一种多接口路由器地理定位方法

朱金玉, 张　宇, 曾良伟, 余卓勋, 张宏莉

哈尔滨工业大学计算机科学与技术学院 计算机网络与信息安全技术研究中心 哈尔滨 中国 150001

摘要　在网络空间测绘中, 如何将虚拟拓扑中设备节点定位到现实世界中地理位置是一个研究难点。以往 IP 地理定位工作多以 IP 地址为单位, 缺少以路由器为单位的定位研究。本文利用同一台路由器上不同接口 IP 地址位置相同, 相连的路由器间地理位置相近这两个事实, 提出一种多接口路由器地理定位方法, 包括接口选举方法、邻居选举方法、综合法, 来定位路由器地理位置。实验结果表明, 与相关数据集相比, 在可定位路由器的覆盖率和定位准确率上都有明显提升。在覆盖率上, 国家级达到 99.84%, 城市级达到 96.00%, 比相关数据集分别高出 0.93%和 36.48%; 在 IXP 数据验证准确率上, 国家级达到 82.51%, 城市级达到 59.45%, 比相关数据集分别高出 9.91%和 27.20%。

关键词　多接口路由器; IP 地理定位; 网络空间测绘

中图法分类号　TP393.4　　**DOI 号**　10.19363/J.cnki.cn10-1380/tn.2018.07.02

Geolocation For Multi-Interface Routers

ZHU Jinyu, ZHANG Yu, ZENG Liangwei , YU Zhuoxun, ZHANG Hongli

Research Center of Computer Network and Information Security Technology, Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001 , China

Abstract　In the Cyberspace mapping, how to map devices in the virtual topology to geolocations in the real world is a difficult problem. In the past, IP geolocation was mostly based on IP address, rather than routers. In this paper, we utilize two facts that different interfaces on the same router are at the same geolocation, and distances between a router and its neighbors are close. We propose a set of geolocation methods for multi-interface routers, including interface election, neighbor election, IE+NE. The experimental results show that compared with the public dataset, the coverage and accuracy of router geolocation are significantly improved. In terms of coverage rate, the country level reached 99.84%, the city level reached 96.00%, higher than the relevant data set 0.93% and 36.48% respectively; in terms of accuracy rate, the country level reached 82.51%, the city level reached 59.45%, higher than the relevant data set 9.91% and 27.20% respectively.

Key words　Multi-interface router; IP geolocation; cyberspace mapping

1　引言

网络空间测绘研究内容之一是实体资源定位, 即确定网络设备实体在地理空间中的位置。路由器作为构建互联网的基石, 对其地理定位技术就成了能否准确绘制网络空间的关键。然而, 网络空间实体多以 IP 地址为标识, 而 IP 地址本身具有地理位置无关性, 同时, 一台路由器拥有多个 IP 地址也增加了定位的不确定性。因此如何对路由器实施定位既是网络空间测绘核心问题之一, 更是一个难点问题。

以往工作通过提取和解码路由器主机名中包含的地理信息字符串定位路由器位置, 或通过建立路径-时延模型来定位路由器地理位置。这些方法通常需要额外测量工作, 或人工辅助解析, 难以持续更新。不仅如此, 测量通常不可重现, 人工参与的解析结果会因人而异, 导致定位过程难以重现, 定位结果难以评估。

本文提出一个种基于公开路由器级拓扑测量数据与商业 IP 地理信息的路由器定位方法--RLoc。该方法“站在巨人的肩膀上”, 充分利用现有网络拓扑测量与 IP 地理定位的成果来实施定位。与以往工作相比, RLoc 无需实施新的大规模网络测量, 或构建及校对时延-距离模型; 无需在人工辅助下解析域名、Whois、网页等 IP 地址相关信息。

通讯作者: 张宇, 博士, 副教授, Email:yuzhang@hit.edu.cn。

本课题得到国家重点研发计划(No.2016YFB0801303); 东莞市引进创新科研团队计划(No.201636000100038)资助。

收稿日期: 2018-03-30; 修改日期: 2018-05-30; 定稿日期: 2018-06-19

RLoc 将拓扑测量与 IP 定位工作获得的数据集作为输入, 但不依赖其获取方法与质量。在具备路由器级拓扑测量数据和 IP 地理信息前提下, 对多接口路由器定位仍充满挑战。如图 1 所示, 由于相邻路由器间共享地址空间, 使路由器各接口 IP 地址定位位置不同。该路由器共有 25 个接口 IP 地址, 位置分别在西班牙、美国、英国、德国。如图 2 所示, 虽然相邻路由器地理位置相近, 但有些路由器逻辑上相连但位置分布较远。该路由器共有 1,269 台邻居路由器, 他们的位置在西班牙、美国、德国、中国、葡萄牙, 还有部分路由器位置无法确定。

RLoc 充分利用同一台路由器拥有多个接口 IP 地址但地理位置相同这一依据, 提出接口选举方法定位; 同时, 利用相连的路由器与路由器间地理位置相近这一依据, 提出邻居选举方法定位; 结合前两种定位依据, 综合路由器接口 IP 地址位置与相邻路由器位置信息定位多接口路由器位置。

本文主要贡献总结如下:

1、提出一种基于公开路由器级拓扑测量数据与商业 IP 地理信息的多接口路由器地理定位方法——RLoc, 利用两点事实: 同一台路由器的不同接口 IP 地址在相同位置; 相连路由器间地理位置相近。

2、将 RLoc 应用于 132,175 台多接口路由器的地理定位。实验结果表明, 在覆盖率上, 国家级达到 99.84%, 城市级达到 96.00%, 比相关数据集分别高出 0.93%和 36.48%; 在 IXP 数据验证准确率上, 国家级达到 82.51%, 城市级达到 59.45%, 比相关数据集分别高出 9.91%和 27.20%。

