文章编号：2095-2163(2020)07-0087-06 中图分类号：TP393.0 文献标志码：A

基于地理信息的自治域级互联网拓扑可视化研究

安宇昊，张 宇
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：本文提出了一种基于地理信息的互联网内自治域拓扑可视化的方法，相比传统的可视化方式，能够清晰的看出互联网内自治域分布与国家地区之间的联系。使用密度聚类的方法计算自治域的地理中心点，通过地图投影等手段可以将自治域的地理信息展示在地图上。本文还将自治域的拓扑位置与地理信息结合，绘制出能体现互联内资源分布与地理和国家之间关系的拓扑图。

关键词：自治域；拓扑可视化；地理信息

Research on Autonomous Domain-level Internet Topology Visualization Based on Geographic Information

AN Yuhao, ZHANG Yu
(School of computer science and technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 This paper presents a method for visualizing the topology of Autonomous Domains in the Internet based on geographic information. Compared with traditional visualization methods, we can clearly see the relationship between the distribution of Autonomous Domains in the Internet and countries and regions. The density clustering method is used to calculate the geographic center point of the Autonomous Domain, and the geographic information of the autonomous domain can be displayed on the map by means such as map projection. This paper also combines the topological location of the Autonomous Domain with geographic information to draw a topological map that reflects the relationship between the distribution of resources within the interconnection and the geography and country.

【Key words】 Autonomous Domain; network visualization; geographic information

0 引 言

自治域是互联网路由架构的关键元素之一。大范围的路由，比如跨国路由不可避免的要使用域际路由，自治域在互联网路由中扮演着重要角色，因而对自治域级别的互联网拓扑可视化将便于对自治域进行了解和掌握。同时，全球互联网还存在自治域分布不均衡，不同国家和地区的自治域数量、规模差别很大等情况，基于地理位置的自治域级拓扑的可视化将有助于认识到国家层面互联网建设的差距。

现有的互联网可视化技术往往更注重互联网路由的拓扑关系，可视化的结果并不能很好的反映自治域的规模、自治域分布与地区的联系等信息；而且互联网拓扑关系数据由于测量的不稳定变化较大，基于拓扑关系形成的互联网拓扑图往往不够稳定。因此选择地理位置作为互联网拓扑可视化的一个重要维度，可以为可视化提供网络背后与国家地区之间的联系，也能提高拓扑图的稳定性。

1 基于地图的可视化方法

1.1 自治域的地理位置计算

在地图上绘制图形是反映地理信息最直观的方式。首先要计算各自治域的地理位置信息。互联网中 IP 地址与地理位置的映射通常称为 IP 地理定位[1]。现有一些 IP 定位数据库，存储互联网中公网 IP 地址范围所对应的位置，一般精确到城市。自治域会在互联网内声明其 IP 前缀地址，即其域内 IP 地址的范围。自治域声明的前缀大小是不确定的，有可能会声明多个前缀，因此自治域的地理位置并不一定是能够精确定位到地点，而是有可能分布在多个位置，需要用一个地理集合来表示。

地理定位数据库存储的是 IP 地址范围内的最小 IP 和最大 IP，由于自治域前缀数量和定位数据量都很大，如果使用二分法确定某一 IP 所属区间，时间复杂度将会很高。为减少计算和查询 IP 定位

基金项目：国家重点研发计划(2018YFB1800702，2016YFB0801303，2016QY01W0103)。

作者简介：安宇昊（1995-），男，硕士研究生，主要研究方向：网络拓扑建模与可视化；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。

收稿日期：2020-05-06

数据的时间,使用前缀树的方式来存储定位数据。为方便绘图,将定位数据中的国家城市信息转换为经纬度进行存储。计算自治域地理位置信息的步骤如下:

Step 1 初始化前缀树 geoTree,读取各国家城市的经纬度数据;

Step 2 逐行读取 IP 地理定位数据库内数据,将 IP 地址范围转换为若干前缀,将城市信息转换为经纬度,写入前缀树中;

