データベースシステム 126－20
（２００２．１．２２）

異なる構造を持つXMLデータを統合するための構造間の関係について

中川 剛　石原 靖哲　藤原 融

大阪大学大学院基礎工学研究科情報数理系

あらまし

XMLはWebアプリケーション等のデータフォーマットとして広く使用され始めている．XMLではユーザが自由にデータ構造を定義できるため，構造が異なる複数のXMLデータを統合して扱うには，それらのデータの構造間の関係を表現できる形式的枠組が必要である，これまでに包摂(subsumption)と呼ばれる関係に基づいた枠組が提案されており，その枠組のもとで構造の異なる複数のXMLデータを統合して効率的に扱えることが知られているが，包摂関係は成立する条件が厳しいため扱えるXMLデータが大きく制限されるという問題がある．そこで従来の包摂関係よりも成立条件の緩い関係を提案し，その性質について考察する．

A Relation on Schemas for Integrating XML Data with Different Schemas

Go Nakagawa　Yasunori Ishihara　Toru Fujiwara

Department of Informatics and Mathematical Science
Graduate School of Engineering Science
Osaka University

Abstract

XML is a promising data format for various applications, especially Web applications. Since XML allows users to define schemas at will, the same data can be represented as XML data with various schemas. In order to manipulate these data, it is important to understand the relationship between these schemas. To capture such relationships, a notion of subsumption for XML has been proposed so far. Using subsumption, some XML data with different schemas can be integrated efficiently. However, subsumption is too strong, that is, subsumption is satisfied by only a few pairs of schemas. In this paper, we propose a relation which is weaker than subsumption, and discuss its properties.

1 まえがき

近年, XML は Web アプリケーション等のデータフォーマットとして広く使用され始めている. XML ではデータをテキスト形式で記述するため, OS やアプリケーションに依らずに処理することができる. また, 関係データベースのように決められた形式のスキーマを用いる必要がないため, データ交換が簡単で semistructured data [1] の処理にも適している. しかし, 自由にスキーマを定義できることによる欠点もある. 例えば, 複数の XML 文書を処理する際に, 同じ種類のデータでもスキーマが異なる場合がある. 問い合わせの構造はスキーマに依存するため, 同じ種類のデータでも別々に処理する必要がある. これらのデータを統合して扱うには, すべてのスキーマを統一しなければならない. このとき, スキーマの表す意味上の包含関係を表す形式的枠組みがあれば, その枠組みのもとですべてのスキーマを意味的に含むスキーマを与えてやればよい.

これまでに包摂 (subsumption) と呼ばれる関係に基づいた枠組が提案されており [3], その枠組のもとで構造の異なる複数の XML データを統合して効率的に扱えることが知られている. しかし, 包摂関係は成立する条件が厳しいため扱える XML データが大きく制限されるという問題がある. 本稿では, 従来の包摂関係よりも成立条件の緩い関係を提案し, その性質を調べる. また, 統合したスキーマに対する問い合わせを統合する前のスキーマに対する問い合わせに変換する方法について考察する.

2 節では, XML 文書のデータモデルとスキーマ, 型割り当て, 問い合わせについて説明する. 3 節で文献 [3] による包摂関係について述べ, 4 節で包摂を拡張した関係を提案する. 5 節では, 問い合わせの自動変換について述べる. 6 節でまとめと今後の課題を述べる.

2 データモデルとスキーマ

文献 [2, 3, 4] に基づいてデータモデルとスキーマを定義する.

2.1 データモデル

参照を含まない XML データはラベルつき順序木で自然にモデル化される. すなわち, 各頂点は対となるタグで囲まれた要素 (オブジェクトと呼ぶ) を表しており, タグ名もしくは値に対応するラベルがつけられている. 頂点 o が o' の子であることは, XML データにおいて o が o' の直接の子要素であることを表す. 根は "Δ" で表され, ラベルを持たない.

本稿では, 文献 [3] に従い, 参照を次のように表現する. オブジェクト o が属性値として o' を参照しているとする. このとき, 新たな頂点 o'' を用意する. o'' のラベルを特別な記号 "&" とし, o から o', および, o'' から o' へ有向辺を引く. なお, 根頂点 Δ はどの頂点からも参照されない.

