伪基站时空轨迹分析与可视化

吕博, 贾月, 吴兆胜

摘要— 本文基于大数据分析对伪基站的时空轨迹进行了分析和可视化，提出监测和治理伪基站的新方法。使用 jieba 分词库对短信内容进行分词和语义分析，提取短信内容的关键词，根据关键词的距离对短信内容进行分类聚类，依托大数据分析与可视化，利用 processing 搭建时空轨迹查询与可视化系统，并通过 Echarts、JS 等对伪基站短信内容进行时域特征分析，为伪基站排查工作的常态化奠定基础。

关键词— 伪基站, 时空轨迹, 可视化, Processing, Echarts,JS

简介

随着信息技术的发展，用“伪基站”发送垃圾短信已成为不法分子诈骗、推销的新手段，伪基站冒用银行、运营商、国家机关或他人号码，强行向用户发送诈骗、色情、赌博、广告等短信息，影响正常通讯，严重侵害了社会秩序。本文基于ChinaVis2017挑战赛开放的垃圾短信样本，开发了伪基站时空轨迹分析与可视化系统，深入挖掘伪基站位置移动与发送内容的时空规律，在深入理解伪基站行为模式的基础上，为伪基站综合治理提出建议。本文首先介绍了伪基站时空轨迹分析与可视化系统的实现方案及关键技术，然后针对伪基站时空轨迹分布、短信内容分类与危害程度进行可视化分析，最后面向伪基站监管与治理的实际需求，提出了切实可行的解决方案。

1 伪基站时空轨迹分析流程

挑战赛开放的数据源为含有文本内容的时空轨迹数据，在对其进行可视化之前需要进行一系列数据处理操作，再提升数据质量之后对数据进行分析，将数据转化为更适合可视化的信息。本文对数据进行分析的流程如图1所示。

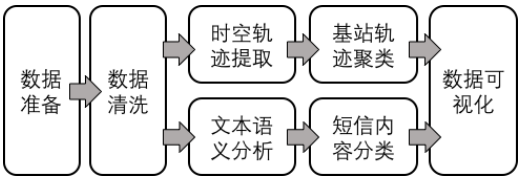


图1 伪基站时空轨迹分析流程

在数据准备部分，除了比赛提供的数据之外，本文还使用了以下的数据源：（1）OpenStreetMap提供的北京市地图数据，针对比赛开放数据与OSM数据存在的坐标偏移进行了人工矫正；（2）为了研究短信内容的分布规律，本文利用百度提供的LBS API对比赛数据中的经纬度坐标进行了映射，获取了每一个经纬度所在市区信息。在数据清洗部分，剔除了经纬度超出地理边界的异常点，剔除了时间超出研究范围（2017年2月23日—2017年4月26日）的数据条目。

在清洗后的数据中，伪基站伪装的发送方电话号码是识别一条时空轨迹的标识，但是有两个问题使我们不能用其作为识别伪基站轨迹的标志。问题一是有多个伪基站可能伪装成了同一个号码（例如95580、10086等），问题二是同一个伪基站可能同时使用了多个伪装号码。为了对伪基站的轨迹进行可视化，首先将每一条用电话号标记的时空轨迹按照空间上的欧氏距离进行聚类，鉴别同一时段同一个号码背后的多条轨迹，从而解决问题一。然后利用轨迹的时空相似度将多条时空规律都极为接近的轨迹聚合为同一个伪基站，从而解决问题二。

为了分析伪基站发送短信内容的时空规律，使用jieba分词库对短信内容进行分词和语义分析，提取短信内容的关键词，根据关键词的距离对短信内容进行分类聚类，参考专家对垃圾短信的分类方法，将短信分为办证发票、冒充身份、色情服务、地产广告、违禁推销、招聘广告、非法博彩、打折促销、教育移民和其他类型等十大类，并对每一类短信的危害程度进行量化评估。

最后，在数据处理与分析的基础上，利用数据可视化的手段对数据进行深入挖掘。由于时空分析是挑战赛的重点，为了实现细粒度、海量坐标点的时变可视化，使用Processing绘制时变的热力图、散点图和轨迹图。为了实现和统计分析数据的交互，使用Echarts绘制河流图、矩形树图、桑基图等，并使用jQuery实现前端可视化界面与数据库的交互。

