

Webの表を対象とした属性の自動識別

大前 信弘[†] 黄瀬 浩一[†]

[†] 大阪府立大学大学院工学研究科 〒599-8531 大阪府堺市学園町1-1
E-mail: [†]ohmae@m.cs.osakafu-u.ac.jp, ^{††}kise@cs.osakafu-u.ac.jp

あらまし 近年のインターネットの普及により、Web上には大量の情報が存在している。この情報を利用するための技術として、情報抽出が期待されている。この技術では、抽出する情報の属性をあらかじめ決めておかなければいけない。しかし、抽出しようとする情報について知識が乏しい人は、どの属性を指定すればよいか分からないという問題がある。そこで、本稿では、属性を自動で決定する手法を提案する。対象は、すでに属性と属性値から構成されている表とする。表には、属性が1行目または1列目に偏って存在するという共通な性質がある。本手法の特徴は、(1) 偏りの推定への χ^2 検定の利用、(2) 属性を表す行・列という構造的制約を用いた属性の発見、(3) 再検索を用いた属性の検証の3点からなる。本稿では、Webから得た13,390個の表(385個の属性を含む)を対象に表の構造解析の実験を行い、F値79%を得た。

キーワード Web, 表, 属性, χ^2 検定

Automatic recognition of attributes from tables in web pages

Nobuhiro OHMAE[†] and Koichi KISE[†]

[†] Osaka Prefecture University 1-1, Gakuenchou, Sakai, Osaka, 599-8531 Japan
E-mail: [†]ohmae@m.cs.osakafu-u.ac.jp, ^{††}kise@cs.osakafu-u.ac.jp

Abstract Information extraction enables us intelligent access to a huge amount of information stored as Web pages. This technique requires the user to determine attributes of information the user needs. It is, however, not easy for the user who only has incomplete knowledge about the information to specify its exact attributes. In this report we propose a method of automated extraction of attributes in response to a topic specified by the user. As the information source, we focus on tables on the Web that contain attributes of the topic. The method is based on the fact that attributes are biased to be in the first column and row of the tables. The characteristic points of the method are as follows: (1) the bias is estimated using the chi-square test, (2) extraction of attributes using structural constraints on rows and columns including attributes, (3) retrieval-based validation of extracted attributes. From the experimental results on 13,390 tables including 385 attributes, the method extracted the attributes with the F-measure of 79

Key words Web, Table, Attribute, χ^2 test

1. はじめに

近年、Web上の情報を利用するための技術として情報抽出への期待が高まっている。情報抽出とは、あらかじめ指定したタイプの情報を抽出する技術である。ここで指定すべき情報とは、例えば学会に興味がある場合には、開催地、日程などである。本稿ではこれらを対象とする情報(学会)の属性と呼ぶ。

属性と属性値から構成されている情報源に表がある。

表は理解しやすく多くのWebページで使われている。このことから、Webページの表は情報抽出の情報源として利用できると考えられる。具体的には、表の要素のうちどれが属性でありどれが属性値であるかということを識別することが情報抽出に相当する。要素のうち属性以外は属性値であることが多いので、表の属性を自動識別することは重要となる。

従来研究には、教師あり学習を用いる方法[1][2]とそれ以外の方法がある。教師あり学習を用いる方法は、未

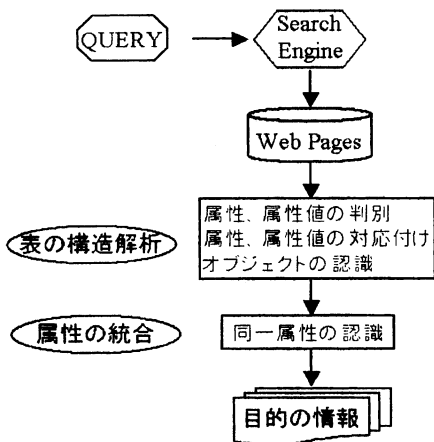


図1 属性の自動識別の処理手順

知のデータに対応できない問題がある。それ以外の方法には、嶋田らの手法 [3]、佐藤らの手法 [4]、Chen らの手法 [5]、Yoshida らの手法 [6] がある。しかし、これらの手法はいずれも表での属性の位置が決まっているという前提条件を必要とする。よって、すべての表には対応できないという問題点がある。

