基于网络结构特征的IP所属区域识别

费高雷* 张亚萌 胡志宇 周 磊 胡光岷
(电子科技大学信息与通信工程学院 成都 611731)

摘 要：现有IP定位技术通过查询IP注册信息数据库或利用测量得到的时延等信息确定IP具体位置，在实际中由于受各种因素的影响，对网络中的大部分IP都无法得到准确、合理的定位结果。为此，该文提出一种基于网络结构特征的IP所属区域识别方法。该方法通过探测节点向待定位的IP发送Traceroute探测包获得两者之间的网络结构特征，并比较待定位节点和已知地理位置节点之间的网络结构特征确定待定位节点所属区域。测试结果表明该文方法和现有的数据库查询的正确率相比有部分提升。

关键词：网络结构特征；特征相似性；IP所属区域识别；网络测量
中图分类号：TP393 **文献标识码**：A **文章编号**：1009-5896(2019)05-1235-08
DOI: 10.11999/JEIT180589

Geographical Location Recognition of IP Based on Network Structure Features

FEI Gaolei ZHANG Yameng HU Zhiyu ZHOU Lei HU Guangmin
(*School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu* 611731, *China*)

Abstract: The existing IP location technology determines the location of IP by querying IP to register information databases or using time-delay information. In fact, due to the influence of various factors, most of the IP in the network can not get accurate and reasonable positioning results. For this reason, a region recognition method of IP is proposed based on network structure features. This method obtains the network topology information between the two nodes by sending the Traceroute detection packet from the detection nodes to the IPs that need to be located Comparing the network structure features between the nodes to be located and the known geographical nodes determines where the nodes located. The actual test shows that this method can achieve better results.

Key words: Network structure features; Feature similarity; Regional identification of IP; Network measurement

1 引言

随着互联网技术的发展与网络应用的不断普及，网络中的设备也越来越丰富，对这些设备的位置进行识别对网络安全、设备识别等具有重要作用。IP地址定位的方法，是在社交网络定位[1–3]、线上定向广告投递以及网络设备定位等定位应用中一种有效的解决方案。

IP定位技术[4]是对网络设备位置进行识别的有效手段，其基本思想是通过多源信息融合得到IP地址的包括地理位置、运营商信息等的基本信息。现有IP定位方法可以分为2大类。第1类是基于特殊设备和基础设施实现的IP定位，第2类是针对普通的互联网中IP地址定位。现有网络测量方法定位IP地址，主要是通过设置地理位置已知的地标IP，网络测量获得时延、跳数、路径等测量信息，最后得到目标IP地址与地标IP地址的地理位置上的关联映射关系。其中典型的方法有GeoTrack[5], GeoPing[5], CBG(Constraint Based Geolocation)[6]技术等。

早期由文献[5]提出了GeoTrack方法，利用Traceroute[7]得到目标IP的可达的最后一跳IP，将其查询出的地理位置近似等于目标IP地理位置。但是该方法准确性高度依赖最后的IP与目的IP的距离以及最后的IP地理位置查询的准确性。同时文献[5]又提出了利用时延信息的GeoPing方法。该方法利用多个探测点测量目标节点的时延信息与地标节

收稿日期：2018-06-13；改回日期：2018-12-17；网络出版：2019-01-07
*通信作者： 费高雷 fgl@usctc.edu.cn
基金项目：国家自然科学基金(61301274, 61471101)，中央高校基本科研业务费(ZYGX2015Z008)
Foundation Items: The National Natural Science Foundation of China (61301274, 61471101), The Fundamental Research Funds for the Central Universities (ZYGX2015Z008)

点的时延信息的相似性。而后，文献[6]利用时延与地理距离存在线性关系提出了基于约束的地理定位方法(Constraint Based Geolocation, CBG)。但是所有基于时延的测量方法都无法回避时延不稳定，以及时延信息与距离之间不存在理想的线性关系[8]对定位准确度的影响。此后，在CBG方法基础上又出现了多种改进算法，文献[9]结合传感器网络定位原理，提出了基于拓扑的定位算法(Topology-Based Geolocation, TBG)。然而该算法也无法避免时延与地理距离非线性关系对定位准确性的影响。文献[10]将IP的定位问题转化为机器学习的分类问题，提出了一种基于贝叶斯估计的定位算法(Naive Bayes IP Geolocation Algorithm, NBIGA)，但是该方法准确性相较于CBG提升不高。

国内对IP定位的研究也较为重视，文献[11]提出了基于路径特征的目标IP区域估计方法，主要是依靠单个路径特征在各个区域的概率，最后将路径的概率和最大的区域定为目标节点的地理位置区域。然而该方法需要大量的地标节点来获取初始的路径特征，缺乏较强的推广性。文献[12]利用目标IP与地标IP重合的最后一跳的跳数来计算目标节点与地标节点的路径相似性来完成IP所属区域估计。然而该方法的缺陷来自于最后重合跳数较大的受到地标节点实际测量情况的限制，以及忽略了目标节点的测量路径中跳数越往后的IP地理位置越接近目标节点的特性以至于导致地理位置结果不准确。

