複雑な構造をもつ統計表における見出し階層の認識と意味解釈手法

Recognition and semantics interpretation of header hierarchies in statistical tables with complicated structures

曽和　修平†　　宮森　　恒††

† 京都産業大学コンピュータ理工学部　〒 603–8047 京都府京都市 北区上賀茂本山
†† 京都産業大学大学院先端情報学研究科　〒 603–8047 京都府京都市 北区上賀茂本山
E-mail: †g1344739@cse.kyoto-su.ac.jp, ††miya@cc.kyoto-su.ac.jp

あらまし　膨大な量の情報が溢れる社会において，情報の集約，解析は極めて重要な課題である．統計データを分析する際にも，情報の集約は最初の課題となる．更に，近年様々な場面で AI の活用が議論されている．例えば経済産業省は AI を用いて国会答弁の下書きを作成する実験を進めている．このような AI を作る上でも，過去の情報の集約，解析は礎となる．情報の集約，解析を考える際に，Excel 統計表の存在は無視できない．総務省統計局は様々な統計データを Excel 形式で公開している．また，企業においても Excel 統計表の利用は盛んである．しかし，これらは人間が読むために作られたものであることが多く，コンピュータでは扱いにくいといった問題がある．そこで，人間が見やすいように作られた，データの再利用性が低い Excel 統計表を，コンピュータで扱いやすい形式に変換する手法が求められる．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複雑なものが多く，これらを適切に意味解釈するには様々な課題が存在する．本稿では，複雑な構造をもつ統計表を，一旦画像化して解析することにより，表の見た目 (appearance) に応じた，柔軟な見出し階層の認識と，表構造の意味解釈を実現する手法を提案する．また，表の各項目の認識を系列ラベリング問題と捉える従来手法と比較評価する．

キーワード　統計表認識，見出し階層認識，ネ申 Excel，機械学習，勾配ブースティング

1. はじめに

膨大な量の情報が溢れる社会において，情報の集約，解析は極めて重要な課題である．過去の統計データを分析する際にも，情報の集約は最初の課題となる．

総務省統計局は様々な統計データを Excel 形式で公開している．(注1) 企業においても Excel の利用は盛んなため，このような Excel 統計表の形で保存されたデータは非常に多いと考えられる．しかし，Excel 統計表は人間が目で見て読み取る事を前提として作成されたものが多い．このような表はセルの大きさやセルの位置，罫線などの視覚的な要素がデータとしての意味を持つ．そのため，Excel 統計表をコンピュータで扱うために，単純に文字列データに変換してしまうと齟齬が生じる．例えば「2017 年」という大きなセルの下に「1 月」「2 月」... といったセルが存在するとする．この場合，視覚的には「1 月」というセルが「2017 年 1 月」を指している事がわかる．しかし，単純に文字列に変換すると，このような視覚的情報が考慮されず「2017 年，1 月，2 月...」などといったように「2017 年」と「1 月」「2 月」があたかも等価であるかのように変換されてしまう．文字列でこの関係を表すとすれば「2017 年 1 月，2017 年 2 月...」とするのが適切である．よって，コンピュータで扱えるように，視覚的な要素を考慮した上で文字列データを作成するには，人間がデータの再入力などの作業をしなければならない．しかし，大量のデータを処理しようとすると多大な手間，時間がかかってしまうという問題がある．そのため，Excel 統計表を解析し，コンピュータに扱いやすいフォーマットに整形するようなシステムが求められる．しかし，Excel 統計表は，見出し階層の表現形態が多様で複雑なものが多く，これらを適切に意味解釈するには様々な課題が存在する．

主な課題として挙げられるのは

- セル種別の識別
- セルの視覚的な情報の取得
- セルの視覚的な情報，セル種別を元にした見出し階層の認識

である．セルの視覚的な情報には，セルの面積，位置や罫線といったものが考えられる．視覚的な要素がデータとしての意味を持つ Excel 統計表の見出し階層を認識するためにはこのような情報の取得が必要不可欠である．

これらの課題を解決するために本稿では，複雑な構造をもつ統計表を一旦画像化して解析することにより，表の見た目 (appearance) に応じた柔軟な見出し階層の認識をした上で，コンピュータが扱いやすい，より単純な表形式に変換する手法を提案する．比較手法として表の各項目の認識を系列ラベリング問題と捉える従来手法を用いる．そして，提案手法と比較手法

(注1)：https://www.e-stat.go.jp/

の比較評価と，様々な見出し階層に対して提案手法にはどのような課題があるか，を考察する．

2. 関 連 研 究

2.1 情報の集約，解析

奥村 [1] は，以前，政府が公開している統計データを用いてデータ解析を行った際，情報の集約，再解析に非常に手間がかかった経験から，データとしての再利用性の困難な Excel データをネ申 Excel と呼び，問題点を指摘している．解決策として，情報教育としてデータリテラシーを学ぶことや，紙文化からデータ文化への移る運動を活発化させることをあげているが，これらは即座にこの問題を解決するものではない．本稿では，人が作成した機械可読でないデータをコンピュータ側で解釈することで，この問題への解決を試みる．

