文章编号:1671-4598(2024)03-0286-07 DOI:10.16526/j.cnki.11-4762/tp.2024.03.042 中图分类号:TP206.3 文献标识码:A

基于云数据中心的多源异构数据治理技术研究

孙 瑜
(中国人民解放军92941部队45分队,辽宁 葫芦岛 125001)

摘要:目前常规的多源异构数据治理方法主要通过对数据属性进行判断,从而实现分区域数据清洗,由于缺乏对非线性数据的分析,导致治理性能不佳;对此,提出基于云数据中心的多源异构数据治理技术;采用关系型数据库中的ETL功能对数据进行清洗,对数据转换模式以及数据清洗规则进行定义;引入互信息系数对数据相关程度进行判定,并进行非线性数据相关性分析;以云数据中心作为载体,对多源异构数据治理体系进行构建;在实验中,对提出的数据治理技术进行了治理性能的检验;最终的实验结果表明,提出的数据治理技术具备较高的查准率,对云数据中心多源异构数据具备较为理想的数据治理效果。

关键词:云数据中心;多源异构数据;数据治理;数据清洗

Research on Multi-source Heterogeneous Data Governance Technology Based on Cloud Data Center

SUN Yu
(Sub Unit 45 Unit 92941 of the PLA, Huludao 125001, China)

Abstract: Currently, conventional multi-source heterogeneous data governance methods are mainly used to judge data attributes to achieve sub-regional data cleaning, which leads to poor governance performance due to the lack of non-linear data analysis. For this reason, a multi-source heterogeneous data governance technique based on cloud data center is proposed. The ETL function of relational database is adopted to clean the data, which defines the data transformation mode and data cleaning rules. The mutual information coefficient is introduced to determine the degree of data relevance, and analyze the data relevance analysis. The cloud data center is used as a carrier to construct the multi-source heterogeneous data governance system. In the experiments, the governance performance of the proposed data governance technique is examined. The final test results show that the proposed data governance technique has a high checking accuracy rate and more ideal data governance effect.

Keywords: cloud data center; multi-source heterogeneous data; data governance; data cleansing

0 引言

在军队数据信息化管理过程中,受到存储方式以及管理系统差异等多方面的影响,数据的来源以及格式通常会存在较大的差异。不同的数据库所存储的数据在量纲和格式上均有所不同,然而在军队信息化管理中,通常需要对数据进行集成化处理,在该过程中,涉及将不同来源的数据进行整合,通过特定的方式对数据进行归一化处理,从而实现信息化数据分析。因此,为提高对多源异构数据的管理效率以及智能化分析水平,需要采取必要的手段对数据进行治理。所谓数据治理指的是通过统一的方式,对多源异构数据中的噪声部分、冗余部分进行剔除,同时对异常数据进行检测,并通过设定检测阈值,实现对异常数据的过滤处理。除此之外,还需要对数据中的不准确部分进行纠正,对时间尺度进行归一化调整,从而保证数据的完整性以及连续性。

对此,文献[1]结合学习事件,针对远程教育系统中产生的多源异构数据进行语义融合处理,并对数据治理技术的实践可能性进行了探讨与分析。文献[2]结合大数据技术,针对高炉数据中的冗余以及过曝问题进行了合理分析,在此基础上提出了具体的数据治理方案。文献[3]结合半监督学习算法,针对多源异构数据的数据特点,对数据治理流程进行了优化。文献[4]以区域教育网格多维数据作为研究对象,通过结合大数据平台,构建出了系统化的数据治理体系,为区域教育数据管理提供了高保真数据。文献[5]针对使用一种异构数据进行深度学习故障诊断不可避免地会导致诊断准确性差的问题,通过设计具有交替优化机制的融合网络,提出了一种深度公共特征提取方法。从两种异构数据中独立提取的粗糙特征用于以另一种优化方式训练所设计的融合网络。建立了替代优化所需的新的损失函数;因此,所有网络都可以进行全局调谐。通过融合网络的交替优化训练过程,可以很好地提取多源异构数

收稿日期:2023-08-20; **修回日期**:2023-09-13。
作者简介:孙 瑜(1980-),女,工学硕士,工程师。
引用格式:孙 瑜.基于云数据中心的多源异构数据治理技术研究[J].计算机测量与控制,2024,32(3):286-292.

