PBSViz：基于伪基站短信数据的可视分析系统

陈斌，栾剑，徐劭斌，汤雅兰，张慧杰

摘要—本文针对 ChinaVis2017 挑战赛第一题的内容，基于伪基站发送垃圾短信的数据进行分析，提供了一套完整的可视分析方案和可视化系统。该系统结合了地图视图、日历矩阵视图、趋势图和气球图等多种可视化方法及交互技术，能够从宏观时空分析的角度出发，对垃圾短信数据进行可视分析，揭示伪基站的总体时空活动规律。然后通过 LDA 主题模型对垃圾短信的具体内容进行分类，针对划分不同类型的垃圾短信，进一步分析伪基站发送不同类型垃圾短信的时空分布规律；并结合该类型垃圾短信的地图视图、堆叠鱼刺图和径向弹壳图可视化的结果对时空分布规律进行总结；同时根据得到的伪基站行为模式，向执法人员提出整治伪基站的有效建议和方案。

关键字—可视分析，交互技术，伪基站，行为模式

简介

不法人员利用伪基站发送短信，不仅干扰公共频率资源，影响人们的正常通讯，而且通过群发短信的方式发送内容不受控的信息，侵犯公民个人隐私，严重侵害了社会秩序。ChinaVis2017挑战赛第一题提供的伪基站发送垃圾短信的数据，其时间跨度为近2个月，包含7个属性维度，共有300多万条记录，具有大数据的规模大、多维及时序性等特点。为了满足此次挑战赛的分析需求，我们的探索过程主要分为三个阶段，第一阶段针对伪基站发送垃圾短信的数据进行预处理，包括从原始数据中删除错误数据和冗余数据、从原始数据集中抽取子集样本作为待处理数据以及通过LDA主题模型对短信内容分类等。第二个阶段是数据可视化，在上一阶段的基础上，我们针对分析需求，结合多种可视化方法及交互技术，提供完整的可视分析方案和可视化系统。第三阶段是可视化结果分析阶段，对垃圾短信数据进行可视分析，揭示伪基站的总体时空活动规律。

本文主要通过三个部分来论述我们的工作，第一部分介绍我们的主要可视化方案，并说明如何解决问题。第二部分主要从实用性和可交互性两方面来分别论述我们可视化方案的特点。第三部分对本文的主要内容进行总结。

1 可视化方案

1.1 探索垃圾短信数据宏观时空活动规律

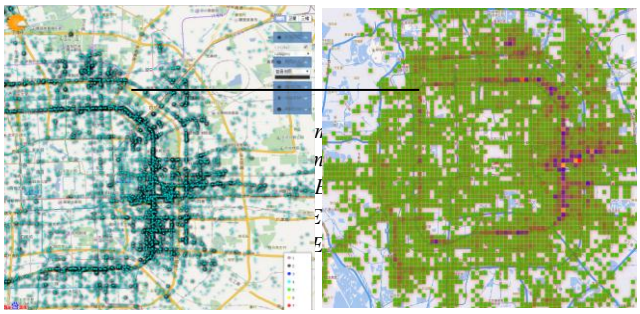


图 1 地图视图可视化伪基站的空间分布

挑战一中QHNet公司提供的伪基站发送垃圾短信的数据，包括伪基站发送短信伪装的电话号、短信内容、垃圾短信接收时间戳、与伪基站连接的时间戳、伪基站发送此条短信时的近似位置经度和维度等7个属性。地图视图能够展现伪基站影响区域的空间规律，并提供热力效果和马赛克效果（如图1）等可视化展示，其中每一个马赛克单元代表所在区域垃圾短信的数量，清楚呈现伪基站的空间分布规律。另外，为了得到更美观的可视化效果，可以通过调整地图视图的底图选择用户更喜欢的可视化方案。

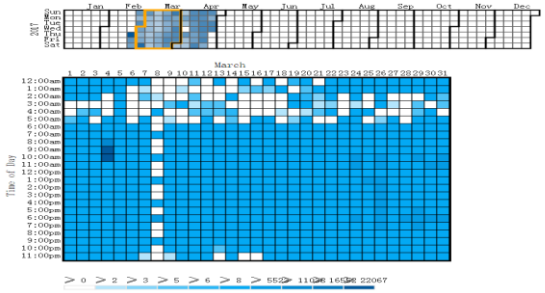


图 2 日历矩阵视图可视化

日历矩阵视图(如图2)是由日历图和矩阵热力图结合而成。该视图展现垃圾短信数量随时间的整体分布规律，该视图横轴为星期的某一天，纵轴为对应的月份，每一个矩形单元代表某一天垃圾短信的数量，颜色深浅编码数量级的多少。通过点击对应月份与矩阵热力图交互联动，细致展示全部垃圾短信的时间分布规律。矩阵热力图的横轴一个月的某一天，纵轴被划分为24h代表具体的时段。我们可以发现伪基站的工作时间主要集中在工作时间段，每天9h-11h相对活跃；在2月23号、3月4号和3月5号，伪基站较活跃，发送了大量的垃圾短信。

