# WWW 検索精度向上の為の HTML 文書中の表構造解析

山本 晋太郎† 松本章代†† 小西 達裕‡

高木朗§ 小山照夫‖ 三宅芳雄¶ 伊東 幸宏‡‡

†静岡大学大学院情報学研究科 〒432-8011 静岡県浜松市中区城北 3-5-1

††青山学院大学理工学部 〒229-8558 神奈川県相模原市淵野辺 5-10-1

§言語情報処理研究所 〒184-0014 東京都小金井市貫井南町 3-6-30

‖国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2

¶中京大学 〒470-0393 愛知県豊田市貝津町床立 101

‡静岡大学情報学部 〒432-8011 静岡県浜松市中区城北 3-5-1

‡‡静岡大学創造科学技術大学院 〒432-8011 静岡県浜松市中区城北 3-5-1

E-mail: riir@inf.shizuoka.ac.jp

**あらまし** HTML 文書中には多くの表構造が出現するが，表のシンタックスを明示する記述や，見出しとなるセルと見出し以外の内容セルとの区別を明示する記述が存在しているものは少ない．そのため，従来の検索エンジンでは表構造を正確に解析することは難しく，表構造の誤解釈が検索精度を落とす一因となっていると考えられる．本研究では，表構造をそのシンタックスから 3 つのタイプに分類し，表構造中のセルの並びに見られる均一性を検出して，その方向と範囲によって表のタイプを自動判定する手法と，表構造中のレイアウト上の特徴とセルの並びの均一性の範囲より，表構造内で見出しとなっているセルを検出する手法による解析を提案し，評価実験を行う．

**キーワード** HTML，表構造，判別分析

# Table structure analysis in HTML document for Improvement in Performance of a WWW Search Engine

Shintaro YAMAMOTO† Akiyo MATSUMOTO†† Tatsuhiro KONISHI‡

Akira TAKAGI§ Teruo KOYAMA‖ Yoshio MIYAKE¶ and Yukihiro ITOH‡‡

†Graduate school of informatics, Shizuoka University

‡Faculty of informatics, Shizuoka University

‡‡Graduate school of science and technology, Shizuoka University

†, ‡, ‡‡3-5-1, Johhoku, Naka-ku, Hamamatsu, Shizuoka, 432-8011 Japan

†† Aoyama Gakuin University 5-10-1, Huchinobe, Sagamihara, Kanagawa, 229-8558 Japan

§NLP Research Laboratory 3-6-30 Nukuiminami-cho,Koganei,Tokyo,184-0014 Japan

‖ National Institute of Informatics 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 Japan

¶Chukyo University 101 Tokodachi, Kaizu-cho, Toyota , Aichi , 470-0393 Japan

E-mail: riir@inf.shizuoka.ac.jp

**Abstract** In this paper, we propose a method to analysis of table structure in HTML document for improving in performance of a WWW search engine. Although there are a lot of table structures in HTML document, they have few plain description of table syntax and role of cell. Because of this problem, it is difficult to analysis of table structure exactly for Search engine. This misunderstanding of table structure brings on lowing of search engine performance. We classify table

structure into three types by syntax of that. And we propose a method to decide of table type and to specify heading cell by focus on uniformity of cell features. We show an experimental evaluation of our method.

**Keyword** HTML, Table structure, Discriminant analysis

## 1. まえがき

WWW 空間上の HTML 文書には，集積した情報を読み手に理解しやすく表示するために表構造が頻繁に用いられている．しかし，Web 検索エンジンでは表構造内の情報を活用している状況には至っていない[1]．そこで本研究の先行研究に当たる松本らの研究[2]の中で，表構造内に検索キーワード 2 語が出現した際，検索キーワード間に意味的係り受け関係のある表の存在する Web ページを優先的に扱うという戦略によって，表構造を用いた検索エンジンの精度向上が検討されている．松本らは意味的係り受け関係の存在しうる表構造内のセルの位置関係を整理することでキーワード間の係り受け関係の有無を判定していた．しかし Web ページ中の表構造のとりうるシンタックスは一意ではなく，見出しセル抽出もセルの出現位置による判定が主であるため，十分な解析がなされていないといえる．本研究では，検索キーワード間の意味的係り受け関係の存在する表構造を優先的に扱う戦略を引き継ぎ，表構造内のキーワード間の意味的係り受け関係を正しく抽出するために，Web ページ中の表のシンタックスの判定と表を構成する各セルの役割の判定を試みる．