据的深层共性特征，提高了深度学习故障诊断方法的准确性。文献［6］从经济、社会和生态 3 个维度建立了可持续发展的标准体系，并给出了 10 个次级标准。然后为了充分表达专家的犹豫和模糊性，使用语言模糊犹豫集来描述主观评价信息，并开发了云模型来处理 LFHS 的随机性和模糊性。文献［7］提出了一个基于资源描述框架的模糊时空 RDF 图模型，该模型由万维网联盟（W3C）提出，以三元组（主语、谓语、宾语）表示数据。其次，对多源异构模糊时空数据的相关异构问题进行分析和分类，并利用模糊时空 RDF 图模型定义相应的规则来解决这些异构问题。此外，还根据 RDF 三元组的特点，分析了 RDF 三元组中多源异构模糊时空数据集成的异构问题，并给出了 FRDFG 的集成方法。上述方法均可以在一定程度上实现多源异构数据的有效治理，但是在治理效果上还有待优化。多源异构数据由于自身的数据结构特点，数据间的信息冗余程度较高，因此常规的治理方法无法去除掉过多的冗余部分。同时，由于多源异构数据的规模较大，导致数据治理效率较低。

为了对数据治理效果进行优化与调整，本文提出了一种基于云数据中心的多源异构数据治理技术。使用 ETL 功能对数据进行清洗和转换，在数据处理过程中引入互信息系数进行非线性数据相关性分析，并构建多源异构数据治理体系。实验结果验证了该技术在数据清洗方面具有较高的查准率，显示对于云数据中心多源异构数据的理想治理效果。该研究提供了一种新的思路和方法，可以为数据治理领域的进一步研究和实践提供有益的参考。

1　多源异构数据治理架构及原理

云数据中心具备强大的大规模数据处理能力和弹性扩展性，能够轻松应对多源异构数据的处理需求。此外，云数据中心提供了丰富的数据集成和协同处理工具，通过统一的平台实现数据的整合和清洗，更有效地发现数据之间的相关性。另外，云数据中心还拥有高级的安全机制和数据备份策略，确保数据的安全性和完整性。基于云数据中心进行多源异构数据治理还能降低运维成本，用户无需投入昂贵的设备和人力资源，由云服务提供商负责维护和管理。因此，结合以上优势，本文基于云数据中心，对多源异构数据的治理架构进行设计，由此构建出的数据中心架构如图 1 所示。

通过上述数据治理中心的架构图可以看出，本文所构建出的数据治理体系主要包括 3 个部分，分别为私有云、边缘侧以及设备端[8-9]。其中，私有云部分主要以云数据中心作为载体，集成了多源异构数据的关系型数据库，对实时数据、历史数据以及非结构化数据进行存储，通过云数据中心平台，可以对多源异构数据进行调取。而边缘侧主要负责提供数据治理服务，具体包括数据采集、数据清洗、数据相关性分析以及数据分类。通过结合机理模型以及数据库的智能诊断功能，对数据格式进行转换，从而实现数据量纲的统一。同时，在边缘侧结构中，还需要对多源异构数据中的异常部分以及冗余部分进行智能化处理，以此优化数据治理效果。设备端的主要功能在于物联网的动态感知，通过配备多种传感器，对当前服务的系统对象进行感知，从而获取实时数据，并将其传送到私有云，实现数据循环治理[10]。

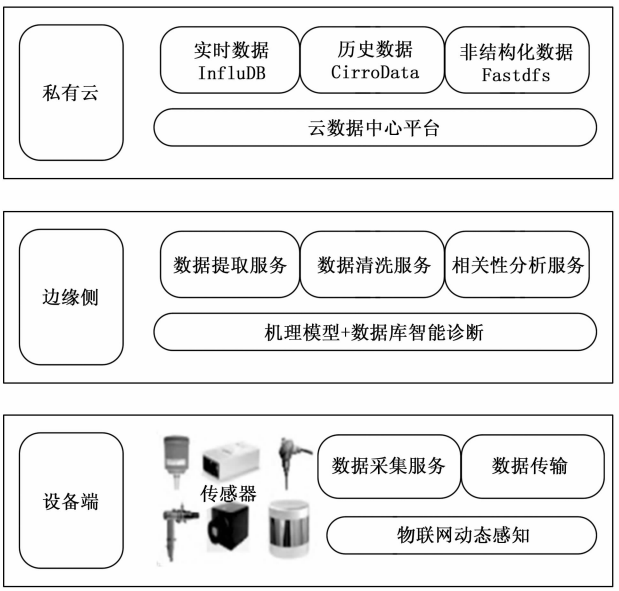

图 1　基于云数据中心的多源异构数据治理体系架构图

2　多源异构数据治理架构设计

2.1　多源异构数据清洗

多源异构数据在实际应用中面临数据质量不高、格式不统一和缺失异常值等问题。由于数据来源的不同和数据传输过程中的扰动，数据可能包含错误、重复、无效值和缺失值等不规范数据。为了解决这些问题，为确保多源异构数据的准确性、完整性和一致性，在多源异构数据在入库存储之前，需要对其中的不规范数据进行清洗处理。数据清洗能够发现和修复数据中的错误和缺失。通过识别和纠正异常值、重复值和无效值等错误，数据清洗可以保证数据的可靠性和准确性。其次，数据清洗可以统一数据的格式和结构，在多源异构数据中消除不一致性，使得数据能够进行有效的整合和对比分析。此外，数据清洗还可以处理缺失值和异常值，填充缺失值、插值等方法能够帮助补全数据并保持数据的完整性。

