DOI:10.20033/j.1003-7241.(2025)02-0075-05.

# 基于DM2的异构数据一致性校验方法研究

朱 文[1]，方文崇[1]，周志峰[1]，李文朝[1]，攀腾飞[1]，吴海勇[2]

（1.中国南方电网有限责任公司，广东 广州 510770；
2.东方电子股份有限公司，山东 烟台 264010）

**摘 要：**随着物联网技术的快速发展，网络中存在大量的异构数据库，其异构性表现在多个方面。当数据库模式不同时，产生的数据类型也不尽相同，为保证数据的一致性，研究基于DM2的异构数据一致性校验方法。在异构系统中捕获异构数据包，通过良性和恶意线程运行模式，分析异构数据的特点。划分异构数据类型，处理并计算数据值。基于DM2技术构建输出图，设定一致性校验原则。以校验原则为基础，构建一致性验证矩阵，实现异构数据的一致性校验。实验结果表明：在不同数据源的并发过程中，所提方法能够有效召回错误数据记录，且召回正确率均在95%以上，具有较好的应用效果。

**关键词：**异构数据；一致性校验；校验方法；DM2技术
中图分类号：TP306 文献标识码：A 文章编号：1003-7241(2025)02-0075-05

# Research on Automatic Consistency Verification of Heterogeneous Data Based on DM2

**ZHU Wen[1], FANG Wen-chong[1], ZHOU Zhi-feng[1], LI Wen-chao[1], PAN Teng-fei[1], WU Hai-yong[2]**

( 1. China Southern Power Grid Company Limited, Guangzhou 510770 China;
2. Dongfang Electronics Co., Ltd., Yantai 264010 China )

**Abstract:** With the rapid development of Internet of Things technology, there are a large number of heterogeneous databases in the network, and their heterogeneity is manifested in many aspects. When the database models are different, the data types produced are also different. To ensure the consistency of data, an automatic consistency verification method for heterogeneous data based on DM2 is studied. It captures heterogeneous data packets in heterogeneous systems, and analyze the characteristics of heterogeneous data through benign and malicious thread running modes, it also divides heterogeneous data types, processes and calculate data values, builds the output map based on DM2 technology and sets the consistency verification principle. Based on the verification principle, a consistency verification matrix is constructed to verify the consistency of heterogeneous data. The experimental results show that in the concurrent process of different data sources, this method can effectively recall the wrong data records, and the recall accuracy is above 95%, which has good application effect.

**Keywords:** heterogeneous data; consistency check; calibration method; DM2 technology

## 0 引言

进入信息化时代以来，计算机技术的快速开发与广泛应用，尤其是数据库技术的开发，为大数据的管理提供了更加有效的手段。在此背景下，计算机数据库的发展越发成熟，并在多个行业中实现应用，特别是军事、银行、保险以及售票等领域[1-2]。从逻辑和系统应用便利性上来讲，分布式计算机信息系统符合现代信息社会的发展要求。在不同行业组织内，其依照自身需求选择适用于自身的数据库模式，通过不同的数据库模式描述数据特性，并通过多重访问机制对数据加以管理。但是由于数据库技术的日益完善，产生的数据也会更加复杂多样，原来集中式的数据库系统已无法适应新的使用要求。为保证数据之间的关联性，需要对数据的一致性进行校验。本文以DM2技术为基础，将其应用在网络节点中，设计异构数据的一致性校验方法，为保证数据之间的一致性提供理论支持。

## 1 异构数据捕获与特点分析

在异构系统中对异构数据进行捕获，分析其是否具有一致性，对其数据特点进行分析，以此设计一致性检验方法。对于异构系统而言，其得益于高性能和高效能的理念，在较低的能耗处理器中，对数据完成处理和存储[3]。异构数据存在安全和一致性问题，为保证异构系统的安全运行，对产生的数据包进行特征提取，以内存溢出为异

