

基于数据世系管理的精准农业不确定性复杂事件处理

聂娟^{1,2} 孙瑞志¹ 邓雪峰¹ 杨昊³

(1. 中国农业大学农业部农业信息获取技术重点实验室, 北京 100083;
2. 北京农学院计算机与信息工程学院, 北京 102206; 3. 首都信息发展股份有限公司, 北京 100029)

摘要: 伴随物联网采集数据流的增加和对复杂事件匹配的准确性、可靠性要求的提高,不确定性复杂事件处理的研究得到广泛的关注。精准农业中存在大量的 RFID 及传感器监测数据,但是监测硬件和无线通信技术不能保证 100% 的可信数据,因此需要一个能应用在精准农业中处理不确定事件的流处理引擎。本文在现有流处理引擎 SASE 基础上,加入概率流理论和数据世系管理理论,提出一种新型的复杂事件处理引擎 PUCEP,可以计算输出复杂事件的概率,同时提高不确定性复杂事件的匹配效率,从而减少计算资源消耗和反应时间,提高整个系统的实时性。实验使用温室传感器采集数据,结果表明所提方法对处理概率事件流的复杂事件是有效的。

关键词: 复杂事件处理; 物联网; SASE; 世系管理; PUCEP; 温室

中图分类号: TP274+.2 **文献标识码:** A **文章编号:** 1000-1298(2016)05-0245-09

Uncertain Complex Event Processing in Precision Agriculture
Based on Data Provenance Management

Nie Juan^{1,2} Sun Ruizhi¹ Deng Xuefeng¹ Yang Hao³

(1. Key Laboratory of Agricultural Information Acquisition Technology, Ministry of Agriculture,
China Agricultural University, Beijing 100083, China
2. College of Computer and Information Engineering, Beijing University of Agriculture, Beijing 102206, China
3. Capital Information Development Limited by Share Ltd., Beijing 100029, China)

Abstract: With the increase of event flow generated from sensor kind electronic devices in IOT(Internet of things) and increasing demand of matching accuracy/confidence of more complex events, uncertain complex event processing is becoming more and more been concerned. A large number of RFID or sensor monitoring data exist in precision agriculture, but current hardware and wireless communication techniques cannot support 100% confident data. One stream processing engine which can process uncertain data in precision agriculture is needed. In this paper, a new type of complex event processing engine PUCEP(Provenance uncertain complex event processing) was proposed, in which probability flow theory and data provenance management theory were added based on the existing flow processing engine SASE. Sufficient approximate lineage query algorithm is used to calculate the probability of an event in order to improve the efficiency of probability calculation of large amount of data and the pattern matching was carried out by using the two fork tree and NFA. This optimized method can not only calculate the probability of outputs of compound events but also improve the matching efficiency of uncertain complex events, thereby reducing the computation cost and response time to a realistic degree. The experiment uses sensor data acquired from an agricultural greenhouse and shows that this method is efficient in processing complex events over probabilistic event streams.

Key words: complex event processing; internet of things; SASE; provenance management; provenance uncertain complex event processing; greenhouse

收稿日期: 2015-11-30 修回日期: 2015-12-23
基金项目: “十二五”国家科技支撑计划项目(2015BAK04B01)
作者简介: 聂娟(1977—),女,博士生,北京农学院讲师,主要从事信息物理融合系统和复杂事件处理研究,E-mail: nie_juan_10@163.com
通信作者: 孙瑞志(1963—),男,教授,博士生导师,主要从事工作流和物联网研究,E-mail: sunrz_cn@sina.com.cn

引言

物联网(IOT)架起了物理世界和信息世界之间的桥梁,被定义为基于标准和互操作通信协议的、具有自我配置能力的动态全球网络。近年来,物联网被广泛应用于重大基础设施、智能电网、物流管理等多个领域,随着信息和通信技术的快速发展,带宽和存储对于 IOT 应用已不再是约束,关键问题变为如何处理这些领域中 RFID、传感器、GPS 等设备产生的海量数据,即大数据。大数据需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力,其中复杂事件处理^[1] (Complex event processing, CEP)技术是研究物联网大数据处理的关键技术。精准农业中随着大规模农场、牧场的增多,RFID 和传感器等设备采集的数据急剧增加,而且监测硬件和无线通信技术的不完善导致采集数据的不确定性,因此需要研究精准农业中面向不确定性数据的复杂事件处理技术。

本文在现有流处理引擎 SASE 基础上,加入概率流理论和数据世系管理理论,提出一种复杂事件处理引擎 PUCEP (Provenance uncertain complex event processing),计算输出复杂事件的概率,同时提高不确定性复杂事件的匹配效率,最后使用农业温室传感器采集数据进行实验。

1 相关工作

复杂事件处理技术在主动数据库中已得到了深入研究,并有一些复杂事件处理引擎相继问世,如 SASE^[2]、Cayuga^[3]、Esper^[4] 和 ZStream^[5] 等,依据的复合事件检测模型包含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等,很多研究对检测模型和处理引擎进行了优化和改进,但面向的事件流都是确定性事件流,原始事件在处理以前都假设是清洗过的。然而精准农业物联网设备类型繁多,传感器误差、网络通信技术等多种原因都会造成事件的不确定性,另外大规模农牧业系统都是分布式部署,不同数据源的事件具有海量、异构和分散等特征,更加剧了事件的不确定性,因而伴随着大数据时代的到来,不确定性复杂事件处理成为该领域中新的研究热点。

周傲英等^[6]对已有的不确定性数据管理技术进行了总结,指出不确定性数据的表现形式,列举了多种针对不确定数据的数据模型,并介绍了数据预处理与集成、存储与索引、查询处理等方面的工作。DALVI 等^[7]进一步从理论角度综述了不确定性数