本文研究IP所属区域识别问题，将IP归属到某个区域(如某个省或某个地区，区域的粒度可根据实际需求定义)，这些信息在网络拓扑结构的可视化、基于IP的位置追踪等很多应用中具有十分重要的作用。本文提出一种基网络结构特征的IP所属区域识别方法。该方法的基本思路是通过在网络中部署一系列探测节点，通过探测节点向待定位的IP发送Traceroute探测包获得探测节点与待定位节点之间的路径网络信息。通过比较待定位IP与不同地区地标IP(已知位置的IP)网络结构特征的相似性，来确定待定位IP所属区域。本文使用planet-lab[13]平台部署探测点，分别对中国和美国IP数据进行测试，在省、州级粒度的识别上取得了较好的效果。

2 问题描述

IP所属区域识别的过程是指提供互联网中拥有IP的设备在区域级粒度上的地理位置。由于互联网及其组织设备是分散式自主组成，缺乏可靠的统一规划，所以现有IP地理位置数据库冲突性较高而准确度较低。本文将借助网络结构特征，提高区域级IP定位的准确性。

2.1 问题模型

本文基于网络探测的方法对IP所属区域进行识别，采用目标IP的路径网路结构特征与各个目标区域中的地标IP的网络结构特征进行对比的方法来确认目标IP的所属区域。本文使用的探测模型可抽象为图1所示，包含3种类型的节点：探测节点、地标节点和目标IP节点。

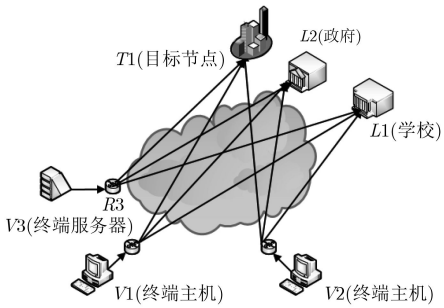


图 1 网络测量模型图

探测节点是指部署在全球范围内，位置固定，且能够向包含地标节点和目标IP在内的探测对象发送探测报文的如主机和服务器的节点。探测节点集合用$V=\{v_1, v_2, \cdots, v_{N_1}\}$表示，其中$N_1$是探测节点的个数。地标节点是指一组地理位置已知的IP地址。本文用$L=\{l_1, l_2, \cdots, l_{N_2}\}$表示地标节点集合，其中$N_2$是地标节点的个数。地标节点可通过对已知地理位置机构主页的IP进行标记获得，例如电子科技大学在成都，其学校主页的服务器也在成都，那么学校主页对应的IP地址就隶属于成都。目标节点是需要进行定位的IP地址，用集合$T=\{t_1, t_2, \cdots, t_{N_3}\}$表示，其中$N_3$是目标节点的个数。本文通过探测节点对地标节点和目标节点进行探测，将所有被探测节点的集合用I表示，$I=L\cup T$；目标区域是指需要被识别的区域，例如中国的各个省或美国的各个州(粒度可根据需求定义)，用$D=\{D_1, D_2, \cdots, D_{N_4}\}$表示目标区域的集合,其中$N_4$是目标区域的个数。

本文的探测模型如图1所示，通过探测节点向已知地理位置的地标节点和未知位置的目标IP发送Traceroute探测包收集数据，然后通过将探测节点到目标IP的路径网络信息的网络结构特征与地标节点的网络结构比较确定目标IP的地理位置。因此首先需要在互联网中部署探测节点，同时通过人工的方式标注一批已知地理位置的地标节点。对于探测节点，本文利用Planet-Lab实验平台，在美国随机选取了7所高校的节点作为探测节点。

IP所属区域的识别问题可概括为式(1)的形式

$$\hat{D}=G(t_i|V,l) \tag{1}$$

其中，$t_i\in T$是需要定位的IP地址；$\hat{D}\in D$是对

T_i进行定位的结果。式(1)表示在已知探测节点和地标节点的条件下，确定t_i的所属区域。

2.2 探测方法

本文利用Traceroute进行网络探测。Traceroute通过设置报文头部的TTL值使得报文到达中间路由器时超时，路由器向源节点反馈ICMP报文，利用该报文地址信息就可以掌握从源到目的路径所有路由器的一个端口的IP地址。

一个探测点测量一个被测节点的一条路径信息是最小单位的网络拓扑信息，一条路径的路径特征也就是最小单位网络结构特征。基于探测节点集合V，获取所有被测节点I的网络拓扑信息集合。首先让探测节点V中任意探测点V_i对各个目标区域的地标节点L和目标节点T发送探测包，获得各目标区域的地标节点L的路径信息和各目标节点T的路径信息。然后让探测节点V中所有探测节点重复以上操作，完成探测节点集合V对获取所有被测节点I的网络拓扑信息集合的收集。之后提取得到网络结构特征，进行相似度计算与比对，最后识别出目标节点所属区域。

