XML 文書における "意味単位" とその役割

鳥井修 † 木村哲郎 † 瀬川淳一 †
東芝 研究開発センター †

1. はじめに

本研究では，XML 文書の要素が "意味単位" と呼ばれるものと，それ以外のものの二種類に分類可能であることを発見した‡．意味単位は，これを削除または分割した結果，元の XML 文書が持っていた意味のまとまりを崩してしまう性質を持つ要素である．ある XML 文書から別の XML 文書への変換が行われる場合，この変換が可逆であるかどうか判定することは一般的には難しいが，"意味単位" に基づく解析により判定が可能となる．本稿では，意味単位について詳細に述べ，これを具体的にどのように用いて変換の可逆性を判定するか示す．

2. XML 文書の要素の分類

まず始めに，具体例に基づき，XML 文書（[1]）の要素が二種類に分類可能であることを示す．

```
<名簿>
  <従業員>
    <名前>
      <姓>鈴木</姓>
      <名>太郎</名>
    </名前>
    <所属>広報</所属>
  </従業員>
  <従業員>
    <名前>
      <姓>山本</姓>
      <名>武史</名>
    </名前>
    <所属>営業</所属>
  </従業員>
</名簿>
```

図 1: 従業員名簿

図 1 に示した XML 文書はある会社の従業員名簿であり，この文書により従業員の名前と所属が管理されている．本節ではこの XML 文書を対象とした変換に基づき要素の分類を行う．一般に XML 文書の変換を行うことで，元の XML 文書とは形式が異なる様々な形式の XML 文書を得ることが可能であることが知られているが（[2]），ここでは要素の削除という変換のみに注目する．

図 1 の XML 文書において『従業員』要素を削除し，元々『従業員』要素の子供要素であった『名前』要素と『所属』要素を，元々『従業員』要素の親要素であった『名簿』要素の子供要素に設定し直した結果を図 2 (1) に，図 1 の XML 文書において『名前』要素を削除し，『姓』要素と『名』要素を『従業員』要素の子供に設定し直した結果を図 2 (2) に示した．

```
<名簿>
  <名前>
    <姓>鈴木</姓>
    <名>太郎</名>
  </名前>
  <所属>広報</所属>
  <名前>
    <姓>山本</姓>
    <名>武史</名>
  </名前>
  <所属>営業</所属>
</名簿>
(1)
```

```
<名簿>
  <従業員>
    <姓>鈴木</姓>
    <名>太郎</名>
    <所属>広報</所属>
  </従業員>
  <従業員>
    <姓>山本</姓>
    <名>武史</名>
    <所属>営業</所属>
  </従業員>
</名簿>
(2)
```

図 2: 要素の削除

『従業員』要素を削除したことにより，意味のまとまりは崩れ，図 1 における『名前』要素と『所属』要素の対応関係が，図 2 (1) においては失われてしまっている．そのために，『従業員』要素を削除する前の要素間の対応関係に関する知識を図 2 (1) 外から得ない限り，図 2 (1) に『従業員』要素を挿入する逆変換を実行して図 1 を得ることは不可能である．

これに対して，『名前』要素を削除しても，意味のまとまりは変わらず，図 1 における要素の対応関係はすべて図 2 (2) において保たれている．したがって，『名前』要素を削除する前の要素間の対応関係に関する知識を図 2 (1) 外から得ることなしに，図 2 (2) に『名前』要素を挿入する逆変換を機械的に実行して図 1 を得ることが可能である．

以上説明を行った通り，XML 文書の要素は，(1) これを削除した結果，元の XML 文書が持っていた意味のまとまりを崩してしまうものと，(2) それ以外のものの，二種類に分類可能である．

3. 意味単位

前節において，具体例を用いて XML 文書の要素が二種類に分類可能であることを説明したが，一般に XML 文書の要素は，以下の性質を満たす E1，E2 の二つのグループに分類することが可能である．

1. E2 に属する任意の個数の要素を削除しても XML 文書の意味のまとまりが崩れない．
2. E1 に属する 1 個以上の要素と，E2 に属する 0 個以上の要素を適切に選んで削除することで XML 文書の意味のまとまりが崩れる．

"Meaningful Unit Elements" of XML Documents, Definition and Roll

† Osamu Torii, Tetsuro Kimura, Junichi Segawa, Corporate Research & Development Center, Toshiba Corporation

‡ 本研究は通信・放送機構の委託研究 “スーパーインターネットプラットフォーム技術の研究開発” の一部として行った.

