DOI:10.3969/j.issn.1001 -4551.2017.12.022

# 面向仓储的 RFID 数据清洗技术研究*

柴文超[1],汤洪涛[1,2*],吴光华[1,3]
(1.浙江工业大学 机械工程学院,浙江 杭州 310032;2.浙江省先进制造技术重点实验室,浙江 杭州 310027;
3.浙江汇智物流装备技术有限公司,浙江 湖州 313028)

**摘要**:针对仓储中无线射频识别(RFID)原始数据的不可靠性问题,对仓储中 RFID 数据冗余和事件流乱序问题进行了研究,提出了一种面向仓储的 RFID 数据清洗模型。首先对仓储中 RFID 数据存在的问题进行了描述,建立了冗余数据和事件流乱序问题产生的抽象场景;然后结合 RFID 仓储数据清洗模型提出了相应的冗余数据清洗和事件流乱序修正算法,并介绍了相应算法的改进之处及具体实现步骤;最后利用读写器检测模型构造了 RFID 原始数据流,通过不同实验参数对记录数量和正确率指标进行了实验测试。研究结果表明:提出的数据清洗方法可以有效去除冗余数据并提高 RFID 事件输出的正确率。

**关键词**:无线射频识别;数据清洗模型;冗余数据清洗;乱序事件流修正
**中图分类号**:TP391 **文献标志码**:A **文章编号**:1001 -4551(2017)12 -1474 -06

# RFID data cleaning technology on warehouse-oriented

CHAI Wen-chao[1], TANG Hong-tao[1,2], WU Guang-hua[1,3]
(1. College of Mechanical Engineering, Zhejiang University of Technology, Hangzhou 310032, China;
2. Key Laboratory of Advanced Manufacturing Technology of Zhejiang Province, Hangzhou 310027, China;
3. Zhejiang Huizhi Logistics Equipment Technology Co., Ltd., Huzhou 313028, China)

**Abstract**: Aiming at unreliability of original RFID data in warehouse, the redundancy and out-of-order event flows of RFID data in warehouse were studied to propose a warehouse-oriented RFID data cleaning model. Firstly, problems of RFID data in warehouse were described and an abstract scenario generated by redundant data and out-of-order event flows was set up. Secondly, combined with the RFID warehouse data cleaning model, the corresponding algorithms were proposed for cleaning redundant data and modifying out-of-order event flows and the improvements of algorithm and detailed implementation procedures were introduced. Finally, by means of the reader inspection model, original RFID data flows were constructed to test the number of records and indicators of accuracy on the basis of different experiment parameters. The results indicate that the proposed data cleaning method can effectively remove redundant data.

**Key words**: radio frequency identification(RFID); datacleaning model; redundantdata cleaning; out-of-order event flow correction

## 0 引 言

无线射频识别作为一种实时数据采集技术,被广泛应用于物流、资产追踪、设备监控等领域[1-2]。但由于 RFID 数据具有客观不可靠性,容易产生冗余数据或时间戳乱序现象[3]。如果由上层应用系统来解析如此巨大的 RFID 数据,不仅会造成系统处理业务逻辑的效率低下,而且开发的系统也不利于扩展[4]。

国内外学者围绕 RFID 不确定性数据清洗问题

**收稿日期**:2017 -03 -26
**基金项目**:国家自然科学基金资助项目(51605442);浙江省先进制造技术重点实验室开放基金资助项目(2016KF03);浙江省教育厅资助项目(Y200909905)
**作者简介**:柴文超(1991 -),男,浙江衢州人,硕士研究生,主要从事 RFID 系统方面研究。E-mail: chaiwenchao@zjut.edu.cn
**通信联系人**:汤洪涛,男,副教授,硕士生导师。E-mail: tanght@zjut.edu.cn

