Web 上の表データの論理構造の発見

大西　香織†　　田島　敬史†,††

† 北陸先端科学技術大学院大学情報科学研究科 〒 923-1292 石川県能見市旭台 1-1
†† 京都大学情報学研究科 〒 606-8501 京都市左京区吉田本町
E-mail: k-oonisi@jaist.ac.jp, tajima@i.kyoto-u.ac.jp

あらまし　Web 上の大きな表形式のデータを携帯端末等の小さな画面上で見やすく表示する方法として，表中の各セルを一つづつ，そのセルに対応する項目名とともに表示するという方法がある．その場合，各セルに対応する項目名をいかにして発見するかという点と，Web 上の表に頻繁に見られる，複数の行や列にまたがるセルの扱いが問題となる．そこで，本研究では，そのような複数行・列にまたがるセルが表現している論理上の構造を自動的に推定し，この情報を利用して，これらのセルの適切な分割および各セルに対応する項目名の発見を行う手法を提案する．

キーワード　HTML，テーブル，セル，項目名，携帯端末，表示

1. ま え が き

現在，Web 上には大量の有用な情報が存在している．また，近年は，インターネットへの接続が可能な携帯電話や携帯端末等のモバイル機器の普及とともに，これらのモバイル機器からの Web へのアクセスも急増している．これらのモバイル機器からの Web へのアクセスと，従来の計算機環境からの Web へのアクセスとを比較した場合，これまで，両者の間には以下のような様々な違いが存在していた．

- **ネットワーク速度の違い**: 現在では，一般家庭においても，高速なインターネットアクセスが可能となっており，Web 上のコンテンツも，そのような高速インターネット接続を前提とした物が主流となっているが，携帯電話等からのインターネット接続は，そのようなブロードバンドサービスによる接続速度に比べると，はるかに遅い環境がまだまだ主流である．
- **メモリ容量や計算能力等の計算資源の量の違い**: Web 上の HTML データの表示にはレンダリングと呼ばれる処理が必要だが，この処理のためのソフトウェアは現在では非常に大きな物になっており，それに応じたメモリサイズが必要となる．また，ユーザにストレスを与えない速度でレンダリング処理を行うために必要となる計算能力も非常に高い．現在では，安価なパーソナルコンピュータでも，これらの処理に十分なメモリ量と CPU 性能を備えるようになっているが，携帯端末，特に携帯電話では，価格，機器のサイズ，電源消費量の問題などから，搭載可能なメモリ量や CPU 性能に限界があり，これらの処理に十分なメモリ量や CPU 性能を持たせることが難しい．
- **表示画面サイズの違い**: 携帯端末などのモバイル機器は，その可搬性のために機器のサイズに制限があり，そのため，多くの機器では，画面サイズが，現在，パーソナルコンピュータ等で主流になっている画面サイズの数分の一から数十分の一程度である．そのため，一般のパーソナルコンピュータ等での表示を想定して作成された HTML データや画像データを表示する場合，非常に見づらい表示となる場合がある．

しかし，これらの点のうち，一つ目のネットワーク速度の違いについては，現在，第三世代の携帯電話が普及しつつあり，近い将来には，さらに高速のインターネットアクセスが携帯電話から可能となる見込みである．そのため，携帯機器からの Web アクセスにおいて従来存在していたネットワーク速度の違いから来る問題は解消されつつある．同様に，二つ目の計算資源の問題についても，メモリの低価格化やバッテリの改良などにより，携帯端末や携帯電話に搭載されるメモリ量や CPU 性能は，年々，向上しており，携帯端末においては，数年前から，通常のパーソナルコンピュータ用の Web ブラウザソフトを搭載するものが登場し，また，携帯電話においても，一昨年，通常のパーソナルコンピュータ用の Web ブラウザソフトとほぼ同等のものを搭載する機器が登場した．

このように，上で挙げた三つの点のうち，最初の二つの点については，現在では解消されつつある．そのため，今後は，通常の PC 用に作成された Web 上のデータに対する携帯端末や携帯電話からのアクセスが，ますます増加するものと思われる．しかし，一方で，上に挙げた三つのうちの最後の点については，この違いが，携帯端末や携帯電話に必須である可搬性を実現するための物理的なサイズ上の制約から来ていることから，今後も，簡単には解消されないと思われる．そのため，モバイル機器からの Web アクセスにおいて，この点から生じる様々な問題の解消が，いっそう重要になると考えられる．

一方，携帯端末や携帯電話等のモバイル機器からアクセスできると便利であるような Web 上の情報としては，以下のようなものが特に需要が高いことが予想される．

- ニュース，天気予報，交通情報などの最新の情報
- 地図情報
- 様々な，時刻表，料金表，スケジュール表など

これらのうち，一つ目のニュース，天気予報，交通情報等の情報については，複雑なレイアウトを伴わない比較的短い文章や，比較的小さい画像によって表現されていることが多く，また，画面の小さい携帯端末向けの専用のサイトも多いことから，画面サイズの小ささは，比較的，大きな問題にならないと思われる．また，二つ目の地図情報の画面表示に関しては，携帯端

末上ので表示に限らず通常の計算機環境での表示においても，地図を様々な縮尺で見易く表示するための技術の研究が，数多く行われている．一方，三つ目に挙げたような情報は，Web 上に表形式で掲載されていることが多く，このような表形式のデータは，小さい画面での表示が特に問題となるデータである．そこで，本研究では，このような表形式のデータ，特に画面サイズに比べて大きい表データを，小さい画面でも見易い形で閲覧できるようにする手法を開発する．

