文章编号：2095-2163(2021)12-0001-07 中图分类号：TP311.13 文献标志码：A

时序数据错误检测与修复研究综述

丁小欧，王宏志，靳贺霖，高 猛
（哈尔滨工业大学 计算机学科与技术学院，哈尔滨 150001）

摘 要：随着数据应用对数据质量要求的提高，时序数据清洗问题得到研究领域和应用领域的更多关注。在数据清洗过程中，错误数据检测和修复是两个关键步骤。近年来，研究人员在这两方面进行了诸多研究与探索。本文从时序数据质量问题、错误数据检测，以及错误数据修复步骤对当前研究技术等进展情况进行介绍，分析了当前时序数据清洗所存在的难点与不足。

关键词：数据清洗；时序数据挖掘；异常检测；数据质量管理

Error data detection and repairing in temporal data：a survey

DING Xiaoou，WANG Hongzhi，JIN Helin，GAO Meng
（School of Computer Science and Technology，Harbin Institute of Technology，Harbin 150001，China）

【Abstract】With the increasing quality requirement in data，time series data cleaning has attracted more attention in both research and application fields. Error data detection and repairing are two key steps in data cleaning process，where researchers have made a lot of research and exploration. This paper introduces the progress of current research techniques from the aspects of time series data quality problems，error data detection and error data repair steps，and reviews the difficulties and deficiencies of the current time series data cleaning.

【Key words】data cleaning；temporal data mining；anomaly detection；data quality management

0 引 言

目前，数据正以前所未有的速度不断产生。随着各种各样数据采集设备的普及使用，海量的时序数据日以继夜的被积累和使用。带有时间标记的时间序列数据，能够帮助分析人员对历史记录进行有效建模，通过整合相应时间段的有关资料，挖掘提炼有价值的信息。但是，实际的时序数据中广泛存在数据质量问题。低质量的时态数据，不仅导致数据预处理环节需消耗大量人力物力成本，而且也会导致在实际应用数据挖掘与知识提取分析结果发生偏差和错误。随着数据应用对数据质量要求的提高，时序数据的质量问题也逐渐引起研究领域和应用领域的更多关注，而数据清洗技术是提高时序数据质量的有效方法。

本文聚焦时序数据清洗研究中错误数据的检测与修复两个主要环节，对时序数据质量管理研究进展及研究现状进行评析；介绍了通用数据清洗流程；分别介绍时序数据中的错误数据检测、以及错误数据修复技术研究进展，并提出未来研究展望。

1 时序数据质量研究评析

2002 年，在综述文献[1]中，国内首个对数据质量和数据清洗进行了全面分析介绍工作。随后，在大数据高速发展的时期，数据可用性问题得到广泛关注。围绕着数据一致性、精确性、完整性、时效性、实体同一性这 5 个核心问题，国内研究人员做了大量工作。2016 年，综述文献[2]对大数据可用性进行了深刻剖析，从表达机理、判定理论、演化原理等方面进行详细介绍，并提出了量质融合管理、劣质容忍计算、深度演化机理等多个具有挑战性的问题。结合数据质量研究中数据劣质的情况，已有多篇综述[3-5]对数据错误类型进行了详细分析。文献[6]对大数据质量从数据质量评估、数据清洗、质量管理等多角度进行了详细介绍；文献[7]提出了“数据工程”的概念，介绍了以数据为主题而展开的各项数据处理技术。

而随着中国现代化进程的发展，各行业领域已

基金项目：国家电网有限公司科技项目(5700-202119176A-0-0-00)。

作者简介：丁小欧(1993-)，女，博士研究生，主要研究方向：数据清洗、数据质量管理、时序数据挖掘等；王宏志(1978-)，男，博士，教授，博士生导师，主要研究方向：大数据管理、大数据分析、智能化数据管理等；靳贺霖(1998-)，男，硕士研究生，主要研究方向：异常检测、时序数据挖掘、机器学习等；高 猛(1998-)，男，硕士研究生，主要研究方向：主动学习、时序数据挖掘、时序数据异常检测等。

