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[J].地理科学,2017,37(2):209-216.[Li Xin, Meng Deyou. Distributed Incremental Traffic Flow Big Data Forecasting Method Based on Road Network Correlation. Scientia Geographica Sinica,2017,37(2):209-216.] doi: 10.13249/j.cnki.sgs.2017.02.006

基于路网相关性的分布式增量交通流大数据预测方法

李欣,孟德友

(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/
河南财经政法大学资源与环境学院,河南 郑州 450046)

摘要:针对城市道路拥堵问题的日益加剧的问题,智能化城市交通管理平台是缓解拥堵问题的有效方法,利用交通流大数据预测结果进行交通诱导,能够指导用户调整出行方案,有效缓解交通压力。研究了交通流大数据的分布式增量聚合方法,对海量交通流数据进行清洗统计,为交通流预测提供数据基础,基于交通流在路网中上下游路段的相关性分析,利用路口转弯率多阶分配将该相关性量化,构建基于路网相关性的空间权重矩阵,完成对于STARIMA模型的改进。通过应用试验证明,该方法能更准确的进行交通流预测,为交通诱导信息发布提供依据。

关键词:交通流;大数据;分布式增量;路网相关性;STARIMA

中图分类号:K909 **文献标识码**:A **文章编号**:1000-0690(2017)02-0209-08

中国的城市化进程正在不断加速,汽车数量增长和城市交通基础设施不足的矛盾日益突出,智能化城市交通管理平台是缓解拥堵问题的有效方法,通过大数据挖掘进行交通流预测和诱导,能够指导用户调整出行方案,有效缓解交通压力。如何利用反映交通活动的大量与时空位置相关的数据[1]进行交通流预测是实现交通诱导的关键问题。

目前国内外已经产生了很多研究成果。在大数据聚合方法方面,如增量并行数据挖掘方法[2],增量降维时序数据处理方法[3],基于MapReduce的样本抽样放回方法[4],基于MapReduce分布内存加速方法[5]等。上述方法的主要特点是通过抽样降维方式减少数据量提高效率,或基于传统规模数据集进行增量挖掘运算,或基于昂贵内存硬件提高计算性能。而针对广域网环境下不断海量增长的时空数据,以上方法仍然无法有效利用已有设备和使用较低成本解决效率问题。在交通流预测分析应用方面,主要研究成果包括历史平均模型[6]、状态空间模型[7]、时间序列模型[8]、神经网络模型[9,10]等。乐阳[11]提出基于时空依赖性的Kalman滤波模型,Kamarianakis等人[12]在交通流预测中引入时空自回归移动平均模型[13]模型,之后国内的一些学者[14~18]在此基础上进行了进一步改进,取得了较好效果。以上研究针对的是固定时段交通流数据的预测,有的仅考虑了单点交通流预测,有的对于交通流的相关性限制较多,未能真实准确的反应交通流在路网中的实际规律,还需要进行深入研究。

本文拟从交通网的时空相关性分析出发,在已有研究成果基础上,改进大数据环境下的城市交通流预测分析模型,利用河南智能交通综合管理平台获取的交通流数据,进行实验性预测分析,为下一步的实际应用提供参考依据。

1 交通流大数据分布式增量聚合管理方法

1.1 分布式增量聚合流程

本文研究实现了一种交通流大数据分布式增量聚合管理方法,该方法将网络中的节点分为中

收稿日期:2016-02-25;**修订日期**:2016-08-05

基金项目:国家自然科学基金项目(41501178)、河南财经政法大学博士科研启动基金项目(800257)资助。[Foundation: National Nature Sciences Foundation of China (41501178), Henan University of Economics and Law Dr. Startup Funds(800257).]

作者简介:李欣(1981–),男,河南郑州人,博士,讲师,主要从事地理信息系统理论研究与实践应用研究。E-mail:lixin992319@163.com

心节点和分布节点：分布节点负责收集和处理交通路网中的不同种类传感器获取的交通流数据，定时将聚合处理完毕的数据推送到中心节点；中心节点负责对历史全集和增量阶段的交通流数据进行存储管理，执行基于路网相关性的交通流预测。图1为交通流大数据分布式增量聚合原理图。

