基于邻近序列的IP地址地理定位方法

郭立轩[1] 卓子寒[2] 何跃鹰[2] 李 强[3] 李舟军[1]

(北京航空航天大学计算机学院 北京 100191)[1] (国家计算机网络应急技术处理协调中心 北京 100029)[2]

(浪潮(北京)电子信息产业有限公司高效能服务器和存储技术国家重点实验室 北京 100029)[3]

摘 要 IP地址地理定位旨在准确地确定给定的IP地址的物理空间位置,通常采用基于测量的技术或者基于数据分析的技术。现有的基于数据分析的IP地址地理定位技术,对IP地址之间的关系考虑较少。考虑到IP地址的聚集特性,提出了一种基于邻近序列的IP地址地理定位方法。首先计算IP地址的邻近序列,并将其转化为对应的经纬度序列,然后建立模型并求解。以IP地址定位库和含有GPS信息的移动流量数据为原始数据,对该方法进行了实验验证。实验结果表明,通过邻近IP序列确实可以确定IP地址的物理空间位置,平均定位误差在20~30km,实现了区县一级的定位。该方法给IP地址地理定位问题提供了新的解决方案,同时该方法也可以与其他基于测量或者基于数据分析的方法相结合,以获得更优的结果。

关键词 IP地址地理定位,邻近序列,数据分析

中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.01.035

IP Geolocation Method Based on Neighbor Sequence

GUO Li-xuan[1] ZHUO Zi-han[2] HE Yue-ying[2] LI Qiang[3] LI Zhou-jun[1]

(School of Computer Science and Engineering, Beihang University, Beijing 100191, China)[1]

(National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China)[2]

(State Key Laboratory of High-end Server & Storage Technology, Inspur, Beijing 100029, China)[2]

Abstract IP geolocation is intended to accurately determine the physical space location of a given IP address, usually based on measurement technology or data analysis. The existing approaches based on data analysis have less consideration of the relationship between IP address. Taking into account the aggregation of IP address, this paper proposed an IP geolocation approach based on neighbor sequence. First, the approach calculates the neighbor sequence of IP address, converts it to the corresponding sequence of latitude and longitude, and then models it based on the sequence and solves. This approach was experimentally verified by using IP address location library and mobile traffic data with GPS information as original data. Result shows that neighbor sequence can determine the physical space location of IP address, and mean error is between 20km and 30km, which means this approach has achieved county level geolocation. This approach provides a new solution and a new idea for the IP geolocation problem, and it can be combined with other approaches based on measurement or based on data analysis to obtain better result.

Keywords IP geolocation, Neighbour sequence, Data analysis

1 引言

当前,针对规模日益扩大的互联网用户群,大量的互联网服务均需要获取用户的位置信息,如何准确地确定用户的地理位置已成为互联网应用中非常重要的课题。在用户设备不提供GPS等定位技术的条件下,使用IP地址的地理定位技术是确定用户位置的首选方法。

现有的IP地址地理定位技术可以分为两类:1)基于测量的方法,其原理是通过traceroute获得网络拓扑、路由跳数或者对目标进行时延探测等,从而推断出IP地址的地理位置。此类方法可以进一步细分为基于空间理论的方法与基于概率估计的方法[1],主要包括GeoPing[2],TBG[3],Posit[4]等。2)基于数据分析的方法,此类方法采用非测量技术,通过对WHOIS数据库、网页等相关数据的处理和分析,得出IP地址对应的地理位置。

在基于数据分析的IP地址地理定位技术中,按照使用数据的结构化程度,可将数据分为结构化数据、半结构化数据与非结构化数据。注册和备案记录是一种常见的包含IP地址(段)及其对应地理信息的结构化数据[5],例如WHOIS数据库、DNS LOC记录、DNS名称等,这些数据可以用于推测IP

到稿日期:2016-12-20 返修日期:2017-02-24 本文受国家自然科学基金项目(61170189,61370126,U1636211),国家863计划项目(2015AA016004),北京成像技术高精尖创新中心项目(BAICIT-2016001)资助。

郭立轩(1991—),男,硕士,主要研究方向为数据挖掘;卓子寒(1987—),男,博士,工程师,主要研究方向为网络安全、大数据分析;何跃鹰(1975—),男,高级工程师,主要研究方向为网络安全、大数据分析,E-mail:hyy@cert.edu.cn(通信作者);李 强(1982—),男,博士,主要研究方向为存储系统、分布式系统;李舟军(1963—),博士,教授,博士生导师,主要研究方向为网络安全、数据挖掘与人工智能。