为了方便描述，首先定义路径信息和路径特征，即最小单位的网络拓扑信息与最小单位的网络结构特征。对于探测节点v_i和目标节点t_j，用$\mathrm{TF}_{v_i\to t_j}$表示从v_i对t_j进行Traceroute所得的路径信息，$\mathrm{TF}_{v_i\to t_j}=\{\mathrm{IP}_1,\mathrm{IP}_2,\cdots,\mathrm{IP}_N\}$，其中$\mathrm{IP}_i$表示路径中的第$i$跳IP地址。由于某些路由器可能拒绝返回ICMP报文以及有的目标节点不可达，因此测量的路径信息中存在匿名路由器。删去路径中的匿名路由器，并且用255.255.255.0作为掩码与IP地址相与得到该IP的子网地址。定义这些子网地址的集合为$\mathrm{PF}_{v_i\to t_j}$，表示从v_i对t_j进行Traceroute所得的路径特征，$\mathrm{PF}_{v_i\to t_j}=\{\mathrm{IP}_{1*},\mathrm{IP}_{2*},\cdots,\mathrm{IP}_{N*}\}$，其中$\mathrm{IP}_{i*}$表示Traceroute拓扑信息中的第$i$跳IP地址的子网。

其次，分别用$\mathrm{PF}_{v_i\to L}$，$\mathrm{PF}_{v_i\to T}$和$\mathrm{PF}_{v_i\to I}$定义v_i对地标节点集合L、目标节点集合T和所有被探测节点I进行Traceroute所得的路径特征，其中$\mathrm{PF}_{v_i\to L}=\cup_{l_m\in L}\mathrm{PF}_{v_i\to l_m}$，$\mathrm{PF}_{v_i\to T}=\cup_{t_n\in T}\mathrm{PF}_{v_i\to t_n}$和$\mathrm{PF}_{v_i\to I}=\mathrm{PF}_{v_i\to L}\cup\mathrm{PF}_{v_i\to T}$，用$\mathrm{PF}_{V\mapsto T}$表示从探测点集合$V$中所有探测点对所有被探测节点$I$进行Traceroute所得网络结构特征，$\mathrm{PF}_{V\mapsto I}=\cup_{v_i\in V}\mathrm{PF}_{v_i\to I}$。最后基于网络结构特征集合$\mathrm{PF}_{V\mapsto I}$计算网络结构特征相似性，依据相似性最大化原则完成IP所属区域识别。

3 IP所属区域识别方法

传统IP地位方法利用测量节点与被探测节点之间的传输时延估计IP所属的位置，但实际网络中两个节点之间的时延主要由报文排队[14]引起，网络中流量的不稳定性会导致测量所得的时延不稳定，导致传统方法所得的定位结果可能存在很大偏差。针对该问题本文提出了利用测量网络结构特征的方法进行IP所属区域识别，本节将对提出的方法进行详细描述。

3.1 地标节点标记

本文方法的基本思路是将基于共同探测点的目标节点的网络结构特征与地标节点的网络结构特征进行比较，以确定目标IP所属区域。因此首先需要对目标区域的部分IP位置进行标记，即确定地标节点。

针对基于网站的地标节点标记方法，网站服务器的地理位置是相对已知与稳定的。本文通过手动查找一个区域内所有权威的组织机构的官网，获得官网服务器的IP地址，并认为该IP地址应该就在该组织机构所在的地理位置。通过这种线下地理位置与IP地址匹配的方式获取到了地标志信息。但是如果出现网站服务器租用的情况，将会导致此类地标IP失效。

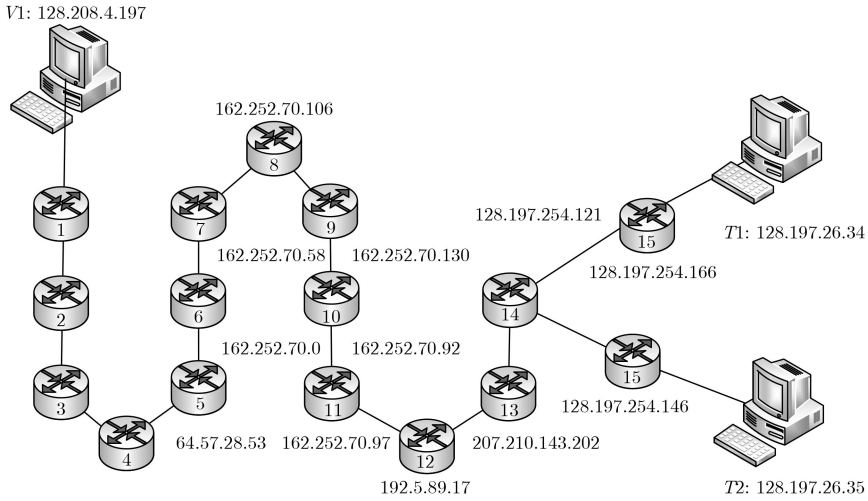
3.2 网络结构特征提取

对于某固定的探测点，开始的几跳IP地址几乎都是相同的，但随着跳数的增加，不同区域的路径的相似性将急剧下降。本文将针对不同目标区域的拓扑信息提取网络结构特征。提取过程可分为如下2种情况：

