第41卷 第6A期 计 算 机 科 学 Vol. 41 No. 6A
2014年6月 Computer Science June 2014

一种基于动态标签的RFID不确定性数据清洗算法

王万良[1] 顾熙仁[1] 赵燕伟[2]
(浙江工业大学计算机科学与技术学院 杭州 310023)[1]
(浙江工业大学特种装备制造与先进加工技术教育部重点实验室 杭州 310012)[2]

摘 要 基于自适应滑动窗口清洗算法SMURF(Statistical sMoothing for Unreliable RFid data)需要手动输入阈值δ,对于静态标签,δ的取值对平滑结果几乎没有影响;对于动态标签,结果会造成巨大的误差。针对以上的缺点,提出一种基于动态标签的RFID不确定性数据清洗算法DSUMRF(Dynamic tags-based SMURF)。另外,SMURF算法主要考虑RFID不确定性数据的漏读和错读,没有涉及到冗余数据的处理。在DSUMRF算法的基础上,提出一种RFID冗余数据清洗框架。对比实验表明,针对动态标签,DSMURF算法具有更好的性能。

关键词 动态标签,RFID,不确定性,数据清洗
中图法分类号 TP391 **文献标识码** A

RFID Uncertain Data Cleaning Algorithm Based on Dynamic Tags

WANG Wan-liang[1] GU Xi-ren[1] ZHAO Yan-wei[2]
(College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)[1]
(Key Laboratory of Special Purpose Equipment and Advanced Manufacturing Technology of Ministry of Education, Zhejiang University of Technology, Hangzhou 310012, China)[2]

Abstract SMURF(Statistical sMoothing for Unreliable RFid data) algorithm based on adaptive sliding-window needs to set a threshold manually. The value of has no effect on the smoothing results to the static RFID tags, but it will cause errors when the tags are dynamic. To solve the shortcomings above, the paper proposed DSUMRF(Dynamic tags-based SMURF) algorithm based on dynamic tags. Above all, SMURF algorithm takes major considerations of lost reading and misreading, it is not related to the processing of redundant RFID data. This paper proposed a framework of redundant data cleaning based on DSMURF algorithm. The results of the experiments show that DSMURF performs better than SMURF to the dynamic tags.

Keywords Dynamic tags, RFID, Uncertain, Data cleaning

1 引言

射频识别(RFID)是一种非接触式的自动识别技术,阅读器通过射频电磁波与标签通信以捕获标签中记录的信息,从而达到识别和跟踪标签物品的目的。与传统的条形码识别技术相比,RFID具有扫描快速、体积小、形式多样化、穿透力强、数据的记忆容量大及安全性好等特点。一个典型的RFID系统包括标签、阅读器、中间件以及应用系统,如图1所示。

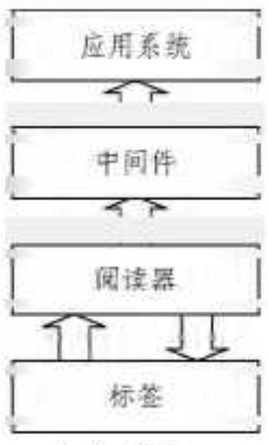

图1 一个典型的RFID系统

当标签接收到阅读器发出的射频信号,就能凭借感应电流所获得的能量发送出存储在芯片中的产品信息。阅读器读取标签信息并解码后,送至计算机系统进行有关数据处理。由于硬件设备固有的限制和环境噪声的影响,RFID数据存在不确定性特点,主要表现在阅读器存在漏读、错读和冗余读等现象。(1)漏读:一个标签位于一个阅读器的探测区域内,但是该阅读器根本没有读取到该标签的信息。造成漏读的原因主要是射频信号碰撞以及周围金属、水等环境的影响。(2)错读:一个标签不在阅读器的探测区域内,但是该阅读器却读取到该标签的信息;阅读器发送错误的EPC编码或者阅读器解码出错导致信息错误等。(3)冗余读:冗余读可以分为时间冗余和空间冗余。一个标签长时间停留在一个阅读器内,那么该阅读器就会不断地对标签进行读取,从而导致大量的冗余信息,称为时间冗余;一个空间区域被多个阅读器所覆盖,那么位于阅读器交叉区域的标签就会同时被多个阅读器读取,称为空间冗余。

