フィルタリングのためのユーザ要求記述言語 FilteringSQL について

澤井 里枝　　塚本 昌彦　　寺田 努　　西尾 章治郎

大阪大学大学院情報科学研究科
{rie,tuka,tsutomu,nishio}@ist.osaka-u.ac.jp

近年，さまざまなデータ放送の普及により，ユーザが必要とするデータのみを自動的に抽出する情報フィルタリングに対する注目が高まってきた．そこで筆者らはこれまで，フィルタリングのためのユーザ要求記述言語であるフィルタリング SQL を提案し，ユーザの嗜好やデータ管理ポリシーを自由に表現できるようにした．しかし，これまでの研究では，フィルタリング SQL の記述構文を設計するのみであったため，ある受信データ集合に対して実際にどのようなフィルタリング結果が得られるかは明確でなかった．そこで本研究では，フィルタリング SQL の意味を明確にすることを目的とする．本研究では，フィルタリングを関数として表現したフィルタリング関数によりフィルタリング SQL の意味を表現する．フィルタリング関数の数学的性質を利用することで，ユーザが記述した要求を実行するフィルタリングの特性が明らかになり，環境に応じた効率的な処理を実現できる．
キーワード 情報フィルタリング，フィルタリング SQL，フィルタリング関数，言語，意味

FilteringSQL: a User Request Description Language for Information Filtering

Rie SAWAI　　Masahiko TSUKAMOTO　　Tsutomu TERADA　　Shojiro NISHIO

Graduate School of Information Science and Technology, Osaka University

In recent years, due to the increasing popularization of data broadcasting, there is an increasing demand for filtering techniques that automatically extract only the necessary data. In our previous works, we have proposed FilteringSQL which is a user request description language for information filtering. FilteringSQL enables users to express their preferences and policy of data management. However, since we designed only sentence structure of FilteringSQL in previous works, it is not clear how filtering results users will get for given data sets. In this paper, we clear the semantics of FilteringSQL. We express the semantics of FilteringSQL using filtering function that represents information filtering as a function. Moreover, by using the mathematical properties of filtering function, we clarify the characteristics of filtering methods that carry out the request users described. From the properties of filtering methods, more efficient processing can be achieved according to environments.
Keywords Information Filtering, FilteringSQL, Filtering Function, Language, Semantics

1　はじめに

近年，ネットワークのブロードバンド化や，放送のデジタル化および多チャンネル化により，さまざまな放送型サービスが提供されるようになった [7, 8]．このような環境では，多様で膨大なデータを受信できるが，一般にユーザが必要とする情報はごく一部に限られているため，受信データから必要なデータを探し出すことはコストの高い作業である．そこで，自動的に受信データの取捨選択をするフィルタリング機構が多数提案されている [2, 3, 6, 9]．しかし，各フィルタリング機構は，キーワードマッチングや関連フィードバックなど，それぞれ独自の手法によってデータのフィルタリングを行っているにもかかわらず，それらの手法を定性的に表現する数学的基盤がなかった．そのため，フィルタリングの性質の定性的な評価や処理手法の最適化，宣言的なフィルタリング言語の設計などができなかった．そこで，筆者らはこれまでにフィルタリングを関数として表す**フィルタリング関数**を定義し，フィルタリングの性質をフィルタリング関数が満たす制約条件として定性的に表現することを可能にした [11, 12]．また，それらの制約条件を満たすフィルタリング関数の数学的性質を明らかにしたことで，さまざまなフィルタリングの特性を明確にした [13, 14, 15]．これまでに構築したフィルタリングの数学的基盤を利用することで，各フィルタリングの性質から，環境に応じてより効率的な処理方法を実現できる．

さらに筆者らは，フィルタリングに対するユーザの要求を記述できる**フィルタリング SQL** を提案した [10]．フィルタリング SQL とは，データベースへの問合せ

言語である SQL(Structured Query Language) をフィルタリングのために拡張した言語である．フィルタリング SQL により，ユーザはコンテンツの嗜好やデータ管理ポリシーなどを自由に表現できる．しかし，文献 [10] は，フィルタリング SQL の記述構文と，それらの構文の用途を示したのみであるため，ある受信データ集合に対して実際にどのようなフィルタリング結果が得られるかは明確でなかった．そこで本稿では，フィルタリング SQL の基本的な構文について，その意味を明確にすることを目的とする．本稿では，フィルタリング SQL 記述の意味をフィルタリング関数で表現することにより，ユーザの要求を実行するフィルタリングの性質を明確にする．また，これまで筆者らが構築してきた数学機基盤を適用することで，そのフィルタリングが満たす性質から，効率的な処理方法を明らかにする．

以下，第 2 章でフィルタリング関数の概要を述べる．第 3 章では，フィルタリング SQL の基本構文についてその意味を明確にし，それを実行するフィルタリングの性質を明らかにする．第 4 章では，本稿で明らかになった結果をもとに，フィルタリング SQL 記述を実行するフィルタリングの処理方法について考察し，関連研究との比較を行う．最後に第 5 章でまとめを行う．

2 フィルタリング関数

本章では，フィルタリング関数とその基本的な性質について述べる [11, 12, 13].