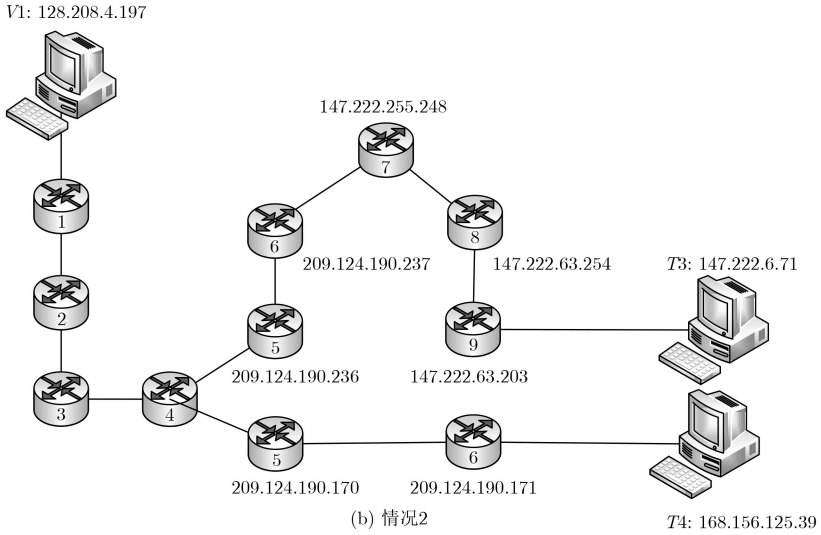
(1) 情况1(目标IP都属于同一地区且同一子网段)：这些IP地址得到的路径拓扑将高度相似，甚至每一跳IP完全一样。以美国波士顿的一对可能处于同一C网的IP地址128.197.26.34和128.197.26.35为例，从华盛顿大学128.208.4.197发出探测，省去前面都一致的4跳，得到的路径信息如图2(a)和表1所示。

(2) 情况2(目标IP属于同一地区但不属于同一子网段)：这是一种更普遍的情况，此时同一目标区域的IP地址的路径相似性将不会像情况1中那样高，但是测量路径中的IP将会有很大可能性出现在相同的C网段中。所以若两个IP的路径中的每跳IP与255.255.255.0相与运算后，得到的子网段的相似性很高，那么它们属于一个区域的概率将会非常大。以美国华盛顿州一对IP地址147.222.6.71和168.156.125.39为示例，从华盛顿大学128.208.4.197发出探测，得到的路径信息如图2(b)和表2所示。

按照2.2小节中路径特征的定义对路径信息进行处理，删除其中匿名IP，最后得到IP地址的最小单位网络特征PF集合。情况1中，128.197.26.34和128.197.26.35对于探测点128.208.4.197的网络结构特征集合如表3所示。情况2中，147.222.6.71和



(a) 情况1



(b) 情况2

图 2 测量拓扑示例

表 1 情况1：华盛顿大学探测波士顿中可能是同一C网的IP对路径信息(可达)

IP	11跳	12跳	13跳	14跳	15跳	16跳
128.197.26.34	162.252.70.97	192.5.89.17	207.210.143.202	128.197.254.121	128.197.254.166	128.197.26.34
128.197.26.35	162.252.70.97	192.5.89.17	207.210.143.202	128.197.254.121	128.197.254.146	128.197.26.35

表 2 情况2：华盛顿大学探测华盛顿州中非同一C网的IP对路径信息(可达)

IP	5跳	6跳	7跳	8跳	9跳	10跳
147.222.6.71	209.124.190.236	209.124.190.237	147.222.255.248	147.222.63.254	147.222.63.203	147.222.6.71
168.156.125.39	209.124.190.170	209.124.190.171	168.156.125.39			

168.156.125.39对于探测点128.208.4.197的网络结构特征集合如表4所示。

3.3 网络结构特征相似性度量

本文方法基于网络结构特征相似性进行IP的所属区域识别，计算网络结构特征相似性是整个方法的核心。对于任意两个IP地址，用$\widehat{S}(\mathrm{IP}_1,\mathrm{IP}_2|V)$表示任意$\mathrm{IP}_1$和$\mathrm{IP}_2$基于探测点集$V$中所有探测点测量得到的网络结构特征相似性。为了获得网络结构特征相似性$\widehat{S}(\mathrm{IP}_1,\mathrm{IP}_2|V)$，需要先计算一个探测点$V_i$测量下$\mathrm{IP}_1$和$\mathrm{IP}_2$间的网络结构特征相似性，用$S(\mathrm{IP}_1,\mathrm{IP}_2|v_i)$表示。