地址的地理位置。Wang 等[1]将基于注册记录的地理定位方法分为 3 类：1）直接查询 WHOIS 数据库；2）通过测量主机名并结合数据库信息进行推断；3）利用网络结构和数据库信息来推断。基于注册记录的地理定位方法可能出现较大偏差，因为某些大型实体机构可能拥有分散在不同地方的多台服务器，但域名均注册为同一地址[4]。在半结构化数据中，地理信息可能作为数据中的某个对象属性存在。Dan 等[5]从移动设备搜索引擎日志中提取 IP 地址和 GPS 坐标，构建了迄今为止最大的基准真实值（Ground Truth）集合，拥有 840 万条 IP 地址的地理位置记录。Web 页面是最常见的非结构化数据，可以使用文本挖掘[15]、数据挖掘和统计学[16]等方法从海量 Web 页面中抽取地理信息，并与特定 IP 地址（段）关联，实现 IP 地址的定位。Guo 等[6]提出了 Structon 方法，将 Web 挖掘、推断和 IP traceroute 等方法结合起来，实现了较准确的 IP 地址定位。Backstrom 等[7]提出了一种基于社交图谱的方法，利用用户的好友位置来确定用户的位置，取得了较好的定位效果。

Guo 等[6]发现，网络管理员倾向于把连续的 IP 地址段分配给相同的区域或相近的地点，这意味着连续的 IP 地址在地理分布上倾向于聚集在一起，即连续的 IP 地址在地理上倾向于相邻。本文称该假设为 IP 地址的聚集特性。基于该假设，若可以获知一个未知定位的 IP 地址相邻的其他 IP 地址的地理位置信息，基于这些相邻 IP 地址的地理位置，可以推断出该目标 IP 地址的物理空间定位范围。

根据上述基本假设，本文提出了一种基于邻近序列的 IP 地址地理定位方法。首先计算 IP 地址的邻近序列，并将其转化为对应的经纬度序列，再建立模型并求解。为此，本文采用了以下两类数据对该方法进行实验验证：1）标定好的开源 IP 定位数据库[8]，该数据库中包含 IP 地址段及其对应的行政区划地址；2）中国教育科研网中采样的中国某省实验数据，该数据为移动应用的网络流量，流量中包含有 IP 地址以及经纬度，并构成对应关系。实验结果表明，通过邻近 IP 序列确实可以确定 IP 地址的物理空间位置，平均定位误差在 20～30km，实现了区县一级的定位。该方法为 IP 地址地理定位问题提供了新的解决方案，同时其还可以与其他基于测量或者基于数据分析的方法相结合，以获得更优的结果。

本文提出的基于邻近序列的 IP 地址地理定位方法的基本原理如图 1 所示。IP 地址有 IPv4 和 IPv6 两个版本[9]，由于 IPv6 地址目前并未广泛使用，因此本文仅讨论 IPv4 地址的地理定位技术。

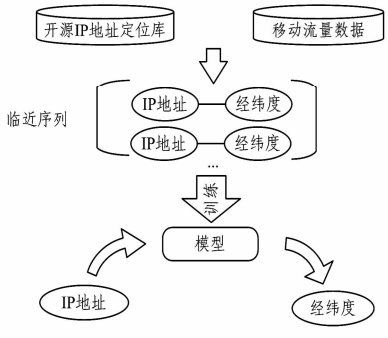

图 1　定位方法的基本原理

Fig. 1　Basic principle of geolocation

2　相关定义

IP 地址[10]的常见表示形式为以英文“·”分隔的 4 个十进制数字，其实质是一个 32 位的二进制整数，可以表示为 $IP=b_{32}b_{31}\cdots b_1$，其中 $b_i(1\leqslant i\leqslant 32)$ 为二进制数 0 或者 1，可以进行数学运算。