Step 3 读取自治域的声明的各个前缀,对每个前缀在前缀树中查询其覆盖的前缀子集,并读取这些子集中的地理信息经纬度;

Step 4 对各自治域的地理信息经纬度进行去重和排序,作为其地理信息。

1.2 地图投影方法

在平面地图中展示呈球状的地球表面需要借助到地图投影,但任何的投影方式都会产生失真[2],因而选用恰当的地图投影方式是必要的。地图投影分圆锥投影、圆柱投影、方位投影等几类,对于大尺度的地图而言,方位投影是最好的选择[2]。选择方位投影中的球心投影,将大圆弧线投影为直线,适合表现世界尺度下的区域关系。借助投影方式,就可以将自治域根据其地理信息绘制在地图上。最简单的方法,就是在基础的地图层上,在每一个地理定位点上绘制出一个小区域,表示出整个自治域的地理信息分布情况。

1.3 自治域地理中心点计算

由于球心投影只能投影出地球表面的一部分,为能够看清自治域在较大范围的分布情况,除了最基本的投影面调整、比例尺切换外,还需要能够定位自治域的主要分布区域,也就是说,在展示单个自治域的地理信息时,需要将投影视角变换为投影面中心,与自治域分布中心重合。然而,自治域分布中心并不能够简单的使用所有地理分布点的几何中心代替,几何中心可能会因为一些特殊点的存在而有较大误差。因此,设计了一种基于密度聚类的地理信息中心点确定方法,其基本思想是使用密度聚类算法,在若干经纬度坐标中找到密集程度较大的区域,以这个区域的球面质心点,作为该自治域的地理中心点。

聚类是数据挖掘中无监督学习的一种,最常见的聚类算法是 k-means 算法,但其在聚簇的形状和密度不同、数据集异常点过多时不能很好的工作。密度聚类算法解决这一问题的方法是密度相似的、彼此连接的区域应属同一聚簇。最终选用 OPTICS 算法进行密集区域发现。OPTICS 算法是对最初的 DBSCAN 密度聚类算法的改进[3],其主要参数为密集点的最小邻居数 minPts 和计算两点距离的距离函数。为减小算法计算量,计算两地理点间的距离时,将地球从椭球抽象为圆球。

将计算两经纬度间球面距离的函数表示为 Distance(point1, point2),计算球面质心的函数表示为 Centroid(points),则计算自治域地理中心点的计算方法可以表示为式(1):

$$center = Centroid(OPTICS(geolist, minPts, Distance)) \quad (1)$$

1.4 自治域可视化方法

为体现不同地区之间自治域分布的差异性,以及不同自治域之间的规模差异等信息,还需要在同一地图上同时展示多个自治域。相比只展示一个自治域时,增加了二个新的挑战:一是单个自治域的覆盖范围可能很大,多个自治域之间如何解决重合问题;二是不同自治域之间如何能够明显的区分。

在同时绘制多个自治域时,并不绘制出自治域所有地理信息,而是使用圆心位于自治域地理中心点的圆形来表示该自治域,可以比较好的解决多自治域重合问题。圆形的半径正相关于自治域的覆盖范围,以表示不同自治域之间的规模差异。图 1 是自治域地理位置覆盖经纬度数量的直方图,从图 1 可以看到,自治域的规模分布非常不均衡,绝大多数自治域只覆盖了极少的位置,而覆盖范围在 100~200 之间的自治域则非常少,因而设定圆形半径与自治域覆盖范围呈 log 函数关系,以使得半径的过渡更加圆滑,不至于出现大多数自治域半径极小的状况。

