携帯端末によるテーブル認識変換システムの構築と評価

安富 大輔† 増田 英孝† 中川 裕志‡
東京電機大学工学部† 東京大学情報基盤センター‡

1 はじめに

近年では、iモードやPDAなどの携帯端末からWebページをブラウズしたいという要求が急激に増加している。しかし現状では、解像度が大きいPCを対象としたページが大多数である。そのため、iモードやPDAなどの解像度が低く、画面の小さい携帯端末では、表示領域の問題、マークアップ言語や画像フォーマットなどが携帯端末によって異なること、さらに <TABLE> タグの取り扱い方が異なるテーブル表示の問題、データ容量の問題、操作の煩雑さなどの問題が発生する。

そこで、本研究はPC向けに作成されたWebページを閲覧するためにテーブル表示の問題に主眼を置き、テーブルを携帯端末に適した形に自動変換するシステムを構築した[1]。

現システムでは、変換対象テーブルを3つのテーブル型のいずれかに判別し、<TABLE> タグを使用しないテキストとして変換表示するが、テーブル型の認識を誤る可能性がある。本稿では、認識精度向上のためにベクトル空間法を用いたテーブル型自動認識について述べる。

2 自動変換システム

現システムでは、テーブル型を縦一覧型 (lengthways)、横一覧型 (sideways)、時間割型 (timetable) に分類し、その型に合わせて属性名部を属性値部に付与したテキストとして表示する。図1は携帯端末のブラウザであるPalmscape[2]で、テーブルページを表示したものである。また、図2は同一のテーブルページを現システムで変換したものである。

現システムで扱うテーブル型の認識では、表から属性名部の列数、行数を求めてテーブル型を判別する。一般的なテーブルの例を、以下の図3に示す。

ここでNumberOfRowAttribute,NumberOfColAttributeは、属性名部が表中で占める行数、列数を示し

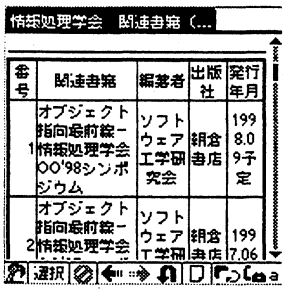

図 1: 携帯端末でのテーブルページ表示

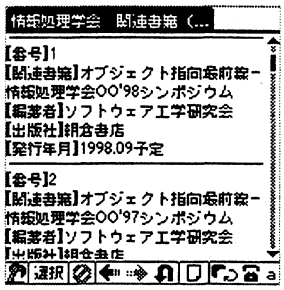

図 2: 現システムを用いた変換結果

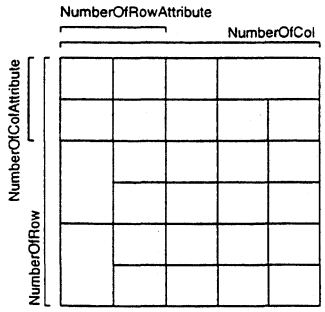

図 3: テーブル認識で得るデータ

ており、NumberOfRow,NumberOfColは、テーブル全体の行数、列数を示している。

3 認識アルゴリズム

認識アルゴリズムは、テーブルデータからまずNumberOfRowAttribute,NumberOfColAttributeの値を算出する。もし、NumberOfRowAttribute, NumberOfColAttributeがどちらも0の場合には、タグの情報を利用した構造からの認識を行う。認識アルゴリズムを、図4に示す。

†Daisuke YASUTOMI, †Hidetaka MASUDA, ‡Hiroshi NAKAGWA
†Department of Electrical Engineering Tokyo Denki University, ‡Information Technology Center The University of Tokyo

