D-030

RDB 技術に基づくストリームデータ問合せ処理

石川 佳治 † ‡ ††　　加藤 翔 ‡
† 名古屋大学情報基盤センター　‡ 名古屋大学情報科学研究科　†† 国立情報学研究所

1　はじめに

今日では，センサ機器の普及に伴い，さまざまな環境下においてセンサ情報を取得できるようになった．そのため，センシングにより得られる大量の情報に対する分析のためのセンサデータ問合せ処理技術が重要となってきている．センサデータは，ストリーム形式で表現されリアルタイムに問合せされることもあれば，蓄積されて後で問合せされる場合もある．従来のストリームデータ処理では主にリアルタイムの処理に焦点が当てられてきたが，取得されたストリームデータに統計処理による加工を行い蓄積し，問合せ時には加工されたストリームを対象とするアプローチもある [4, 5].

本研究では，GPS や加速度センサ等を搭載した携帯機器を持ち移動するユーザの行動履歴を蓄積して分析する状況を考える．センシングにより得られたデータは必ずしも正確であるとは限らず，ノイズやデータの欠損がしばしば発生するという問題があることから，統計モデルなどを用いた生データの処理が行われる．そこで本研究では，加工され蓄積されたストリームデータの問合せに焦点を当てる．特に，*Markovian Streams* [4, 5] で提案されているような，確率モデルに基づきセンサストリームデータを処理した結果である**確率的データストリーム**を対象として考える．確率的ストリームの蓄積と問合せ処理には，リレーショナルデータベース技術を積極的に活用するものとする．本研究は，我々のグループの過去の研究 [3] の拡張と位置づけられるが，RDBMS を直接活用するのではなく，後述するとおり Datalog 言語を用いる点が異なる．

2　確率的データストリームの表現

ここでは、モバイルユーザの行動分析を例としてとりあげる．各種センサ（例：加速度センサ，GPS）により，ユーザの位置や行動がセンシングされたとする．そのような生のセンシングデータに対し，確率モデルを用いたデータ処理を行い，その結果をリレーションの形で表現することを考える．

図 1 は，各ユーザの各時点における存在情報を表すリレーション Loc である．属性 x, t はそれぞれユーザ ID と時刻を表す．属性 c はユーザが位置するグリッドセルの番号を表す．ここでは，**占有格子地図**（occupancy grid map）[6] の考え方を用いており，ある時点においてどのグリッドセルにユーザが存在しているかを表現する．センサ情報が必ずしも正確ではないため，そのセル内に存在している可能性を**存在確率** $p = \Pr_{x,t}(c)$ で表す．たとえばユーザ A は時刻 $t = 5$ の時点において，それぞれ確率 60%, 30%, 10% でセル 10, 12, 8 内に存在すると推定されている．

一方，図 2 に示すリレーション Tran は，各ユーザの**遷移確率**の情報を表している．これは，ある時刻 t にあるグリッドセル c 内にいたユーザが，次の時刻においてどのセル（c'）に移動したかを確率 $p = \Pr_{x,t}(c'|c)$ を付与して表現したものであり，粒子フィルタ（particle filter）などの統計的モデルにより，センシングされたデータを処理して得られる．マルコフ連鎖を想定した表現で内部的なモデルを出力したものといえる．

x	t	c	p
A	5	10	0.60
A	5	12	0.30
A	5	8	0.10
B	5	3	0.80
B	5	8	0.15
B	5	10	0.05
A	6	12	0.45
⋮	⋮	⋮	⋮

図 1　存在情報：Loc

x	t	c	c'	p
A	5	10	10	0.60
A	5	10	12	0.20
A	5	10	8	0.05
A	5	10	9	0.15
A	5	12	10	0.15
A	5	12	12	0.70
A	5	12	8	0.05
A	5	12	9	0.10
⋮	⋮	⋮	⋮	⋮

