软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(1):321−340 [doi: 10.13328/j.cnki.jos.007194] [CSTR: 32375.14.jos.007194]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

客户端独立的 IP 地理定位研究综述*

林金磊[1], 李城龙[1,2], 宋光磊[2], 樊琳娜[3], 王之梁[1,2], 杨家海[1,2]

[1](清华大学 网络科学与网络空间研究院, 北京 100084)

[2](中关村实验室, 北京 100081)

[3](国防科技大学 信息通信学院, 湖北 武汉 430019)

通信作者: 李城龙, E-mail: lichenglong@tsinghua.edu.cn; 杨家海, E-mail: yang@cernet.edu.cn

摘　要: 准确、快速地获取 IP 地理定位信息对于各种网络应用而言至关重要. IP 地理定位指将互联网实体的 IP 地址转换为其地理位置的技术. 然而, 互联网规模的迅速扩大和互联网应用的快速发展, 给 IP 地理定位研究带来了巨大的挑战. 首先, 复杂的网络结构和网络环境导致 IP 定位技术的精确度远远无法满足实际的应用需求. 其次, IP 地理定位在各个领域的作用日益凸显, 如何精准、高效、可靠地计算互联网主机的地理位置正在成为各行关注的焦点. 因此, 通过设备的 IP 地址对其进行地理定位以支撑复杂的上层应用尤为重要. 自 2001 年以来, 学术界和工业界围绕上述问题开展了大量的研究. 系统地梳理客户端独立的 IP 地理定位方面的相关工作, 系统地整理基于网络测量的 IP 地理定位研究分类方法. 根据定位数据是否由主动测量产生, 将相关研究分为主动的 IP 定位技术、被动的 IP 定位技术和主被动结合的 IP 定位技术. 进一步, 对每一类方法进行更细粒度的分类并分析其主要的优缺点. 在此基础上, 总结 IP 地理定位领域的最新进展和研究挑战, 并展望其未来发展方向.

关键词: 网络测量; IP 地理定位; 地理定位地标

中图法分类号: TP393

中文引用格式: 林金磊, 李城龙, 宋光磊, 樊琳娜, 王之梁, 杨家海. 客户端独立的IP地理定位研究综述. 软件学报, 2025, 36(1): 321–340. http://www.jos.org.cn/1000-9825/7194.htm

英文引用格式: Lin JL, Li CL, Song GL, Fan LN, Wang ZL, Yang JH. Survey on Client-independent IP Geolocation. Ruan Jian Xue Bao/Journal of Software, 2025, 36(1): 321–340 (in Chinese). http://www.jos.org.cn/1000-9825/7194.htm

Survey on Client-independent IP Geolocation

LIN Jin-Lei[1], LI Cheng-Long[1,2], SONG Guang-Lei[2], FAN Lin-Na[3], WANG Zhi-Liang[1,2], YANG Jia-Hai[1,2]

[1](Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing 100084, China)

[2](Zhongguancun Laboratory, Beijing 100081, China)

[3](College of Information and Communication, National University of Defense Technology, Wuhan 430019, China)

Abstract: Capturing an accurate view of IP geolocation is of great interest to the networking research community as it has many uses ranging from network measuring and mapping to analyzing the network's infrastructure. However, the scale of today's Internet, coupled with the rapid development of Internet applications, makes it very challenging to acquire a complete and accurate snapshot of the IP geolocation technology. To the best of our knowledge, there is no systematic survey of the relevant research in this field. To fill this gap, this study systematically summarizes the research of client-independent IP geolocation, in which the clients do not participate in the geolocation process. This study aims to examine the major research studies that have been conducted on topics related to IP geolocation in the last 22 years since the first IP-based geolocation technology was proposed. To this end, these prior studies are classified according to the measurement method, that is, active, passive, and hybrid. The main techniques for each category are described, identifying their significant advantages and limitations. Also, the primary experience and lessons learned from these past efforts are presented. After the process, the latest progress in IP geolocation both in academia and industry is shown. Finally, the survey and present promising directions

* 基金项目: 国家重点研发计划 (2022YFB3105001); 国家自然科学基金 (62172251)
收稿时间: 2023-05-29; 修改时间: 2024-02-21; 采用时间: 2024-04-01; jos 在线出版时间: 2024-06-14
CNKI 网络首发时间: 2024-06-17

in the future are concluded, hoping to promote the development of IP geolocation.

Key words: networking measurement; IP geolocation; landmark

互联网是人类社会空间和地理空间的延伸, 已成国家继陆、海、空、天这 4 个疆域之后的“第五疆域”. 互联网上的实体同样具有地理位置属性, IP 地理定位[1]则是通过 IP 地址来确定互联网实体地理位置的技术. 客户端独立的 IP 地理定位表示在地理定位的过程中不依赖用户的参与, 不需要像 GPS[2]和北斗[3]等技术由用户主动发起定位的请求, 或是提前向用户申请定位权限, 只利用 IP 地址这一互联网上的唯一标识来计算其地理位置. IP 地理定位在工业界通常以定位服务或者定位数据库的形式呈现.

随着互联网位置应用的快速发展, 准确推断用户地理位置的能力将为服务和应用提供商带来极大的便利. IP 地理定位的应用非常广泛. 在商业应用方面, 广告和内容推荐是其典型应用, 由于不同地区对广告和内容交付的需求差异很大, 在广告投放和内容推荐时如果能够准确地估计出目标用户的位置, 就可以节省大量的成本和网络资源. 在维护国家安全、社会秩序方面, IP 地理定位与网络犯罪的追踪和预防密切相关, 由于互联网的复杂性和多变性, 网络犯罪的追踪相比于传统的社会犯罪更复杂. 在这种情况下, 快速而准确的 IP 地理定位将有助于减少网络犯罪, 并增加破获网络犯罪案件的可能性. 在网络运维方面, IP 地理定位可以用于构建网络基础设施的位置和 IP 地址之间的映射关系, 网络管理员可以借此优化现有的网络拓扑或指导未来的拓扑设计[4,5].

根据是否依赖用户参与定位, 可以将 IP 地理定位分为客户端独立的和客户端依赖的定位. 客户端独立的 IP 地理定位是指在不依赖用户参与的情况下通过 IP 地址来确定互联网实体的经纬度坐标, 只使用 IP 地址和基于 IP 地址的其他信息来推断主机最可能的位置. 与此相反的是, 客户端依赖的 IP 地理定位是由用户主动发起的, 如 GPS (global positioning system)[2]和 BDS (Beidou navigation satellite system)[3]用于室外 IP 地理定位, Wi-Fi[6]用于室内 IP 地理定位, Cellular 用于移动 IP 地理定位等[7]. 常见的客户端依赖的 IP 地理定位技术在定位的效果上更好[8], 但是硬件和软件的限制使得其在诸多情况下不能使用. 例如, 没有配备 GPS 传感器的互联网设备由于缺少硬件支撑而无法启用 GPS 定位服务; 另外, 即使配备了 GPS 传感器, 如果用户拒绝开启定位权限或者用户在室内时, 这些定位方法也会失效. 本文主要关注客户端独立的 IP 地理定位技术, 调研的文献也是以此进行范围划分, 根据定位数据的来源将 IP 地理定位技术分为主动的 IP 定位技术、被动的 IP 定位技术和主被动结合的 IP 定位技术. 此外, 由于 IP 地理定位地标对 IP 地理定位技术非常重要, 本文对定位地标也进行了综述.

系统的 IP 地理定位研究最早可以追溯到 2001 年, Padmanabhan 等人[1]提出了 3 个定位方法并为后续的定位研究奠定了基础. 虽然此后有大量的定位研究, 但学术界仍有一些未解决的问题. 第一, IP 地理定位技术的精度差异大、应用范围有限, 随着网络应用的日益复杂, 传统的 IP 地理定位技术的精度已经逐渐不能满足实际需要; 第二, IP 地理定位技术没有标准和统一的评价体系, 许多算法无法根据其准确性进行直接比较, 除了精确度, 覆盖率外, 可扩展性和算法复杂度等指标也很重要; 第三, 自 2012 年以来, IP 地理定位研究的重点是使用更高质量的定位数据来提高精确度, 许多地理定位技术很难被复现和推广; 最后, 目前大多数 IP 地理定位研究只支持 IPv4 地址定位, 对 IPv6 地址的支持非常有限. 除了学术界的研究, 工业界还有很多开源和商业的地理位置数据库, 如 IPIP.NET[9]、NetAcuity[10]和 MaxMind[11]. 一方面, 这些地理定位库为 IP 地理定位的研究和发展做出了巨大贡献. 另一方面, 这些数据库通常有商业和开源两个版本, 商业版准确但价格昂贵, 开源版本的准确性和覆盖度都非常有限. 但由于各种原因, 各种版本定位库的准确度都远远达不到实际使用的要求[12−15].

在各种因素的推动下, IP 地理定位经历了 20 多年的发展, 已经有一些综述文章[16,17]和定位算法研究[18−22]对相关工作进行了总结. 王志豪等人[16]将客户端独立的 IP 地理定位分为基于信息推测和基于网络测量, 但是其分类方法不正交且包含的文献不全面, 缺乏系统化的总结. 王占丰等人[17]总结了 2014 年之前的相关工作, 将客户端独立的 IP 地理定位研究分为基于推测的和基于时延的, 这样的分类方法无法准确地分类最新的定位技术. 总结来看, 目前还没有针对该领域最新研究的系统性综述. 首先, 对具有更高精确度的 IP 地理定位技术的需求正在增加, 同时有越来越多新出现的 IP 地理定位方法, 但现有综述缺少对最新研究的讨论[18,21−28]. 其次, 现有的研究工作中采用的定位分类方法存在各类之间有交集、关系不明确的问题. 第三, IP 地理定位已经逐渐发展成为与多个领域交叉的综合性研究, 因此很多与 IP 地理定位相关的研究都被总结到其他领域的综述中[29−31], 但是这些综述并没有结构化地对客户端

独立的 IP 地理定位给出一个完整的描述. 最后, 随着新应用的出现和 IPv6 的快速普及[32], 对更稳定、更准确的 IP 地理定位技术的需求越来越大[33], 需要通过结构化的分类方法来确定研究意义和框架, 推动 IP 地理定位的发展.

本文对现有的 IP 定位研究进行检索并构建文献库, 时间跨度为 2001–2022 年. 文献来源囊括当前主流的文献数据库, 包括: 中国知网、万方数据知识服务平台、维普数据库、IEEE Xplore、ACM Digital Library、Springer Online Library、Wiley InterScience、USENIX、Elsevier ScienceDirect Online Library. 自 2001 年出现第 1 个研究工作[1]以来, IP 地理定位的发展一直处于波动增长之中. 从图 1 可以看出, 2011 年之前只有一些零星的研究, 2012 年后, 随着人工智能、数据挖掘、图像识别等许多新技术的出现, IP 地理定位相关的文献大幅增加, 并在 2022 年达到历史峰值. 如图 1 中虚线所示, 我们对发表的文献数量进行多项式拟合, 从整体趋势来看, IP 地理定位领域的研究越来越多, 未来仍有很大的发展空间.

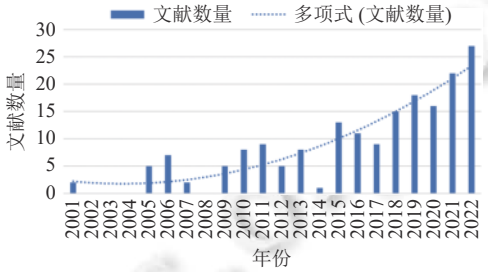

图 1 IP 地理定位研究发表文献趋势

总结: 本文梳理 20 多年来 IP 地理定位的发展, 提出一种结构化[34,35]的分类方法对客户端独立的 IP 地理定位技术研究进行总结. 首先梳理 IP 地理定位地标研究, 然后分别总结主动、被动和主被动结合的定位方法. 基于对这些工作系统的梳理, 总结 IP 地理定位研究在学术界和工业界面临的挑战, 并展望未来发展趋势.