为了度量 IP 地址的邻近程度，本文需要定义两个 IP 地址的距离。IP 距离的定义应满足如下两个条件：

1）两个相同 IP 地址之间的距离为 0。

2）不同 IP 地址之间，若相同前缀越长，则距离越近。该性质保证了同一个网络中的 IP 地址距离总是小于不同网络之间的 IP 地址距离。

定义 1（IP 距离）　IP 距离定义为：

$$Dist(p_1,p_2)=\begin{cases}0, & p_1=p_2\\ Length(p_1\oplus p_2), & \text{否则}\end{cases}\tag{1}$$

其中，$Length(x)$ 表示无前 0 的二进制数 x 的位数。异或运算的结果中，两个 IP 地址中相同的比特位变为 0，不同的比特位变为 1，即两个 IP 地址的相同前缀在结果中是一串 0，并且位于数字的高位。因而 *Length* 函数可以衡量两个 IP 地址中最长的共同前缀的长度，并且其最长的共同前缀越长，则函数值越小。因此，定义 1 满足 IP 距离的定义所需的两个条件。

例如，对于表 1 所列的 3 个 IP 地址，$Dist(A,B)=16$，$Dist(A,C)=24$。这表明 AB 的距离要小于 AC 的距离，这与 IP 地址的十进制给我们的感觉是一致的。

表 1　IP 地址及二进制表示举例

Table 1　Examples of IP address and corresponding binary format

编号	IP 地址	二进制形式
A	203.154.16.34	11001011 10011010 00010000 00100010
B	203.154.154.29	11001011 10011010 10011010 00011101
C	203.18.16.34	11001011 00010010 00010000 00100010

定义 2（邻近 IP 集合和邻近 IP 地址）　对于指定的 IP 地址 p，任何与 p 的距离为 n 的 IP 地址的集合记为 $SE_{p,n}$，称为 p 的 n-邻近 IP 集合。$SE_{p,n}$ 的元素称为 p 的 n-邻近 IP 地址，记为 $N_{p,n}$。对于一个给定的 IP 地址 p，其邻近 IP 集合 $SE_{p,n}$ 是确定的。由 IP 距离的定义可知，$SE_{p,n}$ 中的任意元素与 p 的异或的二进制位数为 n。这说明这些元素的高位与 p 相同，其第 n 位 b_n 与 p 不同，而其最低的 $n-1$ 位的值 $b_i(1\leqslant i\leqslant n-1)$ 可以是任意的 0 或 1。因此 $SE_{p,n}$ 可以使用 IP 地址的闭区间 $[N1_{p,n},N2_{p,n}]$ 来表示，其中 $N1_{p,n}=b_{32}\cdots b_n0\cdots0$，$N2_{p,n}=b_{32}\cdots b_n1\cdots1$。

定义 3（邻近序列）　给定 IP 地址 p、最小距离 n_s 和最大距离 n_e，IP 地址的序列 $N_{p,n_s},N_{p,n_s+1},\cdots,N_{p,n_e}$ 称为 p 的邻近序列。

3　定位算法描述

IP 地址定位的目标是，给定一个 IP 地址，可以快速查询其所在物理空间的位置信息，包括精确的经纬度坐标及粗粒度的行政区划地址[11]。

本文根据 IP 地址的聚集特性，提出了一种基于邻近序列的 IP 地址地理定位方法。该方法根据多个邻近 IP 的经纬度推断目标 IP 地址的定位范围。为此，首先需要有一个由 IP

地址到经纬度的映射关系表作为算法的基准输入。我们拥有两种不同的原始数据:1)开源 IP 地址定位数据库,每个记录的格式为形如(起始 IP 地址,终止 IP 地址,国家,省,市,区县)的六元组,其中区县可能为空;2)从中国教育科研网提取的中国某省的移动应用网络流量,可从数据中解析得到形如(IP 地址,经度,纬度)的三元组。利用百度的地理编码 API[12],将 IP 地址定位数据库的行政地理地址编码为经纬度,形成(起始 IP 地址,终止 IP 地址,经度,纬度)的四元组。对于移动应用网络流量,根据每条记录的 IP 地址进行聚合,最终每个 IP 地址得到一个经纬度集合,计算这些经纬度之间的最大距离,若最大距离在 20km 以上,则认为这个 IP 地址动态分配且分配地址较为分散,本文不予考虑。舍弃最大距离大于 20km 的数据之后,以这些经纬度的集合中心为最终的经纬度,并记录聚集数量。图 2 给出了算法的流程图,表 2 列出了算法相关的参数及其说明。