对此，本文通过构建数据清洗流程，通过定义数据转换模式以及匹配规则，采用中间数据库对多源异构数据进行转换清洗，最后存入到目标数据库中，从而实现数据治理。一般来说，多源异构数据[11-13]的来源主要有两个，分别为不同规模的数据源以及外部文件。其中，数据源中的数据抽取可以通过 SQL 语句实现，而外部文件无法直接通过指令对数据进行调取，因此需要先抽取外部文件中的数据，然后将其传送到中间数据库中进行统一清洗与转换。对此，本文采用关系型数据库中的 ETL 功能对数据进行清

洗，ETL 数据清洗功能如图 2 所示。

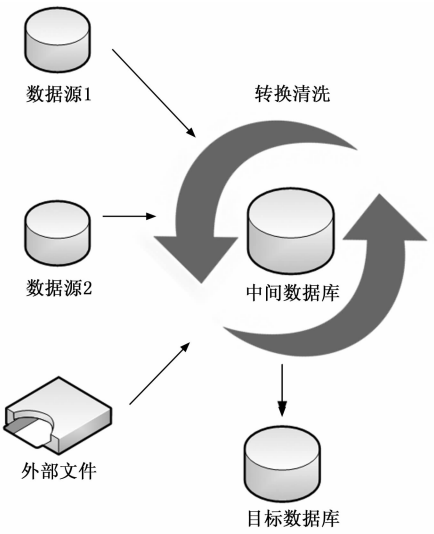

图 2 ETL 数据清洗功能

多源异构数据的清洗流程主要包括 4 个步骤，分别为数据源分析、数据转换模式定义、工作流验证与评估以及执行工作流[14]。

步骤 1：数据源分析。本文首先将手动检查以及程序自动检查两种方式进行结合，对多源异构数据的来源进行筛查，从而明确数据的描述信息以及分布范围等，明确多源异构数据的当前质量水准。手动检查是指专业的数据管理人员或领域专家对多源异构数据进行目视检查和分析。通过手动检查，可以获取数据的直观认识，并发现可能存在的问题或潜在风险。程序自动检查是利用算法、规则或模型来自动检测和评估多源异构数据的质量。通过程序自动检查，可以对大量数据进行快速检测和分析，发现数据中的异常或错误，同时还可以提供数据的描述统计信息、数据分布范围等。综合应用手动检查和程序自动检查的方法，在多源异构数据来源的筛查过程中，可以有效地明确多源异构数据的描述信息、判断数据质量水平和数据分布范围，从而为后续的数据分析、挖掘和决策提供可靠和准确的基础。同时，该方法还能够节约时间和人力成本，提高数据的管理效率和质量。

步骤 2：数据转换模式定义。对元数据的存储格式进行规定，具体格式如表 1 所示。

表 1 元数据的存储格式

列名	是否为主键	数据类型	字段描述
创建时间	否	DATETIME	数据创建时间
安全性	否	INT	数据安全访问级别
描述字段	否	TEXT(128)	数据源描述字段
名称	否	TEXT(16)	数据源名称
标识符	是	TEXT(64)	数据源唯一标识符

数据源唯一标识符在完成数据源分析后，本文对数据转换模式以及数据清洗规则进行定义。由于不同数据源之间的语义规则有所不同，因此在对数据转换模式定义时，可以通过构建等价实体关系表，对不同数据源之间的数据格式进行转换与统一处理。对此，本文选择构建数据库层面索引表，将不同数据源的数据按照数据库进行分类，通过索引的形式进行管理，从而实现数据模式转换。然后对数据清洗[15-17]规则进行定义，针对元数据的数据特点，本文为了方便对多源异构数据进行高效调度，数据清洗规则进行表 2 的定义。

表 2 多源异构数据清洗规则

列名	是否为主键	数据类型	字段描述
错误类别	是	TEXT(16)	格式异常的数据
记录集名	是	TEXT(16)	与记录集合不符的数据
判断条件	否	TEXT(128)	判断错误类型的条件
字段名	否	TEXT(16)	字段名称错误的数据
函数名	是	TEXT(64)	数据清洗算法
权重	否	TEXT(16)	数据属性对应的权重
阈值	否	TEXT(16)	超过阈值的异常数据
缺失值过滤	否	TEXT(128)	空值数据
清洗策略	是	TEXT(64)	手工处理或算法清洗

步骤 3：工作流验证与评估。该步骤中，本文关注的是针对多源数据的异常值判定。异常值是指与样本的大部分观测值显著不同的观测值，可能是由于测量误差、输入错误、系统故障或其他未知因素引起的。箱型图是一种常用的可视化工具，用于展示数据的分布情况和异常值的存在。其主要由一个矩形箱体和两条延伸的触须构成。因此，针对多源数据的异常值判定[18]，本文结合箱型图的方式，通过对数据的边缘鲁棒性对异常数据进行识别。通过边缘鲁棒性分析，能够利用箱型图的特性来识别异常数据。异常数据可能表现为远离箱体的离群点，因为箱体表示了数据的中间 50%的范围。所以，如果数据超出了触须长度的 1.5 倍范围，那么该数据点有可能被视为异常数据。

