实例层数据清洗技术研究

胡文瑜[1,2],应康辉[1,2*]
(1. 福建工程学院 计算机科学与数学学院,福建 福州 350118;
2. 福建省大数据挖掘与应用技术重点实验室,福建 福州 350118)

摘　要:随着科学、技术和工程的迅猛发展,近 20 年来,许多领域诸如光学观测、光学监控、健康医护、传感器、用户数据、互联网和金融公司以及供应链系统等都产生了海量的数据(例如,在医疗检测中,数据都是源源不断而来的,形成了“数据灾难”)。有效的数据分析和数据挖掘建立在数据可用性和数据高质量的基础上,数据高质量的前提是需要对数据进行清洗。数据清洗是对脏数据进行检测和纠正的过程,是进行数据分析和管理的基础,也是常用的提高数据质量的技术。实例层数据清洗是数据清洗的重要组成部分,该文重点对实例层数据清洗技术中属性和重复记录值的检测及清洗方法进行比较和分析总结。介绍了数据清洗技术以电气工程领域、医药领域、交通领域为代表的应用领域结合应用情况,对不同的数据集特点与适用的实例层数据清洗技术提供了有价值的选择建议。最后对实例层数据清洗技术面临的问题与挑战及发展方向进行了展望。

关键词:实例层数据清洗;属性检测;属性清洗;重复记录检测;重复记录清洗
中图分类号:TP309　　**文献标识码**:A　　**文章编号**:1673-629X(2022)05-0022-07
doi:10.3969/j.issn.1673-629X.2022.05.004

Study of Instance-level Data Cleaning Technology

HU Wen-yu[1,2], YING Kang-hui[1,2*]
(1. School of Computer Science and Mathematics, Fujian University of Technology, Fuzhou 350118, China;
2. Fujian Provincial Key Laboratory of Big Data Mining and Applications, Fuzhou 350118, China)

Abstract: With the rapid development of science, technology and engineering, in the past 20 years, many fields such as optical observation, optical monitoring, health care, sensors, user data, Internet and financial companies, and supply chain systems have produced massive amounts of data (For example, in medical testing, data is constantly coming in, forming a "data disaster"). Effective data analysis and data mining are based on data availability and data high quality. The premise of data high quality is the need to clean the data. Data cleaning is the process of detecting and correcting dirty data, is the basis for data analysis and management, and is also a commonly used technology to improve data quality. Instance-level data cleaning is an important part of data cleaning. We focus on comparing, analyzing and summarizing the detection and cleaning methods of attributes and repeated record values in the instance-level data cleaning technology, and introduce the combined application of data cleaning technology represented by the electrical engineering field, the medical field, and the transportation field, and provide valuable selection suggestions for the characteristics of different data sets and the applicable instance-level data cleaning technology. Finally, the problems, challenges and development directions of the instance-level data cleaning technology are prospected.

Key words: instance-level data cleaning; attribute detection; attribute cleaning; repeated record detection; repeated record cleaning

0　引　言

随着信息技术的高速发展,生成、收集和存储大型数据集变得越来越容易。虽然通过大数据分析可获得有价值的信息与智慧见解,但这是建立在数据可用性或脏数据被充分清洗的基础上。影响数据可用性的因素包括:不一致值、重复值、空值和拼写问题等。数据清洗是清洗数据中存在的错误和不一致等问题来提高数据质量[1]。数据质量问题分为数据模式和实例数据,数据清洗也分为模式层清洗和实例层清洗。模式层的清洗主要是完整性约束、异构模式设计和结构冲

收稿日期:2021-06-29　　**修回日期**:2021-10-29
基金项目:国家重点研发计划子课题(2018YFC1201103)
作者简介:胡文瑜(1963-),女,博士,教授,CCF 会员(H0737M),研究方向为数据分析与处理、数据库信息安全;通讯作者:应康辉(1997-),男,硕士研究生,CCF 会员(H0735G),研究方向为数据清洗、数据融合。

