计算机科学
COMPUTER SCIENCE

基于TCN-A模型的高效查询负载预测算法

白文超, 白淑雯, 韩希先, 赵禹博

引用本文

白文超, 白淑雯, 韩希先, 赵禹博. 基于TCN-A模型的高效查询负载预测算法[J]. 计算机科学, 2024, 51(7): 71-79.

BAI Wenchao, BAI Shuwen, HAN Xixian, ZHAO Yubo. Efficient Query Workload Prediction Algorithm Based on TCN-A [J]. Computer Science, 2024, 51(7): 71-79.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231100200

基于 TCN-A 模型的高效查询负载预测算法

白文超[1]　白淑雯[2]　韩希先[3]　赵禹博[3]
1 哈尔滨工业大学计算学部　哈尔滨 150001
2 开封大学信息化管理中心　河南 开封 475004
3 哈尔滨工业大学计算学部　山东 威海 264209

摘　要　针对大数据查询领域中出现的由于查询负载随时间动态变化且难以有效预测所导致的数据库管理系统无法及时优化的问题,提出了一种基于新型时间序列预测模型的查询负载预测算法。首先,该算法采用过滤、时域间隔划分以及查询负载构造等技术对原始的历史用户查询进行预处理,得到便于网络模型分析处理的查询负载序列。其次,所提算法以时间卷积神经网络为核心构建时序预测模型,提取查询负载数据的历史变化趋势及自相关性特征,高效地实现时序预测;同时,融入设计的时域注意力机制,对查询负载序列进行重要性加权,保证模型的分析计算效率,提升算法的预测性能。最后,基于上述时序预测模型,充分利用查询间隔时间完成对未来查询负载的精确预测,使得数据库管理系统得以预先实现自身性能调优,以适应工作负载的动态变化。实验结果表明,设计的查询负载预测算法在多个评价指标中均表现出良好的预测性能,并且能够在查询时间间隔内更加精确地预测未来查询负载的变化。

关键词:时间卷积神经网络;注意力机制;查询负载
中图分类号　TP302

Efficient Query Workload Prediction Algorithm Based on TCN-A

BAI Wenchao[1],BAI Shuwen[2],HAN Xixian[3] and ZHAO Yubo[3]
1 Faculty of Computing,Harbin Institute of Technology,Harbin 150001,China
2 Information Technology Management Center,Kaifeng University,Kaifeng,Henan 475004,China
3 Faculty of Computing,Harbin Institute of Technology,Weihai,Shandong 264209,China

Abstract　The query workload prediction algorithm based on a novel time series prediction model is proposed to address the problem of database management system cannot be optimized in time due to the dynamic change of query workload and the difficulty of forecasting effectively in the field of big data querying. First of all,the algorithm preprocesses the original historical users' queries by filtering,temporal interval partition and query workload construction to obtain the query workload sequence which is convenient for the network model to analyze and process. Secondly,the algorithm constructs a time series prediction model with temporal convolution network as the core,extracts the historical trend and auto-correlation characteristics of query workload,and realizes the time series prediction efficiently. At the same time,the algorithm integrates the designed temporal attention mechanism to weight the important query workloads to ensure that the query workload sequence can be analyzed and calculated efficiently by the model,and thus improving the performance of prediction algorithm. Finally,the algorithm uses the above time series prediction model to make full use of the query interval time to accurately predict the future query workloads,so that the database management system can achieve self-performance tuning in advance to adapt to the dynamic change of the workloads. Experimental results show that the designed query workload prediction algorithm exhibits good prediction performance on several evaluation metrics and is able to predict future query workload accurately over the query time interval.

Keywords　Temporal convolutional network,Attention mechanism,Query workload

1　引言

随着信息技术的快速发展,数据量以爆炸性的速度持续增长,使得传统的数据库管理系统(Database Management System,DBMS)不得不提高自身复杂性以应对规模庞大的查询负载的处理需求,进而增大了自身性能调优的开销[1-3]。而在实际的查询过程中,查询负载随时间动态变化,致使 DBMS 在当前时刻所调整的优化策略难以适用于未来的工作负载

到稿日期:2023-11-29　返修日期:2024-04-26
通信作者:白文超(Baiwenchao1998@163.com)

需求[4-5]。同时,在交互式查询分析任务中,查询负载存在明显的周期性,且相邻查询之间往往存在一定的时间间隔而未能有效利用[6]。因此,如果能够根据历史查询数据捕获查询工作负载的变化趋势,并充分利用查询间隔时间实现对查询负载的精确预测,则可以使得 DBMS 预先进行自身优化策略的动态调整,提高数据库管理系统的查询处理量和处理效率,进而高效地回答用户查询。

查询负载预测是近年来大数据查询领域的一大研究热点,其采用各种机器学习技术,根据历史用户查询数据,在查询间隔时间内对 DBMS 未来的查询工作负载进行预测,使得数据库管理系统得以预先实现自身性能的调优,以适应工作负载的动态变化,从而提升对用户查询的计算分析效率[7-8]。目前查询负载预测方法大致可分为 3 类[9]:第一类是经典的时间序列分析,它以时域和频域分析的方法,实现对历史查询负载的模式转化,以捕获查询工作负载的变化趋势及周期性特征,进而实现对未来查询负载的高效预测。此类方法原理简单,通用性强,但预测性能取决于查询负载数据的质量与稳定性,尤其在面临高度波动的负载数据时,经典的时间序列分析方法无法捕获查询负载的变化规律,难以实现精确预测。第二类是随机过程建模,与经典的时间序列分析不同,该方法侧重于对查询负载的概率属性的提取,并采用诸如马尔可夫模型等技术来预测各负载模式的概率取值,进而自动选择 DBMS 可以使用的优化。但是,随机过程建模方法的实现更多依赖于当前数据,与历史负载数据的联系不够紧密,致使其难以捕获历史查询负载数据的周期性变化趋势,降低了查询负载预测的准确性。第三类是基于深度学习的方法实现查询负载预测,其通过使用循环神经网络、长短期记忆网络等时序网络模型,以历史查询负载数据为驱动,高效地预测未来查询工作负载的变化。但是上述时序模型中循环神经网络存在长期依赖问题,容易发生梯度弥散;长短期记忆网络可以克服长期依赖,但在记忆序列超长的情况下,该模型仍然存在梯度减弱的问题,从而不能很好地捕获长时间负载序列的历史变化。