据统计,原始RFID数据的准确率仅为60%~70%[1],因

本文受国家自然科学基金项目(61070043)资助。
王万良(1957—),男,博士,教授,博士生导师,主要研究方向为智能自动化、数字媒体、无线传感器网络等,E-mail:wwl@zjut.edu.cn;顾熙仁(1989—),男,硕士生,主要研究方向为RFID数据管理和挖掘;赵燕伟(1959—),女,博士,教授,博士生导师,主要研究方向为物流系统智能配送与优化调度、数字化产品现代设计等。

此需要对原始数据进行清洗后才能提供给上层应用。近年来，RFID 不确定性数据清洗技术已经有了很多研究成果。文献[2]提出了一种基于时间粒度和空间粒度的 ESP 清洗机制，通过描述性的查询语句对 RFID 不确定性数据进行处理，该机制基于管道结构被设计为 5 个步骤。ESP 可以根据各类型脏数据的特点，清洗来自不同接收器的数据，但是时间粒度和空间粒度的设置存在一定的困难。文献[3]以机器学习为背景，提出一系列基于 RFID 大规模数据集的时序数据清洗策略，并且介绍了基于动态贝叶斯的清洗方法，通过阅读器历史的观测结果来估算标签下一次可能出现的概率，该算法的准确率受历史数据的质量影响较大。文献[4]提出一种基于统计学习和工作流建模的数据挖掘方法，该方法可以用于简单情况下的以数据为中心的清洗，但没有考虑 RFID 场景的漏读率等重要因素，并且该算法基于数据库，而不是针对 RFID 数据流。文献[5]提出基于定长滑动窗口的 RFID 数据清洗算法，并且在此基础上提出了一种保持结果输出有序的 RFID 清洗策略，但是定长滑动窗口的大小难以确定。文献[6]提出了一种自适应滑动窗口算法 SUMRF，该算法将 RFID 数据流看成是阅读器范围内标签的随机抽样，通过标签的阅读率动态地调整滑动窗口的大小，从而有效地对数据进行清洗。文献[7]提出一种基于伪事件的数据清洗算法，将标签的冗余读看作伪事件，通过设定时间阈值的方法来判断是否为伪事件，丢弃伪事件数据，一定程度上解决了冗余读的问题。文献[8]提出 3 种基于动态概率路径事件模型的数据填补算法，通过挖掘已知的区域事件的顺序相关性来对后续发生的事件进行判断和填补，填补准确率较高。文献[9]提出一种基于卡尔曼滤波的清洗算法，该算法丢弃了滑动窗口而采用卡尔曼滤波方程，通过时间更新方程和测量更新方程进行自回归逼近真实值，从而达到清洗的目的。该方法较好地解决了错读和漏读的问题。

2 SMURF 算法原理及其局限性

SMURF 由 Shawn R. Jeffery、Minos Garofalakis 等人提出[6]。算法改进了定长滑动窗口大小难以确定的缺点。设置滑动窗口的大小需要考虑两个因素：数据的完整性和标签的动态性。数据的完整性是指要将滑动窗口设置得足够大，从而保证能够尽可能多地读到标签信息；标签的动态性是指由于标签会随着物体位置的改变而出现在不同的阅读器内，因此又不能将滑动窗口设置得过大。

