E-025

HTMLの表の構造認識アルゴリズムの提案

塚本 修一† 増田 英孝†† 中川 裕志‡
†東京電機大学大学院工学研究科 ††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

Web ページ中の HTML 文書には、さまざまな情報が表形式で配信されている。HTML 表形式のデータを計算機を用いて自動的に処理する場合には、表の項目名の部分と項目データの部分を識別する必要がある。表の構造認識が可能になると、表からのデータ抽出や、動的な表の出力インターフェースを生成するために利用できる。

これまでに、表に関する研究として、テキスト中の整形された表から構造を認識する研究 [1] 等がある。本研究では、HTML の表から情報を抽出する基礎技術として、表形式データの構造を認識するアルゴリズムを提案する。また、そのアルゴリズムを実装し、評価を行った。多くの研究では対象としていない、複数行、列に渡って項目名持つ表も評価対象としている。

2 表の構造認識アルゴリズム

提案するアルゴリズムは表の項目名の部分と項目データ部分の境界を認識するアルゴリズムである。本研究ではセル間の類似度をベクトル空間法によって計算し、類似度の比を用いて、行と列の項目名と項目データを計算し区別する。

2.1 アルゴリズムの概要

例として行方向での項目名を特定するアルゴリズムの説明する。まず、HTML の表の正規化 [2] を行う。そして、表の各セルデータに対して言語的性質を適用したベクトルのマトリクスを作成する。次に各セルの列方向の類似度をベクトル空間法によって算出する。算出したセルの類似度に対して行ごとの平均値を求め、行の類似度が低くなる部分には、行間に内容的な切れ目があると認識する。行と列を入れ換えて類似度を計算すれば、列間の切れ目を認識することができる。

2.2 ベクトルの要素とベクトルの計算

表の i 行 j 列のセルを $Cell_{ij}$ として、各セルの N 個の言語的性質 $k = 1, \ldots, N$ に対応して、その性質を持てば 1、持たなければ 0 と値 w_k を定義する。w_k を要素とするベクトルを式 (1) のように定義する。

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

以下にベクトルの要素となる言語的性質を列挙する。今回の実験では N は合計で 106 であり、内容は以下に示す。

- 連続する数値データ (1 次元)
- 句読点 (6 次元)
- 文字長 (3 次元)
- 接頭辞 [2](16 次元)
- 接尾辞 [2] (45 次元)
- 単位 (16 次元)
- 特殊文字 (12 次元)
- 文字種 (5 次元)
- テーブルタグの属性 (2 次元)

2.3 ベクトルの計算

例として項目名の行と項目データの行の境界を求める方法を述べる。m 行 n 列の表の行間の類似度を計算するために、まず表の i 行 j 列のセルを $Cell_{ij}$ として表し、同じ列の $Cell_{kj}(k \neq i)$ との類似度の平均 $Sim_{row}(i, j)$ を次式で求める。

$$Sim_{row}(i,j) = \frac{1}{m-1} \sum \frac{\overrightarrow{cell_{ij}} \cdot \overrightarrow{cell_{kj}}}{\left|\overrightarrow{cell_{ij}}\right| \left|\overrightarrow{cell_{kj}}\right|} \qquad (2)$$

図 1 を使い $Sim_{row}(1,1)$ の求め方を解説する。式 (1) により求めた $Cell_{1,1}$ を基準とし、同じ列の $Cell_{1,1}$ 以外の $Cell_{2\cdots m,1}$ との $cosine$ 値を求め、求めた $cosine$ 値の平均値を類似度とする。その結果が図 2(a) の $Sim_{row}(1,1)$ である。この方法で $Cell_{m,n}$ までの類似度を求め、図 2(a) のようなマトリクスを作る。次

図 1: $Sim_{row}(1,1)$ の計算

に、行の切れ目を特定するために、図 2(b) のように、式 (3) を用いて行の類似度の平均を求める。

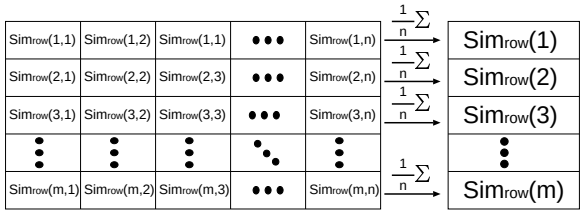

図 2: 式 (3) の計算結果

$$Sim_{row}(i) = \frac{1}{n}\sum_{k=1}^{n} Sim_{row(i,k)} \quad (3)$$

次に、類似度の比 ($Sim_{row}(i)$ と $Sim_{row}(i+1), \cdots, Sim_{row}(m)$ の平均の比) $R(i)$ を式 (4) で定義する。

$$R(i) = \frac{Sim_{row}(i)}{\frac{1}{m-i}\sum_{k=i+1}^{m} Sim_{row}(k)} \quad (4)$$

