E-051

表の属性間の階層関係抽出

Extracting Hierarchical Relationships of Attributes in Tables

山口　智由†　　青野　雅樹†
Tomoyoshi Yamaguchi　Masaki Aono

1．はじめに

Web には表形式によるデータが大量に存在する．表中の属性間には複雑な階層関係がある．このような情報を利用するため，表からのセルの結合や隣接関係などを利用した情報抽出に関する研究がなされてきた[1]．表には，人が見て理解しやすくなるための情報である罫線がある．我々は，罫線に着目することにより，セルの集合関係が得られると考えた．本稿では，この罫線情報を基に，セルの集合関係と階層関係の抽出を試みたので報告する．

2．関連研究と問題点

田仲ら[2]は，表の構造を一般化し，それに基づきオントロジーを獲得している．表からの抽出は，セルの隣接関係を重なる辺の長さにより，接続関係を決定している．しかし，Web には HTML の<table>タグで表現されている表以外に，Excel ファイルによる表も多数存在している．

例えば，総務省統計局では多くの統計データを Excel 形式で公開しているが，図 1 のような複雑な構造を持つものが多く従来の認識手法では上手く扱えない．そこで，セルの結合だけでなく罫線によるセルの集合が重要であると考えた．ここで点線により示されているのがセル 1 つである．このことは，紙媒体の表形式データを電子化する研究として行っている文書画像理解の分野で成瀬ら[3]も罫線情報が重要であると述べている．

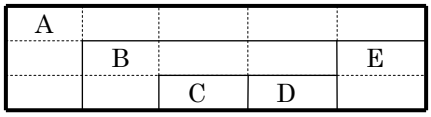

図 1　Excel ファイルの例

3．提案手法

前節で述べた問題を解決するために，罫線によるセルの集合関係に基づく属性の階層関係を抽出するルールを作成した．

3.1　セルの集合

図 1 に示す表を実線により分割したとき以下のようなセルの集合で表現できる．

A = {(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (2, 1), (3, 1)}
B = {(2, 2), (2, 3), (2, 4), (3, 2)}
C = {(3, 3)}
D = {(3, 4)}
E = {(2, 5), (3, 5)}

また，表の一番外枠により以下の集合関係も表現できる．

F = {A, B, C, D, E}

このように罫線に囲まれ領域分けされ，それぞれが意味を持つセルの集合を，以下では「ノード」と呼称する．

3.2　罫線の種類と優先順位

表に含まれる罫線には，実線・破線・鎖線など形状と太さの組み合わせによりいくつかのタイプがあり，一般に太い実線は表全体を，細い破線や鎖線はセルを，という具合に罫線のタイプに自然な優先順位をつけて表をレンダリングするのが通例である。優先順位の高い罫線は，より多くのセルを囲む傾向があるため，囲んでいるセルの数により優先順位を決定することとし，同数の場合は，種類ごとに振られている番号の若い順とした．

3.3　セルの集合と囲み判定

ノード間の階層関係を決定するために，優先順位の高い罫線から順に表に含まれるセルを分割する．これにより，優先順位の高い罫線による集合から徐々に優先順位の低い罫線の集合へと分割し，ノードの包含関係が得られる．また，セル単独，あるいは隣接するセルの集合が罫線により囲まれているか判定するアルゴリズムを考えた．すなわち，対象となるセルからスタートし，右方向に優先順位の高い罫線とぶつかるまで進む．罫線とぶつかった後，反時計回りに進み，最初にぶつかった地点までたどり着くことができれば，囲み判定有りとする．反時計回りに回る際には，下記の手順にて進む．

・下方向の次は左方向
・左方向の次は上方向優先で無ければ下方向
・上方向の次は右方向
・右方向の次は下方向

これにより，図 1 に示す集合 A，B のような形状のセルの集合関係（すなわちノードの階層関係）を決めることができる．

3.4　属性の対象

属性の抽出対象は，文字列を含むセルの中から形態素解析の結果，名詞(数は含まず)が含まれているセルとした．

3.5　抽出手順

3.3 で述べた罫線による分割されたノードを基に次の手順で表の階層抽出を行う．ここで，抽出対象はノードに含まれる属性である．表の階層において親となるのは上または左に隣接しているノードに含まれる属性である．事前の票の分析により，下記に示す順序で親となる属性を決定することとした．ただし，どれにも当てはまらない場合は親ノード無しとする．

1) 上および左に接するノードの属性が同じ場合，今見ているノードの属性の親とする．
2) 上または左に接するノードの親が他方の親である場合，より上位階層のノードの属性を親とする．
3) ノードの包含関係が近い方を親とする．

†豊橋技術科学大学
Toyohashi University of Technology

4) i（＝行方向のセル番号），j（＝列方向のセル番号）
i＞j なら，上を優先．i＜j なら，左を優先する．

他に一つのセルの集合の中で多くの名詞を含む場合は，セルの位置により下記のような順序にて属性の階層関係を決定することとした．

・左上のセルより，他の属性を含むセルにぶつかるまで名詞を右と下方向の空白セルに属性を転写する．
・セル集合の右下から順に i＞j なら上を優先．i＜j なら，左を優先し親を決定する．

