北京大学学报(自然科学版) 第 54 卷 第 6 期 2018 年 11 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 54, No. 6 (Nov. 2018)
doi: 10.13209/j.0479-8023.2018.070

Spark框架下交通流数据高效处理方法及其应用

李欣[1,2]

1. 河南财经政法大学中原经济区“三化”协调发展河南省协同创新中心, 郑州 450046; 2. 河南财经政法大学资源与环境学院, 郑州 450046; E-mail: lixin992319@163.com

摘要 设计并实现基于 Spark 的交通流数据处理与预测分析应用框架, 可以完成交通流数据的高效清洗、统计、存储和查询。利用基于多阶空间权重矩阵的 STARIMA 模型进行交通流预测分析, 可以验证数据处理效率及对预测应用的支撑作用。对比实验结果表明: 1) 交通流数据处理框架运行效率高, 适用于复杂的数据清洗和挖掘算法, 为预测模型建立数据支撑; 2) 交通流预测模型对空间权重矩阵进行了多阶优化, 兼顾高效性和准确性, 预测分析结果可以为交通诱导提供参考。

关键词 Spark; 数据清洗; 语义查询; 空间权重矩阵; 交通流预测

中图分类号 P91

Efficient Traffic Flow Data Processing Method and Its Application Based on Spark Framework

LI Xin[1,2]

1. Collaborative Innovation Center of Three-Aspect Coordination of Central Plain Economic Region, Henan University of Economics and Law, Zhengzhou 450046; 2. College of Resource and Environment, Henan University of Economics and Law, Zhengzhou 450046; E-mail: lixin992319@163.com

Abstract A traffic flow data processing and forecasting framework based on Spark is designed, and it can complete the efficient cleaning, statistics, storage and query of traffic flow data. A multi-order spatial weight matrix STARIMA model is used to predict the traffic flow, and it can verify the efficiency of data processing and the support for the prediction. By comparative experiments, the results show that the traffic flow data processing framework is efficient, and it is suitable for realizing complex data cleaning and mining algorithms and establishing data support for the prediction model. The traffic flow prediction model optimizes the multi-order spatial weight matrix, and it takes both efficiency and accuracy into consideration. The prediction results can provide reference for traffic guidance.

Key words Spark; data cleaning; semantic query; spatial weighting matrix; traffic flow prediction

随着中国城市化进程的不断推进, 一些城市管理问题愈发突出, 其中一、二线城市的交通拥堵问题严重影响人们的生活体验, 迫切需要利用大数据技术及空间信息技术来优化交通管理水平, 对交通流进行预测, 诱导车辆出行。

目前, 国内外已经有一些相关的研究成果。在海量空间数据管理方面, Aji 等[1]通过空间划分以及自定义查询引擎, 建立高性能的空间数据仓库系统 Hadoop-GIS; Witayangkurn 等[2]利用 Java 拓扑套件, 在 Post-GIS 中实现空间自定义函数运算; Abouzeid 等[3]设计 HadoopDB, 在 MapReduce 框架中实现多源数据互连; Plugge 等[4]设计 MongoDB, 并通过 Connector 中间件实现高效数据交换; 温馨等[5]通过自定义空间函数下推, 实现分布式空间查询。以上

国家自然科学基金(41501178, 41771445)和河南财经政法大学博士科研基金(800257)资助
收稿日期: 2017–10–11; 修回日期: 2017–12–01; 网络出版日期: 2018–07–16

研究成果在一定程度上提高了空间数据管理效率，但在框架运算能力和数据管理模式方面还有提升和优化的空间。在交通流预测分析方面，比较成熟的研究成果有状态空间模型[6]、时间序列模型[7]、神经网络模型[8–9]和历史平均模型[10]等。一些学者基于这些模型进行了优化研究，如利用路网结构和实时时序数据构建交通状态多点预测方法[11]，利用时空自回归模型[12]模拟路网时空相关性的预测模型[13–15]，利用路口转弯率提高城市交叉路口交通状态预测性能的网络交通状态模型[16–17]，利用路口可达性和路段连通性分析建立路网时空状态表达模型[18]等。以上研究成果建立了基本的交通流预测功能，但应用限制较多，对时空相关性考虑不够充分。随着数据采集技术的发展，对交通流数据处理运算效率的需求也不断地增加。

本文基于 Spark 框架和分布式空间数据库，设计并实现一种高效的交通流数据获取、清洗、统计和查询框架，完成交通流预测分析的前期数据处理工作。同时，对时空自回归移动平均模型 STARIMA 的空间权重矩阵进行优化改进，并用于交通流预测分析，从而验证本文提出的数据处理框架的效率及其对预测应用的支撑作用，为下一步建设城市智能交通管理平台提供理论依据。