据管理的基础和面临的挑战。CHARU 等^[8]从算法与应用的角度阐述了不确定性数据管理技术。PEI 等^[9]综述了目前国内外不确定性在查询处理方面的进展,并进行了范围查询、Ranking 查询与 Skyline 查询等工作。

不确定性复杂事件的表达和识别主要有 2 种方法:概率方法和符号方法^[10]。概率方法主要包含贝叶斯分类器和隐马尔可夫模型,符号方法主要应用于设计相对简单的布尔识别有关的符号网络。目前对于不确定性复杂事件处理的研究基本采用概率方法。ROMDHANE 等^[11]提出新的事件描述语言和事件识别过程,可以对具有复杂时态约束的复合事件灵活建模,并利用概率论知识为复杂事件识别过程中不确定信息的处理提供一个一致的框架。文献[12]提出链实例队列(CIQ)的数据结构用于检测复杂事件,采用单扫描概率流来满足查询要求,使用条件概率索引树(CPI-树)存储贝叶斯网络的条件概率。KAWASHIMA 等^[13]提出另一种优化方法,不仅计算复合事件的输出概率,也获得了用户对不信任网络设备产生的不确定原始输入数据的复杂模式的可靠值。SHEN 等^[14]设计了一种查询语言,使用了一个新的数据结构 AIG 检测概率数据流的序列模式,允许用户在检测概率事件时表达克林闭包模式。WANG 等^[15]针对分布式概率事件流提出了一种高性能复杂事件处理方法,使用概率不确定性有限自动机和活动实例堆栈,并设计了一个并行算法来提高性能。曹科宁^[16]以大规模 CPS 中数据处理系统为背景,提出不确定事件流的复杂事件处理方法(USCEP)使用事件概率模型来进行概率计算,并结合匹配树和节点缓冲列表来匹配复杂事件。华盛顿大学开发的 Lahar^[17]系统是目前最流行的概率流上的复杂事件处理系统,它能有效地在马尔可夫流上进行事件查询。但是以上方法都假设各个原子事件间没有关联性,忽略了这种关联性对复杂事件概率计算效率的影响,以及利用事件间的关联性回溯初始事件,追踪事件不确定性的来源。

目前对加入数据世系管理理论的不确定性数据管理方法研究的还比较少,主要是基于斯坦福大学的 Trio 系统^[18],这个系统研究基于 ULDB 模型的不确定数据和数据世系关系的管理方法,提出了 TriQL 查询语言,并通过维护世系关系实现对不确定数据概率维的高效计算。但这个系统以数据为中心,且不能处理实时更新的数据流。胡军^[19]将概率数据库中的数据世系概念应用到不确定数据上的复杂事件概率计算中,并利用子事件的重用改进了算法效率。目前针对加入数据世系管理的不确定性复

杂事件处理技术研究的很少,本文在传统的概率流复杂事件处理引擎中加入数据世系管理理论,提高计算效率,评价数据质量,分析不确定性的来源和演化过程。

2 基于世系管理的复杂事件概率计算

从原始数据到带概率的数据流是个复杂的问题,需要经过清洗、过滤和概率抽取等步骤,本文研究重点为复杂事件处理阶段,所以忽略这个过程,研究对象直接是概率数据流。不确定性复杂事件查询可以分为模式匹配和概率计算两部分,如图 1 所示,目前概率计算多采用可能世界模型,本文加入数据世系管理理论进行计算,将两部分解耦,可以提高计算速度,增强系统的实时性。

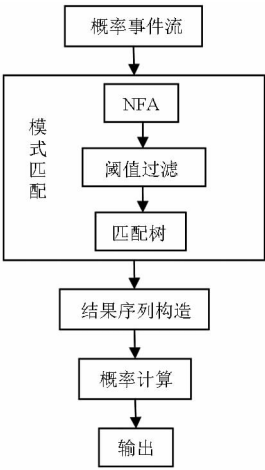


图 1 不确定性复杂事件查询流程图

Fig.1 Uncertain complex event query process

精准农业中需要采集大量的传感器、RFID 等数据,这些数据都是以流的形式进入处理器,数据周期性到达,而流数据在世系分析与管理有其自身的特点,发展到事件流也就有其特殊性。首先,原子事件到达速度很快,对每个事件进行标注和保存会给系统的吞吐量造成巨大压力;其次,事件流只能进行单遍扫描,限制了事件流世系的形式;再次,事件流处理都在内存中进行,而维护世系对空间要求很高;另外,事件流处理系统应该能够响应用户连续查询和即时查询;最后,事件流处理常有时间约束,如滑动窗口模型,数据过期了但世系不一定过期。分析可见,在事件流处理中加入世系管理是很有挑战性的一项任务。

数据产生并随时间推移而演化的整个过程信息被称为数据的世系,包含了在各种异构数据源间的数据演化过程和相同数据源内部数据的演化过程^[20]。斯坦福大学提出的不确定概率数据库 ULDB 是针对不确定性数据世系管理的典型代表,ULDB