1 客户端独立的 IP 地理定位技术分类方法

据我们所知, 目前与 IP 地理定位有关的综述文章并未系统总结其发展历程或者涵盖的文献有限, 大多数研究工作在介绍和背景部分对不同的 IP 地理定位技术进行分类. 这些研究中使用的分类方法是沿用 IP2Geo[1]提出的 3 条技术路线. 基于 IP2Geo 的分类方法的优点在于其分类方法简单, 可以快速地将 IP 地理定位方法分类为基于延迟的、基于拓扑的或基于数据库的. 然而, 这并不是一种结构化的分类方法, 同时随着 IP 地理定位技术的日益复杂, 越来越多的地理定位方案还使用了除延迟、拓扑和数据库以外的数据来计算地理位置[18,26,36,37]. 因此, 基于 IP2Geo 的方法与新出现的技术不能很好地兼容, 各个类别之间存在交叉.

本文使用测量方法作为分类的基础, 以往的工作[38−41]中有过类似的分类方法, 但是没有将其系统化总结. 客户端独立的 IP 地理定位可以根据地理定位计算中使用的数据是否由主动测量产生可以分为 3 类: 主动、被动和主被动结合的.

IP 地理定位技术的 3 种方法描述如下.

1) 主动的 IP 地理定位技术: 主动的 IP 定位技术通过执行主动测量来获取并构建地理定位数据, 在地理定位计算过程中使用的数据最初并不存在. 主动的 IP 定位方法实时性更强, 但需要一定数量的探测节点 (探针), 需要更多的网络和计算资源才能获得较好的定位效果.

2) 被动的 IP 地理定位技术: 被动的 IP 定位方法主要是对现有定位数据的分析并从中提取目标主机的地理位置. 被动的 IP 定位方法相比于主动的 IP 地理定位方法更准确, 但实时性和可扩展性较差, 难以推广.

3) 主被动结合的 IP 地理定位技术: 主被动结合的方法不仅执行主动测量来构建定位数据, 同时还通过对现有定位数据的分析来提升精确度, 结合了主动和被动方法的优点. 近年来越来越多的研究工作都选择了主被动结合的 IP 定位方法.

后续章节将详细介绍 IP 地理定位地标和 3 种 IP 地理定位技术. 由于 IP 地理定位地标在地理定位技术中发

挥着重要的作用, 因此在详细介绍这 3 种方法之前, 先对 IP 地理定位地标相关的研究进行综述, 介绍地标的类型及其评估指标.

2 IP 地理定位地标

IP 地理定位地标[42,43]对 IP 地理定位的计算非常重要. 一般的, 如果在 IP 地理定位的过程当中难以直接对目标主机进行位置计算, 通常会选择一些靠近目标主机的设备进行辅助地理定位计算, 这些主机被称为 IP 地理定位地标. IP 地理定位地标通常具有一个相对固定的 IP 地址和地理位置 (经纬度坐标). 许多 IP 地理定位技术都高度依赖地标, 因此地标的质量在很大程度上决定了 IP 地理定位方法的准确性和覆盖度.

获取 IP 地理定位地标的方法有很多, 如网络服务探测、人工标注等. 可以将常见的地标分为 3 类: 基于网络内容挖掘的定位地标、基于在线地图的定位地标和基于网络测量平台的定位地标, 如表 1. 接下来将详细描述每个类别的定位地标.

表 1 IP 地理定位地标总结

地标类别	代表方法	核心技术	地标数量
基于网络内容的IP地理定位地标	Structon[44]	从网页中挖掘地理位置, 正则化提取	<1000
	Zhu等人[43]	从互联网论坛中抽取地理位置信息构建地标	4 854
	LandmarkMiner[42]	预测可能具有地理位置的域名并分类筛选	9 423
	GeoCAM[18]	从门户网站寻找网络摄像头, 提取位置作为地标	16 863
基于在线地图的IP地理定位地标	Wang等人[45]	从在线网络地图中提取信息, 考虑CDN分布	930
	Ma等人[38]	综合考虑网络拓扑、在线地图, 提取PoP级别地标	8 000
	Maziku等人[46]	机器学习方法提取地图特征, 设计地标提取策略	1 094
	Dan等人[47]	多源异构信息, 融合算法提高地标挖掘效率	7 422
基于网络测量平台的IP地理定位地标	—	研究者部署一部分已知地理位置的探针, 结合CAIDA[48]、Censys[49]、PlanetLab[50]提供的一些有地理位置的网络节点作为地标使用	100–5 000

2.1 基于网络内容挖掘的 IP 地理定位地标

基于网络内容挖掘的 IP 地理定位地标近年来随着网络应用的发展而出现, 是迄今为止最常见的地标类型, 它们来源于互联网上大量公开的、具有地理信息和 IP 地址的网络内容. 随着新技术的发展, 研究人员通过数据挖掘、文本识别、图像处理和语音分析等技术分析网络内容并提取地标.

基于网络内容挖掘的地标的研究相对较多. 早期的网络地标出现在互联网论坛上, 由 Zhu 等人[43]在 2015 年提出, 他们发现了数以千计的城市级地标, 这些地标主要分布在中国. 类似的, Feng 等人[51]研究了物联网的 (Internet-of-Thing, IoT) 设备特征及分布规律, 提出了一个基于规则的引擎来发现物联网设备的地理位置. 此外, Sommers 等人[52]从网络客户端的角度发现 IP 地理定位地标, 从 HTTP Cookies[53]中获得了分布在几个国家和地区的可用地标. Song 等人[54]从地标覆盖度的角度提升了地标的质量, 他们监控在线网络摄像头并将其作为地标. Li 等人[18]在此基础上定制了地标挖掘的方法并提出了优化的 IP 定位算法, 提高地理定位的精确度. 除了从网络应用中获得地标外, Ma 等人[38]从网络基础设施的角度来发现地标, 通过分析公开网络路由器来构建 IP 地理定位地标.

基于网络内容挖掘的 IP 地理定位地标的优势在于网络资源的丰富性和挖掘方法的多样性. 互联网上大量的设备信息和地理位置信息都可以为此类地标提供数据支撑. 随着机器学习和数据挖掘技术的发展, 对于此类地标的提取效率变得更高. 同时, 网络地标种类丰富、位置准确、数量更多, 可以为 IP 地理定位算法提供更大的支持. 如表 1 所示, 网络地标的数量和丰富程度都存在优势. 然而, 这种地标在稳定性方面有一定的局限性, 许多地标在获得后的一段时间内就无法访问. 此外, 这种地标挖掘方法与特定的设备类型密切相关, 很难推广到其他场景.

2.2 基于在线地图的 IP 地理定位地标

基于在线地图的 IP 地理定位地标主要来源于在线地图应用和包含在线地图的网站等. 在线地图应用主要指

高德地图、百度地图、腾讯地图、谷歌地图等 GPS 定位应用, 包含在线地图的网站主要指在页面中嵌入在线地图的网站、网页等, 这些网站通常是调用在线地图的 API 接口用于可视化展示. 在线地图中天然的就带有地理位置属性, 因而具有地标来源的必要条件. 考虑到 IP 地址在地图中存在的数量并不多, 此类方法也有其局限性. 当前, 仍然有一些研究工作从在线地图中提取定位地标. Wang 等人[45]从在线地图中提取地理位置信息, 同时还考虑了 CDN 的分布, 提高了网络地标的可靠性. Ma 等人[38]综合考虑了网络拓扑和在线地图, 从中提取了 PoP (point-of-presence) 级别的地标. Maziku 等人[46]使用机器学习方法提取地图特征, 设计了地标提取的策略. Dan 等人[26]采用多源异构信息, 使用融合算法提高地标挖掘的效率.

基于在线地图的 IP 地理定位地标在数量上非常有限, 另外, 随着在线地图中越来越多地采用安全防护和隐私保护策略, 这类地标的发现难度会更高. 另外, 基于在线地图的 IP 地理定位地标的可扩展性也比较局限.

2.3 基于网络测量平台的 IP 地理定位地标

基于网络测量平台的 IP 地理定位地标主要来源于一些商业公司或网络测量机构. 很多 IP 定位工作中提到的公共地标 (开源地标) 大多来源于此. 自从 IP 地理定位出现以来它们就已经存在了. 到目前为止, 网络测量平台地标并没有一个公认的数据集, 常用的有 CAIDA (cooperative association for Internet data analysis) 数据集[48]、Censys 数据集[49]等. 另外, CAIDA[48]、Censys[49]、PlanetLab[50]也会提供一些有地理位置的网络节点可以作为地标使用. 这些地标的优点是容易获得, 可以作为地标的基准使用. 此外, 与基于网络内容挖掘的地标相比, 这些地标的稳定性和精确度通常更好. 然而其分布和数量远远不能满足实际使用的需要.

2.4 IP 地理定位地标评估指标

IP 地理定位地标的评估指标主要用于衡量地标的质量, 为地标挖掘研究提供性能依据. 值得注意的是, IP 地理定位地标的评价指标与 IP 地理定位方法的评价指标不同. 根据 IP 地理定位技术中地标的特点, 本文总结了 4 个地标评价指标.

1) 数量: 地标的数量越多, 地标挖掘算法的性能就越好. 到目前为止, 最先进的地标挖掘方法[18,54]能够发现几千到上万的地标.

2) 覆盖率: IP 地理定位地标的覆盖率表示地标的地理分布. 地标需要在更细粒度的水平上覆盖尽可能多的国家和地区, 目前的地标通常集中在互联网发达的地区.

3) 稳定性: 正如上面所提到的, IP 地理定位地标的地理位置和 IP 地址需要在一段时间内保持不变. 目前大多数地理定位服务和系统都是每天更新的, 因此具有高稳定性的地标需要和定位服务保持一致.

4) 准确性: IP 地理定位地标和 IP 地理定位技术一样需要在准确性方面进行评估. 通常情况下, 城市级别的地理定位算法需要城市级别的地理定位地标, 使用国家级别的地标是无法实现的. 目前, 最高精确度的 IP 地理定位地标能够达到楼宇级别的精确度, 但其数量相对较少.

3 IP 地理定位精确度和评估

IP 地理定位技术最直观、最重要的要求是准确性. 随着时间的推移, IP 地理定位的准确性越来越高. 一般来说, 根据定位的精确度 IP 地理定位技术可以分为大洲级、国家级、省级、城市级、街道级和楼宇级. 事实上, 许多 IP 地理定位技术并没有明确它们能达到什么样的地理定位精度, 所以本文根据这些文献的评估实验的地理定位精度和实验设置进行总结. 从精确度的角度来看, 早期的 IP2Geo[1]只能达到省级的精确度, 随后的 CBG[55]和 TBG[56]在此基础上有一定的改进, 在性能较好的地区精确度可以达到城市级. SLG[57]是第 1 个可以达到街道级精确度的 IP 地理定位技术. Ding 等人[58]在 SLG 的基础上做了一些改进, 但精确度仍在街道级. 从研究的难度和研究的密度来看, 具有街道级和楼宇级精确度的 IP 地理定位方法很难实现, 通常需要大量的位置数据和高精度的地理定位地标. 现有的研究越来越集中在城市级和街道级的地理定位方法上.

如第 2 节所述, IP 地理定位技术的评价指标与 IP 地理定位地标的评价指标不一样. 本文主要将 IP 地理定位技术的评价指标总结为以下几点.

1) 精确度: 精确度是 IP 地理定位技术评价最关键的指标. 精确度直接决定了一个地理定位算法是否能满足需求. 精确度的表现形式很多, 如平均误差距离、中位误差距离、最佳性能点、误差半径等.

2) 覆盖率: 覆盖率指的是地理定位算法能够支持多少个 IP 地址. IPv4 和 IPv6 的地址空间非常大, 大多数 IP 地理定位算法不能覆盖全部的 IP 地址. 覆盖率越高, IP 地理定位算法的用途就越广. 除了地址空间的覆盖率, 还需要关注地理空间的覆盖率. 一个好的地理定位算法应该覆盖尽可能多的国家的 IP 地理定位计算.

3) 可扩展性: IP 地理定位算法通常有一个算法的适用范围, 如果能够以较小的成本扩展到适用范围之外, 那么该地理定位方案的可扩展性是较强的.

4) 稳定性: IP 地理定位算法通常用于提供在线服务, 这就要求算法能够长期运行并能抵御网络攻击、网络拥堵以及真实网络中的其他情况.

5) 开销: IP 地理定位算法的开销体现在几个方面: 测量开销、计算开销、部署开销和维护开销. 当地理定位算法引入主动测量时, 它会在互联网上产生一定量的负载, 需要将其降到最低. 此外, IP 地理定位的计算速度需要尽可能快, 计算开销过高将导致成本增加, 甚至由于 IP 地址的动态变化而导致不正确的定位结果. 最后, IP 地理定位算法需要被集成到在线服务或定位数据库中, 这个过程中有不可避免的部署成本和维护成本. 例如, 地标、内存、CPU 和数据更新的频率等.