3、RLoc 易于实施与重复, 结果易于更新和评估。为此, 公开了对 132,175 台多接口路由器定位结果。

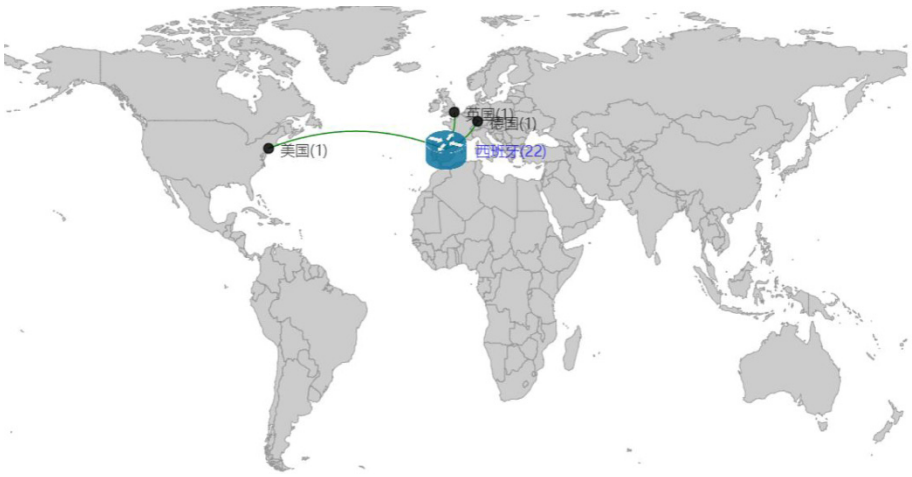

图 1　路由器的接口 IP 地址定位多个位置实例图

Figure 1　Instance of router's interfaces geolocated multi-location

(注: 图中多接口路由器共有 25 个接口 IP 地址, 其中 22 个在西班牙, 1 个在美国, 1 个在德国, 1 个在英国。)

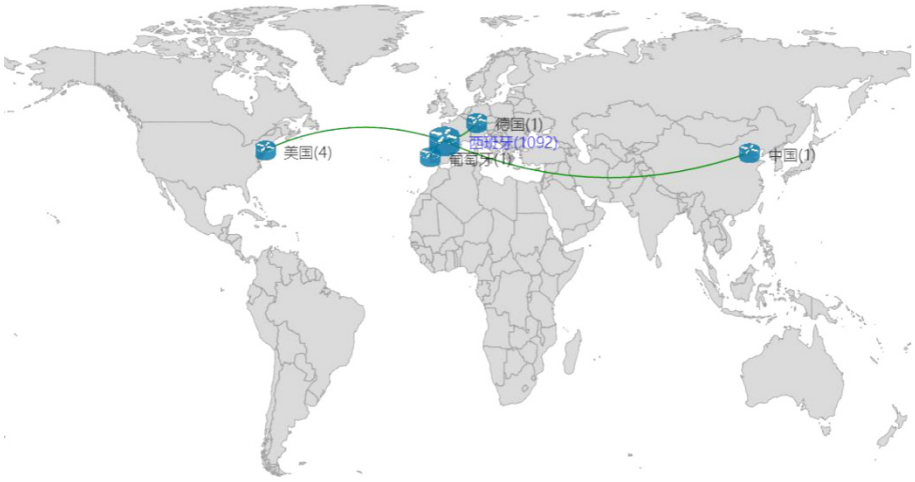

图 2　路由器的邻居路由器定位多个位置实例图

Figure 2　Instance of router's neighbors geolocated multi-location

(注: 图中路由器共有 1,269 个邻居路由器, 其中 1092 个在西班牙, 4 个在美国, 1 个在德国, 1 个在葡萄牙, 1 个在中国, 170 个无法定位。)

(百度网盘下载链接: https://pan.baidu.com/s/1mr6a8vDFMHKaM6-cWwEqwg)

本文的以下章节内容按照以下组织: 第二章介绍相关工作; 第三章提出多接口路由器地理定位方法 RLoc;第四章给出实验及结果分析; 第五章是方法的局限性; 第六章总结了全文并展望未来工作。

2 相关工作

互联网拓扑测量和分析研究方向已有二十多年, 其中包含 AS、POP、路由器、接口级拓扑。IP 地理定位技术伴随着云计算和社交网络等新型网络应用的不断发展, 逐渐受到了越来越多的关注。而本文提出的路由器地理定位方法以路由器级拓扑测量数据和 IP 地理定位数据为基础, 实现对网络空间测绘中实体资源路由器的定位, 从而实现大规模网络逻辑拓扑与地理位置的映射。

路由器级拓扑测量: 路由器级拓扑通常是属于同一台路由器的接口组合, 一个节点代表着一个主机或一个多接口的路由器。如果节点间相互连接, 则节点间一定有接口位于同一个 IP 广播域。在 Traceroute 测量数据中, 存在不可忽略的路由陷阱, 如: 负载均衡和路由改变情况等。路由陷阱需要进行别名解析, 把属于同一台路由器的 IP 地址映射到该路由器上。

Motamedi 等[1]对网络拓扑发现各层级拓扑数据收集的技术和工具深度分析。CAIDA[2]在 2000 年采用相同源地址方法实现了 Iffinder 别名解析工具。Spring 等[3]在 2002 年采用 IPID 计数器实现了 Ally 别名解析工具。Bender 等[4]提出了 RadarGun 工具, 降低了复杂度。Keys 等[5]在 2016 年提出基于 IPID 的 MIDAR 别名解析技术, 比 Ally 和 RadarGun 准确率更高。本文综合考量各类拓扑数据质量, 选择使用 CAIDA 的 Macroscopic Internet Topology Data Kit (ITDK) [6]项目中, MIDAR 和 IFFINDER 共同使用作为别名解析工具所生成的路由器级拓扑数据集, 它较相同 IP 级拓扑数据作为输入的 Kapar 等方法具有高置信度和低误报率, 较其他平台拓扑测量覆盖率更高。

IP 地理定位技术: IP 地址定位是确定一个网络目标节点在某个粒度层次的地理位置, 由于每一个直接与互联网相连的主机都被一个唯一的 IP 地址所表标识, 通常利用 IP 地址来寻找其地理坐标映射。根据近年的研究方法, 定位技术可以分为三类, 分别为基于主动测量、数据挖掘分析和数据库推测的定位技术。

