動向情報の根拠探索を目的とした統計表データの自動認識

松井　侑祐†　　宮森　　恒†

† 京都産業大学先端情報学研究科
〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{i1358112,miya}@cse.kyoto-su.ac.jp

あらまし　近年，インターネットの急速な発展・普及に伴い，膨大な量の情報が溢れている状況となっている．その中には，質が高くて役に立つ情報からデマや誤った情報など，質の高いものから低いものまでが混在し，どれが正しい情報なのかをユーザが選別することはますます困難となっている．筆者らは，ソーシャルネットワーク上の言説，特に，ある対象の時間的変化を記述した動向情報を対象に，その客観的根拠を統計表データから探索するシステムを提案している．しかし，統計表は，見出し内容やその階層構造の表現形態が多様であり，これらの自動認識においては精度の改善が課題となっている．本稿では，表の各項目の認識を系列ラベリング問題として捉え，機械学習に基づく方法で統計表構造を認識する手法を提案する．従来手法であるルールベースに基づく手法と比較評価する．

キーワード　統計表認識, 系列データ, 機械学習, CRF, 根拠探索, 動向情報

1. はじめに

近年，インターネットの急速な発展・普及に伴い，膨大な量の情報が溢れている状況となっている．その中には，質が高くて役に立つ情報からデマや誤った情報など，質の高いものから低いものまでが混在し，どれが正しい情報なのかをユーザが選別することはますます困難となっている．Facebook や Twitter，ブログなど，情報発信手段は発達する一方，信頼度の高い情報を見分けるための手段は現状で十分に整っているとはいえない．

ここで，信頼度を見分ける方法の一つに，客観的根拠の有無に着目した方法が挙げられる．例えば，電子掲示板等に掲載された匿名発信者による真偽の不明確な情報について，何らかの情報源から適切な客観的根拠を見つけることができれば，ユーザ自身がその情報について信頼度を判断するのは格段に容易になると考えられる．しかし，一般に，ユーザ自身が情報を閲覧するたびに適切な情報源を探すことは，ユーザに大きな負担となり現実的でない．また，ユーザ自身で適切な情報源を見つけられないことも十分に考えられる．従って，情報の客観的な根拠を効率よく探索できるシステムの実現が望まれる．

一方で，政府や企業により，膨大な各種データが統計表データとして Web 上で広く公開されている．これらの情報は，政府や企業が調査・集計したものであり，一般に信頼度の高い情報源と考えることができる．ただし，これらのデータは Excel 形式や CSV 形式による表形式で公開されており，中には見出しを階層関係にするなど複雑な形式のものも存在する．コンピュータでこのようなデータを扱うには表の認識や分析といった課題を克服する必要がある．

これらを踏まえ，筆者らは，ソーシャルネットワーク上の言説，特に，ある対象の時間的変化を記述した動向情報を対象に，その客観的根拠を統計表データから探索するシステム [1] を提案した．ユーザが，ソーシャルネットワーク上で見つけた信頼性を判断したい情報をクエリとして入力すると，そのクエリの根拠となる統計表データ（あるいは関連するが矛盾する統計表データ）が探し出され，理解しやすい形に変換された上でユーザに提示される．ただし，統計表データには複雑な見出しをもつものも多く，その構造をより適切に認識するように手法を改善する課題が残っている．

そこで，本稿では，表の各項目の認識を系列ラベリング問題として捉え，機械学習に基づく手法で統計表の構造を認識する手法を提案する．本稿では，CSV 形式で提供されている政府統計データを対象とし，筆者らが以前に提案したルールベースに基づく手法と比較する評価実験を行い，提案手法の利点と問題点について考察する．

本稿の構成は，以下の通りである．2 節では，関連研究を紹介し，3 節では，動向情報の根拠探索システムと本稿の取り組む範囲を述べる．4 節では，統計表構造を認識する提案手法について詳細に述べる．5 節では，評価実験とその結果を示し，考察を行う．最後に 6 節で，まとめと今後の課題を整理する．

