1G3-05

シリーズ型HTML文書のXML文書への半自動変換法の改良

Improvements of a Method for Transforming HTML Documents to XML Documents

倉林寛幸[*1] 正田大輔[*2] 勝野裕文[*2]
Hiroyuki Kurabayashi Daisuke Shoda Hirofumi Katsuno

[*1]太田市立宝泉中学校 [*2]東京電機大学理工学部
Hosen Junior High School Tokyo Denki University

There is a huge amount of information available on the World Wide Web, but it is difficult to locate the information that we intend to obtain. If Web documents become machine-understandable, the difficulty will be overcomed. Therefore, it is important to transform Web documents written in HTML into XML documents, because XML allows to use meaningful tag names. Umehara and et al. showed a case-based, semi-automatic method of transforming a series of HTML documents that are similar in appearance and contents into XML documents. This paper presents some significant changes in their semi-automatic transformation method, and experimental results that show our new method is an improvement in their method.

1. はじめに

コンピュータが文書の意味を理解できれば，インターネット上の大量の情報から人間が意図する情報を的確に検索できる．しかし，現状ではコンピュータが人間と同じように文書の意味を理解するのは困難で．予めその理解を助ける手がかりを文書に付加するなどの補助手段が有効である．

一方，多くの Web 文書は HTML 言語で書かれているが，HTML 言語は文書のレイアウトの記述することを目的とする言語なので，HTML 文書を意味処理するよりは，タグ（要素名）に意味的な情報を付加できる XML 文書の方が機械的に意味処理しやすい．従って，HTML 文書を XML 文書への自動変換が重要になる．

[梅原 01] は HTML 文書から XML 文書への変換手法として，事例に基づく半自動変換法を提案した．彼らは，教員紹介，シラバス，オークションサイトのような，記述されている項目が同じで内容が異なる文書（シリーズ型 HTML 文書と言う）に着目し，HotSpa，JTB，YAHOO，大学の教員紹介などのシリーズ型 HTML 文書に対して XML 文書への変換を試み，80%を超える変換精度を得ている．また，彼らの手法では，HTML 文書をテキストブロック呼ぶ，意味的に一塊と考えられる区画に自動的に分割するが，その精度に問題があるので，[梅原 02] はアラインメント技術を用いて改良を図っている．

我々は，[梅原 01] が提案した手法を追試する過程で，彼らの手法では類似度が高いテキストブロックが複数存在する場合に XML 文書への変換精度が上がらない場合があることに気がついた．我々は，その問題を解決するため，HTML 文書の構文的特徴などを利用して梅原らの手法を改良したので，その結果を報告する．

2. 変換手法

変換全体の流れは以下のようになる．

1. 人間がシリーズ型 HTML 文書群の中から代表的な HTML 文書を一つ選びそれを XML 文書へ変換して，変換事例を作成する．以後，この選んだ HTML 文書を事例 HTML 文書，変換された XML 文書を事例 XML 文書と呼ぶ．
2. 事例 HTML 文書と事例 XML 文書の構造解析を行い，その解析結果を踏まえて変換対象 HTML 文書の構造解析を行う．
3. これらの情報を利用して有益な文字列のみに絞り込んだ後にそれらの意味的な解析を行う．
4. 全体の解析結果を使い、変換対象 HTML 文書に対応する XML 文書を生成する．

以後これらのステップの詳細を述べる．このとき，下に示す説明例の文書を用いる．

[事例 HTML 文書]

```
<HTML>
  <HEAD><TTILE>教員紹介</TITLE></HEAD>
  <BODY>
     <H1>山田太郎教授</H1>
    <TABLE><TR>
      <TR><TD>専門分野</TD>
          <TD>人工知能</TD></TR>
      <TR><TD>所属学会</TD>
          <TD>情報科学学会</TD></TR>
    </TR></TABLE></BODY>
</HTML>
```

