

应用自检机器人与关联规则的大数据全链路追溯算法

李明霞^{1†}, 王鹤森², 赵鑫², 刘佳音²

(1. 国网新疆电力有限公司电力科学研究院, 乌鲁木齐 新疆 830000;

2. 国网新疆电力有限公司营销服务中心, 乌鲁木齐 新疆 830000)

摘要:针对大数据维度高、数据量庞大的特点, 实际应用中大数据全链路追溯算法常面临性能不佳的挑战。为此, 提出了一种结合自检机器人与关联规则的优化算法。对采集的大数据实施数据归约与降维操作, 随后分析其特征。通过计算特征的支持度、置信度及提升度, 确立特征间的关联规则。计算节点综合度量值, 识别关键节点, 初步勾勒大数据全链路追溯路径。引入自检机器人, 实时监控追溯过程, 并评估追溯路径节点与实际节点的匹配度。若匹配度不足, 及时修正, 确保最终追溯路径的准确性。实验结果显示, 该算法在实际应用中表现出优越的追溯性能。

关键词: 自检机器人; 关联规则; 大数据; 全链路; 数据追溯; 追溯算法; 算法设计

中图分类号: TP309.2

文献标识码: A

A Big Data Full Link Traceability Algorithm Using Self Checking Robots and Association Rules

LI Mingxia^{1†}, WANG Heseng², ZHAO Xin², LIU Jiayin²

(1. Xinjiang Electric Power Research Institute State Grid, Urumqi, Xinjiang 830000, China;

2. Marketing Service Center, State Grid Xinjiang Electric Power Co., Ltd., Urumqi, Xinjiang 830000, China)

Abstract: Due to the high dimensionality and large amount of data in big data, the full chain traceability algorithm of big data often faces the challenge of poor performance in practical applications. Therefore, this article proposes an optimization algorithm that combines self checking robots with association rules. Implement data reduction and dimensionality reduction operations on the collected big data, and then analyze its characteristics. By calculating the support, confidence, and enhancement of features, establish association rules between features. Calculate the comprehensive measurement value of nodes, identify key nodes, and preliminarily outline the traceability path of the entire big data chain. Introduce self checking robots to monitor the traceability process in real time and evaluate the matching degree between the traceability path node and the actual node. If the matching degree is insufficient, make timely corrections to ensure the accuracy of the final traceability path. The experimental results show that the algorithm exhibits superior traceability performance in practical applications.

Key words: self checking robot; association rules; big data; full link; data traceability; traceability algorithm; algorithm design

在众多行业中, 从金融交易、医疗健康到物流运输, 大数据的应用已经无处不在, 深刻地改变了人们的生活方式和企业的运营模式。然而, 随着数

据量的激增, 如何高效、准确地管理和利用这些数据, 成为了一个亟待解决的问题。特别是在数据的安全性和可追溯性方面, 更是受到了广泛的关

注^[1]。然而,由于数据量过于庞大,现有的溯源方法仅能对单一数据源进行追溯,且数据追溯效率不高、准确性较低,难以满足数据溯源的实时需求。

因此,不少研究学者针对这一问题展开了研究,并对自己的观点进行了阐述。文献[2]先对电网多源时空数据进行建模,明确数据的来源、类型等关键信息,设计有限状态机模型,明确状态转移的条件和路径,利用区块链技术,构建电源多源时空数据的追溯系统,根据数据的哈希值定位数据位置,由此得到数据的转移路径,从而完成对数据的追溯。该方法应用的区块链技术具有不可篡改性和去中心化的特点,其数据存储和查询效率较低,无法满足数据的实时追溯需求。文献[3]收集新能源汽车在运行过程中的故障数据,并对其进行标准化处理,再从中提取出能够表征故障特征的参数和指标,并对故障进行分析,识别出不同的故障类型,在机器学习的作用下,建立故障类型与运行性能参数之间的映射关系,实现对故障类型的自动追溯。该方法需要处理大量的故障数据,会消耗大量的计算资源,方法的追溯时间较长。文献[4]先收集与特定数据源相关的所有信息,并对数据项进行标记,设计一个分布式区块链网络,并对不同的区块进行定义,且计算区块的哈希值,定义数据溯源的逻辑和规则,将收集到的数据存储到区块链,利用哈希值及数据验证机制对数据进行溯源。该方法在数据溯源的过程中,很容易出现数据泄露的情况,从而影响到溯源结果的准确性。文献[5]先对工业物联网的数据进行详细建模,精确数据的属性和结构,设计一个分布式区块链网络,对数据进行存储和分析,识别数据的内在规律和特征,利用智能合约,对数据进行查询,并设计数据验证机制,确保数据的安全性,并输出相应的溯源结果。该方法应用的区块链技术在实际应用中吞吐量较低,若需要进行处理的数据量过于庞大,区块链将会无法满足其实时处理的需求,导致数据溯源效率低下。