2.1 フィルタリング処理の分類

あるフィルタリング手法が与えられたとき，実際の処理方法は以下に示すいくつかのパターンに分類できる．

データアイテムを受信する度に，新たな受信データと前回のフィルタリング結果を合せてフィルタリングする処理方法を**逐次処理**とよぶ．逐次処理では，一度蓄積されたデータも，データ受信時に再度フィルタリングする．それに対し，放送データを受信側にためておいてから一括してフィルタリングする処理方法を**一括処理**とよぶ．また，データ集合を 2 つ以上の任意の集合に分割して各々フィルタリングし，結果をマージしたものをフィルタリング結果とする処理方法を**分配処理**とよぶ．さらに，分配処理の結果を再びフィルタリングする処理方法を**並列処理**とよぶ．

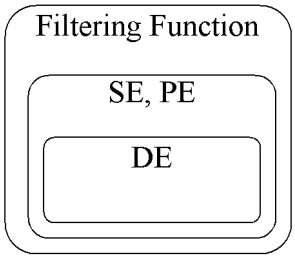

図 1: 等価性間の関係

2.2 フィルタリング関数の性質

データアイテムの集合を $\mathbf{T}$ とする．フィルタリング関数とは，任意の $T \subset \mathbf{T}$ に対し[1]，以下の 2 つの条件を満たす $2^{\mathbf{T}}$ 上の関数 f のことをいう [11, 12].

減少性 (D: Decreasing)
$f(T) \subset T$
ベキ等性 (ID: Idempotent)
$f(f(T)) = f(T)$

減少性 D は，関数を適用した結果が元のデータ集合に含まれるデータアイテムのみであることを表す．ベキ等性 ID は，一度関数を適用すると，何度その関数を適用しても結果が変化しないことを表す．また，フィルタリング関数について以下のような性質が定義されている．

逐次等価性 (SE: Sequential Equivalence)
$f(S \cup T) = f(S \cup f(T))$
分配等価性 (DE: Distributed Equivalence)
$f(S \cup T) = f(S) \cup f(T)$
並列等価性 (PE: Parallel Equivalence)
$f(S \cup T) = f(f(S) \cup f(T))$

ここで，S, T は $\mathbf{T}$ の任意の部分集合とする．逐次等価性は一括処理と逐次処理の結果が等価であることを意味する．同様に，分配等価性は一括処理と分配処理の結果が等価であり，並列等価性は一括処理と並列処理の結果が等価であることを意味する．これまでに筆者らは，これらの等価性間に図 1に示すような包含関係があることを明らかにした [11, 12]．図 1より，一括処理と分配処理の結果が等価であるフィルタリングは，逐次処理や並列処理の結果とも等価となることがわかる．また，一括処理と逐次処理の結果が等価であるフィ

[1]本稿では $A \subset B$ は A が B の部分集合である (A = B の場合を含む) ことを意味するものとする．

ルタリングは，並列処理の結果とも等価となり，一括処理と並列処理の結果が等価であるフィルタリングは，逐次処理の結果とも等価となる．図 1 に示す性質間の関係より，ある性質を満たすフィルタリングが他の性質を満たすかどうかが判断でき，環境に応じてより効率的な処理方法に変換できる．

2.3 セレクション関数

文献 [13] において，**セレクション関数**を次のように定義した．

セレクションとは，各データの取捨選択が潜在的に決まっている手法である．例えば，特定のキーワードを含むデータを蓄積するキーワードマッチングや，データの内容から評価値を計算し，評価値が閾値よりも大きい (あるいは小さい) 場合に蓄積する手法などはセレクションである．ある $X \subset \mathbf{T}$ について，X のセレクション関数 B_X とは，任意の $S \subset \mathbf{T}$ に対して

$$B_X(S) = S \cap X$$

と定義される関数である．X をこのセレクション関数の潜在集合と呼び，蓄積条件を満たすデータの集合を意味する．したがって，キーワードマッチングの潜在集合は，特定のキーワードを含むデータの集合であり，閾値を用いたフィルタリングの潜在集合は評価値が閾値よりも大きい (あるいは小さい) データの集合である．セレクション関数は