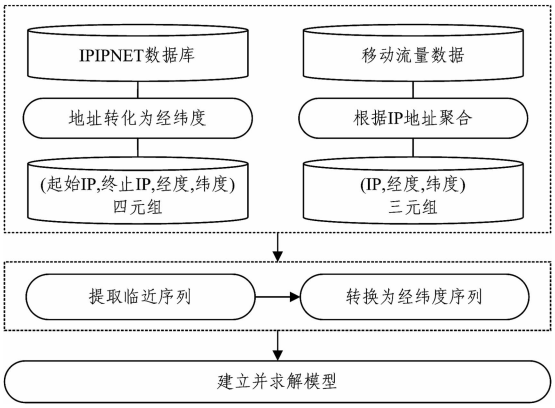

图 2 算法流程图

Fig. 2 Flowchart of algorithm

表 2 参数说明

Table 2 Description of parameters

参数	说明
n_s,n_e	分别表示所取邻近 IP 与目标 IP 的最小距离和最大距离
$N_{p,n}$	与 IP 地址 p 的距离为 n 的 IP 地址
S_a,S_b	分别表示处理后的移动应用网络流量数据集合和 IPIP-NET 数据集合
S_{train},S_{test}	IP 地址训练集和测试集
T	邻近序列对应的经纬度序列的集合

3.1 构建邻近序列对应的经纬度序列

根据定义 2 及式(1),对于目标 IP 地址 p 和 IP 距离 n,邻近集合 $SE_{p,n}$ 的计算公式为:

$$SE_{p,n}=\begin{cases}[p,p], & n=0\\ [N1_{p,n},N2_{p,n}], & \text{否则}\end{cases} \tag{2}$$

对于 S_a 中的每个 IP 地址 p,计算 n_e-n_s+1 个邻近 IP 集合$[N1_{p,n},N2_{p,n}]$,$n_s\leqslant n\leqslant n_e$。然后遍历 S_a,选取在邻近集合内的 IP 地址 $N_{p,n}$,若不存在满足条件的 IP 地址,再遍历集合 S_b,寻找与邻近集合交集最长的 IP 地址段。最后将这些邻近 IP 地址(段)映射为相应的经纬度。算法描述如下所示。

算法 1 计算邻近序列的经纬度

输入:S_a,S_b,n_s,n_e

输出:临近序列对应的经纬度序列的集合 T

```
BEGIN
  T= Ø
  FOR p∈Sa DO
    R=Ø
    FOR n←ns TO ne DO
      计算 p 的距离为 n 的临近 IP 集合 [N1p,n,N2p,n]
      选取 Sa 中在[N1p,n,N2p,n]范围内的 IP 集合 M
      IF Size(M)=1 THEN
        R[n]←M 的元素的经纬度
      ELSE IF Size(M)≠0 THEN
        R[n]←M 的聚集数量最大的元素的经纬度
      ELSE
        选取 Sb 中与[N1p,n,N2p,n]交集最长的 IP 段 L
        R[n]←L 的经纬度
      END
    END
    T[p]←R
  END
  RETURN T
END
```

3.2 推断目标 IP 地址的经纬度

基于以上过程,IP 地址定位问题可以转化为利用邻近序列对应的经纬度序列来推断目标 IP 地址的经纬度。以经纬度序列 X^i 作为输入,目标 IP 地址的经纬度 $\hat{Y}^i$ 作为输出,则有:

$$\hat{Y}^i=f(X^i) \tag{3}$$

实际任务是推测函数 f。假设 Y 和 X 是线性关系,则有:

$$\hat{Y}^i=w^{\mathrm{T}}X^i \tag{4}$$

其中,$\hat{Y}^i$ 是一个 2 维的行向量,w 是一个 n_e-n_s+1 维的列向量,而 X^i 则是 $(n_e-n_s+1)\times 2$ 的矩阵。为了学习该函数,定义损失函数为:

$$Loss(X)=\sum_{i=1}^{m}(\hat{Y}^i-Y^i)\cdot(\hat{Y}^i-Y^i) \tag{5}$$

其中,Y^i 为第 i 组数据 X^i 的真实值。而

$$\hat{w}=\arg\min_{w} Loss(X) \tag{6}$$

为所求模型。本文选取随机梯度下降[13]的方法求解该模型。

算法的主要参数为邻近序列的最小距离 n_s 和最大距离 n_e。当 IP 距离达到一个较大值后,两者之间的地理位置不存在显著相关性。同时,当 IP 距离足够小时,可以认定两者在同一个小的局域网内。但是确定目标 IP 地址的地理位置时不能仅考虑距离足够小的邻近序列,因为彼此之间距离非常小的 IP 地址对在数据中不会大量出现,依赖这种偶尔出现的数据会导致模型的过拟合。