针对上述问题,本文提出了一种新型的查询负载预测算法。首先,本文对给定的历史查询数据进行预处理,去除无效、低质的用户查询,并将其按时域窗口进行切割、划分,以构造查询负载数据。其次,本文设计了一种新型的时间序列预测模型,该模型融合高性能的时间卷积神经网络以及设计的时域注意力机制,从而更好地捕获序列数据在时间维度上的相关性特征,更加高效地完成时序预测。最后,本文基于上述时序模型,充分利用查询间隔时间来完成对未来查询负载的精确预测,使得 DBMS 能够预先实现自身性能调优,从而更好地处理交互式查询。

综上所述,本文的贡献在于:

1)对原始查询数据进行预处理,构造形成便于网络模型分析训练的查询负载序列向量。

2)提出并实现了一种新型的时间序列预测模型,以高效的时间卷积神经网络为核心,融入设计的时域注意力机制,捕获历史查询负载的变化信息及相关性特征,并充分利用查询间隔时间快速地实现查询负载的精确预测。

3)在真实的数据集中进行实验,结果表明,本文提出的查询负载预测算法在多个数据集上均取得了较好的预测性能。

2 相关工作

本章对 3 类查询负载预测方法的相关工作进行了综述,包括经典的时间序列分析、随机过程建模,以及基于深度学习的查询负载预测。

经典的时间序列分析是目前应用极为广泛的查询负载预测算法之一。基于该算法,Taft 等[10]提出使用稀疏周期自回归的时序分析方法来准确预测不同应用场景下查询工作负载随时间的变化趋势,并设计了一种新型的动态规划算法来对数据库管理系统进行调度配置,实现成本节约。Elnaffar 等[11]设计并实现了一种高效的时间序列分析框架 Psychic-Skeptic,通过结合离线分析与在线预测的方法来捕获查询负载的主要变化,使得 DBMS 能够高效地进行自我调优。Higginson 等[12]提出使用自回归综合移动平均的时间序列分析技术来学习数据库自身工作负载的周期性特征和变化模式,从而更加高效地预测未来查询负载的变化情况。Lorido-Botran 等[13]使用时间序列分析的方法来预测云环境下查询负载的未来变化,从而使得 DBMS 能够预先进行自适应的性能优化。

随机过程建模是查询预测领域为提高查询负载预测性能而广泛研究的一大热点。Pavlo 等[14]提出了一种新型的查询负载预测框架 Houdini,其采用马尔可夫模型来估计未来查询负载的概率属性,从而提高预测的灵活性与准确率。Holze 等[15]设计了一种简洁、通用的查询工作负载预测方法,该方法基于 N-Gram 模型来学习数据库管理系统中出现的常见的工作负载模式,以实现 DBMS 的自适应调优。此外,Du 等[16]将马尔可夫模型与 Petri 网络相结合,在工作流驱动的模式下实现对 OLTP 查询工作负载的建模与预测,进而保证 DBMS 自适应优化自身性能。

近年来,采用深度学习方法实现查询负载预测已成为查询预测领域的另一大研究热点。Ma 等[17]采用集成学习的思想,将循环神经网络与回归模型相结合,从而高效地对 DBMS 历史数据中的各种查询负载模式进行提取、预测,以适应查询负载的动态变化;Shahrivari 等[18]基于循环神经网络捕获历史工作负载的变化模式,高效预测未来查询负载,以实现自适应近似查询处理。但上述方法存在长期依赖问题,在面临较长的查询语句序列时容易发生梯度弥散。Durand 等[19]基于强化学习的思想,采用 Q-learning 网络对历史查询负载信息进行建模,实现对未来查询负载的动态预测。Huang 等[20]利用多头注意力机制和卷积操作设计了一种新型轻量级查询负载预测框架,以帮助 DBMS 高效预测未来查询的工作负载模式。Mozafari 等[21]提出了一种新型的查询负载预测机制

DBSeer，其根据查询事务类型对历史查询负载进行聚类，并基于集成预测模型实现对查询负载的高效预测。Jain 等[22]以及 Huang 等[23]均使用优化的长短期记忆网络，以历史查询数据为驱动，捕获查询负载的变化规律，使得 DBMS 能够动态调整自身优化策略，提高查询处理效率。但此类算法由于所用网络自身的固有缺陷，无法解决梯度减弱问题，难以应对较长的查询负载序列。

3 基本概念及形式化定义

1）查询负载

查询负载指数据库管理系统在一段时间间隔内所承载的所有用户查询的数量，是影响数据库系统性能的重要指标。DBMS 会根据当前自身的查询负载情况设置对应的优化策略，实现性能调优，以保证系统对用户查询的处理效率。而将查询负载数据在历史上的各个时刻所对应的实际取值按时间顺序归并组合所构成的序列数据称为查询负载序列。

2）时间卷积神经网络

时间卷积神经网络（Temporal Convolutional Network，TCN）是一种有效的时序预测模型，具有内存开销低、感受野灵活以及可并行化处理等优势。其通过将因果卷积与膨胀卷积相结合，可以很好地克服卷积核大小对时序建模的限制，学习更长时的历史信息，且通过引入残差网络模块，保证了模型梯度的稳定，从而更好地解决序列预测问题。

3）注意力机制

注意力机制（Attention Mechanism，A）是机器学习领域中为提升网络模型分析处理效率而被广泛应用的一种数据处理技术。该方法可在网络模型分析计算能力有限的情况下，令模型关注更为重要的数据特征，以解决数据规模过大而导致的信息超载问题。

定义 1（查询间隔时间） 给定一组连续查询 $Q=\{q_1,\cdots,q_n\}$，对于任意的两个相邻查询 q_i 与 q_{i+1}，假设其各自的查询时刻分别为 t_i 和 t_{i+1}，则查询间隔时间可定义为 $t_{i+1}-t_i$。

定义 2（自相关性特征） 给定任意一个查询负载序列 $s_t=\{s_{t(1)},\cdots,s_{t(n)}\}$，$s_{t(i)}$ 表示在 i 时刻序列 s_t 的实际取值，若存在关系 $F(\cdot)$ 使得 $s_{t(n+1)}=F(s_{t(1)},\cdots,s_{t(n)})$，则称查询负载序列 s_t 存在自相关性，而依赖关系 $F(\cdot)$ 为 s_t 的自相关性特征。

