# 多断面相关性区间预测法在短期交通流预测中的应用

李秀丽,李星毅
(江苏大学 计算机科学与通信工程学院 ,江苏 镇江 212000)

**摘要:** 传统的交通流短期预测方法以点预测为主, 文中提出了基于多断面相关性的区间预测方法。多断面相关性预测是指把两个或者两个以上相邻点看作一个整体考虑,也将考虑被看作一个整体的相邻点的断面交通流的变化规律和各种影响因素,创建预测方法和模型,进而预测这些相邻点的交通流在未来时间段的整体变化。区间预测的概念由传统的点变为由区间预测的确定值,因此,它可以提供更多的信息,进而让用户可以更好地估计未来的不确定性,在可能的范围内做出恰当的处理策略。多断面的相关性区间预测法则是综合使用多个断面数据进行定量分析后,再选取合适的支持向量机(SVM)回归模型进行定量预测,以期提高预测精度。我们在公共的 2011 年 8 月 6 日至 8 日的 SBSJ 数据集上进行多组对比试验,均得到了较好的准确度。

**关键词:** 交通流短期预测; 多断面; 相关性; 区间预测; 支持向量机
**中图分类号:** TN914 **文献标识码:** A **文章编号:** 1674-6236(2017)19-0010-06

## The application of multi-section similarity and interval forecasting in short-term traffic flow forecasting

LI Xiu-li,LI Xing-yi
(*School of Computer Science and Telecommunications Engineering, Jiangsu University, Zhenjiang* 212000,*China*)

**Abstract:** The traditional short-term traffic flow prediction methods to predict the main point,this paper presents a prediction method based on multi-sectional correlation interval.Multi-sectional traffic flow forecasting to two or more adjacent dots to be considered as a whole,taking into account these changes of traffic flow cross-section adjacent dots and a variety of factors, the establishment of forecasting methods and models to predict these changes in traffic flow adjacent point in the future the whole period.The idea is to use the interval prediction interval instead of the dot predicted value is determined, it is possible to provide more information, allowing users to better estimate the uncertainty of the future, in the extent possible, select the appropriate treatment strategy.Correlation Interval prediction rules is a comprehensive multi-section using a plurality of cross-sectional data, quantitative analysis, and then select the appropriate support vector machine (SVM) regression model for quantitative prediction, in order to improve the prediction accuracy.We performed multiple sets of comparative tests on a common data set SBSJ on August 6 to 8th, 2011 have received a good accuracy.

**Key words:** short-term traffic flow prediction; multi-section; correlation; prediction interval; support vector machines

交通系统为一个时变的、非平稳的、非线性系统,具有不确定性。这种不确定性是由多方面原因造成的,不仅包括天气等自然原因,许多出乎意料小概率事件也是不可避免的。同时,它还囊括了许多特殊的人为因素, 比如司机在当时的心理状态下突发的交通事故等。由于上述这些不确定性因素是非常难以判断的, 也就导致交通预测这一工作的难度大大

**收稿日期:**2016-08-12 **稿件编号:**201608090
**作者简介:**李秀丽(1991—),女,山东潍坊人, 硕士研究生。研究方向:智能交通。

提升。

最原始的交通预测并不是一个单一的研究方向，仅仅是给交通控制提供一些数据参考，为它服务。最早期的城市交通控制系统（Urban Traffic Control Systems,UTCS)是一种利用综合考虑交通流的以往信息利用离线预测的方法；随后的城市交通控制系统则使用了实际测量得到的交通流数据来校正以往平均数据的缺陷,进而进行交通流量的预测；而第三代城市交通控制系统虽然利用实测数据进行预测,但是时滞问题尚未解决。总的来说,如今比较流行的交通预测模型主要有历史平均模型，卡尔曼滤波模型,神经网络模型,SVM 回归模型,混沌理论模型,尖点突变模型,小波模型等。