そこで、本研究では、大半の表に共通な性質を用いる方法を考える。属性には、表の1行目または1列目に偏って存在する性質がある。この性質を利用することにより、属性が識別できると考えられる。この方法は、特殊な前提条件を必要とせず、より多くの表に対応できる。本手法の特徴は、(1) 偏りの推定に χ^2 検定を利用 (2) 属性を表す行・列という構造的制約を用いた属性の発見 (3) 再検索を用いた属性の検証の3点からなる。実験の結果 F 値約 79.3% を得て、提案手法の有効性を確認した。

以下、2. で属性の自動識別方法について説明する。3. と 4. で本研究の提案手法について説明し、5. と 6. ではその手法に対する実験とその結果について考察する。7. は本稿のまとめである。

2. 属性の自動識別方法

本稿で提案する Web の表を対象とした属性の自動識別の処理は、図1に示されるように表の構造解析、属性の統合の2つの処理で構成される。以下で、各処理について説明する。

2.1 表の構造解析

表において属性とは、ある事物の性質を問う項目のことをいう。また、属性に対する具体的な性質を表す項目を属性値といい、性質を問われる対象である事物そのものをここではオブジェクトと呼ぶ。図2の例であれば、「学会名」、「開催月日」、「場所」が属性「情報処理学会」、「山形大学」などが属性値。1つ1つの学会（各行に対応）

学会名	開催月日	場所
情報処理学会	5/21(水)	東京大学
機会学会	5/30(金)	山形大学
電気学会	7/10(月)	同志社大学

図2 表の例

がオブジェクトとなる。

もし、パソコンのことにについて全く知識がない人が、パソコンを購入するために情報抽出しようとした場合、「パソコン名」、「価格」程度しか属性として思いあたらないことが考えられるが、それだけでは十分な比較ができない。よって、抽出対象の情報の属性を自動的に決定する処理は、重要であると考えられる。

そこで、この処理では表の各項目の関係を解析し、抽出する属性を自動で決定する。

まず、Web ページから表を抽出する。Web ページでは、TABLE タグを用いて表が書かれている。よって、TABLE タグを取り出すことによって、表を抽出することが出来る。

次に、表の各項目の関係を解析する処理について説明する。表の各項目は、属性と属性値に大きく分けられる。属性と属性値の関係について考えると、各属性値に対する属性はどれなのかを認識する必要がある。次に、属性値同士の関係について考えると、どの属性値が同じオブジェクトを表しているのかを認識する必要がある。つまり、以下の処理が必要となる。まず、表の各要素に対し、属性と属性値の判別をする。次に、属性と属性値の対応付けをする。そして、オブジェクトを認識する。これらの処理によって表の構造を解析できる。

2.2 属性の統合

属性の統合とは、抽出対象の情報に関連し、同じ意味を表す属性をまとめる処理である。以下で統合の必要性について述べる。

表の構造を解析することによって、属性と属性値が対応付けられ、オブジェクトが認識される。しかし、多くの Web ページから表を収集するため、例えば「氏名」と「名前」、「価格」と「値段」のように同じ意味の属性が違った単語で書かれていることがある。このままでは利用者にとって分かりやすい情報を示したとは言えない。

よって、表の構造解析から得られた情報の中で、同一となる属性を認識して属性を統合する必要がある。具体的には、対応する属性値が似た属性同士を同一属性と認識すればよい。すると、同じ意味を持つ複数の属性がない分かりやすい情報が得られる。

3. 表の構造解析法

次に表の構造解析法について詳述する。この手法は、属性候補の抽出、属性の行と列の抽出、再検索という3つの処理からなる。以下で各処理と、属性候補の抽出の処理で用いる χ^2 検定について説明する。

3.1 χ^2 検定

例えば貨幣を10回投げたとする。このとき表が9回出た。この現象は、偶然出たとする考え方と、表が出やすいのではないかとという考え方がある。この現象は偶然なのか、表が出やすいのかを統計的に判断するのに検定というものがある。その中でも χ^2 検定では以下の式(1)の χ^2 値を用いて判断する。

$$\chi^2 \text{ 値} = \sum_{i=1}^k \frac{(N_i - m_i)^2}{m_i} \quad (1)$$

ここで、 k は事象 $A_i (i = 1, \dots, k)$ の数 (貨幣の場合は表か裏なので2)、 N_i は試行を n 回繰り返して事象 A_i が現れる回数、 m_i は事象 A_i の期待度数 (事象 A_i が起こる確率が p_i なら $m_i = np_i$) である。