4 实验结果与分析

根据已有数据,本文选取不同的 n_s 和 n_e 进行对比。考虑到较短的邻近序列更容易受到噪声数据的影响,本文选取的邻近序列的长度为 8~16。表 3 列出了不同参数下算法的准确度。

表 3　不同参数下算法的性能比较

Table 3　Performance comparison of algorithms with different parameters

n_s	n_e	平均误差/km	中位数/km	方差	最小误差/km	最大误差/km
8	16	26.29	16.54	817.18	0.33	214.58
8	20	26.67	14.50	931.25	0.37	250.30
8	24	28.97	17.05	1193.51	0.44	271.90
9	16	27.88	18.86	1049.57	0.09	250.22
9	20	26.61	14.98	1045.82	0.13	243.36
9	24	28.41	18.84	1055.95	0.51	259.11
10	17	28.72	14.31	1182.55	0.23	263.58
10	20	27.47	17.45	877.85	0.57	228.62
10	24	30.19	17.46	1111.28	0.17	237.11
11	18	28.30	19.10	954.34	0.09	225.86
11	20	31.02	23.06	959.17	0.20	231.94
11	24	29.36	21.05	942.20	0.08	228.86
12	19	32.85	24.05	1079.94	0.68	223.11
12	20	32.80	23.93	1077.66	0.60	227.44
12	24	32.30	21.71	1147.81	0.14	232.60

由表 3 可知，n_s 和 n_e 会影响算法的准确度，当 n_s 和 n_e 分别取为 8 和 20 时，算法具有最优结果，平均误差为 26.67km，误差中位数为 14.50km。同时可以看出，随着 n_s 的增加，算法的平均误差和误差中位数均呈现出增加的趋势，这意味着算法的准确度表现出了下降的趋势。这是因为，随着 IP 距离的逐渐增大，IP 地址之间的地理位置的相关性也在逐渐减弱。

由图 3 和图 4 可以看出，不同参数下，误差的分布都呈现出误差距离增大、误差数目快速震荡下降的规律。此外，在不同的参数下，虽然误差距离的增大，误差数目下降的速率不同，但是算法的准确度并没有数量级上的差异，这表明在所选的 n_s 和 n_e 范围内，邻近序列的物理空间位置与目标 IP 地址的物理空间位置具有较大相关性。不过仍然可以看出，在参数 $n_s=8, n_e=20$ 时，相比于 $n_s=8, n_e=24$ 时，在误差距离较小时相同距离之下的误差数目更多，而在误差距离较大时相同距离之下的误差数目则更少，这两点使得 $n_s=8, n_e=20$ 时的算法的平均误差和误差中位数更小，进而使得算法具有更高的准确度。

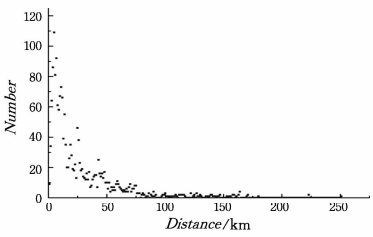

图 3　当 n_s 和 n_e 分别为 8 和 20 时算法的误差分布

Fig. 3　Error distribution of algorithm when n_s is 8 and n_e is 20

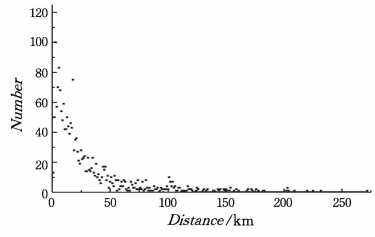

图 4　当 n_s 和 n_e 分别为 8 和 24 时算法的误差分布

Fig. 4　Error distribution of algorithm when n_s is 8 and n_e is 24

由表 3 及图 3、图 4 可以观察到，n_s 和 n_e 会影响算法的准确度。在实际使用中，由于数据的不同，数据具体的分布情况也必然不同，因此应该在充分研究数据的分布情况的基础上进行实验，以确定 n_s 和 n_e 的具体选值，从而获得最优的定位性能。

