IP 地址地理定位算法及其有效性分析

杨云[1]，龚俭[1]

（1. 东南大学计算机科学与工程学院，南京，211189）

摘　要：IP 地址地理定位服务已经被视为未来互联网的一项不可或缺的基础服务，网络管理和网络安全都需要 IP 地址地理定位服务的支持。为满足 IPCIS 系统在网络行为监控和网络管理方面的需求，本文利用 CAIDA 提供的 traceroute 路径信息提出了多跳 IP 地址关联分析算法以及相同接入路由的 IP 地址关联分析算法，并以 IPCIS 现有的 IP 地址使用位置库为第三方数据源，建立统计模型对算法的有效性进行了分析。

关键字：IP 地址地理定位；IP 地址关联分析算法；有效性分析

IP Geolocation Algorithm And Its Efficiency Analysis

Yang Yun[1], Gong Jian[1]

(1. School of Mechanical Engineering, Southeast University, Nanjing, 211189)

Abstract: IP geolocation will be regarded as an indispensable web service in the future of Internet. Network management and security require the support of IP geolocation service. To meet the needs of IPCIS system in network behavior monitoring and network management, we propose two algorithms using multiple hops IP address correlation analysis and same access routing IP address correlation analysis based on CAIDA's network topology data. The validity of the algorithms are analyzed by establishing a statistical model which using IPCIS' IP addresses geolocation database as a third-party data source.

Key Words: IP Geolocation; IP geolocation Co-relation analysis algorithm; efficiency analysis

IP 定位技术是指仅依靠机器的 IP 地址确定机器的实际使用位置，使用位置信息包括 IP 地址所在的国家、省（区域）、城市。

IP 地址地理定位可以应用于许多网络服务，例如，根据地理信息显示当地的天气情况以及当地新闻等相关信息的网络服务，根据地理位置自动选择语言显示网站内容，根据地理位置选择最近的服务器等。此外网络服务与人们的生活越来越密切，网络的安全性不容忽视，许多的互联网管理和安全的应用都需要 IP 地址地理定位服务的支持，确定网络故障或者攻击源头，把网络上的操作对应到实际的操作地点，并通过身份验证进一步找到实际的操作人，从而提高网络的安全系数，有效的遏制网络欺骗。

211 工程三期公共服务体系建设项目“中国教育和科研计算机网主干网和重点学科信息服务体系升级扩容工程”中建设了 CERNET 主干网运行管理和安全保障系统，其中包含有一个 IP 地址综合信息系统 IPCIS（IP Comprehensive Information System），本文利用 CAIDA 提供的 traceroute 路径信息提出 IP 地址地理定位算法，通过关联规则，对 traceroute 数据中的 IP 地址进行关联分析，找到那些在同一地理位置的 IP 地址，形成 IP 地址组，同组内通过使用位置已知的 IP 地址推断使用位置未知的 IP 地址的实际使用位置，之后基于 IPCIS 现有的 IP 地址使用位置库内容对算法结果的可信度进行统计验证，以此作为 IP 地址使用位置库的更新依据，以达到不断维护 IPCIS 系统有效性的目的。

文章主要分为以下四个部分，第一部分介绍实验数据，包括 CAIDA 的 traceroute 路径数据，IPCIS 系统现有的 IP 地址使用位置库。第二部分介绍两种 IP 地址定位算法的原理及其实现。第三部分基于统计学原理建立统计模型对本文提出的两种 IP 地址地理定位算法进行准确性分析。最后一部分对全文进行总结，并提出展望。

