一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
DE2012-34(2012-12)

確率的イベント系列からの複合的イベント検出モデルについて

加藤　翔† 石川　佳治††,†,†††

† 名古屋大学大学院情報科学研究科
†† 名古屋大学情報基盤センター
††† 国立情報学研究所
E-mail: †kato@db.itc.nagoya-u.ac.jp, ††y-ishikawa@nagoya-u.jp

あらまし　大量に発生するイベントの中から，パターン照合などの技術を用いてより高次のイベントを検出しようとする複合イベント処理（CEP）に注目が集まっている．本研究では特に，各イベントに生起確率が付与された確率的イベント系列に対する CEP に着目する．与えられた正規表現のパターンに対し，ひとまとまりの意味をなす照合結果の集合を得るための，二つの複合的イベント検出のセマンティクスを提案し，それらの処理方式について述べる．
キーワード　確率的イベント系列，複合イベント処理（CEP），パターン照合

Complex Event Detection Models for Probabilistic Event Sequences

Sho KATO† and Yoshiharu ISHIKAWA††,†,†††

† Graduate School of Information Science, Nagoya University
†† Information Technology Center, Nagoya University
††† National Institute of Informatics
E-mail: †kato@db.itc.nagoya-u.ac.jp, ††y-ishikawa@nagoya-u.jp

Abstract　Complex event processing (CEP) aims to detect high-level events from a large number of events using the techniques such as pattern matching, and has attracted a lot of interest recently. In this paper, we focus on CEP for probabilistic event sequences, in which each event is associated with an occurrence probability. We propose two types of sememantics for complex event detection to obtain a bundle of coherent matches for a given regular expression pattern and describe the query processing methods.
Key words　probabilistic event sequences, complex event processing (CEP), pattern matching

1. はじめに

ユビキタスコンピューティングや医療情報処理の分野において，センシングに基づいて人々の行動状況をモニタリングしようという試みが盛んに進められている [1]〜[3]．しかし，そのようなセンサから得られたデータには多くのノイズが含まれており，人々の行動を適切に推定することは容易ではない．また，センサデータにノイズが含まれていなくても，似た動きであるが異なる行動を区別することは困難なことがある．行動認識の結果は，たとえば「時刻 1 から 10 まで徒歩，11 から 25 まで階段を上る，25 から 40 まで停止」といったように，検出された行動の系列として一般には与えられるが，認識結果の曖昧性が適切には表現できていない．そこで，このような行動認識結果を**確率的イベント系列**（probabilistic event sequence）として表現することが考えられる．たとえば，行動認識プログラムが図 1 のように各時点ごとに確率が付与された行動データを出力することが考えられる．より多くの情報が含まれていることから，このようなデータを用いた次のステップにおいてより質の高い判断を行うことが可能となりうる．

時刻 1：　walk 70%, sit 20%, car 10%
時刻 2：　walk 80%, sit 20%
時刻 3：　sit 70%, stand 30%
時刻 4：　stand 80%, bicycle 20%
⋮　　⋮

図 1　行動認識の例
Fig. 1　Example of activity recognition

他の例として，RFID タグを用いたユーザ追跡の例を挙げる．たとえばワシントン大の RFID Ecosystem プロジェクト [4] では，ユーザに RFID タグを付与して行動モニタリングを行っている．部屋や廊下などに RFID タグリーダを設置しておき，

ユーザの位置を推定しデータベース化する．その出力の一部には，「時刻 1 において部屋 A に 60 %，廊下 B に 40% の確率で存在」といった形の存在確率の情報が含まれる [5], [6]．このようなデータも，先の行動認識の例と同様，確率的イベント系列として表現することができる．

本研究では，確率的イベント系列の中から，ある特定のパターンに従う部分系列を抽出する問題を考える．ストリーム形式のデータに対するパターン照合は，データストリームに関する研究や，近年話題となっている**複合イベント処理**（complex event processing, CEP）[7] における基盤技術となっている [8]．たとえば，図 1 に示したデータにおいて，「歩いていた人が，しばらく腰かけた後で立ち上がった」という複合イベントを検出したいとする．このような要求は `walk sit`$^+$ `stand` という正規表現で記述できるが，適用した結果は

