

目标网络场景自适应的 IP 定位框架

祖铄迪^{1,2}, 丁世昌^{1,2}, 袁福祥^{1,2}, 罗向阳^{1,2}

(1. 网络空间态势感知河南省重点实验室, 河南 郑州 450001; 2. 信息工程大学, 河南 郑州 450001)

摘要: 通过 IP 来定位目标位置是基于位置的服务的重要基础。当前研究者面向不同的网络场景, 提出了不同实现原理的 IP 定位方法。但在不同的网络场景中, 各类 IP 定位方法难以维持理想的定位效果。介绍了基于网络测量的 IP 定位的 3 种典型实现原理, 分析了各类方法在不同网络场景下的优缺点, 并提出了一种目标网络场景自适应的 IP 定位框架。该定位框架将目标在地标库中进行位置对比, 初步判定目标的的城市级位置并分布式部署探测源; 随后获取目标所在城市的时延、拓扑、同子网地标等属性, 并判定其所处网络场景; 根据目标网络场景, 选择合适类型的定位算法对目标 IP 进行定位, 输出位置估计结果。通过在我国 11 个城市进行模拟定位实验, 得到了在不同数量和分布的地标数据支撑下各类 IP 定位方法的多层次表现, 结果显示所提框架的城市级定位成功率可达 96.16%, 街道级定位中值误差可达 4.13 km, 同时所提框架在不同的同子网地标数量和目标可达性条件下均能保持稳定的定位效果。实验结果验证了所提框架的有效性, 并为 IP 定位研究提供新的思路。

关键词: IP 定位; 网络场景; 网络测量; 网络拓扑; 网络地标

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2023084

Adaptive IP geolocation framework for target network scenarios

ZU Shuodi^{1,2}, DING Shichang^{1,2}, YUAN Fuxiang^{1,2}, LUO Xiangyang^{1,2}

1. Key Laboratory of Cyberspace Situation Awareness of Henan Province, Zhengzhou 450001, China

2. Information Engineering University, Zhengzhou 450001, China

Abstract: The target location can be determined through IP geolocation, serving as a vital foundation for

收稿日期: 2023-01-09; 修回日期: 2023-05-01

通信作者: 罗向阳, luox_y_ieu@sina.com

基金项目: 国家自然科学基金 (1804263, 62172435); 国家重点研发计划 (2022YFB3102904); 中原科技创新领军人才计划 (214200510019); 河南省科技攻关项目 (222102210036, 232102211052); 河南省重点研发与推广专项 (232102211052)

Foundation Items: The National Natural Science Foundation of China (1804263, 62172435), The National Key R&D Program of China (2022YFB3102904), Zhongyuan Science and Technology Innovation Leading Talent Project (214200510019), The Key Science and Technology Project of Henan Province (222102210036, 232102211052), Henan Province Key Research, Development and Promotion Project (232102211052)

引用格式: 祖铄迪, 丁世昌, 袁福祥, 等. 目标网络场景自适应的 IP 定位框架[J]. 网络与信息安全学报, 2023, 9(6): 71-85.

Citation Format: ZU S D, DING S C, YUAN F X, et al. Adaptive IP geolocation framework for target network scenarios[J]. Chinese Journal of Network and Information Security, 2023, 9(6): 71-85.

location-based services. Researchers have proposed various IP geolocation algorithms with different implementation principles to cater to different network scenarios. However, maintaining an ideal geolocation effect proves challenging for these algorithms in diverse network scenarios. Three typical implementation principles for IP geolocation based on network measurement were introduced. The advantages and disadvantages of these methods in various network scenarios were analyzed, and an adaptive IP geolocation framework specifically tailored to target network scenarios was proposed. The geolocation framework was functioned as follows: initially, a preliminary city-level location estimation was obtained by comparing the target's location with the landmark database. Then, detection sources were deployed in a distributed manner, and information such as delay, topology and same subnet landmarks for the target city was gathered to determine the network scenario. Finally, an appropriate geolocation method was employed to accurately estimate the target's location according to the identified network scenario. Through simulation geolocation experiments in 11 cities of China, the multi-level performance of various IP geolocation methods supported by landmark data of different quantities and distributions was evaluated. The results indicate that the proposed framework achieves a city-level geolocation success rate of 96.16% and a median error of 4.13 km for street-level geolocation. Moreover, the framework demonstrates stable geolocation performance across different conditions, including varying same subnet landmark numbers and target accessibility. These experimental findings validate the effectiveness of the proposed framework and offer novel insights for IP geolocation research.

Keywords: IP geolocation, network scenario, network measurement, network topology, network landmark

0 引言

近年来, 基于位置的服务 (LBS, location-based service) 的普及给人们的生活带来了巨大便利。在获得用户位置方面, 除利用全球导航卫星系统 (GNSS, global navigation satellite system) 进行定位外, 通过用户的互联网协议 (IP, internet protocol) 来确定位置也是一种常见的方法。IP 定位适用于家用个人计算机 (PC, personal computer) 用户、各类服务器和路由器等具备公网 IP 的固定互联网接入设备的地理定位^[1-3]。

本文涉及的 IP 定位专注于设备独立的位置, 其中设备的位置由其已知的 IP 地址远程估计, 而不由设备本身来估计, 如使用 GNSS 模块。本文研究无用户协助的设备定位, 满足以下前提: ① 不使用 GNSS 模块; ② 不使用地面无线电定位系统, 如 Wi-Fi 或移动蜂窝网络中的三角测量; ③ 不使用用户输入的位置信息。

网络测量是指用网络探针和各类工具软件, 获取 IP 中时延、流量、拓扑等各类参数的行为^[4], 而基于网络测量的 IP 定位, 是 IP 定位方向的研究热点之一。基于网络测量的 IP 定位可以采用主动测量和被动测量相结合的方式, 获得从探针到待定位 IP 间的路径、时延和拓扑

等信息, 综合分析后判断得出该 IP 所处的地理位置范围^[5-6]。基于网络测量的 IP 定位在运作时, 不需要待定位 IP 主动提供其位置、设备型号、用户数据等信息, 只要该 IP 确实存在于公共互联网, 即可对其进行定位。由于基于网络测量的 IP 定位具备不依赖用户协助的特点, 在商业广告投放、网络设备管理、网络安全等领域得到了广泛应用^[7-10]。

受到基础设施发展程度、设备安全规则、传输负载均衡等因素影响, 互联网中的 IP 设备处于各式各样的网络环境。通信时延能否保证、网络接入服务层是否发达、终端安全策略是否严格等都会影响定位方法对 IP 目标的定位效果。因此, 为应对不同网络场景, 研究者提出了实现途径各异的 IP 定位方法。然而, 现实网络环境较为复杂, 各类 IP 定位算法在不同网络场景中难以维持理想的定位效果^[11]。

为改善在不同网络场景下 IP 定位的效果, 本文介绍了基于网络测量的 IP 定位的 3 种典型实现方式, 并提出了一种目标网络场景自适应的 IP 定位框架。本文的主要工作如下。

1) 介绍基于拓扑与时延分析的 IP 定位、基于同子网地标分布的 IP 定位和基于局部网络结构的 IP 定位 3 种典型实现方法的基本原理,

并分析这3种方法在不同网络环境下的优势和劣势。

2) 提出一种目标网络场景自适应的IP定位框架。该框架通过地标库对目标位置进行初步判定后,部署探测源多源探测目标,根据探测数据分析归类目标所处的网络场景,并调度合适的IP定位方法实施目标位置估计。

3) 分析同子网地标与目标可达性等目标网络场景特性对IP定位过程的影响,并进行相关的数据测试。

4) 在我国11个城市进行了定位实验,得到在不同网络场景下各类定位方法的表现。实验结果显示,本文框架能够有效应对同子网地标数量变化与目标可达性变化的情况。

1 IP定位的3种典型实现方法

1.1 基于拓扑与时延分析的IP定位

基于时延测量的定位原理基于一个假设:报文在两个网络节点间的传播时延与两个节点间的实际地理距离呈正相关关系。然而,受时延抖动的影响,在较小的地理尺度内,时延和地理距离的线性关系并不明显^[12]。因此,部分研究者采用拓扑与时延相结合的方式,在大尺度上使用时延确定位置范围,在小尺度上采用拓扑路径与单跳时延结合分析的方式确定目标位置。而当目标存在多个同级别共同路由器时,再次选择相对时延最小的路由器作为位置估计。代表算法有SLG(street-level geolocation)^[13]、RMP-Geo(router multilevel partitioning based geolocation)^[14]和DDT-Geo(delay deviation tolerance geolocation)^[15]。