$$X = B_X(\mathbf{T})$$

を満たす．セレクション関数に関して，以下の定理が成立する [13]．

[定理 1] フィルタリング関数 f がセレクション関数であることと，f が分配等価性 DE を満たすことは同値である． □

3 フィルタリング SQL

本章では，フィルタリングのためのユーザ要求記述言語であるフィルタリング SQL の意味付けを行う．

まず，フィルタリングの基本構文を以下に示す [10]．

```
EXTRACT   < 属性 >
FROM      < リソース >
WHERE     < プリファレンス >
          < 付加条件 >
```

EXTRACT 句では，抽出されたデータのうち蓄積する属性を指定する．FROM 句にデータ放送名などのデータリソースを指定し，WHERE 句にユーザの嗜好を表現するプリファレンス記述をする．また，必要に応じてデータの蓄積に関するオプションとなる条件を付加できる．

一般にデータ放送では，放送すべき全てのデータアイテムを一度に送信するのではなく，スケジュールにしたがっていくつもの部分集合に分けて送信する．したがって，部分集合の分け方，およびそれらを送信するスケジュールに応じて，ユーザが得られるフィルタリング結果は変化する．ユーザが常に記述した通りのフィルタリング結果を得るには，全てのデータアイテムを一度にフィルタリングする必要がある．そこで，あるフィルタリング SQL 記述が，全てのデータアイテムを一括処理することを**宣言的意味論** (Declarative semantics) とよぶ．しかし，過去に受信したデータアイテムと，将来受信するデータアイテムを集めて一度にフィルタリングすることは現実的でない．そこで，あるフィルタリング SQL 記述が，データアイテムを受信する度に逐次処理することを**手続的意味論** (Procedural semantics) とよぶ．

以下，3.1節で宣言的意味論，3.2節で手続的意味論をそれぞれ明確にする．ただし，本稿ではデータアイテムの取捨選択に焦点を絞った議論を行うため，蓄積する属性やデータリソースの記述は省略する．

3.1 宣言的意味論

データアイテムの集合を $\mathbf{T}$，値の集合を V とすると，属性は

$$a : \mathbf{T} \to V$$

を満たす関数として定義する．特に，受信日時 $received : \mathbf{T} \to D$，最終アクセス日時 $accessed : \mathbf{T} \to D$，蓄積期限 $expire : \mathbf{T} \to D$，メタデータにより予め設定された有効期限 $m\_expire : \mathbf{T} \to D$，データサイズ $size : \mathbf{T} \to R$ とする．D, R は，それぞれ日時 $D \subset V$，数値 $R \subset V$ であり，半順序 $<$ を含むものとする．

ここで，属性 $a : \mathbf{T} \to V$，データアイテム $t \in \mathbf{T}$ に対し，$a(t) \in V$ を $t.a$ と記す．また，キーワード k を $k \subset \mathbf{T}$，キーワード全体の集合を $K \subset 2^{\mathbf{T}}$ とする．

3.1.1 プリファレンスの記述

プリファレンスの記述構文を以下に示す．

```
EXTRACT
WHERE
  PREFER   k1 TO k2
  PREFER   k3 TO k4
    ⋮
  LIMIT    l
```

Σ をフィルタリング SQL 記述の全体，F をフィルタリング関数とすると，セマンティクス関数とは，関数 $S:\Sigma \to F$ である．

ここで，$k_1,\ k_2 \in K$ に対して，プリファレンス記述 p が

$$p:\ \text{PREFER } k_1 \text{ TO } k_2$$

のとき，関数 θ を

$$\theta(p) = (k_1, k_2) \in K \times K$$

と定義する．さらに，

$$P:\ \text{PREFER 文の集合}$$

に対して，

$$\theta(P) = \cup_{p \in P}\{\theta(p)\}$$

とおく．$\theta(P)$ にサイクルがないとき，P は**無矛盾**であるとよぶ．

P が無矛盾であるとき，$\theta(P)$ を含む最小の半順序が一意に存在する．これを $\Theta(P)$ と記す．ここで，$R_P \subset \mathbf{T}^2$ を

$$R_P = \{(d_1,d_2) | d_1 \in k_1,\ d_2 \in k_2,\ (k_1,k_2) \in \Theta(P)\} \quad (1)$$