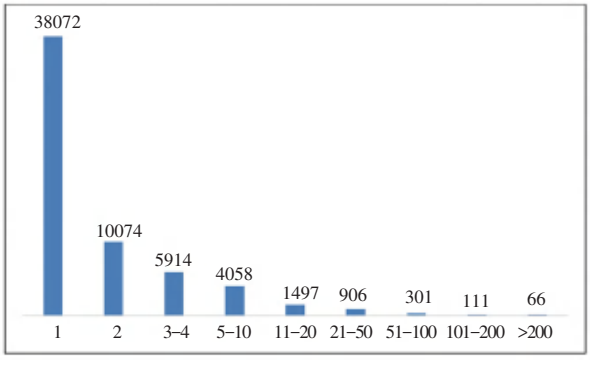


图 1 自治域地理位置覆盖经纬度的数量统计直方图

Fig. 1 The number of geographic locations covered by Autonomous Regions Histogram

绘制不同自治域使用不同的颜色,使自治域容易区分。为体现自治域分布与国家的联系,归属于

同一国家的自治域使用相近的颜色。给拥有自治域数量较大的国家分配基准颜色，自治域的颜色由这些基准颜色随机生成。这样就可以生成基于地图的自治域级互联网分布图，图中用圆形代表自治域，其大小表示自治域规模；颜色表示所属国家，能够反映自治域分布与地区、国家之间的联系。

先根据地图投影绘制出基本的地图层，再依次将要可视化的自治域根据如上所述的圆心、半径、颜色计算方法计算好之后，最终将这些图形根据地图投影绘制在图上，得到基于地图的自治域可视化结果。

2　融合拓扑信息与地理信息的自治域级拓扑可视化

2.1　自治域的拓扑位置计算

自治域（Autonomous System，AS）是在一个或多个网络运营商的控制下的一组相连的 IP 路由前缀集合，该集合对互联网呈现出一个共同明确的路由策略，通常对应运营商、机构、大学、公司等组织。互联网中流量的转发在自治域级别与商业策略结合紧密，自治域之间主要有两种商业关系：供应商-客户（Provider-Customer，P2C）和对等（Peer-Peer，P2P）[4]。P2C 指客户向供应商付费，供应商为客户提供传递 Transit 服务，即为客户接入互联网；P2P 指 AS 间相互传递流量，互不付费。由于有这样的商业关系存在，自治域间就存在了明显的层级关系，供应商 AS 处于客户 AS 的上层。最顶层的自治域通常称为 Tier1 自治域[5]。Tier1 自治域之间通过 P2P 关系连接，不需要 Transit 即可到达整个 Internet[6]。而 Tier2 自治域通常定义为：与某些网络免费对等，但仍需要购买 Transit 服务才能到达整个 Internet 的自治域；Tier3 自治域：仅从其他网络购买中转/对等网络，以参与 Internet 的自治域。

自治域的拓扑位置指的是自治域在互联网拓扑中所处的层级位置。在这里将 Tier1 和 Tier2 自治域都视为顶级自治域，自治域的规模由其拥有的客户自治域数量来表征。拓扑位置的计算方法可以表述为：（1）对于顶级自治域，规模越大的自治域拓扑位置越靠上。（2）对所有的非顶级自治域，拓扑位置在与其最邻近的顶级自治域附近。这里使用最邻近自治域来表示的意义在于这也是其接入 Internet 最主要的供应商。

2.2　融合地理信息的可视化方法

基于地图的可视化方法可以清晰的看出各自治域在地理上的分布情况，但并不能看出各自治域在互联网中的重要程度。因此，将自治域的拓扑位置和地理位置相结合，作为一张图的两个维度进行可视化。

以拓扑位置作为纵轴的维度。顶级自治域根据拓扑位置的高低从上到下依次排列，一个自治域占据一行，每个自治域占据的高度与其自治域规模正相关。Stub 自治域的纵轴位置与其最临近的顶级自治域一致，附属在顶级自治域下方，表示其最主要的流量来源。