展开了一系列研究。JEFFERY, GAROFALAKIS 等人[5]针对标签数据清洗最早提出了一种自适应滑动窗口的数据清洗算法,通过构建标签数据清洗模型对标签漏读、脏读等现象进行处理,从而为上层应用提供可靠的数据[5];MASSAWE, KINYUA 等人[6]在 JEFFERY 研究的基础上针对 RFID 数据的不可靠性提出一种自适应滑动窗口处理方法,有效处理了动态环境和标签下的数据处理;为了解决冗余数据的处理效率问题,贾红梅、李文杰[7]分析了仓储中标签冗余和阅读器冗余问题产生的原因,并给出了解决这两种冗余的数据过滤模型。FAN, WU 等人[8]设计并部署了一种基于标签行为的 RFID 数据清洗系统,并使用了一种通用的 RFID 应用程序来验证该方案的有效性,LI, LIU 等人[9]针对 RFID 系统中发生的数据流乱序现象提出了一种优化序列扫描和构建的方法。

本研究以仓储环境为对象,针对冗余数据清洗问题,提出队列缓存机制实现在线清洗数据。

# 1　仓储数据问题产生场景抽象

由于现场环境和读写器设备本身性能的影响,在每一个读写器采集位置都有可能产生冗余数据和乱序事件。如果能在实际环境中抽象出具有通用特征的场景,则更加便于 RFID 仓储中数据的清洗。

## 1.1　数据冗余

数据冗余主要包括:单数据源冗余和多数据源冗余两方面[10]。

以 RFID 仓储为例其抽象场景如图 1 所示。

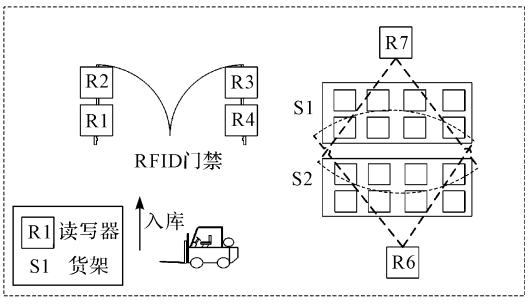


图 1　冗余数据产生场景抽象

单数据源冗余主要指当标签进入读写器 R1 ~ R4 射频范围内,由于 RFID 读周期为 0.2 s 左右,假设时间窗口为 $w$,那么在 $[t, t+w]$ 时间范围会产生大量的具有相同 Tag_ID 的三元组数据。多数据源冗余指不同位置的读写器覆盖范围重叠现象,如图 1 右边所示,R6 和 R7 的读写器覆盖范围重叠,则会产生交叉冗余数据。假设在某一位置存在 $m$ 个读写器,有 $m \times n$ 个固定位置的标签和 $k$ 个等待判定位置的标签,为了方便后续对相关清洗算法地讨论现在对本文相关概念做如下定义:

**定义 1　读写器元组。**为了方便冗余数据过滤论文将标签原始数据扩展为五元组 < Tag_ID, Reader_ID, Reader_Grp, Time_Start, Time_End >, Reader_Grp 属性用于表示读写器所在的位置属性。

**定义 2　确定位置标签信号强度。**假设已知标签 $Tag_i$ 其属于读写器 $\text{Reader}_j$,那么 $\text{Tag}_i$ 相对于 $\text{Reader}_j$ 的信号强度用 $F_{j,i}$ 表示,其中 $i \in (1, \cdots\cdots n)$, $j \in (1, \cdots\cdots m)$。

**定义 3　待判定位置标签信号强度。**假设待判定位置标签 $\text{Tag}_i$ 相对 $\text{Reader}_j$ 信号强度用 $W_{l,j}$ 表示,其中 $l \in (1, \cdots\cdots k)$, $j \in (1, \cdots\cdots m)$。

**定义 4　标签位置相似度。**通过欧式距离公式计算 $W_{l,j}$ 和 $F_{j,i}$ 之间的距离 $D_{l,j}$,如果该距离很小那说明该位置不确定的标签 $Tag_l$ 属于该读写器 $j$,去计算公式为:

$$D_{l,j} = \sqrt{\sum_{i=1}^{n} (W_{l,j} - F_{j,i})^2}, l \in (1, \cdots \cdots k), \quad j \in (1, \cdots \cdots m) \tag{1}$$