4 基于新型时间序列预测模型的查询负载预测算法

本章首先对原始用户查询数据进行预处理，构建查询负载序列；然后，设计了一种新型的时间序列预测模型，并给出了该模型的构建方法与训练过程；最后，基于上述模型充分地利用查询间隔时间，快速地实现查询负载的精确预测。

4.1 数据预处理

在建立查询负载预测模型之前，须对 DBMS 所包含的原始用户查询进行预处理，构造查询负载数据，并将其编码为便于网络模型训练处理的序列向量，具体流程如算法 1 所示。

算法 1 数据预处理算法

输入：各查询起始时间 ST$\{st_1,\cdots,st_n\}$，各查询终止时间 ET$\{et_1,\cdots,et_n\}$，时间间隔阈值 threshold，原始查询负载数据 QD

输出：预处理后的数据 Result

```
1. secList←sectionSplit(QD,threshold,ST,ET)
2. For i←1 to |QD| do
3.     If WhetherInvalid(QDi)= True Then
4.       Continue
5.     End If
6.     Flag←AreaSelect(sti,eti)
7.     For j←1 to |Flag| do
8.       secList[Flagj]←secList[Flagj]+1
9.     End For
10. End For
11. Result←secList
12. Return Result
```

本文在数据预处理阶段对 DBMS 所承载的用户查询数据分别进行时域间隔划分、过滤以及查询负载构造等操作，而时域间隔划分也为本阶段的核心难点。在实现过程中，本文首先设计了时域划分方法 sectionSplit，根据历史查询数据集的起止时间及指定的时域窗口大小对历史查询数据进行时域划分，形成各查询负载数据所对应的时域区间（见算法 1 第 1 行）。然后，依次遍历历史用户查询，基于查询访问情况对原始查询数据进行过滤处理，去除存在语句残缺、语法错误等无效且低质量的查询数据以及由 DBMS 发出的用于检查、测试系统的 SQL 查询（见算法 1 第 3－5 行）。最后，遍历过滤后的各个查询的起止时刻，利用设计的 AreaSelect 函数计算当前查询所涵盖的时域区间（见算法 1 第 6 行），并逐个更新这些时域区间所包含的查询数目，以组成查询负载序列数据（见算法 1 第 7－12 行），从而便于网络模型分析处理，提升算法整体的预测效率。

4.2 模型简介

经典的时间序列预测模型主要包括循环神经网络（Recurrent Neural Network，RNN）以及长短期记忆网络（Long Short-Term Memory，LSTM）。

其中，RNN 是一种常见的时序预测模型，因擅长处理时间序列问题而被广泛应用，但该模型存在长期依赖问题，容易发生梯度弥散。而 LSTM 是对 RNN 的一种有效改进，其引入了细胞单元状态和门限机制，通过利用选择性遗忘与记忆历史信息的方法来改善长期依赖，提升了整体模型的预测效率，但是当面临超长的序列数据时，LSTM 仍然会发生梯度减弱以及信息衰减现象，影响最终的预测结果。

针对上述问题，本文提出了一种新型的时间序列预测模型——基于注意力机制的时间卷积神经网络（Temporal Convolutional Network-Attention，TCN-A）。该模型以时间卷积神经网络为核心，从时间维度上综合提取查询负载序列的自相关性特征，从而更好地捕获查询工作负载的变化趋势，实现对 DBMS 未来查询负载的高效预测；同时，融入设计的时域注意力机制及全连接网络，解决超长序列依赖问题，提高模型整体的预测效率。

4.3 模型构建

4.3.1 模型总体结构

本模型以 TCN 的网络结构为基础，引入设计的时域注意力机制以及全连接神经网络（Full Connect Network，FCN）保证了总体模型的预测性能。具体模型结构如图 1 所示。