4 主动的 IP 地理定位技术

主动的 IP 地理定位技术通过主动测量构建地理定位数据, 进而分析所收集的数据以计算地理坐标. 例如使用 ping[59]测量几个主机之间的延迟/RTT (往返时间)、使用 traceroute[60]测量已知主机周围的拓扑结构、使用 dnsquery 采集包括地理位置信息的 DNS 数据等. 由于较好的实时性能和易用性, 主动的 IP 地理定位技术是最常见和最广泛使用的客户端独立的 IP 地理定位技术, 受到大多数研究人员和地理定位系统的青睐. 但同时, 主动探测的开销以及对网络和道德的影响等, 都是主动的 IP 地理定位技术所面临的挑战[61−63].

许多常见的 IP 地理定位方法都是基于主动测量的, 从表 2 看出, 随着时间的推移, 地理定位的准确性越来越高. 另外, 并非所有的方法都只使用单一的测量工具. 例如, Octant[64]需要 ping[59]来获取距离约束, 需要 traceroute[60]来获取拓扑信息. 大多数 IP 地理定位方法的准确性是通过误差范围、中位数误差和平均误差来评估的. 大部分定位算法在评估时都是使用预测位置和实际位置的距离 (km) 来表示, 但是仍有一部分定位算法由于只预测目标主机的精确度级别, 因此使用百分比来表示预测成功的占比. 例如表 2 中的 Geo-PoP[65]和 RNBG[66]使用百分比来表示预测城市正确的目标在所有目标中的占比, 正确率分别达到 96.67% 和 97.67%.

表 2 主动的 IP 地理定位技术对比

方法类别	代表方法	应用范围	最佳性能 (km)	定位误差 (km)	用到的工具/数据		
					Ping	Traceroute	DNS Query
基于网络延迟的IP地理定位技术	GeoPing[1]	美国/欧洲	382	1 201	√	—	—
	CBG[55]	美国	95	182	√	—	—
		欧洲	22	78			
	Octant[64]	北美	N/A	25	√	√	—
	Spotter[61]	全世界	10	30	√	—	—
	Posit[62]	北美	35	61	√	—	—
	GeoCET[63]	中国/印度/美国/德国	N/A	0.87	√	—	—
基于网络拓扑的IP地理定位技术	GeoTrack[1]	美国/欧洲	102	384	—	√	√
	TBG[56]	美国 (商业网络)	100	209	—	√	√
		美国 (教育网络)	61	194			
	Geo-PoP[65]	中国/美国	N/A	城市级	—	√	√
	RNBG[66]	中国/美国	N/A	城市级	—	√	√

主动的 IP 地理定位技术可以进一步分为基于延迟的方法和基于拓扑的方法. 一方面, 最常用和最直观的地理定位方法是基于延迟的定位技术, 因为网络延迟和地理距离之间存在着自然的联系[67−70]. 尽管延迟-距离函数很难拟合, 但为了克服这一挑战, 解决方案不断出现[40,59,64,71,72]. 另一方面, 基于拓扑结构的方法在基于延迟的方法中增加了网络拓扑结构的领域知识, 基于拓扑结构的方法还使用其他数据来优化地理定位过程, 如 DNS 数据[41,56,73,74]、路由器位置[75,76]等. 接下来, 我们深入讨论两种主动的 IP 地理定位技术, 包括介绍一些典型工作和对这些技术的理解. 另外, 我们选择了几种经典的 IP 地理定位方法来介绍它们的核心思想和定位过程.

4.1 基于网络延迟的 IP 地理定位技术

基于网络延迟的 IP 地理定位技术测量探针和地标以及目标主机之间的网络延迟, 将延迟转换为地理位置约束, 然后求解以获得目标主机的可能位置. 目标主机的位置可以被框定在一个范围内, 范围大小因不同的地理定位方法而异, 这就是地理定位精度的来源. 目标主机所在的范围越小代表越高的精确度. 基于网络延迟的定位方法的核心是求解延迟和地理距离之间的关系. 网络延迟和地理距离之间存在一定的关系[55,61,64,77]. 光纤中信息的传导速度为光速的 2/3, 即 2c/3. 由于光纤中存在一定的信号损失, 考虑到了互联网路径的迂回性, 现有工作[55]通常将 4/9 作为延迟-距离的转换系数, 还有其他的研究工作则使用更复杂的转换模型. 因此较长的地理距离将对应更大的网络延迟. 然而随着网络环境变得越来越复杂, 建立延迟-距离关系变得更加困难. 虽然可以用数据驱动的方法来拟合局部地区的延迟-距离函数[40,62,63], 但目前还没有合适的拟合方法可以在全球范围内通用.

基于网络延迟的 IP 地理定位方法最早来源于 GeoPing[1], 它是 IP2Geo[1]的一个组成部分. GeoPing 的作者提出了 NNDS (nearest neighbor in delay space, 延迟空间中的最近邻) 算法来刻画网络延迟和地理距离之间的关系. 在计算目标主机的位置时, 在 NNDS 算法构建的延迟映射表中找到离目标主机最近的地标, 目标主机的位置由距离目标主机最近的地标来表示. 这是一种直观的定位方法, 目标主机位置的解空间被限制在地标节点的集合内, GeoPing 的定位精确度随着地标数量和覆盖范围的增加而逐渐提高. 实验结果表明, GeoPing 的中位误差约为 382 km. 虽然 GeoPing 的精度与后来的许多算法相比较低, 但它是最早的基于网络延迟的 IP 地理定位方法, 对后来很多的地理定位技术产生了很大影响.

在 GeoPing 出现后的几年里, 许多研究人员都尝试着改进基于网络延迟的 IP 地理定位方法. 其中最经典的方法之一是 CBG[55]. CBG 是在 GeoPing 的核心思想基础上设计和改进的, 其计算过程可以分为 3 个步骤.

1) 具有地理距离约束的多点定位: 给定一组具有详细地理位置的地标, 每个地标推断出它与目标主机的地理距离约束. 如图 2 所示, 推断的地理距离约束是由 $\hat{g}_{it} = g_{it} + \gamma_{it}$ [55]计算得到, 即真实的地理距离 g_{it} 加上一个由 γ_{it} 表示的附加地理距离误差.

2) 构建延迟-距离约束: CBG 使用目标主机和每个地标之间的延迟测量来模拟网络延迟和地理距离之间的关系.

3) 使用分布式延迟-距离约束来计算目标主机的地理位置: CBG 使用一种几何方法来估计一个给定的目标主机的位置. 每个地标推断出与目标主机的地理距离约束 (步骤 1) 的输出), 由于有距离误差的存在, 目标主机的位置位于相交区域内的某个地方. 图 2 中的红色区域对应于目标主机的位置估计.

CBG 解决了之前的地理定位技术的解空间是离散的问题, 并且一定程度上减少了地标数量对定位精确度限制的问题, 使用具有距离约束的多点地理定位[64]来推断互联网主机的地理位置. CBG 的中位误差为 100 km, 它主要的贡献在于地理多点定位思想的引入.

在基于网络延迟的 IP 地理定位方法中, 有较大部分的算法将网络延迟转化为地理定位的各种约束. 在 CBG[55]之后, 地理定位的约束条件和关键步骤有很多改进. 例如, Octant[63]将 CBG 中的约束条件优化为正负约束, 而且约束的区域有任意形状的边界, 而不仅是 CBG 中提出的圆形边界. 从结果上来看, Octant 的中位误差为 35.4 km. 同时, Octant 的精确度不会随着地标的变化而出现大的波动, 使得它可以在不同的场景下对目标主机进行地理定位. Spotter[61]没有以固定的方式描述延迟-距离关系, 而是以概率的方式构建其内部模型, 所有的地标节点被一起建模以得出一个全局的延迟-距离模型, 使 Spotter 具有更高的稳定性和准确性. 从测量成本的角度来看, Posit[62]提出了一种更轻量级的 IP 地理定位方法, 它只需要对目标主机进行少量的延迟测量就可以完成位置计算, 虽然执行

更多次的网络测量, 但是每次测量的开销非常小, 进而总的测量成本大大降低, 定位精确度也得到提升. GeoCET[63]汲取了 Octant 和 Posit 的优点, 结合椭圆轨迹约束和最大对数似然估计对目标位置进行求解. GeoCET 只需要获得少量的单向延迟就可以完成目标主机的地理定位. 在硅谷数据集上, GeoCET 的中位误差在 1 km 以内, 达到了当时最高的精确度. Patel 等人[78]优化了基于网络延迟的地理定位算法在存在拥塞的大型计算机网络中的性能, 在某些情况下将地理定位的误差降低到 0.7%.

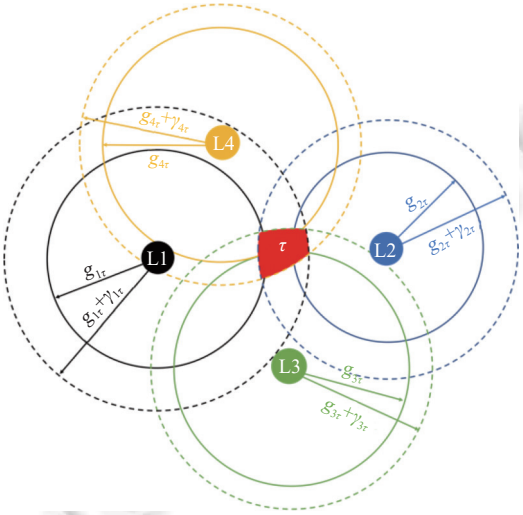

图 2 CBG 算法多点定位示意图

基于网络延迟的方法具有较高的通用性, 主要有以下 3 个原因.

1) 计算成本低, 只需要执行一系列的网络延迟测量就可以进行地理位置的计算;

2) 实时性较好, 可以应对 IP 地址的动态变化[4], 在目标主机移动的场景下也能获得不错的性能;

3) 计算限制较少, 只要探针和目标主机之间有网络路由路径就可以完成地理位置的计算.

但是, 基于网络延迟的定位方法也有一些缺点. 首先, 基于延迟的方法主要是通过改进延迟-距离模型的构建和约束条件的求解来优化原有的算法, 而这些都是复杂的数学问题. 因此, 基于延迟的方法比其他方法需要更多的领域知识. 其次, 与地理空间不同, 延迟-距离的映射受到太多因素的影响, 而延迟-距离关系建模会直接决定约束的类型和内容, 对 IP 地理定位准确性的影响极大. 因此如何建模并解决延迟-地理距离关系是一个很大的挑战. 第三, 延迟-地理距离关系在全球范围内是完全不同的[75]. 正是由于这个原因, 基于延迟的方法的可扩展性较差. 最后, 与其他方法相比, 基于延迟的方法在准确性方面存在不足. 与拓扑结构和其他网络指标相比, 时延是一个更不稳定的指标, 网络拥塞和其他原因都会导致延迟的波动, 使得基于延迟的方法的准确性和稳定性降低.

4.2 基于网络拓扑的 IP 地理定位技术

基于网络拓扑的 IP 地理定位技术在 IP 地理定位计算中考虑了网络拓扑信息, 所以在地理定位过程中会对网络结构有更深的了解[56,73,77,79,80]. 有些方法甚至从子网的划分开始[74,80], 通过寻找子网的位置来完成目标主机的地理定位. 基于网络拓扑的方定位法通常需要其他方法或定位数据的协助, 如使用 DNS 数据寻找到目标主机路径上最后一跳路由器的位置[75,79], 或从子网的网关开始执行延迟主动测量[1,25,56]等. 与基于网络延迟的定位方法相比, 基于网络拓扑的定位方法研究相对较少, 但此类方法的思路和切入点往往很新颖.

基于网络拓扑的定位方法中, 最早的是 IP2Geo 中的 GeoTrack[1]. 网络运营商通常为路由器分配有地理意义的名称, 以方便网络管理[81,82]. GeoTrack 基于这一规律, 根据目标主机或附近其他网络节点的 DNS 名称推断出目标主机的位置. GeoTrack 的计算过程分为 3 个步骤.

1) 使用 traceroute[60]工具测量探针 (vantage points) 和目标主机之间的网络路径;

2) 从 DNS 名称中提取路径上的路由器位置;

3) GeoTrack 使用最后一个路由器的位置来估计目标主机的位置.