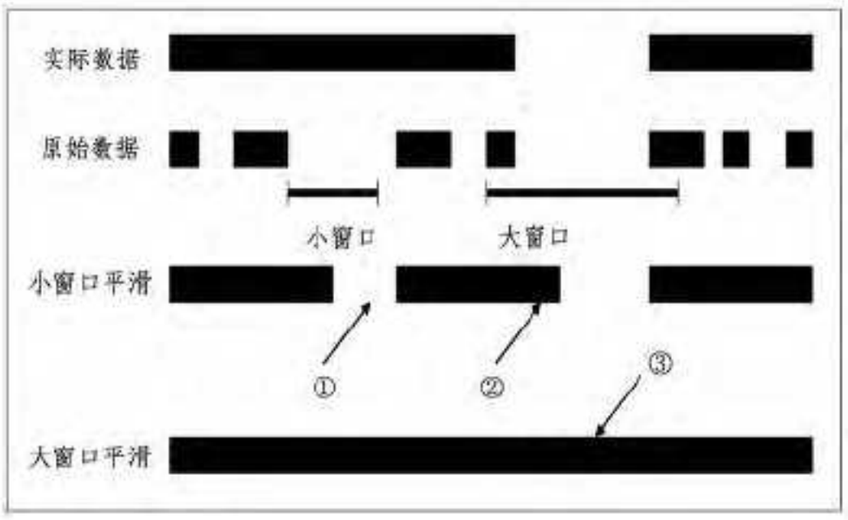

图 2 窗口设置对数据平滑的影响

如图 2 所示，观察实际数据发现数据流间断了一段时间，表明标签在这段时间不在阅读器的阅读范围内，由于射频信号碰撞或者环境的影响，原始数据不能达到理想情况下的实际数据。原始数据通过小窗口平滑出现了①和②两种情况，①表示漏读，②表示多读。所以窗口设置太小不能保证数据的完整性，出现了漏读的情况。而原始数据通过大窗口平滑出现了多读的情况，如上分析，在③这个时间段，标签已经不在阅读器的阅读范围之内，不应该有数据流产生，所以窗口设置太大不能保证标签的动态性。

2.1 标签阅读率影响因素

SMURF 算法涉及到标签的阅读率，阅读率为：$p_{i,t}=\frac{responses}{requests}$，即标签 i 在一个时隙内的阅读率为标签回应的次数和阅读器询问的次数的商。通常一个时隙范围大约为 0.2～0.25 秒。

例：如表 1 所列，为了更好地理解，简化了 Tag ID 的 EPC 编码，以及将 Timestamp 精确到毫秒级。假设时隙为 0.2 秒，一个时隙内阅读器发出 10 个 respects，即阅读周期为 0.02 秒。

表 1 RFID 数据元组

Tag ID	Reader ID	Timestamp
4328	25	12:05:06:0010
4328	25	12:05:06:0050
4328	25	12:05:06:0110
3246	25	12:05:06:0400

标签 4328 在 0.1 秒内读取了 3 次，所以阅读率为 60%。阅读率除了受硬件影响，还受到很多其他因素制约，例如，文献[10]给出了距离和阅读率的关系。除此之外，阅读率还跟天线与标签的角度以及标签的速度有关。为了探讨阅读器与天线角度和标签速度的关系，建立实验平台，采用 RFS2610 阅读器和 RF 软标签。实验距离 1 米，总共 50 个标签，数据点截取 10、30、50 个标签，如图 3 和图 4 所示。