須永ら [3] は，オープンデータの活用を促進するため，データを見つけやすくする仕組みが必要性であると述べている．また，課題として，機械が扱いやすい形式（CSV や RDF）を作成するためのデータ加工作業を人手で持続的に実施するのに非常に手間がかかる，ということを挙げている．武田ら [4] はデータの再利用性，アクセス容易性が高い LinkedData によるデータ公開の重要性やメリットを述べている．そして，浅野ら [5] は統計表データを LinkedData の記述方法の 1 つである RDF に変換するためのテンプレートを提案している．また，データを解析する際には OLAP システムと呼ばれる複雑な集計，分析を素早く実施することが可能であるシステムが用いられることがある．OLAP では多次元データベースが用いられることが多い．豊島 [6] はマネージメントの問題を例に，RDBMS の限界とデータ全体を対象とした分析のための多次元データベースの必要性について述べている．

本稿では，Excel 統計表を対象に，多大な手間がかかるデータ加工作業を自動化し，機会可読な CSV へ変換する手法を提案する．CSV や JSON より機械可読なフォーマットとして RDF が挙げられる．最終的には Excel 統計表から RDF へ，自動変換可能なものが望ましい．浅野らが提案しているテンプレートを用いて統計表データを RDF に変換する手法では，既存のデータを RDF に変換するにはやはり人手が必要であり，それは大きな手間となる．この問題を解決するには，統計表データを浅野らが提案したテンプレートの形に自動変換するといった方法が考えられるが，そのためには統計表の複雑な見出し階層を認識することが必要不可欠である．そのため，本稿で提案する見出し階層の認識手法は RDF への自動変換技術においても重要な要素の 1 つとなる．また，データの解析に関して，OLAP を用いる場合がある．その際にも見出し階層を自動で認識することで，多次元データベースへの適応が容易になると考えられる．

2.2 表 の 認 識

Kieninger ら [7] は文字情報を元に表を認識する手法を提案している．この手法は文書から表のような構造を持っている部分（ログファイルやディレクトリリストなど）を検出する．各単語の連結を抽出し，その連結の形によって表かどうかを認識するものである．松井ら [8] は統計表の構造認識を系列ラベリング問題として解釈し，条件付き確率場 (CRF) を用いた識別モデルで実現する手法を提案している．しかし，この方法はセル種別はある程度の精度で認識可能なものの，統計表の構造を認識するには至っていない．塚本ら [9] は HTML で記述された表（テーブル）に対して解像度の低い携帯端末でも見やすいものに変換する手法を提案している．その過程で表の見出し階層の認識をした上で階層を平坦化する手法，セルの文字情報，HTML のオプション情報を元に列，行見出しの境界認識をする手法を提案している．

本稿で提案する手法は表の領域を認識するものではないが，セル内の文字情報に加えて座標的な情報も考慮してセル種別を判定している．更に，Excel 統計表の各セルの視覚的な情報を取得することで，HTML などの付加情報を必要とせず複雑な見出し階層の認識に挑戦している点で新しい．付加情報を必要としないため，見出し階層の認識については，セル種別とセルの視覚的な情報を取得，もしくは付与できれば，Excel 統計表に限らず本手法を用いることが可能である．

3. 提 案 手 法

3.1 概　　要

本稿では，複雑な構造をもつ Excel 形式の統計表を，コンピュータが扱いやすい，より単純な構造の統計表に変換する手法を提案する．具体的には，見出し階層の内容を平坦化した統計表を CSV 形式で出力することとする．Excel 統計表は列，行の見出しが階層構造を成しているのが一般的である．例えば，「都道府県」という大きなセルの下に「青森県」といった子のセルが続く，といったような形である．これを「都道府県-青森県」のように 1 つにするのが平坦化である．このように列，行の見出しを全て平坦にした上で CSV 化する．

図 1 は処理の流れの概要図である．前処理では，まず，Excel 統計表を画像化する．次に，セルの領域を抽出し，セル内の文字とセル種別の識別に必要な素性を抽出する．識別器では，各セル種別（タイトルのセル，データのセル等）を識別する．見出し階層の認識では，得られたセル種別やセルの座標情報から見出し階層を認識し，それに基づき平坦化された内容を CSV 形式で出力する．

3.2 前　処　理

前処理部ではまず，Excel 統計表を画像化する．その為に本稿では一旦 Excel 統計表を pdf に出力した後，画像に変換する．pdf に変換する際には，Excel の印刷設定を横 1 ページ，縦自動，印刷オプションで「枠線」にチェックを入れ，枠線が出力されるようにする．その後，ImageMagick(注2) により pdf を画像に変換する．後に OCR をかけるため，できるだけ高い解像度で画像化する必要がある．本稿では解像度を指定する density オプションに 600 を指定した．変換された画像は，Excel 統計表のサイズに応じて複数枚の画像として出力される．

次に，画像化された Excel 統計表から，素性を抽出する．本稿で使用した素性を表 1 に示す．

(注2)：http://imagemagick.org/script/index.php

図 1　処理の流れ

表 1　セル種別の識別に用いた素性

素性名	説明
cell_no	左上のセルから順に付与した連続番号
page	画像の何枚目にセルが属しているか (注3)
x	セルの左上の x 座標
y	セルの左上の y 座標
percentage_x	x を表の面積で割った値
percentage_y	y を表の面積で割った値
width	セルの横幅
height	セルの縦幅
area	セルの面積
normalized_area	セルの面積を表の面積で割った値
lower_left_x	セルの左下の x 座標
lower_left_y	セルの左下の y 座標
upper_right_x	セルの右上の x 座標
upper_right_y	セルの右上の y 座標
lower_right_x	セルの右下の x 座標
lower_right_y	セルの右下の y 座標
text_type	セルのテキストの種類（文字，数字，空白）