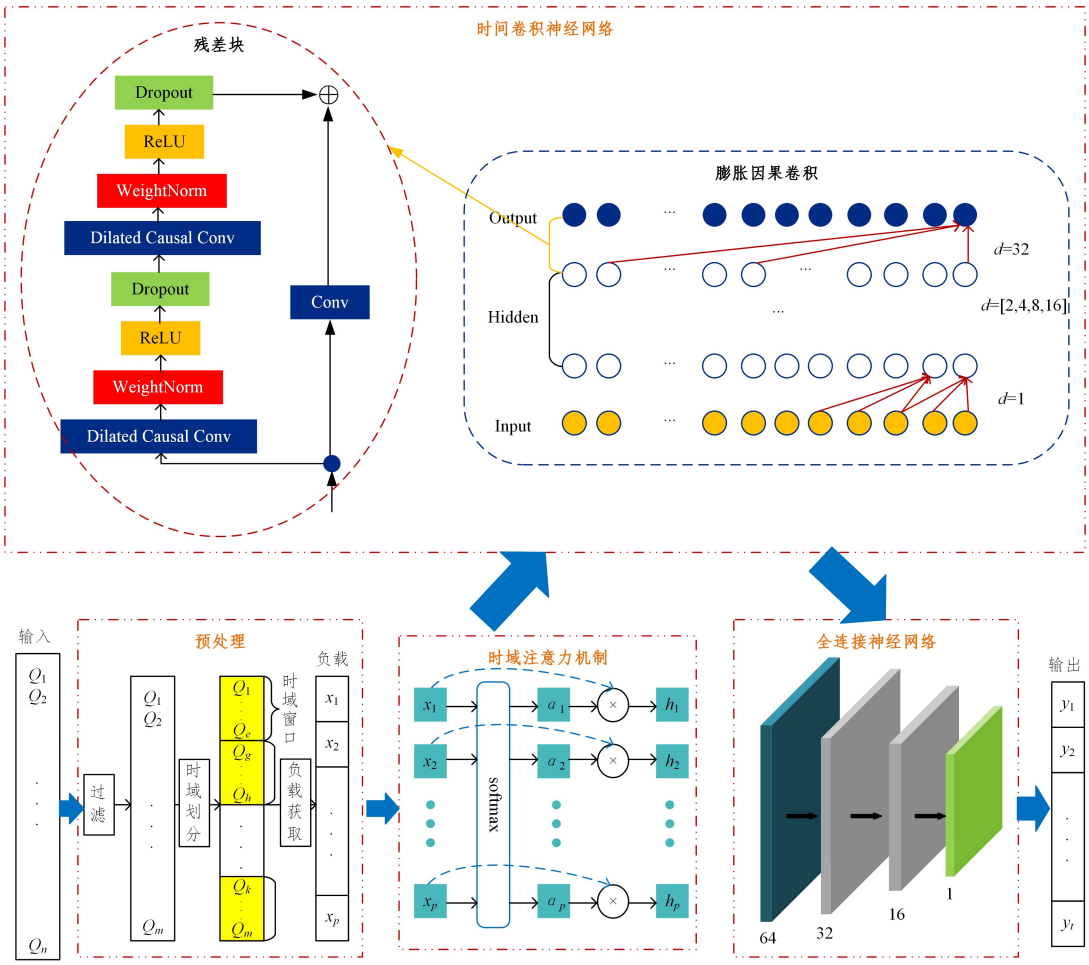


图 1 基于注意力机制的时间卷积神经网络模型结构图

Fig. 1 Structure of attention-based temporal convolutional network model

本模型中各模块的具体描述如下。

1）时域注意力机制

为更好地捕获查询负载序列历史上的有效信息，本模型设计了一种高效的时域注意力机制，从历史查询负载序列中学习各个输入数据特征的重要程度，并按重要性对输入特征赋予不同的权值。本模型将原始查询负载序列作为注意力机制的输入，采用 softmax 函数生成权重矩阵，保证对查询负载序列影响较大的特征被赋予更大的权值，具体计算式如式(1)、式(2)所示：

$$\alpha_t=\frac{\exp(x_t)}{\sum_{i=0}^{n}\exp(x_i)} \tag{1}$$

$$h_t=\alpha_t * x_t \tag{2}$$

其中，α_t 表示注意力机制对输入数据 x_t 进行计算得到的注意力权重；h_t 表示查询负载数据 x_t 经过注意力机制的输出值；n 表示查询负载序列的长度。

2）时间卷积神经网络

为学习查询负载序列的自相关性特征及其变化趋势，本文使用 TCN 网络从时间维度上对经过时域注意力机制加权后的查询负载序列进行预测。本模型的 TCN 模块共含有 7 层网络，分别为输入层、5 层隐藏层、输出层，各卷积层的卷积核大小均为 3，且均以 ReLU 为激活函数；自输入至输出，膨胀因子 d 取值分别为 1，2，4，8，16，32；同时引入残差块结构，克服梯度弥散，保证模型性能。

3）全连接网络

本模型所构建的全连接网络模块共包含 4 层全连接神经网络，用于将 TCN 网络计算得到的特征空间映射为数据标记空间，减小特征位置对预测结果的影响，提升整体时序预测模型的鲁棒性；同时将当前得到的 TCN 的初步预测值加以整合，以得到最终的预测结果，进而实现对未来查询负载的精确预测。其中，各层网络的神经元数量分别为 64，32，16，1，并均采用 Leaky ReLU 作为激活函数。

4.3.2 模型重要参数设置

1）损失函数

本模型的误差分为两部分：TCN 网络误差与 FCN 网络误差。本文根据训练数据的实际特性以及具体的网络结构，选用稀疏类别交叉熵（Sparse Categorical Crossentropy，SCCE）以及均方误差（Mean Squared Error，MSE）函数分别对 TCN 与 FCN 的网络误差进行衡量，并通过最小化上述误差

来优化整体模型，具体表达式如式(3)、式(4)所示：

$$SCCE=\frac{1}{N}\sum_{i=1}^{N}(p_i\log_2 q_i+(1-p_i)\log_2(1-q_i)) \tag{3}$$

$$MSE=\frac{1}{N}\sum_{i=1}^{N}(y_i-y_i')^2 \tag{4}$$

其中，N 为样本个数，y_i 和 y_i' 分别表示真实值与预测值；p_i 和 q_i 表示真实数据与预测数据的概率编码。

2)优化器

本文通过实验对比发现：相比 Momentum 和 SGD 等经典的优化算法，Adam 优化器能适应稀疏梯度变化，缓和梯度震荡问题，从而保证收敛速度，便于模型调参；RMSPropOptimizer 能够更好地保证时间卷积神经网络在训练过程中误差梯度的稳定性，且能够修改传统的梯度积累为指数加权的移动平均，从而能自适应地调节学习率的变化。因此，本文使用 RMSPropOptimizer 与 Adam 分别作为 TCN 和 FCN 网络的优化器，以更好地优化网络模型参数。

本文综合考虑模型的收敛情况以及训练过程中误差大小等因素，将模型的学习率设置为 0.001。

4.4 模型算法描述

本文设计的基于注意力机制的时间卷积神经网络模型的实现过程分为两个阶段：时域注意力加权(Attention Weighting)与时序预测(Temporal Prediction)。TCN-A 模型的具体流程如算法 2 所示。

算法 2 TCN-A 模型

输入：模型 TCN-A{Attention，TCN，FCN}，查询工作负载 QW

输出：负载预测结果 Result

1. Initialize and train the Model
2. Attention Weighting DO：
3. afterData←Attention(QW)
4. Temporal Prediction DO：
5. juniorPredict←TCN(afterData)
6. Result←FCN(juniorPredict)
7. Return Result

在模型具体的执行过程中，本文首先以算法 1 所得到的查询负载序列为输入，通过设计的时域注意力机制计算当前查询负载序列所对应的权重矩阵并与之相乘，完成负载数据的时域注意力加权(见算法 2 第 2－3 行)。之后，利用 TCN 模块对加权后的负载数据进行时序预测，捕获当前查询负载序列的自相关性特征及变化趋势，实现时间维度上查询负载的精确预测(见算法 2 第 4－5 行)。最后，本算法将 TCN 网络得到的当前查询负载序列的初步预测结果进行重塑处理，将其作为全连接网络模块的输入，并利用各层全连接神经网络对得到的初步预测值加以整合，计算出未来查询工作负载的最终预测结果(见算法 2 第 6－7 行)。在上述过程中，如果在 TCN-A 模型预测完成之前，数据库系统的实际查询已到达，则说明当前算法预测失效，终止算法 2 的执行并将实际的查询负载数据返回给数据库系统；否则，完成算法 2，并将 TCN-A 模型得到的查询负载的最终预测值返回给 DBMS，从而使数据库管理系统预先实现自身优化策略的动态调整，以适应查询工作负载的波动性变化，提高 DBMS 对用户查询的处理量和处理效率，进而更加高效地回答用户查询。

