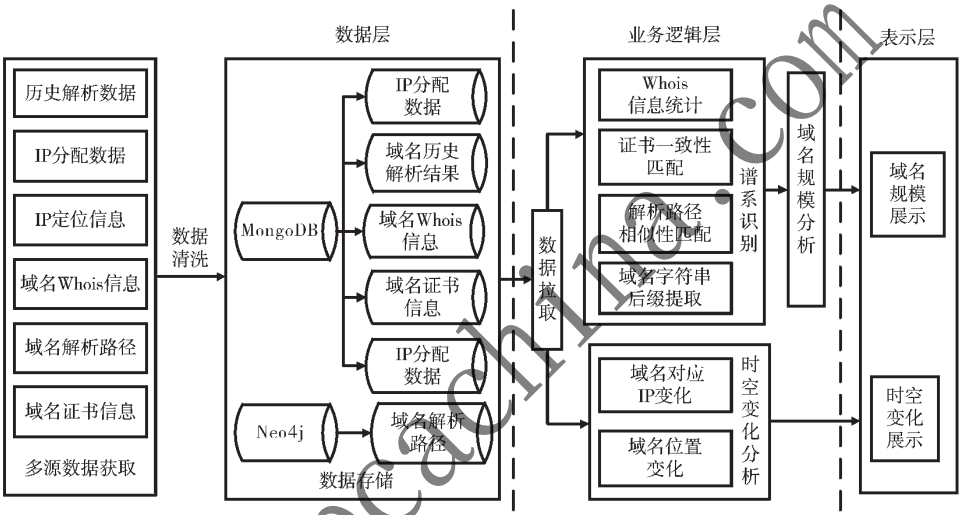

图 3 域名画像实现流程

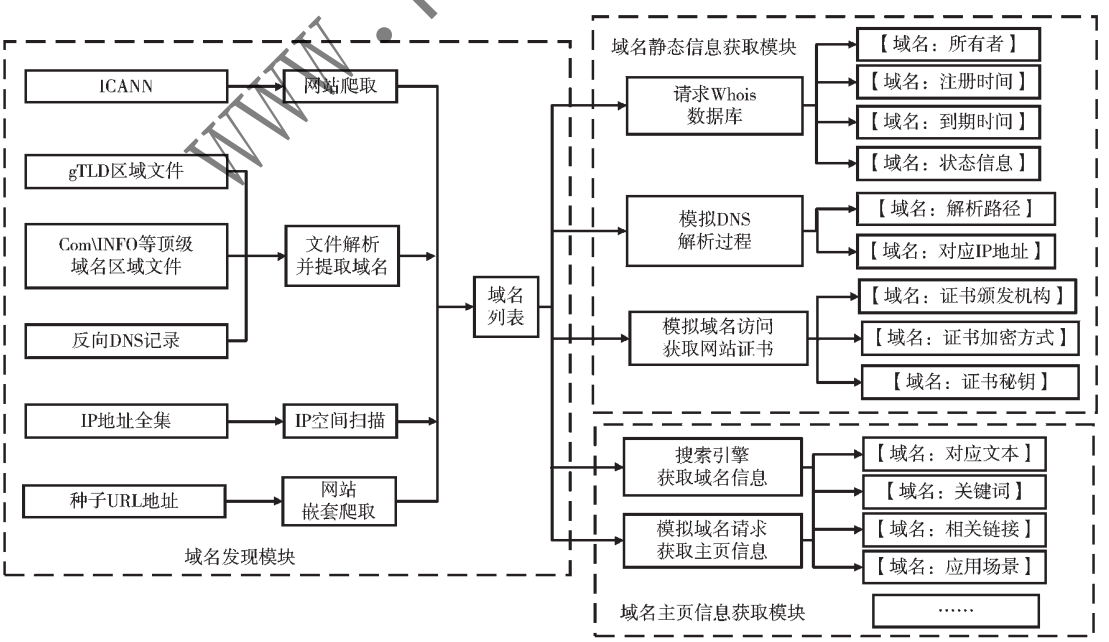

图 4 多源域名数据获取模块

www.pcachina.com

路径和对应 IP 地址属性的填充；三是实现域名对应证书信息的获取，模拟浏览器访问域名过程中主机与网站服务器证书交互过程，获取域名对应证书，实现证书的颁发机构、加密方式、证书秘钥等属性的填充。