整个数据聚合和预测过程，按照时间间隔分为多个周期阶段持续执行。第一阶段，是初次数据聚合和预测分析阶段，称作历史全集数据聚合阶段，基于网络中所有节点的数据全集进行分布式聚合运算，数据整合到中心节点后进行预测分析；后期阶段，称作周期增量数据聚合阶段，基于系统运行周期新产生的增量数据进行聚合运算，增量数据可以用来修正上一阶段预测结果。整个过程通过MapReduce模型执行，其中分布节点的Map运算包含了数据筛选清洗算法，同时在分布节点由Combine运算完成中间统计数据集处理，之后将中间结果推送到中心节点，最终在中心节点使用Reduce运算进行全局数据聚合，最终执行预测模型生成预测结果。

1.2 交通流数据清洗规则

在数据聚合过程中，最为关键的是交通流数据清洗规则，国内外学者经过研究虽然取得了一定的成果[19~22]，但对于复杂的交通流传感器数据，目前还没有统一的清洗规则。

本文利用文献[22]的高维交通流孤立点检测算法，以及依据阈值理论和交通流理论制定的清洗规则，对系统分布节点中的数据集进行数据清洗，步骤如图2。经过实验，通过在分布节点进行交通流数据清洗，可以纠正或丢弃错误、冗余数据90%以上，可以在较大程度上提高数据质量，进而辅助进行精准预测。在进行数据清洗的同时，由于需要对每一条数据记录进行分析，因此在清洗同时即可完成对交通路网中路段流量的统计，并传送到中心节点完成交通流预测。

2 大数据环境下基于路网相关性的交通流预测模型

2.1 交通流在路网中的相关性分析

城市道路网错综复杂，但根据其拓扑结构可以看出，任意两条路段之间的关系存在两种情况：一种是两条路段邻接，即两路段由某个路口相连；另一种是两条路段非邻接，即路段非首尾相连，必须通过一条或多条其他路段和路口相连。

对于邻接路段，上游交通流在到达路口后会重新分配至下游路段，例如典型的十字路口车辆存在直行、左转、右转3种情况。因此，邻接路段交通流之间的关系即为上游交通流的重分配关系。图3为交通路口上下游交通流示意图。

由图3中路段上下游关系，可以将邻接路段，第 k 个交通数据采样时段 $[t_k,t_{k+1}]$ 内上游路段 l_i 与下游路段 l_j 之间的时空相关性 $r_{ij}(k)$ 可以用第 $k-1$ 个时段路段 l_i 与路段 l_j 路口转弯率 $\theta_{ij}(k-1)$ 来量化，表达式为：