時刻 1〜4 において walk, sit, sit, stand と行動した
　確率：$70\% \times 20\% \times 70\% \times 80\% = 7.84\%$
時刻 2〜4 において walk, sit, stand と行動した
　確率：$80\% \times 70\% \times 80\% = 44.8\%$

となり，重複した時間帯において複数のマッチが発生する．確率的データストリームのパターン照合に関する既存研究 [5], [6] では，マッチしたものを確率の高い順に報告するなどのアプローチがとられる．しかし，これらの複数のマッチは，いずれもこの時間帯にパターンに合致した行動が発生したことを表しており，個々を区別することは必ずしも重要ではない．むしろ，これらの照合結果を総合的に判断して，この時間帯に該当する行動があったと判断する方が妥当であると考えられる．

このような考え方から，本稿では与えられたパターンに対する複数のマッチを統合し，ひとまとまりの照合結果としてとらえる，複合的イベント検出モデルを提案する．得られた問合せ結果は，それ自体ユーザに提示されることもあれば，次の段階としてデータマイニング等のさらに高次の処理に渡されることもありうる．後者の場合は，パターン照合処理は，必要な情報を抽出する一種のフィルタリングシステムとして働くことになる．本稿では複数のマッチングのセマンティクスを提案し，それらの処理方式について，特にデータストリーム処理の観点から議論する．

2. 確率的イベント系列

確率的イベント系列を以下のように定義する．

[定義 1] **確率的イベント系列**（probabilistic event sequence）S は，

$$S = \boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_t, \ldots \tag{1}$$

と与えられる無限の系列である．

$$\boldsymbol{e}_t = \{e_{t1}, e_{t2}, \ldots, e_{t|V|}\} \tag{2}$$

は，時刻 t における**イベント集合**（event set）であり，各 $e_{ti} \in \boldsymbol{e}_t$ を**イベント**（event）と呼ぶ．V はイベント値のドメインであり，離散的であるとする．各イベント e_{ti} には生起確率 $\Pr(e_{ti})$ が付与されているとし，

t	a	b	c	d
1	0.9	0.1		
2	0.2	0.8		
3		1.0		
4		0.4	0.6	
5			0.4	0.6

図 2　確率的イベント系列の例（その 1）
Fig. 2　Example of probabilistic event sequence (1)

$$\sum_{i=1}^{|V|} \Pr(e_{ti}) = 1 \tag{3}$$

が成り立つとする．□

以下の例では，表記を簡略化するため，V の要素を a, b, ... というアルファベットで表す．また，イベント集合を表す際には，生起確率が 0 である値は省略して，$\boldsymbol{e}_t = \{\mathtt{a}, \mathtt{c}, \mathtt{d}\}$ のように示すことにする．

先の定義にあるように，本研究では，以下のような想定を行っている．

（1） イベントの値は離散的なドメイン V からとられる．

（2） 単位時間ごとにイベント集合が得られる：イベントの欠落は考えない．また，発生したイベントが発生順には到達しない，いわゆるアウトオブオーダー（out of order）型のイベント系列は考えない．なお，ここでは「単位時間」という用語を用いたが，実際には一定時間きざみである必要はなく，イベント集合間の前後関係が明確に決まればよい．

（3） ある時点におけるイベント値の発生確率を足すと 1 になる：状況によっては，未知の値のイベントを考慮したいこともある．その場合には，$V' = V \cup \{\perp\}$ という，未知の値 $\perp$ で拡張したドメイン V' を考えればよいので，本質的な問題ではない．

（4） 時刻 t におけるイベントの生起確率 $\Pr(e_{ti})$ $(i = 1, \ldots, |V|)$ は，他の時刻におけるイベントの生起確率とは独立している：すなわち，Markovian Streams [5], [6] に見られるような，イベントの生起が時間的な相関を持つような状況は考えていない．

以上のとおり，確率が付与されており，同じ時刻にいくつかの値が同時に発生しうることを除けば，単純なイベント系列を想定している．

3. 確率的イベント系列に対するパターン照合

3.1 単純照合の例

単純な正規表現のパターン $p = \mathtt{ab}^+\mathtt{c}$ が問合せとして与えられたとする．これは，イベント a が発生した後，b というイベントが 1 回以上発生し，c というイベントが発生したというパターンを表現している．ここで図 2 に示されるような確率的イベント系列が与えられたとする．時刻 $t = 1$ から $t = 5$ までの各時刻にイベントが発生し，それぞれに対して表に示すような出現確率が付与されているとする．