*基金项目：南方电网公司科技项目（ZDKJXM20200055）
收稿日期：2023-08-11

构数据捕获基础，分配异构系统中常出现的任务，第一部分为执行主机中的代码，第二部分是执行设备代码，执行过程中将数据在多个线程中流转。

一般情况下，一个任务最少使用32个线程，在足够的并发资源中，异构系统能够执行数据任务，但当资源不足时，则会产生串行执行，产生异构数据集合。将良性和恶意线程作为执行代码，良性线程表示为$P1$，恶意线程表示为$P2$，对其产生的异构数据进行分类，区分不同数据的特征。良性线程内，存在MALLOC分配函数和NEW函数，分别产生$buf$和$fp$数据，恶意线程内，除上述数据外，还存在恶意数据，则：

$$P1=\left[buf1 \quad fp1 \quad length1 \quad input1\right] \tag{1}$$

$$P2=\left[buf2 \quad fp2 \quad length2 \quad input2\right] \tag{2}$$

式中：当$P1$、$P2$共同执行时，在不同线程内异构数据能够产生多种特征元素。$P1$中的元素命名为$buf1$、$fp1$、$length1$、$input1$；$P2$中的元素命名为$buf2$、$fp2$、$length2$、$input2$。以最少的四组元素作为产生介质，并在不断增量下，形成异构数据并行集合，以此对异构数据进行分类，并计算异构数据元素值[4-5]。

## 2 划分异构数据区间

以分布图为基础对异构数据进行排序，按照中位值、上四分位值以及下四分位值，将其之间的距离作为异构数据的离散度，根据计算的数据离散度值，判断其一致性区间[6]。如下：

$$S_A=\begin{cases}S_{\frac{B+1}{2}};B\text{为奇数}\\ \dfrac{S_{\frac{B}{2}}+S_{\frac{B}{2}+1}}{2};B\text{为偶数}\end{cases} \tag{3}$$

式中，$S_A$为中位数，$S_B$为异构数据。在区间$[S_A,S_B]$的中位数，即为上四分位值$S_F$：

$$S_F=\begin{cases}S_{D-fix\left(\frac{B}{4}\right)};\mathrm{mod}\left(B,4\right)\le 1\\ \dfrac{S_{1+fix\left(\frac{B}{4}\right)}+S_{2+fix\left(\frac{B}{4}\right)}}{2};\mathrm{mod}\left(B,4\right)>1\end{cases} \tag{4}$$

下四分位值$S_G$是区间$[S_1,S_A]$的中位数，则：

$$S_G=\begin{cases}S_{1+fix\left(\frac{B}{4}\right)};\mathrm{mod}\left(B,4\right)\le 1\\ \dfrac{S_{1+fix\left(\frac{B}{4}\right)}+S_{2+fix\left(\frac{B}{4}\right)}}{2};\mathrm{mod}\left(B,4\right)>1\end{cases} \tag{5}$$

综合上式，计算离散度$G_S=S_F-S_G$，将其与中间值的偏差数据，认定为异构数据，获取异构数据区间：

$$\begin{cases}H_1=S_G-\dfrac{\chi}{2}*G_S\\ H_2=S_F+\dfrac{\chi}{2}*G_S\end{cases} \tag{6}$$

式中，$\chi$为偏差限值[7]。$[H_1,H_2]$为异构数据的区间，一般将该区间作为一致性数据的有效区间。

## 3 基于DM2设定一致性原则

利用DM2技术提取异构数据中的一致性特征，以其作为验证基础，设定校验的一致性原则，将异构数据认定为一个具有方向的带权图，通过顶点进行数据实体属性标识，解释有向带权下的实体属性关系。

通过异构数据的特点，将给定的校验任务抽象为Q-SCHEMA集合，且存在有一个日志集合$W$，将异构数据在一个图中完成输出，设定输出图为$E=[D,M]$[8]。则异构数据之间的关系问题，需要满足以下约束条件：

$$\forall m\left(y,u\right)\in M,y,u\in D \tag{7}$$

$$\forall m\left(y,u\right)\in M,m^{-1}\left(y,u\right)\notin M\ \left(4-2\right) \tag{8}$$

式中，将校验异构数据一致性任务，看作为确定$D$中元素$d$的问题；将校验异构数据关系，看作为确定$M$中元素的问题[9]。通过DM2组建元模型，指定各组元素的含义，见表1所示。

**表1 异构数据元素含义**

| 元素 | 含义 | 意义 |
|---|---|---|
| E | 表示异构数据包中的有向无环图 | 设定异构数据一致性关系解析原则 |
| D | 表示图中E顶点集合 | 确定异构数据顶点 |
| M | 表示图中E的有向边集合 | 确定异构数据有向边 |
| d | 表示图中E的顶点 | 标定数据顶点 |
| m | 表示图中E的有向边 | 标定数据有向边 |

