表形式からの情報抽出手法

吉田 稔　　鳥澤 健太郎　　辻井 潤一
東京大学大学院理学系研究科情報科学専攻

{mino; torisawa; tsujii}@is.s.u-tokyo.ac.jp

1 はじめに

本稿では、WWW ページ上に存在する表を収集し、表を種類ごとに分類して出力するための手法を提案する。

WWW の魅力は、世界中に存在する多様な情報に容易に触れることができる点にあるが、その多様さゆえに、ユーザーが欲する情報にアクセスするためには、何らかの形でこれらの情報を整理する必要がある。情報をどのように整理するかについては様々な方針が考えられるが、本研究では、表、すなわち、HTML の TABLE タグ<TABLE>,</TABLE> で囲まれた部分に着目し、表に対して分類を行うことを目指した。WWW 上に存在する表を分類して出力することができれば、その出力は、各ページに存在する情報の一覧となり、Yahoo!などのディレクトリ検索と同様の、各表へのインデクスとして機能することが期待される。図 1 に、本手法が目標とする動作例を示す。この例では、自己紹介の表、PC のスペック表の集合から、自己紹介の表同士、PC のスペック表同士をそれぞれクラスタにまとめ、クラスタ内の表を統合して出力している。現在は、表を分類しクラスタにまとめる段階までの手法が完成しており、各クラスタ内の表を統合して出力する作業については実装を進めている段階である。

本手法が目指す表の分類を行うためには、まず、WWW 上に存在する表から、その表がどのような構造でどのような情報を表現しているかの解析が必要となり、さらに、解析された情報を用いて、内容的に類似した表どうしを分類する必要がある。本稿では、この「表の構造解析」「類似した内容の表の発見・分類」という 2 つのタスクに対する解決法を提案し、実験を通してその有効性を確かめる。WWW 上の表の構造解析に関しては、[1] の研究があるが、セルの色やフォント、TH タグなどの表層情報により表の記述者がある程度明示的に表の構造の手掛かりを与えていることが前提となっている。これに対して、本手法は、表に表記される語の出現頻度を手掛かりとしており、表層情報に頼らない解析が行なえる。

次節以降では、まず、一般的に表がどのような性質を持つかを明らかにし、その後、表の性質を利用して表の構造を解析する手法を述べる。さらに、得られた表の構造から、類似した内容の表をクラスタリングするための手法について述べ、最後に、これらの手法を実装したシステムについて実験を行なう。

2 用語の定義

表は、1 つまたは複数の、実世界に存在する**オブジェクト**を、**属性**により表現したものである。図 2 に、自己紹介の表の例を載せる。この表では、2 人の人間オブジェクトを、「名前」「趣味」「血液型」の 3 種類の属性の**属性値**を用いて記述している。各オブジェクトへのインデクスとして働き、その属性の属性値のみでオブジェクトを特定できる属性のことを**キー属性**と呼ぶ。図 2 の例では、「名前」がキー属性となる。

オブジェクトは、一つ以上の**クラス**に属するとし、同一クラスに属するオブジェクトは同一の属性を持つとする。同一の表に表現されたオブジェクトはすべて同一のクラスに属する。図 2 の例では、記述されたオブジェクトはともに「人間」というクラスに属することになる。

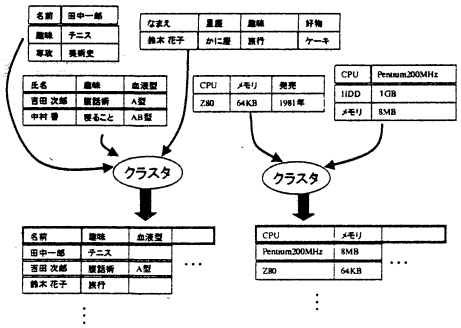
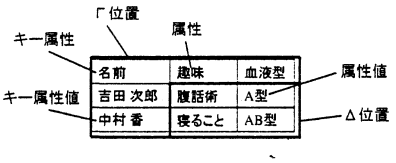


