事例に基づく HTML 文書から XML 文書への半自動変換

— シリーズ型 HTML 文書における類似性の利用 —

A Case-Based Semi-automatic Transformation from HTML Documents to XML Ones

— Using the Similarity between HTML Documents Constituting a Series —

梅原 雅之

Masayuki Umehara

山梨大学大学院工学研究科電子情報工学専攻

Graduate School of Electrical Engineering and Computer Science, Yamanashi University
mon@iw.media.yamanashi.ac.jp

岩沼 宏治

Koji Iwanuma

山梨大学工学部コンピュータ・メディア工学科

Department of Computer Science and Media Engineering, Yamanashi University
iwanuma@iw.media.yamanashi.ac.jp

永井 宏和

Hirokazu Nagai

東芝 IT コントロールシステム株式会社

Toshiba IT Control Systems Corporation
hirokazu-kun@nyc.odn.ne.jp

keywords: HTML, XML, case-based transformation, information retrieval, text processing

Summary

In order to utilize a large quantity of information in Internet, machine processing of HTML documents has been becoming tremendously important. HTML, however, is designed mainly for reading with browsers, thus not suitable for machine processing. XML was proposed as a solution for this problem. Unfortunately, full automatic transformation from HTML to XML is extremely difficult, because it absolutely demands to understand the meaning of HTML documents. On the other hand, there are many series of HTML pages in actual Web sites. Each page of a series usually has a quite similar structure with each other. Therefore a case-based transformation must be a promising method in practice. In this paper, we give a case-based transformation method from HTML documents to XML ones. Given a series of HTML documents and a sample transformation from a selected HTML document into XML one, we first analyze both of the semantic and syntactic information appearing in the sample pair. Next the remaining HTML pages of the series are automatically transformed into XML documents by using the information previously extracted from the sample. We adopt a vector model of term weighted frequency for approximating the meaning of HTML documents, and also use both headlines and a parse tree as syntactical information. Throughout experimental evaluation, we show this case-based method achieved a highly accurate transformation, i.e., 80% of actual 80 pages can be transformed in a correct way.

1. は じ め に

本論文では、シリーズ型の HTML 文書^{*1}群を対象として、事例に基づく XML 文書への変換法を提案し、その実験的評価を行う。

現在インターネット上には極めて多量の情報が溢れており、その有効利用のためには機械処理が極めて重要である。インターネット上の殆どのデータは HTML で記

述されているが、HTML は人が閲覧するデータを記述するための言語であるため、機械処理には向いていない。HTML 文書からの情報抽出法は、ラッパー (wrapper) など幾つかの技術が開発されている (8 章関連研究を参照)。しかし、有用な情報を抽出するためには、HTML 文書の何らかの意味理解がどうしても必要となるため、その自動化と高度化には大きな限界がある。

近年、XML [村田他 98, XML/SGML サロン 98] がデータ交換用の言語として設計され、大きな注目を集めている。XML ではタグを用いて意味記述を行う。もし HTML 文書の意味要素をタグ化し、XML 文書へ変換で

^{*1} ここでは、シリーズ型の HTML 文書を「類似した情報を提示するために作られた文書構造が類似している HTML 文書群」と考える。

```

<HTML>
  <HEAD> <TITLE>温泉案内</TITLE> </HEAD>
  <BODY>
    <H1>白骨温泉</H1>
    <H2> 料金</H2>
      <P>500 円</P>
    <H2> 営業時間</H2>
      <P>10 時から 5 時</P>
  </BODY>
</HTML>

```

図 1 HTML 文書の例.

きたなら、インターネット上の文書の機械処理が格段に容易になる。しかし、人手による変換は多大なコストがかかるため機械的な変換が望ましい。そのためには、HTML 文書の意味理解が必要となり、完全な自動化は殆ど不可能である。

ここで現実の Web サイトを眺めると、シリーズ型の HTML 文書が極めて多いことに気付く。通常、同一のシリーズの文書は、記述されている情報の意味や文書の構造が相互に良く類似しており、この類似性を利用すれば、事例に基づく推論による HTML 文書から XML 文書への自動変換が十分可能と推測される。着目したシリーズ型文書群に固有の意味構造が、事例を通して自然に、変換システム内で取り扱うことが可能となる。

本論文で提案する変換法では、まずシリーズ中の一つの HTML 文書を人手で XML 文書に変換する。次に、この 2 つの文書を変換事例としてシステムに与え、シリーズの残りの HTML 文書を事例に従って自動変換するものである。システムを実装し、評価実験の結果、現実の Web サイトから収集した 5 個のシリーズ、80 件の HTML 文書に対して、本提案手法は約 80%を越える高い変換精度を持つことが確認した。筆者らの知る限りにおいて、現実の Web サイト上の HTML 文書のシリーズ性に着目した研究は、これまでに全く無く、本研究が最初の試みと考えられる。

本論文の構成は以下の通りである。まず 2 章で HTML と XML について概説し、3 章で事例に基づく変換法の概要を述べる。4 章から 6 章で変換法の詳細を述べ、7 章で実験の評価を行う。8 章で関連研究を述べる。9 章は結論と今後の課題である。