中世系采取受限传递的策略,TriQL 是 ULDB 上的查询语言,图 2 直观地展示了 ULDB 的查询过程。

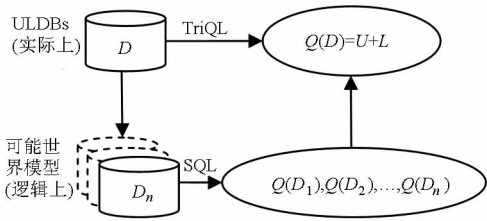


图 2 ULDB 上的 TriQL 查询

Fig.2 TriQL query in ULDB

本文选取农业温室内温湿度、光照强度、CO₂ 浓度和风机、湿帘、补光灯等设备部分行为作为基本事件,它的概率原子事件如表 1 所示。

表 1 农业温室概率原子事件

Tab.1 Agricultural greenhouse probability atom event

事件 ID	描述	发生概率
t_1	温度为 33℃	0.80
t_2	湿度为 80%	0.90
t_3	光照为 4 950 lx	0.85
t_4	CO ₂ 体积分数为 0.499 8%	0.78
t_5	土壤湿度为 66.5%	0.85
t_6	遮光板关闭	0.90
t_7	风机打开	0.89
t_8	湿帘开启	0.87
t_9	补光灯打开	0.86
t_{10}	滴灌开启	0.86

2.1 基于 ULDB 的世系查询

本文采用的世系查询基于 ULDB,并对其进行优化,使之更适合处理流数据,以下先对世系查询进行有关定义。世系使用函数 λ 来表示,是由与它相关的若干事件组成的布尔方程,对于每个基本事件 e ,它的世系可以定义为 $\lambda(e) = e$ 。

定义 1:对于一个固定模式 σ , E 世界是符合模式 σ 的元组的子集,给定一组原子事件 E 和 W 世界,如果 E 包含了与世系函数一致的所有元组,则 W 是由 E 引起的可能世界,将 W 的世系记为 1。本文用 $\lambda(E, W)$ 来定义这个世系函数,即

$$\lambda(E, W) \stackrel{\text{def}}{=} \bigwedge_{t: t \in W} \lambda_t(E) \bigwedge_{t: t \notin W} (1 - \lambda_t(E)) \quad (1)$$

在温室里每次采集环境基本参数,如空气温湿度、光照强度、CO₂ 浓度等都属于原子事件,每个参数值构成了这个元组引起的可能世界。每个参数的变化都会影响植物的生长,因此这种数据世系的管理对植物生长的预测和跟踪有很大帮助。例如室外环境参数的大变化会影响温室里的环境,进而导致一些设备的启动。表 2 列举了一些常见复杂事件的世系管理,其中,因为 t_i 为原子事件,所以 $\lambda_{t_i} = t_i$ 。

表 2 部分复杂事件世系管理

Tab.2 Part of complex event provenance management		
事件 ID	描述	世系
x_1	温度大于 30℃, 关闭遮阳板	$\lambda_{x_1} = t_1 \wedge t_6$
x_2	遮阳板已关闭, 温度大于 30℃, 开启风机	$\lambda_{x_2} = t_1 \wedge t_6 \wedge t_7$
x_3	遮阳板已关闭, 开启风机后温度仍大于 30℃, 则启动湿帘	$\lambda_{x_3} = t_1 \wedge t_6 \wedge t_7 \wedge t_8$
x_4	光照不足, 补光灯打开	$\lambda_{x_4} = t_3 \wedge t_9$
x_5	土壤湿度不足, 开启滴灌	$\lambda_{x_5} = t_5 \wedge t_{10}$
x_6	温度大于 30℃, 关闭遮阳板或者开启风机	$\lambda_{x_6} = (t_1 \wedge t_6) \vee (t_1 \wedge t_7)$

本文用 p 来定义 E 事件集中每个事件发生的概率, 原子事件 e 发生的概率记为 $p(e)$ 。

定义 2: 对一组事件 E , 概率赋值 p 是针对每个原子事件 $e \in E$ 分配的一个 $[0, 1]$ 的概率值, W 世界中的概率分布 μ 表示为

$$\mu(W) \stackrel{\text{def}}{=} \sum \lambda(E, W) \prod_{i: e_i \in E} p(e_i) \prod_{j: e_j \notin E} (1 - p(e_j))$$

(2)

给定 W 世界中的布尔查询 q , q 的边界概率 $\mu(q)$ 被定义为

$$\mu(q) \stackrel{\text{def}}{=} \sum_{q: W \models q} \mu(W)$$

(3)

表 2 中各项复杂事件的概率计算如表 3 所示, 其中已经发生的事件如风机已开等概率为 1。

表 3 复杂事件概率计算

Tab.3 Complex event probability calculation	
事件 ID	概率
x_1	$P_{x_1} = P_{t_1} P_{t_6} = 0.8 \times 0.9 = 0.72$
x_2	$P_{x_2} = P_{t_1} P_{t_6} P_{t_7} = 0.8 \times 1 \times 0.89 = 0.712$
x_3	$P_{x_3} = P_{t_1} P_{t_6} P_{t_7} P_{t_8} = 0.8 \times 1 \times 1 \times 0.87 = 0.696$
x_4	$P_{x_4} = P_{t_3} P_{t_9} = 0.85 \times 0.86 = 0.731$
x_5	$P_{x_5} = P_{t_5} P_{t_{10}} = 0.85 \times 0.86 = 0.731$
x_6	$P_{x_6} = P_{t_1} [1 - (1 - P_{t_6})(1 - P_{t_7})] = 0.8 \times 0.989 = 0.7912$

2.2 充分近似世系查询

精准农业物联网中包含类型繁多的、相互连接的设备, 这些设备不断产生海量的概率原始事件, 产生的世系查询也会呈指数级增长, 给处理系统的吞吐量带来巨大压力, 因此本文采用近似世系查询^[21], 对 Christopher 的充分世系查询进行完善, 在不影响查询结果的情况下, 进行策略追踪, 对完整世系查询的数据进行压缩, 在压缩后的数据上再进行查询, 可以极大地减少计算量从而提高系统实时性。

充分世系查询属于保守近似方法, 每一个世系函数被逻辑上具有相同含义的较小公式所替换, 例如表 2 中 x_6 复杂事件的世系被替换为 $\lambda_{x_6} = t_1 \wedge t_6$ 。充分世系的优点在于它比标准世系占用空间小得