从 GeoTrack 的定位准确性来看, 90% 以上节点的定位平均误差小于 1 000 km, 总体精确度高于 GeoPing[1], 比 GeoCluster[1]差. 然而, 具有地理意义的路由器名称并不是互联网配置路由器的要求或基本属性, 而是一种由经验数据支撑的观察结论, 因此 GeoTrack 的通用性和可扩展性比较差.

另一个最经典的基于网络拓扑的定位方法之一是 TBG[56], 它是在 GeoPing 之后提出的第 1 个基于网络拓扑的系统而完整的定位技术. TBG 的地理定位计算可以分为 3 个部分.

1) 网络拓扑结构生成: TBG 使用 traceroute 工具来获取到目标的 RTT 测量值和中间路径上的网络接口识别码, 并使用端到端延迟来推断的每一跳延迟. TBG 使用这些数据计算网络拓扑结构. 同时, TBG 使用路由器别名识别来提高地理定位的准确性.

2) 约束最优化: TBG 使用基于约束的优化求解方法对目标和所有的中间路由器进行地理定位. 一共有两个约束条件: 硬延时约束和软链路延时约束[56]. 硬约束和软约束分别指的是信号传播速度和网络传输延迟的上限.

3) 将目标位置映射到最后一个受约束的路由器: TBG 通过使用最后一个受限路由器的位置作为位置估计来求解约束问题. 这种方法类似于 CBG 使用交叉区域的中心点来表示目标主机的位置, 但是规模相对较小.

TBG 是为了改进 CBG 而提出的, 基于延迟的定位方法的误差主要取决于距离目标最近的地标[55]. 一方面, TBG 提高了位置估计的一致性, 大大降低了结构化网络带来的误差, 对于结构性约束不足的网络, TBG 还整合了经过测量验证的额外定位数据来提高结果的可靠性. 另一方面, TBG 也有很多局限性和不足之处. 只有当目标主机附近存在足够多的结构约束时, TBG 才能得到高质量的定位结果. 此外, TBG 的测量成本和计算开销比简单的基于延迟的定位技术 (如 CBG 和 GeoPing) 更高. TBG 需要 traceroute 测量来计算目标周围的网络结构, 需要额外的网络测量开销来识别网络接口上的别名. 综上所述, TBG 是利用拓扑信息来推断地理位置的重要尝试, 尽管仍有许多不足之处, 但对后续研究有很大的启发作用.

在 TBG 之后还有其他基于网络拓扑的定位方法, 它们的定位计算框架与 GeoPing、TBG 有所不同. Shavitt 等人[74]通过测量互联网的 PoP (point of presence) 级别的拓扑结构来提高地理定位的准确性, 该算法结合了地理位置数据库和延迟测量的信息, 提供了比地理位置数据库更准确的地理定位结果, 同时避免了延迟测量的缺陷. Li 等人[39]提出了一种基于网络拓扑社区检测的城市级 IP 地理定位方法, 该方法利用了网络社区中的节点通常位于同一个城域网 (metropolitan area network, MAN) 的原理, 将目标 IP 的社区位置作为目标主机的地理位置. Zu 等人[65]提出了一种基于 PoP 网络拓扑的城市级地理定位算法, 该方法提取城市内的 PoP 网络拓扑并记录在 PoP 数据库中. 结果显示, 城市级地理定位的成功率提高到 97.67%, 明显高于 LBG (74.86%)[83]和 SLG (94.14%)[57]. RNBG[66]是一种基于节点排名的 IP 地理定位方法, 它利用复杂网络无标度的特性在网络中找到几个重要、稳定的节点, 然后将这些节点用于不同地区的 IP 地理定位. 在中国和美国的实验结果表明, 即使在弱连接的网络环境中, RNBG 也能达到很高的精确度. 与典型的方法相比, RNBG 的地理定位精确度提高了 2.60%–14.27%, 达到 97.55%. 李明月等人[84]提出了一种基于网络节点聚类的 IP 定位方法 NNC 方法, 该方法基于 IP 地理位置数据库投票规则确定 IP 主机所属网络社区的位置, 并依次来定位目标主机. 赵茜等人[85]将 traceroute 路径中的最后一跳路由器为地标, 根据最后一跳路由器与目标主机的网络拓扑关系确定目标主机的地理位置, 实现了 3.17 km 的平均定位误差.

根据第 3.1 节和第 3.2 节的讨论, 可以得出结论: 基于网络延迟的定位方法和基于网络拓扑的定位方法各有其优点和缺点. 基于网络延迟的方法在很大程度上依赖于地标, 当目标主机和某个地标非常接近时, 地理定位的准确性是很高的, 当目标和地标距离很远时基于网络延迟的定位方法的性能就会受到影响. 同时, 由于网络路径在地理空间可能存在回路, 延迟-距离函数难以构建, 此时基于延迟的方法就不起作用. 相反, 基于网络拓扑的定位方法对地标的需求比基于延迟的方法要低[61]. 所以基于拓扑的方法可以弥补基于延迟的地理定位方法的许多不足. 从性能的角度来看, 基于拓扑的方法使用不同维度的信息, 如拓扑结构、域名、BGP (border gateway protocol) 路由等, 多维数据的使用使得这种方法在很多情况下精确度更高. 然而, 基于拓扑的方法仍有许多缺点. 一方面, 测量开销和计算成本高, 与基于延迟的方法相比, 这种定位方法将增加网络的负担. 另一方面, 基于拓扑的方法的定位效果在一定程度上取决于

网络结构, 在网络连接良好的地区的定位方效果通常会更好, 因此这类方法的可扩展性较差, 难以推广.

5 被动的 IP 地理定位技术

被动的 IP 地理定位技术收集和分析带有地理位置信息的数据, 并依此推断出目标 IP 地址可能的地理位置. 区分主动和被动方法的主要依据是用于地理定位的数据是否原本就存在, 被动的 IP 地理定位技术用到的地理定位数据原本就存在, 而非通过新的网络测量产生. 例如, 从应用数据中获取目标 IP 的地理位置. 各种数据源可用于地理定位, 包括但不限于延迟测量数据、拓扑测量数据、DNS 数据、应用数据等. 从分析方法的角度来看, 可以使用统计方法或基于机器学习的方法. 被动的 IP 地理定位技术的难度主要在于数据处理和信息提取.

如果数据来源可靠且位置数据丰富, 那么被动的定位方法将会具有更高的定位精确度. 数据量和数据维度对于 IP 地理定位的结果都非常重要. 数据量越大, 地理定位的稳定性越高, 越不容易出现定位漂移等偶然误差. 数据维度越多, 地理定位的准确性和可信度就越高. 数据分析技术也是其中必不可少的一部分, 新出现的数据分析和数据挖掘方法对于被动的 IP 地理定位技术来说是新的机遇. 由于 IP 地理定位的对于数据的时效性要求较高, 因此有必要对数据集进行实时维护以保证定位的准确性, 或者研究新的技术来解决地理定位漂移的问题. 表 3 给出了被动的 IP 地理定位技术[1,45,68,83,86–91]对比.

表 3 被动的 IP 地理定位技术对比

方法类别	代表方法	数据集和应用范围	最佳性能 (km)	定位误差 (km)	用到的定位数据类型		
					测量数据	定位库	应用数据
基于数据库的IP地理定位技术	GeoCluster[1]	美国/欧洲	28	226	√	√	—
	LBG[83]	美国	N/A	420	√	√	—
	WBG[86]	巴西	城市级	城市级	—	√	—
	PBG[68]	N/A	3	8	—	√	—
基于网络应用的IP地理定位技术	Checkin-Geo[87]	中国	0.799	7.735	—	—	√
	Dan等人[88]	全世界	N/A	≤10	—	—	√
	DCR[89]	中国	城市级	城市级	—	—	√
	GeoBLR[90]	中国	0.08	0.232	—	—	√
		中国以外	0.179	0.311			
	ONE-Geo[45]	欧洲	N/A	0.463	—	—	√
		美国	N/A	7.795			
基于DNS的IP地理定位技术	Dan等人[91]	全世界	N/A	16.5	√	√	—

从表 3 可以看出, 被动的 IP 地理定位技术的定位精度相比于主动的方法较高, 尤其是 GeoBLR[90]. 然而被动的定位方法的定位范围相对有限, 而且大部分的优势区域集中在中国和美国的部分地区. 另外, 与主动的方法相比, 被动的方法更依赖于带有地理信息的高质量定位数据.

被动的 IP 地理定位技术可以进一步分为基于数据库的 IP 地理定位方法、基于网络应用的 IP 地理定位方法和基于 DNS 的 IP 地理定位方法. 在这些方法中, 基于数据库的 IP 地理定位方法是最容易实现的, 数据来源丰富且容易进行分析. 基于 DNS 的 IP 地理定位方法在原理上并不复杂, 但这种方法的源数据很难获取, 由于域名中包含地理位置的比例不高, 数据分析的难度也很大. 基于网络应用的定位方法非常丰富且定位效果较好, 是当前比较主流的一种定位方法.

接下来将详细介绍这 3 类被动的 IP 地理定位技术, 包括一些经典的定位方法以及对这些方法的理解.

5.1 基于数据库的 IP 地理定位技术

基于数据库的 IP 地理定位是一种朴素且常见的方法, 它通过查询一些公共的 BGP、AS (autonomous system)、WHOIS 或地理定位数据库来得到目标主机的地理地址. 除了直接查询外还有一些特定的查询策略, 结合使用多个地理位置数据库来融合得到目标 IP 的地理位置. 如, 在分配 IP 地址时, IP 地址管理机构会在数据库中记录相应地

址块的所有者信息用于路由申报和管理, 可以利用这些信息将 IP 地址映射到地理位置.

最简单的基于数据库的 IP 地理定位方法之一是直接查询 WHOIS 数据库[92]以获得地理位置, 但其准确性和稳定性较差, 无法推广到大范围使用. 因此, 在此基础上产生了 WBG[86], WBG 是一种基于 IP 地址信息的互联网主机地理定位策略, 它解决了地理定位准确性和域名注册不完整的挑战, 它结合了一系列机制和算法以最大限度地提高准确性和完整性. WBG 结合使用从 WHOIS[92]在线服务器和 CAIDA 数据库[48]中获得的信息. CAIDA 数据库包含 AS 记录, 包括 ASN (autonomous system number)、所有者的名字和它的位置信息 (经度、纬度、国家、州和城市). WBG 的流程图如图 3 所示, 详细说明如下.

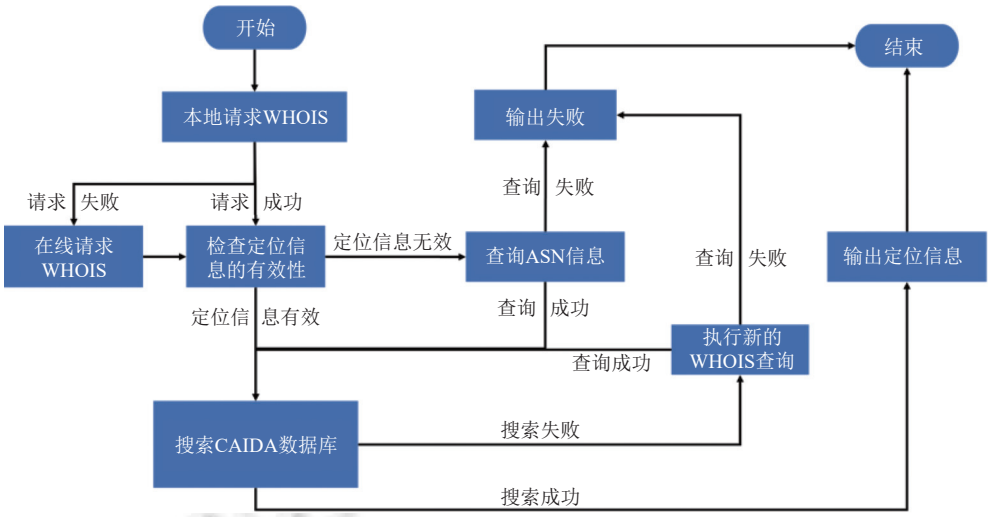

图 3 WBG 定位决策过程

1) 对于任何给定的 IP 地址, WBG 会首先查阅本地的 WHOIS 记录缓存, 如果找不到这个 IP 的条目, WBG 会使用在线 WHOIS 查询来实时搜索;

2) 如果发出的 WHOIS 查询没有返回有效的地理位置信息, WBG 会尝试找到目标 IP 地址对应的 ASN 信息, 如果找到了这样的 ASN 数据, 那么 WBG 就会使用 CAIDA 的数据库来查询获得的 ASN 来获得该 ASN 注册的地理位置信息;

3) 如果 CAIDA 数据库查询失败, WBG 就以 ASN 信息为搜索关键词, 通过发布新的 WHOIS 查询来搜索位置信息;

4) 当上述步骤都不成功时, WBG 将该 IP 地址标记为未知位置.