收稿日期：2021-04-06

经积累并正在产生大量的时间序列数据,例如,在制造领域,现代化的传感器设备、智能仪表[8]等,能够实现对生产环境的感知,并对设备的生产状态进行实时记录。这些数据中隐含丰富的信息和知识,能够为系统运行状态的控制、分析、决策以及规划提供重要的参考依据。然而,时序数据中的数据质量问题也非常普遍,重复记录、异常记录、无效数据、时标不齐等质量问题的存在,限制了对领域数据的深入分析[9]。低质量的工业时间序列数据的清洗,正逐渐成为研究热点、重点和难点。近 5 年来,时间序列清洗的研究呈快速增长趋势。异常检测、序列分类、序列模式挖掘等问题,在理论上和实践中均得到广泛研究。其中,异常和错误数据的识别与检测是数据清洗和故障检测研究领域的重点难点问题。

虽然研究人员在数据清洗上进行了长期的探索,取得了一定的理论成果,并投入应用进行实践。但在时序数据清洗问题上,仍有许多关键问题丞待解决,其主要研究问题包括:

(1)时序数据在数据库系统中存储、拷贝、转移,或者在信息系统中跨数据库调用时,其时间标记(时间戳)可能发生丢失、不可用、时间不统一、时序错乱、不对齐等问题,导致数据的部分时序信息丢失[9]。

(2)数据的多样性和复杂性,导致数据中的错误模式增多。“脏数据”产生缘由复杂,低质量数据中通常存在多种的错误类型, 其中的关联关系仍然缺乏理论认识[10-11],且对于多种错误共存的数据清洗、检测和修复“脏数据”的计算复杂性会大幅增加[2]。

而目前时序数据清洗研究的不足之处主要包括:

(1)时间序列数据与行业联系紧密,大量的时序数据通常体现较为明显的领域特征[12],难以建立通用的理论计算框架。以工业场景为例,时间序列数据的来源广泛,具有大体量、多源性、连续采样、价值密度低、动态性强的特点[13]。不同的时间序列模式难以区分。时间序列的值异常情况难以描述和评价,容易引起误判、漏判。

(2)数据之间相关性考虑不够充分。在传感器采集的多维时序数据中,属性之间具有一定的相关性。多维时序数据中的错误模式常有隐匿性强、成因复杂的特点。因此,若仅对每条序列进行独立的异常检测,会导致无法准确地识别出一些真正的错误问题。序列之间的相关关系和依赖关系有多种类型[14],这些还未被充分地研究。

(3)领域知识尚未得到充分利用。对数据质量问题发生原因和影响关系的挖掘与分析,离不开领域知识的参与。虽然目前知识驱动的异常数据检测技术,如专家系统(expert system ES)[15]、动态贝叶斯网络模型(Dynamic Bayesian Network DBN)[16]、知识库与知识图谱[17]等已开始被应用于异常诊断和数据清洗问题中,但基于知识的方法往往仅适用于具体的问题,模型方法的迁移性较低。除此之外,领域知识的不完整性和模糊性也加大了知识建模的难度。

(4)方法缺少可扩展性。对数据质量问题的定位、产生原因、模式之间关联性的计算分析过程,通常需要大量的计算,且高维度和大规模数据会导致模型的计算代价较高[18-19];另一方面,知识模型的局限性也限制了方法的可扩展能力。数据量之大、数据更新之快,使得方法难以实现在可接受的时间范围内,有效地完成数据质量管理任务。

2 数据清洗典型流程介绍

数据清洗(或称数据净化、数据清洁等)是指,对数据集合中存在的不符合规范数据的检测,并进行数据修复,提高数据质量的过程[6,20]。错误数据(或称脏数据),均造成了对给定数据质量评价的违反结果[4,21]。而错误数据的取值,常被认为与真值存在较大距离。数据清洗通常作为数据预处理环节的必要步骤,是公认的修复数据错误、提高数据质量必要的有效手段[6,21-22]。

近十几年来,已有多篇综述文章及专著(文献[21-25])从不同的角度和场景需求,对数据清洗技术进行总结和介绍。文献[21]中总结了传统数据清洗的典型流程,主要包括错误数据检测和错误数据修复两个核心步骤。此外,数据清洗规则挖掘与提取也是数据清洗流程中可选的重要步骤。传统数据清洗流程如图 1 所示。

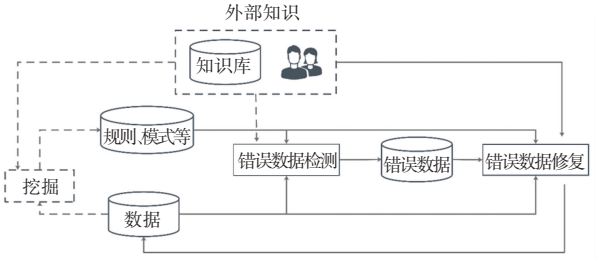

图 1 传统数据清洗流程[21]

