挑战二的可视化方案特点分析

高强，黄健，周欣娜，刘昕蕊，姚羽，东北大学

摘要-本文针对 2017ChinaVis 挑战赛第二题的内容，构建了一套完整、可交互性强的网吧信息的可视化分析系统。采用地图展示了违规网吧的地理位置，利用散点图和热力图等分析了流动人口和未成年人上网行为，利用折线图和拓扑图寻找可能存在的团伙，多维度多形式地分析了全市网吧的上网信息。

关键词-数据清洗、可视化、Echarts、黑网吧

1、数据清洗

分析所给数据，找到几种类型的脏数据，并对其进行清洗。

脏数据的类型主要包含以下几种：(1)人员出生日期 Birthday 不合理(2)登记的网吧信息与上网记录不匹配(3)上下线时间不合理(4)区域代码不合理(5)姓名不合理

2、主要的可视化方案

挑战二所提供的数据主要有:网吧名称及其对应的上网记录（上网用户的姓名，生日，区域代码和上下线时间）。在对所有数据进行清洗之后，我们主要通过分析上网用户的年龄和上下线时间计算出的单次连续上网时长来确定或怀疑有非法经营现象的网吧，并利用百度地图进行直观准确的可视化展示。

同时，利用 Echarts 提供的折线图、散点图和扇形图等多元的统计图，对流动人口、青少年团伙和未成年人上网等进行了简单的用户画像，整体效果如图 2.1 所示。

图 2.1 重庆市网吧可视化分析系统

具体的可视化实现，将在下文一一进行说明。

2.1 非法经营现象

我们把有非法经营现象（接纳未成年人上网）的网吧展示在地图上，如上图 2.1 所示，蓝色代表直接接纳未成年人上网的网吧，黄色代表违规利用某些成年人身份证号帮助未成年人上的网吧，点的大小代表违规上网人数的多少，红色是手动标记的地点。

具体分析过程如下：

•高强，东北大学，1045010454@qq.com，队长

•黄健，东北大学，435219253@qq.com

•周欣娜，东北大学，1020653488@qq.com

•刘昕蕊，东北大学，461593398@qq.com

•姚羽，东北大学，yaoyu@mail.neu.edu.cn，指导老师

① 直接接纳未成年人

部分网吧无视法律法规，直接接纳未成年人上网，在地图上以蓝色圆点表示。

② 根据年龄进行分析

我们设定 40-60 岁年龄段且连续上网时间大于 72 小时或 60 岁以上连续上网时间大于 72 小时的用户为“被套牌”用户，考虑到这些年龄段人群的身体状况、生活背景和普遍兴趣爱好等因素，我们有理由怀疑这些上网记录所在的网吧是利用这些身份信息接纳未成年人上网的非法经营网吧。

③上网单次时长过久

有一些年龄段正常，但上网时长过久的记录也引起了我们的怀疑。我们设定连续上网时长大于 1000 小时的上网记录为被网吧违规利用的身份信息。

分析非法网吧的维度有多种，但由于所给数据的限制，我们只挑选了三个方面进行分析，得到的非法网吧和非法使用的成年人信息请详见**附录 1**。

2.2 流动人口分析

发现流动人口的核心是分析每个上网用户的区域代码，找到相应城市，但是，我们必须明确重庆市在 1998 年改为直辖市之前的身份证区域代码为 **5102**。

下图 2.2（左）展示了重庆外来人口主要来自的省城信息，与重庆市“东邻湖北、湖南，南靠贵州，西接四川，北连陕西”的地理位置信息相符合且外来男性远多于女性。右图则是外来人口热力分布图。

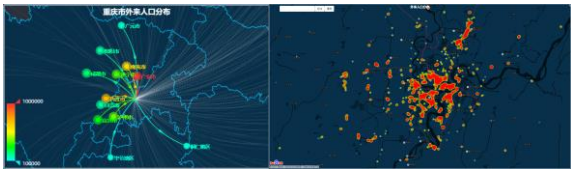

图 2.2 重庆市外来人口省省市分布

图 2.3 是多维动态散点图，数字代表年龄，横坐标是上网人数，可以看出成年之后伤亡人数激增（对数坐标代表成 10 倍增长）。

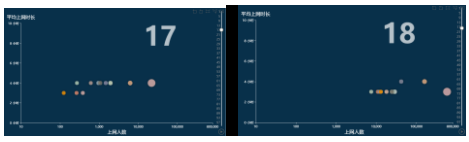

图 4.3 多维动态散点图

2.3 青少年团伙

在确认团伙的部分，我们采取了一种基于可变滑动窗口的数据挖掘方法。

首先，筛选出上网次数大于 5 的用户（所有数据中最大上网次数为 22 次），将这些数据按照上线时间升序排列。

其次，利用可变滑动窗口设置权重并结合 Gephi 得到如下图 2.3 所示的拓扑图，点的大小代表上网次数，线段的粗细代表用户间关系见的强弱（权重），黄色点代表发现的三个团伙。具体分析某一团伙（右图），纵坐标上的每一个数字代表团伙内的一个人，横坐标代表时间，每条折线代表一个团伙成员（用户 id）的上网过程。例如，2016 年 11 月 15 日早 9:30，0 号、4 号、5 号、7 号、8 号和 10 号 6 名团伙成员同时上线（间隔小于 15min）。重叠的部分可以有效且直观地显示出该团伙成员的聚集时间范围，对分析其团伙行为十分有益。

