

RFID 复杂事件处理技术 *

谷 峪, 于 戈⁺, 张天成

GU Yu, YU Ge⁺, ZHANG Tiancheng

东北大学 信息科学与工程学院, 沈阳 110004

School of Information Science and Engineering, Northeastern University, Shenyang 110004, China

+Corresponding author: E-mail: yuge@mail.neu.edu.cn

GU Yu, YU Ge, ZHANG Tiancheng. RFID complex event processing techniques. Journal of Frontiers of Computer Science and Technology, 2007, 1(3): 255-267.

Abstract: With the development of RFID technology, RFID is being applied ubiquitously in every field. With thoroughly processing and analyzing RFID data, more complex events and implicit knowledge could be discovered to effectively support advanced applications such as event monitoring and pre-warning. Due to the speciality of RFID, it is insufficient to take use of the state of art active database and data stream techniques for high-performance RFID event detection and processing. The paper analyzes the characteristics of RFID data, makes a survey of the newest technology over RFID complex event processing, and explores the new problems to be solved urgently, including RFID data cleaning, data-centric event detection techniques, event-centric event detection techniques and complex event processing systems, and finally gives an outlook of the research issues in the future.

Key words: RFID technology; data cleaning; complex event processing; data stream technology

摘 要: 随着 RFID 技术的发展, RFID 应用正无所不在。通过对 RFID 数据的深入处理和分析, 可以发现更复杂的复合事件和隐含知识, 从而有效地支持事件监控、事件预警等先进应用。由于 RFID 的特殊性, 依靠现有的主动数据库技术和数据流管理技术难以实现高效的 RFID 事件检测和处理。分析了 RFID 数据的特点, 归纳和总结了 RFID 复杂事件处理的最新技术, 讨论了一些亟待解决的新问题, 主要有 RFID 数据清洗方法、以数据为中心的检测技术、以事件为中心的检测技术, 以及复杂事件处理系统等, 并对今后的研究重点进行了展望。

关键词: RFID 技术; 数据清洗; 复杂事件处理; 数据流技术

文献标识码: A **中图分类号:** TP391

* the National Natural Science Foundation of China under Grant No.60773220, 60473073(国家自然科学基金).

Received 2007-08, Accepted 2007-10.

1 引言

无线射频识别(Radio Frequency Identification, RFID)技术是一种自动识别和数据获取技术^[1]。RFID的基本工作原理是,首先由阅读器向 RFID 标签发送能量,标签向阅读器返回数据,阅读器解码并向主机返回数据。这种标签也称为无源的被动式标签。高级的 RFID 标签是有源的主动式标签,其感应距离可以更远(超出 10 m),也可以是可读写的,能够将有关数据写回标签。RFID 标签具有低成本、寿命长、不怕污染和适应恶劣环境等特点,有望在将来替代目前流行的条码。国际标准委员会制定的电子产品代码 EPC^[2],为每个产品定义全球唯一的 ID,使每个标签对象携带有唯一的识别码。一个通用的 RFID 应用系统的组成如图 1 所示。其中,RFID 阅读器读取对象的标签数据,该观察值基本形式为: $\langle r, o, t \rangle$,表示阅读器 r 在时刻 t 观察到对象 o 。由于 o 具有全球唯一的 ID,通过这种观察可建立起物理世界与逻辑世界的一对一映射。中间件系统得到底层传来的基本数据后,结合数据库中保存的上下文信息和背景知识,包括 o 的属性、位置、状态等,进行语义上的转换处理,得到可用的业务逻辑数据,最终向上集成到 ERP、CRM 等应用中^[3]。

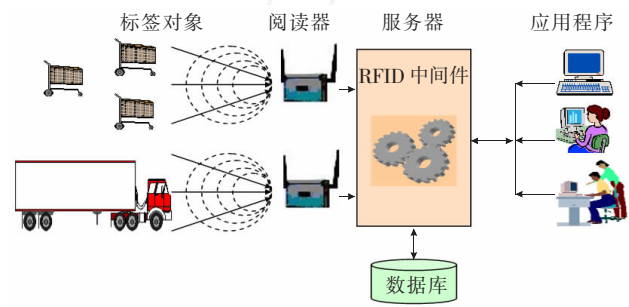


图 1 RFID 应用系统结构

Fig.1 Architecture of RFID application system

RFID 技术的早期应用是为了识别物品、计算数据。例如,在物流配送系统中,当物品带上标签后,可以从一大批物品中快速地追踪和定位目标,并统计数量^[4];在门禁安检系统中,可以快速、准确地识别和统计出入人数^[5]。实际上,RFID 支持事件监控等更高级的应用,这些动态的 RFID 数据的产生代表着某种

基本事件的发生,通过模式匹配、异常发现、关联分析等智能推理手段,可以检测和发现更复杂的事件,即复合事件。例如,在物流配送系统中,可以实时监控物品丢失、配送阻塞等事件;在门禁安检系统中,可以发现异常活动和潜在的不安全因素等;甚至在智能医院中,用于跟踪医生、病人、医疗设备和药品^[6]。因此,实际应用中迫切需要开发支持这些复杂 RFID 应用的事件处理技术。

与普通数据相比,RFID 数据具有如下特点:

(1)时态性、动态性和关联性。现场动态地产生关于标签对象状态变化的观察数据,这些数据包含观察时刻以及对应该时刻的对象位置和对对象状态等数据。例如,物品的入库、出库等状态。一个 RFID 数据不是独立存在的,而是相互关联的,由时态性和动态性衍生出关联性。时态关联表达了事件之间的时序关系,空间关联表达了事件发展的轨迹,时空关联共同表达了与对象有关的事件的变化过程。

(2)语义丰富性。被观察的对象携带有与上下文状态和背景知识有关的信息,这些信息是隐含的,且与上层应用逻辑之间存在密切的关系。利用这些相关信息可进一步导出衍生信息。例如,从产品的 ID 可查出它的型号、价格、产地等,从阅读器的位置可得知物品的存放货架位置等。RFID 数据是一种低级的基础数据,必须上升为高级的业务逻辑数据,与现有的应用相集成,才能真正地发挥作用。

(3)不准确性和异构性。现有的 RFID 阅读器还存在误差问题,如漏读、多读和脏数据等。另一方面,一个阅读器可以识别出多种不同的对象,即 RFID 数据流中可能包含有多种不同性质的观察值。例如,在一个安检入口,既可以识别出进出的人员,也可以识别进出的物品,它们是具有不同性质的对象,所对应的事件语义也不相同。

(4)流特性、批量性和海量性。RFID 数据是以流的形式快速、自动地产生的,需要积累起来以支持跟踪和监控应用。并且,有时具有批量的特点,即多个对象会被集中地观察,例如,当对一个集装箱登记时,同时会读到大批数据。大规模的 RFID 设备的部