[事例 XML 文書]

```
<教員紹介>
  <教員名>山田太郎教授</教員名>
  <専門分野>人工知能</専門分野>
  <所属学会>情報科学学会</所属学会>
</教員紹介>
```

[変換対象 HTML 文書]

```
<HTML>
  <HEAD><TITLE>教員紹介</TITLE></HEAD>
  <BODY>
     <H1>田中花子助教授</H1>
    <TABLE><TR>
      <TR><TD>専門分野</TD>
          <TD>データベース</TD></TR>
      <TR><TD>所属学会</TD>
```

連絡先: 勝野裕文，東京電機大学理工学部情報科学科，埼玉県比企郡鳩山町石坂，katsuno@j.dendai.ac.jp

表 1: 事例文書の解析結果のまとめ

s-ブロック	x-タグ	構文木パス
教員紹介	なし	<HTML><HEAD><TITLE>
山田太郎教授	名前	<HTML><BODY><H1>
専門分野	なし	<HTML><BODY><TR><TD>
人工知能	専門分野	<HTML><BODY><TR><TD>
所属学会	なし	<HTML><BODY><TR><TD>
情報科学学会	所属学会	<HTML><BODY><TR><TD>

```
        <TD>データベース学会</TD></TR>
    </TR></TABLE></BODY>
</HTML>
```

2.1 人間による事例 XML 文書の生成

事例 HTML 文書を人間が XML 文書に変換するときには，XML 文書で使う要素名として HTML 文書中で使われている文字列 (厳密には次節で示す s-ブロック) を用いると仮定する．これは 2.3 節における見出しの検出で述べるように，見出しを確実に抽出するための制限である．HTML 文書中の文字列と異なる要素名を使いたい場合は，本論文で述べる方法で XML 文書を作成した後に，要素名の付け替えを行なえばよい．

2.2 構造的な解析

事例 HTML 文書から開始タグと対応する終了タグで囲まれた文字列を抽出する．抽出した文字列をテキストブロックという．特に，事例 HTML 文書中のテキストブロックを s-ブロックという．メタタグや文字の位置，フォントなどを示すタグは無視し，前後が続いているものとした．テキストブロックの連結にはまだ改良するすべき点があるが ([梅原 02] 参照)，本研究では，変換方法の改良に焦点を当てるため，上の方式を採用した．説明用の事例 HTML 文書からは以下の s-ブロックが抽出される．

・教員紹介　・山田太郎教授　・専門分野
・人工知能　・所属学会　・情報科学学会

抽出した s-ブロックが事例 XML 文書のどのタグに囲まれているかを調べる．事例 XML 文書中でテキストブロックを囲む最も内側のタグの要素名を x-タグと言う．説明用の事例 HTML 文書中の山田太郎教授は，XML 文書中で教員名という要素名を持つタグに囲まれているので，山田太郎教授の x-タグは教員名となる．

次に，各テキストブロックが HTML 文書中のどの位置にあるかを示すために構文木パスを用いる．テキストブロックの構文木パスとは，そのブロックの先頭までに開いていて，まだ閉じていない開始タグの列を指す．説明用の事例 HTML 文書中の「専門分野」の構文木パスは<HTML><BODY><TR><TD>となる．

以上の解析結果をまとめると表 1 の通りとなる．

事例 HTML 文書と同様に，変換対象 HTML 文書も解析を行い，テキストブロックの抽出とその構文木パスを求める．説明用の変換対象 HTML 文書では，表 2 の結果を得る．以下では，変換対象 HTML 文書中のテキストブロックを t-ブロックという．

2.3 見出しと不要文字列の決定

見出しとは，あるテキストブロックが何を示すかが書かれたテキストブロックを指す．見出しを有効に使うことにより精度の高い変換が可能になる．

表 2: 変換対象 HTML 文書の解析結果のまとめ

t-ブロック	構文木パス
教員紹介	<HTML><HEAD><TITLE>
田中花子教授	<HTML><BODY><H1>
専門分野	<HTML><BODY><TR><TD>
データベース	<HTML><BODY><TR><TD>
所属学会	<HTML><BODY><TR><TD>
データベース学会	<HTML><BODY><TR><TD>

