工业时序大数据质量管理

丁小欧，王宏志，于晟健
哈尔滨工业大学海量数据计算研究中心，黑龙江 哈尔滨 150001

摘要
工业大数据已经成为我国制造业转型升级的重要战略资源，工业大数据分析问题正引起重视和关注。时序数据作为工业大数据中一种重要的数据形式，存在大量的数据质量问题，需要设计数据清洗方法对其进行检测和有效处理。介绍了工业时序大数据的特点及工业数据质量管理的难点，并对工业时序大数据质量管理的研究现状加以分析、总结，最后，提出了时序大数据质量管理方法和系统性能的提升方向。

关键词
数据质量管理；时序数据；工业大数据分析；数据清洗

中图分类号：TP311 文献标识码：A doi: 10.11959/j.issn.2096-0271.2019047

Data quality management of industrial temporal big data

DING Xiaoou, WANG Hongzhi, YU Shengjian
Harbin Institute of Technology, Massive Data Computing Research Center, Harbin 150001, China

Abstract
Industrial big data has become an important strategic resource for the transformation and upgrading of China's manufacturing industry, and industrial big data analysis is attracting more and more attention. As an important data form of industrial big data, time series have a lot of quality problems, which is necessary to be detected and handled effectively by designing data cleaning methods. The characteristics of industrial time series big data and the difficulties of industrial data quality management were introduced. Then, the recent developments in the area of that was analyzed and summarized. At last, the quality management method of temporal big data and the improvement direction of system performance were put forward.

Key words
data quality management, time series data, industrial big data analysis, data cleaning

1 引言

随着“工业4.0”和“中国制造2025”等国家战略的提出，我国的大量产业面临着向智能制造转型的重大需求。而5G时代的到来，更是对工业物联网产生的海量大数据质量监控与分析提出了更高层次的精准性和时效性要求[1-3]。

新时期的科学技术革命也给我国制造业的发展带来了新的机遇。目前，我国工业生产过程中已经产生并积累了大量数据，现代化工业制造生产线传感器设备、制造装置监视器等设备能实现对整体工业生产状态和运行参数的感知和记录[4]。而在积累的工业大数据中，时间序列数据是最基本和最普遍的数据形式。对基于采样时间点的时序数据的分析挖掘，能够促进工业大数据分析研究的发展。如果想实现可靠、智能化的工业大数据分析，就需要在高质量的工业数据上进行建模计算。然而，工业时序数据来源广泛，具有大体量、多源性、连续采样、价值密度低、动态性强等特点[5]，导致目前的工业数据质量问题广泛存在。

很多工业生产环境在数据系统智能化中经常遇到瓶颈问题，根据其数据形式，这些问题可归结为与时间序列有关的数据质量问题。但由于数据采集环境不同，不同系统中的针对性解决方案较多，目前学术界的方法仍不是很完备。因此，本文将对近年来数据质量管理和数据清洗的研究现状，尤其是时序大数据质量管理的研究现状进行全面分析。

2 工业背景下的数据质量管理

2.1 时序大数据的特点

与静态数据不同，时间序列数据之间存在大量的依赖关系，对数据依赖关系或相关性的正确处理在时间数据处理中变得至关重要。时间序列数据在统计学上经过了数十年的研究，已经有大量的工作用于检测时间序列数据的离群值和异常值。硬件和软件技术的进步推动了多种应用程序生成的数据集的增长，包括数据流、时空数据、时间网络等。

传感器设备的快速发展导致各个领域对时间序列数据的计算提出大量需求。因此，数据挖掘在时间序列领域的探索也日渐增多。Tang Y等人[6]通过弹性距离测量函数将原始时间序列空间隐式映射到多内核空间，从而实现多内核聚类（multiple kernel clustering，MKC）框架下的时间序列聚类任务；Rawassizadeh R等人[7]设计了一组可伸缩的算法，以时间粒度识别行为模型，即通过智能手机收集的多维时间序列来识别人类日常行为的模式；Zhao J等人[8]采用词袋（bag of words，BoW）框架对时间序列进行分类，即从特征点位置的时间序列采样局部子序列，从而构建局部描述符，并通过高斯混合模型对其分布进行建模后编码，最后使用现有分类器（例如SVM）进行训练和预测；González-Vidal A等人[9]提出了一种用于时间序列分割的BEATS算法，该算法将数据流分成多个块，并按平方矩阵将其分组计算离散余弦变换（discrete cosine transform，DCT），并对其量化，提取子矩阵计算其特征值模数，并删除重复项；Yagoub D等人[10]提出了一种并行索引算法和并行查询处理策略，该策略可以拓展到数十亿个时间序列，并在给定查询的情况下有效地利用索引。