2. 関連研究

これまでにも表やその構造を認識する研究は数多く行われている．

Kieninger ら [2] は，文書画像から表を認識する手法を提案している．ビジネスレター文書の画像を入力として，表中のセルなどのユニットを認識し，それらユニットの空間的配置からレイアウトを分析することで，表の位置とその内部構造を認識している．また，Wang ら [3] は，文書画像を対象に最適化手法を用いた表認識を提案している．予め文書画像中の線や文字の領域が大雑把に分割されたものを入力とし，ページのカラムスタイルのラベリング，表がもつ一貫性評価による表候補の絞り込み，ページ全体の分割を最大化する反復更新による最適化を行っている．本稿では，文書画像ではなく，CSV 形式の表データを解析対象としている点が異なる．

また，塚本ら [4] は，HTML 形式で記述された表を対象とし，

表の項目名と項目データの境界を認識し，解像度の比較的小さい携帯端末でも見やすい形式の表に変換する手法を提案している．表の行間あるいは列間で類似度を定め，類似度が低い場合には行間あるいは列間に内容的な切れ目があると認識する．本稿では CSV 形式で記述された表を対象としている点が異なるが，表項目の階層関係の認識も含めた見出し部分の解釈手法については本研究と共通している．

表を利用した応用システムに関する研究も数多く行われている．

Kerpedjiev ら [5] は，マルチメディアプレゼンテーションを生成するシステム AutoBrief を提案している．アナリストや専門家が大量のデータ集合中のパターンや変化を理解することを支援し，そこで得た情報をテキストと図表によって要約し，プレゼン目的に応じて適切な形式で提示するシステムを，輸送計画のシミュレーションを例に実装している．Wang ら [6] は，HTML 形式で記述された比較的大きな表を，画面サイズの小さい携帯端末でも見やすい形式の表に変換するシステムを提案している．HTML で記述された表を，データテーブルとレイアウトテーブルに分類し，データテーブルに対しては，元々の構造情報が保存されるようにしながら，表全体を 1 カラムビューに収めることで閲覧体験の向上を目指している．松下ら [7] は，統計 DB 等から得られる時系列数値情報と，それに関連する内容の一連のテキスト情報を関連付けて視覚化し，ユーザの探索的データ分析を支援する可視化インタフェースを提案している．佐伯ら [8] は，テレビ番組中の図表画像を利用し，Web 上の動向情報の根拠を探索・提示するシステムを提案している．テレビ番組中に登場する図表画像を信頼性の高い情報源と考え，信頼性の不確かな Web 上の動向情報と関連づけることで，ユーザによる情報の信頼性判断を効率よく支援することを目指している．筆者ら [1] は，政府や企業が公開する統計表データを信頼性の高い情報源と考え，これを信頼性の不確かな Web 上の動向情報と関連づけ，探索・提示するシステムを提案している．本稿では，統計表データをより正確に認識することにより，動向情報の根拠探索の精度向上に繋げることを目指す．

3. 動向情報の根拠探索システム

筆者らが提案している動向情報の根拠探索システム [1] の構成を図 1 に示す．なお，本稿で提案する内容は，図中の「表の意味解釈」の部分に該当する．

本システムは大きく 2 つの部分から構成される．

統計表データの収集・解析・登録では，まず，政府や企業等により Web 上で公開されている信頼性の高い統計表データを収集する．次に，統計表データの構造を認識する．具体的には，表のタイトルや見出し，本体，注釈，単位に当たる項目とそれらの構造を認識する．これには，見出しの階層関係の解釈も含まれる．最後に，解析された統計表データを DB に登録する．

統計表データとの関連づけ・提示では，ユーザから入力された Web 上の動向情報をクエリとして解析し，統計表 DB 中の統計表データと関連づける．関連度の高い統計表から該当部分をグラフ画像化し，ユーザに提示する．