2. HTML と XML

HTML 文書は、図 1 の様に<H1>や<P>などのタグと、“温泉案内”や、“料金”の様な普通のテキストが混在した文書である。HTML の<H1>や<P>等のタグは基本的にレイアウト制御用であり、それぞれの意味は事前に決められており、自由にタグを設けることはできない。文書レイアウト用のタグだけでは情報が不足しており、データの内容が表現しきれず、機械処理が困難である [XML/SGML サロン 98, 村田他 98]。

一方 XML は、図 2 上部の様な文書型を定義する DTD

DTD 部

```

<!ELEMENT 温泉案内 (温泉名, 料金, 営業時間)>
<!ELEMENT 温泉名 (#PCDATA)>
<!ELEMENT 料金 (#PCDATA)>
<!ELEMENT 営業時間 (#PCDATA)>

```

XML インスタンス部

```

<温泉案内>
  <温泉名>白骨温泉</温泉名>
  <料金>500 円</料金>
  <営業時間>10 時から 5 時</営業時間>
</温泉案内>

```

図 2 XML 文書の例.

と、図 2 下の様な実際のタグ付き文書である XML インスタンスからなり、必要に応じて独自のタグを設けることができる。例えば、図 2 の温泉案内の XML 文書では<温泉案内>、<温泉名>、<料金>、<営業時間>などのタグを設けている。これにより、どのデータが温泉名、料金なのか判定できるため、機械処理は容易となる。

XML 文書进行操作するためのライブラリ (パーザ) などは既にならかなり開発されており、機械処理の支援環境は整備されている [XML/SGML サロン 98, 村田他 98]。HTML 文書を XML 文書に変換することにより、インターネット上に存在する多量のデータの機械処理が格段に容易になると考えられる。

3. 事例に基づく変換

多量の HTML 文書の XML 文書への変換は非常に手間のかかる作業である。そこで、シリーズ型 HTML 文書群に対する、事例に基づく半自動変換手法を提案する。

変換手法の概要を図 3 に示す。システムへの入力は、事例の HTML 文書と、それを人手で変換生成した XML 文書、そして新たに事例に基づいて変換する HTML 文書群である。例えば、図 4 のような温泉案内の HTML 文書を変換したい場合、同じシリーズの HTML 文書の図 1 と、それを XML に変換した図 2 の組を変換事例として与える。以後、事例に基づき自動変換する HTML 文書を (事例の HTML 文書と区別して) 変換対象と呼ぶ。

以下、4 章で変換事例の解析について、5 章で変換対象の解析、6 章で XML 文書生成について説明する。

4. 事例解析

変換事例の解析は以下の 3 つの作業から成る。

- (1) 変換事例の HTML 文書中のテキストの分割。
- (2) 変換事例の HTML 文書と XML 文書の対応関係の解析。
- (3) 分割された各テキストの特徴の解析。

以後、図 1 と図 2 を例として用いて、変換事例の解析法を説明する。まず最初に変換事例の HTML 文書のテキ

図3 事例に基づく変換手法の概要.

```

<HTML>
<HEAD> <TITLE>温泉案内</TITLE></HEAD>
<BODY>
  <H1>笛吹川温泉</H1>
  <H2> 料金</H2>
  <P>600 円</P>
  <H2> 営業時間</H2>
  <P>10 時から 17 時</P>
</BODY>
</HTML>

```

図4 温泉案内の HTML 文書の例.

ストを分割する.

4.1 変換事例の HTML 文書中のテキストの分割

変換事例の HTML 文書中のテキストを、複数のブロックに分ける. 本手法では、タグで囲まれているブロックを意味を成す最小のブロックと考え、任意のタグに囲まれているテキスト部分を 1 つのブロックとして分割、処理する. これらのブロックをテキストブロックと呼ぶ. 変換事例の HTML 文書中の各テキストブロックを以下 s-ブロックと呼ぶ.

例えば、図 1 中のテキストは以下の 6 つのテキストブロックに分割される.

- (1) 温泉案内 (2) 白骨温泉 (3) 料金
(4) 500 円 (5) 営業時間 (6) 10 時から 5 時

4.2 変換事例の HTML 文書と XML 文書の対応関係

変換事例の HTML 文書中の s-ブロックが、XML 文書のどこに出現しているかを調べ、XML 文書で囲んでいる XML タグを表にする. s-ブロックを囲む XML タグを x-タグと呼ぶ. 例として、図 1 と図 2 の関係を表 1 に示す.

表1 図1の HTML と図2の XML の対応関係.

s-ブロック	x-タグ
温泉案内	—
白骨温泉	温泉名
料金	—
500 円	料金
営業時間	—
10 時から 5 時	営業時間

4.3 分割された HTML テキストの解析

次に変換事例の HTML 文書における特徴を解析する. 特徴は構造的な特徴と、意味的な特徴を両方考慮する.