就准确性而言, WBG 在整体上要比 IP2Location[93]定位数据库好得多, 尽管 IP2Location 在某些局部地区的准确性要高于 WBG, 但是 WBG 的通用性、实时性和可扩展性都有一定优势.

除 WBG 外, 还有一些基于数据库的 IP 定位技术. 谢波等人[94]提出了一种 IP 定位库评估与融合算法, 并在大数据架构下基于该模型设计实现了 IP 定位库评估和融合系统, 不依赖于验证数据集的构建, 能系统地量化评估 IP 定位库定位数据的准确度, 并融合数据库间不一致的定位数据, 构建了准确度更高的融合 IP 定位库. 除了 IP 定位相关的数据库以外, 还有一些包含地理位置信息的数据库, 如 Wi-Fi 定位库. Wi-Fi 定位库指的是通过 Wi-Fi CPE (customer premise equipment) 设备上的硬件信息来推断位置. IPvSeeYou[23]使用 Wi-Fi 定位库来推断 EUI-64 (64-bit extended unique identifier) IPv6 地址的地理位置, 通过将 IPv6 地址中嵌入的硬件信息转换成 CPE 设备的 MAC 地址, 利用 Wi-Fi 定位库实现了 IP 地理定位. 由于 Wi-Fi 定位库的高精确度, IPvSeeYou 对于部分 IP 地址定位的中位误差达到了 39 m. 这一精确度非常高, 但是由于 IPvSeeYou 只能对 EUI-64 IPv6 地址做定位, 因此局限性较大, 同时, 该方法需要大量的 Wi-Fi 定位库用作分析, 因而实现的成本较高, 可扩展性较差. 郭立轩等人[95]对 IP 定位数据库和移动流量数据进行分析, 使用邻近序列的方法预测目标主机的位置, 实现了 20–30 km 的 IP 定位误差.

基于数据库的方法的最大优点是过程简单, 一些项目收集 BGP 等带有地理位置信息的数据并向所有人开放, 尽管这些数据可能并不完整, 但它们对局部地区的地理定位很有帮助. 此外, 许多 IP 指纹的数据库也会包含地理位置信息. 当然, 这种方法也有弊端. 首先, 存储在 WHOIS 和其他数据库中的信息通常是地址块或地址前缀, 对于地理定位来说其粒度过大. 并且, 很多域名注册信息是不完整的, 使得地理定位的准确性很难达到要求. 即使加上人工处理的数据, 地理定位的准确性还是远远不够的. 其次, 基于数据库的 IP 地理定位在时效性方面存在很大的问题. 因为这些数据库中记录的大部分信息都是在地址申请或变更时产生的, 它无法跟上 IP 地址的动态变化. 最后, 基于数据库的 IP 地理定位不能提供地理定位的误差范围, 这使得地理定位技术的评估更加复杂.

5.2 基于网络应用的 IP 地理定位技术

基于应用的方法是最常见的被动地理定位方法之一. IP 地理定位信息经常被用来帮助应用程序为用户提供高质量的位置服务, 因此许多应用程序数据总是包含很多地理信息, 利用这些数据可以提取许多 IP 地址的地理位置. 虽然基于网络应用的定位方法比传统的 IP 地理定位技术起步要晚得多, 定位的流程也会有所不同, 但是大量的研究工作表明, 这种方法的准确性要比传统方法高得多[12,96].

基于网络应用的 IP 地理定位方法非常多样化, 思路也非常新颖, 有很多典型的例子. Checkin-Geo[87]利用用户在位置共享服务中分享的登录日志来实现实时和准确的地理位置, 与现有的地理定位技术相比, Checkin-Geo 的定位中位误差为 799 m, 这一精确度比当时最好的方法高一个数量级. 此外, Checkin-Geo 的响应时间可以忽略不计, 这可以支撑实时的定位服务和响应时间敏感的应用. Dan 等人[88]利用从搜索引擎日志中收集的数据来评估和改进 IP 地理定位数据库, 这项工作评估了现有的 IP 地理定位数据库, 还从搜索引擎中捕获日志以提高现有地理定位数据库的准确性. Komosny 等人[97]研究了互联网通信特征及其地理定位的使用, 讨论并提出了取决于地理位置的通信属性, 如地理距离、来源国和目的地国的差异、国家人口密度和国家 ICT (information and communications technology) 发展指数. 该方法突破了主流网络性能参数 (如延迟和拓扑结构) 的框架, 整合了国家差异、人口和 ICT 指数等数据用于地理定位. Mun 等人[98]提出了一种基于众包的 IP 地理定位数据库构建方法, 分析了在线二手市场 Ruliweb[99]的信息, 构建了韩国的 IP 地理定位数据库, 可以达到较高的地理定位精度. GeoBLR[90]是一种基于贝叶斯线性回归的动态 IP 地理定位方法, 它利用用户在位置共享服务中愿意分享的位置数据来准确地实时定位动态 IP 地址. ONE-Geo[45]通过提取网络服务器的所有者名称, 进而挖掘出高度可靠的 IP 地理位置. 对于一个给定的目标 IP 地址, ONE-Geo 从网页信息和注册记录中提取真正的所有者名称, 利用这一线索, 通过搜索组织知识地图上的地址信息来确定正确的目标主机位置[45]. ONE-Geo 在 165 个网络服务器上的定位中位误差为 463 m, 在 721 个运行网站服务的节点上的定位中位误差为 7.7 km. 对于网站服务器, ONE-Geo 的地理定位效果优于当时已有方法和一些商业工具, 具体来说, 66.1% 的节点通过 ONE-Geo 实现了地理定位的误差小于 1 km. 张鹤林等人[22]提出一种基于特征选择改进的随机森林城市级 IP 定位方法, 按各特征对分类的贡献度计算权值, 根据权值将特征分为高中低 3 个区间并形成分类效果均匀的特征子集, 实验结果表明具有更高的定位准确率, 算法分类性能更好.

基于网络应用的 IP 地理定位方法的优势主要在于位置信息的来源和分析方法的多样性. 许多用户愿意在位置应用中分享他们的位置信息, 所以它的实现更多样化. 另外, 很多位置数据来源于用户提供的地理定位数据, 所以这种方法的准确性是相当可靠的. 同时, 这种方法也有几个缺点, 其中最大的问题是数据来源问题. 为了保护用户的隐私, 位置信息在很多应用中都是隐藏的, 这给定位技术的应用带来很大的挑战. 此外, 基于网络应用的 IP 地理定位方法可能会受到恶意的攻击, 如果一些错误的位置被注入到应用程序中, 定位的准确性将受到很大影响. 最后, 基于网络应用的方法在很大程度上依赖于上层应用, 所以它们的可扩展性较差, 通用性受限, 定位的覆盖度也被局限在特定范围和有限数量的 IP 地址.

与上层应用密切相关的方法很难给传统的地理定位算法提供反馈, 但它们可以为新的地理定位算法的设计提供思路, 甚至可以为定位算法的评估指标提供参考. 因此, 基于应用的方法是非常有价值的.

5.3 基于 DNS 的 IP 地理定位技术

基于 DNS 的 IP 地理定位方法主要是从 FQDN (full-qualified domain name) 中得到一些带有地理位置的域名

信息[5], 可以通过一些数据处理和查询的步骤将 IP 地址与它的地理位置联系起来.

基于 DNS 的方法既可以作为一种独立的定位技术, 也可以用作其他定位技术的辅助方案. Dan 等人[91]提出了一种系统性的方法, 使用可公开访问的反向 DNS 主机名对 IP 地址进行地理定位, 这种方法旨在与其他地理定位数据源相结合, 并将该任务变成一个机器学习问题. 对于一个给定的主机名, 计算产生一个潜在的候选位置列表并进行排序. 从实验结果来看, 该算法与学术界的经典算法、工业界的地理位置数据库进行了比较, 结果表明该方法优于学术界的方法, 并能与商业数据库形成互补的关系. 另一方面, 基于 DNS 的方法可以被用作辅助方案来帮助其他地理定位技术或服务提高其性能. IPIP.NET[9]使用 DNS 信息来提高地理定位的性能, 其他基于 DNS 的地理定位方法[1,64,75,87,100–102]通常与其他方法混合使用, 用 DNS 信息来提高精度或地理定位性能.

基于 DNS 的方法的优势比较明显. 首先, DNS 的数据源非常全面而且数据分析的方法也很多样. 更重要的是, 这种地理定位方法的准确性很高而且不需要定位地标的帮助. 然而, 基于 DNS 的方法仍然有许多缺点. 首先, 由于不是所有的 IP 地址都能找到它们对应的含有地理信息的 DNS 域名数据, 因此基于 DNS 的定位的通用性并不强. 此外, 由于 CDN (content delivery network, 内容分发网络) 的普及和 HTTP 代理的使用, 许多服务没有部署在本地, 更多的是托管在云端或数据中心. 出于安全考虑, 云网络和数据中心供应商会对真实的 IP 地址进行匿名处理[28], 所以地理定位结果会被转换到网关或其他地方. 因此, 基于 DNS 的地理定位方法的可靠性不能得到保证. 最后, 由于 DNS 数据的更新周期较长, 很多数据仅在申请前期或者发生变更的时候才会更新, 因此基于 DNS 的方法很难应对 IP 地址的动态变化, 使得地理定位数据库的维护很复杂.

6 主被动结合的 IP 地理定位技术

主被动结合的 IP 地理定位技术是一种相对全面的地理定位技术, 利用主动测量产生实时地理定位数据, 结合现有的更多样化的定位数据来提高地理定位性能. 通常来说, 主被动结合的定位方法通常包括地理定位地标挖掘, 进而形成一个完整的地理定位研究框架. 例如, 如表 4 所示, SLG[57]、XLBoost-Geo[20]和 GeoCAM[18]将延迟测量和定位数据库结合使用来完成地理定位. 实验结果表明, 将这些更全面、更复杂的地理定位数据和主动测量结合使用可以带来较好的定位效果.

表 4 主被动结合的 IP 地理定位技术对比

代表方法	应用范围	最佳性能 (km)	定位误差 (km)	用到的工具/数据		
				Ping	Traceroute	数据库
Structon[44]	中国	城市级	省级	—	√	√
SLG[57]	PlanetLab	N/A	0.69	√	—	√
	居民区	N/A	2.25			
	在线地图	N/A	2.11			
Maziku等人[46]	美国	N/A	160.93	√	√	√
GeoCAM[18]	全世界	89	258	√	—	√
Dan等人[47]	全世界	N/A	4.3	√	—	√
Corr-SLG[58]	全世界	N/A	3.34	√	√	√
GraphGeo[21]	上海/纽约/洛杉矶	0.89	3.51	√	√	√

自 2009 年以来, 主被动结合的 IP 定位方法逐渐发展壮大, 大大地推动了 IP 地理定位的发展. 近年来的许多研究工作都属于主被动结合的方法. Structon[44]是一种 IP 地理定位技术, 它利用网络内容挖掘、网络知识推断和 IP traceroute[60]来计算 IP 地址的地理位置, 在大多数情况下比现有的 IP 地理定位方法更准确. SLG[57]自动提取、验证和使用基于网络内容的位置信息以实现较高的地理定位精确度. 此外 SLG 还克服了网络延迟测量时遇到的不准确、不稳定的问题. SLG 主要通过 3 个步骤逐步实现高精确度的地理定位. 每个步骤的细节如下.

1) 使用 CBG 的变体来粗粒度的确定目标 IP 所在的大致区域;

2) 进一步聚焦目标 IP 所在的可能区域. 提出一种基于网络内容的地标挖掘方法, 由于邮政编码和地理位置存

在映射关系, SLG 使用 traceroute 工具和地标的邮政编码获取所有地标和目标之间的延迟约束;

3) 目标 IP 地址的地理定位. 根据所有地标的位置和它们与目标的估计距离, 选择与目标主机距离最小的地标并用该地标的位置表示目标主机的地理位置.

实验结果表明, SLG 的 IP 地址地理定位精度比当时最好的方法提高 50 倍, 在相应的数据集上实现了 690 m 的中位误差.