1 交通流数据处理与应用总体框架

基于 Spark 的交通流数据处理与预测应用的总体框架如图 1 所示，分为以下两个层次。

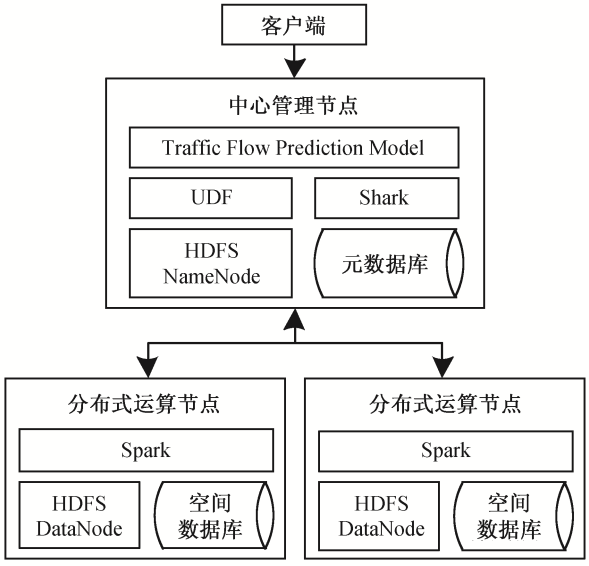

图 1 交通流数据处理与预测应用总体框架

Fig. 1 Traffic flow data processing and prediction framework

1) 分布式运算节点。在利用 Shark[19]完成分布式交通流数据存储的基础上，以 Spark[20–22]作为高效计算引擎，将交通流数据清洗和统计时的查询语句转换为弹性分布式数据集(RDD)进行操作，完成对分布式节点的数据处理。相对于 MapReduce，此方法的优点在于大量低速磁盘 I/O 操作被高速内存存取代替，系统的整体运算效率得到有效的提升。

2) 中心管理节点。分布式节点上经过清洗和统计的交通流中间数据集传输至中心节点后，利用 HDFS 与空间数据库相结合的方式进行管理。基于用户语义查询条件，利用自定义空间函数 UDF 进行高效数据查询，最终使用优化的交通流预测模型生成阶段性预测结果。

基于此框架的交通流数据处理和预测应用流程如下。

1) 基于用户输入的空间范围及其他分析条件，由中心管理节点实现语义查询语句的优化和解译，结合元数据库中的节点信息，完成数据处理初始化任务。

2) 根据数据处理初始化需求，在中心管理节点完成交通流数据清洗和统计工作的任务拆分，并由相应的分布式节点执行分解后的清洗和统计任务。

3) 依据分配到的数据处理任务，使用由孤立点监测算法制定的规则，对各个分布式节点进行数据清洗，同时完成阶段性路网交通流数据统计，并将其传送至中心管理节点。

4) 各个分布式节点的路网交通流数据统计结果传输至中心管理节点后，进行中间统计结果合并，利用基于多阶空间权重矩阵的 STARIMA 模型完成交通流预测分析。

2 交通流数据处理与应用的关键技术

2.1 交通流数据清洗与统计

获取交通流数据的传感设备种类较多，主要有地磁检测器、视频检测器、微波检测器和浮动车 GPS 等，通常通过有线或无限网络进行数据传输，并在分布式节点进行处理和存储。由于设备、网络等突发性故障或系统性的原因，采集到的数据中经常含大量错误、冗余或无效数据，即“脏数据”[23]。为了保证后期交通流预测分析的准确性，必须对此类数据进行纠错清洗。

王晓原等[23]设计了基于孤立点检测算法和阈

值理论的交通流数据清洗方法，可在本文设计的数据处理框架中完成分布式节点上的交通流数据清洗，其步骤如图2所示。

通过实验发现，王晓原等[23]的方法对数据的清洗效果较好，对“脏数据”的识别率可达 90%以上。经过清洗的交通流数据可以更准确地还原实际交通状态，用其进行预测分析的结果会更准确。但是，此方法主要对固定数据集进行运算，且复杂度较高，对分布式海量数据的处理效率不高。

利用本文设计的基于 Spark 的交通流数据处理框架，可以在分布式运算节点上针对有限时段内的交通流数据集部署清洗算法，有效地减少清洗算法在单个节点上处理的数据量。基于 Spark 的高效内存计算，可以有效地提高系统的运行效率。

由于在数据清洗过程中需要对当前时段的数据集进行遍历，对每一条数据记录进行孤立点检测，因此在遍历数据的同时可以完成路网中的交通流信息统计，将其作为中间统计数据集传输至中心节点进行预测分析。

2.2　交通流数据注册与存储

为了提高分布式网络中交通流数据的访问效率，需要在分布式运算节点创建并行服务，将数据存取和清洗运算进行负载均衡配置；还需要在中心管理节点创建数据分区元数据库，用于存储各个分布式节点的数据信息，以便实现数据分区索引，提高存取效率。