与其他数据类型不同的是，时间序列具有高维度、扭曲化、长度不一及多弹性度量集成等特点。尤其是在解决高维度时，即“基数诅咒”的问题上，已有大量

学者提出了各类解决方法。Liu C等人[11]提出了一种“时间骨架化”方法，通过发现重要的时间结构来降低序列基数，关键思想是总结无向图中的时间相关性，并使用图的“骨架”作为更高的粒度。但高维度在带来困难的同时也带来了新的研究方向，Agrawal S等人[12]利用相关网络挖掘时间序列数据中的多维关系，提出了多极点和负相关团的概念，并证明了其实际意义和效用。Batu B B等人[13]提出了一种使用非参数随机de-clustering过程和多元Hawkes模型来定义事件类型内部和事件类型间触发关系的算法。Han M等人[14]提出了一种结构化流形广泛学习系统（structured manifold broad learning system，SM-BLS）揭示动态系统的演化状态，并自动发现变量间的关系。Malensek M等人[15]提出了一种使用分布式散列表进行时间序列数据分析查询的方法，该方法无须索引每个离散值进行后续检索，而是自动学习各维度之间的关系和相互作用而使信息易于用户使用。在解决高维度问题的同时，还可将目光聚焦在特定维度上，如Hao Y等人[16]提出了现象特性变量（phenomenon-specific variable，PV）的概念，即在不同变量中只有较少的变量对特定现象有重大影响，对不同现象起重要作用的变量通常也不同。Chen D等人[17]提出了一种提取时间序列潜在因素的算法，可将其作为检查动态复杂系统的重要手段，即用高维数据中低维和“小”的表现形式来突出数据中的潜在特征。

2.2 时序大数据质量问题

所有的时间数据质量管理都得益于“时间连续性”的存在，这是一个基础，使用数据中的异常变化、序列或时间模式对时间序列进行建模。时序大数据质量问题包括异常检测、异常修复或删除，即数据清洗问题。其中，时间序列异常值分析、更改点检测与事件检测密切相关。笔者将在第3.2节详细阐述其研究现状。异常检测通常被定义为与期望（或预测）的偏差。

（1）时间序列与多维数据

在时间序列数据（例如传感器读数）中，时间连续性至关重要，并且所有分析都要在合理使用较小的时间窗口（上下文变量）的情况下进行。另外，在诸如文本新闻线流之类的多维数据流中，第一层检测之类的应用程序可能并不严重依赖于时间，因此这些方法更接近于标准的多维异常分析。

（2）点与窗口

笔者在时间序列中寻找异常数据点（如心电图读数中，心率突然跳跃）或异常变化模式（连续性心电图模式指示心律失常）。后一种情况通常比前一种情况更具挑战性。即使在多维数据流的上下文中，单点偏差（例如，新闻专线流中的第一个故事）也可能被视为离聚合变化点。

（3）数据类型

不同种类的数据，例如连续系列（如传感器）、离散系列（如网络日志）、多维流（如文本流）、网络数据（如图形和社交流）需要各种专用的分析方法。

2.3 工业时序大数据质量管理的重要性

数据可以帮助人们分析问题，制定决策。然而，数据质量管理仍然是一个主要问题，“脏”数据可能导致不可靠的分析和错误的决策。常见的数据质量问题包括值缺失、格式不一致、数据重复、数据异常和违反业务规则等。分析人员在做任何决策之前，必须考虑脏数据的影响，因此，数据清洗已经成为数据库研究的一个关键领域[18]。Chu X等人[18]将异常检测方法分为两类：