署将产生空前的海量数据。目前,阅读器每秒可捕获 120 个到 400 个标签数据。对于一个部署有 100 个阅读器的中型仓储,每秒可产生 1.2~4 万条数据,若每条数据占 20 字节,则每天产生 1.6 GB~60 GB。因此,需要处理的数据量非常庞大。

在早期的 RFID 应用开发中,将 RFID 数据直接传送给应用程序,由应用程序解释这些原始数据,处理成业务逻辑数据。这种方式导致 RFID 处理部分非常复杂,软件的可重用性、可伸缩性和可适应性都非常低。当前的发展趋势是为 RFID 应用提供基于中间件的平台,在 RFID 代表的物理世界与应用软件之间建立桥梁^[7]。这些方案可分成两类:以数据为中心的系统 and 以事件为中心的系统。以数据为中心的系统采用传统的数据库技术,对 RFID 数据进行建模,保存在数据库管理系统中,在数据库基础上支持事件处理;以事件为中心的系统是一种新型的 RFID 处理技术,基于数据源而不是基于数据库对事件进行建模,对事件直接进行检测和处理,以获得高的效率,处理更复杂的事件。由于 RFID 数据本身的不准确性和异构性,在进行检测和处理前需要对数据进行预处理,这种预处理主要基于 RFID 数据清洗相关技术。因此,RFID 复杂事件处理关键技术主要包括数据清洗和事件检测两个方面。

文章首先对 RFID 数据清洗技术进行总结和比较,然后分别介绍以数据为中心和以事件为中心的事件检测技术,并对一些开放的问题进行讨论和展望。

2 RFID 数据清洗技术

在 RFID 系统中,由于射频干扰和标签读取结构等诸多原因,数据读取的可信度较低,这严重影响了 RFID 技术在各领域中的用途(比如精确的存货跟踪,严格的病人监护等),因此,系统中必须增加对 RFID 数据的预处理环节,使阅读器读取来的数据,无论从时间序列还是从数值上,尽可能地接近真实数据。

2.1 数据清洗的内容

RFID 数据读取的低可信度主要与其采用的无

线射频信号相关,它们极易受环境影响,而且相互干扰,尤其当标签和阅读器数量多时,信号干扰加强,但这属于电磁波的物理特性,很难改变,只能从其他方面加以考虑。RFID 源数据的不准确性,主要归结为下面 3 点:

(1)漏读:当一个标签在一个阅读器阅读范围之内时,该阅读器没有读到该标签。这种情况可能发生在同一时间段内,某个阅读器同时读多个标签时,某些标签被遗漏。经调查表明,通常情况下,阅读器只能读到在它感应范围内 60%~70% 的标签数据,即至少有 30% 的标签数据读取时被遗漏。由此可见,在 RFID 应用中,数据漏读现象比较严重,应重点解决。这种错误也称为拒真(false negative)。

(2)多读:指当一个标签在一个阅读器阅读范围之外时,该阅读器仍然读到了该标签。这种情况主要因为多个阅读器同时存在时的电磁波的干扰,随机性很大。这种错误也称为纳伪(false positive)。

(3)脏数据:指一个标签在一个阅读器阅读范围之内时,该阅读器感应到该标签存在,但读出的 EPC 值不正确。

上面总结的 3 点主要是从数据的准确性方面考虑的,除此之外,还有两个方面需要考虑,即时序性和精简性。RFID 应用对数据的时序性要求很高,而数据在传输过程中,乱序现象比较严重,数据应该按时间戳大小有序输出;而精简性是指消除数据冗余,避免造成存储空间浪费和出现二异性。

目前,对 RFID 应用中数据清洗技术的研究已经取得了一定的进展,对上面提出的问题也有了不同的解决方法,但解决问题的类型主要分为两类,即主要只针对漏读现象的解决办法和针对各类问题的适应性解决办法,下面分别介绍。

2.2 数据漏读的清洗策略

目前,针对数据漏读现象,已经有了许多解决方法,如建立概率模型^[8]、建立 SQL 查询模型^[9]以及利用机器学习方法建模^[10]等,下面依次介绍。

大部分的 RFID 中间件系统采用了“平滑过滤器”策略,即设定一个固定窗口大小用以填补漏读数

据,但由于窗口大小固定不变,致使窗口过小会产生拒真数据,窗口过大会产生纳伪数据,如图2所示。由此可见,大小固定的窗口不能根据数据的特点精确地添补数据,美国加州大学 Berkeley 分校则提出了一种改进方法,即基于概率模型 SMURF 的适应性数据清洗方法^[8]。

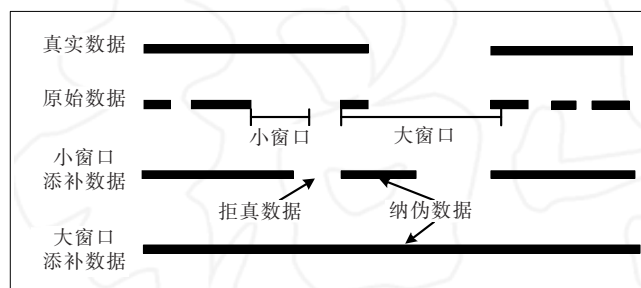


图2 不同平滑窗口大小的影响

Fig.2 Effect of different smoothing-window sizes

该方法主要考虑到了两个因素,完整性和标签动态变化性。完整性是指确保在阅读范围之内的所有标签都能被读到,平滑窗口应该足够大以保证完整性。但是设置一个大的滑动窗口,有可能探测不出窗口内标签移动,即当标签已经离开阅读器的阅读范围之内时,它仍被错误地探测到,这样会产生大量的纳伪数据,不能保证标签动态变化性。由此可见,窗口大小的调整十分关键,既要保证完整性又要保证标签动态变化性。确保完整性的因素主要是阅读器的阅读速率,它主要受阅读器、标签的类型和周围物理环境影响。而确保标签动态变化性的因素主要是标签的移动特性,它主要跟应用类型有关,如放在固定货架和传送带上的两个标签会表现出截然不同的移动特性。

SMURF 模型是在一种新的转换思路基础之上建立起来的,将 RFID 数据流看作统计学中的随机事件,即读出一个标签看作是发生一个随机事件,标签的读出频率看作是事件发生的概率,在平滑窗口内的探测值看作是多次重复随机事件的结果。

在此基础上提出了两个概率模型:二项分布和 π -estimator。

伯努利二项分布模型用来解决单个阅读器探测

单个标签的情况。在平滑窗口内读到标签的次数 N 满足伯努利二项分布 $B(w_i, p_i^{avg})$, w_i 是在平滑窗口内阅读器探测周期 epoch 的次数(窗口大小),即重复试验的次数, p_i^{avg} 是在平滑窗口内标签的平均读出频率,即事件发生概率。当置信度为 δ 时,保证完整性的充分条件是: $w_i \geq \frac{\ln(1/\delta)}{p_i^{avg}}$,当前窗口小于 w_i 时,适

