面向RFID数据处理的复杂事件模式匹配方法

戚 湧 胡 军 李千目
(南京理工大学计算机科学与工程学院 南京 210094)

摘 要 RFID数据具有不确定性,复杂事件处理技术将RFID数据看作不同类型的事件,从事件流中检测符合特定匹配模式的复杂事件。概率事件流分为多项概率事件流和单项概率事件流;针对多项概率事件流,提出NFA-MMG模式匹配方法,亦即使用多个有向无环图结合自动机实现模式匹配。针对单项概率事件流,提出NFA-Tree模式匹配方法,亦即使用匹配树结合自动机实现模式匹配;并提出改进的NFA-Tree方法,即基于概率阈值进行过滤,提高结果过滤效率。实验结果验证了上述模式匹配方法的性能优势。

关键词 RFID数据,复杂事件处理,模式匹配,自动机
中图法分类号 TP391 **文献标识码** A

Pattern Matching Method of Complex Event for RFID Data Processing

QI Yong HU Jun LI Qian-mu
(College of Computer Science and Engineering, Nanjing University of Science & Technology, Nanjing 210094, China)

Abstract RFID data is generally uncertain. Complex event processing (CEP) treats the data as different types of events, queries sequence of events in which match specific patterns of sequence are defined by high-level application from the event stream. Event stream is divided into multiple alternative event stream and single alternatives event stream. NFA-MMG pattern matching method for multiple alternatives event stream was proposed. The method uses combination of directed acyclic graph and automatic machines to achieve complex event pattern matching on the uncertain data. NFA-Tree pattern matching method for single alternatives event stream with the use of matching tree and automatic machines on uncertain data was proposed. The NFA-Tree algorithm was improved by pruning the matching tree to improve the efficiency of query optimization, which filters the results of the match situation based on probability threshold. The complex event processing system prototype uncertain data was developmented to realize the above algorithm, and the experiment examines the validation and performance of the algorithms.

Keywords RFID data, Complex event processing, Pattern matching, Automatic machines

1 引言

近年来,RFID技术在供应链管理[1]、物流[2]、医疗[3]、防伪[4]和工业制造[5]等多个领域得到广泛应用。RFID应用系统结构包括RFID电子标签、RFID读写器、RFID数据处理中间件和上层应用程序,如图1所示。复杂事件处理查询引擎在该系统中处于核心地位,其执行模式匹配连续查询,捕获随时间变化的事件,根据事件模式和趋势等触发响应,发起响应动作序列的调用,将识别出的事件保存至数据库或直接提供给第三方使用。随着RFID技术日益广泛的应用,产生的海量数据处理面临着许多问题,其中一个主要问题是RFID数据的不准确性。一般情况下,原始数据的准确率仅为60%~70%[6],由于阅读器本身的漏读、脏读和多读现象[7],以及RFID数据处理过程中的主观性,造成RFID应用整个生命周期中数据的不确定性。

在RFID应用中,事件可以看作是系统与带RFID标签对象的一次数据交换。RFID事件分为基本事件和复杂事件,基本事件是从RFID原始数据中得到的原始事件,复杂事件由基本事件按时序的某种模式排列构成。将不确定RFID数据抽象为一个带概率的RFID数据流,根据概率事件流中的基本事件是否含有多个可能选项事件将概率事件流分为单项概率事件流和多项概率事件流。

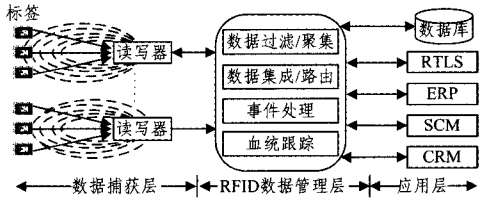

图1 RFID应用系统结构

复杂事件处理(Complex Event Processing,CEP)由斯坦福大学David Luckham等人提出[8],是利用事件属性之间的关联,按模式从事件流中匹配符合定义的事件序列。面向RFID应用的复杂事件检测方法分为基于自动机、基于Petri网、基于匹配树和基于有向图等方法,目前有一些CEP原型