(3)域名主页信息获取模块：该模块实现对域名相关主页信息的获取，主要有两种方式，第一种使用 Google 搜索引擎对域名进行搜索，获取互联网中域名相关文本信息；第二种使用模拟访问域名主页的方式，获取主页信息的文本、关键词、相关链接、应用场景等，完成域名主页相关属性的填充。

2.2 域名谱系构建

针对主动域名关系爬取受限、域名解析路径覆盖率低的问题，通过对域名所有者信息的统计、域名证书一致性判断、解析路径相似性匹配、域名字符串层次结构划分等多种方式实现域名谱系构建，满足用户获取域名站点组织结构关系、域名递归解析关系的需求。在域名谱系构建的基础上实现对域名规模分析、同源特征分析等目标。域名谱系关联分析流程如图 5 所示。主要步骤如下：

(1)从 MongoDB 数据库中获取域名的 Whois 信息，并提取其中的 name、email、registar、domain 等字段，根据这些字段对整体域名集进行统计，以统计结果作为域名谱系关联分析的依据。

(2)在 Whois 信息无法满足域名谱系构建需求时，基于域名证书一致性，对相同证书的域名进行匹配，以匹配结果实现域名的谱系分析。

(3)在证书匹配之后还存在着部分证书没有获取到证书信息的域名，对于这部分域名以解析路径相似比对结果作为域名谱系构建的依据。

(4)在以上三种方案都没有实现域名谱系构建的情况下，本文以域名字符串本身后缀一致性比对实现域名谱系构建。

2.3 域名时空变化分析

针对域名时空变化和发展趋势缺乏分析的现状，从域名数量随时间变化和空间位置变化方面进行分析，为重点域名时空轨迹跟踪提供数据支撑。

通过对域名历史解析数据的获取，抽取域名首次出现时间和域名最后一次出现时间，从而分析域名数量随时间的变化关系；从历史解析数据中获取域名对应 IP 地址和该解析记录的时间戳，分析域名对应 IP 地址的变化及对应 IP 地址数量的变化，从变化情况判断域名是否部署在 CDN 上，结合域名对应 IP 地址的地理位置分析域名位置分布特征等信息。总之，从时间维度而言，以域名对应 IP 数量、域名新增数量、域名消亡数量等为主来进行分析；从空间维度而言，以国家或地区域名分布数量为主进行分析，掌握域名数量、对应 IP 地址、空间变化等特征随时间的变化规律。