**定义 5　冗余数据产生条件。**在某一确定时间范围内,当读写器读到相同的标签时,就有可能产生冗余数据,其条件需同时满足:

$$\text{Tag\_ID}_i = \text{Tag\_ID}_j \tag{2}$$

$$(\text{Time}_i = \text{Time}_j) \mid\mid\mid \text{Time}_i - \text{Time}_j \mid \leqslant \text{Threshold} \tag{3}$$

## 1.2　事件乱序

在实际的 RFID 仓储应用环境中,通常会分布式部署 RFID 读写器设备,然后通过 TCP/IP 或者 RS232 等传输方式将各个位置所采集到的标签数据以事件流的形式聚集到后台处理引擎中。在理想状态下先发生的 RFID 简单事件先到达后台数据处理系统,但是由于仓储现场可能存在通信故障、网络延迟等问题,可能造成发生早的事件反而到达事件处理引擎的时间更晚[11-12],其抽象场景如图 2 所示。

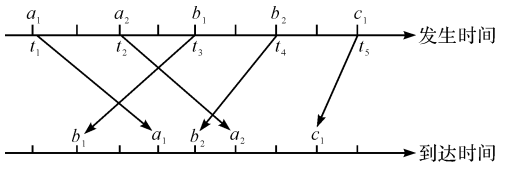


图 2 乱序事件产生场景

以入库流程为例,在质检、入库和上架 3 个位置分别部署 RFID 读写器,假设其触发的事件分别为 A、B 和 C,使用 $T_A$ 代表 A 事件从发生到达系统的时间,$T_{A-B}$ 代表货物从 A 移动到 B 的时间,其他事件时间以此类推。如果满足表达式:$T_A > T_{A-B} + T_B$,则说明先发生的 A 事件后到达,则产生事件流乱序问题[13],在时间戳上先发生的事件后到达则会产生乱序问题。

## 2 面向仓储的 RFID 数据清洗模型

如果需要对 RFID 数据进一步挖掘,那么必须对采集的数据进行处理才能上传至上层应用系统,否则会造成后台系统的业务逻辑设计困难[14]。面向仓储的 RFID 数据清洗模型如图 3 所示。

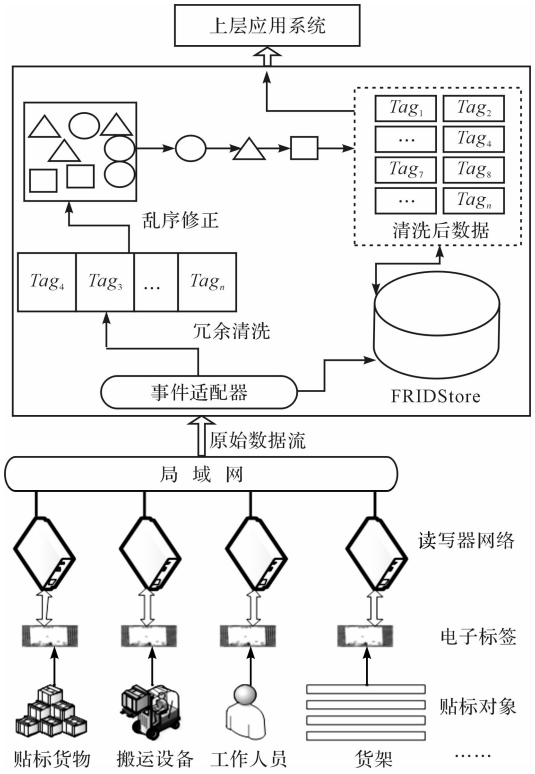


图 3 RFID 仓储数据清洗模型

该数据清洗模型结合冗余数据清洗算法和乱序事件流修正方法,对 RFID 读写器网络所采集到的原始数据进行过滤,最后将干净的数据上传至上层应用系统中进行处理。

### 2.1 冗余数据清洗

本文以多数据源冗余问题为主要对象设计相关算法,引入标签位置相似性的方法来判断标签所属读写器分组,并采用互斥原理来消除不同读写器之间的交叉冗余数据。其核心伪代码为:

算法 1:MD_Cleaning(Multiple Data Source Cleaning)
Input: rfid_stream: <$Tag\_ID_i$, $Reader\_ID_i$,…… >
读写器数量:m
编码规则:RegTag = new Regex("^FA")
冗余条件:RedunCon
Output:清洗后的 RFID 数据流 rfid_CleanStream
1) Init()://读取参数信息
2) BEGIN
3) WHILE(rfid_stream ! = null):
4) IF(RegTag. isMatch($Tag\_ID_i$)):
5) IF(There is Tag wait for arbitration ):
6) ComputerSignalIntensity()
7) FindReaderQueue()
8) END IF
9) For(each $Tag\_ID_j$ in $CQ_m$):
10) IF($Tag\_ID_i$ ! = any $Tag\_ID_j$):
11) $CQ_m$. enqueue($Tag\_ID_i$)
12) Sort(C)$Q_m$
13) ELSE IF(RedunCon )
14) Delete($Tag\_ID_i$)
15) END IF
16) END IF
17) END WHILE
18) MergeData(C)$Q_m$//合并分组队列的数据
19) Output rfid_CleanStream
20) END

算法中输入的参数是读写器原始数据流,输出的参数是清洗后的标签元组映射表,具体步骤为:

(1)读写器网络不断推送原始数据流至系统中,算法首先对标签编码进行判断,如果标签编码符合规则,则进入(2),否则丢弃该标签数据;

(2)根据当前标签的 Reader_Grp 属性判断当前标签所对应的队列是否存在,如果已经存在,则进入对应的标签缓存队列中,否则新建一个属于该读写器组的标签缓存队列 $CQ_i$,如果标签无法判断其位置信息,则进入(3),否则进入(4);

(3)根据定义 4 和公式(1),通过 ComputerSig-

nalIntensity( )计算标签位置相似度,根据欧式距离计算结果大小来判断该标签属于哪个读写器,通过最小化相对位置相似度的方法来裁决该交叉冗余数据所属读写器,判定位置后可以利用互斥原理消除交叉冗余数据;

(4)与当前标签缓存队列中的最后一个标签数据的 Tag_ID 进行对比,如果标签编号满足 $Tag_ID_i = Tag_ID_j$并且时间约束满足$(Time_i = Time_j) \vee |Time_i - Time_j| <= Threshold$,则标签当前标签在短时间内出现了重复读取,那么就剔除旧数据并更新当前数据的最新时间信息,如果不满足约束条件,那么判断当前对比的标签否为队列中的第一个,如果不是就继续循环对比,如果已经对比结束则进入(5);

(5)将新数据存入缓存队列并用 Sort( )方法按到达时间进行排序,如果当前还有数据流需要处理,则继续(2),否则进入(6);