図 1: 表統合の例



スキーマ＝{名前, 趣味, 血液型}

図 2: 例：自己紹介の表

表は、複数の**セル**が 2 次元に配置された行列として捉えることができる。セル内に表記されている語を**要素語**と呼ぶ。

スキーマとは、ある表に表記された属性の集合である。

同一のクラスに属するオブジェクトを表現する表を統合するためには、まず、各表の中で、属性を記述している部分と、属性値を記述している部分を判定する必要がある。表の各セルの要素語が、属性であるか、属性値であるかに関する仮説を、表の**論理的構造**と呼ぶ。図 3 に、図 2 の表に対する論理的構造を示す。

本研究の手法では、表の論理的構造を決定するための、要素語が属性であるか属性値であるかの判定を、要素語の表記される位置の傾向を利用して行う。表の第 1 行目と第 1 列目を合わせて**Γ 位置**、Γ 位置以外の場所を**Δ 位置**と呼ぶ。これらは、経験的に、それぞれ、属性が表記されやすい場所、属性値が表記されやすい場所、と位置づけられる。

それぞれの表には、その表の**記述者**がいる。同一の記述者により記述された表は、同一のクラスに属する可能性が高い。

3 表の分類

本節では、表の論理的構造にどのような種類があるかを説明する。これは、第 5 節において、表の論理的構造を実際に

属性	属性	属性
属性値	属性値	属性値
属性値	属性値	属性値

図 3: 図 2 の表の論理的構造

タイトル	作曲者	価格
交響的舞曲	ラフマニノフ	2500
動物の謝肉祭	サンサーンス	1800
春	ドビュッシー	1500

(a)1型

HN	太郎	本名です(笑)
性別	男	
血液型	O	
生年月日	1974/4/2	もう25歳。

(b) 2-1型

メニュー	価格
コーヒー	700Yen
カレーライス	700Yen
メニュー	価格
ラーメン	700Yen
コーラ	700Yen

(c)2-2型(A種)

名前	花子	血液型	A
性別	女	誕生日	2月22日
国籍	日本	TEL	12-3456

(d)2-2型(B種)

カフェ本郷	
メニュー	価格
コーヒー	300円
カレーライス	700円
トースト	350円

(e)2-3型

図 4: 表の分類

決定する際に必要となる知識である。

表は、その論理的構造により、次の 3 つに分類することができる。

1 型：基本表 すべての属性が Γ 位置にあり、属性以外はすべて属性値である表（図 4(a)）。

2 型：発展表 属性を含む表で、基本表でないもの。

3 型：無属性表 属性を含まない表。

2 型の表は、さらに次の 3 種類に分類できる。

2–1 型：コメント付加表 基本表にコメントが付加された表。一般に、コメントは、属性値の隣に配置される (図 4(b))。

2–2 型：複合表 1 型または 2–1 型の複数の**内部表**が複数個接した状態で配置された表。内部表が同一のスキーマを共有する場合を A 種 (図 4(c))、内部表のスキーマがそれぞれで異なる場合を B 種と呼ぶ (図 4(d))。

2–3 型：Δ 属性表 1 行目にも 1 列目にも、属性でない語が記述されている表 (図 4(e))。

また、3 型の表は、次の 2 種類に分類される。

3–1 型：無属性表 1 型の表から、属性を省略した表。

3–2 型：整形表 3 型の表のうち、3–1 型でないもの全て。単なる整形目的で使われた表などがこれに属する。

一般に、表は、そのほとんどが 1 型に分類される。このため、属性は Γ 位置に表記される可能性が高い。

4 仮定

5,6 節において、表の構造解析や類似した表のクラスタリングを行なう際のアルゴリズムの根拠として、クラスやオブジェクトの性質にいくつかの仮定をおく。

クラスと属性に関する仮定 オブジェクトは、属性値の集合として表現され、その際に使われる属性は、オブジェクトの属するクラスに応じて決定される。クラスの異なるオブジェクト同士が、同一の属性を持つことは少ない。表ではオブジェクトを特定するためにクラスのすべての属性を特定する必要はなく、このことから、実際に表で記述される属性は、クラスの属性の一部である。重要な属性ほど記述される確率は高い。特に、キー属性は、1,2 型の表において、必ず記述されると仮定する。さらに、キー属性値は Γ 位置に表記される可能性が高いと仮定する。