地理位置作为横轴的维度。地理位置由经纬度表示，但作为单一维度时，只能选其一。相比于纬度，经度对地理信息和国家的表征能力更强，因此使用经度表示自治域的地理位置。自治域的地理信息是分布在不同位置的点集。但点集不利于把自治域表示为一个整体，使用覆盖点集的区间线段作为自治域的横轴坐标。区间要尽量覆盖连续的经度，不覆盖间隔较大的经度。计算区间的算法可使用表 1 中的伪代码来描述。算法的输入参数为 X 和 P，X 是自治域的经度信息列表，西经表示成范围为 -180 ~0 的负数，东经表示成范围为 0 ~ 180 的正数；P 是分割粒度，其含义是当经度列表中相邻两个经度差值小于这个分割粒度时，判断自治域覆盖这个经度范围。算法会输出得到的经度区间列表 Intervals，需要注意的是，区间列表中可能会出现恰好越过 180° 的区间，这时需要将其拆分为[……，180]和[-180，……]两部分。

表 1　计算自治域经度区间算法伪代码

Tab. 1　Pseudocode for calculating the longitude interval algorithm

```
Calculate Longitude Interval (X, P)
Input：经度信息数组 X [X1, X2, X3,…, Xn]，分割粒度 P
Output：区间列表 Intervals
1   X = Sort (X)
2   D = X
3   For i = 1; i < n; i ++
4    Di = Xi+1 - Xi
5    Dn = 360 - Xn + X1
6    Dm = max(D)
7   Intervals = [ ]
8   I = [Xm+1]
9   For i = m + 1; ;i = (i = = n?  0:i + 1)
10   If Di < P
11       I.push (Xi+1)
12    Else
13       Intervals.push(I)
14       I = [Xi+1]
15   Return Intervals
```

计算自治域的纵轴坐标和横轴区间列表后，可

以得到一张以拓扑位置为纵轴、地理经度为横轴的拓扑图。但由于东西方互联网发展不均衡,以及太平洋、大西洋的存在,使得自治域在经度的分布不均匀。为尽量减少图中的空白,更明确的表示哪些地区的自治域规模更大,需要为经度坐标轴调整比例,也就是给不同的经度范围不同的权重。经度坐标轴的刻度从-180°~180°,每隔 30°一个刻度,共 13 个刻度,12 个刻度区间,要计算的就是这 12 个刻度区间长短的相对权重。这个权重由各自治域的经度范围落入刻度区间的长度来表达,读取自治域各经度区间的端点[x, y],并找到 x 和 y 分别落入的刻度区间[q_1,q_2] 和[q_3,q_4],那么区间[q_1,q_2] 增加权重 $q_2 - x$,[q_3,q_4] 增加权重 $y - q_3$,如果存在 q_2 和 q_3 之间的区间,则增加等于其区间长度的权重。对每个自治域进行上述操作后,得到权重数组 $W = [w_0, w_1, \cdots\cdots, w_{11}]$。令 d 为坐标轴总长度,t_0 是第一个坐标刻度的位置,则横坐标轴第 i 个刻度的实际坐标可由式(2) 得:

$$t_i = t_{i-1} + d \cdot \frac{w_{i-1}}{\sum_{i=0}^{11} w_i} (i \geqslant 1) \qquad (2)$$

在得到新的横轴坐标刻度后,可以根据横纵轴刻度绘制出完整的拓扑图。

3 实验结果及分析

3.1 基于地图的可视化方法

在可视化技术上,选择了 Web 前端技术进行绘制。使用 Canvas 画布绘制地图和图形,使用 D3.js 框架进行绘图的控制和管理。地图投影方式使用了球心投影,也称为日曼投影(gnomonic projection)。球心投影下的世界地图见图 2。地图以及可视化的所有图形数据都是用 GeoJSON 表示的,GeoJSON 是基于 JavaScript 对象表示法的地理空间信息数据交换格式,这种格式方便将图形移植到其它地图上,诸如 Google 地图、百度地图等知名互联网地图网站都支持绘制 GeoJSON 图形。

图 2 球心投影下的地图

Fig. 2 Map by gnomonic projection

自治域 ChinaNet 骨干网申请的自治域号为 4134,是属于中国的一个 Tier2 自治域,声明了 183.0.0.0/10 等 800 多个前缀,按照自治域地理位置计算方法得到若干经纬度后,为减轻存储压力对经纬度取整,剩下 538 个经纬度,AS4134 的可视化效果如图 3 所示。可以看到 ChinaNet 自治域的地理位置主要分布在中国区域内,说明地理位置计算方法比较准确,落在中国区域外的点可能是一些客户(Customer)自治域的位置。