Gueye 等[7]采用三角定位方法确定被测 IP 地址位置。Katz-Bassett 等 [8]采用网络路径信息作为目的 IP 和中间节点的约束条件从而确定待定位 IP 的位置区间。Wong 等[9]利用时延-距离关系测量城市位置, 使用贝塞尔曲线表示 IP 地址可能出现的区域, 通过不断迭代确定 IP 所在区域。基于主动测量的方法准确度受时延-距离模型影响较大, 且需要大量测量点和地标点, 定位一个目标 IP 地址所需时间也较长。Liu 等[10]利用用户自愿提供的位置信息来定位, 误差中位数为 799m。这类数据挖掘方法虽然可得到更加精确的定位, 但覆盖率和普适性不高。基于数据库推测的定位技术通过推测 DNS、Whois、BGP 等数据中直接或间接提供的位置信息定位 IP 地址。Moore 等[11]通过直接查询 Whois 数据库来推测主机位置信息。Padmanabhan 等[12]通过挖掘主机名字中可能包含的不同粒度的地理位置信息推测主机的位置。

为能满足更广泛的应用, 如定向广告、诈欺监测、网站流量分析、地理目标定位、数字版权管理等, 国内外 IP 商业地理定位数据库, 如: MaxMind、IP2Location、Netacuity、IPMarker、IPIP.NET 等, 利用测量、数据挖掘、数据库推测等各方法对 IP 地理定位, 精确度可为国家、城市、甚至于邮编级。宋健等[13]提出了一种基于 IP 地址库之间差异对比来评估可信程度的方法, 并发现 IP 地址库之间差异所存在的规律。王婷等[14]提出 IPGEL 方法有效提高已有 IP 定位数据库的可用性。Gharaibeh 等[15]评估了多种地理定位数据库在路由器定位的可靠性。本文综合各类商业 IP 地理定位数据库质量, 选择使用 IP2location 数据库为路由器各接口 IP 地址地理定位, 它较 MaxMind 等数据库的国家城市粒度更精细。

路由器定位方法: Huffaker 等[16]提取和解码路由器主机名中包含的地理信息字符串从而定位路由器位置。该方法需要不断收集具有地理字符串的数据, 提出规则将其与实际物理位置联系起来, 需要大量的人工分析编译, 但人工解析结果因人而异且规则只能识别带有主机名信息的 IP 地址的位置, 无法适用于大规模的路由器地理定位。Laki 等[17, 18]提出了一个路径-时延模型来定位路由器地理位置。该方法需要进行大规模的拓扑测量并构建时延-距离模型, 但测量通常不可重现, 模型校对需要大量的地标点和测量工作, 而且时延受网络拥塞等外界条件影响较大。

CAIDA 利用 ITDK 的路由器级拓扑数据, IXP、DNS、MaxMind 数据库信息对路由器进行地理定位。当路由器至少一个接口 IP 地址属于 IXP 地址空间且该地址空间只有一个地理位置, 则定位该路由器; 余下

路由器利用DDec主机名映射, 当路由器至少一个接口IP 地址的主机名解析了地理位置且所有能够解析的接口结果一致, 则定位该路由器; 除此之外的路由器, 根据 MaxMind 数据库定位接口 IP 地址, 当所有接口 IP 地址定位在同一位置, 则定位该路由器位置。同时, CAIDA 公开了路由器位置数据。但该方法过多的依赖于公开数据集的数据, 受到 IXP、DNS、MaxMind 数据库等信息的质量影响较大, 且每个 IXP 有多个互联设施可能分布在不同的地方, 这导致能够利用 IXP 数据定位的路由器数量不多; 使用 MaxMind 数据库定位精度在城市级粒度上能够定位的路由器数量不多。

本文提出了一种新的多接口路由器地理定位方法—RLoc, 基于 IP2Location 商业数据库和路由器级拓扑数据, 减少了大量的测量时间, 与人工收集解析过程, 提出启发式方法, 不仅利用路由器接口 IP 地址映射的位置定位路由器, 而且利用路由器位置与接口 IP 地址位置间关系和邻居路由器间位置关系定位。RLoc 与 CAIDA 定位相同的路由器, 与 CAIDA 公开的数据集相比拥有更高的覆盖率与准确率。

3 多接口路由器地理定位方法

RLoc 利用两点事实: 1.同一台路由器的不同接口 IP 地址在相同位置; 2.相连路由器间地理位置相近。提出了三种启发式方法, 如图 3 所示, 接口选举(Interface Election, IE)、邻居选举(Neighbor Election, NE)和综合法(IE+NE)。

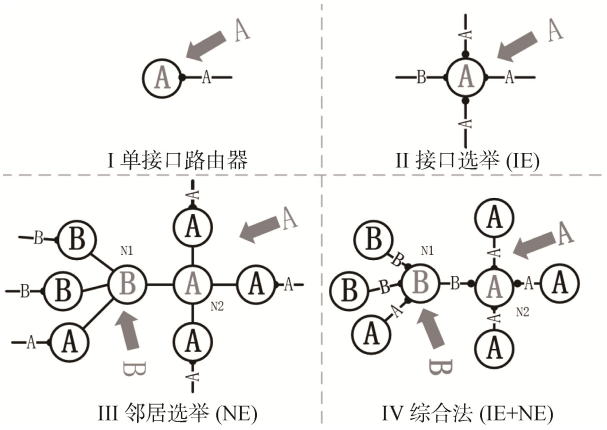

图 3 三种启发式方法定位多接口路由器示意图

Figure 3 Three heuristic methods to geolocate multi-interface routers

(注: 图中四部分分别表示单接口路由器、接口选举方法、邻居选择方法、综合法示意图。圆圈代表路由器,圆圈内的字母代表路由器的位置, 黑色字母表示已知位置, 灰色字母表示经过方法定位后的位置。圆圈的连接线表示路由器的接口 IP 地址, 连线上的字母表示接口 IP 地址的位置。3、4 部分表示对路由器 N1、N2 定位示意图)

3.1 接口选举

RLoc 充分利用同一台路由器上不同接口 IP 地址在相同位置这一事实, 提出 IE 方法。该方法以路由器各接口 IP 地址和商业 IP 地理数据库为输入, 为每台路由器的接口 IP 地址地理定位。路由器的接口 IP 地址由于路由器间交换地址空间导致地理位置不同, 因此为每台路由器建立位置频数矩阵:

$$\boldsymbol{I}=\begin{bmatrix} L_1, F_1, CI_1 \\ L_2, F_2, CI_2 \\ \cdot \\ \cdot \\ \cdot \\ L_n, F_n, CI_n \end{bmatrix}$$