表構造は，そのシンタックスによっていくつかのタイプに分類することができ，表を構成する各セルの持つ役割による表内の位置から得られる意味的な関係の強弱は，表のタイプごとに異なっている．しかし，HTML 文書中において，表のシンタックスを明示する記述や，表構造内で見出しを構成するセルとその他のセルとの区別を明示する記述が存在するものは少ない．このため，WWW 検索の際に検索キーワードが表構造中に存在しても，検索エンジンはその表を正確に解釈し，ユーザの検索意図との適合を判断することが難しい．このとき，表のタイプと，検索キーワードが出現したセルの役割を解析することにより，検索キーワード間の係り受け関係の有無を判定することができ，検索精度の向上に役立てることができると考えられる．

本研究では，HTML 文書中の表構造について，表内のセルに見られる特徴の均一性に着目し，表のタイプの判定と，見出しを構成するセルの判定をおこなうシステムを提案した．表構造は，同質の情報を集積し規則的に並べることで形成されている．集積された情報における共通のクラスに属する情報を持つセル間には，テキスト情報も含めたレイアウト上の特徴の均一性が存在すると考えられる．この均一性の方向，範囲を特定することによって，表のタイプ判定と，見出しセルの判定を目指す．

## 2. 表構造の解析

本章では，表構造とは何かを検討し，本研究における表のシンタックスによるタイプ分類と，表を構成するセルの役割の定義を行う．

### 2.1. 表の役割と関連研究

HTML 文書中には，多くの表構造が用いられている．一般的に，表とは集積した情報を読み手に理解しやすく整理した構造である．表構造とは，情報抽出，情報検索の分野において無視することはできない重要な情報である．

そのため，表構造中から情報を抽出する手法については様々な手法が提案されており，大谷ら[3]，大前ら[4]，板井ら[5] によるものなどがある．これらは，属性，属性値が書かれている位置を特定し属性と属性値の組み合わせを抽出する手法を提案するものである．しかしこれらの研究では，表構造内の属性―属性値の関係を抽出・整理することを目的としており，ウェブ検索の精度向上に利用することを前提とした表構造全体としてのシンタックスの解析には言及してはいない．

### 2.2. 表のタイプの定義

表構造とは共通の構造を持つ情報を集積したものであり，その構造は 2 次元のマトリクスである．このような構造で複数の均一な情報を表現する場合，個々の情報の表現と集積方法について，基本的に可能なタイプは以下の 3 つとなる．

(1) 1 行で一つの情報を表現し，それらを複数行重ねて表を構成するタイプ

(2) 1 列で一つの情報を表現し，それらを複数列重ねて表を構成するタイプ

(3) 1 セルで一つの情報を表現し，それらを縦横 2 方向に集積して表を構成するタイプ

このうち(1)，(2)は行と列を入れ替えるだけで相互に変換可能であるため，同質の構造と考えられる．よって，以下の 2 タイプを表のタイプとして定義する．

タイプ 1：1 行（または 1 列）で一つの情報を表し，それらを複数行(列)重ねて表を構成する

タイプ 2：1 セルで一つの情報を表現し，それらを縦横 2 方向に集積して表を構成する

また，タイプ 1 の中には，内容セルの行(列)が 1 行(列)しか存在しない表も多い．これらを区別するために，

内容セルが複数行(列)集積されている表をタイプ 1－1，内容セルが 1 行(列)のみの表をタイプ 1－2 と定義する．

主なスクリプト言語

| 言語 | 開発者 | 発表年 | 実行速度 |
|---|---|---|---|
| Perl | Larry Wall | 1987 | ○ |
| Python | Guido van Rossum | 1995 | ○ |
| Ruby | まつもとゆきひろ | 1995 | △ |