突,需要通过程序自动发现或者人工实现清洗。实例层的清洗主要对属性值和重复记录进行清洗[2]。数据集中的属性和重复记录问题是脏数据的主要构成,均属于实例层数据清洗目标,所以对实例层数据清洗技术的研究是有意义和价值的。

1　实例层数据清洗

数据清洗主要在数据仓库、数据库知识发现和决策支持这三个领域研究。数据仓库领域中,数据清洗是构建数据仓库的第一步。作为数据清洗中重要的组成部分,实例层数据清洗显得更加重要。实例层数据清洗有以下两个方面研究:(1)属性错误检测与消除:文献[3]采用统计方法来检测数值型属性;聚类方法来寻找出字段级检查不出的孤立点;分箱方法用于清洗异常数据。文献[4]对关系中的数据进行插入、删除等操作来消除数据冗余问题。(2)重复记录检测与消除:采用基于语义和字面的检测方法来检测重复记录;优先队列、近邻排序等方法消除重复记录[5-6]。表 1 介绍了一些国内外研究方法。图 1 描述了实例层数据清洗算法分类。

表 1　现有的一些国内外实例层数据清洗方法

	文献提出的方法
数据异常检测与消除	文献[7]提出一种基于多层架构的油中溶解气体数据清洗与异常识别的方法,首先利用变分模态分解去除时间序列中的趋势项,结合 3δ 准则对时序数据中的噪声值、缺失值、暂时性迁移数据等进行异常识别,然后根据关联分析结果对可清洗的异常数据利用长短期记忆神经网络进行重构清洗 文献[8]提出了基于密度的局部离群点检测算法,该方法利用数据对象周围的相对密度衡量异常因子,相对密度是反映局部的数据分布,可以避免全局数据的影响 文献[9]提出一种基于 K-means 和时间序列分析的变压器异常值检测与清洗的方法。该方法通过灰色关联法从多维的在线监测数据流中筛选关联度高的相关序列,然后基于 K-means 方法对在线监测数据进行异常检测,利用时间序列预测方法完成噪声数据清洗及趋势预测
重复记录检测与消除	文献[10]提出了一种基于改进量子粒子群算法的数据库重复记录检测方法。该方法根据对象之间的相似性构造熵度量,并评估数据库原始数据集中每个属性的重要性,从而去除不重要或噪声属性 文献[11]提出一种基于 IQPSO(Improved Quantum Particle Swarm Optimization)算法的数据库重复记录检测方法。该方法根据对象之间的相似性构建了一个熵度量,并对数据库的原始数据集中每个属性的重要性进行评估,从而消除不重要或噪声的属性

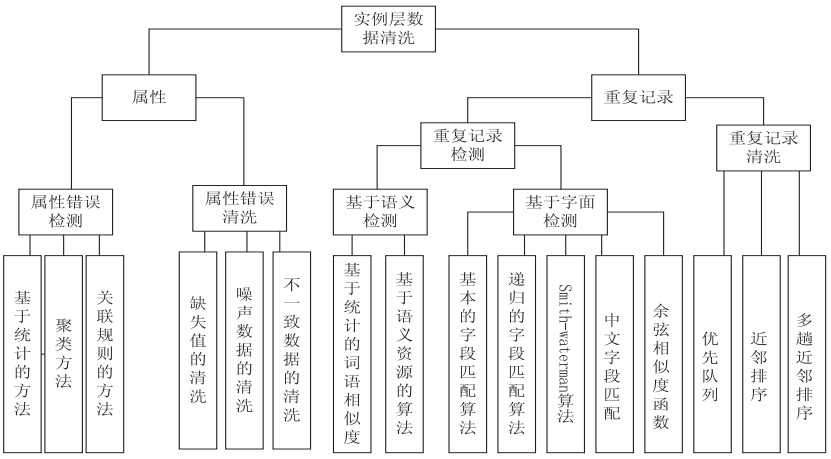

图 1　实例层数据清洗算法分类

1.1　属性错误检测

自动检测属性错误方法减少了人工操作且效率高,具体的方法有基于统计的方法[12]、聚类方法[13]和关联规则方法[14]。表 2 对上述三种方法进行了比较。

表 2　自动检测属性错误方法的比较

方法	优点	缺点	适用的数据集特点
统计分析	随机选取样本数据	准确性低	数值型,小数据集或大数据集(取样),结构化
聚类	找到字段级检查不出的孤立点且区分同一簇中对象	计算复杂度高	字符型,较小的数据集,半结构化,对初始值和噪声数据很敏感
关联规则	置信度和支持度高	计算量大	字符型,大数据集,结构化

