1A4-02

表形式データからのオントロジーの獲得

Ontology Extraction from Tables

田仲正弘　　石田亨
Masahiro Tanaka　Toru Ishida

京都大学情報学研究科社会情報学専攻
Department of Social Informatics, Kyoto University

Previous works on information extraction from tables make use of lexical knowledge bases of tables or prior knowledge such as a cognition model of tables. However, we often need to interpret table structures in each table differently and to treat lexicons in various domains for processing a broad range of tables on the Web. The method proposed in this paper extracts an ontology from a table by using relations represented by structures. Once the interpretations of table structures are given by humans, the table structures are automatically generalized to extract relations from the whole table. We defined a formal representation of generalized table structures based on the adjacency of cells and iterative structures. Our experiments showed that the method extracted class-hierarchies, property-value pairs and other various relations from the tables containing price lists, contact lists and statistics on the Web.

1. はじめに

Semantic Web の実現には大量のメタデータが必要になるが，メタデータの作成には大きなコストがかかるため，既存のデータからメタデータを獲得する方法が必要である．本研究では，ある程度決まった構造を持っている表形式データから，オントロジーを獲得する手法を提案する．表形式データでは表構造により，表中のデータ間の関係が表される．Web では表形式データはオンラインショップのカタログ等に広く用いられており，表構造が表す関係を用いて，表中のデータに関するオントロジーが構築できれば，希望の商品を検索するなどの用途に利用できる．

本研究では幅広く Web から集められた表を対象とすることを考え，以下の 2 点に注目する．

表に応じた構造の解釈 同じ表構造でも，表によって異なる関係を表すことがあり，それぞれの表に応じた表構造の解釈が必要となる．

様々なドメインの表の処理 様々なドメインの語彙を含む表を扱う場合には，ドメインに特化した知識ベースによらない表の解析手法が適する．

従来の研究には，表を解釈するために表構造に関する先見的知識を用いるもの [Chen 00, Pivk 04]，対象ドメインの知識ベースを用いるもの [Tijerino 03] などがある．これらは特定ドメインの表を自動的に処理することに重点を置くものである．

一方，本研究では，人手による表構造の解釈を反映させることで，様々なドメインの表からデータ間の関係を半自動的に獲得することを目的とする．本研究でのアプローチは以下のようになる．

1. 表に応じた表構造の解釈を与える
2. 解釈を与えた表構造を一般化する
3. 表全体からデータ間の関係を獲得する

連絡先: 田仲正弘，京都大学情報学研究科社会情報学専攻，〒 606-8501 京都市左京区吉田本町，TEL 075-753-5396, FAX 075-753-4820, mtanaka@kuis.kyoto-u.ac.jp

表 1: PC 部品の価格表

Processor		
ProductID	ProductName	Price
P4_340	Pentium 4 3.40E GHz	$260
P4_280	Pentium 4 2.80A GHz	$140
A64_320	Athlon 64 3200+	$160

表構造の解釈を表ごとに人手で与えることで，表に応じた表構造の解釈ができる．構造の一般化はセルの隣接関係や繰り返し構造に注目して行う．データ間の関係を表す一般化された表構造を用いることで，表全体からデータ間の関係を得る．また与えた解釈に基づいて表からデータ間の関係を得るため，ドメインに特化した知識ベースを必要とせず，様々なドメインに適用可能である．

以降では，第 2 章で表構造の観察について述べる．次に第 3 章で，表構造の形式化について述べる．第 4 章では，表からのデータ間の関係の獲得の処理について説明する．第 5 章で，提案手法の適用の結果について評価し，第 6 章で結論を述べる．

2. 表構造の観察

本章では，表構造の形式化と一般化のため，表構造で表される関係と表構造による記述についての観察を述べる．

2.1 表構造のセマンティクス

表では，その構造によってセル中のデータ間の関係（クラス-インスタンス関係・クラスの階層関係・プロパティ-プロパティ値の組など）が表される．一つの表の中では同じ構造が表す関係は一定であると考えられる．ここでは表構造が表すデータ間の関係を，その表構造のセマンティクスと呼ぶ．例えば表 1 の構造は,「一行目には表に記載されている個体の属するクラスが，二行目には個体のプロパティが記述され，その下に各個体のプロパティ値が記述される」というセマンティクスを持っている．このような特定の表構造のセマンティクスが得られれば，表中で同じ構造を持つ部分に含まれる多数のデータの関係も得られる．