2 可视化分析与结论

本节首先针对伪基站活动的时空模式进行分析，进而在短信内容分类与危害程度评估的基础上，对短信内容分布的时空规律进行研究，得到了以下可视化结果与结论。

2.1 伪基站移动时空模式分析

2.1.1 伪基站活动总体态势分析

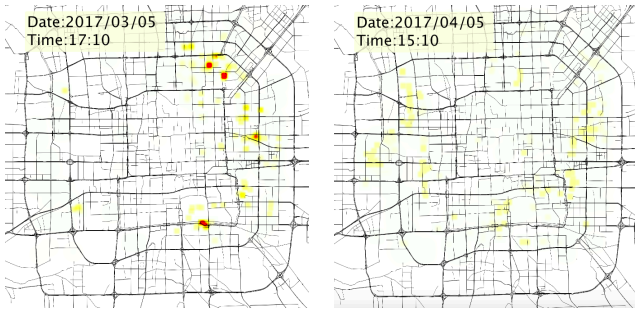


图2 伪基站活动总体态势图

伪基站在选择发送地点和发送时间时会非常谨慎和有针对性，尤其是和市民的日常行为有非常大的关联。从图中可以宏观的看出伪基站分布总体的时空趋势，例如图2所示，3月5日伪基站发送短信数据比较多，而且比较集中在少数几个热点区域（例如写字楼、商圈等），反之4月5日为清明长假，由于市民大多休息在家或出城旅游，伪基站发送短信数明显减少，而且分布更加均匀。

- *吕博,中国电子科学研究院创新中心工程师, E-mail:blyu@csdslab.net*
- *贾月,中国电子科学研究院创新中心工程师, E-mail:yjia@csdslab.net*
- *吴兆胜,中国科学技术大学硕士研究生, E-mail:mega_victor@163.com*

2.1.2 伪基站移动轨迹分析

针对原始数据中存在的多伪基站伪装成同一个号码和同一个伪基站可能同时使用多个伪装号码这两个问题，利用空间聚类和轨迹时空相似度来进行数据处理，基于 Processing 搭建时空轨迹查询与可视化系统，可进行缩小放大、轨迹检索、轨迹可视化等基本功能，同时支持地图风格选取、鼠标取经纬度等辅助功能。使用者可以在交互界面中输入手机号对轨迹进行检索，并指定轨迹的可视化颜色。在可视化轨迹中，轨迹的粗细与饱和度代表了伪基站发送短信的数量和频率。图 3 左图即为同一伪基站冒充 2 个手机号的移动轨迹，右图为两个伪基站冒充同一个手机号的轨迹。

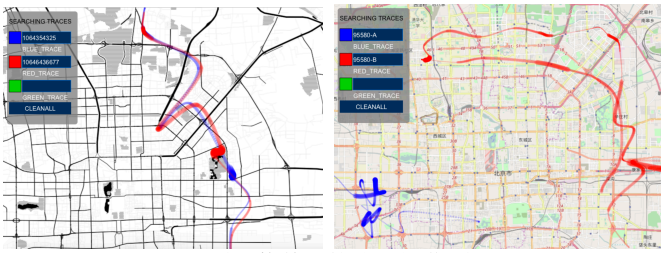


图 3 伪基站轨迹可视化图

2.2 伪基站短信内容分布规律分析

2.2.1 内容整体时空分布规律

针对伪基站发送短信内容，提取短信关键词进行分类聚类以后使用 Processing 将各类短信内容的宏观态势进行直观的展现，发现不同类别的短信时空分布上具有比较强的规律性。例如图 4 左图是选取 2 月 23 日 0 点到 6 点短信分布，发现主要为色情短信，且聚集在三环内繁华商圈地带，右图为 2 月 23 日 6 点到 12 点短信分布，主要为办证发票类短信，且主要集中在各个金融商圈地带。