基于拓扑与时延分析的IP定位示意如图1所示,在获取目标所处周边拓扑及时延信息后,开始对目标实施定位。定位者对目标进行路径探测,得到与目标 T_1 直接相连的路由器 R_1 。统计与 R_1 直接相连的地标,并探测统计这些地标与 R_1 间的时延,挑选时延最小的地标 L_3 作为目标的位置估计。

该类IP定位技术采用时延与拓扑结合分析的方式,在时延膨胀不明显的地区能得到较好的定位效果。末端定位时,采用拓扑与时延相结合的方式,能够避免共同路由地标数量过

多时,可能导致的误差偏大的问题。但是,当时延膨胀普遍存在时,此类方法的误差会受到影响。

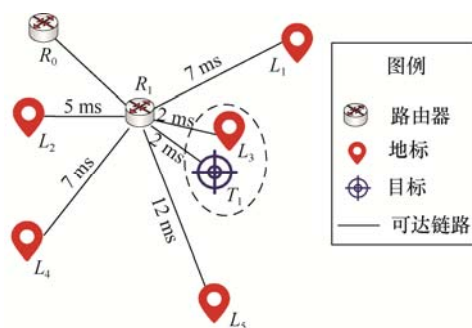


图1 基于拓扑与时延分析的IP定位示意
Figure 1 Diagram of IP geolocation based on topology and delay analysis

1.2 基于同子网地标分布的IP定位

为便于部署、管理和维护网络设施,网络服务提供商往往会将整段IP划分为不同的子网或地址块,并分配给不同的网络用户组。常见的做法是将IP地址前24位相同的IP划分到同一子网,即/24子网。而对于每个子网,通常会设置对应的接入路由器负责子网内终端的网络服务。Mukne等^[16]通过研究发现,同一/24子网中的IP在现实中往往集中在同一地理区域。因此,一些研究者利用该分布特性,在获取待定位IP后,在IP位置库中搜索同一子网IP的记录,并通过这些IP的位置来估计目标的位置。典型算法有GDML(geolocating databases based on machine learning)^[17]、GDMF(geolocation from arbitrary database query towards multi-databases fusion)^[18]。

基于同子网地标分布的IP定位示意如图2所示,在得到目标 T_1 的IP后,查询位置库得到与其处于同一子网的地标 L_1 至 L_7 ,进而计算这些地标坐标分布范围,以此作为目标的位置估计。而在此位置范围内的其他子网地标,如 L_8 至 L_{11} ,此类方法会将其视为噪声并去除。

此类IP定位方法通过查询IP位置数据库、比较各种数据库以及数据库之间的数据融合来获得定位结果。此类方法操作直观、便于理解,且实现步骤较少,方便多平台移植。但此类方法的定位效果同样受到数据库的限制。一方面,很难确保数据库的准确性。数据库只提供查询服务,不

能保证查询信息的真实性。事实上,现有文献指出,不同数据库的查询结果不一致^[19-20]。另一方面,数据库没有及时更新。地理位置数据通常来自多年前的登记数据,数据的及时性很差^[21]。

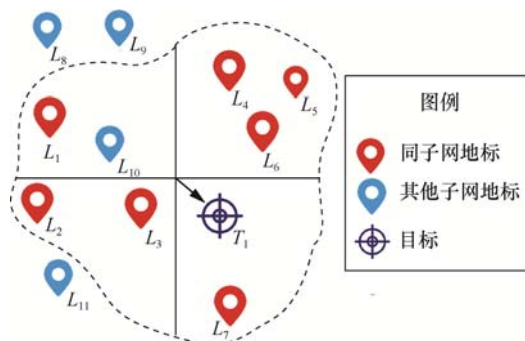


图2 基于同子网地标分布的IP定位示意
Figure 2 Diagram of IP geolocation based on same subnet landmark distribution

1.3 基于局部网络结构的IP定位

在实际互联网中,出于防范网络攻击或节省计算资源的考虑,部分防火墙和安全软件不对ICMP等网络控制报文进行响应。这种情况导致在无用户协作条件下,定位者无法准确获得这类探测不可达IP目标在城市网络拓扑中的位置,进而无法对此类目标进行较高精度的地理定位。部分研究者通过提取目标城市的因特网接入点(PoP, point of presence)网络,并与探测路径对比来估计不可达目标的位置。PoP是介于路由器级网络和城域网级网络间的网络层级,包含特定区域内由同一因特网服务提供者(ISP, Internet service provider)负责网络接入的所有终端。因PoP能够对应一个特定区域的终端用户群,因此被用来对不可达IP目标进行定位。而由于PoP网络中的节点具有紧密的连接特性,因此研究者往往通过在城域网拓扑中搜索特定局部网络结构来提取PoP网络。代表算法有TBG(topology based geolocation)^[22]、PBG(PoP based geolocation)^[23]。

基于局部网络结构的IP定位示意如图3所示, PoP网络 P_1 包含构成紧密结构bi-fan^[24]的4个路由器 $R_1 \sim R_4$ 。对目标 T_1 进行路径探测,并与PoP网络库对比。在确定其路径经过 P_1 后,将 P_1 的位置范围作为目标的位置估计。

此类IP定位技术在目标IP不可达时,使用局部网络结构取代共同路由器对目标进行

定位,提高了定位成功率。但由于PoP层级介于路由器和城域网之间,对应的终端分布范围相对较大,因此定位误差往往大于前两类定位方法。

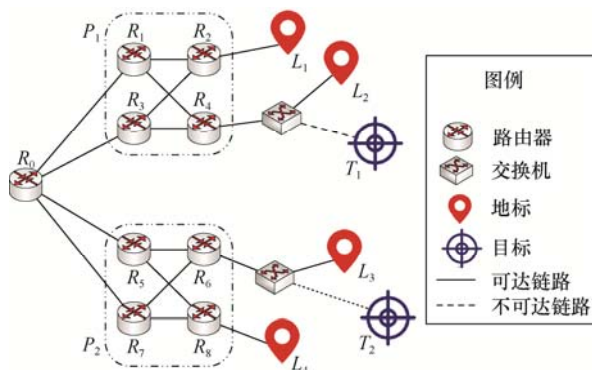


图3 基于局部网络结构的IP定位示意
Figure 3 Diagram of IP geolocation based on local network structure

2 目标网络场景自适应的IP定位

2.1 定位框架

为解决各类定位方法在不同网络场景中定位效果不稳定的问题,本文提出了一种目标网络场景自适应的IP定位框架。现有典型定位方法各自有其适用的网络场景,而当网络场景发生变化时,其定位效果将受到影响。本文定位框架的核心思想是:在实施定位前,通过对目标所处网络环境进行分析,划定其所处的网络场景,并采用相应的IP定位方法对其实施位置估计。如此实施定位,在大多数网络环境下,能够保持较好的定位效果。

在上述思想的指导下,本文定位框架首先根据目标所处的大致位置分配探测源并对目标发起多源探测;然后,探测目标所处城市地标,获取目标城市的时延、路径、拓扑等信息,并确定其所处的网络场景;随后,依次判定同子网地标数量和目标可达性,并调用对应的定位算法;最后,对目标IP进行定位,输出位置估计结果。本文提出的目标网络场景自适应的IP定位框架如图4所示,主要分为如下4个部分。

1) 目标多源探测。获取待定位目标的IP后,在地标库中搜索同子网的地标,并记录这些同子网地标所对应的城市。在探测源库中选择不属于上述同子网地标所处城市的设备,作为定位目标时的探测源,并对目标发起探测。