首先计算探测点v_i探测IP_1和IP_2的网络结构特征

表 3 情况1的最小单位网络结构特征

IP	11跳	12跳	13跳	14跳	15跳	16跳
128.197.26.34	162.252.70.*	192.5.89.*	207.210.143.*	128.197.254.*	128.197.254.*	128.197.26.34
128.197.26.35	162.252.70.*	192.5.89.*	207.210.143.*	128.197.254.*	128.197.254.*	128.197.26.35

表 4 情况2的最小单位网络结构特征

IP	5跳	6跳	7跳	8跳	9跳	10跳
147.222.6.71	209.124.190.*	209.124.190.*	147.222.255.*	147.222.63.*	147.222.63.*	147.222.6.71
168.156.125.39	209.124.190.*	209.124.190.*	168.156.125.39			

$\mathrm{PF}_{v_i\to \mathrm{IP}_1}$和$\mathrm{PF}_{v_i\to \mathrm{IP}_2}$中共有的子网集合和不同的子网集合，分别命名为$\mathrm{insec}(\mathrm{IP}_1,\mathrm{IP}_2|v_i)$和$\mathrm{union}(\mathrm{IP}_1,\mathrm{IP}_2|v_i)$，其计算方法分别为

$$\mathrm{insec}(\mathrm{IP}_1,\mathrm{IP}_2|v_i)=\mathrm{PF}_{v_i\to \mathrm{IP}_1}\cap \mathrm{PF}_{v_i\to \mathrm{IP}_2} \tag{2}$$

$$\mathrm{union}(\mathrm{IP}_1,\mathrm{IP}_2|v_i)=\mathrm{PF}_{v_i\to \mathrm{IP}_1}\cup \mathrm{PF}_{v_i\to \mathrm{IP}_2} \tag{3}$$

借鉴文本相似性中的jacard系数定义，基于已经得到的$\mathrm{insec}(\mathrm{IP}_1,\mathrm{IP}_2|v_i)$和$\mathrm{union}(\mathrm{IP}_1,\mathrm{IP}_2|v_i)$，得到$S(\mathrm{IP}_1,\mathrm{IP}_2|v_i)$的计算如式(4)

$$S(\mathrm{IP}_1,\mathrm{IP}_2|v_i)=\frac{\mathrm{num}(\mathrm{insec}(\mathrm{IP}_1,\mathrm{IP}_2|v_i))}{\mathrm{num}(\mathrm{union}(\mathrm{IP}_1,\mathrm{IP}_2|v_i))} \tag{4}$$

然后对于探测点集V中的多个探测点到IP_1和IP_2的网络结构特征相似度$\widehat{S}(\mathrm{IP}_1,\mathrm{IP}_2|V)$，我们可以通过求得所有探测点到$\mathrm{IP}_1$和$\mathrm{IP}_2$的网络结构特征相似度的算术平均值，计算表达式如式(5)

$$\widehat{S}(\mathrm{IP}_1,\mathrm{IP}_2|V)=\sum_{S(\mathrm{IP}_1,\mathrm{IP}_2|v_i)\in S(\mathrm{IP}_1,\mathrm{IP}_2|V)}\frac{S(\mathrm{IP}_1,\mathrm{IP}_2|v_i)}{\mathrm{num}(S(\mathrm{IP}_1,\mathrm{IP}_2|V))} \tag{5}$$

其中，$S(\mathrm{IP}_1,\mathrm{IP}_2|v_i)$是探测点$V_i$测量$\mathrm{IP}_1$和$\mathrm{IP}_2$间的网络结构特征相似性，$S(\mathrm{IP}_1,\mathrm{IP}_2|V)$是探测点集$V$中所有探测点测量$\mathrm{IP}_1$和$\mathrm{IP}_2$间的网络结构特征相似性集合，具体表示为

$$S(\mathrm{IP}_1,\mathrm{IP}_2|V)=\cup_{v_i\in V}S(\mathrm{IP}_1,\mathrm{IP}_2|v_i) \tag{6}$$

由此得到了IP_1和IP_2在探测点集V测量下的网络结构特征相似性为$\widehat{S}(\mathrm{IP}_1,\mathrm{IP}_2|V)$。若$\widehat{S}(\mathrm{IP}_1,\mathrm{IP}_2|V)$数值越大，$\mathrm{IP}_1$和$\mathrm{IP}_2$的网络拓扑结构相似度越高，$\mathrm{IP}_1$和$\mathrm{IP}_2$同属于一个区域的可能性就越高。