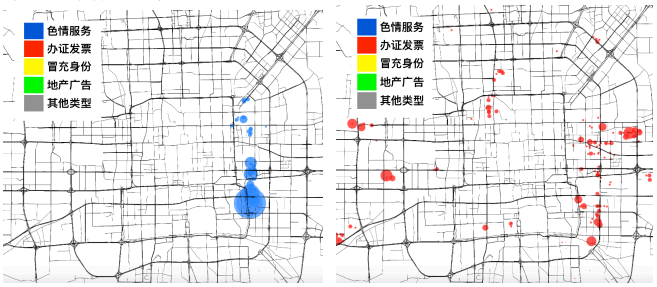


图 4 分类短信总体时空分布

2.2.2 内容分类与比例

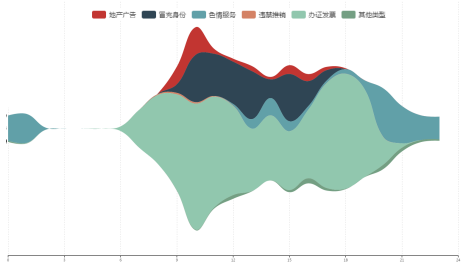


图 5 短信内容分类比例图

通过图 5 可以发现，总体上办证发票短信的数量最多，主要分布在白天；其次数量较多的色情类短信主要活跃于夜间及午间；之后冒充身份及地产类广告类短信活跃于上班期间；剩余少量的其他类型短信零星分布于全天。

2.2.3 内容时间分布规律分析

不同类别的短信出现的时间具有一定的规律性。根据统计各类型短信每天不同时段的数量，构建热力图。如图 6 左图为色情类短信分布图，分析得知，色情类短信主要分布在晚上 18 时至凌晨 3 时，并在法定节假日和执法部门增加执法力度后显著减少。如图 6 右图为办证发票类分布图，分析得知，办证发票类信息主要活跃于 6 时至 21 时，并且有极强的“周期性”：工作日活跃，双休日较收敛。

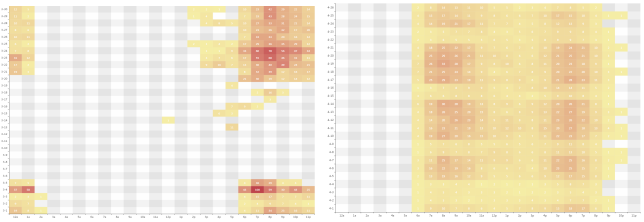


图 6 色情服务及办证发票类信息时间分布图

2.2.4 内容空间分布规律分析

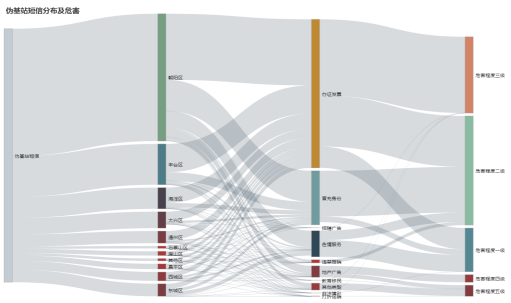


图 7 分类短信空间分布图

伪基站短信主要分布于经济发达、写字楼密集的城六区，以朝阳区为首，其次是丰台区、海淀区、东城区西城区与石景山区，但不与人口与土地面积成正相关。从桑基图中可以直观的看出伪基站短信的流向，例如色情服务类短线主要来自朝阳区，但是东城区色情服务类短信所占的比例同样不可忽视，而海淀区色情服务类短信所占比例最少；石景山区没有教育移民类短信。这些可以由图 7 分类短信空间分布 sankey 图更加直观展示出来

3 伪基站综合治理建议

根据针对伪基站数据的一系列可视化分析， 对伪基站管理提出以下建议。

(1) 伪基站轨迹跟踪与交通摄像头联动实现伪基站溯源

在现实中伪基站通常采用车载的形式，通过将伪基站数据于交通摄像头数据进行关联分析，通过对系统数据库中车牌号进行检索可以获得某个车牌号的移动轨迹，利用本论文提出的时空轨迹相似度关联的方法，可以准确锁定伪基站车辆的车牌号。

(2) 基于分析预测伪基站行为模式实现伪基站精准管控

由于不同的伪基站短信内容分布具有一定的时空分布规律性，利用此分布规律，安排相应的人力检测来实施观察、抓捕。

(3) 完善伪基站短信上报机制，建立伪基站号码查询系统并向社会开放

建议完善伪基站短信上报机制，汇聚数据资源，建立伪基站号码查询系统，具备伪基站号码查询、历史活动轨迹可视化、短信内容模糊检索等功能。当平台积累了一定量数据之后，可以建立伪基站上报与检索平台，面对社会公众开放提供服务。

参考文献

[1] https://processing.org/, 2017.

[2] http://echarts.baidu.com/, 2017.