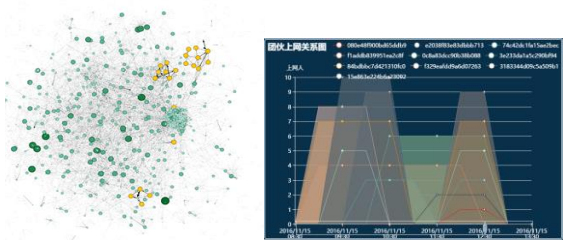

图 2.3 用户关系图

2.4 简单用户画像

综合之前获得的结论，我们可以简要分析违规网吧所处的位置多分布于学校和商业街附近。

我们继续详细地分析用户们的上网高峰时段，根据年龄进行划分。如图 2.4 所示，通过折线图和散点图，分析江北****会所，我们可以看出直接接纳的未成年人的高峰时段为下午 16 时之后。点击一个郊区网吧，如右图所示，高峰时间为下午 20 时之后。

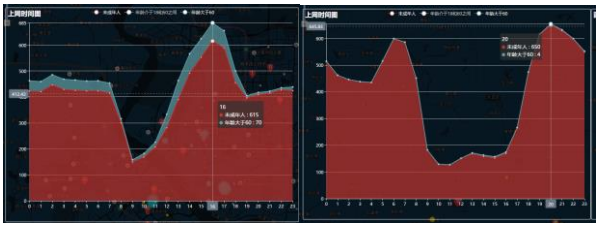

图 2.4 .上网时间图

据此，我们可以分析，由于交通、学校分布、家庭住址等多方面原因，郊区网吧的高峰期到来时间要晚于城区内网吧的高峰期。

同时分析该网吧的人员信息，得到如下图 2.5 所示的扇形图，包括年龄段分类和外来人口分类。

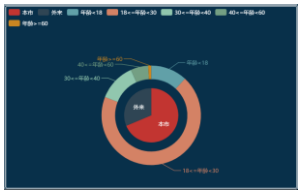

图 2.4 .网吧人员比例图

有关三个月内的日历高峰图，如图 2.5 所示，横坐标代表日期，纵坐标代表星期（第一行代表 Sunday 周日）。将三个月内的每日上网记录数量，按比例显示在图中，选取 Top 12 的点高亮显示。我们可以发现，由于 10 月份有较长的国庆假期，12 月份又临近年底，普遍较忙，日历图也展示了上网的高峰期主要集中在 11 月份的普通周五和周末，这与我们的实际生活规律相符合。

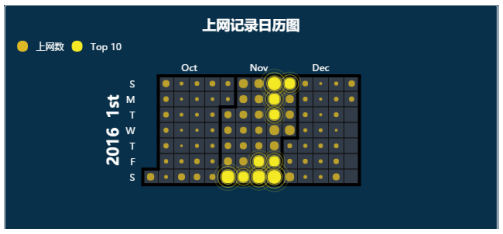

图 2.5 上网日历图

2.5 综合性建议

综合搜集的信息和分析的结果，我们提议公安局加强对学校周边和商区周边网吧的监督力度，对直接接纳未成年人的网吧进行严厉处罚，并对非法利用成年人信息帮助未成年人上网的网吧（如图 2.1 所示）进行调查，并作进一步处理。从我们对非法经营网吧的分析中来看，北城天街周围直接接纳未成年人的网吧最为密集。

根据我们对于上网高峰时间的分析，建议公安局在每天下午 4 时放学/下班之后，加强对商业区和学校周围网吧的查处力度，并引进先进的网吧刷卡系统，如未成年人身份证不能刷卡上网；成年人连续上网超过 5 小时候需要再次刷卡重新上线；对于 60 岁以上的老年人用户限制上网时间等；在防止套牌身份证的同时，也能保证用户们的身心健康。

减少违规经营的网吧数量除了公安部门和网吧营业者的共同努力，还需要全社会的帮助。增强网络安全教育，建立举报机制，对青少年宣传正确的思想价值观念也是减少违规上网行为必不可少的部分举措。

3、结论

本次挑战所选取的可视化方案生动形象且简单直观地完成了挑战赛的要求。同时，该可视化设计方案有着良好的扩展性且已经基本形成一个完整的、可交互性强的可视化分析系统。对于相关部门分析重庆市的网吧信息，提供管理方案有着一定意义。