这一过程中，假设数据上边缘以及下边缘分别为 A 和 B，则异常数据的判定表达式如下所示。

$$A = Q_3 + 1.5(Q_3 - Q_1) \tag{1}$$

$$B = Q_1 - 1.5(Q_3 - Q_1) \tag{2}$$

其中，Q_1 代表多源异构数据的下四分位数，Q_3 代表多源异构数据的上四分位数。箱型图的具体示意图如图 3 所示。

步骤 4：执行工作流。结合上述异常数据的判定方法，对异常值进行剔除，从而完成数据清洗。具体步骤如下：

1）首先，准备待清洗的多源数据。

2）异常数据判定：结合箱型图的方式，对数据的边缘鲁棒性进行分析，识别异常数据。箱型图根据数据的分布情况绘制出最小值、下四分位数、中位数、上四分位数和最大值，利用这些统计量来判断是否存在异常值。将超过上下四分位数 1.5 倍或 3 倍的差距的数据点视为异常值。

3）异常值剔除：根据异常数据的判定结果，将被标记为异常的数据进行剔除。剔除异常值的方式是删除包含异常值的整行数据。

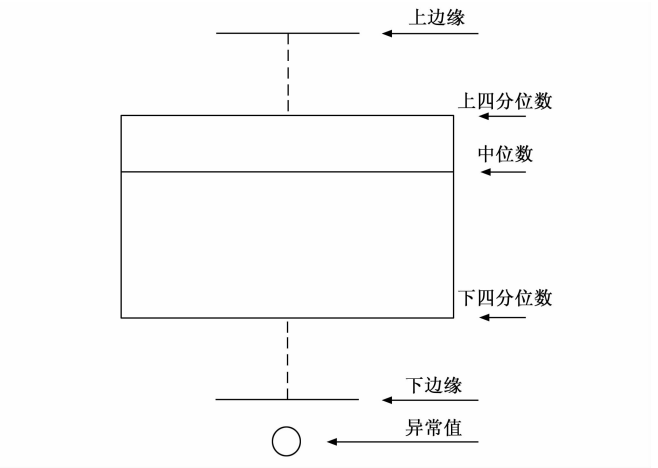

图 3　箱型图示意图

4）数据清洗结果验证：经过异常数据剔除后，使用描述性统计分析方法验证数据清洗的效果，查看异常值是否被成功剔除，并评估数据的质量和一致性。

2.2　数据相关性分析

在完成数据清洗后，尽管数据的质量得到了提升，仍然可能存在冗余的信息。冗余数据可能是指多个数据字段之间存在高度重复或高度相关的情况，这种情况下，一些数据字段可能具有相似的信息，对于分析和决策过程中产生了冗余。为了降低冗余情况并提高数据治理效率，本文采用了皮尔逊相关系数法[19]进行多源异构数据的相关性分析。皮尔逊相关系数是一种衡量两个变量之间线性相关关系强度和方向的统计量。通过计算数据字段之间的皮尔逊相关系数，可以判断它们之间的相关程度，并识别出那些存在较高相关性的数据字段，从而降低数据信息的冗余情况，提高数据治理效率。

假设在多源异构数据库中，存在两个随机数据分别为 X 和 Y，则这两个数据之间的相关性为 $\rho(X,Y)$。为保证分析效果，本文采用线性优化的方式对数据相关性进行计算，具体计算公式如下所示[20]。

$$\rho(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \tag{3}$$

其中：$Var(X)$ 和 $Var(Y)$ 分别代表随机数据 X 和 Y 对应的方差大小，$Cov(X,Y)$ 代表两个随机数据之间的协方差。

上述公式可以针对多源异构数据中呈现线性关系的数据进行相关性分析，例如同一属性下的衍生数据等[21]。但是针对相关性并不明显的随机数据，采用上述公式无法深度挖掘出随机数据之间的非线性关系。因此为了保证相关性分析的效果更为全面，本文引入互信息系数这一参数，对随机数据之间的相关程度进行判定。

互信息系数是一种衡量两个变量之间关联程度的统计量，通过引入互信息系数进行数据相关性分析，可以更准确地识别和量化数据之间的关联关系。引入互信息系数有助于提高数据治理的准确性和效率，同时，互信息系数的值可以根据随机数据的联合概率密度计算而得，具体计算公式如下所示。

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}\log\frac{p(x,y)}{p(x)p(y)} \tag{4}$$

其中，$I(X;Y)$ 代表互信息系数[22]，$p(x,y)$ 代表随机变量 X 与 Y 的联合概率密度函数，$p(x)$ 和 $p(y)$ 代表两个随机变量 X 与 Y 对应的边缘分布密度函数。通过上述公式对互信息系数进行求解，然后结合相关度阈值，对随机数据之间的相关程度进行判定，具体判定规则如图 4 所示。