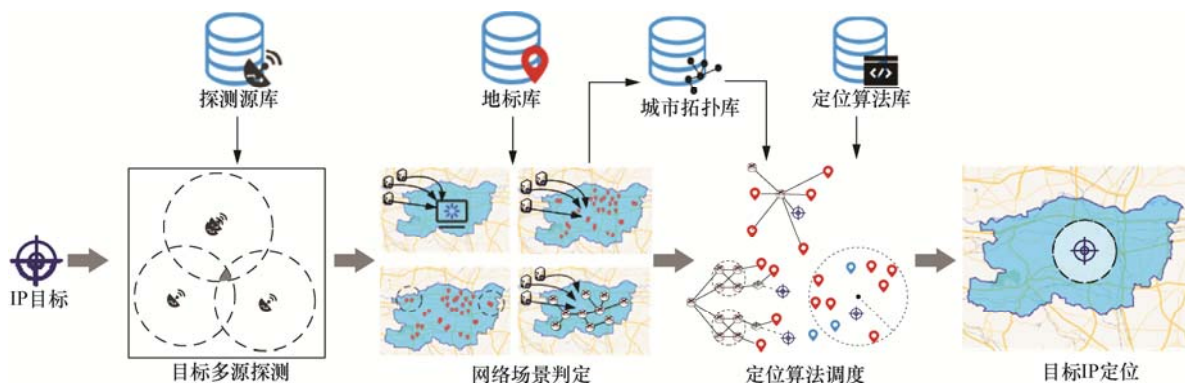


图4 目标网络场景自适应的IP定位框架
Figure 4 Adaptive IP geolocation framework for target network scenarios

2) 网络场景判定。遍历地标库，对处于目标城市的地标发起网络探测，记录时延、路径、可达性等属性，并生成目标城市的城市拓扑库。通过城市拓扑库分别计算目标IP所处拓扑的同子网地标、时延、共同路由地标及局部网络结构，确定其所处的网络场景。

3) 定位算法调度。依次计算或判定目标所处网络场景中同子网地标数量与位置、目标探测路径中共同路由位置、目标探测路径是否抵达终点，按条件优先级在定位算法库中选择合适类型的定位算法。

4) 目标IP定位。定位数据预处理后，使用上一步选择的定位算法对目标进行位置估计，并输出定位结果。

2.2 目标多源探测

在本文的定位框架中，探测源需要对目标IP和地标发起网络探测，并收集目标所处网络环境中时延、路径、拓扑、共同路由地标等信息。而探测路径是上述信息中最基础的信息之一。基于时延和拓扑的定位以及基于局部网络结构的定位均采用路径匹配的方式来确定目标IP在城域网拓扑中的位置，从而实现对目标的定位。路径匹配是指将目标探测路径与城域网拓扑进行对比，找出同时属于两者的共同路由节点的过程。但当探测源与地标或目标处于同一城市时，会在路径匹配后，得到位于探测源附近的共同路由节点。这种共同路由节点可能使得最后的定位结果位于探测源而非目标附近。

为避免此类情况，应当确保探测源与地标或目

标位于不同的城市。现有研究表明，同一/24子网中的IP在现实中往往集中在同一地理区域^[16]。因此，可以采用这一特性，在地标库中寻找与目标属于同一子网的地标，并尽量避免选择与这些地标属于同一城市的探测源。本文通过IPIP官网查询到中国的IP地址段，按整个网段采样了100万个IP，并对这些地标进行了“子网-城市级位置”统计，/24子网包含的城市数量分布如图5所示。

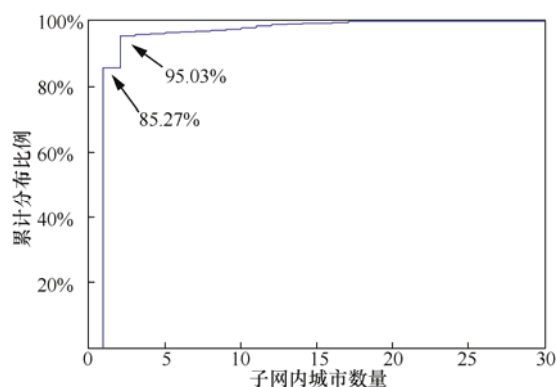


图5 /24子网包含的城市数量分布
Figure 5 Distribution of cities' count included in a /24 subnet

经统计，85.27%的/24子网只包含1个城市级位置，包含至多2个城市级位置的子网占总数的95.03%。而包含2个城市级位置的子网中，包含城市为相邻城市或同省城市的占比为90.48%。

探测源部署流程主要包括两部分。

1) 目标初步位置估计。截取目标IP前24位，记为目标的/24网段。在地标库中搜索同样属于该网段的地标，记录这些地标的城市级位置。按出现次数高低排序这些城市，选择第一个为目标的

可能城市级位置。

2) 探测源分布式部署。在探测源库中选择不属于目标可能城市级位置的设备,启动这些设备作为本次定位的探测源,并向目标发起探测。

探测源部署包括对目标 IP 的初步位置估计和探测源的选择。在执行完本部分后,可以在避免同城市探测的前提下,实现对目标 IP 的多源探测,得到的探测数据可以为合理选用定位算法奠定基础。

2.3 网络场景判定

受现实中运营商服务能力、网络基础设施性能、终端安全设置等影响,互联网中的 IP 设备所处的网络环境各自不同。流行的 IP 定位算法基本是为了定位处于特定网络环境的目标 IP。因此,在不适用的网络环境中,定位性能会有所下降。而在定位前,探测目标 IP 所处的网络环境,并划分为不同的网络场景,从而因地制宜采用合适的定位算法实施目标的位置估计,是本文的核心思想。

有 3 类流行的 IP 定位方法:基于同子网地标的 IP 定位、基于局部网络结构的 IP 定位和基于时延与拓扑的 IP 定位。基于同子网地标的 IP 定位方法便于理解,易于实现与移植,且定位效果较为稳定。但当同子网地标数量足够多后,与其他类型的定位方法相比缺乏优势。基于局部网络结构的 IP 定位通过对网络紧密结构的提取,获取了介于城域网与路由器之间的网络层级——PoP 层,这样做的优势在于扩大了路径匹配的规模,提高了匹配的成功率。这种做法在目标不可达或城域网中存在较多匿名节点时,能够有效提高定位成功率。另外,正因为扩大了路径匹配的规模,此类定位方法获得的共同路由在城域网拓扑中的位置不够精确,进而在统计意义上增大了定位误差。而基于时延与拓扑的 IP 定位是 IP 定位较经典的实现途径,能够应对大部分网络场景,但与此同时,在上述特定网络场景中,该类定位方法的表现不够出色。

为了更好地定位不同网络环境下的 IP 目标,本文的定位框架结合了上述 3 类定位方法,通过对目标所处网络环境选择性地调用不同的定位算法实施目标的位置估计。网络场景判定流程主要包括 3 部分。

1) 城市拓扑库生成。遍历地标库,枚举位于目标可能城市级位置的地标,通过探测源对地标发起探测,收集目标城市的时延、路径、可达性等属性,并生成目标城市的城域网有向拓扑,录入城市拓扑库。

2) 同子网地标数量判定。遍历地标库,枚举目标 IP 的同子网地标。若其城市级位置一致的同子网地标数量大于 20,则目标处于同子网地标丰富网络场景,反之则未处于。

3) 目标可达性判定。检查对目标的多源探测路径。若路径均未抵达目标,则目标处于探测不可达网络场景;反之则未处于,舍弃未抵达目标的探测路径。

网络场景判定通过对 IP 目标网络环境数据的采集与分析,完成了对其所处网络场景的划分,为定位算法调度提供参考依据。

2.4 定位算法调度

在网络场景判定中,得到了 IP 目标所处城市的网络属性,并完成了对目标周边网络场景的判定。在该部分中,需要依据目标网络场景来调用合适的定位算法。

对本文所提 3 类定位方法进行大量测试,发现在目标未处于同子网地标丰富网络场景时,基于同子网地标的定位方法误差会更低一些;在目标处于探测不可达网络场景时,基于时延和拓扑分析的 IP 定位成功率会明显下降。

图 6 展示了定位算法调度与网络场景判定、目标 IP 定位间的关系。如图 6 所示,蓝色框体表示网络场景判定中的操作步骤,绿色框体表示 IP 定位中的操作步骤和定位方法。本节负责按照优先级设置,连接上述两部分内容。定位算法调度的主要步骤如下。