Maziku 等人[46]扩展了基于机器学习的地理定位方法, 从网络测量中提取 6 个特征并设计了新的地标选择策略. Chandekar 等人[103]提出了融合算法结合多个异构来源的位置信息, 同时估计所有主机最可能的区域. HLOC[104]旨在结合数据库的便利性和延迟测量的准确性, 从 rDNS (reverse DNS, 可逆 DNS) 名称中提取位置线索然后进行多层延迟测量. Zhao 等人[105]提出了一种基于路由器相似性和本地延迟分布的 IP 地理定位方法. 由于很多信息缺失, 当共同路由器是匿名的情况下, 所提出的方法可以提高经典的 SLG 方法的地理定位精度. RIPE IPmap[106]是一个由 RIPE NCC 运营的多引擎地理定位平台, 使用已知的地理位置来推断目标 IP 地址的地理坐标. Corr-SLG[58]是一种基于延迟-距离关系和多层共同路由器的街道级 IP 地理定位算法, 它提出了一种新的方法来收集具有街道级精确度的 IP 地理定位地标并能有效地提高地理定位的精确度. GraphGeo[21]使用图神经网络完成街道级 IP 地理定位的计算, 使用延迟、拓扑信息的同时还结合使用节点主机的 IP 知识信息, 包括 IP 地址、AS、BGP 等, 在局部区域的地理定位精确度达到 1 km 以内.

从多年的技术发展来看, IP 地理定位技术正逐渐向更全面的方向发展, 主被动结合的方式逐渐成为 IP 地理定位技术的主流. 主被动结合的 IP 地理定位技术具有明显的优势. 首先是地理定位的准确性. 由于多种地理定位技术路线的融合, 主被动结合的定位方法的精确度得到了明显的提高, 被动的方法提高了准确性, 而主动的方法能有效提高实时性. 此外, 许多主被动结合的定位方案被设计用来形成地理定位系统或在线地理定位服务. 这类方法在设计过程中已经考虑了许多边界条件, 因此具有良好的鲁棒性. 最后, 主被动结合的方法在应用范围和定位覆盖度方面优于主动和被动的定位方法. 与此同时, 主被动结合的 IP 地理定位技术也有一些不足. 由于许多方法对 IP 地理定位问题进行了抽象, 使用的技术也更加多样化, 因此这类方法需要满足更多的条件, 需要更多的测量基础设施和数据来源, 同时还需要掌握多个领域的知识. 此外, 由于对数据源和探测节点的特殊要求, 主被动结合的方法的实施成本和维护开销也相对较高.

7 研究总结及展望

目前, IP 地理定位技术在学术界和工业界有很多新的研究成果.

学术界的 IP 地理定位方法在定位精度和适用范围上存在较大限制. 当前 IP 地理定位的精度已经达到了街道级, GraphGeo[21]和 IPvSeeYou[23]等工作在一些数据集上的定位精度最高可以达到楼宇级, 平均误差达到了百米级别. 这个精度在实际使用中是非常高的, 但不是所有的定位算法都能达到这个精确度. 此外, 任何 IP 地理定位算法都有其适用范围, 在适用范围之外, 这种方法的地理定位精确度是不确定的. 由于缺乏地标的限制或者方法本身的可扩展性较差, 目前还没有可以在全球范围保持其宣称最高精确度的 IP 地理定位技术. 由于许多地理定位算法是根据不同地区的网络条件特点设计的, 要在很大范围内都实现高精度的地理定位是很困难的, 如果没有特定的网络环境作为算法的支持, 算法所达到的精确度是低于预期的. 例如 GeoCAM[18]、Spotter[61]等工作虽然在全球范围的数据集上测试其定位效果, 由于不同地区网络发展、网络资源的差异, 定位效果很难在全球范围内都达到其宣称的精确度级别和精确度范围.

工业界的 IP 地理定位技术形式多样, 目前有许多在线服务和商业数据库. 常见的例子有 IPIP.NET[9]、Maxmind[11]、NetAcuity[10]等. 其中, 中国的 IPIP.NET 是付费服务, 最高可以达到楼宇级的精度. Maxmind 和 NetAcuity 都只能在全球范围内实现城市级的定位精确度. 除了这些商业服务外, 许多研究人员将他们的学术工作开源, 其中许多开源工作已经成为商业应用的一部分. 例如, CBG[55]、TBG[56]、Octant[64]、SLG[57]、HLOC[104]、GeoCAM[18]等都有开源的工具和代码可以使用. 这些服务和开源工具使得 IP 地理定位技术的应用变得多种多样, 可以满足不同精确度的要求.

IP 地理定位的双栈支持还有待提升. 由于互联网上同时在使用 IPv4 和 IPv6 版本的 IP 协议, 因此对 IP 地理定位技术的研究不应该只关注 IPv4 地址, 需要更多的地理定位方法来支持 IPv6 地址的地理定位计算. 目前, IPv6 地理定位的研究还存在很大差距, 许多学术研究和地理定位工具都不支持 IPv6 地址的地理定位, 只有部分的定位算法是专门针对 IPv6 设计的[23,56]. 此外, 许多评估工作只针对 IPv4 地理定位技术和地理定位数据库而设计, 忽略了 IPv6 地址. 在工业应用中, IPv6 地址的地理定位精度只能支持到城市级精确度. 许多地理定位数据库甚至没有存储 IPv6 地址的定位数据, IPv6 的定位算法和服务难以推广.

主被动结合的定位技术逐渐成为主流. 从测量方法的角度来看, 主动的 IP 定位方法在使用场景和实施成本上具有不可替代的优势. 但是, 现有的主动的定位方法还存在难以克服的瓶颈, 如延迟-距离关系拟合和巨大的测量开销, 因此值得在这个方向进行深入研究. 被动的 IP 定位方法在定位的效果和思路上有其独特的先进性, 但这种方法受到的制约较多. 此外, 被动的 IP 地理定位方法在很大程度上依赖于数据分析, 因此, 更多新的数据分析方法的加入将大大促进这种地理定位技术的发展. 主被动结合的方法是近几年来的主流地理定位技术, 它结合了主动和被动方法的优点, 具有更高的地理定位精度和可扩展性.

我们总结了几个未来可能的研究要点和还未解决的问题.

7.1 更高精度的 IP 地理定位技术

为了改进现有的 IP 地理定位技术, 本文认为 IP 地理定位技术研究有两个部分值得深入讨论.

1) 更高的精确度和稳定性: 现有的地理定位技术的精确度在很多场景下都达不到应用要求. 可以使用很多新技术来提高地理定位的准确性. 例如, 使用深度学习和强化学习技术来准确拟合延迟-地理距离关系, 使用图神经网络 (graph neural network, GNN) 来更清楚地描述网络拓扑结构.

2) 完整的 IP 地理定位评估体系: 由于原理和应用范围的不同, 各种 IP 地理定位技术的优劣不能直接通过精确度进行比较. 本文总结了 5 个 IP 地理定位技术的评价指标, 进一步需要构建一个完整的评估流程和评价体系.

7.2 IPv6 地理定位技术

IPv4 地理定位技术已经非常成熟, 但很少有技术能够支持 IPv6 地址的位置计算. 由于 IPv6 和 IPv4 地址之间的巨大差异, 为 IPv4 地址设计的地理定位技术不能直接迁移到 IPv6 地址的定位计算. 因此, 一方面可以研究如何将 IPv4 地址的地理定位技术迁移到 IPv6 地址. 另一方面, 需要研究专门为 IPv6 地址设计的地理定位技术用来克服 IPv6 地址空间巨大的挑战.

7.3 多源信息融合的 IP 地理定位技术

智能决策系统的用途和特点完全符合 IP 地理定位技术的使用场景. 它可以用来快速有效地选择地标、地理定位数据库, 甚至是地理定位算法. 多源信息融合技术能够更好地利用互联网中不同结构、不同来源的位置信息和位置应用数据, 从而提高 IP 地理定位的准确性、稳定性和可扩展性. 智能决策和多源信息融合的结合将为 IP 地理定位提供新的思路和启发.

7.4 IP 地理定位地标挖掘技术

物联网设备的出现可以为新的 IP 地理定位技术和地标挖掘带来新的机会. 例如, 世界各地的网络摄像头能够提供大量可用的地标. 同时, 网络摄像头将提供图像、语音、文本和其他多媒体信息. 在图像识别、语音识别和 NLP (自然语言处理) 的支持下, 来自多媒体的数据也可用于 IP 地理定位计算. 此外, 在地标的选取方面, 深度学习技术可以帮助提取不同地标的特征, 进而为选择地标提供依据, 同时, 可以使用先进的智能决策模型来评估地标的质量.

8 总 结

本文对客户端独立的 IP 地理定位研究进行综述, 关注了不同分类的 IP 地理定位技术研究, 并整理了过去 22 年中关于 IP 地理定位和相关主题的研究. 对于每一类的研究工作, 本文介绍了各种地理定位方法的核心思想及其差异, 解释了不同 IP 定位技术的优点和缺点. 在这个过程中, 展示了该领域的主要研究是如何处理这些已知的 IP

地理定位问题. 同时, 本文回顾了这一领域的现有文献, 着眼于解决互联网主机的地理位置计算并提出了不同维度的 IP 地理定位技术的创新方法.

References:

[1] Padmanabhan VN, Subramanian L. An investigation of geographic mapping techniques for Internet hosts. In: Proc. of the 2001 Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communication. San Diego: ACM, 2001. 173–185. [doi: 10.1145/383059.383073]

[2] GPS: The global positioning system. https://www.gps.gov/

[3] BeiDou navigation satellite system. http://www.beidou.gov.cn/

[4] Padmanabhan R, Dhamdhere A, Aben E, Claffy K, Spring N. Reasons dynamic addresses change. In: Proc. of the 2016 Internet Measurement Conf. Santa Monica: ACM, 2016. 183–198. [doi: 10.1145/2987443.2987461]

[5] Luckie M, Huffaker B, Marder A, Bischof Z, Fletcher M, Claffy K. Learning to extract geographic information from Internet router hostnames. In: Proc. of the 17th Int'l Conf. on Emerging Networking Experiments and Technologies. Virtual Event: ACM, 2021. 440–453. [doi: 10.1145/3485983.3494869]

[6] Wi-Fi alliance. 2023. https://www.wi-fi.org/

[7] World Wide Web Consortium (W3C). 2023. https://www.w3.org/

[8] Ciavarrini G, Luconi V, Vecchio A. Smartphone-based geolocation of Internet hosts. Computer Networks, 2017, 116: 22–32. [doi: 10.1016/j.comnet.2017.02.006]

[9] The best IP geolocation database | IPIP.NET. 2023. https://en.ipip.net/

[10] Industry leading IP location technology | Digital element. 2023. http://info.digitalelement.com/

[11] Industry leading IP geolocation and online fraud prevention | MaxMind. 2023. https://www.maxmind.com/en/home

[12] Livadariu I, Dreibholz T, Al-Selwi AS, Bryhni H, Lysne O, Bjørnstad S, Elmokashfi A. On the accuracy of country-level IP geolocation. In: Proc. of the 2020 Applied Networking Research Workshop. ACM, 2020. 67–73. [doi: 10.1145/3404868.3406664]

[13] Gueye B, Uhlig S, Fdida S. Investigating the imprecision of IP block-based geolocation. In: Proc. of the 8th Int'l Conf. on Passive and Active Network Measurement. Louvain-la-Neuve, Belgium: Springer, 2007. 237–240. [doi: 10.1007/978-3-540-71617-4_26]

[14] Poese I, Uhlig S, Ali Kâafar M, Donnet B, Gueye B. IP geolocation databases: Unreliable? ACM SIGCOMM Computer Communication Review, 2011, 41(2): 53–56. [doi: 10.1145/1971162.1971171]

[15] Callejo P, Gramaglia M, Cuevas R, Cuevas Á. A deep dive into the accuracy of IP geolocation databases and its impact on online advertising. arXiv:2109.13665, 2022.