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

非邻接路段中的交通流重分配可以看做是上游路段到下游路段的多阶重分配，而分配过程涉及到路径选择问题。文献[14]中分析了上游路段

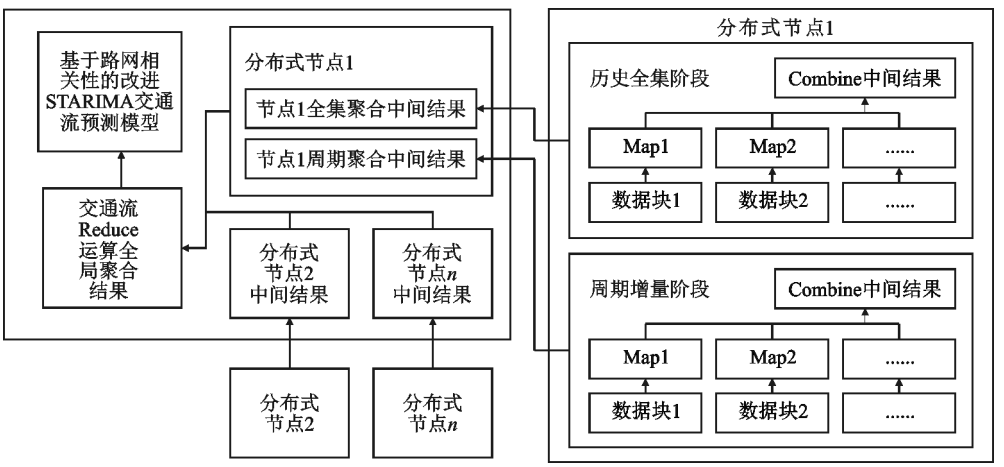


图1 交通流大数据分布式增量聚合原理

Fig.1 Distributed incremental aggregation principle of traffic flow big data

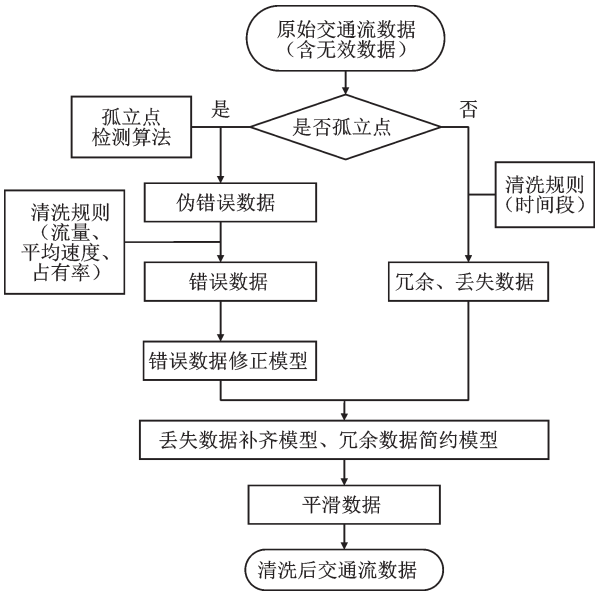
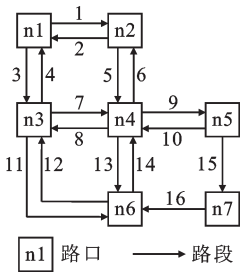


图2　交通流数据清洗步骤

Fig.2　Traffic flow data cleaning steps



注:n4路口下游路段与上游路段对应编号为6→7, 10, 14;
8→5, 10, 14;9→5, 7, 14;13→5, 7, 10

图3　交通路口上下游交通流

Fig.3　Traffic flow upstream and downstream in the intersection n4

和下游路段在分配交通流时存在3种关系。而在实际情况下,影响驾驶员路径选择的原因比较多,例如最短距离,最短时间,个人喜好,中途点选择等等。本文仅考虑正常状态下的交通流预测问题,即驾驶员一般选择最短路径作为驾驶路线,此时上游路段A到下游路段B之间仅有唯一路径,交通流重分配仅在此路径中实现。

在非邻接路段条件下,上游路段段 l_i 与下游路段 l_j 之间经过 n 个路口,其通路为 $R_{ij}=\{l_i,\ l_1,\ l_2,\ \cdots,\ l_{n-1},\ l_j \mid l_1,\ l_2,\ \cdots,\ l_{n-1}\in L,\ n\geqslant 2\}$,则第 k 个交通数据采样时段 $[t_k,t_{k+1}]$ 内的时空相关性即为在 n 个路口的转弯率多阶分配,量化表达式为:

$$r_{ij}^{n}(k)=\theta_{l_{n-1}l_j}(k-1)\,\theta_{l_1l_1}(k-n)\prod_{p=1}^{n-2}\theta_{l_pl_{p+1}}(k-n-p) \quad (2)$$

式中,$r_{ij}^{n}(k)$ 为路段 l_i 经过 n 个路口后与路段 l_j 的时空相关性,θ 为通路中两个相邻路段在对应时段中的路口转弯率;p 为自增变量,取值范围为1到 $n-2$。

通过公式(1)和(2)可以看出,上游路段的交通流会依次分配至下游路段,而下游路段被分配的流量为上游多阶路段累积分配的结果。随着经过路口数量的增加,阶数随之增加,上游路段分配至下游的流量越来越少,其时空相关性也越来越小。根据经验,2阶以内的上游路段和某下游路段的相关性较大,而3阶以上的上游路段与下游路段的相关性已经较小,可以忽略不计。

2.2　基于路网相关性的改进STARIMA模型

2.2.1　改进STARIMA空间权重矩阵

通过分析交通流在路网中的相关性,得到的结论是路网中路段相关性随着上下游阶数增加而减小。本文根据此结论,设计了时空自回归移动平均模型STARIMA(Space-time Autoregressive Integrated Moving Average)模型中的空间权重矩阵,并进行交通流预测。

时空自回归移动平均模型STARIMA是Pfeifer和Deutsch[23~25]提出的,公式如下:

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

式中,μ 为时间延迟;h 为空间间隔;p 为时间自回归延迟;m_k 为第 k 个时间自回归项的空间间隔;ϕ_{kh} 为时间延迟为 k 并且空间间隔为 h 的自回归参数;q 为移动时间平均延迟;n_l 为第 l 个时间移动平均项的空间间隔;θ_{lh} 为时间延迟为 l 并且空间间隔为 h 的移动平均参数;$\varepsilon(t)$ 为随机误差;$W^{(h)}$ 为 h 阶空间权重矩阵;Z 为 n 个路段 t 时段的交通流量组成的交通流量向量。

公式中的 $W^{(h)}$ 为一个 $N\times N$ 阶矩阵,但Pfeifer和Deutsch并没有明确指出矩阵中元素该如何取值,使用者可以根据自己需要解决问题的实际情况,来定义矩阵取值。

因此,可以将该矩阵作为空间权重矩阵进行赋值,而赋值时不仅需要考虑某路段历史和当前交通流量,还需要考虑其上游路段的历史和当前交通流量,而且上游路段距离该路段越近,其相关

性和对流量预测的影响就越大，$W^{(h)}$ 空间权重矩阵即可用来定量表达路段之间的相关性。

考虑文献[23]对于 $W^{(h)}$ 空间权重矩阵的3种限制条件，依据前文交通流在路网中时空相关性分析结论，本文提出的空间权重矩阵元素公式为：

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

式中，$w_{ij}^{(l)}$ 为空间权重矩阵元素值，$r_{ij}^{l}(k)$ 为路段 i 和路段 j 经过 l 个路口的时空相关性。

公式(4)不但反映了上下游路段在交通流重分配过程中的相关性影响，而且也考虑了两者之间的空间拓扑关系，同时还满足文献[23]提出的3种条件。因此，可以作为新的确定空间权重矩阵元素的方法应用到交通流预测分析中。

2.2.2 模型应用步骤

基于路网相关性的改进STARIMA交通流预测模型，其应用步骤可以分为以下6步：① 交通路网拓扑抽象化。根据城市路网的空间拓扑关系，将其抽象为明确表示上下游关系的网状结构，网络中包含方向和长度数据，以此为基础可以建立空间权重矩阵；② 确定空间权重矩阵元素。利用文献[26]中的交通路口动态转弯率预测模型进行估计，使用公式(4)确定一阶和二阶空间权重矩阵，而三阶以上的由于相关性较小，可以忽略不计；③ 时间序列平稳化。实际情况下，交通流时间序列为非平稳序列，可以使用序列图[10]通过差分方式使交通流时间序列平稳化；④ 确定模型阶数和参数。可以使用时空自相关函数[27]与时空偏相关函数[28]确定自回归移动平均阶数，然后利用预测值和实际值残差平方进行参数估计；⑤ 模型校验和诊断。检查预测值和实际值之间的误差序列是否满足随机误差，并检查参数估计的统计显著性，若不满足要求则返回上一步；⑥ 交通流预测。确定模型阶数和参数之后，即可将交通流历史和增量数据代入模型进行预测。

3 应用实例分析

3.1 数据来源及运行环境

本文实验基于智能交通综合管理平台搭建，该平台提供了城市交通指挥系统、智能交通诱导系统、联网视频监控系统、智能交通检测系统等一整套综合管理平台。目前，已经在郑州、开封、洛阳等城市实现了部分应用。

郑州市动态交通流信息采集传感器，包括微波检测器、视频检测器、地磁检测器采集以及浮动车GPS数据。平台数据总量已达到160多亿条，日均增量约2 000万条。本文选取2015年11月9日至11月22日共计14 d的数据进行实验。

实验环境中架设了1台服务器作为中心节点进行交通流预测分析，4台服务器作为分布节点处理增量交通流数据，配置均为Intel5620 2.4GHz，6核，4GB内存，2TB硬盘。

3.2 交通流大数据分布式增量聚合实验

实验设定初始状态下历史交通流数据集合为前4 d的交通流量数据约8 000万条，以15 min为增量周期，每一周期内的数据量约为20万条。由于白天和晚上交通流数据分布不均匀，因此在一个流量高峰周期的数据量可能达到平均量的2倍，即40万条，而且交通流数据是在不断连续增长的，必须在周期时限内完成对数据的快速聚合处理，才能满足中心节点的预测分析的需求。