1) 若目标处于同子网地标丰富网络场景,则进入下一步,否则调用基于同子网地标的定位算法对 IP 目标实施定位。

2) 若目标处于探测不可达网络场景,则调用基于局部网络结构的定位算法对 IP 目标实施定位,否则调用基于时延与拓扑分析的定位算法对 IP 目标实施定位。

在确定需要调用的定位算法后,定位框架将对 IP 目标实施定位。

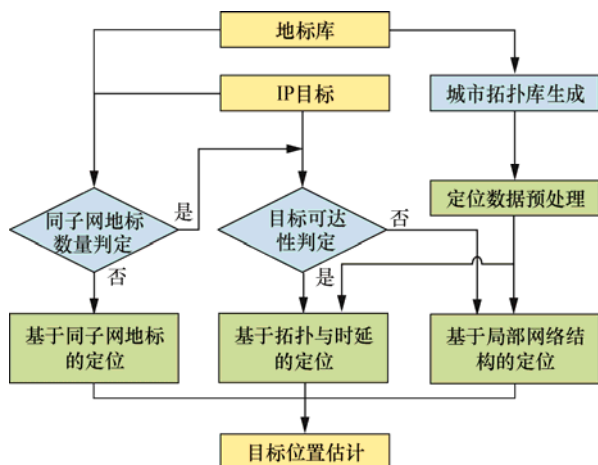


图6 定位算法调度与网络场景判定、目标IP定位间的关系
Figure 6 Connection among geolocation algorithm scheduling, network scenario judgment and target IP geolocation

2.5 IP目标定位

在判定目标所处的网络场景后，本节采用对应的定位方法对IP目标实施位置估计。下面分别介绍几种定位方法的实现流程。

1) 定位数据预处理。将目标探测路径与城市拓扑库进行对比，获得距离目标跳数最小且出现在城市拓扑库的路由节点，记为共同路由。枚举共同路由的所有子节点中的地标节点，记为共同路由地标。对于城域网拓扑，按照PBG^[23]中的方法，进行bi-fan结构的归并，生成PoP网络，并录入城市拓扑库。

2) 基于同子网地标的定位。对于在2.2节中得到的同子网地标，舍弃不属于目标可能城市级位置的地标。计算这些地标的计算中心，并记为目标定位结果。

3) 基于时延与拓扑分析的定位。枚举IP目标的共同路由地标，在城市拓扑库中，计算上述地标与共同路由间的相对时延。当相对时延小于0时，赋值为0；当相对时延最小的地标数量为1时，将该地标位置记为目标定位结果，当数量大于1时，计算这些地标的计算中心，并记为目标定位结果。

4) 基于局部网络结构的定位。对于IP目标的共同路由，在城市拓扑库中查找其包含该节点的PoP网络。对于该PoP网络所包含的路由节点，枚举其所有子节点中的地标节点，记为PoP地标。在去重后，计算这些地标的计算中心，并记为目标定位结果。

在基于时延与拓扑分析的IP定位中，对路径中的时延信息进行改良，即当相对时延小于0时使其为0，这是一种改善网络中时延膨胀的简易方法。在基于局部网络结构的IP定位中，对PBG算法进行优化，通过加入地标计算中心的步骤，进一步提高算法的定位精度。按照本文的定位框架对目标实施IP定位，能够根据其所处的网络环境，尽可能地选择合适的方法完成目标的位置估计。

3 面向IP定位的网络场景特性分析

本节分析了同子网地标地理分布、数量、分布密度和目标可达性对IP定位过程及结果的影响。

3.1 同子网地标地理分布

出于管理开销及维护成本的考虑，网络服务提供商普遍会把整段IP分配给一个特定区域内的终端，这不是学术规范，而是商业公司或基础设施提供者为了降低成本而普遍采用的做法。两类常见的IP分配方式如图7所示。

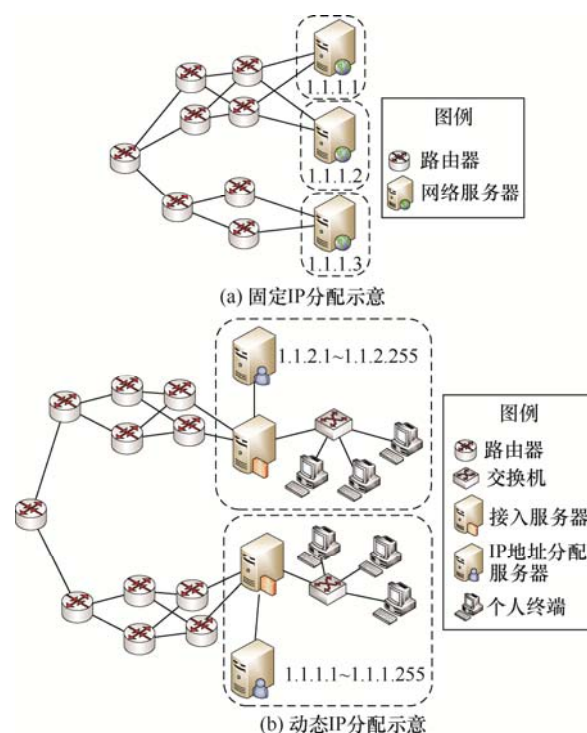


图7 两类常见的IP分配方式
Figure 7 Two common IP allocation approaches

对于需要持续在线并提供服务或有其他特殊

需求的终端或用户,ISP 往往会向其提供固定 IP。而对于大多数普通上网用户,ISP 会提供互联网的动态接入服务,待其申请连接互联网后,再在 IP 池中挑选 IP 并分配给该用户。而 IP 池通常为连续的一整段 IP。为应对日益增加的上网用户数量,部分 ISP 会提供端口转换服务,即一个公网 IP 对应多个终端用户。由于该类服务对本节同子网 IP 的地理分布分析没有影响,在此不多做描述。

对于图 7(a)的情况,同子网终端的地理分布往往取决于 ISP 的地址分配策略。而在大多数情况下,为便于管理,整个网段的 IP 会在被划给固定城市的分公司后,再进行分配。对于图 7(b)的情况,同子网的终端基本由同一网络接入服务器提供网络接入服务。而对应现实中,为方便管理和维护,接入服务器会为其周边区域内的用户提供接入服务。综上所述,同子网地标的地理分布往往趋向于同一地理区域。

为进一步研究同子网终端的分布情况,本文采集了我国 4 个城市的地标,并按子网分组,分别计算每个子网下地标的分布范围。我国 4 个城市同子网地标分布半径统计如图 8 所示。

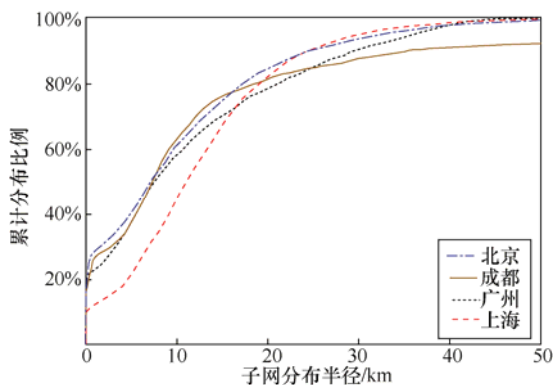


图 8 我国 4 个城市同子网地标分布半径统计
Figure 8 Statistics on the distribution radius of same subnet landmarks in four Chinese cities

可以看出,至少 40%左右子网的 IP 处于以 10 km 为半径的范围内,80%左右子网的 IP 处于以 20 km 为半径的范围内。在我国的大城市中,前者相当于一个或者数个区的面积(位于我国大城市老城区的区普遍面积较小)。因此,本文定位框架的目标多源探测部分,通过对同子网地标的地理分布来初步估计目标的城市级位置。

3.2 同子网地标数量

由上文分析得知,同子网地标趋向于分布在某一特定区域。然而,同子网地标的数量差异会导致该子网分布范围的差异。图 9 展示了地标数量导致的分布范围差异示例。如图 9 左图所示,红色图片代表该子网中被用于计算分布范围及计算中心的 IP,蓝色图标代表未被选择的 IP。而在图 9 右图中,右下的 IP 由未选中变为选中,该子网的分布范围变大,而中心随之移动。因此,同子网地标数量对定位是有影响的。

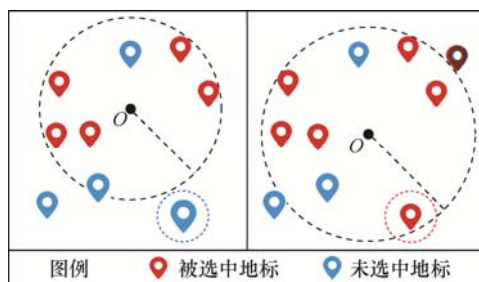


图 9 地标数量导致的分布范围差异示例
Figure 9 Example of difference in distribution range caused by the number of landmarks