[梅原 01] ではテキストブロックの見出しをそのブロックの一つ前の<H1>〜 <H6>に囲まれたテキストブロックとしていた．本研究では，より確実に見出しが取り出せるよう，2.1 節で示したように事例 XML 文書の作成で使える要素名を，事例 HTML 文書中の s-ブロックとした．従って，本論文では s-ブロックが，事例 XML 文書の要素名であり，その要素の内容が別の要素を含まないならば，その s-ブロックと同じ文字列からなるテキストブロックを見出しという．説明用の変換対象 HTML 文書では，専門分野と所属学会が見出しになる．

不要文字列とは，シリーズ内のどの HTML 文書を変換対象として選んでも，必ず出現する t-ブロックで，見出しでないものをさす．説明用の例では，教員紹介が不要文字列になる．不要文字列は，事例 XML 文書に取り込まれていなければ，自動変換結果の XML 文書に取り込まれない．

2.4 意味的な解析

見出しや不要文字列でないテキストブロックに対して意味的な特徴を解析する．各テキストブロックを形態素解析し，名詞を抽出し，その出現数を数え上げる．形態素解析には茶筌 [松本 03] を用いた．説明用の事例 HTML 文書を解析すると表 3 のようになる．

s-ブロックに含まれている名詞数でベクトルを作る．このベクトルを項ベクトルと言う．作り方は，表 3 の結果を順に並べればよい．まとめると表 4 のようになる．

表 4: 項ベクトル

s-ブロック	項ベクトル
山田太郎教授	(1,1,1,0,0,0,0,0)
人工知能	(0,0,0,1,1,0,0,0)
情報科学学会	(0,0,0,0,0,1,1,1)

変換対象文書も同様に形態素解析するが，変換対象 HTML

表 3: 各 s-ブロックにおける名詞の出現数

	山田	太郎	教授	人工	知能	情報	科学	学会
山田太郎教授	1	1	1	0	0	0	0	0
人工知能	0	0	0	1	1	0	0	0
情報科学学会	0	0	0	0	0	1	1	1

文書では，事例 HTML 文書から抽出した名詞のみを数え上げる．t-ブロックにおいて，事例 HTML 文書に含まれる名詞の出現数は表 5 のようになる．事例 HTML 文書と同様に項ベクトルを作成する．たとえば「田中花子教授」の項ベクトルは，(0,0,1,0,0,0,0,0) になる．

二つのテキストブロックは互いの項ベクトルが類似しているほど，意味的な特徴が類似していると考えられる．2つの項ベクトル間の類似度は2つの項ベクトルの内積に両テキストブロックの名詞の個数の比をかけたものを用いた．s-ブロックの項ベクトルを V_i，その名詞数の和を n_i，t-ブロックの項ベクトルを V_j，その t-ブロックの名詞数の和を n_j とすれば類似度は次の式で定義される．

$$Sim(V_i, V_j) = V_i \cdot V_j \times \frac{\min(n_i, n_j)}{\max(n_i, n_j)}$$

2.5 XML 文書の生成

これまでの解析結果を利用して，変換対象 HTML 文書を XML 文書に変換する．事例 XML 文書で x-タグを要素名として持つ要素の内容を取り除き（取り除いて得られる XML 文書を雛形と言う），その代わりの適切な t-ブロックを割り当てることにより，変換対象 HTML 文書に対応する XML 文書を生成する．

本研究では，雛形の要素の内容に適切な t-ブロックを割り当てる方法において，内容を決め易い要素から順に決定する方式をとる．雛形において x-タグを要素名として持つ要素の集合を E，E の要素 N と同じ要素名を持つ，事例 XML 文書の要素の内容を $s(N)$ とする．ここで，事例 XML 文書の作り方から $s(N)$ は s-ブロックの一つである．さらに，変換対象 HTML 文書のすべての t-ブロックからなる集合から，見出しと不要文字列を取り除いた集合を Ct とすれば、決定手順は以下の通りである．