使用约束和规则来检测和修复错误的定性方法；使用统计方法来识别和修复错误的定量方法。

而在工业大数据方面，制造业系统中会存在由产品质量缺陷、设备故障或外部环境突变等因素导致的异常问题[19]。因此，异常和故障工况检测、故障监测、设备健康状态分析等是实现高效生产和智能制造的重要保障，也是工业数据质量管理中重要的具体研究任务[5]。如果在工业生产中出现的异常、故障或危机情况不能被及时地识别和解决，将导致生产环境存在安全隐患，甚至会给整个工厂的智造系统带来难以估量的负面影响，造成重大经济损失。

2.4 研究挑战

通过调研，笔者总结了工业时序数据质量管理的研究难点，具体如下[2,4]。

- 工业大数据除了具有规模大、速度快、类型杂和质量低等基本特征，还具有多模态、强关联和高通量等新特征，即模式多样、多变的工况数据，其数据特点导致了传统数据质量管理模型不能很好地适用于工业时序数据质量管理。
- 工业时序数据质量管理不同于其他数据分析。对于复杂的工业时序数据质量管理，要强调因果性、领域知识和数据分析过程的深度融合、复杂问题简单化。同时，工业数据复杂性的增加也会导致分析工作失败概率增加。因此，前期准备工作和后期评估验证工作显得更为重要。
- 已有的数据分析基础算法变化不大，但将其应用到工业时序数据的过程却非常复杂。因此，工业时序数据质量管理的研究是一个持续改进、修正和完善的过程。
- 工业时序数据具有典型的高维度特征。而目前大部分数据质量管理系统专注于解决单维度、有周期性或简单模式的数据，难以对高维度时间序列数据进行有效的质量管理。

3 时序大数据质量管理研究现状

3.1 数据质量管理和数据清洗方法

异常检测是数据质量管理中的重点工作之一。Liu Y等人[20]提出了一种单目标生成对抗性主动学习（single-objective generative adversarial active learning，SO-GAAL）方法，用于离群值检测，该方法可以基于生成器和判别器之间的最小极大博弈直接生成信息性的离群值。Hu W等人[21]提出了使用局部核密度估计和基于上下文的回归进行异常检测的方法，该方法通过加权邻域密度估计增加对邻域大小变化的鲁棒性，并且组合来自多尺度的邻域信息，以细化样本的异常因子。Sharma V等人[22]提出了自愈神经模糊方法（neuro-fuzzy based horizontal anomaly detection，NHAD），并将其应用于在线社交网络进行水平异常检测，检测的异常内容为允许未经授权的用户访问信息以及伪造信息的在线欺诈，而表现的像无声攻击的异常之一是水平异常。Lu Y等人[23]提出了混合类型鲁棒检测（mixed-type robust detection，MITRE）模型，该方法是一种用于混合类型数据集中异常检测的鲁棒错误缓冲方法，使用了集成嵌套拉普拉斯近似（integrated-nested Laplace approximation，INLA）和变分期望最大化（expectation maximization，EM）的期望传播（expectation propagation，EP）。

在异常检测的同时，可对数据进行清理或修复。Lin X等人[24]提出了基于众包的方法，从而清理不确定图中的边缘，考虑

到众包的时间成本，作者提出了一系列边缘选择算法、优化技术和修剪启发式方法，从而减少计算时间。Hao S等人[25]提出了一种基于成本的数据修复模型，该模型是由两个部分组成的迭代过程，检测违反给定完整性约束（integrity constraint，IC）的错误，并修改各组中的值，以使修改后的数据库满足完整性约束。Dasu T等人[26]提出了一种新的数据清洗评价策略，除了修复错误的效果、修复的花费之外，还考虑了数据的统计信息变化。Bohannon P等人[27]提出了一个基于代价的启发式修复模型，并提出了“最小化修复策略”，指导了后续的研究。Song S等人[28]提出了“宽容约束”，清洗数据使其符合某种约束，且在给定约束的给定相似度内。Li Z等人[29]使用正则表达式代替约束和规则进行数据清洗，因为正则表达式更适合进行语法修改，可以解决结构的增删问题；Khayyat Z等人[30]提出了分布式的数据清洗系统，按照制定质量规则、检测数据错误、修复的步骤进行数据清洗。Jensen S K等人[31]将已经发表的时间序列管理系统（time series management system，TSMS）进行了全面的分析和分类，列举的TSMS均用于物联网（Internet of things，IoT）、时序数据监测、时序数据分析和时序数据评估。