3.4 地标节点的聚类

3.3节中，基于一系列探测点向目标IP发送探测包，可以结算得到这些探测点到任意两个IP网络结构特征的相似性，可以根据相似性对目标IP的所属区域进行识别。但在实际中地标节点与目标IP数目都不少，若直接计算两者间的相似性，计算复杂度将很高。

针对上述问题，本文采取了先将地标节点聚类再与目标节点网络结构特征相似度最大匹配的方法。地理位置相近的IP它们的网络结构特征相似度高，聚类后这些IP属于同一聚类的概率较大。若将目标IP先与聚类后的中心比较，再与聚类内的IP计算相似性，确定其具体位置。这样可以有效地降低计算复杂度。

在目前机器学习的研究中，能够使用的聚类方法很多，本文为了方便起见采用k-中心点聚类方法。但其实基于本文定义的IP之间最小单位网络结构的相似性，任何一种聚类方法都可。在实际操作中，我们先输入某个聚类数值k，将地标节点用k-中心点聚类方法处理后，得到k个聚类簇。本文用$C(L|k)$集合表示地标节点集L中所有节点被划分成的k类簇，用式(7)表示

$$C(L|k)=\{L_1,L_2,\cdots,L_k\} \tag{7}$$

其中，$L_i\subset L$，至此我们将所有地标节点分到了k类簇中。

同时获取k类簇对应的中心对象节点集合C_{core}，具体表示如式(8)

$$C_{\mathrm{core}}=\{C_1,C_2,\cdots,C_k\} \tag{8}$$

其中，l_{c_i}是式(7)中L_i对应的中心对象节点。

3.5 IP所属区域确定

将地标节点聚类后将完成对目标节点所属区域的识别，以某目标节点t_i为例，首先获取探测点集合V测量目标节点t_i得到的最小单位网络结构特征$\mathrm{PF}_{V\to t_i}$，以及测量地标节点L的k类簇的中心对象节点的集合C_{core}的特征$\mathrm{PF}_{V\to L_{\mathrm{core}}}$。

为了确定t_i所属的聚类，计算t_i与C_{core}集合中每个节点基于探测点集合V的平均网络结构特征相似性集合$\widehat{S}(t_i,C_{\mathrm{core}}|V)$，其表示如式(9)

$$\widehat{S}(t_i,C_{\mathrm{core}}|V)=\cup_{l_j\in C_{\mathrm{core}}}\widehat{S}(t_i,l_j|v_i) \tag{9}$$

选取$\widehat{S}(t_i,L_{\mathrm{core}}|V)$中最大的相似性数值对应的地标节点命名为$l_{\max}$，以及其对应的簇为$L_{\mathrm{m}}$，具体表示如式(10)

$$l_{\max}=\arg\max(\widehat{S}(t_i,C_{\text{core}}|V)) \tag{10}$$

为了确认t_i在聚类中具体所属的区域，针对已经获得的L_m，将做以下2种讨论：

(1) 若L_m所有节点都属于一个区域，那么t_i也属于同一区域，完成针对t_i的所属区域识别表示如式(11)

$$\hat{D}=G(t_i|V,l_{\max}) \tag{11}$$

(2) 若L_m不是所有节都属于一个区域，那么将计算t_i与L_m集合中所有节点的平均网络结构特征相似性集合，表示如式(12)

$$\widehat{S}(t_i,L_m|V)=\cup_{l_j\in L_m}\widehat{S}(t_i,l_j|V) \tag{12}$$

再选取$\widehat{S}(t_i,L_m|V)$中最大的相似性数值所对应的地标节点，命名为l_{m}'，以l_{m}'的区域信息即为t_i的所属区域信息，具体表示如式(13)，式(14)

$$l_m'=\arg\max(\widehat{S}(t_i,L_m|V)) \tag{13}$$

$$\hat{D}=G(t_i|V,l_{\text{m}}') \tag{14}$$

4 方法测试分析

4.1 测量点部署与数据收集

利用PlanetLab测量平台作为探测点数据平台，在美国随机选择了7个高校探测点，分别是印第安纳大学、Williams大学、密歇根州大学、亚特兰大市Emory大学、内华达大学(里诺校区)、华盛顿大学、明尼苏达大学。

本文的实验验证数据是已知地理位置的IP数据集，但这样的数据集是较难获得。通过3.1小节提出的权威机构的网页IP与机构地理位置的关联来得到对应的地理位置已知的IP数据集。通过人工收集获取高校主页IP以及借助planetlab平台中的高校节点中的IP数据，初步构建了已知地理位置的IP数据集。

最终，收集了2个IP数据集，分别为中国高校主页的数据集1与美国的高校数据集2，详见表5和表6。通过高校主页的服务器在高校本地以及已知的planetlab节点信息，来确认主页IP的地理位置。数据集1为中国高校，总数945个，涵盖7个地区，如图3所示。数据集2为北美高校，总数496个，涵盖9个州，如图4所示。