根据表1中内容所示，对异构数据的有关验证规则和概念进行设定，将异构数据包$QY$看作数据集合$Y$以及给定任务模式$Q$，则：

$$\begin{aligned}&QY=\left\{Y,Q\right\}\\ &Y=\left\{y_1,y_2,...,y_n\right\}\\ &Q=\left\{q_1,q_2,...,q_n\right\}\end{aligned} \tag{9}$$

式中，异构数据包中所有的记录有序集合为$Y$，按照记录的主键进行排序，其中，$y_n$为第$n$条异构数据记录[10]。$Q$为数据包$Y$的校验任务模式，表示所有约束条件集合，其中，$q_n$为第$n$条约束条件，两者关系为：

$$\forall y\in Y,\forall q\in Q,y\text{满足}q \tag{10}$$

通过上式可知，每一组异构数据记录均满足校验任务中的所有约束[11]。对异构数据的属性关系建模，“键”即为异构数据的属性命名，“值”为异构数据的校验原则。

## 4 构建矩阵校验异构数据

为降低异构数据偏差的影响，将时间序列$K^{(0)}$设为非负序列，构建数据一致性判断矩阵，经处理后生成列表示为：

$$\begin{cases} K^{(0)}=\left\{K^{(0)}(1),K^{(0)}(2),...,K^{(0)}(J)\right\} \\ K^{(1)}=\left\{K^{(1)}(1),K^{(1)}(2),...,K^{(1)}(J)\right\} \end{cases} \tag{11}$$

$$\begin{cases} K^{(1)}(L)=\sum_{i=1}^{L}K^{(0)}(i),L=1,2,...,J \\ K^{(V)}(L)=\sum_{i=1}^{L}K^{(V-1)}(i),L=1,2,...,J \end{cases} \tag{12}$$

式中，累积次数用$V$表示[12]。生成列个数为$L$，其中$i,j\in L$。在构造累加后生成相邻序列$N^{(1)}$：

$$\begin{cases} N^{(1)}=\left\{N^{(1)}(1),N^{(1)}(2),...,N^{(1)}(J)\right\} \\ N^{(1)}(L)=0.5K^{(1)}(L)+0.5K^{(1)}(L-1) \\ N^{(1)}(1)=K^{(1)}(1) \end{cases} \tag{13}$$

将$\delta$和$\gamma$作为待解参数，构建矩阵模型$K^{(0)}(L)+\delta N^{(1)}(L)=\gamma$[13-14]。经过数学建模，将$\alpha=(\delta,\gamma)$作为参数列，展开得到$g=s\alpha$，其中：

$$g=\begin{bmatrix} K^{(0)}(2) \\ K^{(0)}(3) \\ \cdots \\ K^{(0)}(J) \end{bmatrix} \tag{14}$$

$$s=\begin{bmatrix} -\frac{1}{2}\left(K^{(1)}(1)+K^{(1)}(2)\right) & 1 \\ -\frac{1}{2}\left(K^{(1)}(3)+K^{(1)}(3)\right) & 1 \\ \cdots & \cdots \\ -\frac{1}{2}\left(K^{(1)}(J-1)+K^{(1)}(J)\right) & 1 \end{bmatrix}=\begin{bmatrix} -N^{(1)}(2) & 1 \\ -N^{(1)}(3) & 1 \\ \cdots & 1 \\ -N^{(1)}(J) & 1 \end{bmatrix} \tag{15}$$

通过最小二乘法，对$K^{(0)}(L)+\delta N^{(1)}(L)=\gamma$进行求解，得到异构数据的原始序列。如下：

$$\begin{aligned} K^{(0)}(J+1)&=K^{(1)}(J+1)-K^{(1)}(J) \\ &=\left(1-\delta^{\gamma}\right)\left[K^{(1)}-\frac{\gamma}{\delta}\right]\times\delta^{-\gamma J} \end{aligned} \tag{16}$$

在时间响应中对异构数据还原，以此实现对异构数据的一致性校验，至此本文在DM2技术下完成校验方法设计[15-16]。

## 5 实验测试分析

上文中通过DM2技术，设计了校验方法，为验证其能够在异构数据中完成一致性校验，进行实验论证。

### 5.1 准备测试数据

选取4组数据源作为测试样本，共包含120个数据属性。将每个数据源进行编号，从1-600记录数据及各组类型数据的关联属性，共存在有2 400条数据。数据源分别为M1、M2、M3、M4。其中，数据源M1中的各属性参考数据源M2属性，数据源M2属性参考数据源M3属性，数据源M3属性参考数据源M4属性。

