自 然 言 語 処 理153－9
（２００３．１．２０）

webページ中のテキストと表からの重要個所抽出

佐藤　慎哉†　山村　毅‡　工藤　博章†　松本 哲也†　竹内 義則†　大西　昇†
†名古屋大学大学院工学研究科　‡愛知県立大学情報科学部

あらまし　本稿では、情報の信頼度を考慮して低品質なマルチドキュメントであるｗｅｂページ中のテキストと表から重要個所を抽出する手法について述べる。テキストや表に付けられた見出しをテキストや表の内容から抽出した重要個所との類似度で内容を評価してから抽出する重要個所を決めることにより、単純に表示上強調された個所を抽出したり、テキストの表層情報から重要個所を抽出する場合に比べ、より信頼度の高い重要個所が抽出できると考えられる。tf*idf、χ^2値を用いて重要個所を抽出した場合と本手法で用いた上位概念の出現頻度を用いて重要個所を抽出した場合の精度の比較から本手法の有効性を検証する。

Important part extraction from the text and table in a web page

Shinya Sato†　Tsuyoshi Yamamura‡　Hiroaki Kudo†　Tetsuya Matsumoto†
Yoshinori Takeuchi†　Noboru Ohnishi†
†　Nagoya University　‡Aichi Prefectural University
{sato,kudo,matumoto,takeuchi,ohnishi}@ohnishi.nuie.nagoya-u.ac.jp, yamamura@ist.aichi-pu.ac.jp

Abstract　This paper presents a method of extracting an important part from the text and table in a web page, which is a low quality multi-document.It considers information reliability and decides the important part by evaluating the title attached to the text or table with similarity of title and important parts extracted from text or table. We think that this method realizes important part extraction with high reliability, even from a low quality web page. Finally, we compare the result which extracted the important part using the frequency of dominant conception, tf*idf, and χ^2 value, and test the effectiveness of this method.

1.　はじめに

近年、日々増加する膨大な情報源の中から必要な情報をすばやく見つけることが困難になってきている。そのため多くの検索エンジンには、必要な情報をすばやく見つけるために多くの工夫がされている。それらは、キーワードを含む文を提示したり、２次キーワードを用いて検索結果を絞り込んだりするものである。確かに、キーワード情報は情報検索をする上でユーザに大変有益な情報を与える。しかし、低品質な情報があふれているwebページの検索ではあまり参考にならないことも多い。あるキーワードが含まれているからといって必ずしもそれに関連した内容のページであるとは限らない。そこで、webページに書かれている内容をキーワードに関連のある情報とともに提示できれば、より効率のよい検索ができると考えられる。そこで本稿では、webページ中のテキストや表の主題を推定する手法[1,2,3]について述べる。

2.　基本的な考え方

webページは、複数の話題について書かれていることが多いため、本手法ではページ中のテキストや表を表示上のまとまりに分割して[4]、それぞれのまとまりに対して個別に主題の推定を行う。ページの分割にはHTMLタグを利用する。各まとまりにHTMLタグを用いて見出しが付いているも

のもあるが、文字サイズ拡大など別の目的のために使われていることがあるなど、必ずしもそれがそのまとまりを正しく表しているわけではないので、見出しを直接抜き出して主題とするのは不適切な場合がある。そこで、単語の出現頻度などの表層情報を用いて重要個所を求め、それを見出しのように表示上強調されている個所と比べ、内容的にふさわしい方をページ中の各まとまりの主題とする。

3. テキスト要素の主題推定

ここでは、テキスト要素の主題の推定方法について述べる。まず始めにテキスト中に現れる名詞の出現頻度を求める。次に各名詞の上位概念の出現頻度を各名詞の出現頻度と単語間の類似度を考慮して求める。そして、各名詞の上位概念の出現頻度と助詞などの表層情報をもとにテキスト要素中の名詞句の評価値[3,5,6]を求め、評価値の高い複数の名詞句を抜き出す。最後に、テキスト要素に見出しが付いていればその見出しと抽出された名詞句との類似度[6]を求め、類似していたら見出しを、類似していなければ抽出した名詞句をテキスト要素の主題とする。