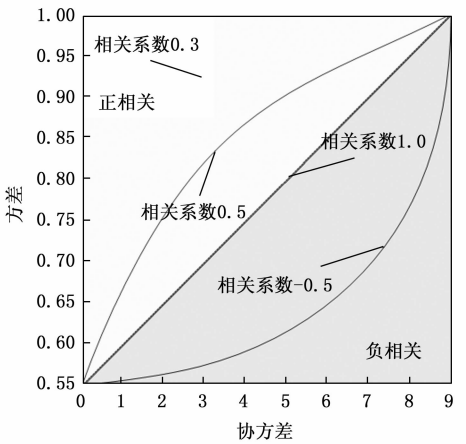

图 4　数据相关性判断规则

根据上述数据相关性判断规则可以看出，相关系数越接近与 1，代表两个随机数据之间的相关性越强。在对多源异构数据相关性进行分析时，首先结合互信息系数，对数据相关性进行判断，然后结合皮尔逊相关系数法[23-24]，计算相关系数，从而实现数据相关性的有效分析。

结合上述提出的数据清洗以及数据相关性分析等内容，即可构建出数据治理机制，通过结合私有云平台，对关系型数据库在内的多种数据进行集成存储，并对数据进行清洗以及相关性分析，从而实现数据治理。至此，基于云数据中心的多源异构数据治理技术设计完成。

3　实验结果与分析

为了证明本文提出的基于云数据中心的多源异构数据治理技术的实际治理效果优于常规的多源异构数据治理技术，在理论部分的设计完成后，构建实验环节，对本文方法的治理效果进行检验。

3.1　实验说明

为验证本文提出的基于云数据中心的多源异构数据治理技术在实际治理性能方面的有效性，本次实验选取了两种常规的多源异构数据治理技术作为对比对象，分别为文献［1］基于学习事件的远程教育多源异构数据语义融合与实践研究和文献［2］基于半监督学习的多源异构数据治理方法。通过构建实验平台，采用 3 种治理方法对同一组多源异构数据进行治理，对比不同方法的实际治理效果。

3.2　实验准备

1）硬件环境准备：

本次实验采用 Hadoop 集群框架对实验平台进行搭建，

通过部署主节点以及从属节点，分别模拟不同数据源之间的数据传输操作。Hadoop 框架版本为 2.5.0，采用 Winows 系统对框架进行开发。

2）数据准备：

为保证实验的可靠性，本次实验的测试数据来自某系统的用户负荷数据。测试系统参数，如表 3 所示。

表 3 系统参数

节点	1	2	3	4	5	6	7	8
额定容量/kVA	156	189	200	169	258	369	147	100
空载损耗/kVA	0.23	0.25	0.48	0.29	0.54	0.258	0.364	0.28
短路损耗/kVA	1.5	2.2	3.16	2.2	3.62	1.5	2.35	2.9

通过对该系统的历史运行数据进行调取，构建原始数据集。

局部异常因子是一种用于检测和识别异常数据的统计方法，它基于数据的局部密度和离群程度来判断数据是否异常。在云数据中心等复杂环境中，数据源可能具有不同的特征和分布，可能会涉及多个数据类型和数据来源。而且，由于数据传输、存储、处理等过程中可能引入噪声、错误或数据异常，这些异常数据对于数据治理和决策过程可能会产生负面影响。因此，在基于云数据中心的多源异构数据治理测试中，为准确测试出不同方法对于异常数据的识别效果，本次实验通过引入局部异常因子，对异常数据进行仿真，样本数据的异常因子经过打乱后会更具备真实性，其具体分布如图 5 和图 6 所示。

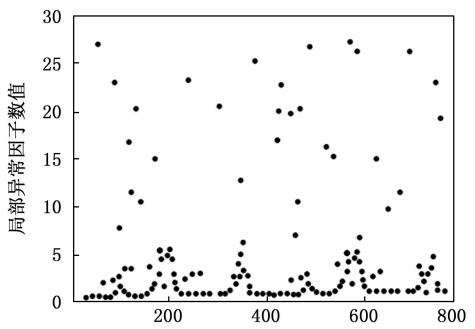

图 5 局部异常因子原始分布图

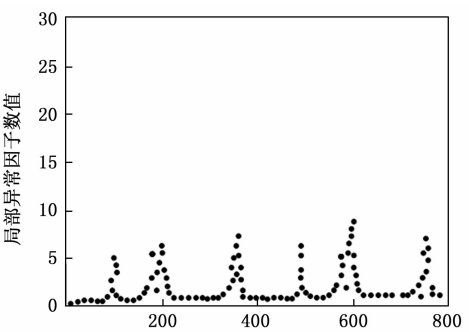

图 6 局部异常因子排序分布图

本次实验共设置了 6 个训练集以及 6 组测试集，每个实验数据集中包含的样本数据均有所不同，具体数据集分配如表 4 所示。

表 4 数据集分配情况

数据集	数据数量/条	文件大小
测试集 A	15 000	2.5 GB
测试集 B	18 000	2.75 GB
测试集 C	20 000	3.42 GB
测试集 D	25 000	3.75 GB
测试集 E	28 000	3.84 GB
训练集 A	10 000	2.01 GB
训练集 B	25 000	3.78 GB
训练集 C	500 000	5.48 GB
训练集 D	38 000	2.89 GB
训练集 E	45 000	3.42 GB