大きな表データを小さな画面で表示する場合，最も単純な方法としては，表のうちの画面に入る範囲の縦横数個 (あるいは一個) のセルのみを表示し，ユーザに上下左右にスクロールさせるという方法が考えられる．しかし，このような手法を取る場合，以下のような二つの問題点が生じる．

まず，第一の問題点として，多くの表は，各列や各行が何の情報を表しているかを表す項目名にあたる列や行と，それらの項目名に対応するデータをあらわしている列や行からなり，項目名にあたる行や列が画面の範囲外にある状態で，データ部分のみを表示すると，非常に意味がわかりにくくなる．この問題に対する解決策として, [1] では，各セルに対応する項目名にあたるものがどれかを自動判定し，各セルの内容をそれに対応する項目名とペアで表示してやるという手法を提案している．

また，第二の問題点として，Web 上で頻繁に見られる，複数の行や列にまたがるセルを含む表をどのように表示するかが問題となる．単純には，そのようなセルは一列一行のセルの集合に分割し各セル中に元のセルの内容をコピーして，格子状の表に変換するという方法が考えられるが，このような単純な分割は，必ずしもそのセルが持つ論理構造に即していない場合があり，その場合，分割によって表がますます読みにくくなりうる．

一方で，第一の問題点に対する解決策として，前述のような各セルに対応する項目名の自動判定を考える場合，第二の問題点であげた複数行・列にまたがるセルの構造は，実は項目名にあたる領域の発見に非常に有用な情報を含んでいる．

そこで，本研究では，複雑なセルの構造を含む表から，それらのセルが表現している論理上の構造を自動的に推定し，その情報を利用して，

- 表中の項目名にあたる部分を判定し
- 複数行や複数列にまたがるセルの適切な分割を行い

その上で，各セルを対応する属性名とともに画面中に表示し，ユーザにスクロールさせるという表示方法を開発する．

以下，第 2 章では関連研究について説明し，第 3 章で，Web 上の表データを小さな画面で表示する場合の問題点について詳しく述べる．第 4 章では，表の項目名にあたる領域とデータにあたる領域の性質について考察し，続いて，第 5 章で本研究のアプローチを説明する．第 6 章では，本研究で提案する表中の論理構造の推定手法を実際の Web 上の表データに適用した場合の現時点までの簡単な実験結果について報告し，最後に，第 7 章でまとめと今後の課題について述べる．

2. 関 連 研 究

表の自動的な認識や解釈に関しては，古くから多くの研究があり，サーベイ論文として [2] があるが，印刷物のスキャン画像やテキストファイルの中に現れる，表にあたる領域やその表の構造を発見する研究が多く，表の物理構造はわかった上でその論理構造を抽出する研究は比較的少ない．これまでの研究では，項目名とデータの境界，対応関係の発見に，セルの色やフォントなどの情報 [3]，隣接セル間の文字数，形式，内容等の観点からの類似度 [1], [4]〜[6]，セル内の語が属性名，属性値の各々に使われる確率 [7] 等を用いる手法が提案されている．一方，複数の行や列にまたがるセルの構造は，項目名にあたる部分を発見するために有用な情報を含んでいるが，この情報を用いる手法は，一部の論文において,「表の上部に下部より列数が少ない行がある場合は項目名である可能性が高い」[8] 等のヒューリスティックスが他の手法による判定の補助として用いられている程度で，セルの構造を主に用いるという手法は見当たらない．

Web 上の大きな表の小画面での表示に関しては，前述の [1] が各セルを対応する項目名とともに表示する手法を提案している．しかし，複数の行や列にまたがるセルがある場合には，これを単純に格子状に分割しているため，階層構造を表現するセルなど複雑なセルに対しては，不適切な分割を行ってしまう．

また，複数の行や列にまたがるセルによって表現された表中の論理構造の解釈に焦点を当てている研究としては，田仲による研究 [9] がある. [9] では，Web 上の大量の表から効率良くオントロジーを獲得するために，ユーザが表中の論理構造の解釈を文法規則の形で与え，これに基づいて計算機に大量の表を処理させる手法について提案している．よって, [9] では，論理構造の解釈はユーザが最初に与える必要があり，表中の構造の自動的な判定については考えていない．

3. 複数行・列にまたがるセルの問題点

小さい画面上で，画面に一度には入らないような大きな表データをスクロールさせる場合，先に述べたように，複数の行や列にまたがったセルの存在が問題となる．この章では，複数行・列にまたがるセルにはどのようなものがあるかについて概観するとともに，それらのセルをどのように取り扱うのが適切かの基本的な方針について考察する．複数行・列にまたがるセルには大きく分けて，値を共有するセルが併合されている場合と，なんらかの論理構造を表現している場合の二つの場合がある．以下，これらについて詳しく説明する．

3.1 値を共有するセルの併合の例

値を共有することを表すセルの併合は，さらに，各セルが同じ値を持つことを表すセルの併合と，複数セルの値の集約値を表すためのセルの併合にわけられる．後者の集約を表すセルの併合については後で述べることとし，ここでは，まず前者について考える．そのような，値が同じであるセルが併合されることで生じる，複数列・行にまたがるセルを含む表の例としては，例えば，図 1 に示すような表が考えられる．この例では，二列目の二つ目のセルは二〜四行目に，三列目の二つ目のセルは二〜五行目にまたがっている．このようなセルは，読み易さのために同一の内容を持つセルを併合したものと考えられる．よって，小さい画面上でこの表のセル一つづつを表示してスク