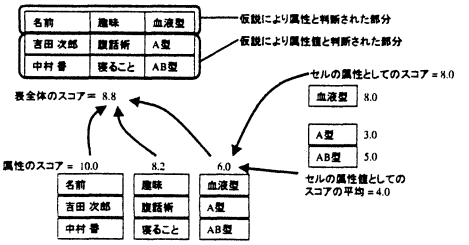


図 5: 表のスコア計算の概要

オブジェクトに関する仮定 表 T が表現するオブジェクトの集合を $O(T)$ とする。表に表現されるオブジェクトは、表が属するクラス C に含まれるオブジェクトの集合から、$|O(T)|$ 個を無作為に非復元抽出したものである。また、オブジェクトとキー属性値は一対一に対応する。

5 表の論理的構造の決定

5.1 概要

本節では、表の論理的構造を決定するための手法の概要について解説する。図 5 に、構造決定の流れを示す。本手法は、表がどの型に分類されるかを決定するに際して、表の論理的構造への可能な仮説をすべて考慮し、各々の仮説について、仮説の尤もらしさを評価する**仮説スコア**を計算する。その結果、最も仮説スコアの高くなった仮説を採用し、論理的構造を決定する。

仮説 h における表 T の仮説スコア $S_h(T)$ は、仮説 h における表 T の各属性 a_i に対する**属性スコア** $S_h(a_i)$ から計算され、属性スコア $S_h(a_i)$ は、仮説によって属性と仮定されたセル、属性値と仮定されたセル、コメントと仮定されたセル、のそれぞれの**セルスコア**の平均値をとることで計算される。

5.2 セルスコアの計算

セルスコアは、属性と仮定されたセルならば属性としての、属性値と仮定されたセルならば属性値としての、コメントと仮定されたセルならばコメントとしての、要素語の尤もらしさを表すスコアである。

仮説 h におけるセル c のスコアは、セル c が、仮説 h において、どの種類のセルと見なされるかに応じて、以下のように定義される。ここで w は、セル c に表記された要素語である。

$$S_h(c) = \begin{cases} s_1(w) - s_2(w) - s_3(w) - s_4(w) \\ \quad (w \text{ が属性と判断された場合}) \\ s_2(w) + s_3(w) + s_4(w) - s_1(w) \\ \quad (w \text{ が属性値と判断された場合}) \\ s_4(w) \times 2 - s_1(w) - s_2(w) - s_3(w) \\ \quad (w \text{ がコメントと判断された場合}) \end{cases}$$

$s_1(w)$ は、w の属性としての尤もらしさを、$s_2(w)$、$s_3(w)$、$s_4(w)$ は、w の属性値或いはコメントとしての尤もらしさを表すスコアであり、以下で定義される。

- $s_1(w) = \begin{cases} \log_2 \lceil \frac{\Gamma(w)}{\Delta(w)} \rceil & (\text{if } ((\Gamma(w) \geq \Delta(w)) \\ & \cap(\Delta(w) > 0))) \\ \log_2 \Gamma(w) & (\text{if } \Delta(w) = 0)) \\ 0 & (\text{otherwise}) \end{cases}$
- $s_2(w) = \begin{cases} \lceil \log_2 \Delta(w) \rceil & (\text{if } \Delta(w) \geq \Gamma(w)) \\ 0 & (\text{otherwise}) \end{cases}$
- $s_3(w) = r \times c_1$
  (r:w における数字の比率、c_1:定数)

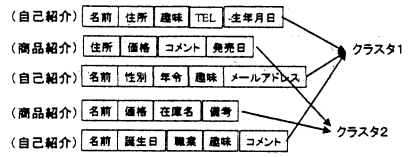


図 6: 表のクラスタリング例

- $s_4(w) = \begin{cases} c_2 & (w \text{ が文章の場合}) \\ 0 & (\text{それ以外}) \end{cases}$

  (c_2:定数。文章であるか否かは、句点「。」や、「!」「?」、語の長さなどを用いて判定される。)