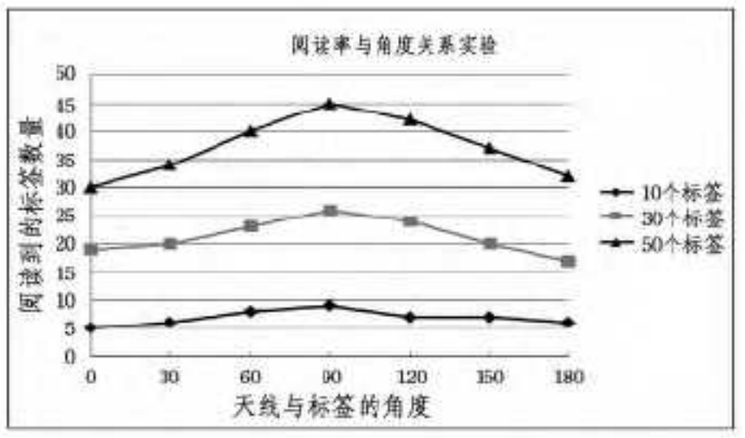

图 3 阅读率与角度的关系

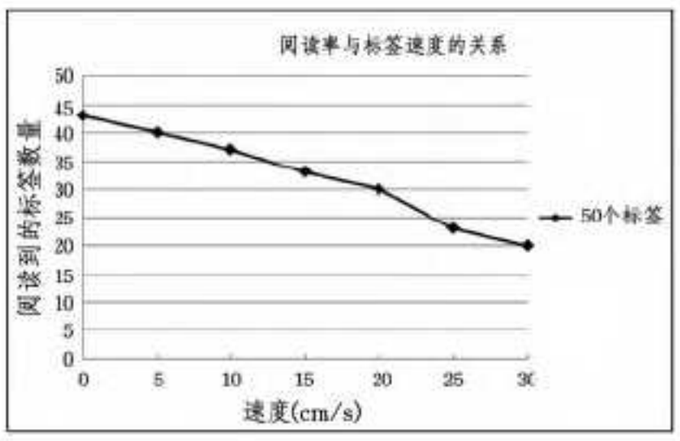

图 4 阅读率与标签速度的关系

从图 3 可以看出天线与标签垂直时，阅读率最大。从图 4 可以看出阅读率与标签的速度密切相关，速度越大，阅读率越低。

2.2 SMURF 算法原理及其局限性

SMURF 算法将观测到的 RFID 数据看作是标签群体的随机样本。假设滑动窗口的大小为 w_i 个时隙，则滑动窗口的范围为 $W_i=(t-w_i,t)$。在每一个时隙标签被阅读到的概率

为 p_i，SMURF 将每一个时隙看作一次伯努利实验，则在滑动窗口内标签被读到的数量服从二项式分布的随机变量，设标签被阅读到的时隙的集合为 $S_i \subseteq W_i$，则标签的平均阅读率为

$$p_i^{avg} = \sum_{t \subset s_i} p_{i,t} / |S_i|$$

为了保证数据的完整性，滑动窗口的大小需要满足：

$$w_i \geqslant \frac{\ln(1/\delta)}{p_i^{avg}} \tag{1}$$

为了保证标签的动态性，滑动窗口的大小需要满足：

$$||S_i| - w_i p_i^{avg}| > 2 \cdot \sqrt{w_i p_i^{avg}(1-p_i^{avg})} \tag{2}$$

SMURF 算法改进了定长滑动窗口大小难以确定的缺点，在静态标签下取得了较高的准确率，但是该算法仍然存在以下不足：

①算法需要输入阈值 $\delta(0<\delta<1)$，对于静态标签，δ 的值对平滑的结果没有太大影响，但是对于动态标签，δ 的取值将会对结果准确率产生一定影响。

②SMURF 算法提供一种自适应的窗口设置方法，极大地减少了漏读和错读的产生，但是算法没有涉及对冗余数据的操作，如表 1 所列，标签 4328 在一个阅读器范围内被读了 3 次，如果标签在阅读器范围内逗留时间更长，将会产生更多的冗余数据。

3 DSMURF 算法

鉴于 SMURF 算法的局限性，对原算法进行了改进。针对局限①，本文提出了基于动态标签的 DSMURF 算法(Dynamic tags-based SMURF)。针对局限②提出了一种 RFID 数据冗余清洗框架。

3.1 DSMURF 算法

对于静态标签，δ 的取值对结果准确率没有太大影响，而对于动态标签，其频繁地进出阅读器的阅读范围，尤其是进入和离开的时候，阅读率特别地低，并且根据前面的实验得知，速度越快，阅读率越低。过低的阅读率导致滑动窗口设置得过大，从而造成错读、多读的概率增大。图 5 所示为阈值 δ 分别对静态标签和动态标签的影响。