入学試験実施要領

学部	試験日	試験会場
法学部	2 月 18 日	本部キャンパス
文学部		
経済学部		
工学部	2 月 19 日	
医学部	2 月 18 日	医学部キャンパス

図 1　複数行にまたがるセルの例

Fig. 1　Cells Spanning Multiple Lines

入学試験実施要領

学部	試験日	試験会場
法学部	2 月 18 日	本部キャンパス
文学部	2 月 18 日	本部キャンパス
経済学部	2 月 18 日	本部キャンパス
工学部	2 月 19 日	本部キャンパス
医学部	2 月 18 日	医学部キャンパス

図 2　複数行にまたがるセルの分割の例

Fig. 2　Decomposition of Cells Spanning Multiple Lines

主要駅からのひかり・こだま特急料金

	品川	新横浜	小田原	熱海	三島	新富士
東京発	840		2190			2920
新横浜発	-		950			

図 3　複数行・列にまたがるセルの例

Fig. 3　Cells Spanning Multiple Lines/Columns

主要駅からのひかり・こだま特急料金

	品川	新横浜	小田原	熱海	三島	新富士
東京発	840	840	2190	2190	2190	2920
新横浜発	-	-	950	2190	2190	2920

図 4　複数行・列にまたがるセルの分割の例

Fig. 4　Decomposition of Cells Spanning Multiple Lines/Columns

ロールする場合，併合されたセルを，同じ内容を持つ複数のセルに分割して考えればよいと考えられる．例えば，図 1 は，図 2 のように分割した上で，一セルずつスクロールしてやればよい．

同様に，複数の列にまたがるセルを含む表や，列と行の双方についてまたがるセルを含む表も存在し，例えば，図 3 のような表が考えられる．このような表を小さい画面上で表示する場合も，同様に考えて，図 4 のように分割して考えればよい．

より一般化して言えば，複数の行や列にまたがったセルがある場合には，全てのセルを同じ列にある最も横幅の小さいセル，同じ行にある最も縦幅の小さいセルにあわせて分割すれば，完全な格子状の表に変形することができる．

3.2　論理構造を表現するセルの例

しかし，複数の行や列にまたがるセルは，様々な論理構造を表現している場合があり，そのような単純な分割の仕方は，必ずしも常に，最も適切な分割にはならない．例えば，図 5 のような例を考える．この表では，項目名にあたる部分のセルの一部が，項目の階層的な分類に対応する構造を表現している．この表に対して，前述の例と同様に，複数列や複数行にまたがるセルを，同じ行や同じ列にある最も小さいセルに合わせて分割することで格子状の表に変形した場合，図 6 のような表となる．

過去の進路状況

年度	民間企業	公務員 国家	公務員 地方	進学 本学	進学 他学
2002	83	23	12	18	5
2003	76	25	16	22	2
2004	88	18	14	16	3

図 5　ある論理構造を表現しているセルの例

Fig. 5　Cells Representing Some Semantic Structure

過去の進路状況

年度	民間企業	公務員	公務員	進学	進学
年度	民間企業	国家	地方	本学	他学
2002	83	23	12	18	5
2003	76	25	16	22	2
2004	88	18	14	16	3

図 6　ある論理構造を表現しているセルの単純な分割

Fig. 6　Naive Decomposition of Cells for Semantic Structure

過去の進路状況

年度	民間企業	公務員 国家	公務員 地方	進学 本学	進学 他学
2002	83	23	12	18	5
2003	76	25	16	22	2
2004	88	18	14	16	3

図 7　ある論理構造を表現しているセルのより適切な分割

Fig. 7　Preferable Decomposition of Cells for Semantic Structure

しかし，元の表のセルの構造が，前述のように，階層的分類を表現していることを考えれば，この表を格子状の表に変形する方法としては，図 7 の表のように変形するのがより適切と考えられる．すなわち，一列目や二列目の一行目と二行目のセルが併合されているのはそのまま残し，一行目の三列目と四列目が併合されているセル (「公務員」のセル) を二つに分割し，かつ，その内容を二行目の三列目 (「国家」) と四列目 (「地方」) の内容と連結してしまうという変形である．一行目の五列目，六列目のセル (「進学」) についても同様である．

3.3　両者の本質的な違い

では，単純に図 2，4 のように分割すればよい場合と，図 5 のようなそうでない場合との違いは何であろうか．上の例だけを考えれば，階層構造をなしているセルがあった場合には，それを階層的分類を表していると考えて，特別扱いするという考え方もできる．しかし，後述のように，論理的な構造を表すセルには，階層構造以外にも，入れ子になっている表や，見出しを表すセルや区切り記号を表すセルなど様々なものがある．

図 2，4 のような場合と，図 5 のような場合の，より本質的な違いは以下のように考えられる．図 2 においては，二列目の二行目以降のセルは，全て「試験日」という項目名に対応する項目を表すセルであり,「日付」という同種の値を含むべきはずのセルである．よって，併合されている二列目の二〜四行目も，同じ値をコピーして分割すればよい．同様に，図 4 の場合では，

過去の進路状況

年度	民間企業	公務員 国家	公務員 地方	進学 本学	進学 他学
2002	83	23	12	18	5
2003	76	25	16	22	2
2004	88	18	14	16	3