形式的には, XML データは以下のように定義される.

定義 1: $\mathcal{O}$ をオブジェクト ID の集合, $\mathcal{L}$ をラベルの集合とする. データベースは $D = \langle O_D, label_D, child_D \rangle$ で表される. ここで,

1. $O_D \subseteq \mathcal{O}$.
2. $label_D$ は O_D から $\mathcal{L}$ への写像.
3. $child_D$ は $O_D \cup \{\Delta\}$ から O_D 上の系列への写像. ただし, $label_D(o) = \&$ なら $child_D(o) \in O_D$.
4. $label_D(o) = \&$ である o の $child_D$ を無視して得られる構造は木である.

例 1: 本稿では次のようなシナリオに基づいて例を示す.

デジタルカメラをオンラインショッピングで購入するために様々なサイトにアクセスし, いくつかの製品カタログを手に入れる. 図 1 の左側のデータは CNN 社の販売店のサイトから手に入れたデータで, 右側は他のサイトから得たデータである.

CNN 社の販売店は CNN 社の製品に関する情報を図 2 で与えられるスキーマに沿って記述している. このスキーマは, CNN 製のデジタルカメラの構成要素を表している. デジタルカメラ要

```
<製品>
  <デジタルカメラ>
    <メーカー>CNN</メーカー>
    <商品名>IXD</商品名>
    <価格>
      <定価>72000</定価>
      <売値>54800</売値>
    </価格>
    <画素数>2110000</画素数>
  </デジタルカメラ>

  <デジタルカメラ>
    <メーカー>CNN</メーカー>
    <商品名>IXE</商品名>
    <価格>
      <定価>52000</定価>
      <売値>39800</売値>
    </価格>
    <ズーム>3 倍</ズーム>
  </デジタルカメラ>

...
```

```
<製品>
  <デジタルカメラ>
    <メーカー>FF</メーカー>
    <商品名>Fp5</商品名>
    <価格>
      <定価>88000</定価>
      <売値>59800</売値>
    </価格>
    <画素数>2110000</画素数>
    <保証期間>1 年</保証期間>
  </デジタルカメラ>

  <デジタルカメラ>
    <商品名>SS4</商品名>
    <価格>
      <定価>58000</定価>
      <売値>44800</売値>
    </価格>
    <画素数>1400000</画素数>
  </デジタルカメラ>

...
```

図 1: XML 文書の例

素はメーカー要素 (値は常に CNN), 商品名要素等を持っている. また, 画素数要素, ズーム要素を持っている場合もある (? は 0 か 1 個, また * は 0 個以上を表す).

図 3 は, CNN 社のデジタルカメラのデータの一部を表しており, 以下のような構造に対応する.

$$
\begin{aligned}
D &= \langle O_D, label_D, child_D \rangle \\
O_D &= \{o_1, o_{11}, o_{12}, \ldots\} \\
child_D(\Delta) &= \cdots o_1 \cdots \\
label_D(o_1) &= \text{デジタルカメラ} \\
label_D(o_{11}) &= \text{メーカー} \\
label_D(o_{111}) &= \text{“}CNN\text{”} \\
child_D(o_1) &= o_{11} o_{12} o_{13} o_{14} \\
child_D(o_{11}) &= o_{111} \\
child_D(o_{111}) &= \epsilon \\
&\cdots
\end{aligned}
$$

CNN スキーマは他のサイトで集められたデータはサポートできない. メーカー要素がない, 他の要素がある等, 集めた全てのデータをサポートするためには, より一般的なスキーマ (例えば図 4) が必要となる.

2.2 スキーマ

XML のスキーマは, 記述するデータの取り得る構造を表す. $\mathcal{T}$ を型名の集合とする. 以下, τ, τ' 等は $\mathcal{T}$ の要素を表す.