表 4 中，测试集 A、测试集 B、测试集 C、测试集 D、测试集 E 分别对应训练集 A、训练集 B、训练集 C、训练集 D、训练集 E。同时，为了满足多源异构数据的多源性和异构性，综合利用手动检查和程序自动检查方法对表 4 中的数据集 A、B、C、D、E 数据类型进行筛选，并将其总结为表 5。

表 5 被分配数据集概况

测试集	数据类型	数据类别	文件大小
测试集 A	时间序列数据	用户请求次数、请求处理时间、服务器负载情况	4.51 GB
测试集 B	结构化数据	不同模块的请求次数、并发用户数、平均响应时间	6.53 GB
测试集 C	日志数据	登录日志、错误日志、交互日志	8.9 GB
测试集 D	网络流量数据	网络带宽、连接数、数据包大小	6.64 GB
测试集 E	自然语言文本数据	用户留言、投诉、建议	7.26 GB

为对原始数据集中的离群点进行检测，对负荷数据的离群点进行聚类分析，分 3 个维度对数据进行展示，从而构建三维散点图，具体如图 7 所示。

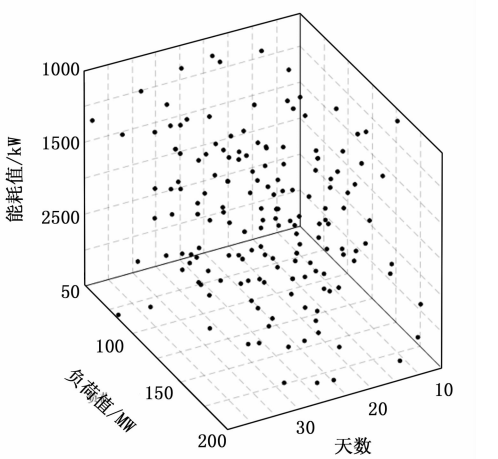

图 7 负荷数据三维散点图

3.3　实验步骤

步骤 1：数据准备：按照 3.2 部分准备多源异构数据，并确保数据经过数据清洗和预处理，以保证数据的质量和一致性。

步骤 2：互信息系数分析：使用互信息系数来判断数据之间的相关性。互信息是一种度量两个随机变量之间的关联程度的统计量。通过计算互信息系数，可以评估两个数据变量之间的非线性关联程度。较高的互信息系数表明两个变量之间具有较强的相关性。

步骤 3：皮尔逊相关系数分析：在确定了可能存在相关性的数据变量后，使用皮尔逊相关系数计算它们之间的相关性。皮尔逊相关系数衡量的是两个变量之间的线性关联程度。皮尔逊相关系数的取值范围从－1 到 1，其中－1 表示完全负相关，1 表示完全正相关，0 表示无相关性。

步骤 4：分别利用本文设计方法、常规方法 1 与常规方法 2 对用户负荷数据进行治理，同时，为提高实验结果的对比性，本次实验设定了两种不同的测试条件，分别为局部异常因子为 4 和 8 的异常数据集。通过采用 3 种数据治理方法对同一组数据集进行处理，对比不同异常数据密度下，3 种方法的实际处理效果。

3.4　数据治理性能对比结果

本次对比实验选取的对比指标为不同方法的数据治理性能，具体衡量指标为数据查全率，计算公式如下所示。

$$R = \frac{N_{TP}}{N_{FP} + N_{FN}} \times 100\% \tag{5}$$

其中，N_{TP} 代表准确检测出异常数据的数量，N_{FN} 代表错误认定为异常数据的数量，N_{FP} 代表未被检测出异常数据的数量。具体实验结果如图 8 和图 9 所示。

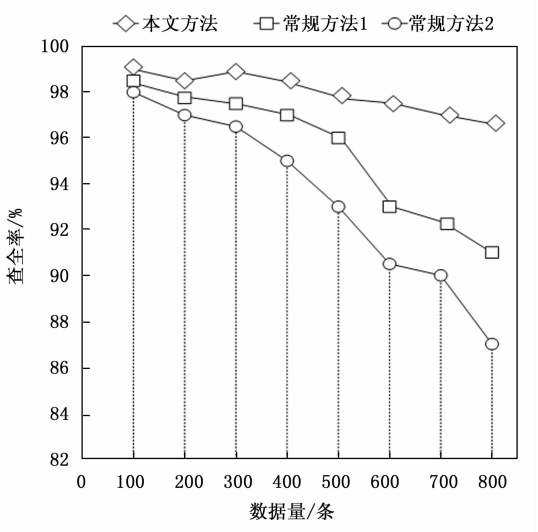

图 8　局部异常因子为 4 的查准率对比结果

通过上述实验结果可以看出，局部异常因子的值会在一定程度上影响方法的治理效果。通过数值上的对比可以看出，本文提出的基于云数据中心的多源异构数据治理技术的查准率明显高于两种常规的治理方法，具备更为理想的治理效果。