SVM 回归模型预测方法在交通流预测中应用非常广泛[1-3]。相比神经网络,SVM 是以研究小样本数据的内在相关性,以统计学理论为基础,利用结构风险最小化可以较好的解决" 小样本"、" 维数灾难"、"过拟合"和"局部极小点"等问题。丁爱玲[4]提出了一种基于统计学理论的交通流量时间序列预测方法，并通过实验证明了其方法的有效性和先进性。Lelitha V.,Laurence R.[5] 把神经网络与 SVM 在交通速度方面预测的效果进行了对比。王继生等[6]研究了在交通流预测中 SVM 的应用,并且将 SVM 与 BP 神经网络进行了比较，进而得出了一个结论:SVM 在交通流预测中的应用是有效的。杨兆升[7]通过实验得出了一种基于 SVM 的预测模型在精度、收敛时间及泛化能力等方面优于 BP 神经网络模型,并称为基于 SVM 的短时交通流预测模型。但是众多的文献均是基于 SVM 的点预测,鲜有文献作区间预测。因此本文基于此,提出基于 SVM 的区间预测方法。

多断面交通流预测是指以道路上的多个数据统计点的断面交通流的数据作为进行研究的对象,同时各个数据之间相互影响的预测。这些相邻点中可以进行划分为同一线路的相邻点和不同线路的相邻点。近几年,有关多断面交通流预测相关的文献主要有以下几个方面。Joe Whittaker 等[8]分析了一个道路交通网中的多个不同点间的交通流量和平均速度的相互影响,并建立了多点交通流模型,使用卡尔曼滤波方法来进行求解和预测。Anthony Stathopoulos 等[9-10]对不同地点的交通流进行了交叉谱分析，结果表明相互关系存在于不同地点的断面交通流。Markos 等[11]利用推广的卡尔曼滤波法结合宏观交通流模型对高速公路进行实时预测，并详细讨论了模型建模和参数估计的过程。诸多文献在使用多断面方法分析时,均采用了比较复杂的模型对交通流进行预测，随之带来的问题则是由于考虑因素比较多且复杂导致泛化性效果较差。而相对简易的模型则相对灵活,因此移植多种场景其泛化性会更好。

为了提高在交通流预测模型的准确率及适应能力，本文提出了基于支持向量机的多断面的相关性区间预测模型,并采用 SBSJ 数据集对模型进行实验和定量分析,以验证模型的可行性与有效性。