当扩大窗口大小;而探测到标签动态变化时的条件是: $|S_i - w_i p_i^{avg}| > 2 \cdot \sqrt{w_i p_i^{avg} (1 - p_i^{avg})}$,其中, $|S_i|$ 表示在平滑窗口内,探测到某标签的次数,此时适当减小窗口大小。由此根据标签当前信息,适应性调整窗口大小,可以保证较准确的填充数据。

π -estimator 模型用来解决单个阅读器探测多个标签聚合的情况,以准确地探测某个时间段内标签的个数。在平滑窗口内,存在的标签个数之和为 $\hat{N}_w =$

$\sum_{i \in S_w} \frac{1}{\pi_i}$,其中 $\pi_i = 1 - (1 - p_i^{avg})^w$,表示在窗口内标签 i 的

读出概率, p_i^{avg} 是平均标签的平均读出频率。保证完整性的充分条件与二项分布模型的相同。探测到标签动态变化条件是: $|N_w - N_{w'}| > 2(\sqrt{Var[N_w]} + \sqrt{Var[N_{w'}]})$,其中 N_w 和 $N_{w'}$ 分别表示在当前窗口 $W = (t-w, t]$ 和当前二分之一窗口 $W' = (t-\frac{w}{2}, t]$ 内,存在标签的个数之和。针对这两个条件的相应具体操作与解决单个标签的情况类似。

与静态窗口填补算法相比,适应性调整窗口大小的填补算法有着明显的优势。但此概率模型只解决了单个阅读器探测的问题,没有考虑空间布置多个阅读器探测的情况。

基于 SQL 查询模型的方法主要提出了一种 ESP 机制^[9],可以用来清洗来自不同接收器的数据,并且针对各类型错误数据的各自特点进行清洗,RFID 错误数据主要是漏读数据。该机制是一个描述性的查询处理工具,设计成管道结构,便于处理数据和建模,如图3所示,其中每一步带有典型描述性查询。

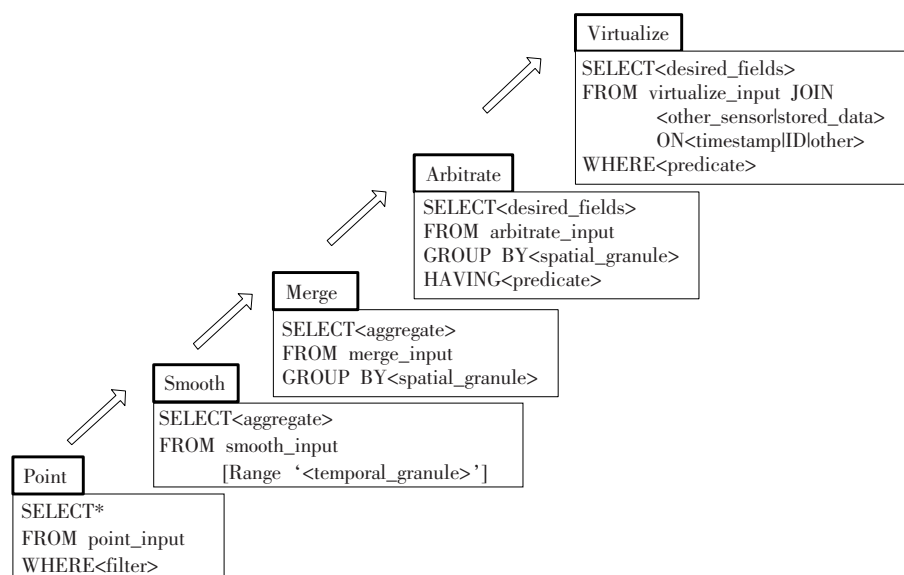


图3 ESP 处理步骤

Fig.3 ESP processing stages

该管道结构由 5 部分连接组成, 分别是 Point, Smooth, Merge, Arbitrate 和 Virtualize。它们是 ESP 机制清洗数据的 5 大步骤, 每个步骤有各自的清洗任务。针对不同的应用, 采取的具体清洗技术也不同, 不一定实现每个步骤。下面进行具体介绍。

Point: 该步骤主要对来自同一个传感器的源数据流中的单个数据进行清洗, 去掉异常数据, 如 RFID 中错误的标签数据、噪声数据, 然后将满足过滤条件的数据输出。

Smooth: 该步骤主要根据数据之间的时间相关性对漏读数据进行填补或探测噪声数据。该方法定义了时间粒度概念, 它是处理数据的时间单位, 如在数据流上定义与之相等的时间窗口 (窗口大小由用户需求决定), 然后在窗口上进行各种聚合操作, 并输出结果。Point 和 Smooth 步骤均可以下推到有处理能力的接收器节点上执行。

Merge: 该步骤主要根据数据的空间相关性对数据进行清洗。该方法定义了空间粒度概念, 它是处理数据的空间单位, 每次将接收到的来自于多个接收器的数据按事先定义好的空间粒度进行分组, 然后逐个处理并输出。

Arbitrate: 多个接收器接收的数据可能会产生冲

突, 例如, 同一数据被位于不同区域的两个接收器同时读到时, 该步骤负责判断应该属于哪个接收器。

Virtualize: 某些应用可能会用到来自不同类型接收器的数据或者已存储的数据。该步骤通过对各种数据的 Join 操作, 提供所要求的数据。连接属性可以是时间戳、ID 或者其他相同属性等。

这 5 个步骤的数据清洗层次逐渐升高, 由对单个接收器数据的清洗, 扩展到对多个接收器的处理; 由对来自同种类型接收器数据的清洗, 扩展到来自不同种类型接收器的处理, 解决数据清洗的范围很广。另外, 该模型处理问题十分灵活, 针对不同的应用, 可采用不同的步骤。

该模型主要根据数据具有的时间相关性和空间相关性对数据进行清洗, 考虑范围比较宽广, 但时间粒度和空间粒度如何设置, 一旦设置可否适应地改变, 如何改变等问题仍在研究中。

上述两种清洗策略的主要衡量标准是准确性, 即清洗后的数据中, 准确数据所占比例。但是当某个 RFID 应用布置规模极大时, 如涉及到数千个阅读器、数万个标签, 这时衡量标准就不能只考虑数据的准确性了, 还要考虑到算法的时间开销问题。因而, 文献[10]提出了以机器学习为背景的清洗算法。它提