图 3 为并行数据存储与注册流程。首先，按照城市路网空间分布及已有交通流量统计信息，设置负载均衡分配规则。由传感器采集的交通流数据经过清洗后，按负载均衡规则进行传输和存储，同时按照交通流预测时间周期，更新分布式节点的元数据信息(主要包括数据的时段、空间范围、数据规模和传感器类型等)，由分布式节点传送至中心节点的元数据库进行注册和存储。

为了保证系统的稳定运行，避免节点状态异常造成的系统故障，还需要进行数据镜像备份。本文使用较成熟的数据复制技术，首先在分布式物理节点创建虚拟节点，在虚拟节点存储经过负载均衡配置的不同数据集，针对不同物理节点上的虚拟节点进行相互备份。系统运行时，可以利用节点探测算法，对所有虚拟节点进行扫描，并进行节点状态实时更新。如果节点状态正常，则提取访问数据；一旦发现异常，则自动切换至镜像备份数据源。交通流数据备份原理见图 4。

2.3　交通流数据语义查询

Spark 虽然可以较大程度地提高大数据运算效率，但其框架不支持地理信息空间查询功能，必须通过自定义空间查询函数 UDF 来实现查询功能。由于语义查询或模糊查询为分布式查询的主要模

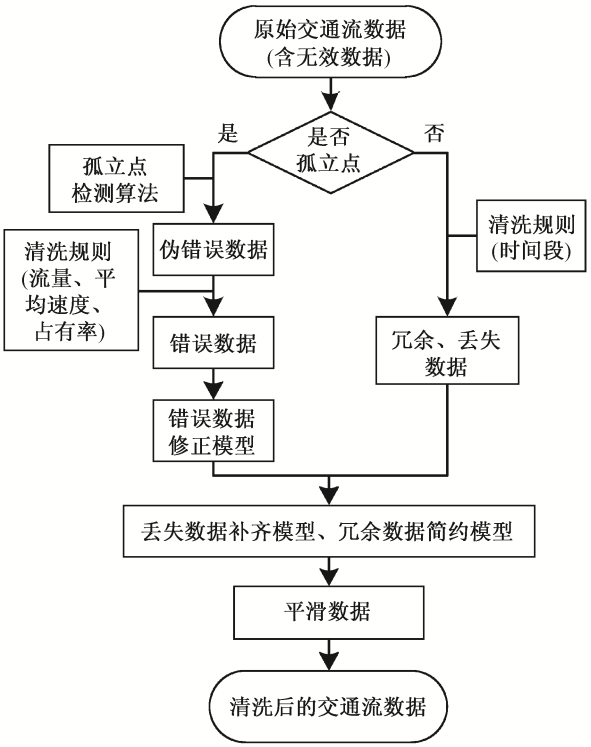

图 2　交通流数据清洗步骤[23]

Fig. 2　Traffic flow data cleaning steps[23]