在以往研究的基础上,应用自检机器人与关联规则的大数据全链路追溯算法。本研究深入探讨大数据追溯的算法原理和实现方法,为相关领域的研究提供新的思路和方法。

1 大数据全链路追溯算法设计

1.1 大数据全链路特征分析

在对大数据进行全链路溯源时,需要先从各个数据源获取大量的数据,并对获取的数据进行优化处理。在优化处理时,先对其进行异常值识别,这个过程中需要设定数据属性值范围,若出现数据属

性值不在该范围内,则说明当前数据为异常数据,需要直接剔除^[6]。

在数据优化处理时,还需要对其进行数据归约和降维处理^[7]。其具体处理过程如下所示:

$$\begin{cases} x'_i = \frac{x_i - k_a}{k_b} \\ y_i = \frac{x'_i \times C_k}{|x'_i|} \times p_c \end{cases} \quad (1)$$

式中, x'_i 表示数据归约的结果, x_i 表示原始的数据, k_a 表示原始数据的均值, k_b 表示原始数据的标准差, y_i 表示数据降维处理的结果, C_k 表示数据协方差矩阵, p_c 表示数据的时间序列值。

将数据进行优化处理后,对大数据进行特征分析^[8]。其具体分析过程如下所示:

$$\begin{cases} K_x^i = A(x_i) \parallel B(x_i) \parallel C(x_i) \parallel D(x_i) \parallel E(x_i) \\ K_x^j = \frac{K_x^1 + K_x^2 + \dots + K_x^N}{N} \times \delta_x \times |y_i| \end{cases} \quad (2)$$

式中, K_x^i 表示数据 x_i 的自身中心特征, $A(x_i)$ 表示数据的度中心性, $B(x_i)$ 表示数据的介数中心性, $C(x_i)$ 表示数据的接近中心性, $D(x_i)$ 表示数据的复合中心性, $E(x_i)$ 表示数据的特征向量中心性, K_x^j 表示数据 x_i 邻近数据 x_j 的特征, N 表示数据 x_i 的数量, δ_x 表示数据的状态量。

1.2 应用关联规则的大数据全链路追溯节点确定

利用关联规则,识别出大数据全链路追溯的关键节点,由此确定大数据全链路追溯路径^[9]。基于数据特征,先建立数据之间的关联规则^[10]。在建立关联规则时,需要计算数据特征间的支持度、置信度和提升度^[11]。其具体计算过程如下所示:

$$\begin{cases} S(K_x^i \rightarrow K_x^j) = \frac{P(K_x^i \cup K_x^j)}{P_t} \\ C(K_x^i \rightarrow K_x^j) = \frac{P(K_x^i \cup K_x^j)}{P(K_x^i)} \\ L(K_x^i \rightarrow K_x^j) = \frac{P(K_x^i \cup K_x^j)}{P(K_x^i)P(K_x^j)} \end{cases} \quad (3)$$

式中, $S(K_x^i \rightarrow K_x^j)$ 表示数据特征的支持度, $P(K_x^i \cup K_x^j)$ 表示特征 K_x^i 和特征 K_x^j 同时出现的概率, P_t 表示所有数据特征出现的概率, $C(K_x^i \rightarrow K_x^j)$ 表示数据特征的置信度, $P(K_x^i)$ 表示特征 K_x^i 出现的概率, $L(K_x^i \rightarrow K_x^j)$ 表示数据特征的提升度, $P(K_x^j)$ 表示特征 K_x^j 出现的概率。

