

データベース処理のデータフロー言語

北海道大学・工学部

田 中 譲

1. 序論

現在著者等が開発中のデータストリーム処理方式データベースマシンのためのデータフロー言語について試案を述べる。このマシンはサーチエンジン (SEE: SEarch Engine) とソートエンジン (SOE: SOrt Engine) と名付けた2種の機能モジュールを用いて、データベース処理のボトルネックとなるデータの転送と、サーチ、ソートのデータベースにおける基本演算の処理とを重畳させ、データベース処理の高速化を図ろうというものである。⁽¹⁾⁽²⁾⁽³⁾ 本論文では、このマシンのアーキテクチャ、SEEとSOEのアーキテクチャに関する説明は省略する。

我々は、このマシンのシステム言語を3つのレベルに分けて考えている。最上位レベルは、更新演算を含むように拡張した関係代数言語で、第2レベルは、これよりやや記述のレベルが低く、よく似ている原始関係代数言語、そして最下位レベルが、データベースマシンの各構成モジュールの機能とよく対応のとれいるデータストリーム言語である。

CODASYL DML インターフェースや QBE インターフェース等は、システム言語の最上位にある関係代数言語との論理的なマッピングの問題なので、ここには触れないおく。

関係代数言語とデータストリーム言語の間に原始関係代数言語を置いたのは、オブジェクトコードの最適化や、スケジューリングの問題をいさる限り、関係代数言語と原始関係代数言語との間のマッピングの問題として論理的に扱えたという心積もりからである。本論文では、最適化およびスケジューリングの問題も避け、原始関係代数言語とデータストリーム言語の間の変換について述べる。さらに、更新演算についても本論文では触れないおく。

2. 原始関係代数言語

原始関係代数は関係代数と同じ演算を持つが、属性を属性名で参照することができず、代わりに指標を用いる。

$R(A, B, C)$, $S(D, E, F)$, $T(G, H, I)$ に対して、関係代数式の例とこれに対応する原始関係代数式をいくつか列挙する。

$R[A, B]$

$[1, 2] R$

$R[B = F] S$

$[2 = 3] R S$

$R[C = 'v']$

$\{3 = 'v'\} R$

$$R[A=B]$$

$$\{1=2\} R$$

$$(R[B=F]S)[C,D]$$

$$[3.1, 1.2][2=3]RS$$

注, R を 1, S を 2 で表わし,
 A, B, C を 1.1, 2.1, 3.1 で
 表わす.

$$((R[B=F]S)[E=I]T)[C,D,H]$$

$$[3.1.1, 1.2.1, 2.2][2.2=3]$$

$$\begin{array}{ccc} \uparrow & \uparrow & \uparrow \\ 1.1 & 2.1 & 2 \\ \hline & \uparrow & \\ & 1 & \end{array}$$

$$[2=3]RST$$

指標 $i_0, i_1, \dots, i_n$ に対して, i_0 をヘッド, $i_1, \dots, i_n$ をテイル
 と呼ぶ。指標 σ のヘッドとテイルを $h(\sigma)$, $t(\sigma)$ で表わす。

通常の結合演算 $R[X=Y]S$ の他に, 半結合演算

$$R[X=Y]S \equiv \{r \mid r \in R \text{ st. } \exists s \in S \text{ } r.X = s.Y\}$$

を考慮する。

原始関係代数言語では, プロジェクションをとらないう限り,
 関係中のタプルの各値を見ることはできないとする。

3. データストリーム言語

データタイプ α のデータストリームとは, データタイプ α
 に属するデータの有限系列のことである。タイプ α のデータ

$a_0, a_1, \dots, a_n$ がこの順に並んだデータストリームを,
 $\langle a_0, a_1, \dots, a_n \rangle$ で表わす。データタイプ α のデータスト
 リームの集合を α^* で表わす。

リストを以下のように定義する。

- (1) a_i をタイプ α_i のデータとするとき, $(a_0, a_1, \dots, a_n)$
 はタイプ $(\alpha_0, \alpha_1, \dots, \alpha_n)$ のリストである。
- (2) b_i をタイプ β_i のデータないしはリストとするとき,
 $(b_0, b_1, \dots, b_n)$ はタイプ $(\beta_0, \beta_1, \dots, \beta_n)$ のリスト
 である。
- (3) (1)(2) を有限回適用して得られるものだけがリスト
 である。

タイプ α のリストのストリームの集合を α^* で表わす。デー
 タタイプ α の要素のみを用いて得られるリストのストリーム
 のすべいを $\mathbb{L}_\alpha$ で表わすことにする。 i を正整数とするとき,

$$\bullet i(a_1, a_2, \dots, a_n) = \begin{cases} a_i & i \leq n \\ \text{undefined} & \text{otherwise.} \end{cases}$$