2.2 表構造の仮定

ある表構造のセマンティクスが明らかなら，表中でその表構造が出てくる箇所からは，その箇所に含まれるデータについての関係が得られる．ただし表中のより多くの箇所からセルのデータ間の関係を得るには，表構造をその特徴に基づいて一般化する必要がある．

表構造の観察の結果から，以下の点に注目し，表構造の記述に関する仮定に基づいて一般化された表構造の形式的表現を定義する．

同じ行や列内のセルの関係 プロパティ名とそのプロパティの値など，互いに関連を持つセルは，ふつう同じ行や列にある．そのため，同じ行や列内のセルの関係は，その行や列の構造と行や列内でのそれらのセルの位置によって表されると考えられる．

行や列の構造 表には複数の行や列にまたがるセルが含まれることがあり，隣接する 2 つのセルで幅が異なる場合，それらのセルにはふつう異なる種類のデータが含まれる．そのため，行や列の構造は，その行や列に含まれるセルとその周囲のセルとの幅の大小関係によって特徴付けられると考えられる．また，同じ行や列内で同じ特徴を持つセル（もしくは複数のセルが集まったブロック）が連続して出現する場合には，それらのセル（ブロック）は出現回数によらず似た種類のデータを表していると考えられる．

異なる行や列にあるセルの関係 表中で異なる行や列にある 2 つのセルが互いに関連を持つ場合は，それらのセルが位置する行や列が交わる部分のセルもまた，その 2 つのセルと関連を持つと考えられる．よって異なる行や列にあるセルの関係は，それらのセルを含む行や列からなる部分の構造で表現できる．

以上の観察に基づいて，特定の関係を表す一般化された表構造の形式的表現を定義する．

3. 表構造の形式化

表構造のセマンティクスを用いて表全体からデータ間の関係を獲得するには，表構造のセマンティクスを決定するような特徴でに基づく表構造の形式的な表現が必要になる．そのため，2.2 節で述べた表構造に関する仮定に基づき，以下に述べるように表構造の形式的表現を定義する．

3.1 同じ行や列内のセルの関係を表す部分

互いに関連するセルは，ふつう同じ行や列にある．そのため，同じ行や列内の関連するセルの関係は，行や列のそれらのセルを含む部分に注目すればよい．例えば表 1 で，“Processor”・“ProductName”・“Pentium 4 2.80A GHz” の関係は “Processor”・“ProductName”・“Pentium 4 3.40E GHz”・“Pentium 4 2.80A GHz” の 4 つのセルからなる部分で表現されると考えられる．

3.2 行や列の構造の表現

表中のセルの関係を表す行や列の構造を，セルの隣接関係や同じ特徴を持つセルによる繰り返し構造に注目して表現する．ここで行や列を，セルの配列とみなし，セルの隣接関係に注目して表現することを考える．

はじめに，一つのセルに相当するボックスと呼ぶ要素を定義する．さらにボックスに対応するセルの周辺のセルとの辺の重

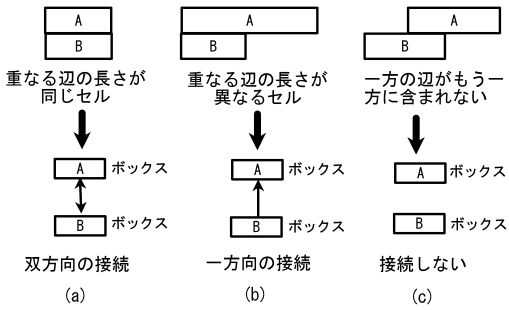

図 1: セルの隣接関係と対応するボックスの接続

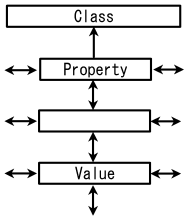

図 2: 表 1 の列の構造

なり方によって，セルの隣接関係を図 1 のように一方向または双方向のボックス間の接続としてあらわす．

図 1 の上部は元の表の隣接する 2 つのセルを示している．図 1 の下部はそれらのセルに対応するボックスとそれらの接続を表している．四角はボックスを表し，上下に並んだ二つのボックス間の接続は，対応するセルの隣接関係を表す．

図 1(a) のように，隣接する二つのセルで重なっている辺の長さが同じ場合には，二つのボックスを双方向に接続して表す．図 1(b) のように，一方の辺がもう一方の辺に含まれる場合には，短い辺を持つセルに相当するボックスから長い辺を持つセルに，一方向に接続して表す．図 1(c) のように，隣接するセルの重なっている辺の一方が，もう一方に含まれない場合には，セルに相当するボックスは接続しない．また，表において隣接する 2 つのセルのどちらが上（左）でどちらが下（右）かということは重要であるため，接続の上下左右の向きも区別する．以上を用いて，行や列の構造を表現する．

