第 34 卷 第 7 期 计 算 机 工 程 2008 年 4 月
Vol.34 No.7 Computer Engineering April 2008
·网络与通信· 文章编号：1000—3428(2008)07—0092—03 文献标识码：A 中图分类号：TP393

RFID 网络的数据清理技术

薛小平[1,2]，张思东[1]，王小平[2]，曹晓宁[1]
(1. 北京交通大学电子信息工程学院，北京 100044；2. 同济大学电子与信息工程学院，上海 200092)

摘 要：结合 RFID 网络数据质量和可靠性研究的最新进展，分类和评述了现有的数据清理技术，分析了平滑和判决方法、流水线方法、基于统计的估计方法、完整性约束的方法等。研究表明，针对不同的应用要求，需要多种数据清理技术的组合才可确保 RFID 阅读可靠性。
关键词：阅读；数据清理；可靠性

Data Cleaning Technologies for RFID Network

XUE Xiao-ping[1,2], ZHANG Si-dong[1], WANG Xiao-ping[2], CAO Xiao-ning[1]
(1. School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing 100044;
2. School of Electronics and Information Engineering, Tongji University, Shanghai 200092)

【Abstract】Combing with the evolution and advances in the data quality and reliability research in RFID network, this paper classifies and analyzes various technologies for improving RFID data reliability and quality, including smoothing and arbitration, pipeline, estimation-based method, integrality-based restraint method, etc. When designing RFID system with high quality for meeting with several requirements, some current data cleaning technologies are properly chosen and fitted together adaptively to improve RFID reading reliability.
【Key words】 reading; data cleaning; reliability

近年来，射频身份标签(RFID)技术正受到普遍关注，并在多个应用领域中取得了成功，如供应链、物流、防伪等。目前，国内外多数的研究集中于解决RFID应用中的一些关键技术，如廉价的标签和阅读器、网络体系结构和阅读器网络的配置等，但作为系统的前端问题，可靠准确的数据流获取是应用成功的首要需求[1-5]。RFID系统因其射频技术无线通信的特点，会受到数据阅读错误的困扰。研究表明，RFID系统仅能正确捕获 60%～70%的标签数据，由阅读错误造成数据的不可靠将给实际应用带来困难。

1 RFID 系统的阅读错误

在 RFID 应用中，标签通常附着在或者内嵌于被识别或跟踪的对象中，其中包含用于标识对象的唯一的标识符(ID)。当标签处于阅读器的识别范围内时，阅读器执行简单的链路层协议来获取标签中的标识符。

定义 1 RFID 阅读(RFID Reading)，指标签在阅读器识读范围内时，阅读器产生的获取标签数据的行为，通常用元组形式表示。只读标签的阅读通常表示为三元组(Reader_ID, Tag_ID, TimeStamp)。可读写标签的阅读通常表示为四元组(Reader_ID, Tag_ID, TimeStamp, Data)。本文将阅读表示为三元组的形式。

普遍认为RFID原始数据流中存在大量的错误[1-5]，引起RFID阅读错误的原因有很多[4],如阅读器和标签的距离过长、阅读和标签间存在障碍物、多标签碰撞、阅读器碰撞、多标签标识同一对象的一致性冲突和人为因素影响等，这些都将使阅读器(或阅读器网络)在其识读范围内没有读到实际存在的标签，或识读了实际上不在阅读器识读范围内的标签。通常阅读错误分为两类：错误阅读(False Positive)，即报告的对象实际上是不存在的；丢失阅读(False Negative)，即没有报告实际存在的对象。

通常，RFID阅读可靠性使用可信度(Confidence)和覆盖率(Coverage)来描述[5]。假设真实数据流R和RFID数据流D，定义：(1)真覆盖集(True Positive, TP)，所有既在R也在D中的阅读的集合；(2)假覆盖集(False Positive, FP)，所有在D中而不在R中的阅读集合；(3)假丢失集(False Negative, FN)，所有在R中而不在D中的阅读集合。

可信度和覆盖率定义如下：

定义 2 可信度指在真实流中每个阅读存在的概率，即对象在指定时间和位置在真实流中存在的概率。理想情况下，可信度对 TP 中的所有阅读是 1；对 FP 中的所有阅读为 0。这种度量可以扩展到整个数据流 D 或任何的阅读子集，这时阅读集合的可信度是每个阅读可信度的平均。