多, 使得查询处理更有效率。另外, 了解充分世系查询的可靠性很容易, 一个充分世系查询的可靠性总是小于或等于原始世系查询的可靠性。不过, 只有单调的世系函数才可以采用充分世系方法, 周期性的世系函数需要采用多项式近似世系查询。

充分近似世系与原始世系具有相同语法, 因此可以直接用于处理现存关系数据流, 无需修改, 另外从定义即可看出一个充分近似世系查询的值只能是一个真值的下界, 不像多项式世系查询的结果比真值可高可低。实际当中原子事件发生的概率波动比较大, 充分近似世系对这种波动不太敏感, 可以取得较好的压缩比例。

定义 3: 定义 $\tilde{\lambda}^s(E, W)$ 为充分近似世系查询函数, 相对于布尔权重这里允许 $\tilde{\lambda}^s$ 分配真值。

$$\tilde{\lambda}^s(E, W) \stackrel{\text{def}}{=} \bigwedge_{t: t \in W} \tilde{\lambda}_t^s(E) \bigwedge_{t: t \notin W} (1 - \tilde{\lambda}_t^s(E))$$

(4)

与定义 1 进行比较, 发现只要将 λ_t 替换为 $\tilde{\lambda}_t^s$ 即得式(4), 同理可以给出以下定义。

定义 4: 针对定义 3, 相应的概率测量定义为 $\tilde{\mu}^s$, 可以任意分配真值权重。

$$\tilde{\mu}^s(W) \stackrel{\text{def}}{=} \sum \tilde{\lambda}(E, W) \prod_{i: e_i \in E} p(e_i) \prod_{j: e_j \notin E} (1 - p(e_j))$$

(5)

以表 3 中 x_6 复杂事件为例, $\tilde{\mu}_{x_6}^s = P_{t_1} P_{t_6} = 0.8 \times 0.9 = 0.72$, 说明温度高于阈值, 将控制设备的操作近似成只有关闭遮阳板, 忽略掉了其他操作, 虽然这种方法无法观察温室设备是否正常运作, 但在计算概率上可以省掉大量运算时间。要想保证算出的近似概率是可信的, 就必须满足以下条件。

定义 5: 对于一组原子事件 E , 给定概率分布 P , 如果

$$L_p(\tilde{\lambda}_t - \lambda_t)^2 \leq \varepsilon$$

(6)

则认为 $\tilde{\lambda}_t$ 是 λ_t 的近似值, 其中 L_p 表示由概率函数 P 引导出的原子事件分配的期望值。这里每一个事件的世系函数都能找到一个 ε 近似, 但是经过计算后发现世系函数的输入不能太少, 通常不少于 3 个, 也就是说表 3 中只有 x_6 事件的世系函数能找到 ε 近似值。

当给定一个计算复杂事件世系的公式, 如何计算其中 top - k 元素对整个结果是最有影响价值的, 并按照影响概率排序, 也是世系管理中一个重要问题。对于充分近似世系查询, 这个问题比较简单, 因为它保留了其中大量的单项式。以下给出对近似世系查询通用的查找 top - k 元素的公式。

定义 6: 查找元素 x_i 对整个 λ_i 世系函数的影响程度, 本文用 $\text{Inf}_{x_i}(\lambda_i)$ 表示, 即

$$\text{Inf}_{x_i}(\lambda_i) \stackrel{\text{def}}{=} P(\lambda_i(E) \neq \lambda_i(E \oplus \{i\})) \quad (7)$$

其中 $\oplus$ 表示对称差。

例如表 3 中 x_6 复杂事件, 它的世系函数有 3 个输入, 当 t_1 为真, t_7 为假, 只有 t_6 能唯一改变 λ_{x_6} 的结果, 这个事件发生的概率是 $0.8 \times (1 - 0.89) =$

0.088, 可见 t_6 影响 λ_{x_6} 的程度不是巧合的系数, 是可以算出的影响因子。

由以上分析可以看出世系最显著的特点是所有世系函数均是特殊类型的布尔公式——1 个 k 值单调 DNF(k-mdnf), 其中, DNF 是析取范式布尔公式。给定一个误差系数 $\varepsilon > 0$ 和 λ_i , 可以给出误差小于 ε 的充分近似世系查询算法 $\hat{\lambda}_i^s$ 的框图, 如图 3 所示。