1.2 属性错误清洗

属性错误清洗包括 3 个方面:(1)清洗空缺值:采用忽略元组、全局变量、属性的平均值和中间值等统计值来填充空缺值[15]。(2)清洗噪声数据:采用分箱法,“箱的深度”表示不同的箱里有相同个数的数据,“箱的宽度”表示每个箱中数值的取值区间为常数,把属性值分配到等深或等宽的“箱”中,用箱中属性值的平均值来替换“箱”中的属性值[16]。(3)清洗不一致数据:采用条件函数依赖、标准函数库和汇总分解函数来清洗。例如属性之间的关系采用函数依赖来清洗不一致数据。表 3 是关于属性错误清洗方法的比较。

表 3 属性错误清洗方法的比较

方法	优点	缺点	适用条件及其数据集特点
忽略元组	简单方便	属性缺少值多时效果差,要考虑数据是否干净	小数据集
使用全局变量填充	简单	产生较大的错误结果	大数据集;适合数值型和字符串型
使用属性的平均值、中间值等统计值填充	准确性较高	可能填充错误的值	适合数值型;平均值填充适合非数值型
分箱	减少每个属性的不同值数量	准确性不高	建立在稳定数据变量基础上

1.3 重复记录检测

检测重复记录的方法有:基本的字段匹配方法[17]、Smith-Waterman 算法、R-S-W 算法、编辑距离方法[18]、基于 N-gram 的字符串匹配算法、中文字段匹配算法和余弦相似度函数[19]、基于统计的词语相似度算法、基于语义资源的算法。

基本的字段匹配方法是把两个分词串中顺序匹配的分词个数除以所有分词个数的平均值,计算出匹配度。如表 4 中两条记录中的地址字段值,字段匹配度 $= k/((|A|+|B|)/2) = 0.85$,$|A|$ 和 $|B|$ 分别为 A,B 中分词的个数。

表 4 重复记录例子

变量	地址
A	Comput sci university califoniasan diego
B	Department of computer science uni califsan diego

文献[19]提出改进后的 Smith-Waterman 算法(R-S-W),该算法对拼写错误和字符串的顺序以及缩写有着深入研究,却无法应用在中文。文献[20]提出基于 N-gram 的字符串匹配算法,对两个字符串中 n 个字符进行排序组合及比较两个字符串的极限阈值,从而得出两个字符串是否相同。文献[21]采用改进编辑距离方法来计算中文句子的相似度。文献[22]提出了基于 PMI-IR 算法,该方法是搜索引擎来获取数据并采用点互信息作为词语相似度计算的指标。文献[22]提出一种基于知网、面向语义的词汇语义相似度计算方法,该方法通过概念切分解决知网中未登录的语义相似度问题。表 5 是关于常用重复记录检测算法的比较。

表 5 常用重复记录检测算法的比较

算法	优点	缺点
基本的字段匹配	算法直观	不能处理不是前缀的缩写情况
Smith - Waterman 算法	识别字符串缩写	不能识别子串顺序颠倒
R-S-W	识别拼写错误、子串颠倒和缩写	不适合中文
基于 N-gram 的字符串匹配	解决字符串排序错误问题	不能识别拼写错误,不能忽略空格以及标点符号
基于编辑距离的字段匹配	可以捕捉拼写错误、单词的插入和删除错误	匹配效果差
余弦相似度函数	解决单词插入和删除的问题	处理不了拼写错误
中文字段匹配	简单直观	不能进行符合汉字表意的切分
基于统计的词语相似度算法	精确度高	计算量大,训练库大,
基于语义资源的算法	简单有效	受人的主观影响较大