Fig. 1 A typical data cleaning workflow[21]

文献[26]从定量和定性两方面，总结了常见的数据错误类型。其中，定性错误数据（Qualitative error）包括重复记录、规则违反、模式违反 3 个类型；定量错误（Quantitative error）主要包括异常值、离群点等类型。

3　时序数据中的错误数据检测

2019 年，综述文章[25]中介绍了时间序列数据上的错误类型，并归纳总结了当前常用的时间序列数据清洗方法。相比于传统关系型数据清洗研究对“数据修复”的侧重，目前的时序数据清洗主要以错误数据检测、异常值、离群点检测为主。

错误数据检测是任何数据清洗任务中首要且关键的一步，只有对“脏数据”检测准确，才能实现有效的修复。文献[21,24]中提出，在数据清洗的检测阶段，关键在于检测什么（What to detect）、怎样检测（How to detect）、何时检测（When to detect）。

在定性错误检测技术研究中，目前形成了以完整性约束作为主要检测标准的方法体系。例如，针对 FDs 的违反检测[27]、基于否定约束的检测方法[28]等。在检测方式上，主要分为算法自动化检测和人工参与检测。典型的自动化检测方法研究的文献有[27]、Holoclean 方法[28,29]、DBRx 方法[22]；人工参与的检测有基于众包方法，例如，CrowdER[30]、Corleone[31]，以及错误数据的解释方法 Scorpion[32]、视图条件因果溯源（View-conditioned causality tracing）[33]等等。

3.1　错误数据类型

对数据进行错误检测阶段，通常采用异常检测技术，目标是发现数据中不寻常的、未知的数据值或模式。在异常检测领域研究中，数据中的异常主要表现为 3 类：点异常（point outliers）、（上下文）内容异常（contextual outliers）、聚集型异常（collective outliers）。文献[34]中将时间序列上的异常分为点异常和结构异常两种。点异常有加性异常点（additive outlier AO）、创新性异常点（innovational outlier IO）两种表现形式。结构异常有水平偏移（level shift LS）、方差变化（variance change VC）两种类型。其中，水平偏移可进一步分为永久性水平变化和瞬时水平变化。文献[25]将时间序列数据错误类型分为 4 种：单点小错误（single small error）、单点大错误（single large error）、连续性错误（continuous errors）和传输偏移错误（translational error）。

由于点异常通常是以较低频率出现在数据中，离散的异常点出现具有偶然性。在实际的数据场景中，异常的表现形式往往复杂多变，这就需要对数据点的“长度”和“宽度”加以分析。

3.2　异常模式与异常数据段检测

2002 年，文献[35]较早地发现，简单的点异常检测方法，难以满足复杂的实际需求，提出了异常模式（anomalous patterns）的概念，并提出了一个基于规则的异常检测算法。将每一种异常模式总结为一条规则，每条规则由若干个组件取逻辑“AND”组成。对于每个组件所包含的每个特征值，比较当前日期的记录与历史记录的差异，并利用统计学方法，对规则上的取值进行评分分析。由于计算限制，文献[35]中使用的每条规则最多只包含 2 个组件。在近十几年的研究中，时序数据上的序列模式挖掘问题（sequence patterns discovery）一直被研究人员关注[36-37]。

2005 年，文献[36]提出 HOT SAX（Symbolic Aggregate approXimation）这一经典的不寻常子序列挖掘方法；文献[37]提出了基于 K-means 模型的阶段性时序数据的异常检测方法。在异常模式检测的研究中，通常以无监督学习模型为主，而由于不同领域的时序数据模式复杂、多样，异常序列片段检测问题仍然面临许多瓶颈性的挑战。

3.3　多维时序数据异常检测

随着单维时间序列数据（Univariate time series UTS）异常检测研究的深入展开，近年来研究人员也将更多的关注点放在多维时间序列数据上（Multivariate Time Series MTS）。相比于单维时序数据，变量（即维数）的增多使得异常模式样式也随之增多，异常数据模式更加多样、复杂，且识别难度也更高。因此，不仅要将单维序列的异常检测方法包含在内，还需进一步考虑多个变量之间的相关性等问题。研究 MTS 的异常检测方法时，通常面临以下问题[38]：一是多维数据的属性可能是异构、多样的；二是不同属性对判断异常（事件）的“贡献性”程度不同；其三，异常事件实例的数量往往非常有限。