§1 構造的な特徴の解析

s-ブロックの構造的な特徴として構文木パスと見出しの 2 つを考慮する.

s-ブロック B の構文木パスとは、変換事例の HTML 文書の構文解析木の根から B までのパスのことである. 例えば、図 1 の構文解析木は図 5 になるので、s-ブロック “白骨温泉” の構文木パスは $\langle \text{HTML} \rangle - \langle \text{BODY} \rangle - \langle \text{H1} \rangle$ となる. 本手法では、同一のタグの出現回数や順序や出現位置を考慮せず、単純なパス構造のみを取り上げる. より複雑な処理をすることも可能であるが、中途半端な処理は全体の変換精度を劣化させると予想されるので、ここでは考慮しない [Ashish and Knoblock 97].

s-ブロック B の見出しとは、HTML 文書中で B の直前に出現した見出し ($\langle \text{H1} \rangle \sim \langle \text{H6} \rangle$ タグで囲まれているテキストブロック) のことである. 例えば、図 1 中の s-ブロック “10 時から 5 時” に対応する見出しは、“営業時間” である.

本手法では、 $\langle \text{H1} \rangle \sim \langle \text{H6} \rangle$ 以外のタグは意味を考慮せずに、テキストブロックを分けるための区切り記号、構文木パスの構成要素として使用する. この意味では、本変換システムは任意の HTML 文書に対応していると言える.

§2 意味的な特徴の解析

s-ブロックの意味的な特徴を表すものとして、項ベクトル [徳永 99] を用いる. まず、各 s-ブロックを形態素解析し、名詞の出現数を数え上げる [長尾 96, 松本他 99]. 例として図 1 の HTML 文書を変換事例としたときの各

図5 図1の HTML 文書の構文解析木.

表 2 図 1 の HTML 文書の各 s-ブロックの名詞の出現数 .

s-ブロック \ 名詞	温泉	案内	白骨	料金	500	円	営業	時間	10	時	5
温泉案内	1	1	0	0	0	0	0	0	0	0	0
白骨温泉	1	0	1	0	0	0	0	0	0	0	0
料金	0	0	0	1	0	0	0	0	0	0	0
500 円	0	0	0	0	1	1	0	0	0	0	0
営業時間	0	0	0	0	0	0	1	1	0	0	0
10 時から 5 時	0	0	0	0	0	0	0	0	1	2	1

表 3 変換事例として HTML 文書 (図 1), XML 文書 (図 2) を与えた時の変換事例の解析結果 .

s-ブロック	x-タグ	構文木パス	見出し	項ベクトル (WIDF)
温泉案内	なし	<HTML><HEAD><TITLE>	なし	(0.5, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
白骨温泉	温泉名	<HTML><BODY><H2>	なし	(0.5, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)
料金	なし	<HTML><BODY><H1>	白骨温泉	(0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)
500 円	料金	<HTML><BODY>	料金	(0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0)
営業時間	なし	<HTML><BODY><H2>	なし	(0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0)
10 時から 5 時	営業時間	<HTML><BODY>	料金	(0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1)

s-ブロックの名詞の出現数を表 2 に示す . 次に , 出現数を基に , 各名詞の HTML 文書における重要度を表す重み付けをする . 重み付け手法は TF・IDF [徳永 99] と WIDF [徳永, 岩山 96] を用いる .

N を HTML 文書中の s-ブロックの総数 , $TF(d_i, t)$ を s-ブロック d_i における語 t の出現数 , $df(t)$ を語 t が出現する s-ブロック数としたとき , s-ブロック d_i における語 t の TF・IDF は式 (1) , WIDF は式 (2) と定義される .

$$TF \cdot IDF(d_i, t) = \log \frac{N}{df(t)} \times TF(d_i, t) \quad (1)$$

$$WIDF(d_i, t) = \frac{TF(d_i, t)}{\sum_{j=1}^N TF(d_j, t)} \quad (2)$$

最後に , 計算した語の重みを用いて各 s-ブロックの項ベクトルを作成する . 例えば , 語の重み付け手法に WIDF を用いたときの変換事例の HTML 文書 (図 1) 中の s-ブロック “温泉案内” の項ベクトルは

$$(0.5, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$$

となる . テキストブロックの項ベクトルが , そのテキストブロックの意味的な特徴を表すと考える . 例として , 語の重み付け手法に WIDF を用いたときの図 1 と図 2 の事例解析結果を表 3 に示す .

5. 変換対象の解析

変換事例の解析の後 , 変換対象の HTML 文書の解析を行なう . 解析は以下の 2 つの作業からなる .

- (1) 変換対象の HTML 文書をテキストブロックに分割する . 変換対象の HTML 文書の各テキストブロックを t-ブロックと呼ぶ .
- (2) 事例解析と同様の方法で t-ブロックの構文木パス , 見出し , 項ベクトルを求める .

図 4 を変換対象としたときの解析結果を表 4 に示す .

6. XML 文書生成

変換事例 , 変換対象の解析結果を用いて , XML 文書生成する . 変換対象の HTML 文書は変換事例の HTML 文書と同様に変換する .

まず , 事例の XML 文書の DTD より XML 文書の骨組みを生成する . このとき x-タグが囲むテキストデータ欄は空欄にしておく . 例えば 図 2 を変換事例とすると , 図 2 上の DTD より図 6 の骨組みが生成される .

次に , 生成した XML 文書の空欄に最適な t-ブロックを割り当てる . x-タグ X の空欄には , 変換事例の XML 文書において X が囲んでいる s-ブロックに最も類似している t-ブロックを割り当てる .