1.4　重复记录清洗

消除重复记录的算法有:优先队列算法[23]、近邻排序算法(SNM)、多趟近邻排序(MPN)、优化的多趟近邻排序算法(OMPN)[24]。优先队列法是由 Monge 提出,首先数据集会根据关键词进行排序,再对排序后的顺序依次扫描数据集。Hernandez[24]提出近邻排序法,该方法通过关键字进行排序,采用固定大小的滑动窗口在排序后的数据集上滑动并重复检测窗口,减少记录的比较次数。文献[13]提出多趟近邻排序法。该方法要求在排序的数据集上使用近邻排序方法,但使用的是不同的关键词和较少的窗口,而且还要对 MPN 算法的结果求传统闭包。文献[25]改进了近邻排序算法,通过比较相似度与阈值来调整窗口值的大小,并加入有效权值来减少字段缺失的影响。文献[25]中 OMPN 算法对 MPN 算法在选取排序关键字时过于依赖专家经验的缺陷进行了改进。衡量三种近邻排序算法的标准是召回率、误识别率和精确度。表 6 是上述各种算法的比较。

表 6　常用重复记录清洗算法比较

算法	优点	缺点
优先队列	减少记录比较的次数,提高匹配效率	算法复杂
近邻排序	使用简单,采用滑动窗口运行速度快	滑动窗口的大小很难控制并且识别精度 依赖于排序所选择的关键词
多趟近邻排序	精确度更高,滑动窗口更小	依赖主键域的记录
优化的多趟近邻排序	关键字提取准确,采用自适应的滑动窗口	时间复杂度高,没有对不完整的排序 键值的记录进行归并

上面分别对实例层数据清洗中属性错误、重复记录各自对应的方法进行比较和分析,属性错误检测中主要针对结构化和半结构化数据,还需要对非结构化数据进行研究。属性错误清洗中分箱和回归的方法都相对简单,容易解决,但准确性不高。重复记录检测中的算法都比较简单直观,但是中英文和语义的关系不能两者都实现,下一步还需要实现中英文和语义的结合。重复记录清洗中近邻排序、多趟近邻排序和优化的多趟近邻排序都是基于窗口大小和排序关键字来判别,而优先队列只是运用关键字排序来判定,效果没有排序算法的好。

2　数据清洗与其他领域的结合

数据清洗可以和其他领域进行结合,获得更好的应用发展。例如,数据清洗中重复记录技术或方法应用于电气工程领域的数据汇聚与数据清洗中。随着数字电厂的不断建设、大数据平台的逐步完善和智能设备的进一步推广,电力系统数据量急剧增长,需要展开大量的数据分析,然而电力数据在采集、汇聚过程中会出现数据质量问题,造成数据融合困难,这就需要结合数据清洗方法,对电力数据进行实时校验和清洗,提高数据的可用性[26]。下面介绍了数据清洗的一些应用领域,表 7 对实例层数据清洗与其他领域的结合情况进行了阐述。

2.1　医疗领域

医疗领域数据是需要进行数据清洗的一个领域,尤其在医疗体检数据方面,由于医疗体检中心只对受检者提供当次的体检报告,缺乏对受检者历史数据的分析,导致医疗机构体检数据库中存在基本信息缺失、体检项目名称不同、体检指标参考值范围不同的问题[27],因此需要对上述出现的问题进行数据清洗,从而保证数据的干净。林予松等人[28]提出基于分词和权重的字段匹配算法,解决了体检数据不一致的问题,但还需要综合考虑部分重心词前移和算法准确性不高的情况。此外医疗体检中数据容易出现唯一标志码缺失问题,是数据清洗的主要障碍之一。