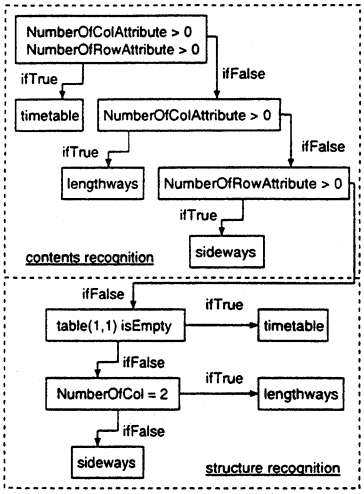

図 4: 現システムの認識アルゴリズム

4 テーブル判別の誤り

現システムでの自動認識アルゴリズムでは、変換精度が約 62%となっている。テーブル判別の誤りとしては、時間割型の認識では colspan,rowspan のタグオプションが多用されているため、NumberOfRowAttribute, NumberOfColAttribute の算出を誤る可能性がある。また、属性値部も属性名部と同様に文字数の少ない簡略化された文字列が使われていることがあるため、属性名部の判別を誤る可能性がある。

さらに、現システムでは意味認識と構造認識を別々に処理しているため、これらの情報を合わせ定式化したものを、認識アルゴリズムとして実装する必要がある。

そこで、本研究ではテーブル認識精度を向上させるために、テーブルの各セルの意味内容をベクトルで表し、ベクトル空間法を適用した認識アルゴリズムを提案する。

5 ベクトル空間法を用いた認識

5.1 ベクトルの要素

テーブルの各セルの言語的性質 x_i に対応して、その性質をもてば 1、もたなければ 0 と値 w_i を定義する。w_i を要素とするベクトルを式 (1) のように定義する。

$$\overrightarrow{Cell_{ij}} = (w_1, w_2, \cdots, w_n) \qquad (1)$$

- 連続データ
  行あるいは列を基準として、“1”,“2”,“3” などのある決まった連続性を持ったデータ列や、“りんご”、“みかん”、“バナナ” など、“果物” として同一カテゴリに含まれるそれぞれのデータ群を、ひとつのベクトルの次元として定義した。

- 句読点

- 文字長が短い
  属性名部は文字長が短いことが多い。文字長が 0(空白)、半角 10 文字以内、半角 11 文字以上をそれぞれベクトルの次元とした。

- 接頭辞 [3]
  “第”,“平成”,“特” など 14 種の接頭辞の各々にベクトルの次元を割り当てる。

- 接尾辞 [3]
  “日”,“課”,“年” など 43 種の接尾辞の各々にベクトルの次元を割り当てる。

- 単位
  “kg”,“人”,“円” など 17 種の単位の各々にベクトルの次元を割り当てる。

- 特殊文字
  属性値データとして、一定の期間を表している “～” や、備考などを示す “(”,“)” などが使われることが多いことから、それら 11 種の各々にベクトルの次元を割り当てる。

- テーブルタグ

  一般的に colspan,rowspan のセル内あるいは、colspan のセルの直下、rowspan のセルの直後のテーブルデータは属性名部となることが多い。

  あるテーブルデータが、colspan あるいは rowspan の構造中に存在するか、あるいは colspan のセルの直下、rowspan の直後のセルであれば、各々をベクトルの次元に割り当てる。

  テーブルタグは構造の決定に重要な役割を持っているため、1 以上の値をとる場合もある。

  これにより、colspan あるいは rowspan に関係するテーブルデータと、そうでないテーブルデータとの距離を離すことができる。

5.2 ベクトルの計算

ひとつのテーブルデータの類似度を決定するために、行あるいは列を基準としてベクトル間の距離を算出する。テーブルデータの類似度 $Sim(cell_{ij})$ を次式で定義する。

$$Sim_{col}(cell_{ij}) = \frac{\sum_{k=1}^{c} \frac{\overrightarrow{cell_{ij}} \cdot \overrightarrow{cell_{ik}}}{|\overrightarrow{cell_{ij}}||\overrightarrow{cell_{ik}}|} - 1}{c} \quad (2)$$

$$Sim_{row}(cell_{ij}) = \frac{\sum_{k=1}^{r} \frac{\overrightarrow{cell_{ij}} \cdot \overrightarrow{cell_{kj}}}{|\overrightarrow{cell_{ij}}||\overrightarrow{cell_{kj}}|} - 1}{r} \quad (3)$$

$$c = NumberOfCol, r = NumberOfRow$$

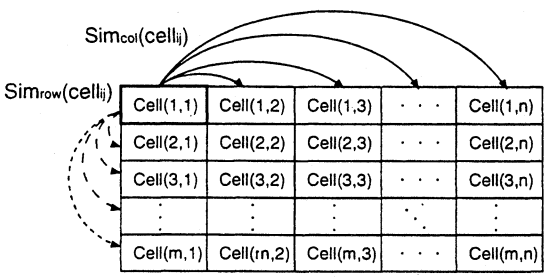

図 5: テーブルデータの比較方法

式 (2) は列を基準として、列のあるひとつのテーブルデータと、他のテーブルデータとの類似度を算出する式である (図 5)。同様に式 (3) は、行を基準としたテーブルデータの類似度を算出する式である (図 5)。

