Operations Research and Fuzziology 运筹与模糊学, 2022, 12(2), 198-203
Published Online May 2022 in Hans. http://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2022.122020

Hans汉斯

景点车流量大数据监测分析及相关预测设计

徐　慧，宋金婷，王兆明

济宁学院，数学与计算机应用技术学院，山东　济宁

收稿日期：2022年3月23日；录用日期：2022年4月30日；发布日期：2022年5月7日

摘　要

游客在景点旅游时的时空信息和行为信息都被蕴含在所获取的多源异构旅游数据中，为了更好地对景点车流量大数据进行监测分析及预测，本文首先对比元旦假期前后车牌数据，进行去重处理后即可对来往车辆是否为自驾游车辆进行判别，然后提出了基于HDFS的景点交通流数据存储，并利用已知模型表示景区内交通流量和密度的关系并提出相关建议。最后，采用K近邻非参数回归等算法来预测短时交通流。

关键词

HDFS存储，大数据分析，Matlab，KNN预测

Big Data Monitoring Analysis and Related Prediction Design of Scenic Spot Traffic Flow

Hui Xu, Jinting Song, Zhaoming Wang

Academy of Mathematics and Computer Application Technology, Jining University, Jining Shandong

Received: Mar. 23rd, 2022; accepted: Apr. 30th, 2022; published: May 7th, 2022

Abstract

The tourists' time-space information and behavior information in scenic spots are contained in the obtained multi-source heterogeneous tourism data. In order to better monitor, analyze and predict the big data of traffic flow, this paper first compares the license plate data before and after the New Year's Day holiday, and can judge whether the passing vehicles are self-driving vehicles after De-duplication processing, then the data storage of scenic spot traffic flow based on HDFS is proposed, and the relationship between traffic flow and density and some relevant suggestions are put forward. Finally, K-nearest Neighbor Nonparametric regression algorithm is used to predict

文章引用: 徐慧, 宋金婷, 王兆明. 景点车流量大数据监测分析及相关预测设计[J]. 运筹与模糊学, 2022, 12(2): 198-203. DOI: 10.12677/orf.2022.122020

short-term traffic flow.

Keywords

HDFS Storage, Big Data Analysis, Matlab, KNN Prediction

1. 引言

为了使景点更好地对来往车辆进行统筹调度[1]，同时也为了让游客更好地安排旅行行程，多源异构数据的出现可以随时提供景区实时热度的相关数据[2]。有了各式各样的数据就需要存储和计算处理，比如：利用 HBASE 进行分布式存储以及运用 MapReduce 进行并行化处理[3]。有了更为简洁的数据之后，便可以建立模型进行预测[4]，之前苏培培对智慧黄山风景区客流量预测系统已进行了详细分析，介绍了支持向量回归(SVR)和 BP 神经网络这两种预测模型[5]，另外，还有文章针对三种情况下的客流分别进行分析建模[6]，这些为旅游景区短期客流量预测工作[7]提供了相关的方法。本篇文章中首先对获取的海量数据进行存储、得出景区内交通流量与密度的关系、并利用历史数据对短时交通流进行预测。我们的研究内容不仅囊括了数据存储、数据处理、数据预测还做出了基于预测结果的预测，设计范围较为广泛。其结果不仅有利于提醒景区内部工作人员及时进行合理调度，而且有利于即将出行的自驾游游客对旅游行程的妥善安排。

2. 自驾游车辆判别

由于旅游交通和日常交通共用城市道路交通系统，采集到的数据中同时包含了自驾游车辆和工作通勤车辆，因此，需要对预处理后的数据进行自驾游车辆的判别[8]。判别方法如下：

1) 将元旦假期前后的车牌数据进行对比，元旦期间和其他时间都有在景区内出现的车牌视为本地车牌，予以剔除，可初步得到自驾车数据，具体过程为：

① 选取元旦假期内的数据，将这些数据中的车牌去重；

② 选取元旦假期以外的数据，将这些车牌号去重；

③ 选取①和②中的车牌数据的交集，即可得本地车牌数据；

④ 在预处理完成的数据中，剔除③中所得的数据，即得到初步被判断为自驾游车辆的数据；

2) 分别对 1)中得到的本地车牌数据和待定的自驾游车辆的车牌数据按照车牌号进行分组统计，遍历分组结果中的时间戳，应用 Python 语句处理日期和时间的标准库 datetime 来比较时间戳的大小，最小值为车辆在景区各监测点第一次出现的时间，最大值则为最后一次出现的时间，将其分别作为开始时间和结束时间，计算自驾游车辆的停留天数。如式(1)、式(2)。

$$\text{stay}_{\text{\_time\_id}} = \left\lfloor \text{end}_{\text{time\_id}} - \text{start}_{\text{time\_id}} \right\rfloor * \text{day} \tag{1}$$

$$\text{stay\_time}_{\text{ratio}_n} = \frac{\text{id\_count}_n}{\sum_{i=1}^{N} \text{id\_count}_i} \tag{2}$$