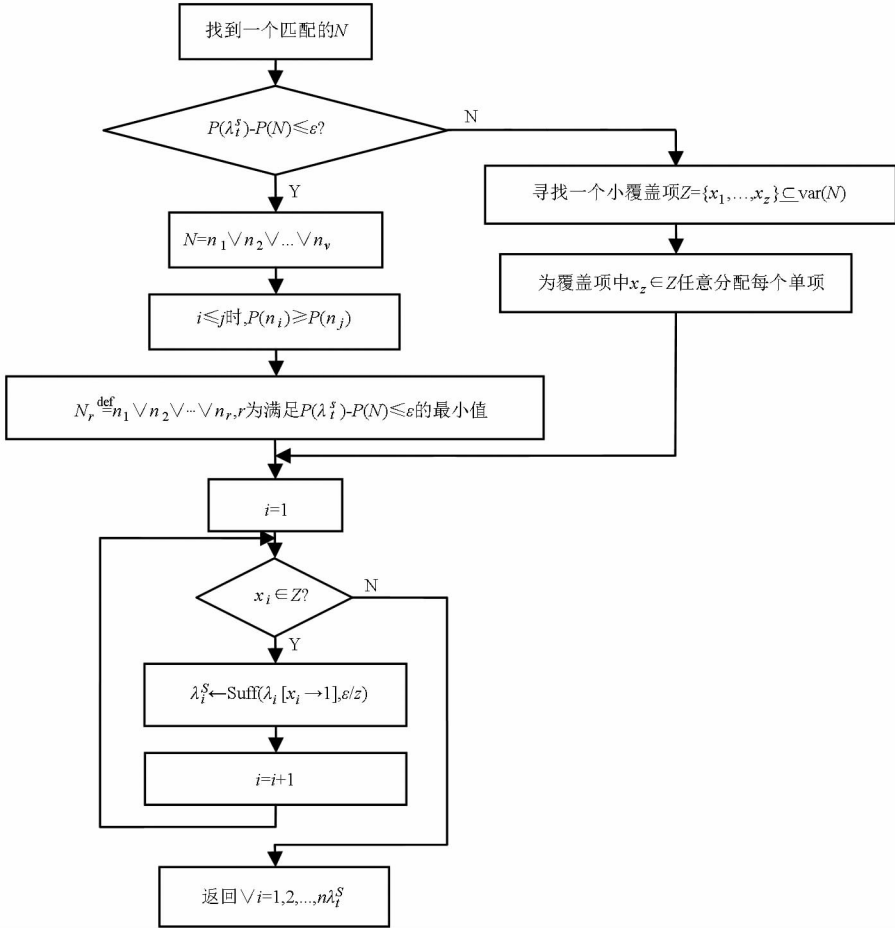


图 3 充分近似世系查询算法流程图

Fig. 3 Sufficient approximate lineage query algorithm diagram

3 不确定性复杂事件处理模型

复杂事件处理技术的一个核心问题是事件流模式匹配方法或者称为复合事件检测的基本模型, 目前主要的事件检测模型包括: 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等。而对于不确定复杂事件模式匹配, 由于数据结构的多样性和匹配情况的多样性, 问题变得极为复杂。自动机是占主导地位的处理不确定性的方法, 因此本文基于不确定性有限自动机(NFA)方法, 配合树等结构实现概率事件流上的复杂事件模式匹配。

3.1 概率复杂事件模型

采用概率方法对不确定性复杂事件进行表达,

概率计算在第 2 节中已经介绍, 以下先对概率事件进行定义。

定义 7: 概率原始事件是原始事件在某个时间的一次发生。概率原始事件表示为 (A, T, P) 。其中 A 为事件属性的集合, T 为事件发生的时间, P 为事件发生的概率, 它代表由原始信号转换为确切事件的概率。

定义 8: 概率复杂事件是由原始事件或者复杂事件按照一定的运算规则形成的事件。一个概率复杂事件表示为 (E, R, T_s, P) 。其中 E 是复杂事件的组成元素, R 为事件合成规则, T_s 为事件的时间跨度。

定义 9: 不确定复杂事件处理引擎 PUCPEP 的查询语法结构为

[FROM <input stream>]
EVENT <event pattern>
[WHERE <pattern matching condition>]
[WITHIN <time window>]
[GROUP BY <attribute_set>]
[HAVING <pattern filtering condition>]
[RETURN <output specification>]

其中,查询语言语法结构以 SASE 为基础,增加 CONFIDENCE () 操作符计算复杂事件概率。CONFIDENCE () 主要有 2 个作用:①放在 RETURN 子句中,在返回复杂事件结果的同时返回其概率。②放在 HAVING 子句中对模式匹配的复杂事件按照概率阈值进行过滤。

FROM 子句表示所查询的输入事件流;EVENT 子句声明了复杂事件的模式结构,例如 SEQ (A a, B + b [], C c) 或 AND (A a, B b); WHERE 子句表示模式匹配中需要满足的约束条件; WITHIN 子句限定了事件流上的时间窗口; GROUP BY 子句表示对符合某种条件的事件进行分组; HAVING 子句对符合匹配模式的事件做进一步过滤; RETURN 子句返回查询的结果。

以温室中温控复杂事件为例,当温度传感器监测 (以 A 表示) 到室内温度大于 30℃ 时,打开风机 (以 B 表示), 10 min 后如果室温仍然大于 30℃,则启动湿帘 (以 C 表示), 查找符合上述序列的概率大于 0.7 的复杂事件。

Q₁:FROM greenhouse stream
 EVENT SEQ (A a, B b, A + a [], C c)
 WHERE a [i]. attribute > 30 AND b. event =
" open " AND c. event = " open "
 WITHIN 10minutes
 HAVING CONF (*) > 0.7
 RETURN *

3.2 复杂事件查询处理方法

如何处理定义 9 中的查询语言,本文在 NFA 的基础上进行了优化。

NFA 自动机可以表示为一个五元组 ($Q, \Sigma, \delta, S_0, F$), Q 表示有限非空的状态集合, Σ 表示有限输入字符集, δ 表示自动机 $Q \times \Sigma \rightarrow Q$ 的状态转移函数, S_0 为自动机的开始状态, F 为自动机的终止状态,也可以叫作接受状态。图 4 所示为上述 Q₁ 示例的 NFA 模型。

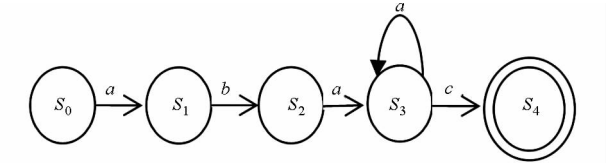


图 4 Q₁ 示例的 NFA 模型
Fig. 4 NFA model of example Q₁

对于确定性事件流的处理自动机可以很容易地进行模式匹配处理。然而,用户需要不确定事件流的输出可信度。不确定性产生过多的可能世界,计算所有可能世界的概率是不可能的,尤其当系统需要处理海量的不确定数据流。本文将二叉树理论与 NFA 结合,构造适于处理不确定事件流的 PUCEP 算法:把树的根节点作为触发状态变化的第 1 个原子事件,子节点映射到一个 NFA 为中间状态的原子事件,触发 NFA 到最终状态的事件映射成叶节点,不能再生成子节点。图 5 所示为 Q₁ 示例的自动机生成树。