面对上述难题，在进行 MTS 异常检测时，通常采用统计指标与机器学习模型相结合[37]的方法。有部分文献利用分类问题模型解决多维时序数据异常检测，例如决策树、支持向量机[39]、神经网络等。这些分类方法通常需要一定规模的异常实例作为训练数据，对训练样本的需求成为朴素分类模型解决异常检测的一个局限因素。文献[38]研究了异构

时间序列上的异常模式挖掘方法,将数据中异构的属性转化成同构的异常评分序列,并用异常评分向量来表示每个故障事件;然后迭代地优化异常模式的特征参数。优化原则主要包括:最大化分离异常模式和正常模式的距离、最大化分离疑似异常实则正常的模式和真实异常模式的距离,以及最小化异常评分向量的平方损失函数。在测试阶段,计算异常评分参数与已训练的异常模式的匹配度,实现对故障的(分类)诊断。

针对多维数据检测的效率问题,文献[40]提出了基于 kNN 方法的多维异常检测算法,并利用异常评分机制对观测点及其邻居数据的"分离度"进行量化,并采用低阶近似方法(low - rank approximation)将高维数据映射到低维空间,实现算法效率上的提升。文献[41]提出了一种基于分支界限(Branch and Border)的非参数的异常检测算法,其能在线性时间内实现对数值型、分类型等混合类型时序数据的检测,并在 11 个时序数据集上进行了实验,验证了该方法具有较强的有效性和可扩展性。

在实际问题中,异常数据和正常数据难以被简单地区分开[42]。目前的异常检测方法通常具有一定的专门性,在通用性和可迁移性的技术突破存在较大难度。虽然近些年异常检测问题在理论和实际应用中得到了大量的研究,但许多关键问题仍未完全解决。相比于传统关系型数据,时序数据属性多以数值型为主。因此,时序数据上具有更多、更复杂的错误模式。此外,时序数据具有时间属性,使得许多错误模式具有一定的聚集和累积效应,这给序列建模和异常识别均带来不少的难度。

4 时序数据中的错误数据修复

相比于检测步骤对错误数据"识别"效果的侧重,修复步骤的关键之处在于如何对错误数据进行合理、有效的修复。类似于错误数据检测步骤,错误数据的修复研究同样聚焦于修复目标(What)、修复形式(How)和修复模型(Where)上。在修复目标上,主要有数据驱动、基于规则、以及基于数据和基于规则的融合方法[24]。

4.1 基于统计的错误数据修复

相比于时序数据的错误检测,时序数据的清洗研究还尚不完备。文献[43]中介绍了大数据上定量数据错误的清洗方法,将时间序列数据的清洗作为一个特殊场景进行讨论。结合时态数据的清洗理论基础和实产经验,目前已有的方法主要以基于统计模型和基于约束为主。平滑噪声(smoothing - based cleaning)是一种常见的数据修复方法。通过调整滑动窗口,计算给定指标,实现对异常部位的平滑修复。常见方法有:滑动平均[44](simple moving average , SMA)利用当前时间点前 K 个数据点上的取值,做不加权计算;指数加权平均[45](exponentially weighted moving average,EWMA)是对简单滑动平均的改进。由于距离近的时间点取值的相关性更强,因此对序列进行两次加权取值计算。平滑的方法虽然简单常用,但其准确率较低。不同类型时间序列数据的情况复杂,仅仅用上下文相关性进行平滑,无法对持续的异常片段进行准确修复。基于 ARIMA 模型也是时间序列上的基础方法,包括了自回归过程(autoregressive)、滑动平均过程(moving-average)分析等。