式(1)中，$\text{stay}_{\text{time\_id}}$ 为第一次出现在景区各监测点的时间；$\text{end}_{\text{time\_id}}$ 为最后一次出现在景区各监测点的时

间；id 为车牌号。如果 $stay_{\_time\_id}$ 为 1，则代表该车停留时间不超过 1 天。式(2)中，$stay\_time_{ratio_n}$ 表示停留 n 天的自驾游车牌量占总车牌量的比例；id_count_n 表示停留 n 天的车牌数量；N 为最大停留天数，这里取 31 天。其中，式(2)是为了验证式(1)所提取本地车牌和我们待定自驾游车牌的准确性，故统计了各个停留天数下的车牌所占比例情况。

3. 自驾游车辆数据实现 HDFS 存储

当筛选出自驾游车辆数据后需对其进行存储，但由于数据量的庞大和数据类型众多，故我们需要有能够稳定存储 GB，TB 级别以上的数据文件，而 HDFS 存储刚好能满足我们这样的需求。因为 HDFS 无需引用任何特定的存储中心就可以建立一个数据共享网络，能够实现数据的永久存储，并且能够预防数据丢失。其次任何数据都是“上链可溯源”。它的另一个优势就是将一条完整的数据链进行分片处理，再将分片数据保存在一定比例的节点中，具有可无限拓展的数据存储能力。

HDFS 的写入操作：

准备前提：① 文件 Data，200 M 大小。客户端将文件 Data 写入到 HDFS 上。② HDFS 按默认配置。③ HDFS 分布在三个机架上 Rack1，Rack2，Rack3。

步骤：

1) 客户端将文件 Data 按 64M 分块。分成四块，block1、block2、block3 和 block4；

2) 客户端向 NameNode 发送写数据请求。

3) NameNode 节点，记录 block 信息。并返回可用的 DataNode。

block1: host2,host1,host3

block2: host7,host8,host4

block3: host5,host9,host6

block4: host11,host10,host12

4) 客户端向 DataNode 发送 block1；发送过程是流式写入。

流式写入过程：

① 将 64M 的 block1 按 64k 的 package 划分。

② 然后将第一个 package 发送给 host2。

③ host2 接收完后，将第一个 package 发送给 host1，同时客户端向 host2 发送第二个 package。

④ host1 接收完第一个 package 后，发送给 host3，同时接收 host2 发来的第二个 package。

⑤ 以此类推，直到将 block1 发送完毕。

⑥ host2,host1,host3 向 NameNode 发送通知，且 host2 向客户端发送通知，即“消息已发送完毕”。

⑦ 客户端收到 host2 发来的消息后，向 NameNode 发送消息，即“我写完了”。

⑧ 发送完 block1 后，再向 host7，host8，host4 发送 block2。

⑨ 以此类推，直至所有 block 全部写入，才意味着文件 Data 已全部存入 HDFS 中。

4. 景区内交通流量与密度的关系

交通流量的准确检测，在景区内各交通要道的管理中十分重要。一般来说，将在单位时间内通过某个交通路口的交通实体数来表示交通流量，实体通常指机动车、非机动车和行人，但这里我们只针对机动车车流量的研究。这里的交通密度[9]指的是景区内的一条道路上单位面积内某一瞬时存在的车辆数。通常根据某个路口的过往车流量来判定该路口交通的拥堵情况，以此有效管控各个路口交通。图 1 为我们演示的某个路段的交通流量与密度的关系图。