I 中 L_i 表示路由器接口 IP 地址的第 i 个位置, F_i 表示第 i 个位置出现的频数, CI_i 表示第 i 个位置作为路由器位置的置信度。其中 CI_i 表示为:

$$CI_i=\frac{F_i}{F_1+F_2+\cdots+F_n}$$

在位置频数矩阵中选择置信度最大的结果作为路由器的位置。在图 3 IE 方法示意图中, 一台路由器有四个接口 IP 地址, 地理位置分别为 A、A、A、B, 则建立路由器位置频数矩阵为:

$$\boldsymbol{I}=\begin{bmatrix} \mathrm{A}, 3, 0.75 \\ \mathrm{B}, 1, 0.25 \end{bmatrix}$$

路由器可能的位置为 A 和 B, 其中 A 的位置置信度更大, 则路由器的位置为 A, 置信度 0.75。

以路由器定位的实例图 1 为例, 一台路由器有 25 个接口 IP 地址, 分别定位地理位置, 并建立位置频数矩阵:

$$\boldsymbol{I}=\begin{bmatrix} \text{西班牙}, 22, 0.88 \\ \text{美国}, 1, 0.04 \\ \text{德国}, 1, 0.04 \\ \text{英国}, 1, 0.04 \end{bmatrix}$$

路由器可能的位置为西班牙、美国、德国、英国, 其中西班牙的置信度更大, 则该路由器位置为西班牙, 置信度为 0.88。

IE 方法会出现位置歧义, 即路由器有两个或几个位置拥有相同的置信度, 导致定位失败的情况。

3.2 邻居选择

RLoc 充分利用相连路由器间地理位置相近这一事实, 提出 NE 方法。该方法以路由器间连接关系、单接口路由器和商业 IP 地理数据库为输入, 为每台单接口路由器地理定位, 通过邻居路由器位置的不断迭代从而定位路由器位置。单接口路由器, 如图 3 中 1 所示, 只有一个接口 IP 地址, 其位置为

接口 IP 地址的位置。根据路由器间关系, 确定每台路由器的邻居路由器, 利用已知的单接口路由器位置, 建立每台路由器的邻居路由器位置频数矩阵:

$$\boldsymbol{N}=\begin{bmatrix} r_1, I_1, CN_1 \\ r_2, I_2, CN_2 \\ \cdot \\ \cdot \\ \cdot \\ r_n, I_n, CN_n \end{bmatrix}$$

N 中 r_i 表示邻居路由器的第 i 个位置, I_i 表示第 i 个位置出现的频数, CN_i 表示第 i 个位置作为路由器位置的的置信度。其中 CN_i 表示为:

$$CN_i=\frac{I_i}{I_1+I_2+\cdots+I_n}$$

单接口路由器的位置填入了初始邻居路由器位置频数矩阵中, 通过不断迭代增加多接口路由器的邻居路由器的位置频数, 直至频数矩阵中不再增加新数据为止。在实际数据中, 路由器可能有上千个邻居路由器(通过物理链路层等相互连接等), 为能够得到完整的频数矩阵会造成无限循环迭代等问题, 使数据规模庞大, 无法得到数据。因此, 本文将 NE 方法迭代次数设置最高阈值为 10。迭代 10 次后, 频数矩阵中置信度最大的位置作为路由器位置。在图 3 NE 方法示意图中, 路由器 N1 与三台单接口路由器位置分别为 B、B、A 和一台多接口路由器 N2 相连; 路由器 N2 与三台单接口路由器 A、A、A 和一台多接口路由器 N1 相连。为 N1、N2 路由器建立初始位置频数矩阵为:

$$\boldsymbol{N1}_1=\begin{bmatrix} \mathrm{B},2,0.5 \\ \mathrm{A},1,0.25 \\ ?,1,0.25 \end{bmatrix}$$

$$\boldsymbol{N2}_1=\begin{bmatrix} \mathrm{A},3,0.75 \\ ?,1,0.25 \end{bmatrix}$$

第一次迭代的位置频数矩阵为:

$$\boldsymbol{N1}_2=\begin{bmatrix} \mathrm{B},2,0.5 \\ \mathrm{A},1,0.25 \\ \mathrm{N2}*0.25 \end{bmatrix}=\begin{bmatrix} \mathrm{B},2,0.5 \\ \mathrm{A},1.75,0.4375 \\ ?,0.25,0.0625 \end{bmatrix}$$

$$\boldsymbol{N2}_2=\begin{bmatrix} \mathrm{A},3,0.75 \\ \mathrm{N1}*0.25 \end{bmatrix}=\begin{bmatrix} \mathrm{A},3.25,0.8125 \\ \mathrm{B},0.5,0.125 \\ ?,0.25,0.0625 \end{bmatrix}$$

不断迭代, 至频数矩阵中不再增加新的位置为止或达到迭代阈值 10 次为止。假设以第一次迭代后结果为最终结果, 则 $N1$ 位置为 B, 置信度为 0.5; $N2$ 位置为 A, 置信度为 0.8125。

以为路由器定位的实例图 2 为例, 一台路由器有 1,269 台邻居路由器, 位置为西班牙、美国、德国、葡萄牙、中国、有 170 台路由器未定位位置, 建立位置频数矩阵:

$$\boldsymbol{N}=\begin{bmatrix} 西班牙,1092,0.8605 \\ 美国,4,0.0032 \\ 德国,1,0.0008 \\ 葡萄牙,1,0.0008 \\ 中国,1,0.0008 \\ ?,170,0.1339 \end{bmatrix}$$

其中有部分邻居路由器没有位置信息, 但置信度不高, 则路由器位置为西班牙, 置信度为 0.8605。

NE 方法会出现两种定位失败的情况: 位置歧义; 邻居缺失, 即邻居路由器无法定位导致的频数矩阵缺失。

3.3 综合法

RLoc 充分利用以上两种方法的事实依据, 提出 IE+NE 方法。该方法以 IE 方法和 NE 方法获得的位置频数矩阵为输入, 结合两个矩阵, 为每台路由器地理定位。两个方法的位置频数矩阵结合:

$$R=I\oplus N$$

R 中 I 与 N 表示各方法获得的位置频数矩阵, $\oplus$ 表示:

当 I 中元素 L_i 和 N 中元素 r_i 相同时, 两个元素对应的频数、置信度分别均为$\frac{F_i+I_i}{2}$、$\frac{CI_i+CN_i}{2}$;