4.2 基于约束的错误数据修复

基于约束的清洗技术不仅广泛应用于传统关系型数据,而且同样适用于时间序列的数据清洗问题。但是,相比于传统关系型数据,针对时间序列数据的约束种类不多。相比于传统关系型数据,时态数据通常具有发展规律、变化趋势、自相关性、季节性等特点。因此,在时态数据上的约束设计也需要考虑更多因素。文献[14]聚焦时态数据质量管理,将用于识别错误数据的约束分为 4 类(见表 1),并提出了时态数据质量管理的通用方法。首先,归纳定义出高质量数据需满足的规则约束条件,将时刻 t 上高质量数据记为 D_t^l;然后,将测试数据 D_t 与高质量数据 D_t^l 进行比较。可利用文献[46]提出的统计失真(statistical distortion)指标,度量测试数据 D_t 和理想数据 D_t^l 之间的差异,即 $SD(D_t) = Dist(D_t, D_t^l)$,找出违反约束的数据。经约束初步检测出的违反数据(violation)也被称为小错误(glitch)[47]。需要注意的是,利用约束检测出来的小错误并不一定是真正的错误,可能是约束参数值设置过于严格所导致的误判。因此,对于违反数据通常需进一步分析、解释后,再决定如何对其进行处理。

表 1 约束种类

Tab. 1 Types of constraints

类型	单列	多列
单行	类型 1:单个实体的单个属性	类型 1:单个实体的多个属性
多行	类型 3:多个实体的单个属性	类型 4:多个实体的多个属性

相比于传统关系型数据,针对时间序列数据的约束种类不多。其中,文献[48]提出了顺序依赖(Order Dependencies ODs),ODs 同样适用于时间序

列的清洗。在顺序依赖的基础上，文献[49]针对序列数据，提出了序列依赖（Sequential dependencies SDs），表示为 $X \rightarrow \rightarrow_g Y$。描述了在以属性 X 排序时，连续两条记录上属性 Y 的取值变化范围为 g。通常，X 是数据库模式 R 上的一个有序属性，记录在属性 Y 上的取值是可比较的。序列依赖是目前可用于时间序列清洗问题仅有的约束之一。

由于序列依赖主要针对有序数据集而设计，未对数据的时间戳属性给予充分考虑。SDs 乃至 CSDs 对于相邻记录之间某属性的取值要求较为理想化，难以满足实际中的数据清洗需求。张奥千[52]等人在序列依赖的基础上，提出了速度约束。文献[50-51]还提出了基于速度约束的时间序列清洗问题，并介绍了基于最小修复原则的数据清洗方法。速度约束在识别并修复时间序列数据中的大错误效果更好，而在对数据上的小错误以及连续性错误的识别上具有局限性。文献[52]提出了基于方差约束的修复方法，以最小修复为原则，找到满足给定方差约束的修正序列。

考虑到修复方法，通常有两类策略：一是完全信任给定的约束集，只对数据进行修改和修复；二是不完全信任已有的约束（参数取值），对数据和约束都会进行修正。对于前者，大部分的修复方法通常每一次只针对同一类错误问题进行修复。一般采用最小修复原则[51]，对原始数据 D 修改为 D'，并使 $Cost(D,D')$ 最小。这里的 $Cost$ 通常是距离函数或具有距离函数类似性质的代价函数。近年来，Ihab F. Ilyas 等人提出了全面数据清洗（Holistic data cleaning）方法[28-29]。文献[53]认为在现实问题中，数据会随时间发生改变，导致既定的约束变得不再准确。因此，提出了约束 & 数据的联合修复模型。运用函数依赖，同时考虑利用约束或者数据本身对违反部分进行修改；并且设计算法实现对约束的更新和修改。由于约束集不总是完备且准确的，文献[54]研究了允许约束小范围变化的数据修复方法，提出一个 θ-容忍的修复模型。在修复时，允许约束谓词在 θ 范围内的增加和删除。

4.3　人机协同的数据修复

数据清洗问题的特点，导致所采用的清洗指标通常是与场景高度相关的，需要用户、领域知识、问题背景等多方面因素的共同作用，才能实现高质量的数据修复[46]。高质量的数据清洗任务离不开人工的参与。因此，近年来研究人员也在数据清洗研究上展开了更多创新性研究，各类人工参与的修复方法逐渐丰富，Yakout 等人较早引入机器学习方法提升数据清洗的可靠性，并提出 GDR（guided data repair）模型[55]；综述文献[23]介绍了该方法主要包括基于 CFD 的脏数据识别、更新策略生成、更新策略排序、基于用于反馈的训练模型步骤。典型的人工参与的数据清洗系统还有 KATARA[17]、Data Tamer[57]等。