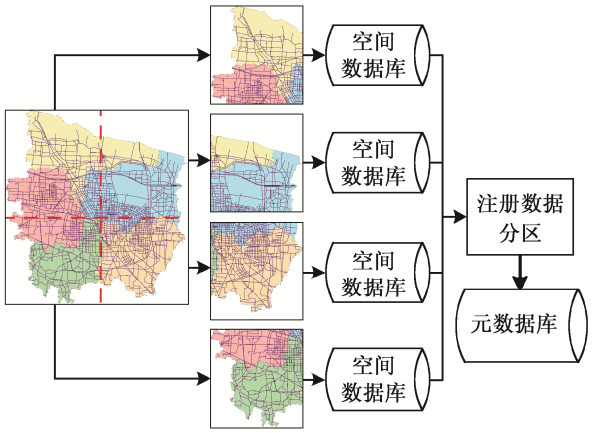

图 3　并行数据存储与注册流程

Fig. 3　Parallel data storage and registration process

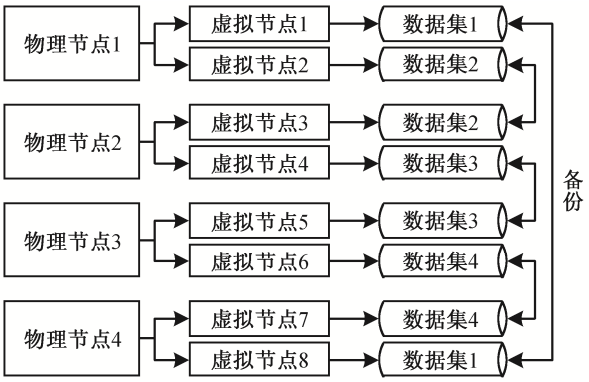

图 4 交通流数据备份原理

Fig. 4 Data backup principle of traffic flow

式, 所以需要对查询语句进行语义解析, 并将查询条件分解为分布式查询任务, 分配至相应节点, 并行完成查询任务后返回结果。

本文实现一套空间查询 UDF 函数库, 并将其进行注册。基于语义的空间查询流程(图 5)如下。

1) 语义解析器: 分析用户输入的查询条件, 按照语义规则对其进行拆分, 按照查询条件的逻辑关系, 将拆分后的基本查询单元组合为查询条件树。

2) 逻辑优化器: 依据查询条件树, 在已注册的自定义空间查询 UDF 函数库中搜索所需函数, 根据优化规则, 将查询逻辑顺序优化, 生成逻辑运行表, 然后进行验证与保存。

3) 物理计划生成器: 读取查询条件及逻辑运行表, 参考分布式节点的元数据信息, 将查询任务转换并分解为分布式节点物理查询操作。

2.4 交通流预测模型

本文设计基于多阶空间权重矩阵的 STARIMA 交通流预测模型, 对数据处理结果进行应用验证。

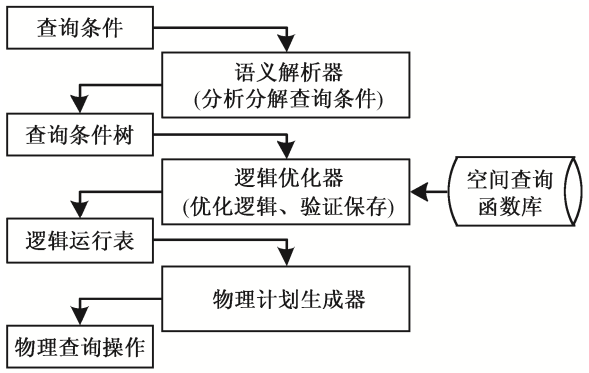

图 5 基于语义的空间查询流程

Fig. 5 Spatial query process based on semantics

在城市路网拓扑关系中, 任意两路段仅存在直接邻接和经过 n 个其他路段间接邻接两种关系, 因此由上游路段到下游路段的交通流分配关系与路口转弯率[24]密切相关, 在仅考虑车辆选择最短路径作为最优路线的情况下, 上下游路段交通流的时空相关性即为经过 n 个路段的多阶路口转弯率重分配关系。结合交通流预测分析的时间周期特征, 设 $R_{ij}=\{l_i,l_1,l_2,\ldots,l_{n-1},l_j\}$ 为上游路段 i 到下游路段 j 的最短路径, 第 k 个交通流预测分析时段 $[k,k+1]$ 内上游路段 i 与下游路段 j 之间的时空相关性表达式为

$$r_{ij}(k)=\theta_{ij}(k-1), \tag{1}$$

$$r_{ij}^{n}(k)=\theta_{l_{n-1}j}(k-1)\theta_{il_1}(k-n)\prod_{p=1}^{n-2}\theta_{l_pl_{p+1}}(k-n-p)。 \tag{2}$$

式(1)中, $r_{ij}(k)$ 为直接邻接路段 i 与路段 j 的时空相关性系数, $\theta_{ij}(k-1)$ 为前一统计时段内二者之间的路口转弯率。式(2)中, $r_{ij}^{n}(k)$ 为非直接邻接路段 i 与路段 j 的时空相关性系数, θ 为最短路径中两个相邻路段的转弯率, p 为取值范围为 1 至 $n-2$ 的自增变量。

从式(1)和(2)可以看出, 路口转弯率 θ 的取值范围为 0~1, 因此时空相关性系数 $r_{ij}(k)$ 和 $r_{ij}^{n}(k)$ 也小于 1, 且最短路径中的中间路段越多, 时空相关性系数的数值越小。经过实验发现, 2 阶以内上下游路段的交通流分配量较多, 相关性高; 3 阶以上路段不仅复杂程度高, 计算量大, 而且交通流分配量少, 相关性低, 因此忽略不计。

根据以上分析, 本文将时空自回归移动平均模型 STARIMA[25−27]的空间权重矩阵与式(1)和(2)进行组合运用, 并且将该模型用于交通流预测分析。STARIMA 模型的计算公式如下:

$$Z(t)=\sum_{k=1}^{p}\sum_{h=0}^{m_k}\phi_{kh}\boldsymbol{W}^{(h)}Z(t-k)-\sum_{l=1}^{q}\sum_{h=0}^{n_l}\theta_{lh}\boldsymbol{W}^{(h)}\varepsilon(t-l)+\varepsilon(t), \tag{3}$$