在建立数据关联规则的过程中,基于支持度、置信度和提升度^[12],若 $S(K_x^i \rightarrow K_x^j)$ 高于所设阈值,则当前数据特征为频繁特征,可以将其作为重要的关联指标。从频繁特征集中筛选出满足置信度阈值和提升度阈值的特征,构建相应的关联规则,逐一识别数据追溯过程中的关键节点。

计算节点的综合度量值^[13]:

$$K_p = \frac{\alpha_S S + \alpha_C C + \alpha_L L}{\delta_p |S + C + L|}$$

(4)

式中, K_p 表示节点综合度量值, α_S 、 α_C 、 α_L 分别表示节点数据特征支持度、置信度和提升度对应的权重占比, δ_p 表示节点的关联系数。

根据计算的 K_p 确定大数据全链路追溯关键节点,为后续的数据追溯路径确定奠定基础。

1.3 应用自检机器人的大数据全链路追溯路径

确定关键节点后,对关键节点进行拟合,再结合数据的变化趋势和特征属性,初步确定大数据全链路追溯路径^[14]:

$$\begin{cases} L_z = \kappa[\sigma(x_i)\mathbf{W}_c + \mathbf{B}_c] \times K_p \times P_c \\ P_c = \frac{e^p}{\sum c^{e^p}} \\ \sigma(x_i) = m_i \odot x_i^m \odot \eta[E(v)\mathbf{W}_c + \mathbf{B}_c] \end{cases}$$

(5)

式中, L_z 表示大数据全链路追溯路径, κ 表示损失平衡系数, $\sigma(x_i)$ 表示分类函数, $\mathbf{W}_c$ 表示权重矩阵, $\mathbf{B}_c$ 表示误差矩阵, P_c 表示数据源判断函数, c 表示数据类别, p 表示数据类型概率值, m_i 表示平衡因子, x_i^m 表示节点维度值, η 表示节点异质特征, $E(v)$ 表示路径差异值。

考虑到式(5)生成的追溯路径可能会受到环境因素的影响,为提高追溯路径的可靠性,应用自检机器人,对大数据全链路追溯路径进行实时修正,从而得到最终的追溯路径。其数据追溯路径确定的具体过程如图 1 所示。

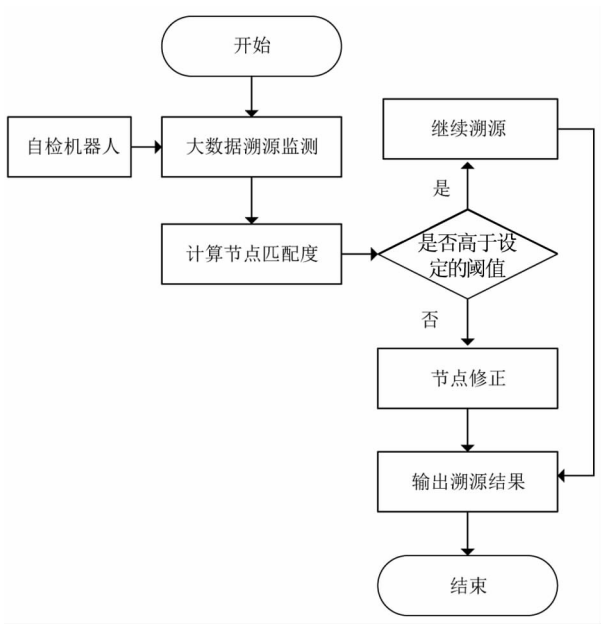


图 1 应用自检机器人的大数据全链路追溯路径确定的具体过程

如图 1 所示,利用自检机器人,实时监测大数

据的溯源过程,自检机器人能够不断感知环境,并计算溯源路径节点与实际路径节点之间的匹配度,若匹配度不高,则需要对其进行修正处理,反之,则不需要修正^[15]。在该过程中,利用自检机器人计算追溯路径节点与实际节点之间的匹配度,从而确定最终的追溯路径。应用自检机器人计算节点匹配度的具体过程如下所示:

$$S_d = \frac{\sum_{i=1}^n L_z^i R^i}{\sqrt{\sum_{i=1}^n (L_z^i)^2} \sqrt{\sum_{i=1}^n (R^i)^2}}$$