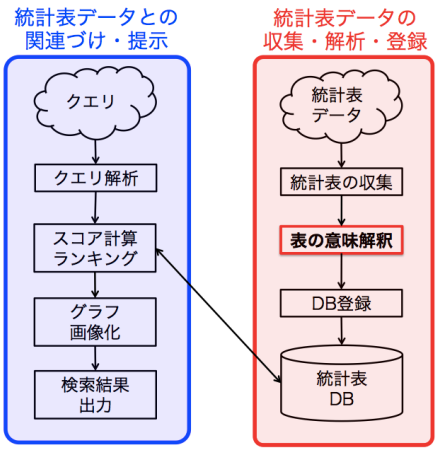

図 1　システム構成図

ユーザは，クエリとして入力した動向情報の内容を，提示されたグラフ画像とを比較することにより，動向情報の信頼度を格段に効率よく判断できるようになると期待される．

4. 提 案 手 法

4.1 提案手法の目的

本手法の目的は，Web 上で公開されている統計表データを正しく意味解釈し，コンピュータが扱いやすいデータ構造に変換することである．ここでは，表を構成する 5 項目を認識対象とし，これらを, 条件付き確率場 (CRF) を用いた識別モデルに基づく機械学習で認識する手法を提案する．

4.2 統計表データの収集

提案システムの目的を実現するためには，信頼性の高い統計表データを収集する必要があるため，本稿では，政府が公開している政府統計 [9] を収集対象とした．

政府統計では，各府省等が収集した各種統計が取りまとめられ，Excel 形式や CSV 形式等で Web 上に公開されている．公開されている政府統計のうち，本稿では日本語で記述されている CSV 形式の統計表データを収集対象とした．さらに，収集した統計表データの中にはライブラリでの読み込みに失敗するデータも存在した．本稿では，それらを除いた CSV 形式の統計表データ 49,961 件を用いることとした．

4.3 統計表構造の認識

4.3.1 認 識 対 象

1 つの統計表には，基本的に，以下の項目が含まれている．

- タイトル
- 注釈
- 列見出し
- 行見出し
- 本体

ここでは，これらの項目に対応するセル (区画) またはセル範囲を認識対象とする．

タイトル，注釈についてはそれぞれ該当するセルを認識する．列見出し，行見出しと本体についてはそれぞれ該当するセル範囲を認識する．なお，見出しについては，表の上側と左側に記

述されていることが一般的であり，これら 2 か所の範囲をそれぞれ認識する．

一般的な統計表において，各認識対象に該当するセルおよびセル範囲を示した例を図 2 に示す．

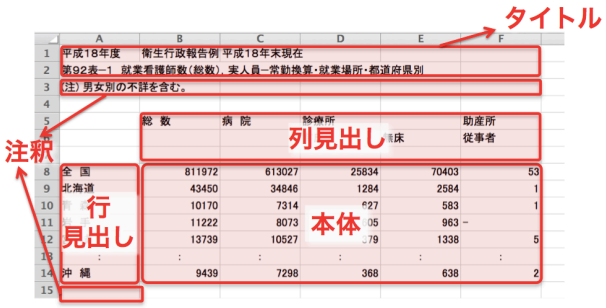

図 2　統計表の認識対象項目に該当するセルおよびセル範囲の例

各項目の認識にあたっては，ルールベースによる手法，および，条件付き確率場 (CRF) を用いた機械学習による手法の 2 つを比較する．まず，筆者らが以前に提案したルールベースに基づく手法について説明する．

4.4　ルールベースによる手法 [1]