出了一个清洗框架,包含大规模的 RFID 数据集和一系列数据清洗策略,并分析了各个策略对应的清洗开销,提了一个适应性调整时间开销和准确性的总体优化算法。

该算法所考虑的开销主要包括三个部分:一是机器学习中,每个元组的训练开销;二是存储开销和运行开销;三是分类错误时所需的开销。具体的清洗框架机制如图 4 所示。

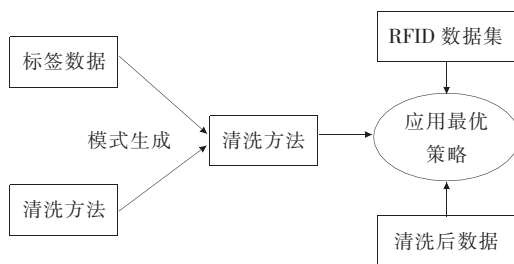


图 4 清洗框架

Fig.4 Cleaning framework

清洗方法是指以标签实例($\langle EPC, t \rangle, \langle f_1, f_2, \dots, f_k \rangle$)作为输入的一个方法分类器 M 。 f_i 是指用来描述标签的各个属性或所处物理环境的特征。共有 4 种特征:(1)标签特征,描述标签属性,如通信协议、销售商、价格、历史探测数据等等;(2)阅读器特征,描述阅读器属性,包括天线数量、协议、价格等等;(3)位置特征,描述该观测值被读出时标签所处位置,如货架、传送带等等;(4)标签所附物体特征,如该物体组成(金属或水质)、物理方位或是否为一容器。这些特征都可以用来作为方法分类的标准,但最终的标准选取需要从训练集数据中学习获得。具体的清洗方法主要为静态平滑窗口填补方法和动态平滑窗口填补方法,还可以是用户自定义的方法。

清洗规则可以具体指明分类条件,根据这些条件,每个元组实例会找到适合它的清洗方法,从而达到总开销最小的清洗策略。一种简单的模型清洗的方法是使用决策树,根据具体应用中得到的标签实例生成对应的决策树,图 5 给出了一个决策树的实例^[10]。在此基础上,对需要清洗的元组实例进行最优策略选择即可。元组实例一般需要在 $\langle r, o, t \rangle$ 的基础上转换而来。

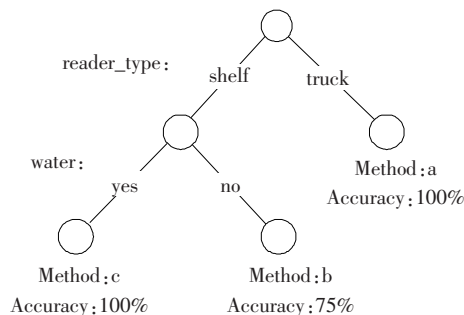


图 5 清洗策略举例

Fig.5 Example of cleaning strategy

文中提到了决策树方法和贝叶斯方法,根据不同数据特征进行最优清洗策略选择,以达到总体开销最小化。由于 RFID 技术采用无线信号传输数据,数据准确性受环境影响很大,该方法抓住了这个关键点,将布署 RFID 的环境特征,标签所附物品特征等作为分类标准。但是训练集如何选取才能覆盖应用中的所有数据特征还有待研究。另外,该技术只解决了漏读现象,如何解决多读、冗余、时间戳乱序等问题;特征值的选取,是否与漏读基本相同等问题也需要进一步研究。

总之,上面这三种技术都针对数据漏读现象在建立模型的基础上提出了清洗策略,研究的已经比较深入,但对于解决其他清洗问题,如去除冗余数据、去除多读数据等,这些模型还有待进一步扩展。

2.3 综合性清洗策略

由于 RFID 数据需要清洗的方面有许多,该节介绍能够进行综合性清洗的两种解决方案。

第一种是网络 RFID 中阅读器的最优化调度策略^[11]。RFID 数据采集可靠性低的一个主要原因是多个阅读器同时感应数据时造成信号干扰。如果根据阅读器空间位置,每一次调度那些互不干扰或干扰较小的阅读器同时感应数据,则会大大减少无线信号的干扰,这样采集数据的质量必然会提高,从预防的角度达到了数据清洗的效果。

第二种是清洗范围非常广的算法^[12],该算法针对不同的数据清洗类型提出了相应的清洗方案。针对漏读现象,该算法提出通过设置高的阅读器读频率

来解决;针对多读现象,该算法提出预先设定一窗口大小,统计在该段时间内各个标签被读到的次数,如果某个标签的统计次数小于阈值,则认为相关数据是多读数据;对于重复数据,将相隔时间太短的相同或相似数据消重,只留下一个数据即可;解决时间戳乱序问题,主要是将接收到的数据进行缓存排序,预先设定一个最大延迟时间,在此时间之后到达的数据给予删除。该算法针对性地解决了多种数据清洗问题,但处理数据的延迟较大,导致效率不高,而且只考虑了单个阅读器探测标签的情况,还有许多地方需要改进。

在 RFID 数据清洗过程中,清洗方法的深度和广度之间存在着制约。针对数据漏读现象提出的各种清洗技术都有各自的模型,对问题的研究已经非常深入,但在此模型基础上,如何解决其他类型的数据清洗问题还有待进一步研究。而对于综合性的数据清洗方法,虽然解决问题的种类很广,但深度不够,仍存在许多不足有待改进。

3 RFID 事件检测技术

RFID 事件处理的主要目的是进行事件检测,即按照给定的规则和模式,检测出指定的事件。下面分别介绍两种解决方案:以数据为中心的方法和以事件为中心的方法。

3.1 以数据为中心的方法

以数据为中心的方法是 RFID 处理系统中最早采用的方法。采用传统的数据库技术,对 RFID 数据进行建模,并将数据保存在数据库中,在 DBMS 基础上支持事件检测,代表性的系统有 Siemens 公司的 RFID 中间件系统^[13]。类似的系统还有美国加州大学伯克利分校开发的 HiFi 系统^[14],在数据流系统 TelegraphCQ 之上增加了事件检测器,具有连续进行事件检测的功能。

Siemens 的 RFID 中间件系统提出了面向时态的 ER 模型(DRER),不仅能描述 RFID 数据的特征,还能够表达业务逻辑,并支持基于分片的存储,具有可伸缩性。基于规则的框架提供自动的 RFID 数据过

滤、转换和聚合,生成高层次的语义数据。在数据库基础之上,利用规则可进行复合事件检测。

DRER 模型是该系统的核心思想,DRER 模型是在时态上扩展的 ER 模型。在普通 ER 模型中,所有的实体和联系都是静态或暂时的。但在 RFID 系统中,实体是静态的,但是所有的联系都是动态的。DRER 的实体包括应用中贴有标签的所有对象、阅读器、位置和事务,而实体间的联系包含对象与对象间的包含关系、对象与阅读器的读写关系、对象与位置的所处关系等(如图 6 所示),这些关系表都存储在磁盘上,当有用户注册查询时,与传统数据库处理数据的方式相同,将磁盘上的数据调入内存进行逐个查找。这种方式能够方便地实现应用所遇到的各种查询,如查询某个商品经过的所有历史位置,检查丢失某个商品的时间,某个商品位置改变的时间,在近 1 小时内卖出的商品信息等。为了便于表示,这些应用被分为对 RFID 物品的跟踪和物品的监测。其中,跟踪包括查询物体状态的改变过程和丢失物体;监测包括时间片查询,时间联合查询,时间聚合和包含检测等。