本文采用了两种方法进行交通流大数据的聚合实验，并对其结果进行了比较：

1）基于MPI的数据聚合。MPI主从模式并行程序中，主进程负责分配任务和数据，从进程完成任务后返回结果。实验利用文献[29]中的MPI方法，主进程设置在中心节点，在4台分布节点设置4个从进程，从进程中运行的计算主要是对于数据的遍历和清洗算法，在遍历数据同时完成交通网络中流量统计，并传送给中心节点的主进程，最终由主进程完成交通流预测。

2）基于分布式增量MapReduce的数据聚合。分布节点的4台服务器存储时空数据全集，并对数据集合进行平均分块，配置48个Map运算和4个Combine运算，在Map运算中包含了交通流清洗算法，由分布节点Combine运算完成中间统计数据集处理，之后将中间结果推送到中心节点，最终在中心节点使用Reduce运算进行全局数据聚合，最终执行预测模型生成预测结果。

针对两种方法，实验选用2万、5万、10万、20万、50万和100万条数据进行聚合，并比较了两种方法的运行效率。实验结果如图4所示。

对比两种方法可以看出，在数据量较小时，MPI方法效率要明显高于MapReduce方法，这是因为MapReduce方法包含了很多应用架构逻辑，

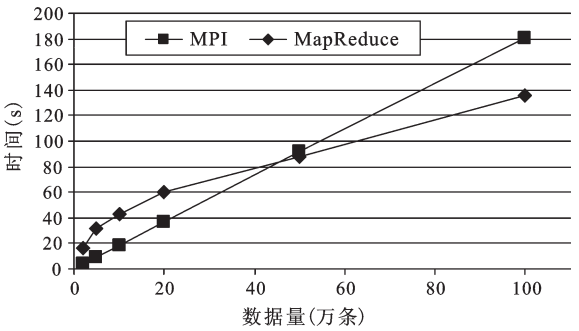


图4　两种算法不同数据量时间对比

Fig.4　The cost of time of two kinds of algorithm for different amount of data

应用逻辑会消耗一定的时间;而在数据量不断增大之后,数据量成为时间消耗的主导因素,应用架构逻辑所占用的时间所占比例越来越小,两者处理数据的时间不断接近,而在数据量达到50万条时,MapReduce方法所用时间反而少于MPI方法,而随着数据量的进一步增加,MapReduce的优势将更加明显。而且两种方法虽然耗费时间的数量级仍然一样,但MapReduce的快速开发周期和连续稳定的运行效果是MPI方法无法达到的。总体上看,实验使用的分布式增量交通流大数据聚合方法,从开发周期和运行效率上都可以满足城市交通预测分析的需要。

3.3　基于路网相关性的交通流预测实验

交通流统计数据集中传送到中心节点后,即可使用基于路网相关性的改进STARIMA模型进行交通流预测。按照模型应用步骤,首先构建交通路网拓扑结构,对研究区域进行抽象化,图5a为郑州市城区道路交通网,图5b是对图5a中矩形框范围内龙子湖高校园区路网抽象化结果示意图。

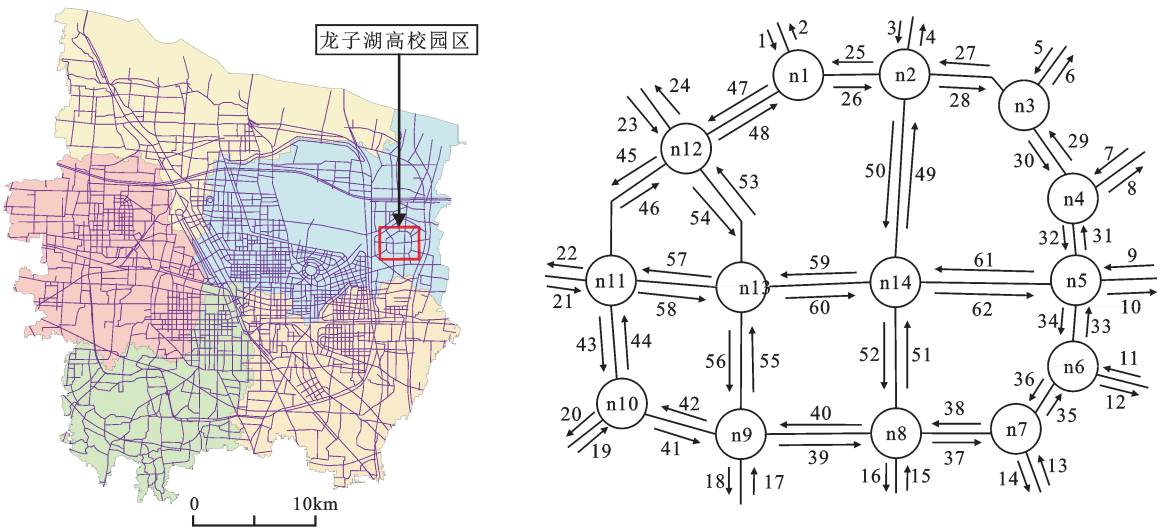
根据文献[30]的研究发现,交通流预测周期以15 min为宜,若时间间隔过小,交通流数据会被信号灯或其他因素影响而出现较大波动,而时间间隔过大,对于交通流预测又无法起到实际的诱导交通作用。因此本文使用15 min作为数据增量周期进行预测。