1. 要素 N の内容の決定しやすさを，0 または 1 からなる 3 つ組 (a, b, c) で表し，この 3 つ組を $\mu(N)$ で表す．
2. 辞書式順序で考えて最大となる $\mu(N)$ の要素 N の内容を決定する．
3. 2 で内容が決まった N を E から取り除き，その内容を Ct から取り除く，$E \neq \phi$ なら 1 に戻り ($\mu(N)$ は Ct により変化する)，$E = \phi$ なら終了する．

ここで，$\mu(N) = (a, b, c)$ は次のように決める．

a) N の要素名が見出しであるならば $a = 1$，さもなければ $a = 0$ とする．

b) $s(N)$ の事例 HTML 文書における構文木パスと同じ構文木パスを持つ t-ブロックが Ct の中で一意に決まるならば $b = 1$，さもなければ $b = 0$ とする．

c) $s(N)$ と Ct の各 t-ブロック間で類似度を計算し，その最大値が 2 番目に大きい値の 2 倍以上あるならば $c = 1$，さもなければ $c = 0$ とする．

すなわち，上のステップ 2 では見出しを一番重要視し，以下構文木パス，意味的類似度の順で重要視して，内容を決めるべき要素 N を決定する．N の内容となる t-ブロックを以下の手順で決める．

(Case 1) $a = 1$ の場合

この場合，N の要素名は見出しかつ s-ブロックであるが，さらに t-ブロックでもあれば，事例 HTML 文書で N と $s(N)$ の間にある s-ブロックの数を k とする．次に，変換対象 HTML 文書で N と k だけ離れた t-ブロックが Ct の中に存在すれば，それを N の内容とする．上の手順で N の内容が決まらなければ,「なし」と決定する．

(Case 2) $a = 0$ かつ $b = 1$ の場合

$s(N)$ の事例 HTML 文書における構文木パスと同じ構文木パスを持つ t-ブロックを N の内容とする．

(Case 3) $a = b = 0$ かつ $c = 1$ の場合

$s(N)$ との類似度が最大の t-ブロックを N の内容とする．

(Case 4) $a = b = c = 0$ の場合

N の内容は「なし」とする．

以上の作業により以下の XML 文書を得る．

```
[変換で得られた XML 文書]
<教員紹介>
  <教員名>田中花子教授</教員名>
  <専門分野>データベース</専門分野>
  <所属学会>データベース学会</所属学会>
</教員紹介>
```

3. 評価実験の結果

提案した変換手法の変換精度を評価する実験を行った．実験では表 6 における 5 サイトのデータを用い，各サーバにある HTML 文書に文法的な誤りがない限りそのまま無作為に使い，各サイトの適当な 1 文書を事例 HTML 文書として用いた．表 7 における文書数は，実験に用いた各サイトの HTML 文書から事例 HTML 文書を除いた変換対象文書数である．変換項目数は，人間により事例 XML 文書に変換した際に，要素の内容となった s-ブロックの項目数を示す．事例 XML 文書での要素名は実験者が決めた．平均テキストブロック数は，本論文の手法で抽出された数である．変換精度は次の式で定義した．

$$\text{変換精度} = \frac{\text{正しい内容を持つ要素の総数}}{\text{内容を割り当てるべき要素の総数}} * 100$$

但し，変換対象 HTML 文書に該当項目がない場合は，なしと出力したとき正しいと判定した．

実験の結果，1〜5 のどのサイトのデータに対しても高い変換精度が得られた．1 の文書はテーブル構造が主体なので見出

表 5: 事例文書で出現した名詞の変換対象 HTML 文書における出現数

	山田	太郎	教授	人工	知能	情報	科学	学会
田中花子教授	0	0	1	0	0	0	0	0
データベース	0	0	0	0	0	0	0	0
データベース学会	0	0	0	0	0	0	0	1