当 I 中元素 L_i 和 N 中元素 r_i 不相同时, L_i 元素对应的频数、置信度分别为$\frac{F_i}{2}$, $\frac{CI_i}{2}$; r_i 元素对应的频数、置信度分别为$\frac{I_i}{2}$, $\frac{CN_i}{2}$。

在位置频数矩阵中置信度最大的结果作为路由器位置。在图 3 IE+NE 方法示意图中, 路由器 *N1* 有三个接口 IP 地址, 与三台单接口路由器和一台多接口路由器相连; 路由器 *N*2 有四个接口 IP 地址, 与三台单接口路由器和一台多接口路由器相连; 基于 IE 和 NE 方法获得的最终的位置频数矩阵, 建立 IE+NE 方法的位置频数矩阵, 路由器 N1 的位置频数矩阵为:

$$\boldsymbol{I1}=\begin{bmatrix} \mathrm{B},2,0.6667 \\ \mathrm{A},1,0.3333 \end{bmatrix}$$

$$\boldsymbol{N1}=\begin{bmatrix} \mathrm{B},2,0.5 \\ \mathrm{A},1.75,0.4375 \\ ?,0.25,0.0625 \end{bmatrix}$$

$$\boldsymbol{R1}=\boldsymbol{I1}\oplus\boldsymbol{N1}=\begin{bmatrix} \mathrm{B},2,0.6667 \\ \mathrm{A},1,0.3333 \end{bmatrix}\oplus\begin{bmatrix} \mathrm{B},2,0.5 \\ \mathrm{A},1.75,0.4375 \\ ?,0.25,0.0625 \end{bmatrix}=\begin{bmatrix} \mathrm{B},2,0.5834 \\ \mathrm{A},1.375,0.3854 \\ ?,0.125,0.0313 \end{bmatrix}$$

则路由器 *N1* 的位置为 B,置信度为 0.5834。

路由器 *N2* 的位置频数矩阵为:

$$\boldsymbol{I2}=\begin{bmatrix} \mathrm{A},3,0.7500 \\ \mathrm{B},1,0.2500 \end{bmatrix}$$

$$\boldsymbol{N2} = \begin{bmatrix} \text{A}, 3.25, 0.8125 \\ \text{B}, 0.5, 0.125 \\ ?, 0.25, 0.0625 \end{bmatrix}$$

$$\boldsymbol{R2} = \boldsymbol{I2} \oplus \boldsymbol{N2} = \begin{bmatrix} \text{A}, 3, 0.7500 \\ \text{B}, 1, 0.2500 \end{bmatrix} \oplus \begin{bmatrix} \text{A}, 3.25, 0.8125 \\ \text{B}, 0.5, 0.125 \\ ?, 0.25, 0.0625 \end{bmatrix} = \begin{bmatrix} \text{A}, 3.125, 0.7813 \\ \text{B}, 1.25, 0.1875 \\ ?, 0.125, 0.0313 \end{bmatrix}$$

则路由器 *N*2 位置为 A,置信度为 0.7813。

图 1-2 是同一台路由器不同方法定位的实例图, 使用 IE+NE 方法为路由器建立位置频数矩阵:

$$\boldsymbol{R} = \begin{bmatrix} 西班牙, 22, 0.88 \\ 美国, 1, 0.04 \\ 德国, 1, 0.04 \\ 英国, 1, 0.04 \end{bmatrix} \oplus \begin{bmatrix} 西班牙, 1092, 0.8605 \\ 美国, 4, 0.0032 \\ 德国, 1, 0.0008 \\ 葡萄牙, 1, 0.0008 \\ 中国, 1, 0.0008 \\ ?, 170, 0.1339 \end{bmatrix} = \begin{bmatrix} 西班牙, 557, 0.8703 \\ 美国, 2.5, 0.0216 \\ 德国, 1, 0.0204 \\ 英国, 0.5, 0.02 \\ 葡萄牙, 0.5, 0.0004 \\ 中国, 0.5, 0.0004 \\ ?, 85, 0.0669 \end{bmatrix}$$

该路由器位置为西班牙, 置信度为 0.8703。

IE+NE 方法会出现位置歧义导致定位失败的情况。该方法消除了 NE 方法中邻居缺失而影响定位的情况。

4 实验及结果分析

4.1 实验过程

4.1.1 实验数据

以 CAIDA ITDK 提供的 2017 年 10 月 MIDAR 和 IFFINDER 别名解析后获得数据作为路由器级拓扑数据。对路由器级拓扑数据 76,520,865 台路由器进行筛选及分析, 删除为主机的 IP 地址, 获得 41,798,800 台路由器, 其中 132,175 台多接口的路由器, 41,666,625 台单接口路由器。以 2018 年 3 月的 IP2location 商业地理数据库定位 IP 地址, 总计定位 42,186,037 个 IP 地址。基于以上数据, 利用本文提出的 RLoc 方法为 132,175 台多接口路由器定位地理位置。

为评价 RLoc 方法的准确率, 以 2017 年 12 月 IXP 数据集、2018 年 3 月 DDEC 工具获取的 IP 地址主机名映射的位置数据作为验证数据。其中, IXP 数据是 CAIDA 发布 IXP 的 json 文件中提取 IP 前缀和位置映射关系, 从而定位路由器。当路由器至少一个接口 IP 地址在 IXP 的 IP 前缀中, 则 IP 前缀的位置定位为路由器真实位置, 有些路由器接口 IP 地址可能对应的多个 IP 前缀有多个位置, 将所有位置均定位为路由器的位置, IXP 可定位 1,929 台路由器。DDEC 数据以 DDEC 工具对路由器接口 IP 地址的反向域名获取带有位置字符串的信息来定位。当路由器至少一个接口 IP 地址解析到位置, 则该位置定位为路由器真实位置, 有些路由器多个接口 IP 均解析地理位置且位置不同时, 将所有可能的位置均定位为路由器的位置, DDEC 数据可定位 26,566 台路由器。

为对比评价方法, 本文与 2017 年 10 月 CAIDA 发布的路由器位置数据集比较。RLoc 以路由器级拓扑数据集和 IP2location 商业地理数据库为输入, 定位路由器级拓扑数据集相同, 可直接比较定位结果数据集。为评价输入数据在地理定位中能够获得准确率的极值, 本文增加两个极值参照数据: 真值极大、真值极小。这两个数据可以获得相同输入数据情况 RLoc 最大、最小的准确率。