其中, m_k 为第 k 个时间自回归项的空间间隔, n_l 为第 l 个时间移动平均项的空间间隔, ϕ_{kh} 是时间延迟为 k 且空间间隔为 h 的自回归参数, θ_{lh} 是时间延迟为 l 且空间间隔为 h 的移动平均参数, $\varepsilon(t)$ 为随机误差, $\boldsymbol{W}^{(h)}$ 为 h 阶空间权重矩阵。

依据 Pfeifer 等[25]对权重矩阵 $\boldsymbol{W}^{(h)}$ 的限制条件, 结合以上时空相关性分析, 该矩阵元素表达式为

$$w_{ij}^{l}=\begin{cases} r_{ij}^{l}(k), l=1, \\ r_{ij}^{l}(k)\Big/\sum_{i=1}^{N} r_{ij}^{l}(k), l>1, \end{cases} \tag{4}$$

其中，w_{ij}^{l} 为矩阵中第 i 行第 j 列元素值，$r_{ij}^{l}(k)$ 为路段 i 和 j 的 l 阶时空相关性系数。

式(4)是在路网拓扑关系的基础上，较充分地考虑了交通流的实际多阶重分配规律，基于式(4)进行的交通流预测分析结果会更准确。

3　实验与分析

3.1　实验数据与运行环境

实验数据来自智能交通管理系统(包含车辆监控、调度指挥、交通监测和流量诱导等子系统)。在郑州市交通管理系统中，通过视频、微波、地磁和 GPS 采集交通流数据，系统运行日均数据增量约为 2000 万条。本文实验数据为 2016 年 12 月 5 日至 12 月 18 日(共计 14 天)的郑州市部分区域交通流数据。

实验中使用 5 台配置相同的服务器，配置均为 Intel XeonE5-2640 2.6 GHz, 8 核, 16 GB 内存。中心管理节点使用 1 台服务器，完成分布式节点元数据管理、用户语义查询任务分配以及交通流周期预测分析等工作。分布式运算节点使用 4 台服务器，完成交通流数据清洗、统计、存储和备份工作。

3.2　数据处理实验

由于交通流数据是由多元传感器实时采集，会随时间推移海量地增长，因此必须验证本文交通流数据处理框架中的清洗、统计、存储和查询效率能否满足预测应用的需求。

根据 Min 等[16]的研究结果，交通流预测分析的时间周期以 15 分钟为宜。按系统日均采集 2000 万条数据计算，15 分钟的平均数据量为 21 万条。考虑到高峰时段的交通流量远大于非高峰时段，兼顾系统的可扩展性，针对本文数据来源，数据处理速度应达到每 15 分钟 50~100 万条。

本文除使用基于 Spark 的方法处理交通流数据外，还使用基于 MPI 的方法和基于 MapReduce 的方法，并对比 3 种方法的数据处理效率。

1) 基于 MPI 的方法[28]。在 4 个分布式节点各配置 1 个从进程，由其并行实现多源交通流数据清洗和统计，并按照 15 分钟的时间周期将数据更新至主进程。主进程配置在中心节点，实现中间统计信息合并汇总，并使用预测模型进行交通流预测。

2) 基于 MapReduce 的方法[13]。将交通流数据按照负载均衡规则平均分为 48 个子集，在 4 个分布式节点设置 48 个 Map 运算和 4 个 Combine 运算，每个 Map 运算对应 1 个数据子集进行数据清洗，Combine 运算完成本节点数据统计并将其传输至中心节点，中心节点的 Reduce 运算对中间统计信息进行合并，并利用预测模型进行分析。

3) 基于 Spark 的方法。基于本文设计的框架，使用 Spark 作为计算引擎，完成交通流数据清洗，并在 Shark 数据仓库进行存储，在中心管理节点实现清洗后交通流统计数据的合并，最终使用预测模型进行分析验证。

实验中整理了 6 个数据集，数据量分别为 2 万、5 万、10 万、20 万、50 万和 100 万条。针对这 6 个数据集，使用以上 3 种方法进行效率对比(图 6)。其中 50 万和 100 万条两个数据集的数据量已超过系统目前在一个预测周期内采集的最大交通流数据量。从图 6 得到以下结论。

1) MPI 方法在处理不同数据量的数据集时，耗费时间与数据量成正比。由于 MPI 框架结构简单，占用计算资源较少，因此当数据量较小时，数据处理时间少于其他两种方法。

2) 相对于 MPI 方法, MapReduce 方法的应用框架更复杂，数据处理耗时占比较大，因此在数据量较小时耗时比 MPI 方法长，但框架耗时基本上不受数据量影响，随着数据量不断增加，数据处理耗时占比不断降低，从而使时间曲线斜率减小，对 50 万条数据的处理时间反而少于 MPI 方法，并且数据量越大，时间优势越突出。