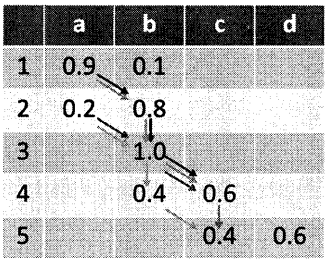

図 3　単純照合の例
Fig. 3　Example of simple match

単純にパターン照合を行うと，次のような四つの照合結果が得られる(注1)．図 3 にこれらを図示する．個々の照合結果のことを，本研究では**マッチ**（match）と呼ぶ．

$$m_1 = \langle(1,\mathtt{a}),(2,\mathtt{b}),(3,\mathtt{b}),(4,\mathtt{c})\rangle \quad (0.432)$$
$$m_2 = \langle(1,\mathtt{a}),(2,\mathtt{b}),(3,\mathtt{b}),(4,\mathtt{b}),(5,\mathtt{c})\rangle \quad (0.1152)$$
$$m_3 = \langle(2,\mathtt{a}),(3,\mathtt{b}),(4,\mathtt{c})\rangle \quad (0.12)$$
$$m_4 = \langle(2,\mathtt{a}),(3,\mathtt{b}),(4,\mathtt{b}),(5,\mathtt{c})\rangle \quad (0.032)$$

たとえば m_1 は，$t = 1,2,3,4$ においてそれぞれ a, b, b, c にマッチしたことを意味している．マッチ m に対応する時区間（time segment）を $\mathrm{ts}(m)$ で表す．たとえば $\mathrm{ts}(m_1) = [1,4]$ となる．また，マッチ m の開始時刻，終了時刻を，それぞれ $\mathrm{start\_ts}(m), \mathrm{end\_ts}(m)$ で表す．マッチ m の時刻 t におけるエントリを $m[t]$ で表し，時刻 t におけるイベントおよび確率を $m[t].\mathrm{ev}, m[t].\mathrm{pr}$ で表す．

マッチに対する確率を計算することができる．そのマッチがどの程度の確率で合致したかを表すことから，マッチの**一致確率**（match probability）と呼ぶ．上に示したマッチ結果の右端の括弧内の数値がこれを表しており，たとえば $\Pr(m_1) = 0.9 \times 0.8 \times 1.0 \times 0.6 = 0.432$ と計算される．以上より，パターン $p = \mathtt{ab^+c}$ に対する問合せ結果として，$\{m_1 : 0.432, m_2 : 0.1152, m_3 : 0.12, m_4 : 0.032\}$ という四つのマッチからなる集合を返すというのが一つの考え方である．このような照合を，本研究では**単純照合**（simple match）と呼ぶことにする．

[定義 2]　確率的イベント系列 S にパターン p が与えられたときの単純照合に基づくマッチの集合を，$\mathrm{simple\_match}(S,p)$ で表す．　□

3.2　単純照合の問題点

単純照合の考え方は分かりやすいものであり，ユーザがそのような照合を望んでいる場合も十分考えられるので，基本となるパターン照合のセマンティクスである．しかし，たとえば行動追跡において，ユーザがひとまとまりの行動を取得するためにパターン照合を行う状況を考えると，単純照合のアプローチは必ずしも適切ではない．図 2 において，イベント a, b, c を行動とみなすと，時刻 $t = 1$ 付近では a，$t = 3$ 付近では b，$t = 5$ 付近では c が発生していることが分かる．すなわち，図に示されている確率的イベント系列は実際にはひとまとまりの行動を表しているが，単純照合ではこれをまとまったものとして抽出できないことがわかる．

そこで本稿では，確率的イベント系列に対する複数のパターン照合のセマンティクスを示し，それらの処理方式について議論する．ユーザは，処理の要求に応じて適切なセマンティクスを選択することになる．

4.　パターン照合のセマンティクス

以下で定義するパターン照合のセマンティクスは，いずれもひとまとまりのイベント系列をどのように取り出すかという考え方に基づいている．どのようにまとまりを構成するかで異なるセマンティクスが得られる．

4.1　完全オーバラップ

完全オーバラップを次のように定義する．

[定義 3]　$M \subseteqq \mathrm{simple\_match}(S,p)$ を単純照合によるマッチの部分集合とする．M が**完全オーバラップ**（complete overlap）の性質を有するとは，M が

$$\forall m, m' \in M \text{ such that } m \neq m', \mathrm{ts\_overlap}(m,m') \quad (4)$$