4.1.2 定位实验

IE 方法以 132,175 台多接口路由器、IP2location 定位数据库为输入, 为每台路由器建立位置频数矩阵, 选择置信度最大的位置作为路由器位置。定位路由器数量在国家级、城市级数量分别为 129,223 和 110,989; 由于位置歧义无法定位的数量分别为 2,952 和 21,186。

NE 方法以 41,666,625 台单接口路由器、路由器间连接关系、IP2location 定位数据库为输入, 为每台路由器建立邻居位置频数矩阵, 选择置信度最大的位置作为路由器位置。定位路由器在国家级、城市级数量分别为 120,990 和 118,004; 由于位置歧义无法定位的数量分别为 35 和 2,415; 由于邻居缺失无法定位的数量分别为 477 和 1083。

IE+NE 方法以 IE 与 NE 方法得到的位置频数矩阵为输入, 结合两个矩阵, 建立新的位置频数矩阵, 选择置信度最大的位置作为路由器位置。定位路由器在国家级、城市级数量分别为 131,967 和 126,893; 由于位置歧义无法定位的数量分别为 208 和 5,282。

4.1.3 数据验证

以 IXP、DDEC 定位的位置作为真实路由器位置, 评估 RLoc 提出的三种方法定位的准确率。当路由器定位结果与 IXP 定位的路由器结果一致时, 定位正确, 反之, 错误。如图 1-2 所示, 路由器使用 RLoc 提

出的三种方法中定位结果为西班牙， IXP 定位位置也是西班牙，则路由器地理定位结果准确。

如下图 4-5 所示，分别为路由器使用 IE、NE 方法定位实例图。从图中可知，该路由器 IE、NE 和 IE+NE 方法定位位置为印度，但在 IXP 定位位置为新加坡，此时该方法定位结果错误。路由器定位虽然结果错误，但在位置频数矩阵中都存在真值位置，新加坡。

为获得 RLoc 方法准确率的最大最小值，利用 EI+NE 方法构建的路由器位置频数矩阵中的位置信息，作为路由器的所有可能位置。真值极大数据为获得最大的准确率，即选择与真实位置一致的位置作为路由器的位置。真值极小数据为获得最小的准确率，即选择与真实数据不一致的位置作为路由器的位置。如图 4-5 所示的路由器定位中, IE+NE 方法得到的位置数据有印度、新加坡、澳大利亚、美国、德国、孟加拉国、肯尼亚、印度尼西亚。真值极大数据选择新加坡为路由器的位置。相反，真值极小数据选择除新加坡以外的位置为路由器的位置。

同时，利用相同的验证数据集评价 2017 年 10 月 CAIDA 发布的路由器位置数据集的覆盖率和准确率并与 RLoc 结果比较。

图 4　在 IE 方法下路由器定位结果实例图

Figure 4　Instance of router geolocation's result in IE

(注: 图中多接口路由器共有 40 个接口 IP 地址，其中 25 个在印度, 7 个在新加坡, 6 个在澳大利亚, 2 个在美国。)

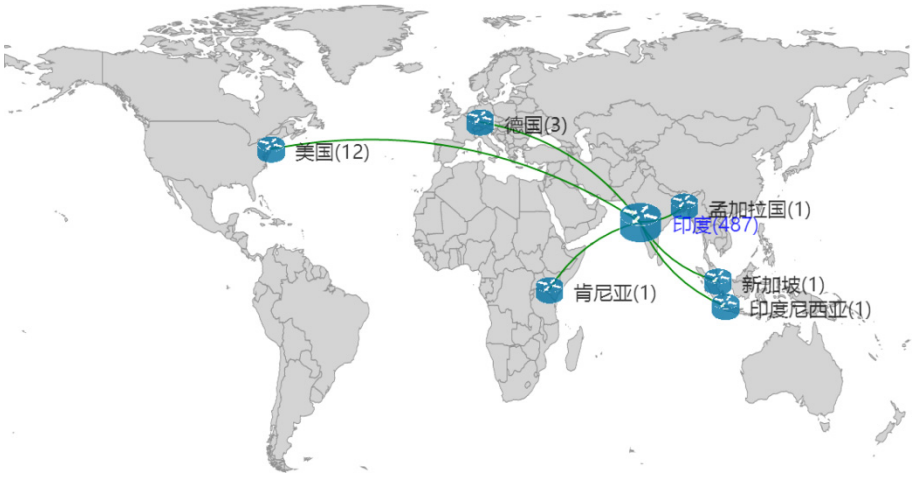

图 5　在 NE 方法下路由器定位结果实例图

Figure 5　Instance of router geolocation's result in NE

(注: 图中多接口路由器共有 965 个邻居路由器，其中 487 个在印度, 12 个在美国, 3 个在德国, 1 个在孟加拉国, 1 个在肯尼亚, 1 个在新加坡, 1 个在印度尼西亚, 459 个无法定位。)

4.2　覆盖率

RLoc 提出的 EI、NI、EI+NI 获得的数据与 CAIDA 数据的路由器定位覆盖率如表 1 所示。

就覆盖率而言，IE+NE 方法更好，能够消除 NE 方法中邻居缺失而导致无法定位的情况，同时也减少了 IE 和 NE 方法出现位置歧义情况。CAIDA 数据

表 1 各方法定位多接口路由器覆盖率

Table 1 The coverage of various methods to geolocate multi-interface routers

方法	国家	城市
IE	97.77%(129,223)	83.97%(110,989)
NE	91.54%(120,990)	89.28%(118,004)
IE+NE	99.84%(131,967)	96.00%(126,893)
CAIDA	98.91%(130,735)	59.52%(78,672)

的国家级覆盖率与城市级覆盖率相差较大, 这是由于 IP 地理定位数据库在城市级对路由器接口定位在同一位置的情况较低, 导致路由器无法在城市级粒度上定位。

IE 在国家级优于 NE,但 NE 在城市级优于 IE。在国家级定位中, IE 方法在 IP 地理定位数据库中对路由器接口 IP 地址定位结果一致性更高, 出现位置歧义情况较小; 而 NE 方法不仅受位置歧义情况影响, 邻居路由器位置缺失也降低了定位的覆盖率。在城市级定位中, 由于 IP 地理定位数据库在城市级定位粒度更加精细, 使路由器可能的城市位置增加, IE 方法产生了较多的位置歧义情况; 而 NE 方法以单接口路由器的位置作为输入, 在原有国家级定位失败的基础上并未产生更多位置歧义情况。