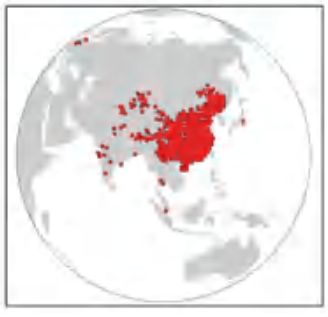


图 3 AS4134 的可视化效果图

Fig. 3 Visualization diagram of AS4134

在展示自治域时需要将地图中心调整到自治域中心点的位置,地图投影的移动方式其实是一个简单的动画,在一秒内绘制出从起点到终点的一帧帧地图,呈现出转换的效果。自治域地理中心点不能简单的用自治域地理位置的几何中心来代替,否则可能会出现如图 4 所示的情况。

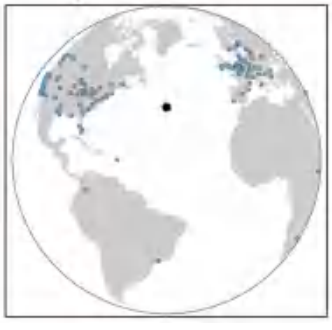


图 4 AS6939 的地理位置的几何中心

Fig. 4 The geometric center of AS6939

图 4 是归属飓风电气有限责任公司(Hurricane Electric LLC.)的自治域 AS6939 的地理位置可视化图,图中的黑点是 AS6939 地理位置的几何中心。由于 AS6939 是一个参与了欧洲所有互联网交换中心的 Tier1 自治域,其地理位置同时分布在美洲和欧洲,其地理位置的几何中心反而落在了大西洋上,并不能代表其中心位置。而使用自治域地理中心点的计算方法之后的效果如图 5 所示(为显示全地理

位置对地图进行了旋转），AS6939 的地理中心点落在了美国西海岸，恰好与其总部位置接近。可以看出自治域地理中心点的计算方法效果显著。

图 5　AS6939 的地理中心点

Fig. 5　The geographic center of AS6939

在展示多个自治域时，使用圆形来表示自治域，圆心为地理中心点，半径 r 与自治域地理位置的覆盖范围 S 成 log 函数关系。设 $r(S) = a \cdot \log(S) + b$，令 $r(1) = 2$，$r(800) = 20$，拟合之后得到圆形半径 r 与自治域覆盖经纬度数量 S 的关系为式（3）：

$$r = 2.5 \cdot \log(S) + 2. \tag{3}$$

选择一些自治域数量较多的国家分配基准颜色，然后按照自治域所属国家根据基准颜色随机生成相近颜色。自治域数量排名靠前的国家分布情况如图 6 所示，可以看到排名前 7 的国家已经占据了互联网内一半多的自治域，给这些自治域以及中国分配的基准颜色见表 2。将这些基准颜色的 RGB 值上下做一些波动，可以得到相近的随机颜色。互联网自治域基于地图的地理位置可视化绘制后的效果如图 7 所示（不同投影角度），可以看出不同地区自治域分布的密集程度区别，也能看出各国家地区自治域分布的差异性。

表 2　国家基准颜色

Tab. 2　Standard color of Nations

国家	颜色	十六进制 RGB 值
美国	蓝色	#87CEEB
俄罗斯	黄色	#FFFF00
巴西	绿色	#008000
波兰	深棕色	#CD853F
英国	紫色	#9400D3
乌克兰	橙色	#FF8C00
德国	浅绿色	#90EE90
中国	红色	#FF0000
其他国家	浅灰色	#D3D3D3