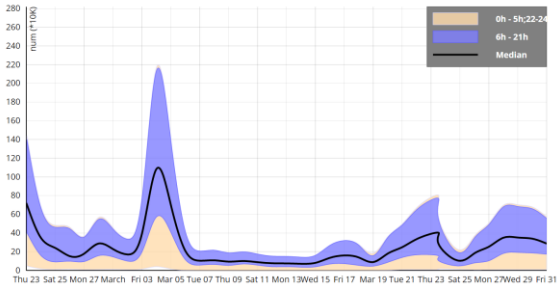


图 3 垃圾短信数量趋势图可视化

趋势图主要呈现垃圾短信数量随时间的变化趋势。为了更好呈现可视化效果，我们在数据初探后将原始数据中每天的各个时段分为热门时段（6h-21h）和非热门时段（0h-6h, 22h-24h），分别通过紫罗兰色和藕荷色编码。趋势图的横轴为具体的每一天，纵轴为垃圾短信的数量，条带的面积越大，表示伪基站越活跃，发送的垃圾短信数量越多。图中的黑色趋势线为总数量的中值趋势线，其落在哪个颜色的条带内，则表示伪基站在一天中的哪个时间段更为活跃。例如，2月23号和3月4号两天伪基站异常活跃，在活跃时间段的垃圾短信发送量更大。

- *陈斌 东北师范大学 研究生 E-mail:82426409@qq.com*
- *栾剑 东北师范大学 研究生 E-mail:2509394085@qq.com.*
- *徐劭斌 长春工业大学 本科生 E-mail:2233935216@qq.com*
- *汤雅兰 东北师范大学 本科生 E-mail:1498926641@qq.com*
- *张慧杰 东北师范大学 副教授 E-mail:zhanghj167@nenu.edu.cn*

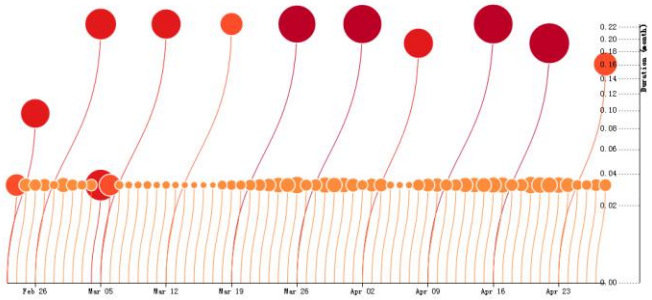


图 4 气球图可视化

此外，我们利用气球图（如图4）主要呈现垃圾短信数量随时间段的变化趋势，横轴为每一天，纵轴为时间跨度，垃圾短信的数量大小用圆编码，半径越大颜色越深则说明该时间段垃圾短信数量越多。我们分别对每一天和每个星期内垃圾短信进行可视化呈现，曲线的弯曲程度表示时间跨度大小。通过该视图发现在3月12-19号和3月19-26号这两周内，伪基站较活跃，并发送了大量的垃圾短信，人们受到垃圾短息的严重骚扰。

1.2 探索不同类型垃圾短信的时空分布规律

如图5所示，我们将原始数据按照短信内容分为：“A”发票短信骚扰（开发票等信息），“B”电信诈骗（银行卡等信息），“C”生活服务骚扰（黄色短信等信息）和“D”其他垃圾短息骚扰。并利用堆叠鱼刺图呈现四种垃圾短信类型的数量变化趋势，其包含展现每个垃圾短信类型总数量的鱼刺图和呈现每天垃圾短信总数量的堆叠图。其中该视图横轴代表某一天，纵轴为垃圾短信的数量，不同的颜色编码不同的垃圾短信类型。通过堆叠效果（如图5），能够清楚地看到每种垃圾短信类型占全天垃圾短信总数量的比例。