综合而言, 在国家级, RLoc 方法定位了 91.54%~99.84%, CAIDA 定位了 98.90%, 两数据基本持平。在城市级, RLoc 方法定位了 83.97%~96.00%, CAIDA 定位了 59.52%, RLoc 比 CAIDA 数据高出了 24.45%~36.48%。

4.3 准确性

以 DDEC 与 IXP 数据分别评价 RLoc 方法获得的数据、极值数据和 CAIDA 数据, 在国家级、城市级路由器定位的准确率, 如下图 6-9 所示: 其中, 堆积柱形图表示各方法验证路由器的总数, 灰色柱形表示验证正确的数量, 白色柱形表示验证错误的数量。折线图表示各方法的准确率。

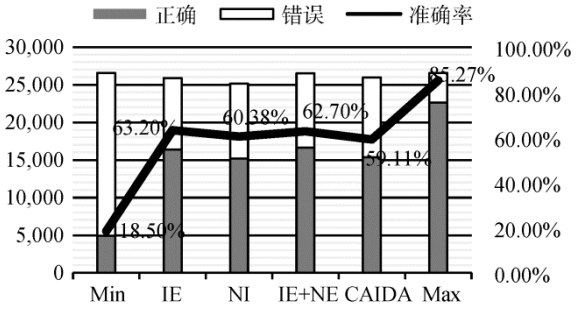

图 6 DDEC 验证国家级各方法的准确率

Figure 6 Country-level accuracy by DDEC validation

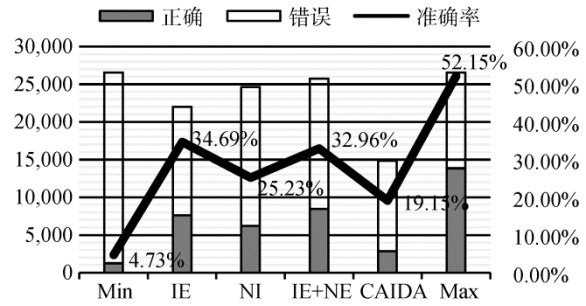

图 7 DDEC 验证城市级各方法的准确率

Figure 7 City-level accuracy by DDEC validation

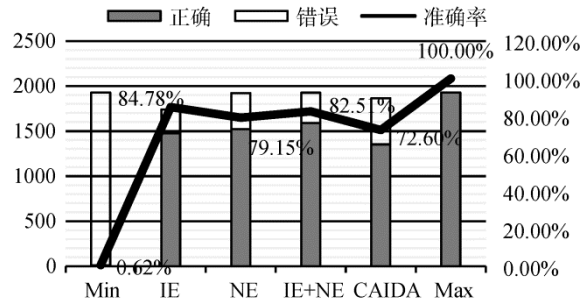

图 8 IXP 验证国家级各方法的准确率

Figure 8 Country-level accuracy by IXP validation

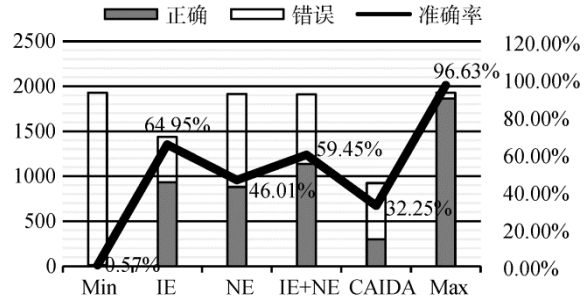

图 9 IXP 验证城市级各方法的准确率

Figure 9 City-level accuracy by IXP validation

就准确率而言, 在国家级利用 DDEC 验证极值范围为 18.50%~85.27%。RLoc 为 60.38%~63.20%, CAIDA 为 59.11%。利用 IXP 验证极值范围为 0.62%~100.00%, RLoc 为 79.15%~84.78%, CAIDA 为 72.60%。在城市级利用 DDEC 验证极值范围为 4.73%~52.15%, RLoc 为 25.23%~34.69%, CAIDA 为 19.15%。利用 IXP 验证极值范围为 0.57%~96.63%, RLoc 为 46.01%~64.95%, CAIDA 为 32.25%。RLoc 数据无论国家级、城市级的准确率均较 CAIDA 的数据有显著提升。对于输入数据获得最大、最小的准确率而言, RLoc 方法的准确率还有一定的上升空间。

在 DDEC 数据验证准确率时, 无论国家级或城市级, 无论各方法和真值的极大均不高。这可能是因为 DDEC 方法对路由器接口 IP 地址定位准确率不高或该方法获得的数据并非路由器真实的位置导致。

城市级定位的准确率均较国家级准确率低。这可能是 IP2location 商业数据库对 IP 地址定位城市级准确率不高导致。

RLoc 提出的方法中, IE+NE 较 IE 准确率低, 这是由于 NE 准确率比 IE 的准确率低, 导致 IE+NE 方法将两个数据结合后准确率比 IE 方法低, 比 NE 方法高。均衡各方法的路由器定位覆盖率和准确率, IE+NE 方法结合了路由器的接口位置信息与邻居路由器的位置信息, 效果更好。

5 局限性

本文提出了一种多接口路由器地理定位方法—RLoc, 虽然定位路由器的覆盖率和准确率均较现有数据有明显提升, 但仍存在局限性。RLoc 方法以路由器级拓扑数据、商业地理定位数据库为输入, 如果没有这两项工作, 该方法无法实现对路由器地理定位。同时, 输入数据的质量也将影响定位结果的质量。

RLoc 方法在实际应用中前提条件有以下几点:

1、每台路由器的各接口 IP 地址均被发现。当 IP 地址被正确别名解析, 且路由器的各接口 IP 地址均被发现时, 保证了路由器的接口的完整性, 使 IE 方法定位路由器定位更加准确。