2.2　电气领域

随着电力系统信息化程度的提高和智能电网的加速建设,用户电力数据量呈指数型增长状态,但是电力数据量的增长而导致电力数据也出现诸多问题。文献[29]将电力数据看成时间序列,用 ARIMA 拟合并迭代检验的方法修复缺失数据,但是该方法利用的信息较少,且不适合修复缺失点连续分布的情况。文献[30]通过训练 RBF 神经网络作为状态转移方程,再利用卡尔曼滤波方法对数据进行滤波并修复,但是该方法计算量较大,在细节上把握不够精确。田英杰等人[31]提出函数型数据分析对错误和缺失数据进行修正和补全。通过函数估计方法,将原有观测个体的离散数据映射到一个新的函数空间,将数据中缺失的成分利用相似的方法修复缺失数据,但是该方法利用的信息较少,且不适合修复缺失点连续分布的情况。随着用户电力数据量逐步的增长,使得该领域数据清洗的任务变得更加困难和富有挑战性。

2.3　交通领域

交通数据的采集和处理技术是智能交通系统的关键性技术,无线电和计算机技术的蓬勃发展使得 RFID

检测技术作为一种新型检测技术广泛应用于道路交通数据采集。由于 RFID 检测设备故障、通信系统故障及环境等异常原因,采集到的交通数据存在冗余、遗漏、错误和不精确的现象,将导致产生不稳定因素,影响交通状态估计、预测及评价,进一步影响交通管理和公众出行信息服务的质量。文献[32]提出一种基于最大频繁模式因子的高位孤立点挖掘算法,能解决孤立点挖掘算法中存在的不容易获取完全频繁模式和时间复杂度高等问题,并且可以减少占用内存,提高运行效率。为了减少数据的错误和冗余,对交通数据的数据清洗得更全面和准确。

表 7 实例层数据清洗的领域应用情况

应用领域	方法	优点
生物环境	文献[33]的基于数据清洗的生物建模方法通过统计分析法筛选数据集中可疑数据,然后利用物料平衡计算法剔除误差,使得实测数据集进一步修正	提升基础建模质量,提高生物模型的可靠性
航空工程	文献[34]的 ADS-B(空管监视技术)数据清洗方法对样本数据集建立航迹模型进行分析,清洗时间戳、经纬度、气压等特征字段,运用密集聚类方法检测离群点	有效清理航迹异常点,准确获得良好的时间戳效果
光电技术	文献[35]的基于滑动标准差的光伏阵列异常数据清洗方法,分析了光伏阵列异常数据分布特征及来源,给出了光伏阵列滑动标准差的计算方法	解决清洗结果容易受到异常数据分布的影响,提高了异常值识别准确率

3 实例层数据清洗面临的挑战及应用研究展望

实例层数据清洗在某些特定领域的数据质量工程中有许多应用需求和应用研究,这些应用研究针对的是特定行业背景的数据清洗任务,但存在着技术局限和不足。文献[36]采用四分位法和 K-means 算法消除异常值,由于 K-means 算法是一种聚类算法,可能会导致正常数据的错误删除,此外 K 值的选择比较复杂,对数据清洗的处理结果有不利影响。文献[37]采用基于密度的局部离群因子算法将足够高密度的区域划分为簇,可以有效地检测出散乱的离群点,但不适用于高密度的堆积离群点。文献[38]是根据异常值的位置分布来检测异常值,它不需要数据样本训练而且是普遍适用的。但是对大量堆积离群值的检测和清除还需要改进,对风速功率曲线离群点的空间分布和形状研究还待深入。文献[39]分析了风力涡轮机中风电异常值的分布特征和分类,并提出了一种基于变点分组和四分位算法的联合数据清理算法。该方法识别风电曲线的叠加异常值和散乱离群值,清洗效果好,效率高,通用性强,可以处理影响数据完整性的异常数据,但是没有考虑数据校正和数据插值来提高数据质量,没有根据实际情况来进行数据修正。文献[40]的方法可以通过文本的重要关键词在一定程度上体现文本的主题,而且统计词频处理相对简单。但词频类算法只统计词语出现的次数,却忽略了关键词所在文档结构上的位置情况和上下文关键词的关联信息。文献[41]提出结合众包数据库的集成机器学习算法,将人类标签的准确性与机器学习分类器的速度与成本效益相结合,该方法可以提取半结构数据中的有效信息,在一定程度上可以更正数据中存在的缺陷,但是对半结构化数据进行规格化还是无能为力。文献[42]提出非结构化数据融合方法,该方法降低数据噪声的干扰,提取剩余数据,整合相关的数据,但是由于非结构化数据本身的特点,多源数据融合分析有很大的难度。