コメントは、主に文章として表記されるという特徴を持つため、要素語が文章であった場合にセルスコアが高くなる。また、数値が含まれる要素語は、属性値となる可能性が高い。Γ 値($\Gamma(w)$ で表記) とΔ 値($\Delta(w)$ で表記) は、属性と属性値の分類を行なうために用いられる尺度であり、以下で定義される。

- 語 w の Γ 値とは、語 w を Γ 位置に表記する記述者の数である。
- 語 w の Δ 値とは、語 w を Δ 位置に表記する記述者の数である。

同一の記述者は、クラスの中で注目するオブジェクトが同一になり、同一のオブジェクト、ひいては同一のキー属性を表記する可能性が高いため、同一記述者の表は区別せず、記述者の数でスコアを定義している。

一般に、w が属性のときは、w は殆どの場合 Δ 位置ではなく Γ 位置に表記される為、$\Gamma(w) \gg \Delta(w)$ となる。また、w が非キー属性値のときは、w は殆どの場合 Γ 位置ではなく Δ 位置に表記される為、$\Delta(w) \gg \Gamma(w)$ となる。

属性とキー属性値は Γ 位置に表記される可能性が高く、非キー属性値は Δ 位置に表記される可能性が高い。また、表 T のオブジェクトがクラス C に属するとき、C 内のそれぞれのオブジェクトのキー属性値が表記される確率は、オブジェクトに関する仮定により、$\frac{|O(T)|}{|C|}$ となるが、通常、あるクラスに属するオブジェクトの数は無数にあり、$|C|$ は非常に大きな値となる。このため、あるキー属性値が表記される回数の期待値は極めて小さくなり、結果的に $\Gamma(w)$ 値も小さくなることが予想される。

以上により、「Γ 値の Δ 値に対する比が高い」「Γ 値が高い」の 2 つの性質を満たす語が、属性である可能性が高いことが予想される。$s_1(w)$ は、この性質に基づき、w が属性である場合に高い値になることを目的として定義されている。また、$s_2(w)$ は、w が属性値である場合に高くなることを目的として定義されている。

6 表のクラスタリング

本節では、決定された各表の論理的構造に基づき、同一のクラスに属する表同士を同一の集合に集めるためのクラスタリング手法について述べる。

図 6 に、クラスタリングの例を示す。図 6 には、3 つの自己紹介の表と、2 つの商品紹介の表の、それぞれのスキーマが示されている。本手法の目的は、類似した表同士 (ここでは「自己紹介」の表同士、「商品紹介」の表同士) をそれぞれ一つのクラスタに集めることである。

6.1 方針

同一クラスに属するオブジェクトは、同一の属性集合を持つ。例えば、人間クラスに属するオブジェクトは、いずれも「年齢」「性別」などの属性を持つ。実際に、表においてオブジェクトを表現するために用いられる属性は、クラスの全ての属性の極一部であるが、スキーマ内の属性が類似している表

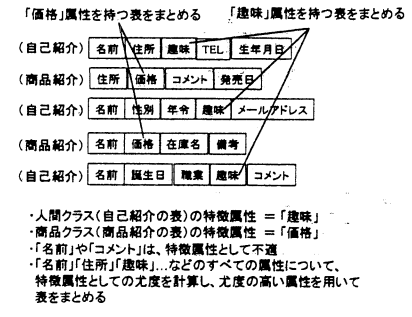


図 7: 特徴属性を用いたクラスタリング

同士が同一のクラスに属している可能性は高い。このため、スキーマの類似度を用いて表のクラスタリングを行なう手法が考えられる。しかし、この方針には、異なるクラス間で共有される属性がうまく扱えないという欠点がある。例えば、「名前」という属性は、人間クラスの他、「犬」クラス、「店」クラスといった、様々なクラスで用いられる。もしも、人間クラスに属するオブジェクトと店クラスに属する表のスキーマがどちらも「名前」「住所」「電話番号」の 3 つの属性で構成されていた場合、これらの表が同一クラスに属する表と判断されてしまう。

この問題を回避するため、本研究では、各クラスの属性のうち、他のクラスで用いられることが少なく、そのクラスに特徴的に現れる**特徴属性**を定義し、特徴属性**のみ**を用いて各クラスに属する表を収集する手法を提案する (図 7)。ここで、あるクラス C の特徴属性とは、クラス C に属する表に表記される頻度が高く、他のクラスに属する表へ表記される頻度が低い属性である。