本次测试同时在4组数据源中执行校验，以异构数据关联示意图为基础，共构造出240条错误记录。分别对错误记录产生的数据源组别进行标记，见表2所示。

根据表2中内容所示，每个数据源由于属性具有相关性，在数据异构中产生的错误量等同，仅在编号上存在差异。将表2中的数据传输至MATLAB测试平台，分别连接三组校验方法，对错误数据进行检测，验证错误数据的召回率和准确率。

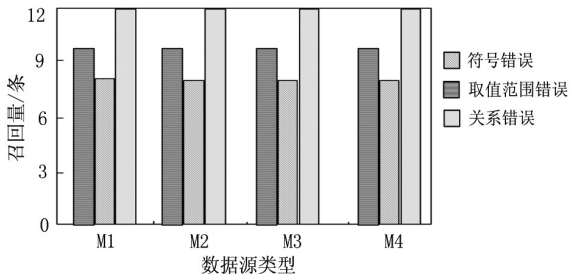
### 5.2 召回率校验

在异构数据一致性校验中，对设定的错误数据进行

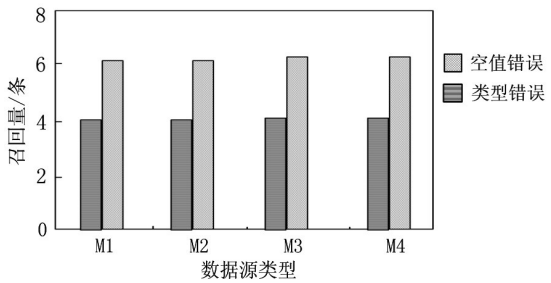
表2 异构数据测试样本分布情况(条)

| 数据源 | M1 | M2 | M3 | M4 |
|---|---|---|---|---|
| 总错误数量 | 40 | 40 | 40 | 40 |
| 取值范围错误数量 | 10 | 10 | 10 | 10 |
| 取值范围错误编号 | 1、4、8、12、25、46、58、112、94、168 | 2、6、7、17、42、68、92、158、302、468 | 3、5、9、18、38、52、76、158、245、540 | 10、14、21、32、45、86、98、142、255、365 |
| 类型错误数量 | 4 | 4 | 4 | 4 |
| 类型错误编号 | 1、51、68、104 | 4、78、78、124 | 6、61、62、134 | 8、46、82、164 |
| 空值错误数量 | 6 | 6 | 6 | 6 |
| 空值错误编号 | 1、21、62、85、102、245 | 2、11、42、75、162、245 | 4、16、52、65、112、255 | 8、18、32、56、122、356 |
| 符号错误数量 | 8 | 8 | 8 | 8 |
| 符号错误编号 | 2、11、32、45、112、212、285、375 | 6、31、42、55、122、222、245、355 | 9、51、52、65、142、205、242、385 | 10、61、72、75、162、235、252、365 |
| 关系错误数量 | 12 | 12 | 12 | 12 |
| 关系错误编号 | 2、3、5、10、22、44、55、64、162、128、295、385 | 3、4、6、14、25、26、28、112、84、178、265、355 | 4、7、9、12、28、46、59、94、114、168、285、375 | 5、8、10、11、35、36、48、84、102、128、275、365 |

召回，测试本文方法的召回率，按照上文中各类型的错误产生情况，采用本文方法对不同数据源的错误记录进行召回，结果见图1所示。




(a) 取值范围错误、符号错误与关系错误




(b) 空值错误与类型错误

图1 异构数据错误记录召回情况

根据图1中内容所示，在不同类型的异构数据错误记录中，本文方法的数据召回量均能与样本数据保持一致，说明本文方法能够实现对异构数据的有效检验，其校验效果较好。

### 5.3 对比校验方法正确率

上述实验对本文方法的数据校验效果进行了验证，为了进一步验证该方法的应用效果，分别选择矢量网络校验方法、非线性校验方法作为对照，与本文方法进行比较。对召回的数据进行正确性验证，对比不同错误数据类型的召回编号是否与样本编号一致。召回正确率$Z$的计算方式如下：

$$Z=\frac{find(X)}{full(C)}\times 100\% \tag{17}$$

式中，$find(X)$为召回的数据占样本总量的比例，$full(C)$为实际错误数据数量。针对上文中全部召回的数据，对照表中的编号情况，以总错误数量为校验基准，代入公式(17)中，结果见图2所示。