図 1 ：タイプ 1-1 例

施設情報

| | |
|---|---|
| 名称 | 浜松駅 |
| 住所 | 静岡県浜松市中区<br>砂山町6-2 |
| 駅構造 | 高架駅 |
| 路線 | 東海道本線<br>東海道新幹線 |
| 開業<br>年月日 | 明治21年<br>9月1日 |

図 2 ：タイプ 1-2 例

郵便物の料金

| | 25gまで | 50gまで | 100gまで |
|---|---|---|---|
| 定形郵便物 | 80円 | 90円 | 100円 |
| 定形外郵便物 | 100円 | 120円 | 140円 |

図 3 ：タイプ 2 例

### 2.3. セルの役割の定義

一般的に，タイプ 1 では 1 セルが集積された情報の持つ一つのクラス情報を表し，タイプ 2 では 1 セルが集積された情報の一つを表している．これらのセルが表の内容を表していることから，これらを内容セルと呼ぶ．図 1 のタイプ 1－1 では 2～4 行目の各セル，図 2 のタイプ 1－2 では 2 列目の各セル，図 3 のタイプ 2 では 2 行 2 列目セルから 4 列目セルと 3 行 2 列目セルから 4 列目セルが内容セルである．

また，タイプ 1 では行(列)方向に集積された内容セルの属するクラス名を表す列(行)見出しセルが存在する．図 1 中 1 行目の各セル，図 2 では 1 列目の各セルがそれぞれ列見出しセルと行見出しセルである．タイプ 2 では内容セルの同行と同列に，それぞれ内容セルと併せて一つの情報を構成する列・行見出しセルが存在し，行見出しセル同士と列見出しセル同士は同じクラスに属する情報を持つ．図 3 中の 1 行目と 1 列目の各セルがそれぞれ列見出しセルと行見出しセルである．

さらに，表全体の見出しとなる情報を表すセルを表見出しセルと呼ぶ．以上のセルの役割を整理する．

- 表見出しセル：表全体が何を表しているかを表すセル
- 行・列見出しセル
  - タイプ 1：行(列)方向に並べられた内容セルの属するクラス名を表すセル
  - タイプ 2：内容セルと併せて一つの情報を構成する要素となり，行（列）見出し同士は同じクラスに属する情報を持つ
- 内容セル　　：集積された共通の構造を持つ情報を表すセル

## 3. 表構造解析アルゴリズム

共通のクラスに属する情報を表すセルを集積したタイプ 1－1 とタイプ 2 の表内には，内容セル間にテキスト内容とレイアウト上の特徴の均一性が存在する．情報の構成より，タイプ 1－1 は均一な特徴を持つ内容セルの並びが，行または列の一方向のみであるという性質を持ち，タイプ 2 は行と列の両方向に内容セルの均一性を持つという性質を持つ．この性質を利用し，内容セルの均一性の方向が行方向であるのか列方向であるのか，または行列両方向に存在するのかを特定することによる表のタイプ判定手法を提案する．また，見出しセルと内容セルでは特徴が異なることから，見出しセルと内容セルの間には内容セル間に見られる特徴の均一性は存在しないため，セル間の均一性が存在する範囲を判定することで，見出しセルを判定するための有益な情報を得ることができる．

セル間の均一性の判定には，各セルの持つレイアウト上の特徴を用い，見出しの判定においては，見出しセルに共通するレイアウト上の特徴と，出現位置に着目した．提案手法ではまず，多変量解析の一手法である判別分析を利用して得られた判別式を用いてセル間の均一性の有無を判定し，セル間均一性情報と見出し候補セルの位置を併せて均一性のある行と列の抽出を行う．続いて，行単位・列単位での均一性情報を用いて表全体での均一性の方向を判定する．最後に，表内の均一性の方向と範囲，見出しセルの位置から表のタイプ判定を試みる．

### 3.1. 見出しセルの判定手法

表構造内の見出しセルに共通する特徴と出現位置より見出しセルを判定する．それぞれの見出しに共通する特徴を以下に示す．

[表見出しセル]