ルールベースによる認識の処理手順を図 3 に示す．

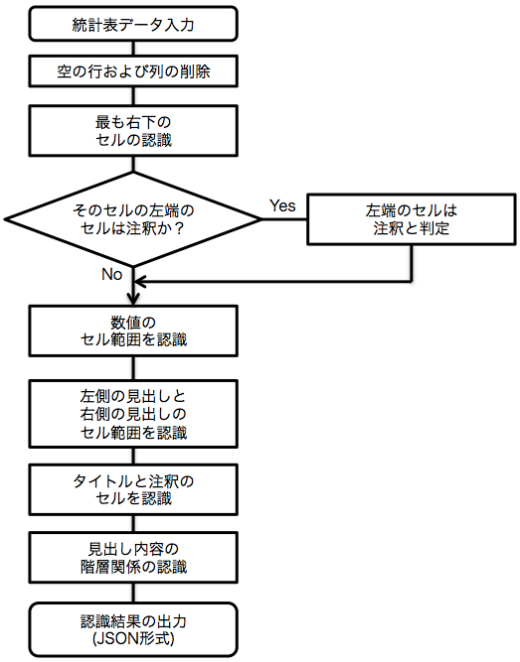

図 3　ルールベースによる認識処理手順

まず，入力された統計表データに対して，空の行および列を削除した後，セル内が空でない表中の最も右下のセルを特定する．その後，表の最下部に注釈が存在するかどうかを判断し，その範囲を特定する．その結果を踏まえて本体のセル範囲を認識し，続けて，行見出しや列見出しのセル範囲，タイトルと注釈のセルを認識する．最後に，見出し内容の階層関係を解析し，認識結果を格納するフォーマット (JSON 形式) で出力する．

4.4.1　空の行および列の無視

表の中には，Excel 等の表計算ソフトで閲覧した際の見栄えを整えるために，1 行全て空のセル，また 1 列全て空のセルとなっている行や列が存在する．このような行や列はデータとしては特別な意味を持たないため，解釈をする際には空の行や列はないものとして処理を続ける．

4.4.2　本体のセル範囲の認識

表における本体の部分は，通常，最も右下の部分に配置され，特定しやすいと考えられるため，まず，本体のセル範囲を認識する．ただし，表の最も左下の部分には注釈が記述される場合があるため，該当箇所が注釈かどうかの判定を行った後，本体のセル範囲を認識する．以下，処理手順を説明する（図 4）．

（1）　入力として与えられた表中の最も右下のセルを特定し，現在の注目セルとする．

（2）　注目セルの左端のセルにおいて，3 文字目以内に「注」の文字が出現する場合，そのセルは注釈であると解釈し，上にある次の非空白行の右端へ移動する．

（3）　この段階での注目セルを，本体のセル範囲の**右下セル**であると決定する．次に，注目セル内の文字列が，数字または記号類 (「・」「-」「…」「X」など) で記述されている場合，そのセルは本体であると判断し，本体ではないと判断されるまで，注目セルを左へ 1 セルずつ移動しながらこの判断を繰り返す．その後，上向きにも，1 セルずつ注目セルを移動しながら，そのセルが本体かどうかの判断をそうでないと判断されるまで繰り返す．この時点の直前で得られた注目セルを，本体のセル範囲の**左上セル**であると決定する．

図 4　本体のセル範囲の認識手順

4.4.3　列見出し・行見出しの認識

一般的な表では，本体のセル範囲の左側に見出しが置かれる．また，行見出しでは，見出し文字列を異なる列のセルに記述することで，見出し内容の階層関係を表す場合がある（図 5）．よって，行見出しは以下のように特定する．まず，本体のセル範囲の左隣の列について，本体のセル範囲の左側に該当するセル群を，現在の注目セル群とする．注目セル群に非空白文字が含まれるセルが 1 つ以上ある場合，その注目セル群は見出しであると判断し，見出しでないと判断されるまで，注目セル群を左へ 1 セルずつ移動しながらこの判断を繰り返す．

なお，列見出しについても，本体のセル範囲の一つ上の行について，本体のセル範囲の上側に該当するセル群を，現在の注

目セル群として処理を開始し，判断と移動を同様に繰り返すことで，特定することができる．

4.4.4 タイトル・注釈の認識

列見出しのセル範囲よりも上にあるセルのうち，非空白文字が含まれる全てのセルを，タイトルまたは注釈であると判断する．ここで，条件に該当するセル内の文字列について，3 文字目以内に「注」の文字が出現する場合，そのセルは注釈であるとし，そうでないセルはタイトルであると判断する．

4.4.5 見出し内容の階層関係の認識

3.5.3 節で述べたように，統計表の中には，見出しを異なる列のセルに記載するなどして，見出し内容の階層関係を表現する例が少なくない．ここでは，見出し内容の階層関係を認識する手順を説明する．

統計表における見出しの階層関係は，多くの場合，以下のいずれかのパターンで表現される．

- 異なる列または行のセルに見出し内容を記述し，段違いにして表現する
- 同じ列のセルについて，見出し内容の開始位置に異なる文字数の空白文字を埋めることで表現する
- 数値が何も記述されていない行を上位の見出し行として表現する