根据图2中结果所示，在召回的全部数据记录中，本文方法在不同错误类型中，正确率均可以达到95%以上，矢量网络校验方法和线性校验方法的正确率最大值分别为91.1%和95.2%。由此可知，本文方法能够对异构数据的错误记录完成准确校验，使其具有一致性和正确率，可以投入使用。

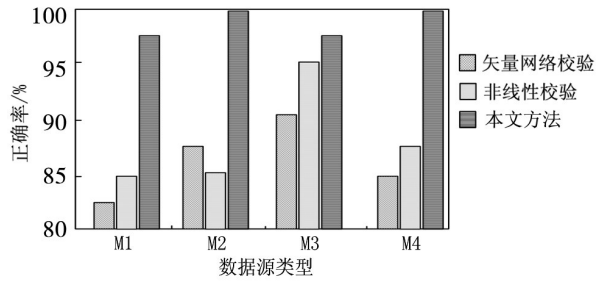



图2 召回数据的正确性对比结果

## 6 结束语

异构数据的同步问题，在近些年被反复提及，此次为保证异构数据的一致性，以捕获的数据包为分析基础，对其特征和类型进行分析，并通过DM2技术设计了校验方法。为保证新方法可以进行实际应用，在不同类型的异构数据中，实现了校验准确性和完整度的测试，并取得了一定的研究成果。但由于此次研究时间有限，在研究中选取的数据规模较小，校验流程过于简洁，在过于复杂的数据类型中，无法保证方法应用的有效性，后续研究中会按照更高的要求，对异构数据进行分析，为保证数据的一致性提供理论支持。

**参考文献**：

[1] 孙群，温伯威，陈欣.多源地理空间数据一致性处理研究进展[J].测绘学报，2022，51(7)：1561-1574.

[2] 廖彬，张陶，李敏，等.基于操作历史图的分布式Key-Value数据库一致性检测算法[J].计算机科学，2019，46(12)：213-219.

[3] 范君健，晁张虎，杨庆娜，等.基于Cadence CHI和IVD VIP的多核SoC系统数据一致性验证[J].电子技术应用，2020，46(8)：72-76.

[4] 徐达，关矗，周诚，等.装备维修性多源验前数据一致性检验方法研究[J].航天控制，2020，38(6)：61-66.

[5] 余安东，翟大海，苏瑾.数据副本一致性的算法研究与实现[J].计算机应用研究，2020，37(1)：63-65，75.

[6] 程子豪，李文强，漆小华，等.连接器接口数据快速校验与自动生成系统开发[J].机械设计与研究，2022，38(1)：4-10.

[7] 夏计强，崔鹏帅，李子勇，等.基于P4的SDN控制——数据平面流规则一致性校验[J].计算机应用研究，2022，39(8)：2479-2483，2489.

[8] 陈海燕.共同因子结构下非平稳面板数据检验的一致性研究[J].数理统计与管理，2019，38(3)：460-472.

[9] 韩圣亚，严莉，刘荫，等.基于XML的自动化异构系统数据一致性校验方法[J].电子设计工程，2021，29(13)：137-141.

[10] 杜岳峰，李晓光，宋宝燕.异构模式中关联数据的一致性规则发现方法[J].计算机研究与发展，2020，57(9)：1939-1948.

[11] 何玉林，金一，戴德鑫，等.混合属性数据集分布一致性度量的新方法[J].深圳大学学报(理工版)，2021，38(2)：170-179.

[12] 王洪申，王道俊.产品MBD数据集三维标注的自动校验与

实现[J].兰州理工大学学报,2021,47(2):48-53.

[13] 岳佳,张磊,鲁江伟.基于证据理论的遥测数据一致性融合判决方法[J].电子技术应用,2019,45(5):43-45.

[14] 王燕玲.共享模式下会计信息化的云数据完整性验证算法[J].现代电子技术,2019,42(5):87-89.

[15] 孙小虎,秦浩,张亚平,等.基于关联规则的电网大数据质量校验方法研究[J].电子设计工程,2020,28(21):145-148,153.

[16] 林培榕,曾海亮,王晨曦,等.小样本类不平衡数据的一致性分析流特征选择[J].小型微型计算机系统,2021,42(11):2252-2258.