为了探究同子网地标数量对定位结果的影响,本文对我国 4 个城市的地标进行了统计,我国 4 个城市同子网地标数量统计如图 10 所示。

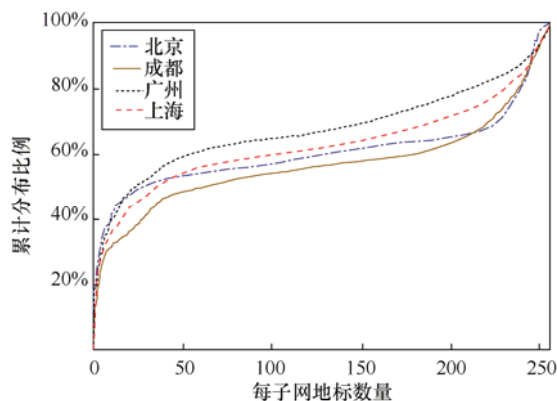


图 10 我国 4 个城市同子网地标数量统计
Figure 10 Statistics on the number of same subnet landmarks in four Chinese cities

从图 10 可以看出,每个城市几乎有 50%以上子网内的地标数量少于 100。随后,在上述城市随机选择 20 个子网,每个子网依次取 1、2、5、10、20、50、75、100、150、200 个地标,并以此为输入,分别使用基于子网和基于时延拓扑的方法进行定位,同子网地标数量与拓扑节点数量

和定位误差间关系分别如图 11 和图 12 所示。

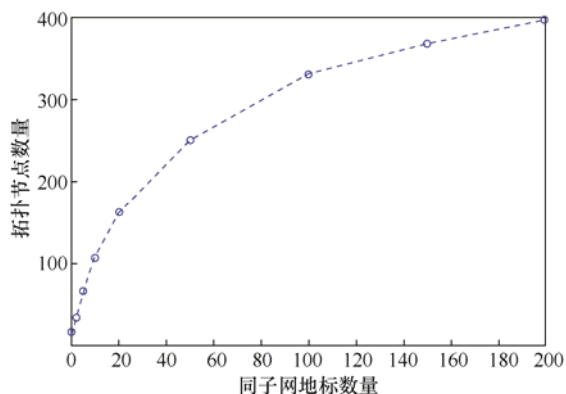


图 11 同子网地标数量与拓扑节点数量间关系
Figure 11 The relationship between the number of same subnet landmarks and the number of topology nodes

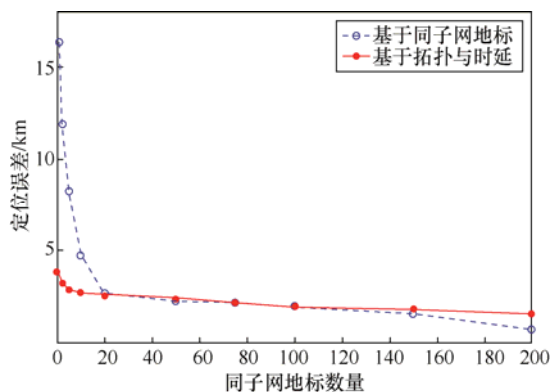


图 12 同子网地标数量与定位误差间关系
Figure 12 The relationship between the number of same subnet landmarks and geolocation errors

从图 11 可以看出, 拓扑节点数量与同子网地标数量基本呈现正相关关系。在同子网地标数量大于 20 后, 拓扑节点数量的增长速度开始放缓。而拓扑节点的数量能够反馈到基于共同路由和基于拓扑时延的定位结果中。从图 12 可以看出, 随着同子网地标数量的增加, 定位误差逐渐下降。同时当同子网地标数量大于 20 后, 基于拓扑与时延的定位误差开始逐渐低于基于同子网地标的定位误差。因此, 在本文的定位框架中, 在调用基于子网地标的定位方法时, 将同子网地标数量上限确定为 20。

3.3 同子网地标密度

上文分析了同子网地标分布和数量对定位结果的影响, 而地标密度为地标数量与分布范围面积之比。本节对同子网地标密度对定位误差的影响进行分析。

为便于理解, 这里以地标均匀分布的理想情况为例。图 13 展示了同子网地标分布范围、数量与密度所对应的不同情况。图 13 左图中有一组均匀分布的地标; 如图 13 中间图所示, 若将蓝色地标从地标组中删除, 则地标分布范围会相应地缩小, 而地标密度与图 13 左图相比没有太大变化; 在图 13 右图中, 若删除相同数量、不同位置的地标, 则地标分布范围不会变化, 而密度会发生较大变化。

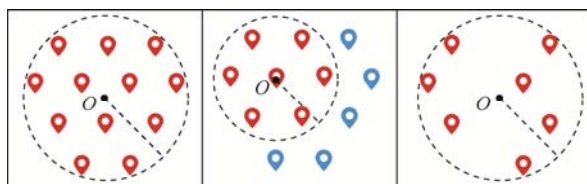


图 13 同子网地标分布范围、数量与密度所对应的不同情况
Figure 13 Different situations corresponding to the distribution area, quantity and density of same subnet landmarks

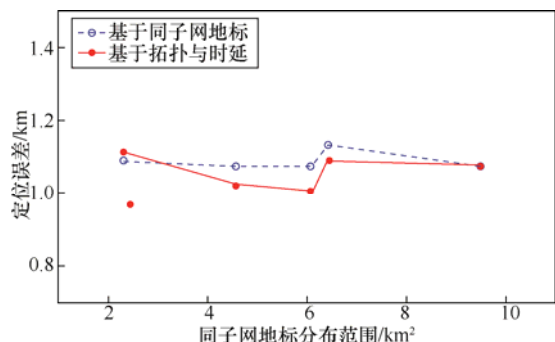
基于同子网地标的 IP 定位通常将目标定位在同子网地标分布范围的中心。与图 13 左图的情况相比, 在图 13 中间图中密度相近条件下的定位误差会有所降低, 而在图 13 右图中密度相差较大条件下的定位误差反而会相对接近。使用此类定位方法进行定位时, 同子网地标密度与定位误差间并不存在强相关关系。

基于拓扑与时延的 IP 定位会将与共同路由器间拓扑跳数或传输时延最小的地标作为该路由器相连目标的定位结果。而在面向细粒度定位时, 时延与地理距离间往往不存在强相关关系, 因此被选作定位结果的地标在该组地标分布范围中的位置是相对随机的。因此, 在使用此类定位方法进行定位时, 同子网地标密度与定位误差间同样不存在强相关关系。

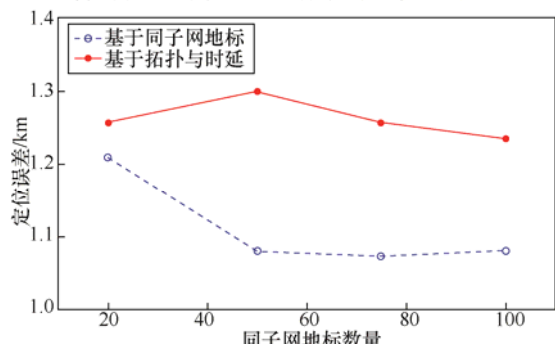
为验证上述分析, 本文在北京市随机选取了 5 组分布较为规则的同子网地标, 分别在数量固定、分布范围可变以及分布范围固定、数量可变的两种条件下对两类定位方法进行测试, 同子网地标密度与定位误差间关系如图 14 所示。

图 14 中的实验结果验证了上述分析。在同子网地标数量固定、分布范围可变条件下, 两类定位方法的误差并未出现与地标密度一致的变化趋势。而在同子网地标分布范围固定、数量分别取

25、50、75、100 时,定位误差与地标密度间呈现微弱负相关关系,但相关性同样不明显。鉴于上述结果,本文选择更直观的同子网地标数量作为判断定位方式的依据。



(a) 同子网地标数量固定、分布范围可变时的定位情况



(b) 同子网地标分布范围固定、数量可变时的定位情况

图 14 同子网地标密度与定位误差间关系

Figure 14 The relationship between same subnet landmark density and geolocation errors

3.4 目标可达性

目标可达性指的是 IP 目标是否对探测报文进行响应。现有的基于网络测量的 IP 定位方法会对目标进行探测,采集相关的信息并基于此定位。然而,在实际互联网中,很多节点出于设备安全或运行成本的考虑,不对探测报文进行响应,这使得基于时延与拓扑的 IP 定位难以确定不可达目标在网络拓扑中的准确位置,进而无法实现准确的目标定位。