3) Spark 方法的时间曲线与 MapReduce 方法近

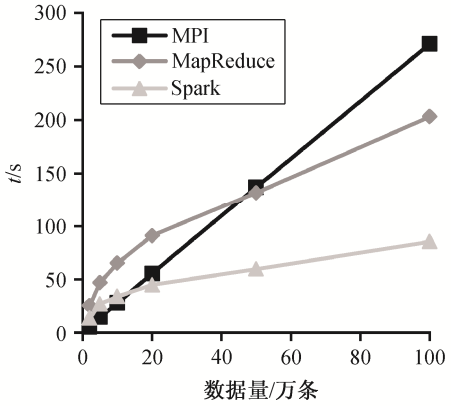

图 6　3 种算法不同数据量时间对比

Fig. 6　Cost of time of three algorithm for different amount of data

似，但总体耗时明显较少，原因在于 MapReduce 在交通流数据清洗过程中需要进行大量 I/O 操作和资源申请，由于磁盘存取效率远远低于内存存取，因此 Spark 方法的内存计算模式大大提高了整体运行效率，更适合实现复杂的数据清洗和挖掘算法。通过实验发现，Spark 方法耗时比 MapReduce 方法少 60%左右。

从总体上看，基于 Spark 的交通流数据处理方法技术成熟，效率较高，适用于交通流预测应用的前期数据处理工作。

3.3 预测应用实验

经过数据清洗和统计后，按照预测的时间周期，各个路段的交通流统计信息被传送至中心管理节点。基于此统计信息，利用动态 STARIMA 模型[1]和本文提出的基于多阶空间权重矩阵的 STARIMA 交通流预测模型进行预测分析。首先对需要预测的区域路网进行抽象化处理。

图 7(a)为郑州市主城区及其路网，图 7(b)为图 7(a)中龙子湖高校园区路网抽象化示意图。以图 7(b)中路网为例，设定 15 分钟为预测分析的时间周期，从实验数据集中选取前 4 天共 384 个时段作为历史数据集，对后 10 天 960 个时段进行预测分析，将分析结果与真实交通流数据进行对比，计算均方误差，定量地评价预测的准确度。

1) 动态 STARIMA 模型。利用时空自回归移动平均模型 STARIMA 来体现交通流在时空维度上的相关性，但此相关性仅包含一阶路段，不能准确地反映交通流运行的真实特征，预测结果会受到影响。

2) 基于多阶空间权重矩阵的 STARIMA 模型。将空间权重矩阵扩展为多阶矩阵，能够更准确地反映上下游交通流的时空相关性。实验结果表明，预测分析中引入二阶空间权重矩阵可以提高分析的准确性；三阶以上路段运算量大，相关性弱，为了保证系统的高效运行，可以忽略不计。

从表 1 看出，本文预测模型比动态 STARIMA 模型准确度更高，原因在于对空间权重矩阵进行了多阶优化，不但能够反映交通流的真实分配规律，而且可以保证系统的运算效率，预测结果也更准确。

4 结语

在 Spark 框架下对各类传感器采集的交通流大数据进行清洗、统计、存储和查询，可以实现交通流数据的高效处理。对交通流进行预测分析，是解决城市拥堵问题的一种技术辅助手段。

本文设计了一种基于 Spark 的交通流数据处理与预测应用框架，利用该框架实现交通流数据的高效清洗、统计、存储和查询。对比实验证明，基于 Spark 的数据处理方法的运算效率优于基于 MPI 和 MapReduce 的方法，可以满足交通流预测分析的应用需求。