在大数据的清洗过程中，有时可靠的数据清洗规则不能轻易获取。因此，研究人员展开了直接在劣质数据上进行学习和分析的方法。文献[58]提出了一种从劣质数据中进行学习的方法 Dirty Learn（DLearn）。大数据技术的普及和发展，使得许多大数据管理与分析问题变得更为领域化、专业化，这对数据的领域知识背景提出了更多的要求。文献[59]介绍了人在回路的数据准备技术研究（human-in-the-loop data preparation）进展。近年来，人在回路数据清洗技术的关键点主要在于人机交互和众包清洗策略（crowdsourcing data cleaning）两方面。在人机交互方面，文献[56]提出了 FALCON 系统，依靠与用户的交互实现对数据修复。在众包方法的研究上，文献[17]提出了基于知识库和众包的数据清洗系统——KATARA，旨在弥补既定完整性约束和模型驱动的数据清洗在效果上的局限性，通过访问主数据或询问领域专家，来解决数据中的歧义问题。

在以时序数据为主要数据类型的清洗研究中，错误数据检测的研究要比错误数据修复的研究更加全面。而在目前的修复方法中，依赖于传统的统计模型方法居多。但是，简单修复结果的精准度难以满足实际要求。由于以 OD、SD、SC 为例的约束模型才被提出不久，时序数据修复的理论方法仍需完善。此外，虽然目前已有人工参与的清洗方法被使用，但人机结合的清洗研究中仍有许多关键问题未被解决。

5　结束语

本文介绍了目前时序数据质量管理以及时序数据清洗的研究进展，针对错误数据检测以及错误数据修复技术进行现状研究分析。在影响时序数据清洗研究的因素中，既有技术层面，又有管理层面，而对数据质量的评估中，既有主观维度，又有客观维度。综合目前的实际需求，时序数据清洗的未来研究方向包括：

（1）继续深入考虑数据、错误模式之间相关性计算问题；

(2)研究多源、复合型的数据质量管理及数据清洗方法体系;

(3)结合机器学习模型研究智能化的错误数据检测方法,实现对未知、复杂模式的错误数据识别;

(4)继续探索人工参与的数据清洗方法,提高错误数据修复的可靠性、准确性;

(5)结合知识工程相关技术,实现对领域知识的有效建模,实现知识融合的数据清洗模型,提高数据清洗结果的可解释性和有效性。

参考文献

[1] 郭志懋, 周傲英. 数据质量和数据清洗研究综述[J]. 软件学报, 2002, 13(11):2076-2082.

[2] 李建中, 王宏志, 高宏. 大数据可用性的研究进展[J]. 软件学报, 2016, 27(7):1605-1625.

[3] ILYAS I F, CHU X. Trends in Cleaning Relational Data: Consistency and Deduplication[J]. Foundations and Trends R in Databases, 2012,5(4):281-393.

[4] KIM W, CHOI B J, HONG E K, et al. A Taxonomy of Dirty Data[J]. Data Mining & Knowledge Discovery, 2003, 7(1):81-99.

[5] RAHM E. Data cleaning : Problems and current approaches[J]. IEEE Data Eng Bull, 2000, 23(23):3-13.

[6] 蔡莉,朱扬勇. 大数据质量[M]. 上海:上海科学技术出版社, 2017:121-127, 167-170.

[7] 岳昆. 数据工程——处理、分析与服务[M]. 北京:清华大学出版社,2013:169-180.

[8] 王建民. 工业大数据技术综述[J]. 大数据, 2017(6):3-14.

[9] 王晨,郭朝晖,王建民. 工业大数据及其技术挑战[J]. 电信网技术, 2017(8):1-4.

[10] FAN W, GEERTS F. Foundations of Data Quality Management [J]. 2012, 4(5):217.

[11] 丁小欧, 王宏志, 张笑影,等. 数据质量多种性质的关联关系研究[J]. 软件学报, 2016, 27(7):1626-1644.

[12] 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,2016:41-42.

[13] 工业互联网产业联盟. 中国工业大数据技术与应用白皮书. 北京, 2017,7.

[14] DASU T, DUAN R, SRIVASTAVA D. Data Quality for Temporal Streams[J]. IEEE Data Eng. Bull., 2016, 39(2): 78-92.

[15] CHIANG L H , RUSSELL E L , BRAATZ R D . Fault Detection and Diagnosis in Industrial Systems[M]. Springer London, 2001: 228-232.

[16] FUJIMAKI R, NAKATA T, TSUKAHARA H, et al. Mining abnormal patterns from heterogeneous time-series with irrelevant features for fault event detection[J]. Statistical Analysis and Data Mining: The ASA Data Science Journal, 2009, 2(1): 1-17.