作者简介:朱文(1987-),男,硕士,工程师,研究方向:调度自动化系统。

(上接第74页)

过程中的连续丢包率,$\eta_l$表示服务信息质量系数;$\varepsilon_{loss}$表示传输过程中的丢包率。

依据式(9)获取本文方法在不同数量的智能终端数量下,随着数据交换执行次数的不断增加,$\vartheta$指标结果如图9所示。实验要求时滞效应结果低于3.5 s。对图9测试结果实行分析后得出:在智能终端数字孪生体生成过程中,在不同的智能终端数量下,数据交换执行次数的不断增加,本文方法的$\vartheta$指标结果也存在一定程度的变化,智能终端数量越多,$\vartheta$指标结果也会发生小幅度的上升,但是结果均低于3.5 s。因此,本文方法在进行智能终端数据交换过程中,具有良好的数据交换性能。

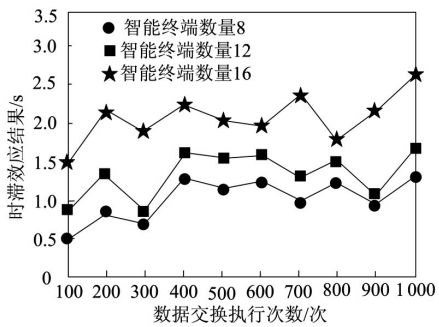



图9 数据交换的时滞效应结果

表1 不同交换情况下数据交换的安全系数结果

| 交换数据量/条 | 情况1 | 情况2 | 情况3 |
|---|---|---|---|
| 50 | 0.977 | 0.962 | 0.963 |
| 100 | 0.968 | 0.974 | 0.959 |
| 150 | 0.959 | 0.983 | 0.974 |
| 200 | 0.966 | 0.964 | 0.983 |
| 250 | 0.974 | 0.958 | 0.986 |
| 300 | 0.982 | 0.968 | 0.966 |
| 350 | 0.991 | 0.972 | 0.975 |
| 400 | 0.979 | 0.982 | 0.955 |
| 450 | 0.994 | 0.961 | 0.967 |
| 500 | 0.982 | 0.995 | 0.988 |
| 550 | 0.973 | 0.989 | 0.991 |
| 600 | 0.969 | 0.967 | 0.987 |

为验证智能终端数据交换安全性,采用安全系数作为衡量标准,获取本文方法在3种数据交换情况下(分别为情况1:物理对象和数字对象之间、情况2:物理对象和物理对象之间、情况3:数字对象和数字对象之间)的安全系数结果,如表1所示。从表1可知:在3种数据交换情况下,随着交换数据量的逐渐增加,安全系数结果均在0.95以上,交换数据量的逐渐增加安全次数结果没有发生明显下降,其中,安全系数的最高值达到0.996。因此,本文方法能够在不同的数据交换应用情况下,实现数据的安全交换。

## 3 结束语

为了实现智能终端的数据的最佳交换效果,提出数字孪生技术下物联网智能终端数据交换方法。测试结果显示:方法的智能终端数据交换时滞效应结果均满足应用需求,在不同的数据交换情况下,数据交换安全性高,能够较好地完成智能物理模型和数字模型之间的数据交换,以此生成其数字孪生体三维场景,应用效果良好。

**参考文献**:

[1] 王慧敏,熊玲,督静雯,等.面向物联网环境的数据安全传输方案[J].计算机应用研究,2020,37(S1):304-305,313.

[2] 杨业平,林德威,黄芳芳,等.基于区块链的物联网安全数据共享系统[J].福州大学学报(自然科学版),2021,49(6):739-746.

[3] 胡亨汶,孟祥印,李丹,等.基于RESTful Web Services的云边数据交换设计与实现[J].现代制造工程,2022(8):25-32.

[4] 李超,韩翔,刘钊,等.基于可信计算的跨网数据安全交换技术[J].计算机工程与设计,2021,42(10):2762-2769.

[5] 饶小康,马瑞,张力,等.基于GIS+BIM+IoT数字孪生的堤防工程安全管理平台研究[J].中国农村水利水电,2022(1):1-7.

[6] 于洋,石振武,王奇,等.寒区高速公路路线三维设计优化研究[J].森林工程,2023,39(6):188-195.

[7] 王庭松,惠小东,曾乔迪,等.基于改进ICP算法的变电站设备三维识别方法研究[J].电测与仪表,2024,61(5):65-70.

[8] 陈楚炼,袁金海,郭智生,等.点云数据处理方法的地铁变电站可视化建模研究[J].自动化技术与应用,2025,44(1):127-131,162.

作者简介:郭敬东(1968-),男,硕士,高级工程师,研究方向:电力信息安全。