分流机制下的RFID不确定数据清洗策略

夏秀峰 玄丽娟 李晓明

(沈阳航空航天大学计算机学院 沈阳 110136)

摘 要 无线射频识别技术(RFID)是物联网的关键技术之一。RFID原始数据的不确定性和海量性严重影响了该项技术的发展。通过对不确定数据特征进行分析,建立了一套分流机制下的RFID数据清洗策略。该清洗策略引入清洗队列的概念,根据清洗节点的判断条件选择最佳的清洗路线,无需遍历清洗系统中的所有清洗节点,从而节省了大量的数据传输和清洗等待时间。实验表明,该策略很好地缓解了数据传输压力,有效地提高了数据清洗的效率。

关键词 RFID,物联网,不确定数据,分流机制,清洗策略

中图法分类号 TP311 **文献标识码** A

RFID Uncertain Data Cleaning Strategy under Shunting Mechanism

XIA Xiu-feng XUAN Li-juan LI Xiao-ming

(School of Computer Science, Shenyang Aerospace University, Shenyang 110136, China)

Abstract Radio frequency identification(RFID) is one of the key technologies in Internet of Things. The uncertainty and mass of RFID original data limit the development of technology seriously. By analyzing the uncertain data characteristics, an RFID data cleaning strategy under shunting mechanism was established. The strategy introduces the cleaning queue concept, and according to the cleaning nodes' judgment conditions it can choose the optimal cleaning line. Since it needn't to travel all the cleaning nodes in cleaning system, it saves much time of data transmission and waiting. The experimental results show that the strategy can ease the pressure of data transmission, and improve the efficiency of data cleaning greatly.

Keywords RFID, Internet of things, Uncertain data, Shunting mechanism, Cleaning strategy

1 引言

RFID(Radio Frequency Identification),即无线射频识别,是20世纪90年代提出来的一种自动识别技术。典型的RFID系统由3部分组成,即标签(Tag)、阅读器(Reader)和天线(Antenna)。首先,阅读器通过天线向自身工作区内发送射频信号;然后,进入阅读器工作区的标签接收到射频信号后被自身产生的感应电流激活,通过天线向阅读器返回一个应答信号;最后,阅读器通过对接收到的应答信号进行解码,实现对标签的自动识别和信息的获取。

近年来,随着该项技术的发展,RFID系统已经广泛应用于物流装配、生产制造、交通管理等众多领域。与此同时,RFID技术的发展也带来了很多数据管理方面的问题。与传统的数据形式相比,RFID数据具有简单性、冗余性、语义丰富性和时空关联性[1]等特点,除此之外,RFID数据还有两个重要的特征,即海量性和不确定性[2]。目前,阅读器每秒可采集120~400个标签数据,对于一个含有100个阅读器的中型仓储,每秒可产生1.2~4万条数据,若每个数据占20个字节,则每天可产生1.6GB~6.0GB的数据量[1]。因此,数据量是非常庞大的。然而,在现实情况中,阅读器采集的原始数据的准确率只为60%~70%[3,4],这样低质量的数据使得原始数据无法直接用于上层应用系统。

根据RFID数据的特点,本文在分流机制的基础上构建了一套完整的RFID数据清洗策略。该策略充分考虑了RFID数据的海量性和不确定性,对进入清洗系统的数据进行分流处理,缓解了海量数据传输所引起的拥塞,从而提高了清洗效率。最后,通过实验证明,该策略有效降低了数据清洗的时间消耗,提高了数据清洗的实时性。

2 分流机制下的RFID数据清洗策略

在RFID系统中,由于标签所处环境的不同以及射频信号物理特性等因素,直接导致了阅读器所采集的原始数据不确定性,通常分为积极读、消极读和冗余读3类,如表1所列。

在RFID不确定数据中,积极读和阅读器冗余读数量较少,且比较随机,大多数和标签周围的环境有关;而消极读则是RFID数据中的常见现象,占据了不确定数据中很高的比例。

文献[5]建立了一种可扩展的传感器数据流处理系统ESP。该系统在充分考虑传感器数据特性的基础上,采用滑动窗口机制,引入时间和空间粒度,运用查询语言对传感器数

夏秀峰(1964—),男,博士,教授,主要研究方向为数据库理论与技术,E-mail:xiaxiufeng@163.com;玄丽娟(1986—),硕士生,主要研究方向为物联网数据管理;李晓明(1980—),男,硕士,工程师,主要研究方向为数据库与数据挖掘。