对域名时空轨迹分析流程如图 6 所示，主要步骤如下：

(1)对域名解析记录进行数据清洗，提取域名、

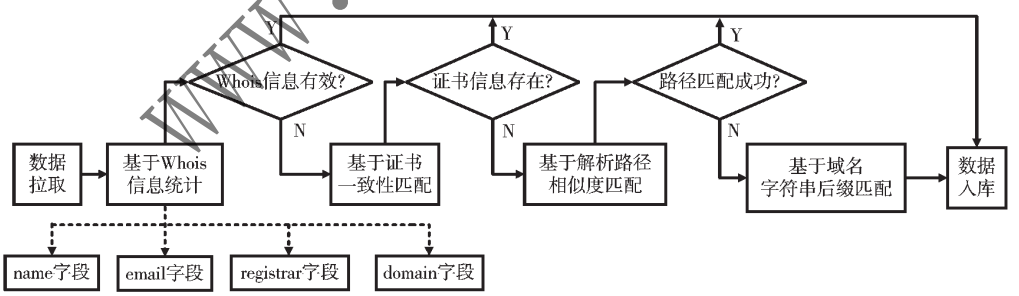

图 5 域名谱系关联分析流程图

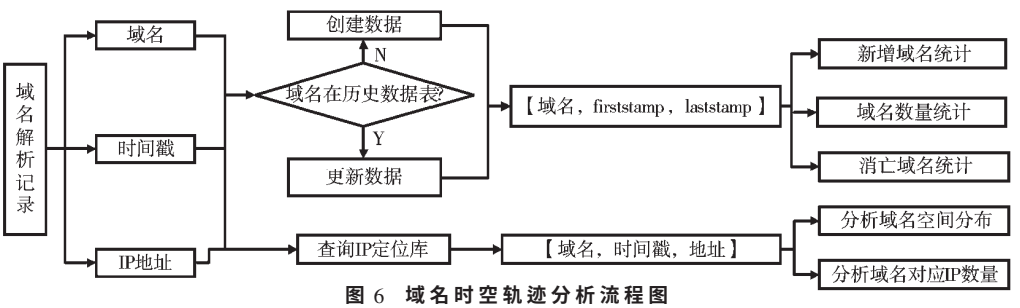

图 6 域名时空轨迹分析流程图

www.pcachina.com

时间戳、IP 地址等字段。

(2)判断当前域名是否存在于域名历史数据表中,如果不在则创建当前域名的记录,如果在则更新当前域名的记录,输出域名、初次出现时间、最后一次出现时间的一条记录,即【域名,firststamp,laststamp】。

(3)通过查询 IP 定位库,获取域名对应 IP 地址所在的物理地址,结合历史解析记录中的时间戳,输出域名、时间戳、物理地址对应的一条记录,即【域名,时间戳,物理地址】。

(4)基于海量【域名,firststamp,laststamp】记录分析某一时间段内新增域名数量、消亡域名数量及域名整体数量等。

(5)基于某一域名的【域名,时间戳,物理地址】记录分析该域名过去一段时间内对应 IP 数量变化和物理地址的变化,预测域名的发展趋势。

3 数据分析结果

3.1 顶级域名谱系识别结果

顶级域名的子域名数量存在较大差别,例如 com 后缀的域名数量明显比 post 后缀域名数量多。为验证域名对应子域名规模的分布规律,本文在域名谱系构建的基础上结合位置信息,研究新顶级域名和通用顶级域名所包含子域名的数量和分布情况。

本文以顶级域名的谱系识别结果为基础,对不同顶级域名下的二级域名数量进行统计,绘制如图 7 所示的顶级域名数量占比图,从图中可以看出,com 顶级域名的数量占据所有顶级域名数量的 80%,而

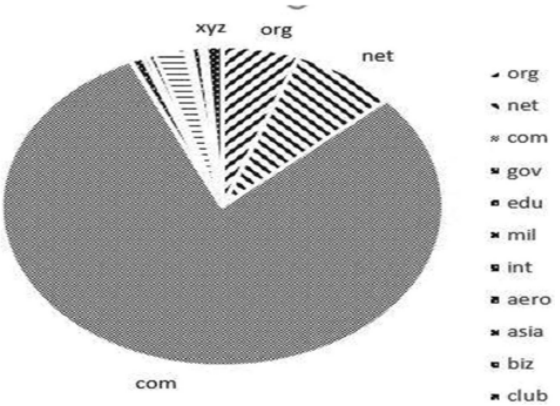

图 7 顶级域名占比统计图

其他新顶级域名的二级域名数量在顶级域名所包含的二级域名数量占比不足 10%,因此可以看出,新顶级域名的出现并没有对通用顶级域名的使用率造成很大影响,通用顶级域名在使用过程中依旧占据顶级域名使用的重要地位。

对于顶级域名谱系识别的结果分析,本文以 2019 年 12 月份的域名解析数据为基础,对通用顶级域名和新顶级域名进行谱系构建,根据谱系构建结果对顶级域名下每一个二级域名所包含子域名数量进行统计,并绘制如图 8 所示顶级域名中二级域名所包含子域名数量分布图,其中纵轴表示不同的顶级域名,横轴表示子域名数量不同的二级域名数量在整体二级域名所占比例。