式 (4) より求める値 $R(i)$ より、項目名と項目データの行の境界 T を次のアルゴリズムで求める。但し、θ は、境界かどうかを判定する閾値である。

```
T = 0;
for(i=1;i<=m;i++){
    if(R(i)< θ){ T = i; }
    else{ break; }
}
if(T==0){ 縦方向に境界なし }
else{T 行までが項目名の行 }
```

以上は項目名の行と項目データの行の境界を求めるアルゴリズムだが、以上の導出において、縦横を交換すれば、$Sim_{col}(j)$ を計算でき、そして項目名の列と項目データの列の境界を認識できる。

2.4 認識アルゴリズムの評価実験

Web ページの中からランダムに抽出した表を 300 用意した。まず、最適な閾値 θ を求めるための教師データとして、この 300 の表を人手によって項目名と項目データの行 (あるいは列) の境界を決めた。この教師データによって、θ の最適化を含め 10 fold 交差検定を用いて評価した。その結果、行、列の閾値 θ はそれぞれ 0.90、0.70 となった。システムは行方向に 82%、列方向に 78% の正解率で表の項目名を認識することができた [2]。行方向の評価の結果を表 1、列方向の結果を表 2 に示す。 また、評価を行った表の大きさの平

表 1: 行方向の結果

データの種類	正解率
トレーニングデータ	83.23%
テストデータ	82.11%

表 2: 列行方向の結果

データの種類	正解率
トレーニングデータ	79.11%
テストデータ	78.11%

表 3: 交差検定によるテストデータとして評価をした 300 表の内訳

切れ目	行	列
0	70	183
1	202	115
2	25	2
3	3	0
合計	300	300

均は 9.2 行 6.3 列であり、それぞれ内訳を表 3 に示す。この表 3 の結果の内、行と列ともに項目名を持つ表は 66 個あり、評価に用いた表の分布の行方向の $R(i)$ の分布を図 3 に示す。図 3 より、切れ目のない表の $R(i)$ は一定となり、切れ目がある表の $R(i)$ は切れる行数までの値が低くなる傾向がある。次に各ベクトルの要素が正解率に与える影響を相対的に評価する。各々の

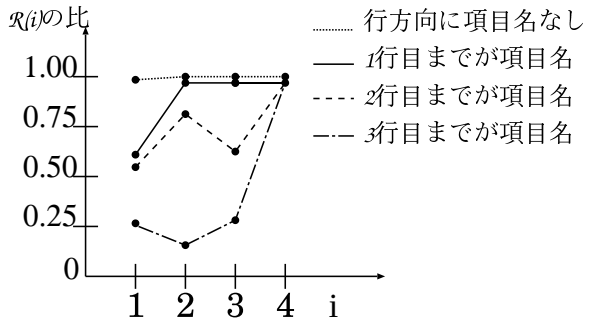

図 3: 行方向での $R(i)$ の分布

ベクトル要素がどの程度認識に影響を与えているのかを調査した。言語的性質の各カテゴリ毎にベクトルの要素群を無効にして正解率の変化を調査した。カテゴリは、1. 文字種,2. 接頭辞 · 接尾辞,3. 句読点 · 単位,4. 特殊文字,5. 文字長とした。その結果を図 4 に示す。この結果から、正

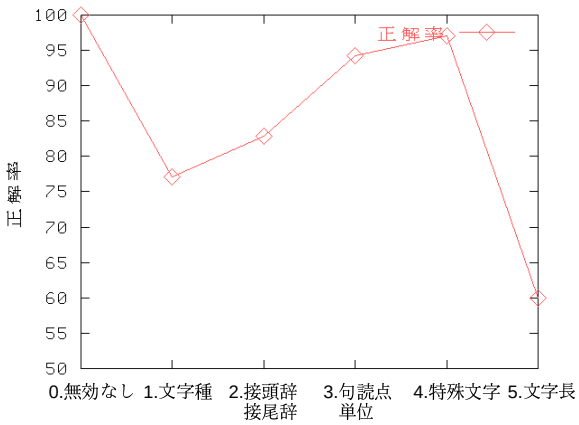

図 4: カテゴリの影響

解率に影響が大きいカテゴリは文字長であることが分かった。続いて文字種、接頭辞 · 接尾辞、句読点 · 単位、特殊文字の順となる。

3 まとめ

本稿では表形式データの構造、を認識するアルゴリズムについて述べた。提案したアルゴリズムを適用したシステムはおよそ 80%の正解率で項目名と項目データを認識することができる。今後の課題はベクトルの各要素の重みを相対的な評価から決定し、認識率の向上を目指す。

参考文献

[1] HURST, M. and DUGLAS, S.: Layout and Language: Preliminary Experiments in Assingning Logical Structure to Table Cells, *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pp. 217–220 (1997).

[2] 塚本修一, 増田英孝, 中川裕志: HTML 表データの構造認識システムとその評価, 言語処理学会第 9 回年次大会論文集, pp. 81–84 (2003).