と定義する。したがって、例えば

$$\bullet 2.3(a, (b, c), (e, (f, g), h)) = (f, g)$$

である。

データストリーム言語で扱うデータタイプには、関係名のタイプ R 、データベースに現われる値のタイプ D 、非負整数のタイプ N と、ストリーム型 D^* 、 N^* 、 $\mathbb{L}_D$ がある。

以下にデータストリーム言語の関数演算子を列挙する。

$$! : R \rightarrow N^*$$

関係 R のタプル数を n とするとき、

$$!R = \langle 0, 1, 2, \dots, (n-1) \rangle$$

である。

$$? : R \rightarrow (N \rightarrow (N^* \rightarrow D^*))$$

$$?R \ i \langle i_1, i_2, \dots, i_j \rangle$$

i が R の属性数より大きいとき空ストリーム、

そうでなければ R の j 番目のタプルの i 属性の値を a_j とするとき

$$\langle a_{i_1}, a_{i_2}, \dots, a_{i_j} \rangle$$

とする。

$$\uparrow : D^* \rightarrow N^* \quad (\downarrow : D^* \rightarrow N^*)$$

$$\uparrow \langle a_0, a_1, \dots, a_n \rangle \quad (\downarrow \langle a_0, a_1, \dots, a_n \rangle)$$

$a_0, a_1, \dots, a_n$ を昇順(降順)にソートした

j のを $a_{i_1}, a_{i_2}, \dots, a_{i_{n+1}}$ とするとき、

$$\langle i_1, i_2, \dots, i_{n+1} \rangle$$

を値とする。

$$S_0: \mathbb{D} \rightarrow (\mathbb{D}^* \rightarrow \mathbb{N}^*)$$

$$S_0 v \langle a_0, a_1, \dots, a_n \rangle$$

$a_i \theta v$ となる i を順に $i_0 < i_1 < \dots < i_k$ とすると

※

$$\langle i_0, i_1, \dots, i_k \rangle$$

を値とする。

$$R_0: \mathbb{D}^* \rightarrow (\mathbb{D}^* \rightarrow \mathbb{N}^*)$$

$$R_0 \langle a_0, a_1, \dots, a_n \rangle \langle b_0, b_1, \dots, b_m \rangle$$

$a_i \theta b_i$ となる $0 \leq i \leq \min(n, m)$ を $i_0 < i_1 < \dots < i_k$ と

すると、

$$\langle i_0, i_1, \dots, i_k \rangle$$

を値とする。

$$X_0^L: \mathbb{D}^* \rightarrow (\mathbb{D}^* \rightarrow (\mathbb{N}, \mathbb{N})^*)$$

$$X_0^R: \mathbb{D}^* \rightarrow (\mathbb{D}^* \rightarrow (\mathbb{N}, \mathbb{N})^*)$$

$$X_0^L \langle a_0, a_1, \dots, a_n \rangle \langle b_0, b_1, \dots, b_m \rangle$$

$$(X_0^R \langle a_0, a_1, \dots, a_n \rangle \langle b_0, b_1, \dots, b_m \rangle)$$

$i = 0, 1, 2, \dots, m$ の順に, $a_{j_i} \theta b_i$ となる最小 (最大)

の j_i を求め (i, j_i) を出力する。 $\langle a_0, a_1, \dots, a_n \rangle$

は $a_0 \leq a_1 \leq \dots \leq a_n$ の場合に限る。

$$A : \mathbb{L}_D \rightarrow (N^* \rightarrow \mathbb{L}_D)$$

$$A \langle a_0, a_1, \dots, a_n \rangle \langle i_0, i_1, \dots, i_m \rangle$$

$j = 0, 1, 2, \dots, m$ の順に, $i_j \leq n$ であれば a_{i_j} を出力する。

$$E : N^* \rightarrow (N^* \rightarrow (N^* \rightarrow (N, N)^*))$$

被演算項である 3 つのタイポ N のデータストリームの各々の k 要素を i_1, i_2, i_3 とする。これら k 要素が $i_2 \leq i_3$ のとき, これらに対して, サイストリーム

$$\langle (i_1, i_2), (i_1, i_2 + 1), \dots, (i_1, i_3) \rangle$$

を出力する。

$$I : \mathbb{L}_D \rightarrow \mathbb{L}_D$$

$s \in \mathbb{L}_D$ に対し, $Is = s$ である。

$$\cdot : N \rightarrow (\mathbb{L}_D \rightarrow \mathbb{L}_D)$$

$i \geq 1$ に対し, $\cdot i$ は前に定義したものに等しい。

$$P : \mathbb{L}_D \rightarrow (\mathbb{L}_D \rightarrow \mathbb{L}_D)$$

$$P \langle a_0, a_1, \dots, a_n \rangle \langle b_0, b_1, \dots, b_m \rangle$$

$$= \langle (a_0, b_0), (a_1, b_1), \dots, (a_{\min(n,m)}, b_{\min(n,m)}) \rangle$$