(6)

式中, S_d 表示追溯路径节点与实际路径节点之间的匹配度, L_z^i 表示确定的追溯路径的第 i 个节点, R^i 表示实际路径的第 i 个节点, P_k 表示自检机器人的参数。

若计算的匹配度低于设定的阈值,则需要对其进行修正。得到最终的追溯路径如下所示:

$$L'_z = L_z \times \beta_c \times \frac{z_k \times l_k}{f_c} \times z_m$$

(7)

式中, L'_z 表示最终的数据追溯路径, β_c 表示修正系数, z_k 表示追溯路径的误差值, l_k 表示追溯路径更新参数, f_c 表示追溯路径当前状态值, z_m 表示自检机器人的调整参数。

利用公式,实现对大数据全链路的追溯,以便更好地实现对各种类型数据的定位。至此,应用自检机器人和关联规则的大数据全链路追溯算法的设计完成。

2 实验

硬件环境如下: Intel Xeon W-2133 处理器,主频 3.6 GHz, 32 GB DDR4 ECC 内存, 1 TB NVMe SSD 硬盘, Cisco Meraki MR33 路由器。

软件环境如下: MATLAB R2023a 编程软件, ROS 机器人操作系统, Visual Studio Code 文本编辑工具, PostgreSQL 数据库管理系统等。

应用的自检机器人型号为 ZR-AutoInspect-01, 技术参数见表 1。

表 1 自检机器人技术参数		
序号	技术参数	参数设置
1	机器人尺寸	120 cm×60 cm×50 cm
2	重量	50 kg
3	电池容量	10000 mAh
4	数据采集频率	5~10 次/s
5	定位精度	±5 cm

设 $S(K_x^i \rightarrow K_x^j)$ 阈值为 0.05, $C(K_x^i \rightarrow K_x^j)$ 阈值为 0.7, $L(K_x^i \rightarrow K_x^j)$ 阈值为 0.5, 迭代次数为 100, 平衡因子 m_i 为 0.88, 损失平衡系数 κ 为 0.45, 修正系数 β_c 为 0.96, 追溯路径更新参数 l_k 为 4.88。

所用数据集为公开数据集, 该数据集数据类型多样, 包含用户行为数据、环境数据、机器人状态数据等, 数据总量达到了 236954 条。计算出多个节点的综合度量值, 如表 2 所示。

本文算法为算法 1, 基于有限状态机的电网多源时空数据的区块链追溯方法研究为算法 2, 新能源汽车运行故障数据分析与自动追溯模型研究为算法 3。利用三种算法进行追溯, 得到其追溯路径, 计算其追溯结果的 AUC 值, 若 AUC 值越接近于 1, 则算法的追溯性能越好, 如图 2 所示。

表 2 多个节点的综合度量值

节点编号	支持度	置信度	提升度	综合度量值
#001	0.44	0.36	0.81	0.88
#002	0.36	0.54	0.12	0.96
#003	0.52	0.58	0.24	0.87
#004	0.45	0.24	0.51	0.45
#005	0.26	0.36	0.26	0.56
#006	0.74	0.25	0.36	0.32
#007	0.36	0.36	0.45	0.91
#008	0.21	0.45	0.41	0.82
#009	0.13	0.24	0.25	0.44
#010	0.24	0.15	0.24	0.86
#011	0.25	0.36	0.23	0.52
#012	0.36	0.39	0.26	0.83