3.1 问题与挑战

通过研究发现,实例层数据清洗是一个相对成熟但又期待有更多突破和创新的领域,技术发展的空间还很大,包括:

(1)数据相似度检测的对象主要是数据库中的短文本,对长文本的数据没有进行充分研究,主要是长文本语言本身的复杂性和文本中的词表结构数据有很强的依赖性。

(2)目前的研究成果主要适用于结构化数据,然而待处理的半结构化数据和非结构化数据的规模远远大于结构化数据。半结构化数据处理难度大且非结构化数据格式多样、缺乏实效性、数据含义比较隐性不容易察觉。

(3)实例层数据清洗的算法目前人工参与度较高,不适合大规模数据的清洗。因此需要普适性好、通用性强、计算机能自动识别的实例层数据清洗算法。

(4)通用的数据清洗算法在专用数据集上都需要结合领域知识,因此需要逐步建立起电气工程、光电技术等特定领域的数据清洗标准规则库。

3.2 未来展望

就现阶段的数据清洗研究工作取得的成绩和存在的问题而言,未来可以通过以下几个方面对数据清洗进行研究:

(1)长文本的相似度检测:随着文本信息的增多,

长文本的检测变得越来越需要,例如论文查重、新闻、大规模网页去重等,采用基于语义信息的相似度检测算法来去除长文本中多余的内容,并且能够解决文本中同义词替换以及一词多义的问题,但是对可利用的信息数量与质量的要求比较高。因此可以结合已有的语义网、深度学习算法和半监督学习算法来提高数据质量。

(2)半结构化和非结构化数据的处理:目前数据结构多种多样,不只是结构化数据的情况,非结构化和半结构化数据现在变得越来越多。采用基于正则表达式的属性集识别方法来识别半结构化数据中的属性集,并进行规则化操作。其实可以将主动学习优化成果应用在基于正则表达式的属性集识别方法中,通过该方法加强机器的学习能力,进一步缩减人工参与。非结构化数据采用非结构化数据分析与决策系统能快速分析出来且发现其中隐藏的价值。可以结合 Hadoop 和机器学习方法来处理非结构化数据中低容错率以及识别活动数据的情况。

(3)自动识别的数据清洗算法:自动的 Web 页面清洗方法可以对相同或相似布局特征的海量 Web 页面进行自行清洗,保存有价值的文本和内容。采用树编辑距离的方法对 Web 页面结构进行分类,可以更好地提高 Web 页面清洗的准确率。

(4)特定领域的数据清洗标准库:目前提出的基于编程语言的反射技术和 python 脚本的银行领域数据清洗规则库,有效降低了数据清洗的复杂度。再结合基于分级规则库的方法来构建规则库的逻辑关系,可以更好地减少数据清洗出错率。

综上所述,需要找到能适用于大数据、流数据、半结构化和非结构化数据集的实例层数据清洗解决方案,能在现有实例层数据清洗技术上找到时空效率高且通用性好的自动化数据错误检测和错误纠正算法,能根据应用领域和数据集特点自动的选择合适的实例层数据清洗技术,允许用户在通用数据清洗技术上定制特定应用领域(比如电气工程领域)的专用数据清洗规则。传统的实例层数据清洗技术仍有研究和发展空间,期待着技术创新、应用创新和突破性进展。

参考文献:

[1] RAHM E,DO H H. Data cleaning:problems and current approaches[J]. IEEE Data Engineering Bulletin,2000,23(4):3-13.

[2] 李学龙,龚海刚. 大数据系统综述[J]. 中国科学:信息科学,2015,45(1):1-44.

[3] 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.

[4] 邓盐婷,曲卫平. 一种消除数据异常的关系代数运算方法[J]. 电脑知识与技术,2014(11):2523-2526.

[5] 张培根,黄树成. 一种用于中文数据清洗的近邻排序算法[J]. 计算机应用与软件,2018,35(8):286-288.