素性を抽出するために，画像から各セルの部分を認識する必要がある．画像処理には OpenCV3 を用いた．

各セルの部分を認識するためにまず，各セルの輪郭を抽出する．この際，輪郭上の全点を保持する必要はないので，上下左右の端点のみを保持する．そして，端点の情報からある程度の面積以上の正方形，長方形領域を切り取り，各セル画像に分割する．この時，各セル画像の座標，面積といった情報を得ることが可能である．ただし，注意しなければならないのが 1 つの Excel 統計表は複数枚の画像（表の部分と余白の部分から成る）に変換されるということである．このため，各セルの座標や表の面積を得る際は画像の余白部分を削除した上で，複数枚の画像の座標が連続しているとして計算する処理を加えている．

前処理の最後に，セル内の文字を得るために各セル画像に対して OCR をかける．OCR は数字のセルに対しては tesseract3.04.01(注4)，文字のセルに対しては Google Cloud Vision API(注5) を用いた．これらを使い分けたのは，双方に利点，欠点が存在するためである．まず，tesseract の利点はオープンソースの OCR エンジンであるため，自身のコンピュータで動作することである．Google Vision API を使用するためには通信が必要であり，応答速度で tesseract に劣る．しかし, 認識精度においては Google Vision API の方が高く，tesseract では日本語の正しい認識は難しい. ただし，数字に関しては tesseract であっても正確に認識することが可能であるため，数字のセルの場合は tesseract を用いた．

3.3　セル種別の識別

3.2 節で作成した素性を元に勾配ブースティングによりセル種別を識別する．勾配ブースティングはアンサンブル学習の手法の 1 つで，逐次的に弱識別器を構築していく．新しい弱識別器を構築する際には，前の弱識別器で間違って識別したデータに対して重みを強くした上で学習する．弱識別器に決定木を用いたものは特に GBDT（Gradient Boosting Decision Tree）と呼ばれる．

勾配ブースティングには XGBoost [10] を用いた．識別するセル種別は以下の通りである．

- タイトル
- タイトルの副題
- 列見出し
- 行見出し
- データ
- その他
- 不必要

その他とはタイトルや見出し，データのいずれにも属さない表の注釈などの部分を指す．不必要というのは表の見た目を調節する等のために挿入された空白のセルを指す．

3.4　見出し階層の認識

各セルの座標情報と識別したセル種別から表構造を解析し，CSV 形式に変換する．CSV 形式に変換する際は表の見た目に応じた柔軟な見出し階層の認識が必要である．見出し階層の認識後，その結果に応じて見出し内容を平坦化する．

3.4.1　列見出しの解析と平坦化

列見出しの解析と平坦化について考える．列見出しの構造には様々な種類が存在する．図 2〜5 は代表的な列見出しの構造である．

人　口　Population			世　帯　数　Households		
総　数	男	女	総　数	一般世帯	施設等の世　帯
Both sexes	Male	Female	Total	Private households	(a)

図 2　列見出しの例 (親となる大きなセルがその下の子となる小さなセルを内包する形の見出し階層)

(注3)：ここでの画像とは，Excel 統計表を画像化したものを指す.Excel 統計表を画像化する際，統計表の大きさによっては複数枚の画像として出力される．

(注4)：https://github.com/tesseract-ocr/tesseract

(注5)：https://cloud.google.com/vision/

世帯数	世帯人員	1世帯当たり 人員	1世帯当たり 延べ面積 (㎡)	1人当たり 延べ面積 (㎡)

図 3　列見出しの例 (階層構造を持たない見出し階層)

親族人員が 人						人以上
1	2	3	4	5	6	7

図 4　列見出しの例 (親となるセルと子となるセルしが離れている見出し階層)

総数	男	女	総数	男	女	総数	男	女
Both sexes	Male	Female	Both sexes	Male	Female	Both sexes	Male	Female
09 栃木県			市部			郡部		

図 5　列見出しの例 (子となる小さなセルの下に親となる大きなセルが存在する形の見出し階層)

本稿で扱った Excel 統計表で，割合が多かったのは図 2 のような，親となる大きなセルがその下の子となる小さなセルを内包する形の階層構造である．それに加え，図 3 のような階層構造を持たないものも存在した．図 4 のパターンは 1，2，3，4.. といった文脈的に連続しているセルが「親族世帯が」という親のセルに対する子のセルになっているという例である．図 5 のパターンは親となる大きなセルの下に子となる小さなセルといった図 2 の形の逆で，子となる小さなセルの下に親となる大きなセルが存在し，階層構造を成している．

図 2 の階層構造を例として，本稿で提案する列見出しの解析，平坦化のアルゴリズムについて記述する．この例では最上位にある「人口」の下に「総数」「男」「女」,「世帯数」の下に「総数」「一般世帯」「施設等の世帯」という 2 段の階層構造を成している．この列見出しを平坦化すると，以下のようになる．

人口 Population-総数-Both sexes
人口 Population-男-Male
人口 Population-女-Female
世帯数 Households-総数-Total
世帯数 Households-一般世帯-Private-households
世帯数 Households-施設等の-世帯-(a)

このような列見出しの平坦化は以下の手順で処理する．

（1） ルート見出しの探索

ルート見出しとは，列見出しにおいて，このセル以上，上の部分に座標的に重なった列見出しのセルが存在しないようなセルのことである．このようなセルを探索する．図 2 の例では「人口」と「世帯数」のセルがルート見出しとなる．