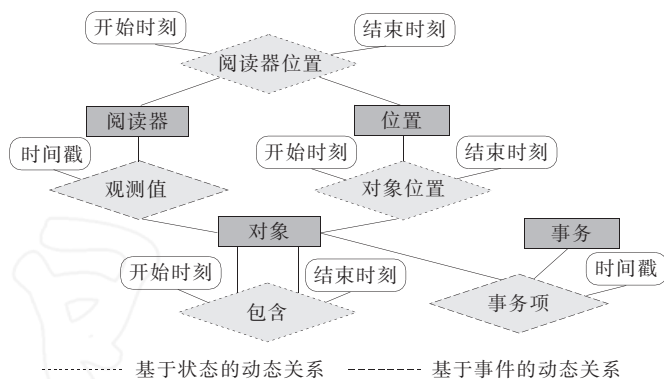


图 6 动态关系 ER 模型

Fig.6 DRER model

RFID 以数据为中心方法的简单性在于它一般是对 ER 模型的简单扩展,仅仅增添动态的联系,与特定的时间相关的属性相关,这样能够具体体现事件和状态的历史,获取数据模型的基本业务逻辑,可以有效地支持跟踪和检测等复杂的查询。此外,在 RDBMS 上容易实现 DRER 模型,能够提供对大量历

史数据的查询,而且查询结果准确。但是,RFID 的应用一般要求从高速、海量的在线事件中及时发现异常事件并进行主动报警。以数据为中心的系统由于性能限制,只适合于进行历史数据的 Ad-hoc 查询或在线监控低速的事件,不适合于针对高速的、海量的事件进行连续查询。

3.2 以事件为中心的方法

以事件为中心的方法是一种新型的 RFID 事件处理技术,基于数据源而不是基于数据库对事件进行处理,对事件直接进行处理,以获得高的效率,处理更复杂的事件。主动数据库技术、Publish/Subscribe 系统技术和数据流系统技术都与这种技术有着一定的相关性。主动数据库当中事件的建模和复合事件处理的方法,Publish/Subscribe 系统对查询数量和流速的高效支持以及数据流系统中的查询语言接口和查询优化方法,都为以事件为中心的 RFID 事件处理技术提供了重要的借鉴。但由于主动数据库的处理技术^[15,16]没有考虑流的特性和复杂的语义,因而不能直接适用。另一方面,RFID 主要考虑同一事件流上不同事件之间的相关性和限制,不同于 Publish/Subscribe 系统^[17,18]中的简单事件过滤和普通数据流处理^[19,20]的基于关系代数的处理,因此需要重新设计能够在内存中完成的高效、实时和增量的事件处理机制。

以事件为中心方法的核心问题是复杂事件处理(CEP,Complex Event Processing),即如何高效地将多个基本事件复合成为有着更加复杂语义的复合事件,包括考虑各种事件之间的限制,甚至在一些应用中需要继续检测复合事件以生成更高层次的复合事件。一些主动数据库的相关工作讨论了进行复合事件检测的基本模型,其中包括:基于有限自动机的模型,基于 Petri 网的模型,基于匹配树的模型和基于有向图的模型。这些模型也是 CEP 问题的基本模型,下面分析这些基本模型。

(1)自动机模型。由于简单的复合事件的表达式与正则表达式有着相似的形式,因此可以使用自动机模型来实现事件表达式。ODE^[21]是最早提出使用自

动机模型进行复合事件检测的系统。每当一个参与复合事件发生的基本事件到达,自动机就会从一个状态跃迁到下一个状态,当自动机进入某个可接受的状态,则说明发生了复合事件。但是,简单的自动机模型在匹配过某个基本事件后无法重新访问该事件,如果需要考虑各个基本事件之间的时间和数值联系,还需要引入额外的数据结构来保存这些信息,就构成了扩展的自动机模型。此外,也可以在自动机跃迁谓词中加入更复杂的基于时间或者数值的条件限制,来设计专门的自动机模型。

(2)Petri 网模型。使用 Petri 网表示和检测一个复合事件。Petri 网的输入位置结点为基本事件,输出位置结点为复合事件。通过输入 Token,计算跃迁守护函数,如果成立,则引发跃迁并标记位置结点,当序列中的最后一个位置结点被标记后,则说明复合事件发生。复合事件的增量检测是通过 Petri 网中标记的位置来描述的。主动数据库系统 SAMOS^[22]和监控系统 HiFi^[14]就采用了这种模型。

(3)匹配树模型。基于匹配树的技术通过匹配树的结构来实现复合事件的过滤器。基本事件为匹配树的叶节点,各个层次的复合事件为匹配树的中间节点,如果到达根节点并且相应的事件被成功地过滤出来,就认为检测到了一个复合事件。READY^[23]和 Yeast^[24]系统使用了这种技术。

(4)有向图模型。与匹配树的方法十分相似,有向图模型采用有向的无环图(DAG)来表示复合事件。节点为对事件描述,边为事件合成的规则。节点通过对相关事件的引用进行标记,除了用边表示事件合成规则,节点也带有相应的规则,对应某节点的事件发生,节点的规则也会触发。Sentinel 系统^[25]和 EVE 系统^[26]采用的就是这种模型。

从上面的分析可以看出,基于自动机和 Petri 网模型的复合事件检测仅仅匹配按照顺序到达的基本事件,而基于树或者图的过滤又没有考虑到匹配基本事件的顺序或者时序距离,例如复合事件中的第一个事件没有发生,第二个事件可能也会被过滤,造成了不必要的开销,在实际应用中都具有一定的局

限性。而且这些对于复合事件的检测方法都将基本事件的过滤和复合事件的检测作为不同的步骤来处理,没有考虑到在复合事件检测的同时过滤那些有价值的基本事件。文献[27]在 CompAS 系统的基础上,考虑了对基本事件和复合事件的集成处理。通过在检测复合事件的同时有选择地过滤基本事件,降低了复合事件检测的响应延迟。

需要说明的是,简单地使用这些模型并不能有效地解决复杂的 RFID 应用,如何处理事件流、如何处理复杂语义、如何处理时态性、如何提高在线处理效率,还有很多问题有待于研究。作者认为下面几个方面是 RFID 事件流中复杂事件处理区别于传统复合事件检测的重要特点:

- (1)语义复杂,数据量大,原始数据不准确,需要在内存中考虑高效的增量维护算法。
- (2)复合事件的时序限制、数值限制等约束大量存在,传统的 ECA 规则不能高效地支持。
- (3)“非自发”的复合事件存在,例如 Negation 和 Repetition,某个事件的发生并不知道复合事件是否发生。
- (4)由于 Reader 位置分散等原因,在集中处理端,基本事件之间的时间戳乱序大量存在。
- (5)针对同样的 RFID 事件流可能同时注册大量不同的查询和分析。
- (6)先前参与匹配的基本事件是否继续参与后面复合事件的检测要视应用而定,也就是说不同的应用可能需要考虑不同的消耗模式。