算法的平均误差为 20～30km，误差中位数在 20km 左右，基本实现了区县一级的定位。表 4 列出了算法的定位结果及其他 IP 地址库给出的结果。可以看出，算法给出的结果具有更细的粒度，精确程度则与各数据库相当。

表 4　定位结果

Table 4　Geoloation results

IP 地址	算法结果	IPIPNET	百度	淘宝
60.190.128.4	嘉兴海宁市	嘉兴海宁市	嘉兴海宁市	嘉兴
61.234.187.0	温州龙湾区	温州	浙江	温州
183.128.1.44	杭州上城区	杭州	杭州	杭州
183.246.99.225	金华金东区	金华金东区	金华东阳区	金华
183.248.96.65	金华婺城区	金华婺城区	金华婺城区	金华
211.138.130.122	杭州萧山区	杭州	杭州萧山区	杭州
211.140.4.79	杭州萧山区	杭州滨江区	杭州西湖区	杭州
220.205.170.211	杭州江干区	杭州	浙江	杭州

结束语　关于 IP 地址的地理定位问题，现有的基于数据分析的方法多关注 IP 地址的外延属性，如行政区划地址、域名注册记录等，而较少关注 IP 地址之间的关系。本文则专注于 IP 地址之间的关系，根据 IP 地址的聚集特性，提出了基于邻近序列的 IP 地址地理定位方法。首先选取 IP 地址的邻近序列，并转换为对应的经纬度序列，然后采用线性模型对经纬度序列进行拟合，选取使损失函数最小的参数作为模型。实验表明，通过选择适当的参数，该方法可以给出较精确的 IP 地址的物理空间位置。受限于实验的数据规模，本文实验的定位误差平均为 20～30km。如果数据量达到百万级规模，预计该方法可以将定位的平均误差限制在 10km 以内。

本方法从 IP 地址之间的关系出发，考虑了 IP 地址的聚集特性，为 IP 地址地理定位问题提供了新的思路，是现有 IP 地址地理定位技术的有力补充。基于数据分析方法，如果在 IP 地址外延属性的基础上引入 IP 地址之间的关系，能够增加大量信息，获取更多的特征，则模型将拥有更多参数，表达能力也将增强，可以期望获得更优的结果。精准的 IP 地址定位结果可以有效地支撑相关互联网产业，如网络安全、广告投放等。本工作是基于 IPv4 地址进行的，IPv6 地址目前虽然应用得并不广泛，却增长迅速，第 37 次中国互联网发展状况报告[14]显示，2014 年到 2015 年，中国 IPv6 地址数量的年增长率达到 9.6%，下一步的工作将致力于 IPv6 地址的地理定位。

参 考 文 献

[1] WANG Z F, FENG J, XING C Y, et al. Research on the IP geolocation technology[J]. Journal of Software, 2014, 25(7): 1527-1540. (in Chinese)
王占丰，冯径，邢长友，等. IP 定位技术的研究[J]. 软件学报，2014，25(7)：1527-1540.

[2] 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.

[3] 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. ACM, 2006: 71-84.

[4] ERIKSSON B, BARFORD P, MAGGS B, et al. Posit: a lightweight approach for IP geolocation [J]. ACM SIGMETRICS Performance Evaluation Review, 2012, 40(2): 2-11.

[5] DAN O, PARIKH V, DAVISON B D. Improving IP Geolocation using Query Logs[C]// Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. ACM, 2016: 347-356.

[6] GUO C X, LIU Y X, SHEN W C, et al. Mining the web and the internet for accurate ip address geolocations [C] // INFOCOM 2009, IEEE. IEEE, 2009: 2841-2845.

[7] BACKSTROM L, SUN E, MARLOW C. Find me if you can: improving geographical prediction with social and spatial proximity [C] // Proceedings of the 19th International Conference on World Wide Web. ACM, 2010: 61-70.

[8] SHAVITT Y, ZILBERMAN N. A geolocation databases study [J]. IEEE Journal on Selected Areas in Communications, 2011, 29(10): 2044-2056.

[9] AL-GADI G, BABIKER A A, MUSTAFA N, et al. Comparison between IPv4 and IPv6 using OPNET simulator[J]. IOSR Journal of Engineering (IOSRJEN), 2014, 4(8): 44-50.