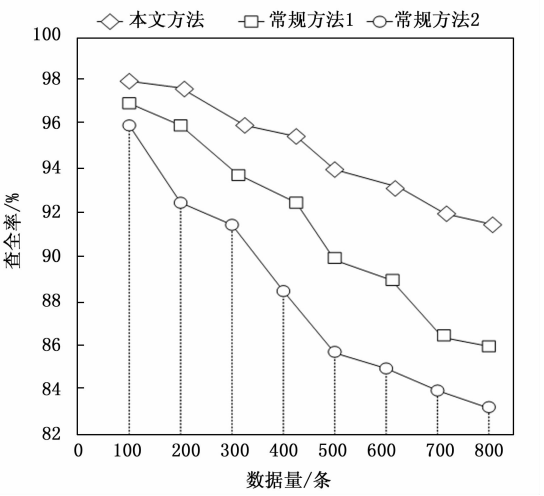

图 9　局部异常因子为 8 的查准率对比结果

3.5　数据治理准确率对比

实验选取治理准确率作为实验的评价指标，表示为公式（6）：

$$p = \frac{p1}{p1 + p2} \times 100\% \tag{6}$$

式中，p 为治理准确率；$p1$ 为多源异构数据被检测出异常的次数；$p2$ 为多源异构数据被检测出正常的次数。

基于以上指标，对 3.2 部分设置的多源异构数据进行进一步细分，得到准确率实验数据集为表 6。

表 6　准确率实验数据集

时间戳	用户 ID	设备 ID	用户负荷
00:00:00	用户 1	Device1	2.5 kW
00:05:00	用户 1	Device1	3.2 kW
00:10:00	用户 1	Device1	2.8 kW
00:00:00	用户 2	Device2	1.7 kW
00:05:00	用户 2	Device2	2.1 kW
00:10:00	用户 2	Device2	2.5 kW
00:00:00	用户 3	Device3	3.8 kW
00:05:00	用户 3	Device3	4.5 kW
00:10:00	用户 3	Device3	4.2 kW

表 6 所示的数据集中包含了不同用户和设备的用户负荷数据，时间戳表示每个测量数据的时间点，用户 ID 标识不同的用户，设备 ID 标识不同的设备，用户负荷表示对应时间点的用户负荷值（以千瓦为单位）。基于以上数据集，分别利用本文设计方法、常规方法 1 与常规方法 2 对多源异构负荷数据进行治理准确率实验。治理结果如图 10 所示。

从图 10 中可以看出，本文设计的方法对比其它两种方法而言，对多源异构数据的治理准确率较高，能够准确治理多源异构数据。

4　结束语

本文针对常规的多源异构数据治理方法在治理性能方面较差的问题进行了研究，提出了一种结合云数据中心的

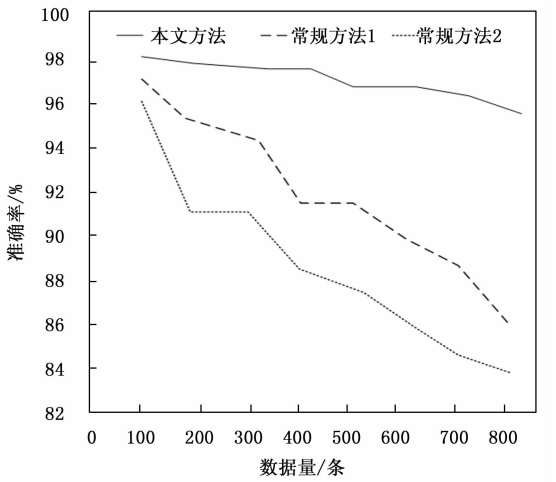

图 10 治理准确率实验结果

新型数据治理技术。在关系型数据库中使用 ETL 功能进行数据清洗的基础上，引入互信息系数进行数据相关性分析，并以云数据中心为基础，构建多源异构数据治理体系，从而实现数据治理的目标。该研究的创新点在于结合了云数据中心的优势。云数据中心具有高效的数据存储和处理能力，可以支持大规模的数据处理任务。通过将数据治理与云数据中心相结合，可以充分利用云平台的资源和计算能力，提高数据治理的整体性能和效果。

在今后的研究工作中，还存在一些需要进一步探索和优化的方面。首先，将对数据治理技术进行约简处理，以减少数据处理的复杂性和冗余性。约简处理可以针对不同类型的数据，采用适当的方法进行特征选择和降维，从而提高数据清洗的效果，并减少计算和存储的开销。其次，将完善多源异构数据治理体系的构建和管理机制。多源数据的种类繁多，数据结构和格式各不相同，需要建立统一的数据模型和标准，以便更好地对数据进行整合和分析。最后，将进一步探索云数据中心与边缘计算的结合方式，实现数据治理的分布式处理。

参考文献：

[1] 刘权纬，王兴辉，蒋红星. 基于学习事件的远程教育多源异构数据语义融合与实践研究 [J]. 成人教育，2023，43 (1)：39 - 48.