对于交通流预测,本文也使用了两种方法进行对比试验:

1）动态STARIMA预测模型。文献[12]中提出了该模型,该模型将上下游路段的相关性引入STARIMA模型,但只考虑了一阶上游路段的影响。

2）基于路网相关性的改进STARIMA模型。本文提出的模型不仅引入路网之间的相关性,而且将n阶上游路段的影响考虑在内。

以图5b中的示意路段为例,将实验数据中的前4 d数据作为历史数据,使用两种预测模型对后10 d的数据做预测。每天24 h按照15 min一个周期进行划分,一天分为96个时段,所以共使用384个时段的历史数据预测后960个时段的交通流量,将预测结果和实际交通流数据对比,计算均方误差作为预测误差指标,两种方法得到的预测误差



a郑州市城区道路交通网　　b龙子湖高校园区路网抽象化结果

图5　郑州市城区道路交通网及龙子湖高校园区路网抽象化结果

Fig.5　Zhengzhou City road network and abstraction of Longzi Lake college area road network

表1 两种方法预测结果均方误差(MSE)对比

Table 1 The MSE comparison of two predict methods results

编号	动态 STARIMA	路网相关 STARIMA	编号	动态 STARIMA	路网相关 STARIMA	编号	动态 STARIMA	路网相关 STARIMA
1	8638.27	6643.68	22	3338.29	2839.48	43	3977.67	3058.83
2	7396.01	6032.39	23	6236.77	5064.66	44	4458.98	3847.45
3	6631.31	4983.57	24	7784.39	5338.54	45	8979.37	7432.82
4	8362.58	5986.44	25	2526.67	1438.33	46	8443.69	6234.73
5	2948.21	1863.49	26	3464.32	1974.34	47	2798.55	1846.49
6	4820.22	2799.35	27	8764.39	5890.43	48	3985.42	2275.45
7	9230.23	7390.32	28	6549.48	5438.93	49	7849.43	5893.37
8	7857.46	6074.45	29	12974.45	8865.29	50	6692.38	5624.78
9	15324.01	12984.61	30	13084.44	10474.63	51	6639.32	4542.43
10	11479.46	8753.05	31	5478.34	3740.35	52	5873.57	4147.57
11	3892.27	2775.73	32	4858.34	3275.39	53	3720.32	2475.54
12	4917.48	3295.48	33	7434.95	6578.36	54	3920.28	2143.27
13	3729.33	2903.57	34	7868.23	5920.49	55	2039.58	1343.47
14	3928.23	2638.57	35	6743.23	4839.33	56	3235.62	1634.57
15	4478.19	3902.29	36	7820.33	5923.67	57	4838.88	3822.44
16	6903.36	4632.84	37	8068.35	6488.38	58	5749.23	4727.28
17	10573.48	7296.23	38	7819.28	6367.35	59	6403.45	4884.74
18	9033.54	7018.83	39	3894.22	2057.75	60	6653.63	4954.54
19	4847.34	3892.33	40	4780.44	3628.65	61	7570.49	6343.45
20	3309.21	3087.88	41	7897.36	4929.38	62	8788.48	6932.43
21	5789.22	4274.49	42	8902.34	5563.37			

结果如表1。

对比两种方法的实验结果可以看出,基于路网相关性的改进STARIMA方法预测效果要明显优于动态 STARIMA方法,原因在于动态STARIMA方法仅考虑了一阶上下游路段之间的相关性,而本文提出的方法还考虑了二阶以上上下游路段之间的相关性,因此更加符合道路网中交通流的分配规则,预测结果也更加准确。