图 5 显示了 NFA 自动机生成树。根节点指示第 1 个输入符号 a_1 触发 S_1 , 并创建一个新的二叉树。根节点在产生 b_1 的过程中生成 2 个子节点,左子节点表示 b_1 发生引起状态改变,而右子节点 $\bar{b}_1$ 表示 b_1 不发生,概率为 $1 - P(b_1)$ 。匹配树一直生成直到它超越时间滑动窗口被丢弃,在此期间,一旦某个输入带一个分支到叶节点 (最终状态),将生成模

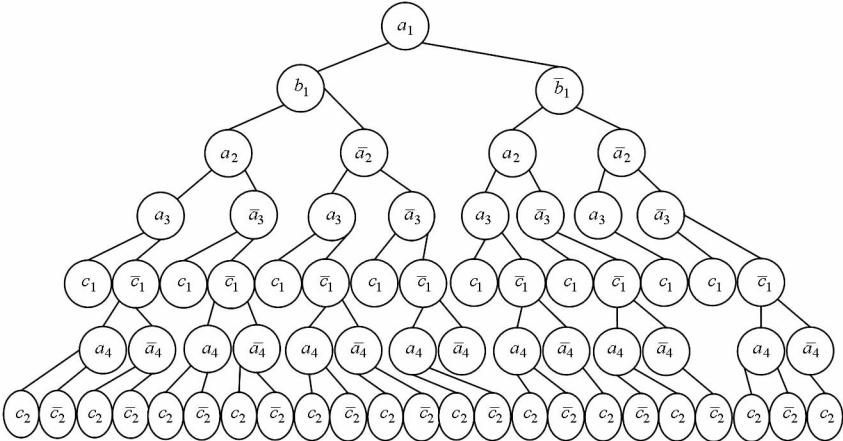


图 5 Q₁ 示例的自动机生成树
Fig. 5 Automaton spanning tree of example Q₁

式序列,并将这个带有置信概率的结果输出给用户。例如当系统处理到 c_1 , 3 个匹配序列被输出,包括: $a_1b_1a_2c_1$ 、 $a_1b_1a_2a_3c_1$ 、 $a_1b_1a_3c_1$;当系统处理到 c_2 , 7 个匹配序列被输出,包括: $a_1b_1a_2a_3a_4c_2$ 、 $a_1b_1a_2c_2$ 、 $a_1b_1a_3c_2$ 、 $a_1b_1a_4c_2$ 、 $a_1b_1a_2a_3c_2$ 、 $a_1b_1a_3a_4c_2$ 和 $a_1b_1a_2a_4c_2$ 。在输出匹配结果之前,该系统将进行统计,对同一时间产生的所有结果进行合计,最后统一算出置信概率。每一个事件序列的置信概率计算方法已在图 3 中给出,根据图 5 中自动机生成树的过程,以下给出 PUCEP 算法的伪代码:

```
1:for each e in eventstream{
2:  if e triggers initial state then
3:    if e. probability confidence then
4:      Create Qtree; Qtree. state = initialstate;
5:    else
6:      Discard e
7:    end if
8:  end if}
9: end for
10: for each childnode do
11:  if e. timestamp - T_start TimeWindow then
12:    if e triggers Qtree to next state then
13:      child node. leftchild = e; child node.
        rightchild =  $\bar{e}$ ;  $\bar{e}$ . probability = 1 - e.
        probability;
14:      update Qtree state;
15:    end if
16:    if confidence of each branch < threshold
        of confidence then prune the branch
```

```
17:      end if
18:    end if
19: end for
20: if Qtree. state = end then
21:   Output all the events stored in positive nodes
        from T_start to end
22: end if
```

以上算法中,2 处引入概率比较:第 1 处为触发初始状态的原子事件的概率比较,只有满足基本概率要求的事件才能作为原始触发事件;第 2 处为每个分支的概率与概率阈值比较,及时将不符合置信条件的分支过滤掉,在建立匹配树的同时根据阈值进行剪枝操作,提高了匹配效率。

4 实验分析

为了测试本文提出的方法,设计了一个不确定性复杂事件处理系统,该系统用 C++实现,只针对结构化数据,面向非结构化数据的复杂事件处理系统目前还处于研究初期,相关研究成果很少,是未来复杂事件处理研究的一个重要挑战。实验使用的数据源为学院温室内采集的环境参数和控制设备继电器状态,温室里安装了 4 组空气温湿度传感器、CO₂ 传感器、土壤湿度传感器和光照传感器,以及相对应的风机、湿帘和滴灌设备,传感器每间隔 10 s 向汇聚节点发送一次数据,数据流量约为 40 B/s,每小时的数据量约为 144 KB^[22]。

为了进行测试,需要将温室原始数据库中增加一个字段“事件标识”,使得所有环境参数和控制设备状态与事件类型一一对应,截取部分内容如表 4 所示。

表 4 温室事件标识对照
Tab. 4 Greenhouse event identification

事件标识	A	B	C	D	E	F	G
温室或设备参数	空气温度	风机状态	湿帘状态	土壤湿度	滴灌设备状态	CO ₂ 浓度	光照强度

为了对本文的 PUCEP 系统进行性能分析,考虑滑动窗口、序列模式、阈值对系统吞吐率、内存占用、响应时间等参数的影响。SASE 是流数据上复杂事件检测的比较著名的原型系统,在性能分析时会用 PUCEP 的指标与 SASE 的指标进行比较。考虑到农业温室的实际情况和本文实验条件,所有的注册查询都是二叉树且深度不超过 6 层。