1 实验数据介绍

1.1 Traceroute 路径信息

CAIDA 是最重要的互联网测量数据提供者之一，CAIDA 的数据收集主要围绕路由寻址，拓扑路径，域名服务，流量分析等几个方面，其中拓扑路径信息是通过 traceroute 方法利用路由追踪技术主动获取网络路由信息得到的，发现探测点和目标主机之间的路由 IP 地址序列，其中的目标 IP 地址是从在互联网上每个 IPv4 /24 前缀中随机选取的。一条 traceroute 数据包含了到指定 IP 的传输路径信息，包括路由器 IP 地址，到各个路由器的往返时延、是否成功追踪等。CAIDA 一个 cycle 的 traceroute 数据的目标地址覆盖了全球所有的 C 类地址，是对全球互联网的一次完整的测量。图 1 是从探测点 213.254.227.26 到目标点 87.180.111.215 的一条 trace 数据实例。

```
T 213.254.227.26          87.180.111.215
7 2921 1388626836 R 264.106 6 59 S 0 C
213.254.227.25,0.185,1   141.136.110.118,211.832,1
62.157.249.49,211.849,1 62.154.58.210,217.057,1
87.186.239.49,217.238,1
```

图 1 CAIDA 中 trace 路径实例

1.2 IPCIS 现有的 IP 地址使用位置库

IPCIS 分层数据库中的 IP 地址使用位置库是开发者通过合并目前几个准确度较高的开源数据库 Maxmind、IP2Location、QQ 纯真，以及 CERNET 自身的 whois 数据库得到的，该数据库包含 3479773 个 IPv4 地址段的使用位置信息，总计覆盖了 242 个国家及地区，对拥有有效 IPv4 地址数量前 6 位的国家的 IP 覆盖率能够达到 95%以上(美国 98.64%，中国 99.78%，日本 98.89%，英国 97.71%，德国 97.87%，韩国 97.91%)。此外，为了提升 IP 地址使用位置库的覆盖率及准确性，使其能够提供更为准确、完整的定位信息，IP 地址使用位置库的开发者还为数据库建立了完善的更新机制对数据库进行不断的更新，因此 IPCIS 现有的 IP 地址使用位置库的准确性和覆盖率是可以保障的。

本文使用 IPCIS 现有的 IP 地址使用位置库作为第三方数据源建立了统计模型对定位算法的准确性进行分析。

2 IP 地址地理定位算法介绍

2.1 连续多跳 IP 地址关联分析算法

(1)算法的原理

定义序列 A=[IP_1，IP_2，…，IP_{n-1}，IP_n](n≥3)为一条 trace 数据的连续多跳 IP 地址序列(一条完整 trace 数据或者 Trace 数据的一部分)，若该序列的首尾 IP 地址(IP_1，IP_n)的使用位置相同，则 trace 序列首尾 IP_1 和 IP_n 之间的 IP 地址的使用位置均相同，我们称序列 A 之为相关 IP 地址序列。

如图 2 所示，序列 A=[IP_{A1}，IP_{A2}，…，IP_{An-1}，IP_{An}]为一个 IP 地址序列，IP_{A1}、IP_{An} 使用位置均为城市 A，则地址序列中其它 IP 地址的使用位置也为城市 A。

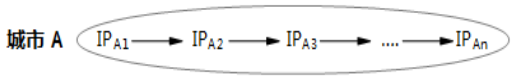

图 2 连续多跳 IP 地址序列

如图 3 所示，即[IP_1，IP_2，…，IP_{n-1}，IP_n]为一条 IP 地址序列。对于同一城市的两个 IP 地址（IP_{A1}, IP_{An}）进行路由选择时，IP_{A1} 放弃了最优路由 IP_{A2}，而选择了与自己不在同一城市的 IP_{B2}，这是不符合路由选择策略，假设不成立。

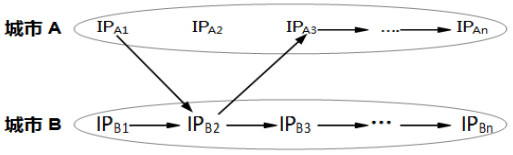

图 3 通信时选择非最佳路由

(2)算法实现

使用双向链表来存储 traceroute 路径信息(数据结构如图 4 所示)，之后使用多跳 IP 地址关联分析算法获取相关 IP 地址序列，最后通过关联分析得到相关联 IP 地址序列中 IP 地址的使用位置信息。