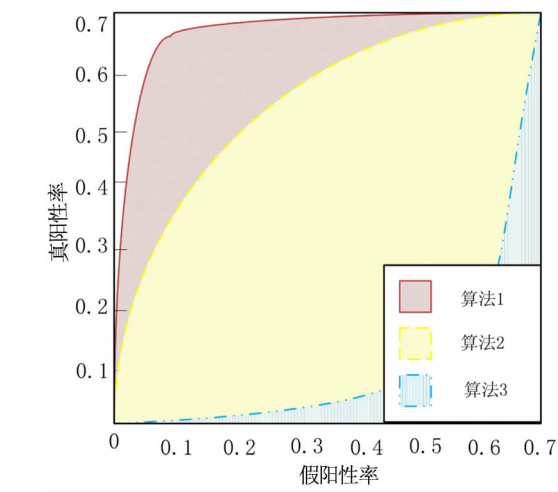


图 2 三种算法的 AUC 值

如图 2 所示, 算法 1 的 AUC 值远高于其余两种算法, 且更为接近 1, 说明其能够准确地对数据

变化趋势进行判断, 追溯性能较好。算法 2 和算法 3 的 AUC 值数值较小, 追溯性能较差。

以覆盖率为评价指标, 即算法追溯到信息占总信息的比例, 验证具体追溯结果, 如表 3 所示。

如表 3 所示, 算法 1 的覆盖率远高于其余两种算法, 说明其在追溯过程中, 能够准确识别路径中的关键节点, 保证节点信息的获取, 追溯性能较好。

表 3 三种算法的覆盖率

数据编号	算法 1/%	算法 2/%	算法 3/%
1	99.85	75.42	56.12
2	98.72	76.33	54.26
3	98.65	75.12	53.22
4	97.12	74.15	58.12
5	96.25	76.25	57.12
6	98.55	73.25	56.25
7	98.78	76.23	54.17
8	99.05	77.14	56.26
9	98.69	72.25	58.24
10	97.15	73.26	57.12
11	98.16	73.14	55.13
12	98.88	74.28	52.36

3 结 论

本研究将自检机器人的自动化检测能力与关联规则的深度挖掘相结合, 对大数据全链路进行了精准追溯, 不仅能够快速定位数据流向, 还能揭示数据之间的潜在关联, 为决策者提供有力支持。在数据安全领域, 这一算法有助于及时发现潜在风险, 确保数据的安全性和完整性; 在业务管理领域, 可以优化流程, 提高运营效率。

参考文献

[1] 张勇, 王艺博, 张进伟, 等. 基于区块链的电网调控数据溯源技术研究[J]. 电力信息与通信技术, 2023, 21 (11): 70—76.

[2] 谷志红, 赵进斌, 王正, 等. 基于有限状态机的电网多源时空数据的区块链追溯方法研究[J]. 电网与清洁能源, 2024, 40 (4): 9—17.

[3] 阮永娇, 陈昕, 孙承臻, 等. 新能源汽车运行故障数据分析与自动追溯模型研究[J]. 辽宁工业大学学报(自然科学版), 2022, 42 (5): 316—319.

[4] 杨叶芬, 何拥军. 大数据视域下区块链技术在数据溯源中的应用探究[J]. 长江信息通信, 2024, 37 (3): 148—151.

[5] 李焕. 基于区块链的工业互联网数据溯源技术实现[J]. 自动化与仪器仪表, 2024, 42(1): 89—92+98.

[6] 常雨晴,贺婉朦,周璐瑶,等. 基于智能合约和雾计算的医疗数据可追溯加密方案[J]. 信息安全研究, 2024, 10 (6): 554—560.

[7] 张华,杨灵运,谷牧,等. 基于标识编码的工业产品数据全程可信追溯研究 [J]. 仪器仪表标准化与计量, 2024, 14(1): 5—8.

[8] 王民涛,陈立钊. 基于区间型数据离散化算法的电力应急物资供应质量可追溯模型研究[J]. 机械设计与制造工程, 2024, 53 (2): 130—134.

[9] 田宏明. 电加热炉变电设备运行质量安全数据追溯方法设计[J]. 工业加热, 2023, 52 (11): 43—47.

[10] 胡牛平,景征骏,史培中,等. 基于区块链的 ETO 型企业生产过程溯源系统研究[J]. 科技和产业, 2023, 23 (11): 120—

128.