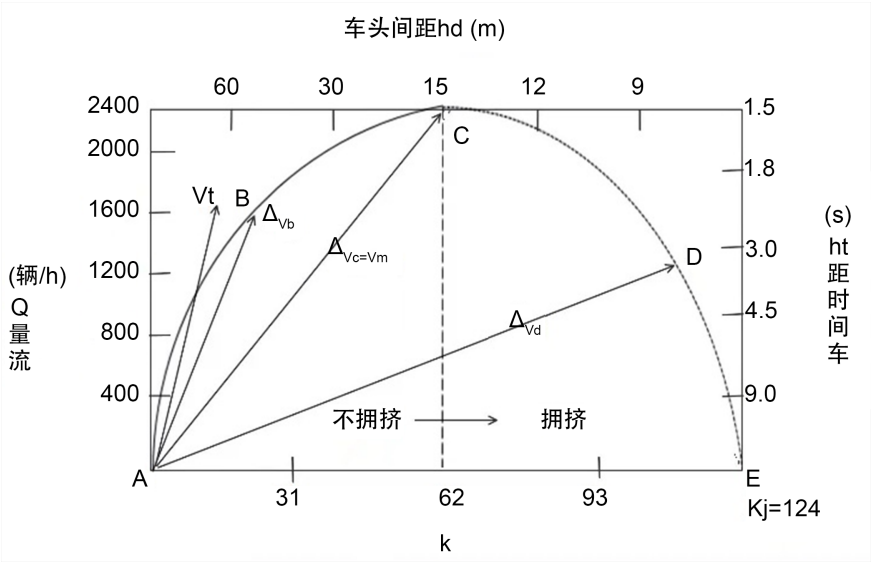

Figure 1. Relationship of traffic density, flow, workshop time distance and distance headway
图 1. 路段的车流密度与流量、车间时距、车头间距的关系图

图 1 中所用公式模型：

$$Q_i = \sum U_i \tag{3}$$

$$K_i = \frac{Q_i}{l_i * S_i} \tag{4}$$

Q_i 表示某一时间间隔内通过 i 路段的交通流量总和(辆)；U_i 表示通过 i 路段的对应小客车的换算系数。K_i 表示 i 路段单位面积内的交通流量密度(辆/m^2)；l_i 表示 i 路段拥有的车道数量；S_i 表示 i 路段上的车道面积(m^2)。

从图 1 中我们可以看出，在景区内某一个路段交通车流量没有达到最大值时，车流密度随着车流量的增长而增大；当交通流量接近或等于最大值时，车头间距逐渐缩小，车速受到限制，出现车辆跟驰，此时景区内交通工作人员应当警惕，防止车辆间发生碰撞；当车流密度大于最佳车流密度后，车辆行驶速度应随着道路交通流量同时逐渐降低，此时车流密度逐渐增大，车辆出现缓慢、匀速低速行驶现象，工作人员应及时做出反应，防止交通拥堵；如果不幸达到这种情况：当车流密度逐渐增大一定值后，车头间距持续变小，车间时距逐渐增大，道路交通发生阻塞，甚至发生停车现象，工作人员应先暂停车辆进入该路段，并尽快有序疏散该路段车辆。

5. KNN 预测设计

定义：如果一个样本在特征空间中的 K 个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别，则该样本也属于这个类别[10]。

优势：KNN 算法和朴素贝叶斯之类的算法比，对数据没有假设，准确度高，对异常点不敏感；在多分类问题上比 SVM 算法好用，因为 KNN 算法主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属的类别，所以 KNN 算法更为适合类域的交叉或重叠较多的待分类样本集；此算法比较适用于样本容量比较大的类域的自动分类。

计算方法[11]：(关键是找到合适的历史值向量维数 m 和近邻个数 K)

1) 计算已知类别数据集中的点与当前点之间的距离；

2) 按距离递增次序排序；

3) 选取与当前点距离最小的 K 个点；

4) 统计前 K 个点所在的类别出现的频率；

5) 返回前 K 个点出现频率最高的类别作为当前点的预测分类。

由于疫情影响，这里采用的仍然是模拟数据集。时间：1 月 1 日；地点：三孔景区。考虑到：① 各个时间段进入景区的车辆数应小于等于剩余停车位总数；② 同时，考虑到景区各个路段的旅客交通方便，故景区内任意路段的车流密度应小于等于流量为峰值时对应的车流密度；