6.2 アルゴリズム

属性 n の、特徴属性としての尤もらしさは、n を含むスキーマ全てによって構成されるスキーマ集合において定義される、**平均重複度**と、**平均特異性**を用いて計算される。

ここで、平均重複度とは、クラスタ内のスキーマに含まれるすべての属性について重複度を計算し、その平均をとることによって求められる値であり、スキーマ集合におけるある属性の重複度とは、その属性がその集合内のいくつのスキーマに含まれているかを表す指標である。例えば、図 7 の例において、クラスタ 1 で「名前」属性は 3 つ全てのスキーマに含まれているため、その重複度は 3 である。また、「年令」属性は、1 つのスキーマのみに含まれているため、重複度は 1 となる。S 内のスキーマの属するクラスの一貫性が高い程、同一の属性が用いられる可能性が高く、平均重複度が高くなる。

また、平均特異性とは、クラスタ内のスキーマに含まれるすべての属性について特異性を計算し、その平均をとることによって求められる値であり、スキーマ集合におけるある属性の特異性とは、その属性を含む集合内のスキーマの数と、その属性を含むすべてのスキーマの数の比で定義される。図 6 の例におけるクラスタ 1 について考えると、「名前」属性は、クラスタ 1 内で 3 つのスキーマに含まれ、全体では 4 つのスキーマに含まれるため、特異性は $\frac{3}{4}$ となる。平均特異性は、クラスタ内のスキーマに含まれるすべての属性について特異性を計算し、その平均をとることによって求められる。異なるクラス間で、属性の共有が起こりにくいという仮定に基づくと、S があるクラスに属するスキーマの大部分を含んでいれば、S に含まれるスキーマと、含まれないスキーマの間で属性の共有が起こりにくく、平均特異性は小さな値となる。

これらの値から、ある属性 n の**優秀度**$goodness(n)$ を、

$$goodness(n) = red(S_n) \times uniq(S_n)$$

として定義する。ただし、S_n は、属性 n を含む全てのスキーマの集合であり、$red(S_n)$、$uniq(S_n)$ は、それぞれ S_n の

```
S = スキーマの集合;
N =S における属性の集合;
それぞれの n ∈ N に対し goodness(n) を計算する;
N の要素を、goodness(n) に関して降順にソートする;
for(i = 1; i ≤ |N|; i = i + 1){
    B = A ∩ S_{n_i};
    if (|B| < |S_{n_i}|  0.2){
        j := j + 1;
        C_j := S_{n_i} − B;
        A := A ∪ B;
    }
}

{C_i} を結果として返す,
```

図 8: クラスタリングのアルゴリズム

型	表の数	正解数	正解率
1	62	42	0.68
2-1	1	1	1.00
2-2	2	1	0.50
2-3	1	1	1.00
3-1	22	15	0.68
3-2	30	26	0.87
計	118	85	0.73

表 1: 表の認識における正解率

平均重複度、平均特異性であり、上記の例で示された方法で計算される[1]。

実際のクラスタリングのアルゴリズムを図 8 で説明する。本手法は、属性の集合から、属性を、優秀度の高いものから先に選択し、選択された属性により表を収集し、クラスタを形成する。この際、新たに形成されたクラスタと、すでに形成されたクラスタとの間で重複する表の割合があるしきい値（現在は 20 %）を越えないことを条件とする。

7 実験と考察

7.1 表の論理的構造の認識率

実験のためのデータとしては、InfoWeb (http://village.infoweb.or.jp/) からユーザーの個人ページを取得し、それらのページから抽出した表 (ここでは、TABLE タグ (<TABLE>,</TABLE>) で囲まれた部分のうち、「内部に再帰的に表を含まない」「2 行以上かつ 2 列以上である」「画像を含まない」「ROWSPAN、 COLSPAN を含まない」といった条件を満たす表) を実験データとして用いる。

本手法を JAVA を用いて実装し、得られた表の集合に適用した。実験セットから無作為に抽出した 118 個の表（うち、<TH> タグを含むものは 7 つであった）についての解析結果を表 1 に示す。正解率は約 70 %であった。