定义 3 覆盖率指在给定的时间周期 T 内，分配给阅读集合的窗口级比值(Windows-Level Value)，即 D 中的阅读占 R 中阅读的比例。对整个流或对任何与 TP,FP,FN 相关的时间窗口来说，有

$$Coverage=\frac{|TP|}{\{TP\bigcup FN\}}$$

可信度和覆盖率描述了 RFID 阅读的质量，可以反映 RFID 数据流与真实数据流之间的关系，并有效评价 RFID 数据质量。

减轻和消除阅读错误的方法可以分为两类：基于软件的数据清理方法和基于 RFID 硬件特性的阅读可靠性改进。前者包括平滑及流水线方法、基于完整性约束的数据清理方法

基金项目：国家自然科学基金资助项目(60572037)
作者简介：薛小平(1963－)，男，副教授、博士研究生，主研方向：网络路由理论，RFID 及传感器网络；张思东，教授、博士生导师；王小平，教授、博士；曹晓宁，硕士研究生
收稿日期：2007-06-02 **E-mail：**xuexp@mail.tongji.edu.cn

等；后者包括抗碰撞算法、天线设计等。本文着重对基于软件的数据清理方法进行分析和比较。

2 基于软件的数据清理方法

2.1 数据清理框架

针对 RFID 系统的阅读错误，文献[6-7]提出了采用流水线数据清理的框架结构以支持普适环境中的应用，称为可扩展的传感器数据流处理(ESP)，它根据阅读器所获得的数据在时间和空间上具有相关性对数据进行处理。数据清理过程是由点(point)处理、平滑(smooth)处理、合并(merge)处理、判决(arbitrate)处理和虚化(virtualize)处理 5 个阶段组成的数据清理流水线，其结构如图 1 所示。ESP 基于阅读器的时空特性来驱动数据清理过程。

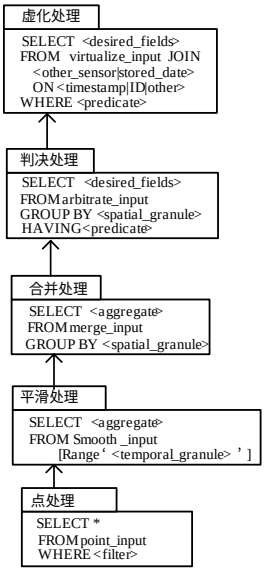
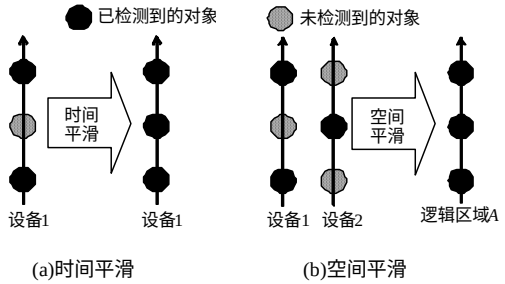


图 1 基于 ESP 流水线的数据清理

点处理阶段对 RFID 数据流中的单个值进行过滤；平滑处理和合并处理则分别由 2.2 节所描述的时间平滑和空间平滑；判决处理阶段协调不同逻辑监测区域的 RFID 阅读器之间的阅读冲突，通常采用应用层逻辑来实现；虚化处理阶段采用时戳的方式组合不同数据源的阅读，提高数据清理质量。判决处理和虚化处理属于高层次清理技术；平滑机制属于低层次清理技术。

2.2 平滑机制及平滑窗口

平滑机制是在时间和空间粒度上补偿阅读丢失、保持 RFID 数据流与真实数据流一致性的有效手段。图 2 给出了时间平滑和空间平滑两种情况。



(a)时间平滑　　(b)空间平滑

图 2 平滑处理技术

2.2.1 平滑技术[7-9]

时间平滑是在时间维度上平滑数据流，多数 RFID 系统采用基本时间平滑算法。算法思想为：如果在窗口大小为 W 的时间内至少有一次阅读，则在窗口结束时刻 t 输出阅读，即如果输入数据流在 $(t-W) \le t' \le t$ 内有阅读(Reader_ID, Tag_ID, TimeStamp)，则阅读(Reader_ID, Tag_ID, TimeStamp)就是时间平滑后的输出值。据此确定对象 O 在时刻 t 时，是否在某个位置，这不仅取决于阅读时刻 t，还依赖于与该对象在时刻 t 相邻的阅读时间窗口 W。主要的改进算法包括：改变窗口的位置，即开始、结束或中心在 t 时刻。时间平滑处理没有消除输入数据流中的阅读，只是增加了额外的阅读量。

空间平滑在空间维度上聚合数据，将监视同一逻辑区域的多个阅读器的阅读数据组合起来。其算法思想为：如果区域中任何一个阅读器报告对象的存在，空间平滑的输出就会含有该对象的阅读信息，即如果逻辑区域中只要存在一个阅读器的输入(Reader_ID, Tag_ID, TimeStamp)，则空间平滑后输出的阅读就是(Reader_ID, Tag_ID, TimeStamp)。

时间平滑和空间平滑均可以单独使用，但不同的平滑措施只能对特定时间和空间有效。在某些应用情况下，采用单独的平滑技术可能达不到预期的目的，需要将时间和空间平滑组合起来使用。由于时间平滑是对单个阅读进行的操作，而空间平滑则是对同一个对象的多个阅读进行的操作，因此当它们组合使用时最好先执行时间平滑，再执行空间平滑。