本稿では，上記の各パターンのどれに該当するかを判断し，そこで表現されている階層関係を，最上位の要素から順に子の要素をスラッシュでつなぎ，順番に記述する形式で表現することで，行見出しおよび列見出しをそれぞれ 1 列，1 行のみとした単純な形式の表に変換する．階層関係で記述されている見出しの例と，変換後の見出しの例を，それぞれ，図 5 と図 6 に示す．

		総数	歯科診療台	
			施設数	台数
2000年				
総数	総数	9026	1758	10867
	国	294	221	2719
	厚生労働省	22	21	43
	その他	272	200	2676
	公的医療機関	1362	489	1833
	都道府県	303	153	574
	市町村	757	238	865
2005年				
総数	総数	9362	1925	11533
	国	298	243	3015
:	:	:	:	:

図 5　階層関係で記述されている見出しの例

		総数/	歯科診療台/ 施設数/	歯科診療台/ 台数/
2000年/総数/総数/		9026	1758	10867
2000年/総数/国/		294	221	2719
2000年/総数/国/厚生労働省/		22	21	43
2000年/総数/国/その他/		272	200	2676
2000年/総数/公的医療機関/その他/		1362	489	1833
2000年/総数/公的医療機関/都道府県/		303	153	574
2000年/総数/公的医療機関/市町村/		757	238	865
2005年/総数/総数/		9362	1925	11533
2005年/総数/国/		298	243	3015
:	:	:	:	:

図 6　階層関係で記述されている見出しの単純な見出しへの変換例

4.5 条件付き確率場 (CRF) を用いた機械学習による手法

統計表データから 4.3.1 節で述べた各項目を認識する問題を系列ラベリング問題ととらえ，条件付き確率場 (CRF) を用いた識別モデルを採用する．

4.5.1 学習用データの準備

統計表データから学習用データを作成する．統計表データは 2 次元上の各セルに文字列が格納されたデータ構造と捉えることができるが，ここでは，1 次元の系列データとみなし，学習用データを作成する．

2 次元から 1 次元へ変換する手順として，水平方向，垂直方向の 2 通りの走査方法を導入する．

水平走査では，表中の最も右下のセルを注目セルとし，そのセルから左方向へ 1 セルずつ走査する．表の左端まで到達した場合は，1 行上の最も右端のセルを次の注目セルとして連結し，同様に左方向へ 1 セルずつ走査する．最も上の行の左端に到達した時点で，全てのセルが 1 次元データとして記述されたこととなる．

垂直走査では，表中の最も右下のセルを注目セルとし，そのセルから上方向へ 1 セルずつ走査する．表の上端まで到達した場合は，1 行左の最も下端のセルを次の注目セルとして連結し，同様に上方向へ 1 セルずつ走査する．最も左の行の上端に到達した時点で，全てのセルが 1 次元データとして記述されたこととなる．

図 7 の表に対して，水平走査および垂直走査のそれぞれで変換する際の注目セルの走査順序を図 8，図 9 に示す．

	A	B	C	D	E
1	A県の人口増減				単位：人
2					
3		○○市		△△市	
4		男	女	男	女
5	1993年	26442	27119	16156	16370
6	1994年	26467	27150	16259	16388
7	1995年	26492	27181	16362	16406
8	1996年	26517	27212	16465	16424
9	1997年	26542	27243	16568	16442
10	1998年	26567	27274	16671	16460
11	1999年	26592	27305	16774	16478