[16] Wang ZH, Zhang WD, Wen H, Zhu HS, Yin LB, Sun LM. A comprehensive survey of IP geolocation and evasion. Journal of Cyber Security, 2019, 4(3): 34–47 (in Chinese with English abstract). [doi: 10.19363/J.cnki.Cn10-1380/tn.2019.05.03]

[17] Wang ZF, Feng J, Xing CY, Zhang GM, Xu B. Research on the IP geolocation technology. Ruan Jian Xue Bao/Journal of Software, 2014, 25(7): 1527–1540 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4621.htm [doi: 10.13328/j.cnki.jos.004621]

[18] Li Q, Wang ZH, Tan DW, Song JK, Wang HN, Sun LM, Liu JQ. GeoCAM: An IP-based geolocation service through fine-grained and stable webcam landmarks. IEEE/ACM Trans. on Networking, 2021, 29(4): 1798–1812. [doi: 10.1109/TNET.2021.3073926]

[19] Wang ZH, Li Q, Song JK, Wang HN, Sun LM. Towards IP-based geolocation via fine-grained and stable webcam landmarks. In: Proc. of the 2020 Web Conf. Taipei: ACM, 2020. 1422–1432. [doi: 10.1145/3366423.3380216]

[20] Wang YC, Zhu HS, Wang JF, Liu J, Wang Y, Sun LM. XLBoost-Geo: An IP geolocation system based on extreme landmark boosting. arXiv:2010.13396, 2020.

[21] Wang ZY, Zhou F, Zeng WX, Trajcevski G, Xiao CJ, Wang Y, Chen K. Connecting the hosts: Street-level IP geolocation with graph neural networks. In: Proc. of the 28th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. Washington: ACM, 2022. 4121–4131. [doi: 10.1145/3534678.3539049]

[22] Zhang HL, Gan Y, Liu YB. An improved random forest city-level IP geolocation method based on feature selection. Modern Computer, 2022, 28(5): 71–75, 81 (in Chinese with English abstract). [doi: 10.3969/j.issn.1007-1423.2022.05.011]

[23] Rye EC, Beverly R. IPvSeeYou: Exploiting leaked identifiers in IPv6 for street-level geolocation. arXiv:2208.06767, 2022.

[24] Kline E, Duleba K, Szamonek Z, Moser S, Kumari W. A format for self-published IP geolocation feeds. RFC 8805. 2020. [doi: 10.17487/RFC8805]

[25] Zu SD, Luo XY, Zhang F. IP-geolocater: A more reliable IP geolocation algorithm based on router error training. Frontiers of Computer

Science, 2022, 16(1): 161504. [doi: 10.1007/s11704-021-0427-4]

[26] Dan O, Parikh V, Davison BD. IP geolocation through geographic clicks. ACM Trans. on Spatial Algorithms and Systems, 2022, 8(1): 2. [doi: 10.1145/3476774]

[27] Saxon J, Feamster N. GPS-based geolocation of consumer IP addresses. arXiv:2105.13389, 2021.

[28] Marder A, Claffy KC, Snoeren AC. Inferring cloud interconnections: Validation, geolocation, and routing behavior. In: Proc. of the 22nd Int'l Conf. on Passive and Active Measurement. Virtual Event: Springer, 2021. 230–246. [doi: 10.1007/978-3-030-72582-2_14]

[29] Roxin A, Gaber J, Wack M, Nait-Sidi-Moh A. Survey of wireless geolocation techniques. In: Proc. of the 2007 IEEE Globecom Workshops. Washington: IEEE, 2007. 1–9. [doi: 10.1109/GLOCOMW.2007.4437809]

[30] Xu GX, Gao SY, Daneshmand M, Wang CG, Liu YB. A survey for mobility big data analytics for geolocation prediction. IEEE Wireless Communications, 2017, 24(1): 111–119. [doi: 10.1109/MWC.2016.1500131WC]

[31] Elgamoudi A, Benzerrouk H, Elango GA, Landry R Jr. A survey for recent techniques and algorithms of geolocation and target tracking in wireless and satellite systems. Applied Sciences, 2021, 11(13): 6079. [doi: 10.3390/app11136079]

[32] IPv6. Wikipedia. 2023. https://en.wikipedia.org/w/index.php?title=IPv6&oldid=1151311659

[33] Koch R, Golling M, Rodosek GD. Geolocation and verification of IP-addresses with specific focus on IPv6. In: Proc. of the 5th Int'l Symp. on Cyberspace Safety and Security. Zhangjiajie: Springer, 2013. 151–170. [doi: 10.1007/978-3-319-03584-0_12]

[34] Kitchenham B. Procedures for performing systematic reviews. Eversleigh: Empirical Software Engineering National ICT Australia Ltd., 2004.

[35] Watson RT, Webster J. Analysing the past to prepare for the future: Writing a literature review a roadmap for release 2.0. Journal of Decision Systems, 2020, 29(3): 129–147. [doi: 10.1080/12460125.2020.1798591]

[36] Wenxin T, Bin C, Fan Z, Ting Z, Goce T, Yong W, Kai C. TrustGeo: Uncertainty-aware dynamic graph learning for trustworthy IP geolocation. In: Proc. of the 29th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. 2023.

[37] Jia DZ, Liu LM, Jia SJ, Lin JQ. VoteGeo: An IoT-based voting approach to verify the geographic location of cloud hosts. In: Proc. of the 38th IEEE Int'l Performance Computing and Communications Conf. London: IEEE, 2019. 1–9. [doi: 10.1109/IPCCC47392.2019.8958736]

[38] Ma T, Liu LF, Zhang F, Luo XY. An landmark evaluation algorithm based on router identification and delay measurement. In: Proc. of the 5th Int'l Conf. on Artificial Intelligence and Security. New York: Springer, 2019. 163–177. [doi: 10.1007/978-3-030-24271-8_15]

[39] Li MY, Luo XY, Shi WQ, Chai LX. City-level IP geolocation based on network topology community detection. In: Proc. of the 2017 Int'l Conf. on Information Networking. Da Nang: IEEE, 2017. 578–583. [doi: 10.1109/ICOIN.2017.7899562]

[40] Ding SC, Luo XY, Ye DP, Liu FL. Delay-distance correlation study for IP geolocation. Wuhan University Journal of Natural Sciences, 2017, 22(2): 157–164. [doi: 10.1007/s11859-017-1229-2]

[41] Yuan FX, Liu FL, Huang DH, Liu Y, Luo XY. A high completeness PoP partition algorithm for IP geolocation. IEEE Access, 2019, 7: 28340–28355. [doi: 10.1109/ACCESS.2019.2902337]

[42] Li RX, Xu R, Ma YY, Luo XY. LandmarkMiner: Street-level network landmarks mining method for IP geolocation. ACM Trans. on Internet of Things, 2021, 2(3): 21 [doi: 10.1145/3457409]

[43] Zhu G, Luo XY, Liu FL, Chen JN. An algorithm of city-level landmark mining based on Internet forum. In: Proc. of the 18th Int'l Conf. on Network-based Information Systems. Taipei: IEEE, 2015. 294–301. [doi: 10.1109/NBiS.2015.46]

[44] Guo C, Liu Y, Shen W, Wang HJ, Yu Q, Zhang Y. Mining the Web and the Internet for accurate IP address geolocations. In: Proc. of the 2009 IEEE INFOCOM. Rio de Janeiro: IEEE, 2009. 2841–2845. [doi: 10.1109/INFCOM.2009.5062243]

[45] Wang YC, Wang X, Zhu HS, Zhao H, Li H, Sun LM. ONE-Geo: Client-independent IP geolocation based on owner name extraction. In: Proc. of the 14th Int'l Conf. on Wireless Algorithms, Systems, and Applications. Honolulu: Springer, 2019. 346–357. [doi: 10.1007/978-3-030-23597-0_28]

[46] Maziku H, Shetty S, Han K, Rogers T. Enhancing the classification accuracy of IP geolocation. In: Proc. of the 2012 IEEE Military Communications Conf. Orlando: IEEE, 2012. 1–6. [doi: 10.1109/MILCOM.2012.6415842]

[47] Dan O, Parikh V, Davison BD. IP geolocation using traceroute location propagation and IP range location interpolation. In: Proc. of the 2021 Web Conf. Ljubljana: ACM, 2021. 332–338. [doi: 10.1145/3442442.3451888]

[48] CAIDA Resource Catalog. 2023. https://catalog.caida.org/search?query=types=%20dataset%20tag:caida%20

[49] Censys Search—Censys. 2023. https://censys.io/data-%20and-%20search/

[50] PlanetLab | An open platform for developing, deploying, and accessing planetary-scale services. 2023. https://planetlab.cs.princeton.edu/

[51] Feng X, Li Q, Wang HN, Sun LM. Acquisitional rule-based engine for discovering Internet-of-Things devices. In: Proc. of the 27th USENIX Security Symp. Baltimore: USENIX Association, 2018. 327–341.

[52] Sommers J. A web client perspective on IP geolocation accuracy. In: Proc. of the 2020 Int'l Symp. on Networks, Computers and Communications. Montreal: IEEE, 2020. 1–8. [doi: 10.1109/ISNCC49221.2020.9297175]

[53] Barth A. HTTP state management mechanism: RFC 6265. Internet Engineering Task Force, 2011. [doi: 10.17487/RFC6265]

[54] Song JK, Li Q, Wang HN, Sun HN. Under the concealing surface: Detecting and understanding live webcams in the wild. In: Proc. of the ACM on Measurement and Analysis of Computing Systems. Boston: ACM, 2020. 77–78. [doi: 10.1145/3393691.3394220]

[55] Gueye B, Ziviani A, Crovella M, Fdida S. Constraint-based geolocation of internet hosts. IEEE/ACM Trans. on Networking, 2006, 14(6): 1219–1232. [doi: 10.1109/TNET.2006.886332]

[56] Katz-Bassett E, John JP, Krishnamurthy A, Wetherall D, Anderson T, Chawathe Y. Towards IP geolocation using delay and topology measurements. In: Proc. of the 6th ACM SIGCOMM Conf. on Internet Measurement. Rio de Janeriro: ACM, 2006. 71–84. [doi: 10.1145/1177080.1177090]

[57] Wang Y, Burgener D, Flores M, Kuzmanovic A, Huang C. Towards street-level client-independent IP geolocation. In: Proc. of the 8th USENIX Symp. on Networked Systems Design and Implementation. Boston: USENIX Association, 2011. 365–379.

[58] Ding SC, Zhao F, Luo XY. A street-level IP geolocation method based on delay-distance correlation and multilayered common routers. Security and Communication Networks, 2021, 2021: 6658642. [doi: 10.1155/2021/6658642]

[59] Wikipedia. Ping. 2023. https://en.wikipedia.org/w/index.php?title=Ping&oldid=1129811227

[60] Wikipedia. traceroute. 2023. https://en.wikipedia.org/w/index.php?title=Traceroute&oldid=1140129266

[61] Laki S, Mátray P, Hága P, Sebők T, Csabai I, Vattay G. Spotter: A model based active geolocation service. In: Proc. of the 2011 IEEE INFOCOM. Shanghai: IEEE, 2011. 3173–3181. [doi: 10.1109/INFCOM.2011.5935165]

[62] Eriksson B, Barford P, Maggs B, Nowak R. Posit: A lightweight approach for IP geolocation. ACM SIGMETRICS Performance Evaluation Review, 2012, 40(2): 2–11. [doi: 10.1145/2381056.2381058]

[63] Du F, Bao XG, Zhang YZ, Yang HH. GeoCET: Accurate IP geolocation via constraint-based elliptical trajectories. In: Proc. of the 15th EAI Int'l Conf. on Collaborative Computing: Networking, Applications and Worksharing. London: Springer, 2019. 603–622. [doi: 10.1007/978-3-030-30146-0_41]

[64] Wong B, Stoyanov I, Sirer EG. Octant: A comprehensive framework for the geolocalization of Internet hosts. In: Proc. of the 4th Symp. on Networked Systems Design and Implementation. Cambridge: USENIX Association, 2007. 313–326.