３．１ テキスト要素の名詞の出現頻度

形態素解析器「Chasen」[7]を用いてテキストを形態素解析して、テキストに含まれる名詞のうち接頭詞や接尾辞のようにそれ自体では意味をなさない名詞と「人」、「もの」のように抽象度が高い名詞を除く全ての名詞の出現頻度を求める。

３．２ テキスト要素の名詞の上概念の出現頻度

EDR のシソーラス辞書[8](tree 構造)を用いて 3.1 で求めた全ての名詞(葉)からのノードの距離が２以下の上位概念を求める。(1)式の w_{ij} は名詞 i と上位概念 j との類似度を表し、シソーラス辞書上での距離が近いほど１に近い値をとる。本手法では、(1)式により求めた類似度を用いて各名詞の頻度との重み付け和を(2)式により計算し、上位概念 j の出現頻度 c_j を求める。これは，複数の名詞に共通の上位概念で、より葉に近い概念ほど大きな値を持つ。

$$w_{ij} = \frac{d_j \times 2}{d_i + d_j} \tag{1}$$

$$c_j = \sum_{i=1}^{N} w_{ij} n_i \tag{2}$$

ここで、d_i、d_j は、それぞれ、名詞 i の *root* からの距離、上位概念 j の *root* からの距離である。また、N は、テキスト中の名詞の種類数、n_i は名詞 i の出現頻度である。なお、(1)で各名詞の上位概念となっていない上位概念との w_{ij} の値は 0 とする。

さて、見出しがついている場合は見出しに含まれる名詞(見出し語)の上位概念は内容を推定する上で重要な概念であると考えられる。そこで、(1)、(2)式から求まった上位概念 j の出現頻度 c_j に上位概念 j の見出し語からのノードの距離に応じた値(現在は 距離が１の概念：3、距離が 2 の概念：2、その他：１)をかけることにする。

３．３ 名詞句の評価値

ここでいう名詞句とは、「名古屋大学工学部」のように名詞が連続しているもの、「名古屋大学の学生」のように名詞(の連続)が助詞の「の」または「での」でつながったもの、「忙しい学生」のように名詞(の連続)に形容詞がついているもののことを指す。名詞句の評価値を(3)式によって求めることにする。

$$NP_i = W_i \sum_{j=1}^{M} \omega_j mc_j \tag{3}$$

W_i は名詞句 i に付いている助詞の種類による重み、ω_j は名詞句中での名詞の位置に対しての重み、

mc_jは名詞句に含まれる名詞jの上位概念のうち(2)で定義される頻度の最も高い上位概念の出現頻度、Mは名詞句に含まれる名詞数を表す。

W_iとしては、文中における助詞の働きを考慮して、例えば次の順序[6]で名詞句を重み付けする。

は/には ＞ が/も/だ/なら/こそ/です
＞ を/に/、/。 ＞ へ/で/から/より (4)

ω_jは、現在は全て1にしているが、接頭詞などでは小さく、名詞句の骨格となる最後の名詞（主辞）では大きくするといった使い方がある。

3．4　見出しとの類似度

3.3節で求めた最も高い名詞句の評価値の8割以上の評価値を持つ名詞句とそれらの名詞句と「AはBです（だ）。」の関係にある名詞句を抜き出し、次にこれらの名詞句と見出しとの類似度[6]を求める。これは、内容に関連した見出しが付いているかどうかを調べるためである。もし、類似度が高ければ見出しは内容に関連があると考え見出しをテキスト要素の主題とし、類似度が低ければ見出しは内容に関連がないとして名詞句を主題とする。ここで、類似度は以下のように計算するものとする。

$$sim1 = \frac{1}{M}\left(\sum_{i=1}^{M} \max_{1\le j\le N}(\omega_{ij})\right) \quad (5)$$

$$sim2 = \frac{1}{N}\left(\sum_{j=1}^{N} \max_{1\le i\le M}(\omega_{ij})\right) \quad (6)$$

$$\omega_{ij} = \frac{L_{ij}\times 2}{L_i + L_j} \quad (7)$$

上式で、Mは見出しに含まれる名詞数、Nは名詞句に含まれる名詞数、ω_{ij}は名詞n_iと名詞m_jの類似度でL_iは概念辞書中での名詞n_iの*root*からの距離、L_jは名詞n_jの*root*からの距離、L_{ij}は名詞n_iと名詞n_jの共通の概念で最も葉に近い概念の*root*からの距離を表す。(5)、(6)式のうち大きいほうの値が設定した閾値(現在は 0.75)以上なら見出しを主題とし、それ以外は抽出された名詞句で最も評価値が高い名詞句を主題とする。