3.2 时序大数据清洗方法

时序数据质量管理和数据清洗同样重要。主流时序大数据清洗方法主要分为基于统计的清洗（statistical-based cleaning）、基于约束的清洗（constraints-based cleaning）和基于机器学习的清洗（machine-learning-based cleaning），本节依次对这3个方法进行阐述。

（1）基于统计的清洗

基于统计的清洗是一类相对传统的方法，通过对时间序列求取统计量和统计规律、模型参数拟合或数据形态转变来达到提取时间序列趋势，检测、清洗低质量数据点的目的。苏卫星等人[32]使用有效分数向量和小波分析统计量来有效地提取时间序列趋势，并利用李氏指数与小波变换的关系构建了同时检测异常点和突变点的时间序列数据清洗框架。Salehi M等人[33]着力于解决使用局部异常因子（local outlier factor）进行异常检测时内存不足的问题，提出了将超出内存限制的历史数据点进行整合的观点，并提出了动态计算整合数量的算法。Cao L等人[34]通过距离空间定义了3种异常，作者基于异常点出现的低概率规律和新数据点的高信息量规律设计了系统，将违反这两种统计规律的数据点识别为异常数据点，并进行捕捉。Yang F等人[35]使用领域知识对历史序列数据进行基线抽取，并使用湮没滤波技术（annihilating filter technique，AFT）对时间序列进行精细重构，以突出基线中的高频模式，借此提升数据质量。Arous I等人[36]对高维时间序列数据库进行线性插值，然后迭代使用中心分解技术进行更新，得到修复数据。Wu S等人[37]针对带有事件标签序列的时间序列中分段缺失的问题，使用两个低阶矩阵的内积和一个事件分量来近似表征时间序列的汉克尔矩阵（Hankel matrix）的方法，对矩阵中相关元素取平均，得到一个缺失值的估计。Feng K等人[38]针对时空数据挖掘问题中的重大事件检测，提出了基于网格搜索和剪枝的方法，以通过实时更新的推特数据检测何时何地出现了异常的重大事件。Ma M等人[39]搭建了检测KPI序列中概念漂移的系统，首先使用基于奇异值分解的奇异谱变换（singular spectrum transformation，SST）进行异常检测，再使用双重差分模型（difference in difference，DID）判断该

异常是不是由软件的变更引起的，最后使用一层线性相关性的过滤逻辑来判断是否出现了概念漂移的情况。Mei J等人[40]使用带有辅助信息的非负矩阵分解对时间序列进行修复和预测，该方法将经典分层交替最小二乘（hierarchical alternating least squares，HALS）算法改进为HALSX（具有异类变量的HALS），对外部特征和响应变量之间的非线性关系进行建模。Rong K等人[41]使用平滑方法来消除小范围的变化，突出显著的偏差，生成高质量的可视化数据。但是基于平滑的方法几乎修改了所有的数据，原本正常的数据也进行了修改，数据准确性不高。Yoon S等人[42]改进了基于滑动窗口的异常检测方法，使用集合代替单一的数据点，这样避免了不必要的数据更新，同时避免了潜在的错误异常点的出现。

（2）基于约束的清洗

作为近年来逐渐兴起的方法，基于约束的清洗方法旨在利用相邻序列的相关性或统计量来确定序列的值是否发生了异常。Song S等人[43]提出了速度约束（speed constraints）的方法，通过对相邻时间戳的速度变化进行计算，使用一个速度约束区间来检测并修复发生了异常的数据。其后，又提出了一种基于顺序约束的清洗方法，在速度约束的基础上，假设连续相邻时间点的值的变化幅度在一定范围内，根据这一规则进行统计，就能够对小幅异常的数据执行更合理的清洗[44]。Yin W等人[45]提出了方差约束，使用一个滑窗确定的区间内方差与阈值的关系来判断是否存在异常，并使用自回归平滑的方法修复异常。Sadik S等人[46]则计算异构和异步的时间序列数据流的相关性，并将违反自相关性约束或与其他属性相关性约束的数据点检测为异常点。

同时，时序数据的时间戳和时效性同样可能存在错误，Song S等人[47]使用时序约束图对时间戳进行修复。Abedjan Z等人[48]挖掘近似的时序函数依赖，以实现数据时效性的检测和修复。