2 つのテキストブロック間の類似度は , 意味的な類似度に構造的特徴を付加した以下の総合的類似度を用いる .

6.1 2 つのテキストブロック間の意味的類似度

2 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程 , 意味的な特徴が類似していると考ええる [徳永, 岩山 96] . 2 つのベクトル間の類似度は 2 つのベクトルのなす角の *cosine* に , ベクトルの大きさの比をかけたものを用いる . ベクトル V_i とベクトル V_j 間の類似度 $Sim(V_i, V_j)$ を以下のように定義する .

$$Sim(V_i, V_j) = \frac{V_i \bullet V_j}{|V_i| \times |V_j|} \times \frac{\min(|V_i|, |V_j|)}{\max(|V_i|, |V_j|)}$$

$Sim(V_i, V_j)$ が大きい程 , 2 つのベクトルは類似していると考ええる . ($0 \leq Sim(V_i, V_j) \leq 1$) .

```
<温泉案内>
  <温泉名>    </温泉名>
  <料金>      </料金>
  <営業時間>  </営業時間>
</温泉案内>
```

図 6 図 2 の XML 文書より生成した XML 文書の骨組み .

表 4 図 4 の HTML 文書を変換対象とした時の解析結果 .

t-ブロック	構文木パス	見出し	項ベクトル (WIDF)
温泉案内	<HTML><HEAD><TITLE>	なし	(0.5, 1, 0, 0, 0, 0, 0, 0, 0, 0)
笛吹川温泉	<HTML><BODY><H2>	なし	(0.5, 0, 0, 0, 0, 0, 0, 0, 0, 0)
料金	<HTML><BODY><H1>	笛吹川温泉	(0, 0, 0, 1, 0, 0, 0, 0, 0, 0)
600 円	<HTML><BODY>	料金	(0, 0, 0, 0, 0, 1, 0, 0, 0, 0)
営業時間	<HTML><BODY><H2>	なし	(0, 0, 0, 0, 0, 0, 1, 1, 0, 0)
10 時から 17 時	<HTML><BODY>	営業時間	(0, 0, 0, 0, 0, 0, 0, 0, 1, 1)

表 5 変換事例の XML 文書 (図 2) の x-タグが囲んでいる s-ブロックと、変換対象の HTML 文書 (図 4) の各 t-ブロックとの類似度 .

x-タグ \ t-ブロック	温泉案内	笛吹川温泉	料金	600 円	営業時間	10 時から 17 時
<温泉名>	0.17	1.17	0	0	0	0
<料金>	0	0	0	3.50	0	1.00
<営業時間>	0	0	0	1.00	0	3.17

6.2 二つのテキストブロックの総合的類似度

2 つのテキストブロックの構造的特徴である構文木パス、見出しが一致したならば、2 つのテキストブロックはより類似していると考えられる。従って、2 つのテキストブロック T_i, T_j 間の総合的類似度 $S(T_i, T_j)$ を以下のように定義する。

$$S(T_i, T_j) =$$

$$\begin{cases} Sim(V_i, V_j) + \alpha & \text{if 見出しが一致} \\ Sim(V_i, V_j) + \beta & \text{if 構文木パスが一致} \\ Sim(V_i, V_j) + \alpha + \beta & \text{if 両方が一致} \\ Sim(V_i, V_j) & \text{その他} \end{cases}$$

V_i, V_j は T_i, T_j の項ベクトル、 α, β は適当な定数である。 $S(T_i, T_j)$ が大きい程、二つのベクトルは類似していると考えられる ($\alpha + \beta \leq Sim(T_i, T_j) \leq 1 + \alpha + \beta$)。本論文では $\alpha = 2.0$, $\beta = 1.0$ とした。

図 1 と図 2 を事例、図 4 の文書を変換対象、重付けに WIDF としたときの、x-タグが囲む s-ブロックと変換対象の t-ブロックの類似度を表 5 に示す。

6.3 テキストブロックの割り当て

作成した XML の骨組み中の各 x-タグに t-ブロックを以下の手順で割り当てる。

- (1) 割り当て方の全ての組合せを考え、組合せ毎に x-タグが事例の XML 中で囲む s-ブロックと、t-ブロックの類似度の合計を計算する。
- (2) 計算した類似度の合計が最も高い割り当て方によって t-ブロックを割り当てる。

```
<!ELEMENT 温泉案内 (温泉名, 料金, 営業時間)>
<!ELEMENT 温泉名 (#PCDATA)>
<!ELEMENT 料金 (#PCDATA)>
<!ELEMENT 営業時間 (#PCDATA)>
<温泉案内>
  <温泉名>笛吹川温泉</温泉名>
  <料金>600 円</料金>
  <営業時間>10 時から 17 時</営業時間>
</温泉案内>
```

図 7 図 4 の HTML 文書の変換結果

- (3) この際、類似度が閾値よりも低い場合には「割り当てるべき適切な t-ブロックが無かった」と判定し、“該当なし”というデータを割り当てる。

この変換法で、事例として図 1 と図 2 を与えて、図 4 の HTML 文書を変換した XML 文書を図 7 に示す。

7. 実験的評価

7.1 語の重み付け手法の評価

語の重み付け法として TF・IDF と WIDF のどちらが本手法に適しているかを決定するために、2 つのシリーズ [HotSpa ツアー 99, JTB 99] の HTML 文書群の 10 個の HTML 文書に対して予備実験を行った。