到稿日期:2012-04-07 返修日期:2012-08-23 本文受国家自然科学基金(61272419)和中国航天CALT创新基金(CALT201102)资助。
戚 湧(1970-),男,博士后,教授,主要研究方向为物联网技术、信息安全,E-mail:qyong@mail.njust.edu.cn;胡 军(1988-),硕士生,主要研究方向为RFID数据处理;李千目(1979-),博士后,副教授,主要研究方向为物联网技术。

系统用于 RFID 应用，如 SASE[9-11]、Cayuga[12] 和 Esper 等。SASE 是针对 RFID 应用的实时数据流上的复杂事件处理系统，其查询处理过程如图 2 所示，查询计划包括序列扫描和构造、选择操作、窗口操作、Negation 非操作和转换操作等步骤。SASE 复杂事件处理系统提供复杂事件处理的基本功能，但没有考虑输入数据流中 RFID 数据的不确定性，对于不确定 RFID 数据不能检测出所有可能的复杂事件。

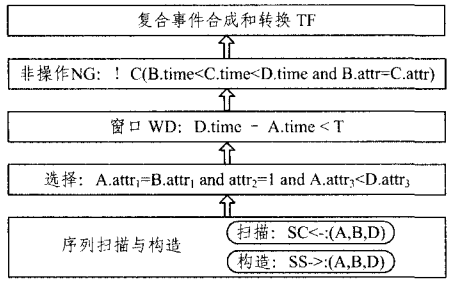

图 2　SASE 复杂事件查询处理过程

传统的数据模型无法准确描述 RFID 数据的不确定性。目前概率数据通常基于可能世界模型(possible world)[13]，该模型中数据的某个状态下的实例 I 是未知的，取而代之的是 n 个可能的实例 $I_1, I_2, \cdots, I_n$，它由若干个可能世界组成。每个可能世界都有一定概率发生，所有可能世界的概率值之和为 1。当输入数据规模急剧扩大时，可能世界的实例数目呈指数倍增长，相应的概率计算问题复杂度急速增大，此时列出所有可能世界实例计算各可能世界实例的概率，进行复杂事件查询处理造成的开销比较大、处理效率不高。

关于 RFID 不确定数据处理，Kimelfeld[14] 等提出在统计学模型上使用马尔可夫序列转换器来实现事件序列匹配。Letcher 等[15]研究马尔可夫流上事件查询的世系问题，通过世系图记录和查询不确定事件匹配过程中的信息。刘海龙与李战怀等[16]对 SASE 复杂事件检测方法进行优化，使用 hash 表来代替堆栈，改进子事件分布不均匀和非事件检测策略。陈群与陈远等[17]提出基于 NFA 带约束检测方法进行不可靠 RFID 数据上的复杂事件处理。谷峪与于戈等提出一种基于动态概率路径事件模型的 RFID 数据填补算法[18]。廖国琼等[19]提出基于核密度估计的 RFID 数据流清洗方法。上述研究主要集中在复杂事件处理技术的优化和不确定 RFID 数据的清洗，关于不确定 RFID 数据上复杂事件处理的国内外研究还较少。

2　基于 NFA 的 RFID 数据复杂事件模式匹配方法

2.1　传统基于自动机的模式匹配

自动机由一个五元组 $(Q, \Sigma, \delta, q_0, F)$ 构成，Q 表示有限非空的状态集合，Σ 为输入字符表，δ 表示自动机 $Q \times \Sigma \to Q$ 的状态转移函数，q_0 为自动机的开始状态，F 为自动机的终止状态。使用自动机模型可以表示正则表达式，复杂事件的表达式与正则表示式具有一致性，因此使用自动机模型表示相应的复杂事件表达式。基于自动机的复杂事件模式匹配分为序列扫描(sequence scan，SS)和结果序列构建(sequence construction，SC)两步，首先构建 NFA 来表示一个复杂事件的事件序列，然后进行序列扫描，构建匹配查询模式的结果序列。