4.2 结果及分析

针对已收集的实验数据集，为了综合测试本文实验方法的有效性，即测试地标节点数占地标节点占总实验数据集比值不一样的情况下的正确率，本文采取了按比例随机选取地标个数的方法。所以选取了多个不同比例的地标IP比率，以说明地标对于该方法效果的影响。在所有的IP数据集中，多次按一定比例随机抽取IP作为地理位置已知的地标IP，将剩余的IP作为待测IP，最终得到平均正确率。

针对实验数据集1和数据集2，随机按照20%，30%和40%的比例抽取地标节点，对于选为地标节点的IP数据都先采用K-中心分层聚类，再根据网络结构特征相似度最大匹配原则，得到区域级地理

表 5 中国高校IP分布情况

省份	安徽	北京	山东	江苏	河南	浙江	广东	辽宁	总计
数目	113	88	135	155	111	102	131	110	945

表 6 美国高校IP分布情况

州	华盛顿	田纳西	佐治亚	密歇根	马萨	密西西比	加州	弗吉尼亚	伊利	总计
数目	52	31	32	40	91	38	126	68	18	496

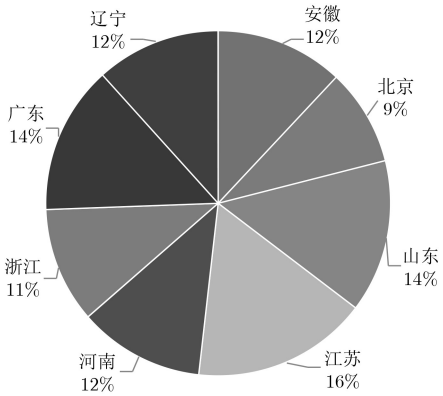


图 3 数据集1：国内IP地址分布情况

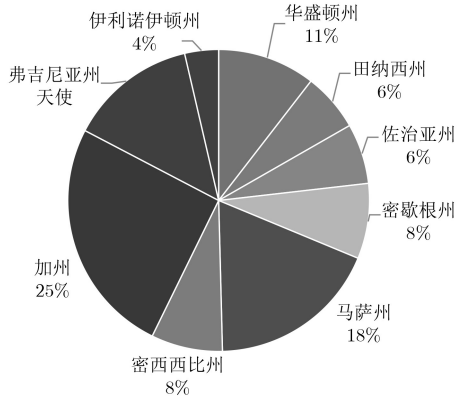


图 4 数据集2：北美IP地址分布情况

位置结果。每种地标的比例重复随机选取1000次，得到了不同比例下的平均正确率。同时，使用综合IP地理数据库——ip2location[15], ipinfo[15], maxmind[15]和AIWEN[16]数据库对待测IP地理位置区域进行查询，得到综合方法的数据库查询的正确率。

针对实验数据集1，综合地理数据库的查询结果与线下结果也有不完全匹配的情况，ip2location, ipinfo, maxmind和AIWEN数据库的正确率为70.73%, 70.39%, 72.54%和79.09%。而本文的实验在随机按照20%, 30%和40%的比例抽取地标节点的情况下的平均正确率分别为80.28%, 84.15%和86.63%。而对于实验数据集2, ip2location, ipinfo, maxmind和AIWEN数据库的正确率为48.18%, 44.95%, 45.36%和63.23%。而本文的实验在随机按照20%, 30%和40%的比例抽取地标节点的情况下的平均正确率分别为63.45%, 64.25%和70.13%。

由本文的实验可知根据网络结构特征相似性可以在很大的程度上判别IP所属区域。但地标IP的选择将会很大程度上影响准确率。原因如下：首先由于多网络多路径存在，处于同一省内的IP可能有不同的网络结构特征的情况。其次地标节点数目不充足，将导致区域的网络结构特征不完备。以及许多目标IP不可达和路径中有大量匿名路由器，将导致路径中有效的信息减少。这些原因都将影响方法的有效性。

5　结束语

本文提出的基于网络特征结构的IP所属区域识别方法，对地标节点分层聚类后，比对目标节点与地标节点的路径，选取最相近的地标节点作为参照，完成区域识别。本文方法显著降低了计算量，在实际应用中有较强的应用价值。但是实际应用中，目标IP不可达、地标节点不足、大量匿名路由器的出现还是会降低方法的效果。如何更好地解决匿名路由器和完备目标节点的问题还待解决。

参 考 文 献

[1] 张少波, BHUIYAN M Z A, 刘琴, 等. 移动社交网络中基于代理转发机制的轨迹隐私保护方法[J]. 电子与信息学报, 2016, 38(9): 2158–2164. doi: 10.11999/JEIT151136.
ZHANG Shaobo, BHUIYAN M Z A, LIU Qin, *et al*. The method of trajectory privacy preserving based on agent forwarding mechanism in mobile social networks[J]. *Journal of Electronics & Information Technology*, 2016, 38(9): 2158–2164. doi: 10.11999/JEIT151136.