从图 8 可以看出,二级域名所包含的子域名的数量即三级域名数量不超过 3 的比例超过 80%,而二级域名所包含的子域名数量超过 8 的比例低于

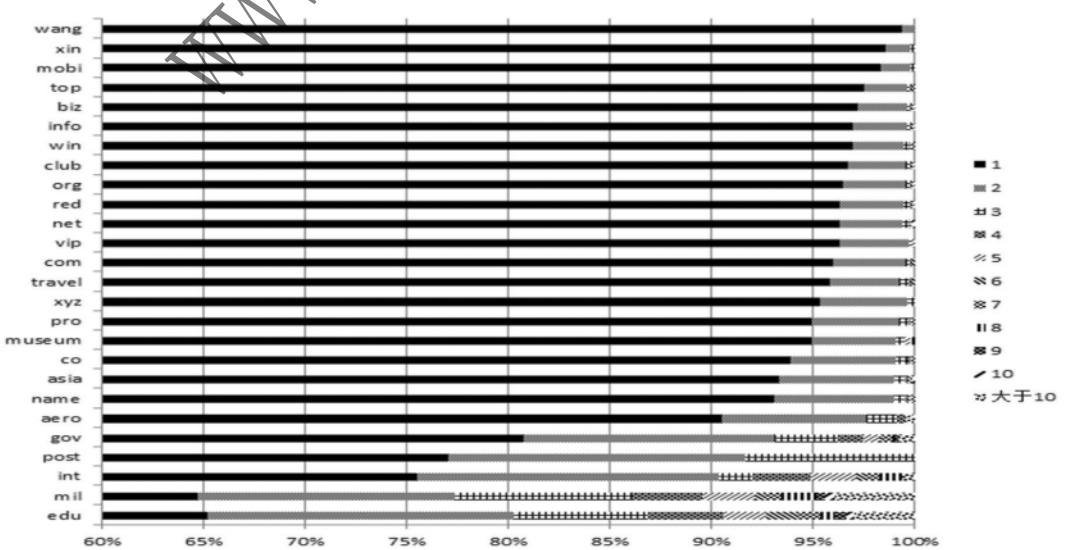

图 8 二级域名所包含子域名数量分布图

www.pcachina.com

10%，甚至 asia、co、club、pro、top 等顶级域名所有二级域名的子域名数量都小于 6，因此可以看出顶级域名所包含的二级域名谱系结构相对简单，甚至谱系规模为 1。此外，由图中结果可见，二级域名的子域名个数大于 10 的占比很少，不超过 5%。

3.2 国家域名数量分析

随着互联网经济的发展，部分地区的域名数量也呈现出一定程度的增加，因此为验证国家或地区域名数量与该地区的经济发展、IP 地址数量是否存在关联关系，本文统计了不同国家和地区的顶级域名数量，并结合国家 IP 地址数量和各个国家的 GDP 发展水平，对域名数量与经济发展、IP 地址数量之间相关性进行分析。

本文以 2019 年 12 月份的域名解析数据为基础，对全球国家域名进行谱系构建，通过对全球225 个国家域名所包含的二级域名进行统计，绘制如图9 所示全球国家域名二级域名数量分布图；以全球 IP 定位数据作为基础，统计各个国家 IP 分布数量，绘制了如图 10 所示全球 IP 数量分布热度图；以世界银行公布的 2019 年全球各个国家 GDP 数据为基础绘制了如图 11 所示全球国家 GDP 热度图。

从图 9 可以看出全球域名数量相对较多的国家或地区以中国、俄罗斯、欧洲、美国、加拿大等为主，而域名数量相对较少的国家和地区主要集中在南美洲、非洲地区；从图 10 可以看出 IP 数量相对较多的国家或地区以中国、美国、日本、欧洲等为主，而 IP 数量相对较少地区主要集中在中亚、非洲、南美洲等；从图 11 可以看出 GDP 相对较高的国家或地区主要以美国、中国、日本、印度、欧洲等为主，而 GDP 相对较低地区主要集中在中亚、非洲、南美洲的部分地区等。

图 9 全球域名数量热度图

图 10 全球 IP 数量热度图

图 11 全球 GDP 热度图

由图 9 和图 10 中所展示的数据占比可以看出，各个国家域名数量与 IP 地址数量之间不具备很强的相关性，例如俄罗斯的域名数量在不同国家域名数量排名相对靠前，而俄罗斯的 IP 地址数量排名则不然；另外由图 9 和图 11 所展示的数据占比可以看出，域名数量占比较高的国家和地区与全球经济较为发达的国家数据相对一致，而域名数量较少的国家则主要集中在经济较为落后的国家和人口相对较少的国家。由此可见国家和地区的域名数量与该地区的经济发展呈现相关性。

4 结论

本文提出了域名画像的概念，对域名进行多个维度的刻画；基于开源域名数据设计并实现了多源域名数据获取系统，实现全球国家 200 多个顶级域名的谱系构建，为域名规模的分析提供数据支持；结合历史解析记录实现了域名的时空变化分析。通过对顶级域名所包含二级域名数量的分析，认为通用顶级域名依旧占据很重要的地位，并且国家顶级域名数量与该地区的经济发展水平具有相关性。通过域名画像有助于用户掌握域名历史发展情况，预测域名发展趋势，实现特定域名目标的发现。

然而本文对域名谱系及历史变化只进行粗粒度的分析，对大规模域名的发现没有进行全面验证，并只对三级域名及以上层级进行谱系构建，因此后

www.pcachina.com

续工作需要对域名发现的规模加以验证，并构建层次分明的多层域名谱系结构，实现更全面、多维、细粒度的域名画像。

参考文献

[1] 郭莉，曹亚男，苏马婧，等.网络空间资源测绘：概念与技术[J].信息安全学报，2018，3(4)：1-14.