図 2　遷移情報：Tran

図 3 には，加速度センサなどのデータをもとに分析された結果としての行動情報を表すリレーション Behav を示す．各ユーザが各時刻においてどのような行動をしていたかの推定が，確率 $p = \Pr_{x,t}(b)$ とともに管理される．行動情報の確率はユーザの位置にも依存するが，ここでは加速度等のセンサのみから行動が推定されたと考え，行動の確率は存在・遷移確率とは独立と考える．

x	t	b	p
A	5	walking	0.60
A	5	running	0.30
A	5	other	0.10
⋮	⋮	⋮	⋮

図 3　行動情報：Behav

このようにアーカイブ表現された大量の確率的ストリームに対して分析問合せを行う機能を実現することが本研究の目的である．関連研究 [4, 5] は確率的データストリームを対象としているが，移動に関するデータ（上記の Loc, Tran）のみを対象として単一ユーザの移動のみを考慮していた（例：「A がオフィス 1 から 2 へいつ移動したかを確率を付与して提示せよ」）．これに対し本研究では，行動状況などの他の情報と統合した問合せを行うことを目的とし，また，大規模なデータの処理のためにリレーショナルデータベースの問合せ機能を効果的に用いることを目指す．

3　行動パターン問合せ

本研究における問合せは，**行動パターン**を図式表現することで与える．図 4 に例を示すが，これは以下のような行動パターンにマッチする．

> 時刻 $t = 5$ にセル 12 内にいて，セル 13 に徒歩で移動して，その後時刻 $t = 10$ まで走った

Query Processing on Stream Data Based on Relational Database Technologies
Yoshiharu Ishikawa† ‡, ††, Sho Kato‡,
†Infomation Technology Center, Nagoya University,
‡Graduate School of Information Science, Nagoya University,
††National Institute of Informatics

```
Vis1(X, 6, C, P, M) ← Loc(X, 5, 12, P1), Tran(X, 5, 12, C, P2), P := P1 * P2, M := (12, C)
Vis1(X, T, C, P, M) ← Vis1(X, T1, C1, P1, M), Tran(X, T1, C1, C, P2), Behav(X, T1, "W", P3),
                      T := T1 + 1, P := P1 * P2 * P3, M := Append(M, C)
Vis2(X, T, 13, P, Y, M) ← Vis1(X, T, 13, P, M), Y := T
Vis2(X, T, C, P, Y, M) ← Vis2(X, T1, C1, P1, Y, M), Tran(X, T1, C1, C, P2),
                    Behav(X, T1, "R", P3), T := T1 + 1, P := P1 * P2 * P3, M := Append(M, C)
Res(X, Y, Z, P, M) ← Vis2(X, 10, C, P, Y, M), Z := C
```

図 5　Datalog を用いた再帰的問合せ

図において，各ノードはグリッドセルに対応し，各ノードには時刻による制約条件が付与されている．セル 13 については $t = Y$ と表されているが，Y はパターンにマッチした行動履歴のインスタンスに応じて値が設定される変数である．ノード 12 には $u = X$ という制約条件があるが，これはそれ以降の 2 つのノードに引き継がれると考える．ノードをつなぐ二つのリンク上には，行動に関する制約条件が付与されている．3 番目のノードには，セルに関する変数 Z が設定されている．

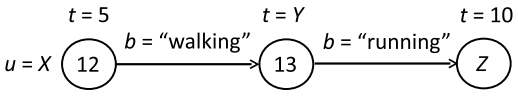

図 4　行動パターンの例

このような問合せパターンが与えられたとき，制約を満たす行動パターンのインスタンスを見つけることが問合せの目的となる．問合せの結果として，指定された変数（この例では X, Y, Z）に値の束縛がなされる．問合せ対象のデータは確率的データストリームであるため，実際には指定された行動パターンにある確率でマッチするインスタンスを探すことになる．確率が高い順にランク付けした結果を返す**ランク付け問合せ**の機能が求められる．