図 8　論理構造を表現しているセルの分割の途中段階
Fig. 8　Intermediate Phase of Cell Decomposition

入学試験実施要領

学部	試験日	試験会場
法学部	2 月 18 日	本部キャンパス
文学部		
経済学部		
工学部	2 月 19 日	
医学部	未定	

図 9　異なる項目名に対応するセルの併合の例
Fig. 9　Merging of Cells for Different Kind of Items

二行目以降，二列目以降の全てのセルは，全て値段を表す数字が入るべきセルになっている．

一方，図 5 の表では，一列目の一行目と二行目のセルが併合されているが，これらのセルは，ある項目名に対応する同種の値が入るべきセルというわけではない．そのため，このセルを二つに分割し，双方に「年度」という値をコピーして格納するのが不適切なのだと考えられる.「民間企業」のセルについても同様である．これらのセルについては，二つのセルが併合されているのではなく，逆に，三列目以降の一〜二行目のセルが，本来,「項目名」に対応する一つの行であるべきだったものが，その値の中にさらに構造を持っているために，その構造に対応する二つの行に分割されている物だと考えられる．

しかし，一行目，三〜四列目の「公務員」のセルについては，これら二つのセルはどちらも「項目名」にあたるものを格納すべきセルであると考えられ，よって,「公務員」という値をコピーして二つに分割すればよいと考えられる.「進学」のセルについても同様である．ここまでの変形を行った時点では，図 8 に示したような表になる．

この表を，さらに最終的な格子状の表へと変形するには,「年度」や「民間企業」のセルを二行に分割するのではなくて，代わりに，三列目以降の一行目と二行目のセルを併合してやればよい．これは，前述のように，一行目と二行目は項目名を表す値の内部構造に対応する行であり，三列目以降の項目名としては，一行目と二行目の値を結合すればよいと考えられるからである．以上の変形を行うと，結果は，前述の図 7 のようになる．

以上のように考えると，セルの適切な分割を行うには，表中のどのセルが項目名を表しており，その項目名に対応していて同種の内容を含むべきセル群がどれであるかを判定する必要があるということがわかる．そこで本研究では，複数の行や列にまたがるセルを含む表に対して，項目名にあたる部分がどの部分かを判定し，その結果に基づいて適切なセルの分割，併合を行い格子状の表へと変形する手法を開発する．

4.　項目名のセルとデータのセル

前章で述べたように，本研究では，表中のどの部分が項目名にあたり，どの部分がその項目名に対応するデータを表すセル群にあたるかの判定に基づいて表の変形を行う．そこで，本章ではまず，項目名と項目内容とは何かについて考察する．

4.1　項目名という概念の定義

まず，最も単純な表の例として，図 1 に示す表を考える．この場合，項目名は「学部」「試験日」「試験会場」であり，一〜三列の二行目以降が，これらの項目名に対応するデータを表すと考えられる．ここで,「法学部」「文学部」なども，対応する各行がその学部に関する情報を記述していると考えると，項目名にあたるのではないかという考え方もあると思われる．しかし，本研究においては「ある項目名に対するデータを表すセル群」として判定したいのは，前述のように，同種の値を格納していて，同じ値を持っている場合は併合され得るようなセル群である．その観点から考えると，行方向の「2 月 18 日」のセルと「本部キャンパス」のセルの並びは，明確に異なる種類の値を格納してあり，これらが同じ値を持つとの理由で併合されることは通常起こらない．(ごく例外的には図 9 に示す表のような例も考えられる．しかし，このような表の出現頻度は非常に低い．) よって，本研究では，この表の一行目のみを項目名とみなし，一行目は項目名とはみなさない．

以上のような考察から，本研究での目的に対しては，項目名を以下のように定義するものとする．

定義: **項目名**
表の上 (左) 端にある一行 (列) または複数行 (列) の各セルは，同じ列 (行) の残りのセルが全て同じ種類の値を格納している場合は項目名を表すセルであり，そうでない場合は，項目名を表すセルではない．　□

上の定義は，何をもって同種の値と見なすかという点において曖昧さを残す定義ではあるが，一定の指針を与えることはできる．よって，これ以後，本研究では，人手によってあるセルが項目名であるかどうかの判定を行う場合には，この定義に基づいて判断するものとする．

また，上の定義に基づいて判定を行う限り，データにあたる行が一行しかない場合は，上端の行が項目名であるかどうかの判定はできないことになる．項目内容にあたる列が一列しかない場合の，左端の列が項目名であるかどうかの判定も同様である．よって，このような場合は，その行または列は項目名とも項目名でないとも，どちらともみなせるものと考える．例えば，図 10 のような表があった場合，一列目は項目名であるともないともみなせると考え，後述の本論文で提案する自動判定手法の評価においては，自動判定手法が一列目を項目名と判定しても，項目名でないと判定しても，正解とみなすことにする．

4.2　行方向，列方向の双方に項目名がある表

従来の関係データモデルにおいては，図 1 の表と同様に，行

入学試験実施要領

学部	試験日
法学部	2 月 18 日
文学部	
経済学部	
工学部	2 月 19 日
医学部	2 月 18 日

図 10　左端を項目名とみなすとデータが一列となる表
Fig. 10　Cells Spanning Multiple Lines

過去の進路状況

年度	民間企業	公務員		進学	
		国家	地方	本学	他学
2002	83	35		18	5
2003	76	25	16	22	2
2004	88	18	14	16	3