（2） ルート見出し以下の子見出しの探索

ルート見出し以下を 1 行ずつ探索し，ルート見出しに属している見出しから多分木を構成していく．ここで「属している」とはセルの左右の端点の x 座標（または上下の端点の y 座標）が，親セル候補の左右の端点の x 座標（または上下の端点の y 座標）内にあるという事を指す．

図 2 の例で多分木を構成すると図 3 のようになる．この多分木を走査し，列見出しを平坦化する．

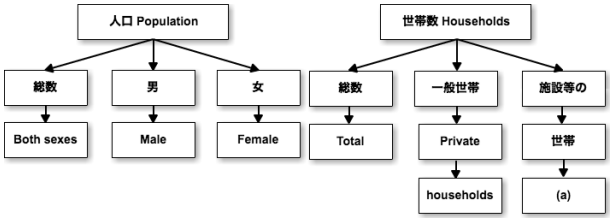

図 6　列見出しの多分木

3.4.2　行見出しの解析と平坦化

次に，行見出しの解析と平坦化について考える．図 7〜9 は代表的な行見出しの構造である．

10	群馬県	Gumma-ken	
	外国人のいる一般世帯数		(e)
		外国人のみ	(f)
		外国人と日本人がいる世帯	(g)
		日本人の親族がいる世帯	(h)
			外国人の親族がいる世帯 (i)
			外国人の親族がいない世帯 (j)
		日本人の親族がいない世帯	(k)
	外国人のいる一般世帯人員		(l)
		外国人のみ	(f)
		外国人と日本人がいる世帯	(g)

図 7　行見出しの例 (1)

全国	Japan
総数	Both sexes
A 親族世帯	
I 核家族世帯	
(1) 夫婦のみの世帯	
(2) 夫婦と子供から成る世帯	
(3) 男親と子供から成る世帯	
(4) 女親と子供から成る世帯	

図 8　行見出しの例 (2)

男	Male
0 〜 4歳	years old
5 〜 9	
10 〜 14	
15 〜 19	

図 9　行見出しの例 (3)

図 7 は代表的な行見出しの例であり，各セルの x 座標によって階層構造を表している．図 8 は各セルの x 座標による階層構造の表現に加え，セル文字の左端の空白によっても階層構造を表している．図 9 は珍しい例で，文字がセルの中心である事で階層構造を表している．

図 8 の階層構造を例として，本稿で提案する行見出しの解析，平坦化のアルゴリズムについて記述する．まず図 8 のような行見出しは以下のように平坦化する事が目標である．

全国-Japan
総数-Both sexes
総数-Both sexes-A 親族世帯
総数-Both sexes-A 親族世帯-I 核家族世帯
総数-Both sexes-A 親族世帯-I 核家族世帯-(1) 夫婦のみの世帯
総数-Both sexes-A 親族世帯-I 核家族世帯-(2) 夫婦と子供から

成る世帯
総数-Both sexes-A 親族世帯-I 核家族世帯-(3) 男親と子供から成る世帯
総数-Both sexes-A 親族世帯-I 核家族世帯-(4) 女親と子供から成る世帯

このような行見出しの平坦化は 1 行ずつ逐次的に解析していく事によって実施する．まず，図 8 の同じ行にあるセルを結合する. 例えば 1 行目は「全国」と「Japan」というセルが同じ行に存在するが，これを「全国-Japan」とし単一のセルとみなす．1 行目は「全国-Japan」で，これはこの状態で完成である．次に 2 行目の「総合-Both sexes」だが，これはセルの左端の x 座標が 1 つ前の「全国-Japan」と等しいので，階層構造にないと判断する. そのため,「総合-Both sexes」もこの状態で完成である．3 行目の「A 親族世帯」は「総数-Both sexes」より左端の x 座標が大きいため，階層構造があるとみなす. そのため,「総数-Both sexes」「A 親族世帯」と平坦化される．

4 行目の「I 核家族世帯」であるが，これは座標的に判断すると「A 親族世帯」と x 座標が同じであるため，階層構造はないと判断されてしまう．しかし，見た目上，セルの左端には空白が存在するため，階層構造が存在するのは明白である．この問題を解決するため，1 文字分の幅以上空白が存在している場合は階層構造が存在するとする．この処理を実現するためには，1 文字分の幅とはどの程度なのか，を推定する必要がある．1 文字の推定幅は，セル内の文字の大きさはセルの高さに比例するという仮定をおき，回帰分析を用いて求める．回帰分析をするには，セルの高さと 1 文字の大きさの組のデータが多数必要である．セルの高さに関しては素性を導出する際に取得可能なので問題ない.1 文字の大きさに関しては別途求める必要があるが，これはラベリング処理により導出した．本稿では，39913 セル分の列見出し，行見出しの文字の大きさをデータとし，回帰直線を求めた．求めた回帰直線から，セルの高さから 1 文字の幅を推定し，先の処理を実現した. 5 行目の「(1) 夫婦のみの世帯」以降の行見出しに関しても先の処理が適用されることにより，行見出しの平坦化が可能である．