重み付け手法による相違を調べるため、テキストブロック間の類似度には総合的な類似度 S ではなく意味的な類似度 Sim を用いた。評価の基準として正確度 (%) を

$$\text{正確度} = \frac{a}{b} \times 100$$

と定義する。ここで a は、生成した XML 文書における正しいテキストデータを持つ x-タグの総数である。一方 b は生成した XML 文書におけるテキストデータをもつ x-タグの総数である。

変換結果を表 6 に示す。表の TF・IDF と WIDF の各欄は正確度を示す。実験より WIDF の方が変換精度が高いことが示された。以下の実験では WIDF を採用する。

7.2 変換手法の評価

以下の 6 種類の HTML 文書群を用いて、事例に基づく変換手法の評価実験を行った。評価の基準は正確度を使用する。実験の結果を表 7 に、各 HTML 文書群の変換事例の解析データを表 8 に示す。

表 6 TF-IDF と WIDF の比較 .

シリーズ	文書数	TF・IDF	WIDF
[HotSpa ツアー 99]	4	85 (%)	89 (%)
[JTB 99]	6	25 (%)	37 (%)
平均		55 (%)	63 (%)

表 7 実験結果

実験番号	サンプル	文書数	平均文書 サイズ (byte)	平均タグ数 (種類数)	t-ブロックに 出現する 語の平均数	b の平均	a の平均	平均 精度
1	[HotSpa ツアー 99] 修正版	4	740	9 (6)	12.6	5.00	4.75	95
2	[YAHOO 99]	12	5992	162 (24)	26.0	3.00	3.00	100
3	[JTB 99] 修正版	4	4382	81 (18)	35.6	8.00	7.75	97
4	[山梨 99]	8	1085	51 (14)	6.00	5.00	4.50	90
5	[HotSpa ツアー 99]	5	5151	97 (20)	25.6	10.0	8.00	80
	[VECTOR 99]	11	7169	130 (18)	9.57	6.00	4.73	79
	[山梨大 99]	8	1788	27 (12)	26.7	3.00	2.13	71
6	[ACM 99] (in English)	20	5079	92(22)	2.00	5.00	4.00	80
	[JTB 00] (in English)	8	12446	205(21)	14.6	8.00	7.10	89
TOTAL		80						85

平均文書サイズ : HTML 文書サイズの平均、但し図は含まずテキストのみ
 平均タグ数 (種類数) : HTML 文書に出現するタグの数と種類の平均。
 b : 生成した XML 文書中のテキストデータを持つタグの総数。
 a : 生成した XML 文書中の正しいテキストデータを持つタグの総数。

表 8 実験 1~5 の変換事例の解析データ (実験 5 は 3 つの内一つのみ示す)。

実験番号	x-タグ	テキストブロック	名詞数	構文木パス	見出し
1	温泉名	白骨温泉	2	<html><body><h1>	なし
	住所	長野県南安曇郡安 ..	8	<html><body>	住所
	料金	5 0 0 円	2	<html><body>	料金
	営業時間	1 0 時から 5 時	4	<html><body>	営業時間
	概要	典型的な露天風呂 ..	47	<html><body>	概要
2	日付	2000 年 2 月 18 日 ..	11	<html><body>	なし
	タイトル	< 男児殺害 > 容疑 ..	13	<html><body><h2>	なし
	記事	— 京都市伏見区 ..	54	<html><body>	< 男児殺 ..
3	県名	秋田県	2	<html><body><center>...	県名
	自然と風土	東北地方を南北に ..	52	<html><body>	自然と風土
	気候と植生	県中央部を南北に ..	35	<html><body>	気候と植生
	地理的特徴	北部の県境に沿っ ..	40	<html><body>	地理的特徴
	人物・歴史	日本の歴史におい ..	39	<html><body>	人物・歴史
	文化・お祭	秋田県は古くから ..	38	<html><body>	文化・お祭
	産業	秋田といえば米ど ..	48	<html><body>	産業
	観光ワンポイント	澄んだ空気に包ま ..	32	<html><body>	観光ワン ..
4	タイトル	大菩薩峠	1	<html><body><H2>	なし
	概要	中里介山の長編小 ..	19	<html><body>	大菩薩峠
	車	中央自動車道勝沼 ..	4	<html><body><table>...	なし
	電車	JR 中央本線塩山駅	4	<html><body><table>...	なし
	バス	塩山駅より	2	<html><body><table>...	なし
5 (3 つの内の 1 つを示す)	タイトル	物語とコンピュータ	7	<html><title>	なし
	名前	電子情報工学科 ..	6	<html><body><center>...	なし
	概要	物語と言えばわ ..	67	<html><body>	なし

テキストブロック : x-タグが変換事例中で持つテキストブロック。
 名詞数 : テキストブロックに出現する名詞の総数。
 構文木パス : テキストブロックの変換事例の HTML 文書中の構文木パス。
 見出し : テキストブロックが変換事例の HTML 文書中で持つ見出し。