[2] VALLINA P，LE POCHAT V，FEAL Á，et al.Mis-shapes，mistakes，misfits：an analysis of domain classification services[C].Proceedings of the ACM Internet Measurement Conference，2020：598-618.

[3] YANG H C，LEE C H.A text mining approach on automatic generation of Web directories and hierarchies[J].Expert Systems with Applications，2004，27(4)：645-663.

[4] AGYEPONG E，BUCHANAN W J，JONES K.Detection of algorithmically generated malicious domain[C].International Conference of Advanced Computer Science & Information Technology，2018.

[5] BILGE L，KIRDA E，KRUEGEL C，et al.EXPOSURE：finding malicious domains using passive DNS analysis[C].Proceedings of the Network and Distributed System Security Symposium，2011.

[6] ANTONAKAKIS M，PERDISCI R.From throw-away traffic to bots：detecting the rise of DGA-based malwar[C].Usenix Conference on Security Symposium，2012.

[7] ANTONAKAKIS M，PERDISCI R，，DAGON D，et al.Building a dynamic reputation system for DNS[C].Usenix Conference on Security.USENIX Association，2010.

[8] OSTERWEIL E，RYAN M，MASSEY D，et al.Quantifying the operational status of the DNSSEC deployment[C].Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement，Vouliagmeni，Greece：ACM，2008.

[9] LIU B J，LU C Y，LI Z，et al. A reexamination of internationalized domain names：the good，the bad and the ugly[C].2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks(DSN)，2018：654-665.

[10] MOWBRAY M，HAGEN J.Finding domain-generation algorithms by looking at length distribution[C].IEEE International Symposium on Software Reliability Engineering Workshops.USA：IEEE，2014：395-400.

[11] 朱迦南.基于DNS日志数据的异常域名检测研究[D].成都：电子科技大学，2018.

[12] LUO X，WANG L M，XU Z，et al.DGASensor：fast detection for DGA-based malwares[C].Proceedings of the 5th International Conference on Communications and Broadhand Networking.ACM，2017.

[13] JARASSRIWILAI T，DAUBER T，BROWNLEE N，et al.Understanding evolution and adoption of top-level domain names[C].Proceedings of IEEE 40th Local Computer Networks Conference Workshops(LCNWorkshops).IEEE，2015.

[14] HALVORSON T，DER M F，FOSTER I D，et al.From academy to zone：an analysis of the new TLD land rush[C].Proceedings of the 2015 Internet Measurement Conference.ACM，2015.

[15] KORCZYNSKI M，WULLINK M，TAJALIZADEHKHOOB S，et al.Cybercrime after the sunrise：a statistical analysis of DNS abuse in new gTLDs[C].Proceedings of the 2018 Asia Conference on Computer and Communications Security.ACM，2018.

(收稿日期：2021-03-15)

作者简介：

包正晶(1996-)，男，硕士研究生，主要研究方向：信息安全。

苏马婧(1985-)，女，博士，高级工程师，主要研究方向：信息安全、网络空间测量。

www.pcachina.com

版权声明

经作者授权，本论文版权和信息网络传播权归属于《信息技术与网络安全》杂志，凡未经本刊书面同意任何机构、组织和个人不得擅自复印、汇编、翻译和进行信息网络传播。未经本刊书面同意，禁止一切互联网论文资源平台非法上传、收录本论文。

截至目前，本论文已经授权被中国期刊全文数据库（CNKI）、万方数据知识服务平台、中文科技期刊数据库（维普网）、JST 日本科技技术振兴机构数据库等数据库全文收录。

对于违反上述禁止行为并违法使用本论文的机构、组织和个人，本刊将采取一切必要法律行动来维护正当权益。

特此声明！

《信息技术与网络安全》编辑部

中国电子信息产业集团有限公司第六研究所

www.ncachina.com