创建识别事件序列模式 $(a, b[], c)$ 不确定自动机 NFA，如图 3 所示。扫描输入事件流，使用该 NFA 检测出匹配模式的事件序列。状态 0 是初始状态，当识别到事件 A 时，状态由 0 迁移到 1，在状态 1 识别到事件 B 时，状态迁移到 2，在状态 2 上接受事件 B 时，则跳回状态 2 保持不变，在状态 2 上接受事件 C 时跳转到状态 3，3 为自动机的终止态，识别过程结束。图 4 是 SASE 系统使用自动机模型查询 $SEQ(A\,a, B\,b, D\,d)$ 的模式匹配过程。

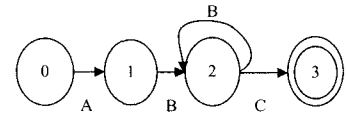

图 3　识别序列 $(a, b[], c)$ 的 NFA

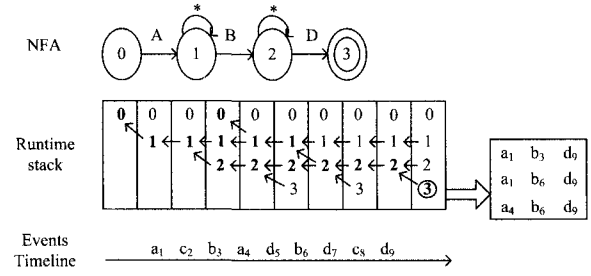

图 4　SASE 查询 (a, b, d) 的过程

以上基于自动机的方法只适用于确定数据复杂的事件模式匹配。当输入数据为不确定数据时，一个基本思路是基于可能世界模型实例的方法，根据数据的不确定性，将每个数据取不同值的情况进行组合，列出所有可能的组合，对这些组合分别进行模式匹配，本文称这种方法为 Naive 方法。当滑动窗口较大时，由于数据的组合数非常大，Naive 方法的效率不高。

2.2　基于 NFA-MMG 的多项概率事件流模式匹配方法

对常规的 NFA 序列扫描和结果序列构建过程进行扩展，提出 NFA-MMG 模式匹配方法，使用多个多源有向无环图(Multiple Multiple-source Directed Acyclic Graph，MMG)的数据结构来保存自动机状态迁移过程中对应的事件序列，使其能够进行概率事件流上的复杂事件处理。表 1 所列概率事件流的第一行为时间戳，第二行为该时间点上的基本事件，“|”用来连接所有可能选项事件，事件后面的括号表示该事件的 ID，如 $a(11) \mid b(12)$ 表示在时间戳 1 上可能发生的基本事件是 ID 为 11 的事件 a 或 ID 为 12 的事件 b。在此事件流中查询复杂事件：$SEQ(A\,a, B{+}b[], C\,c)$。ABC 为事件 abc 所要符合的条件，假设事件流中所有事件都符合 ABC 的条件限制，$b[]$ 表示有一个或多个 b 事件。

表 1　概率事件流

时间戳	1	2	3	4
事件(ID)	a(11) | b(12)	a(21) | b(22)	b(31) | c(32)	a(41) | c(42)

图 5 是当表 4 所列的概率事件流到达时，匹配 $(a, b[], c)$ 事件序列所生成的 MMG，节点中的数据 $(a(21), 1)$ 表示当前节点所包含的事件为 $a21$，对应自动机的状态为 1。

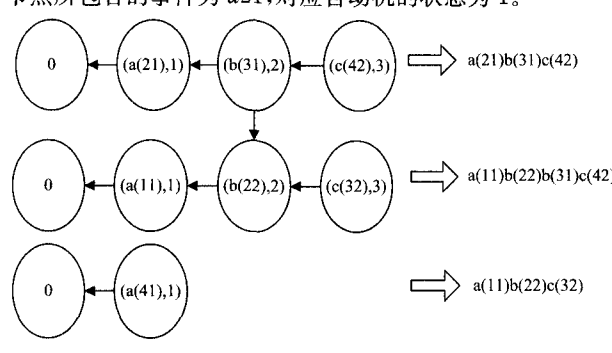

图 5　识别概率事件流的 MMG

2.3 基于 NFA-Tree 的单项概率事件流模式匹配方法

假设复杂事件处理系统中接收的单项概率事件流如表 2 所列，事件流中的每个基本事件都由单个事件及该事件发生的概率值组成。

表 2 单项概率事件流

Timestamp	100	101	103	104	105	106
事件(ID)	a_1:0.8	b_1:0.7	b_2:0.8	c_1:0.4	b_3:0.4	c_2:0.4