## 1　基础介绍

### 1.1　支持向量机

SVM 的主要思想是在给定的训练样本中建立一个超平面作为决策曲面，使得正例和反例之间的间隔最大化[12]。SVM 可用于求解非线性回归问题。

在二维空间中，如果样本数据能被一个线性数据完全分开,则称样本数据是线性可分的,该分类线性函数可表示为：

$$g(x)=w^{\mathrm{T}}X+b \tag{1}$$

式中,$x$ 为样本向量,$w$ 为样本向量的法向量,$b$ 为偏置值。

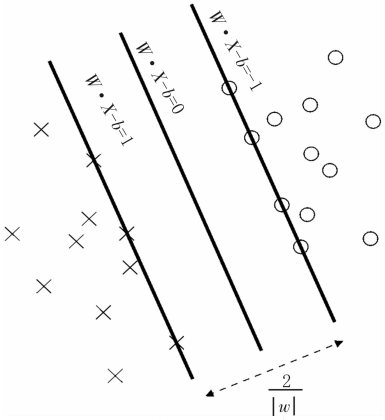


图 1　最优超平面示意图

图1 中,设：

$$H_1: wx-b=1 \tag{2}$$

$$H_2: wx-b=-1 \tag{3}$$

$$H: wx-b=0 \tag{4}$$

$H_1$、$H_2$ 平行于 H 且可以区分各类样本。$H_1$、$H_2$ 的训练样本点称为支持向量(SV)。最优超平面就是使分类间隔最大的平面。

第 $i(i=1,2,\cdots,N;N$ 为训练样本数)个训练样本由一个向量和一个类别组成,可表示为：

$$D_i=(x_i,y_i) \tag{5}$$

式中，$x_i$ 为输入向量，$y_i$ 为类别标记。对于二元线性分类，$y_i$ 只有 1 和−1 两个值。样本点到某个超平面的间隔为：

$$\delta_i=y_i(w^{\mathrm{T}}x_i+b)=|g(x_i)| \tag{6}$$

将式(1)中 $w$ 和 $b$ 进行归一化处理，可得到点到超平面的欧氏距离：

$$\delta_i=\frac{1}{\|w\|}|g(x_i)| \tag{7}$$

将所有样本点中最小的间隔定义为 1，此时相应的两条极端直线的几何间隔为 $\frac{2}{\|w\|}$，如图 1 所示。

由于几何间隔与 $\|w\|$ 成反比，最大化分类间隔可以通过最小化 $\|w\|$ 获得，即：

$$\min\|w\| \tag{8}$$

等价于：

$$\min\frac{1}{2}\|w\|^2 \tag{9}$$

文中规定样本点必须在 $H_1$ 或 $H_2$ 的一侧或者在 $H_1$、$H_2$ 上，由于所有样本点之间的间隔大于 1，因此：

$$y_i(w^Tx_i+b)-1\geqslant 0,i=1,2,\cdots,N \tag{10}$$

最大分类间隔的求解等价于在约束条件(10)下使 $\frac{1}{2}\|w\|^2$ 最小。

### 1.2 多断面相关性

交通流量断面，即单位时间内在某处通过的客流量。现实中交通流量断面之间的联系受到各种因素的影响。从定性分析的角度来看，当两个断面是上下游之间的关系并且距离比较近时相关性会相对较强；随着距离的增加或受其他交通流的交汇或道路断面之间存在交叉口等因素影响时，相关性会相对减弱。当针对交通道路网短时交通流进行预测时，特别是针对较大规模的道路网时，当将道路网中的每一个断面交通流数据都看成一个整体来考虑时，其计算复杂性相对比较高，可能满足不了短时交通流预测的实时性的要求。因此根据断面交通流状态的相关性，将一个较大规模的道路网切割成多个规模较小的子道路网，进而再进行短时交通流预测，这样就可以在保证一定的实时性要求的基础上，提高短时交通流预测的精确性。

因为判别断面交通流数据在总体水平上的相关程度比较困难，也就是说，要区别哪些断面之间相关性比较强，哪些断面的相关性比较弱的问题，是根据整体上的相关性将众多的道路断面进行分组，使一个较大规模的道路网分成几个较小的子路网。因此引进一种自定义度量方法，进行多元数据的分析。该方法主要是考虑每个断面数据采集的地理位置的相关性，主要有以下几点：

1)时间上，当前断面当前时刻交通流大小受上一时刻断面交通流变化的影响。

2)空间上，当前断面的交通流大小受相邻断面交通流变化的影响。

3)当前断面的交通流变化同时在时间与空间上受相关断面的影响。

### 1.3 区间预测

区间预测是在点预测的基础上预测出总体参数一个可能的范围。其理论基础则是区间估计。

设 $\theta$ 为总体 $X$ 的未知参数，$X_1,X_2,\cdots,X_n$ 为来自总体的容量为 $n$ 的简单随机样本，对于预先给定的一个充分小的正数 $\alpha(0<\alpha<1)$，构造两个统计量：

$$\hat{\theta}_1=\hat{\theta}_1(X_1,X_2,\cdots,X_n) \tag{11}$$

$$\hat{\theta}_2=\hat{\theta}_2(X_1,X_2,\cdots,X_n) \tag{12}$$

使得：

$$P(\hat{\theta}_1<\theta<\hat{\theta}_2)=1-\alpha \tag{13}$$

则称区间 $[\hat{\theta}_1,\hat{\theta}_2]$ 为总体参数 $\theta$ 的区间预测或置信区间，$1-\alpha$ 称为置信区间的置信度。