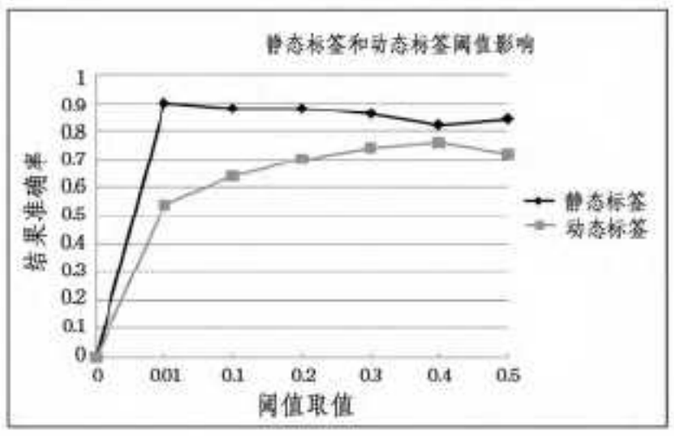

图 5 δ 对静态标签和动态标签的影响

对于静态标签，δ 的取值只要不是太小($\delta<0.01$，因为随着 δ 的值越来越小，滑动窗口会越来越大，导致算法失效)或者太大($\delta>0.5$，因为大于 0.5，滑动窗口则小于 1，没有任何意义)，对准确率基本没有太大影响。而对于动态标签，δ 的值越小，准确率越低，因 δ 越小，根据式(1)知道窗口就设置得越大，而动态标签的阅读率很低，本来就需要大窗口来保证数据的完整性，随着 δ 的值越来越小，在式(2)的范围内，滑动窗口需要设置得越来越大，导致更多的错读。

基于以上分析，需要动态地改变阈值 δ 的大小以提高结果的准确率。设一个标签的速度为 V，阅读器的时隙为 T，阅读器的通信范围半径为 R。则标签被读取的次数是 $f=2R/(V*T)$，根据上面的实验得知参数 δ 与阅读率成正比，与读取次数成反比。所以标签没有被读取的概率为：