[17] CHU X,MORCOS J, ILYAS I F, et al. KATARA: Reliable data cleaning with knowledge bases and crowdsourcing[J]. Proceedings of the VLDB Endowment, 2015, 8(12): 1952-1955.

[18] 丁小欧,王宏志,于晟健. 工业时序大数据质量管理[J]. 大数据,2019,5(6):1-11.

[19] GUPTA M, GAO J, AGGARWAL C, et al. Outlier detection for temporal data[M]. Morgan & Claypool Publishers, 2014: 29-32.

[20] MÜLLER H, FREYTAG J C. Problems, methods, and challenges in comprehensive data cleansing[M]. Professoren des Inst. Für Informatik, 2005.

[21] ILYAS I F, CHU X. Data cleaning[M]. ACM 2019, ISBN 978-1-4503-7152-1. 1-4.

[22] CHALAMALLA A, ILYAS I F, OUZZANI M, et al. Descriptive and prescriptive data cleaning[C]// ACM SIGMOD International Conference on Management of Data. ACM, 2014:445-456.

[23] 郝爽, 李国良, 冯建华,等. 结构化数据清洗技术综述[J]. 清华大学学报(自然科学版), 2018, 58(12):1037-1050.

[24] CHU X, ILYAS I F, KRISHNAN S, et al. Data cleaning: Overview and emerging challenges[C]//Proceedings of the 2016 international conference on management of data, 2016: 2201-2206.

[25] WANG X, WANG C. Time series data cleaning: A survey[J]. Ieee Access, 2019, 8: 1866-1881.

[26] ABEDJAN Z, CHU X, DENG D, et al. Detecting data errors: where are we and what needs to be done? [J]. Proceedings of the Vldb Endowment, 2016, 9(12):993-1004.

[27] Philip Bohannon, Michael Flaster, Wenfei Fan, et al. A Cost-Based Model and Effective Heuristic for Repairing Constraints by Value Modification[C]//SIGMOD Conference, 2005: 143-154.

[28] CHU X, ILYAS I F, PAPOTTI P. Holistic data cleaning. Putting violations into context[C]// IEEE International Conference on Data Engineering. IEEE Computer Society, 2013:458-469.

[29] Theodoros Rekatsinas, Xu Chu, Ihab F. Ilyas, Christopher Ré. HoloClean: Holistic Data Repairs with Probabilistic Inference[J]. Proc. VLDB Endow, 2017, 10(11): 1190-1201.

[30] WANG Jiannan, Tim Kraska, Michael J. Franklin, Jianhua Feng. CrowdER: Crowdsourcing Entity Resolution[J]. Proc. VLDB Endow, 2012, 5(11): 1483-1494.

[31] Chaitanya Gokhale, Sanjib Das, AnHai Doan, Jeffrey F. Naughton, Narasimhan Rampalli, Jude W. Shavlik, Xiaojin Zhu. Corleone: hands-off crowdsourcing for entity matching[C]// SIGMOD Conference, 2014: 601-612.

[32] WU Eugene, Samuel Madden: Scorpion. Explaining Away Outliers in Aggregate Queries[J]. Proc. VLDB Endow, 2013, 6(8): 553-564.

[33] Alexandra Meliou, Wolfgang Gatterbauer, Suman Nath, Dan Suciu. Tracing data errors with view-conditioned causality[C]// SIGMOD Conference, 2011: 505-516.

[34] TSAY R S. Outliers, level shifts, and variance changes in time series[J]. Journal of Forecasting, 2010, 7(1):1-20.

[35] WONG Weng-Keen, MOORE Andrew W, COOPER Gregory F, et al. Rule-Based Anomaly Pattern Detection for Detecting Disease Outbreaks[C]// AAAI/IAAI, 2002: 217-223.

[36] Eamonn J. Keogh, Jessica Lin, Ada Wai-Chee Fu. HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence [C]// ICDM, 2005: 226-233.

[37] Umaa Rebbapragada, Pavlos Protopapas, Carla E. Brodley, Charles R. Alcock. Finding anomalous periodic time series[J]. Mach. Learn, 2009, 74(3): 281-313 .

[38] FUJIMAKI R, NAKATA T, TSUKAHARA H, et al. Mining abnormal patterns from heterogeneous time-series with irrelevant features for fault event detection[J]. Statistical Analysis and Data Mining, 2009, 2(1):1-17.

(下转第 11 页)