- 表の先頭に存在する
- 1 行(列)にわたる連結セルより構成される
- 内容セルに比べレイアウト上の強調が存在

・　見出し内には句点は存在しない

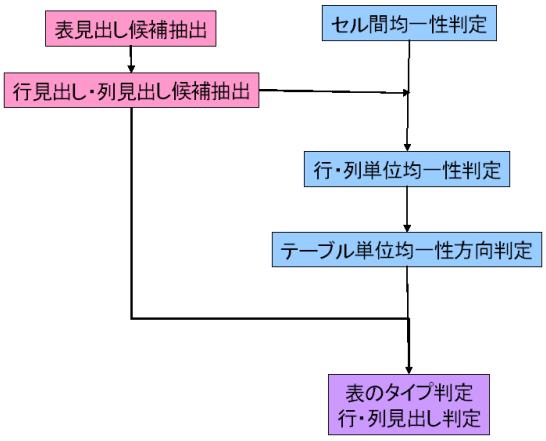


図 4　提案手法の流れ

[行(列)見出しセル]

・　同列(行)に存在
・　表見出しを除き，表の先頭列(行)に存在
・　内容セルに比べレイアウト上の強調が存在
・　内容セルのクラスの数だけ見出しが存在
・　見出し内には句点は存在しない

この特徴を元に，すべての行(列)に対し見出しセルからなる行(列)か否かの判定を行う．また，行・列見出しについては表のタイプ判定で最終的に判定するのでこの段階では見出し候補セルとしておく．

## 3.2. セル間の均一性

本研究では，HTML タグにより与えられた背景色・テキスト強調等といった特徴に加え，テキストの文字種や文字数を含めた特徴を，各セルの持つレイアウト上の特徴と考え，これらの特徴を用いてセル間の均一性の有無の判定を試みる．

### 3.2.1　判別分析を用いたセル間の均一性判定手法

セル間の均一性の判定は，各セルより抽出したレイアウト上の特徴の比較により行う．この際 HTML 文書は半構造データであるため，特徴の完全一致を基準としては多くの均一性を取りこぼす恐れがある．多少の差異を吸収し柔軟な判定を行うために，判別分析という手法を用いる．

判別分析とは，要因となる多変量の量的データと，分類先となる質的データのセットからなる学習データを学習させ，テストデータがどの分類先に属するか判定する手法である．本研究では，学習用データについて目視によりセル間均一性の有無を判定した情報と，セル間の特徴の差異($p_i$)を学習させ，均一性の存在する群と存在しない群を最も良く分ける判別式（$Z = w_1p_1 + w_2p_2 + \cdots + w_np_n + a$ (a:定数)）を得る．

この判別式に均一性の有無のわからないセル間の特徴の差異を与え，結果の値によってセル間の均一性を判定する．

本研究では，レイアウト上の特徴に関連あると考えられる 21 の情報を各セルより抽出し，セル間の差異とセル間の均一性の有無を共に学習させた．

### 3.2.2　行単位・列単位での均一性判定

行単位・列単位での均一性は，行(列)内のセル数のうち，均一性を持つセルの占める割合がしきい値を超えているかにより判定する．しきい値の決定方法は，行(列)内セル数によって場合分け（２セル，３セル，４セル，５セル，６セル，７セル以上の６つの場合）を行い，学習用テーブルを用いた統計により，均一行(列)の取りこぼしを防ぐようしきい値を決定する．

均一セルの占める割合としきい値が等しい場合には，均一セルと見出しセル候補との重複の有無により，行単位・列単位の均一性を判定する（表 1）．

表 1　しきい値と等しい割合の際の判定

| 見出しセルと均一セルの重複 | 行(列)単位均一性の有無 |
|---|---|
| 重複なし | 均一性あり |
| 重複あり | 均一性なし |

### 3.2.2 均一性の方向判定

表全体での均一性の方向は，表の行(列)数のうち均一行(列)がしきい値を超えているかにより判定する．しきい値の決定方法は，学習用テーブルを用いた統計により，均一性の存在する方向を取りこぼさないよう決定する．均一性を持つ方向の持つ均一行(列)数の占める割合の最小値が 0.75，均一性の存在しない方向の持つ均一行(列)数の占める割合の最大値が 0.36 であったため，しきい値はこの間の数値を採用する．今回は 0.5 をしきい値として決定した．