(6)将所属同一读写器组的标签元组数据进行合并,最后输出清洗后的数据。

### 2.2　乱序事件修正

目前提出的乱序事件流修正框架通常采用 Hash 加单链表结构进行数据存取[15],采用双链表结构有利于提高数据的操作效率。

本文提出的乱序事件流修正模型如图 4 所示。

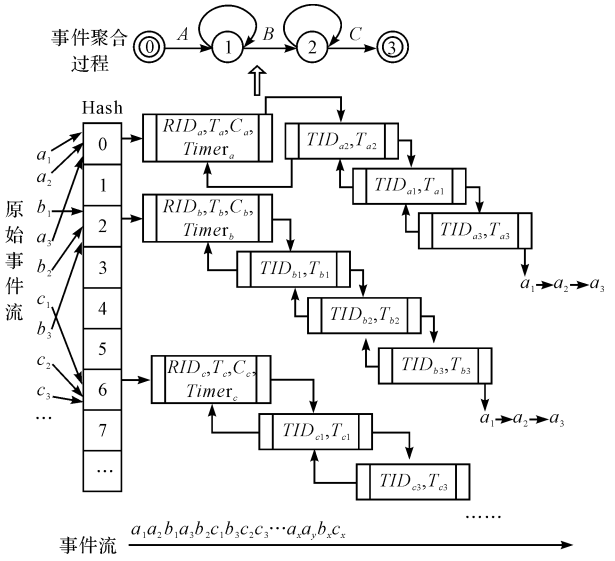


图 4　RFID 乱序事件流修正模型

TID—Tag_ID;RID—Reader_ID;$C$—事件数量;Timer—计时器

Hash 主表中的每一个位置指向一个主链结点,每一个主链结点包含当前的事件类型、事件数量、计时器以及其第一个子链结点的地址。本研究用链地址法将具有相同事件类型的 RFID 事件构成一个双向链表,通过主链结点的地址可以找到所有链表中的子节点,其流程图如图 5 所示。

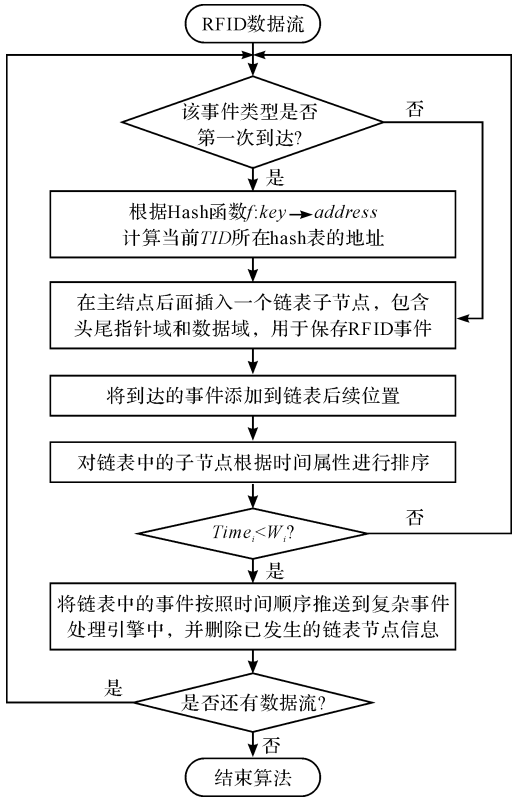


图 5　RFID 乱序事件修正方法流程

## 3　实　验

### 3.1　实验环境与参数设置

实验采用 C#语言在 win764bit、内存为 8 G 的操作系统上实现,标签数据通过多线程随机产生.

读写器检测模型如图 6 所示[16-17]。

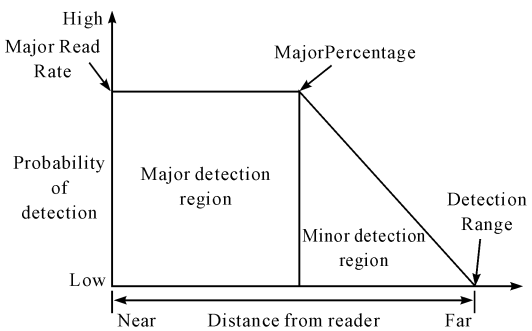


图 6　读写器检测模型

随着标签与读写器距离的变化,相关读写器读取参数也会随之改变,其中横坐标表示标签与读写器之间的距离,纵坐标表示读写器检测到标签的概率。该模型检测区域主要分为 Major detection region 和 Minor

detection region 两部分。当标签与读写器之间的距离超出 DetectionRange 时,读写器将检测不到标签信号,检测概率随着距离的增大而线性递减。实验中设定如下指标验证算法的性能:

(1)记录数。经过冗余数据清洗后的标签存储数量;

(2)正确率。经过乱序修正后可以正确输出的事件流比例。

实验中冗余清洗和乱序修正实验的参数分别如表 1、表 2 所示。

**表 1 冗余清洗实验参数**

| 参数名字 | 参数值 |
|---|---|
| 标签数量 | [100,200,300,400,500] |
| 读写器数量 | 2 ~ 4 |
| 违法编码标签数量 | 5% ~ 20% |
| 最大读取率检测范围 | 0.81 m ~ 3 m |
| 标签移动速度 | 随机变化 |
| 检测概率 | 可变 |

