**06/2021-08/2021**

**数据开发实习生**

**PCI佳都/佳都科技/产品业务群**

**智慧城市产品事业部/智慧城市产品研发中心/智慧城市产品研发数据产品线**

**职责即是负责人车码融合数据的处理和Spark，Hive等大数据基础建设和平台改进。**

**大数据项目部分**

**项目名称：**融合分析系统

**开发环境：**IDEA+maven+scala+jdk+linux

**技术组件:** Spark+Hive+Hdfs+Yarn+Mysql

**项目描述：**

公安视频图像智能化建设是公安智慧城市中一个核心的部分，通过建立视频监控、人脸抓拍、车辆卡口、电子围栏、智能门禁等视频图像数据与公安业务数据的融合数据体系，实现视频图像多要素解析比对和多维度关联分析，提升挖掘视频图像价值信息的能力，最终为态势分析和风险评估等警务决策活动提供循证支持。负责离线弱关系分析模块，实现人脸、车牌号、手机号、IMSI、IMEI等要素之间的关系分析，即通过机器学习算法建立关联关系，数据挖掘后赋予对应的相似度或置信度。

开发步骤：

1. 数据分布统计（以车为例）

使用SparkSQL从Hive读取所有车辆卡口的抓拍数据，利用SparkRDD根据不同的拍摄设备ID对车辆ID进行分类排序，然后在各个设备内进行抓拍时间排序，根据设定的阈值析出同行车辆，如果车辆被抓拍时间差小于设定间隔值（此处设定为30秒），这些满足条件互相关联的两辆车辆关联数据即认为是同行车辆。

1. 数据挖掘

基于初始同行数据，分别用基于轨迹的LCSS和Sparse DTW算法进行迭代计算同行相似度，可选算法还有基于特征权重的熵权法和关联规则FP-growth算法，最终赋予每条同行数据相对应的同行相似度。

1. 融合分析

模型融合算法成功的关键在于能保证弱模型的多样性，融合不稳定的学习算法能解决单个模型过拟合或者准确性不高的问题。融合基础算法包括均值法，投票法，此处采用排名平均法，将同行相似度预测结果进行排名，然后在标准化平均排名在0到1之间后，得出一个均匀分布的预测，最后将关联度最高的一组同行车数据存入Hive当中。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 起止年月(MM/YYYY-MM/YYYY) | 职位 | | 所在公司/单位/项目名称  （如知晓单位英文名称请填入） | 所在部门 | 地址（具体到城市的名字即可） |
| 06/2021-08/2021 | 数据开发实习生 | | PCI佳都/佳都科技/产品业务群 | 智慧城市产品事业部/智慧城市产品研发中心/智慧城市产品研发数据产品线 | 广东省广州市 |
| 职责及获得的成就（可分点描述） | | 融合分析系统：  公安视频图像智能化建设是公安智慧城市中一个核心的部分，通过建立视频监控、人脸抓拍、车辆卡口、电子围栏、智能门禁等视频图像数据与公安业务数据的融合数据体系，实现视频图像多要素解析比对和多维度关联分析，提升挖掘视频图像价值信息的能力，最终为态势分析和风险评估等警务决策活动提供循证支持。  责任描述：   1. 利用linux的shell命令运行脚本和部署大数据集群（包括spark, hive等组件） 2. 通过Spark RDD和Spark sql对人车码等指标进行多维度统计分析（例如基于轨迹的LCSS和Sparse DTW算法计算人车码同行相似度并算法融合分析） 3. 编写SQL语句对mppdb数据库中的数据进行查询，统计等相关数据操作。 4. hugegraph图数据库部署测试和开发 | | | |

**实习体会：**

作为佳都科技公司的大数据开发实习生，实习期间学习了包括Spark,Hive,Kafka在内的一系列大数据组件，并应用在具体的业务分析上。项目是通过建立视频监控、人脸抓拍、车辆卡口、电子围栏、智能门禁等视频图像数据与公安业务数据的融合数据体系，实现视频图像多要素解析比对和多维度关联分析，提升挖掘视频图像价值信息的能力，最终为态势分析和风险评估等警务决策活动提供循证支持。

第一个功能模块是对一段时间里的人车码数据进行分布统计，当时遇到一个困难就是，当统计得出结果时，需要比对重新存入数据库中，数据库中的原表包含数亿级别的数据，当涉及到表与表之间的join操作，会产生笛卡尔积的两两关系矩阵，时间与空间复杂度呈指数级别增加，十分影响执行效率，最终通过分区和使用高效查询的hbase数据库来实现优化。

数据产品部门的一个核心功能则是关联关系分析，当我拿到通过摄像头和车闸卡口捕获的数据时，我需要用大数据计算工具对数据进行轨迹相似度分析，与同事进行商讨，最终采用基于轨迹的LCSS和Sparse DTW算法和关联规则FP-growth算法分别进行人车码同行相似度分析，由于单个模型存在着过拟合或者准确性不高的问题，最终还需要将不同的相似度结果融合起来。常用的融合基础算法有均值法，当平均多个来自不同模型的输出时，各自的权重总是难以分配，导致最终的预测结果十分混乱，改用排名平均法，将同行相似度预测结果进行排名，然后在标准化平均排名在0到1之间后，得出一个均匀分布的预测。

通过此次项目，我对大数据工具的应用变得更加熟练，也使我意识到了数据分析处理的重要性。当学习基于FP-tree的关联规则挖掘FP-growth算法时，我开始慢慢接触数据挖掘，逐渐让我想要深入学习统计学，人工智能，机器学习等方法。使用分析工具，从海量的数据中，挖掘出未知但有价值的信息和知识从而帮助业务运营甚至企业做更好的决策时，我认为这是十分有意义的。