表 1 において，“Processor”・“ProductName”・“Pentium 4 2.870A GHz” のセルはある個体の属するクラス，プロパティ名，プロパティ値を表す．これらの語を含む列の構造と，その表す意味をボックスを用いて表現すると図 2 のようになる．クラス・プロパティ・プロパティ値が記述されたセルに相当するボックスには，それぞれのセルに記述されたデータが何であるかを表すラベルが付けられる．

さらに，同じ行や列内で周囲のセルとの隣接関係が同じセルが連続する場合には，それらのセルは出現回数によらず同じ種類のデータを表していると考えられる．そこで，行（列）で連続して出現するボックスが同じ接続を持ちかつそれぞれのボックスで隣接する行（列）のボックスとの接続が同じ場合には，その出現回数によらずそれらのボックスに相当するセルは同じ種類のデータを表すものとみなす．

以上を表現するために，ボックスを用いた表構造の表現に繰り返し構造を表す+ 記号を導入することで，同じ接続を持つボックスが連続して出現する構造を表現する．図 2 で同じ接続を持つボックスの連続する部分を+記号を用いて表すと，図 3 のように，2 通りの表現が得られる．このようなボックスと+記号による表構造の表現を，一般化構造と呼ぶ．+記号の横の

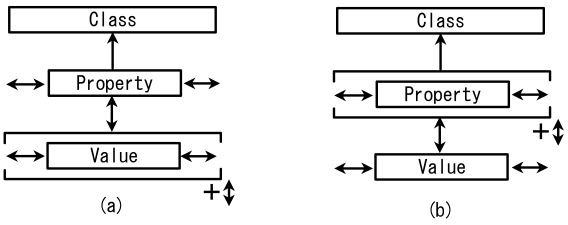

図 3: 表 1 の一般化表構造

表 2: 異なる行や列にある関連する 3 つのセル

	CPU	RAM	HDD
OptiPlex240	P4 1.7GHz	512MB	60GB
Dimension8300	P4 2.4GHz	256MB	80GB
Pavilion505	Cel 2.4Ghz	256MB	40GB

矢印は，括弧の中の “Value” とラベルの付けられたボックスが縦向き双方向の接続で 1 度以上連続して現れることを意味している．ただし，同じ接続を持つ複数のラベル付きボックスが連続して現れる場合でも，一つの繰り返しの中には含めないものとする．これは，“Class”, “Property” のような異なるラベル付きボックスに相当するセルのデータは，その種類を区別する必要があると考えられるためである．

2 通りの一般化構造が得られるのは，プロパティ名である “ProductName” が記述されたセルと，プロパティ値である “Pentium 4 2.80A GHz” が記述されたセル，及びそれらの間に配置されたセルで，周囲のセルとの隣接関係が同じであるため，プロパティ名とプロパティ値のどちらが連続して出現しているのかが表構造からは判断できないためである．しかし図 3(b) は一つのプロパティ値に対して複数のプロパティが対応するという，誤った関係を表している．

3.3 異なる行や列内のセルの関係を表す表構造の表現

表 2 は，表中に “Dimension8300”, “RAM”, “256MB” など，ある個体の名前，個体の持つプロパティ，プロパティ値という関係にあるデータが記述された表であるとする．一般化表構造は，このような異なる行や列内のセルの関係についても考えることができる．まず同じ行にある二つのセル “Dimension8300”, “256MB” について，それらを含む行の構造を得る．同様に，同じ列にある “RAM” と “256MB” について，それらを含む列の構造を得る．2.2 節で述べた仮定から，“Dimension8300”, “RAM”, “256MB” の関係に対応する部分は “Dimension8300”, “256MB” を含む行と “RAM” と “256MB” を含む列を組み合わせた部分である．よってこれらのセルの間の関係を表す構造は表 2 の網掛けした部分であり，ボックスとその隣接関係で表すと図 4 が得られる．

ただし，列の構造と行の構造を組み合わせる場合には，+記号による繰り返し構造の表現をどちらかに限る．これは複数の行や列での繰り返しを考えると，“Dimension8300”,“RAM”

