文章编号：2095-2163(2019)05-0334-05 中图分类号：TP393.07 文献标志码：A

IP 地理定位优化方法初探

曾良伟，张 宇，朱金玉
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要： IP 地址地理定位是网络拓扑测绘的基础，然而如何准确定位 IP 地址是一个难题。为了提高 IP 地址定位的准确性，本文提出了 3 种优化 IP 地理定位准确性的方法。首先综合各个 IP 地理定位数据库的优点，合并新的地理定位数据库，新集成数据库定位一致率较其他数据库提高了 2%。然后获取互联网路由器接口信息并对路由器定位，定位后的路由器可以作为地标点定位接口 IP 地址以及相邻 IP 地址。最后搜集路径信息，从路径中推断出地区的边界网关 IP 地址，得到的网关 IP 地址列表能够对地区内部的 IP 地址定位提供帮助。

关键词： 网络拓扑测绘；IP 地理定位；路由器定位；地区网关

Exploration on IP geolocation optimization method

ZENG Liangwei, ZHANG Yu, ZHU Jinyu
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 IP address geolocation is the basis of network topology mapping, but how to accurately locate an IP address is a problem. In order to improve the accuracy of IP address location, this paper proposes three methods to optimize the accuracy of IP geolocation. First, the paper integrates advantages of each IP geolocation database to merge the newly geolocation database. The newly integrated database geolocation accuracy is 2% higher than other databases. Then the paper collects the Internet router interface information and locates the router. The located router can be used as a landmark point to locate the interface IP address and the adjacent IP address. Finally, the paper collects the path information, infer the regional border gateway IP address from the path. The list of gateway IP addresses can help locate the IP address within the region.

【Key words】 network topology mapping; IP geolocation; router geolocation; region gateway

0 引 言

如何将网络空间与地理信息相互映射，将虚拟、动态的网络空间测绘成可靠、有效的网络空间地图，是一项非常重要的工作[1]。IP 地理定位技术分为 2 类。通过 ISP、查询 whois 信息等基于服务商填写的方法获取 IP 地址对应地理信息[2]，或者利用测量延迟或网络拓扑对 IP 地址进行定位[3]。本文从 3 个角度对 IP 定位准确性进行优化。首先是合并各个公司的 IP 定位数据库，IP 定位数据库包含了 IP 地址段对应的地理信息以及额外信息，提供接口供用户查询 IP 地址定位信息。现在流行的数据库不仅难于被校正，而且由于缺少建立这些数据库的方法的相关信息，其准确度也仍有待商榷；然后是定位路由器，网络空间中有许多路由器节点，而且都是信息传输的中转站。若能准确定位路由器地理信息，对路由器相连的终端节点进行定位将变得非常简单；最后是识别地区的网关 IP 地址列表，这些网关 IP 地址可以作为地理定位中的地标点指导该地区的地理定位，地区外的监测点测量地区内目的 IP 地址时，路径中一定经过该地区的网关 IP 地址且经过网关 IP 地址后的路径 IP 地址均属于该地区。本文主要贡献如下：

（1）将多个流行的地理定位数据库合并为一个更为准确的定位数据库。

（2）提出 3 种路由器定位方法。

（3）提取 Traceroute 数据中的网关 IP 地址信息。

1 相关工作

1.1 IP 地理定位

GeoTrack[4] 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。DRoP 算法[5] 提取和解码路由器接口的主机名中包含的地理信息字符串来给出定位。NetGeo 算法[6] 通过直接查询 Whois 数据库来推测主机位置信息。

基金项目： 国家重点研发计划（2016YFB0801303-2）。

作者简介： 曾良伟（1995-），男，硕士研究生，主要研究方向：网络拓扑测量；张 宇（1979-），男，博士，副教授，主要研究方向：网络测量、网络安全、未来网络；朱金玉（1993-），女，博士研究生，主要研究方向：网络空间测绘、IP 地理定位。

收稿日期： 2019-05-20

MaxMind、IP2Location、埃文、IPMarker、IPIP. NET 等基于数据库的商业定位系统综合各种方法来收集、获取位置信息，定位精确度可为国家、城市、甚至于邮编级。

1.2　路由器定位

CAIDA 维护宏观互联网拓扑数据工具包(ITDK)，ITDK 包含大面积测量全球互联网得到的链接和路由数据。对于在路由器级别研究 Internet 的拓扑结构以及其他用途非常有用。测量数据利用 MIDAR 和 iffinder 工具合并路由器信息，本文使用路由器接口信息和路由器链接信息来对路由器进行定位[7]。