図 11　集約を表すセルの併合の例
Fig. 11　Cell Merging Representing Aggregation

方向は同じ項目に対応するセルが並び，各列に一つ属性名があった．しかし，Web 上で見られる表においては，行と列の双方に項目名がある場合もある．図 3 の表は，そのような例になっている．この表では，一行目と一列目の双方が項目名に対応し，行方向も列方向も金額という同じ種類の値のセルが並んでいる．その結果，二行目以降の二列目以降は，全て同じ種類の値のセルが並んだ表になっている．このように，Web 上の表では，行方向と列方向の双方に項目名が存在する場合があるため，本研究では，従来の関係データモデルで用いられる「属性名」ではなく,「項目名」という呼び方を用いることにする．

図 5 の表も，行方向と列方向の双方に項目名がある表の例である．一般的には，この表は，一〜二行目が項目名を表し，一列目は「年度」という項目名に対応する値が並んでいると捉えられるかもしれないが，本研究での前述の定義に基づけば，一列目を項目名と見なした場合，二行目以降の二列目以降は人数という同じ種類の値が並んでいるので，一列目は項目名と判断される．実際，同様の表で，列方向にセルが併合された，図 11 のような表もありうる．この表は，2002 年度については，国家公務員として就職した人数と地方公務員として就職した人数の詳細は不明で，合計で 35 人であることだけがわかっていることを表す表である．ただし，この表を格子状の表に変形する際に，単純に 35 という値をコピーして二つのセルを生成してしまうと，国家公務員，地方公務員，それぞれの人数が 35 人であったかのような表になってしまい問題がある．これは，この「35」のセルが，値が同じセルの読み易さのための併合を表しているのではなく，前節の冒頭部分でも述べた，複数のセルの内容の「集約」を表しているセルだからである．しかし，このような「集約」を表すセルの併合の出現頻度は非常に低いので，本稿では取り扱わず，今後の課題とする．

また，図 3 の表においては，一列目，一行目のセルが空白であるのに対して，図 15 の表では「年度」と一列目の二行目以降の内容に対応する項目名が書かれているので，一列目は項目名ではないのではないかという捉え方もあると思われるが，実際には，図 3 のような行方向と列方向に対称性があるような表でも,「発駅」等のような一列目，二行目以降の内容に対応する項目名が，一列目，一行目に書かれている表はしばしば見られる．よって，左上角の内容が空白かどうかのみからでは，上端，左端の双方が項目名であるかどうかの判断はできない．

4.3　項目名にあたるセルの判定

本来，HTML4 [10] の規格では，データにあたるセルのためのタグである `<td>` (Table Data cell) とは別に，項目名にあたるセルのためのタグである `<th>` (Table Header cell) が用意されている．しかし，両方の役割を果たすセルは `<td>` を用いるとされており，また，より根本的な問題として，Web 上の多くの表では，`<th>` タグが正しく用いられていない．よって，HTML 中のタグの情報だけから，項目名にあたるセルを判定することはできない．

また，図 1 の表では，一行目のみが太字表示となっている．実際の Web 上の表でも，このようにフォントやセルの背景色などを明示的に変更して，項目名にあたるセルのみを区別している場合は多いが，必ずしも全ての表において，そのような工夫がされてはおらず，また逆に，フォントやセルの背景色が強調やスタイルの都合から変更されていても，項目名に対応しているわけでは無い場合も多い．よって，これらの情報は有用な情報にはなるが，これらだけで，表中のどの行，列が項目名に対応するかを判定することはできない．

5.　本研究の提案方式

本章では，本研究で提案する，複数行・列にまたがるセルを含む表の小画面での表示方式について説明する．

5.1　項目名にあたるセルの自動判定手法

これまでで述べたように，本研究のアプローチでは，どのセルは項目名にあたり，どのセルがその項目名のデータにあたるかの情報を用いて表の適切な変形を行う．そして，前述のように，項目名にあたるセルの判定は簡単ではない．そこで，本節ではまず，項目名にあたるセルの自動判定の手法について述べる．

前述のように，ここでは，項目名にあたるセルは，上端の一行または数行，あるいは左端の一列または数列，あるいは，それらの双方に存在すると仮定する．縦に長い列等において，上端と下端の双方に同じ項目名がある場合があるが，この場合は，上端の項目名を判定できれば，それと中に書かれた文章まで全く同じ内容 (ただし，項目名が数行に渡る場合は上下が反転したもの) が下端にあれば，項目名と判定する．左右両端に項目名がある場合も同様である．ごくまれに，右端のみや下端のみに項目名がある場合があるが，そのような表の出現頻度は極度に低く，ここでは，そのような表の項目名の発見は取り扱わない．

以上の仮定のもと，本研究では以下のような手順で，項目名にあたるセルの発見を行う．

項目名判定手順の概要

（1） 一つのセルからなるような行，一つのセルからなるような列は,「見出し」または「区切り」を表すものとして，以下

駐車料金 (一時間あたり)	
自転車・自動二輪	150
普通車・大型車	300

図 12　見出しを表す一つのセルからなる行の例
Fig. 12　A Line Representing a Caption Consisting of a Single Cell

学内連絡名簿

氏名	内線	e-mail
総務課		
山田一郎	2850	yamada
伊藤伸一	2851	suzuki
経理課		
佐藤　進	3010	sato
太田大介	3011	ohta