[2] 王荣荣. 基于位置的社交网络隐私安全研究[D]. [硕士论文], 华东师范大学, 2016. 15-31.
WANG Rongrong. Research on location based social network privacy security [D]. [Master dissertation], East China Normal University, 2016. 15-31.

[3] 李晴, 叶阿勇, 许力. 社交网络中基于定位欺骗的隐私攻击研究[J]. 信息网络安全, 2017, 1(5): 51–56. doi: 10.3969/j.issn.1671-1122.2017.05.008.
LI Jing, YE Ayong, and XU Li. Research on privacy attack based on location cheating in social network[J]. *Information Network Security*, 2017, 1(5): 51–56. doi: 10.3969/j.issn.1671-1122.2017.05.008.

[4] MUIR J A and OORSCHOT P C V. Internet geolocation: Evasion and counterevasion[J]. *ACM Computing Surveys*, 2009, 42(1): 1–23. doi: 10.1145/1592451.1592455.

[5] PADANABHAN V N and SUBRAMANIAN L. An investigation of geographic mapping techniques for internet hosts[J]. *ACM Sigcomm Computer Communication Review*, 2001, 31(4): 173–185. doi: 10.1145/964723.383073.

[6] 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. doi: 10.1109/TNET.2006.886332.

[7] ZHOU Haifeng, TAN Liansheng, *et al*. Traffic matrix estimation: Advanced—Tomogravity method based on a precise gravity model[J]. *International Journal of Communication Systems*, 2015, 28(10): 1709–1728. doi: 10.1002/dac.2787.

[8] 朱畅华, 裴昌幸, 李建东, 等. 基于线性规划的Internet端到端时延的估计[J]. 电子与信息学报, 2004, 26(3): 446–452.
ZHU Changhua, PEI Changxing, LI Jiandong, *et al*. Linear programming based estimation of internet end-to-end delay[J]. *Journal of Electronics & Information Technology*, 2004, 26(3): 446–452.

[9] KATZBASSET 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, 2006: 71–84. doi: 10.1145/1177080.1177090.

[10] ERIKSSON B, BARFORD P, SOMMERS J, *et al*. A learning-based approach for IP geolocation[C]. Passive and Active Measurement, International Conference, Zurich, Switzerland, 2010: 171–180. doi: 10.1007/978-3-642-12334-4_18.

[11] CHEN Jingning, LIU Fenlin, WANG Tianpeng, *et al*. Towards region-level IP geolocation based on the path feature[C]. International Conference on Advanced Communication Technology IEEE, PyeongChang, South Korea, 2015: 468–471. doi: 10.1109/ICACT.2015.7224839.

[12] REN Lianxing. Method for IP geolocation based on Path Similarity[C]. International Conference on Wireless Communication and Sensor Networks. Boston, USA, 2017:

315–319. doi: 10.2991/icwcsn-16.2017.68.

[13] CHUN B, CULLER D, ROSCOE T, *et al*. PlanetLab: an overlay tested for broad-coverage services[J]. *ACM SIGCOMM Computer Communication Review*, 2003, 33(3): 3–12. doi: 10.1145/956993.956995.

[14] 谢钧, 俞璐, 金凤林. 基于排队时延和丢包率的拥塞控制[J]. 电子与信息学报, 2010, 32(9): 2058–2064. doi: 10.3724/SP.J.1146.2009.01228.
XIE Jun, YU Lu, and JIN Fenglin. Congestion control based on queuing delay and packet Loss probability[J]. *Journal of Electronics & Information Technology*, 2010, 32(9): 2058–2064. doi: 10.3724/SP.J.1146.2009.01228.

[15] SHAVITT Y and ZILBERMAN N. Geolocation Databases Study[J]. *IEEE Journal on Selected Areas in Communications*, 2011, 29(10): 2044–2056. doi: 10.1109/JSAC.2011.111214.

[16] 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. doi: 10.11959/j.issn.2096-109x.2016.00097.
ZHAO Fan, LUO Xiangyang, and LIU Fenlin. Research on cyberspace surveying and mapping technology[J]. *Chinese Journal of Network and Information Security*, 2016, 2(9): 1–11. doi: 10.11959/j.issn.2096-109x.2016.00097.

费高雷：男，1982年生，副教授，研究方向为计算机通信网，网络层析成像.

张亚萌：女，1994年生，硕士生，研究方向为计算机通信网，网络拓扑测量.

胡志宇：男，1992年生，硕士生，研究方向为计算机通信网，网络拓扑可视化.

周 磊：男，1993年生，硕士生，研究方向为计算机通信网，网络行为分析.

胡光岷：男，1966年生，教授，博士生导师，研究方向为计算机通信网、网络行为学和安全.