挑战一可视化方案特点分析

侯伟婷, 林培文, 于阜甲, 张加万

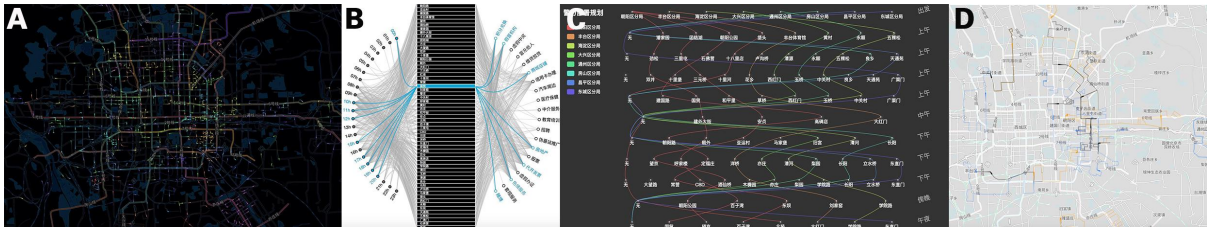

图 1. A：伪基站行为轨迹拟合图，B：各商圈时间点与垃圾短信类型关系图，C：警力部署图；D：执法人员巡逻路线图

摘要—基于 QHNet 公司推出的手机卫士应用软件积累的垃圾短信样本数据，使用垃圾短信发送和接受时间、MD5、经纬度信息对伪基站的时空活动规律进行分析，拟合并以地图形式展示了伪基站的行为轨迹，在此基础上按照短信类型和对个人的经济危害程度对垃圾短信进行分类，并用嵌套环形图展示类别数据的特征，并以概念图的形式展示发送不同垃圾短信类型的伪基站的时空规律，最后根据伪基站的行为模式给出了可视化解决方案。

关键词—伪基站；垃圾短信；可视化分析；地图；概念图

1 简介

我们的完成过程主要分为三个阶段。第一个阶段是对原始数据进行预处理，并根据数据特征分析总结伪基站的活动规律，第二个阶段是对垃圾短信文本进行处理，分析伪基站与短信类型在时空上的综合联系，第三个阶段是针对上述过程得来的伪基站行为模式提出合理化的解决方案，用以打击伪基站活动。

本文的第一个部分是伪基站行为模式的可视化方案，并对展示的方案进行了详细的说明。

本文的第二部分是针对上述得来的伪基站行为模式，给出执法人员在重点行政区的商圈巡查路线，并结合地图进行了详细的说明。

本文的第三个部分是对本文主要内容进行的简单总结。

2 伪基站行为模式可视化设计

在进行数据分析和可视化之前，对原数据集进行清洗，清洗的规则包括：垃圾短信的接收时间早于发送时间；接收时间和连接时间异常，即不在2017/02/23至2017/04/26之间；连接时间早于接收时间前一天的18:00之前；因短信字数限制而分割出的第二条短信。

数据预处理之后，首先利用地图展示垃圾短信的时空分布规律，据此分析总结伪基站在各区数量和它的活动规律，并绘制伪基站行为轨迹拟合图。在此基础上对垃圾短信进行分类，第一种分类过程利用词云展示短信正文关键词，并以嵌套环形图展示各类别间的关系和数量。第二种分类是根据短信对个人的经济危害程度高低对垃圾短信进行分类，并用面积图展示各行政区内不同类型垃圾短信的数量分布。最后概念图展示发送不同短信类型的伪基站活动时间及其覆盖的商圈范围。

2.1 伪基站时空活动规律设计

2.1.1 垃圾短信时间分布

用散点图展示2017/02/23至2017/04/26期间垃圾短信每天24个小时的数量变化。如图2所示，点的颜色反映了在此刻的垃圾短信数量，由蓝变红的渐变颜色表示短信数量由低到高，红色点集中处表示了伪基站相当活跃的时间。