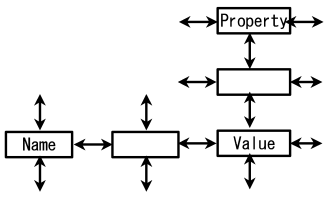

図 4: 表 2 の網掛けした部分の構造

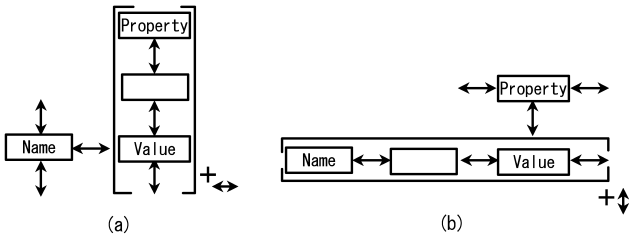

図 5: 表 2 の網掛けした部分の一般化表構造

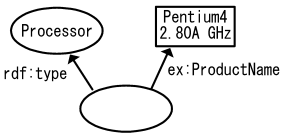

図 6: 表 1 中のデータの関係

のセルが表の端にあるという情報が失われるためである．+の横の矢印は，カッコ中のボックスに相当するセルが繰り返し現れる場合，全てのボックスは+の横に示された接続を持つことを表している．

図 4 は，どの箇所を繰り返し構造として表現するかによって，図 5(a)(b) の 2 通りの一般化表構造が得られる．図 5(a) はある行に注目して横向きに表を読んでいくことに相当し，図 5(b) はある列に注目して縦向きに表を読んでいくことに相当する．

4. 表構造の抽出の処理

第 1 章で述べたように，本研究では表中のデータの関係を獲得するために，以下のようなアプローチを取る．

1. 表に応じた表構造の解釈を与える
2. 解釈を与えた表構造を一般化する
3. 表全体からデータ間の関係を獲得する

以下ではそれぞれの処理について詳しく述べる．

構造の解釈を与える 本研究では，表中の一部のデータ間の関係を RDF ステートメントの集合で記述することにより，それらのデータが含まれる構造の解釈を与える．例えば表 1 では，“Processor” と記述された一行目のセル，“ProductName” と記述された 2 行目 2 列目のセル，“Pentium 4 2.80A GHz” と記述された 4 行目 2 列目のセルはそれぞれある個体のクラス・プロパティ・プロパティ値を表していると解釈できる．この関係を RDF のグラフで表すと図 6 のようになり，2 つの RDF ステートメントによって記述される．

このように表構造が表す関係は，単一の RDF ステートメントでは表現できないことが多い．そこで図 6 のように，特定の表構造があらわす関係を記述する RDF ステートメントの集合を，エピソードと呼ぶものとする．

解釈を与えた構造を一般化する 解釈が与えられた構造から，一般化表構造を得る手順は以下のようになる．まず，与えたエピソードに含まれる RDF ステートメントのリソースやプロパティが表中で出現するセルを探す．次に, 見つかったセルを含む行や列の構造をボックスとその接続で表現し，さらに同じ接続を持つボックスが繰り返し現れる部分を+記号によってまとめる．与えたエピソードのリソースやプロパティが現れ

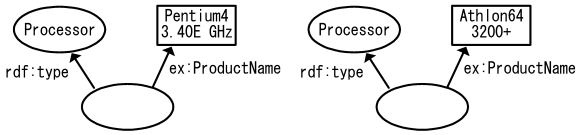

図 7: 表 1 から新たに得られるエピソード

たセルが，ラベル付きボックスとなる．ただし subClassOf や instanceOf のようなプロパティは，プロパティ名が表の中に現れるわけではなく，表の構造それ自体によって表現される．そのためそれらのプロパティは表中で現れるセルを探さず，ボックスのラベルとはしないものとする．

表 1 に対して図 6 のエピソードを与えた場合には，得られる一般化表構造は図 3(a)(b) のようになる．しかし 3 章で述べたように，図 3(b) は一つのプロパティ値に対して複数のプロパティが対応するという，誤った関係を表しているため表の解釈に用いることはできない．また解釈を与えた構造が他の構造と区別できるだけの特徴を持たない場合には，一般化を行うと，全てのボックスが同じ接続を持つ一般化表構造が得られることがある．このような一般化表構造は，特定の関係に対応せず，さまざまな構造に一致するため，そのような一般化表構造も，表の解釈に用いることはできない．