$$(1-P_i^{avg})^w < (P_i^{avg}/f)$$

两边取对数：

$$w_i * \ln(1-P_i^{avg}) < \ln(P_i^{avg}/f)$$

因为 $1-P_i^{avg}<1$，所以有：

$$\ln(1-P_i^{avg}) < 0$$

$$\ln(1-P_i^{avg}) < -P_i^{avg}$$

所以有：

$$w_i > \ln(f/P_i^{avg})/P_i^{avg} \tag{3}$$

通过式(3)，便可以处理因动态标签造成的窗口过大问题。

3.2 冗余数据清洗框架

如表 1 所列，数据流存在时间冗余数据，如果不处理冗余数据而直接存放在数据库中，将会造成严重的空间浪费。所以提出了一种冗余数据清洗框架，如图 6 所示。

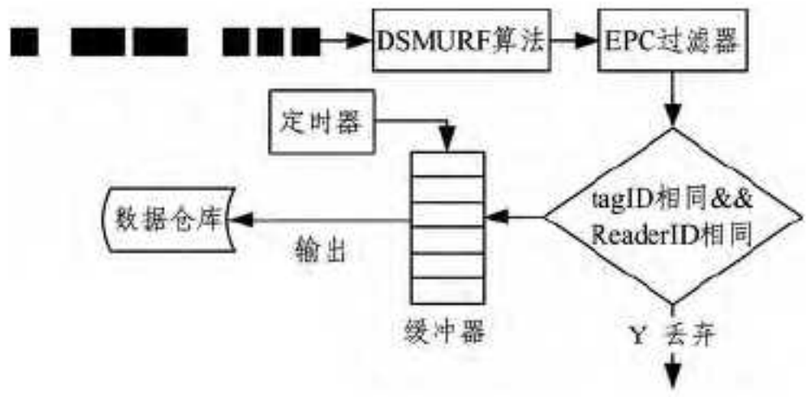

图 6 冗余数据清洗框架

其中 EPC 过滤器过滤因为阅读器等硬件设备产生的标签 ID 不符合 EPC 编码规则的数据元组。原始数据流首先通过 DSMURF 算法进行平滑过滤，再通过 EPC 过滤器，过滤掉编码错误的数据，然后判断新的数据元组是否已经存储在缓冲器中，如果没有，则存入缓冲器中，如果已经存在，则丢弃新元组。缓冲器设置一个定时器，定时器的大小设置为一个滑动窗口的大小，超过时间，则将缓冲器内数据输出到数据仓库。算法的流程图如图 7 所示。

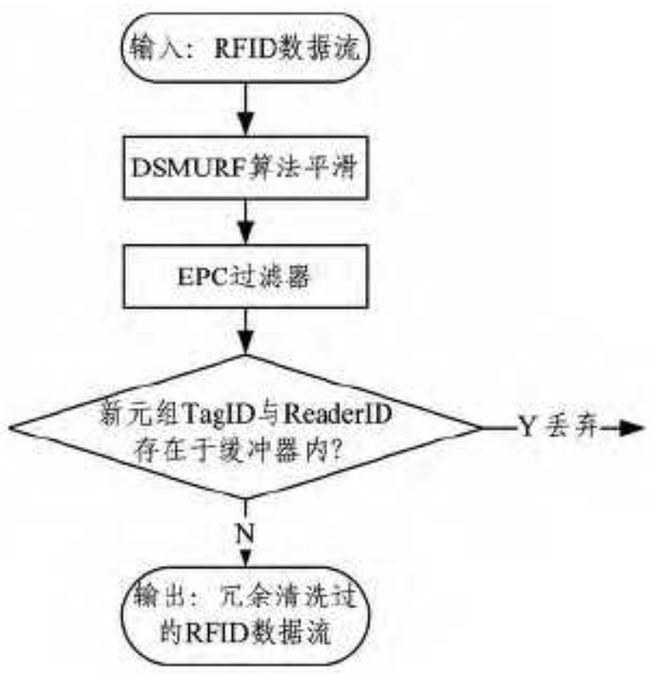

图 7 冗余数据清洗算法流程图

算法判定是否存入缓冲器的条件是新 RFID 元组中的 TagID 和 ReaderID 和缓冲器中的元组是否有重复，即 NewTagID == BufferTagID&&NewReaderID == BufferReaderID。没有重复则存入缓冲器。

4 实验

为了验证算法的有效性，本文采用 RFS2610 射频阅读

器,使用符合 EPC CLASSI G2 标准的 RF 标签,共 50 个标签,每一个标签代表一个物体,让标签以一定的速度通过阅读器。同时,为了验证算法的可扩展性,本文进行了大量的仿真实验,通过如图 8 所示的模型,构造 RFID 数据生成器。通过模型生成器生成大量的仿真数据。