図 7　走査対象の統計表データ

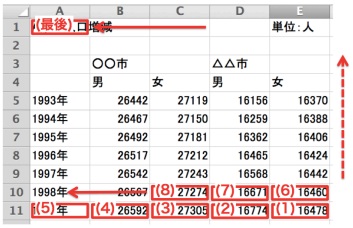

図 8　水平走査による注目セルの走査順序

4.5.2 学習用データの形式

学習用データは，着目する 1 セルから得られる以下の 4 つの

図 9　垂直走査による注目セルの走査順序

要素と正解ラベルをタブ区切りで連結したものを 1 行として，1 次元データのセル数だけ行数をもつ形式となっている．

- セルに含まれる文字列
- セルの垂直方向での区画番号
- セルの水平方向での区画番号
- セルに含まれる文字列の種別
- 正解ラベル

セルの垂直方向での区画番号は，表中の上端から下端までを N 分割し，各セルが垂直方向においてどの位置に存在するかを 0 から N-1 までの ID で表現した番号である．同様に，セルの水平方向での区画番号は，表中の左端から右端までを N 分割し，各セルが水平方向においてどの位置に存在するかを 0 から N-1 までの ID で表現した番号である．

セルに含まれる文字列の種別は，そのセルに含まれる文字列が数値として解釈できるか，そうでないか，もしくは空欄であるかの 3 値で表現する．

正解ラベルは，表 1 のように 11 種類のラベルを定義した．IOB2 タグを用いて，認識対象であるタイトル (TITLE)，注釈 (COMMENT)，行見出し (ROW_HEADER)，列見出し (COL_HEADER)，本体 (BODY) のそれぞれについて，表現の開始を表す B(Beginning)，表現の 2 つ目以降の構成要素を表す I(Inside) を付与し，いずれにも当てはまらないものを該当表現外の要素を表す O(Outside) とする．

表 1　定義した 11 種類の正解ラベル

ラベル名	付与対象のセル
B_TITLE	タイトルの開始点のセル
B_COMMENT	注釈の開始点のセル
B_ROW_HEADER	行見出しの開始点のセル
B_COL_HEADER	列見出しの開始点のセル
B_BODY	本体の開始点のセル
I_TITLE	タイトルの 2 セル名以降のセル
I_COMMENT	注釈の 2 セル名以降のセル
I_ROW_HEADER	行見出しの 2 セル名以降のセル
I_COL_HEADER	列見出しの 2 セル名以降のセル
I_BODY	本体の 2 セル名以降のセル
O	いずれにも該当しないセル

図 7 の統計表データから水平走査で作成した学習用データの例を図 10 に示す．

```
"16478"  4    4    num     B_BODY
"16774"  4    3    num     I_BODY
"27305"  4    2    num     I_BODY
"26592"  4    1    num     I_BODY
"1999年" 4    0    text    B_COL_HEADER
"16460"  4    4    num     B_BODY
"16671"  4    3    num     I_BODY
"27274"  4    2    num     I_BODY
"26567"  4    1    num     I_BODY
"1998年" 4    0    text    B_COL_HEADER
                    :
                    :
```

図 10　学習用データの例

4.5.3　学習および認識手順

学習においては，セルに含まれる文字列，セルの垂直方向での区画番号，セルの水平方向での区画番号，セルに含まれる文字列の種別，正解ラベルのそれぞれ，および，それらを組み合わせたユニグラム，バイグラムを素性として用いた．

認識においては，学習した識別器の出力に基づき各項目を特定する．ただし，行見出し，列見出し，本体については，セル範囲を以下の手順で認識した．

まず，学習結果から得られた BIO のラベルを 2 次元上に配置し，各行において B または I のラベルの左端，右端を特定する．その後，左端の位置，右端の位置においてそれぞれラベルの多数決を取り，最多ラベルの位置を左端，右端と定める．(図 11) そして，得られた左端と右端の位置が完全に一致する行を探し，その中で最も上の行と下の行をそれぞれ上端，下端と特定する（図 12）．このようにして得られた左端，右端，上端，下端からセル範囲を認識する．