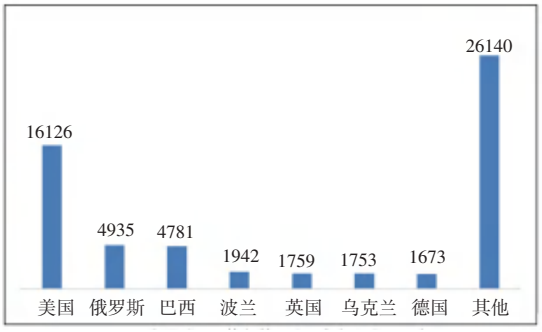
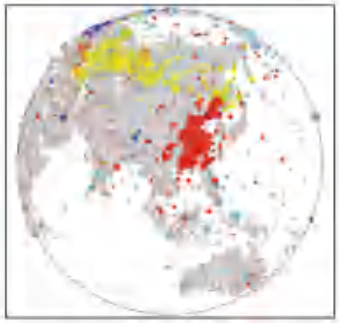


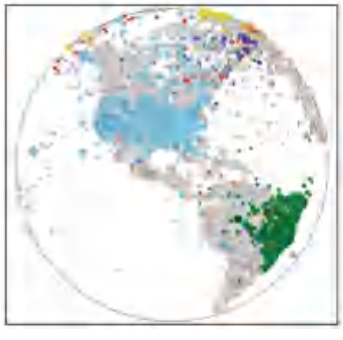
图 6　自治域数量排名前 7 的国家自治域数量统计图

Fig. 6　Top 7 country in the number of Autonomous Domain Bar chart



（a）亚洲中心投影角度

（a）Projection perspective from Asia



（b）美洲中心投影角度

（b）Projection perspective from America

图 7　互联网自治域基于地图的地理位置可视化

Fig. 7　Visualization of geographical location of Internet Autonomous Domain

3.2　融合地理信息的可视化方法

拓扑位置与地理信息融合的可视化方法绘制一张以拓扑位置为纵轴，地理位置为横轴的拓扑图。首先计算纵轴坐标，即自治域的拓扑位置。自治域的拓扑位置计算分为顶级自治域和非顶级自治域两部分。对于顶级自治域，首先要确定哪些自治域是 Tier1 和 Tier2 自治域。Tier1 自治域使用了目前公认的包括 AT&T、Level3 等的 16 个 Tier1 自治域[7]。Tier2 自治域根据自治域的规模，选取了自治域规模大于 100 的自治域作为 Tier2 自治域。将这些

Tier1、Tier2 自治域按照规模排序后,就得到了顶级自治域的拓扑位置。对于非顶级自治域,根据前述算法得到与其最邻近的顶级自治域,拓扑位置位于其顶级自治域下方。

自治域的横轴坐标由地理位置确定,更确定的说是由覆盖的经度确定,但这里需要把经度覆盖范围分割为经度区间。以自治域 AS6830 为例,204 个经纬度坐标,提取出经度后有 49 个经度组成经度列表,在分割粒度为 50°的情况下,使用表 2 中的算法计算得到[-121,-64],[-9,80],[140]这三个区间。

计算所有顶级自治域的横轴坐标后,根据分布情况进行坐标轴调整。根据调整后的坐标轴刻度重新计算自治域横轴坐标,即可以绘制拓扑图,使用 HTML 中的 SVG 图形来绘制,互联网自治域拓扑位置与地理信息融合的可视化效果如图 8 所示。