[10] ANDREW S T, DAVID J. Wetherall, Computer Networks(Fifth Edition)[OL]. https://en.wikipedia.org/wiki/IP_address.

[11] ERIKSSON B, BARFORD P, SOMMERS J, et al. A learning-based approach for IP geolocation [C] // International Conference on Passive and Active Network Measurement. Springer Berlin Heidelberg, 2010: 171-180.

[12] http://lbsyun. baidu. com/index.php?title=webapi/guide/web-service-geocoding.

[13] ZHANG T. Solving large scale linear prediction problems using stochastic gradient descent algorithms[C]// Proceedings of the Twenty-first International Conference on Machine Learning. ACM, 2004: 116.

[14] CNNIC. 第 37 次中国互联网络发展状况统计报告[EB/OL]. https://www. cnnic. net. cn.

[15] ZHANG H, LI Z, CHEN Y, et al. Exploit latent dirichlet allocation for one-class collaborative filtering[C]// Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. ACM, 2014: 1991-1994.

[16] WANG S Z, HE L F, STENNETH L, et al. Citywide traffic congestion estimation with social media [C] // Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2015: 34.

(上接第 182 页)

[2] MUELLER J, STUMME G. Gender Inference using Statistical Name Characteristics in Twitter [OL]. https:/arxiv. org/pdf/1606. 05467v2. pdf.

[3] MARQUARDT J, FARNADI G, VASUDEVAN G, et al. Age and Gender Identification in Social Media [C] // Proceedings of CLEF 2014 Evaluation Labs. 2014: 1129-1136.

[4] WU L, GE Y, LIU Q, et al. Modeling users' preferences and social links in Social Networking Services: a joint-evolving perspective[C]// Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. 2016: 279-286.

[5] MA H, CAO H, YANG Q, et al. A habit mining approach for discovering similar mobile users [C] // Proceedings of the 21st International Conference on World Wide Web. ACM, 2012: 231-240.

[6] ZHU H, CHEN E, XIONG H, et al. Mining mobile user preferences for personalized context-aware recommendation[J]. ACM Transactions on Intelligent Systems and Technology, 2015, 5(4): 1-27.

[7] ZHANG K. Mobile Phone User Profile in Large Data Platform [J]. Information and Communications, 2014(2): 266-267. (in Chinese)
张慷. 手机用户画像在大数据平台的实现方案[J]. 信息通信, 2014(2): 266-267.

[8] HUANG W B, XU S C, WU J H, et al. The Profile Construction of the Mobile User[J]. Journal of Modern Information, 2016, 36(10): 54-61. (in Chinese)
黄文彬, 徐山川, 吴家辉, 等. 移动用户画像构建研究[J]. 现代情报, 2016, 36(10): 54-61.

[9] MA L, TAO L T, XIE J K. Customer demands management based on tho customer portraits[J]. Power Demand Side Management, 2016(A01): 98-100. (in Chinese)
马亮, 陶利涛, 谢骏凯. 基于客户画像的客户诉求管理[J]. 电力需求侧管理, 2016 (A01): 98-100.

[10] YAN Y P, WU G C. Customer Outage Sensitivity based on the Technology of Data Mining Research and Application[J]. New Technology and New Process, 2015(9): 89-93. (in Chinese)
严宇平, 吴广财. 基于数据挖掘技术的客户停电敏感度研究与应用[J]. 新技术新工艺, 2015(9): 89-93.

[11] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system[C]// Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 785-794.

[12] DIERCKX, GOEDELE. Logistic Regression Model[J]. Encyclopedia of Actuarial Science, 2009, 39(2): 261-291.

[13] HEARST M A, DUMAIS S T, OSMAN E, et al. Support vector machines[J]. IEEE Intelligent Systems, 1998, 13(4): 18-28.

[14] QUINLAN J R. C4. 5: programs for machine learning[M]. Elsevier, 2014.

[15] BREIMAN L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.

[16] FRIEDMAN J H. Greedy Function Approximation: A Gradient Boosting Machine[J]. Annals of Statistics, 2000, 29(5): 1189-1232.

[17] BREIMAN L. Stacked regressions[J]. Machine Learning, 1996, 24(1): 49-64.

[18] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.