使用匹配树来保存状态迁移过程中事件的节点和不确定数据流引起自动机转移的整个过程，称为 NFA-Tree 模式匹配方法。NFA 在滑动窗口范围内对输入事件流进行模式匹配时，同时进行结果序列匹配树的创建与维护。当第一个触发自动机状态开始转变的事件 e 到达时，创建一棵结果序列匹配树 Tr，根节点为 e；对每一个后续事件 e，如果 e 使自动机状态转移，则在结果序列匹配树当前层的非终态叶子节点添加子节点 e 和 $\bar{e}$，如果当前到来的事件 e 使自动机进入接受态，则标记新添加的子节点 e 为终态叶子节点。具体如算法 1 所示。

算法 1 基本匹配树维护算法

```
1. for each e in event steam{
2.   if e triggers NFA intial state then
3.   { create a new tree Tr; Tr. root=e;
4.     NFA. state=initial state;}
5.   end if else{
6.     for (each leaf node lnode of Tr){
7.   if e. timestamp-timestampStart<=slideWindow the
8.       if e tiggers NFA to next state then
9.        lnode. leftChild=e;
10.       NFA. state=next state;
11.       if (NFA. state! =final state and lnode. state! =end) then
12          . lnode. ridgtChild=ē;
13.       else e. state=end;}}
```

2.4 NFA-Tree 的模式匹配结果的阈值过滤优化

在某些 RFID 应用环境下，复杂事件查询返回的匹配事件序列概率值必须大于某个阈值。针对不同的 RFID 数据采集环境和对精度要求的不同，该阈值可以在 0 和 1 之间变化。基本的基于阈值的复杂事件查询结果过滤的方法是计算出所有结果序列的概率值，然后剔除低于阈值的结果序列。由于该方法效率不高，本文对其进行改进，在构建匹配结果序列时，在匹配路径中加入概率值约束，在生成结果序列的过程中将概率值低于阈值的复杂事件查询结果进行过滤。

算法 2 针对阈值过滤优化的匹配树生成算法

```
1. for each e in event stream{
2.   if e triggers NFA intial state then
3.     if(e. probability>=Threshold then)
4.       create a new tree Tr; Tr. root=e; Tr. root. probability=e.
         probability
5.       NFA. state=initial state;}
6.     else discard e
7.   else{
8.     for(each leaf node lnode of Tr){
9.   if e. timestamp-timestampStart<=slideWindow then
10.       if e tiggers NFA to next state then
11.       lnode. leftChild=e;
12.       e. probability=e. probability * lnode. probability;
13.       if(e. probability<Threshold) then prune the branch
14.       NFA. state=next state;
15.       if(NFA. state! =final state and lnode. state! =end) then
16.         lnode. ridgtChild=ē;
17.         ē. probability=ē. probability * mode. probability;
18.         if(ē. probability<Threshold) then prune the branch
19.       else e. state=end;}}
```

其优化后如算法 2 所示。观察算法 1 中结果序列匹配树建立维护的过程以及上述匹配结果事件序列概率值的计算方法，当在匹配树建立新节点 n 时，如果该 n 由上述节点概率值算法计算得出该节点需要保存的概率值为 $P(n)$，显然 $P(n)$ 不小于该节点的任一子节点保存的概率值，如果节点 n 的子节点有终态节点，该终态节点代表的输出结果事件序列的概率值为 $P(seq)$，则 $p(n)\geqslant P(seq)$，如果 $p(n)<Threshold$，可得 $P(seq)<Threshold$，该输出事件序列概率值小于过滤阈值，输出结果被过滤。