定義 2: スキーマは $S = \langle T_S, pred_S, re_S \rangle$ で表される. ここで,

1. $T_S = \{\tau_1, \ldots \tau_n\}$ は $\mathcal{T}$ の有限部分集合.
2. $pred_S$ は T_S から ラベル集合 $\mathcal{L}$ のべき集合への写像. ただし, 任意の $\tau \in T_S$ について, $pred_S(\tau) = \{\&\}$ あるいは $\& \notin pred_S(\tau)$ のどちらか一方が成立する.
3. re_S は $T_S \cup \{\Delta\}$ から T_S 上の正規表現への写像. ただし, $pred_S(\tau) = \{\&\}$ ならば, $re_S(\tau)$ は $\tau_1 \mid \ldots \mid \tau_n$ になる.

例 2: 図 2 のスキーマは, 以下の構造に対応する.

$$S = \langle T_{cat}, pred_{cat}, re_{cat} \rangle$$

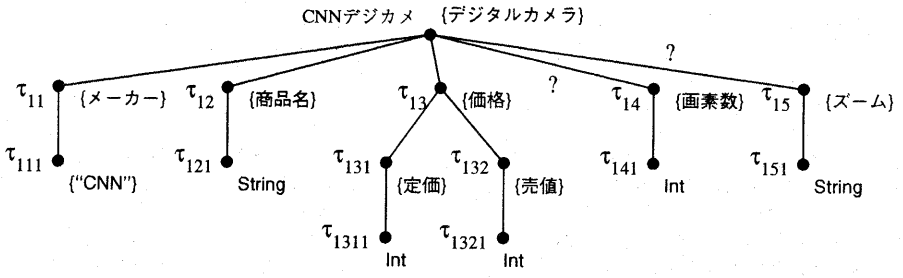

図 2: CNN 製デジタルカメラのスキーマ

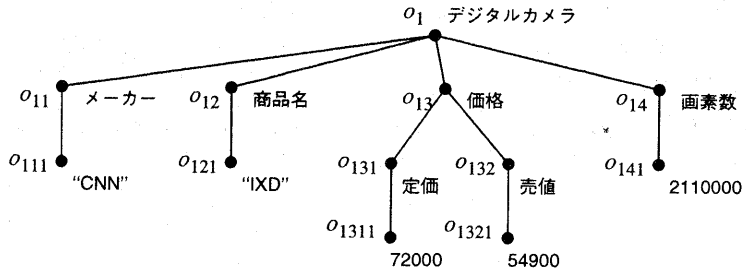

図 3: CNN 製デジタルカメラのデータベース

$T_{cat} = \{catalog, CNN\ デジカメ, \tau_{11}, \tau_{12}, ...\}$

$re_{cat}\{\Delta\} = catalog$

$pred_{cat}(CNN\ デジカメ) = \{\ デジタルカメラ\ \}$

$pred_{cat}(\tau_{131}) = \{\ 定価\ \}$

$pred_{cat}(\tau_{1311}) = \{0, 1, ...\}$

$re_{cat}(CNN\ デジカメ) = \tau_{11}\tau_{12}\tau_{13}\tau_{14}^{?}\tau_{15}^{?}$

$re_{cat}(\tau_{131}) = \tau_{1311}$

$re_{cat}(\tau_{1311}) = \epsilon$

...

2.3 型割り当て

正規表現 r が表す言語を $L(r)$ と表記する.

定義 3: D をデータベース, S をスキーマとする. $O_D \cup \{\Delta\}$ から $T_S \cup \{\Delta\}$ への写像 ρ が以下の条件を満たすとき, D は型割り当て ρ のもとで型 S を持つという.

1. $\rho(\Delta) = \Delta$.
2. 任意の $o \in O_D$ に対し, $label_D(o) \in pred_S(\rho(o))$.
3. 任意の $o \in O_D \cup \{\Delta\}$ に対し, $child_D(o) = o_1 \cdots o_n$ ならば, $\rho(o_1) \cdots \rho(o_n) \in L(re_S(\rho(o)))$.