据进行逐层清洗。该框架虽然也适用于 RFID 系统，但仍然无法避免滑动窗口大小难以确定的缺陷。

表 1　RFID 不确定数据定义及描述

不确定数据	定义	特点描述
积极读	是指那些未出现在阅读器的工作区域内，却被阅读器采集到的标签数据。	数量较少，产生的随机性比较大。通常在标签离开或者进入阅读器的时候最容易发生，同时也容易受到周围环境(如水、金属、磁场等)的影响。
消极读	是指那些出现在阅读器的工作区域内，但未被阅读器采集到的标签数据。	漏读数量非常庞大，属于 RFID 数据当中最常见的现象，比例高达 30%左右，也是数据清洗工作的重点。
冗余读	分为阅读器冗余和数据冗余。前者是指一个标签在同一时刻至少被两个阅读器读到；后者是指一个标签在一个阅读周期内被同一个阅读器多次读取，从而出现大量重复记录。	阅读器冗余多发生在两个阅读器交叉重叠的区域，数据冗余在 RFID 系统中数据量较大，对数据的存储和压缩是个巨大挑战。

目前，国内外专家学者多专注于清洗算法[6-10]的设计，而忽略了不确定数据本身的特征，有关 RFID 数据的清洗策略更是涉及甚少。由于 RFID 不确定数据的特征各异，无法实现对数据的一次性清洗，因此，对不确定数据的分层清洗是绝大多数研究者的研究基础。本文提出了一种分流机制下的 RFID 数据清洗策略，即通过对数据进行检测和条件判断，将数据分流到最佳的清洗节点。

2.1　问题描述

在阅读器采集的原始数据中，数据形式是简单的三元组结构：$\langle EPC, Reader, Timestamp\rangle$，即阅读器 *Reader* 在时刻 *Timestamp* 读到了标签 *EPC*(电子产品代码)。为了更好地描述清洗过程，本文将做如下几个定义。

定义 1(交互，Interaction)　交互为阅读器与其工作区内标签之间的一次通讯过程。即阅读器发送一次射频信号，标签收到射频信号后做出一次应答，记为 $IC_{i,j}^{k}$ ($k=1,2,3,\cdots,n,n\in\mathbf{N}^*$)。

其中，i 表示标签的编号；j 表示阅读器的编号；$IC_{i,j}^{k}$ 指的是标签 Tag_i 完成了与阅读器 R_j 的第 k 次交互。

定义 2(处理单元，Processing Unit)　处理单元为标签与阅读器之间的有限次交互的集合，是进行原始数据清洗的最小单位。

由于标签有静态和动态之分，因此，处理单元与交互之间也存在着静态和动态两种不同的表现形式，如式(1)、式(2)所示。

$$PU_i^m=\{IC_{i,a}^1, IC_{i,a}^2,\cdots, IC_{i,a}^k,\cdots, IC_{i,a}^n\} \tag{1}$$

式中，i 表示标签的编号，a 表示阅读器的编号。由于 Tag_i 属于静态标签，因此，在一个处理单元内，标签与阅读器之间的相对位置保持不变，即阅读器编号为一个固定的值，如式(1)中的 R_a。

$$PU_j^m=\{IC_{j,b}^1, IC_{j,b}^2,\cdots, IC_{j,c}^k,\cdots, IC_{j,c}^n\} \tag{2}$$

式中，j 表示标签的编号，b、c 表示阅读器的编号。由于 Tag_j 在处理单元内位置发生了变化，即运动轨迹发生跃迁，因此在一个处理单元内，存在着不同的阅读器编号，如式(2)中的 R_b 和 R_c。通常，将 R_b 称为前驱阅读器，R_c 为后继阅读器。

在式(1)、式(2)中，m 表示标签的第 m 处理单元；k 代表在处理单元内，标签的第 k 次回应($k=1,2,3,\cdots,n,n\in\mathbf{N}^*$)；$n$ 表示一个处理单元所包含的标签与阅读器之间的交互次数。

为了能够以处理单元的形式对数据进行清洗，本文将对 RFID 原始数据存储形式进行简单的预处理，基本方法是：当原始数据被交付给中央数据库时，将数据按照标签进行分类，之后按照某一大小(如式(1)、式(2)的 n)构造处理单元并对其进行编号。经过简单的处理之后，提取出 *Time_fir* 和 *Time_last*，并增加 *Count* 属性，最终，数据存储形式如(3)所示。

$$\langle EPC, Reader, Time\_fir, Time\_last, Count\rangle \tag{3}$$

式中，*Time_fir* 为在处理单元 PU 内标签首次回应阅读器的时刻，*Time_last* 为在处理单元 PU 内标签末次回应阅读器的时刻，*Count* 表示标签与阅读器之间的回应次数。若在一个处理单元中，标签属于动态标签，则需对前驱阅读器和后继阅读器分别进行统计。