**表 2 乱序修正实验参数**

| 参数名字 | 参数值 |
|---|---|
| Hash 表长度 | 5 ~ 10 |
| 事件类型数量 | 3 ~ 5 |
| 事件查询表达式 | A - > B - > C |
| 查询表达式长度 | 3 |
| 事件流数量 | $1 \times 10^4$ |
| 乱序百分比 | 0 ~ 60% |

### 3.2 实验结果分析

为了验证本研究所提方法的有效性,实验中采用记录数和准确率来分别对比冗余清洗和乱序修正之后的数据。实验结果如下:

(1)记录数指标测试结果如图 7 所示。

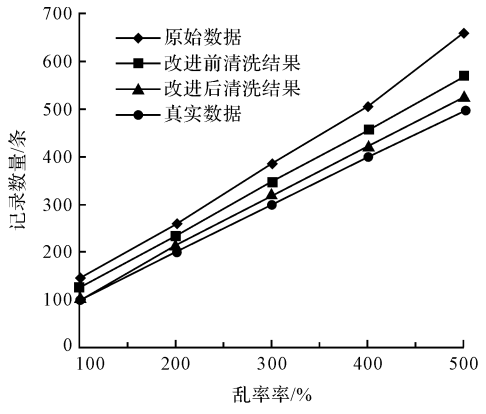


图 7 冗余清洗前后数据对比

记录数量越小表明冗余清洗结果越好。为了对比算法改进后的效果,本文与文献[18]中的 SNM 算法进行对比。标签数量在 100 ~ 500 之间,图中真实数据表示理想状态下标签应该产生的记录数,其一般等于实际的标签数量。

原始数据表示标签实际环境应该产生的数据量,其一般大于真实数据。改进前清洗结果表示文献[18]中 SNM 算法清洗后的标签记录数量。改进后清洗结果表示经过本研究提出算法清洗后的标签记录数量,其越接近真实数据代表算法的清洗效果越好。

通过实验可以发现,在标签数量较少时文中所提算法和 SNM 算法清洗效果相差不大,但是随着标签数量的增加,本研究所提算法具有更好的数据清洗效果,也验证了该算法的有效性。

(2)正确率指标测试结果如图 8 所示。

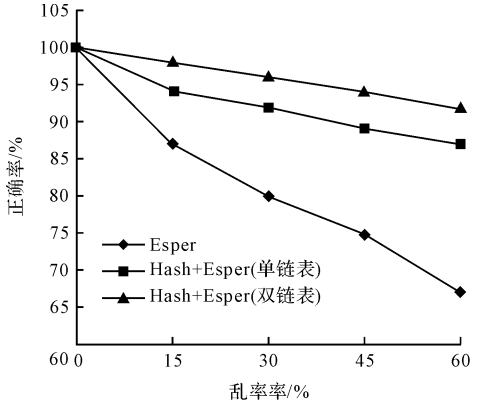


图 8 乱序事件处理前后对比

正确率越高表明乱序事件修正的效果越好。实验中随机生成的乱序比例在 0 ~ 60% 之间,本研究采用开源复杂事件处理引擎 Esper 来进行事件的匹配。图中,Esper 代表的结果是没有经过乱序修正的输出结果,Hash + Esper(单链表)和 Hash + Esper(双链表)分别表示方法改进前后的输出结果。

将未经处理的事件发送到复杂事件处理引擎后,由于事件流的乱序现象会导致原本符合条件的事件不能被匹配到,而通过 Hash 算法进行处理后,可以对原有的乱序事件进行修正。从图中可以:看出采用双链表结构可以明显提高正确匹配的事件数量,符合本研究所提方法的预期效果。

## 4 结束语

本研究结合 RFID 仓储环境中的数据问题设计了面向仓储的 RFID 数据清洗模型,采用不同的实验参数进行了仿真实验。结果表明:当标签数量和乱序率

增大时,通过模型中的数据清洗方法不仅可以有效降低 RFID 仓储中的冗余数据,还可以提高事件输出的正确率。

## 参考文献(References):


[1] LIU H L, CHEN Q, LI Z H. Optimization Techniques for-RFID Complex Event Processing[J]. **Journal of ComputerScience and Technology**,2009,24(4):723-733.