图 15 展示了目标不可达情况下的探测路径情况。在探测目标 T_1 时,路径中所有设备均对探测报文响应,因此能够得到完整的路径信息。而探测不对探测报文响应的目标 T_2 时,探测源只能获取到路由节点 R_2 为止的路径。而在探测目标 T_3 时,由于 T_3 的接入路由不可达,探测源得到更少的路径信息。

因此,本文定位框架引入基于局部网络结构的 IP 定位来解决此类情况。在目标 IP 不可达时,此类 IP 定位技术使用局部网络结构取代共同路由对目标进行定位,提高了定位成功率。

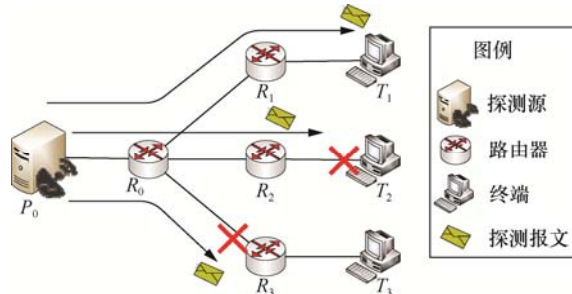


图 15 目标不可达情况下的探测路径

Figure 15 Detection path of target unreachable scenario

本文在我国 11 个城市进行了 IP 终端不可达性统计,结果如图 16 所示。

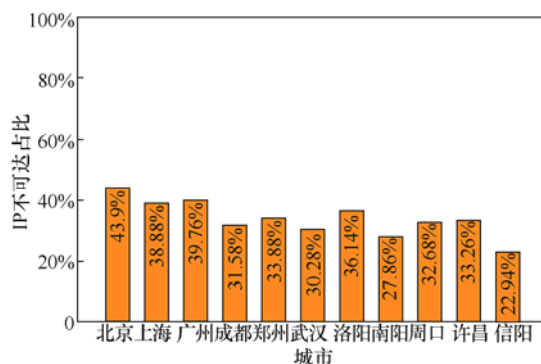


图 16 IP 终端不可达性统计

Figure 16 Statistics on unreachable IP terminals

经统计,对于本文实验城市,IP 不可达占比至少为 22.94%,平均占比为 33.74%,证明了本文的定位框架采用基于局部网络结构的 IP 定位必要性。

4 实验评估

为了验证本文框架的地理定位能力,本文分别测试了城市级定位成功率和街道级定位误差,验证了同子网地标数量与目标可达性对定位结果的影响,并进行了相应的对比实验。

4.1 实验设置

在实验环境搭建过程中,考虑我国地理面积大、网络环境比较复杂,本文选取了城市位置、城市规模、同子网地标数量和目标可达性作为变

量,将本文的定位框架与典型算法 RMP-Geo、DDT-Geo 和 PBG 进行比较分析。表 1 给出了实验详细设置。实验 1 和实验 3 选择我国的省会城市、直辖市和河南省的城市,以测试这些算法在不同规模和分布的城市中的城市级定位能力。实验 2 和实验 4 测试了这些算法在上述城市中的街道级定位能力。而实验 5 和实验 6 验证了同子网地标数量与目标可达性对这些算法的影响。

在定位实验中,本文从现有的公共数据库中大量获取 IP 地标。而后,按照文献[25]中的方法对其中具有街道级位置的 IP 进行评估,并保留位置可靠的 IP 地址,作为方法输入中的地标原始数据。对于实验 5,本文在地标原始数据中抽取了 100 个位于河南省郑州市且包含地标数量大于 200 的/24 子网,将这些子网所包含的地标作为地标集。对于其他实验,本文选择了位于我国 11 个城市的地标,按照同子网地标数量对这些地标对应的/24 子网进行分组,分别在包含 1~65、66~129、130~173、174~255 个地标的子网中各随机选择了 200、20、20、160 个子网,将这些子网所包含的地标作为地标集。上述构建方式使得地标集内同子网地标数量、分布与实际互联网环境较为相似。

本文使用城市级定位成功率和城市内定位误差来判断本文定位框架的性能。城市级定位成功率指的是能够成功返回目标的城市级位置并且结果正确的次数与定位总数之比。城市内定位误差是在对已知位置的 IP 目标进行地理定位时,算法返回的地理位置与目标实际地理位置之间的距离。城市内定位误差涉及的地理尺度较小,平面坐标和球坐标之间的计算误差可以忽略,因此使

用平面坐标系计算两点之间的距离。目前尚未有能够提供街道级位置的基于子网地标分布的定位算法,且给出的位置结果较为模糊,因此没有与此类定位方法进行对比。因为 PBG 算法只能给出城市级位置结果,本文在实验中将城市的位置中心标记为目标定位结果,依此计算定位误差。

本文使用互联网数据分析合作协会(CAIDA, Cooperative Association for Internet Data Analysis)开发的网络测量工具 Scamper^[26]对 IP 目标进行路径探测。此外,在检测目标网络的拓扑信息时,本文综合使用了 5 种协议(ICMP、TCP、UDP、ICMP-Paris 和 UDP-Paris),并通过使用多协议路径检测增加了探测报文成功转发的概率,提高拓扑信息的获取规模。ICMP-Paris 和 UDP-Paris 可以避免生成错误的路径信息^[27]。

4.2 实验结果分析

本节分别对实验设置中提出的几种情况进行对比实验,结果如下。

(1) 城市级定位实验

在对比实验中,考虑到城市的位置与规模可能会对定位结果产生影响,分别对我国 6 个城市和河南省内 6 个市在同一 ISP 环境下进行城市级定位实验,城市级定位成功率对比结果如图 17 所示。

经统计,在我国省会级城市的定位实验中,RMP-Geo、DDT-Geo、PBG 和本文定位框架的平均定位成功率分别为 91.06%、93.93%、88.66% 和 94.04%;在河南省内城市的定位实验中,RMP-Geo、DDT-Geo、PBG 和本文定位框架的平均定位成功率分别为 91.38%、95.51%、92.61% 和 96.16%。实验结果表明,城市位置、规模对定

表 1 实验详细配置
Table 1 Detailed experimental settings

实验	探测报文	定位精度	同子网地标数量	目标可达性	城市	算法
实验 1	ICMP、TCP、UDP、 ICMP-Paris、UDP-Paris	城市级	固定	固定	北京、上海、广州、 成都、郑州、武汉	RMP-Geo ^[14] 、 DDT-Geo ^[15] 、 PBG ^[23] 、本文框架
实验 2		街道级				
实验 3		城市级				
实验 4		街道级	变化	变化	郑州	
实验 5						
实验 6			固定	变化		

位结果影响不大。与此同时,本文定位框架与DDT-Geo相比,增加了通过同子网对目标进行位置估计的步骤,因而进一步提高了城市级定位成功率。

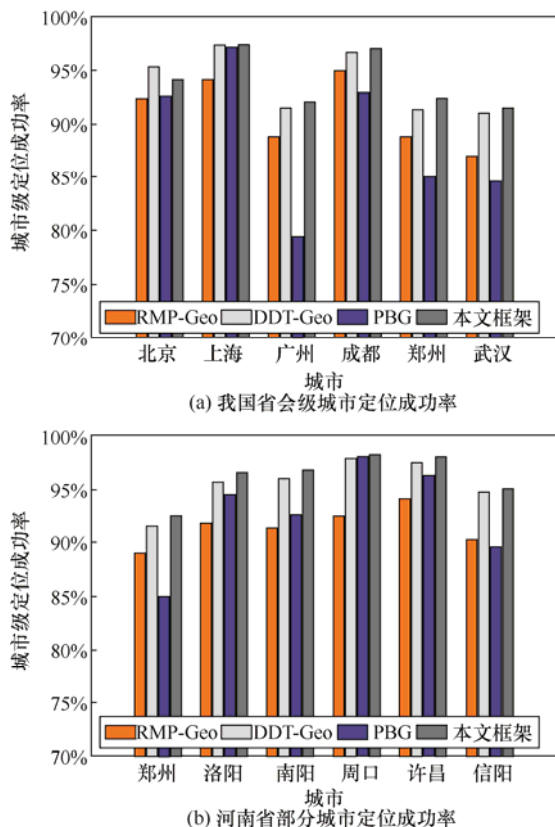
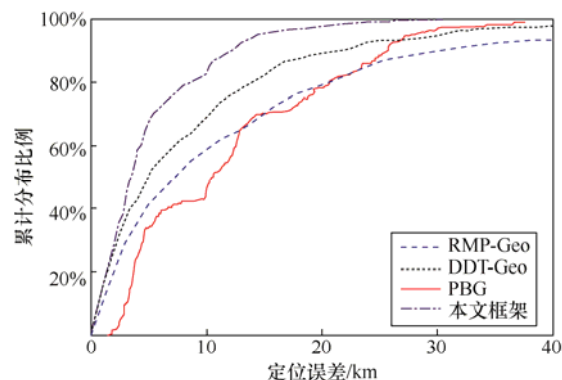