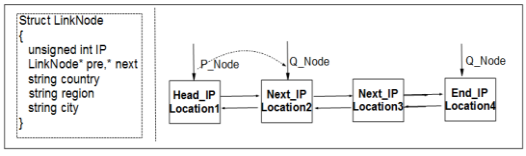

图 4 多跳 IP 地址关联分析算法数据结构

连续多跳 IP 地址关联分析算法的伪代

码如表 1 所示，其目标是为了找到一条 trace 数据中所有首尾节点使用位置相同的子序列。

表 1 多跳 IP 地址关联分析算法

Function:	Get_association_address(L)
Input(L):	A traceroute data sequence
Output(R):	Associate_IP_ sequence

```
Begin
 L←Traceroute link              //input
 P_Node←Head of list
 Q_Node←Tail of list

 while P_Node!=NULL&&P_Node->next!=Q_Node
     if P_Node->Location==Q_Node->Locaiton
         R=[P_Node,…,Q_Node]
         print(R)               //output
         P_Node←Q_Node->next
         Q_Node← Tail of list
     else
         if Q_Node->pre==P_Node
             P_Node=P_Node->next
             Q_Node←Tail of list
         else
             Q_Node←Q_Node->pre
         endif
     end if
 end while
END
```

2.2 相同接入路由的 IP 地址关联分析算法

(1)算法原理

综合考虑多跳 traceroute 路径信息，利用 trace 数据提供的路由拓扑信息作为约束条件结合通信时延约束，提出了相同路由接入的 IP 地址关联分析算法。

如图 5 所示，一般情况下，具有相同接入路由的多个主机 IP 地址的使用位置相同，此结论要求观察对象必须为主机 IP 地址，而 trace 中的 IP 地址不一定为主机 IP 地址，绝大多是是路由器 IP，因此，考虑增加通信时延作为约束，扩展该结论至全体 IP 地址，即具有相同上一跳 IP 地址且与上跳 IP 之间的通信时延小于相关联通信时延阈值 T_m 的多个 IP 地址的使用位置相同。

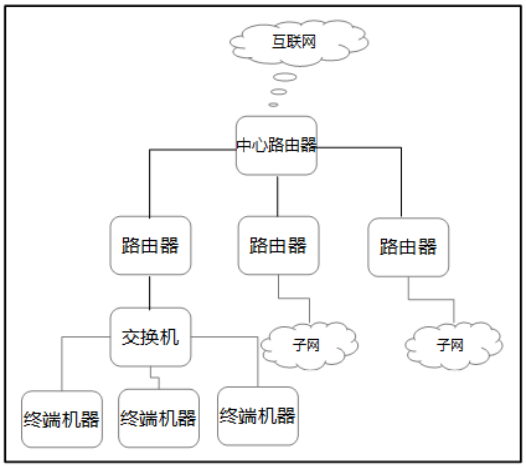

图 5 网络拓扑图

定义序列 A=[IP_1, IP_2, …, IP_{n-1}, IP_n](n≥3)(n≥2)为一条 trace 的连续多跳 IP 地址，把 IP 地址序列转换为如下形式：[IP_1，IP_2，T_1]，[IP_2，IP_3，T_2]，…,[IP_{n-1}，IP_n，T_{n-1}]，我们这种称包含源 IP 和目标 IP 以及他们之间通信时延的单元称为三元组。

把一个 cycle 的 trace 数据均转换为 IP 地址三元组后，存在两个以上的三元组[IP_{n-1}，IP_{1n}，T_1]、[IP_{n-1}，IP_{2n}，T_2]，IP_{1n} 与 IP_{2n} 的上一跳 IP 地址均为 IP_{1n} 且 T_1, T_2 小于通信时延阈值 T_m，那么我们认为 IP_{1n} 和 IP_{2n} 属于同一个 IP 相关地址块，IP 相关地址块内的 IP 地址使用位置相同。