例 3: 図 2, 3 において, デジタルカメラ文書と CNN スキーマの間の型割り当て ρ は次のようになる.

$$\begin{aligned}
\rho(o_1) &= CNN\ デジカメ \\
\rho(o_{11}) &= \tau_{11} \\
\rho(o_{13}) &= \tau_{13} \\
\rho(o_{132}) &= \tau_{132} \\
\rho(o_{14}) &= \tau_{14} \\
\rho(o_{141}) &= \tau_{141} \\
&\cdots
\end{aligned}$$

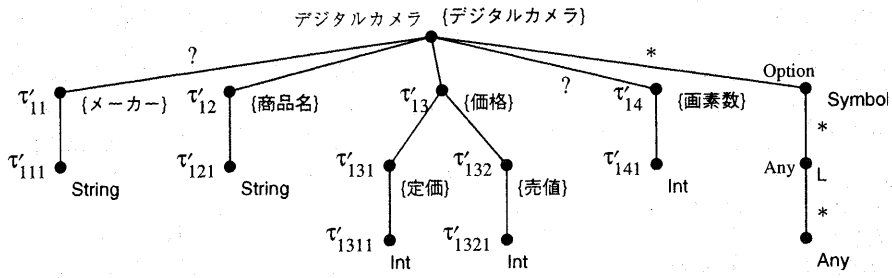

図 4: 一般のデジタルカメラのスキーマ

2.4 問い合わせ

本稿では，データベースへの問い合わせは XPath [5] に基づいた以下のような式で行うものとする．

定義 4: スキーマ S におけるパス式は，以下のように定義される．

1. "/" はパス式．
2. p がパス式のとき，p/τ と $p//\tau$ はパス式．ただし，$\tau \in T_S$．
3. パス式は上のいずれかのみ．

データベース D におけるパス式 p の意味 $E_D(p)$ は，以下で定義される．

- $E_D(/) = \{\Delta\}$．
- $E_D(p/\tau) = \{o \mid o$ は $E_D(p)$ 中のある頂点の子で，かつ，$label_D(o) \in pred_S(\tau)\}$．
- $E_D(p//\tau) = \{o \mid o$ は $E_D(p)$ 中のある頂点の子孫 ($E_D(p)$ 中の頂点自身も含む) で，かつ，$label_D(o) \in pred_S(\tau)\}$．

3 包摂関係

本節では，文献 [3] で提案された包摂関係を説明する．S, S' をスキーマとすると，S が S' を包摂するとは，直感的には，型 S を持つデータの集合が型 S' を持つデータの集合に含まれることである．

定義 5: S, S' をスキーマとする．$T_S \cup \{\Delta\}$ から $T_{S'} \cup \{\Delta\}$ への写像 θ が以下の条件を満たすとき，S は包摂写像 θ のもとで S' を包摂するといい，$S \preceq_\theta S'$ と書く．

1. $\theta(\tau) = \Delta \Leftrightarrow \tau = \Delta$.
2. 任意の $\tau \in T_S$ に対して，$pred_S(\tau) \subseteq pred_{S'}(\theta(\tau))$.
3. 任意の $\tau \in T_S \cup \{\Delta\}$ に対して，$\theta(L(re_S(\tau))) \subseteq L(re_{S'}(\theta(\tau)))$.

$S \preceq_\theta S'$ となるような θ が存在するとき，$S \preceq S'$ と書く．

例 4: 図 2 の CNN スキーマは図 4 のデジカメスキーマを包摂している．このとき，包摂写像 θ' は次のようになる．

$$\begin{aligned}\theta'(CNN\ デジカメ) &= デジタルカメラ\\ \theta'(\tau_{11}) &= \tau'_{11}\\ \theta'(\tau_{111}) &= \tau'_{111}\\ \theta'(\tau_{13}) &= \tau'_{13}\\ \theta'(\tau_{15}) &= Option\\ \theta'(\tau_{151}) &= Any\\ &\cdots\end{aligned}$$

$S \preceq S'$ が成立するとき，型 S を持つ任意のデータベース D は，ある型割り当て ρ のもとで型 S' も持つことが [3] で示されている．よって，複数のスキーマ $S_1, S_2, \ldots$ に対して，$S_1, S_2, \cdots \preceq S'$ となるような S' を見つければ，型 $S_1, S_2, \ldots$ の

```
<デジタルカメラ>
  <商品名>Fp5</商品名>
  <定価>88000</定価>
  <売値>59800</売値>
</デジタルカメラ>                    (a)

<デジタルカメラ>
  <商品名>Fp5</商品名>
  <価格>
    <定価>88000</定価>
    <売値>59800</売値>
  </価格>
</デジタルカメラ>                    (b)
```

図 5: XML 文書の例 2

いずれかを持つ任意のデータベースは型 S' を持つデータとして統合できる.