定义 3(清洗节点，Cleaning node)　清洗节点为清洗系统中对处理单元中的数据进行清洗的环节。在本文提出的清洗策略中，清洗节点分为积极读清洗节点、冗余读清洗节点和消极读清洗节点。

定义 4(清洗队列，Cleaning queue)　清洗队列为进入清洗节点等待被清洗的处理单元的有序排列。根据定义 3，相应地，清洗队列也分为积极读清洗队列、冗余读清洗队列和消极读清洗队列，分别记为 *P_CQ*，*R_CQ*，*N_CQ*。

以积极读清洗队列 *P_CQ* 为例，Tag_i 的第 m、$m+1$ 处理单元和 Tag_j 的第 k、$k+1$ 处理单元都存在积极读现象，则这些处理单元都将进入积极读清洗节点，从而形成积极读清洗队列，如式(4)所示。

$$P\_CQ=\{\cdots, PU_i^m, PU_i^{m+1}, PU_j^k, PU_j^{k+1},\cdots\} \tag{4}$$

式中，i、j 表示标签的编号；m、k 表示该标签的第 m 和第 k 处理单元(其中，$m\in\mathbf{N}^*$，$k\in\mathbf{N}^*$)。

2.2　RFID 数据清洗策略

海量性是 RFID 数据的特征之一。随着 RFID 系统对实时性的要求越来越高，必须尽量减少数据的处理时间，以便将清洗后的数据实时地传送到上层的应用程序中。在 RFID 不确定数据中，只有一部分数据同时含有积极读、消极读和冗余读 3 种现象，大部分数据只含有不确定数据中的一种或者两种，RFID 数据大致分布如图 1 所示。

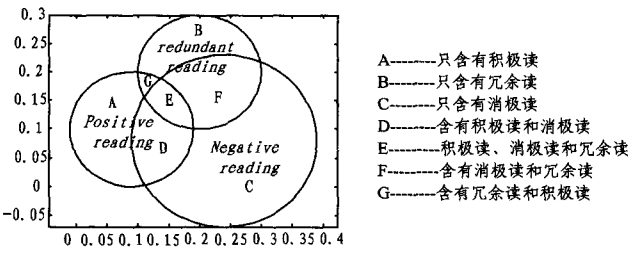

图 1　RFID 数据分布状况

由图 1 可以看出，在 RFID 不确定数据中，比例最大的是消极读，积极读和冗余读占的比例相对较小且两者发生的概率比较接近。如上图所示，交叉区域代表不同种类不确定数据同时存在的比例，其中，冗余读和积极读同时发生的概率最小。

传统的清洗策略是将所有的数据不经过任何检测直接进行清洗，即任何一组数据都会经过清洗系统中的所有清洗环节，因而给数据传输和清洗工作带来巨大的压力，产生了很大的延迟。传统数据清洗模型如图 2 所示。

虽然，传统的清洗策略在一定程度上保证了数据清洗的完整性，但它却没有充分考虑数据自身的比例特征。如果按照传统清洗模型的做法，当数据量比较大时，可能会由于前面数据清洗效率低下而导致后面数据的清洗等待问题，而且在传统的模型中，也没有根据数据特点设计合适的清洗策略，严重影响了清洗的效率。

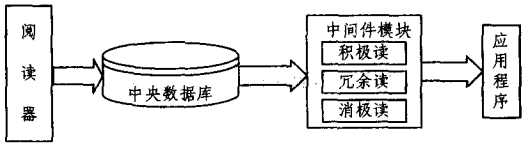

图 2　传统的 RFID 数据清洗模型

针对以上的不足，本文提出了一种分流机制下的 RFID 数据清洗策略(CSUS)，如图 3 所示。

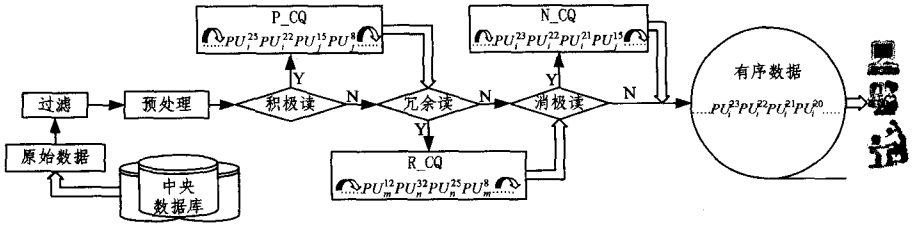

图 3　分流机制下的 RFID 数据清洗策略 CSUS

在 CSUS 中，数据经过简单的预处理后，最终以处理单元的形式进入到清洗系统中。每到一个新的清洗节点，系统会自动检测该处理单元是否符合该节点的清洗条件，如果符合，则进入该节点的清洗队列，等待清洗；如果不符合，则进行下一清洗节点的检测。通过这种方式，系统会将处理单元“分流”到最优的清洗节点，避免了像传统方法那样，每个数据必须要经过清洗系统的所有环节所造成的排队等待。

