**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），将这些满足条件的两辆车辆关联数据预处理后收集起来。

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，最终计算出具有相似轨迹的不同车之间的同行相似度。

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. 部署hugegraph图数据库和模拟数据测试开发 | | | |

**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，最终通过将数据集进行分区的方式，以及采用可以指定版本和索引查询的hbase数据库来实现优化，可以提高并行计算速度。

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。数据挖掘就是从大量的数据中，通过统计学，人工智能，机器学习等方法，挖掘出未知但有价值的信息和知识的过程。我还认识到数据挖掘解决的四类问题：分类、聚类、关联和预测问题，我对关联规则中的Apriori算法和FP-growth算法进行了深入研究，这也正是我的项目中所用到的算法，当这些复杂的算法与我的真实业务结合起来的时候，挖掘出的信息能帮助公安智慧做决策时，这是十分有意义的。我希望学习更深的数据科学方面的知识，成为一位数据科学家，利用我手中的数据和工具，为企业，为社会创造价值。

项目中的问题：

1. 数据来源 数据切分
2. 遇到的困难 为何算法只用两个 大表join笛卡尔积怎么解决 spark配置参数如何确定 算法准确率如何保证 涉及到机器学习吗
3. 一般会做什么
4. 数据已经由前面负责etl的部门进行清洗过了，存在mppdb里之后再放入hive中。
5. 算法只用两个的原因是LCSS算法和Sparse DTW算法已经基本满足准确率的要求，采用更多的算法融合得出结果会很耗时，影响运行效率。
6. 多个大表的join语句数据量太大，运行速度慢，也容易超出内存上限，采用hbase，可以将数据字段进行拆分，只选择需要的那部分字段数据，提高效率。
7. Spark参数一般不用修改，采用瑞帆提供的大数据平台，已经是默认默认最好的配置参数。
8. 数据处理过程中容易导致数据倾斜，三个Executor只有一个在运行程序，数据都到一个分区去了，导致OOM等问题，要及时发现问题将数据进行分区，每台机器分配一个CPU或者3个CPU最终就有3个分区或9个分区，提高并行运算速度。
9. 数据开发岗位现在其实主要有三个不同方向，一个是偏业务的统计数据分析师，用hive用sql语句然后生成报表辅助决策，一个是传统etl清理数据岗位，而一个则我们现在做的根据业务，结合算法的思想，对已经结构化的数据进行挖掘处理实现相应的功能。