## 2 建模过程

### 2.1 支持向量机回归模型参数选取

由文献[13-14]可知，在经验风险最小化原则学习机器的实际风险的组成为经验风险和置信范围，其中置信范围不但受置信度 $1-\alpha$ 的影响，还受 VC 维 $h$ 和训练样本数 $n$ 的影响，并且随着它的增加而单调减少。将其用公式表示如下：

$$R(w)\leqslant R_{\mathrm{enp}}(w)+\Phi(\frac{n}{h}) \tag{14}$$

上式中，$R$ 为实际风险，$R_{\mathrm{enp}}$ 为经验风险，$\Phi$ 为置信区间。当 $\frac{n}{h}$ 较大时，置信范围 $\Phi$ 较小，此时经验风险最小化的最优解就接近实际的最优解。另一方面，若样本数 $n$ 固定，此时 VC 维越高，则置信范围越大，导致实际风险和经验风险之间的可能差就越大。因此我们在选择模型时，不但要使经验风险最小化，还要使 VC 维尽量小，从而缩小置信范围，使期望风险最小。

为了使 SVM 回归模型有效,必须具备一定的推广能力。而式(14)中的置信区间 $\Phi$ 计算比较困难。但是它为模型参数的选取提供了思路,即根据文献[14]可知，对于样本数目较小的数据回归分析,SVM 回归模型的参数选取应该遵循以下原则：

1) 避免$\frac{n}{h}$过小使得模型变复杂,确定其取值范围为 1~2。

2) 损失函数的参数 $\varepsilon$ 要适宜,为了在拟合精度与泛化能力之间平衡,$\varepsilon$ 取值范围一般为（0.000 1~0.01)。

3) 为了控制模型复杂度,惩罚因子 $C$ 应取值偏小,但为了使模型的经验误差不要过大,$C$ 的值又不能过小,因此确定其范围为(1~1 000)。

### 2.2　断面交通流数据处理

根据 1.2 节中描述的多维标度法数据预处理技术,对事先整理好的交通流数据进行预处理。数据按监测点地理位置分为断面 1、断面 2.....,其中每个断面中的数据时间间隔为 10 分钟。利用自定义度量法对数据进行预处理。

1)在当前断面的当前时刻数据中增加上一时刻当前断面的数据。

2)在当前断面的当前时刻数据中增加上下游 $N$ 个断面当前时刻的数据。

3)在当前断面的当前时刻数据中增加上下游 $N$ 个断面上 $M$ 个时刻的数据。

其中,$N$ 与 $M$ 为参数,暂无理论值,需要通过试验确定其最佳取值。

### 2.3　支持向量机预测点的置信区间计算

建立好 SVM 模型以后,对未知点进行预测,可以得到预测点的点预测,再根据点预测,基于主元法[15],借助基于正态分布的区间估计法来计算预测区间。通过文献[15]得知,置信区间的主元法的主要思想是：

1)寻找一个样本函数：

$$Z=Z(\xi_1,\cdots,\xi_n;\theta) \tag{15}$$

此函数只含有待估参数 $\theta$,而不含其他参数,并 $Z$ 且的分布已知且不依赖参数 $\theta$。

2)对给定的置信区间 $1-\alpha$,定出常数 $\lambda_1$、$\lambda_2$,使得：

$$P(\lambda_1\leqslant Z\leqslant\lambda_2)=1-\alpha \tag{16}$$

3)从不等式：

$$\lambda_1\leqslant Z(\xi_1,\cdots,\xi_n;\theta)\leqslant\lambda_2 \tag{17}$$

中解得 $\theta_1<\theta<\theta_2$，即为 $\theta$ 的置信系数为 $1-\alpha$ 的置信区间。

我们暂且认为待处理数据符合正态分布，因此对于任意的真实值 $y$ 和预测值$\hat{y}$之间存在关系为 $y=\hat{y}+\varepsilon$,其中 $\varepsilon$ 为误差,现假设 $\varepsilon\sim N(0,\sigma^2)$,于是 $y_0\sim N(\hat{y},\sigma^2)$,则有：$\frac{y_0-\hat{y}_0}{\sigma}\sim N(0,1)$。所以,根据置信区间的主元法可以得到：