図 8 実験 3 で用いた HTML 文書。

図 9 実験 5 で用いた変換事例の HTML 文書。

§ 1 実験 1: 構造的特徴と意味的特徴双方が類似している場合

最も変換が容易であると思われる HTML 文書群として [HotSpa ツアー 99] のシリーズの修正版を用いた。意図的に、HTML 文書群が構文木パス、見出し、出現する語が類似するように修正した。よって、このシリーズは、意味的特徴と構造的特徴双方が類似したシリーズである。

§ 2 実験 2: 構文木パスの評価

構文木パスが変換にどの程度有効か評価するため、[YAHOO 99] のシリーズを用いた。このシリーズは、見出しと、各 HTML 文書に出現する語が類似せず、構文木パスのみが類似している。よって、変換には構文木パスのみが役に立つ。

§ 3 実験 3: 見出しの評価

見出しが変換にどの程度有効か評価するため、[JTB 99] のシリーズを修正したものを用いた。修正は、各 s-ブロックの見出しと、対応する t-ブロックの見出しの多くを一致させた。このシリーズは、文書中の各テキストブロックに出現する語と構文木パスが互いに類似しているため、テキストブロックの特徴として役に立たない。よって、変換には見出しのみが役に立つ。修正前の、実験に用いた HTML 文書の一つを、図 8 に示す。

§ 4 実験 4: 意味的特徴の評価

意味的特徴が変換にどの程度有効か評価するため、[山梨 99] のシリーズを用いた。このシリーズには見出しが存在しない。また、各テキストブロックの構文木パスが互いに類似しているため、テキストブロックの特徴として役に立たない。出現する語のみが類似している。よって、変換には意味的類似度のみが役に立つ例である。

§ 5 実験 5: 一般の HTML 文書群の変換

構造的、意味的特徴を考慮せず Web 上から任意に収集した 3 つの HTML 文書群 [HotSpa ツアー 99, VEC-TOR 99, 山梨大 99] について変換実験を行った。[HotSpa ツアー 99] は実験 1 と異なり、実際のデータをそのまま用いている。実験に用いた [HotSpa ツアー 99] の HTML 文書の一つを図 9 に示す。

§ 6 実験 6: 英語で記述された HTML 文書群の変換

実験 1 ~ 5 と同様の変換方法で、英語で記述された [ACM 99, JTB 00] のシリーズに対して変換を試みた。

7.3 考察

実験 1 ~ 4 はどれも 90 ~ 100% と高い精度を得た。構文木パス、見出し、意味的特徴の全てがそれぞれ変換に有効であることが示された。本手法は、三つの特徴のいずれかが類似するシリーズに対して、良い精度が得られることが示された。

実験 5 の一般の文書の実験でも [山梨大 99] を除いて約 80% のかなり高い精度を達成している。[山梨大 99] のシリーズは教官の研究紹介のシリーズであるが、内容は各教官に任されている。ブラウザ上での見た目は同じだが、各教官で異なるタグ付けが成されており、構造も異なっている。このような文書群は今回対象としたシリーズ型文書群とは言い難い。ただ、それでも語の数などを頼りに 70% 程の正確度を達成している。

また、実験 6 で平均 84% の正確度を得ていることから、提案手法は英語で記述された HTML 文書に対しても有効であると思われる。

x-タグへの t-ブロックの割当決定の際に、全数探索を

行って厳密に最適解を求めると多少計算に時間がかかる。本論文に掲載した実験データは全数探索を行って求めたものであるが、計算に要した時間の最大値は、実験番号 5 の [Hotspa ツアー 99] に対して、Pentium 500MHz 上で 10 分程度であった。しかし HTML 文書では各テキストブロックの語数は少なく、項ベクトルを用いた類似度の信頼性自体がそれほど高くない。そのため最適解を厳密に求めることはあまり意味が無いと思われる。貪欲算法等のヒューリスティックを用いて近似解を求めれば十分であり、殆んどの場合一致すると推測される。本実験データでも、貪欲算法を用いても殆んどの場合に最適解が求まり、計算時間もかなり短いことが確認されている。

8. 関連研究

WWW サイト上の現実の HTML 文書が持つシリーズ性に着目した研究は、筆者らの知る限りにおいて、他に例が無いと思われる。XML 言語の変換に関する既存の研究は、殆んどが XML から他言語への変換であり、逆方向に関する研究は殆んど無い。また HTML から XML への変換に関する研究も全く無いと思われる。これらは文書の意味構造の理解の難しさに根本的原因がある。