2.2.2 自适应平滑窗口

典型的 RFID 中间件通常采用固定大小的平滑窗口，但采用固定大小的平滑窗口并不能确保标签阅读的完整性并捕获动态标签。文献[10]基于数理统计的方法，提出了一种自适应的不可靠 RFID 数据统计平滑方法(Statistical sMoothing for Unreliable RFID data, SMURF)。它将 RFID 的数据流模型看成是阅读器识读范围内标签数量的随机样本，基于数据的统计特性连续采用平滑策略的统计方式，提出了按标签清理的二项分布模型和基于参量的多标签清理模型。

(1)单标签清理

单标签情况下，阅读器在大小为w_i个时段的平滑窗口内(记为$W_i=(t-w_i,t)$)检测单标签i。假设在窗口的W_i中标签i一直都处于阅读器的识读范围内，并且在W_i的每个时段中，检测到标签的可能性p_i是相同的；标签i被观测到的次数服从二项分布$B(w_i, p_i)$。通常情况下，假设在窗口W_i的所有时段中，只有部分时段S_i能检测到标签的存在，定义p_i^{avg}为所有检测时段中的平均阅读概率：

$$p_i^{avg} = \Sigma_i p_{i,t} / |S_i|$$

对于完整性标签检测而言，假设 p_i^{avg} 是标签 i 在每个时段中被检测到的概率，如果平滑窗口中的时段个数为

$$w_i \ge \left\lceil \frac{\ln(1/\delta)}{p_i^{avg}} \right\rceil$$

则可以保证标签i在窗口W_i被检测到的概率大于 1-δ。

对于标签动态性检测而言，假设当前窗口的大小是w_i，并且样本概率 p_i^{avg} 较大，并服从中心极限定理：若在当前的窗口中没有任何转移发生，则 $|S_i|$ 在 $\pm 2\sqrt{Var[|S_i|]}$ 的范围内的概率期望值会接近于 0.98。这时，如果检测到的标签i的阅读比期望的阅读小且满足

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

则认为标签 i 发生了转移。

(2)多标签清理

在多标签的情况下，假设窗口 $W=(t-\omega,t)$ 中有w个时段，

每个时段都看成是具有成功概率为 p_i^{avg} 的贝努利试验(Bernoulli试验)，窗口W中标签i至少被读到一次的总概率为 $\pi_i = 1-(1-p_i^{avg})^w$，并对标签总数进行统计。SMURF采用了不等概率的随机样本模型，通过 π 估计器来估计总数量，且消除低估偏差。即用 $S_w \subseteq \{1,2,\cdots,N_W\}$ 定义在窗口W中检测到的标签的集合(N_W)表示真实计数，π 估计值定义为

$$\hat{N}_W = \sum_{i \in S_W} \frac{1}{\pi_i}$$

假设不同标签之间也有独立性，则 $\hat{N}_W$ 的方差可写为

$$\hat{Var}[\hat{N}_W] = \sum_{i \in S_W} \frac{1-\pi_i}{\pi_i^2}$$

基于上述思想，文献[10]提出了单标签和多标签清理时的自适应窗口调整算法。

2.3 基于完整性约束的方法

虽然通过底层的平滑机制以及自适应窗口设置的方法可以消除大部分错误阅读，但到目前为止还没有一种机制可以完全消除RFID系统中所有阅读错误。在底层平滑技术的基础上，根据RFID的应用逻辑约束纠正阅读错误，本文将其称为完整性约束方法。完整性约束可以基于不同的约束条件，如基于单个对象的重量、形状、位置和对象运行路径，也可以基于不同对象之间相互关系，如包含和排斥等。下面给出两种典型的完整性约束方法[11-12]。

(1)伴随约束方法

伴随约束(Accompany Constraint)是指在实际的应用中，如果系统知道多个对象同时移动，那么当检测到其中的部分对象时，则可以断定另一部分对象的阅读丢失。即对象组 $G=\{O_1, Q_2,\cdots,Q_n\}$，当$G$通过阅读器时，获取了一个子集$G'$，并且$G' \subseteq G$。这时阅读器可以应用所确定的伴随约束，补偿出$\{G\}$-$\{G'\}$中的所有对象。

伴随约束是最简单的应用逻辑。在伴随约束中，只要伴随约束关系没有解除，所有的对象都有相互的依赖关系，那么就可以通过这些关系来检测出阅读错误。

(2)基于路由约束的方法

路由约束(Route Constraint)是指在实际的应用过程中，对象往往会沿着指定的路由向目的地移动(如在传送带上的产品)，如果系统知道某个(或某组)对象的运动路由，那么当所获得的路由不符合现实世界中的路由时，就可以断定在某些位置有阅读的丢失。