3　清洗策略(CSUS)的工作流程

分流机制下的 RFID 数据清洗策略 CSUS 的工作流程如下。

Step1　从中央数据库获取原始数据，将其送入过滤层，过滤掉其中的脏数据。

Step2　接收来自 Step1 的数据，对其进行简单的预处理。首先，将数据根据标签编号进行分类；然后，对同类标签按照时间戳 *Timestamp* 对数据进行排序；最后，将有序的数据按照某一大小分成不同的处理单元，并对处理单元进行编号。

Step3　接收来自 Step2 的数据，对处理单元内的数据进行积极读的检测，如果不含积极读数据，则转到 Step5。

Step4　将该处理单元加入到积极读清洗队列中，进行积极读清洗。

Step5　接收两方面的数据：(1)未进入积极读清洗队列的处理单元；(2)经过积极读清洗后的处理单元。进行 Reader 冗余的检测，检测的结果也包括两种：如果不是 Reader 冗余，则转 Step7。

Step6　将该处理单元加入到冗余读清洗队列，进行 Reader 冗余清洗。

Step7　接收进入该清洗节点的处理单元，进行消极读清洗节点的判断，如果处理单元内没有消极读数据，则转 Step9。

Step8　将该处理单元加入到消极读清洗队列，进行消极读数据处理。

Step9　接收所有的处理单元，由于在清洗的过程中会造成处理单元的乱序，因此需要对处理单元进行排序，以保证原始数据全局上的有序性。

Step10　将清洗后的数据交付给上层的应用程序。

RFID 的数据传输和清洗等待所产生的延迟是影响系统运行效率的瓶颈之一，在设计清洗策略时必须尽量减少类似的时间消耗。本文提出的分流机制下的 RFID 数据清洗策略，在保证数据进行有效、完整清洗的前提下，尽量避免数据清洗的等待时间，提高了清洗的效率。由于篇幅有限，对于清洗节点的判定条件及清洗队列采用的清洗算法，本文将不再赘述。

4　试验结果及分析

4.1　实验环境

本节将对分流机制下的 RFID 清洗策略 CSUS 的有效性进行实验验证。实验环境为 Intel Pentium Dual E2180 2.00GHz 处理器、2G 内存，操作系统为 Windows XP2 Professional 平台。编程语言为 Java，数据库管理系统为 Mysql。

4.2　实验数据

本文选择 RFID 模拟数据集作为实验数据源。首先，选取一个随机数生成函数，该函数能够以一个固定的频率产生随机数，这就如同阅读器会以固定的频率向其工作区内发送射频信号一样；然后，对生成的随机数进行检测，删除其中的重复记录，以保证数据的唯一性；最后，根据 RFID 原始数据的实际情况调整数据集，使得数据集中存在 30%左右的不确定数据，即原始数据的准确率控制在 60%～70%。根据产生的随机数据，本文选取了 4 种大小不同的数据集，实验数据如表 2 所列。

如表 2 所列，实验数据中包含 16 个由原始数据记录组成的子数据集，其中 DataSet1—DataSet4 数据量为 1000，DataSet5—DataSet8 数据量为 5000，DataSet9—DataSet12 数据量为 10000，DataSet13—DataSet16 数据量为 20000。

表 2　实验数据

数据集名称	实验集大小
DataSet1—DataSet4	1000
DataSet5—DataSet8	5000
DataSet9—DataSet12	10000
DataSet13—DataSet16	20000

4.3　验结果及分析

为了较为全面地比较传统清洗模型(TRCS)与 CSUS 的时间性能，本文将进行两轮实验进行验证。

首先，当数据集中不确定数据比例相同(噪声比相同)时，一共进行 4 组实验，每组实验中含有 4 个数据集，大小分别为 1000、5000、10000、20000 条原始数据记录。TRCS 和 CSUS 的时间消耗如图 4 所示。