しかし, 包摂関係は成立条件が厳しいため, 統合するデータのスキーマで共通点が少ない場合, 統合したスキーマが非常に複雑になってしまう. スキーマが複雑になると, そのスキーマに対する問い合わせも複雑になる. 次節では, より単純なスキーマで統合できるようにするために, 包摂関係よりも成立条件のゆるい関係を考える.

4 包摂関係の拡張

図 5 は, "Fp5" というデジタルカメラのデータを表す. データ (b) の価格要素は, その子要素である定価要素と売値要素をまとめる役割を持っているが, 値をもつ要素ではないので, (a) のように価格要素がないデータでも (b) とほぼ同じ意味を持つと考えられる. このように, XML のスキーマには, あってもなくてもほぼ同じ意味を表す型が存在する場合がある.

このことを考慮して, 包摂よりも成立条件の緩い関係として弱包摂関係を考える. 包摂関係では, 親子の関係を持つ頂点は写像先においても親子でなければならないが, 弱包摂関係では写像先で先祖子孫の関係が成り立っていればよいというように定義する.

まず, スキーマ S に関して拡張した正規表現の集合 $\mathcal{R}_S$ を次のように定義する.

定義 6: $\mathcal{R}_S$ を次のように定義する.

1. T_S上の任意の正規表現は $\mathcal{R}_S$に属する.
2. 任意の $R \in \mathcal{R}_S$と R に現れる任意の$\tau \in T_S$ に対し, R 中の 1 箇所のτを $[R']_\tau$(ただし $L(R') = L(re_S(\tau))$) で置き換えて得られる表現は $\mathcal{R}_S$ に属する.
3. $\mathcal{R}_S$ に属する表現は上のいずれかのみ.

例えば, T_S 上の通常の正規表現 $\tau_1^*\tau_2\tau_3\tau_2$ は $\mathcal{R}_S$ に属する. さらに, $re_S(\tau_2) = \tau_4^*\tau_5^?$ のとき, $\tau_1^*[\tau_4^*\tau_5^?]_{\tau_2}\tau_3\tau_2$ も $\mathcal{R}_S$ に属する.

$R \in \mathcal{R}_S$ が表す言語 $L(R)$ は R から "[" と "]" をすべて取り除いて得られる正規表現が表す言語と定義する.

定義 7: S, S' をスキーマとする. $T_S \cup \{\Delta\}$ から $T_{S'} \cup \{\Delta\}$ への写像 θ と $T_S \cup \{\Delta\}$ から $R_{S'}$ への写像 ϕ が以下の条件を満たすとき, S は 弱包摂写像 (θ, ϕ) のもとで S' を弱包摂するといい, $S \preceq^w_{(\theta,\phi)} S'$ と書く.

1. $\theta(\tau) = \Delta \Leftrightarrow \tau = \Delta$.
2. 任意の $\tau \in T_S$ に対して, $pred_S(\tau) \subseteq pred_{S'}(\theta(\tau))$.
3. 任意の $\tau \in T_S \cup \{\Delta\}$ に対して以下が成り立つ. $\phi(\tau)$ の "[" または "]" で区切られた部分を左から順に $r'_1, \ldots, r'_n$ とする. このとき, $L(re_S(\tau)) = L(r_1 \cdots r_n)$ となる $r_1, \ldots, r_n$ が存在して, $L(\theta(r_i)) \subseteq L(r_i)$ かつ, 任意の $i \neq j$ に対して, r_i と r_j は共通の型を含まない.

例 5: 先の例のデータに加えて, 図 6 のようなスキーマのデータも統合することについて考える.