表 6: 実験で用いたサイト

シリーズ	サイト名	URL
1	東京電機大学理工学部情報科学科教員紹介	http://www.j.dendai.ac.jp/kyoin/
2	東京電機大学理工学部シラバス	http://www.dendai.ac.jp/
3	JTB	http://www.jtb.co.jp/
4	yahoo オークション	http://auctions.yahool.co.jp/
5	asahi.com	http://www.asahi.com/

表 7: 変換精度評価の結果

シリーズ	文書数	平均テキストブロック数	変換項目数	変換精度
1	10	約 15	7	98%
2	50	約 30	12	100%
3	10	約 60	7	100%
4	10	約 60	10	88%
5	10	約 60	3	93%

しが多く，大部分は見出しにより決定された．見出しのない項目も構文木パスが有効に働き，正しく変換された．また，特殊な構造をした文書が 1 つあったが，見出しが発見でき，6/7 の正解率を得た．

２の文書は，決まったフォーマットの空所を埋めて出来る文書なので，見出しだけ存在する項目が多々あったが，見出しの次に見出しの内容があったため，見出しと不要文字列を変換に使う t-ブロックに含めないという制約が機能し，誤変換はなかった．また，講義内容と講義概要，教科書と参考書，時限とオフィスアワーのような意味的に競合するテキストブロックが多かったが，類似度が近い t-ブロックを減らす方法が功を奏し，高い精度を得た．

３の文書は，テーブル構造をしていてすべての項目に見出しがある文書で，誤変換がなかった．

４の文書では商品名以外は変換できた．特に事例文書と変換対象文書が同分類の商品の場合は，100%の結果を得た．しかし，商品名が不要文字列と一緒になる場合は，意味解析に悪い影響があった．

５の文書は，見出しによる決定を行わなかった．テキストブロックが約 60 個あるにもかかわらず，大多数のテキストブロックは，不要文字列となり高い変換精度が出た．また，間違えた文書も，誤変換ではなく t-ブロックが，適切に分離できなかった誤りであった．

全体として，確実に見出しを探し出すことができた点が変換精度の向上に大きく寄与している．さらに，見出しと不要文字列を変換に使う t-ブロックから除外したことが影響し，構造解析，意味解析，両者の精度を向上させた．また [梅原 01] のように，最初に出現する要素からその内容を決めるのではなく，要素の内容の決めやすさを考慮したので，誤変換が減少した．特にテーブル構造をしている場合，[梅原 01] の方法では，構文木パスは同じ，見出しも取得できないという結果になり意味解析に頼らざる得ないが，今回提案した方法では見出し，構文木パスの情報が使え変換精度が上がる．

4. 終わりに

今回の実験では変換対象テキストブロックが，タグ情報で的確に分離できたことも，高い変換精度が得られた要因の一つと考えられる．従って，今後は [梅原 02] が議論しているような，タグ情報だけではテキストブロックを的確に抽出できない場合のテキストブロック抽出方法の検討を進める必要がある．また，本論文で提案した手法はテーブル構造の取り扱いを得意としているが，実験に現れた構造よりもより複雑なテーブル構造に対しては改良の余地があると考えられるので，その検討も今後の課題である．

参考文献

[松本 03] 松本，北内，山下，平野，松田，高岡，浅原：形態素解析システム「茶筌」Version2.3.3 奈良先端科学技術大学院大学 (2003)

[梅原 01] 梅原，岩沼，長井：事例に基づく HTML 文書から XML 文書への半自動変換，人工知能学会論文誌 16 巻 5 号 B, pp.408-416（2001）

[梅原 02] 梅原，岩沼，鍋島：事例に基づくシリーズ型 HTML 文書の意味論理構造の自動認識，人工知能学会論文誌 17 巻 6 号 E, pp.690-698（2002）