を満たす極大な集合（任意の $\mathrm{simple\_match}(S,p)$ の要素を追加した場合に上の式が満たされなくなる）である場合をいう．ここで $\mathrm{ts\_overlap}(m,m')$ は m と m' の時区間（$\mathrm{ts}(m)$ と $\mathrm{ts}(m')$）に交わりがあるときに真になる述語である．M に対応する時区間を，M に含まれるマッチについて最大の時区間をとることで

$$\mathrm{start\_ts}(M) = \min\{\mathrm{start\_ts}(m) \mid m \in M\} \quad (5)$$
$$\mathrm{end\_ts}(M) = \max\{\mathrm{end\_ts}(m) \mid m \in M\} \quad (6)$$

により，

$$\mathrm{ts}(M) = [\mathrm{start\_ts}(M), \mathrm{end\_ts}(M)] \quad (7)$$

と定義する．　□

完全オーバラップのセマンティクスを用いると，確率的イベント系列から，それぞれ完全オーバラップの性質を持つような 0 個以上複数個の単純マッチの集合が得られる．たとえば，先の例に示した確率的イベント系列については，完全オーバラップであるマッチの集合 $\{m_1, m_2, m_3, m_4\}$ がただ一つ得られる．その対応する時区間は $[1,5]$ である．

4.2　完全オーバラップの確率

ここで問題となるのが，完全オーバラップ M についてどのように確率を付与するかである．問合せ結果として意味をなすように確率を与えたい．

ここで図 3 に示したマッチの例を見直してみると，単純照合の経路にオーバラップがあることがわかる．たとえば m_1 と m_2 には，時刻 $t = 2,3$ においてはいずれも b が対応している．行動認識のコンテクストで説明すると，複数のマッチが同じ時刻に同じ行動を共有していることになる．そのため，単純照合におけるマッチの一致確率を単純に足し合わせるのは合理的で

(注1)：データストリームに対するパターン照合では，たとえば a, f, b, ... というデータに ab というパターンのマッチを許したい場合がある．すなわち，パターンに出現しない文字（ここでは f）をスキップするというものである．本稿では簡単のため，このようなスキップを許さない処理を想定する．このような照合戦略は [8] では *strict contiguity* と呼ばれている．

ない．

図を見ると，モニタリングの対象者は，時刻 $t=1$ では a という行動をとり（確率 0.9），$t=2$ では a または b という行動をとり（両者の確率を足すと 1），$t=3$ では b という行動をとり（確率 1），$t=4$ では b または c という行動をとり（両者の確率を足すと 1），$t=5$ では c という行動をとっている（確率 0.4）．そこで本研究では，この場合の完全オーバラップ $M=\{m_1,m_2,m_3,m_4\}$ に対する確率を $0.9\times(0.2+0.8)\times1.0\times(0.4+0.6)\times0.4=0.36$ と与える．この確率は，直観的には，与えられたパターン $p=\mathtt{ab^+c}$ がこの確率的イベント系列においてどの程度成立しているかを表しており，パターンに対するマッチ集合の一致確率と考えることができる．確率が 1 となるのは，その確率的イベント系列が与えられたパターンにすべて合致した場合で，逆に 0 となるのはまったく合致しなかった場合である．この点で，確率的にも意味をなす．

以上の議論をもとに一致確率を以下のように定義する．

［定義 4］ 完全オーバラップに基づくマッチの集合 M の**一致確率**（match probability）を

$$\Pr(M)=\prod_{t\,\in\,\mathrm{ts}(M)}\ \sum_{e\in\{m[t].\mathrm{ev}\mid m\in M\}} e[t].\mathrm{pr} \qquad (8)$$

と定義する． □

ただし，$e[t].\mathrm{pr}$ は，時刻 t におけるイベント e の確率の値を表すとする．たとえば，図 2 においては $\mathtt{a}[1].\mathrm{pr}=0.9$ である．