図 13　区切りを表す一つのセルからなる行の例
Fig. 13　Lines Representing Dilimiters Consisting of Single Cells

の項目名の判定の処理からは取り除く．(見出しを表す行の例としては図 12 に示すような表がある．また，区切りを表す行の例としては，図 13 に示すような表がある．)

（2） 一行目から順に,「その行までを項目名とみなしたとして，後述の規則群に照らして矛盾がないか」を調べていき，最初に,「矛盾なし」と判定された個所までを項目名にあたる行の候補と考える．そのような個所がなければ，上端には項目名はないと判定する．

（3） 同様に，一列目から順に,「その列までを項目名とみなしたとして，後述の規則群に照らして矛盾がないか」を調べていき，最初に,「矛盾なし」と判定された個所までを項目名にあたる列の候補と考える．そのような個所がなければ，左端には項目名はないと判定する．

（4） 上端と左端の双方に，項目名にあたる行と列の候補が発見された場合は，後述の規則群を使って，そのどちら (あるいは双方) が項目名であるかの判定を行う．　□

また，上のステップ (2) で用いる，ある行までを項目名として矛盾がないかを判定する規則群は以下の通りである．(列に関する規則群は，行と列を入れ替えただけで以下と同様なので省略する．)

項目名にあたる行の境界に関する規則群

（1） 項目名にあたる行とデータにあたる行の境界をまたがってつながっているセルはない．

（2） 項目名にあたる行の中で，同じセル数の行が二行以上続くことはない．

（3） 項目名にあたる行の中で，下に進むにつれて列数が減ることはない．

（4） 項目名にあたる行中には現れない列の区切りが，データにあたる行の中に現れる場合はない．ただし，データ中に現れるそのような区切りの左側 (列に関する規則の場合は上側) のセルが複数行にまたがっていてる場合は除く．

（5） 全ての行が項目名にあたり，データにあたる行が一行もないということはない．　□

また，行と列の双方に項目名の候補があった場合に，どちら (あるいは双方) が項目名かの判定のための規則群は以下の通りである．

上端と左端の項目名候補の判定に関する規則群

（1） 表内のデータにあたる部分に，複数行にまたがるセルと複数列にまたがるセルの双方がある場合は，上端も左端も項目名である．

（2） 表内のデータにあたる部分に，複数行にまたがるセルのみがあり，複数列にまたがるセルがない場合は，上端のみが項目名である．

（3） 表内のデータにあたる部分に，複数列にまたがるセルのみがあり，複数行にまたがるセルがない場合は，左端のみが項目名である．

（4） 表内のデータにあたる部分に，複数列や複数行にまたがるセルが全くない場合は，上端が項目名である．　□

ただし，最後の「上端と左端の項目名候補の判定に関する規則群」は，あくまで現時点でのものであり，後述の評価結果の通り，これらの規則群は間違った解を導くことも多いので，さらなる改善が必要であると考えている．単純に考えても，複数行にまたがるセルと複数列にまたがるセルの双方がある場合は，上端，左端とも項目名である可能性が高いということは言えるが，行にまたがるセルがないからといって，上端が項目名でないという根拠にはならない．ただ，Web 上に現れる表の中で，上端，左端の双方に項目名がある表の出現頻度は比較的低く，また，どちらか一方のみに項目名がある場合は上端に項目名がある場合が多いため，複数行にまたがるセルと複数列にまたがるセルがある場合のみ，上端，左端の双方が項目名であると判定し，それ以外の場合は，どちらか一方のみが項目名と判定し，また，上端と左端のどちらが項目名かセルの形からは判断できない場合は上端が項目名であると判定することで，結果としては，ある程度の精度を出すことができる．

より正確な判定には，各セルの内容の類似度や，フォント，セルの背景色，上左角のセルの内容などの情報を使う必要があると考えられる．この点については，今後の課題である．

5.2　自動判定の例

例として，図 14 のような表(注1)を考える．まず，一つのセルからなる行や列があれば取り除くが，この表には存在しない．次に，一行目から順に，その行までを項目名とみなせるかを順に調べていくと，一行目と二行目の間を項目名とデータの境界として，前述の規則群のいずれにも抵触しないことがわかる．よって，この最初に見つかった一行目と二行目の間を，上端の項目名の境界の候補とする．

次に，列についても同様に，一列目から順に見ていく．まず，一列目と二列目の間を項目名の境界だとすると，二行目と三行目の間の区切りは項目名中には存在しないのにデータ中には存在することになり，またデータ中のその区切りの上のセル (二行目，三〜五列目のセル) が複数列にまたがっているわけではないので，規則 (4) に抵触する．よって，次に，二列目と三列目の間を考える．この場合，同じく規則 (4) に抵触すると同時

(注1)：http://www.joc.or.jp/athens/result/athletics.html

日本代表選手団: 陸上競技

種目	エントリー数	選手名	成績	備考
男子 100m	84	末續慎吾	17 位	10.19 2 次予選敗退
		朝原宣治	21 位	10.24 2 次予選敗退
		土江寛裕	39 位	10.37 1 次予選敗退
男子 200m	56	高平慎士	40 位	21.05 1 次予選敗退
		松田亮	52 位	24.59 1 次予選敗退
男子 400m	64	山口有希	32 位	46.16 1 次予選敗退
		小坂田淳	40 位	46.39 1 次予選敗退
		佐藤光浩	45 位	46.70 1 次予選敗退
⋮	⋮	⋮	⋮	⋮