5 实验与分析

5.1 实验设置

5.1.1 实验环境设置

实验硬件环境为 NVIDIA Tesla K80 GPU；16 GB 内存；500 GB 硬盘；操作系统为 Windows 10。本文采用 Pycharm 2020.2 编程环境与 Python 编程语言开发了模拟测试程序，使用 Keras 学习框架构建本文的预测模型。

5.1.2 实验数据集

本文选用两个真实数据集进行实验，分别为 Sky Survey 与 Bus Tracker。

1)Sky Survey 数据集：该数据集来源于斯隆数字化巡天网站(Sloan Digital Sky Survey)的用户查询日志。本文综合考虑实验成本及效果，选用最新的 1 000 万条连续的日志数据，并将其划分为训练集(70%)、验证集(10%)和测试集(20%)进行实验。

2)Bus Tracker 数据集：该数据集收集了 7 000 余万条 2016 年 11 月－2017 年 1 月期间用户在巴士搜索软件中对公交车信息的查询记录。本文按照 7∶1∶2 的比例将该数据集划分为训练集、验证集以及测试集，以进行本文实验。

5.1.3 实验评估指标

为更加全面地评价算法的预测效果，本文选用了 4 种评估指标，具体定义如下。

1)平均绝对百分比误差

平均绝对百分比误差(Mean Absolute Percentage Error，MAPE)是评估预测算法准确性的常用统计指标，可以很好地衡量模型的预测精度，具体表达式如式(5)所示：

$$MAPE=\frac{1}{N}\sum_{i=1}^{N}\frac{|(y_i-y_i')|}{y_i}\times 100\% \tag{5}$$

2)平均绝对误差率和均方根误差率

平均绝对误差(Mean Absolute Error，MAE)和均方根误差(Root Mean Squared Error，RMSE)是查询预测领域常见的用于判断预测模型预测值 y_i' 和实际值 y_i 之间差异的评估指标。但在具体应用中，MAE 和 RMSE 均未考虑实际数据的取值范围，且由于本文所处理的负载数据的具体取值较大，进而导致 MAE 与 RMSE 的实际取值较高，难以直观、精确地衡量算法预测值与实际值之间的偏差。因此，本文令 MAE 与 RMSE 各自除以当前负载数据实际值的平均数，分别得到两者对应的百分率 MAE% 和 RMSE%，实现对算法误差更加精确的评估，具体表达式如式(6)、式(7)所示：

$$MAE\%=\frac{\frac{1}{N}\sum_{i=1}^{N}|(y_i-y_i')|}{\frac{1}{N}\sum_{i=1}^{N}y_i} \tag{6}$$

$$RMSE\%=\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-y_i')^2}}{\frac{1}{N}\sum_{i=1}^{N}y_i} \tag{7}$$

3)均方对数误差

均方对数误差(Mean Squared Log Error，MSLE)是衡量模型预测未知样本效果的常见指标，具体表达式

如式(8)所示：

$$MSLE=\frac{1}{N}\sum_{i=1}^{N}(\log(y_i+1)-\log(y_i'+1))^2 \tag{8}$$

5.2 对比算法

为更好地体现本文提出的查询负载预测算法的准确性与高效性，实验选择以下对比算法。为了避免随机性，本文在实验过程中对设计的 TCN-A 及其他对比算法均重复执行 10 次，计算各指标的平均值。

1)对比算法

ARIMA 算法是一种经典的时间序列分析方法，因实现简单、预测精度较高而被广泛应用于查询负载预测领域。而 RNN 是查询负载预测方向中极为经典的模型算法，具有原理简单、通用性强等优势。此外，LSTM 与 GRU 也是查询负载预测领域中应用十分广泛的时序模型算法，它们均为 RNN 的变体且具有比 RNN 更高的预测性能。

因此，本文选用 ARIMA，RNN，GRU 以及 LSTM 模型作为本文的对比算法，并分别参照文献[12,18,22,24]在本实验平台进行实现。其中，ARIMA 算法的最大差分阶数设置为 3，RNN，GRU 与 LSTM 模型的学习率及迭代次数等参数取值均与 TCN-A 一致，分别设置为 0.001，500，且均使用 Adam 作为网络优化器。此外，RNN，GRU 与 LSTM 模型内部的单元数目均设置为 64，并分别在本实验平台调整其参数至最优状态。

2)消融算法

本文以 TCN 的网络结构为基础，融入设计的注意力机制及全连接网络，进而形成本文所提出的 TCN-A 算法框架。因此，为更好地验证 TCN-A 算法的高效性，本文对算法的注意力机制进行消融处理，获得不含注意力机制的 TCN 模型算法，并将其作为实验对比算法。同时，为更加充分、全面地验证注意力机制对算法的影响，本文在所选的对比算法 LSTM 的基础上融入与本文模型相同的注意力机制，构成 LSTM-A 算法并将其作为本文的对比算法。而且，TCN 与 LSTM-A 算法均在本实验平台实现，其具体参数设置分别与 TCN-A 及 LSTM 相同，并分别在本实验平台训练各自的模型参数至最优状态。

5.3 结果分析

5.3.1 预测性能分析

为了验证本文算法 TCN-A 的可行性和有效性，本文将查询负载的时域窗口大小设置为 1 min，在 5.1.2 节中所述的两个真实数据集上进行实验，并与其他对比算法进行对比，获取 TCN-A 及其他对比算法的 MAPE，MAE%，RMSE%以及 MSLE 等指标值，具体的实验结果如表 1 所列。

分析表 1 中各项实验数据可以发现，在两个真实数据集中，本文提出的 TCN-A 算法的 MAPE，MAE%，RMSE%以及 MSLE 等误差评估指标值均小于其他对比算法。以 Sky Survey 数据集为例，相比实验中的最佳对比算法 TCN，TCN-A 各误差指标值分别下降 0.008 9，0.003 3，0.004 9 以及 0.004 6。这说明本文所设计的 TCN-A 算法拥有更好的时序预测能力，能够更加高效地捕获当前查询负载序列的自相关性特征及历史变化趋势，进而实现更加准确的查询负载预测。