HTML 文書からの情報抽出という観点からは、ラッパー (wrapper) の自動生成や保守ならびに文書要約等の研究が、本研究との関連が深い。文献 [Ashish and Knoblock 97] では文字のサイズや強調、段付け等の簡単な構文的情報だけを用いて、意味を反映した文書の階層・フォーマット構造を認識するラッパーの構成法と、そのためのツールキットについて議論している。また [Kushmerick *et al.* 97] では、目的の Web ページに特化した (汎用ではない) ラッパーの帰納学習による自動生成を試み、併せて学習に必要なサンプル数の理論的解析を行っている。ここで生成するラッパーが取り扱える HTML 文書は、基本的に、データが 2 次元状に並んだ単純なもの (即ち、レコード類似データの並び) だけであり、能力的には不十分なものである。一方 [Cohen 99] では、構文木パス等の繰り返しを調べるヒューリスティックを考察し、併せて対象ページに特化したラッパーの正例のみからの自動合成問題への応用を試みている。Cohen の手法も、箇条書きやハイパリンクのリストのみを対象としており、扱える HTML 文書の範囲は非常に狭い。総じて、これらの研究で生成するラッパーは、かなり単純な HTML 文書しか取り扱うことができない。これは、個別の文書の意味情報の認識の難しさに起因しているものと思われる。これに比べて、本論文で提案した事例ベース変換法は、非常に広い範囲の HTML 文書を、かなり安定的に正しく XML 文書へ変換できる。事例によって文書の意味をシステムに自然に教え込む有用性が顕著に出ていると思われる。

データベースの分野も、Web ページ群を半構造化テキストデータベースと見なして研究が盛んに行われている

([Hsu 99] の参考文献を参照されたい)。そこでは、従来のデータモデルや質問言語の改良や質問処理算法の開発、並びにそれらの形式理論の構築が中心的研究課題となっている。文書の持つ意味的な側面に関しては殆んど研究されていない。また筆者らの知る限り、事例に基づく質問処理のような概念も提案されていない。

[品川他 00] では、WWW 閲覧支援を目的として、ユーザプロファイルに基づく動的ビューページ生成システムを作成している。HTML 文書の意味的な階層構造を把握するために、論理木という概念を定式化し、その有用性を実験的に示している。ビューの作成に事例情報は用いていないが、逆に、品川らの論理木という手法を本変換手法に組み入れることは有効であると考えられる。また本変換法は文書の要約問題と深く関わっている。[Tombros and Sanderson 98] では、検索結果のガイドを目的として、検索条件を考慮した文書の要約法、より具体的には、検索語に深く関わる文の抜き出し手法を議論している。また [Frank *et al.* 99] では領域依存に依存する重要句を、Bayes 学習により自動的に抜き出す手法を示している。どちらも HTML 文書を対象とした技術ではないが、本手法で行う事例と変換対象のテキストブロックのマッチング問題とも関連が深く、興味深い。

文献 [Kushmerick 99] ではラッパーの保守問題、即ち、動的に変化する Web ページに対して、過去に作成したラッパーを適用することの“是非”を判定する問題を考察している。これは一つの事例では表現しきれない複雑なシリーズ型 HTML 文書群の取り扱いに対して、一つの示唆を与えようと考えられる。

9. 結論と今後の課題

本研究では、シリーズ型 HTML 文書群に対して、事例に基づく XML 文書への変換手法を提案し、実験的評価を行った。その結果、実際の Web サイト上から収集した HTML 文書に対して、本手法は、80%を越えるかなり高い変換精度を持つことが確認された。HTML 文書を XML 文書に変換することにより、Web 上の HTML で記述された莫大な情報の機械処理が容易になる。またユーザの負担は、シリーズ型 HTML 文書が n 個ある場合、大雑把に言って、 $\frac{1}{n} + \alpha$ (但し、 α は変換誤りの検出訂正の手間) に軽減されるものと考えられる。これまでの経験からは、誤り検出は容易である。また変換精度が高く、誤っている場所自体が少ないことから、誤り修正も比較的容易であり、作業負担 α はかなり小さいと考えられる。

今後は、より大規模で多様なシリーズ型 HTML 文書群に対して、変換精度と安定性の更なる向上に努める予定である。現在の変換システムでは、見出しは `<h>` タグに限定している。しかし、見出しが `<h>` ではなく他の強調用タグで記述されている文書も多く、この場合、現行シ

ステムで 80%に近い精度を出すことは無理だと考える．部分的にでも他のタグで記述された見出しを検出することにより，変換精度はかなり向上 (or 安定) すると考えられる．今後システムに改良を加えて行きたい．

また表や箇条書を表す<table>，のようなタグは特殊であり，これに対して [Cohen 99] のような特殊技法を用いれば，更に変換精度は上げられると考えられる．また図やリンクを表すタグ，<a>も，その意味を考慮した特殊な手法が適用可能だと考えている．これらも今後の課題である．

複数の事例の利用も今後の重要な検討課題である．但し，複数事例を利用する場合，事例の一般化・融合や適用事例の切り分けなど検討すべき課題は数多い．また，不必要な多数の事例の作成を強要することは，ユーザにとって負担が大きいの．よって「如何なる場合に追加の事例が必要なのか？」また「追加事例は如何なるものが良いのか？」等々の判定・作成作業を支援するシステムがあれば望ましい．

また，この種の変換法には変換誤りは付き物である．宿命的なものであり，完全に無くすことは不可能と考える．よって誤り検出修正を支援するシステムがあれば非常に望ましい．誤りがあれば，追加事例を作成する必要も生じると予想されることから，事例作成支援と誤り修正支援は一体のものとして考える必要がある．今後，対話型変換支援エディタ等の開発を行って，この作業の支援方法を研究する予定である．

