1276 阅读 0 回复

同小盾

级别: P5_安全攻城狮

加关注 写私信

【反欺诈专栏】关于IP，这里有你想知道的一切！中篇

楼主 发表于: 07-13 只看楼主 更多操作

关于IP的所有研究，都可以归结到三个问题上：

1、这个IP在哪儿?

2、这个IP是什么?

3、这个IP干了什么?

关于IP，这里有你想知道的一切！(上篇)

上一篇中，我们介绍了关于IP地址定位的方法，给出了第一个问题的答案。

前面的文章里，提到了国内的一个数据库，能够给出部分IP地址的精确定位，可以定位到某个学校、酒店甚至网吧。

虽然这份依靠人海战术堆积起来的IP地址库在准确性和时效性上无法满足业务需求，但它也反映出了我们对IP地址研究的期望。我们除了想要知道这个IP的精确位置，我们也希望能够知道IP属主或者类别的信息。

这一篇里，我们就来好好聊一聊这个话题。

这个IP是什么?

数据分析从来都不是盲目的。在开始之前，我们需要事先确定把IP地址划分为哪些类型。

网吧、酒店、学校、商场、企业，这种分类实际上是IP属主的类别划分。在不能准确判断IP属主的情况下，这样分类显然是不合适的。

从风控的角度看，我们对IP进行分类，实际上是为了能够优化风控规则。同一类的IP，风险往往会相同，就可以使用相同的风控策略。

比如，基站IP下用户数量非常大，这类IP上不能使用过于严苛的频次限制策略。

机房IP，比如阿里云、腾讯云、运营商数据中心等等。一般情况下，机房IP都会对应到某一台服务器上去。如果你发现某个用户是通过机房IP访问的，那么代理/爬虫访问的可能性很大。

此外，小运营商会通过租赁的方式，使用三大运营商的网络基础设施。他们所使用的线路，就会从机房IP列表中进行分配(机房IP是保证上下行带宽的，其他类型的IP，一般下行带宽高于上行带宽。专用出口使用机房的线路，可以保证足够的带宽。)

专用出口的IP，往往出现在机房IP的列表中，在不能准确排除专用出口IP的情况下，决不能轻易把机房IP拉黑。

比如下面的这个，根据网络位置判断，是广州市电信机房的IP。但是这个IP上的用户数量非常大，而且用户全部分布在广西境内。万一把这个IP拉黑了，投诉电话会被打爆的。

但机房恰恰是垃圾注册、刷单行为、代理行为、作弊行为和爬虫最密集的地方。如果能够准确地把专用出口这个类型识别出来，那么剩下的，就是具有较高风险的机房IP了。为此，我们根据IP地址上的用户行为特征、设备类型分布等信息来判断识别专用出口IP。

能否通过更多的用户特征来区分其他类型的IP呢？比如，判断一个IP是企业还是家用的宽带。

网吧、酒店、学校、商场、企业等等，这些类别，其实都是IP行为位置分析过程中的副产品。如果一个IP能够精确地定位到某一幢建筑物上，我们只需要判断这个建筑物是什么，就能得出结论。

一般的，企业的网络会使用专线，IP在很长的时间里都不会发生变化。随着定位数据的积累，行为位置就会呈现出密集性。

比如下面的这个IP：

定位点在途牛大厦附近聚集，可以确定这是途牛使用的一个固定IP。与之对应的，我们可以判断，通过这个IP上网的人，应该是途牛的员工。

对于一般的家用宽带，虽然IP会频繁变化，但是在特定的一段时间里，IP会固定的出现在某个区域。

举个例子：

这个IP的定位点并没有像前面的例子那样在某一幢建筑物周围聚集，而是随机地分布在南昌市东湖区靠北的一片区域里。这是一个比较典型的家用宽带IP。

但IP只是业务系统的承载，IP定位的分布，会因为实际的业务而呈现出的聚集形式有非常大的差异。单纯通过定位信息的聚类分析，并不能满足所有IP地址的分类需求。

比如，中国邮政储蓄在某市的营业网点，使用专用线路，IP地址固定。每一个定位点的聚簇，都对应一个营业网点。

这个IP下的用户，除了营业网点的工作人员之外，还会有大量到营业厅办理业务的用户。

如果拥有足够的定位数据作为支撑，理论上是可以准确判断这些IP的属主的。

但是这种分析方法要求定位信息有比较高的准确性、时效性和数量级，可不是每家公司都有能力去尝试。

而且，中国范围内共有2.5亿活跃IP，一个月的时间里，平均每个IP会关联上万定位信息，然后做聚类分析。

这个数量级，光想想就觉得可怕.....应该有更简单的办法才对。

为了讲解地更通俗易懂，这里援引《死亡笔记》中的一个片段。

根据作案时间的分布，推断出了作案者是一个学生(作者：都是因为老师布置的家庭作业太少了！)

我们分析IP的方法，和L的分析如出一辙。

如果一个IP是对应某家公司，这个IP下的用户行为，就会呈现出非常明显的工作日和工作时间的密集性，大家都是朝九晚五的上班族，都懂得哈~~

那么反过来，晚上6点以后，以及双休、节假日比较活跃的IP，就应该是普通的家用宽带。

此外，不同类型的IP，对应的用户数量会有所差异。

最简单的，一般基站的覆盖范围是3~5公里(可能存在多个基站公用同一个IP的情况)，那么同一时间内，每个基站IP下面的用户数量可能会超过1~10万。而家庭宽带的IP，一般一个IP对应一户人家，人数在10人以内，某些小规模的营业场所，也会使用宽带的方式来提供网络连接，人数也会在100人以内。

根据这些特征，就可以把不同类别的IP逐步区分出来。最终，形成了今天我们同盾IP地址分类的全部：

编号	类型名称	用户群体	用户基数（一天内）
1	高等院校	主要是大学生	1万~10万
2	普通基站	移动设备用户	1万~5万
3	基站公用出口	移动设备用户	10万~100万
4	普通机房	没有正常用户	<100
5	专用出口	二级运营商/运营商NAT/特定的专线业务	10~50万
6	企业宽带	企业办公人员	50~2000
7	公共场所	没有固定人群	1000~10万
8	家庭宽带	普通家庭/小规模营业场所	5~100

总结一下：

教育网、基站、机房，目前都有比较完整的IP地址列表，通过简单的匹配就可以得出结论。

再根据用户的在不同时间段内的活跃情况，以及每个IP下的用户数量，我们能够准确判断出是家用宽带，还是企业的固定线路。

虽然到目前位置，我们的模型还不能准确区分一个IP到底是酒吧、网吧、酒店或者医院。但从风控的角度而言，我们目前的分类，已经满足绝大部分业务需求。

当然，我们不会满足于今天的成就，对于IP地址分类的研究，一直在进行，我们的覆盖率和准确率也在不断提升。如果你所在的行业，对IP地址类型的判断有更高的要求，不妨和我们分享一下。

我们乐于接受任何能够完善IP画像功能建议！