とする．R_P に対して次の補題が成立する．

[補題 1] R_P は擬順序である．

≪ 証明 ≫ 以下の 2 つの条件を満たすことを示す．

i) $\forall d \in \mathbf{T},\ (d,d) \in R_P$

ii) $\forall d_1,\ d_2,\ d_3 \in \mathbf{T},\ (d_1,d_2),\ (d_2,d_3) \in R_P$ ならば $(d_1,d_3) \in R_P$

i) R_P の定義 (1) より，

$$\{(d,d) | d \in k,\ (k,k) \in \Theta(P)\} \subset R_P$$

なので，

$$\forall d \in \mathbf{T},\ (d,d) \in R_P$$

が成立する．

ii) $\Theta(P)$ は半順序なので，

$$\forall k_1,\ k_2,\ k_3 \in K,$$
$$(k_1,k_2),\ (k_2,k_3) \in \Theta(P) \text{ ならば}$$
$$(k_1,k_3) \in \Theta(P) \quad (2)$$

を満たす．また，(1) の定義より，

$$\forall d_1,\ d_2,\ d_3 \in \mathbf{T},$$
$$(d_1,d_2),\ (d_2,d_3) \in R_P \text{ならば}$$
$$\{(k_1,k_2),\ (k_2,k_3) | d_1 \in k_1,\ d_2 \in k_2,\ d_3 \in k_3\}$$
$$\subset \Theta(P) \quad (3)$$

となるので，(1)，(2)，(3) より

$$(d_1,d_3) \in R_P$$

が成立する．

i)，ii) より，R_P は擬順序である． □

次に，$a,\ b \in \mathbf{T}$ に対して

$$a \sim b \overset{def}{\iff} (a,\ b),\ (b,\ a) \in R_P$$

とする $\sim$ は同値関係である．また，$\mathbf{T}_\sim$ を $\mathbf{T}$ の $\sim$ による商集合とすると，$R_{P\sim} \subset \mathbf{T}_\sim{}^2$ は R_P と同様にして自然に決まる．

$$R_{P\sim} = \{(x,y) | x \in k_1,\ y \in k_2,\ (k_1,k_2) \in \Theta(P)\} \quad (4)$$

ここで，

$$x,\ y \in \mathbf{T}_\sim,\ (x,y) \in R_{P\sim}$$
$$\iff \quad (a,\ b) \in R_{P\sim},\ a \in x, b \in y$$

である．$R_{P\sim}$ について次の補題が成立する．

[補題 2] $R_{P\sim}$ は半順序である．

≪ 証明 ≫ 以下の 3 つの条件を満たすことを示す．

i) $\forall x \in \mathbf{T}_\sim,\ (x,x) \in R_{P\sim}$

ii) $\forall x_1,\ x_2 \in \mathbf{T}_\sim,\ (x_1,x_2),\ (x_2,x_1) \in R_{P\sim}$ ならば $x_1 = x_2$

iii) $\forall x_1,\ x_2,\ x_3 \in \mathbf{T}_\sim,\ (x_1,x_2),\ (x_2,x_3) \in R_{P\sim}$ ならば $(x_1,x_3) \in R_{P\sim}$

i) $R_{P\sim}$ の定義 (4) より，

$$\{(x,x) | x \in k,\ (k,k) \in \Theta(P)\} \subset R_{P\sim}$$

なので，

$$\forall x \in \mathbf{T}_\sim,\ (x,x) \in R_{P\sim}$$

が成立する．

ii) $\forall x_1,\ x_2 \in \mathbf{T}_\sim$ に対して $(x_1,x_2),\ (x_2,x_1) \in R_{P\sim}$ ならば，同値関係 $\sim$ の定義より $x_1 = x_2$ である．

iii) $\Theta(P)$ は半順序なので，

$$\forall k_1,\ k_2,\ k_3 \in K,\quad (k_1,k_2),\ (k_2,k_3) \in \Theta(P) \text{ ならば } (k_1,k_3) \in \Theta(P) \tag{5}$$

を満たす．また，(4) の定義より，

$$\forall x_1,\ x_2,\ x_3 \in \mathbf{T}_\sim,\quad (x_1,x_2),\ (x_2,x_3) \in R_{P\sim} \text{ならば } \{(k_1,k_2),\ (k_2,k_3) | x_1 \in k_1,\ x_2 \in k_2,\ x_3 \in k_3\} \subset \Theta(P) \tag{6}$$