4 RFID 事件处理原型系统

针对传感数据源的 CEP 系统的研究和开发工作正在兴起,有些成果可以适用于 RFID 事件处理的应用场景。典型系统有美国加州大学伯克利分校开发的原型系统 SASE^[28]、Cornell 大学开发的 Cayuga 系统^[29]、Texas 大学 Arlington 分校开发的 EStream 系统^[30],以及 Dartmouth 学院开发的 PQS 系统^[31]。SASE 系统提供扩展的事件语言、事件查询处理器和操作优化策略等,实现了 RFID 设备的数据采集和清洗、

基本事件生成、复合事件处理、事件归档,以及对事件的查询。Cayuga 系统提出了扩展的查询语言,通过自定义的自动机模型和内部命名方法来高效地检测复合事件,并讨论了相关的系统实现的技术细节。EStream 系统在数据流查询器的基础上,集成了连续事件查询器,利用规则检测数据流上的复合事件,考虑了高效的增量维护算法。PQS 系统采用非确定性有限自动机、隐型马尔科夫过程 HMM 等方法对动态过程建模,以事件流为输入参数,通过对模型求解,从而发现产生这些事件的过程,实现过程查询和检测。这些系统可以不同程度地解决 RFID 时间检测的新特点,其中 SASE 和 Cayuga 是针对 RFID 事件检测最有代表性的系统,下面重点介绍 SASE 和 Cayuga 的主要技术。

SASE 是最早提出的针对 RFID 数据以事件为中心进行处理的系统,提出了一种基于查询计划的方法^[32],将时序关系、数值限制、时序限制、Negation 等操作分别作为操作符进行处理。查询处理是由 6 种算符的子集组成:序列扫描、序列构造、选择、窗口、非操作和转换,如图 7 所示,这些运算符由下至上构成了一个查询计划。同时也考虑了大量的优化操作,提出了对操作符下推、基于属性分组等相关的

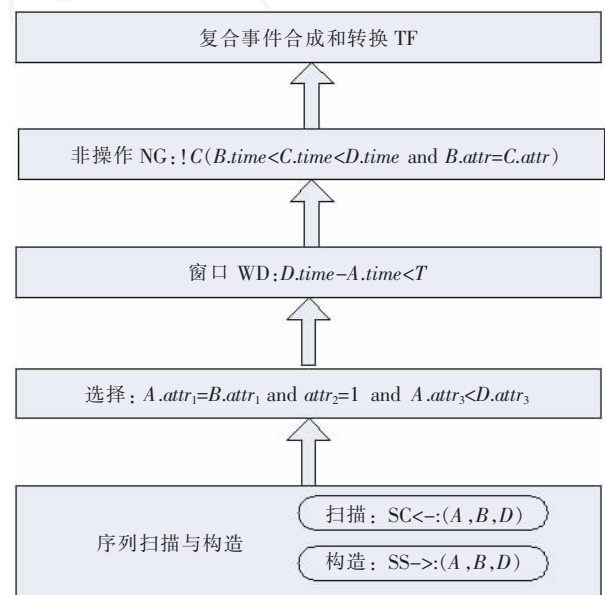


图 7 SASE 系统基于查询计划的处理过程

Fig.7 SASE system process based on query plan

查询优化策略。但是,针对 RFID 新的特点,没有考虑到 Repetition 等语义,消耗模式也仅仅考虑了 exclusive,即已经参与前面复合事件匹配的基本事件还要参与后面复合事件的匹配,而且假设事件流的时间戳满足自然数全序关系,这在很多应用场景下是不适用的。

Cayuga 系统的目标也是对传感数据中的事件实现高效在线的处理。提出了使用 Cayuga 自动机来进行复合事件检测^[33]。与输入流一样,每个自动机的状态设置了一个固定的关系模式。状态 P 与状态 Q 之间的跃迁用 $\langle \theta, F \rangle$ 来标记,其中 θ 代表 $\text{schema}(P) \times \text{schema}(S)$ 上的谓词, F 表示将 $\text{schema}(P) \times \text{schema}(S)$ 映射成为 $\text{schema}(Q)$ 的函数。假设一个自动机当前状态为 P , 数据是 x , x 满足 $\text{schema}(P)$, 如果事件流 S 中某个事件 e 到达, 并且满足 $\theta(x, e)$, 自动机的状态就跃迁为 Q , 存储的数据变为 $F(x, e)$ 。

在 Cayuga 自动机中, 将跃迁分为自循环的跃迁和向前的跃迁, 并且将自循环的跃迁分为用于过滤的跃迁和用于重绑定的跃迁。在自动机图示中, 通过自循环跃迁标在某个状态的上方或者下方来区分表示过滤和重绑定。不同的操作有不同的跃迁类型, Cayuga 自动机定义了一套规则来维护各个跃迁边上的 θ 和 F 。Cayuga 自动机还定义了多种内部命名规则, 使得 F 表示的转换能够顺利完成, 此外 Cayuga 还讨论了内存管理等实现技术。图 8 给出了一个 Cayuga 自动机的示意图。Cayuga 自动机虽然能够解决 SASE 系统所不具备的一些功能, 如简单的重复操作和时序限制; 但是由于这种自动机加入大量的内部命名机制, 不便于用户的理解, 而且只支持简单的数值限制, 也没有考虑时间戳乱序。与 SASE 系统相

比, 不支持重要的 Negation 操作, 消耗模式也仅仅考虑简单的 delete 模式, 即先前参与复合事件匹配的基本事件不再参与后面的匹配。

除了上面的原型系统, 还有一些工作就 RFID 新的特点进行了针对性的研究。下面介绍 SASE 和 Cayuga 都没有很好解决的 Negation 操作和时间戳乱序问题。

对于 Negation 这类非自发的事件, 主动检测技术会提高查询的整体效率。文献[34]提出了一种主动检测 Negation 的方法。首先将事件检测的过程用一个图来表示, 叶结点表述原始事件, 中间结点表示按规则检测出的复合事件。其中, 时序限制作为中间结点的附加信息, 并且自顶向下传递时序限制信息, 最大程度地合并子图, 减少处理开销。通过人工生成“伪事件”, 解决针对非自发事件的检测。例如对于操作 $\text{Within}(\text{Sequence}(\text{Negation}(A), B), 0, 30)$, 表示在某个 B 类型的事件发生之前的 30 s 内, 某个 A 类型的事件并没有发生过。像这种涉及到 Negation 操作的语义, 可以在检测到某个 A 发生时向叶节点传递一个“伪事件” C , $C.\text{timestamp} = A.\text{timestamp} + 30$ 。这样当某个 B 事件到达时, 将 B 的时间戳与 C 相比较, 如果 B 的事件戳小于 C , 则不会产生复合事件; 反之, 产生复合事件, 并将 C 删除。这种方法可以提高复合检测的效率和实时性。特别对于带有各种其他操作和时序限制的 Negation 操作, 具有很好的效率。但是这种技术本身很难解决数值限制和时间戳乱序等问题。