3．5　実験結果・考察

実験にあたり、日本人大学生１２名にwebページ中の３０個所のテキストに対して内容を最も反映している個所を選んでもらった。以下に本手法での重要個所抽出の結果を示す。なお、見出し、助詞による重みは、これまでに経験的に使用してきた値を使用した。

表１：実験結果

最も多くの被験者が選んだ個所を抽出した数	22
その他の被験者が選んだ個所を抽出した数	3
被験者が選ばなかった個所を抽出した数	5

本手法では８3%のテキストから被験者と同じ個所が抽出できた。被験者と同じ個所が抽出できなかった原因としては、形態素解析に失敗したことや被験者が見出しを重要個所として選んだときに、テキストのほうに見出しに結びつく名詞句が存在しなかった場合(図1)などがあった。また、重要個所抽出に失敗した場合でも、重要語候補として被験者が選んだ個所を含んでいた場合が2件あった。

●「Bouquet Amitie」をより知っていただくためのコンテンツ

私のご挨拶から私のここまでの成り立ち、お花についての思いなどなど、このHPを末永く愛していただくために知っていただきたいことばかり。是非ご覧下さい。
メルマガについては、更新情報やお休みのご案内、新しいニュースや 季節のお花のコラムなどを掲載しています。もしよかったら登録してくださいね。

図１：被験者が選ばなかった個所を選んだ例[A]

次に、上位概念の出現頻度の代わりに、tf*idfとχ^2値を使ってテキスト中の名詞句の評価値を求めた場合の結果を示す。tf*idfとχ^2値[9]は以下の式により求めることができる。また、tf*idfとχ^2値の作成には毎日新聞95年版CD-ROMを使用した。

$$tf \times idf = n_i \times \ln\left(\frac{N}{l_i}+1\right) \quad (8)$$

$$\chi^2 = \sum_{j=1}^{n} \frac{\left(x_{ij} - m_{ij}\right)^2}{m_{ij}} \quad (9)$$

$$m_{ij} = \frac{\sum_{j=1}^{n} x_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}} \times \sum_{i=1}^{m} x_{ij} \quad (10)$$

n_i：名詞 i の出現頻度、N：全文献数、l_i：名詞 i の出てくる文献数、m：異なり単語数、n：文献数、x_{ij}：名詞 i の文献 j における頻度、m_{ij}：名詞 i の文献 j における理論度数

表 2：tf*idf を使った結果

最も多くの被験者が選んだ個所を抽出した数	22
その他の被験者が選んだ個所を抽出した数	2
被験者が選ばなかった個所を抽出した数	6

表 3：χ^2 値を使った結果

最も多くの被験者が選んだ個所を抽出した数	22
その他の被験者が選んだ個所を抽出した数	0
被験者が選ばなかった個所を抽出した数	8

tf*idf を使った場合は、上位概念の出現頻度を使った場合とそれほど結果は変わらなかったが、話し言葉やわざと表記をカタカナに変えた個所(「本当」を「ホント」など)があると、それを出現頻度の低い重要語として選んでしまう場合があった。このような書き方をしたwebページは少なくないので tf*idf を使う場合には、web ページを使って tf*idf を作成するなど工夫が必要になると考えられる。誤字に対しても同様のことが言えるので、低品質なwebページ中のテキストから重要個所を抽出するのには不向きかもしれない。また、重要個所候補として選ばれる名詞句に一貫性がなく、内容にそれほど関係のない名詞句が重要語候補として選ばれることもあった。上位概念を使う場合にはこのようなことはなかった。χ^2 値を使った場合は、被験者が選ばなかった個所を選ぶことが他の 2 手法に比べて多かった。これは、上位概念の頻度や tf*idf に比べ、ページ中での名詞の出現頻度の影響が少ないのが原因だと考えられる。この結果により、人がテキスト中で重要だと考える個所は、テキスト中での出現頻度の影響を受けていると考えられ、カタカナ表記や誤字に対して tf*idf よりも堅牢であることから、上位概念の出現頻度から重要個所を求めるのがwebページ中のテキストに対しては有効だと考えられる。