図 6 のデータは価格要素を持たないため, 包摂関係を用いて統合スキーマを作ると構造が複雑になる (図 7). 弱包摂関係を用いると図 4 と同じスキーマで統合できる.

$$\theta(SN\ デジカメ) = デジタルカメラ$$
$$\theta(\tau_{11}) = \tau'_{11}$$
$$\theta(\tau_{16}) = Option$$
$$\theta(\tau_{161}) = Any$$
$$\phi(SN\ デジカメ) = \tau'^{?}_{11}\tau'_{12}[\tau'_{131}\tau'_{132}]_{\tau'_{13}}\tau'^{?}_{14}Option^*$$
$$\ldots$$

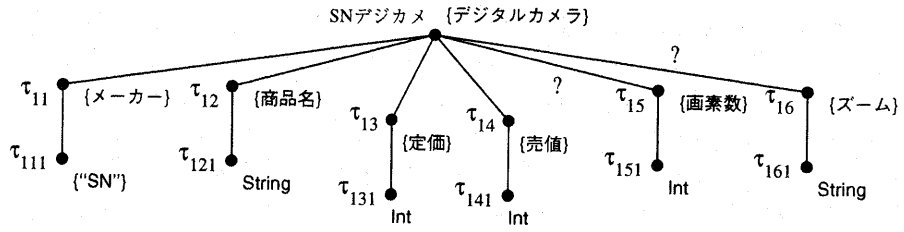

図 6: SN 製デジタルカメラのスキーマ

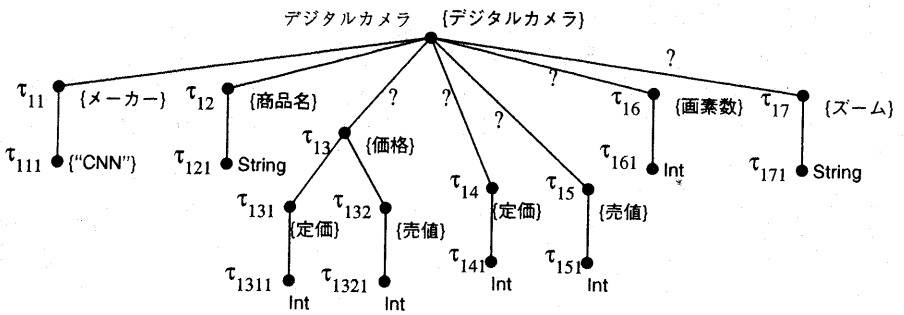

図 7: 一般のデジタルカメラのスキーマ 2

性質: S, S', S'' をスキーマ, D をデータベースとすると以下のことが成り立つ.

1. $S \preceq^w_{(\theta,\phi)} S$
2. $S \preceq^w_{(\theta_1,\phi_1)} S'$ かつ $S' \preceq^w_{(\theta_2,\phi_2)} S''$ ならば, $S \preceq^w_{(\theta_1\circ\theta_2,\phi_1\circ\phi_2)} S''$.
3. $S \preceq_\theta S'$ ならば, ある ϕ が存在して, $S \preceq^w_{(\theta,\phi)} S'$.

$S \preceq^w_{(\theta,\phi)} S'$ となるような θ, ϕ が存在するとき, $S \preceq^w S'$ と書く.

D が型 S を持ち, $S \preceq^w S'$ が成立する場合でも, D が型 S' を持つとは限らないことに注意されたい.

5 問い合わせの変換

弱包摂関係に基づいてスキーマ $S_1, \ldots, S_n$ を S' に統合したとし, それに対して 2.4 節で定義した問い合わせを行うことを考える. 問い合わせを処理する手順として, スキーマ $S_1, \ldots, S_n$ を型として持つデータをすべて, スキーマ S' を型として持つデータに変換する方法が考えられる. しかし, 変換すべきデータの量が膨大である場合, この方法は必ずしも効率的であるとはいえない. そこで本稿では, 問い合わせを元のスキーマに合わせて変換する方法について考える (4 節の最後で述べたように, 元のデータは型 S' を持つとは限らないので, 問い合わせの変換は必須である).

問い合わせの変換を以下の手順で行う.