1.3　边界推测

CFS 算法[8]通过多个约束源缩小一个给定的对等链接可能的位置范围，从而推断对等互连所在的地理位置及互连关系类型。MAP-IT 方法[9]根据多条的 traceroute 路径中提取 IP 地址的接口邻居集，提出启发式推断方法来识别域间连接的接口和所属 AS。Bdrmap[10]方法利用有针对性的 traceroute、traceroute 特性知识和结构化启发式方法集中拓扑约束，从而正确地识别边界路由器的域间链接。

2　改进 IP 地址定位

本文为提高 IP 地址定位准确率，提出了 3 种改进方法。首先是收集多个流行的地理定位数据库，将数据库各自的优点合并到新的地理定位数据库中，提高 IP 地址定位的准确性。其次是对 CAIDA 的 ITDK 中路由器数据进行定位，定位路由器之后相邻 IP 地址地理信息也会更加准确。最后，是对地区网关 IP 地址的识别，这不仅是对网络拓扑到地理位置的映射，更是对网络边界与地理边界间联系的探索。

2.1　定位数据库合并

本文使用的地理定位数据库有 IP2Location、GEOLite、IP2Location Lite、埃文离线数据库，IPMarker 和 IPIPNET 数据库。通过地名翻译、地名映射以及记录合并，最终得到一个融合各个数据库优点的集成数据库。对此可做阐释分述如下。

2.1.1　地名翻译及映射

由于埃文离线库、IPMarkder 和 IPIPNET 数据库地理信息为中文编码，需要先翻译为英文、再融合到新数据库。本文使用百度翻译接口[11]对中文地名进行转换。

各个定位数据库有独特的地名命名方式，不能简单地使用字符串不相同来判断地名指向不同地点，经过统计数据库之间的大部分差异问题均可归结为单词组合或添加删除几个字符方面，因此采用计算编辑距离和字符串长度比较的方法来解决差异问题。在此基础上，通过此算法对不同数据库的地名进行映射，在国家、地区、城市级分别进行映射。

2.1.2　数据合并

由于各个数据库中 IP 地址段可能不相同，研究就需要提取重叠的 IP 地址段，重叠 IP 地址段定义如图 1 所示。

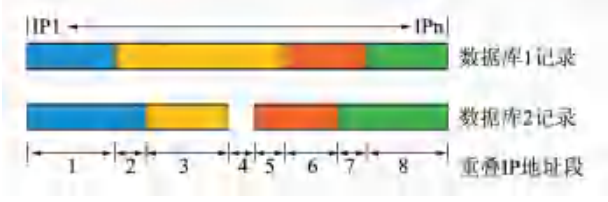

图 1　重叠 IP 地址段

Tab. 1　Overlapping IP address segment

研究中，还要对每个 IP 地址段根据各个数据库定位数据计算出最优最全的定位数据。各个地理定位数据库对 IP 地址段地理信息分别投票，选出票数最多的地理信息作为 IP 地址段的定位结果。

插入函数定义了如何取出选中数据库集合中有用的字段信息写入新数据库中。由于数据库集合中的字段可能重复，定义了数据库记录的优先级可表示为：

IP2Location > IP2Location Lite > GeoLite2 > Aiwen > IPMarker > IPIPnet

选择集合中优先级最高的记录，提取所有属性信息。若此记录缺少合并数据库的一些属性信息，按照优先级依次到其它记录中查找。若均没有，则在未加入集合的记录中按照优先级寻找对应属性信息。将重叠 IP 地址段和现有属性信息组合为记录，插入合并数据库。

至此，完成了多个地理定位数据库的合并。

2.2　路由器定位

定位路由器地理位置可以修正路由器对应接口 IP 地址的定位信息，并且能对相连的终端 IP 地址定位优化提供帮助。根据路由器多个接口应在同一地点以及相邻路由器地理位置较近的事实，提出了综合选举路由器定位方法，本文定位 CAIDA ITDK 推断出的路由器列表[7]。

将路由器对应接口列表以及定位数据库作为输入。建立当前路由器接口位置矩阵。可将其写作如下数学形式：

$$I=\begin{bmatrix} G_1 & C_1 & F_1 \\ G_2 & C_2 & F_2 \\ \cdots & \cdots & \cdots \\ G_n & C_n & F_n \end{bmatrix}, \tag{1}$$