4. 評 価 実 験

4.1 セル種別の識別

4.1.1 実 験 設 定

Excel 統計表を適当な CSV 形式に変換するためには，まず各セル種別を判別する必要がある．そのための識別器の評価実験をする．

実験には総務省統計局からダウンロードした Excel 統計表データ，合計 81 ファイルを使用した．セルの数は合計 4498857 セルである．表 2 は実験に用いるセルの各種類の教師データ，テストデータの数である．これらのデータに関して表 1 で示した素性を導出し，XGBoost によりセル種別を識別した．XGBoost のハイパーパラメータは eta(学習率) が 0.3，max_depth(ツリーの最大の深さ) が 6，min_child_weight(木の分割の閾値) が 1，subsample(教師データの抽出割合) が 1，colsample_bytree(各木を作成するときの列におけるサブサンプルの割合) を 1 とした．

表 2　セルの内訳

セル種別	教師データ数	テストデータ数
タイトル	292	142
タイトルの副題	163	83
列見出し	2,475	947
行見出し	25,600	11,006
データ	196,271	84,614
その他	488	221
不必要	95,676	41,000

4.1.2 実 験 結 果

先のデータを用いて XGBoost によるセル種別の識別を行った. 表 3 は識別結果である．タイトル，タイトルの副題，その他はやや識別精度が低いものの，全体的に高い精度で識別が可能であった．Excel 統計表の CSV 化に必要な情報は列見出し，行見出し，データ，不必要のセルであるので，これらの精度が高いことから，セル種別の識別誤りによる CSV 化の失敗は少ないのではないかと考えられる．

表 3　セル種別識別結果

セル種別	精度	再現率	F 値
タイトル	0.972(138/142)	0.972 (138/142)	0.972
タイトルの副題	0.918(78/85)	0.940(78/83)	0.923
列見出し	0.950(910/958)	0.960(910/947)	0.955
行見出し	0.983(10,909/11,102)	0.991(10,909/11,006)	0.987
データ	0.995(84,129/84,512)	0.994(84,129/84,614)	0.995
その他	0.911(164/180)	0.742(164/221)	0.818
不必要	0.987(40,509/41,034)	0.988(40,509/41,034)	0.988

表 4 は各特徴の重要度の上位 5 件である. この図を見るとセルの位置である percentage_x，percentage_y やセルの面積である normalized_area といった情報が識別するにあたって重要な特徴であることがわかる．

表 4　特徴の重要度

素性名	F 値
percentage_y	2,702
cell_no	2,071
normalized_area	1,830
width	1,523
percentage_x	1,461

4.1.3 比 較 実 験

表の各項目の認識を系列ラベリング問題と捉える従来手法を用いて比較実験をする．系列ラベリング問題を解く手法として CRF(条件付き確率場) を用いた. ライブラリは CRFSuit 0.12(注6) を使用した．各統計表の 1 行を 1 系列とし，不必要セルは事前に削除した．素性として用いるものは以下である．

(注6)：http://www.chokkan.org/software/crfsuite/

- セルの文字列
- 垂直方向での区画番号
- 水平方向での区画番号
- セルの文字の種類 (文字，数値，空白)

垂直，水平方向での区画番号とは，統計表を垂直，水平方向に N 分割した時，どの区画に属しているかの番号である．本稿では N=1〜10 でそれぞれ実験を行った．正解ラベルには表 5 の IOB2 タグを用いた．

表 5　正解ラベルの IOB2 タグ

正解ラベル	付与するセル
B_TITLE	タイトルの開始点のセル
I_TITLE	タイトルの 2 セル目以降のセル
B_SUB_TITLE	タイトルの副題の開始点のセル
I_SUB_TITLE	タイトルの副題の 2 セル目以降のセル
B_COL_HEADER	列見出しの開始点のセル
I_COL_HEADER	列見出しの 2 セル目以降のセル
B_ROW_HEADER	行見出しの開始点のセル
I_ROW_HEADER	行見出しの 2 セル目以降のセル
B_BODY	本体 (データ) の開始点のセル
I_BODY	本体 (データ) の 2 セル目以降のセル
B_COMMENT	その他の開始点のセル
I_COMMENT	その他の 2 セル目以降のセル

実験に用いたデータ数は教師データが 23164 系列，テストデータが 5792 系列であり，教師データの系列，テストデータの系列に含まれる，不必要を除く，各タグの数は表 2 と同数である．但し，各タグの数を集計する際には，開始点のセルと 2 セル目以降のセルは区別していない．実験の際は，1 系列の中に含まれる 1 つデータに対して正しい IOB2 タグが付与出来ていれば 1 つの正解としてカウントしている．

図 10 は系列ラベリングによるセル種別識別の結果である．N=6 辺りで最も高い F 尺度を示している事がわかる．表 6 に N=6 の時の詳細な結果を示す．

提案手法の結果と比較すると，全体的に F 値が低くなっていることがわかる．特にタイトル，タイトルの副題の正答率が低いが，これは系列長の差や区画番号といった大まかな位置での素性では区別がつかないためだと考えられる．