謝 辞

本研究は一部，通信・放送機構 (TAO) 及び電気通信普及団体 (TAF) の援助を受けている．

◇ 参 考 文 献 ◇

- [ACM 99] ACM Home Page: Calendar of Upcoming Events, <http://www.acm.org/events/coe.html> (1999)
- [Ashish and Knoblock 97] N.Ashish and C.A.Knoblock: Wrapper Generation for Semi-Structured Internet Source, *ACM SIGMOD Records*, Vol.26, No.4, pp.8-15 (1997)
- [Cohen 99] W.W.Cohen: Recognizing Structure in Web Pages using Similarity Queries, *Proc. of AAAI-99*, pp.59-66 (1999)
- [Frank et al. 99] E.Frank, G.W.Paynter, I.H.Witten, C.Gutwin and C.G.Nevill-Manning: Domain-Specific Keyphrase Extraction, *Proc. of IJCAI-99*, pp.668-673 (1999)
- [HotSpa ツアー 99] HotSpa ツアー, <http://www.iiijnet.or.jp/act/hotspa/asasina/index.html> (1999)
- [Hsu 99] P.Y.Hsu: Querying Web Pages with Lattice Expressions, *IEICE Trans. Info. & Syst.*, Vol.E82-D, No.1, pp.156-164 (1999)
- [JTB 99] JTB Travel Japan Site, <http://www.jtb.co.jp/TJsite/library/library.html> (1999)
- [JTB 00] JTB: Guides Separated for Each Prefecture, Travel Japan Site, <http://www.jtb.co.jp/TJsite/library/library.e.html> (2000)
- [Knoblock et al. 98] C.A.Knoblock, S.Minton, J.L.Ambite, N.Ashish, P.J.Moer, I.Muslea, A.G.Philpot and S.Tejada: Modeling Web Sources for Information Integration, *Proc. of AAAI-98*, pp.211-218 (1998).
- [Kushmerick et al. 97] N.Kushmerick, D.S.Weld and B.Doorenbos: Wrapper Induction for Information Extraction, *Proc. of IJCAI-97*, pp.729-735 (1997)
- [Kushmerick 99] N.Kushmerick: Regression testing for wrapper maintenance, *Proc. of AAAI-99*, pp.74-79 (1999)
- [松本他 99] 松本, 北内, 山下, 平野: 日本語形態素解析システム「茶筌」version 2.0 使用説明書, 奈良先端科学技術大学院大学松元研究所 (1999)
- [村田他 98] 村田, 門馬, 荒井: XML 入門, 日本経済新聞社 (1998)
- [長尾 96] 長尾真: 岩波講座ソフトウェア科学 15 自然言語処理, 岩波書店 (1996)
- [品川他 00] 品川, 北川, 川田: ユーザプロファイルに基づくユーザページの動的生成による WWW 閲覧支援, 情報処理学会論文誌, Vol.41, No.SIG 6(TOD 7) (2000)
- [徳永, 岩山 96] 徳永, 岩山: 重み付き IDF を用いた文書の自動分類について, 人工知能学会自然言語処理研究会, 100-6 (1994)
- [徳永 99] 徳永健伸: 言語と計算 5 情報検索と言語処理, 東京大学出版会 (1999)
- [Tombros and Sanderson 98] A.Tombros and N.Sanderson: Advantage of Query Biased Summarization in Information Retrieval, *Proc. of ACM SIGIR'98*, pp. 2-10 (1998)
- [VECTOR 99] VECTOR <http://www.vector.co.jp> (1999)
- [XML/SGML サロン 98] XML/SGML サロン: 標準 XML 完全解説, 技術評論社 (1998)
- [YAHOO 99] YAHOO! JAPAN NEWS-トピックス (毎日新聞) <http://www.yahoo.co.jp/headlines/top/> (1999)
- [山梨 99] 山梨情報の森: 観光, <http://www.pref.yamanashi.jp/shouko/index-j.html> (1999)
- [山梨大 99] 山梨大学 工学部 教官研究 <http://www-eng.yamanashi.ac.jp/> (1999)

〔担当委員：外山勝彦〕

2000 年 9 月 21 日 受理

著 者 紹 介

梅原 雅之 (学生会員)

2000 年山梨大学工学部電子情報工学科卒業．現在同大学大学院工学研究科博士前期課程在学中，WEB データからのデータマイニングの研究に興味を持つ．

岩沼 宏治 (正会員)

1983 年東北大学通信工学科卒業．1985 年同大学院修士課程了．同年山形大学情報工学科助手．1992 年より山梨大学コンピュータメディア工学科助教授，博士 (工学)．人工知能とソフトウェアの基礎，特に様相論理，非単調論理，定理自動証明プログラム，ならびに遺伝的アルゴリズムの研究と開発に従事．1987,89,90,91 年人工知能学会全国大会優秀論文賞受賞．

永井 宏和

1997 年 3 月山梨大学電子情報工学科卒業．2000 年 3 月山梨大学大学院卒業．同年 4 月東芝 FA システムエンジニアリング (株) に入社．在学中は定理自動証明プログラム，HTML テキスト処理の研究，開発に従事