4.3 部分オーバラップ

完全オーバラップのセマンティクスにより，与えられたパターンに対し関連するひとまとまりの照合結果をまとめることが可能となる．しかし，完全オーバラップのマッチ集合 M において任意の二つの要素が必ずオーバラップしなければならないという制約は，状況によっては強すぎることがある．例として，図 4 について，先と同様に $\mathtt{ab^+c}$ というパターンの照合を考える．単純照合の結果は

$$m_1=\langle(1,\mathtt{a}),(2,\mathtt{b}),(3,\mathtt{c})\rangle$$
$$m_2=\langle(1,\mathtt{a}),(2,\mathtt{b}),(3,\mathtt{b}),(4,\mathtt{c})\rangle$$
$$m_3=\langle(2,\mathtt{a}),(3,\mathtt{b}),(4,\mathtt{c})\rangle$$
$$m_4=\langle(4,\mathtt{a}),(5,\mathtt{b}),(6,\mathtt{c})\rangle$$

となる．完全オーバラップのセマンティクスに基づけば，$M=\{m_1,m_2,m_3\}$ および $M'=\{m_2,m_3,m_4\}$ が得られる．m_1 と m_4 の時区間がオーバラップしていないため別々のマッチ集合となる．完全オーバラップは，相互に関連が強いという点で一つの意味があるが，M,M' のように重複する要素を含む複数の結果を返してしまうことがある．

そこで，制約を大幅に緩めた部分オーバラップを導入する．

［定義 5］ $M\subseteqq\mathrm{simple\_match}(S,p)$ を単純照合によるマッチの部分集合とする．M が**部分オーバラップ**（partial overlap）の性質を有するとは，M が

$$\forall m\in M,\exists m'\in M \text{ such that } m\neq m',\ \mathrm{ts\_overlap}(m,m') \qquad (9)$$

t	a	b	c	d
1	0.9	0.1		
2	0.2	0.8		
3		0.7	0.3	
4	0.8		0.2	
5	0.1	0.8		0.1
6		0.1	0.7	0.2

図 4 確率的イベント系列の例（その 2）
Fig. 4 Example of probabilistic event sequence (2)

を満たす極大な集合である場合をいう．M に対応する時区間を完全オーバラップの場合と同様に定義する． □

部分オーバラップのセマンティクスでは，先の例の場合，$M=\{m_1,m_2,m_3,m_4\}$ という一つの結果が得られる．部分オーバラップでは，結果のマッチ集合の個数が完全マッチに比べ同等以下になるので，結果を絞り込んで提示することができるという利点もある．

完全オーバラップと部分オーバラップのどちらのセマンティクスを採用するかは，対象とするアプリケーションやユーザの意向に依存する．場合によっては，同じ確率的イベント系列に二つのセマンティクスによる照合を同時に適用し，その結果を比較することも考えられる．

5. 照合処理に関する考察

5.1 マッチの包含関係

これまでの例では問題にならなかったが，与えられたパターンによってはマッチの包含関係も発生する．たとえば $\mathtt{a^+b}$ というパターンが，$t=1$ から開始する $\mathtt{a},\mathtt{a},\mathtt{b},\ldots$ という通常の（確率的でない）イベント系列に適用されたとき，$m_1=\langle(1,\mathtt{a}),(2,\mathtt{a}),(3,\mathtt{b})\rangle$ と $m_2=\langle(2,\mathtt{a}),(3,\mathtt{b})\rangle$ という二つのマッチが得られるが，m_2 が m_1 の部分系列となっている．このようなマッチを両方とも出力するアプローチもありうるが，本研究では冗長であると考え，最長一致である m_1 のみを検出するものとする．最長一致を採用するセマンティクスは *left-maximality* [9] とも呼ばれ，また，すべての他の部分マッチを無視することは *skip-past-last* behavior と呼ばれることがある [10]．いずれもデータストリームやイベント系列処理でよく用いられる考え方である．なお，もし上記 m_2 も照合結果に含めたとしても，本研究の完全オーバラップの確率計算には影響を与えないことに注意する．すなわち，最長一致ではなくマッチをすべて検出するアプローチでも，同じ確率の計算方式が利用できる．

上記 m_1,m_2 のように包含関係にあるマッチが得られてしまうのは，与えられたパターン $\mathtt{a^+b}$ がそのような状況を発生させるパターンであることが一因となっている．たとえば，このようなパターンの指定を扱わないなどの考え方もありうるかもしれない．これについては今後の検討課題としたい．