貨幣のように、起こりえる事象が2つの場合は、イエーツの補正といわれる式(2)を用いて χ^2 値を求める。

$$\chi^2 \text{ 値} = \sum_{i=1}^2 \frac{(|N_i - m_i| - 0.5)^2}{m_i} \quad (2)$$

χ^2 値がある値よりも大きい時、表が出やすいと判断できる。その値の決め方に有意水準が使われる。

3.2 属性候補の抽出

この処理では、 χ^2 検定を用いて属性の候補となる表の要素を抽出する。この処理は以下の4つの手順からなる。各手順と処理を以下で説明する。

step1 Web ページを検索する

必要とする情報が得られるような検索質問で、Web ページを検索する。

step2 TABLE タグを取り出す

Web ページから TABLE タグで囲まれている部分を取り出す。

step3 表の要素の出現頻度を出現位置を考慮してカウントする

表の要素が、表の1行目または1列目に出現する回数と表のその他の部分に出現する回数を分けてカウントする。

step4 χ^2 検定を用いて属性候補を識別する

表の一般的な要素は、表のどの位置に出現するかという偏りはないはずである。しかし、属性となる要素は、表の1行目または1列目に偏って出現すると考えられる。そこで、 χ^2 検定を用いて、表の1行目または1列目に偏って出現する要素を識別する。起こりえる事象は1行目または1列目とその他の2つなので、 χ^2 値は以下の(3)式を用いて計算する。

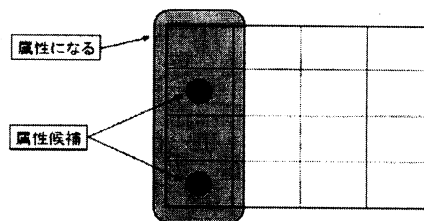


図3 属性の行と列の抽出

$$\chi^2 \text{ 値} = \frac{(|A - EF_A| - 0.5)^2}{EF_A} + \frac{(|B - EF_B| - 0.5)^2}{EF_B} \quad (3)$$

ここで、 A は要素が表の1行目または1列目に出現した観測度数、 EF_A は要素が表の1行目または1列目に出現する期待度数、 B は要素が表の1行目または1列目以外に出現した観測度数、 EF_B は要素が表の1行目または1列目以外に出現する期待度数である。4行4列の表の場合、要素が表の1行目または1列目に出現する期待度数は $\frac{7}{16}$ 、それ以外の部分に出現する期待度数は $\frac{9}{16}$ となる。

この χ^2 値が有意水準 α に対する χ^2 値を上回れば、その表の要素は、表の1行目または1列目に偏って出現するといえるので、属性の候補とする。

3.3 属性の行と列の抽出

この処理では、3.2の処理で得られた属性候補を表と照らし合わせることによって、属性候補の抽出では抽出することの出来なかった属性を抽出する。具体的には、図3に示されるように、属性候補が含まれている行または列の他の要素も属性として抽出する。この処理は、1つの行または列に属性候補が存在すれば、その行または列の要素がすべて属性ではないかという考えに基づいている。

本手法では、1つの行または列に属性候補が2つ以上存在すれば、その他の要素も属性と識別する。

3.4 再検索

3.3の処理により、偶然違う情報の表に属性候補が含まれているために、誤った文字列を属性と判別してしまう場合がある。その誤りを除くために、この処理では、3.3の処理で属性と識別された文字列を元の検索質問に追加して再検索する。そして、再検索して得られたページ内の表の1行目または1列目に追加した文字列が含まれているかどうかを確認する。このとき、前の処理までに用いたページは除く。含まれていれば、その文字列は属性である可能性が高いので属性と識別する。逆に、含まれていなければ属性である可能性が低いので属性ではないと識別する。この処理を行うことによって、3.3の結果の適合率を上げることが出来る。

4. 属性の統合法

この処理では、表の構造解析の結果得られた同じ意味

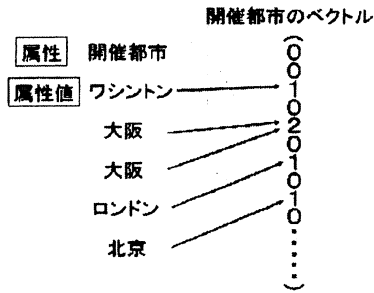


図4 ベクトル化

を持つ属性を統合する。本手法では、属性を持つ属性値の類似度を用いる。この処理は以下の3つの手順からなる。各手順と処理を以下で説明する。

step1 属性値の収集

各属性に対する属性値をすべての表から収集する。

step2 ベクトル化

各属性が持つ属性値を用いて属性をベクトル化する。例えば図4のように「開催都市」という属性が5個の属性値を持つ場合を考える。まず、各属性値が対応するベクトルの要素に出現頻度の値を与える。そして、ベクトルの大きさが1になるように正規化する。属性をベクトルで表したものを A_x は次の式(4)のように表せる。ここで、 n は属性値の総数である。

