HTML 文書における表の携帯端末のための構造変換 *

2Y−02

塚本 修一†, 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、PDA、携帯電話等の携帯端末から Web ブラウズするユーザが急増している。しかし、既存の HTML 文書の大多数が、解像度の高い PC 用に作られている。そのため、これらの Web ページを携帯端末でブラウズすると、1 ページが 1 画面に表示しきれなくなり、単語の途中の折り返しが頻繁に発生し、端末の操作回数が増加する。特に、情報を見やすくなるために使用する表を、携帯端末で表示すると、逆に可読性が低下する。

本研究は、表が含まれている HTML 文書を携帯端末で Web ブラウズすることに注目し、上記の問題を解決するために、表の内容を把握しやすくするための変換システムを提案する [1]。本稿では、HTML 文書の表を、携帯端末で表示する際の問題点、変換の方針、表の調査と検証について述べる。

2 表の自動変換システム

携帯端末で表を表示することは、操作の繁雑さ、可読性の低下、読み誤りといった問題が発生する。そこで、表を携帯端末に適した形に変換するシステムを実装した [1]。表には、属性名と属性値の部分が存在し [2]、それをペアとして表示することによって問題を解消している。図 1 に、携帯端末 (アプリケーションは palmscape[3]) で表を表示したもの、また図 2 に、現システムで変換した結果を示す。

3 変換の問題点

現システムでは表の構造を基に属性名と属性値の判別を行っている為、属性名と属性値の判別を誤ってしまう可能性がある。このため、表の型及び正しく属性名と属性値を区別するために、Web 上に実在する表を調査する必要がある。表の中のデータから、属性名となり得る特徴を調査した。

*Table Transformation in HTMLdocument for Mobile Terminals
†Shuichi TSUKAMOTO, †Daisuke YASUTOMI, †Hidetaka MASUDA, and ‡Hiroshi NAKAGAWA
†Department of Electrical Engineering Tokyo Denki University, ‡Information Technology Center The University of Tokyo

図 1: 情報処理学会のページ

図 2: 変換結果

4 属性名の内容調査

Web ページに実在する表を収集し、502 個の表から、属性名を調査した。1 行目、1 列目には、属性名が入ることが多い [1] ため、これらのセルデータを収集した。

4.1 属性名

以下に 1 行目、1 列目のセルの中身の例を示す。カンマ句切りで、セル 1 つ分のデータを示す。

A. 属性名の候補

(i) 東京都, 茨城県, 神奈川県, 千葉県

(ii) 横綱, 大関, 関脇, 小結,…, 序の口

(iii) 第 1 回, 第 2 回, 第 3 回, 第 4 回, 第 5 回、…

B. 属性名の候補とならないもの

(i) 以下の内容をまとめてダウンロードできます。,1,000〜4,999 人, 中堅企業,——

(ii) ○, △, × (各々単独で出現した場合)

ここで、セルデータを式 (1) の様に定義する。

$$Cell = Head + Core + Tail \quad (1)$$

$Cell$ は 1 セルのデータ全体、$Head$(H と略) は他のセルと共通する先頭からの文字列、$Core$(C と略) は他のセルに比べて変化している文字列、$Tail$(T と略) は他のセルと共通する末尾からの文字列と定義する。本研究では、式 (1) の H を接頭辞、C をグループに属する文字列、T を接尾辞と呼ぶ。

4.1.1 接頭辞、接尾辞

502 個の収集したセルデータから H、T を持つデータ 437 個を抽出した。以下に例の一部を出現頻度、出現種類と共に示す。

- $H+C$ 型 79 回 37 種類

  第$+C$, 貸付$+C$, 特$+C$, 記号 (★、■、○、)$+C$, …

- $C+T$ 型 93 回 33 種類

  C + 郵便局, C + 白書, C + 大学, C + 月分, C + 会, …

- $H+C+T$ 型 265 回 80 種類

  (社)$+C$+雇用開発協会, 第$+C$+回, 平成 11 年$+$ C + 月分, …

以上の 3 つのケースが存在する。

4.1.2 グループ

同様に、同一グループに属する属性値名の候補となる文字列の例を以下に示す。

- C 型

  1, 2, 3, 4, 5, …

  マグロ, サンマ, サバ, イワシ, …

  鉄分, ビタミン, カルシウム, …

これらは、数字、魚の名前、栄養素等のクラスに分けられる。そして、4.1 の相撲の番付の例や数字等の文字列群の様に順序があるものは、上からまたは左から順番に並んでいることが多い。

4.2 類似性

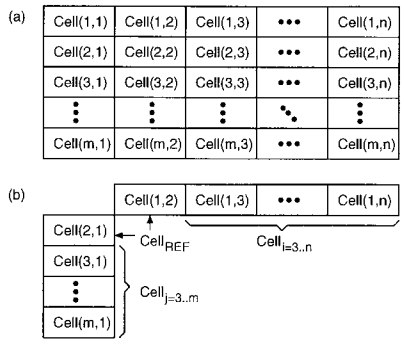

図 3: 表の一般的な形

図 3(a) に表の一般形を示す。図 3(b) に示すように、$Cell(1,1)$ を除いた第 1 行および第 1 列を抽出する。最も左または上のセルを基準 $Cell_{REF}$ として類似度を式 (2) として定義する。

$Sim(Cell_{REF}, Cell_i)$ = 次の 3 つの論理式のうち true となるものの数

$$(H_{REF} = H_i) \vee (T_{REF} = T_i) \vee (C_{REF} \wedge C_i \in G) \Rightarrow Sim(CELL_{REF}, CELL_i) \quad (2)$$

したがって、$Sim(CELL_{REF}, CELL_i) = 0$ は類似性がないことを表し、1 以上の場合は、大きい値ほど高い類似性を示す。

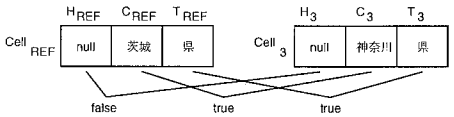

図 4: 4.1 節 A(i) の例

図 4 に 4.1 節 A(i) の例を示す。

5 まとめ

属性名の検証より、2 つの結果が得られた。1 つ目は、各行、各列のセルの類似性が高いときには、属性名の候補となる可能性がある。$H+C+T$ 型、$H+C$ 型、$C+T$ 型の場合がこれに相当する。2 つ目として C 型の場合は、C のセルデータ間に連続性があることで、属性名の候補となる可能性がある。しかし、C 型は連続性がないと属性値となることもある。今後は、属性名と属性値を認識し、表の変換の誤認識を解消するために、現システムに整理した接頭辞、接尾辞、言葉の連続性の情報を使ったルーチンを組み込む予定である。

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] 表形式からの情報抽出手法:吉田稔, 鳥澤健太郎, 辻井潤一, 言語処理学会第 6 回年次大会, pp.252〜255(2000).

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