5.2 確率の閾値の導入

確率的イベント系列を対象としたパターン照合では，生起確率が十分大きいもののみを検出したいという要求が大きい

t	a	b	c	d
1	1.0			
2	0.1	0.9		
3		0.2	0.8	
4	0.8	0.1	0.1	
5		1		
6			0.8	0.2

図 5　確率的イベント系列の例（その 3）
Fig. 5　Example of probabilistic event sequence (3)

と考えられる．そのため，確率に対する閾値を導入する．第一に，全体の照合結果に対する閾値を設定することが考えられる．この閾値を**問合せ閾値**（query threshold）と呼ぶ．たとえば，4.1 に示した完全オーバラップの例において，閾値を 20%と設定することができる．例で述べたとおり，その例では確率 0.36 であったため，閾値を満たすことから結果が出力されることになる．

一方，単純照合により得られる個別のマッチに対する閾値を設定することも考えられる．これを**マッチ閾値**（match threshold）と呼ぶ．たとえば，同じく 4.1 の例においてマッチの閾値を 5%とした場合，m_4 が閾値を満たさないため，完全オーバラップによる問合せ結果は $M = \{m_1, m_2, m_3\}$ となる．M の時区間は $[1, 4]$ になり，一致確率は $0.9 \times 1.0 \times 1.0 \times 0.6 = 0.56$ と変化する．

マッチに対する閾値は，部分オーバラップのセマンティクスにおいて必要以上にオーバラップが発生してしまうことを抑制するために使用することができる．図 5 の例を用いて説明する．

この確率的イベント系列についてパターン $\mathtt{ab^+c}$ による単純照合を行うと，

$$
\begin{aligned}
m_1 &= \langle (1,\mathtt{a}), (2,\mathtt{b}), (3,\mathtt{c}) \rangle && (0.72)\\
m_2 &= \langle (1,\mathtt{a}), (2,\mathtt{b}), (3,\mathtt{b}), (4,\mathtt{c}) \rangle && (0.018)\\
m_3 &= \langle (2,\mathtt{a}), (3,\mathtt{b}), (4,\mathtt{c}) \rangle && (0.002)\\
m_4 &= \langle (4,\mathtt{a}), (5,\mathtt{b}), (6,\mathtt{c}) \rangle && (0.64)
\end{aligned}
$$

という四つのマッチが得られる．これを見ると，m_1, m_4 の確率は高いのに対し，m_2, m_3 は確率がかなり小さいことがわかる．つまり，この場合は m_1 と m_4 が独立したイベント系列であり，m_2, m_3 はノイズのため検出されたのかもしれないと考えられる．しかし，部分マッチングのセマンティクスをそのまま用いた場合，$M = \{m_1, m_2, m_3, m_4\}$ が結果となってしまう．

そこでマッチに対する確率の閾値を設定する．たとえばマッチ閾値を 5%とすると m_1, m_4 のみが検出され，部分マッチングのセマンティクスの下でも $M = \{m_1\}$ と $M' = \{m_4\}$ という分離された結果が得られる．このように，マッチに対する閾値はノイズの影響を差し引き，マッチング結果が過度に連結されてしまうことを防ぐために利用できる．適切な閾値の設定法は今後の課題としたい．

5.3 議　　論

複合イベント処理における言語では，他にもいろいろな機能が存在する．まず，照合結果のパターンの長さに制約を設けることが考えられる．単純照合については，たとえば長さ 4 までのマッチのみを考慮するなどといった制約を取り入れることが考えられる．一方，完全オーバラップおよび部分オーバラップのセマンティクスで得られるマッチ集合 M の長さに制約を設けることは，可能ではあるがその意味についての解釈の問題が生じる．たとえば 4.1 の完全オーバラップの例において，マッチ集合 M の長さの上限を 4 と指定した場合，単純マッチ m_1, m_2, m_3, m_4 のうち m_2 は長さ 5 であり最初に除外される．残りの m_1, m_3, m_4 については，長さ 4 以下で最大数にしようとすると，$M = \{m_1\}$ および $M' = \{m_3, m_4\}$ という結果となる（いずれも長さ 4）．定義自体は可能であるが，意味がある解釈ができるかという点で今後さらに検討が必要である．