(2)算法实现

相同接入路由的 IP 地址关联分析算法使用哈希链表(如图 6 所示)来获取相关 IP 地址块，整个算法实际上就是构造一个哈希链表，哈希链表的构造过程如下：

- 把 IP 地址序列转换为 IP 三元组[PreIP，DesIP，T]形式，其中 PreIP 为上跳 IP，DesIP 为下跳 IP，T 为他们之间的通信时延。
- 对于三元组[$PreIP_m$， $DesIP_n$,T]根据 $PreIP_m$ 的前 16 位二进制数计算 Hash 键值(0-65535)，遍历相应键值对应的 PreIP 序列如果 $PreIP_m$ 不存在，那么插入节点 $PreIP_m$ 否则舍弃节点 $PreIP_m$
- 遍历 $PreIP_m$ 对应的 DesIP 序列如果节 $DesIP_n$ 不存在而且通信时延 T 小于相关

联通信时延阈值 T_m，把 $DesIP_n$ 插入 $PreIP_m$ 对应的 DesIP 序列，否则舍弃。

一个 cycle 的三元组处理完毕之后，Hash 链表构造完成，每个 PreIP 节点对应的 DesIP 序列都是一个以 PreIP 为接入路由的相关联 IP 地址块，算法的伪代码如表 2 所示。

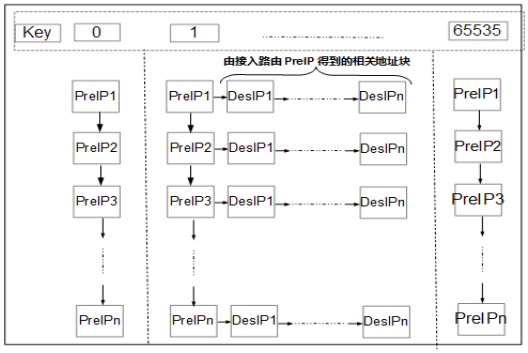

图 6Hash 链表

表 2 相同接入路由 IP 地址相关联分析算法

Function:	Get_association_Block (C)
Input(R):	a cycle traceroute data
Output(R):	Associate_IP_Block

```
Begin
  Hashsize←65536
  PreIP* Array[Hashsize]->next←NULL
  while read PreIP && read DesIP
  /*read from file(a cycle of traceroute data)*/
    index←Getindex(PreIP->IP)
    if PreIP not in Array[index]
        insert PreIP into Array[index]
    endif
    if DesIP not in PreIP's DesIP Link && T<=Tm
        insert DesIP into PreIP's DesIP link
    endif
    Endif
  endwhile
END
```

3 定位算法的有效性分析

本文基于 CAIDA 提供的 traceroute 路径信息提出了两种 IP 地址地理定位算法，但是两种 IP 地址地理定位算法不能保证百分之百的准确性，会存在一定的定位误差。

为了验证本文提出的算法的准确性，我们选取 IPCIS 现有的 IP 地址使用位置库作为第三方的数据源，处理一个 cycle 的数据得到的结果中，与 IP 地址使用位置库相吻合的数量为 M1，相冲突的数量为 M2，那么基于 IPCIS 的 IP 地址使用位置库得到的准确率 P 可以由公式 1 计算得到。

$$P = M1/M2 \quad (1)$$

3.1 连续多跳 IP 地址关联分析算法

在连续多跳 IP 地址地理关联分析算法中，我们忽略了出现环路的情况，如果出现[IP1,IP2,…,IPn,IP1]这样的序列，那么两个 IP1 之间的 IP 地址都会被错误的定位到与 IP1 位置相同的地点。

算法的验证流程如图 7 所示，我们使用连续多跳 IP 地址定位算法处理了 10 个 cycle 的 traceroute 路径信息，并把定位结果与 IPCIS 现有的 IP 地址使用位置库进行对比，得到的结果如图 8 所示，验证率为 85.70%，冲突率为 7.80%，填充率为 6.50%。其中验证率指的是算法的定位结果与数据库相同的部分，冲突率指的是算法的定位结果与数据库不相同的部分，填充率是指 IPCIS 现有的 IP 地址使用位置库中缺失的内容。