本文还基于交通流周期统计信息，设计并实现了基于多阶空间权重矩阵的 STARIMA 交通流预测

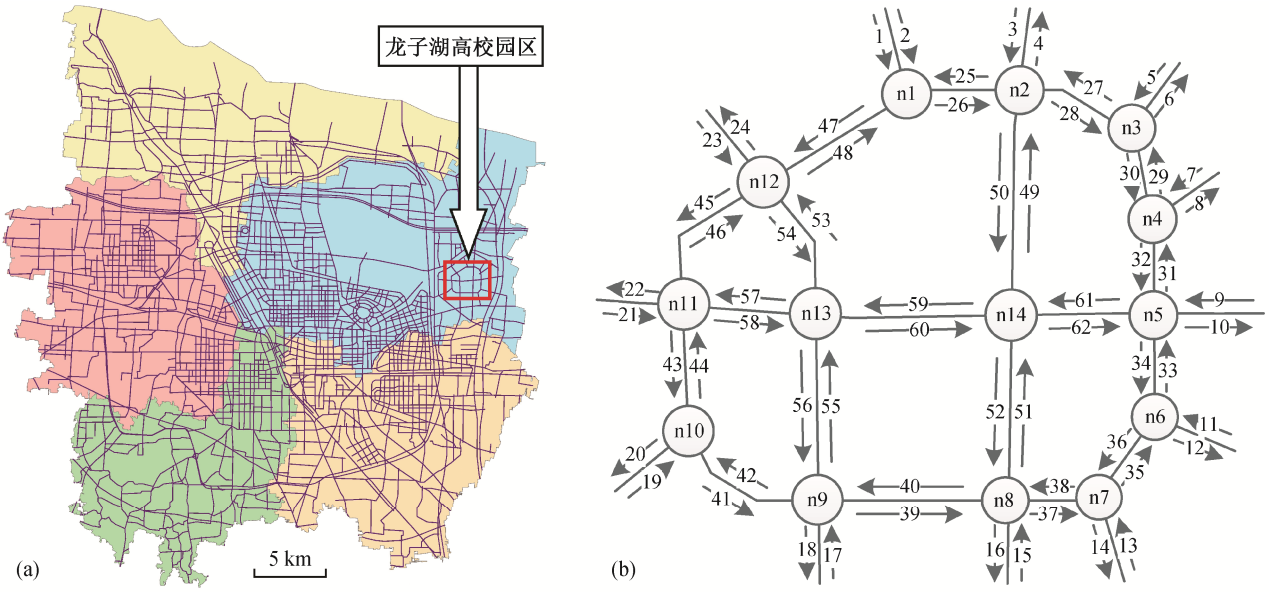

图 7 郑州市城区道路交通网(a)及龙子湖高校园区路网抽象化结果示意图(b)

Fig. 7 Zhengzhou city road network (a) and abstraction of Longzi Lake college area road network (b)

表 1　动态 STARIMA 模型和基于多阶空间权重矩阵的 STARIMA 模型预测结果的均方误差

Table 1　MSE comparison of dynamic STARIMA model and multi-order spatial weight matrix STARIMA model predict results

编号	动态 STARIMA	多阶空间权重 STARIMA	编号	动态 STARIMA	多阶空间权重 STARIMA	编号	动态 STARIMA	多阶空间权重 STARIMA
1	7943.56	6223.66	22	3057.75	2445.67	43	3659.67	2695.67
2	7023.23	5877.23	23	5584.67	4368.68	44	4369.57	3696.78
3	6147.38	4638.43	24	7569.84	5196.54	45	8467.43	7047.69
4	7830.57	5437.67	25	2469.48	1357.63	46	8154.52	6168.31
5	2857.37	1647.68	26	3367.09	1856.36	47	2560.57	1756.57
6	4529.67	2557.64	27	8357.67	5548.72	48	3357.21	2179.74
7	8956.57	6945.68	28	6367.98	5175.66	49	7157.65	5713.67
8	7420.04	5379.28	29	11474.57	8756.23	50	6527.43	5347.55
9	14379.47	12469.57	30	12649.82	9646.72	51	6356.92	4357.68
10	10456.39	8526.63	31	5157.84	3689.72	52	5562.83	4017.01
11	3756.28	2669.56	32	4619.37	3357.71	53	3436.18	2481.47
12	4861.36	3318.81	33	7428.71	6329.16	54	3819.18	2018.76
13	3681.88	2729.57	34	7638.18	5832.11	55	2174.28	1367.28
14	3728.19	2647.57	35	6528.19	4629.19	56	3518.18	1746.82
15	4368.29	3856.28	36	7746.29	5937.18	57	4728.79	3761.47
16	6429.18	4628.18	37	7937.27	6257.18	58	5528.67	4527.63
17	10469.23	7523.87	38	7623.67	6247.67	59	6218.57	4475.71
18	8732.75	6519.48	39	3618.57	2257.67	60	6312.99	4862.53
19	4729.67	3792.66	40	4629.21	3357.99	61	7319.46	6157.38
20	3257.57	2548.37	41	7219.57	5073.38	62	8368.27	7047.26
21	5639.21	4157.57	42	8836.28	5257.67			

模型，通过对空间权重矩阵的多阶优化，可以定量地反映交通流在路网中的真实运行规律，预测结果的准确性优于动态 STARIMA 模型，可以为交通诱导提供参考信息。

本文实验中使用的数据量已具备一定的规模，但由于硬件设备限制，分布式节点数量较少。在未来的工作中，需要进一步验证海量分布式节点环境下系统的运行情况。本文仅模拟了车辆移动时选择最短路径的情况，未对更复杂的最短时间、中途点以及个人偏好等影响路径选择的因素进行分析，需要进一步的研究和验证。

参考文献

[1] Aji A, Wang F, Vo H, et al. Hadoop-GIS: a high performance spatial data warehousing system over mapreduce // Proceedings of the 39th International Conference on VLDB Endowment. Trento, 2013: 1009–1020

[2] Witayangkurn A, Horanont T, Shibasaki R. Performance comparisons of spatial data processing techniques for a large scale mobile phone dataset // Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications. New York: ACM Press, 2012: 1–6