$p_\phi(\tau_1, \tau_2)$ は, ϕ の値の中で τ_1 から τ_2 までのパス式の集合を表すものとする. $\tau' \in \{\theta(\tau) \mid \tau \in T_S\}$, $\tau'' \in T_{S'} - \{\theta(\tau) \mid \tau \in T_S\}$ とする.

1. S' におけるパス式に現れる // の直前あるいは直後にある τ'' を次の手順で取り除く.
   (a) $\tau'//\tau''$ の部分を各 $\tau'//p$ (ただし, $p \in p_\phi(\theta(\tau), \tau'')$) で置き換える. ここで τ は, $\phi(\tau)$ 中に τ'' が現れるような型である.
   (b) $\tau''//\tau'$ の部分を各 $p//\tau'$ (ただし, $p \in p_\phi(\tau'', \tau'_2)$) で置き換える. ここで, τ

を $\phi(\tau)$ 中に τ'' が現れるような型としたとき，τ'_2 は，$\theta(re_S(\tau))$ 中に現れ，かつ τ'' の子孫となる型である．

(c) $\tau''_1//\tau''_2$の部分を各 $p_1//p_2$ (ただし，$p_1 \in p_\phi(\tau''_1, \tau'_1), p_2 \in p_\phi(\theta(\tau_2), \tau''_2)$) で置き換える．ここで，$\tau_1$ を $\phi(\tau_1)$ 中に τ''_1 が現れるような型としたとき，τ'_1 は，$\theta(re_S(\tau_1))$ 中に現れ，かつ τ''_1 の子孫となる型である．また，τ_2 は，$\phi(\tau_2)$ 中に τ''_2 が現れるような型である．

(d) ある τ に対して，$\phi(\tau)$ 中に τ''_1, τ''_2 が現れる場合，$\tau''_1//\tau''_2$を各$p \in p(\tau''_1, \tau''_2)$ で置き換える．

2. 上で得られた各パス式を "/" から順に以下のように変換する．ただし，τは$\theta(\tau) = \tau'$ となる型であり，p' の変換結果を p と書く．

(a) /を/に変換する．

(b) p'/τ'を各 p/τ に変換する．

(c) $p'//\tau'$を各 $p//\tau$ に変換する．

(d) $p'/\tau''_1/\ldots/\tau''_n/\tau'$を各 p/τ に変換する．

例 6: 図 6 のスキーマ S を持つデータ D の定価の値を調べることを考える．統合したスキーマ (図 4) でのパスは，

$$p' = /\text{デジタルカメラ}/\text{価格}/\text{定価}$$

上の手順に従って，問い合わせを変換すると，

$$p = /\text{デジタルカメラ}/\text{定価}$$

となる．このパス式は，スキーマ S における根から定価要素までのパス式と一致する．

6 あとがき

本稿では，異なるスキーマを持つ複数の XML データを統合するための構造間の関係として，[3] で提案された包摂関係をもとにさらに条件の緩い弱包摂関係を提案した．この関係を用いると，包摂関係を用いるより単純なスキーマでデータを統合することができ，また，問い合わせの変換が弱包摂写像を利用して自動的に行えることを示した．

今後の課題としては，統合するスキーマとして最も単純なものを見つける事ができるか，また，スキーマの統合，問い合わせの変換に要するコスト等を調べる，あるいはこれらのコストを最小にするアルゴリズムの検討，等が挙げられる．

参考文献

[1] S. Abiteboul. "Querying semi-structured data." ICDT 1997, LNCS 1186, pp.1-18, 1997.

[2] C. Beeri and T. Milo. "Schemas for integration and translation of structured and semi-structured data." ICDT 1999, LNCS 1540, pp.296-313, 1999.

[3] G. M. Kuper and J. Siméon. "Subsumption for XML Types." ICDT 2001, LNCS 1973, pp.331-345, 2001.

[4] T. Milo and D. Suciu. "Type inference for queries on semistructured data." PODS 1999, pp.215-226, 1999.

[5] XML Path Language (XPath). W3C Recommendation 16 November 1999. http://www.w3.org/TR/xpath .