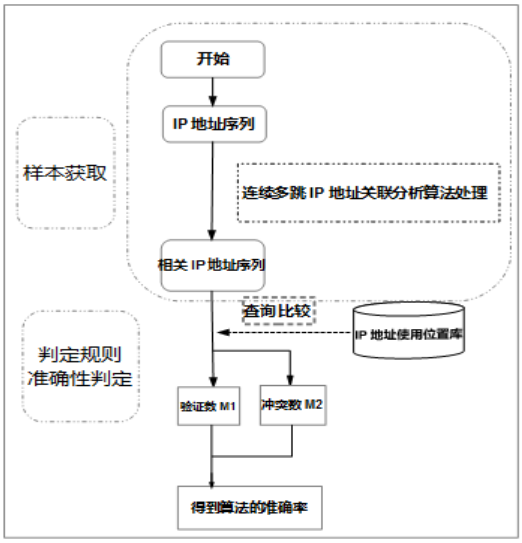

图 7 多跳 IP 地址关联分析算法准确性验证流程

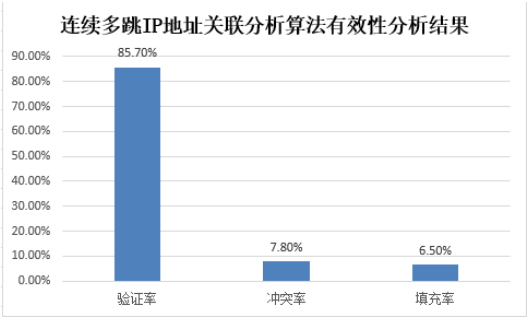

图 8 多跳 IP 地址关联分析算法准确性验证结果

通过处理十个 cycle 的 traceroute 数据的统计结果可以看出，连续多跳的 IP 地址地理定位算法的准确率高于 85.70%，效果比较理想。同时我们建立了基于滑动窗口机制的统计模型，保留处理十个 cycle 的数据所得到的统计结果，并通过求窗口中十个 cycle 的统计结果的平均值作为算法的准确率。当 CAIDA 发布新的 traceroute 路径信息时，分析最新的 trace 数据获取统计结果并加入窗口，并丢弃最老的一个 cycle 的数据，这样就可以动态且长久的观察算法的准确性。

3.2 相同路由接入的 IP 地址关联分析算法

在相同路由接入的 IP 地址关联分析算法中，时延阈值 T_m 是算法的关键，如果时延阈值选取过大，会把使用位置不相同的 IP 地址定位到相同的地方，而且在不同的网络环境中相关联通信时延并不相同，需要进行实验确定通信时延以及对应的算法准确率，从而确定相关联通信阈值。

相同接入路由的 IP 地址关联分析算法的准确性验证流程如图 9 所示，通过相同路由接入的 IP 地址关联分析算法处理 CAIDA 提供的 traceroute 路径信息得到相关联 IP 地址块，并通过关联分析获取组内 IP 地址的使用位置信息，之后以 5 毫秒为时间间隔对相关地址块进行分组，最后以组为单位与 IPCIS 现有的 IP 地址使用位置库进行对比，得到各个时间段相同接入路由的 IP 地址关联分析算法的准确率。

处理十个 cycle 的 trace 数据的统计结果如图 10 所示，相关联通信时延阈值与算法的准确率成反比，时延阈值越小，算法的准确性越高。取相关联通信时延为 5ms 时算法的准确性最高可以达到 78.30%，进一步减小相关联通信时延阈值可以提高算法的准确性，与多跳 IP 地址关联分析算法相同，本算法的准确性验证也需要建立基于滑动窗口机制的统计模型，动态且长久的观察算法的准确率