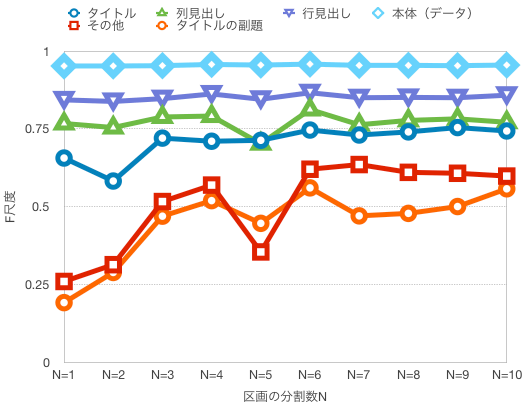

図 10　系列ラベリングによるセル種別識別の結果

表 6　比較手法によるセル種別識別結果（N=6 の場合）

セル種別	精度	再現率	F 値
タイトル	0.854(94/110)	0.662(94/142)	0.746
タイトルの副題	0.755(37/49)	0.446(37/83)	0.561
列見出し	0.926(686/741)	0.724(686/947)	0.812
行見出し	0.940(8,844/9,408)	0.803(8,844/11,006)	0.867
データ	0.948(82,076/86,557)	0.970(82,076/84,614)	0.959
その他	0.767(115/150)	0.520(115/221)	0.620

4.2　統計表の CSV 化

図 14 の Excel 統計表を CSV 化したものが図 15 である．行見出し，列見出しの階層構造の認識は正しく出来ているが，「総数」が「INCA」となっていたりと，全体的に OCR の精度が悪い．他の Excel 統計表に関しても CSV 化を試みたが，やはり OCR の精度の問題は顕著であった．この問題に加えて一部の Excel 統計表では，正しく階層構造の認識ができないものが存在した．このこと関しては考察で言及することとする．

5.　考　　察

実験 1，2 より，セル種別に関しては高い精度で識別が可能になった．これはセルの内の文字種以外に，セルの位置や面積など見た目に関する素性を加えたためだと考える．

一方で，CSV 化に関しては OCR と見出し階層の認識精度が課題である．まず，OCR であるが列見出しや行見出しに含まれる日本語や英語の認識精度が低い．ただし，図 14 の「高齢夫婦と未婚の 18 歳未満の者から成る世帯」などの見出しとしては長い文に対しては正確に認識が可能である．これは Google Cloud Vision API が文脈による補正を行っているためだと推測する．認識精度が特に低いのは数文字程度の短い文である．この問題の解決方法として，Excel のフォントは一括で変換可能なため，特定のフォントに特化した OCR エンジンを独自開発し，短い文に対してはこれを使用する，そして，長い文に対しては，Google Cloud Vision API を使用する，などが考えられる．筆者はこの問題を解決するため TemplateMatching による特定フォントに特化した OCR エンジンを開発しているが，現時点では代替可能な程の精度が出ていない．

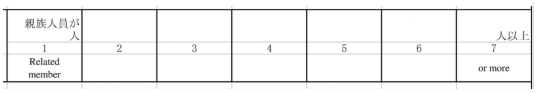

図 11　階層構造の認識が正しくできなかった列見出し例（1）

次に，見出し階層の認識精度に関してである．いくつかの例で階層構造を正しく認識する事ができなかった．まず，図 11 のような列見出しである．この例では，ルート見出し「親族人員が」という部分に対して階層構造を成しているのはその真下の「1」と「Related member」のみだと解釈されてしまう．本来ならば「2」以降の数字のセルも階層構造を成していると判断すべきである．更に，最後の「7」という部分は「親族人員が-7-人-以上-or more」のように解釈されて欲しい．これは座標的な

階層構造の解釈以外にも文脈的な解釈が必要であるということを示している．座標的に見たときに「親族人員が」というセルと階層構造を成しているのは真下の「１」「Related member」だが，文脈的に見たとき,「1」と「2」以降の数値は同じ種類に属するものだと判断し,「1」と同様の階層構造を持っていると解釈するということである．

	A 親族世帯				
	I 核家族世帯			Family nuclei	
総数					
	総数	(1) 夫婦のみ	(2) 夫婦と	(3) 男親と	(4) 女親と

図 12　階層構造の認識が正しくできなかった列見出し例（2）

図 12 のような列見出しでも正しく見出し階層の認識ができなかった．この例では，濃い罫線でセルを分割しているが，セルの結合が正しくされておらず「A 親族世帯」の左右のように薄い罫線が残っている．よって，薄い罫線は無視し，濃い罫線のみよってセルが分割されていると解釈すべきなのだが，本稿では薄い罫線と濃い罫線の区別していない．そのため「A 親族世帯」というセルと「(4)」「女親と」というセルは座標的に階層構造を成していないと誤った解釈してしまっている．更に,「I 核家族世帯」と「Family nuclei」というセルは本来 1 つのセルとして解釈するべきものだが，これも別々のセルとして解釈されている．この問題は一概に薄い罫線を無視し，濃い罫線を基準にセル分割をすれば解決するというものではない．濃い罫線の引き方は Excel 統計表の作成者に依存するため，必ずしも濃い罫線を基準にセルが分割されているとはいえないためである．この問題を解決するには，Excel 統計表によってセル分割の基準を濃い罫線にするか，薄い罫線にするかの判断をするアルゴリズムの提案が求められる．

65歳以上の高齢者１人と未婚の				(b)
18歳未満の者から成る世帯				
	一戸		建	Detached houses
	長屋		建	Tenement houses
	共同住		宅	Apartment houses or flats
	建物全体の階		数	Building
		1 ・ 2 階	建	stories
		3 ～ 5		
		6 ～ 10		
		11 階建以上		and over

図 13　階層構造の認識が正しくできなかった行見出し例

図 13 は見出し階層の認識が正しくできなかった行見出しの例である．これは「65 歳以上の高齢者 1 人と未婚の」と「18 歳未満の者から成る世帯」を別々のセルとして認識してしまうため，以下の階層構造を「18 歳未満の者から成る世帯」-「一戸」というようにしてしまう誤りである．本来ならば「65 歳以上の高齢者 1 人と未婚の 18 歳未満の者から成る世帯」で 1 つのセルとするべきである．この場合でも，文脈による解釈をした上でセルを結合するようなアルゴリズムの提案が求められる．