时间戳乱序在 RFID 系统中尤其普遍, 而且由于 RFID 复合事件中大量涉及时序操作, 而大部分技术仅仅通过到达系统的时间来判断事件的先后发生, 因此错误的时间顺序可能会造成错误的检测结果。文献[35]提出了一种可能的解决时间戳乱序的简单方法, 采用时序自动机模型^[36], 提出在自动机跃迁的谓词上加入判断时间乱序的谓词。例如对于操作 $\text{Sequence}(A, B, C)$, 表示三种类型的事件先后发生。这样可以在自动机的跃迁谓词上分别加入 $A.\text{timestamp} < B.\text{timestamp}$, $B.\text{timestamp} < C.\text{timestamp}$, 这样可

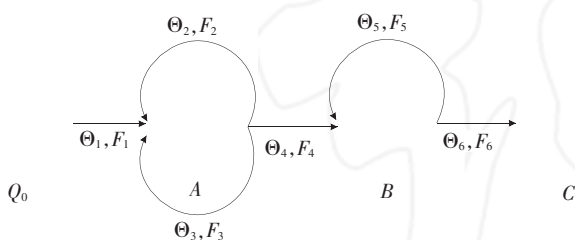


图 8 Cayuga 自动机示意图

Fig.8 Illustration of Cayuga automata

以滤去那些由于事件戳乱序而造成错误时序匹配的复合事件。但是很明显,由于没有修改这种乱序,一些可能参与匹配但乱序的事件也会被删除,并没有检测出所有的复合事件。如果采用缓冲排序的方法,又会降低检测的实时性。因此需要根据不同的应用综合考虑不同的方法。此外,由于系统时间粒度的原因,多个事件可能“同时”到达,处理时要考虑相应的语义。时间戳乱序也可以在预处理阶段进行解决。

5 开放性问题

以事件为中心的 RFID 事件处理技术虽然有了很多创新性研究成果,但在很多方面有待进一步的研究。

首先,从查询语义角度考虑,现有的技术大都只考虑事件的先后发生,而实际应用中,表示“与”、“或”、“同时”等语义也是大量存在的。此外,现有的工作没有重点考虑对于复合事件的进一步复合,特别是例如判断两个复合事件是否重合或包含这样的复杂语义。而且对于前面提出的 RFID 复合事件检测的特点,很少有系统能够比较全面的关注。因此研究和设计高效的通用的以事件为中心的 RFID 管理系统,使之适用于复杂的 RFID 为背景的应用,仍然是一个挑战性工作。

其次,对于大量存在的 RFID 事件流查询共享问题,很少有文献进行讨论。文献[32]提出可以根据查询计划本身的建模特点来共享操作符。但由于时序匹配本身就是独立的一个操作符,这种共享只能检测出时序操作和时序限制完全相同的共享,这种共享的粒度是不够的。例如,对于 $\text{Sequence}(A, B, D)$, $\text{Sequence}(A, B, \text{Negation}(C), D)$ 和 $\text{Sequence}(A, B, C)$ 一些操作是共享的,必须在时序匹配操作的内部检测这种共享。再例如, $\text{Within}(\text{Sequence}(A, B), 0, 40)$ 与 $\text{Within}(\text{Sequence}(A, B), 20, 60)$, 在时序限制上存在着明显的共享。可以参照数据流上滑动窗口查询的相关共享技术,例如进行分桶或索引来处理复杂情况下的共享。有效的共享方法会在很大程度上提高整个系统复合事件检测的效率。

此外,RFID 本身的应用一般是分布式的,现在的研究主要关注集中式的处理方法,如何进行高效的分布式处理是一个关键的问题。对于集中处理端,在多个事件流的环境下,如何定义 RFID 事件流的 QoS,并根据事件流的 QoS 进行调度是一个非常关键的问题。但是,现有的工作还很少有专门针对 RFID 事件分布式处理、QoS 建模和调度算法的研究。

最后,由于 RFID 数据的不准确性,可以在 RFID 事件的建模中加入概率信息,进行基于概率的 RFID 事件流复杂事件处理的研究。可以借鉴数据流上一些基于概率的处理方法^[37,38],考虑在自动机模型中加入统计模型。此外,需要对各种复杂事件处理操作进行概率建模,例如对于复合事件 $\text{Sequence}(A, \text{Negation}(B), C)$, 某个实例的基本事件概率为 $a.Proc=Pa$, $b.Proc=Pb$, $c.Proc=Pc$, 那么直观上讲,复合事件的概率为 $Pa(1-Pb)Pc$ 。在内存受限的情况下,可以考虑高效的算法来首先检测和响应概率较大的复合事件。此外,事件的检测是有层次的,较低的层次可能只能提供满足某种语义的一个范围或一种分布情况^[39]。为了在更高的层次上检测出具有更高语义的事件,有时也需要利用相关的统计信息来进行推断。

6 结论

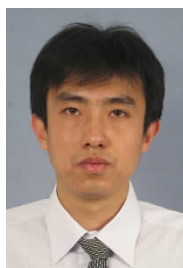
RFID 事件处理是 RFID 技术的重要研究内容,涉及到数据库技术、人工智能技术、实时技术、嵌入式技术、普适计算技术、数理统计理论、现代信号处理与分析理论等领域。文章对 RFID 事件处理中涉及的关键技术,包括数据清洗方法、以数据为中心的检测方法、以事件为中心的事件检测方法,以及复杂事件处理系统等进行了归纳和总结,对一些亟待解决的问题提出了展望。可以看出,RFID 的普及给 RFID 的数据管理方面的研究提供了新的机遇,但研究支持复杂应用的 RFID 处理技术还面临很多挑战。

References:

- [1] Fuhrer P, Guinard D, Liechti O. RFID: from concepts to concrete implementation[C]//Proc of IPSI, 2006.