$$A_x = (x_1, x_2, \dots, x_n) \quad (4)$$

step3 類似度の計算

Jaccard 係数を用いて、属性間の類似度を計る。属性 A_x と A_y の Jaccard 係数 $\sigma(A_x, A_y)$ は次の式(5)を用いて計算する。

$$\sigma(A_x, A_y) = \frac{\sum_{i=1}^n x_i \cdot y_i}{\sum_{i=1}^n x_i^2 + \sum_{i=1}^n y_i^2 - \sum_{i=1}^n x_i \cdot y_i} \quad (5)$$

この値が閾値を上回れば、これらの属性は同じ意味を持つとする。

5. 表の構造解析の実験

属性の自動識別を行う提案手法の有効性を検討するために、属性候補の抽出、属性の行と列の抽出、再検索の実験を行った。

5.1 実験条件

この実験には、Google Web APIs[7]を用いて、4つの検索質問で検索した上位1000件の内、PDFファイルを除くWebページを対象とした。各検索質問とページ数などを表1に示す。

5.2 属性候補の抽出に関する実験

属性候補の抽出手法の有効性を検討するための実験を

表1 実験データの内訳

検索質問	ページ数	表数	列数	行数	属性数
baseball player roster	356	3498	8506	15403	115
international conference date location	251	2354	4580	15575	131
painting title list	310	4475	12129	21499	72
wine name price list	299	3063	7659	53773	67

表2 属性候補の抽出の実験結果

検索質問	適合率 (%)	再現率 (%)	F 値 (%)
baseball player roster	69.4	21.7	33.1
international conference date location	76.3	34.4	47.4
painting title list	47.7	29.2	36.2
wine name price list	45.3	35.8	40.0

F 値の平均 = 39.2%

表3 誤抽出の割合 (属性候補の抽出)

検索質問	誤りの特徴 (1)	誤りの特徴 (2)
baseball player roster	27.3%	72.7%
international conference date location	42.9%	57.1%
painting title list	21.7%	78.3%
wine name price list	38.0%	62.0%

行った。 χ^2 検定での有意水準 α は 0.1% とした。実験の評価には、再現率 $R = |C|/|A|$ 、精度 $P = |C|/|B|$ 、再現率 R と精度 P で表される F 値 $F = \frac{2RP}{R+P}$ を用いた。ここで、 $|A|$ は全属性数、 $|B|$ は結果として得られた属性数、 $|C|$ は $|A|$ 内の正解数である。

実験結果を表2に示す。F 値の平均として 39.2% の結果が得られた。

誤りは誤抽出と未抽出に大別できる。誤抽出の特徴としては、以下の2つが挙げられる。また、その特徴に対する各検索質問の誤りの割合を表3に示す。

(1) 違う情報をあらわす表の属性

(2) 1行目または1列目に多く出現した属性値

(1)の具体例としては、野球選手の情報の属性を抽出する際に、「price」や「author」などの本の情報の属性が抽出されたものがある。(2)の具体例としては、野球選手の情報の属性を抽出する際に、図5の1列目のように野球選手の個人名が抽出されたものがある。

未抽出の属性を、出現頻度で分けた場合の割合を表4に示す。出現頻度が1回の属性が、全体の誤りの内の約45%を占めていることがわかる。本手法では、統計を用いているので、出現頻度が低いものには十分に対応でき

NAME	POS	BAT	THW	AGE	HT	WT	BORN
Jason Anderson	P	L	R	26	6-0	188	Danville, IL
Colter Bean	P	L	R	28	6-6	255	Anniston, AL
Shawn Chacon	P	R	R	27	6-3	220	Anchorage, AK
Jorge DePaula	P	R	R	27	6-1	160	Yamaza, DR
Wayne Franklin	P	L	L	31	6-2	201	Wilmington, DE
Sean Hann	P	R	L	24	6-4	215	Fort Worth, TX
Randy Johnson	P	R	L	42	6-10	231	Walnut Creek, CA
Mike Mussina	P	L	R	36	6-2	190	Williamsport, PA
Carl Pavano	P	R	R	29	6-5	241	New Britain, CT
Scott Proctor	P	R	R	28	6-1	198	Stuart, FL
Mariano Rivera	P	R	R	36	6-2	185	Panama City, Panama