3 实验分析

3.1 实验环境

开发和实验的环境为 AMD Turion 64 X2 1.8GHz CPU、1GB 内存、操作系统为 Red Hat Linux 9.0。在参考 SASE 的基础上，针对不确定数据流上的复杂事件处理，精简部分不适用于不确定数据的功能与模块，增加处理不确定数据需要的功能模块，修改序列扫描与序列构建部分的核心方法，在输出结果前增加概率计算功能，使用 Java 实现不确定数据复杂事件处理原型系统，如图 6 所示。本文将所提出的模式匹配方法与 2.1 节 Naive 方法进行比较分析。

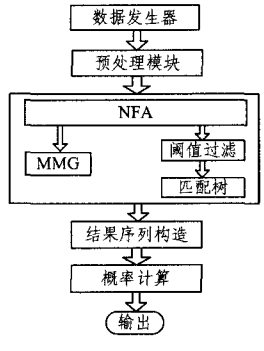

图 6 不确定数据的复杂事件处理原型系统结构

3.2 NFA-MMG 方法实验结果及性能分析

图 7 中的 X 轴为输入事件流中基本事件的可能选项事件数量，Y 轴为系统每秒处理事件的个数。对比 Naive 和 NFA-MMG 方法的吞吐量。当 X 轴为 1 时，每个不确定事件的可能选项事件为 1，输入事件流等同于确定数据流，Naive 和 NFA-MMG 方法的吞吐量相近。当可能选项事件数量增加时，可能世界实例呈指数增加，Naive 方法吞吐量急剧下降。当可能选项事件数量增加时，构建 MMG 图的复杂度也会增加，NFA-MMG 的吞吐量下降也很快。但相对于 Naive 方法，NFA-MMG 方法在同等问题规模下，吞吐量还是要远远高于 Naive 方法，特别是在可能选项事件较少的情况下，吞吐量要高于 Naive 方法几个数量级。在可能选项事件个数不多的情况下，NFA-MMG 方法的吞吐量保持在较高水平，能

比较高效地处理数据。

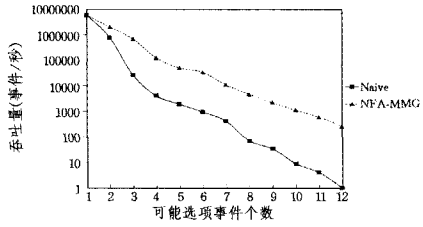

图 7　不同数量可能选项的多项概率事件流

3.3　NFA-Tree 方法实验结果及性能分析

对比 Naive 方法和 NFA-Tree 方法的吞吐量，在相同滑动窗口下 NFA-Tree 方法的吞吐量要远远大于 Naive 方法，如图 8 所示。

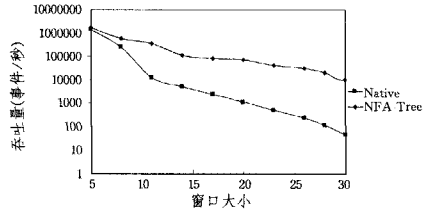

图 8　单项概率事件流复杂事件处理

使用数据发生器产生单项概率事件流，数据流中有 abcdefgh 等 8 种类型的基本事件，每个基本事件的概率值在 0 和 1 之间随机分布。从该概率事件流中查询符合模式$(a,b[],c[],d)$的复杂事件，同时考虑基于阈值的过滤，在上述数据流中的查询结果中，不仅符合模式$(a,b[],c[],d)$，且结果序列的概率值需要大于阈值 0.5。

比较 NFA-Tree 方法和进行剪枝优化后改进的 NFA-Tree 方法的吞吐量，如图 9 所示，可见优化后的 opt NFA-Tree 方法在基于阈值过滤输出结果的查询中，效率确实有明显提升。