4 总结与展望

本文设计了交通流大数据分布式增量聚合管理方法,创新点在于将分布式增量数据聚合方法和交通流数据清洗规则相结合,对海量交通流数据进行聚合管理,通过MPI和MapReduce两种方法的对比试验,证明本文提出的方法在开发周期和稳定运行效果上要优于MPI方法,运行效率上能够满足交通流大数据聚合的需要,最终得到的交通流统计数据可以作为交通流预测的数据基础。而本文设计的大数据环境下基于路网相关性的改进STARIMA模型,创新点在于利用路口转弯率多阶分配将交通流在路网中上下游路段的相关性量化,以此为基础构建了空间权重矩阵,完成对于STARIMA模型的改进,通过对比试验,证明预测结果优于仅考虑一阶相关性的动态STARIMA模型,可以为诱导交通信息发布提供依据。

本文的预测模型在路径选择问题方面仅考虑了最短路径情况,而未考虑实际行车过程中最短时间,中途点选择等问题,还应在下一步的工作中展开研究。

参考文献(References):

[1] 李德仁,马军,邵振峰.论时空大数据及其应用[J].卫星应用,2015,(9):7-11.[Li Deren, Ma Jun, Shao Zhenfeng. The theory of space-time big data and its application. Satellite Application, 2015, (9):7-11.]

[2] Bose J H, Andrzejak A, Hogqvist M. Beyond online aggregation: Parallel and incremental data mining with online Map-Reduce [M]//Proc of Workshop on Massive Data Analytics on the Cloud. New York:ACM,2010

[3] Aghabozorgi S, Saybani M R, Wah T Y. Incremental clustering of time-series by fuzzy clustering. Journal of Information Science and Engineering,2012,28(4):671-688

[4] Laptev N, Zeng K, Zaniolo C. Very fast estimation for result and accuracy of big data analysis:The EARL system[M]//Proc of ICDE.Piscataway,NJ:IEEE,2013:1296-1299.

[5] Zhang S B, Han J Z,Liu Z Y et al. Accelerating MapReduce with Distributed Memory Cache[M]//Proc of ICPADS. Piscataway, NJ:IEEE, 2009:472-478.

[6] Stephanedes Y J, Michalopoulos P G. Improved estimation of traffic flow for real-time control [M]// Transportation Research Record 795, Washington DC: Transportation Research Board, 1981:28-39.

[7] Okutani I, Stephanedes Y J. Dynamic prediction of traffic volume through Kalman filtering theory[J]. Transportation Research Part B:Methodological, 1984,18(1):1-11.

[8] Ahmaed M S, Cook A R. Analysis of freeway traffic time-series data by using Box-Jenkins technique[M]//Transportation Research Record 722. Washington DC: Transportation Research Board, l 979:l-9.

[9] Doughetry M S, Cobbett M R. Short-term inter-urban traffic forecasts using neural networks[J]. International Journal of Forecasting, l997, 13(1):2l-31.

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

[11] Yue Yang. Spatial-temporal dependency of traffic flow and its implications for short-term traffic forecasting[D]. Hong Kong: The University of Hong Kong, 2006.

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

[13] Martin R L, Oeppen J E. The identification of regional forecasting models using space-time correlation functions[J]. Trans Inst Brit Geogr, 1975, 66: 95-118.

[14] 余碧莹,邵春福.基于时空模型的道路网交通状态预测[M]//第四届中国智能交通年会论文集.青岛:全国智能交通系统协调指导小组和山东省人民政府, 2008: 546-551.[Yu Biying, Shao Chunfu. Traffic state forecast of road network based on space-time model//Proceedings of the Fourth China Annual Conference on ITS. Qingdao: The intelligent transportation system to coordinate and guide team and the People's Government of Shandong Province, 2008:546-551.]

[15] Lin Shulan, Huang Hongqiang, Zhu Daqi et al. The application of space-time ARIMA model On traffic flow forecasting[M] // Proceedings of the 8th International Conference on Machine Learning and Cybernetics. Baoding: Hebei University and IEEE SMC Association, 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 [M] // Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems. St. Louis, MO, USA: 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.[Qu Li. Modeling and analyzing the network-level traffic status based on dynamic traffic assignment ratios. Beijing: Tsinghua University, 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[J].清华大学学报:自然科学版,2007, 47(1):157-160. [Zhang Hesheng, Zhang Yi, Hu Dongcheng et al. Spatial-temporal hierarchical model for area traffic state analysis. Journal of Tsinghua University: Sci &Technol. 2007, 47(1):157-160.]