さらに K, C を,

$$K = \lambda x y z. P x z y z, \quad C = \lambda x y z. x z y z$$

と定義する。これらを用いた式がデータストリーム言語のプログラムである。

ここに述べたデータストリーム言語の各演算子は、著者等が開発中のデータベースマシンの各モジュールによって直接実行される。例えば $\uparrow$, $\downarrow$ は SOE で処理され、 $?$, X_0^L , X_0^R , A は SEE で処理される。⁽³⁾ これらの事柄の詳細は省略する。

4. 原始関係代数言語とデータストリーム言語の変換

原始関係代数言語のある式において、指標 σ の指す属性を持つ関係を $r(\sigma)$ で表わすことにする。原始関係代数言語の各演算子は、データストリーム言語の演算子を用いて以下のよう展開できる。 x, y はタイプ $\mathbb{L}_N$ の変数とする。

関係 R	$! R$
射影 $[\sigma_1, \sigma_2, \dots, \sigma_n]$	$\lambda x. (?r(\sigma_1)\sigma_1 x, ?r(\sigma_2)\sigma_2 x, \dots, ?r(\sigma_n)\sigma_n x)$
選択 $\{\sigma \theta v\}$	$\lambda x. A x S_0 v ?r(\sigma)\sigma x$
制約 $\{\sigma \theta \sigma'\}$	$\lambda x. A x R_0 C (?r(\sigma)\sigma) (?r(\sigma')\sigma') x$
半結合 $\langle \sigma \theta \sigma' \rangle$	$\lambda x y. A x .1 X_0^L A ?r(\sigma')\sigma' y \uparrow ?r(\sigma')\sigma' y ?r(\sigma)\sigma x$
結合 $[\sigma \theta \sigma']$	$\lambda x y. K (A x .1) (A y .2) C (\uparrow ?r(\sigma)\sigma x .1) (.1)$

$$X_0 A ? r(\sigma) \circ x \uparrow ? r(\sigma) \circ x \quad ? r(\sigma') \circ y$$

$$X_0 \triangleq \lambda x y. \in (.1 X_0^L x y) (.2 X_0^L x y) (.2 X_0^R x y)$$

上述の関係をを用ゐることにより、原始関係代数学の任意の式をデータストリーム言語に変換することが可能となる。このときのオブジェクトプログラムは、中間結果を関係として求めることなく最終結果を与える。

上の定義において、例えば「 $\sigma \theta \sigma'$ 」のオブジェクトコードで、母式の x と y を入れ換えたものの「 $\sigma \theta \sigma'$ 」のオブジェクトコードと考えることができ、しかも一般に両者は処理に要する時間が異なる。こゝでオブジェクトコードの最適化、スケジューリングの問題が生じる。しかし、今の例で後者のオブジェクトコードを必要とするのであれば「 $\sigma \theta \sigma'$ 」 $x y$ とする所を「 $\sigma' \theta \sigma$ 」 $y x$ とすればよい。このことは、他の演算子についても言え。最適化とスケジューリングの問題は、データストリーム言語のレベルより一段上に引き上げられ、原始関係代数学のレベルで議論することが出来る。したがって、上述のように、各演算子のオブジェクトコードを固定してしまふといふ。

5. 結論

本論文では、現在著者等が開発中のデータストリーム処理方式データベースマシンのデータフロー言語として、原始関係代数言語と、マシンの構成要素の機能に対応した演算子からなるデータストリーム言語の2つを示し、前者の演算子の後者の演算子を用いた一意な展開形を与えた。これにより、検索処理に限っては、関係代数とデータストリーム言語の美しい対応関係がとれた。

今後の課題としては、更新処理を含めた拡張がある。さらに、データストリーム言語の体系そのものの単純化が課題の1つとしてある。このためには、組み合わせ論理で用いられるコンビネータを用いた体系化等が考えられる。

参考文献

- (1) Y. Tanaka, Y. Nozaka, and A. Masuyama, "Pipeline Searching and Sorting Modules as Components of a Data Flow Database Computer," Proc. IFIP Congress 80, 1980 (Tokyo), pp 427-432.
- (2) 田中 譲, "データストリーム処理方式のデータベースコンビネータ," 情報処理研究, 記号処理12-14, 1980 pp 97-103.
- (3) 田中 譲 他, "データストリーム処理方式のデータベースコンビネータ," 情報処理全国大会, 1980, pp 493-494.