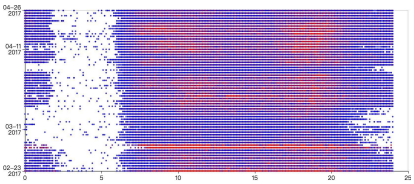

图 2 02-23 至 04-26 期间垃圾短信 24 时数量变化图

2.1.2 垃圾短信空间分布

去掉数据集中重复的经纬度位置，各数据点重复量通过点的明暗程度进行刻画，经此处理后，投影在地图上的点勾勒出了北京各大主干道和高速公路的轮廓，说明伪基站多活动于北京中心的繁华商业区内的各大主干道及高速公路附近。

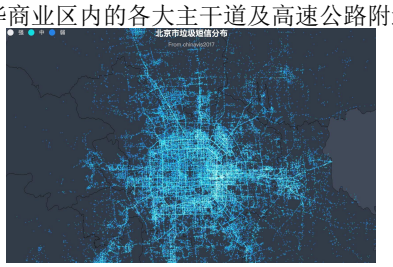

图 3 垃圾短信分布图

2.1.3 伪基站时空分布

综合伪基站时空分布规律，再次对数据进行处理，在同一时间段内，设位置相近的基站的中心点为伪基站在该时段的真实位置，从这些位置中找出一天中每个时段伪基站出没的高频点，并将其映射在地图上，如图4左图所示。图4的右图显示了北京16个行政区在24个时段中的活动猖獗程度指数。可以观察出，伪基站频繁出没于朝阳，东城，西城等地区，其中朝阳区伪基站的猖獗程度无出其右，由于东城、西城的占地面积小，所以虽然伪基站活动频数稍小，但是伪基站活动猖獗等级很高。

- *侯伟婷. 天津大学. E-mail: hhhouwt@163.com.*
- *林培文. 天津大学. E-mail: peiwen051@gmail.com.*
- *于阜甲. 天津大学. E-mail: fujiazhiyu@sina.com.*
- *张加万. 天津大学. E-mail: jwzhang@tju.edu.cn.*

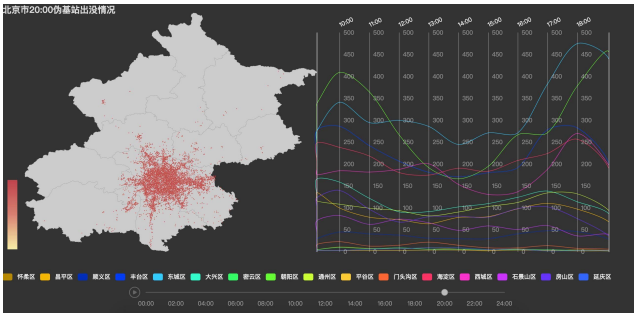

图 4 伪基站时空分布图

2.1.4 伪基站活动路线拟合

由图3可知，伪基站经常出没于某些点，连接某个特定区域内的所有高频点，能够反映出伪基站在该区域的活动规律。基于此想法，对处于同商圈内的高频点做聚类，而后对聚类点做曲线拟合，得到伪基站的活动轨迹，如图1中A图所示。

2.2 伪基站发送不同类型短信时空规律设计

2.2.1 垃圾短信类型分类

处理垃圾短信正文，并使用词云图展示垃圾短信主要涉及到的关键词，如图5左图所示。将垃圾短信分为广告推销、诈骗、非法服务及其他四大类，每一大类包含数量不等的小类，用嵌套环形图展示类别之间的关系和数量比例，如图5右所示，内嵌圆的扇形部分表示某一大类，扇形映射到的外侧环形部分是其包含的短信子类别。