6. ま と め

本稿では Excel 統計表のデータの集約，解析の手間を削減するため，Excel 統計表をコンピュータに扱いやすいフォーマットである CSV に適当な形で変換する手法を提案した．実験により，セル種別は高い精度で識別でき，見た目による見出し階層の認識についてもある程度可能である事がわかった．

しかし，CSV 化に関して OCR や見出し階層の認識については幾つか課題が残った．数値以外の日本語，英語の文字に対して今回，OCR エンジンとして，Google Cloud Vision API を採用したが，特に短い文に関して精度が悪い．これは短い文に対しては特定フォントに特化した OCR エンジンを独自で作成すれば良いのではないかと考える．もしくは，pdf をテキスト化し分割したセル画像に対して完全に対応が取れるような手法が存在すれば OCR の問題は解消される．

見出し階層の認識については考察で述べたように幾つかの例で誤りが見られた．見出し階層の認識の精度を高める方法としては，文脈による解釈とセル分割基準（濃い罫線，薄い罫線のどちらを採用するか）の使い分けが考えられる．これらの手法を導入し，OCR や見出し階層の認識の精度が高まれば，より汎用的に Excel 統計表の CSV 化が可能になることが期待できる．

謝　辞

本研究は京都産業大学総合学術研究所の研究活動 によるものです．

文　献

[1] 奥村晴彦 (2013) “「ネ申 Excel」問題” “http://oku.edu.mie-u.ac.jp/˜okumura/SSS2013.pdf” (最終閲覧日:2016/12/31)
[2] 経済産業政策局 (2015) “ビッグデータ･人工知能がもたらす経済社会の変革” “http://www.meti.go.jp/committee/kenkyukai/sa-nsei/kaseguchikara/pdf/010_03_03.pdf” (最終閲覧日:2016/1/7)
[3] 須永 博行，三浦 仁，山崎 太郎 (2014) “オープンデータの活用を促進するための仕組み” “UNISYS TECHNOLOGY REVIEW 第 121 号，SEP. 2014”
[4] 武田 英明，嘉村 哲郎，加藤 文彦，大向 一輝，高橋 徹，上田 洋 (2011) ‘日本における Linked Data の普及にむけて” “The 25th Annual Conference of the Japanese Society for Artificial Intelligence， 2011 ”
[5] 浅野 優，岩山 真，武田 英明，小出誠二，加藤 文彦，小林巌生 (2013) ‘統計データの RDF 化のためのテンプレート” “第 12 回情報科学技術フォーラム F-034”
[6] 豊島 一政 (1996) ‘多次元データベースと RDB（OLAP の詳解）” “情報処理学会第 52 回全国大会 4-157”
[7] T.G. Kieninger and B. Strieder(1999), “T-Recs Table Recognition and Validation Approach” AAAI Fall Symposium on Using Layout for the Generation， Understanding and Retrieval of Documents.
[8] 松井侑祐，宮森 恒 (2015) “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” DEIM Forum 2015 B4-5.
[9] 増田英孝，塚本修一，安富大輔，中川裕志 (2003) “HTML の表形式データの構造認識と携帯端末表示への応用’ 情報処理学会研究報告.
[10] Tianqi Chen，Carlos Guestrin(2016) “XGBoost: A Scalable Tree Boosting System” “https://arxiv.org/pdf/1603.02754.pdf” (最終閲覧日:2016/12/31)

地域，住宅の建て方（7区分） Area and type of building (7 groups)	主世帯数 Number of principal households	1人当たり延べ面積（㎡） (a)	（再掲）夫婦とも65歳以上の高齢夫婦世帯 (b) 主世帯数 Number of principal households	1人当たり延べ面積（㎡） (a)	（別掲）いずれかが65歳以上の夫婦のみの世帯 (c) 主世帯数 Number of principal households	1人当たり延べ面積（㎡） (a)	（別掲）いずれかが60歳以上の夫婦のみの世帯 (d) 主世帯数 Number of principal households	1人当たり延べ面積（㎡） (a)
全国 Japan								
総数 Total	3,629,622	54.5	2,800,324	54.5	3,937,345	54.0	5,225,767	53.5
一戸建 Detached houses	2,991,904	59.8	2,318,956	59.6	3,209,494	59.6	4,205,996	59.3
長屋建 Tenement houses	136,140	28.2	105,759	28.1	152,239	27.9	202,954	27.8
共同住宅 Apartment houses or flats	494,755	30.0	370,477	30.0	567,948	29.6	806,136	29.4
建物全体の階数 Building stories								
1・2階建	77,146	27.5	59,614	27.9	90,597	26.6	127,194	25.6
3～5	241,962	28.6	180,432	28.7	276,375	28.4	390,386	28.2
6～10	100,436	33.1	75,027	33.0	114,562	32.9	163,493	32.8
11階建以上 and over	75,211	32.8	55,404	32.8	86,414	32.7	125,063	32.7
（再掲）								
世帯が住んでいる階 Living floor								
1・2階 floor	248,091	28.7	190,734	28.8	283,428	28.2	390,244	27.8
3～5	165,011	30.0	119,589	30.0	190,745	29.7	280,097	29.6
6～10	66,149	33.7	48,897	33.7	75,718	33.5	109,262	33.4
11階以上 and over	15,504	35.2	11,257	35.2	18,057	35.0	26,533	34.9
その他 Others	6,823	45.9	5,132	46.7	7,664	45.1	10,681	43.4
（別掲）								
高齢夫婦と未婚の18歳未満の者から成る世帯 (e)	15,296	35.9	11,192	36.5	22,186	33.9	47,313	31.5
一戸建 Detached houses	12,842	39.3	9,525	39.7	17,141	38.6	33,592	37.2
長屋建 Tenement houses	603	16.8	411	17.0	1,007	16.2	2,240	16.1
共同住宅 Apartment houses or flats	1,829	18.6	1,241	18.9	3,990	18.2	11,341	17.8
建物全体の階数 Building								
1・2階建 stories	298	17.1	204	17.1	661	15.4	1,952	14.3
3～5	952	17.7	637	18.0	1,945	17.5	5,377	17.2