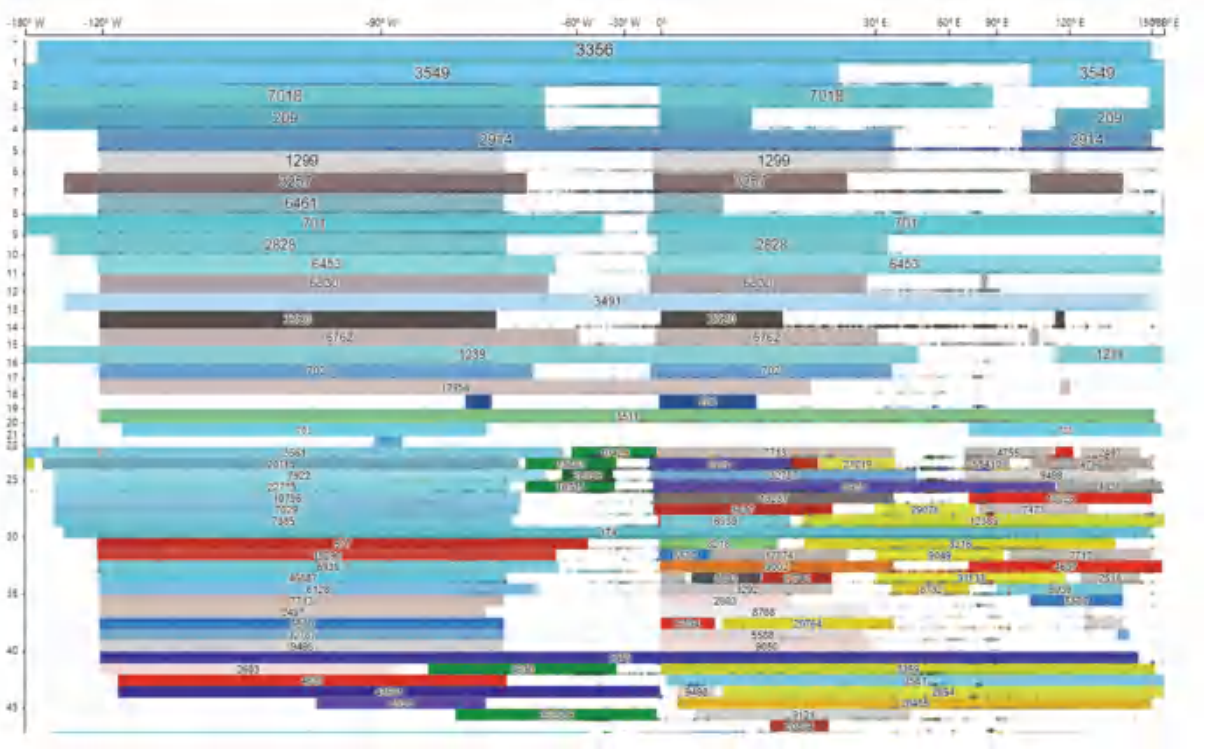


图 8 互联网自治域拓扑位置与地理信息融合的可视化

Fig. 8 Topological Location and Geographic Information Fusion Visualization

4 结束语

本文提出了一种基于地理信息的自治域级互联网可视化方法。相比于传统的网络拓扑可视化方式,本文方法着重体现出了互联网内自治域分布与地理信息之间的联系。基于地图的可视化方法表现地理信息的方式更加直观,能够准确的表现出自治域在地理上的分布;融合地理信息的可视化方法较为抽象,着力于体现顶级自治域的经度覆盖范围。可以看到不同国家、不同区域之间的互联网资源分配情况。在未来,可以使用地理信息对互联网内不同层次的网络拓扑,继续进行可视化方法探索,对网络拓扑测量与拓扑建模工作有一定的应用和指导意义。

参考文献

[1] MUIR J A, OORSCHOT P C V. Internet geolocation: Evasion and counterevasion [J]. ACM Computing Surveys (CSUR), 2009, 42(1): 1-22.

[2] 王英钧. 电子地图的地图投影[J]. 航空电子技术, 1995, (2): 29-33.

[3] SCHUBERT E, GERTZ M. Improving the Cluster Structure Extracted from OPTICS Plots[C]//LWDA. 2018: 318-329.

[4] GAO L. On inferring autonomous system relationships in the Internet[J]. IEEE/ACM Transactions on networking, 2001, 9(6): 733-745.

[5] WINTHER M. Tier 1 isps: What they are and why they are important[J]. IDC White Paper, 2006: 1-13.

[6] BERG R V D. How the 'Net Works: An Introduction to Peering and Transit[J]. Ssrn Electronic Journal, 2009.

[7] https://en.wikipedia.org/wiki/Tier_1_network