4　問合せ処理のアイデアと課題

先のように与えられた問合せを処理するためのアイデアについて述べる．ここで考える問合せは，制約を考慮しつつ再帰的に遷移情報をたどる処理により処理できることから，再帰的問合せの能力が重要となる．そこで，図 4 のように与えられた問合せパターンを，図 5 のような Datalog [1] を用いた再帰的プログラムに変換する．1, 2 行目のルールは idb 述語 `Vis1` を定義しており，時刻 t にセル 12 にいたユーザが，徒歩（`"W"` により表現）により進んだセルとその時刻の情報を収集する．`M` はリストであり，ユーザが通った経路を保持する．3 番目のルールではセル 13 を通過したという制約を導入し，セル 13 以降の情報収集のための idb 述語 `Vis2` を定義している．4 番目のルールでそれを再帰的に処理し，最後のルールで結果をとりまとめる．

複雑な行動パターンに対する問合せについては，たとえば [2] における時空間データベースを対象としたものがあるが，想定するパターンに応じた独自の問合せ処理アルゴリズムが提案されていた．これに対し，宣言的な問合せ言語である Datalog にいったん変換することにより，問合せ処理の見通しが立てやすくなるだけでなく，最適化の戦略を系統だって適用することが可能になると考えられる．また，Datalog はシンプルでありながら表現能力が高いため，さまざまな制約を直接的に表現できるという利点が存在する．

図 4 のような入力から図 5 のような Datalog プログラムを生成する手法を開発することが本研究の当面の目標である．今後の課題としては以下があげられる．

- 行動パターンの表現言語の検討：分析要求を満たす十分な表現能力と，効率的に処理可能というトレードオフを考える必要がある．
- 問合せパターンの Datalog プログラムへの変換手法の開発
- 効率的な問合せ戦略の導入：先に示した問合せを単純に実行すると，正確な結果は得られるものの，効率の面では改善の余地が存在すると考えられる．具体的にはいくつかの方策が考えられる．
  - マジックセット [1] などの Datalog の効率的な実行処理手法の導入
  - 索引の構築：Markovian Stream プロジェクトでは，本研究と問合せ処理のアプローチは違うものの，効率化のための索引を提案している [4]．本研究においても，索引（もしくは実体化ビュー）を効果的に用いることが考えられる．
  - ランク付け問合せへの対応：スコア（確率）がトップ k 件のみの結果を高速に求めたいという要求がしばしば現れる．トップ k 件の候補となるインスタンスを早期に検出し，問合せ結果となりえないものを枝刈りする手法が開発できれば，効率の面で有効であるといえる．
- RDBMS の有効活用：最近の RDBMS には，SQL の再帰問合せ機能を提供しているものが存在している．この機能を積極的に用いれば，Datalog 問合せ処理の多くの部分を RDBMS にプッシュダウンできる可能性が存在する．

謝辞

本研究の一部は，内閣府最先端研究開発プロジェクト（FIRST）による．

参考文献

[1] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995.

[2] M. Hadjieleftheriou, G. Kollios, P. Bakalov, and V. J. Tsotras. Complex spatio-temporal pattern queries. In *Proc. VLDB*, pp. 877–888, 2005.

[3] 堀田 孝司, 石川 佳治. RDB を用いた移動履歴からの移動パターン問合せ処理手法. DEIM フォーラム 2011, 2011 年.

[4] J. Letchner, C. Ré, M. Balazinska, and M. Philipose. Access methods for Markovian streams. In *Proc. ICDE 2009*, pp. 246–257, 2009.

[5] C. Ré, J. Letchner, M. Balazinksa, and D. Suciu. Event queries on correlated probabilistic streams. In *Proc. ACM SIGMOD*, pp. 715–728, 2008.

[6] S. Thrun, W. Burgard, and D. Fox. 確率ロボティクス. 毎日コミュニケーションズ, 2007.