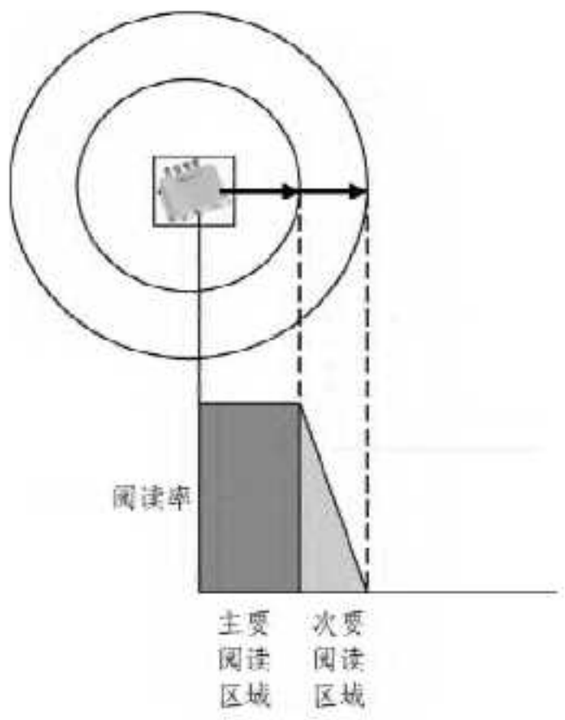

图 8 仿真数据生成模型

4.1 算法准确率比较

运动的标签导致低阅读率,SMURF 算法将会产生过大的窗口,而 DSMURF 算法在保证标签的动态性的情况下,设置合适的窗口大小,如图 9 所示(其中 SMURF 算法 δ 取值为 0.01 和 0.1 两个数据点)。

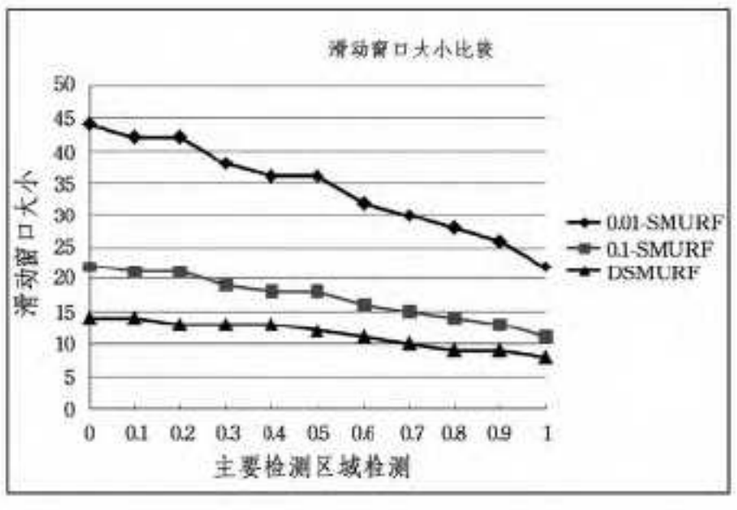

图 9 滑动窗口大小比较

从图 9 可以看出,δ 的值减小 1 个数量级,窗口尺寸大约变成 2 倍,在窗口很大的情况下,δ 的值影响着结果的准确率,而 DSUMRF 算法保证了数据的完整性和标签的动态性,设置了合适的滑动窗口尺寸。

设置标签的数量分别为 10、30、50、100、200、500,比较算法准确率,如图 10 所示。

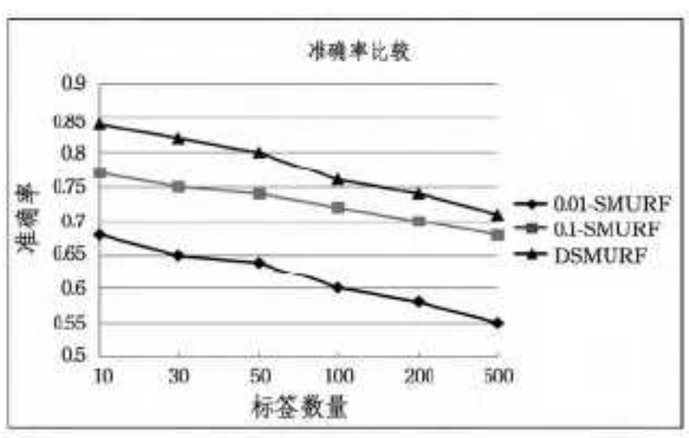

图 10 准确率比较

从图 10 可以看出,针对 SMURF 算法,δ 取 0.1 比 δ 取 0.01 准确率高,说明对于动态标签,δ 的取值对结果有影响。改进后的 DSUMRF 算法取得了较高的准确率,因为 DSMURF 算法有效地设置了滑动窗口的大小,减少了多读、错读的情况,有效地提高了结果的准确性。