- [2] EPC Tag Data Standards Ver.1.1[R]. EPC Global Inc, 2004.
- [3] Bornhoevd C, Lin T, Haller S, et al. Integrating automatic data acquisition with business processes experiences with SAP's auto-ID infrastructure[C]//Proc of VLDB, 2004:1182-1188.
- [4] Lee M, Cheng F, Leung Y A. Quantitive view on how RFID will improve a supply chain, RC23789(W0511-065)[R]. IBM Research Center, 2005.
- [5] Ferguson R B. Logan airport to demonstrate baggage, passenger RFID tracking[J]. eWEEK, 2006.
- [6] Swedberg C. Hospital uses RFID for surgical patients[J]. RFID Journal, 2005.
- [7] Chawathe S S, Krishnamurthy V, Ramachandrany S, et al. Managing RFID data[C]//Proc of VLDB, 2004:1189-1195.
- [8] Jeffery R, Garofalakis M, Franklin M J. Adaptive cleaning for RFID data streams[C]//Proc of VLDB, 2006:163-174.
- [9] Jeffrey R, Alonso G, Franklin M, et al. A pipelined framework for online cleaning of sensor data streams[C]//Proc of ICDE, 2006:773-778.
- [10] Gonzalez H, Han J, Shen X. Cost-conscious cleaning of massive RFID data sets[C]//Proc of ICDE, 2007:1268-1272.
- [11] Deolalikar V, Recker J, Mesarina M, et al. Optimal scheduling for networks of RFID readers[C]//LNCS 3823: Proc of EUC Workshops, 2005:1025-1035.
- [12] Bai Y, Wang F S, Liu P Y. Efficiently filtering RFID data streams[C]//Proc of the 1st Int VLDB Workshop on Clean Databases, 2006:50-57.
- [13] Wang F S, Liu P Y. Temporal management of RFID data[C]//Proc of VLDB, 2005:1128-1139.
- [14] Rizvi S, Jeffrey S, Krishnamurthy S, et al. Events on the edge[C]//Proc of SIGMOD, 2005:885-887.
- [15] Chakravarthy S, Krishnaprasad V, Anwar E, et al. Composite events for active databases: semantics, contexts and detection[C]//Proc of VLDB, 1994:606-617.
- [16] Gehani N H, Jagadish H V, Shmueli O. Composite event specification in active databases: model & implementation[C]//Proc of VLDB, 1992:201-234.
- [17] Fabret F, Jacobsen H A, Llibat F, et al. Filtering algorithms and implementation for very fast publish/subscribe[C]//Proc SIGMOD, 2001:115-126.
- [18] Aguilera M K, Strom R E, Sturman D C, et al. Matching events in a content-based subscription system[C]//Proc PODC, 1999:53-61.
- [19] Carney D, Cetintemel U, Cherniack M, et al. Monitoring streams—a new class of data management applications[C]//Proc of VLDB, 2002:215-226.
- [20] Chandrasekaran S, Cooper O, Deshpande A, et al. TelegraphCQ: continuous data flow processing for an uncertain world[C]//Proc of CIDR, 2003.
- [21] Gehani N, Jagadish H. Ode as an active database: constraints and triggers[C]//Proc of VLDB, 1991:327-336.
- [22] Gatzia S, Dittrich K R. SAMOS: an active object-oriented database systems[J]. IEEE Bulletin on Data Engineering, 1992,15(1/4):23-26.
- [23] Gruber R E, Krishnamurthy B, Panagos E. The architecture of the READY event notification service[C]//Proc of the 19th IEEE Int Conf on Distributed Computing Systems Middleware Workshop, 1999.
- [24] Krishnamurthy B, Rosenblum D S. Yeast: a general purpose event-action system[J]. IEEE Transactions on Software Engineering, 1995,21(10):845-857.
- [25] Chakravarthy S, Mishra D. Snoop: an expressive event specification language for active databases, UF-CIS-TR-93-007[R]. University of Florida, Gainesville, 1993.
- [26] Geppert A, Tombros D. Event-based distributed workflow execution with EVE, ifi-96.05[R]. University Zurich, 1996.
- [27] Hinze A. Efficient filtering of composite events[C]//Proc of the British National Database Conf, 2003:207-225.
- [28] Gyllstrom D, Wu E, Chae H J, et al. SASE: complex event processing over streams[C]//Proc of CIDR, 2007:407-411.
- [29] Brenna L, Demers A, Gehrke J, et al. Cayuga: a high-performance event processing engine[C]//Proc of SIGMOD, 2007:1100-1102.
- [30] Garg V. Estream: an integration of event and stream

- processing[D]. University of Texas at Arlington, 2005.
- [31] Cybenko G, Berk V H. Process query system[J]. IEEE Computer, 2007,40(1):62-70.
- [32] Wu E, Diao Y, Rizvi S. High-performance complex event processing over streams[C]//Proc of SIGMOD, 2006: 407-418.
- [33] Demers A, Gehrke J, Panda B, et al. Cayuga: a general purpose event monitoring system[C]//Proc of CIDR, 2007:412-422.
- [34] Wang F S, Liu S, Liu P Y, et al. Bridge physical and virtual worlds: complex event processing for RFID data streams[C]//LNCS 3889: Proc of Int Conf on Extending Database Technology(EDBT), 2006:588-607.
- [35] Derakhshan R, Orlowska M E, Li X. RFID data management: challenges and opportunities[C]//Proc of IEEE Int Conf on RFID, 2007:175-182.
- [36] Alur R, Dill D. A theory of timed automata[J]. Theoretical Computer Science, 1994,126(2):183-235.
- [37] Dalvi N, Suciu D. Management of probabilistic data foundations and challenges[C]//Proc of PODS, 2007:1-12.
- [38] Jayram T S, McGregor A, Muthukrishnan S, et al. Estimating statistical aggregates on probabilistic data streams[C]//Proc of PODS, 2007:243-252.
- [39] Rizvi S. Complex event processing beyond active databases: streams and uncertainties, UCB/EECS-2005-26[R]. University of California at Berkeley, 2005.



谷峪(1981-),男,辽宁鞍山人,东北大学博士研究生,CCF 学生会员,主要研究领域为数据流查询处理、RFID 数据管理等。

GU Yu was born in 1981. He is a PhD candidate at Northeastern University, China. He is a CCF student member. His research interests include data stream query processing and RFID data management.



于戈(1962-),男,辽宁大连人,1982 年和 1986 年分别于东北大学获得计算机专业学士和硕士学位,1996 年于日本九州大学获计算机工学博士学位;现为东北大学教授,博士生导师,中国计算机学会理事、美国 ACM 和 IEEE 会员,主要研究领域为数据库理论与技术。

YU Ge was born in 1962. He received his BS degree and MS degree in Computer Science from Northeastern University of China in 1982 and 1986, respectively, PhD degree in Computer Science from Kyushu University of Japan in 1996. He is a professor and doctoral supervisor at Northeastern University of China. He is a member of IEEE, ACM, and a senior member of CCF. His major research interests include database theory and technology.



张天成(1969-),男,湖北潜江人,东北大学博士研究生,现为东北大学讲师,CCF 会员,主要研究领域为数据流分析与挖掘,人工智能技术。

ZHANG Tiancheng was born in 1969. He is a PhD candidate at Northeastern University, China. He is a lecturer at Northeastern University of China and CCF member. His research interests include analysis and data mining over data streams, artificial intelligence technology.