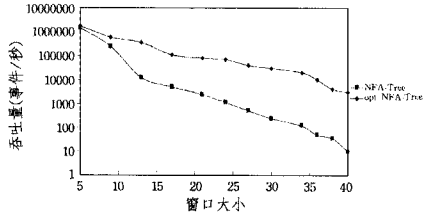

图 9　基于阈值过滤的查询

结束语　本文研究不确定数据上的复杂事件处理技术，分析 RFID 数据的不确定性，针对多项概率事件流和单项概率事件流，分别提出 NFA-MMG 模式匹配方法和 NFA-Tree 的模式匹配方法在确定数据的复杂事件处理自动机方法基础上，使用有向无环图和匹配树结合自动机，在不确定数据流上实现高效的模式匹配。针对模式匹配结果需要基于概率阈值进行过滤的需求，提出改进的 NFA-Tree 方法来提高结果过滤的效率。最后，通过实验验证了提出的模式匹配方法的性能优势。

参 考 文 献

[1] Lee I, Lee B C. An investment evaluation of supply chain RFID technologies: A normative modeling approach[J]. International Journal of Production Economics, 2010, 125(2): 313-323

[2] Shi X, Tao D, Voβ S. RFID Technology and its Application to Port-Based Container Logistics[J]. Journal of Organizational Computing and Electronic Commerce, 2011, 21(4): 332-347

[3] van der Togt R, Bakker P J, Jaspers M W. Methodological Review: A framework for performance and data quality assessment of Radio Frequency IDentification (RFID) systems in health care settings[J]. Journal of Biomedical Informatics, 2011, 44(2): 372-383

[4] Fang Y, BingWu L, LingYu H, et al. Research and Design of a Security Framework for RFID System[C]// Information Technology and Applications (IFITA), 2010 International Forum on. IEEE, 2010, 2: 443-445

[5] Kürschner C, Brintrup A, Bowman P, et al. Implementing RFID in Production Systems: A Case Study from a Confectionery Manufacturer[J]. Pacific Asia Journal of the Association for Information Systems, 2010, 2(2): 4

[6] 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)

[7] Liao G, Li J, Chen L, et al. KLEAP: an efficient cleaning method to remove cross-reads in RFID streams[C]// Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011: 2209-2212

[8] Luckham D C. The power of events: an introduction to complex event processing in distributed enterprise systems[M]. Addison-Wesley Longman Publishing Co., Inc., 2002

[9] Gyllstrom D, Wu E, Chae H J, et al. SASE: Complex Event Processing over Streams[C]// Proc of CIDR. 2007: 407-411

[10] Wu E, Diao Y, Rizvi S. High-performance complex event processing over streams[C]// Proceedings of the 2006 ACM SIGMOD international conference on Management of data. ACM, 2006: 407-418

[11] Agrawal J, Diao Y, Gyllstrom D, et al. Efficient pattern matching over event streams[C]// Proceedings of the 2008 ACM SIGMOD international conference on Management of data. ACM, 2008: 147-160

[12] Brenna L, Demers A, Gehrke J, et al. Cayuga: a high-performance event processing engine[C]// Proceedings of the 2007 ACM SIGMOD international conference on Management of data. ACM, 2007: 1100-1102

[13] Green T J, Tannen V. Models for Incomplete and Probabilistic Information[C]// IEEE Data Engineering Bulletin. 2006

[14] Kimelfeld B, Ré C. Transducing markov sequences[C]// Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems of data. ACM, 2010: 15-26

[15] Letchner J, Balazinska M. Lineage for Markovian stream event queries[C]// Proceedings of the 10th ACM International Workshop on Data Engineering for Wireless and Mobile Access. ACM, 2011: 26-33

[16] 刘海龙，李战怀，陈群，等. RFID 复杂事件检测方法的研究和改进[J]. 计算机工程与应用，2008，44(011)：5-8

[17] 陈远，李战怀，陈群. 不可靠 RFID 数据上的复杂事件处理研究[J]. 计算机应用研究，2009(07)：2537-2539

[18] 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(3)：438-451

[19] 廖国琼，李晶，万常选. 基于核密度估计的 RFID 数据流清洗方法[J]. 计算机研究与发展，2010，47(z1)