[6] 袁　满,穆永豪,王贵友,等. 改进的 SNM 中文语义重复记录检测算法[J]. 吉林大学学报:信息科学版,2021,39(3):348-356.

[7] 袁　帅,王广真,张兴辉,等. 变压器油中溶解气体分析的实验室间比对及结果分析[J]. 变压器,2020,57(6):60-62.

[8] 董　泽,贾　昊. 基于 EWT-LOF 的热工过程数据异常值检测方法[J]. 仪器仪表学报,2020,41(2):126-134.

[9] 陆春光,叶方彬,赵　羚,等. 基于密度峰值聚类的电力大数据异常值检测算法[J]. 科学技术与工程,2020,20(2):654-658.

[10] PAN Zheng,LIU Shuai,FU Weina. A review of visual moving target tracking[J]. Multimedia Tools and Applications,2017,76(16):16989-17018.

[11] YU G. Database repeat record detection based on improved quantum particle swarm optimization algorithm[J]. International Journal of Performability Engineering,2019,15(2):710-718.

[12] MA L,PEI Q,ZHOU L,et al. Federated data cleaning:collaborative and privacy-preserving data cleaning for edge intelligence[J]. IEEE Internet of Things Journal,2021,8(8):6757-6770.

[13] 叶　晨,王宏志,高　宏,等. 面向众包数据清洗的主动学习技术[J]. 软件学报,2020,31(4):1162-1172.

[14] 李　蕾. 大数据环境下相似重复记录数据清洗关键技术研究[D]. 南京:南京邮电大学,2019.

[15] 杨俊闯,赵　超. K-Means 聚类算法研究综述[J]. 计算机工程与应用,2019,55(23):7-14.

[16] 徐搏超. 基于参数关联性的电站参数异常点清洗方法[J]. 电力系统自动化,2020,44(20):142-147.

[17] 谢明吉. 数据清洗中相似记录检测的研究[D]. 广州:华南理工大学. 2010.

[18] 宋　玲,马　军,连　莉,等. 文档相似度综合计算研究[J]. 计算机工程与应用,2006(30):160-163.

[19] 盛怡瑾,张学福,孙　巍,等. 数据匹配算法应用对比研究——以期刊数据融合中作者和机构匹配为例[J]. 数字图书馆论坛,2015(10):14-20.

[20] 周　旭. 哼唱检索系统在音乐播放器中的研究与实现[D]. 呼和浩特:内蒙古大学,2012.

[21] WU Z,LIANG J,ZHANG Z,et al. Exploration of text matching methods in Chinese disease Q&A systems:a method using ensemble based on BERT and boosted tree models[J]. Journal of Biomedical Informatics,2021,115:103683.

[22] 武永亮,赵书良,李长镜,等. 基于 TF-IDF 和余弦相似度的文本分类方法[J]. 中文信息学报,2017,31(5):138-145.