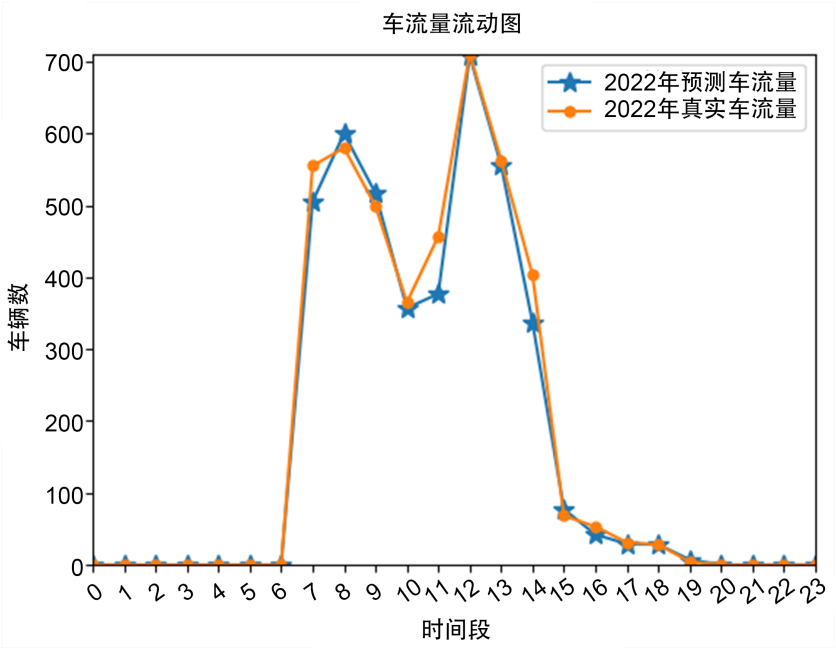

Figure 2. Fitting of predicted and real traffic flow in each period of New Year's Day 2022

图 2. 2022 元旦各时间段预测车流量与真实车流量拟合图

我们最终选取的维数 $m = 24$，最优 $K = 6$ 即选取 2019 至 2021 年元旦的数据集作为样本训练得到预测车辆的数据集，用 matlab 拟合，并且与真实的数据集之间进行误差分析。经过统计，我们发现相对误差和只有 0.00123，绝对误差和只有 0.215，说明取得的预料结果还是相对可靠的。

从图 2 中可以清晰地观察到：1) 从整体分析，这一天的车流量变化是非线性的，有递增状态也有递减状态；2) 从具体分析，在 6 点至 7 点这个时间段内，进入景区的车辆激增，车辆数变化量最大。并且在 7 点至 14 点都属于进入景区的高峰期，其中在 8 点和 12 点都达到了极大值。这些分析表明，在高峰期来临之前，景区内的工作人员应提前就位为高峰期的交通疏通做好准备，防止路段交通拥堵。此外，在 15 点之后，进入景区的车辆逐渐减少直至为零，说明大部分人选择白天到该景区游玩。

6. 结束语

本文通过对景区内自驾游车辆的筛选判别、筛选后数据集的 HDFS 存储、交通流量与密度的关系分析、KNN 算法预测短时交通流，进一步提出了在管理调度时应注意的问题，为景区工作人员和游客双方及时采取更完备的解决方案打下了基础。通过构建模型及求解得出重要时间节点，降低了景区内各路段

的拥堵风险以及自驾游旅客做出选择的时间成本，保证了景区内交通的畅通无碍和游客的旅游愉悦体验。虽然在这其中也有考虑不全面等问题，但也做出了实质性行动，提高了各大景区道路规划和安排调度的效率，使景区内交通通畅度得到具体优化提升。

基金项目

济宁学院省级大学生创新创业训练计划项目：S202010454008；济宁学院教学改革研究项目(项目驱动式的《数学建模》课程实践教学改革研究)。

参考文献

[1] 林树宽, 于伶姿, 乔建忠, 等. 基于 GPS 轨迹数据的拥堵路段预测[J]. 东北大学学报(自然科学版), 2015(36): 1530-1534.

[2] 施元磊. 景区交通流量预测与游客行程规划技术研究[D]: [硕士学位论文]. 西安: 西北大学, 2021.

[3] 朱刘江. 基于 Hadoop 的海量城市交通流数据分布式存储与分析研究[D]: [硕士学位论文]. 扬州: 扬州大学, 2015.

[4] 苏康传, 杨庆媛, 张柏林, 等. 时空协同的城市旅游行程规划模型构建[J]. 地球信息科学学报, 2019, 38(2): 814-825.

[5] 苏培培. 风景区旅游客流量短期预测方法研究[D]: [硕士学位论文]. 合肥: 合肥工业大学, 2013.

[6] 陈荣. 基于支持向量回归的旅游短期客流量预测模型研究[D]: [博士学位论文]. 合肥: 合肥工业大学, 2014.

[7] 王竟成, 张勇, 胡永利, 等. 基于图卷积网络的交通预测综述[J]. 北京工业大学学报, 2021, 47(8): 955-965.

[8] 张瑞敏, 徐红是. 旅游交通研究述评[J]. 桂林旅游高等专科学校学报, 2005(6): 41-44.

[9] 杨贺. 中山陵园风景区节假日交通拥堵治理研究[D]: [硕士学位论文]. 南京: 南京理工大学, 2018.

[10] 闫永刚, 马廷淮, 王建. KNN 分类算法的 MapReduce 并行化实现[J]. 南京航空航天大学学报, 2013(4): 550-555.

[11] 刘彪. 空间数据库中基于 MapReduce 的 KNN 算法研究[D]: [硕士学位论文]. 大连: 大连海事大学, 2012.