となるので，(4)，(5)，(6) より

$$(x_1,x_3) \in R_{P\sim}$$

が成立する．

i)，ii)，iii) より，$R_{P\sim}$ は半順序である． □

同値類 $x \in \mathbf{T}_\sim$ の要素に対して，属性 a の値域 V 上の全順序を

$$D_x^a = \{(d_1,d_2) | d_1,\ d_2 \in x,\ d_1.a > d_2.a\}$$

とする．また，半順序 $Q_P^a \subset \mathbf{T}^2$ を

$$(x,y) \in Q_P^a \Longleftrightarrow \begin{cases} [x]_\sim \neq [y]_\sim \text{ならば } ([x]_\sim,\ [y]_\sim) \in R_{P\sim} \\ [x]_\sim = [y]_\sim \text{ならば } (x,y) \in D_x^a \end{cases}$$

と定義する．

ここで，Q_P^a に対し，ある $y \in \mathbf{T},\ Y \subset \mathbf{T}$ について

$$\cup(y) = \{z | (z,y) \in Q_P^a\}$$
$$\cup(Y) = \cup_{y \in Y} \cup (y)$$

とおく．

ある $x \subset \mathbf{T}$ に対して，

$$X = \cup(x)$$

のとき，X を**σ閉集合**とよぶ．また，X が σ 閉集合かつ

$$\sum_{y \in X} y.size < l \tag{7}$$

を満たすことを $\Box_\sigma X$ と記す．ここで，$S^a(\sigma) : \Sigma \to 2^F$ を

$$S^a(\sigma) = \{X | \Box_\sigma X\} \tag{8}$$

と定義し，これを**aセマンティクス**とよぶ．$\forall S \subset \mathbf{T}$ に対して，$f \in F$ が

$$f(S) = X \cap S$$

とおけるとき，f は**σ適合**であるとよぶ．f がフィルタリング関数であることは，容易に確かめられる．

ここで，$F_\sigma^a \subset F^2$ を

$$F_\sigma^a = \{(f_1,f_2) | \forall x \subset \mathbf{T},\ f_1(x) \subset f_2(x),\ f_1,\ f_2 \text{は}\sigma\text{適合である}\} \tag{9}$$

と定義する．F_σ^a について，次の補題が成立する．

[補題 3] F_σ^a は半順序である．

≪ 証明 ≫ 以下の 3 つの条件を満たすことを示す．

i) $\forall f \in F,\ (f,f) \in F_\sigma^a$

ii) $\forall f_1,\ f_2 \in F,\ (f_1,f_2),\ (f_2,f_1) \in F_\sigma^a$ ならば $f_1 = f_2$

iii) $\forall f_1,\ f_2,\ f_3 \in F,\ (f_1,f_2),\ (f_2,f_3) \in F_\sigma^a$ ならば $(f_1,f_3) \in F_\sigma^a$

i) $\forall f \in F$ に対して，

$$\forall x \subset \mathbf{T},\ f(x) \subset f(x)$$

なので，$(f,f) \in F_\sigma^a$ である．

ii) $\forall f_1,\ f_2 \in F$ に対して，$(f_1,f_2),\ (f_2,f_1) \in F_\sigma^a$ のとき，$f_1 \neq f_2$ と仮定すると，

$$\forall x \subset \mathbf{T},\ f_1(x) \subset f_2(x) \text{ かつ } f_2(x) \subset f_1(x)$$

となり矛盾している．したがって，$f_1 = f_2$ が成立する．

iii) $\forall f_1,\ f_2,\ f_3 \in F$ に対して，$(f_1,f_2),\ (f_2,f_3) \in F_\sigma^a$ ならば，

$$\forall x \subset \mathbf{T},\ f_1(x) \subset f_2(x) \text{ かつ } f_2(x) \subset f_3(x)$$

となる．これより $f_1(x) \subset f_3(x)$ が導き出されるので $(f_1,f_3) \in F_\sigma^a$ となる．

i)，ii)，iii) より，F_σ^a は半順序である． □

特に，

$$\forall x \subset \mathbf{T},\ f^0(x) = \phi$$

を満たす f^0 は (7) を必ず満たすため，常に σ 適合である．また，F_σ^a に極大が存在することは容易に確かめられる．f^m が F_σ^a の極大のとき，**σ極大適合**であるとよぶ．さらに，(7) の左辺を最大とする f^M を**σ最大適合**であるとよぶ．ここで，σ 最大適合である f^M は一意とは限らない点に注意が必要である．