図 14　Excel 統計表の例

	主世帯数--Number of-pri	1人当たり-延べ面積-(r	夫婦とも65歳以上の-ま	夫婦とも65歳以上の-い	いずれかが65歳以上の-う	いずれかが65歳以上の	いずれかが60歳以上(	いずれかが60歳以上の-夫婦のみの世帯-1人当たり-延べ面積-(ni)
INCA-Total	3629622	54.5	2800324	54.5	3937345	54	5225767	53.5
INCA-建戸-Detached houses	2991904	59.8	2318956	59.6	3209494	59.6	4205996	59.3
INCA-建屋長-Tenement houses	136140	28.2	105759	28.1	152239	27.9	202954	27.8
INCA-共同住宅-Apartment houses or flats	494755	30	370477	30	567948	29.6	806136	29.4
INCA-共同住宅-建物全体の階数-階比1-建-Stories	77146	27.5	59614	27.9	90597	26.6	127194	25.6
INCA-共同住宅-建物全体の階数-階比1-53	241962	28.6	180432	28.7	276375	28.4	390386	28.2
INCA-共同住宅-建物全体の階数-階比1-~10	100436	33.1	75027	33	114562	32.9	163493	32.8
INCA-共同住宅-建物全体の階数-11階建以比-LE-and over	75211	32.8	55404	32.8	86414	32.7	125063	32.7
INCA-共同住宅-世帯が住んでいる階比-21-階比-floor	248091	28.7	190734	28.8	283428	28.2	390244	27.8
INCA-共同住宅-世帯が住んでいる階比-21-53	165011	30	119589	30	190745	29.7	280097	29.6
INCA-共同住宅-世帯が住んでいる階比-21-~10	66149	33.7	48897	33.7	75718	33.5	109262	33.4
INCA-共同住宅-世帯が住んでいる階比-11階以比-LE-and over	15504	35.2	11257	35.2	18057	35	26533	34.9
INCA-共同住宅-世帯が住んでいる階比-11階以比-LE-Others	6823	45.9	5132	46.7	7664	45.1	10681	43.4
高齢夫婦と未婚の18歳未満の者から成る世帯(e)	15296	35.9	11192	36.5	22186	33.9	47313	31.5
高齢夫婦と未婚の18歳未満の者から成る世帯(e)-建戸-Detached houses	12842	39.3	9525	39.7	17141	38.6	33592	37.2
高齢夫婦と未婚の18歳未満の者から成る世帯(e)-建屋長-Tenement houses	603	16.8	411	17	1007	16.2	2240	16.1
高齢夫婦と未婚の18歳未満の者から成る世帯(e)-共同住宅-Apartment houses or flats	1829	18.6	1241	18.9	3990	18.2	11341	17.8
高齢夫婦と未婚の18歳未満の者から成る世帯(e)-共同住宅-建物全体の階数-階比1-建-Stories	298	17.1	204	17.1	661	15.4	1952	14.3
高齢夫婦と未婚の18歳未満の者から成る世帯(e)-共同住宅-建物全体の階数-階比1-53	952	17.7	637	18	1945	17.5	5377	17.2
高齢夫婦と未婚の18歳未満の者から成る世帯(e)-共同住宅-建物全体の階数-階比1-6~10	330	20.9	224	21.2	789	20.6	2333	20.3
高齢夫婦と未婚の18歳未満の者から成る世帯(e)-共同住宅-建物全体の階数-11階建以比-上-and over	249	20.6	176	20.8	595	20.5	1679	20.4
高齢夫婦と未婚の18歳未満の者から成る世帯(e)-共同住宅-世帯が住んでいる階-21-階比-floor	971	17.7	677	17.8	1991	17.1	5524	16.5
高齢夫婦と未婚の18歳未満の者から成る世帯(e)-共同住宅-世帯が住んでいる階-21-53	587	18.9	379	19.3	1355	18.4	3939	18.1
高齢夫婦と未婚の18歳未満の者から成る世帯(e)-共同住宅-世帯が住んでいる階-21-10	215	21.1	145	22	511	21.3	1498	20.7
高齢夫婦と未婚の18歳未満の者から成る世帯(e)-共同住宅-世帯が住んでいる階-11階以-上-and over	56	21.3	40	21.4	133	22.2	380	22
高齢夫婦と未婚の18歳未満の者から成る世帯(e)-他のそ-Others	22	29	15	29.1	48	26.7	140	24

図 15　CSV 化の例