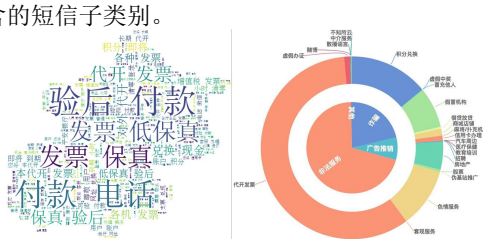

图 5 垃圾短信类型分析图

2.2.2 垃圾短信对个人经济危害程度分类

根据短信对个人的经济危害程度高低对垃圾短信进行分类，各行政区内不同危害程度类型的垃圾短信的数量分布，如图 6 所示。

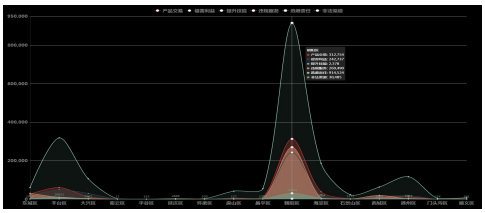

图 6 各行政区不同经济危害程度类别数量图

2.2.3 伪基站发送不同类型垃圾短信时空分布

综合伪基站发送不同类型短信的时间规律和空间分布，绘制如图 1 中 B 图所示的关系图，中间各矩形代表伪基站出没商圈，左侧可以高亮显示伪基站活动时间，右侧可以高亮显示商圈接收到的主要短信类型。当鼠标悬浮在某商圈上时，则会高亮显示该商圈内伪基站的活跃时间和主要短信类型，如图 1 中 B 图所示；当鼠标移到左侧某一时间点上时，则会高亮显示在此时间点内伪基站活跃的商圈列表；当鼠标移到右侧某一短信类型时，则会高亮显示此短信类型被较频繁发送的商圈。

鼠标点击矩形商圈，则显示如图 7 所示，中心圆表示选定的商圈，外围圆数字表示与此类型短信或此活动时间相关的其他商圈个数，点击此外围圆，则显示其他商圈列表，也可在主视图中点击垃圾短信类型和活跃时间，视图结果如图 8 所示。

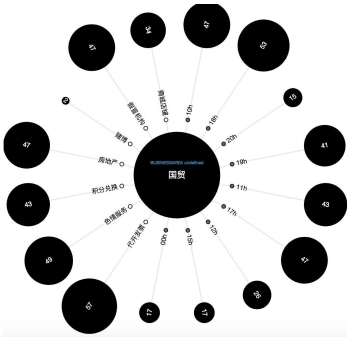

图 7 国贸商圈主要短信类型和伪基站活跃时间图

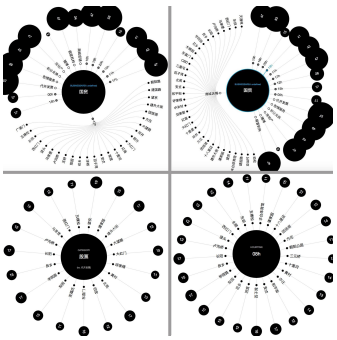

图 8 伪基站发送不同类型短信时空分布其他视图

3 打击伪基站的解决方案可视化设计

根据章节2所得的伪基站行为模式，结合危害程度较高的垃圾短信聚集商圈以及图1中A图的伪基站活动路线，设计伪基站活动较为猖獗的几个行政区内执法人员的巡逻检查路线，如图1中C图，纵轴表示上午、下午和午夜时刻，横轴表示伪基站出没频繁的商圈，颜色表示不同行政区。图1中D图是对上述路线的详细展示，蓝色路线表示上午从各区分局出发的巡逻路线，黄色路线表示下午巡逻路线，黑色路线表示午夜从各区分局出发的巡逻路线。

4 结论

本文中所描述的伪基站行为模式的可视化方案和给执法人员提供的可视化解决方案都具备了准确性、实用性及新颖性的要求，在此基础上形象鲜明的解决了挑战一的所有问题，同时可视化方案具有良好的扩展性，可接受进一步的开发和编写，对用户发现更深层次的信息具有重要意义。

参考文献

[1] 陈伟,沈则潜. 数据可视化. 电子工业出版社, 北京, 2013.
[2] 邱南森. 数据之美. 中国人民大学出版社, 北京, 2014.