表 1 不同算法预测结果的对比

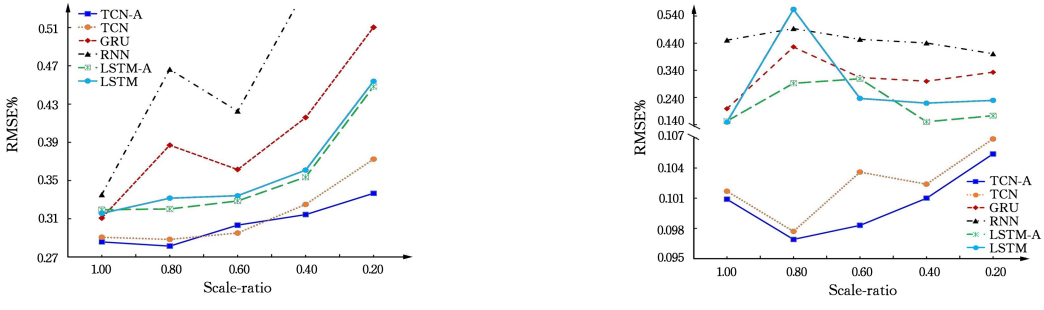
Table 1 Comparison of prediction results of different algorithms

数据集	对比算法	MAPE	MAE%	RMSE%	MSLE
Sky Survey	TCN-A	**0.2084**	**0.1895**	**0.2858**	**0.0617**
	TCN	0.2173	0.1928	0.2907	0.0663
	LSTM	0.3210	0.2382	0.3157	0.1382
	LSTM-A	0.3134	0.2265	0.3195	0.1158
	GRU	0.3081	0.2275	0.3108	0.1109
	RNN	0.4308	0.3009	0.3352	0.1827
	ARIMA	0.5169	0.3473	0.4513	0.2319
Bus Tracker	TCN-A	**0.1089**	**0.0707**	**0.1009**	**0.0186**
	TCN	0.1117	0.0734	0.1017	0.0193
	LSTM	0.1734	0.1157	0.1496	0.1305
	LSTM-A	0.1799	0.1203	0.1531	0.1296
	GRU	0.2279	0.0843	0.1988	0.1190
	RNN	0.6304	0.3939	0.4515	0.2802
	ARIMA	0.4728	0.3761	0.5838	0.2659

此外，算法 TCN-A 在两个数据集中的执行效果均优于 TCN，因此本文提出的 TCN-A 算法所融合的时域注意力机制对提高算法的预测准确性具有重要的意义。而且，TCN-A 在 Sky Survey 以及 Bus Tracker 数据集中的平均预测时间分别为 0.066 3 s 与 0.113 5 s，远低于上述两个数据集中的平均查询间隔时间，即 0.644 7 s 与 0.761 3 s。因此，本文提出的 TCN-A 模型可以快速地完成对未来查询负载的预测，使得 DBMS 能够充分地利用查询间隔时间实现自身性能的动态调优，提高用户查询的处理效率。

5.3.2 不同数据规模下的误差对比分析

为充分验证本文 TCN-A 算法在不同规模的数据集中的预测效果，本文分别在 5.1.2 节所述的两个真实数据集上设置时域窗口大小为 1 min，并逐次按比例减小其训练样本的数目以进行对比实验分析，具体的实验结果如图 2 所示。



(a)Sky Survey 数据集的 RMSE% (b)Bus Tracker 数据集的 RMSE%

图 2 各数据集中不同数据规模下各算法的 RMSE%对比图

Fig. 2 Comparison diagram of RMSE% of each algorithm with different data sizes in each dataset

其中，各轮修改后的训练样本的数目占原始训练数据的比例(Scale-ratio)分别为 0.2，0.4，0.6，0.8 以及 1.0。且受篇幅所限，本文仅选取 RMSE%作为本节的评估指标，并以此为例，分析不同数据规模下算法的预测误差变化。此外，在比例参数的变化过程中，ARIMA 算法的预测误差均远高于本文中的其他对比算法。为精确地表示其他算法的运行效果，本文并未在图 2 中显示 ARIMA 的执行结果，该算法的具体结果如表 2 所列。

表 2　不同数据规模下 ARIMA 算法的 RMSE

Table 2　RMSE of ARIMA algorithm with different data sizes

(%)

数据集	Scale-ratio				
	0.2	0.4	0.6	0.8	1
Sky Survey	0.6986	0.6697	1.0600	0.4225	0.4513
Bus Tracker	0.9226	0.9297	0.4705	0.6720	0.5838

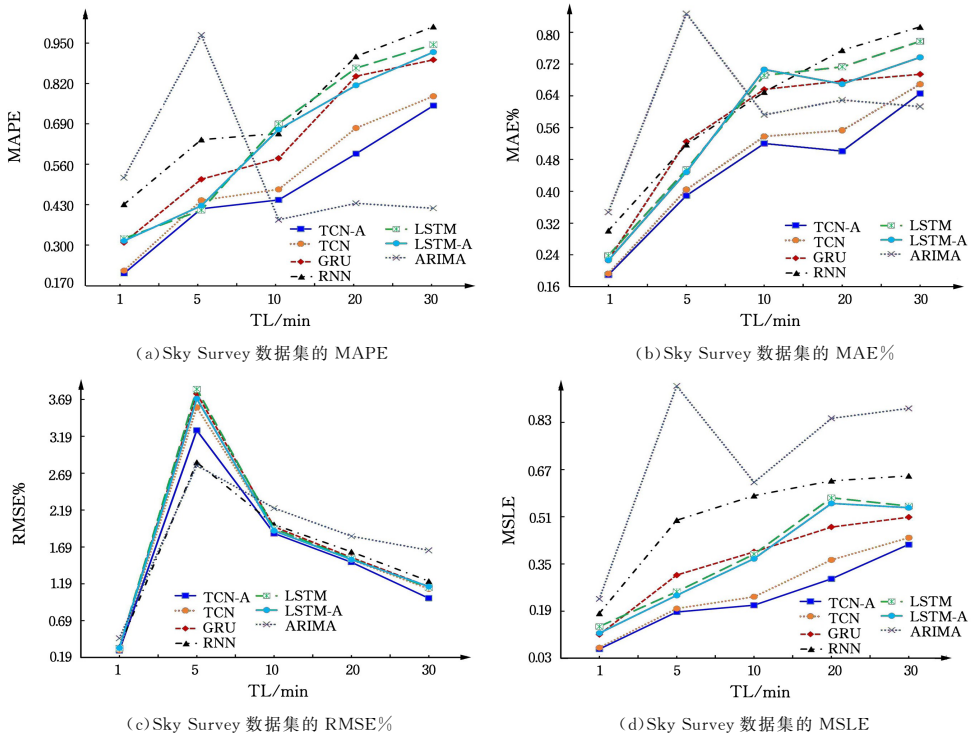
由图 2 可以看出，在 Sky Survey 以及 Bus Tracker 数据集中，随着比例参数 Scale-ratio 的变化，TCN-A 模型的 RMSE%取值整体低于其他对比算法。尤其在 Sky Survey 数据集中，TCN-A 的预测效果远优于其他对比算法，且随着训练样本的减少，TCN-A 与其他对比算法之间的取值差异也在逐步增大，这表明本文所设计的 TCN-A 能够更加有效地弥补训练数据不足所带来的缺陷。而在 Bus Tracker 数据集中，随着数据规模的变化，TCN-A 及 TCN 的预测误差均远低于其他对比算法，且两者之间的 RMSE%值差异也随着训练样本的减少而更加明显。