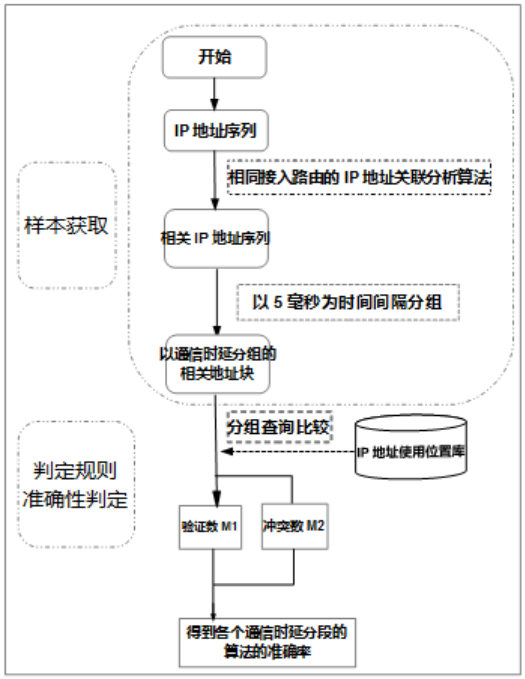

图 9 相同截图路由的 IP 地址关联分析算法准确性验证流程

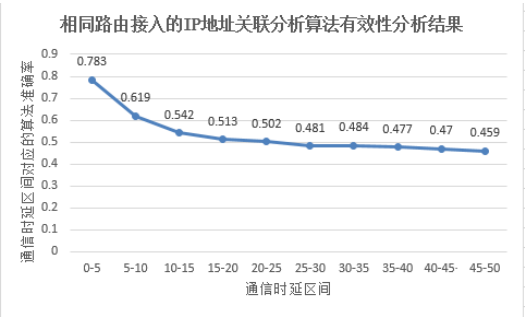

图 10 相同截图路由的 IP 地址关联分析算法准确性验证结果

4 总结与期望

本文利用 CAIDA 提供的 traceroute 路径信息提出了两种 IP 地址地理定位算法，首先详细的介绍了算法的原理和算法的实现，之后利用 IPICS 现有的 IP 地址使用位置库作为第三方数据源验证了算法的准确性，介绍了验证算法准确性的统计模型并得出了统计结果。

本文的不足在于第三方数据源只有IPCIS现有的IP地址使用位置库一个，应该选取更多可靠的数据源验证算法的准确性。算法的准确性验证是在全球互联网环境下得出的，对于相同路由接入的IP地址关联分析算法，相关联通信时延阈值的选取与网络环境有很大的关系，如果在特定的网络环境中进行IP地址地理定位，必须首先通过实验确定相关联通信时延阈值的大小，之后才能使用该算法进行IP地址地理定位。

参考文献

[1] 倪晶晶，IP地址地理定位与管理归属获取方法的研究，东南大学硕士论文，2014.6；

[2] 倪晶晶，龚俭.《一种IP地址地理定位方法的模型》，东南大学校庆报告,2013年

[3] 王占丰，冯径，邢长友，等. IP定位技术的研究[J]. Journal of Software, 2014, 25(7).

[4] 陈鹏.网络实体地理定位IPMapping系统设计与实现[D]. 国防科技大学,2008.

[5] 乐洁，寇晓蕤，罗军勇. 1Traceroute及其在网络拓扑发现中的应用[J]. 微计算机信息, 2005, 21(4): 228-229.

[6] 杨望. CAIDA提供互联网数据共享服务[J]. 中国教育网络, 2008 (5): 27-28.

[7] Jiang Y,Fang B X,Hu M Z, et al. Techniques for determining the geographic location of IP addresses in ISP topology measurement[J]. Journal of Computer Science and Technology, 2005, 20(5): 689-701.

[8] 乐洁，寇晓蕤，罗军勇.Traceroute及其在网络拓扑发现中的应用[J]. 微计算机信息, 2005, 21(4): 228-229.