$$Z=\frac{y_0-\hat{y}_0}{\sigma}\sim N(0,1) \tag{18}$$

则可求得置信度为 95%,即 $\alpha$=0.5 的预测区间为：

$$[\hat{y}_0-1.96\sigma,\hat{y}_0+1.96\sigma] \tag{19}$$

## 3　实　验

本文实验共基于两个数据集。

实验一对象为 46 005 条监控记录,记录时间为 2011 年 8 月 8 日 0 点至 10 点，记录地点为广东省某地区,监控地点为 20 个,每个地点监控 2–4 个不等车道。将以上数据按照时间间隔 5 分钟统计为车流量数据,则每个监控地点包含 120 个数据。

实验二对象为 158 088 条监控记录，记录时间为 2011 年 8 月 6 日 0 点至 10 点，记录地点为广东省某地区,监控地点为 22 个,每个地点监控 2–4 个不等车道。将以上数据按照时间间隔 5 分钟统计为车流量数据,则每个监控地点包含 120 个数据。

### 3.1　单断面与多断面对比实验

本组实验的主要目的是对单断面与多断面在短期交通流实时预测准确度方面进行对比，其中两组实验都是基于点预测，通过计算预测值与实际值的方差进行比较。方差越小,说明预测值与实际值越吻合,即预测精度越高。

本实验中多断面取值为 $N$=1,$M$=1(见 2.2 节中介绍的 $N$ 与 $M$),即上游取 1 个地点,下游取 1 个地点的前 1 个时刻和当前时刻的取值。若无上游或下游,则用自身数据填补。

**表 1　单断面与多断面对比实验结果表**

| 实验组 | 实验一 | 实验二 |
|---|---|---|
| 基于单断面预测方差 | 22163 | 120982 |
| 基于多断面预测方差 | 14211 | 59638 |

从表 1 和图 2 可以看出，基于多断面相似性预测方法要远小于基于单断面预测的方差，其中图 2

为某个点在 120 个时刻的走势图。同时观察图 1 可以看出，基于多断面相似性的预测结果和预测值要更吻合,而且基本趋势也更吻合。从现实角度考虑此问题，上游的车流量肯定会对当前位置的车流量有影响，也即当前位置的车流量也会对下游车辆有影响，因此下游车流量对于当前位置车流量的预测会有一定的帮助。同理历史时刻的车流量也会对当前时刻的车流量有一定的影响也合情合理。可以从图 3 中得出以上假想成立，从图中可以看出这几个断面的数据走势基本趋同。