[23] CAI X,DONG S,HU J. A deep learning model incorporating

part of speech and self-matching attention for named entity recognition of Chinese electronic medical records[J]. BMC Medical Informatics and Decision Making,2019,19(2):101-109.
[24] 张 攀. 面向重复记录检测的数据清洗算法的研究[D]. 西安:西安电子科技大学,2018.
[25] 张培根. 近邻排序算法研究及在中文数据清洗中的应用[D]. 镇江:江苏科技大学,2018.
[26] 冯泽磊,吴美凤. 动态浮箱数据清洗方法在电力系统中的应用[J]. 发电技术,2019,40(S1):109-113.
[27] 张 茜. 大型纵向监测健康管理队列设计及其统计分析策略研究[D]. 济南:山东大学,2017.
[28] 林予松,王培培,刘 炜,等. 医疗体检数据预处理方法研究[J]. 计算机应用研究,2017,34(4):1089-1092.
[29] 严英杰,盛戈皞,陈玉峰,等. 基于时间序列分析的输变电设备状态大数据清洗方法[J]. 电力系统自动化,2015,39(7):138-144.
[30] 俞娜燕,李向超,费 科,等. 基于 Sigma 卡尔曼滤波的光伏电站监测数据修复方法[J]. 数字技术与应用,2018,36(8):32-34.
[31] 田英杰,洪子靖,周 李. 基于函数型数据分析的工商业居民用户电力数据清洗算法[J]. 电测与仪表,2021,58(1):11-19.
[32] 申利民,孙中魁,陈 磊,等. 一种改进的高维孤立点挖掘入侵检测方法[J]. 小型微型计算机系统,2020,41(12):2636-2640.
[33] 李天宇,吴远远,郝晓地,等. 数据清洗对污水处理厂生物建模可靠性影响研究[J]. 环境科学学报,2020,40(10):173-186.
[34] 王 兵. ADS-B 历史飞行轨迹数据清洗方法[J]. 交通运输学报,2020,10(12):173-186.
[35] 时 珉,尹 瑞,胡傲宇,等. 基于滑动标准差计算的光伏阵列异常数据清洗办法[J]. 电力系统保护与控制,2020,48(6):108-114.
[36] DAKI H,EL HANNANI A,AQQAL A,et al. Big data management in smart grid:concepts,requirements and implementation[J]. Journal of Big Data,2017,4(1):1-19.
[37] ZHENG L,HU W,MIN Y. Raw wind data preprocessing:a data-mining approach[J]. IEEE Transactions on Sustainable Energy,2014,6(1):11-19.
[38] WANG Y,INFIELD D G,STEPHEN B,et al. Copula-based model for wind turbine power curve outlier rejection[J]. Wind Energy,2014,17(11):1677-1688.
[39] SHEN X,FU X,ZHOU C. A combined algorithm for cleaning abnormal data of wind turbine power curve based on change point grouping algorithm and quartile algorithm[J]. IEEE Transactions on Sustainable Energy,2018,10(1):46-54.
[40] 姜 雪. 基于 simhash 的文本相似检测算法研究[D]. 绵阳:中国工程物理研究院,2017.
[41] 于溪淼. 基于主动学习的半结构化数据清洗技术研究[D]. 哈尔滨:哈尔滨工业大学,2017.
[42] 郭春霞. 大数据环境下高校图书馆非结构化数据融合分析[J]. 图书馆学研究,2015(5):30-34.

(上接第 21 页)

gence and security (CIS). London:Pentech,2019:406-410.
[10] 戴喜妹,张军峰,赵鹏力,等. 离场航班多目标优化排序研究[J]. 哈尔滨商业大学学报:自然科学版,2019,35(2):241-245.
[11] VESIKAR Y,DEB K,BLANK J. Reference point based NSGA-Ⅲ for preferred solutions[C]//2018 IEEE symposium series on computational intelligence (SSCI). [s. l.]:IEEE,2018:1587-1594.
[12] 李朝芳,苟 英. 基于改进 NSGA-Ⅱ算法测试资源优化配置的研究[J]. 科学咨询,2017(45):38.
[13] 杨 爽. 基于投影面的多目标进化算法 MOEA/P[D]. 沈阳:沈阳化工大学,2021.
[14] FEI Xue,DI Wu. NSGA-III algorithm with maximum ranking strategy for many-objective optimisation[J]. International Journal of Bio-Inspired Computation,2020,15(1):14-23.
[15] LIU Y,GONG D,SUN X,et al. Many-objective evolutionary optimization based on reference points[J]. Applied Soft Computing,2017,50:344-355.
[16] FU X,SUN J,ZHANG Q. A reference-inspired evolutionary algorithm with subrigion decomposition for many-objective optimization[C]//UK workshop on computational intelligence. Berlin:Springer,2017:145-156.
[17] YUAN Y,XU H,WANG B,et al. Balancing convergence and diversity in decomposition-based many-objective optimizers[J]. IEEE Transactions on Evolutionary Computation,2016,20(2):180-198.
[18] 吴伟丽. 基于 NSGA-Ⅲ的复杂成因变压器直流偏磁控制优化算法[J]. 电测与仪表,2018,55(11):89-93.
[19] 胡 涵,李振宇. 多目标进化算法性能评价指标综述[J]. 软件导刊,2019,18(9):1-4.