上記 E1 の要素のことを以下『意味単位』と呼ぶことにする。本節では，意味単位の詳細説明を行い，意味単位を求める方法を述べる．

意味単位を削除することにより，意味のまとまりが崩れてしまう状況は，元々兄弟でなかった同じ名前を持つ要素が，要素を削除したことによって兄弟になる状況によって引き起こされる．例えば，図 1 において兄弟要素でなかった二つの『名前』要素と二つの『所属』要素が，『従業員』要素を削除した結果，図 2 (1) 示した通り兄弟要素となり，意味のまとまりが崩れている．

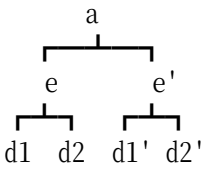

図 4: 意味のまとまり

より厳密に意味のまとまりが崩れる状況を説明すると以下の通りになる．図 4 はある XML 文書を木で表現したものであり，d1，d1' は同一の要素名 D1 を持つ要素，d2，d2' は同一の要素名 D2 を持つ要素，e は d1，d2 の共通の祖先，e' は d1'，d2' の共通の祖先，a は e，e' の共通の祖先であるとする．ただしこの図において，要素名 D2 は D1 と等しくてもよく，また要素 d2' は d1' と同一でもよいとする．図 4 では，要素 d1，d2 が対応関係にあり，これら要素が共通の祖先要素 e によって意味のまとまりを形成し，d1'，d2' が対応関係にあり，これらが e' によって意味のまとまりを形成している．

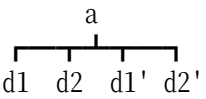

図 5: 意味のまとまりの崩壊

図 4 において，a の子孫であり，かつ d1，d2，d1'，d2' いずれかの祖先になっている要素をすべて削除した結果が図 5 に示してある．この変換により，図 4 では兄弟でなかった d1 と d1'，d2 と d2' が図 5 においてそれぞれ兄弟になったことで，図 4 における意味のまとまりが崩れている．

上記変換では多くの要素を削除し過ぎたために意味のまとまりが崩れた．そこで，変換を行う際に a の子孫であり，かつ e の祖先（e も含む）である要素の中から，少なくとも一つ削除されない要素を選び，a の子孫であり，かつ e' の祖先（e' も含む）である要素の中から，少なくとも一つ削除されない要素を選べば，意味のまとまりは変わらず，d1 と d2，d1' と d2' の対応関係が保たれる．

このことから以下の方法で XML 文書の要素 e にラベルを付けることにより，意味単位を求めることが可能である．

(1) e が子孫要素を持たない場合．
　e のラベルは『通常』と決定する．

(2) e が子孫要素を持つ場合．
　(2-1) e 子孫と e の兄弟の子孫すべてに再帰的にラベルを付ける．
　(2-2) 以下の条件を満たす要素 e’，d1，d2 (≠ d1)，d1'，d2' が存在するとき，e，e' のラベルを『意味単位』に決定する．
- e’は e の兄弟である．
- d1，d2 は e の子孫，d1'，d2' は e' の子孫である．d1，d2 は異ならなければならないが，d1' ,d2' は同一であっても構わない．
- d1，d1' は同じ要素名を持ち，d2，d2' は同じ要素名を持つ．d1，d2，d1'，d2' がすべて同じの要素名を持っても構わない．
- e と d1 を結ぶ最短パス上の要素（e，d1 は除く）は，すべてラベル『通常』を持つ．d2，d1'，d2' に関しても同様．

　(2-3) e にまだラベルが付けられていない場合には，e のラベルを『通常』に決定する．

4．XML 文書変換の可逆性判定

同一の XML 文書を複数のユーザー間で共有する場合など，様々なケースにおいて XML 文書変換の可逆性判定が必要である ([3])．変換によって意味のまとまりが崩れた場合には，もはや逆変換を行って元の XML 文書を得ることは不可能である．変換前の XML 文書と変換後の XML 文書を比較して，もし意味単位を削除する変換を行っていたら，この変換は不可逆と判定される．また，変換によって意味単位の子孫要素が，子孫以外の場所に移動されていたら，この変換も不可逆と判定されるべきである．上記の通り，意味単位を用いて変換の前後における意味のまとまりの変化を調べることで，変換の可逆性を判定することが可能である．

5．おわりに

本稿では，XML 文書の要素が，意味単位と呼ばれるものとそれ以外のものの二種類に分類可能であることを説明し，意味単位を求める方法を述べた．ある XML 文書から別の XML 文書への変換が行われる場合，意味単位を用いることでこの変換が可逆であるかどうか判定することが可能である．

参考文献

[1] Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, "Extensible Markup Language (XML) 1.0 (Second Edition)", http://www.w3.org/TR/2000/REC-xml-20001006, W3C Recommendation, 6 October 2000

[2] James Clark, "XSL Transformations (XSLT) Version 1.0", http://www.w3.org/TR/1999/REC-xslt-19991116, W3C Recommendation, 16 November 1999.

[3] 鳥井 修，木村 哲郎，瀬川 淳一，“XML 文書の双方向変換機構～住所録への適用～”，情報処理学会第 30 回 デジタル・ドキュメント研究会，2001.