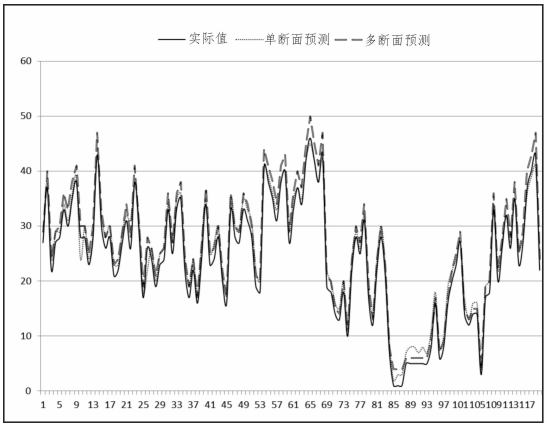


图 2　实验一 20602406 号监测点的实际值、单断面与多断面预测值图

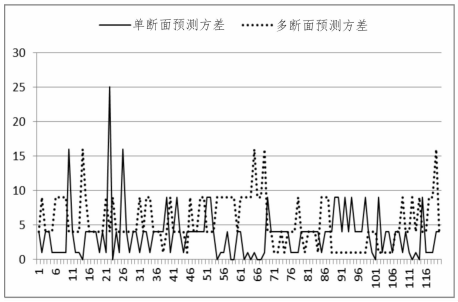


图 3　实验一 20602406 号监测点的单断面与多断面预测值方差图

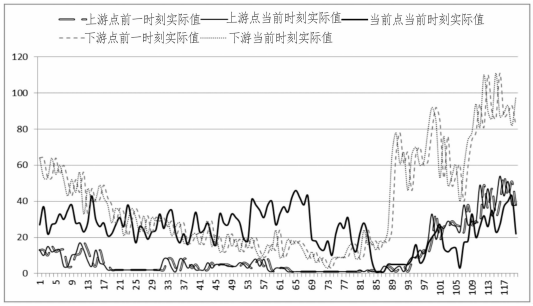


图 4　实验一 20602406 号监测点的多断面实际值图

### 3.2　点预测与区间预测实验

由于区间预测是基于点预测的，因此无法比较两者孰优孰劣。所以本组实验的主要目的是对区间预测在短期交通流实时预测中能提供更多容错信息进行验证,其中两组实验都是基于多断面相似性预测,通过计算实际值位于预测区间的准确率进行验证。

实验一中,经过计算所有样本数据的标准差 $\sigma=3$;实验二中,经过计算所有样本数据的标准差为 $\sigma=7$。所以,根据公式(19)可以计算出数据的预测区间。

**表 2　区间预测实验结果表**

| 实验组 | 实验一(%) | 实验二(%) |
|---|---|---|
| 区间预测准确率 | 88.79 | 84.71 |

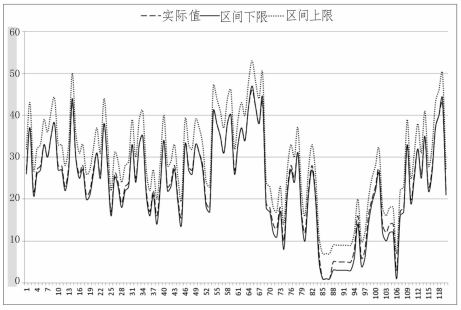


图 5　实验一 20602406 号监测点预测区间与实际值关系图

从图中可以很明显的看到粗虚线的实际值大部分处于区间范围之内，说明区间预测能提供更多的有用信息。

## 4　结束语

文中提出了基于多断面相关性的交通流预测方法。通过综合考虑上游车流量的扩散作用、下游车流量的阻力作用以及历史时刻对当前时刻的车流量的作用,构造多维度训练数据,使得每一个训练样本能尽可能包含多的可变因素,从而提高预测准确度;同时,本文也提出了对交通流的区间预测方法。基于统计学理论,根据点预测计算出区间预测结果,通过给出预测区间的方式,提供更多信息,让使用者能够更好地估计车流量的范围。

文中通过实验验证了方法的可行性。但尚有多点可以改进的地方。下一步计划完善多断面中 $N$ 与 $M$ 的取值方式,因为现在的取值都是经验值。另外一个则是本文默认的数据分布为正态分布，因此计算结果会有一定的误差，所以下一步需要研究的方向则是数据的分布问题。

**参考文献：**

[1] Samia Djemai,Belkacem Brahmi,Mohand Ouamer Bibi.A Primal-Dual Method for SVM Training [J]. Neurocomputing，2016,36(1):132-138.

[2] Zhanquan Sun,Geoffrey Fox.Traffic flow forecasting based on combination of multidimensional scaling and SVM [J]. International Journal of Intelligent Transportation Systems Research, 2014,12(1):20–25.