σ 極大適合である f^m について以下の補題を示す．

[補題 4] σ 極大適合である f^m は，分配等価性 DE を必ずしも満たさない．

≪証明≫ データアイテム集合 $\mathbf{T} = \{d_1,\ d_2,\ d_3,\ d_4\}$ に対して，$d_1.size = 9,\ d_2.size = 2,\ d_3.size = 5,\ d_4.size = 4$，蓄積容量制限 $l = 10$ とする．また，$(d_1,d_2),\ (d_2,d_3),\ (d_3,d_4) \in Q_P^a$ であるとする．

このとき，$S = \{d_1,\ d_2\},\ T = \{d_3,\ d_4\}$ とすると，

$$\begin{aligned} f^m(S) &= \{d_1\} \\ f^m(T) &= \{d_3,\ d_4\} \\ f^m(S) \cup f^m(T) &= \{d_1,\ d_3,\ d_4\} \end{aligned}$$

となる．したがって，分配等価性 $f^m(S \cup T) = f^m(S) \cup f^m(T)$ を必ずしも満たさない． □

[補題 5] $\forall S,\ T$ に対して，$S \subset T$ のとき $X_T \subset X_S$ ならば，σ 最大適合である f^M は逐次等価性 SE を満たす．ただし，σ 閉集合 $X_S,\ X_T$ は $f^M(S) = X_S \cap S,\ f^M(T) = X_T \cap T$ を満たすものとする．

≪証明≫ f^M が $\forall S,\ T \subset \mathbf{T}$ に対して

$$f^M(S \cup T) = f^M(S \cup f^M(T))$$

を満たすことを示せばよい．

ここで，$X_{ST},\ X_S,\ M_T,\ M_{SfT} \in \mathbf{T}$ を

$$\begin{aligned} f^M(S \cup T) &= X_{ST} \cap (S \cup T) \\ f^M(S) &= X_S \cap S \\ f^M(T) &= X_T \cap T \\ f^M(S \cup f^M(T)) &= X_{SfT} \cap (S \cup f^M(T)) \end{aligned}$$

を満たす σ 閉集合とする．

f^M の減少性より $S \cup f^M(T) \subset S \cup T$ であるため，

$$X_{ST} \subset X_{SfT} \tag{10}$$

となる．同様に，$T \subset S \cup T$ より

$$X_{ST} \subset X_T \tag{11}$$

である．

また，f^M のベキ等性より，$f^M(T) = f^M(f^M(T))$ である．したがって，X_T は

$$f^M((T)) = f^M(T) = X_T \cap T$$

を満たす．したがって，$f^M(T) \subset S \cup f^M(T)$ なので，

$$X_{SfT} \subset X_T \tag{12}$$

が成り立つ．ここで，(10)，(12) より

$$\begin{aligned} & X_{SfT} \cap (S \cup f^M(T)) \\ &= X_{SfT} \cap (S \cup (X_T \cap T)) \\ &= (X_{SfT} \cap S) \\ &\quad \cup (X_{SfT} \cap X_T \cap T) \\ &= (X_{SfT} \cap S) \\ &\quad \cup (X_{SfT} \cap T) \quad (\because (12)) \\ &= X_{SfT} \cap (S \cup T) \\ &\supset X_{ST} \cap (S \cup T) \quad (\because (10)) \end{aligned}$$

となり，

$$f^M(S \cup T) \subset f^M(S \cup f^M(T)) \tag{13}$$

が導き出される．ここで，f^M は σ 最大適合であるので，X_{ST} は

$$\sum_{x \in X_{ST} \cap (S \cup T)} x.size < l$$

を満たす最大のデータ集合である．したがって，

$$f^M(S \cup T) \neq f^M(S \cup f^M(T))$$

と仮定すると，(13) より，

$$\sum_{x \in X_{SfT} \cap (S \cup f^M(T))} x.size > l$$

となり，(7) に矛盾する．ゆえに，

$$f^M(S \cup T) = f^M(S \cup f^M(T))$$

が成立する． □

3.1.2 セレクションの記述

フィルタリング SQL 記述 σ_{s_1} が

```
EXTRACT
WHERE      attribute = value
```

であるとき，σ_{s_1} の意味論は

$$S(\sigma_{s_1})(X) \triangleq \{x \in X | x.attribute = value\}$$

とする．また，フィルタリング SQL 記述 σ_{s_2} が

```
EXTRACT
WHERE      PERIOD value
```

であるとき，σ_{s_2} の意味論は，t を現在の日時とすると

$$S(\sigma_{s_2})(X) \triangleq \{x \in X | x.received + value > t\}$$

とする．ここで，次の補題が成立する．

[補題 6] $S(\sigma_{s_1})$, $S(\sigma_{s_2})$ は分配等価性 DE，逐次等価性 SE，並列等価性 PE を満たす．

≪証明≫ $A, B \subset \mathbf{T}$ を

$$\begin{aligned} A &= \{x | x.attribute = value\} \\ B &= \{x | x.received + value > t\} \end{aligned}$$