図 14　項目名の自動判定の例 (1)
Fig. 14　Example of Automatic Detection of Item Names (1)

水族館入場料

区別	金額		
大人			1,320 円
小人			400 円
団体割引	20 人以上	大人	1,190 円
		小人	360 円
	100 人以上	大人	1,060 円
		小人	320 円

図 15　データ領域中の階層構造を持つセル
Fig. 15　Hierarchical Cells in Data Region

に，一列目と二列目は行数が変わらないため，規則 (2) にも抵触する．このような場合，それ以上，先の列を見ても，必ず規則 (2) に抵触することになるので，実際に調べずとも，規則群に抵触しないような，項目名の境界となる列の候補はないということがわかる．

以上から，一行目のみが項目名の候補となり，上端のみに項目名の候補があるので，これを最終的な解とする．この表の場合，人間が前述の基準のもとに判断しても同じ結果が得られるので，自動判定によって正解を得られたことになる．

また，規則 (4) で，項目名中にない区切り線がデータ中にあっても，その左 (上) のセルが行 (列) 方向につながっていれば構わないというのは，図 15 の表(注2)のようなデータ中に階層構造を含む表に対応するためである．この表では，二列四行の表の二列目，四行目のセルの中に入れ子になった表があって，入れ子の表の中にローカルな階層構造があり，そのため，表全体の項目名の部分にはない区切りが入れ子の表中に現れると考えられる．そこで，このような場合に対応するため，項目名中にない区切りがデータ中にあっても，それがローカルな階層構造を表していると考えられる場合は無視するというのが，規則 (4) の主旨である．本来は，階層構造かどうかのより正確な判定をすべきだが，上のような簡単なルールでも，これまでの実験では，間違った解を導く例はなかった．

自動判定の例を，もう一つだけ挙げておく．図 16 の表(注3) に対して自動判定を行うことを考える．まず，一行目と二行目の

国内郵便物の重量・大きさ

区別			重さ	最大	最小
通常郵便物	第二種	通常はがき	2〜6g		
		往復はがき	4〜12g		
	第一種	定形郵便物	50g まで		
		定形外郵便物	4kg まで		
	第三種郵便物		1kg まで		
	第四種郵便物		1kg まで		
			3kg まで		
小包	一般小包・点字小包		30kg まで		
	冊子小包郵便物		3kg まで		
	聴覚障害者用小包				

図 16　項目名の自動判定の例 (2)
Fig. 16　Example of Automatic Detection of Item Names (2)

間を考えると，一〜三列目の区切りが一行目にはないが，これらの区切りの左側のセルは，いずれも複数行にまたがっているため，規則 (4) に抵触せず，一行目と二行目の間が候補となる．

一方，列についても，まず一列目と二列目の間を考えると，一行目が規則 (1) に抵触し，かつ，二行目と三行目の間の区切りなどが規則 (4) に抵触する．次に，二列目と三列目の間は一行目，六行目などが規則 (1) に抵触する．三列目と四列目の間は，七行目と八行目の間の区切りが規則 (4) に抵触する．四列目と五列目の間は，いずれの規則にも抵触しないため，ここが項目名の区切りの候補となる．

さらに，上端，左端の双方に候補が得られたため，どちらが項目名の判定を行うが，この表では，上端，左端を項目名と仮定した場合のデータ領域中に，行方向にまたがるセルも列方向にまたがるセルも存在するため，上端一行，左端四列の双方が項目名となる．この例の場合，人間が前述の基準で判定しても同じ結果となり，自動判定で正解が得られたことになる．

5.3　採用しなかった規則

本稿で提案する手法で用いる様々な規則群について既に述べたが，これ以外にも，検討の結果，採用しなかった規則があるので，紹介する．まず，項目名にあたる行の境界に関する規則としては，上に挙げたものの他に以下のものを検討した．

- 「項目名の中に存在する区切り線で，データ中に現れないものはない」

これは規則 (4) の逆に対応する規則だが，実際には，図 17 のような表(注4)がしばしば存在するため，採用しなかった．図 17 の表では一列目は項目名と見なされるが,「09〜13 時」と「13〜17 時」の間の区切りは，データ中では一度も使われていない．

また，上端と左端の双方に項目名の候補がある場合の判定のための規則としては，以下のものを検討した．

- 「行と列の双方に項目名の候補がある場合，項目名の数が 20 を越える方は項目名ではない」
- 「行と列の双方に項目名の候補がある場合，項目名の数が 1 しかない方は項目名ではない」

これらの規則は，一部の表において，項目名でないものを項目

(注2)：http://www.notoaqua.jp/sougou/index.html
(注3)：http://www.post.japanpost.jp/fee/simulator/service/ookisa.html
(注4)：http://www.naash.go.jp/yoyogi/ichitai_riyou.html

興行的スポーツ及び文化的行事

区分	入場料を徴収しない場合		準備に使用する場合	
	平日	平日以外	平日	平日以外
09〜13 時	611 千円	901 千円	509 千円	751 千円
13〜17 時				
17〜21 時	762 千円	1,130 千円	635 千円	945 千円
09〜17 時	1,080 千円	1,620 千円	903 千円	1,350 千円
13〜21 時	1,220 千円	1,810 千円	1,020 千円	1,510 千円
09〜21 時	1,560 千円	2,250 千円	1,300 千円	1,870 千円