表全体を解釈する　得られた一般化表構造によって構造が表せる箇所からは，その箇所が表す関係が得られる．一般化表構造のラベル付きボックスと一致するセルのデータは，一般化表構造の獲得に用いた元のエピソードの対応するリソースやプロパティと同じ関係にあると考えられる．

表 1 中で図 3(a) の一般化表構造で構造を表せる箇所を探すと，“Class” と “Property” のラベル付きボックスには，それぞれ “Processor” と “ProductName” のセルが対応し，“Value” のラベル付きボックスには “Pentium 4 3.40E GHz”，“Pentium 4 2.80A GHz”，“Athlon 64 3200+” が対応する．“Pentium 4 2.80A GHz” は最初に与えたエピソードに含まれているので，最終的に図 7 のグラフで示した二つのエピソードが新たに得られたことになる．

新たなエピソードの獲得を行った結果，ある語がプロパティであるという記述を含むエピソードと，同じ語がクラスであるという別のエピソードが同時に得られることがある．このように，既知のエピソードや新たに得られたエピソードの間に互いに矛盾する記述がある場合にはそれらを既知のエピソードの集合から除く．さらにこのような結果が得られた原因は一般化表構造を得るときに用いたエピソードに問題があった可能性があるので，一般化表構造を得るのに用いたエピソードも除く．

以上の処理の結果，新たに得られたエピソード中のリソースやプロパティが一般化表構造を用いた構造の解釈によって獲得したのとは別の構造の中で出現していることがある．そのような場合には，新たに得られたエピソードを用いて，そのエピソードの獲得に用いられたのとは違う一般化表構造が得られることがある．そこで，表からより多くの情報を得るために，エピソードを与えて一般化表構造を得て，新たなエピソードを獲得するというサイクルを，新たなエピソードが得られなくなるまで繰り返す．

5. 評価

Web 上の表形式データに対して，4 章で述べたアルゴリズムを適用した．対象としたのは Web 上で入手できる Excel ファイルである．

表 3: 各ドメインでの得られたエピソード数・再現率・適合率

ドメイン	得られたエピソード数	適合率	再現率
価格表	439	91%	97%
連絡先リスト	89	93%	93%
統計データ	269	97%	96%

提案手法はドメインに特化した知識ベースなどを必要としないため，さまざまな表形式データへ適用することができる．各ドメインの表の特徴と得られる関係について調べるため，検索エンジンに関連キーワードを指定することで，価格表・連絡先リスト・統計データの 3 種類の表を収集した．各ドメインについて，得られたファイルのうち，実際にそのドメインに属する上位 20 個の表形式データにアルゴリズムを適用した．表 1 つあたりの獲得されたエピソード数，再現率と適合率の平均を表 3 に示す．ここでは，与えた解釈に基づき表中のデータ間の関係を正しく記述するエピソードの集合を，適合エピソードの集合としている．

ドメインによらず，それぞれの表にあわせたエピソードを与えることで表中のデータに関するクラス階層やプロパティ・プロパティ値を記述する新たなエピソードが高い適合率と再現率で得られた．しかしあまりにも表中のデータが少ない場合にはエピソード記述の手間が無視できなくなる．

6. おわりに

本研究では，表形式データから，オントロジーを獲得する手法を提案した．提案手法の特長は以下の通りである．

人手による表構造の解釈の利用　表構造があらわすデータ間の関係を，表ごとに人手によって解釈して与えることで，表によって同じ構造が異なる関係を意味する場合でも，表中のデータ間の関係が獲得できる．

多様なドメインの表の処理　表構造が表す関係に基づいて，データ間の関係を獲得するため，対象ドメインの知識ベースが不要であり，容易に様々なドメインへ適用できる．

提案手法は，人手による表構造の解釈を反映させることで，様々なドメインの表からデータ間の関係を半自動的に獲得することを目的としたものである．様々なドメインの表を収集して提案手法を適用した結果，与えた解釈に従い，表中のデータについてデータ間の関係が獲得された．

参考文献

[Chen 00] Chen,H., Tsai,S., Tsai,J.: Mining Tables from Large Scale HTML Texts. *18th International Conference Computational Linguistics*, pp.166–172 (2000).

[Pivk 04] Pivk,A., Cimiano,P. Sure,Y.: From Tables to Frames. *3rd International Semantic Web Conference*, pp.166–181 (2004).

[Tijerino 03] Tijerino,Y.A., Embley,D.W., Lonsdale,D.W., Nagy,G.: Ontology Generation from Tables. *4th International Conference on Web Information Systems Engineering*, pp.242–252 (2003).