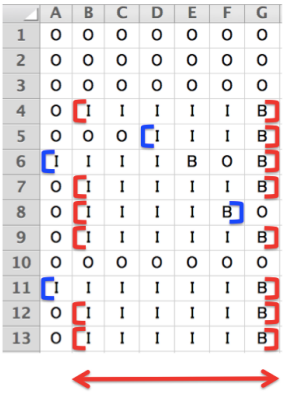

図 11　セル範囲の左端，右端の認識例

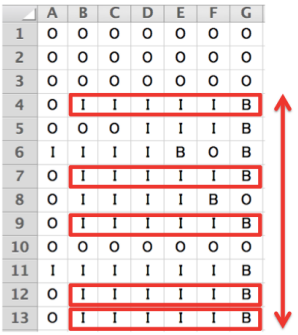

図 12　セル範囲の上端，下端の認識例

5. 評価実験と考察

5.1 実験 1: セルの走査方法，区画の分割数と識別率の関係

本実験では，提案手法の識別率が，セルの走査方法，区画の分割数によってどのように変化するかを明らかにする．

提案手法における，セルの水平走査と垂直走査，および，区画の分割数 N を N=1, 2, 3, … ,10 と変化させた場合の各組み合わせについてそれぞれ識別を行った．結果は，表中の各項目を正しく識別できた割合（識別率）により評価する．

収集した 49,961 件の CSV 形式の統計表データから無作為に選んだ 600 件に対し学習を行い，無作為に選んだ別の 100 件に対し識別を行った．

なお，実験にあたり，各項目の識別結果の正解は，それらが，正解ラベルのセルあるいは正解ラベルのセル範囲と過不足なく一致したかどうかで判定した．

水平走査，垂直走査についての実験結果をそれぞれ図 13，14 に示す．

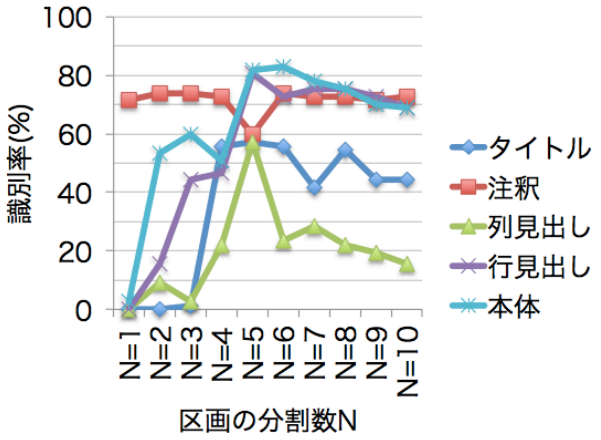

図 13　区画の分割数に対する識別率 (水平走査の場合)

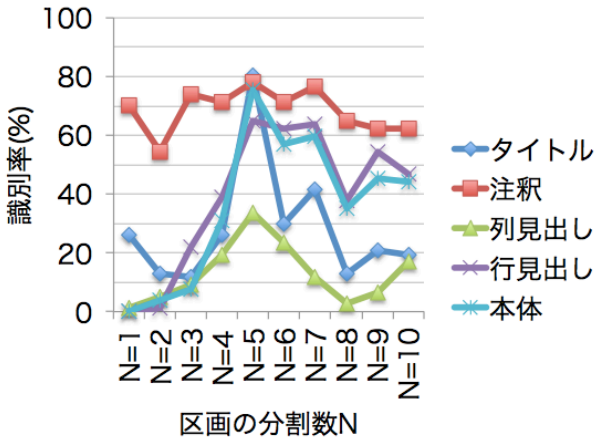

図 14　区画の分割数に対する識別率 (垂直走査の場合)

実験結果から，水平走査，垂直走査ともに多くの項目で区画の分割数 N=5 のときに最も高い精度が得られた．ここで，注釈については，分割数によらずほぼ一定の精度を示している．これは，注釈が記載されていない表がもともと多く存在しており，そのような表では，注釈のセルなしと判断されれば正解と判定されるためと考えられる．

5.2 実験 2: 提案手法およびルールベースに基づく手法による識別率の比較

本実験では，従来手法であるルールベースの手法を方法 1，条件付確率場を用いる提案手法を方法 2 として，両者の識別率を比較する．ただし，方法 2 では，区画の分割数 N=5 とした．