図 17　データ中で一度も使われない区切りの例
Fig. 17　Example of Delimiter Never Used in Data

2005/10/1 現在

氏名		所属	詳細
台長			
海部 宣男	KAIFU, Norio	国立天文台	more >>
副台長			
観山 正見	MIYAMA, Shoken	理論研究部	more >>
櫻井 隆	SAKURAI, Takashi	太陽観測所	more >>
⋮	⋮	⋮	⋮

図 18　項目名中にはないデータ中の区切りの例
Fig. 18　Example of Delimiters not Specified in the Header

名と判定するのを防ぐことに貢献したが，逆に，項目名である側を項目名でないと判定することも多く，採用しなかった．

6. 実　　験

これまでに，複数の行または列にまたがるセルを含む Web 上の表を 70 個収集し，これらに本稿で提案する項目名の自動判定の手法を適用して結果の評価を行った．ただし，この 70 個の表は，ランダムに選んだのではなく，採用すべき規則群の検討のため，作為的に複雑な構造を持ったものを特に選んで収集している．ランダムに表を収集した場合の統計や，モバイル機器からアクセスしたい需要が高いようなページに絞っての実験については，今後の至急の課題である．

まず，70 個の表に対して，どのような項目名を持っているかの判定を人手で行った場合の結果は以下のようになった．

- 上端と左端の双方に項目名を持つもの — 13 個
- 上端のみに項目名を持つもの — 28 個
- 左端のみに項目名を持つもの — 20 個
- 項目名を持たないもの(注5) — 4 個
- その他の特殊な構造を持ったもの(注6) — 5 個

上端や左端に項目名を持つ 61 個に対して，自動判定の手法を適用した結果は以下のようになった．

- 正解 — 44 個
- 行と列の双方に項目名の候補を発見した後，これらのどちらが項目名かを判定するところで誤ったもの — 5 個
- 不正解 — 12 個

この数字だけを見ると精度が悪いように見えるが，前述の通り，これは規則群の検討のために意図的に複雑な構造の表のみを収集したデータを用いた評価の数値であり，ランダムに収集したデータ群を用いた場合，精度が大きく向上すると予想される．

また，前述の項目名の境界に関する規則群のうち，不正解の解の原因となることがあったのは (4) のみであった．規則 (4) のために不正解となるのは，表のデータ領域中に，階層構造にはなっていないローカルな構造があって，項目名には出てこない区切りが出てくる場合で，例えば図 18 に示す表(注7)のような場合である．この表では，項目名「氏名」に対応するデータのセルの中に，項目名のセル中にはない，漢字表記とローマ字表記の間の区切りが使われているため，三行目までが項目名 (ただし二行目は区切り記号として除かれる) と判定されてしまう．

7. ま と め

本論文では，複数行や複数列にまたがるセルを含む Web 上の大きな表データを小画面中で表示するために，単純な格子状の表に変形する手法について提案した．特に，そのような変形を適切に行うために，表中の項目名にあたる領域とデータにあたる領域を自動判定する手法を提案した．提案手法の実験によるより詳細な評価は，今後の至急の課題である．また，本研究では，表中の文字情報の内容には立ち入らずに判定する手法を提案したが，さらに精度を上げるためには，表中の文字情報も合わせて使う必要があり，そのような手法についても今後の課題である．

(注5)：http://www.vjc.jp/j/member.html 等
(注6)：http://www.jasrac.or.jp/network/contents/tariff.html 等
(注7)：http://jouhoukoukai.nao.ac.jp/reslist/index.asp

文　　献

[1] 増田英孝，塚本修一，安富大輔，中川裕志，“HTML の表形式データの構造認識と携帯端末表示への応用，” 情報処理学会論文誌トランザクション「データベース」，vol.44，no.SIG12(TOD19)，pp.23-32，2003.

[2] R. Zanibbi, D. Blostein, and J.R. Cordy, “A survey of table recognition,” International Journal on Document Analysis and Recognitionl, vol.7, no.1, pp.1-16, 2004.

[3] 伊藤史朗，大谷紀子，上田隆也，池田祐治，“属性オントロジーの抽出と統合を用いた実空間と情報空間のナビゲーションシステム，” 人工知能学会誌，vol.14，no.6，pp.1001-1009，1999.

[4] M. Hurst, and S. Douglas, “Layout & language: Preliminary experiments in assigning logical structure to table cells,” Proc. of the 5th Applied Natural Language Processing Conference, pp.217-220, 1997.

[5] H.H. Chen, S.C. Tsai, and J.H. Tsai, “Mining tables from large scale html texts,” Proc. of 18th Intl. Conf. on Computational Linguistics (COLING), pp.166-172, 2000.

[6] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong, “Table structure recognition and its evaluation,” Proc. of Document Recognition and Retrieval VIII, pp.4307:44-55, 2001.

[7] M. Yoshida, K. Torisawa, and J. Tsujii, “A method to integrate tables of the World Wide Web,” Proc. of 1st Intl. Workshop on Web Document Analysis, pp.31-34, 2001.

[8] P. Pyreddy, and W.B. Croft, “TINTIN: A system for retrieval in text tables,” Proc. of 2nd Intl. Conf. on Digital Libraries, pp.193-200, 1997.

[9] 田仲正弘，“表構造の解釈に基づくオントロジーの獲得，” 修士論文，京都大学情報学研究科，Aug. 2005.

[10] I.J. Dave Raggett, Arnaud Le Hors, ed., HTML 4.01 Specification, http://www.w3.org/TR/REC-html40/, Dec. 1999.