選択演算については，通常のイベント処理言語同様，容易に導入し適用することができる．一方，平均などの集約演算については，マッチ集合においてどのように集約をとらえるかについて，再び解釈の問題が発生すると考えられる．これについては今後の課題としたい．

```
PATTERN SEQ(Room a, Room+ b[], Room c)
WHERE [user_id] AND user_id = '0010'
WITH query_threshold = 0.1
  AND match_threshold = 0.05
```

図 6　問合せの例
Fig. 6　Example of query

6. 問合せ言語と問合せ処理方式

6.1 想定する問合せ言語

ここでは，複合イベント処理記述のための標準的機能を持つ SASE+ [8] の問合せ言語をベースに，想定する問合せ言語の例を示す．図 6 に SASE+ のサブセットを拡張した言語による問合せを示す．

ここでは，次のようなシナリオを考えている．`Room` という確率的イベント系列は，各部屋に設置された RFID タグリーダにより，RFID タグを付与したユーザの検出を行っている．イベントは部屋の名前（例：`a`）で表される．イベントには属性 `user_id` があり，その部屋で検出されたユーザの ID（RFID タグの番号）が保持される．問合せではユーザ ID `0010` に関するイベントを追跡し，$\mathtt{ab^+c}$ というパターンに照合する複合イベントを検出する．`[user_id]` は条件の略記法であり，入力イベントの `a`, `b`, `c` の `user_id` 属性が一致していることを意味する．`WITH` 句の閾値の設定の部分が本研究に独自のものであり，確率の閾値を設定している．

6.2 問合せ処理方式

データストリームに対するパターン照合では，有限オートマトン [11] を用いたアプローチが一般的であり，本研究でもこれに従う．特に，SASE+ [8] で用いられている，バッファ領域付きの非決定性オートマトン（$\mathrm{NFA^b}$ と呼ばれる）を処理モデルとして想定する．$\mathrm{NFA^b}$ は，非決定性有限オートマトンとマッチバッファで構成される．マッチバッファ上に照合途中のイベント系列を保持して処理を進める．たとえば，$\mathtt{ab^+c}$ という問合せパターンに対しては，図 7 のような非決定性有限オートマ

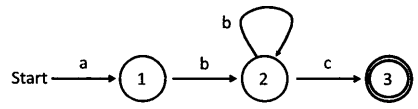

図 7 オートマトンの例
Fig. 7 Example of automaton

t	イベント系列	状態	確率	区間
1	a^1	1	0.9	
2	a^2	1	0.2	
	$a^1 \to b^2$	2	$0.9 \times 0.8 = 0.72$	
3	$a^2 \to b^3$	2	$0.2 \times 1.0 = 0.2$	
	$a^1 \to b^2 \to b^3$	2	$0.72 \times 1.0 = 0.72$	
4	$a^2 \to b^3 \to b^4$	2	$0.2 \times 0.4 = \underline{0.08}$	
	$a^2 \to b^3 \to c^4$	**3**	$0.2 \times 0.6 = 0.12$	[2,4]
	$a^1 \to b^2 \to b^3 \to b^4$	2	$0.72 \times 0.4 = 0.288$	
	$a^1 \to b^2 \to b^3 \to c^4$	**3**	$0.72 \times 0.6 = 0.432$	[1,4]
5	$a^1 \to b^2 \to b^3 \to b^4 \to c^5$	**3**	$0.288 \times 0.4 = 0.1152$	[1,5]

図 8 問合せ処理の例
Fig. 8 Example of query processing

トン（この例の場合は決定性有限オートマトンでもある）を構築する．

図 2 の確率的イベント系列にこのオートマトンを適用する例を考える．ただし，マッチ閾値は 0.1 であるとする．実行の様子を図 8 に示す．$t = 1, 2, \ldots$ と入力が来るにつれ，マッチした部分系列が延びていく様子を示している．なお，a^1 は $t = 1$ おけるイベント a を表す．状態の列で太字になっている状態番号は受理状態を表す．注目すべき点は，$t = 4$ において部分系列 $a^2 \to b^3 \to b^4$ の確率が 0.08 となっている個所である（下線部）．この部分系列については，これ以上追跡してもマッチ閾値を満たす見込みがないため，処理を打ち切ることができる．