(1)不同序列模式对系统吞吐率的影响
图 6 显示了在 PUCEP 和 SASE 系统中,3 种不同的序列模式 abac、aba[]c、aba[]cd[]e 随着滑动窗口的增大,系统吞吐率都在下降,但下降的幅度都趋于平缓,尤其在 20 以后下降幅度非常小,说明这

时候窗口大小对系统的吞吐率影响不大。前 2 个序列代表的含义在第 3 节中已有说明,aba[]cd[]e 表示在温控之后又检测了土壤湿度并开启滴灌。从图中可以看出,随着序列长度的增加和克林闭包模式的增加,系统的吞吐率都有明显的下降,但对于相同的序列,PUCEP 的吞吐率明显好于 SASE,说明对于概率流,二叉树与 NFA 结合的算法优于普通的 NFA。

(2)不同序列模式对内存占用量的影响
图 7 显示在 PUCEP 和 SASE 系统中,3 种不同的序列模式随着滑动窗口的增大,内存占用量缓慢增加,窗口大小为 20 以后,内存占用量迅速增长,说

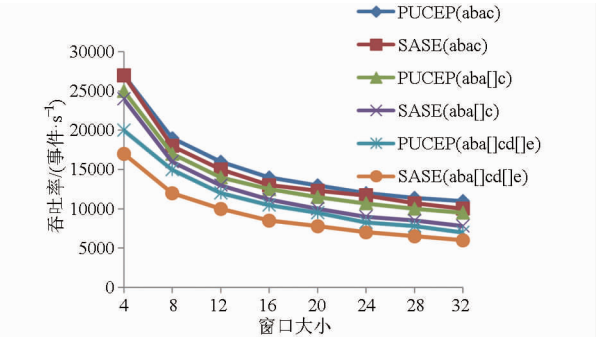


图6 不同序列模式下系统吞吐量

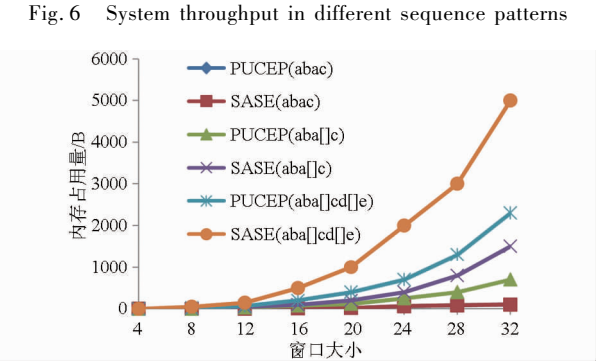


图7 不同序列模式下内存占用量

Fig. 7 Memory size in different sequence patterns

明这个节点对系统影响比较大。对于不含有克林闭包模式的序列,PUCEP不具有明显优势,伴随着序列长度的增加和模式的复杂化,PUCEP的优势逐渐显示,说明伴随着可能事件选项的增加,列出所有可能世界实例再进行匹配,内存占用显著,如果处理海量数据,必须进行算法优化。

(3) 阈值对系统吞吐率的影响

此次实验中窗口大小为 20,通过图 8 可以看出

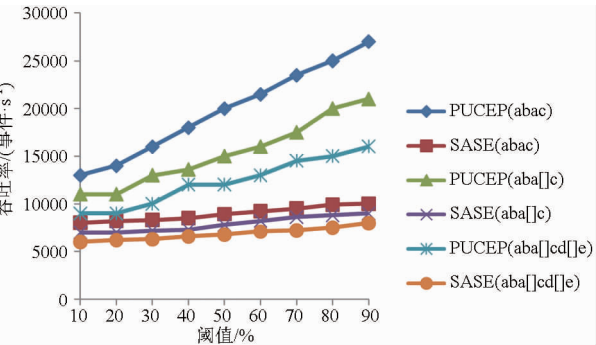


图8 不同阈值下系统吞吐量

Fig. 8 System throughput under different thresholds

采用 PUCEP 算法的 3 种序列模式受阈值影响较大,系统吞吐率随着阈值的增加迅速上升,而 SASE 算法受阈值影响很小,对概率变化不敏感,说明 SASE 算法不适用于处理不确定性复杂事件。

实验结果表明,本文提出的 PUCEP 算法在系统吞吐率、内存占用量和概率阈值响应等方面,都有较好的性能,从一定程度上验证了算法的有效性。然而,面向大量高速事件流的优化算法和世系管理策略,需要进一步研究并进行测试。

5 结束语

精准农业物联网中噪声、传感器误差、网络通信技术等各种原因,均会造成事件的不确定性,伴随着采集数据量的急剧增加,列出所有可能事件再进行模式匹配,会降低整个系统的实时性,并且系统负载迅速增加,为此,本文提出一种基于数据世系管理的不确定性复杂事件处理方法 PUCEP,运用充分近似世系查询进行复杂事件概率计算,将近似世系算法引入到复杂事件的处理中,提高大数据量的采集数据到达时概率计算的效率,另外采用二叉树与 NFA 相结合的方式模式匹配,综合了树形结构和自动机的优点,对单纯的有限自动机进行了优化,最后实验结果证明此方法在精准农业温室环境控制中是有效的。不过温室中取得的数据量很有限,本文的方法放到大规模农场中应用取得的效果需要进一步实验证明。本文的研究基于一个假设的前提,即每个事件类型都是一个单独的事件类别,相互间没有关联性,若考虑同一类别中相互关联的事件类型,事件发生的概率计算、复杂事件的语义和事件模式匹配算法都需要重新定义,涉及到的参数也将极大增加,此类问题需要继续研究。另外本文的研究中没有考虑乱序问题,实际应用中收到的数据可能有乱序现象,如何在保证复杂事件检测完整性和一致性的前提下处理事件流的时间戳问题需要进一步研究。此外本文考虑的原子事件都是基于时间点的,而实际当中有些事件是基于时间区间的,就需要研究基于时间区间的事件流处理模型和相关算法。