図 5 誤りの特徴 (1) の例

表 4 出現頻度による誤りの割合

検索質問	1 回	2 回以上
baseball player roster	43.3%	56.7%
international conference date location	52.3%	47.7%
painting title list	37.3%	62.7%
wine name price list	48.8%	51.6%
平均	45.4%	54.6%

表 5 属性の行と列の抽出の実験結果

検索質問	適合率 (%)	再現率 (%)	F 値 (%)
baseball player roster	94.1	82.6	88.0
international conference date location	66.9	85.4	75.0
painting title list	67.5	75.0	71.1
wine name price list	71.1	80.6	75.6

F 値の平均 = 77.4%

Brilliant Bubbles				
Wine#	Wine Name	Region	Vintage	Price
25	Chateau Gossier	Champagne	N.V.	€50.00
20	Cremant de Bourgogne	Andre Delorine, Burgundy	N.V.	€34.00
320	Picpoul de Pinet	Comte Armand, Burgundy	2002	€24.00

図 6 成功例

ていない。

5.3 属性の行と列の抽出に関する実験

4.2 の結果を用いて、属性の行と列の抽出手法の有効性を検討するための実験を行った。実験の評価には 4.2 と同じものを用いた。

実験結果を表 5 に示す。F 値の平均として 77.4% の結果が得られた。成功例を図 6 に示す。属性である 2 行目を正しく抽出することが出来た。誤抽出の特徴としては以下の 2 つが挙げられる。また、その特徴に対する各検索質問の誤りの割合を表 6 に示す。

(1) 関係のない情報の表に属性候補が含まれていたために誤った属性が抽出されたもの

(2) 属性候補の抽出の時点で誤っていたものに影響を受けて誤った属性が抽出されたもの

(1) の例を図 7 に示す。目的とする情報がワインの

表 6 誤抽出の割合 (属性の行と列の抽出)

検索質問	誤りの特徴 (1)	誤りの特徴 (2)
baseball player roster	0.0%	100.0%
international conference date location	29.1%	70.9%
painting title list	30.8%	69.2%
wine name price list	47.4%	52.6%

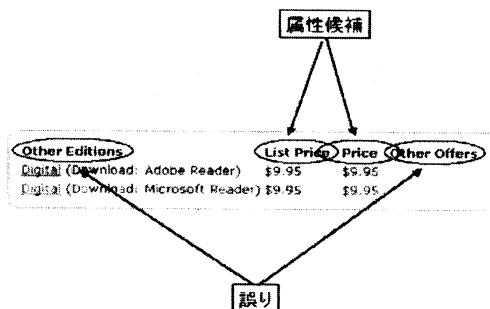


図 7 誤りの例

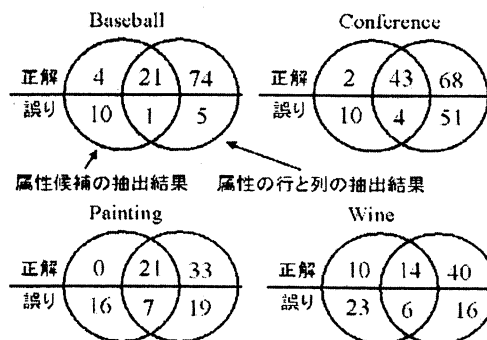


図 8 属性の行と列の抽出処理での属性数の増減

き、図の表は目的とする表ではない。しかし、ワインの情報を表す属性である「price」や「list price」が含まれているので、「Other Editions」と「Other Offers」の要素までもが属性と識別されてしまった。

属性の行と列の抽出の処理を行うことで、属性候補の抽出の処理で得られた属性候補を失う数、新たに得られる属性の数、属性候補で属性の行と列の抽出の処理後も属性と識別された数を各検索質問ごとに図 8 に示す。全体的に、この処理によって属性をあまり失うことなく、新たな属性を得ることができた。また、属性候補の抽出の処理での多くの誤りを取り除くこともできた。

5.4 再検索に関する実験

4.3 の結果を用いて、再検索手法の有効性を検討するための実験を行った。実験の評価には 4.2 と同じものを

表 7 再検索の実験結果

検索質問	適合率 (%)	再現率 (%)	F 値 (%)
baseball player roster	95.4	72.2	82.2
international conference date location	78.9	80.2	79.5
painting title list	78.5	70.8	74.5
wine name price list	82.8	79.1	80.9