（3）基于机器学习的清洗

基于机器学习的清洗方法是一种将传统的分类、聚类、异常检测和深度学习等思想应用在时间序列上，以提高数据质量的方法。由于序列模型相比于向量空间模型更复杂，目前这一类方法在时间序列上的应用仍处于起步状态。陈乾等人[49]基于懒惰学习（lazy-learning，LL）的思想，将距离度量和具有遗忘因子的最小二乘法结合，以补足距离度量对于历史信息模式的缺失问题。Milani M等人[50]通过搭建个人数据清洗系统，融合了时空影响模块、更新预测模块和数据修复模块，以学习数据之间的统计特征和相关性，进而得到缺失数据的估计和劣质数据的代价限制的最大似然修复。Zameni M等人[51]使用滑窗截取时间序列片段，通过最优化窗内信息增益度量来实行时间序列的变点检测，并通过置换检验来判断是否接受模型给出的假设。Souiden I等人[52]则聚焦于云计算中的用户恶意操作这一时间序列异常问题，在云端数据流处理系统（cloud stream generator，CSG）中集成了基于层次聚类模型clus-tree的Anyout模块和基于微聚类（micro-cluster）的MCOD模块，以在极低的响应时间内预测异常数据点，并避免用户恶意操作。Haque A等人[53]针对分类器的表现随时间序列的概念漂移（concept drift）呈现的低效性，将基于滑窗的半监督算法SAND使用类kNN的聚类模型进行集成，并使用联合性（association）和纯净度（purity）来估计模型的置信度（confidence），最终使用阈值法决定何时分类器工作异常。随着深度学习的兴起，基于变分自编码器（variance auto-encoder，VAE）[54]、基于长短期记忆网络（long

short term memory, LSTM)[55]、基于生成式对抗网络(generative adversarial nets, GAN)[20]等深度学习方法也逐渐应用于时间序列的异常检测，但是由于这一类方法需要大量有标签的训练数据以及长足的神经网络训练时间，因此不适用于工业时间序列的场景。

时序数据异常检测是与领域高度相关的问题，从一个领域到另一个领域，异常的组成可能发生很大的变化。Eichmann P等人[56]认为人工干预异常检测能产生较好的效果，设计了一个交互式的工具，供数据科学家部署多个异常检测器，可以方便地比较不同检测器的效果。

4 结束语

已有的时间序列数据质量管理方法和系统的性能仍有许多提升空间。而针对最新出现的工业大数据中遇到的实际问题，包括存在多变的工业机器运行模式、超高的数据属性维度以及极弱周期性的时间序列清洗问题，未来可考虑对相关性机理进行发掘，强调因果性而不是单纯的相关性关系。

同时，要注重在线离线算法的设计，基于历史数据进行工业大数据分析具有极大的局限性：在数据量大、分布完整、质量良好的前提下，可以建立理想的数据模型，但当模型涉及范围广、影响因素多、复杂多维且机理不清晰时，很难有足够的数据来建立和验证模型。因此，要充分利用专业领域知识克服这一局限性。

参考文献:

[1] 张洁，秦威，鲍劲松，等. 制造业大数据[M]. 上海：上海科学技术出版社，2016.
ZHANG J, QIN W, BAO J S, et al. Big data in manufacturing industry[M]. Shanghai: Shanghai Scientific & Technical Publishers, 2016.

[2] 工业互联网产业联盟工业大数据特设组. 工业大数据技术与应用实践[M]. 北京：电子工业出版社，2017.
Industrial Big Data Task Group in Alliance of Industrial Internet. Industrial big data technology and application practice[M]. Beijing: Publishing House of Electronics Industry, 2017.

[3] 国家制造强国建设战略咨询委员会. 《中国制造2025》重点领域技术路线图[Z]. 北京，2015.
National Manufacturing Strategy Advisory Committee. “Made in China 2025” technology roadmap for key areas[Z]. Beijing, 2015.

[4] 工业互联网产业联盟. 中国工业大数据技术与应用白皮书[Z]. 北京，2017.
Alliance of Industry Internet. White paper on big data technology and application in China’s industry[Z]. Beijing, 2017.