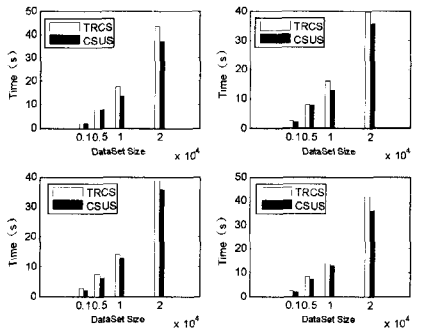

图 4　相同噪声比下 TRCS 和 CSUS 的时间性能比较

在图 4 的分组实验中，不确定数据的比例均控制在 30% 左右。当数据量增大时，两种清洗策略所消耗的时间也随之增加，因此，数据清洗所消耗的时间与数据量大致呈线性关系。另外，随着数据量的不断增大，两种清洗策略在时间消耗上的差距也会随着增大。

其次，当数据量大小相同，不确定数据所占的比例不同(即噪声比不同)时，本轮实验选取 DataSet13—DataSet16 作为实验数据，TRCS 和 CSUS 的时间消耗如图 5 所示。

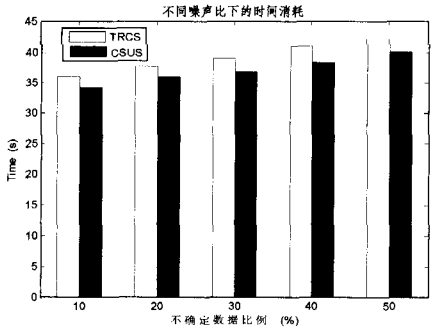

图 5　不同噪声比下 TRCS 和 CSUS 的时间性能比较

由图 5 可以看出，随着数据集中不确定数据比例的提高，即噪声比的不断提高，即使在数据量相同的情况下，数据清洗所用的时间也有所不同。换句话说，数据清洗时间不仅受到数据量大小的影响，而且也受到原始数据噪声比的影响。

通过以上两个实验可以看出，CSUS 在时间性能上比传统模型占有一定的优势，这其中的主要原因是 CSUS 是基于分流机制的，不需要像传统模型那样需要经过清洗系统中的每个环节，而只是选择符合清洗条件的清洗节点，这样在一定程度上就减小了数据传输和清洗的等待时间，从而降低了整个清洗过程的时间消耗。

结束语　本文提出了一种分流机制下的 RFID 数据清洗策略 CSUS，该策略建立在分流机制的基础上，引入了清洗队列的概念，有效解决了数据传输和清洗等待问题。对 CSUS 策略的时间性能进行验证的结果表明，CSUS 策略具有良好的扩展性，在保证清洗准确率的前提下，大大减低了数据清洗的时间消耗，增强了数据清洗的实时性，能够为上层的应用程序提供高效、准确的数据。

参 考 文 献

[1] Gu Y, Yu G, Zhang T C. RFID complex event processing techniques[J]. Frontiers of Computer Science and Technology, 2007, 1(3): 255-267

[2] Xu Jia, Yu Ge, Gu Yu, et al. Uncertain Data Management Technologies in RFID[J]. Frontiers of Computer Science and Technology, 3(6): 561-576

[3] Fishkin K P, Jiang B, Philipose M, et al. I sense a disturbance in the force: Unobtrusive detection of interactions with RFID-tagged objects[C]// Proceedings in Ubicomp. 2004: 268-282

[4] Hahnel D, Burgard W, Fox D, et al. Mapping and localization with RFID technology[C]// International Conference on Robotics and Automation. 2004: 1015-1020

[5] Jeffery S R, Alonse G, Franklin M J, et al. A Pipelined Framework for Online Cleaning of Sensor Data Streams[C]// Proceedings of the 22nd International Conference on Data Engineering (ICDE). Atlanta, Georgia, USA, 2006: 140

[6] Bai Yi-jian, Wang Fus-heng, Liu Pei-ya. Efficiently Filtering RFID Data Streams[C]// The first international VLDB Workshop on Clean Databases (CleanDB) Workshop. Seoul, Korea, 2006: 50-57

[7] Jeffery S R, Garofalakis M, Franklin M. Adaptive Cleaning for RFID Data Streams[C]// Proceedings of the 32nd International Conference Very Large Data Bases (VLDB). Seoul, Korea, 2006: 163-174

[8] Ramanathan C B, Koyutuk M K, Hoffmann M, et al. Redundant Reader Elimination in RFID System, Sensor and Ad Hoc Communications and Networks, 2005[C]// IEEE SECON 2005, 2005 Second Annual IEEECommunications Society Conference. Santa Clara, California, USA, Sep. 2005: 176- 184

[9] 廖国琼，李晶. 基于距离的分布式 RFID 数据流孤立点检测[J]. 计算机研究与发展，2010，47(5)：930-939

[10] 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(Suppl.)：270-274