[2] 张康益,薛继军,王　锐. 基于 RFID 技术的备板备件管理[J]. 兵工自动化,2015,34(5):26-28.

[3] MASSAWE L V, VERMAAK H. Reducing false negative reads in RFID data Streams using an adaptive sliding-window approach[J]. **Sensors**,2012,12(4):4187-4212.

[4] VIJAYARAMAN B S, OSYK B A. An empirical study of RFID implementation in the warehousing industry[J]. **The International Journal of Logistics Management**,2006,17(1):6-20.

[5] MING K L, BAHR W, LEUNG S C H. RFID in the warehouse: a literature analysis(1995-2010) of its applications, benefits, challenges and future trends[J]. **Production Economics**,2013,145(1):409-430.

[6] YANG D Y, ZOU P. Event dirven RFID reader for warehouse management[C]. Proceedings of International Conference on Parallel and Distributed Computing, Applications and Technologies, Los Alamitos: IEEE Computer Society, 2005.

[7] 贾红梅,李文杰. 面向仓储管理的 RFID 数据过滤模型研究[J]. 计算机应用与软件,2014,31(2):75-76.

[8] ZHANG Q, CHENG G, WANG Z, et al. Development of RFID application system in cargo inbound and outbound [C]. IEEE Region 10 Conference 2009, New York: IEEE, 2009.

[9] ZEIMPEKIS V, MINIS I, PAPPA V. Real-time logistics man-agement of dried figs using RFID technology: case study in a greek cooperative company[J]. **International Journal of Logistics Systems and Management**,2010,7(3):265-285.

[10] ZHANG C J, YAO X F, ZHANG J M. Abnormal condition monitoring of workpieces based on RFID for wisdom manufacturing workshops[J]. **Sensors**,2015,15(12):30165-30186.

[11] 李博涵,李东静,王学良,等. 多情境感知环境下 RFID 复合事件检测算法[J]. 南京航空航天大学学报,2015,47(3):413-420.

[12] 曹　原,刘英博,肖　利,等. 状态监测数据流时间乱序问题建模与研究[J]. 计算机集成制造系统,2013,19(12):2961-2967.

[13] LEE M L, LING T W, LOW W L. IntelliClean: a knowledg-ebased intelligent data cleaner[C]. Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Ming, Boston: ACM press,2000.

[14] 王友俊. RFID 技术的应用与发展趋势[J]. 煤炭技术,2011,30(6):220-220.

[15] 刘海龙,李战怀. 基于 ENFA 的乱序 RFID 复杂事件检测算法[J]. 华中科技大学学报,2010,38(1):26-20.

[16] JEFFERY S, GAROFALAKIS M, BERKELEY U C, et al. Adaptive cleaning for RFID data streams[C]. Proceedings of the 32d International conference on Very Large Data Bases, Seoul: ACM,2006.

[17] 高　阳,李　坤,单　静. 基于射频识别的滤棒存储输送控制系统的应用研究[J]. 包装与食品机械,2016(6):43-45.

[18] 陈旭辉,王　馨,柯　铭. 一种改进的基于 RFID 中间件的冗余数据清洗算法[J]. 微电子学与计算机,2013,30(7):155-158.


[编辑:周昱晨]


**本文引用格式:**

柴文超,汤洪涛,吴光华. 面向仓储的 RFID 数据清洗技术研究[J]. 机电工程,2017,34(12):1474 - 1479.

CHAI Wen-chao, TANG Hong-tao, WU Guang-hua. RFID data cleaning technology on warehouse-oriented[J]. Journal of Mechanical & Electrical Engineering, 2017,34(12):1474 - 1479.