图 17 城市级定位成功率对比结果
Figure 17 Comparison results of city-level geolocation success rate

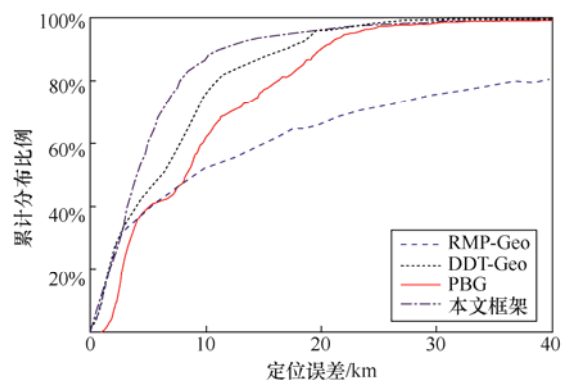
(2) 街道级定位实验

这里对上文实验涉及的城市进行了街道级定位实验。图 18 展示了街道级定位误差对比,即小于给定定位误差占全部定位结果的比例。在图中,曲线越接近左上角,表示误差越小,定位结果越好。

经统计,在我国省会城市及直辖市的定位实验中,RMP-Geo、DDT-Geo、PBG和本文定位框架的定位误差分别为 7.19 km、4.86 km、12.60 km 和 3.45 km;在河南省内城市的定位实验中,RMP-Geo、DDT-Geo、PBG和本文定位框架的中值误差分别为 9.14 km、6.14 km、8.40 km 和 4.13 km。实验结果表明,本文定位框架与RMP-Geo、DDT-Geo和PBG相比,具有更低的定位误差。



(a) 我国省会城市定位误差累计分布



(b) 河南省部分城市定位误差累计分布

图 18 街道级定位误差对比

Figure 18 Comparison of street-level geolocation error

同时,发现一个有趣的现象:与省会城市及直辖市定位结果相比,在河南省内城市进行定位时,RMP-Geo、DDT-Geo和本文框架的定位误差有所上升,而PBG的定位误差则有所下降。在分析结果后发现,我国的大型城市(省会城市及直辖市)中同一子网的用户分布往往更为集中,而中小型城市(河南省内城市)中同一子网的用户分布相对分散,分布范围更广。RMP-Geo、DDT-Geo和本文框架全部或部分采用共同路由地标进行定位,而同一子网的终端有较大概率通过同一组接入路由器连接互联网,因此会受到上述变化的影响。而本文在通过PBG进行定位时,选择用城市中心位置作为定位结果,显而易见地,我国的大型城市通常比中小型城市面积大,因此,PBG在对河南省内城市目标进行定位时误差反而有所降低。

(3) 同子网地标数量可变下的定位实验

考虑到同子网地标可能会对定位结果产生影响,本文在同子网地标数量可变的条件下进行

定位实验，同子网地标数量可变下定位误差对比如图 19 所示。

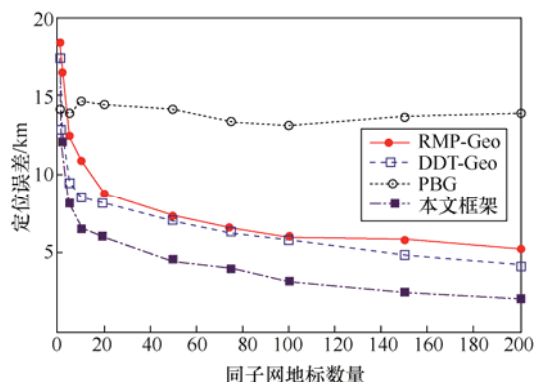


图 19 同子网地标数量可变下定位误差对比
Figure 19 Comparison of geolocation errors with variable number of same subnet landmark

实验结果表明，对于 RMP-Geo、DDT-Geo 和本文框架而言，同子网地标数量会对定位结果产生影响，当地标数量上升时，定位误差会下降。与此同时，由于本文定位框架能够适应各种子网地标环境，在总体条件下具有更好的定位效果。如图 19 所示，当同子网地标数量大于 10 时，本文定位框架的误差明显优于 RMP-Geo、DDT-Geo。而同子网地标数量的变化对 PBG 的定位性能影响则不明显。同子网地标数量可变下的定位误差如表 2 所示。

表 2 同子网地标数量可变下的定位误差
Table 2 Geolocation errors with variable number of same subnet landmark

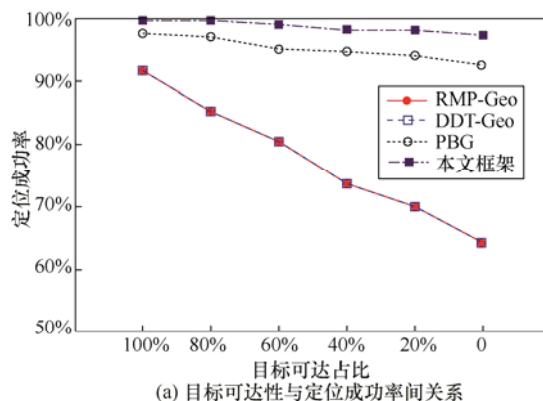
地标数量	定位误差/km			
	RMP-Geo	DDT-Geo	PBG	本文框架
1	18.39	17.39	14.06	16.29
2	16.56	12.77	13.82	12.04
5	12.43	9.41	13.89	8.16
10	10.86	8.47	14.63	6.32
20	8.68	8.16	14.42	5.99
50	7.38	6.95	14.10	4.52
75	6.53	6.29	13.20	3.99
100	5.98	5.82	13.01	3.24
150	5.88	4.79	13.64	2.46
200	5.22	4.16	13.79	2.05

(4) 目标可达性可变下的定位实验

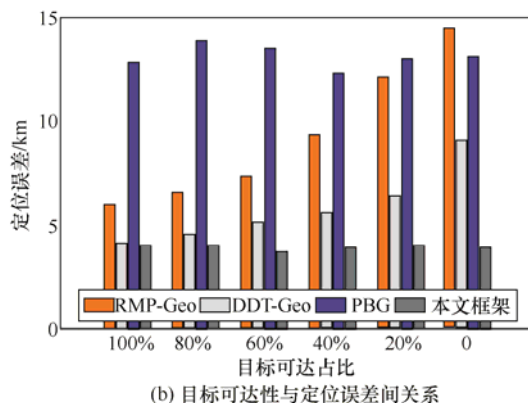
考虑到目标可达性可能会对定位结果产生影

响，本文在地标库中抽取了位于河南省郑州市的 1 000 个可达地标和 1 000 个不可达地标作为待选定位目标。按照目标可达占比分别为 0、20%、40%、60%、80%、100% 在两个地标集中随机采样，随后对其进行定位实验，目标可达性可变下定位结果对比如图 20 所示。

图 20(a)表示在不同目标可达占比下各类方法的定位成功率，其中端点越高，定位方法返回结果且结果正确的次数与总定位次数之比越高，代表定位性能越好。图 20(b)表示在不同可达目标占比下，各类方法的定位误差。柱体越低，定位误差越小，代表定位性能越好。实验结果表明，目标可达占比会影响定位结果。



(a) 目标可达性与定位成功率间关系



(b) 目标可达性与定位误差间关系

图 20 目标可达性可变下定位结果对比
Figure 20 Comparison of geolocation results with variable target accessibility