[3] Abouzeid A, Bajda-Pawlikowski K, Abadi D, et al. HadoopDB: an architectural hybrid of MapReduce and DBMS technologies for analytical workloads // Proceedings of the 35th International Conference on VLDB Endowment. Lyon, 2009: 922–933

[4] Plugge E, Hawkins T, Membrey P. The definitive guide to MongoDB: the NoSQL database for cloud and desktop computing. Berlin: Springer, 2010

[5] 温馨，罗侃，陈荣国，等. 基于 Shark/Spark 的分布式空间数据分析框架. 地球信息科学学报，2015, 17(4): 401–407

[6] Stephanedes Y J, Michalopoulos P G, Plum R A. Improved estimation of traffic flow for real time control. Transportation Research Record: Journal of the Transportation Research Board, 1981, 795: 28–39

[7] Ahmed M S, Cook A R. Analysis of freeway traffic time-series data by using box-jenkins techniques // Transportation Research Record 722. Washington DC: Transportation Research Board, l979: 1−9

[8] Dougherty M S, Cobbett M R. Short-term inter-urban traffic forecasts using neural networks. International Journal of Forecasting, 2004, 13(1): 21−31

[9] Ledoux C. An urban traffic flow model integrating neural networks. Transportation Research Part C: Emerging Technologies, 1997, 5(5): 287−300

[10] Okutani I, Stephanedes Y J. Dynamic prediction of traffic volume through Kalman filtering theory. Transportation Research Part B, 2008, 18(1): 1−11

[11] 余碧莹, 邵春福. 基于时空模型的道路网交通状态预测 // 第四届中国智能交通年会论文集. 青岛, 2008: 546−551

[12] Martin R L, Oeppen J E. The identification of regional forecasting models using space-time correlation functions. Transactions of the Institute of British Geographers, 1975, 66: 95−118

[13] Kamarianakis Y, Prastacos P. Space-time modeling of traffic flow. Computers & Geosciences, 2005, 31: 119−133

[14] 李欣, 孟德友. 基于路网相关性的分布式增量交通流大数据预测方法. 地理科学, 2017, 37(2): 209−216

[15] Lin Shulan, Huang Hongqiang, Zhu Daqi, et al. The application of space-time ARIMA model on traffic flow forecasting // Proceedings of the 8th International Conference on Machine Learning and Cybernetics. Baoding, 2009: 3408−3412

[16] Min Xinyu, Hu Jianming, Chen Qi, et al. Short-term traffic flow forecasting of urban network based on dynamic STARIMA model // Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems. St. Louis: Institute of Electrical and Electronics Engineers, 2009: 461−466

[17] 瞿莉. 基于动态交通流分配参数的网络交通状态建模与分析[D]. 北京: 清华大学, 2010

[18] 张和生, 张毅, 胡东成, 等. 区域交通状态分析的时空分层模型. 清华大学学报(自然科学版), 2007, 47(1): 157−160

[19] Xin R S, Rosen J, Zaharia M, et al. Shark: SQL and rich analytics at scale // Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. New York, 2013: 13−24

[20] Zaharia M, Chowdhury M, Franklin M J, et al. Spark: cluster computing with working sets // Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing. Berkeley: USENIX Association, 2010: 1−10

[21] Tabaa Y, Medouri A, Tetouan M. Towards a next generation of scientific computing in the cloud. International Journal of Computer Science, 2012, 9(6): 177−183

[22] Zaharia M, Chowdhury M, Das T, et al. Fast and interactive analytics over Hadoop data with Spark // Proceedings of the 10th USENIX Conference on File and Storage Technologies. San Jose: USENIX Association, 2012: 45−51

[23] 王晓原, 张敬磊, 吴芳. 交通流数据清洗规则研究. 计算机工程, 2011, 37(20): 191−193

[24] Deng Shuo, Hu Jianming, Wang Yin, et al. Urban road network modeling and real-time prediction based on house holder transformation and adjacent vector // Advances in Neural Networks — ISNN 2009. Berlin: Springer, 2009: 899−908

[25] Pfeifer P E, Deutsch S J. A three-stage iterative procedure for space-time modeling. Technometrics, 1980, 22(1): 35−47

[26] Pfeifer P E, Deutsch S J. Identification and interpretation of first-order space-time ARMA models. Technometrics, 1980, 22(3): 397−408

[27] Pfeifer P E, Deutsch S J. Variance of the sample-time autocorrelation function of contemporaneously correlated variables. SIAM Journal of Applied Mathematics, Series A, 1981, 40(1): 133−136

[28] 牛新征, 佘堃. 面向大规模数据的快速并行聚类划分算法研究. 计算机科学, 2012, 39(1): 134−137