[3] ZHANG Ming–heng,ZHEN Yao–bao,HUI Gang–long,et al. Accurate Multisteps Traffic Flow Prediction Based on SVM [J]. Mathematical Problems in Engineering, 2013(1024–123x):1–8.

[4] 崔立成. 基于多断面信息的城市道路网交通流预测方法研究[D]. 大连:大连海事大学,2012.

[5] Chenyun Yu,Ka Chi Lam. Applying multiple kernel learning and support vector machine for solving the multicriteria and nonlinearity problems of traffic flow prediction[J]. J. Adv. Transp., 2014, 48(3):272–286.

[6] 丁世飞,齐丙娟,谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报,2011(1):8–10.

[7] 周桐,杨智勇,孙棣华.分车型的高速公路短时交通流量预测方法研究[J]. 计算机应用研究, 2015(7):13–15.

[8] Javad Abdi,Behzad Moshiri,Baher Abdulhai,et al. Short–term traffic flow forecasting: parametric and nonparametric approaches via emotional temporal difference learning [J]. Neural Computing and Applications, 2013,23(1):141–159.

[9] Hasan,Md Al Mehedi,Nasser,Mohammed,Pal, Biprodip,Ahmad,Shamim,Support Vector Machine and Random Forest Modeling for Intrusion Detection System (IDS)[J]. Journal of Intelligent Learning Systems and Applications, 2014,6(1): 45–52.

[10]Zhuangquan Sun,Geoffrey Fox. Traffic Flow Forecasting Based on Combination of Multidimen-sional Scaling and SVM [J]. International Journal of Intelligent Transportation Systems Research, 2014, 12(1):20–25.

[11]Wang Y, Papageorgiou M. Real –time freeway traffic state estimation based on extended Kalman filter: a general approach [J]. Transportation Research Part B Methodological, 2005,39(2):141–167.

[12]Haykin Simon.神经网络与机器学习[M].3版.申富饶,徐烨,郑俊,译.北京:机械工业出版社,2011.

[13]刘齐林,曾玲,曾祥艳. 基于支持向量机的区间模糊数时间序列预测 [J]. 数学的实践与认识,2015(11):14–16.

[14]刘岩,张宁,邵星杰. 城市轨道交通断面客流短时预测[J].都市快轨交通, 2015(1):23–25.

[15]袁键,范炳全. 交通流短时预测研究进展[J]. 城市交通, 2012(6):12–14.

(上接第 9 页)

(10):25–35.

[8] 蓝昊慧. 云计算在Web结构挖掘算法中的运用研究[J]. 计算机时代, 2012(10):30–33.

[9] 李明明,李伟. 基于HDFS的高可靠性存储系统的研究[J]. 西安科技大学学报, 2016,36(3):428–433.

[10]王意洁,孙伟东,周松,等. 云计算环境下的分布存储关键技术[J]. 软件学报, 2012,23(4):962–986.

[11]马志强,杨双涛,闫瑞,等. SQL–DFS:一种基于HDFS的海量小文件存储系统 [J]. 北京工业大学学报, 2016,42(1):134–141.

[12]汤羽,王英杰,范爱华,等. 基于HDFS开源架构与多级索引表的海量数据检索mDHT算法 [J]. 计算机科学, 2013,40(2):195–199.

[13]周相兵,马洪江,苗放. 云计算环境下的一种基于Hbase的ORM设计实现[J]. 西南师范大学学报:自然科学版, 2013,38(8):130–135.

[14]陈吉荣,乐嘉锦. 基于MapReduce的Hadoop大表导入编程模型[J]. 计算机应用, 2013,33(9):2486–2489.

[15]李洪敏,卢敏,黄林,等. 基于云计算技术的网络告警融合分析系统的设计与实现[J]. 信息安全与技术, 2014(9):58–63.

[16]张钊,张新峰,郑楠,等. 基于Hadoop平台的LDA算法的并行化实现 [J]. 计算机工程与科学, 2016,38(2):231–239.