4. 表からの重要個所抽出

ここでは、表からの重要個所抽出 [10,11] について説明する。一般に、表には表中のセルの説明として比較対象と比較要素がそれぞれ表の 1 行目または 1 列目に書かれている。例えば図 2 では、1 行目の「水道水」、「電解還元水」などが比較対象で、1 列目の「PH」、「総アルカリ度」などが比較要素となる。比較対象はその表の見出しに関係があると考えられ、比較要素はその表の具体的な内容を知る上で非常に重要な情報であると考えられる。そこで本手法では、まず、表の比較対象に出てくる高頻出語と表に付けられた見だしから表のタイトルを評価・推定する。続いて、表の各行または列に含まれる高頻度語から比較要素を評価・推定し、得られた表のタイトルと比較要素をその表の主題とする。

4．1 処理手順

表を評価するための基本的な処理手順を次に示す。

1．表の 1 行目と 1 列目が表中のセルの説明になっているかどうかを判定
2．セルの説明になっているのが比較対象か比較要素かを決定
3．見出し、比較対象または表の内容から表のタイトルを推定

4．表中の各行または列の内容から比較要素を評価・推定

水道水、ミネラルウオーターの違い

	水道水	電解還元水	ミネラルウォーター
PH(ペーハー)	7.0前後	9～10	7.0前後
総アルカリ度	28	112	31
カルシウム	31.2	56.1	45.1
マグネシウム	5.8	7.8	6.8
カリウム	2.5	4.3	4.1
ナトリウム	6.0	7.5	6.2
塩素	23.4	7.1	59.1
酸化還元電位	＋553mV	－378mV	＋251mV
クラスター(水の分子構造)	117	58	108
浸透圧	中	高	中
溶解力	中	高	中
熱・電気伝導率	中	高	中
表面張力	高	低	高
厚生省医療効果認定	×	○	×

図 2：1 行目と 1 列目にセルの説明がある表[B]

4．2　セルの説明になっているかどうかの判定

web ページ中にある表の中には、表を単なる枠として利用しているもの(図 3)や、比較対象や比較要素の一方が省略された表(図 4)などがある。そこでまず、表の 1 行目と 1 列目が表中のセルの説明としてふさわしいかどうかを判定する必要がある。本手法では、文、画像、リンク、数字(+助数詞)は表中のセルの説明としてふさわしくないと考え、1 行目や 1 列目がこれらを含んでいた場合はセルの説明になっていないとする。

ねこさん	ねこ姫　写真の館5.27UP	掲示板　あかねこの集会
ぐれいさん	―――	あかい日記帳 everydayUP(maybe)
世界猫紀行7.27.UP	旅行の話　ロシア編	キリ番の部屋 NEW 10.31UP
猫の本棚	旅の本棚	リンク　あかねこの輪 NEW 11.22UP

図 3：表を枠として利用しているもの[C]

Aコース	○研修用定食 ○しゃぶしゃぶ ○山荘鍋(牛肉と野菜の味噌煮込み) ○明治鍋(ちゃんこ風寄せ鍋＜季節により内容がかわります＞)	9，800円
Bコース	○会席料理『桂』	10，900円
Cコース	○会席料理『扇』 ○ぼたん鍋(11月～4月)	12，000円
Dコース	○会席料理『錦』	13，200円

図 4：セルの説明の一方が省略されている場合[D]

4．3　セルの説明になっているのが比較対象か比較要素かの決定

表には、1 行目と 1 列目の両方にセルの説明が書かれている表(図 2)と片方が省略されている表(図 4)がある。ここでは、この 2 つの場合に共通した、セルの説明が比較対象か比較要素かを決定する方法について説明する(1 行目と 1 列目の両方にセルの説明がない場合は表を単なる枠と考え、各セルごとに 3 章の方法で重要個所の抽出を行う)。なお，以降の説明で主辞とは，名詞句の最後の名詞を表すものとし、助数詞に関してはそれから連想される名詞(例えば、助数詞「円」なら「料金」、「名」なら「人数」など)を主辞として割り当てるものとする。