其中,G_i为接口定位不同的地理位置信息;C_i为定位 G_i的次数;F_i为定位 G_i的置信度,研究推得其计算公式如下:

$$F_i=\frac{C_i}{C_1+C_2+\cdots+C_n}, \tag{2}$$

此时,还要分离单接口与多接口路由器,将单接口路由器列表、路由器链接关系以及地理定位数据库作为输入,再多次迭代计算定位多接口路由器。每个多接口路由器每次迭代维护一个邻居列表以及邻居位置矩阵。相应地,其数学公式可表示为:

$$L_i=[N_1,N_3,\cdots,N_x], \tag{3}$$

$$\boldsymbol{N}_i=\begin{bmatrix} G_1 & R_1 \\ G_2 & R_2 \\ \cdots & \cdots \\ G_n & R_n \end{bmatrix}, \tag{4}$$

其中,L_i 表示与第 i 个多接口路由器的路由器列表(单接口或多接口路由器);$\boldsymbol{N}_i$ 矩阵对应当前第 i 个路由器的定位情况;G_x为不同的地理位置信息;R_x为对应地理位置的权重信息。

对第 i 个路由器的定位矩阵,可由以下公式计算得出:

$$N*i=\frac{1}{|L_i|}\times\sum_{j\in L_i}\boldsymbol{N}_j\begin{bmatrix} G_x*R_x \\ G_y*R_y \\ \cdots*\cdots \\ G_z*R_z \end{bmatrix}, \tag{5}$$

$\boldsymbol{N}_j$ 的定位矩阵为路由器 j 上次迭代的定位矩阵,将相连的路由器的各个定位结果的权值按比例合并到当前迭代的定位矩阵 N^*i 中。

进一步地,将合并接口位置矩阵和邻居位置矩阵,得到新的定位矩阵,即:

$$\boldsymbol{M}_i=\boldsymbol{I}_i\oplus\boldsymbol{N}_i, \tag{6}$$

其中,$\boldsymbol{M}_i$ 为路由器 i 的新定位矩阵;$\boldsymbol{I}_i$ 为接口选举方法得到的定位矩阵;$\boldsymbol{N}_i$ 为邻居选举方法的到的定位矩阵。⊕运算定义为:计算两矩阵中相同定位信息权重的平均值,具体如下所示:

$$K_{ij}=\frac{(F_{ij}+R_{ij})}{2}. \tag{7}$$

式(7)表示了路由器 i 对应的 j 定位信息权重 K_{ij}为接口选举矩阵权重 F_{ij}和邻居选举矩阵权重 R_{ij}平均值。

2.3 边界网关识别

边界网关识别指从 Traceroute 数据中识别出网关 IP 地址信息,本文使用 Caida 公开的 Traceroute 作为源数据进行实验,通过数据标记以及筛选,最终得到目标地区的网关 IP 地址列表。识别的网关 IP 地址可以用于指导地区内部 IP 地址定位。

2.3.1 数据标记与筛选

为了准确识别路径中的边界网关 IP 地址,利用时延信息、自治域信息、地理信息以及 whois 信息寻找采集 Traceroute 数据中的目标地区网关 IP 地址。

对路径上的 IP 地址标记各个属性,即:

hop|IP 地址|RTT|自治域国家|GEO|WHOIS

标记结束后,开始对错误数据进行筛选。

2.3.2 网关识别

RTT 值作为 traceroute 测量时的真实时延信息,直观地表明了两 IP 地址间距离的远近,因此选择 *RTT* 最大的 IP 地址作为网关 IP 地址,且权值设为 2。AS、WHOIS、GEO 信息提取的 IP 地址的位置信息,均来源于 IP 地址前缀映射,而非单个 IP 地址,并且在边界或网关区域更经常出现共享地址空间的情况,从而导致映射错误,因此文中设置权值为 1。各方法权值见表 1。

表 1 方法对应权值表

Tab. 1 Corresponding weight table of the methods

RTT	rDNS	AS	GeoIP	WHOIS
2.0	1.5	1.0	1.0	1.0

在网关候选集位置矩阵中,每种属性对第一次出现位置为中国的行数按照权重进行投票,过程中,选择 *RTT* 差值最大的行数按照权重投票。统计每行最终得票数,票数最多的行的 IP 地址则为当前网关 IP 地址 G。

3 实现结果与分析

3.1 定位数据库合并

各个数据库记录数、国家级记录、地区级记录、城市级记录数量统计信息详见表 2。

为了验证数据库一致率,本文使用表 2 中的路由器接口数据。将接口选举中均定位在同一地点的多接口路由器判断为定位准确,计算出均定位在同一地点路由器 IP 地址占多接口路由器 IP 地址总量的百分比作为定位数据库一致率。研究得到数据库