5.1 で定義したベクトルを用いて、ベクトル間の距離を計算した結果の例を示す。図 6,7 は、NumberOfColAttribute=1, NumberOfRowAttribute=0 である一覧型のテーブルに対して、各々の行と列を基準にしたテーブルデータの類似度を算出した結果である。NumberOfColAttribute=2,NumberOfRowAttribute=2 の時間割型の別のテーブルに対する計算結果を図 8,9 に示す。

5.3 属性名部の認識

計算結果より、一般的に属性名部は属性値部と比較して類似度が低くなる傾向があることがわかった。この結果を利用して属性名部と属性値部の切り分けを行う。

行を基準とした場合、まず 1 行目のテーブルデータの類似度の平均を $\overline{X_r}$ とし、2 行目以降のテーブルデータの類似度の平均を $\overline{Y_r}$ とする。また、列を基準として 1 列目のテーブルデータの類似度の平均を $\overline{X_c}$ とし、2 列目以降のテーブルデータの類似度の平均を $\overline{Y_c}$ とする (図 10,11)。

0.05	0.69	0.50
0.43	0.77	0.61
0.52	0.77	0.53
0.52	0.77	0.61
0.52	0.77	0.59

図 6: 行基準 (一覧型)

0.31	0.31	0.19
0.27	0.21	0.18
0.35	0.28	0.40
0.35	0.28	0.16
0.28	0.25	0.28

図 7: 列基準 (一覧型)

0.11	0.11	0.54	0.54	0.56	0.56	0.56	0.35
0.11	0.11	0.69	0.81	0.84	0.84	0.84	0.51
0.66	0.66	0.82	0.81	0.84	0.84	0.84	0.51
0.66	0.66	0.82	0.81	0.84	0.84	0.84	0.51
0.66	0.66	0.82	0.81	0.84	0.84	0.84	0.51
0.66	0.66	0.82	0.81	0.84	0.84	0.84	0.51
0.66	0.66	0.82	0.81	0.84	0.84	0.84	0.22
0.66	0.66	0.82	0.81	0.84	0.84	0.84	0.22
0.66	0.66	0.82	0.59	0.84	0.84	0.84	0.22

図 8: 行基準 (時間割型)

0.12	0.12	0.62	0.62	0.62	0.62	0.62	0.62
0.12	0.12	0.51	0.60	0.60	0.60	0.60	0.60
0.81	0.43	0.81	0.81	0.81	0.81	0.81	0.81
0.81	0.43	0.81	0.81	0.81	0.81	0.81	0.81
0.81	0.43	0.81	0.81	0.81	0.81	0.81	0.81
0.81	0.43	0.81	0.81	0.81	0.81	0.81	0.81
0.68	0.37	0.68	0.68	0.68	0.68	0.68	0
0.68	0.37	0.68	0.68	0.68	0.68	0.68	0
0.65	0.40	0.65	0.53	0.65	0.65	0.65	0

図 9: 列基準 (時間割型)

式 (4),(5) を用いて、行あるいは列が属性名部であるかどうかを判別する。このとき、定数 k_r,k_c は経験的に求めた値とする。

$$\overline{X_r} < k_r \times \overline{Y_r} \quad (k_r < 1) \quad (4)$$

$$\overline{X_c} < k_c \times \overline{Y_c} \quad (k_c < 1) \quad (5)$$

式 (4) が真であるならば、$\overline{X_r}$ である列は属性名部となり、式 (5) が真であるならば、$\overline{X_c}$ である行は属性名部となる。1 列目あるいは 1 行目が属性名部と判別された場合、順次、次の列あるいは行を評価し、NumberOfColAttribute, NumberOfRowAttribute を算出する。