参 考 文 献

1 ARTIKIS A, ETZION O, FELDMAN Z. Event processing under uncertainty [C] // Proceedings of the Sixth ACM International Conference on Distributed Event-Based Systems, 2012:32 – 43.

2 WU E, DIAO Y, RIZVI S. High-performance complex event processing over streams [C] // Proceedings of the 2006 ACM SIGMOD Internatinal Conference, 2006: 407 – 418.

3 BRENNA L, DEMERS A, GEHRKE J, et al. Cayuga: a high-performance event processing engine [C] // Proceedings of the 2007 ACM SIGMOD International Conference, 2007:1100 – 1102.

4 EsperTech. Event stream intelligence: Esper & NEsper [EB/OL]. (2008 – 06 – 03) <http://www.esper.codehaus.org/>.

- 5 YUAN Mei, MADDEN Samuel. ZStream: a cost-based query processor for adaptively detecting composite events[C]//Proceedings of the 2009 ACM SIGMOD International Conference, 2009: 193 – 206.
- 6 周傲英,金澈清,王国仁,等. 不确定性数据管理技术研究综述[J]. 计算机学报,2009,32(1):1 – 16.
ZHOU Aoying, JIN Cheqing, WANG Guoren, et al. A survey on management of uncertain data[J]. Chinese Journal of Computers, 2009, 32(1): 1 – 16. (in Chinese)
- 7 DALVI N,SUCIU D. Management of probabilistic data foundations and challenges[C]//Proceedings of the 26th ACM SIGMOD—SIGACT – SIGART Symposium,2007:1 – 12.
- 8 CHARU C, AGGARWAL S, YU P. A survey of uncertain data algorithms and applications[J]. IEEE Transactions on Knowledge and Data Engineering,2009,21(5):609 – 623.
- 9 PEI J,HUA M,TAO Y F. Query answering techniques on uncertain and probabilistic data:tutorial summary[C]//Proceedings of the 2008 ACM SIGMOD International Conference, 2008: 1357 – 1364.
- 10 RYOO M S, AGGARWAL J K. Semantic representation and recognition of continued and recursive human activities[J]. International Journal of Computer Vision, 2009, 82(1):1 – 24.
- 11 ROMDHANE Rim, BOULAY Bernard, BREMOND Francois, et al. Probabilistic recognition of complex event[C]//ICVS 2011, LNCS 6962,2011: 122 – 131.
- 12 XU C, LIN S, LEI W, et al. Complex event detection in probabilistic stream[C]//Proceedings of the 12th International Asia-Pacific Web Conference, 2010: 361 – 363.
- 13 KAWASHIMA H, KITAGAWA H, LI X. Complex event processing over uncertain data streams[C]//Proceedings of the Fifth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 2010: 521 – 526.
- 14 SHEN Z, KAWASHIMA H, KITAGAWA H. Probabilistic event stream processing with lineage[C]//Proceedings of the 26th International Conference on Data Engineering, 2010:1 – 6.
- 15 WANG Yongheng, CAO Kening, ZHANG Xiaoming. Complex event processing over distributed probabilistic event streams[J]. Computers & Mathematics with Applications,2013,66(10):1808 – 1821.
- 16 曹科宁. 面向大规模 CPS 系统的复杂事件处理技术研究[D]. 长沙:湖南大学,2014.
CAO Kening. The research on complex event processing for large-scale CPS[D]. Changsha: Hunan University, 2014. (in Chinese)
- 17 RE C,LETEHNER J,BALAZINSKA M,et al. Event queries on correlated probabilistic streams[C]//Proceedings of the 2008 ACM SIGMOD Conference, 2008:715 – 728.
- 18 DALVI N,SUEIU D. Efficient query evaluation on probabilistic databases[C]//Proceedings of the 30th International Conference on VLDB, 2004: 864 – 875.
- 19 胡军. 不确定 RFID 数据上的复杂事件处理技术[D]. 南京:南京理工大学,2012.
HU Jun. Complex event processing technology on uncertain RFID data[D]. Nanjing: Nanjing University of Science and Technology, 2012. (in Chinese)
- 20 高明,金澈清,王晓玲,等. 数据世系管理技术研究综述[J]. 计算机学报,2010,33(3): 373 – 389.
GAO Ming, JIN Cheqing, WANG Xiaoling, et al. A survey on management of data provenance[J]. Chinese Journal of Computers, 2010,33(3): 373 – 389. (in Chinese)
- 21 CHRISTOPHER Re, DAN Suciu. Approximate lineage for probabilistic databases[C]//Proceedings of the VLDB Endowment (PVLDB 2008), 2008: 797 – 808.
- 22 聂娟,孙瑞志,曹振丽,等. 精准农业信息物理融合系统的事件模型研究[J]. 农业机械学报,2015,46(1):285 – 291.
NIE Juan, SUN Ruizhi, CAO Zhenli, et al. Preliminary study on event model in cyber-physical systems for precision agriculture [J]. Transactions of the Chinese Society for Agricultural Machinery, 2015,46(1):285 – 291. (in Chinese)