2、路由器间连接关系完整并正确。当每台路由器的连接关系完整并正确推断时, 保证了每台路由器的邻居路由器的完整性, 使 NE 方法定位路由器位置更加准确。

3、IP 地理定位数据库的准确性和覆盖率高。每台路由器接口 IP 地址与单接口路由器的地理定位结果直接影响了对路由器地理定位结果。

RLoc 方法更适用于在路由器组网中, 一台路由器的接口 IP 地址大多数被分配来自相同 IP 地址段, 路由器与路由器相互通信、连接, 其位置分布也临近。在这样的网络中, RLoc 方法的准确率和覆盖率更高。

6 结论

RLoc 充分利用同一台路由器上不同接口 IP 地址在相同位置和相连路由器间地理位置相近这两个事实, 具有比现有公开数据的更高的覆盖率和准确率。本文综合定位覆盖率和准确率两个方面, 选择 IE+NE 方法。覆盖率上, 国家级达到 99.80%, 城市级达到 96.00%, 比 CAIDA 数据分别高出 0.93%和 36.48%; 准确率上, IXP 验证国家级达到 82.51%, 城市级达到 59.45%, 比 CAIDA 数据分别高出 9.91%和 27.20%; DDEC 验证国家级到达 62.70%, 城市级达到 32.96%, 比 CAIDA 数据分别高出 3.59%和 13.81%。

RLoc 无需实施新的大规模网络测量, 或构建及校对时延-距离模型; 无需在人工辅助下解析域名、Whois、网页等 IP 地址相关信息; 易于实施与重复, 结果易于更新和评估。适用于对大规模路由器级网络拓扑映射地理位置。

本文在定位多接口路由器时, 发现当一个多接口路由器出现多个位置时, 可能由于路由器间共享 IP 地址空间所导致。那么, 这样的多位置的路由器可能是国家间或城市间的网络边界点的路由器, 也是国家内或城市内的骨干路由器。在未来的工作中, 将着重研究多位置路由器在识别国家边界、识别骨干路由器方面的问题。

参考文献

[1] Motamedi R, Rejaie R, Willinger W. A Survey of Techniques for Internet Topology Discovery[J]. *IEEE Communications Surveys & Tutorials*, 2014, 17(2):1044-1065.

[2] K. Keys, “iffinder, a tool for mapping interfaces to routers,” See http://www. caida.

[3] Spring N, Mahajan R, Wetherall D. Measuring ISP topologies with rocketfuel[J]. *IEEE/ACM Transactions on Networking*, 2004, 12(1): 2-16.

[4] Bender A, Sherwood R, Spring N. Fixing ally's growing pains with velocity modeling[C]// ACM SIGCOMM Conference on Internet Measurement 2008, Vouliagmeni, Greece, October. DBLP, 2008: 337-342.

[5] Keys K, Hyun Y, Luckie M, et al. Internet-scale IPv4 alias resolution with MIDAR[J]. *IEEE/ACM Transactions on Networking*, 2013, 21(2): 383-399.

[6] CAIDA Internet topology data kit http://www.caida.org/data/internet-topology-data-kit/

[7] Gueye B, Ziviani A, Crovella M, et al. Constraint-based geolocation of internet hosts[J]. *IEEE/ACM Transactions on Networking*, 2006, 14(6):1219-1232.

[8] Katz-Bassett E, John J P, Krishnamurthy A, et al. Towards IP geolocation using delay and topology measurements[C]// ACM SIGCOMM Conference on Internet Measurement 2006, Rio De Janeriro, Brazil, October. DBLP, 2006:71-84.

[9] Wong B, Stoyanov I. Octant: a comprehensive framework for the geolocalization of internet hosts[C]// Usenix Conference on Networked Systems Design & Implementation. USENIX Association, 2007: 23-23.

[10] Liu H, Zhang Y, Zhou Y, et al. Mining checkins from location-sharing services for client-independent IP geolocation[C]// INFOCOM, 2014 Proceedings IEEE. IEEE, 2014: 619-627.

[11] Moore D, Periakaruppan R, Donohoe J, et al. Where in the world is netgeo.caida.org?[C]// Internet Society Conference. 2000.

[12] Padmanabhan V N, Subramanian L. An investigation of geographic mapping techniques for internet hosts[J]. *Acm Sigcomm Computer Communication Review*, 2001, 31(4): 173-185.

[13] 宋建, 许可, 宋美娜,等. 一种评估国内 IP 地址库可信度的方法

[C]// 全国开放式分布与并行计算学术年会. 2014.

[14] 王婷, 宋俊德, 宋美娜. 一种基于关联规则挖掘的 IP 定位方法[J]. *东南大学学报(自然科学版)*, 2015, 45(4):657-662.

[15] Gharaibeh M, Shah A, Huffaker B, et al. A look at router geolocation in public and commercial databases[C]// Internet Measurement Conference. 2017:463-469.

[16] Huffaker B, Fomenkov M, Claffy K. DRoP: DNS-based router positioning[M]. ACM, 2014.

[17] Laki S, Matray P, Haga P, et al. A detailed path-latency model for router geolocation[C]// International Conference on Testbeds and Research Infrastructures for the Development of Networks & Communities and Workshops, 2009. Tridentcom. IEEE, 2009: 1-6.

[18] Laki S, Csabai I, Vattay G. A model based approach for improving router geolocation[J]. *Computer Networks the International Journal of Computer & Telecommunications Networking*, 2010, 54(9): 1490-1501.

朱金玉 于 2015 年在四川大学计算机科学与技术专业获得学士学位。现在哈尔滨工业大学计算机科学与技术专业攻读博士学位。研究领域为互联网关键资源安全, IP 地理定位。Email : zhujinyu@hit.edu.cn

张宇 于 2009 年在哈尔滨工业大学计算机系统结构专业获得博士学位。现任哈尔滨工业大学计算机网络与信息安全技术研究中心副教授。研究领域为互联网关键资源安全, 网络拓扑测量, 未来网络体系结构。Email: yuzhang@hit.edu.cn

曾良伟 于 2017 年在哈尔滨工业大学信息安全专业获得学士学位。现在哈尔滨工业大学网络空间安全专业攻读硕士学位。研究领域为互联网关键资源安全。

余卓勋 于 2016 年在哈尔滨工业大学信息安全专业获得学士学位。现在哈尔滨工业大学计算机科学与技术专业攻读博士学位。研究领域为网络测量, 互联网关键资源安全。

张宏莉 女, 博士, 哈尔滨工业大学计算机科学与技术学院教授、博士生导师, 研究方向为网络安全, 网络测量和网络计算。