假设路由约束 $D(V,A)$中没有自环(阅读器对应于 $v \in V$，标签在阅读器之间可能的路由对应于边 $a \in A$)。标签沿着确定的路由$< r_{b1}, r_{b2},\cdots, r_{bk} >$移动，称其为 RFID 标签的实际路径。如果服务器接收的 RFID 标签的路由顺序为$< r_{s1}, r_{s2},\cdots, r_{sl} >, l \le k$，并且任何节点 r_{si} 都包含在实际路径中，则 RFID 标签路由约束例外状态为

$$< r_{s1}, r_{s2},\cdots, r_{sl} > \neq < r_{b1}, r_{b2},\cdots, r_{bk} >$$

在所有被检测到的错误都可以被纠正的理想化条件下，例外状态检测后，可以采用自动或人工的方法来纠错。

3 提高 RFID 阅读可靠性的发展方向

RFID系统阅读可靠性问题是长期困扰RFID应用的关键因素之一。本文讨论了多种提高RFID阅读可靠性的方法和措施，这些措施和方法在一定程度上保证了RFID数据可靠性。然而，到目前为止，还没有任何一种技术可以完全确保RFID阅读的可靠性[1,3]。以下问题有待进一步研究：

(1)阅读器与标签间的可靠通信，包括高性能的标签抗碰撞协议、阅读器配置及抗碰撞算法、天线以及环境适应性研究等，从物理手段上进一步保证 RFID 阅读的可靠和准确。

(2)阅读窗口，固定大小的阅读窗口不能反应标签的动态性和完整性，计算开销小、性能优越的自适应阅读窗口可以在动态环境中保证 RFID 阅读的可靠性。

(3)阅读顺序，由阅读器提交给应用的数据流一般都进行了平滑、判决等处理，而这些处理往往会引起 RFID 阅读的失序，而某些应用需要有序的 RFID 数据流。

(4)纠错和质量估计，从实用的角度看，这两种方法可以组合在一起使用，在纠错的同时，也向应用报告所有阅读的质量。在这种情况下，阅读可以用五元组(Reader_ID, Tag_ID, TimeStamp, Confidence, Coverage)，应用则根据 RFID 阅读中的质量参数，自主地决定是否采用该阅读。

(5)应用差异性[13]，利用应用层逻辑关系纠正RFID阅读中的错误是应用层上常用的方法，但由于应用逻辑各异，对数据可靠性的要求也不尽相同，如在某些应用场合，对RFID阅读的清理采用并不要求很复杂的操作，而某些场合为了保证RFID阅读的可靠性，则要求有十分复杂的数据清理技术，甚至是多种数据清理技术的组合；另一方面，对于可靠性要求高的应用，根据不同的应用逻辑约束及上下文设计合理的面向应用的逻辑约束、例外处理方法等，以充分保证系统的可靠性。

从实用的角度看，单一的数据清理技术无法完全解决 RFID 阅读可靠性问题，多种数据清理技术的组合可能更利于数据阅读可靠性的提高。

4 结束语

本文针对 RFID 系统中的阅读错误及其可靠性进行了研究，分析和总结了现有的 RFID 数据清理技术，并介绍了有关 RFID 阅读质量和可靠性研究的最新进展，从软件算法的角度方面归纳和评述了提高 RFID 数据质量的理论和方法，包括平滑和判决方法、流水线方法、基于统计理论的估计方法等。研究表明：目前 RFID 网络数据质量和可靠性还应侧重于采用综合性的数据清理技术来提高 RFID 阅读的可靠性。

参考文献

[1] Mukhopadyaya S. Data Aware, Low Cost Error Correction for Wireless Sensor Network[C]//Proc. of Wireless Communications and Networking Conference. Hong Kong, China: [s. n.], 2004: 2492- 2497.

[2] Deshpande A, Guestrin C. Model-driven Data Acquisition in Sensor Networks[C]//Proceedings of the 30th International Conference on Very Large Databases. Chicago, Illinois: [s. n.], 2004: 588-599.

[3] Elnabrawy E, Nath B. Cleaning and Querying Noisy Sensors[C]// Proceedings of the 2nd ACM International Conference on Wireless Sensor Networks and Applications. New York, USA: ACM Press, 2003: 78-87.

[4] Floerkemeier C, Lampe M. Issues with RFID Usage in Ubiquitous Computing Application[C]//Proc. of the 2nd International Conference on Pervasive Computing. Paris, France: [s. n.], 2004: 188-193.

[5] Sarma A D, Jeffery S R, Franklin M. Estimating Data Stream Quality for Object-detection Application[R]. EECS Department, University of California, Technical Report: UCB/EECS-2005-23, 2005.

（下转第 97 页）