[5] 王建民. 工业大数据技术综述[J]. 大数据，2017, 3(6): 3-14.
WANG J M. Survey on industrial big data[J]. Big Data Research, 2017, 3(6): 3-14.

[6] TANG Y, XIE Y, YANG X, et al. Tensor multi-elastic kernel self-paced learning for time series clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2019: 10.1109/TKDE.2019.2937027.

[7] RAWASSIZADEH R, MOMENI E, DOBBINS C, et al. Scalable daily human behavioral pattern mining from multivariate temporal data[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(11): 3098-3112.

[8] ZHAO J, ITTI L. Classifying time series using local descriptors with hybrid sampling[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(3): 623-637.

[9] GONZÁLEZ-VIDAL A, BARNAGHI P, SKARMETA A F. BEATS: blocks of eigenvalues algorithm for time series segmentation[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(11): 2051-2064.

[10] YAGOUBI D, AKBARINIA R, MASSEGLIA F, et al. Massively distributed time series indexing and querying[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 32(1): 108-120.

[11] LIU C, ZHANG K, XIONG H, et al. Temporal sclerotization on sequential data: patterns, categorization, and visualization[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(1): 211-223.

[12] AGRAWAL S, STEINBACH M, BOLEY D, et al. Mining novel multivariate relationships in time series data using correlation networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2019: 10.1109/TKDE.2019.2911681.

[13] BATU B B, TEMIZEL T T, DÜZGÜN H Ş. A non-parametric algorithm for discovering triggering patterns of spatio-temporal event types[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(12): 2629-2642.

[14] HAN M, FENG S, CHEN C L P, et al. Structured manifold broad learning system: a manifold perspective for large-scale chaotic time series analysis and prediction[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(9): 1809-1821.

[15] MALENSEK M, PALLICKARA S, PALLICKARA S. Analytic queries over geospatial time-series data using distributed hash tables[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(6): 1408-1422.

[16] HAO Y, CAO H, MUEEN A, et al. Identify significant phenomenon-specific variables for multivariate time series[J]. IEEE Transactions on Knowledge and Data Engineering, 2019: 10.1109/TKDE.2019.2934464.

[17] CHEN D, TANG Y, ZHANG H, et al. Incremental factorization of big time series data with blind factor approximation[J]. IEEE Transactions on Knowledge and Data Engineering, 2019: 10.1109/TKDE.2019.2931687.

[18] CHU X, ILYAS I F, KRISHNAN S, et al. Data cleaning: overview and emerging challenges[C]//The 2016 International Conference on Management of Data, June 26-July 1, 2016, San Francisco, USA. New York: ACM Press, 2016: 2201-2206.

[19] 李杰, 倪军, 王安正. 从大数据到智能制造[M]. 上海: 上海交通大学出版社, 2017.
LI J, NI J, WANG A Z. From big data to intelligent manufacturing[M]. Shanghai: Shanghai Jiao Tong University Press, 2017.

[20] LIU Y, LI Z, ZHOU C, et al. Generative adversarial active learning for unsupervised outlier detection[J]. IEEE Transactions on Knowledge and Data Engineering, 2019 Accepted.

[21] HU W, GAO J, LI B, et al. Anomaly detection using local kernel density estimation and context-based regression[J]. IEEE Transactions on Knowledge and Data Engineering, 2018: 10.1109/TKDE.2018.2882404.

[22] SHARMA V, KUMAR R, CHENG W, et al. NHAD: neuro-fuzzy based horizontal anomaly detection in online social networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(11): 2171-2184.

[23] LU Y, CHEN F, WANG Y, et al. Discovering anomalies on mixed-type data using a generalized student-t based approach[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(10): 2582-2595.

[24] LIN X, PENG Y, CHOI B, et al. Human-powered data cleaning for probabilistic

reachability queries on uncertain graphs[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(7): 1452-1465.

[25] HAO S, TANG N, LI G, et al. A novel cost-based model for data repairing[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(4): 727-742.

[26] DASU T, LOH J M. Statistical distortion: consequences of data cleaning[J]. Proceedings of the VLDB Endowment, 2012, 5(11): 1674-1683.