判定方法

2 行 2 列以降の表中のセルを行方向もくしは列方向に見たとき、主辞の種類の少ない方向を比較対象とする。例えば、図 2 の表の場合、2－10 行目の主辞がいずれも 1 種類なので 1 行目が比較対象、1 列目が比較要素だと判定される。主辞の種類数が行方向と列方向で同じである場合は、1 行目のセル内の名詞句についての意味的なまとまりの数と 1 列目のセル内の名詞句についてのそれを求めて、少ない方を比較対象とする。例えば、図 5 の場合、1 列目の意味的なまとまりの数＜1 行目のまとまりの数なので、1 列目が比較対象、1 行目が比較要素と判定される。セルの説明が片方省略されている場合で、もし行方向と列方向で主辞の種類数が同じ場合(図 5 で 1 行目が省略されている場合など)は、セルの説明になっている個所(図 5 の 1 列目)に含まれるセル内の名詞句が意味的にまとまっているかどうかで比較対象か比較要素かを判定する。図 5 の場合、1 列目は意味的にまとまっているので比較対象となる。

6/2 宿泊料金(税別)	基本料金	インターネット割引 梅　標準プラン	竹	松
大人	8,800円	8,500円	9,500円	10,500円
小人	8,000円	7,600円	8,500円	9,500円
乳幼児	2,000円	1,500円	実費1,500円	実費1,500円
幼児食付	5,500円	5,500円	幼児食5,500円	幼児食5,500円
2名様1室1人	9,500円	8,800円	10,000円	11,000円

図 5：行方向と列方向で主辞の種類数が同じ場合[E]

４．４　表のタイトルの推定

比較対象中のセル内の名詞句の主辞の上位概念の中で、高頻度で比較対象中のセルに出てくる上位概念があれば、その上位概念を持つ主辞と見出しの類似度を 3.2.4 節の方法で求め、似ていれば見出しをその表のタイトルとする。主辞と見出しが似ていない場合や表に見出しがついていない場合は、「主辞の比較」をタイトルとする。例えば図 2 の表の場合、比較対象中に「水」に関する上位概念が高頻度で出てくるので、「水」に関する上位概念を持つ主辞「水」と見出し「水道水、ミネラルウォーターの違い」の類似度を計算すると似ていることが分かるので、見出しをそのままこの表のタイトルとする。もし図 2 に見出しがついていなければ、この表のタイトルを「水の比較」とする。また、比較対象中に高頻度で出てくる上位概念がなければ、表中でセルの説明になっていないセルに高頻度で現れる上位概念があれば、その上位概念を持つ主辞を使って上記と同様の操作を行う。例えば、図 5 でもし、比較対象中に高頻度で出てくる上位概念がなければ、その他のセル中に「～円」という語句が多く出てくるため、「料金」に関する上位概念が高頻度で出てくることになるので、主辞「料金」を使って見出しを「料金の比較」とする。それ以外の場合は表のタイトルはなしとする。

４．５　比較要素の評価・推定

ここでは、比較要素の評価・推定方法について説明する。比較要素の中には、セルの説明として不充分な場合(図 5 の 4、5 列目)や省略されている場合(図 4)がある。そのため、本手法では各比較要素が説明しているセルの内容から比較要素がセルの説明として十分であるかを判定する。また、省略されている場合はセルの説明としてふさわしい名詞を探す。まず、比較要素が説明している行または列のセル中に高頻度で出てくる上位概念を持つ主辞を探す。そして、その主辞と比較要素の類似度を 3.2.4 節の方法で求め、類似している場合は、比較要素はその行または列の内容をよく反映していると考え比較要素をそのまま比較要素とする。そうでなければ比較要素の説明は不十分と考え、比較要素にセルの主辞の情報を付加して新しい比較要素とする。図 5 の 2 列目は、比較要素「基本料金」がその列の内容「料金」を十分に反映していますのでそのまま「基本料金」を比較要素とし、4 列目は比較要素「竹」がその列の高頻度主辞「料金」を反映していませんので「竹(料金)」を新しい比較要素とします。また、図 4 のように比較要素が省略されている場合は、各行または列に高頻度で出てくる上位概念を持つ主辞を比較要素とします。図 4 の 2、3 列目の場合は、それぞれの高頻度主辞「料理」、「料金」が比較要素となる。各行または列から高頻度の主辞が求まらなかった場合は、その行または列の比較要素が、内容的に正しいかどうかの判定ができないので、その行または列の比較要素は表の主題としては不適切と考え、高頻度の主辞を用いて評価・推定された比較要素を優先的に主題に反映する。