从图 20 可以看出，随着目标可达占比下降，RMP-Geo 和 DDT-Geo 的定位误差逐渐增大，定位成功率不断降低，而 PBG 和本文框架的定位性能变化不大。容易得知，在定位不可达目标时，RMP-Geo 和 DDT-Geo 匹配到的共同路由器距离目标跳数更多，对应的地标分布范围更广，导致

误差上升。而本文框架在定位不可达目标时,除了使用局部网络结构进行定位外,在同子网地标数量合适时,会直接通过同子网地标分布进行定位,因而与对比方法相比拥有更好、更稳定的定位效果。同时,DDT-Geo 和 RMP-Geo 均通过目标探测路径与城域网拓扑对比完成目标的城市级定位,且城域网拓扑提取部分基本没有差异,因此具有相同的城市级定位成功率。目标可达性可变下的定位误差和定位成功率分别如表 3 和表 4 所示。

表 3 目标可达性可变下的定位误差
Table 3 Geolocation errors with variable target accessibility

目标可达占比	定位误差/km			
	RMP-Geo	DDT-Geo	PBG	本文框架
100%	5.96	4.19	12.82	4.01
80%	6.56	4.60	13.89	4.07
60%	7.34	5.10	13.52	3.66
40%	9.31	5.68	12.33	3.88
20%	12.14	6.41	13.02	4.07
0	14.53	9.09	13.10	3.94

表 4 目标可达性可变下的定位成功率
Table 4 Geolocation success rate with variable target accessibility

目标可达占比	定位成功率			
	RMP-Geo	DDT-Geo	PBG	本文框架
100%	91.8%	91.8%	97.6%	99.8%
80%	85.4%	85.4%	97.1%	99.8%
60%	80.8%	80.8%	95.2%	99.1%
40%	73.4%	73.4%	94.7%	98.4%
20%	69.8%	69.8%	94.1%	98.2%
0	64.3%	64.3%	92.5%	97.4%

5 结束语

在不同的网络场景中,各类 IP 定位方法难以维持理想的定位效果。为增加 IP 定位对不同网络环境的适应性,本文提出了一种目标网络场景自适应的 IP 定位框架。该定位框架能够在定位前对目标所处网络场景进行判定,并采用合适的定位方法实施定位。本文在我国 11 个城市进行了实验评估,得到了在各类网络环境下不同 IP 定位方法的多层次表现,验证了本文定位框架的有效性,

并为 IP 定位研究提供新的思路。

受工作量限制,本文仅针对两种较为特殊的网络场景进行了分析,采用了 3 类定位方法,这限制了本文框架的适应性。继续细化网络场景的划分并进一步提高定位框架对网络环境的适应能力,是下一步工作的方向。

参考文献:

- [1] NIU B, LI Q, ZHU X, et al. Enhancing privacy through caching in location-based services[C]//2015 IEEE Conference on Computer Communications (INFOCOM). 2015: 1017-1025.
- [2] ZHENG X, CAI Z, LI J, et al. Location-privacy-aware review publication mechanism for local business service systems[C]//2017 IEEE Conference on Computer Communications (INFOCOM). 2017: 1-9.
- [3] 赵帆, 罗向阳, 刘粉林. 网络空间测绘技术研究[J]. 网络与信息安全学报, 2016, 2(9): 1-11.
- [4] BURCH H, CHESWICK B. Mapping the internet[J]. Computer, 1999, 32(4): 97-98.
- [5] CHEN J, LIU F, ZHAO F, et al. A SC-vivaldi network coordinate system based method for IP geolocation[J]. Journal of Internet Technology, 2016, 17(1): 119-127.
- [6] 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.
- [7] GUO L, CAO Y N, SU M J, et al. Cyberspace resources surveying and mapping: the concepts and technologies[J]. Journal of Cyber Security, 2018, 3(4): 1-14.
- [8] LIU C, LUO X, YUAN F, et al. RNBG: a ranking nodes based IP geolocation method[C]//2020 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). 2020: 80-84.
- [9] ZHANG F, LIU F, XU R, et al. Street-level IP geolocation algorithm based on landmarks clustering[J]. Computers Materials & Continua, 2021, 66(3): 3345-3361.
- [10] DING S, LUO X, WANG J, et al. GNN-Geo: a graph neural network-based fine-grained IP geolocation framework[C]//IEEE Transactions on Network Science and Engineering, 2023: 1-18.
- [11] 方滨兴. 从层次角度看网络空间安全技术的覆盖领域[J]. 网络与信息安全学报, 2015, 1(1): 2-7.
- [12] FANG B X. A hierarchy model on the research fields of cyberspace security technology[J]. Chinese Journal of Network and Information Security, 2015, 1(1): 2-7.
- [13] ZU S, LUO X, ZHANG F. IP-geolocator: a more reliable IP geolocation algorithm based on router error training[J]. Frontiers of Computer Science, 2022, 16(1): 1-11.
- [14] ZHAO F, LUO X, GAN Y, et al. IP geolocation based on identification routers and local delay distribution similarity[J]. Concurrency and Computation: Practice and Experience, 2019, 31(22): 4722.
- [15] WANG Y, BURGNER D, FLORES M, et al. Towards street-level

- client-independent IP geolocation[C]//Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation. 2011: 365-379.
- [14] ZHAO F, XU R, LI R, et al. Street-level geolocation based on router multilevel partitioning[J]. IEEE Access, 2019, 7: 59237-59248.
- [15] ZU S, LUO X, DU S, et al. A delay deviation tolerance IP geolocation method with error estimation[J]. Scientific Reports, 2022, 12(1): 1-15.
- [16] MUKNE N, PAFFENROTH R. Probabilistic inference of Internet node geolocation with anomaly detection[C]//2017 IEEE International Symposium on Technologies for Homeland Security (HST). 2017: 1-6.
- [17] WANG T, XU K, SONG J, et al. An optimization method for the geolocation databases of Internet hosts based on machine learning[J]. Mathematical Problems in Engineering, 2015, 2015: 1-17.
- [18] LI H, ZHANG P, WANG Z, et al. Changing IP geolocation from arbitrary database query towards multi-databases fusion[C]//2017 IEEE Symposium on Computers and Communications(ISCC). 2017: 1150-1157.
- [19] POESE I, UHLIG S, KAAFAR M A, et al. IP geolocation databases: Unreliable[J]. ACM SIGCOMM Computer Communication Review, 2011, 41(2): 53-56.
- [20] SHAVITT Y, ZILBERMAN N. A geolocation databases study[J]. IEEE Journal on Selected Areas in Communications, 2011, 29(10): 2044-2056.
- [21] ZHAO F, SONG Y, LIU F, et al. City-level geolocation based on routing feature[C]//2015 IEEE 29th International Conference on Advanced Information Networking and Applications. 2015: 414-419.
- [22] KATZ-BASSETT E, JOHN J P, KRISHNAMURTHY A, et al. Towards IP geolocation using delay and topology measurements[C]//Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement. 2006: 71-84.
- [23] ZU S, LUO X, LIU S, et al. City-level IP geolocation algorithm based on PoP network topology[J]. IEEE Access, 2018, 6: 64867-64875.
- [24] MILO R, SHEN-ORR S, ITZKOVITZ S, et al. Network motifs: simple building blocks of complex networks[J]. Science, 2002, 298(5594): 824-827.
- [25] LI R, SUN Y, HU J, et al. Street-level landmark evaluation based on nearest routers[J]. Security and Communication Networks, 2018, 2018: 1-12.
- [26] LUCKIE M. Scamper: a scalable and extensible packet prober for active measurement of the internet[C]//ACM SIGCOMM Conference on Internet Measurement. 2010: 239-245.
- [27] AUGUSTIN B, CUVELLIER X, ORGOGOZO B, et al. Avoiding traceroute anomalies with paris traceroute[C]//Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement 2006. 2006.

[作者简介]



祖铄迪（1994—），男，河北保定人，网络空间态势感知河南省重点实验室博士生，主要研究方向为网络空间资源测绘和IP定位。



丁世昌（1990—），男，山东日照人，博士，网络空间态势感知河南省重点实验室讲师，主要研究方向为网络空间资源测绘和IP定位。



袁福祥（1991—），男，山东济宁人，博士，网络空间态势感知河南省重点实验室讲师，主要研究方向为网络空间资源测绘和IP定位。



罗向阳（1978—），男，湖北荆门人，博士，网络空间态势感知河南省重点实验室教授、博士生导师，主要研究方向为网络空间安全。