2 行(列)のみからなる表については，とりえる表のタイプと行単位の均一性の組み合わせより，タイプ 1－2 であるかを判定する（表 2）．

表 2　2 行の表のとりうるタイプと行方向均一性

| | 1 行目 | 2 行目 | 行方向均一性 |
|---|---|---|---|
| タイプ 1－1 行見出しあり | 均一 | 均一 | あり |
| タイプ 1－1 行見出しなし | 均一 | 均一 | あり |
| タイプ 1－2 | 均一/なし | なし | |
| タイプ 2 | 均一 | 均一 | あり |

1 行目と 2 行目に均一性なしが存在しえるのは，タイプ 1－2 のみである．よって，この条件を満たす際にはタイプ 1－2 であると判定する．

### 3.2.3　表のタイプ判定・見出しセル判定

均一性の方向判定結果と，見出し候補セルの判定結果を用いて，表のタイプと見出しセルを判定する．均一性の方向と，行・列見出し候補セルの組み合わせを表 3 に整理する．

表 3　均一性の方向と見出し候補セルの組み合わせ

<table>
<tr><th></th><th>行方向</th><th>列方向</th><th>両方向</th><th>均一性なし</th></tr>
<tr><td rowspan="2">列見出し候補あり</td><td>3 行以上<br>①</td><td rowspan="2">タイプ<br>1－1</td><td rowspan="2">⑤</td><td>3 行以上<br>⑦</td></tr>
<tr><td>2 行<br>タイプ<br>1－2</td><td>2 行<br>タイプ<br>1－2</td></tr>
<tr><td rowspan="2">行見出し候補あり</td><td rowspan="2">タイプ<br>1－1</td><td>3 列以上<br>③</td><td rowspan="2">⑥</td><td>3 列以上<br>⑧</td></tr>
<tr><td>2 列<br>タイプ<br>1－2</td><td>2 列<br>タイプ<br>1－2</td></tr>
<tr><td>行・列見出し候補あり</td><td>②</td><td>④</td><td>タイプ 2</td><td>⑨</td></tr>
<tr><td>見出し候補なし</td><td>※1</td><td>※2</td><td>※3</td><td>⑩</td></tr>
</table>

表 3 中の①～⑩は，均一性の方向と，見出し候補セルの位置に矛盾が生じる組み合わせである. このとき，見出し候補セルを元に判定するか，均一性を元に判定するのかにより，表のタイプと見出しセルの位置が異なる．この組み合わせを表 4 に示す.

表 4　均一性の方向と見出し候補位置が矛盾した際の判定

<table>
<tr><th></th><th>見出し判定利用</th><th>均一性利用</th></tr>
<tr><td>①</td><td>タイプ 1－1<br>列方向均一<br>列見出しあり</td><td>タイプ 1－1<br>行方向均一<br>見出し省略</td></tr>
<tr><td>②</td><td>タイプ 2<br>行･列見出しあり</td><td>タイプ 1－1<br>行方向均一<br>行見出しあり</td></tr>
<tr><td>③</td><td>タイプ 1－1<br>行方向均一<br>行見出しあり</td><td>タイプ 1－1<br>列方向均一<br>見出し省略</td></tr>
<tr><td>④</td><td>タイプ 2<br>行･列見出しあり</td><td>タイプ 1－1<br>列方向均一<br>列見出しあり</td></tr>
<tr><td>⑤</td><td>タイプ 1－1<br>列方向均一<br>列見出しあり</td><td>タイプ 2<br>行･列見出しあり</td></tr>
<tr><td>⑥</td><td>タイプ 1－1<br>行方向均一<br>行見出しあり</td><td>タイプ 2<br>行･列見出しあり</td></tr>
<tr><td>⑦</td><td>タイプ 1－1<br>列方向均一<br>列見出しあり</td><td rowspan="4">イレギュラー<br>テーブル</td></tr>
<tr><td>⑧</td><td>タイプ 1－1<br>行方向均一<br>行見出しあり</td></tr>
<tr><td>⑨</td><td>タイプ 2<br>行･列見出しあり</td></tr>
<tr><td>⑩</td><td>イレギュラー<br>テーブル</td></tr>
</table>

表構造内に背景色かテキスト強調のレイアウト属性を持つテーブルでは見出し判定精度が高く，レイアウト属性を持たないテーブルでは均一性判定精度が高いという結果が学習用テーブルによる実験から得られている．よって，表構造内のレイアウト属性の有無によって判定先を決定する.