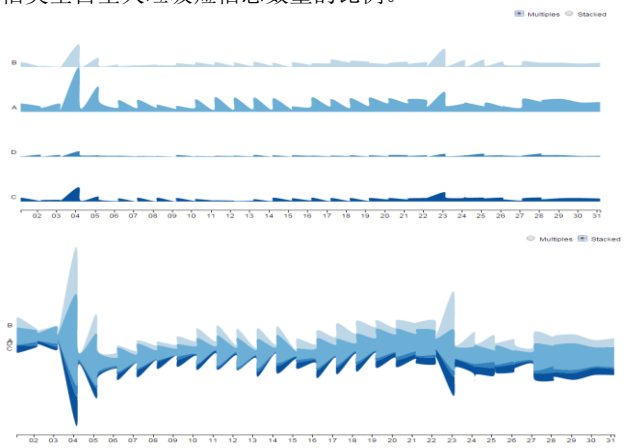


图 5 堆叠鱼刺图可视化

径向弹壳图（如图6）主要呈现发票类短息占总数的比例。其中，日期按照径向分布，每一个弹壳的大小代表当天垃圾短信的总数量。圆环分为两个部分，一部分代表发票垃圾短信类型，另一部分代表其他垃圾短信类型。弹壳在哪个圆环的区域多，则说明对应类型的短信数量较多，反之数量较少。颜色由蓝到深红编码该短信与开发票垃圾短信类型的相关性，蓝色为正相关，深红色为负相关。从图中可以看出，每天存在大量的发票垃圾短信，但在3月6号该类垃圾短信数量比较少。

矩阵图（如图7）：我们通过矩阵图，来发现各种类型垃圾短信数量随时间的变化情况。该视图横轴为每一天，纵轴为具体的时间段，每一个矩形单元代表该类垃圾短信数量的多少。另外，我们还可以利用该视图呈现发送该类型垃圾短信的频繁时间段。

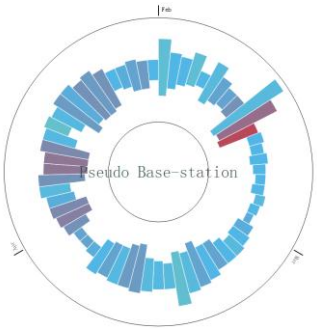
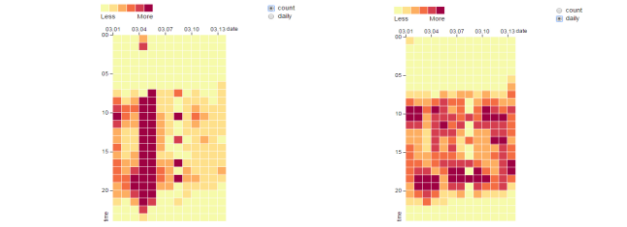


图 6 径向弹壳图可视化效果

进一步结合该类型垃圾短信的地图视图、堆叠鱼刺图和径向弹壳图可视化的结果，我们发现：

- 每天在9h-20h时间段内，发送开发票类短信的伪基站比较活跃；3月4号-5号，开发票类型的伪基站发送了大量的垃圾短信；
- 东三环国贸商圈、大望桥附近成为重灾区，每天都有众多的伪基站在这里出现。



（a）发票骚扰短信的矩阵图可视化　　（b）发票骚扰短信频繁的时间趋势

图 7 发票骚扰类型短信的矩阵图可视化效果

2 实用性与可交互性

在选取可视化方案时，我们以是否满足题目需求为出发点，优先考虑可视化方案的实用性。例如，在展现垃圾短信数量随时间分布规律的日历矩阵图中，支持通过点击日历图与矩阵图进行交互联动，从而细致展示全部垃圾短信的时间分布规律。

我们使用的所有可视化方案都支持交互操作，使用户能够更好地参与对数据的理解和分析，以及查看无法同时展现的数据细节，并从大量信息中选取需要的信息进行深入探索。

3 讨论

基于本文设计的可视化方案，我们对比赛数据进行了可视分析，并根据题目要求对所提出的问题进行了深入讨论。图1反映了伪基站影响区域的空间规律，并提供热力效果和马赛克效果的展示。然后，依据LDA主题模型将短信内容具体分类，并结合地图视图、堆叠鱼刺图和径向弹壳图等多种可视化手段进一步分析伪基站发送不同类型垃圾短信的时空分布规律。最后，根据得到的伪基站行为模式，向执法人员提出整治伪基站的有效建议和方案。

4 结论

针对本次挑战赛的题目要求，我们选取的可视化方案无论从实用性还是可交互性角度，都能够达到完成任务的目的，并获得了很好的效果。此外，这些方案都具有可扩展性，能够在已有图形的基础上，扩展其他的功能或者结合其他的可视化手段来丰富可视内容，从而引导用户发现更多的有价值信息。

参考文献

[1] Chen S, Yuan X, Wang Z, et al. Interactive Visual Discovering of Movement Patterns from Sparsely Sampled Geo-tagged Social Media Data[J]. IEEE Transactions on Visualization and Computer Graphics, 2015, 22(1):270-279.

[2] 陈为,沈则潜,陶煜波. 数据可视化[M]. 电子工业出版社, 2013.