[19] 王晓原,张敬磊.交通流信息挖掘的非参数概率变点模型研究[J].武汉理工大学学报:交通科学与工程版, 2010, 34(4): 801-805.[Wang Xiaoyuan, Zhang Jinglei. Study on Nonparametric Probability Change-point Model for Traffic Flow Exploitation. Journal of Wuhan University of Technology (Transportation Science & Engineering), 2010, 34(4):801-805.]

[20] 郭志懋,周傲英.数据质量和数据清洗研究综述[J].软件学报, 2002, 13(11): 2076-2082.[Guo Zhimao, Zhou Aoying. Research on Data Quality and Data Cleaning: A Survey. Journal of Software, 2002, 13(11): 2076-2082.]

[21] 张敬磊,王晓原.基于非线性组合模型的交通流预测方法[J].计算机工程, 2010, 36(5):202-205.[Zhang Jinglei, Wang Xiaoyuan. Traffic Flow Prediction Method Based on Non-linear Hybrid Model[J]. Computer Engineering, 2010, 36(5):202-205.]

[22] 王晓原,张敬磊,吴芳.交通流数据清洗规则研究[J].计算机工程, 2011, 37(20):191-193.[Wang Xiaoyuan, Zhang Jinglei, Wu Fang. Research on Traffic Flow Data Cleaning Rules. Computer Engineering, 2011, 37(20):191-193.]

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

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

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

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

[27] Bezdek J C, Pal N R. Some new indexes of cluster validity[J]. IEEE Trans Syst Man Cy, 1998, 28: 301-315.

[28] Kamarianakis Y, Prastacos P. Space-time modeling of traffic flow[J]. Comput Geosci-UK, 2005, 31: 119-133.

[29] 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[J].计算机科学,2012,(1):134-137,151. [Niu Xinzheng, She Kun. Study of Fast Parallel Clustering Partition Algorithm for Large Data Sets.Computer Science, 2012,(1):134-137,151.]

[30] Smith B L, Demeisky M J. Traffic flow forecasting: comparison of modeling approaches[J]. Journal of Transportation Engineering, 1997,123(4):261-266.

Distributed Incremental Traffic Flow Big Data Forecasting Method Based on Road Network Correlation

Li Xin, Meng Deyou

(*Collaborative Innovation Center of Three-aspect Coordination of Central Plain Economic Region, Henan University of Economics and Law, College of Resource and Environment, Henan University of Economics and Law, Zhengzhou* 450046, *Henan, China*)

Abstract: Along with the accelerating urbanization, there are more and more contradictions between the number of cars and urban transportation facilities. The congestion time and congested roads in cities are increasing. Intelligent urban traffic management platform is the effective method to alleviate the increasingly serious urban congestion problems. By using prediction results of traffic flow big data, the platform can guide users to adjust the travel plan, and ease the traffic pressure effectively. How to use a large number of spatio-temporal data related to traffic activities to predict the traffic flow is the key to realizing traffic guidance. In this article, a distributed incremental aggregation method for traffic flow data is studied. The method combines the distributed incremental data aggregation method with the traffic flow data cleaning rules, makes cleaning and counting of traffic flow big data, and provides data for traffic flow forecast. With the analysis of traffic flow correlation in the network of upstream and downstream, this article uses the multi-order of turning rate in the intersection to quantize the correlation, builds the spatial weight matrix based on the road network correlation, and improves the STARIMA model. In this article, two groups of contrast experiments were made. Through the contrast experiment between MapReduce method and MPI method, the result proves that the method proposed in this article is better than the MPI method in the development cycle and stable operation. The method's efficiency can meet the need of traffic flow data aggregation. The traffic flow statistics can be used as the basis of traffic flow forecasting. Through the contrast experiment between the Improved STARIMA model and the Dynamic STARIMA model, the result proves that the Improved STARIMA model, which considers the multi-order correlation between the upstream and downstream sections, matches the distribution rules of traffic flow in road network better. Therefore, the forecast results are more accurate. In conclusion, the method of this article is a new method of traffic flow forecasting in the background of big data, and it can realize accurate prediction.

Key words: traffic flow; big data; distributed incremental; road network correlation; STARIMA