[11] 顾斌,邹云峰,单超,等. 基于数字水印的电力数据安全分发与溯源技术研究 [J]. 信息技术, 2023, 35(11): 99—104.

[12] 吴敏,张明达,李盼盼,等. 面向多源遥感影像数据的溯源模型研究[J]. 地球信息科学学报, 2023, 25 (7): 1325—1335.

[13] 景一佳,童一飞,赵建波. 基于区块链的军事物资质量信息溯源系统的设计与研究[J]. 机械设计与制造工程, 2023, 52 (6): 113—118.

[14] 徐滨,翁年凤,樊树海,等. 面向大规模定制的制造业领域数据溯源模型研究[J]. 机床与液压, 2023, 51 (8): 1—7.

[15] 王莹,穆力,宋继红,等. 基于区块链技术的光伏扶贫资金精准溯源方法[J]. 电力大数据, 2022, 25 (12): 69—76.

(上接第 106 页)

4 结 论

目前中国企业已基本完成了自动化改造,拥有大量的自动化设备仪器;在向“云一边一端”数字体系架构转型过程中,大量接口封闭型和无数据接口型仪表的存在使管理者难以准确地感知产品生产过程中的质量态势。针对工业现场中的 LCD 仪表数字识别需求,提出了一种采用 Jetson 嵌入式平台的 LCD 仪表数字识别方法,通过设定数字 ROI 区域减少冗余图像信息,采用帧间差分法过滤显示屏帧切换过程中的数字重影图像,最后基于 PP-OCrv2 架构中的数字检测和识别模型获得准确的仪表数值。该方法每帧图像的识别周期平均为 0.37 s,可以满足工厂大多数场景的仪表数字识别需求。未来的工作中将继续采用 Jetson 嵌入式平台,进行多数据组的识别要求,探索在保持精度的情况下压缩模型以及把设备部署至“云一边一端”架构。

参考文献

[1] 周济,周艳红,王柏村,等. 面向新一代智能制造的人-信息-

物理系统(HCPS)[J]. Engineering, 2019, 5(4): 71—97.

[2] 袁沐坤,于广平,刘坚,等. 基于感知—决策—评估的污水处理智能曝气方法[J]. 工业水处理, 2022, 42(4): 65—72.

[3] 何王金,于广平,郭清达,等. 基于神经网络的锂萃取率软测量模型研究[J]. 计算机仿真, 2021, 38(2): 174—179.

[4] 陈开峰,俞伟聪,唐雁文,等. 基于轻量化深度学习网络的数字仪表识别[J]. 科学技术与工程, 2023, 23(2): 674—680.

[5] 华泽玺,施会斌,罗彦,等. 基于轻量级 YOLO-v4 模型的变电站数字仪表检测识别[J]. 西南交通大学学报, 2024, 59(1): 70—80.

[6] 曾科,高潮,扶新,等. 多参数数显仪表的自动识别方法研究[J]. 中国测试, 2018, 44(12): 122—128.

[7] 杨鑫威,邓春华. 基于巡检机器人的数字仪表自动识别方法[J]. 计算机工程与设计, 2023, 44(5): 1502—1510.

[8] 祝长生,张成元,徐亚超. 温湿度检定仪中的数字式仪表识别[J]. 科学技术与工程, 2020, 20(34):14120—14125.

[9] 寇文博,屈八一,李智奇. 一种改进 Transformer 的仪表字符识别算法[J]. 自动化与仪器仪表, 2022, 7: 284—288.

[10] KARTHICK K, KAVASKAR S. Text detection and recognition in raw image dataset of seven segment digital energy meter display[J]. Energy Reports, 2019, 5: 842—852.

[11] DU Y N, LI C X, GUO R Y, et al. PP-OCrv2: bag of tricks for ultra lightweight OCR system[J]. arXiv 2021, arXiv: 2109.03144.

[12] 付莲莲,丁鑫圣,王映龙. 基于帧间差分法的口罩佩戴检测算法研究[J]. 计算机仿真, 2022, 39(10): 250—254.