４．６　実験結果・考察

以下に表からの主題推定の結果を示す。実験には web ページ中の表 50 個を使用した。表の主題の推定は「何を比較しているか」と「何について比較しているか」が主題に反映されている場合を正解とし、筆者自身が正誤を判定した。

表 4：表からの主題推定の結果

正解	24
「何を比較しているか」が推定できなかった場合	22
主題の推定がうまくいかなかった場合	4

主題の推定がうまくいかなかった場合には、表が沿革だった場合(図 6)や表が小さすぎる場合(図 7)があった。図 6、図 7 から推定された主題は、それぞれ「表(大正 12 年、昭和 23 年、昭和 36 年)」、「年数の比較(毎月(料金))」であった。図 7 のように表が小さすぎる場合は、これ以上の主題を推定するには情報不足だと考えられる。「何を比較しているか」が推定できなかった場合には、図 8 のようにお店や製品を比較している場合や図 9 のようにフレームとして表が利用されている場合などがあった。お店や製品を比較している表は、固有名詞や意味のまとまりのない名詞などが比較対象中に多く含まれるため、単純に上位概念の頻度を見るだけでは「何を比較しているか」が推定できなかった。従って、文献 11 のように名詞に店名、製品などの意味属性を持たせることで、それらが比較対象であっても「何を比較しているか」が推定できるようにするなどの工夫が必要だと考えられる。また、図 9 のようにスロットースロット値を一覧にした「フレーム」となっている表については、内容から「何のフレームか」を推定するのは困難である。なお、今回実験に使用したこのようなスロットタイプの表では、見出しが付いている場合は全て見出しがその表の説明になっていたので、無条件で見出しをその表のタイトルとしても問題はないと考えられる。また、「何を比較しているか」が推定できなかった場合の中には、「高頻度主辞の比較」を表のタイトルとするのでは不充分な場合(図 10)もあった。図 10 の表からは、「表：デーの比較(割引券種、割引料金、発売日等)」が主題として求まるが、「デーの比較」では何を比較しているのか分かりづらい。この点も今後検討する必要がある。

大正12年(1923)	小野正明氏により東京都江戸川区平井にて肉屋として創業。
昭和23年(1948)	有限会社「花　正」を設立。
昭和36年(1961)	平井本店を改装、スーパーマーケット方式に変更。
昭和51年(1976)	千葉県幕張に「千葉工場」開設。

図 6：沿革を表にしている場合[F]

	3年	5年	10年
毎　月(100万円)	58，000円	36，000円	19，000円

図 7：小さい表[G]

店名	お薦めラーメン	特徴	所在地
カノ亭	ゴマ味ラーメン	あっさり味のスープがうまい。お薦め！お店は、ホテル「ヘリティジ」(江南町)から関越東松山インター方面に向かって500メートル位。650円	埼玉県比企郡滑川町
	しょうゆラーメン	これをしょうゆと言って良いのか？うまい！チャーシューも手作りで他にはないうまさ。(写真を見て下さい)500円	
	ネギラーメン	ママはいつもこればっかり。飽きないとのこと。600円	
初代	醤油ラーメン	テレビチャンピオンで見たとおりのマスターを発見。店内は込んでいました。チャーシューが柔らかく、スープもうまい。麺は黄色い縮れ麺でした。バランスがとれてます。閉店時間が早いので注意！(14:00だったかな？)味噌もうまかったです。	小樽市
五丈原	トン塩ラーメン	スープと言い、麺と言い私の食べたラーメンの中でNO.1チャーシューおにぎりも最高！	札幌市
	しょうゆラーメン	あっさり味、これもうまい！旭川ラーメンのファンになりました。	

図 8：お店の比較をしている表[H]

ご利用条件

ご利用いただける方		友の会会員かつ以下のどちらかの条件を充たす方 ①年金受取口座指定者 ②100万円以上の定期性預金預け入れ者
金利	固定金利	年利4．375％
ご融資額		300万円以内
お使いみち		マイカー関連資金、旅行資金、墓地購入および建設(修繕含む)資金
ご返済期間	固定金利	10年以内　※但し、完済時年齢は76歳以下
ご返済方法		毎月払い、隔月払い(2ヶ月サイクル：年金受給月)
担保・保証人		担保は不要。 ＊個人連帯保証人1名以上必要となります。
必要書類	本人	本人確認書類、年収確認書類、資金使途確認書類。 その他の必要書類。
	保証人	本人確認書類、年収確認書類、勤続年数確認書類