F 値の平均 = 79.3%

表 8 取り除いた誤り数と割合

検索質問	取り除いた誤り数	全誤りに対する割合
baseball player roster	1	16.7%
international conference date location	27	49.1%
painting title list	12	46.2%
wine name price list	8	36.4%

全誤りに対する割合の平均 = 37.1%

用いた。

実験結果を表 7 に示す。F 値の平均として 79.3% の結果が得られた。適合率は、すべて検索質問で向上させることができた。しかし、再現率は、すべての検索質問で下がってしまった。その理由としては以下のことが挙げられる。検索質問に追加する文字列が、「special offer price pound」のように多くの語を含む時、再検索により得られるページ数は少なくなる。その結果、1 行目または 1 列目に、追加した文字列を含む表を得ることができなかった例が多く見受けられた。

再検索によって取り除くことができた誤りの数と全誤りに対する割合を表 8 に示す。全誤りに対する割合の平均が約 40% と、多くの誤りを取り除くことができたことが分かる。この処理により全検索質問で、再現率が下がっているが、F 値の平均を向上させることができた。

6. 属性の統合の実験

5. の結果を用いて、属性の統合手法の有効性を確認するための実験を行った。Jaccard 係数 σ の閾値は 0.1 とした。

実験の評価には式 (6) により求められる再現率 R 、適合率 P 、再現率 R と適合率 P で表される F 値を用いた。

$$R = \frac{|X \cap Y|}{|Y|} \quad P = \frac{|X \cap Y|}{|X|} \quad (6)$$

ここで、 $|X|$ は統合される属性数、 $|Y|$ は結果として統合された属性数である。

実験結果を表 9 に示す。F 値の平均として 70.1% の結果が得られた。

この実験で誤って統合してしまった例は、検索質問が野

表 9 属性の統合の実験結果

検索質問	適合率 (%)	再現率 (%)	F 値 (%)
baseball player roster	53.7	92.2	67.9
international conference date location	64.3	85.7	73.5
painting title list	72.6	80.8	76.5
wine name price list	50.0	83.0	62.4

F 値の平均 = 70.1%

球選手のときの「throw」と「bat」である。「throw」は投げ方を表す属性、「bat」は打ち方を表す属性である。これらの属性は両方とも右か左かという属性値である「L」と「R」を多く持つ。その結果、これらの属性を誤って統合してしまった。統合できなかった例は「name」と「wine name」である。ワインの名前は「Chateau Montelena Napa」や「Chateau Montelena Napa(2001)」など表記方法が様々なため、共通な属性値が少なく、統合することができなかった。成功例には次のものがある。野球選手の体重を表す属性には「weight」、「w」、「wgt」、「wt」の 4 つがあるが、これらを正しく統合することができた。

7. おわりに

本稿では Web の表を対象とした属性の自動識別について述べた。本手法の特徴は、(1) 偏りの推定に χ^2 検定を利用 (2) 属性を表す行・列という構造的制約を用いた属性の発見 (3) 再検索を用いた属性の検証の 3 点からなる。実験の結果、表の構造解析で F 値 79.3%、属性の統合で F 値 70.1% を得て、提案手法の有効性を確認した。

今後の課題としては、表以外の情報源にも適用可能な手法について研究を行うことがある。

文 献

- [1] Kumi Itai, Atsuhiko Takasu, Jun Adachi: "Information Extraction from HTML Pages and its Integration", SAINT Workshops, 2003.
- [2] 大谷貴史, 薮々堀正幹, 拓殖寛, 北研二: "HTML 形式の表構造の内容解析手法とその応用に関する研究", 情報処理学会研究報告, NL-154-20, pp. 137-144, 2003.
- [3] 嶋田 和孝, 遠藤 勉: "製品性能表からの特徴データの抽出", 情報処理学会 自然言語処理研究会 99-NL-133, 1999.
- [4] 佐藤慎哉, 山村 毅, 工藤博章, 松本哲也, 竹内義則, 大西 昇: "web ページ中のテキストと表からの重要箇所抽出", 情報処理学会自然言語処理研究会資料, NL153-9, pp.65-72, 2003.
- [5] Hsin-Hsi Chen, Shih-Chung Tsai, Jin-He Tsai: "Mining Tables from Large Scale HTML Texts", 18th International Conference. Computational Linguistics, pp.166-172, 2000.
- [6] Minoru Yoshida, Kentaro Torisawa, Jun'ichi Tsujii: "Extracting ontologies from World Wide Web via HTML tables", Pacific Association for Computational Linguistics, pp.332-341, 2001.
- [7] URL: <http://www.google.com/apis/>