[27] BOHANNON P, FAN W, FLASTER M, et al. A cost-based model and effective heuristic for repairing constraints by value modification[C]// The 2005 ACM SIGMOD International Conference on Management of Data, June 14-16, 2005, Baltimore, Maryland. New York: ACM Press, 2005: 143-154.

[28] SONG S, ZHU H, WANG J. Constraint-variance tolerant data repairing[C]// The 2016 International Conference on Management of Data, June 26-July 1, 2016, San Francisco, USA. New York: ACM Press, 2016: 877-892.

[29] LI Z, WANG H, SHAO W, et al. Repairing data through regular expressions[J]. Proceedings of the VLDB Endowment, 2016, 9(5): 432-443.

[30] KHAYYAT Z, ILYAS I F, JINDAL A, et al. Bigdansing: a system for big data cleansing[C]// The 2015 ACM SIGMOD International Conference on Management of Data, May 31-June 4, 2015, Melbourne, Australia. New York: ACM Press, 2015: 1215-1230.

[31] JENSEN S K, PEDERSEN T B, THOMSEN C. Time series management systems: a survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(11): 2581-2600.

[32] 苏卫星, 朱云龙, 刘芳, 等. 时间序列异常点及突变点的检测算法[J]. 计算机研究与发展, 2014, 51(4): 781-788.

SU W X, ZHU Y L, LIU F, et al. Outliers and change-points detection algorithm for time series[J]. Journal of Computer Research and Development, 2014, 51(4): 781-788.

[33] SALEHI M, LECKIE C, BEZDEK J C, et al. Fast memory efficient local outlier detection in data streams[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(12): 3246-3260.

[34] CAO L, YANG D, WANG Q, et al. Scalable distance-based outlier detection over high-volume data streams[C]// 2014 IEEE 30th International Conference on Data Engineering, March 31-April 4, 2014, Chicago, USA. Piscataway: IEEE Press, 2014: 76-87.

[35] YANG F, SONG H A, LIU Z, et al. Ares: automatic disaggregation of historical data[C]// 2018 IEEE 34th International Conference on Data Engineering (ICDE), April 16-19, 2018, Paris, France. Piscataway: IEEE Press, 2018: 65-76.

[36] AROUS I, KHAYATI M, CUDRÉ-MAUROUX P, et al. RecovDB: accurate and efficient missing blocks recovery for large time series[C]// 2019 IEEE 35th International Conference on Data Engineering (ICDE), April 8-11, 2019, Macao, China. Piscataway: IEEE Press, 2019: 1976-1979.

[37] WU S, WANG L, WU T, et al. Hankel matrix factorization for tagged time series to recover missing values during blackouts[C]// 2019 IEEE 35th International Conference on Data Engineering (ICDE), April 8-11, 2019, Macao, China. Piscataway: IEEE Press, 2019: 1654-1657.

[38] FENG K, GUO T, CONG G, et al. SURGE: continuous detection of bursty regions over a stream of spatial objects[C]// 2018 IEEE 34th International Conference on Data Engineering (ICDE), April 16-19, 2018, Paris, France. Piscataway: IEEE Press, 2018: 1292-1295.

[39] MA M, ZHANG S, PEI D, et al. Robust and rapid adaption for concept drift in software system anomaly detection[C]//2018 IEEE 29th International Symposium on Software Reliability Engineering (ISSRE), October 15-18, 2018, Memphis, USA. Piscataway: IEEE Press, 2018: 13-24

[40] MEI J, DE CASTRO Y, GOUDE Y, et al. Nonnegative matrix factorization with side information for time series recovery and prediction[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(3): 493-506.

[41] RONG K, BAILIS P. ASAP: prioritizing attention via time series smoothing[J]. Proceedings of the VLDB Endowment, 2017, 10(11): 1358-1369.

[42] YOON S, LEE J G, LEE B S. NETS: extremely fast outlier detection from a data stream via set-based processing[J]. Proceedings of the VLDB Endowment, 2019, 12(11): 1303-1315.

[43] SONG S, ZHAO A, WANG J, et al. SCREEN: stream data cleaning under speed constraints[C]// ACM SIGMOD International Conference on Management of Data, May 31-June 4, 2015, Amsterdam, The Netherlands. New York: ACM Press, 2015.