また，表 3 中の※1～※3 では，均一性の方向から表のタイプは一意に決定している．このとき内容セルの均一性の範囲を調べることにより，見出しセルの有無を判定する.

行方向に均一性ありと判定されたテーブルにおいて，各行内の均一セルの位置に着目する．このとき，「表見出し列＋1」列目セルがすべての行において均一セルに含まれていなければ，内容セルの範囲は「表見出し列＋1」列目以降であると判定し,「表見出し列＋1」列目を行見出し列と判定する．表見出し列が存在しない際には 1 列目を対象として判定を行う. また，列方向に均一性が存在していれば同様の処理を行う.

## 4. 評価実験

本章では，提案手法の評価のための実験について述べる. 実験結果の表中, Hit は正しく抽出できたもの，Miss は取りこぼしたもの，FA(Flase Alarm) は誤って抽出したもの，Precision は Hit/(Hit+FA)，Recall は Hit/(Hit+Miss)，F-value は Precision と Recall の調和平均である.

### 4.1. 評価データの作成

以下の手順にて，学習用・評価用データを作成する.

1. 検索キーワード 118 組による検索結果より 200 ページを選定
2. 200 ページから，目視判定によって表構造 159 テーブルを抽出
3. 159 テーブルを対象に，大学生 3 名により「表のタイプ」,「均一性の方向」,「均一性のある行・列」,「見出しセル」を目視判定
4. 3 名の判定の一致した 117 テーブルを採用
5. 117 テーブルを学習用 67 テーブル, 評価用 50 テーブルにランダムに振り分け

### 4.2. 見出しセルの判定精度

評価用 50 テーブルを対象に，表見出しセル判定，行・列見出し判定を行った．表見出しセル判定結果を表 5，行・列見出し判定結果を表 6 に示す.

表 5　表見出しセル判定結果

表見出し数：19

| Hit | FA | Miss | CR | Precision | Recall | F-value |
|---|---|---|---|---|---|---|
| 11 | 0 | 8 | 39 | 100.0% | 57.9% | 73.3% |

表 6　行・列見出し判定結果

行・列見出し数：38

| Hit | FA | Miss | CR | Precision | Recall | F-value |
|---|---|---|---|---|---|---|
| 20 | 5 | 18 | 61 | 80.0% | 52.6% | 63.5% |

### 4.3. 表のタイプの判定精度

評価用 50 テーブルを対象として，表のタイプ判定を行う．表のタイプを正しく判定するためには，均一性の方向を行・列方向共に正しく判定していることと，見出しセルを正しく判定していることが条件である．行方向均一性判定結果と列方向均一性判定結果を表 7, 表 8 に示す．

表 7　行方向均一性判定結果

均一数：11

| Hit | FA | Miss | CR | Precision | Recall | F-value |
|---|---|---|---|---|---|---|
| 10 | 18 | 1 | 23 | 35.7% | 90.9% | 51.3% |

表 8　列方向均一性判定結果

均一数：34

| Hit | FA | Miss | CR | Precision | Recall | F-value |
|---|---|---|---|---|---|---|
| 30 | 13 | 4 | 5 | 69.6% | 88.2% | 77.9% |

実験の結果，評価用 50 テーブル中，上記の条件を満たし表のタイプを正しく判定したのは 20 テーブルであった．

### 4.4. 考察

表のタイプ判定精度を落としている原因について検討する．タイプを誤判定したテーブルを分析したところ，1 行(列)で一つの情報を表すタイプ 1−1 の中でも行(列)数が少ない表における均一性の方向の誤判定と，表中の行見出し列の取りこぼしが多いことを確認した．表 8 に行見出し列の判定結果を示す．