4.2 冗余清洗框架结果分析

数据点设置为 100、200、300、400、500、1000 个标签冗余压缩量与原数据量进行对比,如图 11 所示。

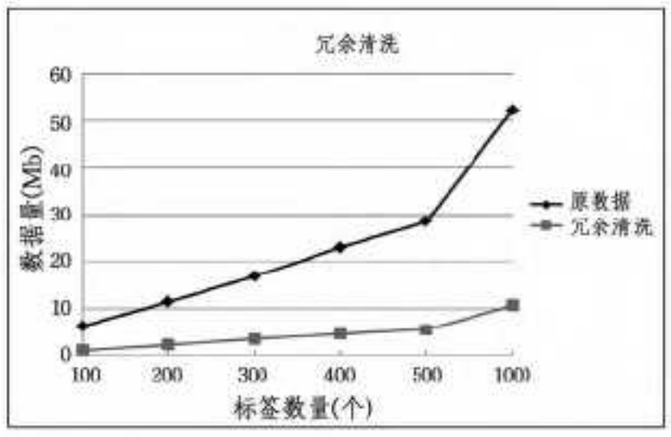

图 11 RFID 时间冗余数据清洗前后比较

从图 11 可以看出,经过冗余数据清洗框架后,RFID 冗余数据大大减少了。

结束语 本文针对 RFID 数据的不确定性,分析了自适应滑动窗口算法 SMURF 的不足之处,提出了基于动态标签的 DSMURF 算法。针对冗余数据,提出一种冗余清洗框架。通过实验分析了阅读率和天线角度以及速度的关系,最后通过部署真实的阅读器环境以及大量的仿真实验验证了 DSMURF 算法和冗余清洗框架的有效性和高效性。

参 考 文 献

[1] Sullivan L. RFID Implementation Challenges Persist, All This Time Later[C]// Information Week. Oct 2005

[2] Jeffrey S R, Alonso G, Franklin M J, et al. A pipelined framework for on line cleaning of sensor data streams[C]// Liu L, Reuter A, et al, eds. Proc. of the 22nd Int'l Conf. on Data Engineering. Atlanta: IEEE Computer Society, 2006: 140-142

[3] Gonzalez H, Han J, Shen X. Cost-conscious cleaning of massive RFID data sets[C]// Proceedings of Internatianal Conference on Data Engineering. ICDE, Istanbul, Turkey, 2007: 1268-1272

[4] Gonzalez H, Han J W, Li X L. Mining compressed commodity workflows from massive RFID data sets[C]// Yu P S, Tsotras V J, eds. Proc. of the 15th ACM Int'l Conf. on Information and Knowledge Management. Arlington: ACM, 2006: 162-171

[5] Bai Yi-jian, Wang Fu-sheng, Liu Pei-ya. Efficiently filtering RFID Data Streams[C]// The First International VLDB Workshop on Clean Databases (CleanDB) Workshop. Seoul, Korea, 2006: 50-57

[6] Jeffery S R, Garofalakis M N, Franklin M J. Adaptive cleaning for RFID data streams[C]// Proceedings of Vary Large Data Bases, VLDB. Seoul, Korea, 2006: 163-174

[7] 王妍,石鑫,宋宝燕. 基于伪事件的 RFID 数据清洗方法[J]. 计算机研究与发展,2009,46(Z2):270-274

[8] 谷峪,于戈,李晓静,等. 基于动态概率路径事件模型的 RFID 数据填补算法[J]. 软件学报,2010,21(3):438-451

[9] 王妍,宋宝燕,付菡,等. 引入卡尔曼滤波的 RFID 数据清洗方法[J]. 小型微型计算机系统,2011,32(9):1794-1799

[10] 马茜,谷峪,张天成,等. 一种基于多阅读器数据冗余的高效 RFID 数据清洗策略[J]. 小型微型计算机系统,2012,33(10):2158-2163