図 15 に，実験結果を示す．

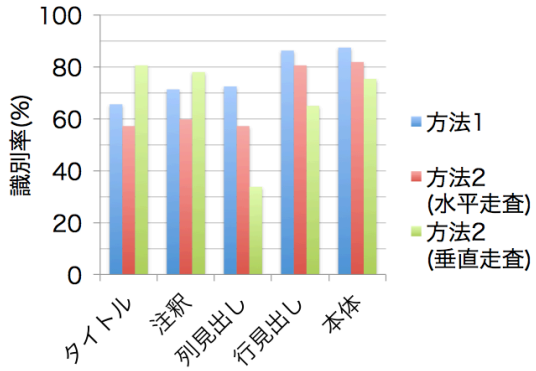

図 15　既存手法と提案手法での識別率の比較

実験結果より，タイトルや注釈においては，垂直走査である方法 2 が最も高い精度を示した．これは，タイトルや注釈は，ルールによる一律な識別が比較的困難な項目であると考えられ，提案手法による条件付確率場を用いた学習により，より柔軟な識別が実現できたためであると考えられる．一方，列見出し，行見出し，本体については，方法 1 が最も高い精度を示した．これらの項目は，セル範囲を認識する必要のある項目であり，提案手法の 2 段階の識別がかえって結果に悪影響を与えた可能性がある．また，提案手法で用いる素性や識別モデルが適切でなかった可能性が考えられる．

また，列見出しについて，方法 2 の識別率が大きく落ち込んでいるのがわかる．これは，列見出しが階層関係で表現されている場合には，空白セルが多く含まれるため，方法 2 ではそれらを誤って'O' と識別してしまったためであることがわかった．

今後は，セル範囲の認識手法や，素性の組み合わせ，識別モデルについて改良していく予定である．

6. まとめと今後の予定

本稿では，統計表データの構造を自動認識する手法を提案した．統計表の構造認識を，系列ラベリング問題として条件付確率場を用いた識別モデルで実現する手法を提案し，従来手法であるルールベースに基づく手法と比較した．

実験の結果，提案手法においては，学習データの生成に用いられる表の区画の分割数 N を 5 とした場合に，ほぼ全ての認識項目で最も高い精度を示すことを確認した．また，ルールベースに基づく従来手法と，条件付確率場に基づく提案手法とでは，識別しやすい項目とそうでない項目があることを確認した．

今後は，素性の種類や組み合わせを変えながら，学習や識別に用いるデータ数をさらに増強して評価実験を行う予定である．また，CSV 形式で記述された統計表データだけではなく，Excel 形式で記述された統計表データに対しても提案手法を適用していく予定である．

また，本稿では，2 次元の表データを 1 次元に変換してから学習するアプローチをとったが，2 次元の形式を活かした学習手法についても今後検討していきたい．

文　　献

[1] 松井, 宮森. ”統計表データを用いた動向情報の根拠探索システムの検討”, 2014 年度情報処理学会関西支部 支部大会 G-02, 2014.

[2] T.G. Kieninger and B. Strieder. T-Recs Table Recognition and Validation Approach. AAAI Fall Symposium on Using Layout for the Generation, Understanding and Retrieval of Documents, 1999.

[3] Yalin Wang, Ihsin T. Phillips, Robert M. Haralick, Table structure understanding and its performance evaluation, Pattern Recognition, Vol.37, 7, pp. 1479-1497, 2004.

[4] 塚本, 増田, 中川. ”HTML の表形式データの変換と携帯端末表示への応用”, 情報処理学会研究報告 2002, 2002.

[5] S. Kerpedjiev, G. Carenini, S.F. Roth, and J.D. Moore. AutoBrief: a multimedia presentation system for assisting data analysis. Comput. Stand. Interfaces 18, 6-7, pp. 583-593, 1997.

[6] C.Wang, X. Xie,W.Wang,W.-Y. Ma. Improving web browsing on small devices based on table classification. Advances in Multimedia Information Processing — PCM, pp.88—95, 2004.

[7] 松下: ”Elucignage : 探索的データ分析のための動向情報可視化インタフェース”, 動向情報の要約と可視化に関するワークショップ第 2 回成果進捗報告会予稿集. pp.17-18, 2007.

[8] 佐伯, 宮森. ”テレビ番組に基づく Web コンテンツの信頼性判断支援システムの提案”, DEIM Forum 2012 B3-5, 2012.

[9] 日本政府, e-Stat 政府統計, https://www.e-stat.go.jp/