表 8　行見出し列判定結果

行見出し列数：17

| Hit | FA | Miss | CR | Precision | Recall | F-value |
|---|---|---|---|---|---|---|
| 4 | 3 | 13 | 32 | 57.1% | 23.5% | 33.3% |

行見出し列の取りこぼしには，2 列からなるタイプ 1−2 が多く，見出し列と内容セル列の間のレイアウト上の特徴の差異が少ないため，見出しを検出できていないものが確認された．

行(列)数が少ない表における均一性の方向の誤判定は，セル数が少ない列(行)における列(行)単位での均一性の誤判定により起こる．このため，行単位・列単位での判定精度を 7 セル以下からなる行(列)と，8 セル以上からなる行(列)に分けて集計を行った．結果を表 9, 表 10 に示す．

表 9　7 セル以下からなる行(列)の判定結果

均一数：205

| Hit | FA | Miss | CR | Precision | Recall | F-value |
|---|---|---|---|---|---|---|
| 82 | 123 | 14 | 405 | 40.0% | 85.4% | 54.5% |

表 10　8 セル以上からなる行(列)の判定結果

均一数：45

| Hit | FA | Miss | CR | Precision | Recall | F-value |
|---|---|---|---|---|---|---|
| 40 | 5 | 7 | 2 | 88.9% | 85.1% | 87.0% |

7 セル以下からなるセル数の少ない行(列)で誤判定が起こるのは，表を構成するセルの中に見出しセルとレイアウト上の特徴の共通点の多いセルや，目視判定では均一性のない内容セル間に、抽出したレイアウト上の特徴に共通点が存在したときに，行中の均一セルの占める割合がしきい値を超えてしまうことが原因である．さらに FA の中には，テキスト以外のレイアウト上の特徴に乏しく，テキストから抽出している文字数や字種といった特徴に差異が少ないために誤判定しているものも多く，レイアウト上の特徴という表層的な情報のみでの判定は困難であるものの存在が確認された．

提案手法では行(列)内のセル数ごとにしきい値を設けたが，十分な効果を得るには至らなかった．

判定精度を向上させるには，セル数の少ない行(列)に属するセル間の均一性判定の改善が必要である．具体的には，

- テキストから得られる字種情報の分類の細分化
- 各セルより抽出する特徴の検討

などが考えられる．

また，表層的な情報での解析が困難な表については，形態素解析やシソーラスの利用に代表される，提案手法では用いていないテキストの意味による解析により得られる情報が必要と考えられる．

## 5. むすび

本研究では，WWW 検索精度向上の為に HTML 文書中の表構造をセルの均一性に着目して解析し，表のタイプと見出しセルを判定する手法を提案した．学習と評価用のデータセットを作成し，評価実験を行った．

今後の課題として，実験結果より得られた知見をもとに，レイアウト上の特徴を用いた手法の改善とテキストの意味解析を用いることによって，判定精度を向上させることが必要である．そして，本研究で提案したシステムを先行研究の検索システムに組み込み，検索性能の評価を行う予定である．

## 6. 参考文献


[1] 岩口義広, 鄭眠洙, 獅々堀正幹, 青江順一: WWW 空間上に存在する表構造の一索引化手法, 情処研報 2001-NL-142, pp. 159− 166 (2001).

[2] 松本章代，小西達裕，高木朗，小山照夫，三宅芳雄，伊東幸宏：表構造における意味的関係に基づく WWW 検索性能の向上，電子情報通信学会論文誌 D，Vol.J91-D，No.3，pp.560-575 (2008.3).

[3] 大谷貴志, 獅々堀正幹, 柘植覚, 北研二: HTML 形式の表構造の内容解析手法とその応用に関する研究，情処研報 2002-NL-154，Vol.2003，No.23，pp.137− 144 (2003).

[4] 大前信弘，黄瀬浩一：Web の表を対象とした属性の自動識別, 情処研報 2006-NL-171, Vol.2006, No.1, pp. 43− 48 (2006).

[5] 板井久美，高須淳宏，安達淳: HTML からの情報抽出と統合, NII journalNo.6, pp. 9− 19 (2003).