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。