接口验证一致率结果如图 2 所示。

表 2　数据库信息统计表

Tab. 2　Statistical table of database information

数据库	记录数	国家级	地区级	城市级
IP2Location	12 365 108	244	3 126	94 954
GeoLite2	2 650 913	247	2 600	83 435
IP2Location Lite	3 327 891	243	3 075	73 321
埃文	16 106 220	236	2 981	91 002
IPMarker	3 283 162	254	2 326	29 243
IPIPNET	1 061 445	811	845	380
合并数据库	18 553 448	253	4 038	91 509

图 2　数据库接口验证一致率

Fig. 2　Consistency rate of database interface verification

3.2　路由器定位

利用搜集到的 1 929 个 IXP 数据对路由器定位结果进行验证，并与理论最小值、最大值以及 CAIDA 定位准确率结果作比较。理论最小值为：当路由器某一定位结果不符合 IXP 定位结果，则判断为错误。理论最大值为：当路由器只要有一个定位结果符合 IXP 定位结果，则判断为正确。最终得到准确率如图 3 所示，可以看出路由器定位方法的准确率有较大的提高。

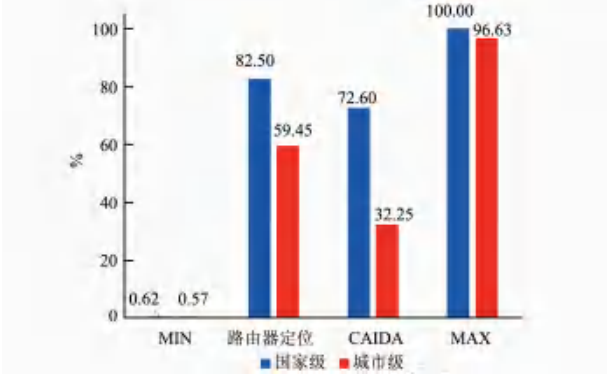

图 3　国家级和城市级定位准确率

Fig. 3　Accuracy of national and city-level positioning

3.3　边界网关识别

此方法共识别出 1 245 个网关 IP 地址，使用国内监测点 Ping 检验，有 1 130 个 IP 地址在国内，45 个检验错误以及 70 个无返回结果。人工检验时，数据中确定为网关的错误个数为 24。研究将识别的网关 IP 地址交给某运营商，验证 28 个网关中，21 个正确，7 个验证错误。对地区内某一目标 IP 地址的测量结果如图 4 所示。

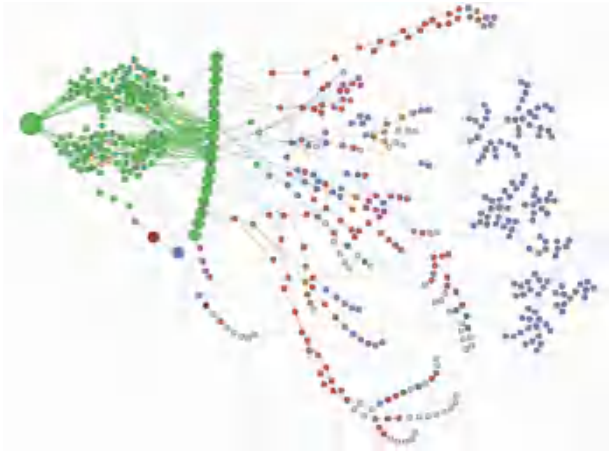

图 4　网关 IP 地址内外部连接示意图

Fig. 4　Diagram of internal and external connection of gateway IP address

图 4 中，每个点代表一个 IP 地址，利用数据库定位的相同地点用相同颜色做出表示。作为网关 IP 地址，则应为地理位置的分界线，那么在网关左侧的非绿色圆圈其定位结果错误，网关右侧的绿色圆圈定位结果错误。由此得到的网关 IP 地址，在一定程度上可以指导地理定位。

4　结束语

本文融合各个地理定位数据库数据，增加了 IP 地址定位的 2% 的一致率。提出了 3 种路由器定位方法，对 Caida ITDK 中的路由器进行定位，并且验证了定位结果的准确率。从公开的 Traceroute 数据中提取目标地区的网关 IP 地址列表，使用 Ping 测量、人工检验以及运营商检验的方式对得到的列表进行验证。在未来，则会利用路由器定位结果对附近的终端 IP 地址进行定位，以及通过得到的地区网关 IP 地址对地区内部 IP 地址定位。

参考文献

[1] 埃文. 网络空间地区测绘的意义[EB/OL]. [2018-01-12]. https://blog.csdn.net/aiwenipgeolocation/article/details/79040485.

（下转第 340 页）