因此，本文提出的 TCN-A 模型能够在不同规模的训练数据下，更加准确地完成对未来查询负载的预测，使得 DBMS 能够基于查询负载预测值实现自身性能的动态调优，提高用户查询的处理效率。

5.3.3　不同时域窗口大小下误差对比分析

为充分验证本文所设计的 TCN-A 模型在不同时域窗口大小的查询负载数据中的预测效果，本文选用 5.1.2 节中所述的 Sky Survey 以及 Bus Tracker 两个真实数据集为基准数据，分别设置时域窗口大小为 1 min，5 min，10 min，20 min 以及 30 min，并根据本文算法 1 构造不同时域窗口大小下的查询负载数据。之后，将本文提出的 TCN-A 模型在上述各查询负载序列中进行实验，获取该方法与其他对比算法的 MAPE，MAE%，RMSE%以及 MSLE 等评估指标值并进行误差对比分析。设计的 TCN-A 及其他算法的具体实验结果分别如图 3、图 4 所示。

分析图 3 可知，在 Sky Survey 数据集中，本文提出的 TCN-A 算法在部分时域窗口下的误差取值高于 ARIMA 等算法，但总体来看，随着时域窗口取值的变化，本文提出的 TCN-A 模型的预测精度整体上均优于对比算法。且以 MAPE，MAE%为例，TCN-A 与 TCN 之间的性能差异随着时域窗口的增大而愈加明显。



(a)Sky Survey 数据集的 MAPE　(b)Sky Survey 数据集的 MAE%

(c)Sky Survey 数据集的 RMSE%　(d)Sky Survey 数据集的 MSLE

图 3　Sky Survey 数据集中不同时域窗口大小各算法性能对比图

Fig. 3　Performance comparison of each algorithm with different time lengths in Sky Survey

根据图 4 可知，在 Bus Tracker 数据集中，随着时域窗口的变化，本文所提出的 TCN-A 模型算法的预测精度均明显优于除 TCN 外的其他所有对比算法，TCN-A，TCN，GRU，LSTM 以及 LSTM-A 等算法的误差结果也远低于 RNN，ARIMA 等经典的时序预测算法，且随着时域窗口大小的增加，两类算法之间的 MAPE，MAE%，RMSE%及 MSLE 等指标的差异也在逐步增大。此外，以 MAE%为例，随着时域窗口的增大，TCN 与 TCN-A 之间的取值差异也由 0.0027 增大至 0.0046，性能差异整体呈增大趋势。而且，在多数时域窗口中，LSTM-A 的 MAPE，MAE%以及 MSLE 等误差指标值均明显优于 LSTM 算法。

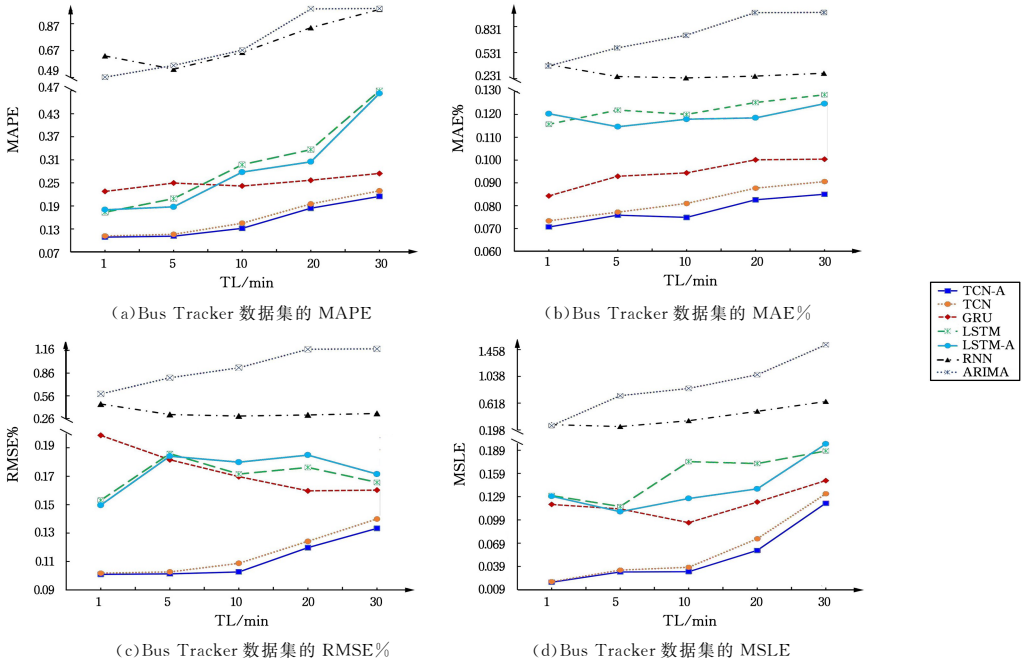


图 4 Bus Tracker 数据集中不同时域窗口大小下各算法性能的对比

Fig. 4 Performance comparison of each algorithm with different time lengths in Bus Tracker

综上所述，随着时域窗口大小的变化，本文所提出的 TCN-A 算法能够更加准确地预测未来查询负载的具体取值，且本文所设计的时域注意力机制对提高算法的预测性能具有重要的意义。

结束语 本文提出了基于 TCN-A 模型的查询负载预测算法。首先，本文采用过滤、时域间隔划分以及查询负载构造等方法对原始的历史用户查询进行预处理，以得到便于网络模型分析处理的查询负载序列数据。其次，本文构建了一种新型的时间序列模型，以实现查询负载预测。该模型以高性能的 TCN 网络结构为基础，融入设计的时域注意力机制，从而更加高效地提取查询负载数据的变化趋势及自相关性特征，并充分利用查询间隔时间来完成对 DBMS 未来查询负载的精确预测，使得数据库管理系统得以预先实现自身性能调优，以适应工作负载的动态变化。实验结果表明，本文模型可以快速地实现对 DBMS 未来查询负载的预测，且该算法的预测精度显著优于其他对比算法。

然而，本文提出的 TCN-A 算法由于网络结构庞大且复杂，使得模型完成单次预测所需的时间长于 RNN 与 GRU 等经典算法。因此，未来将探究对模型进行微调，以优化运行成本较高的问题，进一步降低模型的预测开销。

参 考 文 献

[1] LIU C, MAO W, GAO Y, et al. Adaptive recollected RNN for workload forecasting in database-as-a-service[C]// 18th International Conference Service-Oriented Computing(ICSOC). Berlin: Springer, 2020: 431-438.