図 1 に示す表に上記の抽出ルールを適用すると，B は上と左の両方を A と接していることから，A が親で B が子になる．同様に C も両方を B と接していることから，B が親になる．次に D では，上が B，左が C と接している．この場合は，一方の C の親が他方である B であるため，D の親も B になる．E も同様に，B の親が A であることから A が親になる．以上より，A の子に B と E．B の子に C と D と言うことが得られる．

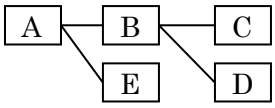

図 2　図 1 に対応するノードの階層関係

4. 実験方法

実験データは，Web より収集した Excel データ 317 ファイルを対象とした．そのうち，ランダムに 20 ファイル選び人手により正解を判断した．正解の基準は，子からみた親が正しいかを判断した．また，Excel ファイルを解析するにあたり Apache Jakarta プロジェクトの「POI」[4]を使用した．これにより CSV 形式による読み込みでは扱えないレイアウト情報（罫線など）を扱うことができ，本提案手法が実現可能となった．

形態素解析には Java で記述されている「sen」[5]を用いた．辞書に関しては，システムのデフォルトを用いた．

4.1 概要

【実験 A】Excel ファイル中のシートを単位として入力とした．シートに含まれる罫線によりセルの集合に分割した包含関係情報を XML で表現した．XML 表現を基に提案手法にて属性の階層関係を抽出した．

【実験 B】3.2 で述べた罫線の優先順位決定方法の有用性を示すため，提案手法である数による優先順位と，表中の罫線をランダムに用いる手法の比較も同様に行った．

実験 B においては，実験 A で用いる優先順位の番号を乱数によりランダムに決めた．罫線が n 種類含まれている場合 1～ n の乱数を n 回発生させ，発生順の罫線を用いた．

4.2 評価方法

本実験を評価するに当たり，再現率および誤り率を下記のように定義した．

$$再現率 = \frac{本実験による正解数}{人手による正解数} \quad \cdots (1)$$

$$誤り率 = \frac{誤った抽出数}{本実験による抽出数} \quad \cdots (2)$$

表 1　階層抽出実験結果

	マイクロ平均		マクロ平均	
	再現率	誤り率	再現率	誤り率
実験 A	0.841	0.104	0.773	0.156
実験 B	0.672	0.081	0.611	0.127

5. 実験結果と考察

本実験結果を表 1 に示す．実験結果から提案手法では，マクロ平均よりマイクロ平均の方が良い結果を得られた．このことから，有効であるデータとそうでないデータが存在していることが分かる．これにより，入力データの傾向に偏りがあったのでないかと考えられる．

また，実験 B に比べ実験 A が良い結果となったことから，囲んでいるセルの数による罫線の優先順位の決め方は有用であることが示せた．

再現率が低下した原因として，3.4 で述べた抽出方法ではカバーしきれないケースがあったことが挙げられる．

例えば，親となるノードが隣接するノードのみに存在すると限定したため，隣接していないノードとの間にある階層関係を抽出できなかった．また，ノード中に複数の属性を含むセルが存在する場合，本来なら組み合わせて一つの属性とすべきところを別々に取り扱ってしまったことも原因の一つと考えられる．

誤り率が高い原因も，再現率低下と同様の理由が考えられる．

6. おわりに

本稿ではセルの集合関係を用いた Web 上にある Excel ファイルの表を対象とした属性間の階層関係の抽出を行った．Excel ファイルの表を扱う上で，罫線により分割されたセルの集合関係を用いることの有用性を示せた．また，罫線の優先順位を決める方法として，罫線に囲まれているセルの数による優先順位の決め方も有用であることが示せた．

今後の課題として，Excel ファイルのみならず，HTML の〈table〉タグにおいても同様にセルの集合関係を用いた手法が有用であるか検証を考える予定である．また，より多くの表を収集･調査・解析し，現状の抽出ルールよりも良い抽出方法の検討も考える．更に，Excel ファイルの中には罫線により閉じていない表も存在する．そのような場合の対応も，今後の課題である．

参考文献

[1]D.W. Embley, D.P. Lopresti, and G. Nagy: Notes on Contemporary Table Recognition, In Proc. 7th Int. Workshop on Document Analysis Systems (DAS), pp. 164- 175, 2006

[2]田仲正弘，石田亨：表構造の一般化に基づくオントロジーの獲得，情報処理学会誌，Vol.47，No．5，pp1530-1537，2006

[3]成瀬博之，渡辺豊英，駱琴，杉江昇：枠罫線情報を用いた帳票文書の構造認識，電子情報通信学会論文誌，D-II，Vol. J75-D-II，No. 8，pp1372-1385，1992

[4] Apache Jakarta プロジェクトの「POI」：http://poi.terra-intl.com/

[5]sen: http://ultimania.org/sen/