とおくと，$S(\sigma_{s_1})$, $S(\sigma_{s_2})$ は

$$\begin{aligned} &\exists A, \forall X \subset \mathbf{T}, \quad S(\sigma_{s_1})(X) = X \cap A \\ &\exists B, \forall X \subset \mathbf{T}, \quad S(\sigma_{s_2})(X) = X \cap B \end{aligned}$$

を満たすため，セレクション関数である．したがって，定理 1，図 1より，$S(\sigma_{s_1})$, $S(\sigma_{s_2})$ は分配等価性 DE，逐次等価性 SE，並列等価性 PE を満たす． □

3.2 手続的意味論

前節までは，宣言的意味論について述べてきたが，実際のフィルタリング処理では，逐次処理を行うのが一般的かつ現実的である．しかし，フィルタリング SQL 記述の要求を逐次処理で実行する，すなわち手続的意味論で解釈できるようにするには，宣言的意味論と手続的意味論が等しくなければならない．両者が等しくなるためには，

$$\text{逐次等価性 } f(S \cup T) = f(S \cup f(T))$$

を満たす必要がある．3.1.1節に示したフィルタリング SQL 記述では，補題 5より，$\forall S, T$ に対して $S \subset T$ のとき $X_T \subset X_S$ であるという条件を満たすならば，σ 最大適合であるフィルタリング関数 f^M が逐次等価性 SE を満たすことが示された．また，3.1.2節に示したフィルタリング SQL 記述では，補題 6より，必ず逐次等価性 SE を満たすことが示された．したがって，これらのフィルタリングでは，手続的意味論と宣言的意味論が一致する．

4 考察

本章では，本稿で取り扱った記述に対して，それを実行するフィルタリングの性質から，可能な処理方法について論じる．また，フィルタリング言語を用いた関連研究について述べる．

4.1 フィルタリング SQL 記述の実現

3.1.1節に示したプリファレンス記述は，受信データ集合 S, T が $S \subset T$ のとき，σ 閉集合 X_S, X_T が $X_T \subset X_S$ であるという条件を満たし，かつ σ 最大適合であるフィルタリング関数 f^M で実行すれば，逐次処理でフィルタリングを行っても宣言的意味論と一致することがわかった．さらに，図 1より，逐次等価性と並列等価性が同値であることから，並列処理でフィルタリングを実行しても，宣言的意味論と等価な結果が得られる．

また，3.1.2節に示したセレクション記述は，常に逐次等価性を満たすことから，逐次処理でフィルタリングを行っても宣言的意味論と一致することがわかった．さらに，図 1より，分配等価性と並列等価性も満たすことから，分配処理や並列処理でフィルタリングを実行しても，宣言的意味論と等価な結果が得られる．したがって，これらのフィルタリング SQL 記述に対して，受信機の処理能力が低いため継続的に受信し続けるデータアイテムを逐一処理できないときは，分配処理や並列処理で実行しても宣言的意味論が保たれることが保証できる．

4.2 関連研究

XML 文書のフィルタリングに XFilter[1] がある．X-Filter では，XPath[4] によってユーザの要求を記述し，XML 文書の構造を利用してセレクション処理を行う．XPath の記述例を以下に示す．

```
//Animal/name=Penguin
```

上の記述例は，「“Animal” というノードの子ノードの属性 “name” が “Penguin” であるデータアイテムを抽出する」というユーザの要求を表す．フィルタリング SQL がデータアイテムの属性 (メタデータの内容) を利用することに対し，XPath はコンテンツの構造を利用することでフィルタリングする．そのため，より精度の高いフィルタリングが可能であるが，ユーザは放送されるデータの構造をより明確に理解しておく必要がある．

また，同じ嗜好をもつユーザの問合せを利用することで協調フィルタリングを行うフィルタリングに Tapestry[5] がある．Tapestry では，フィルタリング SQL と同様に，SQL をフィルタリングのために拡張した TQL を用いてユーザの要求を記述する．TQL の記述例を以下に示す．

m IN Mary.Animal

この記述例は,「“Mary” の “Animal” に関する問合せで抽出されたデータを蓄積する」というユーザの要求を表す．現在のところ，フィルタリング SQL では協調フィルタリングを利用した記述はできない．そこで，今後協調フィルタリングを考慮に入れた構文を定義することで，より柔軟な嗜好の記述が可能となる．ただし，協調フィルタリングには，嗜好の類似したユーザを発見することが困難であるという欠点があるため，その問題を解決する必要がある．