また、上記の結果では、2 型の表の頻度が低いため、2 型の表に対する認識率が不明である。そのため、2 型の表のみを 20 個収集し、それらに対する認識率を測定した。但し、2-1 型 (a) の表については、1 型と同様に認識が可能なため、測定から除外してある。結果を表 2 に示す。正解率は約 80 %であった。

傾向としては、主に 1 型と 3-1 型で正解率が低い。1 型の誤認識は主に、「スキーマ内の属性の Γ スコアが低い」ことに起因し、3-1 型の誤認識は主に、「属性値語の Γ スコアが高い」ことに起因していた。

7.2 クラスタリング結果

実際にクラスタリング手法を表の集合に適用し、その傾向を見た。結果として、「趣味」属性で収集された自己紹介のクラスタ、「CPU」属性で収集された PC のスペック表のクラスタなどが出力された。図 9 に、「資本金」によって収集された会社案内クラスタのスキーマの一部を示す。クラスタリング結果の詳細な解析は、今後行う予定である。

[1]紙面の都合上、計算式は省略する。

型	表の数	正解数	正解率
2-1	7	7	1.00
2-2(B 種)	10	7	0.70
2-3	3	2	0.67
Total	20	16	0.80

表 2: 2 型の表に対する認識の正解率

```
The number of schemata = 103
(本社 tel fax e-mail 代表取締役 設立 資本金 年商 従業員数)
(本社 創立 営業内容 資本金 年商)
(社名 創業 資本金 役員 会計士 主要取引銀行 所在地 関連会社 主要取引先)
(設立 資本金 会員数 本部 会長 副会長 設立目的 活動内容 会員規約 近況報告 備考)
(設立 資本金 本社)
(所在地 tel/fax 設立 資本金 代表者 売上高 主要取引先 取引銀行 事業内容 関連会社 主要生産品目)
(会社名 所在地 電話・fax 代表者 資本金 業務内容 設備 主要取引先 e?mail)
(商号 代表者 創業 資本金 年商 事業内容 従業員数)
...
```

図 9: クラスタリング結果の例

8 おわりに

本稿では、WWW 上の表の構造を解析し、類似した表のクラスタを出力するための手法について論じた。本研究の手法により出力される表のクラスタは、例えば人間クラスのオブジェクトには「趣味」という属性があり、「趣味」属性は「旅行」「音楽鑑賞」などの属性値をとる、といった、一種のオントロジーとしての応用も考えられる。

現在のアルゴリズムは、表の構造解析、クラスタリング共に、未だナイーブなものであり、認識の正解率も 7 割前後であるが、これから手法に改良を加えていくことによりさらなる性能の向上を図る予定である。クラスタリング結果の詳細な解析についても、今後行なっていく予定である。

また、現在、収集された表のクラスタ内で、同一の内容を表す属性（「住所」と「住処」など）を統合する手法の開発も進めている。これは、属性値の類似度から同一内容を表現する属性のクラスタリングを行う手法である。予備的な実験では、「メールアドレス」と「E-mail」「e メール」、「誕生日」と「生年月日」などの属性どうしの統合が確認されている。

参考文献

[1] 伊藤 史朗、大谷 紀子、上田 隆也、池田 祐治. 1999. 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.

[2] 伊藤 史朗、上田 隆也、池田 祐治. 1998. 分散情報源に対する情報エージェントのための事例に基づくフレームマッピング. 電子情報通信学会論文誌 VOL.J81-D-I No.5. pp.433–442.

[3] Matthew Hurst and Shona Douglas. 1997. Layout and Language: Preliminary investigations in recognizing the structure of tables. In *Proceedings of Fourth International Conference on Document Analysis and Recognition (ICDAR)*, pp.1043–1047.

[4] 嶋田 和孝、遠藤 勉. 1999. 製品性能表からの特徴データの抽出. 情報処理学会研究報告 99-NL-133, pp.107–113.

[5] 斎藤 公一. 1999. 数値情報をキーとした新聞記事からの情報抽出. 修士論文. 横浜国立大学大学院工学研究科電子情報工学専攻.