

数据采集过程中的清洗技术及应用

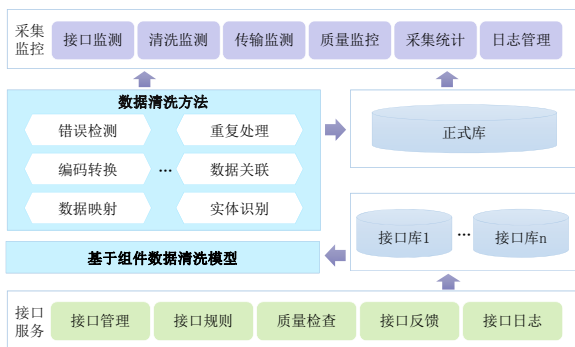
李培军、许舒人、魏克刚、田申、张晓辉、罗剑刚
peijun@iscas.ac.cn, 13911990831

数据清洗和预处理是数据分析的关键，直接决定后续分析的效果，一般数据清洗占分析过程的50%—80%时间，尤其是低质量数据清洗难度大。在一些信息化程度低的行业领域由于数据质量差给数据共享交换和清洗带来较大难度，通过对低质量数据采集过程中清洗和预处理方法的研究来提升自动和智能的数据清洗能力，提升数据采集的质量和效率。

1 关键技术

采集与清洗流程

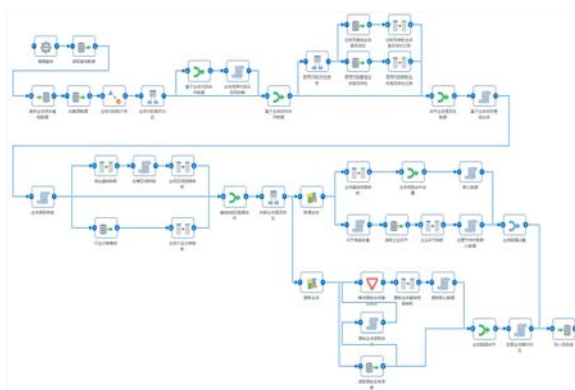
采用基于组件的可视化数据清洗流程，结合数据挖掘和机器学习等处理方法，实现高效、自动和智能的数据清洗和预处理，用于解决低质量数据复杂的清洗需求。



数据处理流程

数据清洗模型

以Metl为基础基于组件（读取器、处理器、写入器、服务、控制）方式建立可自定义的数据清洗模型，针对低质量多源异构数据的复杂清洗需求，实现快速的处理能力，节省处理时间，提升处理质量。



基于组件清洗模型

数据清洗方法

- 基于规则和模型的自动错误检测方法
- 大规模数据量下重复数据检测和处理方法
- 基于文本相似性分析的编码转换方法
- 基于统计分析的数据映射方法
- 基于相似性分析的数据关联方法
- 基于命名实体识别的清洗和转换方法



智能数据清洗方法

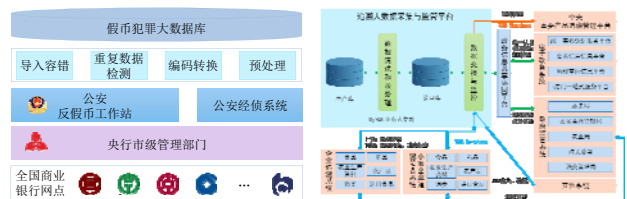
2 应用案例

公安假币犯罪大数据采集与处理

- 高准确率的采集和清洗转换
- 数据类型和信息编码的自动转换与映射
- 案件、嫌疑人、假币相关性分析

追溯大数据采集与监管

- 采用统一数据模型减少60%以上处理工作量
- 解决政务数据交换难题对接5部门多个系统
- 基于数据关联技术的全程追溯链条合成也适用于物流、供应链、政府、金融等领域。



假币犯罪大数据

追溯大数据