5 おわりに

本稿では，フィルタリングのためのユーザ要求記述言語であるフィルタリング SQL の意味を明確にした．言語の意味をフィルタリング関数で表現したことにより，ユーザの要求を実行するフィルタリングの性質を明らかにした．また，その性質から，環境に応じた効率的な処理方法が実現できることを述べた．

今後の課題として，複数の構文を組合せたときの意味を決定することが挙げられる．本稿では，フィルタリング SQL のプリファレンス記述，およびいくつかの記述構文に対し，単一の構文を用いた記述の意味を明確にした．しかし，ユーザが実際にフィルタリングの要求を記述するには，複数の構文を組合せて用いる必要がある．そこで，そのように複雑な要求記述の意味も明確にし，どのような構文を組合せたときに，どのような意味を表すかについて明確にする予定である．

謝辞

本研究は，文部科学省振興調整費「情報フィルタリングの数学的基盤の確立」,「モバイル環境向 P2P 型情報共有基盤の確立」，および文部科学省 21 世紀 COE プログラム (研究拠点形成費補助金)，科学研究費補助金 (基盤研究 (B)(2))「大規模な仮想空間システムを構築する放送型サイバースペースに関する研究」(プロジェクト番号:15300033) の研究助成によるものである．ここに記して謝意を表す．

参考文献

[1] M. Altinel and M. J. Franklin: “Efficient filtering of XML documents for selective dissemination of information,” in *Proc. 26th International Conference on Very Large Data Bases (VLDB2000)*, pp. 53–64 (2000).

[2] N. J. Belkin and W. B. Croft: “Information filtering and information retrieval: two sides of the same coin?,” *Communications of the ACM*, vol. 35, no. 12, pp. 29–38 (1992).

[3] T. A. H. Bell and A. Moffat: “The design of a high performance information filtering system,” in *Proc. 19th International Conference on Research and Development in Information Retrieval (SIGIR1996)*, pp. 12–20 (1996).

[4] J. Clark and S. DeRose: “XML path language (XPath) version 1.0,” W3C Recommendation, http://www.w3.org/TR/xpath (1999).

[5] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry: “Using collaborative filtering to weave an information TAPESTRY,” *Communications of the ACM*, vol. 35, no. 12, pp. 61–70 (1992).

[6] 森田昌宏: “情報フィルタリングに関する研究動向,” JAIST Research Report, IS-RR-93-9I, 北陸先端科学技術大学院大学情報科学研究科 (1993).

[7] 西正, 野村敦子: “多チャンネル放送の衝撃,” 中央経済社 (1997).

[8] Satellite Magazine: http://www.satemaga.co.jp.

[9] 澤井里枝: ミニサーベイ “ブロードバンド時代における情報フィルタリングの動向,” 電子情報通信学会第 13 回データ工学ワークショップ (DEWS2002) 論文集 (2002).

[10] 澤井里枝, 寺田努, 塚本昌彦, 西尾章治郎: “フィルタリング SQL: フィルタリングのためのユーザ要求記述言語,” 電子情報通信学会第 11 回データ工学ワークショップ (DEWS2000) 論文集 (2000).

[11] R. Sawai, M. Tsukamoto, Y. H. Loh, T. Terada, and S. Nishio: “Functional properties of information filtering,” in *Proc. 27th International Conference on Very Large Data Bases (VLDB2001)*, pp. 511–520 (2001).

[12] 澤井里枝, 塚本昌彦, 寺田努, Loh Yin Huei, 西尾章治郎: “情報フィルタリングの関数的性質について,” 電子情報通信学会論文誌 D-I, vol. J85-D-I, no. 10, pp. 939–950 (2002).

[13] 澤井里枝, 塚本昌彦, 寺田努, 西尾章治郎: “フィルタリング関数におけるセレクションとランキングについて,” 情報処理学会論文誌:データベース, vol.43, no. SIG12(TOD16), pp. 80–91 (2002).

[14] 澤井里枝, 塚本昌彦, 寺田努, 西尾章治郎: “合成フィルタリング関数の性質について,” 情報処理学会論文誌: データベース, vol. 44, no. SIG3(TOD17), pp. 43–53 (2003).

[15] 澤井里枝, 塚本昌彦, 寺田努, 西尾章治郎: “情報フィルタリングの実行順序に関する関数的性質について,” 情報処理学会論文誌:データベース, vol. 44, no. SIG3(TOD17), pp. 54–64 (2003).