図 9：スロットタイプの表[G]

名称等	割引券種		割引金(円)	発売日等
シーズンイン割引	大人	一日券	2，500	シーズンイン～平成14年12月27日まで
	小児	一日券	1，100	
スキーキッズデー	小児	一日券	無料	12/15.22。 1/19.26。 2/16.23。 3/16.23。の毎月第3・4日曜日
レディースデー	大人	一日券	3，500	1/6～3/7までの毎週月曜日。女性のみ
岩岳感謝デー	大人	一日券	3，000	1/10,17,31。,2/7,21,28。3/7,14.28(特定日を除いた金曜日)
	小児	一日券	1，100	
岩岳感謝祭	大人	一日券	2，250	2/14　岩岳感謝祭抽選会、振舞い酒、他各種イベント
	小児	一日券	1，100	
岩岳木曜デー	大人	一日券	3，000	1/9,16,30。 2/6,13,20,27。3/6,13.20.27(特定日を除いた木曜日)
	小児	一日券	1，100	
白馬サンクスデー	大人	一日券	2，250	平成15年1月23，24日
	小児	一日券	1，100	
シーズンアウト割引き	大人	一日券	3，500	平成15年3月24日～3月31日
	小児	一日券	1，100	

表 10：本手法で推定した表のタイトルが何を比較しているかを十分に表していない表[I]

5. おわりに

本稿では、web ページ中のテキストと表から主題を推定する方法とその結果について述べた。テキストからの重要個所抽出については、さらに多くのページに対して実験を行い、本手法の有効性を検討する必要がある。表の主題の推定については、より精度よく比較対象から表のタイトルが推定できる手法と、精度を評価する方法を検討する必要がある。

6. 参考文献

1) 岡満美子,小山剛弘,上田良寛,句表現要約の句合成手法,自然言語処理,No.129,pp.101-108,1999

2) 吉見毅彦,奥西稔幸,山路孝浩,福持陽士,表題へのつながりに基づく文の重要度評価,自然言語処理 jan Vol.6 No.1 pp.43-57 1999

3) 奥村学,難波英嗣,テキスト自動要約に関する研究動向,自然言語処理 july Vol.6 No.6 pp.1-26, 1999

4) 山田洋志,福島俊一,松田勝志,web ページからのタイプ別情報抽出・分類方式,自然言語処理 No.136,pp.143-150 2000

5) K.Zechner, Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences,Proc.16th International Conference on Conputational Linguistics vol.2 pp.986-989 1996

6) 長尾真,岩波講座　ソフトウエア科学 15　自然言語処理,岩波書店

7) 茶筌:http://chasen.aistnara.ac.jp/index.html.ja

8) EDR 電子化辞書,概念辞書,CPD-V020.1,株式会社日本電子辞書研究所

9) 長尾真,画像と言語の認識工学,コロナ社,1989

10) 島田和孝,遠藤勉,特徴化された表データからの要約文生成処理, 電子情報通信学会技術研究報告,Vol.99,No487,TL-99,pp.25-31,1999

11) 河合敦夫,塚本雄之,山本勝紀,椎野努,文書構造を利用した箇条書きや表形式文書からの内容抽出,電子情報通信学会論文誌 D-Ⅱ Vol.J81 No.7 pp.1609-1620

7. 参照ページ

A) http://www2.odn.ne.jp/amitie/

B) http://www2u.biglobe.ne.jp/~mizu-oka/treatment.htm

C) http://homepage2.nifty.com/~akaneko/

D) http://www.ikoinoie.com/minoo/price/index.html

E) http://www.page.sannet.ne.jp/taihei/ryoukin.html

F) http://www.coe.co.jp/i/hanamasa/history.html

G) http://www.okinawa-rokin.or.jp/SecondLife.htm

H) http://member.nifty.ne.jp/maichanpapa/ramen.htm

I) http://www.f4.dion.ne.jp/~aburaya/saisinjouhou.htm