[2] SHAHEEN N, RAZA B, SHAHID A R, et al. A novel optimized case-based reasoning approach with k-means clustering and genetic algorithm for predicting multi-class workload characterization in autonomic database and data warehouse system[J]. IEEE Access, 2020, 8(1): 105713-105727.

[3] SHAHEEN N, RAZA B, SHAHID A R, et al. Autonomic workload performance modeling for large-scale databases and data warehouses through deep belief network with data augmentation using conditional generative adversarial networks[J]. IEEE Access, 2021, 9(1): 97603-97620.

[4] QIAN H, WEN Q, SUN L, et al. RobustScaler: QoS-Aware autoscaling for complex workloads[C]// 2022 IEEE 38th International Conference on Data Engineering (ICDE). Piscataway: IEEE, 2022: 2762-2775.

[5] YUAN Z, CHEN H, HUANG Z, et al. A lightweight general adaptive optimization tool for relational DBMSs under HTAP workloads[C]// 2022 IEEE International Conference on Services Computing(SCC). Piscataway: IEEE, 2022: 45-53.

[6] MEDURI V V, CHOWDHURY K, SARWAT M. Evaluation of machine learning algorithms in predicting the next SQL query from the future[J]. ACM Transactions on Database Systems (TODS), 2021, 46(1): 1-46.

[7] ZHI KANG J K, GAURAV, TAN S Y, et al. Efficient deep learning pipelines for accurate cost estimations over large scale query workload[C]// Proceedings of the 2021 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2021: 1014-1022.

[8] TANG C, WANG B, LUO Z, et al. Forecasting SQL query cost at twitter[C]// 2021 IEEE International Conference on Cloud Engineering(IC2E). Piscataway: IEEE, 2021: 154-160.

[9] YAN Z, LU J, CHAINANI N, et al. Workload-Aware performance tuning for autonomous DBMSs[C]// 2021 IEEE 37th International Conference on Data Engineering (ICDE). Piscataway: IEEE, 2021: 2365-2368.

[10] TAFT R,ELSAYED N,SERAFINI M,et al. P-store:An elastic database system with predictive provisioning[C]// Proceedings of the 2018 International Conference on Management of Data. New York:ACM,2018:205-219.

[11] ELNAFFAR S,MARTIN P. The Psychic-Skeptic prediction framework for effective monitoring of DBMS workloads[J]. Data & Knowledge Engineering,2009,68(4):393-414.

[12] HIGGINSON A S,DEDIU M,ARSENE O,et al. Database workload capacity planning using time series analysis and machine learning[C]// Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. New York:ACM,2020:769-783.

[13] LORIDO B T,MIGUEL A J,LOZANO J A. A review of auto-scaling techniques for elastic applications in cloud environments [J]. Journal of Grid Computing,2014,12(4):559-592.

[14] PAVLO A,JONES E P C,ZDONIK S. On predictive modeling for optimizing transaction execution in parallel OLTP systems [J]. arXiv:1110.6647,2011.

[15] HOLZE M,RITTER N. Autonomic databases:Detection of workload shifts with n-gram-models[C]// East European Conference on Advances in Databases and Information Systems. Berlin:Springer,2008:127-142.

[16] DU N,YE X,WANG J. Towards workflow-driven database system workload modeling[C]// Proceedings of the Second International Workshop on Testing Database Systems. New York:ACM,2009:1-6.

[17] MA L,VAN AKEN D,HENFNY A,et al. Query-based workload forecasting for self-driving database management systems [C]// Proceedings of the 2018 International Conference on Management of Data. New York:ACM,2018:631-645.

[18] SHAHRIVARI H,PAPAPETROU O,FLETCHER G. Workload prediction for adaptive approximate query processing[C]// 2022 IEEE International Conference on Big Data(Big Data). Piscataway:IEEE,2022:217-222.

[19] DURAND G C,PINNECKE M,PIRIYEV R,et al. GridFormation:towards self-driven online data partitioning using reinforcement learning[C]// Proceedings of the First International Workshop on Exploiting Artificial Intelligence Techniques for Data Management. New York:ACM,2018:1-7.

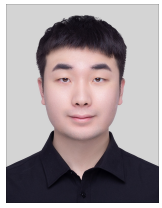
[20] HUANG X,CAO S,GAO Y,et al. LightPro:Lightweight probabilistic workload prediction framework for database-as-a-service[C]// 2022 IEEE International Conference on Web Services (ICWS). Piscataway:IEEE,2022:160-169.

[21] MOZAFARI B,CURINO C,JINDAL A,et al. Performance and resource modeling in highly-concurrent OLTP workloads[C]// Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. New York:ACM,2013:301-312.

[22] JAIN S,HOWE B,YAN J,et al. Query2vec:An evaluation of NLP techniques for generalized workload analytics[J]. arXiv:1801.05613,2018.

[23] HUANG X,CHENG Y,GAO X,et al. TEALED:A multi-step workload forecasting approach using time-sensitive EMD and auto LSTM Encoder-Decoder[C]// 27th International Conference Database Systems for Advanced Applications(DASFAA). Berlin:Springer,2022:706-713.

[24] XU M,SONG C,WU H,et al. esDNN:deep neural network based multivariate workload prediction in cloud computing environments [J]. ACM Transactions on Internet Technology (TOIT),2022,22(3):1-24.



BAI Wenchao, born in 1998, Ph.D, is a member of CCF (No. R5032G). His main research interests include explainable machine learning and intelligent big data processing.

(责任编辑:喻藜)