[65] Zu SD, Luo XY, Liu SQ, Liu Y, Liu FL. City-level IP geolocation algorithm based on PoP network topology. IEEE Access, 2018, 6: 64867–64875. [doi: 10.1109/ACCESS.2018.2878309]

[66] Liu C, Luo XY, Yuan FX, Liu FL. RNBG: A ranking nodes based IP geolocation method. In: Proc. of the 2020 IEEE Conf. on Computer Communications Workshops. Toronto: IEEE, 2020. 80–84. [doi: 10.1109/INFOCOMWKSHPS50562.2020.9162976]

[67] Chen JN, Liu FL, Luo XY, Zhao F, Zhu G. A landmark calibration based IP geolocation approach. In: Proc. of the 10th Int'l Conf. on Availability, Reliability and Security. Toulouse: IEEE, 2015. 411–416. [doi: 10.1109/ARES.2015.52]

[68] Singh SP, Baden R, Lee C, Bhattacharjee B, La R, Shayman M. IP geolocation in metropolitan areas. In: Proc. of the 2011 ACM SIGMETRICS Int'l Conf. on Measurement and Modeling of Computer Systems. San Jose: ACM, 2011. 155–156. [doi: 10.1145/1993744.1993803]

[69] Dong ZQ, Perera RDW, Chandramouli R, Subbalakshmi KP. Network measurement based modeling and optimization for IP geolocation. Computer Networks, 2012, 56(1): 85–98. [doi: 10.1016/j.comnet.2011.08.011]

[70] Hillmann P, Stiemert L, Rodosek GD, Rose O. Dragoon: Advanced modelling of IP geolocation by use of latency measurements. In: Proc. of the 10th Int'l Conf. for Internet Technology and Secured Transactions. London: IEEE, 2015. 438–445. [doi: 10.1109/ICITST.2015.7412138]

[71] Li D, Chen J, Guo CX, Liu YX, Zhang JY, Zhang ZL, Zhang YG. IP-geolocation mapping for moderately connected Internet regions. IEEE Trans. on Parallel and Distributed Systems, 2013, 24(2): 381–391. [doi: 10.1109/TPDS.2012.136]

[72] Zhao F, Luo XY, Gan Y, Zu SD, Liu FL. IP geolocation base on local delay distribution similarity. In: Proc. of the 9th Int'l Symp. on Cyberspace Safety and Security. Xi'an: Springer, 2017. 383–395. [doi: 10.1007/978-3-319-69471-9_28]

[73] Feldman D, Shavitt Y, Zilberman N. A structural approach for PoP geo-location. Computer Networks, 2012, 56(3): 1029–1040. [doi: 10.1016/j.comnet.2011.10.029]

[74] Shavitt Y, Zilberman N. Improving IP geolocation by crawling the Internet PoP level graph. In: Proc. of the 2013 IFIP Networking Conf. Brooklyn: IEEE, 2013. 1–9.

[75] Wang ZH, Chen YL, Wen H, Zhao L, Sun LM. Discovering routers as secondary landmarks for accurate IP geolocation. In: Proc. of the 86th IEEE Vehicular Technology Conf. Toronto: IEEE, 2017. 1–5. [doi: 10.1109/VTCFall.2017.8288146]

[76] Zhao F, Xu R, Li RX, Zhu M, Luo XY. Street-level geolocation based on router multilevel partitioning. IEEE Access, 2019, 7:

59237–59248. [doi: 10.1109/ACCESS.2019.2914972]

[77] Hillmann P, Stiemert L, Rodosek GD, Rose O. Modelling of IP geolocation by use of latency measurements. In: Proc. of the 11th Int'l Conf. on Network and Service Management. Barcelona: IEEE, 2015. 173–177. [doi: 10.1109/CNSM.2015.7367355]

[78] Patel KB, Moukdad N, Anand S. Geolocation of IP hosts in large computer networks with congestion. In: Proc. of the 2020 IEEE Int'l Conf. on Communications. Dublin: IEEE, 2020. 1–6. [doi: 10.1109/ICC40277.2020.9149334]

[79] Wang ZH, Li H, Li Q, Li W, Zhu HS, Sun LM. Towards IP geolocation with intermediate routers based on topology discovery. Cybersecurity, 2019, 2(1): 13. [doi: 10.1186/s42400-019-0030-2]

[80] Yuan FX, Liu FL, Xu R, Liu Y, Luo XY. Network topology boundary routing IP identification for IP geolocation. In: Proc. of the 6th Int'l Conf. on Artificial Intelligence and Security. Hohhot: Springer, 2020. 534–544. [doi: 10.1007/978-3-030-57881-7_47]

[81] Beverly R, Durairajan R, Plonka D, Rohrer JP. In the IP of the beholder: Strategies for active IPv6 topology discovery. In: Proc. of the 2018 Internet Measurement Conf. Boston: ACM, 2018. 308–321. [doi: 10.1145/3278532.3278559]

[82] Rye EC, Beverly R. Discovering the IPv6 network periphery. In: Proc. of the 21st Int'l Conf. on Passive and Active Measurement. Eugene: Springer, 2020. 3–18. [doi: 10.1007/978-3-030-44081-7_1]

[83] Eriksson B, Barford P, Sommers J, Nowak R. A learning-based approach for IP geolocation. In: Proc. of the 11th Int'l Conf. on Passive and Active Measurement. Zurich: Springer, 2010. 171–180. [doi: 10.1007/978-3-642-12334-4_18]

[84] Li MY, Luo XY, Chai LX, Yuan FX, Gan Y. City-level IP geolocation method based on network node clustering. Journal of Computer Research and Development, 2019, 56(3): 467–479 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2019.20170473]

[85] Zhao Q, Chen SH. LRBG-based approach for IP geolocation. Computer Science, 2020, 47(S2): 291–295, 326 (in Chinese with English abstract). [doi: 10.11896/jsjkx.200300078]

[86] Endo PT, Sadok DFH. Whois based geolocation: A strategy to geolocate Internet hosts. In: Proc. of the 24th IEEE Int'l Conf. on Advanced Information Networking and Applications. Perth: IEEE, 2010. 408–413. [doi: 10.1109/AINA.2010.39]

[87] Liu H, Zhang YX, Zhou YZ, Zhang D, Fu XM, Ramakrishnan KK. Mining checkins from location-sharing services for client-independent IP geolocation. In: Proc. of the 2014 IEEE Conf. on Computer Communications. Toronto: IEEE, 2014. 619–627. [doi: 10.1109/INFOCOM.2014.6847987]

[88] Dan O, Parikh V, Davison BD. Improving IP geolocation using query logs. In: Proc. of the 9th ACM Int'l Conf. on Web Search and Data Mining. San Francisco: ACM, 2016. 347–356. [doi: 10.1145/2835776.2835820]

[89] Li H, Zhang P, Wang ZF, Du F, Kuang Y, An Y. Changing IP geolocation from arbitrary database query towards multi-databases fusion. In: Proc. of the 2017 IEEE Symp. on Computers and Communications. Heraklion: IEEE, 2017. 1150–1157. [doi: 10.1109/ISCC.2017.8024680]

[90] Du F, Bao XG, Zhang YZ, Wang Y. GeoBLR: Dynamic IP geolocation method based on Bayesian linear regression. In: Proc. of the 14th EAI Int'l Conf. on Collaborative Computing: Networking, Applications and Worksharing. Shanghai: Springer, 2018. 310–328. [doi: 10.1007/978-3-030-12981-1_22]

[91] Dan O, Parikh V, Davison BD. IP geolocation through reverse DNS. ACM Trans. on Internet Technology, 2022, 22(1): 17. [doi: 10.1145/3457611]

[92] WHOIS-Web.com. 2023. https://www.web.com/whois/index.jsp

[93] IP address to IP location and proxy information. IP2Location. 2023. https://www.ip2location.com/

[94] Xie B. Design and implementation of data fusion algorithm for multi-source IP geolocation databases [MS. Thesis]. Beijing: Beijing University of Posts and Telecommunications, 2019 (in Chinese with English abstract).

[95] Guo LX, Zhuo ZH, He YY, Li Q, Li ZJ. IP geolocation method based on neighbor sequence. Computer Science, 2018, 45(1): 200–204 (in Chinese with English abstract). [doi: 10.11896/j.issn.1002-137X.2018.01.035]

[96] Gouel M, Vermeulen K, Fourmaux O, Friedman T, Beverly R. Longitudinal study of an IP geolocation database. arXiv:2107.03988, 2021.

[97] Komosny D, Voznak M, Bezzateev S, Ganeshan K. The use of European Internet communication properties for IP geolocation. Information Technology & Control, 2016, 45(1): 77–85. [doi: 10.5755/j01.itc.45.1.11062]

[98] Mun H, Lee Y. Building IP geolocation database from online used market articles. In: Proc. of the 19th Asia-Pacific Network Operations and Management Symp. Seoul: IEEE, 2017. 37–41. [doi: 10.1109/APNOMS.2017.8094175]

[99] RULIWEB. https://m.ruliweb.com/

[100] Wei L, Ren GM, Shi L, Tao YC, Cao YJ. How does the recursive undns algorithm affect the accuracy of an IP geolocation system? In: Proc. of the 10th Int'l Conf. on Fuzzy Systems and Knowledge Discovery. Shenyang: IEEE, 2013. 1060–1064. [doi: 10.1109/FSKD.2013.6816353]

[101] Dan O. IP Geolocation [Ph.D. Thesis]. Lehigh University, 2019.
[102] Berger A, Weaver N, Beverly R, Campbell L. Internet nameserver IPv4 and IPv6 address relationships. In: Proc. of the 2013 Conf. on Internet Measurement Conf. Barcelona: ACM, 2013. 91–104. [doi: 10.1145/2504730.2504745]
[103] Chandekar S, Paris BP. Large-scale, discrete IP geolocation via multi-factor evidence fusion using factor graphs. In: Proc. of the 18th Int'l Conf. on Information Fusion. Washington: IEEE, 2015. 1497–1504.
[104] Scheitle Q, Gasser O, Sattler P, Carle G. HLOC: Hints-based geolocation leveraging multiple measurement frameworks. In: Proc. of the 2017 Network Traffic Measurement and Analysis Conf. Dublin: IEEE, 2017. 1–9. [doi: 10.23919/TMA.2017.8002903]
[105] Zhao F, Luo XY, Gan Y, Zu SD, Cheng QF, Liu FL. IP Geolocation based on identification routers and local delay distribution similarity. Concurrency and Computation Practice and Experience, 2019, 31(22): e4722 [doi: 10.1002/cpe.4722]
[106] Du B, Candela M, Huffaker B, Snoeren AC, Claffy K. RIPE IPmap active geolocation: Mechanism and performance evaluation. ACM SIGCOMM Computer Communication Review, 2020, 50(2): 3–10. [doi: 10.1145/3402413.3402415]

附中文参考文献:

[16] 王志豪, 张卫东, 文辉, 朱红松, 尹丽波, 孙利民. IP 定位技术研究. 信息安全学报, 2019, 4(3): 34–47. [doi: 10.19363/J.cnki.Cn10-1380/tn.2019.05.03]
[17] 王占丰, 冯径, 邢长友, 张国敏, 许博. IP 定位技术的研究. 软件学报, 2014, 25(7): 1527–1540. http://www.jos.org.cn/1000-9825/4621.htm [doi: 10.13328/j.cnki.jos.004621]
[22] 张鹤林, 甘勇, 刘渊博. 基于特征选择改进的随机森林 IP 定位方法. 现代计算机, 2022, 28(5): 71–75, 81. [doi: 10.3969/j.issn.1007-1423.2022.05.011]
[84] 李明月, 罗向阳, 柴理想, 袁福祥, 甘勇. 基于网络节点聚类的目标 IP 城市级定位方法. 计算机研究与发展, 2019, 56(3): 467–479. [doi: 10.7544/issn1000-1239.2019.20170473]
[85] 赵茜, 陈曙晖. 基于 LRBG 方法的 IP 定位研究. 计算机科学, 2020, 47(S2): 291–295, 326. [doi: 10.11896/jsjkx.200300078]
[94] 谢波. 多源 IP 定位库数据融合算法设计与实现 [硕士学位论文]. 北京: 北京邮电大学, 2019.
[95] 郭立轩, 卓子寒, 何跃鹰, 李强, 李舟军. 基于邻近序列的 IP 地址地理定位方法. 计算机科学, 2018, 45(1): 200–204. [doi: 10.11896/j.issn.1002-137X.2018.01.035]

林金磊(1998－), 男, 博士生, 主要研究领域为网络测量, 网络安全, 数据挖掘.

樊琳娜(1987－), 女, 博士, 讲师, 主要研究领域为物联网, 机器学习.

李城龙(1985－), 男, 副研究员, CCF 专业会员, 主要研究领域为网络空间测绘, 网络信息安全.

王之梁(1978－), 男, 副教授, CCF 专业会员, 主要研究领域为计算机网络体系结构, 网络形式化验证和测试, 网络测量.

宋光磊(1994－), 男, 博士, CCF 专业会员, 主要研究领域为网络空间测绘, 网络安全.

杨家海(1966－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为计算机网络体系结构, 网络管理与测量, 网络空间安全与测绘, 云计算.