[44] ZHANG A, SONG S, WANG J. Sequential data cleaning: a statistical approach[C]// The 2016 International Conference on Management of Data, June 26-July 1, 2016, San Francisco, USA. New York: ACM Press, 2016: 909-924.

[45] YIN W, YUE T, WANG H, et al. Time series cleaning under variance constraints[C]// International Conference on Database Systems for Advanced Applications, May 21-24, 2018, Gold Coast, Australia. Heidelberg: Springer, 2018.

[46] SADIK S, GRUENWALD L, LEAL E. Wadjet: finding outliers in multiple multi-dimensional heterogeneous data streams[C]// 2018 IEEE 34th International Conference on Data Engineering (ICDE), April 16-19, 2018, Paris, France. Piscataway: IEEE Press, 2018: 1232-1235.

[47] SONG S, CAO Y, WANG J. Cleaning timestamps with temporal constraints[J]. Proceedings of the VLDB Endowment, 2016, 9(10): 708-719.

[48] ABEDJAN Z, AKCORA C G, OUZZANI M, et al. Temporal rules discovery for web data cleaning[J]. Proceedings of the VLDB Endowment, 2015, 9(4): 336-347.

[49] 陈乾, 胡谷雨, 路威. 基于距离和DF-RLS的时间序列异常检测[J]. 计算机工程, 2012, 38(12): 32-35.
CHEN Q, HU G Y, LU W. Outlier detection for time series based on distance and DF-RLS[J]. Computer Engineering, 2012, 38(12): 32-35.

[50] MILANI M, ZHENG Z, CHIANG F. Current clean: spatio-temporal cleaning of stale data[C]// 2019 IEEE 35th International Conference on Data Engineering (ICDE), April 8-11, 2019, Macao, China. Piscataway: IEEE Press, 2019: 172-183.

[51] ZAMENI M, GHAFOORI Z, SADRI A, et al. Change point detection for streaming high-dimensional time series[C]// The 24th International Conference on Database Systems for Advanced Applications, April 22-25, Chiang Mai, Thailand. Heidelberg: Springer, 2019.

[52] SOUIDEN I, BRAHMI Z, LAFI L. Data stream mining based-outlier prediction for cloud computing[C]// The 33rd IEEE International Conference on Data Engineering, April 19-22, 2017, San Diego, USA. Piscataway: IEEE Press, 2017.

[53] HAQUE A, KHAN L, BARON M, et al. Efficient handling of concept drift and concept evolution over Stream Data[C]// 2016 IEEE 32nd International Conference on Data Engineering (ICDE), May 16-20, 2016, Helsinki, Finland. Piscataway: IEEE Press, 2016: 481-492.

[54] XU H, CHEN W, ZHAO N, et al. Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in Web applications[C]// The 2018 Web Conference, April 23-27, 2018, Lyon, France. [S.l.:s.n.], 2018.
[55] MALHOTRA P, RAMAKRISHNAN A, ANAND G, et al. LSTM-based encoder-decoder for multi-sensor anomaly detection[J]. Computer Science, 2016, arXiv: 1607.00148.
[56] EICHMANN P, SOLLEZA F, TATBUL N, et al. Visual exploration of time series anomalies with metro-viz[C]// The 2019 International Conference on Management of Data, June 30-July 5, 2019, Amsterdam, Netherlands. New York: ACM Press, 2019: 1901-1904.

作者简介

丁小欧(1993-),女,哈尔滨工业大学海量数据计算研究中心博士生,主要研究方向为时序数据挖掘与分析、数据清洗、数据质量管理等。

王宏志(1978-),男,博士,哈尔滨工业大学海量数据计算研究中心教授、博士生导师,主要研究方向为数据库管理系统、大数据管理与分析、数据治理等。

于晟健(1997-),男,哈尔滨工业大学海量数据计算研究中心硕士生,主要研究方向为时序数据分析、异常检测、时序数据清洗等。

收稿日期: 2019-09-05
通信作者: 王宏志, wangzh@hit.edu.cn
基金项目: 国家重点研发计划基金资助项目(No.2018YFB1004700);国家自然科学基金资助项目(No.U1509216, No.U1866602, No.61602129)
Foundation Items: The National Key Research and Development Program of China(No.2018YFB1004700), The National Natural Science Foundation of China(No.U1509216, No.U1866602, No.61602129)