[2] 齐月松，储满生，唐 珏，等. 基于大数据技术的高炉数据治理研究进展 [J]. 冶金自动化，2023，47 (1)：43 - 52.

[3] 饶卫雄，高宏业，林 程，等. 基于半监督学习的多源异构数据治理 [J]. 同济大学学报 (自然科学版)，2022，50 (10)：1392 - 1404.

[4] 黄 艳. 基于大数据技术的区域教育网格化多维数据治理体系研究 [J]. 网络安全和信息化，2022 (10)：19 - 22.

[5] ZHOU F，YANG S，HE Y，et al. Fault diagnosis based on deep learning by extracting inherent common feature of multi-source heterogeneous data：[J]. Proceedings of the Institution of Mechanical Engineers，Part I：Journal of Systems and Control Engineering，2021，235 (10)：1858 - 1872.

[6] YUAN J，LUO X，LI Z，et al. Sustainable development evaluation on wind power compressed air energy storage projects based on multi-source heterogeneous data [J]. Renewable Energy，2021，169 (10)：1175 - 1189.

[7] BAI N B H. An integration approach of multi-source heterogeneous fuzzy spatiotemporal data based on RDF [J]. Journal of intelligent & fuzzy systems：Applications in Engineering and Technology，2021，40 (1)：1065 - 1082.

[8] 郑银巧，王 璇，王 蕾. 基于 HAO 治理模型的钢铁企业数据治理平台应用研究 [J]. 冶金经济与管理，2023 (4)：24 - 26.

[9] 陈玲玲，陈静霖，杨玉贤. 面向电力多源异构的数据治理及共享服务研究与应用 [J]. 中国高新科技，2021 (21)：76.

[10] 骞 巍，刘莎莎，孙晶莹，等. 多源异构数据治理技术在跨境电力运营监管过程中的应用与实践 [J]. 电力大数据，2021，24 (2)：55 - 60.

[11] 刘 宇，黎 琳，鲁 放，等. 轨道交通“四网融合”数据治理关键技术研究 [J]. 铁路计算机应用，2023，32 (6)：82 - 86.

[12] 吕仲琪，董卓达，刘晓丽，等. 多源异构数据特征的智能结构化方法仿真 [J]. 计算机仿真，2022，39 (7)：451 - 455.

[13] 孔亚宁，李春山，初佃辉. 面向多源异构数据的跨模态存储与检索系统 [J]. 南京大学学报 (自然科学)，2022，58 (3)：377 - 385.

[14] 于显浩，邱 伟，黄文龙，等. 多类型基建现场大数据治理平台 [J]. 水电与抽水蓄能，2023，9 (3)：85 - 89.

[15] 程红云，方 亮，李亚军，等. 钢铁企业 IT&OT 数据融合治理研究 [J]. 冶金自动化，2023，47 (s1)：415 - 417.

[16] 王 影，李柯景. 基于最小哈希的网络多路虚假数据清洗算法 [J]. 计算机仿真，2023，40 (5)：511 - 514，519.

[17] 李寅龙，杨森帆. 基于财务共享的企业大数据资源治理机制研究 [J]. 价格理论与实践，2023 (4)：104 - 108.

[18] 徐胜超，宋 娟，潘 欢. 云数据中心基于皮尔逊相关系数的虚拟机选择策略 [J]. 电子技术应用，2021，47 (10)：77 - 81.

[19] 谭章禄，王美君. 智能化煤矿数据治理概念模型及技术架构研究 [J]. 矿业科学学报，2023，8 (2)：242 - 255.

[20] 张 凯，薛嗣媛，周建设. 语言智能技术发展与语言数据治理技术模式构建 [J]. 语言战略研究，2022，7 (4)：35 - 48.

[21] 李俊丽. Spark 平台下类别数据互信息计算的并行化 [J]. 计算机工程与应用，2021，57 (7)：95 - 100.

[22] 熊菊霞，吴尽昭，王秋红. 邻域互信息熵的混合型数据决策代价属性约简 [J]. 小型微型计算机系统，2021，42 (8)：1584 - 1590.

[23] 骆菁菁，唐卫贞，丁继婷. 基于皮尔逊系数的管制仿真训练数据独立化与因子分析下的数据可视化研究 [J]. 计算机科学，2021，48 (z1)：623 - 628.

[24] 张文磊. 基于分布式数据治理技术的声像数据处理系统设计与实现 [J]. 广播电视信息，2022，29 (1)：91 - 94.