完全マッチングと部分マッチングでは処理の進め方に違いがある．上の例で完全マッチングを考える場合，たとえば $m_1 = a^2 \to b^3 \to c^4$ に対しては時区間が [2,4] であるため，その時区間とオーバラップする可能性があるマッチが発生する可能性がある間，バッファ中に留めておく必要がある．$t = 4$ で受理状態に達したとき，同時点で受理状態に達した $m_2 = a^1 \to b^2 \to b^3 \to c^4$ と，その時点で進行中の $a^1 \to b^2 \to b^3 \to b^4$ のみが，将来を通じてオーバラップしうる全てである．よって，$t = 5$ で $m_3 = a^1 \to b^2 \to b^3 \to b^4 \to c^5$ が受理されると，m_1, m_2 に関する出力 $\{m_1, m_2, m_3\}$ が行え，バッファ上から m_1, m_2 が削除できる．m_3 については一般には削除できないが，この場合には $t = 5$ で他に進行中の処理がないことから同じく削除できる．

一方，部分マッチングに関する処理はより単純である．オーバーラップの連鎖が途切れた時点で，現在バッファ上に保持されているマッチの集合を一度に出力しバッファ上から削除する．上の例では，$t = 5$ で継続中の処理がなくなった時点で出力できる．

実際の問合せ処理では，先に述べた最大一致によるパターン照合を行うことや，他の条件に対する対応などが発生するためより複雑なものとなる．今後アルゴリズムとして体系化したい．

7. 関連研究

関連研究について簡単に紹介する．データストリームに対するパターン照合については多くの研究がある．本研究は SASE+ [8] における問合せ処理方式（特に NFA^b）をベースとしている．[8] では多くの最適化のアプローチが提案されているが，その一部は本研究にも適用可能と考える．確率的イベントストリームに対するパターン照合は Markovian Stream プロジェクト [5], [6] において研究されている．特に時間的な相関を持つマルコフ遷移による確率的推移データを対象としている．しかし，彼らの研究では照合結果をひとまとまりで検出することはできない．

8. まとめと今後の課題

本稿では，確率的イベント系列に対するパターン照合において，ひとまとまりの結果を抽出するための二つのセマンティクスについて述べた．今後はアルゴリズムの詳細化と実装技術の開発を行いたい．また，本稿ではストリーム的な処理を対象としたが，蓄積された確率的イベント系列に遡及的に問合せを行うことも考えられ，履歴ストリームに対する索引を用いること [12] などが考えられる．これについても今後の課題としたい．

謝　辞

本研究の経費の一部は内閣府最先端研究開発プロジェクト（FIRST）による．

文　献

[1] HASC (Human Activity Sencing Consortium) ホームページ, http://www.hasc.jp/.
[2] Y. Hattori and S. Inoue: "A large scale gathering system for activity data using mobile devices", Journal of Information Processing, **20**, 1, pp. 177–184 (2012).
[3] 栗原, 福田, 菅原："センサ情報からの系列パターンマイニング", 人工知能学会誌, **27**, 2, pp. 112–119 (2012).
[4] The RFID Ecosystem, http://rfid.cs.washington.edu/.
[5] J. Letchner, C. Ré, M. Balazinska and M. Philipose: "Challenges for event queries over Markovian streams", IEEE Internet Computing, **12**, 6, pp. 30–36 (2008).
[6] C. Ré, J. Letchner, M. Balazinksa and D. Suciu: "Event queries on correlated probabilistic streams", Proc. ACM SIGMOD, pp. 715–728 (2008).
[7] G. Cugola and A. Margara: "Processing flows of information: From data stream to complex event processing", ACM Comput. Surv., **44**, 3 (2012).
[8] J. Agrawal, Y. Diao, D. Gyllstrom and N. Immerman: "Efficient pattern matching over event streams", Proc. ACM SIGMOD, pp. 147–159 (2008).
[9] R. Sadri, C. Zaniolo, A. Zarkesh and J. Adibi: "Optimization of sequence queries in database systems", Proc. ACM PODS, pp. 71–81 (2001).
[10] F. Zemke, A. Witkowski, M. Cherniak and L. Colby: "Pattern matching in sequences of rows", ANSI Standard Proposal (2007).
[11] エイホ, ラム, セシィ, ウルマン："コンパイラ [第 2 版]：原理・技法・ツール", サイエンス社 (2009).
[12] J. Letchner, C. Ré, M. Balazinska and M. Philipose: "Access methods for Markovian streams", Proc. ICDE, pp. 246–257 (2009).