$Sim_{row}(Cell_{11})$	$Sim_{row}(Cell_{12})$	$Sim_{row}(Cell_{13})$	$\overline{X_r}$
$Sim_{row}(Cell_{21})$	$Sim_{row}(Cell_{22})$	$Sim_{row}(Cell_{23})$	
$Sim_{row}(Cell_{31})$	$Sim_{row}(Cell_{32})$	$Sim_{row}(Cell_{33})$	
$Sim_{row}(Cell_{41})$	$Sim_{row}(Cell_{42})$	$Sim_{row}(Cell_{43})$	$\overline{Y_r}$
$Sim_{row}(Cell_{51})$	$Sim_{row}(Cell_{52})$	$Sim_{row}(Cell_{53})$	

図 10: NumberOfColAttribute の計算

$Sim_{col}(Cell_{11})$	$Sim_{col}(Cell_{12})$	$Sim_{col}(Cell_{13})$
$Sim_{col}(Cell_{21})$	$Sim_{col}(Cell_{22})$	$Sim_{col}(Cell_{23})$
$Sim_{col}(Cell_{31})$	$Sim_{col}(Cell_{32})$	$Sim_{col}(Cell_{33})$
$Sim_{col}(Cell_{41})$	$Sim_{col}(Cell_{42})$	$Sim_{col}(Cell_{43})$
$Sim_{col}(Cell_{51})$	$Sim_{col}(Cell_{52})$	$Sim_{col}(Cell_{53})$
$\overline{X_c}$		$\overline{Y_c}$

図 11: NumberOfRowAttribute の計算

5.4 属性名部の算出

式 (4),(5) を使用して、テーブルの NumberOfColAttribute, NumberRowAttribute を算出し、それぞれのテーブル型を判別する。ここでは、$k_c = 0.85, k_r = 0.85$ とした。

5.4.1 縦一覧型

まず、1 列目が属性名部であるかどうかを判別する。図 6 から $X_r \fallingdotseq 0.41, Y_r \fallingdotseq 0.61$ となり式 (4) が真になることから、1 列目は属性名部と決定される。さらに、2 列目が属性名部であるかどうかを判別することになるが式 (4) が偽となるため、属性名部ではないと決定される。

結果として、NumberOfColAttribute=1 となる。一方、行を基準とした場合は NumberOfRowAttribute=0 となるためテーブル認識の結果、一覧型として認識される。

5.4.2 時間割型

時間割型も縦一覧型と同様に、列または行を基準にして属性名部と属性値の切り分けを行う。この結果 NumberOfColAttribute=2,NumberOfRowAttribute=2 となり時間割型と認識される。

5.4.3 横一覧型

横一覧型は縦一覧型を転置した形であり、NumberOfRowAttribute>0 である場合、横一覧型として認識される。

以前行ったテーブル調査から、一般的に横一覧型は 2 列である場合が多い。2 列のテーブルではテーブルデータ間の距離は同一となることは明らかであり、類似度を決定することができない。そこで、NumberOfColAttribute = 0,NumberOfRowAttribute = 0 でありかつ 2 列のテーブルである場合は、横一覧型として認識する。

6 まとめ

本研究では、携帯端末でのテーブル表示に対する問題として携帯端末に適した形にテーブル自動変換するシステムを構築した。システムのテーブル型の認識精度を向上させるため、ベクトル空間法を利用したテーブル認識アルゴリズムを提案した。今後は、ベクトルの各次元と各係数の最適化を行い、認識精度の評価を行う予定である。

参考文献

[1] Hidetaka MASUDA, Daisuke YASUTOMI, and Hiroshi NAKAGAWA:
"How to Transform Tables in HTML for Displaying on Mobile Terminals", NLPRS2001 Workshop, pp.29～36(2001).

[2] 株式会社イリンクス:Palmscape3.1,
http://www.ilinx.co.jp/

[3] 塚本修一, 安富大輔, 増田英孝, 中川裕志:
"HTML 文書における表の携帯端末のための構造変換", 第 64 回情報処理学会全国大会 2Y-06,(2002).